SocialMediaDataAnalysis

August 20, 2024

1 Clean & Analyze Social Media

1.1 Introduction

Social media has become a ubiquitous part of modern life, with platforms such as Instagram, Twitter, and Facebook serving as essential communication channels. Social media data sets are vast and complex, making analysis a challenging task for businesses and researchers alike. In this project, we explore a simulated social media, for example Tweets, data set to understand trends in likes across different categories.

1.2 Prerequisites

To follow along with this project, you should have a basic understanding of Python programming and data analysis concepts. In addition, you may want to use the following packages in your Python environment:

- pandas
- Matplotlib
- …

These packages should already be installed in Coursera's Jupyter Notebook environment, however if you'd like to install additional packages that are not included in this environment or are working off platform you can install additional packages using `!pip install packagename` within a notebook cell such as:

- `!pip install pandas`
- `!pip install matplotlib`

1.3 Project Scope

The objective of this project is to analyze tweets (or other social media data) and gain insights into user engagement. We will explore the data set using visualization techniques to understand the distribution of likes across different categories. Finally, we will analyze the data to draw conclusions about the most popular categories and the overall engagement on the platform.

1.4 Step 1: Importing Required Libraries

As the name suggests, the first step is to import all the necessary libraries that will be used in the project. In this case, we need pandas, numpy, matplotlib, seaborn, and random libraries.

Pandas is a library used for data manipulation and analysis. Numpy is a library used for numerical computations. Matplotlib is a library used for data visualization. Seaborn is a library used for statistical data visualization. Random is a library used to generate random numbers.

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import random
```

1.5 Step 2: Generate random data for the social media data

[2]:
```
# Define the list of categories
categories = ['Food', 'Travel', 'Fashion', 'Fitness', 'Music', 'Culture',␣
 ↪'Family', 'Health']

# Define the number of data points
n = 500

# Generate the data dictionary
data = {
    'Date': pd.date_range('2021-01-01', periods=n),
    'Category': [random.choice(categories) for _ in range(n)],
    'Likes': np.random.randint(0, 10000, size=n)
}

# Convert the data dictionary to a pandas DataFrame
df = pd.DataFrame(data)
```

1.6 Step 3: Load the data into a Pandas DataFrame and Explore the data

[3]:
```
# Print the head of the DataFrame (first 5 rows)
print("DataFrame Head:")
print(df.head())

# Print DataFrame information
print("\nDataFrame Info:")
print(df.info())

# Print DataFrame description
```

```python
print("\nDataFrame Description:")
print(df.describe())

# Print the count of each 'Category' element
print("\nCategory Counts:")
print(df['Category'].value_counts())
```

```
DataFrame Head:
        Date Category  Likes
0 2021-01-01  Fashion   1069
1 2021-01-02   Travel   1502
2 2021-01-03  Fitness   2492
3 2021-01-04  Fitness   6251
4 2021-01-05  Culture   2362

DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      500 non-null    datetime64[ns]
 1   Category  500 non-null    object
 2   Likes     500 non-null    int64
dtypes: datetime64[ns](1), int64(1), object(1)
memory usage: 11.8+ KB
None

DataFrame Description:
             Likes
count   500.000000
mean   4809.860000
std    2885.366397
min       2.000000
25%    2366.000000
50%    4601.500000
75%    7367.750000
max    9897.000000

Category Counts:
Fitness    85
Food       71
Music      65
Fashion    63
Travel     62
Health     54
Family     50
```

```
Culture    50
Name: Category, dtype: int64
```

1.7 Step 4: Clean the data

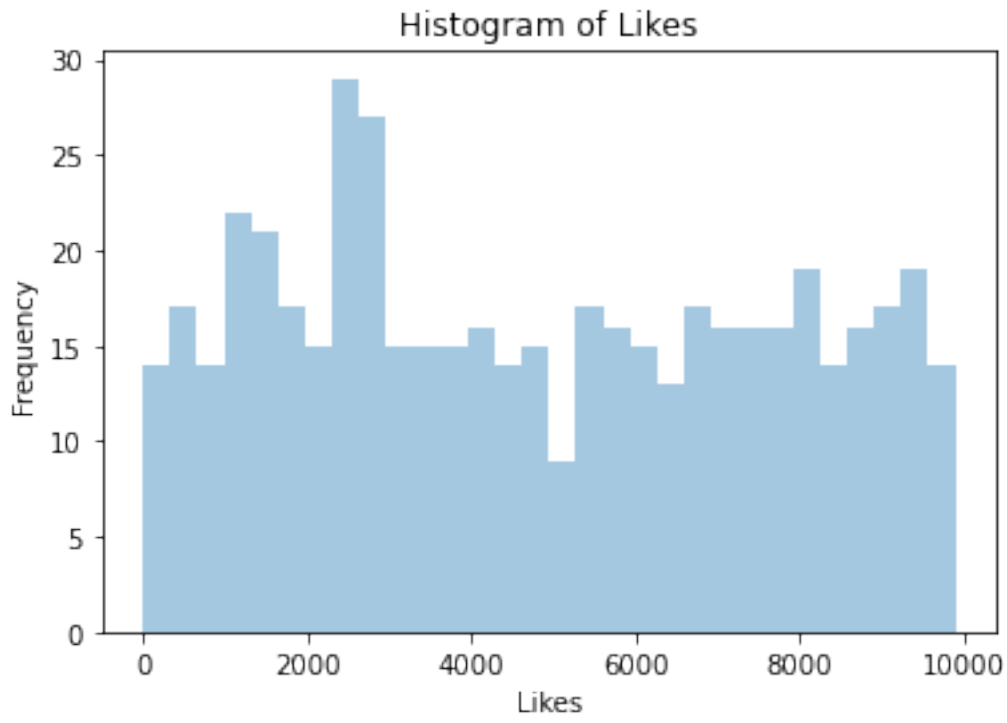
[4]:
```
# Step 1: Remove null data
df = df.dropna()

# Step 2: Remove duplicate data
df = df.drop_duplicates()

# Step 3: Convert 'Date' field to datetime format
df['Date'] = pd.to_datetime(df['Date'])

# Step 4: Convert 'Likes' field to integer
df['Likes'] = df['Likes'].astype(int)
```

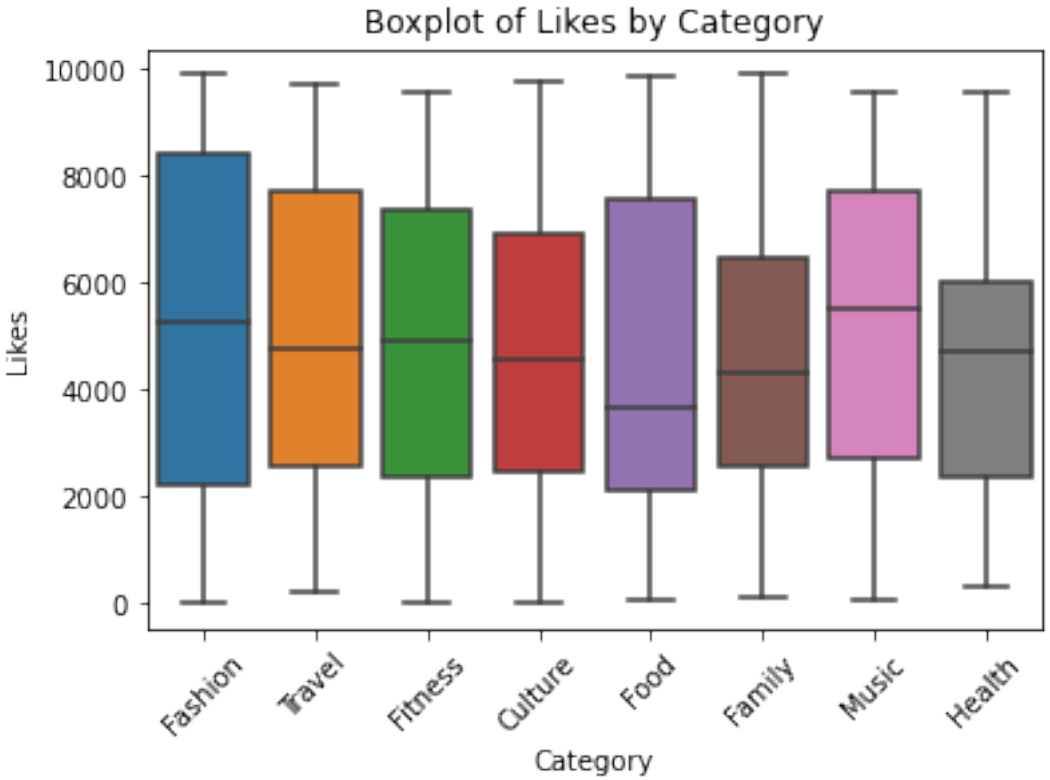
[5]:
```
# Display the cleaned DataFrame
print("\nCleaned DataFrame Head:")
print(df.head())

# Verify the data types to ensure the conversions were successful
print("\nDataFrame Info After Cleaning:")
print(df.info())
```

```
Cleaned DataFrame Head:
        Date Category  Likes
0 2021-01-01  Fashion   1069
1 2021-01-02   Travel   1502
2 2021-01-03  Fitness   2492
3 2021-01-04  Fitness   6251
4 2021-01-05  Culture   2362

DataFrame Info After Cleaning:
<class 'pandas.core.frame.DataFrame'>
Int64Index: 500 entries, 0 to 499
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      500 non-null    datetime64[ns]
 1   Category  500 non-null    object
 2   Likes     500 non-null    int64
dtypes: datetime64[ns](1), int64(1), object(1)
memory usage: 15.6+ KB
None
```

1.8 Step 4: Visualize and Analyze the data

[6]:
```
# Step 1: Create a histogram of the 'Likes'
sns.distplot(df['Likes'], bins=30, kde=False)
plt.title('Histogram of Likes')
plt.xlabel('Likes')
plt.ylabel('Frequency')
plt.show()

# Step 2: Create a boxplot with 'Category' on x-axis and 'Likes' on y-axis
sns.boxplot(x='Category', y='Likes', data=df)
plt.title('Boxplot of Likes by Category')
plt.xlabel('Category')
plt.ylabel('Likes')
plt.xticks(rotation=45)
plt.show()

# Analyze the Data

# Step 1: Print the mean of the 'Likes' category
mean_likes = df['Likes'].mean()
print(f"Mean of Likes: {mean_likes}")

# Step 2: Print the mean of 'Likes' for each 'Category'
mean_likes_by_category = df.groupby('Category')['Likes'].mean()
print("\nMean of Likes by Category:")
print(mean_likes_by_category)
```


Histogram of Likes
Frequency
Likes
0
5
10
15
20
25
30
0
2000
4000
6000
8000
10000


Boxplot of Likes by Category
Likes
Category
0
2000
4000
6000
8000
10000
Fashion
Travel
Fitness
Culture
Food
Family
Music
Health

```
Mean of Likes: 4809.86

Mean of Likes by Category:
Category
Culture    4720.140000
Family     4453.060000
Fashion    5153.301587
Fitness    4839.305882
Food       4726.478873
Health     4426.777778
Music      5093.492308
Travel     4912.387097
Name: Likes, dtype: float64
```

1.9 Conclusion

In this project, I analyzed social media data using Python, Pandas, Seaborn, and Matplotlib. The process involved generating synthetic data, cleaning it, visualizing key aspects, and drawing statistical conclusions.

1.9.1 Process Overview:

- **Data Generation:** I created realistic social media data with categories like Food, Travel, and Music, using tools like `pandas.date_range` and `numpy.random.randint`.
- **Data Cleaning:** I removed nulls and duplicates, converted data types, and ensured data consistency, preparing it for analysis.
- **Visualization and Analysis:** I used Seaborn to create a histogram and boxplot, which revealed the distribution and variability of 'Likes' across different categories. I calculated the mean 'Likes' overall and by category.

1.9.2 Key Findings:

- Most posts receive a moderate number of likes.
- Significant variation in 'Likes' was observed across categories.

1.9.3 Challenges and Solutions:

Ensuring realistic data generation and mastering Seaborn functions were challenging. I overcame these by researching best practices and experimenting with different approaches.

1.9.4 Future Improvements:

Incorporating more complex data generation methods and machine learning models could enhance the project.