

Statistical-Modelling- Project

By: Ibadet Azemi



FOURSQUARE

PROJECT/GOALS:

- ❑ Determine the strength of the relationship between number of bikes + POI's (Characteristics: reviews, ratings, distance, location)



PART 1: CONNECTING TO CITYBIKES API

- Send a request to CitiBikes for city of your choice
- Find longitude/latitude + number of bikes
- Find bike stations
- Create a dataframe of results

	Latitude	Longitude	Number of Bikes
0	40.734786	-74.050444	12
1	40.737604	-74.052478	19
2	40.724605	-74.078406	0
3	40.728745	-74.032108	13

- ❑ Retrieve API to get info for networks (New York City)
- ❑ Parsing my response into a dataframe/database
- ❑ Convert my JSON file into a database that will go into SQLite database



Part 2: Connecting to Foursquare and Yelp APIs



- ❑ On terminal creating env variable
- ❑ Send a request to Foursquare with a small radius (1000) for all the bike stations in NYC
- ❑ Parse through the response to get the POI (such as restaurants, bars, etc) details you want (ratings, name, location, etc)
- ❑ Comparing Results: Which API provided you with more complete data? Provide an explanation.

(Yelp API (Easy to use, Specific to detail)

	Bike Stations	Ratings	Restaurants	Number of Bikes
0	W 100 St & Manhattan Ave	4.5	Anitas	12
1	7 Ave & Central Park South	4.5	Taste Of Italy	19
2	Bedford Ave & Bergen St	4.5	Taqueria Gardenias	0
3	28 St & 41 Ave	5	Golden Diner	13

TOP 10 RESTAURANTS ACCORDING TO THEIR RATING:

- ☐ Anitas Restaurant
- ☐ A Taste Of Italy
- ☐ Taqueria Gardenias
- ☐ Hot-Dog-Jays-Little-Falls
- ☐ ME Mediterranean Eatery
- ☐ First Wok Chinese Restaurant
- ☐ Camellia Milk Tea
- ☐ Halal Cart
- ☐ Pistache NYC
- ☐ John's Market



PART 3: JOINING DATA

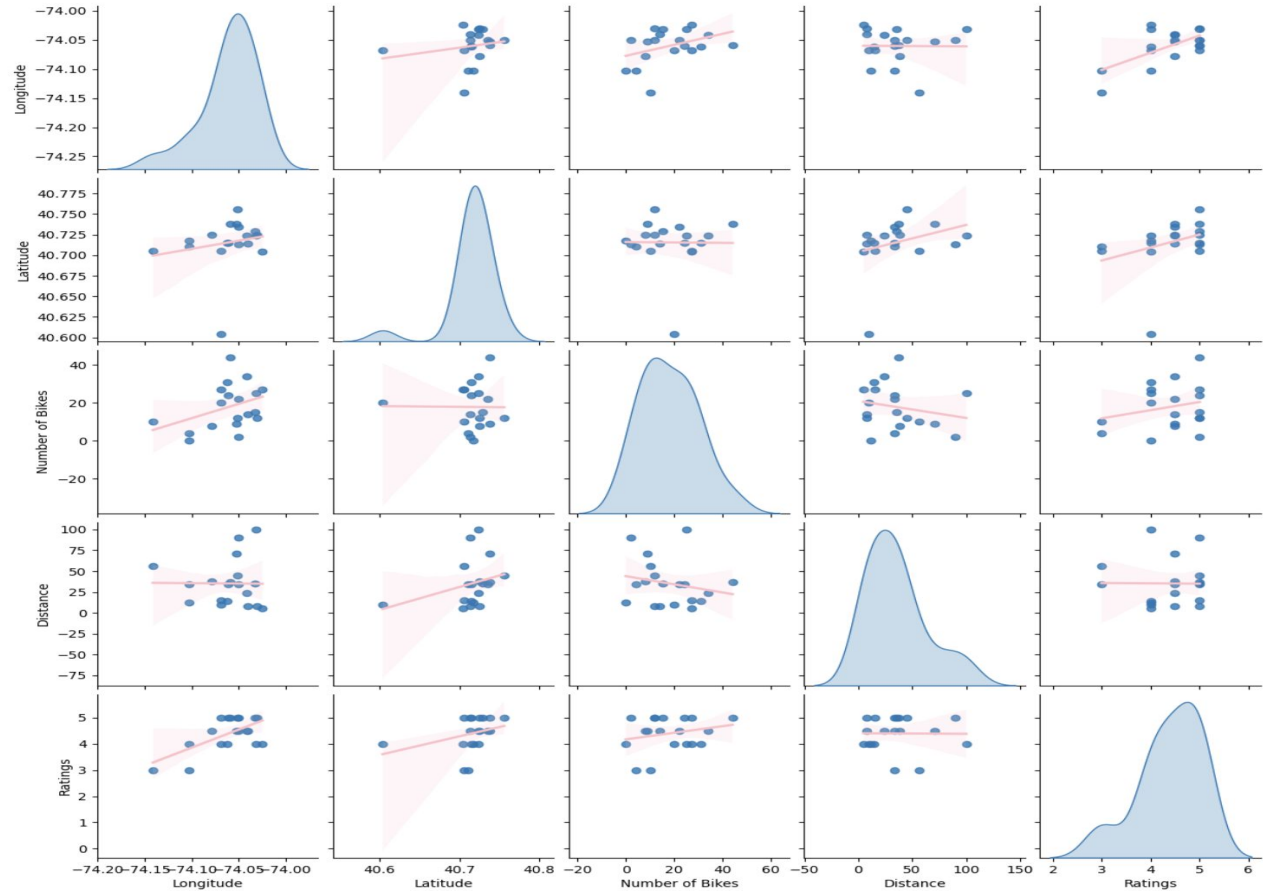
JOINING DATABASES/TABLES

	Longitude	Latitude	Number of Bikes	Restaurants	Distance	Ratings	Bike Stations	Returned	Slots	Reviews	Count_POI
0	-74.050444	40.734786	22	Anitas	34	4.5	W 100 St & Manhattan Ave	0	2	100	20
1	-74.052478	40.737604	9	Taste Of Italy	71	4.5	7 Ave & Central Park South	10	0	45	20
2	-74.078406	40.724605	8	Taqueria Gardenias	38	4.5	Bedford Ave & Bergen St	3	9	12	34
3	-74.032108	40.728745	15	Hot-Dog-Jays	35	5.0	28 St & 41 Ave	4	10	45	45
4	-74.061107	40.714528	24	ME	34	5.0	Frederick Douglass Blvd & W 112 St	3	34	34	18
5	-74.141107	40.705302	10	First Wok	56	3.0	W 113 St & Broadway	21	23	16	16
6	-74.103217	40.717643	0	Camellia Milk Tea	12	4.0	8 Ave & W 31 St	12	5	1	12
7	-74.041386	40.723786	34	Pistache NYC	24	4.5	Bergen St & Vanderbilt Ave	9	0	0	15
8	-74.031466	40.723574	25	Johns Market	100	4.0	Monroe St & Bedford Ave	11	3	150	19
9	-74.050444	40.713317	2	Jajaja Mexicana	90	5.0	35 Ave & 37 St	0	6	56	6
10	-74.062305	40.714675	31	Kiki's	14	4.0	Bank St & Hudson St	2	9	78	17
11	-74.030446	40.724605	12	Betty	8	5.0	Greenwich St & W Houston St	7	34	200	13
12	-74.024708	40.704201	27	Double Chicken Pleas	5	4.0	Madison St & Clinton St	3	16	67	17
13	-74.068408	40.604301	20	Ye's Apothecary	10	4.0	Stanton St & Mangin St	13	6	89	23
14	-74.040554	40.713675	14	Wu's Wonton King	8	4.5	Clinton Ave & Flushing Ave	34	7	90	34
15	-74.103217	40.710675	4	Wayla	34	3.0	Rivington St & Ridge St	45	2	11	29
16	-74.051105	40.755773	12	Bacaro	45	5.0	E 114 St & 1 Ave	9	5	159	19
17	-74.068506	40.705302	27	Golden Diner	15	5.0	Columbia St & Lorraine St	6	0	200	20
18	-74.059377	40.737604	44	La Contenta	37	5.0	Riverside Dr & W 104 St	4	0	123	20

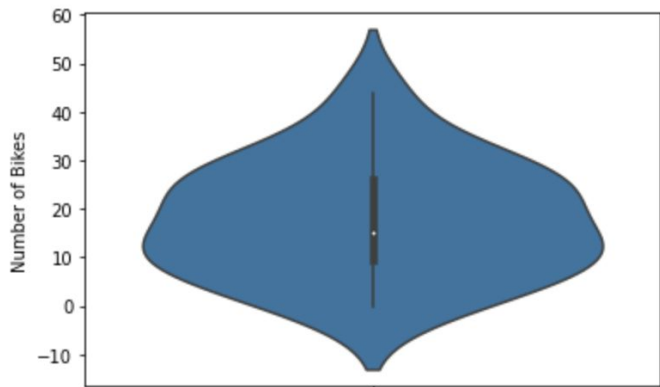
VISUALIZATIONS: EDA

PAIRPLOT:

- ❑ No correlation was found
- ❑ R-Squared: is weak

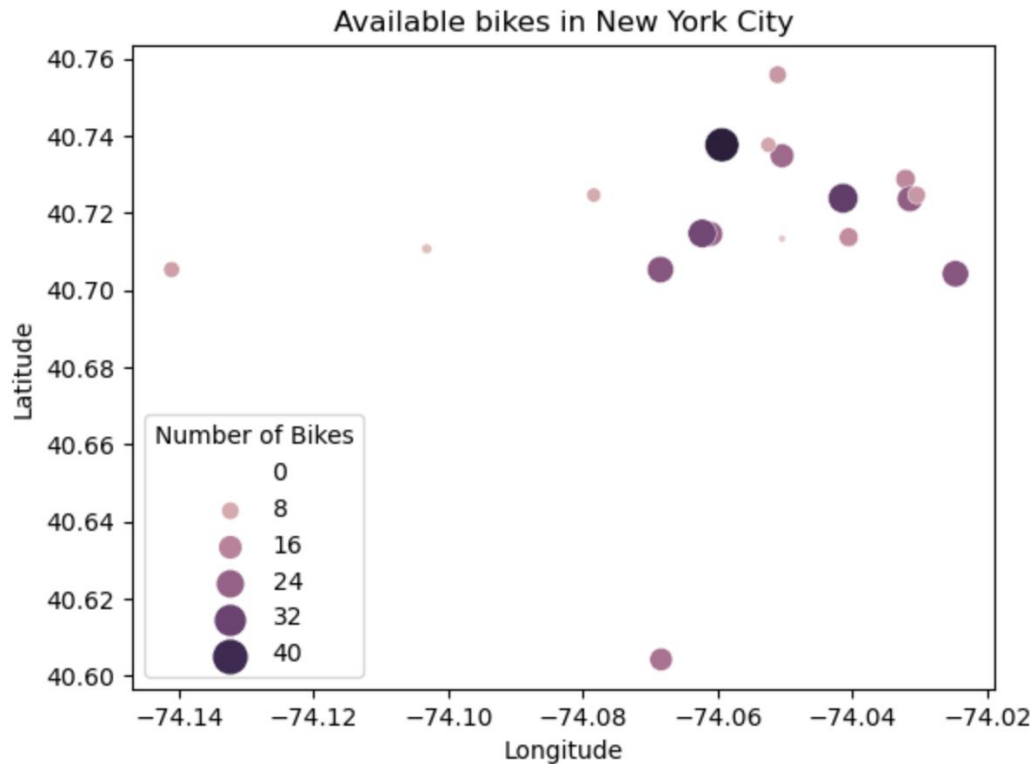


VISUALIZATIONS:EDA



Visual EDA Analysis:

- ❑ There does not seem to be a relationship
- ❑ There are outliers



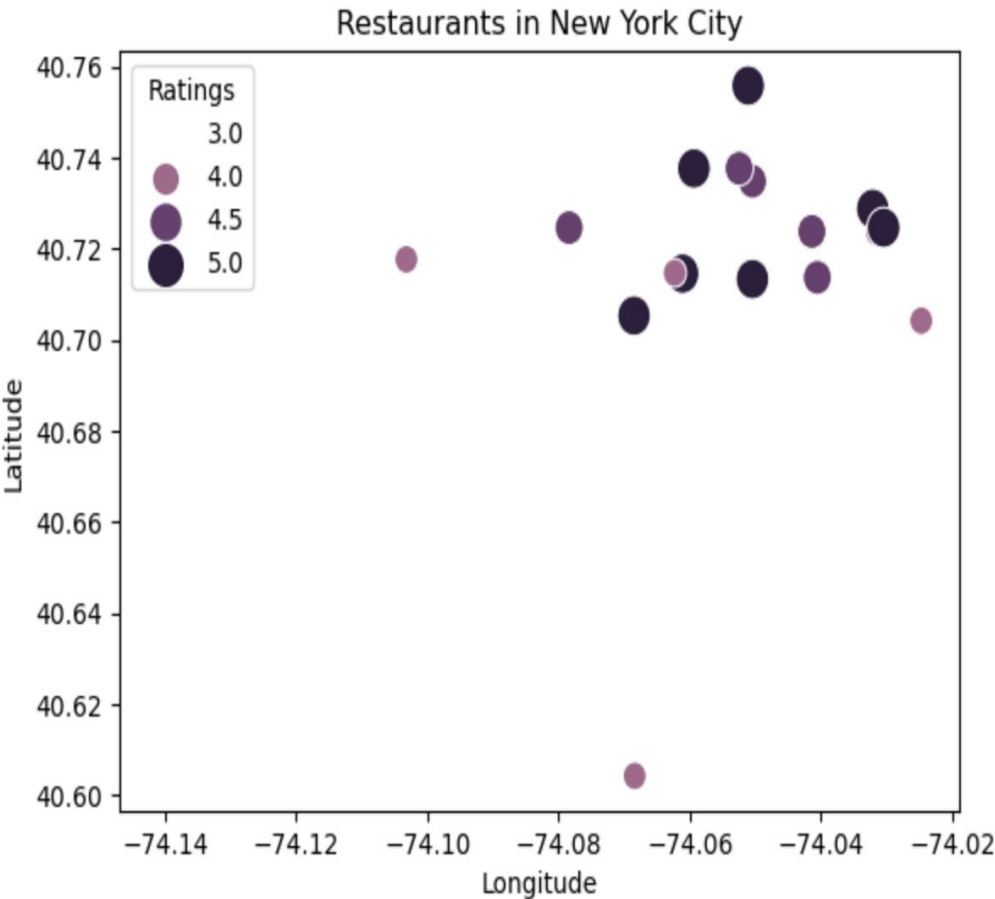
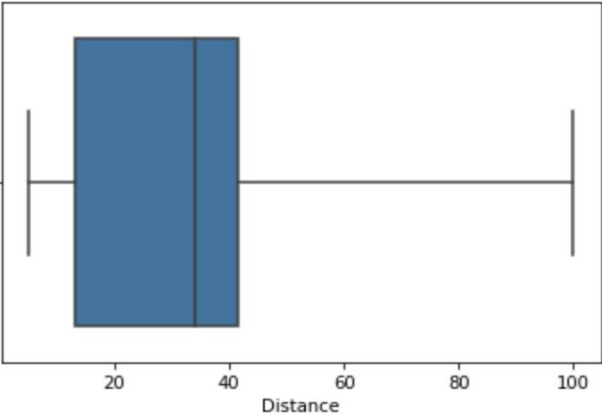
VISUALIZATIONS:EDA



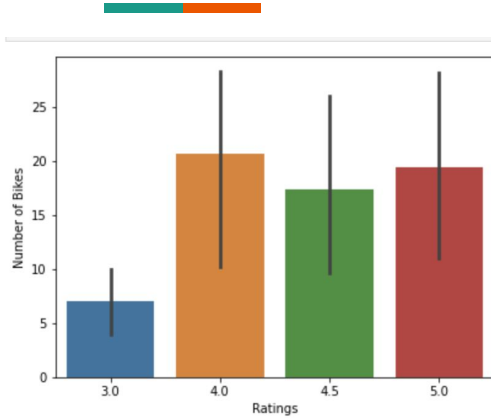
Visual EDA Analysis:

- ☐ There does not seem to be a relationship
- ☐ There are outliers

<Axes: xlabel='Distance'>

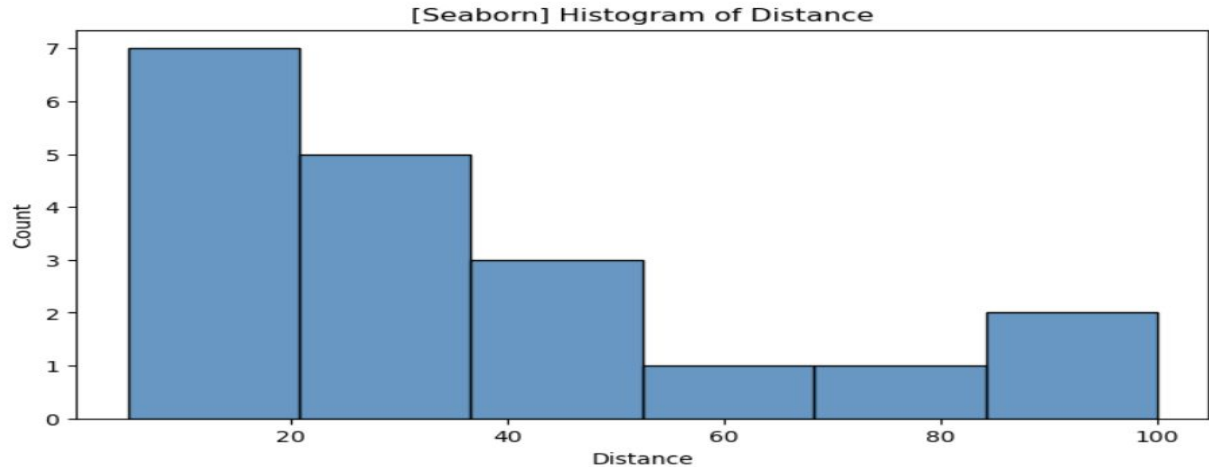
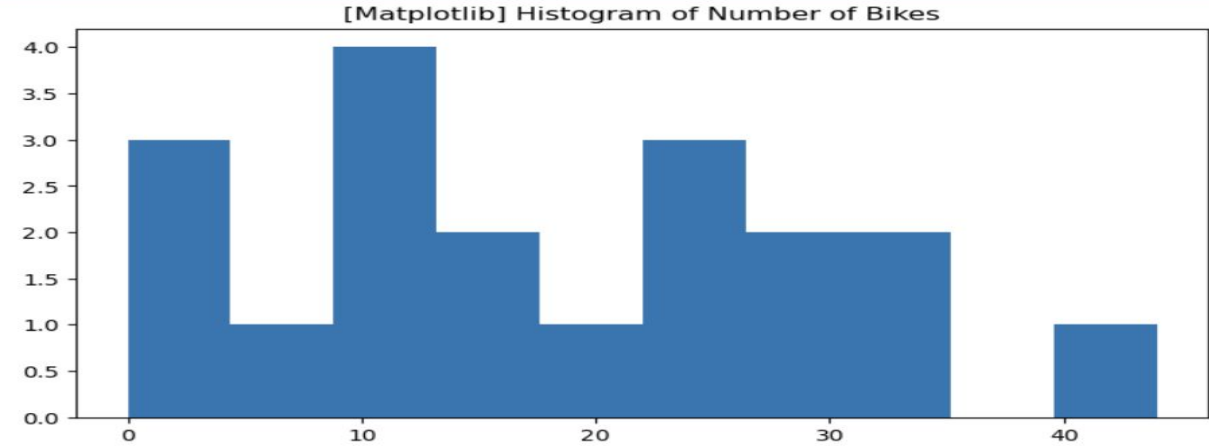


VISUALIZATIONS:EDA



Visual EDA Analysis:

- ❑ There does appear to be a relationship
- ❑ Outliers



Step 4.) Model Building

OLS REGRESSION RESULTS:

Interpretation


Based on the interpretation of the results, it seems that the regression model that was built to predict number of bikes, ratings and distance is performing very well.

The low R-squared value appears to reflect that there is a strong correlation

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified

OLS Regression Results						
Dep. Variable:	Number of Bikes		R-squared:	0.097		
Model:	OLS		Adj. R-squared:	-0.016		
Method:	Least Squares		F-statistic:	0.8584		
Date:	Mon, 21 Aug 2023		Prob (F-statistic):	0.442		
Time:	05:11:46		Log-Likelihood:	-72.294		
No. Observations:	19		AIC:	150.6		
Df Residuals:	16		BIC:	153.4		
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	2.3139	19.832	0.117	0.909	-39.729	44.357
Ratings	4.2563	4.387	0.970	0.346	-5.045	13.557
Distance	-0.0886	0.102	-0.871	0.396	-0.304	0.127
Omnibus:	0.875		Durbin-Watson:	2.087		
Prob(Omnibus):	0.646		Jarque-Bera (JB):	0.850		
Skew:	0.399		Prob(JB):	0.654		
Kurtosis:	2.339		Cond. No.	331.		

R-SQUARED/ADJ R. SQUARED & F-STATISTICS



R-squared:	0.097
Adj. R-squared:	-0.016
F-statistic:	0.8584
Prob (F-statistic):	0.442
Log-Likelihood:	-72.294
AIC:	150.6
BIC:	153.4

R-SQUARED: Is a weak correlation

ADJ. R-SQUARED: Shows model is good at predicting as is equal to R-squared

F-STATISTIC: Is greater than Prob(F-Statistic)

Prob (F-Statistic): Is less than F-Statistic

COEFF & P-VALUE

	coef	std err	t	P> t	[0.025	0.975]
const	2.3139	19.832	0.117	0.909	-39.729	44.357
Ratings	4.2563	4.387	0.970	0.346	-5.045	13.557
Distance	-0.0886	0.102	-0.871	0.396	-0.304	0.127

<> Coefficient: Is positive and has a positive affect on Y

<> T-Statistic: Strong coefficient

<> P-Value: Significance of each and is a strong coefficient



CHALLENGES:

- ❑ Time consuming for a large city
- ❑ 1119 bike stations in NYC
- ❑ Lack of bikes
- ❑ Limited calls for Yelp
- ❑ Limited API calls
- ❑ Learning and understanding API's
- ❑ Converting json response to csv
- ❑ Combining multiple data frames/csv



RESULTS:

- ❑ <> Coefficient: Is positive and has a positive affect on Y
- ❑ <> T-Statistic: Strong coefficient
- ❑ <> P-Value: Significance of each and is a strong coefficient

The regression model that was built to predict number of bikes, ratings and distance is performing very well.

The low R-squared value appears to reflect that there is a strong correlation



FUTURE GOALS:

- ❑ More data cleaning
- ❑ Going more in depth with the process
- ❑ More POI's
- ❑ Better/more visualizations(EDA)



THANK YOU!



FOURSQUARE