Decoding Restaurant Opinions

UNDERSTANDING REVIEWS AND RATING

Kelly Choy
COMM 599
November 20, 2023

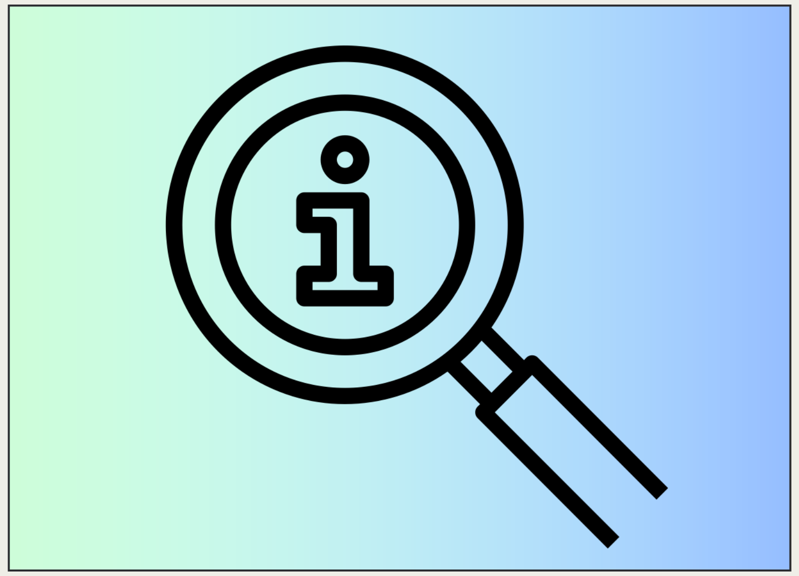
AGENDA

- The Problem
- Solution
- Project Methods
- Project Progress
- Next Steps

THE PROBLEM

Subjectivity in Ratings

People express their opinions in diverse ways, and sentiment analysis alone may not capture the subtle aspects within their reviews



Lack of Insights from Restaurant Reviews

Existing rating systems may not provide a comprehensive understanding of the factors influencing overall restaurant ratings, and manual analysis of reviews is time-consuming

*Sentiment analysis is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral

SOLUTION - PROJECT IDEA

- Utilizing **Yelp Dataset**
- LDA Topic Modeling
- Create a model to predict the ratings of the restaurant
- Why do I want to create this:
  - Want to utilize machine learning tools in a project
  - I’m a foodie, and I want to extract main topics from reviews and create a new approach to predicting rating

PROJECT METHODS

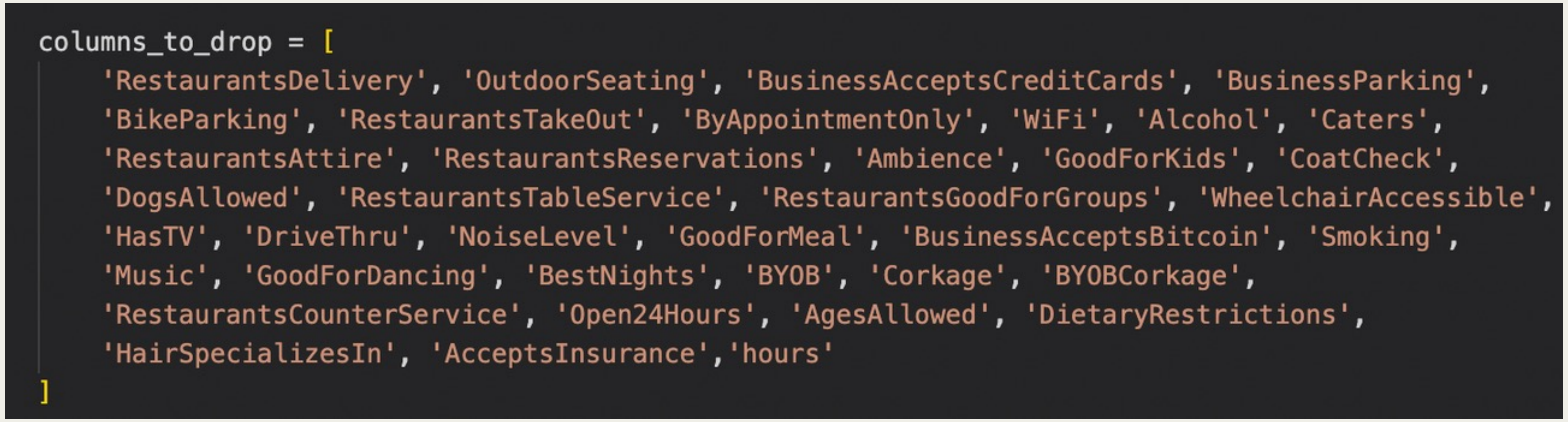
- LDA Topic Modeling:
  - Latent Dirichlet Allocation (LDA)
    - Generative statistical model used for topic modeling
    - Identifies patterns of words and infers the underlying topics that characterize the corpus
- Purpose of Topic Modeling:
  - Identify several topics or themes presented in a collection of text
  - Content summarization

- Predictive Modeling
  - Regression Analysis
    - Goal is to establish a relationship between one dependent variable and one or more independent variables

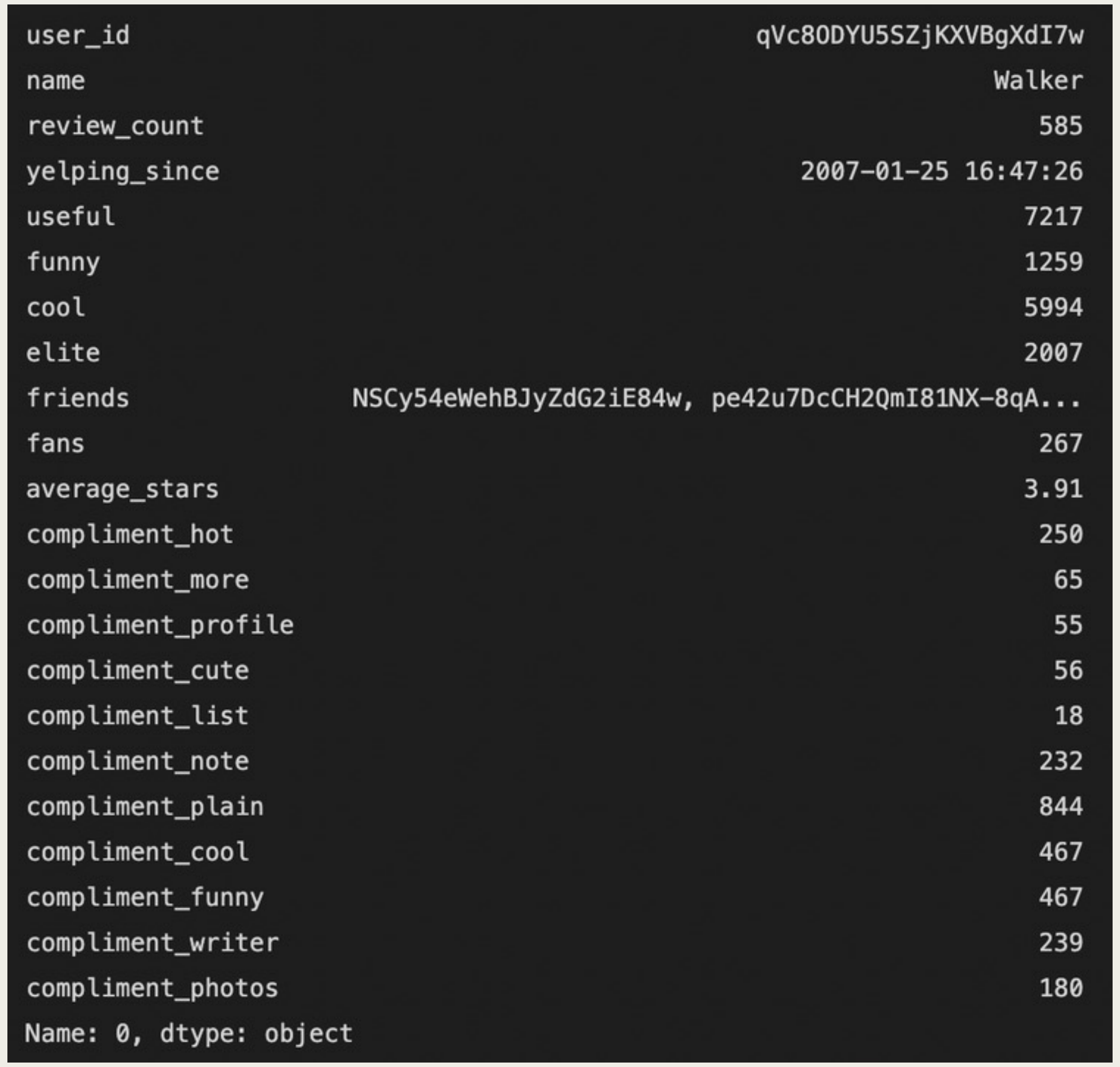
STEP 1: GATHERING THE DATA

Combine Files
Business, Users, Reviews

- Overall folder ~8.8GB
- Filter business file:
  - State: California
  - Remove extra columns
- Merge business file with Reviews file on "Business _ID"
- Then merge user file on "User_ID"
- Dataset:
  - 211,749 rows x 31 columns



```
columns_to_drop = [
    'RestaurantsDelivery', 'OutdoorSeating', 'BusinessAcceptsCreditCards', 'BusinessParking',
    'BikeParking', 'RestaurantsTakeOut', 'ByAppointmentOnly', 'WiFi', 'Alcohol', 'Caters',
    'RestaurantsAttire', 'RestaurantsReservations', 'Ambience', 'GoodForKids', 'CoatCheck',
    'DogsAllowed', 'RestaurantsTableService', 'RestaurantsGoodForGroups', 'WheelchairAccessible',
    'HasTV', 'DriveThru', 'NoiseLevel', 'GoodForMeal', 'BusinessAcceptsBitcoin', 'Smoking',
    'Music', 'GoodForDancing', 'BestNights', 'BYOB', 'Corkage', 'BYOBCorkage',
    'RestaurantsCounterService', 'Open24Hours', 'AgesAllowed', 'DietaryRestrictions',
    'HairSpecializesIn', 'AcceptsInsurance','hours'
]
```



```
user_id                                        qVc8ODYU5SZjKXVBgXdI7w
name                                                           Walker
review_count                                                      585
yelping_since                                     2007-01-25 16:47:26
useful                                                           7217
funny                                                            1259
cool                                                             5994
elite                                                            2007
friends             NSCy54eWehBJyZdG2iE84w, pe42u7DcCH2QmI81NX-8qA...
fans                                                              267
average_stars                                                    3.91
compliment_hot                                                    250
compliment_more                                                    65
compliment_profile                                                 55
compliment_cute                                                    56
compliment_list                                                    18
compliment_note                                                   232
compliment_plain                                                  844
compliment_cool                                                   467
compliment_funny                                                  467
compliment_writer                                                 239
compliment_photos                                                 180
Name: 0, dtype: object
```

STEP 2: DATA CLEANING

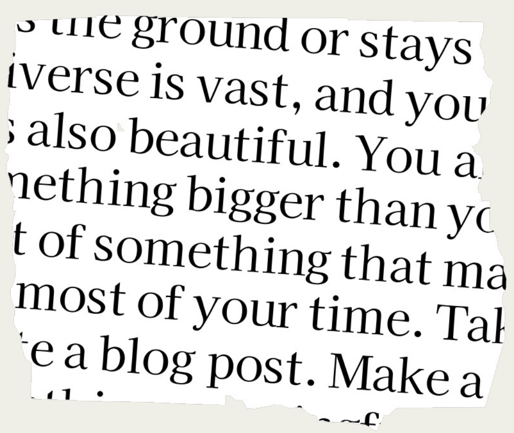
Prepare Data

- Utilize NLTK (Natural Language Toolkit)
  - ***Text Preprocessing:*** manipulating text, making it easier to analyze and extract information from natural language data
  - ***Tokenizing:*** breaking text into words or sentences
  - Remove stopwords
    - a, the, he, she, it, in, etc
  - Remove punctuation
  - All text in English

STEP 2: DATA CLEANING

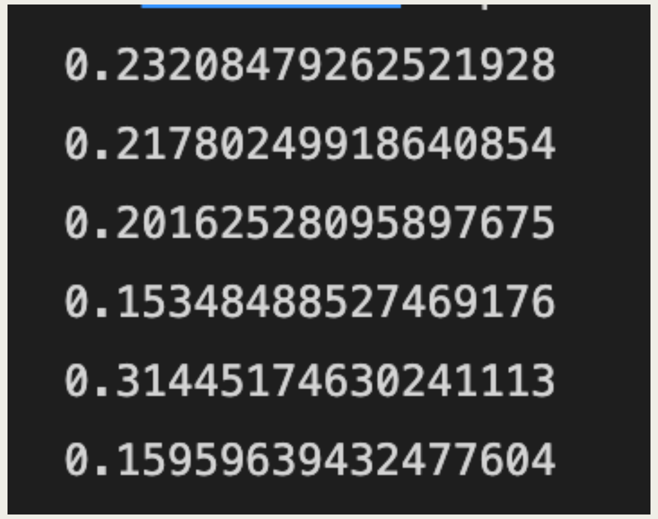
Text Vectorization

- Utilize scikit-learn's TfidfVectorizer
  - TF-IDF = Term Frequency-Inverse Document Frequency
  - Transform your text data into TF-IDF vectors
    - Numerical statistic that reflects the importance of a word in a document relative to a collection of documents
- TF-IDF vectors are used for LDA Topic Modeling
  - Machine learning algorithms typically require numerical input



s the ground or stays
iverse is vast, and you
s also beautiful. You a
nething bigger than yo
t of something that ma
most of your time. Tak
e a blog post. Make a

→



```
0.23208479262521928
0.21780249918640854
0.20162528095897675
0.15348488527469176
0.31445174630241113
0.15959639432477604
```

STEP 3: LDA TOPIC MODELING

Topic 1: room, hotel, stay, resort, pool, rooms, property, bacara, spa, stayed

Topic 2: burger, good, fries, place, burrito, food, like, sandwich, cheese, chicken

Topic 3: thai, food, great, amazing, service, delicious, ramen, best, place, wine

Topic 4: beer, food, good, place, beers, great, masks, bar, games, like

Topic 5: food, service, order, minutes, time, asked, table, said, didnt, place

Topic 6: sushi, food, good, roll, place, chicken, rolls, rice, indian, spicy

Topic 7: tacos, pizza, food, mexican, good, salsa, great, place, best, taco

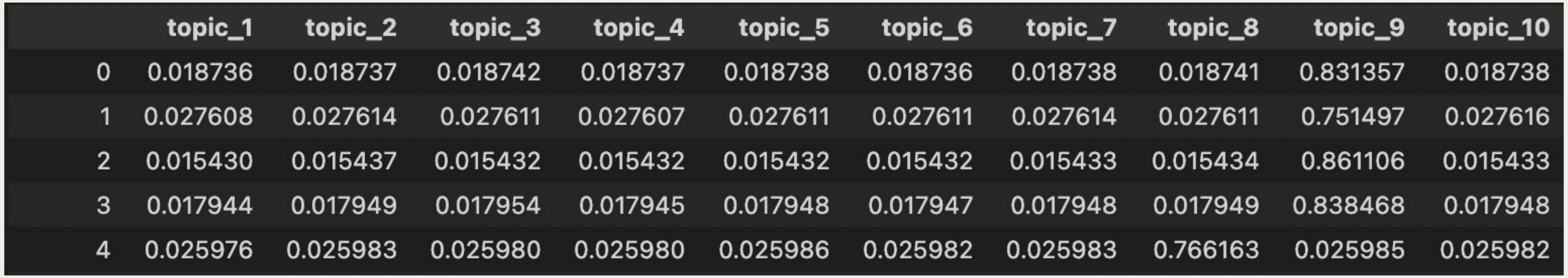
Topic 8: great, food, service, place, good, friendly, amazing, staff, nice, santa

Topic 9: coffee, great, place, breakfast, good, sandwich, delicious, love, sandwiches, best

Topic 10: good, great, food, ordered, salad, place, service, fish, delicious, sauce

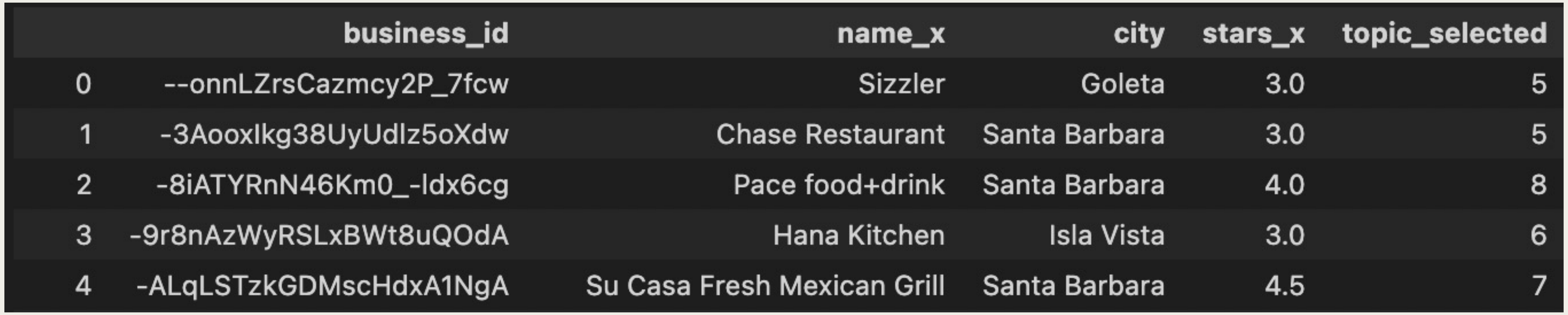
STEP 3: LDA TOPIC MODELING

TF-IDF Values for each individual review



	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8	topic_9	topic_10
0	0.018736	0.018737	0.018742	0.018737	0.018738	0.018736	0.018738	0.018741	0.831357	0.018738
1	0.027608	0.027614	0.027611	0.027607	0.027611	0.027611	0.027614	0.027611	0.751497	0.027616
2	0.015430	0.015437	0.015432	0.015432	0.015432	0.015432	0.015433	0.015434	0.861106	0.015433
3	0.017944	0.017949	0.017954	0.017945	0.017948	0.017947	0.017948	0.017949	0.838468	0.017948
4	0.025976	0.025983	0.025980	0.025980	0.025986	0.025982	0.025983	0.766163	0.025985	0.025982

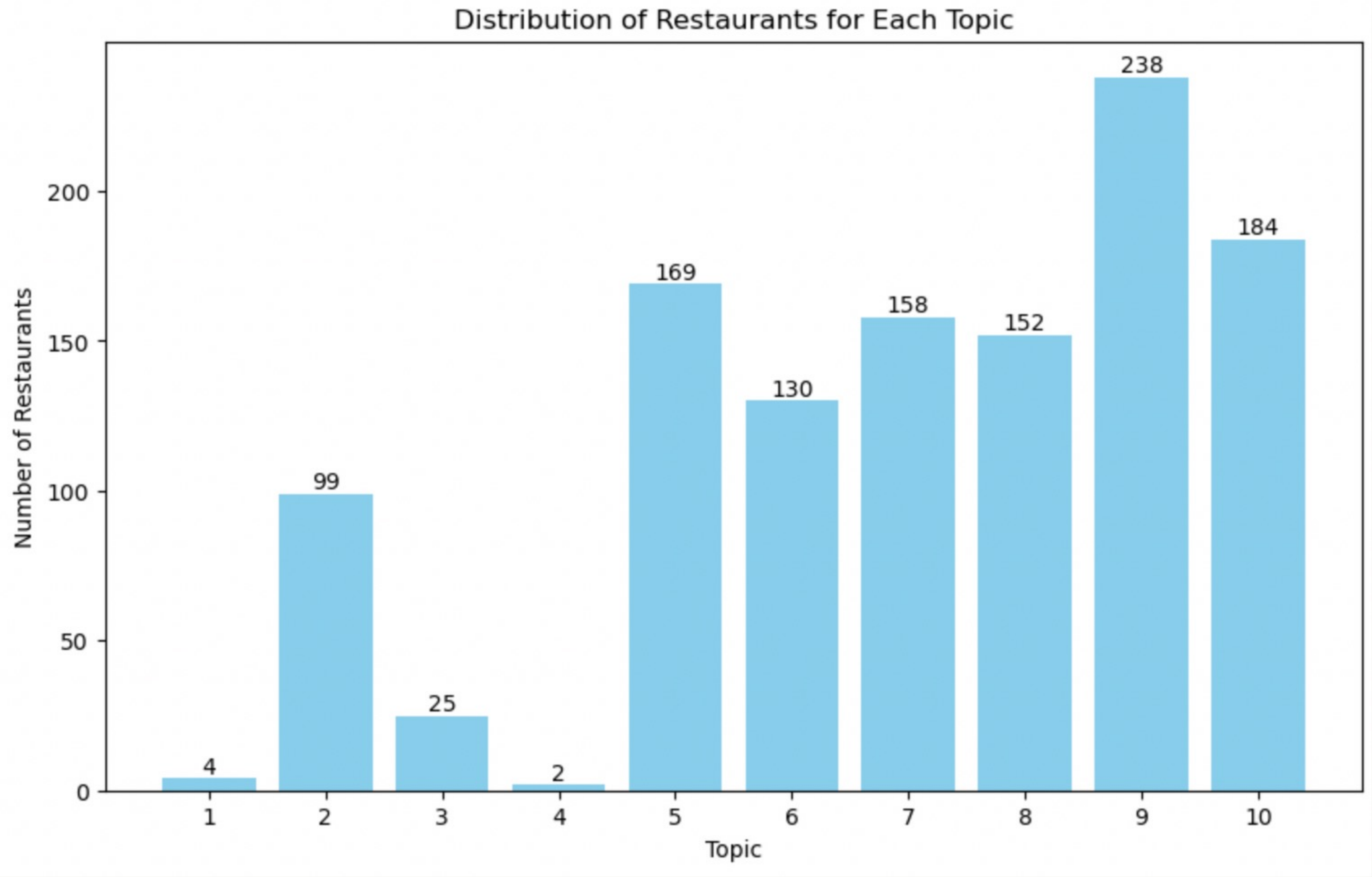
Select the common topic out of all the reviews to be represented



	business_id	name_x	city	stars_x	topic_selected
0	--onnLZrsCazmcy2P_7fcw	Sizzler	Goleta	3.0	5
1	-3AooxIkg38UyUdlz5oXdw	Chase Restaurant	Santa Barbara	3.0	5
2	-8iATYRnN46Km0_-Idx6cg	Pace food+drink	Santa Barbara	4.0	8
3	-9r8nAzWyRSLxBWt8uQOdA	Hana Kitchen	Isla Vista	3.0	6
4	-ALqLSTzkGDMscHdxA1NgA	Su Casa Fresh Mexican Grill	Santa Barbara	4.5	7

Topic 7: tacos, pizza, food, mexican, good, salsa, great, place, best, taco

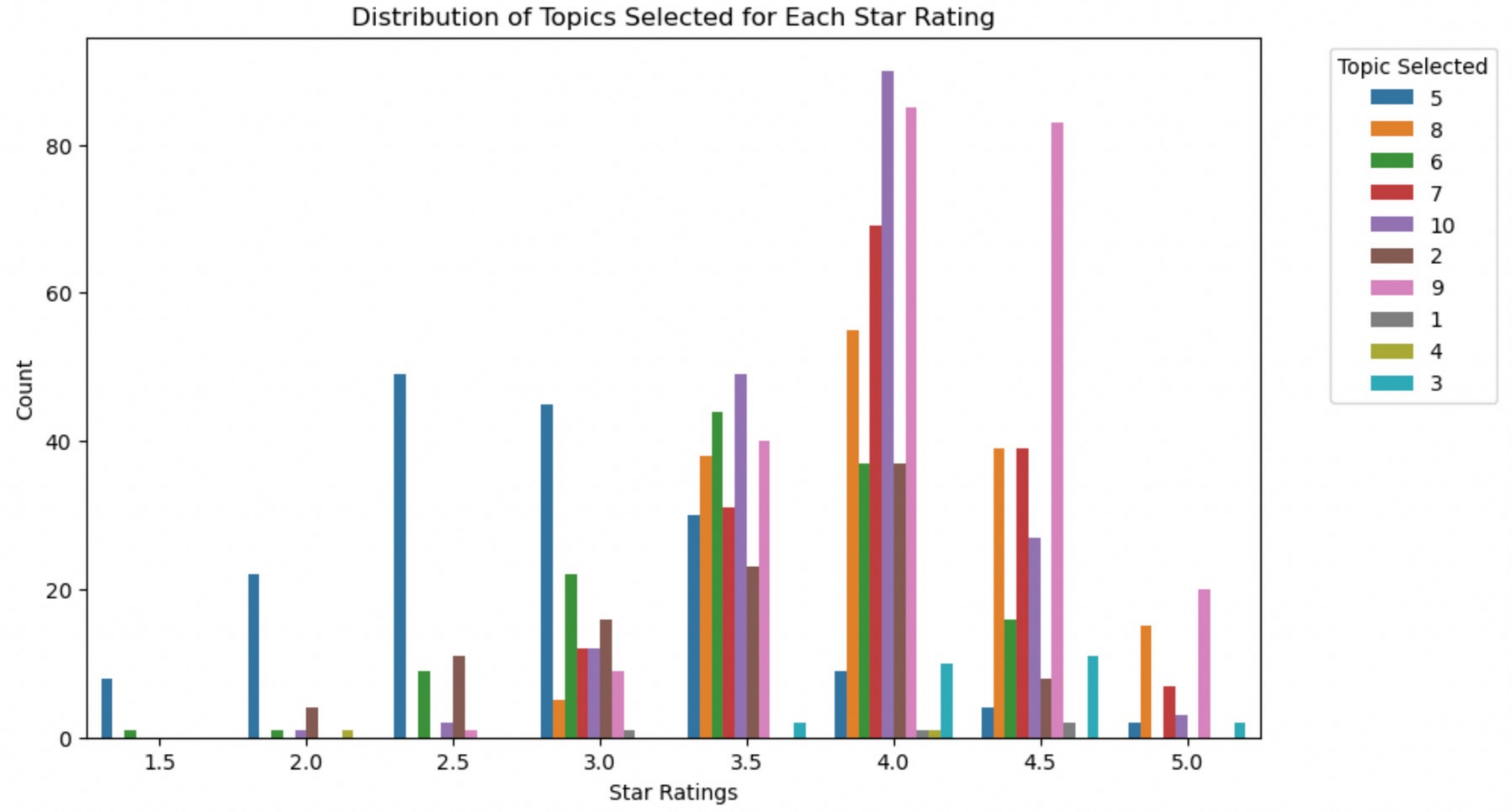
VISUALIZATIONS



1161 restaurants

10 topics

VISUALIZATIONS



STEP 4: PREDICTIVE MODELING

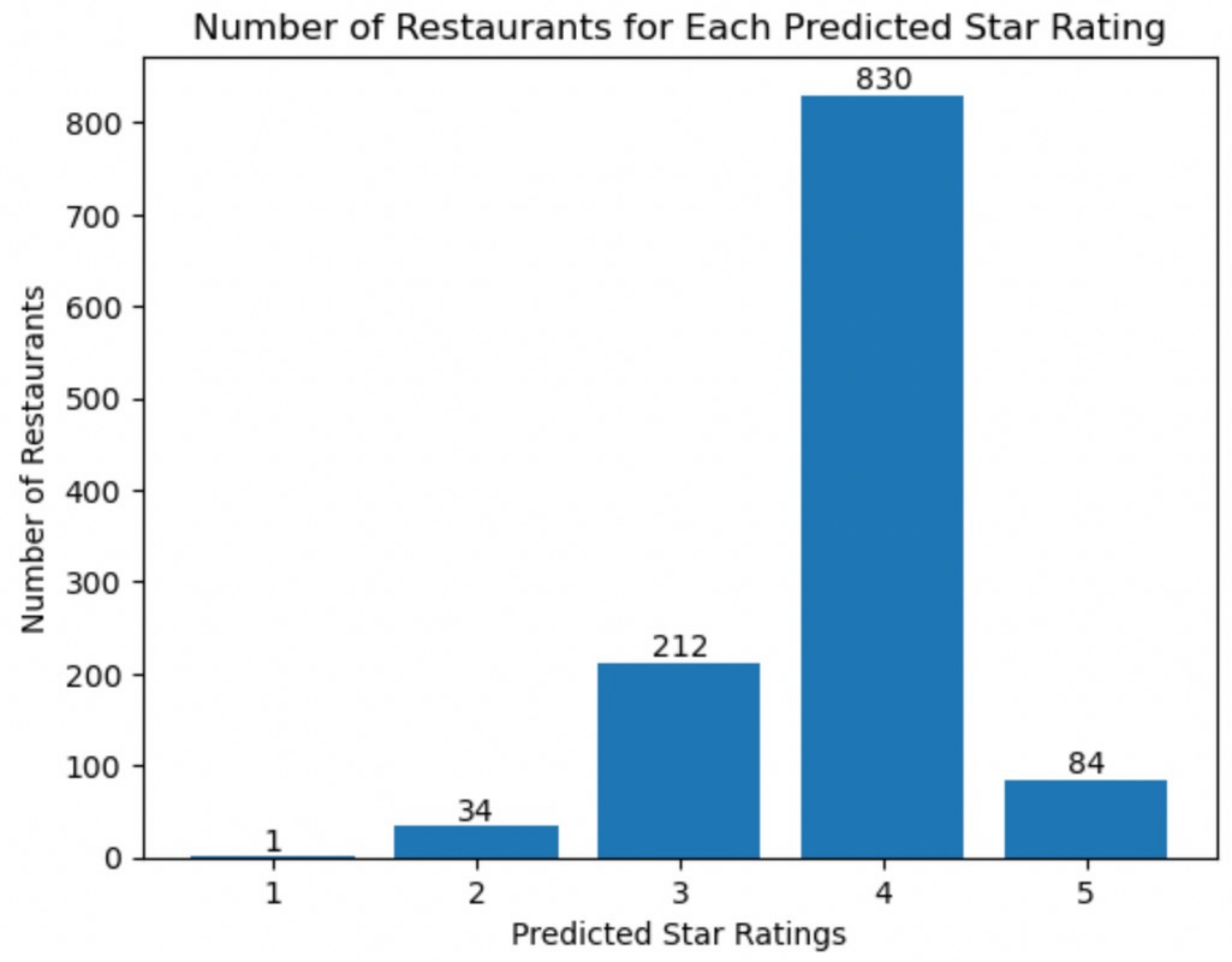
- Linear Regression vs. XGBoost Regressor
  - **XGBoost:**
    - Can model complex interactions, adapt to non-linear patterns, and capture high-order relationships, making it more flexible in handling diverse data patterns
    - More powerful for *large dataset*
  - Metric Used:
    - **Mean Absolute Error (MAE):** a lower value = predicted values are closer to the actual values, indicating better accuracy in the model's predictions.
    - Mean Squared Error, R-Squared (R^2), Root Mean Squared Error (RMSE)

STEP 4: PREDICTIVE MODELING

- Use the model created to predict the restaurant's overall rating

Count of Restaurants for Each Predicted Star Rating:

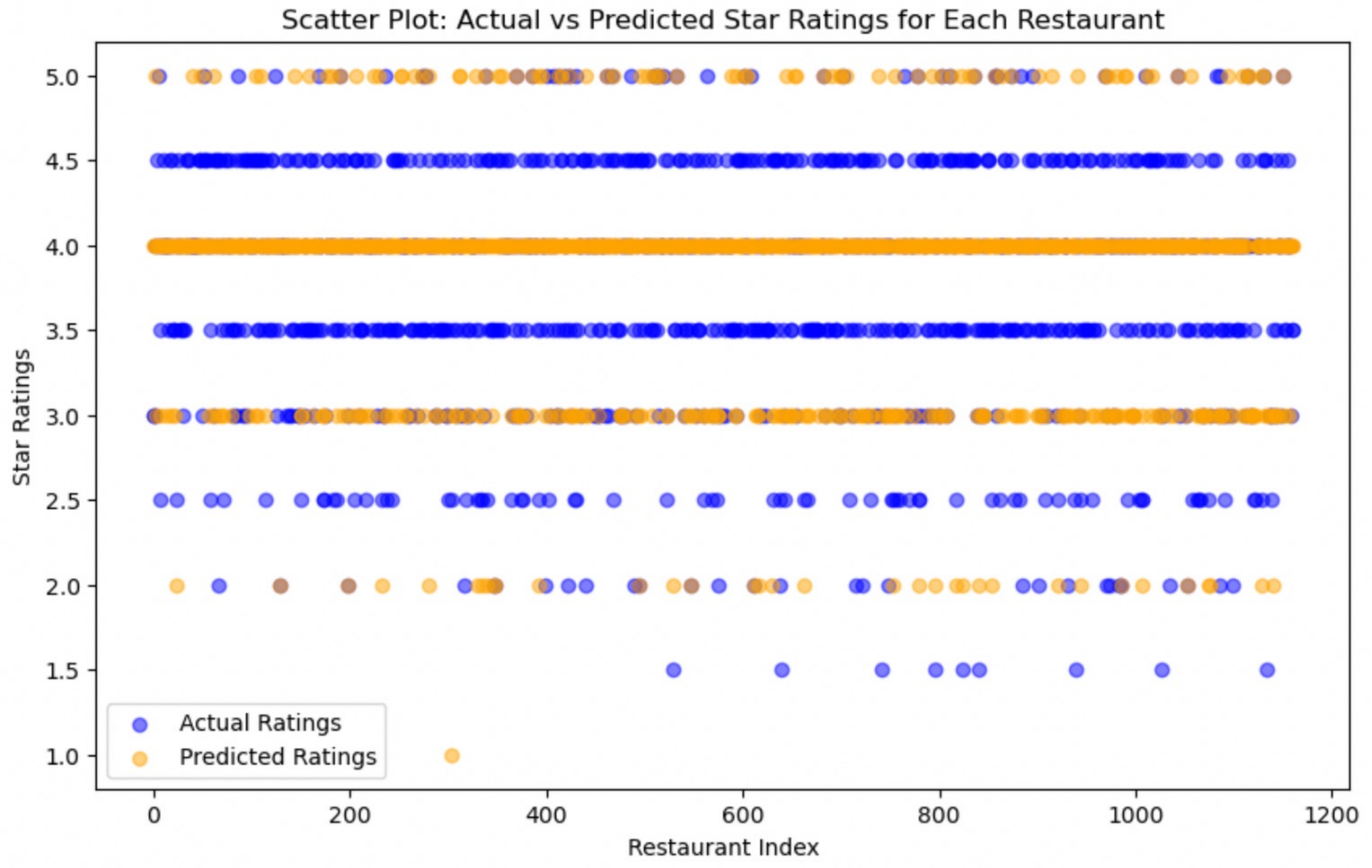
Predicted Star Rating: 1.0, Count: 1
Predicted Star Rating: 2.0, Count: 34
Predicted Star Rating: 3.0, Count: 212
Predicted Star Rating: 4.0, Count: 830
Predicted Star Rating: 5.0, Count: 84



VISUALIZATIONS

Metric:
Mean Absolute Error

On average, the model's predictions are off by approximately **0.80 units** from the true values

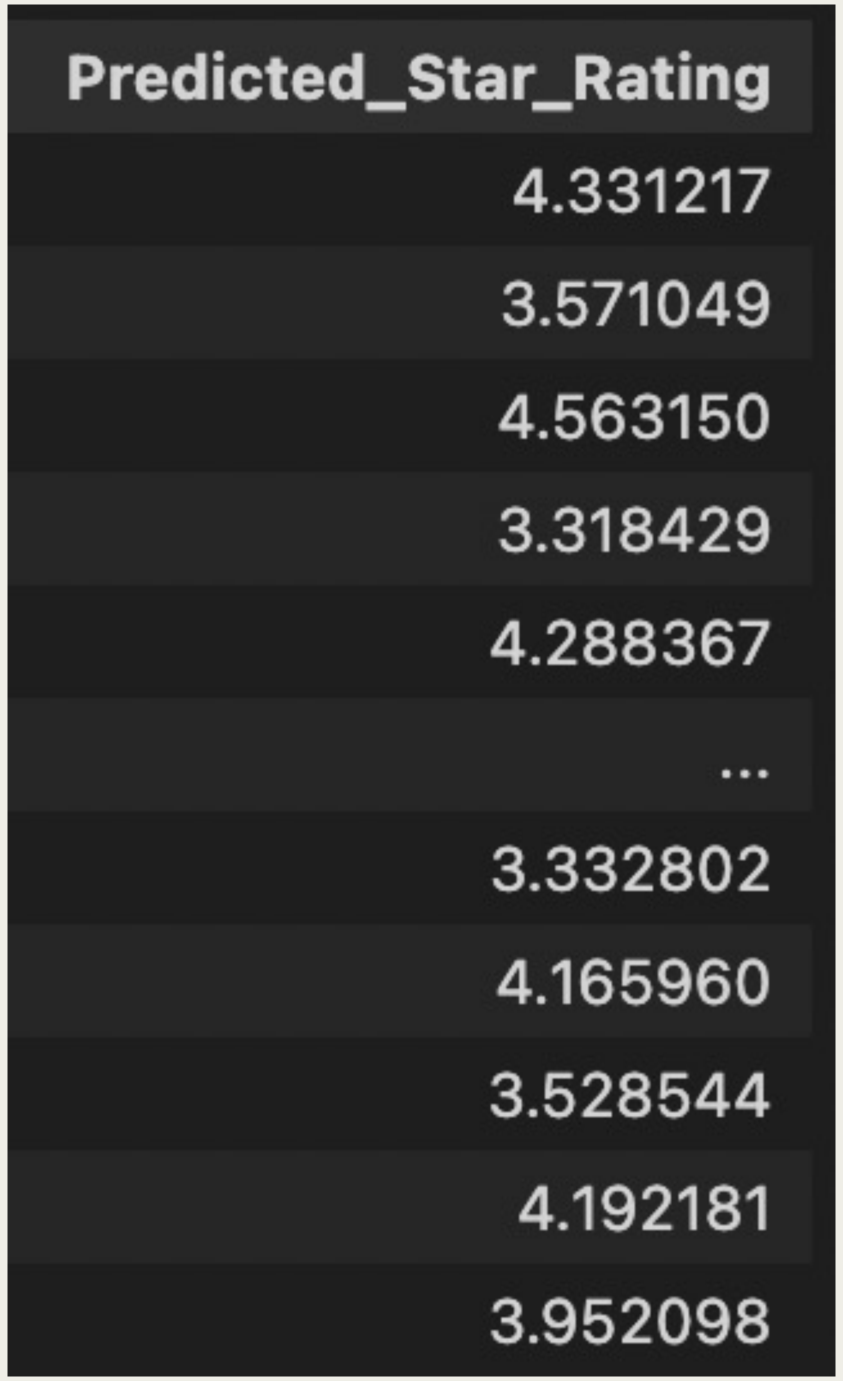


Mean Squared Error: 1.076
Mean Absolute Error: 0.804
R-squared: 0.403
Root Mean Squared Error: 1.037

NEXT STEPS

- Testing out the rounding values from the predictive model
  - Currently rounding to the nearest whole number
- Try out different techniques to improve the predictive modeling outcome
  - Binary classification
    - 1-3 stars = Dislike
    - 4-5 stars = Like
  - Multi-class classification

Without Rounding



Predicted_Star_Rating
4.331217
3.571049
4.563150
3.318429
4.288367
...
3.332802
4.165960
3.528544
4.192181
3.952098

Thank you!