For this project I analyzed the correlation of certain words with more expensive menu prices, specifically at places that serve pizza.

According to my results, the words used to predict the prices made a lot of sense to me. Restaurants with “resort” and “spa” in the name are likely to be expensive, as those places are often more upscale. Italian words such as “il”, “cucina” and “brio” also indicate it will be more expensive.

The words used to predict the prices of cheaper restaurants are words I associate with being casual, such as “deli”, “fried” and “bar.”

For this final week I added step 17 to determine the best model.
Some surprises:

- I was able to get a pretty clear distinction in the most important words predicting expensive and cheap restaurants, despite quite a small dataset. It would be cool to run this on a much larger dataset.
- The 18th most common restaurant name in my dataset is called “7 Day 24 Hours Emergency Locks.” I have never seen a locksmith that served pizza, but apparently, this is a thing.

Here are my steps for completing the graph analysis:

1. Load data and create the dataframe
2. Check dataframe dimensions
3. Examine the variables and their types
4. Draw histograms of appropriate variables
5. Visualize the zip codes of the restaurants in my data using folium
6. Use agate to determine outliers in price columns (> 3 std deviations from the mean)
7. View the distribution of the length of each restaurant name using a histogram
8. Remove stop words (Step added week 9 based on feedback.)
9. View the most common restaurant names, as well as the distribution of all words used
10. Bar plot the 20 most common words in order to visualize them for better understanding
11. Drop the rows that contain duplicate restaurant names
12. Find the midpoint price range for each restaurant and transform into target
13. Use TFIDF-Vectorizer on restaurant names to create feature variables
14. I am using the Random Forest algorithm so that I can view which words contribute the most to the decision. Use one-vs-rest classifier and gridsearchCV to evaluate model performance and determine the best hyperparameters
15. Train Random Forest algorithm using best hyperparameters, and use scikit-learns random forest method of .feature_importances_ to find the words that contribute most to classifying a restaurant as expensive (mid-point > $40)
16. Repeat steps 13 and 14 to find words that contribute to classifying a restaurant as cheap (mid-point < $15)
17. Select Best Model from Multiple Learning Algorithms (New step added this week.)

The dimension of the table is: (3510, 21)

	id	...	province
0	AVwc_6KEIN2L1WUfrKAH	...	OR
1	AVwc_6KEIN2L1WUfrKAH	...	OR
2	AVwc_6qRByjofQCxkcxw	...	Brentwood
3	AVwc_6qRByjofQCxkcxw	...	Brentwood
4	AVwc_6qRByjofQCxkcxw	...	Brentwood

[5 rows x 21 columns]

Describe Data

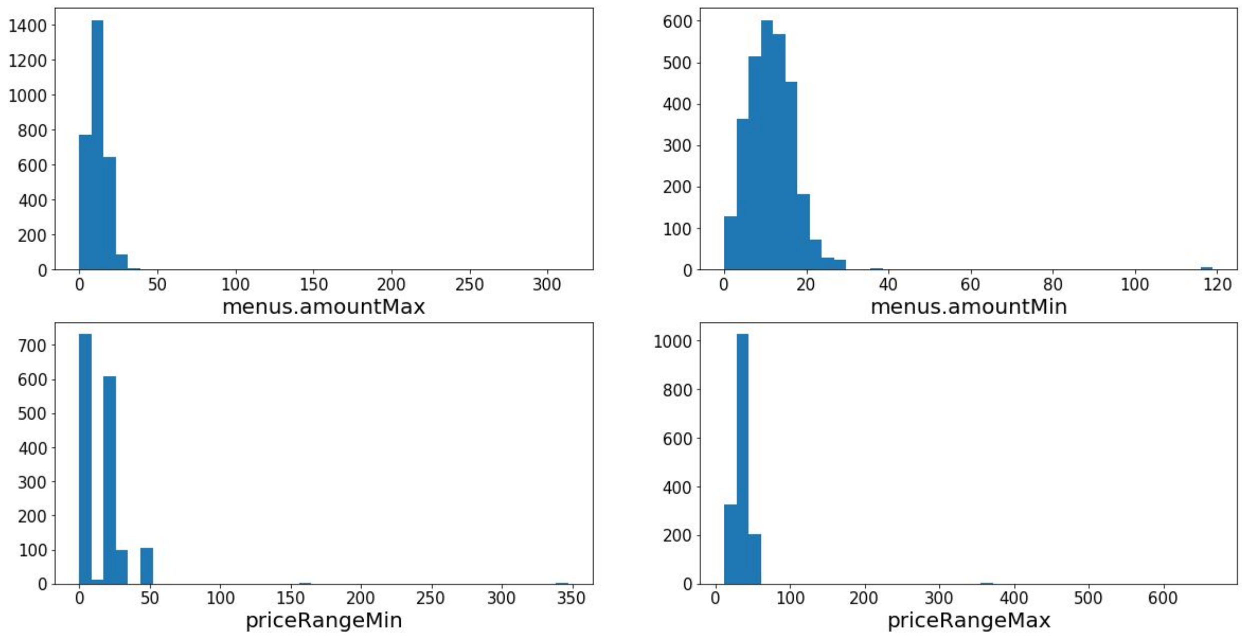
	latitude	longitude	...	priceRangeMin	priceRangeMax
count	3510.000000	3510.000000	...	1557.000000	1557.000000
mean	38.555114	-87.472055	...	15.597945	36.566474
std	4.651092	16.430008	...	18.495854	21.737839
min	18.411826	-157.837461	...	0.000000	12.000000
25%	35.769852	-94.202573	...	0.000000	30.000000
50%	40.020710	-81.675414	...	25.000000	40.000000
75%	41.455179	-74.743820	...	25.000000	40.000000
max	64.854370	-66.024871	...	347.000000	666.000000

[8 rows x 6 columns]

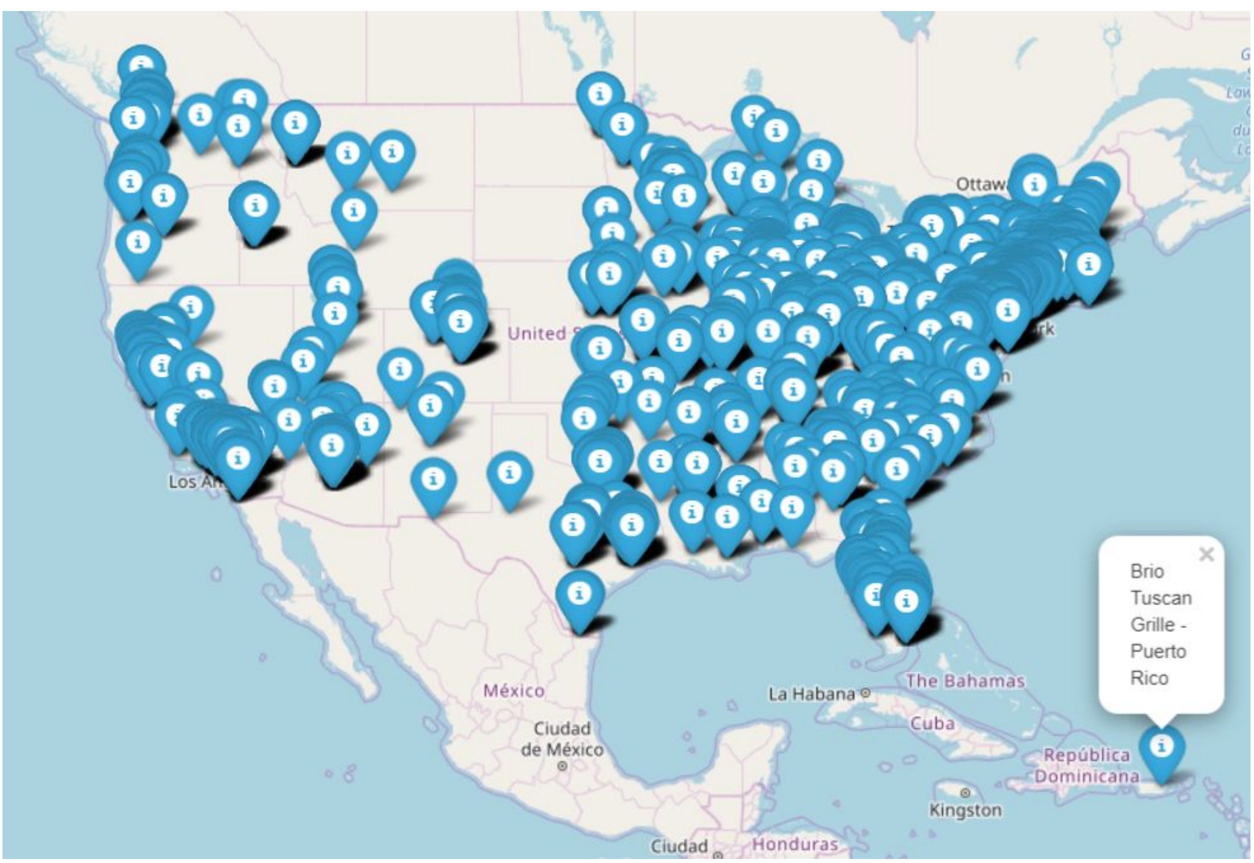
Summarized Data

	id	address	...	priceRangeCurrency	province
count	3510	3510	...	1557	3510
unique	989	984	...	1	281
top	AVwdIsuzkufWRAb52p9M	1605 Kanawha Blvd W	...	USD	CA
freq	64	64	...	1557	256

[4 rows x 15 columns]



Menus.amountMin / Max = Price range for specific menu items that I have data for. (There may be multiple menu items per restaurant.)
priceRangeMin / Max = Price range for the restaurant as a whole.

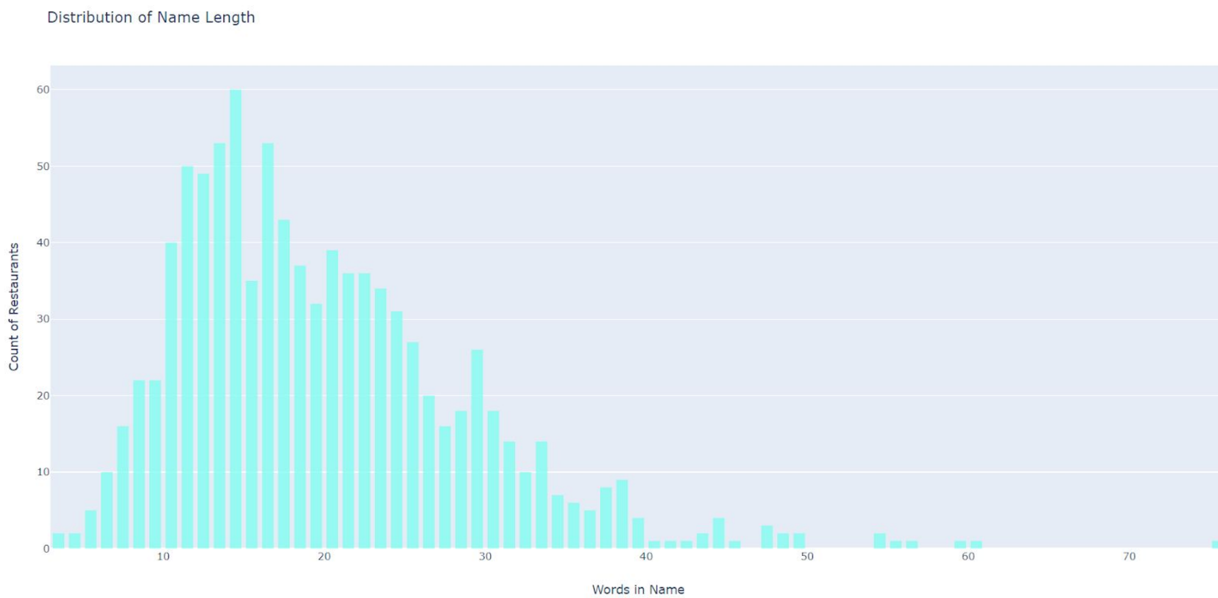


menus.amountMax: 11 Outliers
Mean 12.479
116.99
116.99
116.99
118.99
100.0
116.99
312.95
310.95
311.95
312.95
69.95

menus.amountMin: 14 Outliers
Mean 11.427
37.99
116.99
116.99
116.99
118.99
100.0
116.99
35.99
36.99
47.5
39.99
50.99
44.0
69.95

priceRangeMin: 3 Outliers
Mean 15.597
164.0
164.0
347.0

priceRangeMax: 3 Outliers
Mean 36.566
363.0
363.0
666.0

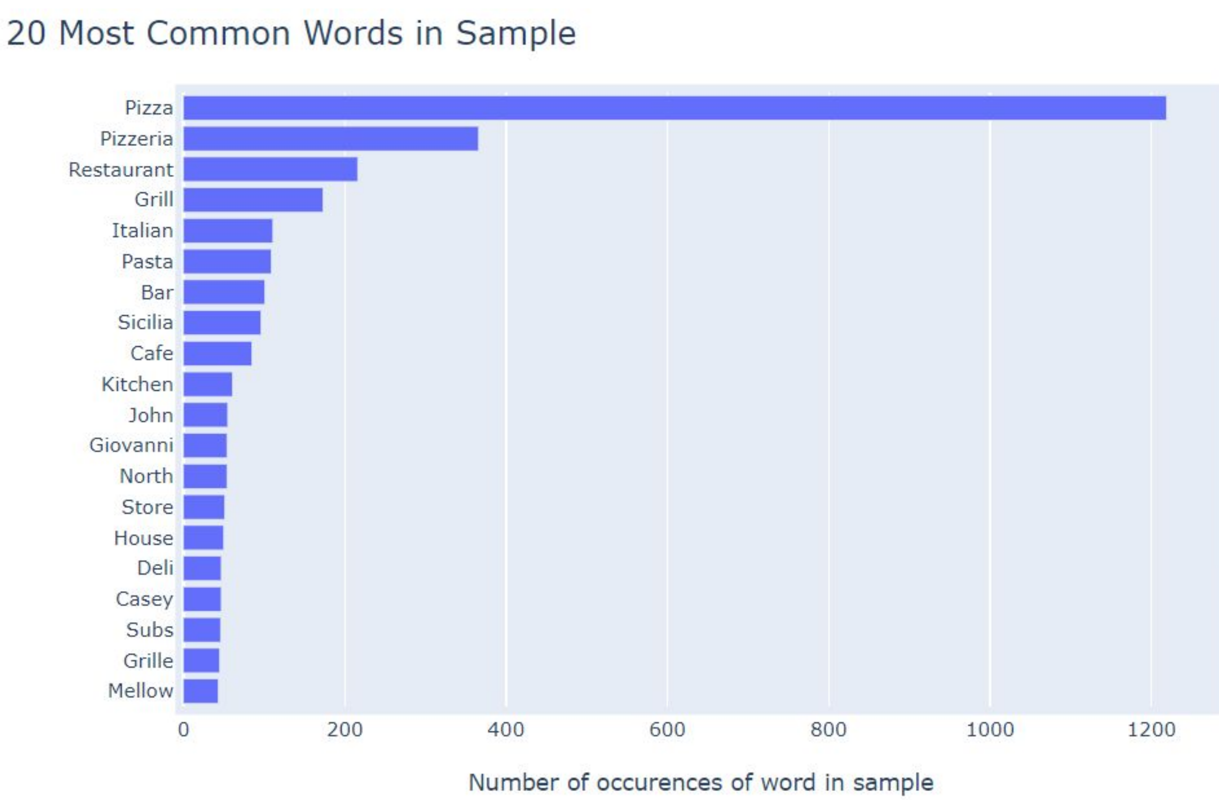


The most common restaurants in the dataset (with counts) are:
[('Sicilia Pizzeria', 96), ('J G Restaurant', 55), ('Casey General Store', 43), ('Pizza Joint', 36), ('North End Pizzeria', 34), ('Labella Pizza Pasta', 31), ('Giovanni Pizzeria', 30), ('Nino Trattoria Pizzeria', 28), ('Papa John Pizza', 27), ('Takka Grill', 26), ('Marco Pizza', 26), ('Stone Paddle', 24), ('Hungry Howie Pizza', 22), ('Original Giorgio', 20), ('Palace Pizza Bartow', 20), ('Pronto Pizza', 19), ('Bertucci', 19), ('7 Day 24 Hours Emergency Locks', 18), ('Ameci Pizza Pasta', 18), ('Valentino Pizza', 18)]

The number of unique restaurants in our training sample is 930.

The most commonly used words (with counts) are:
[('Pizza', 1219), ('Pizzeria', 366), ('Restaurant', 216), ('Grill', 173), ('Italian', 111), ('Pasta', 109), ('Bar', 101), ('Sicilia', 96), ('Cafe', 85), ('Kitchen', 61), ('John', 55), ('Giovanni', 54), ('North', 54), ('Store', 51), ('House', 50), ('Deli', 47), ('Casey', 47), ('Subs', 46), ('Grille', 45), ('Mellow', 43)]

The number of unique words in our training sample is 1399.
After removing stop words, the number of unique words in our training sample is 1302.



Stop words have been removed.

For Restaurant is expensive (mid-point > $40):
Target Value Counts:
0.0 452 (not expensive)
1.0 85 (expensive)

Best Score: 0.8286778398510242
Best Params: {'estimator__min_samples_leaf': 4, 'estimator__min_samples_split': 10}

Top 15 most important words in name for predicting an expensive pizza restaurant:

Importance	Word
0.07630498713198094	pot
0.06462667823153828	melting
0.0287585653865833	house
0.025518956856531223	resort
0.02488870357903771	il
0.023633080189634073	pizzeria
30.021084436545314434	italian
0.02027845150538192	cafe
0.01992961773732572	grill
0.019681611399994445	ristorante
0.019411682499474555	cucina
0.01853942935076896	steak
0.018174151216137108	lounge
0.01802266077383435	spa
0.016703303598256076	brio

For Restaurant is cheap (mid-point < $15):
Target Value Counts:
1.0 423 (not cheap)
0.0 114 (cheap)

Best Score: 0.819366852886406
Best Params: {'estimator__min_samples_leaf': 2, 'estimator__min_samples_split': 5}
Score: 0.7962962962962963

Top 15 most important words in a name for predicting a cheap pizza restaurant:

Importance	Word
0.07593321390417072	pizza
0.041764656125596836	deli
0.04113194212482545	john
0.035668115610345975	chicken
0.03509176883116404	papa

0.034380077371007456	kitchen
0.031883498964207994	shop
0.02683791469219333	fried
0.025309961939515704	sports
0.02292627127640406	street
0.021068872974467107	ristorante
0.01879707511214269	bar
0.018319442414161106	bagel
0.017388615439130898	subs
0.016753406555392213	grill

Finding the best model from from Logistic Regression, Random Forest and SVC:

```
Best Score:  0.8603351955307262
Best Params:  {'classifier': RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
        max_depth=None, max_features=3, max_leaf_nodes=None,
        min_impurity_decrease=0.0, min_impurity_split=None,
        min_samples_leaf=1, min_samples_split=2,
        min_weight_fraction_leaf=0.0, n_estimators=300, n_jobs=None,
        oob_score=False, random_state=None, verbose=0,
        warm_start=False), 'classifier__max_features': 3, 'classifier__n_estimators': 300}
```