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

avod_2017= pd.read_csv("avod_2017.csv")
avod_2018= pd.read_csv("avod_2018.csv")
avod_2016= pd.read_csv("avod_2016.csv")
avod_2019= pd.read_csv("avod_2019.csv")
avod_2020= pd.read_csv("avod_2020.csv")
avod_2021= pd.read_csv("avod_2021.csv")
avod_2022= pd.read_csv("avod_2022.csv")
avod_2023= pd.read_csv("avod_2023.csv")
avod_missing= pd.read_csv("missing_starttimepst.csv")

# to get average streams per user by year
# List of dataframes
dataframes = [avod_2019, avod_2020, avod_2021, avod_2022, avod_2023,
avod_missing]

# Iterate through each dataframe
for idx, df in enumerate(dataframes):
    # Calculate the stream session count for each user
    extid_sessionid_counts = df.groupby('extid')['streamsess'].count()
    # Calculate the average stream session count for each user
    average_session_count = extid_sessionid_counts.mean()
    # Print the average stream session count for each user
    print(f"Average Stream Session Count for Each User in
avod_{2019+idx}:")
    print(average_session_count)

# to count unique IDs by year

unique_extid_values2019 = avod_2019['extid'].unique()
unique_ids_count2019 = len(unique_extid_values2019)
print("Number of unique IDs in 2019:", unique_ids_count2019)

unique_extid_values2020 = avod_2020['extid'].unique()
unique_ids_count2020 = len(unique_extid_values2020)
print("Number of unique IDs in 2020:", unique_ids_count2020)

unique_extid_values2021 = avod_2021['extid'].unique()
unique_ids_count2021 = len(unique_extid_values2021)
print("Number of unique IDs in 2021:", unique_ids_count2021)

unique_extid_values2022 = avod_2022['extid'].unique()
unique_ids_count2022 = len(unique_extid_values2022)
print("Number of unique IDs in 2022:", unique_ids_count2022)

unique_extid_values2023 = avod_2023['extid'].unique()
```

```
unique_ids_count2023 = len(unique_extid_values2023)
print("Number of unique IDs in 2023:", unique_ids_count2023)

# to get overall instances for each platform by year
dataframes = [avod_2019, avod_2020, avod_2021, avod_2022, avod_2023,
avod_missing]

# Loop through each DataFrame
for idx, df in enumerate(dataframes, start=2019):
    print(f"Year {idx}:")
    platform_counts = df['plat'].value_counts()
    print(platform_counts)
    print()

# test code to find percentage for streaming count by viewers in
general

# Assuming avod_2023 is your DataFrame
# Group by 'extid' and count occurrences of 'streamsession'
extid_sessionid_counts = avod_2023.groupby('extid')
['streamsess'].count()

# Count how many times each count of 'streamsession' appears
count_counts = extid_sessionid_counts.value_counts()

# Calculate the total number of users
total_users = count_counts.sum()

# Calculate the percentage for each count
percentage_counts = (count_counts / total_users) * 100

# Create a new DataFrame
percentage_dataframe = pd.DataFrame({'Number of Stream Sessions':
percentage_counts.index, 'Percentage of Users':
percentage_counts.values})

# Print the new DataFrame
print(percentage_dataframe)

# to count stream acitivity of a specific platform

# test code to find percentage for stream count by viewers for those
on tv

tv_data = avod_2023[avod_2023['plat'] == 'tv']

# Group by 'extid' and count the number of 'streamsession' for each
unique 'extid'
extid_tv_sessionid_counts = tv_data.groupby('extid')
['streamgsess'].count()
```

```
# Count how many times each count of 'streamsession' appears
count_counts = extid_tv_sessionid_counts.value_counts()

# Calculate the total number of users
total_users = count_counts.sum()

# Calculate the percentage for each count
percentage_counts = (count_counts / total_users) * 100

# Create a new DataFrame
percentage_dataframe = pd.DataFrame({'Number of Stream Sessions':
percentage_counts.index, 'tv': percentage_counts.values})

top_10_rows = percentage_dataframe.head(20)

# Print the new DataFrame
print(top_10_rows)
```