Debunking Myths: What affects COVID deaths and what doesr

Joseph Wobus, Justin Sher, Megan Kee

Introduction

Coronavirus disease 2019 (COVID-19) is a respiratory disease caused by severe acute respiratory syr
appeared at the end of December in 2019 in Wuhan, China. This virus has its origins in bats and the e
reported to have been linked to a large seafood and live animal market. The virus was originally sugg
to person contact, and as later cases emerged, the virus was suggested to have been spread througl
started in Wuhan, China has become a global pandemic, with an extremely high transmission rate an
Up to date globally, there has been a total of 4.69 million confirmed cases, with 1.72 million recoverec
has the highest number of both confirmed cases and deaths compared to the rest of the world. Com
nations (Russia and the United Kingdom respectively), the United States has a whopping 1.51 million
282,000 confirmed cases, and the United Kingdom has only 243,000 confirmed cases. In this project,
are correlating factors that are causing such high rates in the United States.

With the rise of the COVID-19 Pandemic comes the inevitable rise of the spread of misinformation. In
myths surrounding the COVID-19 crisis and test which demographics are more susceptible to infectio
analysis, and null-hypothesis testing, we set out to prove and disprove several assumptions made ab
COVID-19 death rates. We will look at the following variables: Smoking Rate, Income, and Population
on COVID-19 Deaths.

Import Libraries/Data

```
import pandas as pd
import numpy as np
from plotnine import *
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df = pd.read_csv("COVID19_state.csv")
df = df.set_index("State")
df
```

Analysis

Test 1: Does Smoking Affect COVID-19 Death?

Smoking continues to be one of the leading causes of death around the world. In the United States, tl
per year. As previously mentioned, Coronavirus is a respiratory disease, so one would expect that pec
at greater risk of infection and death. To test whether there was or was not a correlation between sm
smoking rates of each state and compared them to the number of Coronavirus deaths of each state a
plot. We used a linear regression because if there was a correlation, it would intuitively be a linear rela
calculations.

Null Hypothesis: Smoking has no effect

Alternate Hypothesis: There is a significant correlation between smoking and COVID-19 Deaths

a=0.05

```
smoking_df = df.drop(columns = ['Population','Pop Density','Gini','ICU Beds','Income','GDP','
                                'Health Spending','Pollution','Med-Large Airports','Temperatur
```

```
smoking_df
```

```
slope, intercept, r_value, p_value, std_err = stats.linregress(smoking_df['Smoking Rate'], sm
print("slope: " + str(slope))
print("p value: " + str(p_value))
smoking_plot = sns.regplot(x='Smoking Rate', y = 'Deaths', data = smoking_df)
smoking_plot.set_title("Smoking rate vs COVID-19 Deaths")
plt.show(smoking_plot)
```

Since the p-value of 0.115 is greater than 0.05, we failed to reject the Null Hypothesis and thus we ca
correlation between smoking rate and COVID-19 Deaths

Test 2: Does Income Affect COVID-19 Death?

Since income is a great indicator of many things in the United States, we are curious to see if it is an
test, we removed New York from the data set because it was an outlier to get a more accurate repres

Null Hypothesis: Income has no effect

Alternate Hypothesis: There is a significant correlation between incomelink text and COVID-19 Death

a=0.05

```
income_df = df.drop(columns = ['Population','Pop Density','Gini','ICU Beds','GDP','Unemployme
                               'Health Spending','Pollution','Med-Large Airports','Temperatur

#Since New York is a great outlier in our data, we will exclude New York
income_df = income_df.drop("New York", axis=0)
income_df.head()
```

```
slope, intercept, r_value, p_value, std_err = stats.linregress(income_df['Income'], income_df
print("slope: " + str(slope))
print("p value: " + str(p_value))
income_plot = sns.regplot(x='Income', y = 'Deaths', data = income_df)
income_plot.set_title("Income vs COVID-19 Deaths")
plt.show(income_plot)
```

Since the p-value of 0.00038 is less than 0.05, meaning we can reject the null hypothesis and say tha between Income and Deaths. This is interesting because it is showing that a higher income is an indi to the fact that those with higher income live in more populated cities and thus are more exposed to

Test 3: Does Population Density Affect COVID-19 Death?

Since COVID-19 is easily spread through person-to-person contact, it's reasonable to assume that sta lead to more cases of COVID-19 and thus deaths caused by it. In this specific test, we removed D.C. f to get a more accurate representation of the data.

Null Hypothesis: Population density has no effect

Alternate Hypothesis: There is a significant correlation between population density and COVID-19 De

a=0.05

```
pop_den_df = df.drop(columns = ['Income','Population','Gini','ICU Beds','GDP','Unemployment',
                                'Health Spending','Pollution','Med-Large Airports','Temperatur

#Since the District of Columbia is a great outlier in our data, we will exclude the District
pop_den_df = pop_den_df.drop("District of Columbia", axis=0)
pop_den_df
```

```
slope, intercept, r_value, p_value, std_err = stats.linregress(pop_den_df['Pop Density'], pop
print("slope: " + str(slope))
print("p value: " + str(p_value))
pop_den_plot = sns.regplot(x='Pop Density', y = 'Deaths', data = pop_den_df)
pop_den_plot.set_title("Population Density vs COVID-19 Deaths")
plt.show(pop_den_plot)
```

Based on the results of our linear regression, we have found that there is a meaningful relationship b deaths. Our resulting p-value of 0.00158 is way below our threshold of 0.05 thus we are able to reject with our plot leads us to conclude that population density **DOES** increase a person's risk of dying fror because with a higher population density, more people come into contact with each other allowing th

Test 4: Does Pollution Affect COVID-19 Death?

Just like with smoking, pollution is known to negatively affect people's respiratory health. The followi pollution with deaths caused by COVID-19. In this specific test, we removed both New York and New were outliers to get a more accurate representation of the data.

Null Hypothesis: Pollution has no effect

Alternate Hypothesis: There is a significant correlation between pollution levels and COVID-19 Death

a=0.05

```
poll_df = df.drop(columns = ['Income','Population','Gini','ICU Beds','GDP','Unemployment','Se
                               'Health Spending','Pop Density','Med-Large Airports','Temperat

#Since New Jersey and New York are great outliers in our data, we will exclude New Jersey anc
poll_df = poll_df.drop(["New Jersey", "New York"], axis=0)
poll_df
```

```
slope, intercept, r_value, p_value, std_err = stats.linregress(poll_df['Pollution'], poll_df[
print("slope: " + str(slope))
print("p value: " + str(p_value))
poll_plot = sns.regplot(x='Pollution', y = 'Deaths', data = poll_df)
poll_plot.set_title("Pollution vs COVID-19 Deaths")
plt.show(poll_plot)
```

Since our resulting p-value of 0.02 is below our threshold of 0.5, we reject the null hypothesis and prc pollution and COVID-19 deaths. We suspect the elevated levels of pollution negatively affect people's from the Coronavirus.

Standardize the pollution statistics

```
import statistics
poll_df.head()

mean = 0
for row,col in poll_df.iterrows():
  curr_pollution = int(poll_df.loc[row,'Pollution'])
  mean += curr_pollution

mean = mean/len(poll_df)

print(mean)
std_dev = statistics.mean(poll_df['Pollution'].values)

for row,col in poll_df.iterrows():
  curr_pollution = int(poll_df.loc[row,'Pollution'])
  poll_df.loc[row,'Standardized Pollution'] = float((curr_pollution - mean) / float(std_dev))


poll_df.head()
```

```
# slope, intercept, r_value, p_value, std_err = stats.linregress(poll_df['Standardized Pollut

poll_plot = sns.scatterplot(x='Standardized Pollution', y = 'Deaths', data = poll_df)
poll_plot.set_title("Standardized Pollution vs COVID-19 Deaths")
plt.rcParams["figure.figsize"] = (10,5)
plt.show(poll_plot)
```

Predicting the Future

Up until 2016, the EPA has reported massive trends in reduced air pollution in the United States, but v repealed, the percentage of Americans living in areas with higher levels of pollution than the EPA reco years, the Washington Post reports that the concentration of fine-particle air pollution has risen by 5. laws and climate change. If the pandemic has not been dealt with and pollution continues to rise, usi more people will die per state.

```
# Average fine-prticle air pollution percentage rise in the past 2 years
a = 5.5/2

print("Average additional death count due to pollution per state per year")
y = slope * a
print(y)
print("Average additional death total due to pollution per year")
print(y*50)
print("Average deaths per state from COVID-19")
y0 = slope * (mean) + intercept
print(y0)
print("Average death total from COVID-19")
print(y0*50)
print()
print("Expected average deaths per state in 1 year from COVID-19")
y1 = slope * (mean + a) + intercept
print(y1)
print("Expected average death total in 1 year from COVID-19")
print(y1*50)
print()
print("Expected average deaths per state in 2 years from COVID-19")
y2 = slope * (mean + 2*a) + intercept
print(y2)
print("Expected average death total in 2 years from COVID-19")
print(y2*50)
```

```
Average additional death count due to pollution per state per year
760.1876580324166
Average additional death total due to pollution per year
38009.38290162083
Average deaths per state from COVID-19
787.3310533148135
Average death total from COVID-19
39366.55266574067

Expected average deaths per state in 1 year from COVID-19
1547.51871134723
Expected average death total in 1 year from COVID-19
77375.9355673615

Expected average deaths per state in 2 years from COVID-19
2307.7063693796463
Expected average death total in 2 years from COVID-19
115385.31846898231
```

Every year we would add another 760 deaths per state to our count which is another 38,000 America