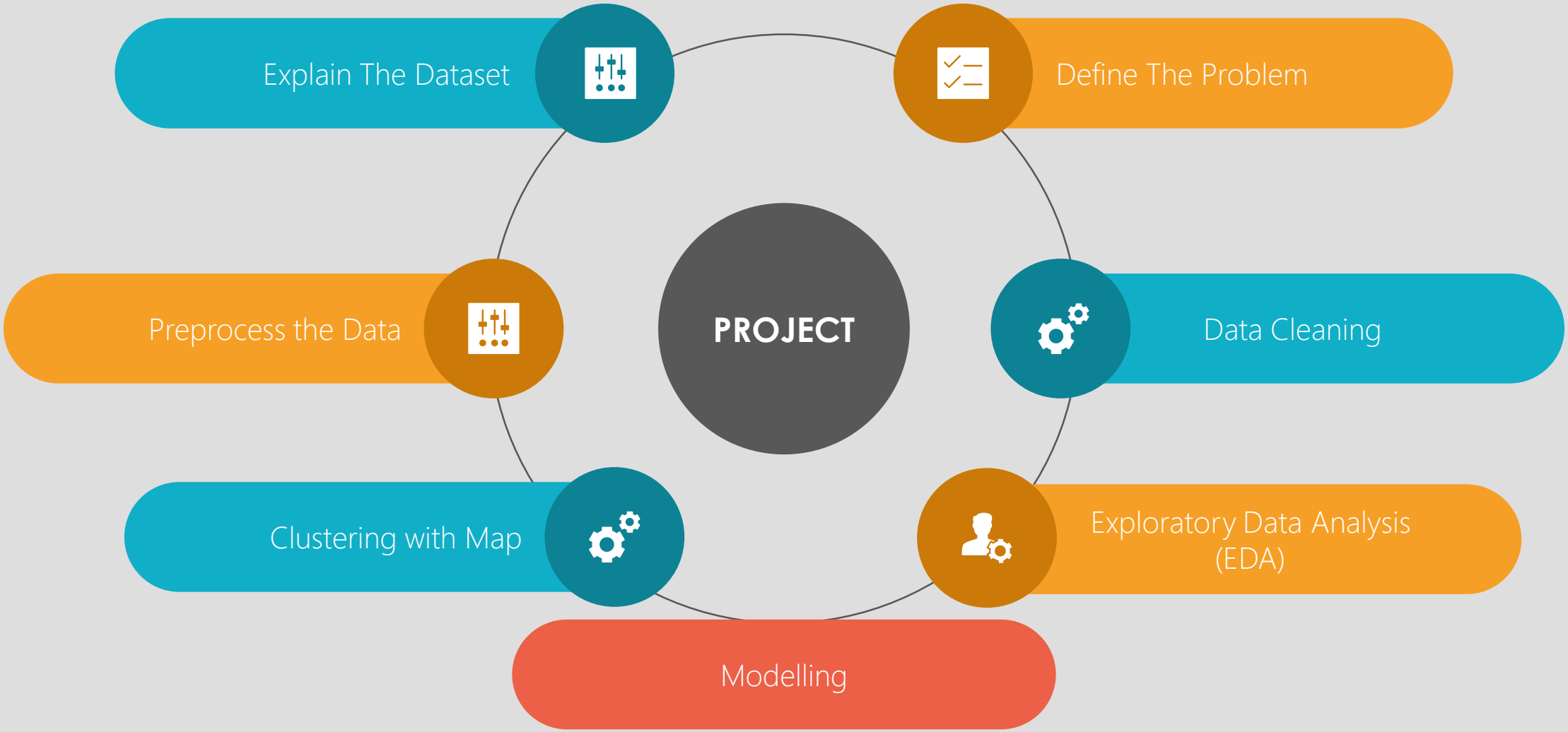
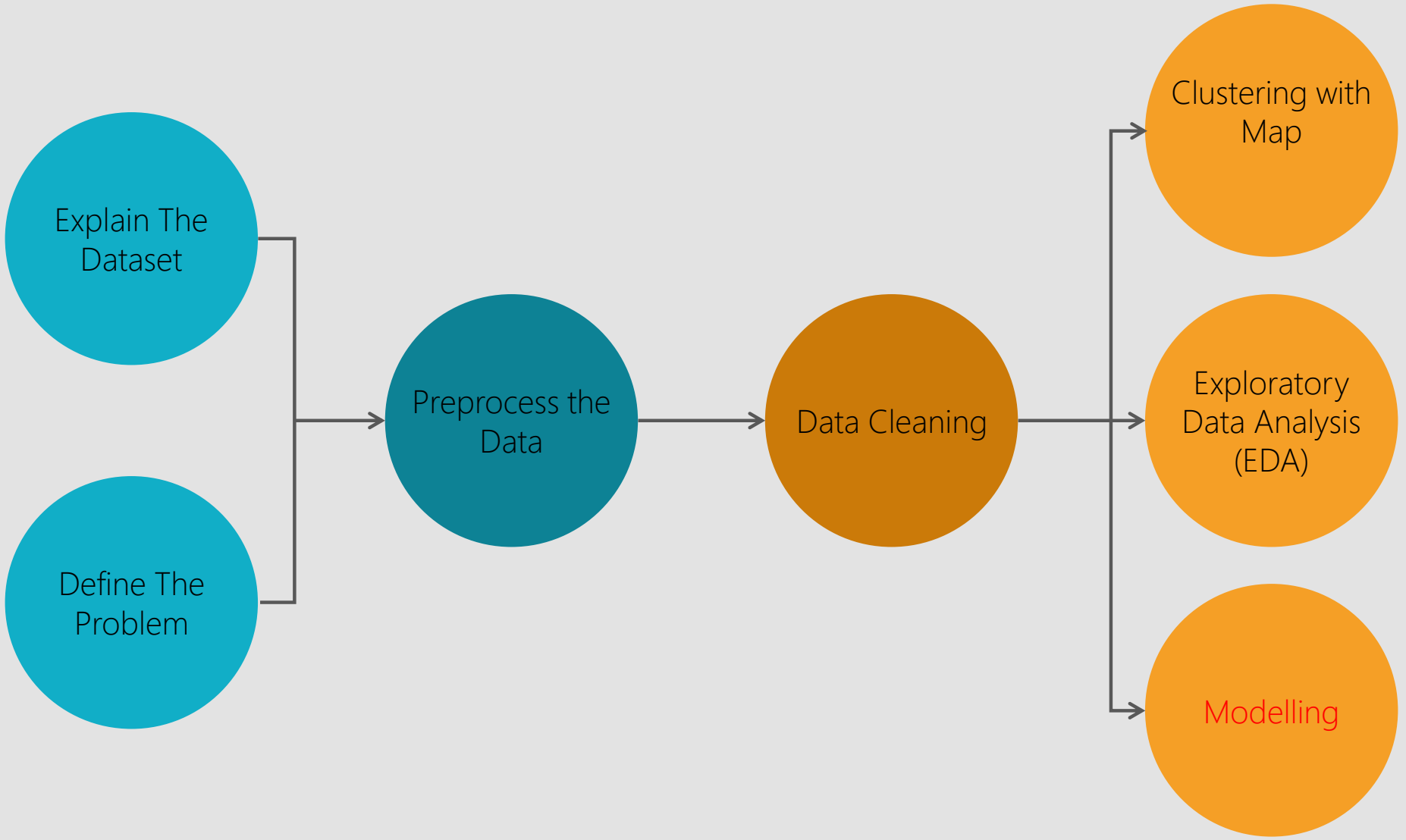
NY City Vehicle Analysis

Statistical Learning Presentation

Project Analysis



Project Analysis



Explain The Dataset

The Motor Vehicle Collisions crash table contains details on the crash event.
The Motor Vehicle Collisions data tables contain information from all police-reported motor vehicle collisions in NYC.
The police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage

A brief summary of data with *head()* function

	ACCIDENT.DATE	ACCIDENT.TIME	BOROUGH	ZIP.CODE	LATITUDE	LONGITUDE	LOCATION	ON.STREET.NAME	CROSS.STREET.NAME
2019-08-05T00:00:00.000		16:30	QUEENS	11434	40.67605	-73.79018	{'type': 'Point', 'coordinates': [-73.790184, 40.676052]}		
2019-08-27T00:00:00.000		16:02	BROOKLYN	11225	40.65778	-73.95110	{'type': 'Point', 'coordinates': [-73.951096, 40.65778]}		
2019-08-15T00:00:00.000		17:57	MANHATTAN	10002	40.71814	-73.99384	{'type': 'Point', 'coordinates': [-73.993835, 40.718143]}	CHRYSTIE STREET	GRAND STREET
2019-08-30T00:00:00.000		21:53	BRONX	10460	40.84053	-73.86661	{'type': 'Point', 'coordinates': [-73.86661, 40.840534]}		
2019-08-06T00:00:00.000		9:45	MANHATTAN	10016	40.74544	-73.97540	{'type': 'Point', 'coordinates': [-73.9754, 40.74544]}	EAST 35 STREET	2 AVENUE
2019-08-29T00:00:00.000		12:28	BROOKLYN	11222	40.72652	-73.94639	{'type': 'Point', 'coordinates': [-73.94639, 40.726524]}		

	OFF.STREET.NAME	NUMBER.OF.PERSONS.INJURED	NUMBER.OF.PERSONS.KILLED	NUMBER.OF.PEDESTRIANS.INJURED	NUMBER.OF.PEDESTRIANS.KILLED	NUMBER.OF.CYCLIST.INJURED
150-08	123 AVENUE	0	0	0	0	0
288	HAWTHORNE STREET	0	0	0	0	0
		1	0	0	0	0
1837	EAST TREMONT AVENUE	0	0	0	0	0
		1	0	0	0	1
38	JEWEL STREET	1	0	0	0	1

	NUMBER.OF.CYCLIST.KILLED	NUMBER.OF.MOTORIST.INJURED	NUMBER.OF.MOTORIST.KILLED	CONTRIBUTING.FACTOR.VEHICLE.1	CONTRIBUTING.FACTOR.VEHICLE.2	CONTRIBUTING.FACTOR.VEHICLE.3
	0	0	0	Passing Too Closely	Unspecified	
	0	0	0	Passing Too Closely	Unspecified	
	0	1	0	Driver Inattention/Distraction		
	0	0	0	Unspecified	Unspecified	
	0	0	0	Driver Inattention/Distraction	Driver Inattention/Distraction	
	0	0	0	Unspecified	Unspecified	

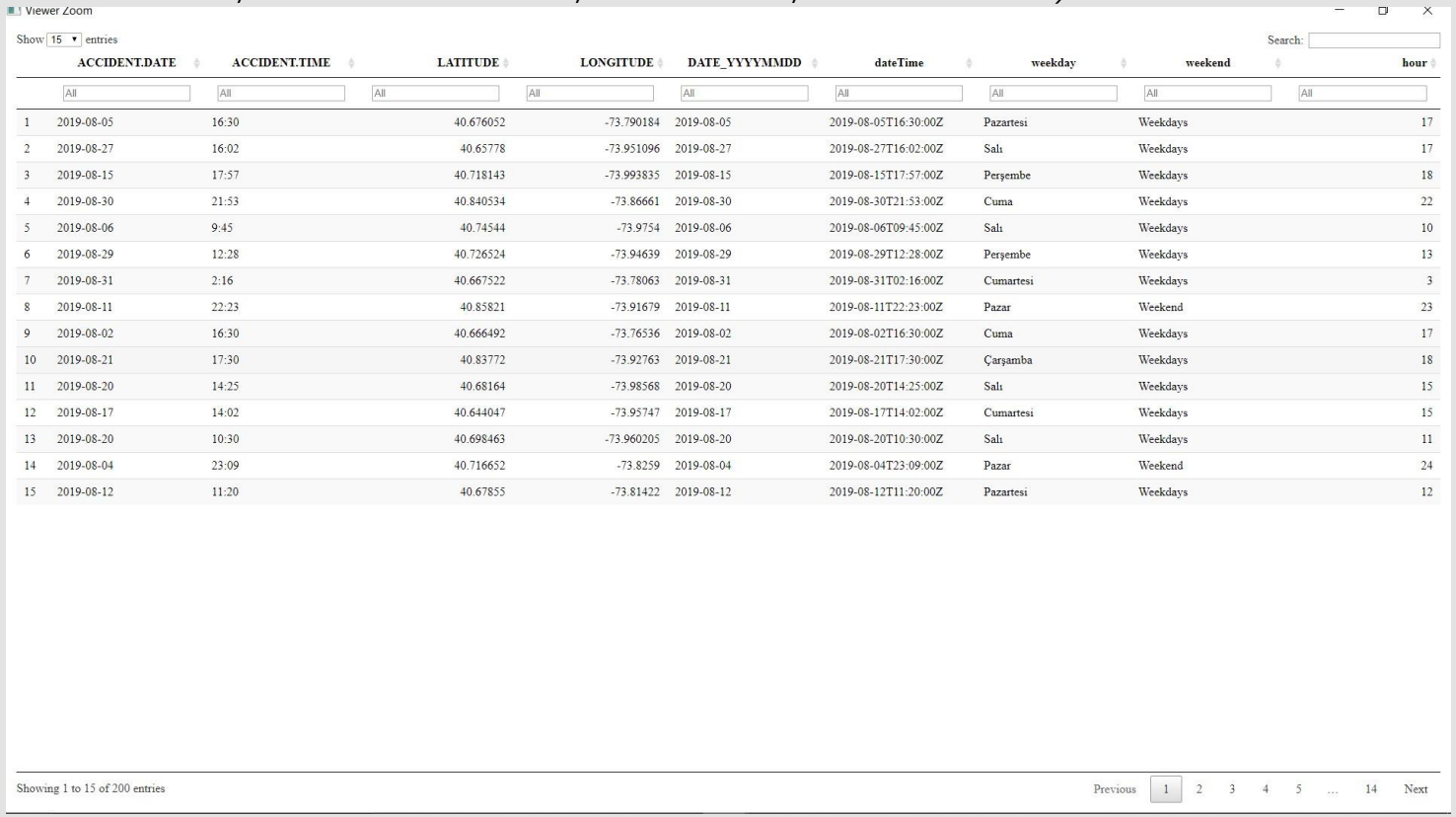
	CONTRIBUTING.FACTOR.VEHICLE.4	CONTRIBUTING.FACTOR.VEHICLE.5	COLLISION_ID	VEHICLE.TYPE.CODE.1	VEHICLE.TYPE.CODE.2	VEHICLE.TYPE.CODE.3	VEHICLE.TYPE.CODE.4
			4184637	Sedan	Pick-up Truck		
			4195773	Station Wagon/Sport Utility Vehicle	Station Wagon/Sport Utility Vehicle		
			4202457	Sedan			
			4198749	Taxi	Station Wagon/Sport Utility Vehicle		
			4183798	Station Wagon/Sport Utility Vehicle	Bike		
			4196772	Station Wagon/Sport Utility Vehicle	Bike		

	VEHICLE.TYPE.CODE.5

Objectives

1. **The Problem** : Big amount of accidents on daily basis
2. Download data on the kaggle
3. Import the data to R Studio
4. Visualize accidents and locations.
5. Select the variables in the dataset from the dataset as required. (date/location/time etc...)
6. **Solution** : Estimate the potential traffic accident area based on current locations.

With *data <- data_try %>% select(ACCIDENT.DATE, ACCIDENT.TIME, LATITUDE, LONGITUDE)*
We select our variables



Viewer Zoom

Show 15 entries Search:

	ACCIDENT.DATE	ACCIDENT.TIME	LATITUDE	LONGITUDE	DATE_YYYYMMDD	dateTime	weekday	weekend	hour
	All	All	All	All	All	All	All	All	All
1	2019-08-05	16:30	40.676052	-73.790184	2019-08-05	2019-08-05T16:30:00Z	Pazartesi	Weekdays	17
2	2019-08-27	16:02	40.65778	-73.951096	2019-08-27	2019-08-27T16:02:00Z	Salı	Weekdays	17
3	2019-08-15	17:57	40.718143	-73.993835	2019-08-15	2019-08-15T17:57:00Z	Perşembe	Weekdays	18
4	2019-08-30	21:53	40.840534	-73.86661	2019-08-30	2019-08-30T21:53:00Z	Cuma	Weekdays	22
5	2019-08-06	9:45	40.74544	-73.9754	2019-08-06	2019-08-06T09:45:00Z	Salı	Weekdays	10
6	2019-08-29	12:28	40.726524	-73.94639	2019-08-29	2019-08-29T12:28:00Z	Perşembe	Weekdays	13
7	2019-08-31	2:16	40.667522	-73.78063	2019-08-31	2019-08-31T02:16:00Z	Cumartesi	Weekdays	3
8	2019-08-11	22:23	40.85821	-73.91679	2019-08-11	2019-08-11T22:23:00Z	Pazar	Weekend	23
9	2019-08-02	16:30	40.666492	-73.76536	2019-08-02	2019-08-02T16:30:00Z	Cuma	Weekdays	17
10	2019-08-21	17:30	40.83772	-73.92763	2019-08-21	2019-08-21T17:30:00Z	Çarşamba	Weekdays	18
11	2019-08-20	14:25	40.68164	-73.98568	2019-08-20	2019-08-20T14:25:00Z	Salı	Weekdays	15
12	2019-08-17	14:02	40.644047	-73.95747	2019-08-17	2019-08-17T14:02:00Z	Cumartesi	Weekdays	15
13	2019-08-20	10:30	40.698463	-73.960205	2019-08-20	2019-08-20T10:30:00Z	Salı	Weekdays	11
14	2019-08-04	23:09	40.716652	-73.8259	2019-08-04	2019-08-04T23:09:00Z	Pazar	Weekend	24
15	2019-08-12	11:20	40.67855	-73.81422	2019-08-12	2019-08-12T11:20:00Z	Pazartesi	Weekdays	12

Showing 1 to 15 of 200 entries Previous 1 2 3 4 5 … 14 Next

Summary of The Data

```
> data %>% summary()
 ACCIDENT.DATE      ACCIDENT.TIME        LATITUDE        LONGITUDE
 Length:1612178     Length:1612178     Min.   : 0.00   Min.   :-201.24
 Class :character   Class :character   1st Qu.:40.67   1st Qu.: -73.98
 Mode  :character   Mode  :character   Median :40.72   Median : -73.93
                                       Mean   :40.69   Mean   : -73.87
                                       3rd Qu.:40.77   3rd Qu.: -73.87
                                       Max.   :42.32   Max.   :   0.00
                                       NA's   :196285  NA's   :196285
```

Data Cleaning

In fact, we wanted to do a data manipulation here because the location variables in the data were too large.
For this we would use knn, random forest to do an undefined value tuning.
However, this process takes too long (because we do not have enough processing power). Here is a simple data editing

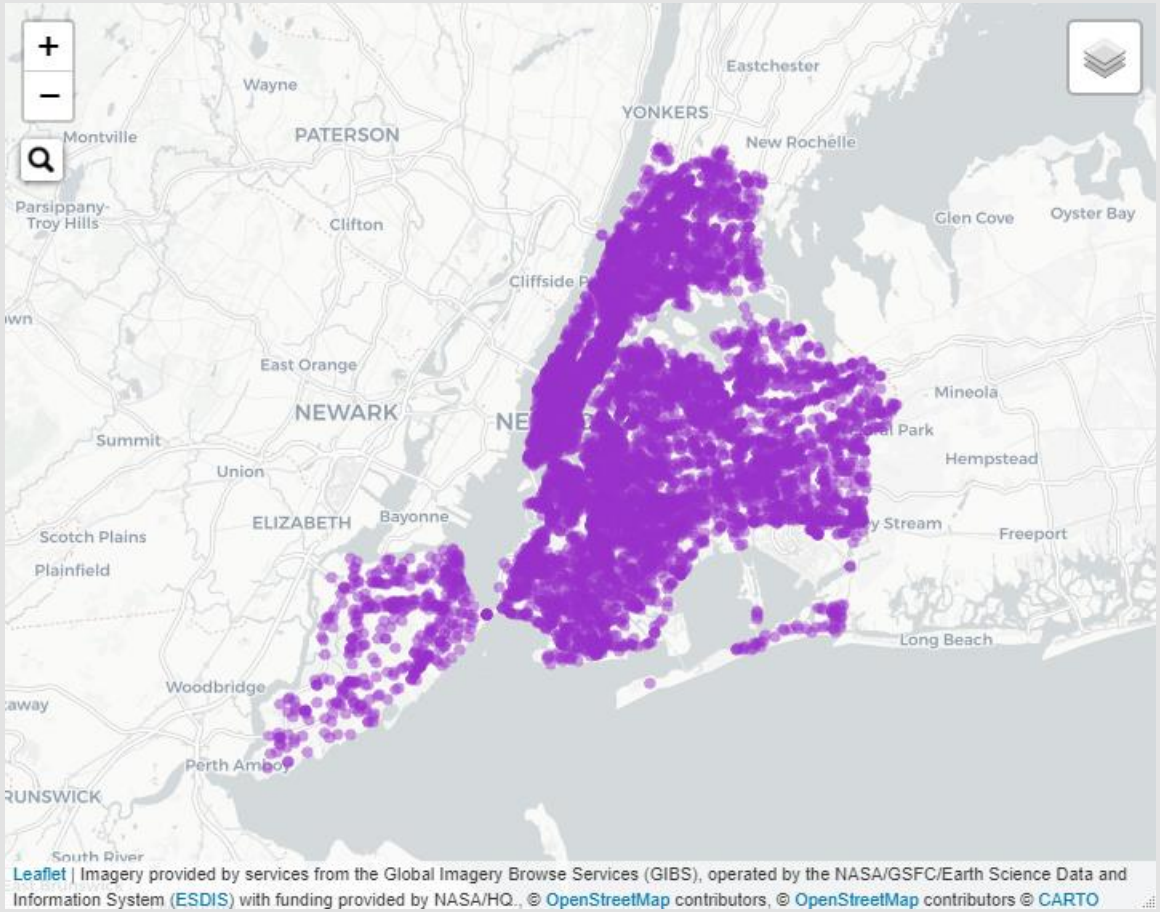
```
### Fill with Random Forest
library(missForest)
summary(na_data)
rf_source <- na_data %>%
  select(LATITUDE, LONGITUDE,NUMBER.OF.PERSONS.INJURED,NUMBER.OF.PERSONS.KILLED)

rf_data <- missForest(rf_source, ntree = 5)

l <- sapply(na_data, function(x) which(is.na(x)))
cln_na_data[c(l$LATITUDE),]$LATITUDE
cln_na_data[c(l$LONGITUDE),]$LONGITUDE

rf_data <- rf_data$ximp
rf_data[c(l$LATITUDE),]$LATITUDE
rf_data[c(l$LONGITUDE),]$LONGITUDE

mean(cln_na_data[c(l$LATITUDE),]$LATITUDE - rf_data[c(l$LATITUDE),]$LATITUDE)
summary(cln_na_data$LATITUDE)

mean(cln_na_data[c(l$LONGITUDE),]$LONGITUDE - rf_data[c(l$LONGITUDE),]$LONGITUDE)
summary(cln_na_data$LONGITUDE)
```

data <- data %>% filter(LATITUDE>0, LONGITUDE<-72, LONGITUDE>-75)

Create a Map

When I look at the summary results,

We took the map on the right and it with 7500 samples.

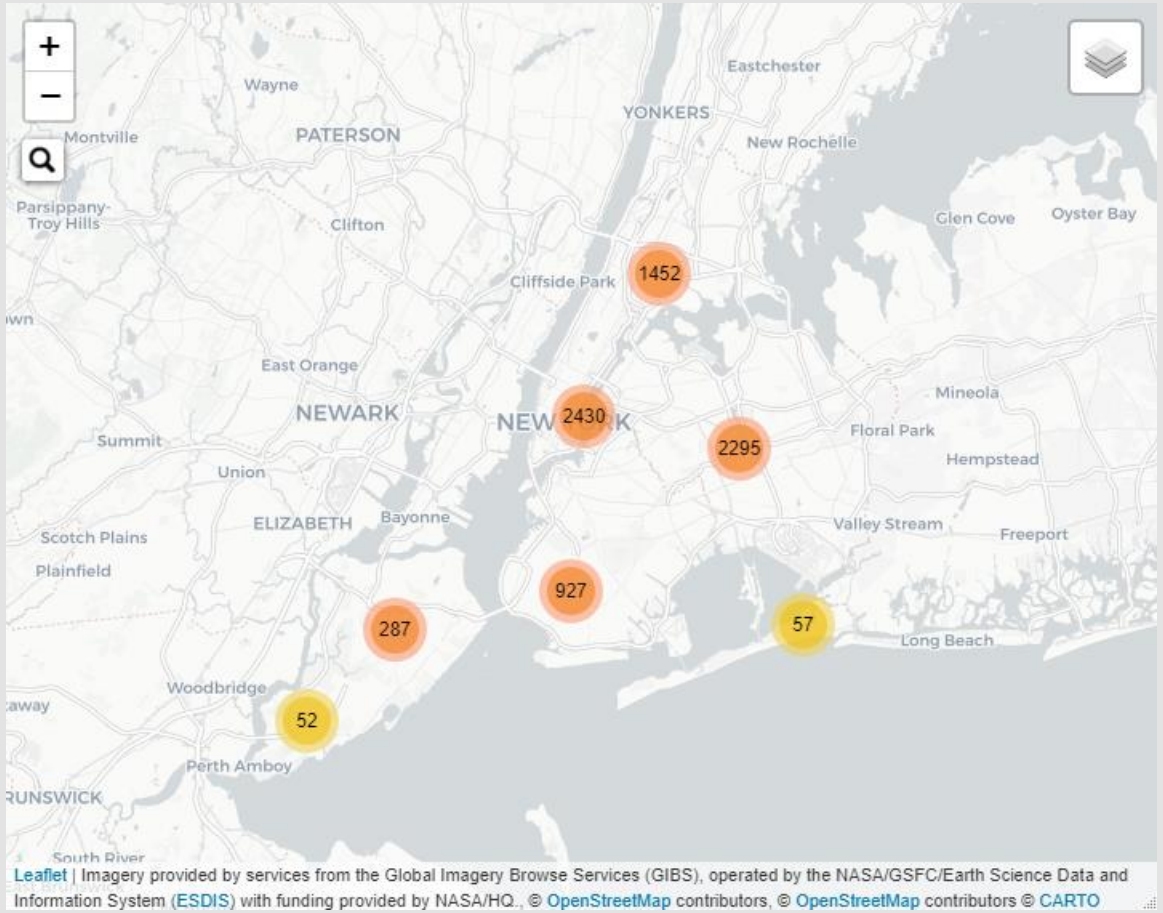


Map with Clustering

Thanks to that piece of code

addCircleMarkers(~LONGITUDE, ~LATITUDE, radius = 1,
color = "brown", fillOpacity = 0.005,
clusterOptions = markerClusterOptions()) %>%

We can organize our maps data.



Making and Converting

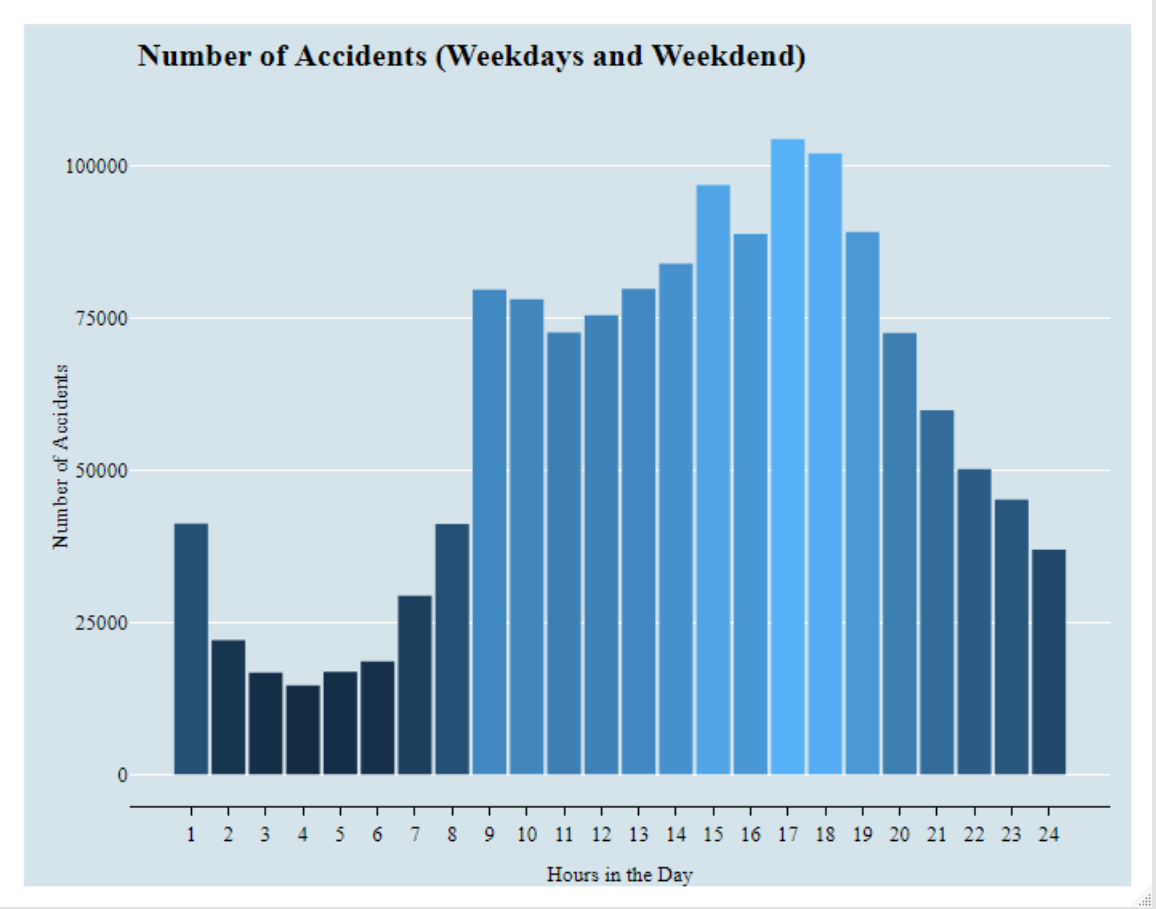
- We converted date time to time series data and created variables such as hour, weekday, weekend, etc.
- Hour is denoting 24 hours a day.
- However, we had to write the days in Turkish because the system language of the computer we applied was Turkish.

So: Saturday = Cumartesi, Sunday = Pazar,

```
library(lubridate)
format(seq.Date(as.Date('2000-01-01'), by = 'month', len = 12), "%b")
1] "Oca" "Şub" "Mar" "Nis" "May" "Haz" "Tem" "Ağu" "Eyl" "Eki" "Kas" "Ara"
data <- data %>%
  mutate(DATE_YYYYMMDD = str_extract(data$ACCIDENT.DATE, "\\d{4}-\\d{2}-\\d{2}"))
data$ACCIDENT.DATE <- data$DATE_YYYYMMDD
data <- data %>%
        mutate(dateTime = ymd_hm(paste(ACCIDENT.DATE, ACCIDENT.TIME, sep = ' ')),
        weekday = as.factor(weekdays(dateTime)),
        weekend = if_else(weekday == 'Cunartesi'|weekday=='Pazar','Weekend','Weekdays'),
        hour = hour(dateTime)+1)
```

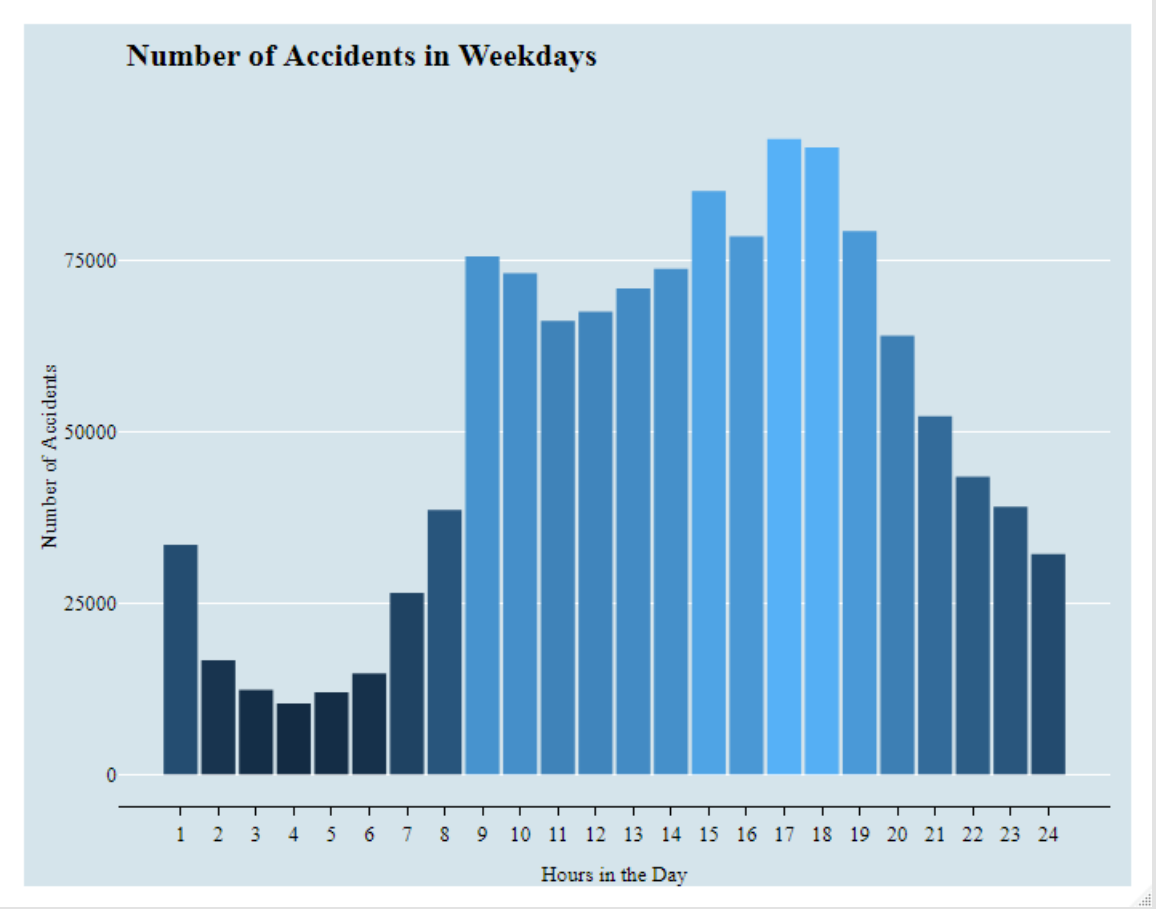
Exploratory Data Analysis (EDA)

- When we look at the number of accidents according to the time schedule of the day, there is an increase in accidents almost regularly.
- However, we can see that this increase ended at 17 ~ 18 hours.



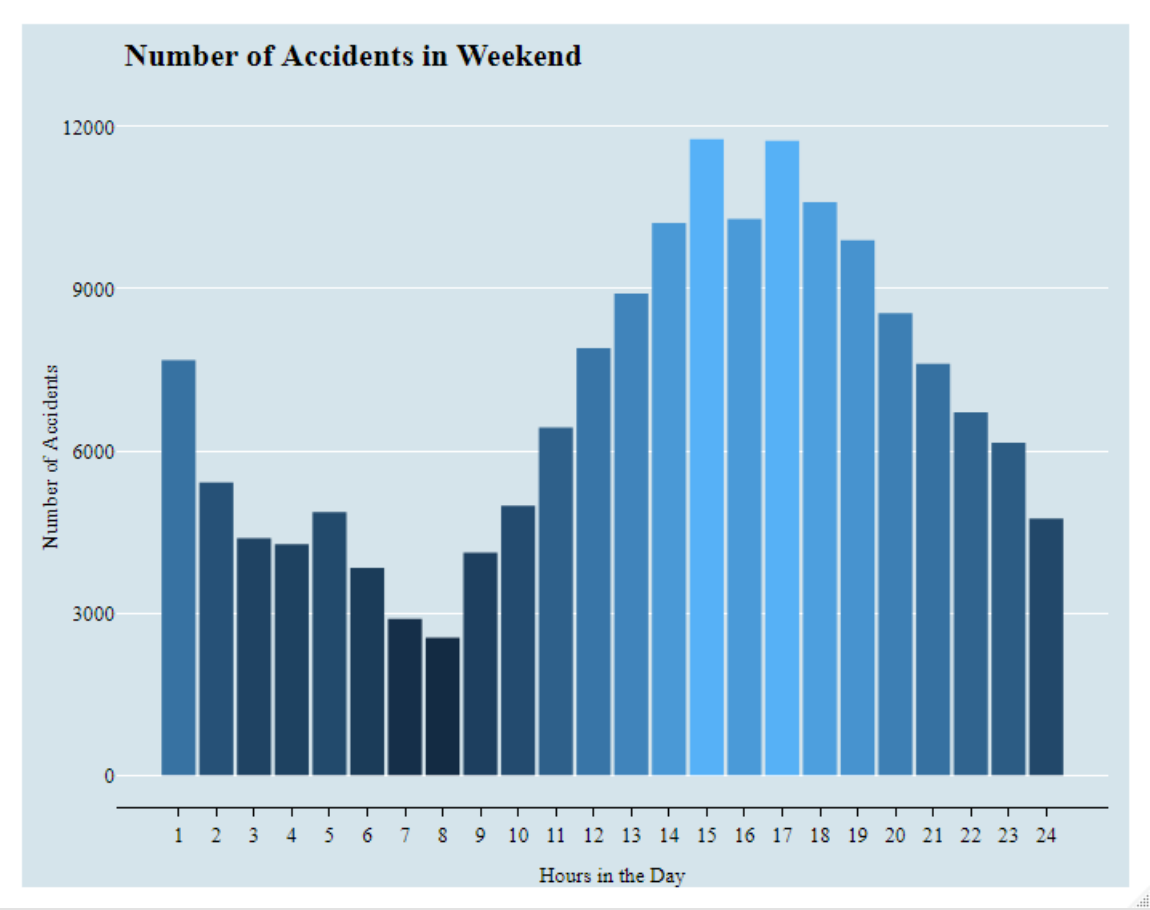
Accident in Weekdays

- It is the same as the observations from the data set, which is slightly higher in the morning.



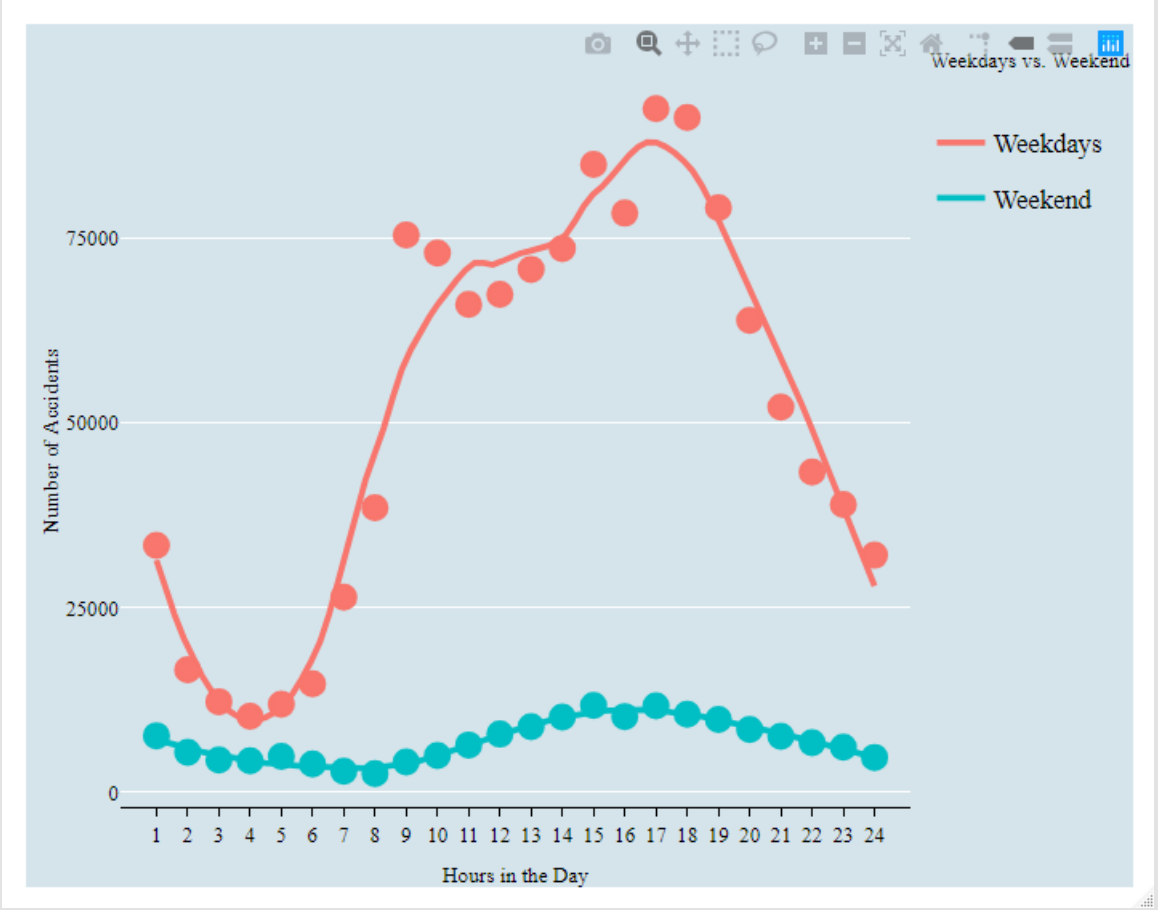
Accident in Weekends

- For the weekend, the graph has changed in appearance, and the peak takes place between 15 and 17 hours.



Weekdays vs Weekend

- If we compare the weekdays and weekend tables, we can say that people drive less risky vehicles on weekends.



Top 10 Dangerous Areas in Weekdays

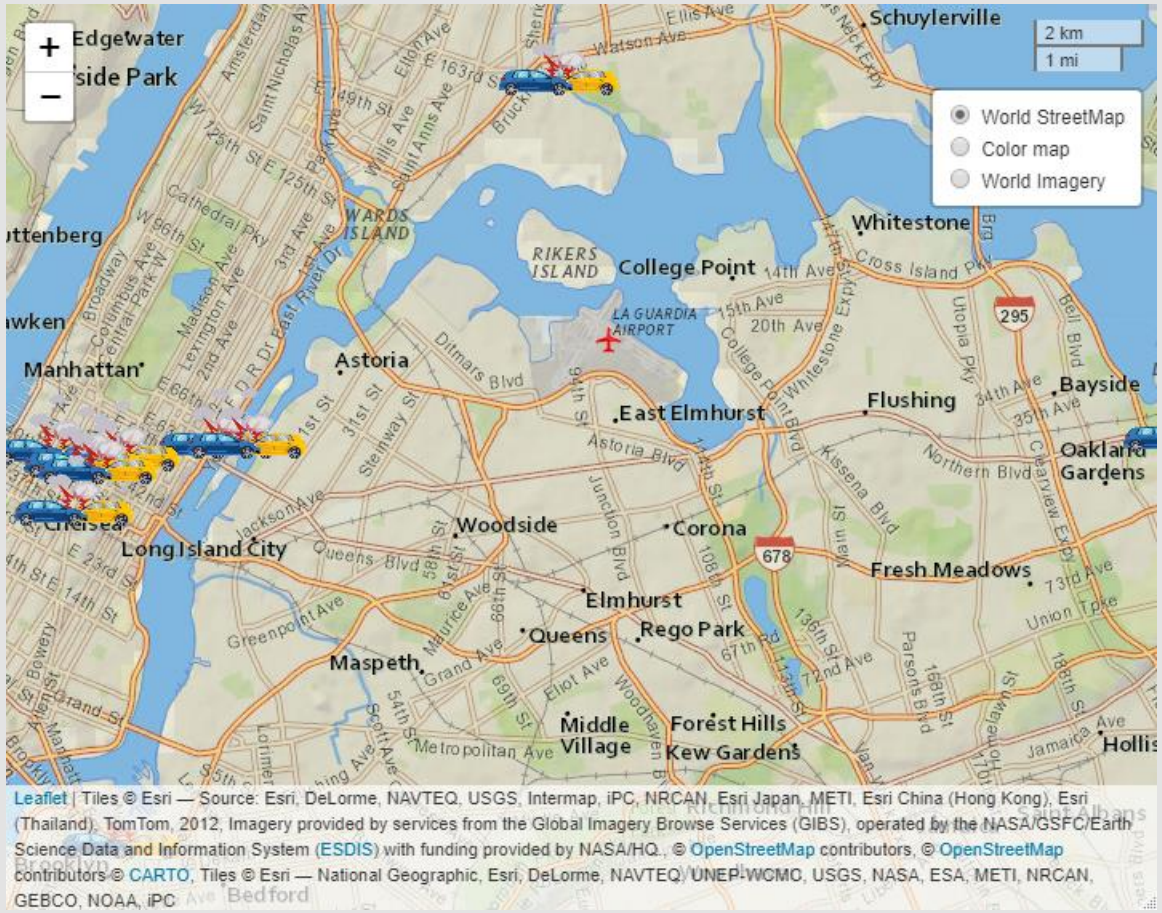
- Instead of directing the calculation of the top 10 accident locations, we processed the data a bit.

- If We use direct longitude and latitude data, the same collection point with slightly different coordinates is considered to be different collection locations and will definitely deviate from the actual result.

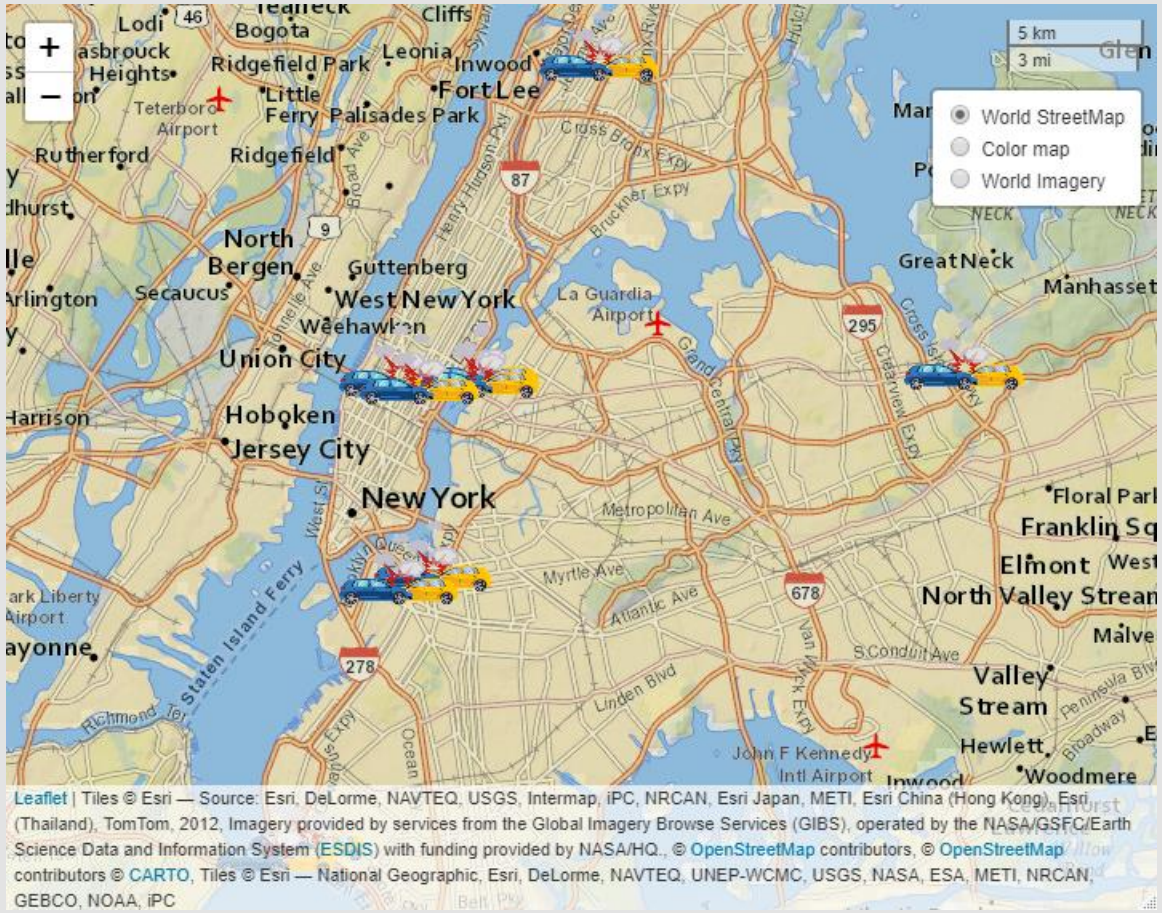
- Therefore, I rounded up to 3 decimal places without making a huge change to longitude and latitude. We also uploaded a picture to Ogulcan's blog to indicate the accident sites and used that picture as an icon. The graphic is interactive and can zoom in and out. If you place the mouse over the crashed car icon, it shows how many accidents are in the position according to the data set.

Top 10 Dangerous Areas in Weekends

- The locations of the vehicles, namely the regions where the most accidents occur, change directly on weekends and weekdays.



WEEKDAYS



WEEKEND

Modelling

Modelling

After examining all the accidents, we prepared a simple kmeans model to divide the accident points into 50 groups to suggest a potential accident point. With unsupervised learning, we got a simple but functional result.

```r
data_coordinates <- data %>% select(LONGITUDE, LATITUDE)
data_fm <- data
```

```r
set.seed(0)
data_kmeans <- data_coordinates %>%
  kmeans(50, nstart=20)
save(data_kmeans, file = "input/kmeans_data.rda")
load("input/kmeans_data.rda")

data_fm$cluster <- data_kmeans$cluster


colors_in <- colorNumeric(
  palette = "magma",
  domain = data$cluster)
```

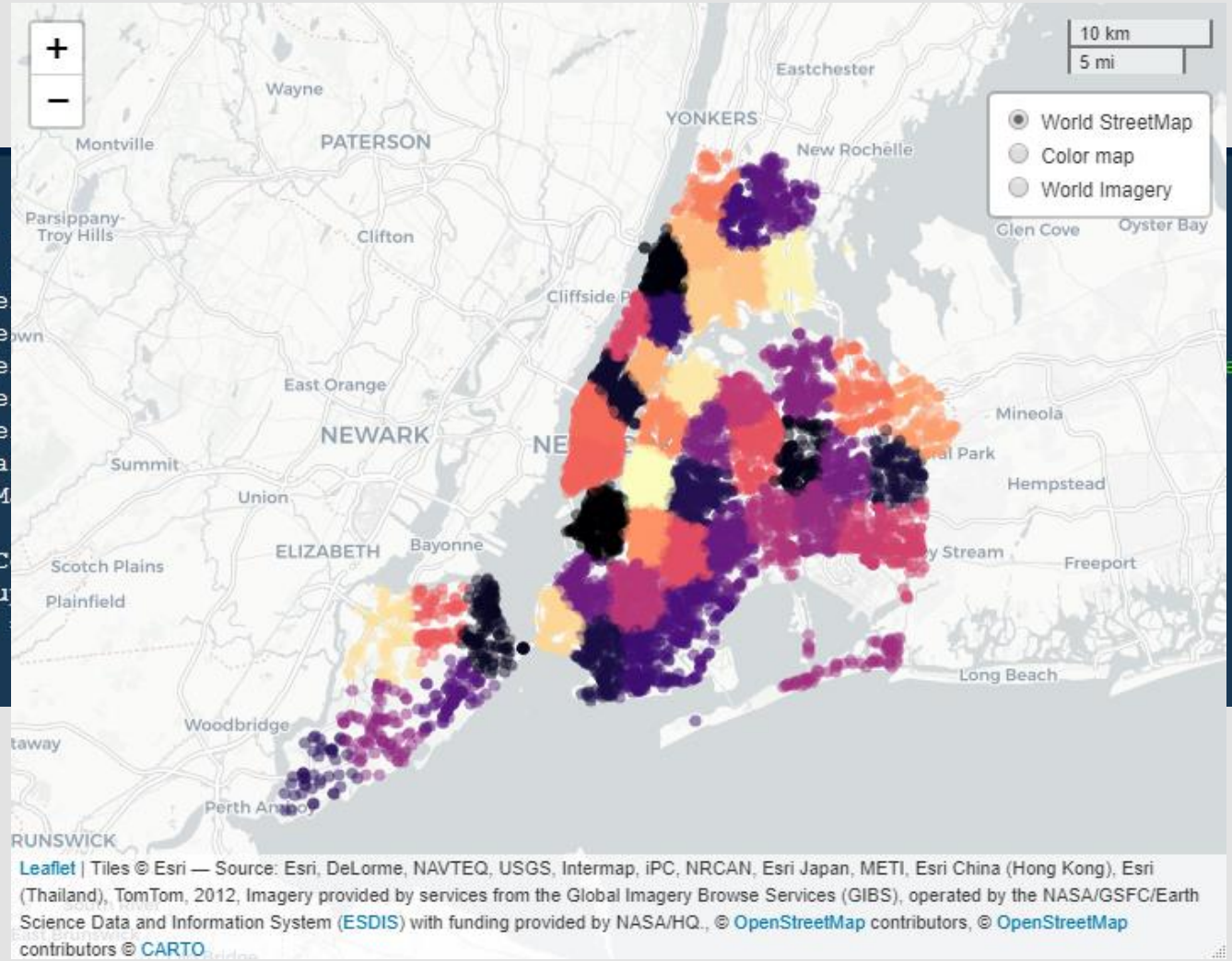
Modelling

New Map

Finally, we created a new car accident map using the kmeans model.

```
```{r}
set.seed(1)
data_fm %>%
  leaflet()
  addProvide
  addProvide
  addProvide
  addProvide
  addProvide
  addScaleBa
  addCircleM

  addLayersC
    baseGrou
    options
  )
```
```



Modelling

New Top 10 Estimated Dangerous Area Map on Weekdays

Modelling

New Top 10 Estimated Dangerous Area Map

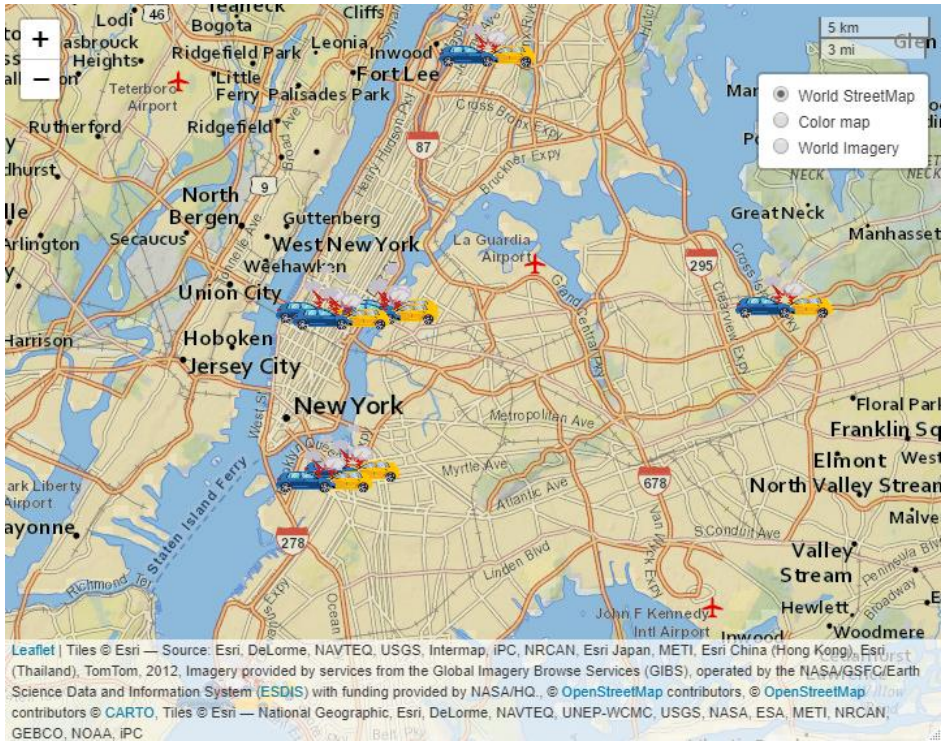
WEEKENDS

WEEKDAYS

Modelling

Estimated Dangerous Area Map on Weekends vs Dangerous Area from data

ESTIMATED



ORIGINAL

References

1. https://www.kaggle.com/new-york-city/nypd-motor-vehicle-collisions
2. https://stackoverflow.com/questions/21382681/kmeans-quick-transfer-stage-steps-exceeded-maximum
3. https://github.com/tidyverse/lubridate/issues/669
4. http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf
5. https://stackoverflow.com/questions/49951416/how-to-use-colornumeric-within-addcircles-in-leaflet
6. https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb
7. https://medium.com/@mueller.johannes.j/use-r-and-gganimate-to-make-an-animated-map-of-european-students-and-their-year-abroad-517ad75dca06

Thank You

Ogulcan Ertunc