**The Battle of Neighborhoods**

Kirandeep Kaur

August 4, 2020

**1. Introduction**

**1.1 Background discussion**

Due to the COVID-19 pandemic , students are forced for studying at home but after this whole situation ends, a student is going to continue his/her studies at a College in Southampton. So this problem belongs to all those students who want to go abroad(Southampton) for higher studies . He/She is interested in buying the cheapest house in Southampton, UK where he/she gets all kinds of restaurants and fast-food chains in vicinity which are the cheapest in town because he/she is bad at cooking and won't get much time to cook due to higher studies and part-time jobs and plans to never cook in your life. In our scenario, we will be exploring the cheapest houses in Washington, DC near the cheapest and many fast-food chains. We will recommend the perfect place with all the desired and required facilities.

**1.2 Business discussion**

We are low on cash due to admission fee payments and we are not selected for any jobs yet. As a result, the business problem is where can we buy a suitable house in Southampton, UK when we are having a housing boom which drives prices ever higher and when you don't have time to cook and you don't like cooking. As a data scientist/house broker we will do a cluster of Southampton neighborhoods to get the venues with the best prices and we will be looking for cheap nearby fast-food chains and elementary schools for your children.Here the use of Foursquare - an Api which provides most updated data regarding the places to visit , food junctions , daily needs availability will help providing a detailed view of the luxuries and facilities preent in Southampton. The use of Foursquare will simplify the huge task of viewing and extracting useful information regarding the best and cheapest facilities available.

**1.3 Interest**

Obviously, students would be vey much interested in a reccomendation for a house with price less than 120k where moe food junstions,pubs and daily need stores are available for financial benefits.

**2. Data acquisition and cleaning**

**2.1 Data sources**

For my analysis I will be using "HM Land Registry" of UK for 2020 where you can find records on the Price Paid Data and you can download it as .csv or .txt file. I will be using .csv file . The following fields comprise the address data included in Price Paid Data:  
1.Postcode  
2.PAON : Primary Addressable Object Name ( Typically the house number or name)  
3.SAON : Secondary Addressable Object Name ( If there is a sub-building, for example, the building is divided into flats, there will be a SAON.)  
4.Street  
5.Locality  
6.Town   
7.District  
8.County  
I chose this data set as it is published by UK government and therefore it has better probability of correctness as compared to other datasets. It has all the attributes regarding locality , Postcode that simplifies the search . I will also be accessing Foursquare Api for detailed facilities available at the various streets to provide a clear comparison. It will help in getting data for the fast food chains around the future property and when combining it with the price paid data per house to get the best property for our needs. This is the link to the site : <https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads> This is the link to the price data for the whole 2018 : <http://prod.publicdata.landregistry.gov.uk.s3-website-eu-west-1.amazonaws.com/pp-monthly-update-new-version.csv> . In the pricedata provided by above link the Postcode will be used to get Longitude and latitudes of the streets in Southampton. Then using Foursquare we will check prices of varius Houses and explore restaurants available along with their categories.

**2.2 Data cleaning**

The dataset downloaded had several problems due to which it was cleaned following certain steps as per the customer requirements:

First, the column names were quite difficult to understand. From the website which provided this dataset, an info about the attributes was provided. So the attribute names were changed from :

|  | {A96E4 ACB-D1DA-9205E053-6C04A8C0DA09} | 240000 | 2004-06-22 00:00 | S | N | F | 3 | Unnamed: 8 | CAMBRIDGE COURT | WRINGTON | BRISTOL | NORTH SOMERSET | NORTH SOMERSET.1 | | | A | A.1 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |

To :

| ID | Price | | TransferDate | | | Postcode | TypeProp | New-Old | frhld\_lease | PrimAdrs | ScndAdrs | Street | Locality | Town\_City | District | County | stnrd\_addtnl | | site |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  |  |  |  |  |  |  |  | | | |

Second, since we have to buy house, we are only interested in recent rates. The older prices cannot be used to predict new houses as the prices change from year to year. So we will keep those rows which have been transfered after 2018. To do so we need to access Tansfer Date column. But we see that its datatype is not in accessible format so we will also change its datatype . Finally we will sort the the rows in descending order to see the ones with highest prices and ignore them. Earlier we had 77588 rows and 16 columns but after this step we will have 76397 and 16 columns.

Now our data set is ready to be analyzed. We have almost removed all the flaws which could cause hinderance in data analysis.

**2.3 Feature Selection**

Third , since we require a house in Southampton , we donot require rest of the cities . So we will now filter out rows associated with Southampton only. Then we will evaluate Average price for each street to select outthe ones with lowest prices.

|  |  |  |
| --- | --- | --- |
| **NO.** | **Street** | **Avg\_Price** |
| 10  29  77  149  157  163  172  175  178  194  206  226  251  262  294  296  318  328  361  363  373  383  388  434  436  439 | ANSON DRIVE  BERNARD STREET  CHERWELL CRESCENT  GOLDEN GROVE  GROVE GARDENS  HAMPTON LANE  HERRICK CLOSE  HIGH STREET  HILL LANE  HOWARD ROAD  KENSINGTON FIELDS  LOWER CANAL WALK  MILLBROOK ROAD EAST  NEWTON ROAD  PARK ROAD  PAXTON CLOSE  PYLEWELL ROAD  RIVER VIEW ROAD  SOUTH EAST ROAD  SOUTH STREET  ST MONICA ROAD  TATE MEWS TATE ROAD  THAMES CLOSE  WHITWORTH CRESCENT  WIMPSON LANE  WINDERMEREAVENUE | 97500.0  40000.0  88000.0  90000.0  105000.0  110000.0  112000.0  82000.0  110000.0  89000.0  110000.0  115000.0  119500.0  77000.0  90000.0  10000.0  60000.0  80000.0  97000.0  118000.0  2100.0  120000.0  80000.0  115000.0  85500.0  70000.0 |

**3.METHODOLOGY**

**3.1Exploratory Data Analysis**

**3.1.1 Evaluating the top 5 cheapest streets**

Since the specified budget includes prices less than 120k, so we will retrieve the corresponding rows only. The following table describes the cheapest 5 streets.

| **NO.** | **STREET** | **AVG\_PRICE** |  |  |
| --- | --- | --- | --- | --- |
| **206** | KENSINGTON FIELDS | 110000.0 |  |  |  |
| **226** | LOWER CANAL WALK | 115000.0 |  |  |  |
| **251** | MILLBROOK ROAD EAST | 119500.0 |  |  |  |
| **318** | PYLEWELL ROAD | 60000.0 |  |  |  |
| **436** | WIMPSON LANE | 85500.0 |  |  |  |

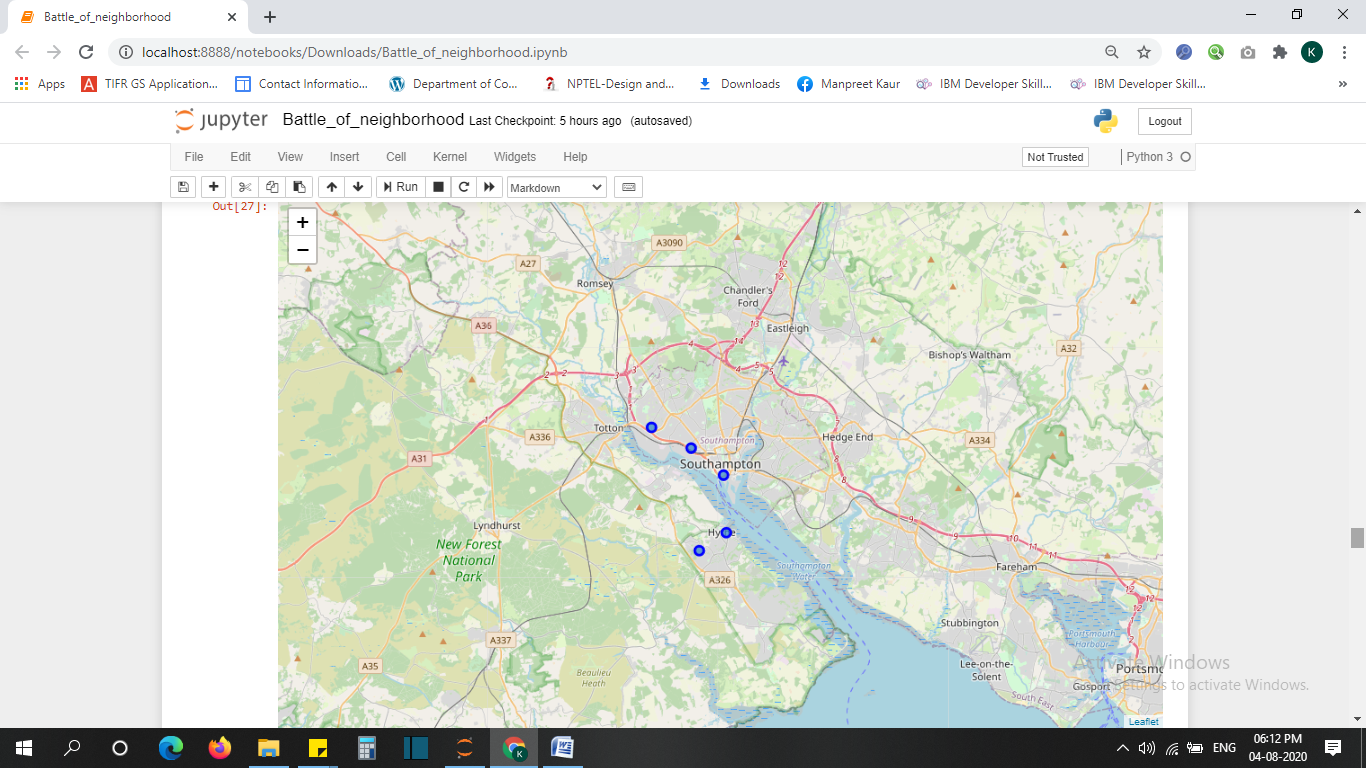
**3.1.2 Evaluating geolocations**

Now we will calculate the geolocations ( latitude , longitude ) of the retrieved streets . The latitude and longitude values of the streets will help us to see where exactly on the Southampton map does these streets lie and we can easily estimate the distance ofthese streets from each other and other mportant venues nearby that we would like to explore. The following table shows the geolocations of the streets retrieved above.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NO.** | **Street** | **Avg\_Price** | **city\_coord** | **Latitude** | **Longitude** |
| 206 | KENSINGTON FIELDS | 110000.0 | (50.8604243, 1.4207463) | 50.860424 | -1.420746 |
| 226 | LOWER CANAL WALK | 115000.0 | (50.897111, -1.4022897) | 50.897111 | -1.402290 |
| 251 | MILLBROOK ROAD EAST | 119500.0 | (50.9098907, 1.4266578) | 50.909891 | -1.426658 |
| 318 | PYLEWELL ROAD | 60000.0 | (50.8694403, 1.4000328) | 50.869440 | -1.400033 |
| 436 | WIMPSON LANE | 85500.0 | (50.9201304, 1.4571668) | 50.920130 | -1.457167 |

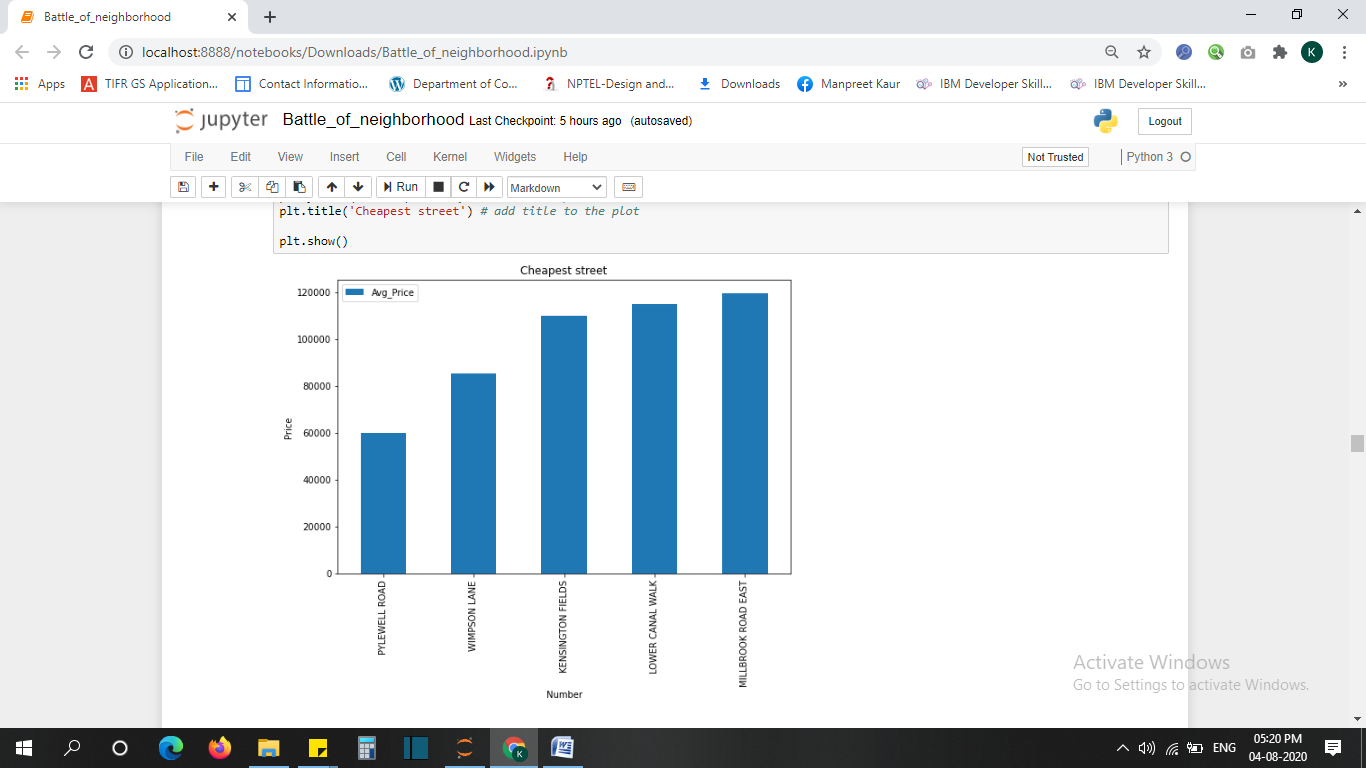
**3.1.3PLOTTING THE STREETS**

Lets visualize the above streets on Sothampton map using the geolocations generated above .



**3.1.4 Visualizing the cheapest street**

We will use a bar graph to see the street which is cheapest and the street which is the most costly . Even if the customer doesnot agree to choose the cheapest street in case the facilities he/she requires are missing then this graph visualization to choose the street which is the one next to the cheapest. Thus bar graph provides a clear visualization for the cheapest street.



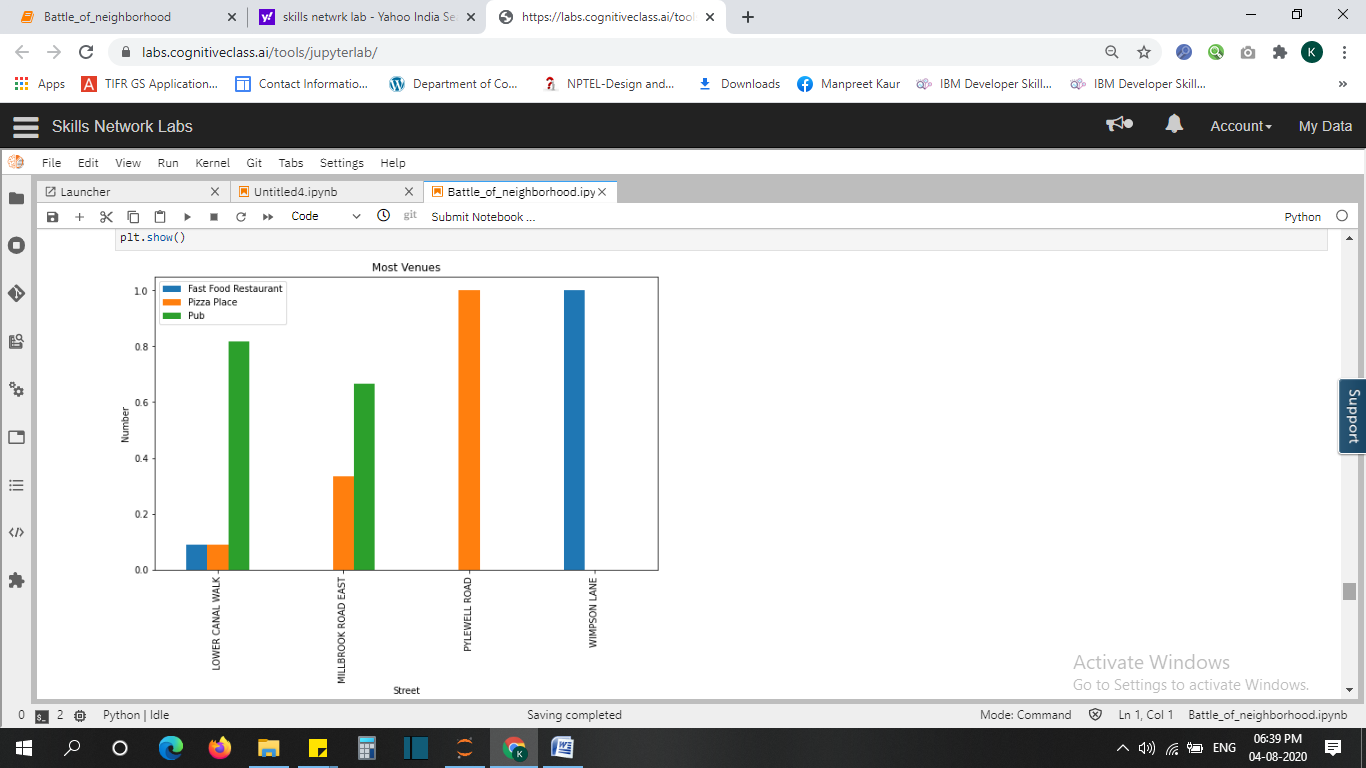
From the bar graph titled ‘Cheapest Street’ we can clearly see that street – ‘PYLEWELL ROAD’ is the cheapest of all the streets and the street – ‘MILLBROOK ROAD EAST’ is the one with highest average price. So we would like to explore the cheapest street avoiding the ones with higher costs for further analysis.

**3.2 Using FOURSQUARE**

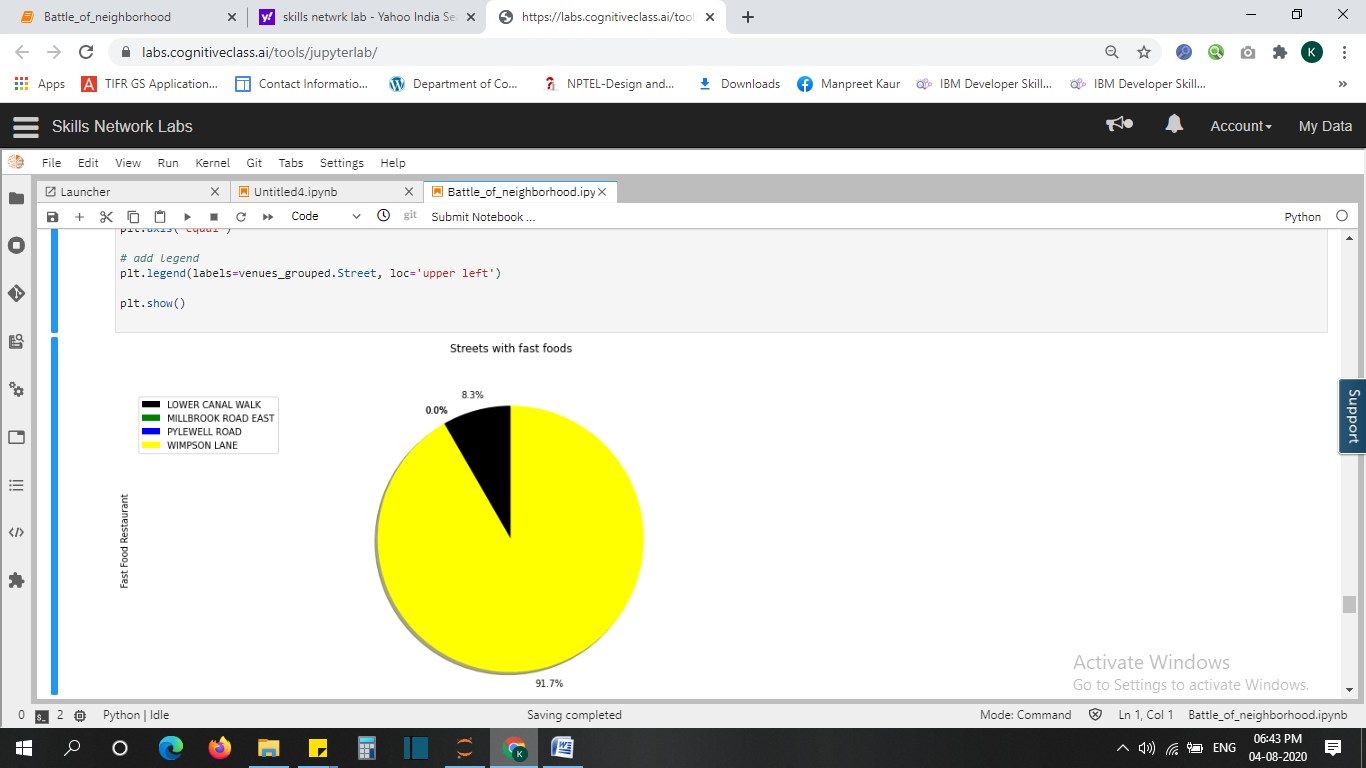
We will signup for Foursquare and generate credentials which include Id and SECRET. Using these credentials we will query the Foursquare Api. Our main query is to explore venues in the above 5 streets , get the categories of these venues. Then these venues will be shown to the customer as facilities of streets respectively which will help in narrowing our search for the most suitable street.

Since the customer was interested in places like pub , pizza places and food restaurants, we will eliminate rest of the rows and keep the ones satisfying above conditions.

Using a bar graph we will compare the streets with more facilities.



We will use a pie chart to get an overview of the fractions that we evaluated using one-hot encoding means to visualize the street which has maximum maount of fast food venues.



Now we will create a data frame using sorting methods to see which city has which venue in the highest amount i.e the most common venues within the chosen streets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NO.** | **Street** | **1st Most Common Venue** | **2nd Most Common Venue** | **3rd Most Common Venue** |
| 0 | LOWER CANAL WALK | Pub | Pizza Place | Fast Food Restaurant |
| 1 | MILLBROOK ROAD EAST | Pub | Pizza Place | Fast Food Restaurant |
| 2 | PYLEWELL ROAD | Pizza Place | Pub | Fast Food Restaurant |
| 3 | WIMPSON LANE | Fast Food Restaurant | Pub | PizzaPlace |

**3.3 CLUSTERING METHOD**

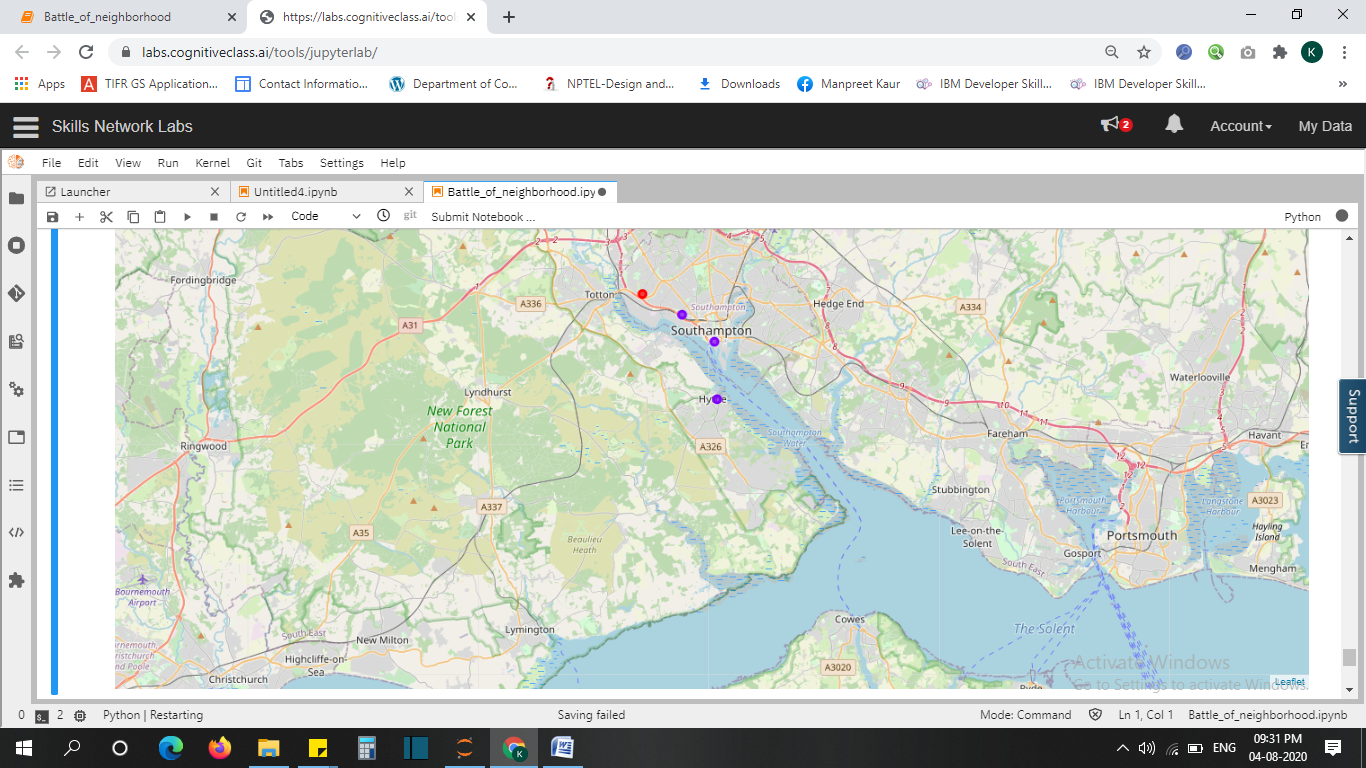
**3.3.1 k-means**

Segmentation is the practice of partitioningthe streets based into groups of streets that have similar characteristics. One of the algorithms that can be used for segmentation is K-means clustering. K-means can group data when it isunsupervised based , on the similarity of streets to each other. K-means is a type of partitioning clustering. It divides the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. K-means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.

Here we will assume K=2 and see the result:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Street** | **Avg-price** | **city\_coord** | **Latitude** | **Longitude** | **ClusterLabels** |
| PYLEWELL ROAD | 60000.0 | (50.8694403, -1.4000328) | 50.869440 | -1.400033 | 0 |
| WIMPSON LANE | 85500.0 | (50.9201304, -1.4571668) | 50.920130 | -1.457167 | 1 |
| LOWER CANAL WALK | 115000.0 | (50.897111, -1.4022897) | 50.897111 | -1.402290 | 0 |
| MILLBROOK ROAD EAST | 119500.0 | (50.9098907, -1.4266578) | 50.909891 | -1.426658 | 0 |

Here the column clusterlabels define the cluster no. Assigned to each street. We will see this insight in the Southampton map as below , where Wimpson Lane is assigned red label as it belongs to cluster 1 where as rest three streets are grouped under segment 0 . Therefore they will have a blue label.



**4. Result Section**

Its not so hard to find a cheap place to live in Southampton. Wimpson Lane seems to be the most appropriate one.

**5. Discussion Section**

Wimpson Lane with cluster label as one and an average price of 85,500 seems to be the appropriate place to be recommended because it has as 1st Most Common Venue : Fast Food Restaurants and 2nd Most Common Venue is Pubs. Pylewell Road was cheaper but the facilities of pub,food restaurants were not much available.

**6. Conclusion**

In thisstudy i used k-means clustering to segment the streets data collected from an official site of UK government and Foursquare Api. I was able to provide a clear comparison among all the streets based on houseing prices and facilities available using various graphs like bar graph and pie chart .This model was successfully used to provide a satisfactory recommendation to the customer who seeks a cheap house in Southampton, UK. Thus this model recommends a house in Wimpson Lane as the best solution to the problem described by the customer.