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

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

# 1. Intervention Type Analysis
# The 'Interventions' column often starts with the type, e.g., 'Diagnostic
Test: ...' or 'Drug: ...'
def extract_intervention_types(interv_str):
    if pd.isna(interv_str):
        return []
    # Interventions are separated by '|'
    intervs = interv_str.split('|')
    types = []
    for item in intervs:
        if ':' in item:
            types.append(item.split(':')[0].strip())
    return types

df['Intervention_Types'] = df['Interventions'].apply(extract_intervention_types)
all_types = df.explode('Intervention_Types')['Intervention_Types'].value_counts()

# 2. Demographic Analysis: Age and Gender
gender_dist = df['Gender'].value_counts()

# Age is often '18 Years and older (Adult, Older Adult)' or similar.
# Let's see the most common age requirement strings
age_dist = df['Age'].value_counts().head(10)

# 3. Study Results Availability
results_dist = df['Study Results'].value_counts()

# Plots
# Intervention Types
plt.figure(figsize=(10, 6))
sns.barplot(x=all_types.head(10).values, y=all_types.head(10).index,
palette='magma')
plt.title('Top 10 Types of Interventions')
plt.xlabel('Count of Occurrences')
plt.savefig('intervention_types.png')

# Gender Distribution
```

```
plt.figure(figsize=(8, 6))
gender_dist.plot(kind='pie', autopct='%1.1f%%', colors=['skyblue', 'salmon',
'lightgreen'])
plt.title('Gender Eligibility for COVID-19 Trials')
plt.ylabel('')
plt.savefig('gender_distribution.png')

# Output some stats
print("Top Intervention Types:")
print(all_types.head(10))
print("\nGender Eligibility:")
print(gender_dist)
print("\nStudy Results Status:")
print(results_dist)

# 4. Sponsor/Collaborator Analysis
# The 'Sponsor/Collaborators' column is often a list separated by '|'. The first
one is usually the lead sponsor.
df['Lead_Sponsor'] = df['Sponsor/Collaborators'].str.split('|').str[0]
top_sponsors = df['Lead_Sponsor'].value_counts().head(15)

# 5. Study Design: Randomization and Masking
# The 'Study Designs' column contains strings like "Allocation: Randomized|
Intervention Model: Parallel Assignment|Masking: Double (Participant,
Investigator)|Primary Purpose: Treatment"
def extract_design_element(design_str, element_prefix):
    if pd.isna(design_str):
        return 'Unknown'
    elements = design_str.split('|')
    for e in elements:
        if e.strip().startswith(element_prefix):
            return e.split(':')[1].strip()
    return 'Not Specified'

df['Allocation'] = df['Study Designs'].apply(lambda x: extract_design_element(x,
'Allocation'))
df['Masking'] = df['Study Designs'].apply(lambda x: extract_design_element(x,
'Masking'))

allocation_dist = df['Allocation'].value_counts()
masking_dist = df['Masking'].value_counts()

# 6. Age Group Analysis
# Looking for 'Child', 'Adult', 'Older Adult' in the Age column
def categorize_age(age_str):
    if pd.isna(age_str):
```

```
        return 'Unknown'
    tags = []
    if 'Child' in age_str: tags.append('Child')
    if 'Adult' in age_str: tags.append('Adult')
    if 'Older Adult' in age_str: tags.append('Older Adult')

    if not tags: return 'Other'
    return '/'.join(tags)

df['Age_Category'] = df['Age'].apply(categorize_age)
age_cat_dist = df['Age_Category'].value_counts()

# Visualizations
# Top Sponsors
plt.figure(figsize=(10, 8))
sns.barplot(x=top_sponsors.values, y=top_sponsors.index, palette='viridis')
plt.title('Top 15 Lead Sponsors of COVID-19 Trials')
plt.xlabel('Number of Trials')
plt.tight_layout()
plt.savefig('top_sponsors.png')

# Masking Distribution
plt.figure(figsize=(10, 6))
sns.barplot(x=masking_dist.values, y=masking_dist.index, palette='cool')
plt.title('Distribution of Masking in Study Designs')
plt.xlabel('Count')
plt.tight_layout()
plt.savefig('masking_distribution.png')

# Age Category Distribution
plt.figure(figsize=(10, 6))
sns.barplot(x=age_cat_dist.values, y=age_cat_dist.index, palette='autumn')
plt.title('Age Group Eligibility')
plt.xlabel('Number of Trials')
plt.tight_layout()
plt.savefig('age_eligibility.png')

print("Top 10 Lead Sponsors:")
print(top_sponsors.head(10))
print("\nAllocation Distribution:")
print(allocation_dist)
print("\nAge Categories:")
print(age_cat_dist)
```