Name:Shelomith Anyango

Email:shelomith42@gmail.com

Country:Kenya

College:Kibabii University

Specialization: Data Science

Data Cleaning Report for Cab Dataset

1. Introduction

This report outlines the data cleaning steps performed on the Cab dataset sourced from DataGlacier. The primary goal was to handle missing values and outliers in key numerical columns to prepare the data for further analysis.

2. Dataset Overview

Source: Cab_Data.csv

Key Columns Analyzed:

- **Price Charged**
- **Cost of Trip**
- **Company**

3. Data Cleaning Steps

3.1 Missing Value Treatment

Initial Checks: Missing values were found in both Price Charged and Cost of Trip columns.

a)Imputation Strategy:

Applied group-wise mean imputation based on the Company column. This ensured missing values were replaced with the mean of the respective company, preserving group-level integrity.

```
df['Price Charged'] = df.groupby('Company')['Price
Charged'].transform(lambda x: x.fillna(x.mean()))
```

```
df['Cost of Trip'] = df.groupby('Company')['Cost of
Trip'].transform(lambda x: x.fillna(x.mean()))
```

3.2 Outlier Detection and Removal

Boxplots were used to visualize outliers in Price Charged and Cost of Trip.

Interquartile Range (IQR) method was applied to remove outliers:

```
Q1 = column.quantile(0.25)
Q3 = column.quantile(0.75)
IQR = Q3 - Q1
Lower Bound = Q1 - 1.5 * IQR
Upper Bound = Q3 + 1.5 * IQR
```

Rows outside these bounds were filtered out for both numeric columns.

3.3 Final Check

All missing values were successfully handled.

Outliers were removed.

Dataset shape reduced slightly, indicating successful data cleaning without affecting core integrity.

4. Final Dataset Summary

a) **Shape after cleaning: (353,434 rows × 7 columns)**
b) **Missing values: None**
c) **Columns cleaned: Price Charged, Cost of Trip**

5. Conclusion

The dataset is now clean, consistent, and ready for analysis. Company-based imputation and robust outlier detection ensured that the data retained its structure while improving reliability.