Machine learning Project

Master's in Data Science and Advanced Analytics

**NOVA Information Management School**

Universidade Nova de Lisboa

**To Grant or Not to Grant:  
Deciding on Compensation Benefits**

**Report**

**Group 16**

Sarah Leuthner, 20240581

Bárbara Capitão, 20211532

Ricardo Kayseller, 20230450

Ludovica Collé, 20241449

Elisa Pacheco, 20240600

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

[Abstract 3](#_Toc184075258)

[I. Introduction 4](#_Toc184075259)

[I.1. Overview and Main Goals 4](#_Toc184075260)

[I.2. State of the Art 4](#_Toc184075261)

[II. Data Exploration and Preprocessing 5](#_Toc184075262)

[II.1. Key Data Insights 5](#_Toc184075263)

[II.2. Cleaning and Preparing the Data (Ludovica) 6](#_Toc184075264)

[III. Multiclass Classification 6](#_Toc184075265)

[III.1. Feature Selection (Elisa) 6](#_Toc184075266)

[III.2. Model Assessment Strategy (Barbara / Ricardo) 6](#_Toc184075267)

[III.3. Comparing Performances and Choosing the Best Model (Barbara / Ricardo) 6](#_Toc184075268)

[III.4. Model Optimization (Barbara / Ricardo) 7](#_Toc184075269)

[IV. Open-Ended Section 7](#_Toc184075270)

[IV.1 Objectives 7](#_Toc184075271)

[IV.2 Description 7](#_Toc184075272)

[IV.3 Results and Discussion 8](#_Toc184075273)

[V. Conclusion 8](#_Toc184075274)

[Bibliographical References 10](#_Toc184075275)

[Appendix Graphs 11](#_Toc184075276)

[Annexes (Optional, Not included in page limit) 20](#_Toc184075277)

# Abstract

* Summary of work (200-300). The abstract should five an overview of the work: What is the context? What are your goals? What did you do? What were your main results, and what conclusion did you draw from them?
* Write at THE END!

# Introduction

The New York Workers’ Compensation Board (WCB), as regulating authority, processes insurance claims whenever it becomes aware of an injury at the workplace. With more than 200,000 cases and hearings in the last year[[1]](#footnote-1), the state agency needs assistance in their decision-making process. The following report will describe the strategy and approach to achieve this optimization with a predictive model as well as an interface to use for single case predictions that will automate the process and, therefore, will forecast the WCB’s decision.

## Overview and Main Goals

For this optimization in the WCB’s process, several steps are taken, structured in the following chapters. In *Chapter II – Data Exploration and Preprocessing* the received data and its key insights are described. Then the data is cleaned and prepared for the model, which is described in *Chapter III – Multiclass Classification*. This main part focuses on selecting the appropriate features and assessing the best model and its metrics by comparing the performances between several algorithms. In *Chapter IV – Open-Ended Section*, for additional insights, the created interface for the WCB’s workers and possibly claimants is explained. Lastly, in *Chapter V – Conclusion*, the findings are compared and discussed with the main goals.

All defined steps are taken, to achieve these three main objectives:

* **Multiclass Classification Benchmarking**: Classification model to accurately predict the WCB’s final decision on what type of injury should be given to a claim.
* **Model Optimization**: Improving the best model from the multiclass classification.
* **Additional Insights**: An analytical interface in form of a website that returns predictions given input for a single new insurance case.

## State of the Art

As work injuries increase[[2]](#footnote-2) the process of administration and especially settling the claims become increasingly resource-intensive and prone to inefficiencies leading to high time and cost efforts. Therefore, predictive modelling is becoming more frequently used to optimize insurance claims management processes, giving many similar use cases as the one on hand. Logistic regressions are ideal for binary results, while decision trees excel at capturing non-linear relationship and interactions between variables. Neural Networks recognize complex patterns managing high-dimensional datasets. All three model types are being used in the insurance branch.[[3]](#footnote-3) However, the question is, which model will most accurately predicts results. When comparing Tree and Neural Networks-based models, tree-based models, especially Gradient Boosting (XGBoost) and Random Forests, provide better results than Neural Networks.[[4]](#footnote-4) The implementation XGBoost is widely popular for its ability to handle missing values in the dataset.[[5]](#footnote-5) Though the Tree-based models give an overall more accurate prediction, Multi-Layer Perceptrons provide better results in some datasets.[[6]](#footnote-6) This leads to the conclusion, that there is no single best model for every use case and that each individual dataset must be tested with several models in order to find the best one. Additionally, even among tree-based models different applications need to be analyzed. A higher prediction accuracy can be observed when using multivariate trees instead of simple univariate trees. Generally, multivariate tree boosting tends to deliver a better performance than any other predictive model. Yet, Gradient Boosting and Random Forests provide almost equal values of validation measure to multivariate tree boosting. Regression trees usually underperform compared to ensemble models as well.[[7]](#footnote-7)

Based on this information and the use case of this report, it is possible for the best model to be tree-based, most likely being an ensemble model. This model could be either some form of Gradient Boosting / Random Forest or ani multivariate tree, where weak learners are turned strong. Depending on the dataset, there is a chance to find a fairly accurate Neural Network models as well.

# Data Exploration and Preprocessing

## Key Data Insights

The training dataset consists of almost 60.000 rows of claims with 32 different attributes from the start of 2020 till the end of 2022. The majority of people in the claims are between 31 and 54 years. Analyzing the dataset, several anomalies were found. The dataset has an imbalance regarding gender and industry. Most claimants are male and work in the health industry. Further imbalance can be seen in the attributes Alternative Dispute Resolution, Attorney/Representative and COVID-19 Indicator, as the majority of the values are “No”. The attribute WCB Decision only has “not work related” cases. Therefore, this attribute is a constant and can be discarded. Almost all attributes have **missing values**. For low amounts of missing data, the attributes can be treated by using the median or mean for metric values and the mode for categorical values. The median will be used in case of outliers, as it is not affected by them. There are also attributes with a high number of missing values: OIICS Nature of Injury Description (100%), IME-4 Count (77%), First Hearing Date (73%), C-3 Date (67%). Treating these missing values can introduce bias or noise, so removing the variable might be preferable.

Furthermore, **outliers** can be observed. According to the data description the cases should only range from 2020 to 2022. However, there are Accident Dates found from the beginning of September 1961 to end of September 2023, indicating invalid values. The Average Weekly Wage with a range of 0 to 2,8 million, even though the mean is 491,09$, shows either extreme cases or high outliers. The average claimants age is 42 years. Yet the range goes from 0 to 117, also indicating at least outliers for the high ages. Similar observations can be found for the attribute Birth Year. The attribute Zip Code also shows invalid or placeholder zip codes (e.g. 0, 99999), which need to be treated during preprocessing.

When analyzing accidents per months, there is no clear **trend** visible, but a slight decrease in number of accidents in April, May and November (Figure 2). However, less accidents occur on the weekend, as there are probably less workers during those days (Figure 3). Two peaks for Age at Injury can be observed (Figure 4). A person around the age of mid-30 and mid-50 is more likely to injure, than any other. Most accidents occur in the Health Care and Social Assistance as well as the Public Administration industry (Figure 5) being a strain/tear or contusion (Figure 6) caused by lifting (Figure 7). The specified location for the most accidents are in New York, Queens and Kings (Brooklyn) as well as Suffolk (Figure 8, Figure 9). Overall, the majority of the claims will be declared as a non-compensable injury type (Figure 10).

There are no significant **correlations** between the attributes, except for the Birth Year and Age of Injury (0,99) of the claimant. To reduce redundant information, one of these attributes needs to be discarded (Figure 11). Furthermore, no significant **multivariate relationships** between the Claim Injury Type and Gender can be noted, as they follow the general trend for both. There are more male injuries recorded than female. The most declared Injury Type is non-compensable, followed by temporary disability (Figure 12). Similar observations are made with the Accident Date. However, when comparing the top 10 Accident Causes across the top 10 counties, it can be seen that the highest occurrence is for lifting and pandemic in Suffolk. A high accident cause of lifting is also observed in Queens (Figure 13).

In general, younger people submitting a claim either will cancel their claim / get their claim CANCELLED, won’t be compensated (NON-COMP), only will be compensated for their medical expenses (MED ONLY) or have a temporary disability. Very young people in the management or mining industry will most likely cancel their claim or receive a cancelation. People older than 45 will most likely receive the injury types Permanent Patrial Disability (PPD) (Scheduled or Non-Scheduled Loss), Permanent Total Disability (PTD) or Death (Figure 14). But not only the age is an indicator for the Claim Injury Type. The wage in specific industries can also indicate results. People with no wage more likely get a classification of CANCELLED, NON-COMP or MED ONLY. Workers in public administration with a high wage mostly receive a PTD. High wage in the Information industry indicates several claims: temporary disability or any type of PPD. An injury in the industry mining pared with a high wage will most likely lead to a PDD or DEATH claim (Figure 15).

## Cleaning and Preparing the Data (Ludovica)

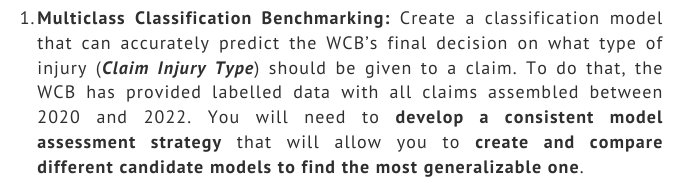
- Steps taken to clean and prepare the data

# Multiclass Classification

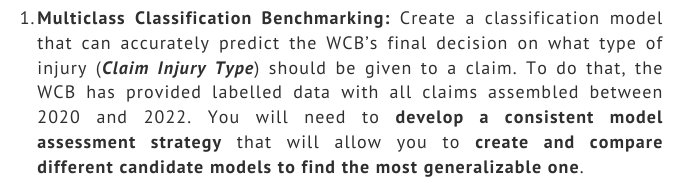
* Additional preprocessing steps adopted
* Feature Selection Strategy
* Explanation of model assessment strategy and metrics used
* Comparison of performance between candidate algorithms
* Optimization efforts: presentation, results and discussion

## Feature Selection (Elisa)

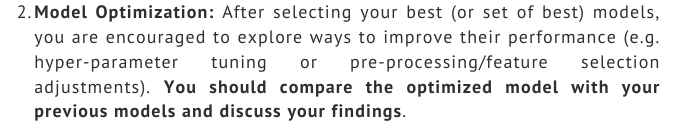
## Model Assessment Strategy (Barbara / Ricardo)



## Comparing Performances and Choosing the Best Model (Barbara / Ricardo)



## Model Optimization (Barbara / Ricardo)



# Open-Ended Section

## IV.1 Objectives

For additional insights, regarding the *open-ended section*, the objective is to create an interface that as a result returns a prediction given the new input. This should help the WCB to fasten their insurance management process making it more efficient but could also give any claimant clarity on its process and its probabilities, given this interface will be published for the public. With this tool, the WCB could concentrate on more difficult cases as well as the hearings instead of going through every claim manually saving time, enhancing the workers productivity. It will also increase the trustworthiness towards the claimant, making the process and all the considered aspect more transparent.

Regarding the tool, the goal is to save the created model via object serialization into a “Pickle” file. This is needed, as the model should not be retrained every time using the interface. The interface consists of multiple HTML templates. However, the mail goal is one template for all input data from the users (WCB worker and/or the claimant) and a separate one to receive the prediction as result. A “Flask” app creates the connection between the model and the HTML templates, making them interactive.

In addition, as a soft goal, the interface should be user friendly, easy to navigate and visually appealing for a good user experience.

## IV.2 Description

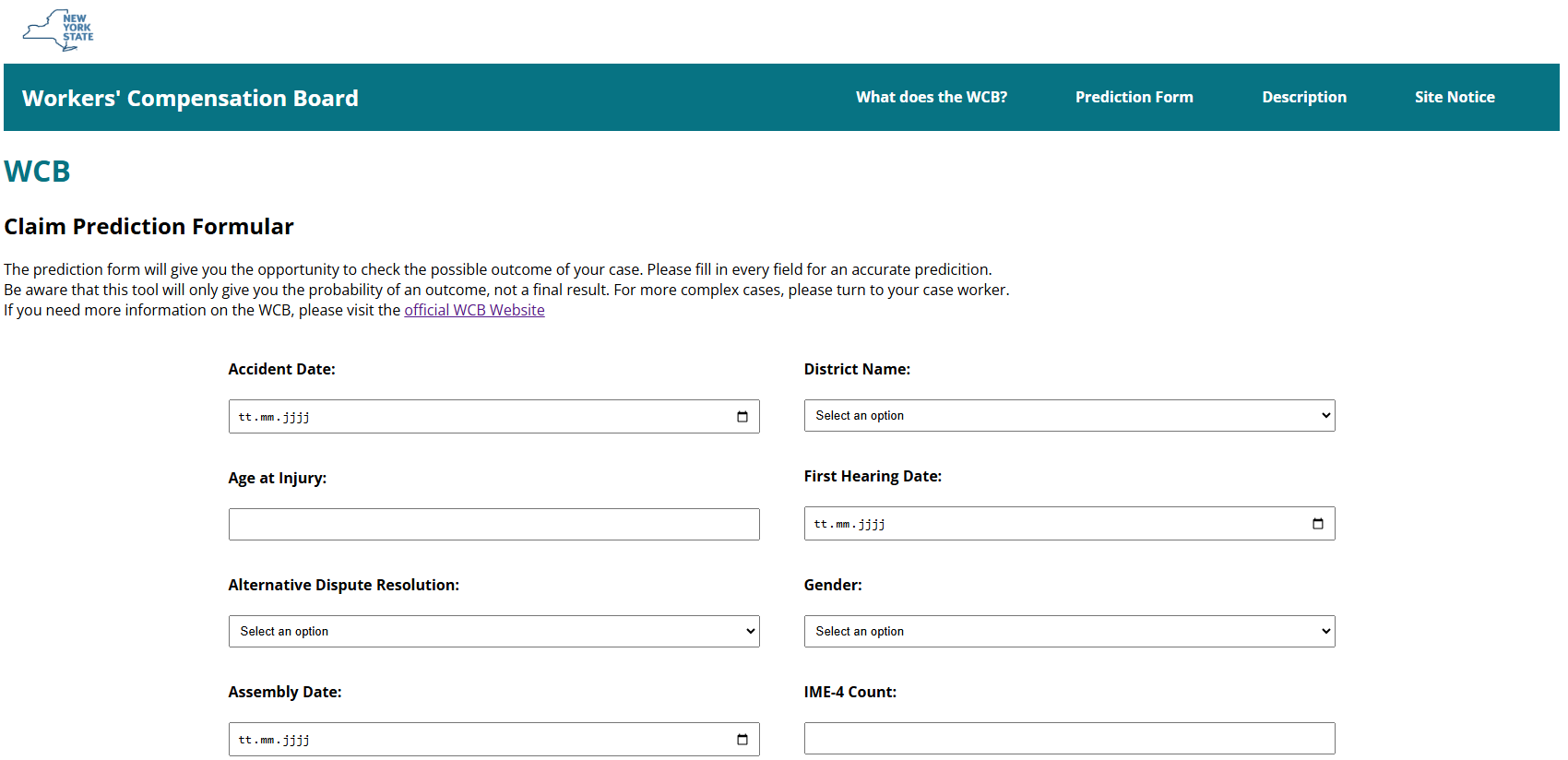


Figure 1: HTML prediction template extract

The HTML templates together combine to form a website. The general interface is replicated to look similar as the official website[[8]](#footnote-8) (Figure 1), matching the corporate identity with the color scheme. The website consists of 4 sites that can be accessed through the navigation bar:

* **What does the WCB?** Short description taken from the official WCB website
* **Prediction Form:** formular with input fields to predict the outcome of an individual case
* **Description:** Definition of every input field headline
* **Site Notice:** Lists the corporation between the WCB and NOVA IMS project members

The prediction formular contains different data entry fields. Depending on the demanded data it is either a date form, a drop-down menu or an input field for integers. Any other data types are not possible to enter, to reduce wrong input and adjustments afterwards to the data.

Once the template user fills in the input and activates the “Predict”-button, the “Flask” app will receive the data, adjust the values (if needed, for example by encoding) and then analyze the inputs using the saved model to return a prediction. This prediction is sent back to the prediction HTML template, which is shown to the user.

## IV.3 Results and Discussion

The results of the analytics interface, can be best seen by testing the template. It is a good way to predict single instances without having to training and running the entire model. An application of this tools is possible for both sides of the case. The WCB can work through their case load more efficiently getting first results and the claimant can get an idea on the expected result.

However, there are multiple possibilities to make this tool more beneficial. At the moment the model as its interface only uses the given data and therefore only has the given inputs available to select and can only make predictions regarding those. To include all possible codes of the WCIO[[9]](#footnote-9) and still predict accurate results, the model needs to be trained with these new possible feature values in addition. With this adjustment all possible cases are covered to make future predictions.

Furthermore, the use cases of this tool need to be defined. At the moment, the website is designed for web application. If a mobile application is needed, it needs to be responsive. Also, if the WCB workers want more insights or analytical methods, the interface could show similar cases with its prediction to review for the workers. This could be useful especially for new or complex cases and could work similar as the process of case-based reasoning. For more advanced analytics, an interactive analytical dashboard with graphs and filters could be created using the “Python” libraries “Bokeh” or “Plotty”.

# Conclusion

* Summary of initial objectives and discussion of corresponding findings
* DO the findings match what you initially expected? How?
* Discussion of limitations of your work (e.g. what could you have done differently)
* Suggestions for possible work to follow on your work.

# Bibliographical References

Kanchetti, D. (2021). Optimization of Insurance Claims Management Processes Through the Integration of Predictive Modeling and Robotic Process Automation

<https://www.researchgate.net/publication/383987572>

Michael Lawrence Varon, PLLC (2018 – 2024). Workplace Injury And Illness Increasing In New York. <https://www.nycompensationlaw.com/workplace-injuries-and-illness-increasing-in-new-york/>

New York State Workers’ Compensation Board (2023). New York State Workers’ Compensation Board Annual Report. <https://www.wcb.ny.gov/content/main/TheBoard/2023AnnualReport.pdf>

Quan, Z., Valdez, E. A. (2018). Predictive analytics of insurance claims using multivariate decision trees. <https://doi.org/10.1515/demo-2018-0022>

Rusdah, D. A., Murfi, H. (2020). XGBoos in handling missing values for life insurance risk prediction.   
<https://doi.org/10.1007/s42452-020-3128-y>

Tjahjono, S., Murfi, H., Devila, S. (2024). Claim Severity Prediction of Workers’ Compensation Using Tree and Neural Networks-Based Models. DOI 10.1109/SDS60720.2024.00039

Use APA Style for the entire document

We suggest that students use a reference manager system (Zotero, Mendeley, EndNote),

Please review the style guide at: <https://apastyle.apa.org/style-grammar-guidelines/references/examples>:

Author, A. A., Author, B. B., & Author, C. C. (Year). Title of article. *Title of Periodical, volume number* (issue number), pages.

# Appendix Graphs

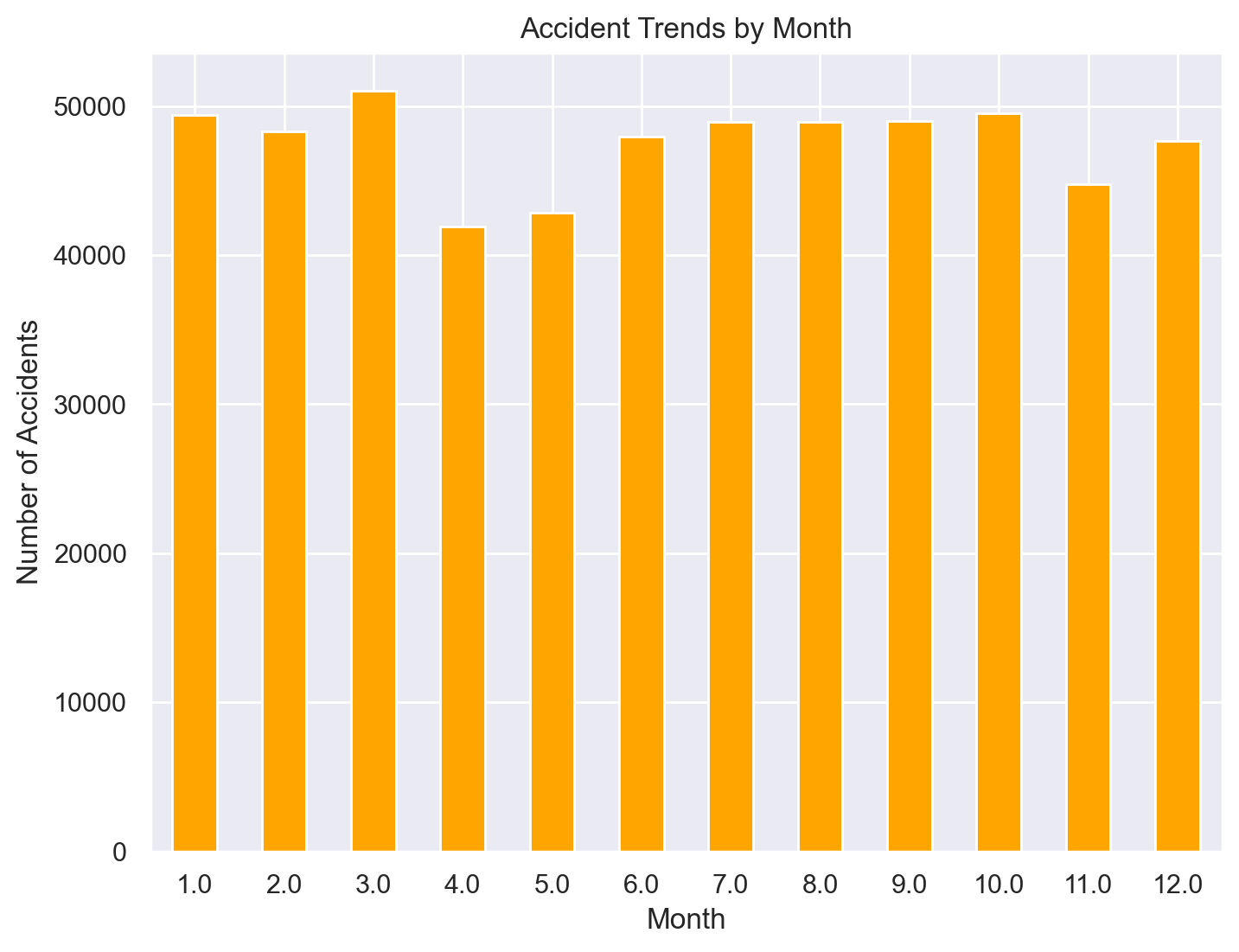


Figure 2: Accidents by Month

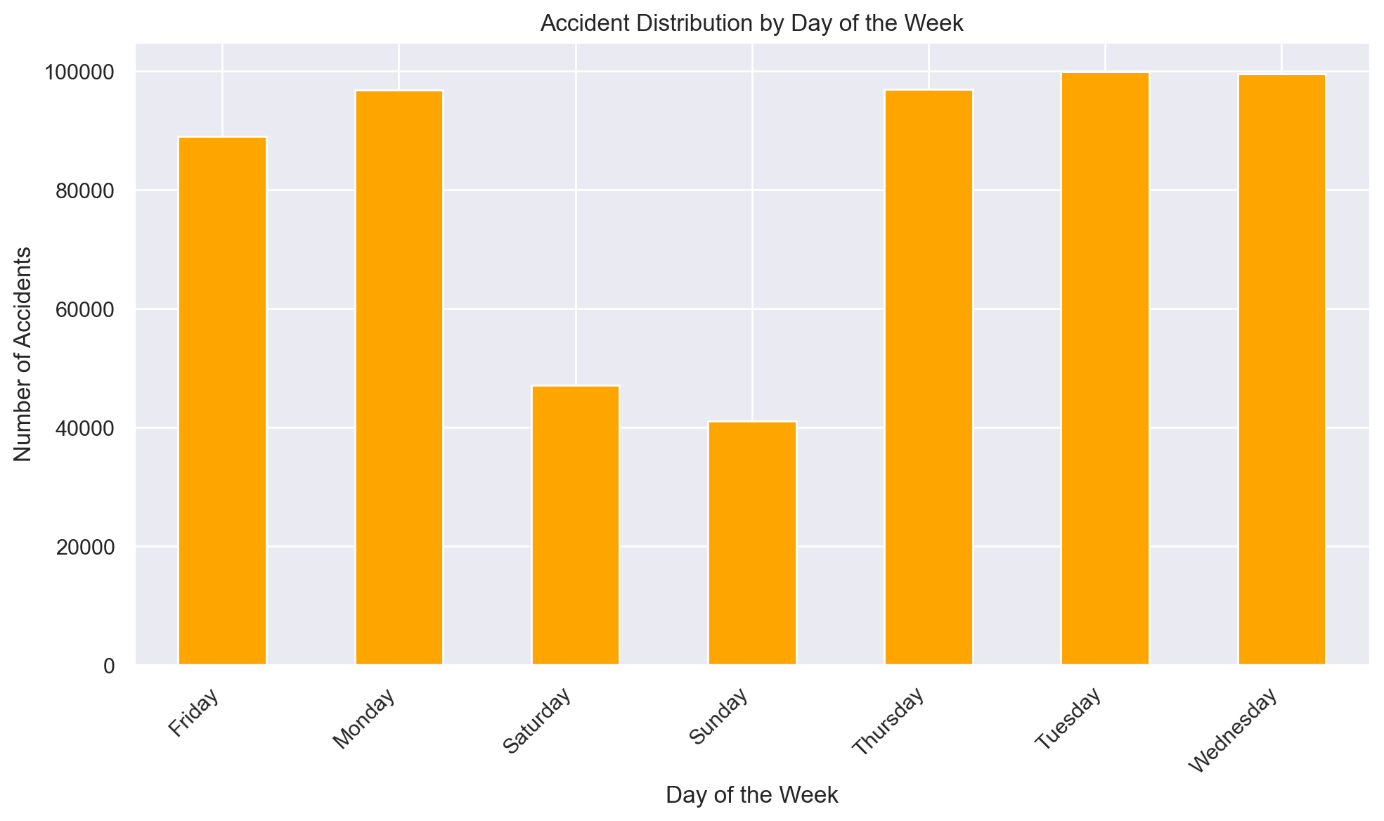


Figure 3: Accidents per Day

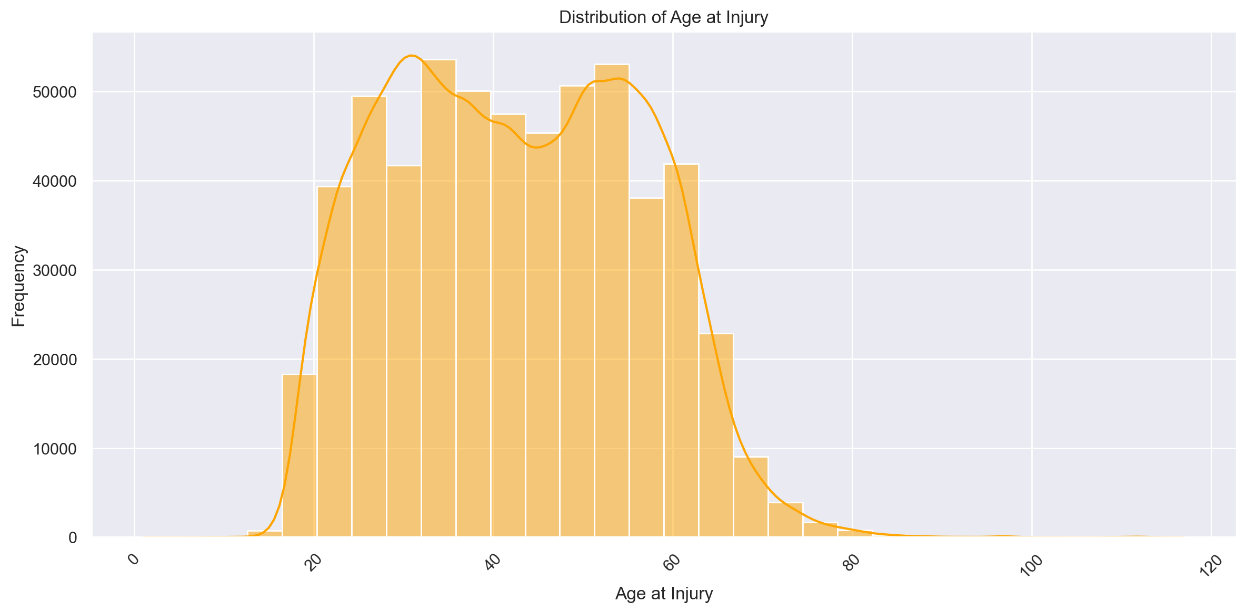


Figure 4: Distribution of Age at Injury

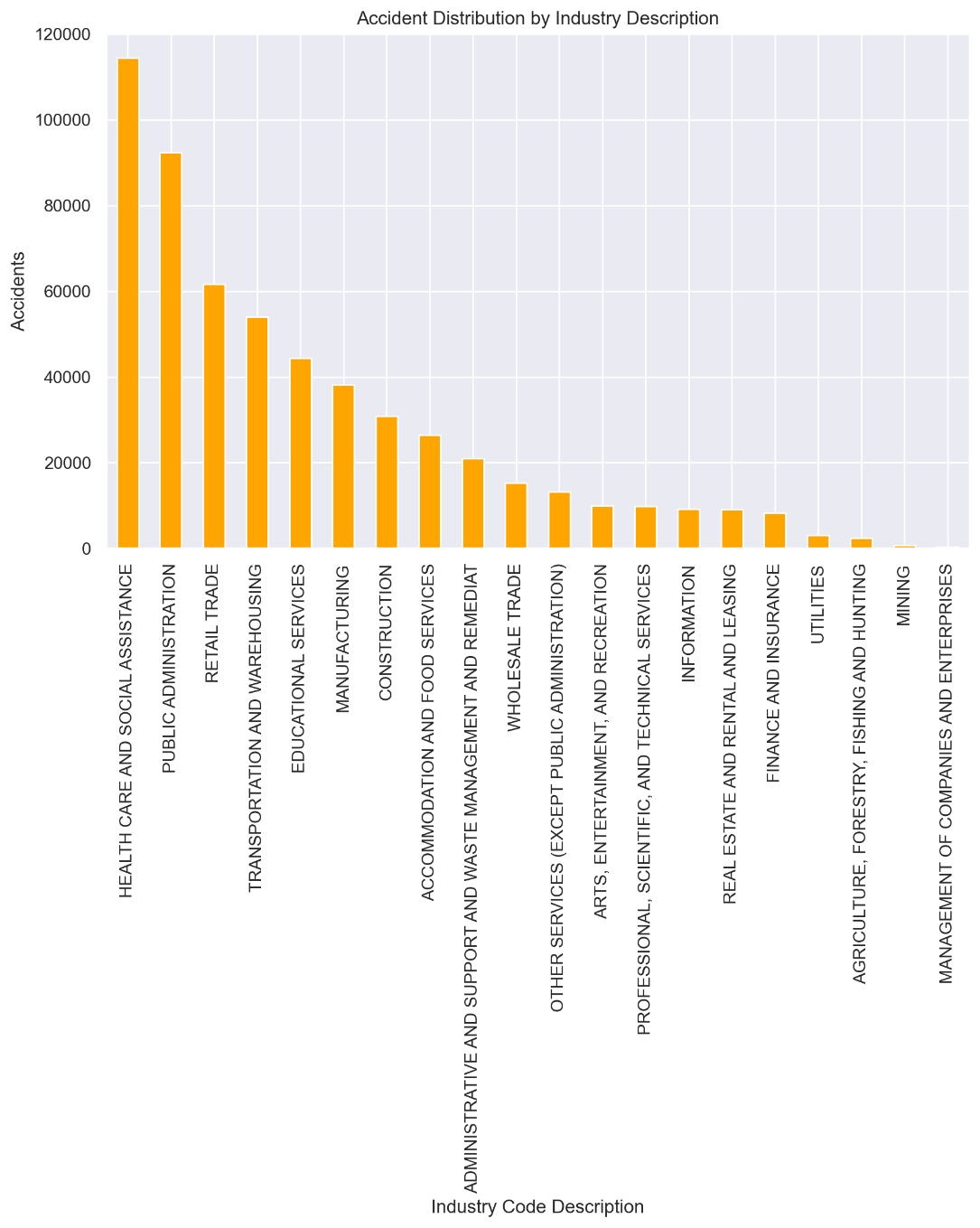


Figure 5: Accident Distribution by Industry

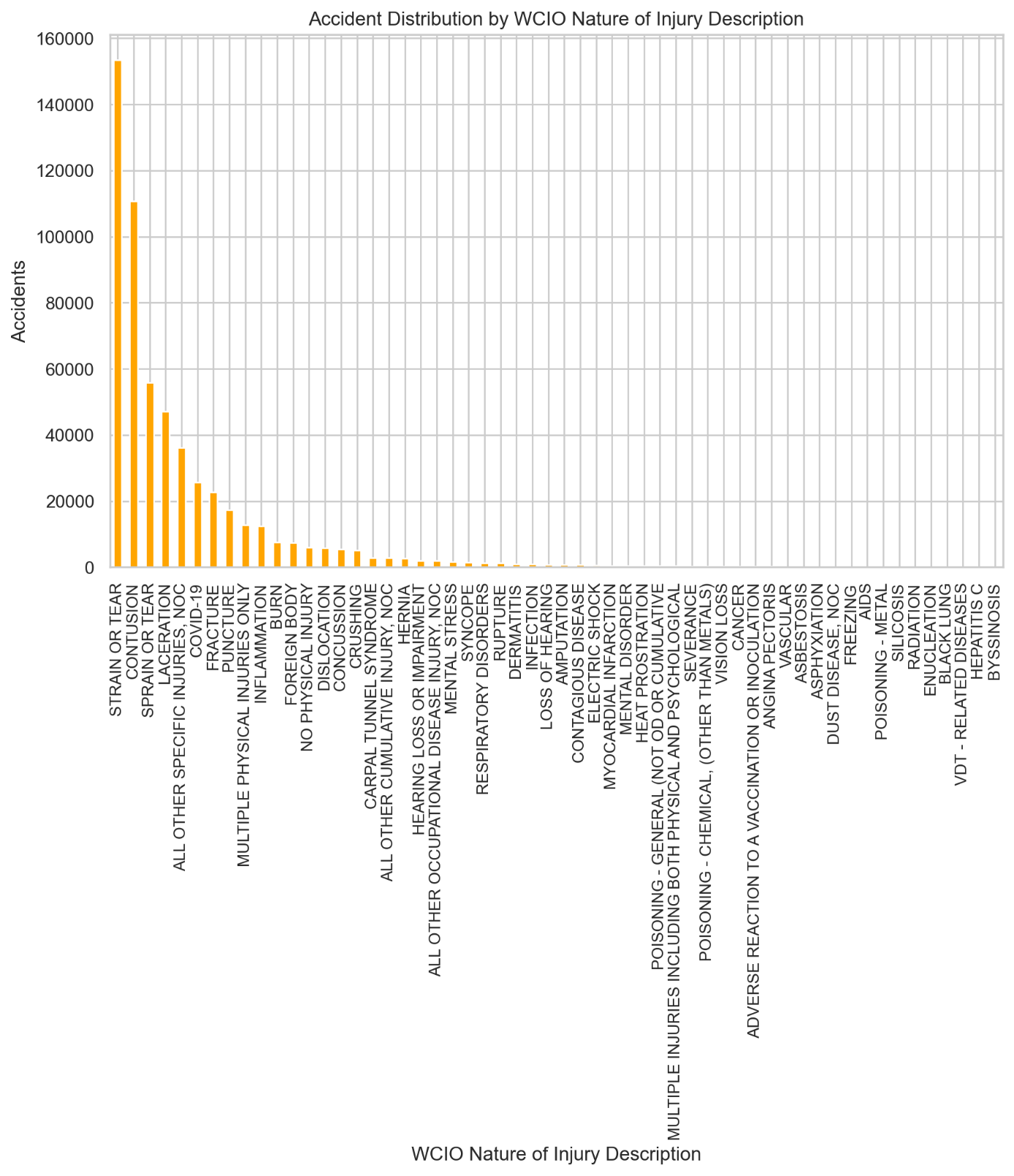


Figure 6: Distribution by WCIO Nature of Injury Description

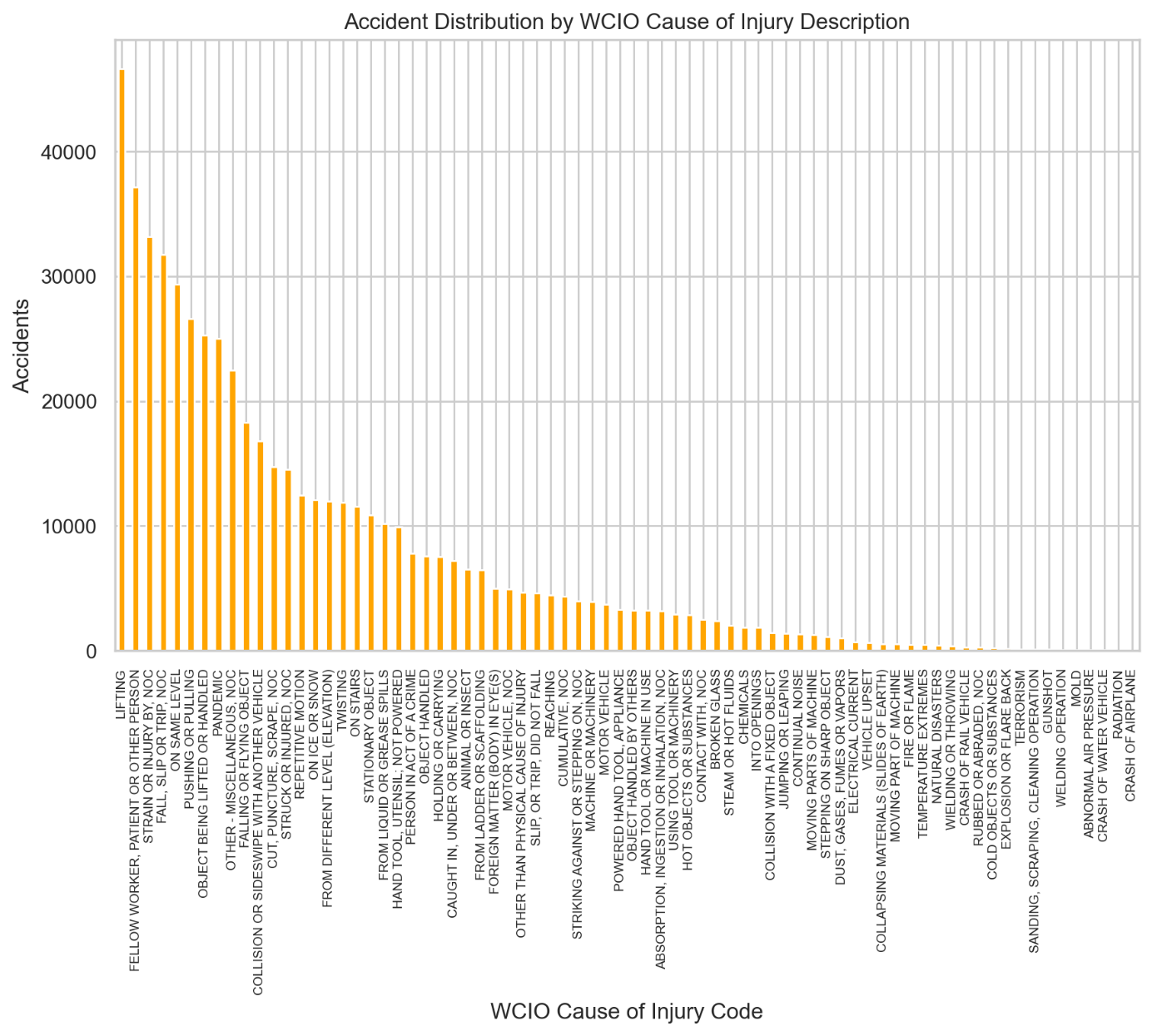


Figure 7: Distribution by WCIO Cause of Injury Description

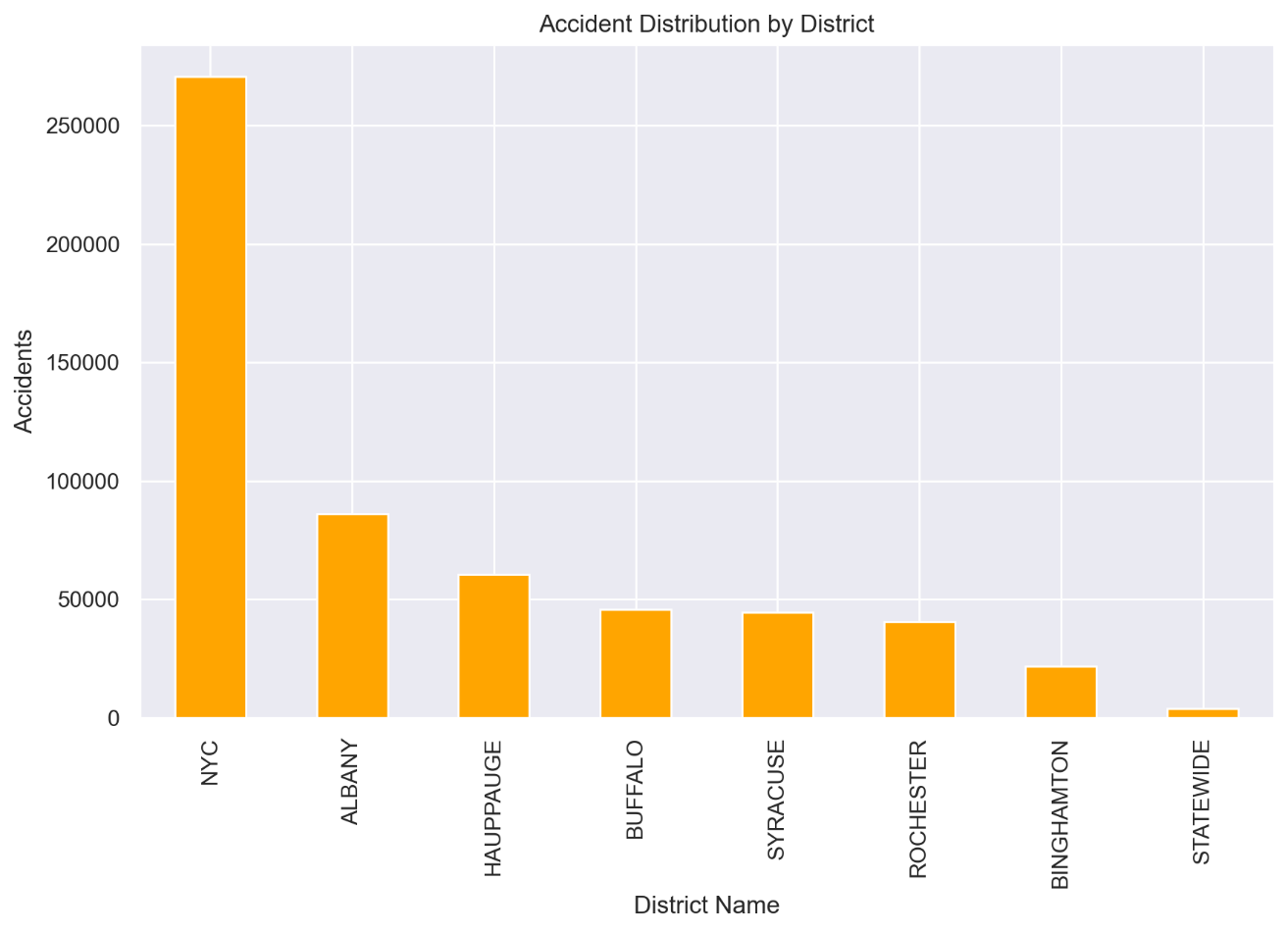


Figure 8: Distribution by District

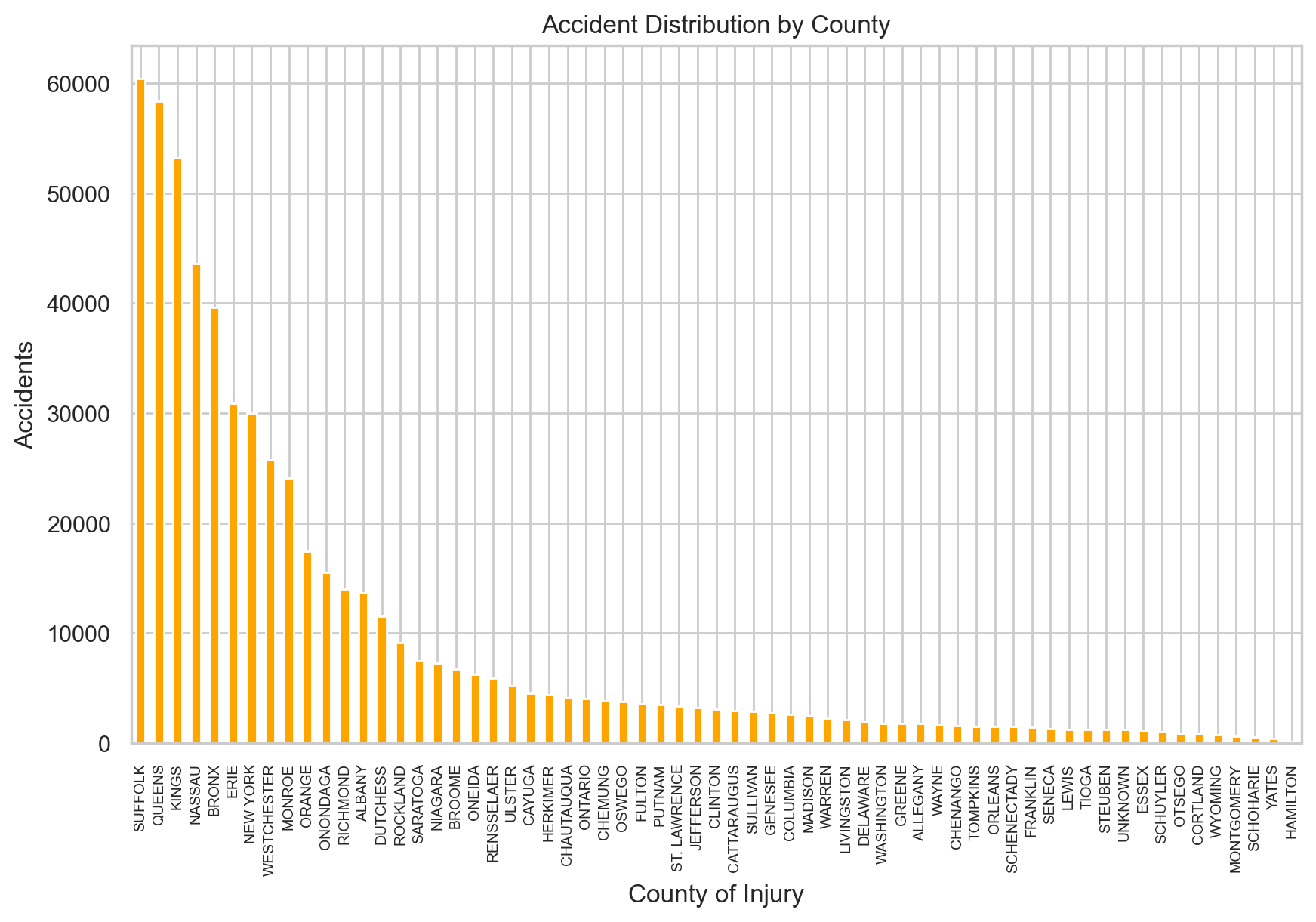


Figure 9: Distribution by County

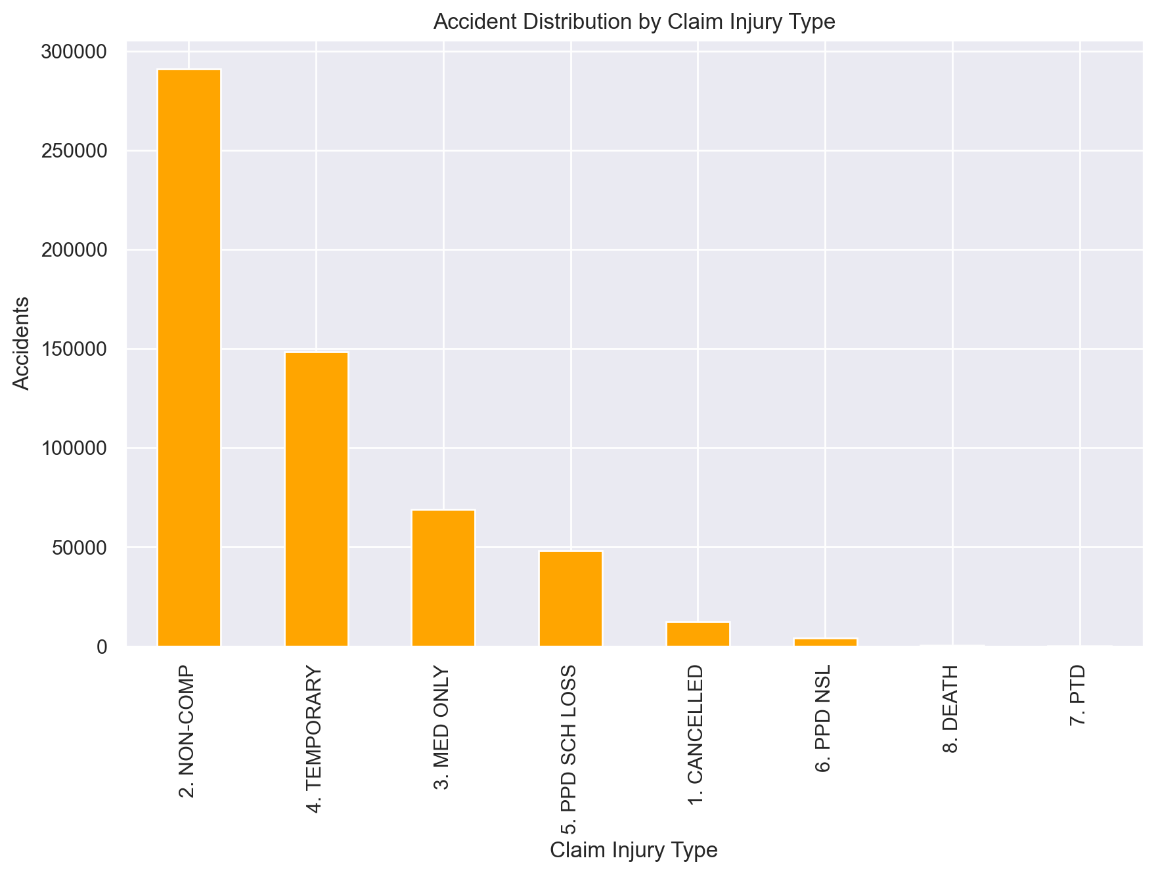


Figure 10: Distribution by Claim Injury Type

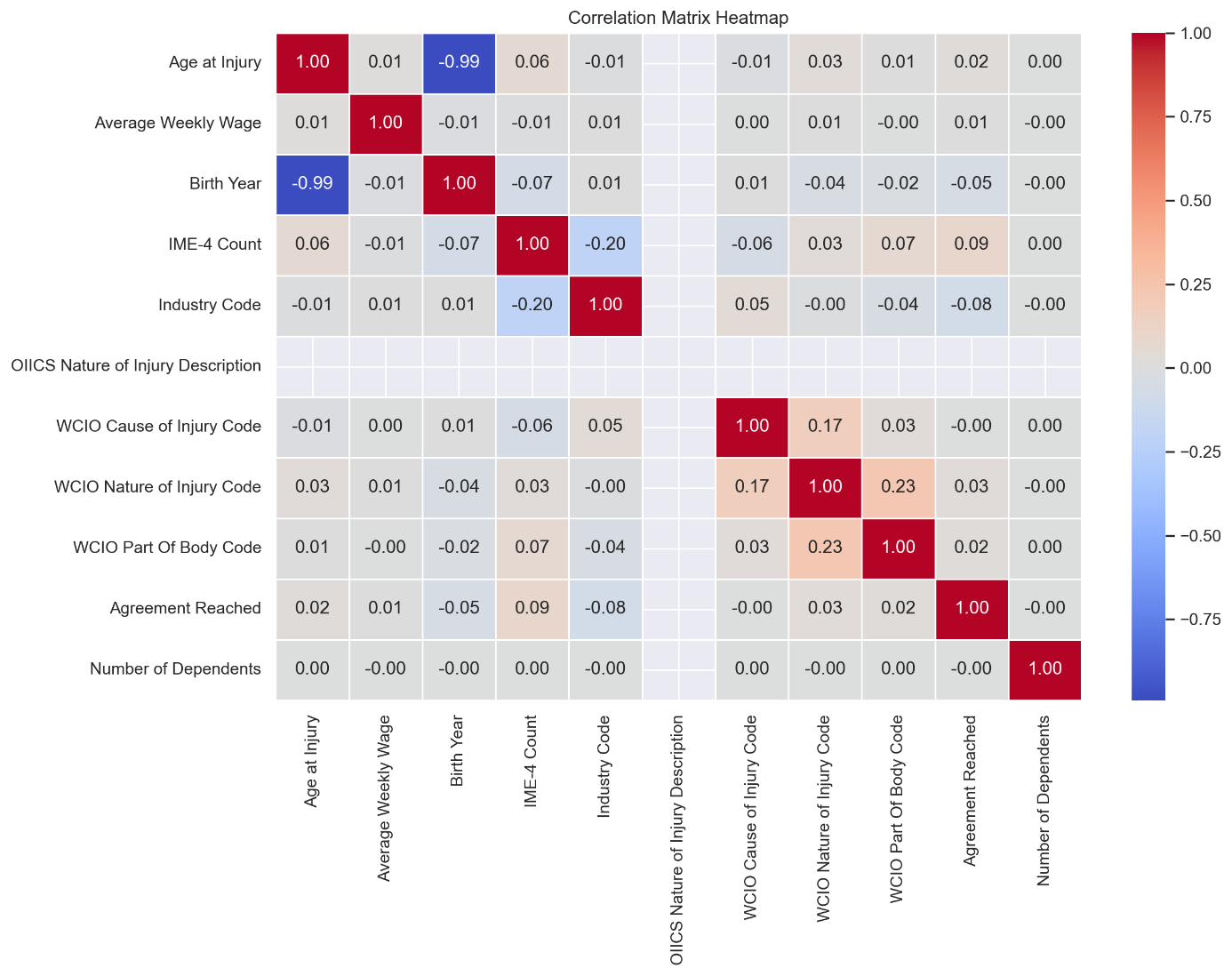


Figure 11: Correlation Matrix Heatmap

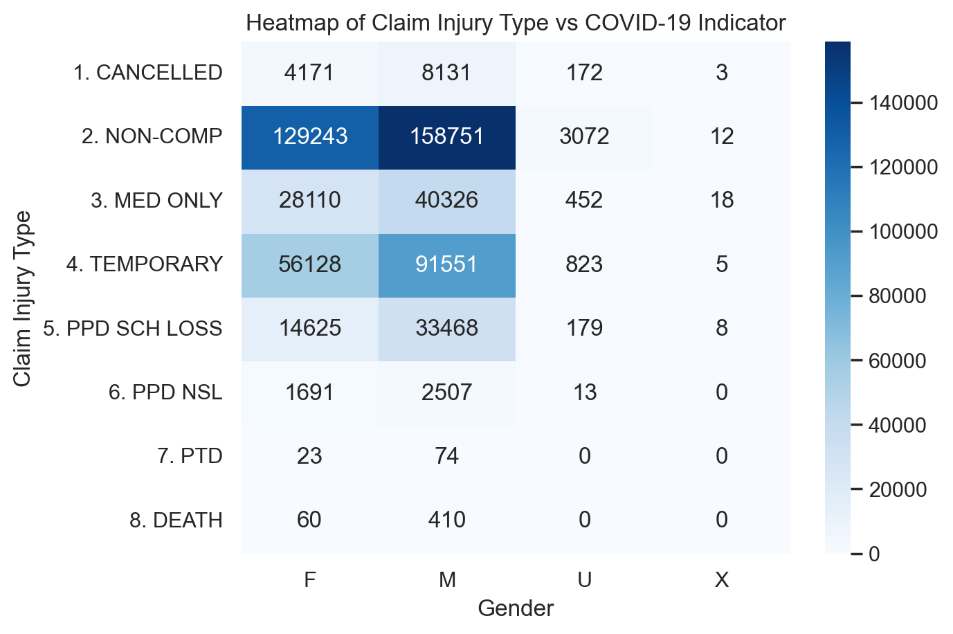


Figure 12: Heatmap of Claim Injury Type vs. Gender

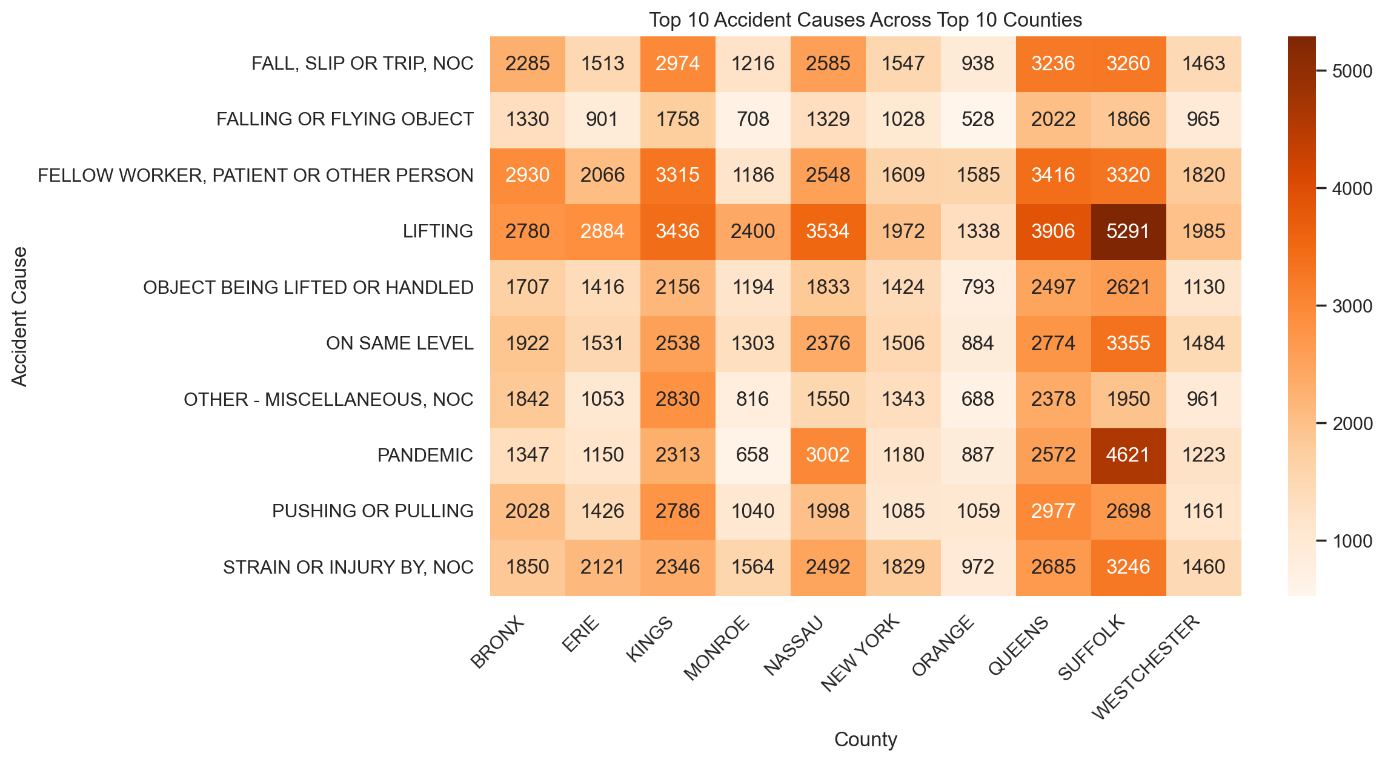


Figure 13: Top 10 Accident Causes Across Top 10 Counties

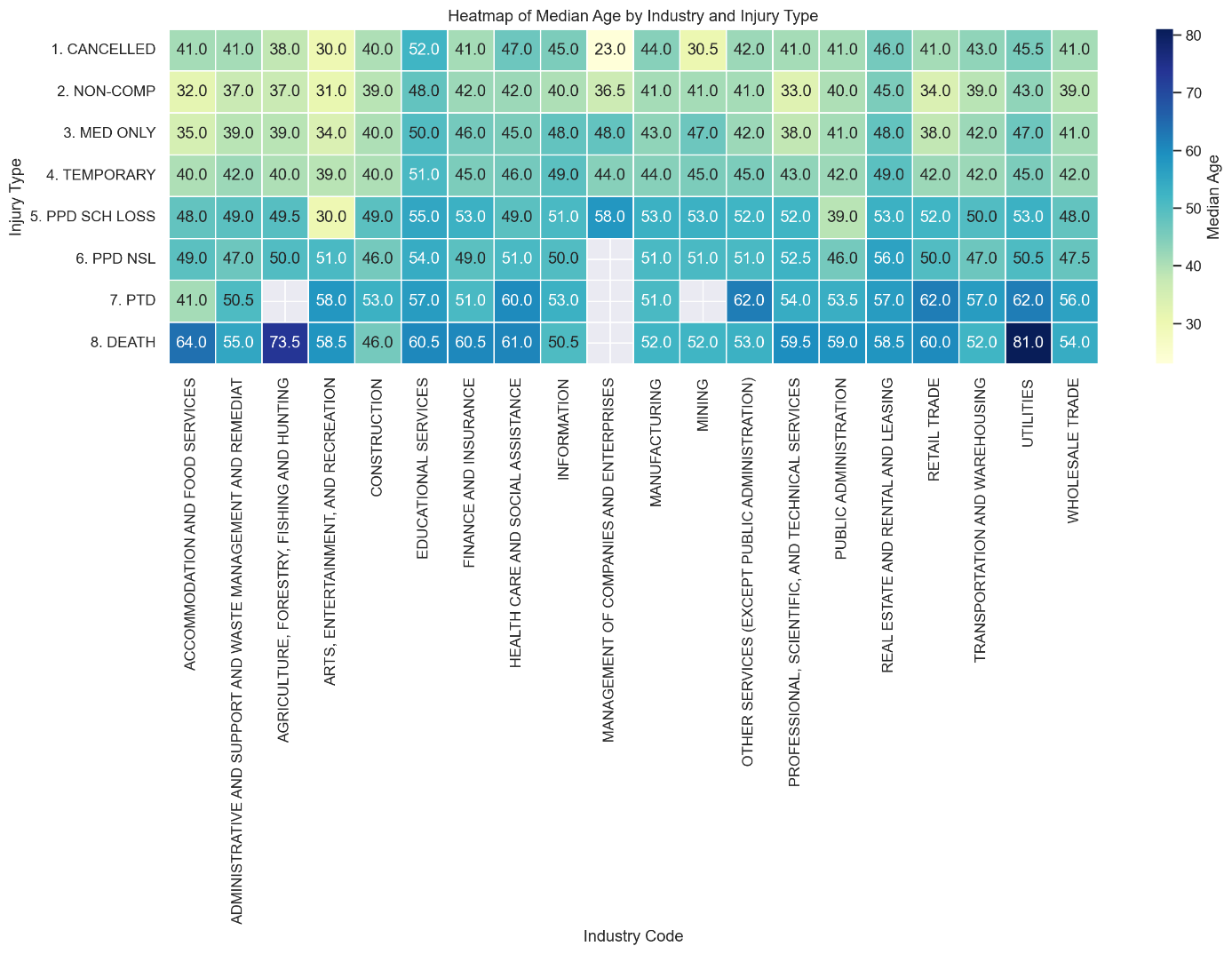


Figure 14: Heatmap of Median Age by Industry and Injury Type

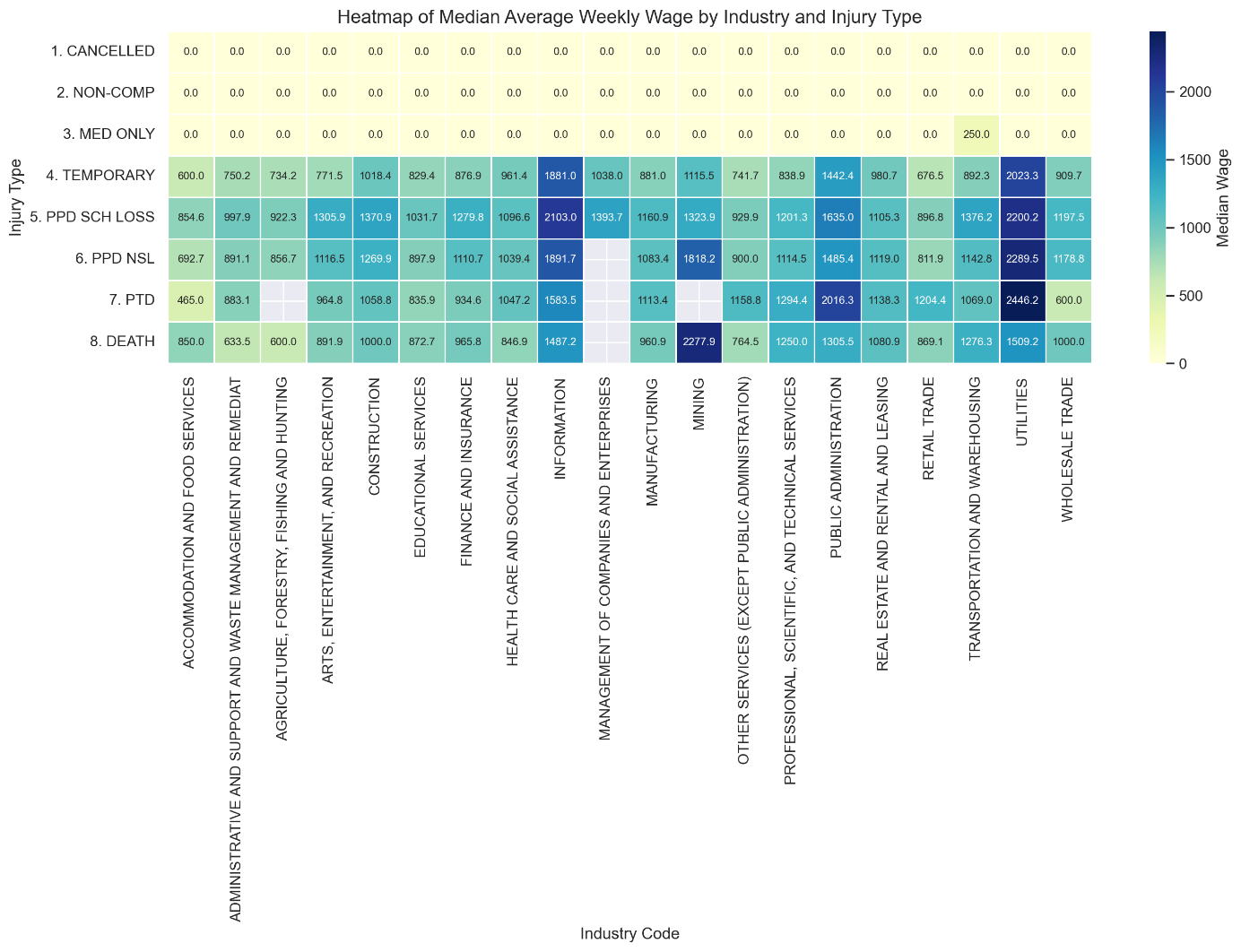


Figure 15: Heatmap of Median Average Weekly Wage by Industry and Injury Type

# Annexes (Optional, Not included in page limit)

[Annexes are optional, since they have material and sources not developed by the students, so in most cases referencing them is enough]

1. New York State Workers’ Compensation Board (2023). [↑](#footnote-ref-1)
2. Michael Lawrence Varon, PLLC (2018 – 2024). [↑](#footnote-ref-2)
3. Kanchetti, D. (2021) [↑](#footnote-ref-3)
4. Tjahjono, S., Murfi, H., Devila, S. (2024) [↑](#footnote-ref-4)
5. Rusdah, D. A., Murfi, H. (2020). [↑](#footnote-ref-5)
6. Tjahjono, S., Murfi, H., Devila, S. (2024). [↑](#footnote-ref-6)
7. Quan, Z., Valdez, E. A. (2018) [↑](#footnote-ref-7)
8. <https://www.wcb.ny.gov/> [↑](#footnote-ref-8)
9. <https://www.wcio.org/> [↑](#footnote-ref-9)