Smart Local Guide: A Recommender System Leveraging Google Reviews and Metadata

G9 - The Eagles

Data Science Capstone Project
Exploratory Data Analytics Report

Date:

03/09/2025

Team Members:

Rushikesh Gholap (rg954@drexel.edu)

Hassaan Alvi (haa76@drexel.edu)

Jelitta James (jj982@drexel.edu)

Priyanka Choudhary (pc845@drexel.edu)

Krish Balar (kb3842@drexel.edu)

The purpose of this report is to describe the exploratory data analytics. It includes five major sections:

- ## Google Local Reviews : Exploratory Data Analysis

1. Analysis the basic metrics of variables

In this section, we identify all the variables in the dataset and conduct the basic metrics of the variables. What are the data types (numerical/categorical, discrete or continuous, ordinal or nominal) and size? Provide the descriptive statistics of the variables such as mean, standard deviation, min, max, percentiles, etc.

The reviews dataset has the following variables user_id, name, time, rating, text, pics, resp and gmap_id. The metadata dataset has a total of 15 columns :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 190816 entries, 0 to 190815
Data columns (total 15 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   name              190816 non-null  object
 1   address           188090 non-null  object
 2   gmap_id           190816 non-null  object
 3   description       31045 non-null   object
 4   latitude          190816 non-null  float64
 5   longitude         190816 non-null  float64
 6   category          189922 non-null  object
 7   avg_rating        190816 non-null  float64
 8   num_of_reviews    190816 non-null  int64
 9   price             35935 non-null   object
 10  hours             144070 non-null  object
 11  MISC              154252 non-null  object
 12  state             133096 non-null  object
 13  relative_results  177663 non-null  object
 14  url               190816 non-null  object
dtypes: float64(3), int64(1), object(11)
memory usage: 21.8+ MB
```

There are 4 numerical variables Latitude, Longitude, Avg rating and num of reviews.

	latitude	longitude	avg_rating	num_of_reviews
count	190816.000000	190816.000000	190816.000000	190816.000000
mean	40.463714	-77.000711	4.311211	115.005042
std	0.542237	2.675941	0.633938	314.023287
min	28.647890	-122.151620	1.000000	1.000000
25%	40.048368	-79.148584	4.000000	8.000000
50%	40.320494	-76.275641	4.400000	28.000000
75%	40.688839	-75.328971	4.800000	95.000000
max	45.959974	180.000000	5.000000	9998.000000

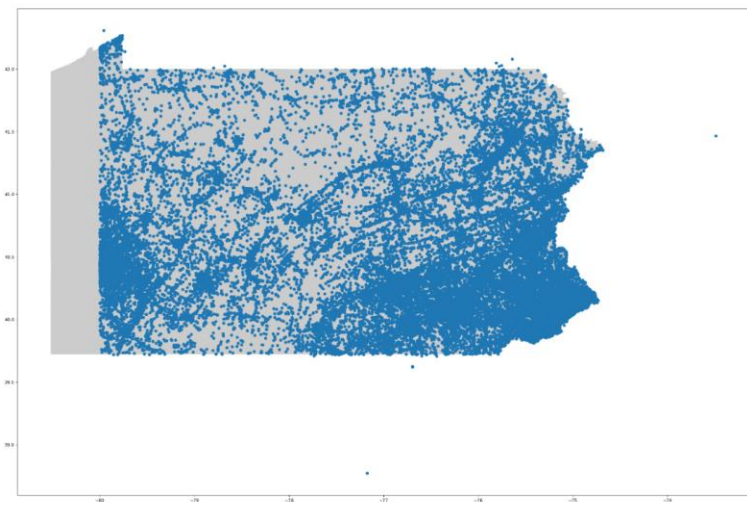
gmap_id is a continuous variable, Name, address, description are textual features.

MISC is a dictionary type column with multiple categorical features like Service options, Health & safety, Offerings, Amenities, Atmosphere, Crowd, Accessibility, Planning, Payments, Highlights, Popular for, Dining options, From the business, Recycling, Getting here, Activities and Health and safety.

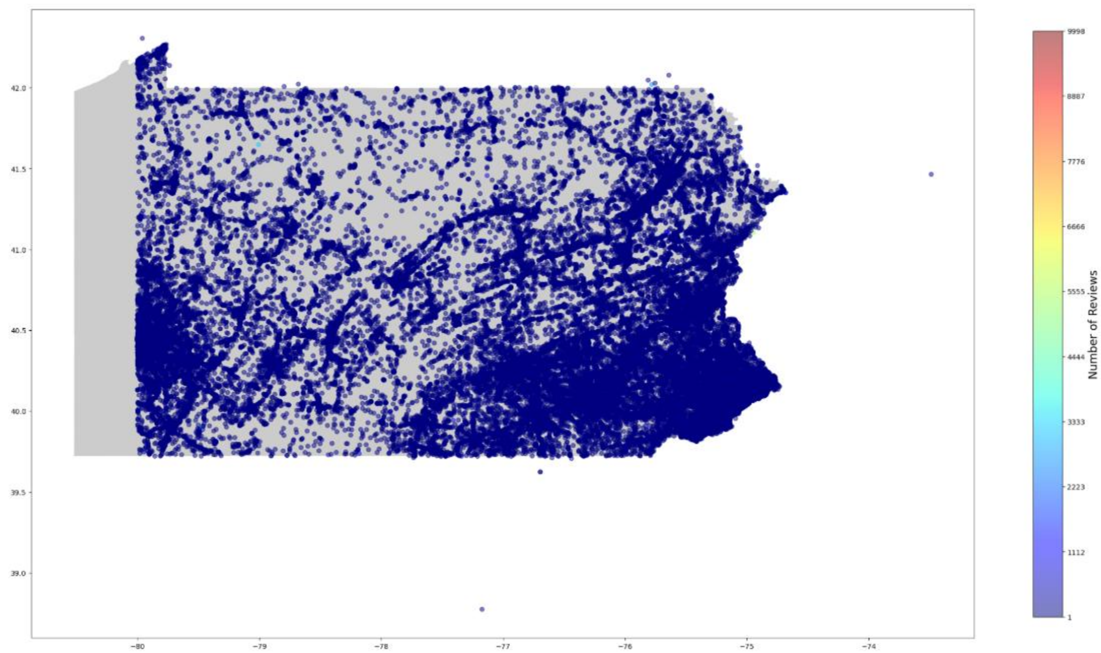
2. Non-graphical and graphical univariate analysis

In this section, we identify the list and number of unique values for each variable and provide the histogram and box plots to understand the distribution of the data.

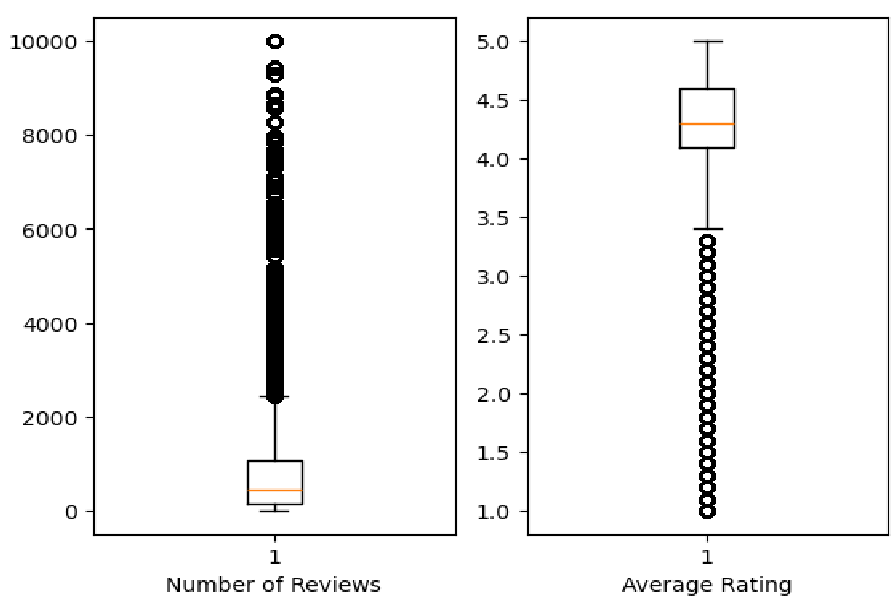
There is a total of 189836 unique gmap ids. The businesses are spread throughout PA state identified by latitude and longitude as shown below



Business with number of reviews



Few of the businesses have a lot of ratings as shown below with average rating relatively positive or above average than lower as shown below



```
5.0     23833
4.5     15390
4.6     15294
4.7     15156
4.8     14172
4.4     14099
4.3     13624
4.2     11226
4.9     10977
4.0     10001
```

```
2.1     207
1.8     155
1.9     146
1.7     128
1.5     110
1.6      93
1.4      70
1.3      64
1.2      23
1.1      12
```

name	num_of_reviews	
Citizens Bank Park	9998	1
Hershey's Chocolate World	9998	1
The Franklin Institute	9998	1
PNC Park	9998	1
Rivers Casino	9998	1

Category is a categorical variable with 3484 unique values. Each business is associated with multiple categories. There are a total of 3484 unique categories with certain categories distributed more than others as shown below

```
([('Restaurant', 11801),
  ('Auto repair shop', 6094),
  ('Fast food restaurant', 4935),
  ('Pizza restaurant', 4902),
  ('Bar', 4653),
  ('Grocery store', 4288),
  ('Takeout Restaurant', 3966),
  ('Gas station', 3812),
  ('Convenience store', 3718),
  ('American restaurant', 3602)],
 3484)
```

State has 1827 unique values.

```
(array(['Open ⋅ Closes 9PM', 'Open ⋅ Closes 7PM', None, ...,
        'Closed ⋅ Opens 3AM Wed', 'Closes soon ⋅ 11PM ⋅ Opens 12PM Wed',
        nan], dtype=object),
 1827)
```

Missing value analysis and outlier analysis

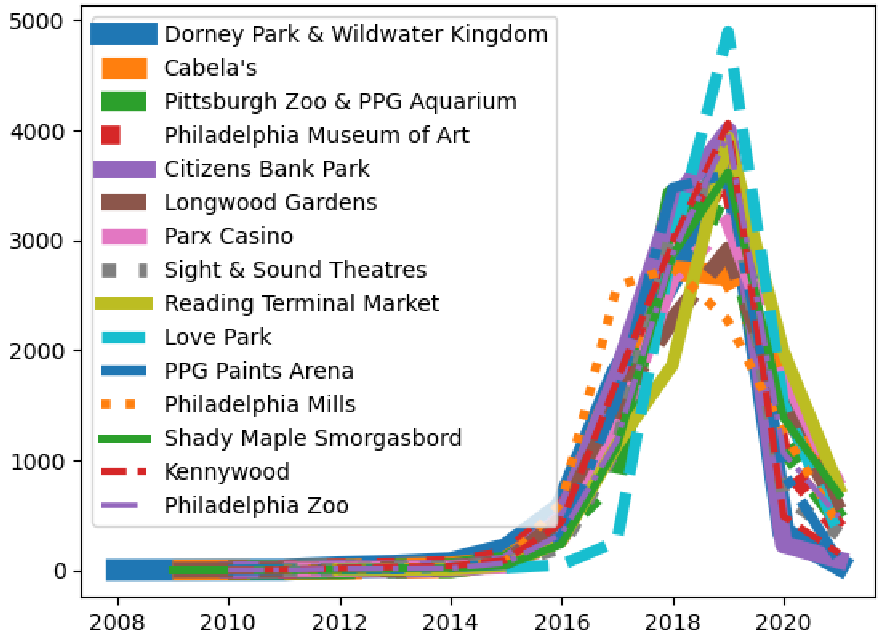
In this section, we identify the missing values and outliers and determine how we handle these values before analysis.

For the pics and response columns by users and businesses, more than 85% of the data is missing so dropped the columns

```
user_id     0.647265
name        0.000000
time        0.000000
rating      0.647265
text       43.518470
pics       97.765175
resp       86.697396
gmap_id     0.000000
```

There are 4957916 unique users, with 43743 users with no ids. So dropped all the rows with no user id and rating.

The reviews are accounted more for the period of 2016 to 2020, with very less reviews from 2008 to 2015.



Same users reviewed the same businesses multiple times. So, considering only the last review and dropping the previous reviews.

	user_id	business_name	latitude	longitude	gmap_id
0	1.141847e+20	ARCannabisClinic.com - Same Day Online Card Ap...	40.994593	-77.604989	3
1	1.091304e+20	Pocono Mountain Villas by Exploria Resorts	41.076850	-75.026198	2
2	1.093564e+20	Philadelphia Dental Associates	39.953289	-75.173645	2
3	1.134716e+20	Pocono Mountain Villas by Exploria Resorts	41.076850	-75.026198	2
4	1.005712e+20	City Cruises Philadelphia	39.941598	-75.141250	2
...	...	...	...	...	...
775	1.150284e+20	Marmont Steakhouse and Bar	39.949878	-75.144574	2
776	1.079580e+20	AutoZone Auto Parts	40.423227	-79.662546	2
777	1.124910e+20	Bourbon Mill	39.861455	-77.072725	2
778	1.107251e+20	Moviltech Cell Phone Repair & Accessories- Cus...	40.272313	-78.856489	2
779	1.172141e+20	Concord Steaks	39.852628	-75.539581	2

780 rows × 5 columns

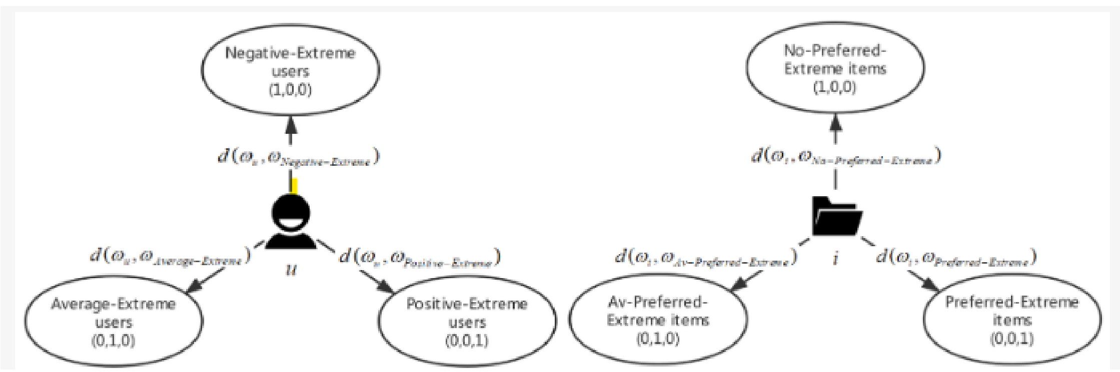
Natural Noise Detection

Based on paper Manage Natural noises [1] calculated a 3 vector rating for each user and business using below equations

$$x_u = \frac{|UR_{low}|}{|UR|}, y_u = \frac{|UR_{middle}|}{|UR|}, z_u = \frac{|UR_{high}|}{|UR|}, \qquad x_i = \frac{|IR_{low}|}{|IR|}, y_i = \frac{|IR_{middle}|}{|IR|}, z_i = \frac{|IR_{high}|}{|IR|},$$

Where UR denotes the set of all ratings provided by user u, and UR_{low}, UR_{middle}, and UR_{high} are the subsets of UR.

Based on the above matrix, for each user and business the distance is calculated to extreme users to identify if the user/item is Positive/Negative/Extreme.



After calculation based on the rules below, each rating is detected as a natural noise

	Low Ratings	Medium Rating	High Ratings
(Negative users, No-Preferred items)	-	Natural noise	Natural noise
(Average users, Av-Preferred items)	Natural noise	-	Natural noise
(Positive users, Preferred items)	Natural noise	Natural noise	-

T

For the dataset, around 1825967 records got identified as noise based on above approach. The solution is either to drop the rows or fix the rating based on a revised formula.

	user_id double precision	gmap_id text	rating double precision
1	1.055497151364822e+20	0x89c90c859e8762b9:0x6030983e3eb16c...	2
2	1.0930237408789242e+20	0x8834e6183dbbc639:0x148abd2ce6ee2e	4
3	1.0398889003227596e+20	0x89c6b356c95aabbf:0x3d96f572f571ee55	5
4	1.0475033350806992e+20	0x89c8a89d6de537e5:0x37f2a8bb631cf0a8	4
5	1.0595935226147788e+20	0x89c85f87876f97eb:0x4b22d334a14299d9	4
6	1.0952330381607603e+20	0x89c5897277278ed1:0x84ac8c27b0278c...	4
7	1.165897663010511e+20	0x89c4d81bea232023:0x897f098a51647718	4
8	1.1240940421385147e+20	0x89c6c40fc2468d51:0x44adea6f529dd00	3
9	1.1445246704054224e+20	0x8834f2f79bea23db:0x3c028b0a5c3f7ead	4

Description column is missing more than 80% of the data, so dropped the column. Based on the geospatial analysis the dataset has businesses with latitude and longitude range outside of the PA state range. The latitude and longitude coordinates for Pennsylvania, United States are approximately 41.2033° N to 42° N and -77.1945° W to 80° 31° W1. Dropped the longitude range > -70 and <-80

For MISC column except for Service options and Accessibility more than 70% of the data is missing. So dropped the columns

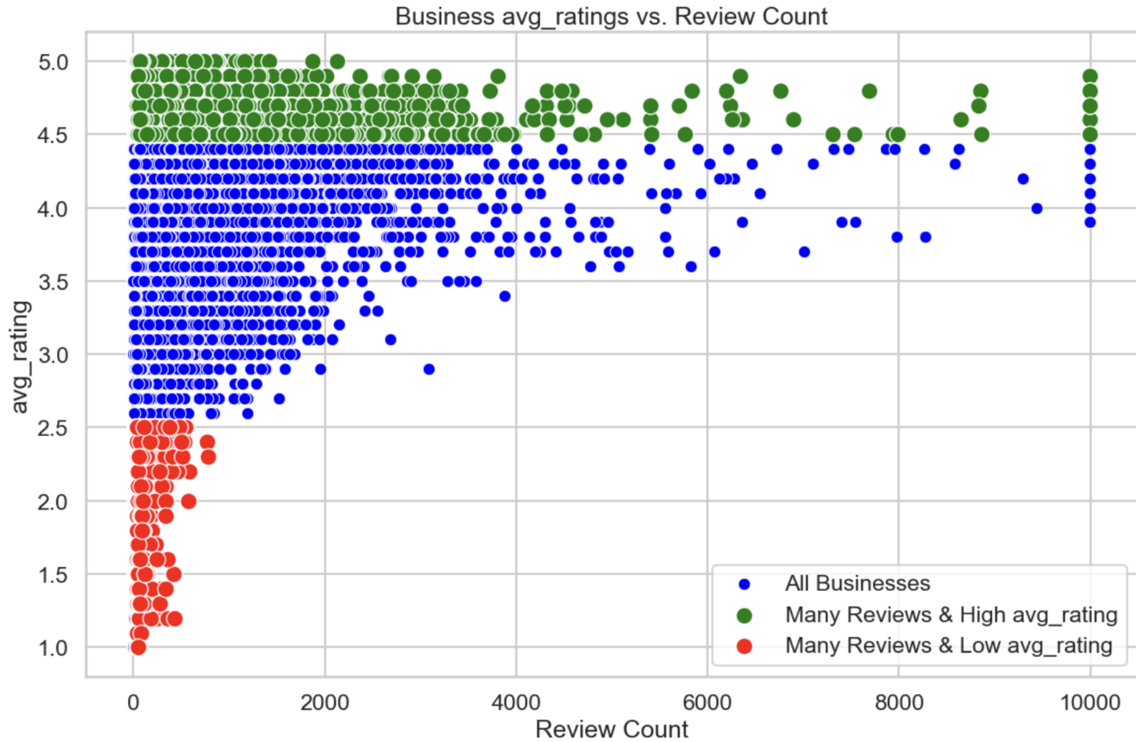
```
Service options       59.298487
Health & safety       78.469835
Offerings             76.843661
Amenities             74.717005
Atmosphere            86.479121
Crowd                 88.134119
Accessibility         34.038026
Planning              69.529809
Payments              79.906821
Highlights            86.955496
Popular for           88.397723
Dining options        88.797061
From the business     97.240273
Recycling             99.948642
Getting here          99.936588
Activities            99.990567
Health and safety     99.824438
dtype: float64
```

For Service options there are 20 different categorical features

```
array(['Curbside pickup', 'Delivery', 'Dine-in', 'Drive-through',
       'In-store pick-up', 'In-store pickup', 'In-store shopping',
       'Language assistance', 'No-contact delivery', 'On-site services',
       'Online appointments', 'Online care', 'Online classes',
       'Online estimates', 'Onsite services', 'Outdoor seating',
       'Outdoor services', 'Same-day delivery', 'Takeaway', 'Takeout'],
      dtype=object)
```

State column has as specified 1827 unique values but the feature is very sparse. However, one of the values in state is “Permanently Closed”. There are 10118 business with that value and 556459 reviews associated with that business. So dropped all the rows and the state column.

3. Business Ratings vs. Review Count:



Insights:

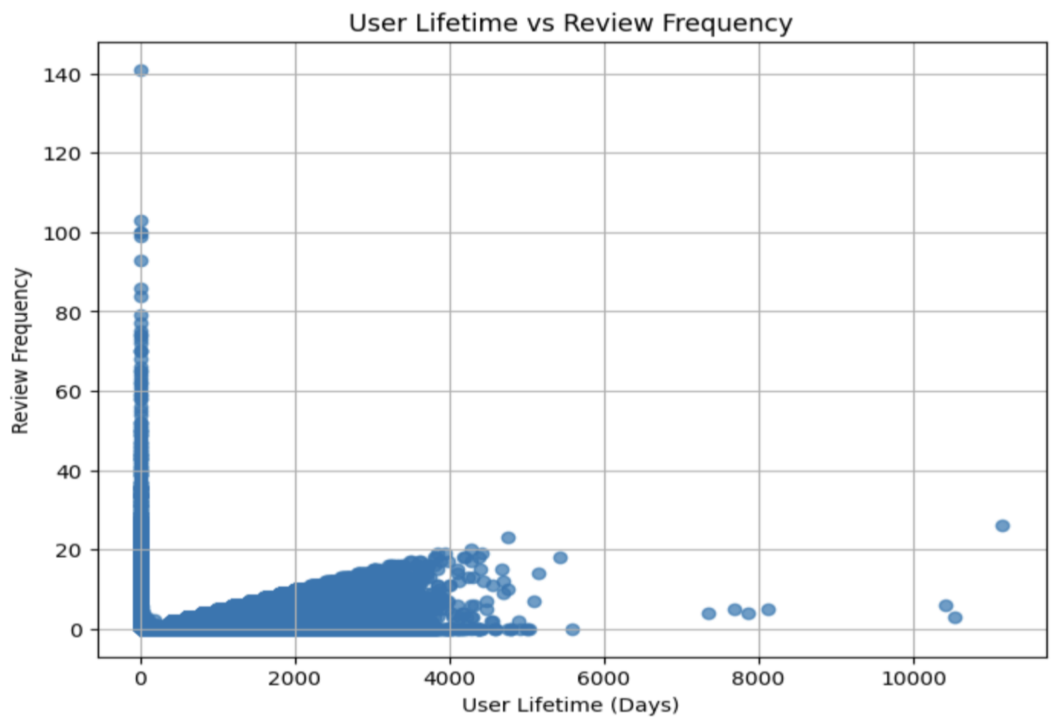
- **Cluster of Highly Rated Businesses:** The plot reveals a dense cluster of businesses with an average rating between 4.0 and 5.0. This indicates that many businesses receive high ratings from users, with varying numbers of reviews. Businesses clustered in the green area (high ratings, many reviews) are strong candidates for initial recommendations. These are generally well-regarded and popular places.
- **Impact of Review Count on Rating:**
  - **Segregation of Ratings:** The red points clearly show that businesses with low average ratings (1.0 to 2.5) are mostly concentrated with a small number of reviews. As the review count increases, it's rare to find businesses maintaining such low ratings.
  - **Outliers with High Review Count:** There are several businesses with a high review count (up to 10,000 reviews) that maintain a rating between 4.0 and 5.0, shown in green. These represent consistently highly-rated establishments that have garnered a large number of reviews.
  - **Blue dots:** Represent all business showing relation between average rating and review count.

Use in Recommender System:

- **Initial Ranking:** We can prioritize businesses based on a combination of average rating and review count. Weight review count to favor businesses with a more established reputation.

- **Filtering:** Filter out businesses with low average ratings or an insufficient number of reviews to maintain a baseline level of quality.
- **Cold Start Mitigation:** For new businesses (few reviews), we need to consider incorporating data from similar businesses (e.g., location, category) to make initial recommendations.
- **Avoid Low-Rated Businesses:** Businesses in the red area (low ratings, many reviews) should generally be avoided for recommendations unless specific user preferences indicate otherwise (e.g., someone explicitly looking for a "dive" experience).

4. User Lifetime vs. Review Frequency:



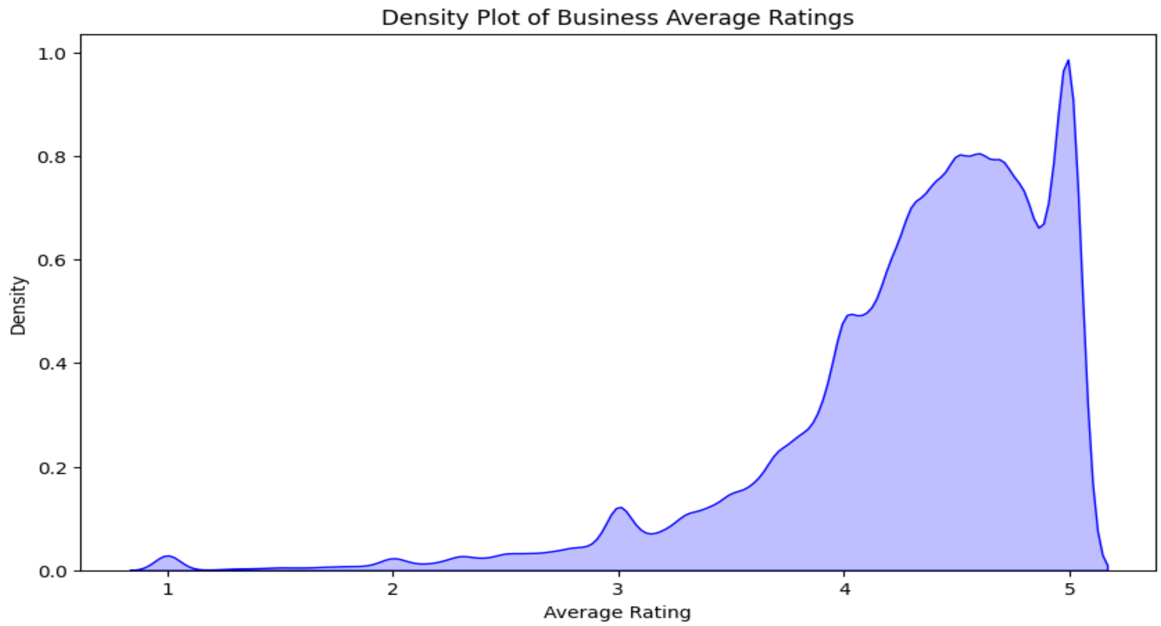
- **High Review Frequency for Short User Lifetimes:** The majority of users with high review frequency (e.g., 80–140 reviews) have very short lifetimes (close to 0 days). This suggests that these users are highly active in a brief period, possibly shortly after account creation or during specific events. Users tend to be most active early in their lifecycle, potentially due to initial enthusiasm or specific incentives for new accounts.
- **Declining Review Frequency Over Time:** As user lifetime increases (moving along the x-axis), review frequency tends to decrease significantly. Most users with longer lifetimes contribute fewer reviews, indicating reduced engagement over time.
- **Sparse Activity Among Long-Lifetime Users:** Users with lifetimes exceeding 4000 days (approximately 11 years) exhibit very low review frequencies, with only a few outliers contributing more than 20 reviews. This could indicate a small subset of highly dedicated or consistent users.

- **Cluster of Moderate Activity:** There is a noticeable cluster of users with moderate review frequencies (0–20 reviews) and lifetimes ranging from 0 to 4000 days. This represents the bulk of the user base, who contribute occasionally over their lifetime.
- **Outliers:** A few users with extremely long lifetimes (>8000 days) and relatively high review frequencies stand out as exceptions. These may represent power users or individuals who remain consistently engaged over many years.

Possible Interpretations:

- **Early Activity Matters:** By focusing on capturing user preferences and behavior early in their "lifetime" (initial interactions with the app). New users are often the most active, providing valuable data for personalization.
- **Engagement Decline:** Recognize that user engagement will likely decrease over time. Implement strategies to keep users active and provide ongoing value to maintain the relevance of recommendations.
- **Identify "Power Users":** The outliers with long lifetimes and high review frequencies represent potentially valuable sources of information. Consider soliciting their expertise to improve the recommender system (e.g., asking them to curate lists, provide feedback on recommendations).
- **Use in Recommender System:**
  - **Preference Learning:** Use the initial burst of activity from new users to quickly learn their preferences. Capture implicit data (e.g., places viewed, interactions with maps) in addition to explicit ratings and reviews.
  - **Adaptive Recommendations:** Adjust recommendations based on the user's evolving behavior. If a user's activity declines in a particular category (e.g., restaurants), shift recommendations to other areas of interest.

5. **Density plot of business reviews**



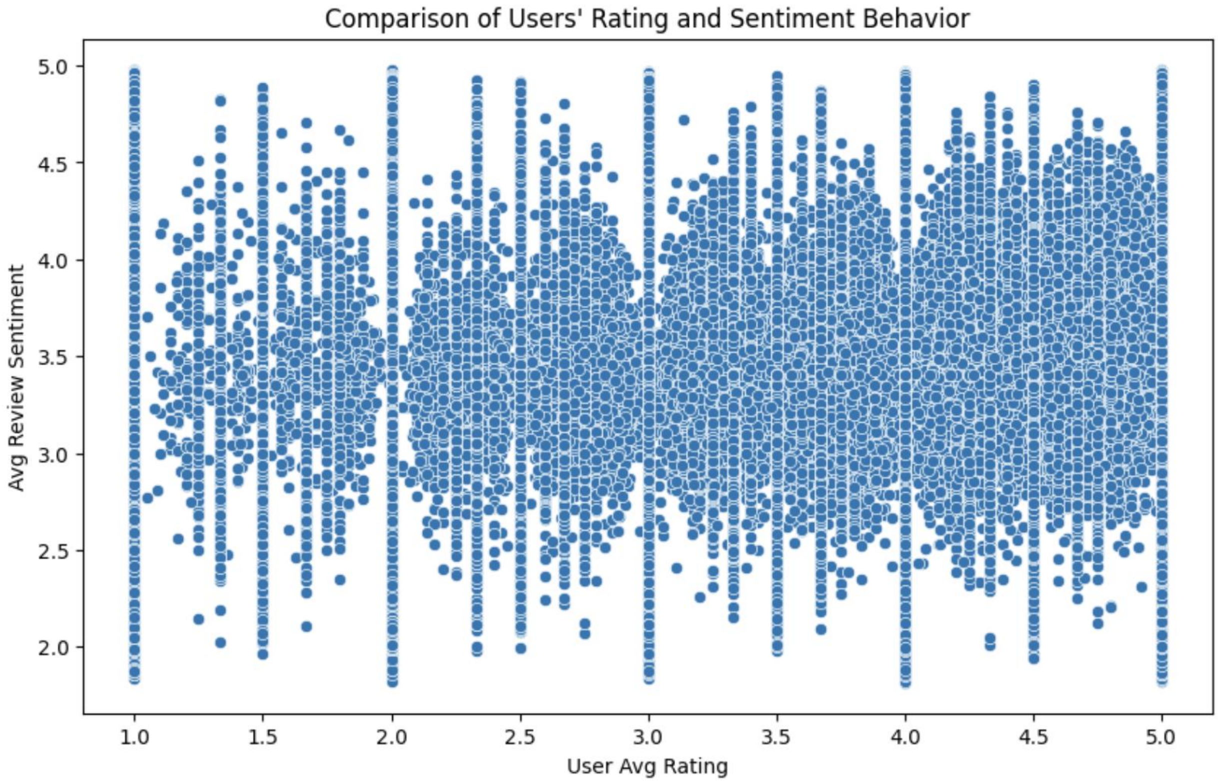
Insights:

- Most businesses have an average rating near 5, indicating generally positive customer feedback.
- A smaller, secondary peak appears around 4, suggesting a decent number of businesses are good but not perfect.
- Ratings below 3 are relatively uncommon, which means truly negative experiences are rare.
- The distribution is heavily skewed to the right, with the majority of businesses in the 4–5 range.
- A small number of businesses have very low average ratings around 1 or 2, pointing to consistently poor experiences.

Use in Recommender System:

- Highlight top-rated businesses for users seeking the best-reviewed options.
- Consider using rating normalization or Bayesian averaging, given the skewed distribution.
- Pay special attention to businesses with very low ratings, as they might need service improvements.
- Differentiate among the many 4–5 star businesses by factoring in review volume or recency.
- Flag potential rating anomalies (e.g., extremely low or high averages) for deeper investigation or personalized recommendations.

6. Users' Rating and Sentiment Behavior:



Insights:

- **Concentration around Average Sentiment:** The majority of data points are concentrated around an average review sentiment score of 3.0 to 4.5. This suggests that most users tend to express moderately positive to positive sentiments in their reviews, regardless of their average rating.

- **Vertical Bands:** The chart displays distinct vertical bands across different user average ratings. These bands indicate that for any given average user rating (from 1.0 to 5.0), there is a wide range of average review sentiment scores. This could reflect that users' sentiments can vary even if their overall rating tendency remains consistent.

- **Low Sentiment Scores at High Ratings:** Even for users with high average ratings (4.0 to 5.0), there are instances of low average review sentiment scores (around 2.0 to 3.0). This could indicate that while these users generally give high ratings, some of their individual reviews may still contain criticisms or negative feedback.

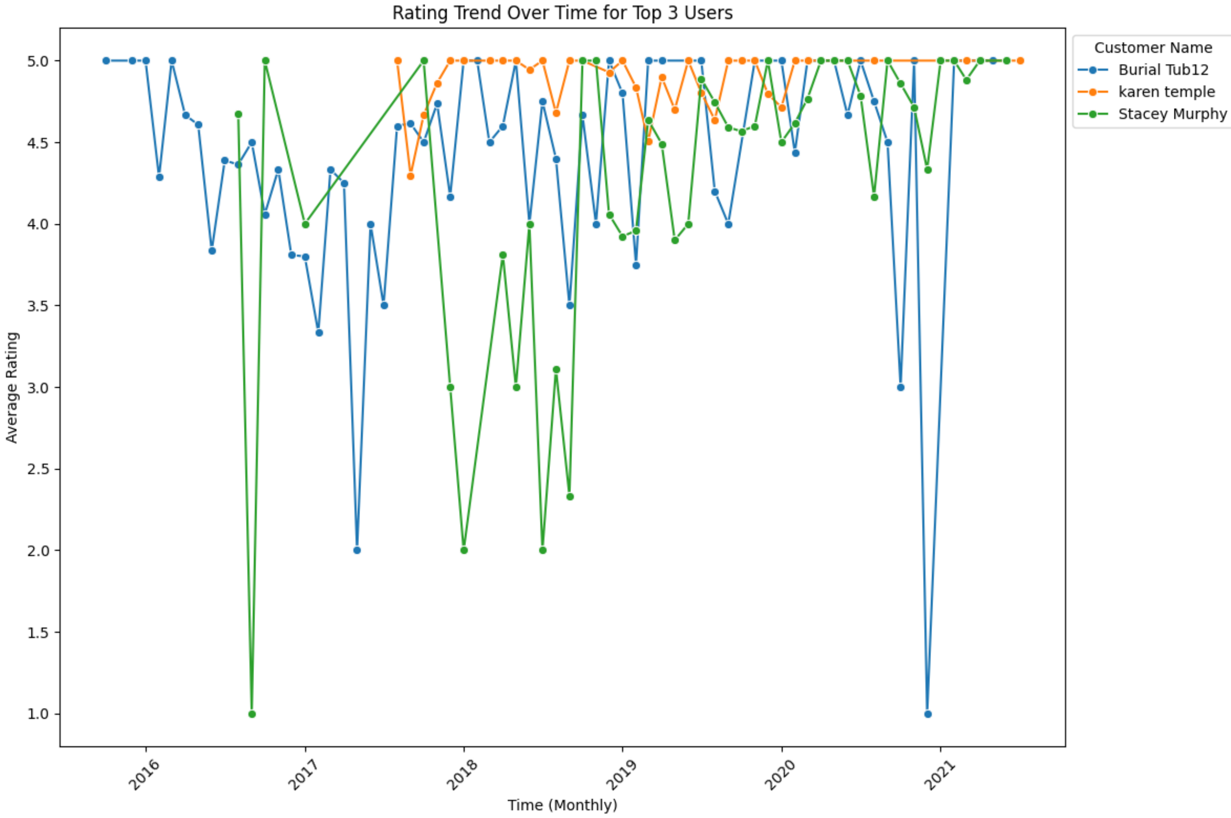
- **High Sentiment Scores at Low Ratings:** Similarly, for users with low average ratings (1.0 to 2.0), some reviews have high average sentiment scores (around 4.0 to 5.0). This could represent cases where users, despite giving a low overall rating, still express positive sentiments about specific aspects of their experience.

- **No Clear Correlation:** There is no strong positive or negative correlation between user average rating and average review sentiment. The scatter plot appears quite random, suggesting that sentiment analysis scores may not directly align with the numerical ratings users provide. While users are giving low ratings, they still express positive sentiments. This means that, they are still expecting more.

Use in Recommender System:

- **Fine-Grained Preference Modeling:** Incorporate sentiment analysis to understand why user likes or dislikes a place. Extract specific features or aspects that contribute to positive or negative sentiments.
- **Explainable Recommendations:** Provide users with explanations for why a particular place is being recommended. Highlight positive sentiments expressed by other users, focusing on aspects relevant to the current user's preferences.
- **Refine Recommendation Logic:** Use sentiment analysis to identify and address potential shortcomings in the recommendation algorithm. If the system consistently recommends places that elicit negative sentiments, adjust the ranking criteria accordingly.

7. Preference patterns of top 3 users based on time



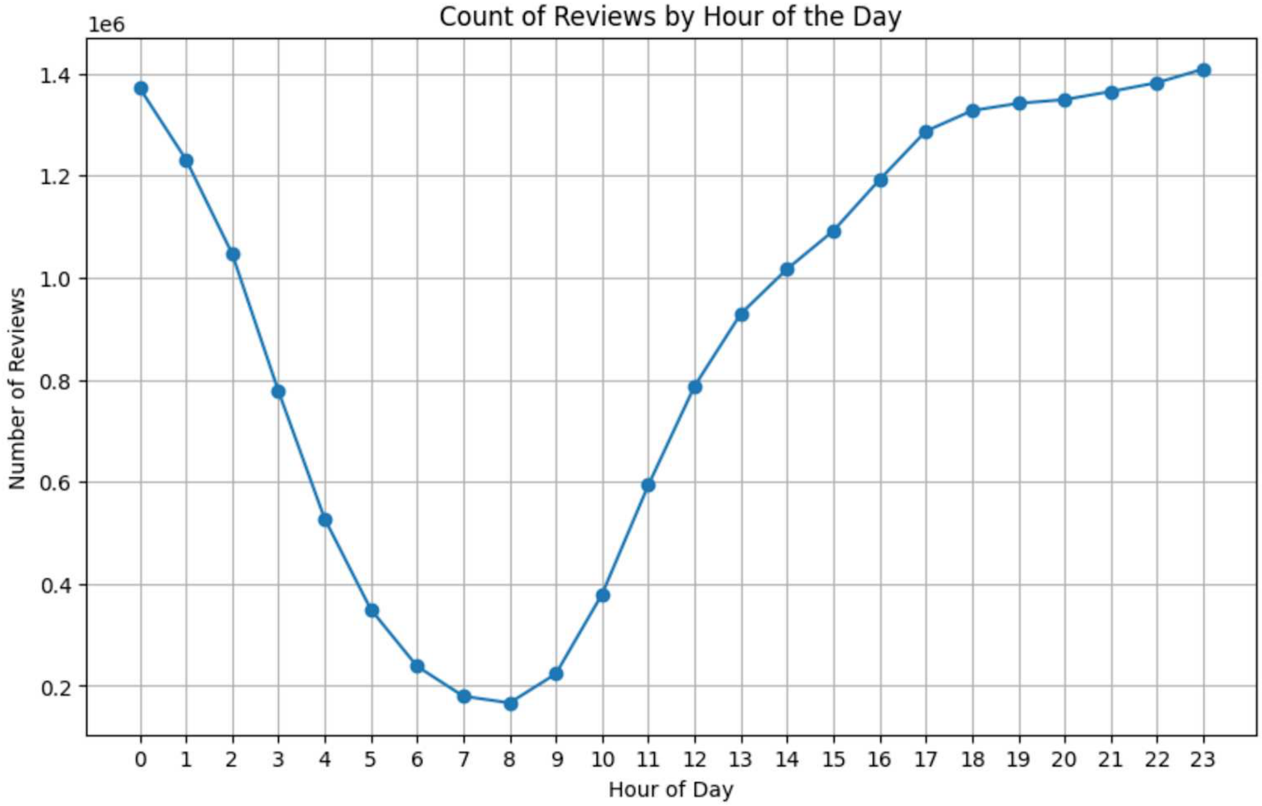
Insights:

- Most users tend to give high ratings (around 4–5 stars).
- Occasionally, there are sudden drops to low ratings, which may indicate a bad experience.
- Some users show a wide range in their ratings, while others are very consistent.
- Overall, users' ratings remain high over time, with only a few exceptions.
- Changes in ratings can sometimes be linked to specific times or events.

Using These Insights in Our Recommender System:

- Give more weight to recent reviews if a user's taste appears to change.
- Recommend similar, trusted businesses to users who consistently give high ratings.
- Offer a diverse set of options to users whose ratings vary widely.
- Adjust recommendations quickly if a user suddenly starts giving low ratings.
- Factor in time-based trends, like seasonal changes, to refine suggestions.

8. Analysis of Review Frequency by Time of the Day (User Perspective)



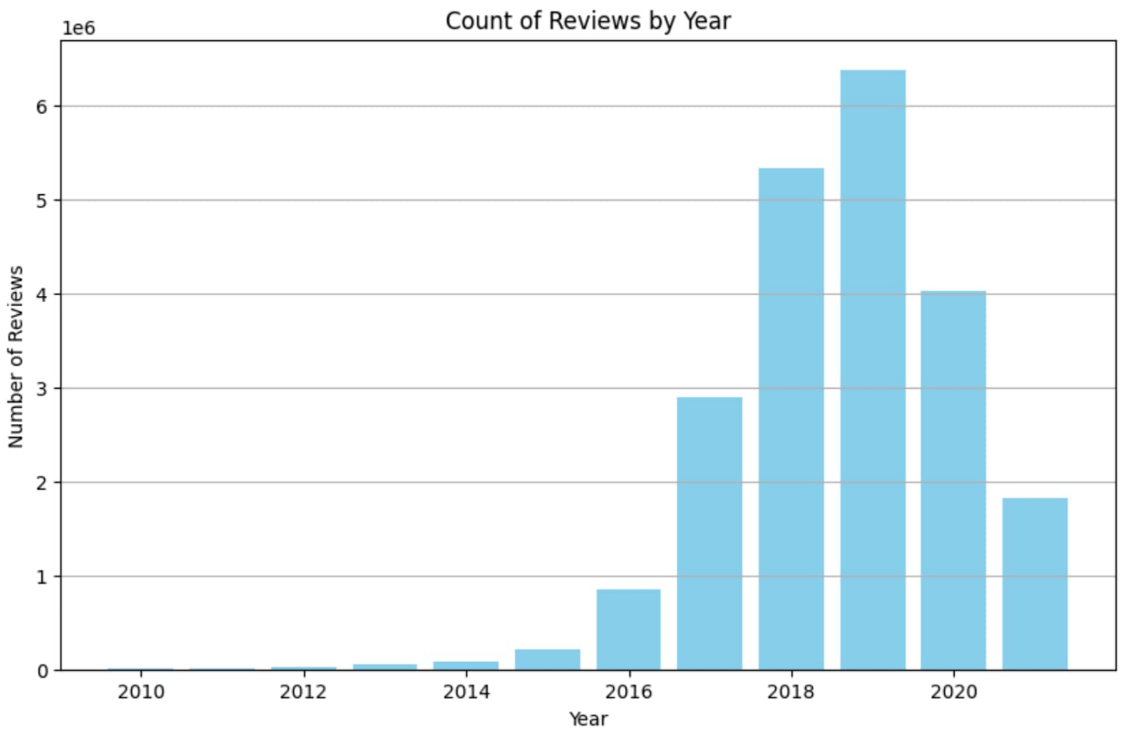
Insights:

- **Peak Review Times:** The graph shows that the highest volume of reviews occurs during the late evening and overnight hours (18:00 to 00:00). A smaller peak is observed in the morning (00:00 to 1:00).
- **Lowest Review Times:** The lowest review activity happens in the early morning (around 6:00 to 9:00).
- **Gradual Increase:** There's a gradual increase in review frequency from late morning (10:00) through the afternoon and into the evening.

Use in Recommender System:

- **Real-Time Updates:** We conclude form the graph that most reviews are written in the evening. Prioritize processing and incorporating these new reviews into the recommender system quickly to ensure that users see the most up-to-date information.
- **Timing of Recommendations:** Consider the time of day when making recommendations. If a user is active in the app during the evening, they may be more receptive to recommendations for places open late or with evening activities.
- **Activity-Based Recommendations:** Leverage this data to infer user activities and make relevant recommendations.

9. Analysis of Review Volume by Year



Insights:

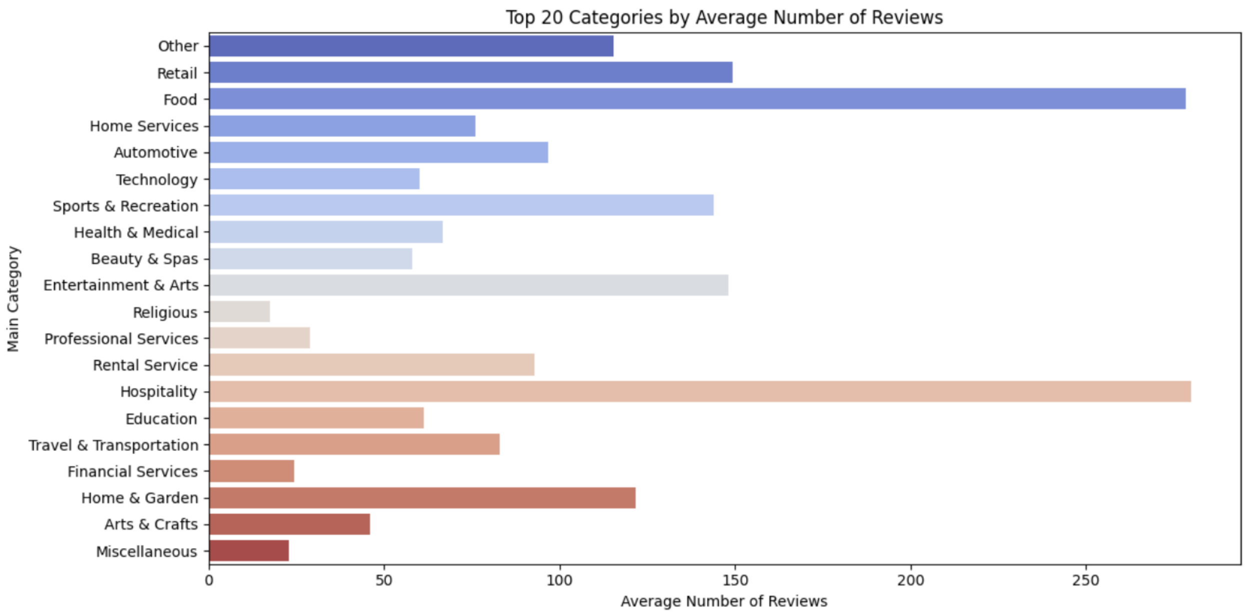
- **Exponential Growth (2017-2019):** There's a significant surge in the number of reviews between 2017 and 2019. This period represents a peak in user engagement and review activity.
- **Plateau/Decline (2020-2021):** After 2019, the review count plateaus and even declines slightly. This could indicate saturation of the market, a shift in user behavior, or external factors affecting tourism.
- **Limited Data Before 2017:** Data from years prior to 2017 is sparse, making it less reliable for training or evaluating the recommender system.

Use in Recommender System:

- **Weighting Recent Data:** Give more weight to recent reviews in the recommender system's algorithms. This ensures that recommendations are based on the most up-to-date information and user preferences.
- **Seasonality and Trend Analysis:** Analyze the reviews within each year to identify seasonal trends and predict future patterns. This allows the system to proactively adjust recommendations based on the time of year (e.g., suggesting outdoor activities in the summer, indoor attractions in the winter).

- **Monitoring for Anomalies:** Continuously monitor the review volume for any sudden spikes or drops. These could indicate new trends, crises, or changes in user behavior that require adjustments to the recommender system.
- **Cold Start Strategy (Historical Data):** Because of limited historical data before 2016, implement a solid cold start strategy using the most recent data trends and incorporate initial data from similar business.
- **Model Retraining Schedule:** Establish a regular schedule for retraining the recommender system models based on the annual review data. This ensures that the system adapts to evolving user preferences and market conditions.

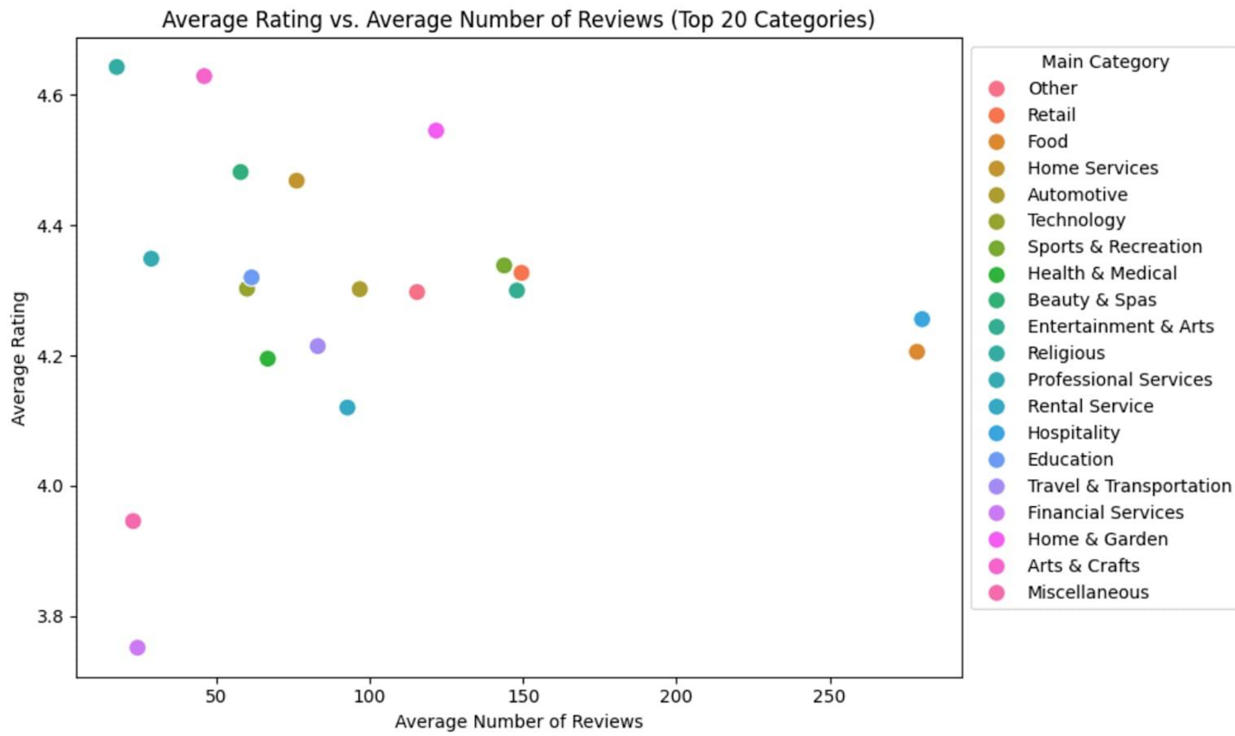
10. Top 20 Categories by Average Number of Reviews:



Insights:

- **"Hospitality" and "Food" Dominate:** These categories have the highest average number of reviews, suggesting that these are the most reviewed categories on the platform. This means users are highly engaged with providing feedback for these types of businesses.
- **"Other," "Retail," and "Sports & Recreation" See Moderate Review Volume:** These categories also see a sizable average number of reviews, but less than "Hospitality" or "Food".
- **Niche Categories Have Few Reviews:** Categories like "Financial Services," "Arts & Crafts," and "Miscellaneous" have the lowest average number of reviews, indicating less feedback for these types of businesses.
- **Review Volume as Engagement Indicator:** Review volume is a good indicator of user engagement and interest in different categories.

11. Average Rating vs. Average Number of Reviews (Top 20 Categories)



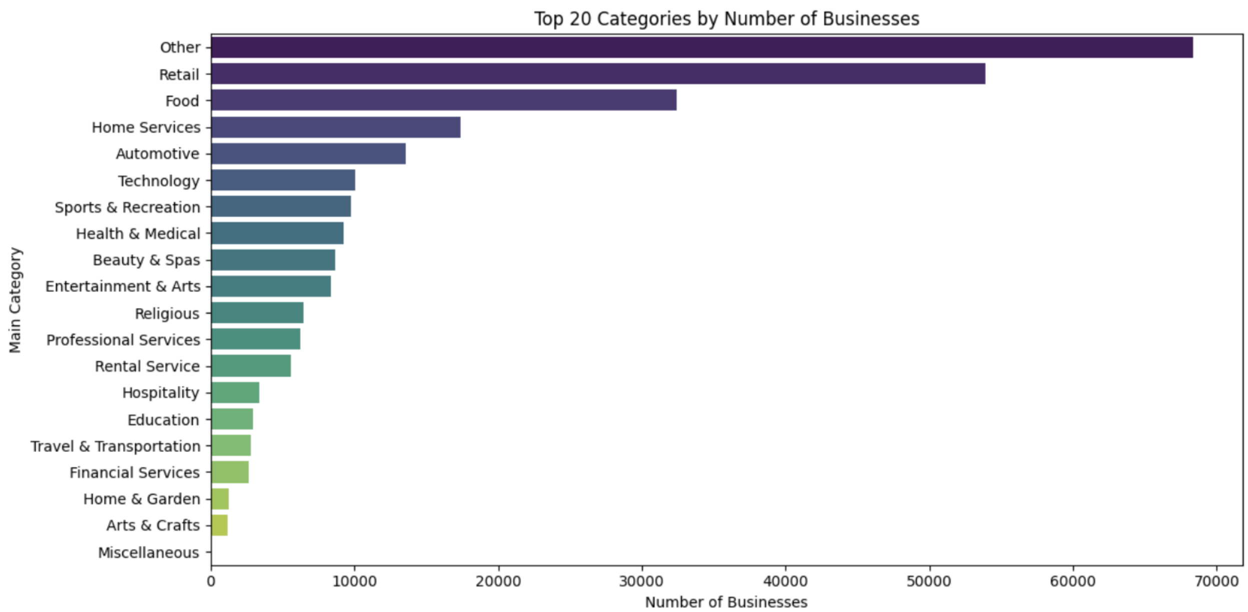
Insights:

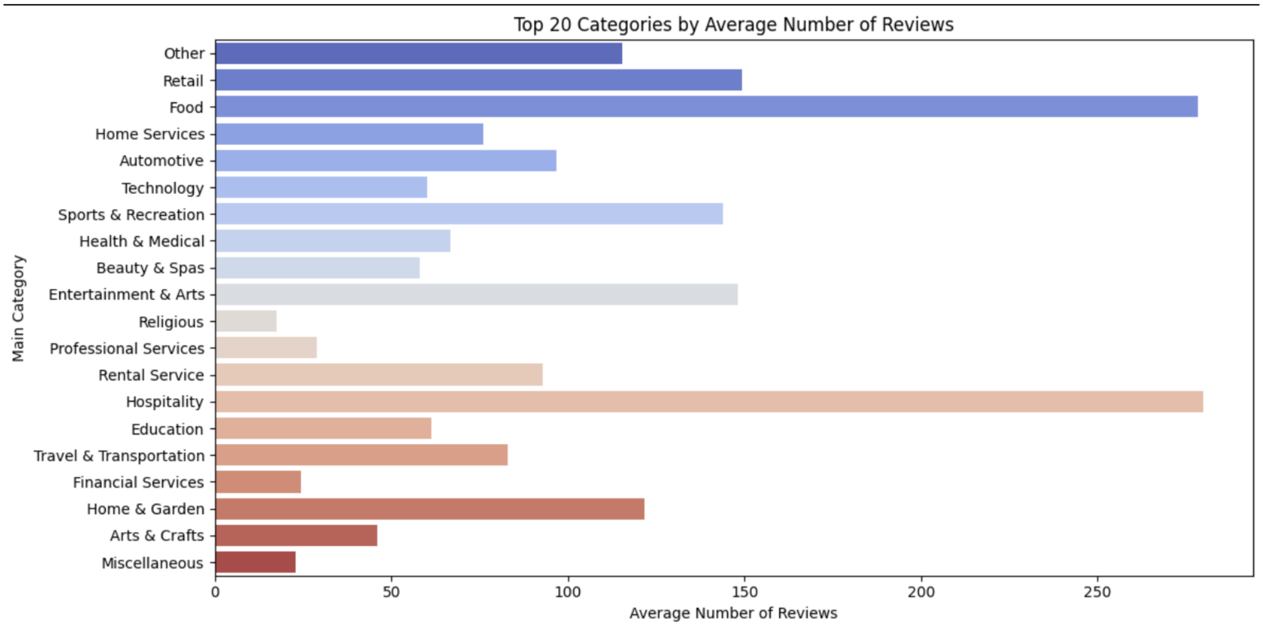
- **No Strong Correlation:** There isn't a clear linear relationship between the average number of reviews and the average rating. High review volume doesn't necessarily translate to higher or lower average ratings.
- **Categories with High Ratings:** "Arts & Crafts" and "Home & Garden" have a relatively high average rating, even though their review volume is lower than other categories.
- **Categories with Low Ratings:** "Miscellaneous" has low average rating as well as low number of reviews.
- **Cluster of Moderate Ratings:** Many categories cluster around an average rating of 4.2 to 4.4, regardless of the average number of reviews. This suggests that most categories maintain a reasonably positive perception.
- **Categories with Many Reviews Don't Dominate Ratings:** "Food" and "Hospitality", despite having the most reviews, don't necessarily have the highest average ratings. Their ratings are within the middle cluster.
- **Rating Stability:** Average rating is relatively stable as average number of reviews changes.

➢ Feature Engineering and Analysis

In this section, we identify the variables that are useful for predictive modeling and machine learning through correlation analysis. You may also reduce the dimension or derive new variables so that the predictive modeling can be more efficient and effective.

1. Redefining the Categories: From Un-Structured to Structured:





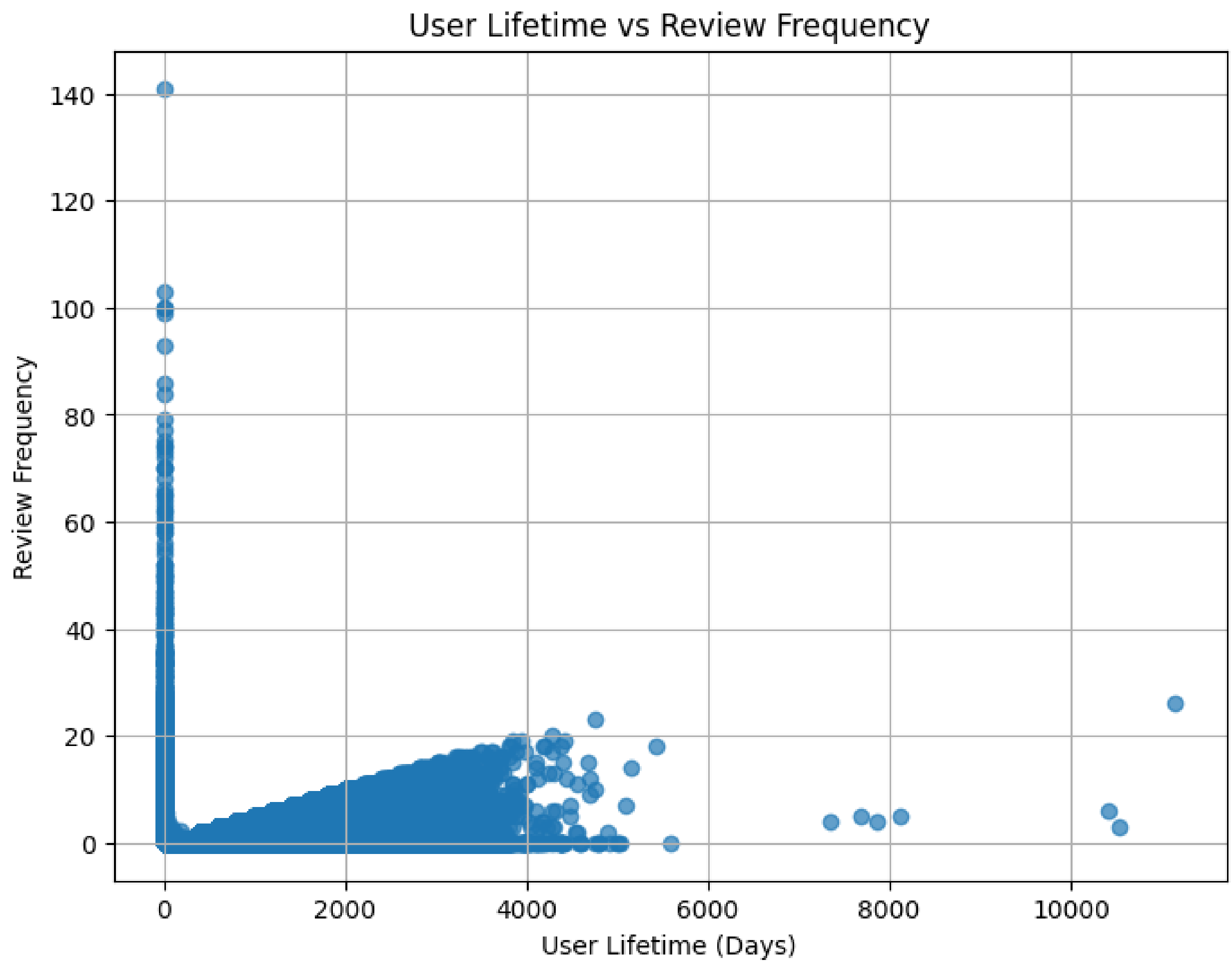
Process Involved:

- **Unstructured to Structured Conversion:** The initial "Categories" column contained unstructured text with various tags. To convert this into a usable format for machine learning, the following steps were taken:
- **Unique Tag Extraction:** All unique tags present in the "Categories" column were identified.
- **Category Consolidation:** Certain categories can be combined or refined to reduce dimensionality or improve interpretability.
- **One-Hot Encoding:** One-hot encoding was applied to create new binary columns for each unique tag. If a business was tagged with a particular category, the corresponding column would have a value of 1; otherwise, it would be 0.
- **Top Category Selection:** The top 20 categories based on the number of businesses were selected for analysis and visualization.

Insights:

- **Dominant Categories:** The chart reveals that "Other" and "Retail" are the most prevalent categories by a significant margin. This suggests that these sectors have a large number of businesses represented in the dataset. "Food" is also a major category.
- **Mid-Range Categories:** Categories like "Home Services," "Automotive," and "Technology" represent a substantial number of businesses, falling in the mid-range of the distribution.
- **Niche Categories:** Categories like "Financial Services," "Home & Garden," "Arts & Crafts," and "Miscellaneous" have relatively few businesses compared to the top categories.
- **Distribution Shape:** The distribution of categories is skewed, with a few dominant categories and many less frequent categories.
- The "Other", and "Retail" categories have the highest number of businesses but do not have the highest average number of reviews. Categories like "Food", "Hospitality", “Entertainment”, and "Sports" tend to have high average number of reviews
- **Based on this analysis we have decided that we will only include top 3-5 categories with highest number of reviews.**

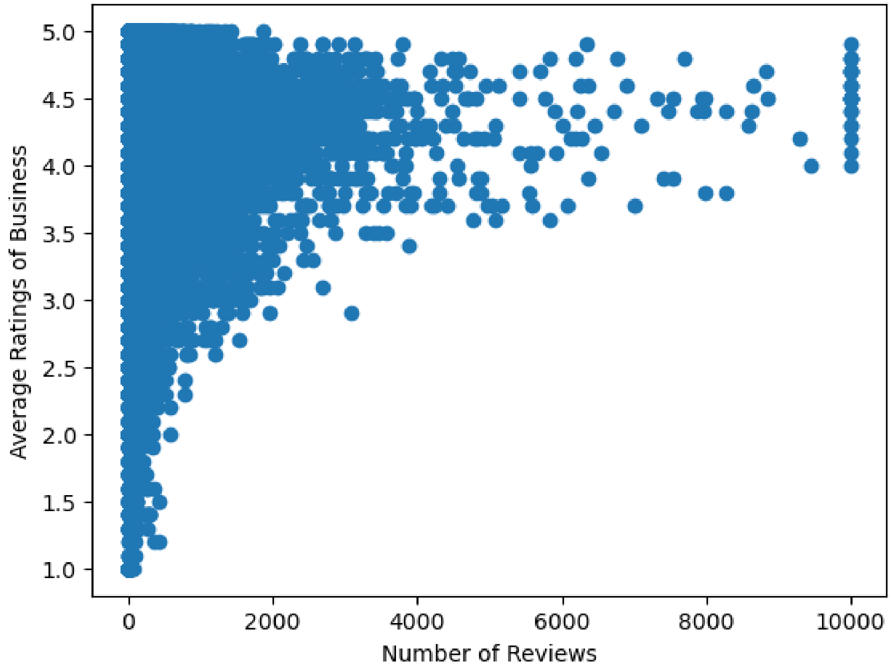
Determining Outliers based on the Users Involvement



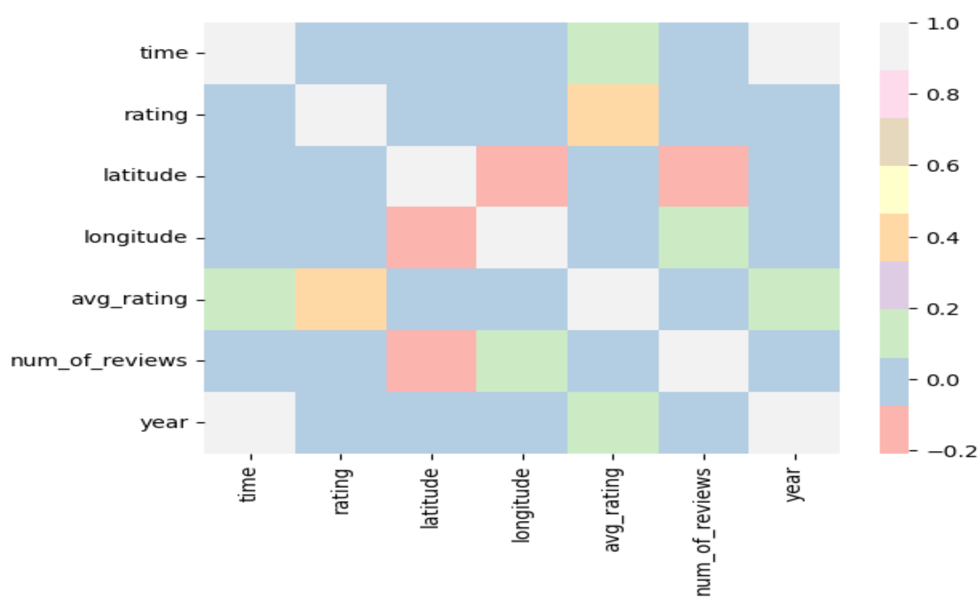
Actions taken:

- The above plot shows the review frequency of a user based on their lifetime, lifetime is basically the number of days the users was being active (last review date – first review date).
- The plot shows multiple users which have lifetime of 1 days but has more than 20 reviews, and users with 6000+ days has only 7-8 reviews.
- Based on this analysis we have decided that we will remove users which have 1 lifetime day and has 20+ reviews, and users with 6000+ days with less than 9 reviews.

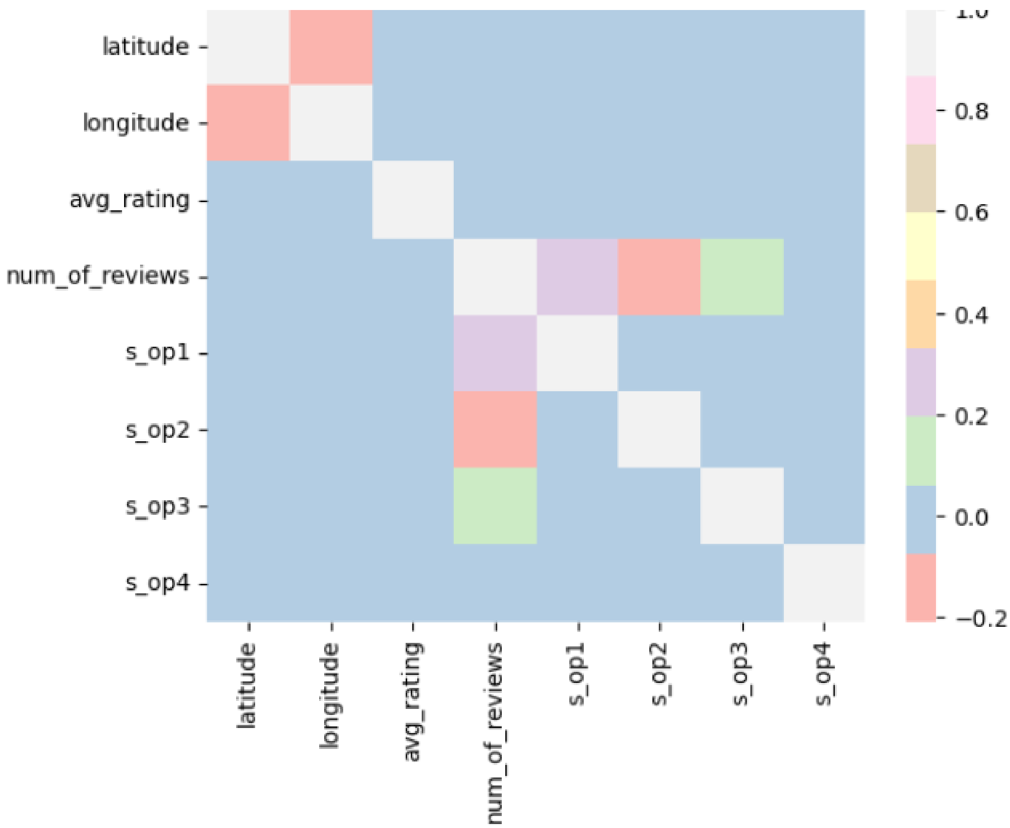
Analysis between number of reviews and average rating



Correlation Analysis



For the MISC column as identified in the outliers all the columns except Service Options did not have enough data. The column has 20 different features which are sparse. So performed a truncated SVD algorithm to reduce the number of the features to 4. It explained around 80% of the variance. The correlation with the new features


latitude
longitude
avg_rating
num_of_reviews
s_op1
s_op2
s_op3
s_op4
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Appendix

[1] Luo C, Wang Y, Li B, Liu H, Wang P, Zhang LY. An Efficient Approach to Manage Natural Noises in Recommender Systems. *Algorithms*. 2023; 16(5):228. https://doi.org/10.3390/a16050228

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	**Writing**	**Editing**
1	**Analysis the basic metrics of variables**	**Rushikesh, Jelitta**	**Hassaan, Priyanka**
2	**Non-graphical and graphical univariate analysis**	**Krish, Jelitta, Rushikesh**	**Hassaan, Priyanka**
3	**Missing value and outlier analysis**	**Priyanka, Krish, Hassaan**	**Jelitta, Krish**
4	**Feature engineering and analysis**	**Hassaan, Rushikesh, Krish**	**Jelitta, Priyanka**
5	**Appendix**	**Priyanka**	**Jelitta**