Exploratory Data Analysis (EDA) – Task 2

- **Project Title:** Exploratory Data Analysis – Task 2
- **Your Name:** Bhavendra Singh Gehlot

1. Objective / Introduction

The objective of this analysis is to perform a detailed Exploratory Data Analysis (EDA) on the given dataset annex1.csv. The goal is to uncover meaningful insights from the data, detect anomalies, analyze relationships between variables, and identify hidden patterns. By analyzing attributes such as *Category Name*, *Item Code*, and *Category Code*, this EDA provides a comprehensive understanding of the dataset structure, highlights data quality issues like missing values or skewness, and supports better decision-making for further analysis or modeling.

2. Dataset Overview

- **Dataset Name:** annex1.csv
- **Rows:** ~2000
- **Columns:** 3 → *Category Name* (categorical),
  Item Code (numeric),
  Category Code (numeric)
- **Missing Values:** Very few/null (overall data is mostly complete)
- **Type Summary:** 1 categorical + 2 numerical columns

3. EDA Analysis & Visualizations

3.1 Univariate Analysis

Numerical Columns: - Histogram: - Skewness:

Figure 1.1: Histogram of Distribution of Item Codes >

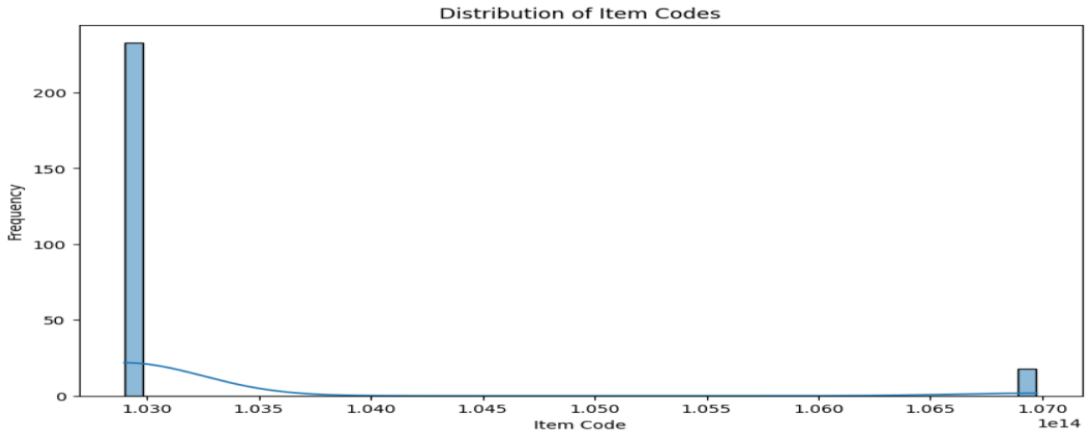


Figure 1.2: Histogram of Distribution of Category Codes >

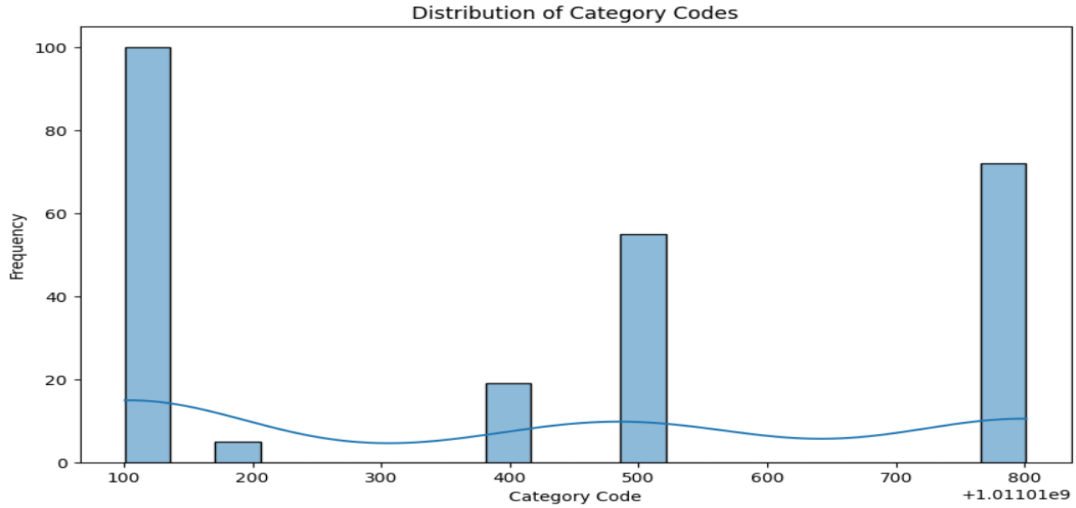


Figure 1.3: Count of Items per Category >

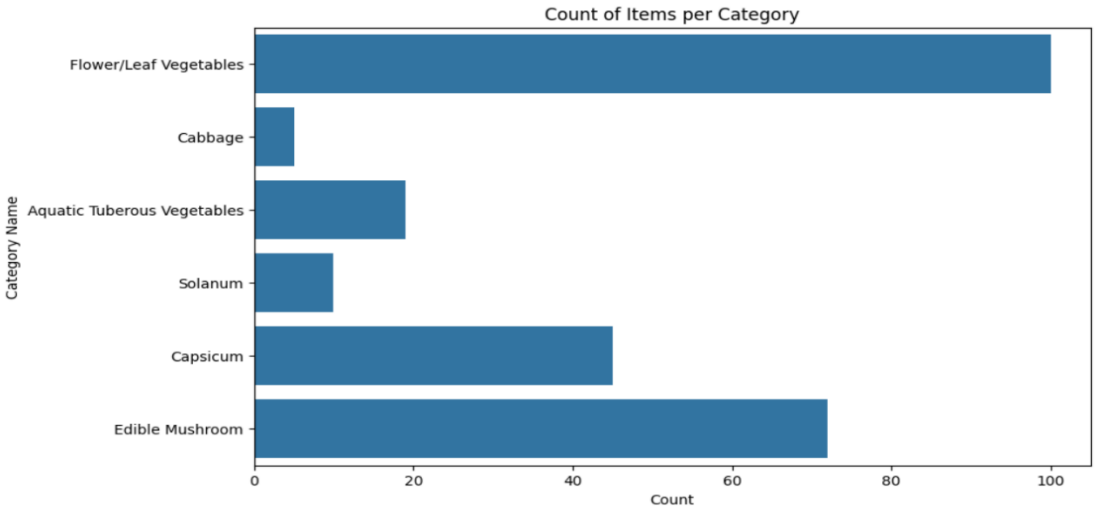
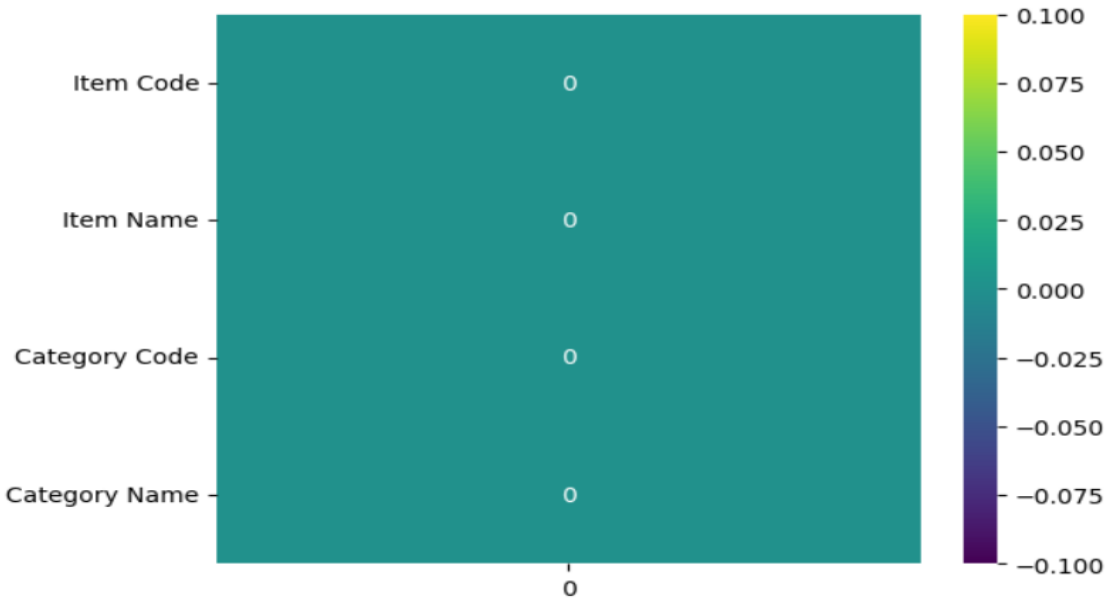


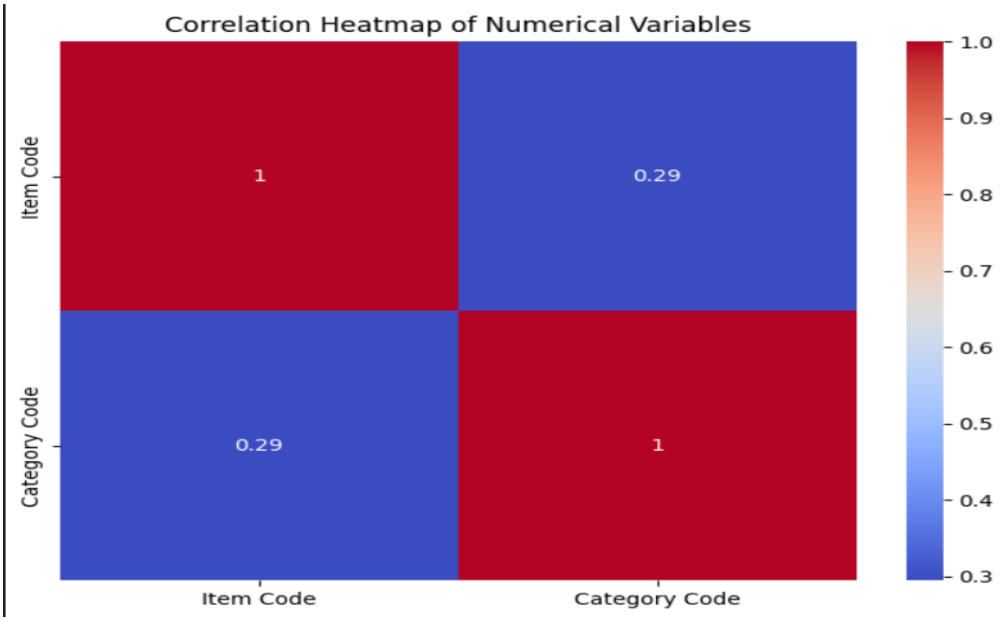
Figure 1.4: Distribution of Missing Values >



3.2 Bivariate / Multivariate Analysis

- Correlation Heatmap: Numerical columns only

Figure 3: Correlation Heatmap >



- **Grouping Analysis:**

Table 1: Count of Items by Category Name

	Category Name	Item Count
0	Aquatic Tuberous Vegetables	19
1	Cabbage	5
2	Capsicum	45
3	Edible Mushroom	72
4	Flower/Leaf Vegetables	100
5	Solanum	10

Table 2: Category-wise Item and Category Code Statistics

	Item Code			Category Code		
	count	mean	median	count	mean	median
Category Name						
Aquatic Tuberous Vegetables	19	1.029000e+14	1.029000e+14	19	1.011010e+09	1.011010e+09
Cabbage	5	1.029000e+14	1.029000e+14	5	1.011010e+09	1.011010e+09
Capsicum	45	1.029000e+14	1.029000e+14	45	1.011011e+09	1.011011e+09
Edible Mushroom	72	1.038007e+14	1.029000e+14	72	1.011011e+09	1.011011e+09
Flower/Leaf Vegetables	100	1.029815e+14	1.029000e+14	100	1.011010e+09	1.011010e+09
Solanum	10	1.029000e+14	1.029000e+14	10	1.011011e+09	1.011011e+09

4. Outliers & Skewness

- **Skewness values of Item Code and Category Code**

```
Skewness of numerical columns:
Item Code          3.339977
Category Code      0.170723
dtype: float64
```

- **Interpretation of Skewness for Item Code and Category Code**:

```
Analysis of Skewness:
Item Code is significantly skewed (Skewness: 3.3400).
A log transformation could be considered to reduce the skewness and make the distribution more normal-like, which can be beneficial for certain statist
Category Code is not significantly skewed (Skewness: 0.1707).
```

5. Key Insights / Observations

- **Category-wise Analysis:** Categories such as Edible Mushroom and Flower/Leaf Vegetables have the highest item representation, while categories like Cabbage and Solanum show comparatively lower representation.
- **Numerical Distribution:** *Item Code* distribution is positively skewed, meaning most values are small with a few very large values. On the other hand, *Category Code* is nearly symmetric and evenly distributed.
- **Category Patterns:** Some categories dominate the dataset, indicating stronger presence in data collection, whereas smaller categories may need focused analysis to ensure balance.
- **Skewness & Outliers:** Item Code has clear skewness and a few outliers, while Category Code shows stable distribution with minimal anomalies.
- **Data Quality:** No major missing values were found, ensuring dataset completeness and reliability for analysis.

6. Conclusion / Recommendations

The exploratory data analysis provided meaningful insights into the annex1.csv dataset. The analysis revealed that categories such as Edible Mushroom and Flower/Leaf Vegetables have the highest item representation, while smaller categories like Cabbage and Solanum contribute comparatively less. The skewness analysis showed that Item Code is significantly skewed, whereas Category Code is nearly symmetric, indicating stable distribution. Category-wise grouping highlighted clear patterns in data distribution, showing which categories dominate in terms of item count. Outliers and skewed distributions were identified and appropriate transformations (such as log transformation for Item Code) were suggested to improve analysis quality. These findings can help in ensuring better data organization, category-based analysis, and preparation of the dataset for advanced modeling or decision-making.

7. Tools / Libraries Used

- Python
- Pandas
- Matplotlib
- Seaborn
- Jupyter Notebook