Questions:

Preprocess the dataset prior to analysis for accurate insights

1. What are the top 10 YouTube channels based on the number of subscribers?

```
import pandas as pd

file_path = r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['subscribers'] = pd.to_numeric(df['subscribers'], errors='coerce')
    df = df.dropna(subset=['subscribers'])
    df_sorted = df.sort_values(by='subscribers', ascending=False)
    top_10_channels = df_sorted.head(10)
    print(top_10_channels[['Youtuber', 'subscribers']])

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

2. Which category has the highest average number of subscribers?

```
import pandas as pd

file_path =  r'C:\Globalytstat\globalyt.xlsx'

try:

    df = pd.read_excel(file_path, engine='openpyxl')
    df['subscribers'] = pd.to_numeric(df['subscribers'], errors='coerce')
    df = df.dropna(subset=['subscribers'])
    category_avg_subs = df.groupby('category')['subscribers'].mean()
    highest_avg_subs_category = category_avg_subs.idxmax()
    highest_avg_subs_value = category_avg_subs.max()

    print(f"The category with the highest average number of subscribers is '{highest_avg_subs_category}' with an average of {highest_avg_subs_value:.2f} subscribers.")

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

3. How many videos, on average, are uploaded by YouTube channels in each category?

```
import pandas as pd

file_path =  r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['uploads'] = pd.to_numeric(df['uploads'], errors='coerce')
    df = df.dropna(subset=['uploads'])
    category_avg_uploads = df.groupby('category')['uploads'].mean()
    print(category_avg_uploads)

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
```

```
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

4. What are the top 5 countries with the highest number of YouTube channels?

```
import pandas as pd


file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    country_channel_counts = df['Country'].value_counts()
    top_5_countries = country_channel_counts.head(5)
    print(top_5_countries)

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if
you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

5. What is the distribution of channel types across different categories?

```
import pandas as pd
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    channel_type_distribution = pd.crosstab(df['category'], df['channel_type'])
    print(channel_type_distribution)

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if
you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

6. Is there a correlation between the number of subscribers and total video views for YouTube channels?

```
import pandas as pd
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['subscribers'] = pd.to_numeric(df['subscribers'], errors='coerce')
    df['video views'] = pd.to_numeric(df['video views'], errors='coerce')
    df = df.dropna(subset=['subscribers', 'video views'])
    correlation = df['subscribers'].corr(df['video views'])
    print(f"The correlation between the number of subscribers and total video views is
{correlation:.2f}")

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if
you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

7. How do the monthly earnings vary throughout different categories?

```
import pandas as pd
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
   df = pd.read_excel(file_path, engine='openpyxl')
   df['lowest_monthly_earnings'] = pd.to_numeric(df['lowest_monthly_earnings'],
errors='coerce')
   df['highest_monthly_earnings'] = pd.to_numeric(df['highest_monthly_earnings'],
errors='coerce')
   df = df.dropna(subset=['lowest_monthly_earnings', 'highest_monthly_earnings'])
   df['average_monthly_earnings'] = (df['lowest_monthly_earnings'] +
df['highest_monthly_earnings']) / 2
   earnings_summary = df.groupby('category')['average_monthly_earnings'].describe()
   print(earnings_summary)

except PermissionError:
   print("Permission denied: Please check if the file is open in another application or if
you have the necessary permissions.")
except FileNotFoundError:
   print("File not found: Please check the file path.")
except UnicodeDecodeError:
   print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
   print(f"An error occurred: {e}")
```

8. What is the overall trend in subscribers gained in the last 30 days across all channels?

```
import pandas as pd
import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'


try:
   df = pd.read_excel(file_path, engine='openpyxl')
```

```
    df['subscribers_for_last_30_days'] = pd.to_numeric(df['subscribers_for_last_30_days'], errors='coerce')
    df = df.dropna(subset=['subscribers_for_last_30_days'])
    summary_stats = df['subscribers_for_last_30_days'].describe()
    print(summary_stats)

    plt.figure(figsize=(10, 6))
    plt.hist(df['subscribers_for_last_30_days'], bins=50, edgecolor='k', alpha=0.7)
    plt.title('Distribution of Subscribers Gained in the Last 30 Days')
    plt.xlabel('Subscribers Gained')
    plt.ylabel('Frequency')
    plt.grid(True)
    plt.show()

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

9. Are there any outliers in terms of yearly earnings from YouTube channels?

```
import pandas as pd
import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['lowest_yearly_earnings'] = pd.to_numeric(df['lowest_yearly_earnings'], errors='coerce')
    df['highest_yearly_earnings'] = pd.to_numeric(df['highest_yearly_earnings'], errors='coerce')
    df = df.dropna(subset=['lowest_yearly_earnings', 'highest_yearly_earnings'])
    df['average_yearly_earnings'] = (df['lowest_yearly_earnings'] +
df['highest_yearly_earnings']) / 2
    Q1 = df['average_yearly_earnings'].quantile(0.25)
    Q3 = df['average_yearly_earnings'].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = df[(df['average_yearly_earnings'] < lower_bound) |
(df['average_yearly_earnings'] > upper_bound)]
    print("Outliers in terms of yearly earnings:")
    print(outliers[['Youtuber', 'average_yearly_earnings']])
    plt.figure(figsize=(10, 6))
    plt.boxplot(df['average_yearly_earnings'])
    plt.title('Boxplot of Average Yearly Earnings')
    plt.ylabel('Average Yearly Earnings')
    plt.grid(True)
    plt.show()

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if
you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

10.What is the distribution of channel creation dates? Is there any trend over time?

```
import pandas as pd
import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['created_date'] = pd.to_datetime(df['created_date'], errors='coerce')
    df = df.dropna(subset=['created_date'])
    df['created_year'] = df['created_date'].dt.year
    df['created_month'] = df['created_date'].dt.to_period('M')
    plt.figure(figsize=(12, 6))
    df['created_year'].value_counts().sort_index().plot(kind='bar', color='skyblue')
    plt.title('Distribution of YouTube Channel Creation Dates by Year')
    plt.xlabel('Year')
    plt.ylabel('Number of Channels Created')
    plt.grid(True)
    plt.show()

    plt.figure(figsize=(12, 6))
    df['created_month'].value_counts().sort_index().plot(kind='line', color='skyblue',
marker='o')
    plt.title('Trend of YouTube Channel Creation Dates Over Time')
    plt.xlabel('Year-Month')
    plt.ylabel('Number of Channels Created')
    plt.grid(True)
    plt.show()

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if
you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

11.Is there a relationship between gross tertiary education enrollment and the number of YouTube channels in a country?

```
import pandas as pd
import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['Gross tertiary education enrollment (%)'] = pd.to_numeric(df['Gross tertiary education enrollment (%)'], errors='coerce')
    df = df.dropna(subset=['Gross tertiary education enrollment (%)', 'Country'])
    country_channel_counts = df['Country'].value_counts().reset_index()
    country_channel_counts.columns = ['Country', 'Number of Channels']
    merged_df = pd.merge(country_channel_counts, df[['Country', 'Gross tertiary education enrollment (%)']].drop_duplicates(), on='Country')
    correlation = merged_df['Gross tertiary education enrollment (%)'].corr(merged_df['Number of Channels'])
    print(f"The correlation between gross tertiary education enrollment and the number of YouTube channels is {correlation:.2f}")

    plt.figure(figsize=(10, 6))
    plt.scatter(merged_df['Gross tertiary education enrollment (%)'], merged_df['Number of Channels'], alpha=0.7, color='skyblue')
    plt.title('Relationship between Gross Tertiary Education Enrollment and Number of YouTube Channels')
    plt.xlabel('Gross Tertiary Education Enrollment (%)')
    plt.ylabel('Number of YouTube Channels')
    plt.grid(True)
    plt.show()

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

12. How does the unemployment rate vary among the top 10 countries with the highest number of YouTube channels?

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['Unemployment rate'] = pd.to_numeric(df['Unemployment rate'], errors='coerce')
    df = df.dropna(subset=['Country', 'Unemployment rate'])
    country_channel_counts = df['Country'].value_counts().reset_index()
    country_channel_counts.columns = ['Country', 'Number of Channels']
    top_10_countries = country_channel_counts.head(10)
    merged_df = pd.merge(top_10_countries, df[['Country', 'Unemployment rate']].drop_duplicates(), on='Country')
    print("Top 10 countries with the highest number of YouTube channels and their unemployment rates:")
    print(merged_df[['Country', 'Number of Channels', 'Unemployment rate']])
    plt.figure(figsize=(12, 6))
    plt.bar(merged_df['Country'], merged_df['Unemployment rate'], color='skyblue')
    plt.title('Unemployment Rate Among the Top 10 Countries with the Highest Number of YouTube Channels')
    plt.xlabel('Country')
    plt.ylabel('Unemployment Rate (%)')
    plt.grid(True)
    plt.show()

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

13. What is the average urban population percentage in countries with YouTube channels?

```
import pandas as pd
file_path =r'C:\Globalytstat\globalyt.xlsx'
```

```
try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['Urban_population'] = pd.to_numeric(df['Urban_population'], errors='coerce')
    df = df.dropna(subset=['Urban_population'])
    average_urban_population = df['Urban_population'].mean()
    print(f"The average urban population percentage in countries with YouTube channels is {average_urban_population:.2f}%")

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

14. Are there any patterns in the distribution of YouTube channels based on latitude and longitude coordinates?

```
import pandas as pd
import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['Latitude'] = pd.to_numeric(df['Latitude'], errors='coerce')
    df['Longitude'] = pd.to_numeric(df['Longitude'], errors='coerce')
    df = df.dropna(subset=['Latitude', 'Longitude'])
    plt.figure(figsize=(12, 6))
    plt.scatter(df['Longitude'], df['Latitude'], alpha=0.7, c='skyblue', edgecolors='w', linewidth=0.5)
    plt.title('Distribution of YouTube Channels Based on Latitude and Longitude')
    plt.xlabel('Longitude')
    plt.ylabel('Latitude')
    plt.grid(True)
    plt.show()

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you
have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

15. What is the correlation between the number of subscribers and the population of a country?

```
import pandas as pd
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['subscribers'] = pd.to_numeric(df['subscribers'], errors='coerce')
    df['Population'] = pd.to_numeric(df['Population'], errors='coerce')
    df = df.dropna(subset=['subscribers', 'Population'])
    country_subscribers = df.groupby('Country')['subscribers'].sum().reset_index()
    country_population = df[['Country', 'Population']].drop_duplicates()
    merged_df = pd.merge(country_subscribers, country_population, on='Country')
    correlation = merged_df['subscribers'].corr(merged_df['Population'])
    print(f"The correlation between the number of subscribers and the population of a
country is {correlation:.2f}")

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you
have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
```

```
except Exception as e:
    print(f"An error occurred: {e}")
```

16. How do the top 10 countries with the highest number of YouTube channels compare in terms of their total population?

```
import pandas as pd
import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['Population'] = pd.to_numeric(df['Population'], errors='coerce')
    df = df.dropna(subset=['Country', 'Population'])
    country_channel_counts = df['Country'].value_counts().reset_index()
    country_channel_counts.columns = ['Country', 'Number of Channels']
    top_10_countries = country_channel_counts.head(10)
    merged_df = pd.merge(top_10_countries, df[['Country', 'Population']].drop_duplicates(),
on='Country')
    print("Top 10 countries with the highest number of YouTube channels and their total
population:")
    print(merged_df[['Country', 'Number of Channels', 'Population']])
    plt.figure(figsize=(12, 6))
    plt.bar(merged_df['Country'], merged_df['Population'], color='skyblue')
    plt.title('Total Population of Top 10 Countries with the Highest Number of YouTube
Channels')
    plt.xlabel('Country')
    plt.ylabel('Population')
    plt.grid(True)
    plt.show()

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you
have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
```

```
except Exception as e:
    print(f"An error occurred: {e}")
```

17. Is there a correlation between the number of subscribers gained in the last 30 days and the unemployment rate in a country?

```
import pandas as pd
#import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'



try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['subscribers_for_last_30_days'] = pd.to_numeric(df['subscribers_for_last_30_days'],
errors='coerce')
    df['Unemployment rate'] = pd.to_numeric(df['Unemployment rate'], errors='coerce')
    df = df.dropna(subset=['subscribers_for_last_30_days', 'Unemployment rate'])
    country_subscribers_30_days =
df.groupby('Country')['subscribers_for_last_30_days'].sum().reset_index()
    country_unemployment_rate = df[['Country', 'Unemployment rate']].drop_duplicates()
    merged_df = pd.merge(country_subscribers_30_days, country_unemployment_rate,
on='Country')
    correlation = merged_df['subscribers_for_last_30_days'].corr(merged_df['Unemployment
rate'])

    print(f"The correlation between the number of subscribers gained in the last 30 days and
the unemployment rate in a country is {correlation:.2f}")

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you
have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

18. How does the distribution of video views for the last 30 days vary across different channel types?

```
import pandas as pd
import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['video_views_for_the_last_30_days'] =
pd.to_numeric(df['video_views_for_the_last_30_days'], errors='coerce')
    df = df.dropna(subset=['video_views_for_the_last_30_days', 'channel_type'])
    plt.figure(figsize=(14, 7))
    df.boxplot(column='video_views_for_the_last_30_days', by='channel_type', grid=False,
vert=False)
    plt.title('Distribution of Video Views for the Last 30 Days Across Different Channel Types')
    plt.suptitle('')
    plt.xlabel('Video Views for the Last 30 Days')
    plt.ylabel('Channel Type')
    plt.show()

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you
have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different
encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```

19. Are there any seasonal trends in the number of videos uploaded by YouTube channels?

```
import pandas as pd
import matplotlib.pyplot as plt
file_path =r'C:\Globalytstat\globalyt.xlsx'
```

```
try:
  df = pd.read_excel(file_path, engine='openpyxl')
  df['created_date'] = pd.to_datetime(df['created_date'], errors='coerce')
  df = df.dropna(subset=['created_date'])
  df['year'] = df['created_date'].dt.year
  df['month'] = df['created_date'].dt.month
  monthly_uploads = df.groupby(['year', 'month']).size().reset_index(name='num_uploads')
  monthly_uploads_pivot = monthly_uploads.pivot(index='month', columns='year', values='num_uploads')
  plt.figure(figsize=(14, 7))
  for year in monthly_uploads_pivot.columns:
   plt.plot(monthly_uploads_pivot.index, monthly_uploads_pivot[year], label=year)
  plt.title('Monthly Uploads of YouTube Videos Over Years')
  plt.xlabel('Month')
  plt.ylabel('Number of Videos Uploaded')
  plt.legend(title='Year')
  plt.grid(True)
  plt.xticks(range(1, 13))
  plt.show()

except PermissionError:
  print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
  print("File not found: Please check the file path.")
except UnicodeDecodeError:
  print("Unicode decode error: Please check the file encoding or try using a different encoding.")
except Exception as e:
  print(f"An error occurred: {e}")
```

20. What is the average number of subscribers gained per month since the creation of YouTube channels till now?

```
import pandas as pd
file_path =r'C:\Globalytstat\globalyt.xlsx'

try:
    df = pd.read_excel(file_path, engine='openpyxl')
    df['subscribers'] = pd.to_numeric(df['subscribers'], errors='coerce')
    df['created_date'] = pd.to_datetime(df['created_date'], errors='coerce')
    df = df.dropna(subset=['subscribers', 'created_date'])
    df['months_since_creation'] = ((pd.Timestamp.now() - df['created_date']) / pd.offsets.MonthEnd(1)).astype(int)
    df['subscribers_per_month'] = df['subscribers'] / df['months_since_creation']
    overall_avg_subscribers_per_month = df['subscribers_per_month'].mean()
    print(f"The average number of subscribers gained per month since the creation of YouTube channels till now is {overall_avg_subscribers_per_month:.2f}")

except PermissionError:
    print("Permission denied: Please check if the file is open in another application or if you have the necessary permissions.")
except FileNotFoundError:
    print("File not found: Please check the file path.")
except UnicodeDecodeError:
    print("Unicode decode error: Please check the file encoding or try using a different encoding.")
except Exception as e:
    print(f"An error occurred: {e}")
```