MACHINE LEARNING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Machine Learning Handout

To Grant or Not to Grant

Group 26

Pedro Santos, 2040295

Diogo Correia, 20211586

Oumaima Hfaieh, 20240699

Rita Morgadito, 20240611

Duarte Miguel, 20240608

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

We developed a multi-classification model to assist New York's WCB in compacting their decision-making process for work-related claims, significantly reducing the need for manual review. This initial phase features a functional model, along with a concise overview of the methodology used in its development as well as our proposed plans for the next phase.

2. EDA

2.1. GRAPH VISUALIZATION AND ANALYSIS

We began by thoroughly understanding the task at hand, creating a comprehensive Exploratory Data Analysis (EDA) to explore relationships across numerical and categorical data types in connection with our target variable, *“Claim Injury Type.”* For this, we developed three specialized functions (for numerical values, dates, and categorical values) to generate visualizations that would help us identify inconsistencies, outliers, missing values, and other data quality issues. To ensure accurate analysis, we adjusted data types to allow each function to process values correctly.

In examining numerical features, we employed histograms and boxplots, focusing particularly on the target variable. Most features exhibited outliers except for a few, such as *“Number of Dependents”* (**Annex A: Figure A1**), *“WCIO Nature of Injury Code”* (**Annex A: Figure A2**), *“WCIO Cause of Injury Code”* (**Annex A: Figure A3**), and *“Industry Code”* (**Annex A: Figure A4**). Some fields, like *“Age at Injury”* (**Annex A: Figure A5**), also contained unusual values that require further handling. For date-related features, outliers were prevalent in most fields, with exceptions like *“First Hearing Date”* (**Annex B: Figure B1**) and *“Assembly Date”* (**Annex B: Figure B2**). Lastly, categorical features visualized with histograms and pie charts revealed low frequencies in fields like *“Alternative Dispute Resolution”* (**Annex C: Figure C1**), *“Carrier Type”* (**Annex C: Figure C2**), *“County of Injury”* (**Annex C: Figure C3**), *“Gender”* (**Annex C: Figure C4**), and *“Agreement Reached”* (**Annex C: Figure C5**), which we’ll need to manage carefully due to the data’s sparseness.

3. DATA PREPROCESSING (WITH SCALING & ENCODING)

Our EDA revealed significant missing data and an imbalance in class distributions, leading us to choose the macro F1 score as our primary metric to promote balanced model performance. We used various imputation methods to address missing values, but around 20,000 rows contained only *“Claim Identifier”* and *“Assembly Date”* data and were deemed unsuitable for imputation. To prevent data leakage, we split the data into training and validation sets early in the process. In preprocessing, we streamlined certain features through encoding and aggregation to reduce noise while capturing key patterns, although only one engineered feature was ultimately retained for this phase.

Robust Scaling was applied to numerical features to mitigate the impact of outliers by scaling based on the median and IQR, a method well-suited to KNN imputation. For categorical data, we used an Ordinal Encoder, and label-encoded the target variable, fitting all transformations solely on the training set to avoid leakage. Afterward, we combined scaled and encoded datasets for feature selection.

3.1. FEATURE SELECTION:

For feature selection, we applied all the methods covered in class. These included Spearman's correlation matrix for identifying correlated features, Recursive Feature Elimination (RFE) with Logistic and Random Forest classifiers to optimize feature count, and Lasso regression. Feature selection was decided by a majority vote across methods, though some trial and error were involved, with adjustments based on model performance.

4. Model Assessment:

The purpose of this first submission was to develop a simple model. We elaborated a function that allowed us to check the F1 Macro baseline scores for several models, with cat boost performing the best out of our selection. We also tuned several models individually, with the purpose of upgrading the model performance. It’s important to be aware that after hyper parameterization, models have the tendency to overfit, an issue that might cause a drop off in our scores due to worse generalization of the model (In some cases).

We elaborated various submissions on Kaggle, with different variations of models and parameters. We quickly realized that checking if our model is overfitting or not is essential for its generalization on our test set! The best scoring submission ended up being the Cat Boost with hyper parameter tuning. Other models were also kept on the notebook for future reference if needed, although some modifications were entirely dropped due to overfitting.

5. Conclusion and Open-Ended Section For Final Delivery:

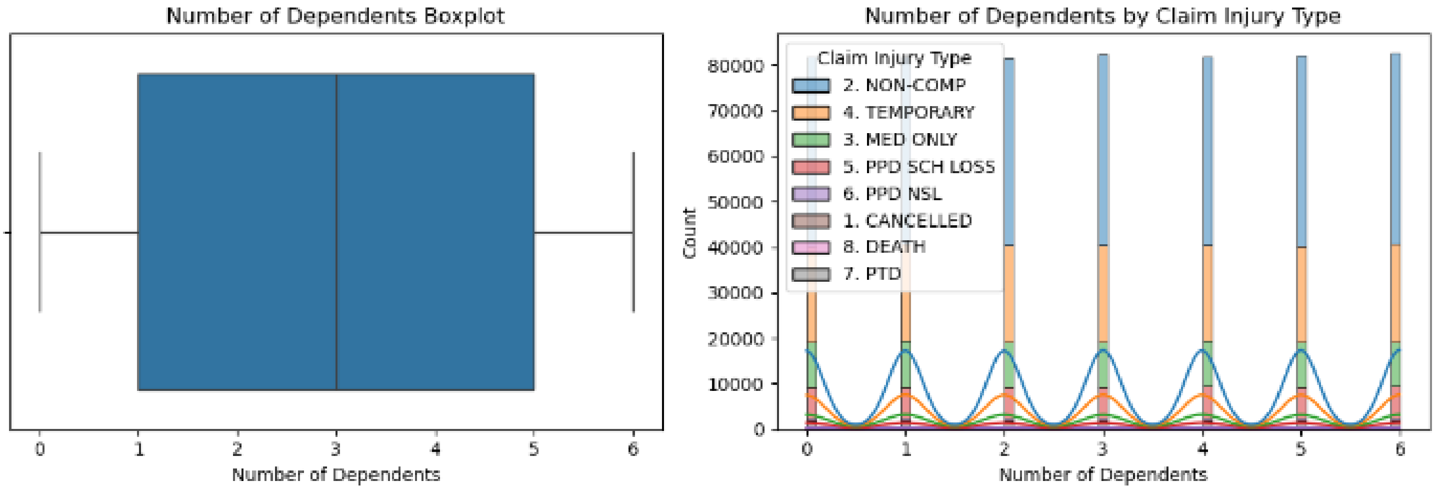
It’s important to note that, although we fine-tuned some models, the models in itself are not extremely complex. For the second part of our work, we could perhaps elaborate a Neural Network which should perform better. In terms of the preprocessing, although the scores at the time of writing are promising, it might be subject to some changes, depending on the task at hand.

Regarding the open ended section, we will try to develop an interface for New York's WCB in order to provide all users with an experience that allows them, like a credit simulator, to simulate potential results for injuries that they may actually have, so the primary objective would be for the user to submit their characteristics, type of injury, among other features present in the model, and the user would obtain their “predicted” type of compensation, or possibly create a model with a binary variable, something that still has to be defined.

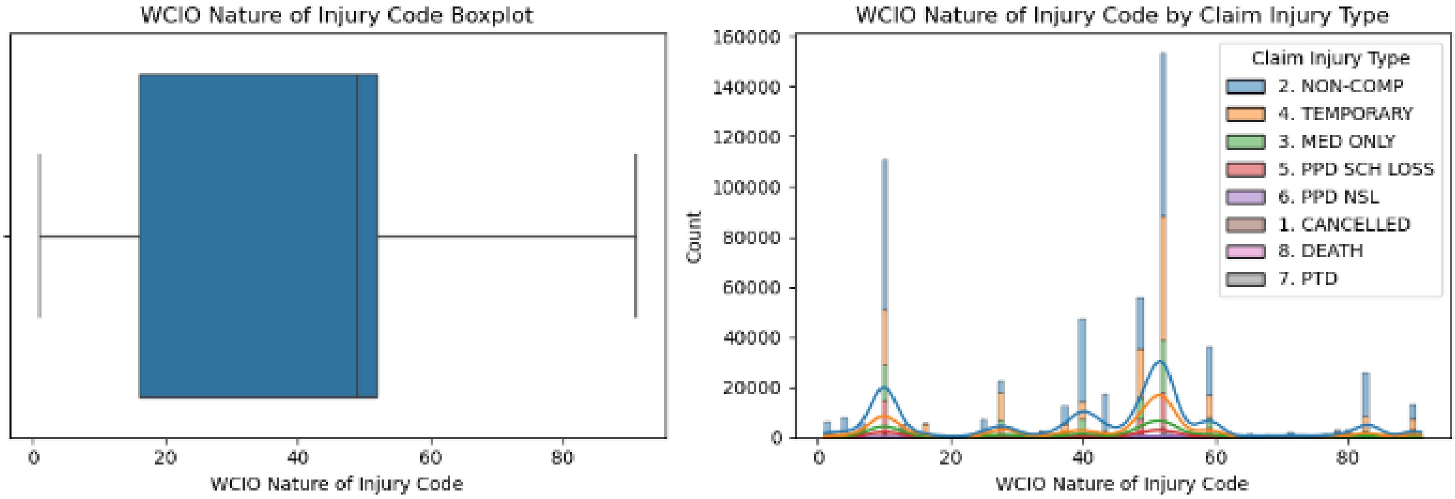
6. Annexes:

6.1. Numerical Variables (A):

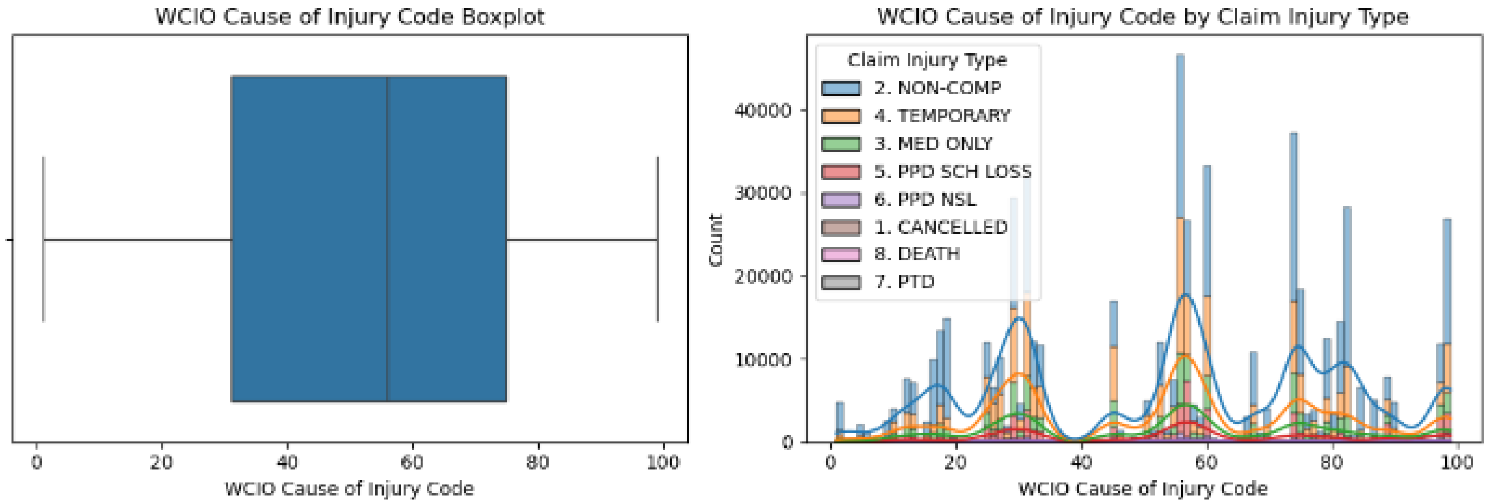
6.1.1. Figure A1 (Number of Dependents):



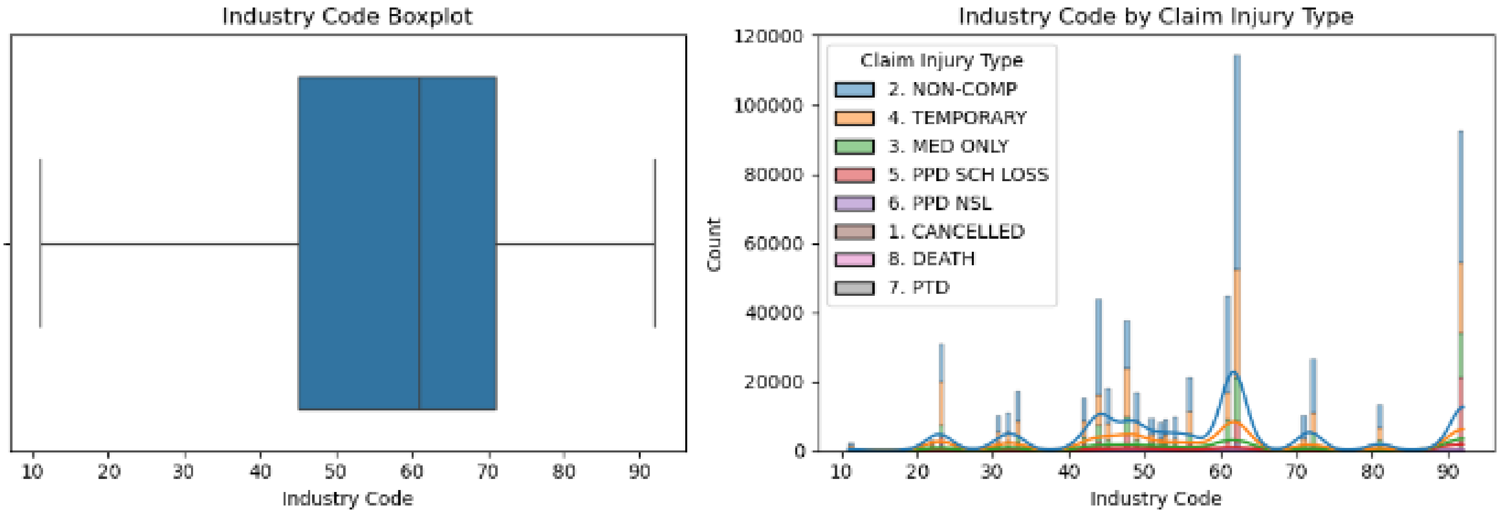
6.1.2. Figure A2 (WCIO Nature of Injury Code):



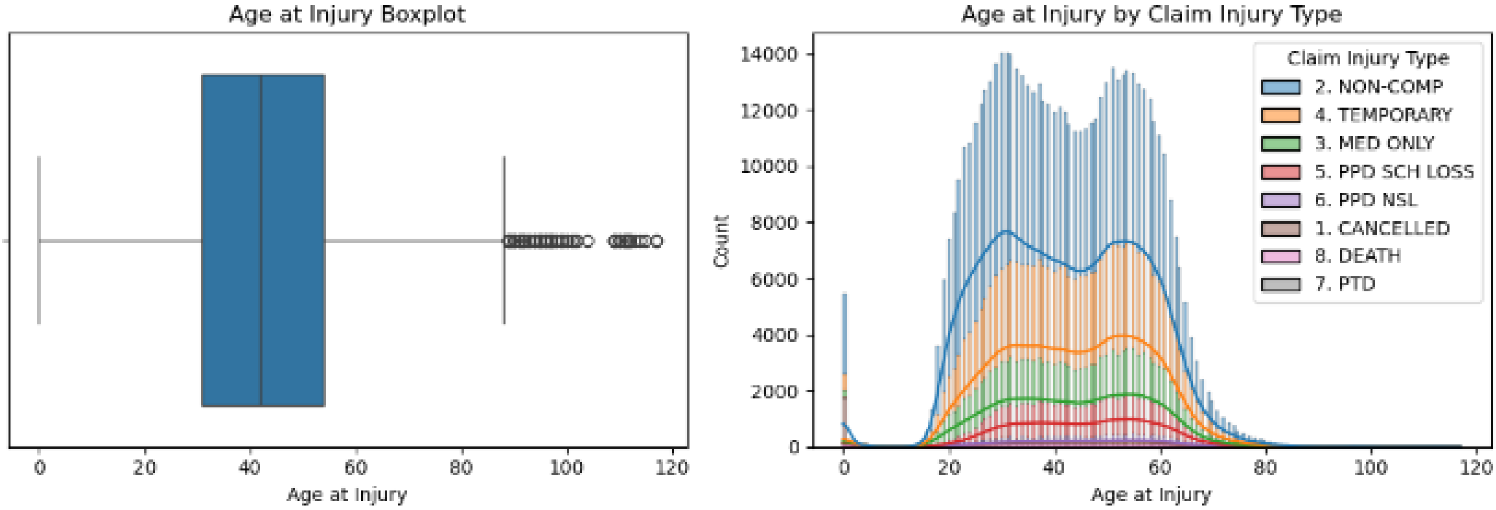
6.1.3. Figure A3 (WCIO Cause of Injury Code):



6.1.4. Figure A4 (Industry Code):

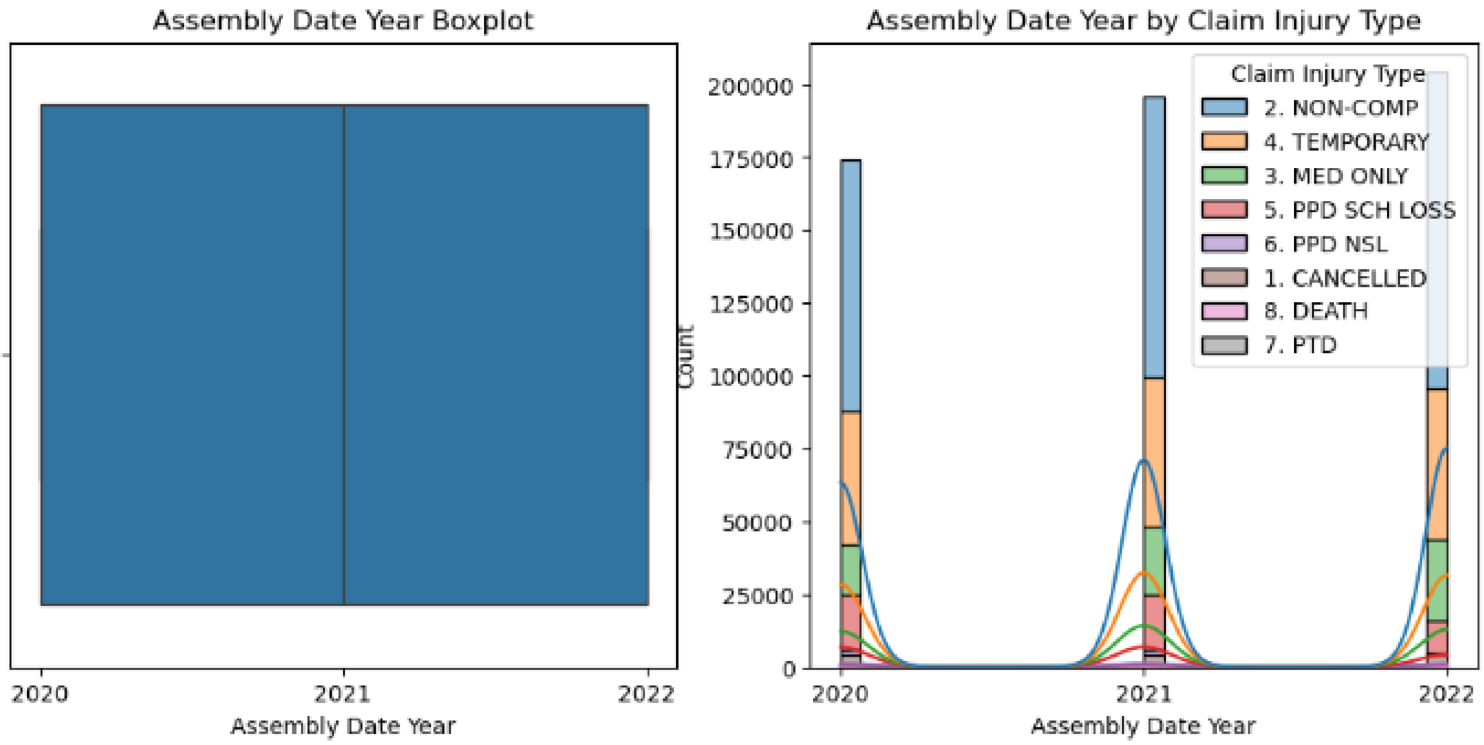


6.1.5. Figure A5 (Age at Injury):

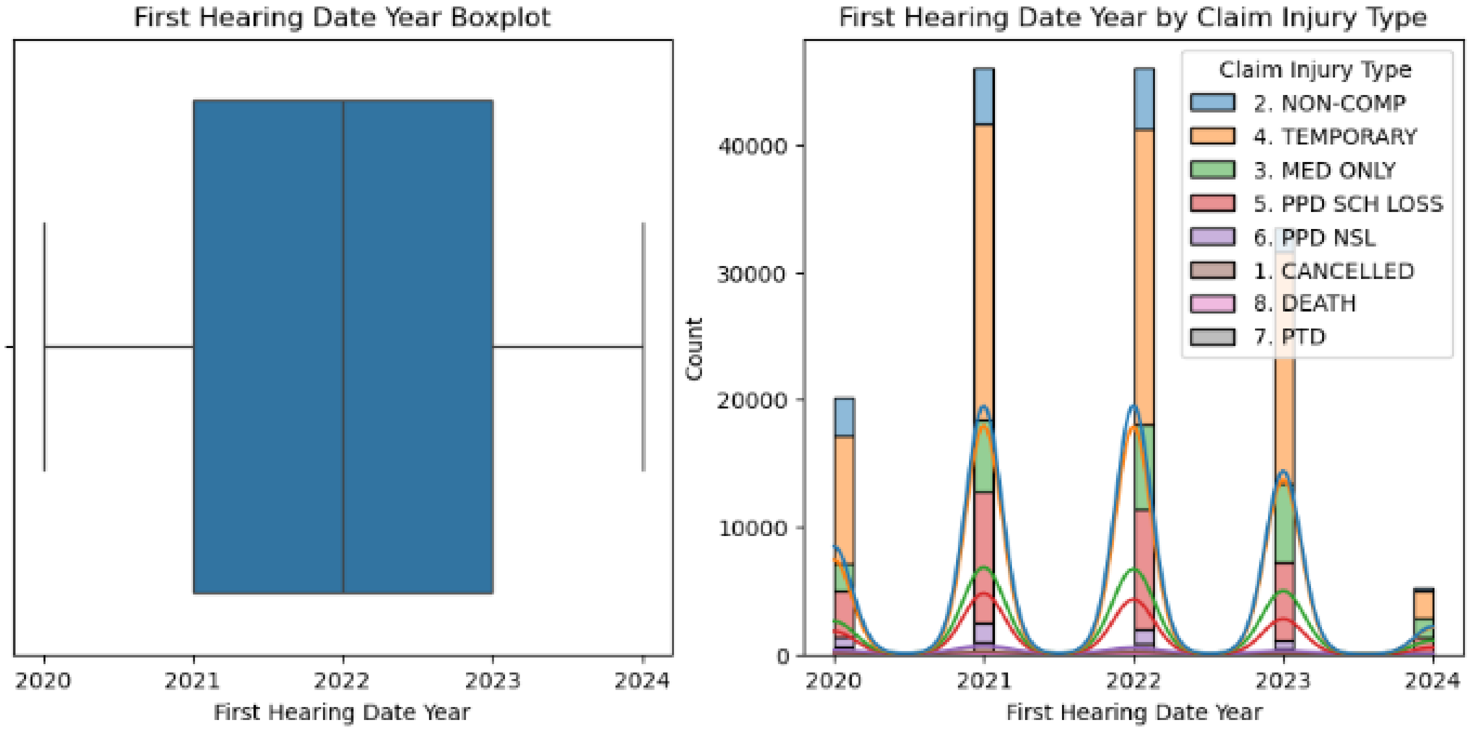


6.2. Date Variables (B):

6.2.1. Figure B1 (Assembly Date):

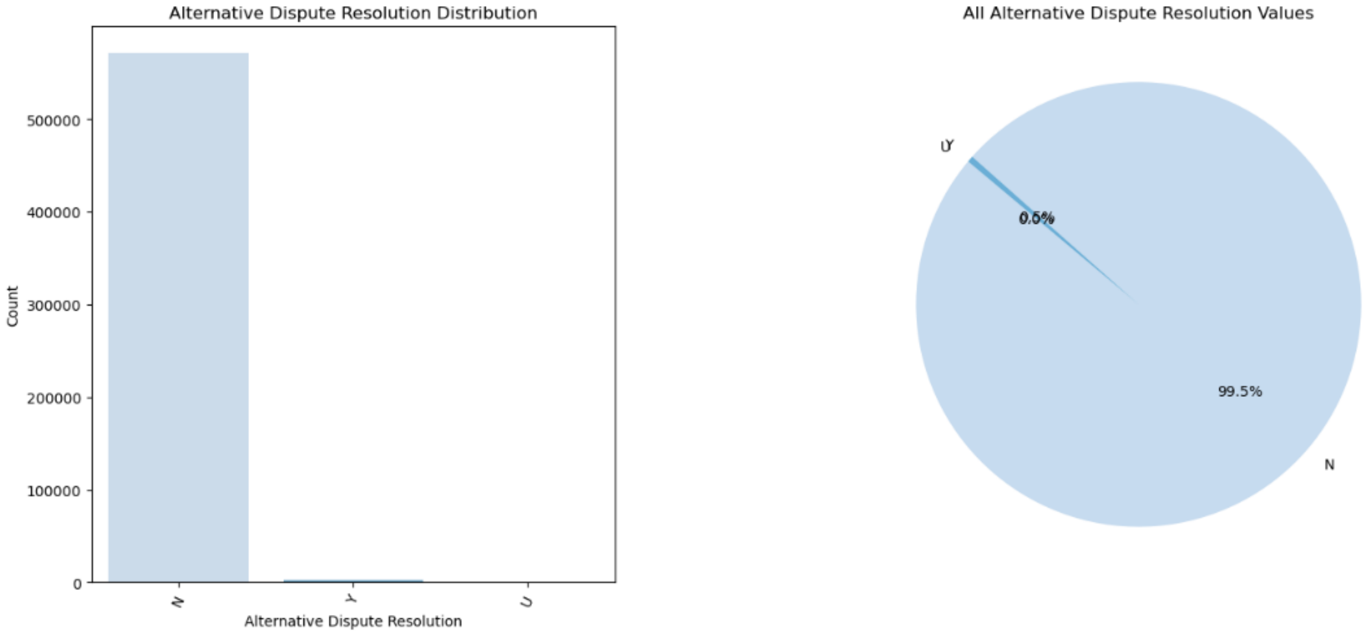


6.2.2. Figure B2 (First Hearing Date):

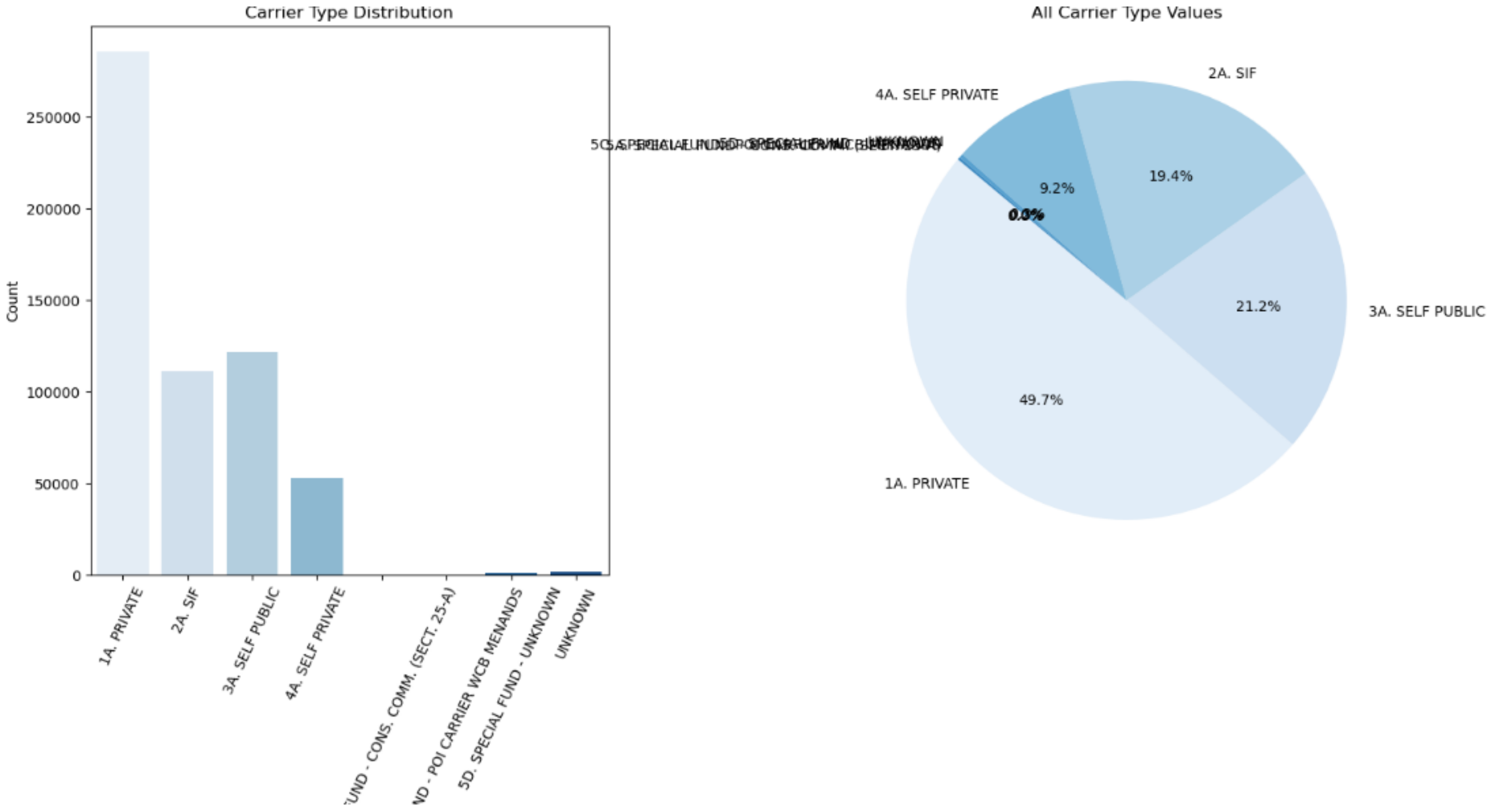


6.3. Categorical Variables (C):

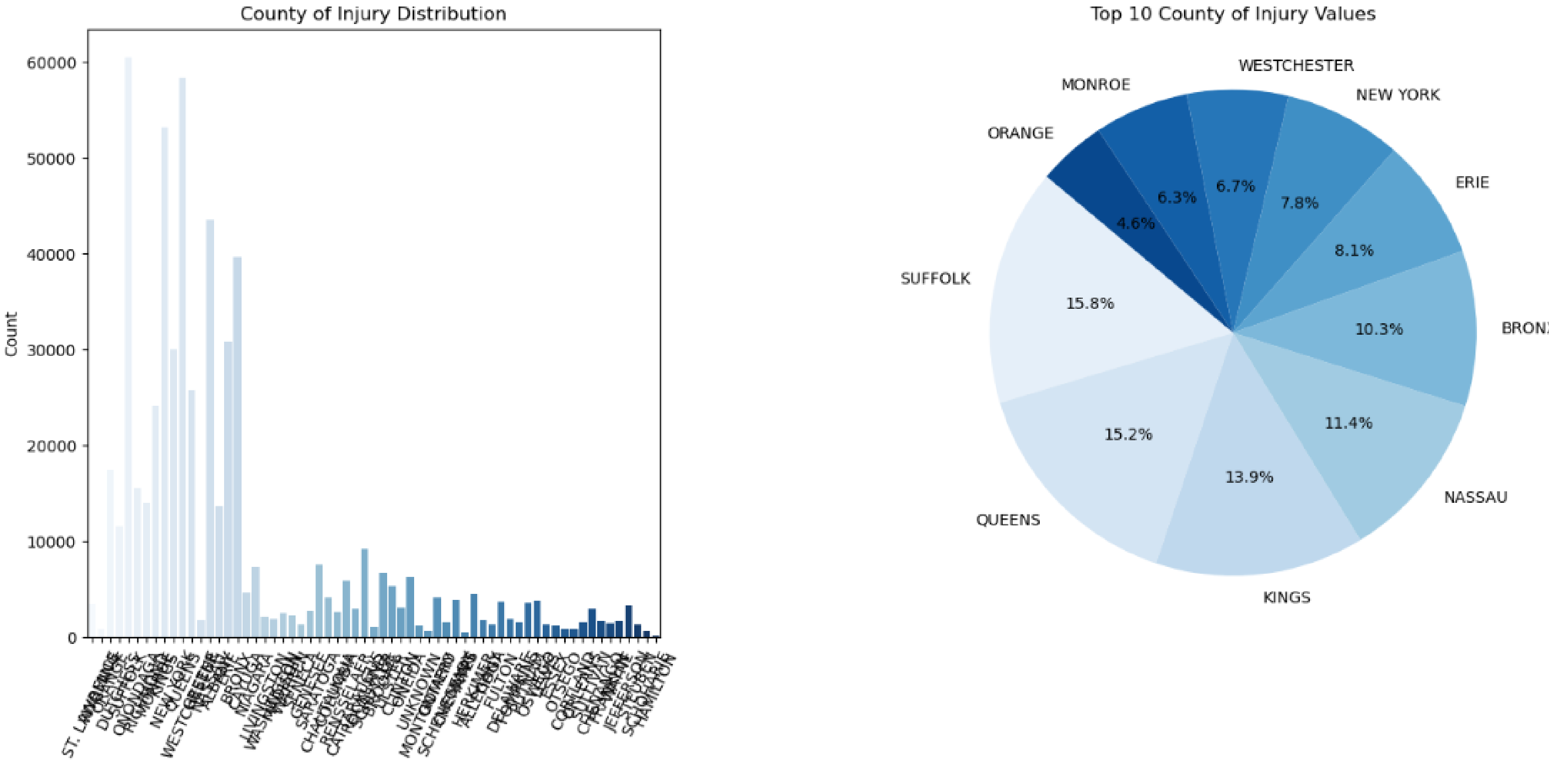
6.3.1. Figure C1 (Alternative Dispute Resolution):



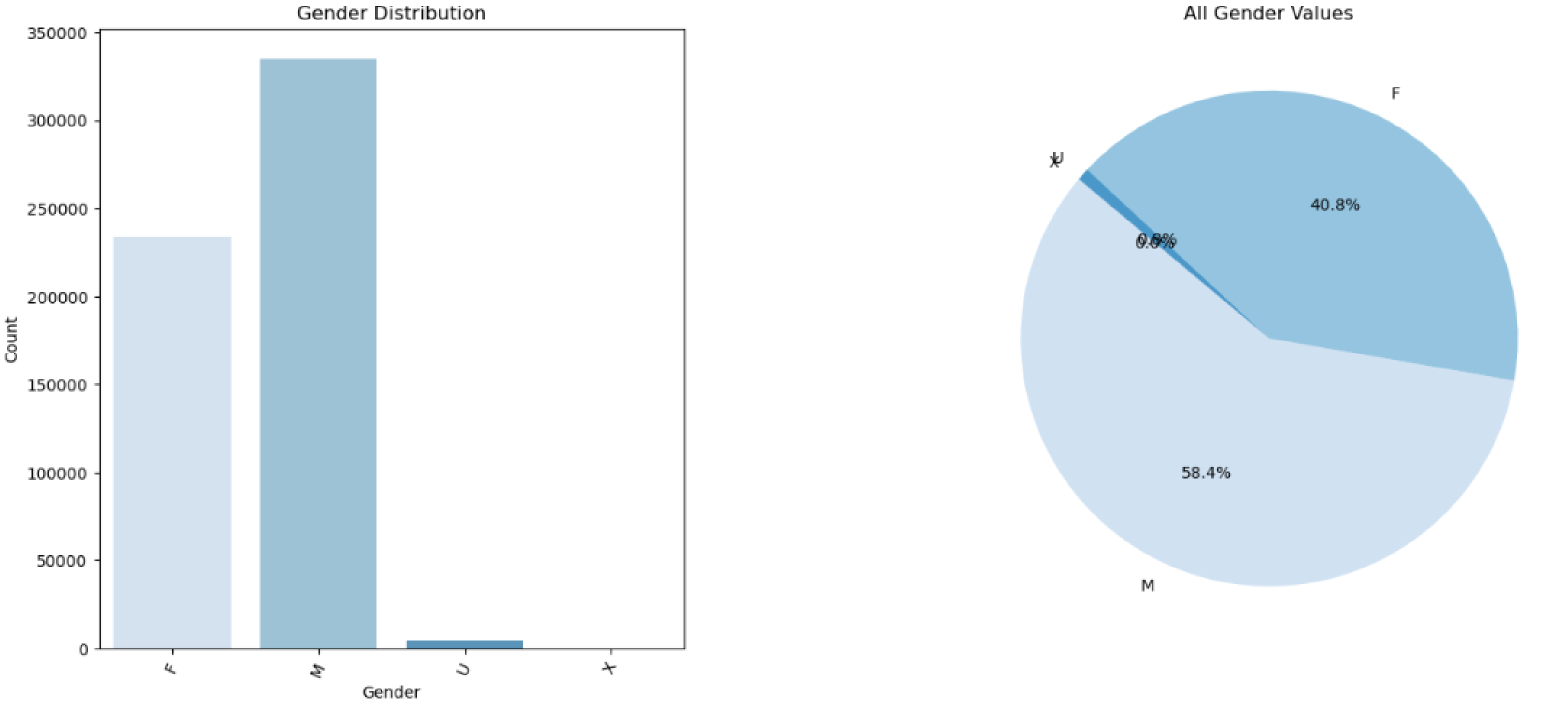
6.3.2. Figure C2 (Carrier Type):



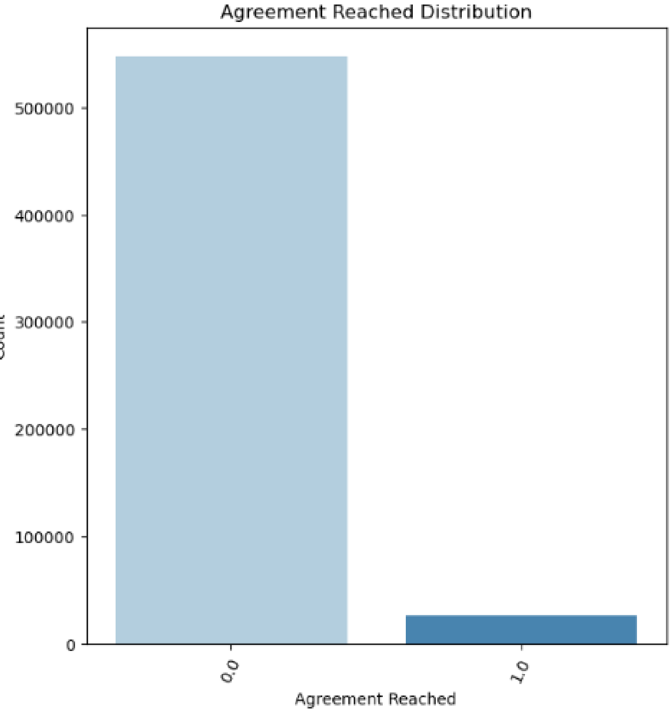
6.3.3. Figure C3 (County of Injury):

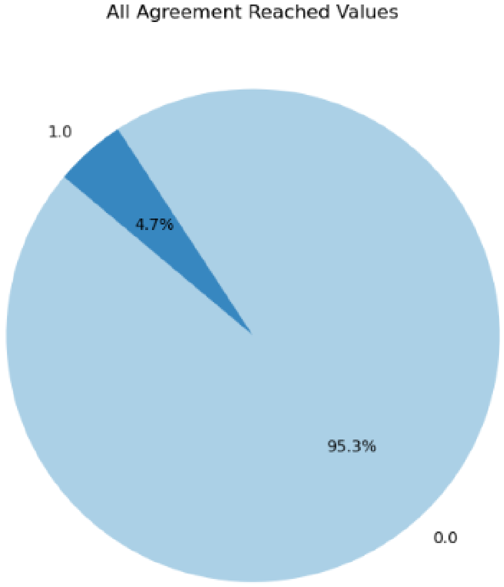


6.3.4. Figure C4 (Gender):



6.3.5. Figure C5 (Agreement Reached):


Agreement Reached Distribution
Count
500000
400000
300000
200000
100000
0
0.0
1.0
Agreement Reached


All Agreement Reached Values
1.0
4.7%
95.3%
0.0