Create Part 3 of your Analysis Case Study project. Part 3 should consist of Model Evaluation and Selection. You can use any methods/tools you think are most appropriate.

Write the step-by-step instructions for completing the Model Evaluation and Selection part of your case study.

I would like to analyze the correlation of certain words with more expensive menu prices, specifically at places that serve pizza.

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 number of words in each restaurant name using a histogram
8. View the most common words names, as well as distribution of all words used (minus stopwords)
9. Bar plot the 20 most common words in order to visualize them for better understanding
10. Drop the rows that contain duplicate restaurant names
11. Find the midpoint price range for each restaurant and transform into target
12. Use TFIDF-Vectorizer on restaurant names to create feature variables
13. 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
14. 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)
15. Repeat steps 13 and 14 to find words that contribute to classifying a restaurant as cheap (mid-point < $15)

**(See next page for output.)**

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

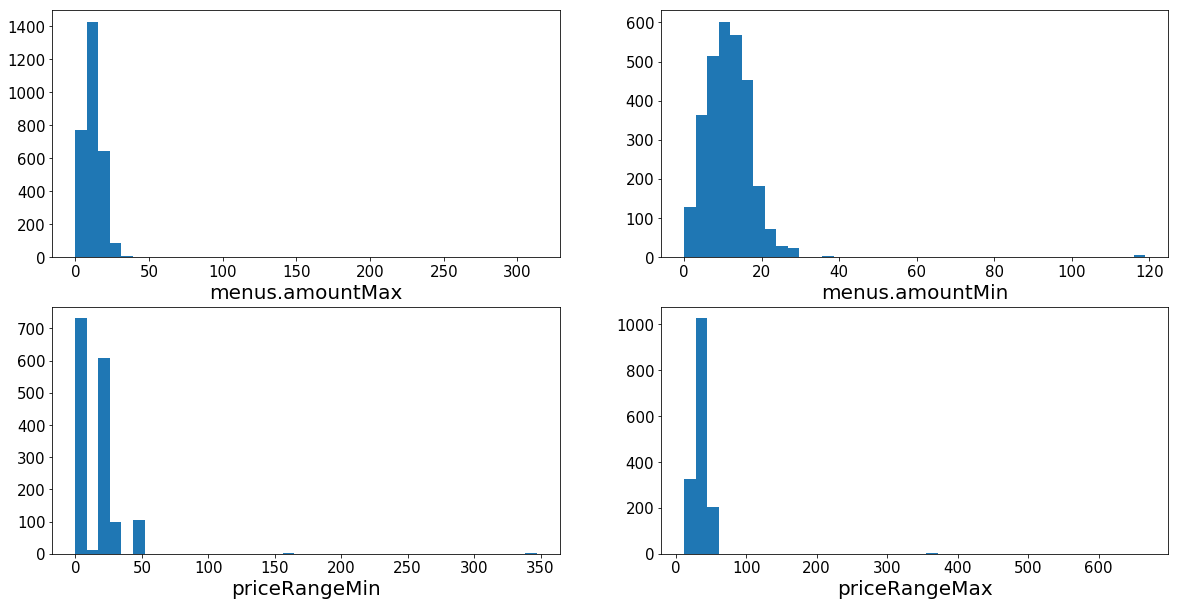
count 3510 3510 ... 1557 3510

unique 989 984 ... 1 281

top AVwdIsuzkufWRAb52p9M 1605 Kanawha Blvd W ... USD CA

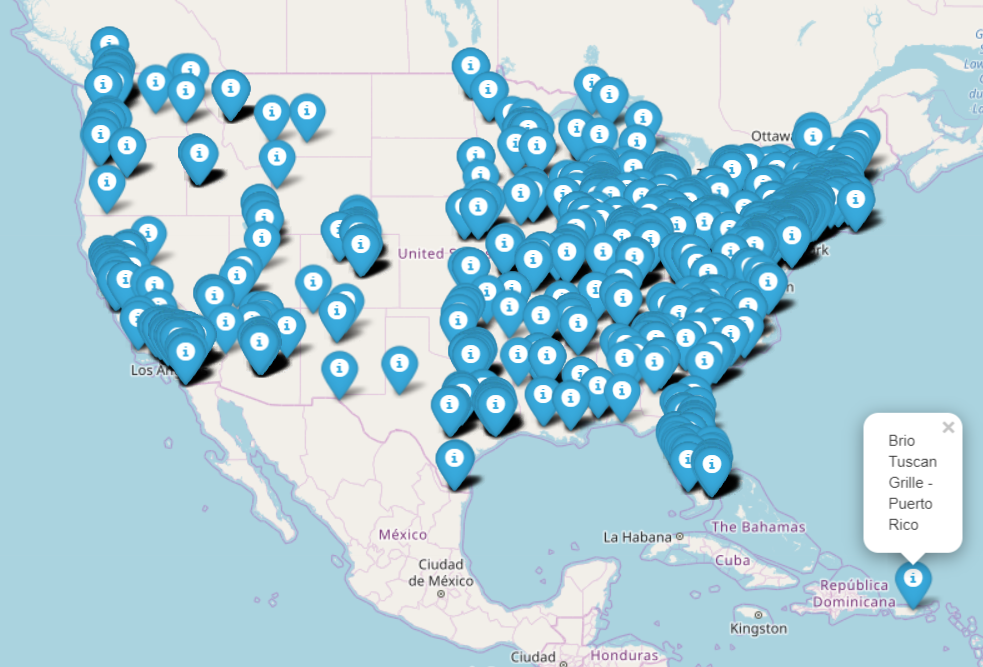
freq 64 64 ... 1557 256

[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

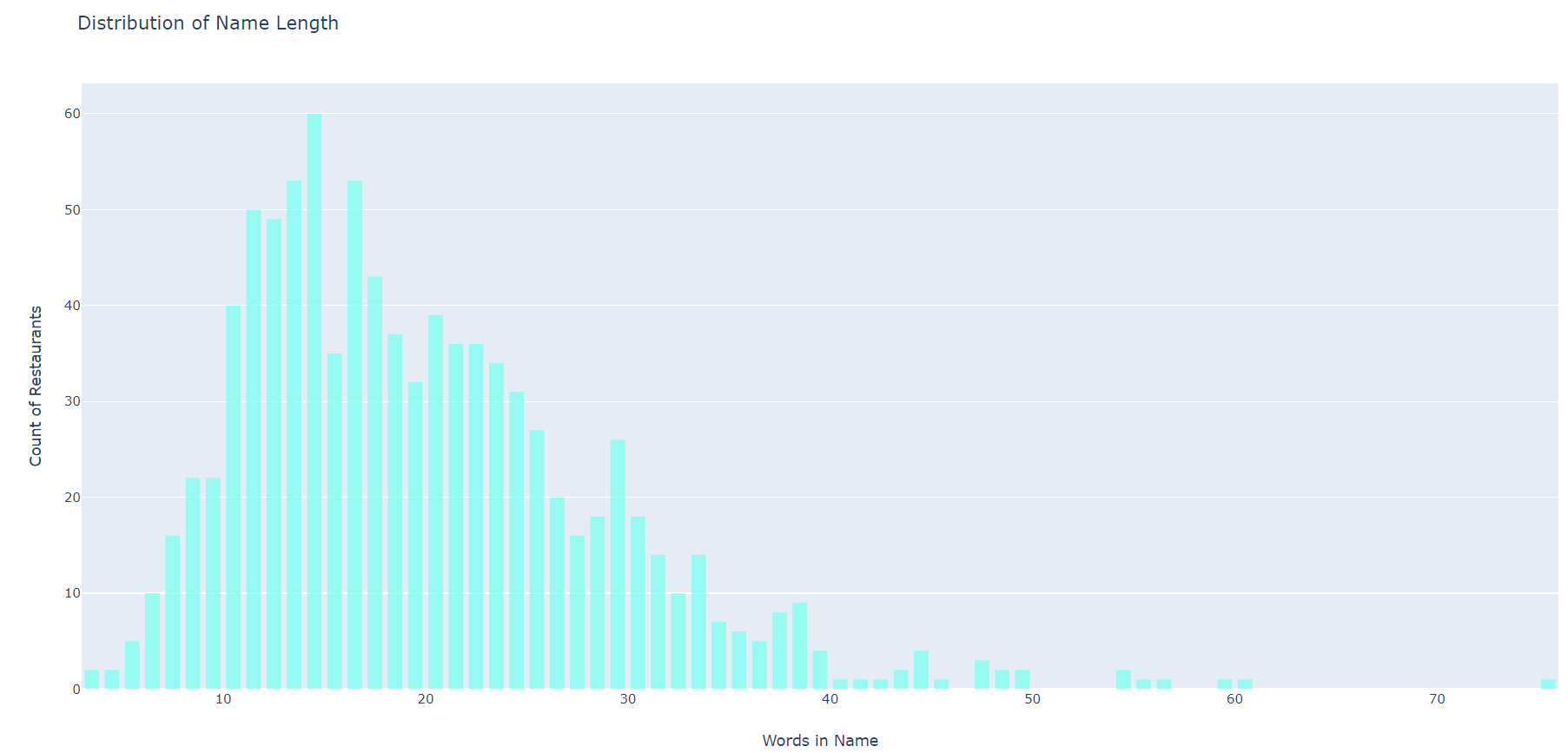
priceRangeMax: 3 Outliers

Mean 36.566

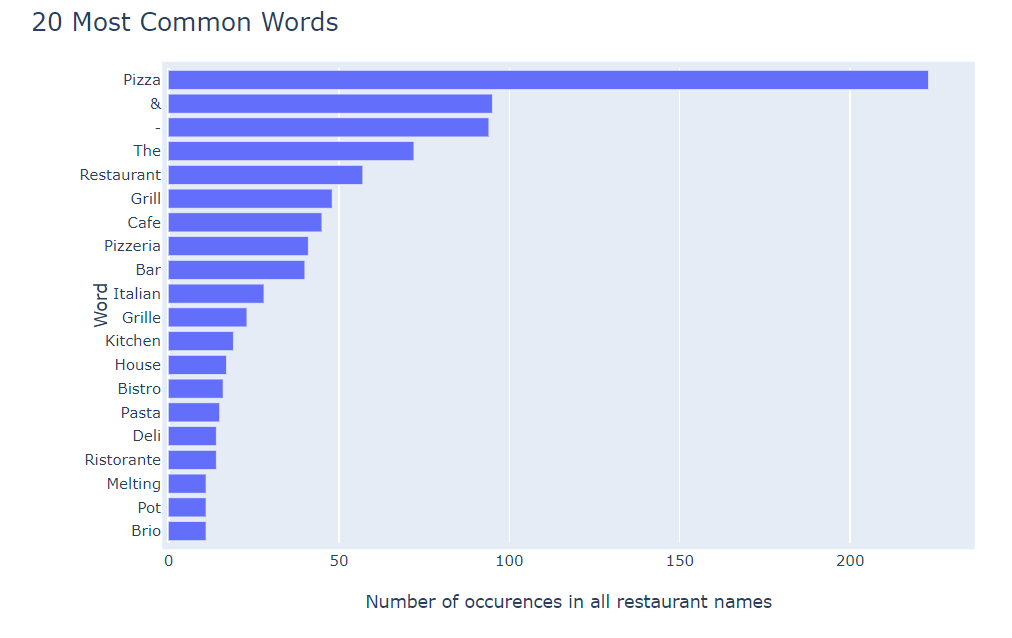
363.0

363.0

666.0



The number of unique words in our training sample is 1399.



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}

importance feat

0.1026411598864373 pizza

0.067721892210266 ristorante

0.058189625761467675 melting

0.04566060842563853 pot

0.035793965230999494 trattoria

0.026892039344862347 pizzeria

0.02426068629363222 il

0.02243633654651342 house

0.022126212500813307 cucina

0.02170337103157211 resort

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 10 most important words in name for predicting a cheap restaurant:

importance feat

0.04849982504948452 deli

0.03938377983723025 kitchen

0.033269117323561775 papa

0.029278150258440307 john

0.027149766826985496 chicken

0.02275737687481527 bar

0.022283032382184906 fried

0.02159708470951138 shop

0.020023471387833683 sports

0.018034960566606258 subs