Yelp Restaurant Star Rating Prediction

Zi Leng Chua, Anh Nguyen, Jiacheng Yu

Problem Statement

To predict restaurants' star rating in Pennsylvania & Florida.

Background Information

67% of customers will consider leaving a review for a positive experience, while **40%** will consider leaving a review for a negative experience. And **98%** of customers read online reviews for local businesses.

Only **3%** of customers said they would consider using a business with an average star rating of **2** or fewer stars.

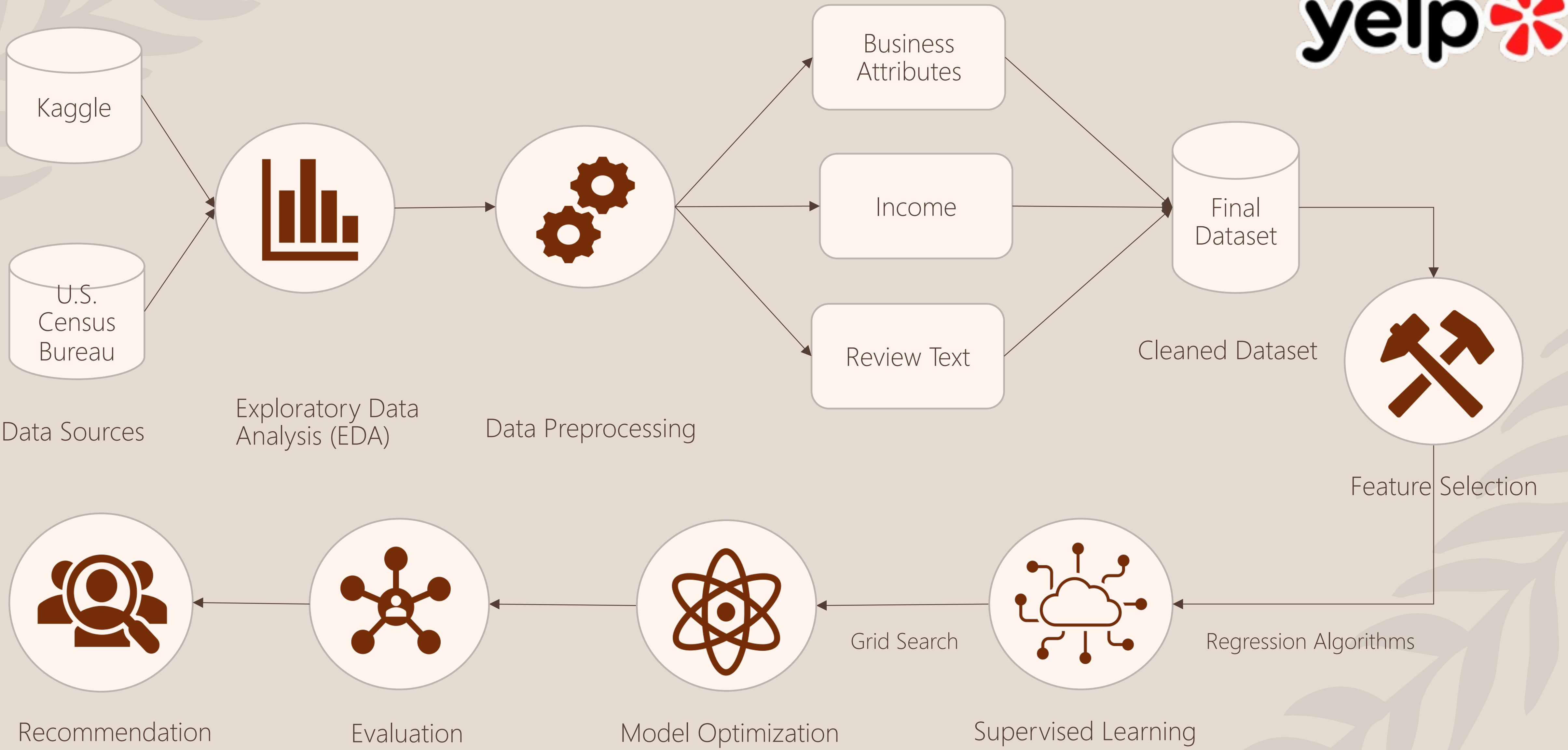
More consumers use Yelp to evaluate local businesses than ever before. In 2021, **53%** did, but the year before that, only **32%** did.

All statistics in this page are from Local Consumer Review Survey 2022: Customer Reviews and Behavior (brightlocal.com)

Contribution

Our focus of the study can be widely applied by many food businesses to educate owners on factors that have great influences on their ratings. This can help businesses improve their customer satisfaction and attract new customers.

Process Flow Chart



Kaggle

U.S. Census Bureau

Data Sources

Exploratory Data Analysis (EDA)

Data Preprocessing

Business Attributes

Income

Review Text

Final Dataset

Cleaned Dataset

Feature Selection

Regression Algorithms

Supervised Learning

Grid Search

Model Optimization

Evaluation

Recommendation

Data

Yelp Dataset

Business

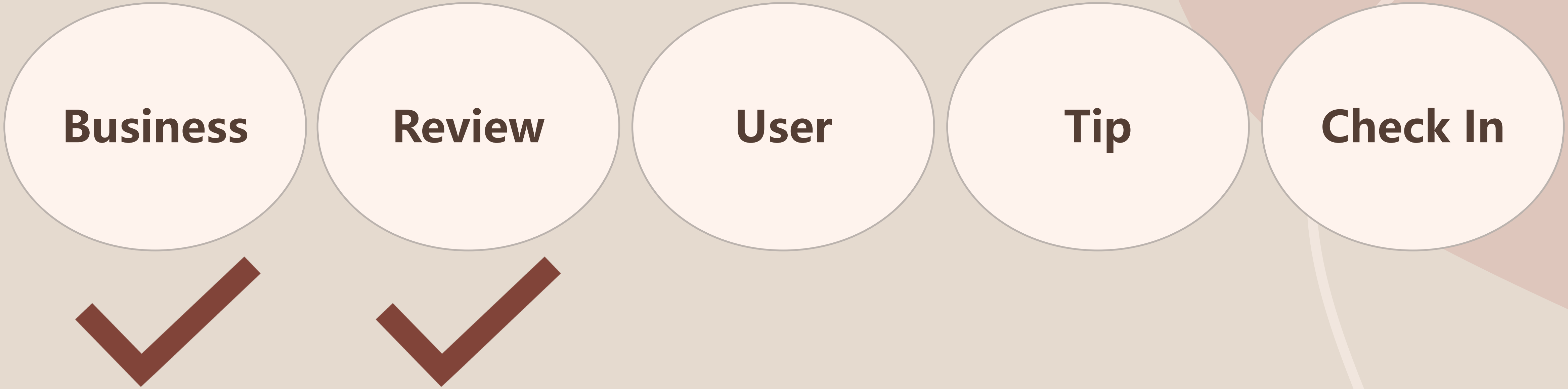
Review

User

Tip

Check In

Yelp Dataset



- Business ✓
- Review ✓
- User
- Tip
- Check In

Data Description

- Yelp Dataset:
  - 8 metropolitan areas in the USA and Canada.
  - 6 files including 5 JSON data files and 1 PDF file that is about user agreement.

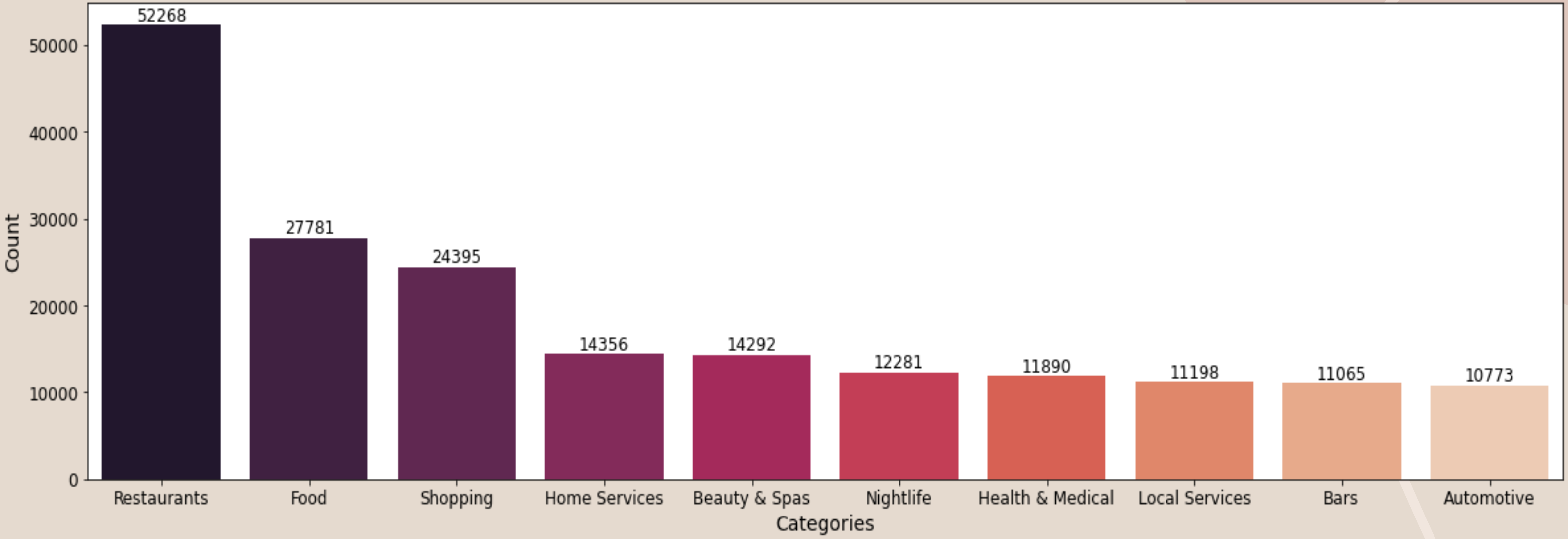
Data File	Data Size	Observations	Features
business	119 MB	150,346	14
check-in	287 MB	131,930	2
review	5.34 GB	6,990,280	9
tip	181 MB	908,915	5
user	3.36 GB	1,987,897	22
Total Size	9.3 GB		

- 2020 Census Income Data for PA and FL postal zip codes: (3 columns x 2781 rows)
  - 3 columns:
  - postal_code: all 5-digit postal zip codes fully/partially contained within PA and FL
  - Total number of households in the zip code
  - ACS 5-year estimate average household income of each zip code

Exploratory Data Analysis

Exploratory Data Analysis (Business)

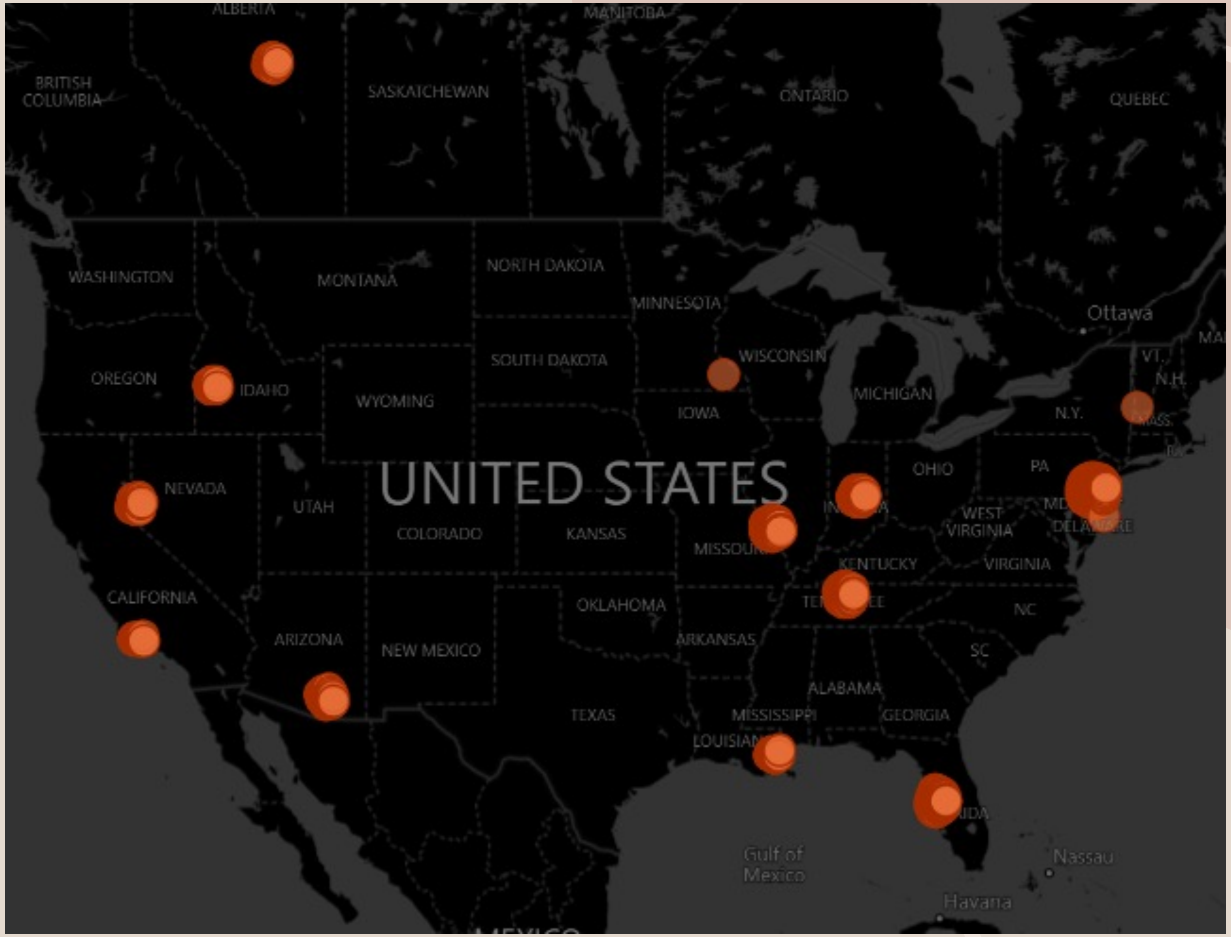
Top 10 Business Categories by Number of Businesses



Exploratory Data Analysis (Business)

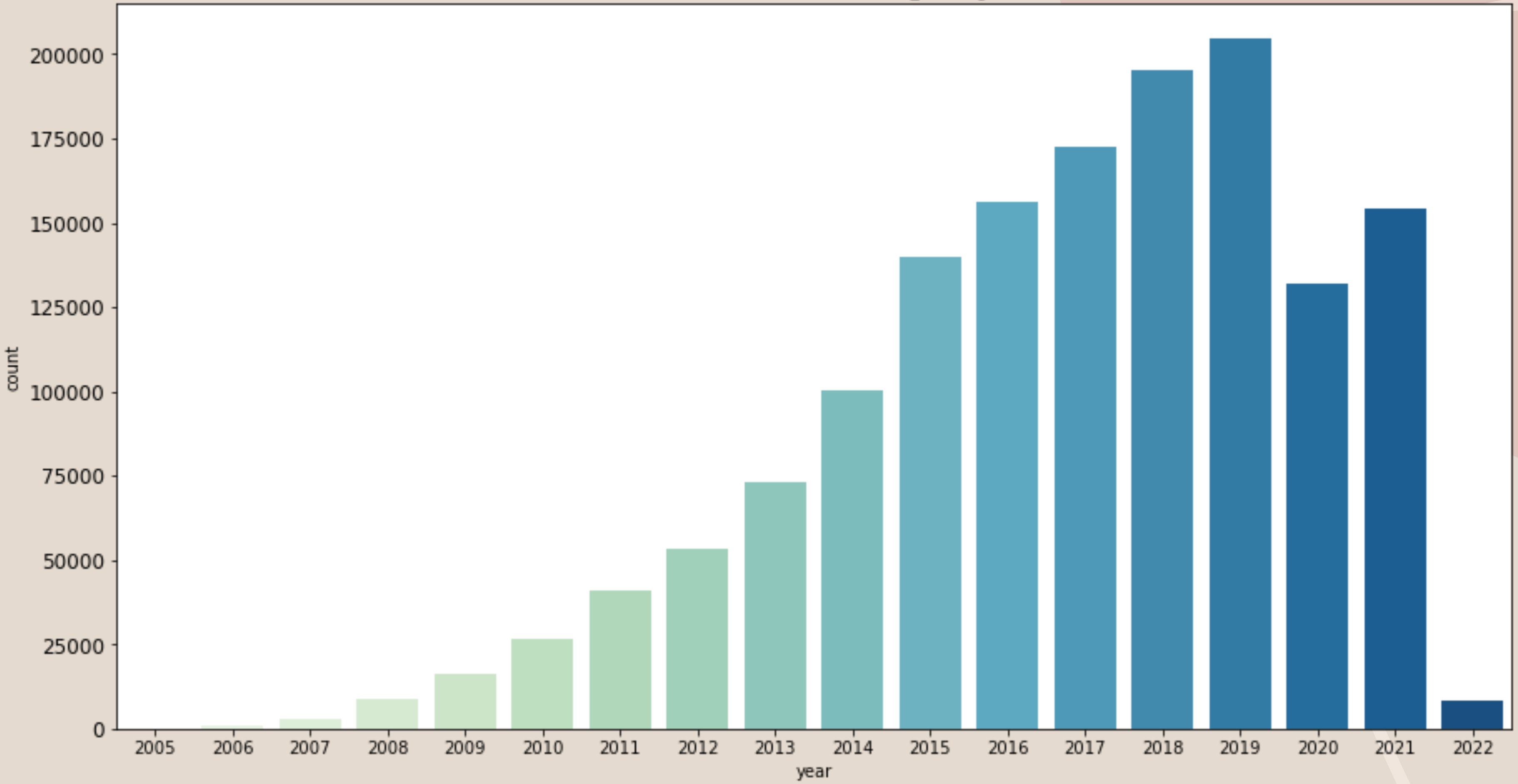
Most Popular States by Number of Businesses

- 23% Pennsylvania
- 18% Florida
- <10% Other States



Exploratory Data Analysis (Review)

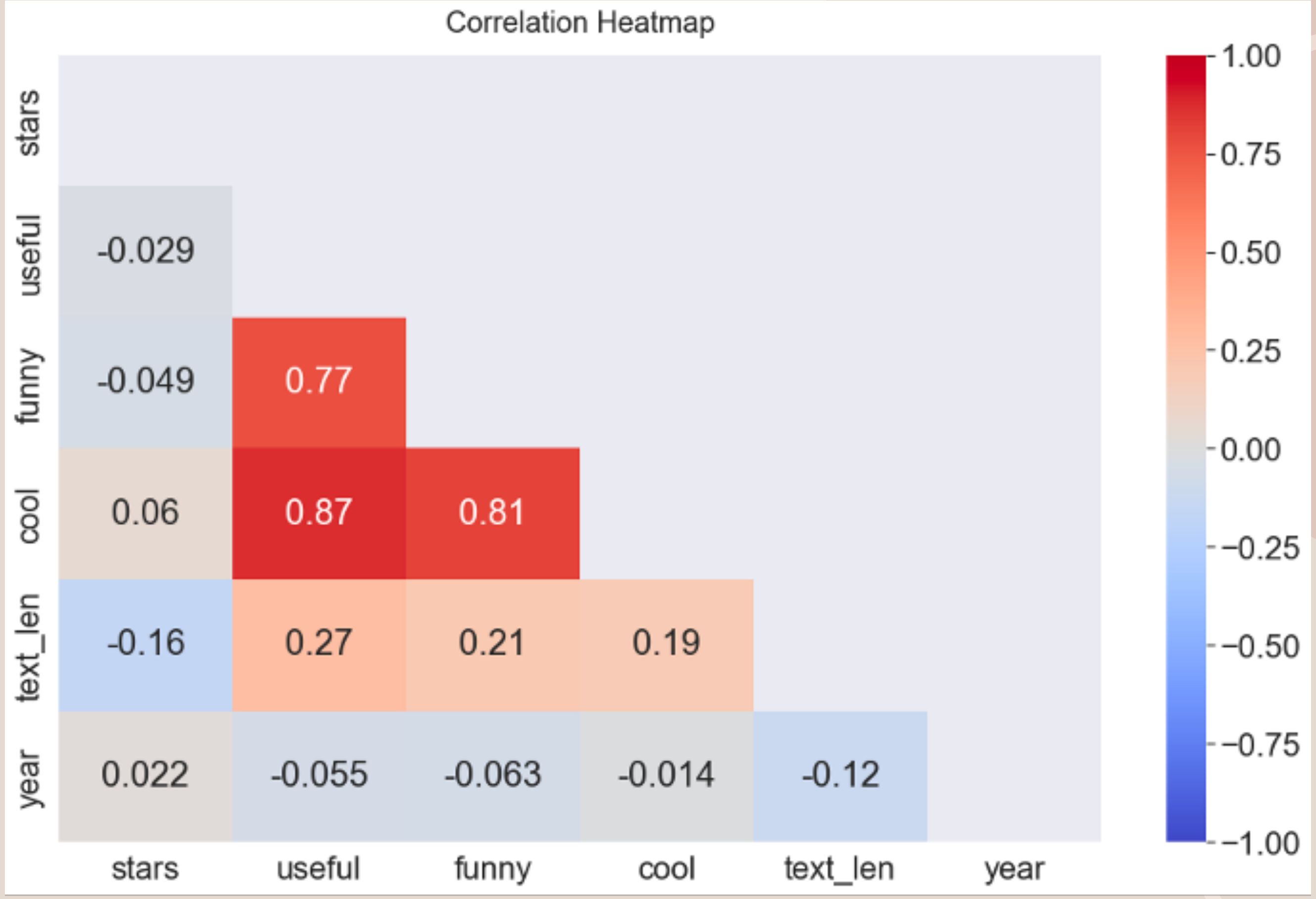
Distribution of Users' Reviews



Exploratory Data Analysis (Review)

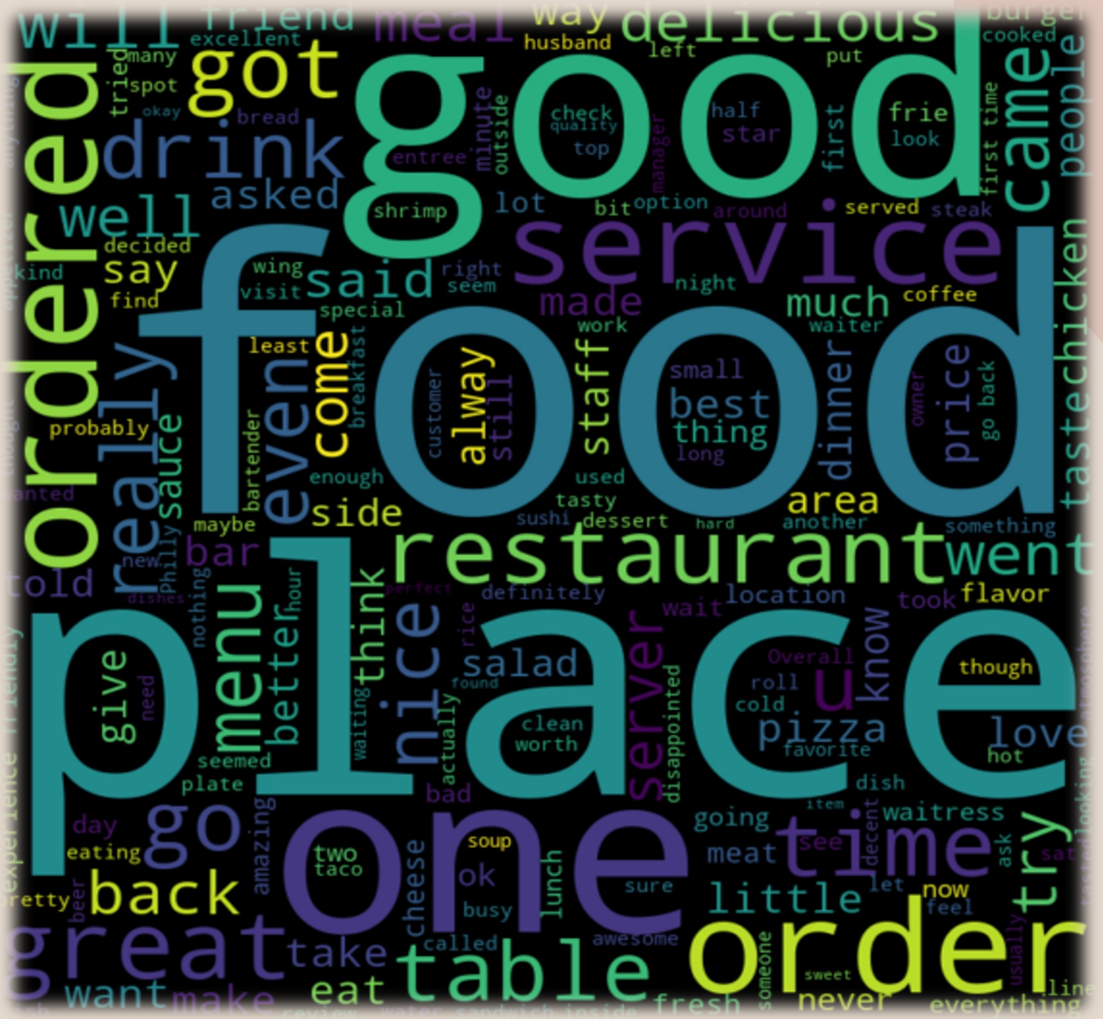
Review Correlation Analysis

- High correlation among votes (funny, useful, cool)



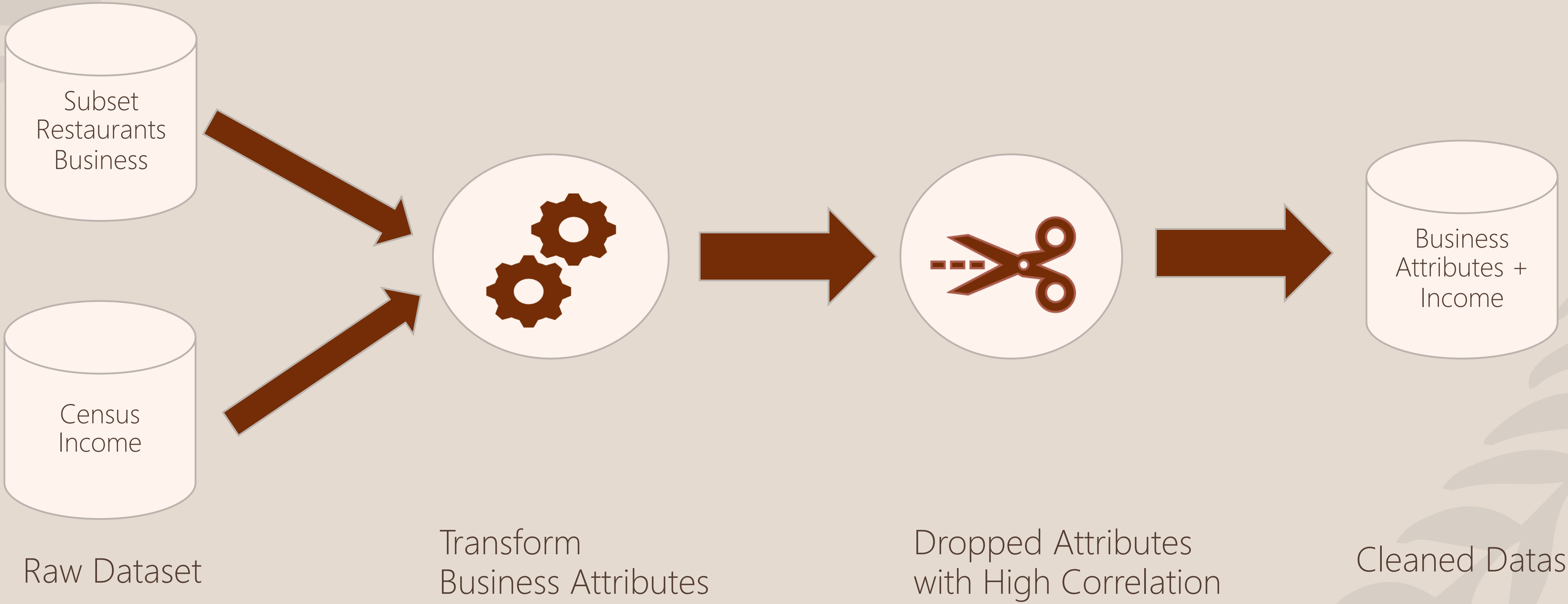
Exploratory Data Analysis (Review)

Top Words in Reviews



Pre-Processing

Pre-Processing (Business)



Subset Restaurants Business

Census Income

Raw Dataset

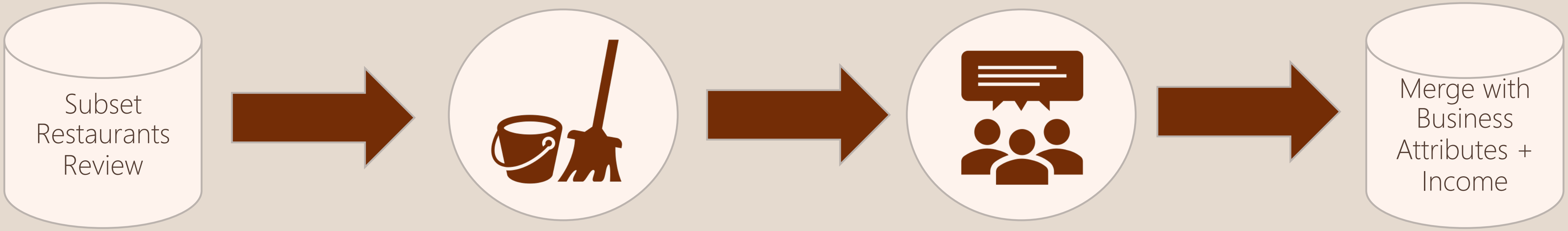
Transform Business Attributes

Dropped Attributes with High Correlation

Business Attributes + Income

Cleaned Dataset

Pre-Processing (Review)



Subset Restaurants Review

Merge with Business Attributes + Income

User Review Dataset

Text Cleaning

Sentiment Analysis

Final Dataset

- Convert to Lowercase
- Remove Numbers, Punctuation, Stop Words, URL
- Convert Contraction Expression to Formal Auxiliary Verbs
- Word Tokenization
- Lemmatization

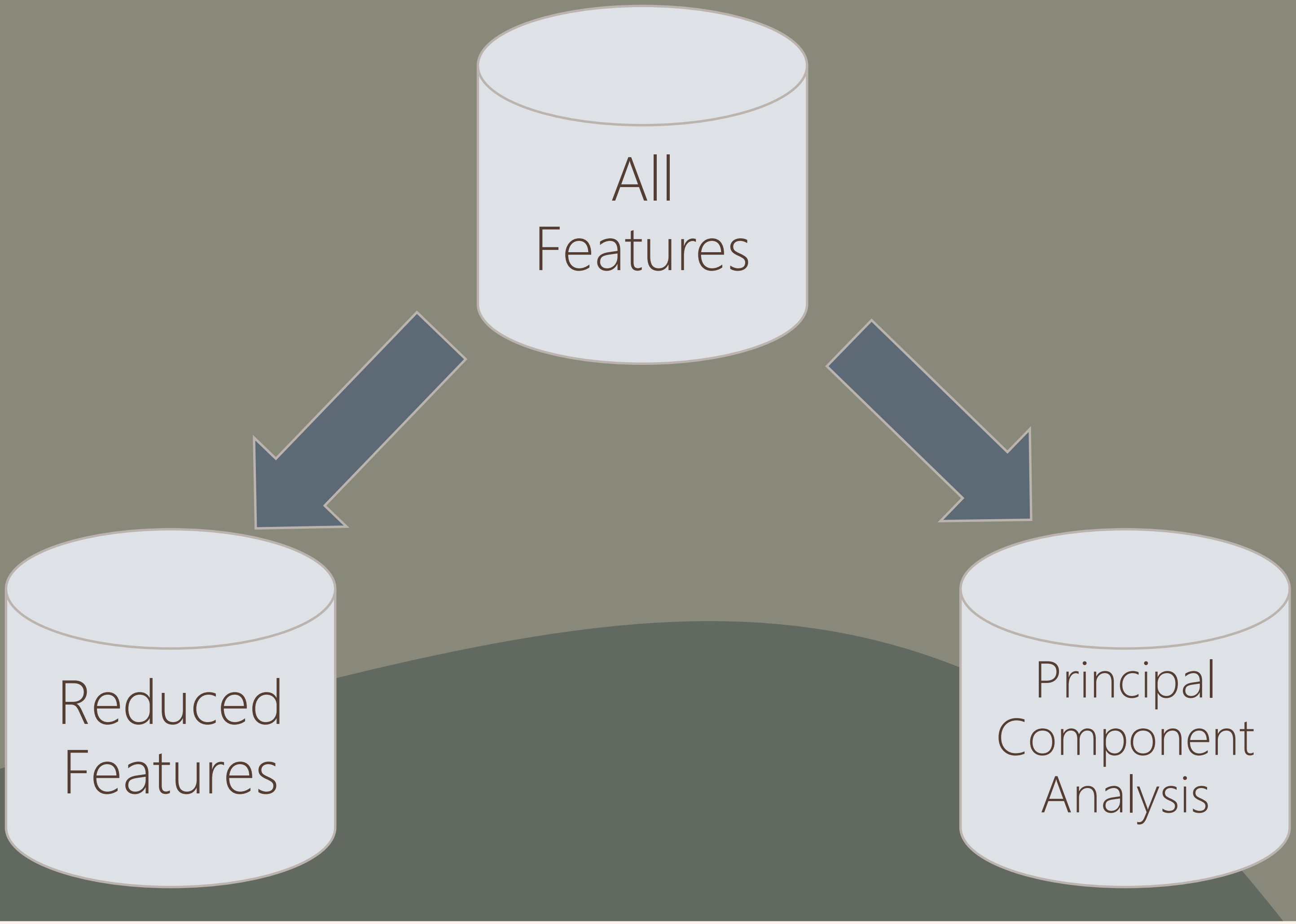
Polarity Scores

- Positive
- Negative
- Neutral

Models
&
Evaluations

Feature Sets

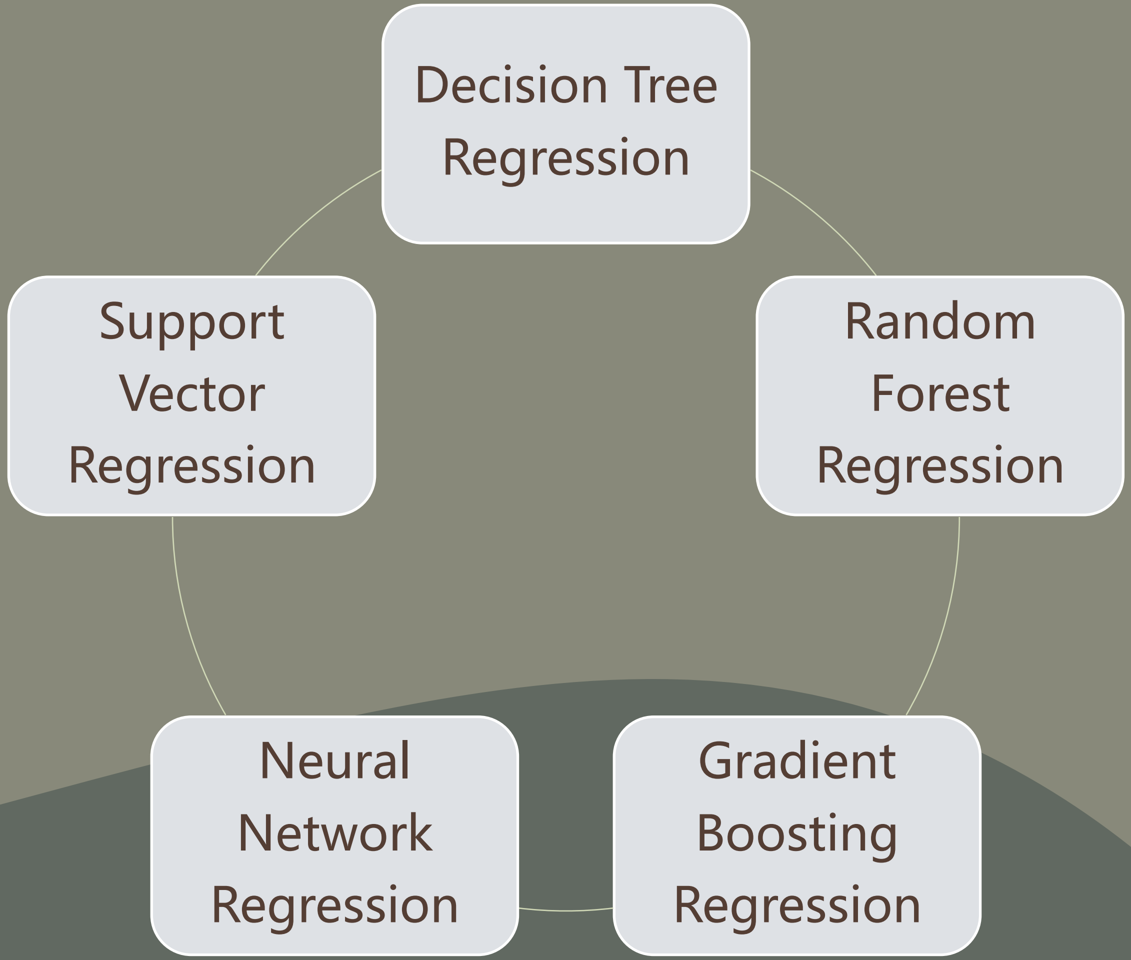
47 features



- Selected 5 to 7 features

- Selected 30 Components

Algorithms



Decision Tree Regression

Random Forest Regression

Gradient Boosting Regression

Neural Network Regression

Support Vector Regression

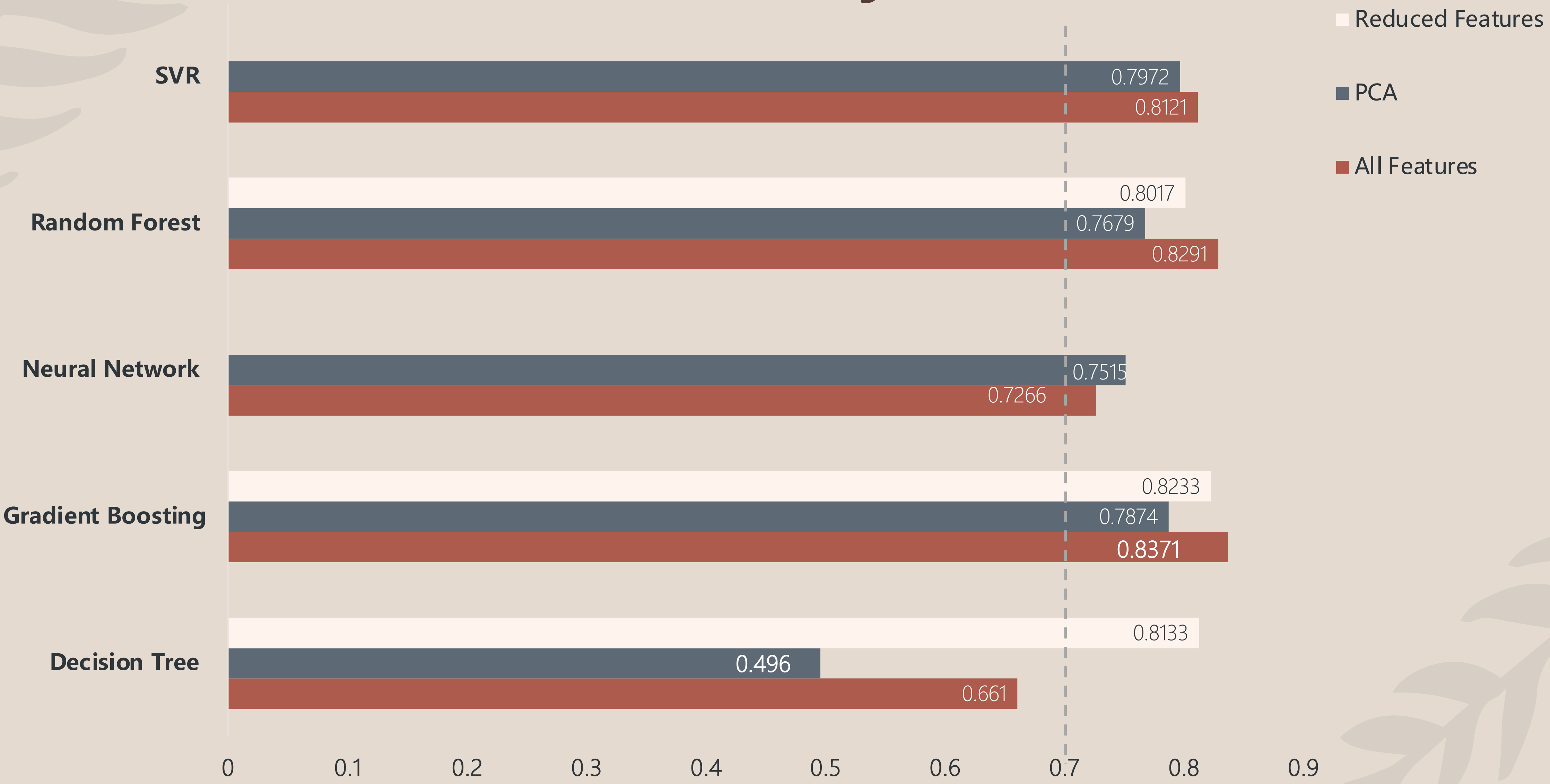
Other Techniques

K-fold Cross Validation

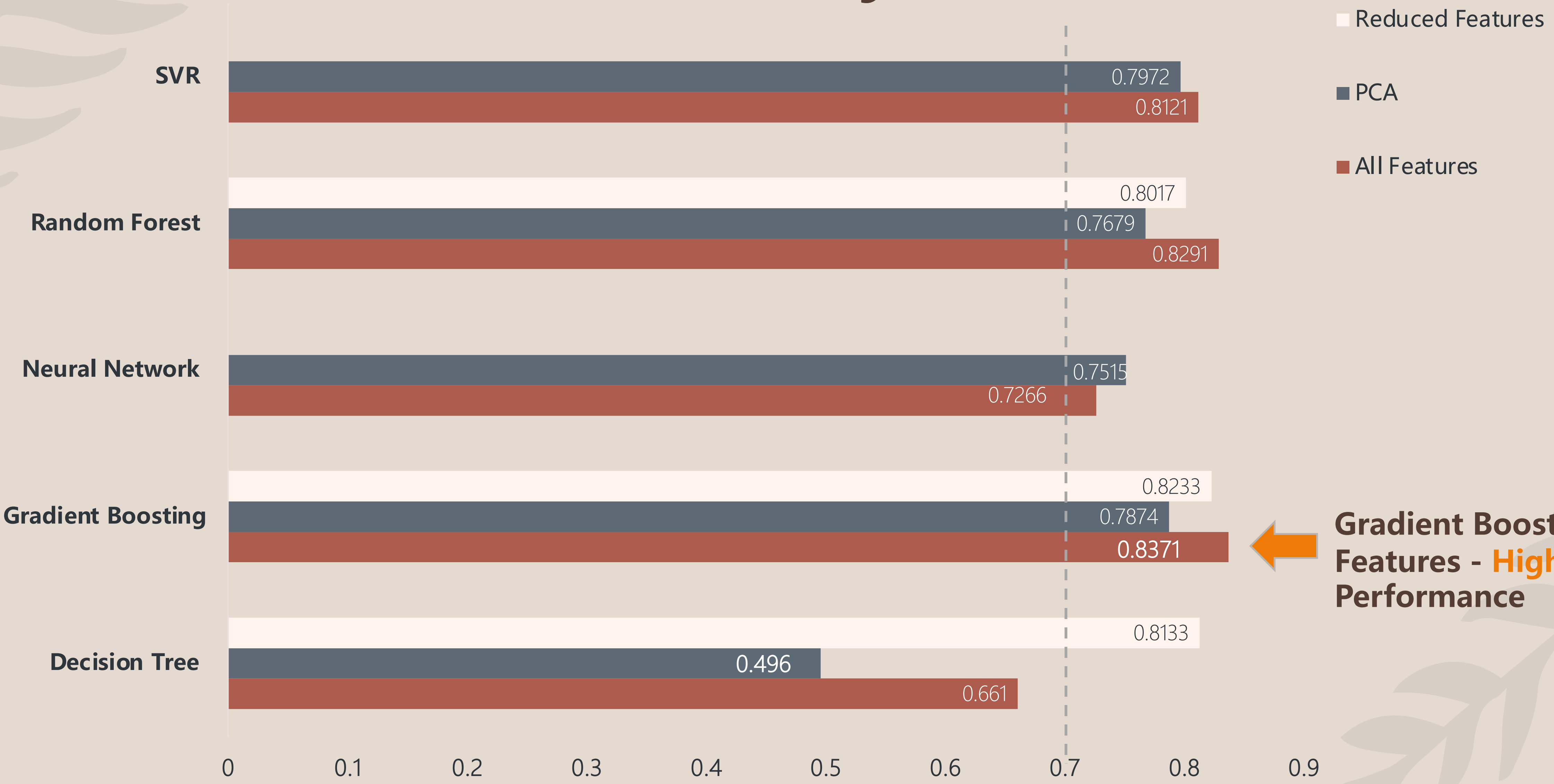
Hyperparameter Tuning

Clustering

Evaluation Summary

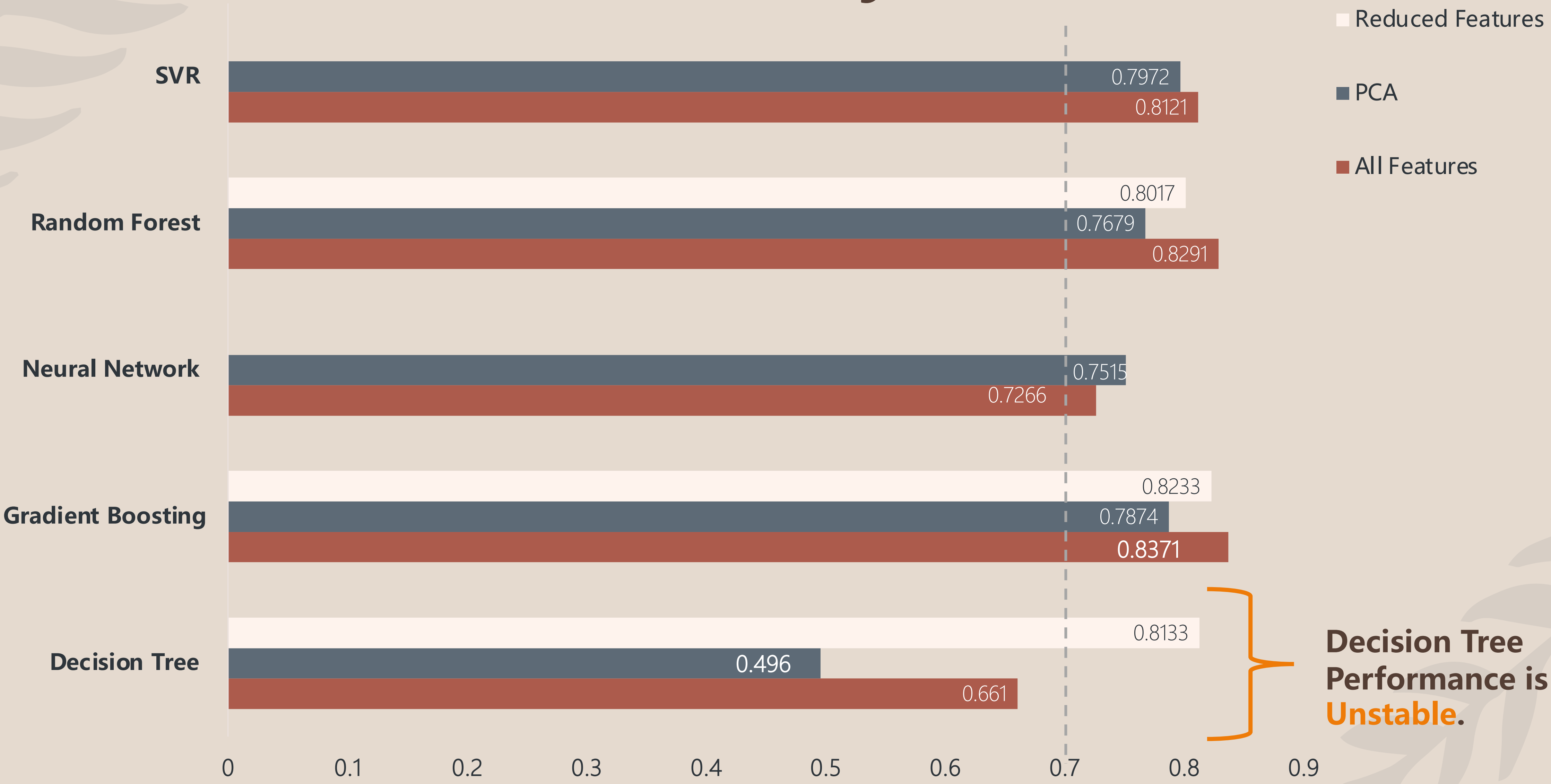


Evaluation Summary



Gradient Boosting All Features - Highest Performance

Evaluation Summary



Decision Tree Performance is Unstable.

Evaluation Summary

Models	Datasets	K-fold Score	MSE	RMSE	R Squared
Decision Tree	All Features	0.6440	0.1520	0.3899	0.6610
Gradient Boosting	**All Features**	**0.8361**	0.0731	0.2703	**0.8371**
Neural Network	All Features	0.7321	0.1226	0.3502	0.7266
Random Forest	All Features	0.8257	0.0766	0.2768	0.8291
SVR	All Features	0.8042	0.0843	0.2903	0.8121
Decision Tree	**PCA**	**0.5045**	0.2261	0.4755	**0.4960**
Gradient Boosting	PCA	0.7805	0.0954	0.3088	0.7874
Neutral Network	PCA	0.7171	0.1114	0.3338	0.7515
Random Forest	PCA	0.7693	0.1041	0.3226	0.7679
SVR	PCA	0.7915	0.0909	0.3016	0.7972
Decision Tree	Reduced Features	0.8097	0.0837	0.2893	0.8133
Gradient Boosting	Reduced Features	0.8219	0.0793	0.2815	0.8233
Random Forest	Reduced Features	0.7992	0.0890	0.2983	0.8017

Discussion & Conclusion

Discussion and Conclusion

DOMAIN KNOWLEDGE

METHODOLOGICAL CONTRIBUTIONS

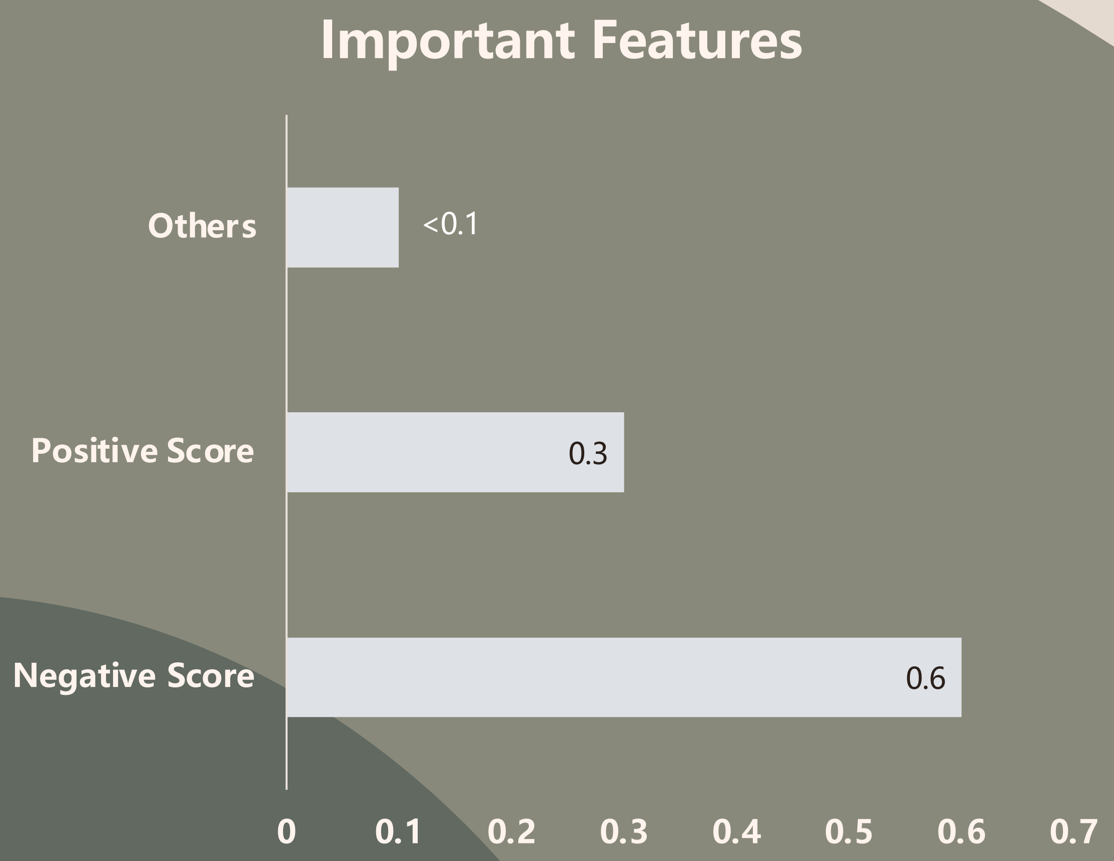
LIMITATIONS

FUTURE PROJECTS

Domain Knowledge / Business Insights

Sentiment Scores, Review Counts, Business Attributes

- **Important in influencing Business Star Rating**
- **Encourage User to Leave Review** to understand Customers' Opinions & Expression
- **Give Promotions or Referral Program**
- **Alcohol Full Bar** is shown as one of the most important features in Business Attributes



Methodological Contributions

HANDLING DATASET

- Store file as a list of dictionaries in the memory
- Subset data
- Split attributes

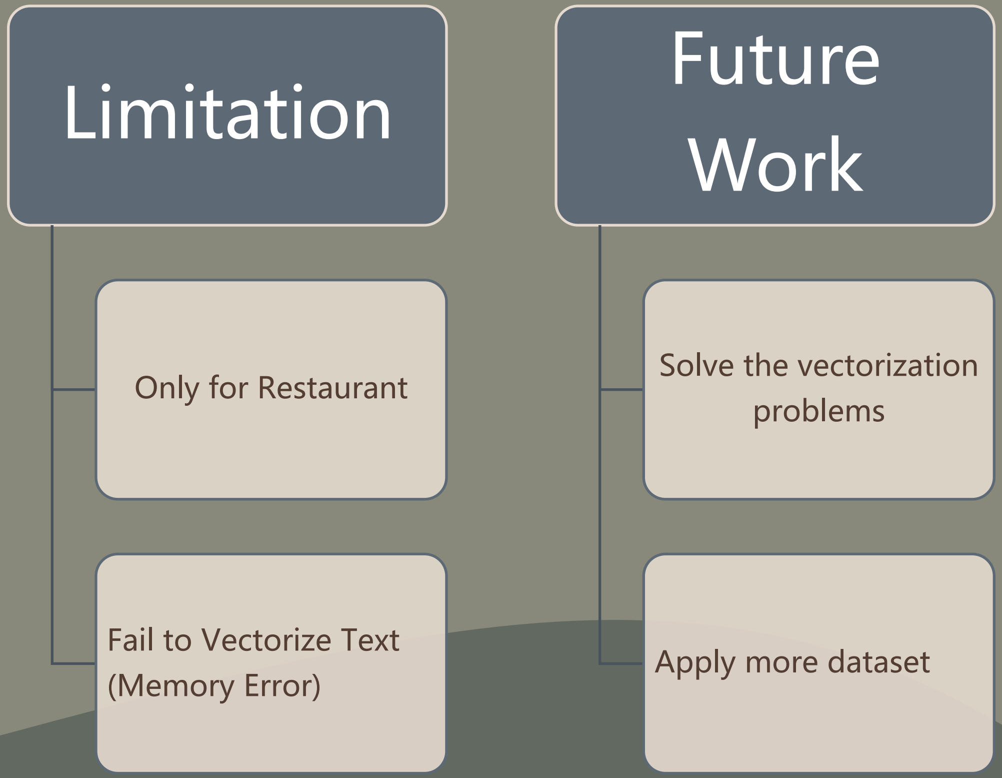
TEXT PROCESSING

- Time Consuming
- Contraction Expressions increase processing run time

STATISTICAL TECHNIQUES

- Does not meet expectation
- Generate lower performance
- PCA
- Feature Selection
- K-fold Cross Validation
- Clustering

Limitations and Future Work



Limitation

- Only for Restaurant
- Fail to Vectorize Text (Memory Error)

Future Work

- Solve the vectorization problems
- Apply more dataset

Q&A

THANK YOU