report

March 13, 2025

1 Assignment 1 : DAI 101

1.1 Report on Exploratory Data Analysis

- **Name:** Kartik Goyal
- **Enrolment Number:** 23114046
- **Course:** Data Science (DAI 101)
- **Instructor:** Mrs. Shalini Priya
- **Institution:** IIT Roorkee
- **Date:** 13.03.2025

1.1.1 Objective of the Report

The objective of this report is to conduct a thorough **Exploratory Data Analysis (EDA)** on the given dataset. The dataset consists of numerical and categorical variables related to order transactions, including order prices, customer satisfaction, delivery charges, and warehouse locations. The analysis will focus on **data cleaning, univariate and bivariate analysis, outlier detection, and multivariate analysis** to extract meaningful insights.

1.2 Difference Between `dirty_data.csv` and `missing_data.csv`

1.2.1 Dataset Structure

Both datasets have the same columns and data types but differ in data quality: - `dirty_data.csv` focuses on incorrect/messy data (potential typos, outliers). - `missing_data.csv` mainly contains missing values that need to be handled.

1.3 Data Cleaning

1.3.1 Initial Data Inspection

- The dataset was loaded and inspected for missing values, duplicate records, and inconsistent formatting.
- Initial observations revealed that some categorical fields contained inconsistent text formatting and numerical fields had outliers that needed treatment.

1.3.2 Cleaning Steps Performed:

Handled Missing Values: Removed or imputed missing data in key variables like customer reviews.
Removed Duplicates: Ensured each transaction was unique to prevent bias in analysis.
Standardized Categorical Data: Fixed formatting issues (e.g., capitalizing category names).
Detected and Removed Outliers: Used the IQR method to filter extreme values in numerical fields.

1.3.3 After cleaning we are left with 481 values in our dataset (out of 500) !

1.4 Exploratory Data Analysis (EDA)

1.4.1 Univariate Analysis

Univariate analysis was performed on both categorical and numerical variables to understand their distributions.

Key Findings:

- **Order Price & Order Total:** Right-skewed distributions, suggesting the presence of high-value transactions.
- **Delivery Charges:** Some warehouses showed higher-than-average delivery costs.
- **Seasonal Trends:** Transitionery seasons like Spring and Autumn had more sales than extreme seasons like Summer and Winter
- **Warehouse Trends** Nickolson and Thompson warehouse had roughly equal and significantly higher sales than Bakers

Visualizations Used:

- **Histogram & KDE Plot:** Analyzed the distribution of `order_price` and `order_total`.
- **Bar Chart:** Showed the distribution of orders across seasons.
- **Pie Chart:** Showed the distribution of orders across warehouses.

1.4.2 Bivariate Analysis

Scatter Plot: Order Price vs Order Total (Outliers Removed)

Observations:

- **A strong positive correlation was observed**—higher order prices led to higher order totals.
- **Outlier removal helped in revealing the actual pattern**, which was previously hidden due to extreme values.

Regression Plot: Distance to Warehouse vs Delivery Charges

Observations:

- **Longer distances were generally associated with higher delivery charges**, but the relationship was **not perfectly linear**.
- **Some points deviated from the trend**, suggesting that factors other than distance also influence delivery costs.

Box Plot: Expedited Delivery vs Order Price (Box Plot)

Observations:

- **Expedited orders tend to have a higher median order price**.

Order Price Distribution Across Seasons (Violin Plot)

Observations:

- **Order price distribution differs across seasons**, suggesting seasonal trends in purchasing behavior.
- Some seasons show **higher median order prices**, possibly indicating high-demand periods.

Order Price Distribution vs Customer Satisfaction

Observations:

- Happy customers tend to have **higher order prices**, indicating that better service could be linked to premium purchases.

Coupon Discount Distribution vs Customer Satisfaction

Observations:

- **Satisfied customers tend to receive higher average discounts**, possibly as a retention strategy.
- **Discount distribution for unhappy customers is more spread out**, indicating inconsistent promotional strategies.

Line Plot: Delivery Charges vs Warehouses

Observations:

- Some warehouses consistently **charge higher delivery fees**, which might indicate regional pricing differences.
- Fluctuations in delivery charges suggest **warehouse-specific factors influencing costs**.

Stacked Column Chart: Warehouse vs Customer Satisfaction

Observations:

- Certain warehouses have a **higher proportion of happy customers**, indicating better service or faster deliveries.

Heatmap: Average Order Total by Warehouse & Season

Observations:

- **Order total varies significantly across warehouses and seasons**, suggesting seasonality in demand.
- **Some warehouses perform consistently well across seasons**, while others see fluctuations.
- **Peak order totals appear in specific seasons**, highlighting potential seasonal trends in customer behavior.

1.4.3 Multivariate Analysis

Pair Plot for Key Numerical Variables

Observations:

- **Order price and order total are strongly correlated**, with a clear upward trend in scatter plots.
- **Delivery charges show significant spread**, indicating varying policies across warehouses.
- **Distance-to-warehouse vs delivery charge distribution highlights exceptions**, suggesting additional pricing factors.

Cluster Heatmap: Warehouse & Seasonal Order Trends

Observations:

- **Some warehouses consistently have high order totals year-round**, while others are highly seasonal.
- **Warehouse-specific trends exist**, which could be linked to regional demand patterns.

1.5 Conclusion

1.5.1 Key Insights from the Analysis:

Data Cleaning Enhanced Data Quality:
- Removing duplicates, handling missing values, and correcting categorical inconsistencies ensured accurate analysis.

Order Price & Total Show a Strong Positive Correlation:
- Expensive orders result in higher order totals, validating expected business trends.

Warehouse & Delivery Charges Impact Customer Satisfaction:
- Warehouses that charged **higher delivery fees had lower satisfaction ratings**.

Seasonality Affects Order Totals and Delivery Charges:
- **Peak seasons show increased order values and shipping costs**, which could influence logistics planning.