EXPLORATORY DATA ANALYSIS USING PYTHON

Summary and Recommendations

📂 Dataset Overview

The dataset used in this analysis comprises **1,284 records**, with each row representing data related to the usage of wellness centers across different cities. The data is clean, with **no missing values**. The primary columns of interest include:

- **cityName**: The city where the wellness center is located.
- **card_type**: The type of user accessing the center (e.g., Pensioner, Employee).
- **centre_name**: The name of the wellness center.
- **count**: Represents the number of times a center was accessed.

🔍 Initial Data Exploration

The initial exploration of the dataset was focused on understanding the data types, distributions, and basic statistics. Key findings include:

- The **count** variable has a wide range and shows potential outliers.
- The **card_type** field has multiple categories, with **Pensioner** being the most common.
- Several cities such as **Ahmedabad**, **Mumbai**, and **Delhi** dominate in terms of total usage count.

📊 Visual Analysis and Insights

A variety of plots and charts were used to uncover hidden trends and distributions within the data:

1. Histogram of Count

- Revealed a **right-skewed distribution** of counts, indicating that while many centers have low activity, a few have significantly high usage.

2. Boxplot and Violin Plot

- **Boxplot** exposed outliers with extremely high usage counts.
- **Violin plot** provided a visual understanding of count distribution across different **card types**.

3. Bar Plot - Total Count per City

- Showed that **Ahmedabad**, **Mumbai**, and **Delhi** lead in terms of total user visits.

- Other cities had significantly lower counts, showing disparity in service usage.

4. Top 10 Wellness Centres

- Identified wellness centers with the highest usage, allowing focus for resource allocation or case-specific study.

5. Heatmap - City vs Card Type

- Highlighted the relationship between different card types and their activity across cities.
- Some card types are more active in specific cities.

6. Scatter Plots

- Visualized individual centers against their count to highlight concentration and spread.
- A jittered version made overlapping points more visible.

7. Pie Chart - Top 6 Cities

- Illustrated the proportional share of total visits by the most active cities.

8. KDE Plot with Log Scale

- Smoothed density plot showing that the count variable is heavily skewed.
- The log scale allowed better visibility into low-count areas.

9. Count Plot for Card Types

- Reinforced that **Pensioners** dominate usage.
- Helped quantify the distribution of different user categories.

📈 Grouped Analysis

Two key groupings provided further insights:

Grouped by City and Card Type:

- Enabled a granular view of how different user types use services in each city.
- Useful for targeting interventions or campaigns.

Top Centre per Card Type:

- Identified which centers are most frequently used by each user category.
- Can inform improvements or promotions tailored to specific groups.

📝 Final Summary and Recommendations

- The data is **clean and structured**, ready for deeper statistical or predictive analysis.
- **Pensioners** are the most frequent users, suggesting targeted services for this demographic.
- **Ahmedabad, Mumbai, and Delhi** are the highest traffic cities, and should be prioritized for quality improvements.
- Some centers are **outliers** in terms of usage and warrant individual attention.
- The overall **count distribution is right-skewed**, indicating a need to support underutilized centers.

Recommendations:

1. Investigate outlier centers to understand the reasons behind high traffic.
2. Focus on expanding successful center models to lower-traffic cities.
3. Develop city-wise and card_type-wise marketing or engagement strategies.
4. Use this data to inform future infrastructure investments and policy decisions.
