Seattle City Traffic Accident Segmentation with K-Mean Clustering

Applied Data Science Capstone by IBM/Coursera

Ma Yu | 2020-10-03

1. Introduction

Traffic accidents have been a great threat to public health and security. It causes the loss of properties and lives, for both individuals and society. Traffic accident segmentation research aims at categorizing different types of accidents and summarising their main characteristics. Successful segmentation can have significant benefit to the road safety and traffic efficiency by a series of measurements, such as design the new road system, reinforce aged infrastructure in critical spots, 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 collision data of Seattle from 2004 till 2020, and identifies some of the typical dangerous situations for drivers in different areas of the city in Seattle. It has a special focus on the severity of the accident applying the K-Mean clustering method. The aim of this research is to have a better understanding of the current traffic situation in the city to formulate appropriate prevention strategies and actions.

2. Data preparation

2.1. Data Source

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

2.2. 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.

There are 22 out of 37 features that may contribute to categorisation and Table.1 summarized the different types of selected features. They have to be transformed into appropriate format for further processing and exploratory

Table.1 Types of features

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

2.3. Data Cleaning

2.3.1. Formatting Date/Time Features

we can extract time related information, like “year”, “month”, “day in the week” and “hour”, by formatting date time attributes. It allows us to observe how the occurrence of accidents change over time and the hourly accident frequency can be helpful to verify the light condition in further steps. However, accident rate does not appear to be largely influenced by “month” or “day in the week”, and we do not have particular interest in long-term trends over years in this project, so we would not dig into much detail.

2.3.2. Encoding Binary Features

Most models are not able to read non-numeric features and use them as the proper input, thus it is fundamental to transform the non-numeric features into numeric ones. Non-numeric features in general include binary and categorical features, we would first start with binary ones that are relatively straightforward, since the results are boolean values that are either “1” or “0”. All the values 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

2.3.3. Filling Missing Values in Categorical Features and encoding

A crucial step before encoding categorical features is filling the unknown values and dropping the blank row. Here we cross compare similar features and assign the none value with the most frequent value in the confront 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 “Clear” to “unknown”, 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 cross comparison methods should not be applied when the most frequent value is not much larger than the rest, or the top-n values do not appear to have a notable difference, or the none value is the largest itself, 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 6.28% o f the original record is removed. “INC_hour” will not be used to train the model because it has too many missing values. In addition, the less frequent value in the feature, for instance the one less than 1000 rows, will be merged into “other”, because too many values may reduce the generalisation capacity of the model.

3. Data Exploratory

3.1. Continuous Features

The involved pedestrian and cycle number in an accident demonstrate similar characteristics, only very few cases have involved more than one so the features can be considered as binary attributes. In Figure.1, the count plots compare the total quantity of car accidents and injuries since 2004 when different numbers of pedestrians or cycles were involved. It is clear that the absolute majority of accidents do not touch upon a 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 contrary, incidents involving other pedestrians or cyclists have 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

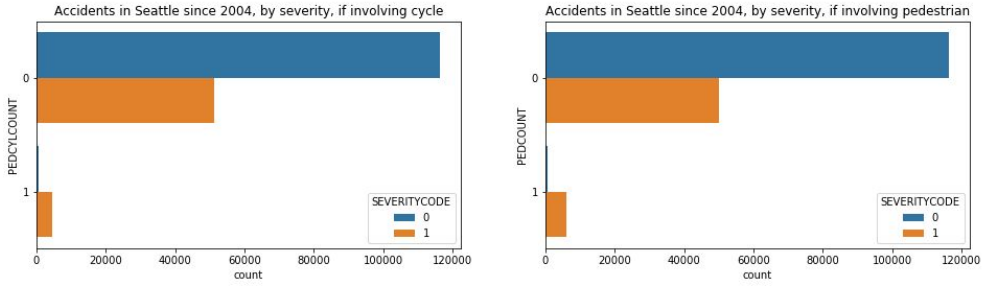

Figure 1

Figure.2 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, 0.26, when the accident involves 1 person. Whenever people are involved in accidents, the injury rate increases as growing numbers of people are involved, and it nears 50% when more than 5 people get involved.

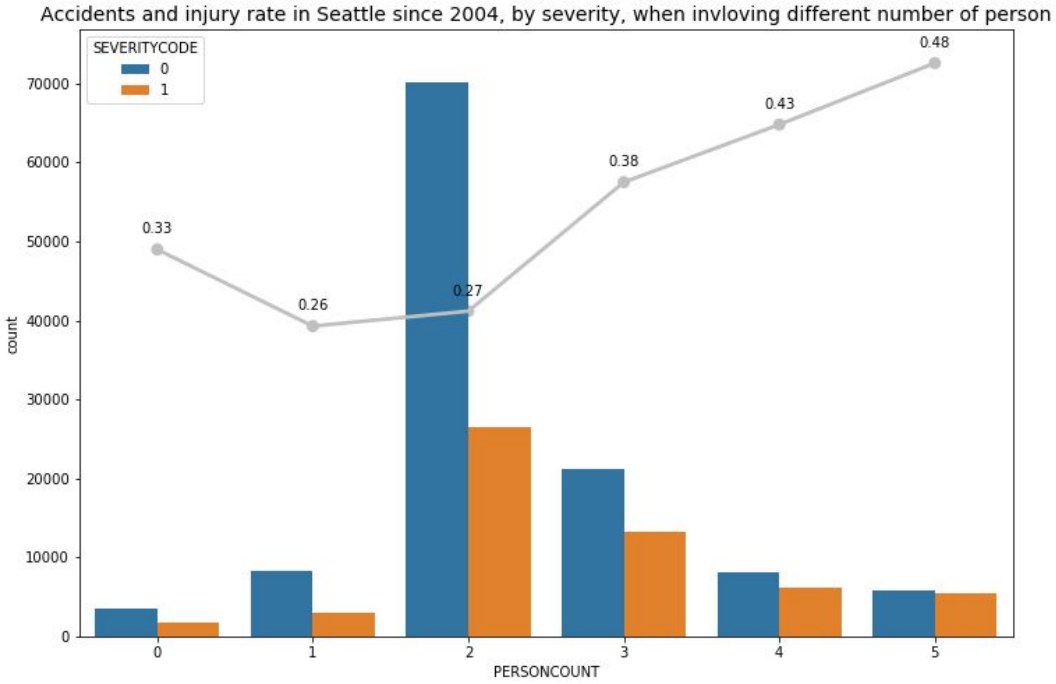

Figure 2

In Figure.3, the relationship between accidents, injury rate and involvement of the car shows comparable trends with the previous chart, while the injury rate is extremely high when the car number happens to be 1(0.98) and 2(0.55). we may get some clues from Table.2 for the reason, when an incident involves one or two cars, it often accompanies the collision with pedestrians or cycles to different extent. From previous plots, we know the involvement of bikes/pedestrians are more likely to cause injury.

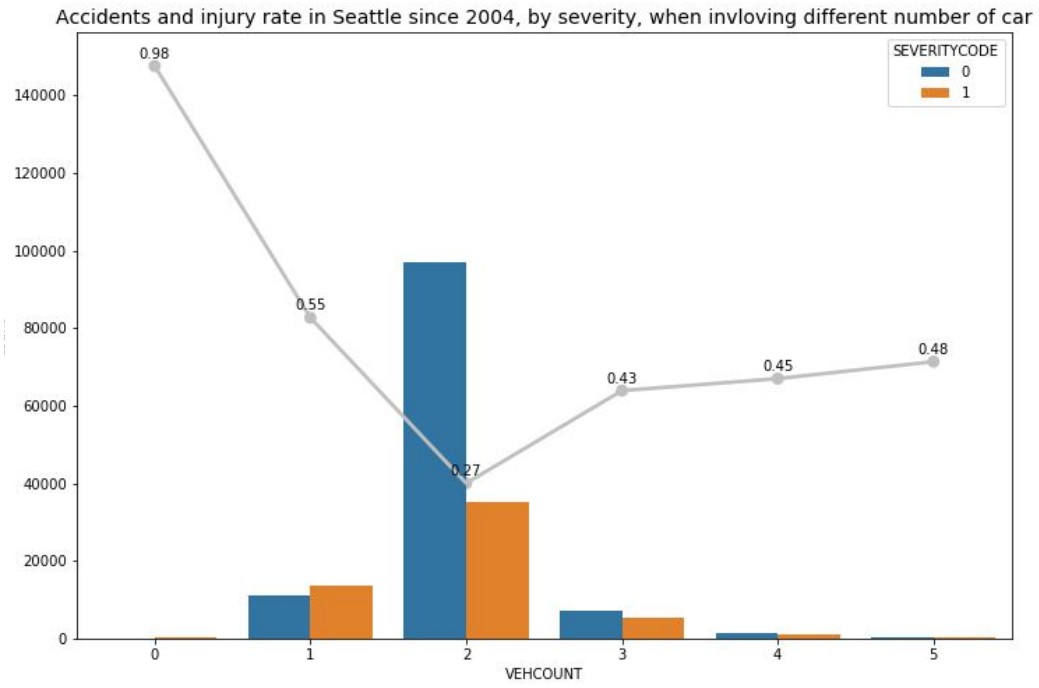

Figure 3

Table.2 Count of number of involved car groupby collision type

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

Figure.4 compares the mean hourly frequency of accidents in a day in different severity. The two lines illustrate a roughly proportional relationship and both of them are more frequent during the day. There are

two peak hours, 8 am and 5 pm, which are correspondent to the rush hour. However, due to the large quantity of missing value, 24k that account for 15% of the total records, this feature will not be used for training.

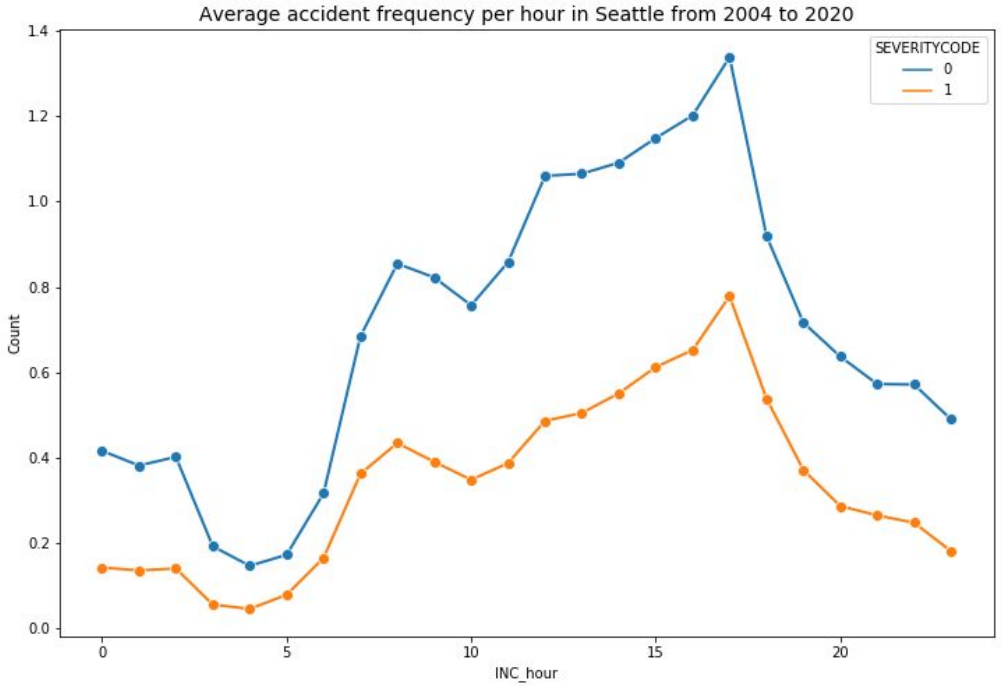

Figure 4

3.2. Correlation Matrix and Feature Selection for The Modelling

The most familiar measure of dependence between two quantities is the Pearson's correlation coefficient. Given a series of n measurements of the pair (X_i, Y_i), the correlation coefficient can be defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ 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[3].

Pearson's correlation coefficient is widely used to compare the relationship between two continuous variables, Cramér's V is a better choice in the case of two categorical variables or mix of continuous and categorical variables[4]. However, for the limited time available, I would simply conduct Base N encoding for categorical variables[5] and generate a Pearson's correlation matrix for further analytics. The correlation matrix is visualised with a heatmap and will be used 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 in Figure.5 provides a reference for modeling feature selection. 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 as well, due to the weak correlation with other

features especially the “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”.

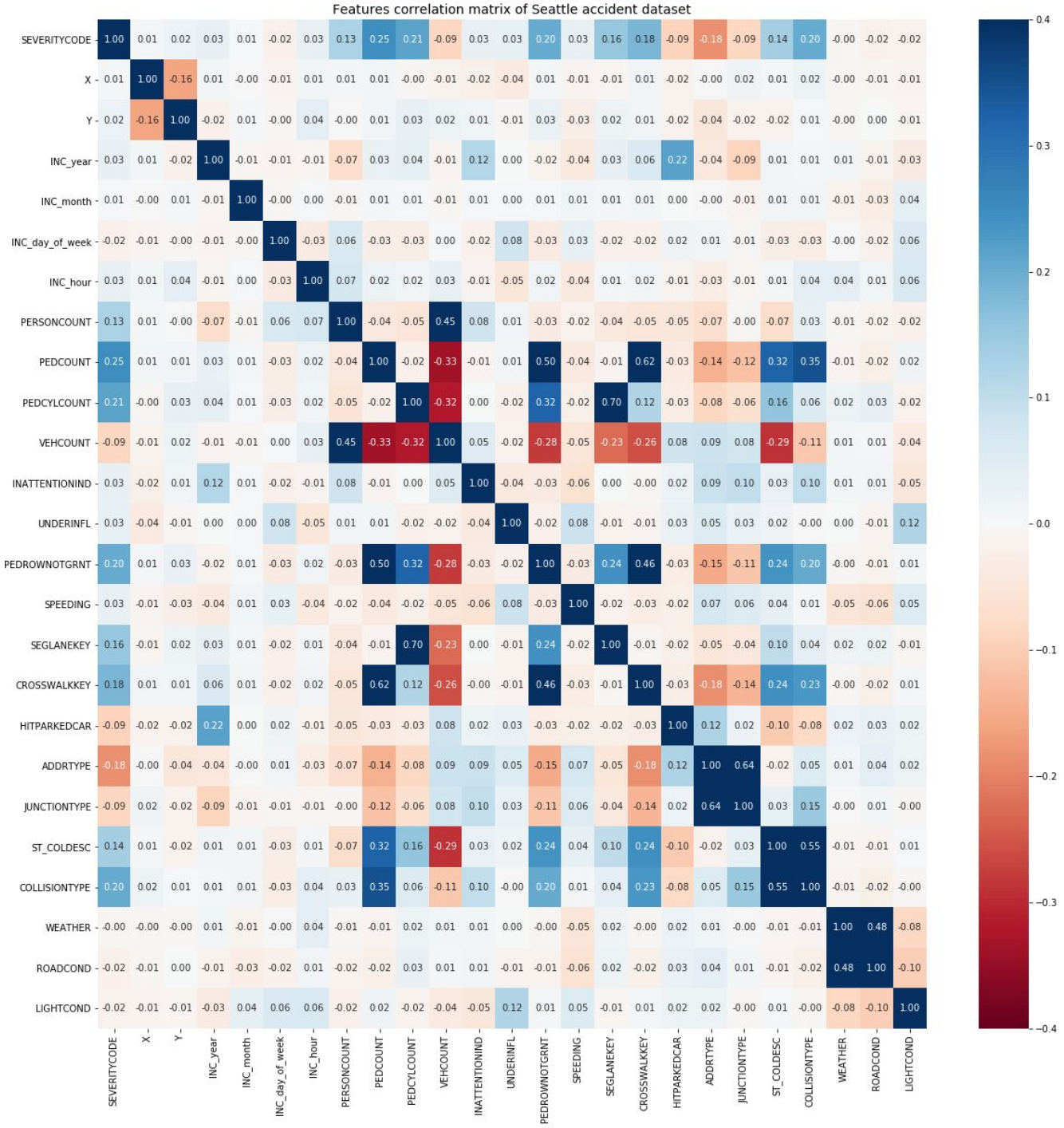

Figure 5

3.3. Geography Characteristics

Although geography factors will not be taken into consideration in the segmentation, it is still useful to have some idea about the distribution of accidents(Figure.6) and injury(Figure.7) rate among different

neighbourhoods in the city. Due to the limited capacity of my laptop, I just selected the data after 2019 which is approximately 10k rows. The choropleth map below shows representatively the geographical distribution of accidents and jury rate. It is clear that a higher density of incidents in the central districts, where the CBD and urban centre. Compared to the inland neighbourhoods, the ones along the waterside have less accidents in record. In contrast, the injury rate map shows a different situation. As the district becomes farther to the central districts, the percentage of injuries in accidents are likely to grow, although less accidents are reported in these areas.

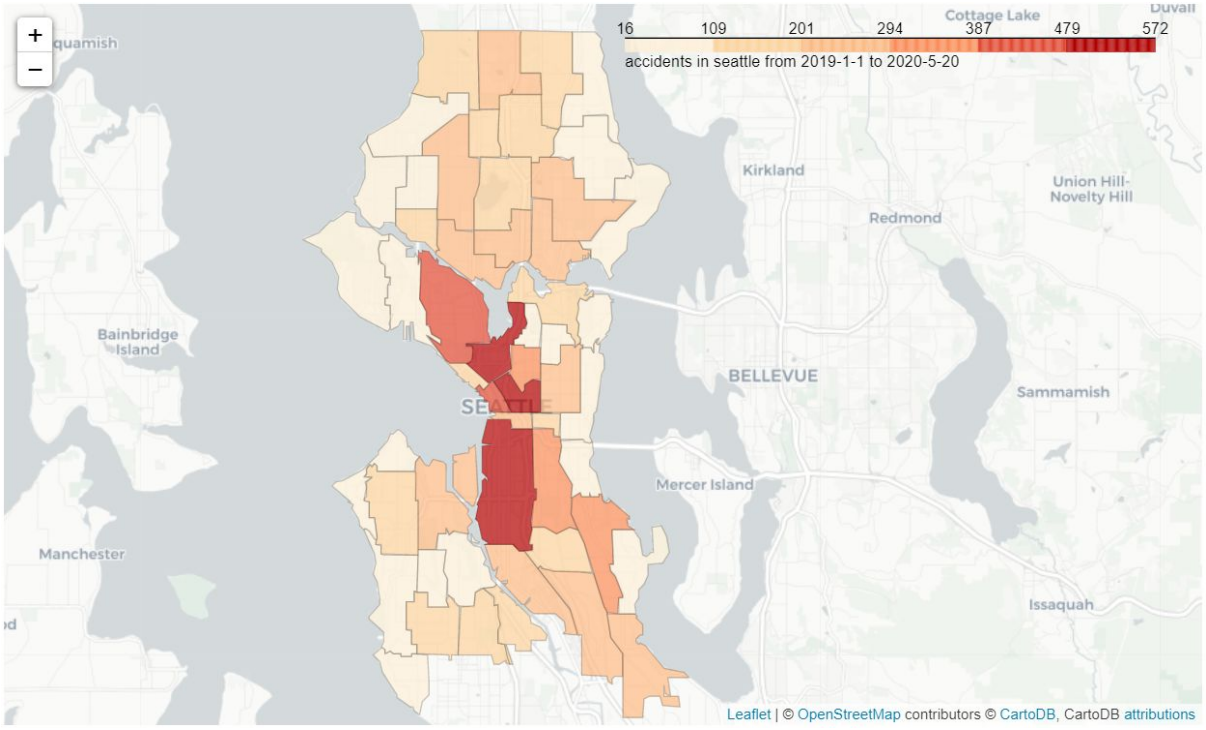

Figure 6

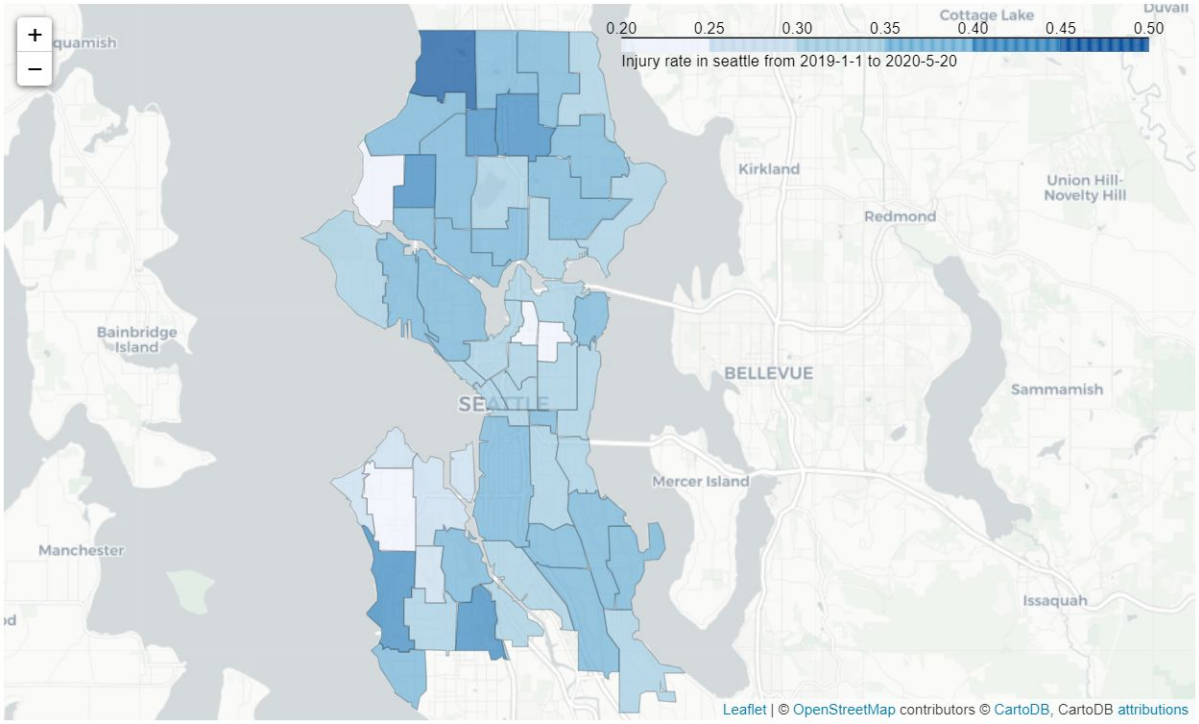

Figure 7

4. Data Modelling

4.1. K-Mean Clustering

K-means is a type of unsupervised learning and one of the popular methods 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 $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets $S = S_1, S_2, \ldots, S_k$ so as to minimize the within-cluster sum of squares:

$$arg\,min \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2 = arg\,min \sum_{i=1}^{k} |S_i| \, Var S_i$$

where μ_i is the mean of points in S_i [6].

4.2. 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. Firstly, I normalise all features in the data set that is previously processed, and scale the continuous features using a min-max scaler, as the feature matrix is a mix of binary and continuous features. Then for each k value, I 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. Figure.8 is the plot of the 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 as the candidate of the best k[7].

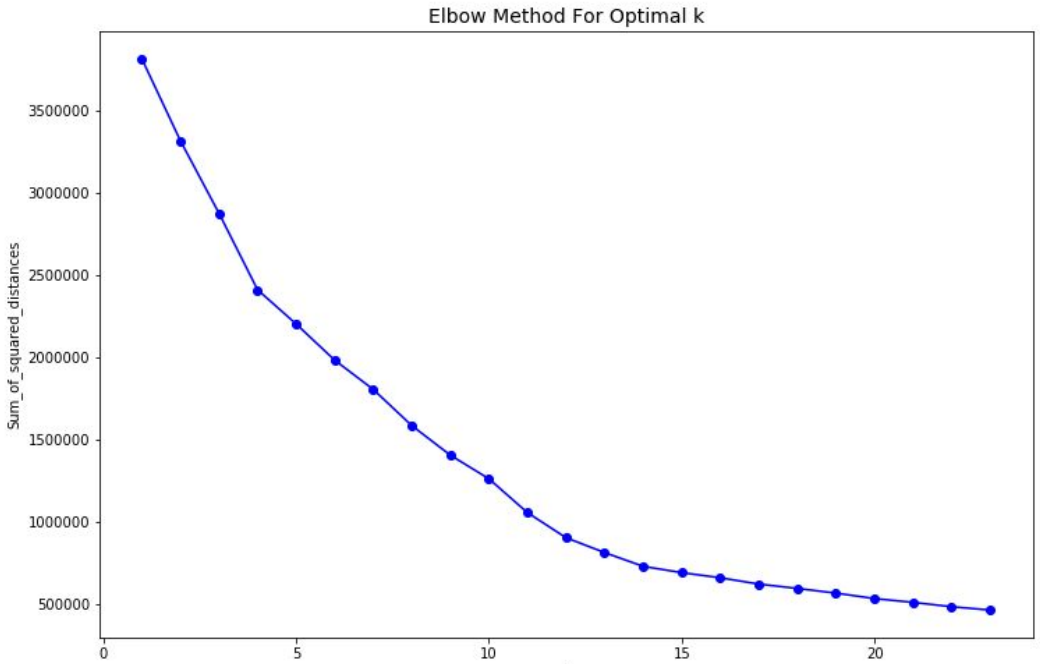

Figure 8

4.3. 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 also coloured differently.

From Figure.9, I conclude that the primary characteristics is “COLISIONTYPE”, “ADDRTYPE”, “SEVERITYCODE”, because they are widely scattered on all categories. In 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.3 summarised characteristics of different categories.

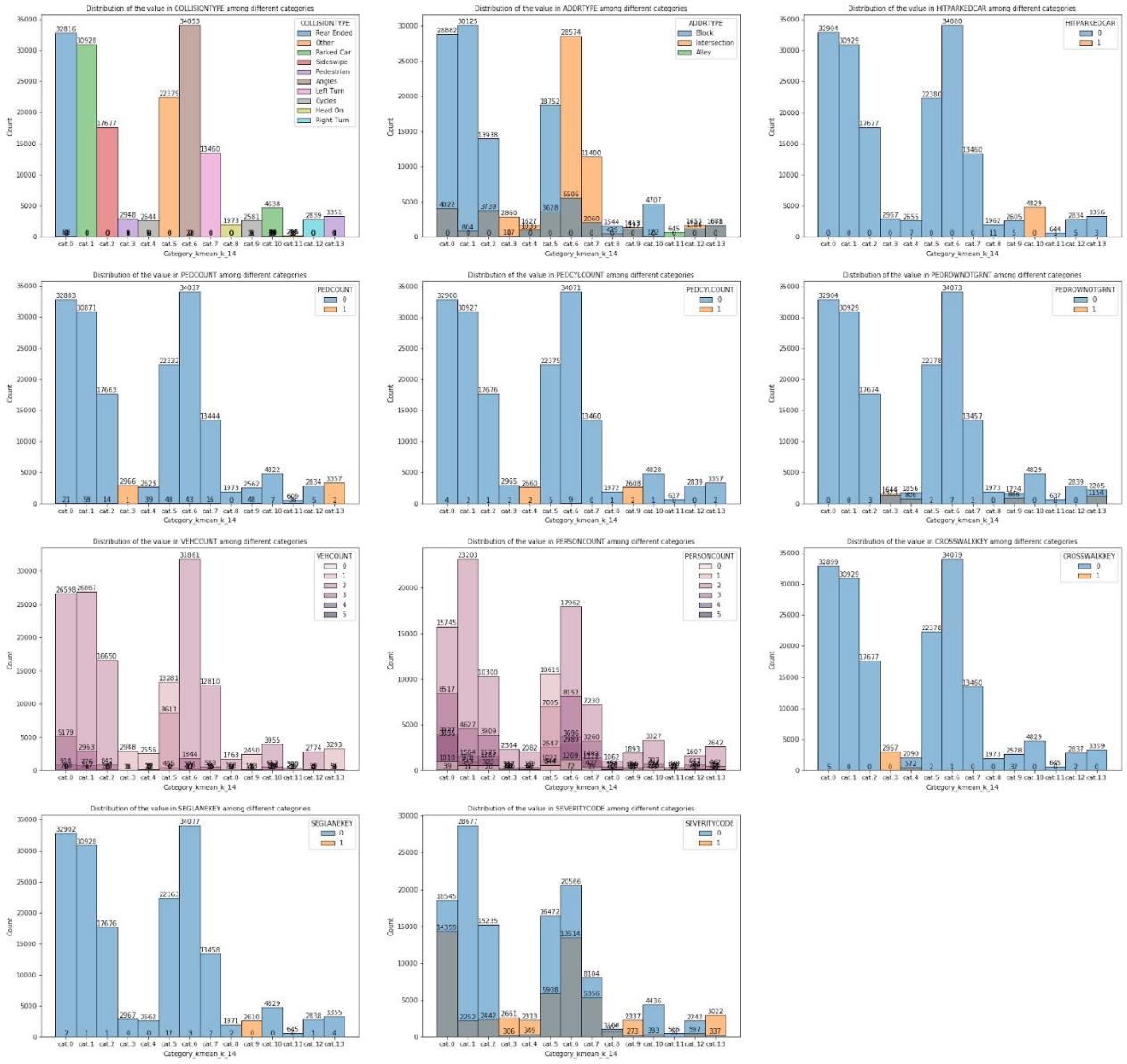

Figure 9

Table.3 Category

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

Finally, let's take another step and have a look at what is the spatial distribution of these categories in the city. Table.4 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 neighbourhoods, 33 out of 53 neighbourhoods has reported it as the most frequent incident. They usually perform as angle collisions, happen mostly in intersections and have a 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 collisions involving hitting parked cars, usually take place in blocks and seldom lead to injury. In contrast, Category 0 and 5 are minorities, and have the highest frequency in only 7 and 2 neighbourhoods.

Table.4 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 at local level would be more helpful to understand the specific situation and take correspondent improving measurements. In Figure.10 I integrate this information into the accident choropleth map. In this way, the distribution of accidents and the most frequency type is visualised in a single map, and their relationship becomes explicit, which could be constructive for urban planners and administrators to improve road security in their policy-making.

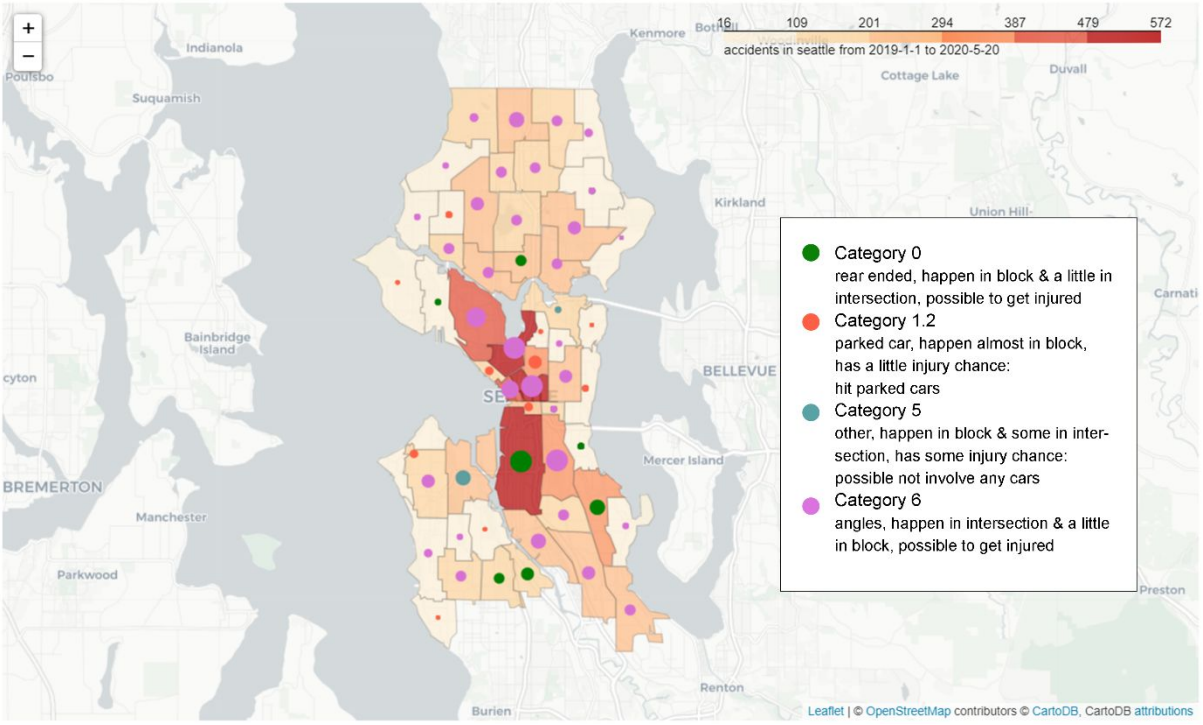

Figure 10

5. Future Studies

In the correlation research, I calculated correlation between each other using Pearson's correlation coefficient, indifferent if categorical, continuous or binary attributes. As a matter of fact, this method works better for continuous or binary attributes than categorical ones. A more accurate research should conduct Cramér's V for categorical attributes, and the correlation matrix may present differently and impact the partitioning to some extent. Moreover, I applied the segmentation in only one type of clustering in this research. Further research may conduct different clustering models with the same input and compare their results

6. Reference

[1] Collisions - Seattle GeoData - ArcGIS Online: https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

[2] Seattle Neighbourhoods Geo-JSON: https://opendata.arcgis.com/datasets/fbf6ca85b6b0408da346c8896b6f8aef_0.geojson

[3] Correlation and dependence, Wikipedia: https://en.wikipedia.org/wiki/Correlation_and_dependence

[4] The Search for Categorical Correlation, Shaked Zychlinski, 2018-02: https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9

[5] 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/

[6] k-means clustering, Wikipedia: https://en.wikipedia.org/wiki/K-means_clustering

[7] 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