R Notebook

DATA 602 PROJECT

Zheyu (Jerry) Song, Hao (Leo) Su, Kelly Wu (Group 11)

Introduction & Background:

For our project we explored the relationship between COVID-19 Pandemic outcomes, such as positive cases of COVID-19 and related deaths to COVID-19 related to obesity.

We are interested in obesity as a factor in COVID-19 related outcomes because obesity is a non-communicable chronic disease that can impact one’s overall health and their ability to defend themselves against other illnesses, such as COVID-19. Already, a study from 2021 on the USA population found that the the risk for severe COVID-19 related outcomes increased with higher BMI (Kompaniyets et al., 2021).

Before we look into the relationship between COVID-19 outcomes and obesity directly, we wish to understand the relationship between Gross Domestic Product (GDP) and COVID-19. The reason for looking into the GDP’s relationship with obesity, is to understand if the type of country, i.e. developed/developing matters when looking at COVID-19 outcomes and the the obesity rate. In early 2022, a research paper by Oshakbayev et al., discussed findings that of analyses that showed significant correlations between GDP and obesity, and overweight prevalence (2022).

Datasets Used:

Overweight and obesity based on measured body mass index, by age group and sex (Open Government, 2021)

COVID-19 Healthy Diet Dataset (Kaggle, 2020)

Cleaned data set from DATA 601 (dervied from the COVID-19 Healthy Diet Dataset)

Guiding Questions

Question 1:

Firstly, we analyze the trends of obesity in Canada and it’s relationship with Canada’s Gross Domestic Product (GDP) value. We will do this using statistical hypotheses.

Question 2:

Secondly, we will analyze and model the relationship between obesity and each COVID-19 outcome, deaths as a result of COVID-19 and confirmed positive cases of COVID-19. We will do this using linear regression and bootstrapping methods.

Reading the data

```
food = read.csv("Food_Supply_Quantity_kg_Data.csv")
canada = read.csv("13100373.csv")
processed = read.csv("haha.csv")
head(food, 3)
```

	Country	Alcoholic.Beverages	Animal.fats	Animal.Products	Aquatic.Products..Other	▸
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	
1	Afghanistan	0.0014	0.1973	9.4341	0	

	Country <chr>	Alcoholic.Beverages <dbl>	Animal.fats <dbl>	Animal.Products <dbl>	Aquatic.Products..Other <dbl>	▶
2	Albania	1.6719	0.1357	18.7684	0	
3	Algeria	0.2711	0.0282	9.6334	0	

3 rows | 1-6 of 33 columns

```
head(canada, 3)
```

	REF_DATE <int>	GEO <chr>	DGUID <chr>	Measures <chr>	Sex <chr>	Age.group <chr>	▶
1	2009	Canada	2016A000011124	Overweight	Both sexes	Ages 5 to 79	
2	2009	Canada	2016A000011124	Overweight	Both sexes	Ages 5 to 79	
3	2009	Canada	2016A000011124	Overweight	Both sexes	Ages 5 to 79	

3 rows | 1-7 of 19 columns

```
head(processed, 3)
```

	Country <chr>	Alcoholic.Beverages <dbl>	Animal.fats <dbl>	Aquatic.Products..Other <dbl>	Cereals...Excluding.Be <db
1	afghanistan	0.0028	0.3946	0	49.619
2	albania	3.3438	0.2714	0	11.563
3	algeria	0.5422	0.0564	0	27.363

3 rows | 1-6 of 31 columns

Data wrangling

```
food <- select(food, c('Country', 'Obesity', 'Undernourished', 'Confirmed', 'Deaths', 'Recovered'))
food <- na.omit(food)
head(food, 3)
```

	Country <chr>	Obesity <dbl>	Undernourished <chr>	Confirmed <dbl>	Deaths <dbl>	Recovered <dbl>
1	Afghanistan	4.5	29.8	0.1421342	0.006185779	0.1233739
2	Albania	22.3	6.2	2.9673009	0.050951374	1.7926357
3	Algeria	26.6	3.9	0.2448971	0.006558153	0.1675722

3 rows

```
canada <- select(canada, c('REF_DATE', 'Measures', 'Sex', 'Age.group', 'Characteristics', 'VALUE'))
head(canada, 3)
```

	REF_D... <int>	Measures <chr>	Sex <chr>	Age.group <chr>	Characteristics <chr>	VA... <dbl>

	REF_D…	Measures	Sex	Age.group	Characteristics	VA…
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>
1	2009	Overweight	Both sexes	Ages 5 to 79	Percent	33.6
2	2009	Overweight	Both sexes	Ages 5 to 79	Low 95% confidence interval, percent	30.8
3	2009	Overweight	Both sexes	Ages 5 to 79	High 95% confidence interval, percent	36.6

3 rows

```
canada.percent <- filter(canada, Measures == 'Obese' & Sex == 'Both sexes' & Age.group == 'Ages 5 to 79' & Characteristics == 'Percent')

canada.number <- filter(canada, Measures == 'Obese' & Sex == 'Both sexes' & Age.group == 'Ages 5 to 79' & Characteristics == 'Number of persons')

head(canada.percent, 3)
```

	REF_DATE	Measures	Sex	Age.group	Characteristics	VALUE
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>
1	2009	Obese	Both sexes	Ages 5 to 79	Percent	22.4
2	2011	Obese	Both sexes	Ages 5 to 79	Percent	23.8
3	2013	Obese	Both sexes	Ages 5 to 79	Percent	24.2

3 rows

```
head(canada.number, 3)
```

	REF_DATE	Measures	Sex	Age.group	Characteristics	VALUE
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>
1	2009	Obese	Both sexes	Ages 5 to 79	Number of persons	6479500
2	2011	Obese	Both sexes	Ages 5 to 79	Number of persons	7174000
3	2013	Obese	Both sexes	Ages 5 to 79	Number of persons	7424000

3 rows

```
processed <- select(processed, c('Country', 'Obesity', 'Undernourished', 'Confirmed', 'Deaths', 'region'))
head(processed,3)
```

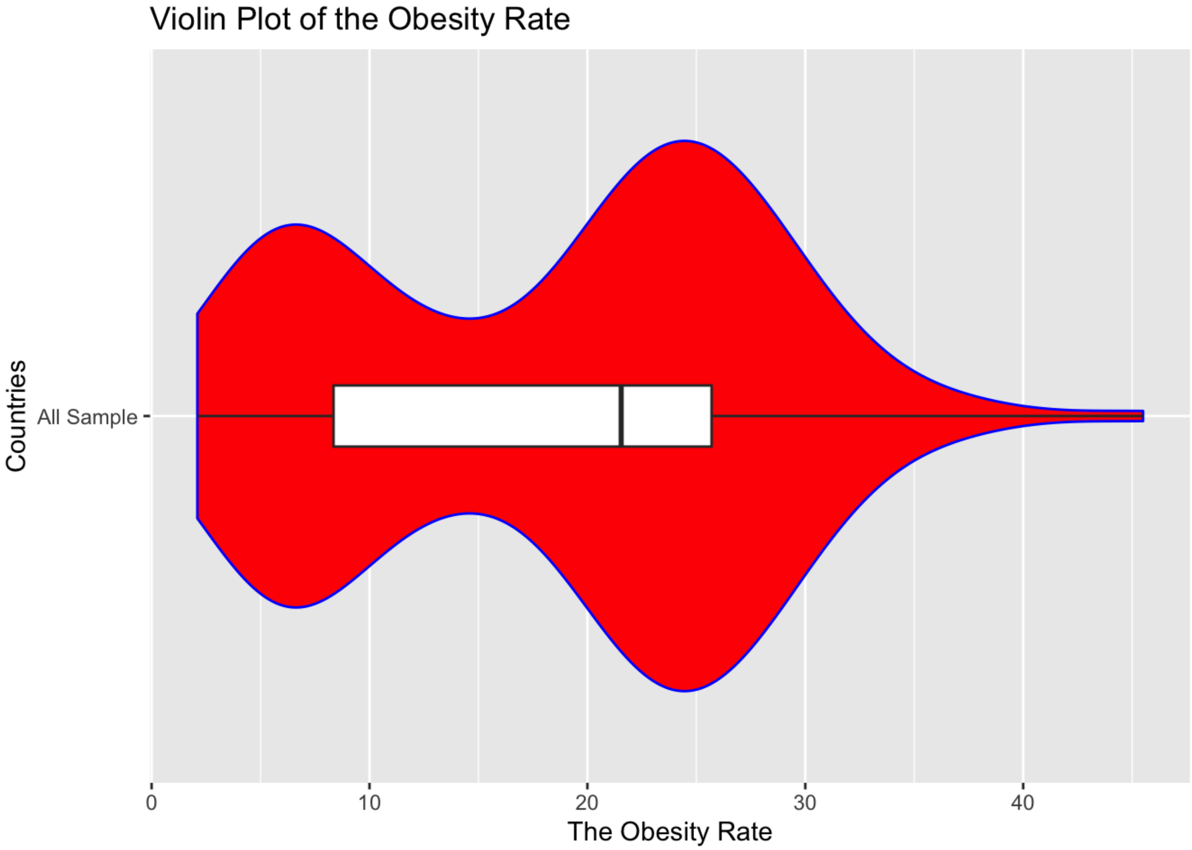
	Country	Obesity	Undernourished	Confirmed	Deaths	region
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	afghanistan	4.5	29.8	0.1421342	0.006185779	Asia
2	albania	22.3	6.2	2.9673009	0.050951374	Europe
3	algeria	26.6	3.9	0.2448971	0.006558153	Africa

3 rows

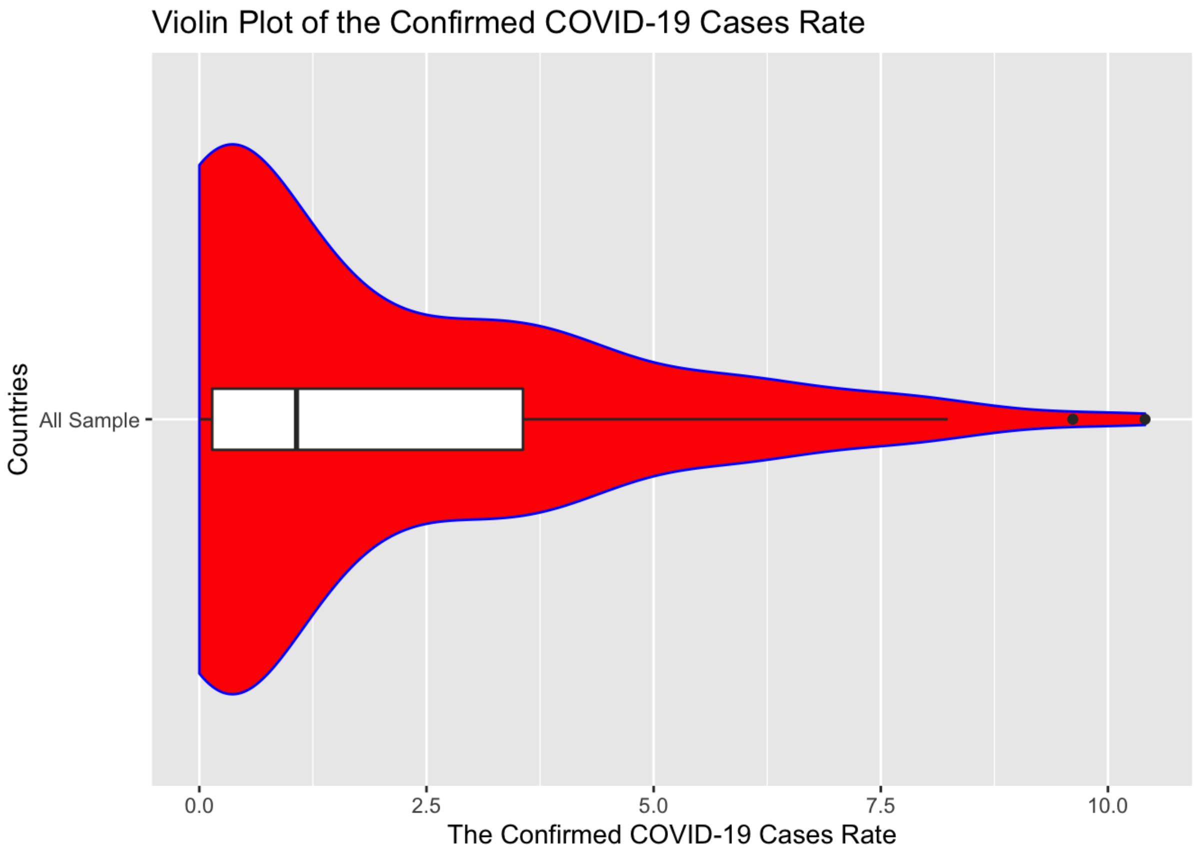
Visualizing the overall data from our data sets

Violin Plots to Demonstrate the Distribution of countries for each variable: Obesity Rate, Confirmed COVID-19 Positive Cases, Deaths as a result of COVID-19

```
ggplot(data = food, aes(x = "All Sample", y = Obesity)) + geom_violin(col="blue", fill="red", na.rm
=TRUE) + geom_boxplot(width=0.1, na.rm=TRUE) + xlab("Countries") + ylab("The Obesity Rate") + ggtit
le("Violin Plot of the Obesity Rate") + coord_flip()
```

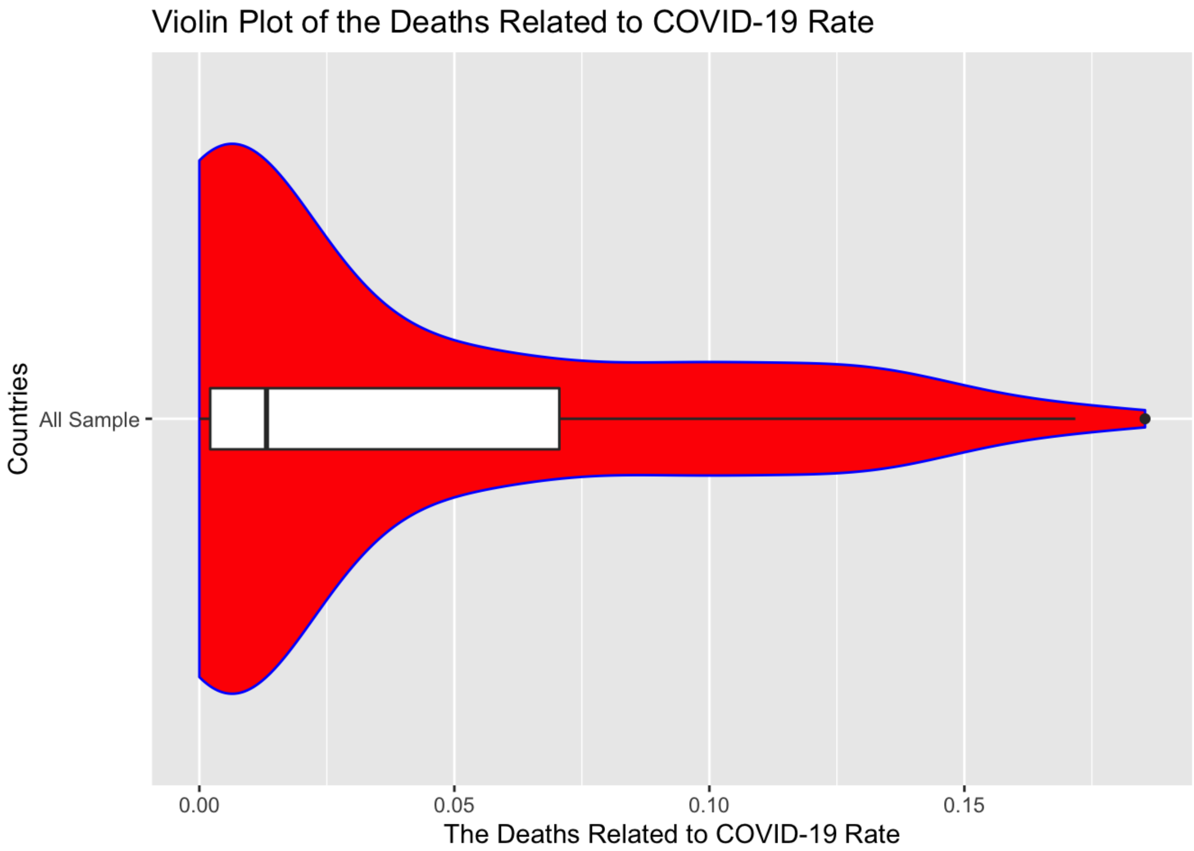


```
ggplot(data = food, aes(x = "All Sample", y = Confirmed)) + geom_violin(col="blue", fill="red", na.
rm=TRUE) + geom_boxplot(width=0.1, na.rm=TRUE) + xlab("Countries") + ylab("The Confirmed COVID-19 C
ases Rate") + ggtitle("Violin Plot of the Confirmed COVID-19 Cases Rate") + coord_flip()
```



Violin Plot of the Confirmed COVID-19 Cases Rate

```
ggplot(data = food, aes(x = "All Sample", y = Deaths)) + geom_violin(col="blue", fill="red", na.rm=
TRUE) + geom_boxplot(width=0.1, na.rm=TRUE) + xlab("Countries") + ylab("The Deaths Related to COVID
-19 Rate") + ggtitle("Violin Plot of the Deaths Related to COVID-19 Rate") + coord_flip()
```



Violin Plot of the Deaths Related to COVID-19 Rate

Describing the Violin Plots

At initial glance, we can see that many countries have low rates of COVID-19 related deaths and confirmed cases. We may attribute this to the large number of smaller countries in our data set that would have smaller populations and in turn, have lower COVID-19 related outcomes (deaths and positive cases).

Wider sections of the violin plot represent a higher probability that members of the population will take on the given value Looking at the violin plot that shows the obesity rate, we see that there is a higher probability that countries will have an obesity rate that is greater than the median of the sample.

```
obesitymean=aggregate(processed$Obesity, by=list(type=processed$region),mean)
obesitymean
```

type	**x**
<chr>	<dbl>
Africa	10.64884
Americas	23.67143
Asia	15.93611
Europe	24.79722
Oceania	30.31667

5 rows

```
undernourishedmean=aggregate(processed$Confirmed, by=list(type=processed$region),mean)
undernourishedmean
```

type	**x**
<chr>	<dbl>
Africa	0.4329477
Americas	2.1122347
Asia	1.6908908
Europe	4.5599475
Oceania	0.0280809

5 rows

```
confirmedmean=aggregate(processed$Confirmed, by=list(type=processed$region),mean)
confirmedmean
```

type	**x**
<chr>	<dbl>
Africa	0.4329477
Americas	2.1122347
Asia	1.6908908
Europe	4.5599475
Oceania	0.0280809

5 rows

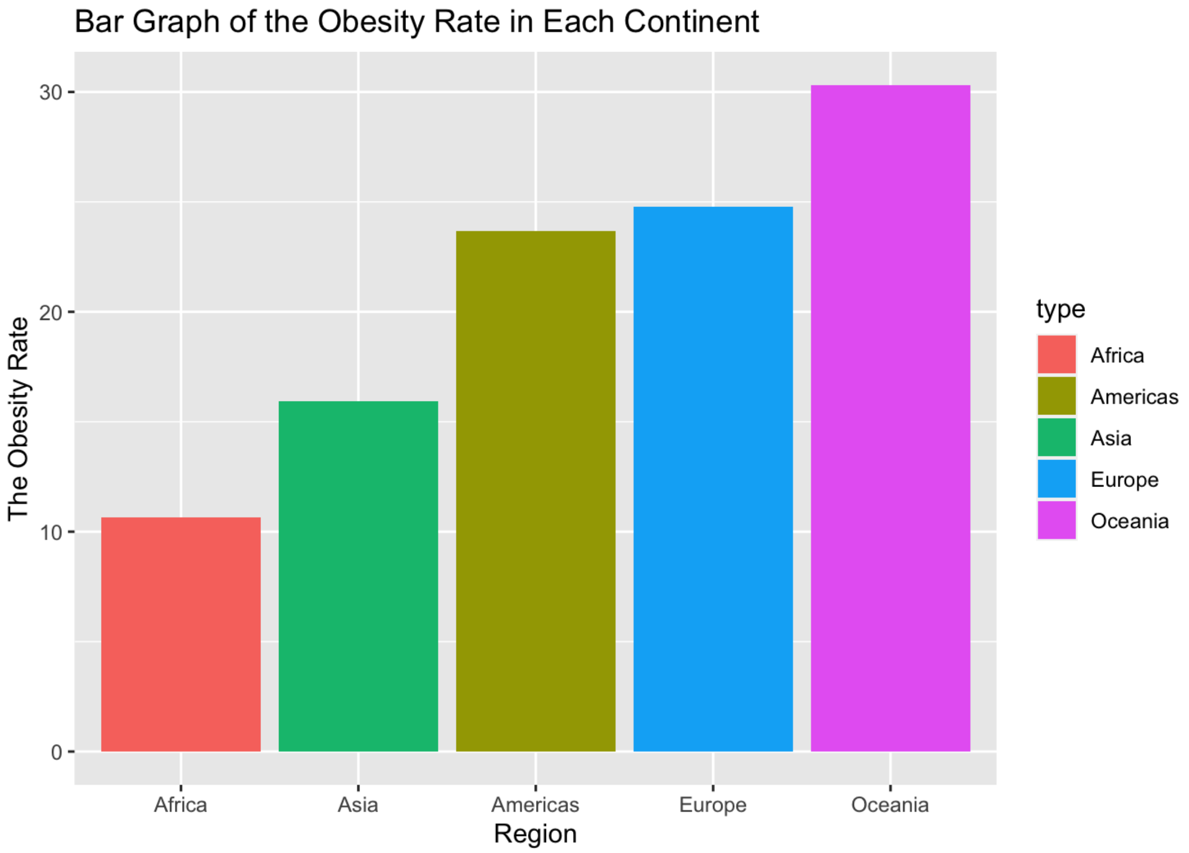
```
deathsmean=aggregate(processed$Deaths, by=list(type=processed$region),mean)
deathsmean
```

type <chr>	x <dbl>
Africa	0.0089824182
Americas	0.0508787541
Asia	0.0201625284
Europe	0.0916222103
Oceania	0.0007090111

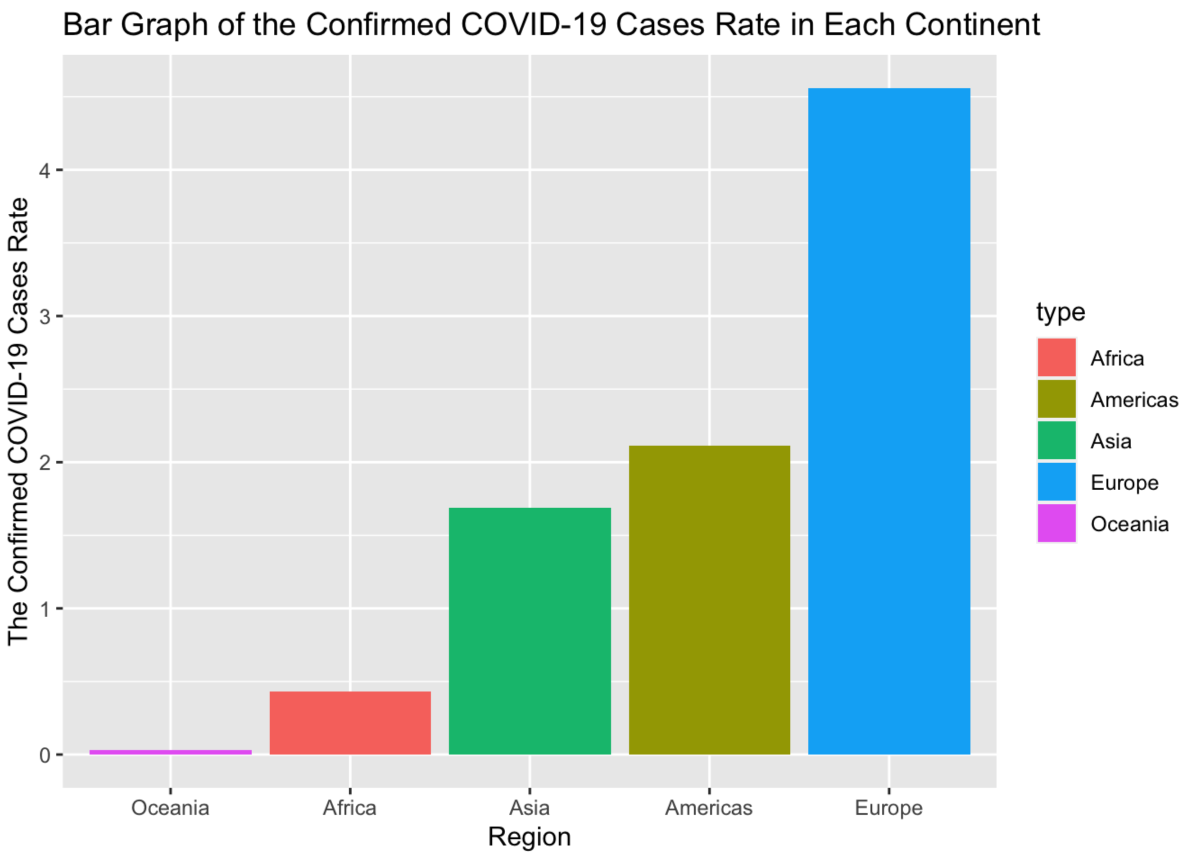
5 rows

Bar Graphs to Demonstrate the Distribution of Each Variable: Obesity Rate, Confirmed COVID-19 Positive Cases, Deaths as a result of COVID-19 Over the Number of Countries

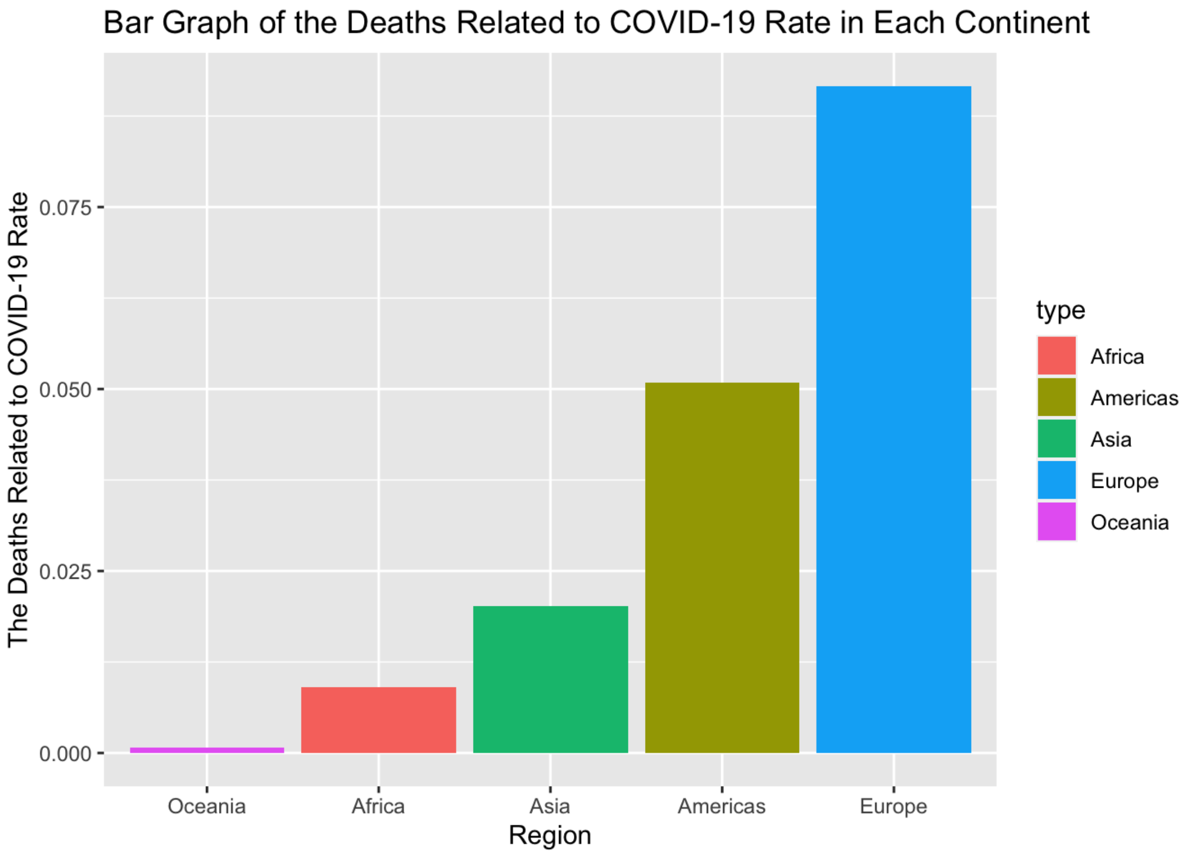
```
ggplot(data=obesitymean, aes(x=reorder(type,x), y=x, fill=type)) + geom_bar(stat="identity") + xlab
("Region") + ylab("The Obesity Rate") + ggtitle("Bar Graph of the Obesity Rate in Each Continent")
```



```
ggplot(data=confirmedmean, aes(x=reorder(type,x), y=x, fill=type)) + geom_bar(stat="identity") + xl
ab("Region") + ylab("The Confirmed COVID-19 Cases Rate") + ggtitle("Bar Graph of the Confirmed COVI
D-19 Cases Rate in Each Continent")
```



```
ggplot(data=deathsmean, aes(x=reorder(type,x), y=x, fill=type)) + geom_bar(stat="identity") + xlab(
"Region") + ylab("The Deaths Related to COVID-19 Rate") + ggtitle("Bar Graph of the Deaths Related
 to COVID-19 Rate in Each Continent")
```



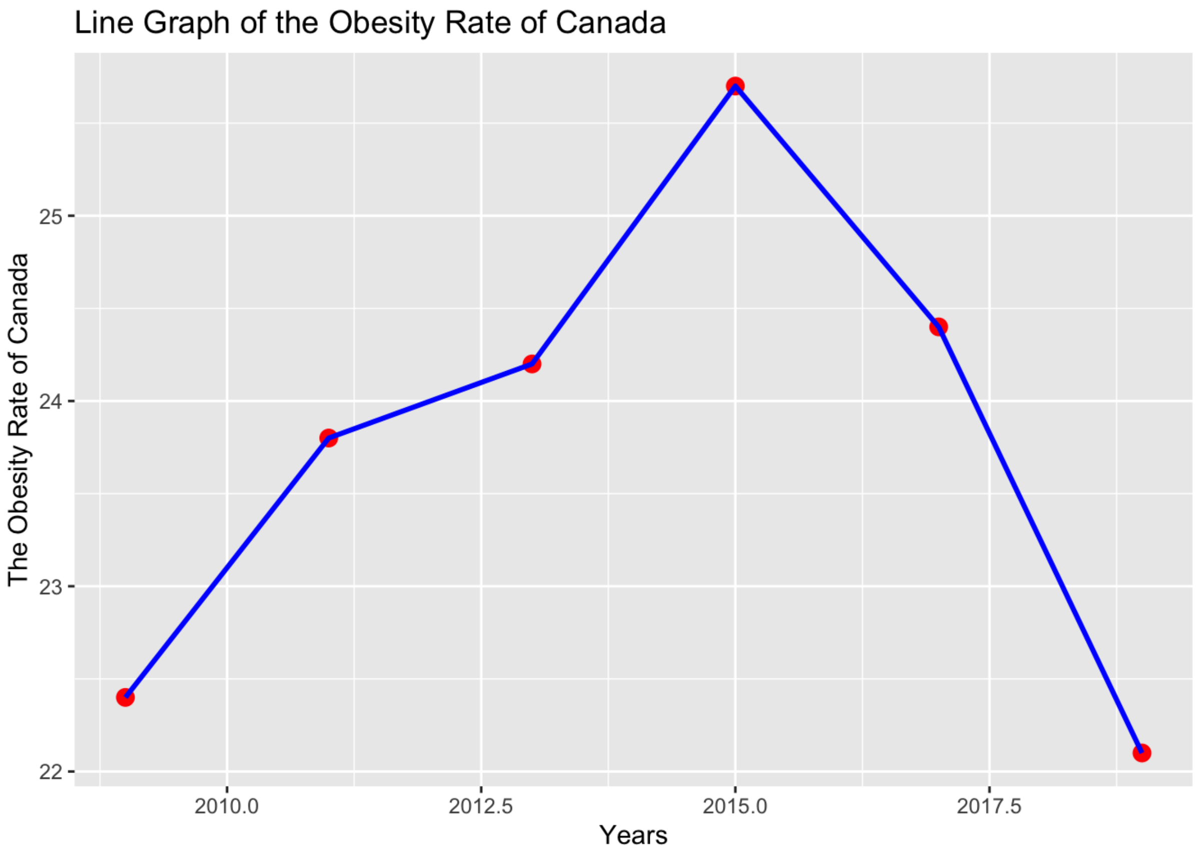
Describing the Bar Graphs

As you can see, the bar graphs showing the amount of COVID-19 related deaths and confirmed positive cases of COVID-19 are very similar, including the order of the countries. A large gap can also be found between the Europe and the next continent.

In the graph showing the obesity rate among continents, Oceania leads with the highest rate of obesity, followed by Europe, Americas, Asia and then Africa. The gaps between these continents are much smaller for this graph than the others.

Simple Line Graph to Demonstrate the obesity rate in Canada between the years of:

```
ggplot(canada.percent, aes(x=REF_DATE, y=VALUE)) + geom_point(color="red", size=3) + geom_line(color="blue", size=1) + xlab("Years") + ylab("The Obesity Rate of Canada") + ggtitle("Line Graph of the Obesity Rate of Canada")
```



Describing the Line Graph

The line graph visualizes the trend of obesity in Canada between the years: 2009 and 2019.

Question 1: Analyze the trends of obesity in Canada and it's relationship with Canada's Gross Domestic Product (GDP) value.

Part A: Compare obesity rate for years 2009 and 2013

Reasoning for the years chosen: Canada's GDP per capita is the lowest after the 2008 recession and Canada's GDP in 2013 GDP is the second highest after 2008 recession. We also chose the year 2013 becuase our data set has data for that year.

A a random selection of $n_{2009} = 6479500$ Canadians aged 5 to 79 years old, of which 22.4 of the sample size are obese. A similar poll by the Government of Canada in 2013 had a sample size of $n_{2013} = 7424000$ of Canadians aged 5 to 79 years old found that 24.2 of the sample size are obese.

We want to know if this data indicates that the obesity rate in Canada in 2009 is higher than the obesity rate in 2013. Carry out the statistical test.

The statistical hypotheses to be tested are:

$$\begin{aligned} \mathrm{H}_0 &: p_{Obe\_2013} \leq p_{Obe\_2009} \quad \text{(The obesity rate in 2013 is not higher than the obesity rate in 2009)} \\ \mathrm{H}_A &: p_{Obe\_2013} > p_{Obe\_2009} \quad \text{(The obesity rate in 2013 is higher than the obesity rate in 2009)} \end{aligned}$$

We set up the $\alpha = 0.05$

Next we compute the value of the test statistic.

```
p_hat2009 = 0.224
p_hat2013 = 0.242
n2009 = 6479500 #sample size selected in 2009
x2009 = n2009 * p_hat2009 #the number of random selected Canadians aged 5 to 79 years old who is ob
ese in 2009
n2013 = 7424000 #sample size selected in 2013
x2013 = n2013 * p_hat2013 #the number of random selected Canadians aged 5 to 79 years old who is ob
ese in 2013
x2009
```

```
## [1] 1451408
```

```
x2013
```

```
## [1] 1796608
```

In this computation, we assume a “common proportion” under H_0. The estimate of this common proportion, the “pooled sample proportion” $\widehat{p}$ is:

```
pooled.p1 <- (x2009 + x2013)/(n2009 + n2013)
pooled.p1
```

```
## [1] 0.2336114
```

$$\widehat{p} = \frac{X_{Obe\_2009} + X_{Obe\_2013}}{n_{Obe\_2009} + n_{Obe\_2013}} = \frac{1451408 + 1796608}{6479500 + 7424000} = 0.2336114$$

Compute the test statistic Z_{Obs}

$$Z_{obs} = \frac{\widehat{p}_{Obe\_2013} - \widehat{p}_{Obe\_2009} - (p_{Obe\_2013} - p_{Obe\_2009})}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_{Obe\_2013}} + \frac{1}{n_{Obe\_2009}}\right)}} = \frac{0.242 - 0.224 - (0)}{\sqrt{0.2336114(1 - 0.2336114)\left(\frac{1}{7424000} + \frac{1}{6479500}\right)}} = 79.12772$$

```
z.obsnumerator1 <-  p_hat2013 - p_hat2009
z.obsdenominator1 <- sqrt(pooled.p1 * (1 - pooled.p1)*((1/n2009) + (1/n2013)))
z.obs1 <- z.obsnumerator1/z.obsdenominator1

z.obs1
```

```
## [1] 79.12772
```

```
1-pnorm(z.obs1)
```

```
## [1] 0
```

The P-value is then $P(Z > 79.12772)$ which is computed to be 0

Because our $p - value < \alpha$, Reject H_0. We conclude that The obesity rate in 2013 is higher than the obesity rate in 2009.

The computation of the test statistic above can be completed with the `prop.test` command.

```
prop.test(c(x2013,x2009), c(n2013,n2009), alternative="greater", correct=FALSE)
```

```
## 
##  2-sample test for equality of proportions without continuity correction
## 
## data:  c out of cx2013 out of n2013x2009 out of n2009
## X-squared = 6261.2, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.0176266 1.0000000
## sample estimates:
## prop 1 prop 2 
##  0.242  0.224
```

Part B: Compare obesity rate for years 2013 and 2019

Comparing the biggest gap between the lowest point with the highest point before the COVID-19 pandemic.

We want to know if these data indicate that the obesity rate in 2013 is higher than the obesity rate in 2019.

A a random selection of $n_{2019} = 7169800$ Canadians aged 5 to 79 years old, of which 22.1 the sample size are obese. A similar poll by the Government of Canada in 2013 had a sample size of $n_{2013} = 7424000$ Canadians aged 5 to 79 years old found that 24.2 of the sample size are obese in 2013.

The statistical hypothesese to be tested are:

$$\begin{aligned} \mathrm{H}_0 &: p_{Obe\_2013} \leq p_{Obe\_2009} & \text{(The obesity rate in 2019 is not higher than the obesity rate in 2013)} \\ \mathrm{H}_A &: p_{Obe\_2013} > p_{Obe\_2009} & \text{(The obesity rate in 2019 is higher than the obesity rate in 2013)} \end{aligned}$$

We set up the $\alpha = 0.05$ Next we compute the value of the test statistic.

```
p_hat2019 = 0.221
p_hat2013 = 0.242
n2019 = 7169800 #sample size selected in 2009
x2019 = n2019 * p_hat2019 #the number of random selected Canadians aged 5 to 79 years old who is obese in 2009
n2013 = 7424000 #sample size selected in 2013
x2013 = n2013 * p_hat2013 #the number of random selected Canadians aged 5 to 79 years old who is obese in 2013
x2019
```

```
## [1] 1584526
```

```
x2013
```

```
## [1] 1796608
```

In this computation, we assume a “common proportion” under H_0. The estimate of this common proportion, the “pooled sample proportion” $\hat{p}$ is:

```
pooled.p2 <- (x2019 + x2013)/(n2019 + n2013)
pooled.p2
```

```
## [1] 0.2316829
```

$$\widehat{p} = \frac{X_{Obe\_2019} + X_{Obe\_2013}}{n_{Obe\_2019} + n_{Obe\_2013}} = \frac{1451408 + 1796608}{6479500 + 7424000} = 0.2336114$$

Compute the test statistic Z_{Obs}

$$Z_{obs} = \frac{\widehat{p}_{Obe\_2013} - \widehat{p}_{Obe\_2019} - (p_{Obe\_2013} - p_{Obe\_2019})}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_{Obe\_2013}} + \frac{1}{n_{Obe\_2019}}\right)}} = \frac{0.242 - 0.224 - (0)}{\sqrt{0.2336114(1-0.2336114)\left(\frac{1}{7424000} + \frac{1}{6479500}\right)}} = 79.12772$$

```
z.obsnumerator2 <-  p_hat2013 - p_hat2019
z.obsdenominator2 <- sqrt(pooled.p2 * (1 - pooled.p2)*((1/n2019) + (1/n2013)))
z.obs2 <- z.obsnumerator2/z.obsdenominator2

z.obs2
```

```
## [1] 95.05837
```

```
1-pnorm(z.obs2)
```

```
## [1] 0
```

The P-value is $P(Z > 79.12772)$ which is computed to be 0.

Because $p - value < \alpha$, Reject H_0. We conclude that The obesity rate in 2013 is higher than the obesity rate in 2019.

The computation of the test statistic above can be completed with the `prop.test` command.

```
prop.test(c(x2013,x2019), c(n2013,n2019), alternative="greater", correct=FALSE)
```

```
## 
##  2-sample test for equality of proportions without continuity correction
## 
## data:  c out of cx2013 out of n2013x2019 out of n2019
## X-squared = 9036.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.02063694 1.00000000
## sample estimates:
## prop 1 prop 2 
##  0.242  0.221
```

Part C: Is there ample statistical evidence to confirm that, in the year 2019, the proportion of Canadians aged 5 to 79 years old who were obese, was less than 31.3%?

From the “Overweight and obesity based on measured body mass index, by age group and sex” (Open Government, 2021), we know that out of the random sample of $n_{2019} = 7169800$ Canadians aged 5 to 79 years old, 1584526 were obese in 2019. Compared to the “COVID-19 Healthy Diet Dataset” (Kaggle, 2021) obesity values taken from the year 2021, 31.3% of Canadians were obese.

$$\begin{array}{ll} \mathrm{H}_0 : p \geq \ 0.313 & \text{(the proportion of Canadians aged 5 to 79 years old who are Obese, is not less than 31.3\%,in 2019)} \\ \mathrm{H}_A : p < \ 0.313 & \text{(the proportion of Canadians aged 5 to 79 years old who are Obese, is less than 31.3\%, in 2019)} \end{array}$$

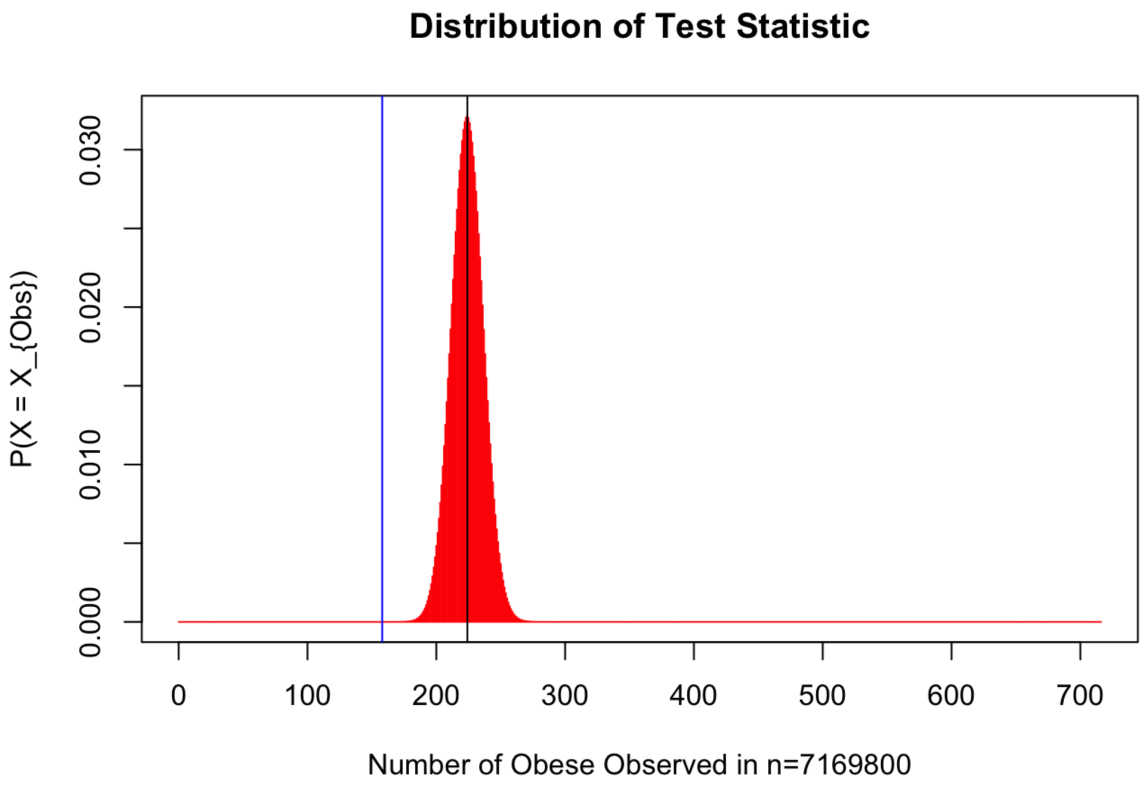
Set up $\alpha = 0.05$.

The data we are looking at has the sample size of $n_{2019} = 7169800$ From this same data, the observed number who are obese is $X_{Obs} = 1584526$, which is the test statistic. The distribution of X, which can be modeled by the Binomial probability model is provided below and is conditional upon the null hypothesis being true or $p = 0.313$.

If the null hypothesis is true, then the distribution of X will look like this:

The Distribution of the Statistic X_{Obs}:

```
plot(0:716, dbinom(0:716, 716, 0.313), xlab="Number of Obese Observed in n=7169800", ylab="P(X = X_
{Obs})", type="h", col='red', main="Distribution of Test Statistic")
abline(v = 158, col="blue")
abline(v = 716 * 0.313 , col="black")
```



And the resulting mean/expected value is: $E(X) = \mu_X = 2244147$ with a standard deviation of $SD(X) = \sigma_X = 1241.664$.

```
ex = 7169800 * 0.313
sd = sqrt(7169800 *0.313*(1-0.313) )

ex
```

```
## [1] 2244147
```

```
sd
```

```
## [1] 1241.664
```

Compute the P-value.

$$\begin{aligned}P - \text{value} =& P(X < \overbrace{1584526}^{X_{Obs}} | p = 0.313) \\ =& \sum_{x=0}^{1584525} \binom{7169800}{x} (0.313)^x (0.687)^{7169800-x} \\ =& 0 \ \text{ (see R code below)}\end{aligned}$$

```
pbinom(1584525, 7169800, 0.313)
```

```
## [1] 0
```

Our P-value is 0, which means there is no trust in the null hypothesis (and P-value of $0 < 0.05$). As a result, we decide to reject the null hypothesis and conclude from this data that the proportion of Canadians aged 5 to 79 years old who are obese is less than 31.3%, in 2019.

Question 2:

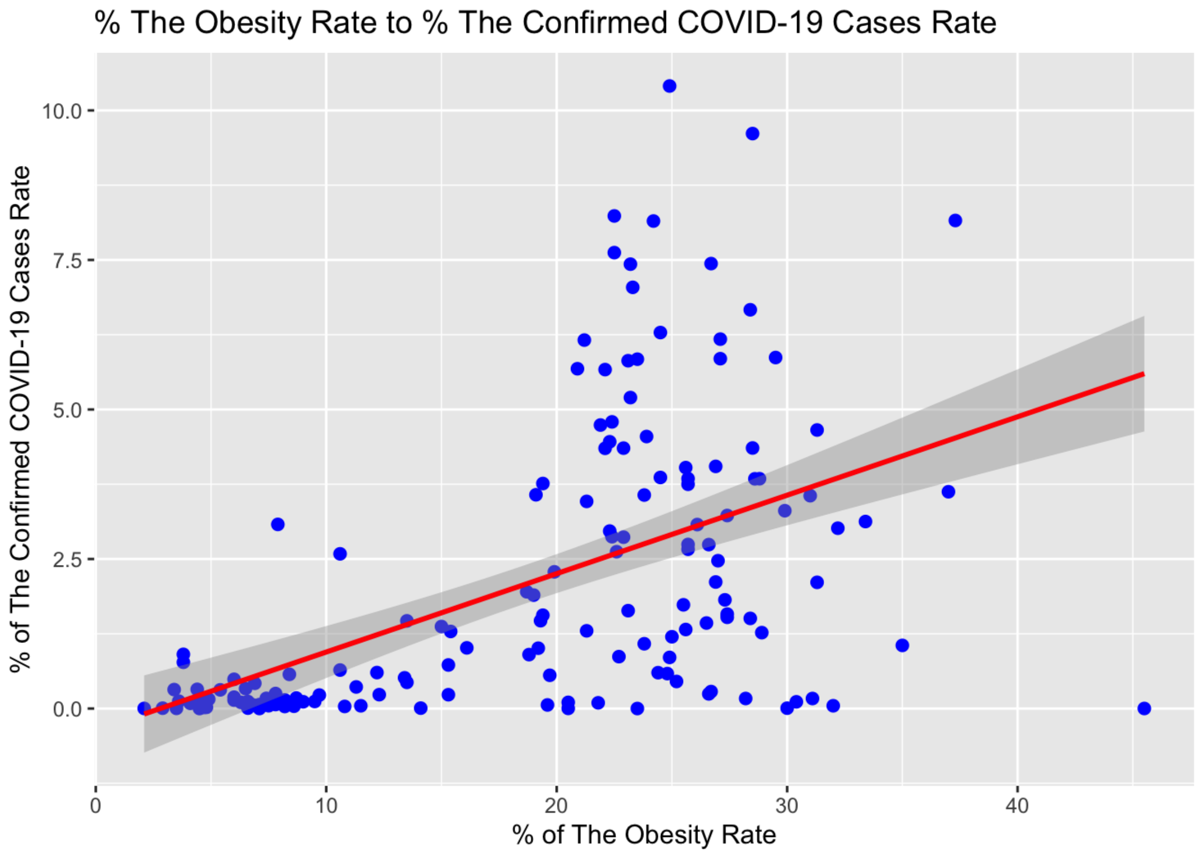
Part A: Analyze and model the relationship between obesity and each COVID-19 outcome, deaths as a result of COVID-19 and confirmed positive cases of COVID-19.

Visualization of The Obesity Rate and The Confirmed Rate

Create scatterplot of **The Obesity Rate** to **The Confirmed** Rate:

```
ggplot(food, aes(x = Obesity, y = Confirmed)) + geom_point(col="blue", size = 2) + xlab("% of The O
besity Rate") + ylab("% of The Confirmed COVID-19 Cases Rate") + ggtitle("% The Obesity Rate to % T
he Confirmed COVID-19 Cases Rate") + geom_smooth(method="lm", col="red")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Quantifying the Relationship between The Obesity Rate and The Confirmed Rate by Using the Correlation Coefficient

```
cor(~Obesity, ~Confirmed, data=food)
```

```
## [1] 0.5248714
```

The correlation coefficient between the obesity Rate and the confirmed positive COVID-19 cases rate is 0.5248714. This shows that there is a strong positive linear relationship between the rate of obesity and confirmed positive COVID-19 cases.

```
favstats(~ Obesity, data=food)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
2.1	8.35	21.55	25.7	45.5	18.59808	9.549116	156	0

1 row

```
favstats(~ Confirmed, data=food)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
0.0003115265	0.1419854	1.069339	3.561377	10.4082	2.0719	2.385296	156	0

1 row

From this output,

$$\overline{X} = 18.59808 \quad S_X = 9.549116 \quad \overline{Y} = 2.0719 \quad S_Y = 2.385296 \quad r = 0.5248714$$

## Modeling the Relationship Between The Obesity Rate and The Confirmed Rate

The *statistical model* is:

$$\underbrace{ConfirmedRate_i}_{Y-variable} = A + (B * \underbrace{ObesityRate_i}_{X-variable}) + e_i$$

We use the values of r, S_X, and S_Y computed.

$$b = r\left(\frac{S_Y}{S_X}\right) = (0.5248714)\left(\frac{2.385296}{9.549116}\right) \approx 0.1311088$$

We now compute the value of a:

$$a = \overline{Y} - (b * \overline{X}) = 2.0719 - (0.1311088 * 18.59808) \approx -0.3664725$$

and the estimate of the model is then

$$\widehat{ConfirmedRate}_i = -0.3664725 + 0.1311088 * ObesityRate_i$$

Compute the estimate with R using the `lm` function.

```
predictdemovote = lm(Confirmed ~ Obesity, data=food)
```

```
options(scipen=999)
predictdemovote$coef
```

```
## (Intercept)     Obesity
##  -0.3664725   0.1311088
```

The value of a and b ar $a = -0.3664725$ and $b = 0.1311088$. From the training data, we have an estimate of the model.

$$\widehat{ConfirmedRate}_i = -0.3664725 + (0.1311088 * ObesityRate_i)$$

Notes about the values A & B:

A: when the obesity rate is 0, the confirmed rate is -0.3664725%. This is meaningless because most countries do not have an obesity rate of 0.

B: As the obesity rate increases by 1%, then the confirmed rate will increase by an *average* of 0.1311088%.

The F-Test of Linear Appropriateness

Checking if the linear model meets the condition of linear appropriateness and if $ConfirmedRate_i = A + (B * ObesityRate_i) + e_i$ is a valid model by checking whether the slope term β_1 is non-zero or not.

Test the statistical hypotheses:

$$\begin{array}{c} \mathrm{H}_0 : B = 0 \ \ (ConfirmedRate \text{ cannot be expressed as a linear function of } ObesityRate) \\ \mathrm{H}_A : B \neq 0 \ \ (ConfirmedRate \text{ can be expressed as a linear function of } ObesityRate) \end{array}$$

Set up the $\alpha = 0.05$. F-Test of Linear Appropriateness

```
summary(aov(predictdemovote))
```

```
##              Df Sum Sq Mean Sq F value           Pr(>F)    
## Obesity       1  243.0  242.95   58.56 0.00000000000201 ***
## Residuals   154  638.9    4.15                             
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value = 0.00000000000201. $p - value < \alpha$, Due to the p-value being below the alpha value, we can reject the null hypothesis. We can conclude that the confirmed positive cases of COVID-19 rate can be expressed as a linear function of the Obesity Rate.

Normality of the Residuals of the linear model

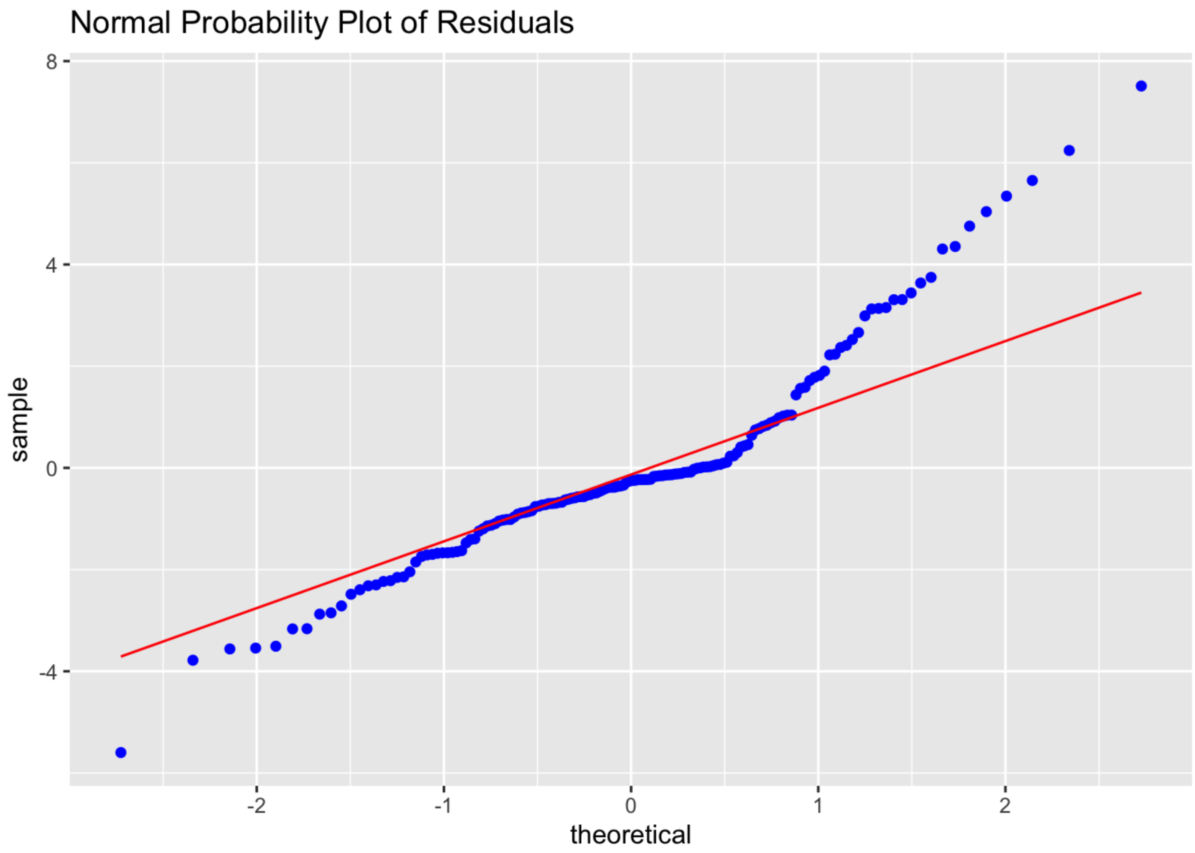
Checking if the model meets the condition of normality of the residuals.

```
predicted.values.demovote = predictdemovote$fitted.values
eisdemovote = predictdemovote$residuals
diagnosticdf = data.frame(predicted.values.demovote, eisdemovote)
head(diagnosticdf)
```

	predicted.values.demovote	eisdemovote
	<dbl>	<dbl>
1	0.2235173	-0.08138306
2	2.5572547	0.41004626
3	3.1210227	-2.87612560
4	0.5250676	-0.46338012
6	3.3701295	0.98601790
7	2.3737023	3.30752235

6 rows

```
ggplot(diagnosticdf, aes(sample = eisdemovote)) +  stat_qq(col='blue') + stat_qqline(col='red') + g
gtitle("Normal Probability Plot of Residuals")
```

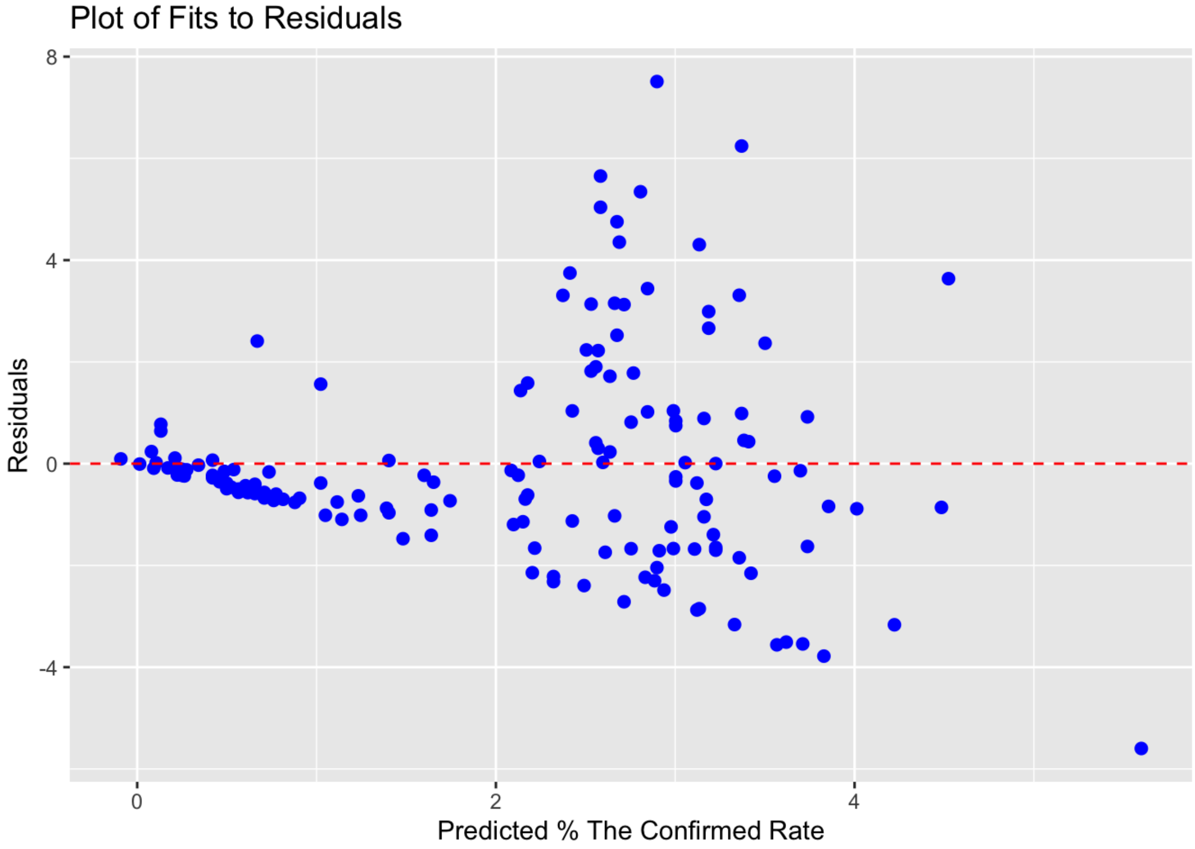


The plot shows that the residuals are normally distributed and meets the condition.

Homoscedasticity of the Residuals of the linear model

Checking if the linear model meets the condition of homoscedasticity.

```
ggplot(diagnosticdf, aes(x = predicted.values.demovote, y = eisdemovote)) +  geom_point(size=2, col='blue', position="jitter") + xlab("Predicted % The Confirmed Rate") + ylab("Residuals") + ggtitle("Plot of Fits to Residuals") + geom_hline(yintercept=0, color="red", linetype="dashed")
```



Based on our plot above, we can say that the homoscedasticity condition is satisfied.

Analysis of Error and Coefficient of Determination of The Linear Model

```
aov(predictdemovote)
```

```
## Call:
##    aov(formula = predictdemovote)
##
## Terms:
##                   Obesity Residuals
## Sum of Squares   242.9528  638.9409
## Deg. of Freedom         1       154
##
## Residual standard error: 2.036901
## Estimated effects may be unbalanced
```

From this output,

$$SSE = 638.9409 \quad SSR = 242.9528 \quad df = 154 \quad SST = SSE + SSR = 881.8937 \quad S_e = 2.036901$$

Coefficient of Determination

```
rsquared(predictdemovote)
```

```
## [1] 0.2754899
```

The coefficient of determination is 0.2754899. This coefficient of determination is too small, which means that our model is not good.

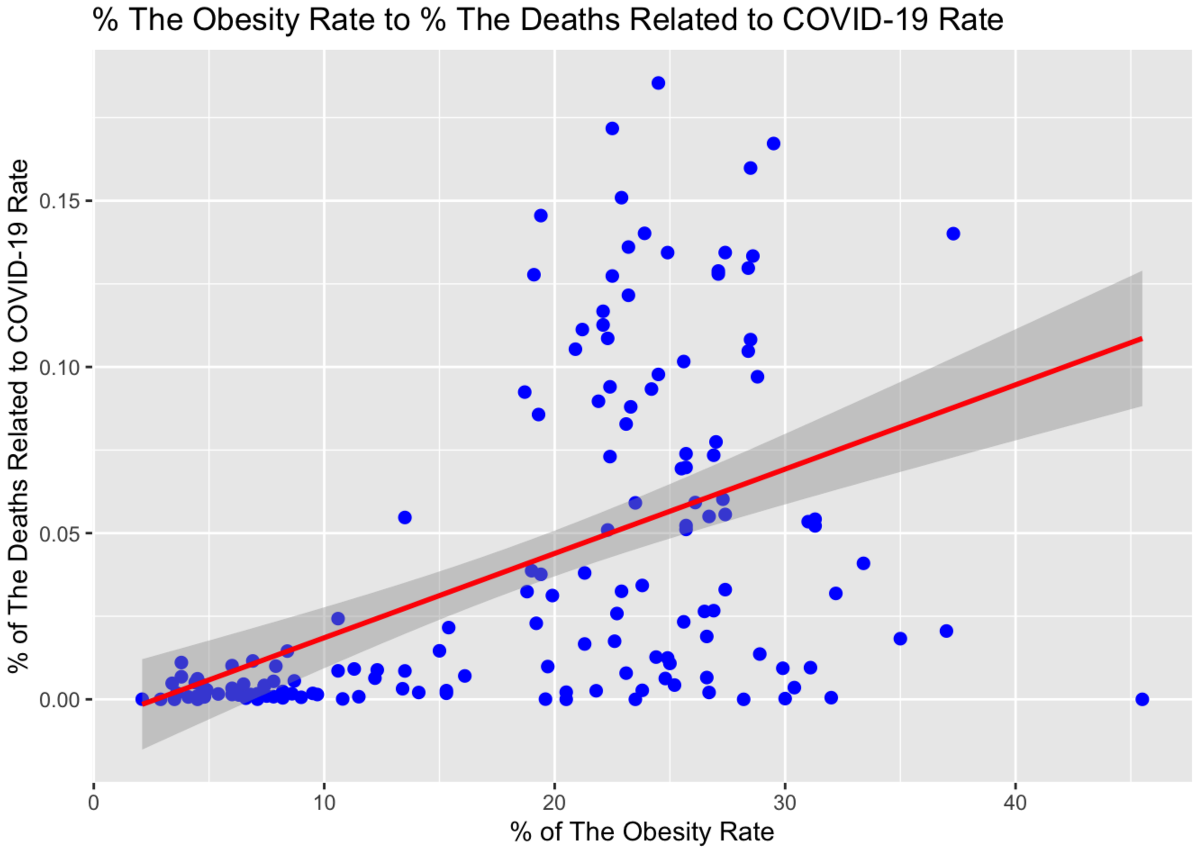
Part B:

Visualizing of The Obesity Rate and The Deaths

Scatterplot of **The Obesity Rate** to **The COVID-19 Related Deaths Rate** Rate:

```
ggplot(food, aes(x = Obesity, y = Deaths)) + geom_point(col="blue", size = 2) + xlab("% of The Obesity Rate") + ylab("% of The Deaths Related to COVID-19 Rate") + ggtitle("% The Obesity Rate to % The Deaths Related to COVID-19 Rate") + geom_smooth(method="lm", col="red")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The F-Test of Linear Appropriateness

Checking if the linear model meets the condition of linear appropriateness and if $DeathsRate_i = A + (B * ObesityRate_i) + e_i$ is a valid model by checking whether the slope term β_1 is non-zero or not.

Test the statistical hypotheses:

$$\text{H}_0 : B = 0 \;\; (DeathsRate \text{ cannot be expressed as a linear function of } ObesityRate)$$
$$\text{H}_A : B \neq 0 \;\; (DeathsRate \text{ can be expressed as a linear function of } ObesityRate)$$

Set up the $\alpha = 0.05$.

```
predictdemovote = lm(Deaths ~ Obesity, data=food)
summary(aov(predictdemovote))
```

```
##              Df  Sum Sq Mean Sq F value          Pr(>F)    
## Obesity       1 0.09096 0.09096    49.3 0.0000000000659 ***
## Residuals   154 0.28412 0.00184                            
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value = 0.00000000000201 $p - value < \alpha$. Since our p-value is less than our alpha value, we can reject the null hypothesis. We can conclude that the COVID-19 related deaths rate can be expressed as a linear function of the obesity rate.

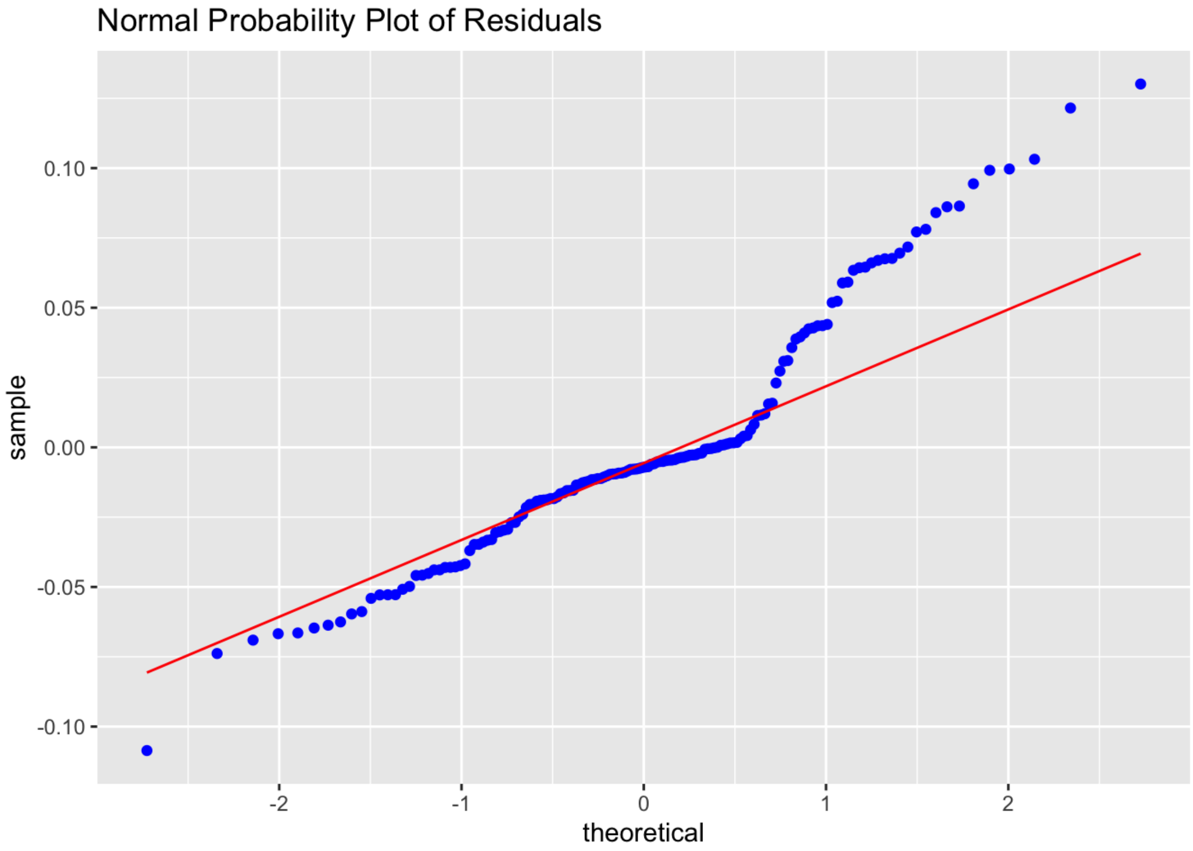
Normality of the Residuals of the linear model**

```
predicted.values.demovote = predictdemovote$fitted.values
eisdemovote = predictdemovote$residuals
diagnosticdf = data.frame(predicted.values.demovote, eisdemovote)
head(diagnosticdf)
```

	predicted.values.demovote <dbl>	eisdemovote <dbl>
1	0.004575211	0.001610568
2	0.049731430	0.001219944
3	0.060639955	-0.054081802
4	0.010410003	-0.008949454
6	0.065460001	0.042766634
7	0.046179817	0.059165243

6 rows

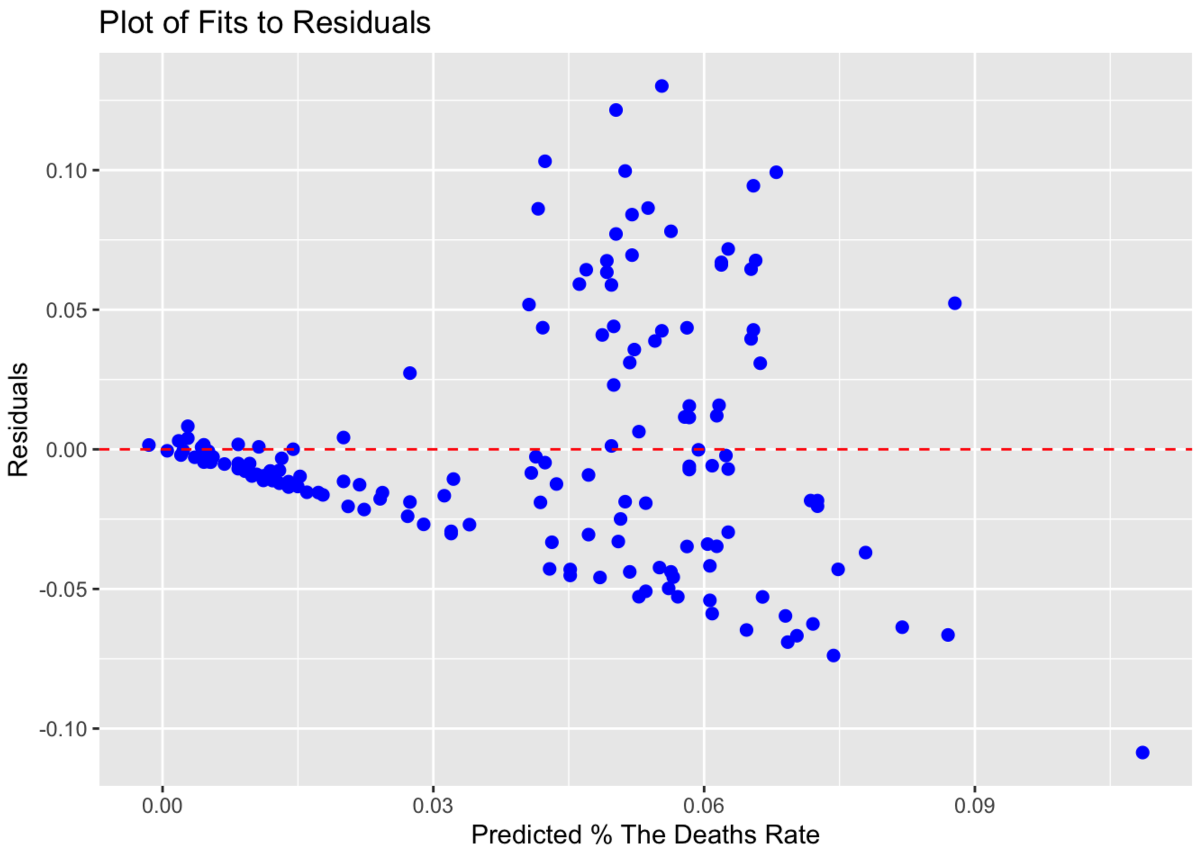
```
ggplot(diagnosticdf, aes(sample = eisdemovote)) +  stat_qq(col='blue') + stat_qqline(col='red') + g
gtitle("Normal Probability Plot of Residuals")
```



The plot shows that the residuals are normally distributed.

Homoscedasticity of the Residuals of the linear model

```
ggplot(diagnosticdf, aes(x = predicted.values.demovote, y = eisdemovote)) +  geom_point(size=2, col
='blue', position="jitter") + xlab("Predicted % The Deaths Rate") + ylab("Residuals") + ggtitle("Pl
ot of Fits to Residuals") + geom_hline(yintercept=0, color="red", linetype="dashed")
```



Based on our graph above we can conclude that the homoscedasticity condition is satisfied.

Modeling the Relationship Between The Obesity Rate and the COVID-19 Related Deaths Rate by Bootstrapping

```
Nbootstraps = 1000 #resample n =  200, 1000 times
cor.boot = numeric(Nbootstraps) #define a vector to be filled by the cor boot stat
a.boot = numeric(Nbootstraps) #define a vector to be filled by the a boot stat
b.boot = numeric(Nbootstraps) #define a vector to be filled by the b boot stat
```

```
nsize = dim(food)[1]  #set the n to be equal to the number of bivariate cases, number of rows

#start of the for loop
for(i in 1:Nbootstraps)
{   #start of the loop
    index = sample(nsize, replace=TRUE)  #randomly picks a number between 1 and n, assigns as index
    demovote.boot = food[index, ] #accesses the i-th row of the regressionex1.df data frame

    cor.boot[i] = cor(~Obesity, ~Deaths, data=demovote.boot) #computes correlation for each bootstr
ap sample
    votedemocrat.lm = lm(Deaths ~ Obesity, data=demovote.boot)  #set up the linear model
    a.boot[i] = coef(votedemocrat.lm)[1] #access the computed value of a, in position 1
    b.boot[i] = coef(votedemocrat.lm)[2] #access the computed valeu of b, in position 2

}
#end the loop
#create a data frame that holds the results of teach of he Nbootstraps
bootstrapresultsdf = data.frame(cor.boot, a.boot, b.boot)
```

```
head(bootstrapresultsdf, 3)
```

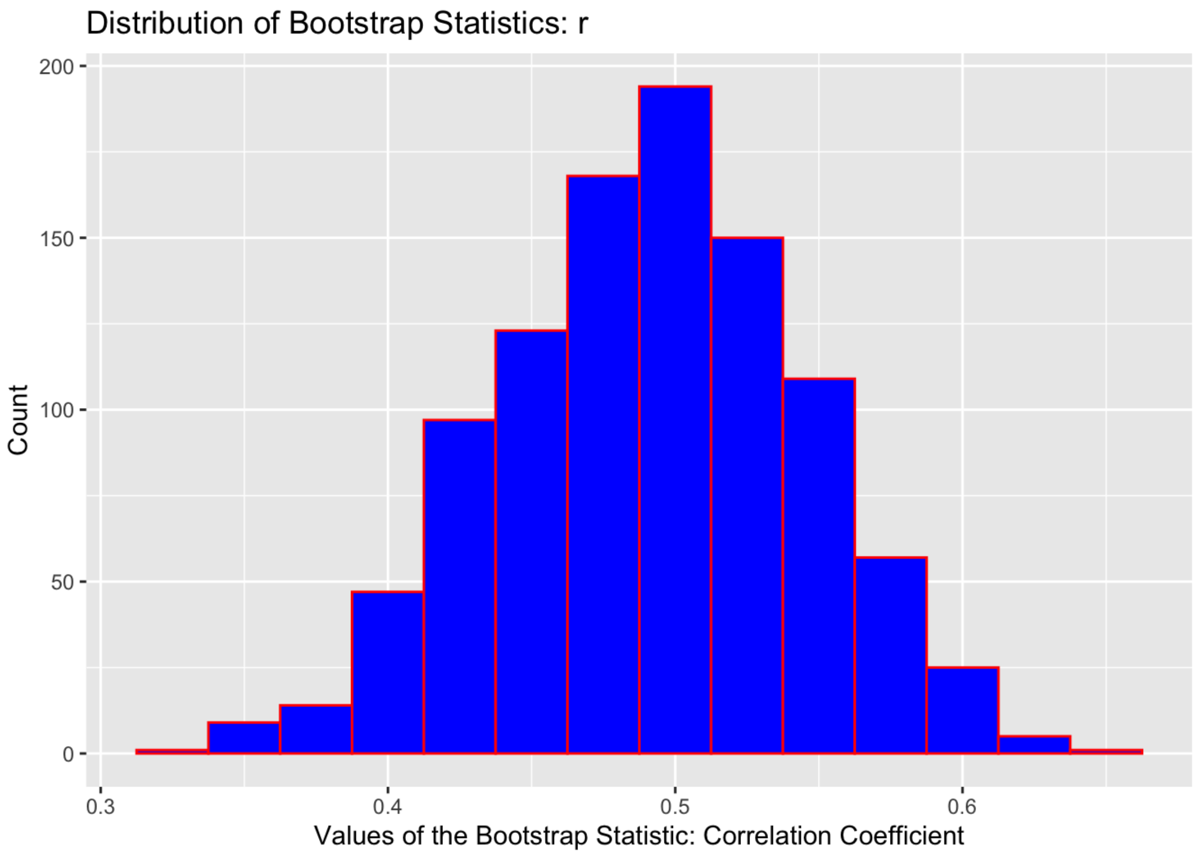
	cor.boot <dbl>	a.boot <dbl>	b.boot <dbl>
1	0.4520348	-0.00747102084	0.002731791
2	0.4344365	-0.00004523288	0.002139069
3	0.4849796	-0.01077467774	0.002542653

3 rows

Bootstrap Distribution of r_{boot}

Coefficient of Determination

```
ggplot(bootstrapresultsdf, aes(x = cor.boot)) + geom_histogram(col="red", fill="blue", binwidth=0.025) + xlab("Values of the Bootstrap Statistic: Correlation Coefficient") + ylab("Count") + ggtitle("Distribution of Bootstrap Statistics: r")
```



```
favstats(~cor.boot, data=bootstrapresultsdf)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
0.3301075	0.4534791	0.4921449	0.5298323	0.6398375	0.4913533	0.05304403	1000	0

1 row

```
qdata(~cor.boot, c(0.025, 0.975), data=bootstrapresultsdf)
```

```
##      2.5%     97.5%
## 0.3891116 0.5928044
```

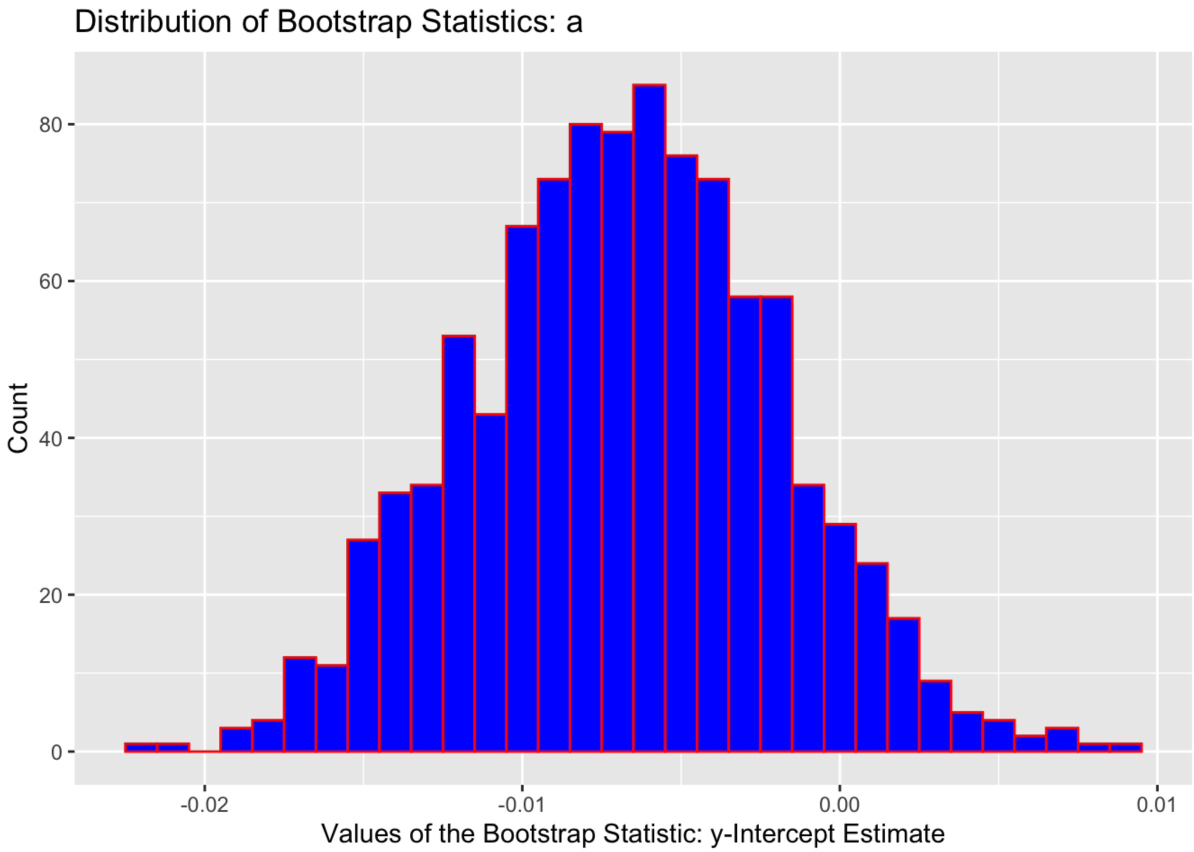
```
rsquared = (favstats(~cor.boot, data=bootstrapresultsdf)$mean)^2
rsquared
```

```
## [1] 0.2414281
```

The mean of all r_{boot} is 0.4933598. with a 95% confidence interval $0.3998882 r_{boot} $ The coefficient of determination is 0.2436653.

Bootstrap Distribution of a_{boot}

```
ggplot(bootstrapresultsdf, aes(x = a.boot)) + geom_histogram(col="red", fill="blue", binwidth=0.001
) + xlab("Values of the Bootstrap Statistic: y-Intercept Estimate") + ylab("Count") + ggtitle("Dist
ribution of Bootstrap Statistics: a")
```



```
qdata(~a.boot, c(0.025, 0.975), data=bootstrapresultsdf)
```

```
##         2.5%        97.5%
## -0.015914617  0.002479049
```

```
favstats(~a.boot, data=bootstrapresultsdf)
```

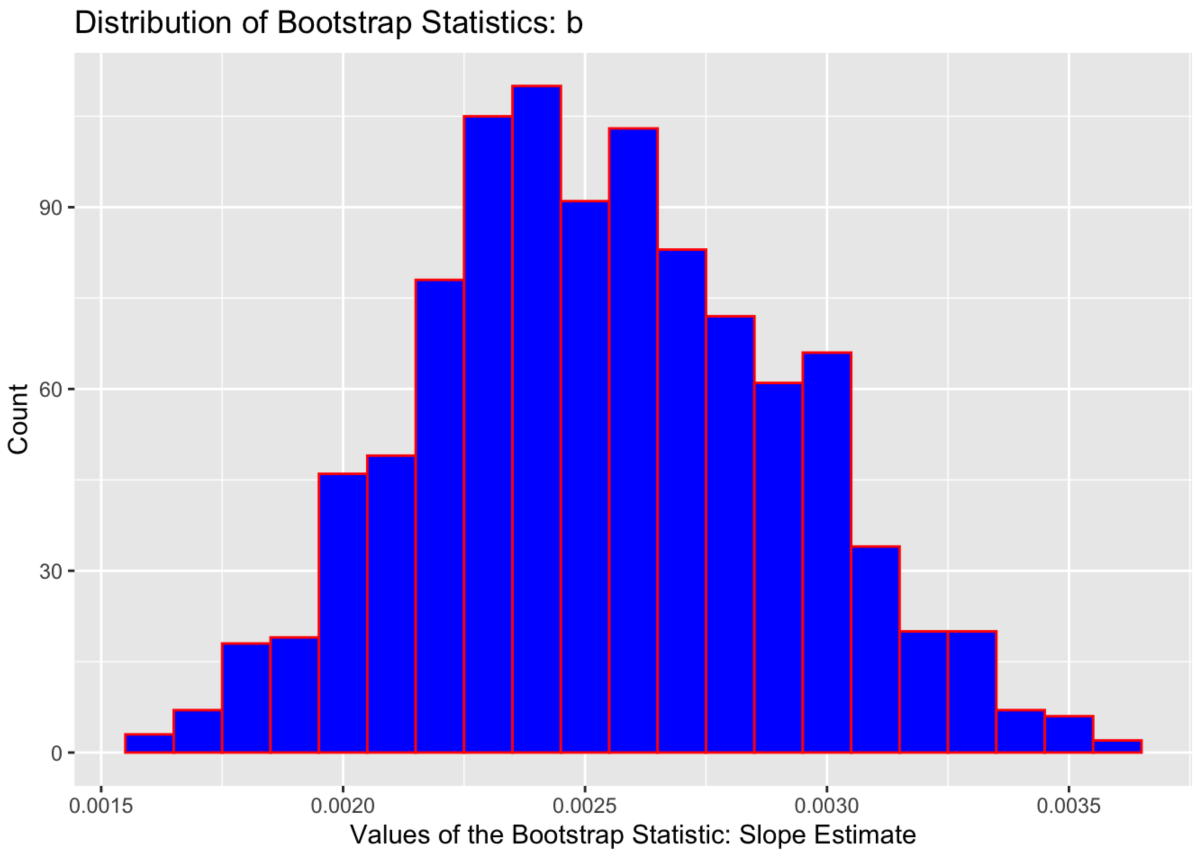
min	Q1	median	Q3	max	mean	sd	
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<ir
-0.02186768	-0.01010724	-0.006789153	-0.003601387	0.009213171	-0.006827313	0.004850294	10

1 row | 1-9 of 10 columns

The mean of all a_{boot} is -0.006948612 with a 95% confidence interval $-0.01606346 a_{boot} $

Bootstrap Distribution of b_{boot}

```
ggplot(bootstrapresultsdf, aes(x = b.boot)) + geom_histogram(col="red", fill="blue", binwidth=0.000
1) + xlab("Values of the Bootstrap Statistic: Slope Estimate") + ylab("Count") + ggtitle("Distribut
ion of Bootstrap Statistics: b")
```



```
favstats(~b.boot, data=bootstrapresultsdf)
```

min	Q1	median	Q3	max	mean	sd	n	r
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	
0.001616457	0.002280948	0.00251836	0.002808373	0.003584204	0.002541658	0.0003724964	1000	

1 row

```
qdata(~b.boot, c(0.025, 0.975), data=bootstrapresultsdf)
```

```
##        2.5%       97.5% 
## 0.001841977 0.003291035
```

The mean of all b_{boot} is 0.002547277. with a 95% confidence interval $0.001901773 \leq b_{boot} \leq 0.003281046$.

Using the means of a_{boot} and b_{boot}, our estimate of the model is:

$$\widehat{DeathsRate}_i = -0.006948612 + (0.002547277 * ObesityRate_i)$$

A: when the obesity rate is 0, the deaths rate is -0.006948612%. This is meaningless because most countries do not have an obesity rate of 0.

B: As the obesity rate increases by 1%, then the deaths rate will increase by an *average* of 0.002547277%.

Conclusions

We can conclude that the obesity rate in Canada is higher in 2019 than 2013, and higher in 2013 than in 2009. Contrary to some economical theories and ideas that suggest that lower GDP results in poorer eating habits (fast food, junk food, etc.), we found higher GDP rates in Canada were related to higher obesity prevalence. However, the study mentioned in the introduction also found that obesity was actually more prevalent in countries that had a higher GDP. This suggests that our findings are in agreement with some past findings (Oshakbayev et al., 2022).

We can also conclude that there is a positive linear relationship between obesity and COVID-19 related deaths and between obesity and confirmed COVID-19 positive cases. Furthermore, we were also able to model this relationship via linear regression. This supports previous findings in a study by Kompaniyets et al. which found COVID-19 death rates were 10 times higher in countries where more than half of the adult population is classified as overweight (2020).

Finally, we can conclude that obesity is definitely an important factor when looking at the outcomes of COVID-19 related outcomes.

Limitations and Future Directions

There are some limitations to our project analysis which I will discuss in this section and some future directions we can take based on these limitations.

For first question, we use the variable GDP as our indicator of how well the population is doing economically. However, this may not be the best indicator of how well the people are doing and it may not capture how well pecific populations are doing. In fact, in a study by Templin et al., they looked at obesity's relationship with one's household income and found that obesity was most prevalent among the poor (2019). We may want to consider different economical variables to measure “how well” a population is doing.

Additionally, because we only looked at Canada when analyzing the relationship between GDP and obesity, we do not if this relationship holds for other countries. It may be beneficial to look at different countries and compare countries obesity rates by their GDP value (group them as a developed/developing country).

Lastly for question one, our Canadian sample size included people from ages of 5-79, which is a big age gap for our sample. This does not allow us to look at specific age populations within the country, which possibly could skew data.

For the second question, although we found a relationship found between COVID-19 deaths and confirmed positive cases, we cannot specifically say if those who were obese were the ones who had a positive case of COVID-19 or died as a result of COVID-19. Deeper and further analysis will be needed to determine if this relationship still exists for specific populations.

References

Kompaniyets, L., Goodman, A.B., Belay, B., Freedman, D.S., Sucosky, M.S., Lange, S.J., Gundlapalli, A.V., Boehmer, T.K. and Blanck, H.M., 2021. Body mass index and risk for COVID-19–related hospitalization, intensive care unit admission, invasive mechanical ventilation, and death—United States, March–December 2020. Morbidity and Mortality Weekly Report, 70(10), p.355.

Open Government. (2021) Government of Canada. Overweight and obesity based on measured body mass index, by age group and sex, open.canada.ca, accessed 17 October 2022.

Oshakbayev, K., Zhankalova, Z., Gazaliyeva, M., Mustafin, K., Bedelbayeva, G., Dukenbayeva, B., Otarbayev, N. and Tordai, A., 2022. Association between COVID-19 morbidity, mortality, and gross domestic product, overweight/obesity, non-communicable diseases, vaccination rate: A cross-sectional study. Journal of infection and public health, 15(2), pp.255-260.

Ren, M. (2020) Kaggle, COVID-19 Healthy Diet Dataset, kaggle.com, accessed 17 October 2022.

Templin, T., Cravo Oliveira Hashiguchi, T., Thomson, B., Dieleman, J. and Bendavid, E., 2019. The overweight and obesity transition from the wealthy to the poor in low-and middle-income countries: A survey of household data from 103 countries. PLoS medicine, 16(11), p.e1002968.