

```

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data
df = pd.read_csv('COVID clinical trials.csv')

# Preprocessing
df['Enrollment'] = pd.to_numeric(df['Enrollment'], errors='coerce').fillna(0)
df['Start Date'] = pd.to_datetime(df['Start Date'], errors='coerce')
df['Start YearMonth'] = df['Start Date'].dt.to_period('M')

# 1. Enrollment by Study Type
study_type_enrollment = df.groupby('Study Type')['Enrollment'].agg(['sum',
'mean', 'median', 'count']).sort_values(by='sum', ascending=False)

# 2. Conditions and Enrollment
# We need to split and explode conditions
df_exploded = df.copy()
df_exploded['Conditions'] = df_exploded['Conditions'].str.split('|')
df_exploded = df_exploded.explode('Conditions')
df_exploded['Conditions'] = df_exploded['Conditions'].str.strip()

# Standardize common COVID names
covid_variants = ['COVID-19', 'Covid19', 'COVID', 'Covid-19', 'SARS-CoV-2',
'Coronavirus Infection',
'Coronavirus', 'COVID 19', 'COVID19', 'Corona Virus Infection', 'SARS-
CoV2', 'SARS-CoV 2']
df_exploded['Conditions'] = df_exploded['Conditions'].apply(lambda x:
'COVID-19' if x in covid_variants else x)

# Top conditions by total enrollment
top_conditions_enrollment = df_exploded.groupby('Conditions')
['Enrollment'].sum().sort_values(ascending=False).head(15)

# 3. Trends over time
# Average and Total enrollment per month
enrollment_trend = df.groupby('Start YearMonth')['Enrollment'].agg(['sum',
'mean']).reset_index()
enrollment_trend['Start YearMonth'] = enrollment_trend['Start
YearMonth'].astype(str)
# Filter for pandemic period (2020 onwards)
enrollment_trend = enrollment_trend[enrollment_trend['Start YearMonth'] >=
'2020-01']

```

```

# Visualizations

# Plot 1: Total Enrollment by Study Type
plt.figure(figsize=(10, 6))
sns.barplot(x=study_type_enrollment.index, y=study_type_enrollment['sum'],
palette='viridis')
plt.title('Total Planned Enrollment by Study Type')
plt.ylabel('Total Enrollment (Log Scale for better visualization)')
plt.yscale('log')
plt.savefig('enrollment_by_study_type.png')

# Plot 2: Top Conditions by Enrollment
plt.figure(figsize=(10, 6))
sns.barplot(x=top_conditions_enrollment.values,
y=top_conditions_enrollment.index, palette='rocket')
plt.title('Top 15 Conditions by Total Planned Enrollment')
plt.xlabel('Total Enrollment')
plt.savefig('enrollment_by_condition.png')

# Plot 3: Enrollment Trend Over Time
plt.figure(figsize=(12, 6))
sns.lineplot(data=enrollment_trend, x='Start YearMonth', y='sum', marker='o',
label='Total Enrollment')
plt.title('Trend of Total Planned Enrollment (Trials starting each month)')
plt.xticks(rotation=45)
plt.ylabel('Total Enrollment')
plt.tight_layout()
plt.savefig('enrollment_trend_total.png')

plt.figure(figsize=(12, 6))
sns.lineplot(data=enrollment_trend, x='Start YearMonth', y='mean', marker='o',
color='red', label='Average Enrollment per Trial')
plt.title('Trend of Average Planned Enrollment per Trial')
plt.xticks(rotation=45)
plt.ylabel('Average Enrollment')
plt.tight_layout()
plt.savefig('enrollment_trend_avg.png')

print("Enrollment by Study Type Statistics:")
print(study_type_enrollment)
print("\nTop 5 Conditions by Total Enrollment:")
print(top_conditions_enrollment.head(5))

```