# Significance Of Exploratory Data Analysis (EDA) for the Sentiment Analysis Project:

**Data Overview**

* **Code**

data.sample(9)

data.tail(9)

* This code displays random samples (9 rows) and the last 9 rows of your dataset. It helps you quickly view the structure of your data and understand its content.

**Checking for Missing Values**

* **Code**

data.isnull()

data.isnull().sum()

* These lines of code check for missing values in your dataset and provide a summary of the count of missing values in each column. It's essential for data cleaning and imputation.

**Data Information**

* **Code**

data.info()

data.dtypes

data.columns

* These code snippets provide an overview of your dataset. `data.info()` gives information about data types, non-null values, and memory usage, while `data.dtypes` shows the data types of each column, and `data.columns` lists the column names.

**Duplicates Removal**

* **Code**

data.duplicated().sum()

data = data.drop\_duplicates()

data.duplicated().sum()

* The code first checks for and reports the number of duplicate rows in the dataset. Then, it removes the duplicates using `drop\_duplicates()`. This ensures your data contains unique records, avoiding data inconsistencies.

**Scatter Plot**

* **Code**

plt.scatter(data.tweet\_id, data.airline\_sentiment\_confidence)

* This code creates a scatter plot between "tweet\_id" and "airline\_sentiment\_confidence." It helps visualize the distribution and confidence of airline sentiments.

**Box Plot**

* **Code**

plt.boxplot(data['airline\_sentiment\_confidence'], vert=False)

* The box plot visualizes the distribution of the "airline\_sentiment\_confidence" column. It shows the median, quartiles, and potential outliers.

**Pie Chart**

* **Code**

plt.pie(sentiment\_counts, labels=sentiment\_counts.index, autopct='%1.1f%%', startangle=140)

* This code creates a pie chart to display the distribution of different airline sentiments. It offers a visual representation of sentiment proportions.

**Histograms**

* **Code**

plt.hist(data['airline\_sentiment\_confidence'], bins=20, edgecolor='k')

plt.hist(data['retweet\_count'], bins=20, edgecolor='k')

* These snippets generate histograms to visualize the distribution of "airline\_sentiment\_confidence" and "retweet\_count." Histograms provide insights into data distribution.

**Bar Charts**

* **Code**

plt.bar(airline\_counts.index, airline\_counts.values)

plt.bar(negativereason\_counts.index, negativereason\_counts.values)

* These code snippets create bar charts to display the distribution of airlines and negative reasons. They help understand which airlines are mentioned more and why customers express negative sentiments.

**Seaborn Count Plot**

* **Code**

sns.countplot(data=data, x='airline\_sentiment')

* This snippet uses Seaborn to create a count plot of sentiment classes. It provides a clear view of the distribution of sentiment classes.

**Word Clouds**

* **Code**

wordcloud = WordCloud(width=800, height=400, background\_color='white').generate(text)

* These code snippets generate word clouds to visualize the most common words in tweets. They help identify frequent keywords in the dataset.

**Tweet Length Distribution**

* **Code**

plt.hist(data['tweet\_length'], bins=20)

* This snippet creates a histogram to visualize the distribution of tweet lengths, helping you understand the length of tweets in your dataset.

**Top Hashtags and Mentions**

* **Code**

top\_hashtags.plot(kind='bar', title="Top Hashtags")

top\_mentions.plot(kind='bar', title="Top Mentions")

* These snippets display bar charts of the top hashtags and mentions in the tweets, allowing you to identify popular topics and user mentions.