Statistical Modelling of City Bikes and Points of Interest (POIs)

Approach and findings

1) Goals
2) Process
3) Findings
4) Challenges
5) Future Goals

Goals

- The project aim is to statistically model a relationship (if any) between city bike availability and Points of Interest (POIs) in the area.
- Join bike and venue data, explore any relationships and model these with linear regression.

- City chosen: Toronto, Ontario
- Tools:
    - City Bike API (bike station data)
    - Foursquare API (venue/location data)
    - Yelp API (venue/location data)

- See city_bikes.ipynb

Process

- Foursquare's API used to obtain various places of interest in 800m (city density and constraints from API) of a city bike station

- Began with a sample request of one bike station latitude/longitude location and manipulating the data returned by Foursquare
    - Number of bars/restaurants, parks, live venues and cafes, total number of POIs.
    - Sampled the Foursquare API calls on bike different stations to build a catch-all word search (Regex).

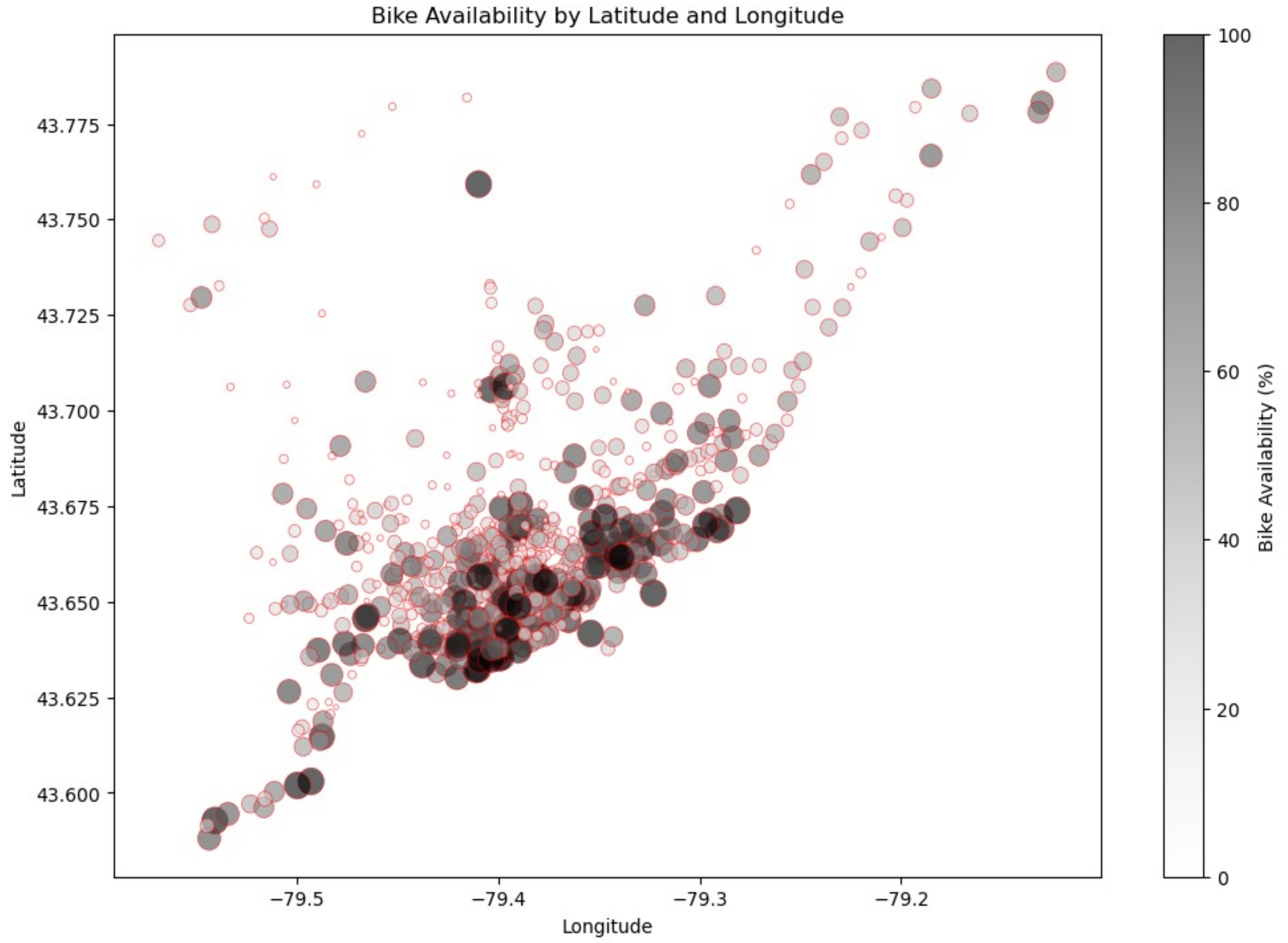
Process

- For each of the 826 stations, Foursquare obtained up to 50 POIs in the bar/restaurant, park, cafe and live venue categories. Build the model and perform a linear regression later.
- The individual venue data such as name, address, venue category for saving into a SQL .db file, as with the city bikes stations.
- We now have a database holding tens of thousands venues and hundreds of bike stations, we can search by ll, search by bike station, search by venue where there is a bike station nearby.
- EDA to find correlations between number of POIs, bike availability, correlation heatmaps and scatter graphs
- See joining_data.ipynb.

Findings

- -0.51 corr between latitude and n_POIs
- -0.31 corr between latitude and bike availability (image on right)
- 0.17 corr between n_POIs and bike availability

All three of the above are statistically significant ($p < 0.05$)



Findings

Simple linear regression:

- Adjusted R^2 of 0.096 for bike availability / latitude. Latitude only explains 9.6% of the variance
- Adjusted R^2 of 0.27 for n_POIs / latitude. Latitude only explains 27% of the variance of n_POIs. I expected higher because south is downtown, more densely packed than northern Toronto.
- Multilinear regression: Adj. R^2 of 0.096 – predicting bike availability against latitude and n_POIs

Conclusion: Latitude and n_POIs do contribute to bike availability but they are by no means the only factors.

model_building.ipynb

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      bike_availability   R-squared:                       0.097
Model:                            OLS   Adj. R-squared:                  0.096
Method:                 Least Squares   F-statistic:                     88.47
Date:                Thu, 08 Aug 2024   Prob (F-statistic):           4.98e-20
Time:                        17:31:55   Log-Likelihood:                -3928.8
No. Observations:                 826   AIC:                             7862.
Df Residuals:                     824   BIC:                             7871.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.196e+04   1267.362      9.434      0.000    9468.269    1.44e+04
latitude    -272.9633     29.020     -9.406      0.000    -329.926    -216.001
==============================================================================
Omnibus:                       80.969   Durbin-Watson:                   1.920
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               57.577
Skew:                           0.538   Prob(JB):                     3.14e-13
Kurtosis:                       2.282   Cond. No.                     5.65e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.65e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Challenges

- Avoiding multicollinearity in this is very difficult with n_POIs and bike availability/latitude

- API is damaging the power of the n_POIs data, need > 50 returns. This impacted the radius that I was calling as well

Future Goals

- Run the model on other categories of POI, or only with the specific categories. This would still have problems with collinearity though (in >= 50 POIs, more bars will mean less cafes are pulled)
- Run the Monday afternoon data instead of Saturday afternoon for different findings?
- Classification problem – categorize high and low density / high and low traffic stations and try and predict using other features in the data.
- Use the n_POIs, bike availability and other features (Yelp distance from venue to bike station) to identify stations that are underserving the area – build more bike stations?