COVID-19
Data Analysis Project

Blue Team
Nicole Gee, Nicholas Richmond, Matthew Clark, Swati Kohli, Joseph Carozza

Introduction

Project Outline

Step 1:

Data Collection

Step 2:

Data Cleaning

Step 3:

Descriptive Analysis

Step 4:

Regression Analysis

A. Cases
B. Deaths

Step 5:

Recommendations and Conclusion

Step 1: Data Collection

- The team brainstormed about different variables that may affect the number of COVID cases and deaths.
- Various sources of data were researched and added to a central document
- Those data files were downloaded or pulled directly from web pages to be cleaned and combined in R

Step 2: Combining and Cleaning Data

- All of the .csv and .xlsx files were loaded into the environment in RStudio
- Country names and labels were ‘cleaned’ to match the primary JHU data source
- Finally, desired variables from the researched data sets were joined with the JHU data

Final Data Set

Finance & Sustainability

GDP/Capita (USD)

Median Household Income

Median Percapita Income

Per Capita Spending on Health

R&D Expenditure (% of GDP)

Population & Trends

Population 2020

% Urban Population

People 65 & Over

Median Age

Proportion of Female Population

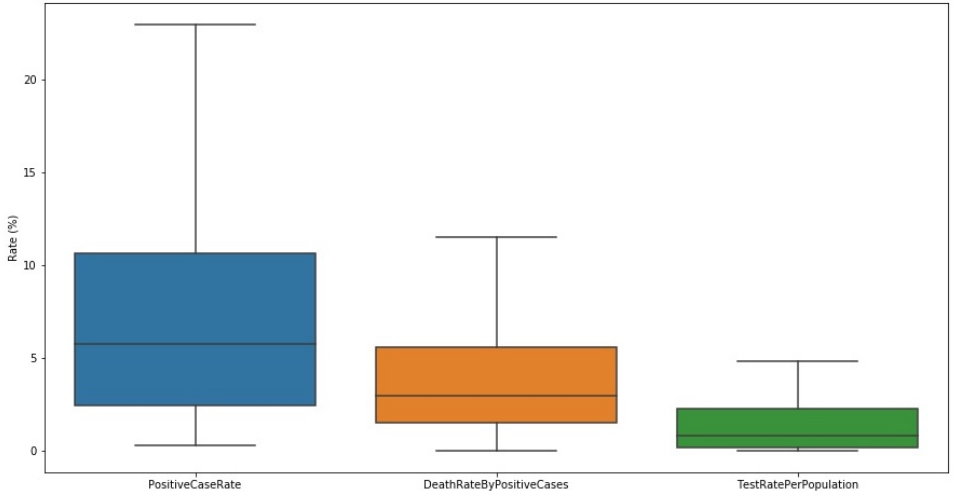
Smoking prevalence (% of adults)

Human & Capital Mobility

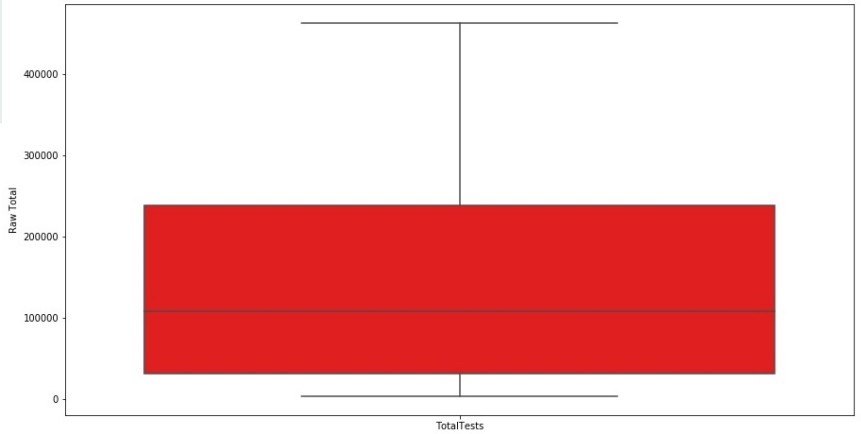
International Inbound Tourists(thousands)

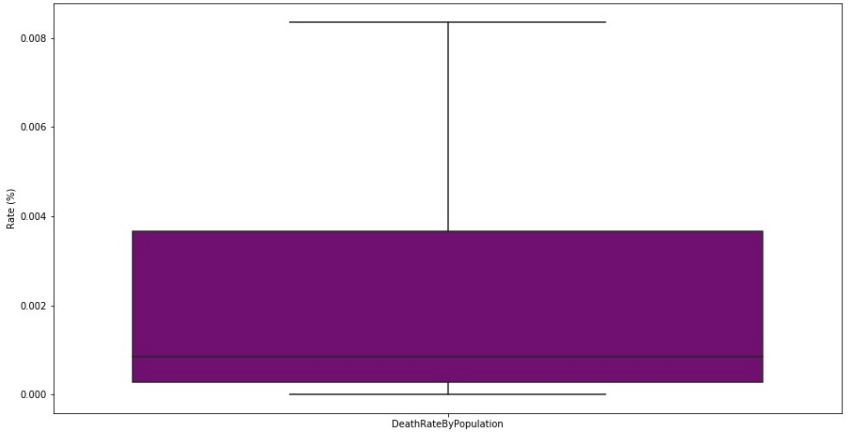
Import/Export (% of GDP)

Step 3: Descriptive Analytics

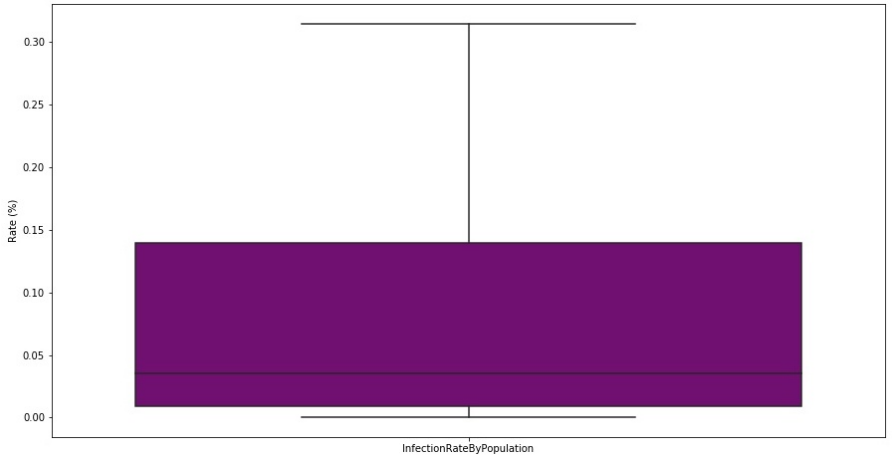


Rate Field	Calculation Method
InfectionRateByPopulation	Total Cases / Total Population
DeathRateByPopulation	Total Deaths / Total Population
DeathRateByPositiveCases	Total Deaths / Total Cases
PositiveCasesPerTestsGiven	Total Cases / Total Tests
TestRatePerPopulation	Total Tests / Total Population

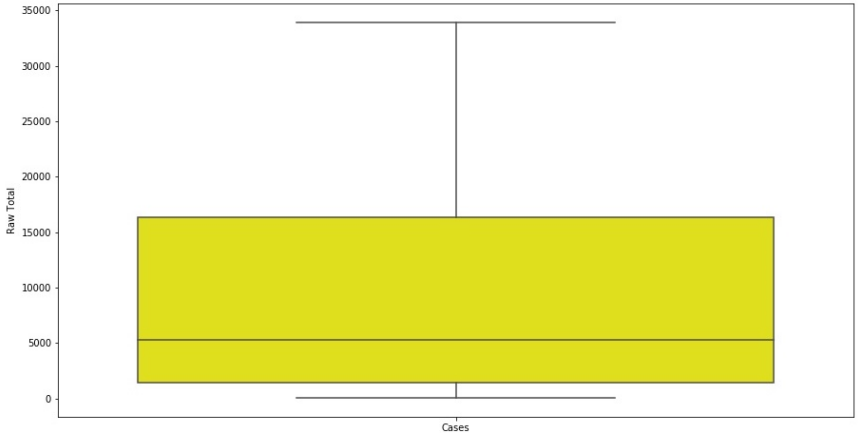




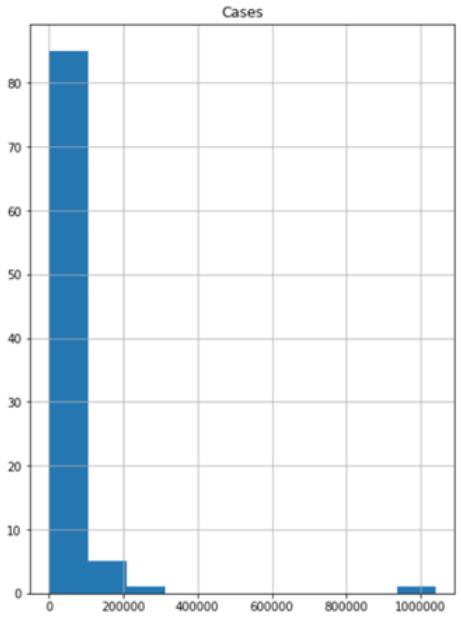
Raw (Total) Cases & Tests

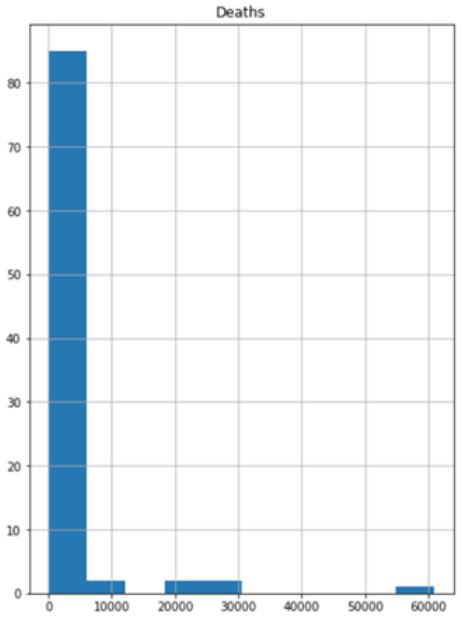


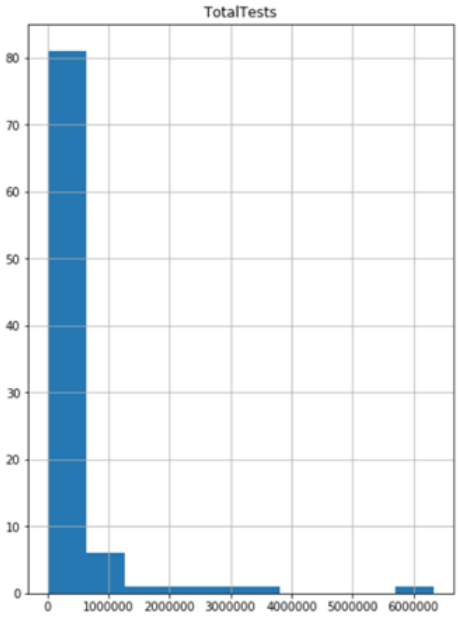
Calculated (Rate) Fields

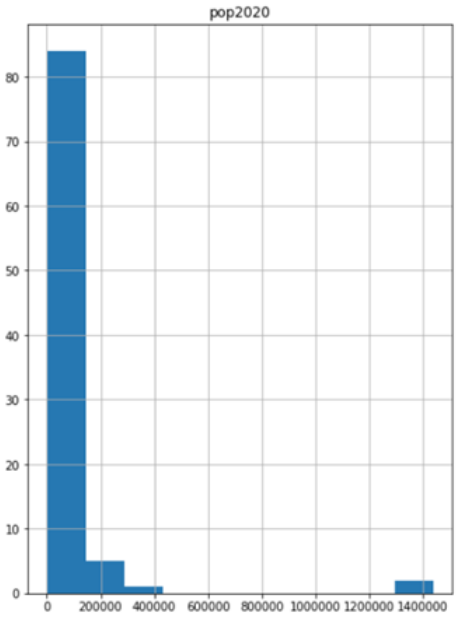


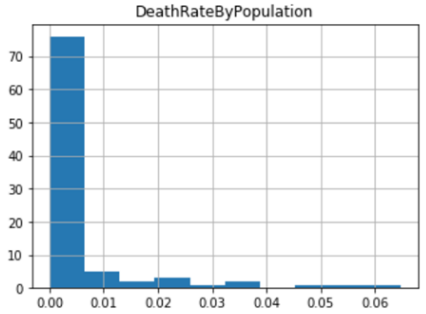
Histogram Plots

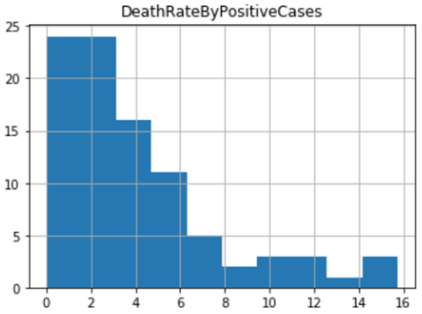


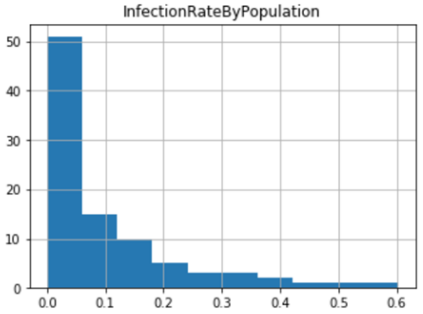


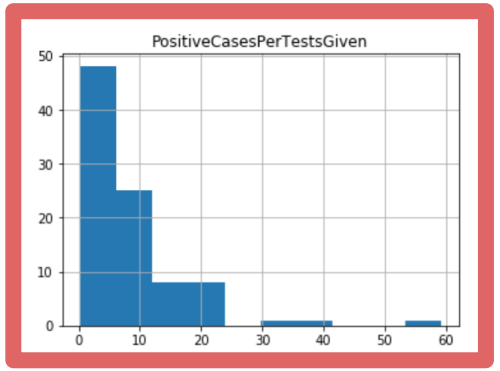


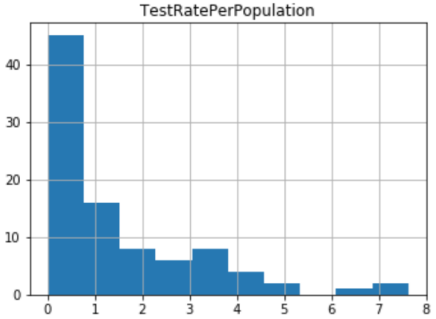












Correlations

Pt. 1

Financial Variables

1. Household Income
2. Per Capita Income
3. GDP
4. Healthcare Spending
5. R&D Spending
6. Import/Export %
7. International Students

Financial Variables

	pop2020Total	TotalTests	Cases	Deaths	TestRatePerPopulation	InfectionRateByPopulation	PositiveCasesPerTestsGiven	DeathRateByPopulation	DeathRateByPositiveCases	medianHouseholdIncome	medianPerCapitaIncome	USGDPCapita	PerCapitaSpendingonHealth	R&DExpenditure%ofGDP	ImportExport%ofGDP	InternationalinboundTourists
pop2020Total	1	0.26	0.19	0.15	-0.19	-0.12	0.031	-0.05	0.071	-0.14	-0.12	-0.12	-0.072	0.067	-0.24	0.32
TotalTests	0.26	1	0.9	0.79	0.13	0.31	0.083	0.3	0.16	0.27	0.3	0.24	0.41	0.32	-0.2	0.64
Cases	0.19	0.9	1	0.93	0.072	0.36	0.2	0.37	0.25	0.27	0.29	0.25	0.46	0.27	-0.18	0.65
Deaths	0.15	0.79	0.93	1	0.082	0.45	0.26	0.6	0.47	0.29	0.32	[illegible]	[illegible]	[illegible]	[illegible]	0.74
TestRatePerPopulation	-0.19	0.13	0.072	0.082	1	0.58	-0.18	0.23	-0.11	0.58	0.52	0.63	0.47	0.29	0.56	0.046
InfectionRateByPopulation	-0.12	0.31	0.36	0.45	0.58	1	0.28	0.73	0.33	0.6	0.58	0.77	0.65	0.4	0.43	0.38
PositiveCasesPerTestsGiven	0.031	0.083	0.2	0.26	-0.18	0.28	1	0.33	0.52	0.034	0.036	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DeathRateByPopulation	-0.05	0.3	0.37	0.6	0.23	0.73	0.33	1	0.73	0.4	0.45	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DeathRateByPositiveCases	0.071	0.16	0.25	0.47	-0.11	0.33	0.52	0.73	1	0.19	0.28	0.17	0.3	0.27	-0.085	0.44

GDP per Capita vs. Test Rate per Population

GDP per Capita vs. Infection Rate per Population

Correlations

Pt. 2

Demographic Variables:

1. Median Age
2. People Over 65
3. Urban Population %
4. Female Population %
5. Smoking Prevalence

Demographic Variables

	pop2020Total	TotalTests	Cases	Deaths	TestRatePerPopulation	InfectionRateByPopulation	PositiveCasesPerTestsGiven	DeathRateByPopulation	DeathRateByPositiveCases	MedianAge	People65andOverTotal	UrbanPop%	PopProportionFemale	SmokingPrevalence%ofAdults
pop2020Total	1	0.26	0.19	0.15	-0.19	-0.12	0.031	-0.05	0.071	-0.082	0.94	-0.19	-0.068	-0.067
TotalTests	0.26	1	0.9	0.79	0.13	0.31	0.0	[illegible]	[illegible]	[illegible]	0.4	0.15	0.061	0.12
Cases	0.19	0.9	1	0.93	0.072	0.36	0.2	0.37	0.25	0.15	0.34	0.15	0.028	0.044
Deaths	0.15	0.79	0.93	1	0.082	0.45	0.26	0.6	0.47	0.21	[illegible]	[illegible]	[illegible]	[illegible]
TestRatePerPopulation	-0.19	0.13	0.072	0.082	1	0.58	-0.18	0.23	-0.11	0.47	-0.14	0.45	-0.23	0.27
InfectionRateByPopulation	-0.12	0.31	0.36	0.45	0.58	1	0.28	0.73	0.33	0.42	-0.038	0.45	-0.25	0.14
PositiveCasesPerTestsGiven	0.031	0.083	0.2	0.26	-0.18	0.28	1	0.33	0.52	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DeathRateByPopulation	-0.05	0.3	0.37	0.6	0.23	0.73	0.33	1	0.73	0.37	0.029	0.29	0.069	0.13
DeathRateByPositiveCases	0.071	0.16	0.25	0.47	-0.11	0.33	0.52	0.73	1	0.22	0.12	0.17	0.14	0.059

65 and older vs. Total Test

Urban Population vs. Test Rate per Population

Urban Population vs. Infection Rate per Population

Step 4a: Regression Analysis on Positive Cases

Why are richer countries seemingly more impacted?

Analysis Prep

Country “richness” was benchmarked by GDP per Capita. Also tested household income, income per capita, and healthcare spending per capita.

To correct for size, case numbers were converted from raw count to cases per million residents.

Removed country lines without GDP, population, or testing information.

Leaves 106 countries for analysis.

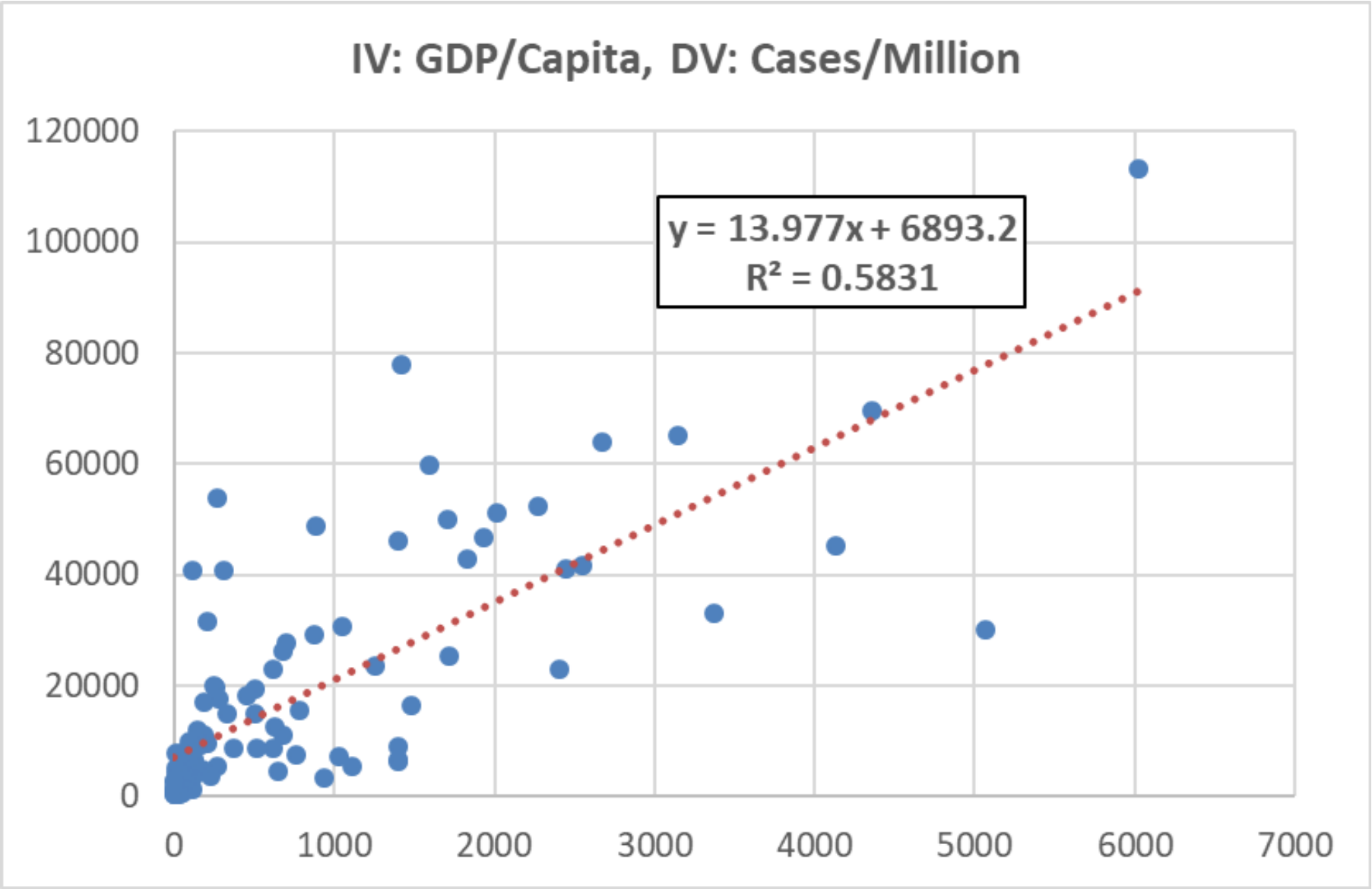
Why are richer countries more impacted?

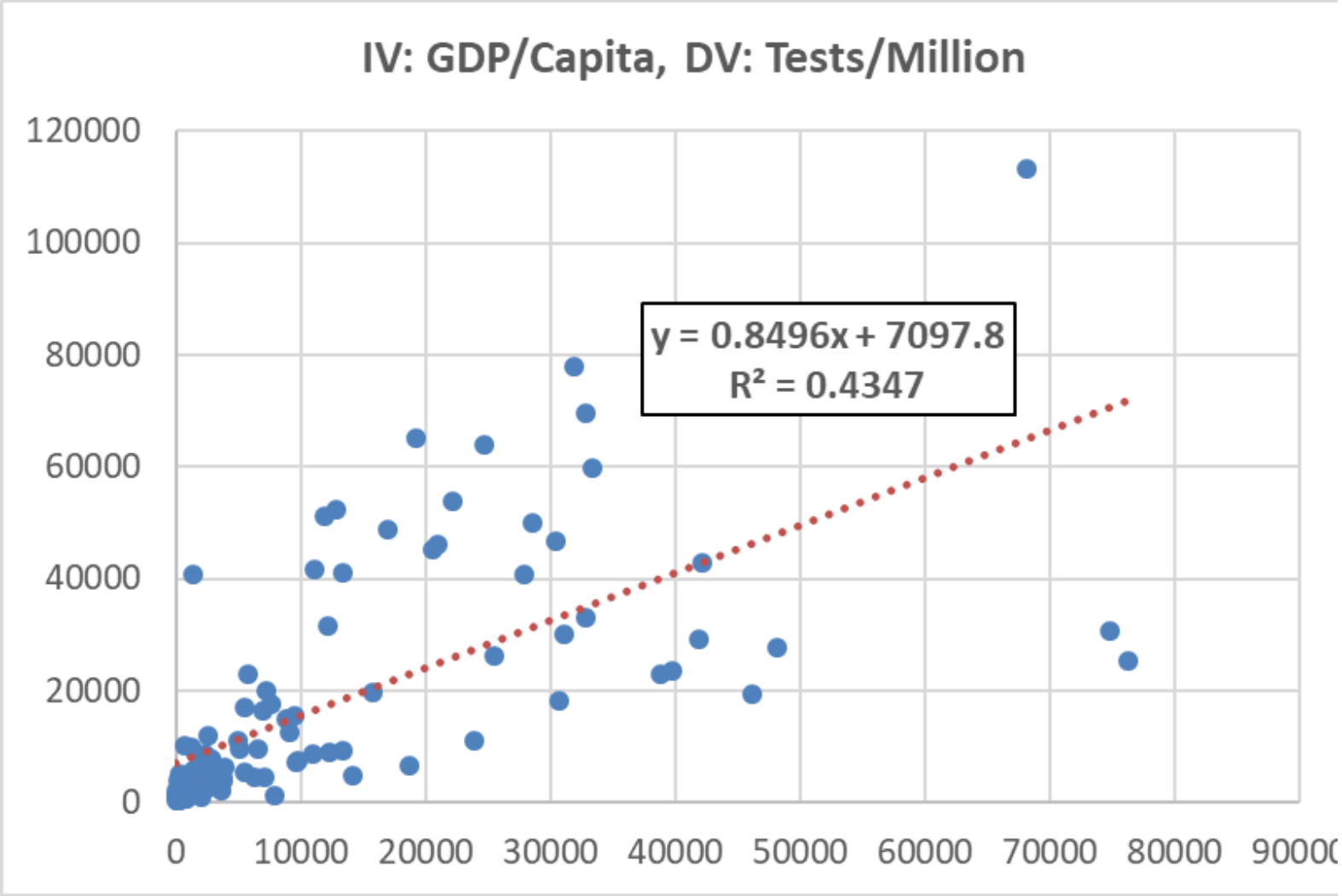
- Of the top 10 countries with the most cases/million:
  - 4 are within the top 10% of GDP/capita
  - 8 are within the top 20% of GDP/capita
  - All are within the top 30% of GDP/capita
- Of the 10 countries with the least cases/million:
  - 6 are within the lowest 10% of GDP/capita
  - 8 are within the lowest 20% of GDP/capita
  - All are within the lowest 30% of GDP/capita
- But **WHY?** Are richer countries more infectious?

	Country	Cases/M Ranking	GDP/Capita Ranking
Top 10 →	Luxembourg	1	1
	Spain	2	23
	Qatar	3	3
	Belgium	4	14
	Italy	5	20
	United States	6	4
	Singapore	7	5
	France	8	16
	United Kingdom	9	17
	Portugal	10	29
← Bottom 10	Haiti	97	100
	Zambia	98	95
	Madagascar	99	104
	Vietnam	100	83
	Laos	101	84
	Nepal	102	97
	Uganda	103	101
	Mauritania	104	93
	Burundi	105	106
	Yemen	106	98

Regression Analysis

Are cases/million correlated with GDP/capita?



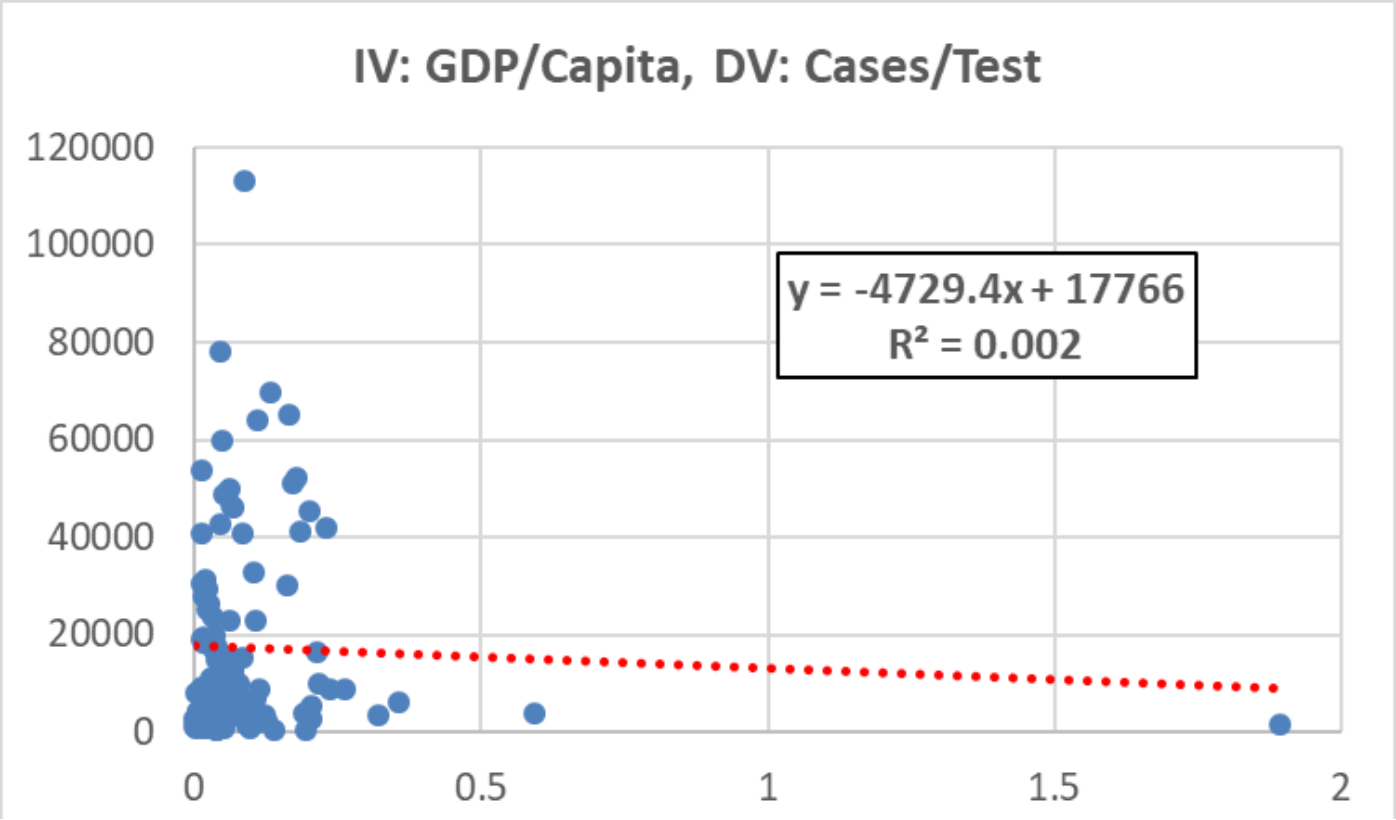


Is testing the confounding variable?

- Will we get a similar correlation if we replace cases per million with tests per million?
- Not quite as strong, but still some correlation.

Is Testing the Confounding Variable?

- Reminder: Original regression was IV GDP vs. DV cases/M.
- If we run the regression again, correcting population for number of tests, there is no correlation with GDP.
- Why? It is likely that the number of tests is the confounding variable.
- Therefore, richer countries are NOT more infectious, but the correlation MAY be due to the rate of testing.



Questions asked...

7.2% Deaths in total positive cases!

Why are nations spending more on health or R&D showing more deaths?

Does Covid hit women and men equally? Why do men appear to be dying at higher rates?

Am I safe because I am young?

More Tourism more deaths??

Step 4b- Regression Analysis on Deaths

Can relevant financial, demographic and mobility variables predict the rate of Covid-19 deaths ?

Analysis Prep

Cleaning

Countries with Null Values

Assumptions

Key Assumptions for 5 significant countries

Standardize

Scale variables to avoid overlooking of predictors with smaller numbers and ease of comparison

Dataset

Dataset with 92 countries

Response Variables

1. Deaths/Million
2. Deaths/ +ve Cases

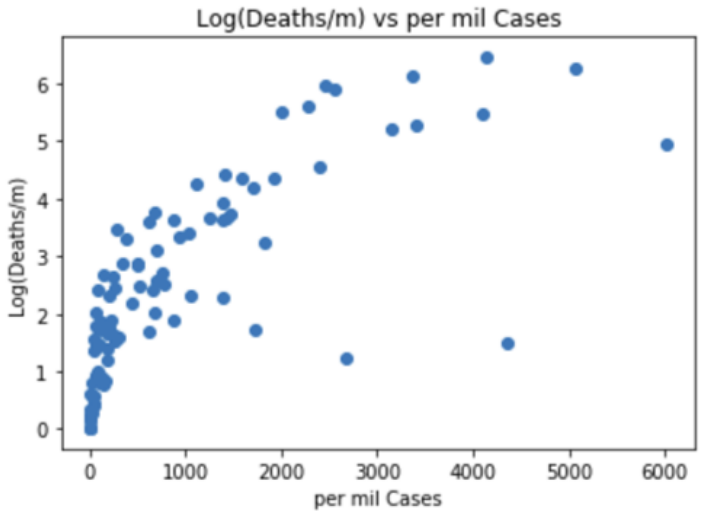
China	• Total Tests=1000,000 (random assumption)
South Korea	• Female proportion= 50 (equal male : female)
Switzerland	• Median Per Capita Income(derived from median per capita Household income/ avg household size *assuming 3.2*)
Czechia	• Female proportion 50 (equal male : female) • smoking prevalence (% of adults) 21.6 (average across the world)
Ireland	• Median Per Capita Income

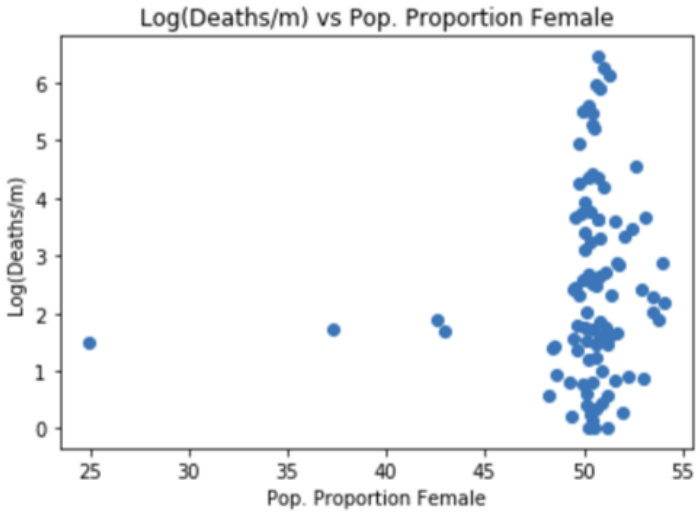
Approach

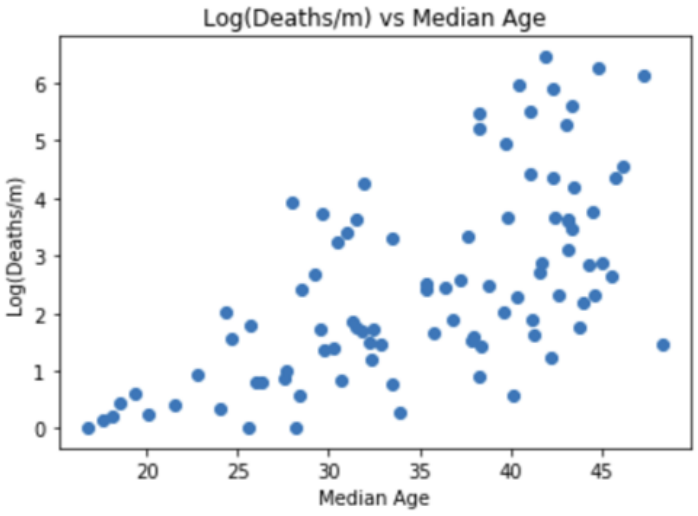
Regression Analysis

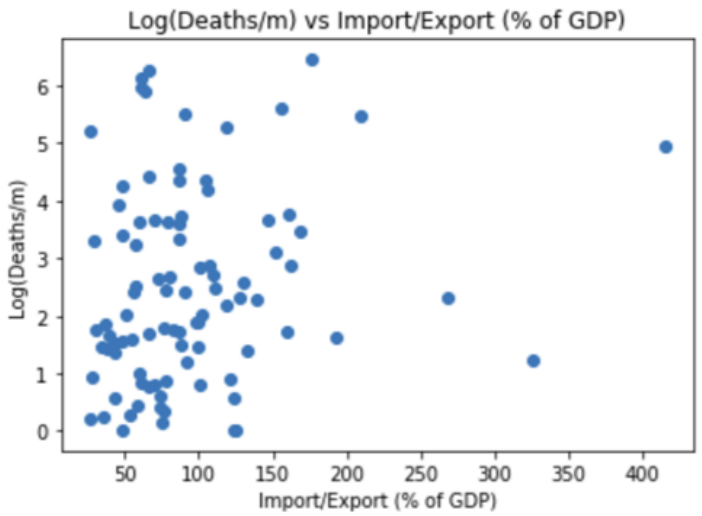
- Supervised Machine Learning for training the models
- Lasso Regression for backward regression to remove insignificant variables
- Multiple Linear Regression with
- Hypothesis Testing for Significance at 10% significance level or P value = 0.1 (increase the range of probability due to dynamics.

Analysis, Results & Challenges

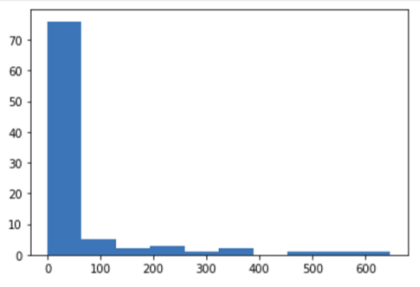


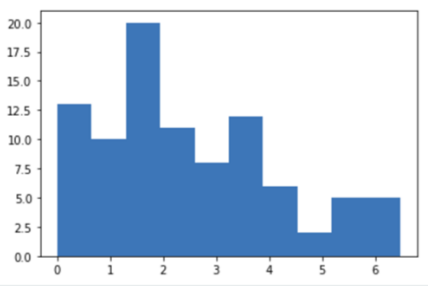






Deaths ---> deaths per million from raw data





Uneven distribution of deaths/million

Transformation y = Log(Death/Million)
more normal or symmetric moving the big countries closer together and space out the smaller ones

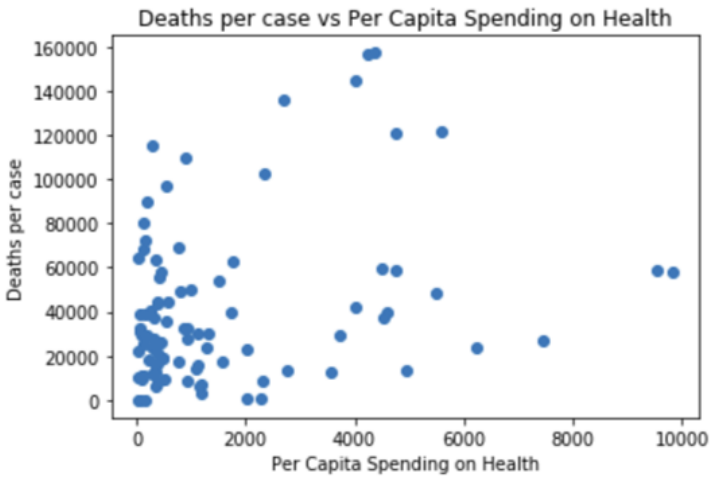
R^2 = 0.82

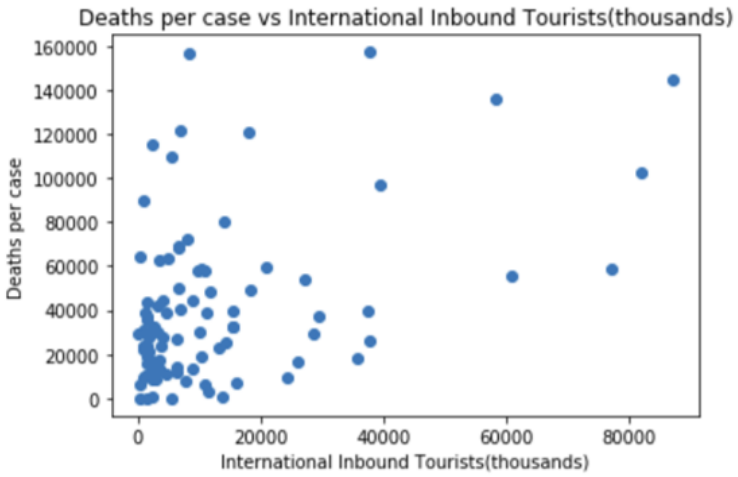
probable overfitting scenario

Significant Predictors

- Positive Cases (+)
- Female proportion from population (+)
- Median Age (+)
- Import Export (% of GDP) (-)

Analysis, Results & Challenges





Challenge - Finding right data due to missing values of countries

MICE technique can be used to find missing values

Deaths ---> deaths per Cases from raw data

R^2 = 0.36

Underfitting scenario

Significant Predictors

- Per Capita spending on Health (+)
- International Inbound Tourists (+)

Questions we asked and what we found

Not really but +ve cases with more median age show a considerable correlation (0.75 and 0.62) and significance to Covid-19 fatality rate.

Why are nations spending more on health showing more deaths?

Contagious-> Mobility->Risk Intrnl inbound tourists considerable correlation 0.44 and significance. Inverse relation Import /Export (%of GDP), low correlation 0.19. Needs more research.

Does Covid hit women and men equally? Why do men appear to be dying at higher rates?

Am I safe because I am young?

Analysis shows more testing more detection implies some percentage of those dying.

More Tourism more deaths??

Contradicting results female ratio predominantly between 46-55%. Need disaggregated data by sex to improve real-time targeted forecasting.

Recommendations & Conclusion

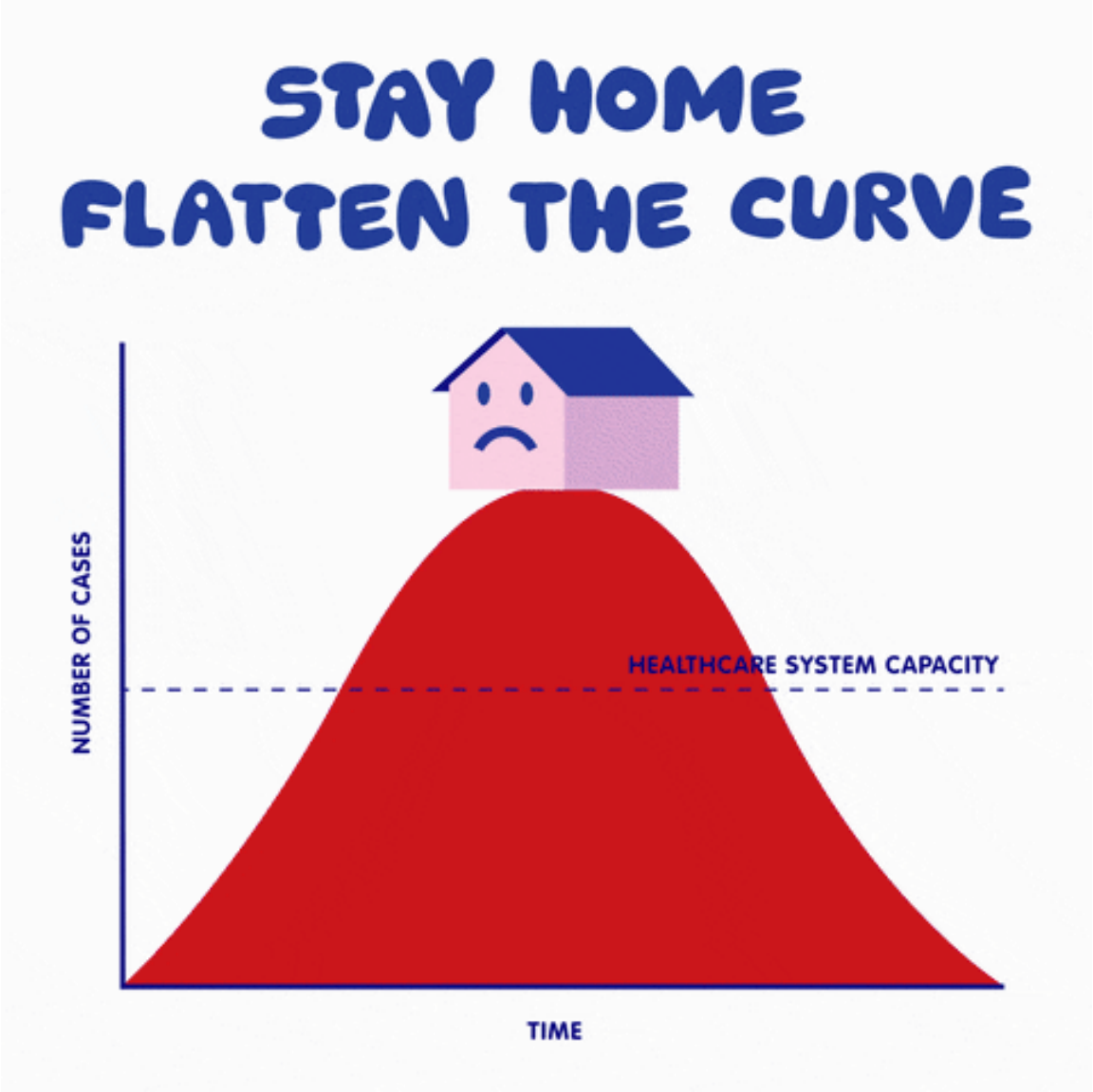
- Case/ death rates must be taken with a grain of salt.
- Richer countries aren't more infectious, but likely have better means for testing and tracking.
- To better understand case rates, it is imperative that countries:
  - Test more
  - Find common definitions for testing numbers
- To assist poorer countries with testing, the UN should develop a fund devoted to testing.

Adam Kucharski
@AdamJKucharski

Look at the Wuhan line on this new graph from @jburnmurdoch. The lockdown was introduced there on 23rd Jan – 69 days ago – which means this entire Wuhan curve has happened since then. It shows how long it can take to see the effect of control measures on the number of deaths.

3:10 PM · Apr 1, 2020

3.6K 2.4K people are talking about this



THANK YOU

In addition, we would like to acknowledge Professor Rex for his help.