STA 207 Project (Title TBD)

Eric Chagnon

Abstract

TODO

Introduction

The Coronovirus pandemic that started in 2019 has had a lasting impact on the world. As of writing this report over 5.9 million people lost their lives to the Coronavirus, better known as Covid-19. Covid-19 affected each country differently, this could be due to a multitude of things like vaccination rates, differences in mandated public health policies, public opinion on Covid-19, etc. The goal of this paper is to try and determine how well a country protected its citizens from the pandemic by analyzing countries' health metrics. In order for large and small countries to be compared more fairly, the response variables used will be the percentage of the country's population that died from Covid-19, and the predictor variables will be Gross National Income (GNI), and the Literacy Rate.

Background

The data sources for this project consist of the World Health Organization (WHO), the World Bank, and the Central Intelligence Agency (CIA). The WHO maintains a Covid-19 dashboard that updates daily cases and deaths for countries around the world and keeps a count of cumulative deaths and cases starting from January 3, 2020. The CIA maintains a dataset for metrics of individual countries such as GPD, Area, Net Migration, etc. and is available on Kaggle. Finally, every year the World Bank updates a dataset containing the GNI of individual countries.

Descriptive Analysis

A moderate amount of data cleaning is required to obtain the final dataset. In order to obtain the response variable the Cumulative Deaths of each country from the WHO dataset is divided by the country's Population from the CIA dataset. This provides a percentage of the population that had Covid-19 related deaths. Secondly, the GNI of each country from the World Bank dataset, the Literacy Rate from the CIA dataset, and the new response variable need to be joined together via an inner join on the country name. After removing NA values, the final amount dataset contains 157 countries. The GNI treatment effect was split into four categories: Low, Medium-Low, Medium-High, and High as per (Hamadeh, et. al, 2020), and the Literacy Rate was spliting into three categories: Low, Medium, and High as per (World Population Review, 2022). The amount of countries in each level of GNI and Literacy rate can be seen in the table below:

	High	Low	Medium
High	38	0	12
Low	0	18	4
Medium_High	14	3	28
Medium_Low	5	13	22

The literacy rate was split into three categories: Low, Medium, and High. The distributions of the GNI and the Literacy Rates of each country can be seen in the figures below.

Figure 1: Percentage Dead by GNI

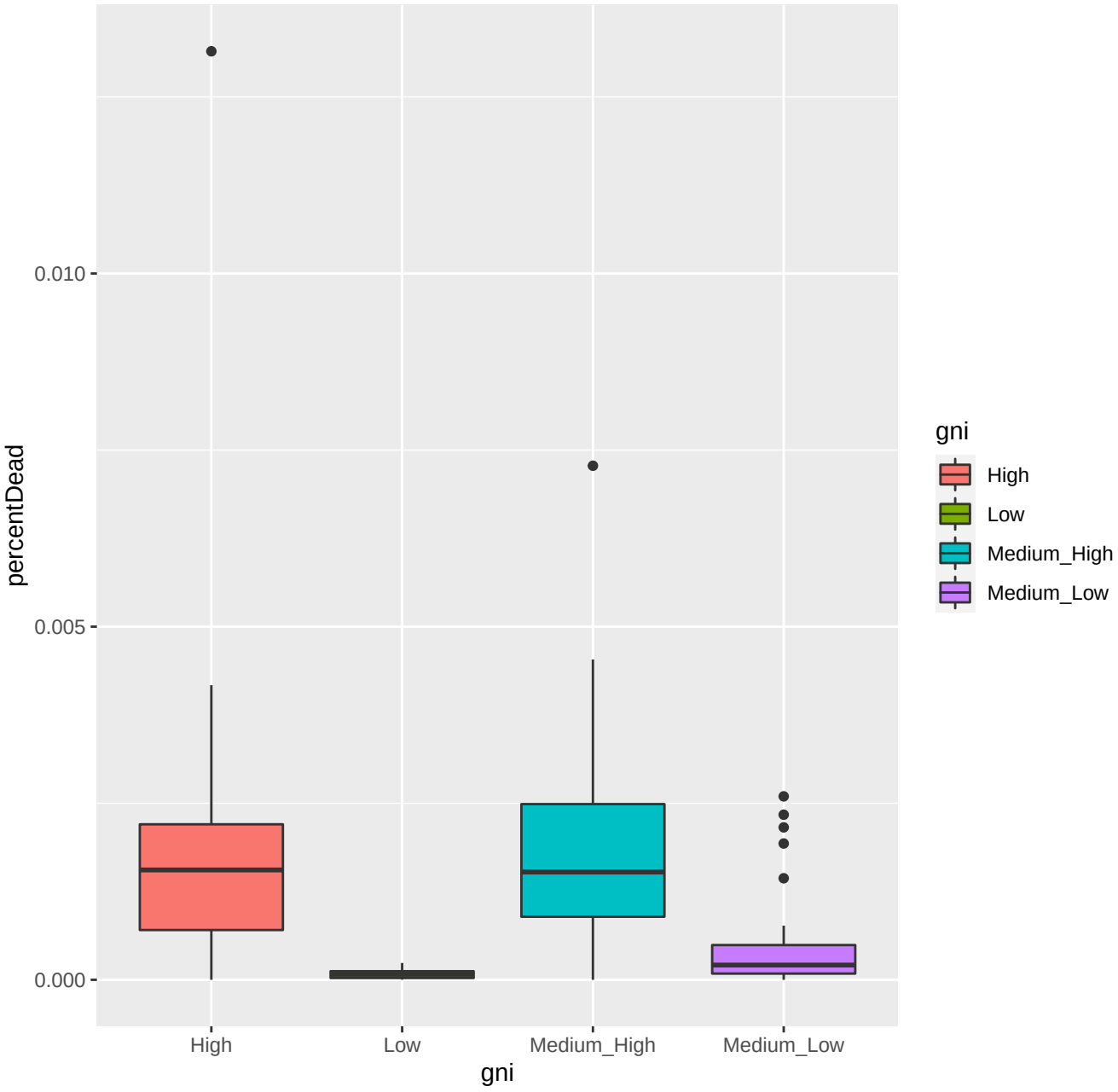
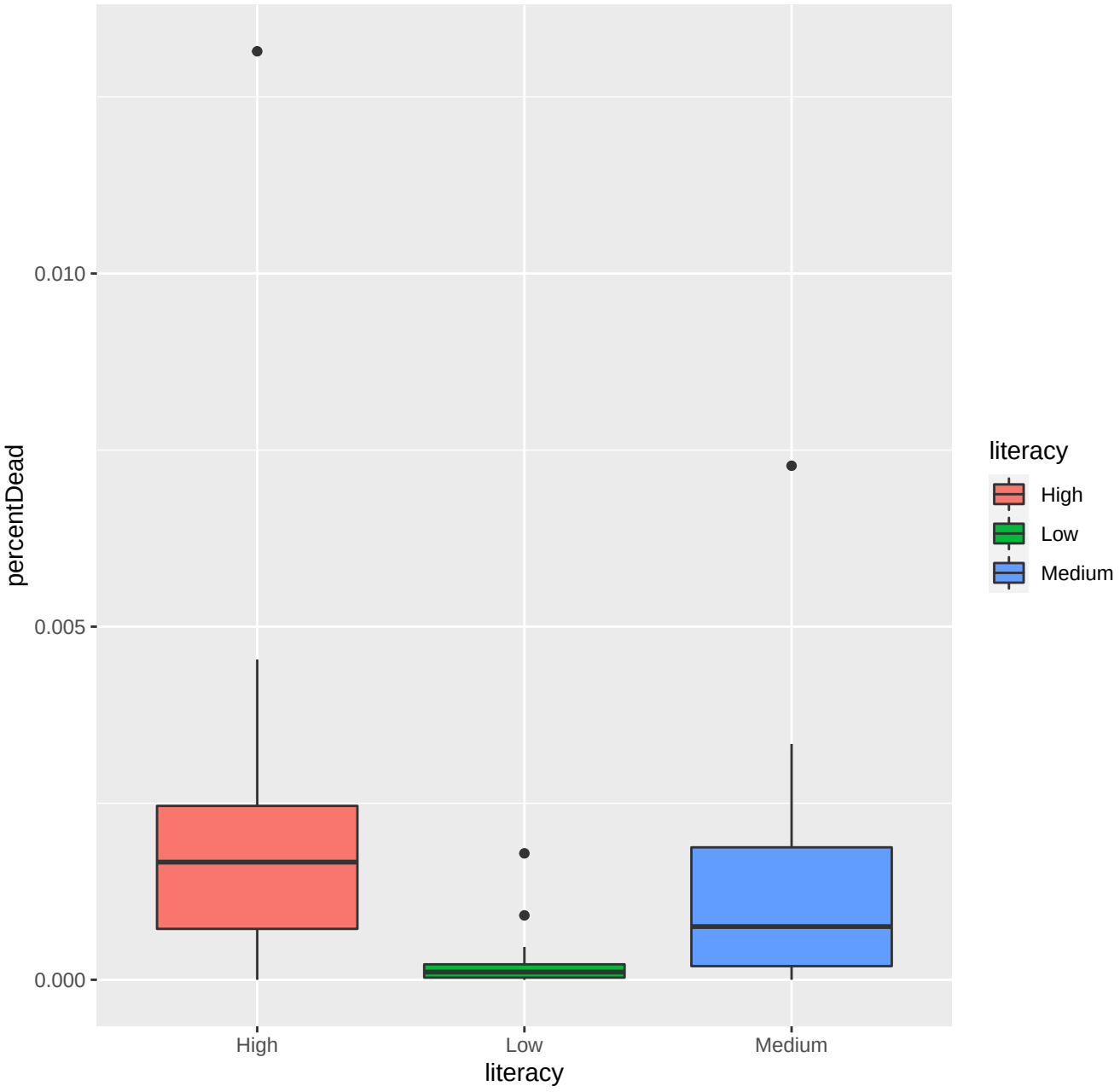


Figure 2: Percentage Dead by Literacy Rate



From the above Figures it seems as though the countries with Low GNI and Low Literacy rate have the smallest mean of Percentage of Population Deaths.

Predictive/Inferential Anaysis

In order to determine the effects of GNI and Literacy rate on the Percentage of Population Deaths, and two-way ANOVA model was constructed with the form:

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \epsilon_{i,j,k}, i = 1, ..., 4, j = 1, .., 3, k = 1, 2, ..., n_{i,j}$$

Where α_i satisfies $\Sigma_{i=1}^{4} \alpha_i = 0$, β_j satisfies $\Sigma_{j=1}^{3} \beta_j = 0$, and $\epsilon_{i,j,k}$ are i.i.d $N(0, \sigma^2)$.

In this model μ is the average Percentage of a county's population that died to Covid-19, α_i is the main effect of GNI, and β_j is the main effect of Literacy Rate.

```
Analysis of Variance Table
```

```
Response: percentDead
           Df     Sum Sq    Mean Sq F value    Pr(>F)
gni         3 7.5921e-05 2.5307e-05 14.1644 3.497e-08 ***
literacy    2 8.5250e-06 4.2624e-06  2.3857   0.09549 .
Residuals 151 2.6979e-04 1.7867e-06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

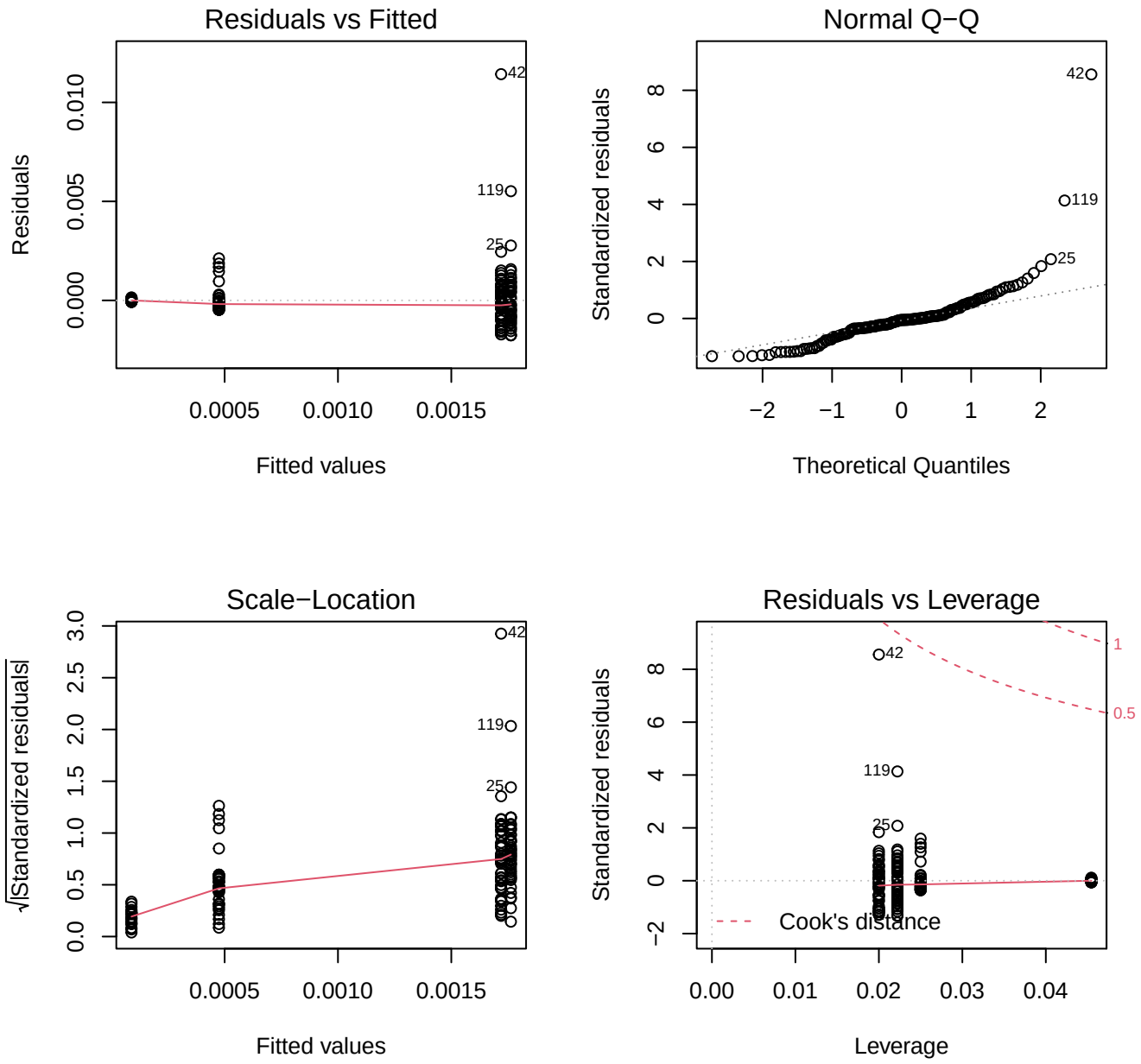
Here it is evident that the Literacy Rate does not have a significant effect. So removing it from the model yields the following:

```
Analysis of Variance Table

Response: percentDead
           Df     Sum Sq    Mean Sq F value    Pr(>F)
gni         3 7.5921e-05 2.5307e-05  13.912 4.549e-08 ***
Residuals 153 2.7831e-04 1.8190e-06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sensitivity Analysis

The above model has the following diagnostic plots.



The Normal Q-Q plot shows that the approximate normal distribution of errors assumption may be violated. A Levene Test was carried out with $H_0 : E[d_1] = E[d_2] = ... = E[d_r]$ where $d_r = |Y_{ij} - \bar{Y}_i|$.

```
data$res.abs = abs(aov.fit2$residuals)
summary(aov(res.abs~percentDead, data=data))
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
percentDead   1 1.026e-04 1.026e-04   179.3 <2e-16 ***
Residuals   155 8.871e-05 5.700e-07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the test statistic has a p-value $< 2e16$ there is sufficient evidence to reject H_0, and conclude that the errors are not from the same distribution.

Conclusion

Since the assumptions required for ANOVA are violated another approach should be taken in order to determine the relationship between GNI and Percentage of Population that died.

Refrences

World Health Organization. (n.d.). Covid-19 Dashboard. World Health Organization. Retrieved March 4, 2022, from https://covid19.who.int/table

GNI per capita, Atlas method (current US$). Data. (n.d.). Retrieved March 4, 2022, from https://data.worldbank.org/indicator/NY.GNP.PCAP.CD

Abhishek252. (2021, April 14). CIA country dataset for unsupervised learning. Kaggle. Retrieved March 4, 2022, from https://www.kaggle.com/abhishek252/cia-country-dataset-for-unsupervised-learning

Hamadeh, N., Van Rompaey, C., & Metreau, E. (2021, July 1). New World Bank country classifications by Income Level: 2021-2022. World Bank Blogs. Retrieved March 4, 2022, from https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2021-2022

World Population Review. Literacy Rate by Country 2022. Retrieved Mar 4, 2022, from https://worldpopulationreview.com/country-rankings/literacy-rate-by-country

Appendix

```
library(ggplot2)
setwd("~/Downloads")
data = read.csv("final_data.csv")

table(data$gni, data$literacy)

ggplot(data=data, aes(x = gni, y = percentDead, fill = gni)) + geom_boxplot() + ggtitle("Figure 1: Percentage

ggplot(data=data, aes(x = literacy, y = percentDead, fill = literacy)) + geom_boxplot() + ggtitle("Figure 2: P

aov.fit = aov(data = data, percentDead ~ gni + literacy)
anova(aov.fit)

aov.fit2 = aov(data = data, percentDead ~ gni)
anova(aov.fit2)
par(mfrow=c(2,2))
plot(aov.fit2)
par(mfrow=c(1,1))

data$res.abs = abs(aov.fit2$residuals)
summary(aov(res.abs~percentDead, data=data))
```