DATA 602 - Project Report

Team Data Nerds (Group 10) MDSA, Fall - 2022 University of Calgary

#Introduction

Initially, we explored and visualized the trends for the Mental Health disorders during our DATA 601 project. Mental health disorders are one of the biggest prevailing issues in the society and it has been a belief in the society that this issue has been increasing significantly over the years. We wanted to understand this in detail and hence, from our study in DATA 601, wanted to understand patterns, across various parameters such as age, gender, education etc., on how this issue has emerged over the years.

As per the scope for DATA 602, we want to understand and perform some additional studies wherein we involve some of our hypothesis based on our prior exploration, determine the statistical aspect of our study and understand the validity of this data set.

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected
by this.
```

```
##
## Attaching package: 'mosaic'
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
#reading data-sets
data_disease_type = read_excel("Mental health Depression disorder Data.xlsx", sheet = "prevalence-by-mental-and-substa")
data_depression_by_education = read_excel("Mental health Depression disorder Data.xlsx", sheet = "depression-by-level-of-educatio")
data_depression_by_age = read_excel("Mental health Depression disorder Data.xlsx", sheet = "prevalence-of-depression-by-age")
data_depression_by_gender = read_excel("Mental health Depression disorder Data.xlsx", sheet = "prevalence-of-depression-males-")
data_suicide_rates = read_excel("Mental health Depression disorder Data.xlsx", sheet = "suicide-rates-vs-prevalence-of-")
data_depression_by_count = read_excel("Mental health Depression disorder Data.xlsx", sheet = "number-with-depression-by-count")
```

Question 1

What does the data reveal about a specific group (age, gender) suffering from mental illness?

Part A

Is depression higher in men or women?

This section of the study involves the prevalence of depression among the population based on gender i.e. is there a difference in how depression builds up between different genders in the society (for the study we are only considering two genders i.e. Male and Female). Also, we will try to identify if there is any correlation between the depression levels of the two genders under observation.

As a next step, we try to determine if there is any relation between the two groups. We Build up the hypothesis where we initially deny for any relation between the depression levels of the two genders and try to determine the prevalent hypothesis statistically.

The Hypothesis areas follows: H0: There is no relationship between the depression levels in Males and Females. H1: There is a relationship between the depression levels in Males and Females and can be expressed in terms of a linear function.

```
cat("H0:B=0(Y cannot be expressed as a linear function of X)\n")
```

```
## H0:B=0(Y cannot be expressed as a linear function of X)
```

```
cat("HA:B≠0(Y can be expressed as a linear function of X)\n")
```

```
## HA:B≠0(Y can be expressed as a linear function of X)
```

```
predictdepression = lm(`Prevalence in males (%)` ~ `Prevalence in females (%)`,data=data
_depression_by_gender)
predictdepression
```

```
## 
## Call:
## lm(formula = `Prevalence in males (%)` ~ `Prevalence in females (%)`, 
##     data = data_depression_by_gender)
## 
## Coefficients:
##                 (Intercept)  `Prevalence in females (%)`  
##                      0.6559                       0.5175
```

```
summary(aov(predictdepression))
```

```
##                                Df Sum Sq Mean Sq F value Pr(>F)    
## `Prevalence in females (%)`    1 1222.8  1222.8   12679 <2e-16 ***
## Residuals                    6466  623.6     0.1                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 41339 observations deleted due to missingness
```

```
cat("\nThe p value for the given data distribution is close to 0, therefore we can rejec
t the null hypotheses. In the context of the data of Male vs Female depression affected
 population, the data is linear i.e. as the % population affected by depression increase
s for either gender, based on a linear model, it increases for the other as well")
```

```
##
## The p value for the given data distribution is close to 0, therefore we can reject th
e null hypotheses. In the context of the data of Male vs Female depression affected popu
lation, the data is linear i.e. as the % population affected by depression increases for
either gender, based on a linear model, it increases for the other as well
```

The analysis of the data based on F-test resulted in the below mentioned values:

F-value = 12679

P-value < 2*10^{-16}

Such a low p-value indicates that the null hypothesis cannot be supported which signifies that there exists a linear relationship between the depression levels of males and females.

```
data_depression_by_gender = na.omit(data_depression_by_gender)
data_depression_by_gender$Male_count = (data_depression_by_gender$`Prevalence in males
(%)` * data_depression_by_gender$Population)/100
data_depression_by_gender$Female_count = (data_depression_by_gender$`Prevalence in femal
es (%)` * data_depression_by_gender$Population)/100
data_depression_by_gender
```

Entity <chr>	C... <chr>	Year <chr>	Prevalence in males (%) <dbl>	Prevalence in females (%) <dbl>	Popul <
Afghanistan	AFG	1990.0	3.499982	4.647815	1241
Afghanistan	AFG	1991.0	3.503947	4.655772	1329
Afghanistan	AFG	1992.0	3.508912	4.662066	1448
Afghanistan	AFG	1993.0	3.513429	4.669012	1581
Afghanistan	AFG	1994.0	3.515578	4.673050	1707
Afghanistan	AFG	1995.0	3.522763	4.674739	1811
Afghanistan	AFG	1996.0	3.524278	4.678305	1885
Afghanistan	AFG	1997.0	3.529116	4.678696	1935
Afghanistan	AFG	1998.0	3.534040	4.680175	1973
Afghanistan	AFG	1999.0	3.537596	4.683704	2017

1-10 of 5,488 rows | 1-6 of 8 columns Previous **1** 2 3 4 5 6 ... 549 Next

```
ggplot(data_depression_by_gender, aes(x = `Prevalence in males (%)`, y = `Prevalence in
females (%)`)) + geom_point(col = "red", alpha = 0.3, size = 0.5) + geom_smooth(method=
"lm", col="green")
```

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

The above chart is a representation of the linear relationship between the two genders that was statistically proven by the hypothesis testing done earlier.

```
predicteddepression= predictdepression$fitted.values
eisdepression = predictdepression$residuals
gender.df = data.frame(predicteddepression, eisdepression)
gender.df
```

	predicteddepression <dbl>	eisdepression <dbl>
191	3.061000	0.4389821877
192	3.065118	0.4388298717
193	3.068374	0.4405378986
194	3.071969	0.4414603317
195	3.074058	0.4415199686
196	3.074932	0.4478311023
197	3.076778	0.4475000750
198	3.076980	0.4521358006
199	3.077745	0.4562948328
200	3.079571	0.4580245012

1-10 of 6,468 rows Previous **1** 2 3 4 5 6 ... 647 Next

```
ggplot(gender.df, aes(x = predicteddepression, y = eisdepression)) + geom_point(size=0.3
, col='blue', position="jitter") + xlab("Predicted Depression Values") + ylab("Residual
s") + ggtitle("Plot of Fits to Residuals") + geom_hline(yintercept=0, color="red", linet
ype="dashed")
```

To inspect the homoscedasticity condition, we plot the fitted.values with the residuals in the above graph. The values are densely populated near to the origin line(x-axis) hence suggesting that the difference between the predicted and the actual values is not high and the reliability of the prediction model is high

```
N = 1000
cor = numeric(N)
nsize = dim(data_depression_by_gender)[1]
for(i in 1:N)
{
 index = sample(nsize, replace=TRUE)
 sample = data_depression_by_gender[index, ]
 cor[i] = cor(~`Prevalence in males (%)`, ~`Prevalence in females (%)`, data=sample)
}
bootstrapresultsdf = data.frame(cor)
bootstrapresultsdf
```

cor
<dbl>
0.8162953
0.8115053
0.8067872
0.8068443
0.8166480
0.8081095
0.8103728
0.8133650
0.8059440
0.8072014

1-10 of 1,000 rows Previous **1** 2 3 4 5 6 ... 100 Next

```
ggplot(bootstrapresultsdf,aes(cor)) +geom_histogram(col="pink",fill="purple")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

In this section of the study, we ran a bootstrap model to calculate the correlation index between the data for the two genders and plotted them over a histogram. We also infer that with a 95% confidence interval, the value of the correlation index is between the range of 0.798 and 0.820, which depicts a strong correlation between the two parameters

```
gender_data = gather(data_depression_by_gender,"Gender","Affected_Percentage",4:6)
gender_data = na.omit(gender_data)
gender_data = gender_data %>%
  filter(Gender != "Population")

ggplot(data = gender_data, aes(x = Gender,y = Affected_Percentage)) + geom_bar(col = "bl
ue",fill = "blue",stat = "identity", position = "dodge") + xlab("Gender") + ylab("Affect
ed Population (%)") + ggtitle("Depression statistics across Genders")
```

Lastly, we plot the overall data for the Affected population between the two genders and infer that females have a higher level of depression as compared to males

Part B

Which age group has more depression?

```
age_data = gather(data_depression_by_age,"Age Group","Affected_Percentage",4:13)
age_data
```

Entity	Code	Year	Age Group	Affected_Percentage
<chr>	<chr>	<dbl>	<chr>	<dbl>
Afghanistan	AFG	1990	20-24 years old (%)	4.4178018
Afghanistan	AFG	1991	20-24 years old (%)	4.4335243
Afghanistan	AFG	1992	20-24 years old (%)	4.4536892
Afghanistan	AFG	1993	20-24 years old (%)	4.4645167
Afghanistan	AFG	1994	20-24 years old (%)	4.4629596
Afghanistan	AFG	1995	20-24 years old (%)	4.4566334
Afghanistan	AFG	1996	20-24 years old (%)	4.4438388
Afghanistan	AFG	1997	20-24 years old (%)	4.4272566
Afghanistan	AFG	1998	20-24 years old (%)	4.4170786
Afghanistan	AFG	1999	20-24 years old (%)	4.4221761

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 ... 1000 Next

```
age_data_filtered = age_data %>%
  filter(`Age Group`=="10-14 years old (%)" | `Age Group`=="15-49 years old (%)" | `Age
 Group`=="50-69 years old (%)" | `Age Group`=="70+ years old (%)")
age_data_filtered
```

Entity	Code	Year	Age Group	Affected_Percentage
<chr>	<chr>	<dbl>	<chr>	<dbl>

Entity <chr>	Code <chr>	Year <dbl>	Age Group <chr>	Affected_Percentage <dbl>
Afghanistan	AFG	1990	10-14 years old (%)	1.5946763
Afghanistan	AFG	1991	10-14 years old (%)	1.5883556
Afghanistan	AFG	1992	10-14 years old (%)	1.5779797
Afghanistan	AFG	1993	10-14 years old (%)	1.5772008
Afghanistan	AFG	1994	10-14 years old (%)	1.5708456
Afghanistan	AFG	1995	10-14 years old (%)	1.5745642
Afghanistan	AFG	1996	10-14 years old (%)	1.5747433
Afghanistan	AFG	1997	10-14 years old (%)	1.5706078
Afghanistan	AFG	1998	10-14 years old (%)	1.5755239
Afghanistan	AFG	1999	10-14 years old (%)	1.5751997

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 ... 1000 Next

```
ggplot(data = age_data_filtered, aes(x = `Age Group`,y = `Affected_Percentage`)) + geom_
boxplot(col = "purple", fill = "pink") + xlab("Age Group") + ylab("Affected Population
 (%)") + ggtitle("Box Plot across various Age Groups")
```

The above graph visualizes how the depression percentage varies in the overall population across various age groups. From this, we infer that with the increase in ages, a person falling into different age groups, the depression level increases

```
age_data_filtered$`Depression Category` = ifelse(age_data_filtered$Affected_Percentage >
=4.5,"High","Low")
head(age_data_filtered,5)
```

Entity <chr>	C... <chr>	Y... <dbl>	Age Group <chr>	Affected_Percentage <dbl>	Depression Category <chr>
Afghanistan	AFG	1990	10-14 years old (%)	1.594676	Low
Afghanistan	AFG	1991	10-14 years old (%)	1.588356	Low
Afghanistan	AFG	1992	10-14 years old (%)	1.577980	Low
Afghanistan	AFG	1993	10-14 years old (%)	1.577201	Low
Afghanistan	AFG	1994	10-14 years old (%)	1.570846	Low

5 rows

```
tail(age_data_filtered,5)
```

Entity	C…	Y…	Age Group	Affected_Percentage	Depression Category
<chr>	<chr>	<dbl>	<chr>	<dbl>	<chr>
Zimbabwe	ZWE	2013	15-49 years old (%)	3.133858	Low
Zimbabwe	ZWE	2014	15-49 years old (%)	3.153508	Low
Zimbabwe	ZWE	2015	15-49 years old (%)	3.179233	Low
Zimbabwe	ZWE	2016	15-49 years old (%)	3.206184	Low
Zimbabwe	ZWE	2017	15-49 years old (%)	3.233777	Low

5 rows

Based on the initial inferences from the above section, we now try to statistically determine a relation between depression percentage and age groups.

```
print("H0: There is no relationship between Depression percentages and Age Group")
```

```
## [1] "H0: There is no relationship between Depression percentages and Age Group"
```

```
print("H0: There is a relationship between Depression percentages and Age Group")
```

```
## [1] "H0: There is a relationship between Depression percentages and Age Group"
```

```
population_data = subset(data_depression_by_gender,select = c("Entity","Year","Populatio
n"))
population_data$Year= as.double(population_data$Year)

age_data_modified = left_join(age_data_filtered,population_data,by=c("Entity","Year"))

age_data_modified$`Affected Population`=(age_data_modified$Population*age_data_modified
$Affected_Percentage)/100
age_data_modified
```

Entity	C…	Y…	Age Group	Affected_Percentage	Depression Category
<chr>	<chr>	<dbl>	<chr>	<dbl>	<chr>
Afghanistan	AFG	1990	10-14 years old (%)	1.5946763	Low
Afghanistan	AFG	1991	10-14 years old (%)	1.5883556	Low
Afghanistan	AFG	1992	10-14 years old (%)	1.5779797	Low
Afghanistan	AFG	1993	10-14 years old (%)	1.5772008	Low
Afghanistan	AFG	1994	10-14 years old (%)	1.5708456	Low
Afghanistan	AFG	1995	10-14 years old (%)	1.5745642	Low
Afghanistan	AFG	1996	10-14 years old (%)	1.5747433	Low

Entity	C...	Y...	Age Group	Affected_Percentage	Depression Category
<chr>	<chr>	<dbl>	<chr>	<dbl>	<chr>
Afghanistan	AFG	1997	10-14 years old (%)	1.5706078	Low
Afghanistan	AFG	1998	10-14 years old (%)	1.5755239	Low
Afghanistan	AFG	1999	10-14 years old (%)	1.5751997	Low

1-10 of 10,000 rows | 1-7 of 8 columns Previous **1** 2 3 4 5 6 ... 1000 Next

```
age_group_chi = age_data_modified %>% group_by(`Entity`,`Age Group`,`Depression Category
`) %>% summarise(`Average Affected Population`=mean(`Affected Population`))
```

```
## `summarise()` has grouped output by 'Entity', 'Age Group'. You can override
## using the `.groups` argument.
```

```
age_group_chi=na.omit(age_group_chi)
age_group.df = age_group_chi %>% group_by(`Depression Category`,`Age Group`) %>% summari
se(`Average Affected Population`=mean(`Average Affected Population`))
```

```
## `summarise()` has grouped output by 'Depression Category'. You can override
## using the `.groups` argument.
```

```
age_group.df
```

Depression Category	Age Group	Average Affected Population
<chr>	<chr>	<dbl>
High	15-49 years old (%)	1479586.9
High	50-69 years old (%)	4316902.0
High	70+ years old (%)	4573306.1
Low	10-14 years old (%)	788778.4
Low	15-49 years old (%)	2876807.6
Low	50-69 years old (%)	609464.3
Low	70+ years old (%)	1051963.0

7 rows

```
age_group.df = spread(age_group.df, `Age Group`,`Average Affected Population`)
age_group.df[is.na(age_group.df)]=0
age_group.df = age_group.df[-c(1)]
rownames(age_group.df)=c("High","Low")
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
chisq.test(age_group.df, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  age_group.df
## X-squared = 5141975, df = 3, p-value < 2.2e-16
```

```
cat("Since the p value is close to zero, it can concluded that the null hypotheses fail
s. Therefore the two are not independant and there is a relationship between the Age Gro
up and the Depression Percentages of the population.")
```

```
## Since the p value is close to zero, it can concluded that the null hypotheses fails.
Therefore the two are not independant and there is a relationship between the Age Group
and the Depression Percentages of the population.
```

We ran a Chi-Squared test to determine if there is a relationship between depression and age groups. As per the results, the p-value obtained is insufficient to support the null hypothesis and hence we reject the same. This statistically proves that there is a relationship between the Age groups and the depression percentages

```
age_data_all = subset(data_depression_by_age,select=c("Entity","Year","All ages (%)"))
age_data_all
```

Entity <chr>	Year <dbl>	All ages (%) <dbl>
Afghanistan	1990	3.218871
Afghanistan	1991	3.203468
Afghanistan	1992	3.156559
Afghanistan	1993	3.120655
Afghanistan	1994	3.082179
Afghanistan	1995	3.039739
Afghanistan	1996	2.994851
Afghanistan	1997	2.952654
Afghanistan	1998	2.915789
Afghanistan	1999	2.877621

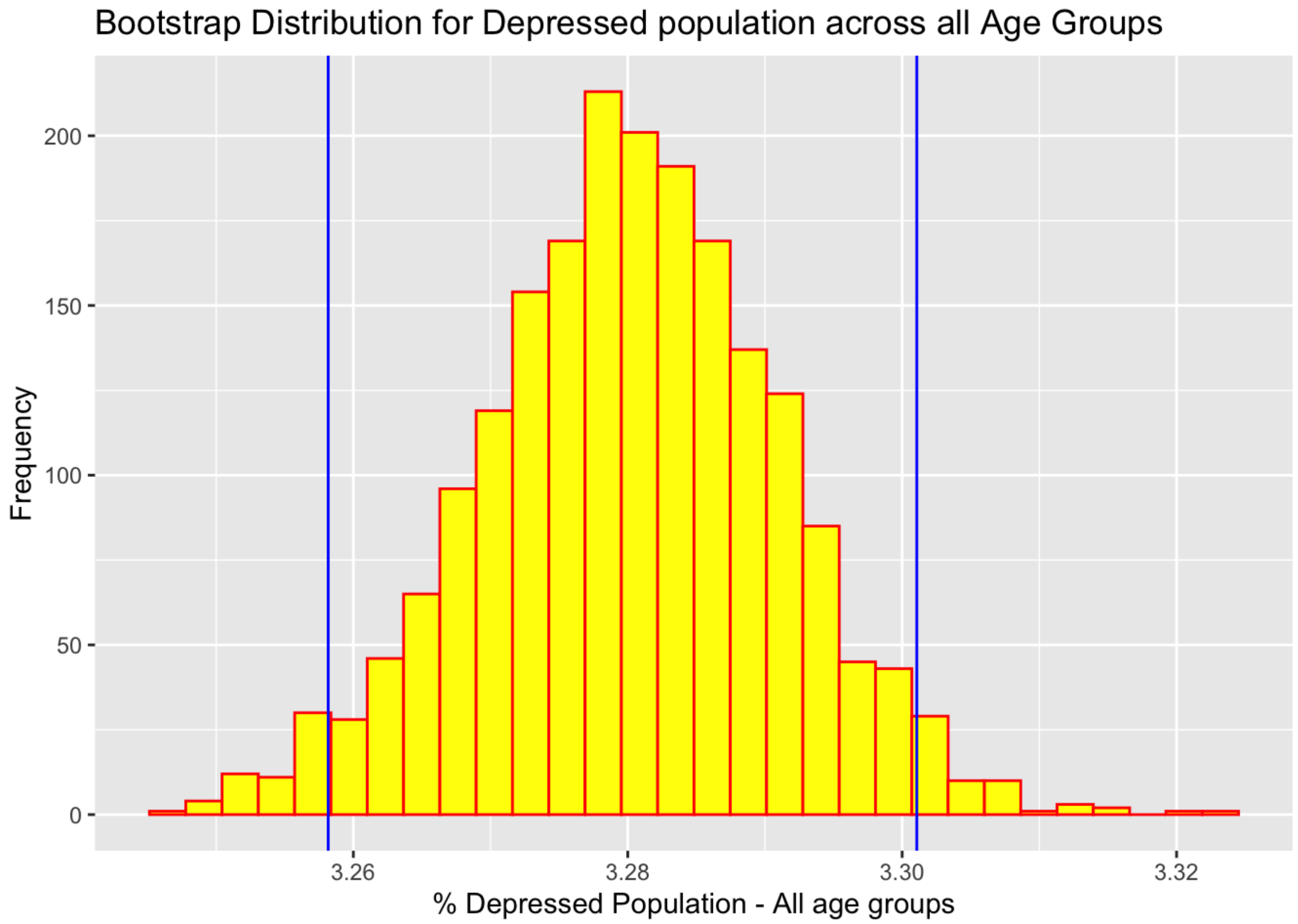
1-10 of 6,468 rows Previous **1** 2 3 4 5 6 ... 647 Next

```
nsamples = 2000
all_level_means = numeric(nsamples)
for(i in 1:nsamples){
  all_level_sample = sample(na.omit(age_data_all$`All ages (%)`), 6468, replace=TRUE)
  all_level_means[i] = mean(all_level_sample)
}
all_level.df=data.frame(all_level_means)
head(all_level.df,5)
```

	all_level_means <dbl>
1	3.281238
2	3.283510
3	3.263169
4	3.279225
5	3.293042

5 rows

```
a=quantile(all_level.df$all_level_means,c(0.025,0.975))
lower = as.numeric(a[1])
upper = as.numeric(a[2])
ggplot(all_level.df,aes(x=all_level_means)) + geom_histogram(col='red',fill='yellow',bin
s=30) +geom_vline(xintercept = lower, col="blue")+geom_vline(xintercept = upper, col="bl
ue") + xlab("% Depressed Population - All age groups") + ylab("Frequency") + ggtitle("Bo
otstrap Distribution for Depressed population across all Age Groups")
```



Finally we run a bootstrap statistic to determine the overall mean depression affected population percentage across all age group levels. We infer from the above graph that with a 95% confidence, the mean of the affected depression percentage lies between the range 3.258 and 3.300 percent

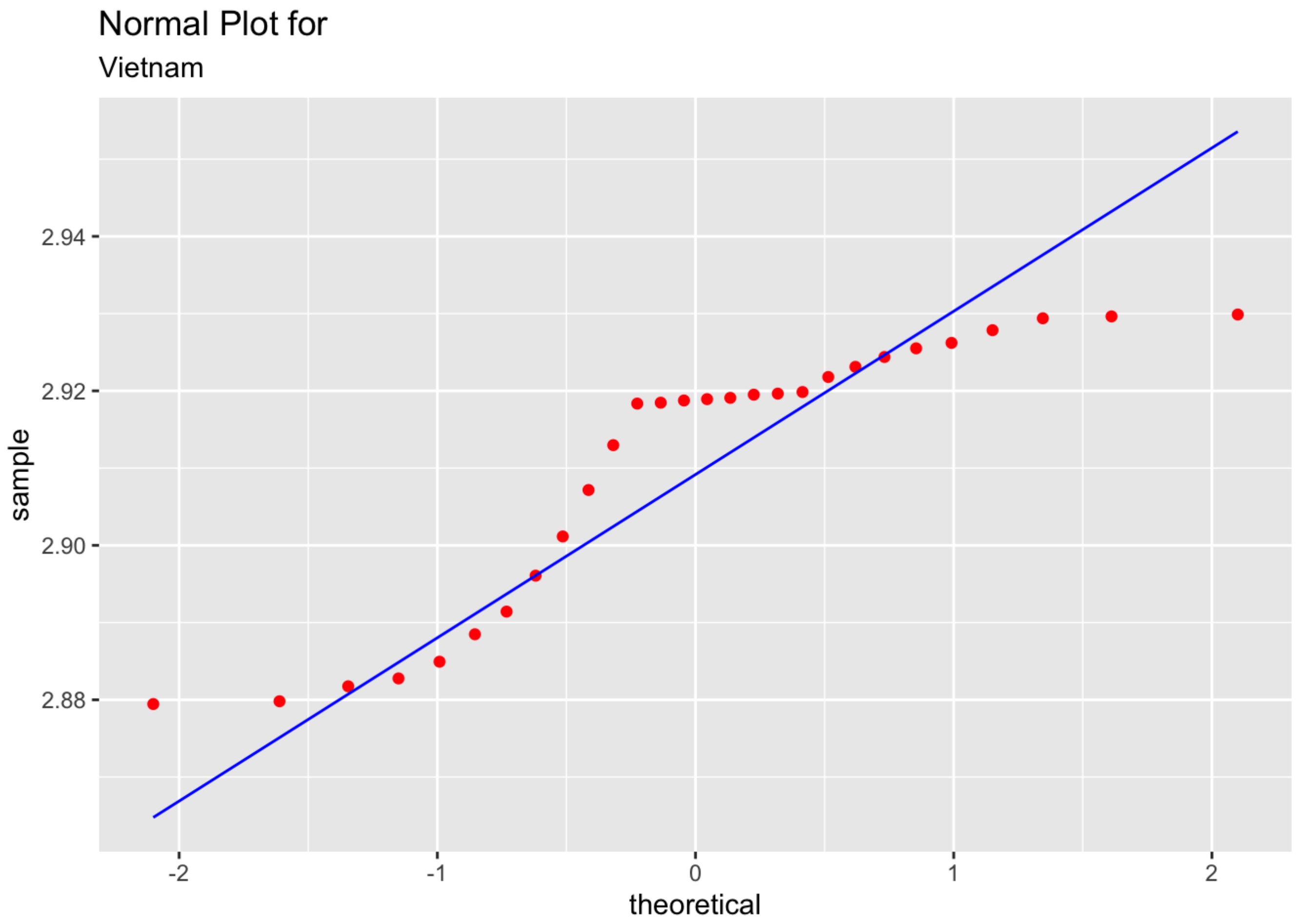
Question 2

Is the mean of the affected population percentage for the chosen country is equal to the overall population mean for depression?

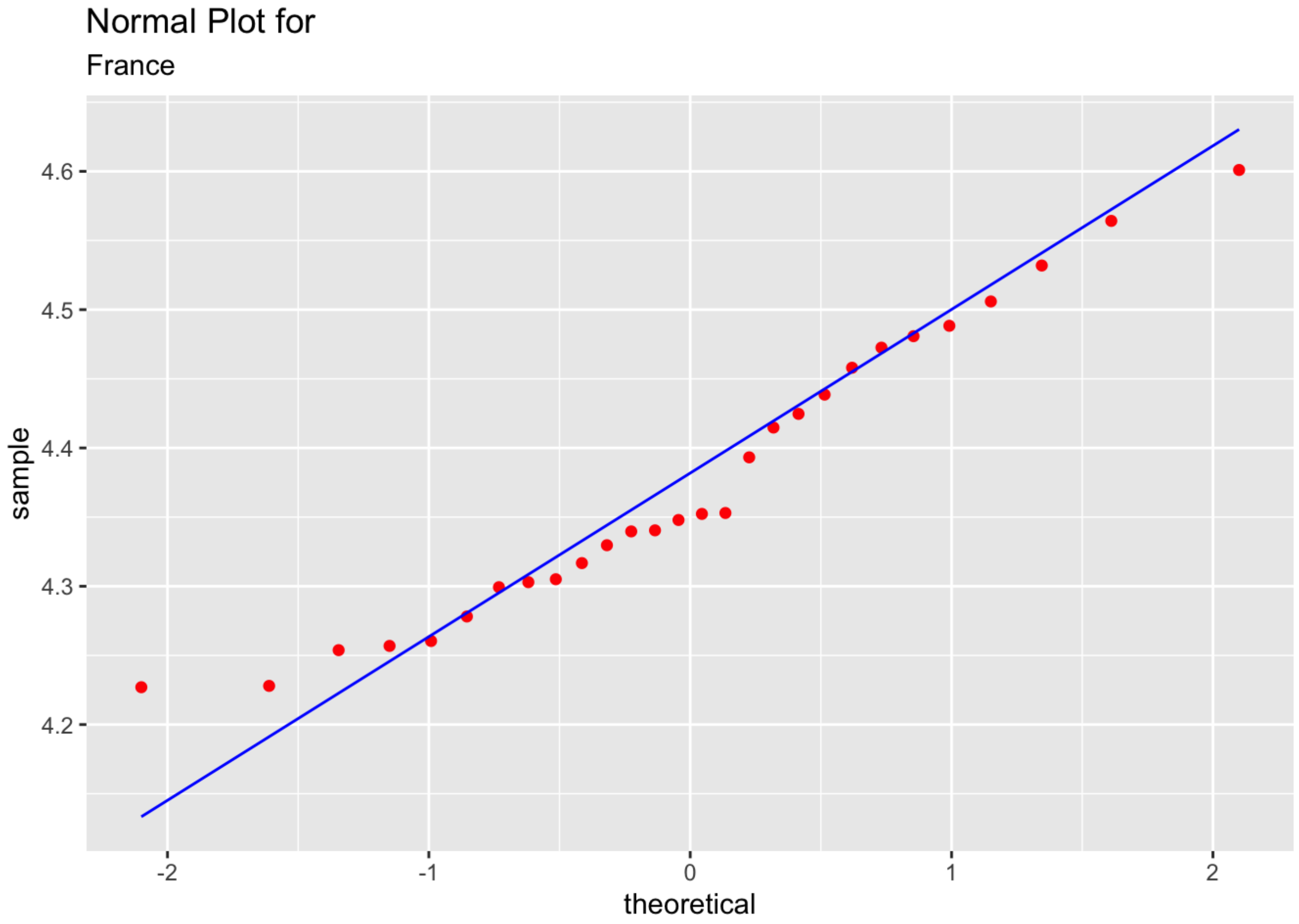
```
Disease_summar_by_country = function(x){

  filtered_dataset = data_disease_type %>%
    filter(Entity == x)

  print(favstats(filtered_dataset$`Schizophrenia (%)`))
  print(favstats(filtered_dataset$`Bipolar disorder (%)`))
  print(favstats(filtered_dataset$`Eating disorders (%)`))
  print(favstats(filtered_dataset$`Anxiety disorders (%)`))
}
country_1 = c("Vietnam")
country_2 = c("France")
all_countries = as.list(unique(data_disease_type$Entity)) # Contains a list of all count
ries in the dataset

countrywise_summary = Disease_summar_by_country(country_1)
```

```
##        min        Q1    median        Q3       max      mean          sd  n
##  0.2227357 0.2274408 0.2341545 0.2390236 0.2446628 0.2335535 0.006813693 28
##  missing
##        0
##        min        Q1    median        Q3       max      mean          sd  n
##  0.5391997 0.5408801 0.5454858 0.5491571 0.5526533 0.5453234 0.004510197 28
##  missing
##        0
##         min         Q1    median        Q3       max      mean         sd  n
##  0.08348448 0.09031541 0.1011236 0.1135386 0.1284665 0.1026424 0.01396906 28
##  missing
##        0
##       min       Q1   median      Q3      max     mean         sd  n missing
##  2.023393 2.028385 2.034345 2.03639 2.066871 2.035264 0.01015182 28       0
```

```
country_1_data = data_disease_type %>%
    filter(Entity == country_1)

ggplot(data = country_1_data, aes(sample = `Depression (%)`)) + stat_qq(col = "red") +st
at_qqline(col = "blue") + ggtitle("Normal Plot for",country_1)
```



```
country_2_data = data_disease_type %>%
    filter(Entity == country_2)

ggplot(data = country_2_data, aes(sample = `Depression (%)`)) + stat_qq(col = "red") +st
at_qqline(col = "blue") + ggtitle("Normal Plot for",country_2)
```



```
cat("H0: The mean of the afftected population percentage for the chosen country is equal
to the overall population mean for depression
H1: The mean of the afftected population percentage for the chosen country is lesser tha
n the overall population mean for depression")
```

```
## H0: The mean of the afftected population percentage for the chosen country is equal t
o the overall population mean for depression
## H1: The mean of the afftected population percentage for the chosen country is lesser
than the overall population mean for depression
```

```
t.test(country_1_data$`Depression (%)`,data_disease_type$`Depression (%)`,alternative =
"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  country_1_data$`Depression (%)` and data_disease_type$`Depression (%)`
## t = -66.718, df = 1174.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5725866
## sample estimates:
## mean of x mean of y
##  2.910582  3.497654
```

```
cat("Since the p-value is approximately zero, we reject the null hypothesis and infer th
at for the chosen country, the average percentage of population affected via depression
 is lesser than the world average")
```

```
## Since the p-value is approximately zero, we reject the null hypothesis and infer that
for the chosen country, the average percentage of population affected via depression is
lesser than the world average
```

```
x = c("Vietnam")
countrywise_data = function(x) {
  country_mean = data_disease_type %>%
    filter(Entity == x)
}
country_filtered_data = countrywise_data(x)
country_filtered_data_for_prediction = country_filtered_data %>%
  filter(Year != 2017)
predict_depression_percentage = lm(`Depression (%)`~Year, data=country_filtered_data_for
_prediction)
predict(predict_depression_percentage, data.frame(Year=2017))
```

```
##        1
## 2.888798
```

```
country_filtered_data_predicted_year = country_filtered_data %>%
  filter(Year == 2017)

country_filtered_data_predicted_year$`Depression (%)`
```

```
## [1] 2.87945
```

Question 3

Is there a relation between the education levels and the depression percentages in the given population ?

This study focuses on analyzing the relationship between the various education levels of people in 25 different countries and look into how it relates to the depression percentages. In this data sheet we are dealing with data from 2014, this data features depression statistics in the form of percentage of people suffering from depression for three different education levels- below upper secondary, upper secondary and tertiary. Additionally we also have employement and job seeking statistics for these given education levels. That is, the percentages of people in certain education level who are looking for jobs as well as the people in these categories that are already employed.

Data Sources The following data sources from the OurWorldIndata Our World in Data. Available at: https://ourworldindata.org/mental-health#data-sources (https://ourworldindata.org/mental-health#data-sources) to work on the statistical analysis:- 1. https://ourworldindata.org/mental-health#data-sources (https://ourworldindata.org/mental-health#data-sources) CSV name: Mental Health Depression Disorder Data Sheet name: depression-by-level-of-education The resource, as mentioned by the author, is open to use freely with proper citations under the Creative Commons BY License (link).

```
head(data_depression_by_education,5)
```

Entity <chr>	**C…** <chr>	**Y…** <dbl>	**All levels (active) (%)** <dbl>	**All levels (employed) (%)** <dbl>	**A**
Austria	AUT	2014	6.5	4.7	
Belgium	BEL	2014	5.0	4.1	
Czech Republic	CZE	2014	3.0	2.6	
Denmark	DNK	2014	6.7	5.7	
Estonia	EST	2014	3.8	3.8	

5 rows | 1-6 of 15 columns

```
tail(data_depression_by_education,5)
```

Entity <chr>	**C…** <chr>	**Y…** <dbl>	**All levels (active) (%)** <dbl>	**All levels (employed) (%)** <dbl>	**A**
Slovenia	SVN	2014	7.6	6.0	
Spain	ESP	2014	5.5	4.1	
Sweden	SWE	2014	8.4	8.0	
Turkey	TUR	2014	10.2	9.6	
United Kingdom	GBR	2014	7.4	6.3	

5 rows | 1-6 of 15 columns

From the head and tail values above we can understand the data better. In the data we have 26 countries, mostly in the European continent. The percentage values have been provided as the overall values for all education levels as well as segregated education levels.

```
education_levels=gather(data_depression_by_education,"Education Levels","Affected_Percen
tage",4:15)
education_levels
```

Entity	C...	Y...	Education Levels	Affected_Percentage
<chr>	<chr>	<dbl>	<chr>	<dbl>
Austria	AUT	2014	All levels (active) (%)	6.5
Belgium	BEL	2014	All levels (active) (%)	5.0
Czech Republic	CZE	2014	All levels (active) (%)	3.0
Denmark	DNK	2014	All levels (active) (%)	6.7
Estonia	EST	2014	All levels (active) (%)	3.8
Finland	FIN	2014	All levels (active) (%)	8.5
France	FRA	2014	All levels (active) (%)	5.2
Germany	DEU	2014	All levels (active) (%)	10.3
Greece	GRC	2014	All levels (active) (%)	2.8
Hungary	HUN	2014	All levels (active) (%)	2.8

1-10 of 312 rows Previous **1** 2 3 4 5 6 ... 32 Next

```
education_data_filtered = education_levels %>%
  filter(`Education Levels`=="Upper secondary & post-secondary non-tertiary (total) (%)"
| `Education Levels`=="Tertiary (total) (%)" | `Education Levels`=="Below upper secondar
y (total) (%)")
education_data_filtered
```

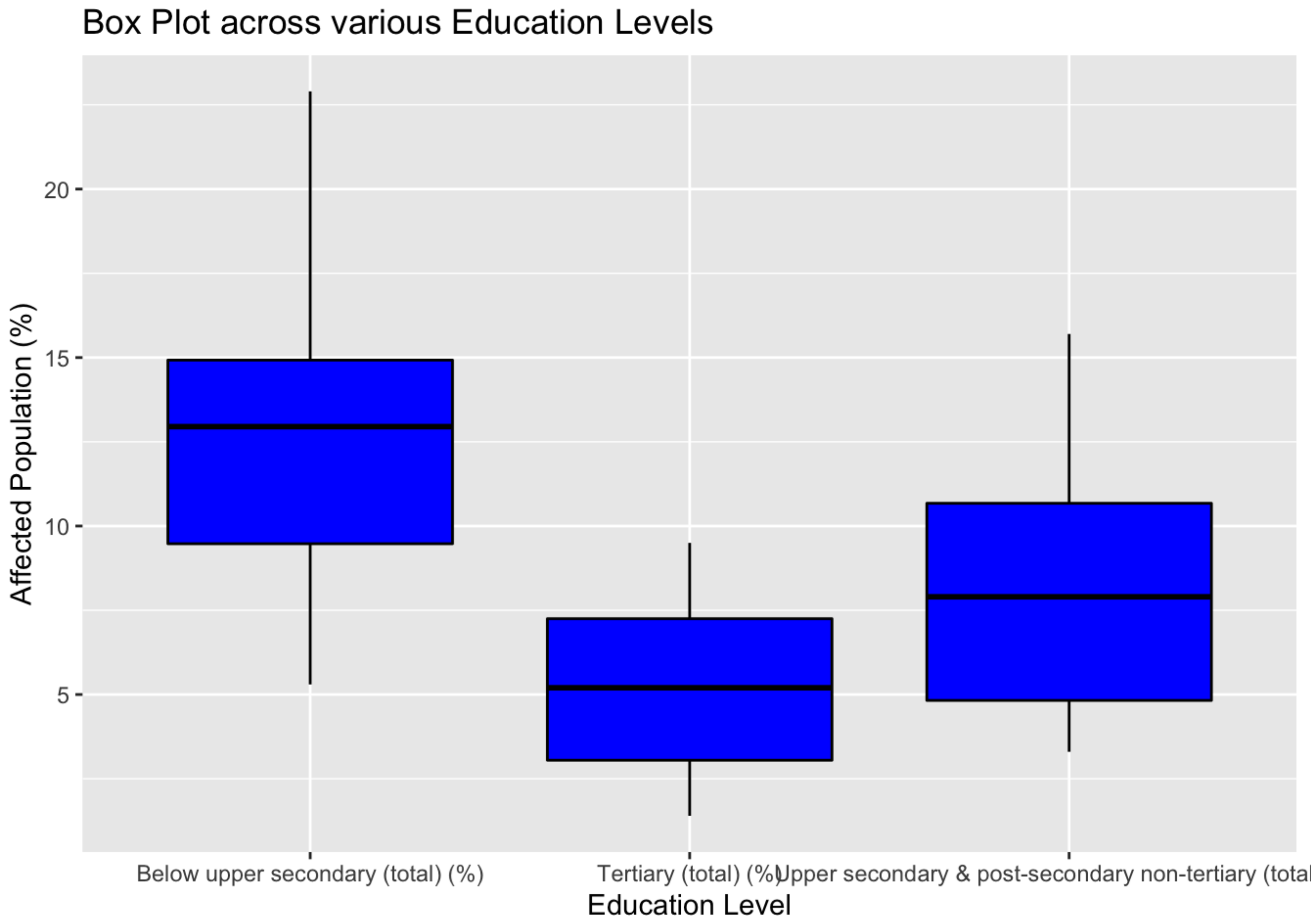
Entity	C...	Y...	Education Levels
<chr>	<chr>	<dbl>	<chr>
Austria	AUT	2014	Below upper secondary (total) (%)
Belgium	BEL	2014	Below upper secondary (total) (%)
Czech Republic	CZE	2014	Below upper secondary (total) (%)
Denmark	DNK	2014	Below upper secondary (total) (%)
Estonia	EST	2014	Below upper secondary (total) (%)
Finland	FIN	2014	Below upper secondary (total) (%)
France	FRA	2014	Below upper secondary (total) (%)

Entity <chr>	C… <chr>	Y… <dbl>	Education Levels <chr>
Germany	DEU	2014	Below upper secondary (total) (%)
Greece	GRC	2014	Below upper secondary (total) (%)
Hungary	HUN	2014	Below upper secondary (total) (%)

1-10 of 78 rows | 1-4 of 5 columns Previous **1** 2 3 4 5 6 ... 8 Next

Visualizing the Education Levels over All Countries For ideal plotting,we need to rearrange the data available into one column “Education Levels” which is done through the tidyr and msaic packages. The visualization below plots the values of “Affected Depression Percentages” over the 25 countries provided for the three education categories. From the plot, we can infer a few things, 1. A person pursuing his tertiary education is the least likely to deal with depression in comparison to the other groups 2. Tertiary education level has the least spread of data that is Standard Deviation, which is confirmed in the next r chunk 3. Below Upper Secondary Education level has the most of amount of spread of data as well as highest depression percentages in the three groups

```
ggplot(data = education_data_filtered, aes(x = `Education Levels`,y = `Affected_Percenta
ge`)) + geom_boxplot(col = "black", fill = "blue") + xlab("Education Level") + ylab("Aff
ected Population (%)") + ggtitle("Box Plot across various Education Levels")
```



The favtstats function provides us with some basic statistical metrics for the three groups. As infered from the plot above the below upper secondary has the highest standard deviation value of 4.610483, followed by upper secondary with 2.662656. Another thing to notice in the data would be Q1 values in the below upper secondary data is approximately same as the max value seen in the tertiary data.

```
favstats(data_depression_by_education$`Below upper secondary (total) (%)`)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
5.3	9.475	12.95	14.925	22.9	12.28462	4.610483	26	0

1 row

```
favstats(data_depression_by_education$`Upper secondary & post-secondary non-tertiary (to
tal) (%)`)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
3.3	4.825	7.9	10.675	15.7	8.1	3.460751	26	0

1 row

```
favstats(data_depression_by_education$`Tertiary (total) (%)`)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
1.4	3.05	5.2	7.25	9.5	5.457692	2.662656	26	0

1 row

CHI Squared Analysis The guiding question for this specific study looks to inquire about the relationship between the education levels and the depression rates. For this we need to prepare the data in the before we utilize the CHI test. The “Affected Percentages” for the different classes are classified into “High” or “Low” categories based on the approximate median of the data. These categories are added to the dataframe as the field “Depression Category”.

```
education_data_filtered$`Depression Category` = ifelse(education_data_filtered$Affected_
Percentage >=4.5,"High","Low")
head(education_data_filtered,5)
```

Entity	C…	Y…	Education Levels	Affected_Percentage	Depr
<chr>	<chr>	<dbl>	<chr>	<dbl>	<chr
Austria	AUT	2014	Below upper secondary (total) (%)	15.2	High
Belgium	BEL	2014	Below upper secondary (total) (%)	11.6	High
Czech Republic	CZE	2014	Below upper secondary (total) (%)	6.0	High
Denmark	DNK	2014	Below upper secondary (total) (%)	15.5	High

Entity	C…	Y…	Education Levels	Affected_Percentage	Depr
<chr>	<chr>	<dbl>	<chr>	<dbl>	<chr
Estonia	EST	2014	Below upper secondary (total) (%)	6.4	High

5 rows

```
tail(education_data_filtered,5)
```

Entity	C…	Y…	Education Levels
<chr>	<chr>	<dbl>	<chr>
Slovenia	SVN	2014	Upper secondary & post-secondary non-tertiary (total) (%)
Spain	ESP	2014	Upper secondary & post-secondary non-tertiary (total) (%)
Sweden	SWE	2014	Upper secondary & post-secondary non-tertiary (total) (%)
Turkey	TUR	2014	Upper secondary & post-secondary non-tertiary (total) (%)
United Kingdom	GBR	2014	Upper secondary & post-secondary non-tertiary (total) (%)

5 rows | 1-4 of 6 columns

```
population_data = subset(data_depression_by_gender,select = c("Entity","Year","Populatio
n"))
population_data$Year= as.double(population_data$Year)

education_data_modified = left_join(education_data_filtered,population_data,by=c("Entit
y","Year"))

education_data_modified$`Affected Population`=(education_data_modified$Population*educat
ion_data_modified$Affected_Percentage)/100
education_data_modified
```

Entity	C…	Y…	Education Levels
<chr>	<chr>	<dbl>	<chr>
Austria	AUT	2014	Below upper secondary (total) (%)
Belgium	BEL	2014	Below upper secondary (total) (%)
Czech Republic	CZE	2014	Below upper secondary (total) (%)
Denmark	DNK	2014	Below upper secondary (total) (%)
Estonia	EST	2014	Below upper secondary (total) (%)
Finland	FIN	2014	Below upper secondary (total) (%)
France	FRA	2014	Below upper secondary (total) (%)

Entity	C...	Y...	Education Levels
<chr>	<chr>	<dbl>	<chr>
Germany	DEU	2014	Below upper secondary (total) (%)
Greece	GRC	2014	Below upper secondary (total) (%)
Hungary	HUN	2014	Below upper secondary (total) (%)

1-10 of 78 rows | 1-4 of 8 columns Previous **1** 2 3 4 5 6 ... 8 Next

```
education_group_chi = education_data_modified %>% group_by(`Entity`,`Education Levels`,`
Depression Category`) %>% summarise(`Average Affected Population`=mean(`Affected Populat
ion`))
```

```
## `summarise()` has grouped output by 'Entity', 'Education Levels'. You can
## override using the `.groups` argument.
```

```
education_group_chi=na.omit(education_group_chi)
education_levels.df = education_group_chi %>% group_by(`Depression Category`,`Education
Levels`) %>% summarise(`Average Affected Population`=mean(`Average Affected Population
`))
```

```
## `summarise()` has grouped output by 'Depression Category'. You can override
## using the `.groups` argument.
```

```
education_levels.df
```

Depression Category	Education Levels
<chr>	<chr>
High	Below upper secondary (total) (%)
High	Tertiary (total) (%)
High	Upper secondary & post-secondary non-tertiary (total) (%)
Low	Tertiary (total) (%)
Low	Upper secondary & post-secondary non-tertiary (total) (%)

5 rows | 1-2 of 3 columns

```
education_levels.df = spread(education_levels.df, `Education Levels`,`Average Affected P
opulation`)
education_levels.df[is.na(education_levels.df)]=0
```

```
education_levels.df
```

Depression Category <chr>	Below upper secondary (total) (%) <dbl>	Tertiary (total) (%) ▸ <dbl>
High	2577926	1787443.5
Low	0	541827.1

2 rows | 1-3 of 4 columns

From the data above, we can see that the data has been transformed into a matrix with 2 rows and 3 columns. The row values are providing us the depression categories which we computed above. On the column end , we have the populations of people affected by depression in each category available. The population values have been computed by utilizing the population sample taken along with the percentages with each category. One interesting thing to notice in the data would would be how “Below Upper Secondary” category has zero low depression cases, which is different from the common perception. This could be because of the limitations of the data for the specific year, additionally the data we are using is mostly from countries based in Europe which could skew the data.

Hypothesis

The purpose of this study was to look into various relationships between the depression percentages and the educational categories available, from the above have been able to see some common trends of these educational classes and how they compare with each other. This assertion was tested using the following statistical hypothesis:

H0: There is no relationship between the a persons education level and depression percentage HA: There is a relationship between the a persons education level and depression percentage

```
education_levels.df = education_levels.df[-c(1)]
rownames(education_levels.df)=c("High","Low")
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
education_levels.df
```

	Below upper secondary (total) (%) <dbl>	Tertiary (total) (%) ▸ <dbl>
1	2577926	1787443.5
2	0	541827.1

2 rows | 1-3 of 4 columns

```
chisq.test(education_levels.df, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  education_levels.df
## X-squared = 956193, df = 2, p-value < 2.2e-16
```

The p value using the chi squared test is 0.00000000000000022~0, hence the the likeliness of the null hypotheses being true is very low. The p value fails at 1%, 5%, 10% significance level tests as well. Therefore we can confidently reject the null hypotheses, meaning the alternative hypotheses is true, that is there is a relationship between the educational levels and the depression rates for the given population.

Employed VS Job Seekers in Education Levels

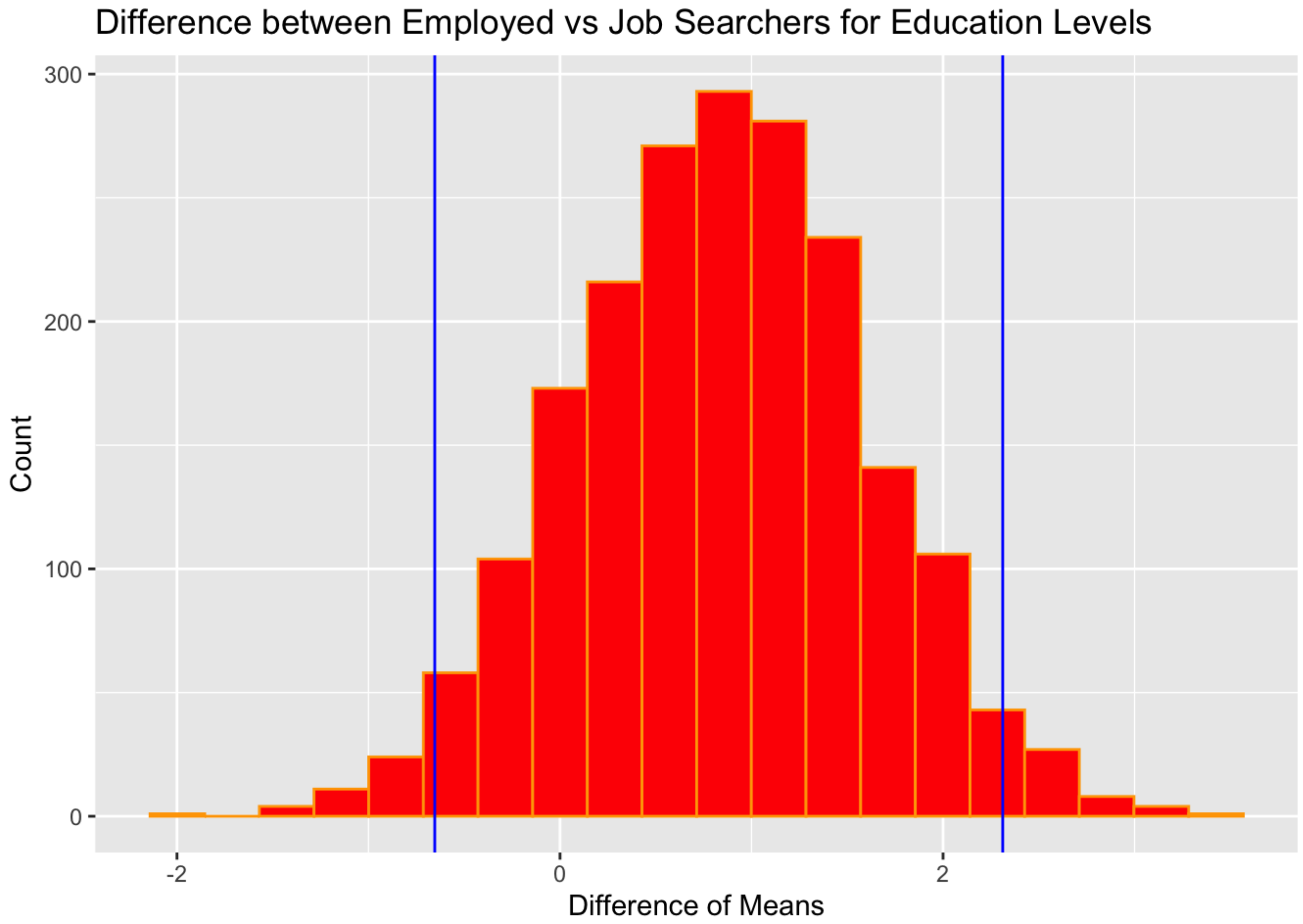
Another part of the data to consider is the available depression statistics for people who are employed with their depression percentages and unemployed poeople who are looking for jobs. This can be computed using bootstrapping over the difference of he means of the employed and the unemployed population.

```
nsamples = 2000
overall_country_active = numeric(nsamples)
overall_country_employed = numeric(nsamples)
education_diff = numeric(nsamples)
for(i in 1:nsamples)
  {
  all_level_active_mean = sample(data_depression_by_education$`All levels (active) (%)`,
26, replace=TRUE)
  all_level_employed_mean = sample(data_depression_by_education$`All levels (employed)
 (%)`, 26, replace=TRUE)
  overall_country_active[i] = favstats(all_level_active_mean)$mean
  overall_country_employed[i] = favstats(all_level_employed_mean)$mean
  education_diff[i] =overall_country_active[i] - overall_country_employed[i]
}
all_level_education.df=data.frame(education_diff)
all_level_education.df
```

education_diff
<dbl>
1.253846154
1.623076923
0.569230769
0.480769231
1.150000000
1.215384615
0.423076923
-0.026923077
1.988461538

```
                                                         education_diff
                                                                  <dbl>
                                                            0.942307692
1-10 of 2,000 rows            Previous  1  2  3  4  5  6 ... 200  Next
```

```
a=quantile(all_level_education.df$education_diff,c(0.025,0.975))
lower = as.numeric(a[1])
upper = as.numeric(a[2])
ggplot(all_level_education.df,aes(x=education_diff)) + geom_histogram(col='orange',fill=
'red',bins=20) +geom_vline(xintercept = lower, col="blue")+geom_vline(xintercept = uppe
r, col="blue") + xlab("Difference of Means") + ylab("Count") + ggtitle("Difference betwe
en Employed vs Job Searchers for Education Levels")
```



From the observation above we can see that the majority of the values are lying between the 0 and 2 difference values between employeed and unemployed population. From this we can infer that that in a randomly sampled data, we are getting the difference values as positive that is in majority of the cases we have the unemployed population with more depression compared to the people who are employed

Form the 95% confidence value we are getting the differential values as the following :- (-0.735 VALUE_{Unemployeed-Employeed} \2.346)

Question 4

Does suicide rates and depressive disorder rates have any correlation ?

When we talk about suicide, one of the most common cause can be attributed to depression, mental disorders and other critical situations. In this study we plan to look how the suicide rates and the disorder rates relate to each other.

```
data_suicide_rates = na.omit(data_suicide_rates)
data_suicide_rates$Year = as.numeric(data_suicide_rates$Year)
```

```
head(data_suicide_rates)
```

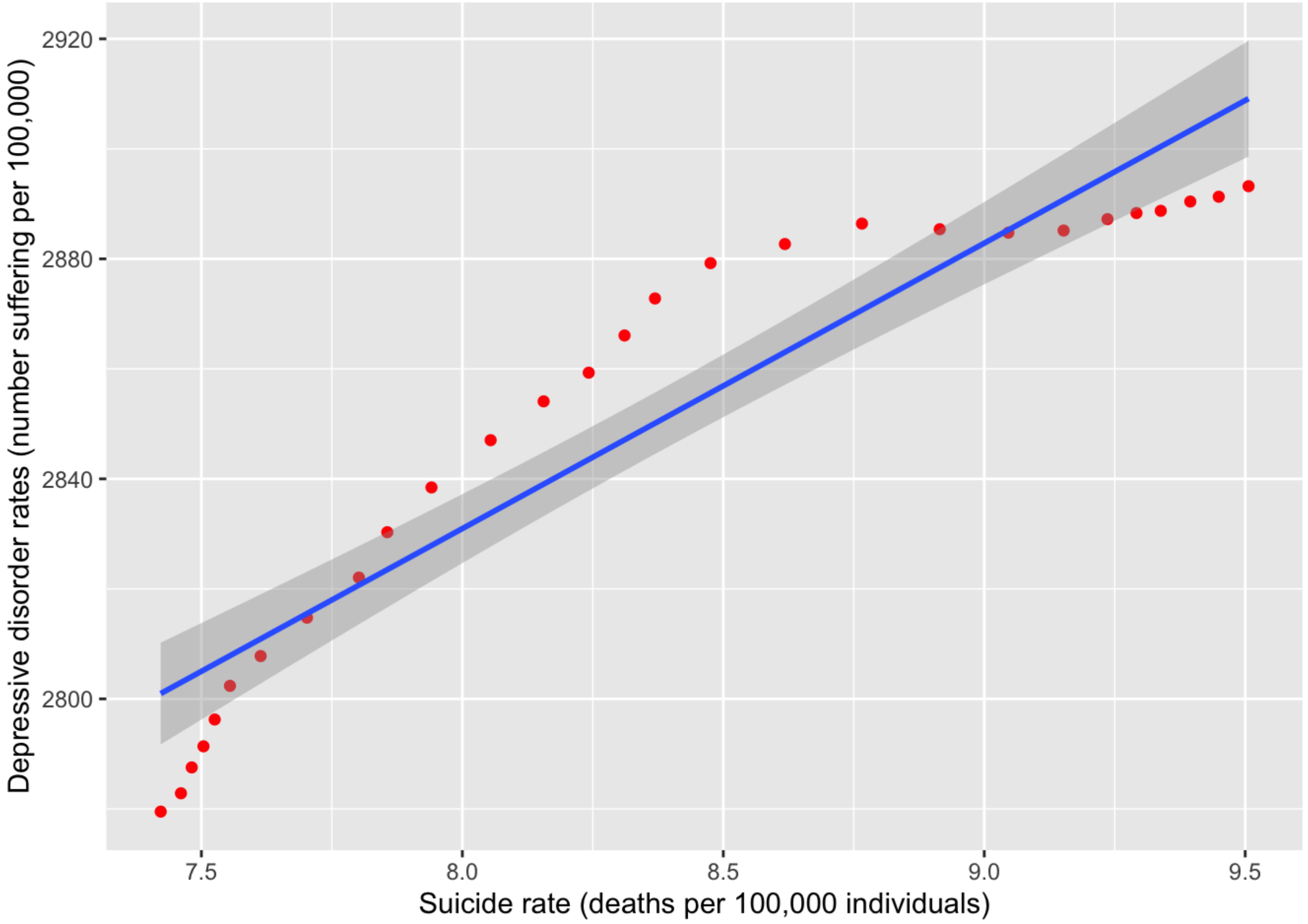
Entity	C…	Year	Suicide rate (deaths per 100,000 individuals)
<chr>	<chr>	<dbl>	<dbl>
Afghanistan	AFG	1990	10.31850
Afghanistan	AFG	1991	10.32701
Afghanistan	AFG	1992	10.27141
Afghanistan	AFG	1993	10.37612
Afghanistan	AFG	1994	10.57591
Afghanistan	AFG	1995	10.68235

6 rows | 1-4 of 6 columns

In the below chunk, we are building the prediction model for suicide rates vs depressive disorders. The prediction data is provided for a specific country which is given in the vecctor x in this case “Vietnam”. The data is fed till the year

```
x = c("Vietnam")

country_filtered_data_suicide_rates = data_suicide_rates %>%
    filter(Entity == x)
ggplot(country_filtered_data_suicide_rates, aes(x = `Suicide rate (deaths per 100,000 in
dividuals)`, y = `Depressive disorder rates (number suffering per 100,000)`)) + geom_poi
nt(col="red") + geom_smooth(method ="lm")
```

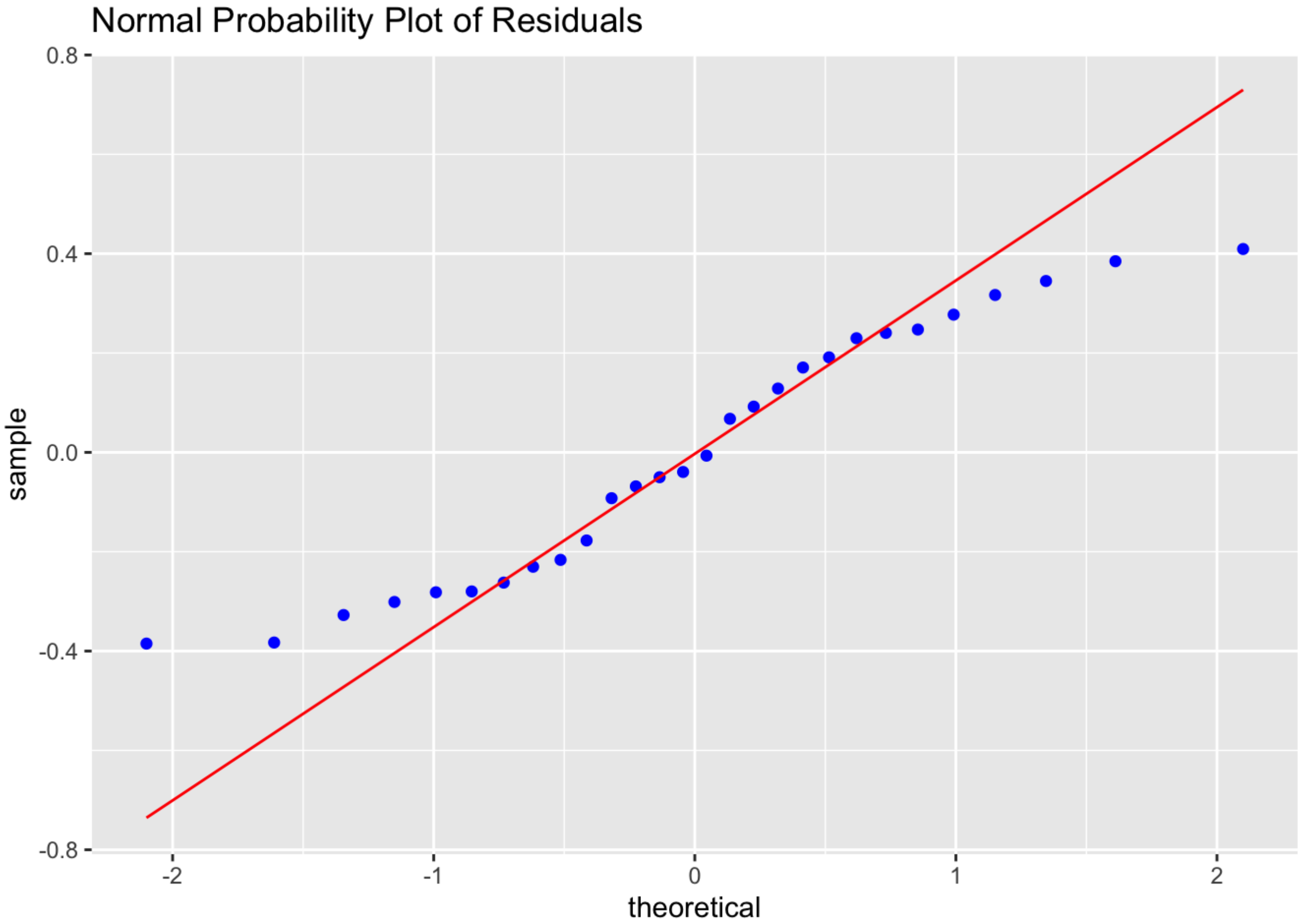
```
## `geom_smooth()` using formula 'y ~ x'
```



```
predictsuicide = lm(`Suicide rate (deaths per 100,000 individuals)`~`Depressive disorder
rates (number suffering per 100,000)`, data=country_filtered_data_suicide_rates)
predictsuicide
```

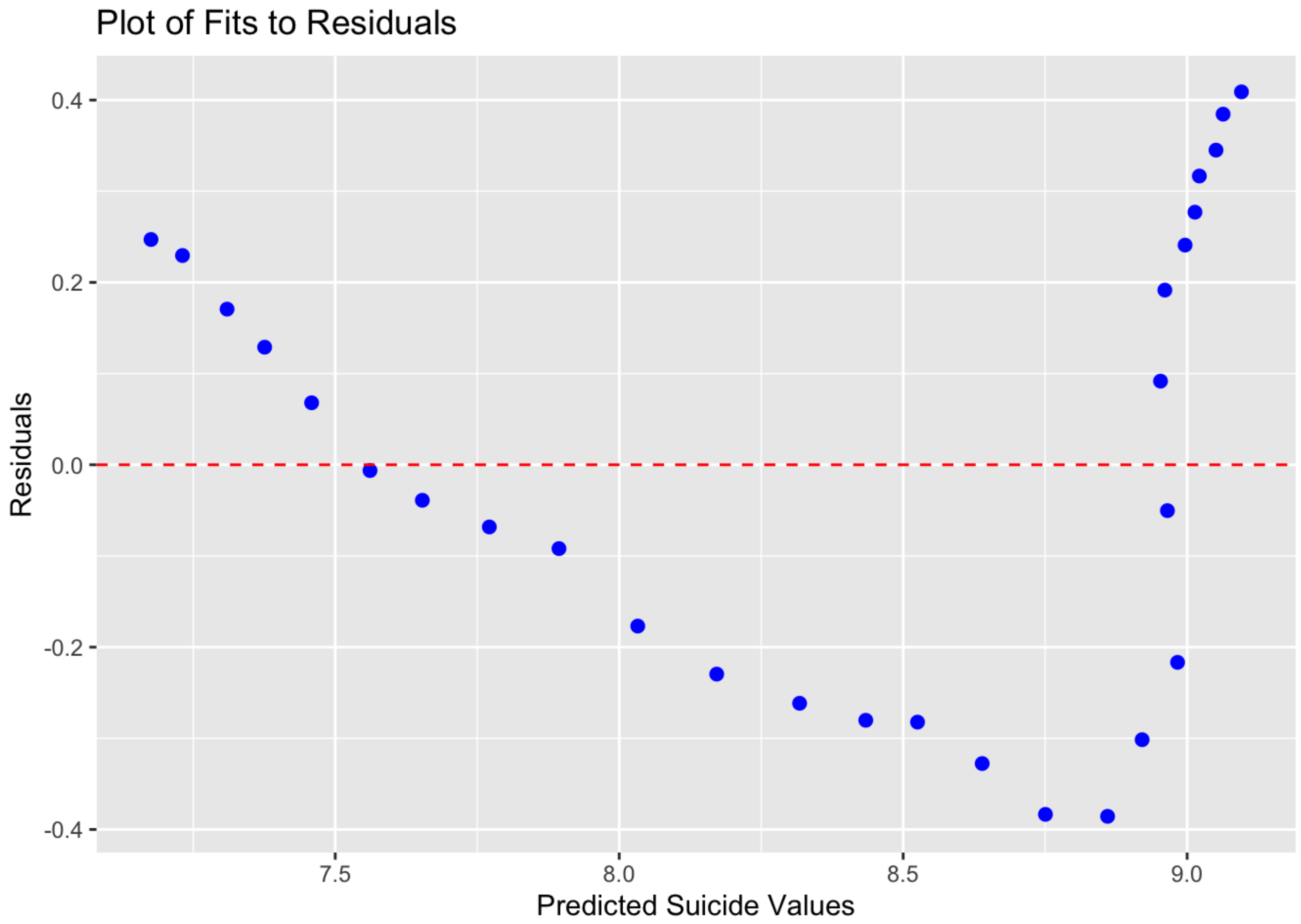
```
##
## Call:
## lm(formula = `Suicide rate (deaths per 100,000 individuals)` ~
##     `Depressive disorder rates (number suffering per 100,000)`,
##     data = country_filtered_data_suicide_rates)
##
## Coefficients:
##                                                 (Intercept)
##                                                    -39.8114
## `Depressive disorder rates (number suffering per 100,000)`
##                                                      0.0169
```

```
predicted.values.suicide= predictsuicide$fitted.values
eissuicide1 = predictsuicide$residuals
suicide.df = data.frame(predicted.values.suicide, eissuicide1)
ggplot(suicide.df, aes(sample = eissuicide1)) +  stat_qq(col='blue') + stat_qqline(col=
'red') + ggtitle("Normal Probability Plot of Residuals")
```



From the plot above the residual points for the prediction model follows the trend of the normal probability plot. Therefore the data that we are working with can be stated as normal.

```
ggplot(suicide.df, aes(x = predicted.values.suicide, y = eissuicide1)) +  geom_point(size=2, col='blue', position="jitter") + xlab("Predicted Suicide Values") + ylab("Residuals") + ggtitle("Plot of Fits to Residuals") + geom_hline(yintercept=0, color="red", linetype="dashed")
```



Hypothesis Development The purpose of this study was to look into relationships between the depression disorder rates and the suicide rates. This assertion was tested using the following statistical hypothesis:

H0:B=0(Y cannot be expressed as a linear function of X) HA:B≠0(Y can be expressed as a linear function of X)

```
summary(aov(predictsuicide))
```

```
##                                                             Df Sum Sq Mean Sq
## `Depressive disorder rates (number suffering per 100,000)`  1 12.488  12.488
## Residuals                                                   26  1.747   0.067
##                                                             F value   Pr(>F)
## `Depressive disorder rates (number suffering per 100,000)`   185.9 2.36e-13 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From using summary(aov(predictsuicide)) we are getting the Fobs value as 185.9 and p value as 0.000000000000236, the degree of freedom as 1. The p value fails to pass the significance level test, hence we can state that the null hypotheses is false and there is a linear relationship between the two values x and y. So, by doing the above test we can confirm our initial perception that the there is a relationship between suicide rates and depressive disorder rates.

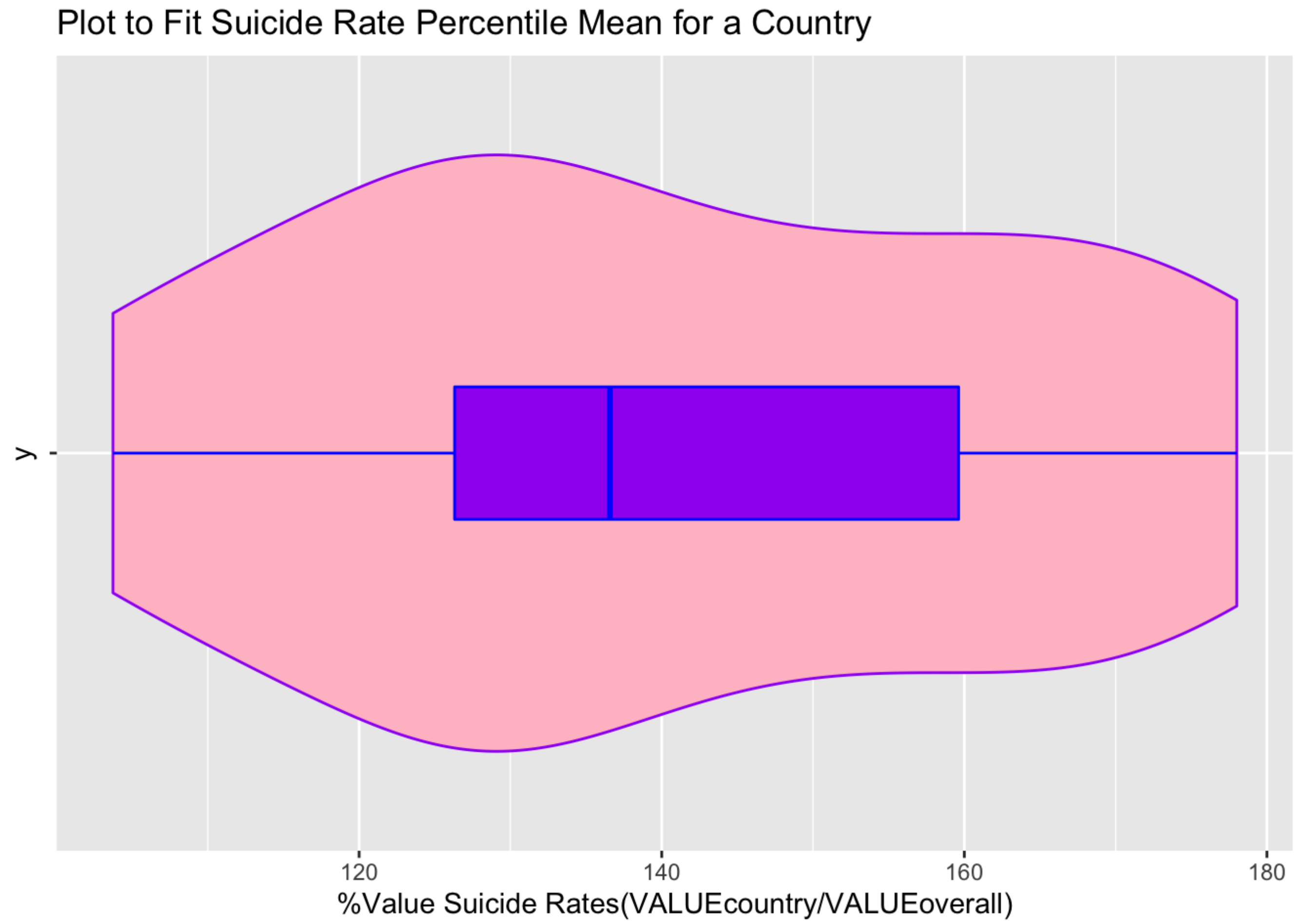
Suicide Rate Data We are working with data 1990 to 2017, there is good possibility that the data collection in the earlier years could be skewed based on how the data was collected and based on the reported values. Therefore we can normalize and visualize the data from 1990 to 2

```
x=c("France")
overall_mean=favstats(data_suicide_rates$`Suicide rate (deaths per 100,000 individuals)`)$mean

country_filtered_data_suicide_rates = data_suicide_rates %>%
    filter(Entity == x)

country_data = country_filtered_data_suicide_rates$`Suicide rate (deaths per 100,000 individuals)`
mean_percentile_country = (country_data/overall_mean)*100
country.df=data.frame(mean_percentile_country)
country.df
```

mean_percentile_country <dbl>
178.0087
176.0792
174.0150
173.6700
168.2260
164.5971
162.2983
158.7304
156.9667
152.0468

1-10 of 28 rows Previous **1** 2 3 Next

```
country.df %>%
  ggplot(aes(x=mean_percentile_country, y=""))+geom_violin(col="purple",fill="pink") +xlab("%Value Suicide Rates(VALUEcountry/VALUEoverall)") + geom_boxplot(width = 0.2,col="blue",fill="purple") + ggtitle("Plot to Fit Suicide Rate Percentile Mean for a Country")
```



Depressive Disorder Data Similar to the case of Suicide Rates, we will normalize the depressive data . The calculations below lead the graph for one country over the period of 1990 to 2017. In this case the country choosen is “France”, this can be modified by changing the value in the x vector.

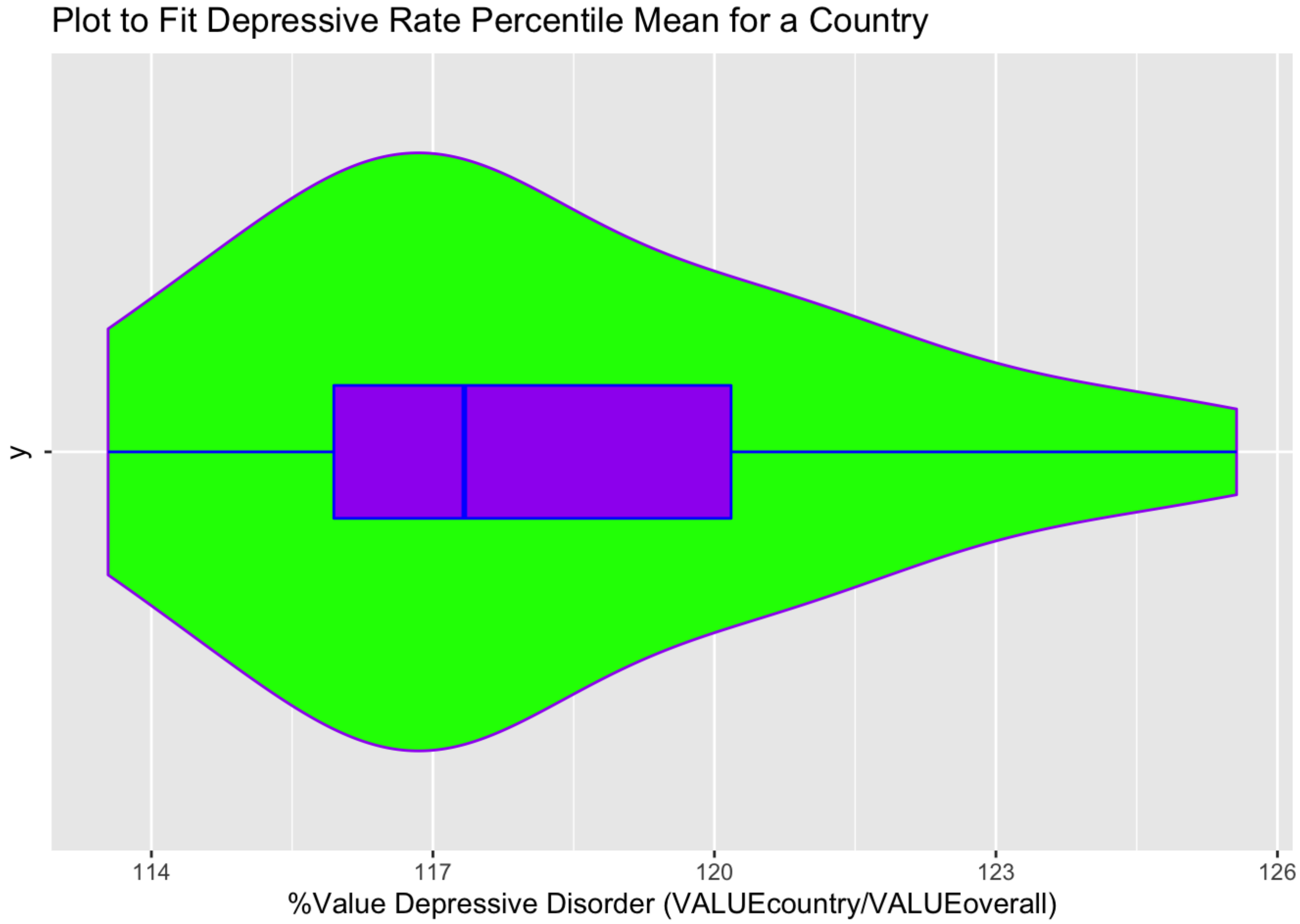
```
x=c("France")
overall_mean=favstats(data_suicide_rates$`Depressive disorder rates (number suffering pe
r 100,000)`)$mean

country_filtered_data_depressive_rates = data_suicide_rates %>%
    filter(Entity == x)

country_data = country_filtered_data_depressive_rates$`Depressive disorder rates (number
suffering per 100,000)`
mean_percentile_country = (country_data/overall_mean)*100
countryDepression.df=data.frame(mean_percentile_country)
countryDepression.df
```

mean_percentile_country <dbl>
125.5646
124.1799
122.9416

mean_percentile_country
<dbl>
121.9259
121.2022
120.8565
120.5413
120.0561
119.4534
119.0120

1-10 of 28 rows Previous **1** 2 3 Next

```
countryDepression.df %>%
  ggplot(aes(x=mean_percentile_country, y=""))+geom_violin(col="purple",fill="green") +x
lab("%Value Depressive Disorder (VALUEcountry/VALUEoverall)") + geom_boxplot(width = 0.2
,col="blue",fill="purple") + ggtitle("Plot to Fit Depressive Rate Percentile Mean for a
 Country")
```



From the above observations we have proved a linear relationship between the suicide rates and the depressive disorder rates . Furthermore to better visualize the data we have computed the percentile value for a country from the data available from 1990 to 2017.

Question 5

What kind of relationship patterns do we see between mental disorders?

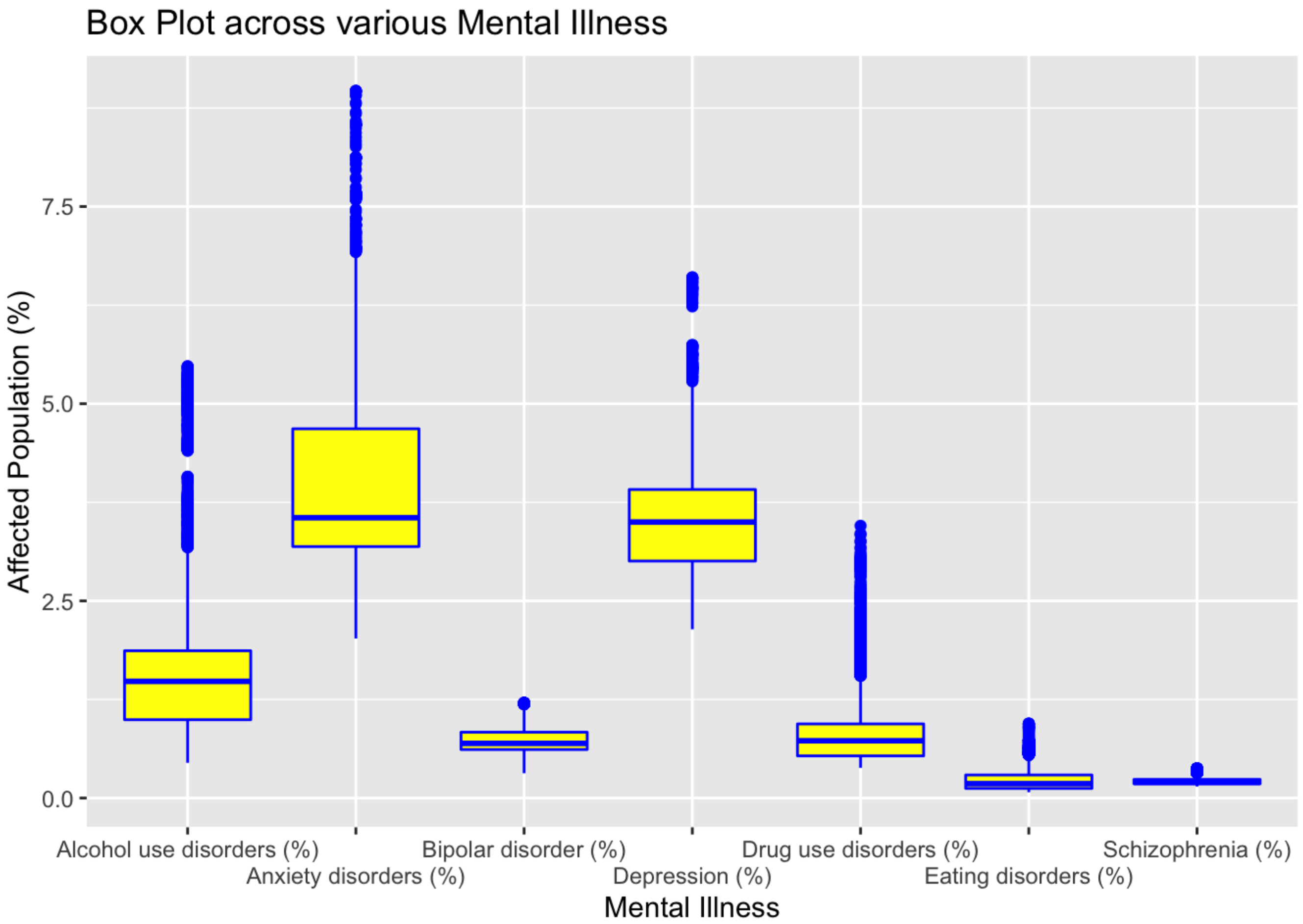
This dataset includes the pervalance of seven different kinds of Disorders among the population from various countries across the globe from the years 1990-2017. In this dataset the disorder among population is brokendown into percentages. We will be doin statistical analysis of the differences across different types of mental Illness to reveal any patterns.

Hypothesis

For this need to form a hypothesis

“H0: Anxiety Percentages and Depression percentages are equal to each other across all the continets” “Ha: Anxiety Percentages are higher than depression rates across all the contients”

```
box_plot_data = gather(data_disease_type,"Disease","Affected_Percentage",4:10)
ggplot(data = box_plot_data, aes(x = Disease,y = Affected_Percentage)) + geom_boxplot(co
l = "blue", fill = "yellow") + scale_x_discrete(guide = guide_axis(n.dodge=2)) + xlab("M
ental Illness") + ylab("Affected Population (%)") + ggtitle("Box Plot across various Men
tal Illness")
```



```
cat(" Statistical Summary for the disease Schizophrenia")
```

```
##  Statistical Summary for the disease Schizophrenia
```

```
favstats(data_disease_type$`Schizophrenia (%)`)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.1469018	0.1815296	0.1995631	0.2363652	0.3751096	0.2116436	0.0442528	6468	0

1 row

```
cat(" Statistical Summary for Bipolar disorder")
```

```
##  Statistical Summary for Bipolar disorder
```

```
favstats(data_disease_type$`Bipolar disorder (%)`)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.3145345	0.6155322	0.6931345	0.8350626	1.206597	0.7191452	0.1715886	6468	0

1 row

```
cat(" Statistical Summary for Anxiety disorder")
```

```
##  Statistical Summary for Anxiety disorder
```

```
favstats(data_disease_type$`Anxiety disorders (%)`)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
2.023393	3.188824	3.554373	4.682163	8.96733	3.989921	1.167526	6468	0

1 row

```
cat(" Statistical Summary for Depression disorder")
```

```
##  Statistical Summary for Depression disorder
```

```
favstats(data_disease_type$`Eating disorders (%)`)
```

min	Q1	median	Q3	max	mean	sd	n	missir
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<in
0.07390753	0.1223872	0.1825245	0.2926665	0.9439906	0.2399984	0.1581412	6468	

1 row

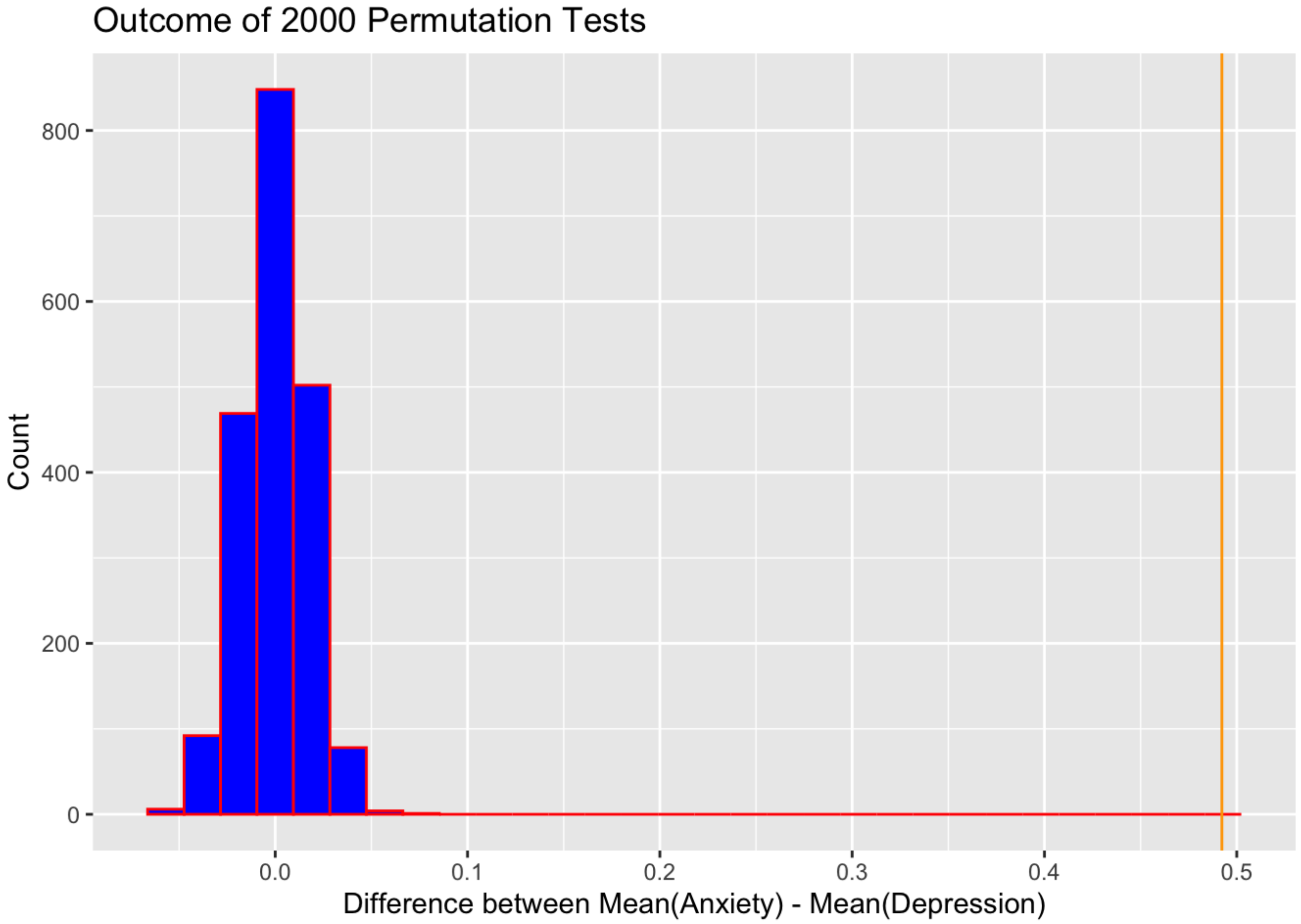
```
# Importing the anxiety and depression columns from the dataset to test the hypothesis
anxiety = data_disease_type$`Anxiety disorders (%)`
depression = data_disease_type$`Depression (%)`

percent = c(anxiety,depression)
name_disease = c(rep("A", length(anxiety)), rep("D", length(depression)))
disease.data = data.frame(name_disease, percent)

# Finding the difference in means betwee the disorders
disdiff = mean(~percent, data=filter(disease.data, name_disease=="A"))  - mean(~percent,
data=filter(disease.data, name_disease=="D"))
```

```
nperms <- 2000
perm.outcome <- numeric(nperms)
mean.A <- numeric(nperms)
mean.D <- numeric(nperms)
for(i in 1:nperms){
  index <- sample(12936, 6468, replace=FALSE)
  mean.A[i] <- mean(disease.data$percent[index])
  mean.D[i] <- mean(disease.data$percent[-index])
  perm.outcome[i] <- mean.A[i] - mean.D[i]
}
```

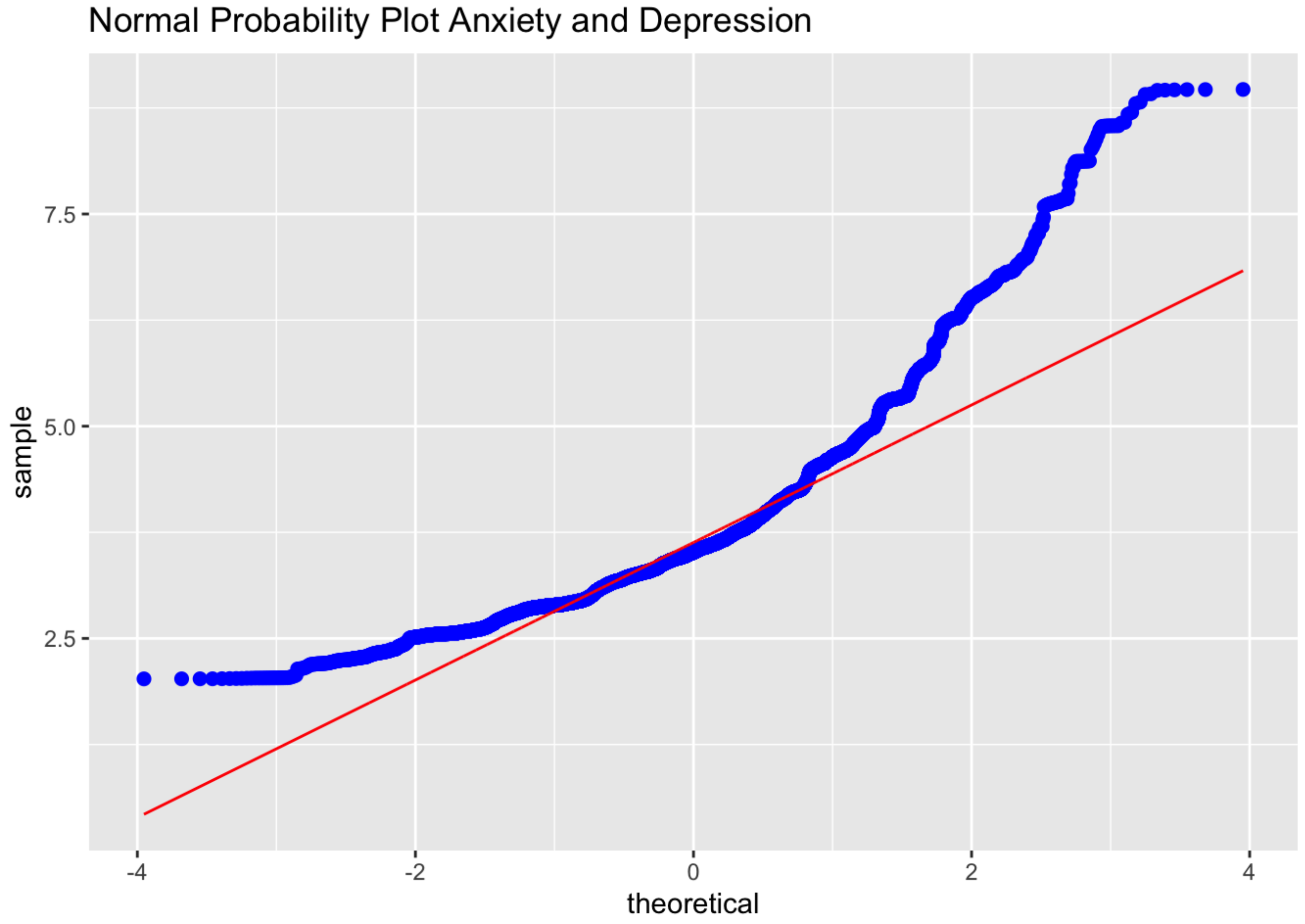
```
permutationtest1 <- data.frame(mean.A, mean.D, perm.outcome)
```

```
ggplot(permutationtest1, aes(x = perm.outcome)) + geom_histogram(col="red", fill="blue")
+ xlab("Difference between Mean(Anxiety) - Mean(Depression)") + ylab("Count") + ggtitle(
"Outcome of 2000 Permutation Tests") + geom_vline(xintercept = disdiff, col="orange")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data=disease.data, aes(sample = percent)) + stat_qq(size=2, col="blue") + stat_qq
line(col="red") + ggtitle("Normal Probability Plot Anxiety and Depression")
```



After testing to see if Anxiety and Depression follows a normal distribution, we see that it does follows a normal distribution, since most of the points are along the straight line, as a result we can do a t-test find the p value and the confidence interval

```
ttest_pval1 = t.test(~percent|name_disease, conf.level=0.95, alternative = "greater", va
r.equal=FALSE, disease.data)
ttest_pval1
```

```
##
##  Welch Two Sample t-test
##
## data:  percent by name_disease
## t = 29.564, df = 10179, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group A and group D is great
er than 0
## 95 percent confidence interval:
##  0.4648769       Inf
## sample estimates:
## mean in group A mean in group D
##        3.989921        3.497654
```

Because our p-value is <2.2e-16 which is significantly low, so we reject our null hypothesis. As a result we can conclude that Anxiety rates are higher than depression rates

#Conclusion

At an overall level, for the gender based statistics, the % affected population for males is higher over the years as compared to females. Along with this, there exists a linear relationship between the two groups with a high correlation index.

When considering the age-wise statistical tests performed above, we infer that with the increase in ages, a person falling into different age groups, the depression level increases. Also, the mean of the depression affected percentage population lies between the range 3.258 and 3.300 percent

Overall from the results above we can see that there is a relationship between a persons education level and the depression percentages in the provided countries. Further we have also observed that the a person from any one of these education groups are more likely to be depressed if they are actively looking for a job compared to a person who is already employed. The above has been confirmed from both our initial inference and statistical analysis.

After doing the above statistical analysis on the dataset that includes various Illnesses, we have concluded that across all the continents Anxiety and Depression are the leading mental health concern for the population. Based on the results of our permutation test on finding the difference between Anxiety and Depression, we found out that no matter the sample across the world Anxiety is always going to be higher than depression since our confidence interval is always positive. We were also curious and wanted to predict the 2017 depression prevalence in Vietnam. Since France is the number 1 tourist destination in the world with over 90 million visitors in 2019, we wanted to analyze if the depression rates were lower compared to the mean of the world. And as we hoped that France being a well-known tourist destination, the depression rates were lower compared to the mean average of depression in the world.