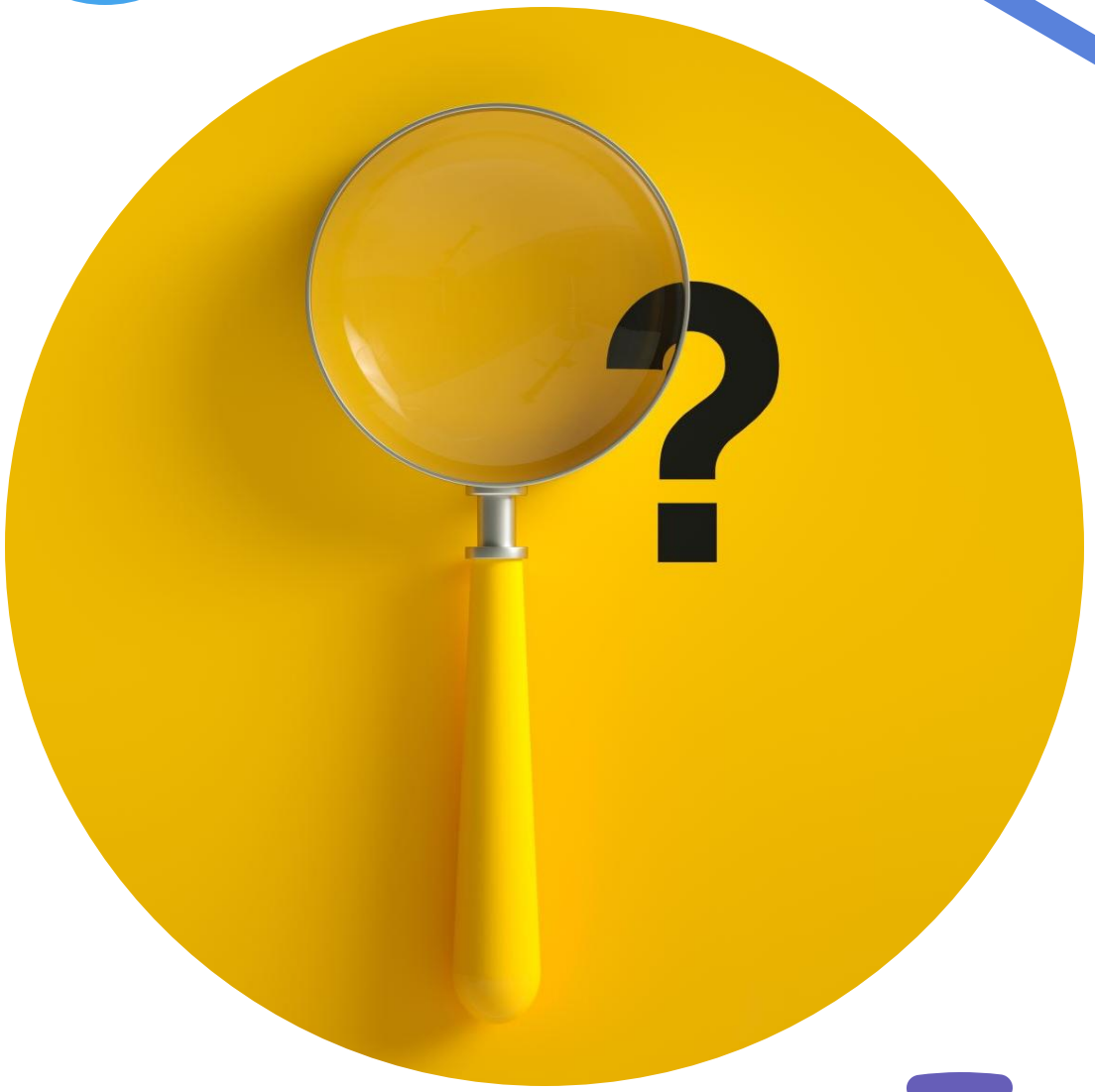


Factors of a Yelp Review and Their Influence

Group Members:

- Dean Papadopoulos
- Kripanjali Dhungana
- Norman Morris
- Zheding Zhao

Member Introduction

Dean Papadopoulos

- BS – Cyber Security
- MBA – Business Analytics

Kripanjali Dhungana

- MBA- Business Analytics
- Formerly in Logistics Operations

Norman Morris

- BBA - Computer Information Systems
- Formerly Healthcare IT

Zheding Zhao

- BS: Quantitative Finance
- BA: Mathematics and Economics

Overview

ABOUT THE DATASETS

INITIAL QUESTIONS

RESEARCH PROCESS

VISUALIZATIONS

MACHINE LEARNING MODEL

CONCLUSION

Datasets

- Yelp datasets containing reviews, business and user information.
- Income and Education datasets.
- Unemployment rate dataset.

Initial Questions

1. What are some keywords for good reviews and bad reviews?
2. Is there a correlation between review score and length of review text?
3. What is the relationship between median income and ratings on expensive restaurants?
4. What is the relationship between median income and number of restaurants?
5. Are people who leave more reviews likely to be more critical than people who leave less reviews?
6. Does lower unemployment result in more favorable reviews on Yelp?
7. Can classification machine learning models be used to predict a business' rating based on its available attributes?

EDA

150346 Businesses

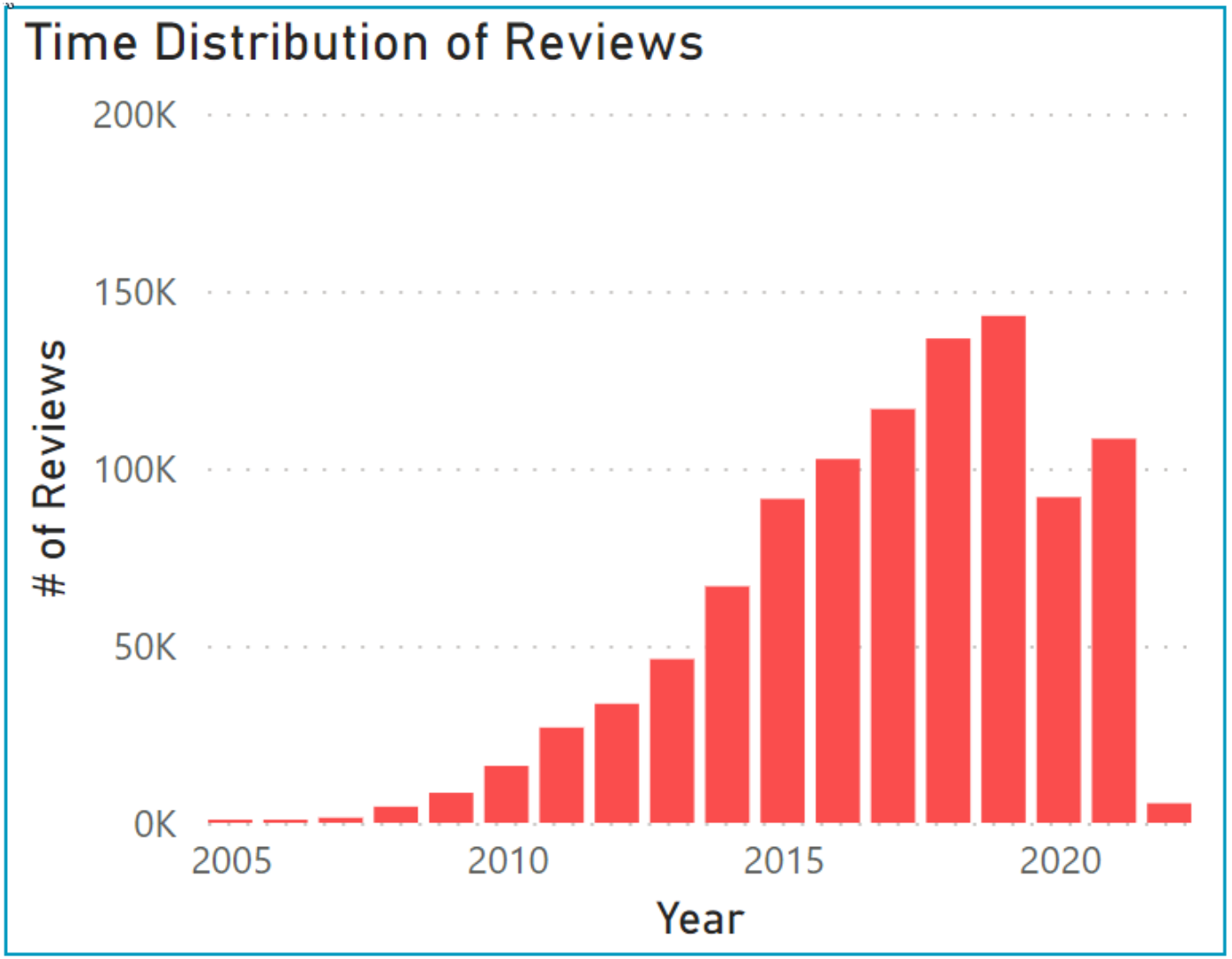
- 21 States, Territories & Provinces

6990280 Reviews

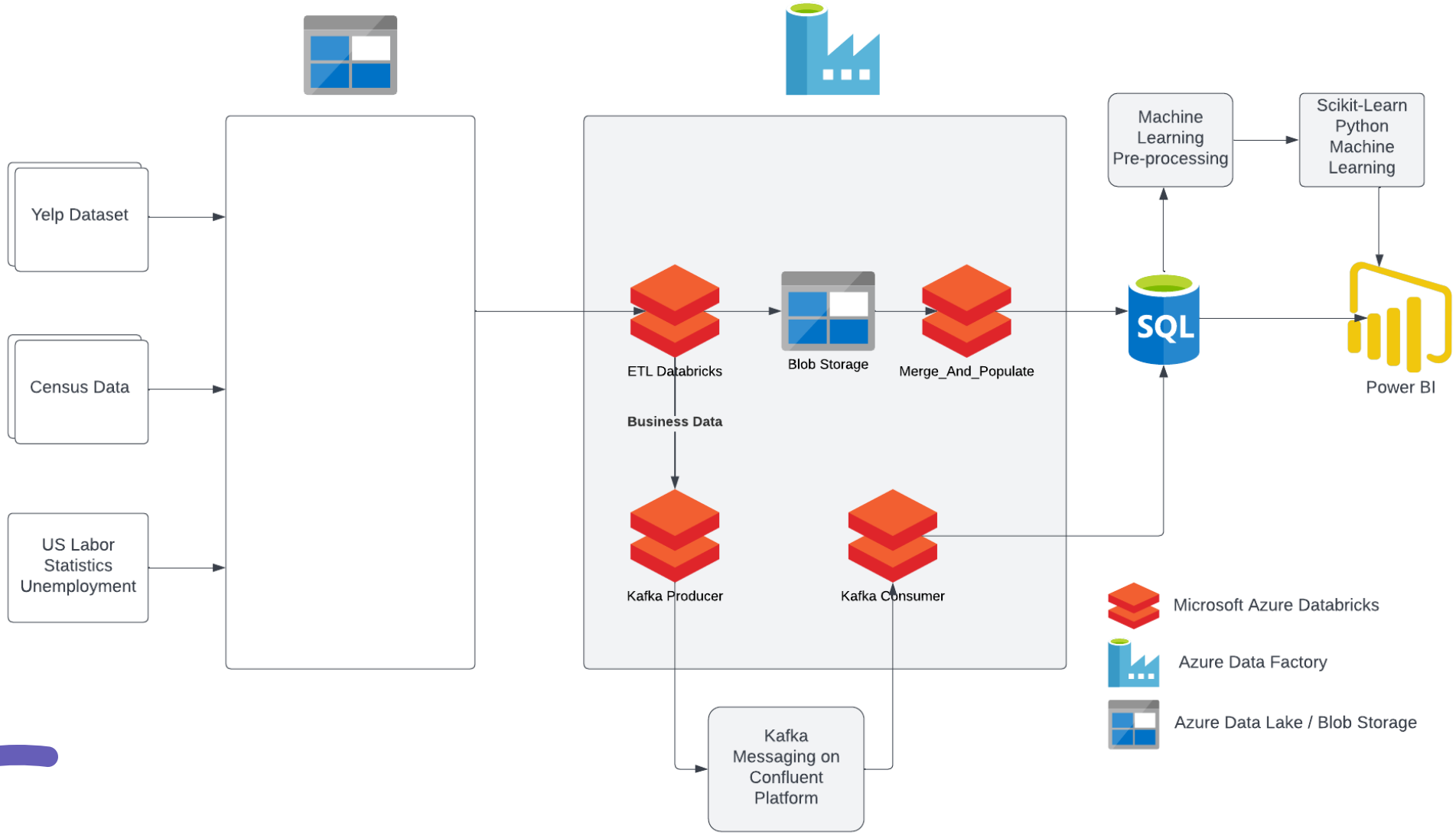
- 2005 to 2022
- Avg Rating: 3.75

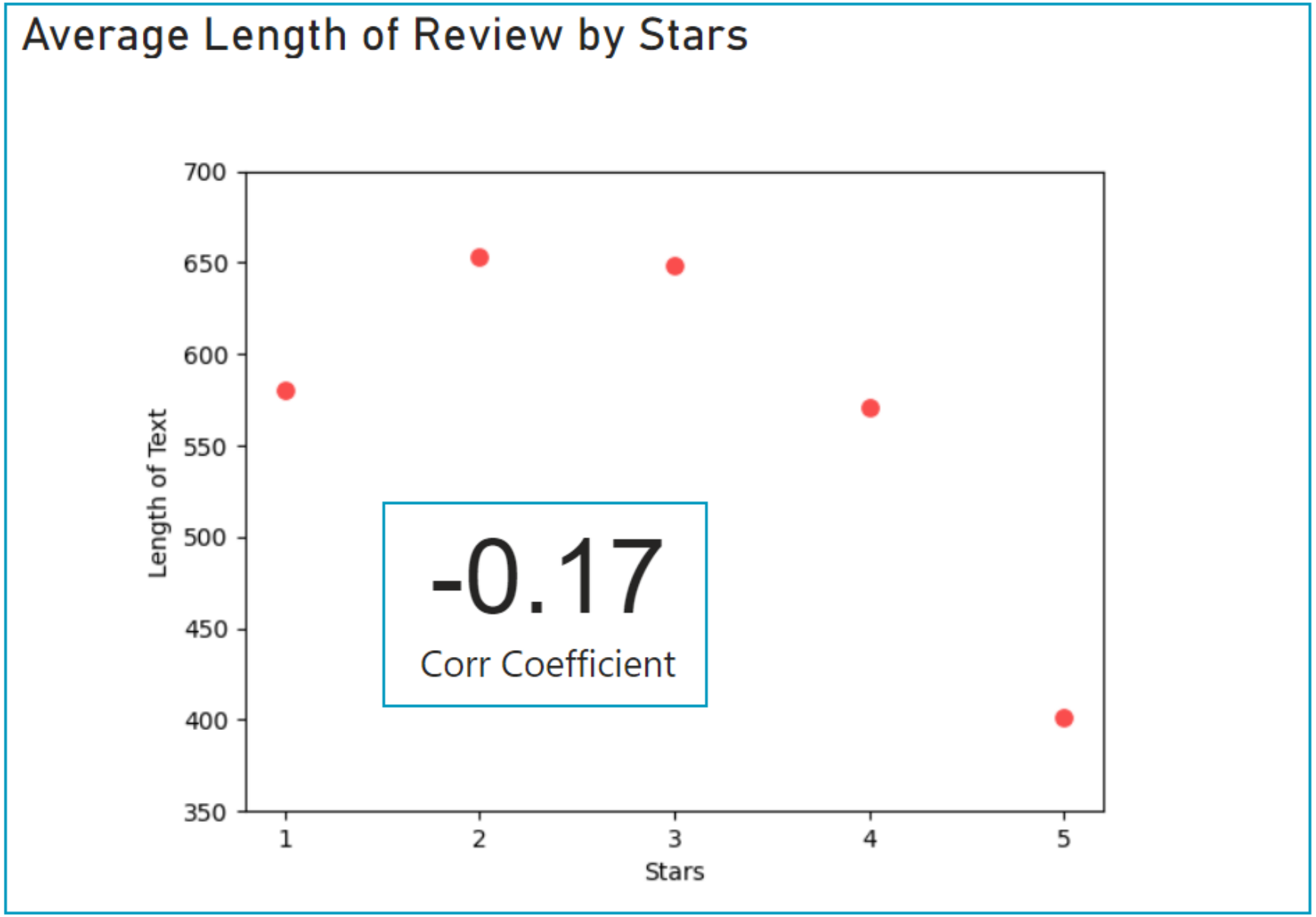
1987897 Users

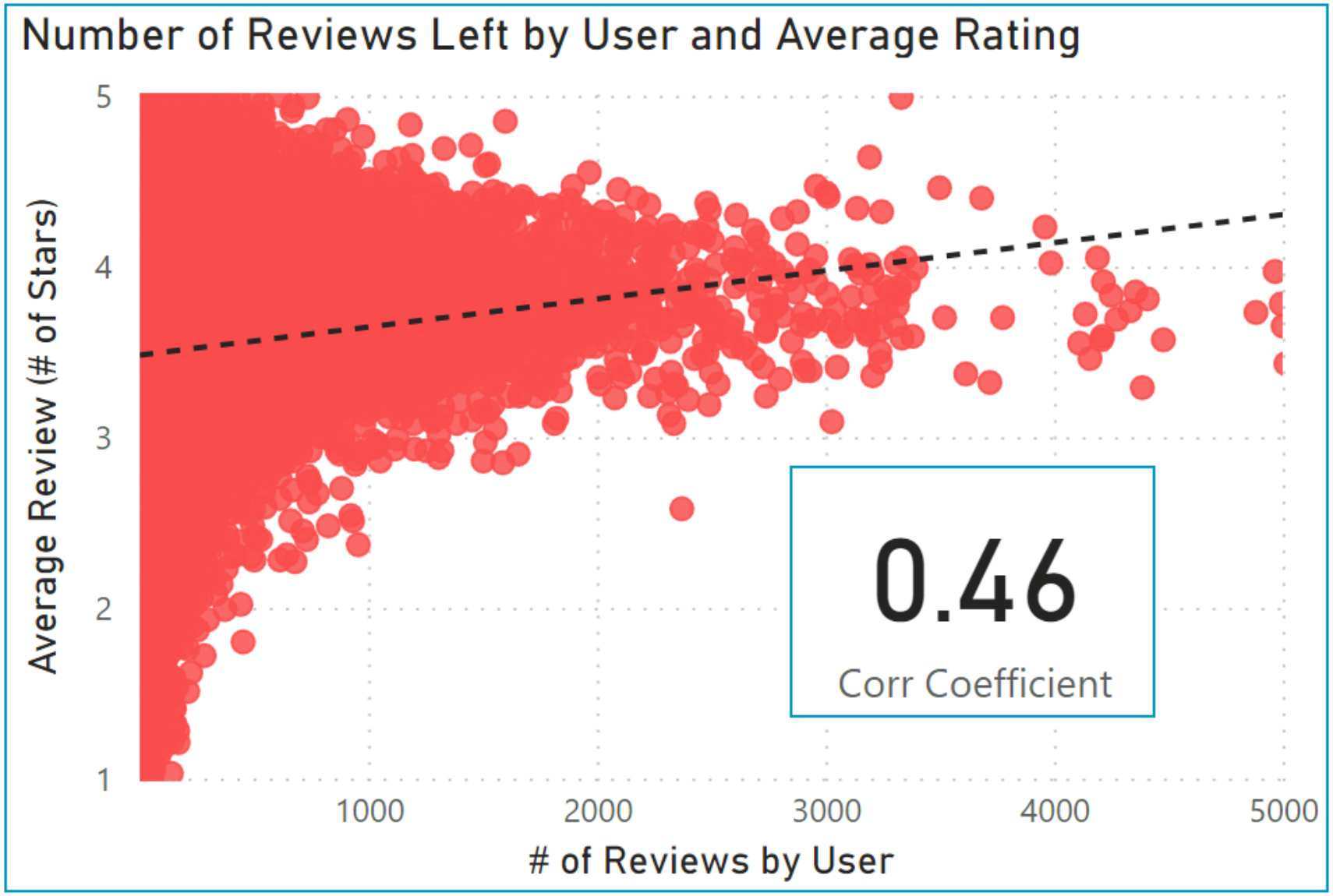
- Avg Number of Reviews: 23

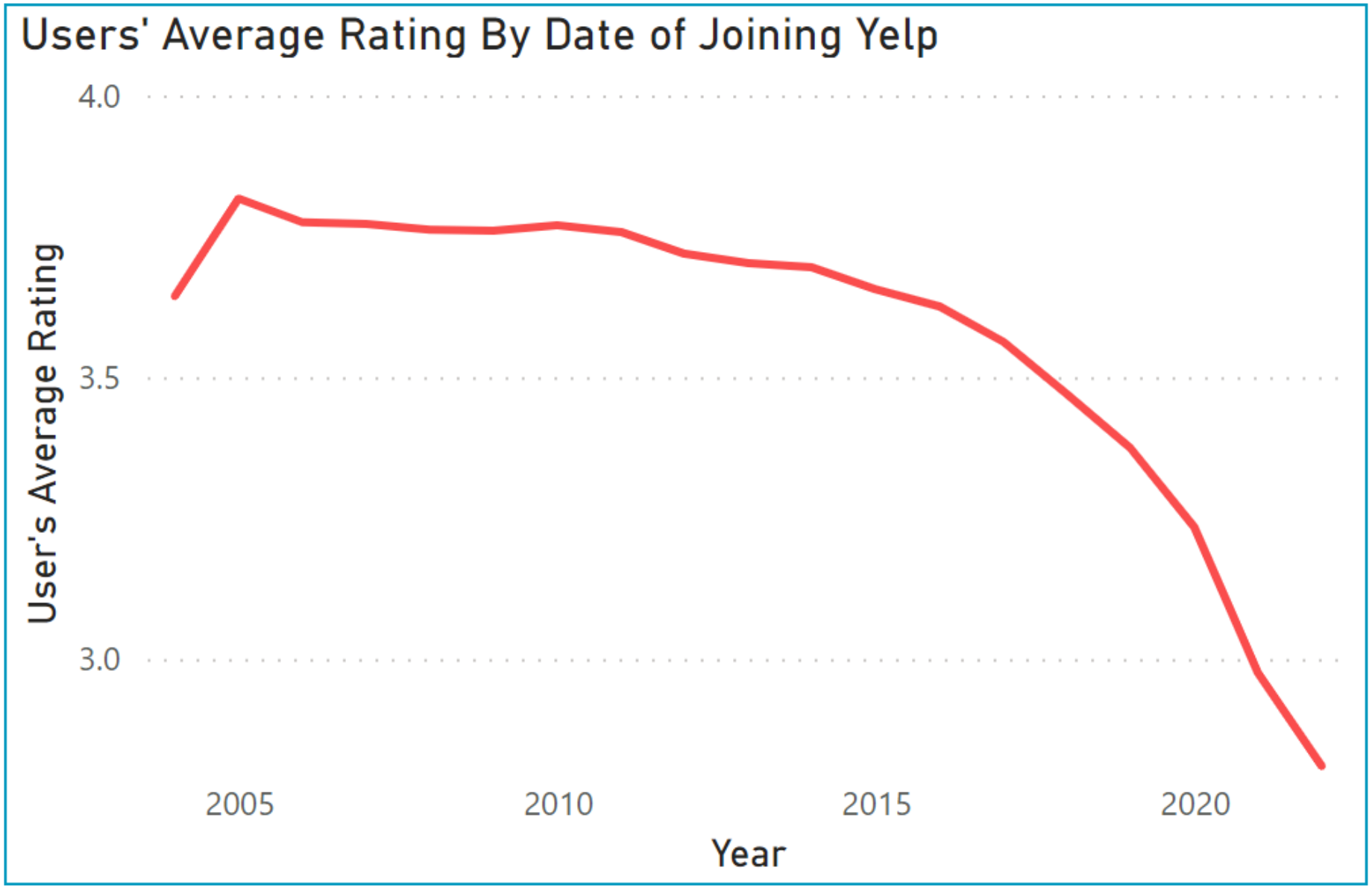


Cloud Platform & ETL




Average Length of Review by Stars
700
650
600
550
500
450
400
350
Length of Text
1
2
3
4
5
Stars
-0.17
Corr Coefficient


Number of Reviews Left by User and Average Rating
Average Review (# of Stars)
5
4
3
2
1
1000
2000
3000
4000
5000
# of Reviews by User
0.46
Corr Coefficient


Users' Average Rating By Date of Joining Yelp
4.0
3.5
3.0
User's Average Rating
2005
2010
2015
2020
Year

Machine Learning

- **Objective:** Use Business Attributes to Predict Ratings on Yelp.com

- **Attributes Used:** Alcohol, Credit Card Acceptance, Noise Level, Delivery, Price Range... (15 independent variables in total).

Machine Learning Cont'd

- **Model Selected:** Extreme Gradient Boosting (XGBoost)
- **Advantages (Datasets):** Build-in Algorithm that Handles Missing Values
- **Advantages (Objective):** Build-in Objective for Multi-class Classification

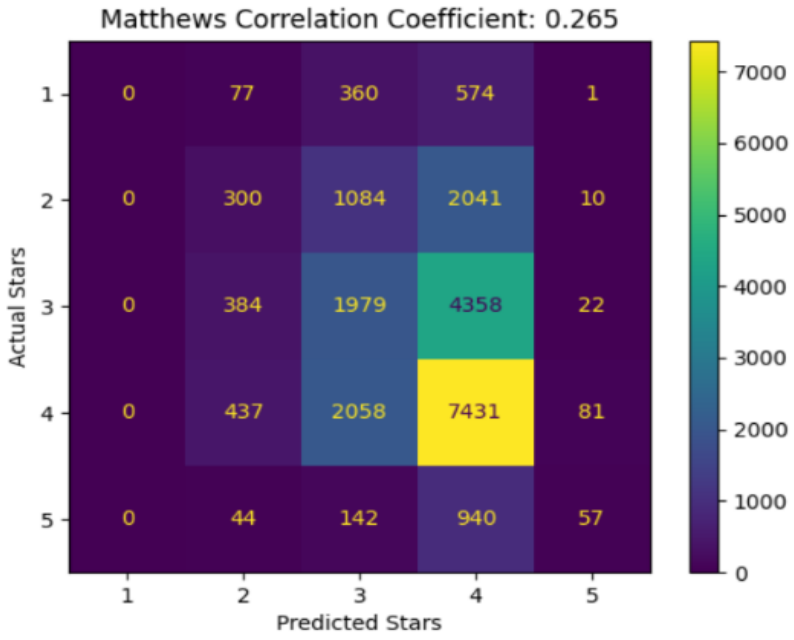
Machine Learning Cont'd

- **Model Optimization:** Grid Search
- **Parameters Tuned:**

```
▾                                XGBClassifier
XGBClassifier(base_score=0.5, booster='gbtree', callbacks=None,
              colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
              early_stopping_rounds=None, enable_categorical=False,
              eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
              importance_type=None, interaction_constraints='',
              learning_rate=0.1, max_bin=256, max_cat_to_onehot=4,
              max_delta_step=0, max_depth=5, max_leaves=0, min_child_weight=3,
              missing=nan, monotone_constraints='()', n_estimators=140,
              n_jobs=4, nthread=4, num_class=5, num_parallel_tree=1,
              objective='multi:softmax', predictor='auto', ...)
```

Machine Learning Cont'd

Confusion Matrix on Prediction of Ratings Based on Business Attributes



Machine Learning Conclusion

- The *Matthew Correlation Coefficient* is **0.265**, which is relatively low. It means that the prediction is not accurate.

- A possible explanation could be that the Yelp categorized some numeric variables to categorical variables, which loses some accuracy.

- There are other factors that affect rating of restaurants not included in Business Attributes.

Recommendations & Takeaways

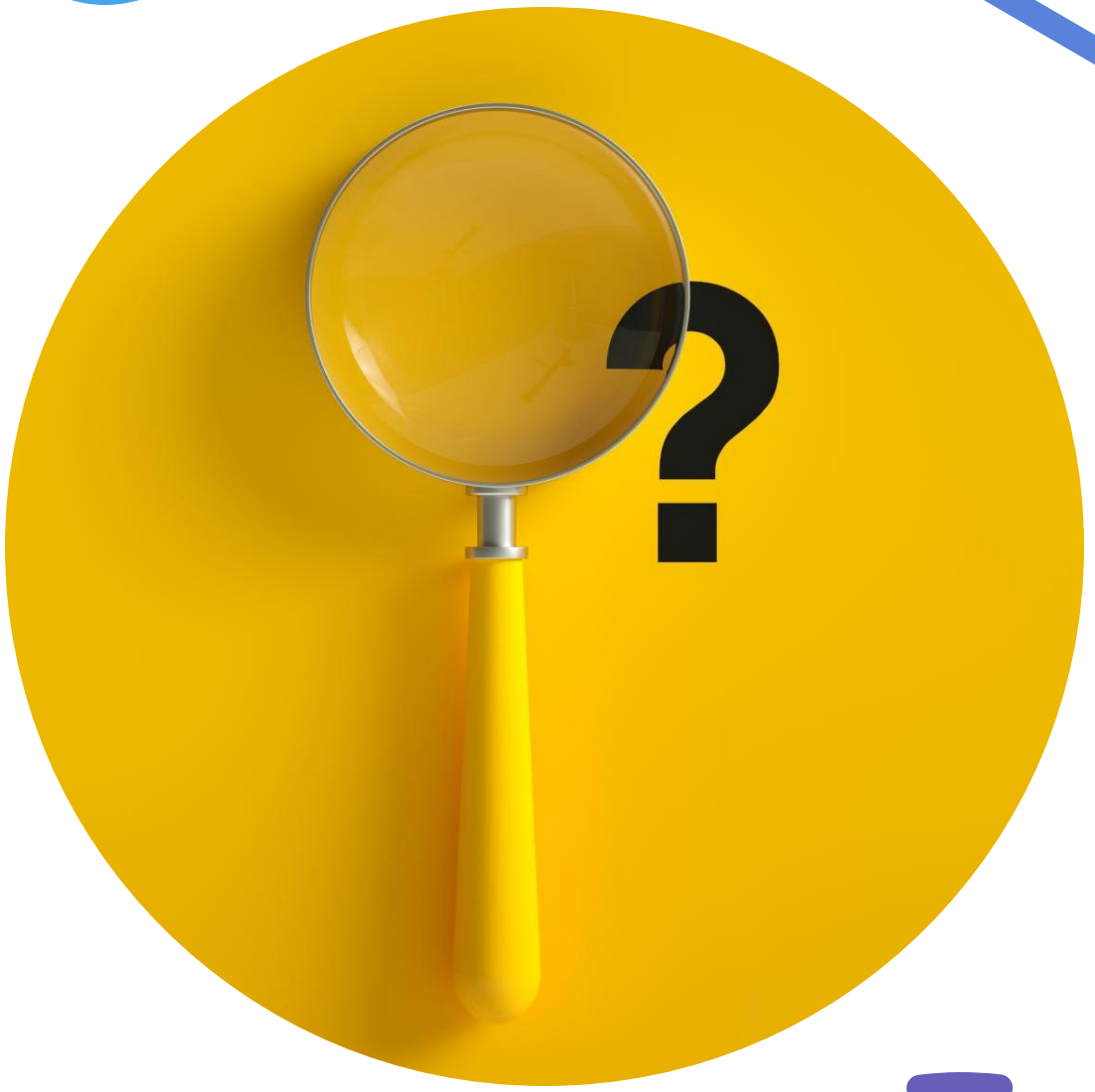
- Restaurants should focus on their individual advantages instead of adding features in business attributes

- User reviews could be promoted

- Besides quality of food, service is also another important part of a restaurant ratings

Data Sources

- Yelp Educational Dataset
- U.S. Bureau of Labor Statistics Unemployment Rate
- U.S. Census Educational Attainment Dataset
- U.S. Census Income Dataset



Q&A