P3 Data Enrich/Format/Blend (Revised)

Team 1: Adrita Anika, Anantha Sai Sreelekha, Soham Sarda

2022-03-25

We have four datasets: Collision, Person, Vehicles and Weather datsets. The following steps are taken for enrichment, blending and blending the datasets.

Enrichment and Formatting:

1. Creating time-based features: We have one variable named as "CRASH_DATE' which contains the date, month and year information. From that, Crash Year', 'Crash Month', 'Crash Day', and 'Time of day' variables were derived. These variables will be used to visualize the temporal distribution of crashes and look for patterns, if any.

```
library(lubridate)
crash_data <- read.csv("NYC_crashes.csv")
crash_data$CRASH_DAY <- day(crash_data$CRASH_DATE)
crash_data$CRASH_MONTH <- month(crash_data$CRASH_DATE)
crash_data$CRASH_YEAR <- year(crash_data$CRASH_DATE)
```

2. We want to see patterns based on weekdays (Sat, Sun,.. etc). Hence, we enhance the data with another variable containing the weekday information.

```
crash_data$CRASH_WDAY <- weekdays(as.Date(crash_data$CRASH_DATE))
```

3. For time analysis, we divided 24 hours of the day time into 12 timeslots. Like: 00:00-01:59, 2:00-03:59,. . . , 22:00-23:59. Later we represented this with numbers starting from 0 to 11 as shown later in this report.

4. We have the geographic information: longitude and latitude and zip code. We will consider the zip codes as location IDs in our analysis and place a marker on the map for each zip code, so we need to fix a longitude and latitude value for the marker to be placed for each zip code. We considered the average of all available longitudes under a zip code for placing a marker on the map. Same thing is done for latitude as well. This data is required for spatial analysis, we save this as "cluster_data.csv'.

```
library(dplyr)
cluster_data <- crash_data %>% group_by(ZIP_CODE) %>% summarise(LAT = mean(LATITUDE),
                                                                LNG = mean(LONGITUDE),
                                                                Crash = n(),)
```

5. All the column headers are once again formatted with a consistency followed - Upper case letters with underscores between words.

6. For our analysis we created another variable named as "VEHICLE_TYPE" for the vehicles dataset. We categorized all the vehicles depending on their sizes into four categories: very_large, large, medium, small. This has been done following this paper [Paper Link]

7. The weather dataset does not have any common variable apart from geodetic reference. The Lat-Long coordinates were used to find the nearest weather station to each crash activity, and was then the weather data for that station on the particular day of crash was linked to the crash data.

8. We are considering a classification problem of traffic collision severity based on ten features: emotional status, driver license status, vehicle type, two contributing factors, position in vehicle, safety equipment, day and time. We converted some of these variables into numerical values with one hot encoding. This is done only for the classification problem. Following code chunk shows the one-hot encoding process for only one variable: position in vehicle.

```
perX <- read.csv("NYC_persons.csv")
per <- perX[, c("COLLISION_ID","POSITION_IN_VEHICLE")]
per <- filter(per, per$POSITION_IN_VEHICLE!="Unknown")
library(caret)
#define one-hot encoding function
dummy <- dummyVars(" ~ .", data=per)
#perform one-hot encoding on data frame
final_df <- data.frame(predict(dummy, newdata=per))
```

Merging/Blending:

1. All of the files will be stored in csv format. As we have huge datasets, we created seperate csv files for some of the analysis. For the classification of Traffic collision based on severity into three classes (No hurt, Injury, Lethal) we will use ten attributes from the three datasets: collision, person and vehicle. All these datasets contain one common attribute "COLLISION ID". We inner joined these datasets and kept the ten required features. First few rows of the dataset is shown below. The result is presented after doing one hot encoding.

```
clf_data <- read.csv("dfx.csv")
head(clf_data, 3)
```

```
##   X CONTRIBUTING_FACTOR_VEHICLE_1 EMOTIONAL_STATUS POSITION_IN_VEHICLE
## 1 1                              15                7                   4
## 2 2                              15                5                   2
## 3 3                              15                5                   2
##   SAFETY_EQUIPMENT CRASH_TIME VEHICLE_TYPE DRIVER_SEX DRIVER_LICENSE_STATUS
## 1                8      22:20          178          2                     1
## 2                2      22:20          178          1                     0
## 3                2      22:20          178          2                     1
##   CONTRIBUTING_FACTOR_1 Day label
## 1                    28   2     1
## 2                    15   2     1
## 3                    28   2     1
```

2. As mentionied earlier, "cluster_data.csv" wil be used for spatial analysis for finding collision hotspots.

```
head(cluster_data,3)
```

```
## # A tibble: 3 x 4
##   ZIP_CODE   LAT   LNG Crash
##      <int> <dbl> <dbl> <int>
## 1    10001  40.8 -74.0     1
## 2    10012  40.7 -74.0     1
## 3    10016  40.7 -74.0     3
```

3. For time analysis, we created another csv file that has the information of weekdays and timeslot of the day and corresponding number of collisions. We need this is data for two analysis (forecasting and visual analysis)

```
time_data <- read.csv("refined_time.csv")
head(time_data,3)
```

```
##      day new_time numCrash   time_slot
## 1 Friday        0      430 00:00-01:59
## 2 Friday        1      150 02:00-03:59
## 3 Friday        2      164 04:00-05:59
```

4. As mentioned earlier, we categoried vehicles into four groups. The dataset is shown below:

```
vehicle_data <- read.csv("veh.csv")
head(vehicle_data,3)
```

```
##        X  BOROUGH NUMBER_OF_PERSONS_INJURED NUMBER_OF_PERSONS_KILLED
## 1 126446 BROOKLYN                         0                        0
## 2 126447   QUEENS                         0                        0
## 3 126448 BROOKLYN                         0                        0
##   NUMBER_OF_PEDESTRIANS_INJURED NUMBER_OF_PEDESTRIANS_KILLED
## 1                             0                            0
## 2                             0                            0
## 3                             0                            0
##   NUMBER_OF_CYCLIST_INJURED NUMBER_OF_CYCLIST_KILLED NUMBER_OF_MOTORIST_INJURED
## 1                         0                       0                          0
## 2                         0                       0                          0
## 3                         0                       0                          0
##   NUMBER_OF_MOTORIST_KILLED                VEHICLE_TYPE_CODE_1   TYPE
## 1                         0 Station Wagon/Sport Utility Vehicle  small
## 2                         0                              Sedan  small
## 3                         0                       Pick-up Truck medium
```

5. Finally, our weather dataset is shown:

```
weather_data <- read.csv("weather.csv")
head(weather_data, 3)
```

```
##       X COLLISION_ID       DATE  TIME   BOROUGH    lat_x     lon_x
## 1 37250      4273965 2020-01-14 18:10    QUEENS 40.70177 -73.90070
## 2 69952      4128277 2019-05-08 14:24 MANHATTAN 40.76617 -73.95425
## 3 69963      4128012 2019-05-08 10:30 MANHATTAN 40.75931 -73.96530
##               LOCATION NUMBER_OF_PERSONS_INJURED NUMBER_OF_PERSONS_KILLED
## 1    40.70177, -73.9007                         0                        0
## 2 40.766167, -73.954254                         0                        0
## 3   40.759308, -73.9653                         0                        0
##   NUMBER_OF_CYCLIST_INJURED NUMBER_OF_CYCLIST_KILLED NUMBER_OF_MOTORIST_INJURED
## 1                         0                       0                          0
## 2                         0                       0                          0
## 3                         0                       0                          0
##   NUMBER_OF_MOTORIST_KILLED  CONTRIBUTING_FACTOR_VEHICLE_1 VEHICLE_TYPE_CODE_1
## 1                         0          Following Too Closely               Sedan
## 2                         0 Driver Inattention/Distraction               Sedan
## 3                         0 Driver Inattention/Distraction                 Bus
##   VEHICLE_TYPE_CODE_2 YEAR MONTH                        NAME PRCP SNOW SNWD
## 1               Sedan 2020     1 NY CITY CENTRAL PARK, NY US 0.07    0    0
## 2               Sedan 2019     5 NY CITY CENTRAL PARK, NY US 0.00    0    0
## 3           Box Truck 2019     5 NY CITY CENTRAL PARK, NY US 0.00    0    0
```