

# Predicting Restaurant Rating Star

Presentation by Hoda Shoghi



# Predicting Restaurant Success

## 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 Stakeholders with **insights** into how certain factors might influence their new establishment's ratings.





## Yelp Open Dataset

An all-purpose dataset for learning



The Yelp dataset is a subset of our businesses, reviews, and user data for use in connection with academic research. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

### The Dataset



6,990,280 reviews



150,346 businesses



200,100 pictures



11 metropolitan areas

# Dataset

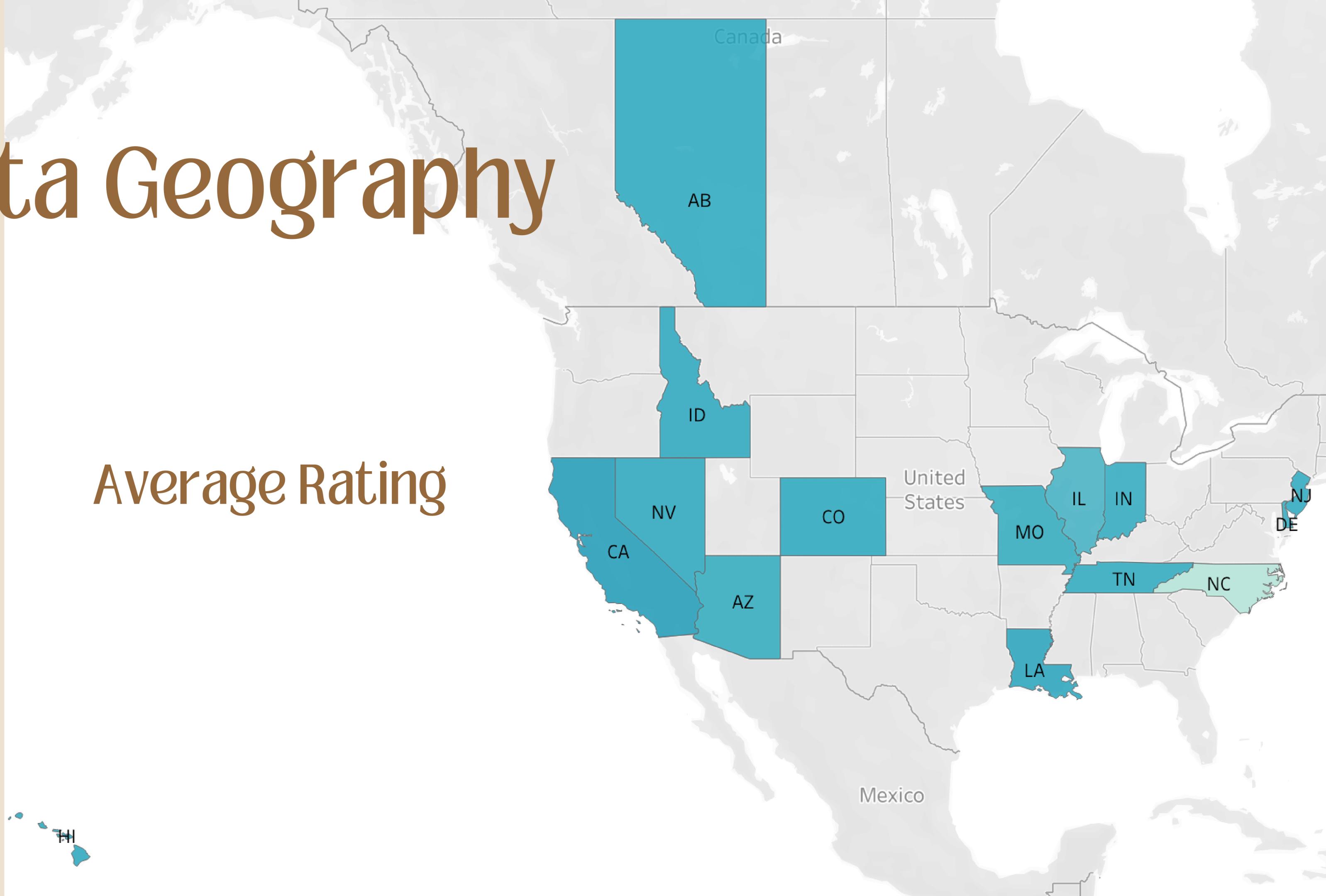
Yelp's business data as my primary dataset, specifically focusing on restaurants.

**Shape  
(150346, 14)**

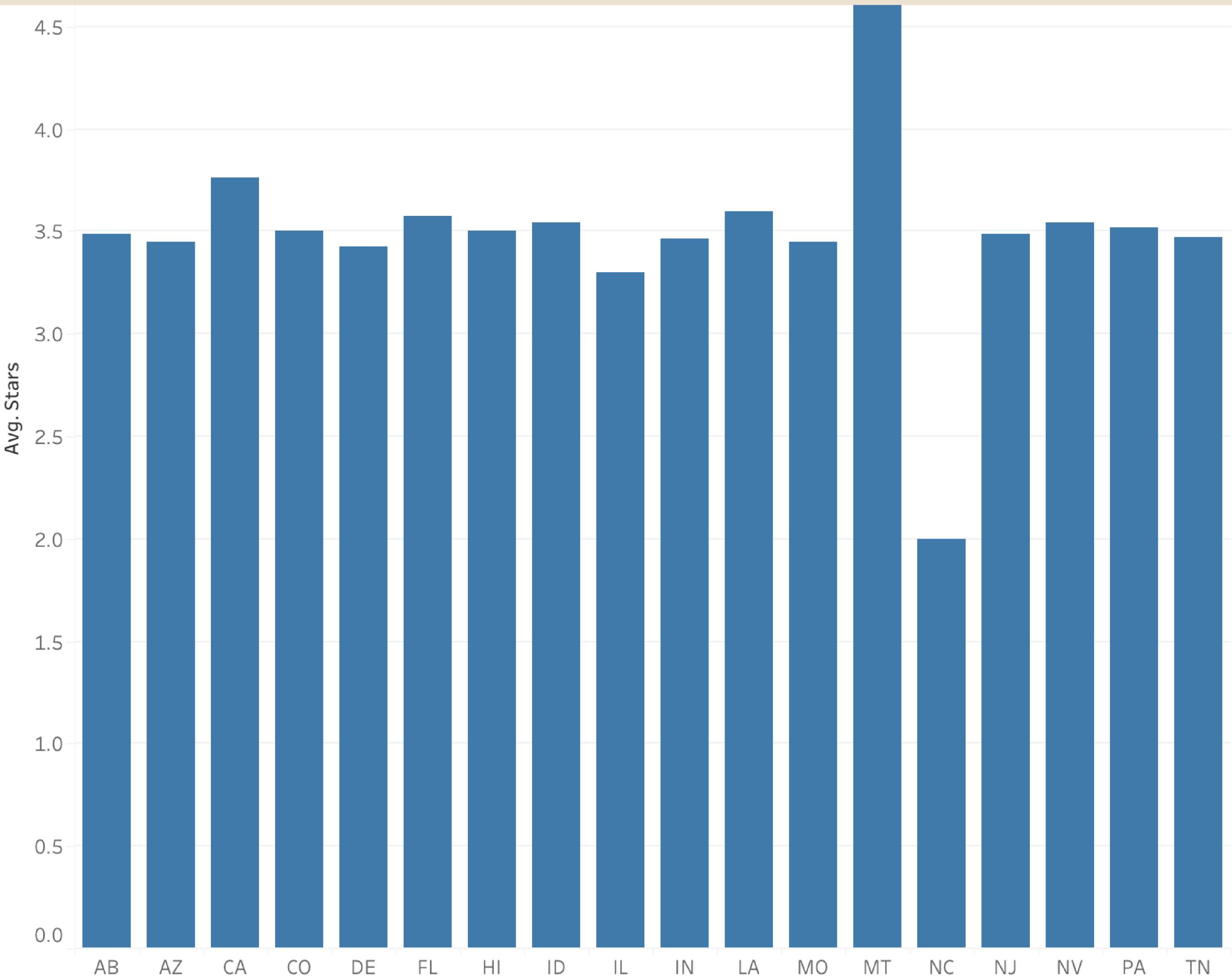
COLUMN NAME	DESCRIPTION	DATA TYPE
business_id	Unique identifier for the business	string
name	Name of the business	string
address	Address where the business is located	string
city	City where the business is located	string
state	State where the business is located	string
postal_code	postal_code where the 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

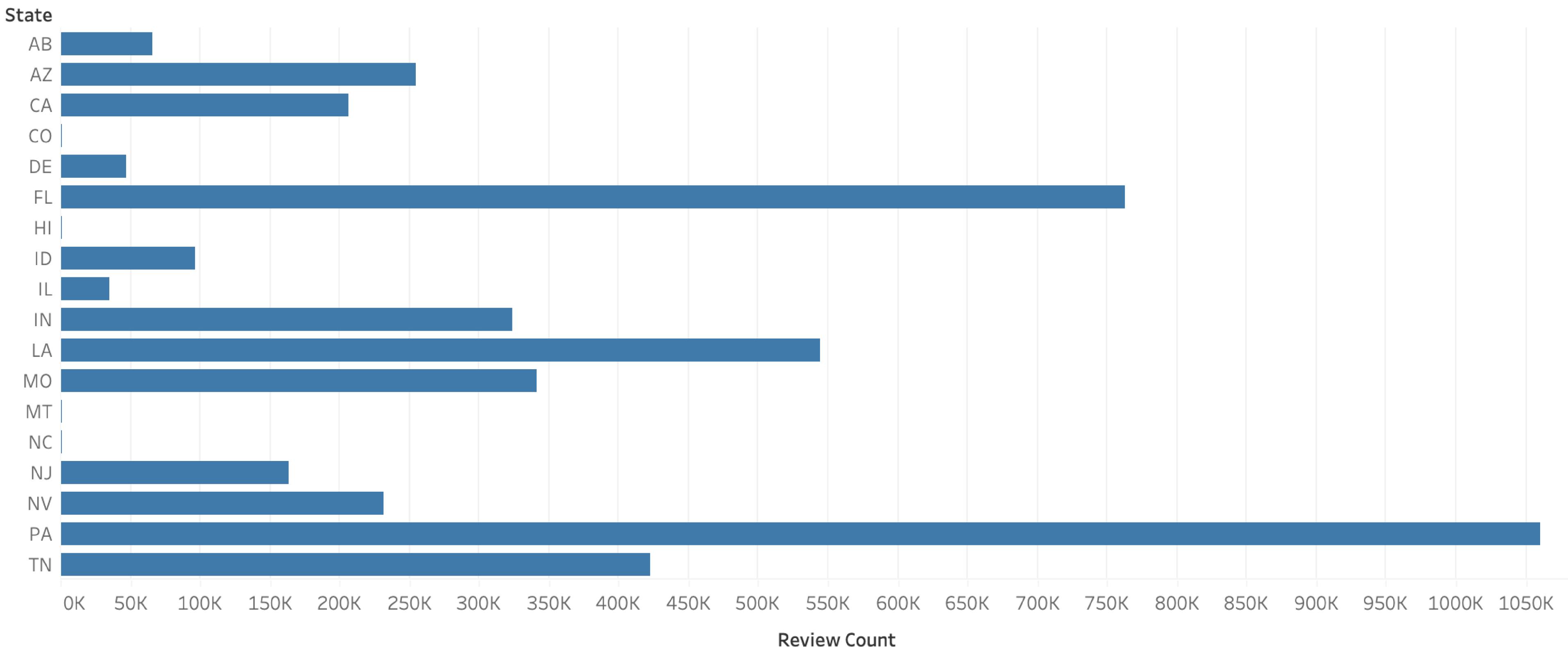
Average Rating



# Data Geography

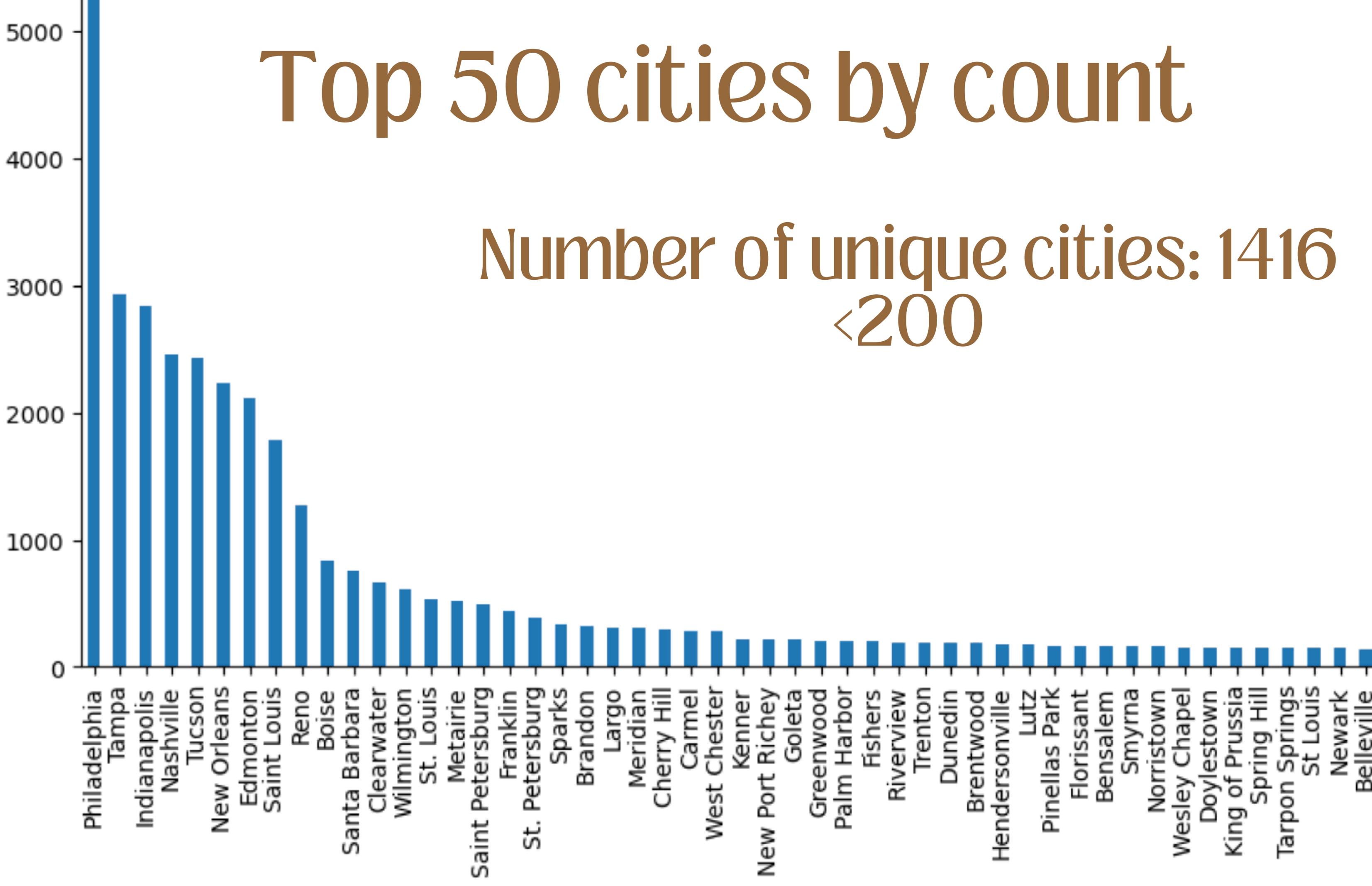


# Review Counts per State



# 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)

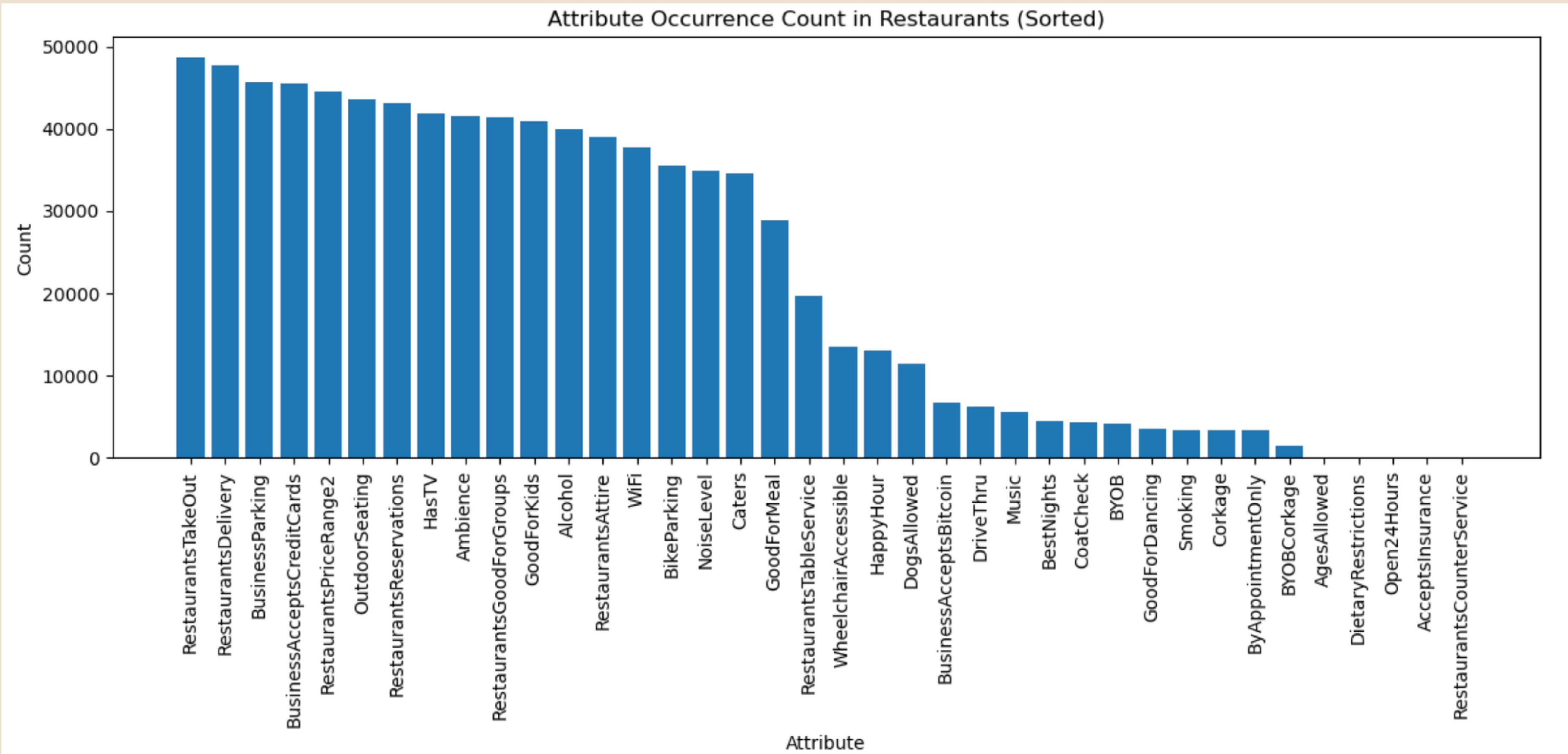
lambda x: x.get(key)

MultiLabelBinarizer

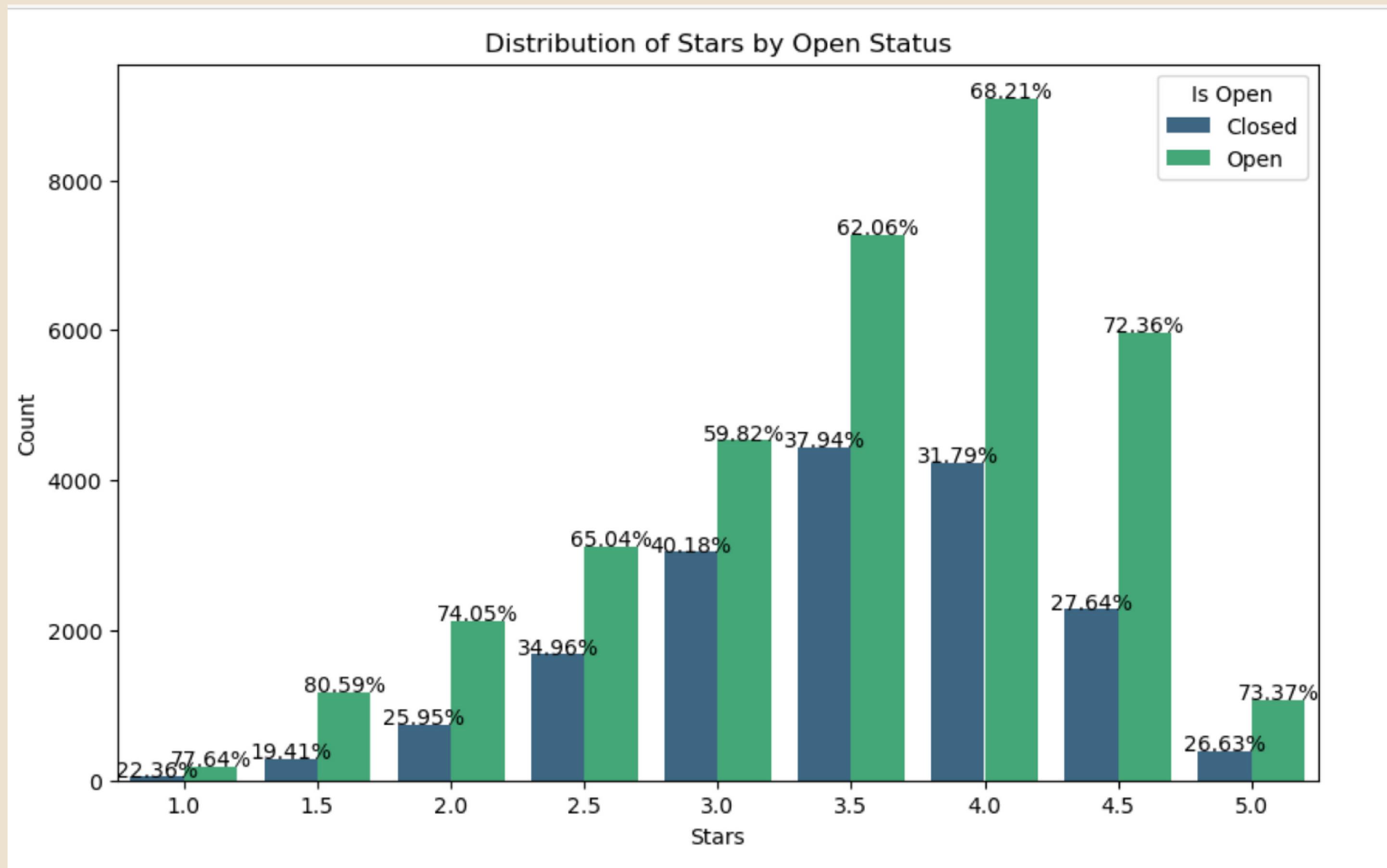
combine or delete  
based on correlation  
between categories

>0.50

# Flattered Attributes

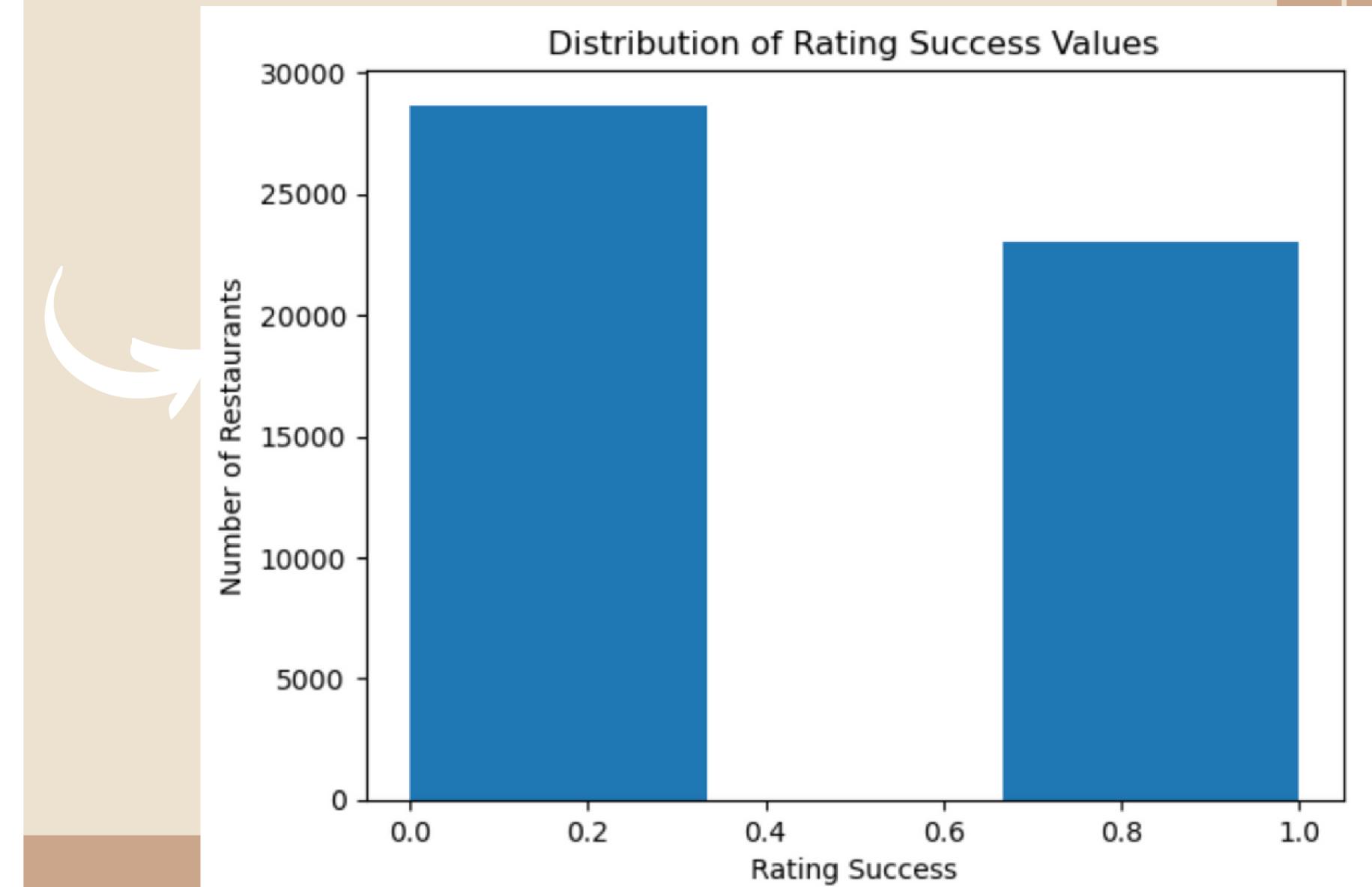
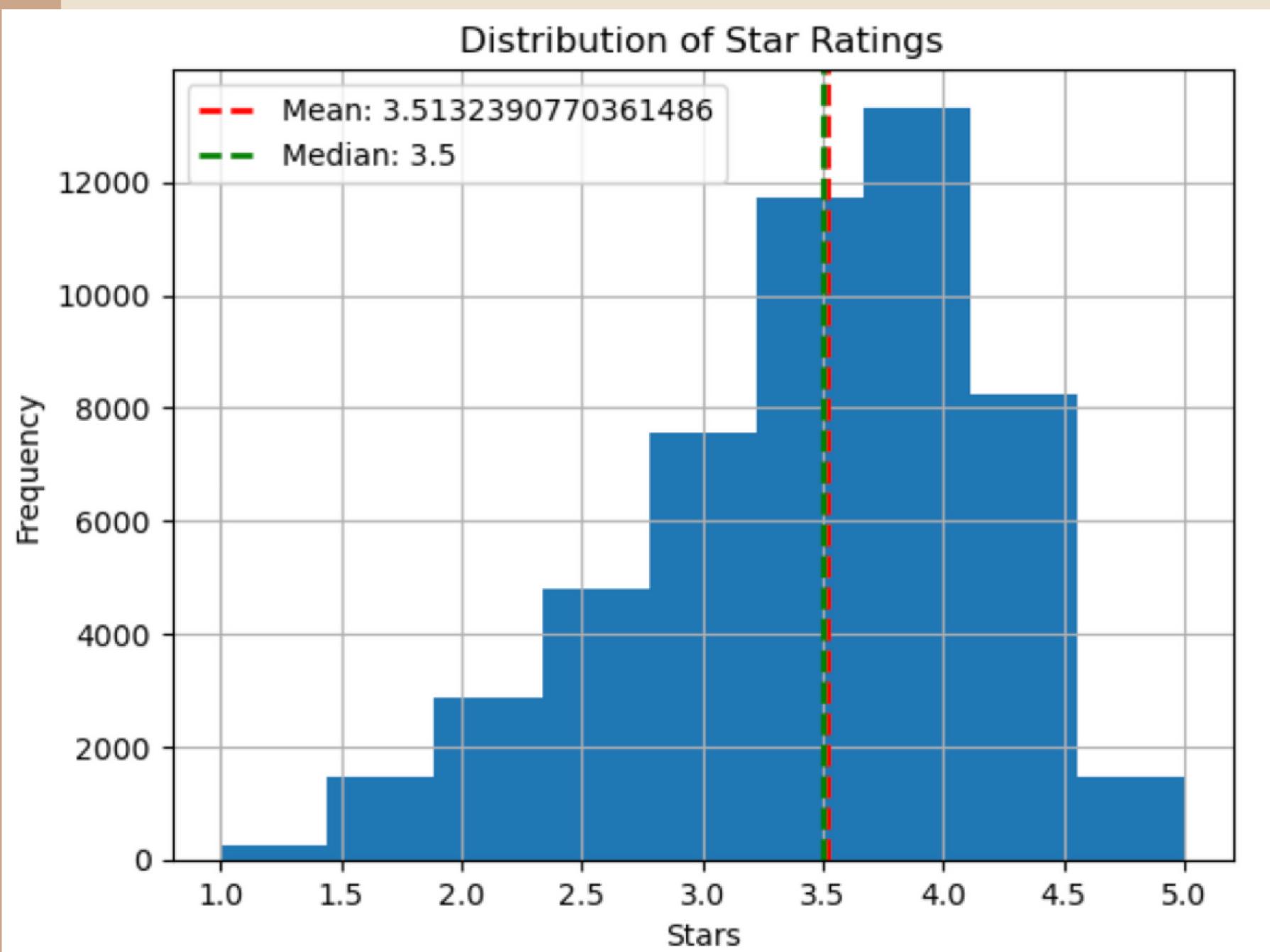


# Permanently Closed

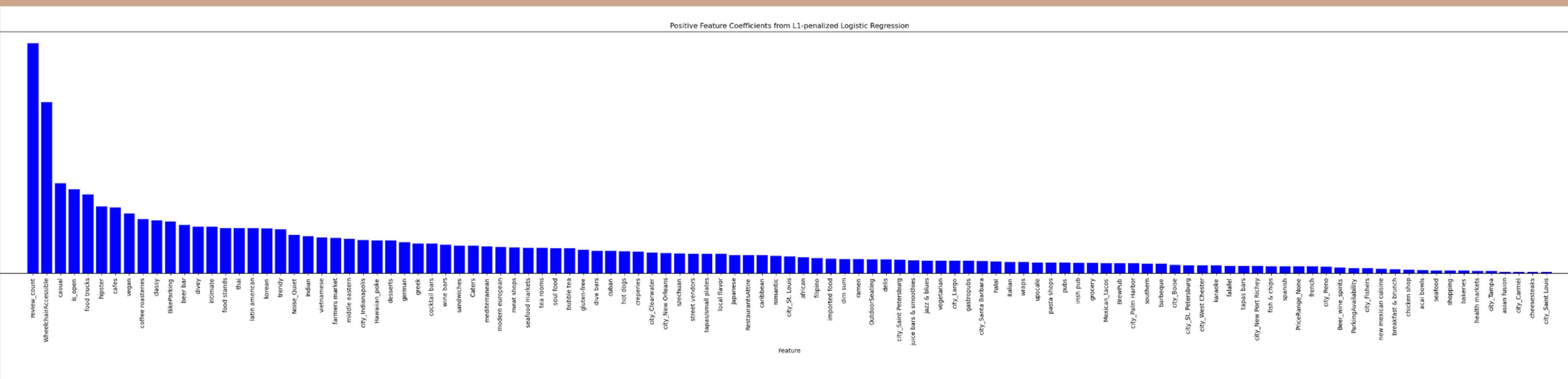


# Target Variable

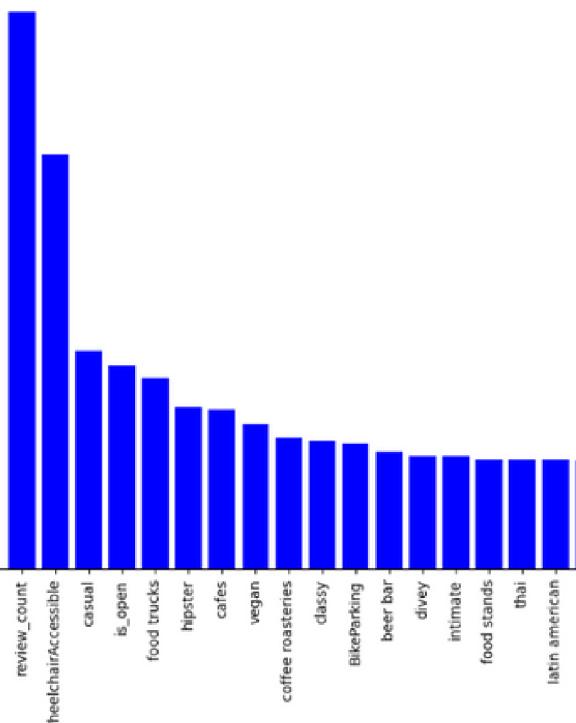
Final Shape(51703, 175)



# Positive Feature Coefficients from Logistic Regression

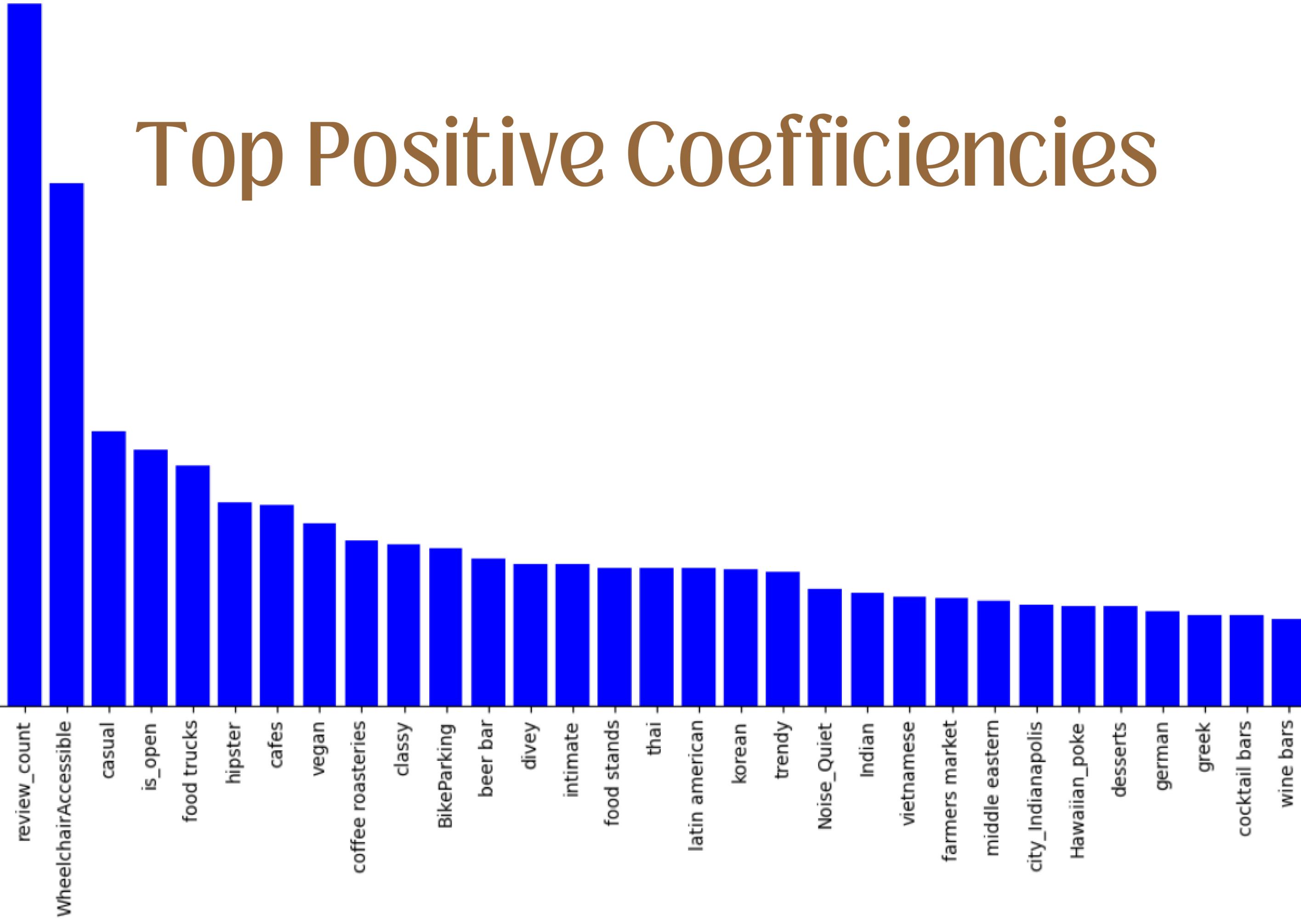


# wanna Zoom in?

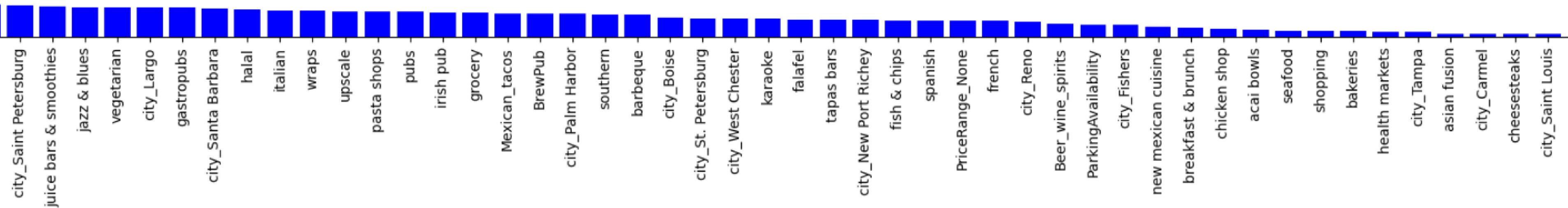


city\_St\_Petersburg  
city\_West\_Chester  
karaoke  
falafel  
tapas bars  
city\_New\_Port\_Richey  
fish & chips  
spanish  
PriceRange\_None  
franch  
city\_Reno  
Beer\_wine\_spirits  
ParkingAvailability  
city\_Fishers  
new\_mexican\_cuisine  
breakfast\_brunch  
chicken\_shop  
acai\_bowls  
seafood  
shopping  
bakeries  
health\_markets  
city\_Tampa  
asian\_fusion  
city\_Carmel  
cheesesteaks  
city\_Saint\_Louis

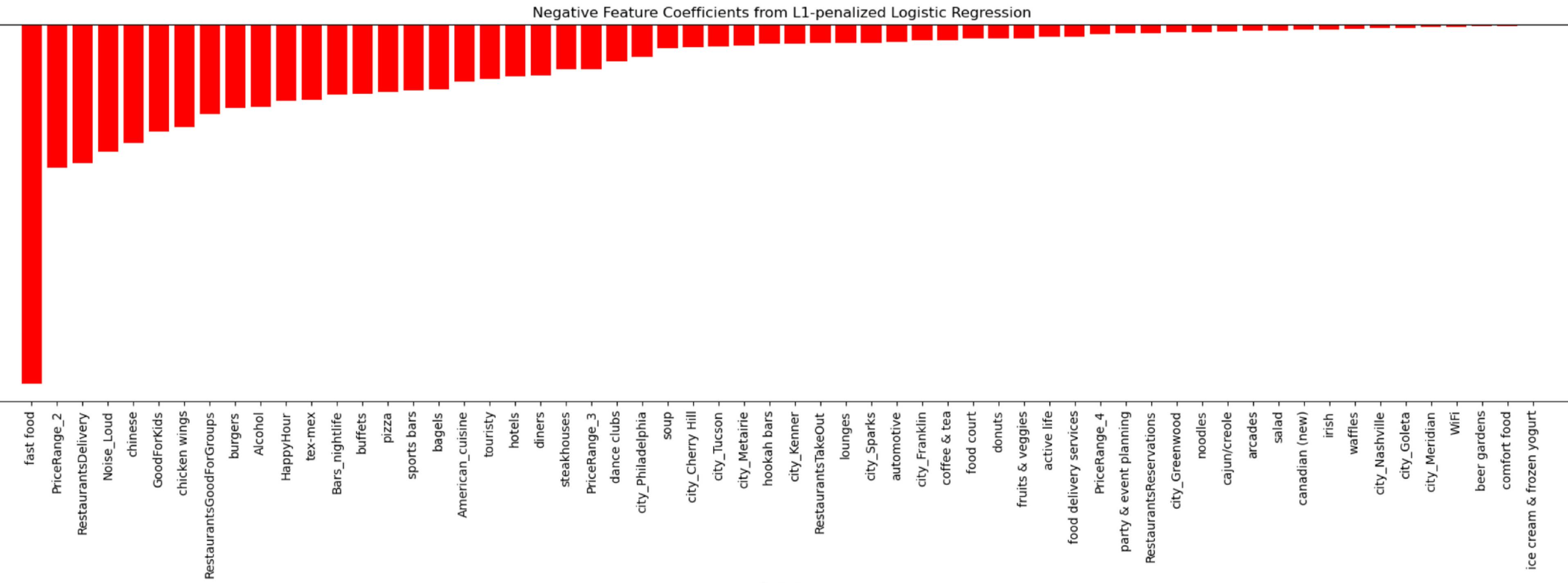
# Top Positive Coefficiencies

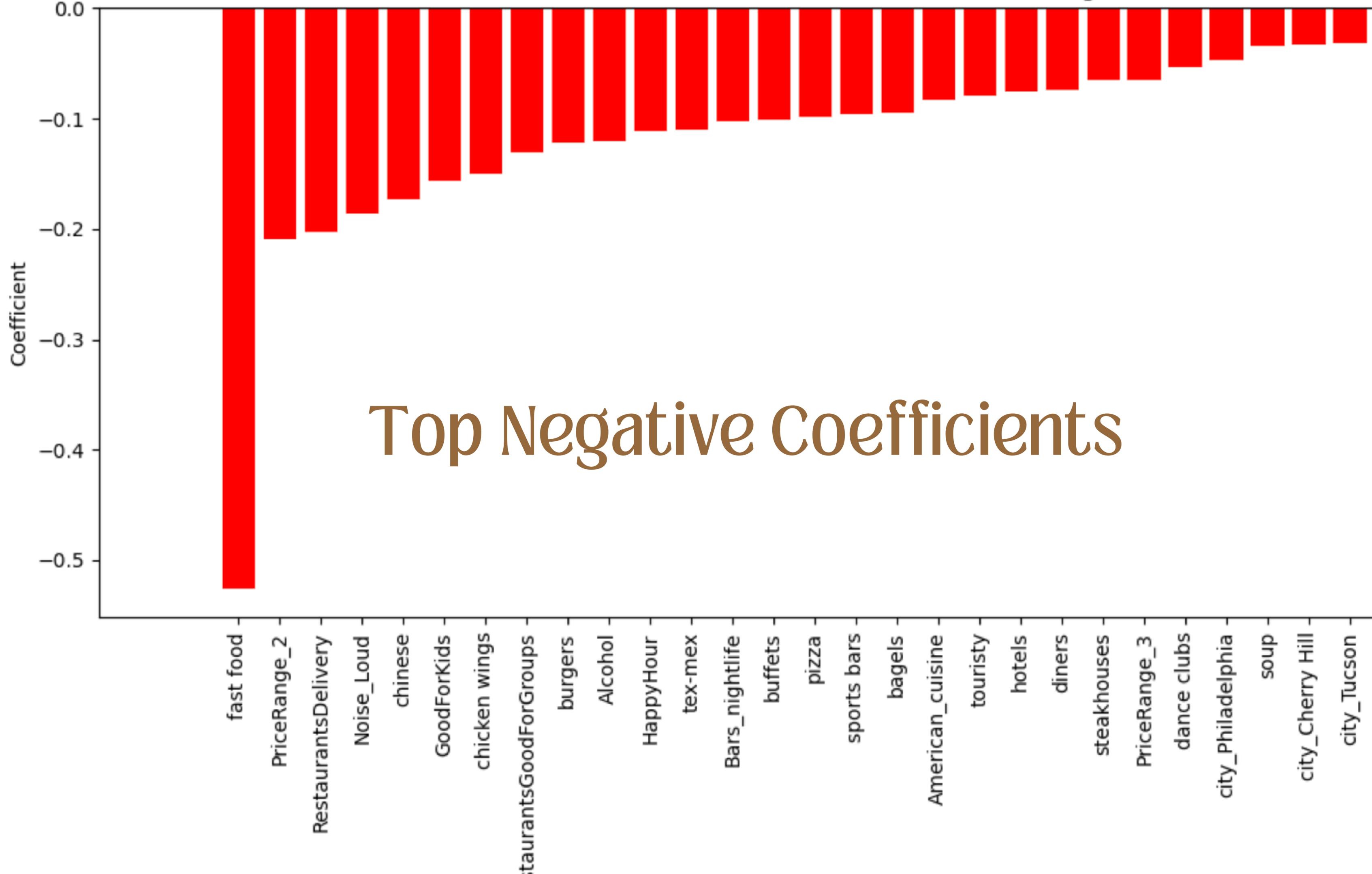


# Less Positive Coefficiencies



# Negative Feature Coefficients from Logistic Regression





from L1-penalized Logistic Regression



# Nagative near to zero Coefficiency

hookah bars	-
city_Kenner	-
RestaurantsTakeOut	-
lounges	-
city_Sparks	-
automotive	-
city_Franklin	-
coffee & tea	-
food court	-
donuts	-
fruits & veggies	-
active life	-
food delivery services	-
PriceRange_4	-
party & event planning	-
RestaurantsReservations	-
city_Greenwood	-
noodles	-
cajun/creole	-
arcades	-
salad	-
canadian (new)	-
irish	-
waffles	-
city_Nashville	-
city_Goleta	-
city_Meridian	-
WiFi	-
beer gardens	-
comfort food	-
ice cream & frozen yogurt	-

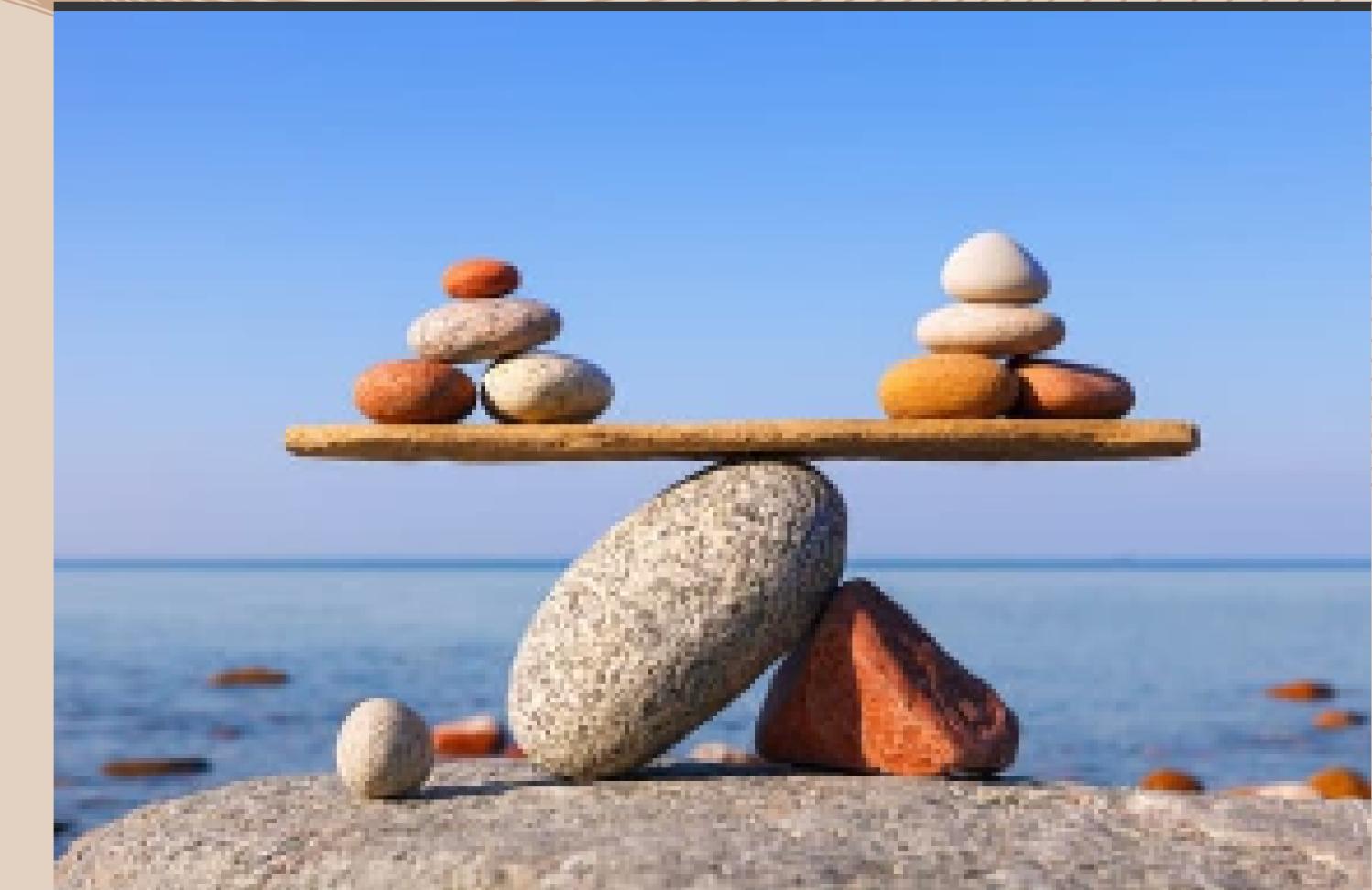
from L1-penalized Logistic Regression



# Nagative near to zero Coefficiency

hookah bars	-
city_Kenner	-
RestaurantsTakeOut	-
lounges	-
city_Sparks	-
automotive	-
city_Franklin	-
coffee & tea	-
food court	-
donuts	-
fruits & veggies	-
active life	-
food delivery services	-
PriceRange_4	-
party & event planning	-
RestaurantsReservations	-
city_Greenwood	-
noodles	-
cajun/creole	-
arcades	-
salad	-
canadian (new)	-
irish	-
waffles	-
city_Nashville	-
city_Goleta	-
city_Meridian	-
WiFi	-
beer gardens	-
comfort food	-
ice cream & frozen yogurt	-

# Logistic Regression & Power of Scaling



Train Accuracy: 0.4460101679929266

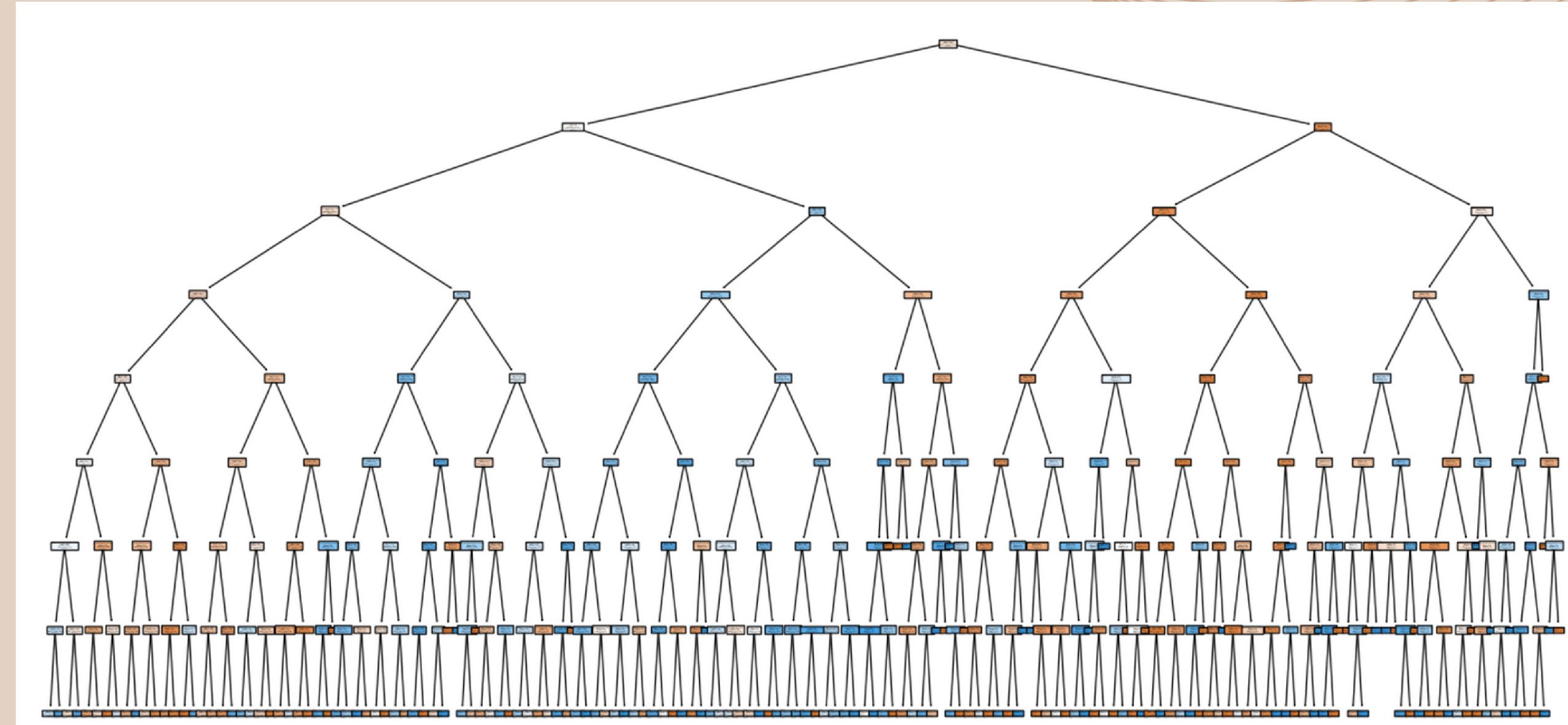
Test Accuracy: 0.44445877119463606

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

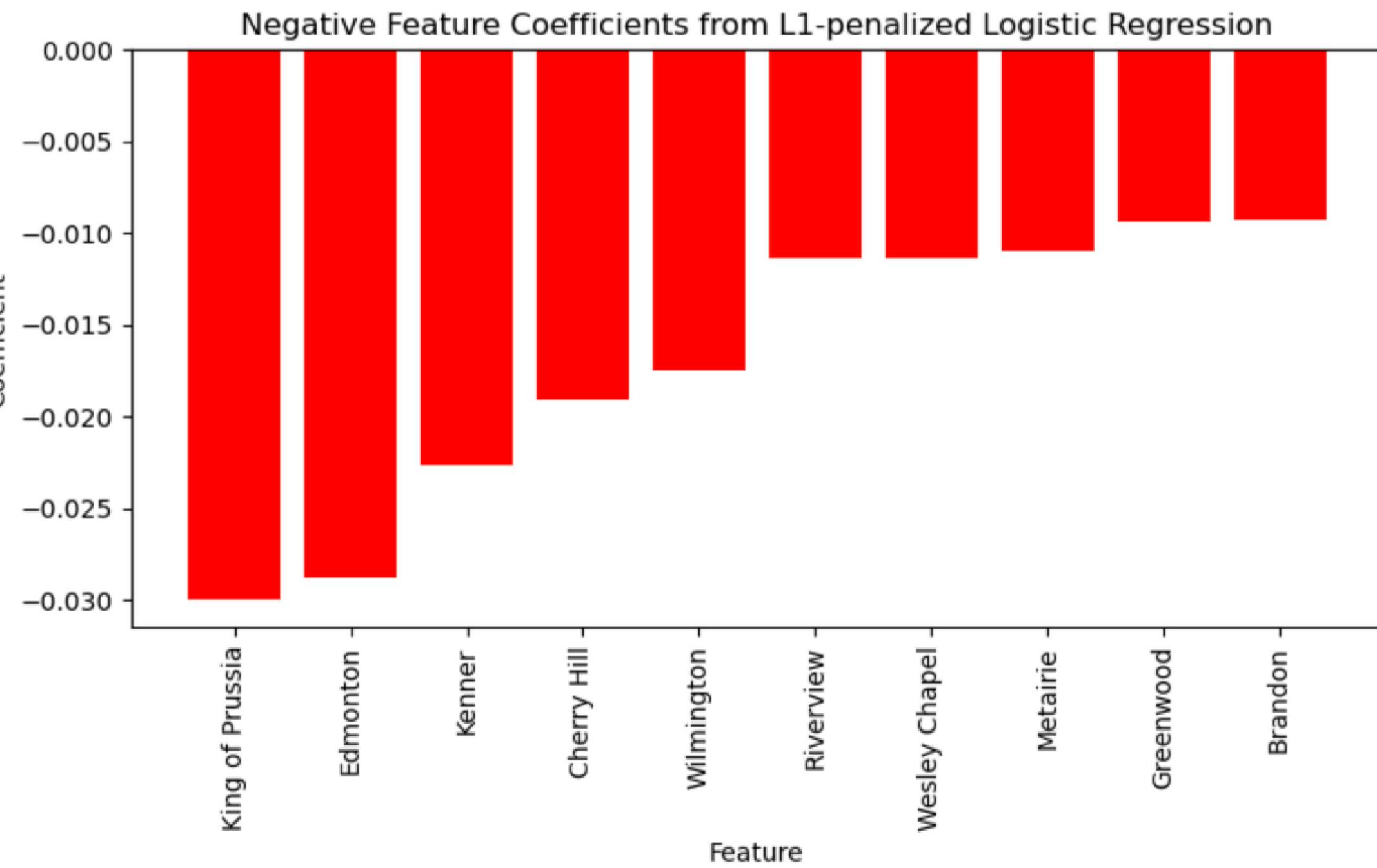
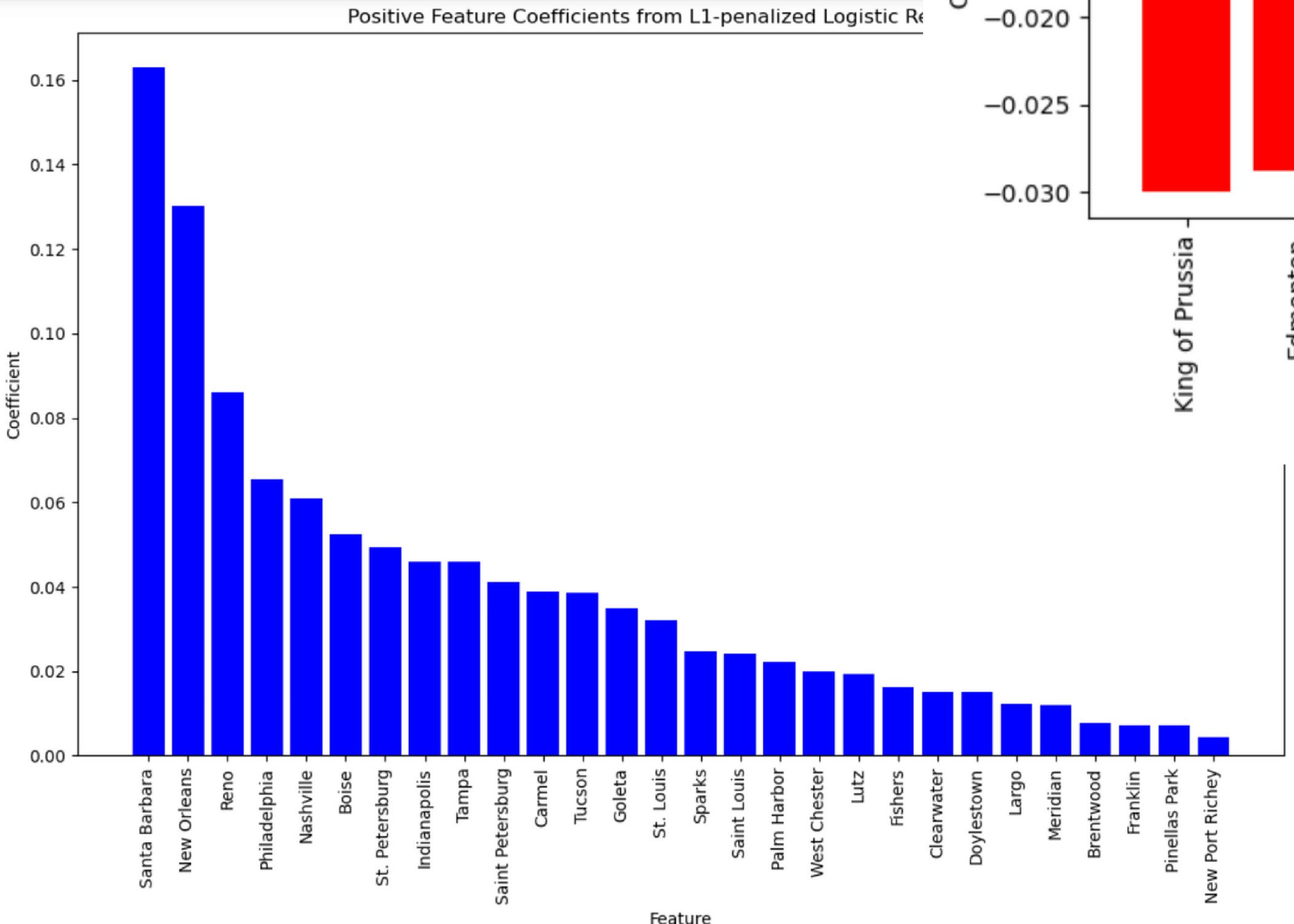
DecisionTree Scaled Train Accuracy: 0.6930813439434129

DecisionTree Scaled Test Accuracy: 0.671974727612662

# Is City a Bad Player !?



# City & Target



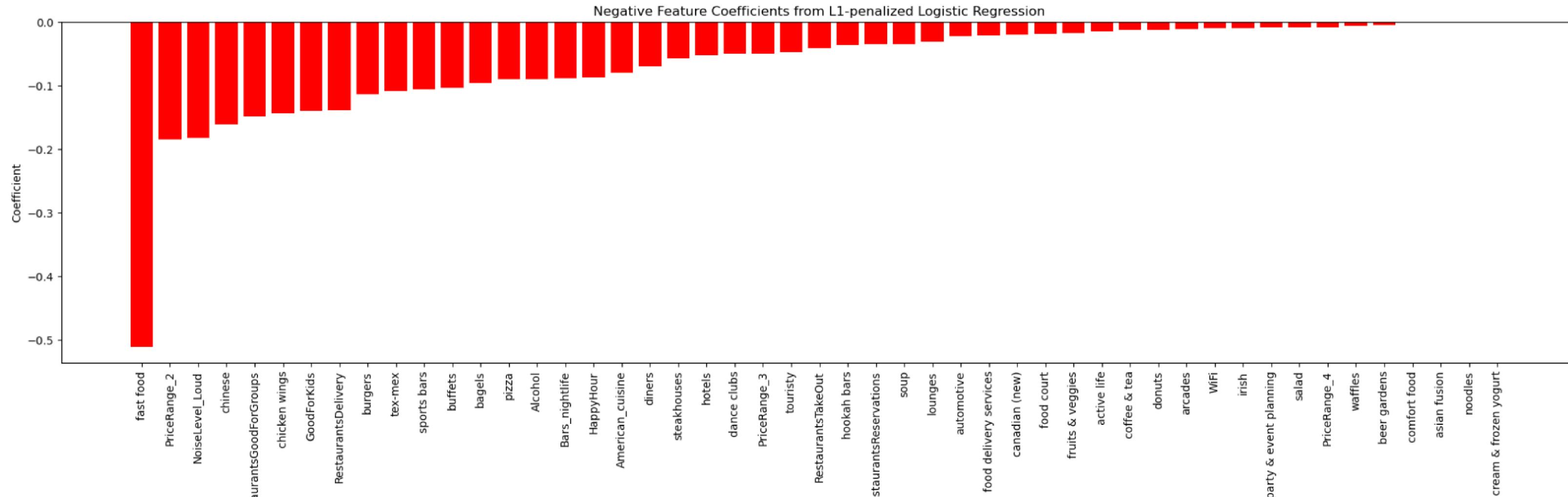
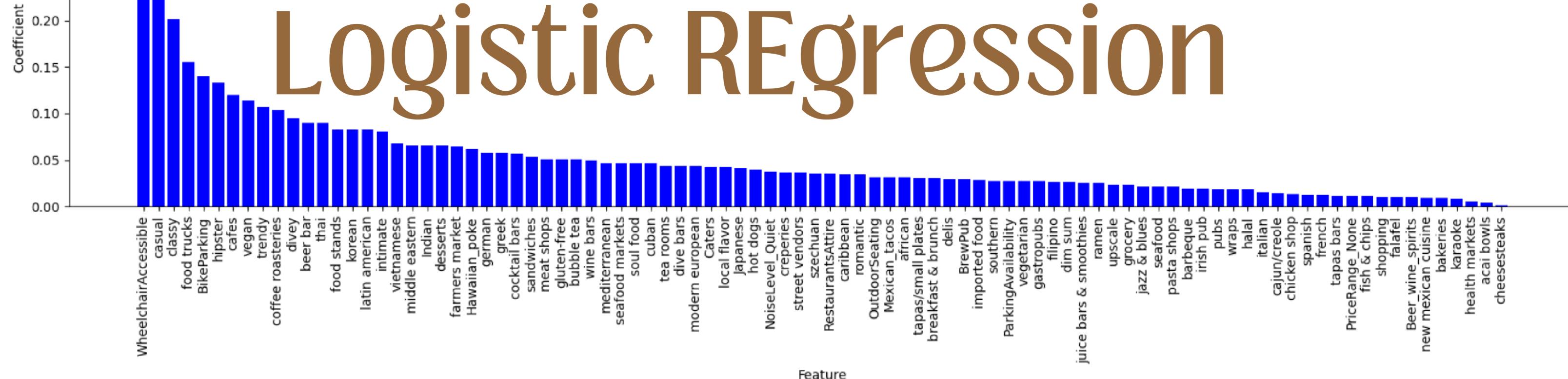
Train Accuracy: 0.54158985956177

Test Accuracy: 0.537646328485278

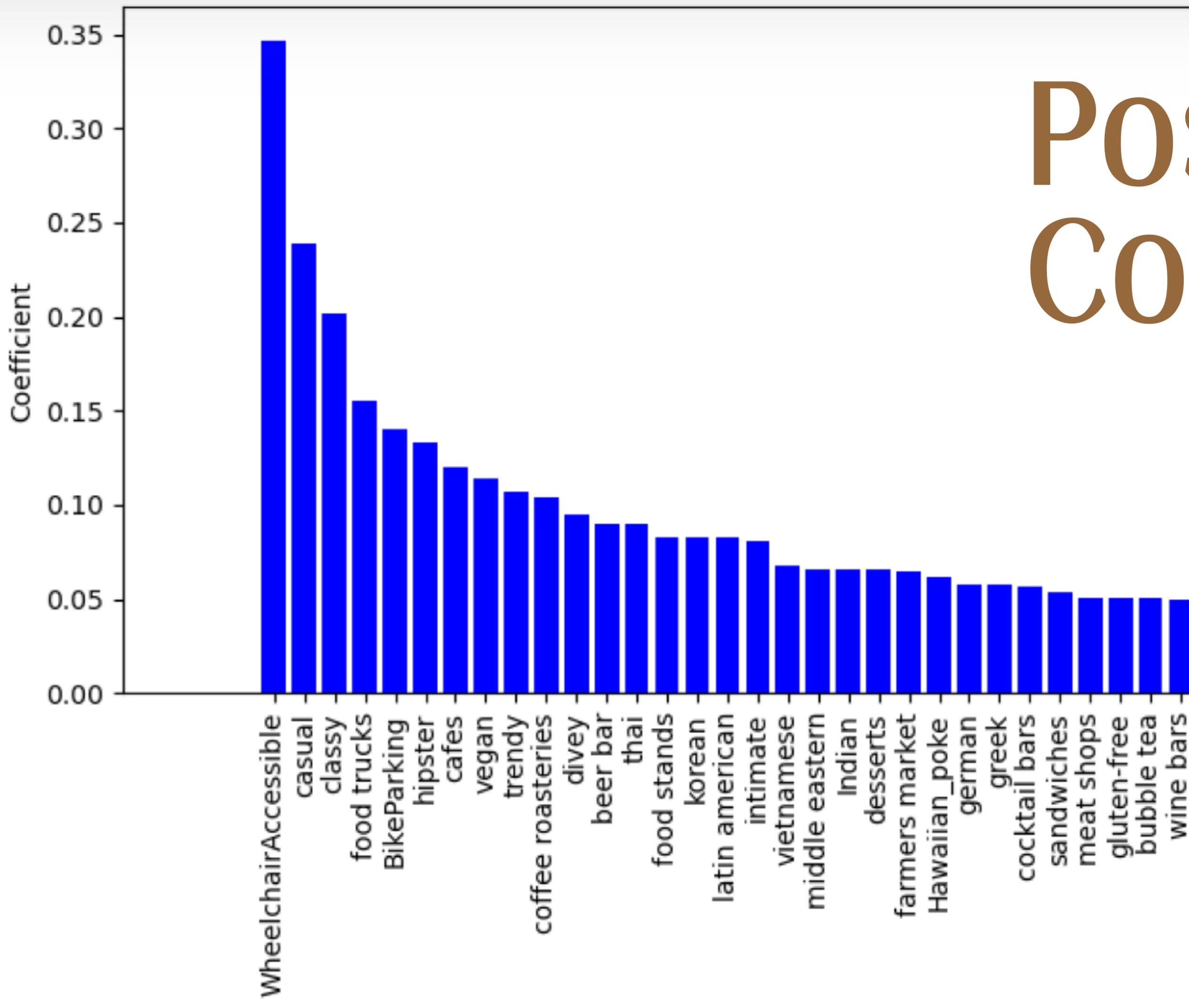
Scaled Train Accuracy: 0.5433097

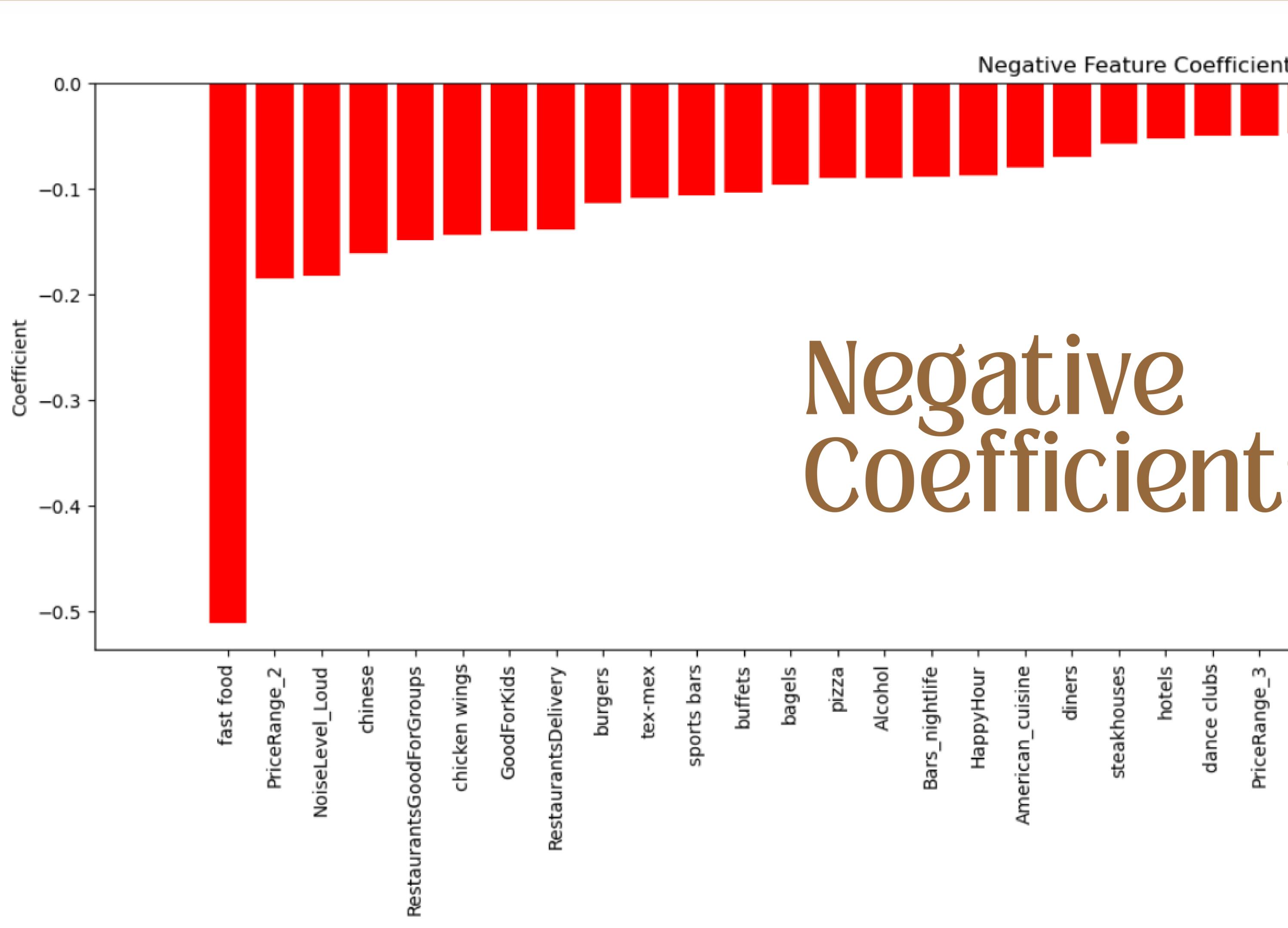
Scaled Test Accuracy: 0.53735810

# Feature Coefficients from Logistic Regression



# Positive Coefficient





Negative  
Coefficients

# Logistic Regression

Train Accuracy: 0.6115992484526968

Test Accuracy: 0.6093095222745148

Scaled Train Accuracy: 0.7065097259062776

Scaled Test Accuracy: 0.6960221778092966

# Decision Tree

DecisionTree Train Accuracy: 0.6218225022104332

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