BIKE STATIONS AND BUSINESSES STATISTICAL MODELLING PROJECT

MELISSA NIELSEN

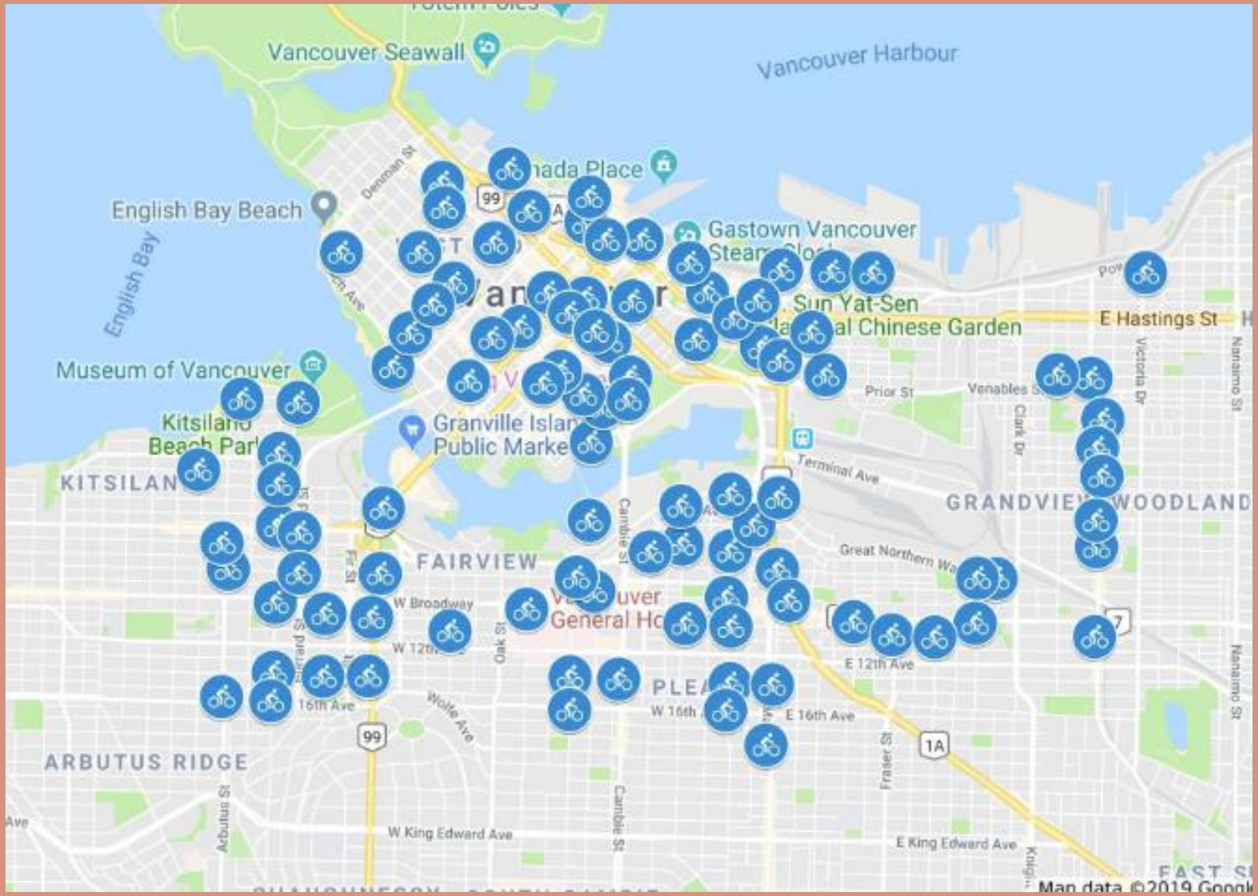
NOVEMBER 7, 2022

PROJECT OVERVIEW

1. Get data from CityBikes API
2. Get data from Foursquare and Yelp APIs
3. Join data and create database
4. Create regression model

PROJECT SCOPE

1. Looked at Mobi bikes in Vancouver, BC
2. Investigated following business types within 100 m of every bike station:
    - Bars
    - Restaurants



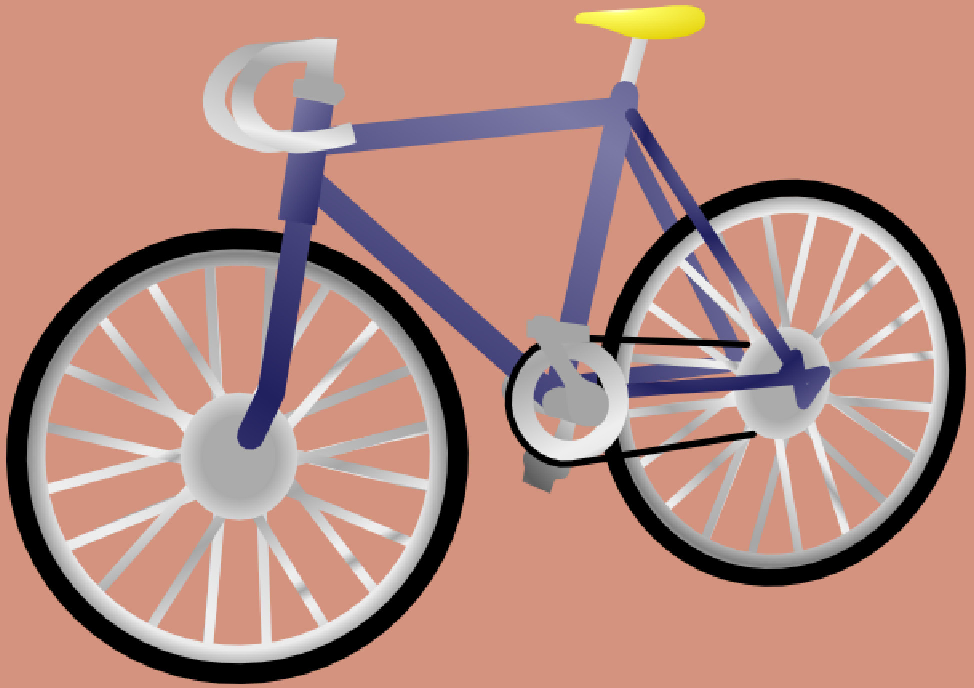
KEY QUESTIONS

1. Are business attributes correlated with the proportion of available bikes?

2. Is the number of nearby bars and restaurants correlated with the proportion of available bikes?

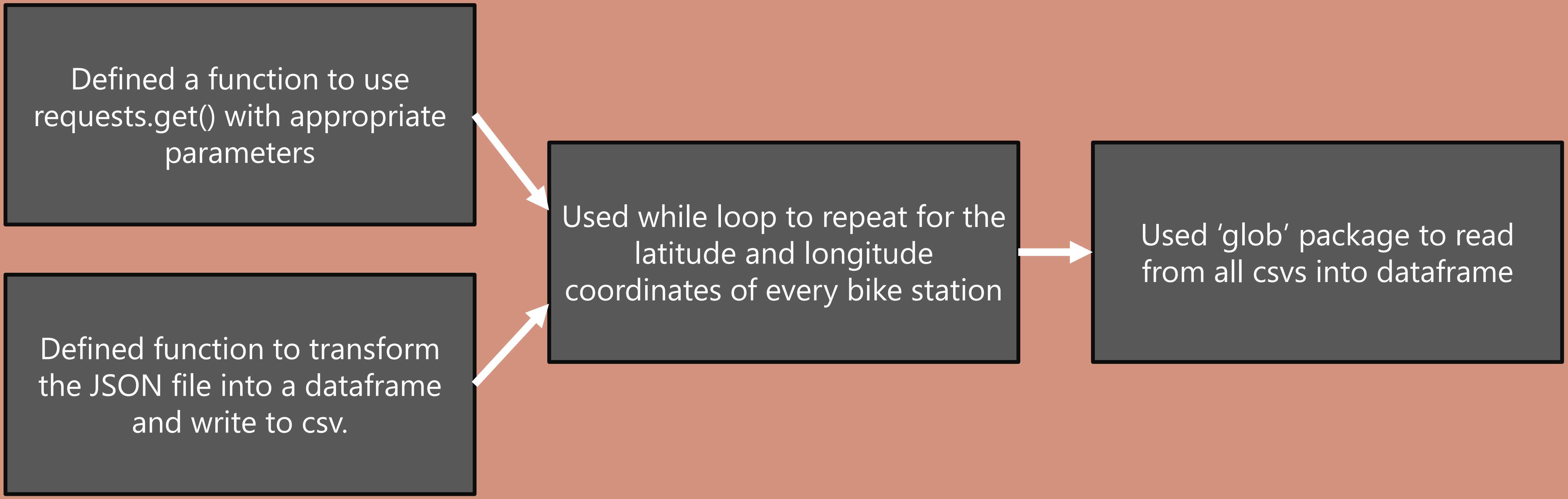
STEP 1: CITY BIKE API

1. Parsed JSON file
2. Removed stations with status “offline”



STEP 2: GET DATA FROM FOURSQUARE AND YELP APIS

1. Used requests.get() function
2. For each of the Yelp and Foursquare APIs, created the following:



Defined a function to use requests.get() with appropriate parameters

Defined function to transform the JSON file into a dataframe and write to csv.

Used while loop to repeat for the latitude and longitude coordinates of every bike station

Used ‘glob’ package to read from all csvs into dataframe

STEP 2: GET DATA FROM FOURSQUARE AND YELP APIS

1. Used data from previous step to create 3 tables (for each Foursquare and Yelp):

Bike Station Table:
information about **unique** bike stations

key = station_id

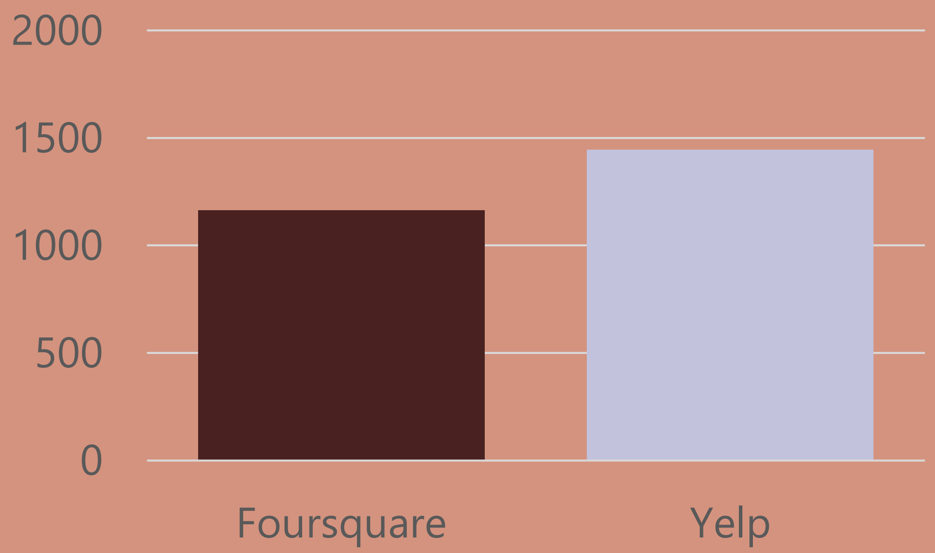
Intermediary Table:
two columns only:
business id & bike station id

key = two above columns combined

Business Table:
information about **unique** businesses

key = business id (fsq_id for Foursqare and id for Yelp)

STEP 2: COMPARE RESULTS FROM FOURSQUARE AND YELP



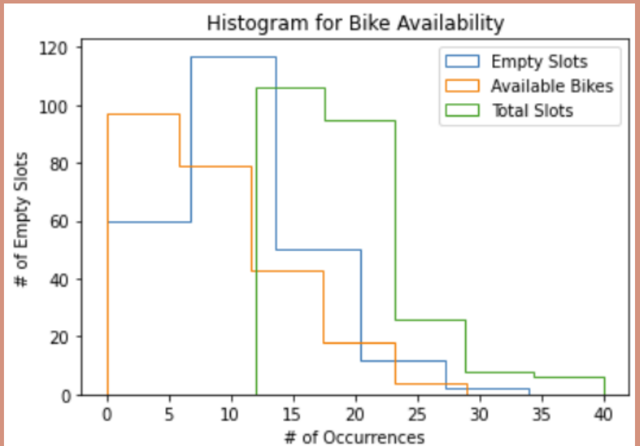
Yelp had more businesses, so used yelp data for further analysis

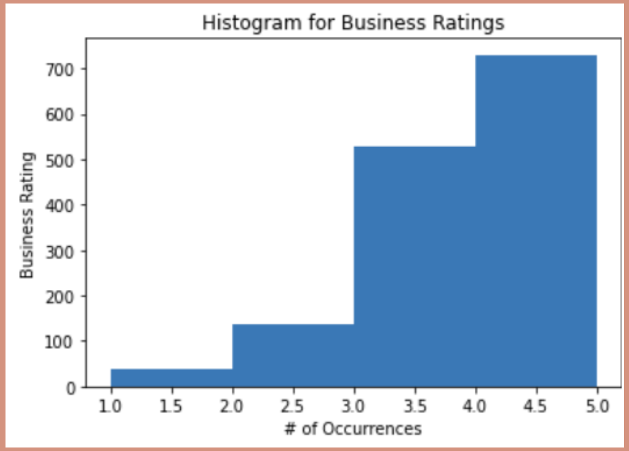
STEP 3: JOINING AND EXPLORING DATA

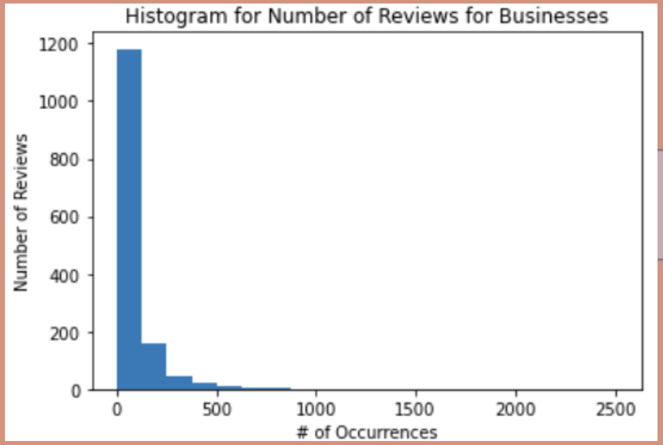
1. Combined Yelp and Citybike data using pd.merge
2. Since there was a ‘many-to-many’ relationship between bike stations and restaurants, used an intermediary table to join them.

STEP 3: EXPLORATORY DATA ANALYSIS

1. Explored data using:
   - histograms



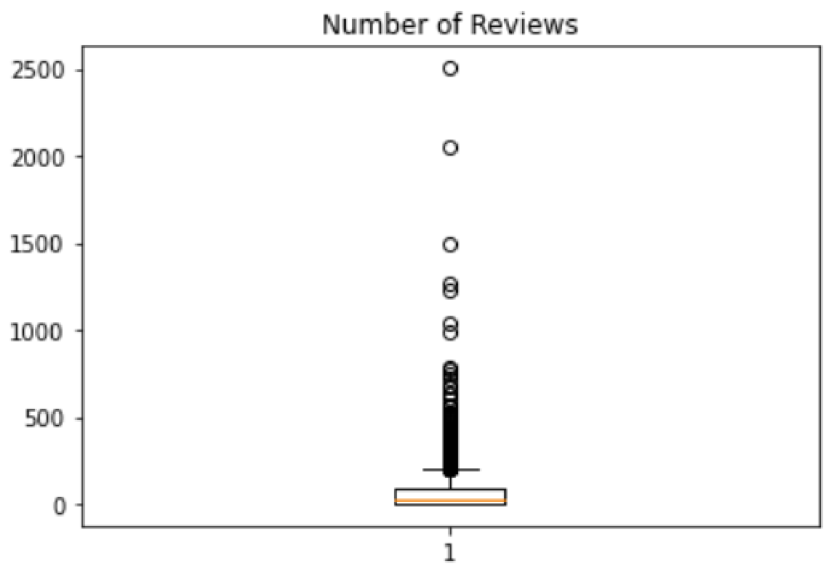




Data not normally distributed

STEP 3: JOINING AND EXPLORING DATA

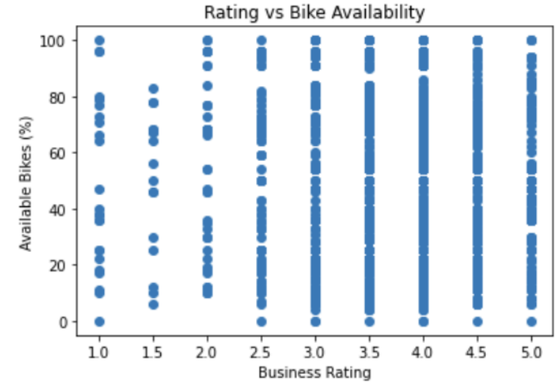
1. Explored data using:
   - boxlots

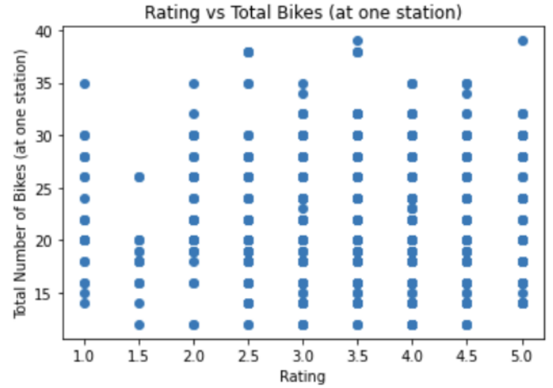


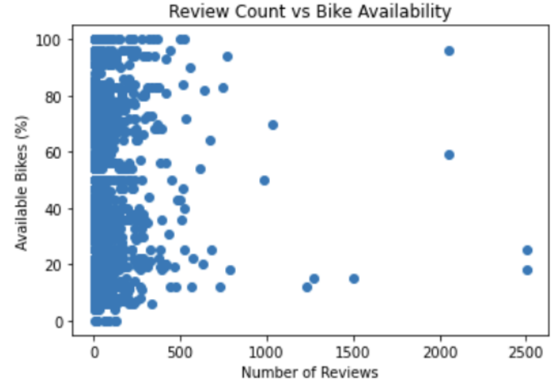
Many outliers!

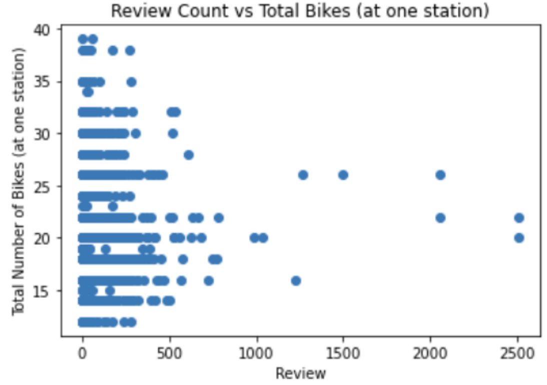
STEP 3: JOINING AND EXPLORING DATA

1. Explored data using:
   - scatter plots





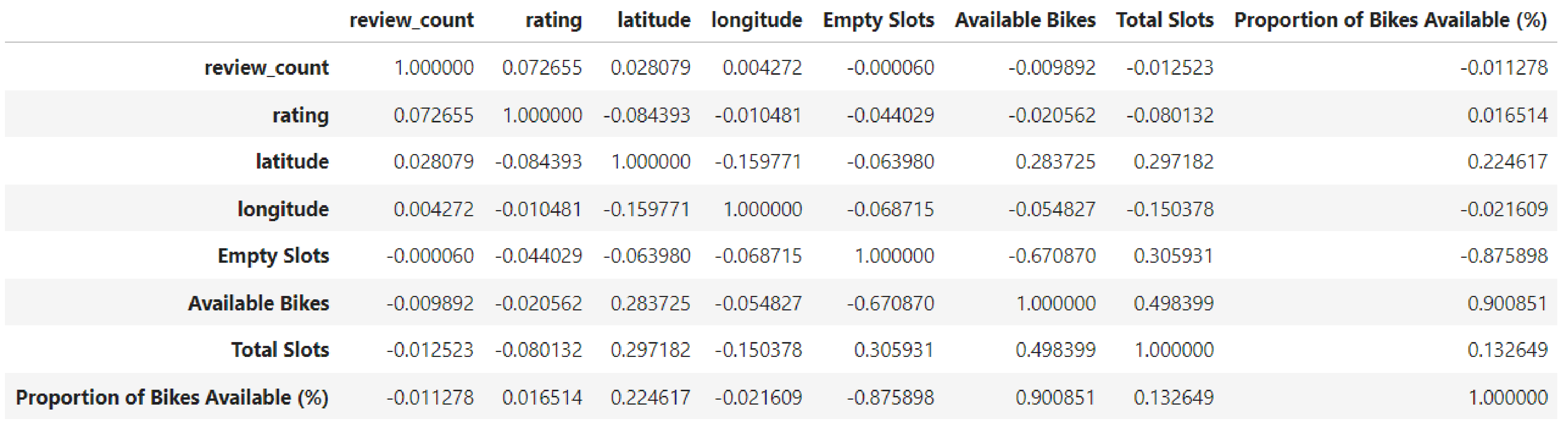




No clear linear relationships

STEP 3: JOINING AND EXPLORING DATA

1. Explored data using:
   - correlation coefficients



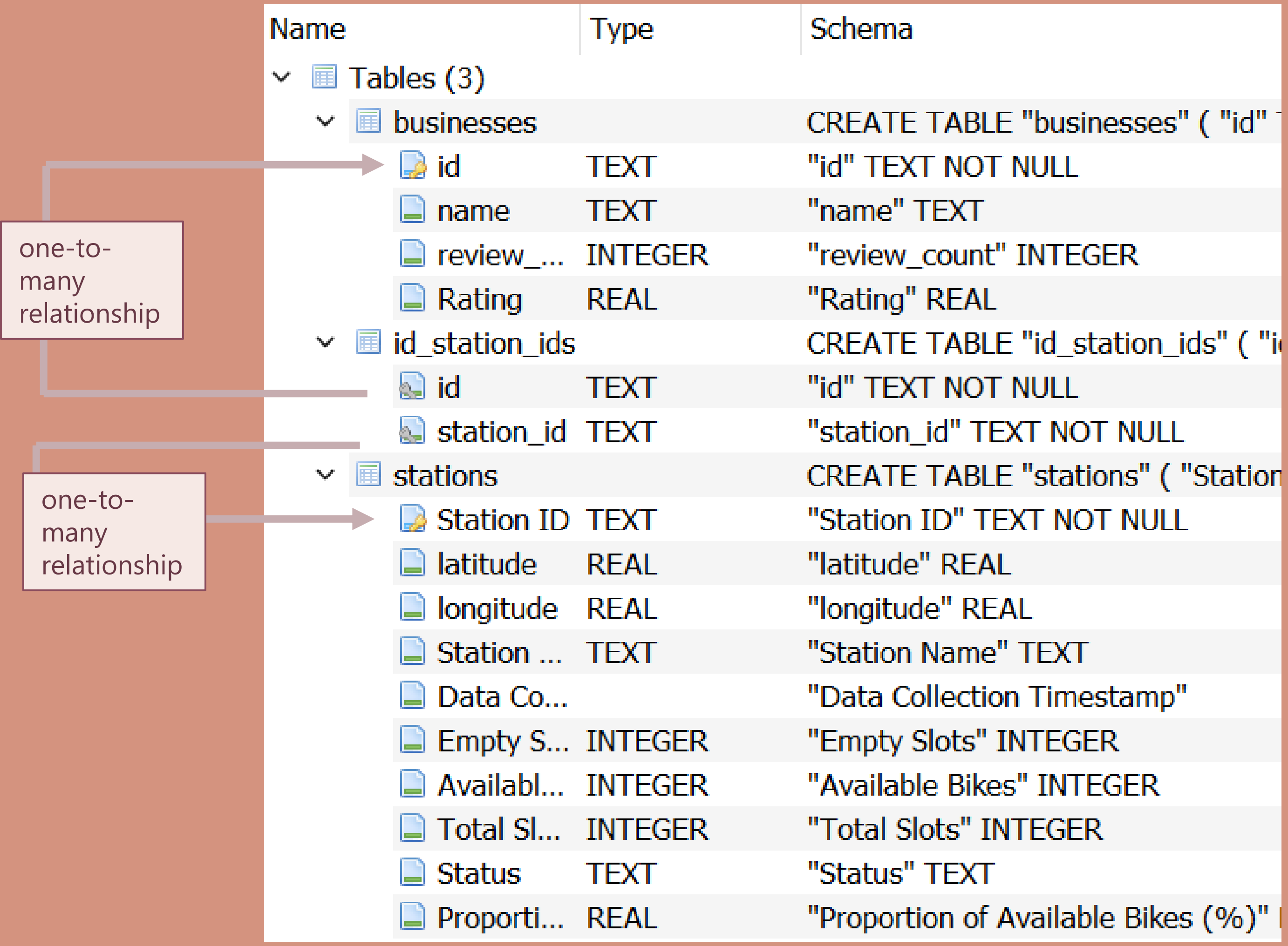
	review_count	rating	latitude	longitude	Empty Slots	Available Bikes	Total Slots	Proportion of Bikes Available (%)
review_count	1.000000	0.072655	0.028079	0.004272	-0.000060	-0.009892	-0.012523	-0.011278
rating	0.072655	1.000000	-0.084393	-0.010481	-0.044029	-0.020562	-0.080132	0.016514
latitude	0.028079	-0.084393	1.000000	-0.159771	-0.063980	0.283725	0.297182	0.224617
longitude	0.004272	-0.010481	-0.159771	1.000000	-0.068715	-0.054827	-0.150378	-0.021609
Empty Slots	-0.000060	-0.044029	-0.063980	-0.068715	1.000000	-0.670870	0.305931	-0.875898
Available Bikes	-0.009892	-0.020562	0.283725	-0.054827	-0.670870	1.000000	0.498399	0.900851
Total Slots	-0.012523	-0.080132	0.297182	-0.150378	0.305931	0.498399	1.000000	0.132649
Proportion of Bikes Available (%)	-0.011278	0.016514	0.224617	-0.021609	-0.875898	0.900851	0.132649	1.000000

Low correlation coefficients!

INITIAL FINDINGS FROM EDA:

1. No correlation between business attributes an proportion of available bikes
2. No correlation between quantity of open businesses and proportion of available bikes

STEP 3: CREATE SQLITE DATABASE



Name	Type	Schema
Tables (3)		
businesses		CREATE TABLE "businesses" ("id"
id	TEXT	"id" TEXT NOT NULL
name	TEXT	"name" TEXT
review_...	INTEGER	"review_count" INTEGER
Rating	REAL	"Rating" REAL
id_station_ids		CREATE TABLE "id_station_ids" ("i
id	TEXT	"id" TEXT NOT NULL
station_id	TEXT	"station_id" TEXT NOT NULL
stations		CREATE TABLE "stations" ("Station
Station ID	TEXT	"Station ID" TEXT NOT NULL
latitude	REAL	"latitude" REAL
longitude	REAL	"longitude" REAL
Station ...	TEXT	"Station Name" TEXT
Data Co...		"Data Collection Timestamp"
Empty S...	INTEGER	"Empty Slots" INTEGER
Availabl...	INTEGER	"Available Bikes" INTEGER
Total Sl...	INTEGER	"Total Slots" INTEGER
Status	TEXT	"Status" TEXT
Proporti...	REAL	"Proportion of Available Bikes (%)"

one-to-many relationship

one-to-many relationship

Businesses Table:

- used 'id' (automatically assigned by Yelp) as primary key

Bike Stations Table:

- Used "Station ID" as primary key

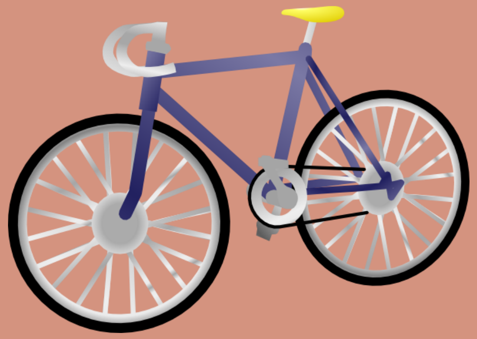
Intermediary Table (called 'id_station_ids':

- id and station_id combined as key
- id was foreign key linked to businesses table
- station_id was foreign key linked to stations table

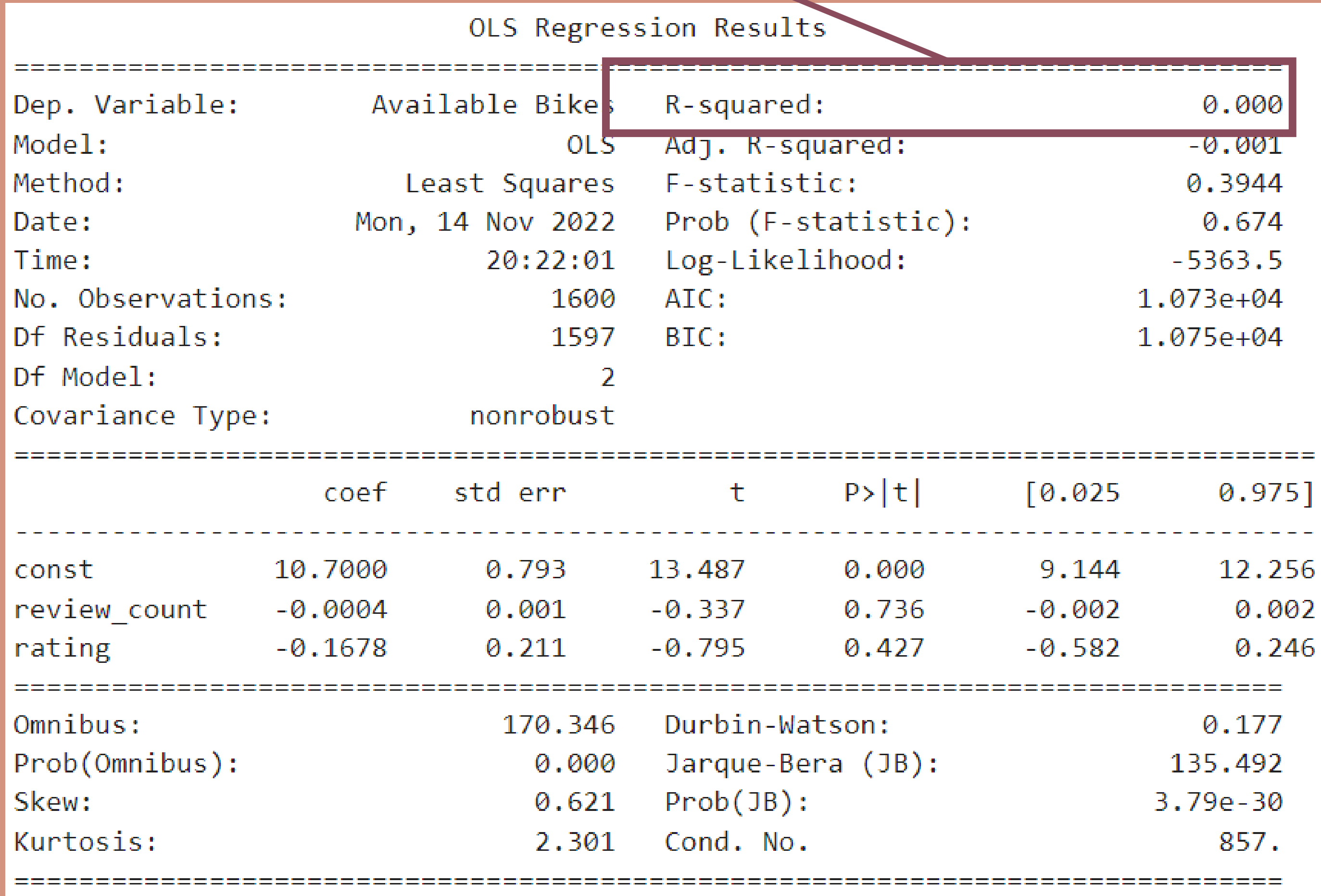
STEP 4: CREATE MODEL

Model 1: number of available bikes as a function of attributes of nearby restaurants and bars?

- Dependent variable is **Available Proportion of Bikes (%)**
- Independent variables (x_1, x_2)
  - review_count
  - ratings



R-squared: 0.000
Model does not predict bike availability whatsoever

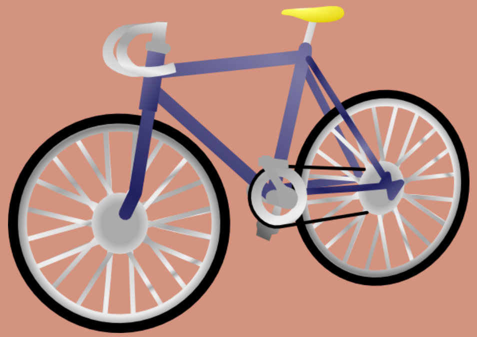


```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Available Bikes   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.3944
Date:                Mon, 14 Nov 2022   Prob (F-statistic):              0.674
Time:                        20:22:01   Log-Likelihood:                -5363.5
No. Observations:                1600   AIC:                         1.073e+04
Df Residuals:                    1597   BIC:                         1.075e+04
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           10.7000      0.793     13.487      0.000       9.144      12.256
review_count    -0.0004      0.001     -0.337      0.736      -0.002       0.002
rating          -0.1678      0.211     -0.795      0.427      -0.582       0.246
==============================================================================
Omnibus:                      170.346   Durbin-Watson:                   0.177
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              135.492
Skew:                           0.621   Prob(JB):                     3.79e-30
Kurtosis:                       2.301   Cond. No.                         857.
==============================================================================
```

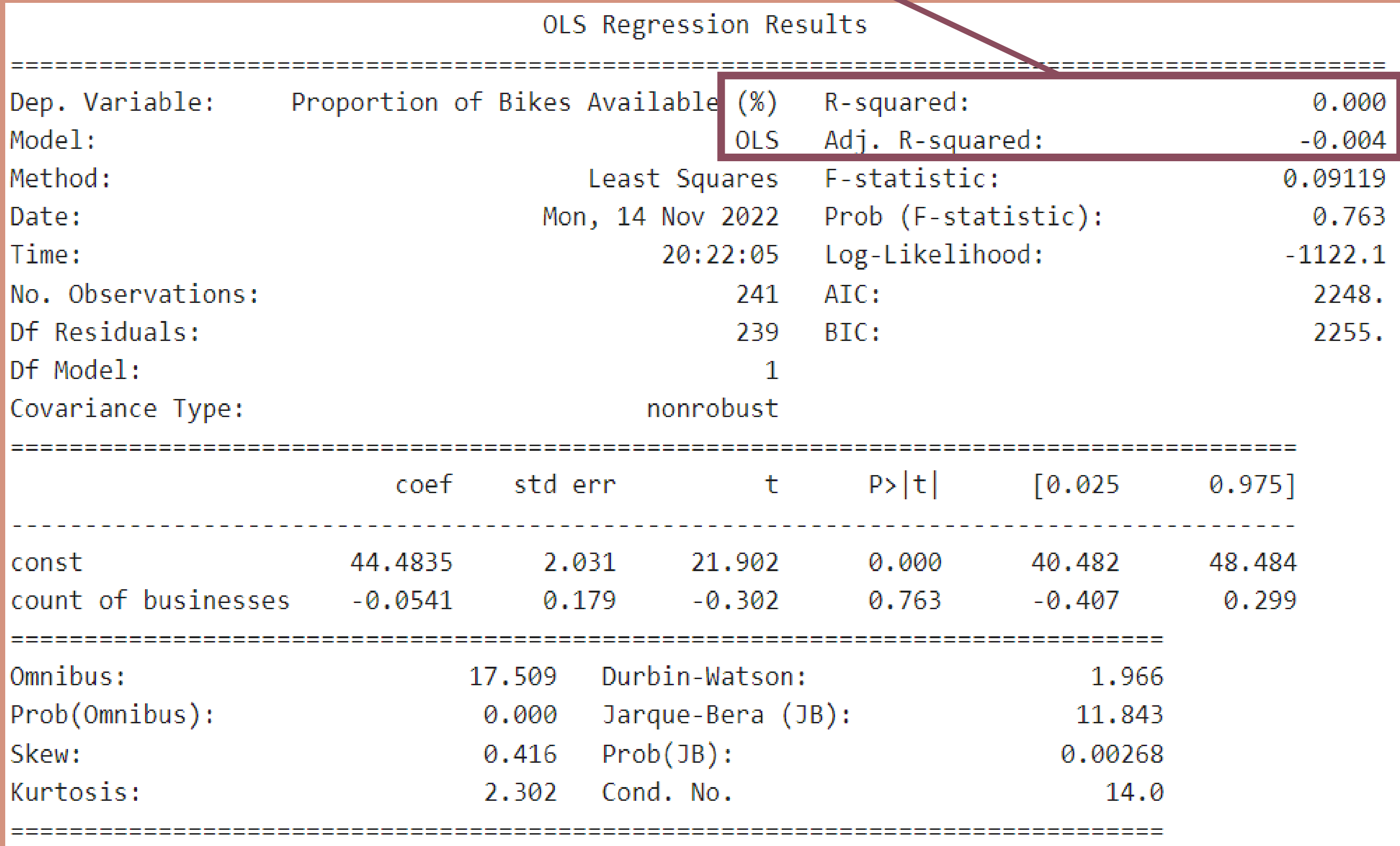
STEP 4: CREATE MODEL

Model 2: number of available bikes as a function of count of nearby restaurants and bars?

- Dependent variable is **Available Proportion of Bikes (%)**
- Independent variables (x_1)
  - number of nearby businesses



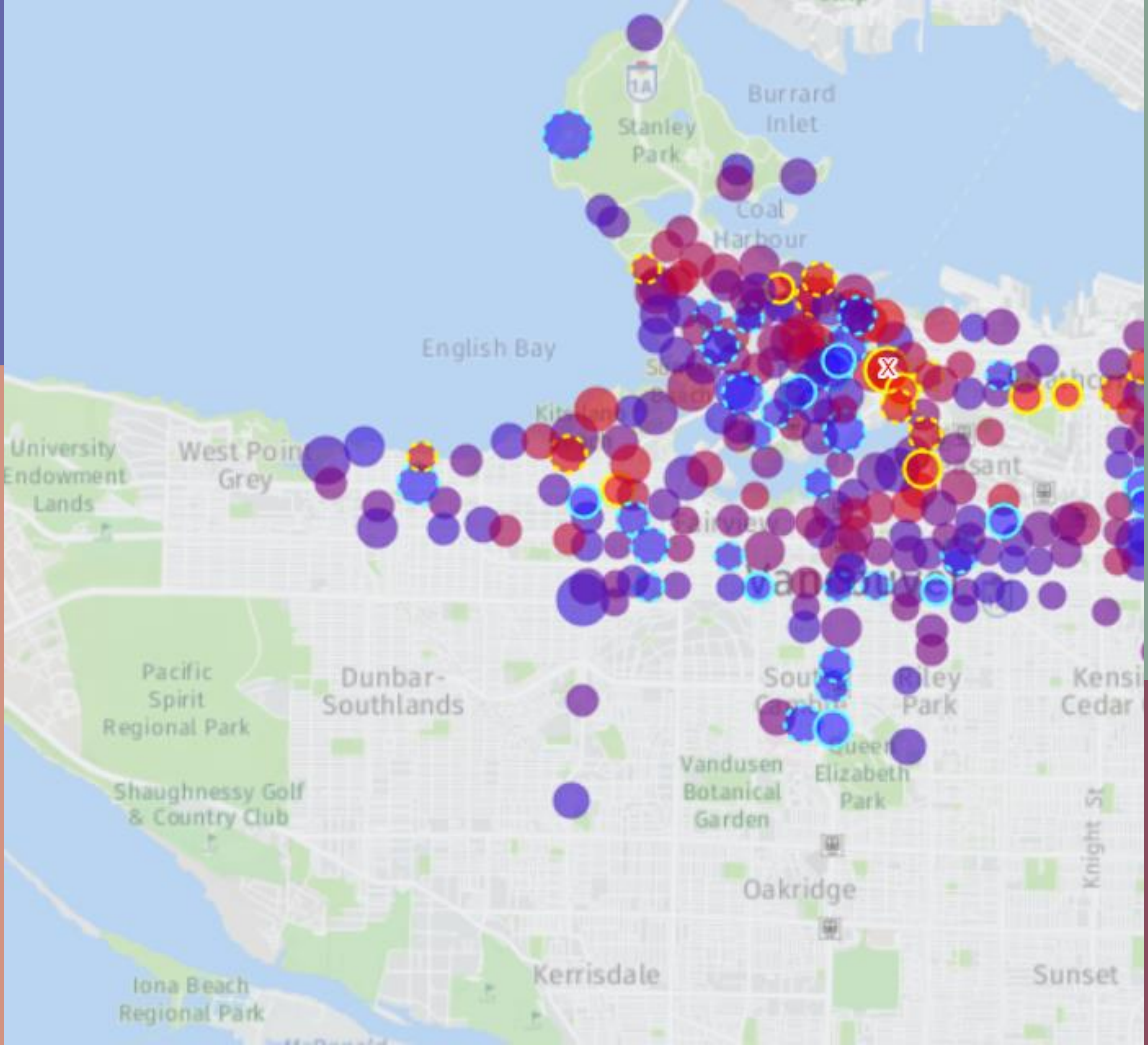
R-squared: 0.000
Model does not predict bike availability whatsoever



```
                                OLS Regression Results
==============================================================================================
Dep. Variable:      Proportion of Bikes Available (%)   R-squared:                       0.000
Model:                                            OLS   Adj. R-squared:                 -0.004
Method:                                 Least Squares   F-statistic:                   0.09119
Date:                                Mon, 14 Nov 2022   Prob (F-statistic):              0.763
Time:                                        20:22:05   Log-Likelihood:                -1122.1
No. Observations:                                 241   AIC:                             2248.
Df Residuals:                                     239   BIC:                             2255.
Df Model:                                           1
Covariance Type:                            nonrobust
=============================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
const                44.4835      2.031     21.902      0.000      40.482      48.484
count of businesses  -0.0541      0.179     -0.302      0.763      -0.407       0.299
==============================================================================
Omnibus:                       17.509   Durbin-Watson:                   1.966
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               11.843
Skew:                           0.416   Prob(JB):                      0.00268
Kurtosis:                       2.302   Cond. No.                         14.0
==============================================================================
```

IF I HAD MORE TIME

- See how open-ness affects availability of bikes by looking at bike availability at different times
- Would look at density of total bikes, not just bikes per station


Burrard Inlet
Stanley Park
Coal Harbour
English Bay
University Endowment Lands
West Point Grey
Pacific Spirit Regional Park
Dunbar-Southlands
Shaughnessy Golf & Country Club
Vandusen Botanical Garden
Queen Elizabeth Park
Riley Park
Kensi Cedar
Oakridge
Knight St
Kerrisdale
Sunset
Iona Beach Regional Park