People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of Oran 2
Institute of Maintenance and Industrial Safety

DISSERTAION

For the award of the Master's degree
In Industrial Safety and Environment

The Use of Data Analysis to Enhance Safety

Realized by:

Terbeche Abderrahmane

Sihamdi Hakim

ZOUAIRI SAIM	MAA	IMSI	Supervisor
TITAH MOULOUD	MCB	IMSI	PRESIDENT
BELKHOUDJA LEILA	MCB	IMSI	EXAMINATOR

29 Septembre 2022

Acknowledgements

First of all, we would like to thank Almighty God for giving us the encouragement to complete our studies.

This endeavor would not have been possible without the help and support of our parents and families.

Many thanks to our supervisor Dr. ZOUAIRI Saim, and all the teachers who helped us in our educational journey.

We are also grateful to my classmates and our dear friends, and to the Kaggle community.

I would like to acknowledge and dedicate this work to my beloved mother "Ms. Ben Zahia Khaira" رحمها الله وأسكنها فسيح جنانه

— *Sihamdi Hakim*

Lastly, I would be remiss in not mentioning my second family "Algerian Positive Vibes", their support and inspiration have kept my spirits and motivation high during this process.

— *Terbeche Abderrahmane*

Abstract

Data analysis is a process that consists of several stages and steps that take us from the information we have to generate insights, and these insights help us make the right decisions.

To most people, data analysis may appear to be an area of interest only to those in this science. However, in fact, there is a role for all people, and there is certainly a large role for occupational safety and health professionals.

In recent years, data analysis has become one of the most important tools in occupational safety and health, due to the technological development that touched all fields and industries, which contributed to the availability of data. Nevertheless, the problem remains that we are not making the best use of it in order to improve safety levels, which separates the leading companies from the others.

Unfortunately, many companies and industries do not give much value to the use of data analysis in occupational safety and health, because either they are ignorant of the field, or because they see that the process takes considerable time and resources. Here comes the role of occupational safety and health leaders to spread awareness and the right safety culture.

Our work aims to highlight the use of data analysis in the field of occupational safety and health in order to improve safety levels. On our way, we will learn about these two areas and the most prominent concepts related to them. We also touched on the two most important types of data analysis, which are exploratory data analysis, and predictive data analysis.

In the end, we conducted a case study that illustrates the process, and we concluded by mentioning the obstacles we faced and providing suggestions and recommendations in order to facilitate matters for occupational safety and health specialists so that they approach the goal that we all seek, which is to maintain the safety of employees.

Keywords: Occupational safety and health, Data analysis, Exploratory data analysis, Predictive data analysis, The use of data analysis to enhance safety.

Abstract

تحليل البيانات هو عملية تتكون من عدة مراحل وخطوات تأخذنا من المعلومة التي نملكها إلى توليد رؤى، وهذه الرؤى تساعدنا على اتخاذ القرارات الصحيحة.

لدى عامة الناس قد يبدو تحليل البيانات على أنه مجال يهتم به فقط المختصون في هذا العلم. لكن في الواقع هنالك دور لكافة الناس، وبالتأكيد هنالك دور كبير لمختصي السلامة والصحة المهنية.

في السنوات الأخيرة أصبح تحليل البيانات من أهم الأدوات في السلامة والصحة المهنية بسبب التطور التكنولوجي الذي مس كافة الميادين والصناعات، مما أسهم في توفر البيانات. مع ذلك يبقى المشكل هو أننا لا نستغلها أحسن استغلال من أجل تحسين مستويات السلامة، وهو الأمر الذي يفصل بين الشركات الرائدة عن غيرها.

مع الأسف، العديد من الشركات والمؤسسات لا تعطي قيمة كبيرة لاستعمال تحليل البيانات في السلامة والصحة المهنية، إما لجهلها أصلا بالمجال، أو لأنها ترى بأن العملية تأخذ وقتا ومواردا معتبرة. وهنا يأتي دور رواد السلامة والصحة المهنية من أجل نشر الوعي وثقافة السلامة الصحيحة.

عملنا هذا يهدف إلى تسليط الضوء على استخدام تحليل البيانات في مجال السلامة والصحة المهنية من أجل تحسين مستويات السلامة. في طريق استعراض هذين المجالين وأبرز المفاهيم المتعلقة بهما، كما تطرقنا إلى أهم نوعين من تحليل البيانات، وهما تحليل البيانات الاستكشافي، وتحليل البيانات التنبئي.

في النهاية قمنا بدراسة حالة توضح العملية، وختمنا بذكر العقبات التي واجهتنا وتقديم اقتراحات وتوصيات من أجل تسهيل الأمور على مختصي السلامة والصحة المهنية، حتى يقتربوا من الهدف الذي نسعى إليه جميعا، ألا وهو الحفاظ على سلامة الموظفين.

الكلمات المفتاحية: السلامة والصحة المهنية، تحليل البيانات، تحليل البيانات الاستكشافي، تحليل البيانات التنبئي، استعمال تحليل البيانات من أجل تحسين مستويات السلامة.

Abbreviations

OSH	Occupational safety and health
PPE	Personal protective equipment
SIP	Strategic Improvement Process
ASSE	the American Society of Safety Engineers
LTI	Lost Time Injury
EDA	Exploratory data analysis
NLP	Natural Language Processing
HR	Human Resources

Figures Table

Table of contents

Table of contents

Table of contents

Chapter I General Introduction

Many industries and business functions are taking advantage of their "big data" sets by performing advanced analytics to make predictions about the future. When applied correctly, data analytics allows leaders to gain deep insight into their business and deploy their scarce resources in an optimal way.

Advanced and predictive analytics has revolutionized many industries. From biotechnology and mapping of the genome, to banking and market research, and is the foundation of Internet search engines such as Google search.

Data analytics is now also available to safety professionals to predict and prevent workplace injuries.

This thesis reviews the use of data analytics methods, especially exploratory and predictive analysis to enhance occupational safety and health performance, and outlines the safety inspection data used to fuel the predictive models (leading indicators), and why this type of data is preferred over other safety data (lagging indicators). It also includes a case study for manufacturing plants database using Python.

1. Problematic

In the light of the dominance and growth of the role of technology in every sector, it generates huge amounts of information that can yield valuable insights into the field of occupational safety and health.

Big data and analytics have shown promise in predicting safety incidents and identifying preventative measures directed towards specific risk variables.

However, the safety industry is lagging in big data utilization due to various obstacles, including a lack of data readiness (e.g., disparate databases, missing data, low validity) and personnel competencies. In addition, a large part of the available data is not exploited in order to obtain insights about decision-making and precautions, which may lead to many avoidable losses for businesses and industries.

Our goal with this study is to find ways that we can exploit the available data in order to improve the safety and health levels of companies and industries.

How can we achieve it?

By studying and analyzing the data using the appropriate methods, which we will address in the coming chapters.

What are the benefits of this study?

- Avoid and reduce the proportion of material and human losses for businesses and industries.
- Make timely precautionary decisions, based on precise studies.
- Make informed, data-driven safety decisions.
- Gain knowledge in the field of data analysis, which has become a requirement for occupational safety engineers.

Chapter II Occupational safety and health

IN THIS CHAPTER

- ✓ Discovering the occupational safety and health field and its importance.
- ✓ Who are OSH professionals, where do they work? What are their roles and expertise?
- ✓ OSH professionals and anticipation.
- ✓ Getting to know the definition of safety and some related terms.

1. Introduction

To know where you are going, you need to know where you have been; going back in the history of occupational safety and health there are two main milestones.

Before the Industrial Revolution began in 1760, it was the norm to make a living through agriculture or the making and selling of products from home. With new developments in machinery and manufacturing processes, Britain, followed by parts of Europe and the US, began moving towards a society fuelled by mass production and the factory system. Unfortunately, working conditions in the 1800s were poor...

During this movement, Workers formed unions and began to demand better working conditions. Government organizations responded by regulating the workplace and forcing safer work practices.

Figure 1 vintage empire state building construction - photo by Lewis Wickes Hine

Like the Industrial Revolution, the digital revolution has changed where and how people and industries work.

The field of occupational health and safety, in turn, kept pace with these changes by introducing modern sciences and technologies such as data analysis, machine learning, and artificial intelligence...

Rising global access to the internet and rapid developments in technology have made managing health and safety in the workplace more efficient than ever before. Health and safety systems like smartphone apps are revolutionizing how we keep staff safe, especially those who work alone. In fact, according to a new research report from Berg Insight, the number of lone workers using connected safety solutions in Europe and North America is expected to reach 1.8 million by 2025. (Don Cameron, 2020)

2. Definition

Occupational safety and health (OSH) is generally defined as the science of the anticipation, recognition, evaluation, and control of hazards arising in or from the workplace that could impair the health and well-being of workers, taking into account the possible impact on the surrounding communities and the general environment. This domain is necessarily vast, encompassing a large number of disciplines and numerous workplace and environmental hazards. A wide range of structures, skills, knowledge and analytical capacities are needed to coordinate and implement all of the "building blocks" that make up national OSH systems so that protection is extended to both workers and the environment. (Benjamin O. Alli, 2002)

And according to (Robyn Correll, 2022) occupational health and safety is the field of public health that studies trends in illnesses and injuries in the worker population and proposes and implements strategies and regulations to prevent them. Its scope is broad, encompassing a wide variety of disciplines—from toxicology and epidemiology to ergonomics and violence prevention.

3. Why is occupational safety and health needed?

Usually, as (Reese, 2017) indicated in his book, consequences of not addressing safety and health have to do with why OSH is addressed by employers. These consequences include but are not limited to the following:

- Injury or illness to members of the workforce
- Loss of profit
- Loss of credibility as a responsible company
- Liability of not addressing safety and health
- Loss of productivity
- Loss of employees due to danger, risk, injury, illness, death, or unsafe/unhealthy work environment
- Decrease of employee morale
- Damage to or loss of capital investment (e.g., equipment or facilities)
- Decrease in reputation and integrity of the company
- View that the company does not exhibit good business practices

4. Occupational safety and health professionals

In general terms, an occupational safety and health (OSH) professional is there to protect workers from harm and prevent damage to equipment, property, the environment, and the public. This is accomplished by analysis, design, and implementation of programs to prevent occupationally related injuries and illnesses. Many safety and health professional specialize in specific areas, with expertise in engineering, industrial hygiene, system safety, loss control, and ergonomics, while others handle all facets of OSH. These professionals may work for the public sector, private sector, or government, or as

consultants. They are found in a host of industry sectors such as mining, military, service industries, construction, hazardous waste, chemical handling and processing, manufacturing, insurance, transportation, long shoring, agriculture, energy source production, research and development, and a multitude of other domains. (Reese, 2017)

5. Why is it important to anticipate?

In 1996, the American Society of Safety Engineers (ASSE) published a pamphlet entitled Scope and Functions of the Professional Safety Position. This pamphlet provides a superb presentation of why a safety and health professional is essential to the arena of OSH by explaining why the safety and health professional must anticipate, recognize, evaluate, control, and communicate safety and health as their main function.

Why is it important to anticipate? The safety and health professional needs to develop methods for anticipating and predicting hazards based on their experience, historical data, and other pertinent sources of information. This will allow for identifying and recognizing hazards in existing and future systems, equipment, products, software, facilities, processes, operations, and procedures during the life expectancy of these various facets of the workplace.

As part of this, they must evaluate and assess the probability and severity of loss events and accidents/incidents that may result from the actual or potential hazards.

This requires applying these methods and conducting hazard analyses and interpreting results.

The analysis and interpretation are accomplished by reviewing, with the assistance of specialists where needed, entire systems, processes, and operations, and any subsystems or components, for failure modes, causes, and effects, due to the following:

- System, subsystem, or component failures.
- Human error.
- Incomplete or faulty decision-making, judgments, or administrative actions.
- Weaknesses in proposed or existing policies, directives, objectives, or practices.

It is important to review, compile, analyze, and interpret data from accident and loss event reports and other sources regarding injuries, illnesses, property damage, environmental effects, or public impacts. This is why a safety and health professional must identify causes, trends, and relationships to ensure completeness, accuracy, and validity of required information; evaluate the effectiveness of classification schemes and data collection methods, and initiate investigations.

The duty of the safety and health professional is to provide advice and counsel about compliance with safety, health, and environmental laws, codes, regulations, and standards, including conducting research studies of existing or potential safety and health problems and issues.

It is also expected that he/she will determine the need for surveys and appraisals that help identify conditions or practices affecting safety and health. including those that require the services of specialists, such as physicians, health physicists, industrial hygienists, fire protection engineers, design and process engineers, ergonomists, risk managers, environmental professionals, psychologists, and others while assessing environments, tasks, and other elements to ensure that physiological and psychological capabilities, capacities, and limits of humans are not exceeded.

6. Safety

1. Definition

It is the state of being safe, the conditions of being protected from health, economic, and physical losses, damage accident, error, harm, or other types of consequences of failure.

Safety can also be defined to be the control of recognized hazards to achieve an acceptable risk level.

2. Hazards

According to (Reese, 2017) hazards are defined as a source of danger that could result in a chance event such as an accident/incident. A danger itself is a potential exposure or a liability to injury, pain, or loss. Not all hazards and dangers are the same. Exposure to hazards may be dangerous, but this is dependent upon the amount of risk that accompanies it. The risk of water contained by a dam is different compared to being caught in a small boat in rapidly flowing water.

Risk is the possibility of loss or injury/illness or the degree of the possibility of such loss. Incidents do not occur if a hazard does not exist that presents a danger to those working around it. If the potential exposure is high, there is a greater risk that an undesired event will occur.

3. Energy

Energy is classified in one of two ways. It is either potential or kinetic energy. Potential energy is defined as stored energy such as a rock on the top of a hill. There are usually two components to potential energy. They are the weight of the object and its height above another surface. The rock resting at the bottom of the hill has little potential energy as compared to the one at the top of the hill.

The other classification is kinetic energy, which is best described as energy in motion. Kinetic energy is dependent upon the mass of the object. Kinetic energy is a function of an object’s mass and its speed of movement or velocity.

Chapter III Data analysis

IN THIS CHAPTER

- ✓ Discovering the data analysis science and the two main types of data.
- ✓ Exploring the data analysis process.
- ✓ Getting to know the types of data analysis.
- ✓ Addressing the importance of data analysis to the OSH field.
- ✓ Getting to know the big data and its three Vs.

1. Definition

Data analysis is the process of collecting, modeling, and analyzing data to extract insights that support decision-making. There are several methods and techniques to perform analysis depending on the industry and the aim of the investigation.

All these various methods are largely based on two core areas: *quantitative* and *qualitative research.*

2. The difference between quantitative and qualitative data

When it comes to conducting research and data analysis, you will work with two types of data: quantitative and qualitative. Each requires different collection and analysis methods, so it is important to understand the difference between the two.

Generally speaking, quantitative analysis involves looking at the hard data, the actual numbers. Qualitative analysis is less tangible. It concerns subjective characteristics and opinions – things that cannot be expressed as a number.

What is quantitative data?

Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it is quantitative data. Quantitative data can tell you "how many," "how much," or "how often" — for example, how many workplace incidents occurred last month?

To analyze and make sense of quantitative data, you will conduct statistical analyses.

What is qualitative data?

Unlike quantitative data, qualitative data cannot be measured or counted. It is descriptive, expressed in terms of language rather than numerical values.

Researchers will often turn to qualitative data to answer "Why?" or "How?" questions. For example, if your quantitative data tells you that workers are not using their PPE, you'd probably want to investigate why—and this might involve collecting some form of qualitative data from them.

3. Data analysis process

As shown in Figure 2, when we talk about analyzing data, there is an order to follow in order to extract the needed conclusions. The analysis process consists of five key stages.

Here is a rundown of the five essential steps of data analysis according to (Bernardita Calzon, 2022):

- **Identify:**

Before we get our hands dirty with data, we first need to identify why we need it in the first place. The identification is the stage in which we establish the questions we will need to answer. For example, what is the site most at fire hazard in our factory? Once the questions are outlined, you are ready for the next step.

- **Collect:**

As its name suggests, this is the stage where we start collecting the needed data. Here, we define which sources of information we will use and how we will use them. The collection of data can come in different forms such as internal or external sources, surveys, interviews, and questionnaires, among others. An important note here is that the way we collect the information will be different in a quantitative and qualitative scenario.

- **Clean:**

Once we have the necessary data it is time to clean it and leave it ready for analysis. Not all the data we collect will be useful, when collecting big amounts of information in different formats it is very likely that we will find ourselves with duplicate or badly formatted data. To avoid this, before we start working with our data we need to make sure to erase any white spaces, duplicate records, or formatting errors. This way we avoid hurting our analysis with incorrect data.

- **Analyze:**

With the help of various techniques such as statistical analysis, regressions, neural networks, text analysis, and more, we can start analyzing and manipulating our data to extract relevant conclusions. At this stage, we find trends, correlations, variations, and patterns that can help us answer the questions we first thought of in the identify stage. Various technologies in the market assist researchers and safety managers and engineers with the management of their data. Some of them include business intelligence and visualization software, predictive analytics, and data mining, among others.

- **Interpret:**

Last but not least, we have one of the most important steps: it is time to interpret our results. This stage is where the researcher comes up with courses of action based on the findings. For example, here we would understand why the workers of a sector do not use their PPE, or why the number of lost time injuries is high in the last 3 months. Additionally, at this stage, we can also find some limitations and work on them.

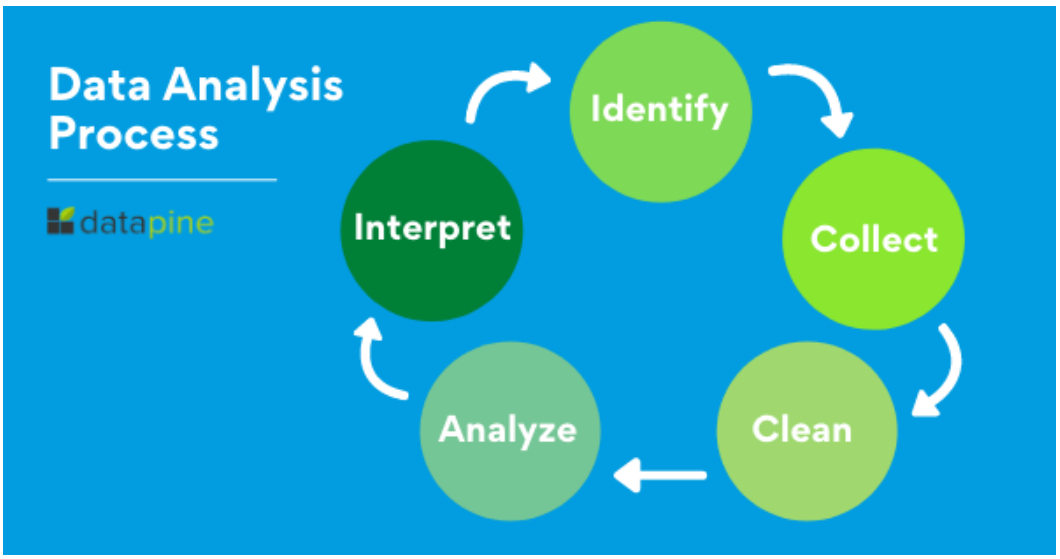

Figure 2 Data analysis process

4. Types of data analysis

Definitions from (Bernardita Calzon, 2022)

1. Descriptive analysis - What happened?

The descriptive analysis method is the starting point to any analytic reflection, and it aims to answer the question of what happened? It does this by ordering, manipulating, and interpreting raw data from various sources to turn it into valuable insights for your organization.

Performing descriptive analysis is essential, as it allows us to present our insights in a meaningful way. Although it is relevant to mention that this analysis on its own will not allow you to predict future outcomes or tell you the answer to questions like why something happened, *it will leave your data organized and ready to conduct further investigations.*

2. Exploratory analysis - How to explore data relationships?

As its name suggests, the main aim of the exploratory analysis is to explore. Prior to it, there was still no notion of the relationship between the data and the variables. Once the data is investigated, the exploratory analysis enables us to find connections and generate hypotheses and solutions for specific problems. A typical area of application for it is data mining.

3. Diagnostic analysis - Why it happened?

Diagnostic data analytics empowers analysts and executives by helping them gain a firm contextual understanding of *why something happened.* If you know why something happened as well as *how it happened,* we will be able to pinpoint the exact ways of tackling the issue or challenge.

Designed to provide direct and actionable answers to specific questions, this is one of the world's most important methods in research, among its other key organizational functions such as retail analytics.

4. Predictive analysis - What will happen?

The predictive method allows us to look into the future to answer the question: *what will happen?* In order to do this, it uses the results of the previously mentioned descriptive, exploratory, and diagnostic analysis, in addition to machine learning (ML) and artificial intelligence (AI). Like this, we can uncover future trends, potential problems or inefficiencies, connections, and casualties in our data.

With predictive analysis, we can unfold and develop initiatives that will not only enhance our various operational processes but also help us gain an all-important edge on the competition. If we understand why a trend, pattern, or event happened through data, we will be able to develop an informed projection of how things may unfold in particular areas of the business.

5. Prescriptive analysis - How will it happen?

Another of the most effective types of analysis methods in research. Prescriptive data techniques cross over from predictive analysis in the way that it revolves around using patterns or trends to develop responsive, practical business strategies.

5. Why do we need analysis?

According to (Faisal Osama, 2020), the answer to this question boils down to four main points:

1. Take correct decisions

- Training needs analysis.
- Change HR policy.
- Incentive and violation scheme.
- Change in inspection frequency.
- Change in monitoring method.
- Change in work method.
- Change in the management system.
- Prioritizing resources.
- PPE selection.

2. Make studies

- Epidemiology study.
- Toxicology study.
- Risk assessments.

3. Identify root causes

- Accidents.
- Poor behavior.
- Failure in the management system.

4. Calculate probabilities

- Accident occurrence.
- Absence.
- Sickness.
- Failure of equipmentent.

6. Big data

1. Definition

Big data has made its way into virtually every aspect of our lives, and that now includes the workplace. When it comes to occupational safety, big data can improve safety to create a healthier work environment for employees. Big data can give us insight into what we are doing wrong so that we can take steps to reduce injuries and illnesses in the workplace.

The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. It is an accumulation of data that is too large and complex for processing by traditional database management tools.

2. The three Vs of big data

Definitions from (Oracle.com, 2022)

Volume

The amount of data matters. With big data, you will have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.

Velocity

Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real-time or near real-time and will require real-time evaluation and action.

Variety

Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semi-structured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

Chapter IV Safety analysis

IN THIS CHAPTER

- ✓ Exploring the safety analysis procedure and some related concepts.
- ✓ Discovering the leading and Lagging Indicators of Safety Performance.
- ✓ Exploring safety data sources.

1. Definition

According to (Harms-Ringdahl, 2001), Safety analysis is a systematic procedure for analyzing systems to identify and evaluate hazards and safety characteristics.

This definition is wide, and it includes both qualitative and quantitative methods.

1. Data rich, time limited information

The objective of the safety analysis process is to extract useful information from the stored safety data and then display this information using graphs, tables, dashboards, and presentations so that managers can make informed decisions on safety.

Imagine that our safety database contains a large number of reports, which should be complete, correct, and appropriately classified. This large amount of data is valid for a time-limited period, and it is important that we are able to determine useful safety information quickly and then present it to support data-driven decision-making purposes.

The process of turning safety data into useful safety information is achieved by using various statistical and technical analysis methods.

A well-developed analysis process will help organizations to:

- Refine safety objectives.
- Establish effective safety performance targets.
- Establish effective safety performance indicators.
- Alert safety decision-makers, based on safety triggers.
- Establish safety presentation capabilities (e.g. safety dashboard) for ready interpretation of safety information by decision-makers.
- Monitor safety performance of a given department, system, or process.
- Highlight safety trends.
- Identify factors that cause change.
- Identify correlations between or among various factors.
- Test assumptions.
- Develop predictive modeling capabilities.

(*Information is taken from ICAO Document 9859)*

2. Risk analysis

Within the area of dependability and reliability, there is an international standard (IEC, 1995) that defines "risk analysis" and a number of related terms.

According to this standard: Risk analysis is the systematic use of available information to identify hazards and to estimate the risk to individuals or populations, property or the environment.

In the standard, risk is defined as a combination of the frequency, or probability, of occurrence and the consequence of a specified hazardous event. Risk analysis is also

sometimes referred to as probabilistic safety analysis (PSA), probabilistic risk analysis (PRA), quantitative safety analysis, and quantitative risk analysis (QRA).

3. The systematic approach

One of the keywords in the definition of safety analysis presented above is "systematic". If an analysis is of good quality, it is essential to consider the points below.

Let us suppose that a particular production system is to be analyzed. The analysis might apply to existing installation or production facilities still at the planning stage. According to (Harms-Ringdahl, 2001) there are several different aspects to a systematic approach:

- A general procedure for the analysis is defined.
- Gathering of information on the system provides the basis for the analysis and must be carried out systematically.
- The entire system and the activities within it should be included in the analysis. The analysis needs to be designed so that important elements are not overlooked. The main thread must be identified and followed.
- A systematic specified methodology is required for the identification of hazards.
- The risks to which these hazards give rise need to be assessed in a consistent manner.
- A systematic approach is required when safety proposals are to be generated and evaluated.

4. Safety analysis procedure

A safety analysis consists of a number of coordinated steps, which jointly make up a procedure.

These steps are according to the "Safety Analysis, Principles, and practice in occupational safety" book by (Harms-Ringdahl, 2001), it might be a slight difference in other sources.

1. Introductory part of the analysis

Plan

One of the first steps is to take the decision to conduct a safety analysis. This involves consideration of:

- What is to be analyzed, what limits to the analysis are to be set, and what assumptions are to be made.
- The aim of the analysis. This might be finding ways to increase the level of safety, or a general evaluation of safety. In the latter case, the stage "Proposals for safety measures" disappears from the analysis.
- Choice of methods and manner of approach.

Gather information

Information on the system to be analyzed is needed. This applies to its technical design, how the system functions, and which activities are undertaken. To a great extent, the need for information is governed by the choice of methods to be employed.

Other useful information may concern accidents that have occurred, near accidents and disturbances to production. If probabilistic analyses are to be conducted, data on frequencies of failure for the components used in the system are also needed.

In the cases of analyses of installations that have been in operation for some time, information is relatively easily accessible. When the analysis concerns production facilities that are still at the planning stage, it is more difficult. Information can then be obtained from drawings, written and oral descriptions, and from experiences of similar installations.

2. Central parts of safety analysis

Identify hazards

The central component of most safety analyses is the identification of hazards and other factors in the system that might lead to accidents. One aim should be to discover the major sources of danger and which factors might trigger off an accident.

The method selected determines how the process of hazard identification proceeds. When a specialized method is used, certain types of hazards are discovered, but others may be overlooked.

Assess risks

An assessment is made of risks in the system. Such assessments can take different forms. One application of risk assessment is to judge whether a system is safe enough, or if safety measures are necessary.

In a quantitative analysis, values for probabilities and consequences are estimated. In the case of qualitative analysis, an evaluation is made without numeric values.

Propose safety measures

If needed, risks can be reduced through one or several safety measures. The reduction can apply to either consequences or the probability that such negative events will occur. Some of the methods include a systematic procedure for the identification of potential safety measures.

Supplementary analysis

In the course of conducting the analysis, it might be discovered that more detailed examinations are required, or that a supplementary method is appropriate.

Summarize

The results of an analysis are summarized so as to provide a basis for decision-making. The summary might include a list of the hazards observed, proposals for safety measures, and an account of the assumptions and conditions under which the analysis was conducted. This will finish the analysis.

3. After the analysis

Make decisions

We assume that the summary is then used as a basis for decision-making. Usually this is not a part of the analysis, and decisions are made somewhere else in the company.

Implement safety measures

For an analysis to have an effect, the safety measures decided upon must, of course, be implemented.

Follow up the analysis

It is also a good idea to make plans to follow up the analysis. This can involve making certain checks on the analysis, establishing that measures have been implemented, and examining results at a later date.

For example: Has the number of accidents fallen? How has production been affected?

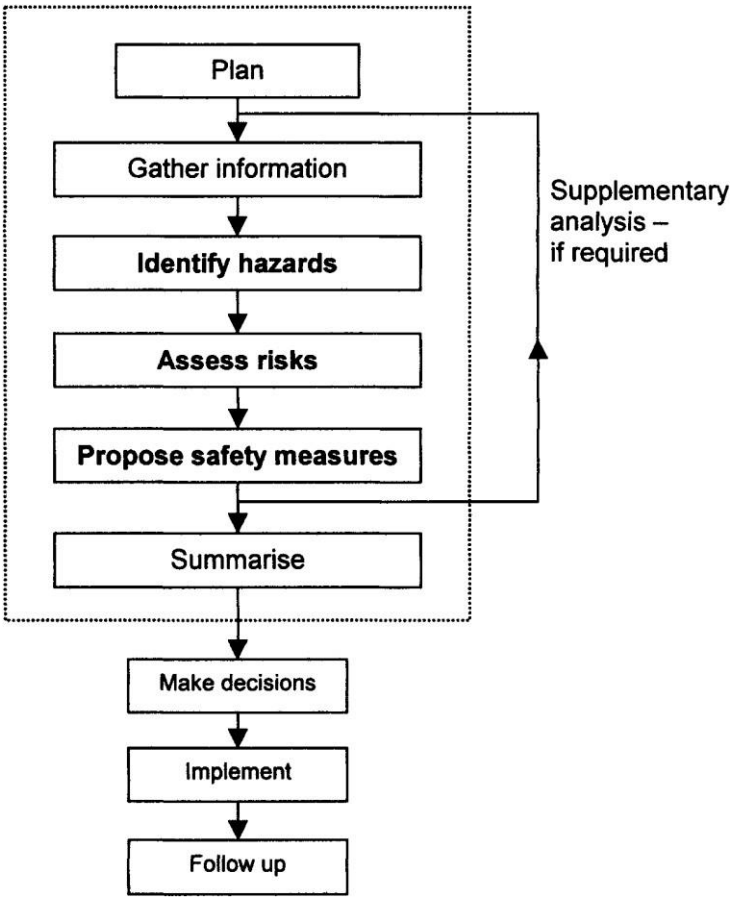

Figure 3 Example of stages of safety analysis procedure, and how results are used.

5. Leading and Lagging indicators of safety performance

Safety leading indicators are proactive measures that measure prevention efforts and can be observed and recorded prior to an injury. As opposed, safety lagging indicators are reactive measures that track only negative outcomes, such as an injury, once it has already occurred. (Cary Usrey, 2016)

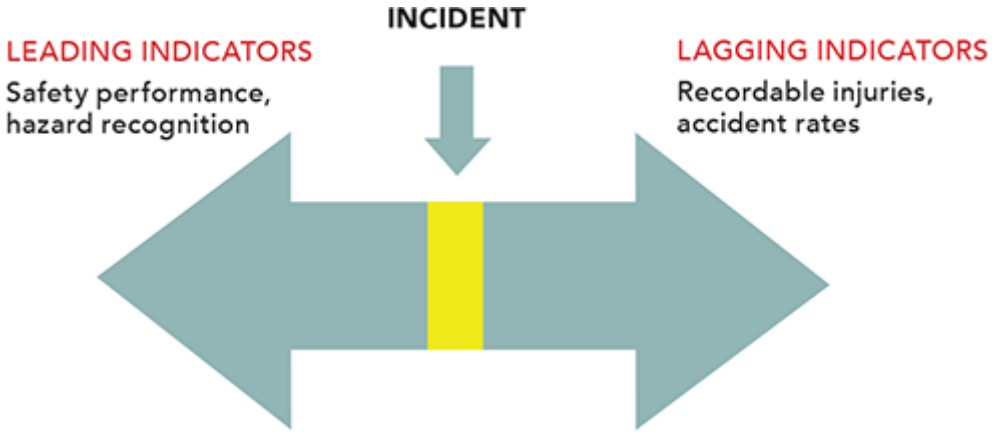

Figure 4 Leading and Lagging Indicators of Safety Performance

1. Lagging indicators of safety performance

Lagging indicators measure a company's incidents in the form of past accident statistics. Examples include:

- Injury frequency and severity
- OSHA recordable injuries
- Lost workdays
- Worker's compensation costs

Why use lagging indicators?

Lagging indicators are the traditional safety metrics used to indicate progress toward compliance with safety rules. These metrics evaluate the overall effectiveness of safety at the facility. They tell us how many people got hurt and how badly.

2. Leading indicators of safety performance

A leading indicator is a measure preceding or indicating a future event used to drive and measure activities carried out to prevent and control injury. Examples include:

- Safety training
- Ergonomic opportunities identified and corrected
- Reduction of MSD risk factors
- Employee perception surveys
- Safety audits

Why use leading indicators?

Leading indicators are focused on future safety performance and continuous improvement. These measures are proactive in nature and report what employees are doing on a regular basis to prevent injuries.

Using Leading Indicators to Improve Safety and Health Outcomes

Leading indicators can play a vital role in preventing worker fatalities, injuries, and illnesses and strengthening other safety and health outcomes in the workplace. Leading indicators are proactive and preventive measures that can shed light on the effectiveness of safety and health activities and reveal potential problems in a safety and health program. (OSHA, 2022)

Leading indicators are a valuable tool regardless of whether we have a safety or health program, what we have included in our program, or what stage we may be at in our program. OSHA encourages employers to get started today.

Leading indicators can improve organizational performance in a variety of ways. Employers may find that leading indicators can:

- **Prevent** workplace injuries and illnesses.
- **Reduce** costs associated with incidents.
- **Improve** productivity and overall organizational performance.
- **Optimize** safety and health performance.
- **Raise** worker participation.

One example of a company that use leading indicators is the Caterpillar Company; Caterpillar is one of the world's leading manufacturer of construction and mining equipment, diesel and natural gas engines, industrial turbines...

An article on EHS Today titled, "Caterpillar: Using Leading Indicators to Create World-Class Safety" recaps an interview with two Caterpillar executives who explained how they were able to successfully transition to a culture that utilizes leading indicators for safety.

According to the execs at Caterpillar, "… traditional metrics can help companies tell the score at the end of the game, but they don't help employers understand the strengths and weaknesses of their safety efforts and cannot help managers predict future success."

By utilizing a Safety Strategic Improvement Process (SIP) that emphasized leading indicators of safety, they saw an 85% reduction in injuries and $450 million in direct/indirect cost savings.

According to the article, the critical elements of the SIP included:

- Enterprise-wide statement of safety culture.
- Global process, tools, and metrics.
- Top-down leadership of and engagement with the process.
- Clearly defined and linked roles and responsibilities.
- Clearly defined accountability.
- Consistent methods establishing targets and reporting performance.
- Consistent criteria for prioritizing issues and aligning resources.

- Recognition for positive behavior and performance.

3. Conclusion

To improve the safety performance of the facility, we should use a combination of leading and lagging indicators.

When using leading indicators, it is important to make our metrics based on impact. For example, do not just track the number and attendance of safety meetings and training sessions – measure the impact of the safety meeting by determining the number of people who met the key learning objectives of the meeting/training. (Mark Middlesworth, 2022)

6. Safety data sources

According to (James Pomeroy, 2019), the profession's use of data has historically been dominated by injury and ill health data, primarily because this information has been readily available. However, even the best analysis of incident data is only telling half the story because it ignores the potentially influential factors such as changes in work type, hazards, or exposure periods. Many of the factors that are potentially influencing performance are non-OSH information and much of this data is stored in different business systems or external sources. Data sources are also frequently stored within different formats and systems, making them very difficult to access. Advancements in data science enable us to access and mine different sources of information at speeds and methods not previously possible.

Companies today have an unprecedented opportunity to leverage disparate data sources and commercially-viable analytic tools to support and inform their strategic safety decisions.

This enables organizations to extend their analysis beyond conventional injury case management reports to other data sources not directly associated with workplace safety.

Fusing disparate data sources can also help organizations view workplace safety incidents from a variety of different analytic perspectives. Rather than analyzing workplace safety from a traditional employee-focused perspective by trying to determine, for example, what employee attributes contributed to the incident (e.g., fatigue, training and engagement, age, tenure), companies can begin to focus on other parts of the workplace ecosystem to construct a more holistic model of the incident.

By refocusing the analytic perspective, organizations can consider variables such as weather, aspects of the job site, maintenance schedules, production measures, financial data, etc., to identify other causal factors not associated with the worker at all. This positions companies to take preventative actions to reduce the non-worker-related risk, such as adjustments to equipment maintenance schedules, placement of machines and vehicles, or scheduling of particular tasks at different times of the day. (Deloitte, 2012)

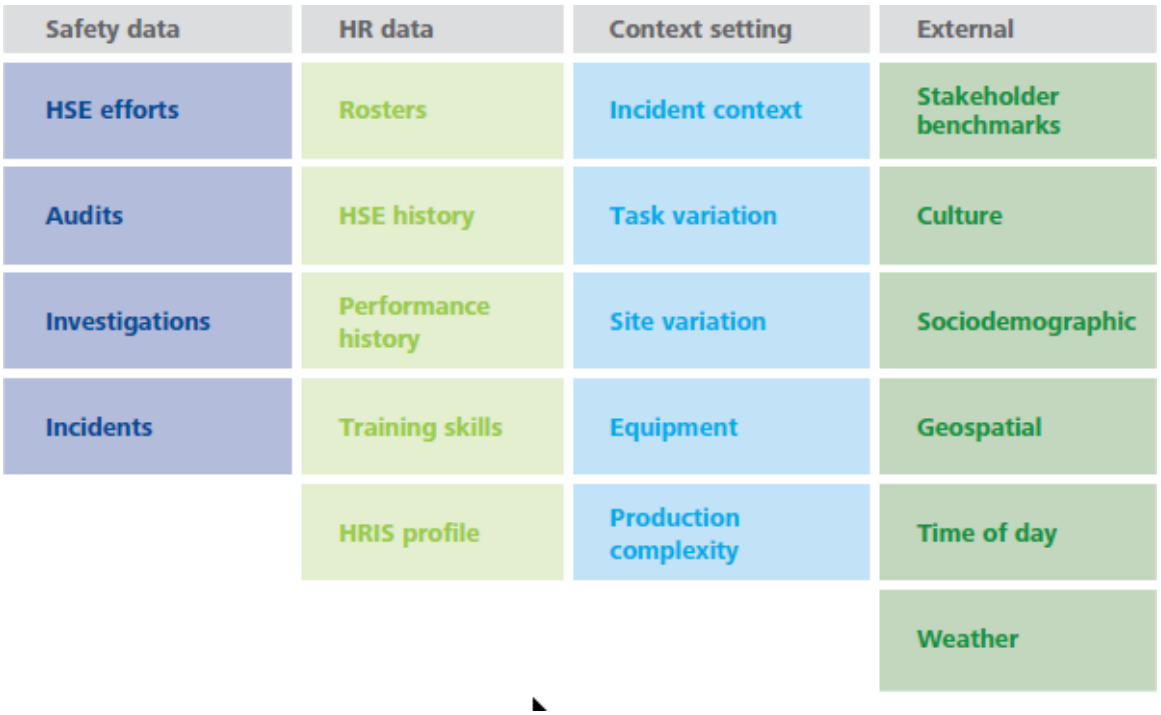

Figure 5 Examples of safety analytic data sources

The critical aspect of a safety analytics initiative is the ability to act on findings in a timely manner. This is where predictive modeling, advanced machine learning, and data visualization technologies come into play.

Chapter V Exploratory data analysis

IN THIS CHAPTER

- ✓ Discovering the exploratory data analysis method and it is important for occupational safety and health.
- ✓ Getting to know the types of exploratory data analysis and some common data science tools used to create an EDA.
- ✓ Discovering the basic steps to conduct an exploratory analysis.

1. Definition

In addition to the definition that we covered earlier, according to (IBM Cloud Education, 2020), exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers we need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis-testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques we are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

2. Why is exploratory data analysis important for occupational safety and health?

The exploratory analysis of data is the first important step in any research study. The main objective of conducting exploratory analysis is to achieve maximum insights into the data by employing a variety of techniques. Typically, these techniques are employed before building a statistical model or conducting more analysis that are advanced. Exploratory data analyses are mainly performed to discover variable patterns, identify outliers, test hypotheses, and to examine data assumptions using summary statistics and graphical representations. Through the visualization of data, exploratory analyses can reveal the underlying structure of the data and help select one or more models for analyzing the data. (Lord et al., 2021)

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, and find interesting relations among the variables.

OHS professionals can use exploratory analysis to ensure the results they produce are valid and applicable to any desired safety outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including *predictive analytics models and machine learning*.

According to (Terence Shin, 2020), conducting an EDA allows us to turn an almost useable dataset into a completely useable dataset. It does not that EDA can magically make any dataset clean — that is not true. However, many EDA techniques can remedy some common problems that are present in every dataset.

Exploratory Data Analysis does two main things:

1. It helps clean up a dataset.

2. It gives us a better understanding of the variables and the relationships between them.

3. Types of exploratory data analysis

According to (IBM Cloud Education, 2020) There are four primary types of EDA:

- **Univariate non-graphical**

This is the simplest form of data analysis, where the data being analyzed consists of just one variable. Since it is a single variable, it does not deal with causes or relationships. The main purpose of the univariate analysis is to describe the data and find patterns that exist within it.

- **Univariate graphical**

Non-graphical methods do not provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:

- Stem-and-leaf plots, which show all data values and the shape of the distribution.
- Histograms are a bar plots in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
- Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.

- **Multivariate nongraphical**

Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.

- **Multivariate graphical**

Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

4. Exploratory Data Analysis Tools

Some of the most common data science tools used to create an EDA are:

- **R Language**

R is an open-source language and environment for statistical computing and graphics. It provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering …) and graphical techniques, and is highly extensible. (The R Foundation, 2022)

The R language is widely used among statisticians in data science in developing statistical observations and data analysis.

- **Python**

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development and for use as a scripting or glue language to connect existing components. Python's simple, easy-to-learn syntax emphasizes readability and reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed. (Python Software Foundation, 2022)

According to (Yuli Vasiliev, 2022) the easy-on-the-brain Python programming language is an ideal choice for accessing, manipulating, and gaining insight from data of any kind. It has both a rich set of built-in data structures for basic operations and a robust ecosystem of open source libraries for data analysis and manipulation of any level of complexity.

Python and EDA can be used together to identify missing values in a data set, which is important so we can decide how to handle missing values for machine learning.

Python Libraries for Data Analytics

One of the main reasons why Data Analytics using Python has become the most preferred and popular mode of data analysis is that it provides a range of libraries for numerical computation, data manipulation, graphics and data visualization to build plots.

- **NumPy**

  NumPy supports n-dimensional arrays and provides numerical computing tools. It is useful for Linear algebra and Fourier transformation.

- **Pandas**

  Pandas provides functions to handle missing data, perform mathematical operations, and manipulate the data.

- **Matplotlib**

  Matplotlib library is commonly used for plotting data points and creating interactive visualizations of the data.

- **SciPy**

  SciPy library is used for scientific computing. It contains modules for optimization, linear algebra, integration, interpolation, special functions, signal and image processing.

- **Scikit-Learn**

Scikit-Learn library has features that allow you to build regression, classification, and clustering models.

5. Exploratory data analysis process

(Indeed Editorial Team, 2021) suggest that before we begin exploratory data analysis, it is important to understand a few key terms:

- **Value:** A data value is a piece of information, such as a number or a date.
- **Variable:** A data variable is a characteristic that you can measure, such as weight or income.
- **Distribution:** The distribution of a dataset is how the dataset is spread out. We can visualize a dataset's distribution by observing its shape on a graph.
- **Outlier:** An outlier is a data value that is significantly different, including much higher or lower, from the rest of a dataset.
- **Data model:** A data model is a method of organizing data and relationships between values in a dataset.

There is no specific process for EDA, as it varies from one source to another, and from case to case as necessary. These are the basic steps for exploratory analysis:

1. Data Collection

Data collection is the process of gathering information in an established systematic way that enables one to test hypotheses and evaluate outcomes easily.

After getting data, we need to check the data type of features. There are the following types of features:

- numeric
- categorical
- ordinal
- DateTime
- coordinates

2. Data Cleaning

Data cleaning is the process of ensuring that your data is correct and useable by identifying any errors in the data, or missing data by correcting or deleting them. Refer to this link for data cleaning.

Once the data is clean, we can go further for data preprocessing.

3. Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. It includes normalization and standardization, transformation,

feature extraction, selection, etc. The product of data preprocessing is the final training dataset.

4. Data Visualization

Data visualization is the graphical representation of information and data. It uses statistical graphics, plots, information graphics and other tools to communicate information clearly and efficiently.

Figure 6 Data Visualization in Python

Data Visualization in Python

Python offers several plotting libraries, namely Matplotlib, Seaborn, and many other such data visualization packages with different features for creating informative, customized, and appealing plots to present data most simply and effectively, as shown in Figure 6.

Chapter VI Predictive data analysis

IN THIS CHAPTER

- ✓ Discovering the predictive data analysis method
- ✓ Addressing the question, can workplace injuries be predicted?
- ✓ Exploring the predictive analytics process.
- ✓ Discovering the predictive analytics models.
- ✓ Understanding the concept of the predictive questions.

1. Introduction

The impact of technology on health and safety has added a whole new depth to managing health and safety and new solutions providers have taken up the mantel when it comes to developing new tools and techniques for keeping workers safe.

Predictive analytics as a method has been around for many years and has been used in a variety of different industries to try to improve performance as time goes on. The increasing sophistication of techniques used means it is now able to be applied in a safety context in a much more methodical way.

(SAS company, 2022) defines predictive analytics as the use of data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to provide the best assessment of what will happen in the future.

2. Can workplace injuries really be predicted?

The simple answer is yes, workplace injuries and safety incidents can be predicted before they happen. This has been confirmed by research conducted by teams from Predictive Solutions Corporation and Carnegie Mellon University (CMU) - the same CMU team that helped develop the Watson supercomputer that originally gained fame by beating the top "Jeopardy" champions and has since been applied to helping doctors diagnose rare and complicated diseases. Using a subset of Predictive Solutions data set of over 112 million safety observations and their associated safety incidents recorded from over 15,000 individual worksites, the researchers proved that workplace incidents could indeed be predicted before they happen, with high levels of accuracy.

They also found that the safety inspection and observation data from these worksites was a strong predictor of future incidents. The researchers developed a number of predictive models with accuracy levels between 80 and 97% in predicting injuries at actual worksites. The research also found a high degree of correlation - r-squared as high as 0.754 - between predicted and actual incidents. (Predictive Solutions, 2012)

3. Predictive analytics process

According to (insightsoftware, 2022b), any successful predictive analytics project will involve these steps:

Identify what you want to know based on past data

First, we should identify what questions we want to answer. What are some of the important decisions we will make with the insight? Knowing this is a crucial first step to applying predictive analysis.

Consider if you have the data to answer those questions

Next, we should ask these questions: is our operational system capturing the needed data? How clean is it? How far in the past do we have this data, and is that enough to learn any predictive patterns?

Train the system to learn from your data and predict outcomes

When building our model, we will have to start by training the system to learn from data. By establishing the right controls and algorithms, we can correlate that data to safety predictions.

Another key component is to regularly retrain the learning module. Set a timeline, maybe once a month or once a quarter to regularly retrain our predictive analytics learning module to update the information.

Schedule your modules

Predictive analytics modules can work as often as we need. For example, if we get new safety data every Tuesday, we can automatically set the system to upload that data when it comes in.

Use the insights and predictions to act on these decisions

Predictive analytics is only useful if we use it. We will need leadership champions to enable activities to make change a reality. These predictive insights can be embedded into our safety system for everyone in your organization to use.

4. Predictive Analytics Models

Predictive analytics tools are powered by several different models and algorithms that can be applied to a wide range of use cases. Determining what predictive modeling techniques are best for our company is key to getting the most out of a predictive analytics solution and leveraging data to make insightful decisions.

Let us go over the top and common predictive analytics models:

1. Classification Model

The classification model is, in some ways, the simplest of the several types of predictive analytics models we are going to cover. It puts data in categories based on what it learns from historical data.

Classification models are best to answer yes or no questions, providing broad analysis that is helpful for guiding decisive action. These models can answer questions such as:

- Will an accident happen in this sector during the next four weeks?

The breadth of possibilities with the classification model, and the ease by which it can be retrained with new data, means it can be applied to many different industries.

2. Clustering Model

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabeled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group. (geeksforgeeks, 2021)

This process includes a number of different algorithms and methods to make clusters of a similar kind.

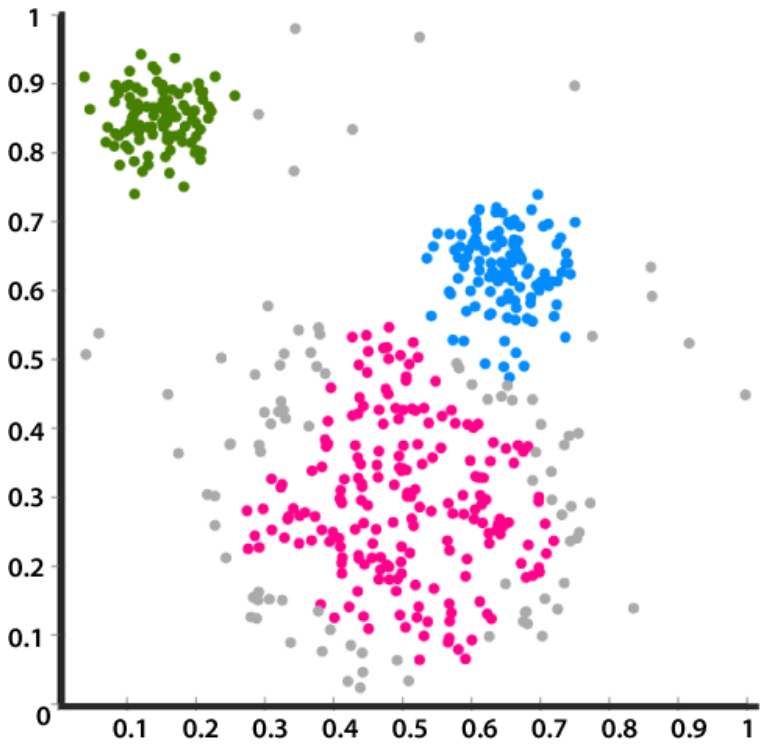

Figure 7 an example of a cluster graph

Clusters should exhibit high internal homogeneity and high external heterogeneity. What does this mean? When plotted geometrically, objects within clusters should be very close together and clusters will be far apart.

Properties of Clustering:

- **Clustering Scalability:** In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable if it is not scalable, then we cannot get the appropriate result which would lead to wrong results.
- **High Dimensionality:** The algorithm should be able to handle high dimensional space along with the data of small size.
- **Algorithm Usability with multiple data kinds:** Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical, and interval-based data, binary data, etc.

- **Dealing with unstructured data:** Databases that contain missing values, and noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. Therefore, it should be able to handle unstructured data and give some structure to the data by organizing it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.
- **Interpretability:** The outcomes of clustering should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

Clustering Methods:

The clustering methods can be classified into the following categories:

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

3. Forecast Model

One of the most widely used predictive analytics models, the forecast model deals in metric value prediction, estimating the numeric value for new data based on learnings from historical data.

This model can be applied wherever historical numerical data is available. The forecast model also considers multiple input parameters.

Any safety situation in which historical data can be processed can make use of this predictive analytics model. For example, a safety inspector or manager could predict the amount of missed or faulted risk assessments that may occur within a specific period.

4. Outliers model

Accident data sets can include some unusual data points that are not typical of the rest of the data.

The outliers model is oriented around anomalous data entries within a dataset. It can identify anomalous figures either by themselves or in conjunction with other numbers and categories.

This model is particularly useful in health and safety as it identifies anomalous data within a series. Not only can it categorize information, but also if the input data is correct it can identify anomalies that may indicate a specific area where a safety concern needs to be addressed.

Take, for example, a series of factories or plants in which incidents are recorded over time, if a spike is recorded during a certain period, an outliers model could identify the data in addition to potential anomalies in parameters (like air quality) which may be present in different factories. (hse-network, 2020)

5. Time series model

According to (tableau.com, 2022), time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting—predicting future data based on historical data.

Why organizations use time series data analysis?

Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time. Using data visualizations, safety managers can see trends and dig deeper into why these trends occur.

When organizations analyze data over consistent intervals, they can also use time series forecasting to predict the likelihood of future events.

What is time series forecasting?

Time series forecasting is the process of analyzing time series data using statistics and modeling to make predictions and inform strategic decision-making. It is not always an exact prediction, and the likelihood of forecasts can vary wildly—especially when dealing with the commonly fluctuating variables in time series data as well as factors outside our control. However, forecasting insight about which outcomes are more likely—or less likely—to occur than other potential outcomes.

Time Series Analysis Types

Because time series analysis includes many categories or variations of data, analysts sometimes must make complex models. However, analysts cannot account for all variances, and they cannot generalize a specific model to every sample. Models that are too complex or that try to do too many things can lead to a lack of fit. Lack of fit or overfitting models lead to those models not distinguishing between random error and true relationships, leaving analysis skewed and forecasts incorrectly.

Models of time series analysis include:

- Classification: Identifies and assigns categories to the data.
- Curve fitting: Plots the data along a curve to studying the relationships of variables within the data.
- Descriptive analysis: Identifies patterns in time series data, like trends, cycles, or seasonal variation.
- Explanative analysis: Attempts to understand the data and the relationships within it, as well as cause and effect.
- Exploratory analysis: Highlights the main characteristics of the time series data, usually in a visual format.
- Forecasting: Predicts future data. This type is based on historical trends. It uses the historical data as a model for future data, predicting scenarios that could happen along with future plot points.
- Intervention analysis: Studies how an event can change the data.
- Segmentation: Splits the data into segments to show the underlying properties of the source information.

6. Classification model

One of the most common predictive analytics models is the classification model. These models work by categorizing information based on historical data. Classification models are used in different industries because they can be easily retrained with new data and can provide a broad analysis for answering questions.

Classification begins with a training dataset where each piece of data has already been labeled. The classification algorithm learns the correlations between the data and labels and categorizes any new data. Some popular classification model techniques include decision trees, random forests, and text analytics.

7. Summary

How do we determine which predictive analytics model is best for our needs? We need to start by identifying what predictive questions we are looking to answer, and more importantly, what we are looking to do with that information. Consider the strengths of each model, as well as how each of them can be optimized with different predictive analytics algorithms, to decide how to best use them for our organization.

5. Are We Asking the Right Predictive Questions?

Predictive analytics works by learning the patterns that exist in our historical data, then using those patterns to predict future outcomes.

The process of feeding in historical data for different outcomes and enabling the algorithm to learn how to predict is called the training process (as shown in figure 7).

Once the algorithm determines a pattern, we pass on new safety data and it will make a prediction. However, the first step is deciding what predictive questions we want to answer.

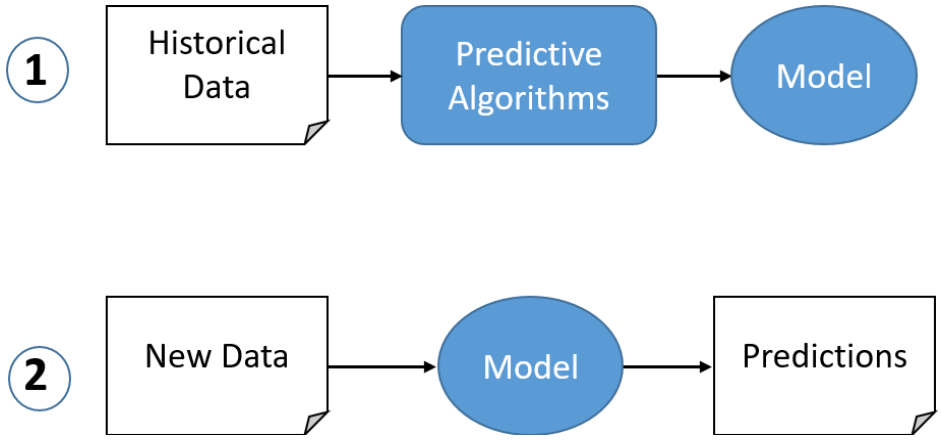

Figure 8 Training process

How do we know which predictive questions to ask?

When determining a predictive question, the rule of thumb is to base it on what we want to do with the answer.

Following that logic, if we are forecasting Lost Time Injuries for a company, our question might be: "How many LTIs will I get in the next six months?" That is a forecasting/regression question. However, we could also ask a binary question such as: "Will I get more than 10 LTIs in the next six months?" That is a classification question because the answer will either be yes or no.

The predictive question we should ask will depend on what we are going to do with the information. If we have the staff to handle 10 LTIs, then we will likely want to know if we will get 10 LTIs or not (so we would ask the classification question). However, if our goal is to identify how many LTIs we are going to get in the next six months so that we can staff accordingly, we would ask the forecasting question.

Over time, we will be able to run multiple algorithms to pick the one that works best with our data, or even use an ensemble of algorithms. We will also want to regularly retrain our learning model to keep up with fluctuations in our data based on several variables. (insightsoftware, 2022a).

Chapter VII Case study — exploratory data analysis for manufacturing plants database

1. Overview

The database comes from one of the biggest industries in South America and in the world. It is an urgent need for industries/companies around the globe to understand why employees still suffer some injuries/accidents in plants. Sometimes they also die in such environment.

We chose to work with the exploratory data analysis because it better suits our data set, allowing us to achieve maximum insights into the data, discover variable patterns, identify outliers, test hypotheses, and examine data assumptions using summary statistics and graphical representations.

2. Objectives

- Analyze real labor accident data aiming to help manufacturing plants to find solutions that improve the level of safety by reducing the rate of accidents, and even saving lives.
- This case study does not intend to use any machine learning technique, but to help people to take valuable insights from a few lines of data, and showcase the use of data analysis to enhance safety.

3. Details

- We will use two basic concepts about data exploration: Highlight and Insight
  - **Highlight:** Observations and facts extracted from summarizing and reviewing the data.
  - **Insight:** ideas and questions that come from the Highlights.
- Programming Language for data analysis: **Python**.
- Data handling tools: **pandas** helping with data-frame operations.
- Data Visualization tools: **Seaborn and Plotly**.
- Environment used to write and execute the code: **Google Colab**.
  A free Jupyter notebook environment runs entirely in the cloud. Most importantly, it does not require a setup, and the notebooks that the user create can be simultaneously edited by the team members.

4. Data Description:

The database is basically records of accidents from 12 different plants in 03 different countries from South America.

The dataset source is from: (www.kaggle.com, 2018).

Columns description:

- Data: timestamp or time/date information.
- Countries: which country the accident occurred, country 2 and 3 are anonymized, we know that country 1 is Brazil.
- Local: the city where the manufacturing plant is located (anonymized).
- Industry sector: which sector the plant belongs to.
- Accident level: from I to VI, it registers how severe was the accident (VI is the highest level of accidents).
- Potential Accident Level: Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors involved in the accident).
- Genre: if the person is male or female.
- Employee or Third Party: if the injured person is an employee or a third party.
- Critical Risk: some description of the risk involved in the accident.
- Description: Detailed description of how the accident happened.

5. Why data analytics using Python?

There are many programming languages available, but Python is popularly used by statisticians, engineers, and scientists to perform data analytics.

Here are some of the reasons why we choose to use Python in our study:

- Python is easy to learn and understand and has a simple syntax.
- The programming language is scalable and flexible.
- It has a vast collection of libraries for numerical computation and data manipulation.
- Python provides libraries for graphics and data visualization to build plots.
- It has broad community support to help solve many kinds of queries.

6. Application of EDA using Python

First, we started with importing the necessary libraries and adjusting the settings options.

After that, we followed three steps to conduct our exploratory data analysis:

- Data collection
- Data cleaning
- Univariate analysis

Then we did some research about the calendar related insights to study the causes and find potential solutions.

Last but not least, we wrapped up the process with a conclusion, and we mentioned some limitations that faced us.

The Python source code for this study will be delivered separately.

1. Data collection

Load dataset

	Unnamed: 0	Data	Countries	Local	Industry Sector	Accident Level	Potential Accident Level	Genre	Employee or Third Party	Critical Risk	Description
0	0	2016-01-01 00:00:00	Country_01	Local_01	Mining	I	IV	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 f...
1	1	2016-01-02 00:00:00	Country_02	Local_02	Mining	I	IV	Male	Employee	Pressurized Systems	During the activation of a sodium sulphide pum...
2	2	2016-01-06 00:00:00	Country_01	Local_03	Mining	I	III	Male	Third Party (Remote)	Manual Tools	In the sub-station MILPO located at level +170...
3	3	2016-01-08 00:00:00	Country_01	Local_04	Mining	I	I	Male	Third Party	Others	Being 9:45 am. approximately in the Nv. 1880 C...
4	4	2016-01-10 00:00:00	Country_01	Local_04	Mining	IV	IV	Male	Third Party	Others	Approximately at 11:45 a.m. in circumstances t...

Figure 9 the top five rows

Shape of the data

```
Number of rows = 425 and Number of Columns = 11 in the Data frame
```

Outcome:

- Small data set but with some relevant information.

The data type of each attribute

Outcome:

- We found that except first column all other columns' data type is object.
- Categorical columns - 'Countries', 'Local', 'Industry Sector', 'Accident Level', 'Potential Accident Level', 'Genre', 'Employee or Third Party', 'Critical Risk', 'Description'
- Date column - 'Data'

Data Collection Summary:

- There are about 425 rows and 11 columns in the dataset.
- We noticed that except for a 'date' column all other columns are categorical columns.

2. Data cleaning

Remove 'Unnamed: 0' and Rename - 'Data', 'Countries', 'Genre', 'Employee or Third Party' columns

	Date	Country	Local	Industry Sector	Accident Level	Potential Accident Level	Gender	Employee type	Critical Risk	Description
0	2016-01-01 00:00:00	Country_01	Local_01	Mining	I	IV	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 f...
1	2016-01-02 00:00:00	Country_02	Local_02	Mining	I	IV	Male	Employee	Pressurized Systems	During the activation of a sodium sulphide pum...

Figure 10 the top row after the changements

Check Duplicates

- We found seven duplicates.

View the duplicate records

	Date	Country	Local	Industry Sector	Accident Level	Potential Accident Level	Gender	Employee type	Critical Risk	Description
77	2016-04-01 00:00:00	Country_01	Local_01	Mining	I	V	Male	Third Party (Remote)	Others	In circumstances that two workers of the Abrat...
262	2016-12-01 00:00:00	Country_01	Local_03	Mining	I	IV	Male	Employee	Others	During the activity of chuteo of ore in hopper...
303	2017-01-21 00:00:00	Country_02	Local_02	Mining	I	I	Male	Third Party (Remote)	Others	Employees engaged in the removal of material f...
345	2017-03-02 00:00:00	Country_03	Local_10	Others	I	I	Male	Third Party	Venomous Animals	On 02/03/17 during the soil sampling in the re...
346	2017-03-02 00:00:00	Country_03	Local_10	Others	I	I	Male	Third Party	Venomous Animals	On 02/03/17 during the soil sampling in the re...
355	2017-03-15 00:00:00	Country_03	Local_10	Others	I	I	Male	Third Party	Venomous Animals	Team of the VMS Project performed soil collect...
397	2017-05-23 00:00:00	Country_01	Local_04	Mining	I	IV	Male	Third Party	Projection of fragments	In moments when the 02 collaborators carried o...

Figure 11 the duplicate records

Outcome:

- We do not have to worry about data retention, it's already part of the industry dataset and we can just remove or drop these rows from the cleaned data.

Drop Duplicates

Get the shape of Industry data

```
Number of rows = 418 and Number of Columns = 10 in the Data frame after removing the duplicates.
```

Check Outliers

As we know, there is no concept of outliers detection in categorical variables (nominal and ordinal), as each value is counted as a label. Let us check the uniqueness and frequency (mode) of each variable.

Outcome:

- We observed that there are records of accidents from 1st Jan 2016 to 9th July 2017 in every month. Therefore, there are no outliers in the 'Date' column.
- There are only three country types so there are no outliers in 'Country' column.

- There are 12 Local cities where the manufacturing plant is located and its types are in sequence so there are no outliers in 'Local' column.
- There are only three Industry Sector types that are in sequence so there are no outliers in 'Industry Sector' column.
- There are only five Accident Level types that are in sequence so there are no outliers in 'Accident Level' column.
- There are only six Potential Accident Level types that are in sequence, so there are no outliers in 'Potential Accident Level' column.
- There are only two Gender types in the provided data so there are no outliers in 'Gender' column.
- There are only three Employee types in the provided data so there are no outliers in 'Gender' column.
- There are quite a lot of Critical risk descriptions and we do not see any outliers but with the help of SME, we can decide whether this column has outliers or not.

Check Missing Values

Outcome:

- No missing values.

Data Cleansing Summary:

- Removed 'Unnamed: 0' column and renamed - 'Data', 'Countries', 'Genre', 'Employee or Third Party' columns in the dataset.
- We had seven duplicate instances in the dataset and dropped those duplicates.
- There are no outliers in the dataset.
- No missing values in the dataset.
- We are left with 418 rows and 10 columns after data cleansing.

Variable Identification

- Target variable: 'Accident Level', 'Potential Accident Level'
- Predictors (Input variables): 'Date', 'Country', 'Local', 'Industry Sector', 'Gender', 'Employee type', 'Critical Risk', 'Description'

3. Univariate Analysis

1. Country

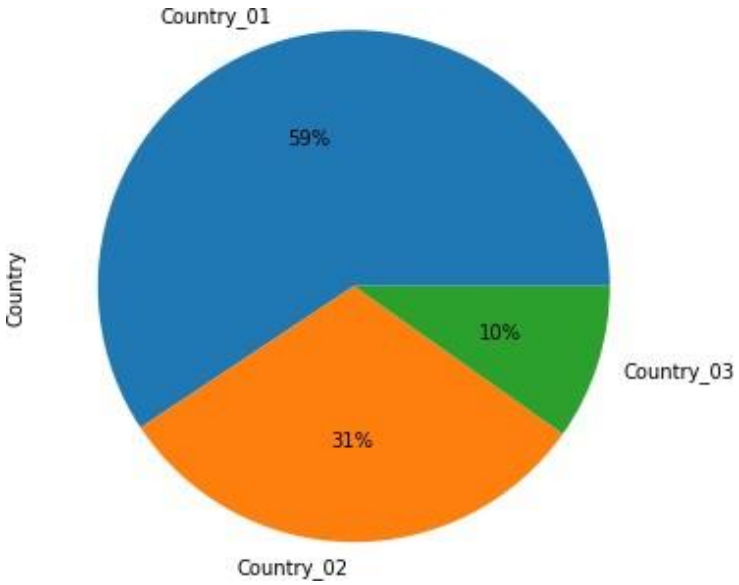

Figure 12 Accidents occurrence by country

Highlights

- 59% of accidents occurred in Country_01
- 31% of accidents occurred in Country_02
- 10% of accidents occurred in Country_03

Insights

- Why Country_01 has the most number of accidents?

2. Local

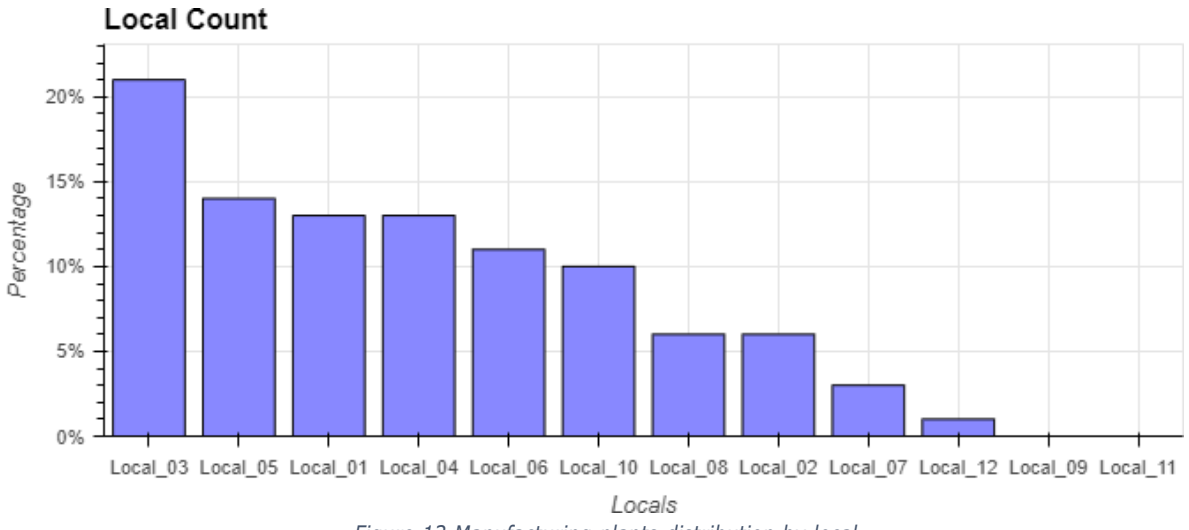

Figure 13 Manufacturing plants distribution by local

Highlights

- The highest manufacturing plants are located in Local_03 city.
- The lowest manufacturing plants are located in Local_09 & Local_11.

3. Industry Sector

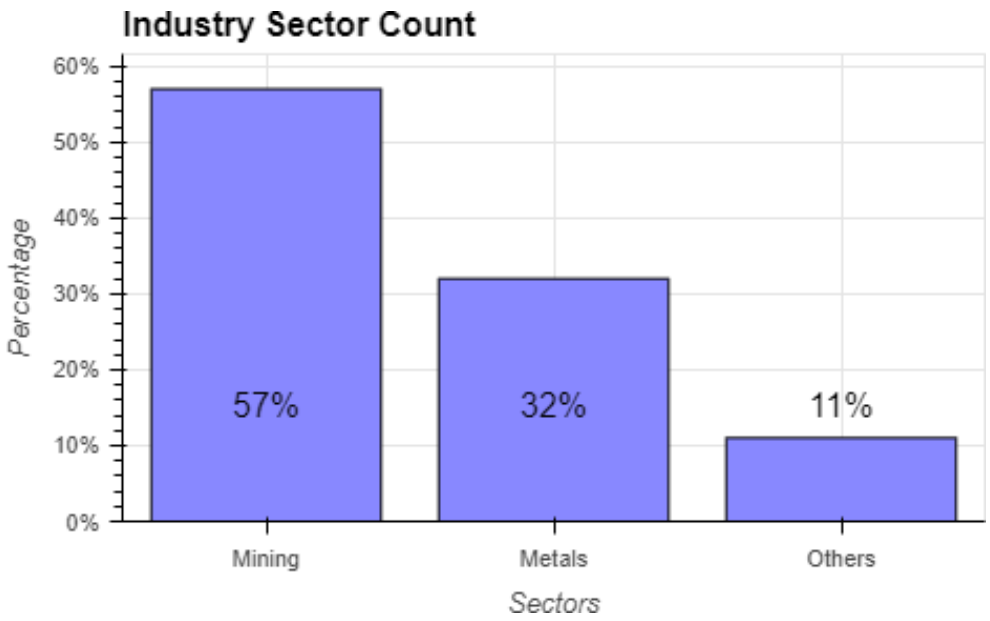

Figure 14 Distribution of "Industry Sector" label

Highlights

- 57% of manufacturing plants belong to "Mining" sector.
- 32% of manufacturing plants belong to "Metals" sector.
- 11% of manufacturing plants belong to "Others" sector.

4. Accident level

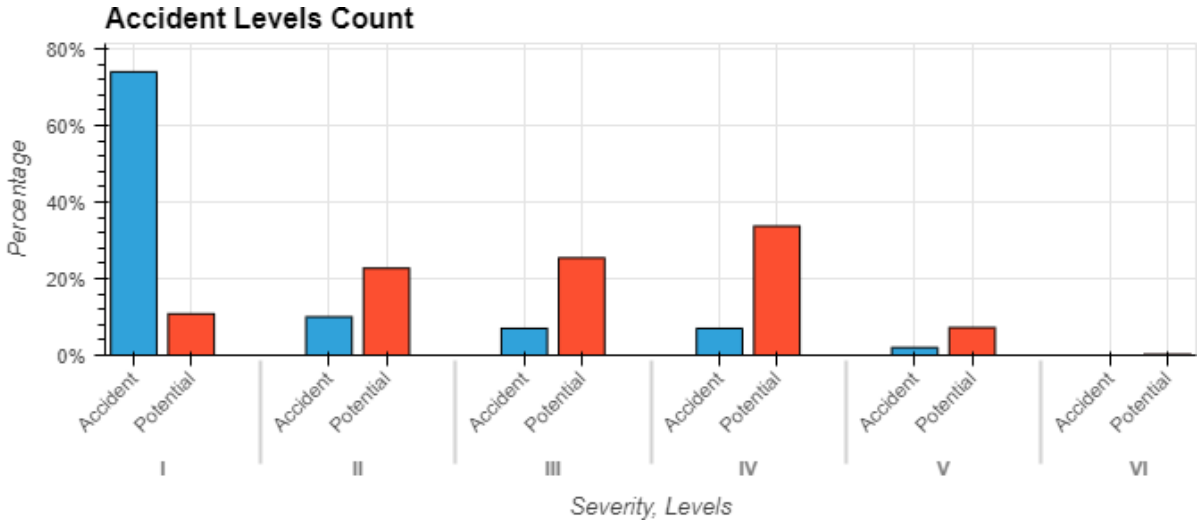

Figure 15 Distribution of "Accident Level" and "Potential Accident Level" label

Highlights

- The number of accidents decreases as the Accident Level increases.
- The number of accidents increases as the Potential Accident Level increases.

Insights

- Accidents with level of severity 1 are the most common!

5. Gender

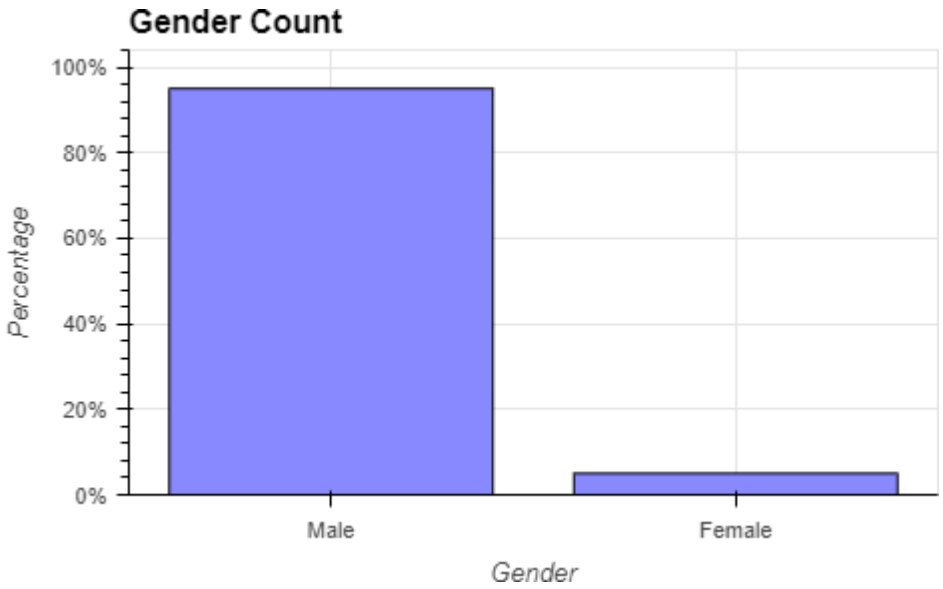

Figure 16 Distribution of workers by gender

Highlights

- The number of working men in this industry is much larger than the number of women.

6. Employee type

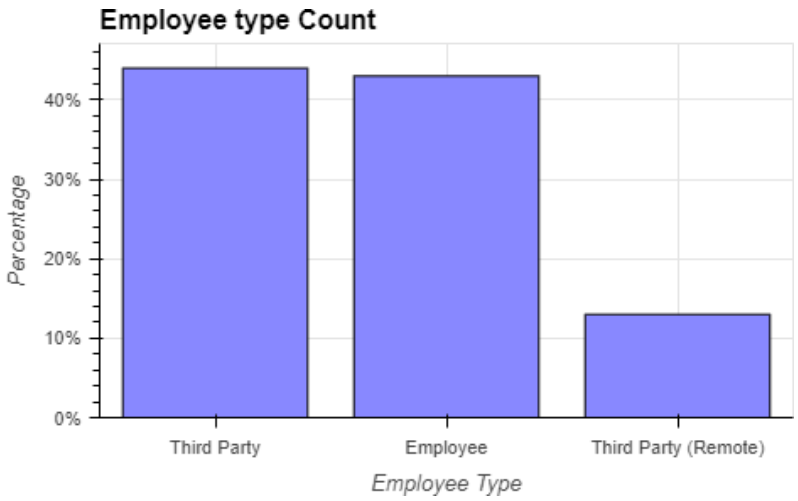

Figure 17 Distribution of workers by employee type

Highlights

- 44% Third party employees working in this industry.
- 43% own employees working in this industry.
- 13% Third party (Remote) employees working in this industry.

Insights

- A large part of the staff are from a third party, are they well trained?

7. Critical risk

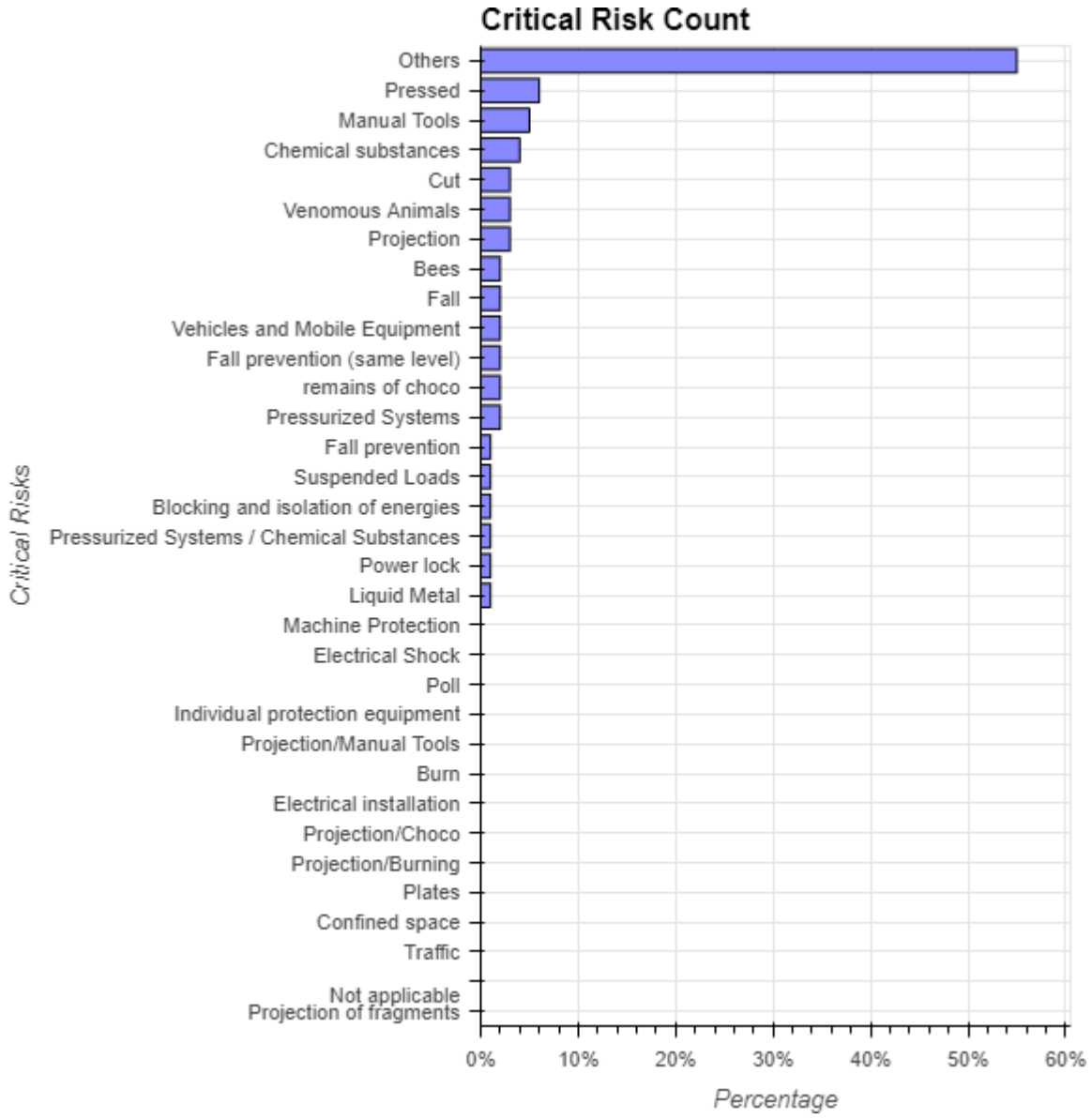

Figure 18 Percentage of critical risk classification

Highlights

- The Critical Risks classified as 'Others' are the most frequent.

Insights

- We need to focus on the column "Others", as it represents most of the Critical risks.

8. Calendar

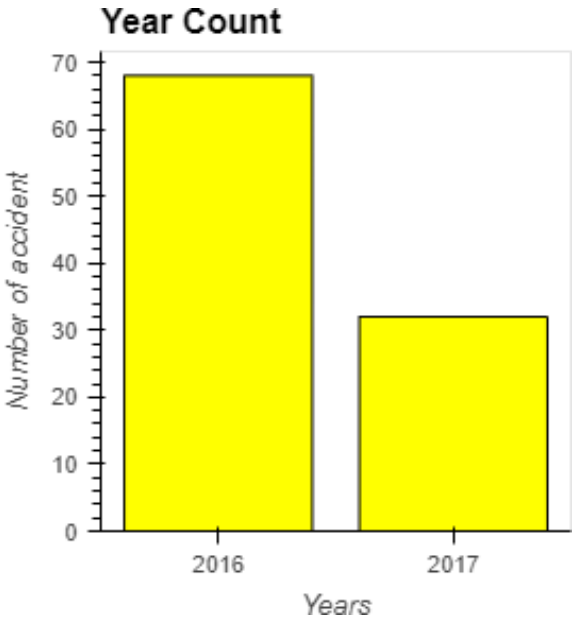

Figure 19 Accidents distribution by years

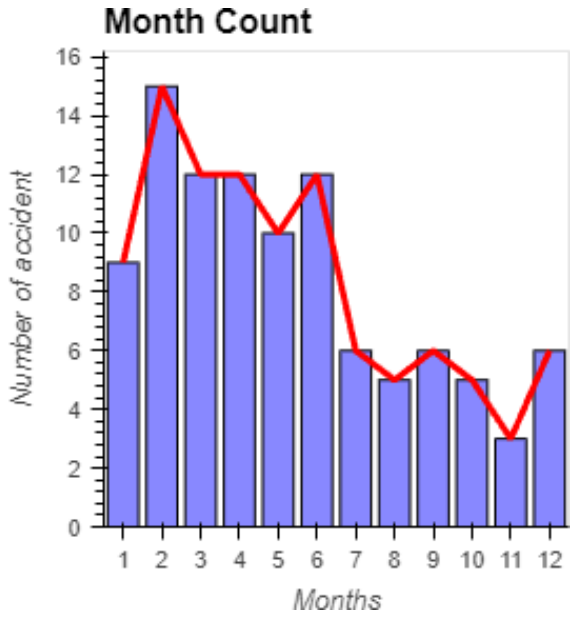

Figure 20 Accidents distribution by months

Highlights

- Accidents are recorded from 1st Jan 2016 to 9th July 2017 every month; there is a high number of accidents in 2016 and fewer in 2017.
- The number of accidents is high at beginning of the year and it keeps decreasing later.

Insights

- Why is the number of accidents in 2017 almost halved compared to the previous year?
- What are the reasons and routines that make the beginning of the year peak number of accidents?

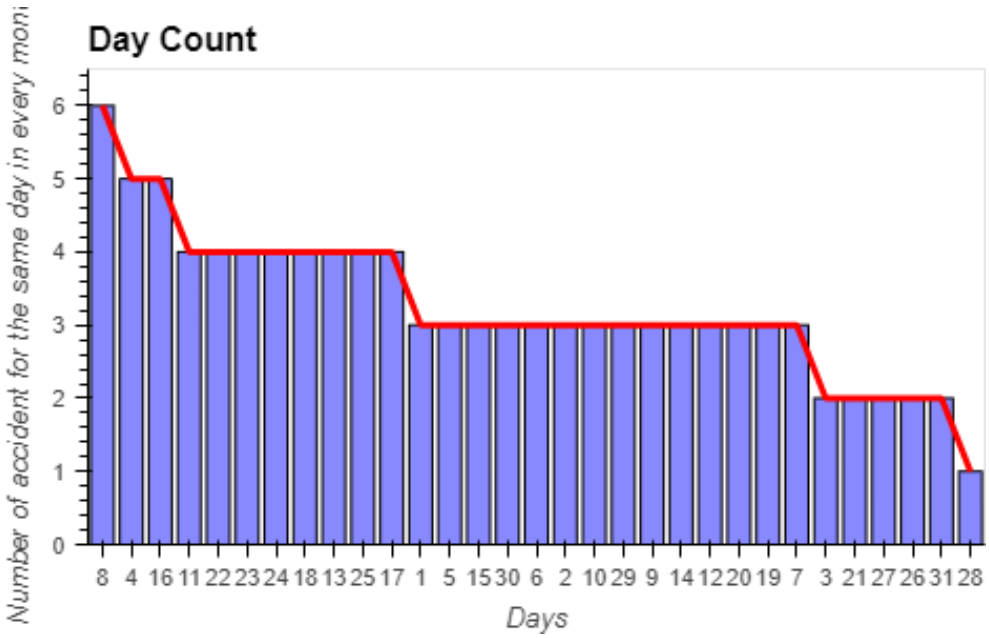

Figure 21 Accidents distribution by days

Highlights

- The number of accidents is very high on particular days like 4, 8, and 16 every month.

Insights

- What are the reasons behind these patterns?

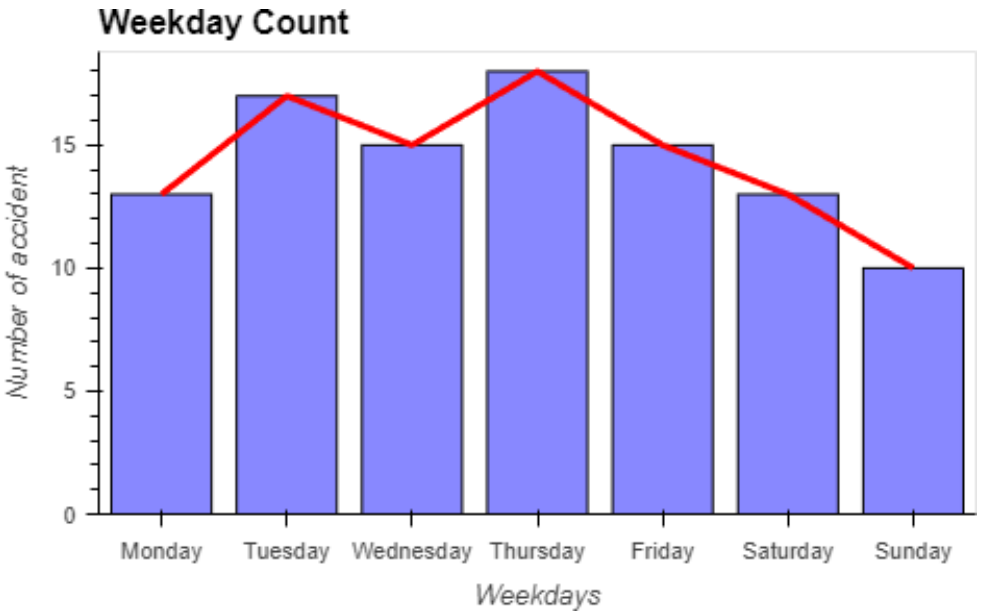

Figure 22 Accidents distribution by weekdays

Highlights

- The number of accidents increased during the middle of the week and declined after.

Insights

- Another pattern!

4. Expansion from the insights

Because of the lack of further information and details, we decided to break down the "calendar" insights, as we can find valuable data through research to study the causes and find potential solutions.

The analysis revealed that accidents are high at the beginning of the year! Therefore, we did traditional research on the topic and came out with the following reasons:

- Holidays and vacations

Brazil is famous for its variety and abundance of holidays, especially in the first half of the year, from the New Year's holidays to the carnivals. For most people that is a good thing, but from an HSE view, it represent a pain point, as it leaves the company with less number of personnel to work with, especially in the industrial factories. Alternatively, in other cases to work with inexperienced replacements.

LIST OF HOLIDAYS IN BRAZIL IN 2016

Day	Date	Holiday Name	Type	Comments
Friday	Jan 01	New Year's Day	National Holiday	
Monday	Feb 08	Carnival	National Holiday	Ponto facultativo
Tuesday	Feb 09	Carnival	National Holiday	Ponto facultativo
Wednesday	Feb 10	Ash Wednesday	National Holiday	Ponto facultativo
Friday	Mar 25	Good Friday	National Holiday	
Thursday	Apr 21	Tiradentes Day	National Holiday	
Sunday	May 01	Labour Day	National Holiday	International Labour Day
Thursday	May 26	Corpus Christi	National Holiday	Ponto facultativo. Second Thursday after Whitsun

Figure 23 List of holidays in Brazil in 2016

- Weather

As Brazil lies in the Southern Hemisphere, its seasons are the exact opposite of what Northern Hemisphere residents are used to summer is December through March, and winter is June through September. Within the country, the climate varies considerably from region to region. In most of Brazil, the summers are very hot. Temperatures can rise to 43°C (110°F) with high humidity. (www.frommers.com, n.d.)

Weather conditions can be very important at job sites, especially for those working outside or in small areas without air-conditioning. Construction workers, HVAC technicians, and manufacturers are among many professions whose health and safety are impacted by the weather.

Hot working conditions can be dangerous; one of the primary risks is heat stress. Heat stress (or heat injury) occurs when hot, humid temperatures prevent the body from properly sweating and cooling itself. It can be exacerbated by factors such as age, increased weight, and pre-existing conditions such as heart disease, diabetes, and respiratory illnesses like asthma. To prevent heat stress, it is important to hydrate and practice safe sun practices such as wearing a hat and taking planned breaks in the shade.

In addition to heat stress, hot temperatures also increase the risk of other workplace injuries. Dehydration and heat-related fatigue can impair cognitive functions, causing workers to lose focus or have slow reflexes. High temperatures increase the risk of injury due to falls, collisions with an object, overexertion, and more.

In support of the case, we can cite the study that (Sheng et al., 2018) did, the summary of which was that they found a higher risk of work-related injuries due to hot weather. This study provides important epidemiological evidence for policy-makers and industry that may assist in the formulation of occupational safety and climate adaptation strategies.

- Temporary workers

The holiday season can be one of the most profitable times for many businesses, but the increased activity requires employers to hire temporary or seasonal workers to keep up with the demand.

Temporary workers are more likely to be involved in workplace accidents because they are inexperienced and unfamiliar with the company's safety measures. It is important for all employers to provide temporary workers with clear and precise safety policies and ensure they are thoroughly trained on how to operate equipment to avoid injury.

- Fatigue

The New Year period can be a stressful time. In addition to the added responsibilities associated with the holidays, many people work overtime to keep up with production demands, meet end-of-year goals, or earn extra money to help buy gifts. This added stress leaves many workers fatigued.

Fatigue is a common contributor to workplace accidents. Fatigued drivers are more likely to be involved in traffic accidents and can also cause people to become complacent, which is particularly hazardous for workers who operate heavy machinery. Workers need to make sure they are getting proper rest at night and taking scheduled breaks during the workday to avoid making careless mistakes that could lead to serious injury.

- Higher workloads

Part of the reason fatigue can be so common around the New Year period is that most industries are busier during this time. This increased workload can also increase the risk of workplace injury.

When things are rushed, safety procedures may be the first thing out the window. People are trying to get their work done quickly, so they may not take time to make sure they are following full and proper procedures. They may also not pay attention to details that can cause an increased risk of danger.

- Fires

It may seem surprising, but fires are one of the most common sources of workplace injuries. Some offices like to decorate for the holidays, including setting up trees decorated with lights. Unfortunately, these lights can spell danger.

The National Fire Prevention Association estimates that three people die and an additional 34 are injured because of holiday decoration fires every year. Most of these issues are related to electrical outlets being overloaded. Christmas trees can also pose a major fire hazard in the workplace.

5. Potential solutions

After analyzing the data and studying the causes, it is time to research and discuss possible solutions in order to improve safety levels.

The solutions are divided into three categories, security solutions, strategic/policy solutions, and external solutions.

OSH related solutions

These are the measures and procedures that concern the OSH manager and workers, from planning and awareness to actions in the work environment.

- Provide proper clothing for the workers.
- Abiding the use of PPE.
- Necessary breaks and hydration.
- Use Heat Stress Tools to Identify Hazards.

One way to identify heat stressors is by using the following tools:

- OSHA's Heat Safety App
- Heat Stress Monitor
- National Weather Service Heat Index

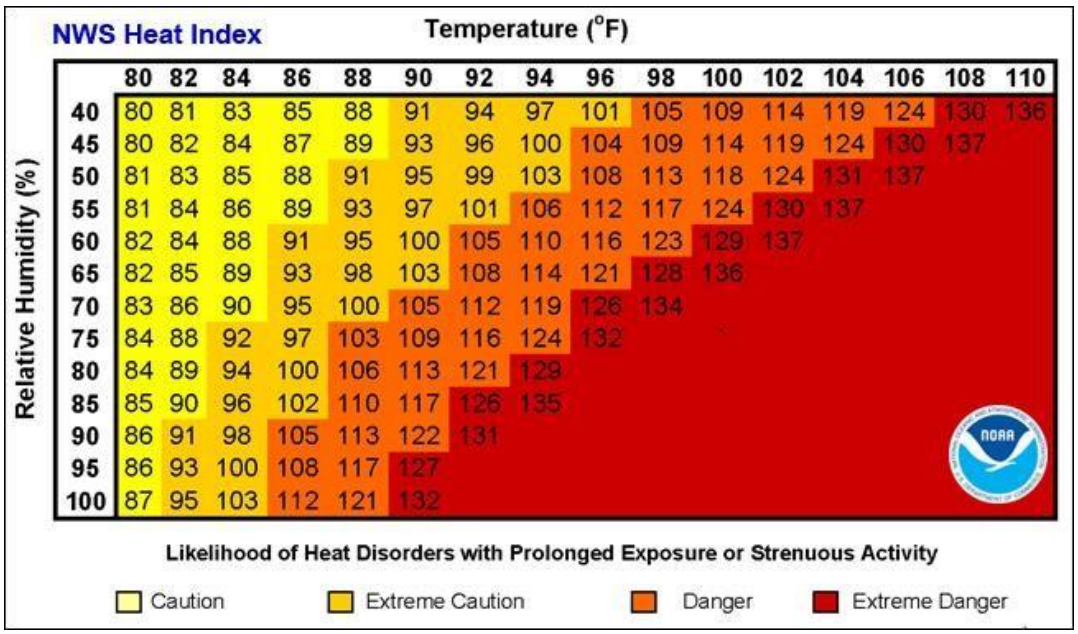

Figure 24 Heat index chart

- Monitor At-Risk Workers

Supervisors should monitor workers who are at risk of heat stress. Workers who are 65 years of age or older, are overweight, have heart disease or high blood pressure, or take medications may be affected by extreme heat.

Strategy/policy related solutions

These solutions concern the HR department and company policies.

- Offer Training

Provide heat-stress training that includes information about worker risk, prevention, symptoms, and the importance of monitoring oneself and coworkers for symptoms.

- Help the occupational health and safety team to instill a culture of "safety first", by enacting strict procedures regarding security violations and providing the necessary equipment.
- Provide adequate training for temporary workers.
- Adjust the appropriate working hours according to the seasonal temperature.
- Providing the appropriate number of labor to carry out the increasing tasks.

External solutions

These are potential solutions related to the global system, such as "global warming" and climate change.

In this regard, (Spector et al., 2019) suggest that we should prioritize methods that reduce local and global disparities and better address the source of heat exposure, including conservation informed land-use planning, built environment, and prevention through design approaches. Participation of occupational health experts in transdisciplinary development and integration of these approaches is needed.

6. Conclusion

We conducted an exploratory analysis of the data using Python; overall, the study was fairly successful as we came up with the following points:

- We cleaned the dataset and make it more useful.
- We had some useful insights and detected patterns in the dataset.
- It was possible to explore the data using Excel, but with Python, the process is easier and fancier, especially for presentations.
- We can do further exploration using NLP analysis. Moreover, even further by applying predictive analysis algorithms.
- We expanded from the insights (calendar related insights) and did a research to find the reasons behind the patterns and the analysis results.
- We concluded the study by providing and discussing the potential solutions to enhance the safety levels.

Limitations

- Fewer features are available in the data set.
- Lack of access to richer data.
- The database goes back four years from the time the study was conducted.

- It takes some time and effort to be comfortable using Python in your data analysis process.
- We do not have direct contact with the company, which prevents us from applying the study in the field.

Chapter VIII General conclusion

In this thesis, we were able to get a general idea about data analysis, occupational safety and health, and the relation between these two fields. As we were acquainted with exploratory and predictive analysis methods, and how the safety specialists can leverage them to enhance their field.

The objective of our study was to find ways that we can exploit the available data to improve the safety and health levels of companies and industries. We applied the exploratory data analysis for the manufacturing plants database using Python.

We achieved most of our goals as we use the available data and extract valuable insights that can be a handful in safety decision-making.

If there is one thing that we can conclude from this study, it is that data analysis has an important role in the modern occupational safety and health field, and it is our responsibility as safety specialists to gain knowledge in this science and implement it in our organizations.

Limitations

On our way to accomplish this research, we encountered many obstacles and limitations, which represent a problem for the field in general, including:

- Lack of access to safety and health databases (especially in our country).
- Neglecting the collection of safety data within organizations and making it available to specialists and researchers.
- As safety professionals, we do not have enough knowledge in the field of data analysis, perhaps because we have not received the necessary training in this important field.

Recommendations

- Giving high priority to collecting safety data and making it available to specialists and researchers.
- Educate and train safety professionals in the science of data analysis, which has become a requirement of the profession.

The results of this research, applied to workplace safety, bring us one-step closer to the vision many of us share of sending every employee home safe, every day. After all, if workplace injuries can be predicted, they can be prevented.

Bibliography

Benjamin O. Alli. (2002). Fundamental principles of occupational health and safety. *Choice Reviews Online*, *39*(07), 39-3997-39–3997. https://doi.org/10.5860/choice.39-3997

Bernardita Calzon. (2022). *Your Modern Business Guide To Data Analysis Methods And Techniques*. The Datapine Blog. https://www.datapine.com/blog/data-analysis-methods-and-techniques/

Cary Usrey. (2016). *The Campbell Institute: What are safety leading indicators?* Www.Safetyandhealthmagazine.Com. https://www.safetyandhealthmagazine.com/articles/13821-the-campbell-institute-what-are-safety-leading-indicators

Deloitte. (2012). Workplace safety analytics: Save lives and the bottom line. *Deloitte.Com*, 1–16.

Don Cameron. (2020). *History of Workplace Healthy and Safety - The Evolution | StaySafe*. Staysafeapp.Com. https://staysafeapp.com/blog/history-workplace-health-and-safety/

Faisal Osama. (2020). *EHS Data Analysis - لحظة السر*. A&F for Safety. https://www.youtube.com/watch?v=S2Yk1HOsqyk

geeksforgeeks. (2021). *Data Mining - Cluster Analysis - GeeksforGeeks*. Geeksforgeeks.Org. https://www.geeksforgeeks.org/data-mining-cluster-analysis/

Harms-Ringdahl, L. (2001). *Safety Analysis Principles and Practice in Occupational Safety* (T. & Francis (ed.)). TAYLOR & FRANCIS.

hse-network. (2020). *3 different predictive analytics models used in safety | HSE Network*. Hse-Network.Com. https://www.hse-network.com/3-different-predictive-analytics-models-used-in-safety/

IBM Cloud Education. (2020). *What is Exploratory Data Analysis? | IBM*. Ibm.Com. https://www.ibm.com/cloud/learn/exploratory-data-analysis

Indeed Editorial Team. (2021). *How To Conduct Exploratory Data Analysis in 6 Steps | Indeed.com*. Www.Indeed.Com. https://www.indeed.com/career-advice/career-development/how-to-conduct-exploratory-data-analysis

insightsoftware. (2022a). *Are You Asking the Right Predictive Questions? - insightsoftware*. Insightsoftware.Com. https://insightsoftware.com/blog/are-you-asking-the-right-predictive-questions/

insightsoftware. (2022b). *What Is Predictive Analytics? - insightsoftware*.

Insightsoftware.Com. https://insightsoftware.com/blog/what-is-predictive-analytics/

James Pomeroy. (2019). *Safety by Numbers: The power of data science to improve performance.* Www.Lr.Org. https://www.lr.org/en/insights/articles/safety-by-numbers/

Lord, D., Qin, X., & Geedipally, S. R. (2021). Exploratory analyses of safety data. *Highway Safety Analytics and Modeling*, 135–177. https://doi.org/10.1016/B978-0-12-816818-9.00015-9

Mark Middlesworth. (2022). *A Short Guide to Leading and Lagging Indicators of Safety Performance*. ErgoPlus. https://ergo-plus.com/leading-lagging-indicators-safety-preformance/

Oracle.com. (2022). *What Is Big Data? | Oracle*. Oracle.Com. https://www.oracle.com/big-data/what-is-big-data/

OSHA. (2022). *Using Leading Indicators to Improve Safety and Health Outcomes*. Www.Osha.Gov. https://www.osha.gov/leading-indicators

Predictive Solutions. (2012). Predictive Analytics in Workplace Safety: Four "Safety Truths" that Reduce Workplace Injuries. *Predictive Solutions*.

Python Software Foundation. (2022). *What is Python? Executive Summary | Python.org*. Www.Python.Org. https://www.python.org/doc/essays/blurb/

Reese, C. D. (2017). Occupational safety and health: Fundamental principles and philosophies. In *Occupational Safety and Health: Fundamental Principles and Philosophies*. https://doi.org/10.1201/b21975

Robyn Correll. (2022). *What Is Occupational Health and Safety?* Verywell Health. https://www.verywellhealth.com/what-is-occupational-health-and-safety-4159865

SAS company. (2022). *Predictive Analytics: What it is and why it matters | SAS*. Www.Sas.Com. https://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Sheng, R., Li, C., Wang, Q., Yang, L., Bao, J., Wang, K., Ma, R., Gao, C., Lin, S., Zhang, Y., Bi, P., Fu, C., & Huang, C. (2018). Does hot weather affect work-related injury? A case-crossover study in Guangzhou, China. *International Journal of Hygiene and Environmental Health*, *221*(3), 423–428. https://doi.org/10.1016/J.IJHEH.2018.01.005

Spector, J. T., Masuda, Y. J., Wolff, N. H., Calkins, M., & Seixas, N. (2019). Heat exposure and occupational injuries: Review of the literature and implications. *Current Environmental Health Reports*, *6*(4), 286. https://doi.org/10.1007/S40572-019-00250-8

tableau.com. (2022). *Time Series Analysis: Definition, Types & Techniques | Tableau*. Tableau.Com. https://www.tableau.com/learn/articles/time-series-analysis

Terence Shin. (2020). *An Extensive Step by Step Guide to Exploratory Data Analysis | by Terence Shin | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-

ddd99a03199e

The R Foundation. (2022). *R: What is R?* Www.r-Project.Org. https://www.r-project.org/about.html

www.frommers.com. (n.d.). *When to Go in Brazil | Frommer's*. Www.Frommers.Com. Retrieved September 11, 2022, from https://www.frommers.com/destinations/brazil/planning-a-trip/when-to-go

www.kaggle.com. (2018). *Industrial Safety and Health Analytics Database*. Www.Kaggle.Com. https://www.kaggle.com/datasets/ihmstefanini/industrial-safety-and-health-analytics-database

Yuli Vasiliev. (2022). Python for Data Science: A Hands-On Introduction. In *No Starch Press*.