**STAT 628 Module 3 Summary**

*Srivats Kumar Tharanilath, Yanrun Lu, Yuchen Dou*

**Introduction:**

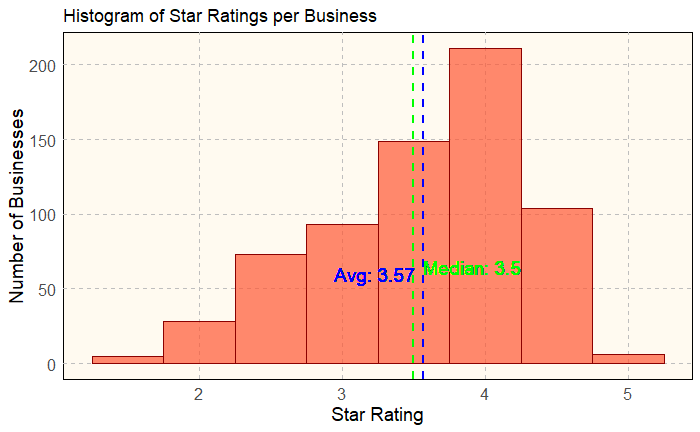
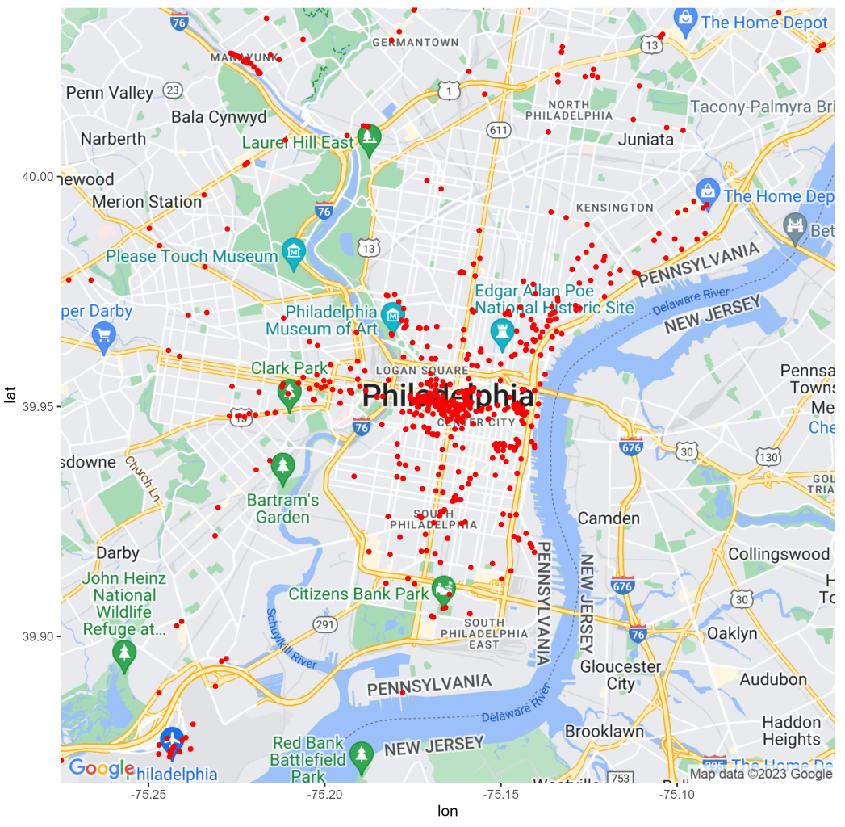
This study focuses on understanding the factors that contribute to American restaurants’ success or failure in Philadelphia. Our analysis adopts a two-pronged approach. Firstly, we conduct hypothesis testing to discern which attributes significantly impact restaurant ratings. Secondly, we delve into sentiment analysis to parse through reviews. Then, in order to make our findings accessible and interactive, we have developed a web-based application to show the data.

**Background & Data Cleaning:**

We believe that the best manner by which we can assess a businesses success or failure is by evaluating the review stars it has. So, we need to process the raw data on Yelp's platform. The initial dataset of 244 million reviews was narrowed down to include only open restaurants in Philadelphia with over 10 reviews. This data was merged with the 'business.json' file and US Census Bureau data, correlating restaurant information with respective zip codes. Data cleaning involved removing blank data and closed businesses. The dataset was then categorized into 'cheap' and 'expensive' restaurants based on price range, and zip codes were classified into 'high' and 'low' income neighborhoods for predictive analysis. The final dataset contained 168,284 observations across 112 variables, including 53 columns covering 669 restaurants (24 high-priced, 176 low-priced, 469 unknown-priced). The review data consisted of 7,847 reviews for high-priced restaurants, 43,422 for low-priced, and 117,015 for unknown-priced restaurants.

**Exploratory Data Analysis:**

We also do a brief EDA to have an overview of the data. Firstly we made the histogram of the star ratings per business of the whole dataset( Plot 1). We can see that the number of 4 stars is the highest, the average rating of the whole data is 3.57, and the median is 3.5.

**Plot 1 Plot 2**

Plot 2 shows the map of restaurants in Philadelphia. We can find that most of them are located in the city center, and some of them are concentrated in the traffic arteries and hubs. In order to have a brief knowledge on what features may play an important role in the comments, these eight words that appeared most frequently in the comments were selected after deleting some uncorrelated words like ‘and’, so in the following NLP part we may focus on these words.

|  | service | chicken | brunch | atmosphere | sandwich | salad | steak | burger |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| frequency | 47982 | 19229 | 15302 | 14986 | 14958 | 12716 | 12564 | 12114 |

**Key Findings About American Restaurants in Philadelphia:**

Hypothesis Testing:

We decided to evaluate each of the attributes we got from ‘business.json’ and census data. We used the Wilcox test to find which of the grouped attributes were significant, for this method is suitable for large samples and there is no requirement for the normality. The attributes we found that really matters are as follows:

| **Cheap Restaurants** | **W-value** | **p-value** | **Expensive Restaurants** | **W-value** | **p-value** |
| --- | --- | --- | --- | --- | --- |
| HappyHour | 7104087 | 6.865e-08 | Dessert | 161221 | 2.2e-16 |
| Drive Thru | 1485917 | 5.073e-07 | Upscale | 245730 | 2.2e-16 |
| Divey | 9630394 | 1.063e-07 | Brunch | 206364 | 2.2e-16 |
| Trendy | 3370217 | 6.167e-08 | Attire | 229656 | 3.775e-06 |

We can clearly see that there are different attributes that are more important for expensive restaurants compared to cheap ones. When it comes to cheap restaurants what seems to matter more to customers, ease of obtaining food. They are less focused on the price of the food itself. We can see this with the Drive Thru and Happy Hour attributes. However this doesn’t mean that ambience is entirely excluded, as we do see in general that divey restaurants do very poorly. There seems to be a certain degree of comeliness required by the customers regardless of prices offered. Trendy cheap restaurants seem to be the only kind of ambience that brings about a more significant change in the rating. However when it comes to expensive restaurants the general ambience and food quality are much more important. We observed that those restaurants that serve brunch and dessert do much better than those in a similar price range that do not. The upscale ambience and formal attire required clearly indicate that when it comes to more expensive restaurants.

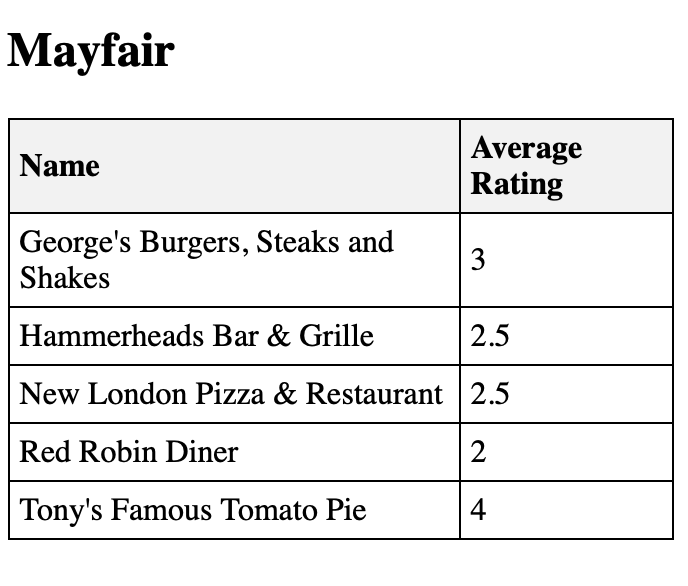
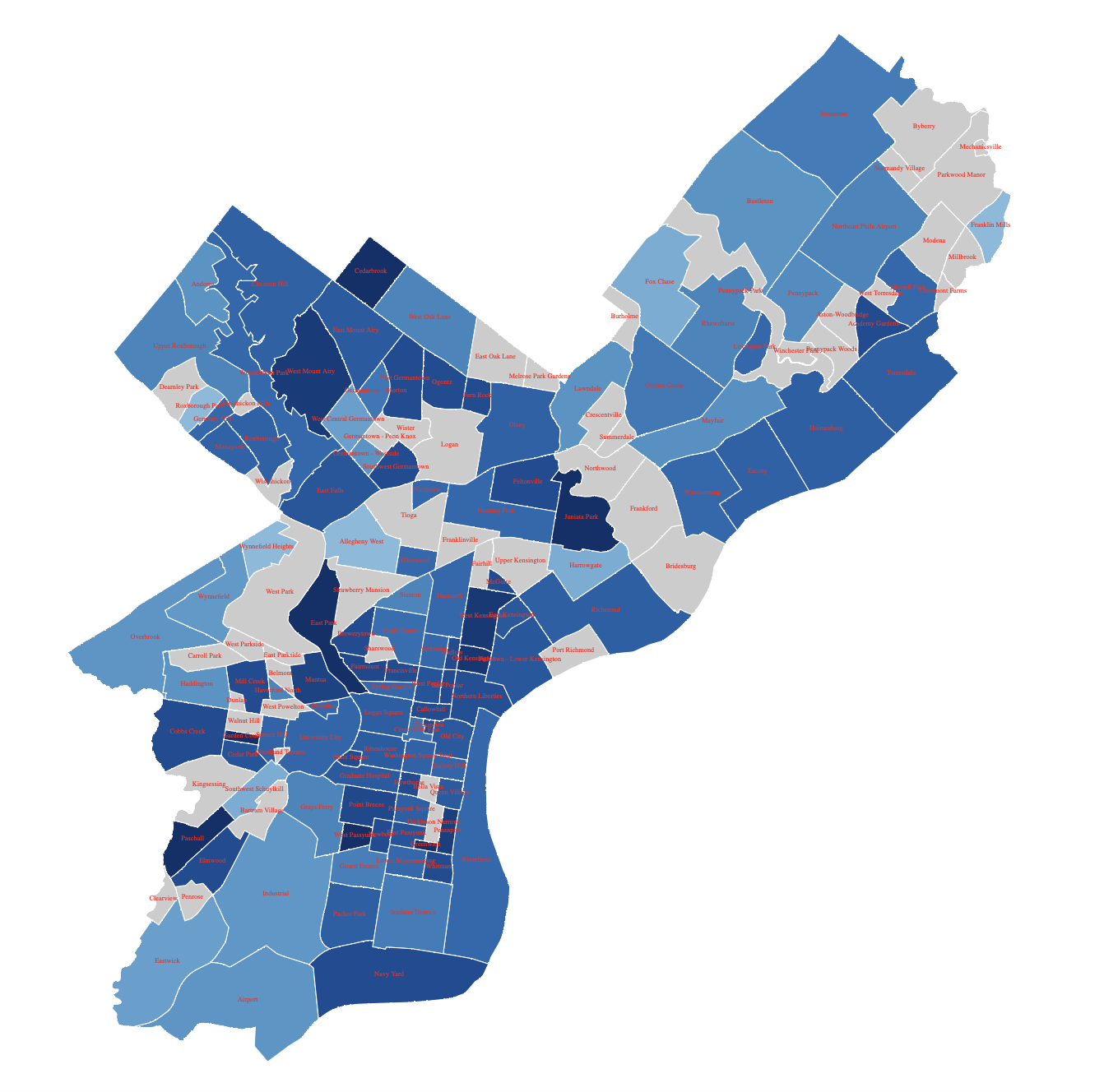
Sentiment Analysis:

Based on the results in the EDA part, we choose the 8 words as the main research object for the sentiment analysis. We used SentimentIntensityAnalyzer which is a class from the nltk.sentiment.vader module in the Natural Language Toolkit (NLTK) library in Python to calculate the average sentiment score (ranges from -1 to 1) of the 8 words separately in the reviews , and then find out whether there is strong correlation between the average score and the average review\_stars. At the same time, we compared the average sentiment score of four types of restaurants, namely high-priced high-rating restaurants and high-priced low-rating restaurants, etc. From the table below we can see that sentiment scores of all the 8 words have correlation effects with the review\_stars and in which service has the strongest correlation. It is worth noting that lower-priced restaurants tend to receive higher ratings for service, while higher-priced restaurants seem to have more demanding and sensitive customers for the average score of both low-priced restaurants are higher than the expensive ones. From our research, we also find that for low rating reviews containing service, 44.28% of the reviews mentioned ‘time’ which may suggest that unreasonable serving time may be a major deduction. When it comes to the food, we can see that chicken and salad have higher average scores than the others, and seem to get a better comment in high rating reviews. Burger also does well in high price business, while steak and sandwich have better performance in low price ones. Brunch is another key factor to increase the rating from the table, it also has a high score in high rating reviews and as is mentioned above the average rating of restaurants serving brunch is 0.3 higher than the ones not serving. Last but not least, the atmosphere similarly has a quite high score in high rating reviews which means if you want to get a high rating, paying attention to the restaurant environment is an effective choice.

|  | Average of Sentiment score | Correlation coefficient | Variance of Sentiment Score | High Rating Reviews | Low Rating Reviews | High price & High Rating | High price & Low Rating | Low price & High Rating | Low price & Low Rating |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| service | 0.701 | 0.621 | 0.295 | 0.908 | 0.347 | 0.900 | 0.322 | 0.911 | 0.358 |
| chicken | 0.764 | 0.562 | 0.212 | 0.907 | 0.443 | 0.924 | 0.543 | 0.902 | 0.438 |
| brunch | 0.832 | 0.466 | 0.129 | 0.911 | 0.559 | 0.928 | 0.537 | 0.905 | 0.574 |
| atmosphere | 0.860 | 0.419 | 0.097 | 0.920 | 0.652 | 0.911 | 0.621 | 0.918 | 0.681 |
| sandwich | 0.739 | 0.500 | 0.212 | 0.874 | 0.456 | 0.890 | 0.432 | 0.906 | 0.515 |
| salad | 0.794 | 0.537 | 0.192 | 0.926 | 0.506 | 0.933 | 0.610 | 0.927 | 0.511 |
| steak | 0.709 | 0.501 | 0.227 | 0.854 | 0.421 | 0.827 | 0.402 | 0.923 | 0.481 |
| burger | 0.757 | 0.518 | 0.206 | 0.897 | 0.492 | 0.923 | 0.475 | 0.898 | 0.546 |

Web Based App:

Based on the analysis we made before, we implement a web based interactive map to help the user to understand the dataset better. Our visualization is based on d3.js. In the initial stage, a map of Philadelphia bounded by neighborhoods is shown. Regions are colored by the level of average rating of restaurants in the region.



We can notice that High-rated areas like East Park, West Mount Airy, and Cedar Park contrast with lower-rated ones like Airport, Industrial, and Mayfair. Key attributes affecting ratings include noise level, restaurant delivery, drive-through, and outdoor seating. Detailed data, including boxplots of these attributes, is available for each region. For instance, in Mayfair, low-rated restaurants often lack delivery or outdoor seating, whereas in high-rated areas like Mantua, these features are more common. Noise level is also crucial, with quieter restaurants in high-rated regions and louder ones in lower-rated areas. Clicking on a restaurant in the provided data frame reveals more details such as location and number of reviews.

Limitations & Suggestions:

Our research actually has some limitations. Firstly, the study primarily relies on data from a specific dataset (business.json and census data), which may not comprehensively represent all American restaurants in Philadelphia. Secondly, the study identifies correlations between various factors (like ambiance, food type, and service) and restaurant ratings. However, it does not establish causation, meaning it is unclear if these factors directly cause changes in ratings or if other, unexamined variables play a role. Last but not least, we didn’t take time dimension into consideration, this temporal limitation restricts the study's long-term applicability, as restaurant trends and customer expectations can change rapidly. In the future study, we could include a broader range of data sources, such as different online review platforms or direct customer surveys, to mitigate selection bias and provide a more holistic view of the restaurant industry.At the same time, conducting longitudinal research would allow for tracking changes in consumer preferences and restaurant performance over time, providing a more dynamic understanding of the industry.

**Conclusion:**

From our study, iIt reveals that while convenience factors are key in cheaper restaurants, upscale restaurants need to focus more on ambiance and food quality. The significant role of service across all restaurant types cannot be overstated. Moreover, the sentiment analysis highlights specific food items and aspects like brunch and atmosphere that resonate well with customers. The regional analysis through the web-based app offers a strategic tool for future restaurant owners, suggesting that choosing the right location and aligning with local customer preferences can significantly influence a restaurant's success.

**Contributions and References:**

| Contributions | Srivats Kumar Tharanilath | Yanrun Lu | Yuchen Dou |
| --- | --- | --- | --- |
| Presentation 1 | Responsible for slides 1-4 (introduction and data cleaning). | Responsible for slides 5-9  Reviewed and edited slides 1-4. | Reviewed and edited and provided feedback on all slides. |
| Presentation 2 | Responsible for slides 3-5 | Responsible for slides 5-8  Reviewed/edited and provided feedback on all slides. | Responsible for slides 2,9,10 |
| Summary | Responsible for introduction, data cleaning, and hypothesis testing  Reviewed & edited data analysis section. | Responsible for Background & Data Cleaning, Sentiment Analysis, Limitations & Suggestions and conclusion  Reviewed/edited and provided feedback on the whole document. | Responsible for Web Based Apps.  Reviewed & edited the introduction, Limitations & Suggestions and conclusion |
| Code | Responsible for data cleaning code and the code for the hypothesis testing section | Responsible for Sentiment Analysis  Reviewed & edited data cleaning code. | Responsible for Web Based Apps. |
| Shiny App | Provided feedback on Web Based Apps | Provided feedback on Web Based Apps | Responsible for Web Based Apps. |