BATTLE OF NEIGHBORHOODS

THE FOODIES

COURSERA CAPSTONE PROJECT

BY:

GURPUR NAMITHA KAMATH

INTRODUCTION

- AN ENTREPRENEUR MAKES SACRIFICES, INCURS LOSSES IN THE BEGINNING
- LOCATION IS THE BOSS
- BEST LOCATION SHOULD BE ABLE TO DRAW CROWDS, ACCESSIBLE AND HAVE GROWTH POTENTIAL

OBJECTIVE:

ANALYZE AS TO WHETHER MANHATTAN AS A LOCATION FOR AN INDIAN RESTAURANT IS FEASIBLE OR NOT.

NEW YORK CITY

- MOST POPULATED CITY IN THE UNITED STATES
- FINANCIAL, MEDIA AND CULTURAL CAPITAL OF THE WORLD.
- KNOWN FOR ITS SKY SCRAPERS AND TOURIST PLACES
- LAND OF OPPORTUNITIES AT THE SAME TIME EXTREMELY COMPETITIVE AND FAST PACED.
- STANDARAD OF LIVING AND THE COST OF A STARTING A BUSINESS IS EXTREMELY HIGH
- CULTURALLY DIVERSE AS IT IS THE LEADING GATEWAY FOR IMMIGRANT POPULATION.
- THERE ARE FIVE BOROUGHS NAMELY - BROOKLYN, QUEENS, THE BRONX, MANHATTAN AND STATEN ISLAND.

MANHATTAN

- MOST DENSELY POPULATED OF ALL THE BOROUGHS
- HAS TOP ATTRACTIONS LIKE EMPIRE STATE BUILDING, CENTRAL PARK, CENTRAL PARK, THE CHRYSLER BUILDING ETC.
- HOME TO THE WORLD'S LARGEST TWO STOCK EXCHANGES - THE NASDAQ AND NEW YORK STOCK EXCHANGE.

DATA REQUIREMENTS

SOURCE OF THE DATA:

- THE DATA IS SOURCED FROM THE FOLLOWING LINK:HTTPS://GEO.NYU.EDU/CATALOG/NYU_2451_34572
- THE DATA OF NEW YORK CITY HAS BEEN USED
- THE DATA FRAME HAS 5 BOROUGHS AND 306 NEIGHBORHOODS.
- THE DATA WILL BE USED TO ANALYZE THE BOROUGH MANHATTAN AND ITS NEIGHBORHOODS
- FOURSQAURE API IS USED FOR GATHERING THE DATA RELATING TO NEARBY VENUES, RESTAURANTS BY LEVERAGING ON THE GEOGRAPHICAL COORDINATES OF MANHATTAN
- THE IMAGE SHOWS THE DATA OF MANHATTAN GROUPED BY ITS NEIGHBORHOODS AND VENUES

HENCE, DATA IS THE KEY INGREDIENT FOR THE PREPARATION OF A RECIPE CALLED DATA ANALYSIS

2]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Battery Park City	100	100	100	100	100	100
Carnegie Hill	100	100	100	100	100	100
Central Harlem	44	44	44	44	44	44
Chelsea	100	100	100	100	100	100
Chinatown	100	100	100	100	100	100
Civic Center	100	100	100	100	100	100
Clinton	100	100	100	100	100	100
East Harlem	44	44	44	44	44	44
East Village	100	100	100	100	100	100
Financial District	100	100	100	100	100	100
Flatiron	100	100	100	100	100	100
Gramercy	100	100	100	100	100	100
Greenwich Village	100	100	100	100	100	100

METHODOLOGY

BUSINESS UNDERSTANDING:

- IT IS THE CORE QUESTION
- A CLEARLY DEFINED QUESTION DIRECTS THE ANALYTICAL APPROACH THAT WILL BE REQUIRED TO SOLVE THE PROBLEM.
- MAIN QUESTION HERE IS TO ANALYZE WHETHER MANHATTAN AS AN AREA IS FEASIBLE TO OPEN A RESTAURANT OR NOT.

ANALYTIC APPROACH:

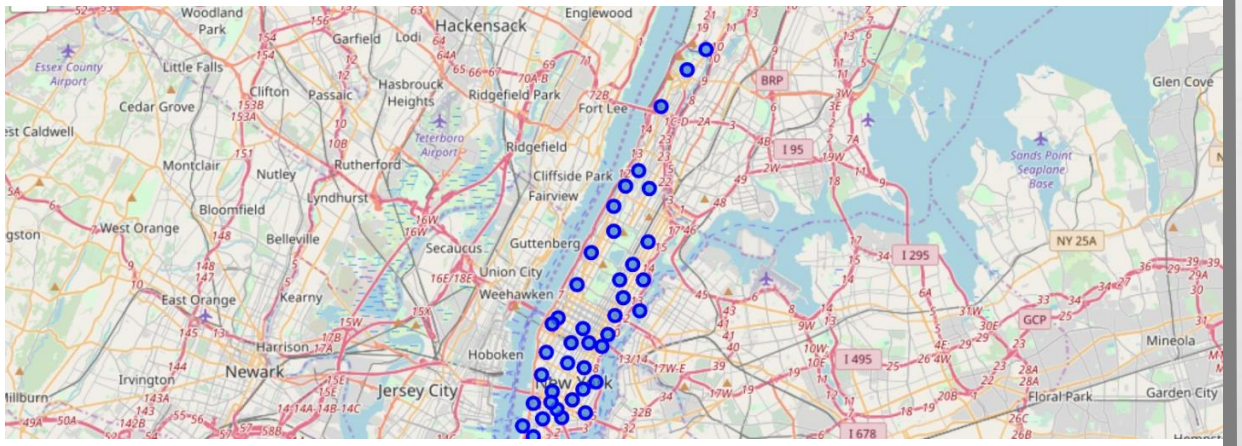
- EXPLORATORY DATA ANALYSIS WAS USED: TECHNIQUES SUCH AS PREDICTIVE OR DESCRIPTIVE STATISTICS AND VISUALIZATION CAN BE APPLIED TO THE DATA SET TO ASSESS THE CONTENT, QUALITY AND OFFER INITIAL INSIGHTS ABOUT THE DATA.

METHODOLOGY (CONT)

- THE IMAGE ON THE RIGHT SHOWS THE MAP OF NYC WITH ALL ITS BOROUGHS
- THE NEXT IMAGE SHOWS THE BIFURCATION OF THE BROUGHS
- MANHATTAN HAS 40 NEIGHBORHOODS

Borough	
Queens	81
Brooklyn	70
Staten Island	63
Bronx	52
Manhattan	40

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688



METHODOLOGY [CONT]

- THE IMAGE ON THE LEFT SHOWS THE NEIGHBORHOODS OF MANHATTAN WITH ITS GEOGRAPHICAL COORDINATES
- THE OTHER IMAGE IS THE MAP OF MANHATTAN WITH ALL ITS NEIGHBORHOODS
- 3302 VENUES WITH 331 VENUE CATEGORIES WERE GENERATED FOR THE 40 OF ITS NEIGHBORHOODS.
- FOURSQUARE WAS USED

METHODOLOGY (CONT)

- MACHINE LEARNING IS THE SCIENTIFIC STUDY OF ALGORITHMS
- IT HAS VARIOUS CATEGORIES
  - ✓ SUPERVISED LEARNING
  - ✓ SEMI-SUPERVISED LEARNING
  - ✓ UNSUPERVISED LEARNING
- UNSUPERVISED LEARNING HAS BEEN USED HERE
- IT BUILDS A MATHEMATICAL MODEL FROM A SET OF DATA WHICH CONTAINS ONLY INPUTS AND NO DESIRED OUTPUT LABELS
- SILHOUETTE (CLUSTERING) REFERS TO A METHOD OF INTERPRETATION AND VALIDATION OF CONSISTENCY WITHIN CLUSTERS OF DATA
- SILHOUETTE ANALYSIS SHOWS THE CLUSTER QUALITY AND HELPS TO FIND THE K CLUSTERS THROUGH THE MEANS.

METHODOLOGY(CONT)

- I USED THE SILHOUETTE SCORE FOR ANALYSIS.
- 9 CLUSTERS HAVE BEEN USED AND SCORES OF EACH HAVE BEEN PROVIDED
- IT CAN INFERRED THAT THE SCORES FOR N_CLUSTERS 2,6,7,8,9 ARE BELOW AVERAGE
- THE SCORES FOR N_CLUSTERS 3 & 5 IS ABOVE AVERAGE BUT NOT SATISFACTORY
- BUT THE N_CLUSTER 4 HAS THE HIGHEST SCORE AND IS THE BEST

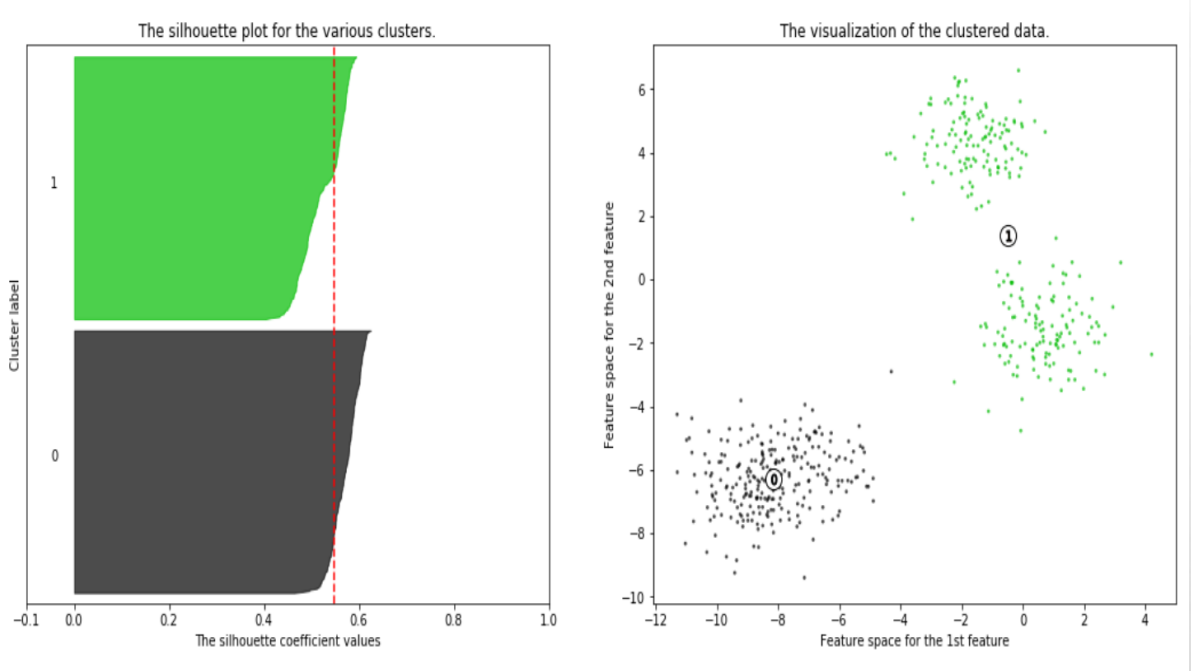
```
Automatically created module for IPython interactive environment
For n_clusters = 2 The average silhouette_score is : 0.547358312599
For n_clusters = 3 The average silhouette_score is : 0.679029294409
For n_clusters = 4 The average silhouette_score is : 0.813771753455
For n_clusters = 5 The average silhouette_score is : 0.632702179746
For n_clusters = 6 The average silhouette_score is : 0.453070706527
For n_clusters = 7 The average silhouette_score is : 0.282396769658
For n_clusters = 8 The average silhouette_score is : 0.102367146321
For n_clusters = 9 The average silhouette_score is : 0.101872995931
```

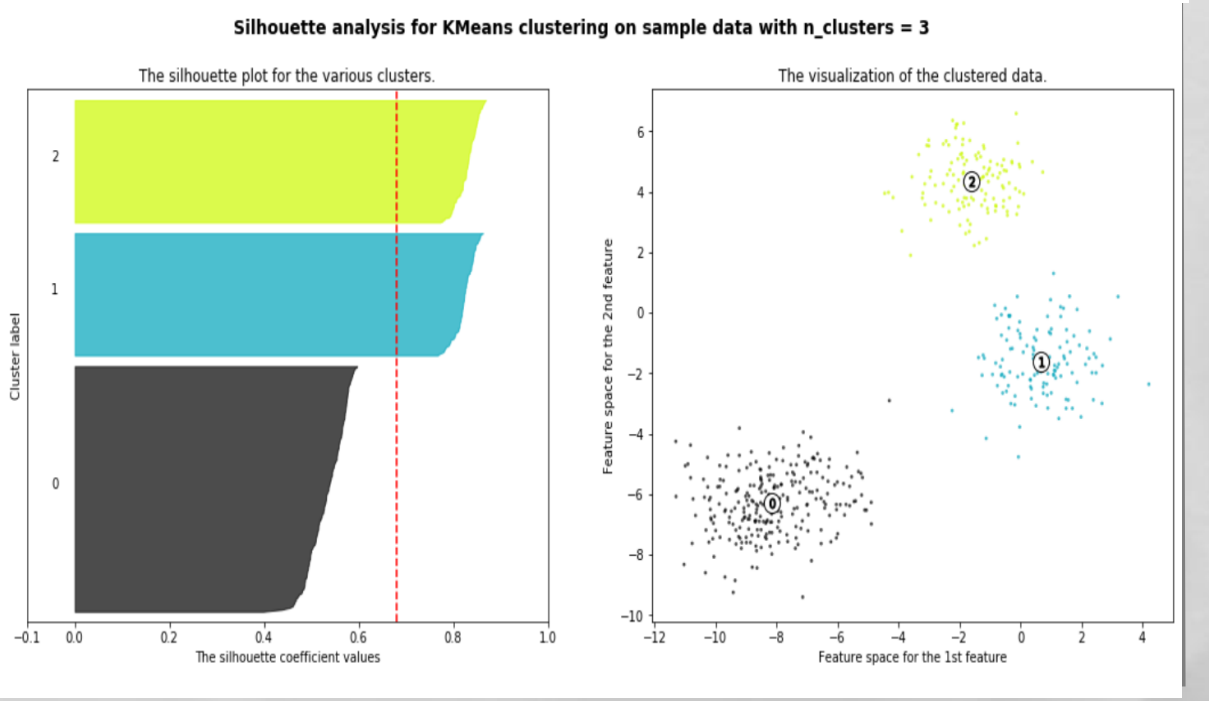
METHODOLOGY(CONT)

- THE PLOTS ON THE RIGHT SHOW THE SILHOUETTE ANALYSIS FOR K MEANS CLUSTERING ON SAMPLE DATA WITH N_CLUSTERS:

✓ THE FIRST ONE IS N_CLUSTERS=2 WHERE THE SCORE IS BELOW AVERAGE

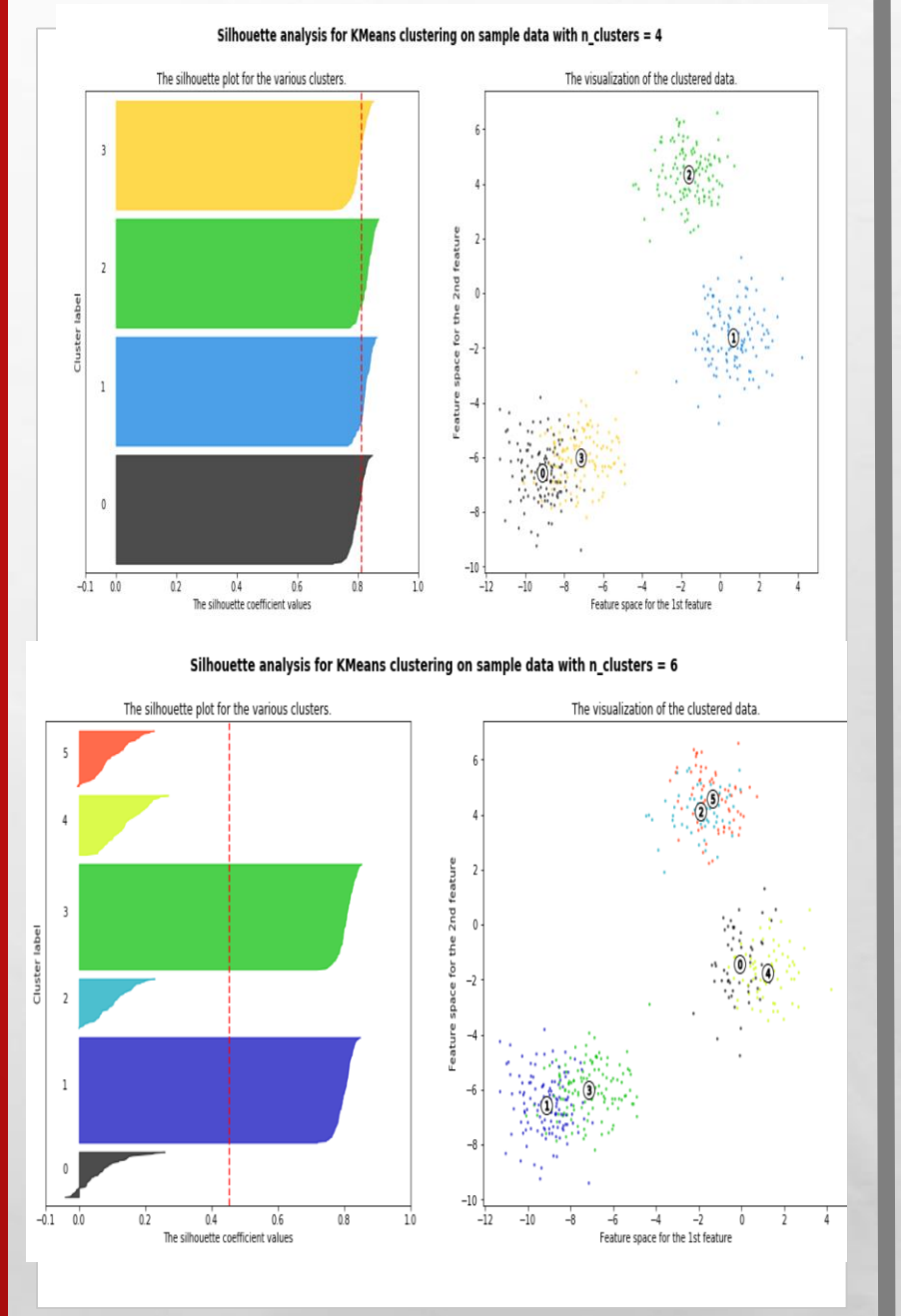
✓ THE SECOND ONE IS N_CLUSTER=3 WHERE THE SCORE IS SLIGHTLY ABOVE AVERAGE BUT NOT QUITE SATISFACTORY





METHODOLOGY(CONT)

- THE PLOTS ON THE RIGHT SHOW THE SILHOUETTE ANALYSIS FOR K MEANS CLUSTERING ON SAMPLE DATA WITH N_CLUSTERS =4 & 6
- N_CLUSTERS =6 HAS A SCORE BELOW AVERAGE AND THERE ARE WIDE FLUCTUATIONS IN THE PLOT
- N=_CLUSTER=4 HAS THE HIGHEST SCORE AND THE PLOT IS OF SIMILAR THICKNESS AND SIZE



METHODOLOGY(CONT)

- FOR DETERMINING THE OPTIMAL VALUE OF K FOR OUR DATASET, I HAVE USED THE SILHOUETTE COEFFICIENT METHOD.
- THE MAXIMUM COEFFICIENT IS FOR CLUSTER 2 THAT IS 0.9254248520430671
- 2 IS THE OPTIMAL NUMBER OF CLUSTERS

```
For n_clusters=2, The Silhouette Coefficient is 0.9254248520430671
For n_clusters=3, The Silhouette Coefficient is 0.7678253176858705
For n_clusters=4, The Silhouette Coefficient is 0.7900175659521259
For n_clusters=5, The Silhouette Coefficient is 0.8025055551640486
For n_clusters=6, The Silhouette Coefficient is 0.8186954158820021
For n_clusters=7, The Silhouette Coefficient is 0.8350223550134503
For n_clusters=8, The Silhouette Coefficient is 0.8473675449993284
For n_clusters=9, The Silhouette Coefficient is 0.8544630907778618
```

RESULTS

- CLUSTERING ANALYSIS WAS DONE THE BASIS OF THE RESTAURANTS IN THE NEIGHBORHOODS OF MANHATTAN
- CLUSTER 0 HAS A POSITIVE VALUE WHILE CLUSTER 1 HAS A NEGATIVE VALUE
- THE POSITIVE VALUE IS NOT VERY HIGH AND INDICATES THAT THE MARKET IS NOT SATURATED
- THE MAP SHOWS THE CLUSTERS OF MANHATTAN NEIGHBORHOODS

DISCUSSIONS

- THERE ARE DIFFERENT TYPES OF RESTAUARANTS IN MANHATTAN
- MANHATTAN AS A LOCATION IS NOT VERY SATURATED AS WE CAN SEE FROM THE ANALYSIS
- A RESTAUARANT WITH GREAT MENU AND TASTY CUSINES CAN BE OPENED.

CONCLUSION

- ANALYSIS IS ON LIMITED DATA
- IT CAN BE SAFELY CONCLUDED THAT AN INDIAN RESTAURANT CAN BE OPENED
- THERE ARE DIFFERENT TYPES OF RESTAUARANTS WHICH OFFER DIFFERENT TYPES OF CUISINES FROM AROUND THE GLOBE IN MANHATTAN.
- THERE IS DEMAND FOR GOOD FOOD AND DIFFERENT CUISINE TYPES
- THERE IS COMPETITION, BUT COMPETITION IS A FEATURE OF A HEALTHY MARKET CONDITION
- LOCATION IS THE BOSS FOR ANY BUSINESS, MANHATTAN SEEMS TO SATISFY THAT CONDITION
- I CAN CONCLUDE THAT THE COMBINATION OF THE LOCATION PLUS MY CLIENT'S BRAND AND HIS CUISINE CAN TOGETHER CREATE A SUCCESS STORY HERE IN MANHATTAN AS WELL.


THANK
YOU