Project-1 (Media and Technology)

Name:- Mehnaz Shafeek

Code transcript

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt


df = pd.read_csv(r"C:\Users\shafe\OneDrive\Desktop\STUDIES\YEAR 3\finlatics\Media and Technology\Media and Technology\Global YouTube Statistics2.csv", encoding='latin1')

 #Preproccessing the Data


df['subscribers']=df['subscribers'].fillna(df['subscribers'].median())

df['Abbreviation']=df['Abbreviation'].fillna(df['Abbreviation'].mode()[0])

df.dropna(subset=['Country'],inplace=True)

df.dropna(subset=['Country of origin'],inplace=True)

df['video_views_rank']=df['video_views_rank'].fillna(df['video_views_rank'].median())

df['channel_type_rank']=df['channel_type_rank'].fillna(df['channel_type_rank'].median())

df['video_views_for_the_last_30_days']=df['video_views_for_the_last_30_days'].fillna(df['video_views_for_the_last_30_days'].median())

df['created_year']=df['created_year'].fillna(df['created_year'].median())

df['subscribers_for_last_30_days']=df['subscribers_for_last_30_days'].fillna(df['subscribers_for_last_30_days'].median())

df['created_date']=df['created_date'].fillna(df['created_date'].median())

df['Gross tertiary education enrollment (%)']=df['Gross tertiary education enrollment (%)'].fillna(df['Gross tertiary education enrollment (%)'].median())

df['Population']=df['Population'].fillna(df['Population'].median())

df['Unemployment rate']=df['Unemployment rate'].fillna(df['Unemployment rate'].median())

df['Urban_population']=df['Urban_population'].fillna(df['Urban_population'].median())

df['Latitude']=df['Latitude'].fillna(df['Latitude'].median())

df['Longitude']=df['Longitude'].fillna(df['Longitude'].median())
```

```
df['category']=df['category'].fillna(df['category'].mode()[0])
df['channel_type']=df['channel_type'].fillna(df['channel_type'].mode()[0])
df['country_rank']=df['country_rank'].fillna(df['country_rank'].median())
df['created_month']=df['created_month'].fillna(df['created_month'].mode()[0])

print(df.isnull().sum())

print(' What are the top 10 YouTube channels based on the number of subscribers?')
top_10_channels=df.sort_values(by='subscribers',ascending=False).head(10)
print(top_10_channels[['Youtuber','subscribers']])

print('Q2. Which category has the highest average number of subscribers?')
grp_data=df.groupby(['category','subscribers'])['subscribers'].mean()
highestavg=grp_data.sort_values(ascending=False).head(1)
print(highestavg)

print('Q3.	How many videos, on average, are uploaded by YouTube channels in each category?')
grpby_2=df.groupby(['category','uploads'])['uploads'].mean()
print(grpby_2)

print('Q4.What are the top 5 countries with the highest number of YouTube channels?')
grpby_3=df.groupby(['Abbreviation'])['Abbreviation'].value_counts()
top_5_countries=grpby_3.sort_values(ascending=False).head(5)
print(top_5_countries)

print('Q5.What is the distribution of channel types across different categories?')
plt.figure(figsize=(12, 6))
sns.countplot(data=df, x='category', hue='channel_type')
plt.title('Distribution of Channel Types Across Categories')
plt.xticks(rotation=45)
plt.xlabel('Category')
plt.ylabel('Count')
plt.legend(title='Channel Type')
plt.show()

print('Q6.Is there a correlation between the number of subscribers and total video views for YouTube channels?')
plt.figure(figsize=(8,6))
sns.heatmap(df[['subscribers','video views']].corr(),annot=True,cmap='coolwarm')
plt.title('correlational Matrix')
plt.show()

print('Q7.How do the monthly earnings vary throughout different categories?')
plt.figure(figsize=(12, 6))
df.groupby('category')[['highest_monthly_earnings', 'lowest_monthly_earnings']].mean().plot(kind='bar')
plt.title('Average Monthly Earnings by Category')
plt.xlabel('Category')
plt.ylabel('Earnings (in $)')
plt.xticks(rotation=90)
plt.legend(title='Earnings Type')
plt.show()

print('Q8.What is the overall trend in subscribers gained in the last 30 days across all channels?')
plt.figure(figsize=(10, 6))
df.groupby('channel_type')[['subscribers_for_last_30_days']].mean().plot(kind='bar')
plt.title('Overall trend in Subsribers')
plt.xlabel('channel_type')
plt.ylabel('subscribers')
plt.show()
```

```
print('Q9.Are there any outliers in terms of yearly earnings from YouTube channels?')
sns.boxplot(y='highest_yearly_earnings',data=df)
plt.title('Box Plot for highly yearly earnings')
plt.show()
sns.boxplot(y='lowest_yearly_earnings',data=df)
plt.title('Box Plot for lowest yearly earnings')
plt.show()

print('Q10.What is the distribution of channel creation dates? Is there any trend over time?')
df['created_date'] = pd.to_datetime(df['created_date'], errors='coerce')
plt.figure(figsize=(12, 6))
sns.countplot(x='created_year', data=df, order=df['created_year'].value_counts().index)
plt.title('Channels Created Each Year')
plt.xticks(rotation=45)
plt.xlabel('Year')
plt.ylabel('Count')
plt.show()

print('Q11.Is there a relationship between gross tertiary education enrollment and the number of YouTube channels in a country?')
channel_count = df['Country'].value_counts().reset_index()
channel_count.columns = ['Country', 'channel_count']
education_data = df[['Country', 'Gross tertiary education enrollment (%)']].drop_duplicates()
merged_data1 = pd.merge(channel_count, education_data, on='Country', how='inner')
plt.figure(figsize=(8, 6))
sns.heatmap(merged_data1[['Gross tertiary education enrollment (%)', 'channel_count']].corr(), annot=True, cmap='coolwarm')
plt.title('Relationship between Education Enrollment and Number of YouTube Channels')
plt.show()

print('Q12.	How does the unemployment rate vary among the top 10 countries with the highest number of YouTube channels?')
```

```
channel_count = df['Country'].value_counts().reset_index()
channel_count.columns = ['Country', 'channel_count']
top_10_countries = channel_count.head(10)
unemployment_data = df[['Country', 'Unemployment rate']].drop_duplicates()
merged_data2=pd.merge(top_10_countries,unemployment_data, on = 'Country' , how='inner')
plt.figure(figsize=(10, 6))
sns.barplot(x='Country', y='Unemployment rate', data=merged_data2)
plt.title('Unemployment Rate in Top 10 Countries with the Most YouTube Channels')
plt.xlabel('Country')
plt.ylabel('Unemployment Rate (%)')
plt.xticks(rotation=45)
plt.show()

print('Q13.	What is the average urban population percentage in countries with YouTube channels?')
population_data=df[['Country','Urban_population']].drop_duplicates()
avg_pop=population_data['Urban_population'].mean()
print(f'average urban population percentage in countries with YouTube channels is {avg_pop} ')

print('Q14.Are there any patterns in the distribution of YouTube channels based on latitude and longitude coordinates?')
plt.figure(figsize=(10,8))
sns.scatterplot(x='Longitude',y='Latitude',data=df,hue='Country',palette='coolwarm',s=100)
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.title('Patterns in the distribution of YouTube channels based on latitude and longitude')
plt.legend(title='Country',bbox_to_anchor=(1.05,1),loc='upper left')
plt.grid(True)
plt.show()

print('Q15.What is the correlation between the number of subscribers and the population of a country?')
```

```
plt.figure(figsize=(8,6))

sns.heatmap(df[['subscribers','Population']].corr(),annot=True,cmap='coolwarm')

plt.title('Correlation between the number of subscribers and the population')

plt.show()


print('Q16.	How do the top 10 countries with the highest number of YouTube channels compare in terms of their total population?')

channel_count = df['Country'].value_counts().reset_index()

channel_count.columns = ['Country', 'Channel_Count']

top_10_countries = channel_count.head(10)

population_data = df[['Country', 'Population']].drop_duplicates()

merged_data = pd.merge(top_10_countries, population_data, on='Country', how='inner')

#  Plot the results

plt.figure(figsize=(12, 6))

sns.barplot(x='Country', y='Population', data=merged_data, palette='Blues_d')

plt.title('Population of Top 10 Countries with the Most YouTube Channels')

plt.xlabel('Country')

plt.ylabel('Population (in billions)')

plt.xticks(rotation=45)

plt.show()


print('Q17.	Is there a correlation between the number of subscribers gained in the last 30 days and the unemployment rate in a country?')

plt.figure(figsize=(10,8))

sns.scatterplot(x='subscribers',y='Unemployment rate',data=df,hue='Country',palette='PuBuGn',s=100)

plt.xlabel('subscribers')

plt.ylabel('Unemployment rate')

plt.title('Patterns in the number of subscribers and unemployment rate of a country')

plt.legend(title='Country',bbox_to_anchor=(1.05,1),loc='upper left')

plt.grid(True)

plt.show()
```

```
print('Q18.	How does the distribution of video views for the last 30 days vary across different channel types?')

plt.figure(figsize=(12, 6))

sns.barplot(x='channel_type', y='video_views_for_the_last_30_days', data=df, palette='Blues_d')

plt.title(' distribution of video views for the last 30 days vary across different channel types')

plt.xlabel('Channel type')

plt.ylabel('Video Views')

plt.xticks(rotation=45)

plt.show()


print('Q19. Are there any seasonal trends in the number of videos uploaded by YouTube channels (Quarterly Analysis)?')

plt.figure(figsize=(12,6))

sns.lineplot(x='created_month',y='uploads',hue='channel_type',data=df)

plt.title('Seasonal trends in the number of videos uploaded by YouTube channels')

plt.show()


print('Q20.What is the average number of subscribers gained per month since the creation of YouTube channels till now?')

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

current_date = pd.to_datetime('today')

df['channel_age_months'] = ((current_date - df['created_date']).dt.days) / 30

df['subscribers_per_month'] = df['subscribers'] / df['channel_age_months']

plt.figure(figsize=(12, 6))

sns.histplot(df['subscribers_per_month'], kde=True, color='teal')

plt.title('Distribution of Average Subscribers Gained Per Month')

plt.xlabel('Subscribers per Month')

plt.ylabel('Count of Channels')

plt.grid(True)

plt.show()

overall_avg = df['subscribers_per_month'].mean()
```

```
print(f'Overall Average Subscribers Gained Per Month: {overall_avg:.2f}')
```