Seattle City Traffic Accident Segmentation with K-Mean Clustering

Ma Yu | 2020-09-30

# Introduction

Traffic accident has been a great threat to public health and security. It causes the loss of properties and lives, for both individual and society. Traffic accident segmentation research categorize different types of accidents, summarise their main characteristics. Successful segmentation can have significant benefit to the public traffic safety and transportation efficiency by multiple measurements, such as design the new road system, reinforce aged infrastructure in critical spots to reduce the accidental risk, redistribute assistance resource for timely rescue in case of emergency, alert divers to pay more attention to accident-prone condition and so on.

This project examines the collisions data of Seattle since 2004 till 2020, and identifies some of the typical dangerous situations for drivers in different area of the city in Seattle, with a special focus on the severity of the accident, applying K-Mean clustering method. The aim of this research is to have a better understanding of the current traffic situation and formulate appropriate prevention strategies and actions.

# Data preparation

## Data Source

The master dataset is acquired from Seattle Geo-information portal – Seattle GeoData[1], where provide the complete record of the collision information in the city, including the date, time, location, type of collision, involvement and severity and so on. There are also query service [2] and attribute information [3] available on that site. However, to perform geographical analytics, it still requires Geo-JSON data of the city, which could be find in the same portal [4]

## Feature Selection

The main source contains all types of collisions data in Seattle city from 2014/01/01 to 2020/05/20, There are 37 attributes and 194673 records in total and 184920 out of 194673 valid records are selected by ignoring "Unmatched" values in "STATUS" column and "Not Enough Information, or Insufficient Location Information" values in "EXCEPTRSNDESC" column. As a result, around 5% records cannot be used for the training and are removed from the datasets.

In the 37 features, there are 22 may contribute to accidental severity in certain way, and they have to be transformed into appropriate data format for further processing and exploratory. The table below summarized the treatment of different types of the feature.

|  |  |
| --- | --- |
| **Type** | **Selected Features** |
| Binary | INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR, SEVERITYCODE |
| Float | X, Y |
| Date/Time | INCDATE, INCDTTM |
| Categorical | ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND, ST\_COLCODE |
| Int | PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT |

## Data Cleaning

### Formatting Date/Time Features

we could extract time related information and create new attributes of year, month, day in the week and hour by formatting date time attributes. It allows us observe how the occurrence of accidents would change along time. It appears to be more accident happening in certain period of time in a day than others, these features may be helpful to verify the light condition in further steps. However, accident rate does not indicate to be largely influence by month or day in the week and we do not have particular interest on the same trends over years in this project, so we would not dig into much details.

### Encoding Binary Features

Most models are not able to read non-numeric feature and use as them the proper input, thus it is fundamental to transform the non-numeric features into numeric ones. Non-numeric features in general includes binary and categorical features, we would first start with binary ones that is relatively more straightforward, since the result are Boolean values and can be either “1” or “0”. All the value such as “Yes”, “true”, “Exist” can be defined as “1”, instead “0” represent “No”, “false”, “Not Exist”. sometimes, the value of a feature can be multiples, they can be treated as binary features if we just need to know their positive/negative status

### Filling Missing Values in Categorical Features and encoding

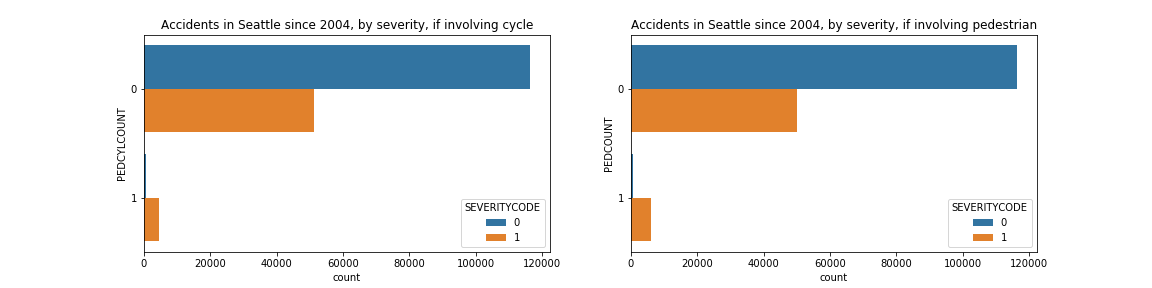
The important step before encoding categorical features is filling the unknow values and drop the blank row. Here we compare similar corelated features, cross-group them and assign the none value with the most frequent value in the group. For example, aggregate “WEATHER” feature grouping by” ROADCON”: there are 104902 “Clear”, 16037 “Overcast” and 871 “unknown” in the “ROADCON” group, it is reasonable to assign the nan value as “Clear”, as a dry road is more probably given by a clear weather. The same method works in the other way around and can be applied to all the related features and reduce the overall null values.

But there are exceptions, when the most frequent value is not much larger than the rest, or the top n values do not appear notable difference, or the none value is the largest itself, the cross comparing should not be applied to avoid the bias. I drop all the rows with none value existing in my selection (“ADDRTYPE” and “JUNCTIONTYPE”, “ST\_COLDESC” and “COLLISIONTYPE”, “WEATHER” and “ROADCOND”, “LIGHTCOND” and “INC\_hour”) and as the result 6.28% record is removed. The “INC\_hour” has too many missing values so it will not be used to train the model. Moreover, the less frequent value in the feature, for instance the one less than 1000 rows, will be merged into “other”, because too many values would increase the complexity and reduce the generalisation capacity of the model.

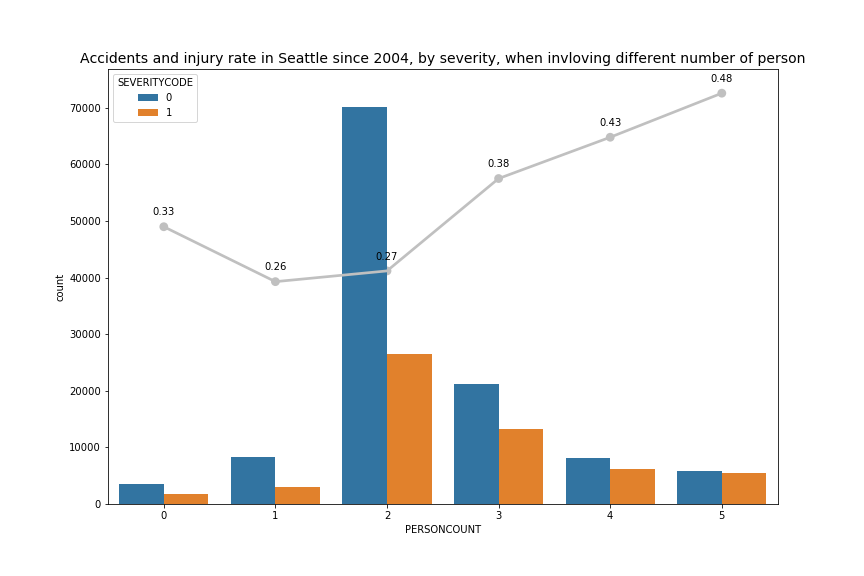
# Data Exploratory

## Continue Features

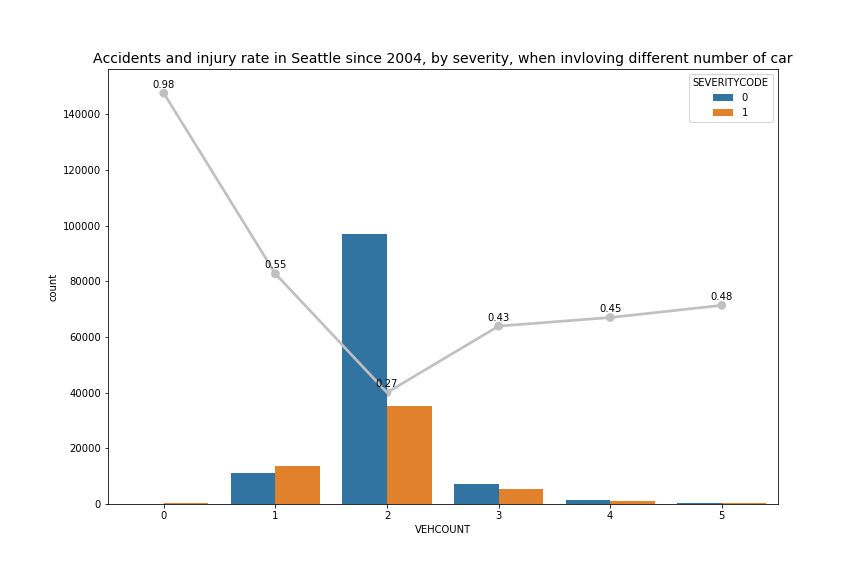
The involved pedestrian and cycle number in an accident show similar characters, very few cases have more than one involved so the features can be considered as binary attributes. The count plots below compare the total quantity of car accident and injuries since 2004 when different number of pedestrian or cycle involved. It is clear that the absolute majority of accidents do not touch upon bike or pedestrian, and the injury rate is relatively low. In this case, around 120k only has property damage and 50k get injured. On the contrast, incidents involve other pedestrian or cycler has much lower records, no more than 5k altogether, but it is almost sure that someone could get hurt. the high injury rate may due to the fragility of bicycle or pedestrian when encountering collision with cars



The plot below compares the accidents and injury rate as increasing people are involved in the collision. Since the number of accidents with more than 5 people is ignorable, they are considered as 5 people. It is noticeable that the number of accidents involving 2 people are far larger than the rest, whereas the injury rate arrives at bottom when the accident involves 1 people, 0.26. whenever exist people involved in accidents, the injury rate increases as growing numbers of people are involved, it nears 50% when more than 5 people get involved.

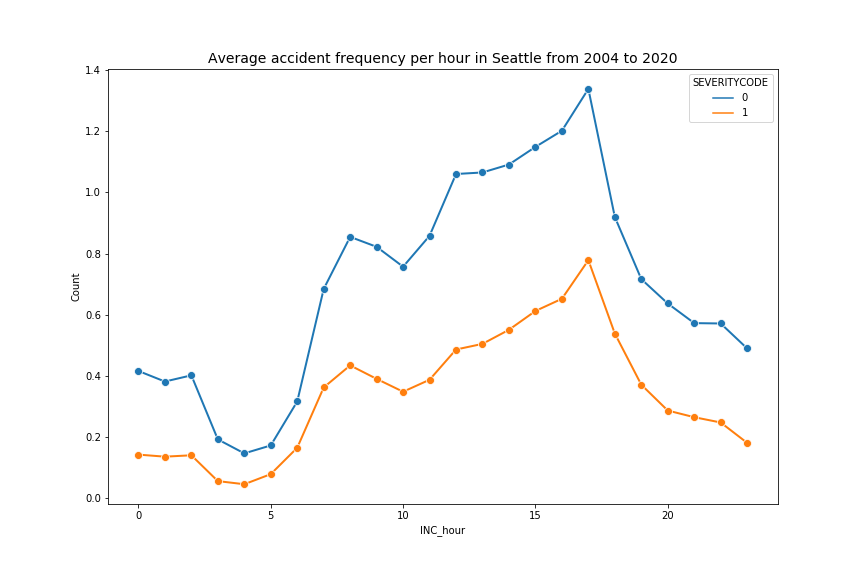


The relationship between accidents, injury rate and involvement of the car has comparable trends with previous chart, while the injury rate is extremely high when the car number happen to be 1(0.98) and 2(0.55). we may have some clues from Table, when an incident involves one or two cars, it often accompanies the collision with pedestrian or cycle to different extent. From previews plots, we know the involvement of bike/pedestrian are likely to cause injury, this may explain this phenomenon.



|  |  |  |
| --- | --- | --- |
| **VEHCOUNT** | **COLLISIONTYPE** | **Count** |
| 0 | Cycles | 189 |
|  | Other | 1 |
| 1 | Other | 13503 |
|  | Pedestrian | 6262 |
|  | Cycles | 4974 |
|  | Angles | 23 |
| 2 | Angles | 31931 |
|  | Parked Car | 30931 |
|  | Rear Ended | 26641 |
|  | Sideswipe | 16717 |
| 3 | Rear Ended | 5183 |
|  | Parked Car | 3565 |
|  | Angles | 1841 |
|  | Sideswipe | 850 |
| 4 | Parked Car | 944 |
|  | Rear Ended | 874 |
|  | Angles | 264 |
|  | Sideswipe | 128 |
| 5 | Parked Car | 373 |
|  | Rear Ended | 171 |
|  | Angles | 82 |
|  | Sideswipe | 53 |

The line chart compares the mean hourly frenquecy of accident in a day in different severity. The two lines illustrate a roughly porpotional relationship and both of them are more frequent during the day. There are two peak hour, 8 am and 5 pm, which are corrisponde to the rush hour. However, due to the large quantity of missing value, 24k that acount 15% of the total records, this feature will not be used for trainning.



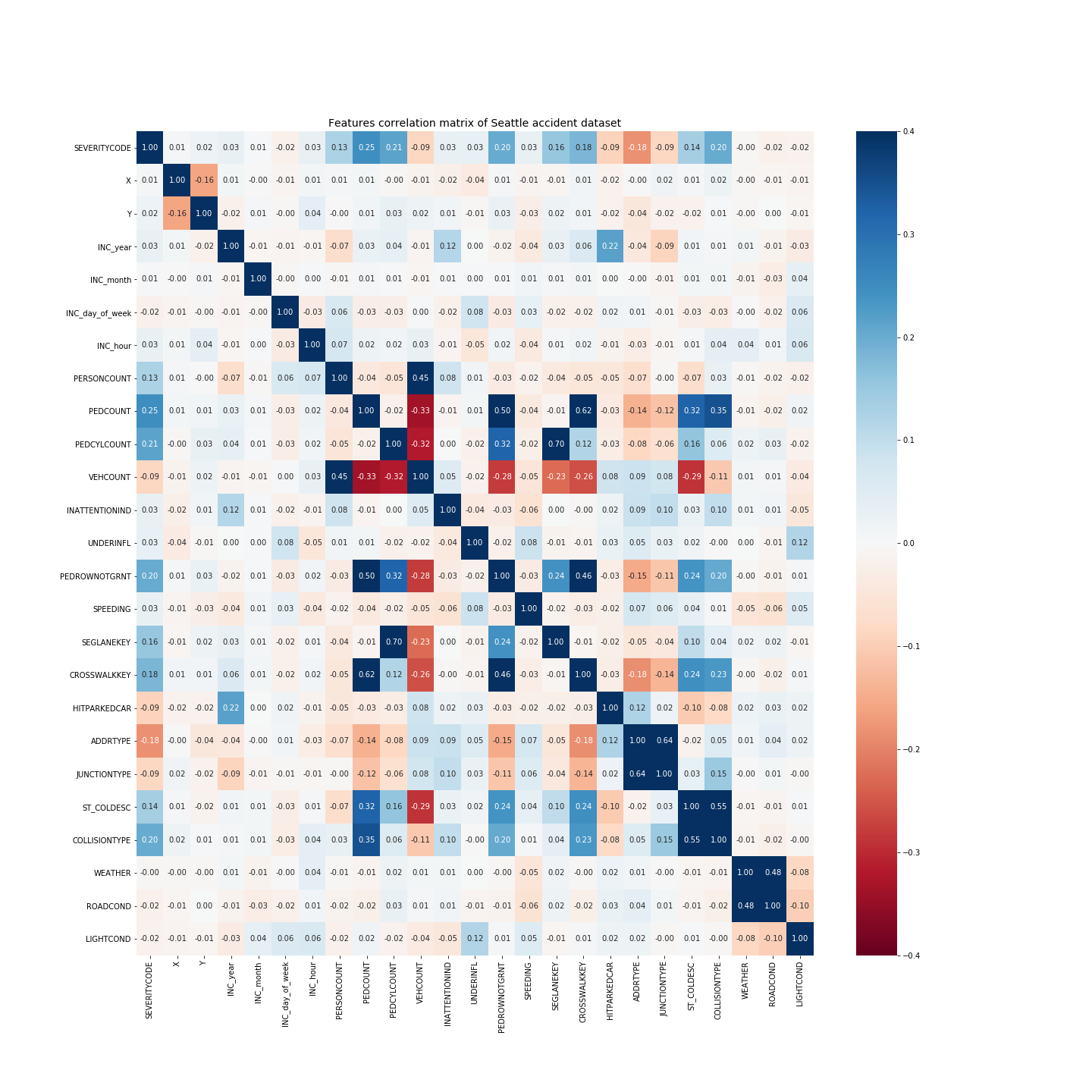
## Correlation Matrix and Feature Selection for The Modelling

correlation commonly refers to the degree to which a pair of variables are linearly related. The most familiar measure of dependence between two quantities is the Pearson's correlation coefficient. Given a series of n measurements of the pair indexed by , the correlation coefficient can be defined as

where and are the sample means of X and Y. The value of a correlation coefficient ranges between -1 and +1. The correlation coefficient is +1 in the case of a perfect direct linear relationship, −1 in the case of a perfect inverse linear relationship, 0 when the variables are independent.

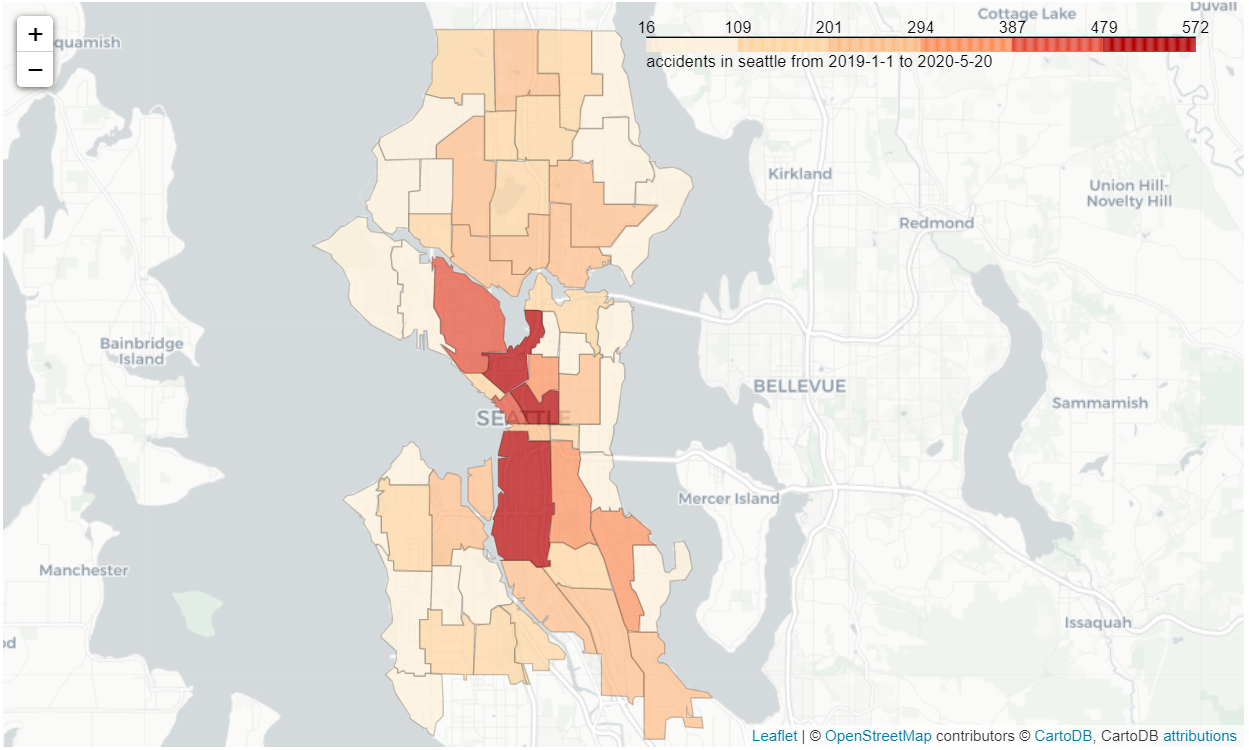
Pearson's correlation coefficient is widely used to compare the relationship between two continue variables, whereas in the case of two categorical variables or mix of continue and categorical variables, Cramér’s V is a better choice. however, for limit time available, I would simply conduct Base N encoding for categorical variables and generate a Pearson's correlation matrix for further analytics. The correlation matrix is visualised with a heatmap and will help to select the most relevant features for data modelling. In this graph the darker red means higher negative correlation and darker blue represents higher positive correlation. The relationship between “SEVERITYCODE” and other variables has more weight in my consideration, as injury rate is the most important focus of this project.

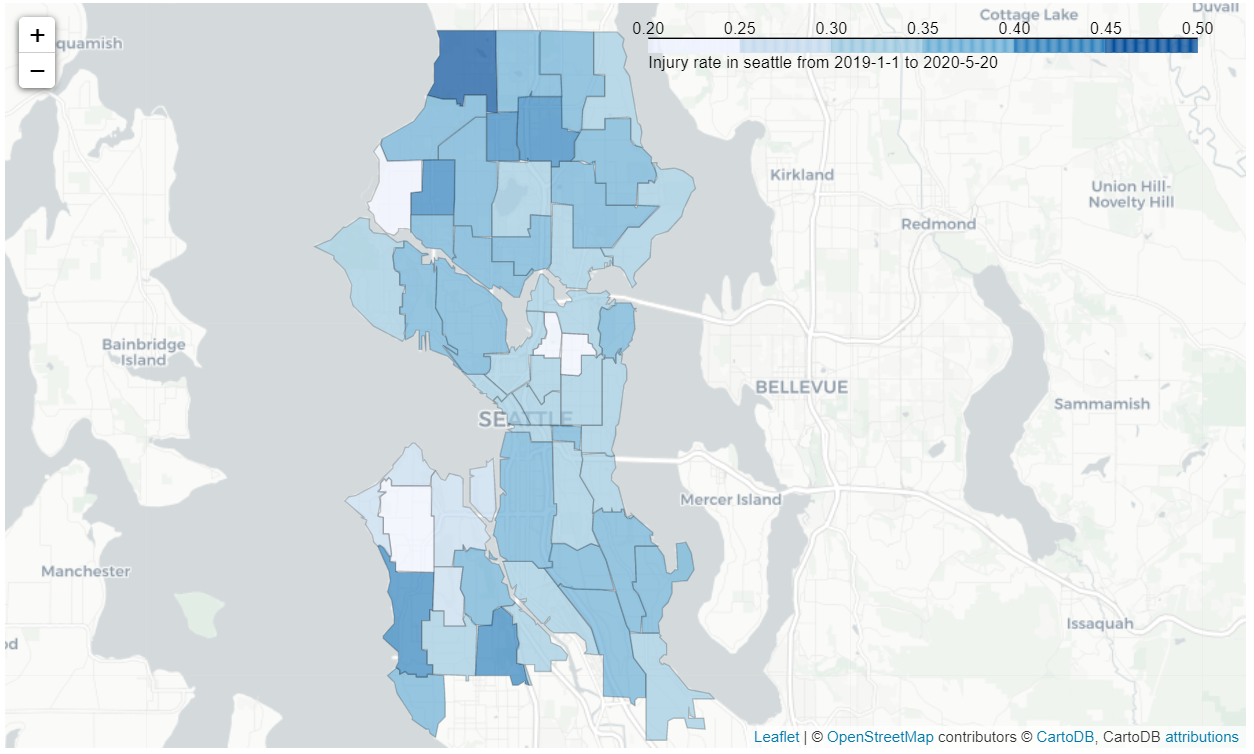
The heatmap provide a base for further feature selection in the modelling. First of all, I decide to exclude the time and location factors in the segmentation and drop “X”, “Y”, “INC\_year”, “INC\_month”, “INC\_day\_of\_week”, “INC\_hour” columns. Then, “INATTENTIONIND”, “UNDERINFL”, “SPEEDING”, “WEATHER”, “ROADCOND”, “LIGHTCOND” are removed either due to the weak correlation with other features especially “SEVERITYCODE”. At last, Categorical features will be encoded with one-hot encoding for the modelling, to reduce the dimensions and the risk of overfitting, I prefer the features with less unique values. as a result, “ST\_COLDESC” and “JUNCTIONTYPE” are abandoned because their information can be better generalised by “COLLISIONTYPE” and “ADDRTYPE”.



## Geography Characteristics

Although geography factor will not be taken into the consideration in the segmentation, it is still useful to have some idea about the distribution of accidents and injury rate among different neighbourhood in the city. Due to the limit capacity of my laptop, I just select the data after 2019 which is approximately 10k rows. The choropleth map below shows reprehensively the geographical distribution of accidents and jury rate. It is clear that a higher density of incidents in the central districts, where locate the CBD and urban centre. Compared to the inland neighbourhoods, the ones along the waterside has less accidents in record. In the contrast, the injury rate map shows a different situation. As the district become farther to the central districts, the percentage of injuries in accidents are likely to grow, although less accidents are reported in these areas.





# Data Modelling

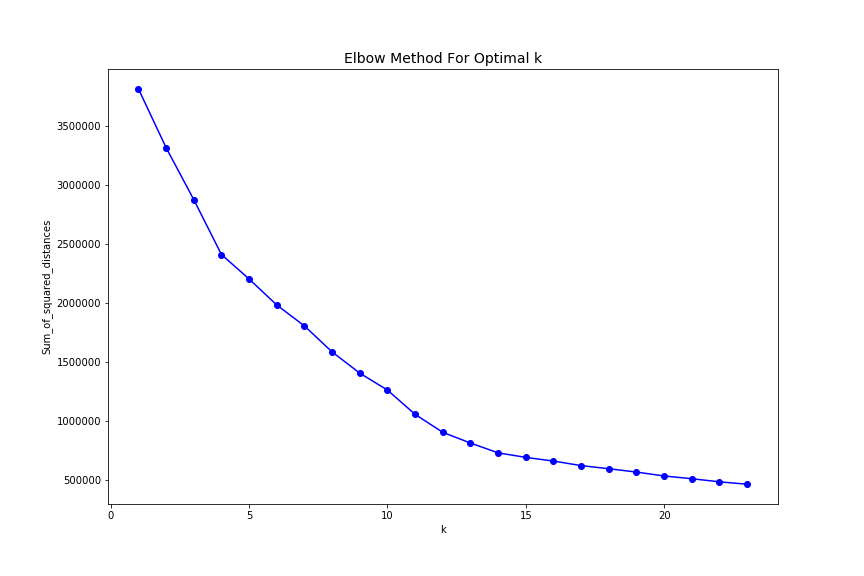
## K-Mean Clustering

K-means is a type of unsupervised learning and one of the popular methods of for cluster analysis in data mining. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean i.e., k-means clustering minimizes within-cluster variances (squared Euclidean distances). Given a set of observations, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets so as to minimize the within-cluster sum of squares (i.e. variance). The objective is to find:

where is the mean of points in .

## Find the Best K

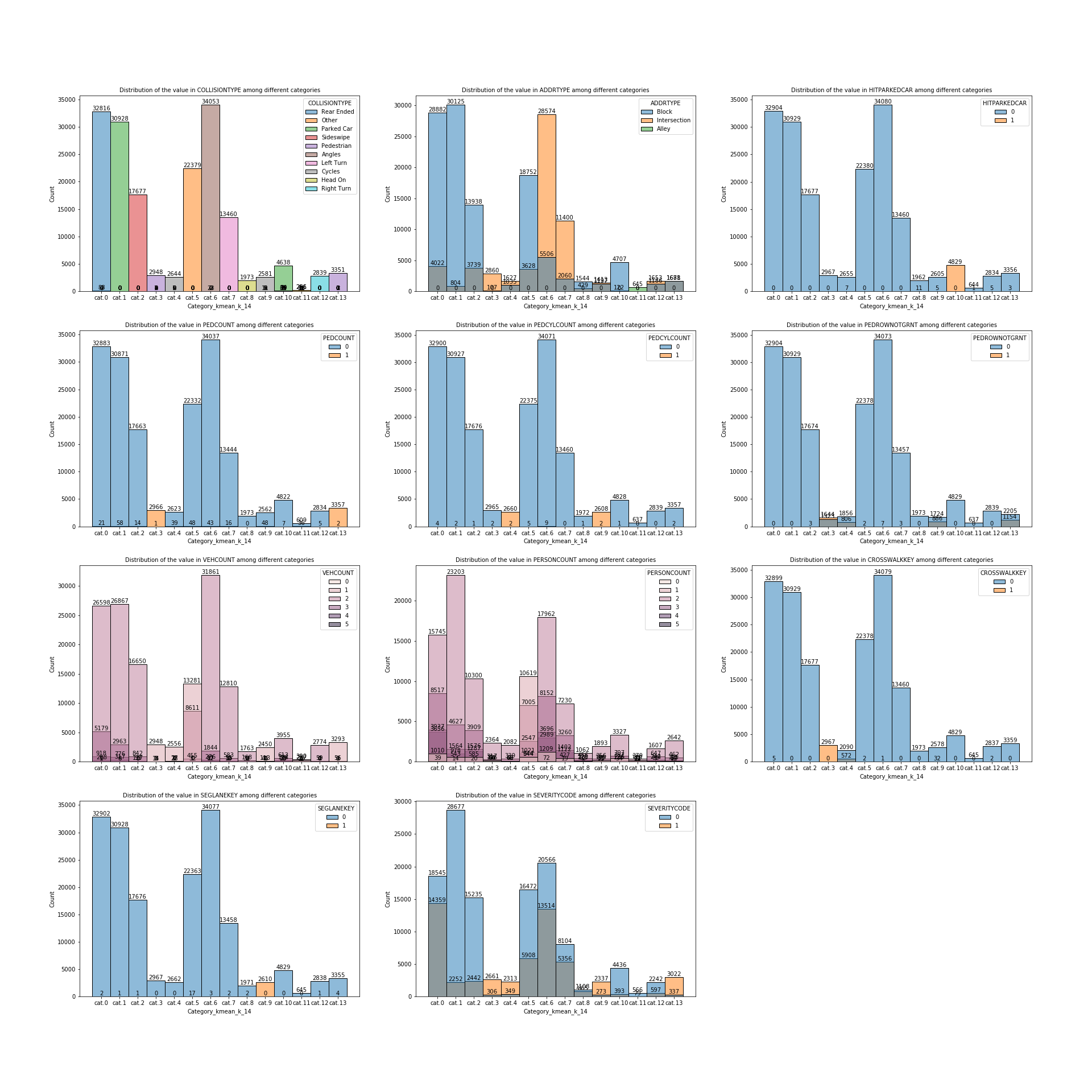
Determining the number of clusters in a data set is a frequent problem in data clustering. We commonly use the elbow method to find the best k in the k-Means algorithm. With the previously processed data set, we normalise all features and scale the continuous features using min-max scaler, as the feature matrix is a mix of binary and continuous features. For each k value, we will initialise k-means to identify the sum of squared distances of samples to the nearest cluster centre. As k increases, the sum of squared distance tends to zero. Below is the plot of sum of squared distances for k in the predefined range. If the plot looks like an arm, then the elbow on the arm is optimal k. I launched the model multiple times and the result may slightly vary, finally I chose 11 and 14 to be the candidate of the best k.



## Insight and Segmentation

The selected k is the number of partitions and each row in the original dataset are assigned a label as its new category feature. To better characterise these categories, I generated a matrix of histogram plot of the accidents by imported features on different categories, different colour represents the proportion of each features compared to the rest in the same category. It is clear that each plot has the same distribution of the total quantity on different categories, while the proportion of each features in different categories vary significantly, which coloured differently.

From the matrix of plot below, I gain insights and conclude that the primary characteristics is “COLISIONTYPE”, “ADDRTYPE”, “SEVERITYCODE”, because they are widely scattered on all the categories. On the contrast, other features tend to concentrate on one or seldom categories thus defined as secondary features. As a consequence, the 14 initial categories can be merged into 11 major types with 3 subsets. Table indicates the summary and characteristics of different categories.



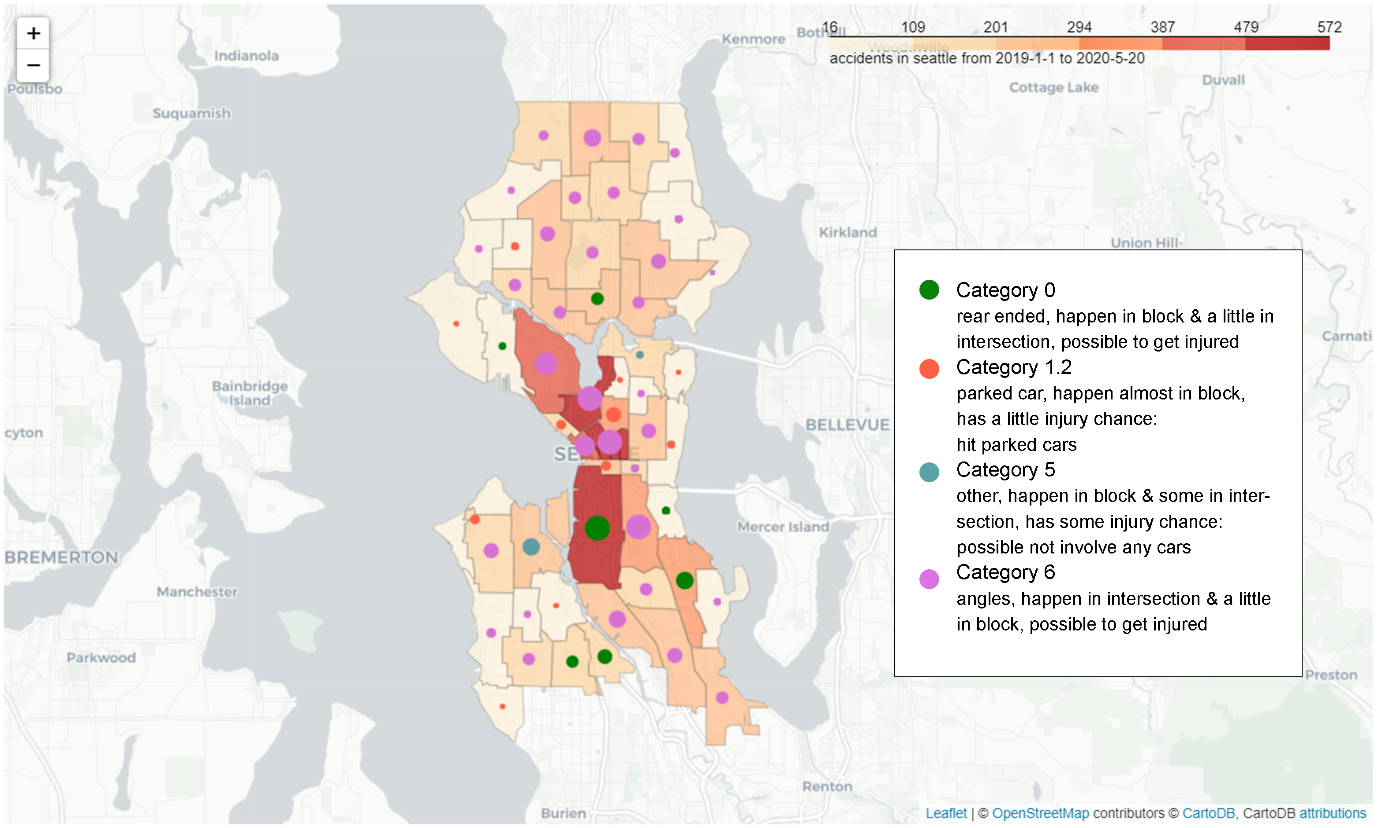
|  |  |
| --- | --- |
| **Category** | **Description** |
| Category 0 | rear ended, happen in block & a little in intersection, possible to get injured |
| Category 1.1 | parked car, happen almost in block, has a little injury chance |
| Category 1.2 | parked car, happen almost in block, has a little injury chance:  hit parked cars |
| Category 2 | sideswipe, happen in block & a little in intersection, has a little injury chance |
| Category 3.1 | pedestrian, happen in block & a little in intersection, highly possible to get injured:  involve pedestrian, possible when pedestrian right of way was not granted, almost not involve any cars, have a key for the crosswalk |
| Category 3.2 | pedestrian, happen in block & intersection, highly possible to get injured:  involve pedestrian, some pedestrian right of way was not granted, almost not involve any cars |
| Category 4.1 | cycle, happen almost in intersection, highly possible to get injured:  involve cycle, some pedestrian right of way was not granted, almost not involve any cars, some chance to have a key for the crosswalk |
| Category 4.2 | cycle, happen in block & intersection, highly possible to get injured:  involve cycle, some pedestrian right of way was not granted, almost not involve any cars, have a key for the lane segment |
| Category 5 | other, happen in block & some in intersection, has some injury chance: possible not involve any cars |
| Category 6 | angles, happen in intersection & a little in block, possible to get injured |
| Category 7 | left turn, happen in intersection & a little in block, possible to get injured |
| Category 8 | head on, happen in block & a little in intersection, possible to get injured |
| Category 9 | other & parked car, happen in alley, has a little injury chance |
| Category 10 | right turn, happen in block & intersection, has some injury chance |

At last, let’s take another step and have a further look on what is the spatial distribution of these categories in the city. The Table summarises the frequency of the most frequent categories in each neighbourhood. It is obvious that the Category 6 in the prominent type of collision in most the neighbourhood, 33 out of 53 neighbourhoods has reported it as the most frequent incident. They usually perform as angle collision, happen mostly in intersection and has certain chance to get hurt. the second most widespread is category 1.2, they are top 1 accident type in 11 areas, and are defined as parked collision involving hitting parked car, usually take place in block and seldom lead to injury. In the contrast, Category 0 and 5 are minorities, and has the highest frequency in only 7 and 2 neighbourhoods.

Count of most frequent categories in each neighbourhood

|  |  |
| --- | --- |
| **Category** | **Count** |
| Category 6 | 33 |
| Category 1.2 | 11 |
| Category 0 | 7 |
| Category 5 | 2 |

A metropolis such as Seattle has complex urban structure, the social-economic and physical condition can differ hugely across areas. A detailed categorization in each area would be more helpful to understand the specific local situation and take correspondent improving measurements. When integrating this information into the accident choropleth map previously create, the distribution of accidents and the most frequency type is visualised in a single map, in this way their relationship is also well presented. It could be constructive for urban planners and administrators when concluding any decisions to improve road security in the area.



# Future Studies

In the correlation research, I calculated correlation between each other of all the attributes using Pearson's correlation coefficient, indifferent the categorical, continue or binary. As matter of fact, this method function perfectly for continue or binary attribute but not for the categorical independents. A more accurate research should conduct Cramér’s V for categorical attributes, and the correlation matrix may present in differently and impact on the partitioning to some extent. Moreover, I applied the segmentation in only one type of clustering in this research, the k-mean clustering. Further research may conduct different clustering models with the same input and compare their results

# Reference

Collisions - Seattle GeoData - ArcGIS Online

<https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0>

Seattle Neighbourhoods Geo-JSON:

<https://opendata.arcgis.com/datasets/fbf6ca85b6b0408da346c8896b6f8aef_0.geojson>

Correlation and dependence, Wikipedia

<https://en.wikipedia.org/wiki/Correlation_and_dependence>

The Search for Categorical Correlation, Shaked Zychlinski, 2018-02

<https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9>

Here’s All you Need to Know About Encoding Categorical Data (with Python code), SHIPRA SAXENA, 2020-08

<https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/>

k-means clustering, Wikipedia

<https://en.wikipedia.org/wiki/K-means_clustering>

Tutorial: How to determine the optimal number of clusters for k-means clustering, Tola Alade, 2018-05

<https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f>