Module 3: Executive Summary

Introduction

The focus of this report is to explore the preprocessing, exploratory data analysis, modeling, and statistical tests that were performed on a slew of datasets including but not limited to Yelp business and review datasets, Google Maps data, GIS data, etc. Our aim in the project was to **(1)** Come up with actionable insights for Italian Restaurant business owners in the state of Pennsylvania that can help them increase their success on Yelp and overtime lead to success in their business overall, **(2)** Look for potential locations for opening a new restaurant. In our project, we used Python to create a comprehensive, composite 'success metric' for judging restaurant success instead of relying only on direct stars/reviews Along with that, we employ a combination of both statistical tests and machine learning models to solidify our findings.

Data Employed (in addition to what was necessary):

1. GIS dataset – Downloaded from Overpass Turbo and implemented using ArcGIS Pro software and GeoPandas Library.
2. Google Maps dataset – To enrich our Yelp dataset in places where there was a lack of reviews and business information. (per Google data usage policy)
3. Google Trend dataset
4. Demographic data (all 5-digit ZIP Code Tabulation Areas fully/partially within Pennsylvania) –
    a. DP03: Selected Economic Characteristics
    b. DP04: Selected Housing Characteristics
    c. DP05: American Community Survey Demographic and Housing Estimates
5. Shape Files – To help in visualizing administrative and ZIP Code boundaries in PA.

What is 'success' in our project?

Success for our project is quantified through a composite **'success score'** that evaluates Italian restaurants on multiple dimensions beyond basic metrics like star ratings and review counts. This holistic approach integrates:

- **Weighted Ratings**: Combines star ratings with the volume of reviews, adjusted by the highest review count to give prominence to widely reviewed establishments.
- **Time Decay**: Incorporates a temporal aspect by applying a decay factor to older reviews, reflecting the idea that more recent reviews are more indicative of the current state of the restaurant.
- **Sentiment Analysis**: Goes deeper into the text of reviews to extract sentiment polarity, categorizing them into positive, neutral, or negative sentiment scores and normalizing these scores for comparison.
- **Normalized Success Metrics**: Combines the normalized weighted rating and normalized sentiment score with respective weights, acknowledging that both popularity and customer sentiment play crucial roles in a restaurant's success.
- **Final Success Score**: The composite score is normalized out of 5 to provide a consistent, comparative measure of success.

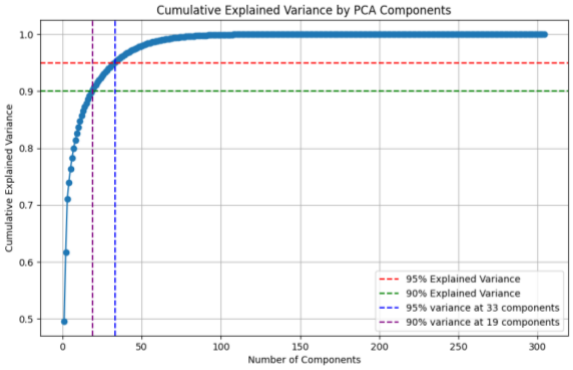
Data Preprocessing, Exploratory Data Analysis, and Involvement of Machine Learning:

Data Preprocessing -

- We started by weeding out insignificant columns in the reviews and business data using firstly the basic removal steps (handling null values, outliers, constant values, etc.) and then a combination of correlation analysis (with the target being the count of reviews and stars), **ANOVA**, and domain research. In the business data, all columns were standardized (stars, number of reviews), lowercased (address, city, categories), and duplicate business IDs were dropped. In the review data, techniques like tokenization, basic text processing (stop words, de-emojize, punctuation removal, lemmatization, and stemming), and negation handling.
- The mobility pattern dataset was cleaned with a strategy based more on domain knowledge and facts (e.g., considering the size of PA and inferring that trips of more than 500 miles are most likely out-of-state travel, etc.), than anything else. Redundant columns like Row ID, Level, State Postal Code, etc. were dropped. Along with that, the day of the week was extracted from the 'date' to later visualize daily trends.
- The census datasets were dealt with, which, when downloaded as a CSV, had numerous indentation issues and non-numeric characteristics, among other trivial complications that were handled in a single script. Upon numerous iterations of cleaning, it was realized that some issues had to be cleaned manually. For instance, there

is a hierarchical structure in all demographic datasets that did not follow a definite pattern and hence there was a fair bit of manual cleaning done.

- **PCA -**
  - In total, our demographic Yelp, and GIS datasets had 300+ features which made dimensionality reduction a necessity. To move forward with this, we performed **PCA** on our demographic dataset to try and reduce the dimensionality. The results from PCA were then juxtaposed with the results from our modeling stage along with domain research and **case studies**, to come up with the final set of features that played the most influential role concerning an Italian restaurant's success.



Exploratory Data Analysis -

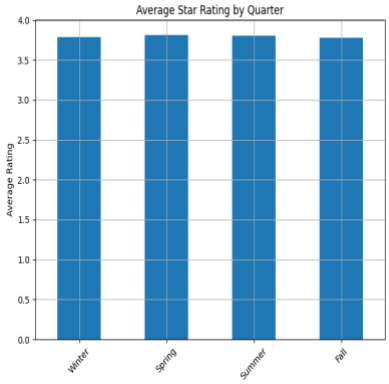
- **Correlation analysis and Factor analysis –**
  - These were the very first steps in EDA we performed to get a sense of what recurring features (or indirect features) we should focus on. Both these analyses were done against the success metric we defined earlier. The results from these, although not imposing, did give us a good idea of what kind of features we should be looking into, specifically, 'Employment Status', 'Moving-in patterns of the population over time', 'Construction of new units over time' and 'Income-related factors.'

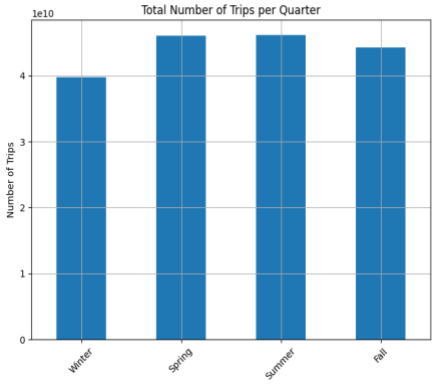
- **Sentiment Analysis –**
  - The sentiment analysis offered us great insights to build on for Italian restaurant owners:
    - **Aspect-Based Sentiment Analysis:** High sentiment scores across aspects like food, service, and ambiance suggest areas of strength. Owners should maintain quality here while also reviewing lower scores to identify and address specific concerns. (ABSA viz)
    - **Bi-gram/Tri-gram Analysis:** Common positive phrases like "go back" and "highly recommend" indicate customer satisfaction and loyalty. Owners should capitalize on this by encouraging customers to share their experiences and by targeting marketing efforts to highlight these strengths. (bi-gram, tri-gram)
    - **Sarcasm Detection:** With sarcasm present in over a quarter of reviews, owners should scrutinize these for hidden negative feedback, ensuring they don't overlook critical insights masked by sarcasm. (sarcasm detection viz)
    - **Sentiment Distribution:** A skew towards positive reviews underscores a generally favorable customer perception. Owners should continue to foster the elements that contribute to this positivity. (sentiment distribution viz)
    - **Word Embeddings Visualization:** Semantic relationships displayed in the embeddings can help owners understand which attributes customers frequently discuss together in a positive light, such as "excellent" and "friendly." (interactive viz for word embeddings)
  - Owners should continue to enhance what customers love, address negative sentiments, leverage loyal customer advocacy, and stay vigilant to the subtleties of customer feedback, such as sarcasm.

- **Geospatial Analysis –**
  - The focus of this was to have a look at the distribution of restaurants throughout the state, create choropleth maps (using demographic features and such) with overlapping restaurant distributions, and try to find relations. There was a need to reverse geocode to populate the column with 'county name', in

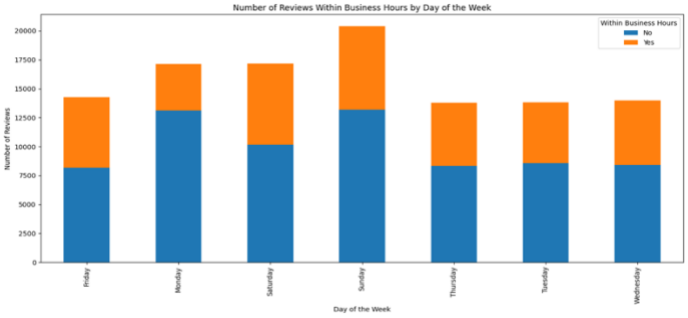
more than one dataset, which was done using zip codes. Detailed visualizations for this have been provided in our application/dashboard. (interactive viz for choropleth to inspect restaurant distribution)

- **Trend Analysis using Mobility Patterns and Reviews Dataset –**
  - We saw that overall, the different mobility patterns had either **insignificant** or **no effect** on the average ratings of Italian reviews, over time. The trend was constant when observed both quarterly and monthly. This was surprising since there should have been some seasonal effects based on the time of the year being a holiday season, or certain events in town.
  - Upon further inspection, it was clear that since there is no seasonal trend, Italian Restaurants, at least in PA, must maintain a year-round scheme (like loyalty programs) and keep up their pace throughout the year with there being no crests or troughs in customer engagement. There was an obvious decrease in the number of trips during the wintertime which can be accounted for due to the inclement weather seen on average during the winters in PA.





- **Time series analysis –**
  - **Focus on Business Hours:** The number of reviews within business hours is consistently higher than those outside of business hours. This suggests that most customers visit during these hours. Therefore, ensuring excellent service during business hours can lead to more positive reviews.
  - **Prepare for Sundays:** The highest number of reviews is on Sundays. This indicates that Sundays are likely the busiest day of the week (which checks out). Restaurant owners could consider staffing up on Sundays to handle the increased customer volume and maintain a high level of service.
  - **Improve Monday Experience:** The lowest number of reviews during business hours is on Mondays. This could be an opportunity to attract more customers on this day. Restaurant owners could consider special promotions or events on Mondays to increase customer visits and reviews.



Machine Learning -

- We performed **random forest regression**, **feature elimination** using the drop-column technique, **linear regression**, and **RFE** (Recursive Feature Elimination) to support our findings from PCA and have a comprehensive reasoning behind choosing the features that would eventually be converted into a composite metric representing a 'success index' for Italian restaurants in PA.
- From our PCA, we observed that 19 features out of all, explained 90% of the variance. Moreover, with all the above-mentioned ML techniques, statistical tests, and PCA, we found out that the top 35 features had

commonalities that were logically sound and allowed us to further reinforce our results from PCA. We decided to choose 20 of them based on domain knowledge and taking into consideration computational efficiency and ease of visualization.

Conclusion/Discussion:

(1) The actionable suggestions that we recommend to an Italian restaurant business owner are as follows –

1. **Updating Customer Feedback Strategy –**
   a. We noticed that customers are mostly categorized as follows –
      i. The reviews are posted too late vis-à-vis the time of having the meal hence hampering the authenticity of the reviews and introducing discrepancy.
      ii. The reviews are never posted which introduces **non-selection** bias since the reviews which are eventually analyzed, are not representative of the whole customer base.
   b. We recommend introducing a scheme that motivates a table to give a Yelp review before paying the bill to be eligible for a discount on their next visit. This not only generates around 6,000+ reviews a year but can also lead to customer retention. This can be further enhanced by asking for an additional, well-tailored survey to drive specific decision-making.

2. **Regular Trend Analysis –**
   a. Our EDA, with its many insightful trends and sometimes no trend (like with mobility data) nevertheless imparted important information that was then converted to actionable insights or at least assisted in drawing further analysis plans.
   b. We recommend investing in technology such as
      i. Association mining to understand customer behavior and update the menu for increased local engagement.
      ii. Review analysis to get near-real-time suggestions about the improvements customers are looking for, are some of those areas to invest in.
      iii. Competition analysis to understand if there has been any change in the success score of nearby restaurants and understand the reasons behind that to avoid pitfalls or get inspiration.

3. **Strengthen Online Presence –**
   a. The case study we referred to showed that over 94% of U.S. foodies refer to restaurant websites and reviews to decide where to eat. This, coupled with a strong positive coefficient associated with the presence of a website on Yelp with the star rating of a restaurant, as seen in lasso regression and random forest regression, puts the business on a steady track to increase its success and eventually, dollar profit.
   b. We recommend that the owners should invest in a basic website that showcases their culture and history. If not a website, working on social media presence, and engaging with other local businesses and communities, can have a similar effect.

(2) In the culmination of our extensive analysis, we devised a **composite metric** that robustly encapsulates a multifaceted view of restaurant success. By meticulously merging critical datasets related to demographic, economic, and consumer data and cross-referencing against a set of highly influential features identified through our stack of preprocessing, ML, and rigorous EDA, we ensured that each variable contributed significantly to the overarching narrative of restaurant viability and popularity. This metric, normalized across pertinent variables and weighted equally to tackle bias, provides a comprehensive gauge by which we measure the potential for success within various ZIP codes. Rescaled to a 0-100 range for practical interpretability, this composite score is not merely a derivative of singular indicators like foot traffic or average income but a sophisticated combination of factors that offer a likely glance at the success of Italian restaurants.

Contributions:

Contribution	Jinchen Gong	Shourya Maheshwari	Xiaoyang Dong
Presentation	Introduction and till slide 4	Slide 6 till the end (12)	Slides 5 and 6
Executive Summary (section-wise)	-	Complete Summary	-
Model Scripting	-	Data cleaning, preprocessing, EDA, and modeling	Significantly assisted in modeling and preprocessing
Application	Completed Majority of the app/dashboard	-	Significantly assisted in app/dashboard building

References:

[1] Brian. (2022, October 10). *Yelp Statistics, Demographics, Users, And Facts For 2021 | SaaS Scout (formerly SoftwareFindr)*. SaaS Scout (Formerly SoftwareFindr). https://saasscout.com/statistics/yelp-statistics/

[2] *Restaurant Reputation Management – Building a Positive Online Presence*. (2023, June 26). https://restaurant.eatapp.co/blog/restaurant-reputation-management

[3] Rivera, R. (2023, May 9). *Restaurant marketing: Building a strong local presence*. Marketing 360® Blog. https://blog.marketing360.com/case-studies/restaurant-marketing-case-study-building-a-strong-local-presence/

[4] Shanbhag, A. (2021, December 14). Association Rule Mining - Analytics Vidhya - Medium. *Medium*. https://medium.com/analytics-vidhya/association-rule-mining-7f06401f0601

[5] *Unselecting features that are selected—ArcMap | Documentation*. (n.d.). https://desktop.arcgis.com/en/arcmap/latest/extensions/production-mapping/unselecting-features-that-are-selected.htm

[6] Yelp for Business. (2023, October 19). *Study shows that high-intent consumers are contacting businesses on Yelp*. https://business.yelp.com/resources/study-shows-high-intent-consumers-are-contacting-businesses-quickly-on-yelp/