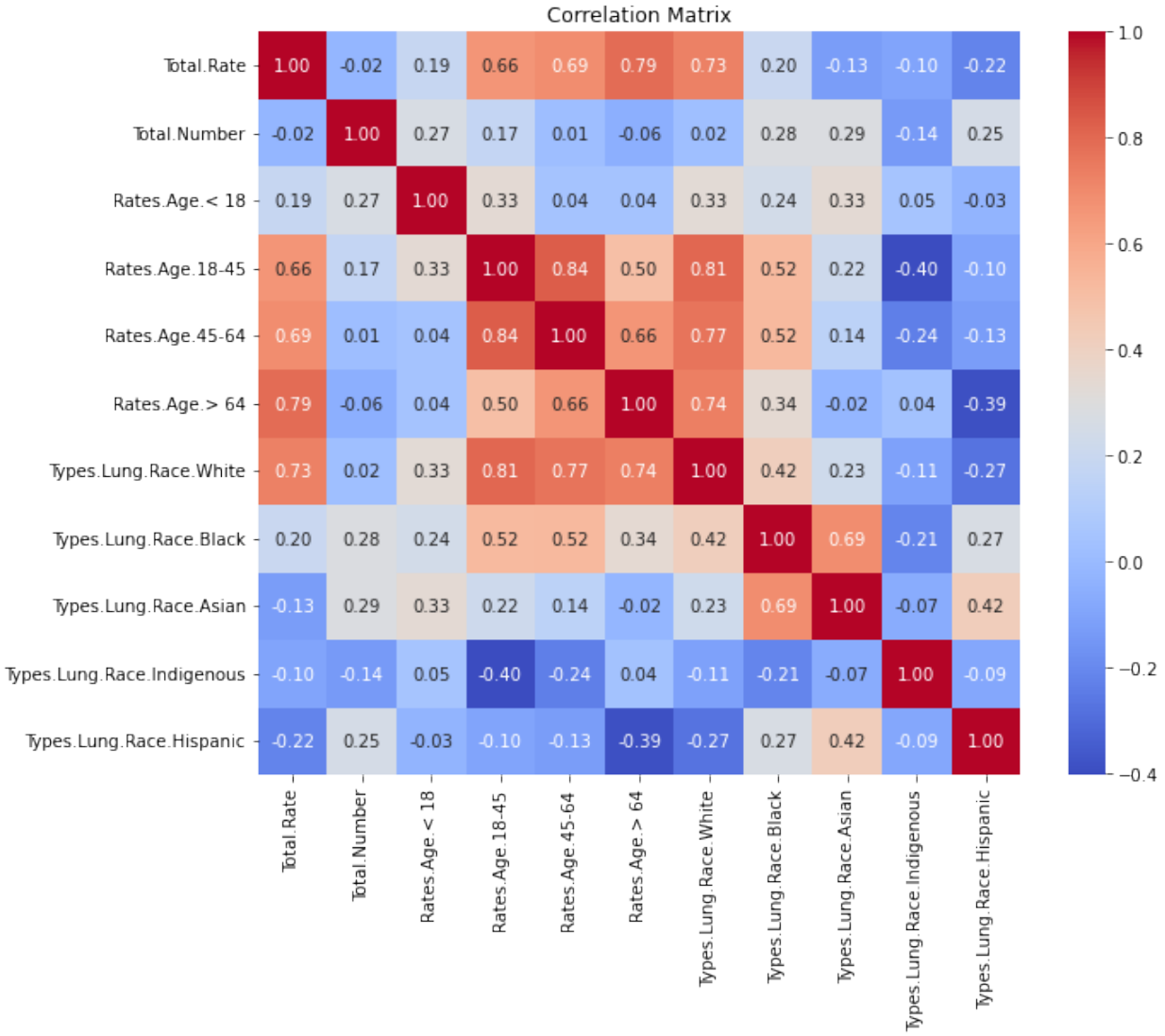
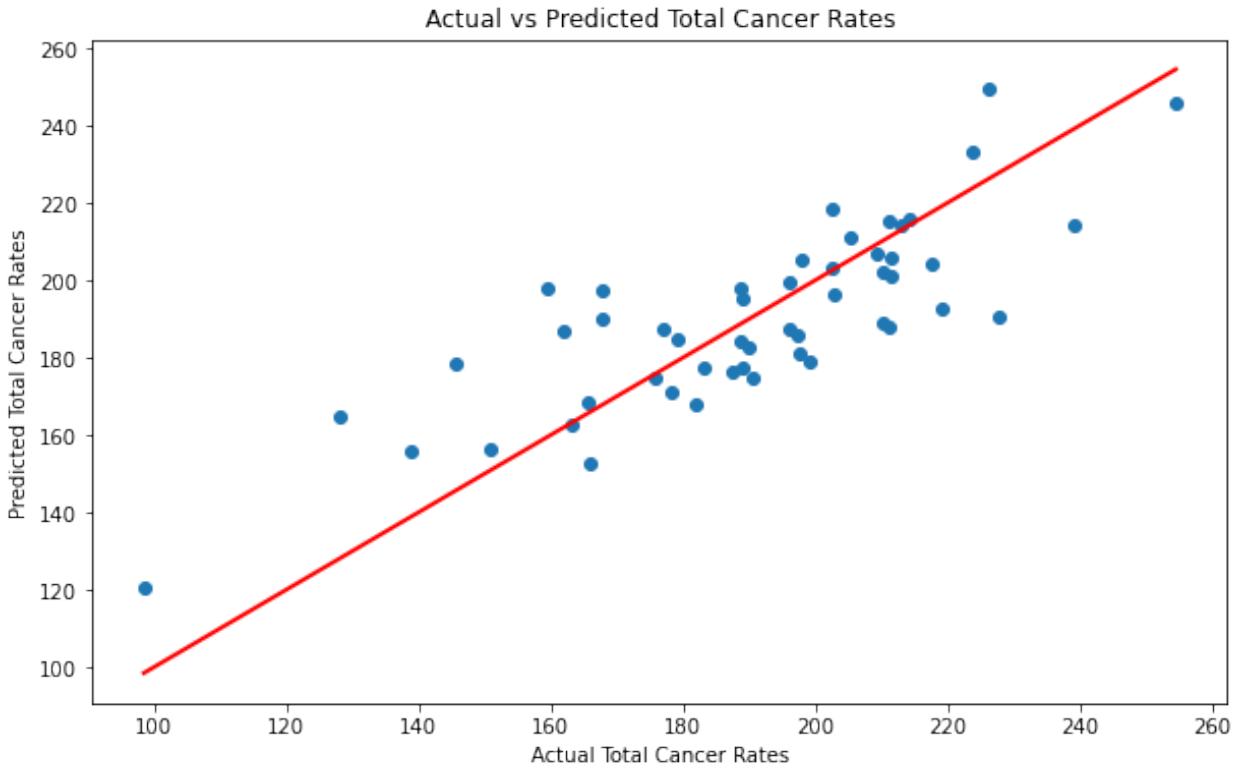
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression

# Load the dataset

cancer_data = pd.read_csv('/Users/khadijaarslan/Desktop/cancer.csv')

# Data Cleaning
# Remove any rows with null values
cancer_data_cleaned = cancer_data.dropna()

# Focusing on columns relevant to race and geographic location
relevant_columns = [
    'State', 'Total.Rate', 'Total.Number', 'Rates.Age.< 18',
'Rates.Age.18-45',
    'Rates.Age.45-64', 'Rates.Age.> 64', 'Types.Lung.Race.White',
    'Types.Lung.Race.Black', 'Types.Lung.Race.Asian',
'Types.Lung.Race.Indigenous',
    'Types.Lung.Race.Hispanic'
]
cancer_data_relevant = cancer_data_cleaned[relevant_columns]

# Statistical Analysis
# Descriptive Statistics
descriptive_stats = cancer_data_relevant.describe()

# Correlation Matrix
correlation_matrix = cancer_data_relevant.corr()

# Regression Analysis
# For simplicity, let's analyze the relationship between Total Rate
and lung cancer rates by race
# Preparing data for regression
X = cancer_data_relevant[['Types.Lung.Race.White',
'Types.Lung.Race.Black', 'Types.Lung.Race.Asian',
'Types.Lung.Race.Indigenous', 'Types.Lung.Race.Hispanic']]
y = cancer_data_relevant['Total.Rate']

# Create and fit the model
model = LinearRegression()
model.fit(X, y)

# Coefficients
coefficients = model.coef_
intercept = model.intercept_

# Predictions
```

```
predictions = model.predict(X)

# Visualizations
# Heatmap for correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
fmt='.2f')
plt.title('Correlation Matrix')
plt.show()

# Scatter plot of actual vs predicted values
plt.figure(figsize=(10, 6))
plt.scatter(y, predictions)
plt.xlabel('Actual Total Cancer Rates')
plt.ylabel('Predicted Total Cancer Rates')
plt.title('Actual vs Predicted Total Cancer Rates')
plt.plot([y.min(), y.max()], [y.min(), y.max()], color='red',
linewidth=2)
plt.show()

# Displaying descriptive statistics
print("Descriptive Statistics:\n", descriptive_stats)
print("\nModel Coefficients:\n", coefficients)
print("\nModel Intercept:\n", intercept)
```


Correlation Matrix
Total.Rate 1.00 -0.02 0.19 0.66 0.69 0.79 0.73 0.20 -0.13 -0.10 -0.22
Total.Number -0.02 1.00 0.27 0.17 0.01 -0.06 0.02 0.28 0.29 -0.14 0.25
Rates.Age.< 18 0.19 0.27 1.00 0.33 0.04 0.04 0.33 0.24 0.33 0.05 -0.03
Rates.Age.18-45 0.66 0.17 0.33 1.00 0.84 0.50 0.81 0.52 0.22 -0.40 -0.10
Rates.Age.45-64 0.69 0.01 0.04 0.84 1.00 0.66 0.77 0.52 0.14 -0.24 -0.13
Rates.Age.> 64 0.79 -0.06 0.04 0.50 0.66 1.00 0.74 0.34 -0.02 0.04 -0.39
Types.Lung.Race.White 0.73 0.02 0.33 0.81 0.77 0.74 1.00 0.42 0.23 -0.11 -0.27
Types.Lung.Race.Black 0.20 0.28 0.24 0.52 0.52 0.34 0.42 1.00 0.69 -0.21 0.27
Types.Lung.Race.Asian -0.13 0.29 0.33 0.22 0.14 -0.02 0.23 0.69 1.00 -0.07 0.42
Types.Lung.Race.Indigenous -0.10 -0.14 0.05 -0.40 -0.24 0.04 -0.11 -0.21 -0.07 1.00 -0.09
Types.Lung.Race.Hispanic -0.22 0.25 -0.03 -0.10 -0.13 -0.39 -0.27 0.27 0.42 -0.09 1.00
Total.Rate
Total.Number
Rates.Age.< 18
Rates.Age.18-45
Rates.Age.45-64
Rates.Age.> 64
Types.Lung.Race.White
Types.Lung.Race.Black
Types.Lung.Race.Asian
Types.Lung.Race.Indigenous
Types.Lung.Race.Hispanic
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

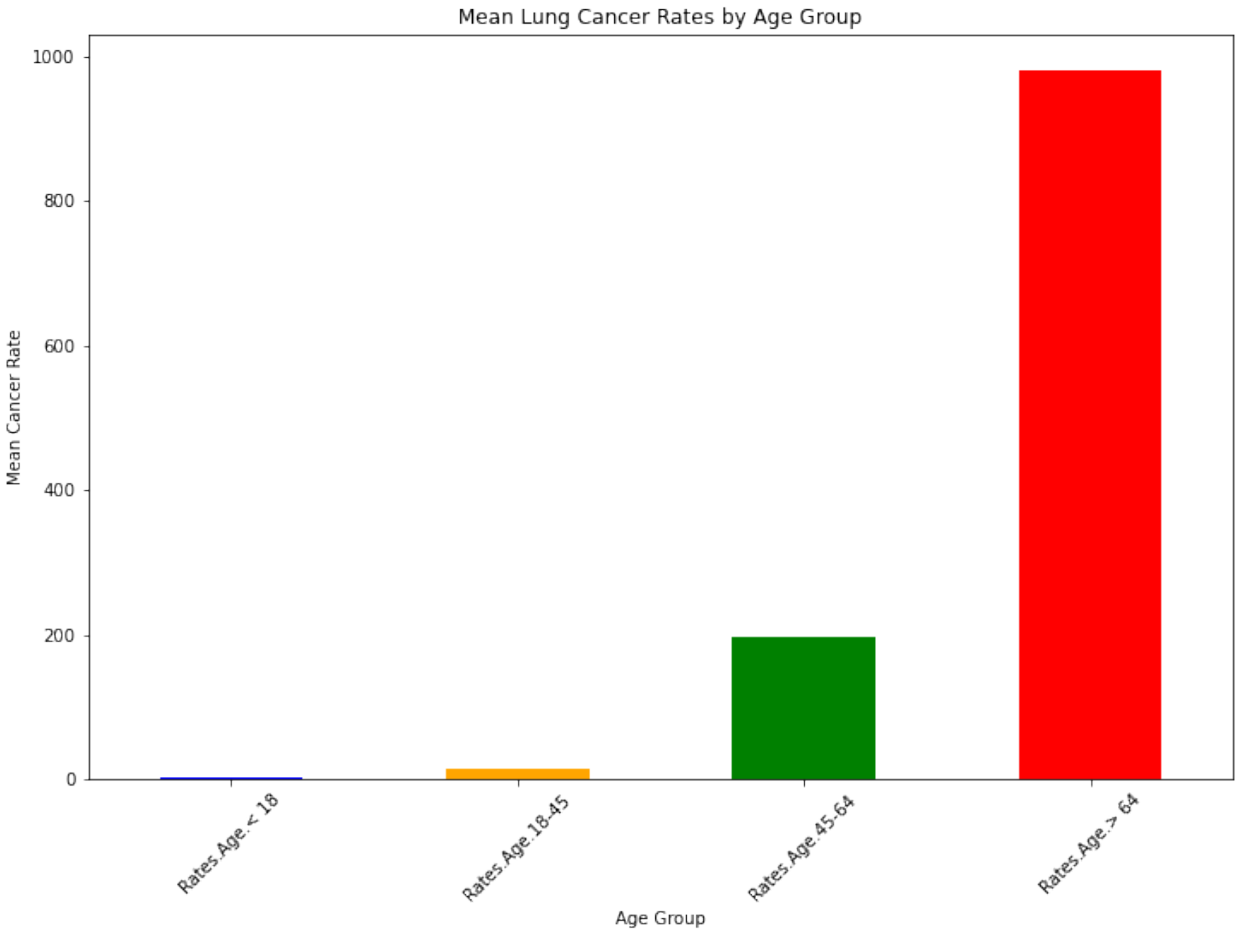



```
Descriptive Statistics:
       Total.Rate   Total.Number  Rates.Age.< 18  Rates.Age.18-45  \
count   51.000000      51.000000       51.000000        51.000000
mean   190.656863   78723.725490        2.119608        14.764706
std     28.592155   80861.292676        0.502004         2.204343
min     98.500000    6361.000000        0.000000        10.000000
25%    176.500000   20631.000000        2.050000        13.350000
50%    196.100000   54930.000000        2.200000        14.600000
75%    210.750000   93328.000000        2.400000        16.250000
max    254.600000  393980.000000        2.700000        20.300000

       Rates.Age.45-64  Rates.Age.> 64  Types.Lung.Race.White  \
count        51.000000       51.000000              51.000000
mean        197.576471      980.949020              48.158824
std          31.261040       75.189556               9.446125
min         132.300000      735.800000              20.400000
25%         175.000000      943.550000              42.750000
50%         189.300000      999.600000              48.100000
75%         217.850000     1031.450000              54.500000
max         263.900000     1110.200000              71.500000

       Types.Lung.Race.Black  Types.Lung.Race.Asian  \
count              51.000000              51.000000
mean               44.056863              19.323529
std                20.517283              10.427801
min                 0.000000               0.000000
```

```
25%                  38.150000                18.350000
50%                  47.000000                22.200000
75%                  60.000000                25.500000
max                  73.200000                33.900000

       Types.Lung.Race.Indigenous  Types.Lung.Race.Hispanic
count                   51.000000                51.000000
mean                    27.309804                16.184314
std                     26.435236                 8.041079
min                      0.000000                 0.000000
25%                      4.450000                13.700000
50%                     20.900000                18.100000
75%                     44.250000                20.450000
max                     87.500000                35.200000

Model Coefficients:
 [ 2.49216661  0.20924566 -1.34004955  0.00758393  0.59416707]

Model Intercept:
 77.48952841384431
```

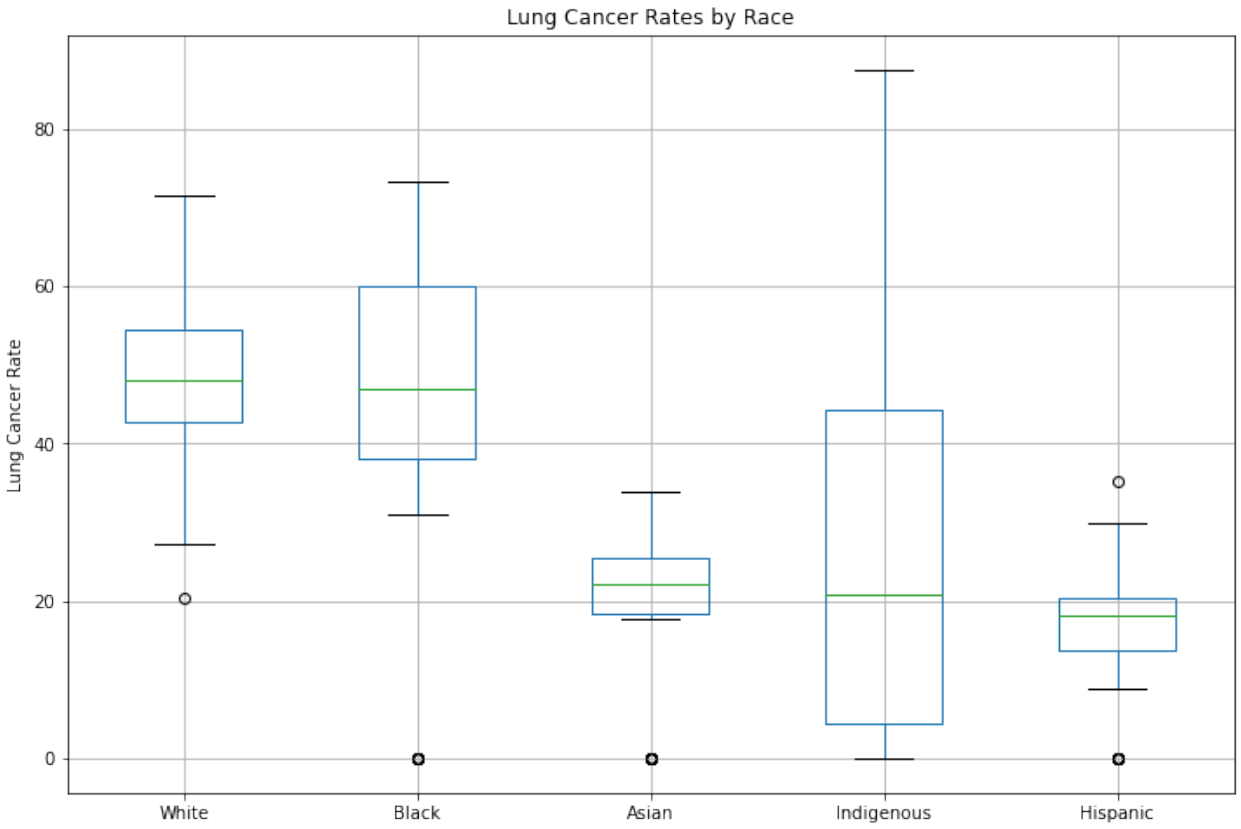
```python
import matplotlib.pyplot as plt

# Focusing on cancer rates by age and race
columns_of_interest = [
    'Rates.Age.< 18', 'Rates.Age.18-45', 'Rates.Age.45-64',
'Rates.Age.> 64',
    'Types.Lung.Race.White', 'Types.Lung.Race.Black',
'Types.Lung.Race.Asian',
    'Types.Lung.Race.Indigenous', 'Types.Lung.Race.Hispanic'
]

# Creating a subset of the relevant data
cancer_rates_by_age_race = cancer_data_relevant[columns_of_interest]

# Calculate the mean cancer rates for each category
mean_cancer_rates = cancer_rates_by_age_race.mean()

# Plotting the trends in cancer rates by age
plt.figure(figsize=(12, 8))

# Age-based cancer rates
age_categories = ['Rates.Age.< 18', 'Rates.Age.18-45', 'Rates.Age.45-
64', 'Rates.Age.> 64']
mean_age_rates = mean_cancer_rates[age_categories]
mean_age_rates.plot(kind='bar', color=['blue', 'orange', 'green',
'red'])

plt.title('Mean Lung Cancer Rates by Age Group')
plt.xlabel('Age Group')
```

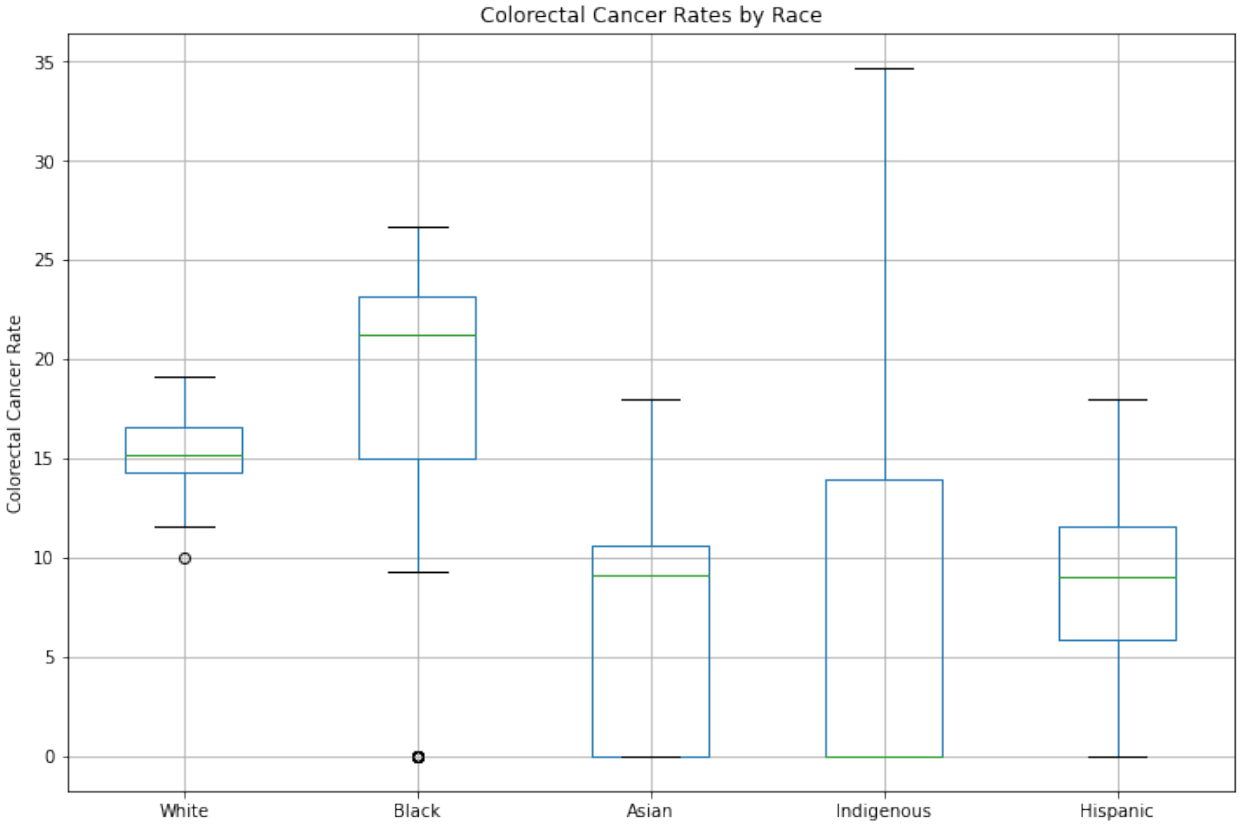
```
plt.ylabel('Mean Cancer Rate')
plt.xticks(rotation=45)
plt.show()
```

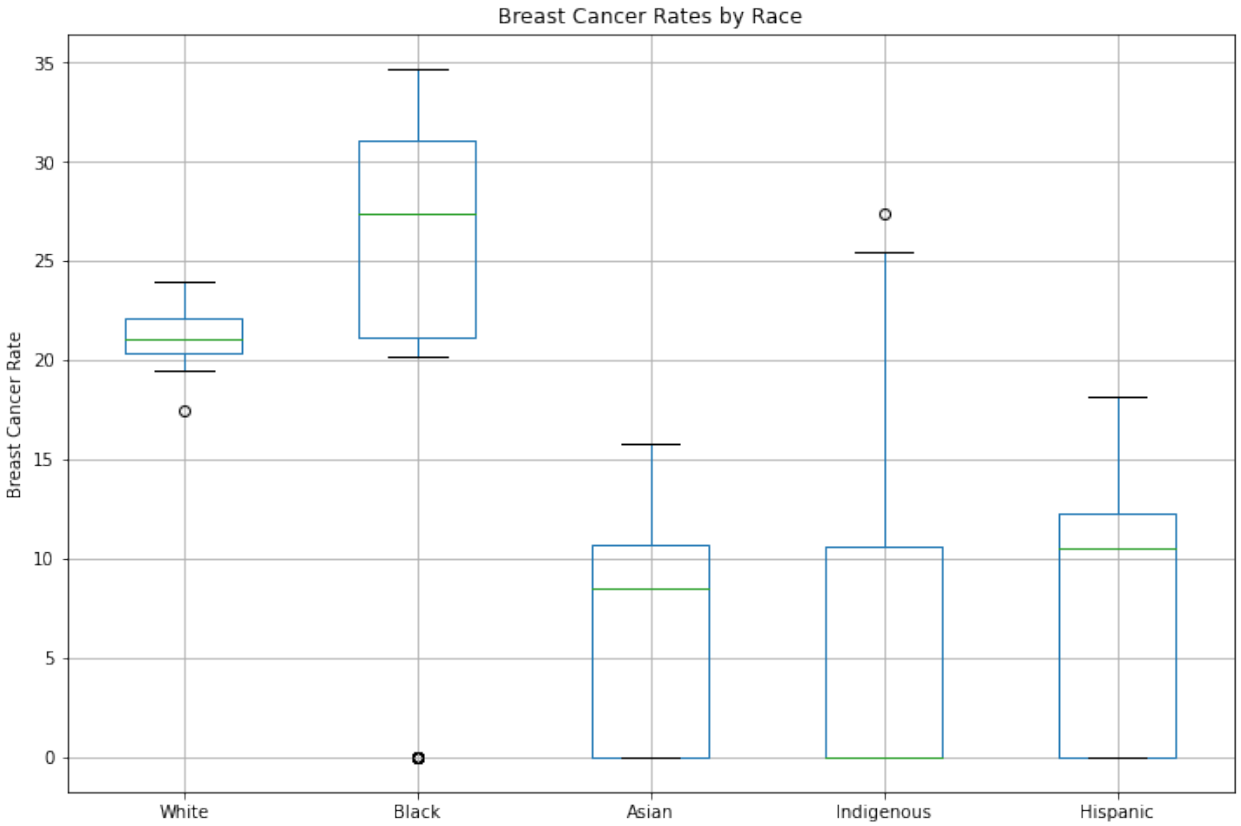



```
lung_cancer_columns = ['Types.Lung.Race.White',
'Types.Lung.Race.Black', 'Types.Lung.Race.Asian',
'Types.Lung.Race.Indigenous', 'Types.Lung.Race.Hispanic']
colorectal_cancer_columns = ['Types.Colorectal.Race.White',
'Types.Colorectal.Race.Black', 'Types.Colorectal.Race.Asian',
'Types.Colorectal.Race.Indigenous', 'Types.Colorectal.Race.Hispanic']
breast_cancer_columns = ['Types.Breast.Race.White',
'Types.Breast.Race.Black', 'Types.Breast.Race.Asian',
'Types.Breast.Race.Indigenous', 'Types.Breast.Race.Hispanic']

# Create boxplots for each cancer type by race

# Lung Cancer
plt.figure(figsize=(12, 8))
cancer_data[lung_cancer_columns].boxplot()
plt.title('Lung Cancer Rates by Race')
```

```
plt.ylabel('Lung Cancer Rate')
plt.xticks([1, 2, 3, 4, 5], ['White', 'Black', 'Asian', 'Indigenous',
'Hispanic'])
plt.show()

# Colorectal Cancer
plt.figure(figsize=(12, 8))
cancer_data[colorectal_cancer_columns].boxplot()
plt.title('Colorectal Cancer Rates by Race')
plt.ylabel('Colorectal Cancer Rate')
plt.xticks([1, 2, 3, 4, 5], ['White', 'Black', 'Asian', 'Indigenous',
'Hispanic'])
plt.show()

# Breast Cancer
plt.figure(figsize=(12, 8))
cancer_data[breast_cancer_columns].boxplot()
plt.title('Breast Cancer Rates by Race')
plt.ylabel('Breast Cancer Rate')
plt.xticks([1, 2, 3, 4, 5], ['White', 'Black', 'Asian', 'Indigenous',
'Hispanic'])
plt.show()
```

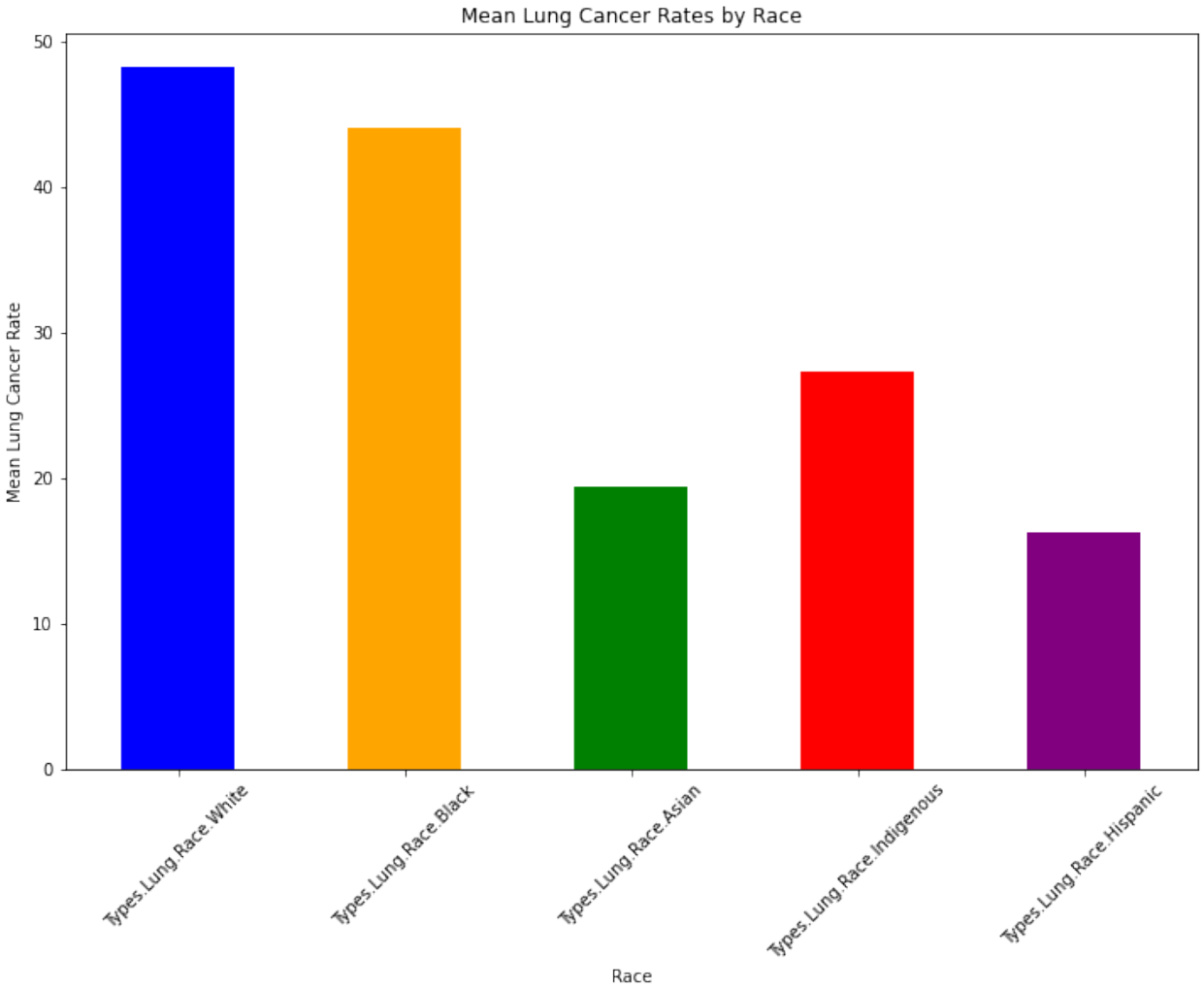


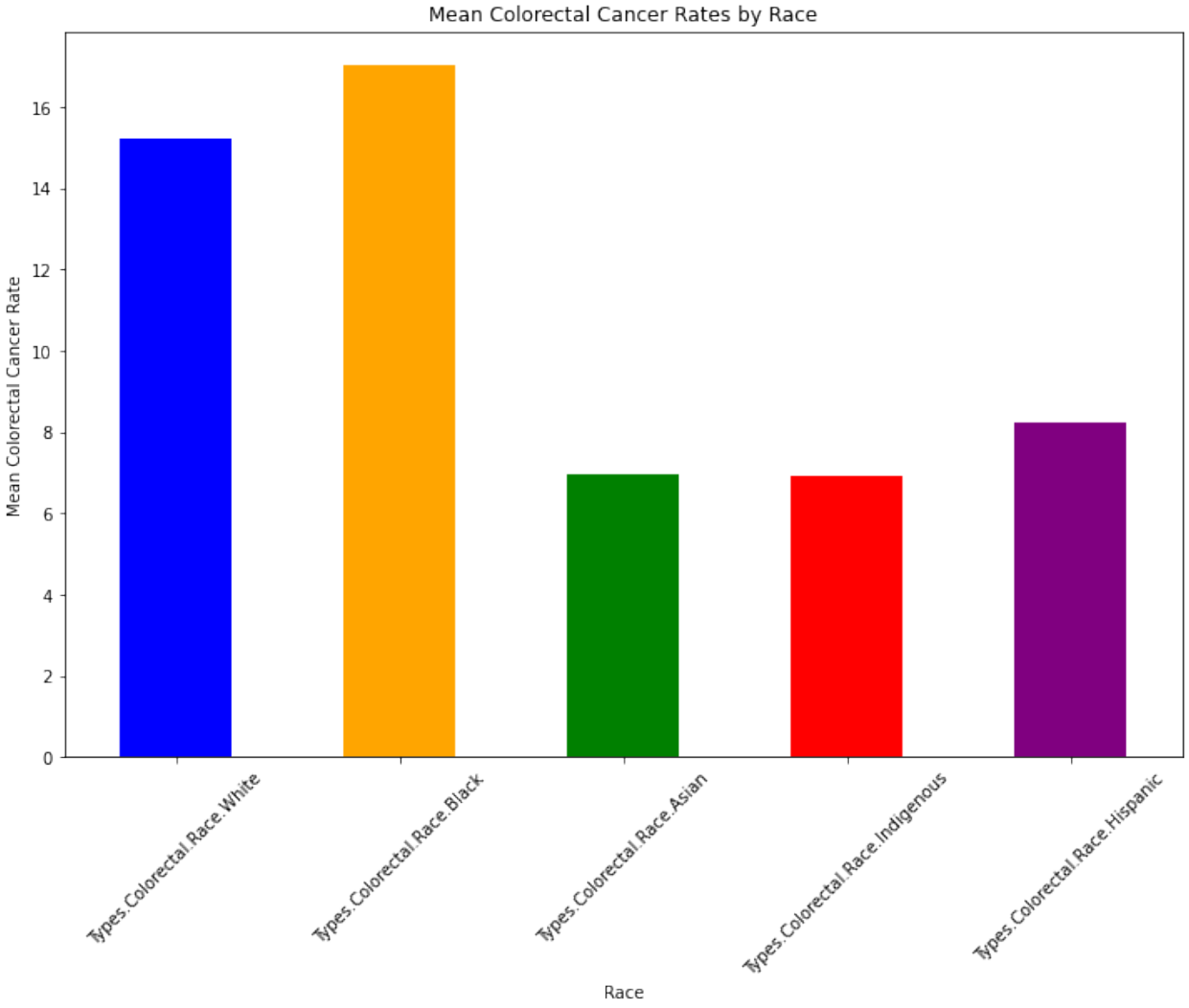

Colorectal Cancer Rates by Race
Colorectal Cancer Rate
0
5
10
15
20
25
30
35
White
Black
Asian
Indigenous
Hispanic

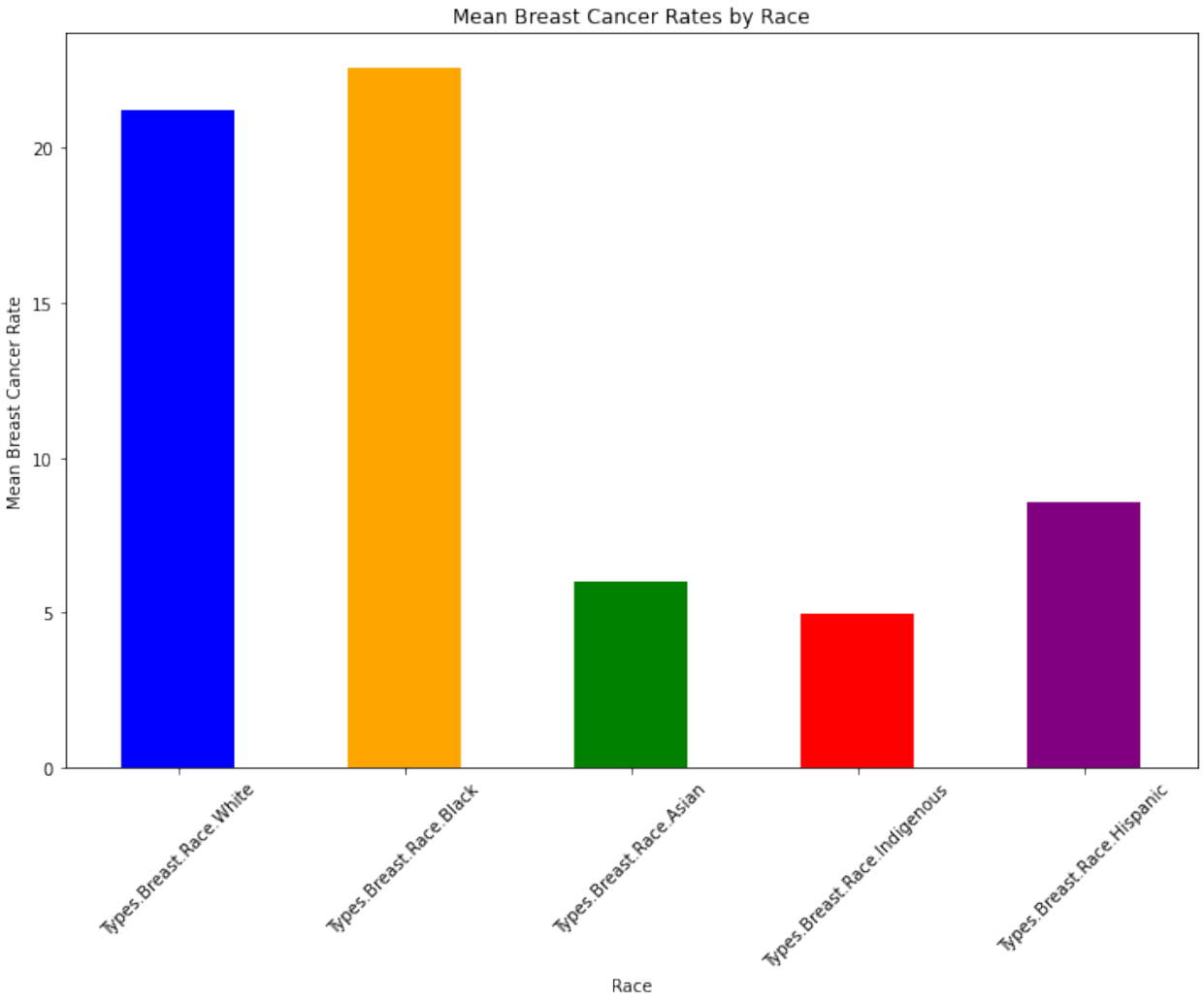

Breast Cancer Rates by Race
Breast Cancer Rate
0
5
10
15
20
25
30
35
White
Black
Asian
Indigenous
Hispanic

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



# Define column names for different types of cancer by race
lung_cancer_columns = ['Types.Lung.Race.White',
'Types.Lung.Race.Black', 'Types.Lung.Race.Asian',
'Types.Lung.Race.Indigenous', 'Types.Lung.Race.Hispanic']
colorectal_cancer_columns = ['Types.Colorectal.Race.White',
'Types.Colorectal.Race.Black', 'Types.Colorectal.Race.Asian',
'Types.Colorectal.Race.Indigenous', 'Types.Colorectal.Race.Hispanic']
breast_cancer_columns = ['Types.Breast.Race.White',
'Types.Breast.Race.Black', 'Types.Breast.Race.Asian',
'Types.Breast.Race.Indigenous', 'Types.Breast.Race.Hispanic']

# Calculate mean cancer rates by race
mean_lung_cancer_rates = cancer_data[lung_cancer_columns].mean()
mean_colorectal_cancer_rates =
cancer_data[colorectal_cancer_columns].mean()
mean_breast_cancer_rates = cancer_data[breast_cancer_columns].mean()

# Create bar charts for mean cancer rates by race

# Mean Lung Cancer Rates by Race
plt.figure(figsize=(12, 8))
mean_lung_cancer_rates.plot(kind='bar', color=['blue', 'orange',
'green', 'red', 'purple'])
plt.title('Mean Lung Cancer Rates by Race')
plt.xlabel('Race')
plt.ylabel('Mean Lung Cancer Rate')
plt.xticks(rotation=45)
plt.show()

# Mean Colorectal Cancer Rates by Race
plt.figure(figsize=(12, 8))
mean_colorectal_cancer_rates.plot(kind='bar', color=['blue', 'orange',
'green', 'red', 'purple'])
plt.title('Mean Colorectal Cancer Rates by Race')
plt.xlabel('Race')
plt.ylabel('Mean Colorectal Cancer Rate')
plt.xticks(rotation=45)
plt.show()

# Mean Breast Cancer Rates by Race
plt.figure(figsize=(12, 8))
mean_breast_cancer_rates.plot(kind='bar', color=['blue', 'orange',
'green', 'red', 'purple'])
plt.title('Mean Breast Cancer Rates by Race')
plt.xlabel('Race')
```

```
plt.ylabel('Mean Breast Cancer Rate')
plt.xticks(rotation=45)
plt.show()
```

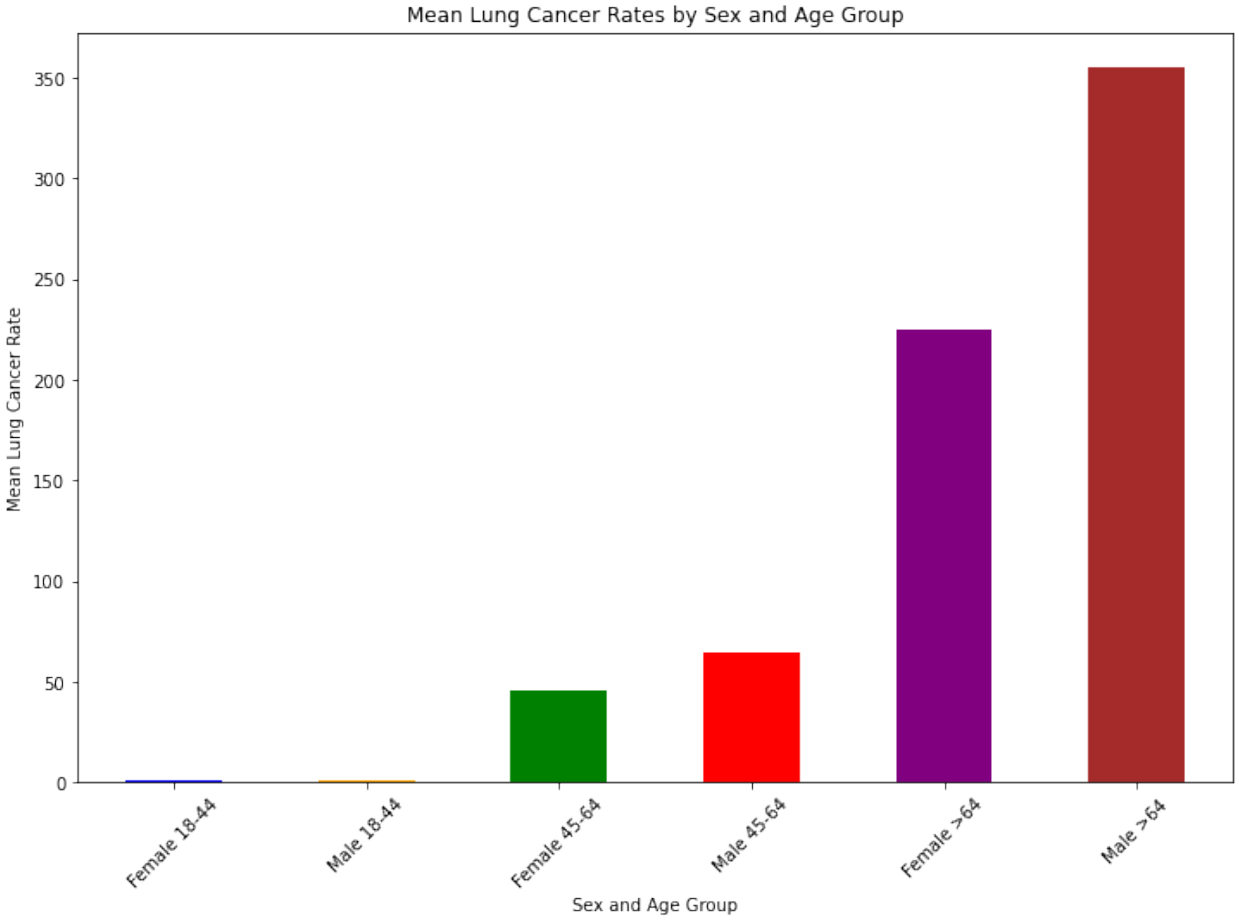


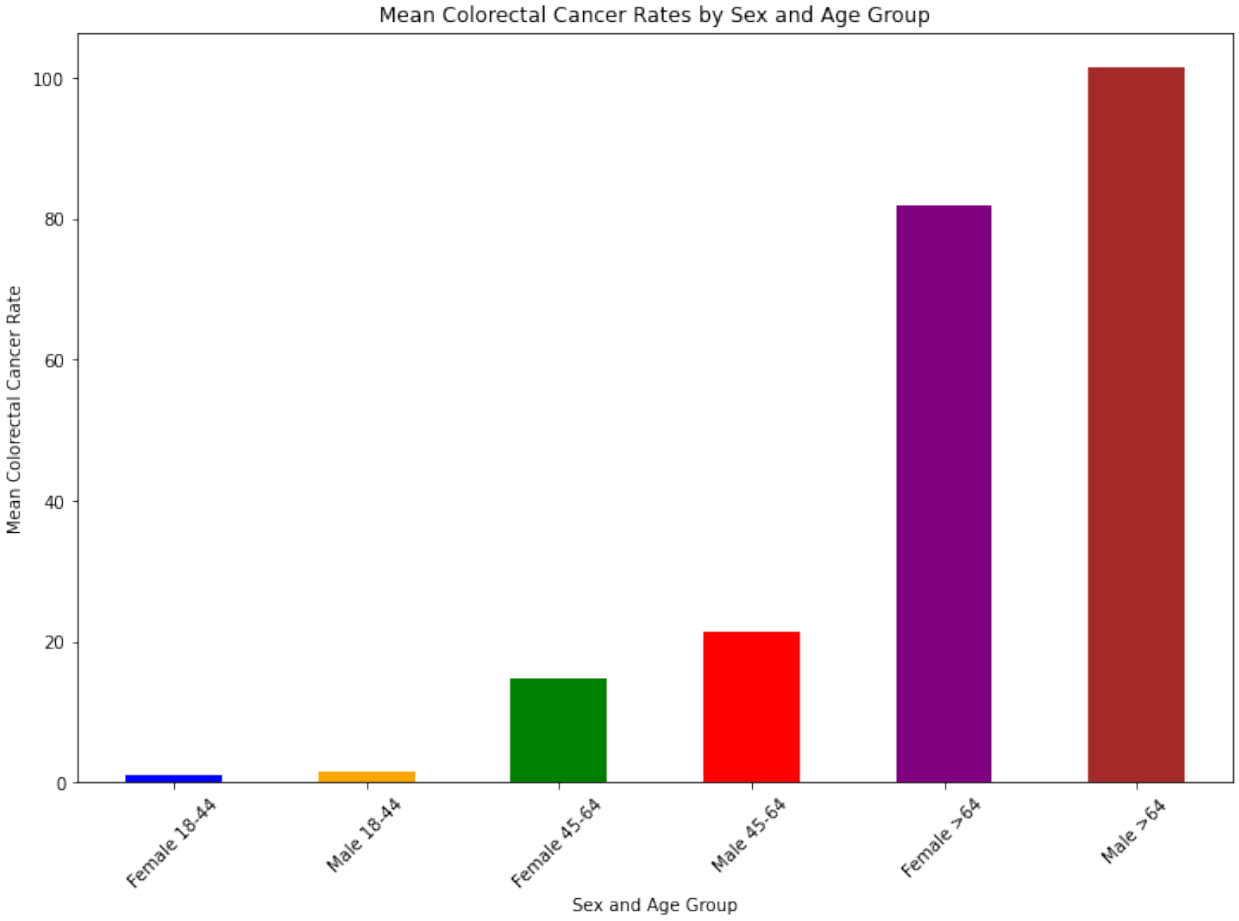

Mean Colorectal Cancer Rates by Race
Mean Colorectal Cancer Rate
0
2
4
6
8
10
12
14
16
Types.Colorectal.Race.White
Types.Colorectal.Race.Black
Types.Colorectal.Race.Asian
Types.Colorectal.Race.Indigenous
Types.Colorectal.Race.Hispanic
Race

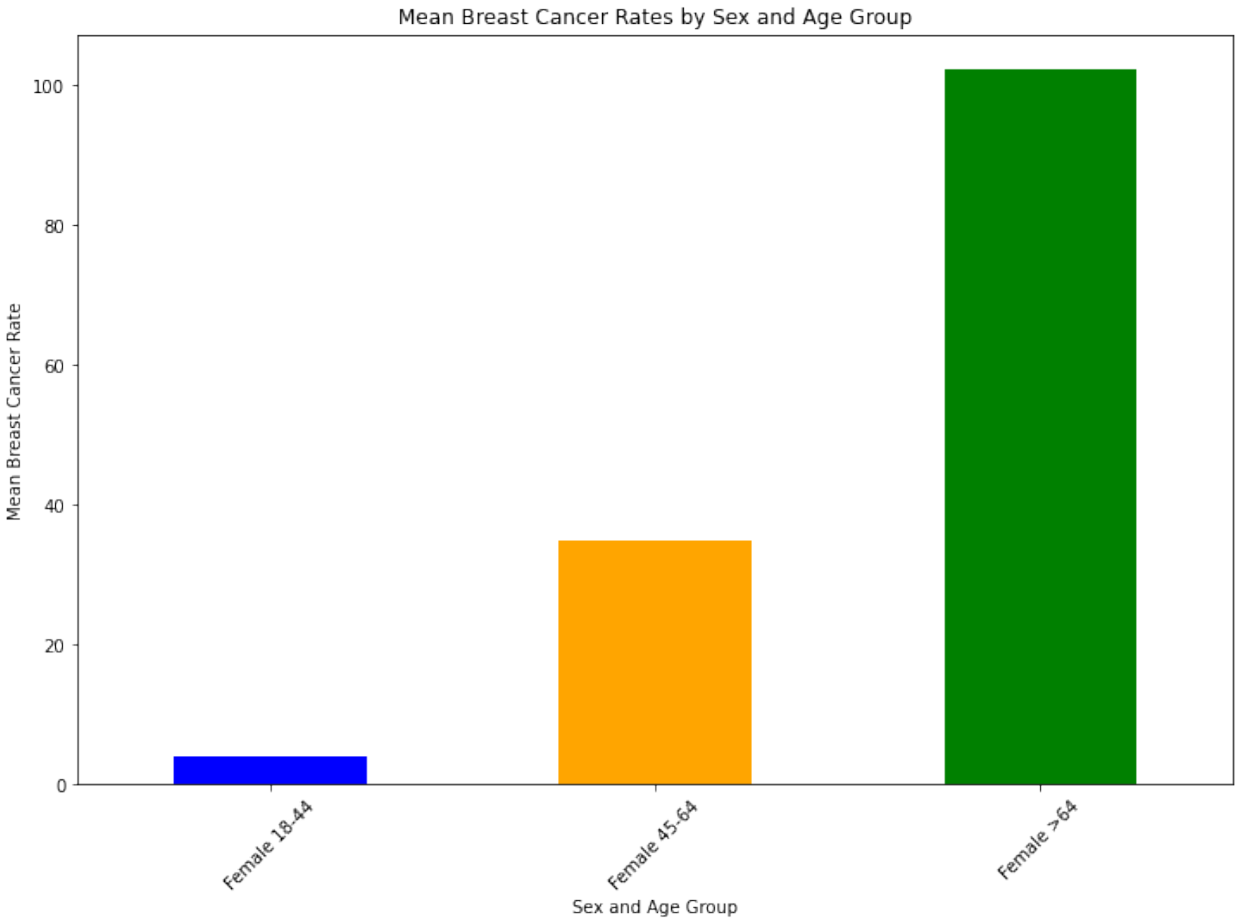



```
import pandas as pd
import matplotlib.pyplot as plt



# Define column names for different types of cancer by sex and age
lung_cancer_sex_columns = ['Types.Lung.Age and Sex.Female.18 - 44',
'Types.Lung.Age and Sex.Male.18 - 44',
                           'Types.Lung.Age and Sex.Female.45 - 64',
'Types.Lung.Age and Sex.Male.45 - 64',
                           'Types.Lung.Age and Sex.Female.> 64',
'Types.Lung.Age and Sex.Male.> 64']
colorectal_cancer_sex_columns = ['Types.Colorectal.Age and
Sex.Female.18 - 44', 'Types.Colorectal.Age and Sex.Male.18 - 44',
                                 'Types.Colorectal.Age and
Sex.Female.45 - 64', 'Types.Colorectal.Age and Sex.Male.45 - 64',
                                 'Types.Colorectal.Age and
Sex.Female.> 64', 'Types.Colorectal.Age and Sex.Male.> 64']
```

```
breast_cancer_sex_columns = ['Types.Breast.Age.18 - 44',
'Types.Breast.Age.45 - 64', 'Types.Breast.Age.> 64']

# Calculate mean cancer rates by sex for each type of cancer
mean_lung_cancer_sex = cancer_data[lung_cancer_sex_columns].mean()
mean_colorectal_cancer_sex =
cancer_data[colorectal_cancer_sex_columns].mean()
mean_breast_cancer_sex = cancer_data[breast_cancer_sex_columns].mean()

# Create bar charts for mean cancer rates by sex for lung, colorectal,
and breast cancer

# Mean Lung Cancer Rates by Sex and Age Group
plt.figure(figsize=(12, 8))
mean_lung_cancer_sex.plot(kind='bar', color=['blue', 'orange',
'green', 'red', 'purple', 'brown'])
plt.title('Mean Lung Cancer Rates by Sex and Age Group')
plt.xlabel('Sex and Age Group')
plt.ylabel('Mean Lung Cancer Rate')
plt.xticks([0, 1, 2, 3, 4, 5], ['Female 18-44', 'Male 18-44', 'Female
45-64', 'Male 45-64', 'Female >64', 'Male >64'], rotation=45)
plt.show()

# Mean Colorectal Cancer Rates by Sex and Age Group
plt.figure(figsize=(12, 8))
mean_colorectal_cancer_sex.plot(kind='bar', color=['blue', 'orange',
'green', 'red', 'purple', 'brown'])
plt.title('Mean Colorectal Cancer Rates by Sex and Age Group')
plt.xlabel('Sex and Age Group')
plt.ylabel('Mean Colorectal Cancer Rate')
plt.xticks([0, 1, 2, 3, 4, 5], ['Female 18-44', 'Male 18-44', 'Female
45-64', 'Male 45-64', 'Female >64', 'Male >64'], rotation=45)
plt.show()

# Mean Breast Cancer Rates by Sex and Age Group
plt.figure(figsize=(12, 8))
mean_breast_cancer_sex.plot(kind='bar', color=['blue', 'orange',
'green'])
plt.title('Mean Breast Cancer Rates by Sex and Age Group')
plt.xlabel('Sex and Age Group')
plt.ylabel('Mean Breast Cancer Rate')
plt.xticks([0, 1, 2], ['Female 18-44', 'Female 45-64', 'Female >64'],
rotation=45)
plt.show()
```


Mean Lung Cancer Rates by Sex and Age Group
Mean Lung Cancer Rate
0
50
100
150
200
250
300
350
Female 18-44
Male 18-44
Female 45-64
Male 45-64
Female >64
Male >64
Sex and Age Group


Mean Colorectal Cancer Rates by Sex and Age Group
Mean Colorectal Cancer Rate
0
20
40
60
80
100
Female 18-44
Male 18-44
Female 45-64
Male 45-64
Female >64
Male >64
Sex and Age Group


Mean Breast Cancer Rates by Sex and Age Group
Mean Breast Cancer Rate
0
20
40
60
80
100
Female 18-44
Female 45-64
Female >64
Sex and Age Group