Final Project: Accident Severity Prediction

6 ottobre 2020

Indice

1 Business Understanding

1.1 Background

The amount of accidents in a city is a matter that interest many sides as the infrastructure engineering, police and hospital assistance to the citizens. Such entities collaborate together handling the problems related to the event of incidents throughout the city.

1.2 Problem

Dealing with all the incidents requires for a consistent investment of money, resources and time to try to guarantee the most responsive and efficient service for the citizens. Without a clear understanding of the most crucial conditions that influence the development the accident it is impossible to know how to invest the right attentions risking to lose the focus and the overall efficiency.

1.3 Interest

This report collects the data that describe the conditions present at the moment of the incident and the various specifics of the latter in order to develop a model that could learn from such data. This would help into understanding better what are the most crucial factor in the developing of the accidents as well as the most critical zones where the most critical incidents take place. A better understanding of these factors will be used by the interested entities to understand what are the most important problems which requires the most urgent attention and also to asset where to increase the spending of resources and where to reduce the efforts where is not necessary.

2 Data Understanding

This phase consists of different steps of work on the data:

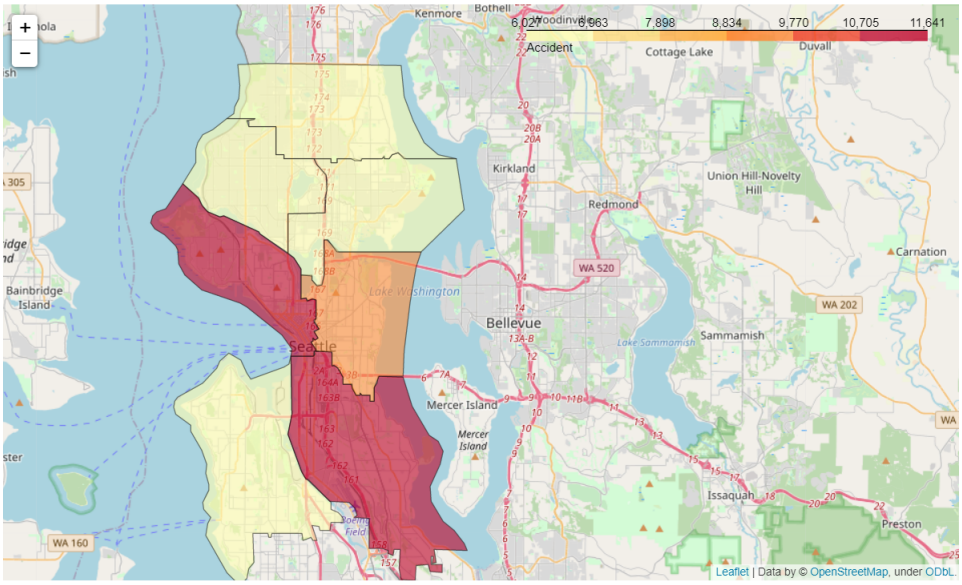
1. Selecting the column that tracks the description of the accident's severity 'SEVERITYCODE'
2. Select the most important features that may influence the final result:
    - Road, Light and Weather Conditions ('ROADCOND', 'LIGHTCOND', 'WEATHER')
    - The city district in which the accident happened, obtained from the coordinates columns ('X','Y')
    - The kind of spot in which the accident took place as 'Interception', 'Block', 'Alley' ('ADDRTYPE')
    - The number of people, pedons and cyclists involved ('PEDCOUNT', 'PERSONCOUNT', 'PEDCYLCOUNT')

- The day of the week in which the accident took place ('INCDATE')

3. Adjust the data format changing from text to numerical values
4. Take care of the NaN values spread around the dataframe
5. Remove the duplicates columns and the ones that don't contribute in any way to the model development
6. Check for columns with strong correlation between each other so to clear and simplify the dataset

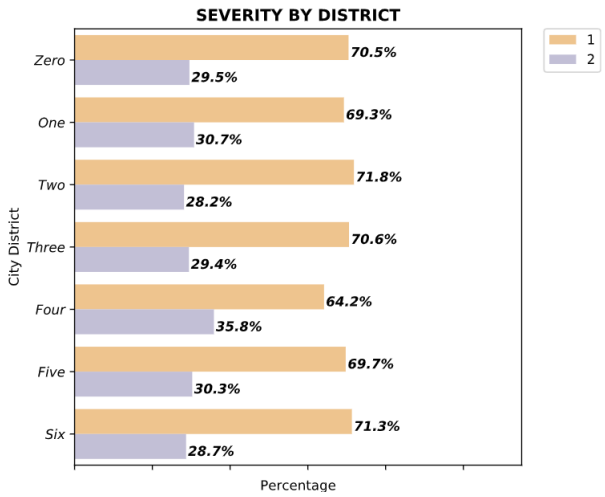
2.1 District Analysis

In this part the coordinates of the incidents are extracted from the dataset. Each incident with its position is classified and associated with correspondent city's district



The map shows how in the city are present districts with a significative higher number of accidents with respect to the others.

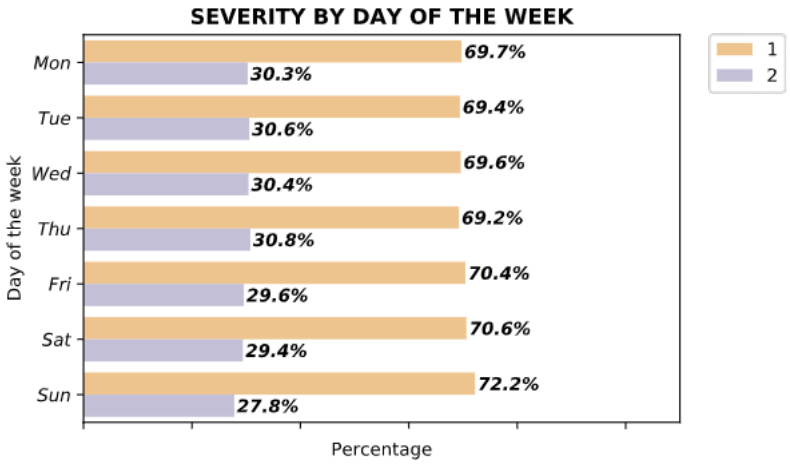
An analysis of the distribution of the presence of critic/non critic accidents will show if some of the districts present a particular behaviour or not.



The graph shows that each district presents the same distribution and so the districts present the same problems in the handling of the traffic and the difference is only in the quantity of cases.

2.2 Day of the Week

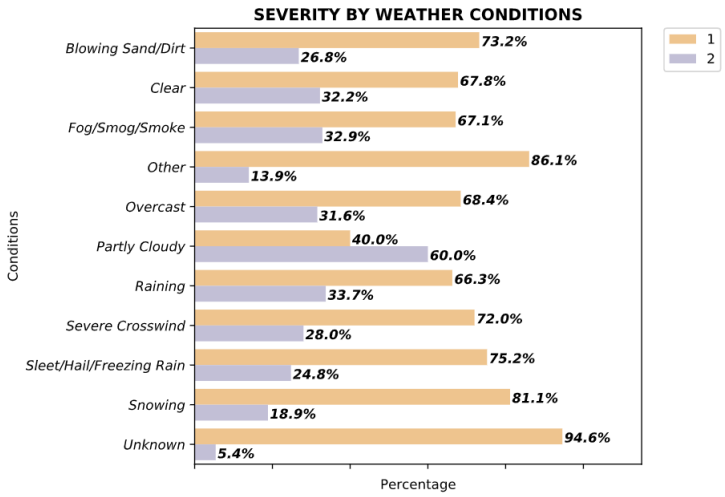
I extract the column of the date of the incident and transform it into a set of values for each value of the week. This allows to check if there's a particular day of the week in which more accidents are due to happen or if more critic incidents happen in particular days.



It can be seen as the day of the week doesn't imply any particular variation in the statistics of the incidents and that the evenience of critics/non critic accidents remains constant throughout the week.

2.3 Weather Conditions

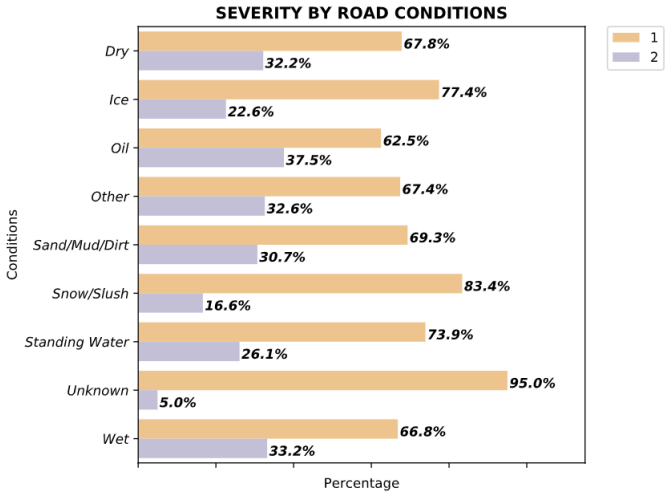
The next graph shows the percentage distribution of the kind of accidents given the different weather conditions.



Given the size of the various conditions and the resemblance in percentage it is possible to gather more weather conditions together under the group 'Other' to reduce the dataset size. Taking care of the nan values assimilating them under the 'Unknown' group.

2.4 Road Conditions

In this part the same approach used for the weather conditions is applied to the road ones

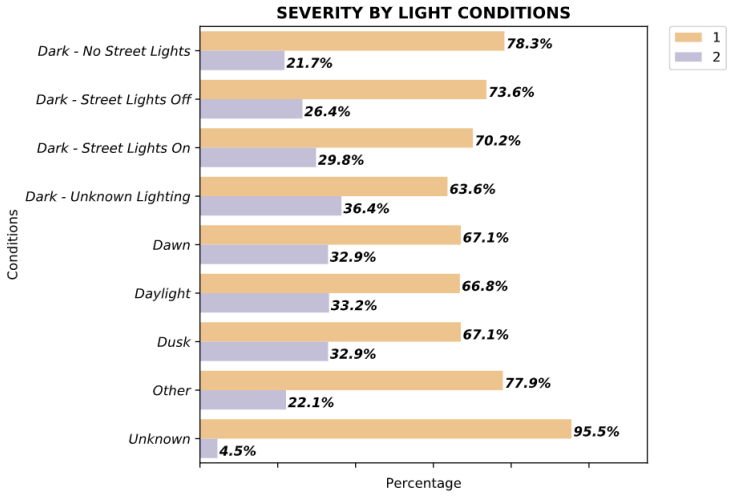


It is possible to categorize the conditions into groups that comprehend two or more that share the same percentages.

1. The NaN values are classified as 'Unknown'
2. The 'Dry' and 'Wet' conditions presenting the same distribution are grouped together
3. The 'Ice' and 'Snow/Slush' conditions too can be assembled
4. The remaining small groups are collected as part of the 'Other' group

2.5 Light Conditions

In the same way the sequent graph shows the influence of the light conditions onto the severity of the accident.

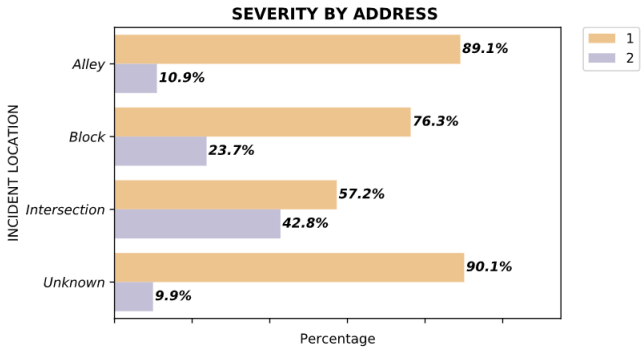


It is possible to categorize the conditions into groups that comprehend two or more that share the same percentages.

1. The NaN values are classified as 'Unknown'
2. The Dark related columns Values are classified as 'Dark'
3. The Light related columns Values are classified as 'Light'

2.6 Address Location

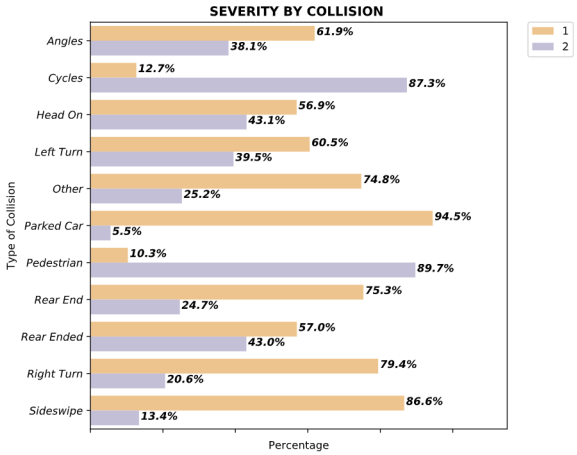
The graph shows the influence of the incident "address" onto the severity of the latter.



By the graph is clear as 'Unknown' and 'Alley' values share same percentages together with having the same low number of cases and thus they can be collected as one

2.7 Collision Type

The type of collision influence onto the severity of the incident is showed in the next graph.

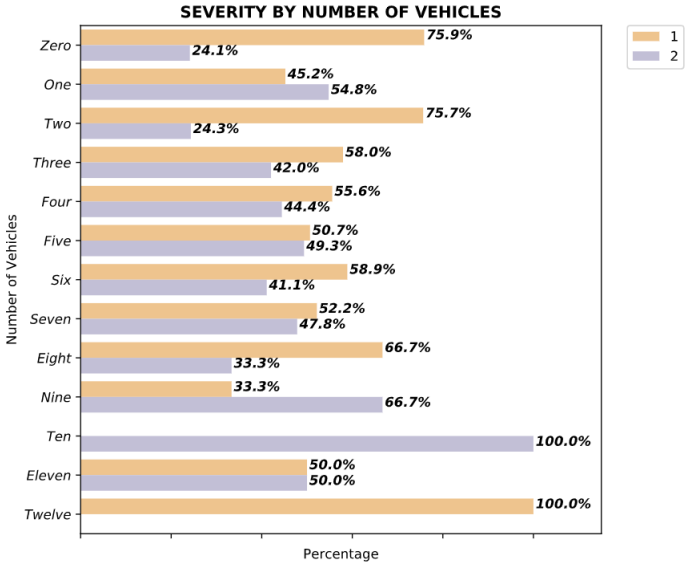


Given graph and value counts for the 'COLLISIONTYPE' column it is possible to group some of the characteristics that share same percentages

1. 'Pedestrian' and 'Cycles' incidents are regrouped as 'Ped/Cyc'
2. 'Angles' and 'Left turn' can be regrouped as the 'Angle/Left' group
3. 'Head On' and 'Rear Ended' become 'Front/Rear'

2.8 Vehicles Involved

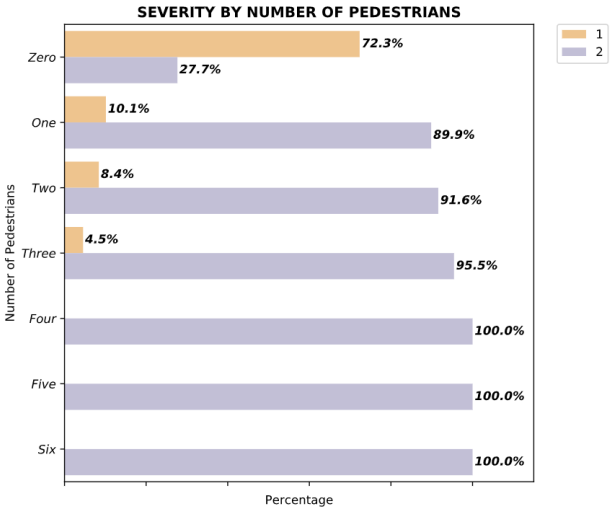
The severity of the incident in function of the number of vehicles involved is showed in the next graph.



From the percentage graph and the info from the dataset it is possible to group all accidents that involve more than 3 cars under the same category. It has to be noted that even if higher number of vehicles included in the accident don't show a percentage similar to the others the low amount of cases makes it difficult to build a reliable statistic. Given these conditions it is safe to assume that accidents that involved more than three cars can be considered critic.

2.9 Pedestrians Involved

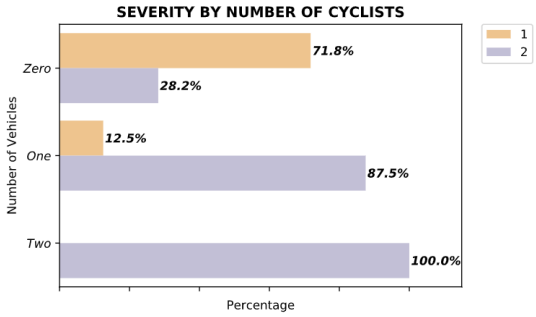
The severity of the incident in function of the number of pedestrians involved is showed in the next graph.



The graph shows clearly as whenever a pedestrian is involved there's a high chance for the incident to be critic. The 'PEDCOUNT' column can be then considered as a boolean. The column will display one if at least one pedestrian is involved and zero otherwise.

2.10 Cyclists Involved

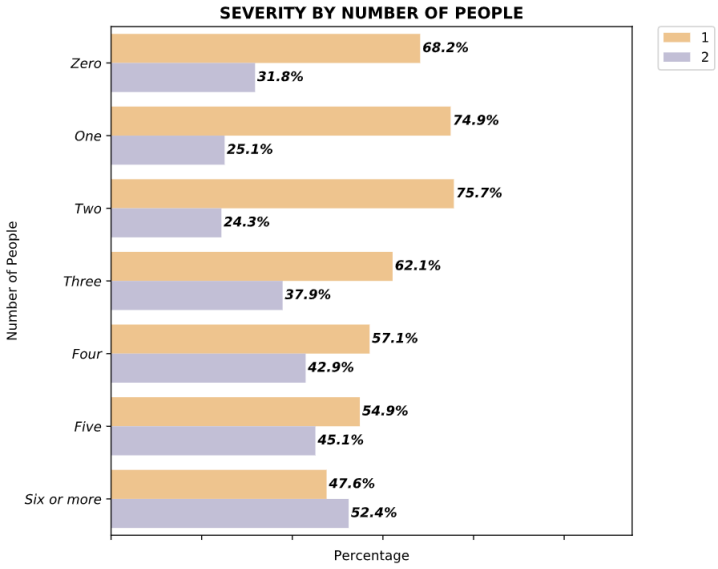
The severity of the incident in function of the number of cyclists involved is showed in the next graph.



The graph shows clearly as whenever a cyclist is involved there's a high chance for the incident to be critic. The 'PEDCYLCOUNT' column can be then considered as a boolean. The column will display one if at least one cyclist is involved and zero otherwise.

2.11 People Involved

The severity of the incident in function of the number of persons involved is showed in the next graph.

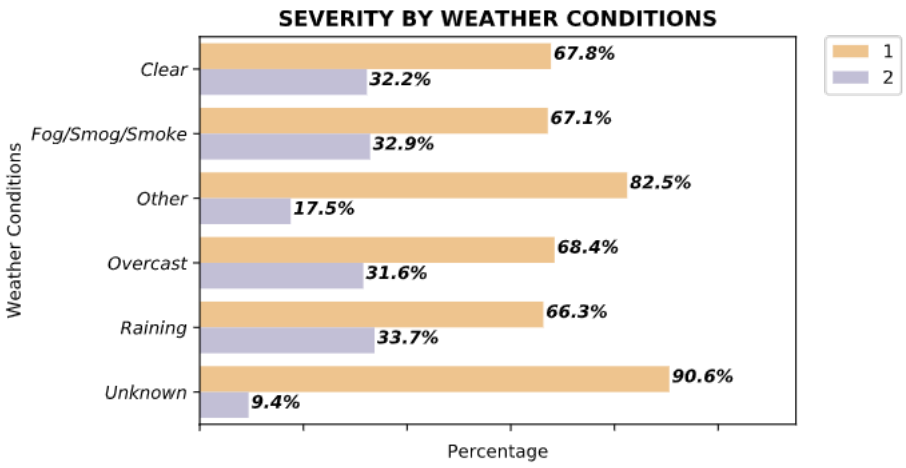


The graphs shows that there are groups that share similar percentages and thus can be grouped together as the groups 1,2 and 4,5.

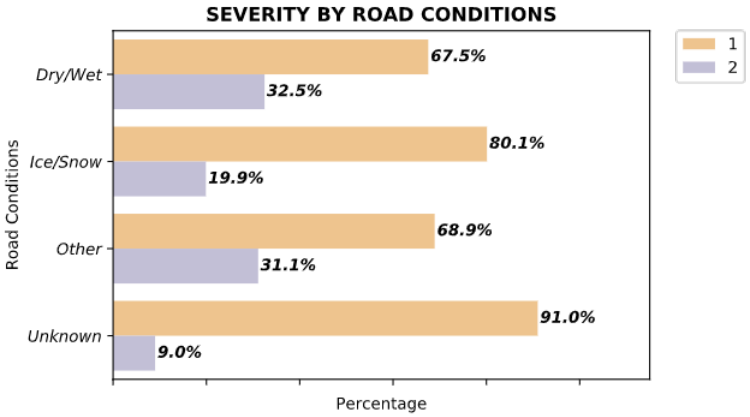
3 DATA PREPARATION

In this phase the data have been modified according to the observation made during the previous one. The columns of the dataset are simplified, some of the values are grouped together and the missing values are taken care of. The next graphs show the resulting data obtained once the transformations have been implemented onto the columns according to the previous analysis. Here is presented the new distribution for the various columns of the dataset:

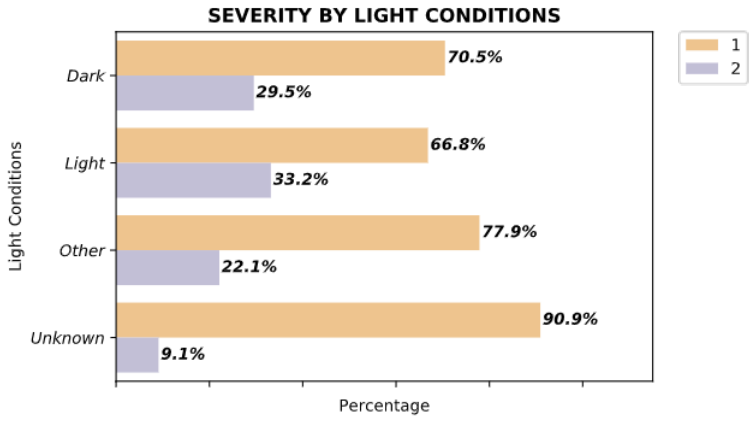
Weather:



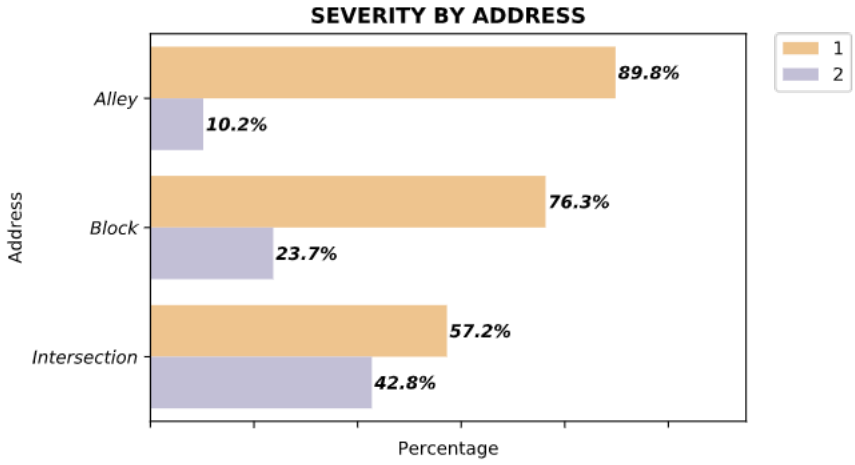
Road Conditions:



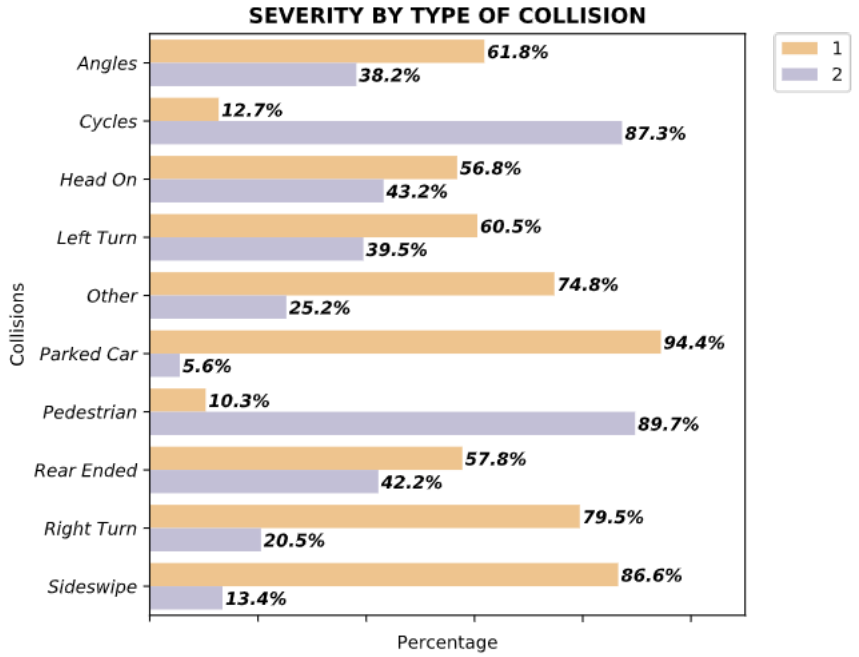
Light Conditions:



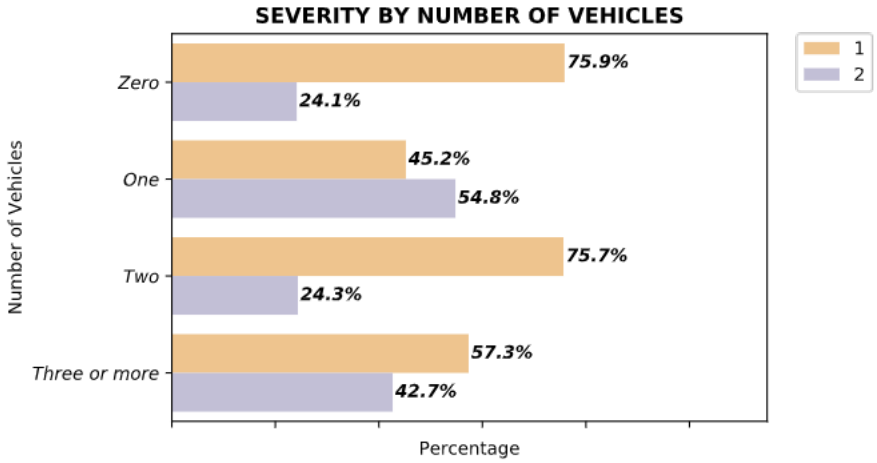
Address:



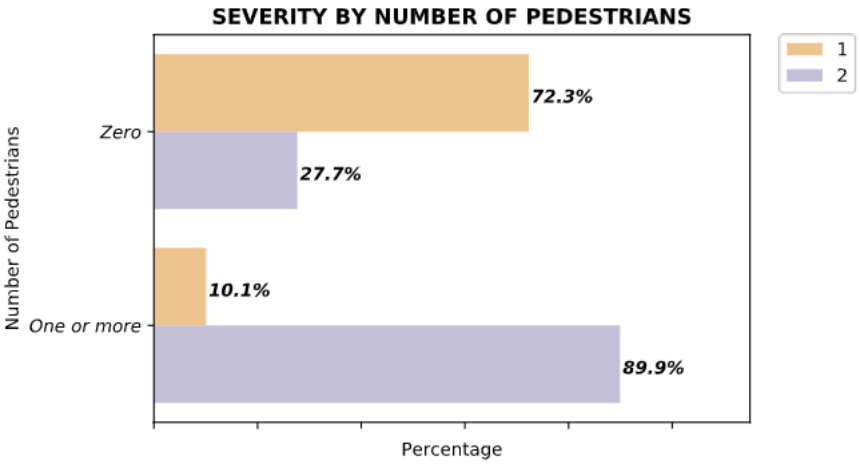
Collision Type:



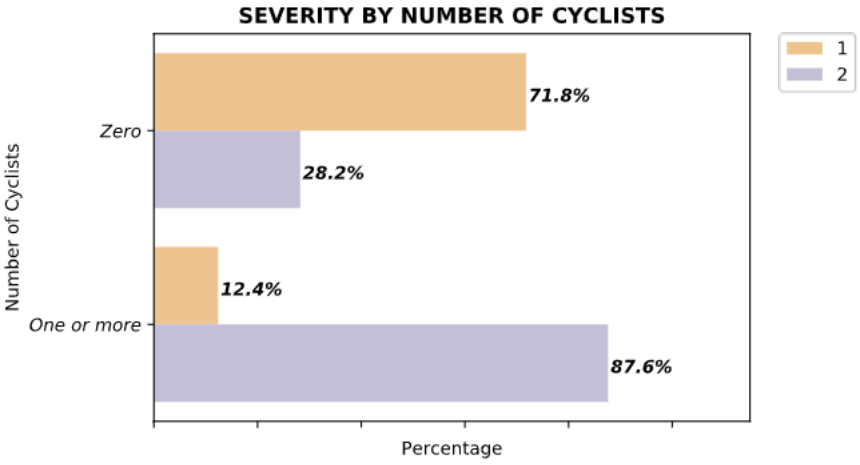
Number of Vehicles:



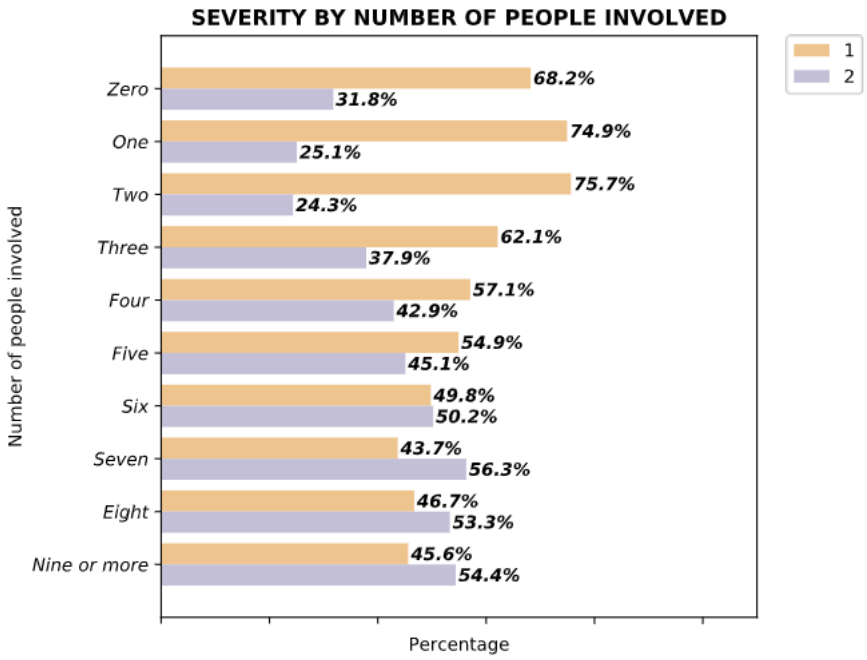
Number of Pedestrians:



Number of cyclists:



Number of people:



4 Modeling Phase

In this phase a model is built that could reproduce the results. Being the problem a classification one based on unbiased dataset a logistic regression algorithm will be implemented. For first the dataset is split into a train and test part.

Secondly the logistic regression algorithm is applied to solve the binary classification problem

5 Model Evaluation

Having to deal with an unbiased data set the measurement of the accuracy of the model is done applying the F-SCORE analysis. The final F-Score value obtained lies around the 0.85.

6 Model Deployment

Once the model is accepted it can be put in act on the street analyzing the conditions present at the moment of the accident. The feeding of these info to the model makes possible to obtain an immediate answer about the criticity of the event and give more time to the interested entities such as hospital, policemens etc.. to activate and response in a fast and proper way to the emergency.