Intermediate project status report

Customer Segmentation And Analysis with Yelp Dataset

Sakshi M. Mukkirwar, Keerthana Raskatla, Swati, Akanksha Tyagi, and Anjali H. Ojha

Department of Applied Data Science, San Jose State University

DATA 228: Big Data Technologies and Applications

Dr. Vishnu Pendyala

October 27, 2023

TABLE OF CONTENTS

1. Introduction

In today's data-driven world, restaurants strive to acquire a competitive advantage by analyzing diner preferences and behavior. The Yelp restaurant reviews dataset, which contains millions of diner reviews, ratings, and restaurant profiles, is a significant resource for such customer insights. This intermediate status report highlights our progress towards using Big Data analytics to do consumer segmentation utilizing this rich Yelp dataset.

2. Problem Statement

Developing a solution for precise audience targeting to facilitate restaurants in gaining a deeper understanding of their clientele, thereby enhancing the overall impact on their business. There are many aspects of this data that can be leveraged to quantify the customer's behavior and expectations. This dataset provided by Yelp is real data coming from real businesses and reviews. Our approach for this will not just be restricted to analysis, we are going to explore this data keeping restaurants in mind for customer segmentation.

3. Dataset

We are using the Yelp Review dataset. This dataset not only contains the user review but other useful information like Customers, Business details, and different attributes related to that, Customer check-ins, and tip amounts. All the files are in JSON format. Data has 6,990,280 reviews 6 for 150,346 restaurants in 11 metropolitan areas given by 1,987,897 customers. It also contains information about the cuisine served in the restaurant, the ambiance, and other information. The data size is 11.8 GB in compressed (.tar) format.

Dataset Name	Record Counts
yelp_academic_dataset_business.json	businessDf.count() = 150346
yelp_academic_dataset_user.json	userDf.count() = 1987897
yelp_academic_dataset_checkin.json	checkinDf.count() = 131930
yelp_academic_dataset_review.json	reviewDf.count() = 6990280
yelp_academic_dataset_tip.json	tipDf.count() = 908915

4. Progress to date

4.1. Project Proposal: We have crafted a proposal that outlines our data selection process and provides a broad overview of what we want to implement and achieve through Big Data tools and technology.

4.2. Data collection and refinement: We have developed scripts that will facilitate the cleaning and formatting of data required for further analysis.

4.3. Data Cleaning: Most of the data was clean since its flexible schema missing fields were automatically handled. To make different fields consistent we have to do some data transformation.

4.4. Data Transformation: It's real-world data and it keeps changing. So capturing some of the information in structure is not ideal. Here is an example of some business attributes present in the data, we can see that earlier versions of the data have listed all the different attributes as columns in the schema. But this is not ideal as there can be a new attribute tomorrow which will need system-wide change to accommodate the new column. So, we need to change such fields into a generic type, which can handle new attributes. We can see the code example below.

```python
# Convert struct field 'attributes' to MapType
cols = businessDf.select("attributes.*").columns
businessDf.select("attributes").printSchema()


print("after transformation - ")
df_with_map = businessDf.selectExpr("map(" + ", ".join(["attributes."+col for col in cols[:-1]]) + ") as attribut
df_with_map.printSchema()

```

```
root
 |-- attributes: struct (nullable = true)
 |    |-- AcceptsInsurance: string (nullable = true)
 |    |-- AgesAllowed: string (nullable = true)
 |    |-- Alcohol: string (nullable = true)
 |    |-- Ambience: string (nullable = true)
 |    |-- BYOB: string (nullable = true)
 |    |-- BYOBCorkage: string (nullable = true)
 |    |-- BestNights: string (nullable = true)
 |    |-- BikeParking: string (nullable = true)
 |    |-- BusinessAcceptsBitcoin: string (nullable = true)
 |    |-- BusinessAcceptsCreditCards: string (nullable = true)
 |    |-- BusinessParking: string (nullable = true)
 |    |-- ByAppointmentOnly: string (nullable = true)
 |    |-- Caters: string (nullable = true)
 |    |-- CoatCheck: string (nullable = true)
 |    |-- Corkage: string (nullable = true)
 |    |-- DietaryRestrictions: string (nullable = true)
 |    |-- DogsAllowed: string (nullable = true)
 |    |-- DriveThru: string (nullable = true)
 |    |-- GoodForDancing: string (nullable = true)
 |    |-- GoodForKids: string (nullable = true)
 |    |-- GoodForMeal: string (nullable = true)
 |    |-- HairSpecializesIn: string (nullable = true)
 |    |-- HappyHour: string (nullable = true)
 |    |-- HasTV: string (nullable = true)
 |    |-- Music: string (nullable = true)
 |    |-- NoiseLevel: string (nullable = true)
 |    |-- Open24Hours: string (nullable = true)
 |    |-- OutdoorSeating: string (nullable = true)
 |    |-- RestaurantsAttire: string (nullable = true)
 |    |-- RestaurantsCounterService: string (nullable = true)
 |    |-- RestaurantsDelivery: string (nullable = true)
 |    |-- RestaurantsGoodForGroups: string (nullable = true)
 |    |-- RestaurantsPriceRange2: string (nullable = true)
 |    |-- RestaurantsReservations: string (nullable = true)
 |    |-- RestaurantsTableService: string (nullable = true)
 |    |-- RestaurantsTakeOut: string (nullable = true)
 |    |-- Smoking: string (nullable = true)
 |    |-- WheelchairAccessible: string (nullable = true)
 |    |-- WiFi: string (nullable = true)

after transformation -
root
 |-- attributes: map (nullable = false)
 |    |-- key: string
 |    |-- value: string (valueContainsNull = true)
```

There are many similar types of transformation performed for the other fields in different datasets. There are other transformations being done like the categories field, which will be converted to a list type of field, etc.

```python
from pyspark.sql.functions import from_json, col, map_keys
from pyspark.sql.functions import split

businessDf.select("categories").printSchema()

print("After shcema change - \n")
businessDf = businessDf.withColumn("categories", split(col("categories"), ", "))
businessDf.select("categories").printSchema()
```

```
root
 |-- categories: string (nullable = true)

After shcema change -

root
 |-- categories: array (nullable = true)
 |    |-- element: string (containsNull = false)
```

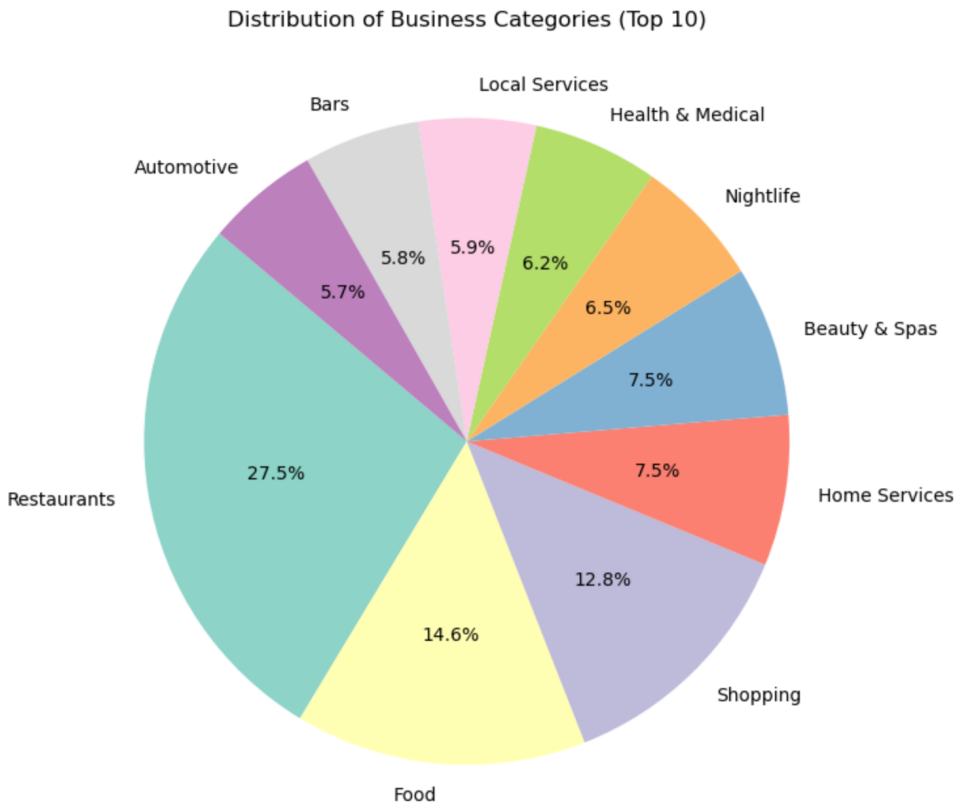
4.5. Exploratory Data Analysis: We've delved into the data to discover valuable attributes and comprehend how each of these attributes can be harnessed for effective targeting.

4.6. Spark Jobs creation: We are writing Spark Jobs to create a data pipeline that will generate aggregates, which later be saved into RDBS and used in Tableau.

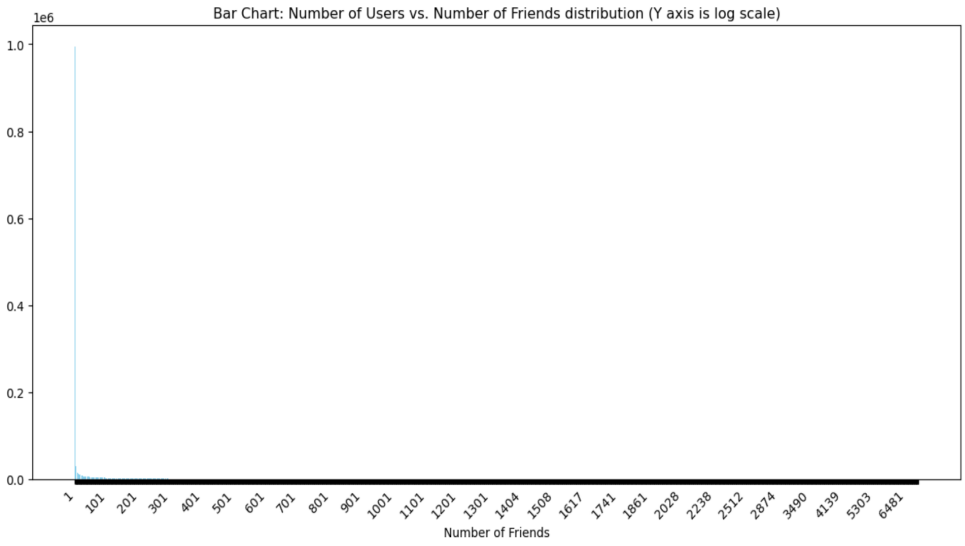
5. Results

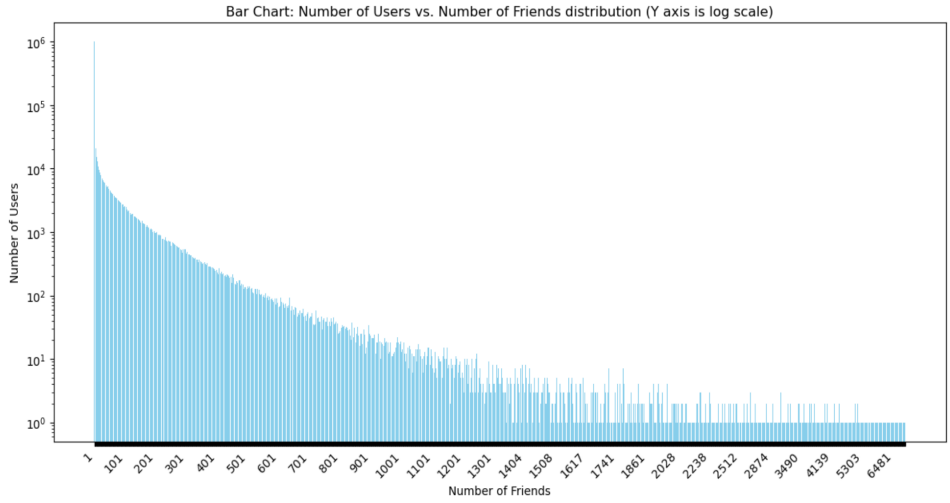
Once we cleaned and transformed the data, we did a quick data analysis to understand the data. Here are some findings -

- Top 10 categories businesses are categorized as. This shows that the data is not just for restaurants, it also has information from other Margaret segments.

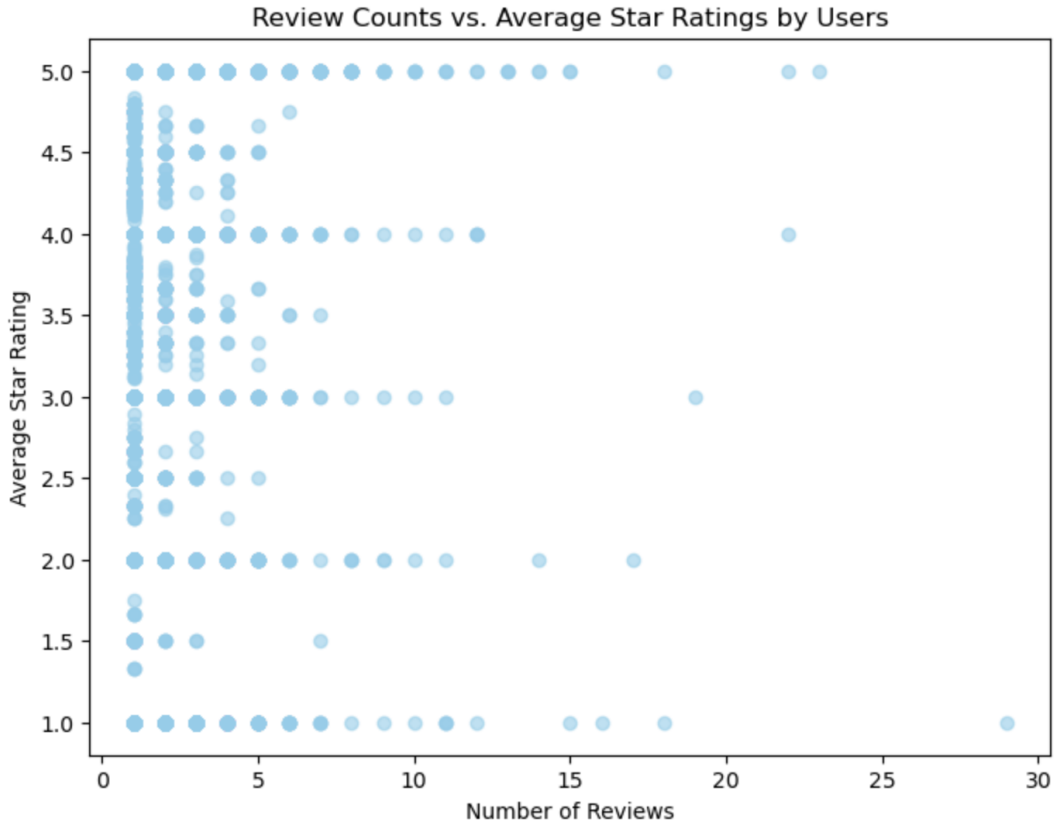


- As the Yelp data captures the individual's social network of friends, we tried to see what the average number of friends people have. Since the y-axis is logarithmic, we can see that most people have only 1 friend.

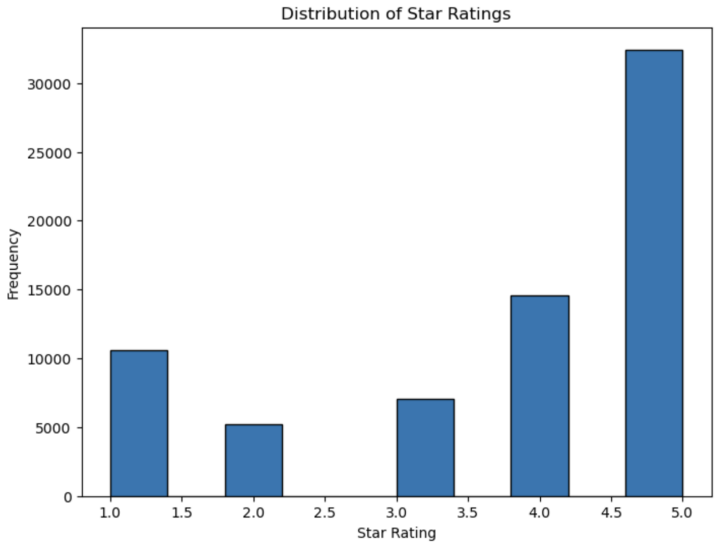




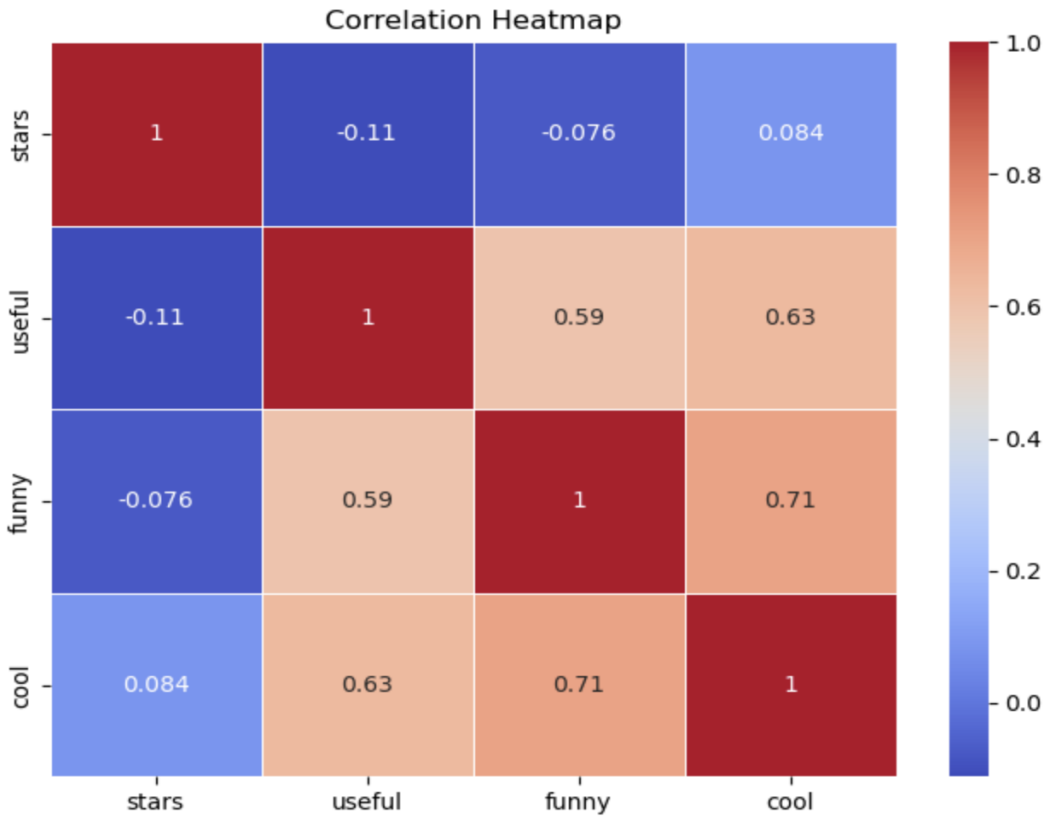
- Distribution of people rating. As we can see most users give very few reviews, even though there are a lot of users but they are not actively reviewing places. And we can also see that most of the reviews are above 3, which can suggest people don't give low ratings often.



And the same can be seen here also.



- Here is the analysis of how other tags related to customer's experience tell about how they feel.



6. In progress tasks

6.1. Cluster And Environment Setup: We have initiated the configuration of cloud infrastructure, which will serve as the hosting environment for our big data application and enable us to conduct advanced analyses.

6.2. Data Aggregating: We are writing spark jobs to aggregate the data and creating the data pipeline for the data to be ingested into the system.

6.3. Understanding the data for meaningful business insights and customer behavior. Which will be later used in visualization.

7. Difficulties encountered and solutions

The set we used for the project is customer and business-facing, and it relies on the people to provide data points. Here are the following difficulties we are facing.

1. Scale of Data - The size of the data is big enough to operate on the local computer. To get some meaningful insights, it took some time for the system to process. To tackle this problem we are using sampling of the dataset, once the code is working we are scaling it for the full data.
2. The sparsity of Data - Since the data is a sample shared by Yelp, it's very sparse. The sparsity of the data makes it difficult to capture prominent patterns and correlations. For example, lots of restaurants have missing attribute information. To solve this problem we are focusing on small segments of data using filters.

	number_of_friends	number_of_users
0	1	993654
1	2	57923
2	3	38827
3	4	30168
4	5	24389

3. Unstructured Data - The data is provided in JSON format, so it has some structure to work with, but information like categories and attributes needs some data transformation before we can use it. To handle this problem, we are writing custom transformations to unify the fields.

8. Planned work

8.1. Creating Tableau Visualizations - After processing and aggregating the data, our next step is to develop a Tableau dashboard that incorporates all relevant information.

8.2. Create Data Cube for Segmentation - As we are focusing on customer segmentation, the next step will be to bring all the users and business attributes together as a data cube. This data structure will help us to create the segment faster. With this approach, we are targeting faster query performance. Meek et al. (2021)[1] study the contextual and descriptive features of online restaurant reviews that influence users' perceptions of them as useful.

8.3. Performing end-to-end system testing: We will be performing end-to-end checks on the project to ensure the proper functioning of each component that is part of the system life cycle. This involves data loading, processing, storing, and visualization modules.

8.4. Deployment at Cloud Infrastructure - once we complete the development, we will deploy the whole system on the cloud where it can leverage large compute resources for enterprise-level systems.

8.5. Delivering the Final Report and Presentation - This involves the process of writing the final report and preparing the presentation, where we gather and organize valuable insights for effective customer segmentation.

9. Additional information

Completed other sections required for a technical report like introduction, problem statement, motivation, goals, and contributions of each member towards the successful delivery of the project.

10. References

[1] Meek, S., Wilk, V., & Lambert, C. (2021). A big data exploration of the informational and normative influences on the helpfulness of online restaurant reviews.