Yelp Business Reviews:
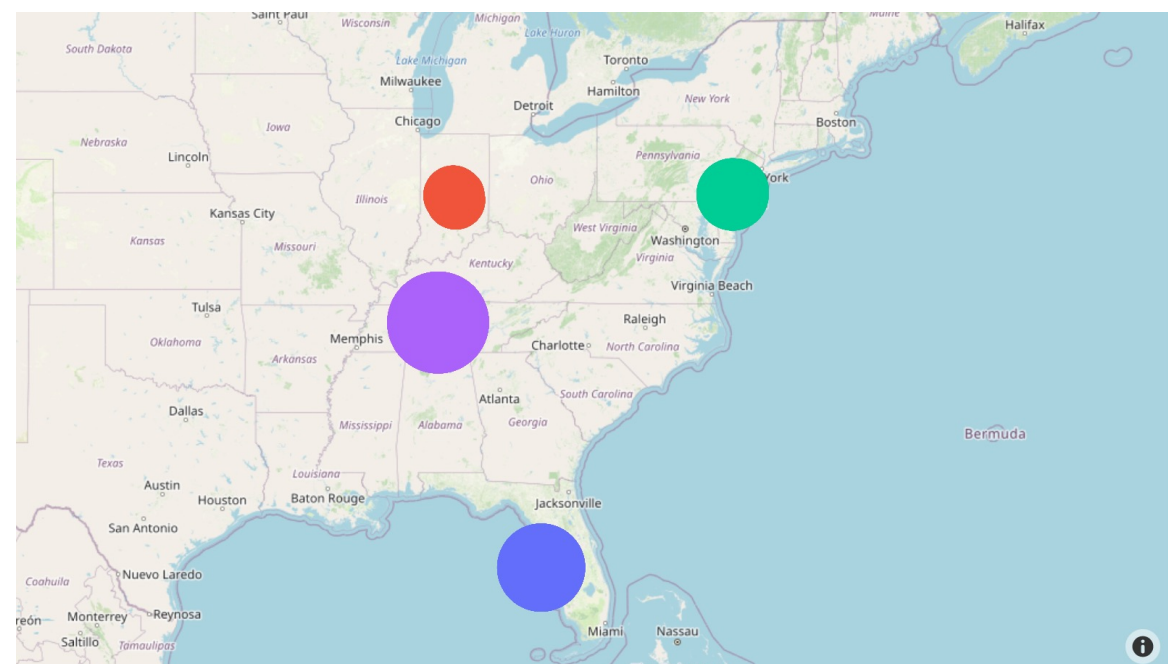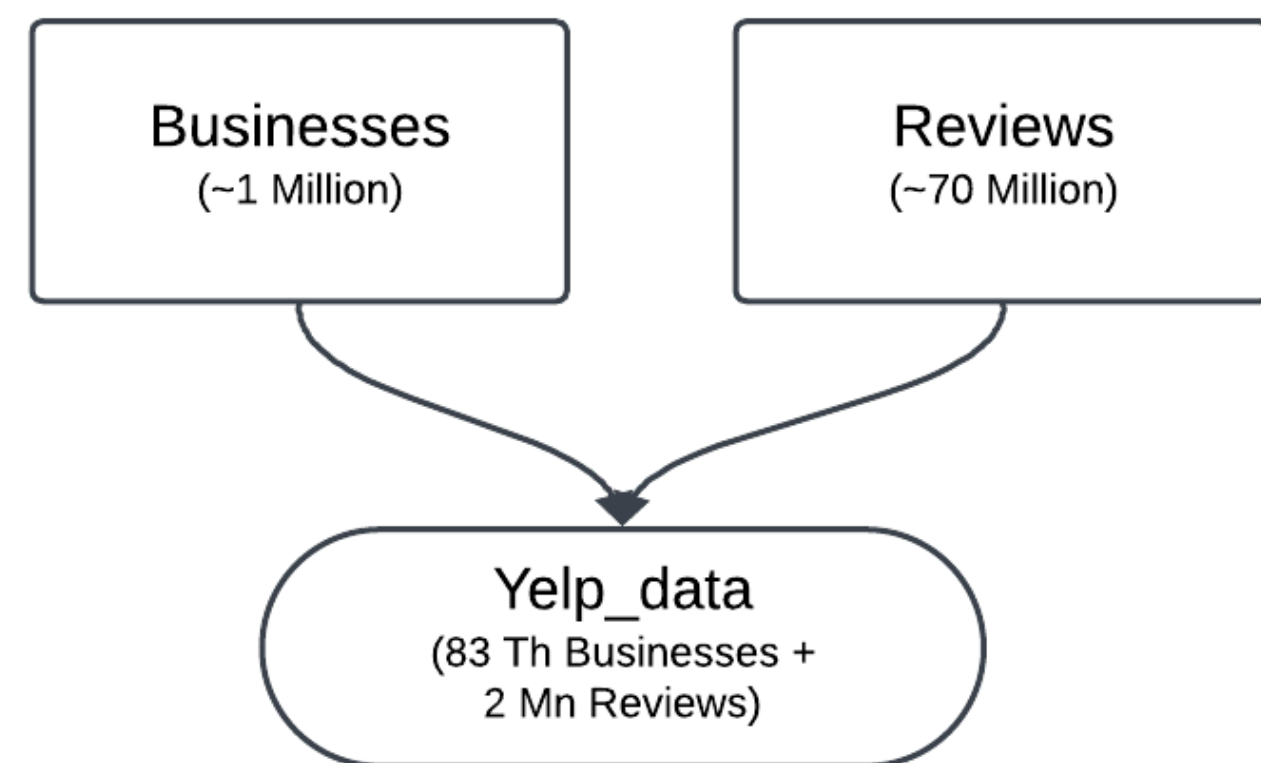# Indiana, Florida, Tennessee, and Pennsylvania

Devanshi Patel, Nikita Bedekar, Sneh Bhandari, Tanvi Mahajan
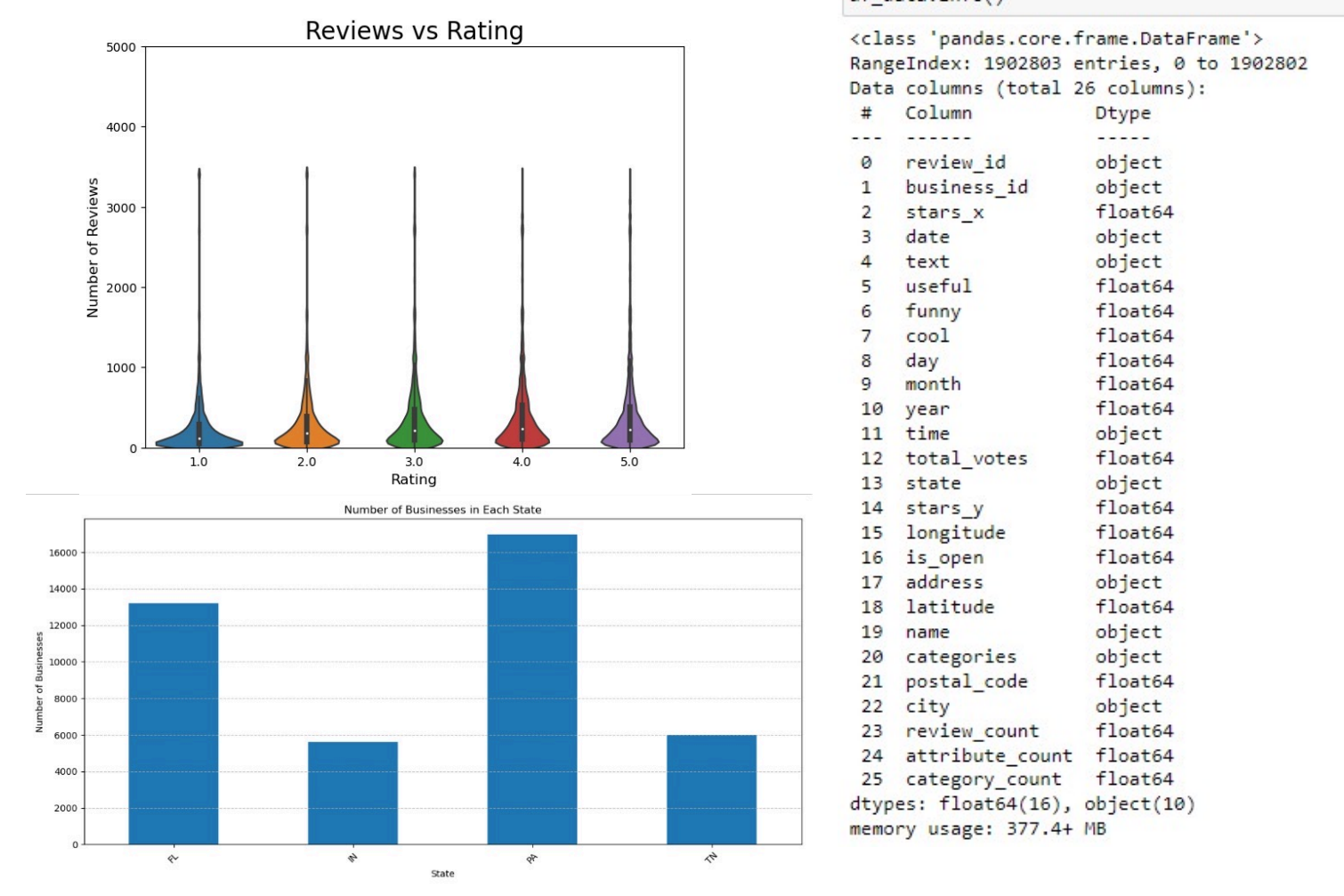Department of Statistics, Rutgers University

Abstract

In this project, we explore Yelp's rich dataset, sourced directly from their official website, to uncover insights within business reviews. Our focus is on four diverse states: Indiana (IN), Florida (FL), Tennessee (TN), and Pennsylvania (PA). Through meticulous data wrangling, we clean and prepare the data for analysis. Utilizing sentiment analysis techniques, we delve into the underlying sentiments expressed in the reviews. Additionally, employing n-gram analysis, we dissect the textual data to identify prevalent phrases and patterns. By amalgamating these methodologies, we aim to provide a comprehensive understanding of the trends and sentiments prevalent across businesses in the selected states on Yelp.
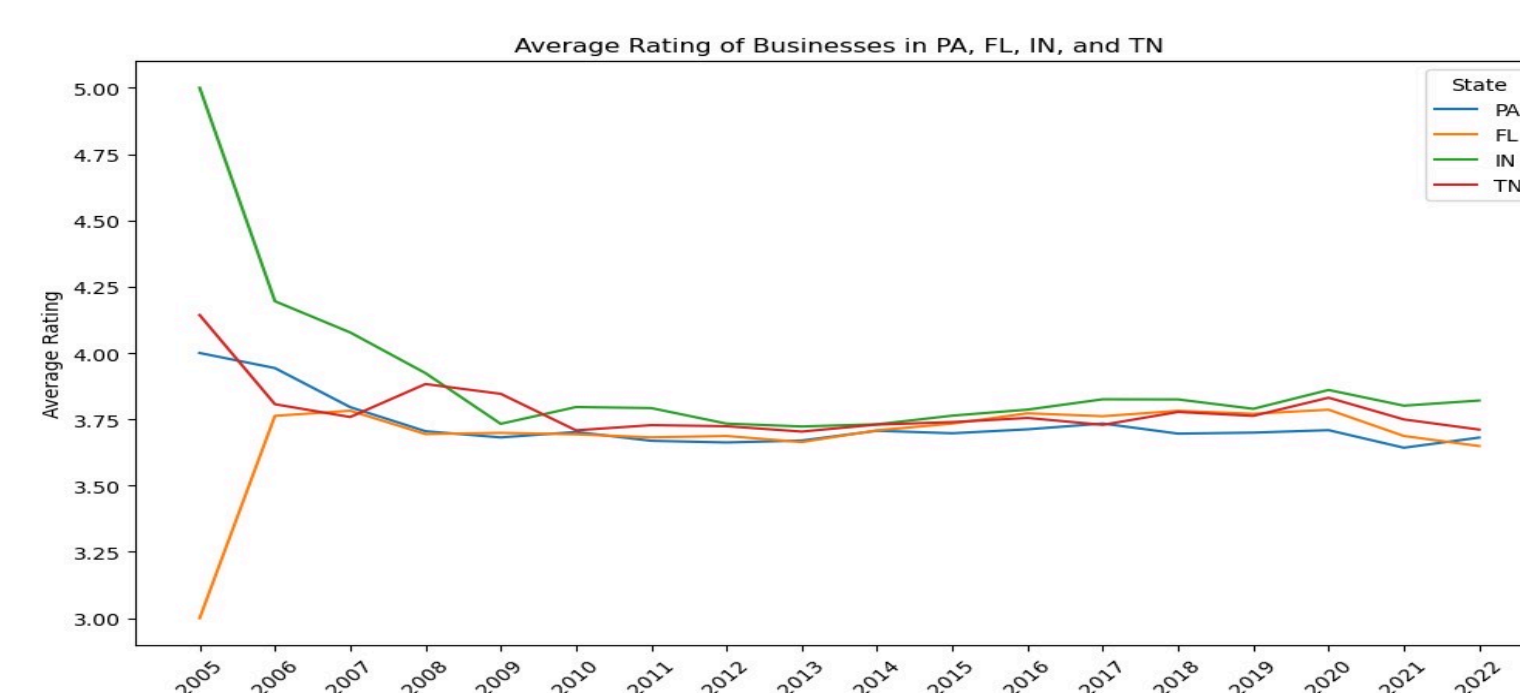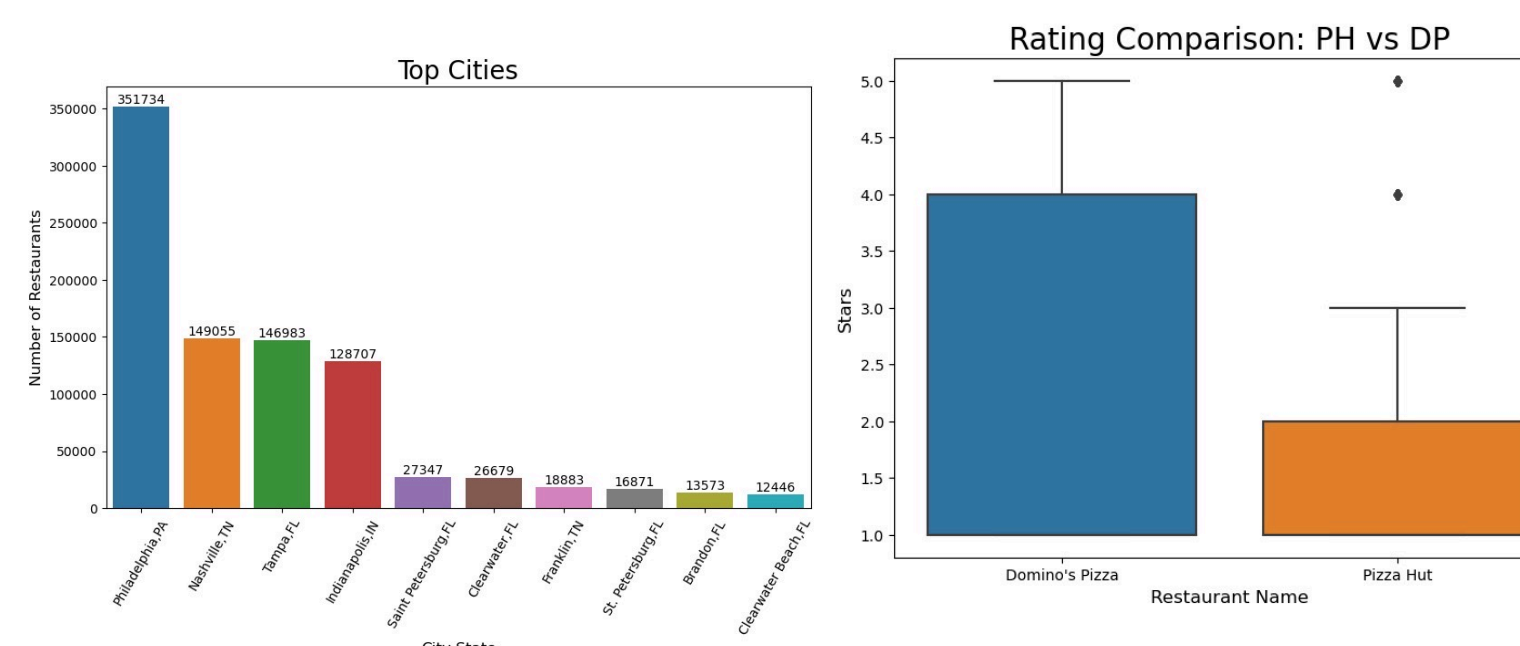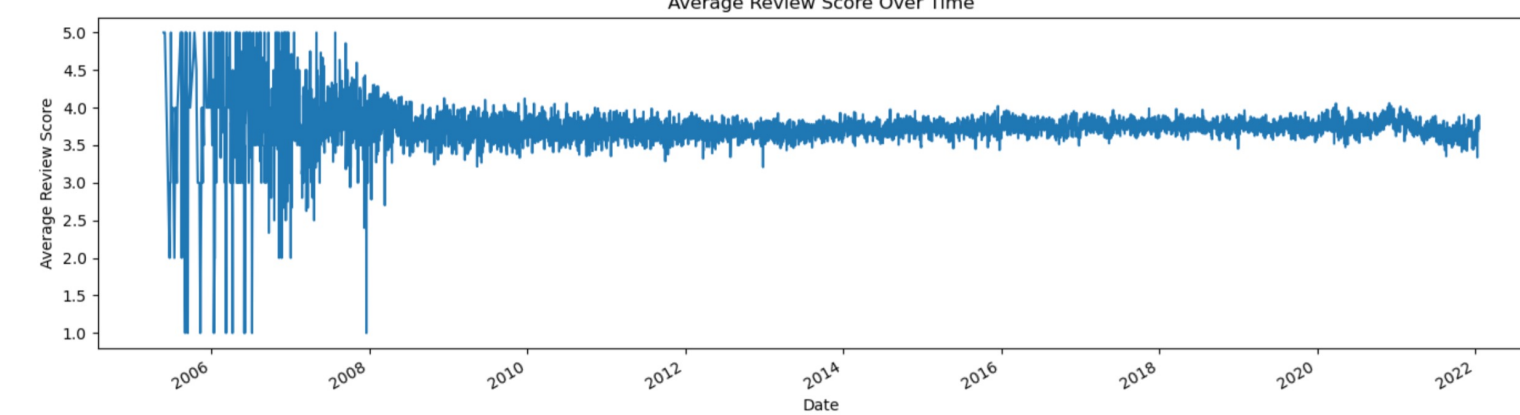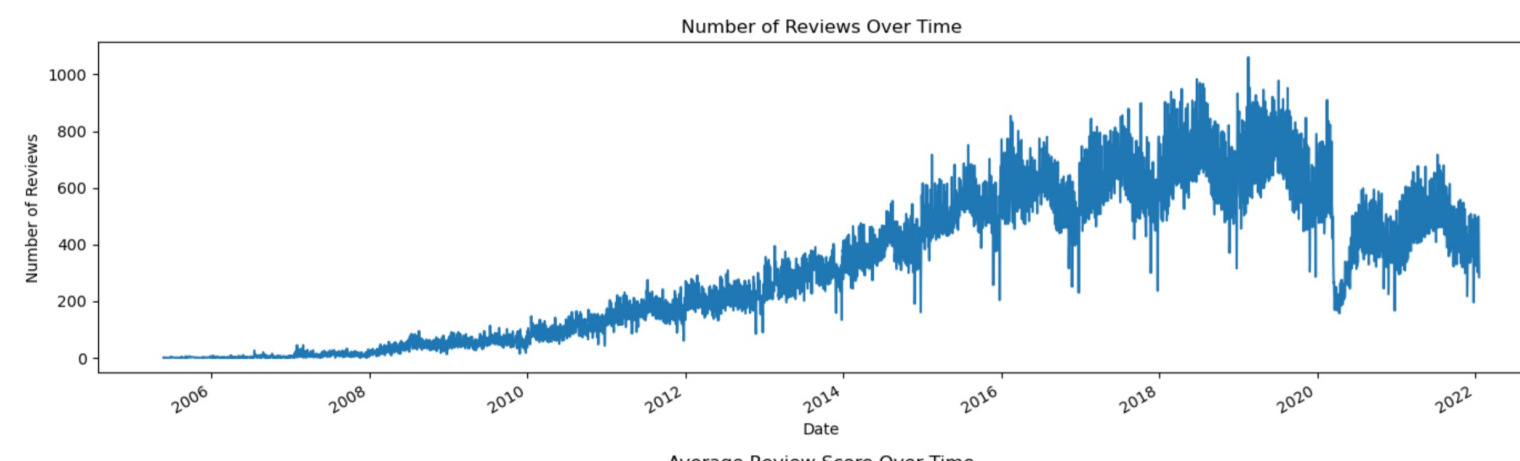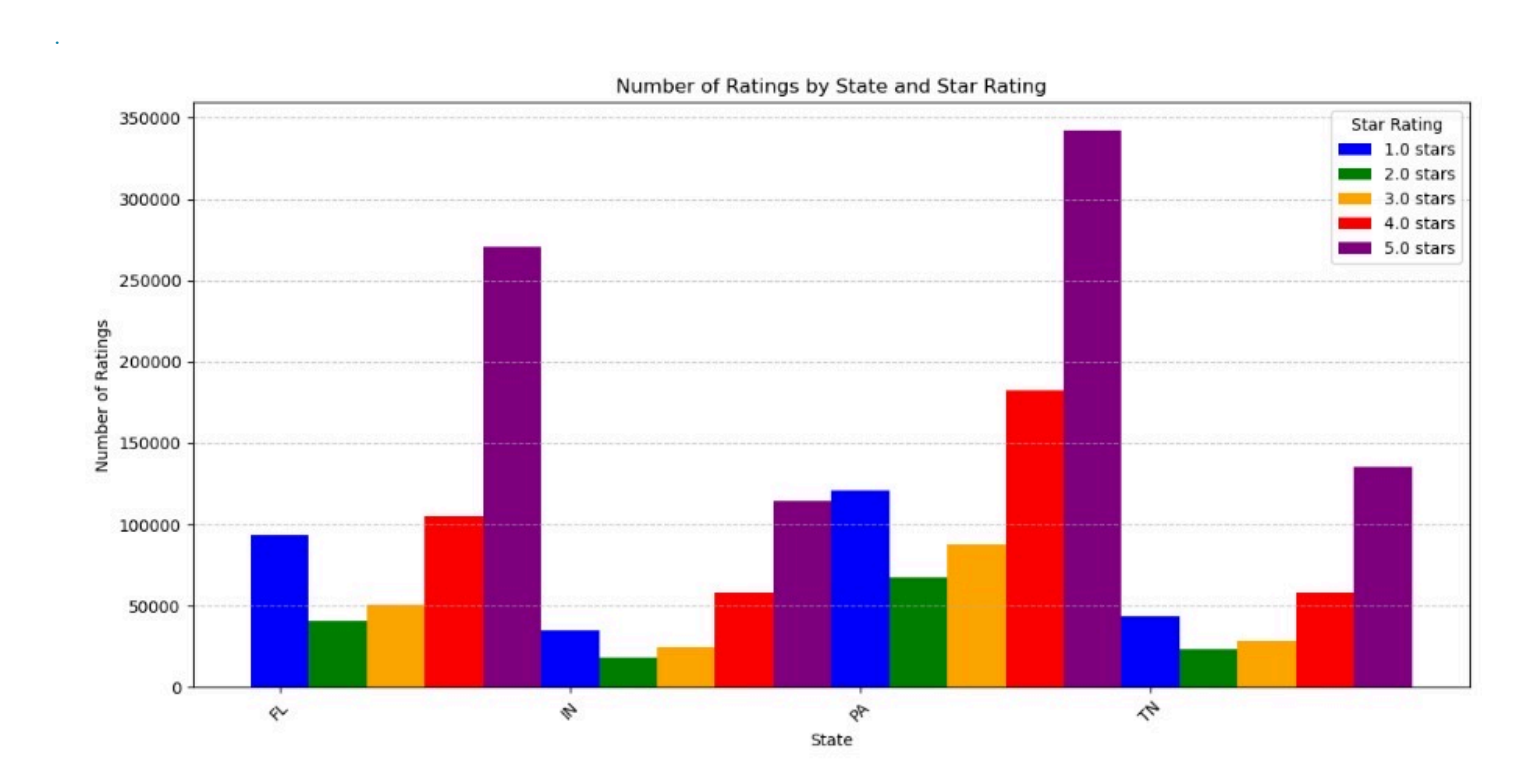
Data

Yelp is a prominent platform where users share reviews and opinions about businesses ranging from restaurants to salons.

Summary Statistics

Methodology

Data Enrichment: The missing addresses and postal codes were filled using latitude-longitude values with the help of geopy library.

Data Storage: Data was stored in three different formats: csv, Amazon S3, and SQLite.

Sentiment Analysis: It is the process of extracting emotional insight from text data, categorizing it as positive, negative, or neutral. It was conducted for all 4 states and the results were summarized in a word cloud for all three sentiments: positive, negative, and neutral.

N-Gram Analysis: N-gram analysis is a technique in natural language processing that involves breaking down text into sequences of contiguous words or characters of length 'n'. Uni-gram and Bi-gram analysis were performed to find the most used 1 and 2 sequence words.
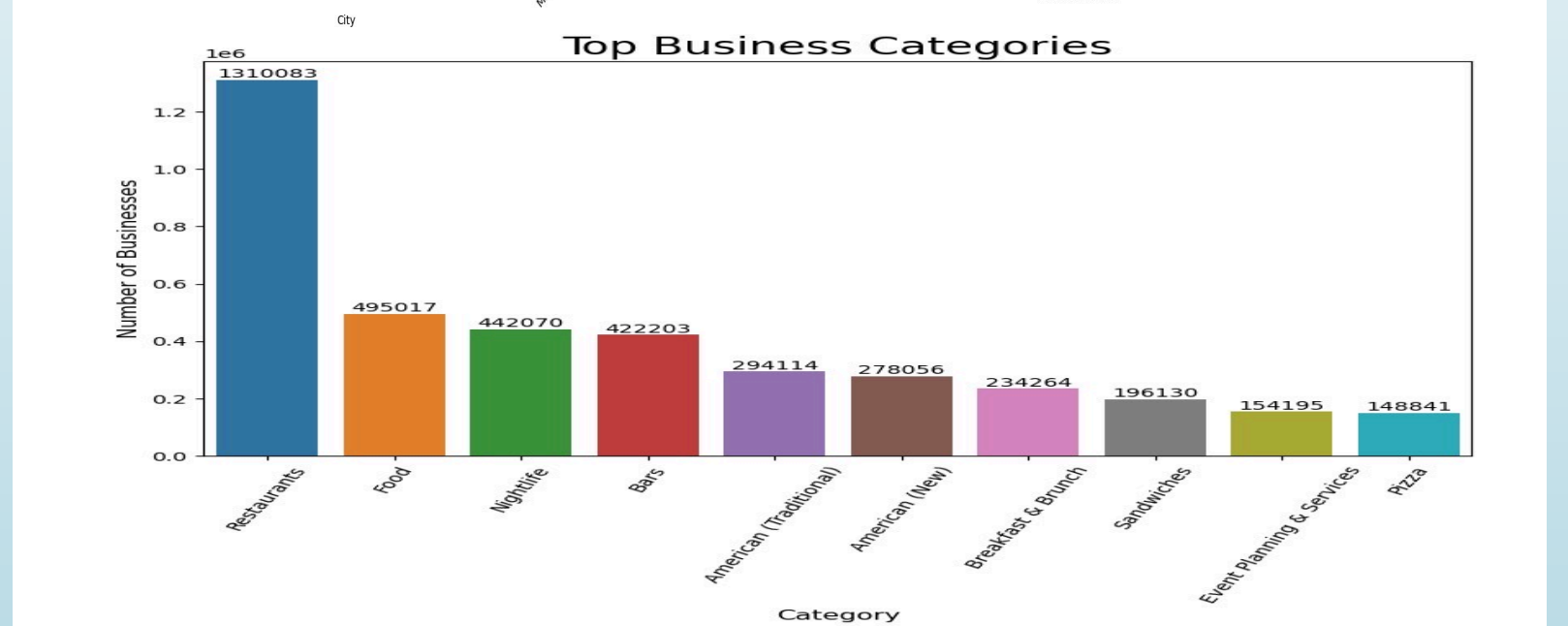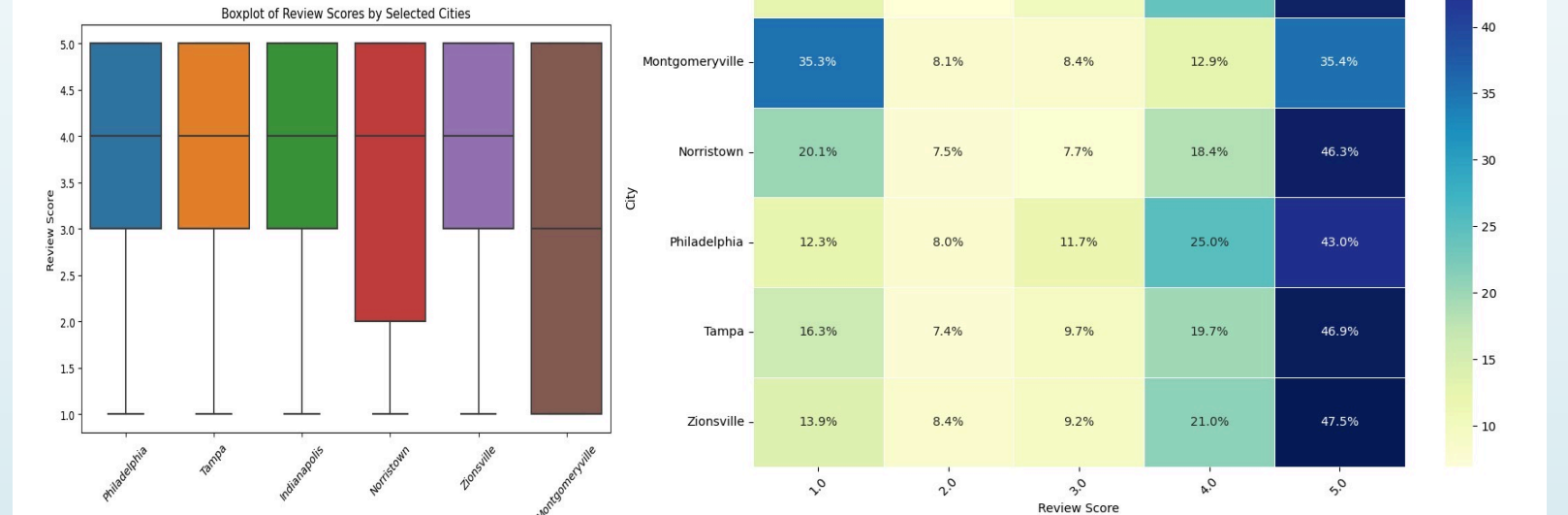

Four States Comparison

Top Cities

Sentiment Analysis

- Positive Reviews

- Negative Reviews

- Neutral Reviews

Future Work

- Sentiment Analysis can be performed for each state over the years.
- Customers can be categorized based on review sentiment.
- Outlier detection can be performed to identify customers with most extreme reviews.