**An Analysis of the San-Diego Traffic Collisions Dataset**

Neil Kingdom – 040967309  
Computer Engineering Technologies, Algonquin College  
CST8390: Business Intelligence and Data Analytics  
Dr. Anu Thomas  
April 14, 2023  
  
  
  
  
  
  
  
  
  
  
Question: *Which factors contribute the most towards vehicle-related incidents involving injury and/or homicides?*

# Link to Dataset: https://data.sandiego.gov/datasets/police-collisions-details/

Table of Contents

[Link to Dataset: https://data.sandiego.gov/datasets/police-collisions-details/ 1](#__RefHeading___Toc9104_2374119336)

[1.0 Introduction 1](#__RefHeading___Toc2721_2374119336)

[2.0 Data Understanding 2](#__RefHeading___Toc2723_2374119336)

[2.1 Initial Dataset 2](#__RefHeading___Toc2725_2374119336)

[2.2 Data Exploration 5](#__RefHeading___Toc2727_2374119336)

[8](#__RefHeading___Toc9086_2374119336)

[9](#__RefHeading___Toc9088_2374119336)

[3.0 Data Preparation 9](#__RefHeading___Toc14008_4068621970)

[3.1 Data Selection 10](#__RefHeading___Toc14010_4068621970)

[3.2 Cleaning Data 12](#__RefHeading___Toc14012_4068621970)

[3.3 Constructing Data 15](#__RefHeading___Toc14014_4068621970)

[3.4 Integrating Data 19](#__RefHeading___Toc14016_4068621970)

[3.5 Formatting Data 19](#__RefHeading___Toc14018_4068621970)

[4.0 Modeling and Evaluation 19](#__RefHeading___Toc14020_4068621970)

[4.1 Using a Decision Tree 20](#__RefHeading___Toc9090_2374119336)

[4.2 Using a Random Forest 23](#__RefHeading___Toc9092_2374119336)

[5.0 Discussion of Results 25](#__RefHeading___Toc2729_2374119336)

[25](#__RefHeading___Toc2731_2374119336)

[6.0 Conclusion 28](#__RefHeading___Toc9094_2374119336)

[7.0 References 30](#__RefHeading___Toc9096_2374119336)

# 1.0 Introduction

The city of San-Diego provides public statistics regarding information about its citizens on the internet website data.sandiego.gov. One of these datasets is the Traffic Collisions dataset (with people and vehicles involved). This report will attempt to answer a simple question by analysing this dataset. The question is as follows: Which factors contribute the most towards vehicle-related incidents involving injury and/or homicides? Prior to answering these questions, a few hypotheses can be explored. Out of the handful of features provided within the dataset, it can be assumed that the following may have some correlation with injury and homicides: the time at which the incident occurred, the type of vehicles involved in the incident, the location of the incident (indicated by the San-Diego Police beat), the type of road that the incident occurred on (indicated by numerous features, such as street number, street name, street suffix, and intersection name), the violation type of the primary collision factor, and the hit and run level. It is estimated that incidents will be more likely to occur on Fridays, and during the weekend, as these are typically the busiest times of the week. Considering that rush hour occurs around 7am to 10am and 4pm to 6pm, it can be estimated that incidents will be more likely to occur during these times, increasing the risk of injury and/or death. Streets which are busier and have a higher level of traffic congestion are estimated to produce more severe collisions. Finally, hit and runs associated with felonies are expected to have a higher correlation with injuries and homicides in comparing to misdemeanors, considering that hit and runs marked as felonous are more severe by definition (whereas misdemeanors are less severe by definition).

# 2.0 Data Understanding

The San-Diego dataset consists initially of 119, 297 instances and 22 features. Each instance within the dataset represents a single person involved in a vehicle-related incident which occurred within the city of San-Diego, and was reported by the San-Diego Police Department. Thus, the same collision may be reported multiple times (once per person involved). If this is the case, the same report ID will be reused for each person involved in the collision. The dataset is updated yearly by the city of San-Diego. As of current date, it contains records from between 2015-2023. As noted on data.sandiego.gov, reports are *generally* not made for property damage-only incidents that do not involve a hit and run or DUI. Additionally, the dataset only records incidents that occurred on a street, as incidents occurring on the freeway are handled by the California Highway Patrol.

## 2.1 Initial Dataset

The city’s website (data.sandiego.gov) provides two additional resources alongside the dataset. The first resource is the ‘person\_veh\_type dictionary’. This CSV file contains a list of nominal values that may be attributed to the person\_veh\_type feature. The second resource is the ‘Traffic collisions details dictionary’, which provides a table of attributes and their descriptions, respectively. As stated on the city’s website for the dataset, beginning in 2018, the San-Diego Police Department began recording more details for each collision. The two new attributes which were introduced were person\_veh\_type, which provides a description of the vehicle, and veh\_type, which provides a nominal value that represents the broad category of vehicle. The person\_veh\_type dictionary is essentially obsoleted by veh\_type, as the categories are too specific to be of any use when analysing the dataset. A table containing each attribute within the dataset and their respective descriptions is displayed below:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| report\_id | Numeric | The collision report number |
| date\_time | String | The date and time at which the collision was reported |
| person\_role | Nominal | A nominal value describing the person’s involvement in the collision |
| person\_injury\_level | Nominal | A nominal value describing the level of injury inflicted upon the victim |
| person\_veh\_type | String | A description of the vehicle |
| veh\_type | Nominal | A nominal value describing the broad category of vehicle |
| veh\_make | Nominal | The make of the vehicle |
| veh\_model | Nominal | The model of the vehicle |
| police\_beat | Numeric | A numeric value associated with the San-Diego Police beat, within which the incident occurred |
| address\_no\_primary | Numeric | The street number of the collision location (if applicable) |
| address\_pd\_primary | Nominal | The direction of the street at which the collision occurred (if applicable) |
| address\_road\_primary | Nominal | The name of the street at which the collision occurred |
| address\_sfx\_primary | Nominal | The type of street at which the collision occurred |
| address\_pd\_intersecting | Nominal | The direction of the cross street at which the collision occurred (if at an intersection) |
| address\_name\_intersecting | String | The street name at which the collision occurred (if at an intersection) |
| address\_sfx\_intersecting | Nominal | The type of street at which the collision occurred (if at an intersection) |
| violation\_section | Nominal | The section code that proceeds the violation\_type, describing the exact type of violation for the primary collision factor |
| violation\_type | Nominal | A nominal value describing the broad category of violation for the primary collision factor (preceeds violation\_section) |
| charge\_desc | String | A description of the violation type for the primary collision factor |
| injured | Numeric | A numeric count of the number of people injured from the incident |
| killed | Numeric | A numeric count of the number of people killed from the incident |
| hit\_run\_level | Nominal | A nominal value describing the level of violation (if the collision was a hit and run) |

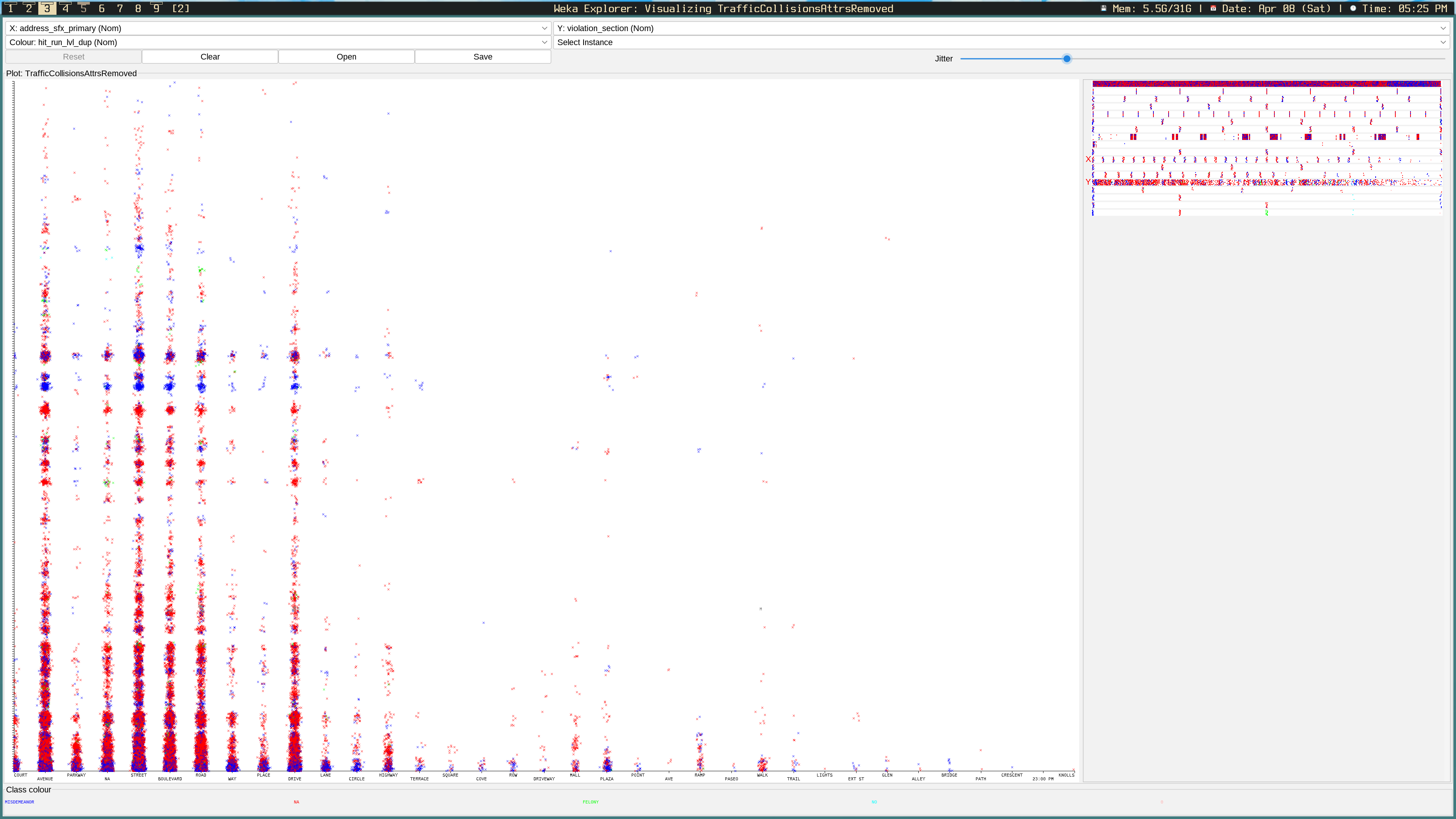
Many of the instances within the dataset were missing one or more attribute values. The methodology in terms of how missing values were handled is explained in section 3.0. The table below reveals the ratio of known to unknown attribute values for each attribute, without taking into account values which were set incorrectly.

|  |  |  |
| --- | --- | --- |
| Attribute | Not Missing : Missing | Percentage Missing |
| report\_id | 119, 297 : 0 | 0% |
| date\_time | 119, 297 : 0 | 0% |
| person\_role | 119, 297 : 3, 205 | 2.69% |
| person\_injury\_lvl | 119, 297 : 90, 295 | 75.69% |
| person\_veh\_type | 119, 297 : 68, 887 | 57.74% |
| veh\_type | 119, 297 : 54, 395 | 45.60% |
| veh\_make | 119, 297 : 12, 595 | 1.06% |
| veh\_model | 119, 297 : 37, 622 | 31.54% |
| police\_beat | 119, 297 : 0 | 0% |
| address\_no\_primary | 119, 297 : 11, 219 | 0% |
| address\_pd\_primary | 119, 297 : 114, 140 | 95.68% |
| address\_road\_primary | 119, 297 : 51 | 0.43% |
| address\_sfx\_primary | 119, 297 : 7, 160 | 6.00% |
| address\_pd\_intersecting | 119, 297 : 118, 988 | 99.74% |
| address\_name\_intersecting | 119, 297 : 109, 829 | 92.06% |
| address\_sfx\_intersecting | 119, 297 : 110, 782 | 92.86% |
| violation\_section | 119, 297 : 33 | 0.28% |
| violation\_type | 119, 297 : 0 | 0% |
| charge\_desc | 119, 297 : 0 | 0% |
| injured | 119, 297 : 29 | 0.24% |
| killed | 119, 297 : 3 | 0% |
| hit\_run\_lvl | 119, 297 : 70, 772 | 59.33% |

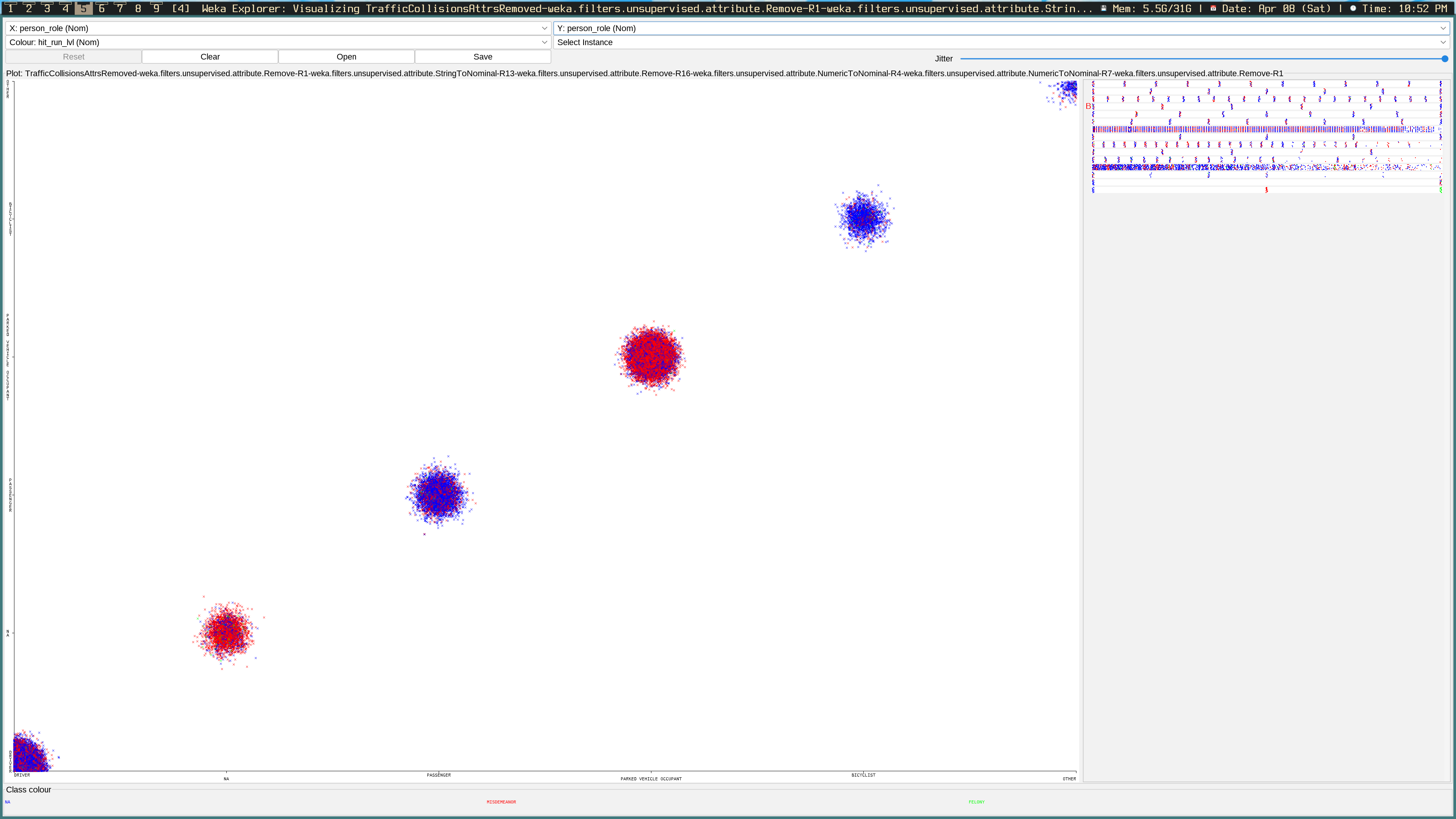
## 2.2 Data Exploration

The following section covers a few images which showcase the data exploration phase of the CRISP-DM standard. Data exploration was performed to gain a better understanding pertaining to the relationships between attributes within the dataset.

Figure 1 displays the primary address suffix of the collision (y axis) over the various violation sections (x axis). The graph is interesting for a couple of reasons. First, we can see that certain road types are much more prevalent than others judging by the large pillars on the left-hand-side. For example, we can see avenues, parkways, streets, boulevards, etc. as having many more collisions than plazas, ramps, rails, etc. Second, many of the violation sections only apply towards the more collision-heavy street types. Finally, looking at the colorization, which represents the hit and run level, we can see that there must be a specific category of violation sections for hit and runs classified as felonies, judging by the fact that there is a row of blue (representing hit and runs associated with felonies) closer to the peaks of the more traffic-heavy street types.

Figure 1: Primary Address Suffix vs. Violation Section

The plot seen in Figure 2 is not concerned with modelling anything over the x and y axes (hence the reasoning for why person role is selected for both), as much as it is with the colorization, which represents the hit and run level. Interestingly, person roles NA and Parked Vehicle Occupant have might higher associations with hit and runs marked as misdemeanors, whereas drviers, passengers, bicyclists, and pedestrians have a much higher association with hit and runs marked as NA. Looking very closely, it would appear as though Parked Vehicle Occupants have a lower association with hit and runs marked as felonies than the other person roles.

Figure 2: Person Role With Hit and Run Level Highlighted

# 

Figure 4: Top 5 Violation Sections Related to Homicides

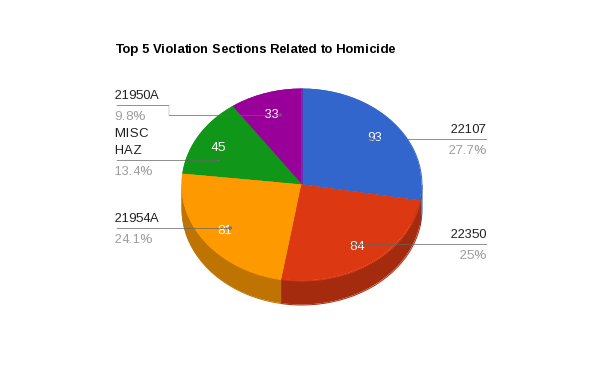
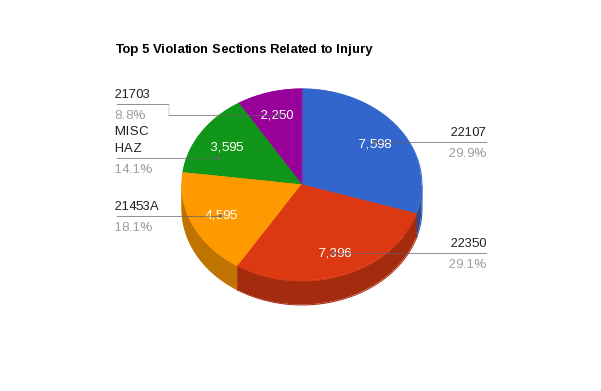
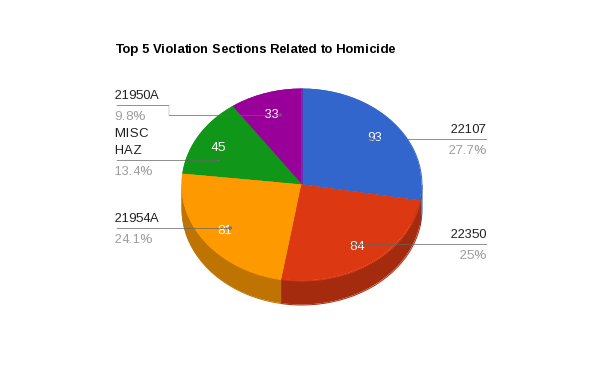
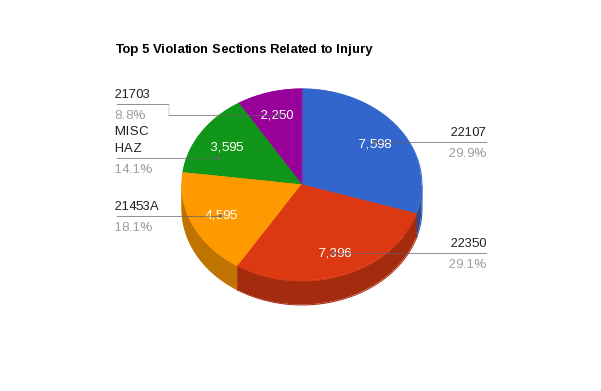


Figure 3: Top 5 Violation Sections Related to Injury



Figures 3 and 4 display pie charts for top 5 violation sections related to both injury and homicides, respectively. Both pie charts share the same two violation sections for both injuries and homicides, which are sections 22107 (unsafe turning), followed by 22350 (speeding). In third place for injuries is violation code 21453A (not stopping for a red light) and for homicides is 21954A (pedestrians not yeilding the right of way e.g., jaywalking). Fourth place in both cases goes to MISC-HAZ, which stands for miscellaneous hazards. This could be things like icey roads, or a deer standing in the freeway. Finally, in fifth place for injuries is 21703 (tailgating), and for homicides it goes to 21950A (cars not giving the right of way to pedestrians).

# 

Figure 5: Ratio of Collisions With Injuries/Deaths to Non-Injuries/Deaths (%)

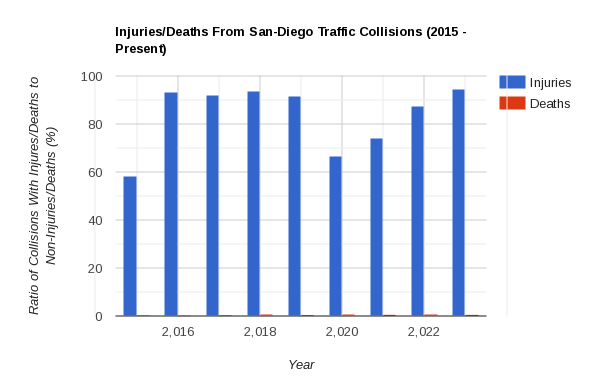
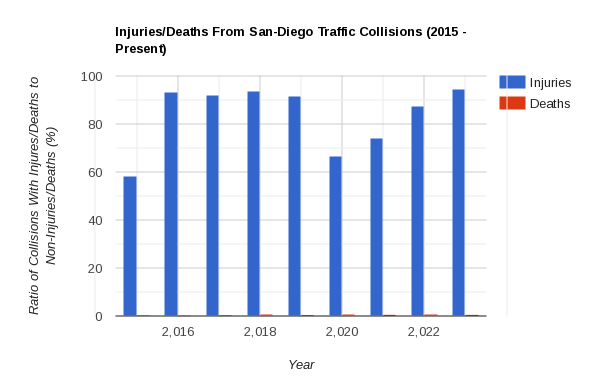


Figure 5 displays the *ratio* of collisions associated with injury and homicides to collisions which were not associated with injury or homicides over the recorded year. Note that 2015 is an outlier, as there are only 19 instances recorded for this year, thus making it a bad sample size for ratio. This chart is interesting for two primary reasons: the first is that it puts into perspective how few homicides are actually caused by car collisions in relation to those which only sufferred from some level of injury. The second reason that this chart is facinating is because of the sudden dip in collisions in the year 2020, with a steady rise up until current year. It is estimated that this dip is primary due to restrictions being placed on travel due to the Covid-19 pandemic which began in 2019.

# 3.0 Data Preparation

The dataset was prepared by following the guidelines of the CRISP-DM standard. Data preparation, according to CRISP-DM, is split into five stages: data selection, data cleaning, data construction, data integration, and data formatting. The following sections cover each stage of the process.

## 3.1 Data Selection

In order to create a predictive model that may answer the question outlined within the Introduction of the report, attributes that have a correlation with injury and/or homicides ought to be preserved, and attributes which are not correlated with these class values must be discarded. It should also be noted that many of the algorithms which will be used for classification are only capable of interpreting nominal or numeric values.

The list of attributes which are preserved, and the reasoning behind their preservation, are detailed below:

* **date\_time:** The date and/or time can have an effect on both the probability of injury and death, considering that the levels of traffic fluxuate in accordance with certain hours of the day. The day of the week and/or month may also have an effect on the afformentioned class values considering that weekends are busier, and certain months of the year may be associated with higher levels of tourism, and thus, more traffic.
* **person\_role:** The persons role within the collision is estimated to have a high correlation with the probability of injury and death. For instance, pedestrians and cyclists are much more likely to be killed than drivers or passengers.
* **veh\_type:** Larger vehicles such as heavy-duty trucks, have a higher chance of causing injury and death in opposition to, for example, motorcyclists, who have a higher change of being the victims of a collision due to the fact that they have less external protection and perform riskier manouvers in general.
* **police\_beat:** Certain police beats may have busier streets than others, resulting in higher probabilities of injury and death in comparison to police beats that contain a smaller populous.
* **address\_no\_primary:** Depending upon the way that street numbers are assigned in San-Diego, the address number of the street at which the collision occurred may or may not have an association with injury and death.
* **address\_pd\_primary:** San-Diego is located in the South-West corner of California, on the coast of the Pacific Ocean. For this reason, it is possible that traffic would tend to increase towards the North-East, as this is where people would be leaving or entering the State’s border.
* **address\_sfx\_primary:** The type of street may have a significant impact upon the probability of injury and homicides. Avenues, for example, are typically longer and wider than a typical Street, meaning that they may be more prone to speeding and/or other traffic violations.
* **address\_pd\_intersecting:** This is listed for the same reasons as the address\_pd\_primary attribute.
* **address\_sfx\_intersecting:** This is listed for the same reasons as the address\_sfx\_primary attribute.
* **violation\_section:** Certain violation sections, such as the ones associated with speeding, running a red light, or DUIs, may have a higher association with injury and death in comparison to less offensive violations.
* **violation\_type:** Certain violation types, such as Vehicle Code (VC), are more likely to have a correlation with traffic-related incidents than, for example, Health and Safety (HS).
* **hit\_run\_lvl:** The hit and run level will naturally have a high degree of correlation with injury and death, considering that misdemeanors are generally defined as hit and runs associated with less severe or indirect incidents, whereas felonies are generally defined as hit and runs associated with high levels of injury and or death.

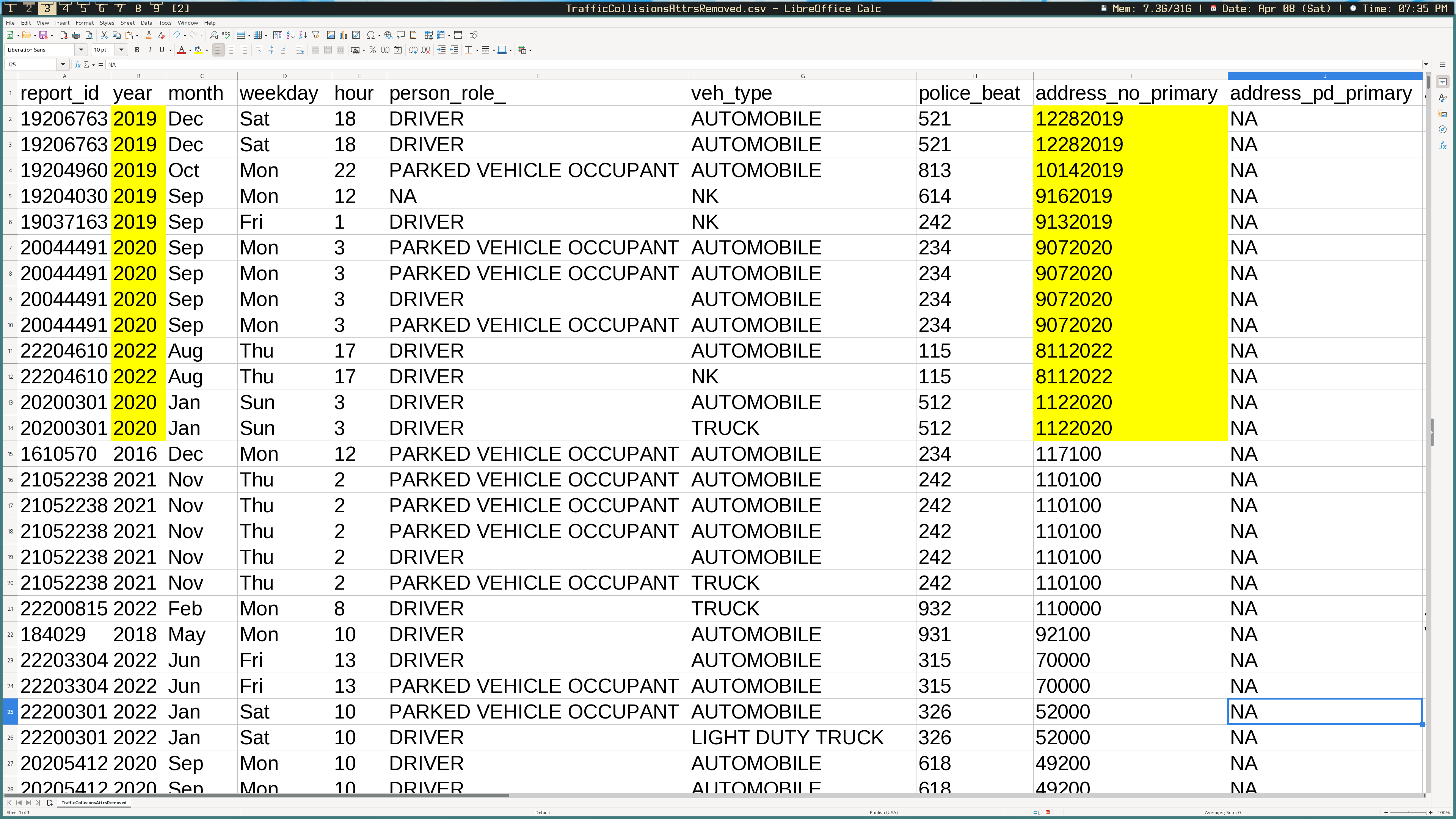
The list of attributes which are not considered as candidates for the remainder of data preparation, and the reasoning for their disqualifications are detailed below:

* **report\_id:** The report ID is used to identify specific records and can be used to sort data within a list. It does not, however, have any significance in relation to injury or homicides.
* **person\_injury\_lvl:** Inclusion of the victims’ injury level would skew the results of the predictive model. If injury is selected as the class value, then an injury level that is not NA would indicate someone was injured. Likewise, since deaths are not counted as injuries, an injury level that is not NA would indicate that no one was killed. To avoid overfitting the data, person\_injury\_lvl is ommitted.
* **person\_veh\_type:** The person vehicle type is a description of the type of vehicle. Although it is a string value, it could technically be converted to a nominal value; however, the type descriptions are too specific to be applicable to a large amount of vehicles. Considering that veh\_type essentially obsoletes this attribute, it is not considered.
* **veh\_make:** The vehicle type can be interpreted from the vehicle make and vehicle model. Considering that there are too many vehicle makes (some of which are very nuianced), the inclusion of this attribute is not necessary.
* **veh\_model:** Vehicle model is ommitted for the same reasons as veh\_make. The reasoning is even more applicable to this attribute, since there are so many different vehicle models that classification by this attribute would essentially have no effect on the predictive model (it would be near random).
* **address\_road\_primary:** Street names are unique strings, and therefore, provide no meaningful information for the predictive model.
* **address\_name\_intersecting:** This attribute is ommitted for the same reasoning as address\_road\_primary.
* **charge\_desc:** The charge description is a string value that provides additional explanatory information about the violation type, but has no association with injury or homicides.
* **injured/killed:** Depending upon which attribute is selected as the class value, the other shall be ommitted to avoid skewing the results of the predictive model.

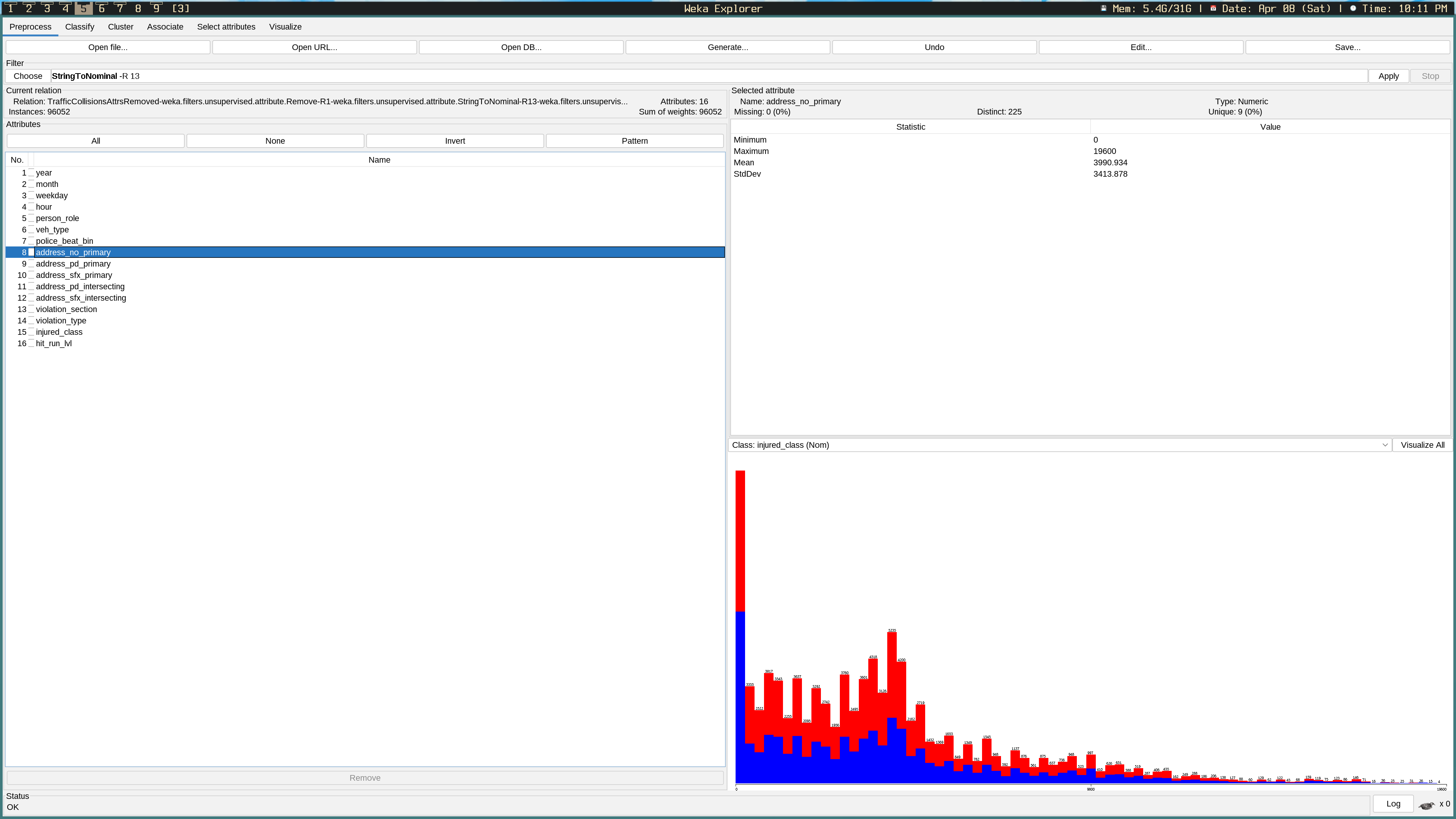
## 3.2 Cleaning Data

Data cleaning is an essential part of the data preparation phase. During data cleaning, we remove or modify instances or attributes that may be incorrect or unnecessary. This section highlights some of the processes that were performed in order to clean the dataset.

When the dataset was imported into Weka for further analysis, it was noticed that the address\_no\_primary attribute had at least one outlier. Returning to Excel, the dataset was sorted by the address\_no\_primary attribute in descending order. Performing this operation revealed that there were more than one outlier for this attribute. After closer inspection, it would appear that the corresponding year had been appended to these specific address numbers for whatever reason.

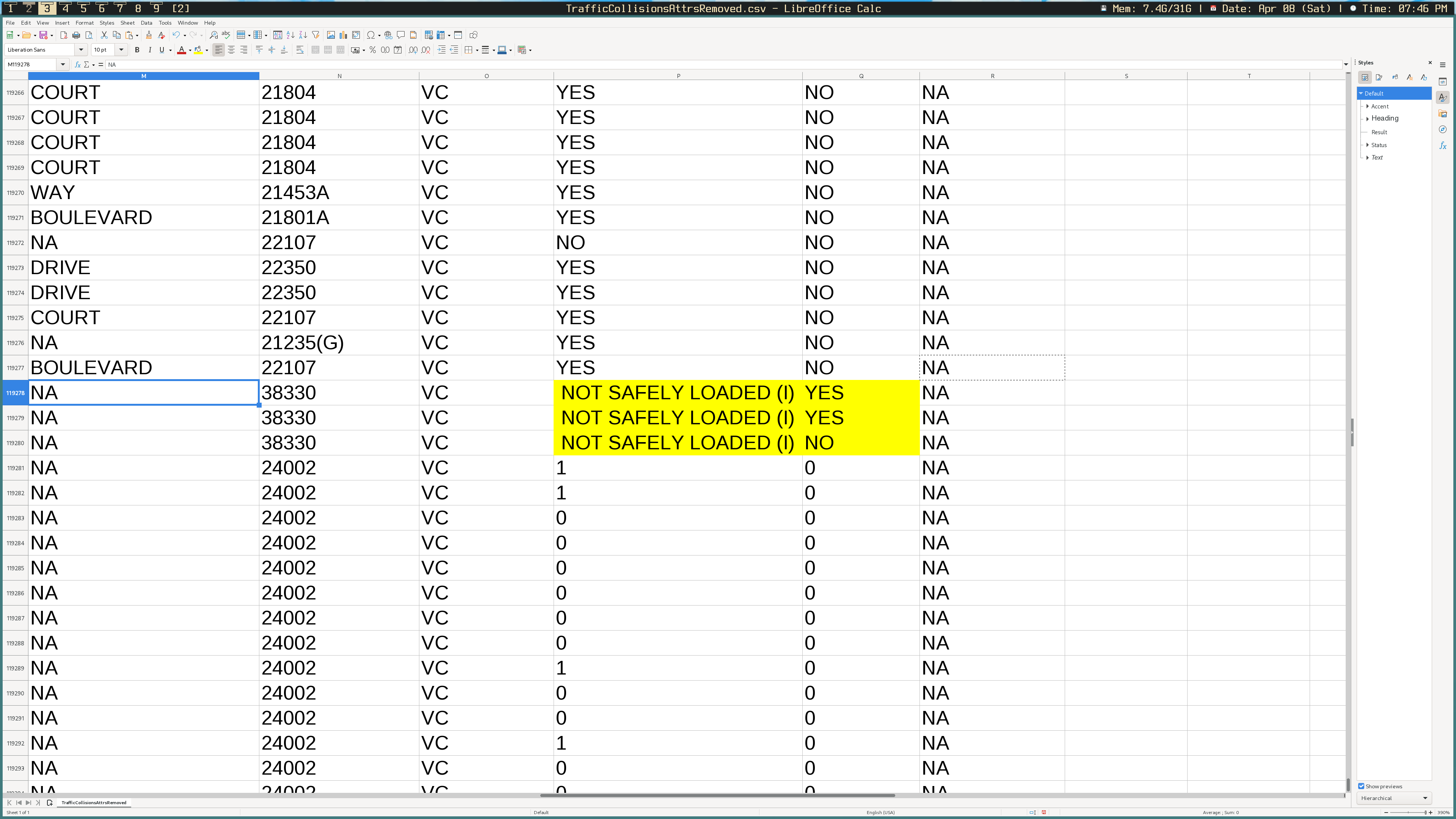
Figure 6: Address Number Outliers

The address numbers for these instances were corrected by removing the year from the end of each address. Returning to Weka, it still appeared as thought there might be outliers for the address\_no\_primary attribute. Upon closer inspection of the data, it was noticed that there was a sudden jump from addresses numbers 19, 600 to 20, 000. This seemed too unnatural to be coincidence. Google Maps was used to enter in the addresses in the range 19, 600 and below. The way that Google Maps functions, is that it will return an vague location if the address number does not exist, but it will return the exact location if the address number does exist. All addresses sampled from 19, 600 and below returned exact locations. All addresses sampled from the range 20, 000 and above returned only the general location. After only a bit of experimentation, it was discovered that all addresses from 20, 000 and above had trailing 0s appended at the ends, as removing these trailing 0s returned exact matches in Google Maps. Thus, these instances were corrected. The resulting range of addresses appeared to look much more reasonable after this fix.

Figure 7: Address Numbers (Weka)

It was noticed that there were 656 instances which had an invalid nominal value of “RANGER” for veh\_type. Each of these instances’ vehicle makes were Ford, and each of their vehicles models were Ranger. In other words, all instances with Ford Rangers had “RANGER” set as the nominal vehicle type. Since the Ford Ranger is a truck, the nominal value was set to “TRUCK” for each of these instances.

A few instances within the injured\_class were marked as “Not Safely Loaded (I)”. The best way of determining the most probable values for these instances is to refer to the killed class. If killed is set to “YES”, then injured shall be “NO”, and vice versa.

Figure 8: Injured Class Anomalies

While analysing the dataset in Weka, it was noticed that address\_sfx\_primary has a single nominal value of “23:00 PM”, which does not make sense for address suffix. The instance is missing values for all other attributes related to the address of the collision, except for address\_road\_primary, which simply contains a value of “3”, and is not a valid street name. Due to the fact that original address could not be found, a value of “NA” was set for address\_sfx\_primary. This same instance also had a value of “MISC\_HAZ” placed for the violation type instead of the violation code. This was corrected so that MISC\_HAZ was set for violation\_code. According to other instances in the dataset, MISC\_HAZ falls under the VC violation type, so a value of VC was placed for violation\_type.

## 3.3 Constructing Data

Considering that the dataset is missing many values for numerous attributes, it is necessary to construct new data to fill in the gaps. Additionally, there are a few useful attributes which can be added to help the predictive model.

The date\_time attribute follows the format “YYYY-MM-DD HH:MM:SS”. In the current format, date\_time is essentially useless, considering that (as mentioned previously) many of the algorithms that will be used to create the model require each attribute to be nominal at the very least. To begin, a new column is created after the date\_time column, titled ‘time’. The date\_time column is renamed to ‘date’. By selecting the entire date column and going to Data > Text to Columns, and then splitting by whitespace, the times will be split from the dates and placed into the time column. Once this step is completed, four new columns are created: year, month, weekday, and hour. It should be ensured that year, month, and hour are formatted as type Number. The date column ought to be formatted as the Date type in the format 1999-12-31. In order to calculate year, the following Excel formula is used: **=TEXT($B2, “yyyy”)**. In order to calculate the month, the following Excel formula is used: **=TEXT($B2, “mmm”)**. In order to calculate the weekday, the column must first be formatted as a custom User-Defined Type, with the format specifier “ddd”. Then, the following Excel fomula is applied: **=TEXT($B2, “ddd”)**. For hour, it must first be ensured that the time column is formatted as type Number. The following Excel formula is then applied: **=LEFT($C2, 2)**. This will simply retrieve the first two digits of the timestamp, providing us with a number between 0 and 23, with 0 representing 12am, and 23 representing 11pm. Once year, month, weekday, and hour have been constructed, the date and hour columns may be removed.

The following attribute that will be examined is the veh\_type column. Ths column is missing many values, and contains many misclassifications of vehicle. The ‘Traffic collisions details dictionary’ provides a pre-defined list of nominal values that can be found within the veh\_type attribute. It was decided that some of these values were unnecessary for the final dataset. The table below displays the old list of nominal values on the left, and the new selection of nominal values on the right.

|  |  |
| --- | --- |
| **Old Nominal Values** | **New Nominal Values** |
| AIRCRAFT |  |
| ALL TERAIN VEHICLE |  |
| AUTOMOBILE | AUTOMOBILE |
| BICYCLE | BICYCLE |
| FARM EQUIPEMENT | FARM EQUIPEMENT |
| LIGHT DUTY TRUCK | LIGHT DUTY TRUCK |
| MOTORCYCLE | MOTORCYCLE |
| TRAILER | TRAILER |
| TRUCK | TRUCK |
| UNLISTED CONSTRUCTION EQUIPEMENT | UNLISTED CONSTRUCTION EQUIPEMENT |
|  | NK |

Certain vehicles within the dataset are classified as aircrafts, but in actual fact, there are no aircrafts within the dataset if the veh\_make and veh\_model attributes are to be trusted. The nominal value “ALL TERAIN VEHICLE” is ambiguous and shows up very infrequently within the dataset. For this reason, if the vehicle’s type cannot be interpolated (e.g., if veh\_make and veh\_model are empty) and the veh\_type is listed as “AIRCRAFT” or “ALL TERAIN VEHICLE”, it is replaced with the new nominal value “NK”, which stands for “Not Known”. The San-Diego Police Department does not appear to have a standard for determining which vehicle makes correspond to which nominal values aside from common sense. For the purposes of this report, a standard was decided upon based upon review of which vehicles were most commonly attributed to their respective types. The standard used for this report is displayed below:

|  |  |
| --- | --- |
| **New Nominal Values** | **Vehicle Categories** |
| AUTOMOBILE | cars, hatchbacks, vans, golf carts, SUVs, etc. |
| BICYCLE | non-electric bicycles |
| FARM EQUIPEMENT | tractors or other farming-related vehicles |
| LIGHT DUTY TRUCK | commercial vans or other large vehicles that do not fit into the other categories |
| MOTORCYCLE | electric bicycles, electric scooters, dirt-bikes, motorcycles, mopeds, etc. |
| TRAILER | mobile homes, RVs, coaches, trailer attachments, boat haulers, etc. |
| TRUCK | pickup trucks, 18-wheelers, tankers, trailers trucks, etc. |
| UNLISTED CONSTRUCTION EQUIPEMENT | dump trucks, cement trucks, bulldozers, front loaders, backhoes, cranes, loaders, etc. |
| NK | Vehicles for which the type is not known and cannot be interpolated |

To make the process easier, the dataset was sorted with veh\_make as the primary sorting key, and veh\_model as the secondary sorting key. Discression must be used when determining the vehicle type. The veh\_make, veh\_model, and person\_veh\_type attributes are all useful indicators. If the vehicle’s make and model are known, a Google Image search is the most reliable method of finding the vehicle’s type. If veh\_model is not known, veh\_make can sometimes provide enough assurance to determine the vehicle’s type. For example, if veh\_model is “Harley-Davidson”, it can be reasoned that veh\_type will be “MOTORCYCLE”. If person\_veh\_type is not blank, it may attest to the vehicle’s make, however, person\_veh\_type is usually unreliable, and should only be depended upon as a last resort. Many instances within the dataset are incorrectly classified as “FARM EQUIPEMENT” and must be corrected. This process is repeated for each instance in the dataset who’s value is missing for veh\_type. Once completed, the person\_veh\_type, veh\_make, and veh\_model columns may be removed from the dataset.

The police\_beat attribute is a continuous numeric type that ranges from values 3 to 999. To avoid outliers, equal frequency binning is applied to the police\_beat attribute. Since the range of possible values is approximately 0 to 1000, it makes logical sense to select a value of 100 for the bin size. A new column called police\_beat\_bin was created, and a long series of IF statements was used to check if the police beat number was between 0 to 99, 100, to 199, 200 to 299, etc. until 999. The result was a column of numeric values ranging from 1 to 10, representing each of the 10 bins. Once this was completed, the police\_beat column could be removed.

A handful of attributes are intentionally missing values where including a nominal value would not be applicable. Due to the fact that our data processing tools cannot accept blank values, these must be replaced with a nominal value such as “NA”, representing a state of being “Not Applicable”. The attributes which require such treatment are as follows: person\_role, address\_pd\_primary, address\_sfx\_primary, address\_sfx\_intersecting, and hit\_run\_lvl. A new column is created beside each of these with matching attribute names, but appended with “\_dup” to differentiate between the old and new attribute columns. The following Excel formula is applied to each of the newly created columns: **=IF(TRIM($<COL>2) = “”, “NA”, $<COL>2)**. The TRIM() function is required due to the fact that the dataset contains many random whitespace characters that cause the equality comparison with the empty string to fail. In essence, the formula simply checks if a reference cell is blank, and if sosets the corresponding cell to “NA”, otherwise, it fills it with the corresponding value. Once this is done, each of the old columns may be removed.

Depending upon whether injuries or deaths is chosen as the class value, new columns must be created to translate the numeric counts to nominal attributes. These columns are named either injury\_class or killed\_class depending upon whether injuries or deaths are being analysed, respectively. In order to convert the existing counts to nominal values, the following Excel formula can be used for both injuries and killed: **=IF(<COL>2 = 0, “NO”, “YES”)**. Once the new class column has been populated, the old attribute column may be removed. If injuries is selected as the class, killed is removed to avoid skewing the results of the predictive model and vice versa if killed is selected as the class.

## 3.4 Integrating Data

Data integration is used when multiple data sources apply to the same business problem. In this report, only one dataset was used, so this step is not applicable.

## 3.5 Formatting Data

Once data cleaning and construction had been completed, the dataset could now be loaded into Weka for formatting. To begin, duplicate instances were removed using the *RemoveDuplicates* filter under Filters > Unsupervised > Instance. Hour was converted from numeric to nominal using the *NumericToNominal* filter under Filters > Unsupervised > Attribute. Likewise, police\_beat\_bin had to receive the same treatment. The violation\_section attribute had to be converted from string to nominal using the *StringToNominal* filter under Filters > Unsupervised > Attribute. Rather than remove report\_id and the opposing class value (either injured or killed) from the CSV file, it was simpler to remove them from Weka and export the file to ARFF format for both the injured class and killed class.

# 4.0 Modeling and Evaluation

In order to create a predictive model that may be able to reveal which attributes are most heavily influencial upon injury and death in San-Diego collisions, the J48 algorithm will be used to create a decision tree that models the most probable outcomes of a collision given the set of attributes that are provided. The benefits of a decision tree are that it is relatively easy to read and interpret, and that it works well with most datasets. This is important in our case, as there are many nominal attributes that would not play nice with other algorithms such as Linear Regression, which requires calculating the Sum of Square Errors, meaning that all attributes would need to be converted into numeric types.

## 4.1 Using a Decision Tree

To begin, the injured class ARFF file is loaded into Weka for classification. The Classify tab is selected, and the classifier algorithm is set to J48, which can be found under Classifiers > trees. There are a few parameters that were considered for updating. A description of each is listed below:

* **confidenceFactor:** The confidence factor dictates the agressiveness of pruning on the dataset. The confidence factor is a number between 0 and 1, with 0 meaning no certainty about the reliability of the dataset, and 1 meaning total certainty in the reliability of the dataset. The default value is 0.25. The higher the confidence factor is, the less the tree will be pruned, which may lead to overfitting. This parameter was left unchanged in order to avoid overfitting the dataset.
* **minNumObj:** The minimum number of objects represents the minimum number of instances that are required to be present for each leaf node in the tree. Typically, increasing the number of minimum objects from the default value (2) will result in overfitting, as it increases bias and decreases variance. For these reasons, this parameter was left untouched.
* **unpruned:** Setting unpruned to true means that the tree will not be pruned. The default value for this parameter is false, meaning that the tree will be pruned by default. Pruning helps to reduce overfitting, and is coupled with the confidence factor mentioned earlier. Once again, to reduce overfitting, this parameter was left unmodified.
* **seed:** The final parameter that was considered was the seed value. The default value for the seed is 1. This parameter determines the exact sequence of psuedo-random numbers that will be generated and used to create the decision tree. This parameter was changed to 10 so that it can be reused for subsequent tests to receive the same results. This is not technically necessary, since the default seed value will always be 1 regardless.

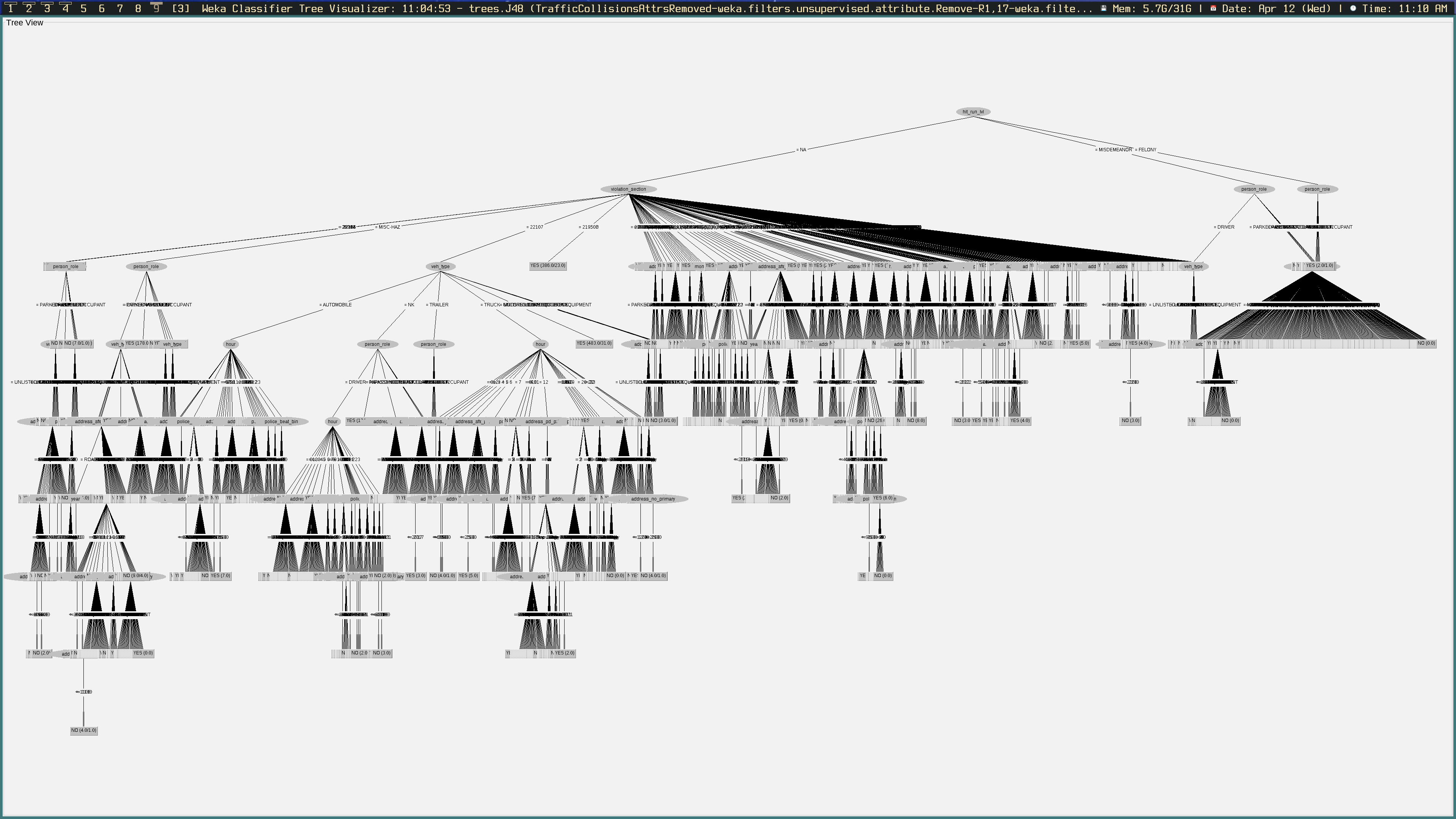
Once the proper parameters were selected, injured\_class was selected as the attribute that classification would be performed upon and k-folds was selected as the test option with a value of 10 for k. It took 6.8 seconds to build the initial model, and then additional time to build the model for each of the 10 folds. The results are shown below:

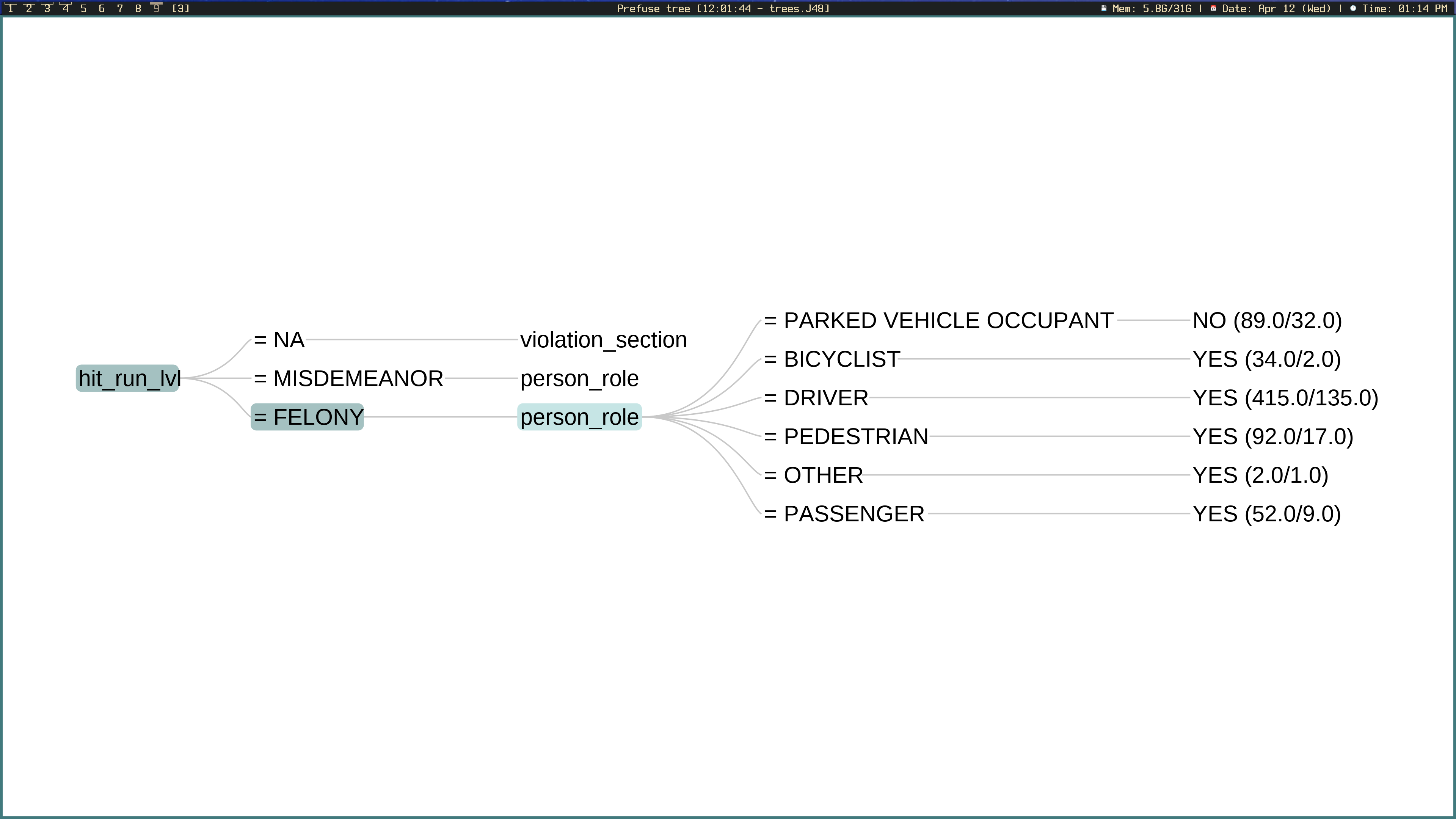
|  |  |  |
| --- | --- | --- |
| Correctly Classified Instances | 79047 | 82.296% |
| Incorrectly Classified Instances | 17005 | 17.704% |

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix** | | |
| Actually Injured | Not Actually Injured |  |
| 35558 | 8482 | Predicted Injured |
| 8523 | 43489 | Predicted Not Injured |

Once the predictive model had been built, the result buffer was right clicked and “Visualize Tree” was selected in order to view a graphical representation of the decision tree.

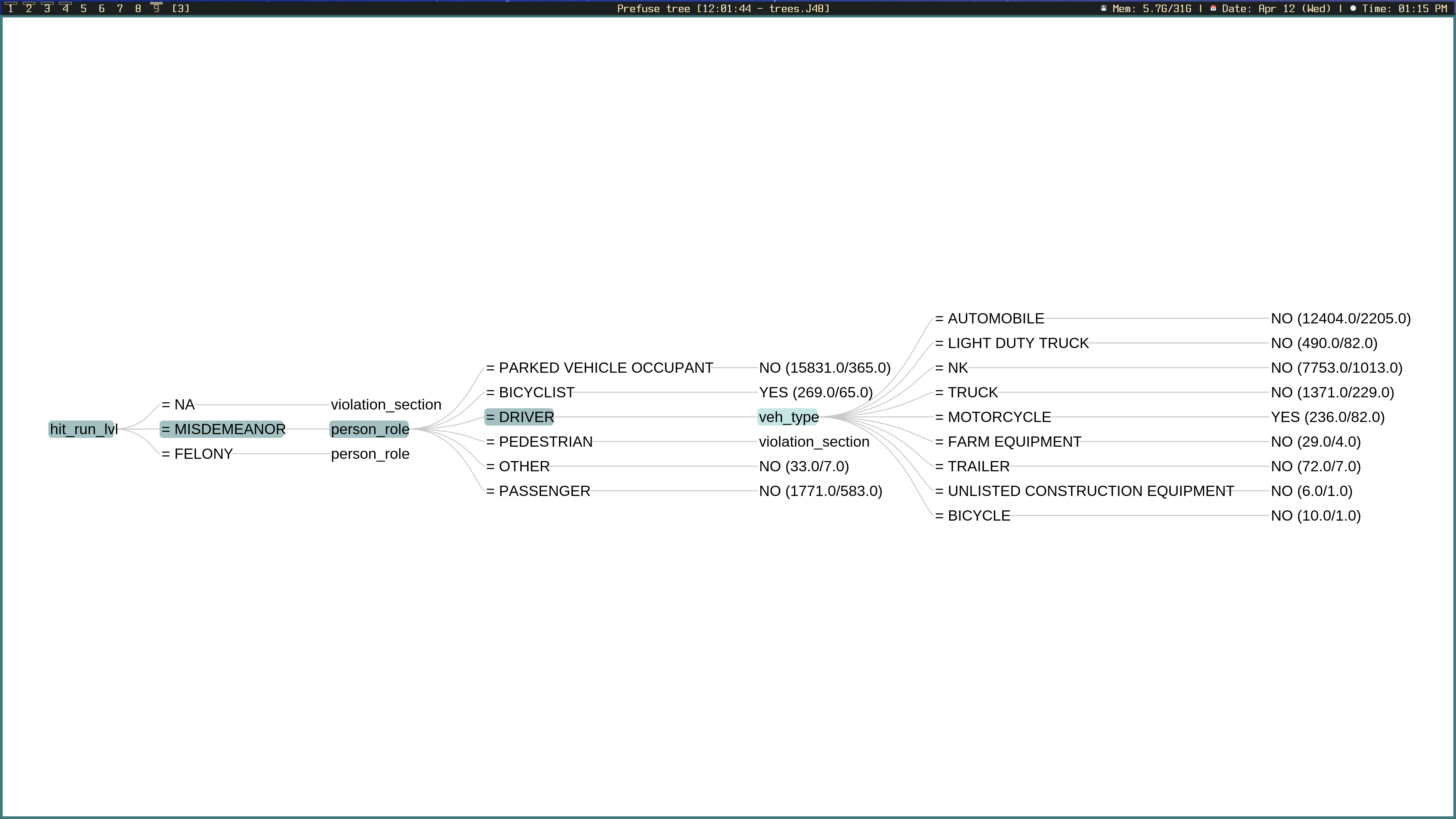
It is immeditately apparent that there are too many nominal values (particularly in the violation\_section attribute) to be analysed properly. For this reason, a plugin was installed using Weka’s built-in package manager called prefuseTree.

Figure 9: Decision Tree in Weka (Injured Class Selected)

Figure 10: Hit and Run Level - Felony Category

The prefuseTree plugin, seen in Figure 10, will display one branch at a time. Transitionatory rules are represented as comparison operators e.g., the equality operator symbol indicating that the attribute must contain that specific nominal value, or the greater than symbol, indicating that the attribute must be greater than some numeric or nominal constant. The root node at each branch in the decision tree is determined by calculating the attribute which produces the most entropy i.e., information gain. Figure 10 indicates that the hit and run level produces the most entropy, and thus, indicates that it has the highest correlation with injury. If the hit and run was a felony, the attribute which produces the second greatest amount of entropy is person role. As can be seen, the predictive model determined that injury was the most probable outcome for drivers, passengers, cyclists, pedestrians, and people classified in the N/A category that were victims of a hit and run marked as a felony. Parked vehicle occupants were much less likely to suffer from injury due to hit and runs marked as felonies. The numbers on the right-hand side of the predicted class value indicate the ratio of actual instances that reached the leaf node to the number of incorrectly classified instances that reached the leaf node.

Figure 11 displays collisions that were the result of hit and runs at the level of misdemeanor. Once again, parked vehicle occupants were predicted to not suffer injury from hit and runs marked as misdemeanors, this time, with a much higher degree of confidence. Pedestrians were also deemed to not be victims of injury, unlike with hit and runs of type felony. Bicyclists, on the contrary, were predicted to suffer from injury. Lastly, it was noted that the only type of driver to suffer from injury from hit and runs marked as misdemeanors were motorcyclists. In the case of both cyclists and motorcyclists, the reasoning for why they were predicted to suffer from injury is most likely due to the fact that both have very little protection in comparison to the other vehicle types, which are all enclosed vehicles.

Figure 11: Hit and Run Level - Misdemeanor Category

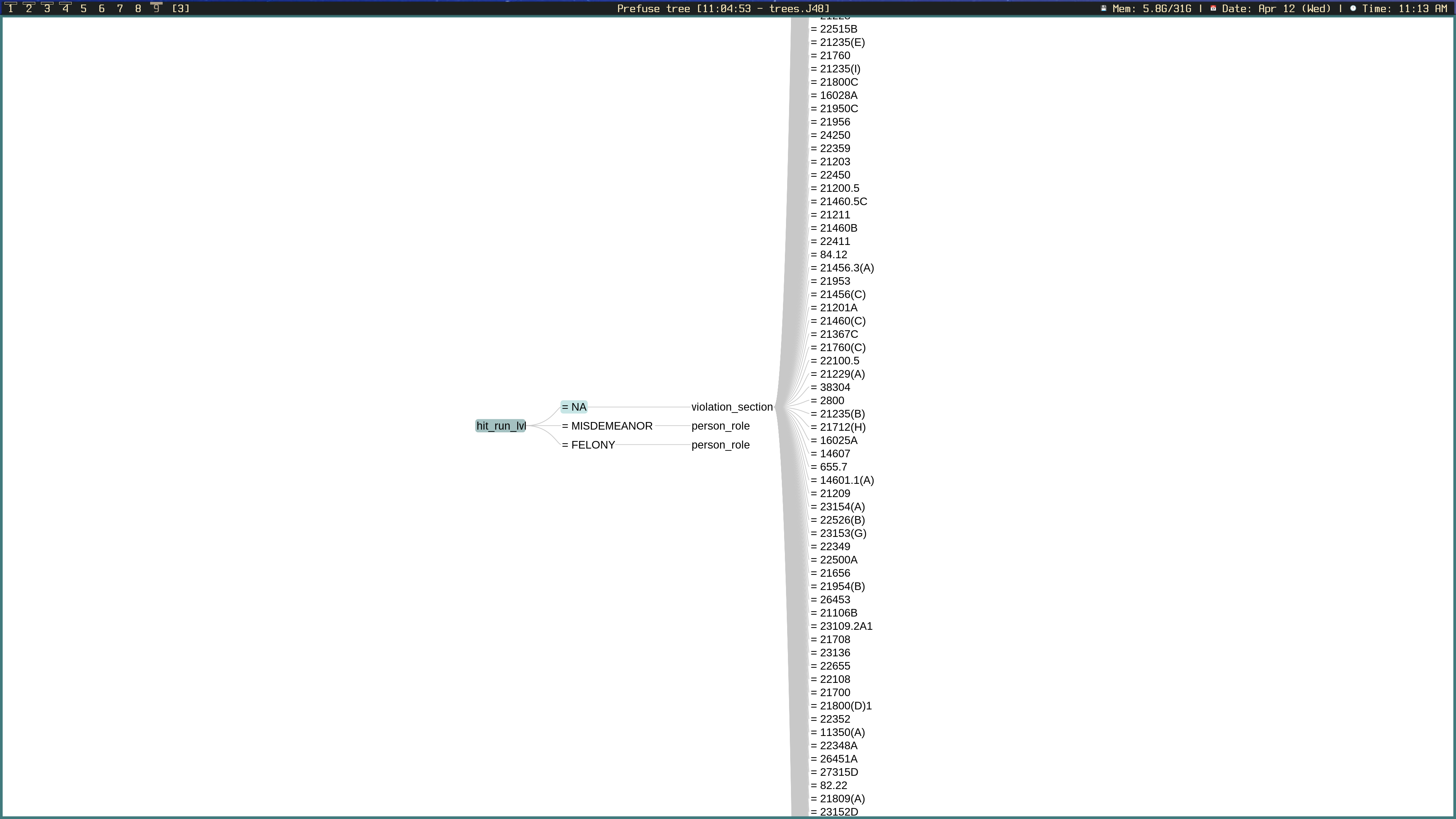
Figure 12: Hit and Run Level - Not Applicable Category

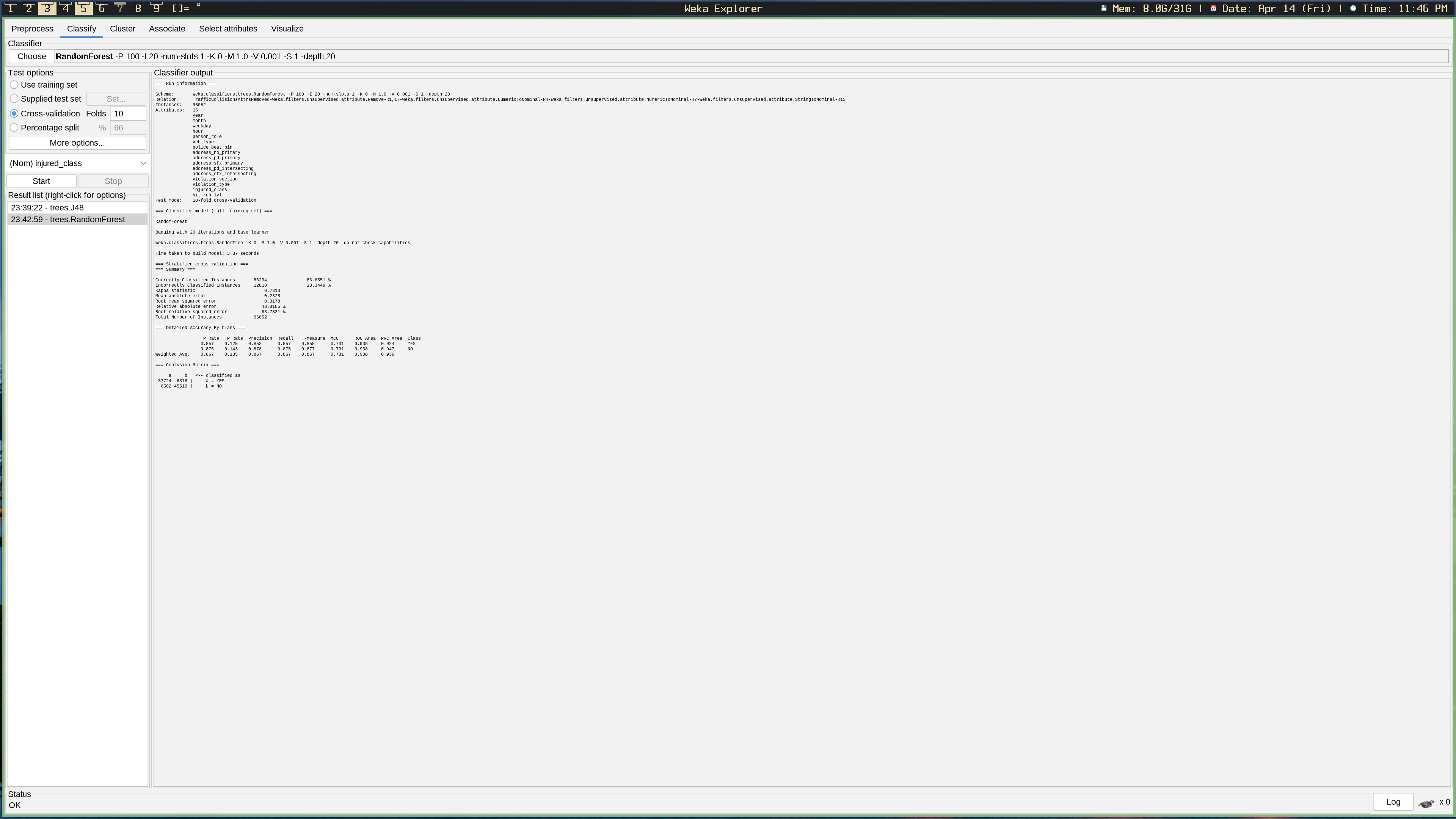
Figure 12 displays the transition rule for collisions which were not marked as hit and runs. The attribute with the second highest degree of entropy in this case was the violation section.

## 4.2 Using a Random Forest

One of the downsides to using decision trees is their tendency to overfit on the training set, so that they appear to have a high degree of predictive power, but proceed to perform poorly when given a test set. Random forests alleviate this issue by using feature bagging, which adds more diversity to the dataset and reduces the correlation among decision trees. A random forest is essential a collection of decision trees that each take a random sample of data. Each individual decision tree performs classification on the given attribute, and then the majority vote decides the final prediction. A description of the important parameters is given below:

* **maxDepth:** The max depth dictates the maximum number of levels for each tree in the random forest. The default value is 0, meaning that the tree will split infinitely if necessary. This was changed to 20 to reduce computational cost.
* **numIterations:** The number of iterations dictates the number of decision trees that will be in the random forest. Essentially the more trees, the better, however there reaches a point of diminishing returns. An ideal number of trees for this dataset might be 100, however, due to computational limitations, this was set to 20.
* **seed:** For the same reasons that were mentioned in section 4.1, the seed value was set to 10 prior to running the classification algorithm.

Injured class was selected as the class that the model would predict for. The algorithm was then ran. The results of the random forest can be seen below:

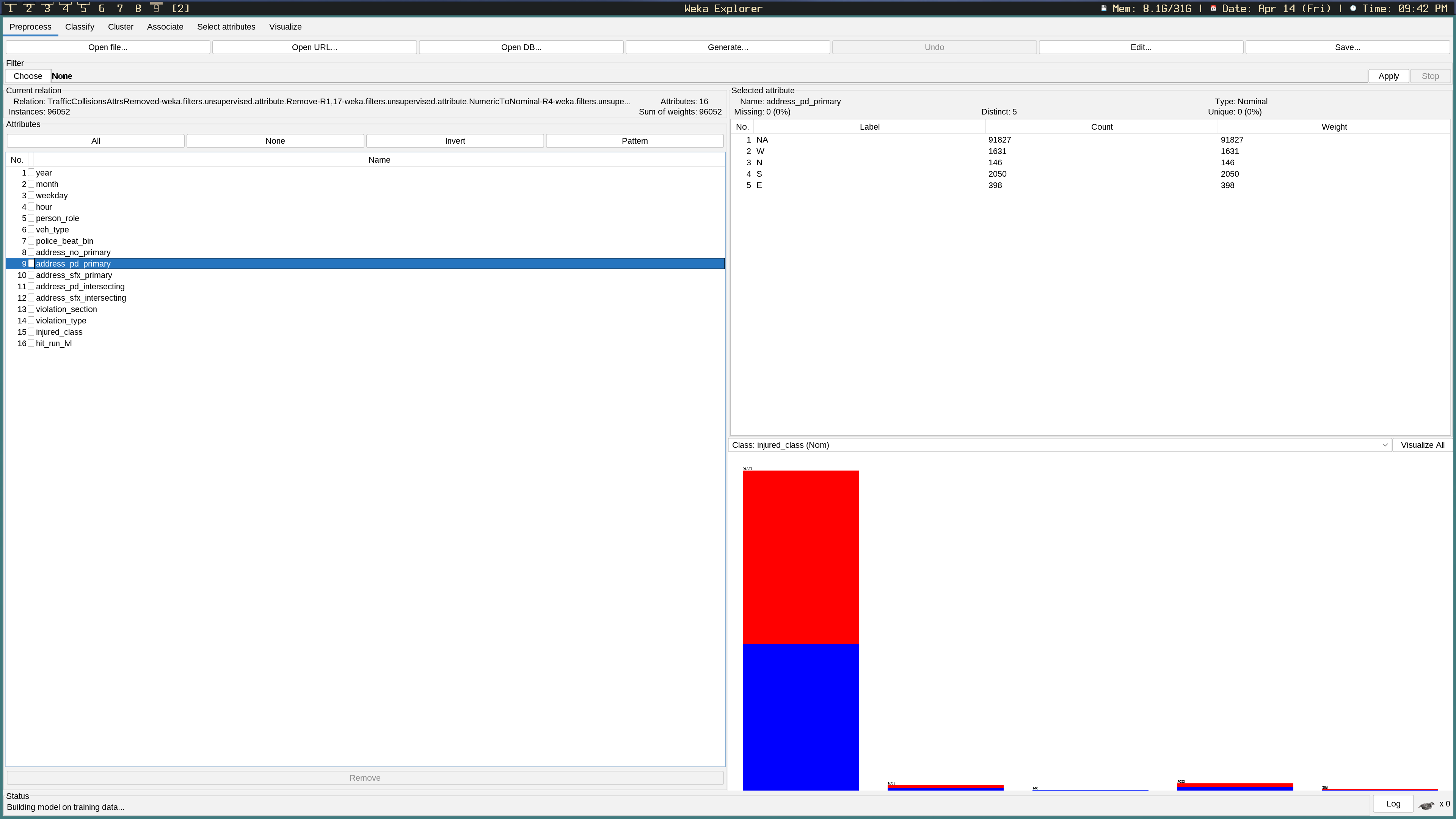
Figure 13: Random Forest Output

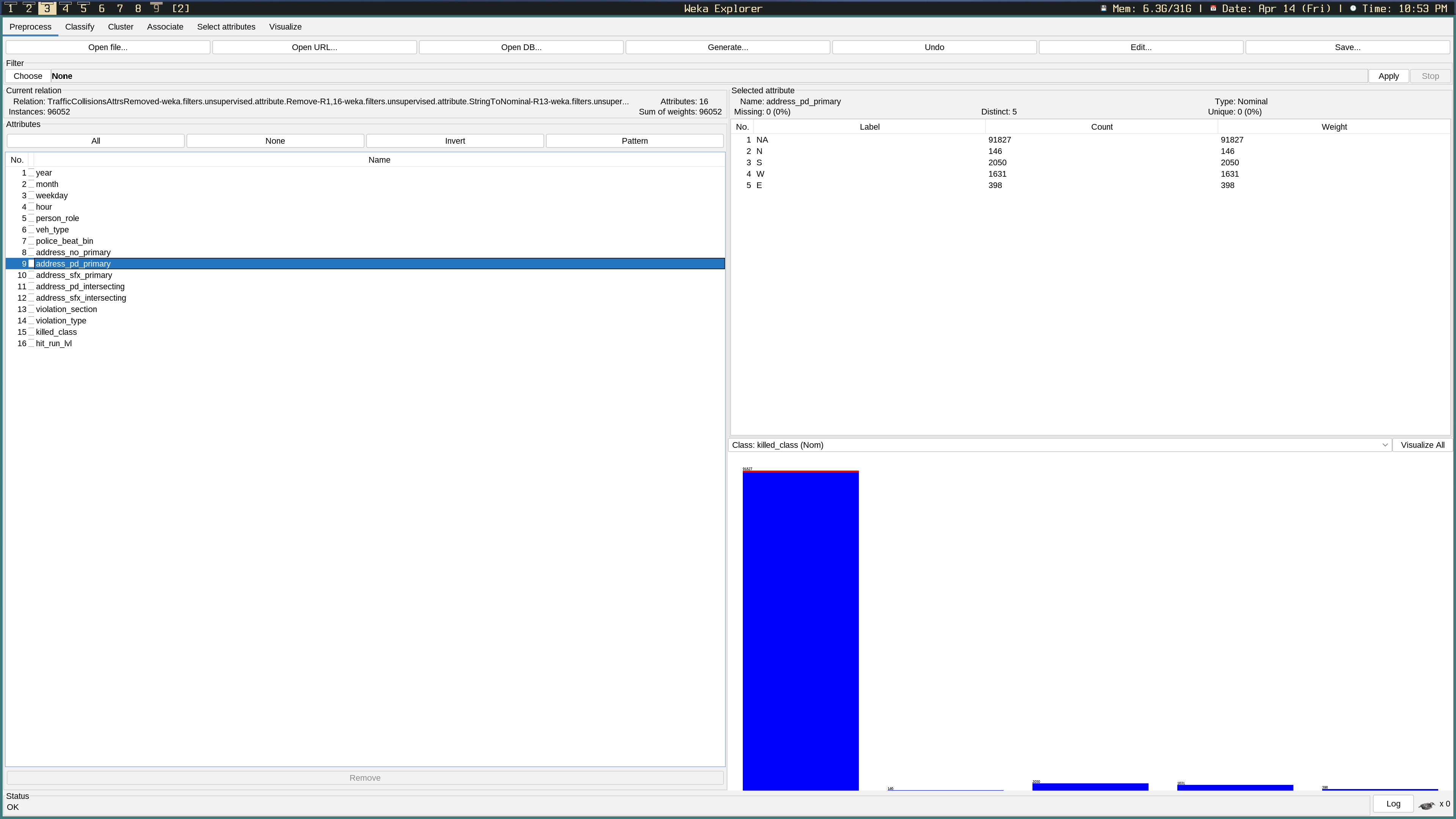
|  |  |
| --- | --- |
| Correctly Classified Instances | 83234 |
| Incorrectly Classified Instances | 12818 |

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix** | | |
| Actually Injured | Not Actually Injured |  |
| 37724 | 6316 | Predicted Injured |
| 6502 | 45510 | Predicted Not Injured |

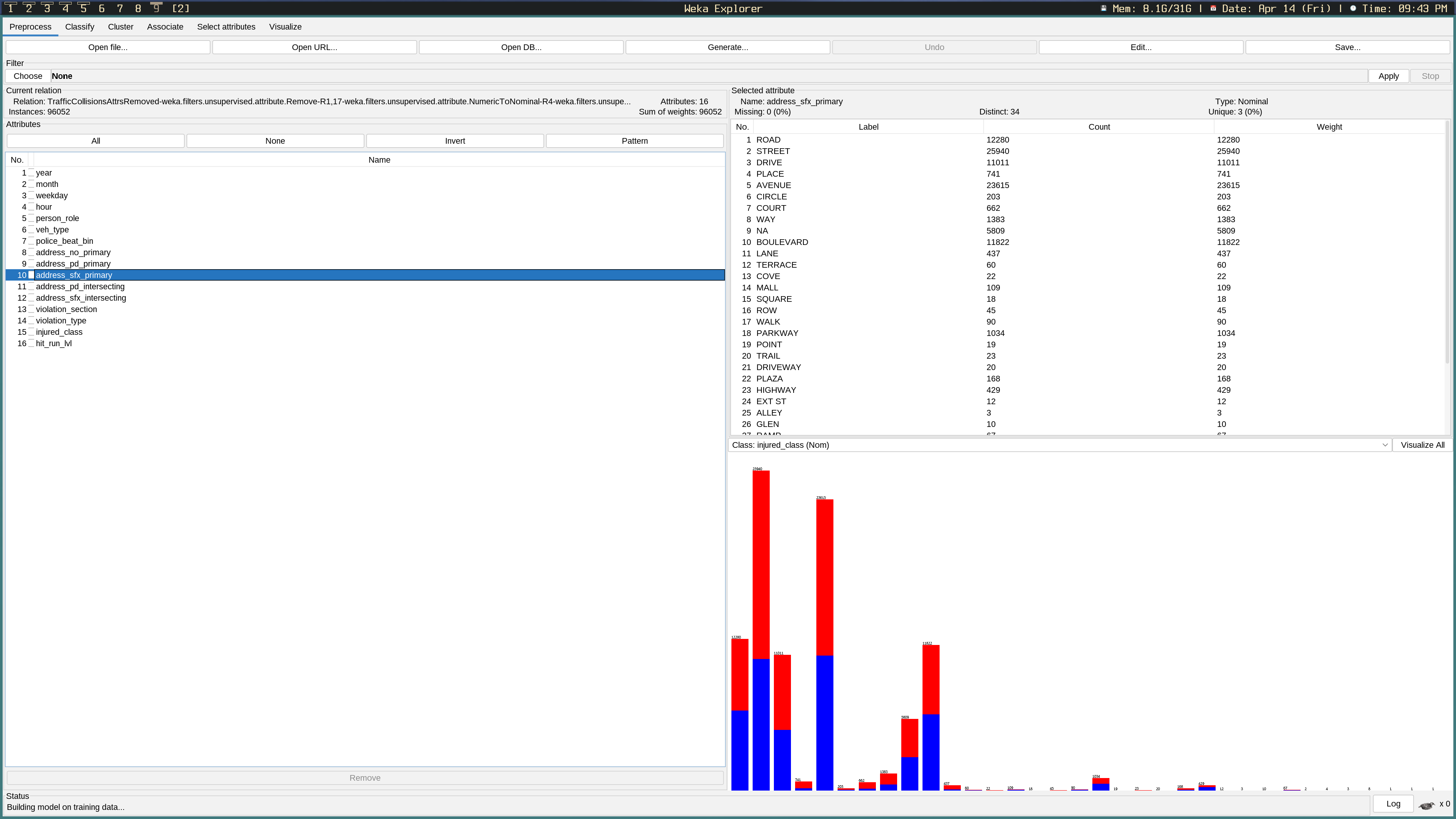
# 5.0 Discussion of Results

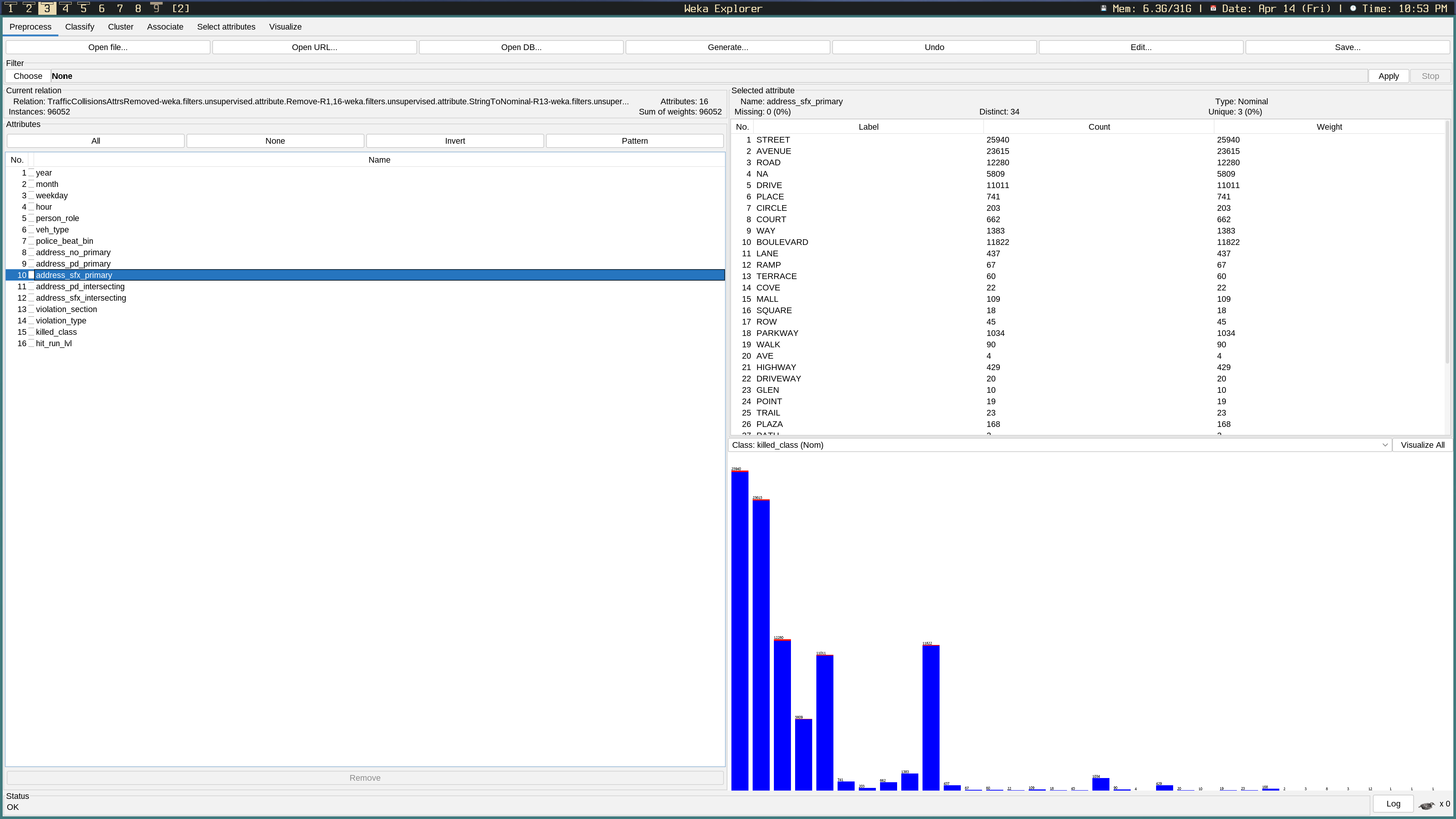
# 

Figure 14: Primary Address Direction (Injured)

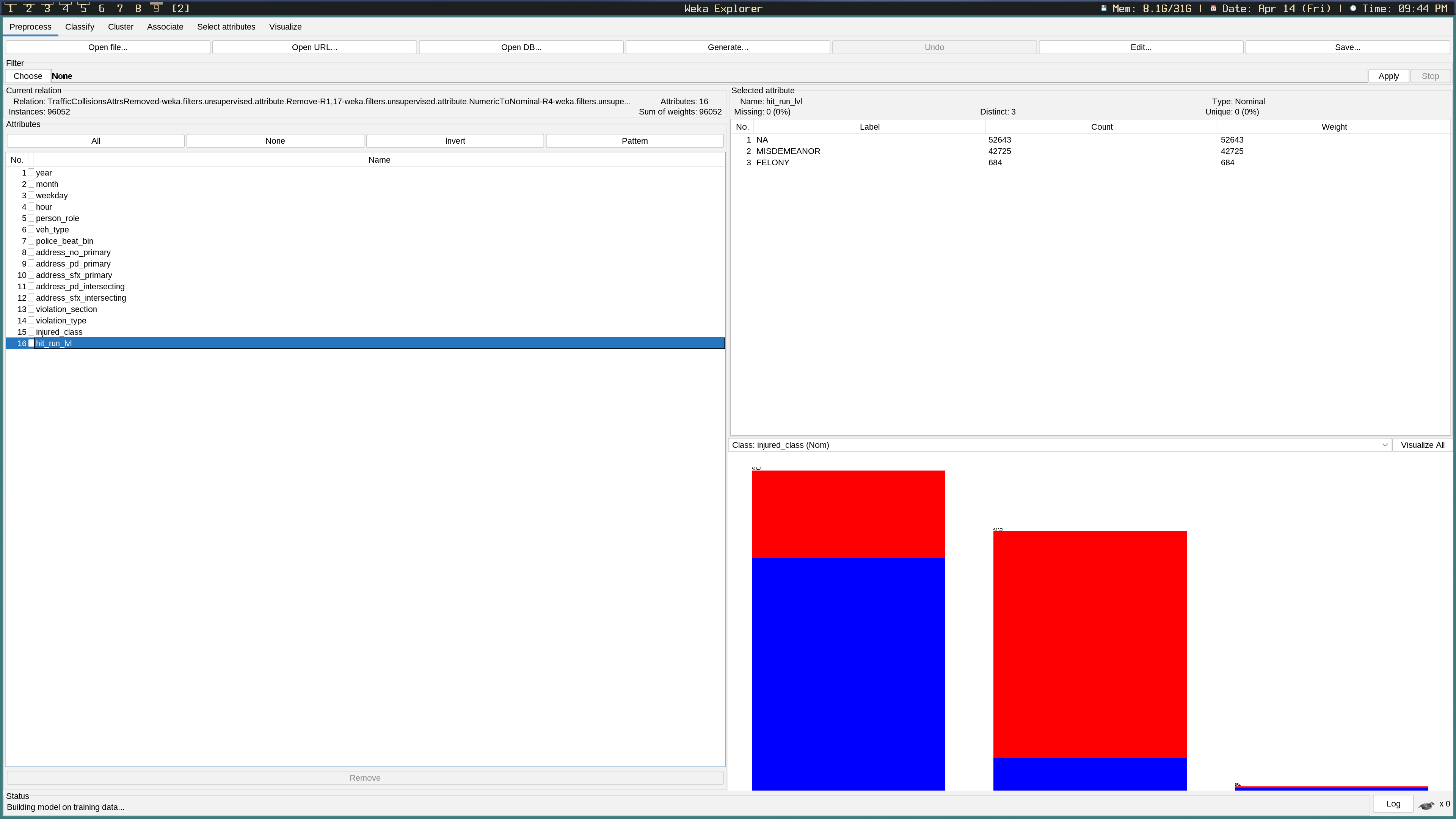
Figure 15: Primary Address Direction (Killed)

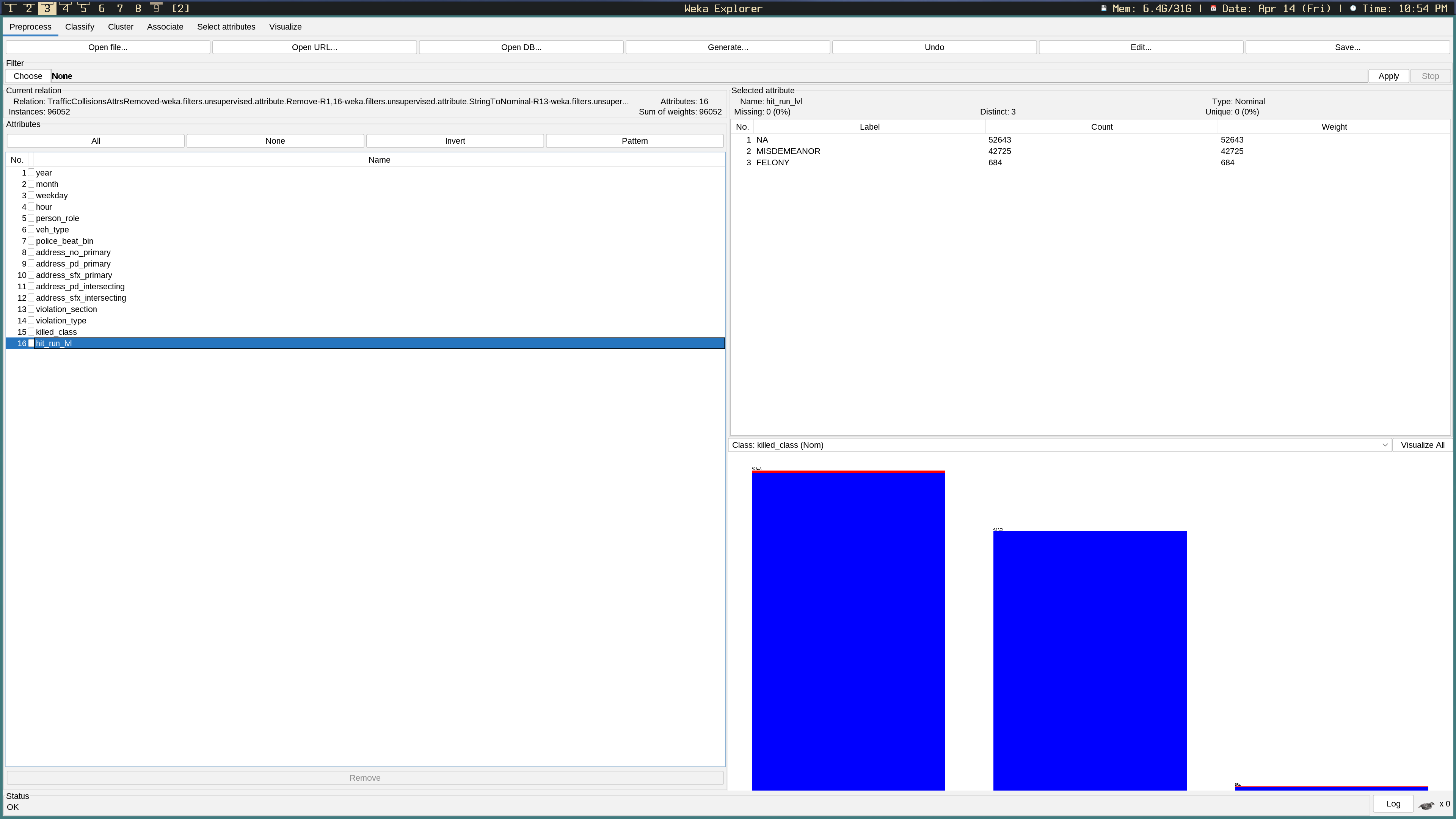
Figures 13 and 14 display the number of collisions, injuries, and homicides for each street in the dataset that contained a direction. Confusingly, blue represents the number of injuries in Figure 13, but the number of survivors in Figure 14. There are practically next to no deaths recorded in Figure 14, as there simply were not enough instances to find any associations with street direction. Interestingly though, we can see more injuries occurring towards the South-West.Figures 15 and 16 display the amount of injuries (marked as blue in Figure 15) and the number of deaths (marked in red in Figure 16) per suffix associated with the address name at which the collision occurred. Although perhaps difficult to tell visually, both injuries and homicides are most common amongst streets, then avenues, and then roads.

Figure 16: Primary Address Suffix (Injuries)

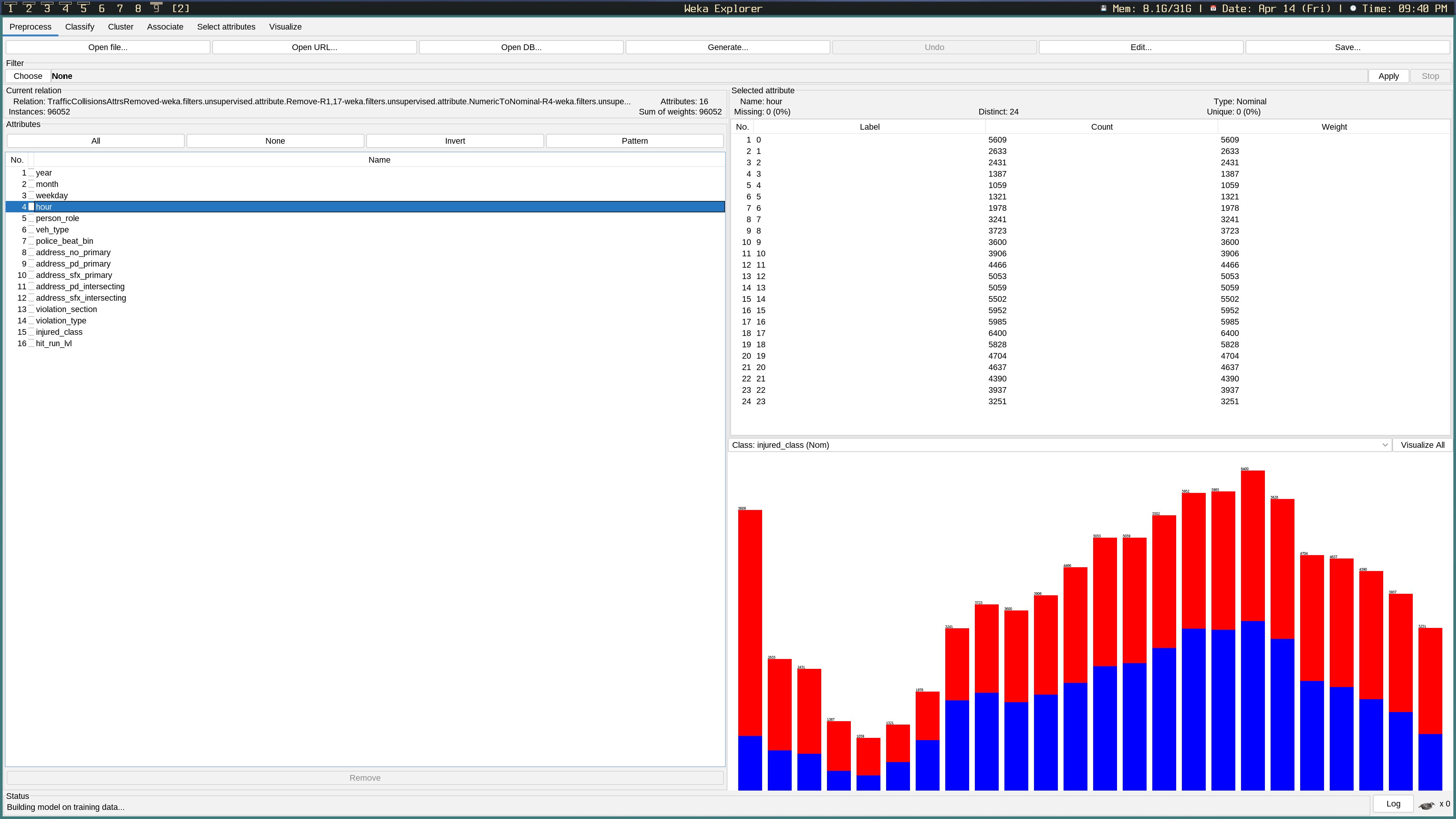
Figure 17: Primary Address Suffix (Killed)

Here we get a better look at the level of injuries vs homicides in regards to the hit and run level. In both Figures 17 and 18, hit and runs marked as felonies are represented on the right-most bar. Misdemeanors are represented by the middle bar, and non-hit and runs are represented on the left-most bar. In terms of injuries, there are significantly more for non-hit and runs, which is somewhat unexpected, but I suppose that what we can take away from this is that many victims of vehicle collisions are responsible enough to wait for a police officer to arrive at the scene. For hit and runs marked as felonies, the injury rate is approximately 50%. Similarly for homicides, the vast majority occur within collisions which are not marked as hit and runs. There are virtually no deaths for hit and runs marked as misdemeanors, which makes sense, as a misdemeanor is likely something akin to hitting someones bumper in a parking lot. There are a few homicides within the hit and runs marked as felonies, although not to the same degree as the injury rate for the same category.

Figure 18: Hit and Run Level (Injuries)

Figure 19: Hit and Run Level (Killed)

Taking a look at Figure 19, we can note a couple of things. First is the sudden spike in collisions around 12am. Although there are more collisions at this time, it also appears to have the lowest ratio of injuries to collisions out of any hour. Notice that there is an ever so slight peak in injuries at 8am, during rush hour. The general trend however, is that injuries increase towards 5pm before falling back down during the evening. There also appears to be a bit of a sudden dip after 6pm when, presumably, rush hour ends.

Figure 20: Injuries Per Hour

# 6.0 Conclusion

To conclude, predictions towards the beginning of the report will be compared and contrasted to the actual outcomes which were determined during the data exploration phase and confirmed as being accurate by the predictive model. The first prediction that was made was the the date time would affect collision rates. Specifically, I mentioned that Friday, Saturday, and Sunday were busier, and thus, were more likely to result in collisions. Coinciding with the weekday, I also predicted that rush hour (approximately 7am to 10am and 4pm to 6pm) would, in the same vain, result in a higher collision rate. This prediction appears to have been mostly accurate, as the findings discussed in section 5.0 demonstrated that Friday, Saturday, and Wednesday were the weekdays with the most amount of collisions, with Sunday being very close in number to Wednesday. Likewise, the data indicates that, although there does appear to be a spike in the number of collisions at around 8am, the general trend is that collisions become more frequent as the hour approaches 5pm, where it then peaks, and falls back down. Additionally, there was an unexpected spike in collisions at around 12am. It was predicted that bicyclists, motorcyclists, and pedestrians would have a higher rate of injury and/or fatalities. Motorcycles and bicycles did appear to have a ratio of injuries and fatalities that was much higher in comparison to other vehicle types, however, the data does not appear to support the idea that pedestrians are more likely to suffer from injury. It was predicted that the violation section and type would both have an effect on the severity of the collision. There were so few instances of reports that were not placed under the Vehicle Code violation type, that they can likely be considered as outliers. On the contrary, the violation section, as discussed in section 2.0, had specific codes which were more heavily associated with injury and homicides. The Police beat was suspected to have a potential correlation with collision rates due to certain beats being home to busier streets than others. This prediction appears to have been accurate, as seen by the large spikes in collisions within certain beats e.g., beat 122, for example. It was predicted that certain address suffixes would be correlated with a higher degree of injuries and/or homicides. As discussed in section 5.0, streets, followed by avenues, and then by roads, appear to be the top 3 producers of both injury and homicides. Due to the fact that the state borders were towards the North-East, it was predicted that a higher rate of both injury and fatalities would occur towards those directions. As was previously discussed, the polar opposite was true, with more injuries and deaths occuring towards the South-West, towards the heart of San-Diego. In retrospect, this would seem to make more sense than my initial prediction. Finally, it was predicted that hit and runs marked as felonies would have a higher rate of injury and homicides in comparison to misdemeanors. This was, of course, correct.

The San-Diego dataset was a fascinating learning opportunity to visualize and understand which factors are most prevalent towards both injury and fatalities in terms of vehicle collisions. I believe that a lot can be gleaned from the findings demonstrated within this report. It would be beneficial for the Municipal authorities of San-Diego to consider some of the most dangerous factors and attempt to find ways to reduce or eliminate them. For example, more speed cameras or traffic lights may be required to reduce speeding on busy roads. Additionally, we’ve learned that long stretches of road are often-times the most dangerous, thus, finding a way to enforce San-Diegos driving laws even more would be beneficial for the safety of the public.

# 7.0 References

Chang, Edward. “Traffic Collisions in San Diego.” Kaggle.com, 2020, www.kaggle.com/code/edchang1992/traffic-collisions-in-san-diego/notebook. Accessed 12 Apr. 2023.

“Cleaning Data.” Www.ibm.com, 2015, www.ibm.com/docs/en/spss-modeler/SaaS?topic=preparation-cleaning-data.

Ellis, Christina. “Number of Trees in Random Forests.” Crunching the Data, 27 Aug. 2022, crunchingthedata.com/number-of-trees-in-random-forests/.

Open, Diego. “City of San Diego Open Data Portal.” City of San Diego Open Data Portal, 2018, data.sandiego.gov/datasets/police-collisions-details/.

“San Diego Traffic Collision Analysis.” Datathon2019, nhtsai.github.io/datathon2019/. Accessed 12 Apr. 2023.

webcontroller. “Is a Hit & Run a Felony? - Potential Charges & Jail Time.” Valiente Mott, 20 Mar. 2020, valientemott.com/legal-terminology/is-a-hit-and-run-a-felony/. Accessed 12 Apr. 2023.

Yıldırım, Soner. “Decision Tree and Random Forest — Explained.” Medium, 11 Feb. 2020, towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd.