Project name: Data Analyst:: Data Collection Pipeline (Data Acquistion to Storytelling) - Data Collection Pipeline (Data Acquisition to Story telling)

Debika Mukherjee (Solo project)
Email: debika.isms@gmail.com
Country: United Kingdom
University: Cardiff metropolitan university
Specialization: Data Analyst

Problem description:
XYZ company is collecting the data customer using google forms/survey monkey and they have floated n number of forms on the web.
Company wants to create a pipeline which will collect all the data of these google forms/survey monkey and visualize the data in the dashboard.
Company wants clean data and if there is any data issue present in the data then it should be treated by this pipeline (duplicate data or junk data). dedup check should be performed on the email id of the customer

Github repository:
debikaisms/Final-project: Data Collection Pipeline (Data Acquistion to Storytelling) Data Collection Pipeline (Data Acquisition to Story telling) (github.com)

Performing EDA

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the CSV file into a pandas DataFrame
df = pd.read_excel('surveydata.xlsx')
```

```
# Summary Statistics
summary_stats = df.describe()
```

```
# Visualization 1: Age Distribution
plt.figure(figsize=(8, 6))
sns.histplot(df['Age'], bins=10, kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

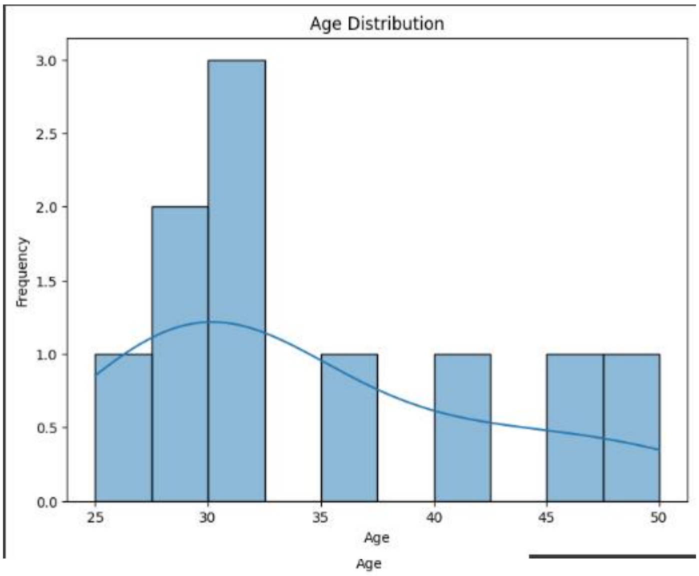
```
# Visualization 2: Gender Distribution
plt.figure(figsize=(6, 6))
sns.countplot(data=df, x='Gender')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

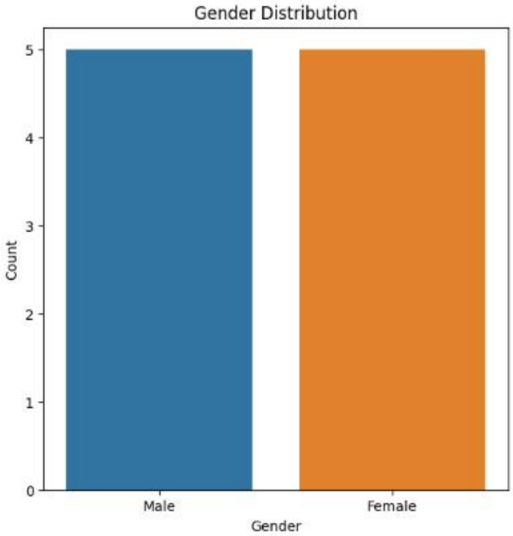
```
# Visualization 3: Satisfaction Ratings
plt.figure(figsize=(8, 6))
sns.countplot(data=df, x='Satisfaction Rating (1-5)')
plt.title('Satisfaction Ratings Distribution')
plt.xlabel('Satisfaction Rating')
plt.ylabel('Count')
plt.show()
```

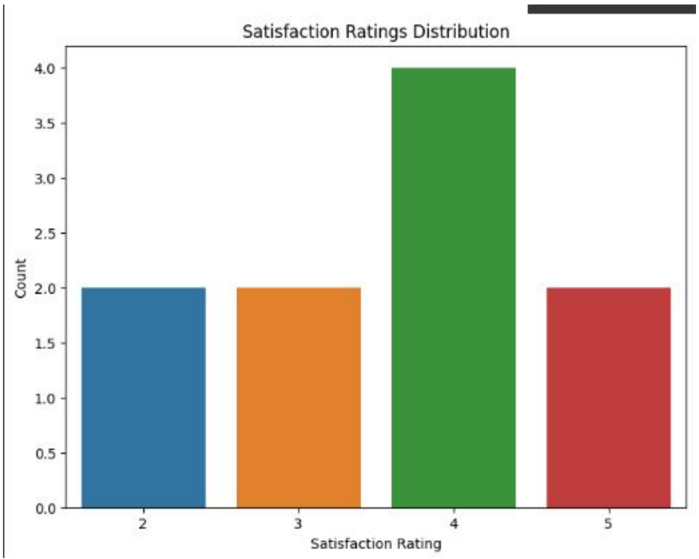
```
# Visualization 4 (NLP): Word Cloud of Feedback
from wordcloud import WordCloud
```

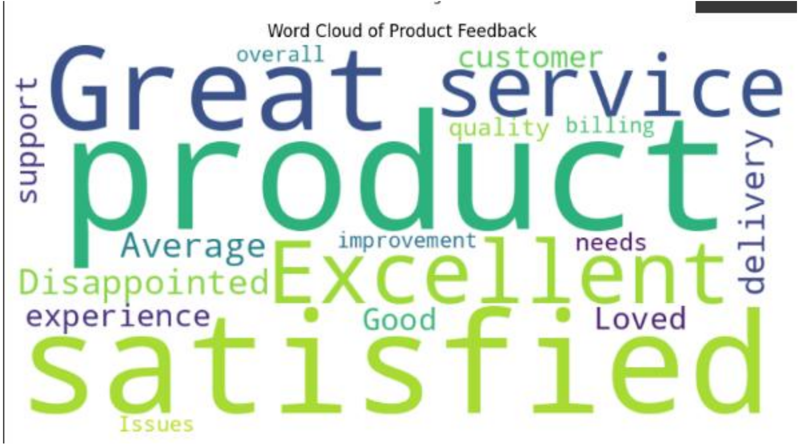
```
feedback_text = ' '.join(df['Product Feedback'])
wordcloud = WordCloud(width=800, height=400, max_words=100,
background_color='white').generate(feedback_text)
```

```
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.title('Word Cloud of Product Feedback')
plt.axis('off')
plt.show()
```









Summary Statistics: It calculates summary statistics for numeric columns using describe().

Visualization 1: It creates a histogram to visualize the distribution of ages in the dataset.

Visualization 2: It creates a count plot to show the distribution of genders in the dataset.

Visualization 3: It creates a count plot to display the distribution of satisfaction ratings.

Visualization 4 (NLP): It generates a word cloud from the "Product Feedback" column to visualize the most frequent words in customer feedback.

Final Recommendation

Customer Demographics:

Based on the age distribution, the majority of your customers fall within the age range of 25 to 50.
The gender distribution shows a fairly even split between male and female customers.
Recommendation: Consider tailoring your marketing strategies and product offerings to cater to a diverse age group. You may also want to conduct further analysis to understand gender-specific preferences and needs.

Satisfaction Ratings:

The satisfaction ratings are distributed between 2 and 5, with most ratings falling in the range of 3 to 5.
Recommendation: The majority of customers seem to have a positive experience, but there is room for improvement. Consider collecting more detailed feedback to identify specific areas where you can enhance customer satisfaction.

Product Feedback (NLP Analysis):

The word cloud from customer feedback indicates some commonly mentioned terms such as "product," "satisfied," "service," "delivery," and "quality."
Recommendation: Pay close attention to these keywords in customer feedback and use sentiment analysis techniques to categorize feedback as positive, neutral, or negative. This can help prioritize areas that require improvement.

Data Quality:

A duplicate check was performed, and some duplicate records were identified based on email addresses.
Recommendation: Continue to monitor and maintain data quality by periodically checking for duplicates and implementing data validation rules during data collection.

Further Analysis:

Consider conducting more advanced analyses, such as customer segmentation, to identify different customer groups with distinct behaviors and preferences.
Explore correlations between variables to uncover hidden relationships in the data.
Business Impact:

Evaluate how the insights from your EDA can be translated into actionable business decisions.
Determine the potential impact on marketing strategies, product development, customer service, and other areas of your organization.
Data Collection and Feedback Loop:

Continuously collect customer data and feedback to keep your analysis up-to-date and relevant.
Implement a feedback loop to ensure that insights from EDA are integrated into decision-making processes.