MADA Project Manuscript

Dawson Dobash

2021-10-29

# 1 Summary/Abstract

*Write a summary of your project.*

# 2 Introduction

## 2.1 General Background Information

COVID-19 hit the world by storm and is a possible ever-lasting change to the way society acts. It also had an impact on each countries funding towards health care. Actions within the past two years have been taken to help fight against this virus.

## 2.2 Description of data and data source

The data I plan on using is from a mixture of locations. Two sources are from kaggle.com and the other is from the Organisation for Economic Co-operation and Development (OECD) website. My data consists of eight data sets with information over each country in the world (for those that data was collected for). Six of these data sets provide the same information just for different countries where each data set is a different region. There is a data set that includes all of the countries but I want to include a region indicator variable for each country. These data sets includes total covid-19 cases, active cases, total deaths, total recovered, total tests, and population size. This data was updated on kaggle as of Sept 16th. Another data set I am using is a vaccination data set for each country. This is also data as of Sept. 16th. Finally, the last data set is of each countries healthcare funding for the past 4-5 years. Some countries provide 2020 data but others only provide up to 2019.

I plan on combining the data sets into one with the variable combining them all is the country. If there are missing data within the variables used in analysis, I will exclude those countries. Ill have to change the raw data as one data set uses 3 letter abbreviations for countries while the others use the whole name of the country. There will be a lot of cleaning for this data just to make it as I want it.

I included 3 data sets out of the 8 within the processingscript for now because one shows the same information as the other data sets not included just in a different region of the world. This is just to show the data I will be using.

Links to the locations I received the data are below:

<https://data.oecd.org/healthres/health-spending.htm>

<https://www.kaggle.com/iamsouravbanerjee/covid19-dataset-world-and-continent-wise?select=Covid+Data+-+World.csv>

<https://www.kaggle.com/anandhuh/latest-worldwide-vaccine-data>

## 2.3 Questions/Hypotheses to be addressed

1. Since this data is cumulative and does not break down the amount of cases per month per country, I want to compare the recovery proportion to the death proportion and if this difference is based on variables such as the number of vaccines pushed out to society, the amount of funding the health care system has, the amount of tests taken. This will hopefully show if the amount of health care support really helps patients recover.
2. I will also look at how the percent of COVID-19 cases for each country is affected by multiple predictors such as the ones addressed in the above paragraph.

Some issues I can run into are: 1) Developed countries might have more cases as they are in contact with the world more as travelers commonly travel to certain countries. Therefore could show an inverse relationship to what I expect with more developed countries have more recoveries. 2) Unknown dates of when a country received vaccines and amount.

# 3 Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

## 3.1 Data import and cleaning

For the data importing and cleaning section, I first had to import 8 different data sets into R from multiple sources specified in the Description of data and data source. After this importation, The first thing to do before combing the data set was to create indicator variables for each country. After this I wanted to make sure that each variable was the right class as some of the numeric variables were character variables. This will help in making sure that each variable was able to be used in the right way.

I then was able to combine each continent data set into one world data set to make the importing go easier. After combining the data sets, I need to only include the variable for each data set needed for analysis and make sure that all na values were removed from each data set as I only want full information observations. However this would delete some countries that were large in population such as China and Phillipines as they did not have vaccine information.

The last problem I ran into was that the country names for all datasets were different. One included a three letter code for countries, then another one used different abbreviations/names for countries. I had to use the countrynames package to help change into names I wanted. Since I could not change all of the countries names to the way I wanted it, I had to unfortunately go into the raw data csv files to change them manually. After all of this was good to go, I used the merge function to combine all of the data sets the way I wanted to.

One disclosure I have for this is that one data set, health care funds, only had 48 countries when I cleaned the data. Therefore, there are a lot of na values for only those specific variables that came from the health care funds data set in the merged one.

To look at a more descriptive explanation of all of the steps I took in data cleaning and importing, go to the processing script where it has comments for every line of code.

## 3.2 Exploratory analysis

Table 1 shows a table summarizing the data.

Table 3.1: Data summary table.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Country | Total Cases | Total Deaths | Total Recovered | Active Cases | Total Tests | Population | location | Doses administered per 100 people | Total doses administered | % of population vaccinated | % of population fully vaccinated | TIME | Value |
|  | Length:164 | Min. : 4 | Min. : 1.0 | Min. : 3 | Min. : 0 | Min. : 14689 | Min. :1.073e+05 | Length:164 | Min. : 0.10 | Min. : 31090 | Min. : 0.100 | Min. : 0.100 | Min. :2018 | Min. : 257.4 |
|  | Class :character | 1st Qu.: 29810 | 1st Qu.: 541.5 | 1st Qu.: 25709 | 1st Qu.: 1878 | 1st Qu.: 386297 | 1st Qu.:2.857e+06 | Class :character | 1st Qu.: 14.25 | 1st Qu.: 439333 | 1st Qu.: 9.025 | 1st Qu.: 4.875 | 1st Qu.:2019 | 1st Qu.: 2160.3 |
|  | Mode :character | Median : 193204 | Median : 2761.0 | Median : 166708 | Median : 8217 | Median : 2338400 | Median :1.021e+07 | Mode :character | Median : 60.50 | Median : 3384675 | Median :36.500 | Median :26.000 | Median :2019 | Median : 3547.0 |
|  | NA | Mean : 1343590 | Mean : 26684.9 | Mean : 1207378 | Mean : 109527 | Mean : 20194232 | Mean :3.721e+07 | NA | Mean : 66.49 | Mean : 21763125 | Mean :37.130 | Mean :29.080 | Mean :2019 | Mean : 3778.2 |
|  | NA | 3rd Qu.: 721395 | 3rd Qu.: 13124.0 | 3rd Qu.: 694092 | 3rd Qu.: 37644 | 3rd Qu.: 11712582 | 3rd Qu.:3.245e+07 | NA | 3rd Qu.:112.75 | 3rd Qu.: 12816936 | 3rd Qu.:62.250 | 3rd Qu.:50.000 | 3rd Qu.:2020 | 3rd Qu.: 5298.3 |
|  | NA | Max. :42634054 | Max. :688486.0 | Max. :32598424 | Max. :9597842 | Max. :615684393 | Max. :1.396e+09 | NA | Max. :196.00 | Max. :761350715 | Max. :92.000 | Max. :83.000 | Max. :2020 | Max. :10948.5 |
|  | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA’s :120 | NA’s :120 |

Table 2 shows a table of the top 10 countries with the highest percentage of cases by population. We can see that there are a variety of locations and that Montenegro and Czech Republic are the highest 2 countries with the most cases by population.

|  |  |  |
| --- | --- | --- |
| Country | location | totcase\_percent |
| Montenegro | Europe | 19.81734 |
| Czech Republic | Europe | 15.69815 |
| Bahrain | Asia | 15.45937 |
| Maldives | Asia | 15.08056 |
| Georgia | Asia | 14.81760 |
| French Polynesia | Oceania | 14.20375 |
| Aruba | N. America | 14.14261 |
| Slovenia | Europe | 13.49234 |
| Israel | Asia | 12.95736 |
| United States | N. America | 12.78968 |

Table 3 shows a table of the top 5 and bottom 5 recovery rates also sharing the total cases numbers.It looks that Asia dominates the top 5 countries for recovery rate. For the bottom 5, there is a mixture of locations. We can also see that the recovery rate and death rate for the bottom 5 teams do not add up to 100 or close to it. This means there are a lot of active cases.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Country | location | recov\_rate | death\_rate | Total Cases |
| 1 | Bhutan | Asia | 0.9984592 | 0.0011556 | 2596 |
| 2 | Bahrain | Asia | 0.9917312 | 0.0050649 | 274041 |
| 3 | Qatar | Asia | 0.9899229 | 0.0025682 | 235187 |
| 4 | UAE | Asia | 0.9882621 | 0.0028292 | 731307 |
| 5 | Uruguay | S. America | 0.9800516 | 0.0156107 | 387299 |
| 160 | Laos | Asia | 0.3083227 | 0.0008860 | 18059 |
| 161 | Honduras | N. America | 0.3034003 | 0.0264979 | 356707 |
| 162 | Grenada | N. America | 0.2789700 | 0.0141018 | 3262 |
| 163 | Mauritius | Africa | 0.1317416 | 0.0034108 | 14073 |
| 164 | New Caledonia | Oceania | 0.0204010 | 0.0049244 | 2843 |

Table 4 shows a table of the top 5 and bottom 5 countries for death rate. We can see that Vanuatu only has a total of 4 cases but Yemen also has a significantly large rate then the next top country, Mexico. We can also see that Asia dominates the lowest death rates.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Country | location | recov\_rate | death\_rate | Total Cases |
| 1 | Vanuatu | Oceania | 0.7500000 | 0.2500000 | 4 |
| 2 | Yemen | Asia | 0.6222975 | 0.1892018 | 8557 |
| 3 | Mexico | N. America | 0.8164215 | 0.0761703 | 3549229 |
| 4 | Sudan | Africa | 0.8398947 | 0.0756579 | 38000 |
| 5 | Syria | Asia | 0.7602892 | 0.0697775 | 30153 |
| 160 | Maldives | Asia | 0.9758072 | 0.0027268 | 83248 |
| 161 | Qatar | Asia | 0.9899229 | 0.0025682 | 235187 |
| 162 | Bhutan | Asia | 0.9984592 | 0.0011556 | 2596 |
| 163 | Laos | Asia | 0.3083227 | 0.0008860 | 18059 |
| 164 | Singapore | Asia | 0.9252218 | 0.0007883 | 74848 |

Figure 1 shows a scatterplot figure of total cases by population. United States, India, and Brazil are some outliers of the countries.

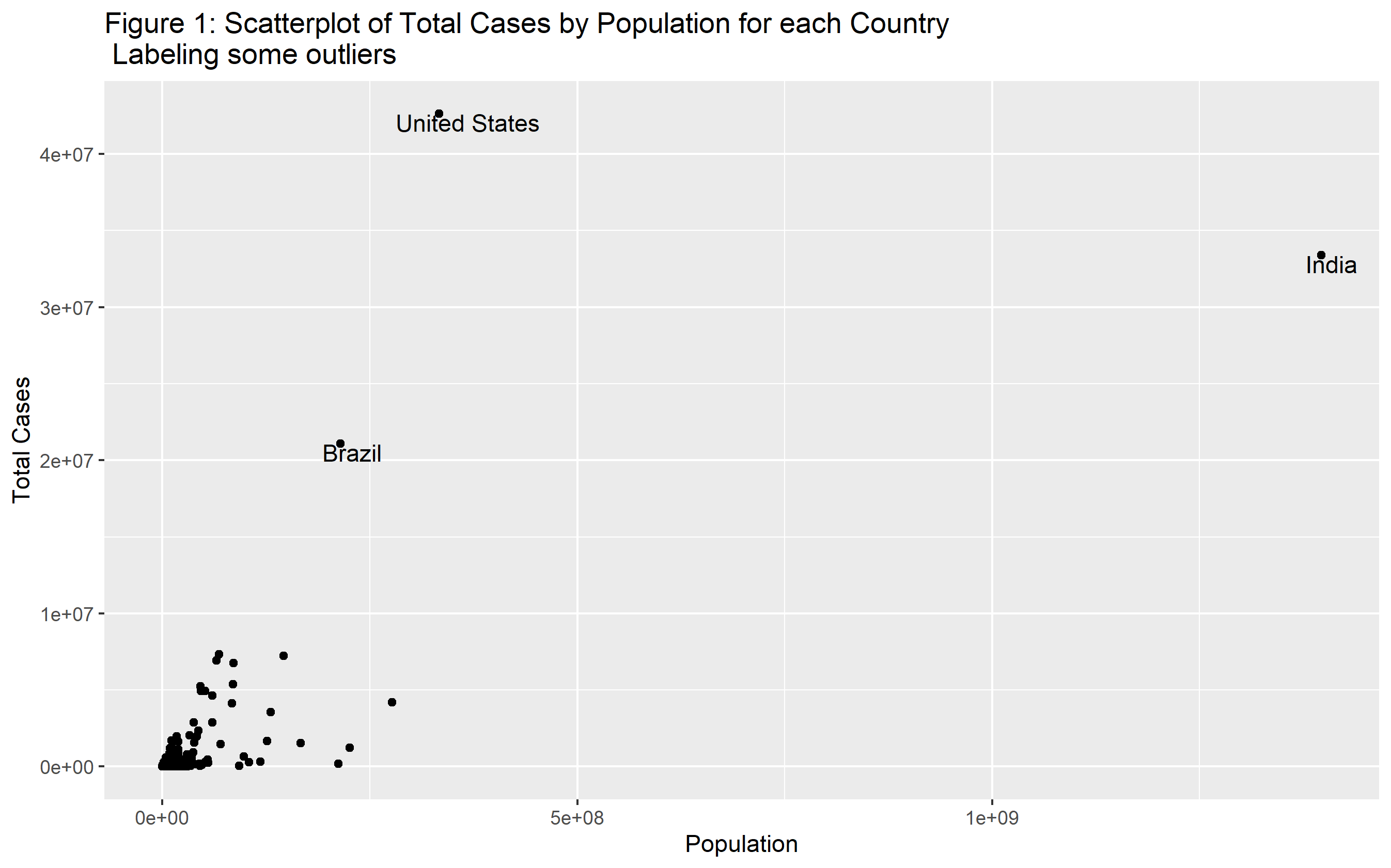


Figure 3.1: Analysis figure.

Figure 2 shows a scatterplot figure of total cases by population filtered down to countries with less than 40 million population. We can see that the total cases are spread out.

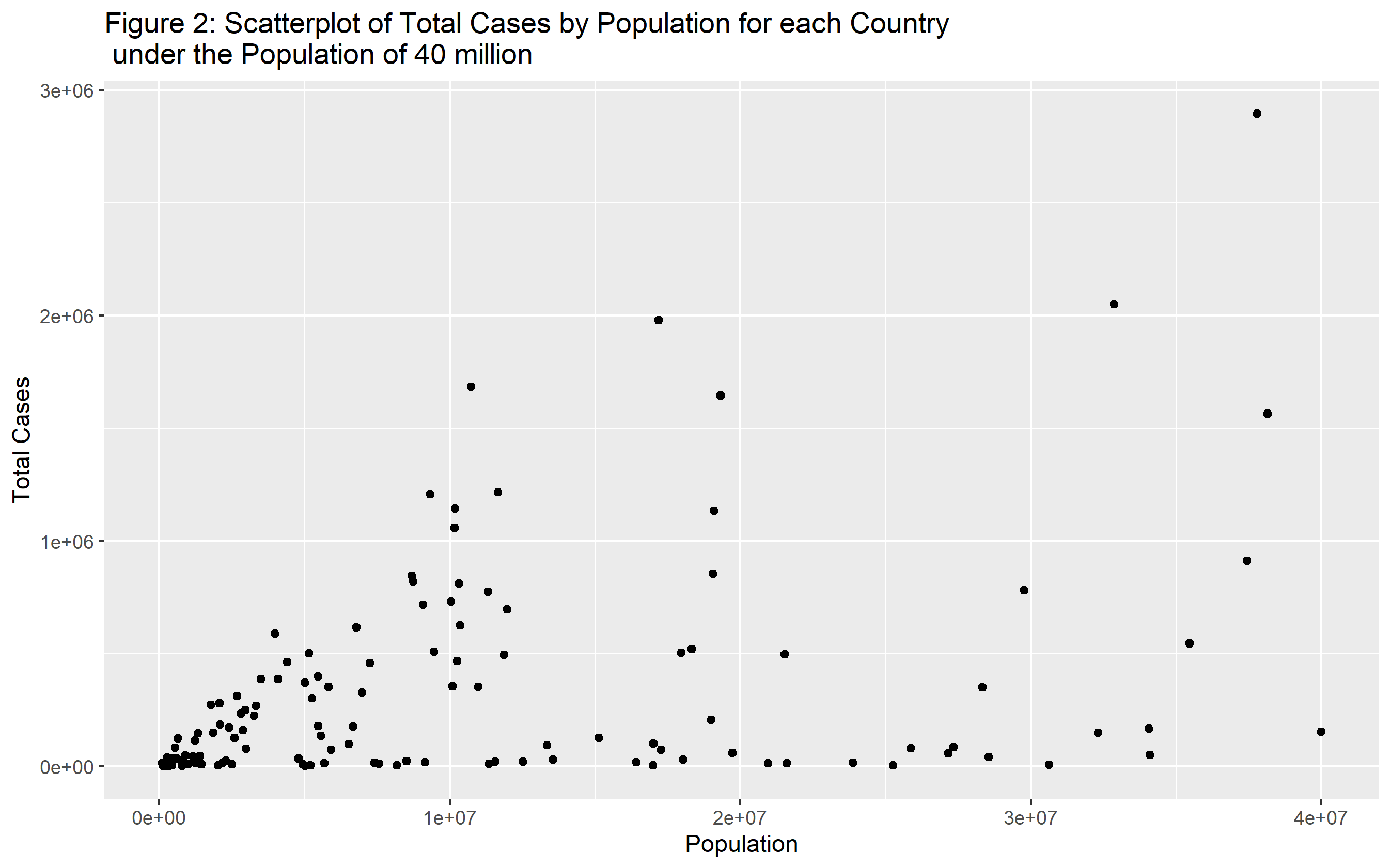


Figure 3.2: Analysis figure.

Figure 3 shows a scatterplot figure of total recovered patients by total cases. We can see just like in figure 1 that the United States, India, and Brazil have a overwhelmingly more cases than other countries. Therefore in figure 4, we will explore the majority of the countries.

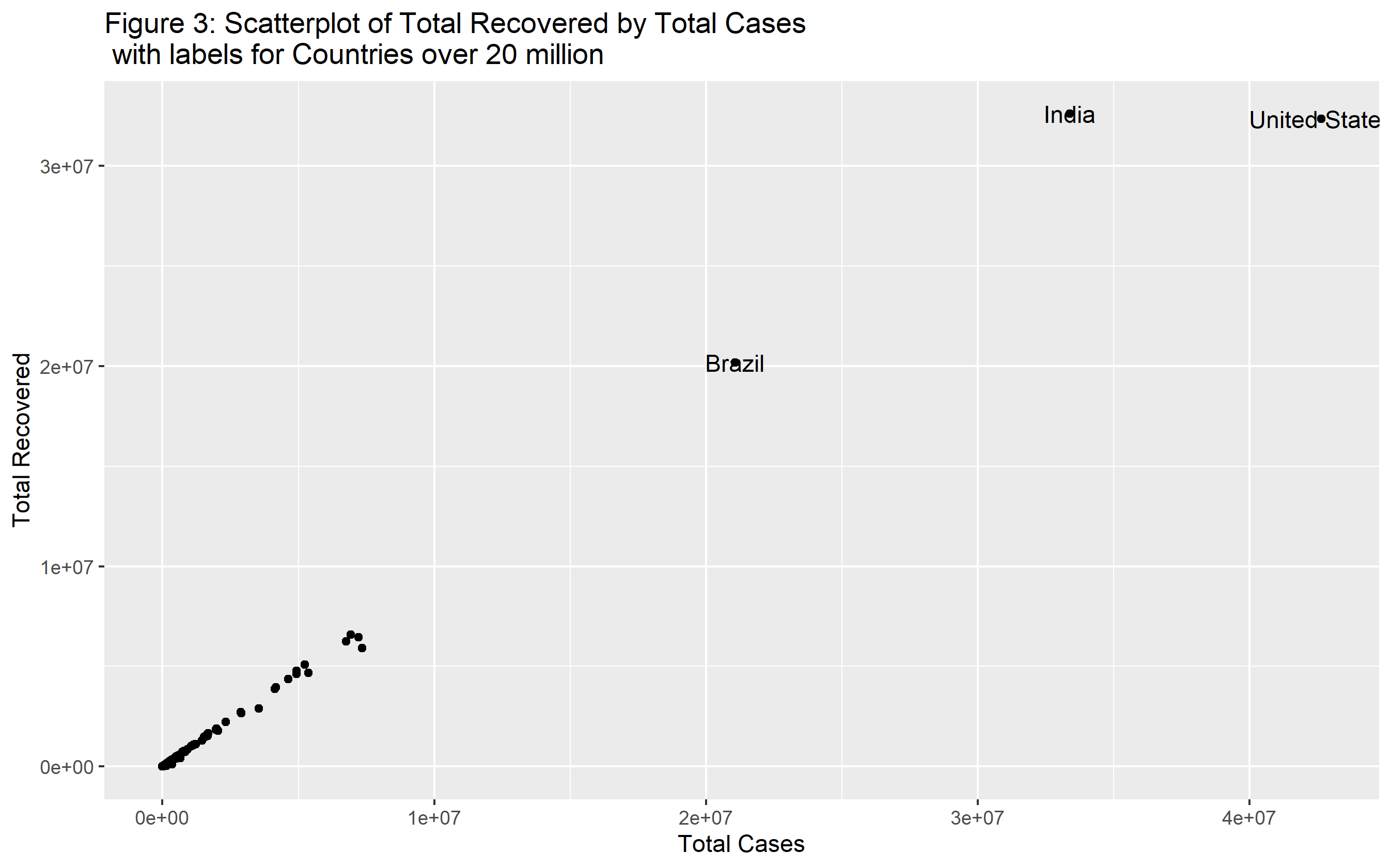


Figure 3.3: Analysis figure.

Figure 4 shows a scatterplot figure of the same plot as before just filtering countries with less than 20 million cases. We can see a linear relationship between the recovered to total cases. It looks like a slope of 1.

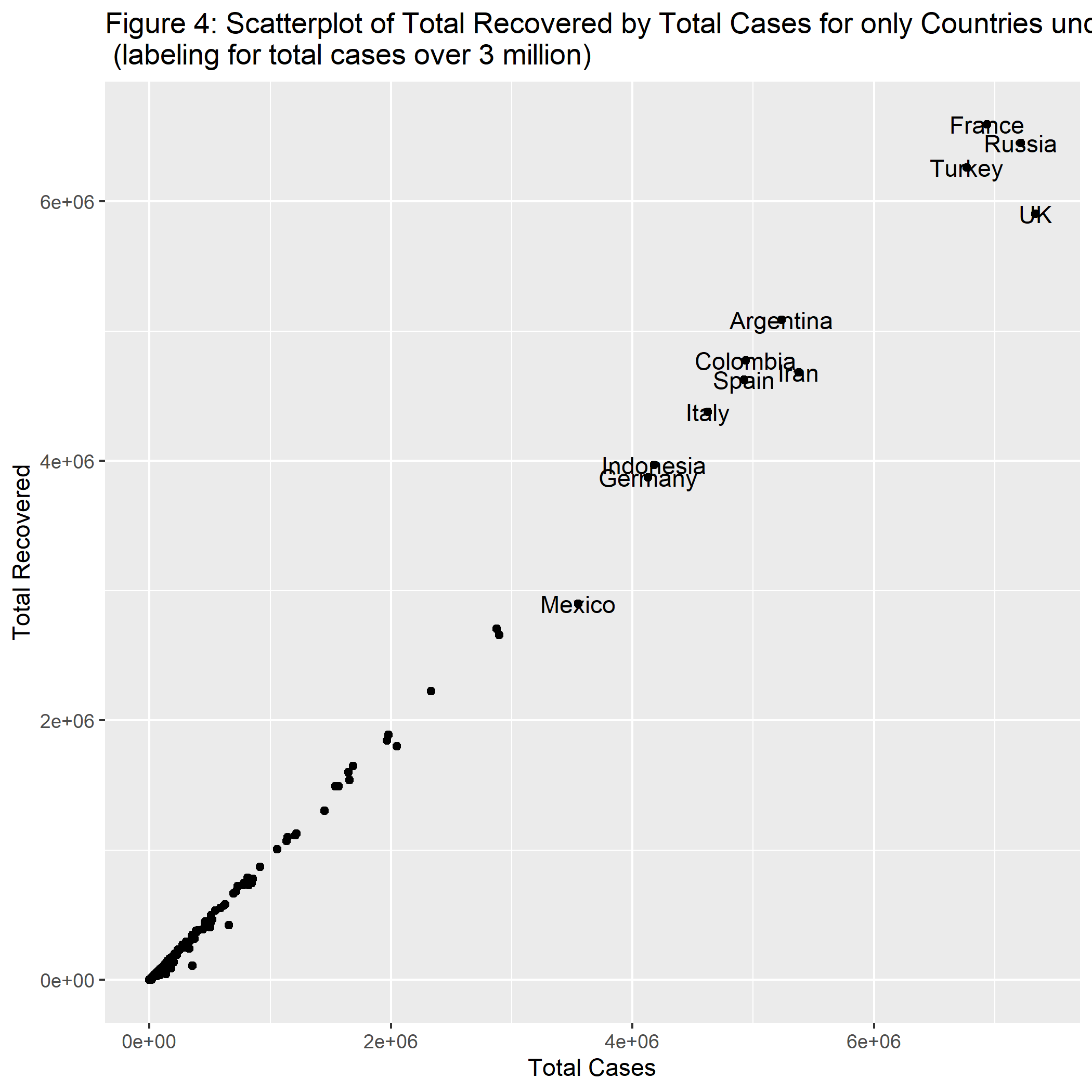


Figure 3.4: Analysis figure.

Figure 5 shows a histogram figure of the percent of fully vaccinated people. We can see that the histogram is right skewed where most of the countries have less than 25% of the population fully vaccinated.

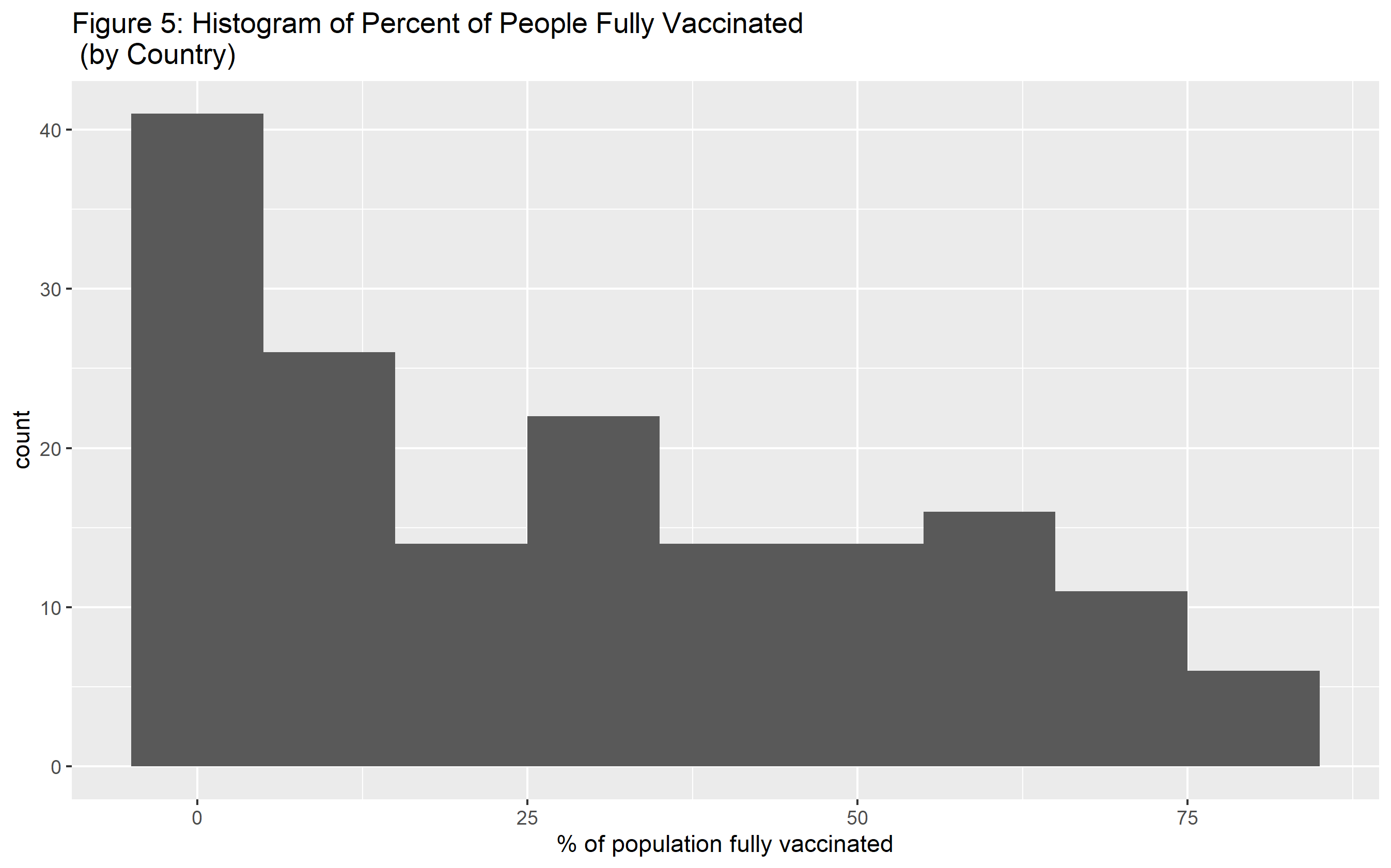


Figure 3.5: Analysis figure.

Figure 6 shows a scatterplot figure of percent of people fully vaccinated by health care funding. We can see that it is a linear relationship with a positive correlation. We can also see that the United States is out of the norm where they spend a lot more on health care funding compared to the percent of people fully vaccinated.

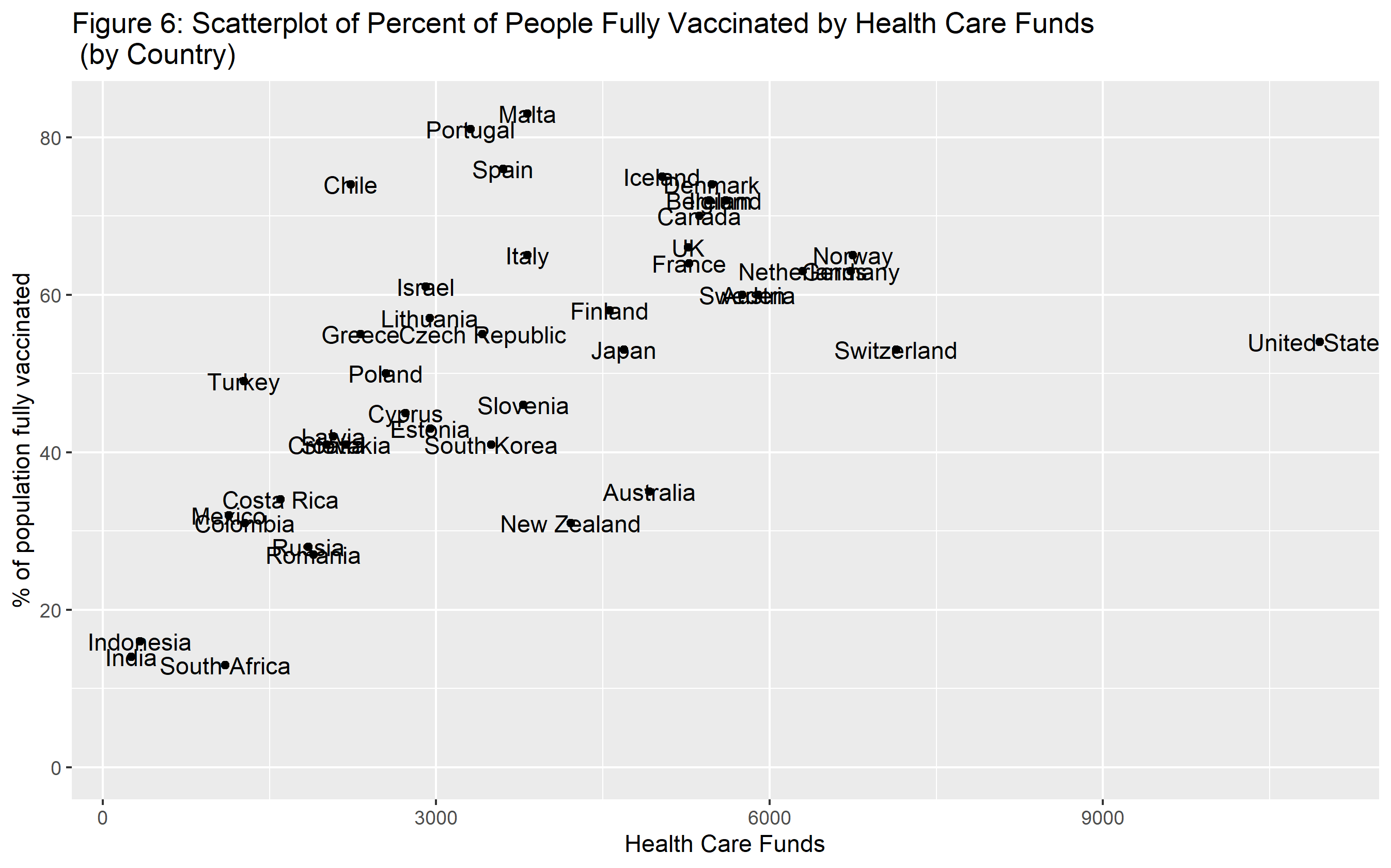


Figure 3.6: Analysis figure.

Figure 7 shows a histogram figure that describes the difference in proportions between those who recovered versus the proportion of those who died. From the histogram we can see that this difference is left skewed as most countries have a much hagher recovery rate with only a some having more of a lower difference meaning lower recovery rates.

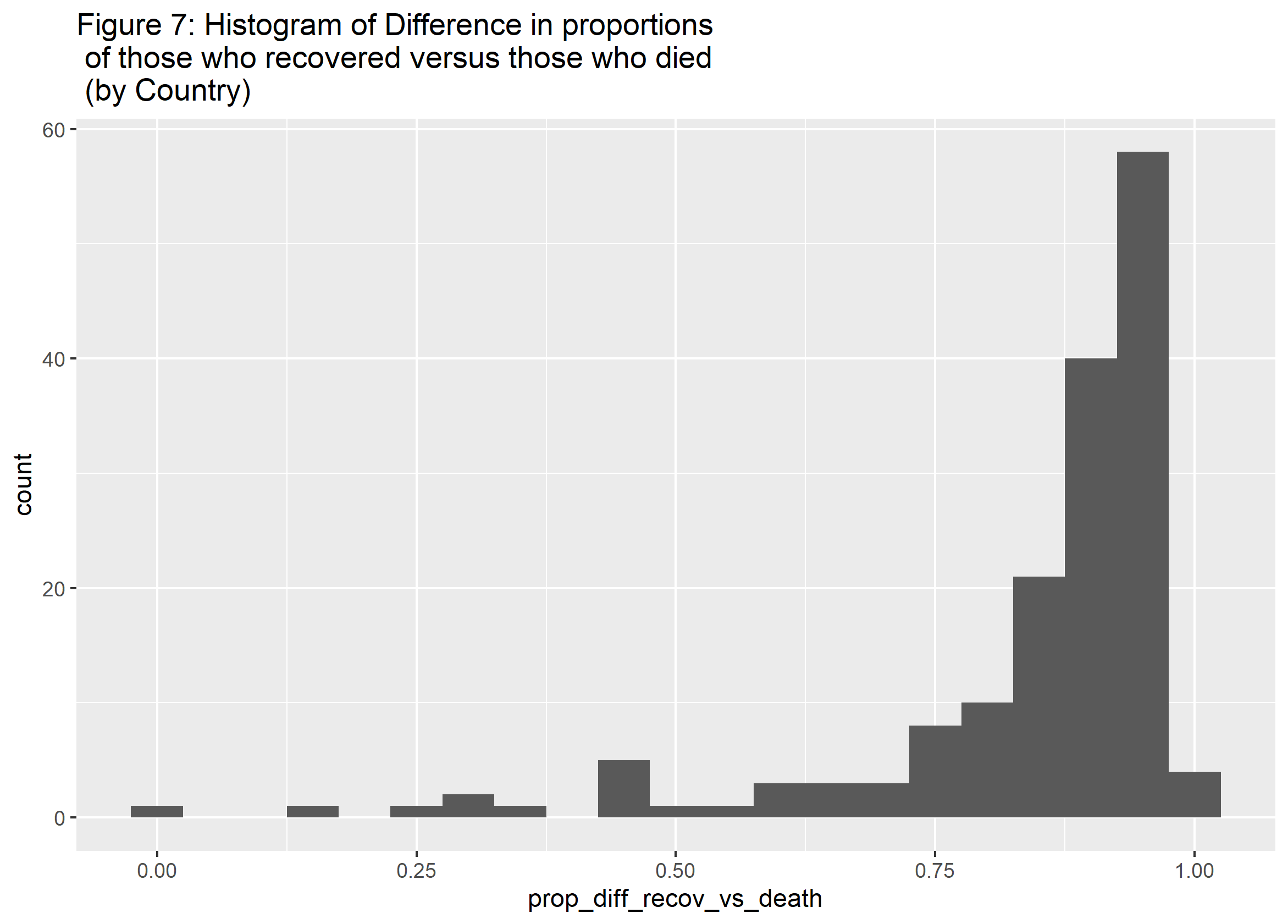


Figure 3.7: Analysis figure.

Figure 8 shows a histogram figure of the percentage of cases per population. We can see that this data is highly skewed to the right.

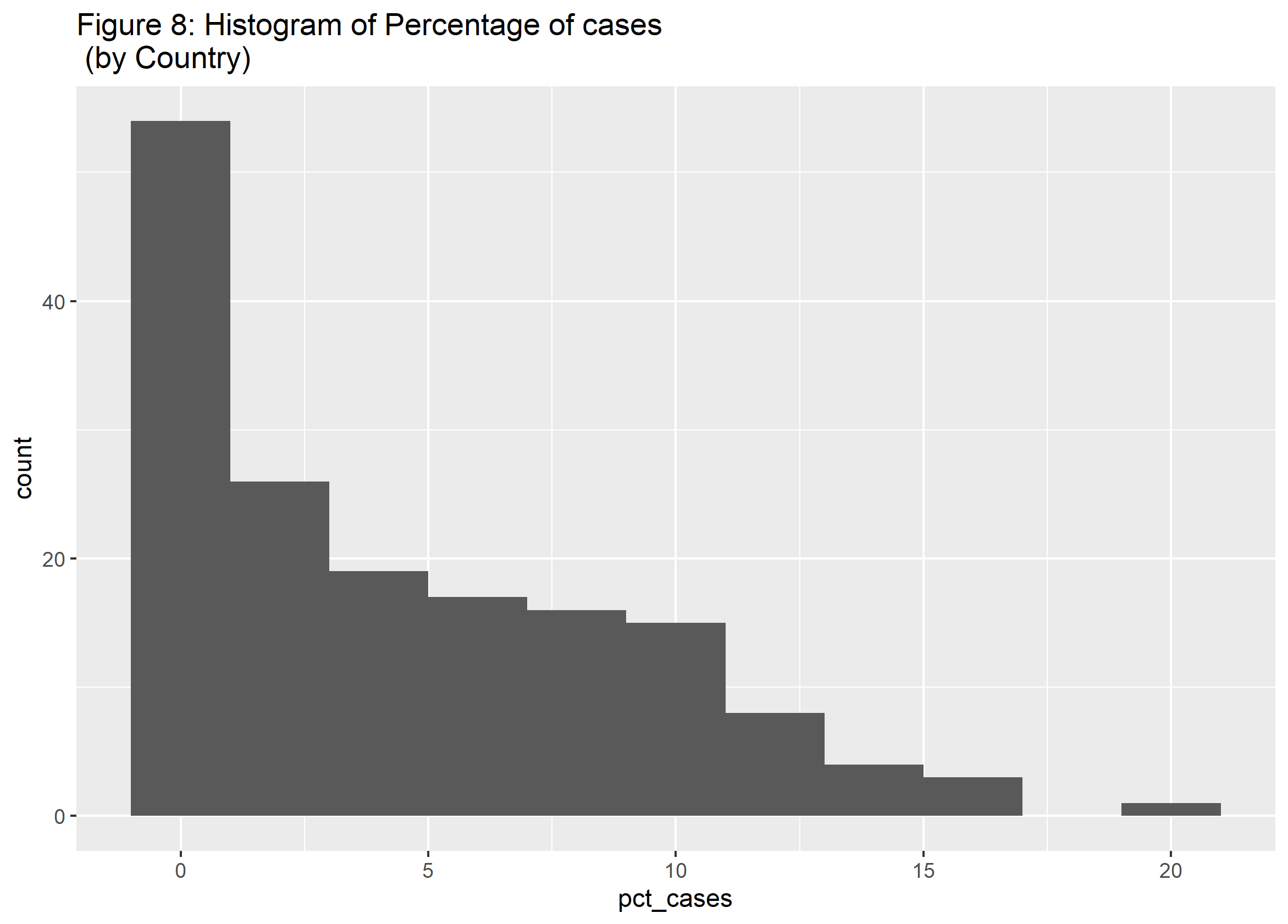


Figure 3.8: Analysis figure.

## 3.3 Full analysis

### 3.3.1 Difference in proportions of recovery and deaths

For my first question, we are looking at the difference in proportions of those who have recovered versus those how died. We can see in figure 7 that this data is fairly skewed. Therefore, I will do a logit transformation of the outcome to make it more normal as shown in figure 9.

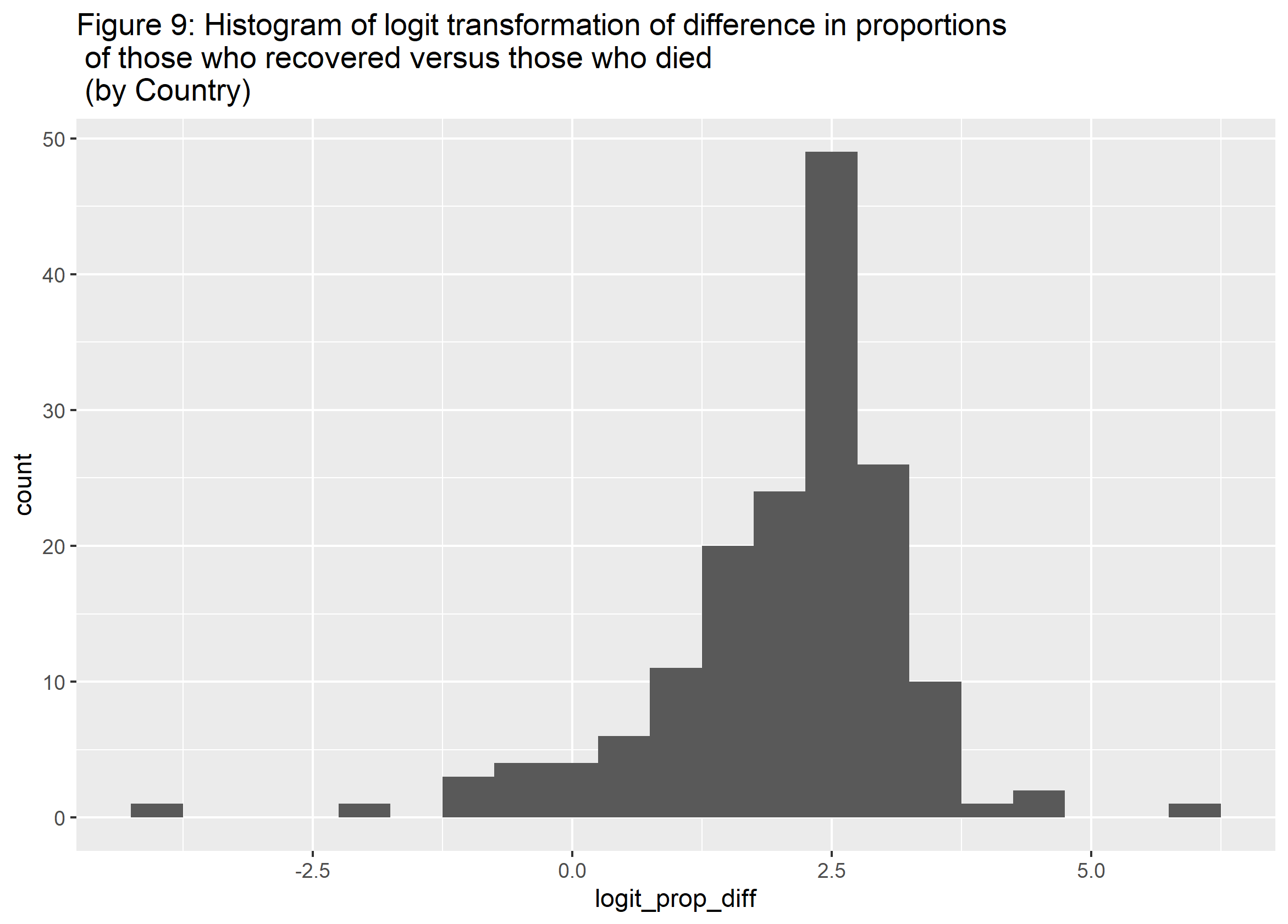


Figure 3.9: Analysis figure.

Before I decided on the logit transformation, I first decided to try this difference in proportions outcome I made within a binomial model and then a quasibinomial using the values of total cases as my weight to see how the deviance and dispersion values are. After running two simple models, the results were showing very high values for these statistics. Therefore this is how I decided on using the logit transformation.

Using the logit transformation, I ran 4 simple linear regression models. Below I provide the statistics of each of the models within Tables 5-8.

Table 5 is showing statistics for the predictor vaccine percentage.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.0403567 | 0.0343962 | 1.184012 | 6.770678 | 0.0101312 | 1 | -257.8129 | 521.6257 | 530.907 | 225.7033 | 161 | 163 |

Table 6 is showing statistics for the predictor health care funds.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.0482893 | 0.0256295 | 0.8814947 | 2.131057 | 0.1517811 | 1 | -55.85986 | 117.7197 | 123.0723 | 32.63538 | 42 | 44 |

Table 7 is showing statistics for the predictor number of tests per person.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.0554038 | 0.0495367 | 1.174692 | 9.443193 | 0.0024895 | 1 | -256.5248 | 519.0497 | 528.3309 | 222.1643 | 161 | 163 |

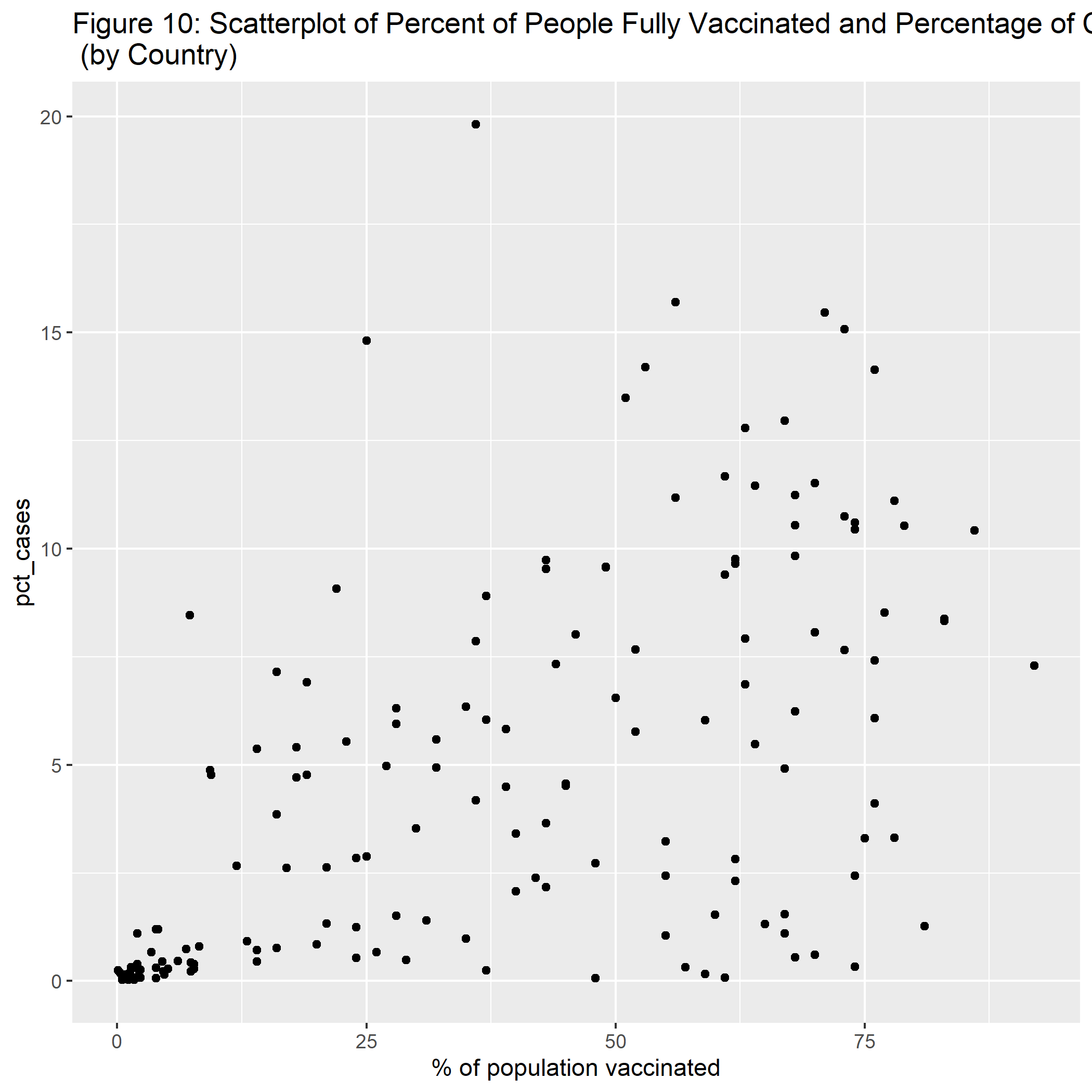
Table 8 is showing statistics for the predictor location which specifies the continent for each country.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.1094214 | 0.0810591 | 1.155049 | 3.857979 | 0.0025218 | 5 | -251.7256 | 517.4512 | 539.1075 | 209.4596 | 157 | 163 |

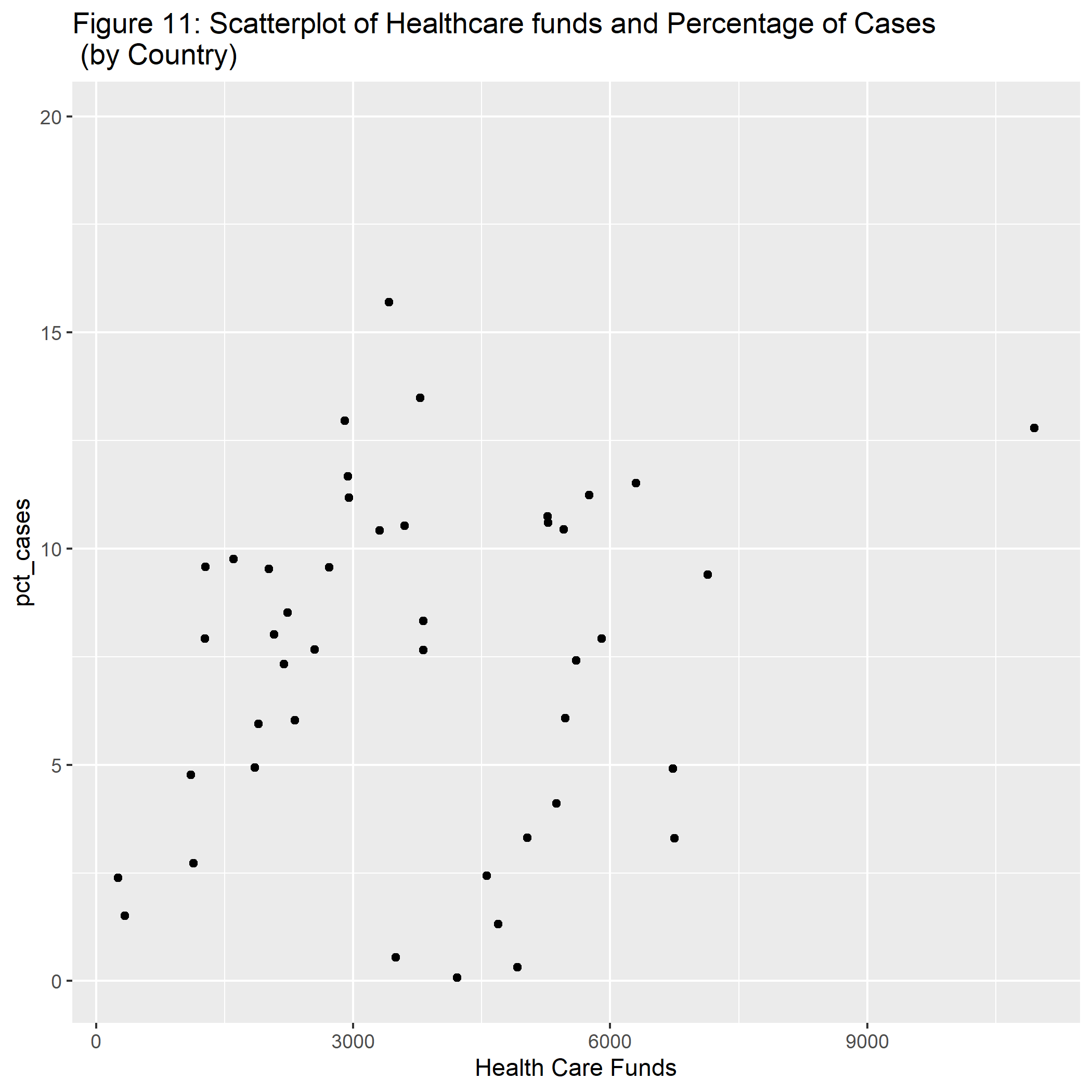
Looking at all of these tables, we can see that the R^2 values are very low. Therefore none of these predictors shows any association with the logit transformation outcome. Since not even one predictor is good for this outcome, I am going to spot my analysis with this outcome and go to my second question which uses the percent of cases within each country.

### 3.3.2 Percent of cases for each country

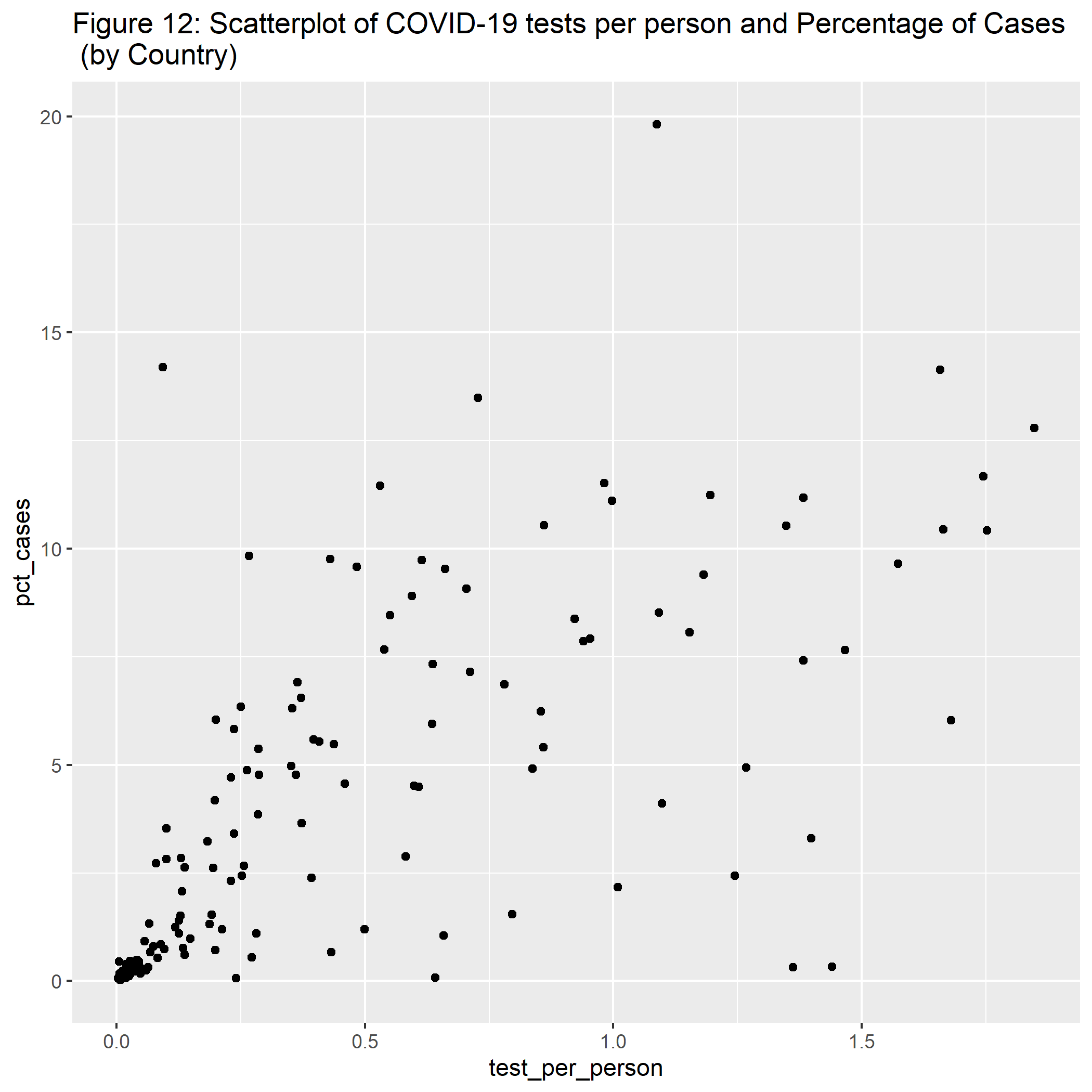
We can see with looking at figure 8 that the percentage of cases are extremely right skewed. For now, I will proceed with doing some simple linear regression. Below shows model statistics for each model and a visual showing the relationship between the two variables.

Figure 10 is showing statistics for the predictor vaccine percentage. 

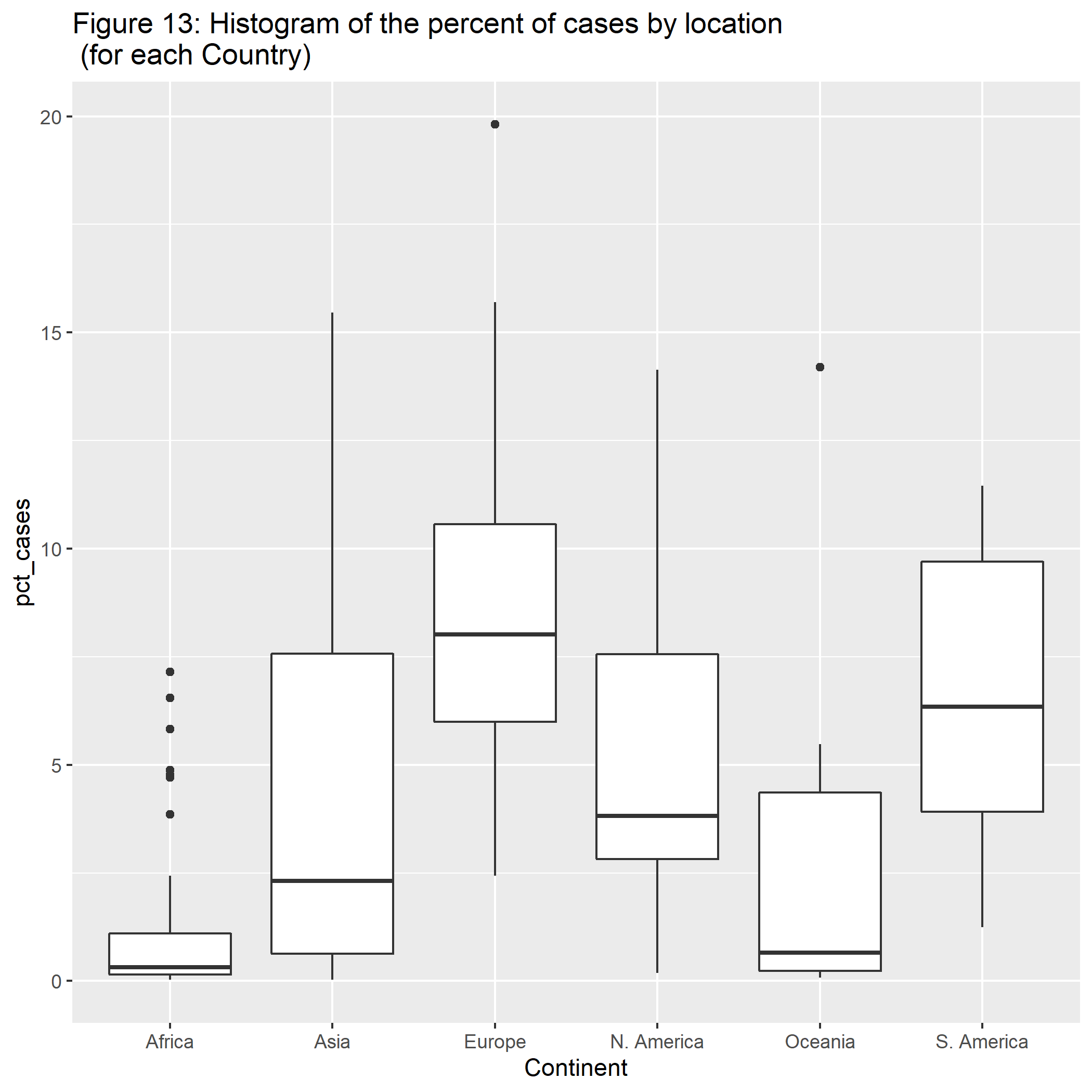
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.3312598 | 0.3271062 | 3.652049 | 79.7512 | 0 | 1 | -441.4128 | 888.8256 | 898.1069 | 2147.332 | 161 | 163 |

Figure 11 is showing statistics for the predictor health care funds. 

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.0253885 | 0.0021834 | 3.942663 | 1.094093 | 0.3015489 | 1 | -121.7715 | 249.5431 | 254.8956 | 652.8729 | 42 | 44 |

Figure 12 is showing statistics for the predictor number of tests per person. 

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.1366386 | 0.1312761 | 4.149581 | 25.48043 | 1.2e-06 | 1 | -462.231 | 930.462 | 939.7432 | 2772.262 | 161 | 163 |

Figure 13 is showing statistics for the predictor location which specifies the continent for each country. 

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
| 0.3704216 | 0.3503714 | 3.588359 | 18.47465 | 0 | 5 | -436.4947 | 886.9893 | 908.6456 | 2021.583 | 157 | 163 |

From these visuals, we can see that the R^2’s for these models are still relatively low but a simple linear regression of using percent vaccination or the location of each country have R^2’s of around 0.35 which is an improvement from the linear regression models from using logit transformation of difference in proportions as the outcome.

# 4 Discussion

## 4.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 4.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 4.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

# 5 References