L.A. Traffic Collision Analysis

Presented to:
Dr. Daniel Soper
Department of Information Systems & Decision Sciences
Mihaylo College of Business and Economics
California State University, Fullerton
FALL 2019

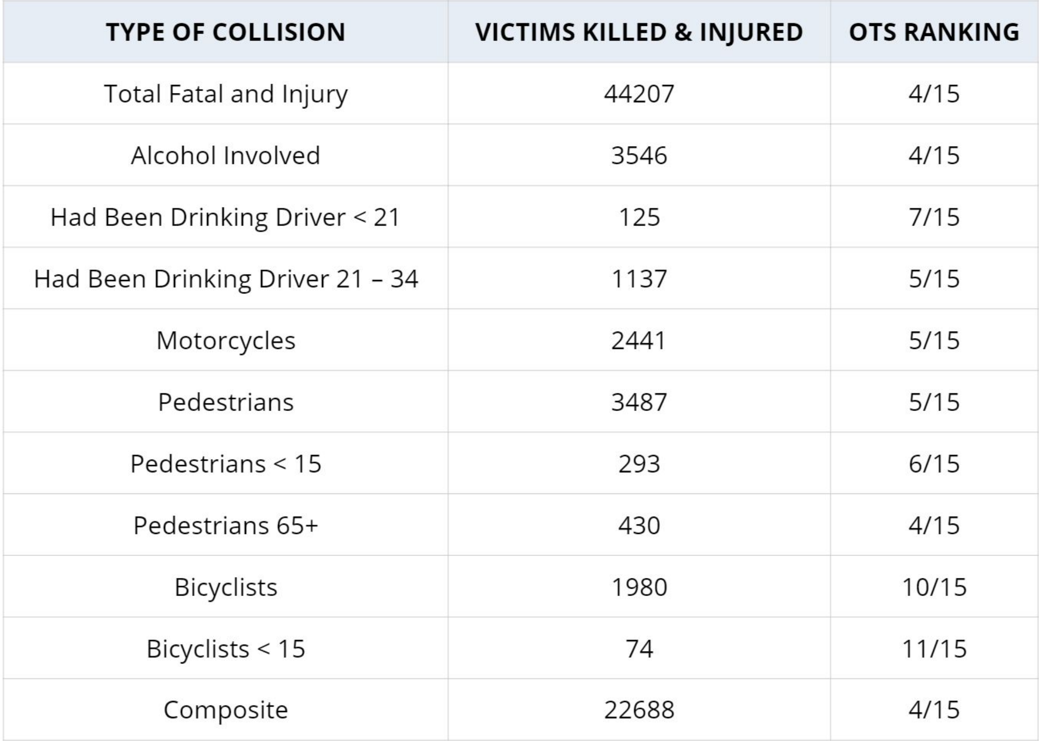
Team Members
Jung, Wone Eui
Marri, Manveetha Krishna
Nguyen, Chi
Nguyen, Vy
Saini, Priyanka

CALIFORNIA STATE UNIVERSITY FULLERTON™

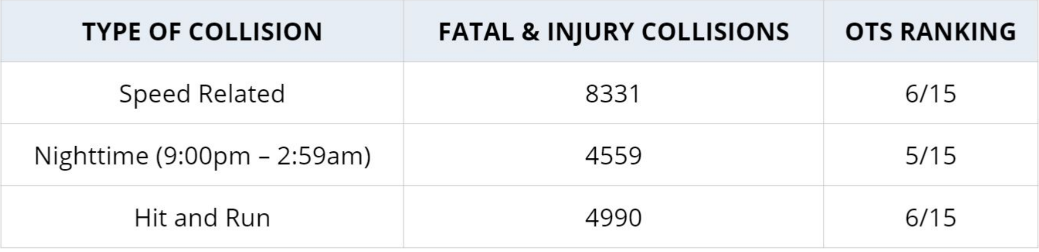
TABLE OF CONTENTS

I. Introduction

Car accidents happen every day with an alarmed increasing rate in California. According to the California Office of Traffic Safety (OTS), traffic fatalities increased 7% from 3,387 in 2015 to 3,623 in 2016. Below is a summary statistics for traffic collision in Los Angeles city in 2016 to give us some other perspectives:



TYPE OF COLLISION	VICTIMS KILLED & INJURED	OTS RANKING
Total Fatal and Injury	44207	4/15
Alcohol Involved	3546	4/15
Had Been Drinking Driver < 21	125	7/15
Had Been Drinking Driver 21 – 34	1137	5/15
Motorcycles	2441	5/15
Pedestrians	3487	5/15
Pedestrians < 15	293	6/15
Pedestrians 65+	430	4/15
Bicyclists	1980	10/15
Bicyclists < 15	74	11/15
Composite	22688	4/15



TYPE OF COLLISION	FATAL & INJURY COLLISIONS	OTS RANKING
Speed Related	8331	6/15
Nighttime (9:00pm – 2:59am)	4559	5/15
Hit and Run	4990	6/15

We want to provide a solution for LAPD (Los Angeles Police Department) and OTS to help create a safer city with fewer traffic collisions.

II. Questions of Interests and Descriptions of Variables

1. The Questions of Interests

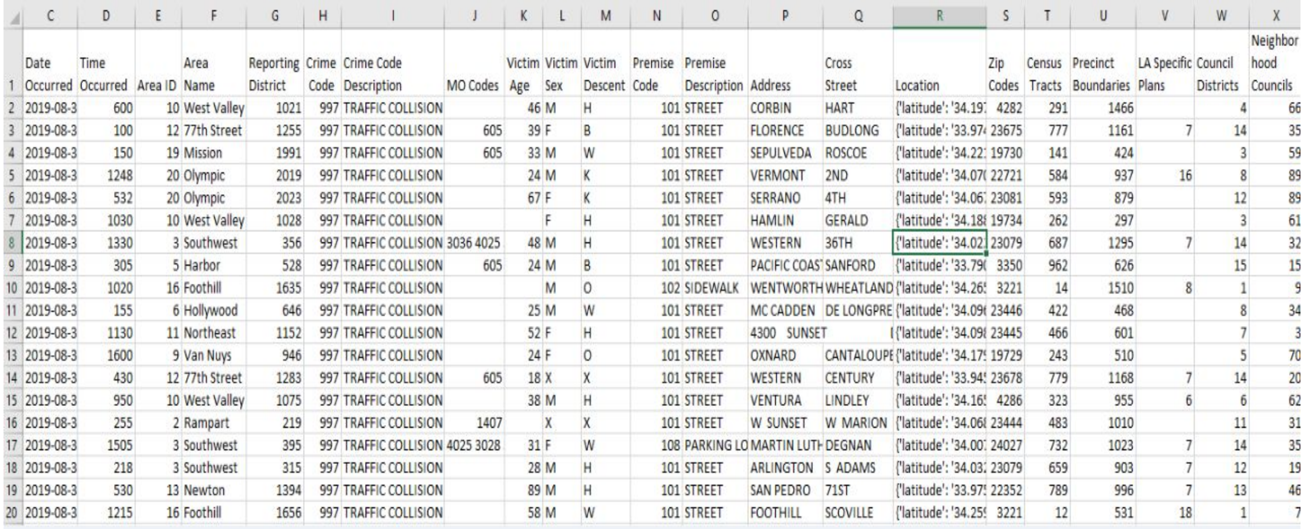
There are four research interests that our research is based on:

- Which factor contributes to the accidents happening on the weekends or on holidays.
- Which factor predicts a DUI (driving under influence) and hit and run.
- Which cluster of variables identifies incidents violated traffic laws.
- Forecasting model for the next 3 months (September to November) of 2019.

These findings will give an insight to create a better policy and/or education program to lower the traffic collision rate, thus reduce, fatality rate.

2. The Dataset

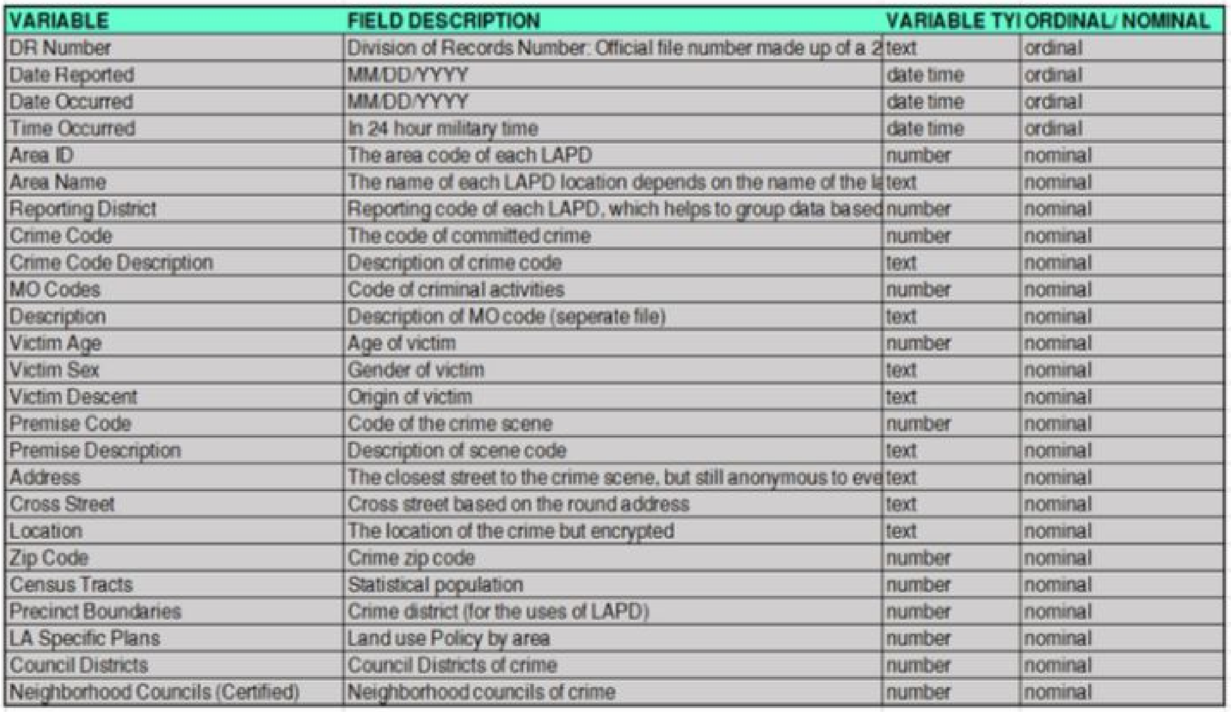
Los Angeles (L.A.) is one of the busiest urban metropolitan cities where people with various modes of transportation are on the road. The dataset was retrieved from Kaggle.com, which further directed us to the lacity.org website where a broad layer of information was available. The dataset includes records from 2010 to August 2019, when the data was pulled. There are total 18 attributes with 488,384 observations. The sample data of L.A. Traffic Collision data is provided below:



	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Date Occurred	Time Occurred	Area ID	Area Name	Reporting District	Crime Code	Crime Code Description	MO Codes	Victim Age	Victim Sex	Victim Descent	Premise Code	Premise Description	Address	Cross Street	Location	Zip Codes	Census Tracts	Precinct Boundaries	LA Specific Plans	Council Districts	Neighbor hood Councils
2	2019-08-3	600	10	West Valley	1021	997	TRAFFIC COLLISION		46	M	H	101	STREET	CORBIN	HART	{'latitude': '34.19	4282	291	1466		4	66
3	2019-08-3	100	12	77th Street	1255	997	TRAFFIC COLLISION	605	39	F	B	101	STREET	FLORENCE	BUDLONG	{'latitude': '33.97	23675	777	1161	7	14	35
4	2019-08-3	150	19	Mission	1991	997	TRAFFIC COLLISION	605	33	M	W	101	STREET	SEPULVEDA	ROSCOE	{'latitude': '34.22	19730	141	424		3	59
5	2019-08-3	1248	20	Olympic	2019	997	TRAFFIC COLLISION		24	M	K	101	STREET	VERMONT	2ND	{'latitude': '34.07	22721	584	937	16	8	89
6	2019-08-3	532	20	Olympic	2023	997	TRAFFIC COLLISION		67	F	K	101	STREET	SERRANO	4TH	{'latitude': '34.06	23081	593	879		12	89
7	2019-08-3	1030	10	West Valley	1028	997	TRAFFIC COLLISION			F	H	101	STREET	HAMLIN	GERALD	{'latitude': '34.18	19734	262	297		3	61
8	2019-08-3	1330	3	Southwest	356	997	TRAFFIC COLLISION	3036 4025	48	M	H	101	STREET	WESTERN	36TH	{'latitude': '34.02	23079	687	1295	7	14	32
9	2019-08-3	305	5	Harbor	528	997	TRAFFIC COLLISION	605	24	M	B	101	STREET	PACIFIC COAS	SANFORD	{'latitude': '33.79	3350	962	626		15	15
10	2019-08-3	1020	16	Foothill	1635	997	TRAFFIC COLLISION			M	O	102	SIDEWALK	WENTWORTH	WHEATLAND	{'latitude': '34.26	3221	14	1510	8	1	9
11	2019-08-3	155	6	Hollywood	646	997	TRAFFIC COLLISION		25	M	W	101	STREET	MC CADDEN	DE LONGPRE	{'latitude': '34.09	23446	422	468		8	34
12	2019-08-3	1130	11	Northeast	1152	997	TRAFFIC COLLISION		52	F	H	101	STREET	4300 SUNSET		{'latitude': '34.09	23445	466	601		7	3
13	2019-08-3	1600	9	Van Nuys	946	997	TRAFFIC COLLISION		24	F	O	101	STREET	OXNARD	CANTALOUPE	{'latitude': '34.17	19729	243	510		5	70
14	2019-08-3	430	12	77th Street	1283	997	TRAFFIC COLLISION	605	18	X	X	101	STREET	WESTERN	CENTURY	{'latitude': '33.94	23678	779	1168	7	14	20
15	2019-08-3	950	10	West Valley	1075	997	TRAFFIC COLLISION		38	M	H	101	STREET	VENTURA	LINDLEY	{'latitude': '34.16	4286	323	955	6	6	62
16	2019-08-3	255	2	Rampart	219	997	TRAFFIC COLLISION	1407		X	X	101	STREET	W SUNSET	W MARION	{'latitude': '34.06	23444	483	1010		11	31
17	2019-08-3	1505	3	Southwest	395	997	TRAFFIC COLLISION	4025 3028	31	F	W	108	PARKING LO	MARTIN LUTH	DEGNAN	{'latitude': '34.00	24027	732	1023	7	14	35
18	2019-08-3	218	3	Southwest	315	997	TRAFFIC COLLISION		28	M	H	101	STREET	ARLINGTON	S ADAMS	{'latitude': '34.03	23079	659	903	7	12	19
19	2019-08-3	530	13	Newton	1394	997	TRAFFIC COLLISION		89	M	H	101	STREET	SAN PEDRO	71ST	{'latitude': '33.97	22352	789	996	7	13	46
20	2019-08-3	1215	16	Foothill	1656	997	TRAFFIC COLLISION		58	M	W	101	STREET	FOOTHILL	SCOVILLE	{'latitude': '34.25	3221	12	531	18	1	7

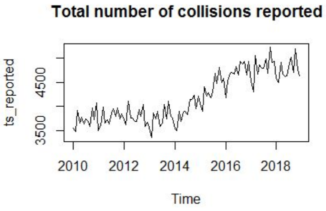
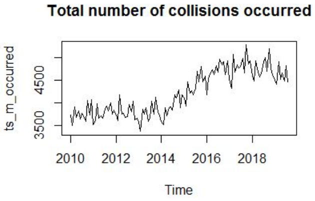
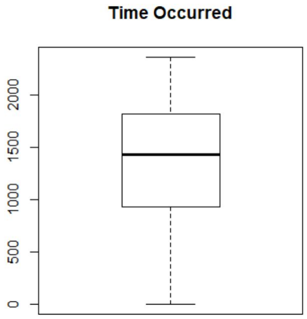
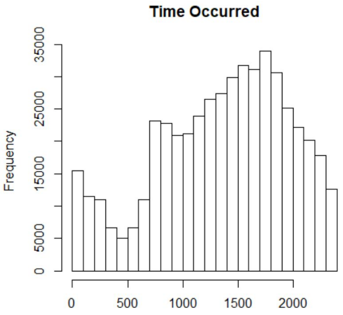
Exhibit 2.1: The sample data of L.A. Traffic Collision data.

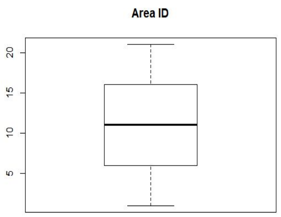
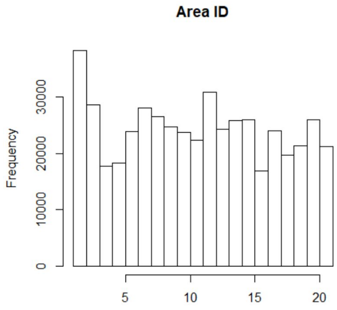
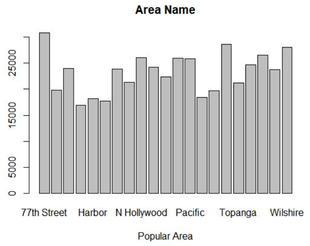
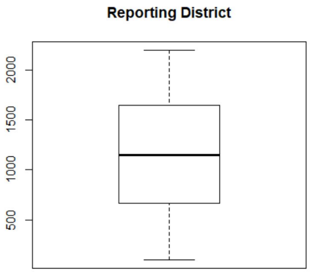
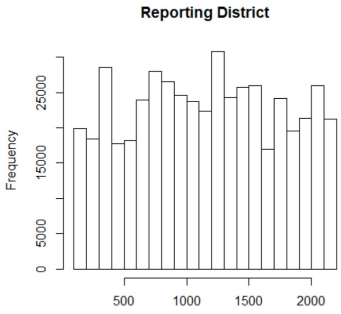
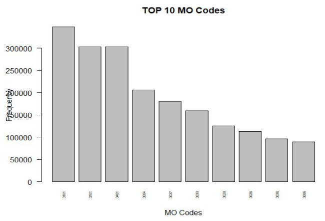
3. *The Descriptions of Variables*

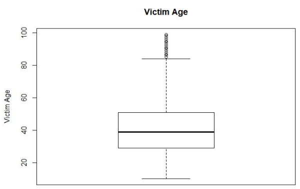
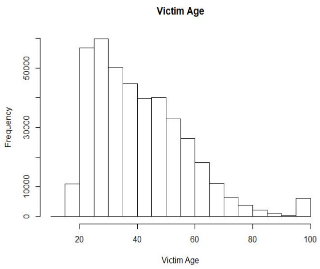
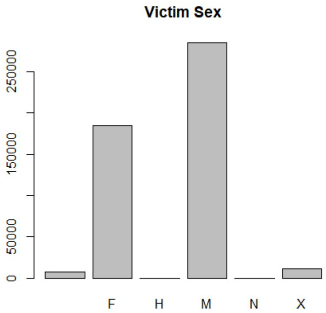
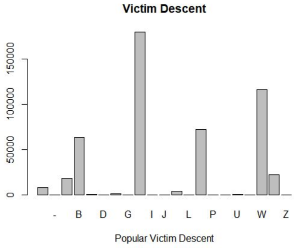


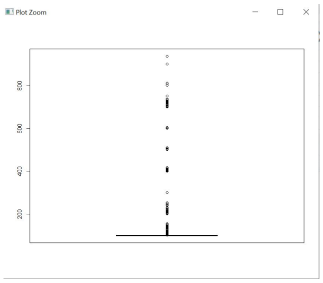
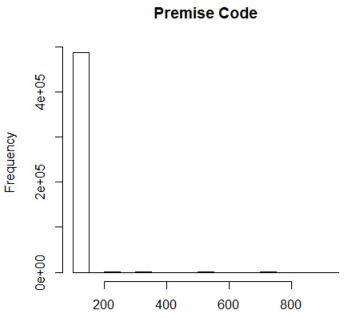
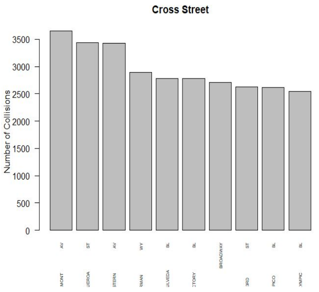
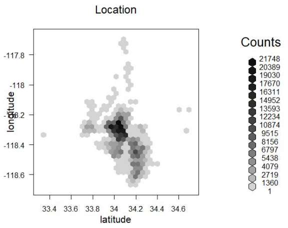
VARIABLE	FIELD DESCRIPTION	VARIABLE TYI	ORDINAL/ NOMINAL
DR Number	Division of Records Number: Official file number made up of a 2	text	ordinal
Date Reported	MM/DD/YYYY	date time	ordinal
Date Occurred	MM/DD/YYYY	date time	ordinal
Time Occurred	In 24 hour military time	date time	ordinal
Area ID	The area code of each LAPD	number	nominal
Area Name	The name of each LAPD location depends on the name of the la	text	nominal
Reporting District	Reporting code of each LAPD, which helps to group data based	number	nominal
Crime Code	The code of committed crime	number	nominal
Crime Code Description	Description of crime code	text	nominal
MO Codes	Code of criminal activities	number	nominal
Description	Description of MO code (seperate file)	text	nominal
Victim Age	Age of victim	number	nominal
Victim Sex	Gender of victim	text	nominal
Victim Descent	Origin of victim	text	nominal
Premise Code	Code of the crime scene	number	nominal
Premise Description	Description of scene code	text	nominal
Address	The closest street to the crime scene, but still anonymous to eve	text	nominal
Cross Street	Cross street based on the round address	text	nominal
Location	The location of the crime but encrypted	text	nominal
Zip Code	Crime zip code	number	nominal
Census Tracts	Statistical population	number	nominal
Precinct Boundaries	Crime district (for the uses of LAPD)	number	nominal
LA Specific Plans	Land use Policy by area	number	nominal
Council Districts	Council Districts of crime	number	nominal
Neighborhood Councils (Certified)	Neighborhood councils of crime	number	nominal

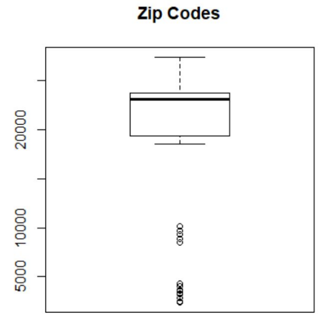
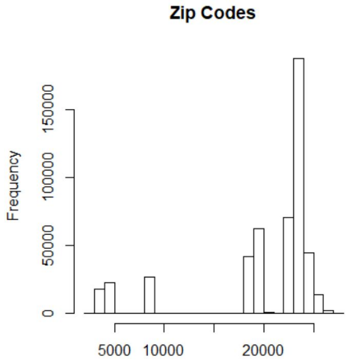
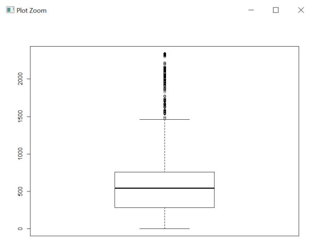
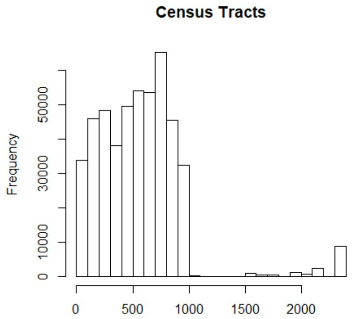
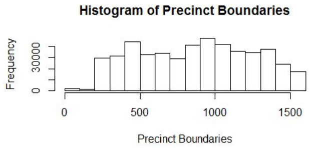
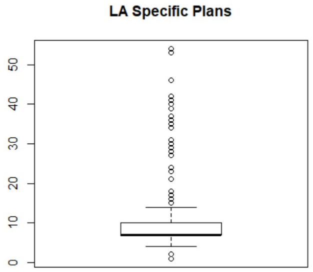
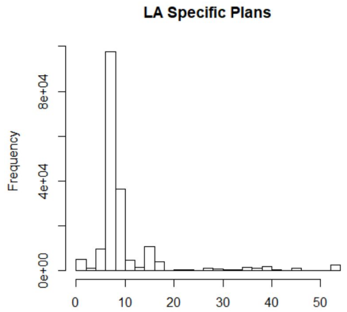
4. Data Exploration

Variable	Graph 1	Graph 2	Comment
DR Number	N/A	N/A	There is no missing value. 889 DR Number repeated 2 times, the rest showed only one time (MAX=2, MIN=1).
Date Reported			There is no missing value. From the graph, we can depict that the number of collisions was reported as the collision occurred. Note: 2019 is not considered as the data for the complete year is not available.
Date Occurred			There is no missing value. From the graph, it is evident that the number of collisions is increasing every year. Note: 2019 is not considered as the data for the complete year is not available.
Time Occurred			Through Graph 1, no outlier is indicated. There is no missing value. Graph 2 shows that there is a strong correlation between time occurred and the chance of the collision happened.

Area ID			Through Graph 1, no outlier is indicated. There is no missing value. There are 21 Area IDs.
Area Name			There is no missing value. There are some major area names, such as 77th street, Habor, N Hollywood, Pacific, Topanga, and Wilshire.
Reporting District			There are no missing values. There are no outliers.
Crime Code	N/A	N/A	There is only one constant value for this field. It will be deleted during data cleaning.
Crime Code Description	N/A	N/A	There is only one constant value for this field. It will be deleted during data cleaning.
MO Codes			There are 85,096 missing values.

Victim Age			There are 77,907 missing values. The age of 85 or larger is indicated as an outlier from graph 1.
Victim Sex			There are 5,770 missing values and 5 categories as follow (there is no description for H and N categories. Moreover, these 2 categories don't have a lot of observations and it would not affect the analysis): F (female): 184,950 obs H: 133 obs M (male): 284,766 obs N: 11 obs X (unknown): 11,328 obs
Victim Descent			There is no missing value, but there are 2 obs were represented by "-". There are some major descents, the 1st one is H - Hispanic/Latin/Mexican, following by W - White, O - Other, and B - Black.

Premise Code			There are 25 missing values.
Premise Description	N/A	N/A	There are 25 missing values. This variable is an enumerated variable.
Address	N/A	N/A	There are no missing values and 11,472 unique values. This variable will be useful for geographic map analysis but will not be used in prediction models.
Cross Street			There are 21,945 missing values. The most frequent location of collisions are at Vermont AV recorded as 3650.
Location			There is no missing values.

Zip Code			There are 396 missing values. Zip codes are usually 5-digit or 9-digit formats. However, there are 421,864 5-digit zip codes and 66,124 4-digit zip codes.
Census Tracts			There are 6,591 missing values. There are few outliers as shown in graph 1.
Precinct Boundaries			There are 3,114 missing values in this variable. Since these are area codes, outlier treatment will not be an appropriate step to explore.
LA Specific Plans			There are 308,578 missing values, which take about 63.18% of the observation. We will drop this variable in the preprocessing step.

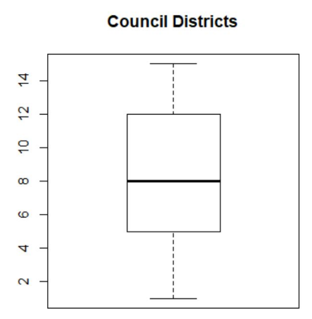
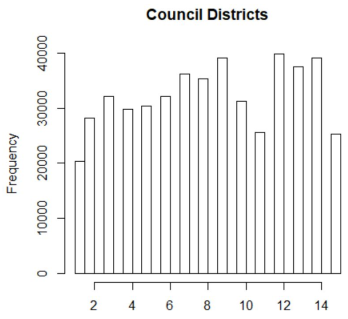
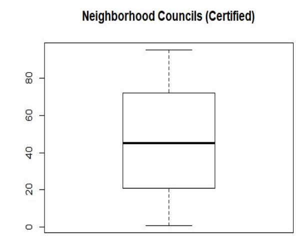
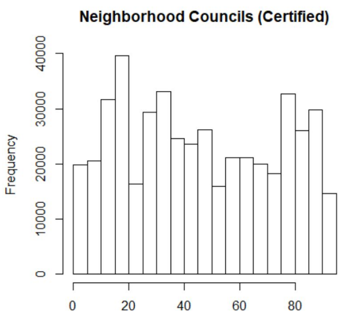
Council Districts			Through Graph 1, no outlier is indicated. There is no missing value and there are 15 council districts.
Neighborhood Councils (Certified)			Through Graph 1, no outlier is indicated. There are 24,381 missing values.

Exhibit 3.1: Exploratory Data Analysis of traffic collision variables

III. Data Preprocessing

1. *Create a binary 'Day Off' variable:*

Since we are interested in whether the accident happens on the weekend/holidays, we create a binary response variable based on 'Date Occurred' predictor combining with the US holiday/weekend calendar through manipulation in Excel.

2. *Missing values:*

When we deal with data preprocessing, we would first check whether values are missing from the dataset. Though missing data is common, it can cause a huge problem to the study on the dataset. Missing values can lower the representativeness of the sample, which can mislead people about the population of the dataset. If the dataset has missing values, we have to eliminate those values, or carry forward the values of the previous records. When a predictor has a lot of missing data (more than 30%) and, thus, becomes irrelevant for our prediction models, we will delete that predictor.

3. *Detect and delete outliers:*

We check if the variables that have outliers in order to decide whether we should leave the outliers in the dataset or take them off the dataset. When a dataset has too many outliers, those outliers will affect other analytical results dramatically. We run boxplots to have a closer look at the outliers in each variable. Then, we calculate lower and upper whiskers of the boxplot to identify the values of outliers. Lower whisker is equal to the smallest value greater than Q1-1.5IQR (Interquartile Range). Lower whisker represents the lower bound of the dataset. Upper whisker is calculated by the greatest value smaller than Q3+1.5IQR. It indicates the upper bound of the dataset. There are some outliers in Victim Age, Zip Code, and Census Tracts. Zip Code and Census Tracts's data is not in their right respective format. For example, the Zip Code should always have 5 digits, but our data have values less than 10000. As a result, we drop these two variables. On the other hand, Victim Age has outliers at 81 and above. People who have age 81 or higher are less likely to drive a vehicle. Therefore, we drop all the outliers for Victim Age.

4. *Frequence table of categorical variables - combine values:*

We run table function for all predictors to see if there is any category within each predictor that has a small number of observations. Usually, we combine with predictor's categories that have similar patterns. Since there are a lot of values in 'Area Name', we combine the areas into 5 main territories (South LA, North LA, West LA, East LA, and Central LA). In contrast, we take a different approach for 'Victim Descent'. White and Hispanic descent have the highest number of observations in our data. Thus, we combine other descents into one category called 'Other Descent' within 'Victim Descent'.

5. Collinearity:

Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable but also to each other. In other words, it results when you have factors that are a bit redundant. Serve multicollinearity is a major problem, because it increases the variance of the regression coefficients, making them unstable. The more variance they have, the more difficult it is to interpret the coefficients. If there is a strong indicator of collinearity between two predictors, we will delete one that is less appropriate for our response variable. There is no collinearity in our data as shown below:

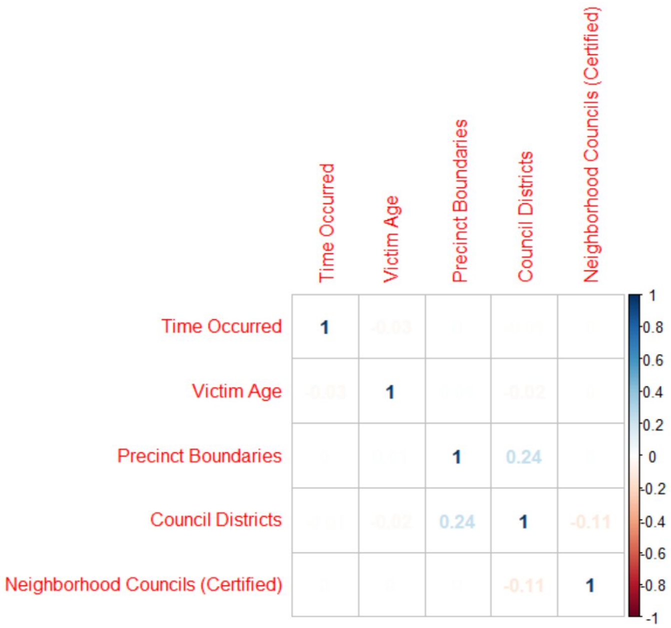


Exhibit 4.1: Correlation matrix to check multicollinearity in continuous variables

5. Data Reduction

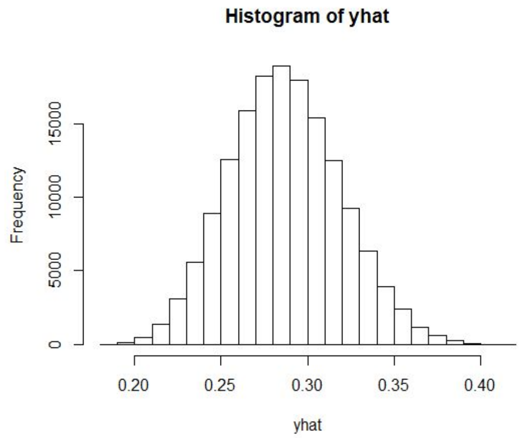
We will partition data randomly into standard 60% for training and 40% for validation to test the accuracy of the model.

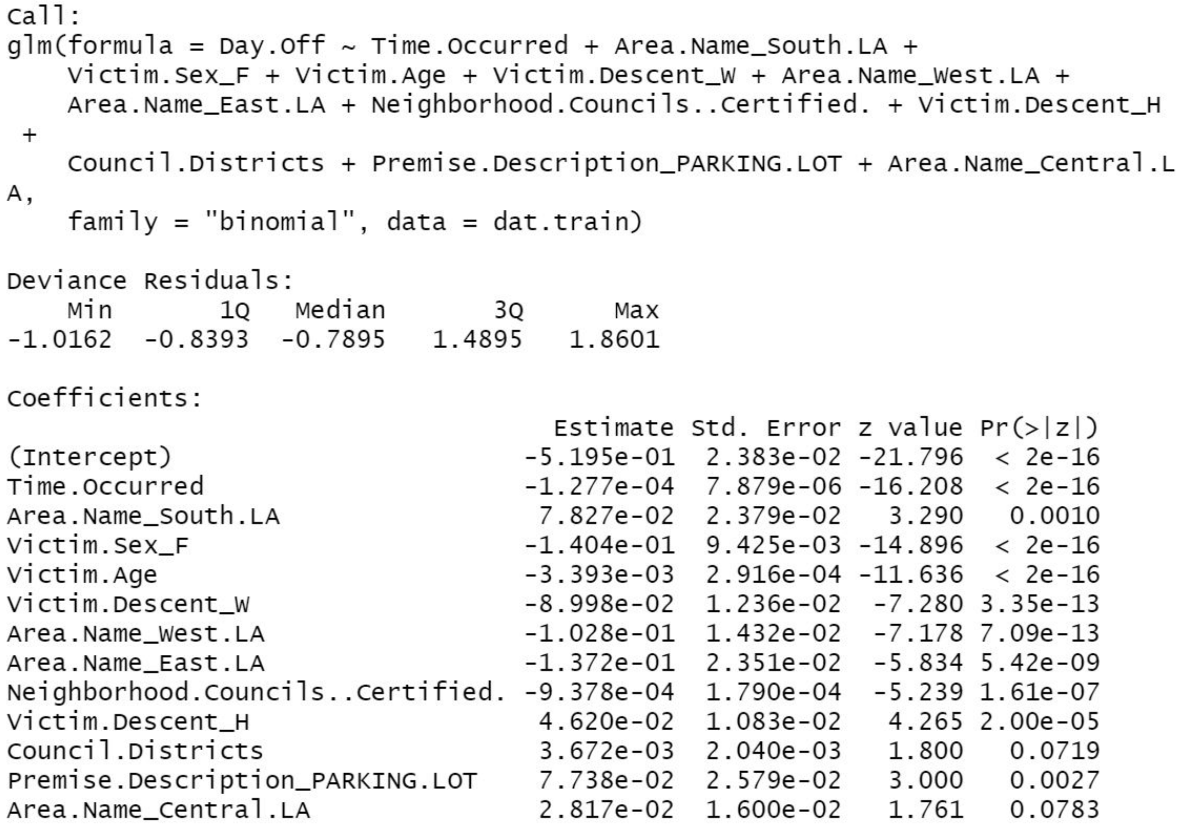
IV. Data Mining - Prediction Models

1. Logistic Regression

Logistic Regression is a statistical machine learning classification model used to predict the probability of a categorical dependent variable. In Logistic Regression, instead of using the dependent variable, it is expressed as a function of logit. In logistic regression, the output is modeled as a binary value (0 or 1) rather than a numeric value and the probability of the outcome lies between 0 and 1. The logit is modeled as a linear function of the predictors, which reflects the probability. An example logistic regression equation can be written as:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x)) where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). To implement logistic regression, all the categorical predictors are transformed into dummy variables (0/1). Dummy variables have been created for these categorical predictors - Area Name, Victim Sex, Victim Descent and Premise Description. The areas were categorized into South LA, North LA, Central LA, West LA and East LA. The sex of the victim was categorized into Male (M), Female (F) and Other (X). Majority of traffic collisions have occurred on the street and parking lot. Therefore, we have categorized Premise Description into Street, Parking Lot and Other Premises. Victim Descent was categorized into H, Other Descent and W. We divided our data into training data 60% and validation data 40%. There are 20 predictors or variables and the logit is modeled as a linear function of the predictors, which reflects the probability. The accuracy of the model is around 71.15 %.





```
Call:
glm(formula = Day.Off ~ Time.Occurred + Area.Name_South.LA +
    Victim.Sex_F + Victim.Age + Victim.Descent_W + Area.Name_West.LA +
    Area.Name_East.LA + Neighborhood.Councils..Certified. + Victim.Descent_H
 +
    Council.Districts + Premise.Description_PARKING.LOT + Area.Name_Central.L
A,
    family = "binomial", data = dat.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0162  -0.8393  -0.7895   1.4895   1.8601

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -5.195e-01  2.383e-02 -21.796  < 2e-16
Time.Occurred                    -1.277e-04  7.879e-06 -16.208  < 2e-16
Area.Name_South.LA                7.827e-02  2.379e-02   3.290   0.0010
Victim.Sex_F                     -1.404e-01  9.425e-03 -14.896  < 2e-16
Victim.Age                       -3.393e-03  2.916e-04 -11.636  < 2e-16
Victim.Descent_W                 -8.998e-02  1.236e-02  -7.280 3.35e-13
Area.Name_West.LA                -1.028e-01  1.432e-02  -7.178 7.09e-13
Area.Name_East.LA                -1.372e-01  2.351e-02  -5.834 5.42e-09
Neighborhood.Councils..Certified. -9.378e-04  1.790e-04  -5.239 1.61e-07
Victim.Descent_H                  4.620e-02  1.083e-02   4.265 2.00e-05
Council.Districts                 3.672e-03  2.040e-03   1.800   0.0719
Premise.Description_PARKING.LOT   7.738e-02  2.579e-02   3.000   0.0027
Area.Name_Central.LA              2.817e-02  1.600e-02   1.761   0.0783
```

2. K-Nearest Neighbor

K-Nearest Neighbor (kNN) is an algorithm, which can be used for classification or prediction. kNN classification works best for categorical output and kNN prediction is more suitable for numerical response. In this project we applied both kNN prediction and kNN classification.

There are pros and cons when using kNN method. One of the pros of kNN is: no model-driven. kNN is data-driven. It means that users don’t have to fit a data model like linear regression. Besides that, users don’t have to make any assumptions about the data, which seems to be more objective. However, kNN does take a long time to calculate and produce k value.

We started kNN analysis process by selecting the variables which can be fitted in the kNN model and have great impacts on the prediction process, and normalizing those variables to avoid bias analysis results. The results of our kNN method is the majority of the car accidents in LA area happened on weekends or holidays. The accuracy of our kNN classification is approximately 70.7%.

3. *Classification Tree*

Classification tree is one of the most popular methods used for prediction. It's flexible and data-driven. It results in a set of rules by dividing observations into subgroups based on predictor values. The diagram below is a minimum error tree where the validation data will have the lowest error rate. If an accident occurs from midnight to 5:04 am, it is more likely happened on the weekend/holiday.

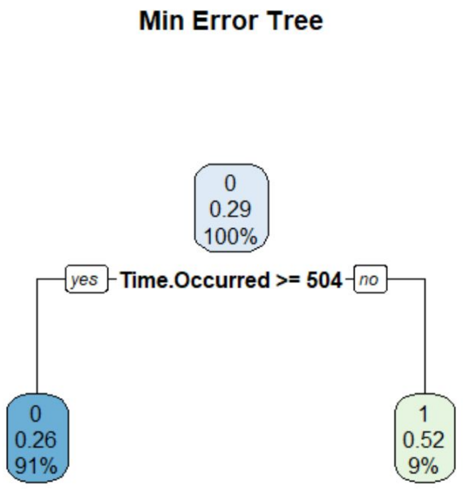


Exhibit 6.3: Classification tree model

4. *Prediction results*

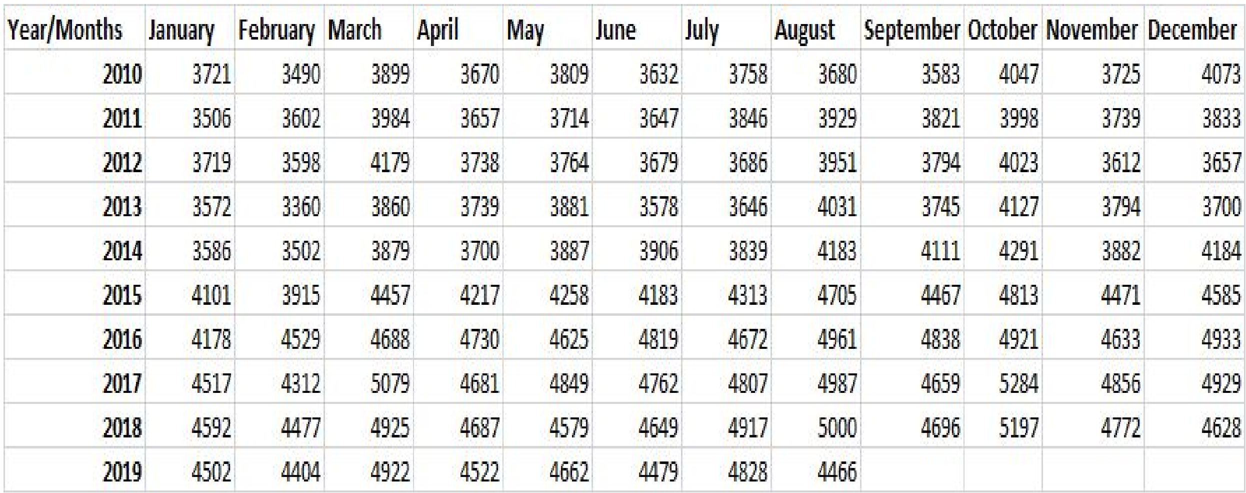
We use accuracy rate to evaluate the performance of three prediction models. Logistic Regression, kNN, and classification have the accuracy of 71.5%, 70.7%, and

71.85%, respectively. Since the classification tree has the highest accuracy, it is our best prediction model. Compared to the other two models, the classification tree has only one variable, ‘Time Occurred’, which makes it even more preferable.

V. Forecasting

1. Data Transformation for forecasting

“It is likely that unlikely will happen. But, if you wait long enough, the unlikely will happen.” (Aristotle, 2400 B.C). Hence, forecasting is a method that takes past values into consideration to predict the estimates of the future values. It helps to develop strategies to make a better decision. The question of interest is to forecast monthly collisions from September 2019 to November 2019. The variable "Date Occurred” is the daily date of the collision which was aggregated to monthly data. After the transformation, there are a total of 116 observations. A snapshot of the data can be viewed as followed:



Year/Months	January	February	March	April	May	June	July	August	September	October	November	December
2010	3721	3490	3899	3670	3809	3632	3758	3680	3583	4047	3725	4073
2011	3506	3602	3984	3657	3714	3647	3846	3929	3821	3998	3739	3833
2012	3719	3598	4179	3738	3764	3679	3686	3951	3794	4023	3612	3657
2013	3572	3360	3860	3739	3881	3578	3646	4031	3745	4127	3794	3700
2014	3586	3502	3879	3700	3887	3906	3839	4183	4111	4291	3882	4184
2015	4101	3915	4457	4217	4258	4183	4313	4705	4467	4813	4471	4585
2016	4178	4529	4688	4730	4625	4819	4672	4961	4838	4921	4633	4933
2017	4517	4312	5079	4681	4849	4762	4807	4987	4659	5284	4856	4929
2018	4592	4477	4925	4687	4579	4649	4917	5000	4696	5197	4772	4628
2019	4502	4404	4922	4522	4662	4479	4828	4466				

2. Model Criteria

We first identified the four components of the time series which are trend, seasonality, cyclical and random variations. Trend is a general tendency that increases or decreases in a predictable manner. Trend methods involves determining the speed and direction of data over time. Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every period (usually calendar year). While cyclical component indicates recurrent variation in the time series, a fluctuation in data that is caused by uncertain or random occurrences. Random variations or noise is unexplainable variability and it is something which does not fall under any of the above three described. Visual analysis will help us to understand the time-series and produce various forecasting methods to understand various components of the time-series in the given data.

We are going to select the best model based by evaluating performance measure. There are various accuracy measures like Mean Error (ME), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE). However, we are focusing on MAPE as it is unit free, making the interpretation easier, as well as it is useful when comparing two entirely two different models. Lower the MAPE, better the model.

3. *Time Series Analysis*

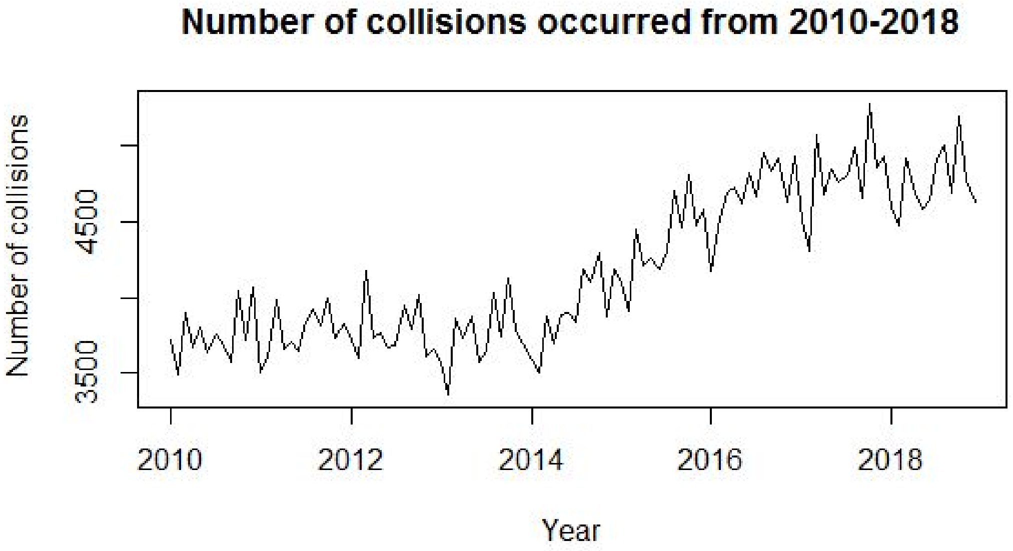


Exhibit 6.5: Time-series from January 2010 to August 2019.

From the exhibit 6.5 above, there seems to be an upper trend starting from 2014 and a seasonality in the data. Also, there seems to be a multiplicative seasonality as seasonal cycle seems to be growing over time. Multiplicative seasonality is when the seasonality influences increases or decreases with the increase and decrease in the level of the series. the seasonality. To confirm the presence of these components, we will decompose the time-series into its components as shown in Exhibit 6.6.

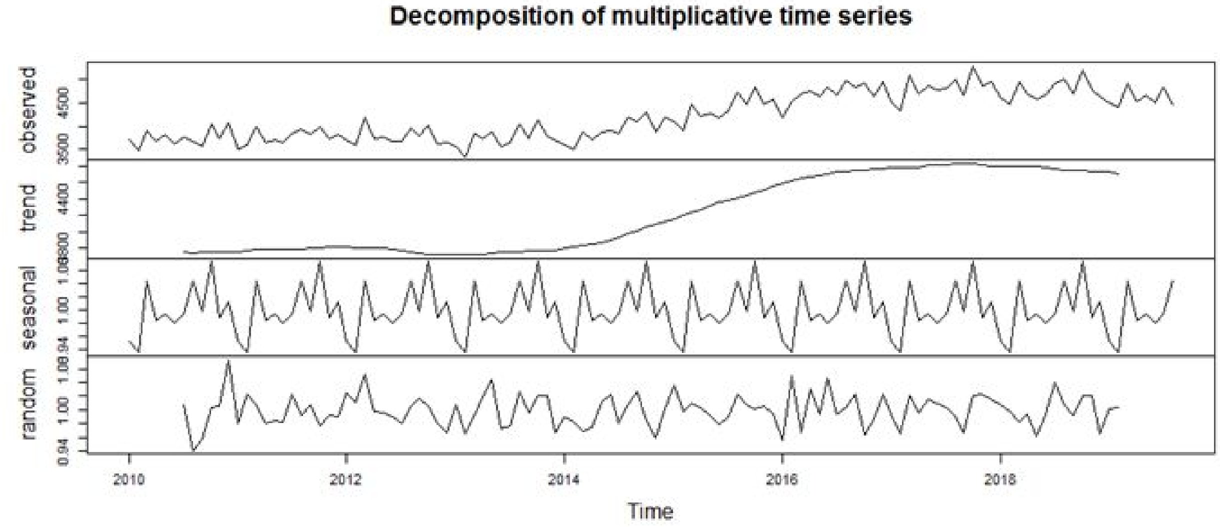


Exhibit 6.6 - Decomposition of time series

From Exhibit 6.6, we see that the time series has a gradual growing trend. Hence, it is not stationary. To understand the type of trend (linear, damped, exponential trend), we will cover in the next topic. The repetitive pattern in the seasonal variations confirms the presence of seasonal pattern. Seasonality is influenced by climate change, human behavior, holidays, etc.

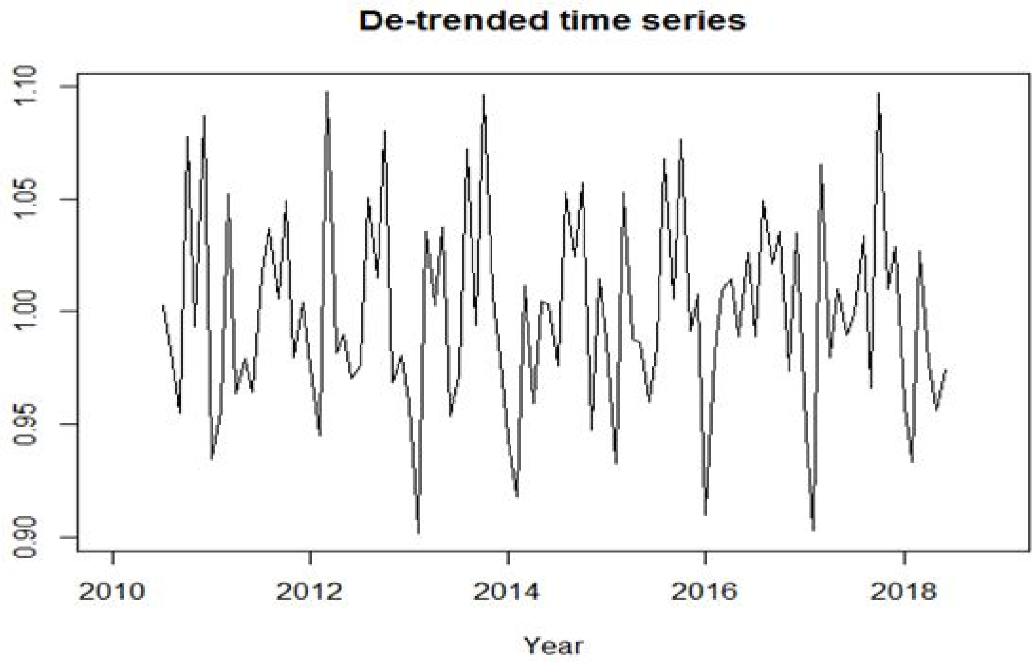


Exhibit 6.7: Detrended time series of the number of collisions from 2010 to 2019

This detrended time series forces its mean to zero and reduces overall variation. It helps removes any kind of distortion, provide a clear picture of the data as well as focus on other important factor(s) (if present).

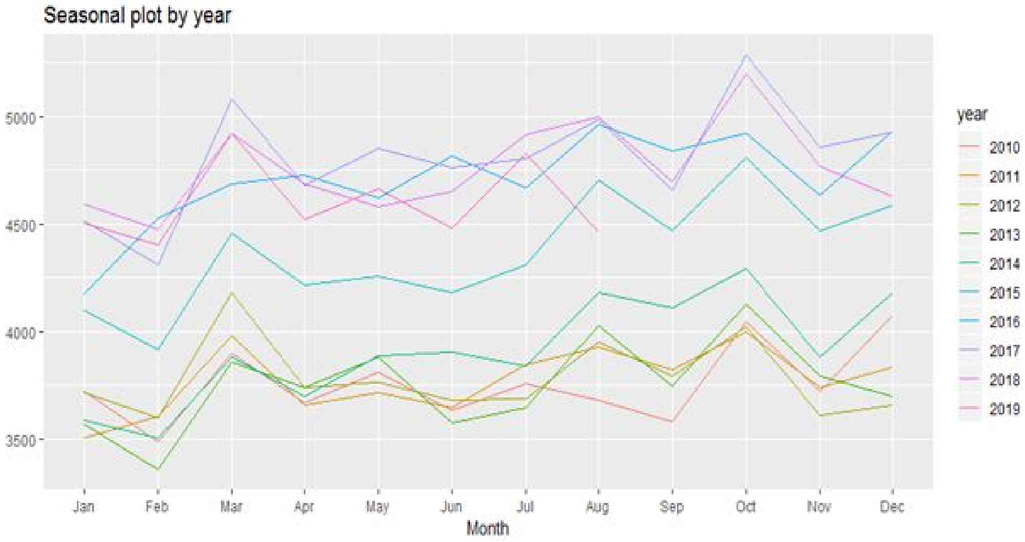
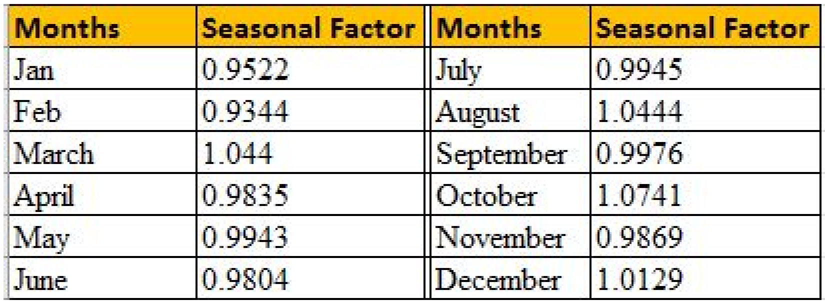


Exhibit 6.8: Seasonal plot of number of collisions by year

The above visualization graph (Exhibit 6.8) shows the seasonality by year starting from 2010 to August 2019. The line graph shows that the seasonality has increased from 2010 to 2019. Hence, this confirms the presence of multiplicative seasonality in the number of collisions. Also, to understand the concentration of accidents based on certain seasons, we focused on seasonal factor. A seasonal factor greater than 1 indicates that the number of collisions for that month was above yearly average. On the other hand, a seasonal factor below 1 indicates the number of collisions was below yearly average number of collisions. From the exhibit 6.9 and 6.10, October averaged the highest number of collisions. In fact, March and August had the second highest whereas January and February averaged the lowest rate. High number of collisions in October could be due to on-set of fall season. Weather conditions do play an important role in traffic conditions, hence rise and fall in the seasonal factor can be observed. Other factor such as holiday seasons, number of road trips/journeys made which also depend on the weather could be the contribution to fatal accidents.



Months	Seasonal Factor	Months	Seasonal Factor
Jan	0.9522	July	0.9945
Feb	0.9344	August	1.0444
March	1.044	September	0.9976
April	0.9835	October	1.0741
May	0.9943	November	0.9869
June	0.9804	December	1.0129

Exhibit 6.9: Seasonal factors by month

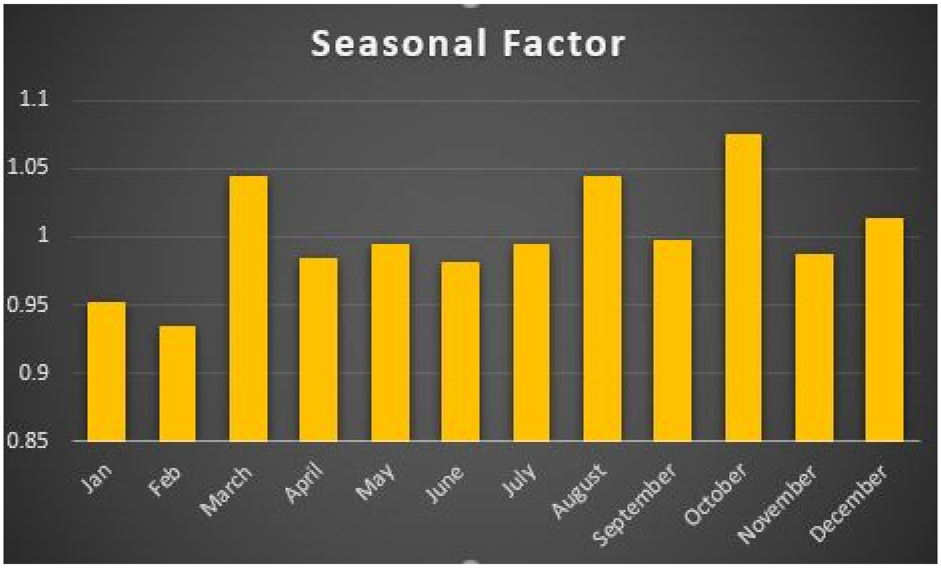


Exhibit 6.10: Bar chart of seasonal factors by month

4. *Exponential smoothing models*

Exponential Smoothing method is one of the forecasting methods. They use weighted averages of past observations to forecast new values. They give more importance to recent values in the time series. Thus, as observations get older (in time), the importance of these values get exponentially smaller.

For exponential smoothing model, we have again, considered the data from the year 2010 to 2019 (August). We will be forecasting for the next three months (i.e. September 2019, October 2019 and November 2019) giving more importance to the recent values over the older observations.

(a) Simple exponential smoothing model (SES): It assumes there is no trend and seasonality in the data. It requires only one parameter called alpha or smoothing factor which has a range from 0 to 1. Smoothing factor controls the rate at which the influence of the observations at prior time helps decay exponentially. A value close to 1 indicates the most recent values are weighted heavily related to older past observations whereas value close to 0 indicates the older observations influence the forecasts. The graph in Exhibit 4.8 shows that forecasted values are going to be the same for the next 3 months which can be confirmed by viewing the summary in this exhibit.

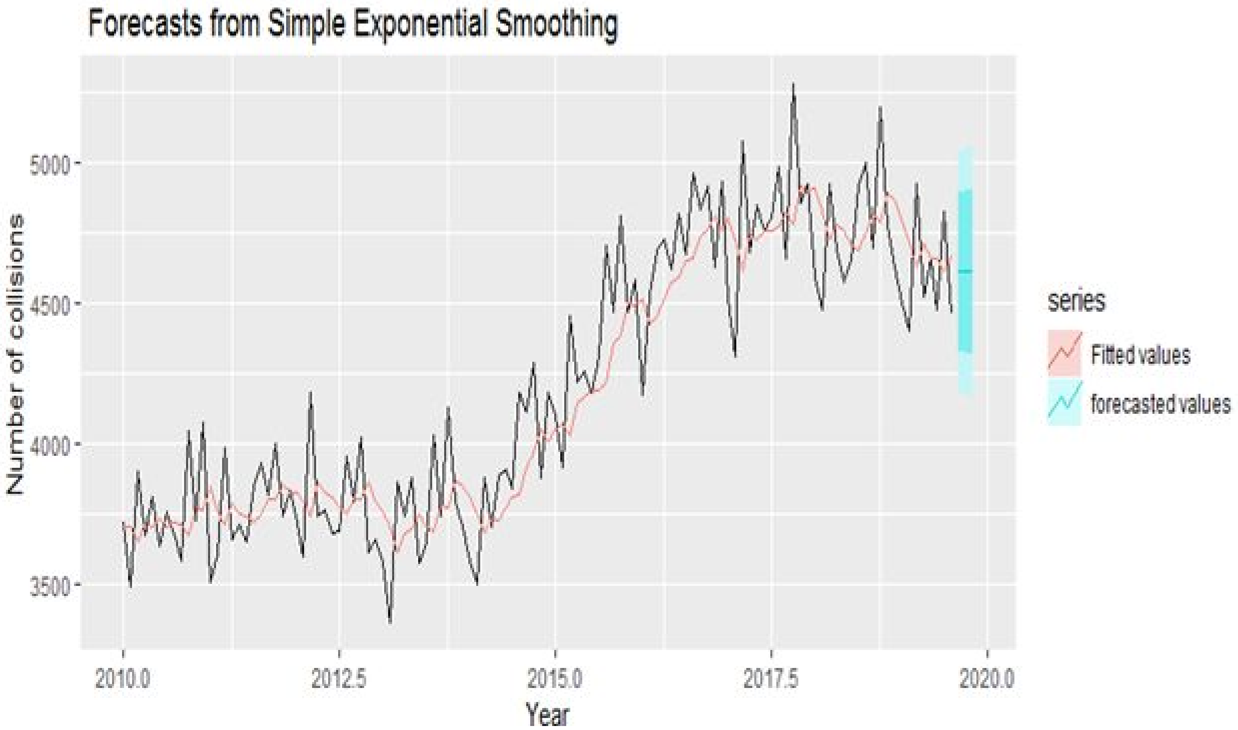
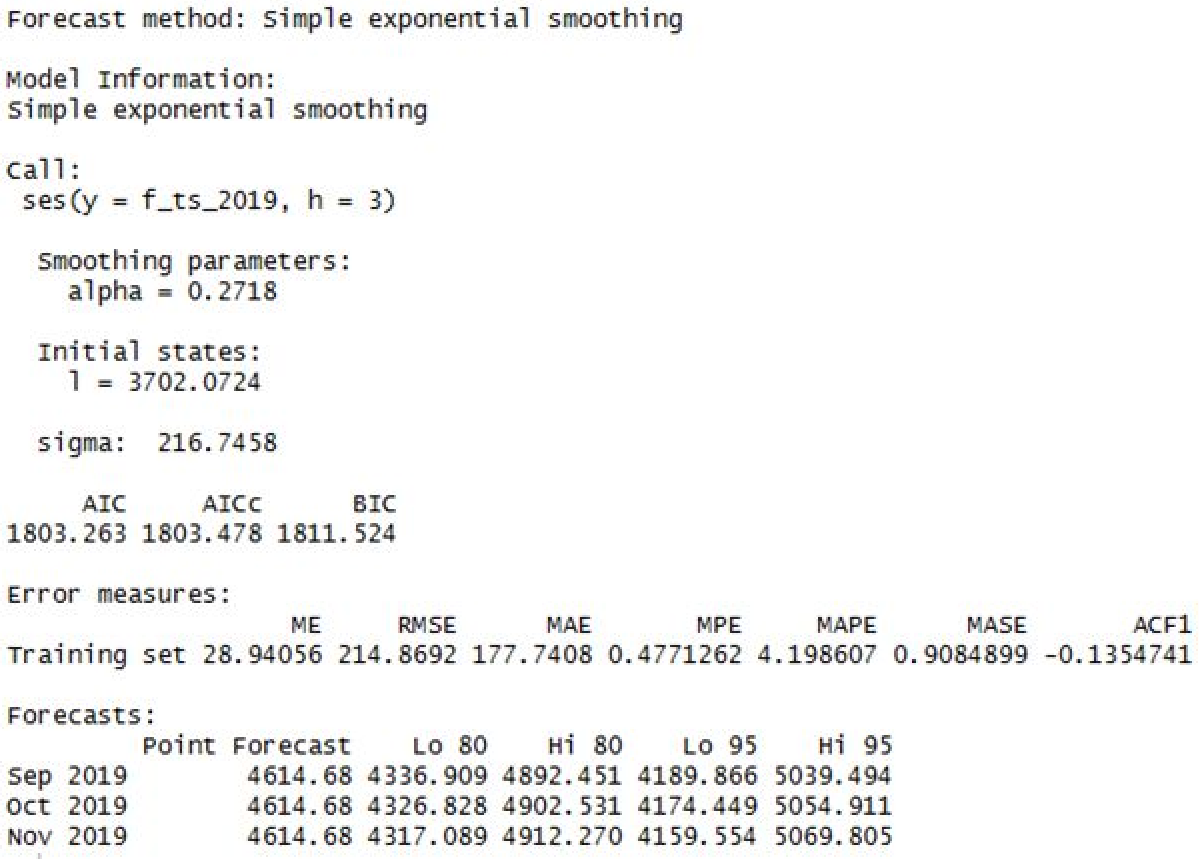
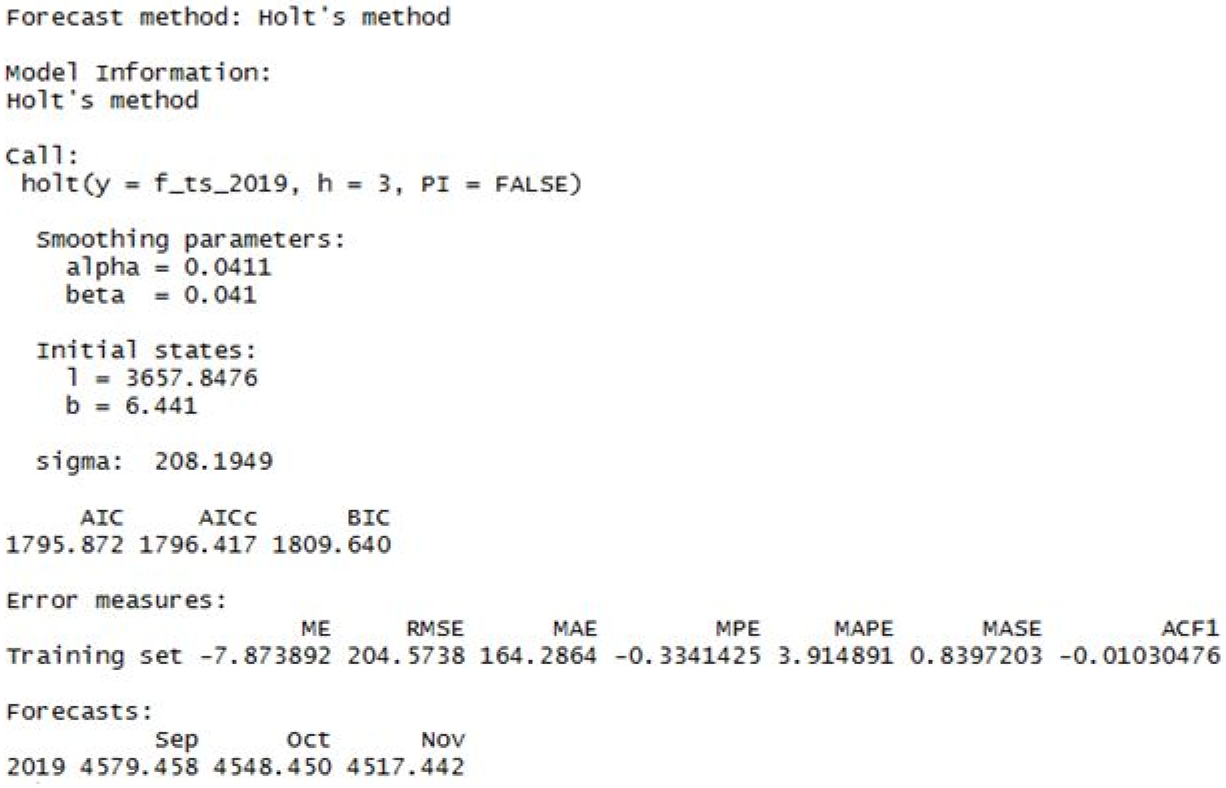


Exhibit 6.11: Graph and Summary of Simple Exponential Smoothing model



```
Forecast method: Simple exponential smoothing

Model Information:
Simple exponential smoothing 

Call:
 ses(y = f_ts_2019, h = 3) 

  Smoothing parameters:
    alpha = 0.2718 

  Initial states:
    l = 3702.0724 

  sigma:  216.7458

     AIC     AICc      BIC 
1803.263 1803.478 1811.524 

Error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 28.94056 214.8692 177.7408 0.4771262 4.198607 0.9084899 -0.1354741

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Sep 2019        4614.68 4336.909 4892.451 4189.866 5039.494
Oct 2019        4614.68 4326.828 4902.531 4174.449 5054.911
Nov 2019        4614.68 4317.089 4912.270 4159.554 5069.805
```

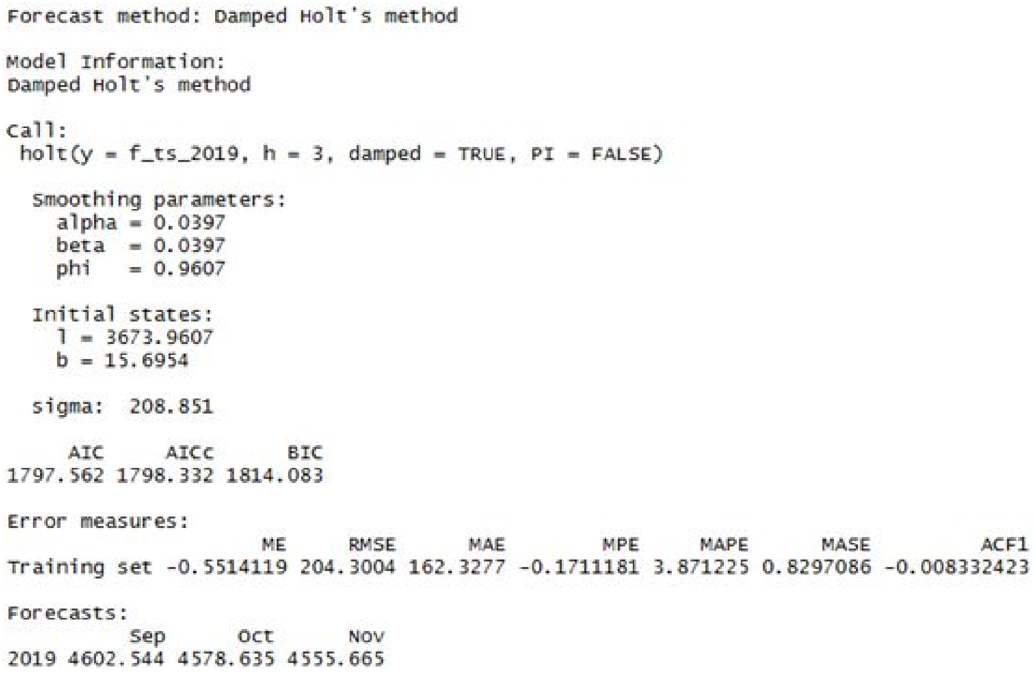
(b) Holt's Linear method: This method assumes that there is a presence of linear trend and there is no seasonality in the data. The forecasts from linear trend extrapolate the last estimate of the trend without limit. It involves two smoothing factors: alpha for the level and beta for the trend. The value of Beta also ranges from 0 to 1. The value of beta close to 0 reduces to simple exponential smoothing model.



```
Forecast method: Holt's method

Model Information:
Holt's method

Call:
 holt(y = f_ts_2019, h = 3, PI = FALSE)

  Smoothing parameters:
    alpha = 0.0411
    beta  = 0.041

  Initial states:
    l = 3657.8476
    b = 6.441

  sigma:  208.1949

     AIC     AICc      BIC
1795.872 1796.417 1809.640

Error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
Training set -7.873892 204.5738 164.2864 -0.3341425 3.914891 0.8397203 -0.01030476

Forecasts:
          Sep      Oct      Nov
2019 4579.458 4548.450 4517.442
```

Exhibit 6.12: Summary of Holt's Linear method

(c) Damped Holt's method: This method assumes that there is a damped trend and there is no seasonality. The forecast from damped trend starts almost linearly but dies off exponentially until they reach a constant level.



```
Forecast method: Damped Holt's method

Model Information:
Damped Holt's method

Call:
 holt(y = f_ts_2019, h = 3, damped = TRUE, PI = FALSE)

  Smoothing parameters:
    alpha = 0.0397
    beta  = 0.0397
    phi   = 0.9607

  Initial states:
    l = 3673.9607
    b = 15.6954

  sigma:  208.851

     AIC     AICc      BIC
1797.562 1798.332 1814.083

Error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE         ACF1
Training set -0.5514119 204.3004 162.3277 -0.1711181 3.871225 0.8297086 -0.008332423

Forecasts:
         Sep      Oct      Nov
2019 4602.544 4578.635 4555.665
```

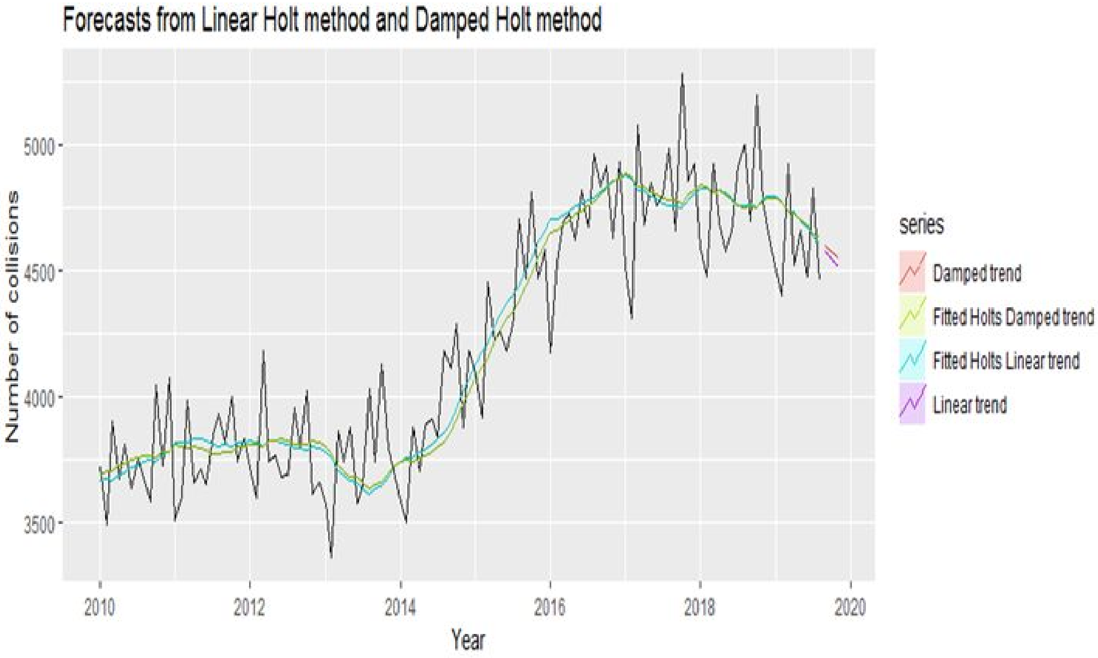
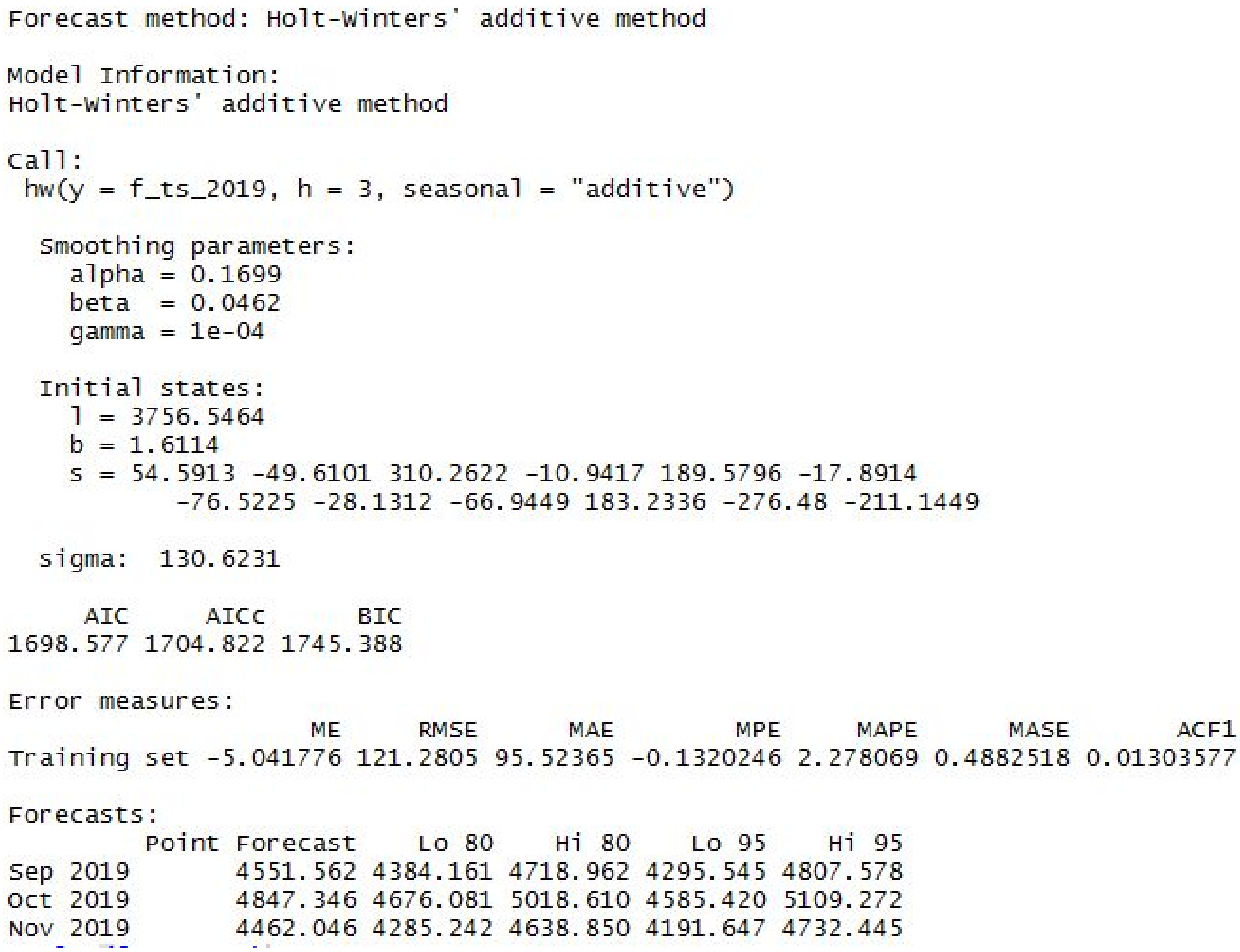


Exhibit 6.12: Summary of Damped Holt's method

(d) Holt-Winter's additive method: This model is an extension of Holt's exponential model. It assumes that there is a linear trend and an additive seasonality. There are three smoothing factors: alpha for level, beta for trend and gamma for seasonal adjustment. The value of gamma also ranges from 0 to 1.



```
Forecast method: Holt-Winters' additive method

Model Information:
Holt-Winters' additive method

Call:
 hw(y = f_ts_2019, h = 3, seasonal = "additive")

  Smoothing parameters:
    alpha = 0.1699
    beta  = 0.0462
    gamma = 1e-04

  Initial states:
    l = 3756.5464
    b = 1.6114
    s = 54.5913 -49.6101 310.2622 -10.9417 189.5796 -17.8914
           -76.5225 -28.1312 -66.9449 183.2336 -276.48 -211.1449

  sigma:  130.6231

     AIC     AICc      BIC
1698.577 1704.822 1745.388

Error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set -5.041776 121.2805 95.52365 -0.1320246 2.278069 0.4882518 0.01303577

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Sep 2019       4551.562 4384.161 4718.962 4295.545 4807.578
Oct 2019       4847.346 4676.081 5018.610 4585.420 5109.272
Nov 2019       4462.046 4285.242 4638.850 4191.647 4732.445
```

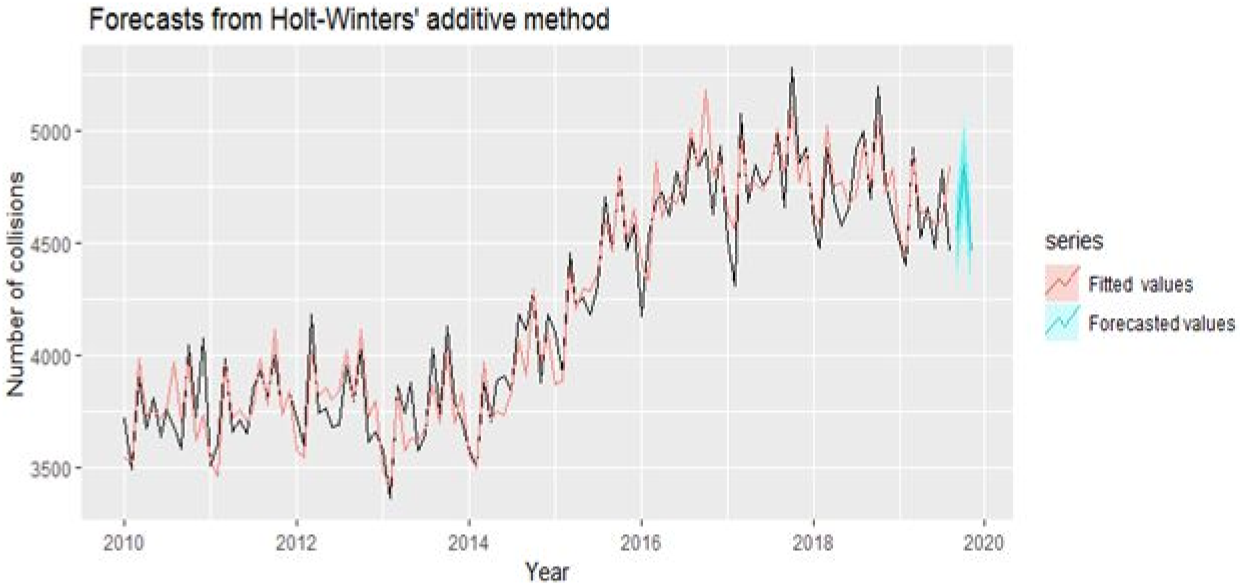
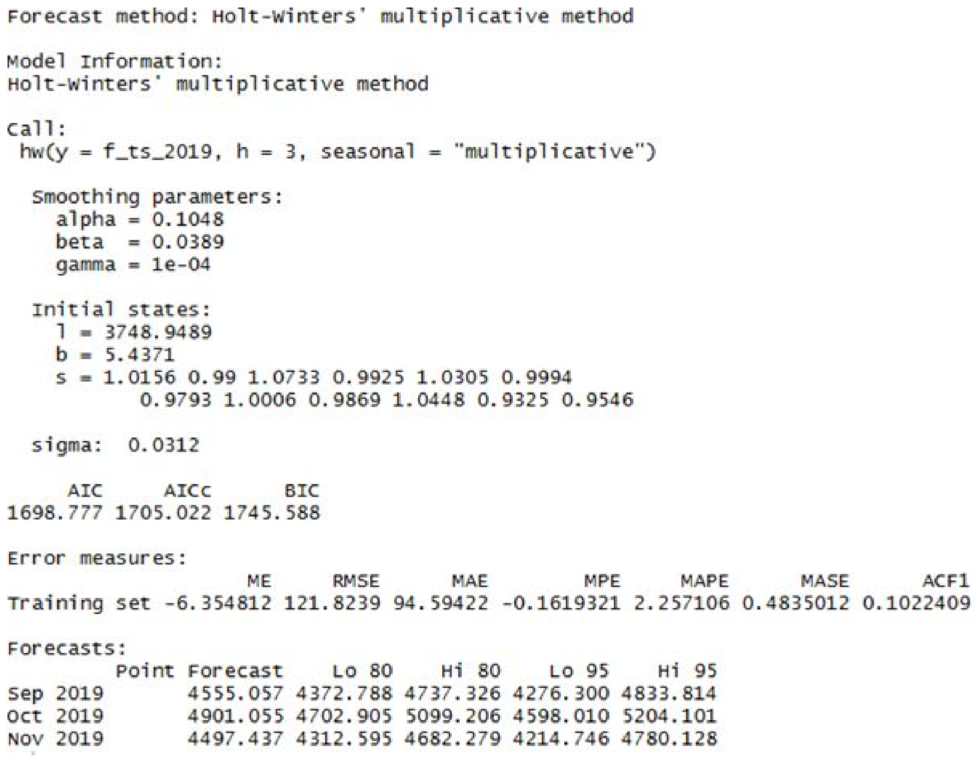


Exhibit 6.14: Graph and Summary of Holt-Winter's additive model

(e) Holt-Winter's Multiplicative model: It is similar to the above model. The only difference is it assumes that there is a linear trend and a multiplicative seasonality in the data.



```
Forecast method: Holt-Winters' multiplicative method

Model Information:
Holt-Winters' multiplicative method

Call:
 hw(y = f_ts_2019, h = 3, seasonal = "multiplicative")

  Smoothing parameters:
    alpha = 0.1048
    beta  = 0.0389
    gamma = 1e-04

  Initial states:
    l = 3748.9489
    b = 5.4371
    s = 1.0156 0.99 1.0733 0.9925 1.0305 0.9994
           0.9793 1.0006 0.9869 1.0448 0.9325 0.9546

  sigma:  0.0312

     AIC     AICc      BIC
1698.777 1705.022 1745.588

Error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE      ACF1
Training set -6.354812 121.8239 94.59422 -0.1619321 2.257106 0.4835012 0.1022409

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Sep 2019       4555.057 4372.788 4737.326 4276.300 4833.814
Oct 2019       4901.055 4702.905 5099.206 4598.010 5204.101
Nov 2019       4497.437 4312.595 4682.279 4214.746 4780.128
```

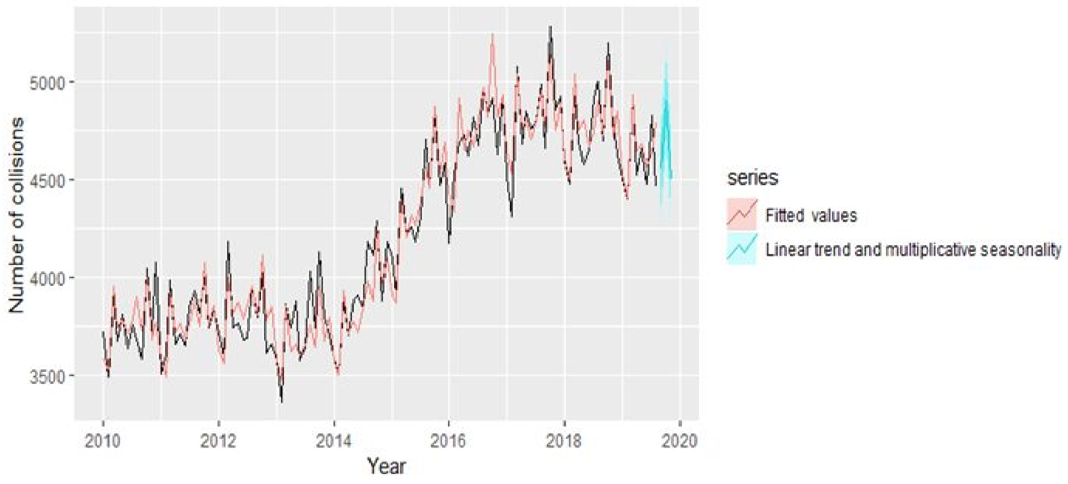
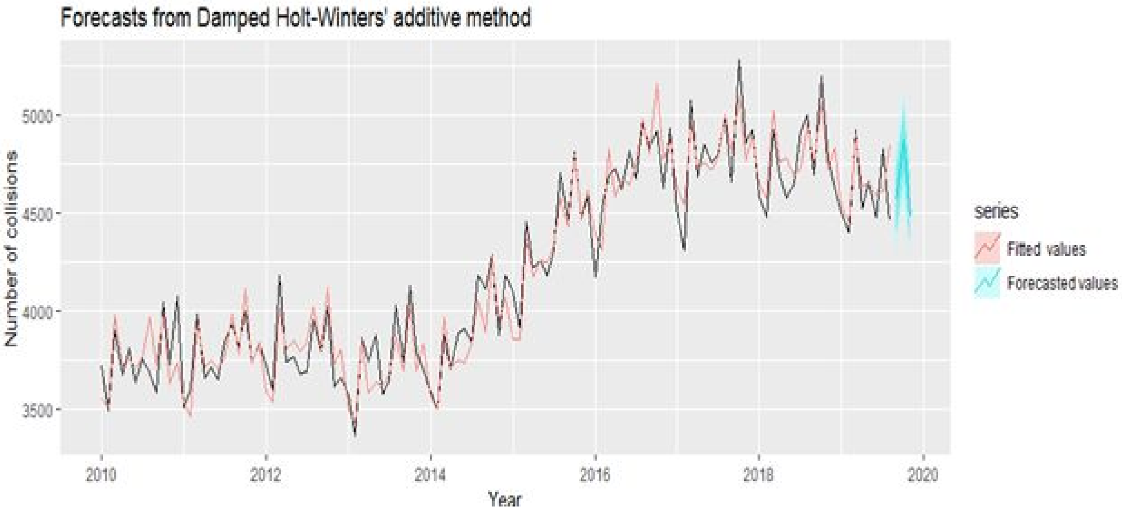
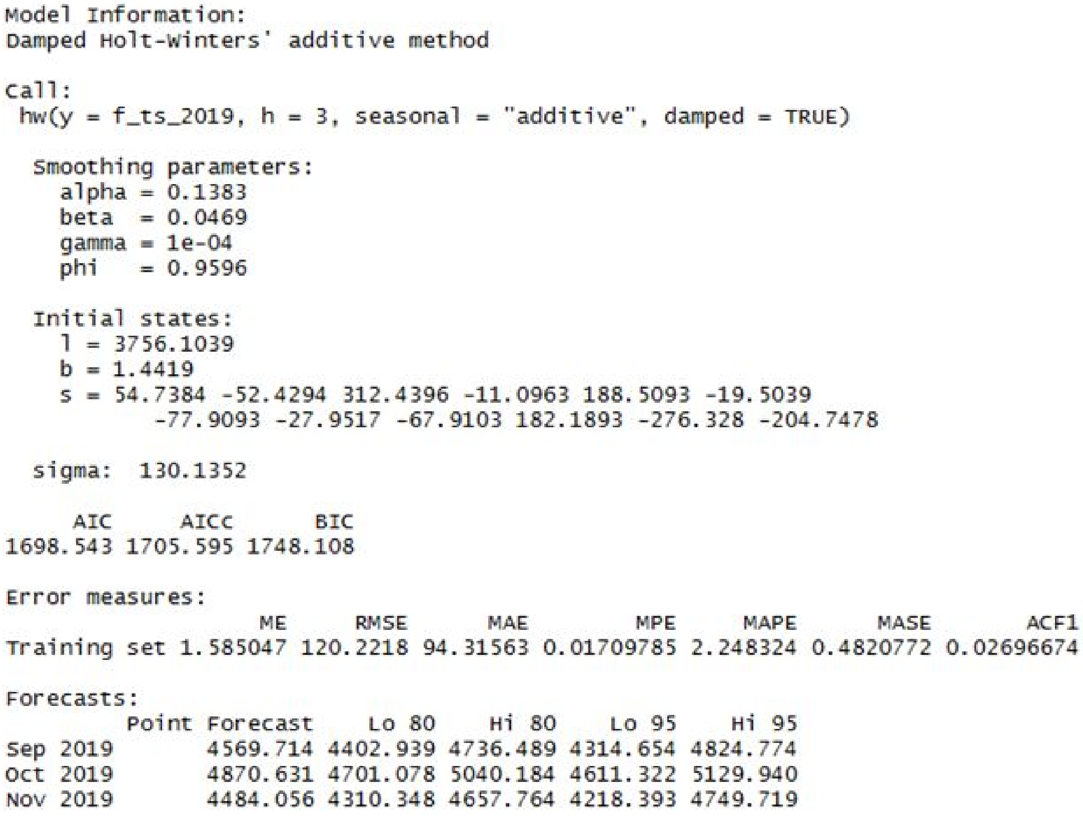


Exhibit 6.15: Graph and Summary of Holt-Winter's multiplicative model

(f) Damped Holt-Winter's additive method: This exponential model considers that there is a damp trend and additive seasonality in the time series.

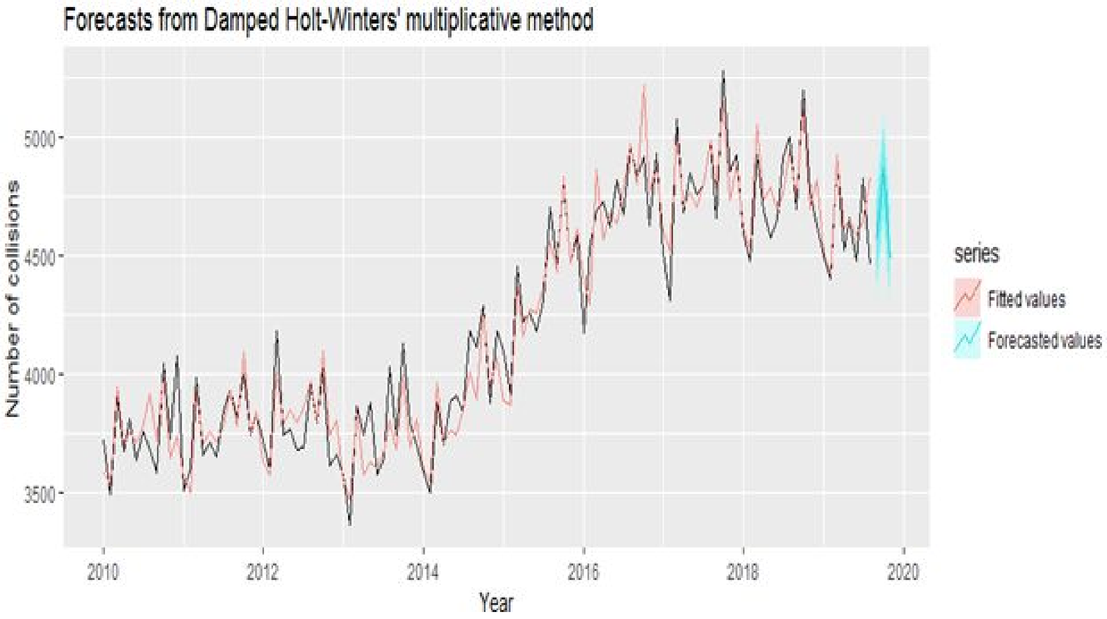


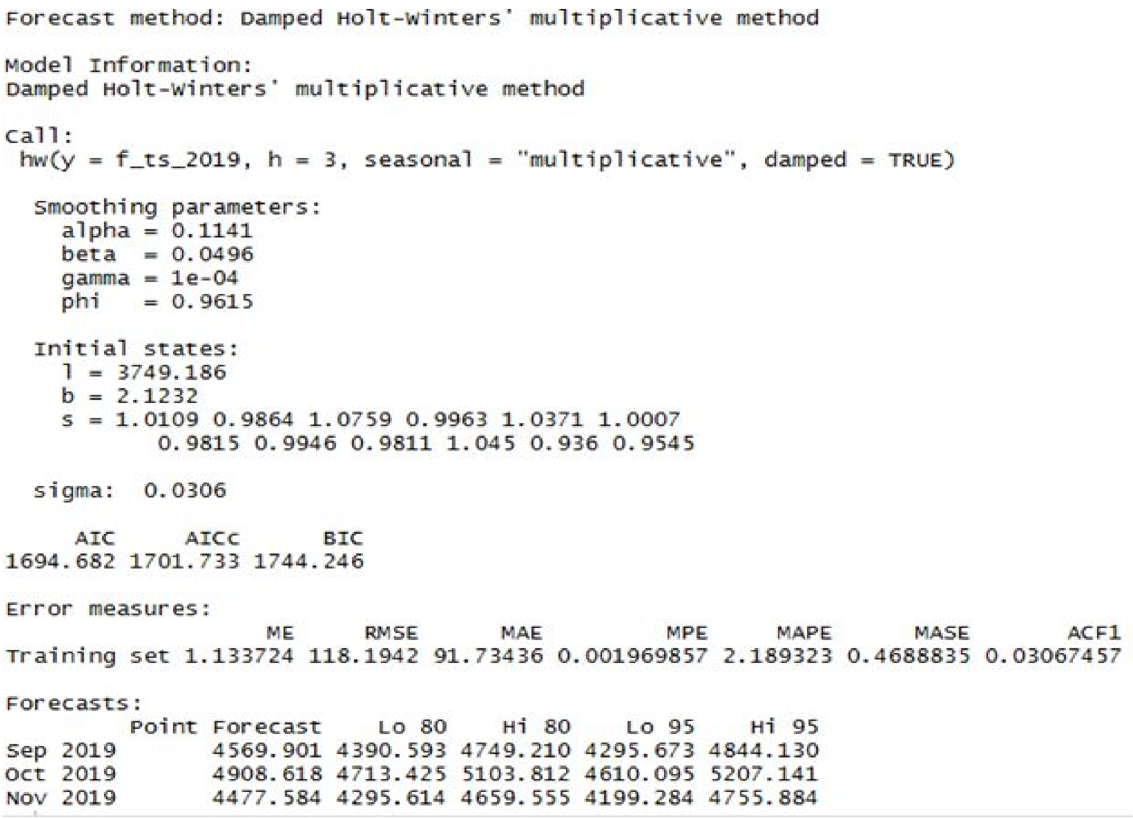


```
Model Information:
Damped Holt-Winters' additive method

Call:
 hw(y = f_ts_2019, h = 3, seasonal = "additive", damped = TRUE)

  Smoothing parameters:
    alpha = 0.1383
    beta  = 0.0469
    gamma = 1e-04
    phi   = 0.9596

  Initial states:
    l = 3756.1039
    b = 1.4419
    s = 54.7384 -52.4294 312.4396 -11.0963 188.5093 -19.5039
           -77.9093 -27.9517 -67.9103 182.1893 -276.328 -204.7478

  sigma:  130.1352

     AIC     AICc      BIC 
1698.543 1705.595 1748.108 

Error measures:
                   ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set 1.585047 120.2218 94.31563 0.01709785 2.248324 0.4820772 0.02696674

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Sep 2019       4569.714 4402.939 4736.489 4314.654 4824.774
Oct 2019       4870.631 4701.078 5040.184 4611.322 5129.940
Nov 2019       4484.056 4310.348 4657.764 4218.393 4749.719
```

Exhibit 6.16 Graph and Summary of Damped Holt-Winter's additive model

(g) Damped Holt-Winter's Multiplicative method: This model assumes that there is a damped trend and multiplicative seasonality in the time series.

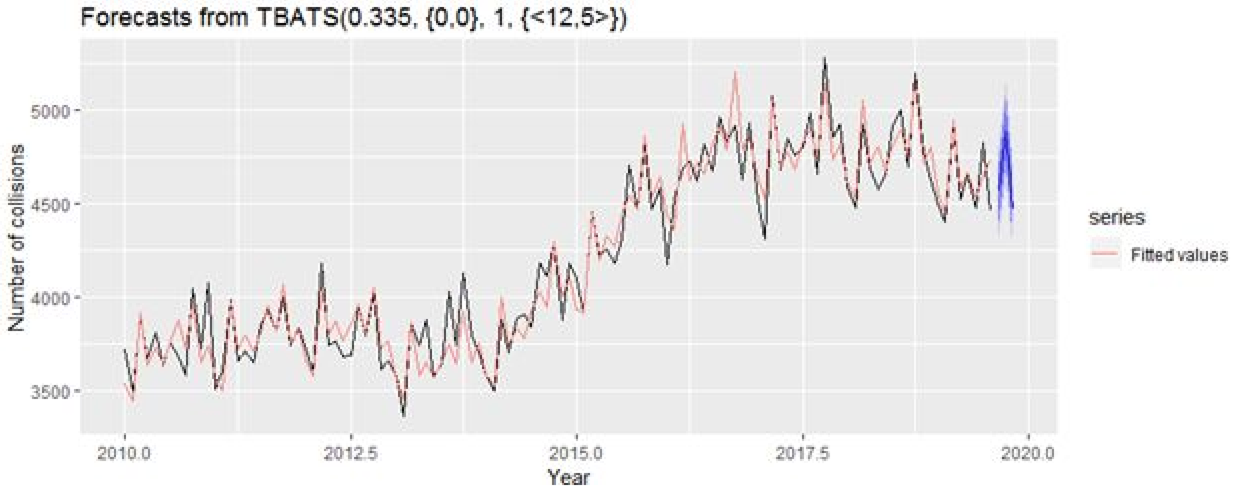


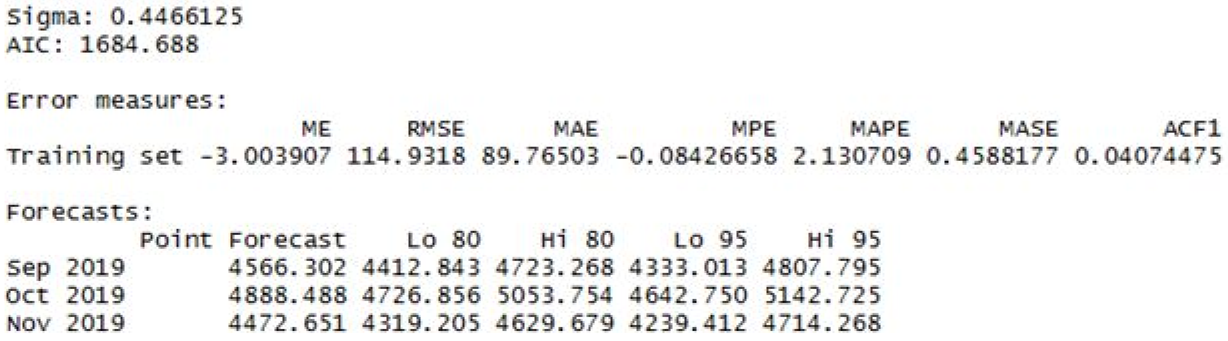


```
Forecast method: Damped Holt-Winters' multiplicative method

Model Information:
Damped Holt-Winters' multiplicative method

Call:
 hw(y = f_ts_2019, h = 3, seasonal = "multiplicative", damped = TRUE)

  Smoothing parameters:
    alpha = 0.1141
    beta  = 0.0496
    gamma = 1e-04
    phi   = 0.9615

  Initial states:
    l = 3749.186
    b = 2.1232
    s = 1.0109 0.9864 1.0759 0.9963 1.0371 1.0007
           0.9815 0.9946 0.9811 1.045 0.936 0.9545

  sigma:  0.0306

     AIC     AICc      BIC
1694.682 1701.733 1744.246

Error measures:
                   ME     RMSE      MAE         MPE     MAPE      MASE       ACF1
Training set 1.133724 118.1942 91.73436 0.001969857 2.189323 0.4688835 0.03067457

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Sep 2019       4569.901 4390.593 4749.210 4295.673 4844.130
Oct 2019       4908.618 4713.425 5103.812 4610.095 5207.141
Nov 2019       4477.584 4295.614 4659.555 4199.284 4755.884
```

Exhibit 6.17 Graph and Summary of Damped Holt-Winters multiplicative model

(h) TBATS method: This method is completely automated method and it is a modified exponential smoothing state space model. TBATS stands for Trigonometric seasonality, Box-Cox transformations, ARMA models for residuals, Trend and Seasonality.

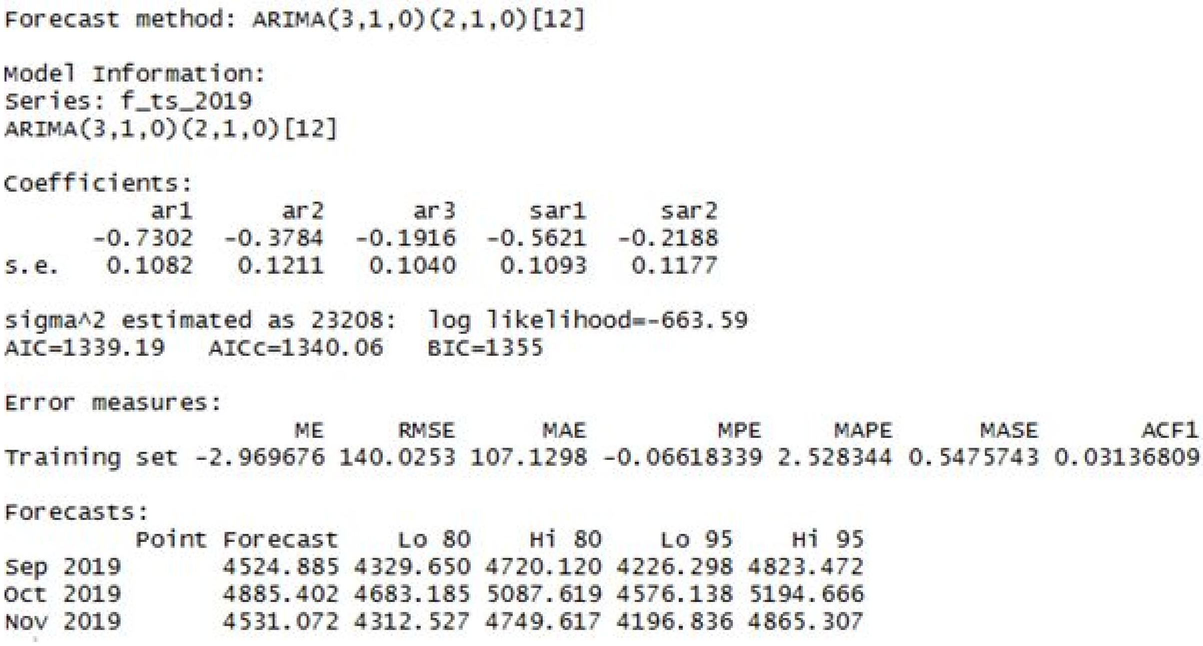




```
Sigma: 0.4466125
AIC: 1684.688

Error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE       ACF1
Training set -3.003907 114.9318 89.76503 -0.08426658 2.130709 0.4588177 0.04074475

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Sep 2019       4566.302 4412.843 4723.268 4333.013 4807.795
Oct 2019       4888.488 4726.856 5053.754 4642.750 5142.725
Nov 2019       4472.651 4319.205 4629.679 4239.412 4714.268
```

Exhibit 6.18 Graph and Summary of TBATS model

(i) ARIMA model: ARIMA models is technically sophisticated way of forecasting a time series variable by looking only at the past patterns of the time series. They do by exploiting the autocorrelation structure of the time series.



```
Forecast method: ARIMA(3,1,0)(2,1,0)[12]

Model Information:
Series: f_ts_2019
ARIMA(3,1,0)(2,1,0)[12]

Coefficients:
          ar1      ar2      ar3     sar1     sar2
      -0.7302  -0.3784  -0.1916  -0.5621  -0.2188
s.e.   0.1082   0.1211   0.1040   0.1093   0.1177

sigma^2 estimated as 23208:  log likelihood=-663.59
AIC=1339.19   AICc=1340.06   BIC=1355

Error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE       ACF1
Training set -2.969676 140.0253 107.1298 -0.06618339 2.528344 0.5475743 0.03136809

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Sep 2019       4524.885 4329.650 4720.120 4226.298 4823.472
Oct 2019       4885.402 4683.185 5087.619 4576.138 5194.666
Nov 2019       4531.072 4312.527 4749.617 4196.836 4865.307
```

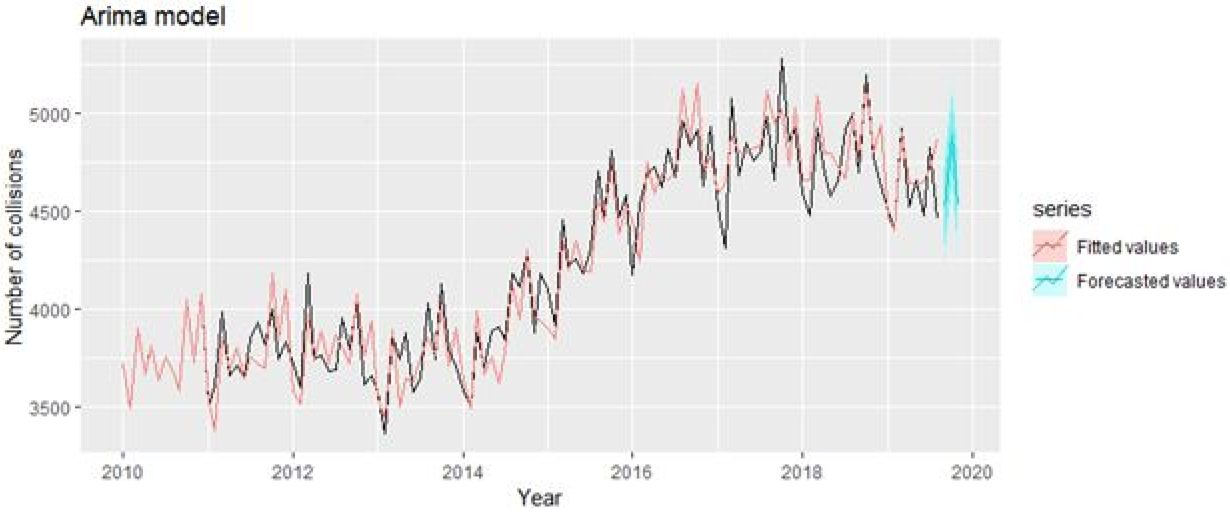
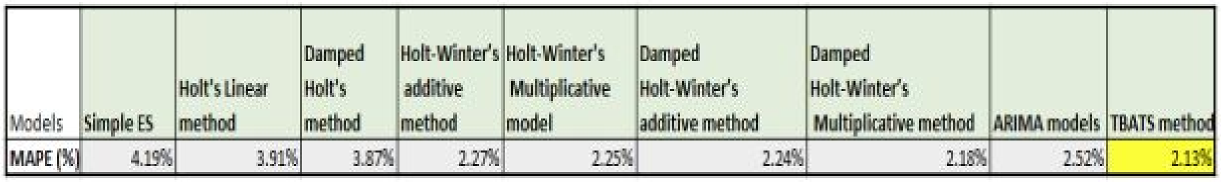


Exhibit 6.19 Graph and Summary of ARIMA model

5. *Result from forecast:*



Models	Simple ES	Holt's Linear method	Damped Holt's method	Holt-Winter's additive method	Holt-Winter's Multiplicative model	Damped Holt-Winter's additive method	Damped Holt-Winter's Multiplicative method	ARIMA models	TBATS method
MAPE (%)	4.19%	3.91%	3.87%	2.27%	2.25%	2.24%	2.18%	2.52%	2.13%

Exhibit 6.20 Performance evaluations using MAPE of all forecasting models

In conclusion, TBAT model has the lowest MAPE. Thus, it is the best model for forecasting the number of collisions for next three months. This model makes sense because the frequencies of the seasonality increases over time in data. Also, this method works fine for short-term predictions.

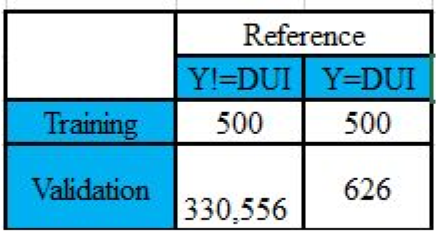
VI. Predict DUI (Driving under influence)

1. *Random forest*

Random forest combines several trees for better performance, which is one of the ensemble methods. it uses bootstrap which draws multiple random samples with a replacement that selected data can be repeated, and fits each sample for separate model and compute (average) the predictions to obtain enhanced prediction result. For the classification, the final prediction selected with the majority vote among models. Unlike a single tree, the results from a random forest can not be displayed like a dendrogram; it only provides the variable importance scores which measure the relative contribution of different predictors.

The purpose of this test is to predict DUI (driving under influence) using MO Codes variable, which contains 3038 (DUI felony) and 3039 (DUI misdemeanor). Since there were only a few DUI cases (DUI 0.34% of total 332,182 data set), under-sampling was used for the test. If the original data used for the training, it may not detect y=1

(DUI) data well since y=0 (non-DUI) dominates the training data set. We split the selected independent variables, which are "Time Occurred", "Victim Age" in a range of 0, 20, 35, 50, 100 respectively, "Victim Sex", "Victim Descent" and "Council Districts". Among those independent variables, the "Time Occurred" variable was rescaled within a range from 0 to 1, and other ones were transformed into dummy variables.



	Reference	
	Y!=DUI	Y=DUI
Training	500	500
Validation	330,556	626

Exhibit 6.21: Frequency table of the output label, DUI or not.

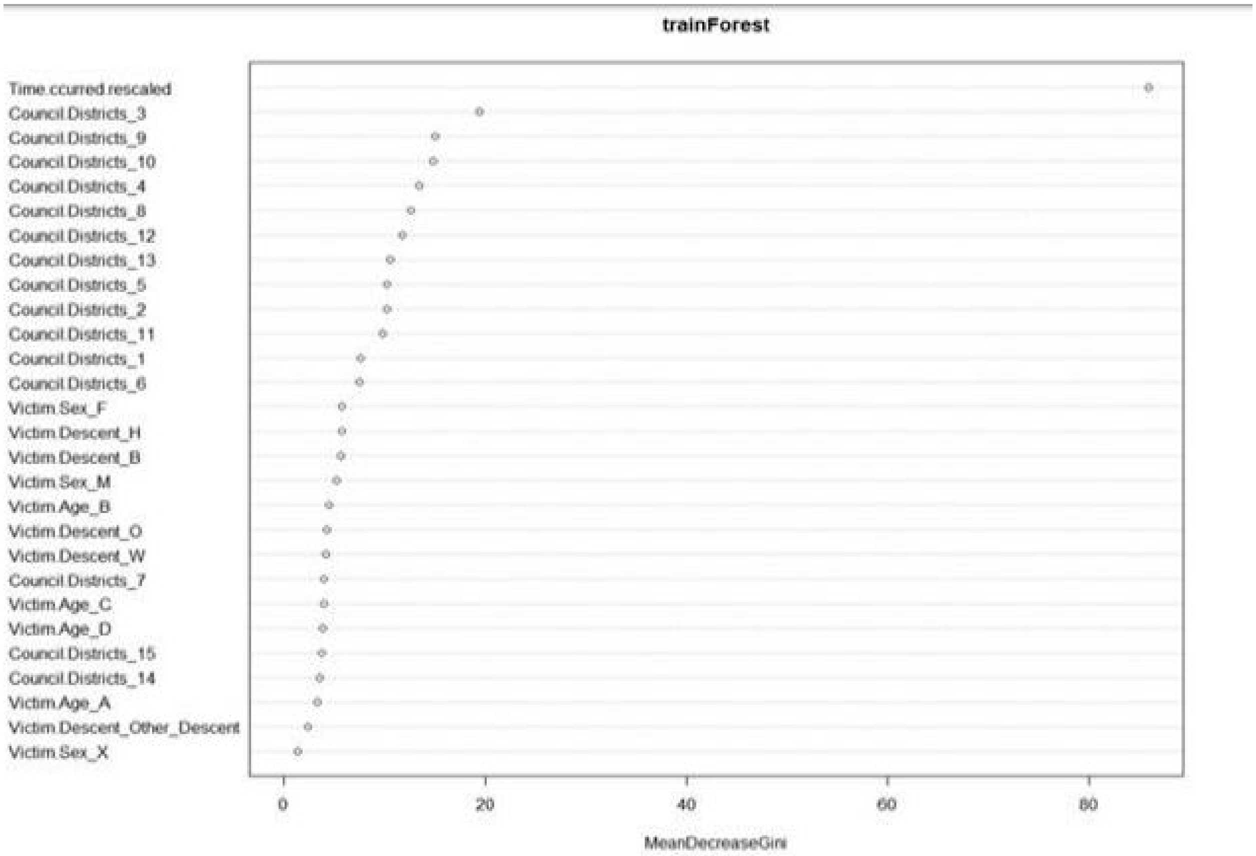
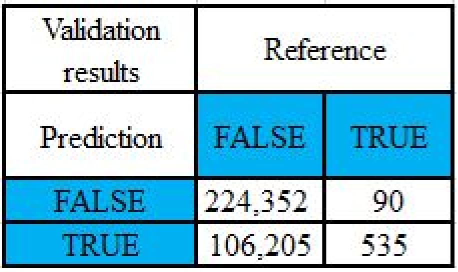


Exhibit 6.21: Significant variables from random forest



Validation results	Reference	
Prediction	FALSE	TRUE
FALSE	224,352	90
TRUE	106,205	535

Exhibit 6.22: Confusion Matrix from Random Forest Model

From the chart of Variable Importance, the "Time Occurred" variable dominated the data. The differences among other variables are not significant. The accuracy, sensitivity, and specificity of Random Forest are 67.90%, 85.60%, and 67.87% respectively.

2. Up sampling and SMOTE

Since my target data sample (1126, 0.34%) was very small compared to the whole data, we tried upsampling and SMOTE sampling to make larger training data. Up-sampling randomly replicates instances in the minority class. Synthetic minority sampling technique (SMOTE) decreases samples of the majority class and synthesizes new minority instances by interpolating between existing ones.

Up-sampling: Produced 9985 TRUE and selected 9818 FALSE data for training.

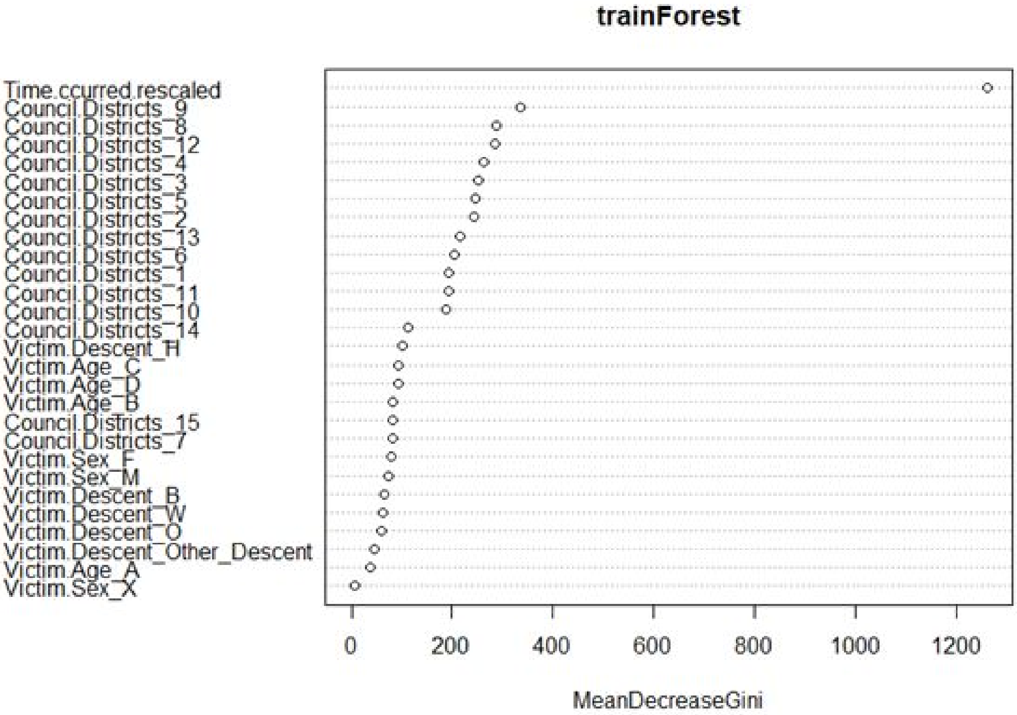
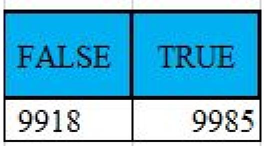
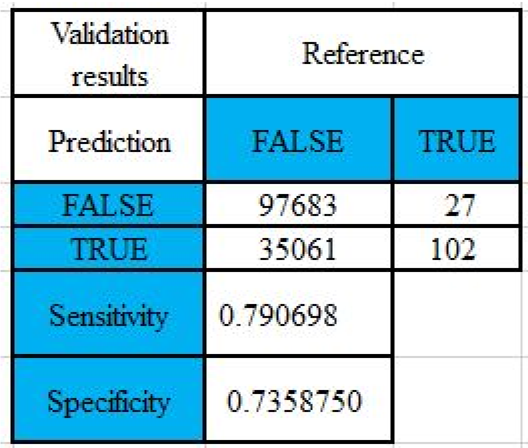


Exhibit 6.23: Significant variables from Up-Sampling



FALSE	TRUE
9918	9985

Exhibit 6.24: Frequency table from Up-Sampling



Validation results	Reference	
Prediction	FALSE	TRUE
FALSE	97683	27
TRUE	35061	102
Sensitivity	0.790698	
Specificity	0.7358750	

Exhibit 6.25: Confusion Matrix and Statistics from Up-Sampling

The sensitivity rate for the of the up-sampling was 6.6% lower than the under-sampling model. There were 273 y=1 (DUI) in the original training data, and the size of y=1 data increased about 30 times the scale of the x-axis for the up-sampling increased nearly 14 times.

SMOTE: produced 819 TRUE data and selected 1092 FALSE data for training.

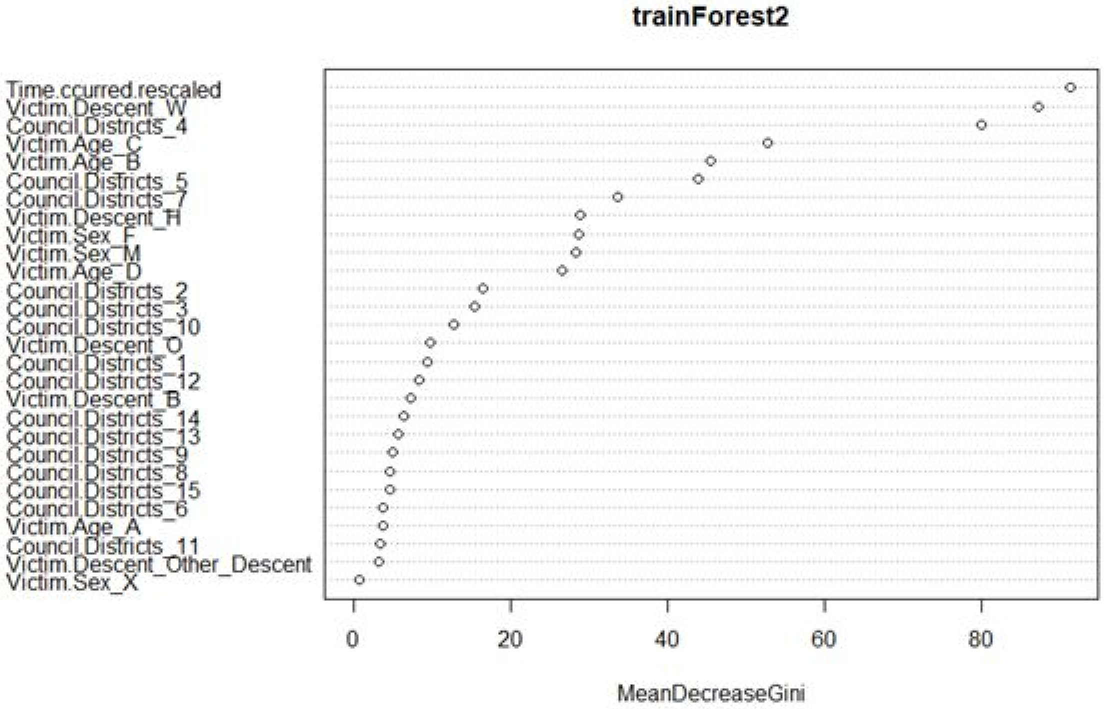
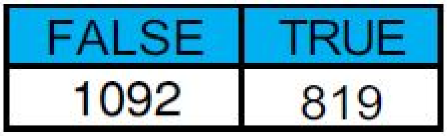
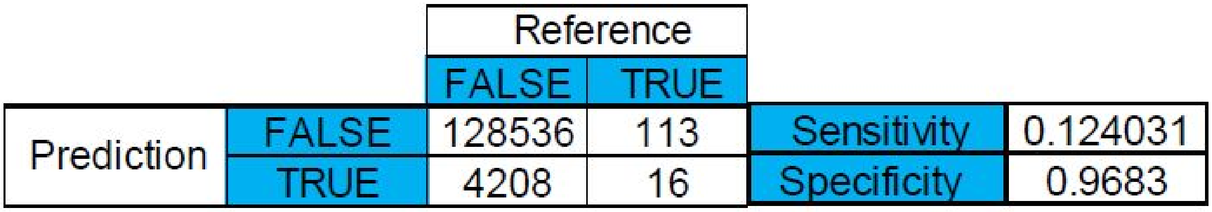


Exhibit 6.26: Significant variables from SMOTE



FALSE	TRUE
1092	819

Exhibit 6.26: Frequency table from training data for SMOTE



		Reference			
		FALSE	TRUE		
Prediction	FALSE	128536	113	Sensitivity	0.124031
	TRUE	4208	16	Specificity	0.9683

Exhibit 6.27: Confusion Matrix and Statistics from SMOTE

Even though the SMOTE increased specificity a lot from 67.87% (under-sampling) to 96.83%, the sensitivity of SMOTE was very low as 12.4%. Since the purpose of this research is to predict the target y=1, SMOTE is not recommended. From the importance chart, the pattern of the distribution became distorted than the up-sampling. It can be inferred that producing instances by interpolating between existing values produced more distortion than the up-sampling for this distortion for this case.

3. XGBoost

XGBoost is a type of Boosting model which gives higher selection probabilities to misclassified records. The most important feature of XGBoost is the capability of managing sparse data; it stores data without storing zeros that can save memory and time. It has a distributed weighted quantile sketch algorithm to effectively handle weighted data. And it can execute multiple threading that has the effect of running multiple machines. Finally, it can handle missing data and use for the regression as well.

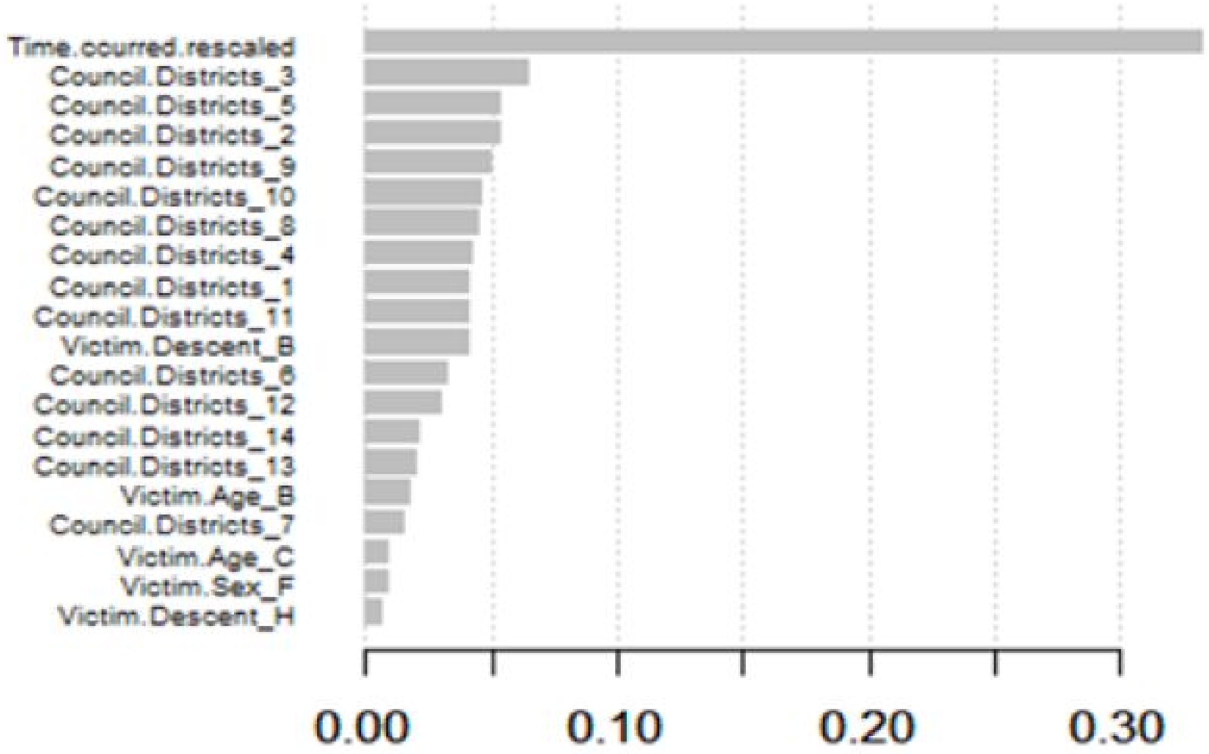
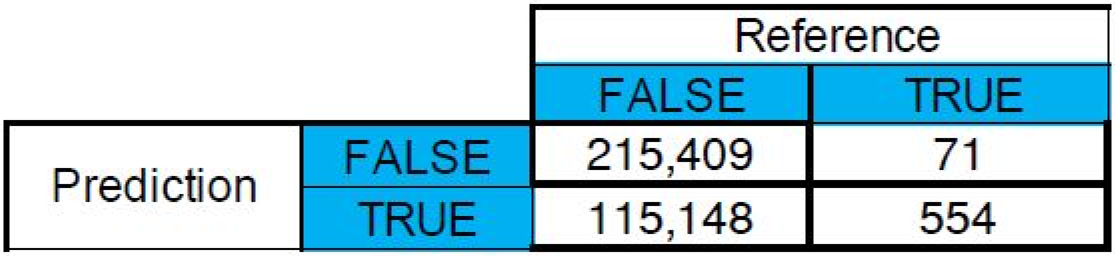


Exhibit 6.28: Significant variables from XGBoost



		Reference	
		FALSE	TRUE
Prediction	FALSE	215,409	71
	TRUE	115,148	554

Exhibit 6.29: Confusion Matrix from XGBoost

The results of XGBoost are Accuracy 65.21%, Sensitivity 88.64%, Specificity 65.16%. The Variable Importance chart of XGBoost showed a very similar pattern with the random forest since it uses tree method as well. Because XGBoost is the most enhanced model among tree methods, it had a 3% higher sensitivity compared to random forest. However, XGBoost had a lower specificity of about 2.7% (8943 data more False-true) than the random forest. Therefore, to choose the better model, if the LA transportation department thinks to search 8943 data (115,148 XGBoost False positive-106,205 random forest) more than the random forest model is worth searching 19 cases (554 XGBoost True positive -535 random forest) of DUI then they can select a XGBoost model to predict the DUI case.

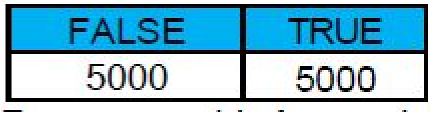
From the importance variable chart, the time occurred, and locations such as district 3 are important factors. It can suggest that LA CITY can strengthen monitor of DUI for district 3 at late hours to prevent DUI cases.

4. Predict Hit and Run Felony

The purpose of this test is to predict Hi and Run Felony using MO Codes variable, the cleaned data set contained 3029 Hit and Run Felony. Since there were few Hit and Run Felony cases (29259, 8.8% of total 332,182 data set) in the data set, under-sampling was used for the test.

(a) Random forest (Predict Hit and Run Felony)

y='hit and run' 5000 and y ! hit and run' 5000, 10000 data set in the training data.



FALSE	TRUE
5000	5000

Exhibit 6.30: Frequency table from training data for Random Forest

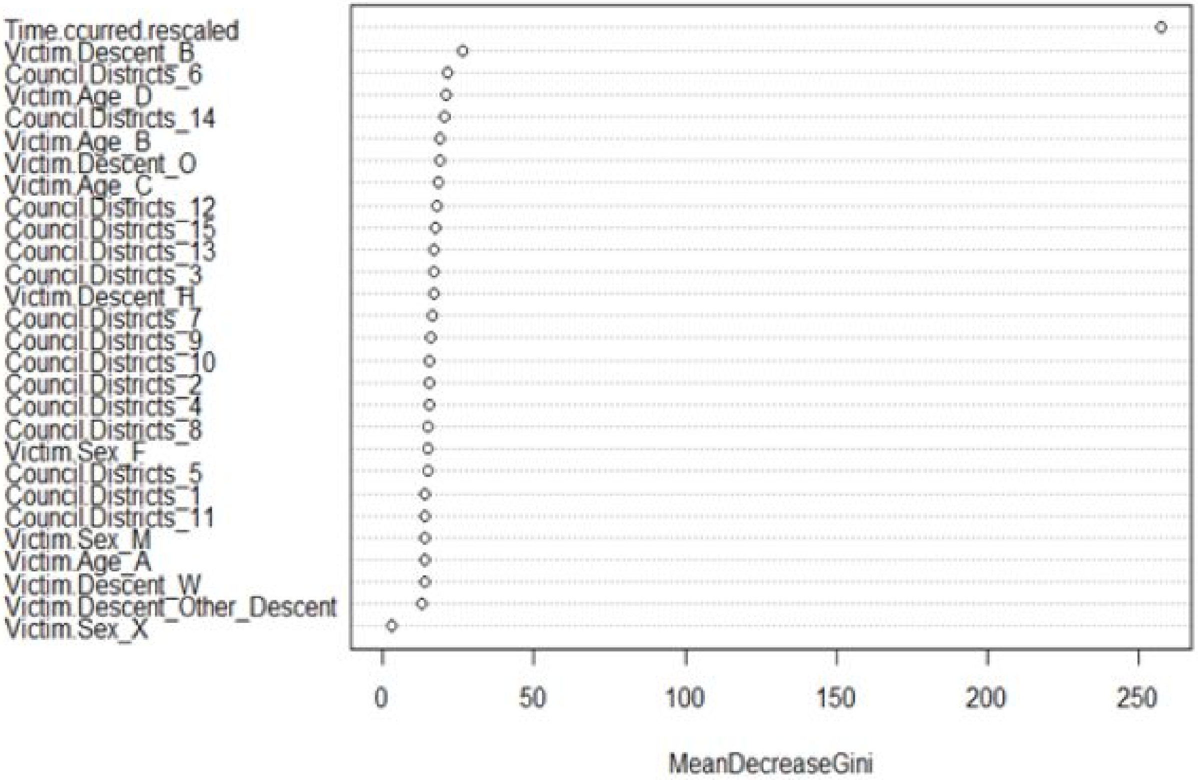


Exhibit 6.31: significant variables from Random Forest to predict hit and run

From the importance variable chart, time was the dominant factor, and the victim descent B was the second important variable for predicting hit and run felony prediction.

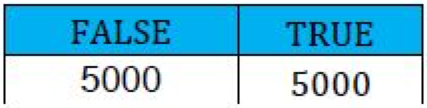
		test							
		0	1	0	1	0	1	0	1
	Cut off	0.5		0.45		0.4		0.3	
y_predict	0	166107	10866	145637	9195	124599	7599	83311	4758
	1	131816	13393	152286	15064	173364	16660	214612	19501

Cut off	Sensitivity	Specificity
0.5	0.55208	0.55755
0.45	0.62097	0.48884
0.4	0.68676	0.41809
0.3	0.8039	0.2796

Exhibit 6.32: Confusion matrix, sensitivity and specificity with different cut-off values

Unlike to predict the DUI test, the sensitivity of “hit and run felony” was only 55% (compared to the sensitivity of random forest predicting DUI 85%). Therefore to predict hit and run felony's cut off 0.45 or 0.4 can be selected for random forest model.

(b) XGBoost (Predict Hit and Run Felony)



FALSE	TRUE
5000	5000

Exhibit 6.33: Frequency table from training data for XGBoost for hit and run

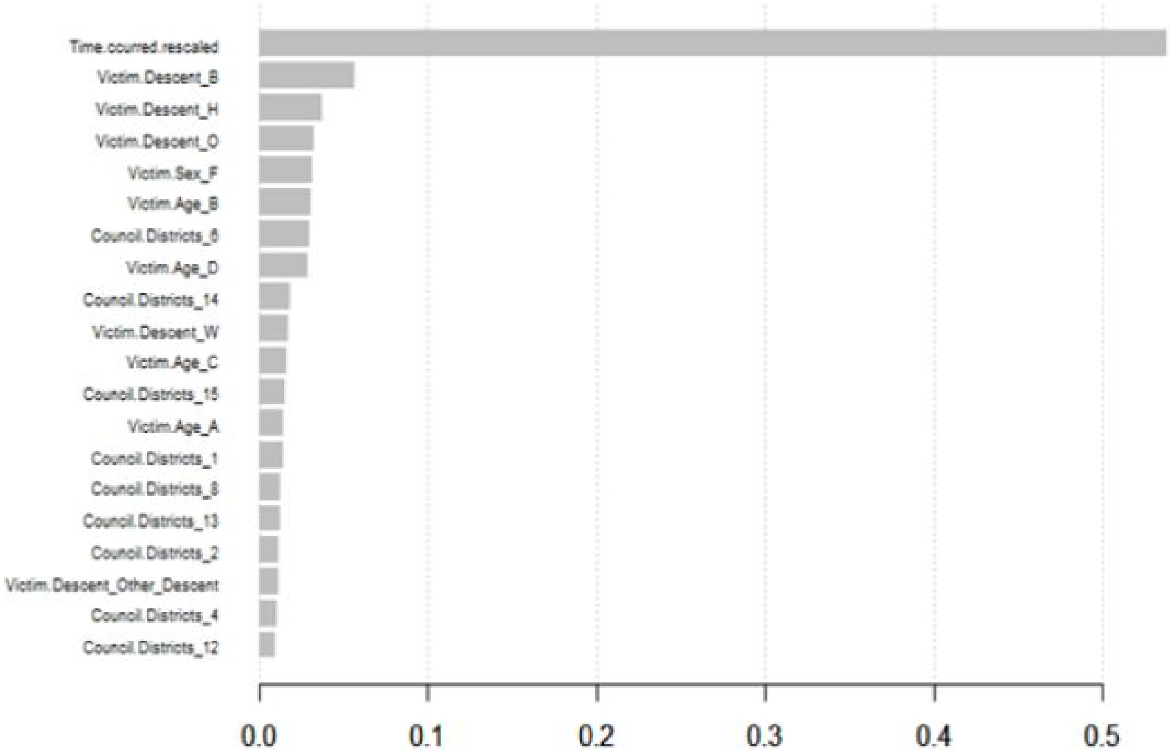
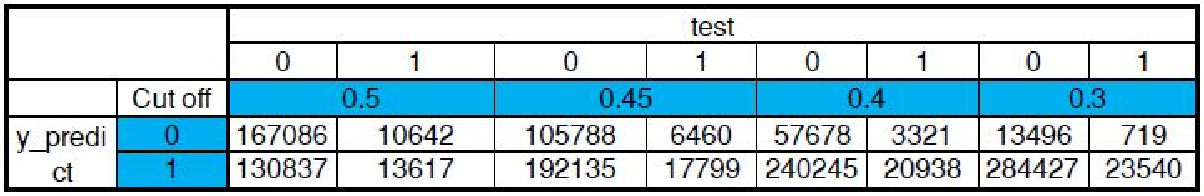
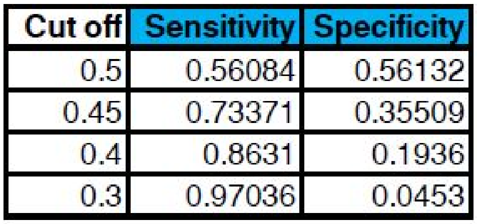


Exhibit 6.35: significant variables from XGBoost to predict hit and run.

From the Variable Importance chart, time was the dominant factor, and the victim descent B was the second important variable for predicting hit and run felony prediction similar to the random forest method. However, compared to the random forest, the council district ranked lower than the random forest model.



		test							
		0	1	0	1	0	1	0	1
	Cut off	0.5		0.45		0.4		0.3	
y_predict	0	167086	10642	105788	6460	57678	3321	13496	719
	1	130837	13617	192135	17799	240245	20938	284427	23540



Cut off	Sensitivity	Specificity
0.5	0.56084	0.56132
0.45	0.73371	0.35509
0.4	0.8631	0.1936
0.3	0.97036	0.0453

Exhibit 6.36: Confusion matrix,sensitivity and specificity with different cut-off values.

Despite the XGBoost model return much higher sensitivity result random forest(at cut off 0.45 random forest 0.62, XGBoost 0.73) , its specificity was much lower than the random forest. This means that to detect the true hit and run felony case, the XGBoost model should predict more cases for the hit and run felony for false positive. For 0.45 and 0.4, the test produced higher returns within the cutoff values. To prevent hit and run felony, LA transport department can emphasize education about the crime, especially for the upper ranked groups, which are Decent Black, Hispanic, and female, and age 20 to 35 range than other groups.

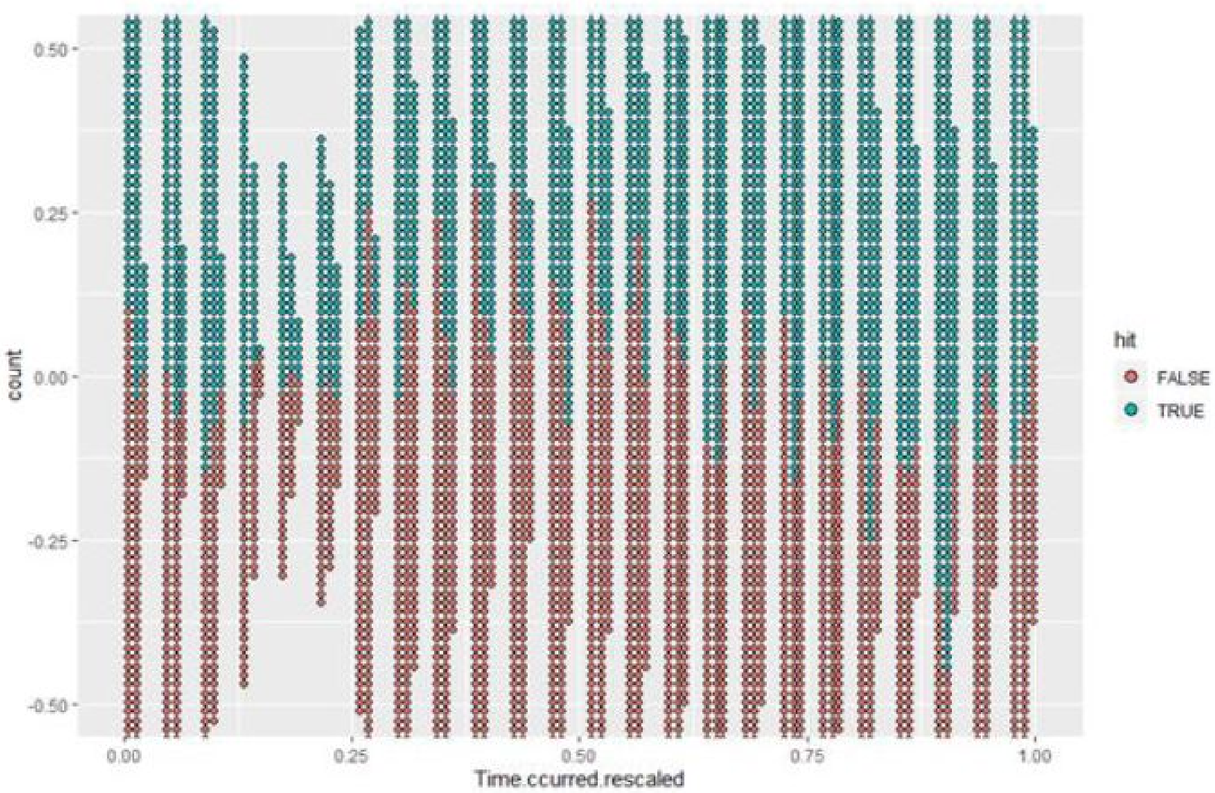


Exhibit 6.37: "Time Occurred' distribution

The hit and run felony was evenly spread across time; it can be explained why the sensitivity was lower using the time variable as a dominant factor. Since the dominant variable (time occurred rescaled) looked hard to split vertically for prediction, the tree-related method may not work well for this type of distribution.

VII. Clustering

Clustering is unsupervised learning (no answers are given). The goal of clustering is to segment the data into similar clusters to generate insight. Clustering is popular for business applications such as customer segmentation for industry analysis. For this test, the k-means method used to assign k clusters to minimize dispersion within the cluster using Euclidean distances.

The object of clustering research is to find the distinctive traffic law violations for each cluster group. First, it found that clustering five groups for the data set gives a reasonable sum of within-cluster distance. Second, it searched traffic law violations for each five groups. Last, it showed relative traffic law violations against population data set (% of law violation for the individual groups subtracted by % of law violation of total population). The variables for clustering were Victim.Sex, Victim.Age, Victim.Decsent, which contains demographic characteristics.

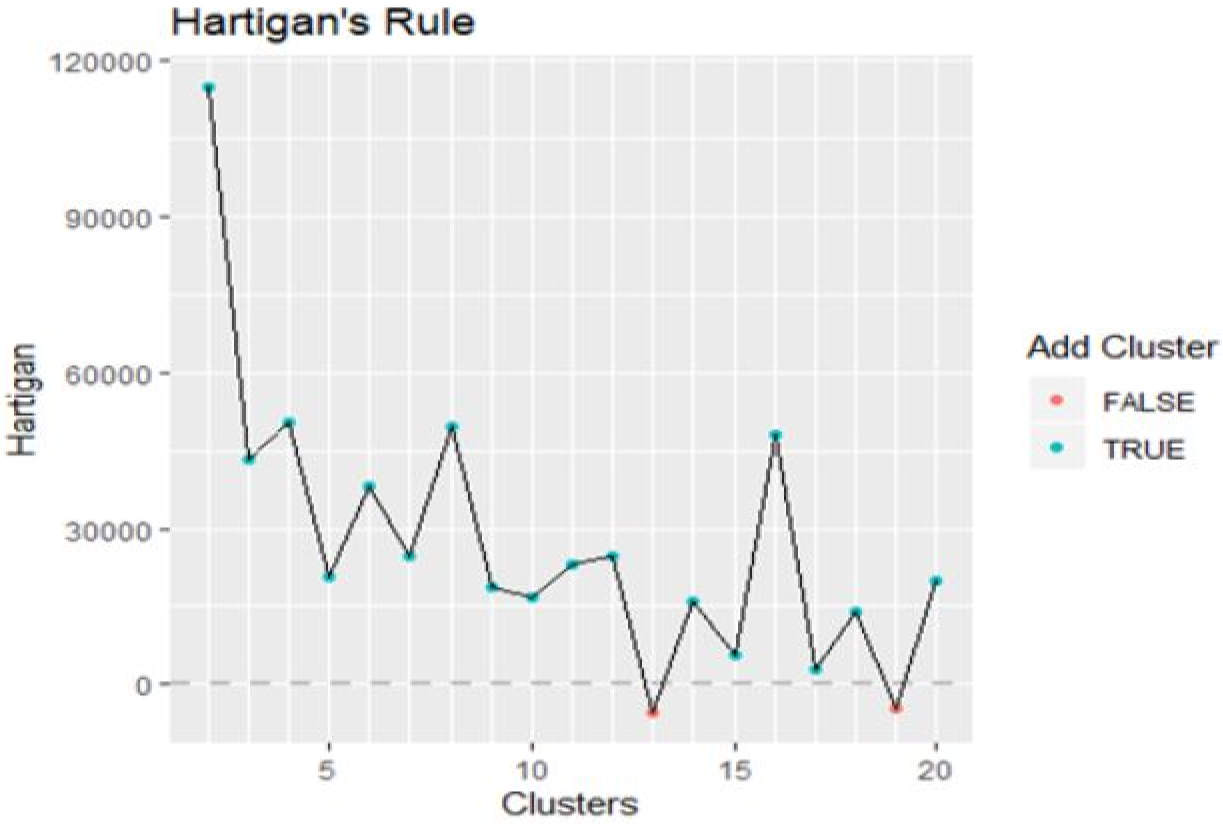


Exhibit 6.38: Graph of Hartigan's rule

Hartigan's rule compares the values of the within-cluster sum of squares for a clustering to 5 groups gives a reasonable low value of the within-cluster sum of squares.

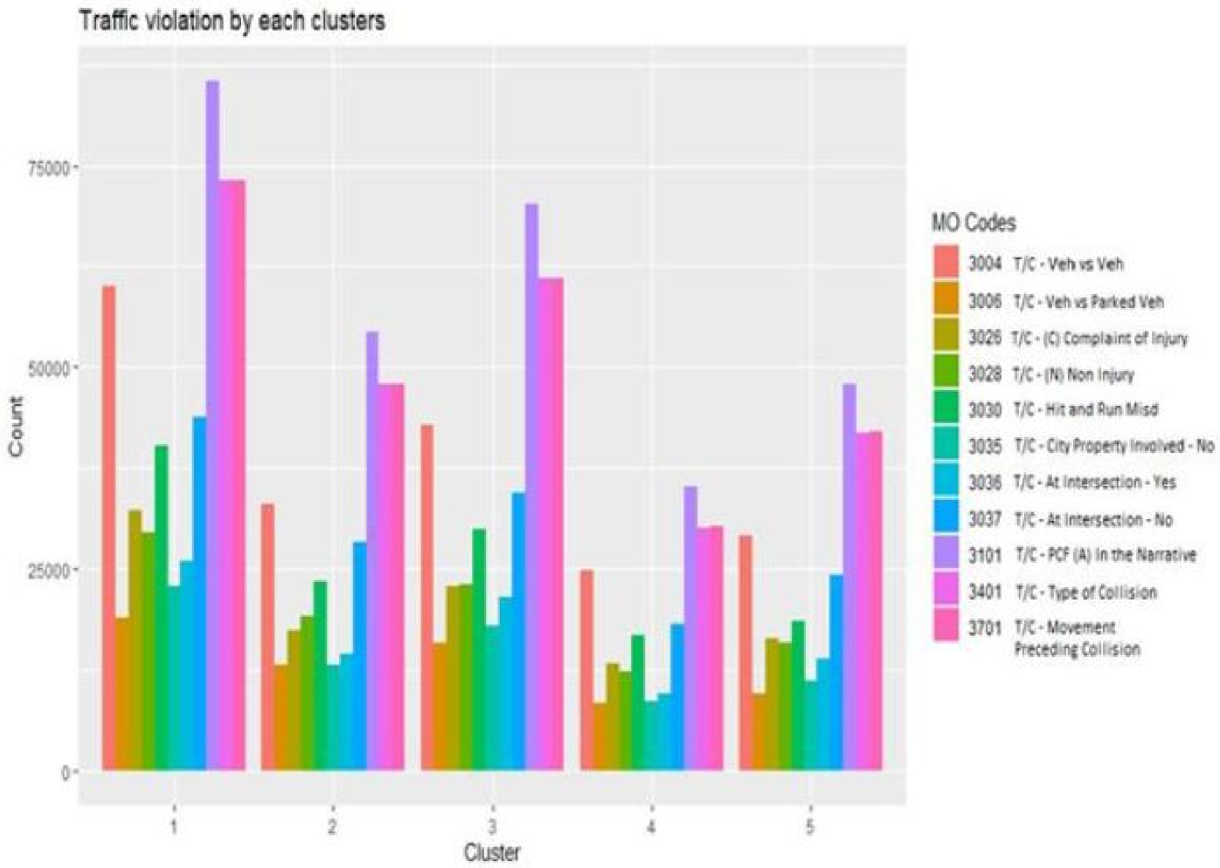
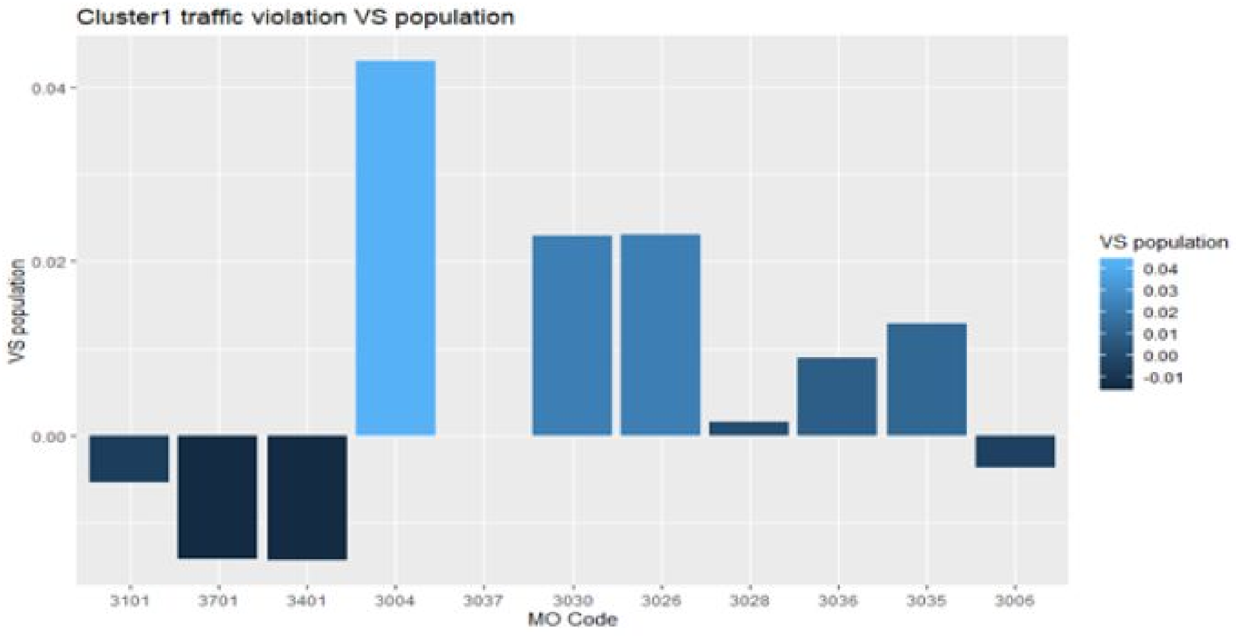
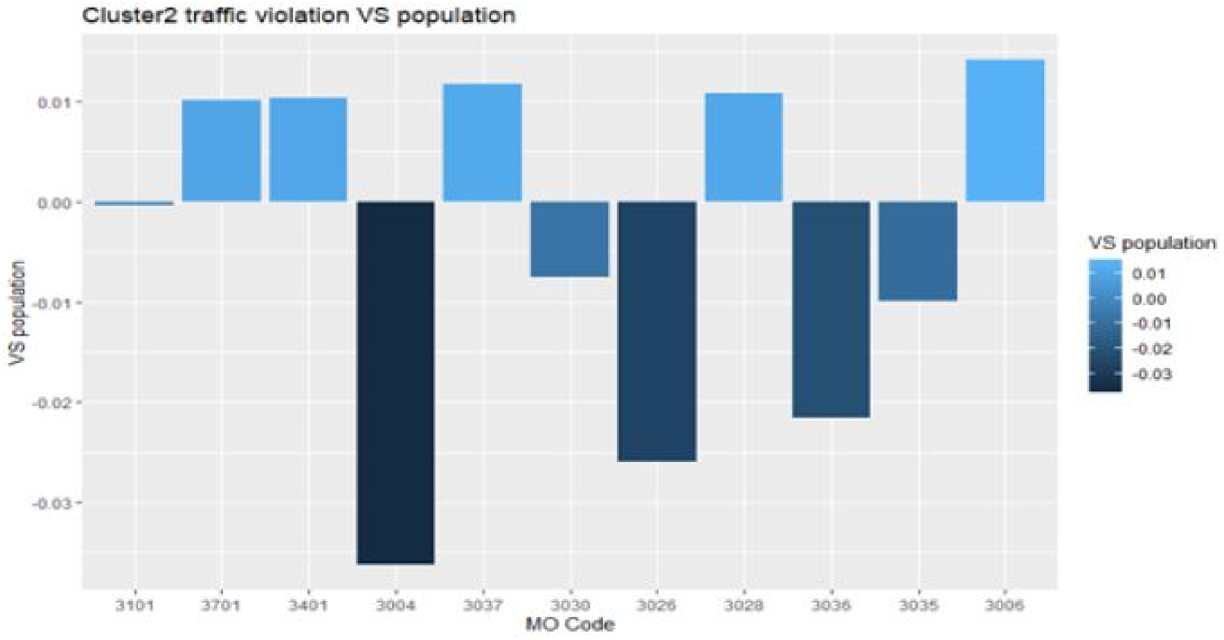


Exhibit 6.39: Traffic violation by each cluster

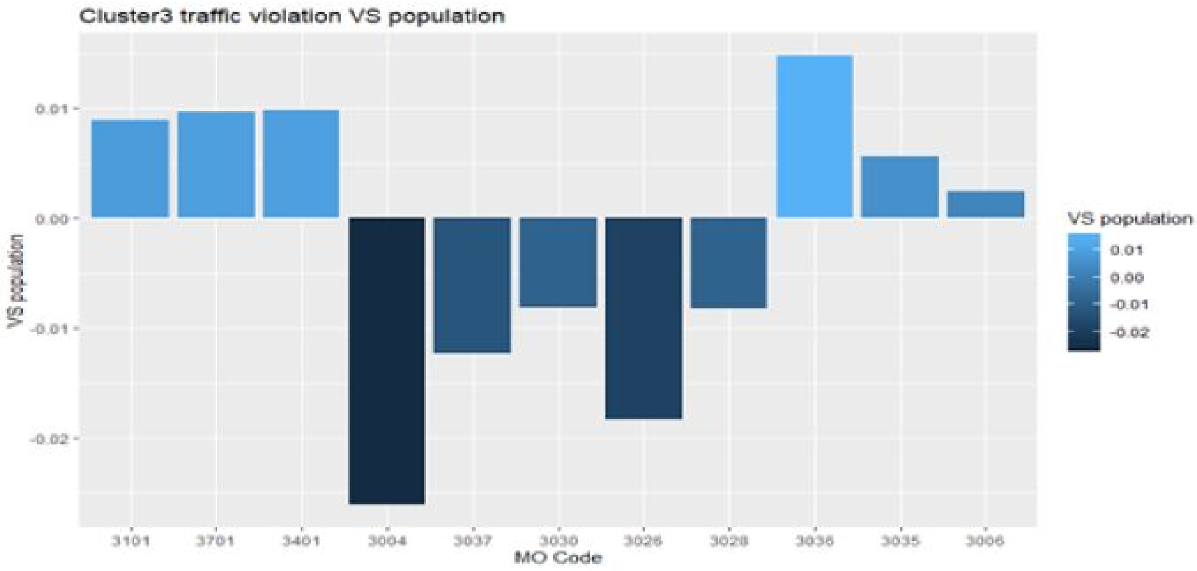
The 3004 (T/C vehicle vs vehicle), 3401(T/C type of Collision), 3701 (T/C Movement preceding Collision) were the most common traffic violations.

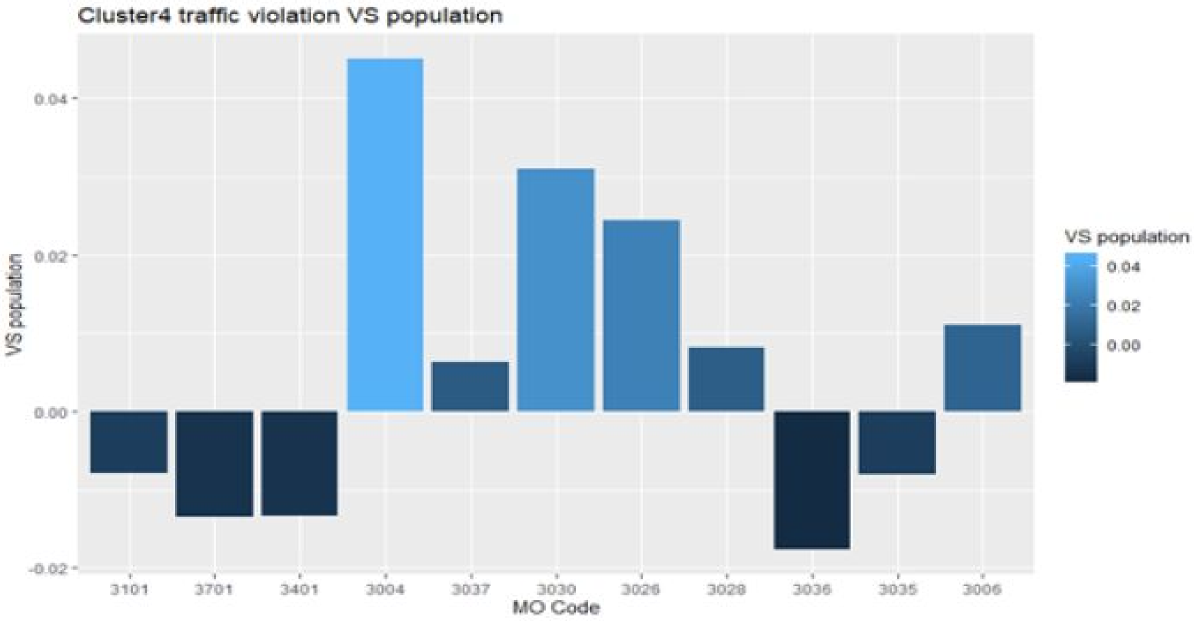


Cluster1 traffic violation VS population



Cluster2 traffic violation VS population


Cluster3 traffic violation VS population
VS population
0.01
0.00
-0.01
-0.02
3101
3701
3401
3004
3037
3030
3026
3028
3036
3035
3006
MO Code
VS population
0.01
0.00
-0.01
-0.02


Cluster4 traffic violation VS population
VS population
0.04
0.02
0.00
-0.02
3101
3701
3401
3004
3037
3030
3026
3028
3036
3035
3006
MO Code
VS population
0.04
0.02
0.00

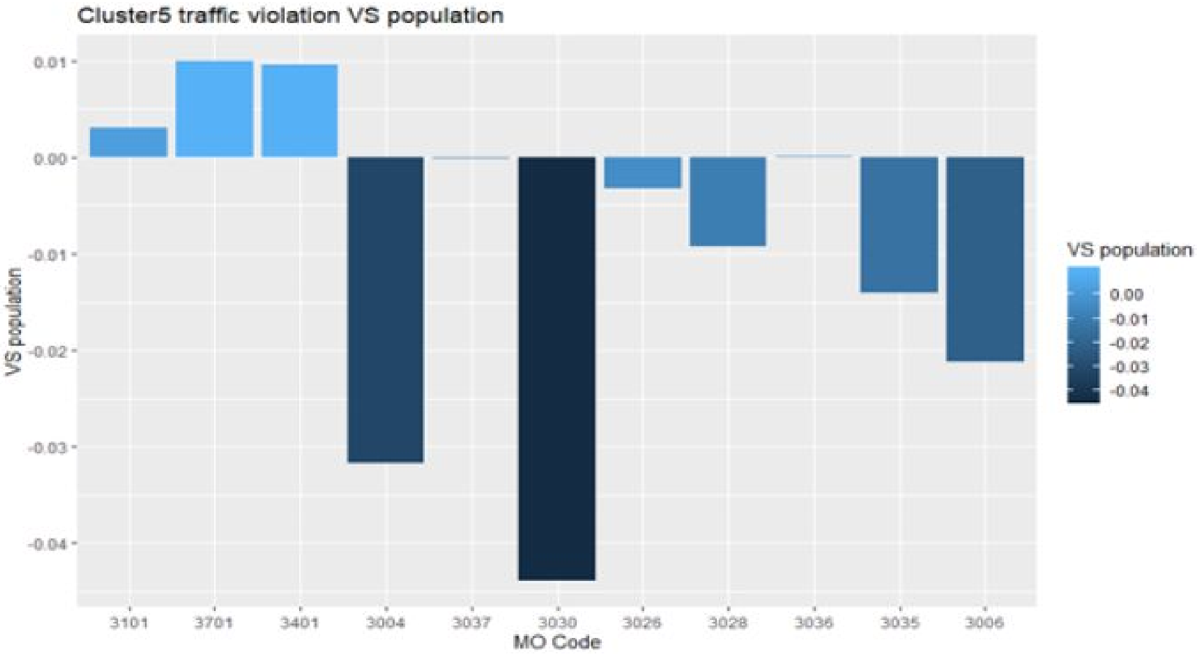


Exhibit 6.40: Various clusters for violation vs population

Percentage of traffic violation by cluster VS percentage of violation by population data set. The most frequently violated law versus population are as below:

1. Cluster 1: age 50-100, Female, Black 3004 (T/C - Veh vs Veh)
2. Cluster 2: age 35-50, Male, Hispanic 3006 (T/C - Veh vs Parked Veh)
3. Cluster 3: age 20-35, Male, Hispanic 3036 (T/C - At Intersection Yes)
4. Cluster 4: age 35-50, Female, O (race-Unknown) 3004 (T/C - Veh vs Veh)
5. Cluster 5: age 50-100, Male, White 3701 (T/C - Movement Preceding Collision)

For example, the LA transportation department can give additional warning for driving in the parking lot to cluster 2 members who is the age range 35-50, Male, and Hispanic to prevent 3006 Vehicle versus parked vehicles.

For further research, separating a dataset into clusters is useful for improving performance of supervised methods by modeling each cluster separately. For instance, random forest and XGboost test in the previous section can proceed for each 5 cluster.

VIII. Conclusion

Our team has used three classification prediction models (Logistic Regression, kNN, and Classification Tree) to predict whether an accident will happen on a day off. Classification Tree provides the best accuracy rate with the simplest model. If the time is from midnight to 5:04 am, the accident is more likely to happen on a day off. The increase probability of accident occurrence at night on a day off happens because people usually go to party, drink alcohol, sleep late and/or their night vision is bad. The California Office of Traffic Safety (OTS) can inform people to not go out too late on the weekend.

For prediction, XGBoost was a better model for finding higher sensitivity. However, the specificity was lower than random forest means that it has more false-positive data. For data sampling, three methods used, under-sampling, up-sampling, and SMOTE, under-sampling returned the most precise prediction in the models. Clustering method used for grouping by demographic characteristics, there was much variation for the frequent accidents among groups, LA city can provide customized education by each cluster for preventing the accident effectively.

IX. References

Alice, Michy (2015, October 13). *How to Perform a Logistic Regression in R*. Retrieved October 13, 2015, from https://datascienceplus.com, https://datascienceplus.com/per
form-logistic-regression-in-r/

ALTERYX, INC. (2004 - 2019). Alteryx Designer (2019.1.4.57073). Retrieved from https://www.alteryx.com/why-alteryx/alteryx-for-good/students.

Bartlett. (2016). Better Decisions Demand Forecast Accuracy - Forecast Pro. Retrieved October 13, 2019, from Forecast Pro website: https://www.forecastpro.com/

Charpentier, Arthur (2013, September 26). *Logistic Regression and Categorical Covariates.* Retrieved from https://www.r-bloggers.com/logistic-regression-and-categorical-covariates/

ggplot2 package | R Documentation. (2019). Retrieved from Rdocumentation.org website: https://www.rdocumentation.org/packages/ggplot2/versions/3.2.1

glm function | R Documentation. (2019). Retrieved from Rdocumentation.org website: https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/glm

Hilbe, Joseph M., & Hilbe, Joseph M. (2009). Logistic regression models (A Chapman & Hall Book). Hoboken: CRC Press.

Hyndman, R. J., & Athanasopoulus, G. (2018, April). Forecasting: Principles and Practice. Retrieved from https://otexts.com/fpp2/.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). Chapter 4.3: Logistic Regression. An Introduction to Statistical Learning with Application in R(1st ed., pp.130-138). New York: Springer Science+Business Media, LLC.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). Chapter 4.6.5: K-Nearest Neighbors. An Introduction to Statistical Learning with Application in R(1st ed., pp.163-164). New York: Springer Science+Business Media, LLC.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). Chapter 10.3: Clustering Methods. An Introduction to Statistical Learning with Application in R(1st ed., pp.385-401). New York: Springer Science+Business Media, LLC.

Kaggle (2010). Los Angeles Traffic Collision Data. Retrieved from Kaggle.com website: https://www.kaggle.com/cityofLA/los-angeles-traffic-collision-data

Le, James (2018, April 10). *Logistic Regression in R Tutorial.* Retrieved from https://www.datacamp.com/community/tutorials/logistic-regression

Los Angeles City (2017, June 8). Traffic Collision Data from 2010 to Present. Retrieved October 13, from Lacity.org website: https://data.lacity.org/A-Safe-City/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w

National Center for Statistics and Analysis (2019, October). Estimate of motor vehicle traffic crash fatalities for the holiday periods of 2019. (Traffic Safety Facts Research Note. Report No. DOT HS 812 823). Washington, DC: National Highway Traffic Safety Administration.

Package Forecast. (2019, August 22). Retrieved from https://cran.r-project.org/web/ packages/forecast/forecast.pdf.

Shmueli, G., Patel, N. R., Bruce, P. C. (2010). Logistic Regression. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner* (2nd ed., pp. 347-365). Hoboken, NJ: John Wiley & Son Inc.

Shmueli, G., Patel, N. R., Bruce, P. C. (2010). Classification Tree. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner* (2nd ed., pp. 301-306). Hoboken, NJ: John Wiley & Son Inc.

Shmueli, G., Patel, N. R., Bruce, P. C. (2010). k-NN Classifier (categorical outcome). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner* (2nd ed., pp. 250-259). Hoboken, NJ: John Wiley & Son Inc.

Shmueli, G., Patel, N. R., Bruce, P. C. (2010). Chapter 14: Cluster Analysis. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner* (2nd ed., pp. 489-523). Hoboken, NJ: John Wiley & Son Inc.

Shumway, Robert & Stoffer, David (2011). *Time Series Analysis and Its Applications With R Examples*. (3rd Ed.). New York: Springer Science+Business Media, LLC.

RStudio.(2019). Retrieved October 13, 2019, from Rstudio.com website: https://rstudio.com/sqldf package | R Documentation. (2013). Retrieved October 13, 2019, from Rdocumentation.org website: https://www.rdocumentation.org/packages/ sqldf/versions/0.4-11

Torgo, Luis (2010, October 1). *K-Nearest Neighbour Classification*. Retrieved October 1, 2010, from https://www.rdocumentation.org, https://www.rdocumentation .org/packages/DMwR/versions/0.4.1/topics/kNN

Wickham H., Retrieved from https://www.rdocumentation.org/packages/ggplot2/ versions3.2.1

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2018). Efficient kNN Classification With Different Numbers of Nearest Neighbors. *IEEE Transactions on Neural Networks and Learning Systems, 29*(5), 1774-1785.

https://www.ots.ca.gov/media-and-research/collision-rankings-results/?wpv-wpcf-year=2016&wpv-wpcf-city_county=Los+Angeles&wpv_filter_submit=Submit