Data Analytics for Smart City & Transportation

FINAL PROJECT

Roberto Batista

December 7th, 2019
Version 1.3

Table of Contents

Table of Figures

List of Tables

1. Introduction

The city of San Francisco has an area of 47 square miles with a population of 883,305 inhabitants in 2018, and it is located in the east coast of the USA in the peninsula of San Francisco Bay [1][2]. The economic growth of the Bay Area attracted and created wealth, especially in the technological field. In consequence, positive impacts flourished into society but also was created several new challenges to the inhabitants as high traffic.

Transportation Authority's Congestion Management Program (TACMP) points to the worsening of traffic in 2010 and 2016. The TACMP indicates that average AM peak arterial travel speeds decreased by -26%, and the PM peak arterial speeds decreased by -27% in the same period. Vehicle delay on the significant roadways increased by 40,000 hours on a weekday; also, the vehicle miles traveled on major roadways increased by over 630,000 miles on a weekday [3].

The traffic information is an excellent allied to the citizen to avoid traffic jams and consequently the decay of quality of life. In this work, it will be presented a machine learning model as proof of the concept of an application that can cooperate with San Francisco people to predict the Free Flow Travel Time ("FF_TIME"). The algorithm proposed is a Linear Regression Model designed to predict the free-flow travel time ("FF_TIME"), which is the travel time for roadway conditions like 60 mph in freeways and 35mph on principal arterials.

The dataset used in this project is supplied by the San Francisco County Transportation Authority (SFCTA) [11], which also provided the data dictionary, which is in Appendix A of this document.

2. Methodology

The project was performed using the R programming language in the RStudio Integrated Development Environment (IDE) [4]. The following libraries were used in this project:

- *Tidyverse* [5] data science library collection:
- *Dplyr* [6]: data manipulation.
- *Ggplot2* [7]: graphic generation.
- *Readr* [8]: data parser.
- *DataExplorer* [9] library used to create the EDA graphics. The library uses *ggfortify* and *ggplot2* to plot its graphics.
- *Caret* [10] library used to create the test and training datasets for the machine learning models.
- *Ggfortify* [11] library is used by *DataExplorer* to create plots.

Exploratory Data Analysis (EDA)

Initially, it was verified the data types of each variable to assure the right importation data process into the R notebook. The exploratory data analysis shows that the dataset contains 7116 observations (records) and 29 variables (features), of which twenty-three are quantitative, and six are qualitative variables, without missing data.

The quantitative variables were inspected through descriptive analysis such as minimal, maximal, mean, median, first and third quartiles, and histogram plots. The qualitative variables were verified using the bar plots. The diagnostic plots were also used to evaluate the linear regression assumptions: normality, homoscedasticity, and linearity of the dataset.

Variables selection

It was identified the variables which are not relevant to be included in the regression model based on the data dictionary information and the statistical analysis.

- **Unique ID:** The variables related to unique id like "X1", "ID," and "ModifiedTMC" was disregarded.
- **Variable that does not vary:** The variable "YEAR," which indicates the year 2016, and the variable "BASE_INRIX_VOL_PRESIDIO" with 99.43% of its content holds the same value was disregarded.

- **Correlated variables:** The variable "SPEED_20TH" was disregarded because it derives from INRIX speed data and has a high correlation with it.
- **Further investigation.** For further variable investigation was used the statistical significance calculation results provided by the linear regression model of the variables (Table 1).

The histograms (figures 1 and 2) created from continuous variables point right-skewed distribution in high or low degree, motivating the application of natural logarithm to "normalize" the distribution. The results of the log function application did not improve the model and was abandoned. Better results were reached, removing the outliers of the dataset using the result of *autoplot()* function in multiples times. The analysis resulted in 139 records removed. Also, the boxplot of the dependent variable "FF_TIMES" and the predictor variable "CHAMP_LINK_COUNT" helped to remove 687 "FF_TIMES" additional outliers.

Table 1-Variable significance analysis.

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-5.915e-01	1.741e-01	-3.397	0.000686	***
TODEA	4.568e-03	4.139e-03	1.104	0.269823	
TODEV	6.021e-03	4.846e-03	1.243	0.214074	
TODMD	-1.665e-02	5.001e-03	-3.329	0.000876	***
TODPM	-2.373e-02	3.860e-03	-6.149	8.22e-10	***
CHAMP_LINK_COUNT	2.041e-02	7.014e-04	29.093	< 2e-16	***
PRESIDIO1.0	-1.376e-02	2.229e-02	-0.618	0.536883	
ALPHA	1.735e+00	3.602e-01	4.815	1.50e-06	***
BETA	-7.953e-02	1.296e-02	-6.137	8.88e-10	***
AT3.0	-7.845e-03	3.663e-03	-2.142	0.032264	*
AT2.0	-7.286e-03	3.486e-03	-2.090	0.036642	*
AT0.0	-3.073e-02	4.106e-03	-7.484	8.07e-14	***
FT24	4.575e-01	6.490e-02	7.049	1.97e-12	***
FT21	-2.808e-01	5.705e-02	-4.923	8.74e-07	***
FT23	2.323e-02	4.198e-03	5.534	3.25e-08	***
LANES	-6.479e-03	2.775e-03	-2.335	0.019565	*
DISTANCE	6.784e-01	1.189e-02	57.041	< 2e-16	***
CAPACITY	-7.048e-06	1.212e-06	-5.813	6.38e-09	***
FFS	-2.026e-02	4.222e-04	-47.979	< 2e-16	***
INRIX_SPEED	2.052e-02	5.265e-04	38.977	< 2e-16	***
INRIX_TIME	3.974e-01	4.084e-03	97.307	< 2e-16	***
INRIX_VOL	8.496e-06	1.217e-06	6.983	3.14e-12	***
CHAMP_PCE	-2.272e-05	3.892e-06	-5.838	5.53e-09	***
CHAMP_VOL	2.806e-05	4.374e-06	6.415	1.50e-10	***
AVG_DUR	-7.991e-03	1.455e-03	-5.494	4.07e-08	***
AVG_DUR_MAJOR_ARTERIALS	6.746e-03	1.510e-03	4.466	8.07e-06	***
AVG_DUR_MINOR_ARTERIALS	6.080e-03	1.678e-03	3.623	0.000293	***

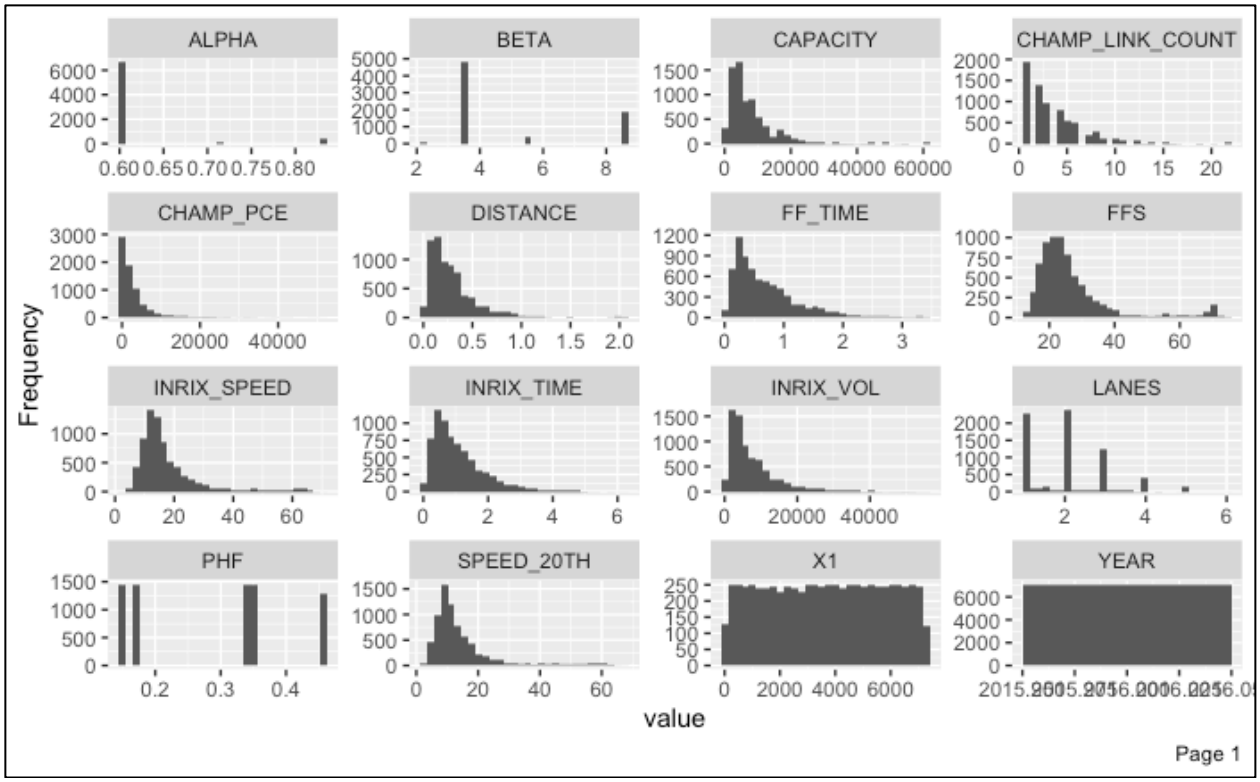

Figure 1-Continuous variables histograms I

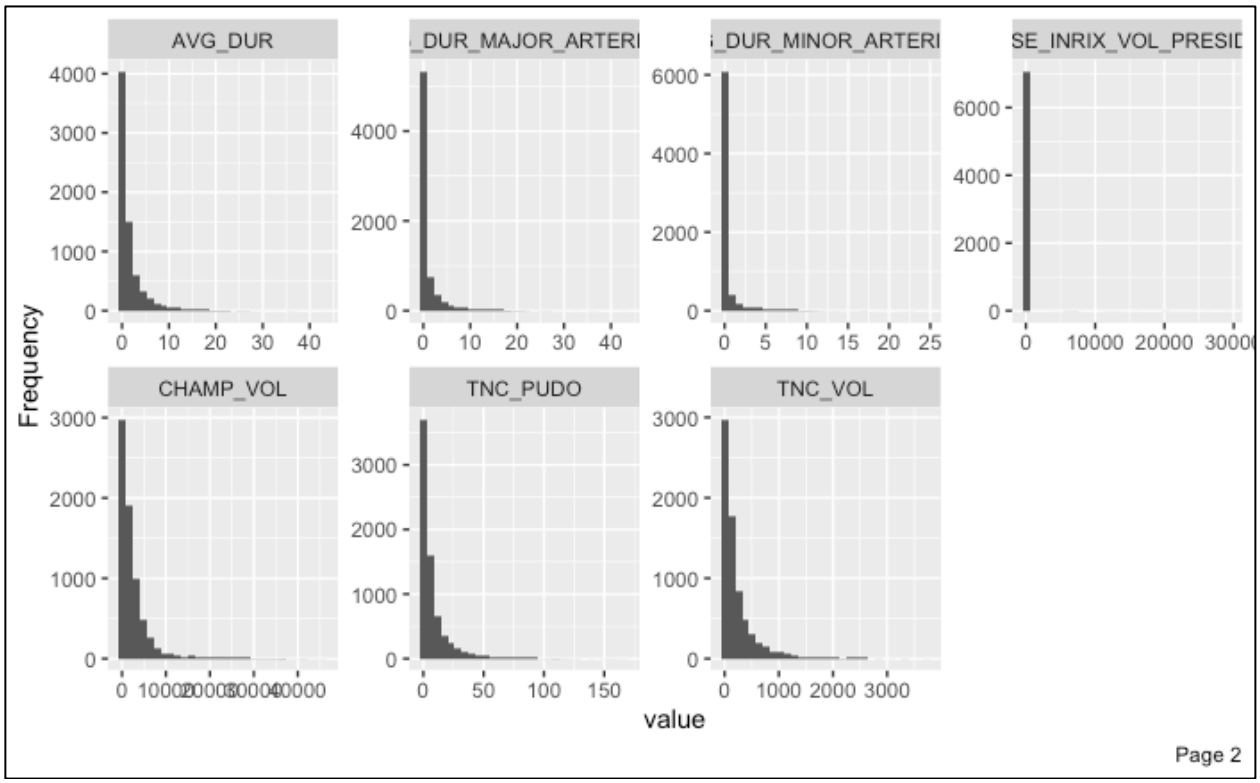

Figure 2-Continuous variables histograms II

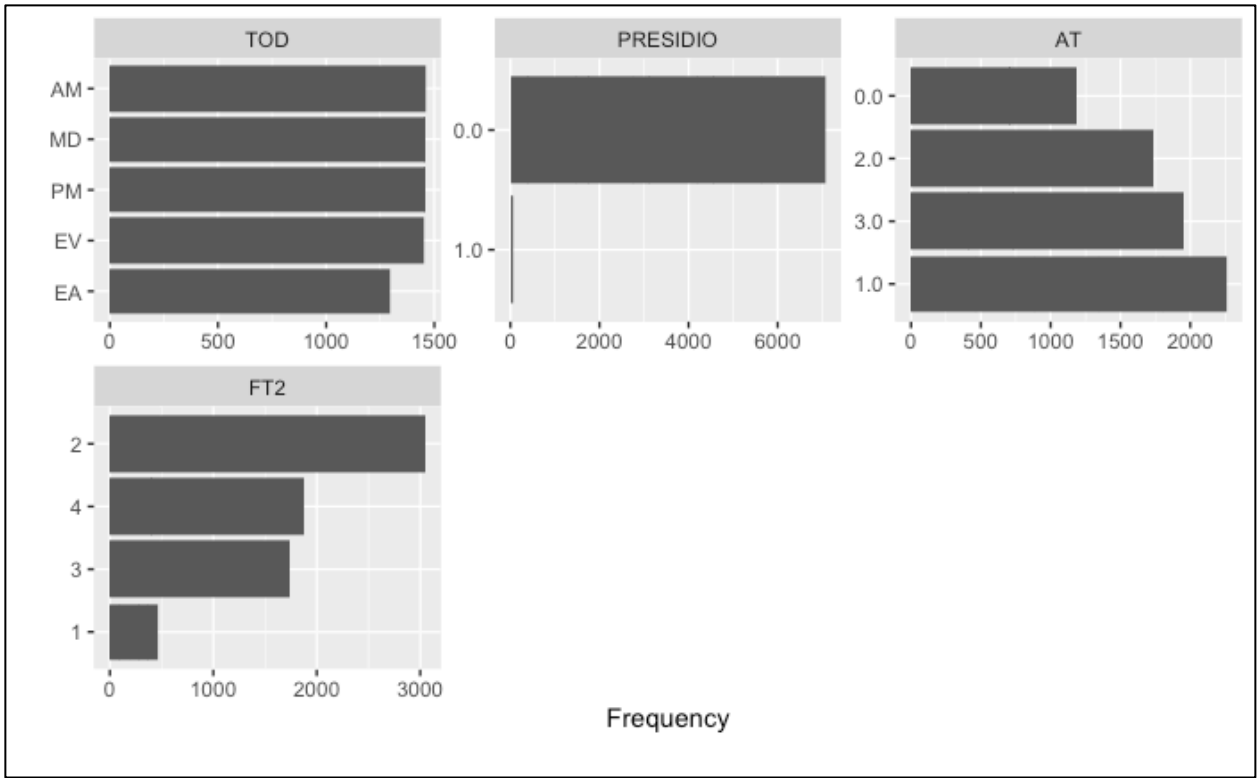

Figure 3-Categorical variables bar plots.

3. Models

It was created twenty-six linear regression models combining the predictor variables according to its significance and impact on R-Square and RMSE results. Table 5 presents all the models created and its evaluation, which was used to select the best predictive model. The best two models among the twenty-six tested will be described in this report.

Table 2-Models results.

Model	RMSE	R2	Variables Removed
01	0.1039624	0.9668751	
02	0.0722703	0.9590644	AVG_DUR_MINOR_ARTERIALS
03	0.0725117	0.9587976	AVG_DUR_MAJOR_ARTERIALS
04	0.07236	0.9589663	AVG_DUR
05	0.0743496	0.9567421	CHAMP_VOL
06	0.0742657	0.9568337	CHAMP_PCE
07	0.0723347	0.959003	INRIX_VOL
08	0.0971243	0.9260216	INRIX_TIME
09	0.0761874	0.9544939	INRIX_SPEED
10	0.0867117	0.9410225	FFS
11	0.0723359	0.9589997	CAPACITY
12	0.0969026	0.9263568	DISTANCE
13	0.0722239	0.9591313	LANES
14	0.0732021	0.9580171	FT2
15	0.0727604	0.9585233	AT
16	0.0730803	0.9581463	BETA
17	0.0728767	0.9583879	ALPHA

18	0.0777029	0.9527795	CHAMP_LINK_COUNT
19	0.0719598	0.9594457	ALPHA + BETA + LANES
20	0.072104	0.959273	ALPHA + BETA + LANES + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS
21	0.0782265	0.9521839	ALPHA + BETA + LANES + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS+ AT + CAPACITY + INRIX_VOL + CHAMP_PCE + FT2 + CHAMP_VOL
22	0.0835436	0.9454653	ALPHA + BETA + LANES + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS + AT+ CAPACITY + INRIX_VOL + CHAMP_PCE + FT2 + CHAMP_VOL + CHAMP_LINK_COUNT
23	0.0973295	0.9256439	ALPHA + BETA + LANES + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS + AT+ CAPACITY + INRIX_VOL + CHAMP_PCE + FT2 + CHAMP_VOL + CHAMP_LINK_COUNT + **INRIX_SPEED**
24	0.1310861	0.8655102	ALPHA + BETA + LANES + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS + AT+ CAPACITY + INRIX_VOL + CHAMP_PCE + FT2 + CHAMP_VOL + CHAMP_LINK_COUNT + **INRIX_TIME**
25	0.116347	0.8937649	ALPHA + BETA + LANES + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS + AT+ CAPACITY + INRIX_VOL + CHAMP_PCE + FT2 + CHAMP_VOL + CHAMP_LINK_COUNT + **FFS**
26	0.1059189	0.912117	ALPHA + BETA + LANES + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS + AT+ CAPACITY + INRIX_VOL + CHAMP_PCE + FT2 + CHAMP_VOL + CHAMP_LINK_COUNT + **DISTANCE**

Linear Regression Assumptions

The linear regression assumptions analysis is the same for the models 20 and 21 as the results did not differ significantly.

Linearity. The linearity of the model was analyzed using the Residuals vs. Fitted plot (figures 4 and 5). The blue reference line plotted is sufficiently close to zero and horizontal, confirming the linear relationship between the predictors (independent) and outcome (dependent) variables.

Homogeneity. The homogeneity assumption of the model was analyzed using the scale-location plot (figures 4 and 5), which indicated that the residuals of variance have no constant variance. The possible solution is the continuous process of outliers removal pointed by *autoplot()*, which was done up to 860 records in total, explained in the variable selection section.

Normality. The normality of residuals was verified using a Q-Q plot, which points for the problematic skewed right distribution of the variables. The use of logarithm was tested to adjust the distribution of variables to normal but unfortunately caused more distortion in the Normal Q-Q plot and Residuals vs. Fitted plot.

Model 19

The model 19 (model_19) was the best model fitted. It was included fifteen predictor variables: "TOD", "CHAMP_LINK_COUNT", "AT", "FT2", "DISTANCE", "'CAPACITY", "FFS", "'INRIX_SPEED", "INRIX_TIME", "INRIX_VOL, CHAMP_PCE", "CHAMP_VOL", "AVG_DUR", "AVG_DUR_MAJOR_ARTERIALS", and "AVG_DUR_MINOR_ARTERIALS".

Call: lm(formula = FF_TIME ~ TOD + CHAMP_LINK_COUNT + AT + FT2 + DISTANCE + CAPACITY + FFS + INRIX_SPEED + INRIX_TIME + INRIX_VOL + CHAMP_PCE + CHAMP_VOL + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS, data = train_data)

Residuals

Min	1Q	Median	3Q	Max
-1.24369	-0.03551	-0.00272	0.03081	0.45474

Table 3-Model 19

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	1.531e-01	1.262e-02	12.132	< 2e-16	***
TODEA	4.591e-03	3.423e-03	1.341	0.1800	
TODEV	8.982e-03	3.689e-03	2.435	0.0149	*
TODMD	-6.572e-03	3.730e-03	-1.762	0.0781	.

TODPM	-1.552e-02	3.279e-03	-4.735	2.25e-06	***
CHAMP_LINK_COUNT	1.818e-02	7.492e-04	24.266	< 2e-16	***
AT1.0	2.487e-02	3.342e-03	7.441	1.17e-13	***
AT3.0	2.269e-02	3.968e-03	5.719	1.13e-08	***
AT2.0	1.784e-02	3.770e-03	4.732	2.28e-06	***
FT22	1.876e-02	7.911e-03	2.372	0.0177	*
FT24	7.682e-02	9.393e-03	8.179	3.60e-16	***
FT23	4.462e-02	9.132e-03	4.886	1.06e-06	***
DISTANCE	8.164e-01	1.412e-02	57.800	< 2e-16	***
CAPACITY	-1.218e-06	9.764e-07	-1.247	0.2123	
FFS	-1.780e-02	3.548e-04	-50.173	< 2e-16	***
INRIX_SPEED	1.560e-02	4.403e-04	35.431	< 2e-16	***
INRIX_TIME	3.326e-01	4.309e-03	77.203	< 2e-16	***
INRIX_VOL	1.792e-06	9.572e-07	1.872	0.0612	.
CHAMP_PCE	-3.939e-05	3.367e-06	-11.699	< 2e-16	***
CHAMP_VOL	4.729e-05	3.866e-06	12.233	< 2e-16	***
AVG_DUR	-7.189e-03	1.155e-03	-6.224	5.23e-10	***
AVG_DUR_MAJOR_ARTERIALS	7.678e-03	1.211e-03	6.341	2.48e-10	***
AVG_DUR_MINOR_ARTERIALS	6.399e-03	1.344e-03	4.759	2.00e-06	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4- Model 19 RSE, Multiple R-Squared, and F-Statistic.

Residual standard error: 0.07243 on 4987 degrees of freedom
Multiple R-squared: 0.9593, Adjusted R-squared: 0.9591
F-statistic: 5339 on 22 and 4987 DF, p-value: < 2.2e-16

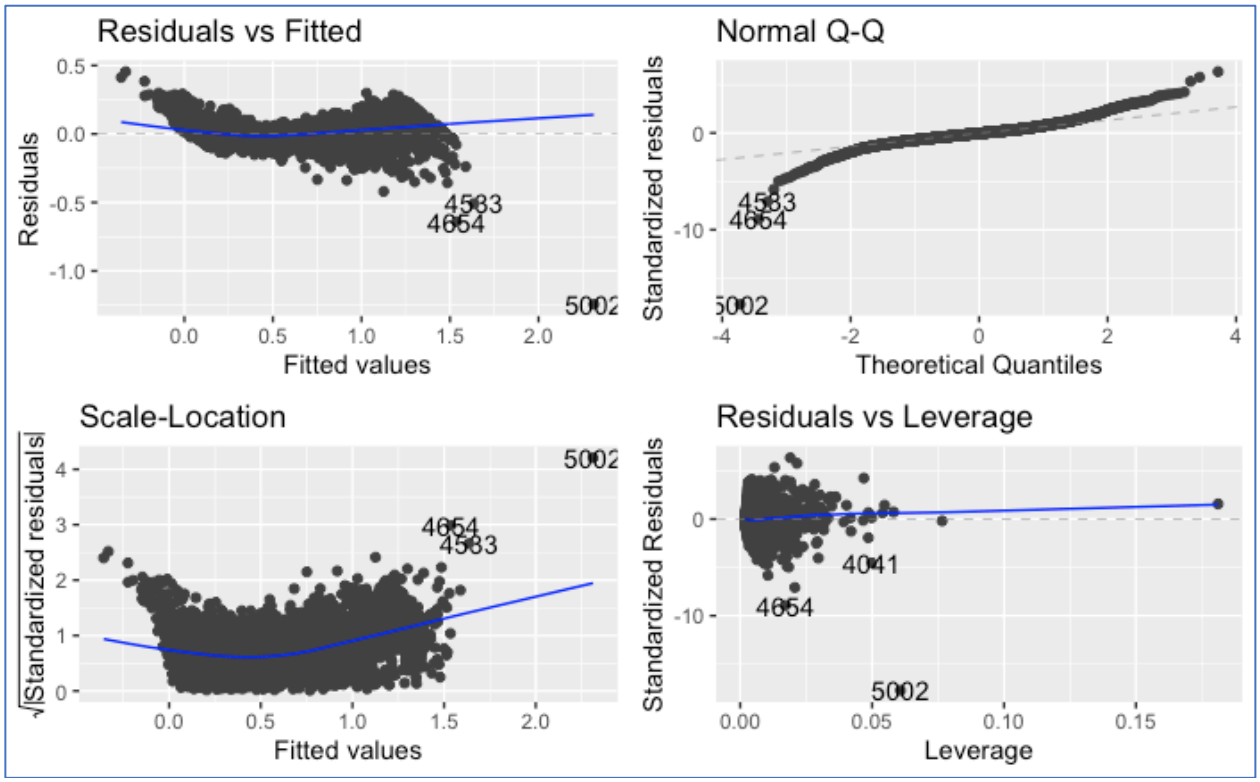

Figure 4-Model 19 diagnosis plots.

Model 20

The model_20 was the second-best model fitted. It was included twelve predictor variables: FF_TIME as outcome variable, and the following predictors: "TOD", "CHAMP_LINK_COUNT", "AT", "FT2", "DISTANCE", "CAPACITY", "FFS", "INRIX_SPEED", "INRIX_TIME", "INRIX_VOL", "CHAMP_PCE", "CHAMP_VOL".

Call: lm(formula = FF_TIME ~ TOD + CHAMP_LINK_COUNT + AT + FT2 + DISTANCE + CAPACITY + FFS + INRIX_SPEED + INRIX_TIME + INRIX_VOL + CHAMP_PCE + CHAMP_VOL, data = train_data)

Residuals

Table 5-Model 20 residuals

Min	1Q	Median	3Q	Max
-1.23004	-0.03597	-0.00261	0.03052	0.45575

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	1.461e-01	1.254e-02	11.651	< 2e-16	***
TODEA	4.729e-03	3.415e-03	1.385	0.16623	
TODEV	5.485e-03	3.571e-03	1.536	0.12463	
TODMD	-8.865e-03	3.715e-03	-2.386	0.01706	*
TODPM	-1.580e-02	3.290e-03	-4.802	1.62e-06	***
CHAMP_LINK_COUNT	1.875e-02	7.379e-04	25.405	< 2e-16	***
AT1.0	2.474e-02	3.255e-03	7.599	3.55e-14	***
AT3.0	2.480e-02	3.807e-03	6.515	7.98e-11	***
AT2.0	1.894e-02	3.640e-03	5.203	2.04e-07	***
FT22	2.482e-02	7.878e-03	3.150	0.00164	**
FT24	7.786e-02	9.412e-03	8.273	< 2e-16	***
FT23	4.920e-02	8.998e-03	5.467	4.79e-08	***
DISTANCE	8.093e-01	1.413e-02	57.277	< 2e-16	***
CAPACITY	-1.884e-06	9.495e-07	-1.984	0.04733	*
FFS	-1.794e-02	3.498e-04	-51.300	< 2e-16	***
INRIX_SPEED	1.589e-02	4.349e-04	36.545	< 2e-16	***
INRIX_TIME	3.320e-01	4.306e-03	77.088	< 2e-16	***
INRIX_VOL	2.585e-06	9.155e-07	2.824	0.00476	**
CHAMP_PCE	-3.779e-05	3.320e-06	-11.384	< 2e-16	***
CHAMP_VOL	4.561e-05	3.814e-06	11.957	< 2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6-Model 20 RSE, Multiple R-Squared, and F-Statistic.

Residual standard error: 0.07271 on 4990 degrees of freedom
Multiple R-squared: 0.9589, Adjusted R-squared: 0.9588
F-statistic: 6132 on 19 and 4990 DF, p-value: < 2.2e-16

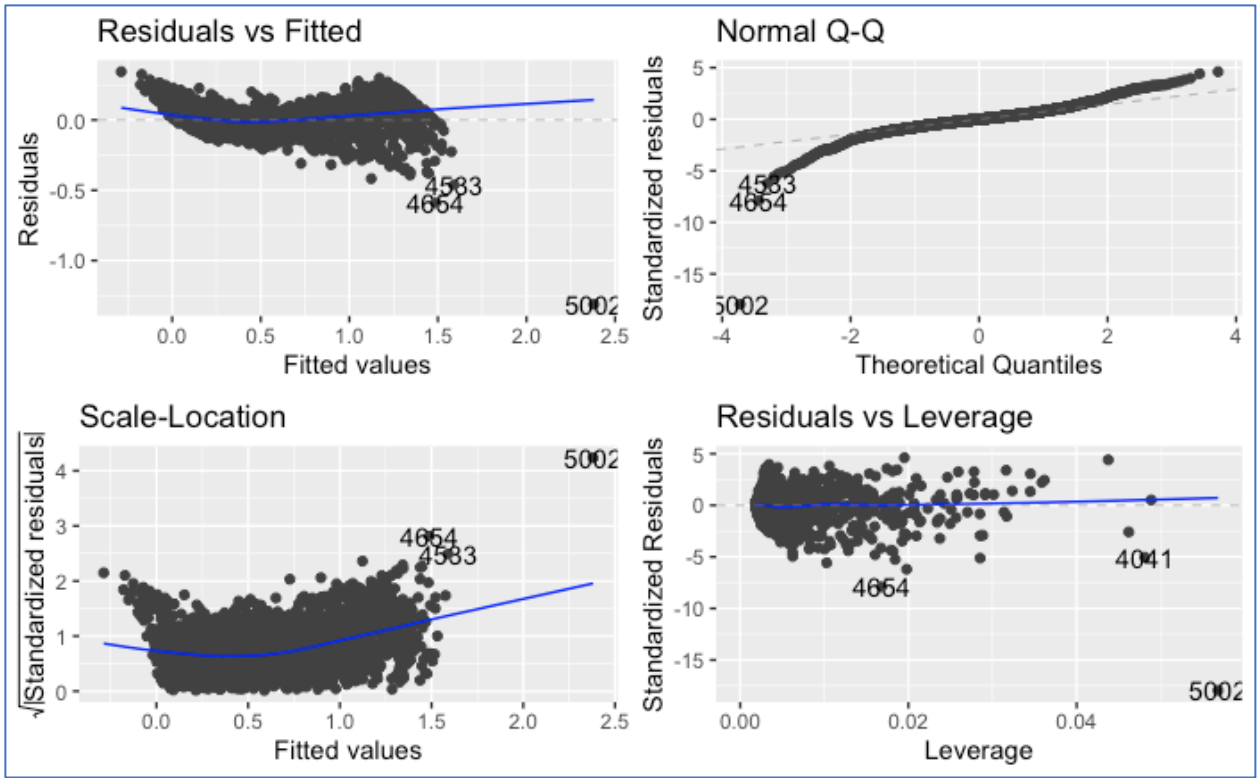

Figure 5-Model 20 diagnosis plots.

4. Conclusion

The SFCTA dataset used in this project has inherent problems which prevent the satisfactory use of the linear regression. The linearity assumptions is quite satisfied, but homoscedasticity was not satisfied as the variance of the residuals have not constant variance across the x-axis, and the normal distribution is not reached. The dataset has a good potential for use, but additional transformation effort is necessary.

The dataset was created based on multiple sources, including estimate data, which is a possible reason for the problematic in the dataset in general. Despite this challenge, the dataset was useful to demonstrate that a travel time prediction can have several predictor variables involved with different levels of influence in the dependent variable.

The dataset is fascinating, and this project is a proof of concept that could be extended for an advanced model and application to be available on the SFCTA dashboard website [11].

5. References

[1] Encyclopedia.com. (n.d.). San Francisco | Encyclopedia.com. Retrieved September 14, 2019, from https://www.encyclopedia.com/places/united-states-and-canada/us-political-geography/san-francisco

[2] United States Census Bureau. (n.d.). QuickFacts: San Francisco city, California; United States. Retrieved November 27, 2019, from https://www.census.gov/quickfacts/fact/table/sanfranciscocitycalifornia,US/PST045218

[3] San Francisco County Transportation Authority. (2018). TNCs & Congestion - Final Report. San Francisco County Transportation Authority.

[4] Allaire, J. (2012). RStudio: integrated development environment for R. Boston, MA, 770.

[5] Wickham, H. (2017). The tidyverse. R package ver. 1.1, 1.

[6] Wickham, H., Francois, R., Henry, L., & Müller, K. (2015). dplyr: A grammar of data manipulation. R package version 0.4, 3.

[7] Wickham, H. (2016). ggplot2: elegant graphics for data analysis. Springer.

[8] Wickham, H., Hester, J., & Francois, R. (2017). Readr: Read rectangular text data. R package version, 1(1).

[9] Cui, B. (n.d.). Introduction to DataExplorer. Retrieved November 28, 2019, from https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html

[10] Kuhn, M. (2015). Caret: classification and regression training. Astrophysics Source Code Library.

[11] San Francisco County Transportation Authority (SFCTA). (n.d.). TNCs Today. SFCTA. Retrieved November 28, 2019, from https://www.sfcta.org/projects/tncs-today

APPENDIX A

Data Dictionary

Feature Name	Description
ID	a unique ID, which is a combination of the next two fields
ModifiedTMC	ID for the spatial unit of analysis, a directional section of roadway
TOD	Time-of-day: AM=6-9 AM, MD=9 AM-3:30 PM, PM=3:30-6:30 PM, EV=6:30 PM-3:00 AM, EA=3-6 AM
YEAR	The year, either 2010 or 2016
CHAMP_LINK_COUNT	The number of SF CHAMP links that aggregate to this ModifiedTMC
PRESIDIO	Binary flag indicating whether the ModifiedTMC is on the Presidio Parkway or Veterans Blvd
PHF	Peak Hour Factor, the share of the total period volume that occurs in the highest 1hour period.
ALPHA	alpha term for use in VDF
BETA	beta term for use in VDF
AT	Area Type: 0=Regional Core, 1=Central Business District, 2=Urban Business, 3=Urban
FT2	Facility Type: 1=Freeway, Expressway or Ramp, 2=Major Arterial, 3=Minor Arterial, 4=Local or Collector
LANES	Number of lanes (can be non-integer due to averaging across SFCHAMP links)
DISTANCE	distance in miles
CAPACITY	capacity in vehicles for the period as a whole
FFS	free flow speed
SPEED_20TH	20th percentile speed, as measured by INRIX data
FF_TIME	free-flow travel time
INRIX_SPEED	average speed, as measured by INRIX data
INRIX_TIME	average travel time, as measured by INRIX data
INRIX_VOL	implied volume
CHAMP_PCE	SF CHAMP passenger car equivalents (PCEs)
CHAMP_VOL	SF CHAMP volume (vehicles)
TNC_VOL	TNC volume
TNC_PUDO	TNC pick-ups and drop-offs
AVG_DUR	TNC average duration variable, calculated as: (CAPACITY/LANES) * (TNC_PUDO / (3600/PHF]))
AVG_DUR_MAJOR_ARTERIALS	TNC average duration variable on major arterials, zero elsewhere
AVG_DUR_MINOR_ARTERIALS	TNC average duration variable on minor arterials, zero elsewhere
BASE_INRIX_VOL_PRESIDIO	In 2016, the base year (2010) implied volume on the Presidio Parkway or Veterans Blvd, zero elsewhere.
TNC_PUDO	TNC pick-ups and drop-offs
AVG_DUR	TNC average duration variable, calculated as: (CAPACITY/LANES) * (TNC_PUDO / (3600/PHF]))
AVG_DUR_MAJOR_ARTERIALS	TNC average duration variable on major arterials, zero elsewhere
AVG_DUR_MINOR_ARTERIALS	TNC average duration variable on minor arterials, zero elsewhere
BASE_INRIX_VOL_PRESIDIO	In 2016, the base year (2010) implied volume on the Presidio Parkway or Veterans Blvd, zero elsewhere.
BASE_INRIX_VOL_PRESIDIO	In 2016, the base year (2010) implied volume on the Presidio Parkway or Veterans Blvd, zero elsewhere.

APPENDIX B

R Notebook

---
title: "Smart City Project - Final"
Author: Roberto Batista - rbatista7484@floridapoly.edu

Loading Libraries

```{r, warning = FALSE, message=FALSE}
library(tidyverse)
library(stats)
library(mlogit)
library(DataExplorer)
library(caret)
library(ggfortify)
```

Loading the data

```{r, warning=FALSE, message=FALSE}
setwd("~/Desktop/SMARTCITIES/FINAL/tncs_congestion_empirical_dataset")
year2016 <- read_csv("ESTFILE_2016.csv", col_types = "iccfiifddddffddddddddddddddddddd")
```

Exploratory Data Analysis (EDA)

Checking data structure

```{r, eval=TRUE}
str(year2016)
```

```{r, eval=TRUE}
summary(year2016)
```

```{r, eval=TRUE}
year2016 %>%
  head(3)
```

```{r, eval=TRUE}
plot_missing(year2016)
```

Histograms to check quantitative variables.

```{r, eval=TRUE}
year2016 %>%
  plot_histogram()
```

Bar plots for qualitative variables analysis.

```{r, eval=TRUE}
plot_bar(year2016)
```

```{r, eval=TRUE}
year2016 %>%
  count(BASE_INRIX_VOL_PRESIDIO) %>%
  mutate(perc = (n/sum(n)))
```

Variables Selection

Let's remove the obvious variables which are not relevant to the model.
Removing variables related to ID's (`X1`, `ID`, `ModifiedTMC`), `YEAR` column, as all the data is related to year 2016, and `BASE_INRIX_VOL_PRESIDIO` which 99.4% of data is `0`.

```{r}
subset_var <- year2016 %>%
  select(-c(X1, ID, ModifiedTMC, YEAR, BASE_INRIX_VOL_PRESIDIO))
```

Now checking the statistical relevance of the variables.

```{r, eval=TRUE}
test_model <- lm(FF_TIME ~., data = subset_var)
summary(test_model)
```

The test_model summary presents 'NA's for PHF variables, indicating high collinearity with the dependent variable. Let's confirm it checking colinearity using alias().

```{r, eval=TRUE}
alias(test_model)
```

The variable `PHF` has high collinearity and will be removed from the model. In addition, the variables `TNC_PUDO` and `TNC_VOL` will be removed due to being not significant. According to the dictionary, the variable `SPEED_20TH` is a deviation from the same data as `INRIX_SPEED` is and will also be removed.

```{r}
subset_var <- subset_var %>%
  select(-c(PHF, TNC_PUDO, TNC_VOL, SPEED_20TH))
#dim(subset_var)
```

```{r, eval=TRUE}
test_model <- lm(FF_TIME ~., data = subset_var)
summary(test_model)
```

Once removed the variables the variable `PRESIDIO` became also statistical not significant and it will be removed.

```{r}
subset_var <- subset_var %>%
  select(-PRESIDIO)
```

```{r, eval=TRUE}
test_model <- lm(FF_TIME ~., data = subset_var)
summary(test_model)
```

Let's change the `AT` categorical variable factor levels for a better understanding of model_analysis summary.
Check the current levels

```{r}
levels(subset_var$AT)
```

Change the "0.0" level as the first level to change the reference.

```{r}
subset_var$AT <- relevel(subset_var$AT, "0.0")
levels(subset_var$AT)
```

Let's change the `FT2` categorical variable factor levels for a better understanding of model_analysis summary.
Check the current levels

```{r}
levels(subset_var$FT2)
```

Change the "0.0" level as the first level to change the reference.

```{r}
subset_var$FT2 <- relevel(subset_var$FT2, "1")
levels(subset_var$FT2)
```

Computing the LR model with the variables of interest.

```{r, eval=TRUE}
model_analysis <- lm(FF_TIME ~ . , data = subset_var)
summary(model_analysis)
autoplot(model_analysis)
```

```{r, eval=TRUE}
subset_var %>%
  ggplot(aes(x = CHAMP_LINK_COUNT, y = FF_TIME))+
  geom_boxplot()
```

Removing FF_Time outliers

```{r}
subset_var %>%
  filter(FF_TIME >1.5) %>%
  count()
```

```{r}
subset_var <- subset_var %>%
  filter(FF_TIME <1.5)
subset_var %>%
  dim()
```

Removing the outliers.

```{r}
subset_var <- subset_var %>%
slice(-c(7100, 6969, 2307, 5514, 5914, 7104, 5049, 1310, 3865, 1937, 4772, 4767, 4685, 4686, 4094, 5123, 4763, 4101, 4764, 5122,
4105, 4663, 4763, 1869, 4715, 4764, 5122, 4105, 4663, 5116, 4760, 4623, 5114, 4621, 4091, 5110, 4073, 4615, 4297,4301, 4750,
5108, 4080, 5105, 4750, 4619, 4089, 4620, 4745, 5105, 4093, 5102, 4788, 4604, 5098, 5096, 4743, 4277, 4605, 4738, 5096, 4601,
4734, 5092, 4729, 4600, 5090, 4603, 5090, 4780, 6359, 5890, 5721, 5053, 6354, 5886, 5718, 5716, 5883, 6350, 5881, 5715, 6347,
5052, 5878, 5713, 6343, 5051, 5875,  571, 6389,5050, 5709, 5872, 6385, 6332, 5870, 5708, 6329, 5868, 5707, 6326, 5866, 5706,
5048, 6322, 5863, 5704, 5861, 5703, 6319, 5047, 5701, 5858, 6315, 6312, 5856, 5700, 5046, 5853, 5698, 6308, 5697, 5851, 6305,
```

```
6302, 5849, 5656, 5045, 6298,5846, 5694, 5693, 5844, 6295, 5692, 5842, 6292, 5691, 5840, 6289, 5044, 2304, 5782, 5313, 5626,
5766, 4444, 6171, 5685, 4367, 6217, 6222, 5861, 5059, 6219, 6222, 3228, 5710, 5202, 3715, 5862, 3715, 5012, 5011, 5014, 5012,
4546, 4659, 5013, 5011, 4510, 5010))
```
```{r}
subset_var <- subset_var %>%
  slice(-c(5000, 5001, 5002, 5003, 5004, 5005, 5006, 5007, 5008, 5009, 5010, 5011, 5012))
```
```{r, eval=TRUE}
model_analysis <- lm(FF_TIME ~ . , data = subset_var)
summary(model_analysis)
autoplot(model_analysis)
```
```{r}
subset_var %>%
  dim()
```
COMPUTING LINEAR REGRESSION PREDICTION
Testing Linear Regression Models for FF_TIME (free flow travel time) Prediction Separing the data into training and test set.
```{r}
set.seed(2019)
training_samples <- subset_var$FF_TIME %>%
  createDataPartition(p=0.8, list = FALSE)
train_data <- subset_var[training_samples,  ]
test_data  <- subset_var[-training_samples, ]
```
Computing the model with train_data
```{r}
train_data %>%
  colnames()
```
Model 1
```{r}
model_1 <- lm(FF_TIME ~ TOD + CHAMP_LINK_COUNT + ALPHA + BETA + AT +  FT2 + LANES + DISTANCE + CAPACITY + FFS  +
INRIX_SPEED  +   INRIX_TIME + INRIX_VOL + CHAMP_PCE + CHAMP_VOL + AVG_DUR +  AVG_DUR_MAJOR_ARTERIALS +
AVG_DUR_MINOR_ARTERIALS , data = train_data)
summary(model_1)
#Make predictions with test_data
predictions <- model_1 %>% predict(test_data)
```
Model performance
```{r}
#Prediction Error - RMSE
RMSE(predictions, test_data$FF_TIME)
#R-Square
R2(predictions, test_data$FF_TIME)
autoplot(model_1)
```
Model 19
```{r}
model_19 <- lm(FF_TIME ~ TOD +  CHAMP_LINK_COUNT +   AT +  FT2 +  DISTANCE + CAPACITY + FFS  + INRIX_SPEED  +
INRIX_TIME  +   INRIX_VOL  +  CHAMP_PCE  +   CHAMP_VOL  +  AVG_DUR  +   AVG_DUR_MAJOR_ARTERIALS  +
AVG_DUR_MINOR_ARTERIALS  , data = train_data)
summary(model_19)
predictions <- model_19 %>% predict(test_data)
```
Make predictions with test_data
```{r}
predictions <- model_19 %>% predict(test_data)
print("Model performance")
print("Prediction Error - RMSE:")
RMSE(predictions, test_data$FF_TIME)
print("R-Square:")
R2(predictions, test_data$FF_TIME)
autoplot(model_19)
```
Model 20
```{r}
model_20 <- lm(FF_TIME ~ TOD +  CHAMP_LINK_COUNT + AT +  FT2 + DISTANCE +  CAPACITY + FFS  + INRIX_SPEED  + INRIX_TIME
+  INRIX_VOL + CHAMP_PCE + CHAMP_VOL , data = train_data)
summary(model_20)
predictions <- model_20 %>% predict(test_data)
```
Make predictions with test_data

```{r}
predictions <- model_20 %>% predict(test_data)
print("Model performance")
print("Prediction Error - RMSE:")
RMSE(predictions, test_data$FF_TIME)
print("R-Square:")
R2(predictions, test_data$FF_TIME)
autoplot(model_20)
```

Testing the application of log in the predictor variables and the outcome variable, both in train and test datasets.

```{r}
train_data$FFS        <- log(train_data$FFS)
test_data$FFS         <- log(test_data$FFS)
train_data$FFS %>%
  plot_histogram()
```

Model 19 Log

```{r}
model_19_log <- lm(log(FF_TIME) ~ TOD + CHAMP_LINK_COUNT + AT + FT2 + DISTANCE + CAPACITY + FFS + INRIX_SPEED + INRIX_TIME + INRIX_VOL + CHAMP_PCE + CHAMP_VOL + AVG_DUR + AVG_DUR_MAJOR_ARTERIALS + AVG_DUR_MINOR_ARTERIALS, data = train_data)summary(model_19_log)
predictions <- model_19_log %>% predict(test_data)
autoplot(model_19_log)
```

Make predictions with test_data

```{r}
predictions <- Model_26_log %>% predict(test_data)
#Model performance
#Prediction Error - RMSE
RMSE(predictions, test_data$FF_TIME)
# R-Square
R2(predictions, test_data$FF_TIME)
```