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 through
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 recovere
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, population dens
influence 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
```

California	991897	67939	2770	39937489	256.3727	0.4899	7338	62586
Colorado	106761	19879	987	5845526	56.4011	0.4586	1597	56846
Connecticut	132508	33765	3008	3563077	735.8689	0.4945	674	74561
District of Columbia	30261	6389	328	720687	11814.5410	0.5420	314	47285
Delaware	31928	6565	225	982895	504.3073	0.4522	186	51449
Florida	561057	40982	1735	21992985	410.1256	0.4852	5604	49417
Georgia	251288	34002	1444	10736059	186.6719	0.4813	2508	45745
Hawaii	35216	634	17	1412687	219.9419	0.4420	201	54565
Iowa	77792	12373	271	3179849	56.9284	0.4451	545	48823
Idaho	32518	2260	70	1826156	22.0969	0.4503	314	43155
Illinois	442425	79007	3459	12659682	228.0243	0.4810	3144	56933
Indiana	146688	24627	1411	6745354	188.2810	0.4527	1861	46646
Kansas	54109	7116	158	2910357	35.5968	0.4550	767	50155
Kentucky	104001	6677	311	4499692	113.9566	0.4813	1392	41779
Louisiana	220830	31815	2242	4645184	107.5175	0.4990	1289	45542
Massachusetts	394728	78462	5108	6976597	894.4355	0.4786	1326	70073
Maryland	164780	33373	1683	6083116	626.6731	0.4499	1134	62914
Maine	23554	1462	65	1345790	43.6336	0.4519	256	48241
Michigan	308233	47552	4584	10045029	177.6655	0.4695	2423	47582
Minnesota	115781	11799	591	5700671	71.5922	0.4496	1171	56374
Missouri	121296	9918	488	6169270	89.7453	0.4646	1888	46635
Mississippi	95885	9674	435	2989260	63.7056	0.4828	824	37994
Montana	22572	459	16	1086759	7.4668	0.4667	165	47120
North Carolina	195865	15045	550	10611862	218.2702	0.4780	2227	45834
North Dakota	47014	1518	36	761723	11.0393	0.4533	238	54306
Nebraska	48019	8572	100	1952570	25.4161	0.4477	440	52110
New Hampshire	35561	3160	133	1371246	153.1605	0.4304	242	61405
New Jersey	425933	139945	9310	8936574	1215.1991	0.4813	1822	67609
New Mexico	106721	5069	208	2096640	17.2850	0.4769	340	41198
Nevada	60084	6152	312	3139658	28.5993	0.4577	900	48225
New York	1204650	337055	21640	19440469	412.5211	0.5229	3952	68667

New York	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Ohio	209153	24777	1357	11747694	287.5038	0.4680	3314	48242
Oklahoma	106559	4613	274	3954821	57.6547	0.4645	1064	46128
Oregon	77542	3286	130	4301089	44.8086	0.4583	659	49908
Pennsylvania	288858	57154	3731	12820878	286.5449	0.4689	3169	55349
Rhode Island	11633	2256	113	1056161	1021.4323	0.4781	279	54523
South Carolina	93332	11450	430	5210095	173.3174	0.4735	1225	42736
South Dakota	89968	7792	346	903027	11.9116	0.4495	152	50141
Tennessee	24578	3614	34	6897576	167.2748	0.4790	2209	47179
Texas	273277	15544	251	29472295	112.8204	0.4800	6199	49161
Utah	525697	39869	1100	3282115	39.9430	0.4063	565	45340
Virginia	150585	6362	68	8626207	218.4403	0.4705	1654	56952
Vermont	167758	25070	850	628061	68.1416	0.4539	94	53598
Washington	20871	927	53	7797095	117.3272	0.4591	1265	60781
Wisconsin	252108	17122	945	5851754	108.0497	0.4498	1159	50756
West Virginia	118451	10418	409	1778070	73.9691	0.4711	653	40578
Wyoming	64165	1369	57	567025	5.8400	0.4360	102	60095

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
```

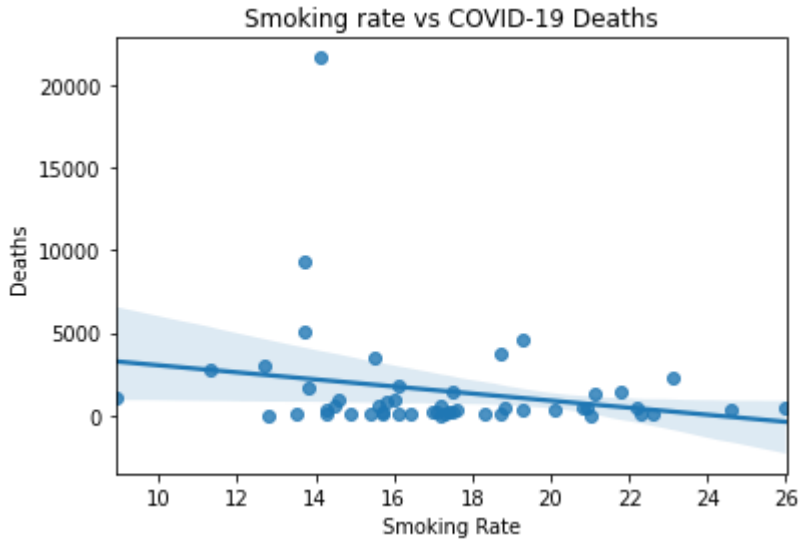
```
smoking_df
```

State	Tested	Infected	Deaths	Smoking Rate
Alaska	28680	381	10	21.0
Alabama	129444	10164	403	20.9
Arkansas	70323	4043	94	22.3
Arizona	150241	11380	542	15.6
California	991897	67939	2770	11.3
Colorado	106761	19879	987	14.6
Connecticut	132508	33765	3008	12.7
District of Columbia	30261	6389	328	14.3
Delaware	31928	6565	225	17.0
Florida	561057	40982	1735	16.1
Georgia	251288	34002	1444	17.5
Hawaii	35216	634	17	12.8
Iowa	77792	12373	271	17.1
Idaho	32518	2260	70	14.3
Illinois	442425	79007	3459	15.5
Indiana	146688	24627	1411	21.8
Kansas	54109	7116	158	17.4
Kentucky	104001	6677	311	24.6
Louisiana	220830	31815	2242	23.1
Massachusetts	394728	78462	5108	13.7
Maryland	164780	33373	1683	13.8
Maine	23554	1462	65	17.3
Michigan	308233	47552	4584	19.3
Minnesota	115781	11799	591	14.5
Missouri	121296	9918	488	20.8
Mississippi	95885	9674	435	22.2
Montana	22572	459	16	17.2
North Carolina	195865	15045	550	17.2

```
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)
```

```
slope: -213.80574316163757
p value: 0.11519081952108609
```



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

Virginia	150585	6362	68	16.4

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

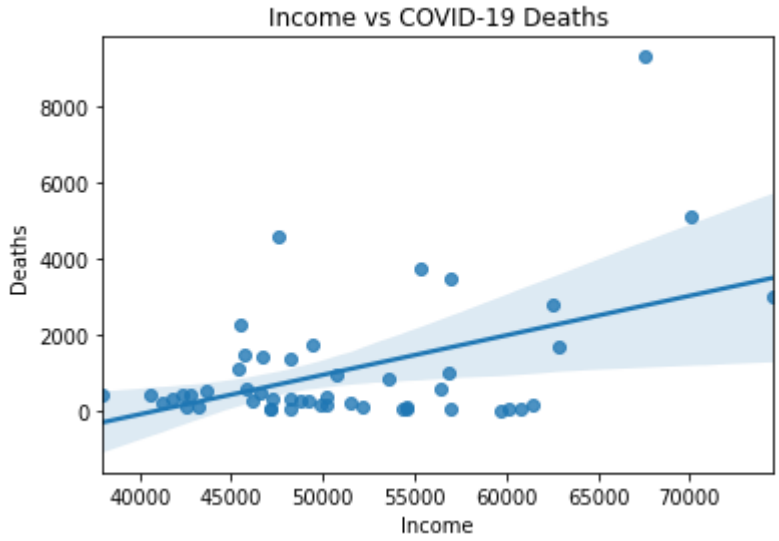
a=0.05

```
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()
```

State	Tested	Infected	Deaths	Income
Alaska	28680	381	10	59687
Alabama	129444	10164	403	42334

```
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)
```

```
slope: 0.10401338392425985
p value: 0.0003797685185062452
```



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

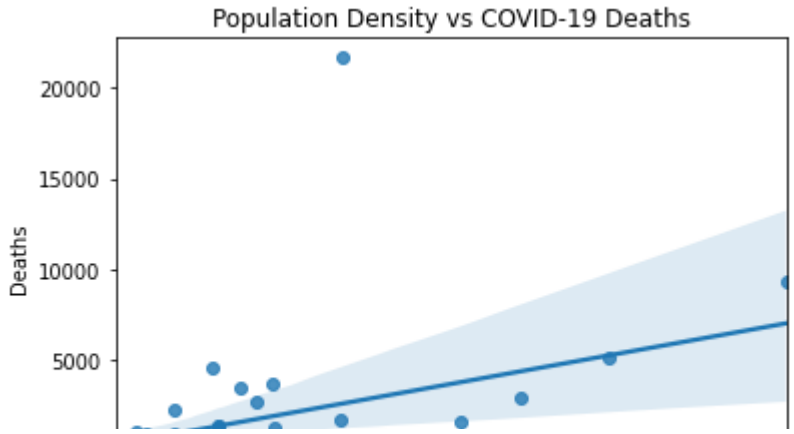
a=0.05

```
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
```

State	Tested	Infected	Deaths	Pop Density
Alaska	28680	381	10	1.2863
Alabama	129444	10164	403	96.9221
Arkansas	70323	4043	94	58.4030
Arizona	150241	11380	542	64.9550
California	991897	67939	2770	256.3727
Colorado	106761	19879	987	56.4011
Connecticut	132508	33765	3008	735.8689
Delaware	31928	6565	225	504.3073
Florida	561057	40982	1735	410.1256
Georgia	251288	34002	1444	186.6719
Hawaii	35216	634	17	219.9419
Iowa	77792	12373	271	56.9284
Idaho	32518	2260	70	22.0969
Illinois	442425	79007	3459	228.0243
Indiana	146688	24627	1411	188.2810
Kansas	54109	7116	158	35.5968
Kentucky	104001	6677	311	113.9566
Louisiana	220830	31815	2242	107.5175
Massachusetts	394728	78462	5108	894.4355
Maryland	164780	33373	1683	626.6731
Maine	23554	1462	65	43.6336
Michigan	308233	47552	4584	177.6655
Minnesota	115781	11709	591	71.5922

```
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)
```

```
slope: 5.486002712103553
p value: 0.0015837574492009985
```



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 fron because with a higher population density, more people come into contact with each other allowing th

Tennessee 24578 3014 54 107.2746

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 Deaths

a=0.05

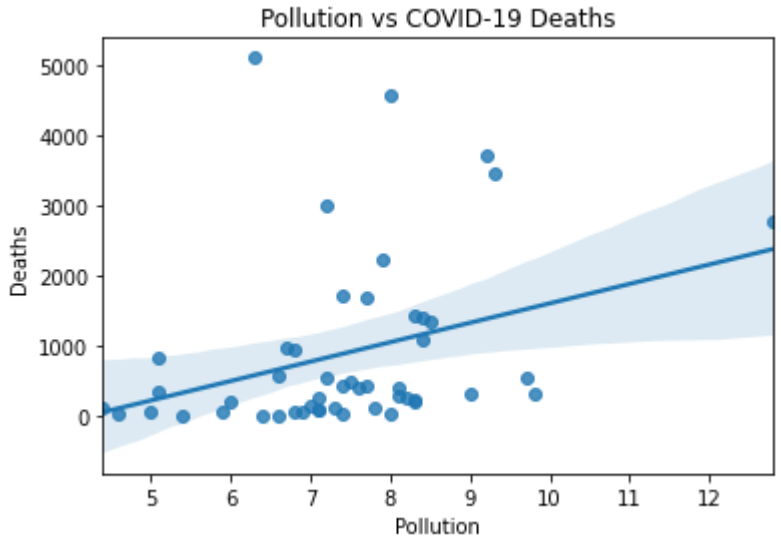
```
poll_df = df.drop(columns = ['Income','Population','Gini','ICU Beds','GDP','Unemployment','Se
                             'Health Spending','Pop Density','Med-Large Airports','Temperat

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

State	Tested	Infected	Deaths	Pollution
Alaska	28680	381	10	6.4
Alabama	129444	10164	403	8.1
Arkansas	70323	4043	94	7.1
Arizona	150241	11380	542	9.7
California	991897	67939	2770	12.8
Colorado	106761	19879	987	6.7
Connecticut	132508	33765	3008	7.2
District of Columbia	30261	6389	328	9.8
Delaware	31928	6565	225	8.3
Florida	561057	40982	1735	7.4
Georgia	251288	34002	1444	8.3
Hawaii	35216	634	17	5.4
Iowa	77792	12373	271	7.1
Idaho	32518	2260	70	6.8
Illinois	442425	79007	3459	9.3
Indiana	146688	24627	1411	8.4
Kansas	54109	7116	158	7.0
Kentucky	104001	6677	311	8.1
Louisiana	220830	31815	2242	7.9
Massachusetts	394728	78462	5108	6.3
Maryland	164780	33373	1683	7.7
Maine	23554	1462	65	5.9
Michigan	308233	47552	4584	8.0
Minnesota	115781	11799	591	6.6
Missouri	121296	9918	488	7.5
Mississippi	95885	9674	435	7.7
Montana	22572	459	16	6.6
North Carolina	195865	15045	550	7.2
North Dakota	47014	1518	36	4.6

Nebraska	48019	8572	100	7.1
New Hampshire	35561	3160	133	4.4
New Mexico	106721	5069	208	6.0
Nevada	60084	6152	312	9.0
Ohio	209153	24777	1357	8.5
Oklahoma	106559	4613	274	8.2
Oregon	77542	3286	130	7.8
Pennsylvania	288858	57154	3731	9.2
Rhode Island	11633	2256	113	7.3
South Carolina	93332	11450	430	7.4
South Dakota	89968	7792	346	5.1
Tennessee	24578	3614	34	7.4
Texas	273277	15544	251	8.3
Utah	525697	39869	1100	8.4

```
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)
```

```
slope: 276.4318756481515
p value: 0.02046859732762195
```



Since our resulting p-value of 0.02 is below our threshold of 0.5, we reject the null hypothesis and pro
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()
```

```
7.020408163265306
```

State	Tested	Infected	Deaths	Pollution	Standardized Pollution
Alaska	28680	381	10	6.4	-0.137589
Alabama	129444	10164	403	8.1	0.132086
Arkansas	70323	4043	94	7.1	-0.002752
Arizona	150241	11380	542	9.7	0.266924
California	991897	67939	2770	12.8	0.671436

```
# 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)
```