Dataset Analysis of an Online Retailer Based in Brazil Using Python

Objective:

The goal of this analysis is to evaluate the distribution of delivery times for an online retailer based in Brazil. By examining key statistical metrics such as mean, median, mode, and standard deviation, we aim to understand whether the distribution is positively or negatively skewed. This analysis will provide insights into the delivery efficiency and customer satisfaction levels, potentially guiding logistics improvements.

Summary:

In this analysis, we focused on the distribution of delivery days for customer orders. Using descriptive statistics and visualization techniques, we explored the nature of delivery time distribution and identified any skewness present in the data. Understanding skewness is vital in interpreting the delivery efficiency and customer experience, as it reveals potential delays or inconsistencies.

Data Collection:

1. **Data Source:**
    - Order Data: Contains various timestamps, including purchase and delivery dates.
2. **Data Import:**
    - Used Pandas to import CSV data directly from GitHub.
3. **Data Preprocessing:**
    - Calculated the difference between order_delivered_customer_date and order_purchase_timestamp to derive delivery days.
    - Converted timedelta objects into integer days for analysis.

Analysis:

Descriptive Statistics:

- **Mean Delivery Days:**
    - Calculated the average delivery time, representing the central tendency of delivery days.
- **Median Delivery Days:**
    - Determined the middle value, indicating the typical delivery experience without the influence of extreme values.
- **Mode Delivery Days:**
    - Identified the most frequently occurring delivery time, highlighting common patterns.
- **Standard Deviation:**
    - Assessed the spread of delivery times, reflecting variability in delivery experiences.

- **Skewness Analysis:**
    - Analyzed the skewness of the delivery days distribution to determine whether it is positively or negatively skewed.

Visualization:

- **Bar Plot:**
    - Created a bar plot to visualize the frequency of delivery days, highlighting the most common delivery timeframes.
- **Distribution Insights:**
    - Examined the distribution shape to identify skewness, focusing on the tail behavior.

Results:

Delivery Days Statistics:

- **Mean Delivery Days:** 12.2
- **Median Delivery Days:** 10.0
- **Mode Delivery Days:** 10
- **Standard Deviation:** 10.8

Skewness:

- **Skewness Analysis:**
    - The mean delivery time is greater than the median, and the distribution has a long right tail.
    - This indicates a **positively skewed distribution**, where a few orders take significantly longer to deliver than the majority.

Visualization Insights:

- **Bar Plot:**
    - The bar plot reveals that most deliveries occur within 10 days, with decreasing frequency for longer delivery times.
    - The presence of longer delivery times stretches the distribution to the right, confirming positive skewness.

Conclusion:

1. **Key Findings:**
    - The distribution of delivery days is positively skewed, indicating that most deliveries are completed in a reasonable time frame, but a few deliveries experience substantial delays.
    - The mean delivery time (12.2 days) is skewed by longer delivery times, while the median (10 days) provides a more typical representation of delivery performance.

Code and Analysis :

```
#Reading Data from Github

import statistics


import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats

import numpy as np

#Read the csv file

order_df =
pd.read_csv('https://raw.githubusercontent.com/swapnilsaurav/OnlineRetail/master/orders.csv')

#Display all the column names

print(list(order_df.columns))

X=pd.to_datetime(order_df['order_delivered_customer_date']) -
pd.to_datetime(order_df['order_purchase_timestamp'])


for i in range(0,len(X)):

   X[i]=X[i].days

plt.figure(figsize=(10,5))


sns.barplot(x=X.value_counts().sort_values(ascending=False).head(30).index,y=X.value_counts().sort
_values(ascending=False).head(30).values)

plt.xlabel('Delivery Days')

plt.ylabel('Frequency')

plt.show()

info=X.describe()


print("Mean Value of Delivery Days: {:0.1f}".format(np.mean(X)))

print("Median Value of Delivery Days: " , np.median(X))

print("Mode Value of Delivery Days: " , statistics.mode (X))

print("Standard Deviation of Delivery Days: {:0.1f}".format(X.std()))
```