Occupational Hazard Data Analytics:

Predicting Workplace Injury Trends

Veronica Morales

**Table of Contents**

**CHAPTER 1 – INTRODUCTION……………………………………………………………………………………3**

**CHAPTER 2 – LITERATURE REVIEW……………………………………………………………………………6**

**CHAPTER 3 – METHODOLOGY: DATA PREPARATION…………………………………………………12**

**CHAPTER 4 – METHODOLOGY: EXPLORATORY DATA ANALYSIS………………………………..26**

**CHAPTER 5 – METHODOLOGY: MODELING……………………………………………………………….32**

**CHAPTER 6 – MODEL EVALUATION ………………………………………………………………………….43**

**CHAPTER 7 – CONCLUSION………………………………………………………………………………………44**

**REFERENCES ……………………………………………………………………………………………………………48**

Introduction

Background

According to the International Labour Organization (ILO), there is an estimation of 340 million victims of occupational injuries and 160 million victims of work-related diseases each year. The ILO surmises an average of 2.3 million lives lost due to work-related accidents annually. These figures identify a dire need for systemic change to better protect vulnerable workers from preventable incidents and tragedies. Occupational related events impact several ripples of society, starting with individual families, to affecting the productivity of organizations (Khairuddin *et. al.* 2022).

In the United States, the Occupational Safety and Health Association (OSHA), a branch of the U.S. Department of Labor, is responsible for establishing and enforcing safety and health regulations for American employees. OSHA’S main purpose is preventing work related injuries by conducting on-site inspections and developing safety awareness education. According to OSHA, the top three safety and health violations among US employers are found to be insufficient fall protection standards, inadequate hazardous substance communication, and minimal or non-existent respiratory protection (OSHA, 2023). Considering these unfortunate trends, identifying the gaps in occupational hazard prevention through empirical research should be of paramount priority to all organizations.

Health Hazard Analytics

The role of a Health Hazard Analyst is to exhibit the value and the urgency of protecting employees in the workplace by analyzing statistical trends to ultimately reduce occupational dangers. Companies that serve the public with manual labor in industrial settings such as construction, urban development, or factory work have occupational hazards and possess an obligation to their employees to deploy health hazard prevention research. Health Hazard analytics would be beneficial in identifying gaps in safety resources so that policies can be implemented that would prevent occupational injuries and fatalities. Companies have much to gain from investing in health hazard analytics in the form of ensuring the safety of their employees, protecting themselves from litigation, and avoiding hefty fines due to OSHA/Workman’s Compensation Board violations.

The information found in the *Assembled Worker’s Compensation Claims* by Selected Characteristics would be used to identify gaps in occupational resources so that policies can be implemented that would prevent occupational fatalities and injuries. The Workman's Compensation Board (WCB) is a government organization with several jurisdictions and serves the oversight of worker’s benefits and compensation entitlements. WCB services are provided across a vast range of labor industries, including construction, healthcare, manufacturing, retail, hospitality, and transportation. Employees that are injured or develop an illness from their employment setting must submit a claim to the WCB to be compensated for medical costs, time missed at work, or potential permanent disability benefit arrangements. The claim must include details about the injury incident and any medical attention that has occurred. Once the claim paperwork has been properly submitted, the WCB will conduct an investigation on the incident and make a judgment on the compensational benefits owed. Workers that are generally not eligible for WCB services include some small businesses employees, seasonal workers in agriculture and farming, domestic workers, independent contractors, and volunteers/ unpaid workers. The WCB’s main purpose is to oversee and provide resources and support to workers that become injured or ill due to their employment. Some of the main tasks of the WCB include benefit provision, dispute resolution, and employer compliance.

Research Problem and Research Question

According to the data documenting occupational injuries by the WCB, workplace injuries are a statistical certainty. However, by following the trends of injuries by demographics, occupations, and incidence rates, we can observe which occupations need further research to identify precautions and implement safety protocols. The research problem is identifying gaps in current occupational safety protocols by using machine learning practices to analyze workplace injury data. The goal is to minimize workplace injuries and fatalities while ultimately preserving the health and wellness of employees. The extensive dataset of WCB injury claims provides a valuable platform for addressing essential questions that can contribute to the improvement of public policy and safety protocols. By analyzing injury patterns and identifying areas in need of improvement, this research can propel meaningful change in safety practices.

**Research question:** *Is it possible to predict the probability of severe injuries? Which demographical and nature of injury factors contribute to likelihood of severe injury?*

**Hypothesis:** Demographical and nature of injury data can be used to classify the severity of a WCB injury claim.

The research question and hypothesis are highly plausible, supported by the dataset, *Assembled Worker’s Compensation Claims* obtained from the WCB. Utilizing the variety of available data, machine learning models will be developed to identify connections and patterns between injury severity and exploratory variables such as injury type, physical labor levels, and body part injured. By establishing the potential to predict severe injuries and confirming the hypothesis that demographical and nature of injury data can be used to classify injury severity, companies are propelled toward improving occupational safety practices. With insights obtained from the research, companies can identify weaknesses in their safety policies while enhancing preventative measures and ultimately reducing occupational injuries.

**Literature Review**

Exploring and analyzing workplace injury trends is a crucial practice in addressing and combatting social issues through a scientific discipline using quantitative research methods. Numerous methodological approaches including survey research and machine learning models have been developed to analyze occupational injury trends throughout different market segments. Analyzing workplace safety statistics serves infinite purposes and spans through several different organizations including large businesses to government agencies. The goals and intentions of workplace safety research may vary from cost-management, as high injury incidents rates may be costly to organizations, to social justice efforts to improve the working conditions of physical laborers. Regardless of the intention, various entities have approached workplace safety data in interesting ways. Existing occupational hazard research has established the importance of understanding the trends that lead to workplace injuries. By using the current literature as a platform, the research question, **“***Is it possible to predict the probability of severe injuries? Which demographical and nature of injury factors contribute to likelihood of severe injury?”* can be addressed.

Cross-sectional Commonalities in Occupational Hazard Research

An important element to understand regarding workplace injury data, is that most organizations acquire their data from government sources. A common source for workplace injury data is the Bureau Labor Statistics (BLS) and its child organization, the Occupational Health and Safety Administration (OSHA). According to the U.S. Bureau of Labor Statistics (BLS), in 1970, the Occupational Safety and Health Act was passed by President Nixon to "preserve our human resource" by ensuring "healthful working conditions" (BLS, 2020). To accomplish its goals, the OSH Act propelled the development of the BLS to collect, maintain, and analyze occupation and health statistics. The main purpose of establishing the BLS was to utilize data to improve national occupational injury statistics. For the last 50 years, researchers, regulation agencies, and policy makers have utilized the data to identify, address, and enhance workplace safety measures. When the BLS was established, injury incidence rates were at 10.9 cases per 100 fulltime laborers. The incidence rate has steadily improved to 2.8 cases per 100 workers in 2018 (BLS, 2020). The success of the BLS's efforts demonstrates that quantitative methods are vital to the collection of workplace safety research.

Machine Learning Models in Health Hazard Research

A comprehensive and thorough example of machine learning models analyzing workplace injury outcomes, is the medical journal, *Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance* written by Mohamed Zul Fadhli Khairuddin *et. al.* This study uses OSHA occupational data to forecast severity because of workplace injuries. The machine learning models utilized in this study include support vector machines, K-Nearest Neighbors, Naïve Bayes, Decision Tree, and Random Forests to predict injury severity (Khairuddin *et. al.* 2022).

In *Learning Approach and Feature Optimization for Smart Workplace Surveillance,* The methodology of variable selectionfor the development of the machine learning models demonstrates the effectiveness of using demographical data, incidents report data, and nature of injury data. The researchers ultimately decided that the most important features were “nature of injury”, “type of event”, and “affected body part.” The medical journal compares the measures of each machine learning model to identify the best performer to address the research goal (Khairuddin *et. al.* 2022). From the models designed, the best performing machine learning model was the random forest model (Khairuddin *et. al.* 2022). The research study is a great example of the process of data management, model construction, and evaluation. This study will be used as a framework for the research of this project.

Business Applications of Research Methods

There are several questions that researchers ask when approaching workplace injury data. Several studies conducted, focus on corporate outcomes to occupational injuries, such as, employee recovery time, absenteeism, financial consequences, and public image outcomes. An example of a research study conducted to address corporate impacts after an injury, is *How Workplace Injuries Impact your Business* by HR service company, 02 Employment Services (Corrigan, 2018)*.* This study conducts several methods of statistical analysis to estimate the loss of productivity and financial burden that companies have because of workplace injuries (Corrigan, 2018)*.*  There are several research efforts dedicated to risk management and minimizing financial consequences of lack of safety efforts.

A notable attribute of the data analysis conducted by 02 Employment Services is the variable selection to answer their research questions. Outcome variables explored by this analysis included Lost Productivity, Cost to Replace Workers, and OSHA Fines (Corrigan, 2018)*.* 02 Employment services performed a variable selection that was resourceful of their company data and ultimately used variables such as Workplace Injury Occurrence, Safety Vulnerabilities, Staffing Levels, and Staff Experience Levels (Corrigan, 2018)*.*

The Social Science and Medicine book Vol. 269 published in 2021, includes the journal, *The Future of Research on Work, Safety, Health, and Wellbeing: A Guiding Conceptual Framework* by authors Glorian Sorenson et. al. The research in this journal is based on the belief that "work plays a central role in determining health," and that "work related injuries and illnesses can lead to chronic disease" (Sorenson, 2020). The purpose of the journal is to contribute to the growing amount of research that aims to understand the key factors between working conditions and health by designing a conceptual framework model for the use of building research hypothesis. The conceptual framework model implies that physical and organizational conditions, such as workplace policies, impact employee well-being, which ultimately effects company metric outcomes such as turnover rates, healthcare costs, and overall productivity. External factors are also considered, such as social, political, and economic environment, and employment & labor patterns.

Some companies may be interested in enhancing workplace safety for more empathetic reasons, such as enhancing employee satisfaction. Alert Media company is a data-driven company that conducts an aggressive acquisition of employee satisfaction data. As it is challenging to acquire quantitative satisfaction data, they conduct survey research in the form of occupation safety related questioning with a numerical ranking system measuring their response (Alert Media, 2022). Some of the most impactful statistics acquired in the surveys include:

* 83% of employees have experienced an emergency while working.
* 84% of employees believe that their employers aren’t doing enough to prepare them for emergencies.
* 40% of employees believe that safety training should be often and mandatory.
* 58% of employees don’t feel that their employers have made improvements in efforts to enhance their safety.
* 46% of surveyors are more likely to stay at a workplace where they feel that employers care about their safety regardless of economic factors (Alert Media, 2022).

The importance of these findings is that workplace emergencies are a statistical certainty and American workers do not feel prepared to respond to such emergencies. Large samples of workers do not feel prepared to respond to workplace emergencies, and they believe that their employers are not doing enough to prepare them (Alert Media, 2022). In addition to lacking preparedness to respond to emergencies, workers do not feel as though their employers are making efforts to elevate worker safety in ways such as safety education and training(Alert Media, 2022).

The research assessing the sentiments of workplace safety by workers is important to the compilation of workplace safety literature because it identifies gaps in safety measures that can only be identified through exploring the experiences of individual workers. The research also establishes an urgency for employers to address employee concerns so that they can preserve employee satisfaction and ultimately reduce expensive turnover rates. According to Alert Media, *"Business continuity depends on employee safety."* An interesting development of this data would be to conduct multivariate analysis with other attributes of their occupational safety and injury data.

Gaps in the Research

This literature review aimed at identifying the factors assessed in modern workplace injury statistics, as well reviewing quantitative, and mixed method approaches used to conduct research. Some of the gaps in the research included quantitative approaches to measuring the impacts that workplace injuries have on the financial well-being of workers. Other gaps in the research included lack of analysis regarding worker demographics to identify groups of workers that are most vulnerable to injuries. Demographical factors may include age or sex. To address and combat workplace injury occurrences, it is important to explore which industries, work conditions, and business practices contribute to injuries and fatalities but research in these areas were minimal in the last 7 years.

Ultimately, the literature review provides a platform to address certain gaps in the research as well as providing an informed understanding of the industrial environment in which workplace injuries occur. The benefit of pursuing a more thorough analysis of demographical and conditional factors as predictors of workplace occurrences is that identifying key factors can help policy makers and employers develop parameters for safer work practices, ultimately enhancing work conditions. The literature evaluated in this report establishes a foundation of valuable research and effective methodology that contributes to future analysis. Methodology must continuously evolve to address everchanging trends and practices in workplace settings, which results in high demand for up-to-date research.

**Methodology: Data Preparation**

**Software**

Data preparation, manipulation, exploration, and modeling was conducted using **SAS JMP Pro 17** and Tableau 2022. SAS JMP Pro markets itself as a *"predictive analytics software for scientists and engineers"* (JMP, 2023). JMP capabilities include advanced statistical modeling, predictive modeling, machine learning, data exploration, and visualization. Tableau is a visual analytics platform with multiple data analytics tools. For this project, Tableau was used for exploratory data analysis and data visualization.

**Data Collection**

The dataset, New York State’s Assembled Workers’ Compensations Claims

was collected from Data.NY.Gov ( <https://data.ny.gov/Government-Finance/Assembled-Workers-Compensation-Claims-Beginning-20/jshw-gkgu> ). The website contains a sophisticated built-in querying program that allows you to filter and download the data that you want to use. Data from January 2018 to December 2022 was filtered and downloaded for this study. This time frame was selected to capture the most recent injury trends within the parameters of pre-covid and post-covid work environments.

|  |  |
| --- | --- |
| **Dataset Summary** | |
| Organization | NYS WCB Data Management & Analytics |
| Dataset Owner | New York State Workers’ Compensation Board |
| Coverage | Statewide |
| Time Period | Beginning year 2,000 |
| Date Created | 4-Dec-14 |
| Posting Frequency | Quarterly |
| Rows | 4.77M |
| Rows from 2018-2022 | 1.3M |
| Columns | 54 |

**About the Data**

The dataset collected from the New York State’s Workman’s Compensation Board, titled Assembled Worker’s Compensation Claims documents the injury claims filed to the WCB. The dataset was created by the NYS WCB Data Management & Analytics team in 2014 and documents data starting in the year 2,000 and is updated daily. According to the WCB, the data is obtained from the boards Enterprise Data Warehouse (EDW) where the data is transformed and cleaned to obtain reliability and accuracy. The dataset consists of 54 columns and 4.71 million rows of injury claims. The dataset will be condensed methodologically which will ultimately reduce the number of rows used in the research project.

The dataset consists of three different types of data: demographical data, nature of the injury, and resolution outcomes. The demographical data includes variables/characteristics of the injured workers, such as age, sex, the worker’s weekly wage, and the industry in which the individual worked. Demographical data is important because it can be used to identify at-risk individuals, and weaknesses in safety protocols.

Nature of injury data is very important to the dataset because it documents characteristics and elements of the incidents of the injuries. By analyzing nature of injury data, organizations can better understand the injury patterns, their causes, identify common injury types, and risk factors. Understanding this information can be essential for policy makers and organizations in strategizing and implementing prevention and safety measures in workplace environments. In this dataset, variables of nature of injury data include injury type, part of body injured, cause of injury, injury source, and event exposure.

Resolution outcome data can be very useful in assessing the severity of an injury and whether the appropriate and fair protocols were conducted. Resolution outcome data can be used by organizations and government agencies to anticipate injury costs, obtain the appropriate insurance type, develop prevention strategies, and develop response protocols for the events of an injury. In this dataset, resolution outcome data is represented by the following variables: whether an attorney was obtained, alternative dispute resolution, current claim status, court hearing info, and appeals.

**Data Wrangling:**

**Eliminating Columns, Rows, and Outliers**

Several variables were initially eliminated due to the entirety of the column being blank, or most of the inputs being blank. It is assumed that when the injury claims were submitted, either these portions of the paperwork were not required, or they were simply not loaded into the data base.

|  |  |
| --- | --- |
| **Eliminated Columns** | **Reason for Elimination** |
| ANCR Date | Majority of cases left blank in this field |
| C2 & C3 dates | Not useful and most cases left blank |
| IME-4 Count | Not useful and most cases left blank |
| Interval assembles to ANCR | Not useful and most cases left blank |
| Controverted Date | Most cases left blank |
| Occupational Injury and Illness Classification system (OIICS) (14 rows) | Entirety of 14 columns left blank. Likely that OIICS data was failed to be synced from another data source. |

Outliers in numerical variables and observations with several incomplete rows were eliminated. the AGE column, outliers included workers over the age of 100, with the highest data point being 117. For this study, the age parameter is set to 100 which resulted in 114 rows being eliminated. There were several age inputs at 0 years of age, along with several age inputs between 1 and 16. According to the New York State Department of Labor, minors under 17 require special permits to work, so the age limit for this study is set to 17. 4,962 rows were eliminated for age inputs between 0 and 17, accounting for 3.5% of the data.

**Data Reduction**

Data reduction is the process of minimizing the data needed to represent important information while preserving the fundamental attributes. The practice of data reduction is necessary in this study due to large number of categories representing certain variables. Empirical evidence implies that recategorizing these values to focus on relevant distinctions of the overall injury, rather than focusing on scarce representation between sub-categories, will improve the overall interpretability and accuracy of the data analysis (Khairudden et. al. 2022). Data reduction was also used to address redundancy within the coding of the data such as combining repetitive categories like “hearing loss” and loss of hearing”.

The Cause of Injury data had 74 categories describing causes of injury. Data reduction was performed on the Cause of Injury Data for the purpose of enhancing statistical analysis, improving visualization, and increasing interpretability. To reduce and simplify this column, the variables were condensed into 8 categories. The 8 categories were obtained from the BLS. The BLS has determined that all workplace injuries can be grouped into the following broad categories. The descriptions of the injuries provided by the WCIO were used to determine which categories each value belonged to.

|  |  |
| --- | --- |
| **CAUSE OF INJURY DATA** | |
| **New Values (9)** | **Old Values (74)** |
| VIOLENCE/ OTHER INJURIES BY PERSONS OR ANIMALS | ·    ANIMAL/INSECT |
| ·    PERSON IN ACT OF CRIME |
| ·    GUNSHOT |
| ·    TERRORISM |
| ·    FELLOW WORKER/PATIENT/OTHER PERSON |
| ·    OBJECT HANDLED BY OTHERS |
| OVEREXERTION AND BODILY CONTACT | ·    PUSHING/PULLING |
| ·    TWISTING |
| ·    REPETITIVE MOTION |
| ·    REACHING |
| ·    JUMPING/LEAPING |
| ·    LIFTING |
| ·    OBJECT BEING LIFTED/JHANDLED |
| ·    OBJECT HANDLED |
| ·    HOLDING OR CARRYING |
| ·    WIELDING OR THROWING |
| ·    WELDING OPERATING |
| CONTACT WITH OBJECTS AND EQUIPMENT | ·    HAND TOOL; UTENSIL |
| ·    POWERED HAND TOOL/APPLIANCE |
| ·    MACHINE/MACHINERY |
| ·    HAND TOOL/MACHINE IN USE |
| ·    USING TOOL/MACHINE |
| ·    MOVING PARTS OF MACHINE |
| ·    SANDING, SCRAPING, CLEANING OPERATION |
| ·    CAUGHT IN, UNDER, OR BETWEEN |
| ·    INTO OPENING |
| ·    CUT, PUNCTURE, SCRAPE |
| ·    BROKEN GLASS |
| ·    RUBBED OR ABRADED |
| ·    STRUCK OR INJURED |
| ·    STATIONARY OBJECT |
| ·    CONTACT WITH NOC |
| ·    COLLISION WITH A FIXED OBJECT |
| ·    STRIKING AGAINST OR STEPPING ON |
| ·    STEPPING ON SHARP OBJECT |
| ·    FALLING OR FLYING OBJECT |
| ·    COLLAPSING MATERIALS |
| EXPOSURE TO HARMFUL SUBSTANCES OR ENVIRONMENTS | ·    CONTINUAL NOISE |
| ·    ELECTRICAL CURRENT |
| ·    HOT OBJECTS/SUBSTANCES |
| ·    STEAM/HOT FLUIDS |
| ·    TEMPERATURE EXTREMES |
| ·    COLD OBJECTS/SUBSTANCES |
| ·    ABNORMAL AIR PRESSURE |
| ·    FOREIGN MATTER IN EYES |
| ·    FROM LIQUID OR GREASE SPILLS |
| ·    NATURAL DISASTERS |
| ·    PANDEMIC |
| FALLS/TRIPS | ·    FALL, SLIP, OR TRIP |
| ·    ON SAME LEVEL |
| ·    ON ICE/SNOW |
| ·    ON STAIRS |
| ·    FROM DIFFERENT LEVELS |
| ·    SLIP OR TRIP (DID NOT FALL) |
| FIRE/EXPLOSION | ·    FIRE/FLAME |
| ·    EXPLOSION/FLARE BACK |
| HAZARDOUS/CHEMICAL SUBSTANCE | ·    ABSORBTION, INGESTION, INHALING |
| ·    CHEMICALS |
| ·    DUST, GASES, FUMES, OR VAPORS |
| ·    MOLD |
| ·    RADIATION |
| OTHER/MISC. | ·    OTHER MISC |
| ·    OTHER THAN PHYSICAL CAUSE OF INJURY |
| TRANSPORTATION INCIDENTS | ·    COLLISION/SIDESWIPE WITH ANOTHER VEHICLE |
| ·    MOTOR VEHICLE NOC |
| ·    MOTOR VEHICLE |
| ·    VEHICLE UPSET |
| ·    CRASH OF RAIL VEHICLE |
| ·    CRASH OF WATER VEHICLE |
| ·    CRASH OR AIRPLANE |

Another variable column with several categories is the Nature of Injury column with 56 values. This column was also recoded and re-organized for simplification with the purpose of enhancing statistical analysis, improving visualization, and improving interpretability. This column was successfully reduced from 56 values to 25 values while consolidating the causes of injury. Because these categories address medical terms, to promote accurate categorization, the recoding was performed using descriptions of each medical term provided by the Workman’s Compensation Insurance Organization Board.

|  |  |
| --- | --- |
| **NATURE OF INJURY DATA** | |
| **New Values (25)** | **Old Values (56)** |
| BURN | ·    BURN |
| CANCER | ·    CANCER |
| CARDIOVASCULAR | ·    SYNCOPE |
| ·    MYOCARDIAL INFARCTION |
| ·    ANGINA PECTORIS |
| ·    VASCULAR |
| CARPEL TUNNEL SYNDROME | ·    CARPEL TUNNEL SYNDROME |
| CONCUSSION | ·    CONCUSSION |
| CONTUSION | ·    CONTUSION |
| CRUSHING | ·    CRUSHING |
| DERMATITIS | ·    DERMATITIS |
| DISLOCATION | ·    DISLOCATION |
| DISMEMBERMENT | ·    LACERATION |
| ·    AMPUTATION |
| ·    SEVERENCE |
| ·    ENUCLEATION |
| ELECTRIC SHOCK | ·    ELECTRIC SHOCK |
| FOREIGN BODY | ·    FOREIGN BODY |
| FRACTURE | ·    FRACTURE |
| HERNIA | ·    HERNIA |
| INFECTIOUS DISEASE | ·    COVID-19 |
| ·    INFECTION |
| ·    CONTAGIOUS DISEASE |
| ·    AIDS |
| ·    HEPATITIS C |
| INFLAMMATION | ·    INFLAMMATION |
| MENTAL INJURY | ·    MENTAL STRESS |
| ·    MENTAL DISORDER |
| ·    NO PHYSICAL INJURY |
| OTHER/MISC. | ·    ALL OTHER SPECIFIC INJURIES |
| ·    MULTIPLE PHYSICAL INJURIES ONLY |
| ·    ALL OTHER CUMMULATIVE |
| ·    ALL OTHER OCCUPATIONAL DISEASE |
| ·    MULTIPLE INJURIES |
| ·    ADVERSE REACTION TO VACCINATION |
| ·    RADIATION |
| ·    VDT RELATED DISEASES |
| POISIONING | ·    POISONING- GENERAL |
| ·    POISONING- CHEMICALS |
| ·    POISONING- METALS |
| PUNCTURE | ·    PUNCTURE |
| REPIRATORY | ·    RESPIRATORY DISORDERS |
| ·    ASBESTOSIS |
| ·    DUST DISEASE |
| ·    ASPHYXIATION |
| ·    SILICOSIS |
| ·    BLACK LUNG |
| ·    BYSSINOSIS |
| RUPTURE | ·    RUPTURE |
| SPRAIN OR TEAR | ·    SPRAIN OR TEAR |
| ·    STRAIN OR TEAR |
| THERMAL INJURY | ·    HEAT PROSTRATION |
| ·    FREEZING |
| VISION/HEARING LOSS | ·    HEARING LOSS OR IMPAIRMENT |
| ·    LOSS OF HEARING |
| ·    VISION LOSS |

The last column to be significantly consolidated was the Injured Body Part column. Each body part was recategorized using the same criteria proposed in the medical publication, *Occupational Injury Risk Mitigation* (Khairudden et. al. 2022), where “Injured Body Part” was used as a predictor in their machine learning model to predict injury severity.

|  |  |
| --- | --- |
| **INJURED BODY PART** | |
| **NEW VALUES (8)** | **OLD VALUES (55)** |
| HEAD | ·    EYES |
| ·    MULTIPLE HEAD INJURY |
| ·    SKILL |
| ·    FACIAL BONES |
| ·    EARS |
| ·    NOSE |
| ·    MOUTH |
| ·    BRAIN |
| ·    TEETH |
| INTERNAL ORGANS | ·    LUNGS |
| ·    INTERNAL ORGANS |
| ·    HEART |
| ·    TRACHEA |
| LOWER EXTREMITIES | ·    KNEE |
| ·    ANKLE |
| ·    FOOT |
| ·    UPPER LEG |
| ·    TOES |
| ·    MULTIPLE LOWER EXTREMITIES |
| ·    GREAT TOE |
| MISC. | ·    NO PHYSICAL INJURY |
| ·    ARTIFICIAL APPLIANCE |
| MULTIPLE BODY PARTS | ·    MULTIPLE BODY PARTS |
| ·    MULTIPLE BODY SYSTEMS |
| ·    WHOLE BODY |
| NECK/THROAT | ·    MULTIPLE NECK INJURY |
| ·    LARYNX |
| TRUNK | ·    LOWER BACK AREA |
| ·    CHEST |
| ·    UPPER BACK AREA |
| ·    ABDOMEN INCLUDING GROIN |
| ·    HIP |
| ·    DISC |
| ·    MULTIPLE TRUNK |
| ·    BUTTOCKS |
| ·    PELVIS |
| ·    SACRUM AND COCCYX |
| UPPER EXTREMITIES | ·    FINGERS |
| ·    SHOULDERS |
| ·    HAND |
| ·    WRIST |
| ·    LOWER ARM |
| ·    THUMB |
| ·    LOWER LEG |
| ·    UPPER ARM |
| ·    ELBOW |

In addition to the other recoding efforts, the ***AGE AT INJURY*** values were divided into age groups from 17-19, 20-29, 30-39, and so on until the age limit of 99. This reduction method is meant to simplify the data analysis and visualization for age as a descriptor variable, ultimately enhancing predictability of exploratory analysis and machine learning models.

**Variable Creation Methods**

Variable creation methods were utilized in this study by generating new variables using existing ones. The dataset did not include an official ranking system to measure the severity of an injury. However, the dataset does include the variable “Claim Injury Type”, which categorizes the long-term medical consequence that the injury had on the employee’s ability to continue working. The variable “Claim Injury Type” is divided into five categories that are described as follows:

|  |  |
| --- | --- |
| Claim Injury Type | |
| **Med Only** | ·    Requires medical treatment but does not result in disability or loss of bodily functions. |
| ·    Implies little to no time lost in the workplace. |
| **Temporary** | ·    Requires medical treatment. |
| ·    Results in temporary loss of work. |
| ·    Expected full recovery |
| **Permanent Partial Disability** | ·    Requires medical treatment. |
| ·    Permanent loss of partial bodily functions. |
| ·    Permanently limits an employee’s capacity to perform certain activities at work. |
| **Permanent Total Disability** | ·    Requires medical treatment. |
| ·    Injury results in permanent loss of bodily function |
| ·    Rendered permanently unable to work |
| **Death** | ·    Fatality caused by workplace injury. |

Injury Severity, as an outcome of workplace injuries has been calculated by various organizations using different variables and measures. Organizations such as OSHA and the National Safety Council measure injury severity in two ways. The first measure of injury severity by OSHA is the level of loss of bodily function described as, “impairment of the body in which part of the body is made functionally useless or is substantially reduced in efficiency” (OSHA, 2016). The second measure of severity used by OSHA is a calculation in which severity is the outcome of time lost at work divided by total time an employee works. This calculation implies monetary significance to workplace severity and commonly used by for-profit corporations for risk-mitigation purposes.

Developing a measure of severity is crucial for predicting and understanding workplace injury trends and outcomes. Because the dataset does not include any documentation of work hours lost due to injury, our best resource for measuring the severity is the “Claim Injury Type” column which directly relates to OSHA’S definition of severity measured by bodily impairment/loss of bodily function. Although there isn’t a numerical measure of work hours lost, the “Claim Injury Type” implies a theoretical measure of time lost with the descriptors of “temporary, partial, and permanent” limitations to work abilities. Following OSHA’s guidelines for measuring injury severity, the Injury Severity variable column was created.

|  |  |  |
| --- | --- | --- |
| **New Variable: Injury Severity** | | |
| **Med Only** | Due to no loss of ability, this injury type is classified as “Low". | Low |
| **Temporary** | Due to the temporary loss of function with full expected recovery, this injury is classified as “Moderate”. | Moderate |
| **Permanent Partial Disability** | Due to the permanent nature of this injury, this classification will be ranked as "Severe". | Severe |
| **Permanent Total Disability** | Due to the permanent complete loss of physical function and ability to work, this injury classification will be ranked as "Very Severe" | Very Severe |
| **Death** | Because Death is absolute, the regular terminology of "low, moderate, etc." will not suffice to express the gravity of the outcome. This injury result will be classified as "Fatal". | Fatal |
|  |

Another value that is not directly documented by the data, yet is essential to analyzing workplace injury trends, is the measure of Physical Labor Levels in the workplace. Although the data does not contain a ranking system directly measuring the physical labor levels of each occupation, the data documents each occupation associated with the injury case. Each occupation is easily categorized into physical labor levels using the Occupation Requirements Survey (ORS) published by the Bureau of Labor Statistics. The ORS is a compilation of job-specific information documenting factors such as physical demands and environmental conditions, which is used to calculate the “strength levels” of each job. The ORS information is used in this study to classify each industry by “Physical Labor Levels” as displayed below:

|  |  |
| --- | --- |
| **New Variable: Physical Labor Levels** | |
| **Low Physical Labor** | ·    Health Care and Social Assistance |
| ·    Educational Services |
| ·    Information, Professional, and Technical Services |
| ·    Real Estate, Rental, and Leasing |
| ·    Finance and Insurance |
| ·    Enterprise MGMT |
| **Moderate Physical Labor** | ·    Public Administration |
| ·    Entertainment and Recreation |
| ·    Hospitality and Food Services |
| ·    Wholesale Trade |
| ·    Utilities |
| ·    Retail Trade |
| **High Physical Labor** | ·    Transport and Warehousing |
| ·    Manufacturing |
| ·    Construction |
| ·    Waste MGMT Admin |
| ·    Farm and Agriculture |
| ·    Mining |

**Exploratory Data Analysis (EDA)**

**Data Description**

After the initial cleaning of data including the reduction of outliers, and elimination of incomplete columns and rows, there are 34 variables in the dataset that can be used for exploratory data analytics and model building.

**Response Variable**

***Injury Severity*** is a categorical variable that indicates the severity of the injury outcome. The injury severity is measured by the level of disability that results from the injury event and is ranked from **low** to **fatal.**

**Univariate Analysis: Distribution of Key Variables**

To initiate the EDA, an evaluation of variable distributions was conducted to assess descriptive attributes of the data. The focus of the distribution analysis was placed on the variables being considered as potential candidates for the predictive analysis model. The following variables distributions were analyzed: Severity, Age Groups, Attorney/Representative, Physical Labor Level, Industry Cody Description, Nature of Injury Description. The analysis explores measures such as counting the frequency and central tendencies of each occurrence. The frequencies describe the summations of each occurrence per category in the variable and are accompanied by the probability of each category occurring. The following graphs illustrate the distributions for the response variable and each exploratory variable candidate.

The following distribution charts illustrate the dispersion and frequency rates of each variable that is being considered for the statistical models. Some of the key observations of these charts are the distribution of the result variable, severity. There are very few instances within the fatal and very severe categories of the variable and the frequency is centered around the moderate category. With the exception of variables Age and Attorney Presence, the other variable categories are listed from least frequent to most frequent.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Bivariate Analysis**

The next phase of the EDA phase focuses on identifying trends and relationships between candidate predictor variables, also known as bivariate analysis. The purpose of this phase is to identify relationships between variables and discover correlations and associations. Exploring correlations between variables can assist in identifying patterns and trends among the variables, ultimately enhancing the data analysis. Because most of the variables are categorical, the variable correlation is assessed using cross-tabulation analysis in JMP. Cross tabulation analysis is a technique that explores the relation of two categorical variables by arranging the data into a contingency matrix that displays the frequencies for the combinations of the two variables. Because the matrixes are large, they won’t be included in the visual results, but we will display the graphical representation of each cross-tabulation, along with the test results tables.

A screenshot of a computer

Description automatically generated

This analysis examines correlation between Severity and Age Groups with 32 degrees of freedom. The R-Square value of 0.0055 suggests that the combination of both variables explains only a small amount of variation in the data. The likelihood ratio test resulted in a chi-square value of 4584.758 and a p-value less than 0.0001, which confirms that the relationship between Severity and Age is statistically significant.

This contingency analysis model assesses the relationship between Severity and Attorney/Representative. According to the R-Squared analysis, this correlation accounts for 8.85% of the total variation of the data. The likelihood Ratio and the Pearson Ratio both resulted in p-values of less than 0.0001, which indicates that the relationship between the variables is statistically significant.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

This table represents the correlation analysis between Severity and Sex. The high LogLike value determines a positive goodness of fit within the model. The low R-Square value indicates that this correlation amounts for a small portion of the variation in the data. The Likelihood Ration and Pearson test both indicate that the relationship between the variables is statistically significant.

Severity by Physical Labor Level is analyzed by this contingency analysis model. This plot shows a significant relationship between both variables, as indicated by the Likelihood Ratio test and the Pearson test. However, according to the R-Square result, this correlation accounts for a very small portion of the variation in the data.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The relationship between Severity and Industry description is analyzed in this contingency analysis. The high LogLike indicates a positive goodness of fit between the data and the model. The Likelihood and Pearson test ratios indicate a significant relationship between the variables. The R-Square value suggests a moderate amount of variation is explained by this correlation model.

A screenshot of a computer screen

Description automatically generated

The contingency table analysis measuring the correlation between Severity and Cause of Injury demonstrates a positive goodness of fit between the variables and the model. The likelihood ratio and Pearson test show a high correlation, which is exhibited by the high Chi-Square values in contrast to the p-value, The P-value being less than .0001 demonstrates a high significance between the variables.

This contingency table demonstrates the relationship between Severity and Nature of Injury Description. This model has a positive goodness of fit, as demonstrated by the LogLike value of 13194.193. The R-Square value 0.0317 indicates that 3.17% of the variation in the data is explained by this correlation which is considered a moderate effect. The Likelihood ratio and Pearson tests express a statistically significant relationship between the variables.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

This contingency analysis model demonstrates the correlation relationship between Severity and Body Part Categories. The Loglike test suggest a positive goodness of fit between the variables and the data, the R-Square result indicates that 2.8% of the datasets variation is explained by these variables. The P-value being less than .0001 determines that the relationship between the two variables is statistically significant.

**Modeling**

**Multiple Logistical Regression Analysis**

To address the research question: *Is it possible to predict the probability of severe injuries? Which demographical and nature of injury factors contribute to likelihood of severe injury?* Multiple Logistical Regression Analysis is used. Multiple Logistical Regression Analysis is a machine learning formula used to “model the relationship” of “continuous or categorical explanatory variables” with categorical response variable (JMP, 2022).

The Multiple Regression Model was performed to analyze the relationship between the categorical response variable, Injury SEVERITY and the explanatory variables, Attorney/Representative, Body Part Categories, Physical Labor Level, BLS Cause of Injury, and Age at Injury. The analysis aims to explore the significance of these variables in ultimately predicting injury severity.

The model was run twice using the validation method where 60% of the data was used for training and 40% of the data was used for testing. The validation approach allows us to test the reproducibility of the model’s performance on unseen data. The null hypothesis states that there is no significance between the output variable and the exploratory variables. The figure below displays the imputations to run the model using the JMP software.

A screenshot of a computer

Description automatically generated

Training Testing**A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated**

The tables above demonstrate the training and testing model results of the nominal logistic regression analysis measuring the relationship between the response variable, Severity, and the exploratory variables. On all measures, the testing data set performed better than the training set with higher accuracy and lower misclassification, which indicates that the model generalizes well to unseen data. The better performance of the testing model shows that the model has learned the patterns in the training data, has captured the underlying relationships, and can be applied to new examples of data.

**Performance Metrics:**

* Entropy R-Square: 0.2151 & Generalized R-Square: 0.3995
  + Higher values of Entropy R-Square and Generalized R-Square indicate better fit of the model. However, Entropy and Generalized R-Square is low, which is common in models explaining human behavior (Frost, 2018).
* Mean -Log p: 0.7660
  + The Mean -Log p: is higher than the desired and a lower rate would indicate more accurate predictability.
* RASE:
  + The root averaged squared error (RASE) is within the desired parameters, as a lower value indicated better prediction accuracy.
* MEAN ABS DEV: 0.3913
  + The value is within the desired parameters, as lower values suggest better performance of the model.
* Misclassification Rate: 0.2676
  + The misclassification rate indicates that the model is misprediction approximately 27% of the predicted severity instances. This accuracy score is favorable.

**Variable Logworth P-Value**

The logworth is the “-log transformations of each model effect's p-value” (Upton, 2022) and is used to better indicate the importance of each variable toward severity prediction. The variables are listed in ranking order of most to least importance. High logworth values indicate stronger importance of the variable in the model. Researchers commonly use the logworth P-value measure to indicate whether a variable is useful to the model or if it should be removed to enhance interpretability (Upton, 2022). The logworth P-value results indicate that each variable has a strong association with the response variable, Severity.

According to the table, the strongest variable associate with Severity is Attorney/ Representative and the weakest association is with Age Groups.

A screenshot of a graph

Description automatically generated

The results of the performance metrics of the nominal regression analysis demonstrate that the exploratory variables analyzed in this study have a significant relationship with the prediction of severity levels within the injury data. Each variable returned very low P-values suggesting that the positively significant relationships discovered between the exploratory variables and injury severity are not occurring randomly, rather by correlation. According to the performance metrics, the model suggests an acceptable fit to the data provided. Furthermore, the null hypothesis stating that there is no relationship between the exploratory variables and injury severity can be rejected.

After establishing the positive significance between the variables, separating the outcome variable to the highest severity categories provides a deeper analysis of the relationship between the exploratory variables and worst-case outcomes. Isolating the outcome categories, Very Severe and Fatal in the non-linear regression model, the highest contributing variables can be identified. According to the parameter estimates table, the variables that are the most associated with Very Severe and Fatal outcomes include bodily injuries associated with the Head, Lower extremities, Multiple body parts, and the Neck/Throat. The cause of injury most associated with very severe and fatal outcomes are Fire/Explosions. Finally, the physical labor level contributing to very severe and fatal injury outcomes is High Manual Labor.A screenshot of a computer

Description automatically generated

Decision Tree:

According to JMP, Decision Trees, also known as Classification or Partition Models, are used to identify the “most important factors that predict a categorical outcome” (JMP, 2022). Decision trees have tree-like structures consisting of root nodes, decision nodes, and leaf nodes (Khairuddin *et. al,* 2023). The root node is the beginning point of the model, the decision nodes are used for decision-making and developing the model into various branches. Lastly, the leaf node is the outcome result of the decisions made by the decision node (Khairuddin *et. al,* 2023). The classification element of the tree come from the splitting process from the root node into the leaf node until a tree-like decision structure is created.

To approach the research question: *Is it possible to predict the probability of severe injuries? Which demographical and nature of injury factors contribute to likelihood of severe injury?* The same variable selection where Severity is the response variable, will be used to construct the Decision Tree Model. To improve the simplicity and interpretability of the model, the Severity response variable was re-coded into a binary classification where 0= low severity and 1=high severity. The low severity category consolidates the low and moderately severe portions of the data, while the high severity consolidates the severe, very severe, and fatal portions of the data. The tree below demonstrates the small tree version of the decision tree as a simple representation of the points in which the decisions splits occurred.

A diagram of a structure

Description automatically generated

Because we specifically wanted to evaluate which variables contributed to high severity outcomes, the decision tree followed the classification pattern that pursued the High Severity variables for the splits.

**Performance Metrics:**

**A screenshot of a computer program

Description automatically generated**

* Entropy R-Square: 0.2120 & Generalized R-Square: 0.2972
  + The lower entropy and generalized R-Square results are not within the desirable parameters, as higher levels express better fits for the model. As mentioned before, low R-Square values are common when evaluating human behavioral data (Frost, 2018).
* Mean -Log p: 0.3731.
  + The score is within the desirable parameters below 0.5. Lower scores suggest better accuracy within the predicted values.
* RASE: 0.3349
  + Lower RASE scores indicate better accuracy, and the score is within the desired parameters.
* MEAN ABS DEV: 0.2379
  + The value is within the desired parameters, as lower values suggest better performance of the model.
* Misclassification Rate: 0.1686
  + The misclassification rate indicates that the model is misprediction approximately 17% of the predicted severity instances. This accuracy score is favorable.

**A screenshot of a computer

Description automatically generated**

The table above illustrates the contributions that each exploratory variable generated for classifying high severity outcomes. The highest contributing variable is the Attorney/ Representative category, accounting for approximately 64% of the column. The lowest ranking variable is the Age Groups category with two splits.

**Neural Network Model:**

According to JMP, Neural Networks are network-based models used to “describe the impact that multiple predictor variables have on an outcome” as well as making “predictions of a categorical or continuous outcome”. (JMP, 2022). To analyze the research question, *Is it possible to predict the probability of severe injuries? Which demographical and nature of injury factors contribute to likelihood of severe injury?* The Injury Severity variable will be used as the output layer of the neural model and input variables will be the exploratory variables that have been used for t other models. The neural network model will be using the k-fold validation with 5 hidden folds. The neural model will use a random seed of 5.

The diagram below shows the visual construction of the neural model. The first layer of the neural network model contains the input variables, also known as the exploratory variables. The second layer is the hidden layer which is responsible for transforming the extracted data from the input layers to make predictions and classifications. Finally, the output layer demonstrates the categorization of severity. Unfortunately, the visual representation is limited and doesn’t include an output icon for each of the severity categories.

**A diagram of a diagram

Description automatically generated**

**Performance Metrics: A screenshot of a computer

Description automatically generated**

* Entropy R-Square: 0.1546111 & Generalized R-Square: 0.30888904
  + The lower entropy and generalized R-Square results are not within the desirable parameters, as higher levels express better fits for the model.
* RASE: 0.5505864
  + Lower RASE scores indicate better accuracy. The RASE score is higher than the desired accuracy parameters for this performance measure.
* MEAN ABS DEV: 0.520258
  + The value is higher than the desired parameters, as lower values suggest better performance of the model.
* Misclassification Rate: 0.41218
  + The misclassification rate indicates that the model is misprediction approximately 41% of the predicted severity instances. The misclassification score is very high.

**Actual vs. Predicted Confusion Matrix and Prediction Profiler**

The confusion matrix table shows the predictive capability of each category within the output variable, Severity. The predicted count shows the summation of each predicted classification occurrence per category and the actual summation for each classification category. The predicted rate shows the rate each of prediction occurrence and the actual rate occurrence for each category. The Prediction Profiler is a visualization of the neural model’s predictions across the exploratory variables.

**A screenshot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Model Evaluation**

To determine the research question, *Is it possible to predict the probability of severe injuries? Which demographical and nature of injury factors contribute to likelihood of severe injury?* three different models were constructed.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic Regression** | **Decision Tree** | **Neural Model** |
| Entropy R-Square: | 0.2151 | 0.212 | 0.15461 |
| Generalized R-Square: | 0.3995 | 0.2972 | 0.3089 |
| Mean -Log p | 0.766 | 0.3731 |  |
| RASE | 0.4954 | 0.3349 | 0.5506 |
| MEAN ABS DEV | 0.3913 | 0.2379 | 0.520258 |
| Misclassification Rate: | 0.2676 | 0.1686 | 0.41218 |

According to the comparison of metrics for each model, the best performing model is the logistic regression model. The logistic regression model demonstrates the higher of the R-Squared values in combination with a good misclassification rate. The Decision Tree model was a close second in ranking and had the best misclassification rate of the models. The Neural Model performed poorly on all metrics and had a very high misclassification rate with the lowest R-Squared values. In addition to having the best metric results from the three models, the multiple logistic regression model was also the easiest of the models to build and execute and most efficiently illustrated the levels of significance between the response and exploratory variables.

**Conclusion**

The research question, *Is it possible to predict the probability of severe injuries? Which demographical and nature of injury factors contribute to likelihood of severe injury?*  and the hypothesis: *Demographical and nature of injury data can be used to classify the severity of a WCB injury claim* were explored using three predictive models. The findings were that it is indeed possible to predict the probability of severe injuries and the factors that contributed to the likelihood of severe injury included Attorney Presence, Body Part, Age, Physical Labor Level, Cause of Injury, and Injury description. According to each of the predictive models, each variable had a statistically significant relationship with the predictor variable, Severity. The performance metrics of the models suggests that the significance of the relationships between the models are not occurring by random chance, rather by correlation.

According to the machine learning models, the highest contributing variable within the severity classification outcome was if lawyers were involved in the Workman’s Comp claim process. This demonstrates that when an employee is severely injured, litigation is required to achieve disability status, which is costly in time and resources for organizations. This observation should serve as a motivating factor for companies when approaching improvement within occupational safety for their employees.

The second most important determining factor of injury severity outcomes, as suggested by the models, is Body Part. Notably, all three models determined that the body part injuries that are most indicative of high severity are the head and the upper extremities. These findings highlight that there are opportunities for improvement within the resources for protecting these body parts.

After confirming the significance between the exploratory and outcome variables, deeper analysis was conducted to focus on the leading contributors of Very Severe and Fatal injuries. By isolating the highest severity cases in the regression model, it was discovered that the worst bodily injuries were trauma to the head, neck/throat, upper extremities, and multiple body parts. Fire and explosions were the leading cause of high severity outcome. Lastly, as expected, high physical labor levels were a prominent factor in predicting high severity cases.

**Limitations of the Research:**

Some of the limitations of the research involved gaps within the data. The variable of Weekly Income would have provided desired insight regarding the disparities within income brackets and occupational safety resources. However, the raw data column for Weekly Income was left 75% blank and several values were erroneously inputted as one or two dollars per 40-hour work week.

The dataset would be well served to include more medical details such as if hospitalization was required, or the exact number of days that were missed from work. Variables such as these would contribute to better understanding the outcomes of the incidents and how they impact individuals and organizations.

Lastly, a limitation within the research is that there may be biases within the dataset that result from the unreliability of those that input the data. The data is collected as a legal requirement of the state of New York and employers are required to report the data. The amount of erroneously inputted data, such as the vast majority of weekly wage inputs, implies possible unreliability of data accuracy.

**Recommendations for Future Research and Enhancing Safety:**

As certain indicators for high severity trends are discovered, there are new opportunities for deeper research. An approach to better understand the indicators of injury severity outcomes would be to use the top contributing attributes of high severity outcomes and run machine learning models on them. Since several bodily injuries were determining factors in high severity cases, further research must be conducted, such as constructing models that explore the environments in which certain injuries occur. This approach would explore relationships between certain body part injuries and find correlations within different industries and other environmental factors. The results of such research could be used to identify the need for certain head protection gear, as well as determining the incidence rates and how to avoid injuries. The models also identified that high physical level occupations pose the highest risk for severe injury outcomes. This finding establishes an urgency to focus on the occupational hazards associated with high physical level fields.

The extensive dataset of WCB injury claims provides a valuable platform for addressing essential questions that can contribute to the improvement of public policy and safety protocols. By analyzing injury patterns and identifying areas in need of improvement, this data can propel meaningful change in safety practices.

References

Bureau of Labor Statistics. “Strength Levels.” U.S. Bureau of Labor Statistics, www.bls.gov/ors/factsheet/strength.htm. Accessed 17 July 2023.

Bureau of Labor Statistics. “Using Workplace Safety and Health Data for Injury Prevention: Monthly Labor Review.” U.S. Bureau of Labor Statistics, www.bls.gov/opub/mlr/2013/article/using-workplace-safety-data-for-prevention.htm. Accessed 18 July 2023.

“Department of Labor Logo United States department of Labor.” Definitions for “near Proximity” and “Serious Injury”. | Occupational Safety and Health Administration, www.osha.gov/laws-regs/standardinterpretations/1993-09-02. Accessed 15 July 2023.

Frost, Jim. “How to Interpret R-Squared in Regression Analysis.” Statistics By Jim, 7 July 2023, statisticsbyjim.com/regression/interpret-r-squared-regression/.

Glorian Sorensen a b, et al. “The Future of Research on Work, Safety, Health and Wellbeing: A Guiding Conceptual Framework.” *Social Science & Medicine*, 8 Dec. 2020, www.sciencedirect.com/science/article/pii/S0277953620308121.

“How to Benchmark Occupational Injury and Illness Rates.” Injury Facts, 9 May 2023, injuryfacts.nsc.org/work/industry-incidence-rates/how-to-benchmark/#:~:text=The%20basic%20formula%20is%20(N,%2C%2050%20weeks%20per%20year).

“Improving the National Statistics.” *U.S. Bureau of Labor Statistics*, July 2020, www.bls.gov/opub/btn/volume-9/nearly-50-years-of-occupational-safety-and-health-data.htm.

International Labor Organization. “Occupational Safety and Health (OSH) (Eastern Europe and Central Asia).” Occupational Safety and Health (OSH) (Eastern Europe and Central Asia), www.ilo.org/moscow/areas-of-work/occupational-safety-and-health/lang--en/index.htm. Accessed 15 July 2023.

JMP. “Multiple Logistic Regression - JMP.” JMP.Com, www.jmp.com/content/dam/jmp/documents/en/academic/learning-library/05-multiple-logistic-regression.pdf. Accessed 18 July 2023.

“New Study Finds Workplace Safety More Important than Ever for Employed Americans.” *AlertMedia*, 15 Mar. 2022, www.alertmedia.com/news/employee-safety-report-2022/.

Nystrom, Michelle. “How Workplace Injuries Impact Your Business.” *O2 Employment Services*, 27 Feb. 2023, o2employmentservices.com/how-workplace-injuries-impact-your-business.

“Severity Meaning &amp; Definition.” EcoOnline, www.ecoonline.com/glossary/severity. Accessed 16 July 2023.

Vasudev, Rakshith. “How Are Logistic Regression &amp; Ordinary Least Squares Regression (Linear Regression) Related?” Medium, 5 June 2018, towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5.

Upton, Jeff. “What Does Logworth Measure That Is Not Included in the ANOVA Table of a Fit Model Analysis?” JMP User Community, 9 June 2023, community.jmp.com/t5/Discussions/What-does-logWorth-measure-that-is-not-included-in-the-anova/td-p/544036.

Zafar, Shehzad. “How Government Can Promote Standards of Health & Safety.” *Occupational Health and Safety Blog*, 23 Apr. 2023, www.hseblog.com/national-government-contribution-towards-good-standards-of-health-and-safety/.