Data_Exploration_Report

Contents

1.Purpose and Deliverable

This document provides a data exploration report for the Yelp Open dataset. It provides following deliverables:

1. A report of raw data that allows potential users to understand the data structure, content, and variable definitions.
2. Publication-ready presentation that is rendered in a clean, formatted format with text, tables, and figures.

The purpose of this report is to ensure transparency and reproducibility, helping readers understand the cleaned dataset that serves as input for our subsequent regression analysis.

2.Data Overview and Sources

The project utilizes two main datasets from the Yelp Open dataset which are retrieved from Google Drive links for reproducibility:

1. business.csv contains information about businesses,including name, category, star ratings,and review counts.
2. photos.csv contains metadata for photos uploaded on Yelp, including their associated business_id,caption,and label.

Only restaurant businesses were retained for this study, following a systematic filtering and cleaning process.

3.Data Preparation Summary

The datasets were merged and cleaned using R (dplyr, tidyr, ggplot2) following a structured workflow:

1.Merge by Business_id:

photos.csv and business.csv were merged on business_id to match photo information with business-level attributes.

2.Variable Selection:

Retained business_id, review_count, name, attributes, categories, stars, photo_id, caption, and label variables that are relevant for analysis

3.Photo Recategorization:

Photo labels were simplified into three categories:

-Food & Drink → for "food" and "drink" labels

-Environment → for "inside" and "outside" photos

-Menu → for "menu" photos

4.Aggregation by Business:

For each restaurant (business_id), the number of photos in each category was counted and reshaped so that each became a column:

-food_and_drink

-environment

-menu

-total_photos

5.Filter for Restaurants:

Using the categories column, only businesses containing the keyword "Restaurants" were retained.

6.Duplicate Removal:

Identical rows were removed to ensure each restaurant appears only once.

7.Output:

The cleaned dataset, named final_dataset.csv, was created with 29,374 observations and 10 variables.

4.Variable Description

Variable	Description	Data Type
`business_id`	Unique Yelp business ID	Character
`name`	Business name as displayed on Yelp	Character
`attributes`	Map of amenities, services, and policies	List
`categories`	Yelp classification of cuisines and type	Character
`stars`	Average Yelp rating (1–5 scale)	Numeric
`review_count`	Number of Yelp reviews per restaurant	Numeric
`environment`	Number of environment-related photos	Numeric
`food_and_drink`	Number of food & drink photos	Numeric
`menu`	Number of menu photos	Numeric
`total_photos`	Total number of photos for the business	Numeric

This structured data forms the foundation for statistical modeling and visualization

5.Data Integrity Checks

Several validation steps ensured the dataset's accuracy and consistency:

1.Missing Data: Missing photo categories were treated as 0 (absence of photos, not missing data). This approach maintains data completeness and interpretability.

2.Value Ranges: The variables stars and total_photos were inspected for outliers and logical consistency (no stars > 5 or < 1).

6.Exploratory Analysis

6.1 Summary Statistics

-The dataset includes 29,374 restaurants.

-The average Yelp rating is around 4.0 stars, with most ratings between 3.5 and 4.5.

-Total photos per restaurant vary widely, from 0 up to over 400.

-Food & Drink photos dominate the dataset, followed by Environment and Menu photos.

6.2 Visual Exploration

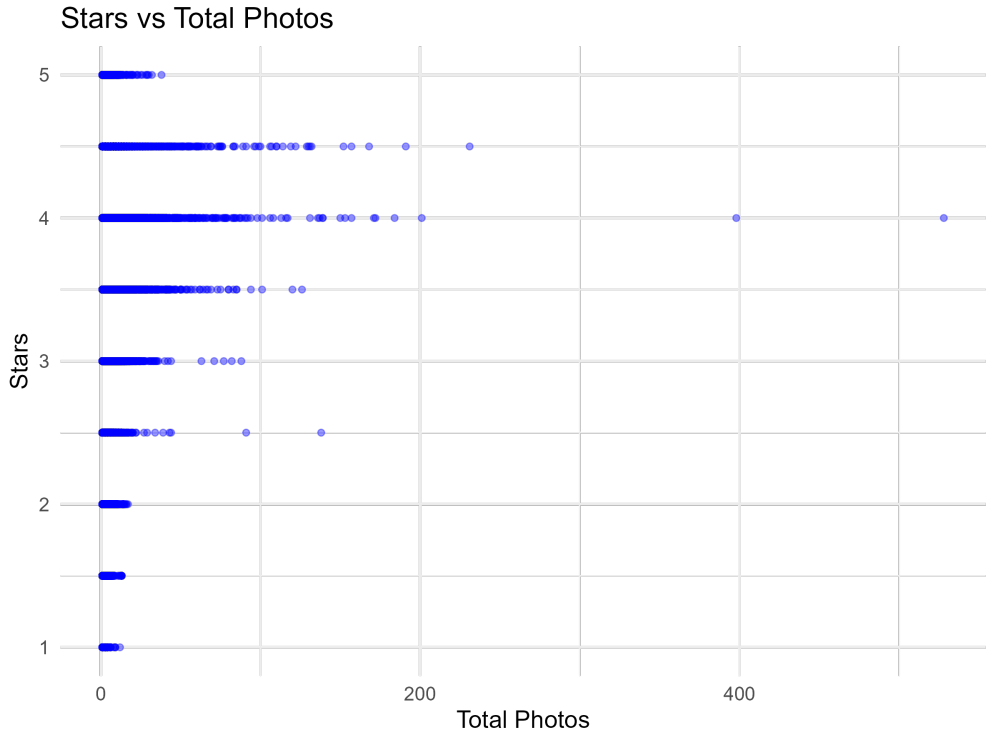

Figure 1. Stars vs. Total Photos: *A scatter plot illustrating a slightly positive relationship between the total number of photos and restaurant ratings.*

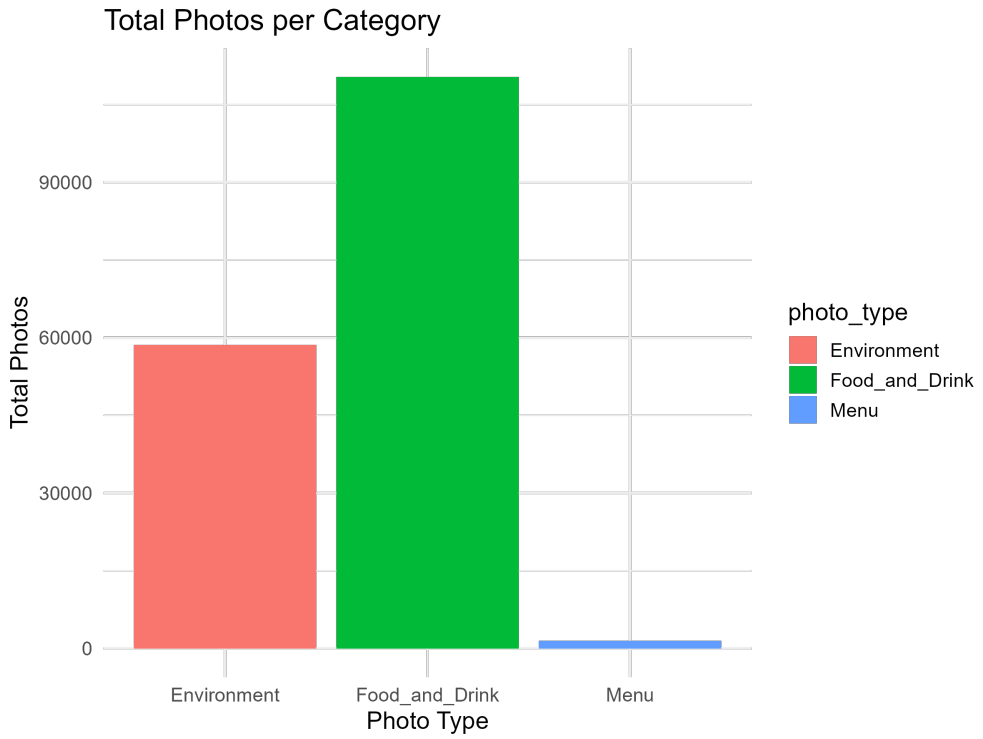

Figure 2. Total Photos per Category: *Bar plot demonstrating that Food & Drink photos are by far the most common type, followed by Environment and Menu photos.*

7.Key Insights

1.The dataset contains a substantial sample of restaurants.

2.Ratings concentrate in the mid-to-high range , consistent with prior Yelp research.

3.Restaurants with more photos tend to have slightly higher average ratings.

4.Food & Drink photos dominate total imagery, suggesting visual presentation of dishes is the most common user-generated content.Menu photos are less frequent but may indicate greater transparency or professionalism.

6.A modest positive correlation is visible between photo counts and ratings, warranting regression tests that control for review volume and category.

7.The data are suitable for regression analysis, showing adequate variation in both independent and dependent variables.

8.Next Steps and Planned Modelling

Following the completion of exploratory analysis, the next stage of the project involves estimating a series of regression models to formally test the research question:

"How does the total number of photos included in Yelp reviews influence a restaurant's average rating, and to what extent does the type of photo (food, environment, menu) moderate this relationship?"

To address this question, a set of **Multiple Linear Regression (MLR)** models will be developed using the cleaned dataset of 29,374 restaurant observations.

Each regression model will be summarized and exported as a figure in PNG format to ensure reproducibility and clarity.
Visualizations of model diagnostics and descriptive summaries will be in the final report along with results.

The purpose of this section is to outline the analytical framework guiding the next phase of the study. Final results are stored in final_paper.pdf.