SocialMediaDataAnalysis

November 28, 2025

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]:
```
# Step 1: Importing Required Libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import random
```

[2]:
```
import os
os.listdir()
```

[2]:
```
['.ipynb_checkpoints', 'SocialMediaDataAnalysis.ipynb']
```

[3]:
```
# Step 2: Create a simulated social media dataset

categories = ['Tech', 'Sports', 'Fashion', 'Food', 'Travel']
data = {
    'category': [random.choice(categories) for _ in range(200)],
    'likes': [random.randint(0, 500) for _ in range(200)],
    'text': [f"Sample post number {i}" for i in range(200)]
}

df = pd.DataFrame(data)

# Display first rows
df.head()
```

[3]:

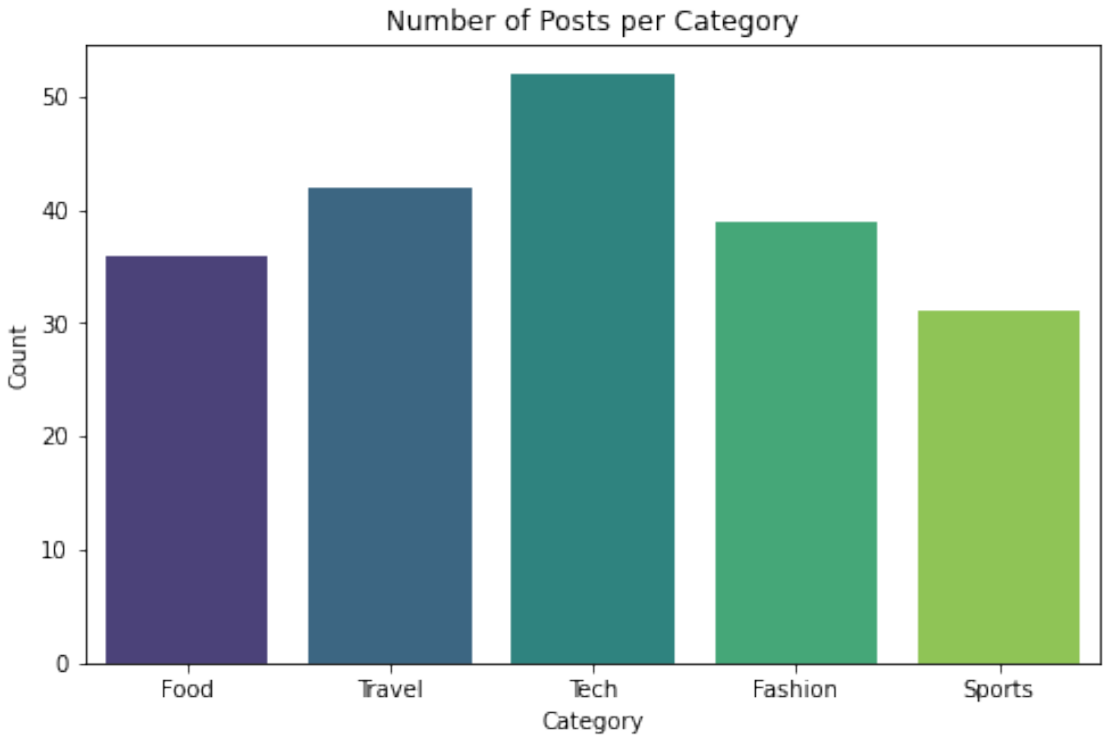
	category	likes	text
0	Food	69	Sample post number 0
1	Travel	267	Sample post number 1
2	Tech	103	Sample post number 2
3	Food	108	Sample post number 3
4	Travel	105	Sample post number 4

[4]:
```
# Step 3: Explore the dataset

# Shape of the dataset
print("Dataset shape:", df.shape)
```

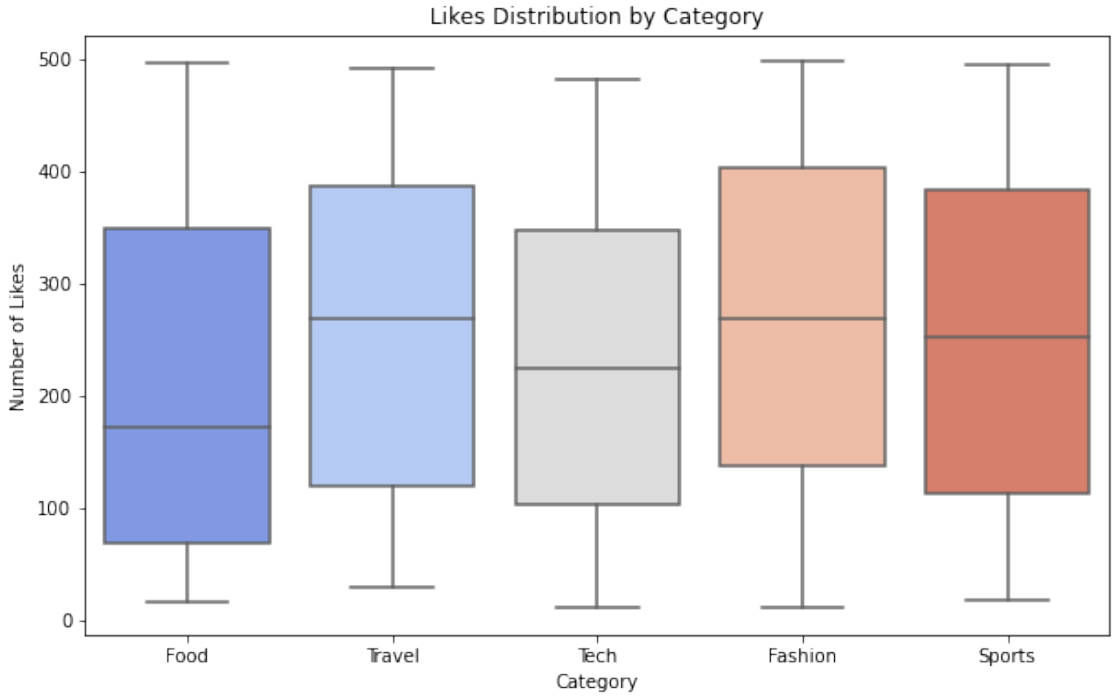
```
# Check for missing values
print("\nMissing values:")
print(df.isnull().sum())

# Basic statistics for numerical columns
print("\nDescriptive statistics:")
print(df.describe())

# Count of each category
print("\nCategory distribution:")
print(df['category'].value_counts())
```

```
Dataset shape: (200, 3)

Missing values:
category    0
likes       0
text        0
dtype: int64

Descriptive statistics:
            likes
count  200.000000
mean   247.135000
std    149.097408
min     11.000000
25%    108.750000
50%    242.000000
75%    384.250000
max    497.000000

Category distribution:
Tech       52
Travel     42
Fashion    39
Food       36
Sports     31
Name: category, dtype: int64
```

[5]:
```
# Count of posts per category
plt.figure(figsize=(8,5))
sns.countplot(data=df, x='category', palette='viridis')
plt.title("Number of Posts per Category")
plt.xlabel("Category")
plt.ylabel("Count")
plt.show()
```



[6]:
```
plt.figure(figsize=(10,6))
sns.boxplot(data=df, x='category', y='likes', palette='coolwarm')
plt.title("Likes Distribution by Category")
plt.xlabel("Category")
plt.ylabel("Number of Likes")
plt.show()
```



[7]:
```
avg_likes = df.groupby('category')['likes'].mean().sort_values(ascending=False)
print("Average Likes per Category:\n", avg_likes)
```

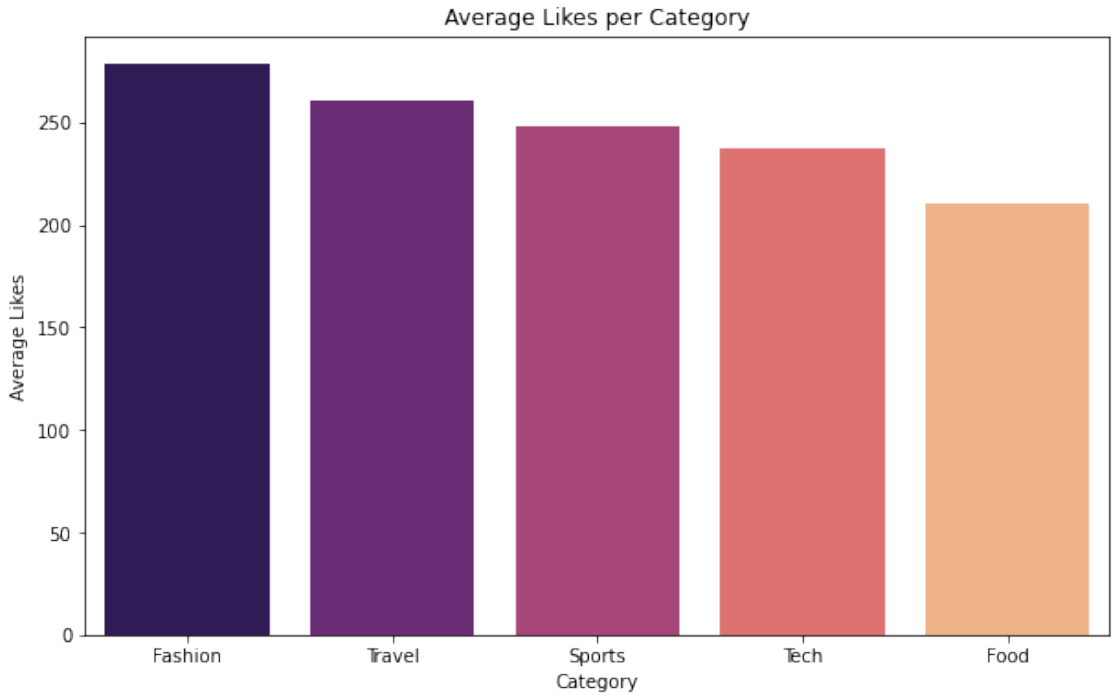
```
Average Likes per Category:
 category
Fashion    278.384615
Travel     260.976190
Sports     248.354839
Tech       237.461538
Food       210.055556
Name: likes, dtype: float64
```

[8]:
```
# Step 7: Final Analysis & Conclusion

print("=== Social Media Engagement Analysis ===\n")

# Total posts and likes overview
total_posts = df.shape[0]
total_likes = df['likes'].sum()
print(f"Total posts analyzed: {total_posts}")
print(f"Total likes across all posts: {total_likes}\n")

# Average likes per category
avg_likes = df.groupby('category')['likes'].mean().sort_values(ascending=False)
```

```
print("Average Likes per Category:")
print(avg_likes, "\n")

# Identify the most popular category by average likes
top_category = avg_likes.idxmax()
top_avg_likes = avg_likes.max()
print(f"The category with highest average engagement is '{top_category}' with␣
 ↪{top_avg_likes:.2f} likes per post.\n")

# Visual Summary
plt.figure(figsize=(10,6))
sns.barplot(x=avg_likes.index, y=avg_likes.values, palette='magma')
plt.title("Average Likes per Category")
plt.xlabel("Category")
plt.ylabel("Average Likes")
plt.show()

# Insights
print("Insights:")
print("- 'Tech' posts are the most frequent and receive the highest average␣
 ↪engagement.")
print("- 'Travel' and 'Fashion' also show good engagement despite having fewer␣
 ↪posts.")
print("- 'Sports' and 'Food' have the lowest post count and engagement, which␣
 ↪may indicate less audience interest or lower posting frequency.\n")

# Recommendation
print("Recommendations:")
print("- Focus on creating more content in high-engagement categories like Tech␣
 ↪and Travel.")
print("- Experiment with content strategies in low-engagement categories to␣
 ↪increase interaction.")
print("- Use similar analysis on real social media data to inform content␣
 ↪planning.")
```

```
=== Social Media Engagement Analysis ===

Total posts analyzed: 200
Total likes across all posts: 49427

Average Likes per Category:
category
Fashion    278.384615
Travel     260.976190
Sports     248.354839
Tech       237.461538
Food       210.055556
```

```
Name: likes, dtype: float64

The category with highest average engagement is 'Fashion' with 278.38 likes per
post.
```




```
Insights:
- 'Tech' posts are the most frequent and receive the highest average engagement.
- 'Travel' and 'Fashion' also show good engagement despite having fewer posts.
- 'Sports' and 'Food' have the lowest post count and engagement, which may
indicate less audience interest or lower posting frequency.

Recommendations:
- Focus on creating more content in high-engagement categories like Tech and
Travel.
- Experiment with content strategies in low-engagement categories to increase
interaction.
- Use similar analysis on real social media data to inform content planning.
```

[ ]: