PREDICTING RESTAURANT RATINGS IN PHILADELPHIA

CAPSTONE PROJECT

27/01/21

RESTAURANT SCENE IN THE US

- Restaurant market size in US is $800 billion
- One of the county’s top employers
- High failure rate among new restaurants
- Severely affected by COVID-19

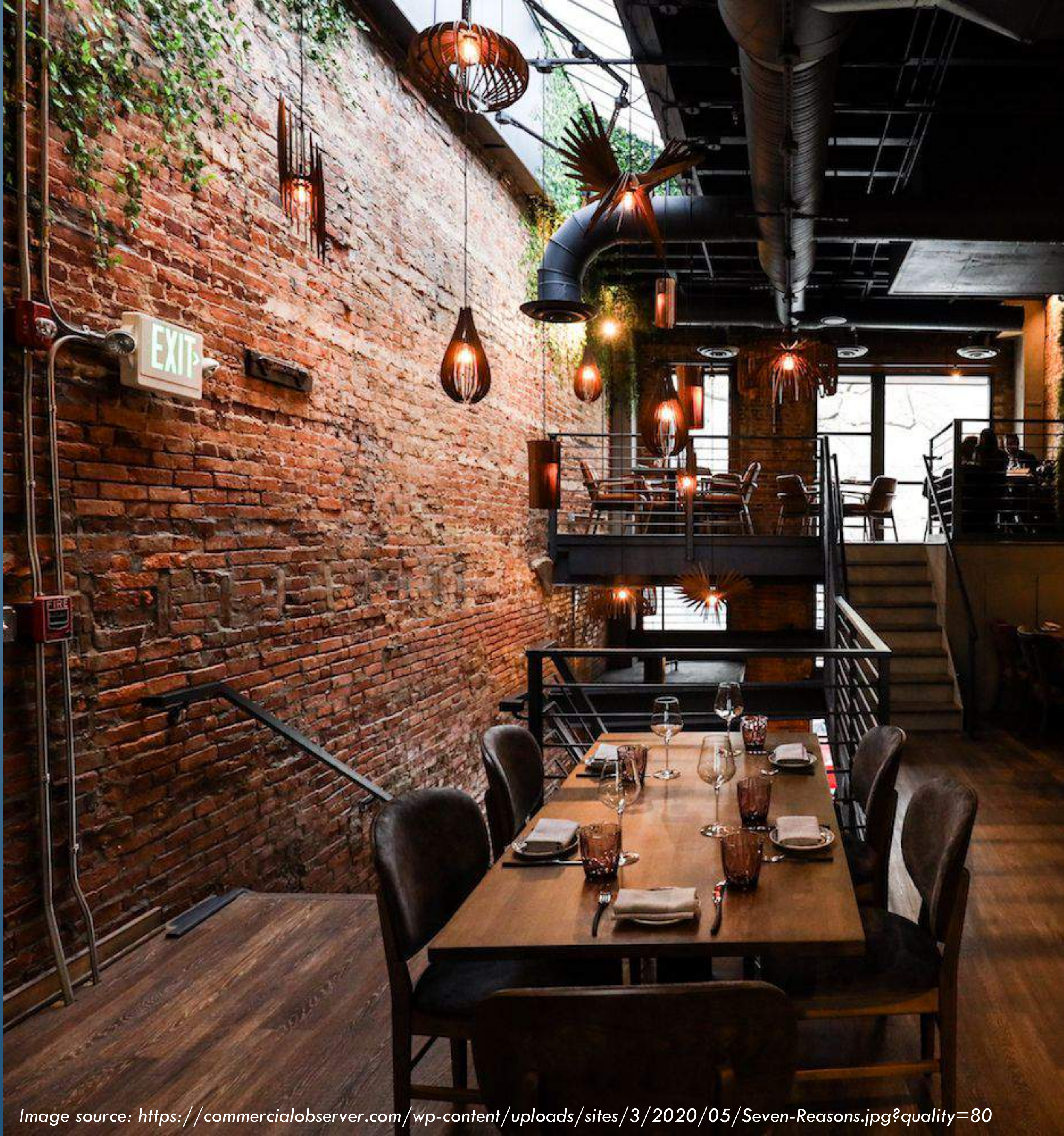


Image source: https://commercialobserver.com/wp-content/uploads/sites/3/2020/05/Seven-Reasons.jpg?quality=80

PURPOSE - PREDICT RESTAURANT RATINGS IN PHILADELPHIA

- Create a model that predicts rating using readily available data
- Yield insight regarding drivers of customer satisfaction factors
- Existing restaurants: revitalize / improve / adapt
- New restaurants: learn from successful examples

DATA ACQUISITION

Philly zip codes and geometry obtained from opendataphilly.org → 48 Zip codes and their locations

Restaurant IDs obtained from foursquare’s endpoint *explore* → 1460 unique restaurants (rows)

Detailed restaurant information obtained from foursquare’s endpoint *venue details* → 118 variables (columns)

FINAL DATASET

Cleaning:

- Removed imputable missing values
- Transformed into numeric values
- Created new columns (distance center, dummy variables)

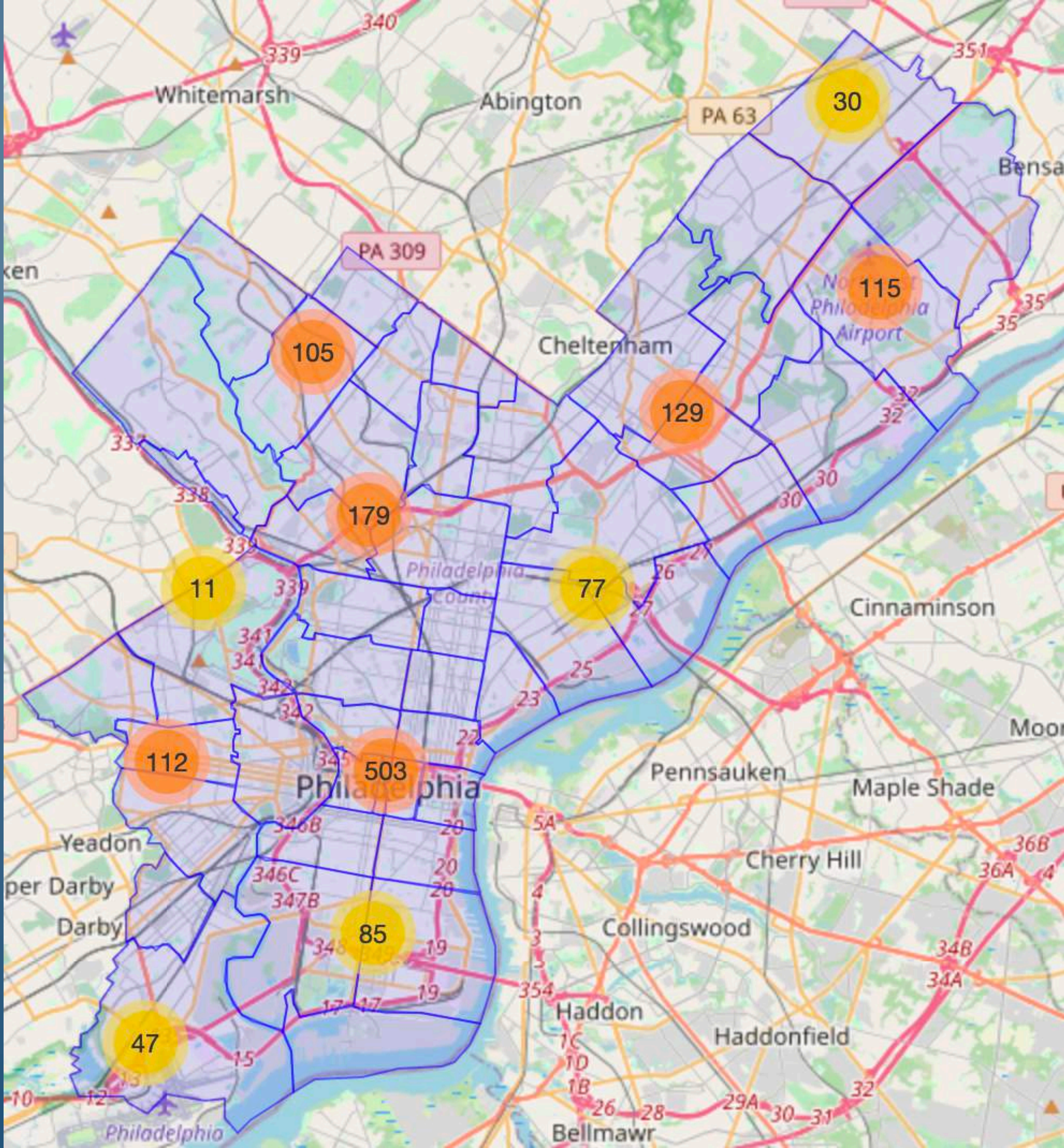
Final dataframe shape 1393 rows, 118 columns:

- 9 general data columns
- 32 attribute columns (Wi-Fi, Alcohol, Credit Cards etc.)
- 76 category columns (Fast Food, Sushi, Chinese etc.)

ID	Name	Zip	Lat	Long	Category	Distance	Rating	Price Tier
4ab2ac0bf964a520d66b20e3	Del Frisco's Double Eagle Steak House	19102	39.9510	-75.1655	Steakhouse	0.00234	8.7	4.0
4a281e64f964a520f4941fe3	Oyster House	19102	39.9504	-75.1665	Seafood Restaurant	0.00353	9.3	3.0
56cc831ccd10c5927d30dc1d	Snap Custom Pizza	19102	39.9504	-75.1662	Pizza Place	0.00322	8.8	1.0
4af2d4cef964a520a9e821e3	The Capital Grille	19107	39.9507	-75.1639	American Restaurant	0.00170	8.5	4.0
4a4268fdf964a520d4a51fe3	Fogo De Chão	19107	39.9509	-75.1630	Churrascaria	0.00160	8.8	4.0

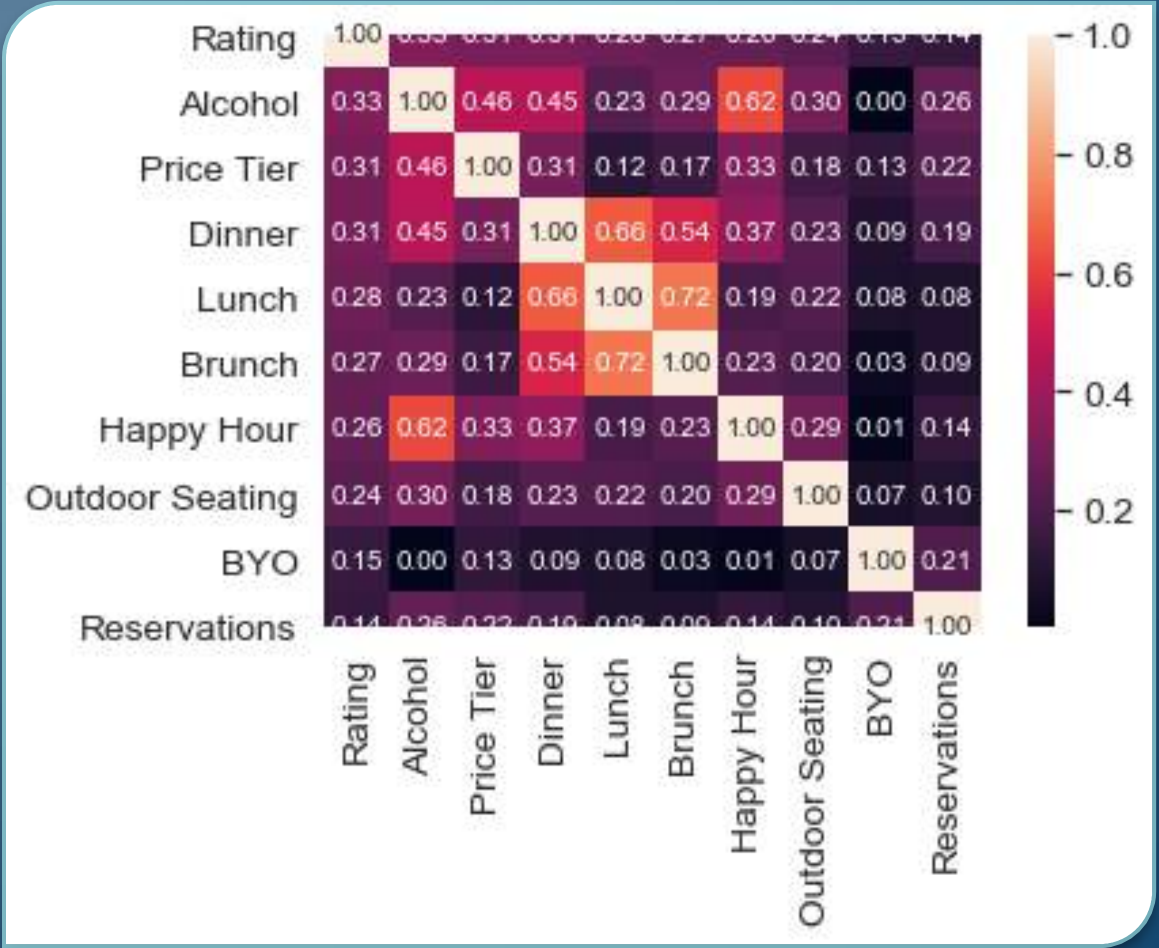
DATA VISUALIZATION: *RESTAURANT LOCATIONS*

- Aspired to collect an even number of restaurants/zip
- Density in center city due to smaller zip codes



DATA VISUALIZATION: CORRELATIONS WITH RATINGS

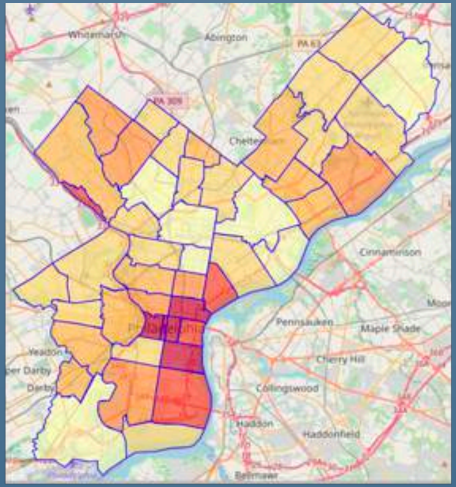
1. Alcohol
2. Price Tier
3. Dinner
4. Lunch



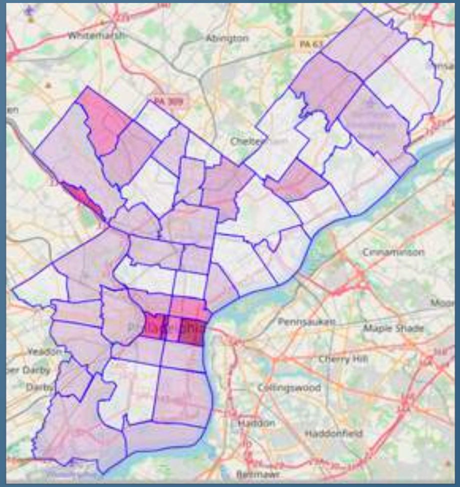
	Rating	Alcohol	Price Tier	Dinner	Lunch	Brunch	Happy Hour	Outdoor Seating	BYO	Reservations
Rating	1.00	0.33	0.31	0.31	0.28	0.27	0.26	0.24	0.15	0.14
Alcohol	0.33	1.00	0.46	0.45	0.23	0.29	0.62	0.30	0.00	0.26
Price Tier	0.31	0.46	1.00	0.31	0.12	0.17	0.33	0.18	0.13	0.22
Dinner	0.31	0.45	0.31	1.00	0.66	0.54	0.37	0.23	0.09	0.19
Lunch	0.28	0.23	0.12	0.66	1.00	0.72	0.19	0.22	0.08	0.08
Brunch	0.27	0.29	0.17	0.54	0.72	1.00	0.23	0.20	0.03	0.09
Happy Hour	0.26	0.62	0.33	0.37	0.19	0.23	1.00	0.29	0.01	0.14
Outdoor Seating	0.24	0.30	0.18	0.23	0.22	0.20	0.29	1.00	0.07	0.10
BYO	0.15	0.00	0.13	0.09	0.08	0.03	0.01	0.07	1.00	0.21
Reservations	0.14	0.26	0.22	0.19	0.08	0.09	0.14	0.10	0.21	1.00

DATA VISUALIZATION: *CHOROPLETHS*

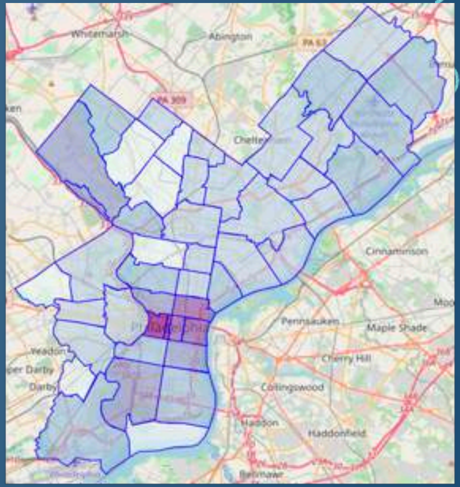
- Areas around center and west have higher more popular restaurants
- Areas around center and south have more highly rated restaurants
- Areas around center are more expensive
- Delivery is slightly more common further from center
- Areas around center have more restaurants offering alcohol and dinner



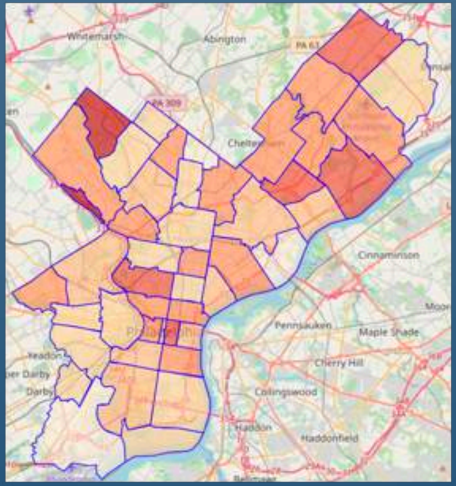
RATING



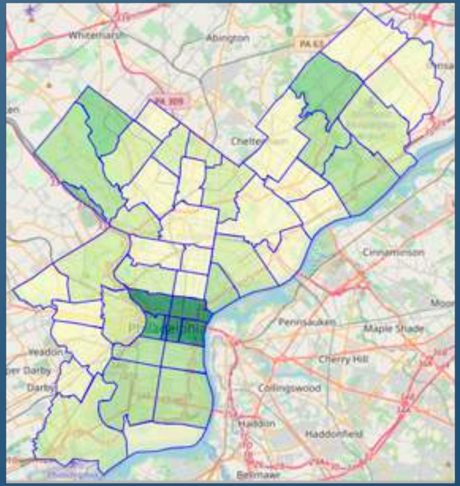
PRICE TIER



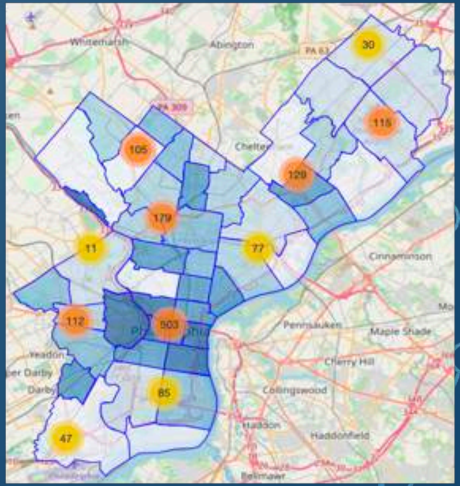
DINNER



DELIVERY



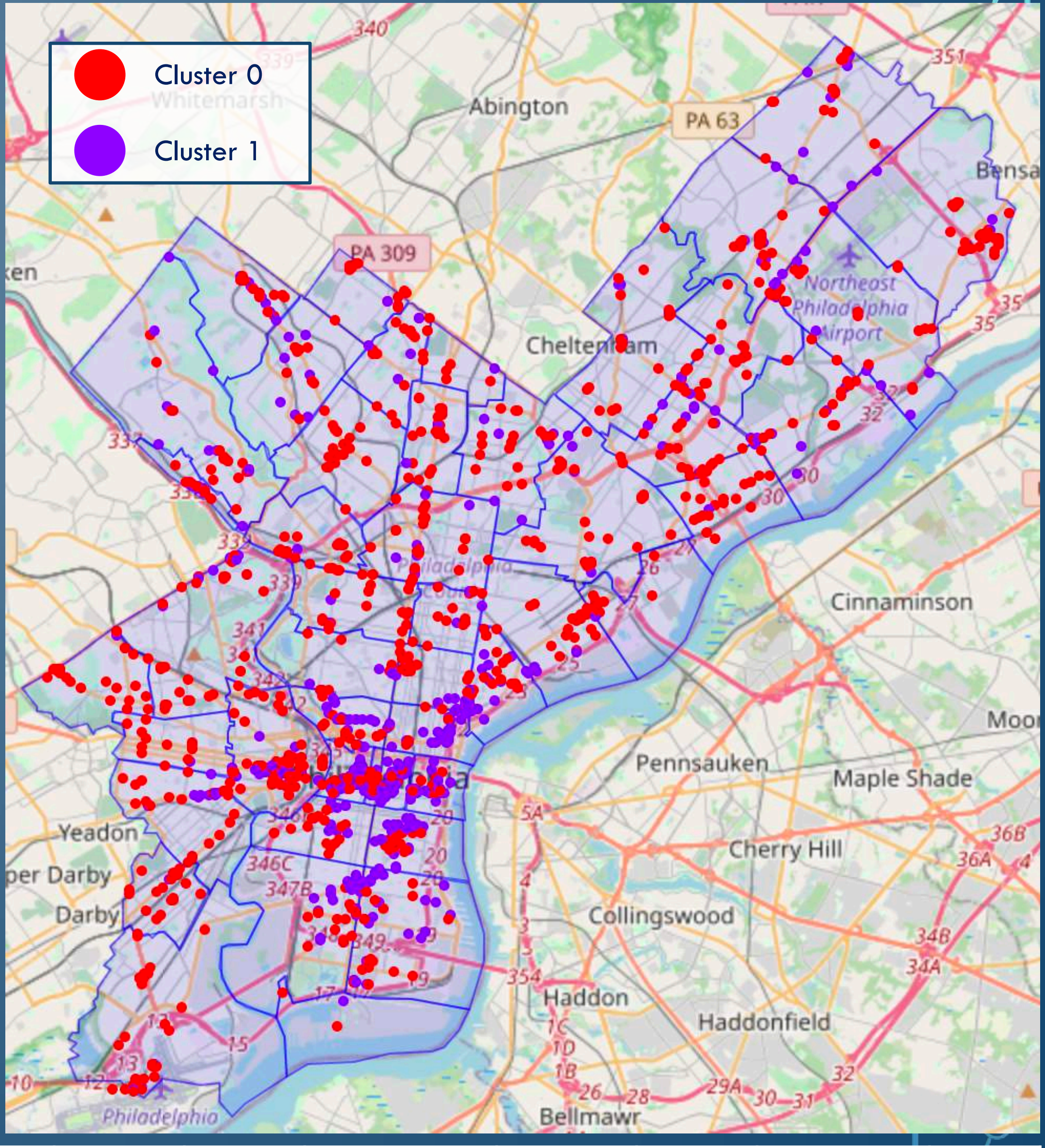
ALCOHOL



DENSITY

MODELS: *K MEANS CLUSTERING*

- High separation on rating and attributes important to rating
- Cluster 1 represents high rated restaurants
- Cluster 1 mostly around center



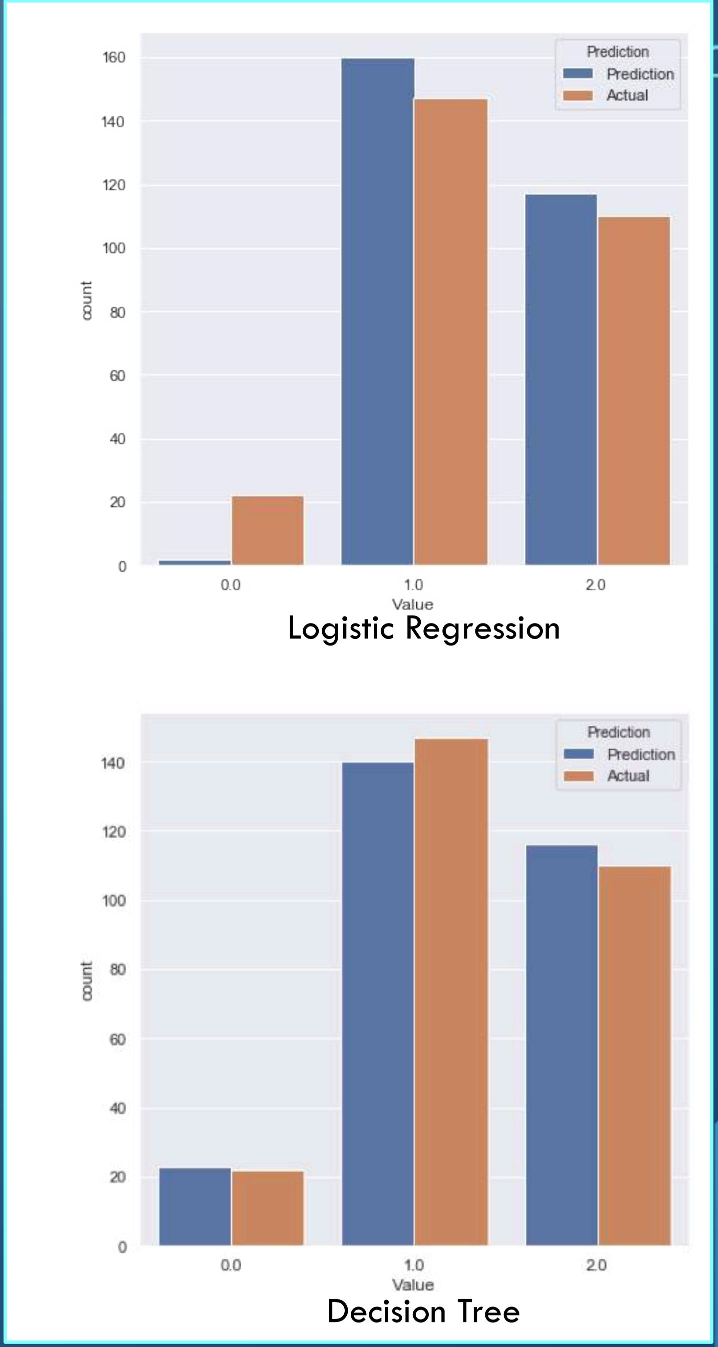
Clst	Distance	Rating	Price Tier	Delivery	Alcohol	Credit Cards	Outdoor Seating	Wi-Fi	Breakfast	Brunch
0	0.0894	7.0926	1.2323	0.3732	0.0293	0.3720	0.0633	0.0633	0.1725	0.0892
1	0.0541	8.2338	1.8687	0.5027	0.3900	0.7264	0.3216	0.2033	0.1959	0.5693

MODELS: CLASSIFIERS

	Accuracy	F1 Score	Jaccard Score	Precision	Recall
SVM	0.7025	0.7155	0.5674	0.7512	0.7025
DTree	0.7491	0.7492	0.6104	0.7503	0.7491
Log Reg	**0.7706**	**0.7449**	**0.6196**	**0.7512**	**0.7706**
Ridge Reg	0.6667	0.6856	0.5339	0.7376	0.6667

COUNT PLOT OF PREDICTIONS VS. ACTUALS

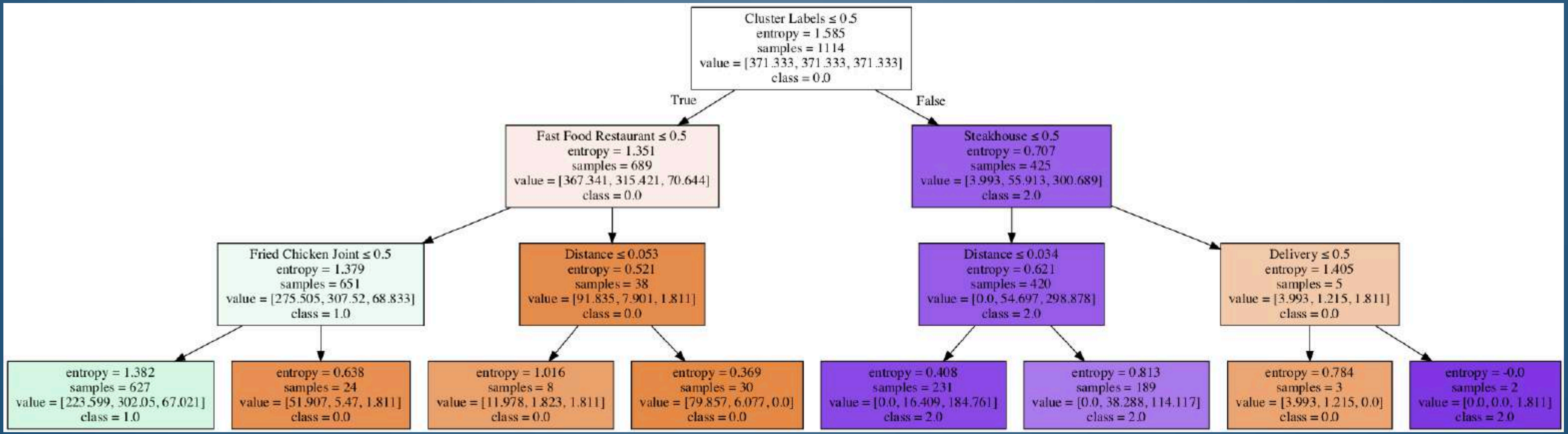
- Decision tree handled class imbalance better despite lower scores
- Distribution of predictions more closely matches distribution of actual values



Logistic Regression

Decision Tree

DECISION TREE 1: WITH CATEGORIES



Cluster Labels ≤ 0.5
entropy = 1.585
samples = 1114
value = [371.333, 371.333, 371.333]
class = 0.0

True

False

Fast Food Restaurant ≤ 0.5
entropy = 1.351
samples = 689
value = [367.341, 315.421, 70.644]
class = 0.0

Steakhouse ≤ 0.5
entropy = 0.707
samples = 425
value = [3.993, 55.913, 300.689]
class = 2.0

Fried Chicken Joint ≤ 0.5
entropy = 1.379
samples = 651
value = [275.505, 307.52, 68.833]
class = 1.0

Distance ≤ 0.053
entropy = 0.521
samples = 38
value = [91.835, 7.901, 1.811]
class = 0.0

Distance ≤ 0.034
entropy = 0.621
samples = 420
value = [0.0, 54.697, 298.878]
class = 2.0

Delivery ≤ 0.5
entropy = 1.405
samples = 5
value = [3.993, 1.215, 1.811]
class = 0.0

entropy = 1.382
samples = 627
value = [223.599, 302.05, 67.021]
class = 1.0

entropy = 0.638
samples = 24
value = [51.907, 5.47, 1.811]
class = 0.0

entropy = 1.016
samples = 8
value = [11.978, 1.823, 1.811]
class = 0.0

entropy = 0.369
samples = 30
value = [79.857, 6.077, 0.0]
class = 0.0

entropy = 0.408
samples = 231
value = [0.0, 16.409, 184.761]
class = 2.0

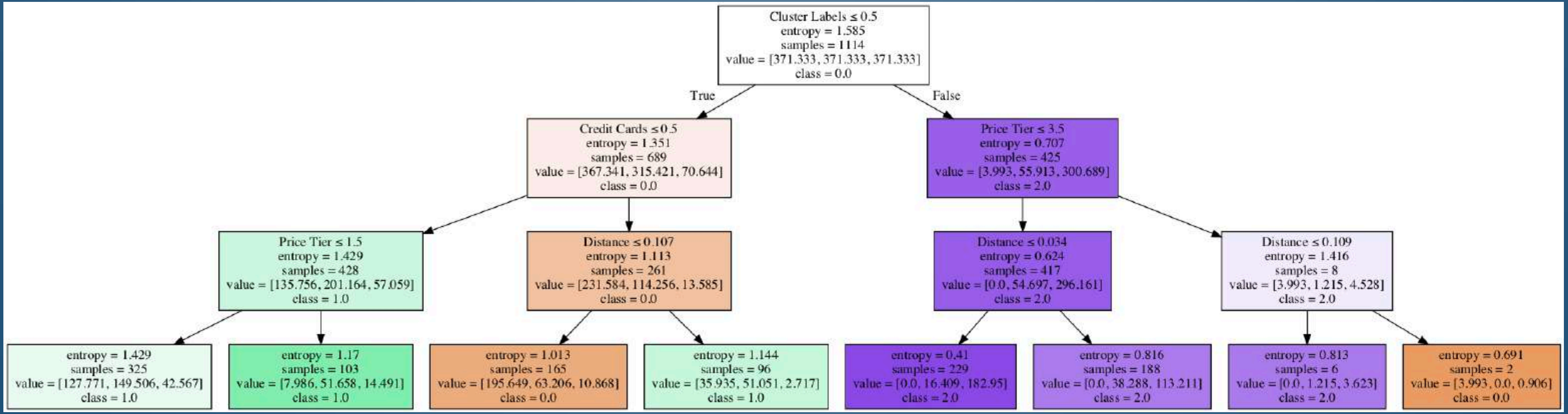
entropy = 0.813
samples = 189
value = [0.0, 38.288, 114.117]
class = 2.0

entropy = 0.784
samples = 3
value = [3.993, 1.215, 0.0]
class = 0.0

entropy = -0.0
samples = 2
value = [0.0, 0.0, 1.811]
class = 2.0

- Cluster Label leads to highest information gain
- Steakhouse – highest rated category
- Fast Food & Fried Chicken lowest rated categories
- Overfitting? Our data only featured highly rated steakhouses, but surely this isn't representative of reality.

DECISION TREE 2: W/O CATEGORIES



CONCLUSION

- (Multiclass) Logistic regression model achieved 77% accuracy
- Important factors for rating: Price Tier, Alcohol, Distance, Dinner

- Sparsity and missing values affected insights
- Probably would have to supplement with more complete and qualitative data for future study