Data Science 2: Statistics for Data Science

King St. Transit Pilot

(October & November 2017)

Statistical Data Analysis Project

Lamjed Lounissi

Date: April 18th, 2022

Introduction

The King Street Transit Pilot is a joint undertaking by City of Toronto and TTC to explore bold, transformative ideas for how to redesign King Street to achieve three broad city-building objectives:

- Move people more efficiently on transit
- Improve placemaking
- Support economic prosperity

King Street travels through a highly urbanized environment. The 504 King streetcar route is the busiest surface transit route in Toronto, carrying over 60,000 riders on an average weekday. The streetcar currently operates in mixed traffic with transit signal priority at intersections, sharing the street with about 20,000 vehicles each day. Streetcar operations currently suffer from slow travel speeds, delays caused by traffic signals and turning vehicles, unreliable headways, and frequency leading to bunching, and general overcrowding of vehicles [1].

The transit pilot project went into effect in November 2017 and saw the construction of a corridor built between Jarvis and Bathurst. While private vehicles are allowed to use the corridor during the test with some restrictions, the intent is to optimize streetcar and pedestrian performance, reducing traffic time especially during rush hours [2].

Objectives

In this study, our primary goal is to determine whether the King St. Transit Pilot improved reliability, speed and capacity on King Street by prioritizing streetcars over private vehicles. The secondary objective is to predict the traffic volume of different transportations using information such as intersection name, direction, and time of the day etc. With an effective model, it should be possible to predict the traffic volume at any given time of a day at any intersection, hence help with decision making accordingly.

Datasets

For this analysis, we used the *King St. Transit Pilot - Detailed Traffic & Pedestrian Volumes* and the *King St Transit Pilot-Bluetooth Travel Time* both from the City of Toronto Open Data Portal [3 & 4]. The first data spanned from early November 2017 to June 2018 containing information about traffic volumes at different time segments across different intersections of King St. The data was segmented in 3 different categories: cyclists, pedestrians, and vehicles. We were also given information about baseline traffic volume as a benchmark. The second dataset contained travel time over the same duration and again used a baseline for comparison.

Data Preparation

To begin our analysis, we imported the data sets and checked for null values. The sets were robust and complete. Afterwards, we inspected columns and made modifications by dropping certain categories such as passenger identification id and renaming some features for greater clarity.

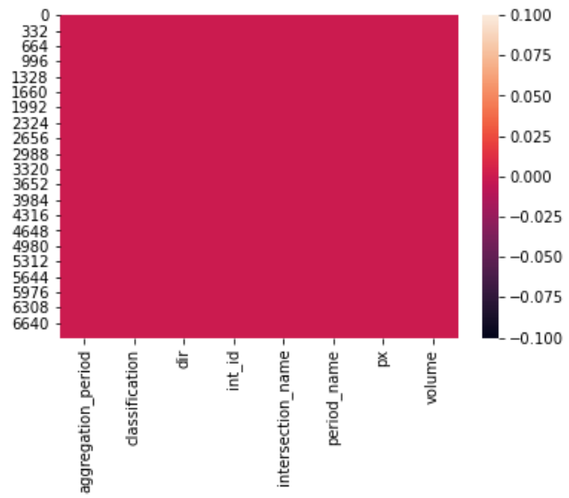
```
data.columns
```

```
Index(['_id', 'aggregation_period', 'classification', 'dir', 'int_id',
       'intersection_name', 'period_name', 'px', 'volume'],
      dtype='object')
```

```
# drop _id columns
data = data.drop('_id', axis=1)
```

```
# check to see if there are any blanks
sns.heatmap(data.isnull())
```

```
<AxesSubplot:>
```




```
data
```

	aggregation_period	classification	dir	int_id	intersection_name	time_period	px	passengers
0	18-Apr	Cyclists	EB	13466110	Queen St E / Jarvis St	14 Hour	6	171
1	18-Apr	Cyclists	EB	13466110	Queen St E / Jarvis St	Afternoon (12:00-17:00)	6	67
2	18-Apr	Cyclists	EB	13466110	Queen St E / Jarvis St	AM Peak Period (07:00-10:00)	6	19
3	18-Apr	Cyclists	EB	13466110	Queen St E / Jarvis St	Evening (17:00-22:00)	6	81
4	18-Apr	Cyclists	EB	13466110	Queen St E / Jarvis St	Midday (10:00-16:00)	6	67
...	...	...	...	...	...	...	...	...
6951	18-May	Vehicles	WB	13468126	Bathurst St / Front St W	Afternoon (12:00-17:00)	297	493
6952	18-May	Vehicles	WB	13468126	Bathurst St / Front St W	AM Peak Period (07:00-10:00)	297	560
6953	18-May	Vehicles	WB	13468126	Bathurst St / Front St W	Evening (17:00-22:00)	297	840
6954	18-May	Vehicles	WB	13468126	Bathurst St / Front St W	Morning (08:00-12:00)	297	413
6955	18-May	Vehicles	WB	13468126	Bathurst St / Front St W	PM Peak Period (16:00-19:00)	297	1333

Initial Data Analysis

After preparation, we decided to do some initial analysis to understand more about traffic volume and time. Specifically, our intent was to understand how traffic volume varied per time of day/direction, and how it trended over the span of the pilot.

We grouped the time periods together, then reset the index to display the results by category.

```
# exclude overlapped time-period
data_all_day = data[(data['time_period']=='Morning (08:00-12:00)') | (data['time_period']=='Afternoon (12:00-17:00)')|(data['
data_day_sum=data_all_day.groupby(['aggregation_period','classification']).passengers.sum()
data_day_sum=data_day_sum.to_frame()
```

```
data_day_sum = data_day_sum.reset_index()
data_day_sum
```

From Chart1A, we determined that baseline traffic volume between pedestrians and vehicles are approximately equal at 230,000 passengers while cyclist volume is a much smaller proportion at 15,000. There is a gradual trend of more pedestrian passengers as the study progresses. Pedestrian volume outpaces vehicles by 2x about 5 months into the study. It is worth noting though that over the winter months the gap was much narrower, although pedestrian traffic was still higher with the exception of January.

```
# The bar graph shows the total passangers of each time in each month
sns.set(rc = {'figure.figsize':(15,8)})
ax=sns.barplot(x="aggregation_period",y="passengers",hue="classification",data=data_day_sum,order = ['Baseline', '17-Dec',
```

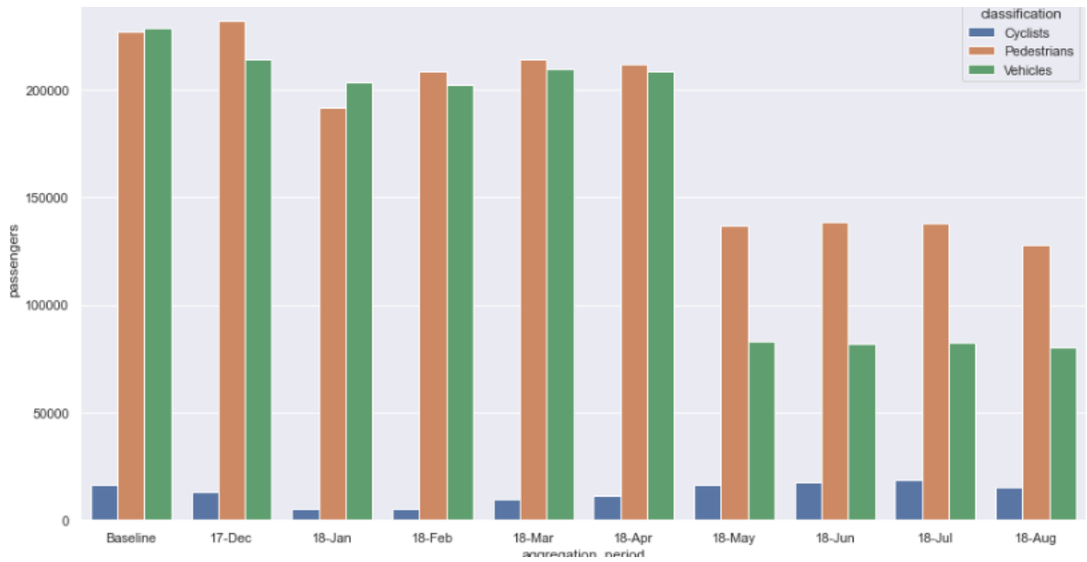


Chart 1A: Distribution of different types of traffic volume over the aggregation period

We also decided to compare the baseline traffic volume against the test pilot set within each time of day (Chart1B). In general, we found baseline traffic volume was heavier in off hours during the morning, afternoon, and evening; however, the test pilot period saw greater traffic during peak rush hour periods. This suggests that the test pilot was a success in optimizing traffic flow at key times during the day.

```
#This bar grouph is comparison of the two direction
sns.set(rc = {'figure.figsize':(15,8)})
ax=sns.barplot(x="aggregation_period",y="passengers",hue="dir",data=data_dir_sum,order = ['Baseline', '17-Dec', '18-Jan', '18
```

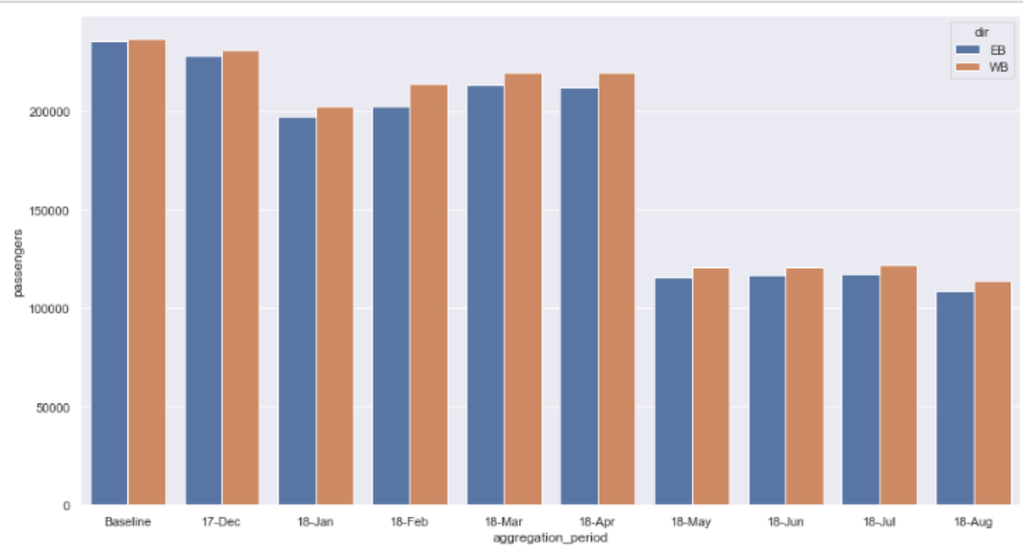


Chart 1B: Distribution of Baseline vs test pilot set of the traffic volume over the aggregation period

We analyzed this deeper by comparing the mode of traffic during the daily periods (Chart1C). Although pedestrian volume is always nearly higher, the largest gap occurred during the PM rush hour.

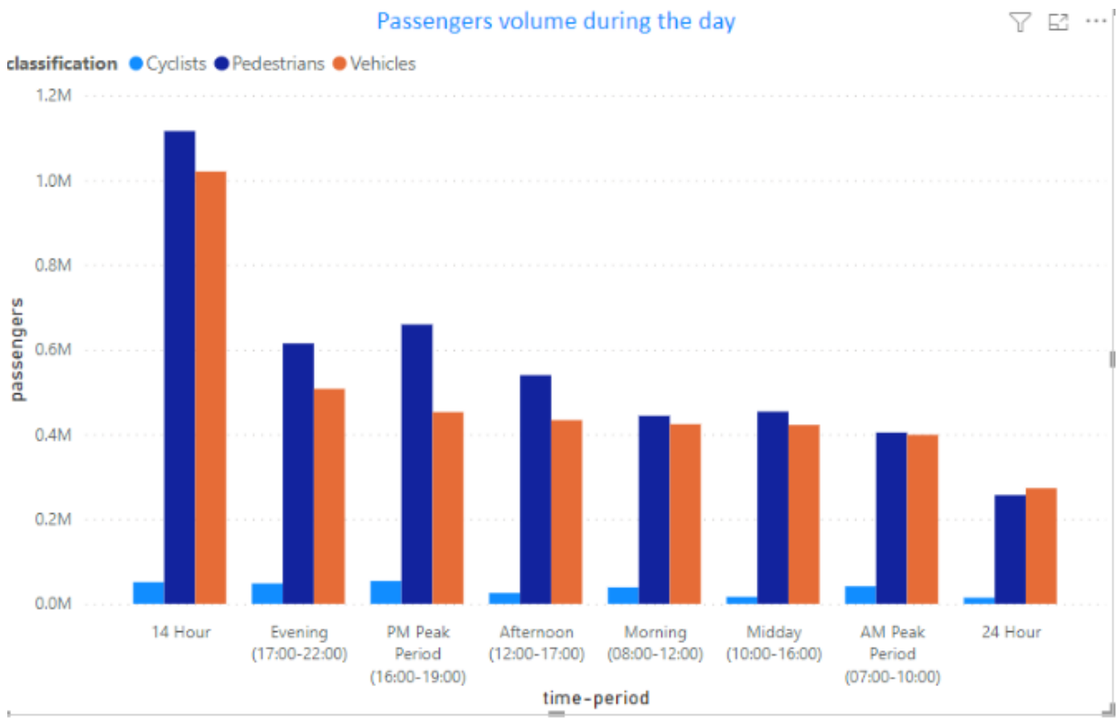


Chart 1C: Distribution of different types of traffic volume over specific time period

We also analyzed traffic volume by direction (eastbound by westbound) and didn't find any meaningful difference (Chart1D). Intuitively this makes sense. If most passengers in the city are commuting for a purpose such as going to work, then whichever direction they travel they will need to return by the opposite orientation.

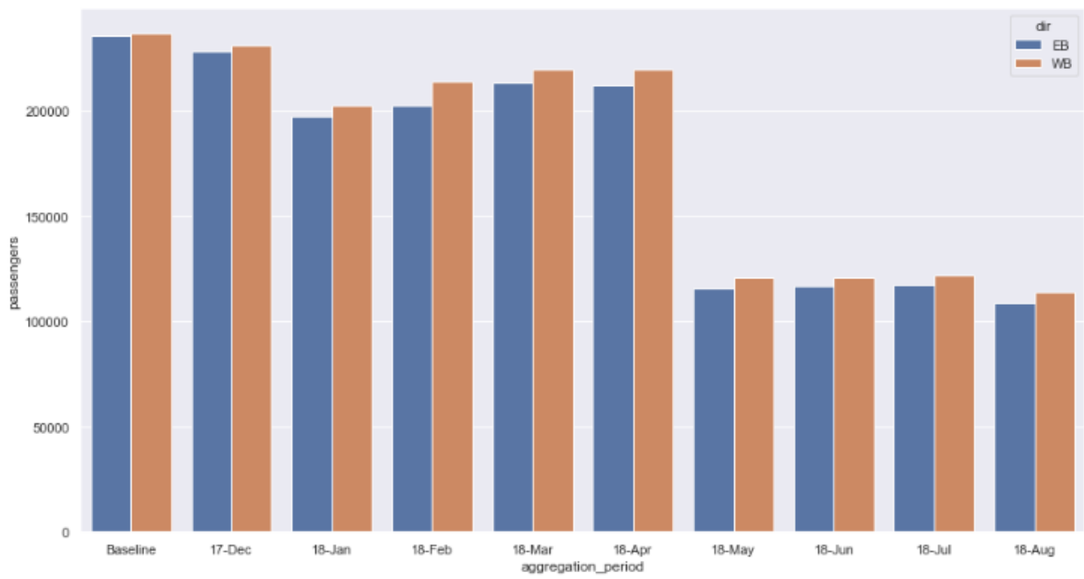


Chart 1D: Distribution of eastbound and westbound volume of traffic over the aggregation period

In these charts (1E/1F) we analyzed the travel time in the two directions east and west during the PM and AM time period. We also examined the wait time reliability, the wait time reliability by date range and the difference between low and high travel time by time. We determined that travel time is more important during the afternoon while the wait time reliability is more important in the morning.

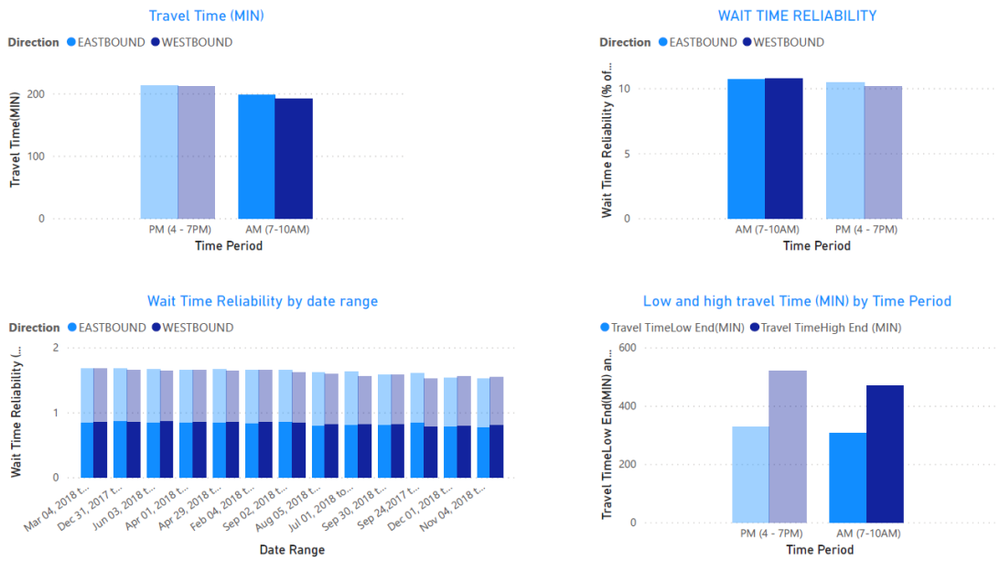


Chart 1E : Travel Time and Wait Time Reliability

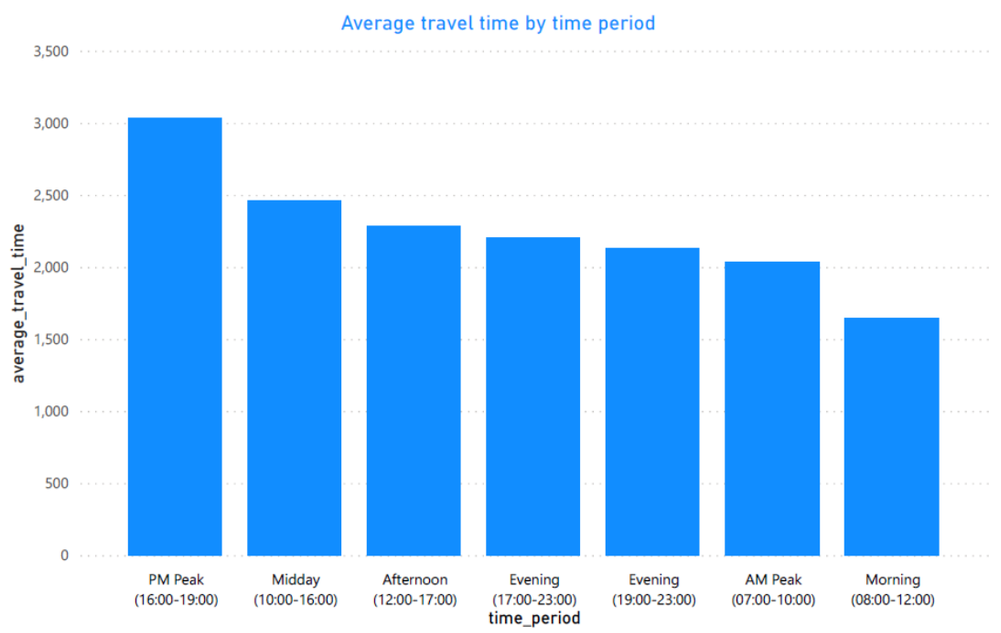


Chart 1F: Distribution of average travel time by time period

Finally, we also analyzed travel time over the duration of the test pilot (Chart1G). Since inception of the test to final data collection, there is a reduction of nearly 10 minutes. As the pilot project progressed, travel time continued to drop suggesting again that the pilot was a success in not only optimizing traffic flow during rush hour, but also reducing commute time for Toronto passengers.

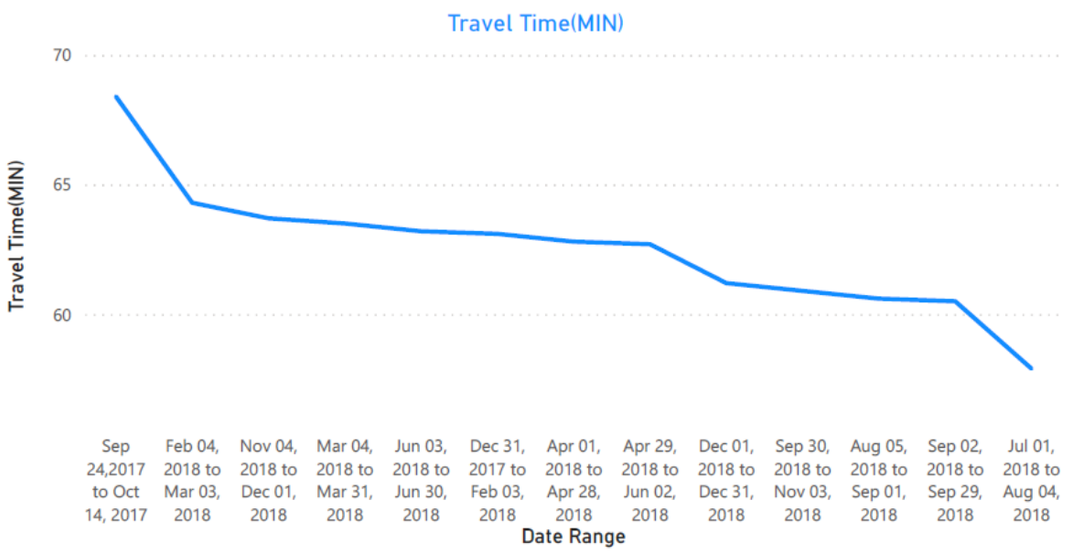


Chart 1G: Travel Time over the duration of test pilot over date ranges

In summary, given the reduction in traffic time, and greater passenger flow during peak rush periods it does appear that the test pilot was a success. We utilized different modeling techniques to explore further.

Predictive Modeling (Detailed Scripts in Appendix)

As discussed, the goal of our modeling is to determine if we can predict the traffic volume of different transportations given an intersection name, direction, and time of the day. We also want to know how accurate these predictions could be.

We randomly shuffled the data and split the training and testing sets by an approximate 80/20 ratio. Two models, OLS and K-NN were trained on the training set and tested on the testing (validation) set. The metric to compare the two models is RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|^2}{N}},$$

The model with lower RMSE is the better model.

The first step is to one-hot encode all categorical variables, so the models can understand and take them as input. That is converting categorical variables with n levels to $n - 1$ columns of 0s and 1s.

The second step is to drop id, as ids don't contain any useful information, and it's likely to introduce bias or over-fitting when id is used as an explanatory variable.

Linear Model:

As the response variable here (volume) is numerical, it makes sense to use a linear model (scrips in appendix). As a result, the RMSE is 1558.

K-NN:

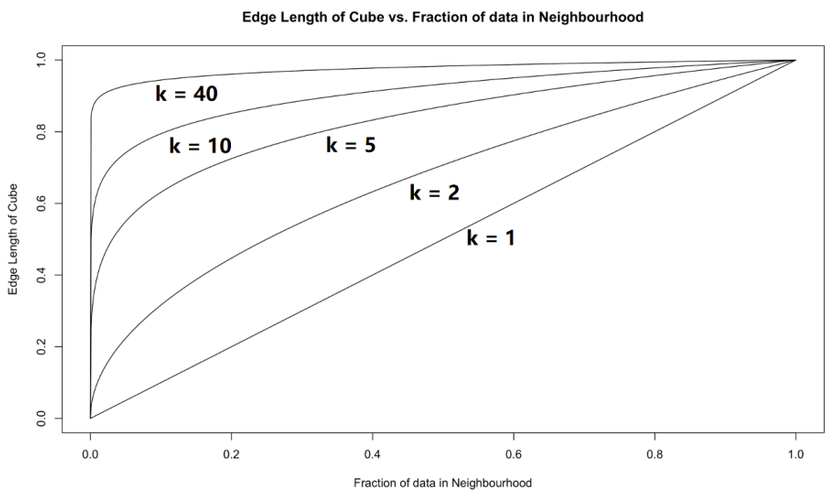
The second model used is K-NN regression. K-NN works by calculating the Euclidean distances between all observations in the test dataset and all observations in the training dataset, therefore for each value to be predicted in the testing set, the prediction is the average of the k closest observations in the training set. In our case, data is already one-hot encoded, so we can directly feed it into K-NN.

From the output, 23 seems to be the best k-value.

However, given the number of dimensions of the data, K-NN is not a good choice since K-NN doesn't work very well in high dimensions due to the "Curse of Dimensionality". The basic assumption K-NN holds is that similar observations (smaller Euclidean distance) share similar value. However, in high dimensions, points are almost never close to each other. If they are not close to each other, they won't share the same label. Imagine we have a unit square (2-dimension) with uniformly distributed points in it, and we want to circle 10% of the total data with a smaller square, we will need the smaller square to have an edge length of 0.33, while that number is 0.1 in 1-dimension space. In 3-dimension space, if we want to circle (contain) 10% of the points with a smaller cube in a unit cube, the edge of the smaller cube will be 0.46. The relationship between the edge length of the smaller cube and dimension follows:

$$E(f) = f^{1/D},$$

where E is the edge length, f the fraction of observations, and D the dimension (plot below).



There is no way to solve this problem. However, there are ways to get around it.

1. Reduce the dimension
2. Use a smaller value of k. This will lead to overfitting. Say if we fit a 1-NN, it might perform super well on the testing set. However, that's a "fake " good performance, the model is essentially over-fitting itself. This is also sensitive to noise and outliers. The key of k-NN is to allow the model to take the average/majority vote of different individuals in the training set, not to look for exact same observations. Normally $\sqrt{N}$ is a good start to avoid over-fitting.

In our case, the k-value we should choose should be around 80, instead of 23, which gives the best RMSE but may lead to overfitting. RMSE when k is around 80 is about 1700. That means, the linear model is performing better on this dataset with a smaller RMSE.

Therefore, with our model, it is possible to predict the traffic volume with a pretty high accuracy given the variables in the dataset. So, for any timeframe in the future, with the

information collected for each intersection, we will be able to predict what the volume for different types of transportations will be, hence aid with decision making.

Conclusion

Ultimately, we determined that the test pilot was a success and created a model in order to predict traffic volume to aid with municipal planning. Ultimately this is what came to fruition. In April 2019, the city of Toronto elected to extend the pilot and other large cities including New York have taken inspiration to build their own private corridors [5]. Large cities were not designed with the infrastructure to accommodate modern day traffic volumes, but by utilizing statistical analysis and data science techniques we can optimize and improve urban development.

References

[1] Annual Summary Dashboard published as a part of Data Reports and Background Materials by the City of Toronto Data Reports and Background Materials – City of Toronto

[2] CBC Article " King Street Pilot to Launch In November But Not Everyone's Happy About It." Published Oct 26 , 2017.

https://www.cbc.ca/news/canada/toronto/king-street-pilot-to-launch-in-november-but-not-everyone-s-happy-about-it-1.4371960

[3] Toronto Transportation Services, "King St. Transit Pilot - Detailed Traffic & Pedestrian Volumes" (October & November 2017). Distributed by City of Toronto Open Data. https://open.toronto.ca/dataset/king-st-transit-pilot-traffic-pedestrian-volumes-summary/#sent/_blank

[4] Toronto Transportation Services, " King St.Transit Pilot-Bluetooth Travel Time" (October & November 2017). Distributed by City of Toronto Open Data.

https://ckan0.cf.opendata.inter.prod-toronto.ca/sk/dataset/king-st-transit-pilot-detailed-bluetooth-travel-time

[5] CBC Article " King Street Pilot Project Extended Until End Of July 2019 " Published December 13 ,2018.

https://globalnews.ca/news/4759182/king-street-pilot-project-extended/

Appendix

Modeling Code

```
types = []
newdata = data["_id"]
#handeling categorical variables (one-hot)
for i in range(0, data.shape[1]):
    if type(data[data.columns[i]][1]) == str:
        temp = pd.get_dummies(data=data[data.columns[i]], drop_first=True)
        col_names = []
        for j in range(0, temp.shape[1]):
            col_names.append(";".join([data.columns[i], temp.columns[j]]))
        temp.set_axis(col_names, axis=1, inplace=True)
    else:
        temp = data[data.columns[i]]
    newdata = pd.concat([newdata, temp], axis=1)

newdata = newdata.drop(labels='_id', axis=1)

#linear model
X = np.array(newdata.iloc[:, 0:newdata.shape[1]-1])
Y = np.array(newdata.iloc[:, newdata.shape[1]-1])
x_train, x_test, y_train, y_test = sms.train_test_split(X, Y, shuffle = True)
lm = sm.OLS(y_train,x_train)
regressior = lm.fit()
y_pred = regressior.predict(x_test)
print(regressior.summary2())

sq_diff = []
abs_diff_pct = []
for i in range(0,len(y_pred)):
    sq_diff.append((y_pred[i] - y_test[i])**2)
    abs_diff_pct.append(abs((y_pred[i] - y_test[i])/y_test[i]))
rmse = np.sqrt(sum(sq_diff)/len(sq_diff))
mape = sum(abs_diff_pct)/len(abs_diff_pct)
rmse
mape

#knn
to_test = [num for num in range(round(np.sqrt(x_train.shape[0])-50),
round(np.sqrt(x_train.shape[0])+50)) if num %2 == 1]
rmse_knn = []

for i in range(0,len(to_test)):
    sq_diff = []
    diff = []
    for j in range(0,10):
        X = np.array(newdata.iloc[:, 0:newdata.shape[1]-1])
        Y = np.array(newdata.iloc[:, newdata.shape[1]-1])
        x_train, x_test, y_train, y_test = sms.train_test_split(X, Y, shuffle =
True)
        model = KNeighborsRegressor(n_neighbors=to_test[i]).fit(x_train, y_train)
        predicted = model.predict(x_test)
        for k in range(0, len(predicted)):
            sq_diff.append((predicted[k] - y_test[k])**2)
            abs_diff_pct.append(abs((predicted[k] - y_test[k])/y_test[k]))

    rmse_knn.append(np.sqrt(sum(sq_diff)/len(sq_diff)))
    mape_knn.append(sum(abs_diff_pct)/len(abs_diff_pct))

best_k_val = to_test[np.argmin(rmse_knn)]
```