Toronto

Module 5 Assignment – Sponsor Project

Dhimahi Patel - 002985259

Parva Patel - 002195186

Parth Savaliya - 002982302

Pratik Malaviya - 002963548

Vishu Sangwan - 002110398

College of Professional Studies**,** Master of Professional Studies in Analytics.

Subject: ALY 6980 Capstone

Current Academic term – 5th Quarter [Spring 2023]

Assignment Completion Date – 9th May 2023

Under the guidance of

Prof. Jay Qi

Team Members	Dhimahi Patel - 002985259 Parva Patel - 002195186 Parth Savaliya - 002982302 Pratik Malaviya - 002963548 Vishu Sangwan - 002110398
Team Lead	Parth Savaliya
Team Member Roles and Responsibilities	Dhimahi Patel - Documentation Parva Patel - Finding datasets, joining table in SQL Query, and finding project proposal question. Parth Savaliya - Help parva and Pratik Pratik Malaviya – Created Exploratory data analysis and data Visualization also, he was involved in joining tables. Vishu Sangwan - Helped Dhimahi in Documentation Every team member is permitted to make suggestions outside of their assigned role.
Mission, Vision Objectives and Core Value	**Mission** - We will try to tackle Sofie's top-level project. **Vision** - To create a predictive modeling strategy for proactively identifying prospective incidents and hazards to prevent them from happening and to respond promptly and efficiently when incidents do happen. **Core** value -The project's fundamental value might be safety. The project aims to make individuals and communities safer by recognizing and resolving occurrences and dangers.
Internal Checks, Balances, and Reviews	**Internal Checks** - Dhimahi and Vishu **Balances** - Parva, Pratik, and Parth **Reviews** - All the team members

Operation: • **Assignments** • **Meetings** • **Communication Guidelines** • **Status Updates** • **Deadlines**	**Assignments -** Team leader would distribute the task among members according to the Calibri of the member. **Meetings -** Meeting will be arranged twice a week according to the availability of the team member and the mode of the meeting would be Microsoft teams. **Communication Guidelines -** All the team members need to communicate in the same language and need to be transparent. Clearly define roles and responsibilities and need to follow that. Encourage all members to communicate openly and provide feedback. **Status Updates -** Each week the different members would update the status of the ongoing project within the group. **Deadlines -** The assignment would be submitted by the team leader before the deadline.

Project Description

The incident and hazard domain of the project involves using machine learning algorithms to predict and prevent incidents and hazards from occurring, as well as resolving them quickly and effectively when they do occur. The analysis of incident trends over time, the prediction of incident likelihood based on historical data and employee/supervisor information, and the identification of patterns in the data to generate suggestions for reducing risks and incidents are all included in projects. Projects also entail defining the different sorts of hazards, the necessary actions, and the time frames for completion. Each initiative aims to make people and communities safer by making sure that events and dangers are handled and resolved in a proactive manner.

Communication Plan

Date	**Communication Method**	**Purpose of Communication**
16th April 2023	Microsoft Team	Discuss the information about our sponsor and what our sponsor does.
19th April 2023	Microsoft Team	Discuss about the main topic which the group would be performing as a final project

20th April 2023	Microsoft Team	Discuss the bibliographies and discuss which all the projects we can cover in the project along with the dataset.
23rd April 2023	Microsoft Team	Finalize the bibliography and give tasks to each team member to sort a few projects which we can share with the professor and our sponsor along with the dataset (tables).
29th April 2023	Microsoft Team	Discussion regarding selection of dataset and business proposal.
4th May 2023	Microsoft Team	Meeting with the sponsor regarding data selection, target variable, EDA and Data Visualization.

Key Stakeholder

Name	Title	Project Role
Workers	Frontline Employees, Operators, Technicians, Staff	Testers and End-Users
Management	CEO, COO, CFO	Providing Resources, Decision-Making, Monitoring
Safety Professionals	Health and Safety Officer	Identifying and Assessing Hazards, Providing Guidance and Input, Monitoring
Regulators	OSHA Representative	Reviewing Compliance, Providing Feedback and Recommendations, Monitoring
Customers	Customer	Providing Feedback, Suggesting Improvements or Changes, Testing
Suppliers	Supplier Representative	Providing Input on Product Safety Features, Minimizing Risks, Advising
Investors	Financial Investor	Advising, Contributing to Financial Viability
Local Community	Community Members	Stakeholders, Ensuring Alignment with Values and Goals

Desired Results and Final Deliverable

The entire team was helpful and encouraging. Every member accomplished their allocated tasks. The datasets were created by merging several pieces of data, yielding 14k rows and 90 columns. Only 24 variables were chosen from among all the variables to complete our project. Team members worked on Eda and data visualization. We also scheduled the meeting with the sponsor. Finally, we were able to select our project topic, "Use machine learning algorithms to predict the completion time for hazards based on historical data."

Potential Timeline:

Week 1: Some research on our sponsor and selecting our topic for the project.

Week 2: Finding some bibliography on our research topic and selecting some project which would be performed in the future.

Week 3: Finalizing the project and the dataset (tables, Schema)

Week 4: Performing Eda, Data cleaning.

Week 5: Data Cleaning

Week 6: Performing Machine learning Model and Data cleaning, Documentation.

Week 7: Performing Machine learning Model and Documentation

Week 8: Final report and Presentation.

Abstract

In this paper, the data understanding of a Sofvie is conducted. This paper presents an in-depth exploration of the workplace safety Hazard dataset of Sofvie, a leading organization in the field of occupational safety. The objective of this study is to gain a comprehensive understanding of the data, uncover meaningful patterns, and derive insights to enhance workplace safety practices and Hazard completion time. The dataset comprises hazard records, employee information, and associated variables collected over a specified time period. The data understanding phase encompasses various techniques and approaches to analyze and interpret the dataset. Descriptive statistics are employed to provide summary measures and gain insights into the distribution, central tendency, and variability of the variables. Data visualization techniques are utilized to visually represent the data, allowing for the identification of trends, patterns, and potential outliers.

Keywords: Data understanding, Hazard, Sofvie, Incident, Employee, Completion Time, workplace safety, descriptive statistics, data visualization, correlation analysis, data profiling.

Data Understanding and Exploration

The objective of the data understanding phase in data exploration is to gain a comprehensive understanding of the dataset by analyzing and interpreting it using various techniques and approaches. This phase involves exploring the data to identify patterns, relationships, potential outliers, and other anomalies that can guide further analysis, feature engineering, or modeling techniques to address specific research questions and objectives. Descriptive statistics, data visualization, and correlation analysis are among the techniques used in this phase to gain insights into the data's distribution, central tendency, variability, and relationships between variables. Overall, the data understanding phase aims to provide a solid foundation for subsequent data analysis and modeling.

In this analysis, the key variables are completion_time_bucket as the target variable and 11 explanatory variables as the main X variables. The explanatory variables include hazard_type, hazard_identification_score, further_action_required, potential_risk_score, complete_action_score, action_status, site, sitelevel, supervisor, workplace, and immediate_action_requirement_and_performed.

Completion_time_bucket is the target variable that represents the time required to complete an action related to workplace safety incidents. The explanatory variables are factors that may potentially influence the completion time. These variables include hazard_type, which refers to the type of hazard involved in the incident, and hazard_identification_score, which measures the degree of hazard identification before the incident occurred.

Further_action_required is an indicator variable that denotes whether additional action was needed after the incident, and potential_risk_score measures the level of risk associated with the incident. Complete_action_score represents the completeness of the action taken in response to the incident, and action_status indicates the current status of the action taken.

Other explanatory variables include site, sitelevel, supervisor, and workplace, which describe the location and managerial aspects of the workplace. Immediate_action_requirement_and_performed are variables that measure the need for immediate action and whether such action was taken.

Data exploration is a systematic process that involves analyzing and understanding a dataset to gain insights into patterns, relationships, and potential influencing factors related to workplace safety incidents. The available schemas and tables in the Sofvie dataset were explored and studied by querying the database catalog or using a database management tool to view the schema structure. The table names, column names, and table descriptions were inspected to identify tables that were specifically related to hazards. The tables that were directly related to hazards were selected, noting their names and relevant columns that may need to be joined or queried later. SQL queries were used to join the selected hazard-related tables based on their common key columns, ensuring the integrity and consistency of the data. Appropriate SQL join operations such as JOIN, INNER JOIN, LEFT JOIN, and others were used depending on the specific requirements and data relationships.

Descriptive statistics such as mean, median, mode, standard deviation, and range were used as a first step in data exploration to provide summary measures to understand the central tendency, spread, and distribution of the data. Data visualization through plots and charts such as histograms, scatter plots, box plots, and bar charts was also employed to provide insights into patterns, trends, and relationships within the dataset. Correlation analysis was utilized to identify the relationship between variables and calculate correlation coefficients such as Pearson's correlation coefficient to indicate the strength and direction of the linear relationship between pairs of variables. This analysis helped to explore the relationships between variables and identify factors that significantly impact hazard and its completion time. Outliers in the data were identified using box plots, which were useful to uncover unusual or erroneous observations that may affect subsequent analysis. During the data exploration process, interesting patterns, relationships, and other anomalies in the data were uncovered. These insights can guide further analysis, feature engineering, or modeling techniques to address specific research questions and objectives.

Exploratory Data Analysis

1. The summary statistics of the quantitative variables are shown in the table below, which was generated using the describe () function in Python. The *Figure 7 Summary* provides information on the mean, minimum, maximum, 25th percentile, 50th percentile (median), 75th percentile, standard deviation, and count for each variable. Additionally, the missing value for each variable is as per the *Figure 8 Null values*.
2. To investigate potential relationships between variables associated with hazard identification and potential risk scores, we did a correlation plot analysis in this study [*Figure 1 Correlation Plot*]. The objective of this study was to find trends and linkages that would shed light on the connections between various variables and assist in guiding risk management tactics.
3. The *Figure 2 Bar Plot* represents the distribution of different hazard types present in the dataset. It shows that the most frequent hazard type in the dataset is 'Stop & Correct', followed by 'Unsafe Condition' and 'Quality'. The chart provides an at-a-glance view of the relative frequency of different hazard types and helps stakeholders identify the most common hazards. By examining the chart, decision-makers can prioritize their risk mitigation strategies and allocate resources accordingly to tackle the identified hazards. The chart also enables stakeholders to recognize any patterns or trends in hazard occurrences, facilitating them in taking appropriate measures to mitigate potential risks effectively. Overall, the bar chart serves as a useful tool to gain a quick understanding of the hazard types' of distribution and can help in making informed decisions to ensure a safe and hazard-free environment.
4. The *Figure 3 Hazard type distribution chart* provides an excellent summary of the frequency and occurrence of different hazard types in the dataset. The visualization highlights the most common hazards and identifies areas that require greater attention or mitigation efforts. The 'Stop & Correct', 'Unsafe Condition' and 'Quality' hazard types are the most frequent, while the counts gradually decrease for the remaining hazard types. This information can guide risk mitigation strategies, resource allocation, and hazard trend analysis. Furthermore, it can facilitate informed decision-making and help ensure effective hazard prevention and management efforts. Overall, the bar chart is a useful tool for stakeholders to gain a comprehensive understanding of the distribution and frequency of different hazard types in the dataset.

5. The *Figure 4 Distribution of hazards* displays the hazard distribution based on the time it takes to complete them. This visualization provides valuable insights into the count of hazards across different completion time ranges. Each bar represents a specific completion time bucket, and its height represents the count of hazards in that bucket. As completion time is the target variable for the machine learning model, this chart is crucial for understanding how completion time affects the count of hazards. By analyzing the chart, we can identify patterns in completion time and prioritize urgent hazards accordingly. It also helps in predicting completion time, which is essential for allocating resources effectively. Furthermore, the distribution of hazards based on completion time can aid in developing effective strategies for managing and mitigating hazards. Overall, this visualization provides a clear picture of how hazards are distributed based on completion time and helps in making data-driven decisions.
6. The box plot is a powerful tool for visualizing the distribution of continuous data, and it has been used here to explore the variation in hazard identification scores based on different hazard types. The plot [*Figure 5 Box plot*] is organized by hazard type, with each box representing the range of scores for that type. The height and spread of each box offer valuable insights into the distribution and variability of hazard scores for each type. By comparing the boxes, we can identify the hazard types with higher or lower median scores and those with greater score variability. This information can help stakeholders understand the differences in hazard identification scores among different hazard types and prioritize efforts to improve hazard identification practices for specific types. Ultimately, this can lead to more effective risk management and mitigation strategies, as stakeholders can allocate resources and focus their efforts on the areas that require the most attention.
7. The histogram displays the distribution of hazard identification scores in the dataset. It provides a visual representation of the concentration and spread of scores across the range of values. The x-axis represents the score range, and the y-axis represents the count of hazards falling within each range [*Figure 6 Histogram Plot*]. By analyzing this histogram, we can gain insights into the distribution of hazard identification scores and their frequency in the dataset. The concentration of scores in specific ranges can indicate potential areas for improvement in hazard identification practices. Additionally, it can help in setting benchmarks and evaluating the effectiveness of hazard identification processes. Histograms can also be used to visualize

the distribution of other continuous variables in the dataset, such as potential risk score and immediate action score. These histograms provide valuable information for understanding the range and distribution of scores and identifying patterns or trends that can inform risk management and mitigation strategies. The histograms in the appendix can be referred to for further analysis of the distribution of other variables in the dataset.

Conclusion

In conclusion, the data understanding and exploration process provide valuable insights into workplace safety hazards and their associated variables. The use of descriptive statistics, data visualization, correlation analysis, and data profiling helped to identify meaningful patterns, trends, and potential outliers within the dataset. The results obtained from this analysis can guide further investigation and modeling techniques to address specific research questions and objectives related to workplace safety.

The correlation plot analysis showed potential linkages between variables associated with hazard identification and potential risk scores, which can guide risk management tactics. The bar charts provided a comprehensive view of the distribution and frequency of different hazard types and helped stakeholders prioritize their risk mitigation strategies and allocate resources accordingly. The distribution of hazards based on completion time aided in developing effective strategies for managing and mitigating hazards and predicting completion time to allocate resources effectively.

The box plot analysis provided insights into the variation in hazard identification scores based on different hazard types, which can guide the development of tailored safety training programs and hazard identification protocols. Overall, the data understanding and exploration process provides a foundation for data-driven decision-making and enhancing workplace safety practices.

References

1. McKinney, W., & others. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56). Retrieved May 7, 2023, from https://conference.scipy.org/proceedings/scipy2010/mckinney.html
2. Home - Sofvie Inc. | Health and Safety Software. (n.d.). Sofvie Inc. Retrieved April 18, 2023, from https://sofvie.com/
3. Matplotlib: A 2D Graphics Environment. (n.d.). Retrieved May 08, 2023, from https://ieeexplore.ieee.org/document/4160265

Annotated Bibliography

Tamascelli, N., Solini, R., Paltrineiri, N., & Cozzani, V. (2022, June). Learning from major accidents: A machine learning approach. Computers & Chemical Engineering, 162. https://doi.org/10.1016/j.compchemeng.2022.107786.

The article "Learning from Large Accidents: A Machine Learning Approach" investigates the application of machine learning methods to analyze data from major accidents and pinpoint underlying trends and causes. The authors contend that conventional accident investigation techniques, which rely on human judgment and knowledge, are constrained in their capacity to pinpoint intricate causes and underlying patterns. The article explains how machine learning may be used to information about accident causes. The authors go over various machine learning methods and how they might be used to analyze accident data, such as decision trees, random forests, and neural networks. They also underline how crucial feature engineering and data pre-treatment are for enhancing the effectiveness of machine models. The article includes two case studies in which accident data were analysed using machine learning. The first case study involves a Canadian train crash, where machine learning was utilized to determine the contributing elements and create a prediction model to avoid similar mishaps in the future. In the second case study, machine learning was utilized to pinpoint the underlying reasons of a gas explosion that occurred in the UK and create suggestions for enhancing safety.

According to the study, machine learning can be an effective tool for accident investigation and prevention since it can spot intricate patterns and contributing elements that might not be immediately obvious using more conventional inquiry techniques. The authors also stress the importance of working together between machine learning algorithms and human specialists to enhance the precision and utility of the analysis. All in all, by offering insights into intricate causal elements and patterns, this work shows the potential of machine learning to enhance accident investigation and prevention. The work emphasizes the value of feature engineering and data treatment in enhancing machine learning model performance and advises more investigation in this area to further improve safety in diverse domains.

Khairuddin, M. Z. F., Lu Hui, P., Hasikin, K., Abd Razak, N. A., Lai, K. W., Mohd Saudi, A. S., & Ibrahim, S. S. (2022). Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. International Journal of Environmental Research and Public Health, 19(21), 13962. https://doi.org/10.3390/ijerph192113962.

A machine learning strategy to reduce occupational injury risks at work is presented in this article, "Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance." The goal of the study is to create a system that can forecast worker injury risk based on numerous environmental conditions and offer suggestions to reduce these risks. The authors explain how they created a machine learning model that can forecast the risk of injury for employees using information from a smart workplace surveillance system. The information covered both worker actions and movements as well as ambient aspects like noise level, temperature, humidity, and illumination. The authors also go over how they developed new features based on domain expertise and selected the most pertinent data points to improve the features utilized in the machine learning model. To get the model's highest prediction performance, they combined feature selection and engineering strategies. According to the study, the machine learning algorithm could correctly forecast a worker's risk of harm based on surrounding characteristics and working activity. Based on the model's projections, the authors also offer suggestions for reducing these risks.

This study demonstrates how machine learning can reduce the risk of occupational accidents at work and increase workplace safety. The strategy utilized in this study could be used to identify and mitigate hazards in real-time in many other sectors, including healthcare and transportation. The study emphasizes the significance of feature optimization in enhancing machine learning models' capacity for prediction and advises additional investigation in this area to further improve workplace safety.

Chu, C., Jain, R., Muradian, N., & Zhang, G. (2016, January). Statistical analysis of coal mining safety in China with reference to the impact of technology. The Journal of Southern African Institute of Mining & Metallurgy, 116. http://www.scielo.org.za/pdf/jsaimm/v116n1/17.pdf.

The safety of coal mining in China is statistically examined in the article "Statistical Analysis of Coal Mining Safety in China with Reference to the Impact of Technology," which also examines the effects of technology on safety outcomes. The study aims to pinpoint patterns in coal mining fatalities and accidents and assess the efficiency of technology in lowering safety risks. The authors explain how they gathered information on coal mining incidents and fatalities in China between 2005 and 2015 and how they utilized statistical methods including regression analysis and time series analysis to spot patterns and trends in the data. By examining statistics on the use of technology in the mining industry, they also looked at the effect of technology on safety results.

Although there has been an overall decrease in coal mining accidents and fatalities in China over the past ten years, the study concluded that there are still very real safety dangers in the sector. The implementation of technology, according to the authors, has improved safety outcomes overall, particularly in the fields of ventilation and gas detection. The authors propose applying machine learning methods to the data to further recognize trends and causal elements causing accidents and fatalities in the mining industry. They also stress the significance of continued review and monitoring of technological adoption to guarantee continuous advancements in safety outcomes. By finding trends and causal factors that contribute to accidents and deaths, this study shows the potential of statistical analysis and machine learning techniques to enhance safety results in the mining industry. The report also emphasizes the value of utilizing technology to lower safety risks and proposes further investigation in this area to further improve safety procedures and avoid accidents in the mining industry.

Bahn, S. (2013, August). Workplace hazard identification and management: The case of an underground mining operation. Safety science, 57, pg 129-137. https://doi.org/10.1016/j.ssci.2013.01.010.

The literature review highlights the unique hazards present in underground mining operations, such as roof falls, explosions, and exposure to harmful substances. The authors discuss the importance of effective hazard identification and management in preventing accidents and injuries and provide an overview of various strategies for achieving this, including hazard identification checklists, risk assessments, and safety audits.

To identify hazards and assess the effectiveness of hazard management measures, the paper employed a combination of interviews with key individuals, site inspections, and document reviews. A hazard matrix was created to categorize dangers based on their likelihood and severity. It was shown that workers are frequently the most aware of the hazards they experience and are best placed to identify and provide solutions to hazards in their workplace. Companies should prioritize worker engagement in the hazard identification and management process and consider developing a hazard reporting system that allows workers to report dangers in real time. The study gives useful insights into the sorts of hazards present in the mining industry and the tactics used to address those hazards. The study emphasizes the need to involve workers in the hazard identification and management process, as well as the requirement for continual hazard monitoring and management to guarantee workplace safety.

APPENDIX

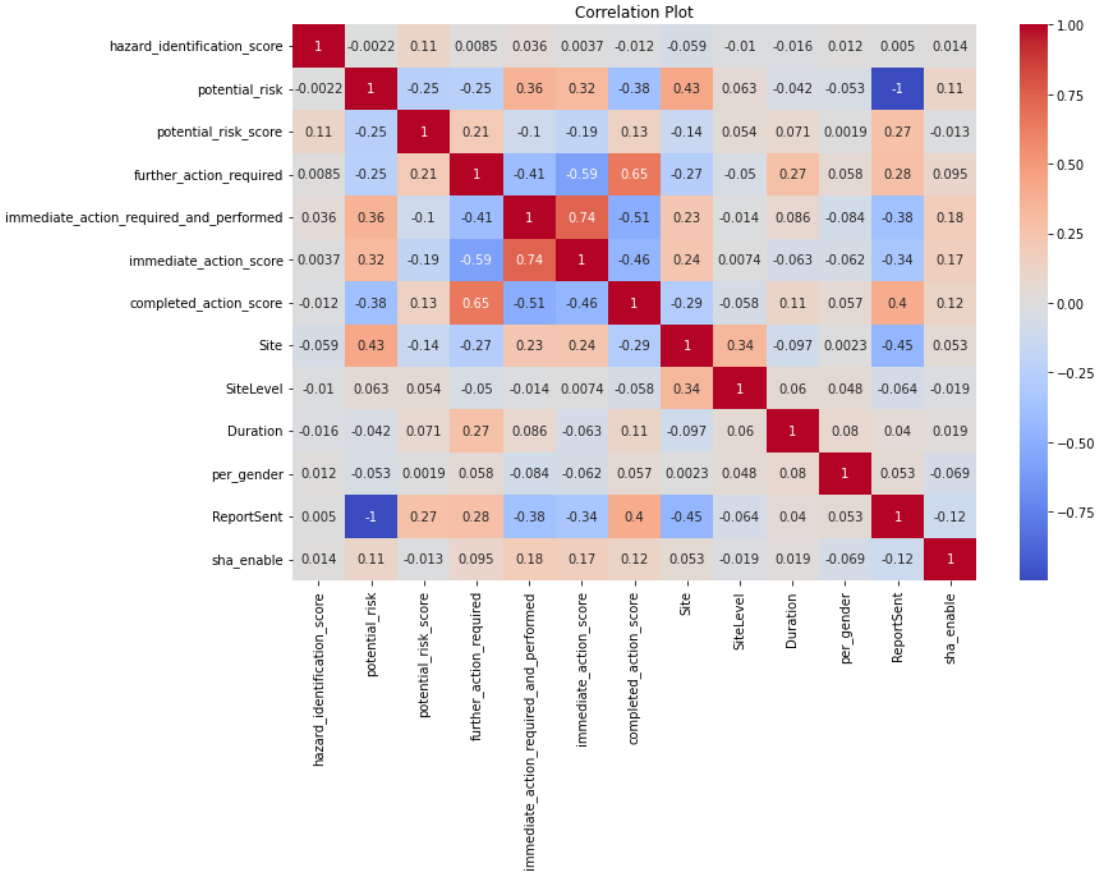

Figure 1 *Correlation Plot*

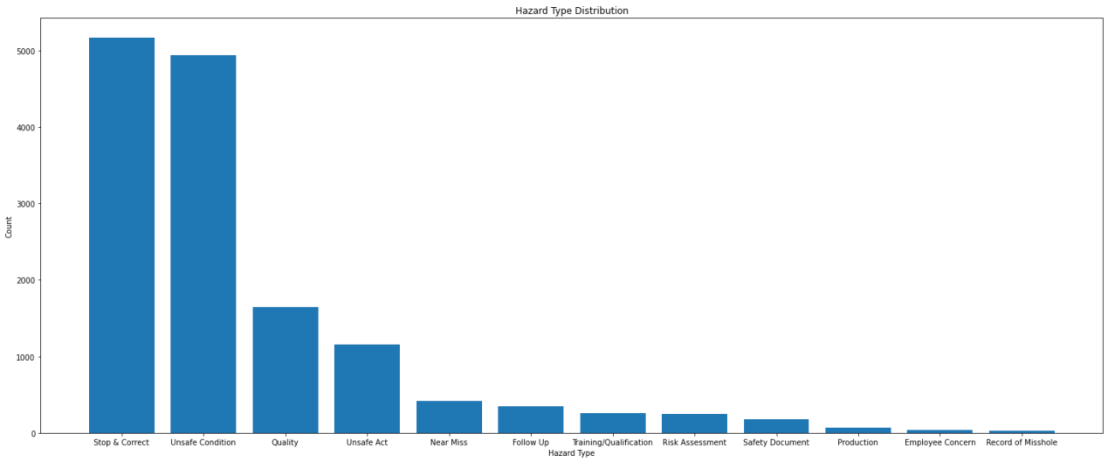

Figure 2 *Bar Plot*

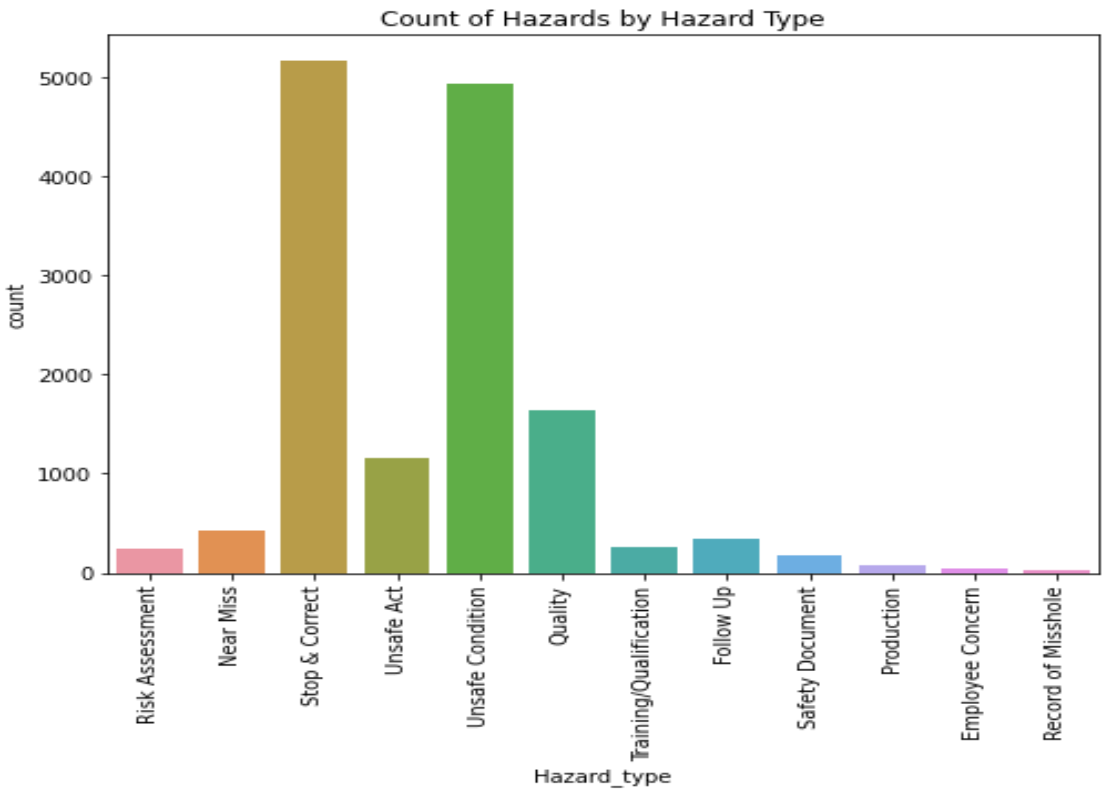

Figure 3 *Hazard type distribution chart*

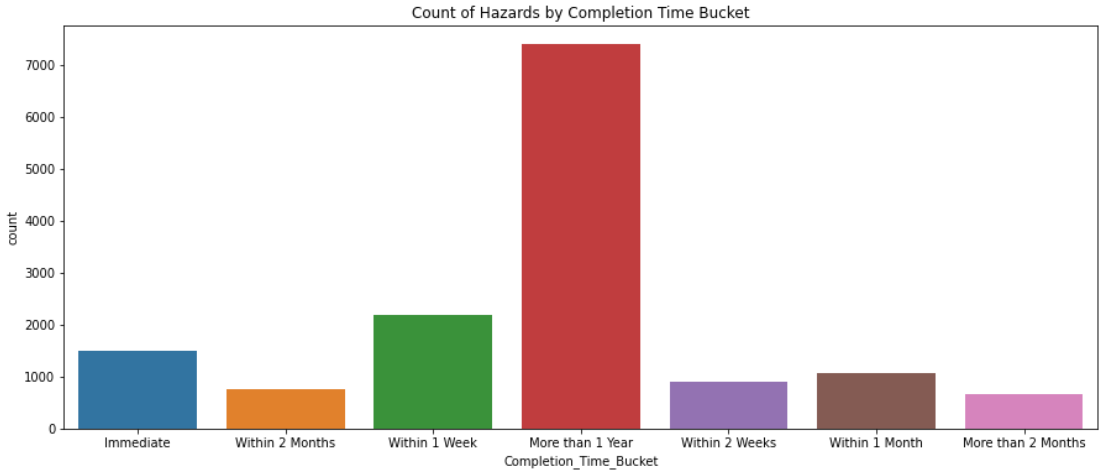

Figure 4 *Distribution of hazards*

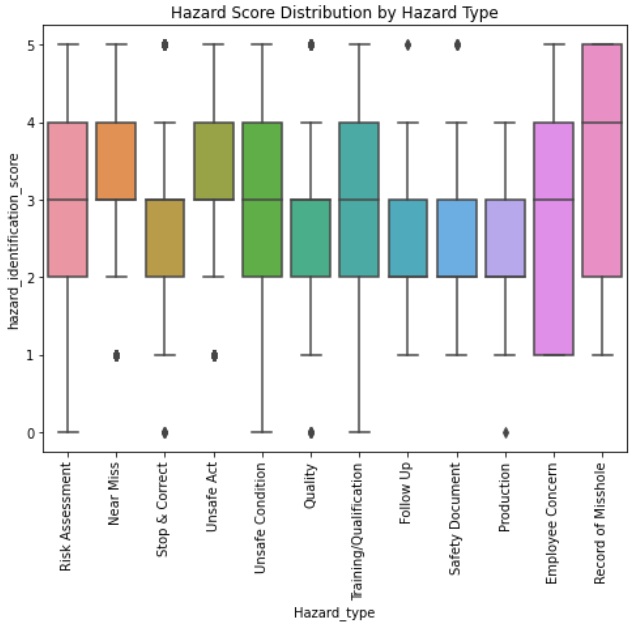

Figure 5 *Box plot*

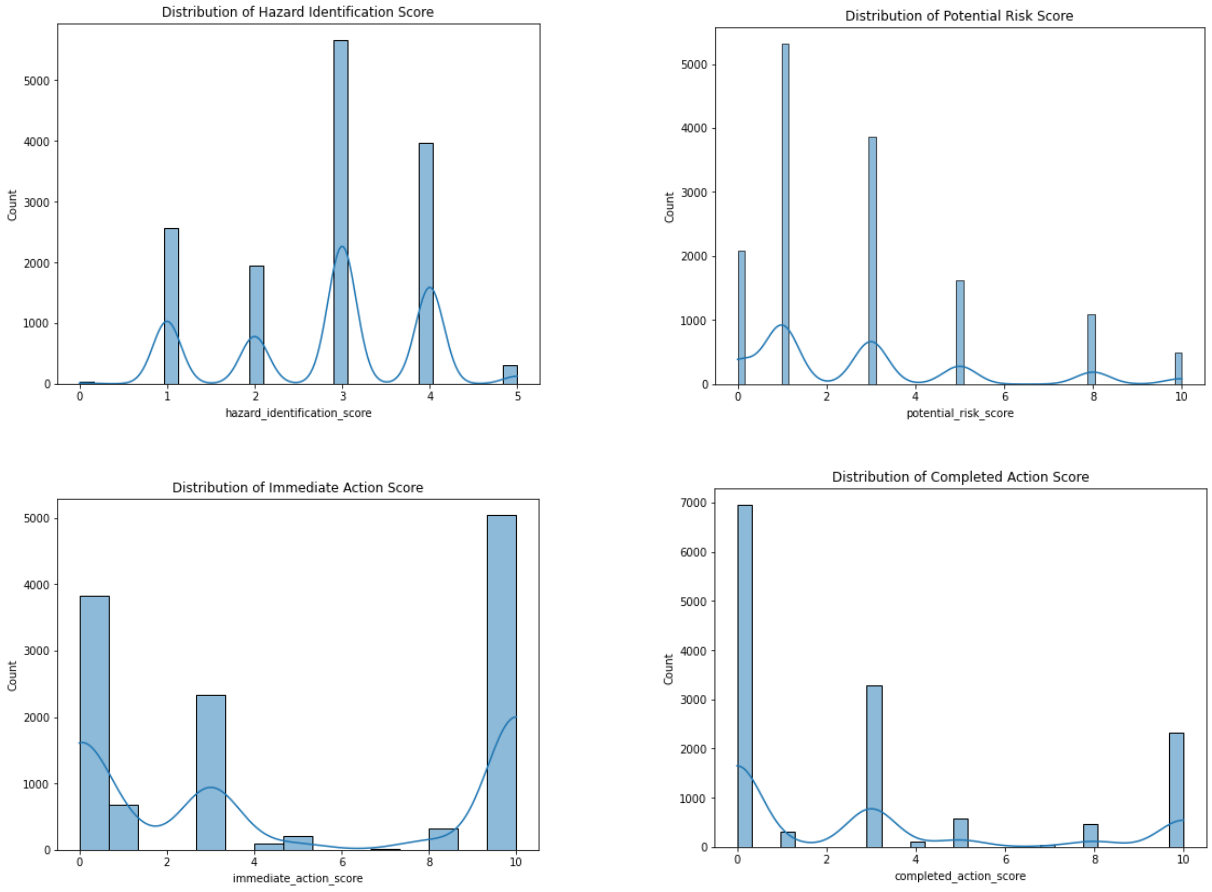

Figure 6 *Histogram plot*

```
# data descirption
hazardData.describe()
```

0.1s Python

	hazard_identification_score	potential_risk	potential_risk_score	further_action_required	immediate_action_required_and_performed
count	14479.000000	14429.000000	14475.000000	14467.000000	14471.000000
mean	2.821466	2517.651535	2.676891	0.393447	0.602101
std	1.091272	444.218443	2.584162	0.488531	0.489481
min	0.000000	2198.000000	0.000000	0.000000	0.000000
25%	2.000000	2199.000000	1.000000	0.000000	0.000000
50%	3.000000	2201.000000	1.000000	0.000000	1.000000
75%	4.000000	3138.000000	3.000000	1.000000	1.000000
max	5.000000	3141.000000	10.000000	1.000000	1.000000

Figure 7 Summary

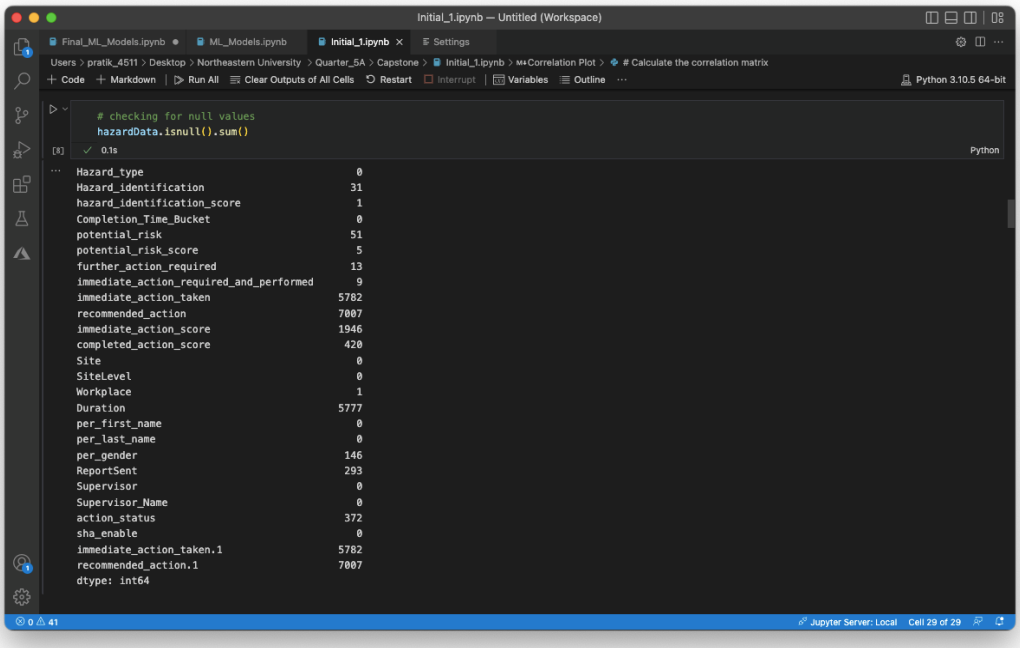

Figure 8: Null Value

Code for gathering data from database.

```
import pandas as pd

data = pd.read_sql("""SELECT haz1.ltr_text AS Hazard_type , haz.ltr_text AS Hazard_identification, hap.hazard_identification_score,

    CASE

        WHEN DATEDIFF(hap.action_completed_date, sh.FormSubmissionDate) < 1 THEN 'Immediate'

        WHEN DATEDIFF(hap.action_completed_date, sh.FormSubmissionDate) BETWEEN 1 AND 7 THEN 'Within 1 Week'

        WHEN DATEDIFF(hap.action_completed_date, sh.FormSubmissionDate) BETWEEN 8 AND 14 THEN 'Within 2 Weeks'

        WHEN DATEDIFF(hap.action_completed_date, sh.FormSubmissionDate) BETWEEN 15 AND 30 THEN 'Within 1 Month'

        WHEN DATEDIFF(hap.action_completed_date, sh.FormSubmissionDate) BETWEEN 31 AND 60 THEN 'Within 2 Months'

        WHEN DATEDIFF(hap.action_completed_date, sh.FormSubmissionDate) BETWEEN 61 AND 365 THEN 'More than 2 Months'

        ELSE 'More than 1 Year'

    END AS Completion_Time_Bucket, hap.potential_risk, hap.potential_risk_score, hap.further_action_required, hap.immediate_action_required_and_performed,

  hap.immediate_action_taken, hap.recommended_action, hap.immediate_action_score,

  hap.completed_action_score, sh.Site, sh.SiteLevel, sh.Workplace, sh.Duration, p.per_first_name, p.per_last_name, p.per_gender, sh.ReportSent, sh.Supervisor,

  CONCAT(p.per_first_name, ' ', p.per_last_name) AS Supervisor_Name, hap.action_status, hap.sha_enable, hap.immediate_action_taken,

  hap.recommended_action



  FROM SubmissionHAP hap

  INNER JOIN SubmissionHeader sh
```

```
    ON hap.SubmissionHeaderID = sh.ID
    JOIN person p
    ON sh.SubmittedBy_SupervisorID = p.per_id
    LEFT JOIN employee e
    ON e.emp_id = p.per_id
    LEFT JOIN (
        SELECT
            DISTINCT hap.hazard_identification,
            lang.ltr_text
        FROM
            SubmissionHAP hap
        INNER JOIN ref_list_detail ref
                ON
            ref.rld_id = hap.hazard_identification
        INNER JOIN language_translation lang
                ON
            lang.ltr_tag = ref.rld_name
            AND lang.ltr_tag_type = ref.rld_tag_type
            AND lang.ltr_lng_id = 1
        ORDER BY
            lang.ltr_text
) haz
        ON
        hap.hazard_identification = haz.hazard_identification
```

```
    JOIN (
        SELECT
            DISTINCT hap.hazard_type,
            lang.ltr_text
        FROM
            SubmissionHAP hap
        INNER JOIN ref_list_detail ref
            ON
            ref.rld_id = hap.hazard_type
        INNER JOIN language_translation lang
            ON
            lang.ltr_tag = ref.rld_name
            AND lang.ltr_tag_type = ref.rld_tag_type
            AND lang.ltr_lng_id = 1
) haz1
        ON
        hap.hazard_type = haz1.hazard_type;""", con=con)
data.head(5)
df2 = pd.DataFrame(data)
print(data)
```