Predict whether to Invest a Pub Business in Canada

Eric M. H. Goh

SVBook.com and EMHAcademy.com

Author Note

[Include any grant/funding information and a complete correspondence address.]

Abstract

This data mining project aims to predict or classify whether we can invest in a Pub business in Canada. The project will use data from Wikipedia and the FourSquare Database and use data mining process like the CRISP DM process model to create the classification model. The project will go through the webscraping and data extraction from Wikipedia and FourSquare, and data understanding using data visualizations and clustering, creating models using Decision Tree, SVM, Logistic Regressions and other algorithms. The project will then proceed to evaluation of the models and select the most accurate models.

Keywords: [Click here to add keywords.]

Predict whether to Invest a Pub Business in Canada

# Introduction

This project will create a prediction or classification model to predict whether we should invest in the pub business in Canada. The project will go into the data mining process CRISP DM process model, where we will understand the business first, followed by data understanding, then data preparation, modeling, evaluation and deployment.

The link to the Python file is at: <https://github.com/gohminghui88/Coursera_Capstone/blob/master/Coursera_Capstone_final1.ipynb> Data is where we will be showing the two data preparation stages. The result portion is the data understanding stage and the model evaluation stage.

## Business Understanding

We would like to invest a pub business in Canada, and we would need to know about the risk and whether it is safe to invest in the location. We would like to know whether there is a way to predict whether it is safe to invest in a Pub business based on a location

People Interested

* Decision Makers
* Management

Who to be doing this data mining project

* Data Analyst, Data Scientist
* Domain Expert in the data sets

### Data.

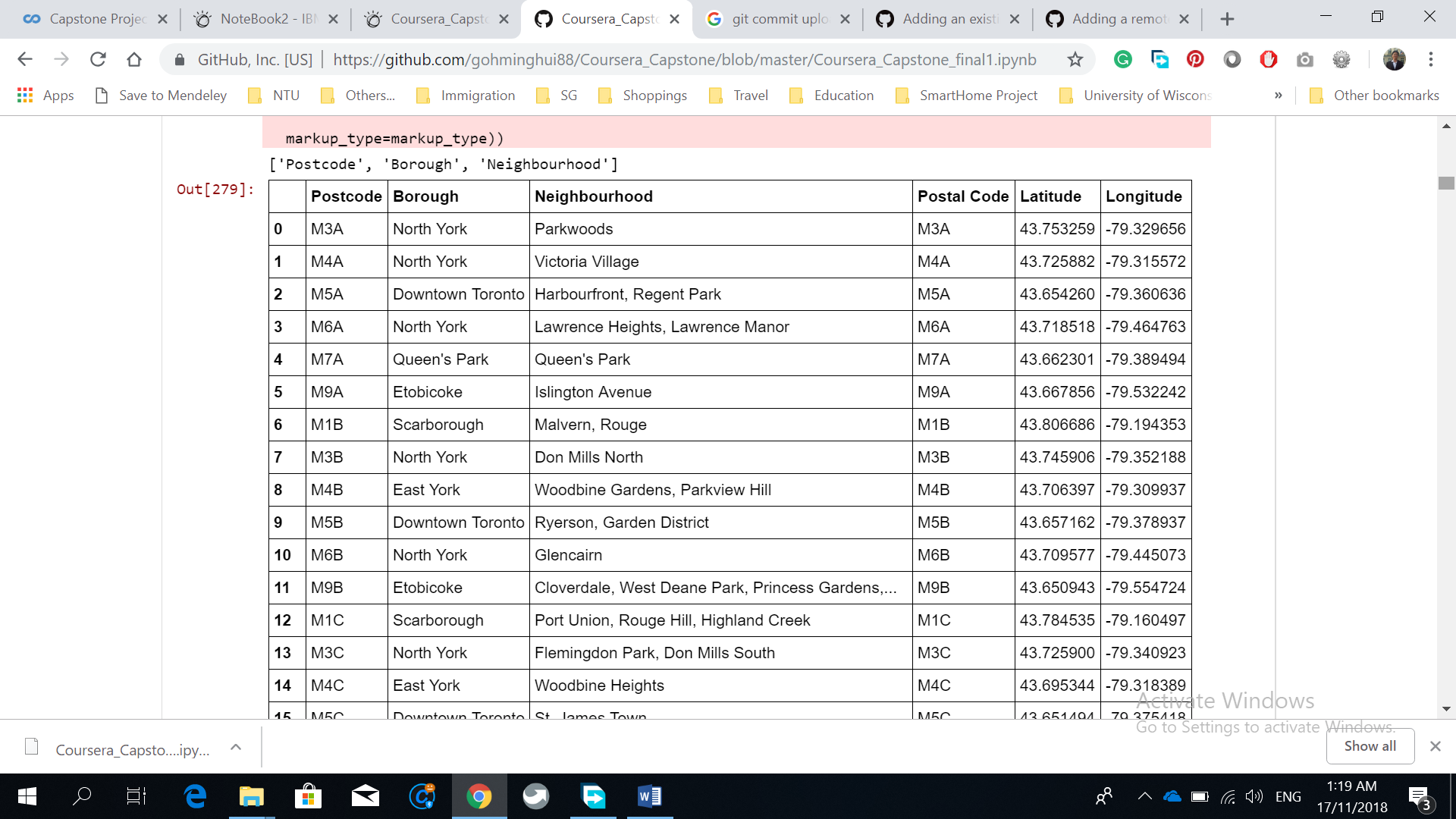
The data will be extracted form the internet, the Wikipedia, using webscraping libraries from Python. We will then extract the longitude and latitude based on another dataset, and we will extract data about shoppings, restuarants, pub, massage from the FourSquare Database using the Lagitude and Longitude. The above is the first Data Preparation. For second data preparation, based on the data, we will then filter out the Pub data, and count how many Pub companies are there in each borough and neighbourhood. We will then based on the count of Pub companies and apply a few rules

* IF number of Pub is less than 10, THEN it is risky to invest a pub business
* IF number of Pub is more than 20, THEN it is risky to invest a pub business due to competitions

After we apply the rules, we will get the isSetUp variables. We will then select the Longitude and Latitude variables, and the iSetUp variables as our data to train a prediction or classification models. When we enter the Longitude and Latitude, the model can predict or classify whether we should invest a Pub business there.

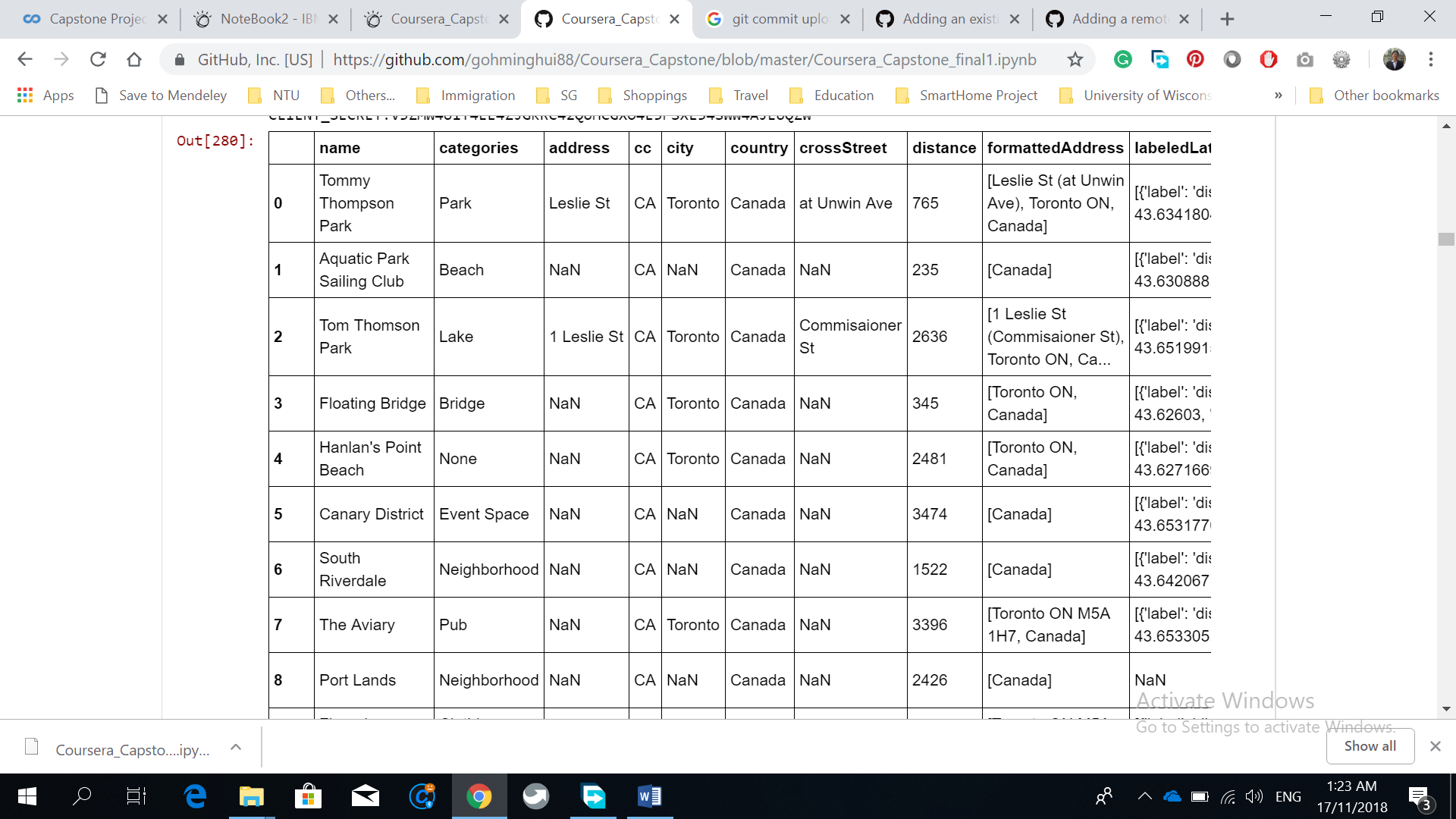
Data scraped from the "https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M"

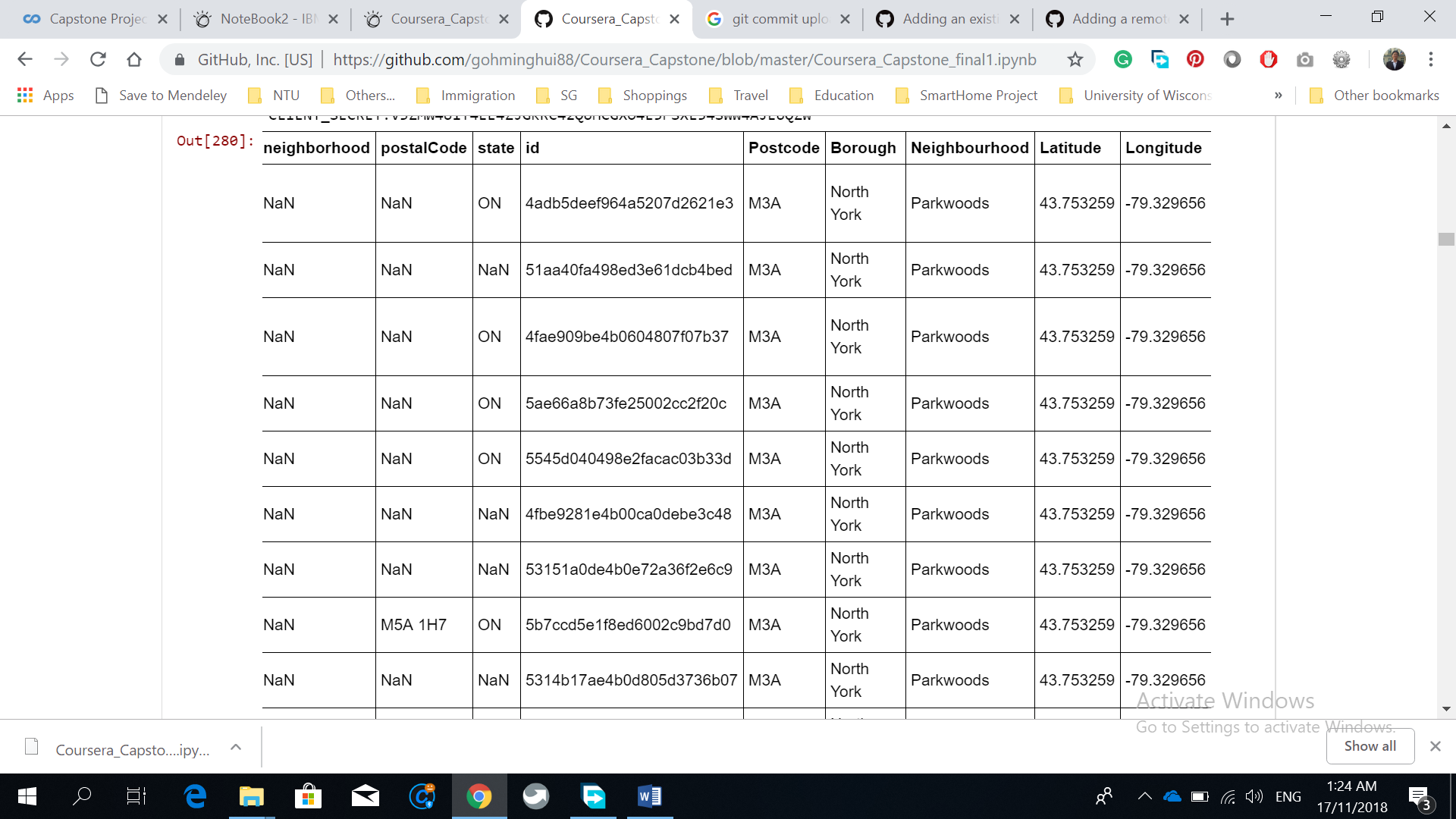
and combined with the Latitude and Longitude file based on the postcode. The longitude and latitude can also be retrieved using geocode libraries also.



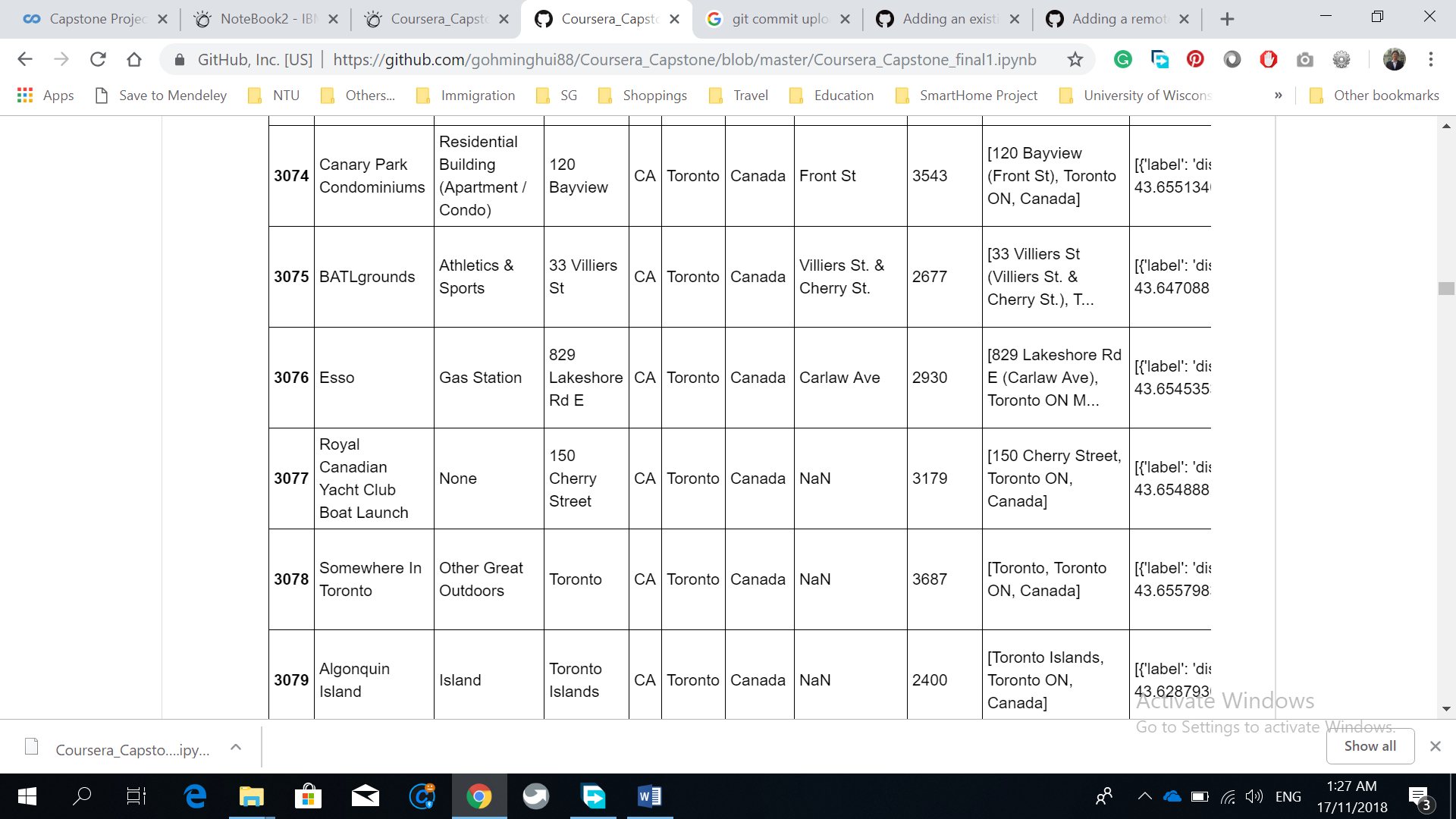
Consequently, each Longitude and Latitude is used to search for the shoppings, restaurants and Pubs from the FourSquare Data.

Latitude 43.75, Longitude -79.32



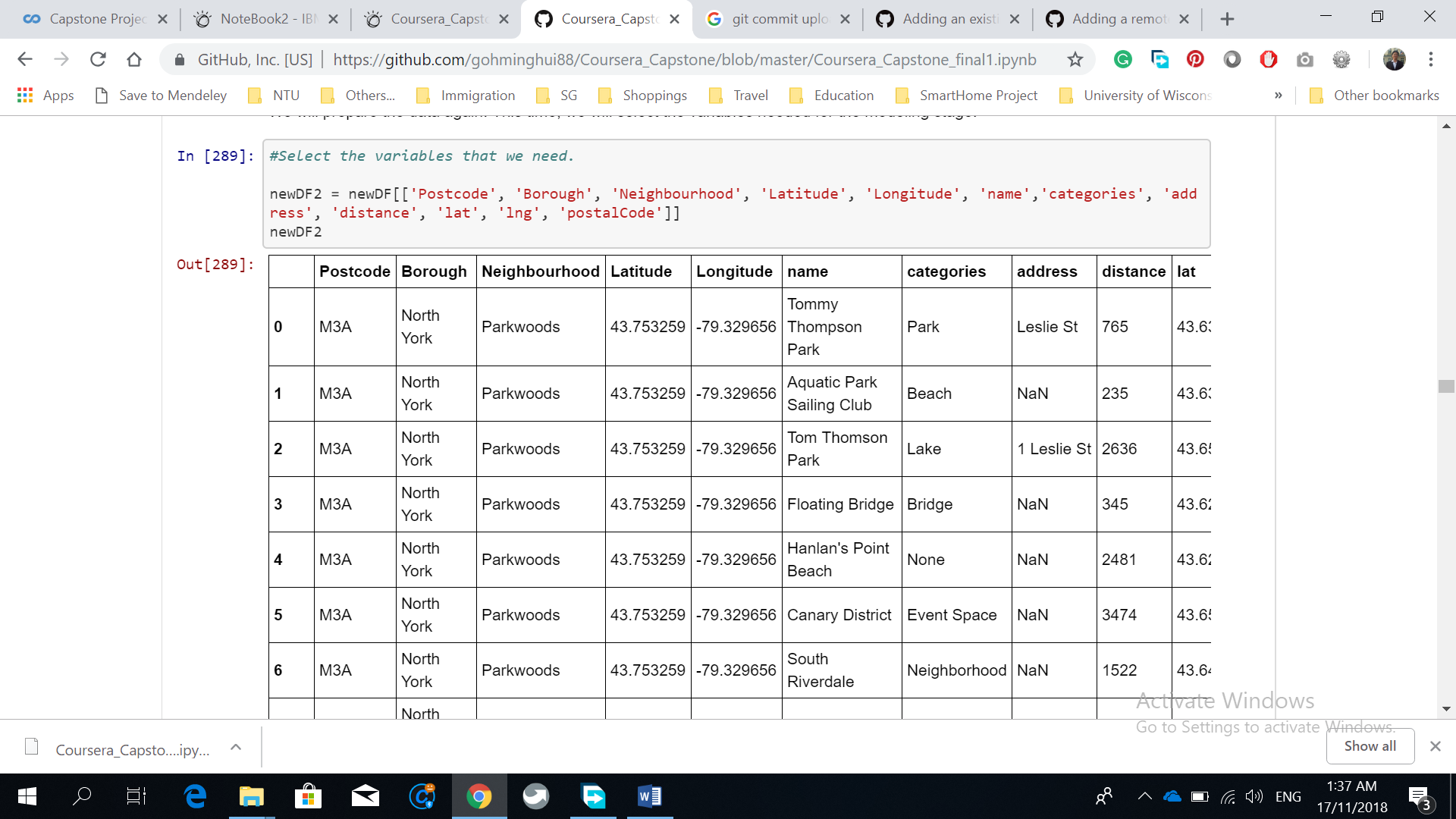


Latitude 43.62, Longitude -79.52

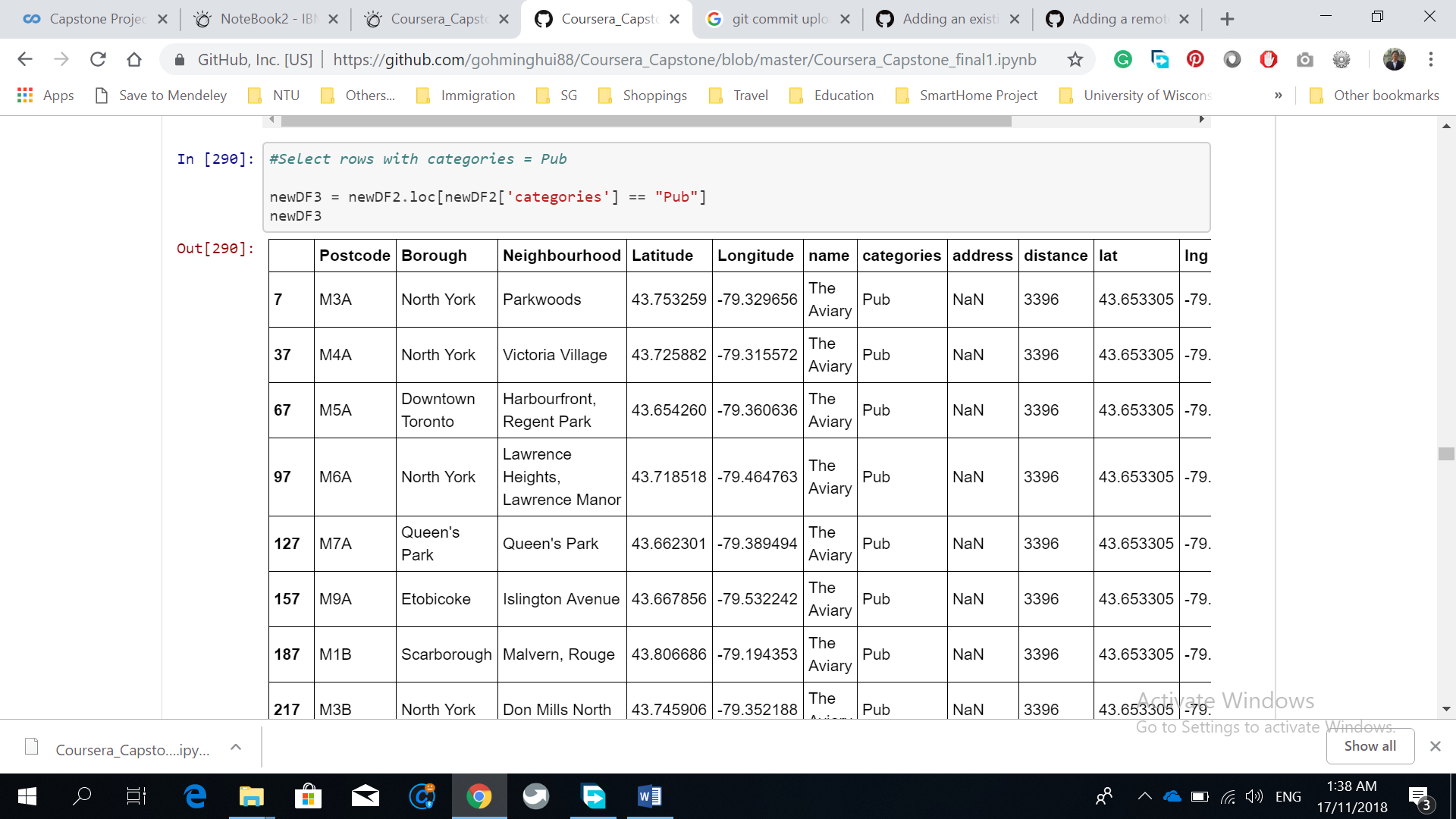




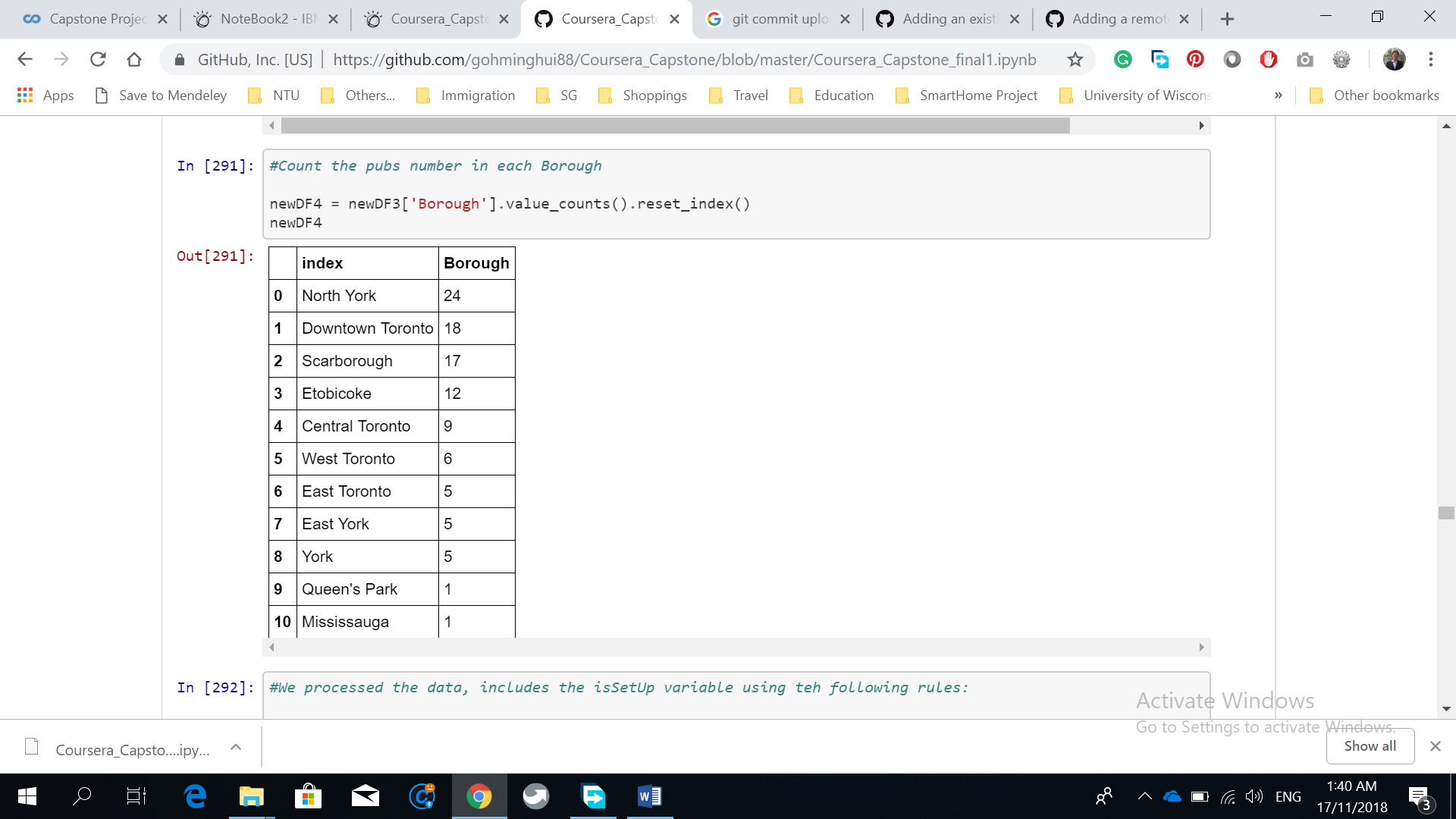
We will then Select the variables and arrange them in the order we want



We will then Select the observations for Pub only



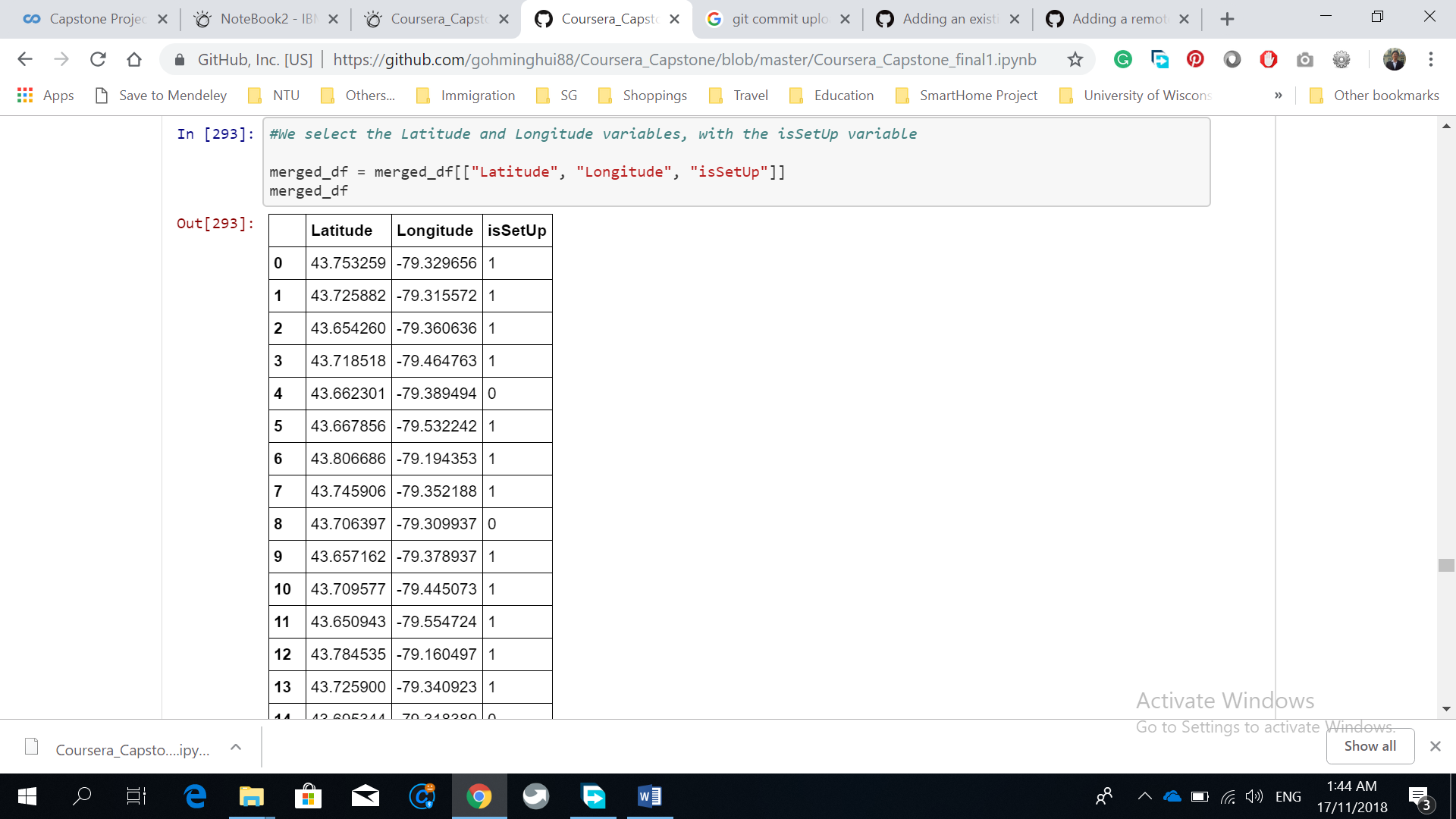
We will then Count the pubs number in each borough



We will then Add in the isSetUp variable



We will then Select the Latitude, Longitude and isSetUp variables to train our models

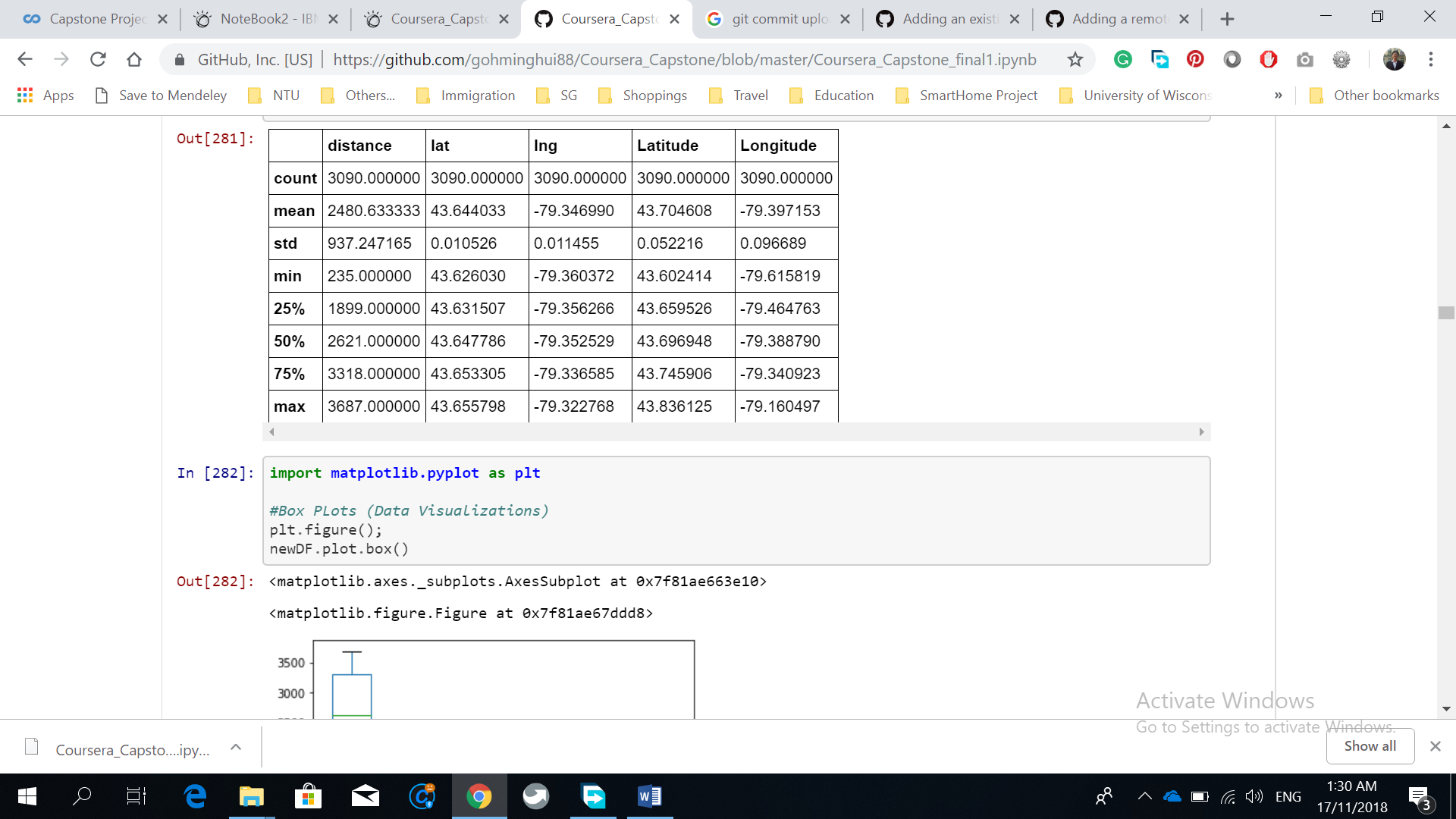


#### Methdology.

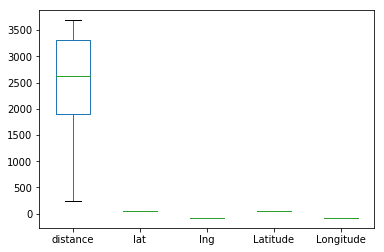
We will use the CRISP DM model to create the data mining models. We will understand the business first, then we will do some data understanding and data exploration to understand about the data, then will go into data preparation to prepare or process our data for creating the prediction or classification models. At the modeling stage, we will create KNN, Decision Tree, SVM, and logistic models from the dat. We will then evaluate the models, getting the confusion matrix, the precision, recall, the accuracy, and the we will select the most accurate models for creating into a software to predict whether we should invest in a pub business based on the longitude and Latitude or the locations.

##### **Results**

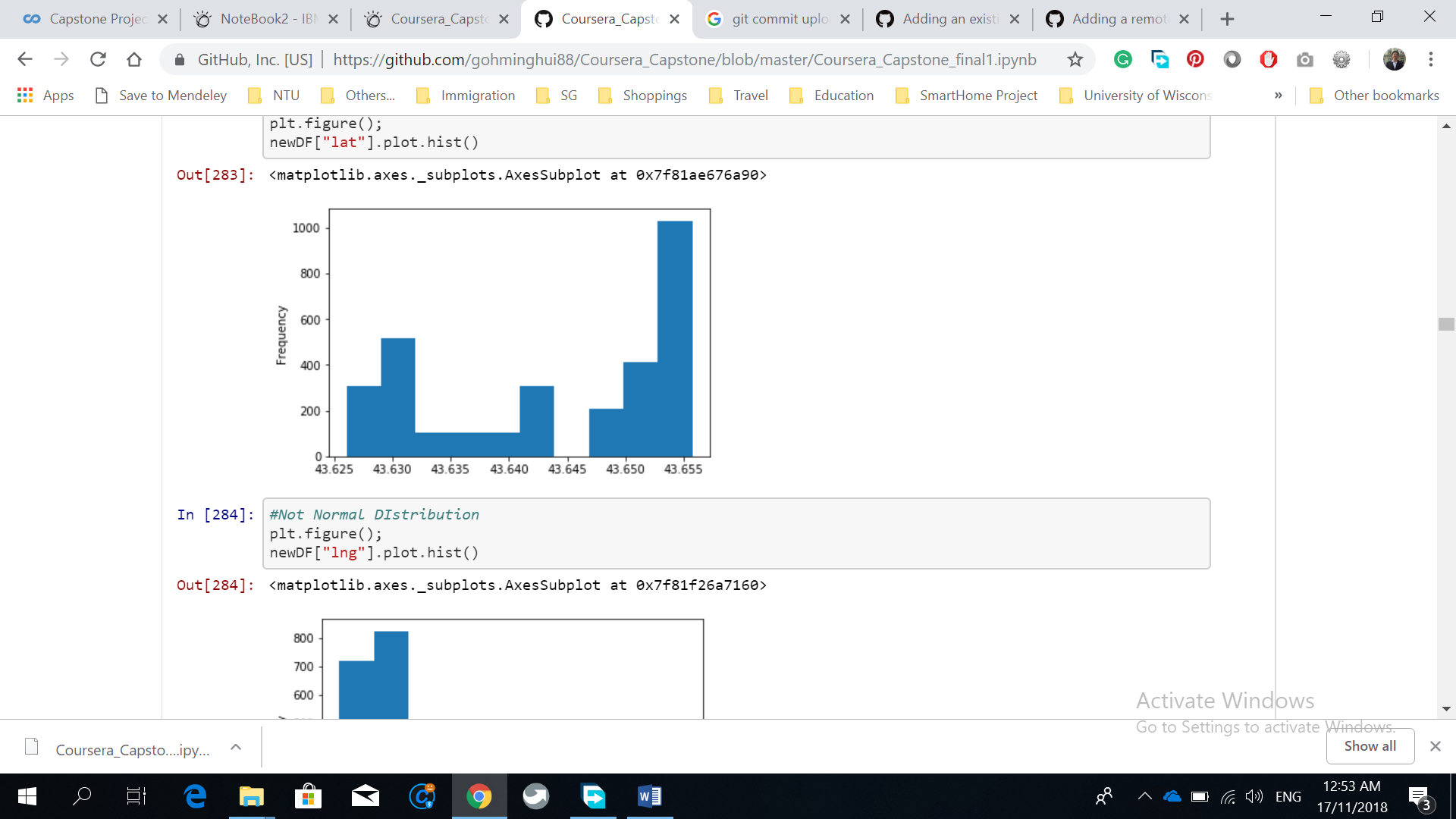
Descriptives Statistics



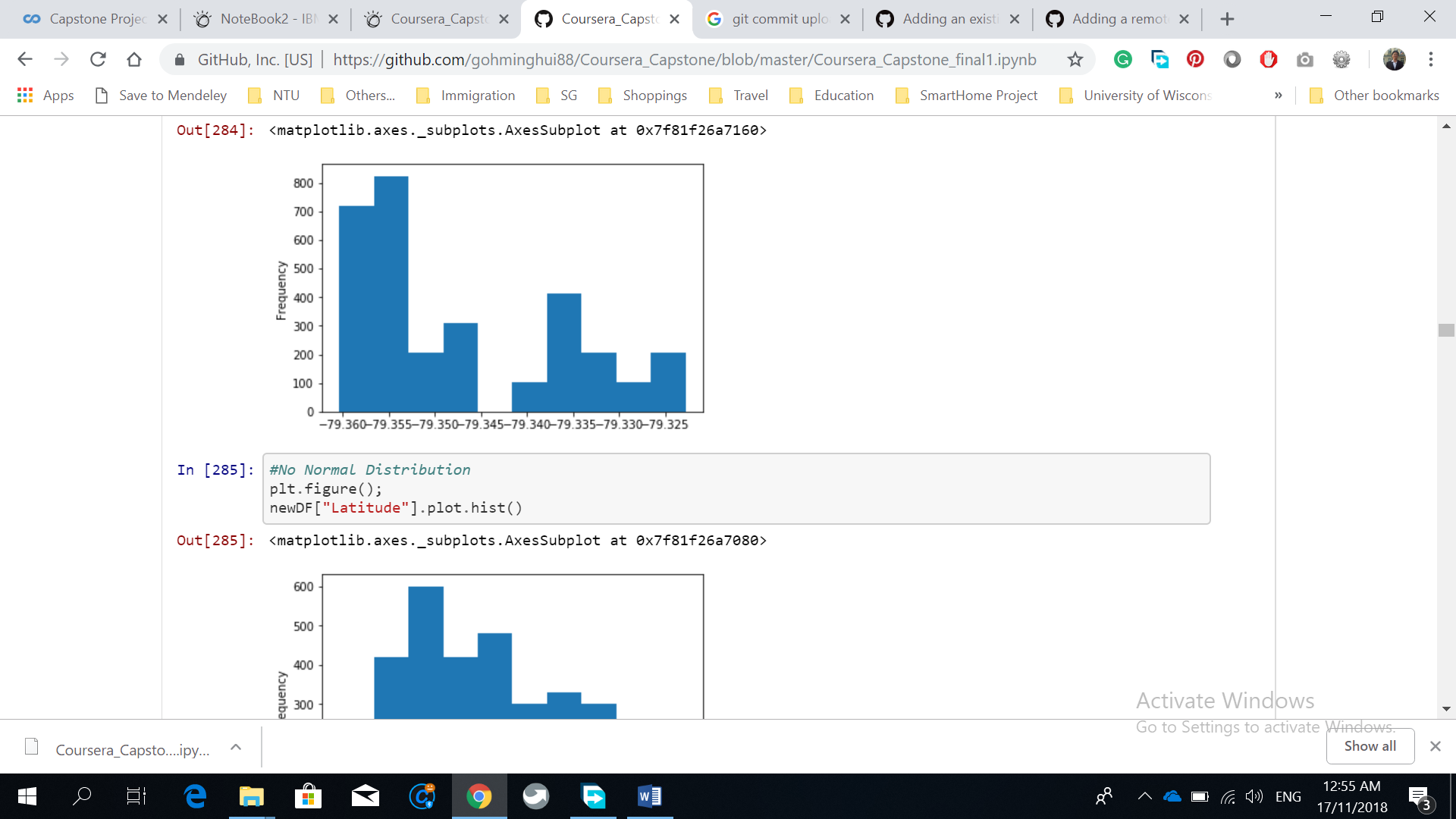
Boxplot of all variables:



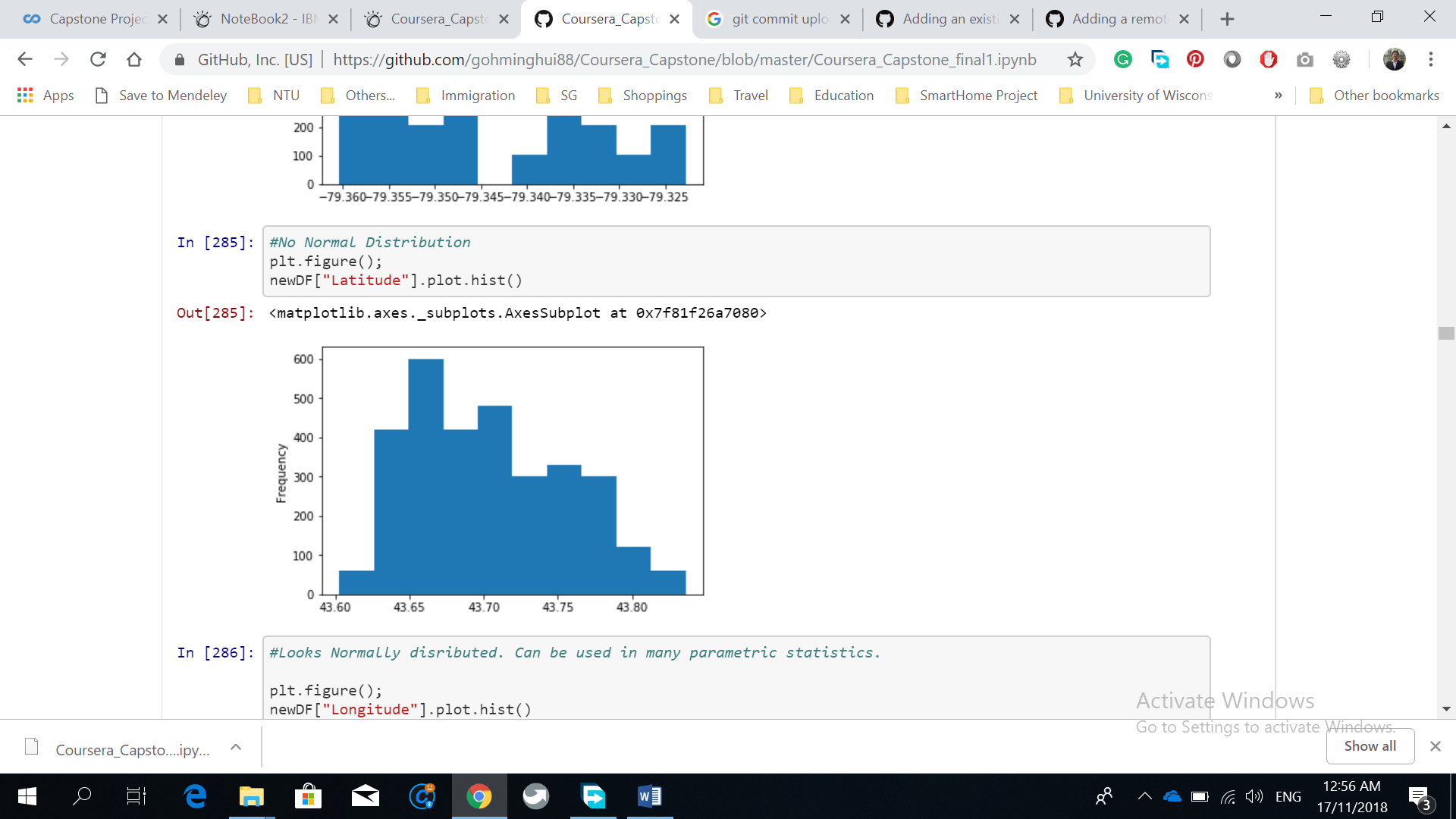
Histogram of Lat variables



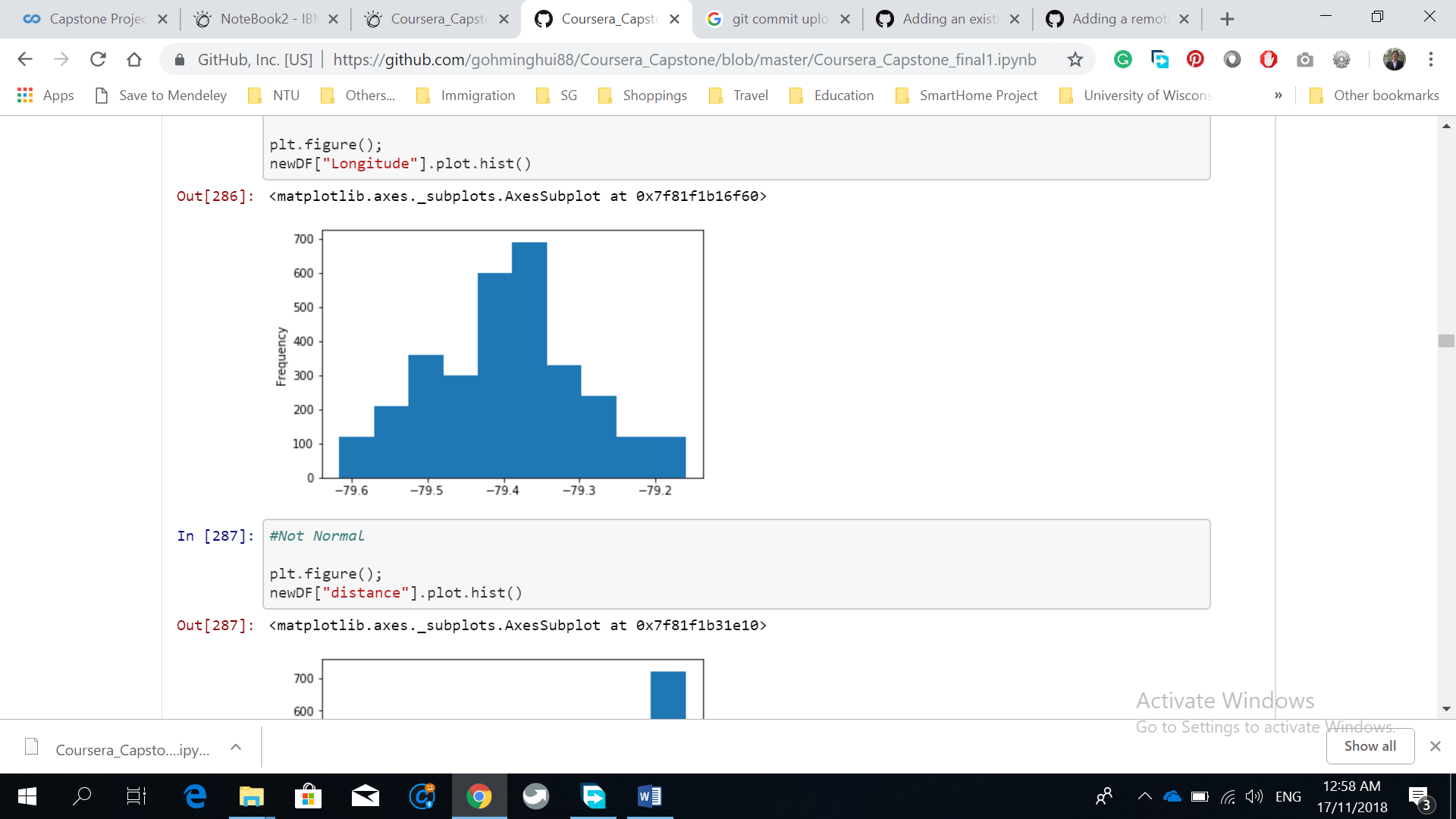
Histogram of lng variables



Histogram of Latitude variable



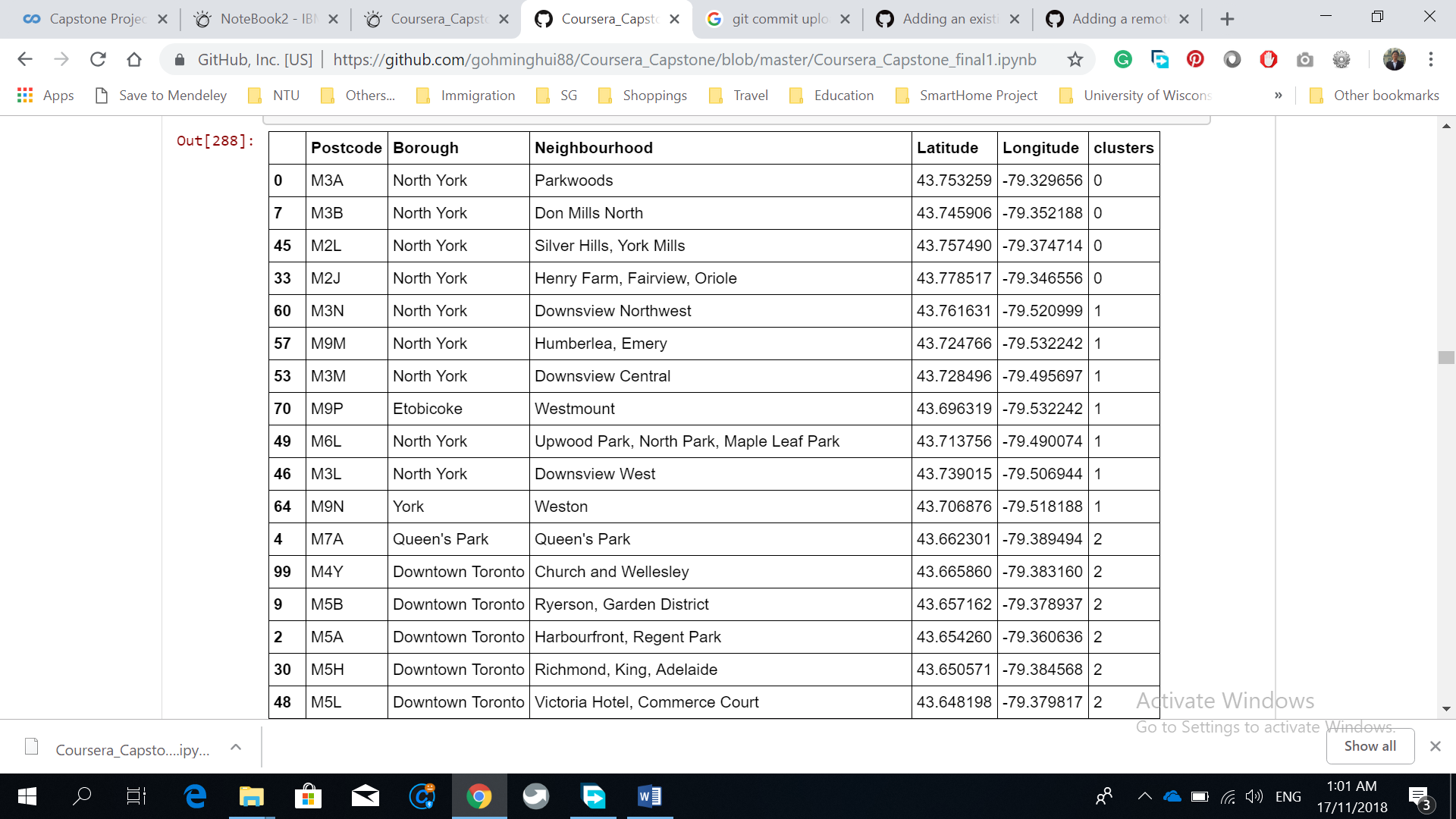
Histogram of Longitude variable



Histogram of Distance Variable

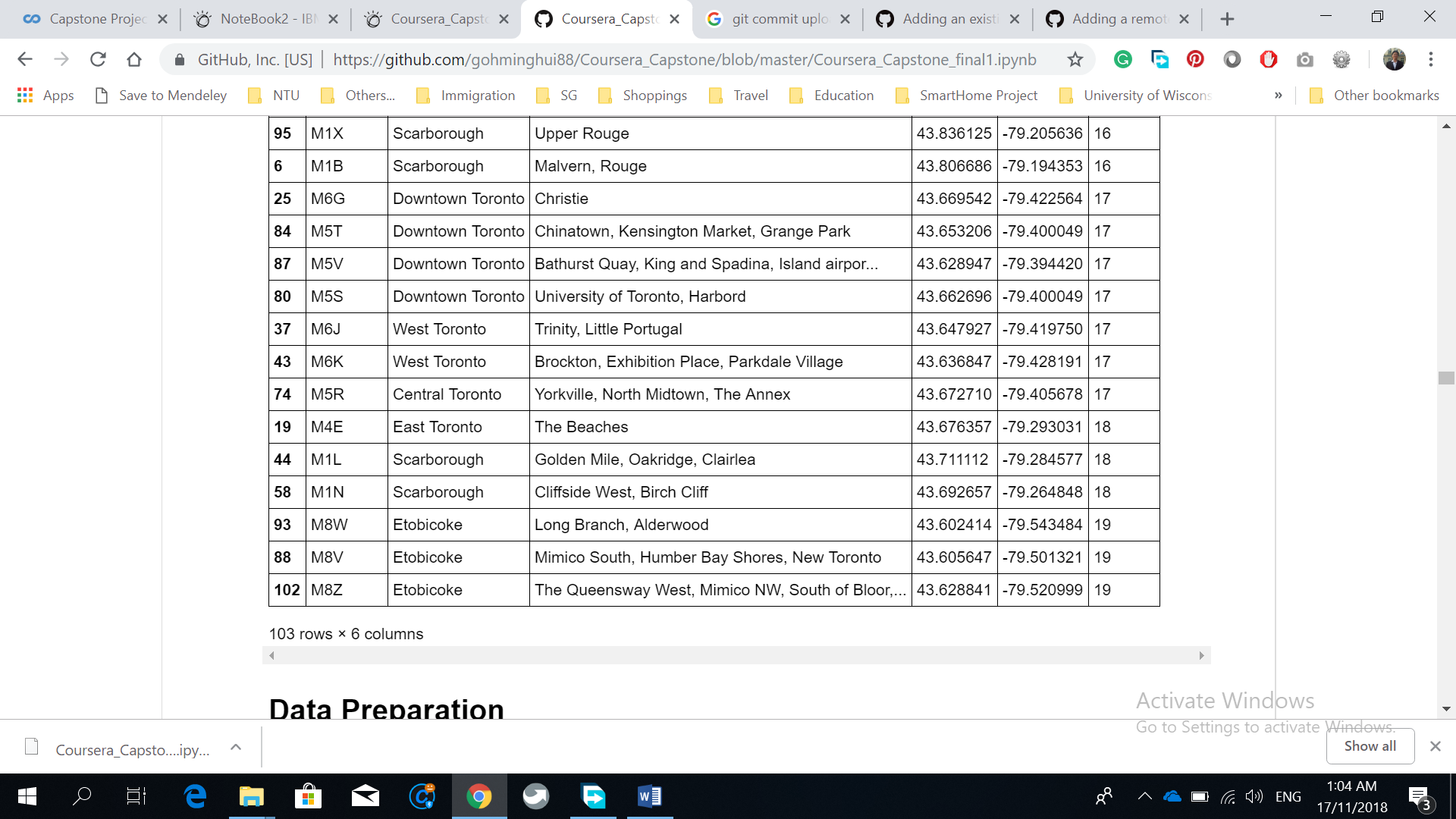


Clustering

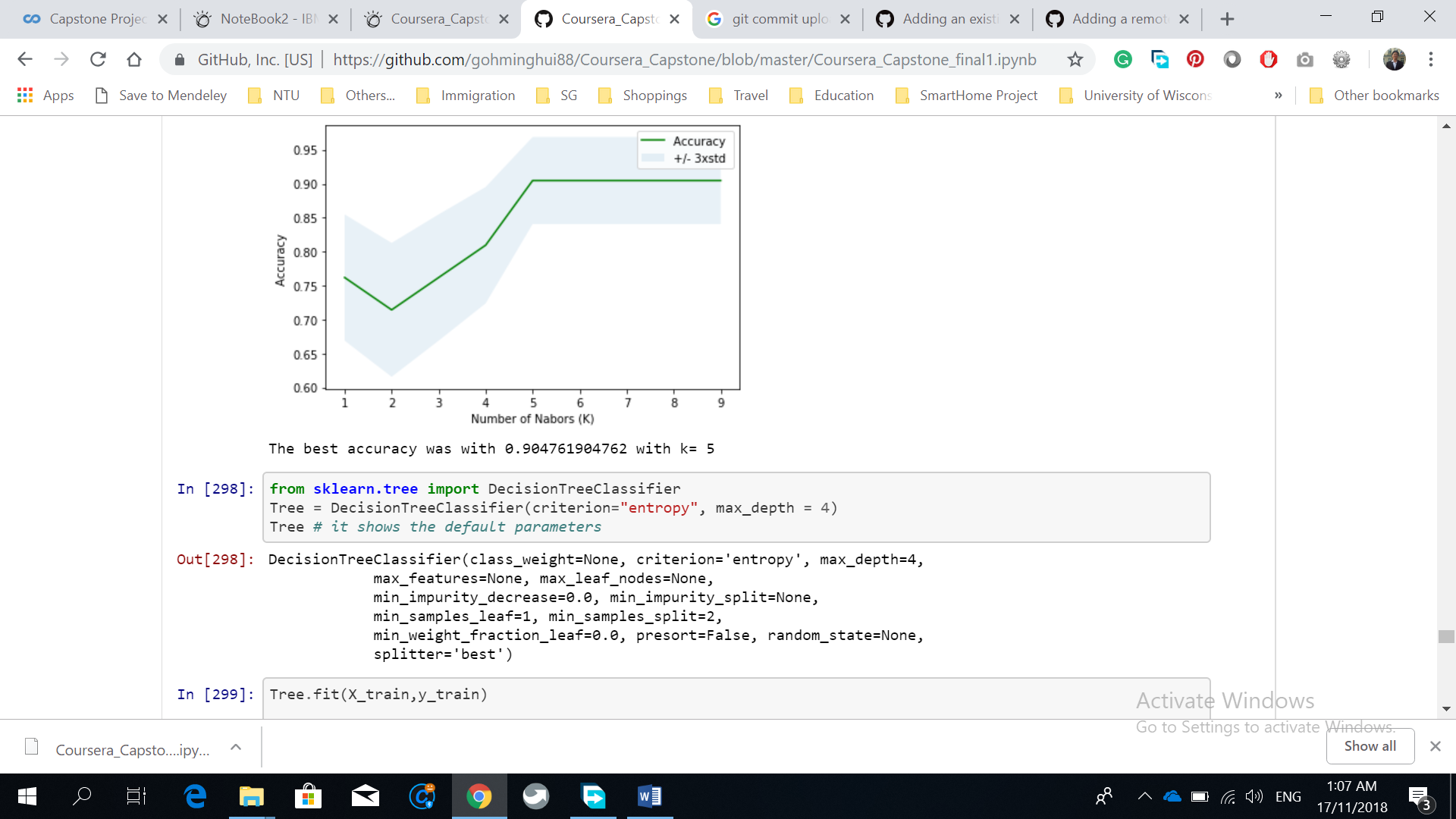




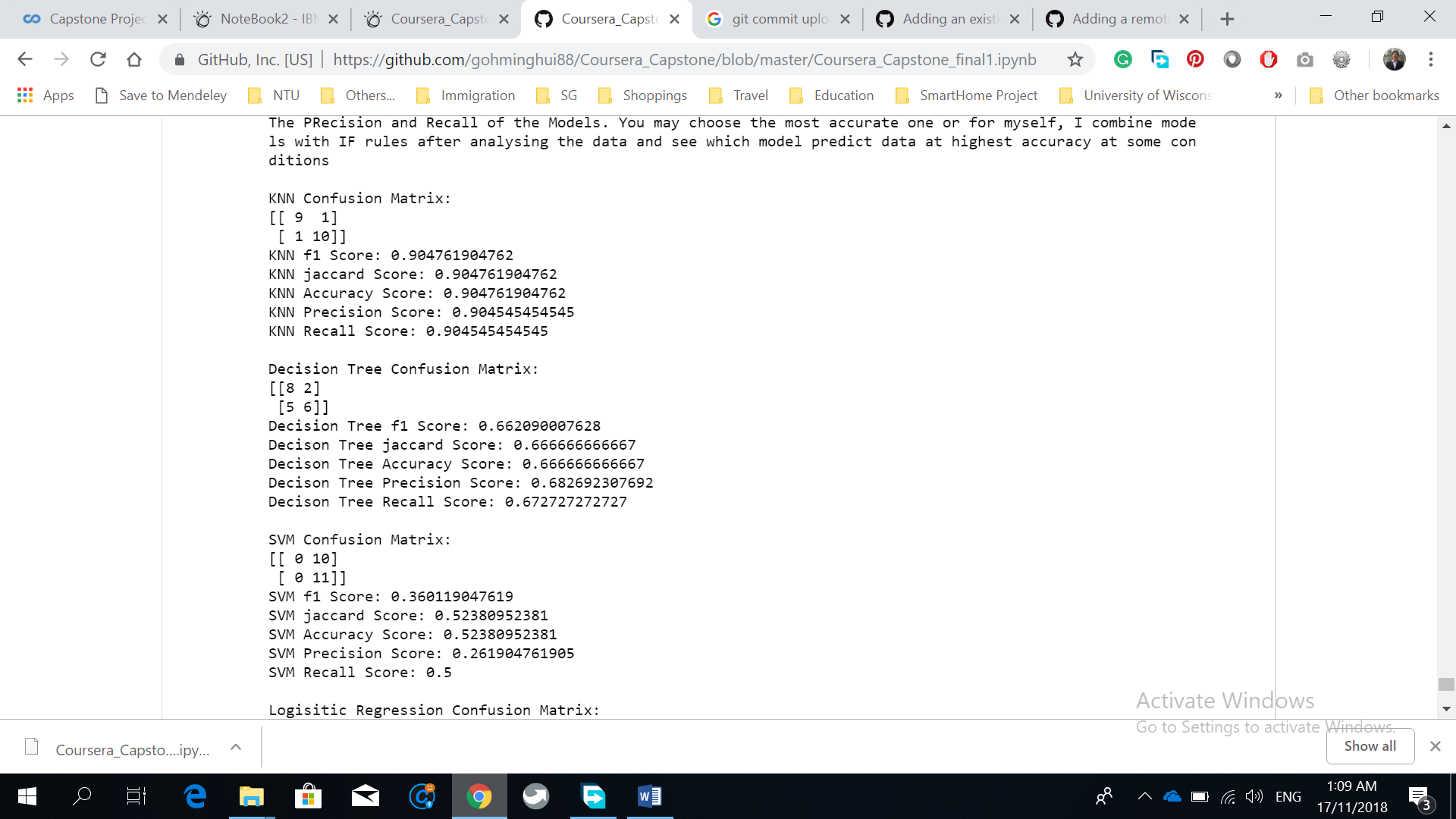


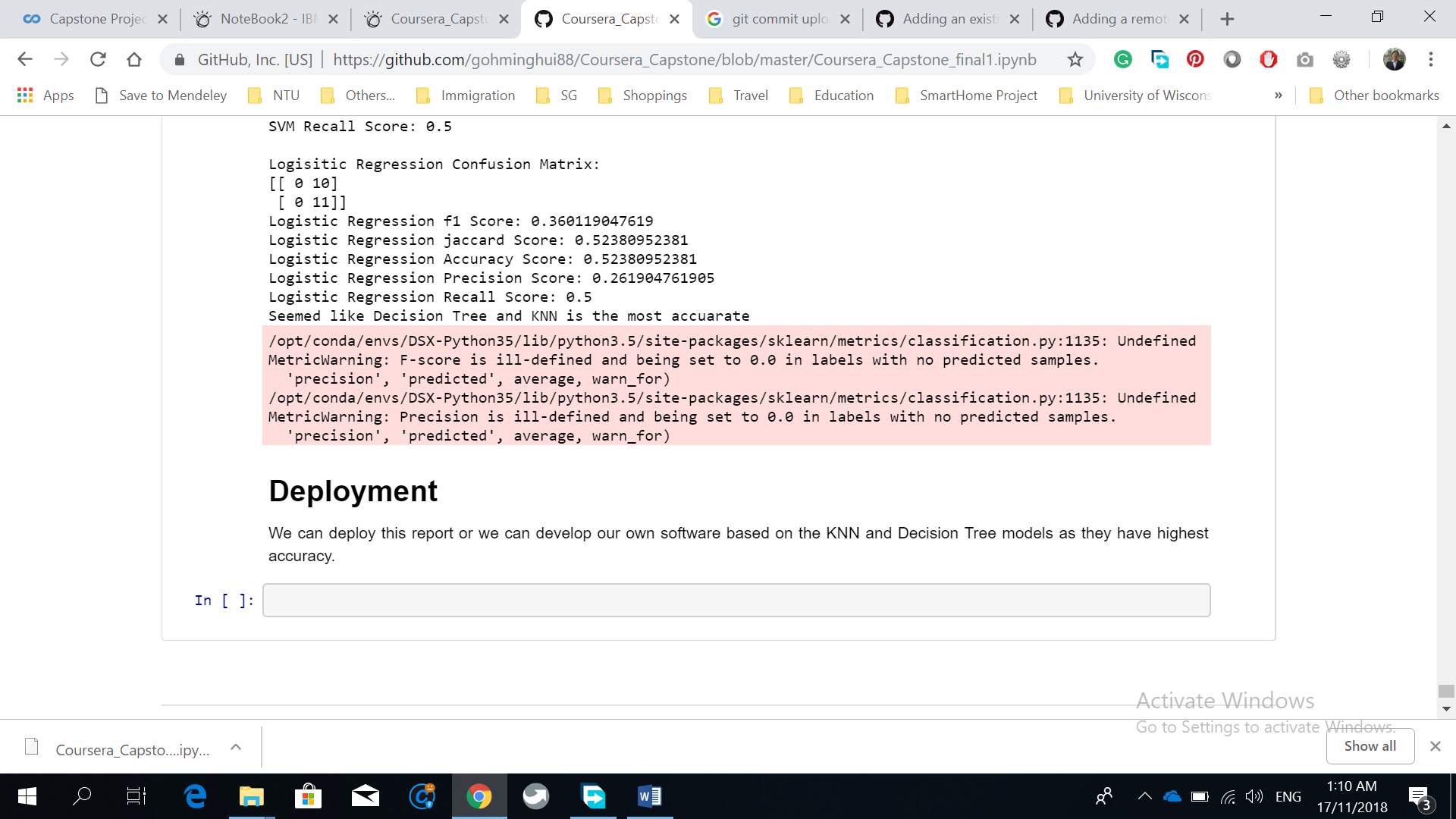


K for KNN



Model Evaluation Results



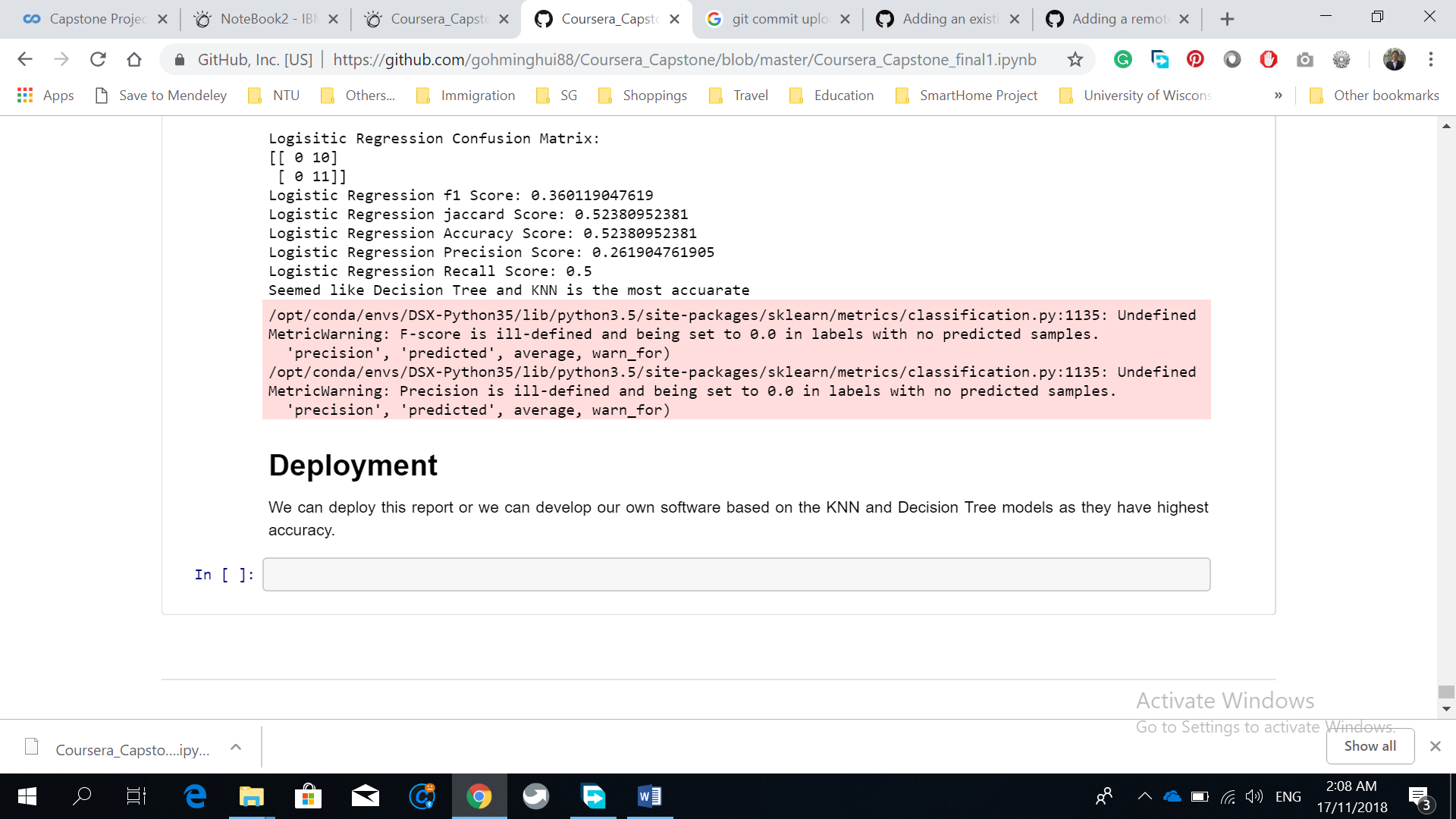


#### Discussions

In the clustering, we can see the in the 20 clusters, there are some neighbourhoods from other borough being clustered in the same cluster, hence, these neighbourhoods has higher similarity in terms of Latitude and Longitude. They may be closer together.

For variables, only the Longitude variable is more normally distributed, hence, parametric statistics and other algorithms may works well on longitude variable. Other variables are not normally distributed, hence, algorithms may not works well.

From the models evaluation results, we can see that the decision tree and KNN has higher accuracy. We can convert the models into a software to predict whether we should invest in the pub business based on the location. Based on the [recision, recall, and accuracy, we may choose the most accurate models, but for myself, I recommend combine models with IF rules after analysing the data and see which model predict data at highest accuracy at some conditions.



#### Conclusion

In this data mining project, we used CRISP DM model to create prediction and classification models. We go into business understanding, data understanding, data preparation, modeling, evaluation and deployment. The result is from data understanding and we can see that only longitude variable is normally distributed. The result also shows the model evaluation results and it seemed that the decision tree and KNN is more accurate. The result also show clustering of top 20 clusters. Clustering can be a model, but we used it at the data understanding stage. Data preparation is used to scraped the data from Wikipedia and extract from FourSquare. A second Data Preparation is used to reduce the variables (variable selection), add in the isSetUp variable, and retrieve only the rows with pub businesses. Both the data preparation is shown in the Data portion. We used the Decision Tree, KNN, Logistic Regression, and etc. to create our models.

References

Some codes is from the course. The python notebook link is at: https://github.com/gohminghui88/Coursera\_Capstone/blob/master/Coursera\_Capstone\_final1.ipynb

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.