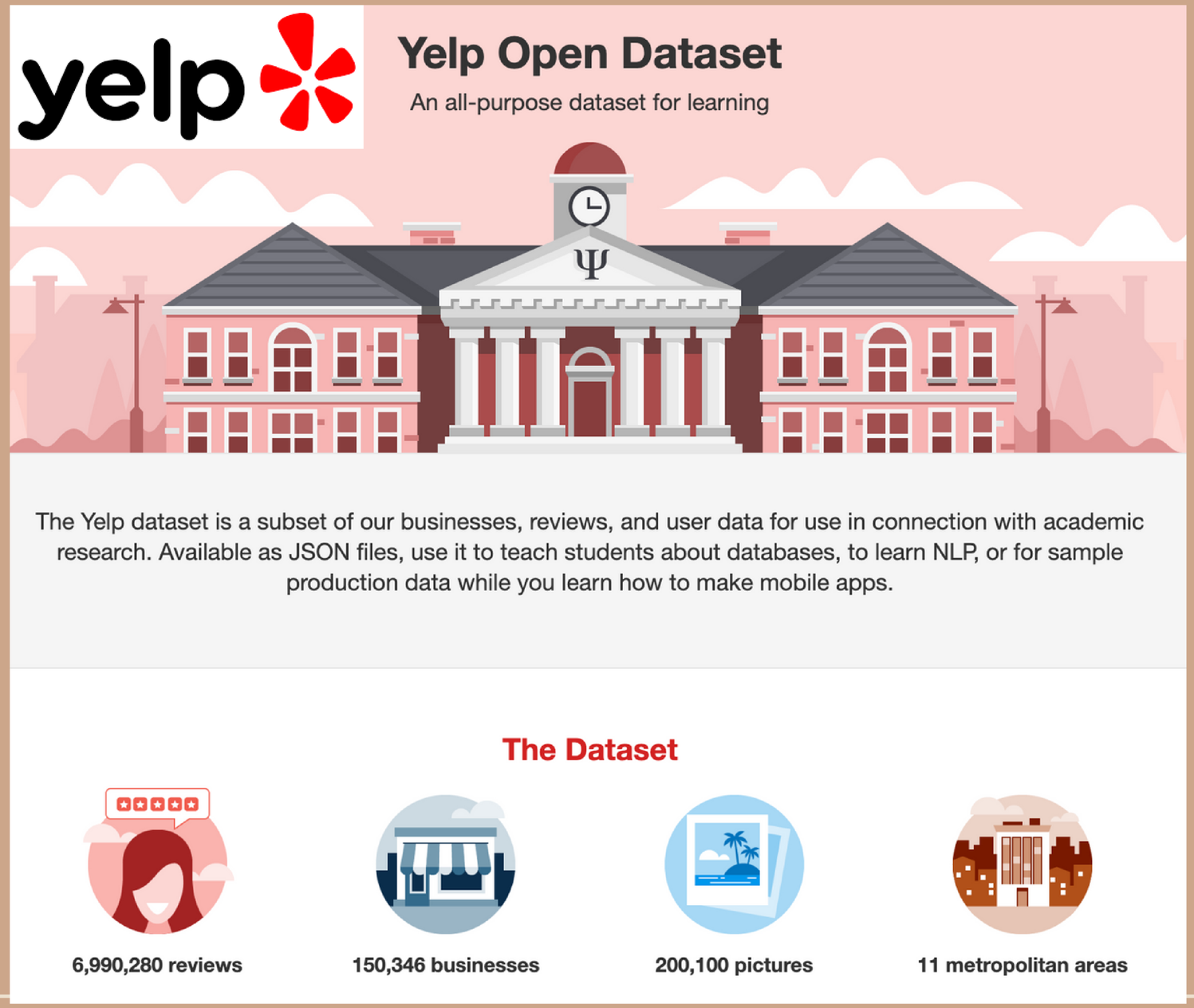
Predicting Restaurant Rating Star

Presentation by Hoda Shoghi

Predicting Restaurant Succes

OBJECTIVE

To build a **predictive model** that can accurately predict a restaurant's star rating by leveraging the vast amount of information in the dataset. This model can provide prospective Steakholders with **insights into how certain factors might influence their new establishment's ratings.**



Dataset

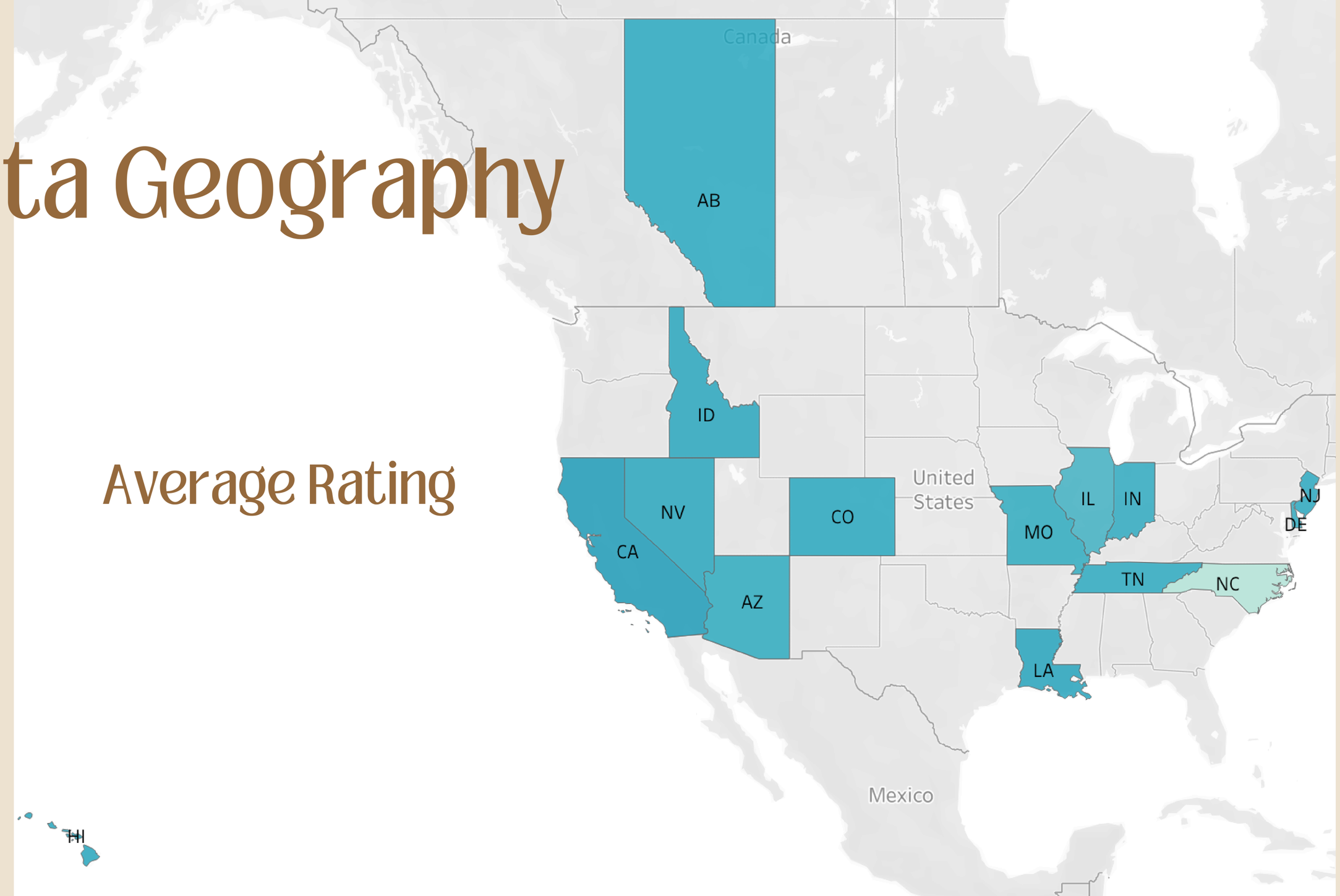
Yelp's business data as my primary dataset, specifically focusing on restaurants.

Shape (150346, 14)

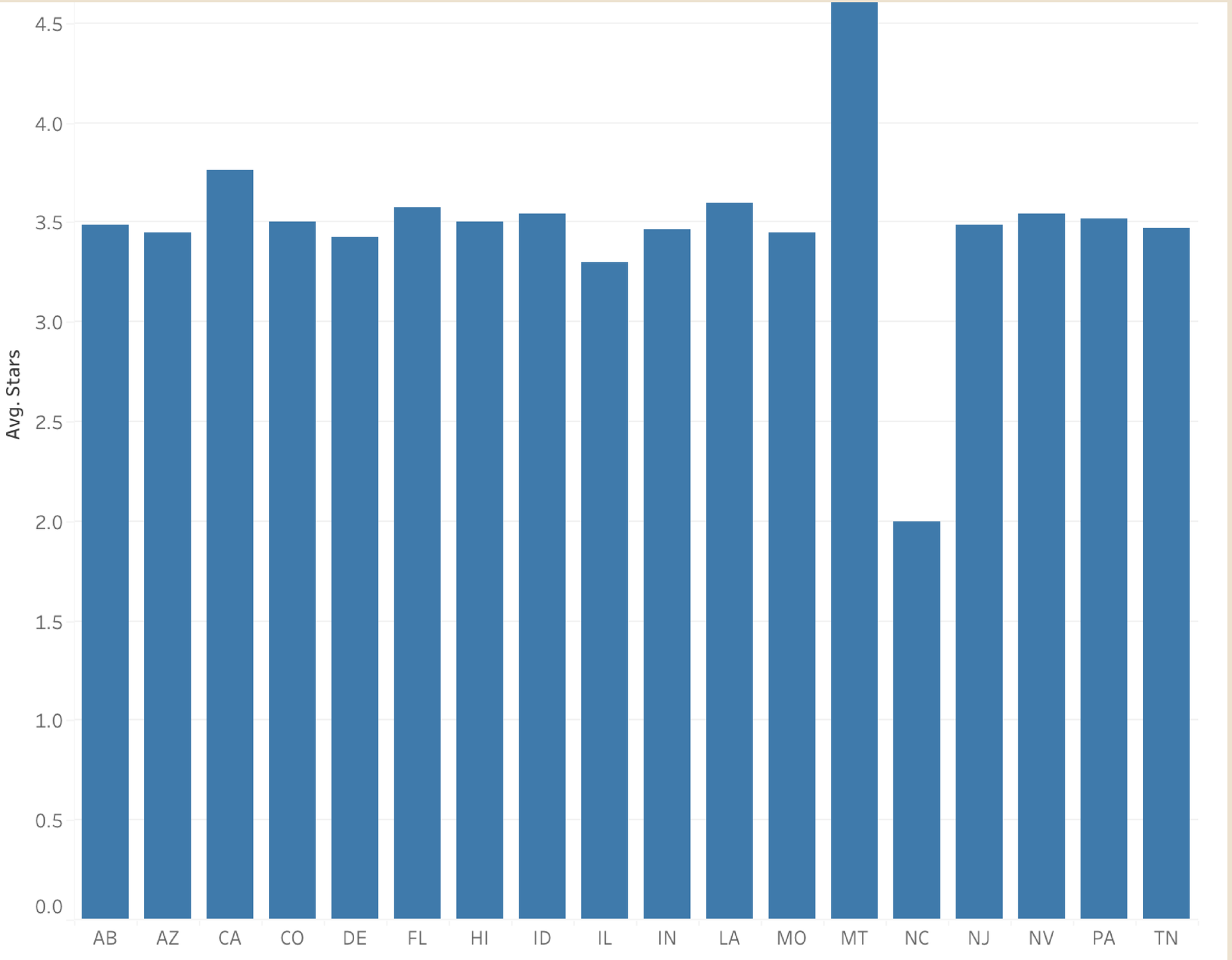
COLUMN NAME	DESCRIPTION	DATA TYPE
business_id	Unique identifier for the business	string
name	Name of the business	string
address	Address where the business is located	string
city	City where the business is located	string
state	State where the business is located	string
postal_code	postal_code where the business is located	string
latitude	Geographical latitude of the business	float64
longitude	Geographical longitude of the business	float64
stars	Star rating of the business	float64
review_count	Number of reviews the business has received	int64
is_open	0 is closed and 1 is open	int64
attributes	Dict-different attribute like payment method, , etc	string
categories	Categories the business falls under	string
hours	Dict-Hours of operation	string

Data Geography

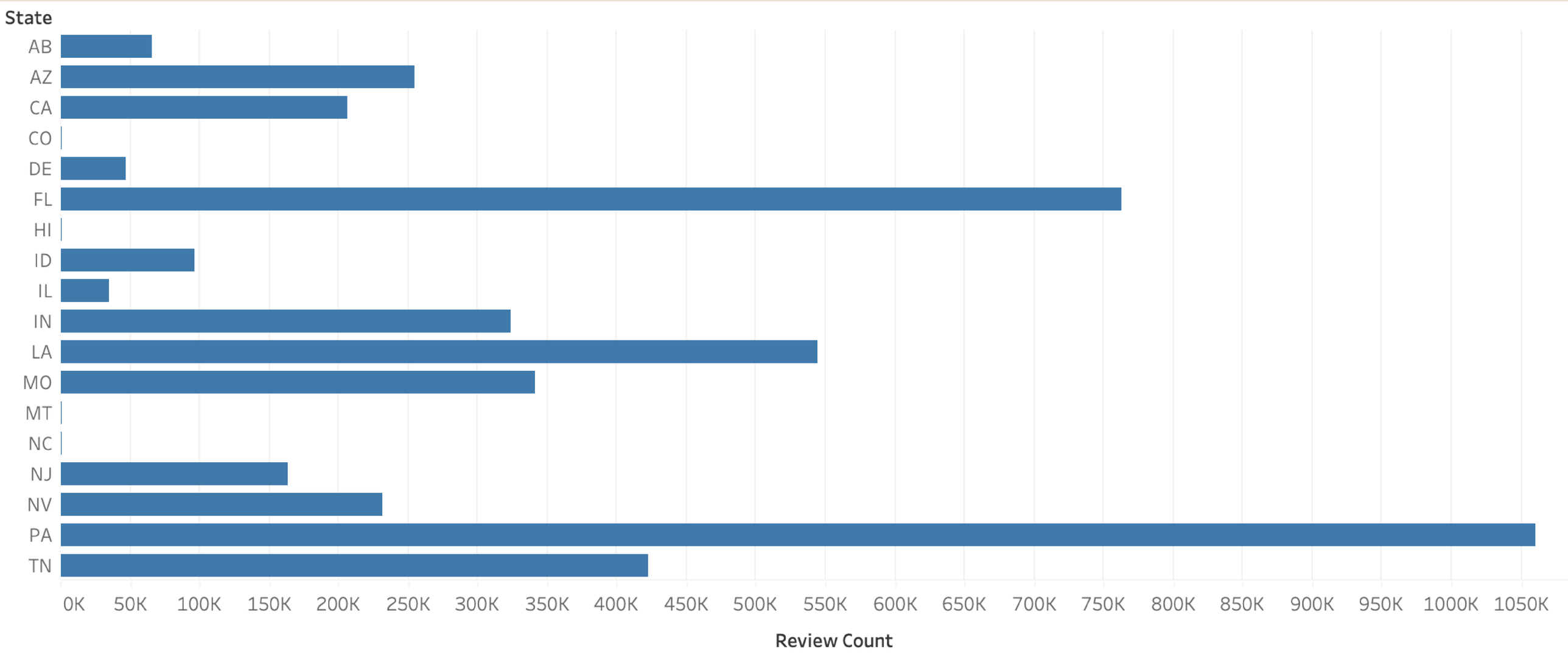
Average Rating



Data Geography



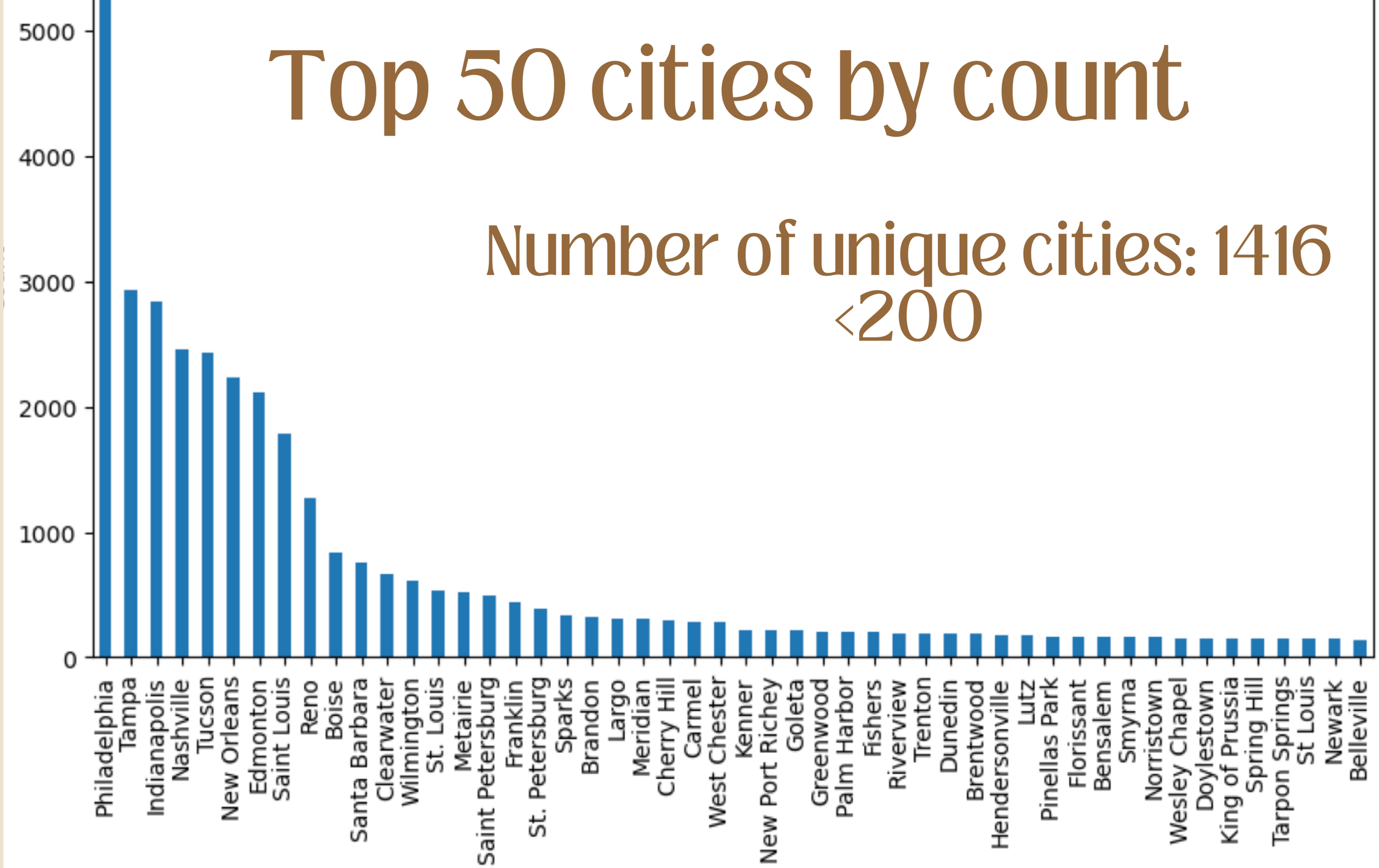
Review Counts per State



Top 50 cities by count

Number of unique cities: 1416

<200



Multiple value in one cell

	city	stars	review_count	is_open	attributes	categories
3	Philadelphia	4.0	80	1	{'RestaurantsDelivery': 'False', 'OutdoorSeati...	restaurants, food, bubble tea, coffee & tea, b...
5	Ashland City	2.0	6	1	{'BusinessParking': 'None', 'BusinessAcceptsCr...	burgers, fast food, sandwiches, food, ice crea...
8	Affton	3.0	19	0	{'Caters': 'True', 'Alcohol': 'u'full_bar'', '...	pubs, restaurants, italian, bars, american (tr...
9	Nashville	1.5	10	1	{'RestaurantsAttire': ''casual'', 'Restaurants...	ice cream & frozen yogurt, fast food, burgers,...
11	Tampa Bay	4.0	10	1	{'Alcohol': ''none'', 'OutdoorSeating': 'None'...	vietnamese, food, restaurants, food trucks

pd.get_dummies()

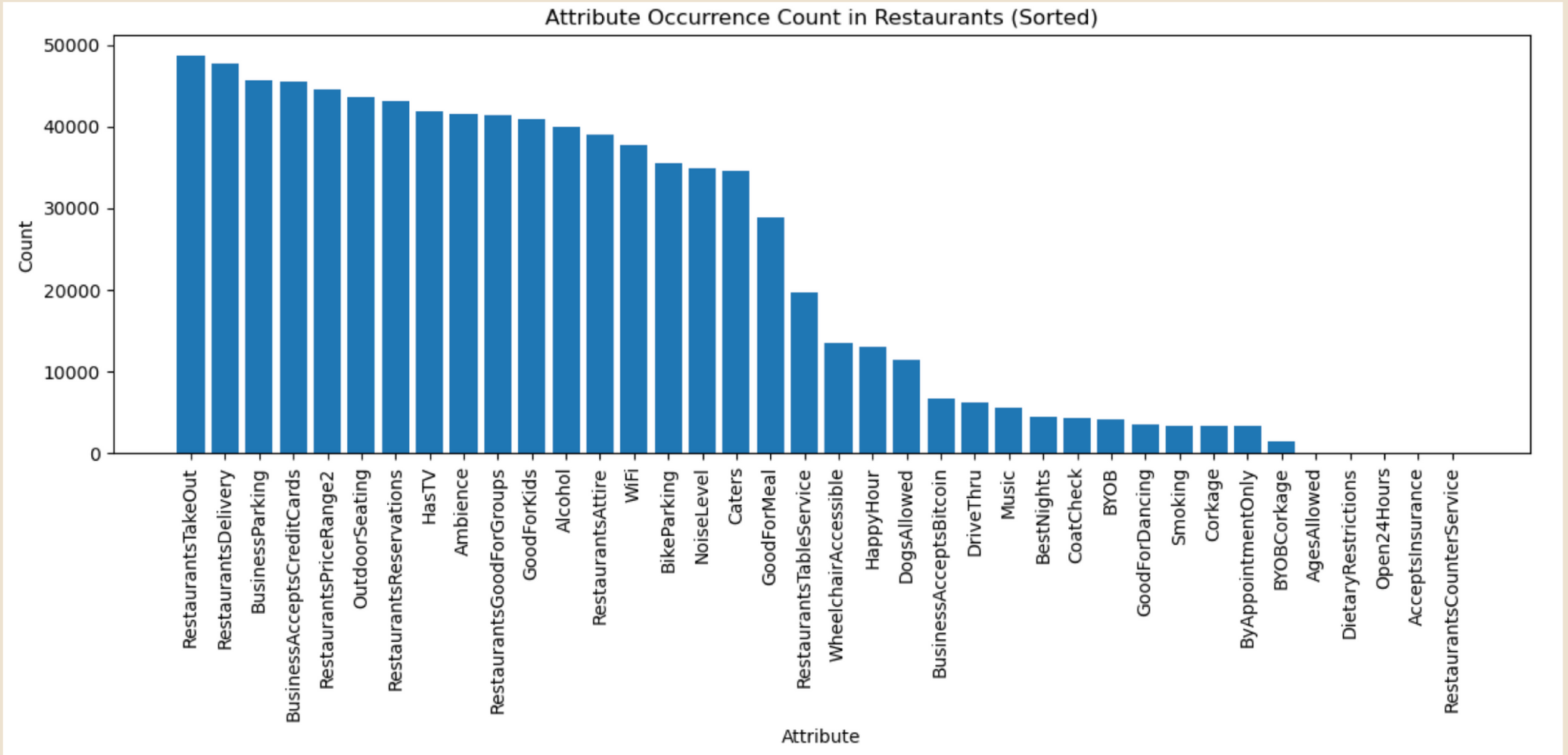
ast.literal_eval(x)

lambda x: x.get(key)

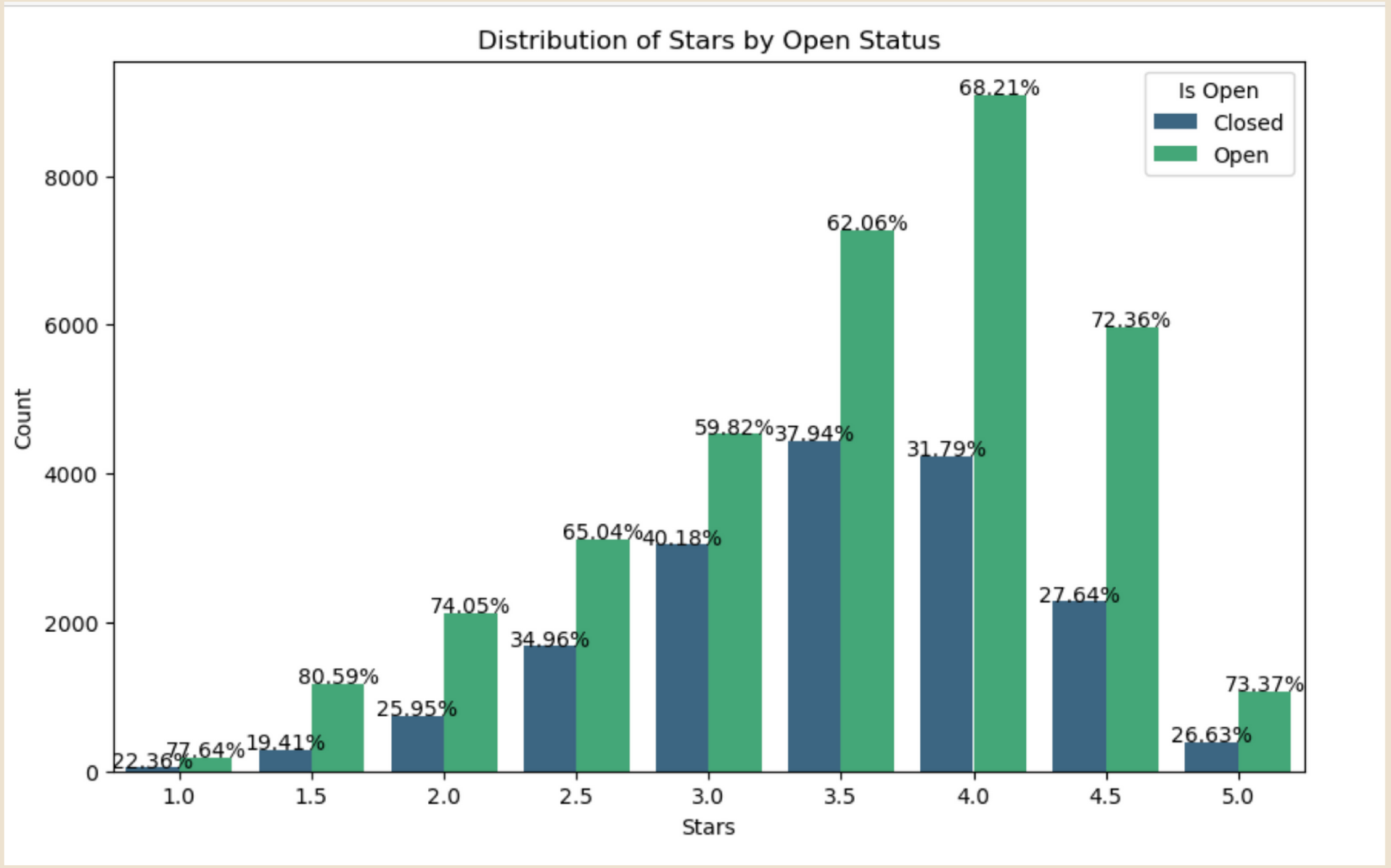
MultiLabelBinarizer

combine or delete based on correlation between categories >0.50

Flatterned Attributes

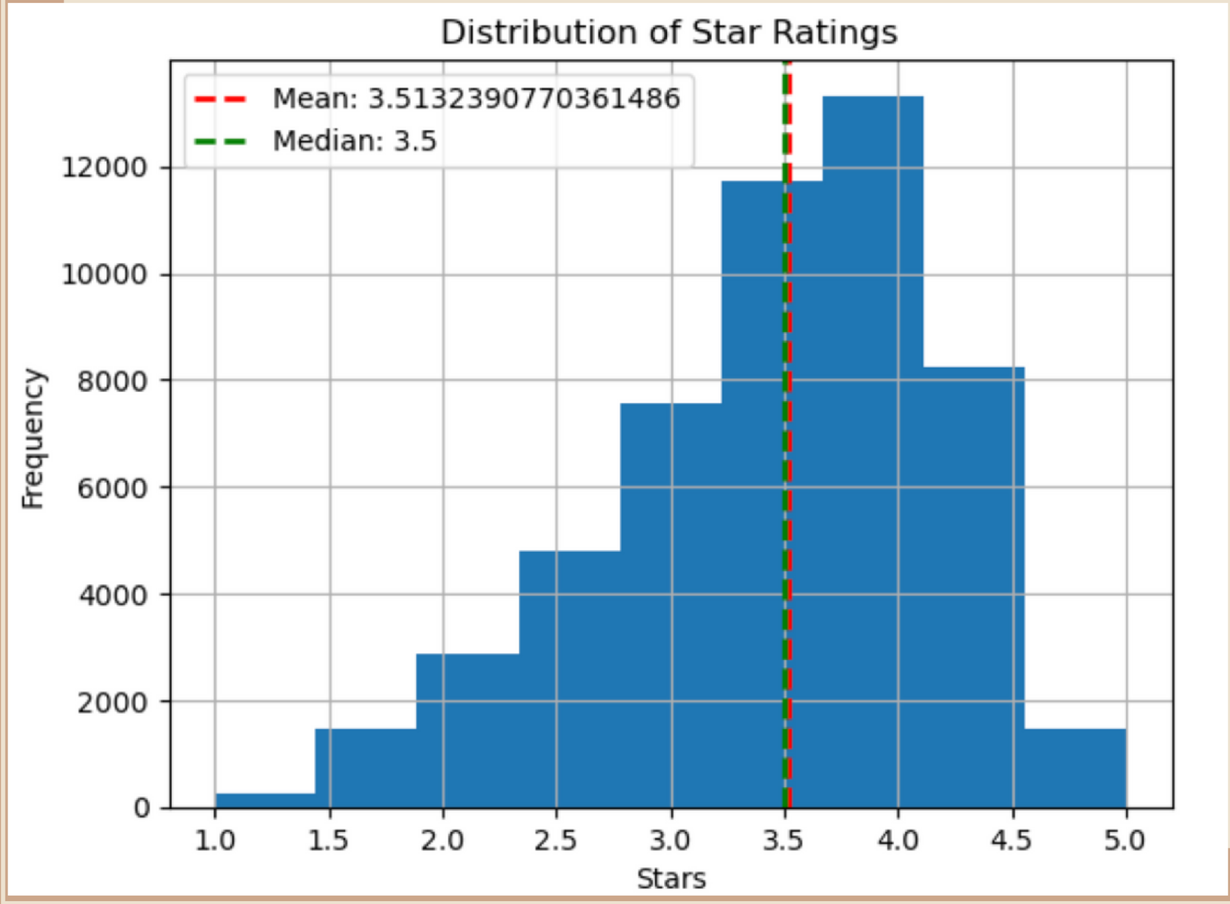


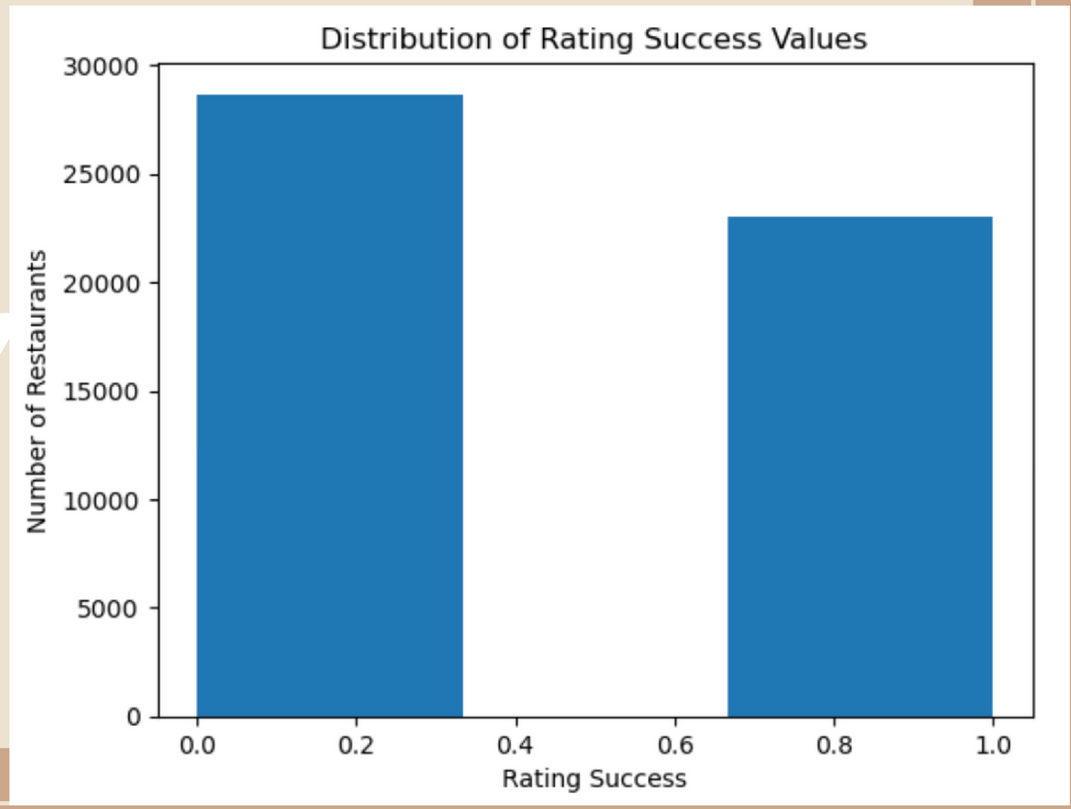
Permanently Closed



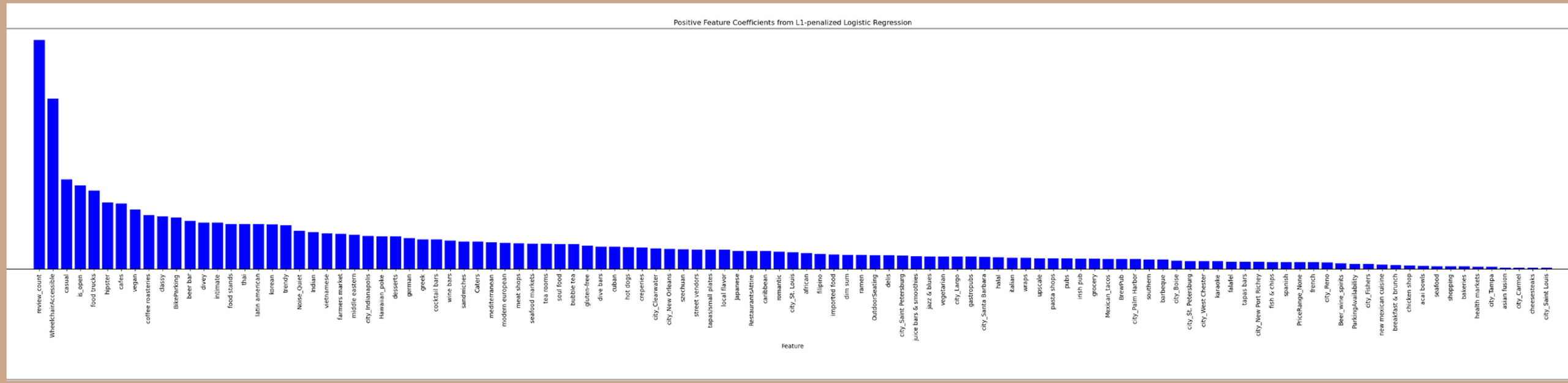
Target Variable

Final Shape(51703, 175)

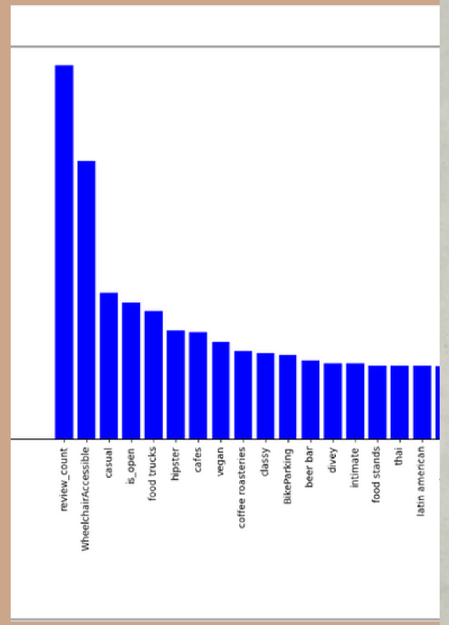


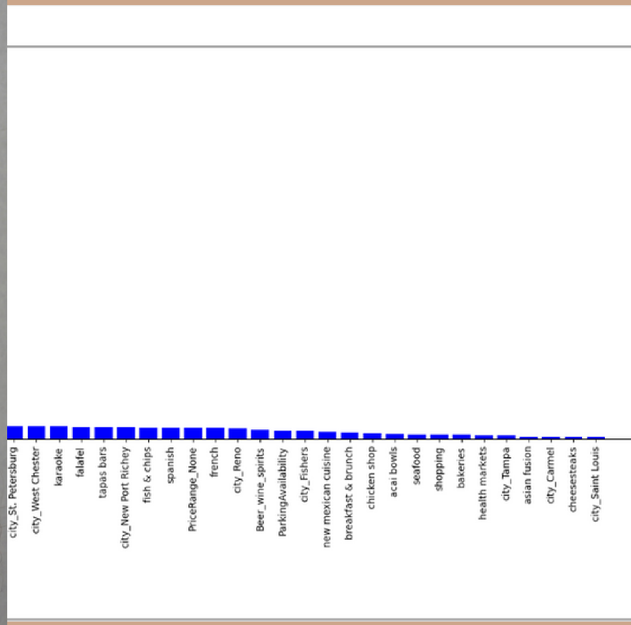


Positive Feature Coefficients from Logistic Regression

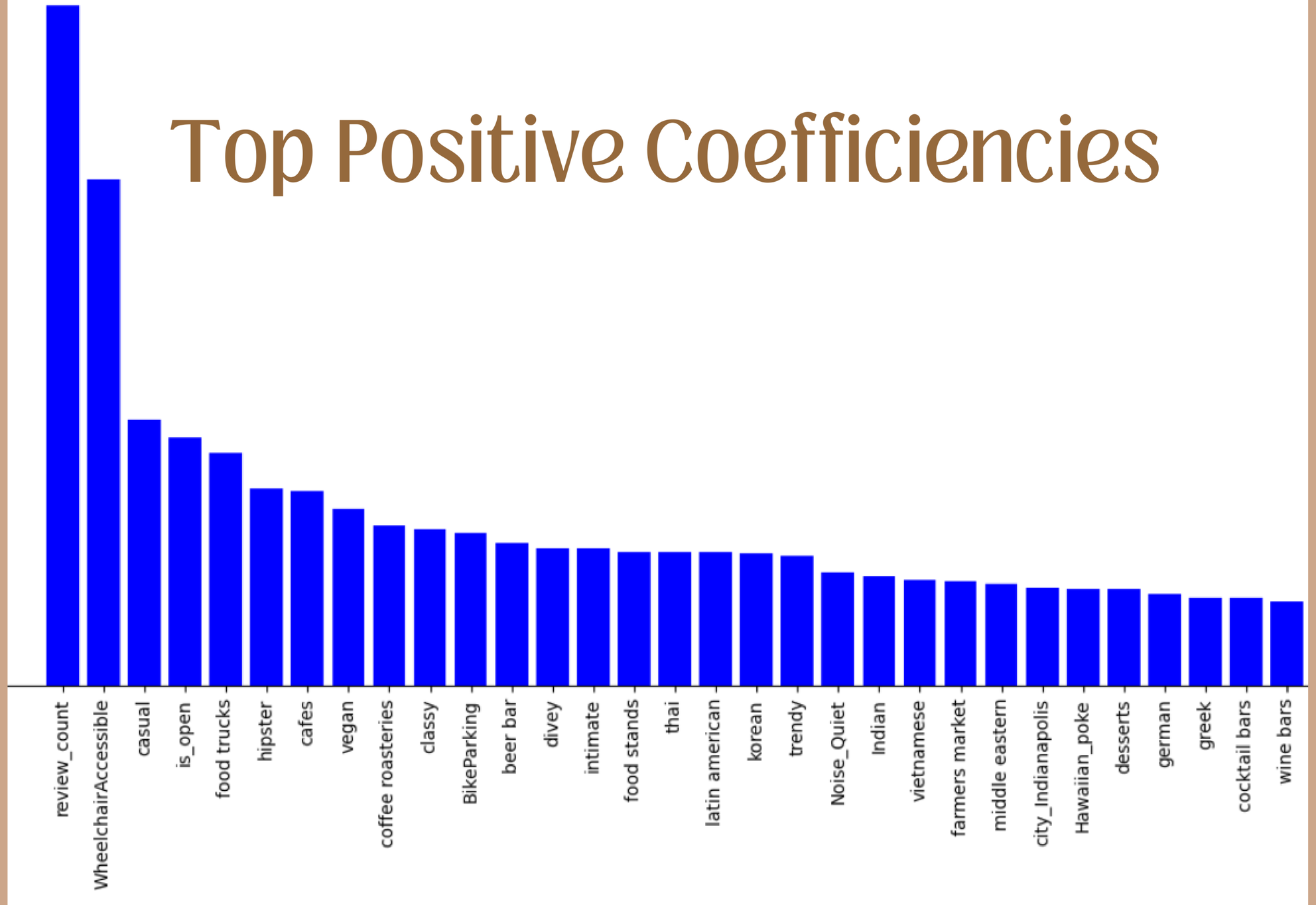


wanna Zoom in?

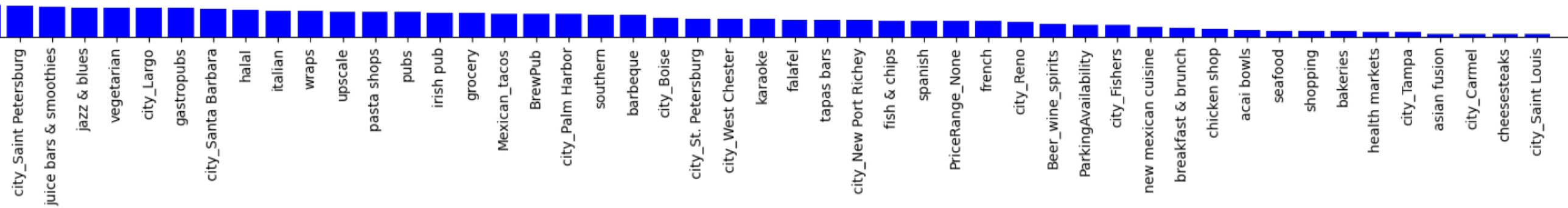




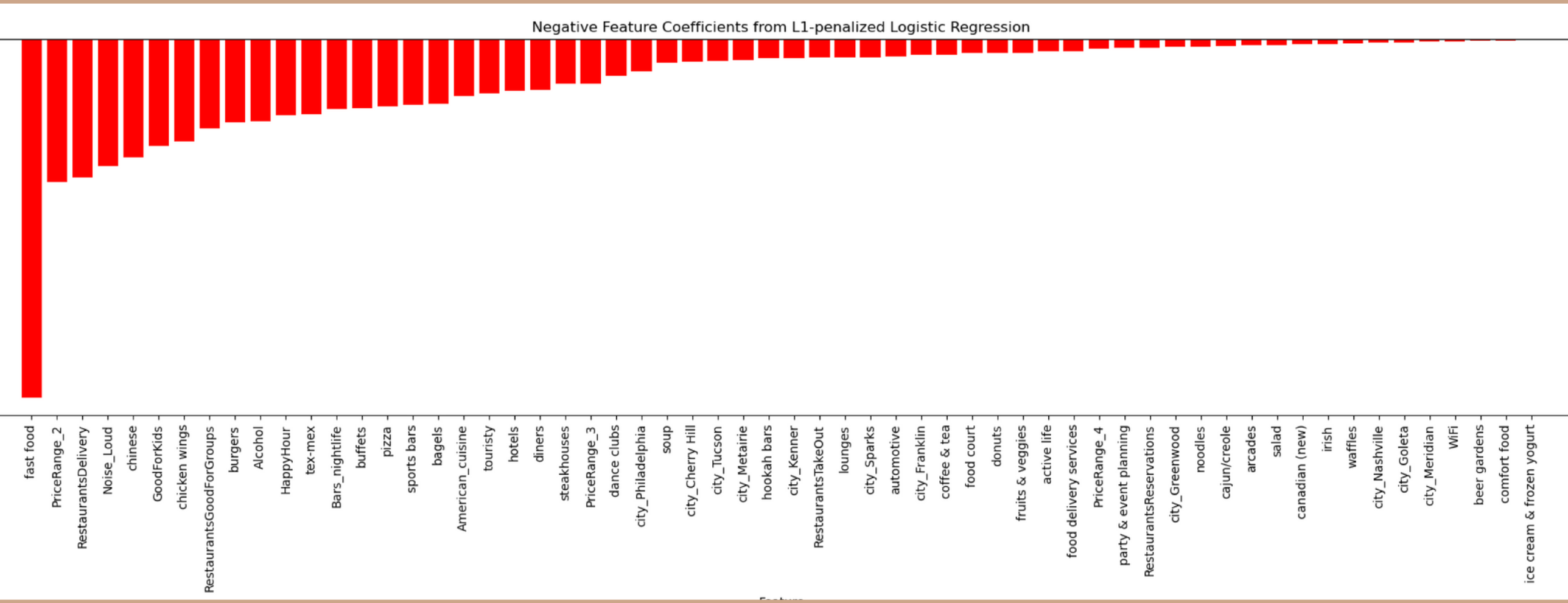
Top Positive Coefficiencies

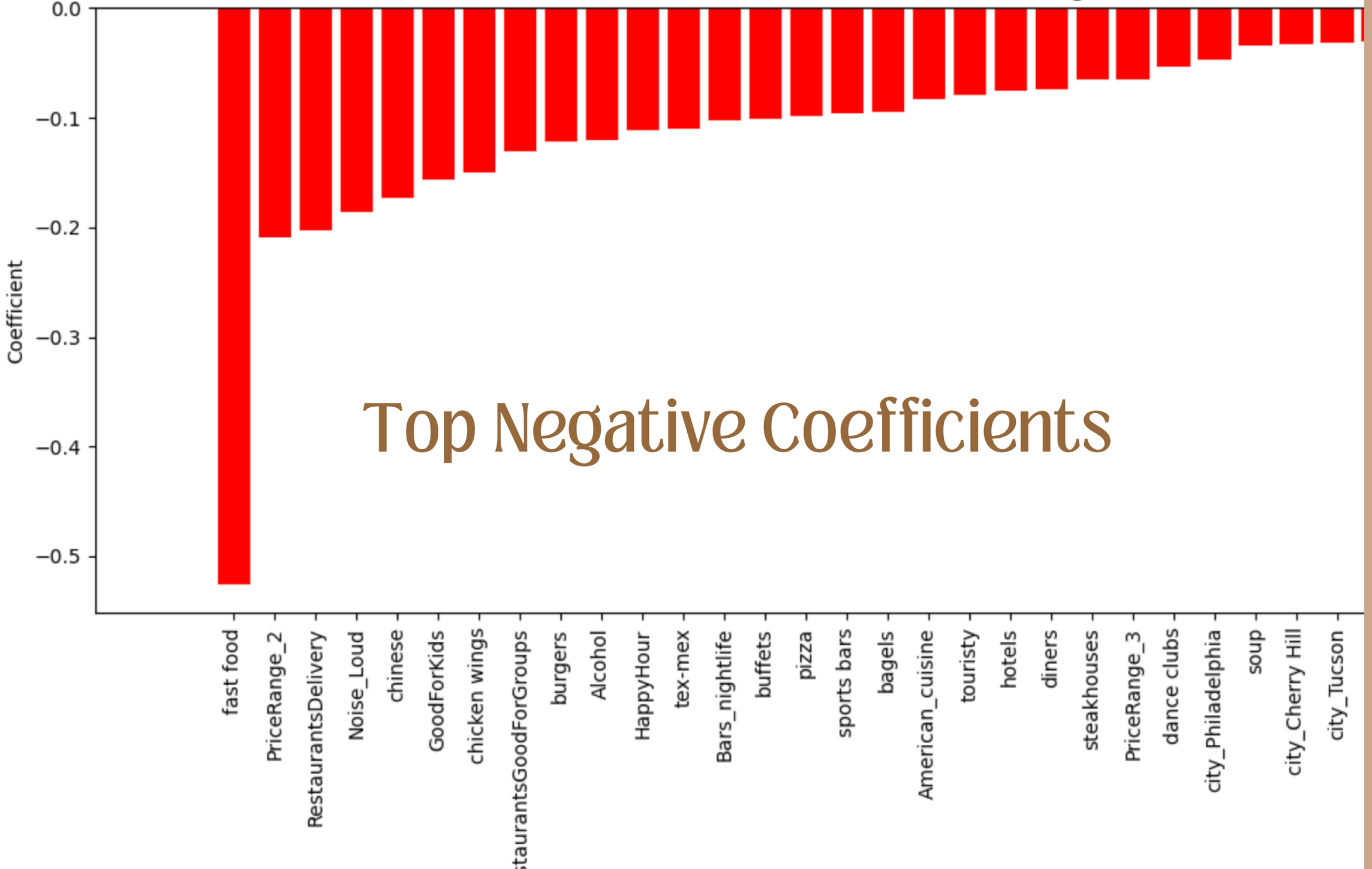


Less Positive Coefficiencies



Negative Feature Coefficients from Logistic Regression





Top Negative Coefficients

Nagative near to Zero Coefficiency

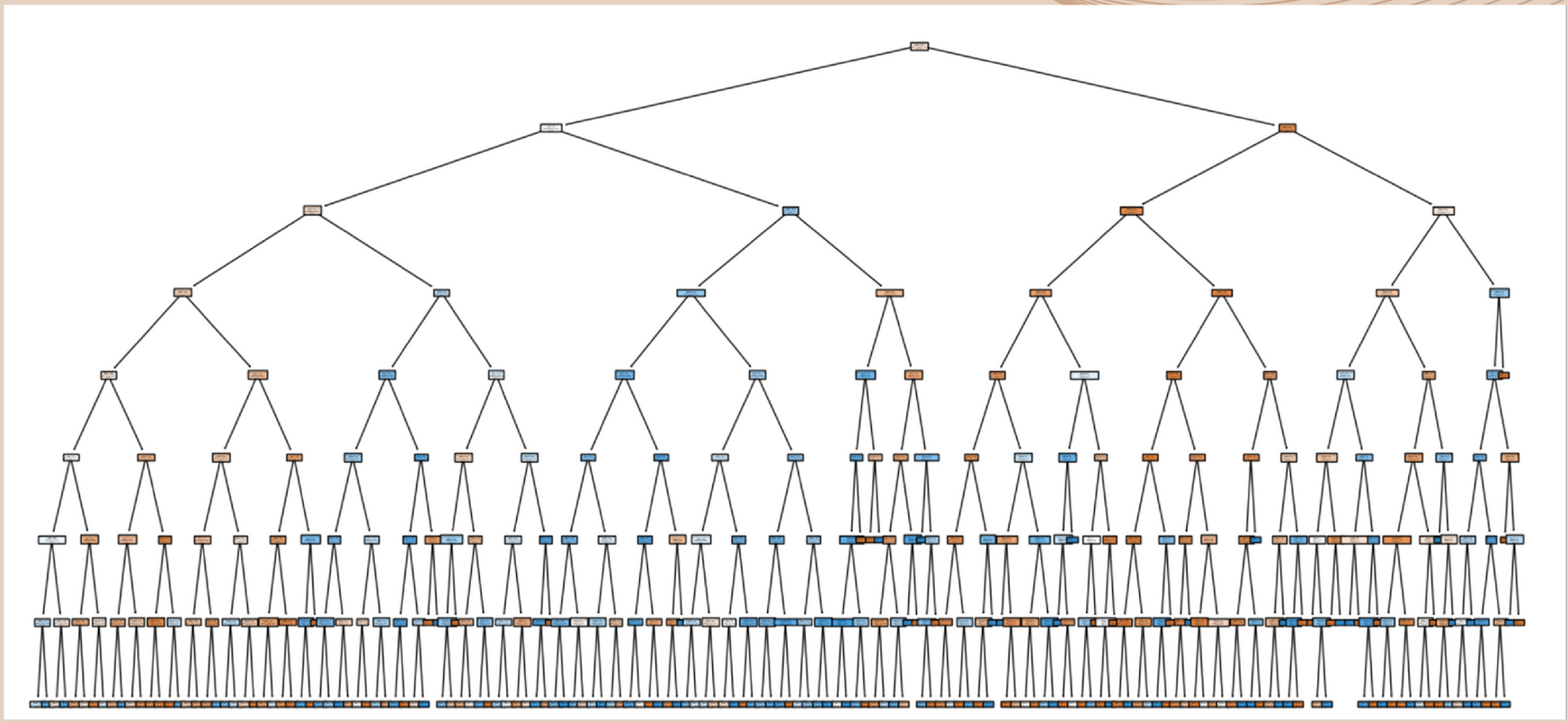
Nagative near to Zero Coefficiency

Logistic Regression & Power of Scaling

```
Train Accuracy: 0.4460101679929266
Test Accuracy: 0.4444587711946360 6

Scaled Train Accuracy: 0.7193578691423519
Scaled Test Accuracy: 0.7059506156920895
```

Decision Tree

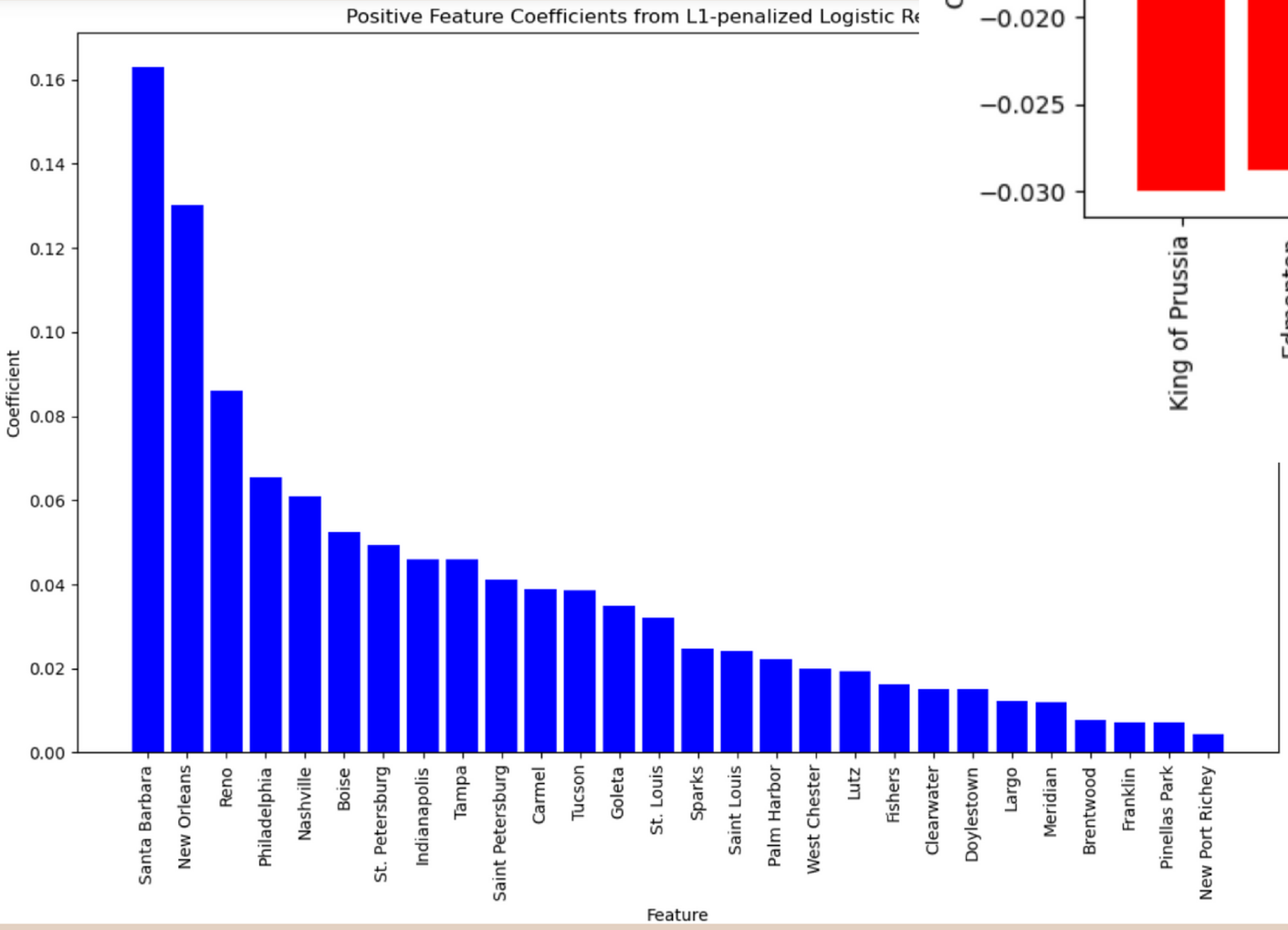
Max_Dept=8

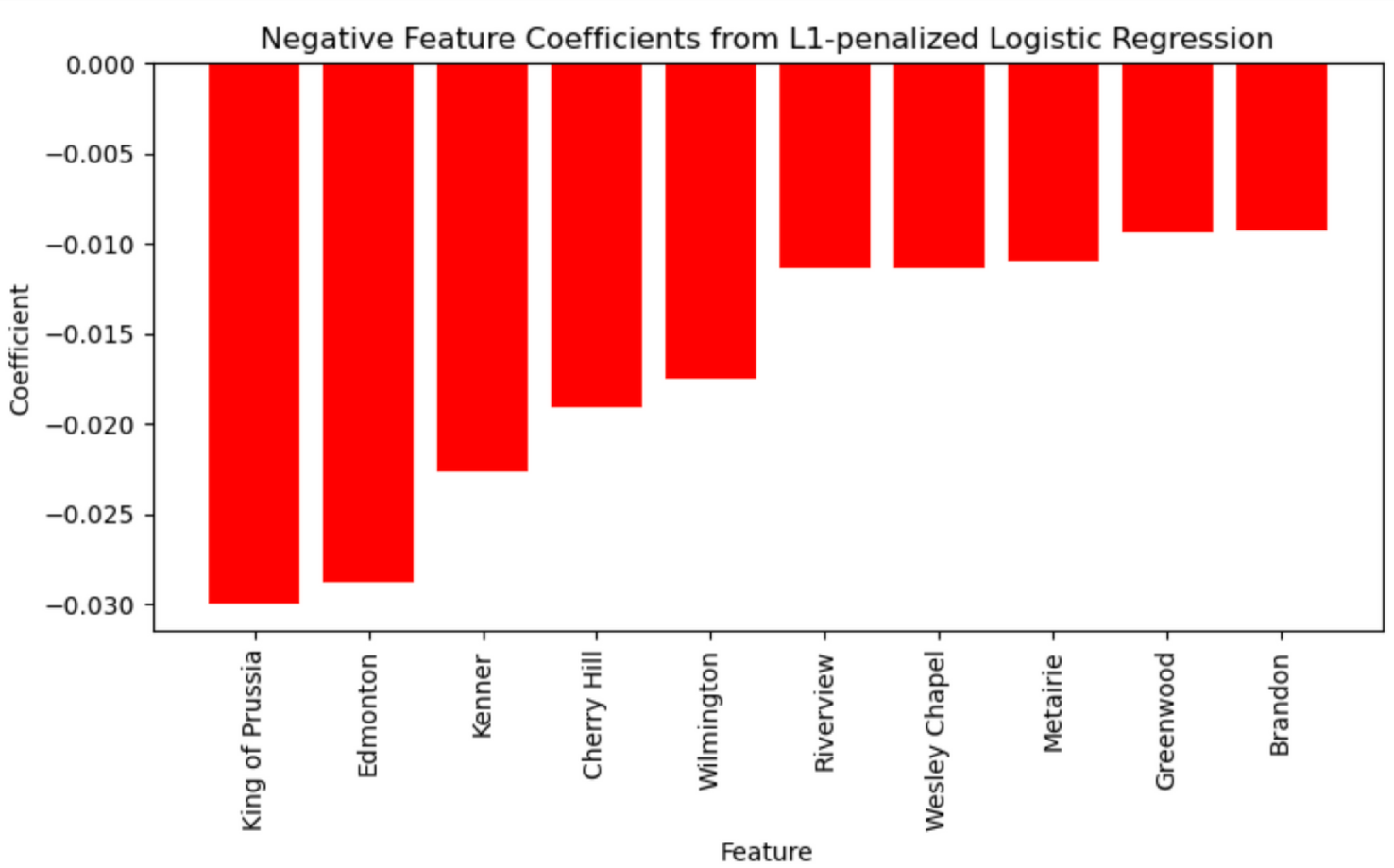


```
DecisionTree Train Accuracy: 0.4460101679929266
DecisionTree Test Accuracy: 0.44445877119463606

DecisionTree Scaled Train Accuracy: 0.6930813439434129
DecisionTree Scaled Test Accuracy: 0.671974727612662
```

Is City a Bad Player !?

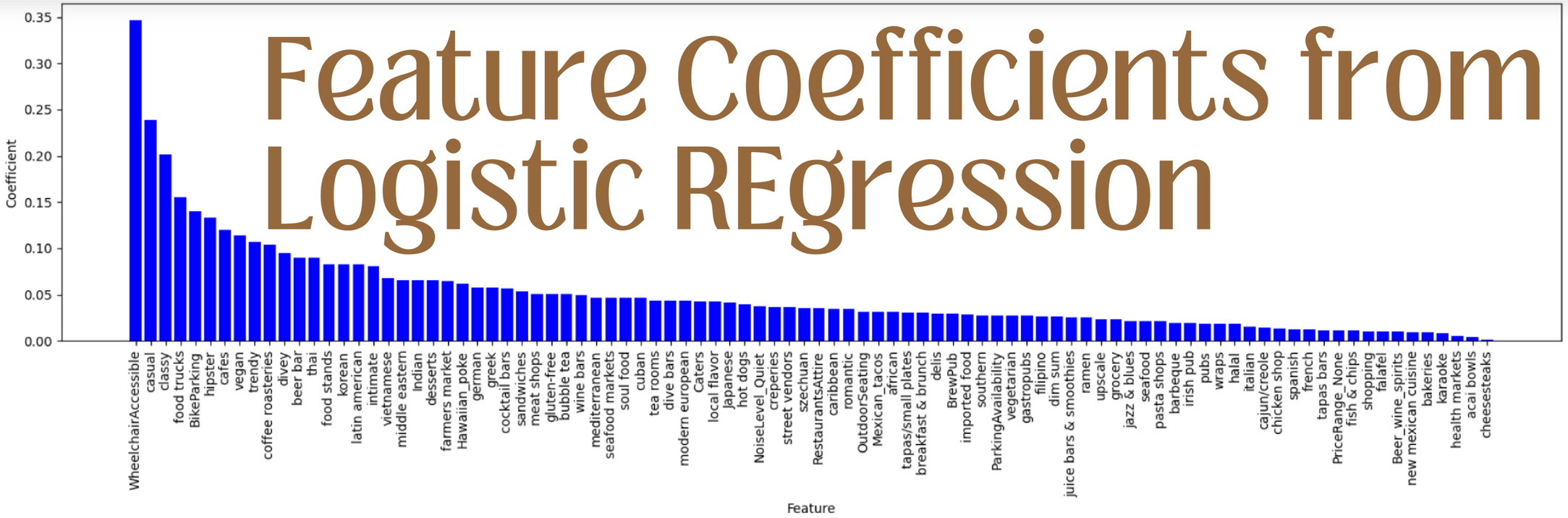
City & Target

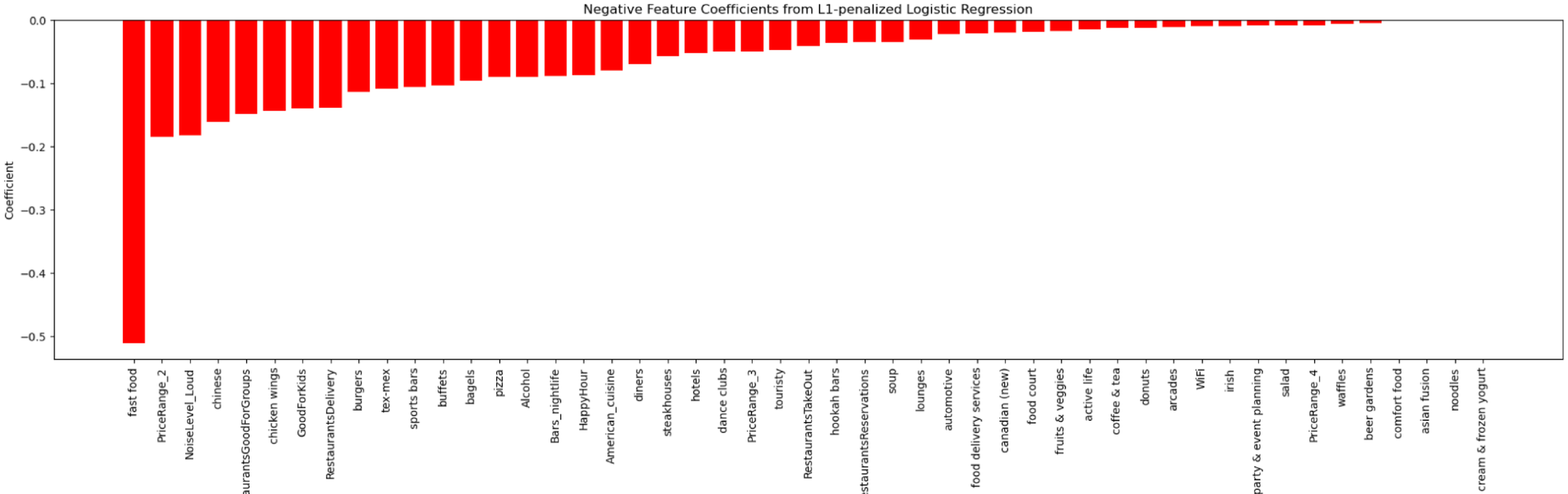




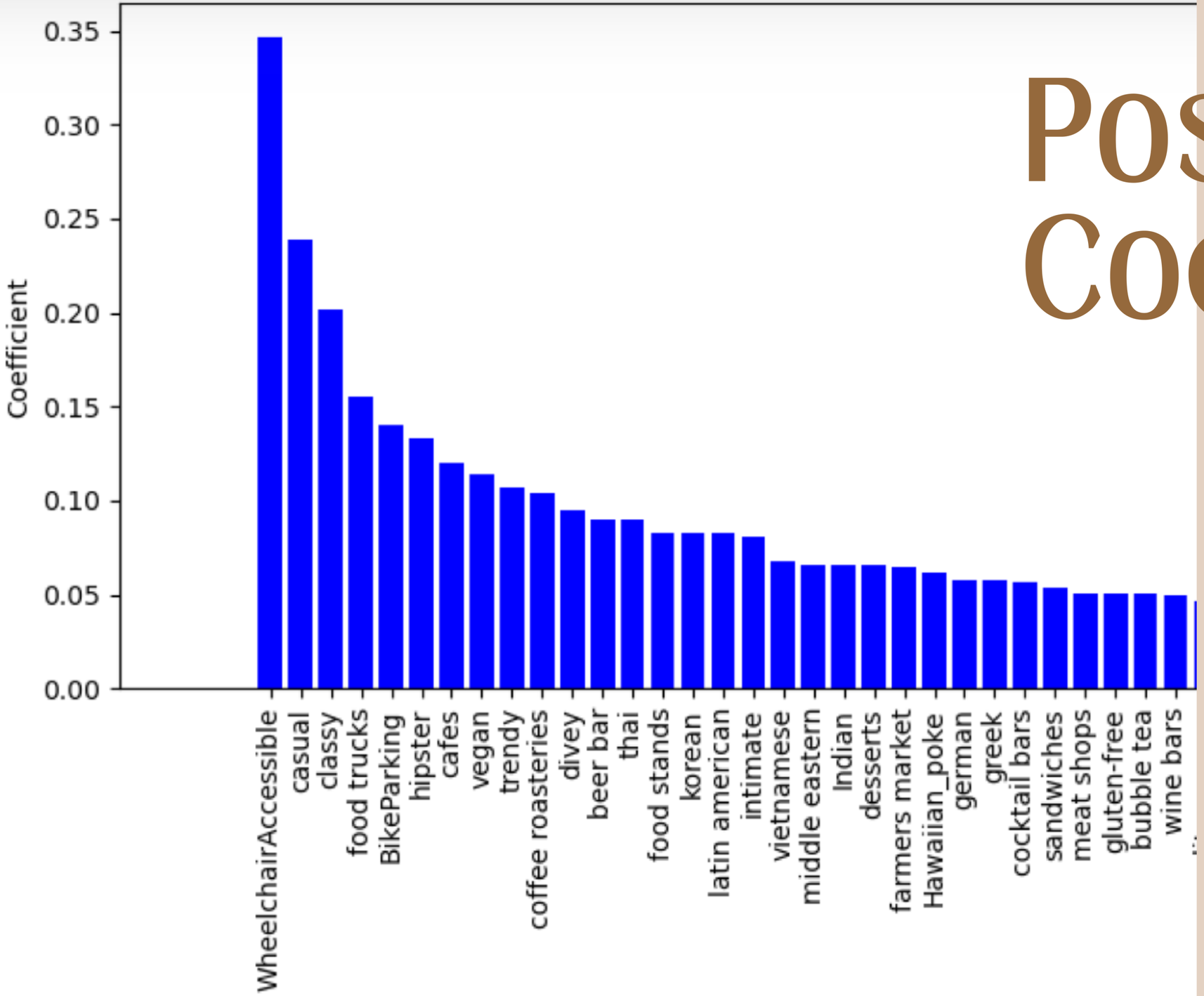

```
Train Accuracy: 0.541589859561773
Test Accuracy: 0.5376463284852788

Scaled Train Accuracy: 0.5433097
Scaled Test Accuracy: 0.5373581
```

Feature Coefficients from Logistic REgression





Positive Coefficiency



Negative Coefficients

Logistic Regression

```
Train Accuracy: 0.6115992484526968
Test Accuracy: 0.609309522274514 8

Scaled Train Accuracy: 0.7065097259062776
Scaled Test Accuracy: 0.6960221778092966
```

Deciecion Tree

```
DecisionTree Train Accuracy: 0.621822502210433 2
DecisionTree Test Accuracy: 0.6167236154986784

DecisionTree Scaled Train Accuracy: 0.672883510167993
DecisionTree Scaled Test Accuracy: 0.660692411836761
```

Forging Ahead

1 Cross Validation - PCA

2 Random Forest - Feature importance

3 Hyper Parameter Optimization & Pipeline

4 X-Gboost- Shably


Thank
You