Snapp! Food

Data Cleaning Challenge

Interview Task Report

Presenter:
Hamed Araab

Introduction

- **Objective:** Showcase expertise in data cleaning, analysis, visualization, and reporting for restaurant data in Ahvaz.
- **Dataset:** Restaurant data with columns: Name, Address, Rate, Review, Phone, Super Type, Marketing Area.
- **Task Overview:**
  - Clean the provided dataset.
  - Add "Subtype" and "Type ID" columns.
  - Assign grades (A to C) based on a scoring system using Rate and Review.
  - Create and visualize a comprehensive report.

Data Cleaning

Objective: Ensure data integrity by removing incomplete or incorrect entries.

Data Cleaning

- **Stripping Strings:** Removed leading/trailing whitespace from all string columns to ensure consistency.
- **Duplicate Check:** Verified no duplicate restaurant names in the dataset.
- **Normalization:**
  - **Rate:** Converted Persian digits and decimal separators to English (e.g., "۳٫۸" to 3.8).
  - **Review:** Removed "نفر" and parentheses, converting to numeric values (e.g., "(۲۲ نفر)" to 22).
  - **Phone:** Used `phonenumbers` library to format valid phone numbers in national format (e.g., "tel://0613222611" to "061 3222 6611") and marked invalid entries as "Unknown".
- **Dropping Missing Values:** Removed rows with null values in any column, reducing dataset from 49 to 46 entries.

Adding Subtype and Type ID

Objective: Derive "Subtype" and "Type ID" columns based on Super Type.

Adding Subtype and Type ID

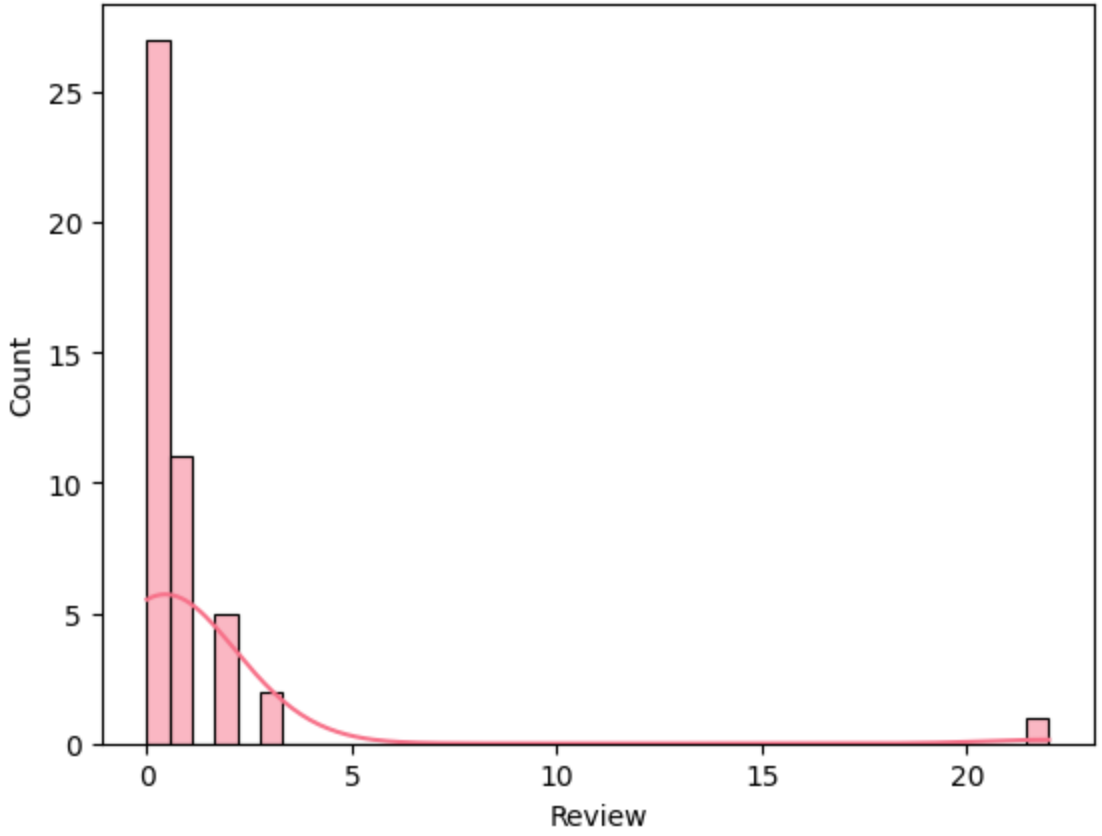
- **Subtype:**
  - If the record's Name contains its Super Type (e.g. "رستوران") and it is preceded or succeeded by a word with a "و" in between, that word is assumed to be Subtype.
  - Otherwise, a map is created to find a set of keywords corresponding to each predefined Subtype. If the record's Name contains any keyword, the corresponding Subtype is assigned to it.
  - The type pairings included an unreasonable pair: "فست‌فود" and "سنتی". In such cases, "فست‌فود" is replaced with "رستوران".
- **Type ID**: Things are kept simple. Thus, the Type ID feature is created by joining the Super Type and Subtype values with an underscore (e.g., "رستوران_سنتی").

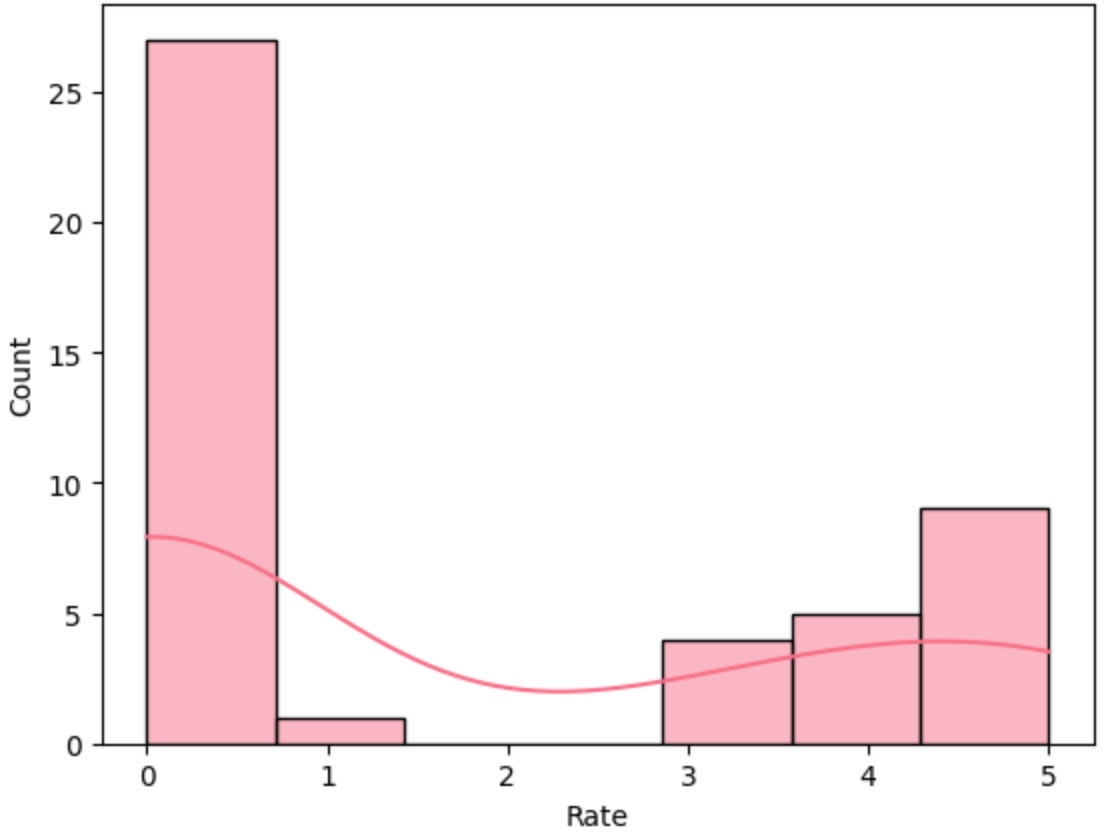
Grading System

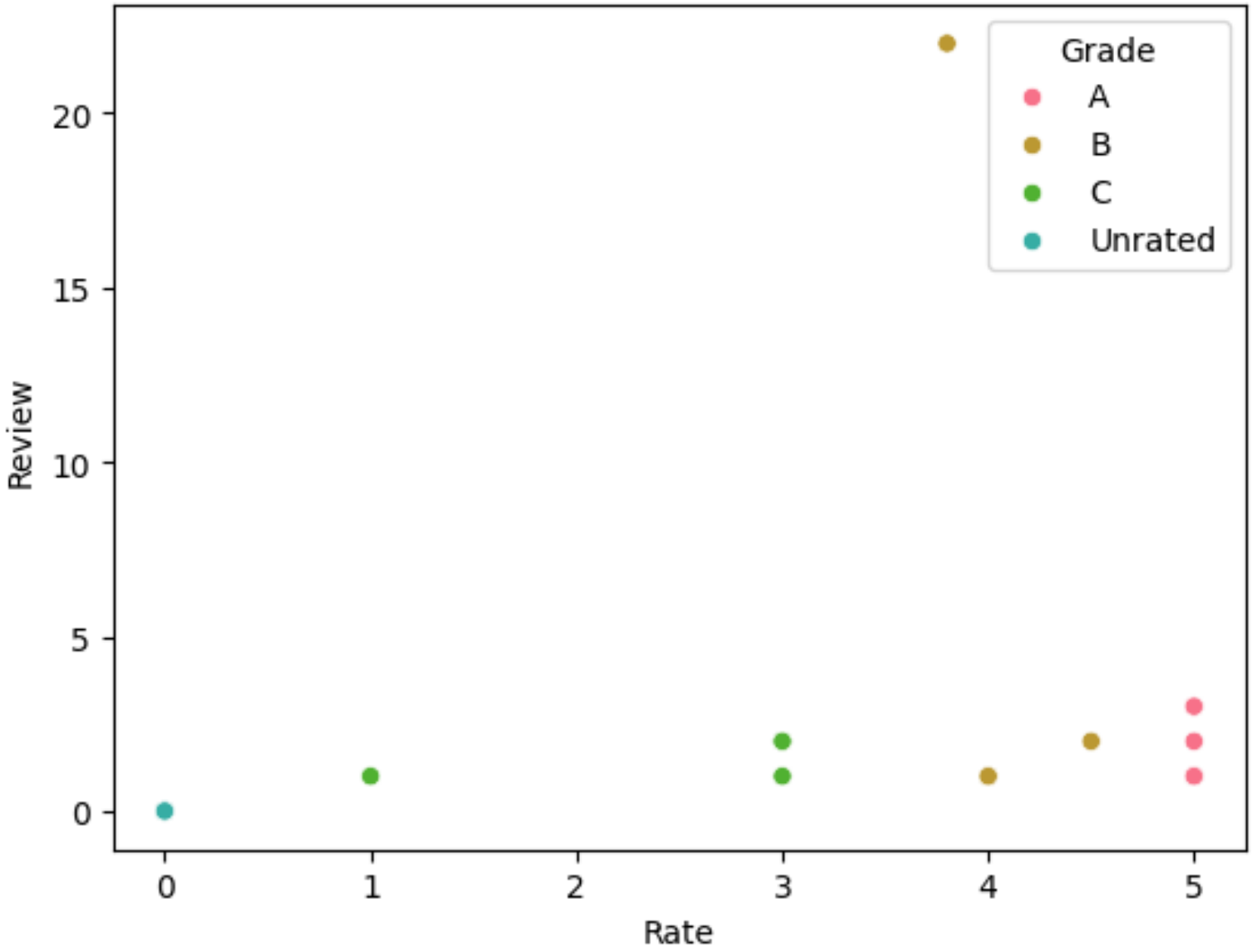
Objective: Assign grades (A to C) to restaurants based on Rate and Review.

Grading System

- **Input Data:** A copy of the dataset containing only the Rate and Review features with no outliers.
- **Clustering:**
  - Train a 3-means clustering model with the input data.
  - Sort the cluster centers according to the sum of standardized Rate and Review.
  - Label the sorted cluster centers as A, B, and C.
- **Grading:** For each record,
  - If Rate = 0, then grade = Unrated
  - Otherwise, grade = the closest cluster center

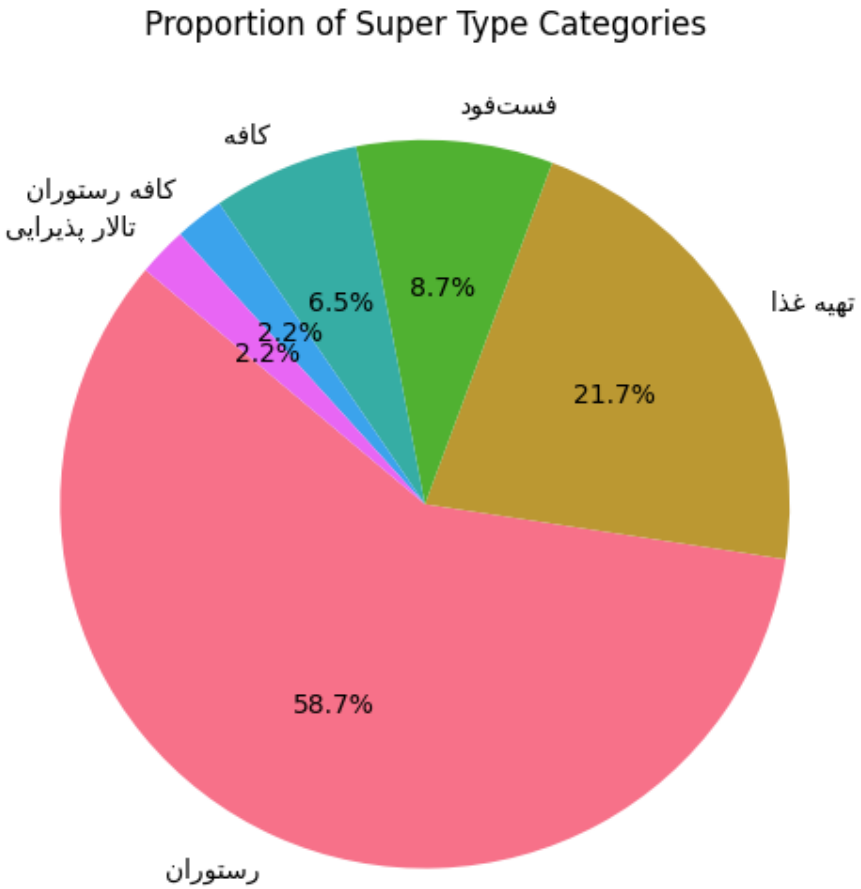

Count
Review
0
5
10
15
20
25

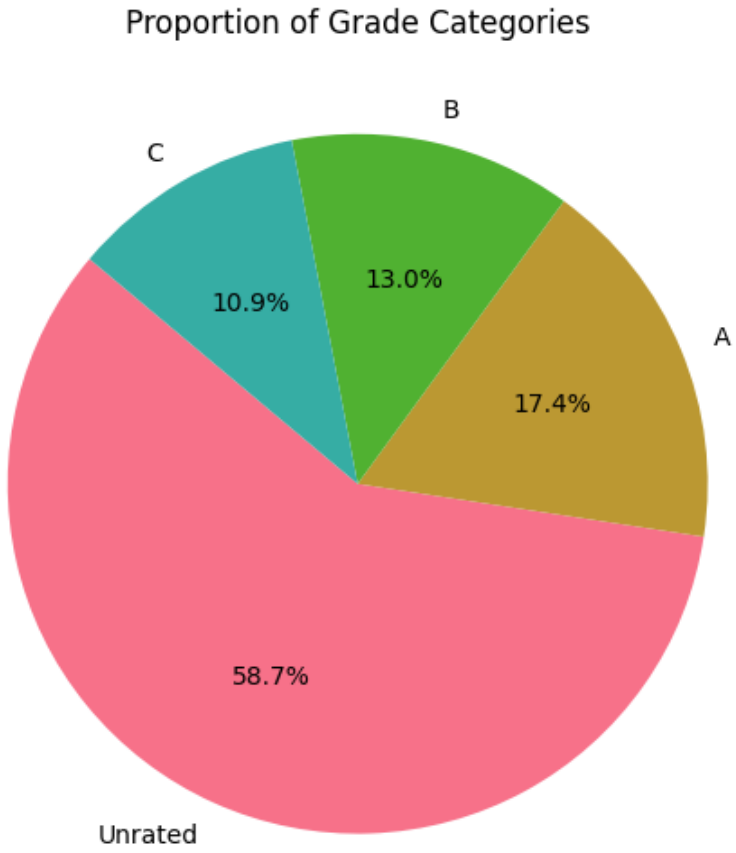

Count
Rate
0
1
2
3
4
5
10
15
20
25

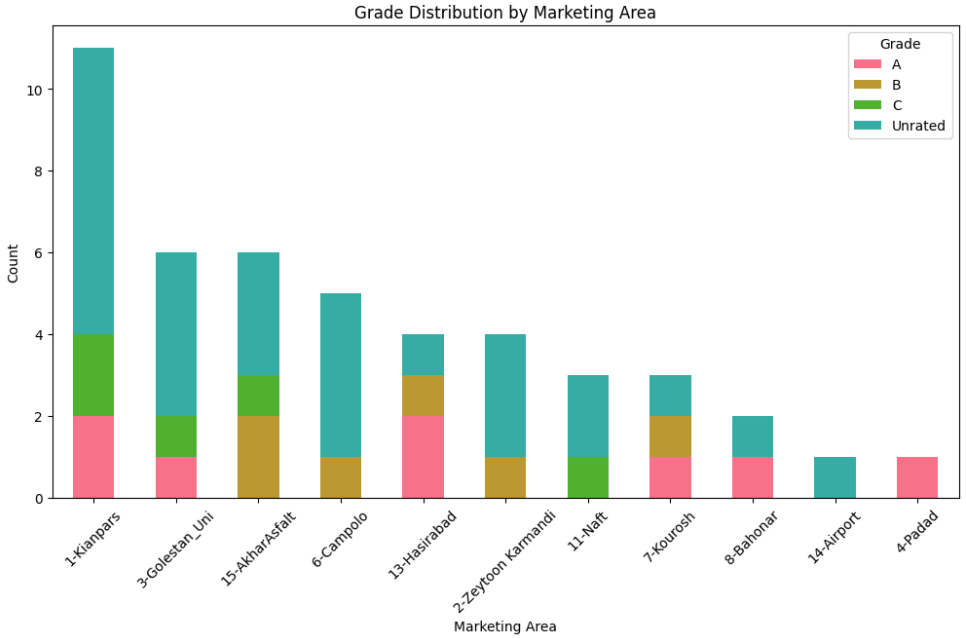

Grade
A
B
C
Unrated
Review
Rate
0
5
10
15
20
0
1
2
3
4
5

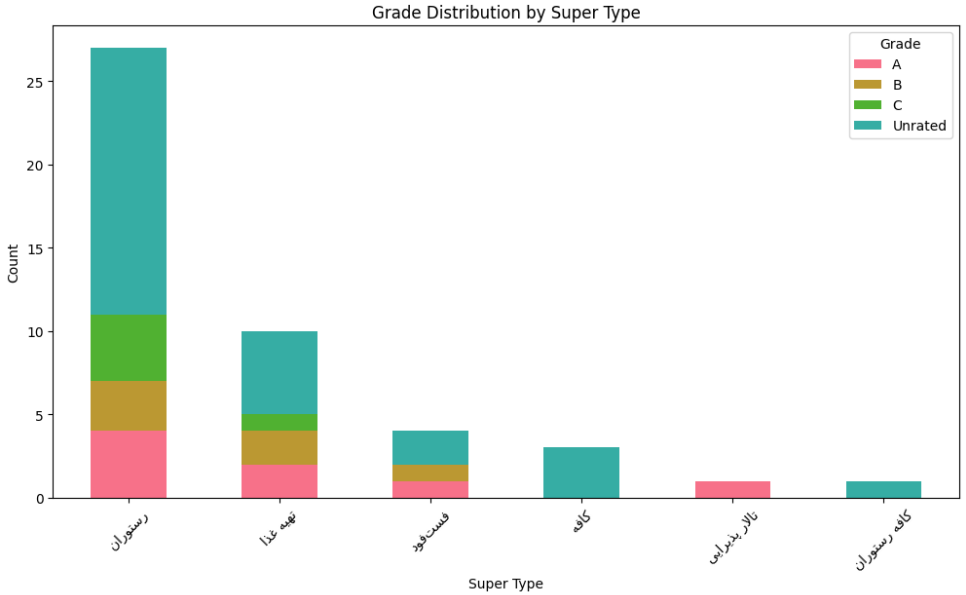
Data Visualization

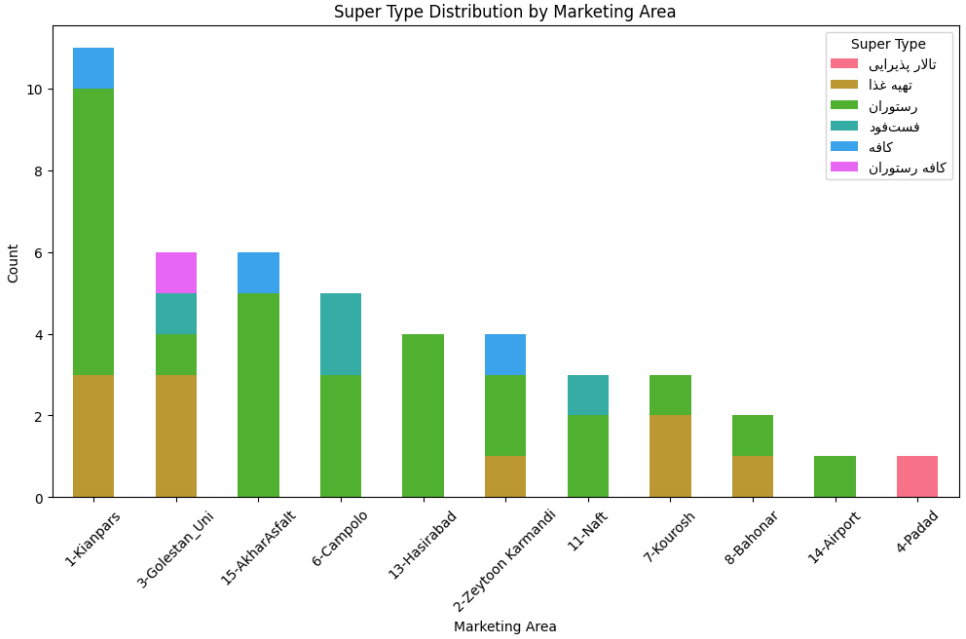
Objective: Present data insights visually for clarity.

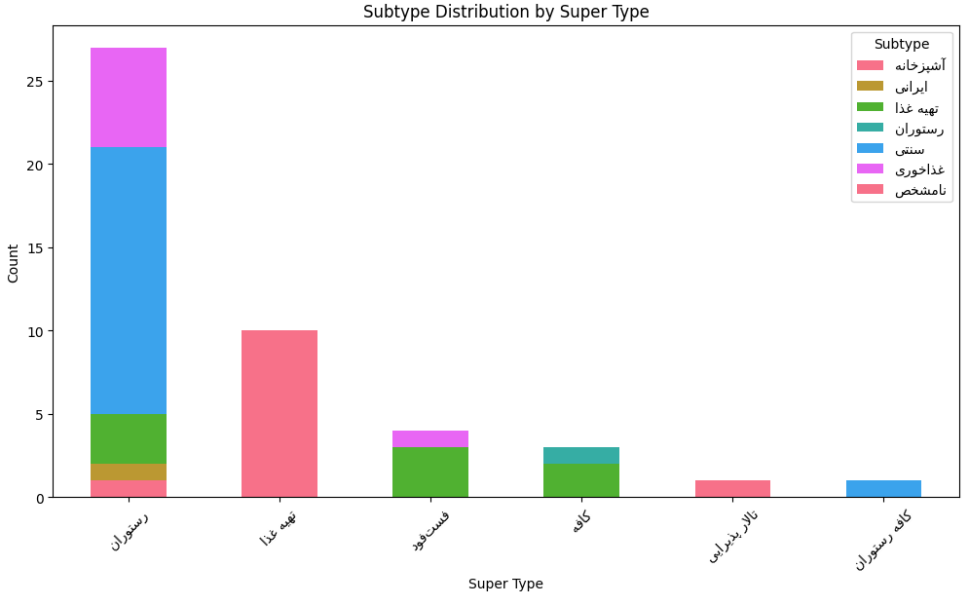

Proportion of Super Type Categories
فست‌فود
کافه
کافه رستوران
تالار پذیرایی
تهیه غذا
رستوران
8.7%
6.5%
2.2%
2.2%
21.7%
58.7%


Proportion of Grade Categories
B
C
A
Unrated
13.0%
10.9%
17.4%
58.7%


Grade Distribution by Marketing Area
Grade
A
B
C
Unrated
Count
Marketing Area
1-Kianpars
3-Golestan_Uni
15-AkharAsfalt
6-Campolo
13-Hasirabad
2-Zeytoon Karmandi
11-Naft
7-Kourosh
8-Bahonar
14-Airport
4-Padad


Grade Distribution by Super Type
Grade
A
B
C
Unrated
Count
Super Type
رستوران
تهیه غذا
فست‌فود
کافه
تالار پذیرایی
کافه رستوران


Super Type Distribution by Marketing Area
Super Type
تالار پذیرایی
تهیه غذا
رستوران
فست‌فود
کافه
کافه رستوران
Count
Marketing Area
1-Kianpars
3-Golestan_Uni
15-AkharAsfalt
6-Campolo
13-Hasirabad
2-Zeytoon Karmandi
11-Naft
7-Kourosh
8-Bahonar
14-Airport
4-Padad


Subtype Distribution by Super Type
Subtype
آشپزخانه
ایرانی
تهیه غذا
رستوران
سنتی
غذاخوری
نامشخص
Count
Super Type
رستوران
تهیه غذا
فست‌فود
کافه
تالار پذیرایی
کافه رستوران

Data Analysis Insights

Data Analysis Insights

- **Marketing Area Concentration:** Over 23% of entries are concentrated in Kianpars, followed by Golestan_Uni and AkharAsfalt (each ~13%), indicating geographic focus in a few key regions.
- **Grade Distribution:** A majority (58.7%) of businesses are unrated, with only 17.4% receiving an A grade, suggesting a need for more evaluations.
- **Super Type Breakdown:** Restaurants dominate the market at 58.7%, followed by Catering Services at 21.7%, showing a strong preference for dining establishments.
- **Subtype Insights:** The most common subtype is Traditional (37%), while Unspecified entries remain high at 23.9%, pointing to potential improvements in the subtype assignment approach.
- **Diversity of Offerings:** Less common categories like Cafes and Reception Halls represent small shares, possibly indicating niche opportunities or underrepresentation.

Thank You