SocialMediaDataAnalysis

August 12, 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.

[2]:
```
# your code here
```

[1]:
```
!pip install pandas
!pip install matplotlib
!pip install numpy
!pip install seaborn
!pip install random
```

```
Requirement already satisfied: pandas in /opt/conda/lib/python3.7/site-packages
(1.0.3)
Requirement already satisfied: pytz>=2017.2 in /opt/conda/lib/python3.7/site-
packages (from pandas) (2020.1)
Requirement already satisfied: numpy>=1.13.3 in /opt/conda/lib/python3.7/site-
packages (from pandas) (1.18.4)
Requirement already satisfied: python-dateutil>=2.6.1 in
/opt/conda/lib/python3.7/site-packages (from pandas) (2.8.1)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.7/site-
packages (from python-dateutil>=2.6.1->pandas) (1.14.0)
WARNING: You are using pip version 21.3.1; however, version 24.0 is

available.

You should consider upgrading via the '/opt/conda/bin/python3 -m pip install

--upgrade pip' command.
Requirement already satisfied: matplotlib in /opt/conda/lib/python3.7/site-
packages (3.2.1)
Requirement already satisfied: kiwisolver>=1.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib) (1.2.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib) (2.4.7)
Requirement already satisfied: python-dateutil>=2.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib) (2.8.1)
Requirement already satisfied: numpy>=1.11 in /opt/conda/lib/python3.7/site-
packages (from matplotlib) (1.18.4)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.7/site-
packages (from matplotlib) (0.10.0)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from cycler>=0.10->matplotlib) (1.14.0)
```

```
WARNING: You are using pip version 21.3.1; however, version 24.0 is
available.
You should consider upgrading via the '/opt/conda/bin/python3 -m pip install
--upgrade pip' command.
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages
(1.18.4)
WARNING: You are using pip version 21.3.1; however, version 24.0 is
available.
You should consider upgrading via the '/opt/conda/bin/python3 -m pip install
--upgrade pip' command.
Requirement already satisfied: seaborn in /opt/conda/lib/python3.7/site-packages
(0.10.1)
Requirement already satisfied: scipy>=1.0.1 in /opt/conda/lib/python3.7/site-
packages (from seaborn) (1.4.1)
Requirement already satisfied: pandas>=0.22.0 in /opt/conda/lib/python3.7/site-
packages (from seaborn) (1.0.3)
Requirement already satisfied: numpy>=1.13.3 in /opt/conda/lib/python3.7/site-
packages (from seaborn) (1.18.4)
Requirement already satisfied: matplotlib>=2.1.2 in
/opt/conda/lib/python3.7/site-packages (from seaborn) (3.2.1)
Requirement already satisfied: python-dateutil>=2.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=2.1.2->seaborn) (2.8.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=2.1.2->seaborn) (2.4.7)
Requirement already satisfied: kiwisolver>=1.0.1 in
/opt/conda/lib/python3.7/site-packages (from matplotlib>=2.1.2->seaborn) (1.2.0)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.7/site-
packages (from matplotlib>=2.1.2->seaborn) (0.10.0)
Requirement already satisfied: pytz>=2017.2 in /opt/conda/lib/python3.7/site-
packages (from pandas>=0.22.0->seaborn) (2020.1)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages
(from cycler>=0.10->matplotlib>=2.1.2->seaborn) (1.14.0)
WARNING: You are using pip version 21.3.1; however, version 24.0 is
available.
You should consider upgrading via the '/opt/conda/bin/python3 -m pip install
--upgrade pip' command.
ERROR: Could not find a version that satisfies the requirement random (from
versions: none)
ERROR: No matching distribution found for random
```

```
WARNING: You are using pip version 21.3.1; however, version 24.0 is
available.
You should consider upgrading via the '/opt/conda/bin/python3 -m pip install
--upgrade pip' command.
```

[11]:
```
import pandas as pd
import seaborn as sns

import numpy as np
import matplotlib as mpl

import matplotlib.pyplot as plt
import random as rnd
plt.show()
```

[ ]:
```
import pandas as pd
import random
import numpy as np
categories=['Food','Travel','Fashion','Fitness','Music','Culture','Family','Health']
n=500
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)
}

df = pd.DataFrame(data)
print(df.head())
```

[3]:
```
import pandas as pd
import random
import numpy as np
categories=['Food','Travel','Fashion','Fitness','Music','Culture','Family','Health']
n=500
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)
}

df = pd.DataFrame(data)
print("DataFrame Head:")
print(df.head())
print("\nDataFramw Description:")
print(df.describe())
print("\nCategory Counts:")
```

```
print(df['Category'])
```

```
DataFrame Head:
        Date Category  Likes
0 2021-01-01  Fashion   6121
1 2021-01-02  Fashion    340
2 2021-01-03   Family   4185
3 2021-01-04  Fashion    293
4 2021-01-05   Health   5510

DataFramw Description:
             Likes
count   500.000000
mean   5111.572000
std    2794.430122
min      30.000000
25%    2944.000000
50%    5169.000000
75%    7402.500000
max    9970.000000

Category Counts:
0      Fashion
1      Fashion
2       Family
3      Fashion
4       Health
        …
495      Music
496      Music
497    Culture
498    Fashion
499     Travel
Name: Category, Length: 500, dtype: object
```

[4]:
```
import pandas as pd
import random
import numpy as np
categories=['Food','Travel','Fashion','Fitness','Music','Culture','Family','Health']
n=500
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)
}

df = pd.DataFrame(data)
```

```
print("Original Dataframe Head:")
print(df.head())

df=df.dropna()

df=df.drop_duplicates()

df['Date']=pd.to_datetime(df['Date'])

df['Likes']=df['Likes'].astype(int)

print("\nCleaned Dataframe Head:")
print(df.head())

print("\nDataFrame Info:")
print(df.info())

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

print("\nCategory Counts:")
print(df["Category"].value_counts())
```

```
Original Dataframe Head:
        Date Category  Likes
0 2021-01-01  Fitness   2726
1 2021-01-02   Family   1025
2 2021-01-03    Music   2207
3 2021-01-04  Culture   7823
4 2021-01-05   Family   4833

Cleaned Dataframe Head:
        Date Category  Likes
0 2021-01-01  Fitness   2726
1 2021-01-02   Family   1025
2 2021-01-03    Music   2207
3 2021-01-04  Culture   7823
4 2021-01-05   Family   4833

DataFrame Info:
<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

DataFrame Description:
             Likes
count   500.000000
mean   4877.324000
std    2933.541675
min       0.000000
25%    2243.250000
50%    4694.000000
75%    7330.750000
max    9989.000000

Category Counts:
Music      77
Travel     69
Culture    66
Family     64
Fashion    62
Health     57
Fitness    55
Food       50
Name: Category, dtype: int64
```

[11]:
```python
import pandas as pd
import random
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

categories=['Food','Travel','Fashion','Fitness','Music','Culture','Family','Health']
n=500
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)
}

df=pd.DataFrame(data)
df=df.dropna()
df=df.drop_duplicates()
df['Date']=pd.to_datetime(df['Date'])
df['Likes']=df['Likes'].astype(int)
```

```
sns.histplot(df['Likes'], kde=True)
plt.title('Histogram of Likes')
plt.xlabel('Likes')
plt.ylabel('Frequency')
plt.show()


plt.figure(figsize=(10,6))
sns.boxplot(x='Category', y='Likes', date=df)
plt.title('Boxplot of Likes by Category')
plt.xlabel('Category')
plt.ylabel('Likes')
plt.xticks(rotation=45)
plt.show()

mean_likes=df['Likes'].mean()
print(f"Mean of Likes: {mean_likes}")

mean_likes_by_category=df.groupby('Category')['Likes'].mean()
print("\nMean Likes by Category:")
print(mean_likes_by_category)
```

```
␣
↪---------------------------------------------------------------------------

       AttributeError                            Traceback (most recent call␣
↪last)

       <ipython-input-11-8cab549044fa> in <module>
        19 df['Likes']=df['Likes'].astype(int)
        20
   ---> 21 sns.histplot(df['Likes'], kde=True)
        22 plt.title('Histogram of Likes')
        23 plt.xlabel('Likes')


       AttributeError: module 'seaborn' has no attribute 'histplot'
```

[2]:
```
import seaborn as sns
import matplotlib.pyplot as plt

sns.histplot(df['Likes'], kde=True)
plt.title('Histogram of Likes')
plt.xlabel('Likes')
```

```
plt.ylabel('Frequency')
plt.show()
```

```
      ␣
↪---------------------------------------------------------------------------

       AttributeError                            Traceback (most recent call␣
↪last)

       <ipython-input-2-271b4a401b30> in <module>
         2 import matplotlib.pyplot as plt
         3
   ----> 4 sns.histplot(df['Likes'], kde=True)
         5 plt.title('Histogram of Likes')
         6 plt.xlabel('Likes')


       AttributeError: module 'seaborn' has no attribute 'histplot'
```

[ ]: