Coursera Capstone – “School-Classification”
Comparative analysis using location data via Foursquare

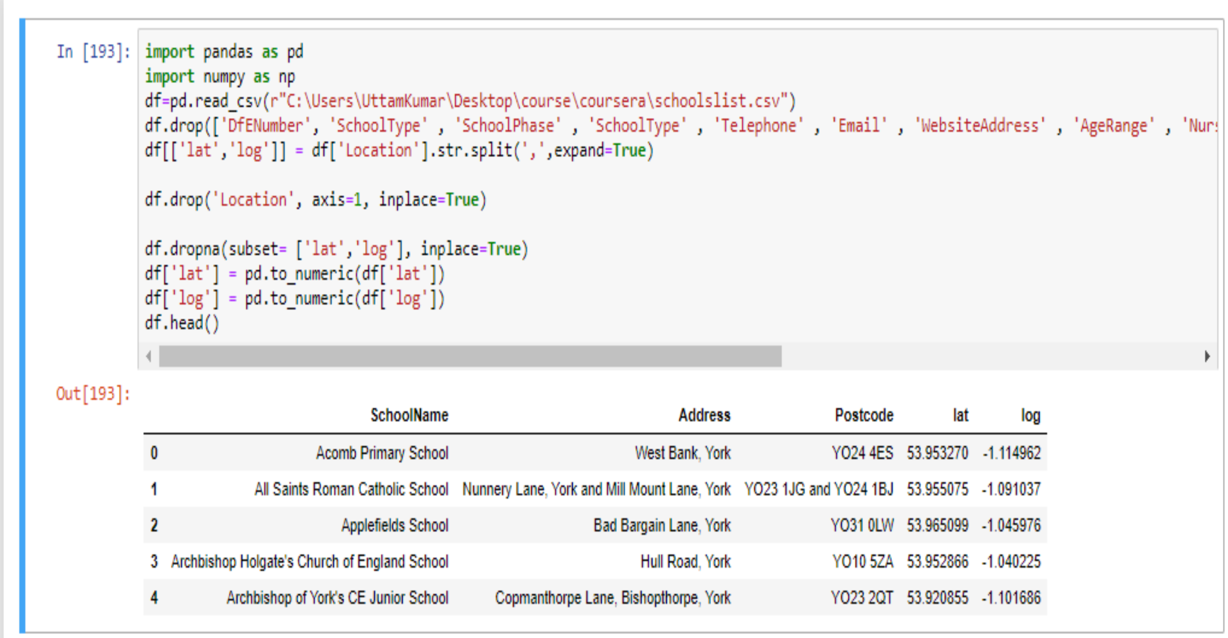
Introduction

In today’s world, it is a headache for parents to decide which school is good for their children. It is very important for a child to have his/her education at a better place so that he/she should be influence by bad things nearby. The mission of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the ‘best’ school in York for a child. My proposal, then, is an analysis of the neighborhoods of schools in York for the consideration of safe place for children. The objective is to evaluate the neighborhoods of schools and analyses whether there is any venue nearby which can be harmful for a child.

This project will be interesting for parents those are willing to send their children to school and yet little bit of confuse about which school will be safe for their children.

Data Overview

The data that we will use for this analysis is a combination of a CSV file that has been prepared for the purposes of the analysis from multiple sources (schoolslist.csv) and the location/venue information in foursquare. Schoolslist.csv has list of schools in York along with their latitude and longitude. Using foursquare APIs, we can get the details and category of the neighborhood venues and using this we can classify the schools whether it is safe or unsafe for a child using SVM linear classification algorithm.



In [193]:

```
import pandas as pd
import numpy as np
df=pd.read_csv(r"C:\Users\UttamKumar\Desktop\course\coursera\schoolslist.csv")
df.drop(['DfENumber', 'SchoolType' , 'SchoolPhase' , 'SchoolType' , 'Telephone' , 'Email' , 'WebsiteAddress' , 'AgeRange' , 'Nur
df[['lat','log']] = df['Location'].str.split(',',expand=True)

df.drop('Location', axis=1, inplace=True)

df.dropna(subset= ['lat','log'], inplace=True)
df['lat'] = pd.to_numeric(df['lat'])
df['log'] = pd.to_numeric(df['log'])
df.head()
```

Out[193]:

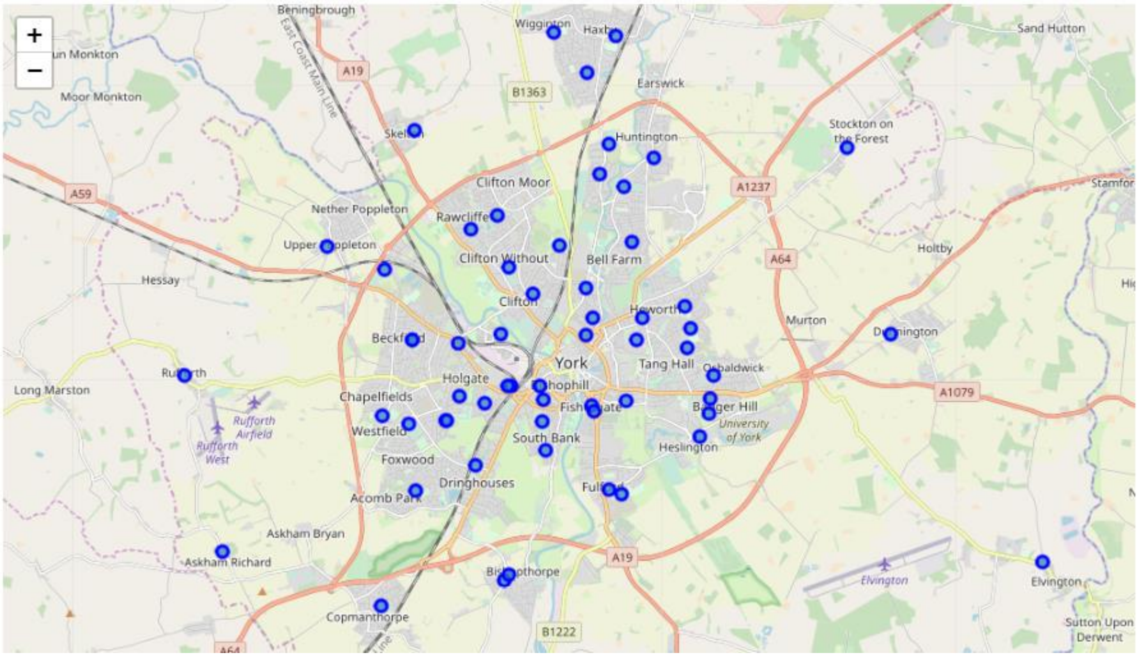
	SchoolName	Address	Postcode	lat	log
0	Acomb Primary School	West Bank, York	YO24 4ES	53.953270	-1.114962
1	All Saints Roman Catholic School	Nunnery Lane, York and Mill Mount Lane, York	YO23 1JG and YO24 1BJ	53.955075	-1.091037
2	Applefields School	Bad Bargain Lane, York	YO31 0LW	53.965099	-1.045976
3	Archbishop Holgate's Church of England School	Hull Road, York	YO10 5ZA	53.952866	-1.040225
4	Archbishop of York's CE Junior School	Copmanthorpe Lane, Bishopthorpe, York	YO23 2QT	53.920855	-1.101686

Methodology section

Exploratory data analysis:-

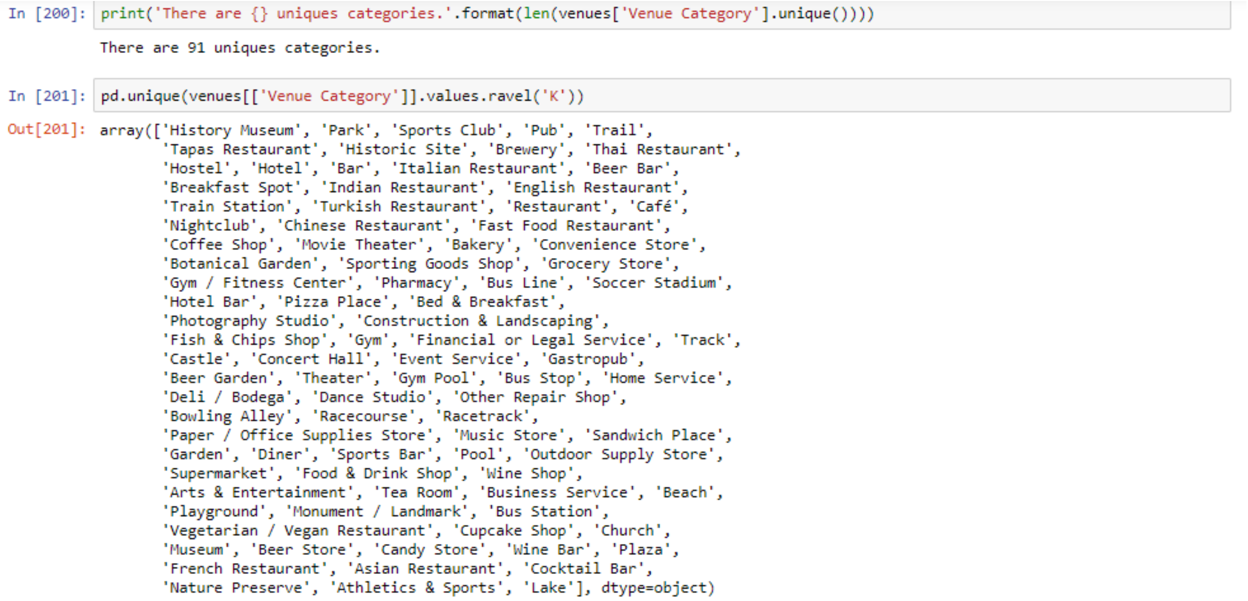
I have done the data cleaning n plotted the schools on the map of York by using folium library of python.

Out[195]:

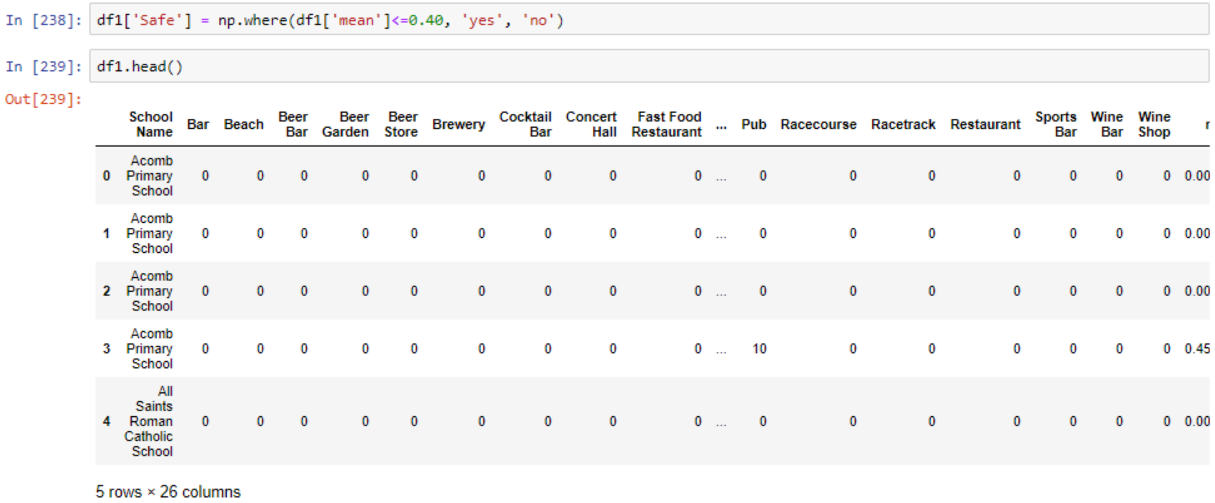


After getting the coordinates of schools, I have use foursquare APIs to get the details of nearby venues of schools. There are total 91 different categories of venues near school. Out of these categories, we will consider only those venues which will give bad impact on the school children such as wine shops, bar, pub etc.

After removing all the unwanted venue, I have done one-hot encoding for each school. If a school surrounded by any of the following venue, than the corresponding value becomes 10 otherwise 0. After this, I have calculated the mean value for each school and by considering that if any of the following venue is present nearby a school, it will be classified as unsafe for children.



```
In [200]: print('There are {} uniques categories.'.format(len(venues['Venue Category'].unique())))

          There are 91 uniques categories.

In [201]: pd.unique(venues[['Venue Category']].values.ravel('K'))

Out[201]: array(['History Museum', 'Park', 'Sports Club', 'Pub', 'Trail',
                 'Tapas Restaurant', 'Historic Site', 'Brewery', 'Thai Restaurant',
                 'Hostel', 'Hotel', 'Bar', 'Italian Restaurant', 'Beer Bar',
                 'Breakfast Spot', 'Indian Restaurant', 'English Restaurant',
                 'Train Station', 'Turkish Restaurant', 'Restaurant', 'Café',
                 'Nightclub', 'Chinese Restaurant', 'Fast Food Restaurant',
                 'Coffee Shop', 'Movie Theater', 'Bakery', 'Convenience Store',
                 'Botanical Garden', 'Sporting Goods Shop', 'Grocery Store',
                 'Gym / Fitness Center', 'Pharmacy', 'Bus Line', 'Soccer Stadium',
                 'Hotel Bar', 'Pizza Place', 'Bed & Breakfast',
                 'Photography Studio', 'Construction & Landscaping',
                 'Fish & Chips Shop', 'Gym', 'Financial or Legal Service', 'Track',
                 'Castle', 'Concert Hall', 'Event Service', 'Gastropub',
                 'Beer Garden', 'Theater', 'Gym Pool', 'Bus Stop', 'Home Service',
                 'Deli / Bodega', 'Dance Studio', 'Other Repair Shop',
                 'Bowling Alley', 'Racecourse', 'Racetrack',
                 'Paper / Office Supplies Store', 'Music Store', 'Sandwich Place',
                 'Garden', 'Diner', 'Sports Bar', 'Pool', 'Outdoor Supply Store',
                 'Supermarket', 'Food & Drink Shop', 'Wine Shop',
                 'Arts & Entertainment', 'Tea Room', 'Business Service', 'Beach',
                 'Playground', 'Monument / Landmark', 'Bus Station',
                 'Vegetarian / Vegan Restaurant', 'Cupcake Shop', 'Church',
                 'Museum', 'Beer Store', 'Candy Store', 'Wine Bar', 'Plaza',
                 'French Restaurant', 'Asian Restaurant', 'Cocktail Bar',
                 'Nature Preserve', 'Athletics & Sports', 'Lake'], dtype=object)
```



```
In [238]: df1['Safe'] = np.where(df1['mean']<=0.40, 'yes', 'no')

In [239]: df1.head()

Out[239]:
```

	School Name	Bar	Beach	Beer Bar	Beer Garden	Beer Store	Brewery	Cocktail Bar	Concert Hall	Fast Food Restaurant	...	Pub	Racecourse	Racetrack	Restaurant	Sports Bar	Wine Bar	Wine Shop	r
0	Acomb Primary School	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0.00
1	Acomb Primary School	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0.00
2	Acomb Primary School	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0.00
3	Acomb Primary School	0	0	0	0	0	0	0	0	0	...	10	0	0	0	0	0	0	0.45
4	All Saints Roman Catholic School	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0.00

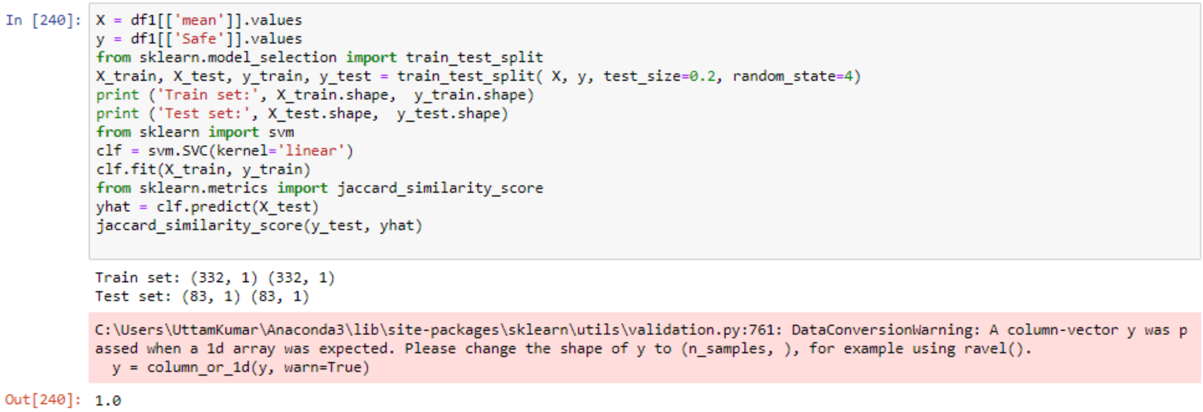
5 rows × 26 columns

Classification model:-

In classification model, I have use Linear SVM which classifies the data by making a linear separator between different classes.

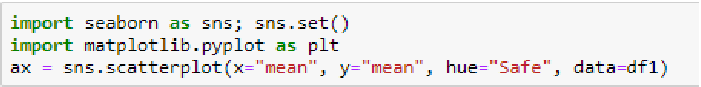
As we need to classify the data into two group i.e. whether a school is safe or not (binary classification) based on the nearby venues, I have use SVM machine learning algorithm because it is the best algorithm for binary classification.

It also has very high accuracy which is another reason to use this algorithm.

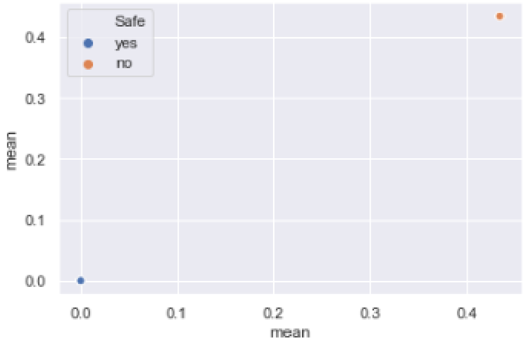


```
In [240]: X = df1[['mean']].values
          y = df1[['Safe']].values
          from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4)
          print ('Train set:', X_train.shape,  y_train.shape)
          print ('Test set:', X_test.shape,  y_test.shape)
          from sklearn import svm
          clf = svm.SVC(kernel='linear')
          clf.fit(X_train, y_train)
          from sklearn.metrics import jaccard_similarity_score
          yhat = clf.predict(X_test)
          jaccard_similarity_score(y_test, yhat)

          Train set: (332, 1) (332, 1)
          Test set: (83, 1) (83, 1)

          C:\Users\UttamKumar\Anaconda3\lib\site-packages\sklearn\utils\validation.py:761: DataConversionWarning: A column-vector y was p
          assed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
            y = column_or_1d(y, warn=True)

Out[240]: 1.0
```

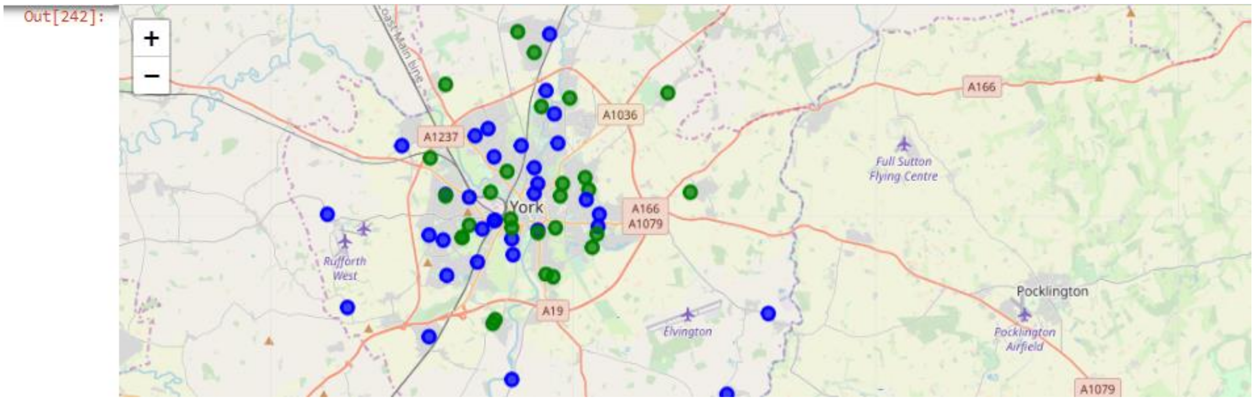
Hence, the accuracy here is 1.0



```
import seaborn as sns; sns.set()
import matplotlib.pyplot as plt
ax = sns.scatterplot(x="mean", y="mean", hue="Safe", data=df1)
```



Result



From the above result, it shows some schools with green color and some schools with blue color. By using SVM linear model, I have classified all the unsafe schools by labelling them with blue color and all the schools with green color label are safe for children as there is no harmful venue around the school.

Discussion Section

As we observed that schools are classified on the basis of surrounding venues, so to improve the no. of safer schools, we can suggest some policies to government so that there is no harmful venues around 500 meter radius of a school.

Conclusion

In this study, I have analyzed the relation between schools and their surrounding with the help of foursquare APIs. I have identified the safer schools in the York city based on the different venues around school. I have built SVM linear model as classification model which can classify schools whether it is safe or not with a very high accuracy.