BIKE STATIONS AND BUSINESSES STATISTICAL MODELLING PROJECT

MELISSA NIELSEN

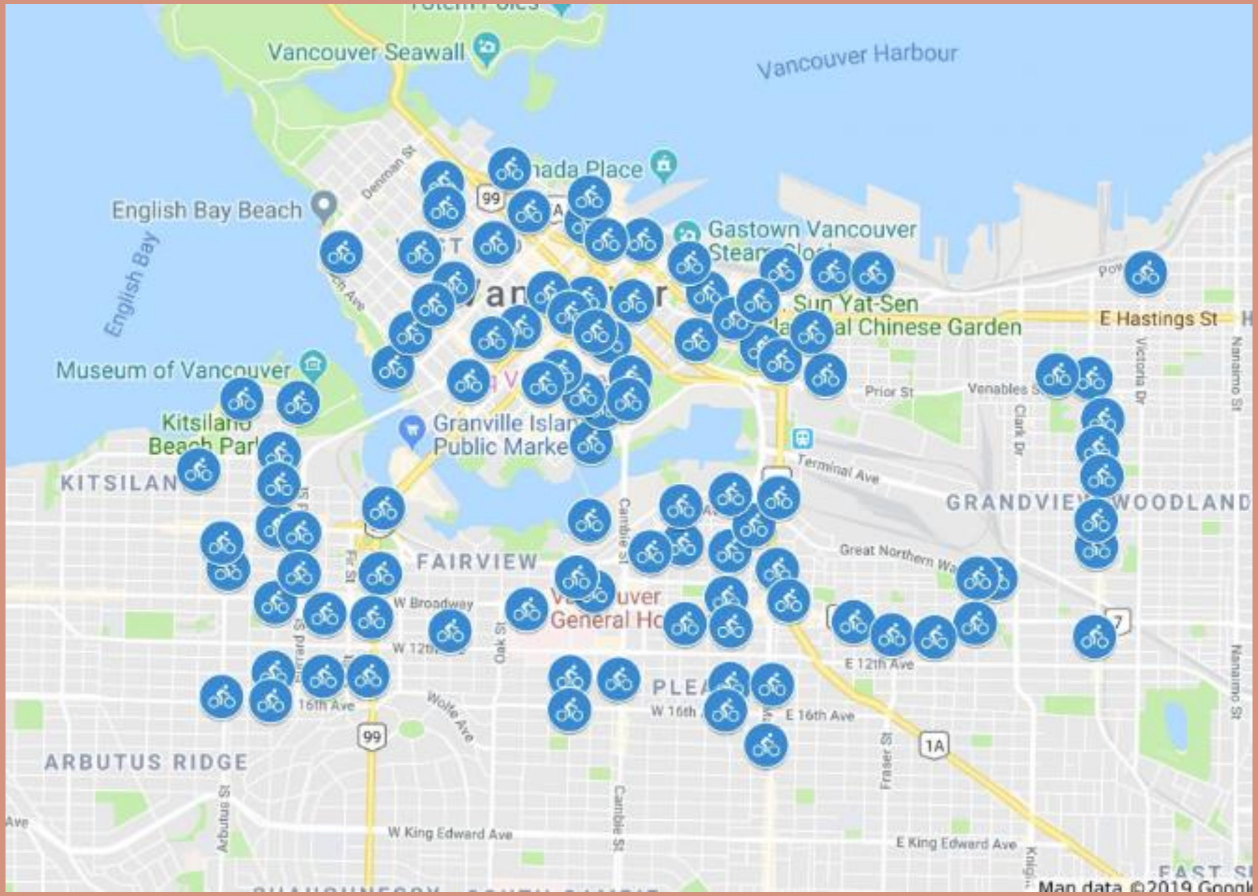
NOVEMBER 7, 2022

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
   - Shopping
   - Education
   - Arts and Entertainment
3. Looked at total numbers of ***open*** and ***closed*** businesses



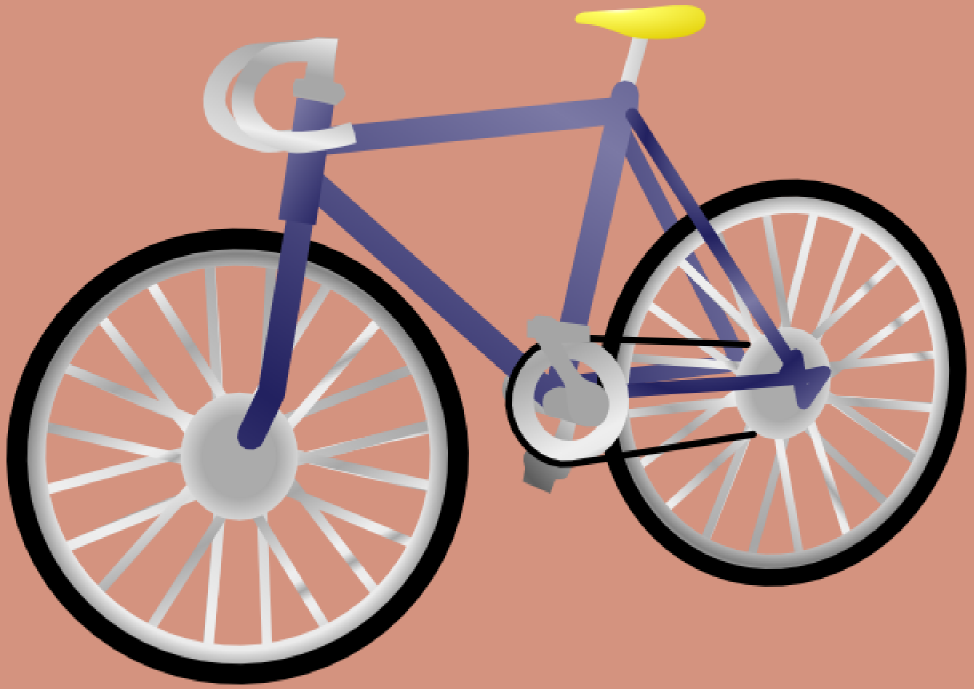
KEY QUESTION

Is the number of open businesses correlated with the proportion of available bikes?

CITY BIKE API

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

There were 240 online bike stations in Vancouver

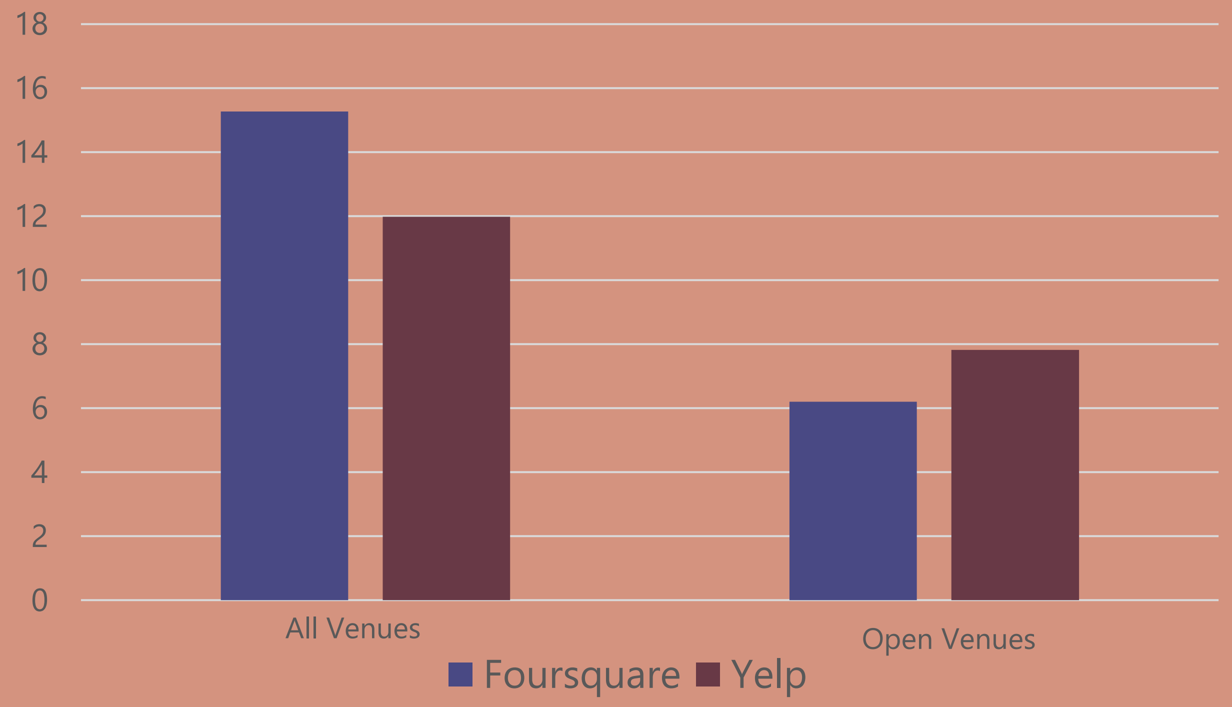


STEP 2: GET DATA FROM FOURSQUARE AND YELP APIS

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

- For each lat/long coordinate:
  - Look at each of the five chosen business types
    - Count all businesses
    - Count open businesses
  - Append results to list
- Reshape list into dataframe and merge dataframe with station ids

STEP 2: COMPARE RESULTS FROM FOURSQUARE AND YELP



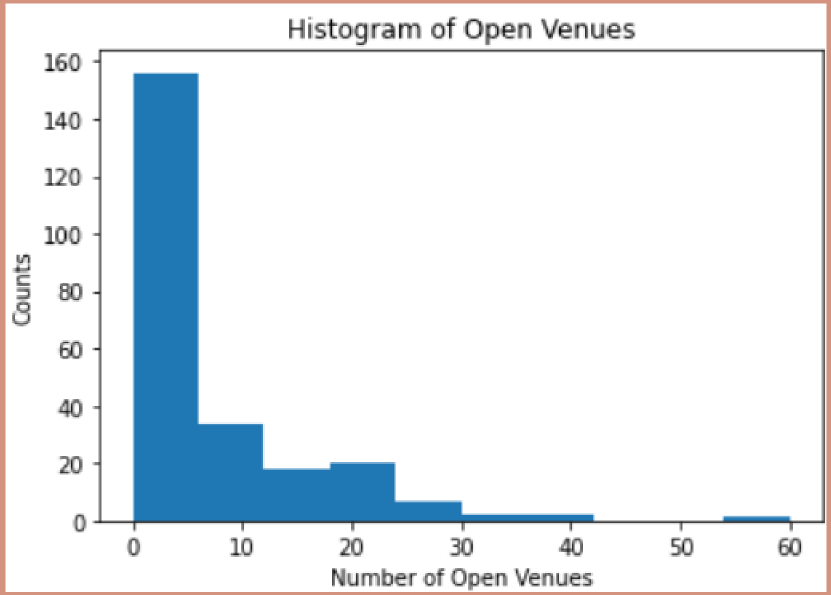
Foursquare had more overall restaurants, so used foursquare data for further analysis

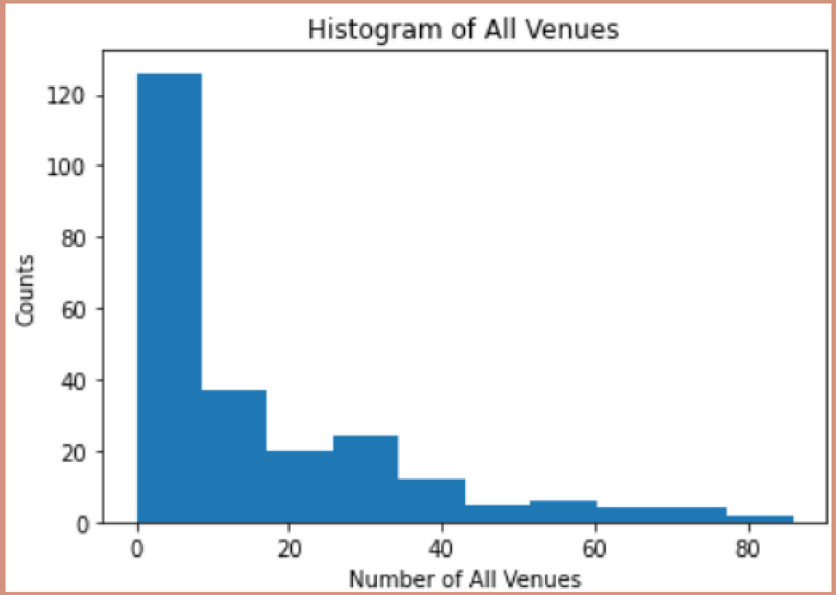
STEP 3: JOINING AND EXPLORING DATA

1. Combined Foursquare and Citybike data using pd.merge and a left join
2. Joined on the key “Station ID”

STEP 3: JOINING AND EXPLORING DATA

1. Explored data using:
   - histograms
   - correlation coefficients
   - scatter plots

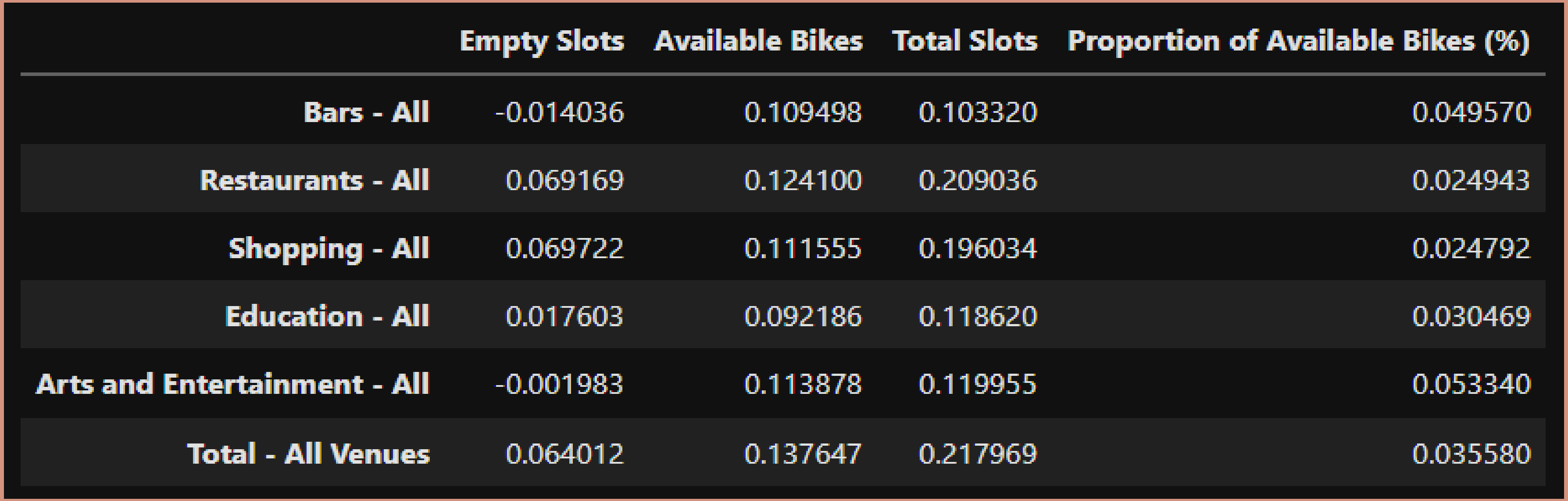




Data not normally distributed

STEP 3: JOINING AND EXPLORING DATA

1. Explored data using:
    - correlation coefficients

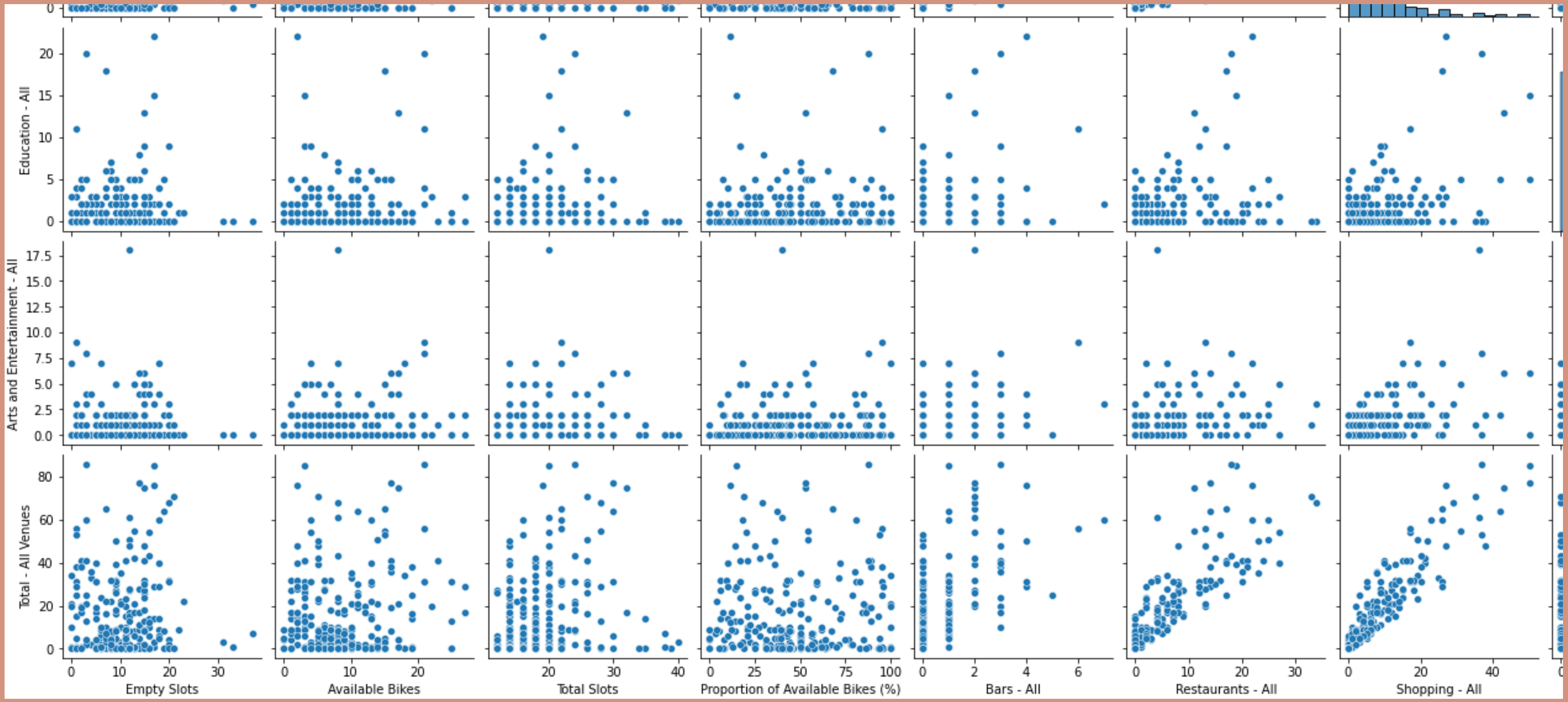


	Empty Slots	Available Bikes	Total Slots	Proportion of Available Bikes (%)
Bars - All	-0.014036	0.109498	0.103320	0.049570
Restaurants - All	0.069169	0.124100	0.209036	0.024943
Shopping - All	0.069722	0.111555	0.196034	0.024792
Education - All	0.017603	0.092186	0.118620	0.030469
Arts and Entertainment - All	-0.001983	0.113878	0.119955	0.053340
Total - All Venues	0.064012	0.137647	0.217969	0.035580

Low correlation coefficients!

STEP 3: JOINING AND EXPLORING DATA

1. Explored data using:
   - scatter plots

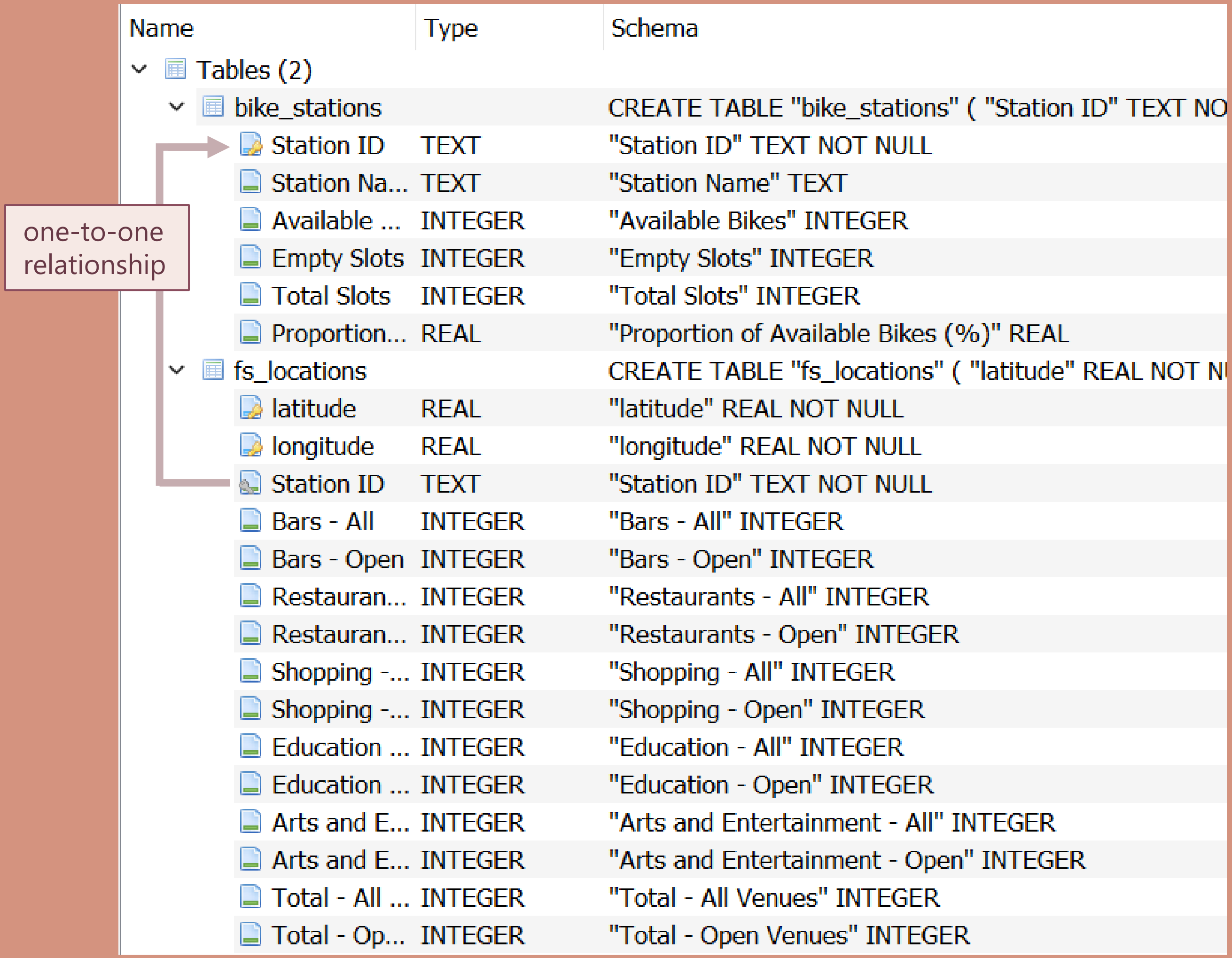


Relationships between bikes and businesses not linear

INITIAL FINDINGS FROM EDA:

1. No correlation between quantity of open businesses and proportion of available bikes
2. Low correlation between total slots and number of total businesses
3. Low correlation between total slots and number of restaurants

STEP 3: CREATE SQLITE DATABASE



Name	Type	Schema
Tables (2)		
bike_stations		CREATE TABLE "bike_stations" ("Station ID" TEXT NO
Station ID	TEXT	"Station ID" TEXT NOT NULL
Station Na...	TEXT	"Station Name" TEXT
Available ...	INTEGER	"Available Bikes" INTEGER
Empty Slots	INTEGER	"Empty Slots" INTEGER
Total Slots	INTEGER	"Total Slots" INTEGER
Proportion...	REAL	"Proportion of Available Bikes (%)" REAL
fs_locations		CREATE TABLE "fs_locations" ("latitude" REAL NOT N
latitude	REAL	"latitude" REAL NOT NULL
longitude	REAL	"longitude" REAL NOT NULL
Station ID	TEXT	"Station ID" TEXT NOT NULL
Bars - All	INTEGER	"Bars - All" INTEGER
Bars - Open	INTEGER	"Bars - Open" INTEGER
Restauran...	INTEGER	"Restaurants - All" INTEGER
Restauran...	INTEGER	"Restaurants - Open" INTEGER
Shopping -...	INTEGER	"Shopping - All" INTEGER
Shopping -...	INTEGER	"Shopping - Open" INTEGER
Education ...	INTEGER	"Education - All" INTEGER
Education ...	INTEGER	"Education - Open" INTEGER
Arts and E...	INTEGER	"Arts and Entertainment - All" INTEGER
Arts and E...	INTEGER	"Arts and Entertainment - Open" INTEGER
Total - All ...	INTEGER	"Total - All Venues" INTEGER
Total - Op...	INTEGER	"Total - Open Venues" INTEGER

one-to-one relationship

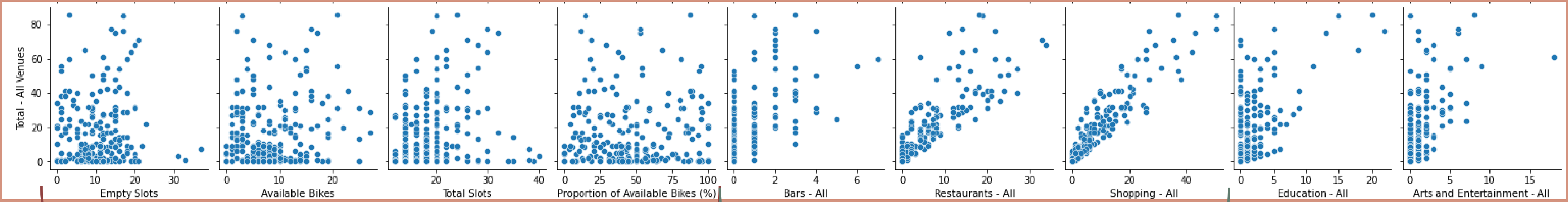
Bike Stations Table:

- Used “Station ID” as primary key

Foursquare Locations Table:

- Used “latitude” and “longitude” (combined) as primary key
- Used “Station ID” as foreign key

STEP 4: CREATE MODEL



No linearity found between bikes and businesses

Linearity found between:

- Number of restaurants and total number of businesses
- Number of bars and total number of businesses

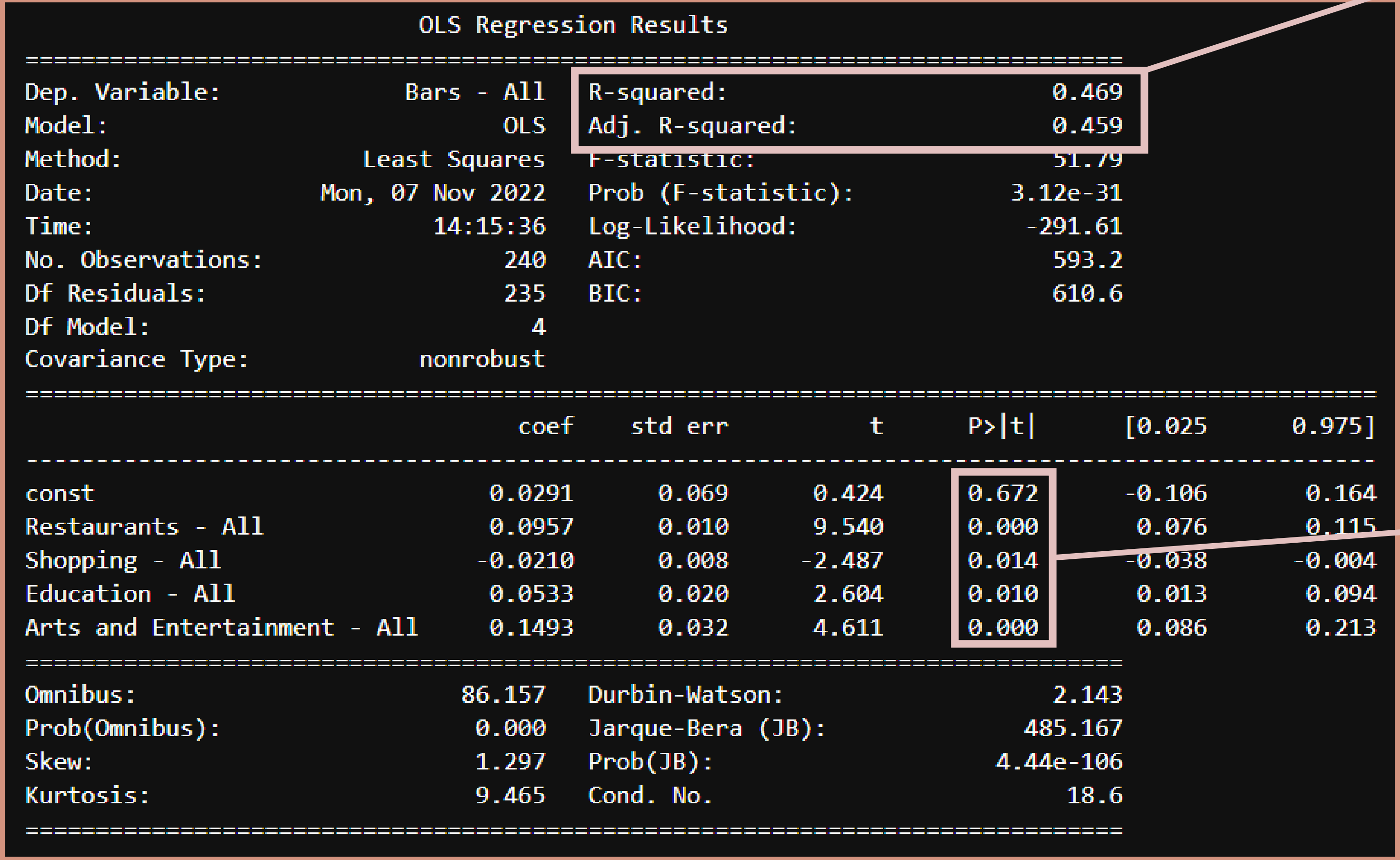
STEP 4: CREATE MODEL

How many bars in an area based on number of other business types?

- Dependent variable is **number of bars (y)**
- Independent variables:
  - number of restaurants
  - number of stores
  - number of arts and entertainment businesses
  - number of education businesses

STEP 4: CREATE MODEL

Model results:



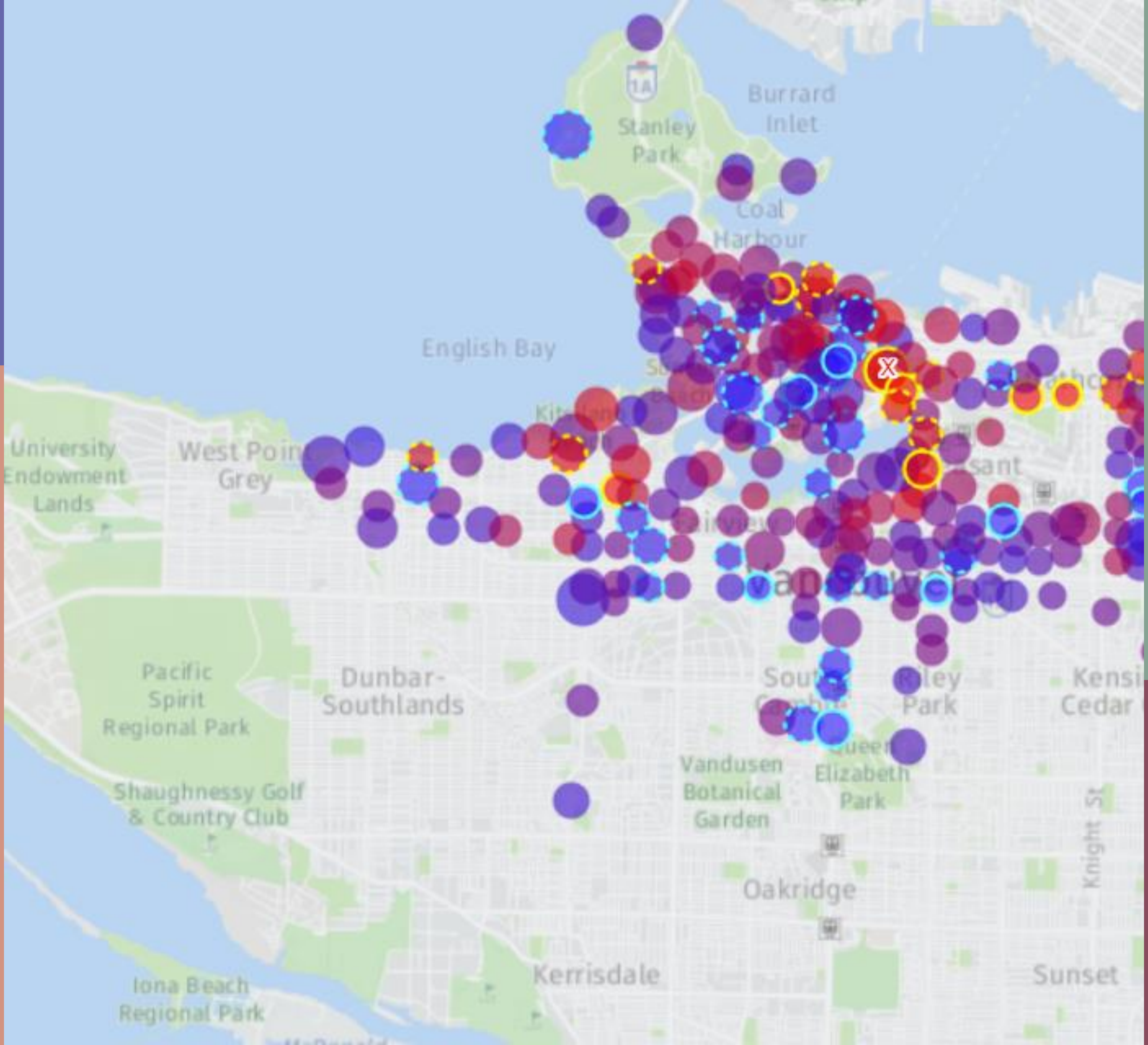
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Bars - All   R-squared:                       0.469
Model:                           OLS   Adj. R-squared:                  0.459
Method:                Least Squares   F-statistic:                     51.79
Date:               Mon, 07 Nov 2022   Prob (F-statistic):           3.12e-31
Time:                       14:15:36   Log-Likelihood:                -291.61
No. Observations:                240   AIC:                             593.2
Df Residuals:                    235   BIC:                             610.6
Df Model:                          4
Covariance Type:           nonrobust
===============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          0.0291      0.069      0.424      0.672      -0.106       0.164
Restaurants - All              0.0957      0.010      9.540      0.000       0.076       0.115
Shopping - All                -0.0210      0.008     -2.487      0.014      -0.038      -0.004
Education - All                0.0533      0.020      2.604      0.010       0.013       0.094
Arts and Entertainment - All   0.1493      0.032      4.611      0.000       0.086       0.213
==============================================================================
Omnibus:                      86.157   Durbin-Watson:                   2.143
Prob(Omnibus):                 0.000   Jarque-Bera (JB):              485.167
Skew:                          1.297   Prob(JB):                     4.44e-106
Kurtosis:                      9.465   Cond. No.                         18.6
==============================================================================
```

R-squared: 0.469
R-squared: 0.45
...not a great model

p-values: under 0.05 so kept all independent variables

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
1A
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
Kerrisdale
Sunset
Knight St
Iona Beach Regional Park