MARKETING ANALYTICS

Research on Yelp Data

Abstract

Research and Analysis of Yelp datasets to find meaningful insights based on factors such as user reviews, income and education, unemployment, and food service.

Group 1:

Dean Papadopoulos
Kripanjali Dhungana
Norman Morris
Zheding Zhao

Executive Summary

This project proposes to perform a marketing analysis on Yelp datasets and census datasets and attempt to answer hypothetical questions based on the combination of these chosen datasets. There will be data exploration, data analyzation and presentation of the findings. There is an introduction section that talks about the datasets followed by the project details where we talk about the main proposal, technical specifications, use of the analysis, tasks to be performed and project timeline.

Introduction

Food is the way to every human's heart and trying different cuisines is one of our biggest adventures where some of us realize this but some of us do not. People prefer to eat outside at least 2 to 3 times a week, however, a majority of us do not know what kind of cuisine we want to try, or which restaurant would be willing to try. Should we stick to our go to restaurant again or try something new? Once we decide on a specific cuisine and a specific restaurant, we want to check out the reviews to see what other people think about the place. These are the people who have already tried the place and the food. Everyone is different and thus seek different things when trying out a place. Quality food or taste is one thing that is in most cases non-negotiable for everyone, but besides that some people want ambience whereas others want fast service. Some people want to go to a place where they have live music whereas some people want to sit and have conversation while they dine. It has become a trend these days for people to sit in cafes and get their work done while sipping on their beverage or grabbing a bite.

It is all about preferences and for some it is quality service. From a user's perspective these are the obvious questions that comes to mind when choosing a restaurant. However, from a business perspective what are the concerns that may arise? Or just as a fun project what would we want to know about who is writing these reviews or see if there is a certain region where people tend to eat out more. How about where the demographic of more health-conscious population resides over not so health conscious, or how can Yelp improve their page given tough competitors such as TripAdvisor or Foursquare. Hence, to get to the core of questions as such we are conducting a marketing analysis specifically on the restaurant reviews on Yelp.

Project description

Main proposal:

The main goal of this marketing analysis project we are conducting is to attempt to answer some questions based on the datasets we chose. We have Yelp's datasets lined up which we will combine with the population and demographic information from the census data. We also plan to use the unemployment rate and inflation rate on our hypothesis to check if a correlation exists with the check-ins or the reviews. We plan to do an initial exploratory analysis to learn about the datasets and how we can formulate questions we could later get answers to. We brainstormed some questions that could be interesting to look at, below are those questions that we plan to answer after conducting the analysis. Please note the last question is a stretch goal that we may attempt to answer using a different model.

- What are some keywords for good reviews and bad reviews?
- Is there a correlation between review score and length of review text?

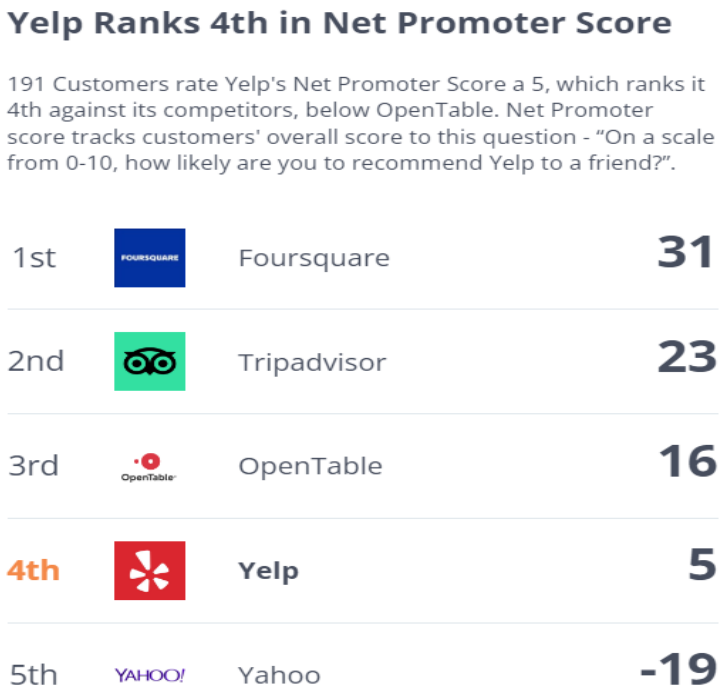
- What is the relationship between state income and ratings on expensive restaurants?
- What is the relationship between state income and number of restaurants?
- Is there a correlation between the number of categories and average score?
- Are people who leave more reviews likely to be more critical than people who leave less reviews?
- Does higher inflation result in less check-ins on Yelp?
- Does lower unemployment result in more reviews on Yelp?
- Can ratings be predicted by business attributes?
- Can clustering be used to group or find patterns in businesses or reviews?

Technical specifications:

Throughout the project, we plan on using Kafka messaging, Data Factory Pipeline, Data Lake Usage, create a Machine Learning Model and create a dashboard using Power BI that will incorporate the Machine Learning predictions. We will also be creating an Entity Relationship Diagram (ERD), Data Definition Language (DDL) for SQL server and Data Flow Diagram for data exploration purposes.

Use of current research:

The current research/analysis we are performing can be used to better understand Yelp users. The hypothetical questions we are attempting to answer above under the main proposal can be used by Yelp to not only understand the demography of their users but also consider other factors such as unemployment, income and inflation has. On a survey conducted by Comparably, we can see below that Yelp scores a 4th place on a question “how likely are you to recommend Yelp to a friend?” This analysis can help Yelp further investigate and look into the reasons why its competitors are ahead of the game.



Tasks to be performed:

For this project we are using Agile methodology where we have planned sprints for two weeks. We have divided the whole project into three sections, data exploration, ETL and analysis and presentation.

- For the data exploration part:
    - Start off by finalizing on the datasets
    - Power BI environment setup
    - Initial Exploratory Questions
    - Entity Relationship Diagram
    - Kafka Messaging
    - DDL for SQL Server
    - Data Flow Diagram
- For Extraction, Transformation and Loading (ETL), we setup:
    - Rough ETL report
    - Business ETL
    - Unemployment ETL
    - State ETL
    - User ETL
    - Census Economic ETL
    - Census Income ETL
    - Review ETL
    - Final ETL Report
- Analysis and Presentation:
    - Create Machine Learning Model
    - Enhance Machine Learning Model

- Draft Technical Report
- Data Factory
- Dashboard Drawings
- Visualization Drawings
- Drawings Feedback
- Power BI Dashboard
- Review Report
- Final Report
- Create ReadMe in GitHub

Project Timeline:

The timeline for this project is two weeks. Starting from August 8, 2022, we plan on wrapping the project by August 19, 2022.