2020

Data Exploration Report

POURIA EBRAHIMNEZHAD

STUDENT ID: 30035678

TUTOR NAME: MOHAMMAD HAQQANI

FIT5147 DATA EXPLORATION AND VISUALISATION | S1 APRIL 2020

Contents

1.Introduction

As our lives have been turned upside down in the recent months by a global pandemic and the spread of the COVID-19 virus across the globe it was of a great interest to me to explore the data obtained from the spread of this disease and use this opportunity to explore if I can find any interesting patterns or relations from a data perspective between the rate of the spread of the disease in various countries versus different health, economic or environmental attributes. For this reason, I initially proposed three rather large questions but after consulting with my tutor I decided to focus on fewer attributes to allow for exploration within our given time frames.

1. Is there any relationship between the spread of COVID-19 and the median age of the countries impacted?
2. Is there any relationship between the spread of COVID-19 and the population density of the countries impacted?
3. Is there any relationship between the spread of COVID-19 and different health and governmental factors (i.e. smoking, life expectancy, Government spending on health)?

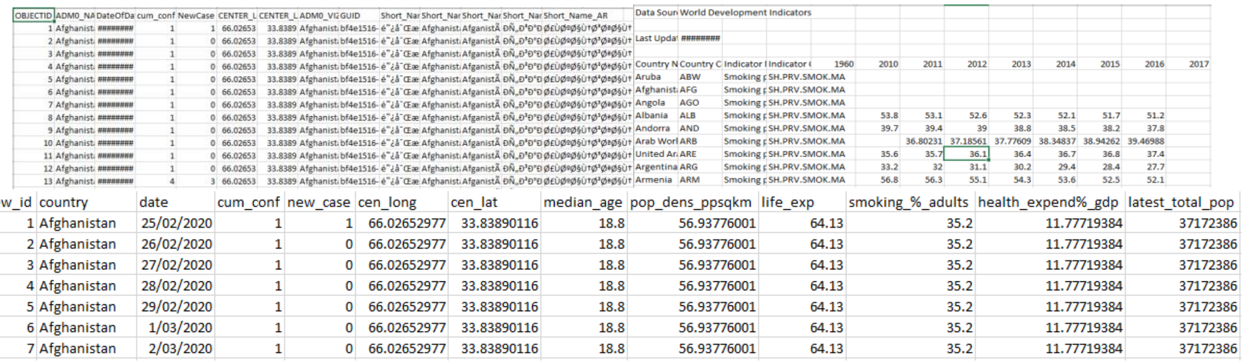
2.Data Wrangling

The hardest part of this exercise as with every other Data science related task was the wrangling of the data. As I used the dataset provided by WHO from data.humdata.org (OCHA, 2020) for the COVID-19 statistics in csv format and for the purpose of exploring above questions I needed to source other data for each attribute by country (mostly from world bank datasets in csv format) (Group, 2020) and ultimately construct a tabular dataset merging the COVID-19 data with all the other attributes for each country. This proposed its own challenges, country names for example was not the same across multiple datasets, handling missing data for each attribute, data format and reshaping, etc.

I used mostly excel spreadsheets and Python pandas to wrangle the data in to my desired format. below main steps were taken to produce the final tabular data

1. Dropping unnecessary columns from covid-19 and other data sets keeping only desired columns and renaming
2. Cleaning other data sets such as missing data as most data was reported by year and some countries had missing data for most recent years hence, I used the latest reported record for each country where possible. Smoking data (Review, 2020) proposed the greatest number of missing records and I replaced those countries values by averaging all countries figures
3. Also, I removed any country from this study which had most attributes missing and I couldn’t find a trusted source to replace the values
4. Merging the main Covid-19 dataset with each attribute dataset and producing the tabular data ready for exploration

I have used sample snap shots of multiple datasets csv and final format which I used to explore the data below



OBJECTID	ADM0_NA	DateOfDa	cum_conf	NewCase	CENTER_L	CENTER_L	ADM0_VI	GUID	Short_Nar	Short_Nar	Short_Nar	Short_Nar	Short_Name_AR
1	Afghanist	########	1	1	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
2	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
3	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
4	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
5	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
6	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
7	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
8	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
9	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
10	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
11	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
12	Afghanist	########	1	0	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†
13	Afghanist	########	4	3	66.02653	33.8389	Afghanist	bf4e1516-	é˜¿å¯Œæ	Afghanist	AfganistÃ	ÐÑ„Ð³Ð	Ø£ÙØºØ§Ù†Ø³ØªØ§Ù†

Data Soun	World Development Indicators											
Last Upda	########											
Country N	Country C	Indicator I	Indicator (	1960	2010	2011	2012	2013	2014	2015	2016	2017
Aruba	ABW	Smoking p	SH.PRV.SMOK.MA									
Afghanist	AFG	Smoking p	SH.PRV.SMOK.MA									
Angola	AGO	Smoking p	SH.PRV.SMOK.MA									
Albania	ALB	Smoking p	SH.PRV.SMOK.MA		53.8	53.1	52.6	52.3	52.1	51.7	51.2	
Andorra	AND	Smoking p	SH.PRV.SMOK.MA		39.7	39.4	39	38.8	38.5	38.2	37.8	
Arab Worl	ARB	Smoking p	SH.PRV.SMOK.MA			36.80231	37.18561	37.77609	38.34837	38.94262	39.46988	
United Ar	ARE	Smoking p	SH.PRV.SMOK.MA		35.6	35.7	36.1	36.4	36.7	36.8	37.4	
Argentina	ARG	Smoking p	SH.PRV.SMOK.MA		33.2	32	31.1	30.2	29.4	28.4	27.7	
Armenia	ARM	Smoking p	SH.PRV.SMOK.MA		56.8	56.3	55.1	54.3	53.6	52.5	52.1	

new_id	country	date	cum_conf	new_case	cen_long	cen_lat	median_age	pop_dens_ppsqkm	life_exp	smoking_%_adults	health_expend%_gdp	latest_total_pop
1	Afghanistan	25/02/2020	1	1	66.02652977	33.83890116	18.8	56.93776001	64.13	35.2	11.77719384	37172386
2	Afghanistan	26/02/2020	1	0	66.02652977	33.83890116	18.8	56.93776001	64.13	35.2	11.77719384	37172386
3	Afghanistan	27/02/2020	1	0	66.02652977	33.83890116	18.8	56.93776001	64.13	35.2	11.77719384	37172386
4	Afghanistan	28/02/2020	1	0	66.02652977	33.83890116	18.8	56.93776001	64.13	35.2	11.77719384	37172386
5	Afghanistan	29/02/2020	1	0	66.02652977	33.83890116	18.8	56.93776001	64.13	35.2	11.77719384	37172386
6	Afghanistan	1/03/2020	1	0	66.02652977	33.83890116	18.8	56.93776001	64.13	35.2	11.77719384	37172386
7	Afghanistan	2/03/2020	1	0	66.02652977	33.83890116	18.8	56.93776001	64.13	35.2	11.77719384	37172386

Figure 1 - multiple data formats and final tabular merged data

3.Data Checking

I used summarized statistical view of the data frame and also excel spreadsheets to check the data, various issues I faced was missing values, which I made the assumption that are mostly as MCAR because most of the missing data is either survey not carried out in those countries or the value doesn't exist for specific country in a given year I used the latest reported values in those cased if the value existed for previous years. In the case of smoking percentage of adults where a lot of the countries either had missing values or I couldn't find the value online I replaced the values with the average value from the dataset.

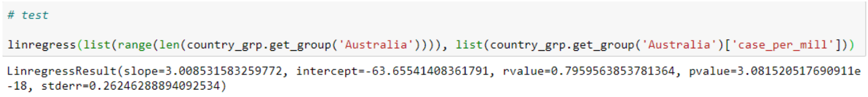
Also used Tableau to check the coordinates of the countries for any issues visually and made sure the final data frame doesn't have any missing data using pandas in python

4.Data Exploration

In order to explore the data further I looked at a few issues with respect to what I am trying to answer, in order to be able to compare the rate of the spread of the disease in each country as well as its impact I needed to statistically make some adjustments and define and calculate some new fields from my data.

4.1 defining rate of spread

The approach I took to continue this study further was to use the latest reported total population of each country year 2018 in worldbank (Group, 2020) to calculate a new column (***case_per_mill***) which reflects the total cases in each country per 1 million population each date. Then I used the linregress function from scipy.stats library (community, 2017) and the two dimensions one being the case_per_mill column and the other, the number of reported dates (number of rows) for each country to calculate the best line of regression that fits this 2D data in each country's case. I then added the slope of that line and for the rest of this document when I refer to rate, I am referring to this slope value. use rate or slope as a new parameter to my data set. figure 2 shows a sample of this function for Australia



```
# test

linregress(list(range(len(country_grp.get_group('Australia')))), list(country_grp.get_group('Australia')['case_per_mill']))

LinregressResult(slope=3.008531583259772, intercept=-63.65541408361791, rvalue=0.7959563853781364, pvalue=3.081520517690911e
-18, stderr=0.26246288894092534)
```

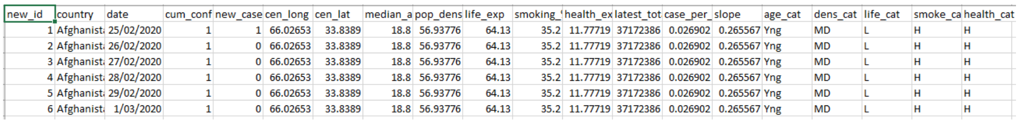
Figure 2 - function used to estimate the slope of best fit line to each country data

Now my dataset allows me to have a column (***slope***) for each country with the given slope of the best fit linear line to have some representation of the rate of the spread of the disease and what is further is that this provides a weighted by population view and not just the raw figure per country which levels the playing field for further comparison.

4.2 defining categories

The next step I took was to introduce categories with respect to the attributes for each country, this would later allow me to slice my data in order to create more meaningful visualisations as I am dealing with a large number of countries in my final dataset (169) and can aid in exploring the impact of various attributes on the spread of the disease. In order to do this, I used summary statistics from the dataframe to allow for an Equal Frequency approach for categorisation

For each attribute I defined categories and added to my dataset, resulting in 3 median age (Review, 2020) categories (Old, Mid, Yng), 4 categories for population density (Group, 2020) (Extreme Dense, High Dense, Mid Dense, Low Dense), 3 for life expectancy (High, Mid, Low), 3 for smoking percentage of adults (Heavy, Mid, Light) and 3 categories for countries GDP percentage on health expenditure (High, Mid, Low). Figure 3 shows the final view of the spreadsheet which I used in Tableau for exploring further.



new_id	country	date	cum_conf	new_case	cen_long	cen_lat	median_a	pop_dens	life_exp	smoking_'	health_ex	latest_tot	case_per_	slope	age_cat	dens_cat	life_cat	smoke_ca	health_cat
1	Afghanist:	25/02/2020	1	1	66.02653	33.8389	18.8	56.93776	64.13	35.2	11.77719	37172386	0.026902	0.265567	Yng	MD	L	H	H
2	Afghanist:	26/02/2020	1	0	66.02653	33.8389	18.8	56.93776	64.13	35.2	11.77719	37172386	0.026902	0.265567	Yng	MD	L	H	H
3	Afghanist:	27/02/2020	1	0	66.02653	33.8389	18.8	56.93776	64.13	35.2	11.77719	37172386	0.026902	0.265567	Yng	MD	L	H	H
4	Afghanist:	28/02/2020	1	0	66.02653	33.8389	18.8	56.93776	64.13	35.2	11.77719	37172386	0.026902	0.265567	Yng	MD	L	H	H
5	Afghanist:	29/02/2020	1	0	66.02653	33.8389	18.8	56.93776	64.13	35.2	11.77719	37172386	0.026902	0.265567	Yng	MD	L	H	H
6	Afghanist:	1/03/2020	1	0	66.02653	33.8389	18.8	56.93776	64.13	35.2	11.77719	37172386	0.026902	0.265567	Yng	MD	L	H	H

Figure 3 - final dataset for Tableau exploration

4.3 Visualisations

First, I looked at if there are any outliers with respect to the Slope field that I created in previous step. I used Tableau to develop a statistical test using z-core calculations (Teggart, 2016) based on slope and visualised this (figure 4), I used this and removed 5 countries with the highest rate (Andorra, Iceland, Luxembourg, San Marino, Switzerland) from my dataset. These countries would significantly impact any trends I am trying to visualise and would make it easier to see any relationships in the rest of the data without them.

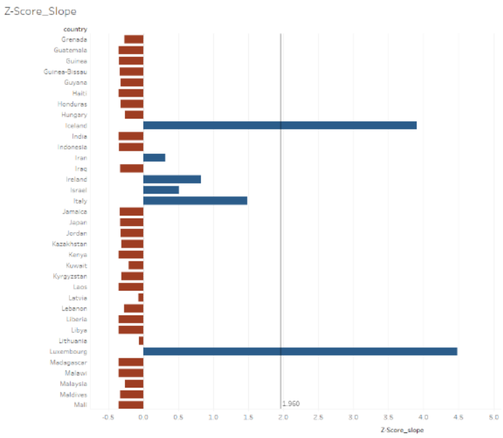


Figure 4 - Z-score for slope showing outlier countries

The first visualisation I used was to view map of the world and at the trend size for each country using slope and use colour for different categories, for this I used tableau visualisation visible in figure 4

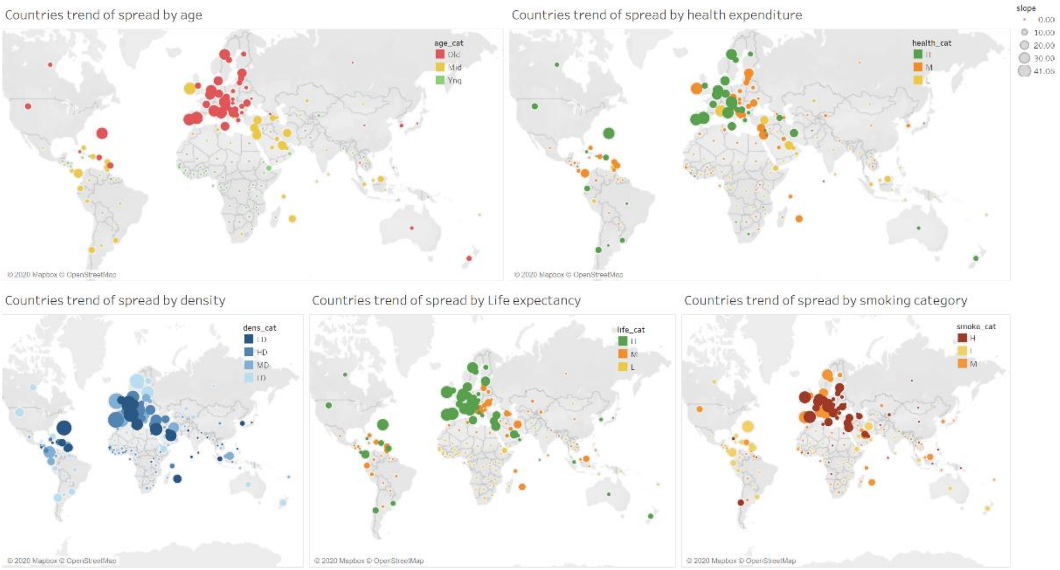

Countries trend of spread by age
age_cat
Old
Mid
Yng
© 2020 Mapbox © OpenStreetMap
Countries trend of spread by health expenditure
health_cat
H
M
L
slope
0.00
10.00
20.00
30.00
41.05
© 2020 Mapbox © OpenStreetMap
Countries trend of spread by density
dens_cat
LD
HD
MD
ID
© 2020 Mapbox © OpenStreetMap
Countries trend of spread by Life expectancy
life_cat
H
M
L
© 2020 Mapbox © OpenStreetMap
Countries trend of spread by smoking category
smoke_cat
H
L
M
© 2020 Mapbox © OpenStreetMap

Figure 5 - trend map based on feature category

Based on some initial instant observations we can see some positive relationships between these categories and the rate of the disease (size), I will explore these further using other methods of visualisations to see how significant each case actually is.

4.3.1 Rate vs Median age

Next step was to see if there is actually any relationship between median age of countries and the rate of the growth of cases, for this I plotted a few visualisations which I think can allow for visualising if there are any relationships. Figure 6 demonstrates the various approach I took in visualising my data

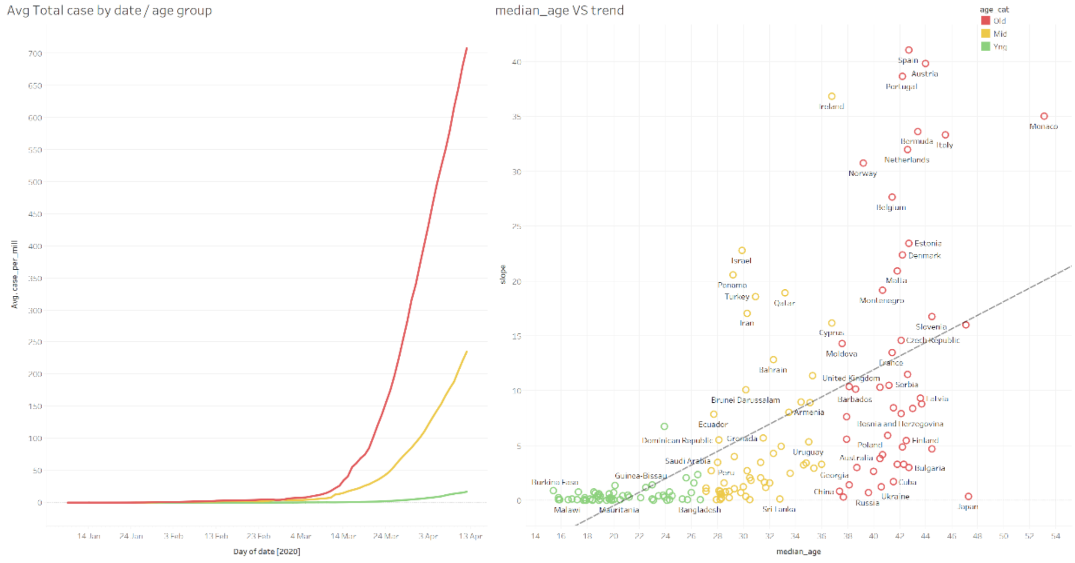
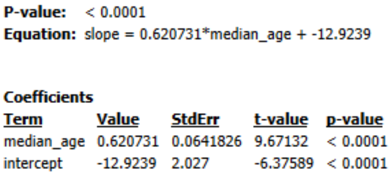


Figure 6 - rate VS median age

First graph on the left shows the average of cases per million by calendar date and categorised by old, middle and young groups we can see that the average cases for each category clearly are separated and show distinction between the age groups with older age countries experiencing more cases on average than middle and young age. The second graph on the right shows scatter plot of the median age of each country plotted on the x-axis vs the rate of growth in that country, I have used a linear trend line to capture the correlation between these two variables which shows there is a clear positive correlation between the two, also have printed the statistical significance of this trend line here.



P-value: < 0.0001
Equation: slope = 0.620731*median_age + -12.9239

Coefficients

Term	Value	StdErr	t-value	p-value
median_age	0.620731	0.0641826	9.67132	< 0.0001
intercept	-12.9239	2.027	-6.37589	< 0.0001

Another way to visually study the relationship of age and the spread of the disease in each country was to create bar charts for the slope value for each country and sort them by age. However, in order to visually be able to see the trend I segmented the countries using density category so I have four groups demonstrated in figure 8, we can see that the sorting by age results in the higher rate countries to line up on the right side of each category. This can allow for quick comparison of countries themselves in each density category. And discovering interesting trends as well as exception cases. For example, if we look at extreme dense countries, most with higher median age show a higher rate of the spread of the disease and exception cases are Japan or maybe Korea which show a much lower trend.

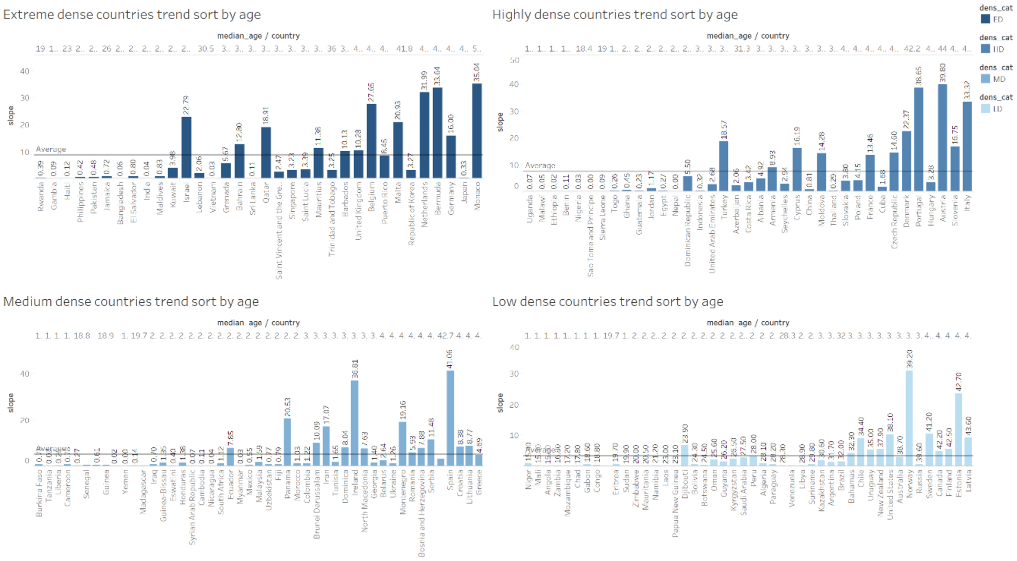
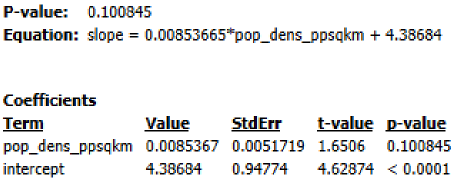


Figure 7 - relationship of age and spread of disease for countries segmented by density

4.3.2. Rate vs population density

here I have explored the relationship between density (people per square kilometre) and the rate of the spread of disease using 4 categories of extreme, high, medium and low density for countries. As visible in figure 7 the left hand graph demonstrates that the extreme dense and high dense countries show to have a higher average than that of middle and low density but the relationship between the rate of increase of the disease and these categories are not as clear looking at the right hand side graph of figure 8, there is still a positive trend but not as significant as that of age. Given the statistics scores shown here.



P-value: 0.100845
Equation: slope = 0.00853665*pop_dens_ppsqkm + 4.38684

Coefficients

Term	Value	StdErr	t-value	p-value
pop_dens_ppsqkm	0.0085367	0.0051719	1.6506	0.100845
intercept	4.38684	0.94774	4.62874	< 0.0001

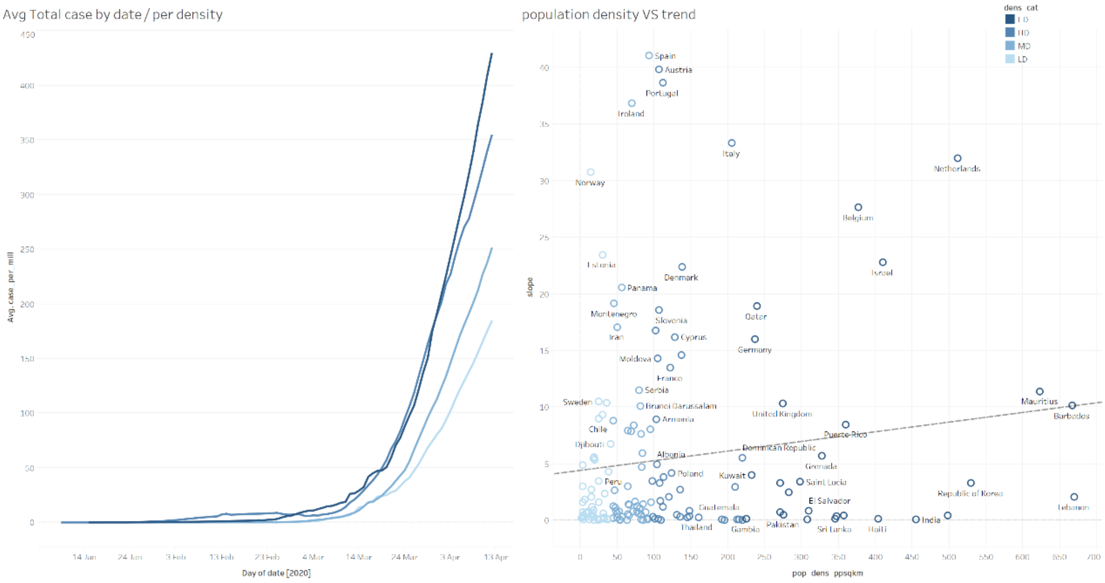
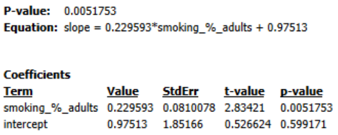


Figure 8 - rate VS population density

4.3.3. Rate vs Smoking percentage of adults

Now looking at the smoking categories in my dataset (figure 9) I can see that maybe the smoking data doesn’t provide as clear relationship as age but still better than density, even though in first look there seems to be some higher average figures for high smoking countries the trend line in the scatter plot on the right hand side of figure 9 doesn’t show a large significance in the relationship with statistics presented here



P-value: 0.0051753
Equation: slope = 0.229593*smoking_%_adults + 0.97513

Coefficients

Term	Value	StdErr	t-value	p-value
smoking_%_adults	0.229593	0.0810078	2.83421	0.0051753
intercept	0.97513	1.85166	0.526624	0.599171

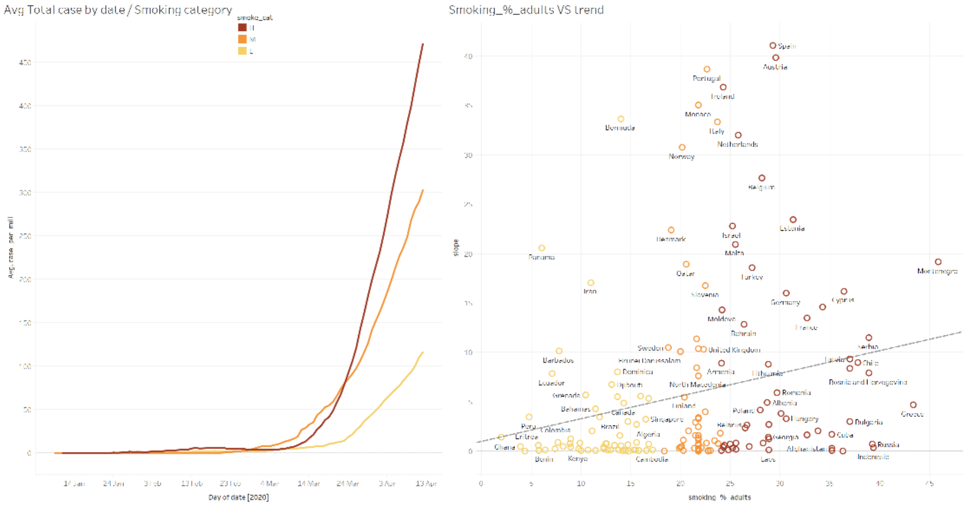


Figure 9 - rate vs smoking percentage of adults

4.3.4. Rate vs Health expenditure

The next attribute which I looked at was governments health expenditure as percentage of GDP and explore its relationship with the rate of the spread of disease in different countries, looking at figure 10, we can see a rather interesting and more clear relationship between this factor and trend of the spread. The countries with higher health expenditure seem to be showing a faster rate of cases per million population. This is an interesting observation considering at first thought this wouldn't be expected

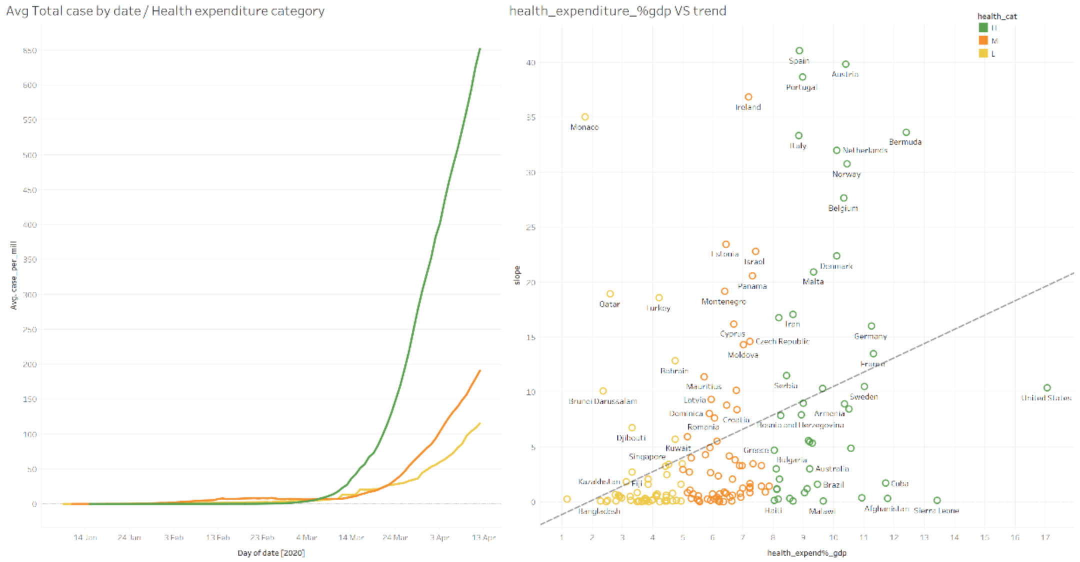
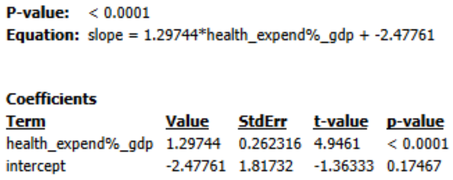


Figure 10 - rate VS Gov health expenditure as % of GDP

And the relationship seems to be more significant than that of all other categories explored so far, and we can see from the left-hand side graph that the trend is consistent for lower health expenditure compared to middle as well.



P-value: < 0.0001
Equation: slope = 1.29744*health_expend%_gdp + -2.47761

Coefficients

Term	Value	StdErr	t-value	p-value
health_expend%_gdp	1.29744	0.262316	4.9461	< 0.0001
intercept	-2.47761	1.81732	-1.36333	0.17467

4.3.5. Rate vs Life expectancy

The next attribute to explore was the life expectancy and its relationship with slope or rate. Figure 11 demonstrates this relationship, again in the case of life expectancy we see that there is a clear positive relationship between the two attributes meaning the higher he life expectancy of a country is the higher the rate of the spread of the disease.

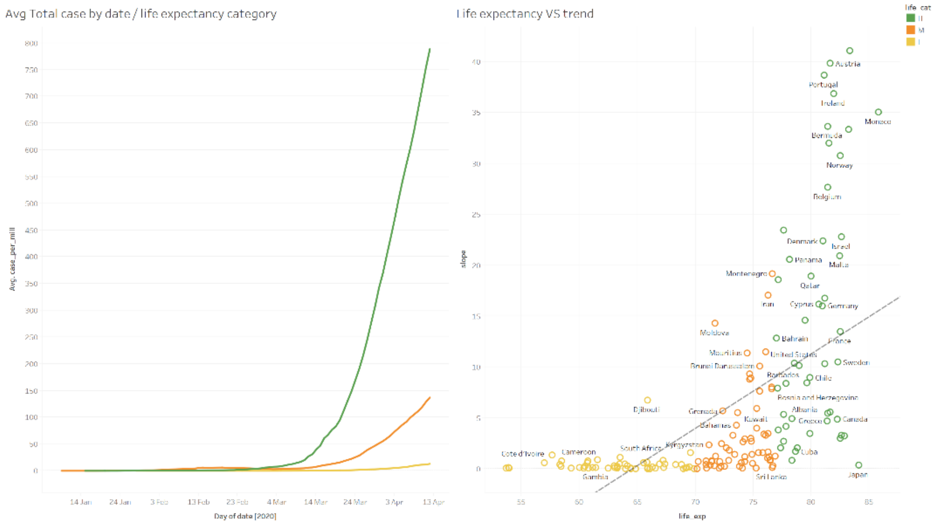


Figure 11 - rate VS Life expectancy

The relationship seems to be comparable to that of age and a significant one compared to smoking or density.

P-value: < 0.0001
Equation: slope = 0.731737*life_exp + -47.2754

Coefficients

Term	Value	StdErr	t-value	p-value
life_exp	0.731737	0.0808675	9.04859	< 0.0001
intercept	-47.2754	5.90218	-8.00983	< 0.0001

Again for exploring the impact of life expectancy in figure 12 I have used smoking category to split the countries into groups and then visualise the relative slope of each country in each group using bar charts, but I have sorted the countries by their life expectancy and clearly we can see the higher rate countries align on the right hand side of each graph which emphasises the existence of this relationship in the data more visually.

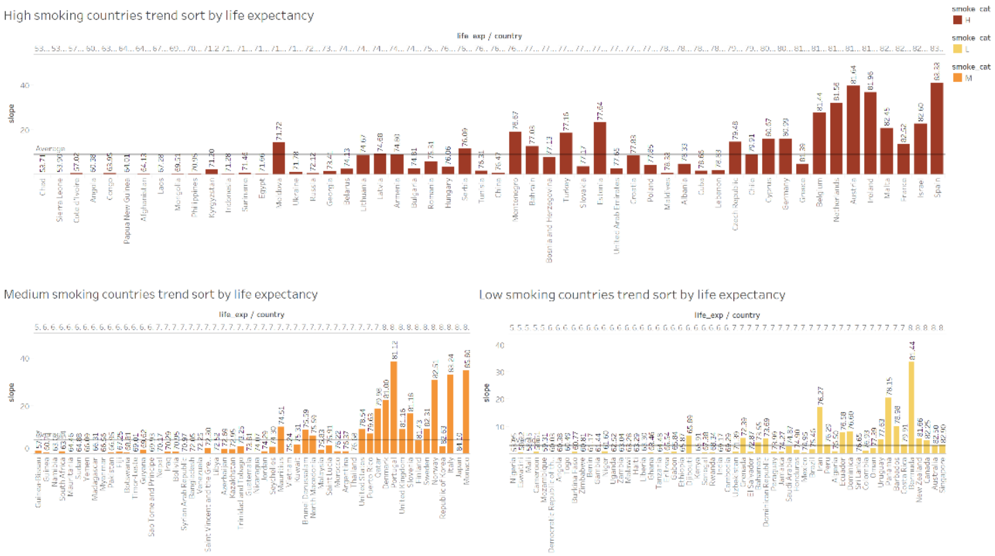


Figure 12 - The relationship of life expectancy and rate of disease each graph represents countries segmented by smoking category

5 Conclusion

Comparing the relationship of these factors on the linear trend of the spread of the disease we can see some positive relationship in all cases with the order of their importance shown in below table

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
health_expend%_gdp	1.29744	0.262316	4.9461	< 0.0001
life_exp	0.731737	0.0808675	9.04859	< 0.0001
median_age	0.620731	0.0641826	9.67132	< 0.0001
smoking_%_adults	0.229593	0.0810078	2.83421	0.0051753
pop_dens_ppsqkm	0.0085367	0.0051719	1.6506	0.100845

Table 1 - Coefficients for linear trend lines desribing the relationship with trend of disease

At face value it’s an interesting and maybe an unexpected observation to see that governments who have spent a larger percentage of their GDP on health have shown a faster trend of spread of the COVID-19, also those who have a higher life-expectancy.

On the other hand one may suggest that age plays the underlying factor here and the relationship between government health expenditure and how old a country is and the life expectancy of that country all have age at the centre of it and seems to show a positive correlation with the slope of the spread of disease in countries, but we have to be careful in drawing this conclusion as a lot of other factors may have impacted the trend which has not been studied here.

6 Reflection

I was expecting to find a more meaningful relationship between smoking data and its relationship to the spread of the disease which doesn’t seem to be as significant here and also the population density of countries which is surprising to see.

However there maybe caveats here in the data, for one I replaced the missing values of smoking data in a lot of the countries with the average of this value which could have impacted the relationship, also not all countries had most recent data for this parameter which makes the observations questionable. With regards to population density I used the data provided in data.worldbank.org which provides description for this data as people per square kilometre of land area in each country. If I could’ve sourced data regarding average density of the countries with respect to their cities, I would have expected different results using the same analysis I have done.

Another point I would like to make here is that I used a linear model in this comparison in all cases one could argue that if I had used a different model (i.e. exponential, polynomial) the relationship observations may have differed, but due to time and simplicity constraints I wasn’t able to include this model in my comparison.

Also, what I would have liked to have the chance to do further was to study relationship of other environmental and socioeconomic factors such as Cardiovascular disease, respiratory disease, income, education, average temperature, etc. these would have required further data sourcing, wrangling and may have led to discovering other interesting trends.

Bibliography

community, T. S. (2017, March). *scipy.org*. Retrieved from docs.scipy.org: https://docs.scipy.org/doc/scipy-0.19.0/reference/generated/scipy.stats.mstats.linregress.html

Group, W. B. (2020). *The World Bank*. Retrieved from worldbank.org: https://www.worldbank.org/

OCHA. (2020). *HUMANITARIAN DATA EXCHANGE*. Retrieved from data.humdata.org: https://data.humdata.org/dataset/coronavirus-covid-19-cases-data-for-china-and-the-rest-of-the-world

Review, W. P. (2020). *World Population Review*. Retrieved from https://worldpopulationreview.com/countries/smoking-rates-by-country/

Teggart, A. (2016, February). *The Data School*. Retrieved from https://www.thedataschool.co.uk/anuka-teggart/tipweek-calculating-z-scores-in-tableau/