Statistical Modeling with Python

Korinne Angela White

Introduction

City of Choice:
Hamilton, ON

Bike Stations
190

Points of Interest (POIs):
- Restaurant
- Amusement Parks
- Museums

Project Flow Structure

1. Problem Definition
2. Data Collection
3. Data Cleaning and Integration
4. Exploratory Data Analysis
5. Model Building

Results

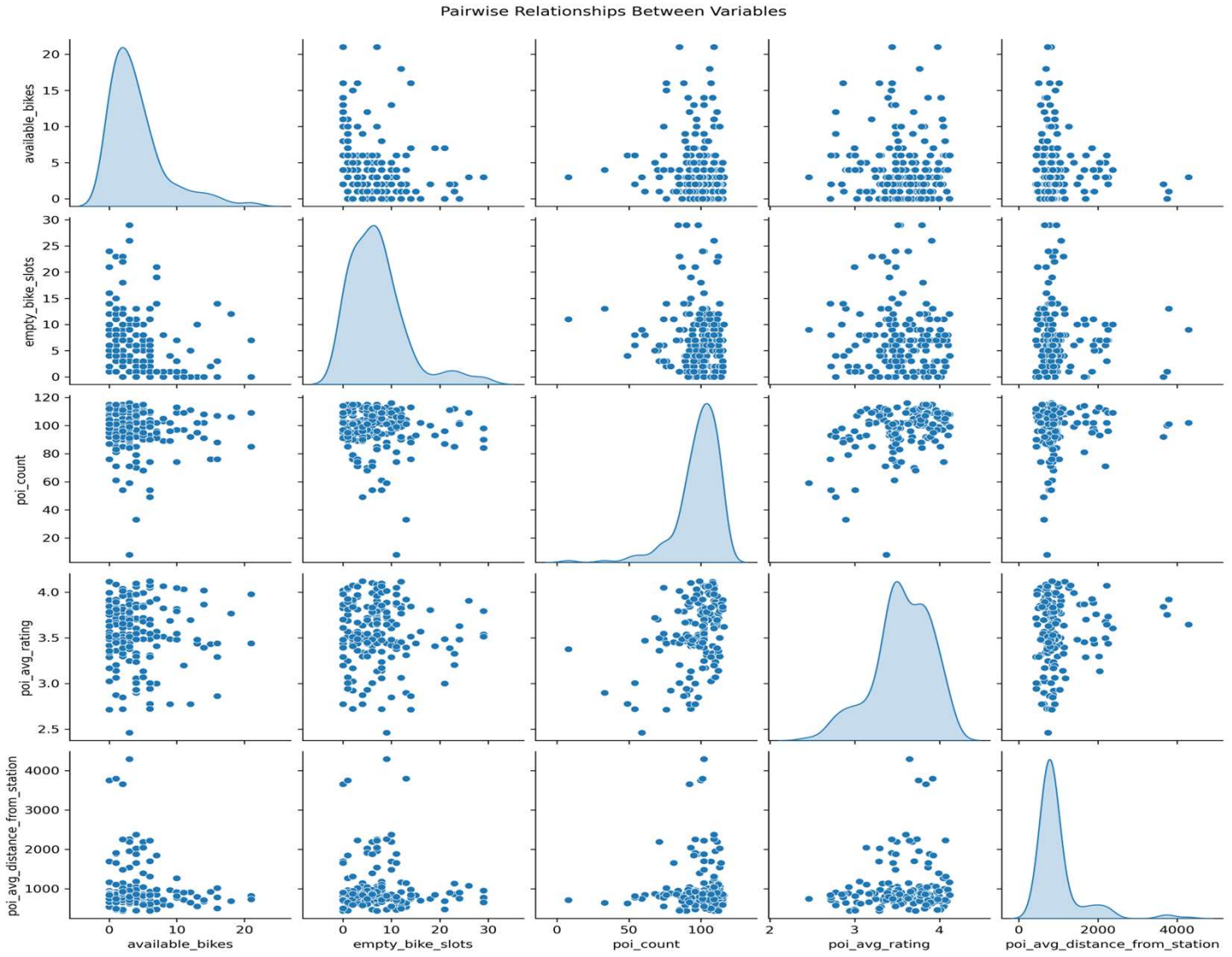
FOURSQUARE

- More Unique POIs
  (586 vs Yelp has 485)

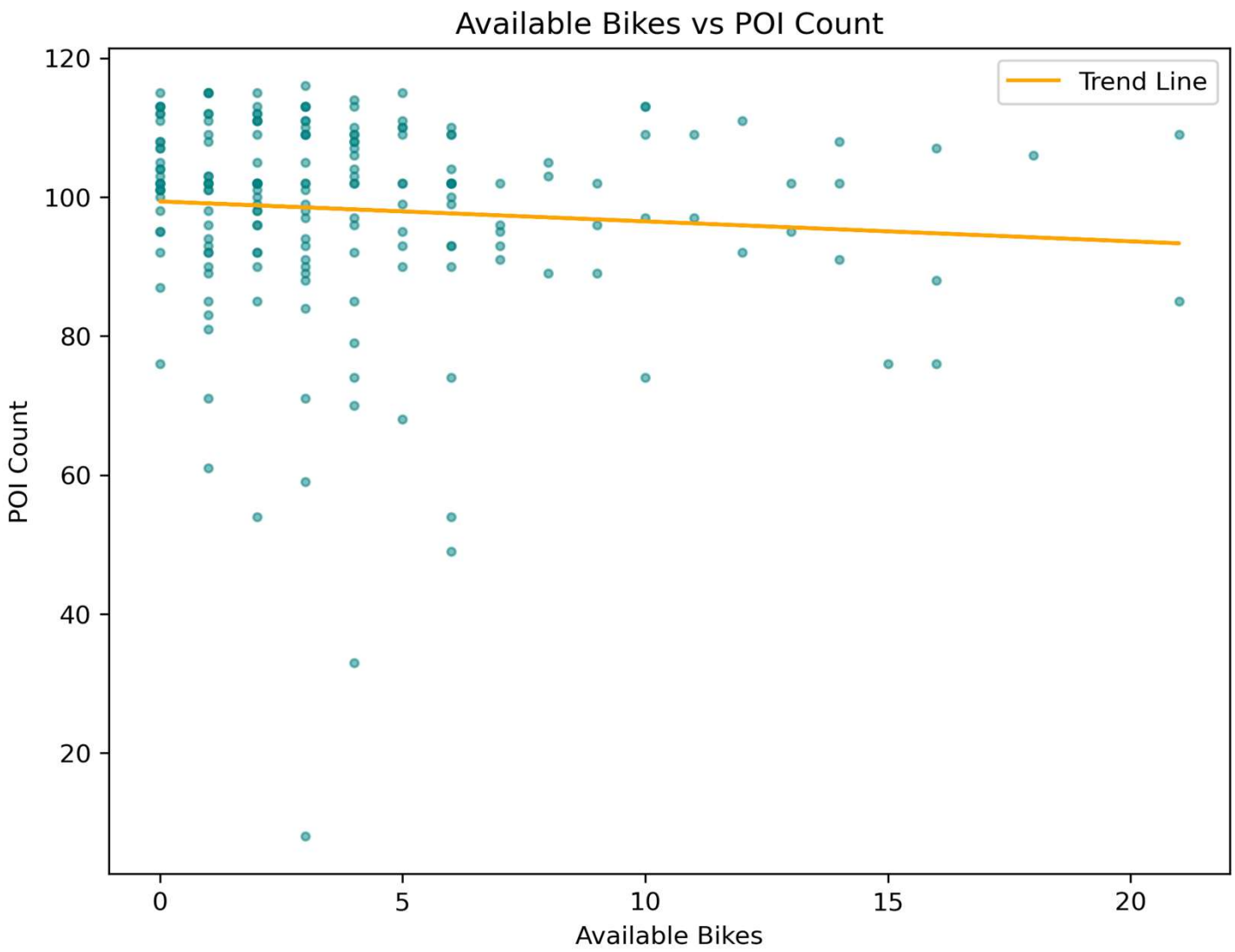
YELP

- Provides ratings and reviews
- Higher precision in distances

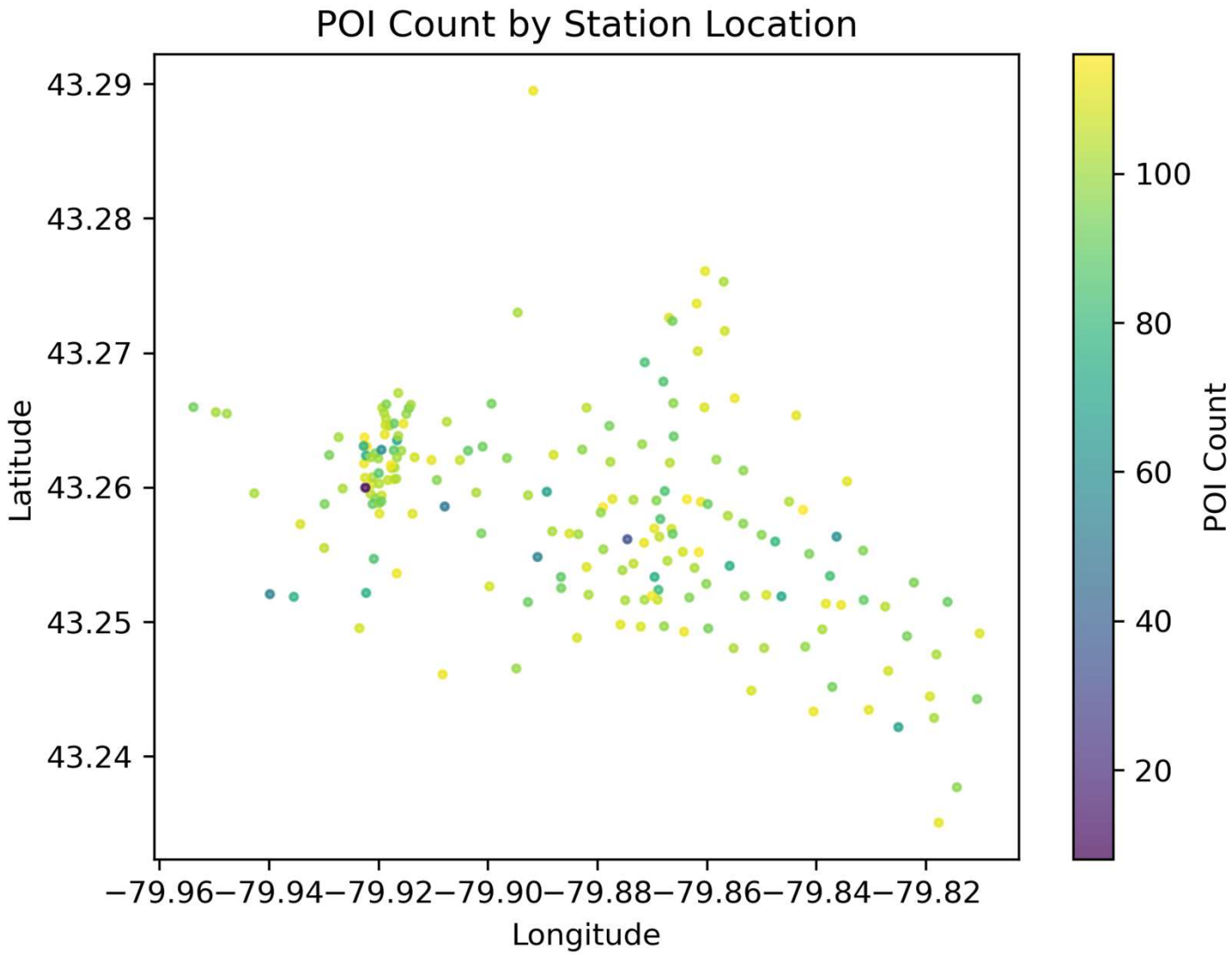
Results


Pairwise Relationships Between Variables
available_bikes
empty_bike_slots
poi_count
poi_avg_rating
poi_avg_distance_from_station

Results


Available Bikes vs POI Count
Trend Line
POI Count
Available Bikes
120
100
80
60
40
20
0
5
10
15
20

Results


POI Count by Station Location
Latitude
43.29
43.28
43.27
43.26
43.25
43.24
−79.96
−79.94
−79.92
−79.90
−79.88
−79.86
−79.84
−79.82
Longitude
POI Count
100
80
60
40
20

Results

OLS Regression Results

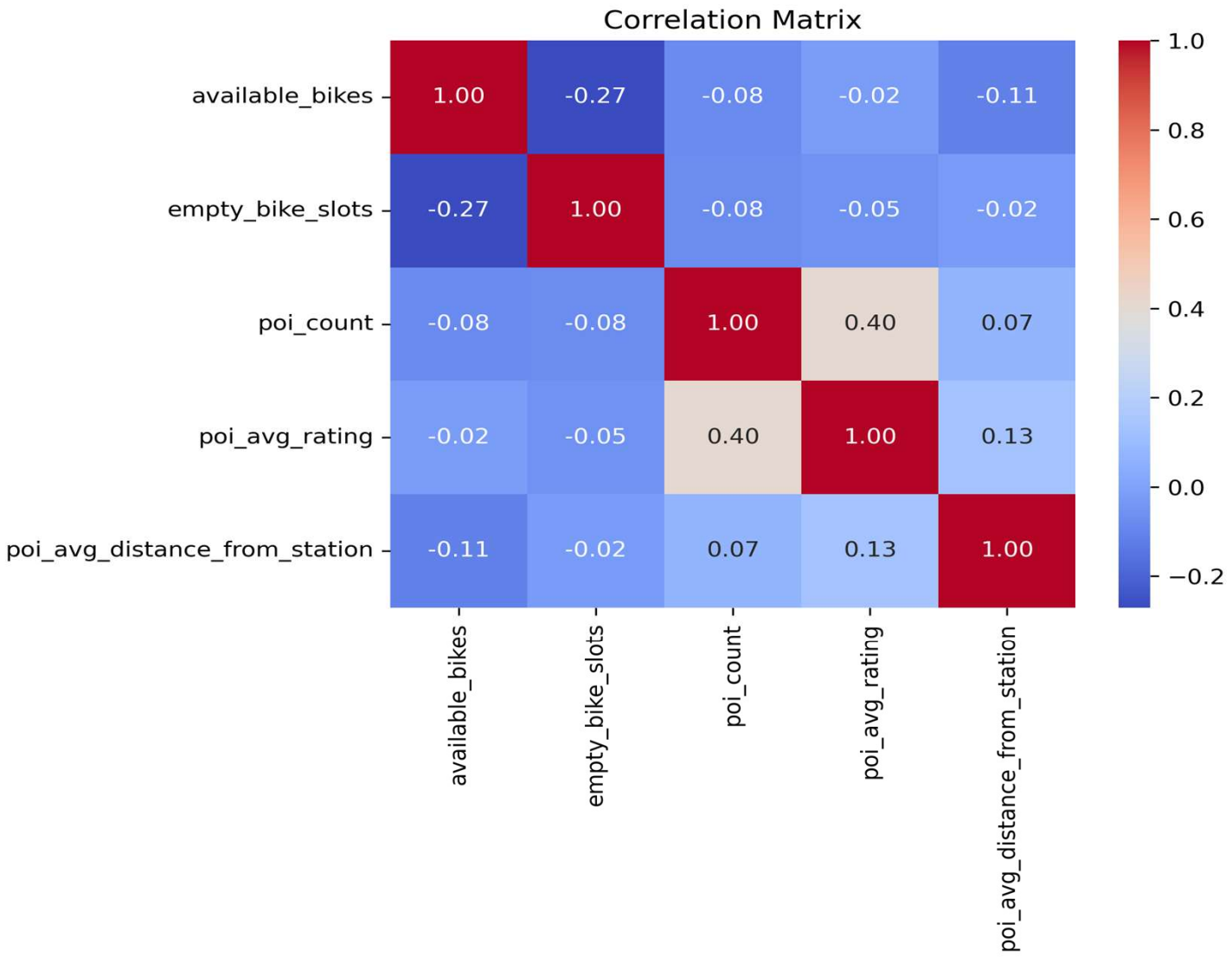
Dep. Variable:	poi_count	R-squared:	0.175
Model:	OLS	Adj. R-squared:	0.157
Method:	Least Squares	F-statistic:	9.783
Date:	Sun, 15 Dec 2024	Prob (F-statistic):	3.36e-07
Time:	19:29:25	Log-Likelihood:	-765.83
No. Observations:	190	AIC:	1542.
Df Residuals:	185	BIC:	1558.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	37.7003	10.957	3.441	0.001	16.084	59.317
available_bikes	-0.3312	0.253	-1.311	0.191	-0.830	0.167
empty_bike_slots	-0.2006	0.176	-1.141	0.255	-0.547	0.146
poi_avg_rating	17.7282	3.011	5.888	0.000	11.788	23.668
poi_avg_distance_from_station	0.0001	0.002	0.079	0.937	-0.003	0.004

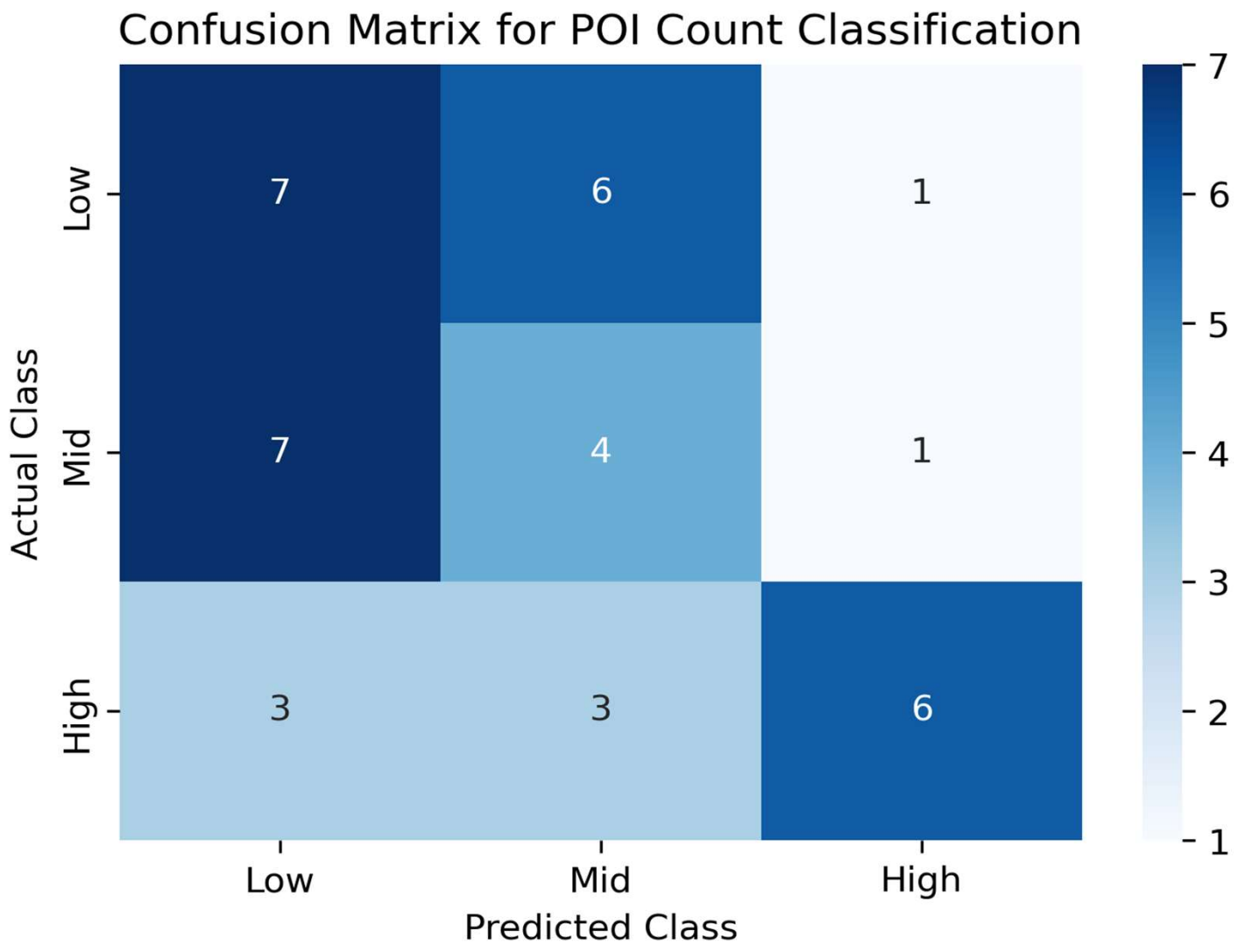
Omnibus:	110.265	Durbin-Watson:	1.803
Prob(Omnibus):	0.000	Jarque-Bera (JB):	741.997
Skew:	-2.142	Prob(JB):	7.54e-162
Kurtosis:	11.681	Cond. No.	1.29e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.29e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Results


Correlation Matrix
available_bikes
empty_bike_slots
poi_count
poi_avg_rating
poi_avg_distance_from_station
1.00 -0.27 -0.08 -0.02 -0.11
-0.27 1.00 -0.08 -0.05 -0.02
-0.08 -0.08 1.00 0.40 0.07
-0.02 -0.05 0.40 1.00 0.13
-0.11 -0.02 0.07 0.13 1.00
available_bikes
empty_bike_slots
poi_count
poi_avg_rating
poi_avg_distance_from_station
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Results


Confusion Matrix for POI Count Classification
Actual Class
Low
Mid
High
7
6
1
7
4
1
3
3
6
Low
Mid
High
Predicted Class
7
6
5
4
3
2
1

Conclusion

Challenges

1. API Call Limitations and Time Constraints
2. Data Iterations and Integration

Future Goals

1. Expand to Multiple Cities
    - Comparative Analysis
    - Comprehensive POI Lists
2. Temporal Analysis
    - Bike Availability Patterns
    - POI Activity Correlation

End.