Statistical Modelling with Python

Nickolas Stello

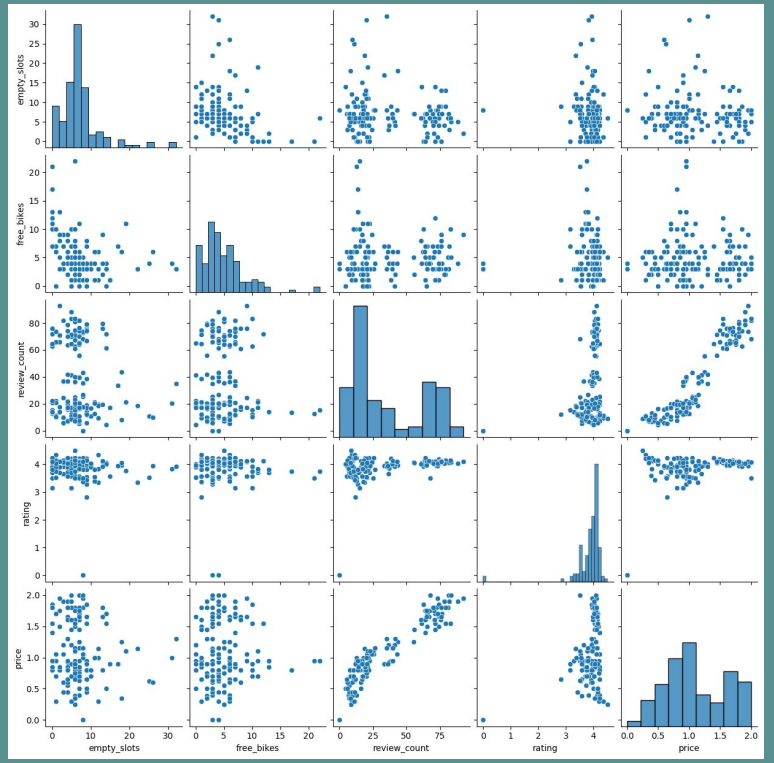
Project Flow Structure

- Connected to the Citybikes API
- Connected to Foursquare API and Yelp API
- Analyzed the results from both APIs
- Decided to use just the Yelp data for the analysis.
- Joined the data from Citybikes with the data from Yelp
- Explore the data
- Created a SQLite database to store the data collected
- Built a Multivariate Linear Regression
- Interpreted results and derived insights from the model.

Multivariate Linear Regression Model

Multivariate Linear Regression was selected as the method to illustrate a potential relationship between the number of bikes in a specific location and the characteristics of points of interest in that location, including review_count, rating, and price.

However, the outcomes from the model did not yield significant insights. This could be attributed to the dataset's nature, suggesting a lack of correlation among the numerical variables.



Model Output

OLS Regression Results

Dep. Variable:	free_bikes	R-squared:	0.027
Model:	OLS	Adj. R-squared:	0.006
Method:	Least Squares	F-statistic:	1.309
Date:	Wed, 06 Dec 2023	Prob (F-statistic):	0.274
Time:	10:53:52	Log-Likelihood:	-388.18
No. Observations:	145	AIC:	784.4
Df Residuals:	141	BIC:	796.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.6021	2.332	1.545	0.125	-1.007	8.212
review_count	-0.0606	0.032	-1.901	0.059	-0.124	0.002
rating	-0.0524	0.593	-0.088	0.930	-1.225	1.120
price	3.4168	1.733	1.972	0.051	-0.009	6.842

Omnibus:	60.031	Durbin-Watson:	1.705
Prob(Omnibus):	0.000	Jarque-Bera (JB):	187.630
Skew:	1.605	Prob(JB):	1.81e-41
Kurtosis:	7.555	Cond. No.	379.

Model Output

- **Adj. R-squared:** The multivariate model explains only 0.6% of the variance in the data, suggesting it is not a good fit.
- **Prob (F-statistic):** The p-value for the hypothesis test is greater than 0, indicating that the independent variables do not significantly impact the dependent variable.
- **coef:** The average POI price has the strongest positive impact on the number of bikes per station, while review_count has the largest negative impact.
- **P>|t|:** All p-values are >0.05, indicating that the review_count, rating, and price attributes of a point of interest do not significantly affect the number of bikes in a bike station.

Challenges

Interpreting and deriving insights from the model proved challenging due to the lack of interconnectedness within the data.

Thank you!

 github.com/nickolas-s

 linkedin.com/in/nickstello