**How to determine the best bank branch location depending on Demographics Dataset& Foursquare API Case study(Toronto Neighborhoods)**

**By : Abdelaziz Lila**

**Project –capstone course coursera**

**I-Introduction**

**I-A: Background**

Market conditions determine the overall opportunity and probability of selecting the right location which is highly important like selecting the market place. Success in determination the location grantee how well you can serve that market by taking advantage of natural consumer shopping patterns on the other hand site quality will determine how visible and accessible you are to those people. Bank branches perform well when aligned with retail centers visited at similar frequency to bank visits on that way you are convenient to your customer’s existing shopping patterns and not a stand-alone destination .This is the reason why community shopping centers with large grocery store anchor are the most favored locations. Locations lack significant retail focal point underperforms due to lack of convenience and draw impacting sales by as much as 20-30%.

Once we select a location depending on our Data Analysis where data tells us the highly performance locations then it will be plotted in a map and we will be able to describe the advantages and disadvantage for each particular location and we will be able to determine the feasibility of our expansion plan , closing old branches or transformation of branches depending on these criteria:-

1-Customers segmentations (lifestyle-behavior especially their consuming and purchase pattern)

2-population &demographics (age –gender-households)

3-financial indicator for each neighborhood (income level)

4-Location of competitors branches.

5-Easiness to reach and access.

6-Distance between branches.

7-Cost of opening a new branch .

8-Average expected transaction volume**.**

Consist Locational profiles of the existing customer base in the area under consideration by segmenting & clustering neighborhoods techniques .That is necessary in determining not only the optimum placement or modifications in branches and equipment but to determine the likelihood of financial success of branch and equipment under various scenarios .We can use foursquare as main player in determining our bank branch location feasibility study in corporation with some datasets to reach the optimal decision and help decision makers to put their convenient marketing strategies and plans .

Our journey will start with analyzing each Neighborhood Data to determine what data looks like and which variables is important to take into account when predicting our target variable - Neighborhoods distinct with highly prospected clients - then building our Machine Learning model ,feeding it with a meaningful variables & choosing the best classifier. Finally display the clusters for each nominated neighborhood on maps- by foursquare API and by analyzing clusters we will define the best locations

**I-B:Problem**:-

A multinational bank adopted expansion plan in Canada . We are a consulting company we were assigned a mission to prepare a feasibility study to determine the best locations in Toronto neighborhoods. Depending on our distinctive approach “ Mapping customers & competitors segmentation and clustering in neighborhoods by using Foursquare incorporation with needed datasets by Building a Decision making Model to predict the best locations for new branch in Toronto neighborhoods.

Board of directors decided to open new branch in Toronto neighborhoods. Upon Bank policy bank target is increasing retail savings and lending portfolios according to increase bank accounts, credit cards and consumer loans sales volume. We have to determine the best location for a retail bank branch.

**I-3:Interest**:- Banks Board of directors & Mortgage Brokers

**II Data Description:-**

**A:-Data Acquisition**

The type of data collected depend on the pre-set criteria which will be used in our model . The datasets is multivariate and attributes are real data related to Toronto neighborhoods which consist of:-

1-Dataset for Toronto borough and neighborhood locations its attitudes and longitudes.

2-Dataset for Toronto neighborhoods geospatial data .

3- Dataset for Toronto neighborhoods demographics includes numbers of populations distributed by age groups, origin, certification, employment status, apartment ownership ,average households income in each in excel format.

4- Dataset for Toronto neighborhoods average income level in excel format .

5-Dataset for Toronto neighborhoods education levels.

6-we will depend on foursquare API to retrieve venues in nominated Toronto neighborhoods . Like banks, ATMs, restaurants, tourism locations, hospitals .

**II-B:Data Source**:-

1-Wikipedia (Toronto City).

2-City of Toronto Open data portal Home.

**II-C:Data Type**:-

1-EXCEL files for Main datasets .

2-CSV.file for Locations and geospatial data.

**II-D: Read Data:-**

We use pandas library tool to read Datasets in Data frames on our Jupyter Notebook on IBM Watson .

-Main datasets in excel for simplification were merged in one excel file and unneeded data columns dropped and renamed file to be Toronto Neighborhood Dataset on my PC and then dropped on my Jupyter Notebook to read it in pandas data frame.

**II-E: Data understanding:**-

**Dataset size**: 140 rows x columns23

**Basic insight of Dataset**:-

1. **HoodID**  :- Number of Neighbourhood - int64.
2. **Neighbourhood:**- Name of each Toronto Neighbourhood-Object.
3. **Pop2016:**-number of population -int64.
4. **employed**:-number of employed labour on each Neighbourhood - int64
5. **unemployed**:- number of unemployed labour on each Neighbourhood -int64.
6. **Not in the labour force** :- number of retired or non-able to work -int64
7. **uneducated** :-number of no educated still under education or out of education system- int64
8. **college** :-number of population certified from college - int64
9. **educated** :-number of population certified and ready to join workforce- int64
10. **households** :-number of population represents households -int64
11. **Owner**:-number of population on each neighborhoods has as ownership-int64
12. **Renter:**-number of population in each neighborhood depend on mortgage rent- int64
13. **households spending**%:-spending as a percentage of average income -float64
14. **income level** :-household average income on each neighborhood- int64
15. **distribution of population from 20 years old to 64 years old which will represents**

**prospected clients**

20 to 24 years int64

25 to 29 years int64

30 to 34 years int64

35 to 39 years int64

40 to 44 years int64

50 to 54 years int64

55 to 59 years int64

55 years and over int64

60 to 64 years int64

**II-F :Data Wrangling**

1-No Missing values

2-added columns as prospected clients by aggregating all distributed populations by age

3-be sure that all data types is int64

4-Binning:-rearrange data in (Pop 2016, income level ,prospecting clients, households, educated ,employed )in three bins high , low, medium and add a binned columns to the data frame.

5-indicator variables:- we will label binned categories replacing high by 3 medium by 2 finally low by 1

**III Methodology**

**III A:Exploatry Data Analysis:**

**IIIA-1:Analyzing individuals features patterns by using visualization**

**Our target**:-To understand the type of variables & relationship between variables and how the impact of relaying on those variables will be.

**Continuous numerical Data Model**

Dependent variable:-

1. Total\_prospected\_clients

**Independent variables** :-

1. Total Population
2. Households
3. Educated
4. income level
5. employed

**Visualization :**

1-Scatterplot with fitted lines for a continuous numerical data.

2-Boxplot to visualize the distribution of data .

**The relationship between the dependent V. independent variables :**

1. **Total\_Prospected\_clients v. Total Population**:

As the total population goes up the prospected clients increased that indicates positive relationship and direct correlation between those two variables .Total population seems pretty good predictor of total prospected clients since the regression line is almost a perfect diagonal line.

1. **Total\_Prospected\_clients v. educated:**

As the total educated population goes up the prospected clients increased that indicates positive relationship and direct correlationbetween those two variables. Total educated population seems pretty good predictor of total prospected clients

since the regression line is almost a perfect diagonal line.

1. **Total\_Prospected\_clients v. households:**

As the total households number goes up the prospected clients increased that indicates positive relationship and direct correlation between those two variables. Total educated population seems pretty good predictor of total prospected clients since the regression line is almost a perfect diagonal line.

1. **Total\_Prospected\_clients v.employed**:

As the total employment goes up the prospected clients increased that indicates positive relationship and direct correlation between those two variables. employment seems pretty good predictor of total prospected clients since the regression line is almost a perfect diagonal line.

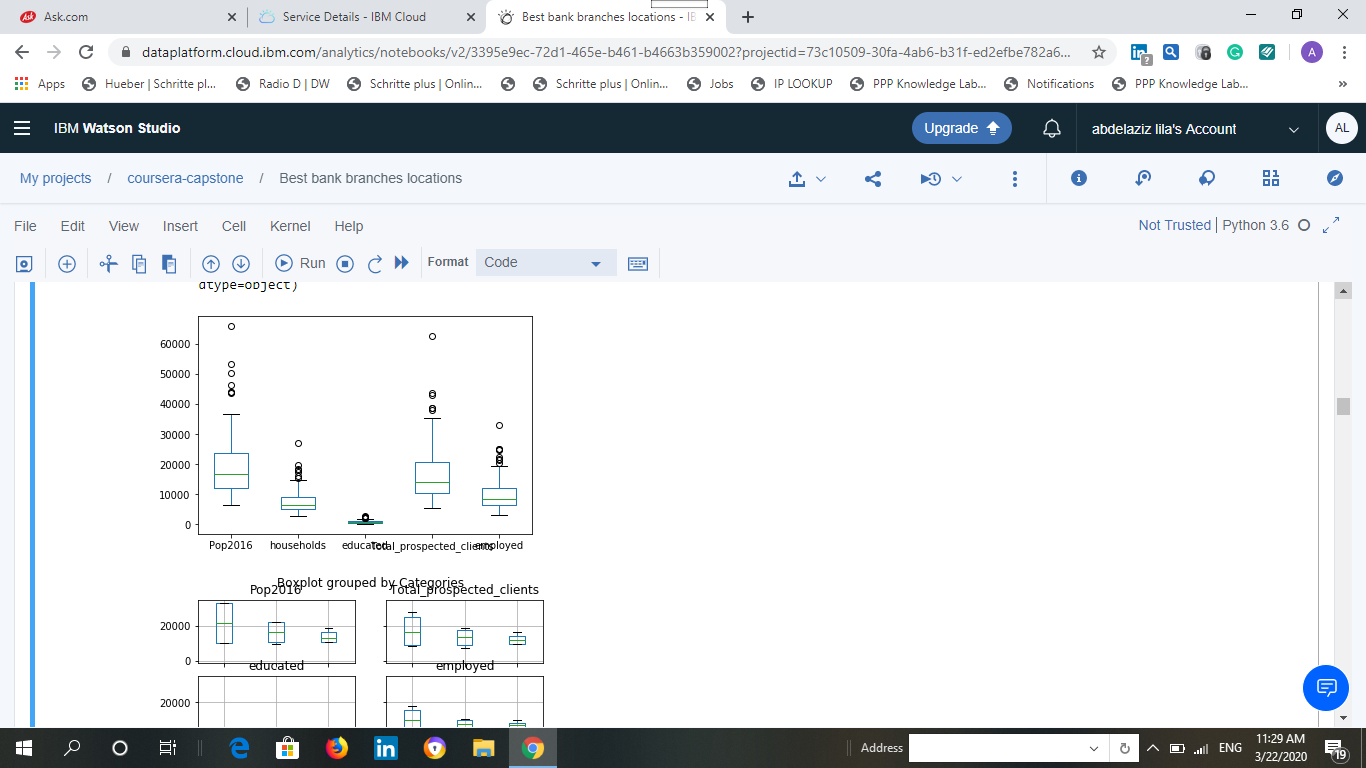
1. **Total\_Prospected\_clients v. income level :**

As the income level goes up the prospected clients decreased that indicates negative relationship and weak correlation between those two variables. Income level not seem a good predictor of total prospected clients since the regression line is almost a perfect diagonal line but the data points are very scattered and far from the fitted line showing some extend of variability therefore it is not a reliable variable .

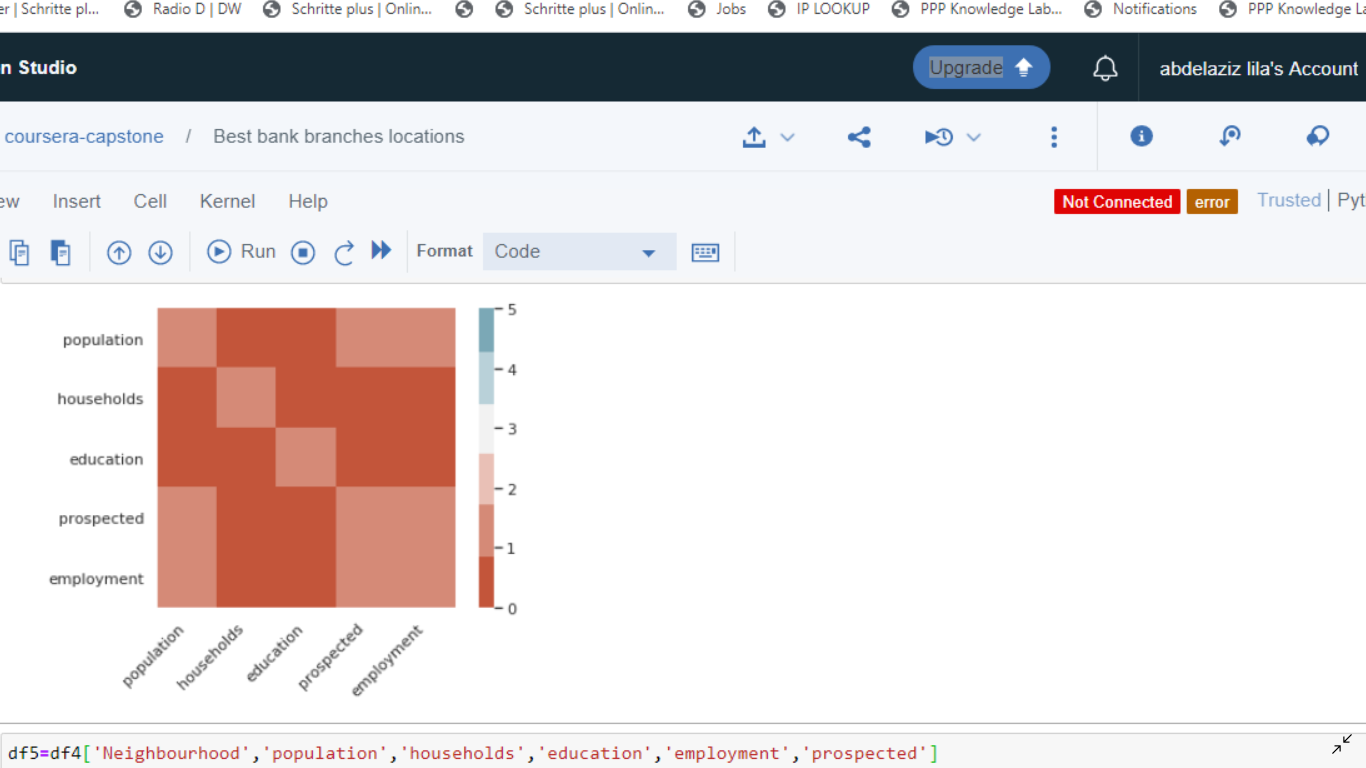
**III A:2 Descriptive Statistical Analysis**

-**First** we used describe() function :to understand the basic statistics for our continuous variables.

-**Boxplot** :- shows that data is a little bit skewed left represent asymmetric data distribution with outliers .



**Heat Map** :- For clarity As all variables positively and strongly correlated our map was built centered around 3 which means it starts from zero correlation and the darkness starts from respectively 0 to less than one in two degrees of colors the first dark represent lesser correlation between variables and the lighter one represents correlation close to 1or 1



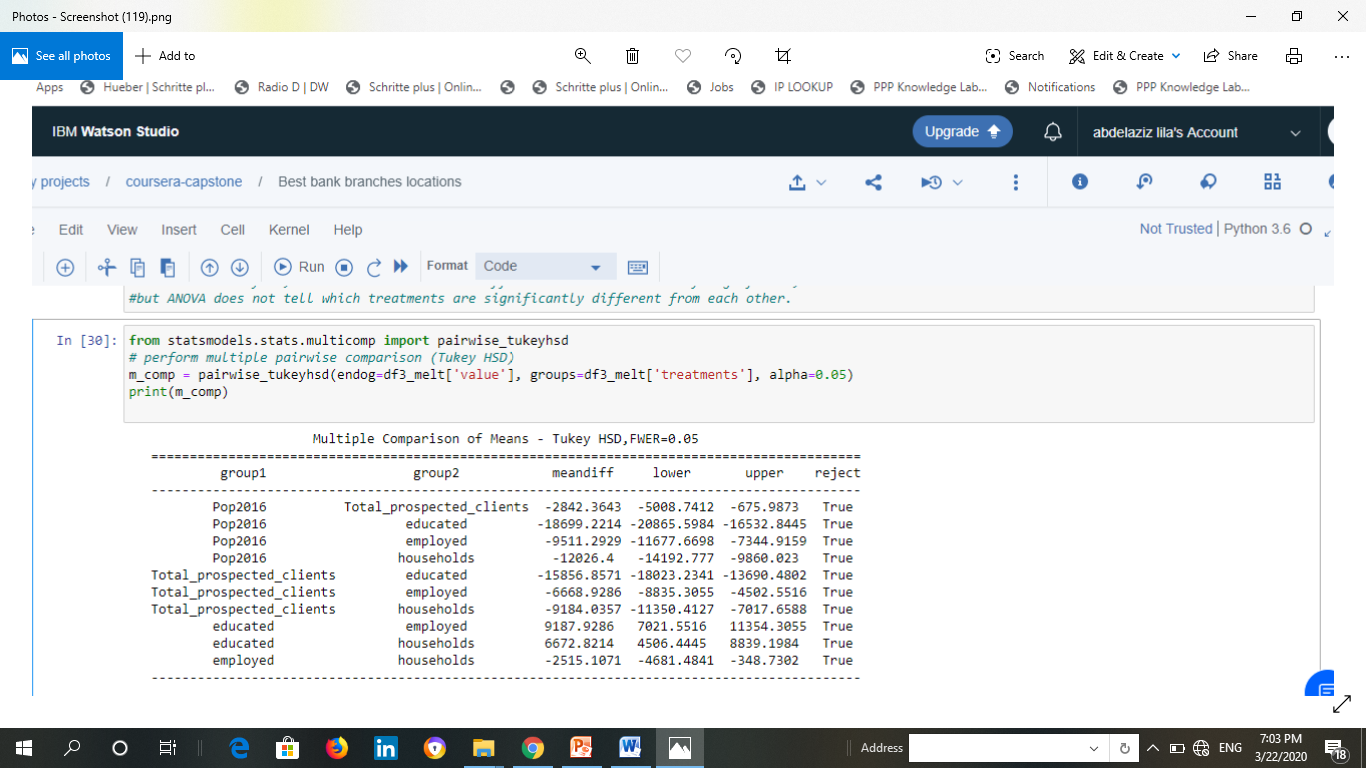
-.

**ANOVA analysis**

**Interpretation**: The P-value obtained from ANOVA analysis for Data, and interaction are statistically significant (P<0.05). We conclude that all independent variables affects the dependent variable significantly, and interaction between independent variables significantly affects the yield outcome.

Now, we know from ANOVA analysis, that treatment differences are statistically significant, but ANOVA does not tell which treatments are significantly different from each other. To know the pairs of significant different we have to perform multiple pairwise comparison (Post-hoc comparison) analysis using Tukey HSD test.

**Tukey HSD Test result:-**

****

Above results from Tukey HSD suggests that there is a significant difference between the two means , all pairwise comparisons for treatments rejects null hypothesis and indicates statistical significant differences which support the preceding result from ANOVA analysis **.**

**Test ANOVA assumptions**

Similar to one-way ANOVA, we can use **Levene** and **Shapiro-Wilk test** to validate the assumptions for homogeneity of variances and normal distribution of residuals

Can be used to check the normal distribution of residuals null hypothesis data is drawn from normal distribution. As P –value less than 0.05 imply significant& the [*p* value](https://en.wikipedia.org/wiki/P-value) is less than the chosen [alpha level](https://en.wikipedia.org/wiki/Alpha_level), then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. So we succeeded to reject null hypothesis and that support the ANOVA results.

**Bartlett’s Test**

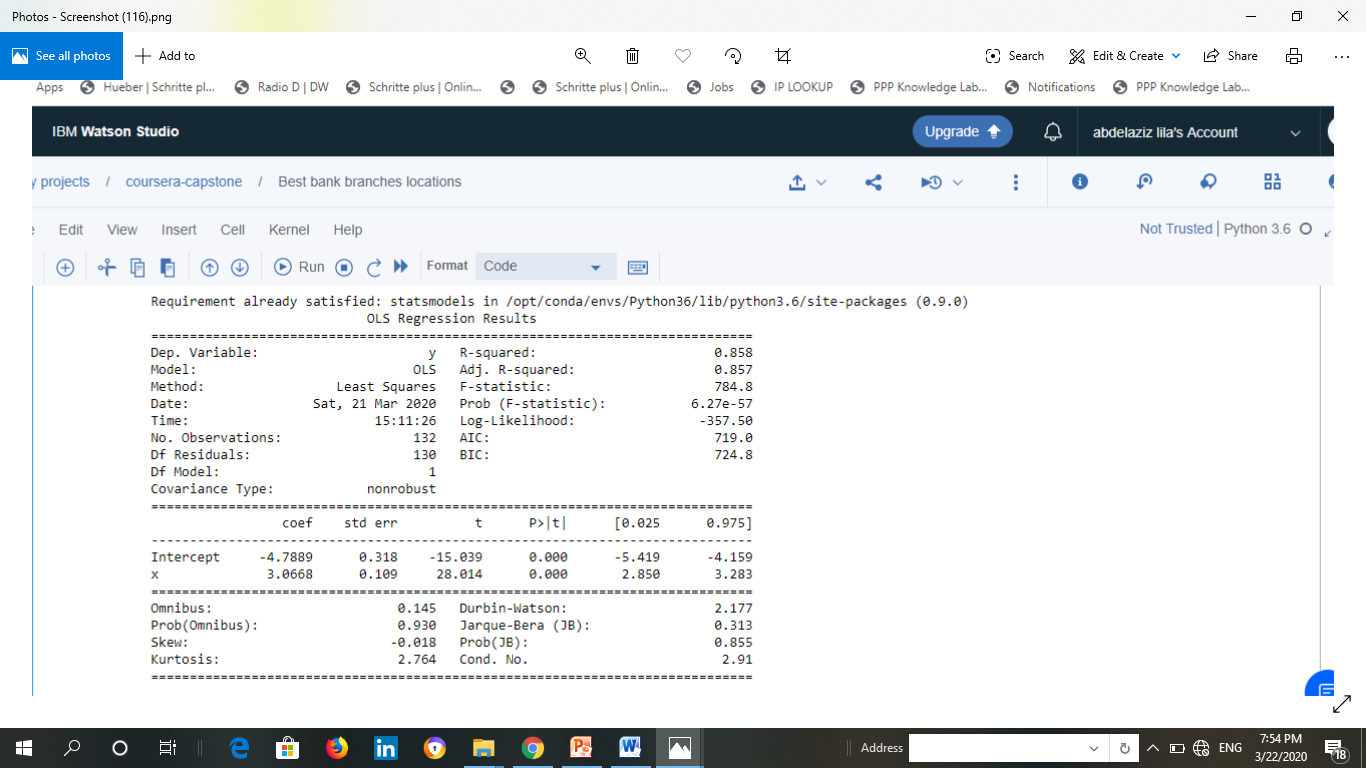
As null hypothesis assumption the data is drawn from normal distribution, use Bartlett’s test to check the Homogeneity of variances.as the p-value still less than 0.05 we reject the null hypothesis and samples in the population have different variance.

**Validity conclusion**

Our model assumption is valid and there is a strong relation among variables and our model is reliable to predict the prospected clients depending on the demographics variables

Now let’s go through the regression analysis

**Ordinary least square Linear Regression**

****

The section we are interested in is at the bottom. The summary provides several measures to give you an idea of the data distribution and behavior. From here we can see if the data has the correct characteristics to give us confidence in the resulting model. We aren't testing the data, we are just looking at the model's interpretation of the data. If the data is good for modeling, then our residuals will have certain characteristics. These characteristics are: The data is "linear", Errors are normally distributed across the data, There is "homoscedasticity". This means that the variance of the errors is consistent across the entire dataset. The results of the linear regression model run above are listed at the bottom of the output and specifically address those characteristics as follow:-

1. **R –squares**: Ordinary Least Squares (OLS) Linear Regression analysis is the best fit for Data modeling shows R2=0.858 imply good accuracy to predict values and all independent variables have impact in the population represents 86% we could add another variables
2. **Prob (F-statistic)**per the above results, probability is close to zero. This implies that overall the regression is meaningful.
3. **P –values** is close to zero we could reject the null hypotheses and all variables are reliable
4. **Omnibus/Prob(Omnibus)**– a test of the skewness and kurtosis of the residual. We hope to see a value close to zero which would indicate normalcy. The Prob (Omnibus) performs a statistical test indicating the probability that the residuals are normally distributed. We hope to see something close to 1 here. In this case Omnibus is low and the Prob (Omnibus) is high so the data is normal,. A linear regression approach would probably be better than random guessing but likely is better than nonlinear approach.
5. **Skew**– a measure of data symmetry. We want to see something close to zero, indicating the residual distribution is normal. Note that this value also drives the Omnibus. This result has a small, and therefore good, skew.
6. **Kurtosis**– a measure of "peakiness", or curvature of the data. Higher peaks lead to greater Kurtosis. Greater Kurtosis can be interpreted as a tighter clustering of residuals around zero, implying a better model with few outliers.
7. **Durbin-Watson**– tests for homoscedasticity .We hope to have a value between 1 and 2. In this case, the data is close, but within limits.
8. **Jarque-Bera (JB)/Prob(JB)**– like the Omnibus test in that it tests both skew and kurtosis. We hope to see in this test a confirmation of the Omnibus test. In this case we do.
9. **Condition Number**– This test measures the sensitivity of a function's output as compared to its input. When we have multicollinearity, we can expect much higher fluctuations to small changes in the data, hence, we hope to see a relatively small number, something below 30. In this case we are well below 30.

In looking at the data we see an "OK" set of characteristics. This would indicate that **the OLS approach has validity,**

**This is the end of Data analysis section**

**IIII Classification Models by Machine Learning –the best Classifier**

**N.A:**

1. **Data binned and transferred to categorical Data and scaled as 1 for low ,2 for medium &3 for high numbers**
2. **Data split to train & test sets by ratio 30% and fit by scalar**
3. **K set to 9 in KNN model the best k was 3 at high accuracy**
4. **F1- TEST for KNN model ill-defined and set to 0 IN some training means that test Precision affected.**

**Evaluation of each classification models result**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Model**  **Evaluation tool** | **KNN K=3** | **SVM** | **Decision tree** | **Logistic regression** |
| **Jaccard-similarity -score** | **0.952** | **0.976** | **0.976** | **0.905** |
| **F1 test** | **0.976** | **0.975** | **0.975** | **0.830** |
| **Log Loss** | **NAN** | **NAN** | **NAN** | **0.858** |

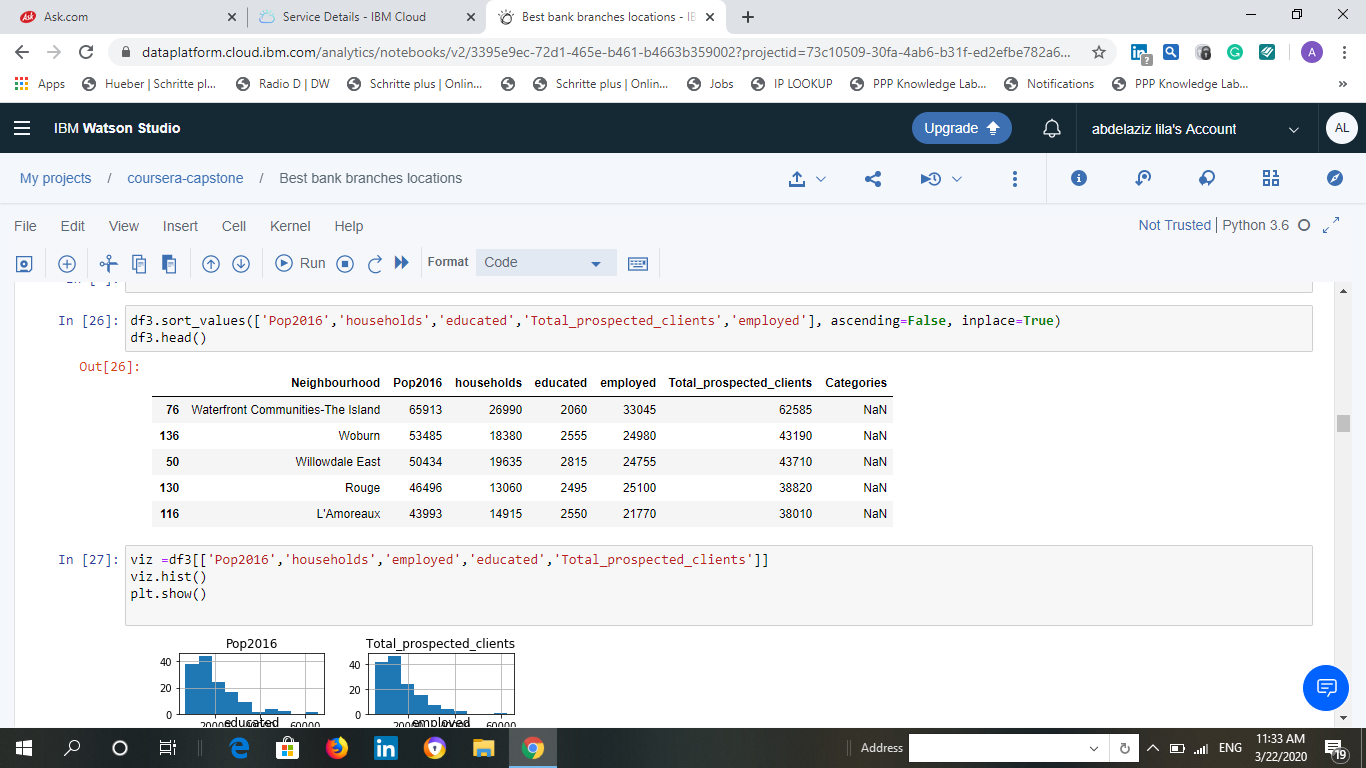
**Despite the best results for evaluation test in the first three models we will chose the logistic regression model as our classifier for the following reasons:-**

1. It is the best predictor model used for any customer churn prediction That is our case for Toronto prediction of prospected clients on each neighborhood depending on demographics Data to determine the best location for our branch
2. The Data is categorical data and it is the best fit model deals with this type of data
3. Logistic regression provides lots of ways to regularize your model, and you don’t have to worry as much about your features being correlated
4. We also have a nice probabilistic interpretation, and we can easily update model to take in new data, unlike decision trees or SVMs.
5. We can use it if we want a probabilistic framework or especially when we expect to receive more training data in the future that you want to be able to quickly incorporate into model.
6. Logistic regression can also help you understand the contributing factors behind the prediction, and is not just a black box method.

V:Conclusion

Upon Data analysis& Classifier Model what is the best nominated neighborhoods?

As our model is valid and there is a strong correlation between dependent variable which is prospected clients and our reliable independent variables educated population , number of home owner households ,number of population & number of employed population ,as they increase the number of clients will increases we assess and prove the validity of our assumption and we could depend on it to predict or clients market share in neighborhood



.By sorting data ascending we choose the best two neighborhoods result

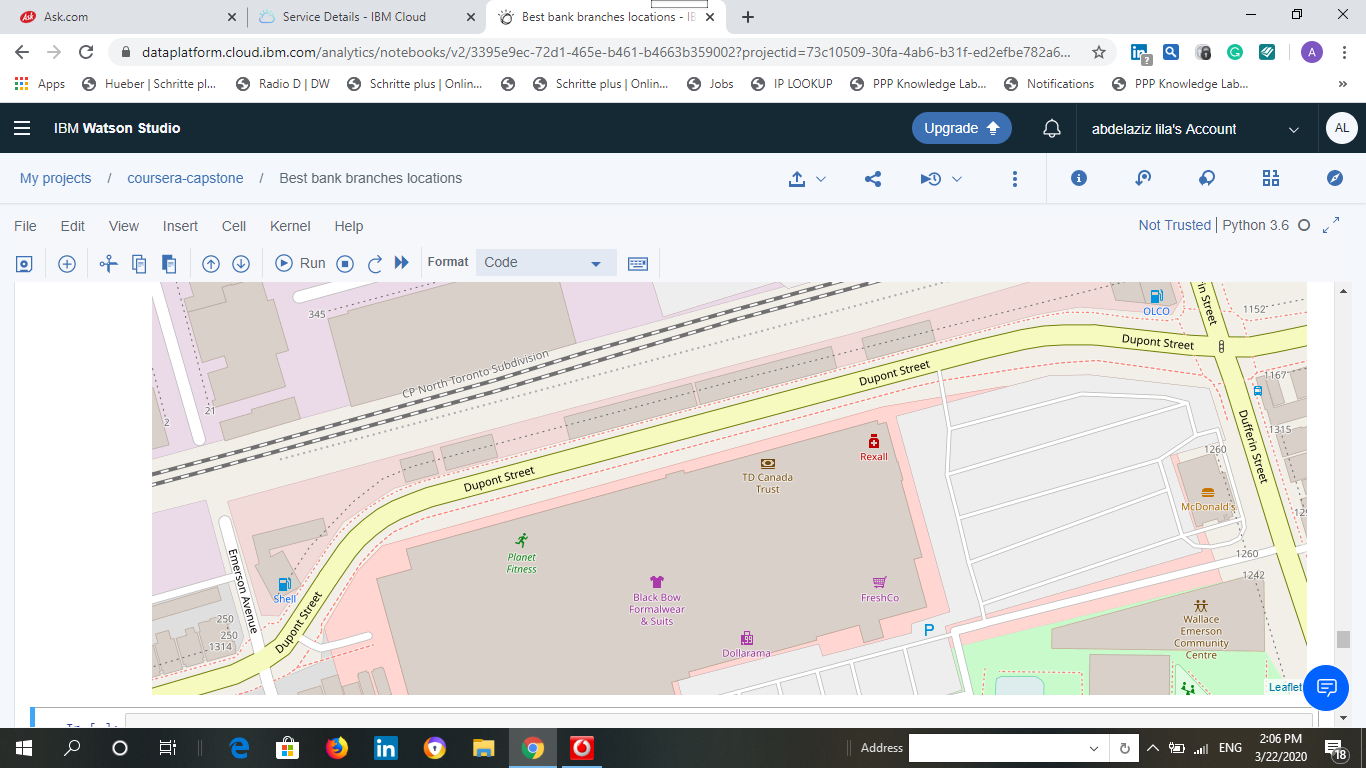
1-**Waterfront Communities – The Island** : Expected clients in the market 62585 &The highest scored numbers in all independent variables .

2-**Willowadle East :**Expected clients in the market 43190 &the almost second highest score in all independent variables .

Now the second part performing Location analysis by Foursquare API and proof the validity of our final decision .

**VI :Foursquare API location analysis** :

VI-A:-waterfront communities-the island



**Advantage:**

1-highly venues concentration distinct with purchasing patterns and retailers which is the suitable location for retail branch.

2-Location is on main street easiness to enter branch granted .

3-Fitness club& McDonalds branch exist near our branch location which attract most of youth in the arena especially millennium generation our clients’ cornerstone .

4-shopping & restaurant area grant attracting more clients .

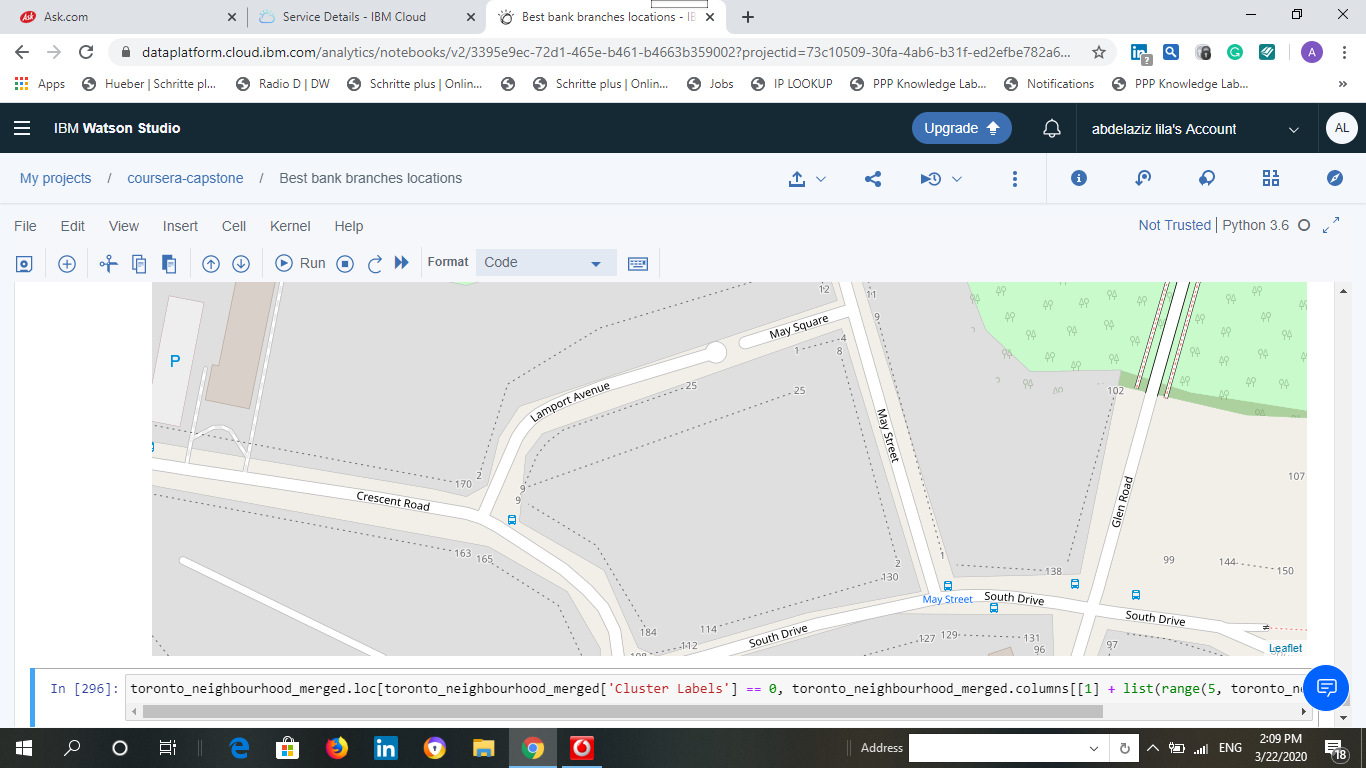
5-Parking area is nearest the branch location.

1. It is so easy to reach the branch location by bus .

**Disadvantage:**

There is already a competitive bank branch but the market could afford another share as our preceding analysis approve still not saturated.

**VI-B:Willowdale East**

****

**Advantage:**-

1. Best concentration for entertainment areas as it near stadium

And parks best location for attracting youth customers

2-location is a square and a bus station is closest more easy to reach .

3-There is a nearest parking area

4-main street which grant easiness of entrance

5-no competitive bank branch in the area

**Disadvantage**

It lake the retailers concentration it is not a commercial area but we might benefit from the population concentration .

**VII: Results and final Decision** :

As per as our expletory data analysis which prove the validity of our assumption , variables, model & logistic regression as our classifier and after Foursquare API location analysis and as our main target to choose the best location for a retail branch in a growing society distinct with highly education , population, employment& households and we could predict the volume of our branch transaction and set the suitable service mix depending on prediction which will support marketing plans we recommend Waterfront Communities –The Island as the best location for our new branch in Toronto neighborhoods especially if it will be retail automated branch .

**Viii: Discussion and future recommendation :**

* As we observed from regression analysis result r2 =0.858 denoted that other variables could be added one of important variable in investment decision is the cost of investment the price of land in location but as a result of insufficient trusted Data set I excluded it though its importance so for any future researcher could add this variable in the model
* As I decided to exclude a highly important variable (income level) because of its weak correlation which will affect the model results We have to consider that its an average income so if the population increases this average decreases and it is inversely related with the other variables in contrast of common reality as educated employed people increase the income increase this dilemma need further study the best indicator instead is the volume of wealth in each neighborhood .

**Viiii: Conclusion**

Demographics is a core dominant Factor in determining best bank location depending on the prediction of the expected volume of transactions especially retail banking. Foursquare API is one of the best Locator to support analysis still this findings could be improved by adding another variables such as the cost of investment depending on land prices Trustworthy Dataset