DATA 602 Final Project

Group 8

2022-10-16

Contents

Libraries

```
library(mosaic)
library(dplyr)
library(ggplot2)
```

Introduction

Mental health constitutes our emotions and mental, and social well-being. It can greatly affect how we feel, think, behave, and deal with stress. According to Canadian Mental Health Association (2022), by the age of 40, about 50% of the population will have or had experienced mental illness at some point in their life. There are different types of mental disorders, including bipolar, eating, personality, post-traumatic stress disorders, etc (WHO, 2022).

However, depression is one of the most common mental disorders as it has a significant public health implication. Depression is a complex mental disorder. It is characterized by consistent feelings of sadness that lasts for a long period of time that negatively affects their daily life. Other characteristics of depression are loss of pleasure in everyday activities or those once enjoyed (WHO, 2022), feelings of emptiness and hopelessness, and even sleep and eating disturbances. Depression is a major source of distress for individuals and it can be considered as a threat to human health (Razzak et al., 2019). It affects about 5% of the population worldwide (WHO, 2022). There is a 20% lifetime chance of developing depression in the general population (Brigitta, 2002). Overall, women experience depression more likely than men, with a ratio of 5:2 (Brigitta, 2002). Although depression can develop at any age, older individuals are at more risk of experiencing depression (National Council on Aging, Inc, 2022).

Depression has various causes, including but not limited to age, major life events (i.e., trauma, abuse), genetics, brain chemistry, unhealthy lifestyle, etc (Bruce, 2021). Depression can result in functional disabilities, which leads to a significant decrease in people's quality of life (Cho et. al., 2019). However, it is important to consider other risk factors, such as environment, education, employment, and etc. Furthermore, depression can highly increase one's risk of suicidal behavior. Around 50% of individuals attempting suicide have the criteria for a depression diagnosis (Cummins et. al., 2015).

Since depression is a heterogeneous disorder, it is difficult to determine the exact factor that has brought it on in patients. It has fixed symptoms, however, there are a variety of different analyses that show little symptom overlap (University of Amsterdam, 2016), making it difficult to determine a single or primary root cause, as well as hindering the treatment of depressed individuals. Moreover, there are several risk factors that increase the chance of an individual becoming clinically depressed. Nowadays, depression rates, as well as many other mental health disorder rates are increasing in societies due to chronic daily stress, improper coping mechanisms (i.e., substance abuse), unhealthy lifestyles, isolation, and improvements in diagnosis (Hidaka, 2012). In addition, our society has increased accessibility to mental health resources significantly compared to even 25 years ago, leading to new mental illness criterion being brought forward, as well as an in increase reports of mental health disorders. Therefore, many researchers have tried to study this disorder more thoroughly. Our project aims to use data analytics and visualizations to determine the prevalence of depression among different demographics and its correlation with factors such as substance use and suicide.

Guiding Questions

Our guiding questions for the project are as below:

1. Which demographics (i.e., gender, age groups) are more susceptible to depression?

a) `Are these different for different parts of the world or change over time?`

2. Are there different behaviours (i.e., drug use) that correlate with the prevalence of mental disorders?

a) `Does this change when separated by gender and age groups?`

b) `Does this change for different parts of the world or change over time?`

3. Which attribute from our dataset primarily explains the variance in suicide rates?

Dataset

The dataset being used is Mental Health Depression Disorder Data, which is structured and in tabular format. The dataset source is from Our World in Data, which is a project of a United Kingdom non-profit organization known as the Global Change Data Lab. This data was found on a database that is intended for public use, and thus the team is permitted to make use of this dataset. The dataset has continuously been updated by its relative authors and contributors over the course of 2020 – 2021; this indicates that the dataset is quite recent and is not antiquated.

It is important to keep in mind that there is a widespread issue when it comes to examining this dataset of under reporting accurate and representative data for mental illnesses. There are reasons for why this is, including: outdated measurement criterion, inaccessibility to health and mental health services and other social supports (especially in underdeveloped countries), inaccurate measurements, high cost associated with mental health services, and negative social stigma associated with mental illness.

The majority of data included in this dataset is from the Institute for Health Metrics and Evaluation which is an independent global health research centre at the University of Washington (Institute for Health Metrics and Evaluation, 2022), and reports global data from 1990 to 2016. The dataset uses entries from the Global Burden of Disease (GBD) which provides worldwide data on mental health and substance abuse disorders (GBD Results, 2019), whereas organizations such as the World Health Organization (WHO) only publish statistics related to depression disorder (Ritchie & Roser, 2018). This is advantageous for the purposes of this project as it will allow the team to manipulate a larger amount of data and examine various distinct relations between mental disorders other than mainstream topics, such as depression.

Summary of Data set:

Table 1: Mental Disorder and Substance Abuse

- Columns: 11
- Rows: 26
- Data Reported from the Years: 1990 - 2017
- Number of Countries Reported on: 231
- **Additional Notes:** Notable columns include the country, year, percentages of various mental disorders (ie: anxiety disorders, bipolar disorder, drug use disorders, etc).

Table 2: Depression by Level of Education

- Columns: 15
- Rows: 6468
- Data Reported from the Years: 2014
- Number of Countries Reported on: 231
- **Additional Notes:** Our team has opted to **not** make use this table for our project as the column names do not make it clear what the values represent.

Table 3: Prevalence of Depression by Age

- Columns: 13
- Rows: 6468
- Data Reported from the Years: 1990 - 2017
- Number of Countries Reported on: 231
- **Additional Notes:** Notable columns include the year, percentages for age groups broken down into sub-groups from 10-14 years old to all the way up to 70+ years old, and also includes a column for all ages.

Table 4: Prevalence of Depression in Males and Females

- Columns: 6
- Rows: 47807
- Data Reported from the Years: 1800 - 2019
- Number of Countries Reported on: 274
- **Additional Notes:** There is a significant amount of missing data from earlier years. Noteworthy columns include the country, year, the prevalence of depression in percentages, and the population of the country.

Table 5: Prevalence of Depression vs. Suicide Rates

- Columns: 6
- Rows: 47807
- Data Reported from the Years: 1800 - 2019
- Number of Countries Reported on: 274
- **Additional Notes:** There is a significant amount of missing data from earlier years. Notable columns include the country, year, suicide rate per 10,000 individuals, depressive disorder rate per 100,000 individuals, and the population of the country.

Table 6: Prevalence of Depression vs. Suicide Rates

- Columns: 4
- Rows: 6568
- Data Reported from the Years: 1990 - 2017
- Number of Countries Reported on: 231
- **Additional Notes:** Notable columns include the country, year, prevalence of depressive disorders in both sexes for people of all ages.

Data Cleaning

The team has:

- **Checked for NAs:** Handling missing data (NAs) have been dealt with depending on the proprtion missing from the data. NAs can be removed, or replaced with the appropriate statistic.
- **Checked for duplicates in the data:** Duplicate observations will be removed where they should not be, using the Pandas library. For example, in the context of the selected dataset, this means a table that has one observation per country per year.
- **Ensured consistent formatting in the data:** Using the Pandas library, formatting must be consistent across variables. In the context of the selected dataset, this means percentages and decimals.

Data Wrangling

The team has:

- **Merge tables:** The team utilized the Pandas library for merging tables in order to make analysis easier. Each row has a unique key that identifies it, and each table has a consistent amount of rows.
- **Create new columns:** The team had been interested in a calculation from two existing columns into a new column that the team can then plot on a graph using NumPy and Python.

Question 1

1. Which demographics (i.e., gender, age groups) are more susceptible to depression?

a) `Are these different for different parts of the world or change over time?`

Question 1 Background:

We are interested in seeing if certain demographics including gender and age are more susceptible to depression.

We will try to answer this by first looking at the linear relationships between variables in our data set and depression disorders through correlation and visualizations. We will construct line graphs as our data is time-series, as well as include box and violin plots. We will also investigate where in the world depression is the highest based on gender and age, and plot histograms.

We will be using data from 1990 to 2019.

```
# Load the data set
data <- read.csv("Mental Health Merged.csv")
```

Adding "Development" column. Developed countries are defined as countries with HDI > 0.72.

```
data$Development = ifelse(data$hdi2019>0.72, "Developed", "Undeveloped")
data_developed = data[(data$Development == "Developed"), ]
data_undeveloped = data[(data$Development == "Undeveloped"), ]
```

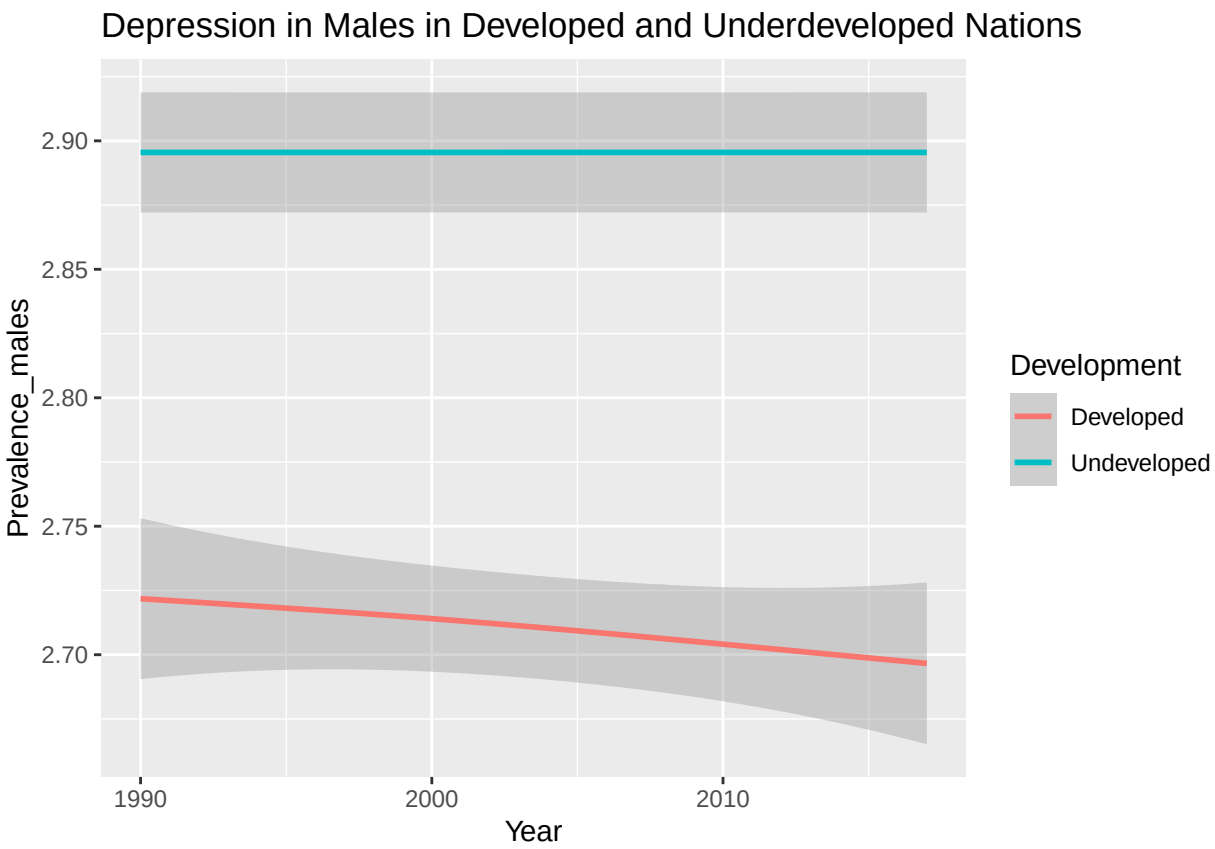
Depression Prevalence Between Male & Female Genders

Create and display data frame for Gender & Depression Data

```
data_gender = data[, c("Country", "Year", "Prevalence_males", "Prevalence_females", "Prevalence_females
```

Figure 1.1: Line Graph for the Prevalence of Depression in Males, sorted by Developed and Underdeveloped Nations

```
ggplot(data_gender, aes(x = Year, y = Prevalence_males, color = Development)) + geom_smooth() + ggtitle
```



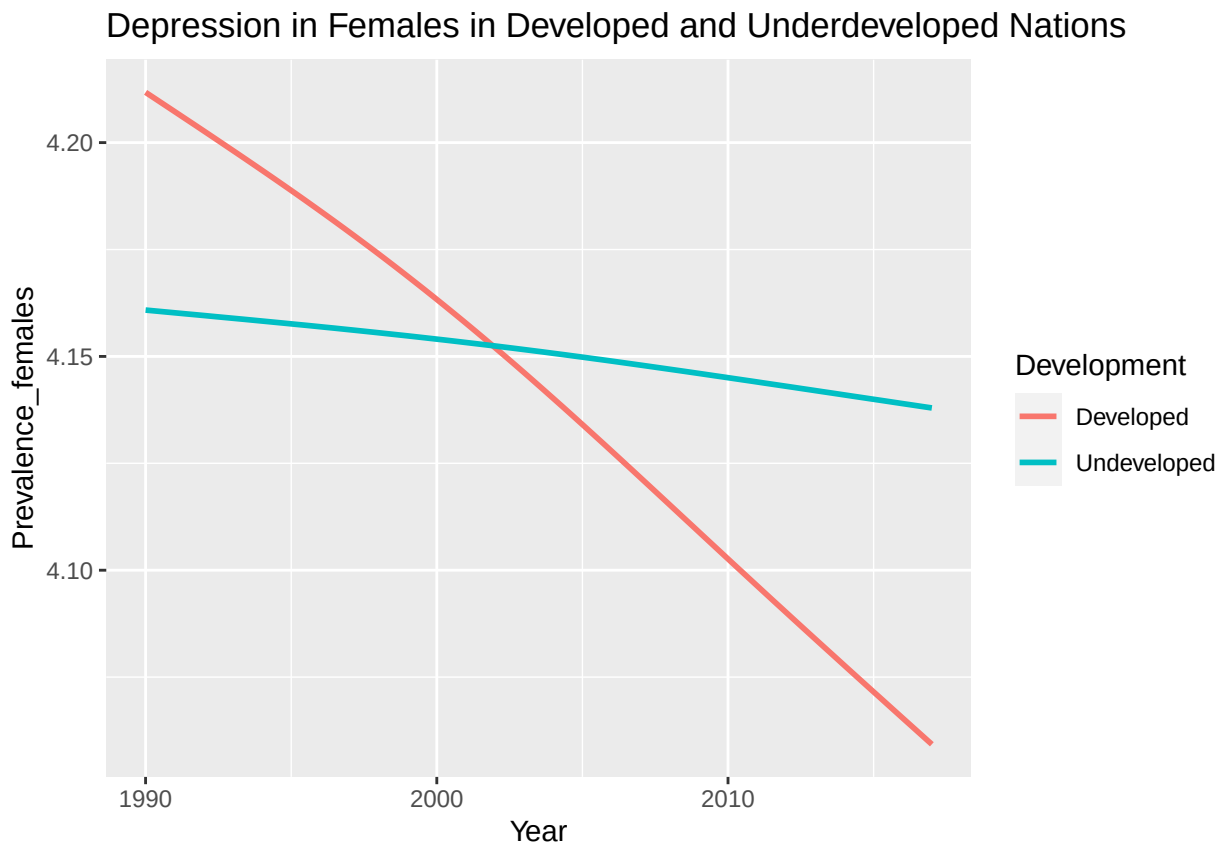
The above plot shows that the prevalence of depression in males is consistently higher in developed nations compared to undeveloped nations. This may be explained by a variety of different reasons:

- Highly developed nations on average have more accessibility to healthcare and mental health resources.
- The screening and criterion for mental health disorders are more accurate in developed nations.
- There has been a push to eliminate negative social stigma associated with mental health in developed nations and create more awareness. As a result, the population may be more encouraged to seek out mental health resources than compared to underdeveloped nations.
- Less negative social stigma towards mental illness in developed nations may lead to more reported cases of mental illness.

Figure 1.2: Line Graph for the Prevalence of Depression in Females, sorted by Developed and Underdeveloped Nations

```
ggplot(data_gender, aes(x = Year, y = Prevalence_females, color = Development)) + geom_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



On average, it can be concluded from both the above plot (**Figure 1.2**), as well as the previous (**Figure 1.1**) that depression rates seem to be consistently higher in females than in males throughout time (1990 - 2019). This could possibly be attributed to men being socialized differently than women, and thus resulting in men being discouraged from seeking out help with their mental health.

It would be interesting to how these rates compared to suicide rates between men and women, however this is not provided in the dataset.

Moreover, it can also be observed from the above plot, (**Figure 1.2**) that depression rates appear to be decreasing at a constant rate in developed nations, and the same is true (although at a lesser extent) for underdeveloped nations.

Figure 1.3: Scatter Plot for the Prevalence of Depression in Males, sorted by Developed and Underdeveloped Nations

```
ggplot(data_gender, aes(x = Year, y = Prevalence_males, color = Development)) + geom_point(alpha=0.5) +
```

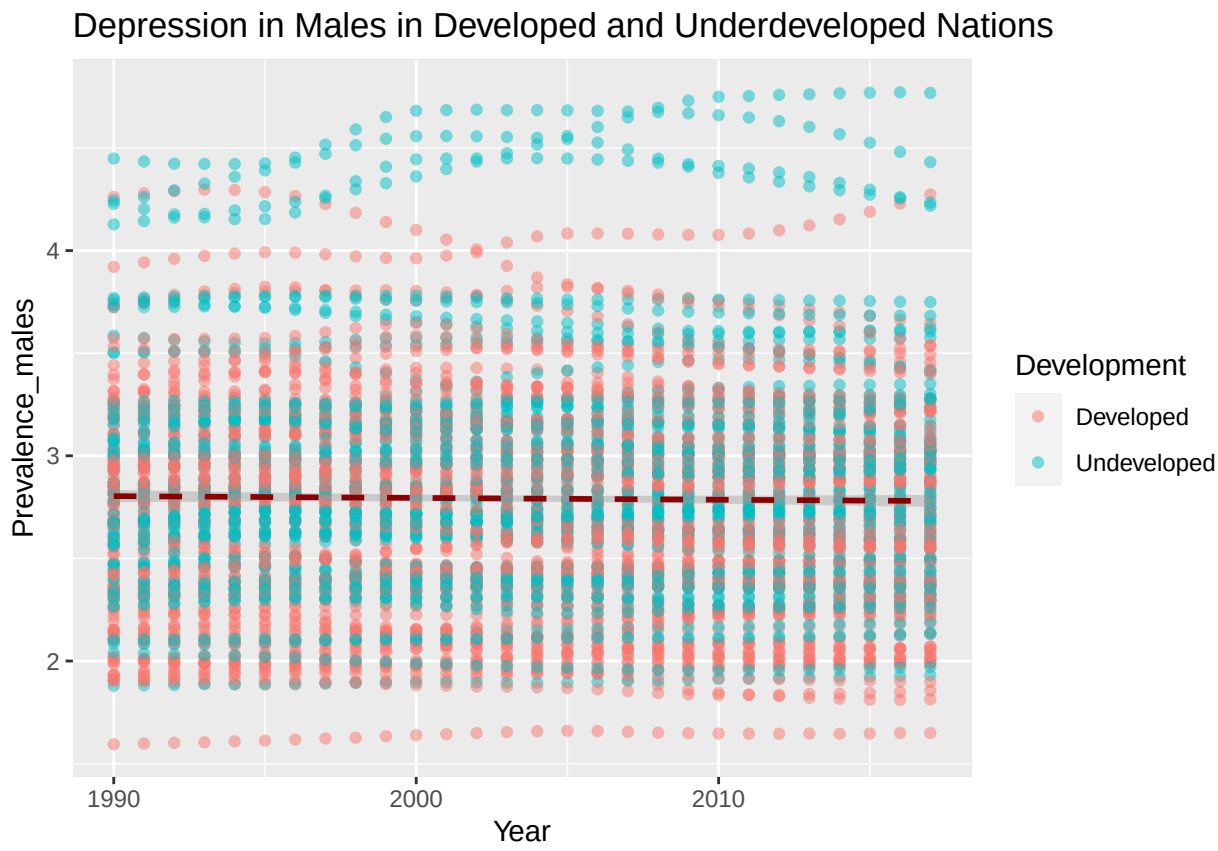
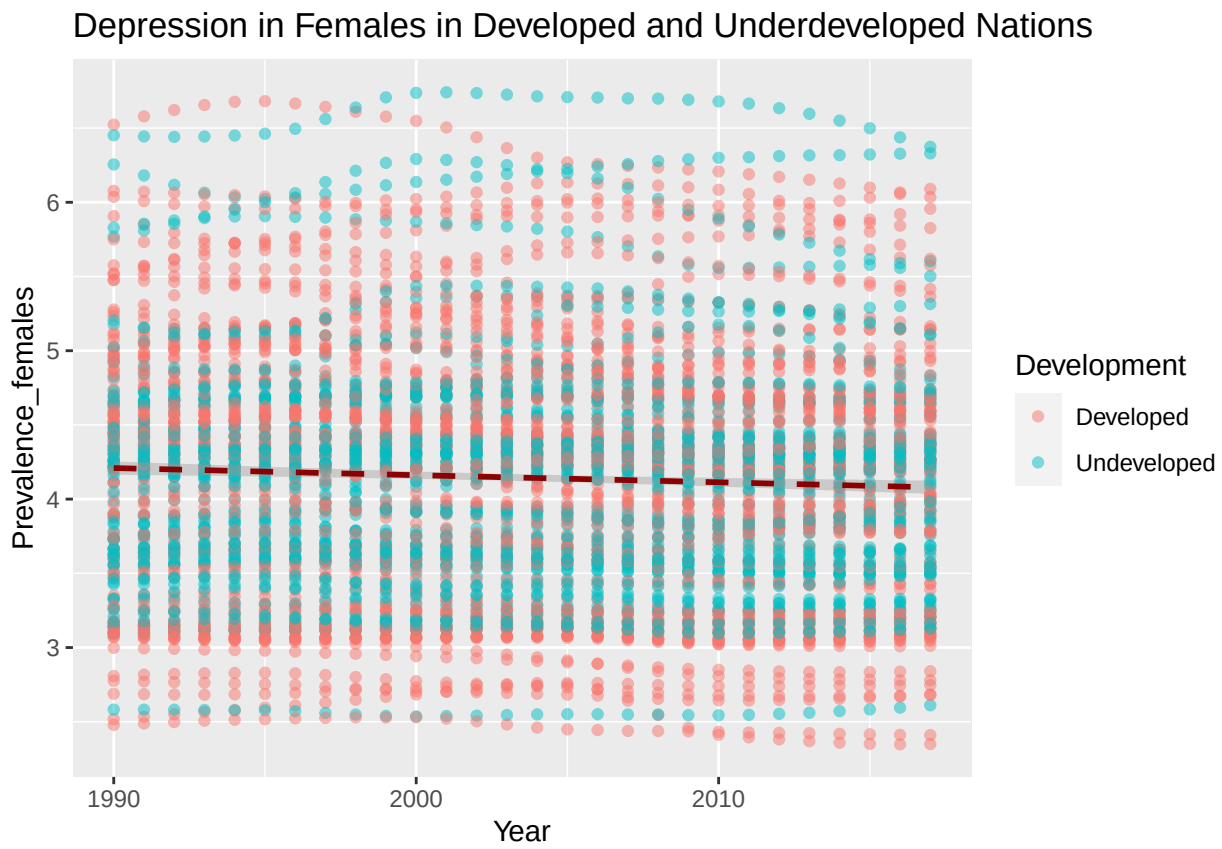


Figure 1.4: Scatter Plot for the Prevalence of Depression in Females, sorted by Developed and Underdeveloped Nations

```
ggplot(data_gender, aes(x = Year, y = Prevalence_females, color = Development)) + geom_point(alpha=0.5)
```



The scatter plots (Figure 1.3 and Figure 1.4) are generally very similar, with some notable key points to be made: * This is clearly not a normal distribution, and no correlation is being made between the x-variable (Year) and the Y-variable (Depression Prevalence in Males or Depression Prevalence in Females). Rather, we can observe the depression rates between Developed & Underdeveloped countries. * The smoother in Figure 1.3 is lower than the smoother in Figure 1.4. The smoother helps to visualize the patterns in the presence of over plotting. The result from the smoother lines remains consistent with the previously ascertained fact that depression prevalence is higher on average in females than in males.

Figure 1.5: Box and Violin Plot for the Prevalence of Depression in Males, sorted by Developed and Underdeveloped Nations

```
ggplot(data_gender, aes(x = Year, y = Prevalence_males, color = Development)) + geom_violin() + geom_bo
```

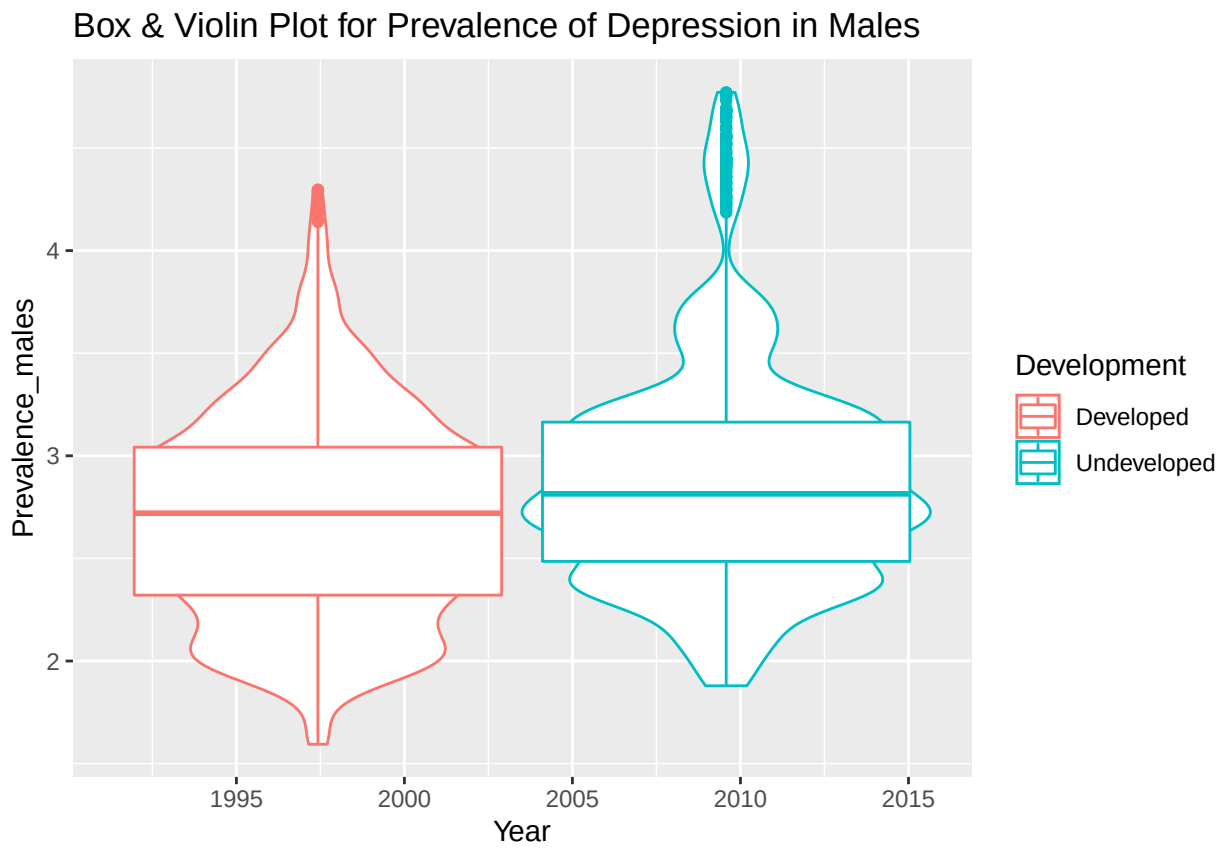
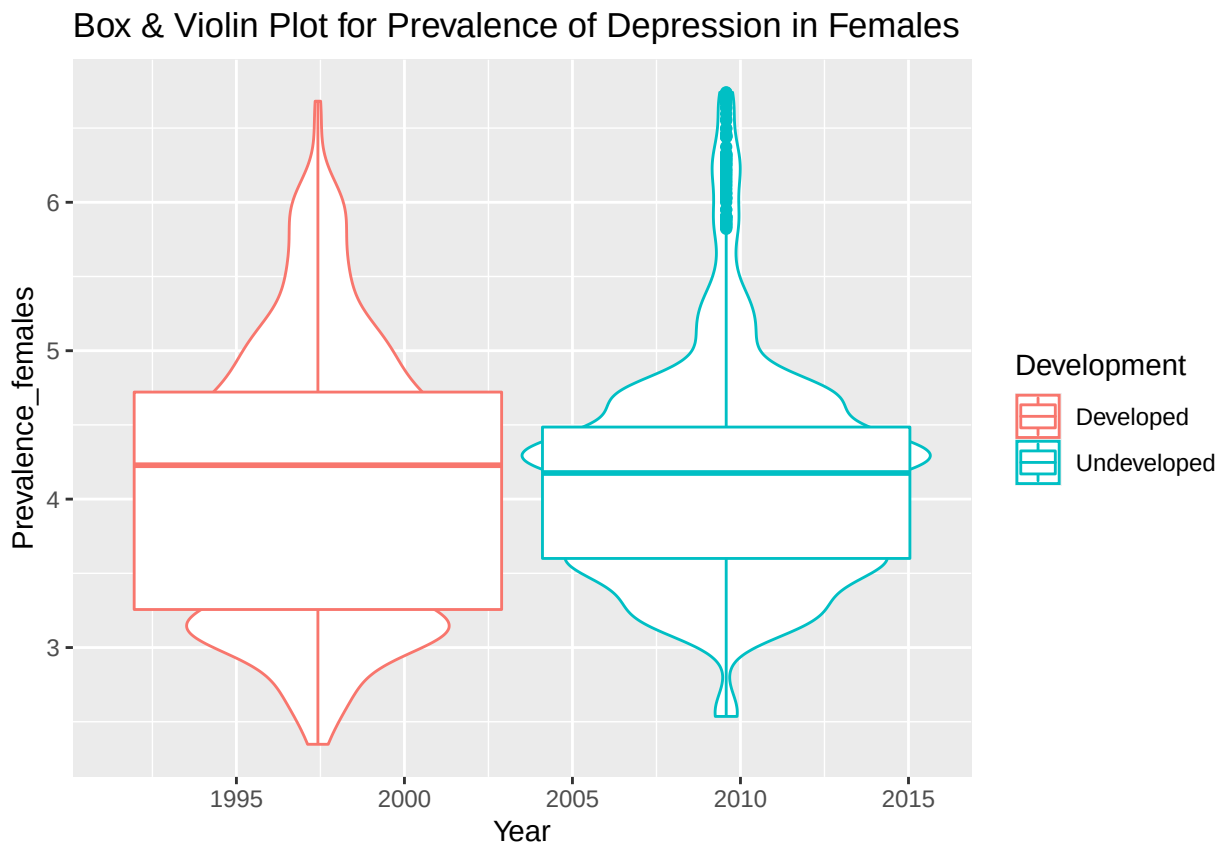


Figure 1.6: Box and Violin Plot for the Prevalence of Depression in Females, sorted by Developed and Underdeveloped Nations

```
ggplot(data_gender, aes(x = Year, y = Prevalence_females, color = Development)) + geom_violin() + geom_
```



The above plots (Figure 1.5 and Figure 1.6) show clearly the:

- 1.5x Interquartile Range
- Interquartile Range
- Median

This data allows us to better visualize the distribution of the numeric data for the prevalence of depression between males and females in developed and underdeveloped nations. It allows us to view the peaks in data. With the addition of a box-plot overlaying the violin plot, we are able to also see the summary statistics as well as the density of the variable (depression prevalence).

In the context of the data being examined, Figure 1.5 shows that male depression rates in underdeveloped countries have a lot of outliers, compared to those of developed countries. Similarly, Figure 1.6 also shows that female depression rates in underdeveloped nations have many more outliers compared to those of developed nations. There seem to be practically no outliers for females in developed countries, meaning the depression rate is pretty standard across the years of 1990 - 2019. Additionally, a lot of the data points are stored in the third quartile, median, and first quartile.

Interestingly, when we observe Figure 1.5 and Figure 1.6 together, we can see that the overall shape and distribution of the tips are similar for males in developed countries and females in underdeveloped countries.

Figure 1.7: Histogram with Density Plot and Mean Line for the Prevalence of Depression in Males

```
ggplot(data_gender, aes(x=Prevalence_males)) + geom_histogram(aes(y=..density..), colour="black", fill=
```

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

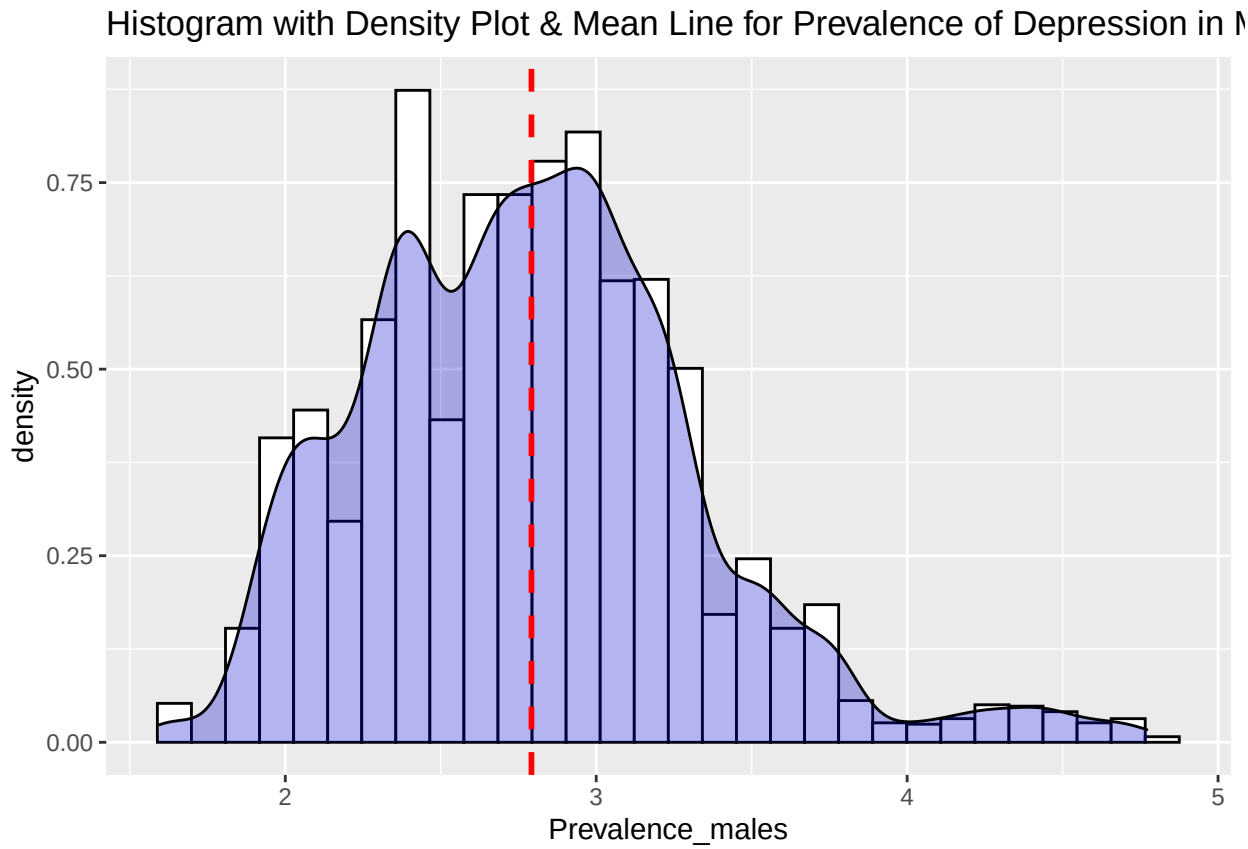


Figure 1.7 above describes the data as a right-skewed distribution and has more peaks (frequently occurring values) to the right of the distribution. With the density curve, the mean and median can be observed. The mean is greater than the median because the plot is right-skewed. Since we can observe more than one peak, the graph can be said to be multi-modal. In the context of the data, this indicates that for the depression prevalence in males, there are a number of data points (including outliers) that are greater than the mode.

Figure 1.8: Histogram with Density Plot and Mean Line for the Prevalence of Depression in Females

```
ggplot(data_gender, aes(x=Prevalence_females)) + geom_histogram(aes(y=..density..), colour="black", fil
```

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

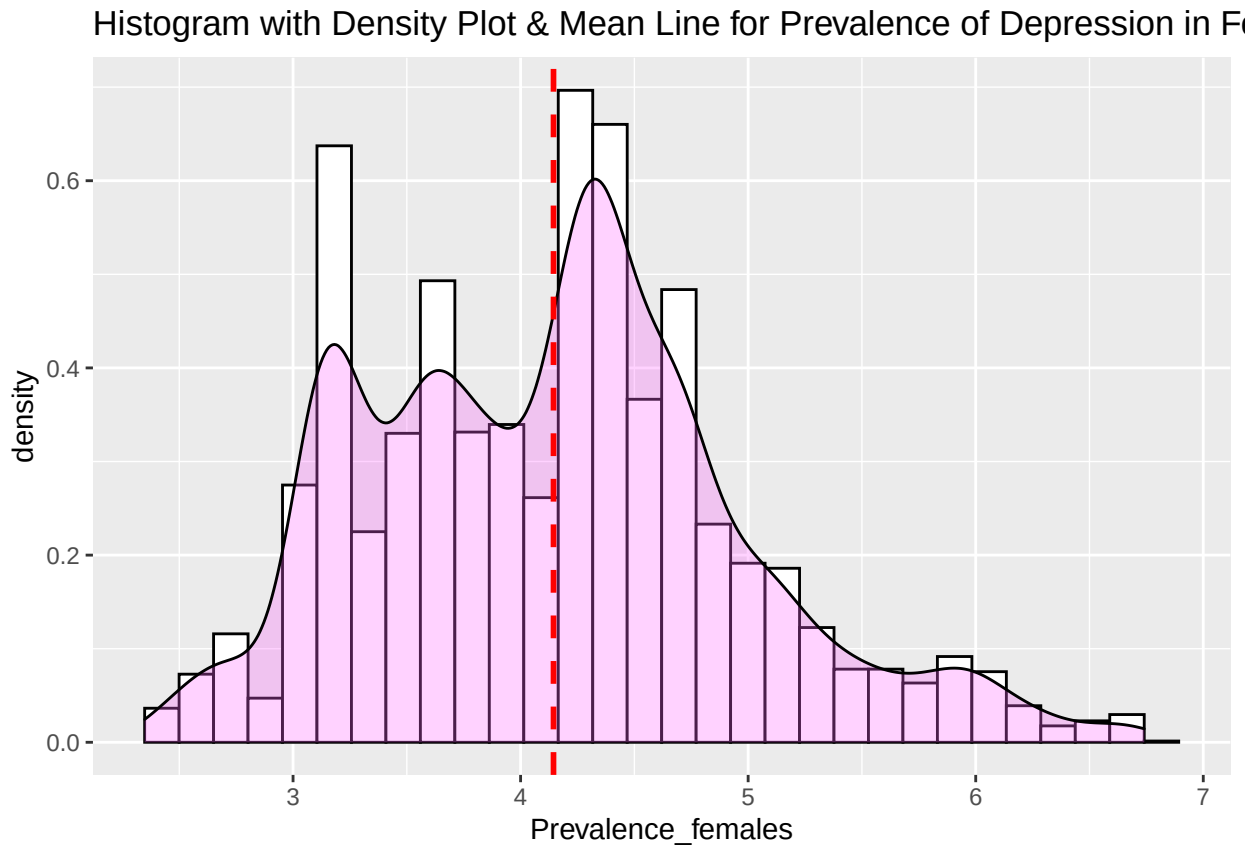


Figure 1.8 above describes the data as a right-skewed distribution and has more peaks (frequently occurring values) to the right of the distribution. With the density curve, the mean and median can be observed. The mean is greater than the median because the plot is right-skewed. Since we can observe more than one peak, the graph can be said to be multi-modal.

Depression Prevalence Among Various Age Groups

Create and display data frame for Age Groups and Depression Data

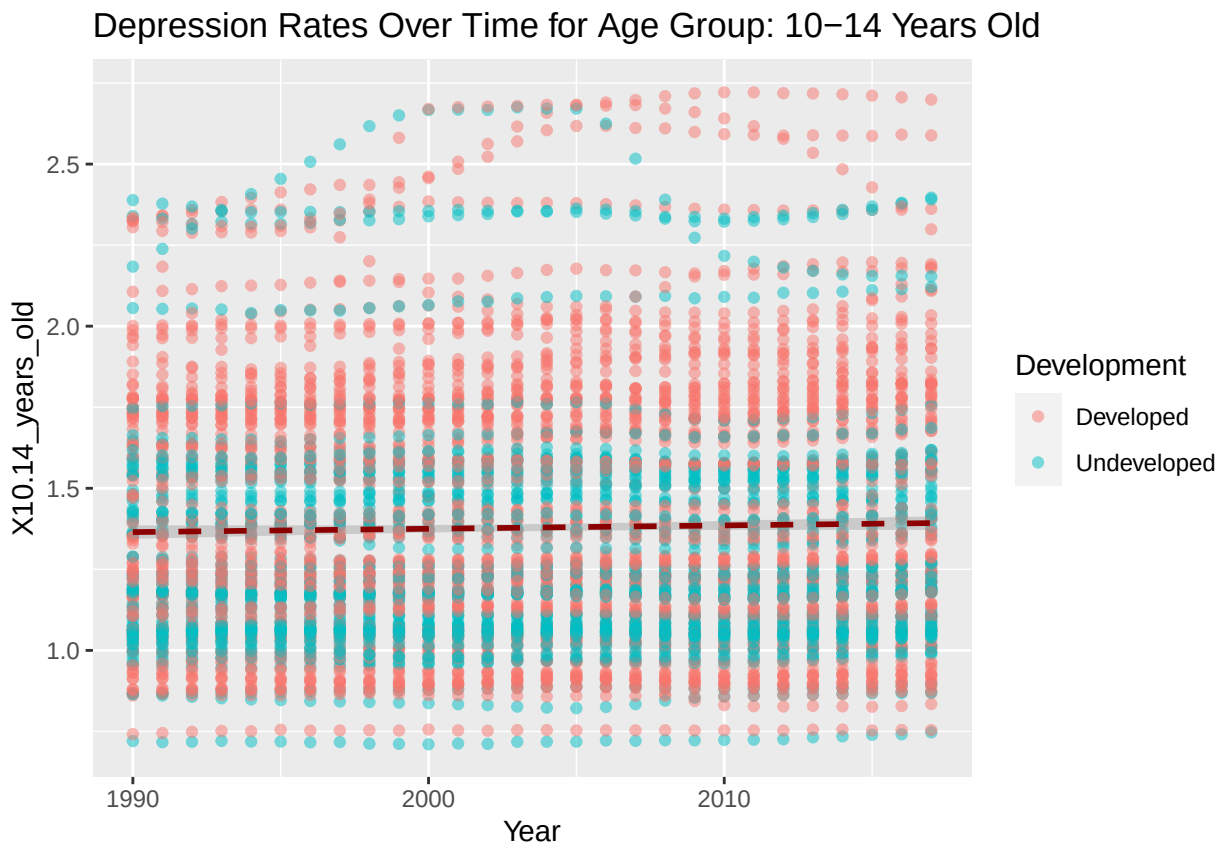
```
data_age = data[, c("Country", "Year", "X10.14_years_old", "X15.19_years_old", "X20.24_years_old", "X25
```

_Visualizations for the Prevalence of Depression through Various Age Groups, sorted by Developed and Underdeveloped Nations

Visualizations for Depression Prevalence between 10-14 Years Old

Figure 2.1: Scatter Plot for Prevalence of Depression for Age Group: Between 10-14 Years Old, sorted by Developed & Underdeveloped Nations

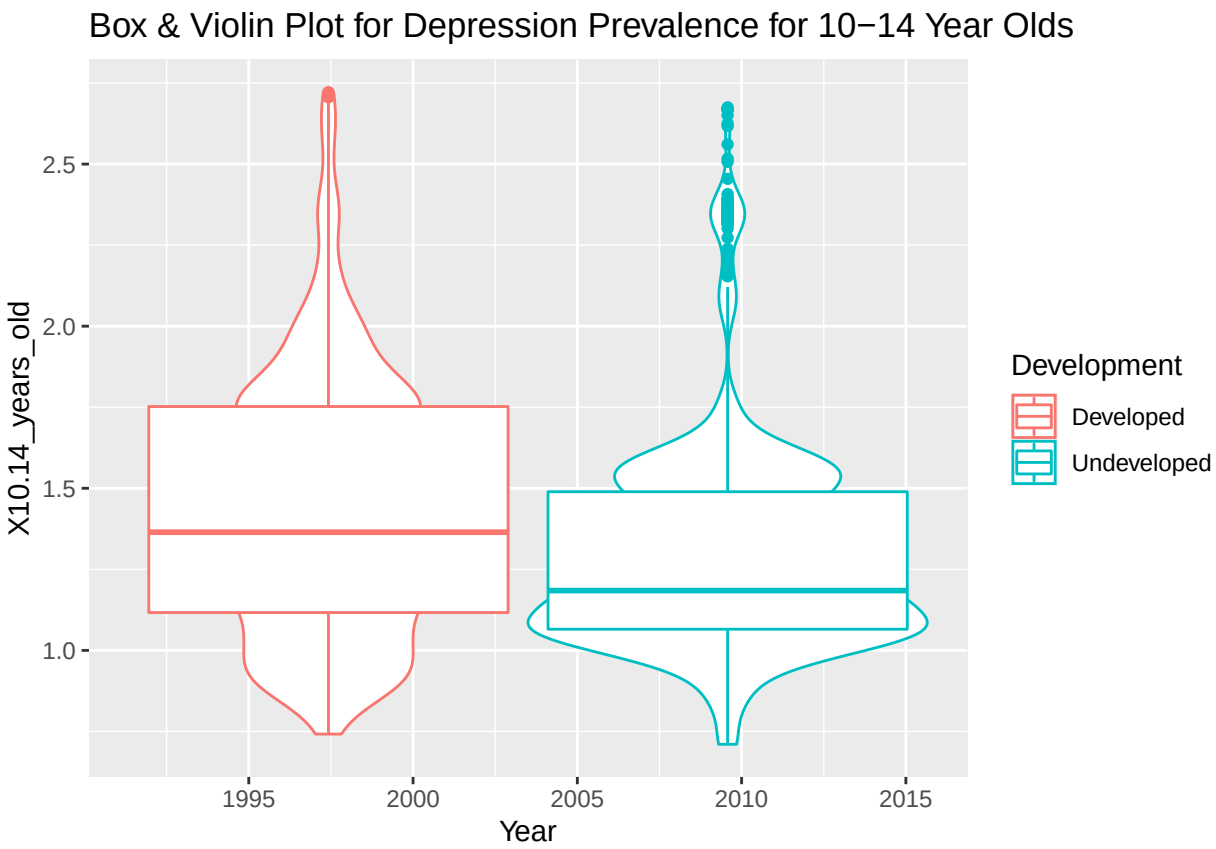
```
ggplot(data_age, aes(x = Year, y = X10.14_years_old, color = Development)) + geom_point(alpha=0.5) + ge
```



The scatter plot above (Figure 2.1) shows: * This is clearly not a normal distribution, and no correlation is being made between the x-variable (Year) and the Y-variable (Depression Rates in Age Groups of 10-14). * The smoother in Figure 2.1 helps to visualize the patterns in the presence of over plotting. The result from the smoother lines remains consistent with the finding that depression rates are higher in developed countries compared to underdeveloped.

Figure 2.2: Box & Violin Plot for Prevalence of Depression for Age Group: Between 10-14 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x = Year, y = X10.14_years_old, color = Development))+ geom_violin() + geom_boxplo
```



In the context of the data being examined, Figure 2.2 shows that depression rates in underdeveloped countries have a lot of outliers, compared to those of developed countries. There seem to be significantly less outliers for 10-14 year olds in developed countries, meaning the depression rate is pretty standard across the years of 1990 - 2019.

Figure 2.3: Histogram with Density Plot and Mean Line for Prevalence of Depression for Age Group: Between 10-14 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x=X10.14_years_old)) + geom_histogram(aes(y=..density..), colour="black", fill="wh
```

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

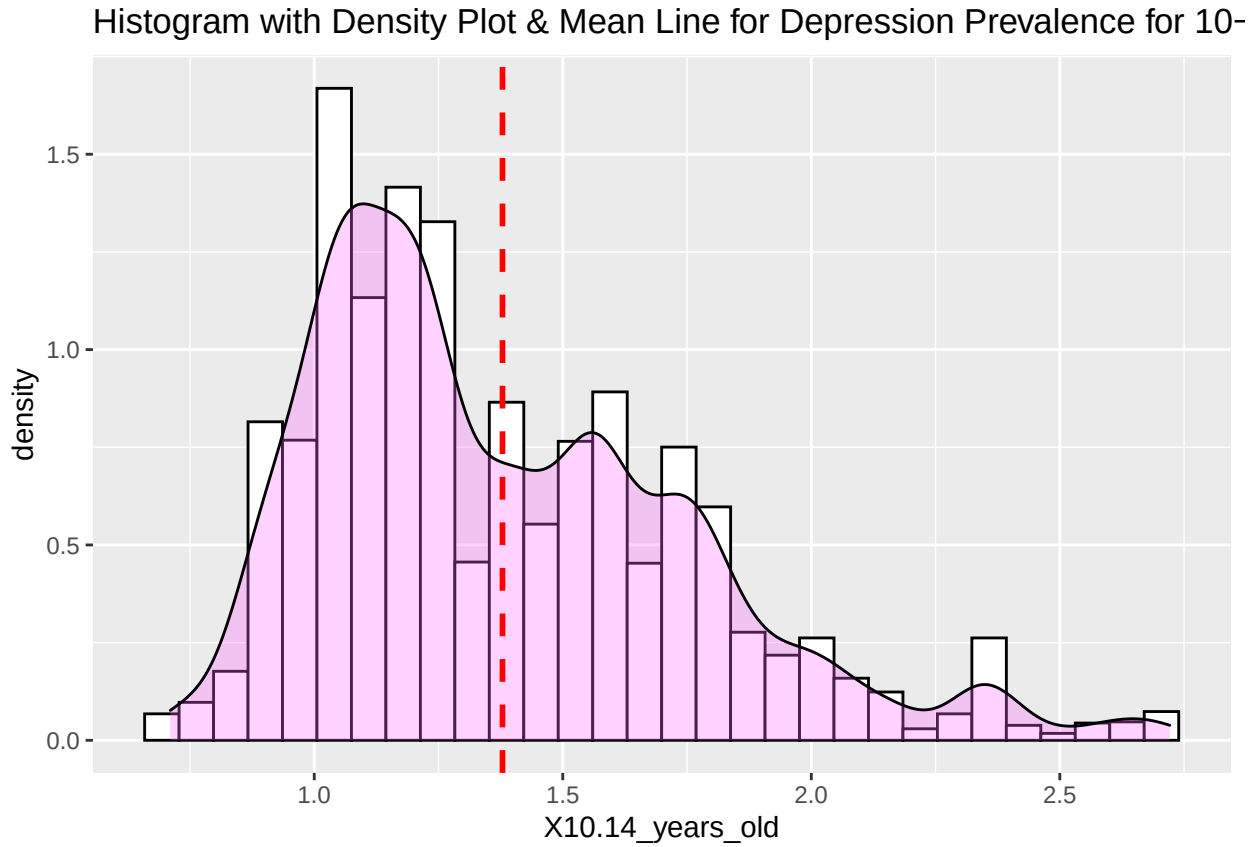
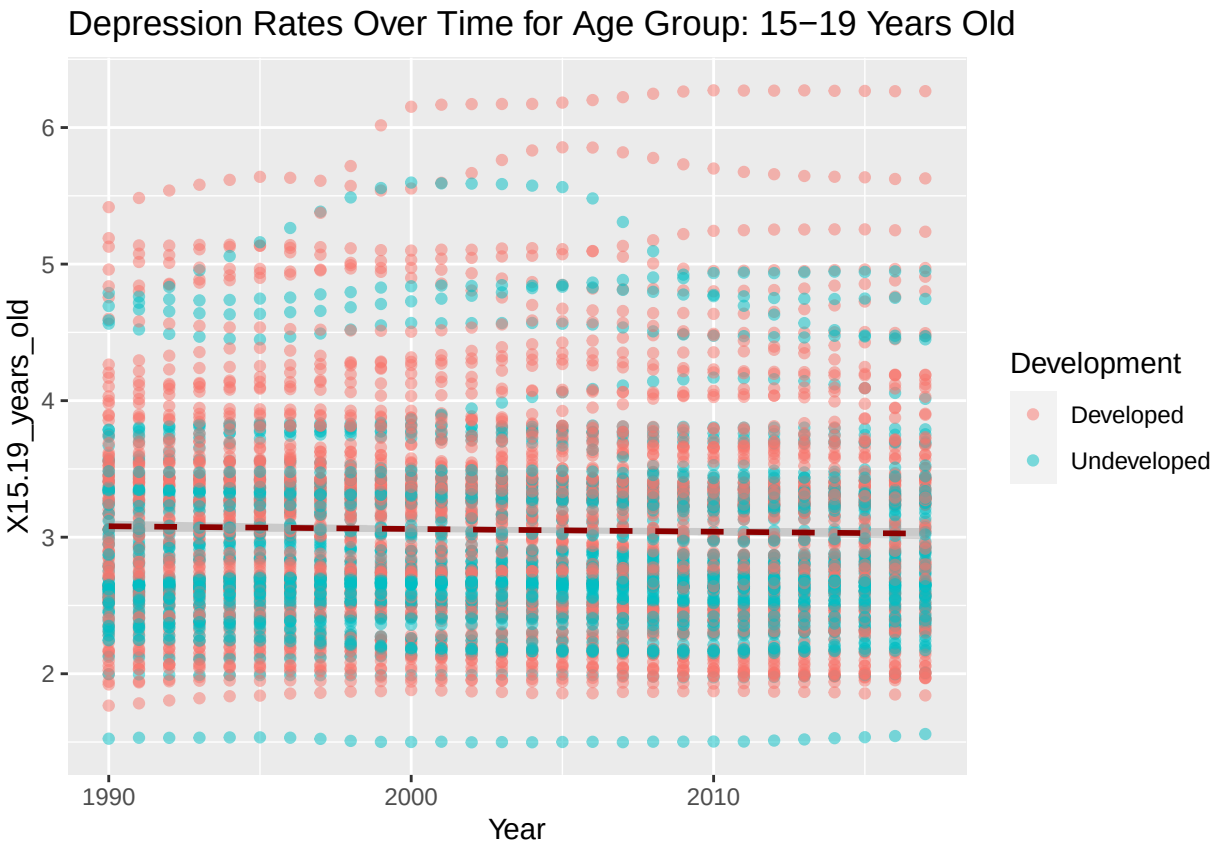


Figure 2.3 above describes the data as a right-skewed distribution and has more peaks (frequently occurring values) to the right of the distribution. With the density curve, the mean and median can be observed. The mean is greater than the median because the plot is right-skewed. Since we can observe more than one peak, the graph can be said to be multi-modal. In the context of the data, this indicates that for the depression prevalence in the age group of 10-14, there are a number of data points (including outliers) that are greater than the mode.

Visualizations for Depression Prevalence between 15-19 Years Old

Figure 3.1: Scatter Plot for Prevalence of Depression for Age Group: Between 15-19 Years Old, sorted by Developed & Underdeveloped Nations

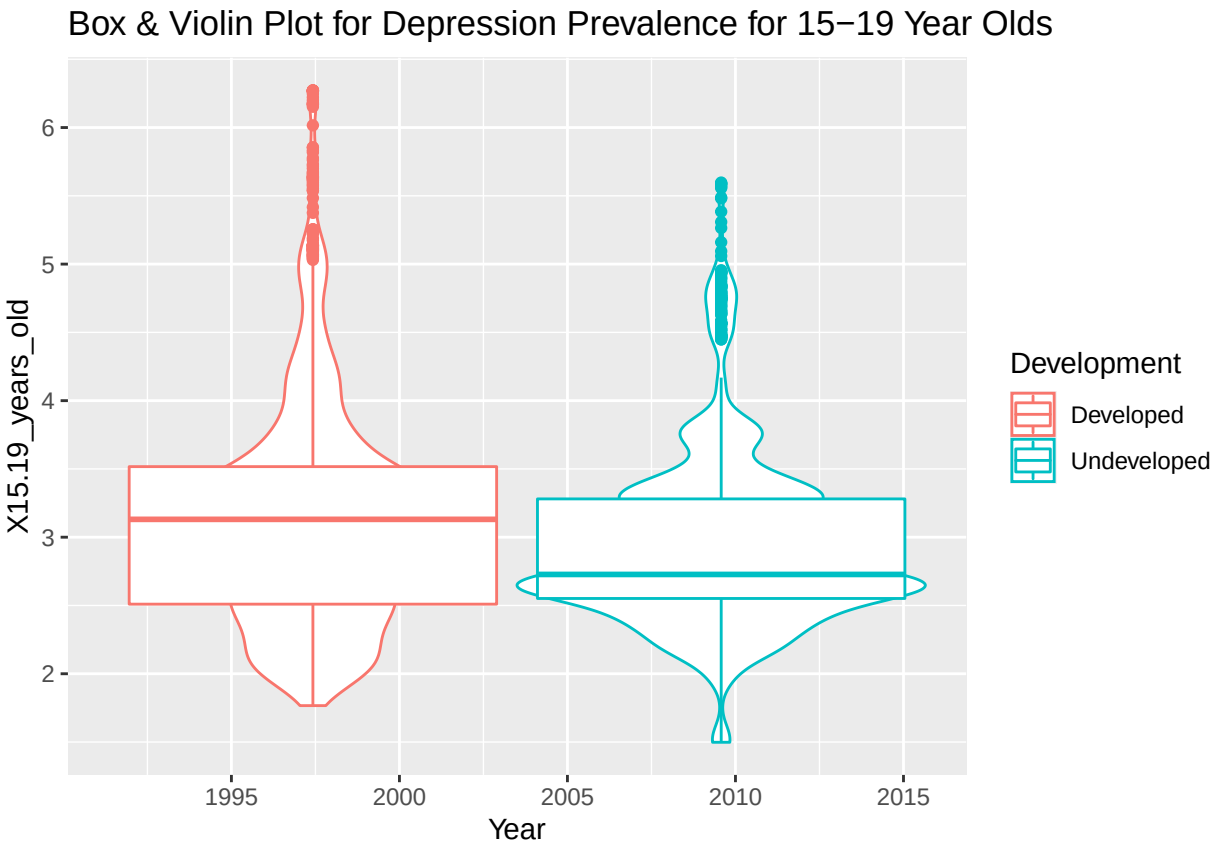
```
ggplot(data_age, aes(x = Year, y = X15.19_years_old, color = Development)) + geom_point(alpha=0.5) + ge
```



The scatter plot above (Figure 3.1) shows: * This is clearly not a normal distribution, and no correlation is being made between the x-variable (Year) and the Y-variable (Depression Rates in Age Groups of 15-19). * The smoother in Figure 3.1 helps to visualize the patterns in the presence of over plotting. The result from the smoother lines remains consistent with the finding that depression rates are higher in developed countries compared to underdeveloped. * The scatter plot for Figure 2.1 and Figure 3.1 is very similar, which makes sense as when we combine the age groups, it is for between the ages of 10 - 19 years old, representing early teens to early adults.

Figure 3.2: Box & Violin Plot for Prevalence of Depression for Age Group: Between 15-19 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x = Year, y = X15.19_years_old, color = Development)) + geom_violin() + geom_boxpl
```



In the context of the data being examined, Figure 3.2 is extremely similar to Figure 2.2, with the exception that in Figure 3.2, there are many outliers for both developed and underdeveloped countries. The similarity is that the mean of depression rates in both Figure 3.2 and 2.2 are higher for developed countries than underdeveloped. This could be explained as in developed nations, there is higher accessibility to mental health resources and better awareness to break down negative social stigma associated with depression, as well as more accurate measurement processes.

Figure 3.3: Histogram with Density Plot and Mean Line for Prevalence of Depression for Age Group: Between 15-19 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x=X15.19_years_old)) + geom_histogram(aes(y=..density..), colour="black", fill="wh
```

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

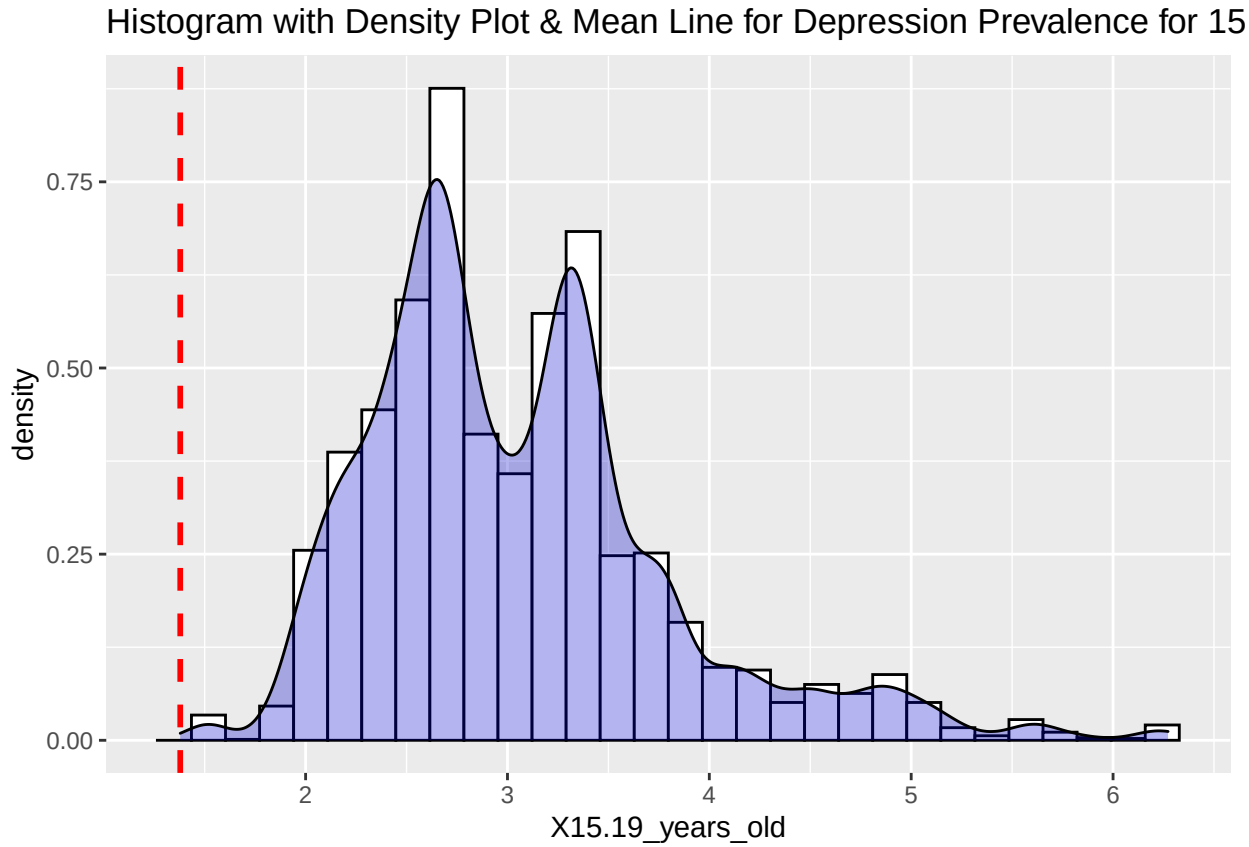
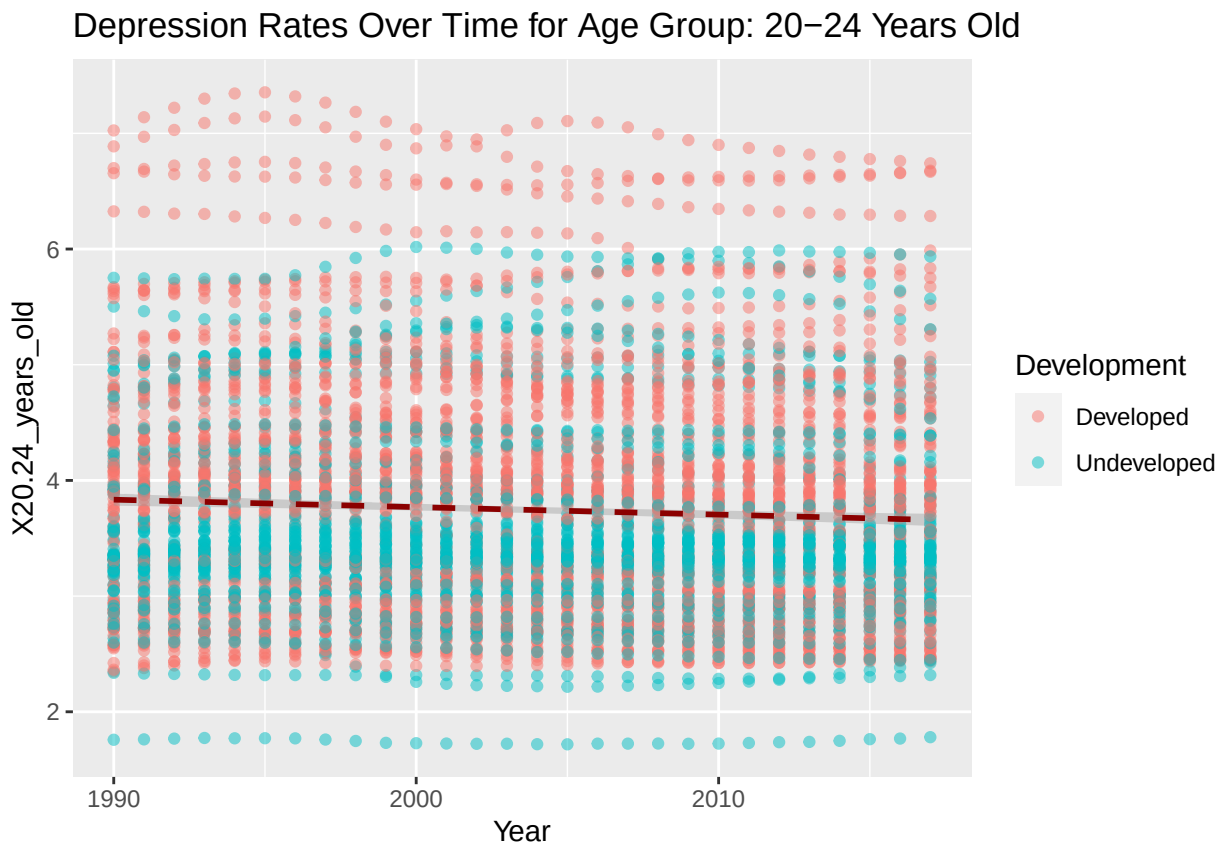


Figure 3.3 above has more peaks (frequently occurring values) to the right of the mean. Thus it is right-skewed. With the density curve, the mean and median can be observed. Since we can observe more than one peak, the graph can be said to be multi-modal. Compared to Figure 2.3, the graph looks very different for 15-19 year olds than it was for 10-14 year olds. This is a very asymmetrical graph, and making some assumptions, depression is an outlier until 2 on the x-axis, then steeply declines during 3 and is at an extreme low for 4.

Visualizations for Depression Prevalence between 20-24 Years Old

Figure 4.1: Scatter Plot for Prevalence of Depression for Age Group: Between 20-24 Years Old, sorted by Developed & Underdeveloped Nations

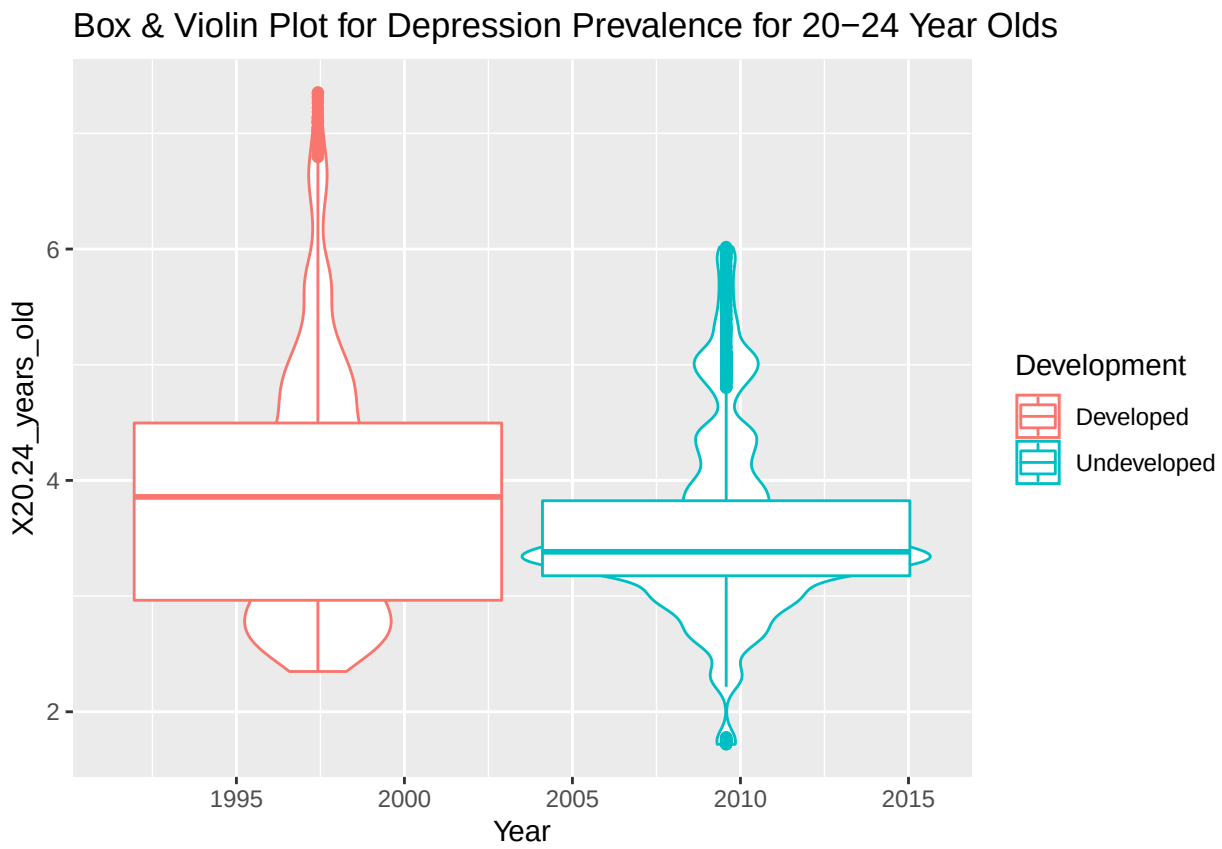
```
ggplot(data_age, aes(x = Year, y = X20.24_years_old, color = Development)) + geom_point(alpha=0.5) + ge
```



The scatter plot above (Figure 4.1) shows: * This is clearly not a normal distribution, and no correlation is being made between the x-variable (Year) and the Y-variable (Depression Rates in Age Groups of 20-24). * The smoother in Figure 2.1 helps to visualize the patterns in the presence of over plotting. The result from the smoother lines remains consistent with the finding that depression rates are higher in developed countries compared to underdeveloped.

Figure 4.2: Box & Violin Plot for Prevalence of Depression for Age Group: Between 20-24 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x = Year, y = X20.24_years_old, color = Development)) + geom_violin() + geom_boxpl
```



In the context of the data being examined, Figure 4.2 shows a very drastic difference in distribution shape when looking at developed versus underdeveloped countries. There are also a significantly higher rate of outliers in underdeveloped countries which may be explained by inaccurate reporting methods for depression disorder.

Figure 4.3: Histogram with Density Plot and Mean Line for Prevalence of Depression for Age Group: Between 20-24 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x=X20.24_years_old)) + geom_histogram(aes(y=..density..), colour="black", fill="wh
```

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

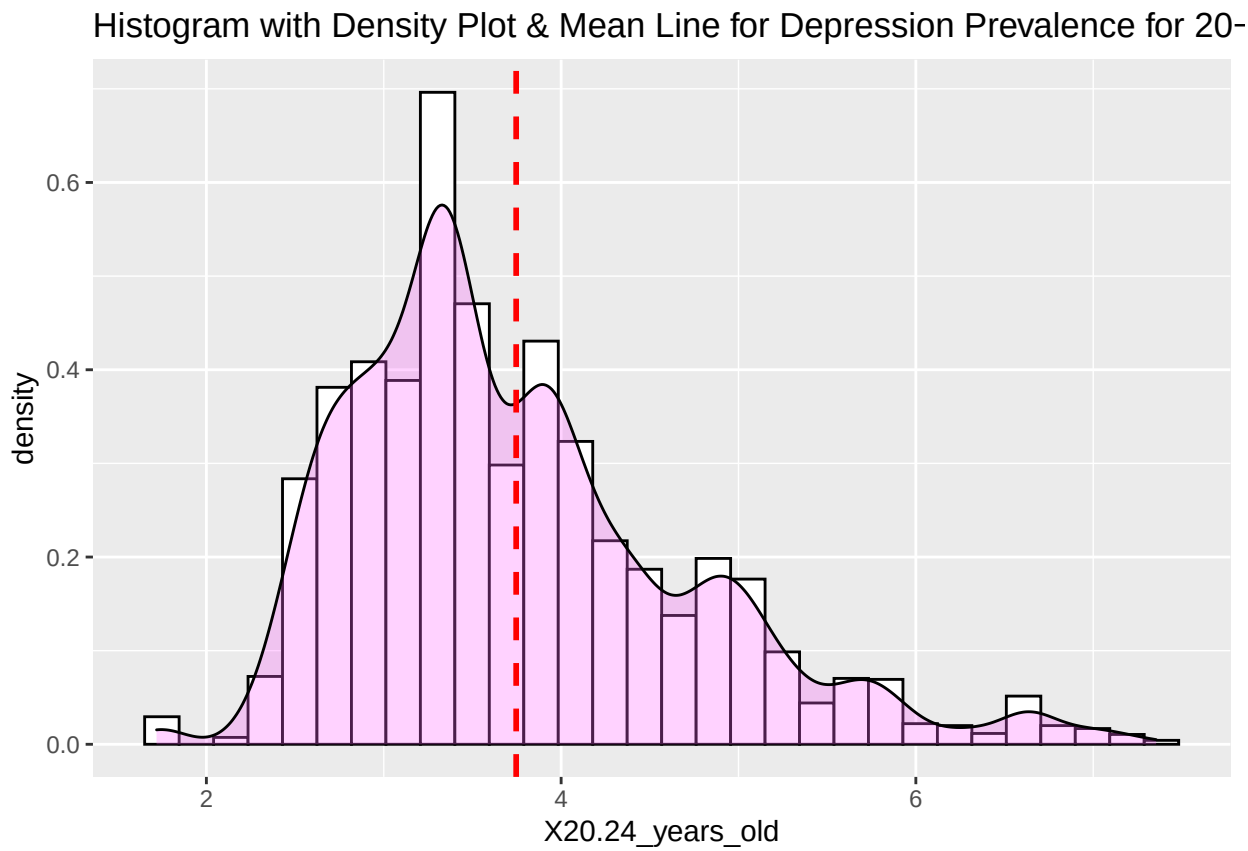
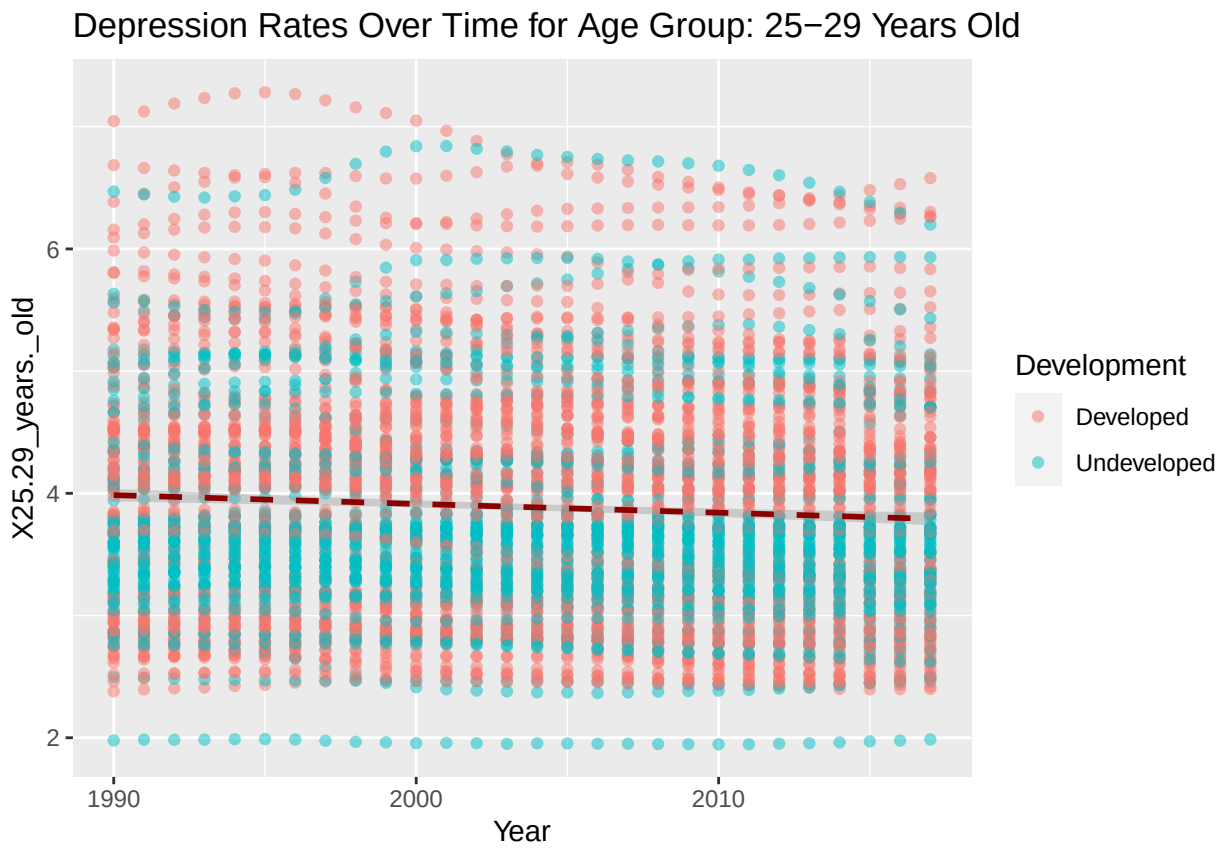


Figure 4.3 is a right skewed distribution, with majority of the peaks occuring to the right of the mean. A more conclusive explanation will be provided later in the report.

Visualizations for Depression Prevalence between 25-29 Years Old

Figure 5.1: Scatter Plot for Prevalence of Depression for Age Group: Between 25-29 Years Old, sorted by Developed & Underdeveloped Nations

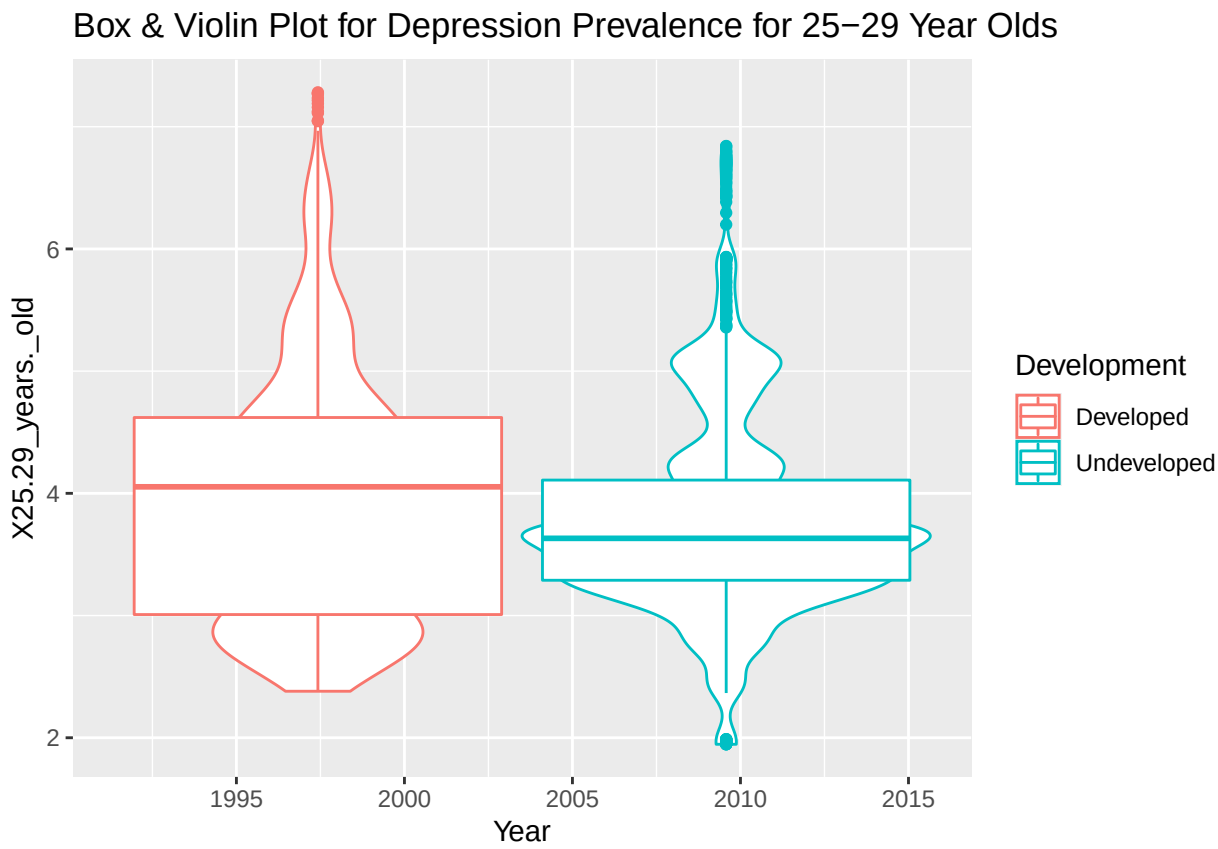
```
ggplot(data_age, aes(x = Year, y = X25.29_years._old, color = Development)) + geom_point(alpha=0.5) + g
```



The scatter plot above (Figure 5.1) shows: * This is clearly not a normal distribution, and no correlation is being made between the x-variable (Year) and the Y-variable (Depression Rates in Age Groups of 25-29). * The smoother in Figure 3.1 helps to visualize the patterns in the presence of over plotting. The result from the smoother lines remains consistent with the finding that depression rates are higher in developed countries compared to underdeveloped. * The scatter plot for Figure 4.1 and Figure 5.1 is very similar, which makes sense as when we combine the age groups, it is for between the ages of 20 - 29 years old, representing early teens to early adults. This is interesting as many mental illnesses, including depression are often diagnosed when a person is in their mid-early twenties.

Figure 5.2: Box & Violin Plot for Prevalence of Depression for Age Group: Between 25-29 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x = Year, y = X25.29_years._old, color = Development)) + geom_violin() + geom_boxpl
```



In the context of the data being examined, Figure 5.2 is similar to Figure 4.2. The aggregated conclusions for this will be explained later on in the report.

Figure 5.3: Histogram with Density Plot and Mean Line for Prevalence of Depression for Age Group: Between 25-29 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x=X25.29_years._old)) + geom_histogram(aes(y=..density..), colour="black", fill="wh
```

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

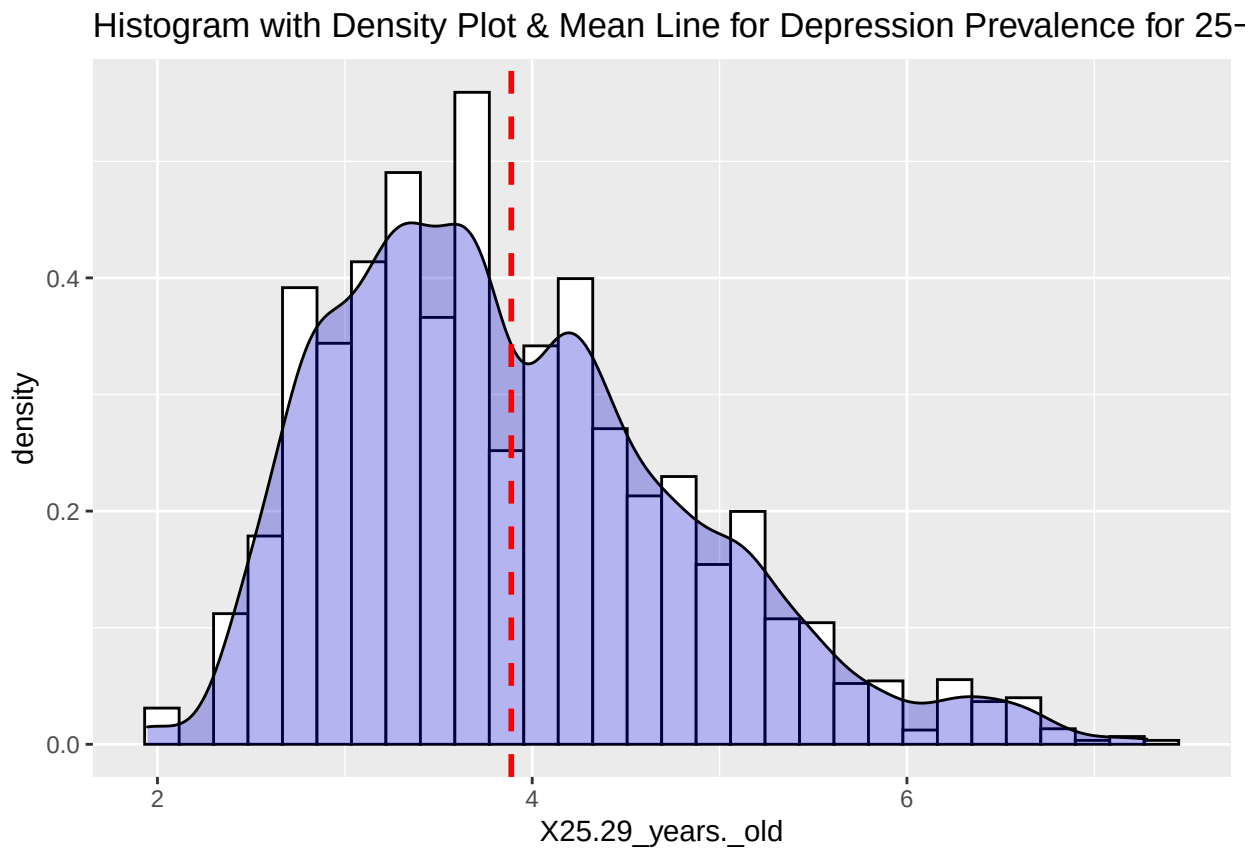
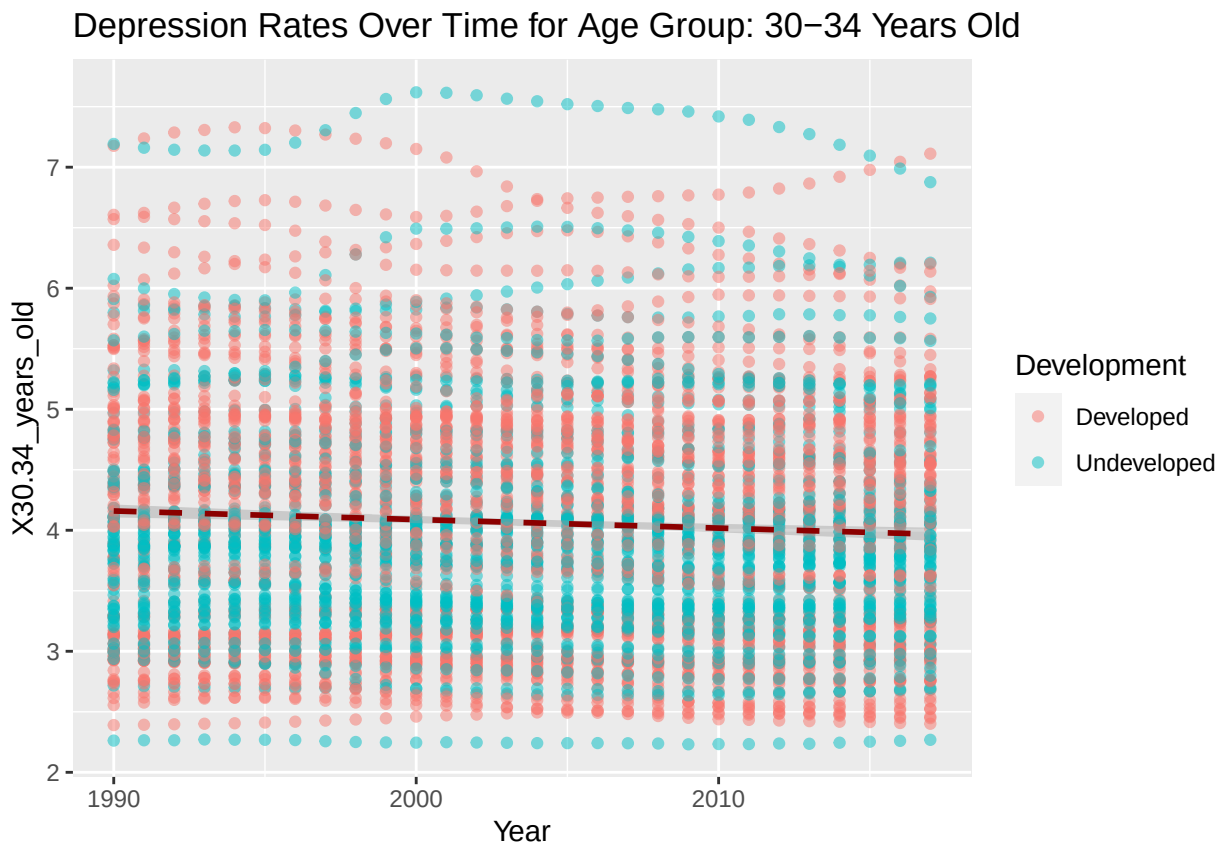


Figure 5.3 is a right skewed distribution, with majority of the peaks occuring to the right of the mean. A more conclusive explanation will be provided later in the report.

Visualizations for Depression Prevalence between 30-34 Years Old

Figure 6.1: Scatter Plot for Prevalence of Depression for Age Group: Between 30-34 Years Old, sorted by Developed & Underdeveloped Nations

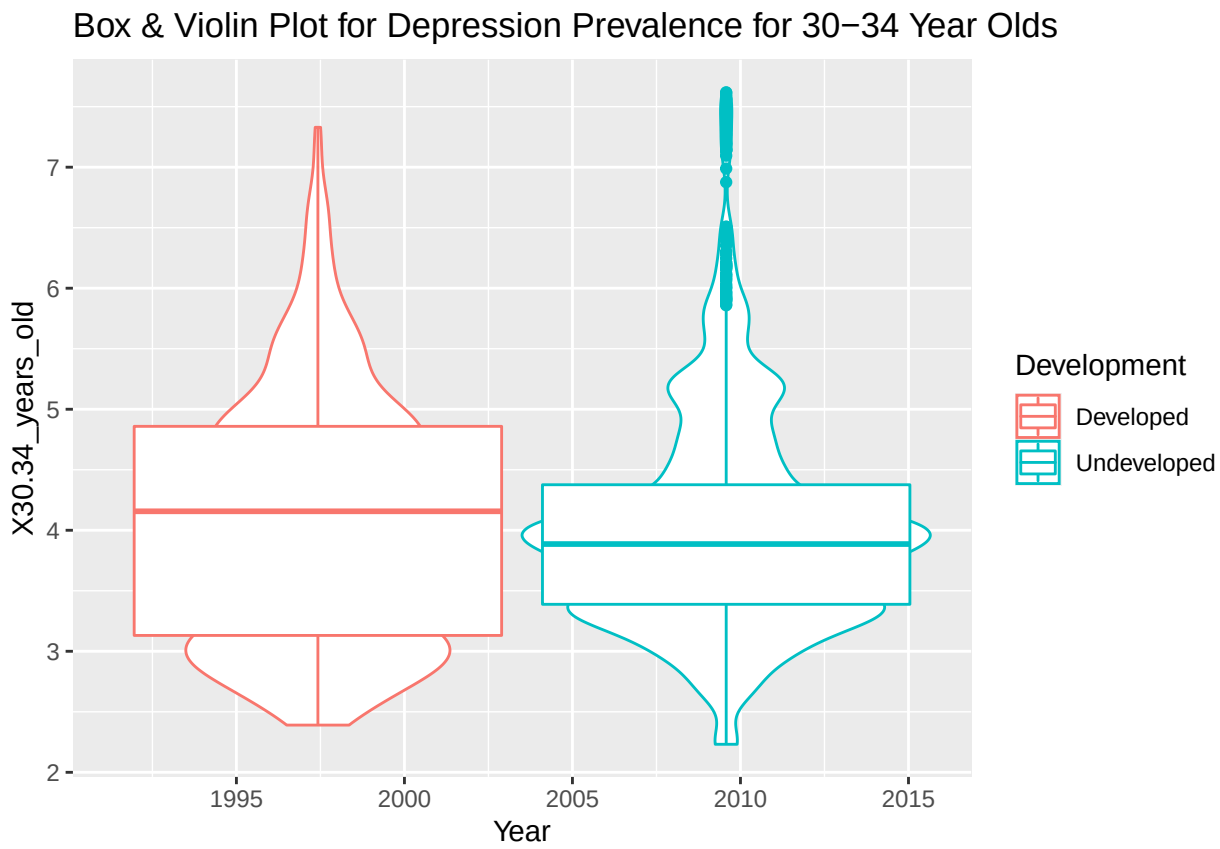
```
ggplot(data_age, aes(x = Year, y = X30.34_years_old, color = Development)) + geom_point(alpha=0.5) + ge
```



The scatter plot above (Figure 6.1) shows: * This is clearly not a normal distribution, and no correlation is being made between the x-variable (Year) and the Y-variable (Depression Rates in Age Groups of 30-34). * The smoother in Figure 6.1 helps to visualize the patterns in the presence of over plotting. The result from the smoother lines remains consistent with the finding that depression rates are higher in developed countries compared to underdeveloped. * Over time, depression rates for individuals between the ages 30-34 increased for those in underdeveloped countries when compared to developed countries.

Figure 6.2: Box & Violin Plot for Prevalence of Depression for Age Group: Between 30-34 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x = Year, y = X30.34_years_old, color = Development)) + geom_violin() + geom_boxpl
```



In the context of the data being examined, Figure 6.2 shows more data for the developed nations compared to underdeveloped as developed nations have a higehr mean rate. The aggregated conclusions for this will be explained later on in the report.

Figure 6.3: Histogram with Density Plot and Mean Line for Prevalence of Depression for Age Group: Between 30-34 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x=X30.34_years_old)) + geom_histogram(aes(y=..density..), colour="black", fill="wh
```

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

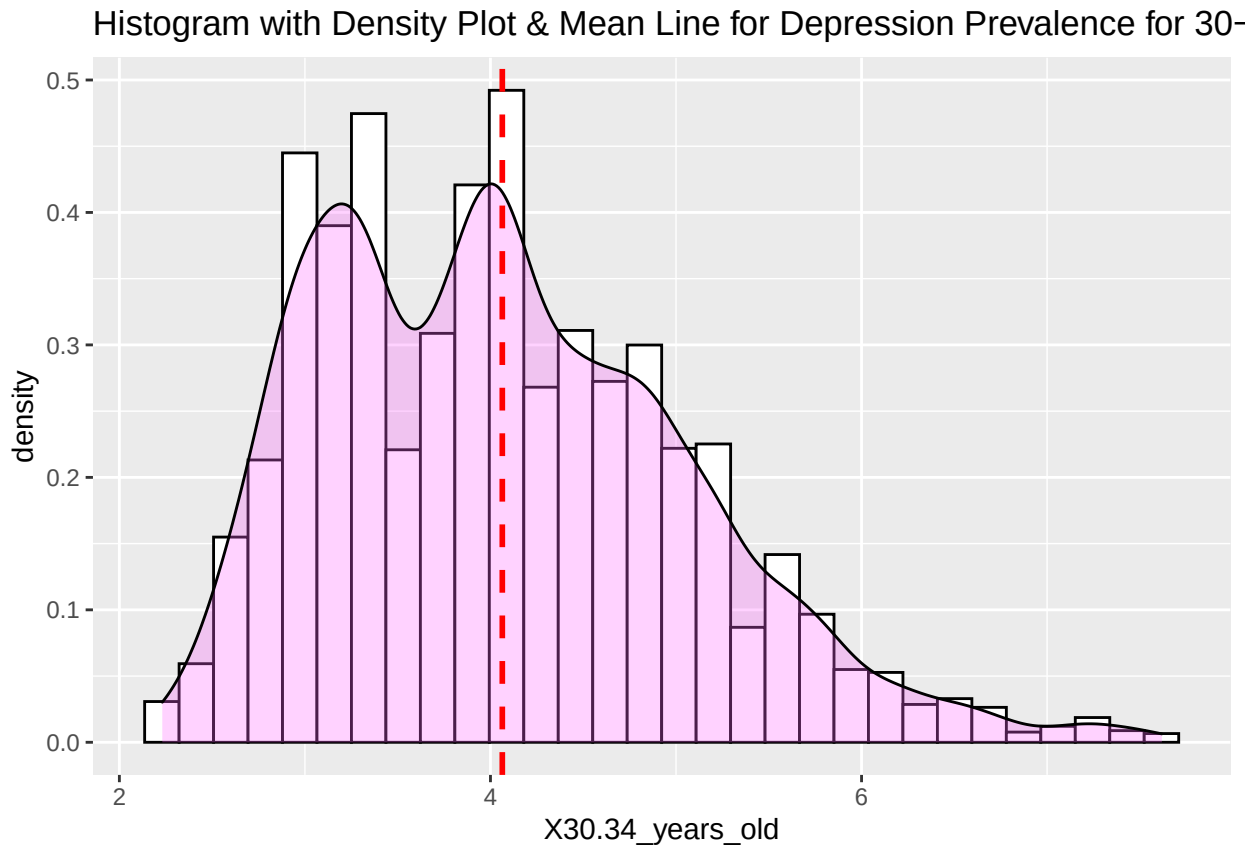


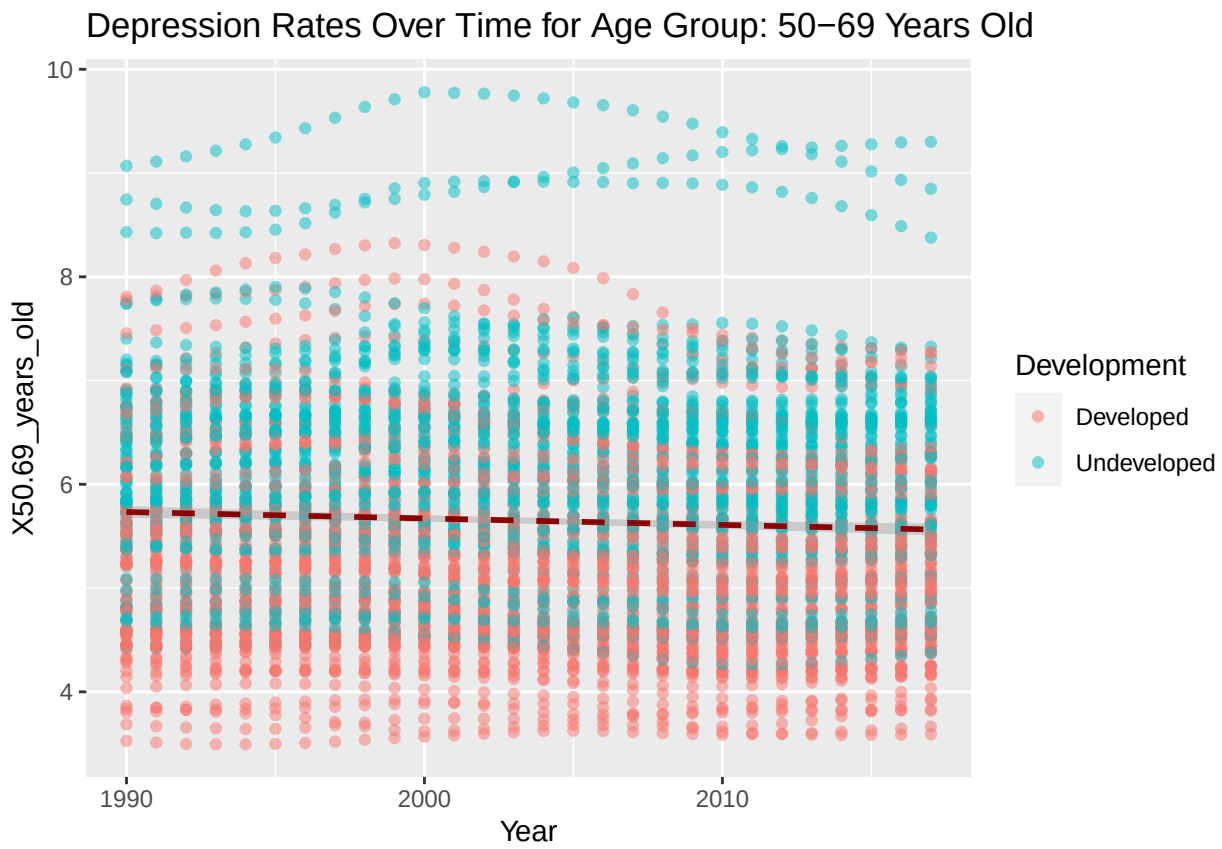
Figure 6.3 is a right skewed distribution, with majority of the peaks occuring to the right of the mean. A more conclusive explanation will be provided later in the report.

Figure 6.3 above is very similar to Figure 3.3. Figure 6.3 has more peaks (frequently occurring values) to the right of the mean. Thus it is right-skewed. With the density curve, the mean and median can be observed. Since we can observe more than one peak, the graph can be said to be multi-modal. Making some assumptions, depression is an outlier until after 2 on the x-axis, then steeply declines after 4 and is at an extreme low for after 6.

Visualizations for Depression Prevalence between 50-69 Years Old

Figure 7.1: Scatter Plot for Prevalence of Depression for Age Group: Between 50-69 Years Old, sorted by Developed & Underdeveloped Nations

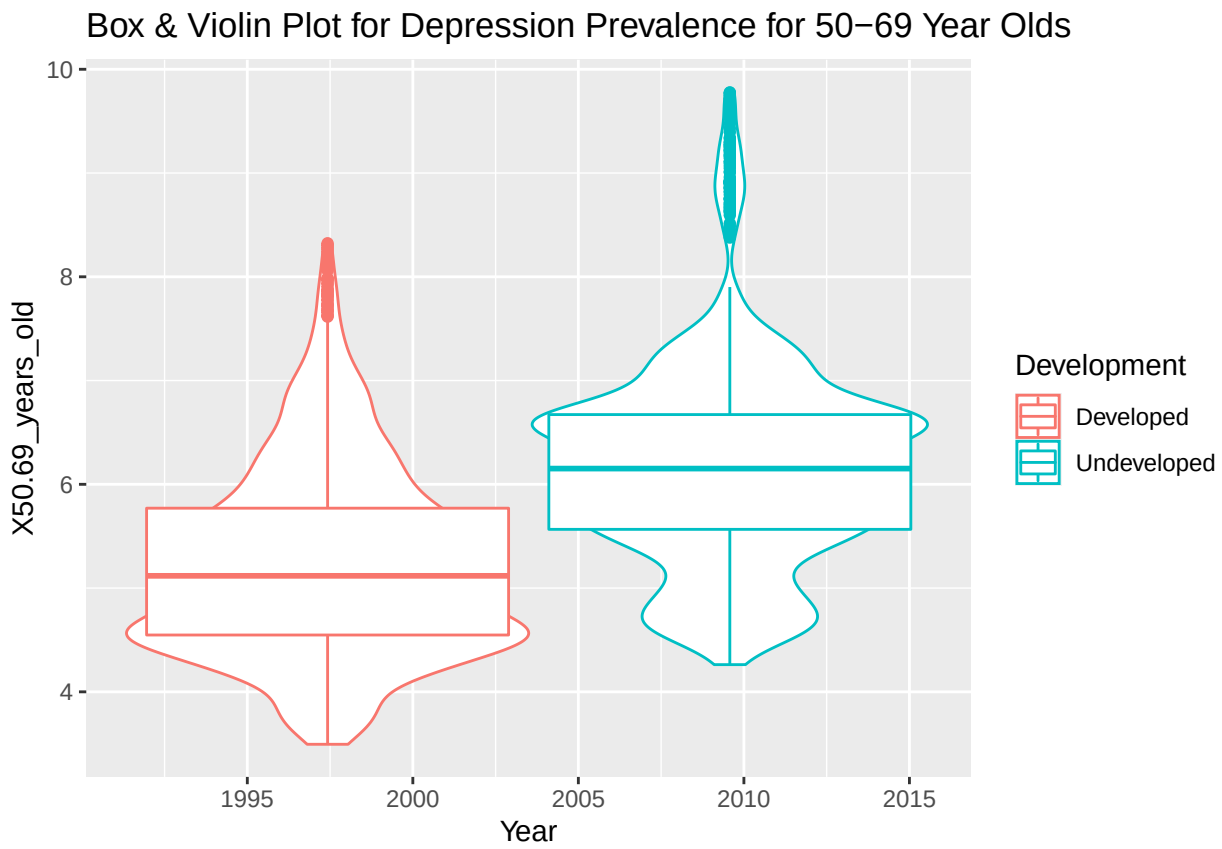
```
ggplot(data_age, aes(x = Year, y = X50.69_years_old, color = Development)) + geom_point(alpha=0.5) + ge
```



The scatter plot above (Figure 7.1) shows: * This is clearly not a normal distribution, and no correlation is being made between the x-variable (Year) and the Y-variable (Depression Rates in Age Groups of 50-69). * The smoother in Figure 6.1 helps to visualize the patterns in the presence of over plotting. The result from the smoother lines remains consistent with the finding that depression rates are higher in developed countries compared to underdeveloped. * Over time, depression rates for individuals between the ages 50-69 increased significantly for those in underdeveloped countries when compared to developed countries.

Figure 7.2: Box & Violin Plot for Prevalence of Depression for Age Group: Between 50-69 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x = Year, y = X50.69_years_old, color = Development))+ geom_violin() + geom_boxplo
```



In the context of the data being examined, Figure 7.2 shows that the depression rates for the age groups of 50-69 year olds are higher in underdeveloped countries. More detailed findings will be explained later on in this report.

Figure 7.3: Histogram with Density Plot and Mean Line for Prevalence of Depression for Age Group: Between 50-69 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x=X50.69_years_old)) + geom_histogram(aes(y=..density..), colour="black", fill="wh
```

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

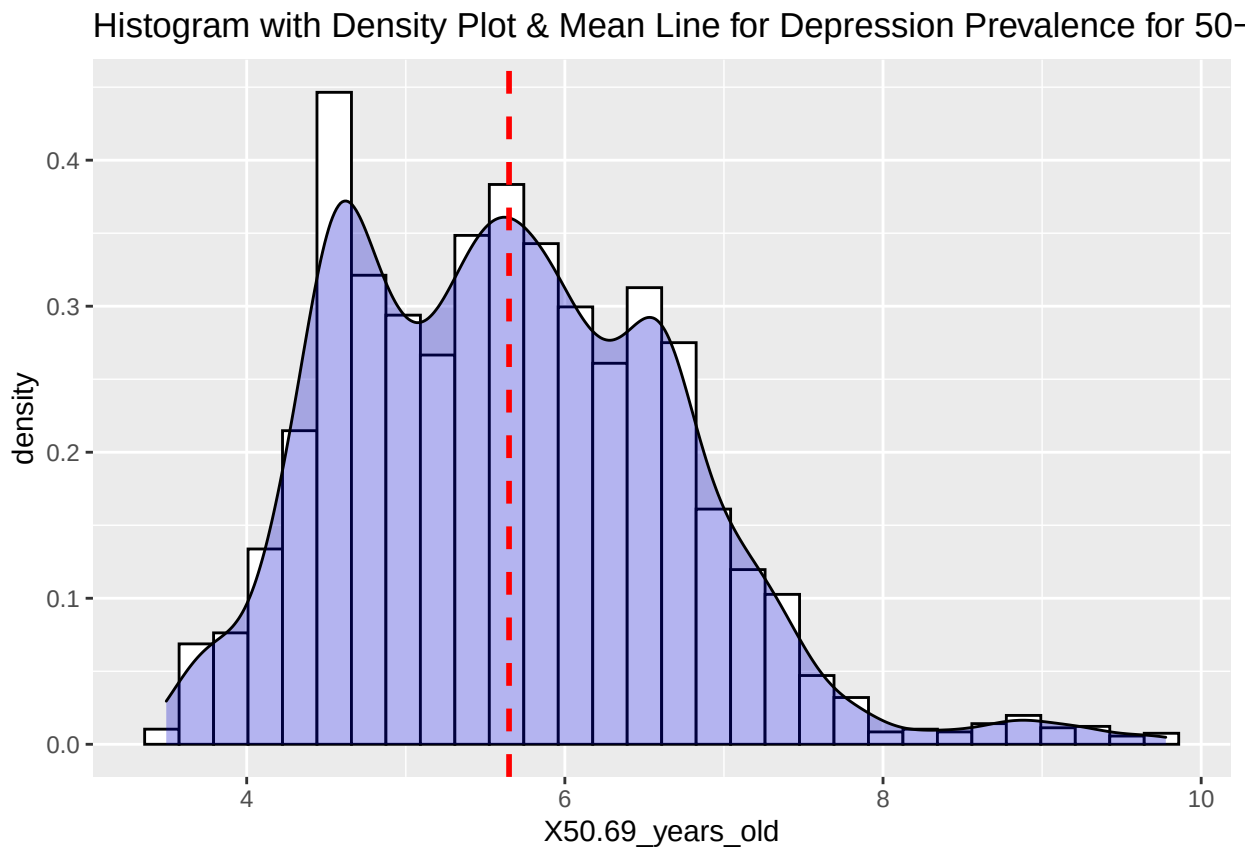
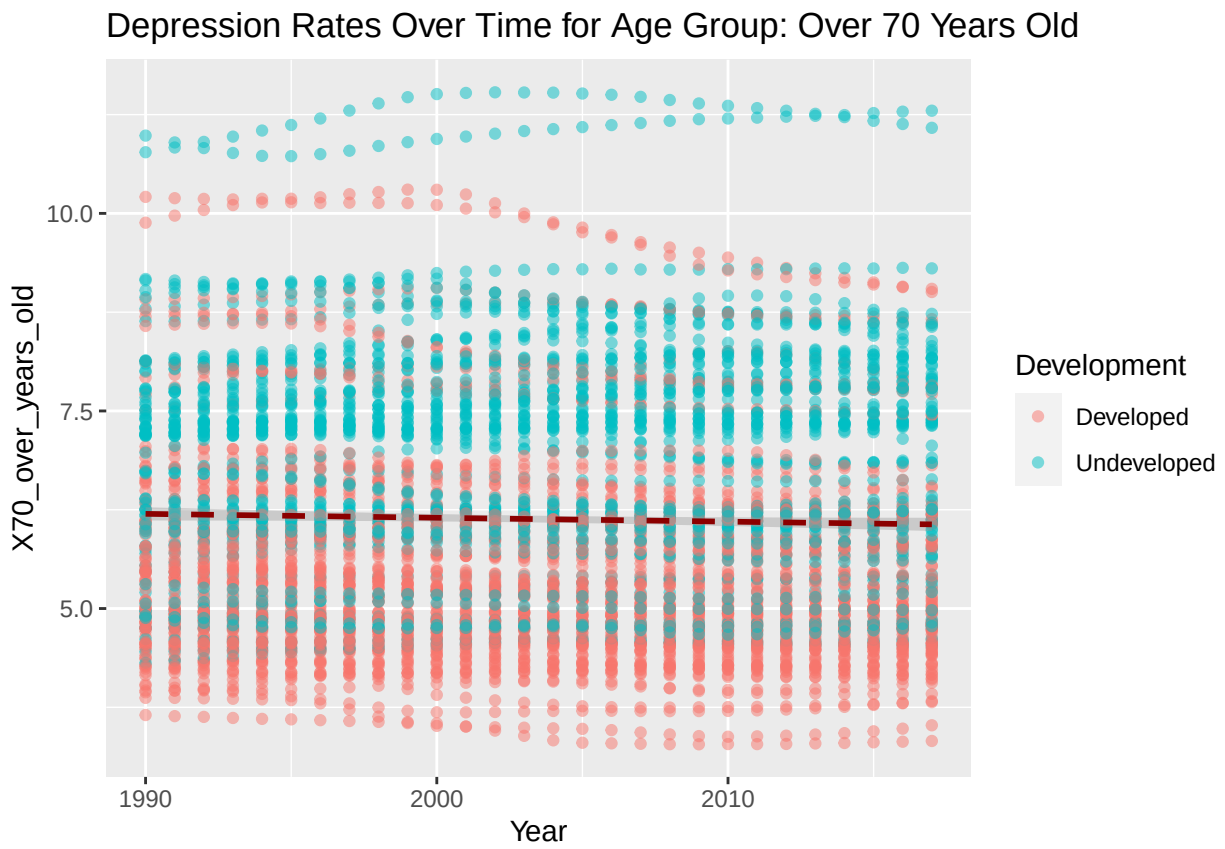


Figure 7.3 is a right skewed distribution, with majority of the peaks occurring to the right of the mean. A more conclusive explanation will be provided later in the report.

Visualizations for Depression Prevalence between 70+ Years Old

Figure 8.1: Scatter Plot for Prevalence of Depression for Age Group: Over 70 Years Old, sorted by Developed & Underdeveloped Nations

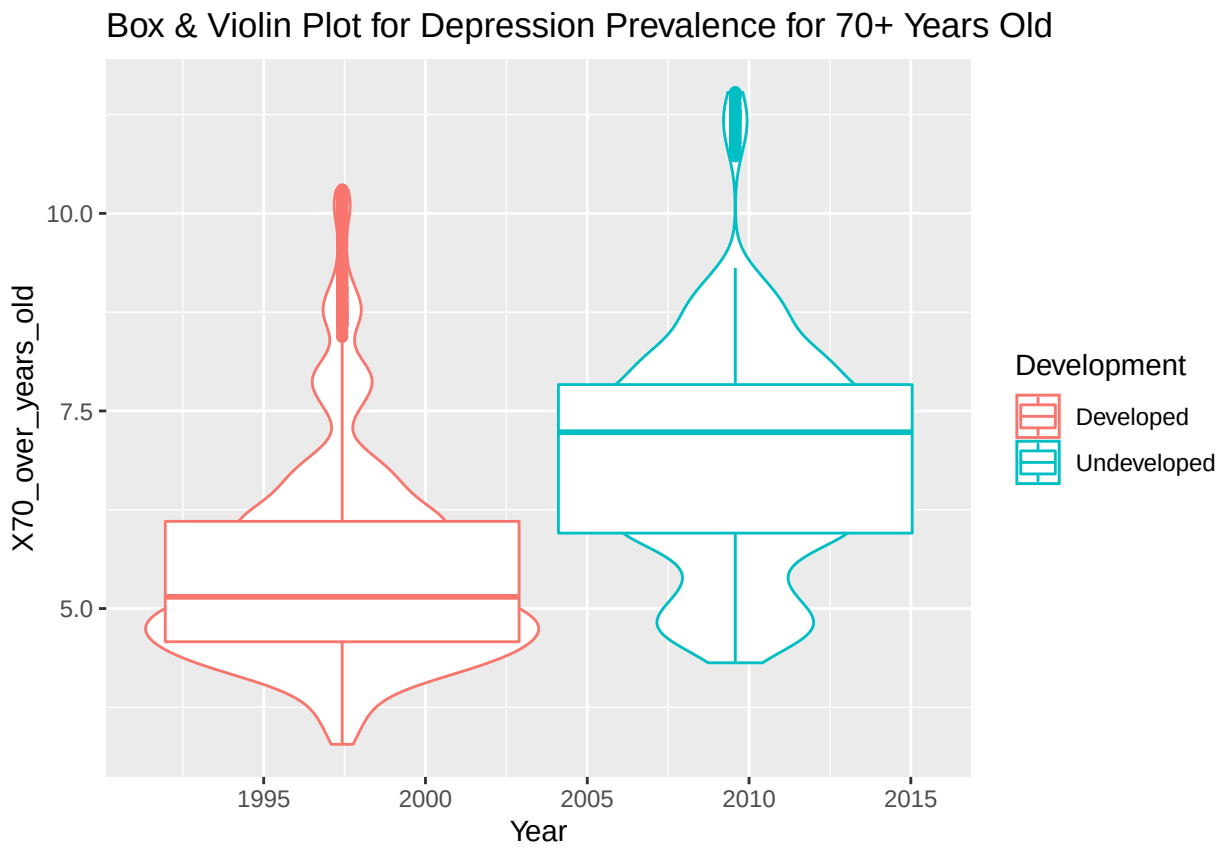
```
ggplot(data_age, aes(x = Year, y = X70_over_years_old, color = Development)) + geom_point(alpha=0.5) + g
```



The scatter plot above (Figure 8.1) shows: * This is clearly not a normal distribution, and no correlation is being made between the x-variable (Year) and the Y-variable (Depression Rates in Age Groups of 70+). * The smoother in Figure 6.1 helps to visualize the patterns in the presence of over plotting. The result from the smoother lines remains consistent with the finding that depression rates are higher in developed countries compared to underdeveloped. * Over time, depression rates for individuals that are 70+ increased significantly for those in underdeveloped countries when compared to developed countries. * The findings are extremely similar to that of Figure 7.1, which could indicate that after the age of 50, depression on average levels out to a common level.

Figure 8.2: Box & Violin Plot for Prevalence of Depression for Age Group: Over 70 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x = Year, y = X70_over_years_old, color = Development)) + geom_violin() + geom_boxp
```

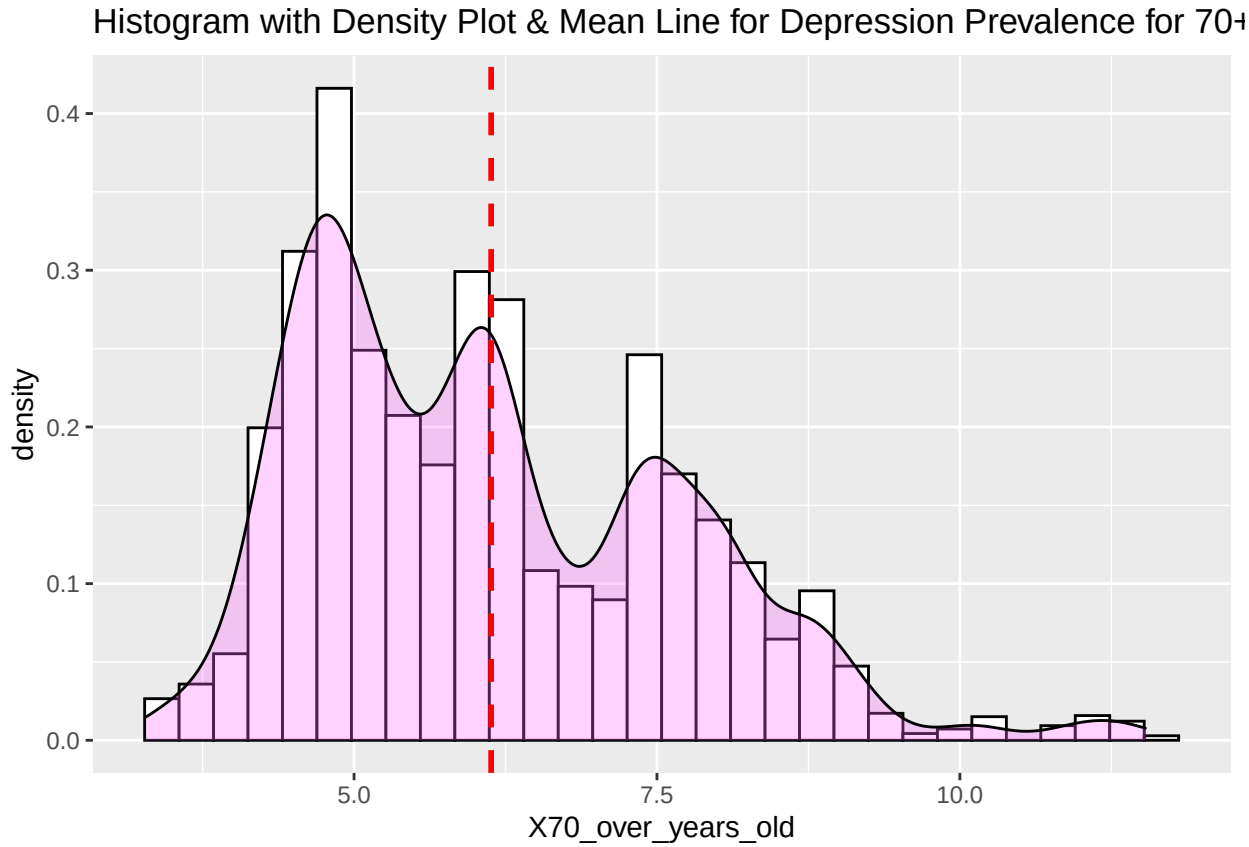


In the context of the data being examined, Figure 8.2 shows very high similarity to Figure 7.2, in that depression rates are higher in underdeveloped countries in the 70+ age group. More detailed findings will be explained later on in this report.

Figure 8.3: Histogram with Density Plot and Mean Line for Prevalence of Depression for Age Group: Over 70 Years Old, sorted by Developed & Underdeveloped Nations

```
ggplot(data_age, aes(x=X70_over_years_old)) + geom_histogram(aes(y=..density..), colour="black", fill="
```

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



Conclusion to Question 1

1. Which demographics (i.e., gender, age groups) are more susceptible to depression?

Based on the above analyses, it can be determined that in all countries the median estimate for the prevalence of depression is higher for females than for males throughout time. Thus, females are more susceptible to depression. This could possibly be attributed to men being socialized differently than women, and thus resulting in men being discouraged from seeking out help with their mental health.

Regarding age group demographics, it can be determined that ages 70 and older, globally, have a higher rate of depression relative to any other age group and are thus more susceptible to depression. Moreover, from the age group visualizations, it can also be determined that the modal onset age of depression is at 19, although majority of people develop depression later than this (as seen with the figures for age groups between 25-29, 30-34, and 50-69).

a) `Are these different for different parts of the world or change over time?`

When examining gender, the depression rate for males is consistently higher in developed nations compared to undeveloped nations. This may be explained by a variety of different reasons:

- Highly developed nations on average have more accessibility to healthcare and mental health resources.
- The screening and criterion for mental health disorders are more accurate in developed nations.
- There has been a push to eliminate negative social stigma associated with mental health in developed nations and create more awareness. As a result, the population may be more encouraged to seek out mental health resources than compared to underdeveloped nations.

- Less negative social stigma towards mental illness in developed nations may lead to more reported cases of mental illness.

When examining age groups, the depression rate for age groups are always higher in developed countries than underdeveloped, with the exception of the age groups 50-69 and 70+. Over time, we can also see a trend that the depression in developed countries tends to decrease over time across all age groups. The reason for this could be explained by the changing diagnoses methods and criterion of depression over time, especially in developed nations. It can be inferred that in underdeveloped nations, these updated measurement systems for accurate diagnoses are lagging behind, resulting in under-reporting of depression across all age groups.

Load the necessary datasets for Mental Health / Gender / Substance / Suicide Rates

```
data3 = read.csv("Mental Health Merged.csv")
```

Question 2

Concepts Covered in Each Question:

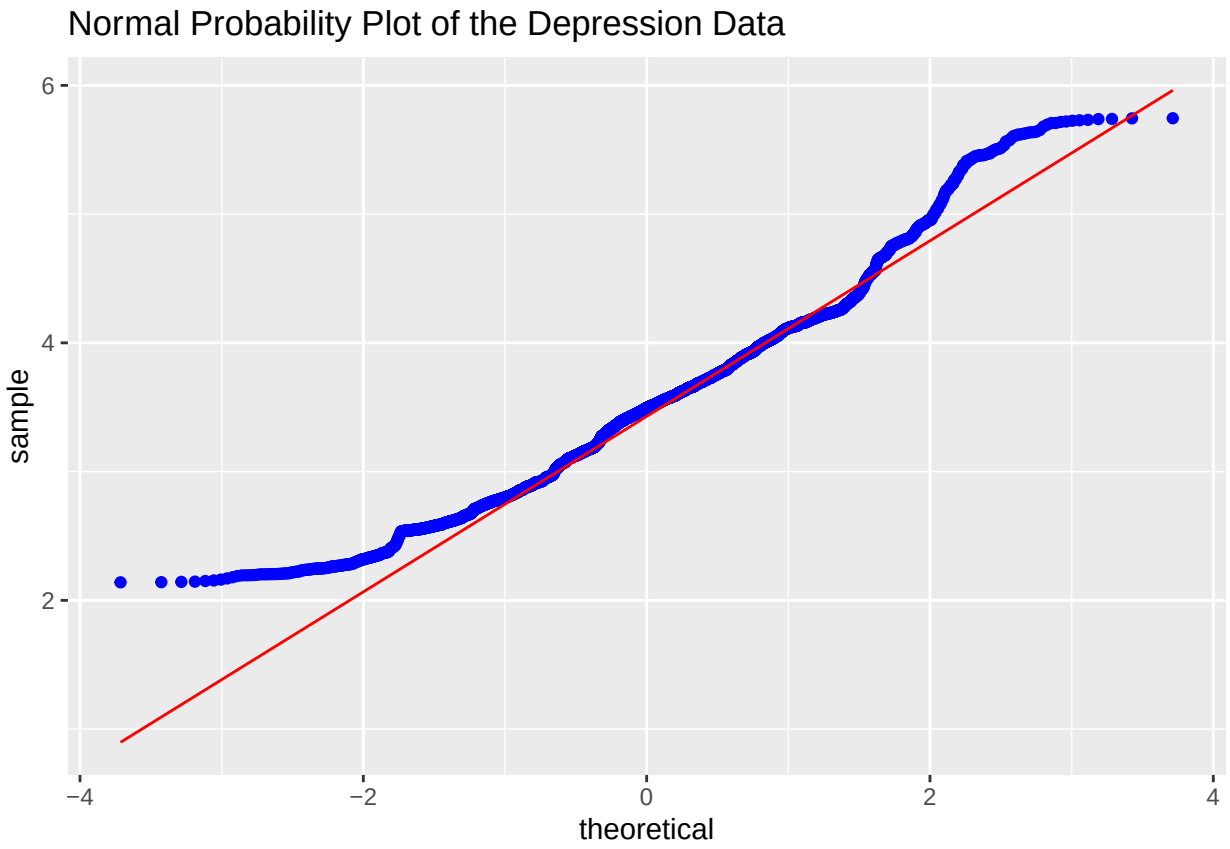
1. Normality Test
2. Correlation Matrix
3. Scatter Plots
4. Hypothesis Test
5. Linear Regression

Question 2A:

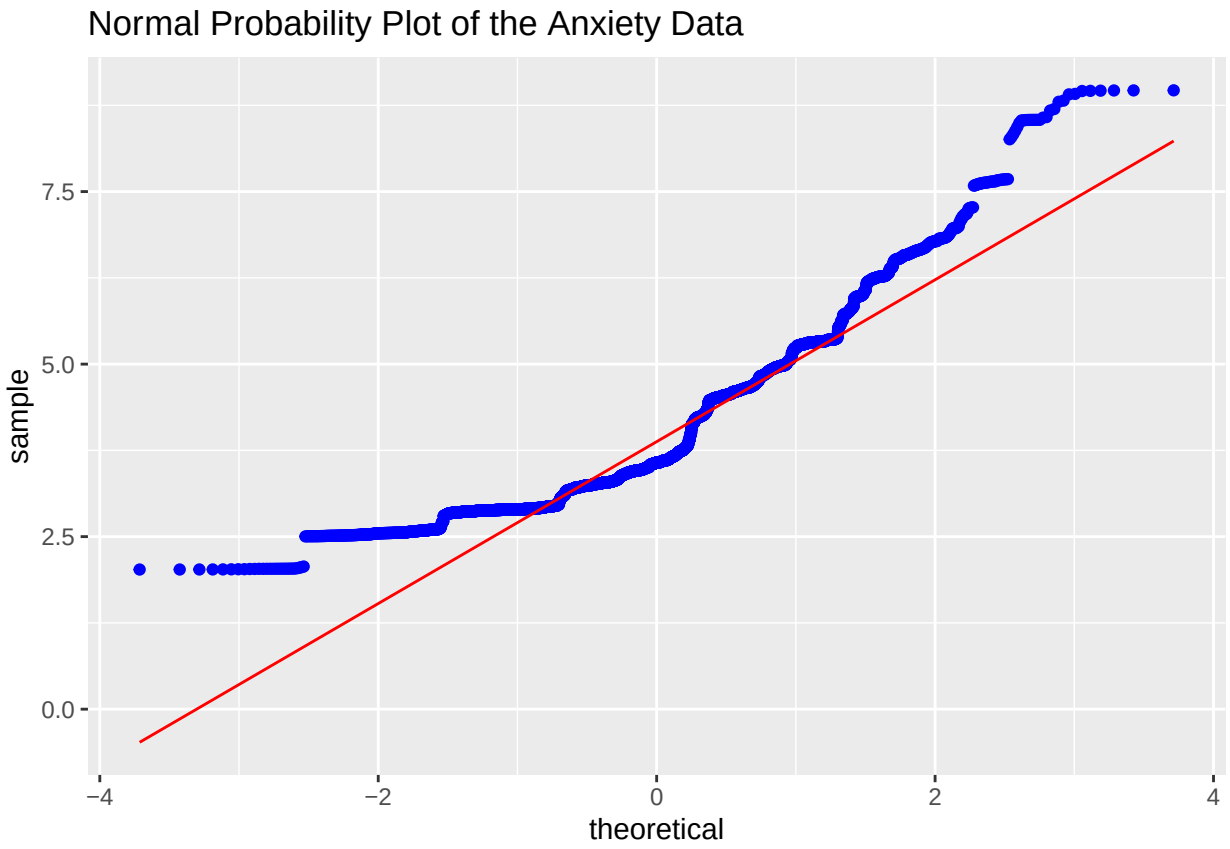
Are there different behaviors (i.e., drug use) that correlate with the prevalence of mental disorders? (Use Variable - data3)

Test General Normality of the Data set Based off of Depression and Anxiety Samples

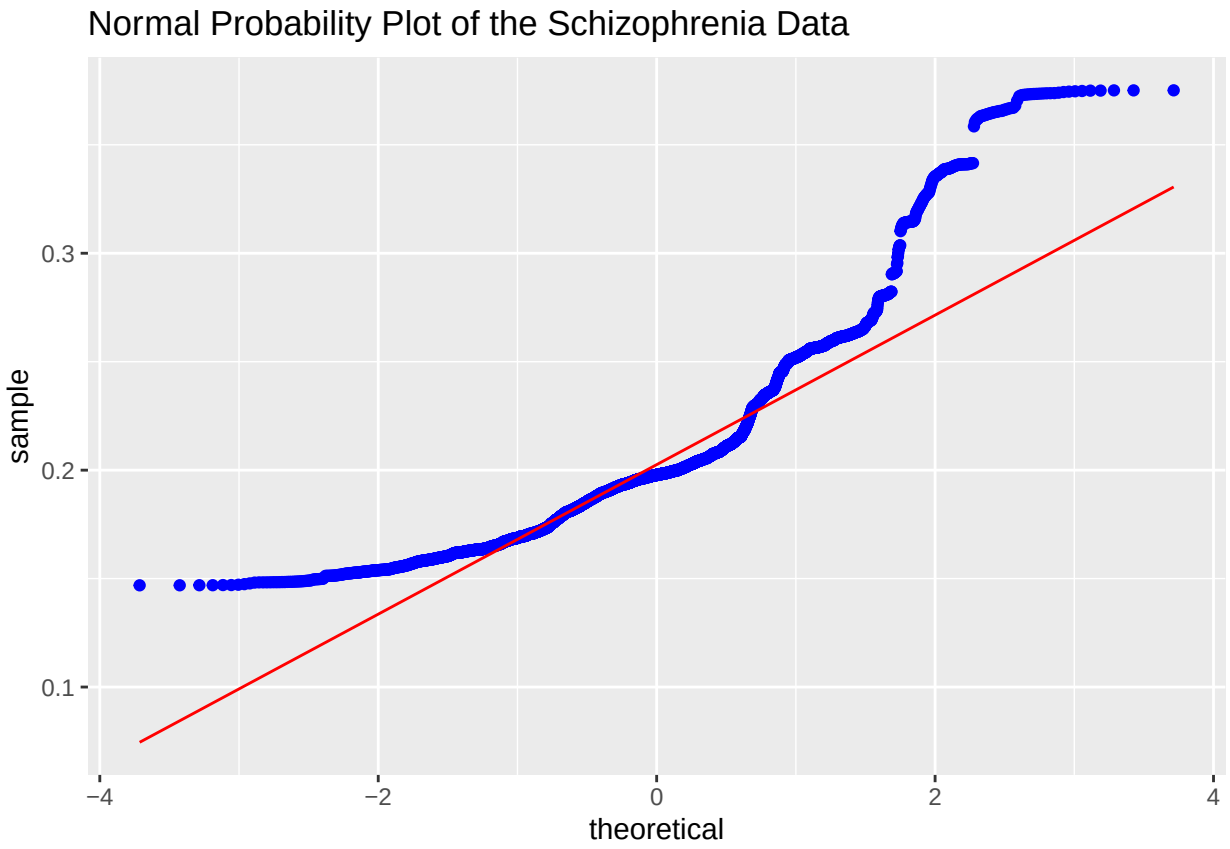
```
ggplot(data3, aes(sample = Depression)) + stat_qq(col="blue") + stat_qqline(col="red") + ggtitle("Norma
```



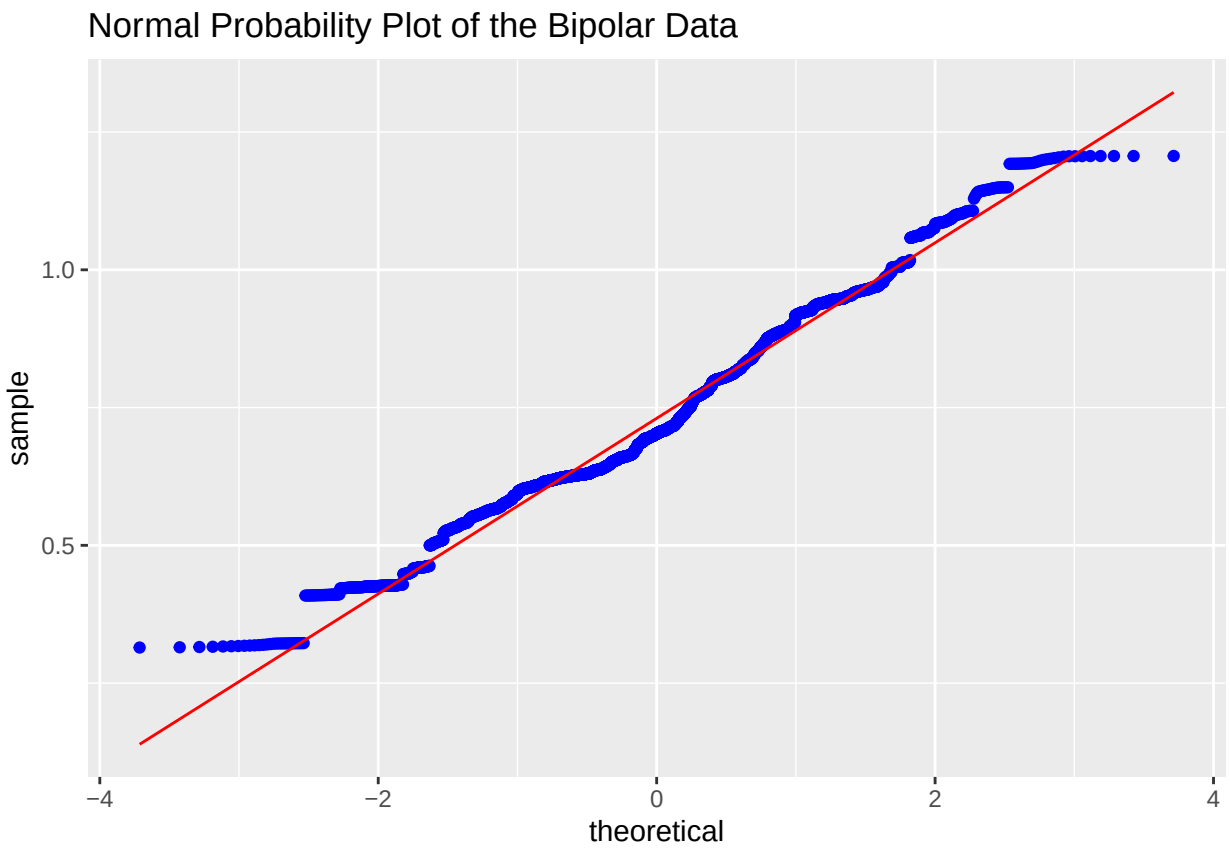

```
ggplot(data3, aes(sample = Anxiety.disorders)) + stat_qq(col="blue") + stat_qqline(col="red") + ggtitle
```




```
ggplot(data3, aes(sample = Schizophrenia)) + stat_qq(col="blue") + stat_qqline(col="red") + ggtitle("No
```




```
ggplot(data3, aes(sample = Bipolar_disorder)) + stat_qq(col="blue") + stat_qqline(col="red") + ggtitle(
```



What we can infer from these four sample datasets that they are approximately normally distributed with n being >= 25.

Test Correlation (Select One From Each Group With the Highest Correlation To Be Examined Further)

```
cor_matrix.df3 = data.frame(cor(data3[, unlist(lapply(data3,is.numeric))], use = "complete.obs"))
cor_matrix.df3 <- cor_matrix.df3 %>% mutate(across(is.numeric, round, digits=2))
cor_matrix.df3$names = rownames(cor_matrix.df3) #Create a correlation matrix for only the numeric data
```

NOTE: We only want to know the relationships between the mental health issues and the potentially self-destructive behaviors.

NOTE: For the purposes of this assignment we will treat each mental health issue as separate.

The relationship between Depression and Drug Usage had the highest correlation (0.3159474) which indicates a positive linear relationship among the two variables. It can be inferred that the two variables are leaning towards a slightly powerful linear relationship. (In the Depression Category).

The relationship between Anxiety and Eating Disorder had the highest correlation (0.6764022) which indicates a very strong positive linear relationship among the two variables. It can be inferred that the two variables have a powerful linear relationship. (In the Anxiety Category).

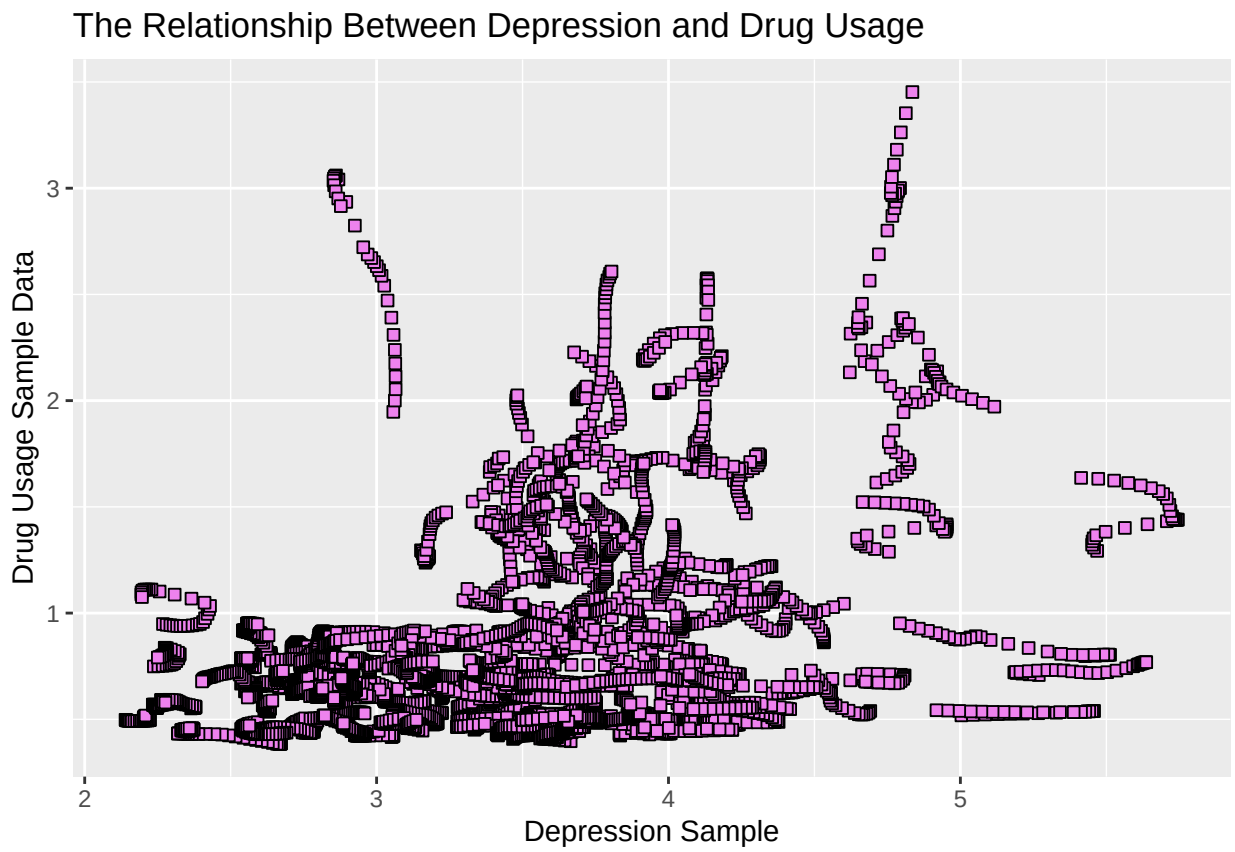
The relationship between Schizophrenia and Eating Disorder had the highest correlation (0.68469612) which indicates a very strong positive linear relationship among the two variables. It can be inferred that the two variables have a powerful linear relationship. (In the Schizophrenia Category).

The relationship between Bipolar Disorder and Eating Disorder had the highest correlation (0.70401875) which indicates a very strong positive linear relationship among the two variables. It can be inferred that the two variables have a powerful linear relationship. (In the Schizophrenia Category).

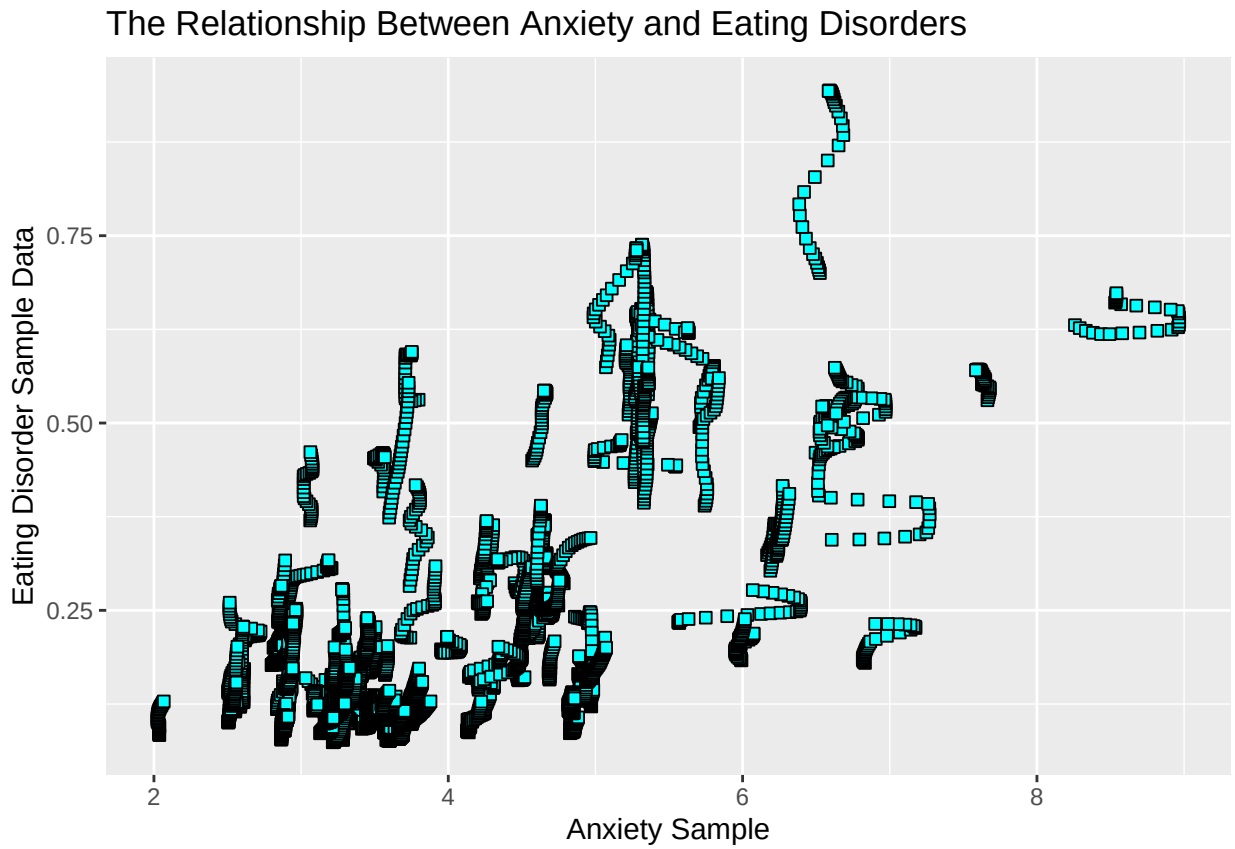
Represent the Strength of the Correlations Using Scatter Plots

Figure 9.1 / 9.2 / 9.3 / 9.4: Relationships Between Variables

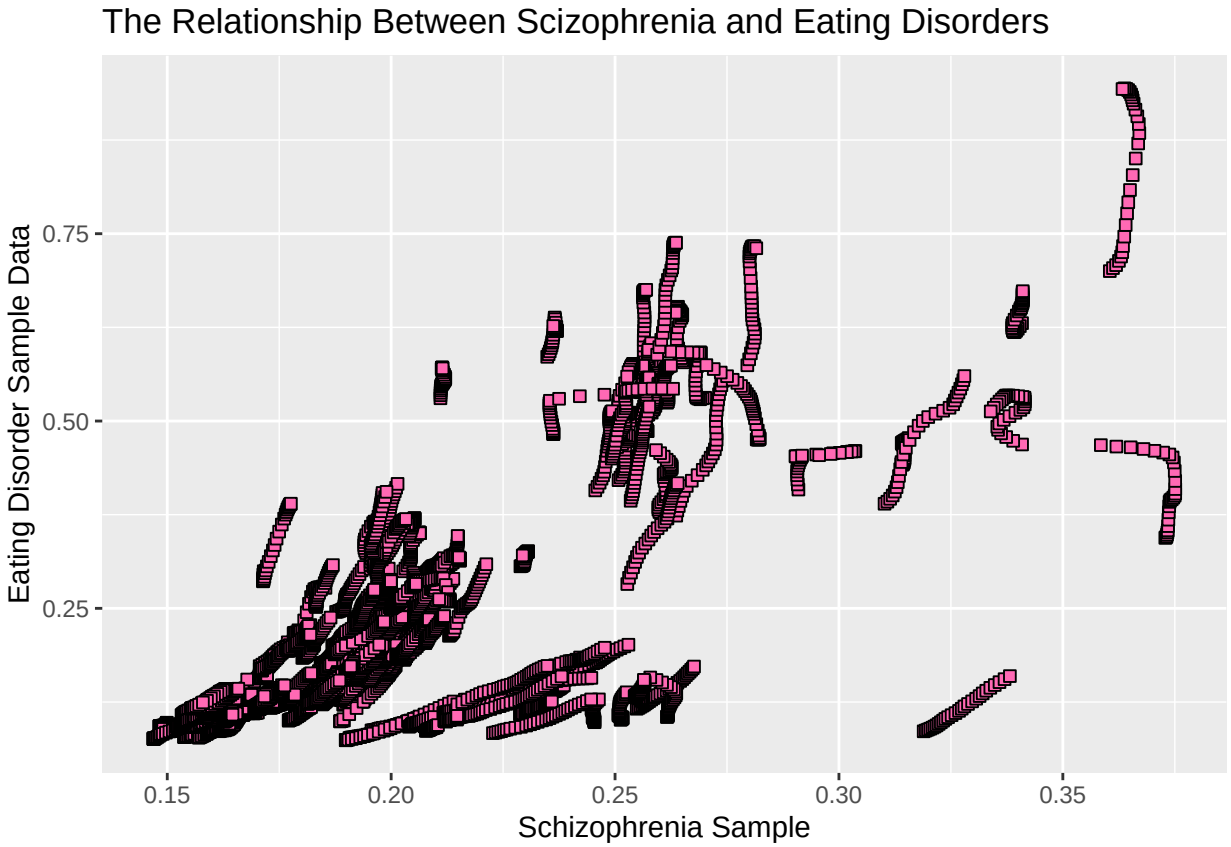
```
ggplot(data3, aes(x = Depression, y = Drug_use_disorders)) + geom_point(size=2, shape=22, fill="violet")
```



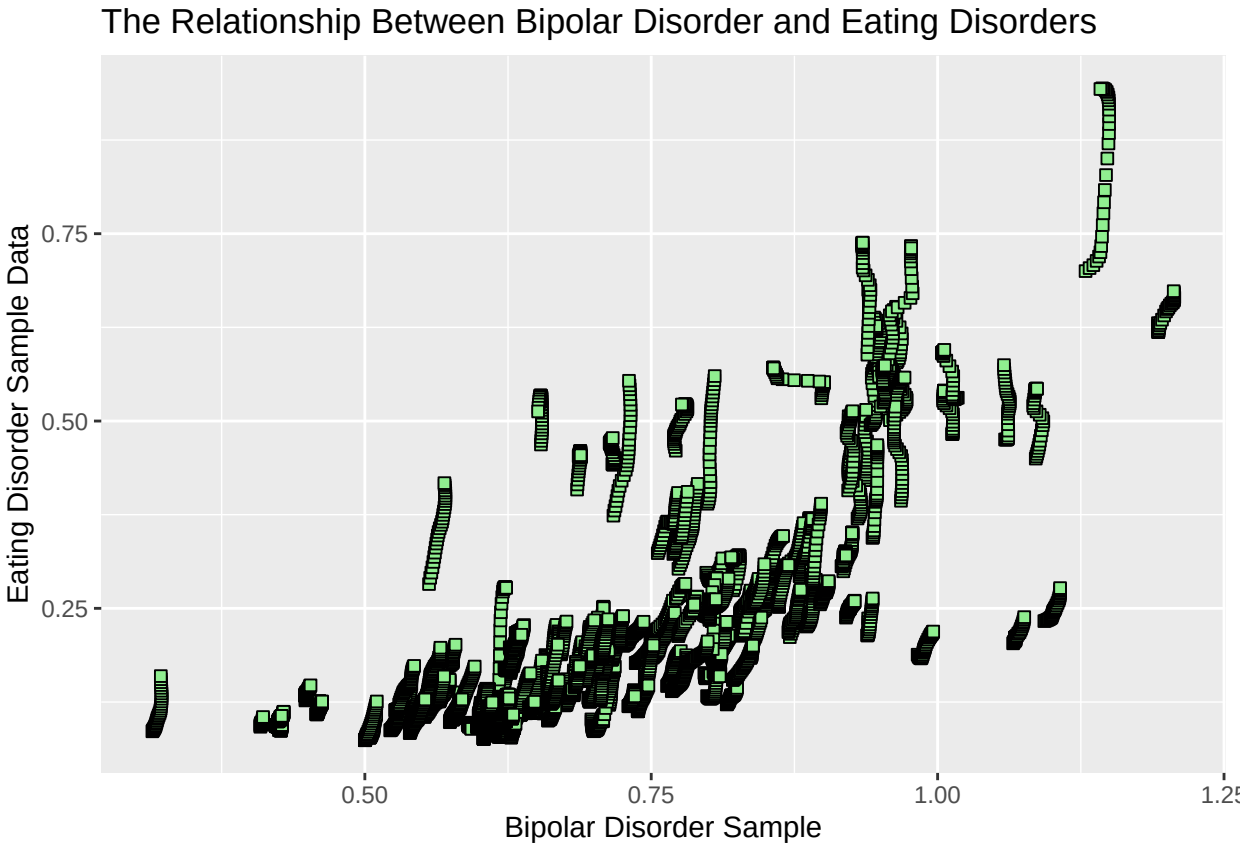

```
ggplot(data3, aes(x = Anxiety.disorders, y = Eating_disorders)) + geom_point(size=2, shape=22, fill="cya
```




```
ggplot(data3, aes(x = Schizophrenia, y = Eating_disorders)) + geom_point(size=2, shape=22, fill="hot pi
```




```
ggplot(data3, aes(x = Bipolar_disorder, y = Eating_disorders)) + geom_point(size=2, shape=22, fill="lig
```



We can observe from the first visualization that although the relationship between Depression and Drug Usage is centered more towards the middle there are still data points moving towards the upper right corner of the graph (indicating a slight positive relationship).

We can observe from the second visualization that although the relationship between Anxiety and Eating Disorders is less centered towards the middle there are a lot more data points moving towards the upper right corner of the graph (indicating a strong positive relationship).

We can observe from the third visualization that although the relationship between Schizophrenia and Eating Disorders is less centered towards the middle there are a lot more data points moving towards the upper right corner of the graph (indicating a strong positive relationship).

We can observe from the fourth visualization that although the relationship between Bipolar and Eating Disorders is less centered towards the middle there are a lot more data points moving towards the upper right corner of the graph (indicating a strong positive relationship).

Question 2B:

A study found that the sample population's Depression Rate in 2017 was 3.4, for Anxiety Rate 3.8, for Schizophrenia 0.3, and for Bipolar Disorder 0.6 (https://ourworldindata.org/mental-health). Create a Statistical Hypothesis Test to see if the average Depression Rate was greater than this study's claim; that is, what is the CI and p-value of the data supporting or rejecting this claim?

Ho: 2017 Rates From Sample Data = Rates From Study

HA: 2017 Rates From Sample Data != Rates From Study

```r
data3_2017 <- data3 %>% #Filter the dataset to find all the depression rates in 2017
  group_by(Depression) %>%
  filter(Year == "2017") #Concatenate the data for depression data in 2017

mu = 3.4 #Population parameter

t.test(data3_2017$Depression, mu=mu, alternative = "two.sided") #Find the test statistic p-value
```

```
##
##  One Sample t-test
##
## data:  data3_2017$Depression
## t = 0.95321, df = 174, p-value = 0.3418
## alternative hypothesis: true mean is not equal to 3.4
## 95 percent confidence interval:
##  3.352057 3.537509
## sample estimates:
## mean of x
##  3.444783
```

```r
data3_2017 <- data3 %>% #Filter the dataset to find all the depression rates in 2017
  group_by(Anxiety.disorders) %>%
  filter(Year == "2017") #Concatenate the data for depression data in 2017

mu = 3.8 #Population parameter

t.test(data3_2017$Anxiety.disorders, mu=mu, alternative = "two.sided") #Find the test statistic p-value
```

```
##
##  One Sample t-test
##
## data:  data3_2017$Anxiety.disorders
## t = 2.2785, df = 174, p-value = 0.02391
## alternative hypothesis: true mean is not equal to 3.8
## 95 percent confidence interval:
##  3.826852 4.174584
## sample estimates:
## mean of x
##  4.000718
```

```r
data3_2017 <- data3 %>% #Filter the dataset to find all the depression rates in 2017
  group_by(Schizophrenia) %>%
  filter(Year == "2017") #Concatenate the data for depression data in 2017

mu = 0.3 #Population parameter

t.test(data3_2017$Schizophrenia, mu=mu, alternative = "two.sided") #Find the test statistic p-value
```

```
##
##  One Sample t-test
##
```

```
## data:  data3_2017$Schizophrenia
## t = -28.904, df = 174, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.3
## 95 percent confidence interval:
##  0.2040189 0.2162891
## sample estimates:
## mean of x
##  0.210154
```

```
data3_2017 <- data3 %>% #Filter the dataset to find all the depression rates in 2017
  group_by(Bipolar_disorder) %>%
  filter(Year == "2017") #Concatenate the data for depression data in 2017

mu = 0.6 #Population parameter

t.test(data3_2017$Bipolar_disorder, mu=mu, alternative = "two.sided") #Find the test statistic p-value
```

```
##
##  One Sample t-test
##
## data:  data3_2017$Bipolar_disorder
## t = 11.076, df = 174, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.6
## 95 percent confidence interval:
##  0.7092762 0.7566640
## sample estimates:
## mean of x
## 0.7329701
```

From the first statistical test since we got a p-value of 0.042 $>$ 0.05 we can say with 95% confidence that there is statistical evidence to suggest that we should accept the Null Hypothesis. That is we can say that the Depression Rate in our Sample Dataset is larger than the rate provided by the Study. Specifically, with 95% confidence we can say the difference in how much larger the rate differs by is found within the range of 3.35 and 3.53.

From the second statistical test since we got a p-value of 2.356 x 10 ^ -13 $<$ 0.05 we can say with 95% confidence that there is statistical evidence to suggest that we should reject the Null Hypothesis. That is we can say that the Depression Rate in our Sample Dataset is larger than the rate provided by the Study. Specifically, with 95% confidence we can say the difference in how much larger the rate differs by is found within the range of 3.80 and 4.12.

From the third statistical test since we got a p-value of 2.2 x 10 ^ -16 $<$ 0.05 we can say with 95% confidence that there is statistical evidence to suggest that we should reject the Null Hypothesis. That is we can say that the Depression Rate in our Sample Dataset is larger than the rate provided by the Study. Specifically, with 95% confidence we can say the difference in how much larger the rate differs by is found within the range of 0.21 and 0.22.

From the fourth statistical test since we got a p-value of 2.2 x 10 ^ -16 $<$ 0.05 we can say with 95% confidence that there is statistical evidence to suggest that we should reject the Null Hypothesis. That is we can say that the Depression Rate in our Sample Dataset is larger than the rate provided by the Study. Specifically, with 95% confidence we can say the difference in how much larger the rate differs by is found within the range of 0.70 and 0.74.

Explanation for Q2A:

The correlations that were conducted here were to focalize our larger dataset into a few key variables that would help with the remainder of this analysis. That is dividing the variables into 2 categories mental-illnesses (Depression / Anxiety / Schizophrenia / Bipolar Disorder) and behavioral (Alcohol Usage / Drug Usage / Eating Disorders). When analyzing the relationships, we find that the most common behaviors associated with each mental-illness are indicative of all the major symptoms that individual will be exhibiting. For example, if they are Schizophrenic and have an eating disorder issue their most probable triggers are both psychological and even physical. When we identify all of the symptoms that individual patients face it becomes much more complex whilst simultaneously more expansive of the types of treatments (prescriptions and conditioning) that will be required to help lessen the symptoms so that the individual can function at a relative normal. The combinations of these mental illnesses and behavioral patterns also make it easier to prevent future self-destructive actions (given the right supervision and prevention) but with obviously larger added costs.

Explanation for Q2B:

The 4 hypothesis tests conducted were incorporated into the analysis to help identify how accurate and reliable the overall 'Mental Health Substance' dataset was. To test its accuracy I utilized the sample means found from another study and by using the function t.test() was able to conduct the hypothesis testing for the Depression / Anxiety / Schizophrenia / Bipolar Disorders mental illnesses. From the aforementioned analysis we found that in each hypothesis test that the sample rates were slightly larger than each of the respective study rates. From this we can conclude that although there wasn't an exact match between the two, our sample data still holds quite a lot of accuracy and validity; as having similar values to the much larger study data sets presents a more accurate inference of our sample data on the larger global population. Having accurate data allows us to ensure that the insights we derive from the sample data is actually applicable within the real world and that it can be used to help enhance the treatment processes associated with diagnosing mental-illnesses (and their respective prescription models).;

Question 3:

Which Disorders in our data set can help predict the Anxiety Disorder Rates?

Intro:

We are interested in seeing if a prevalence of one mental health disorder can help predict the prevalence of Anxiety disorders in a population using a linear model. In other words, are Anxiety disorders and other mental health disorders associated in some way?

We will try to answer this by first looking at the linear relationships between variables in our data set and Anxiety disorders through Pearson's correlation and visualizations. Using the variables with the highest linear correlations, we will build a linear regression model and test that model to see if it is valid and has any predictive power.

R^2 is defined as the proportion of variation in the dependent variable that is explained, or can be predicted, by the independent variable in a regression. R^2 can be calculated by squaring the correlation coefficient.

```
# Load the data set
data <- read.csv("Mental Health Merged.csv")
```

Adding "Development" column. Developed countries are defined as countries with HDI > 0.72.

```
data$Development = ifelse(data$hdi2019>0.72, "Developed", "Undeveloped")
data_developed = data[(data$Development == "Developed"), ]
data_undeveloped = data[(data$Development == "Undeveloped"), ]
```

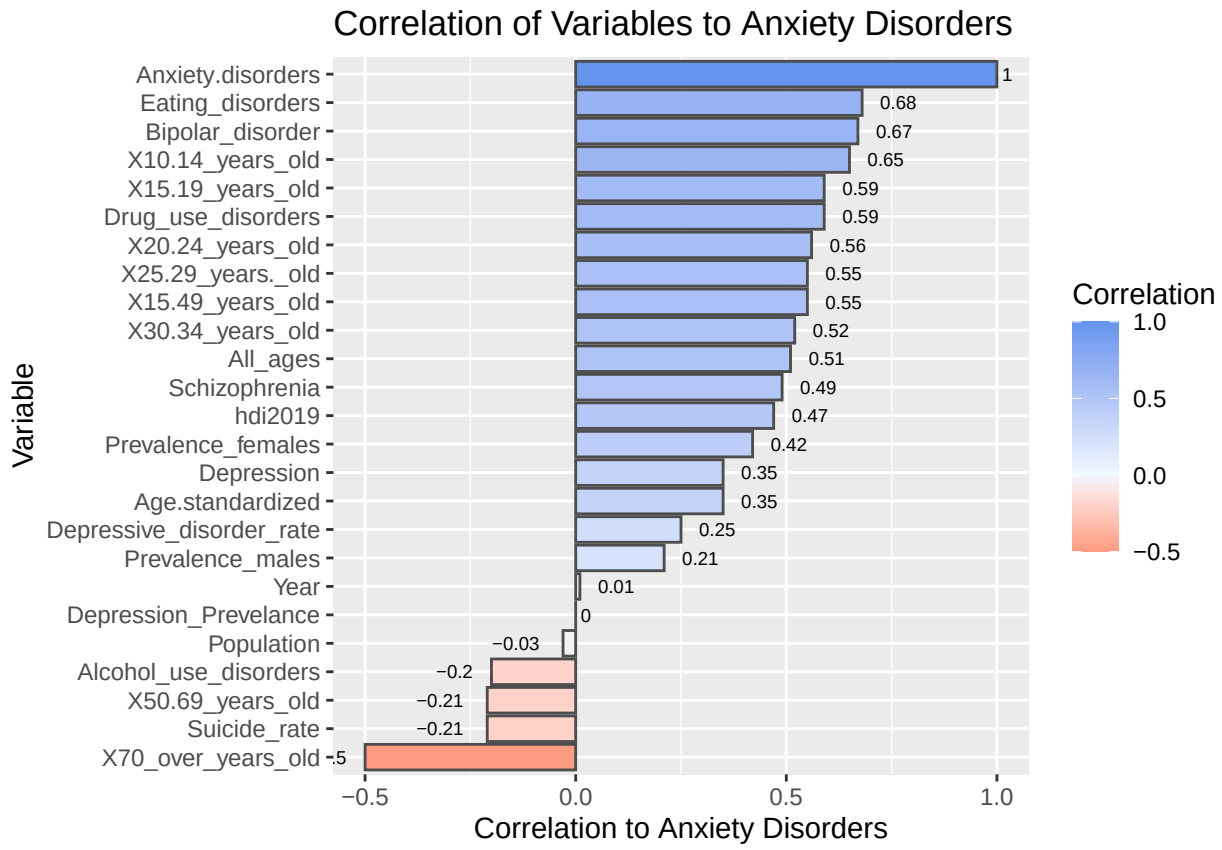
Calculating the correlation matrix:

```
cor_matrix.df = data.frame(cor(data[, unlist(lapply(data,is.numeric))], use = "complete.obs"))
cor_matrix.df <- cor_matrix.df %>% mutate(across(is.numeric, round, digits=2))
cor_matrix.df$names = rownames(cor_matrix.df)
```

A correlation heat map is difficult to see due to the amount of variables. Since we are interested in the correlations with Suicide Rate, we can extract those and visualize them.

Figure 10.1: Correlation of Variables with Anxiety Disorder Rates

```
ggplot(cor_matrix.df, aes(y = reorder(names, Anxiety.disorders), x = Anxiety.disorders, fill = Anxiety.
```



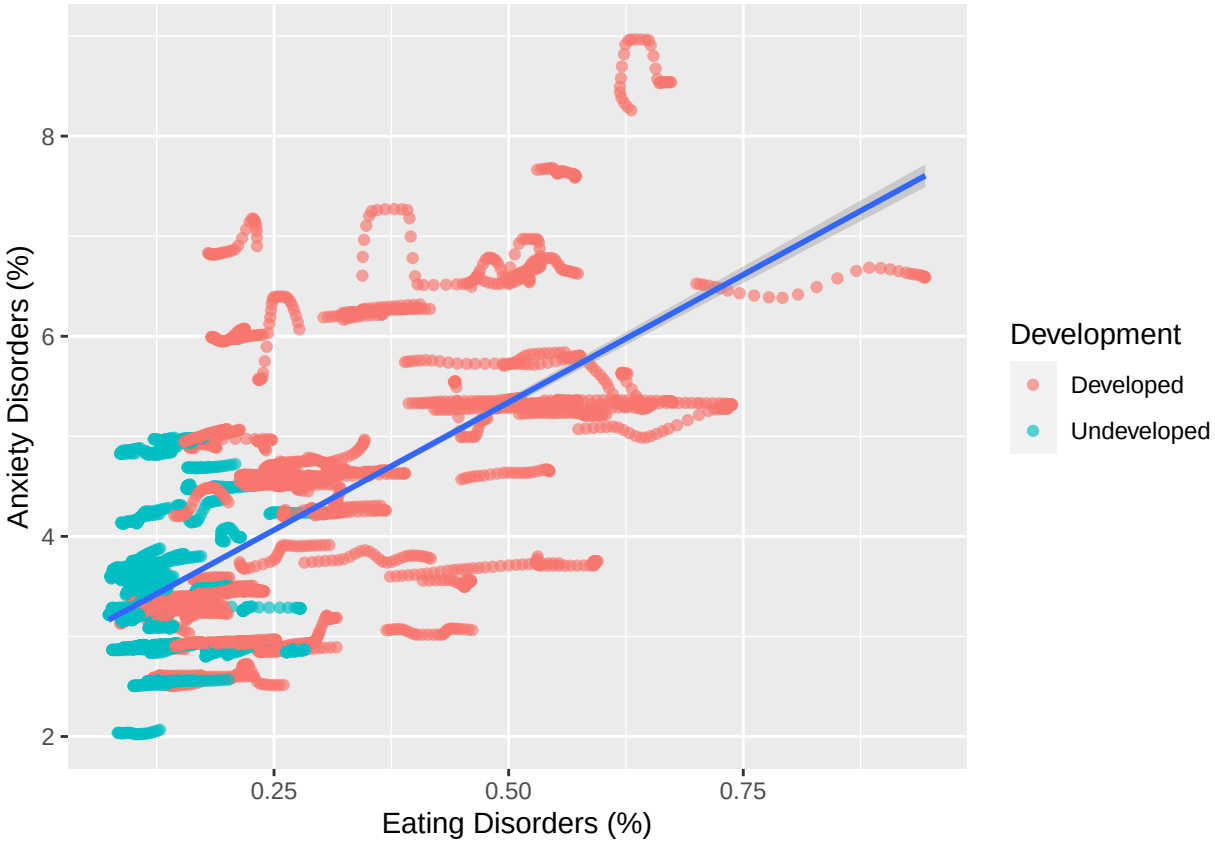
From our visualization, we can see that Eating disorders, Bipolar disorders and drug use disorders all have moderately high correlation coefficients. We will look into these disorders as possible variables that can help predict Anxiety disorders.

Eating Disorders:

Figure 10.2: Relationship between Anxiety and Eating Disorders

```
ggplot(data)+geom_point(aes(y = Anxiety.disorders, x = Eating_disorders, color = Development), alpha =
```

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



From the scatter plot of Anxiety disorders and Eating disorders, we see these "smears" of data points for each country where anxiety disorders stay relatively constant over time while eating disorders increase. Because of this, we are doubtful that eating disorders have any significant predictive power of anxiety disorders.

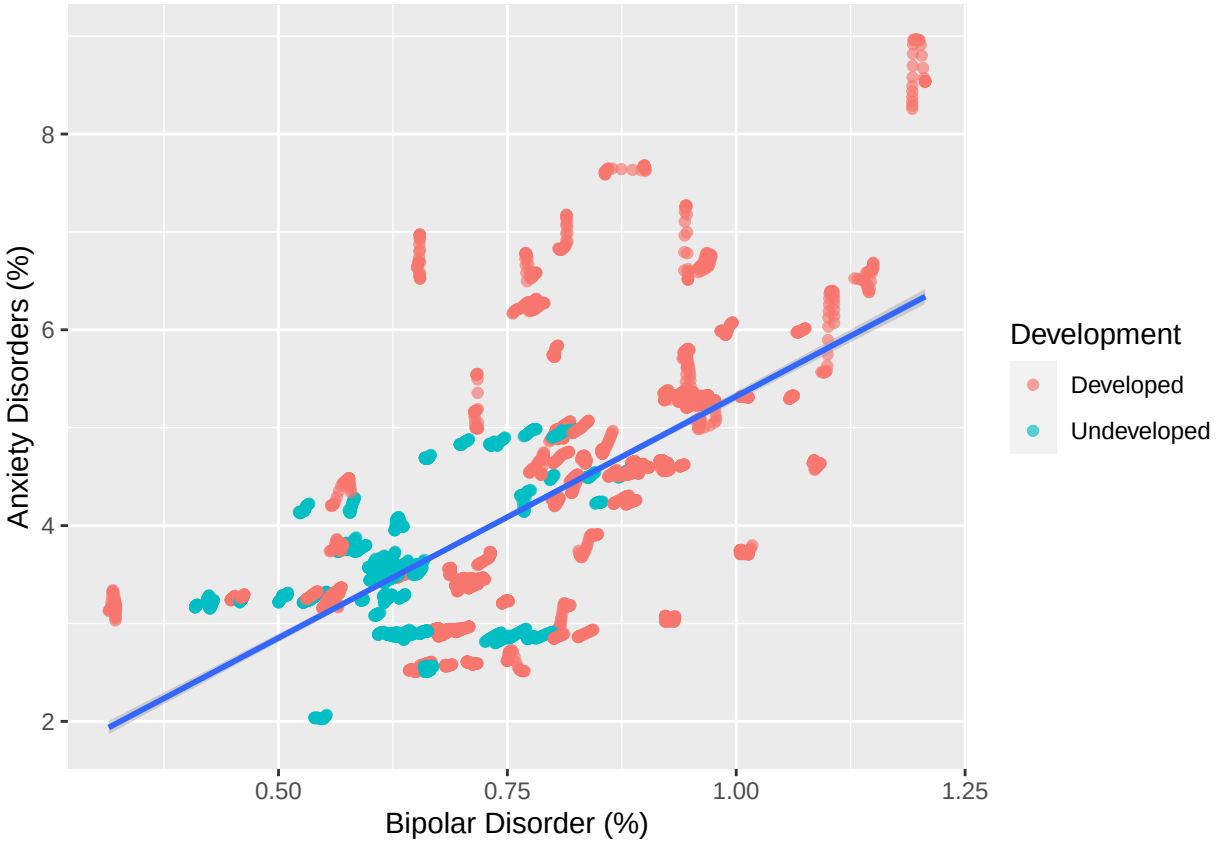
One interesting find is that the smears are much longer for developed countries compared to undeveloped countries, meaning developed countries have a higher prevalence of eating disorders.

Bipolar Disorder:

Figure 10.3: Relationship between Anxiety and Bipolar Disorder

```
ggplot(data)+geom_point(aes(y = Anxiety.disorders, x = Bipolar_disorder, color = Development), alpha = (
```

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



We can see from the scatter plot that there seems to be a positive linear relationship between Anxiety and Bipolar disorder. We grouped the data by developed/undeveloped to see if we see anything interesting. It looks like developed nations have higher rates of anxiety and bipolar disorders.

We will go ahead and create a linear regression model to predict anxiety disorder rates with bipolar disorder rates.

```
predicted_anxiety_bipolar = lm(Anxiety.disorders ~ Bipolar_disorder, data=data)
predicted_anxiety_bipolar$coef
```

```
##      (Intercept) Bipolar_disorder
##        0.3850484        4.9355663
```

Our linear regression equation is: $\hat{Anxiety} = 0.38505 + 4.9356 Bipolar$. The R-squared for this model is $0.67^2 = 0.45$. If there is a valid linear relationship between Bipolar Disorder and Anxiety Disorder, then we can say about 45% of the variance in the prevelance Anxiety Disorders can be explain by the prevalence in Bipolar Disorder.

Meaning of the intercept: If there was no Bipolar disorders, there would still be approx. 0.385% prevalence of Anxiety Disorders.

Meaning of the coefficient: For every 1% increase in Bipolar disorders, Anxiety Disorders increase by 4.9356%.

Before we can use this model to make predictions, we must check two conditions to ensure it is valid.

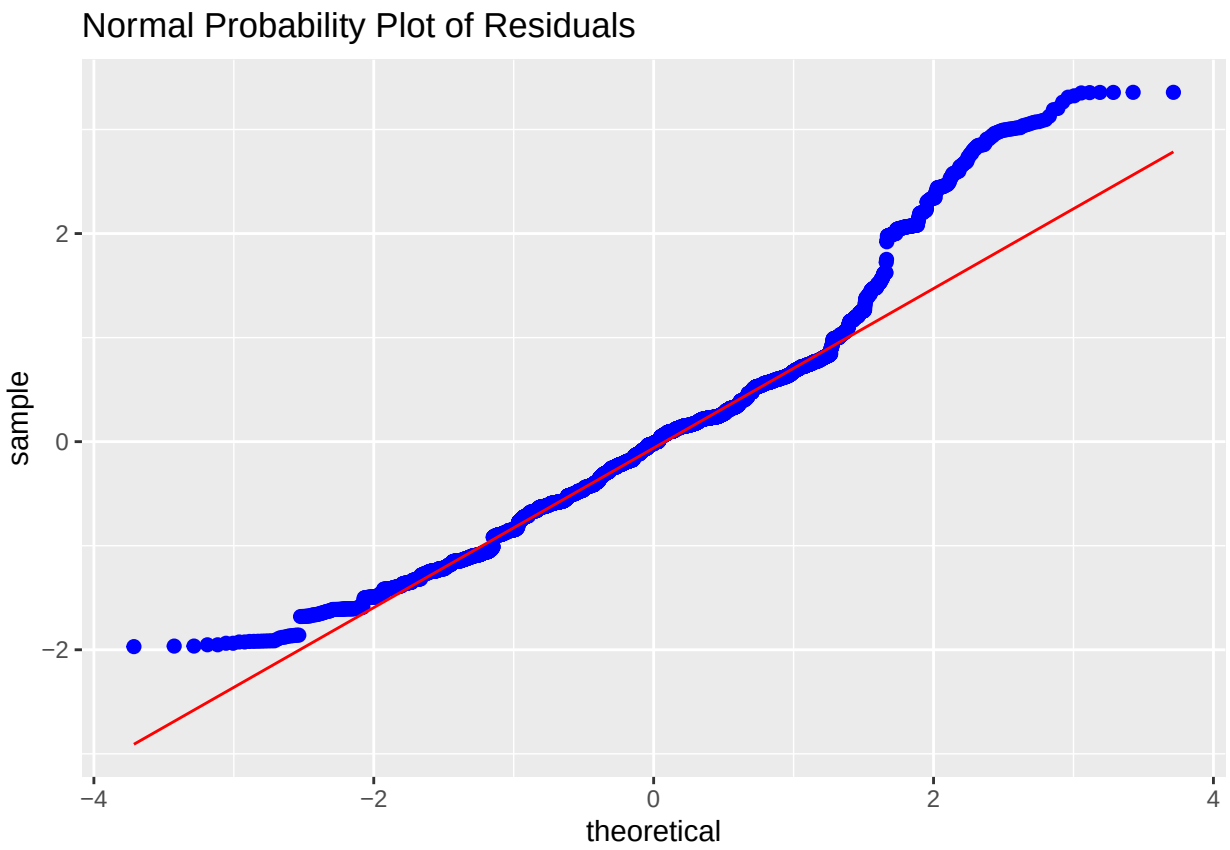
1. Normality of residuals: The dependent variable (Anxiety Disorders) must be normally distributed with a mean of μ and standard deviation of σ. To check this we will plot a stat_qq plot of the residuals since $e_i = y_i - \hat{y}_i$, if y is normally distributed, so will the residuals.

2. Homoscedasticity: For each distinct value of the independent variable (Bipolar Disorder), the dependent variable (Anxiety Disorder) has the same standard deviation σ. To check this, we will plot a scatter plot of the fitted values and the residuals.

```
predicted_anxiety_bipolar.fit = predicted_anxiety_bipolar$fitted.values
ei_anxiety_bipolar = predicted_anxiety_bipolar$residuals
diagnostic_bipolar.df = data.frame(predicted_anxiety_bipolar.fit, ei_anxiety_bipolar)
```

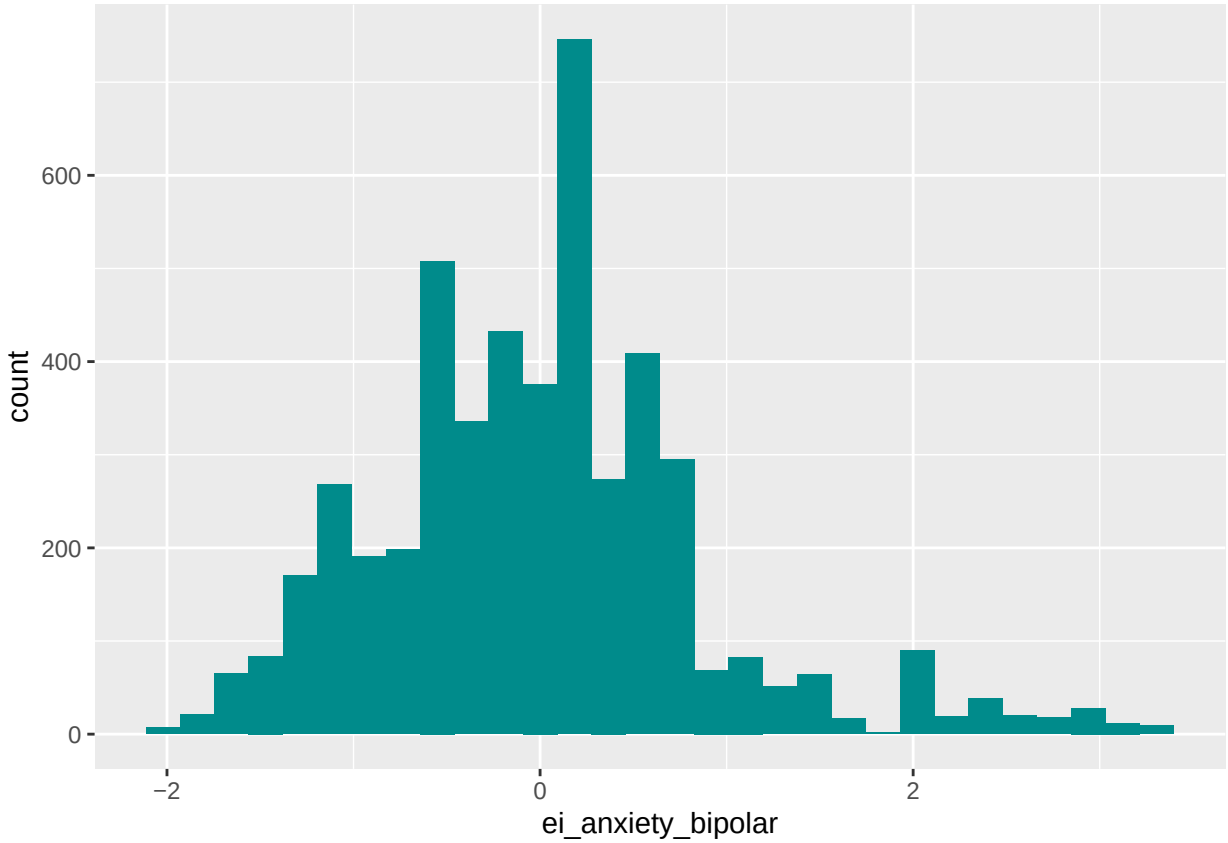
Normality of residuals:

```
ggplot(diagnostic_bipolar.df, aes(sample = ei_anxiety_bipolar)) + stat_qq(size=2, col='blue') + stat_qq
```




```
ggplot(diagnostic_bipolar.df, aes(x=ei_anxiety_bipolar)) + geom_histogram(fill = "darkcyan")
```

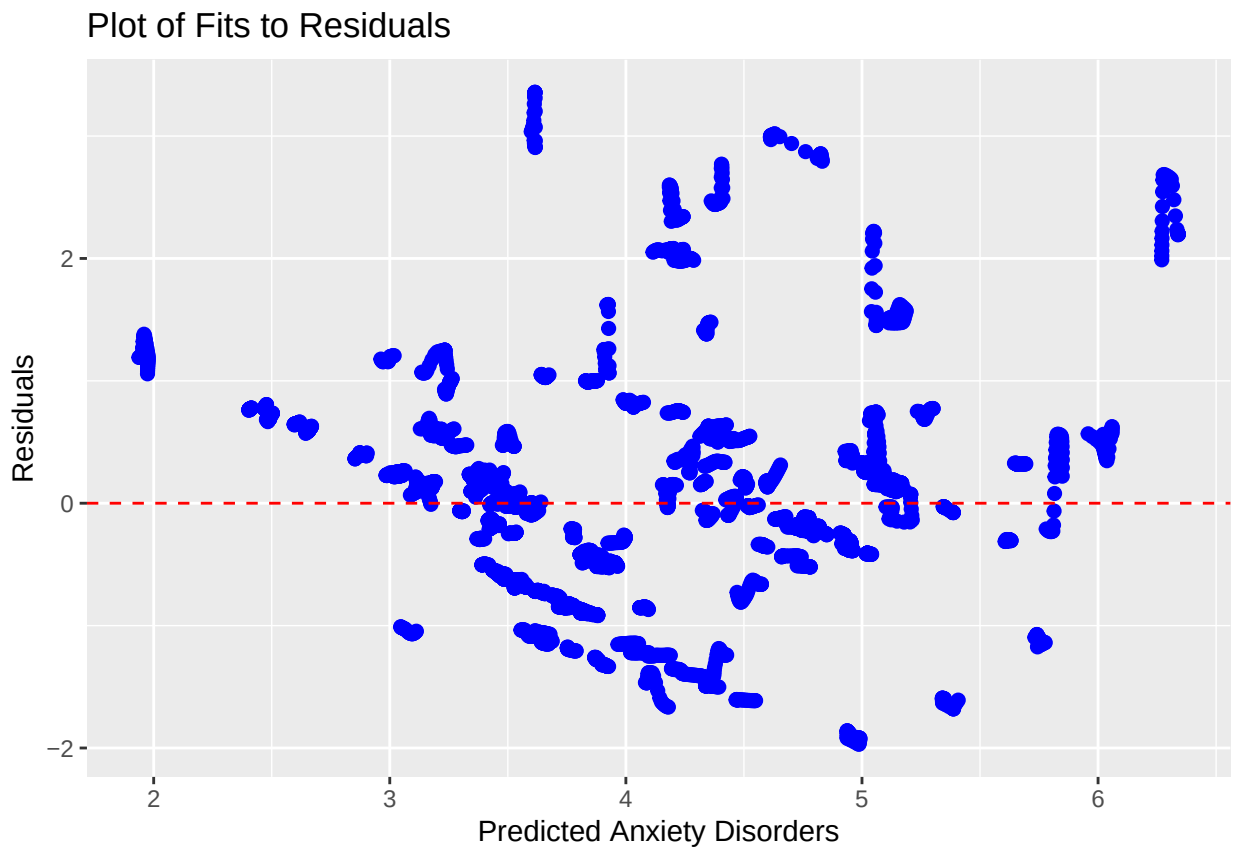
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Looking at our normal probability plot, the residuals seem to be distributed normal enough as the majority of the residuals fall on the line. Therefore we will say that the normality of residuals condition holds.

Homoscedasticity:

```
ggplot(diagnostic_bipolar.df, aes(x = predicted_anxiety_bipolar.fit, y = ei_anxiety_bipolar)) +  geom_p
```



From the above plot, we can see that the majority of the data points are evenly distributed between 2 and -2. Therefore the condition of homoscedasticity holds.

Since both conditions hold, our linear regression model is valid and we can use it to predict values of Anxiety Disorders based on levels Bipolar Disorder.

We will now do a hypothesis test to test whether the regression coefficient of Bipolar Disorder is different from 0. Since we are only testing the coefficient of one variable, we will use a t-test.

Null hypothesis: $H_0 : \beta_{Biploar} = 0$

Alternative hypothesis: $H_A : \beta_{Bipolar} \neq 0$

We will set the alpha value to 0.05.

```
coef(summary(predicted_anxiety_bipolar))
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      0.3850484 0.05842326  6.590669 4.839505e-11
## Bipolar_disorder 4.9355663 0.07828096 63.049386 0.000000e+00
```

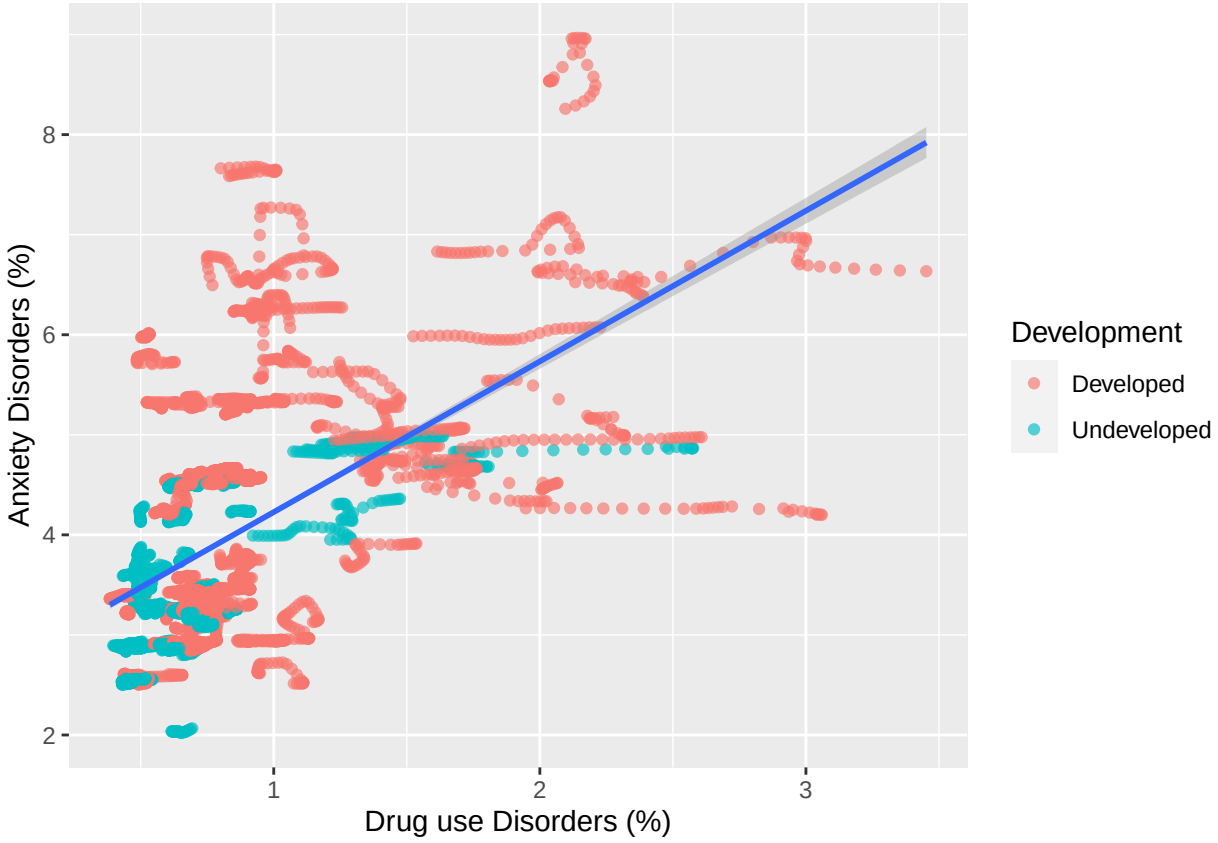
From our t-test, we get a test statistic of 63.0494 and a p-value of 0. This is smaller than the set alpha value of 0.05, so we reject our null hypothesis that the linear regression coefficient for Bipolar Disorder is 0. And therefore, we can conclude that the Prevalence of Anxiety Disorders can be expressed as a linear function of the prevalence of Bipolar Disorder, and since $\beta_{Bipolar} > 0$, we can say it is also positive.

Drug Use Disorder:

Figure 10.3: Relationship between Anxiety and Drug use Disorders

```
ggplot(data)+geom_point(aes(y = Anxiety.disorders, x = Drug_use_disorders, color = Development), alpha =
```

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



We can see from the scatter plot that there seems to be a positive linear relationship between Anxiety and Drug use disorders.

We will go ahead and create a linear regression model to predict anxiety disorder rates with drug use disorder rates.

```
predicted_anxiety_drug = lm(Anxiety.disorders ~ Drug_use_disorders, data=data)
predicted_anxiety_drug$coef
```

```
##         (Intercept) Drug_use_disorders
##            2.721955           1.505773
```

Our linear regression equation is: $\hat{Anxiety} = 2.7220 + 1.5058 Drug$ The R-squared for this model is $0.59^2 = 0.35$. If there is a valid linear relationship between Drug use Disorders and Anxiety Disorder, then we can say about 35% of the variance in the prevalence Anxiety Disorders can be explain by the prevalence in Drug use Disorders.

Meaning of the intercept: If there was no drug use disorders, there would still be approx. 2.72% prevalence of Anxiety Disorders.

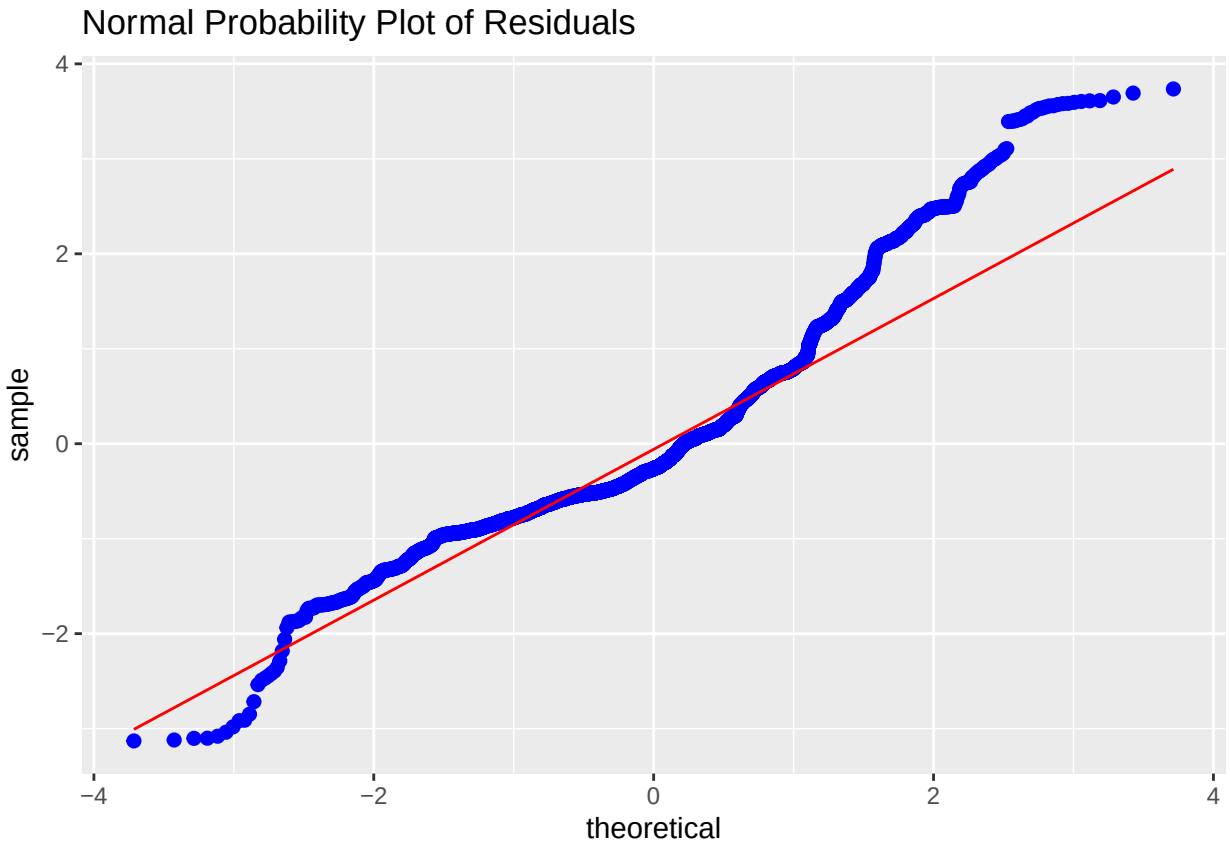
Meaning of the coefficient: For every 1% increase in drug use disorders, Anxiety Disorders increase by 1.5058%.

Before we can use this model to make predictions, we must check two conditions to ensure it is valid.

```
predicted_anxiety_drug.fit = predicted_anxiety_drug$fitted.values
ei_anxiety_drug = predicted_anxiety_drug$residuals
diagnostic_drug.df = data.frame(predicted_anxiety_drug.fit, ei_anxiety_drug)
```

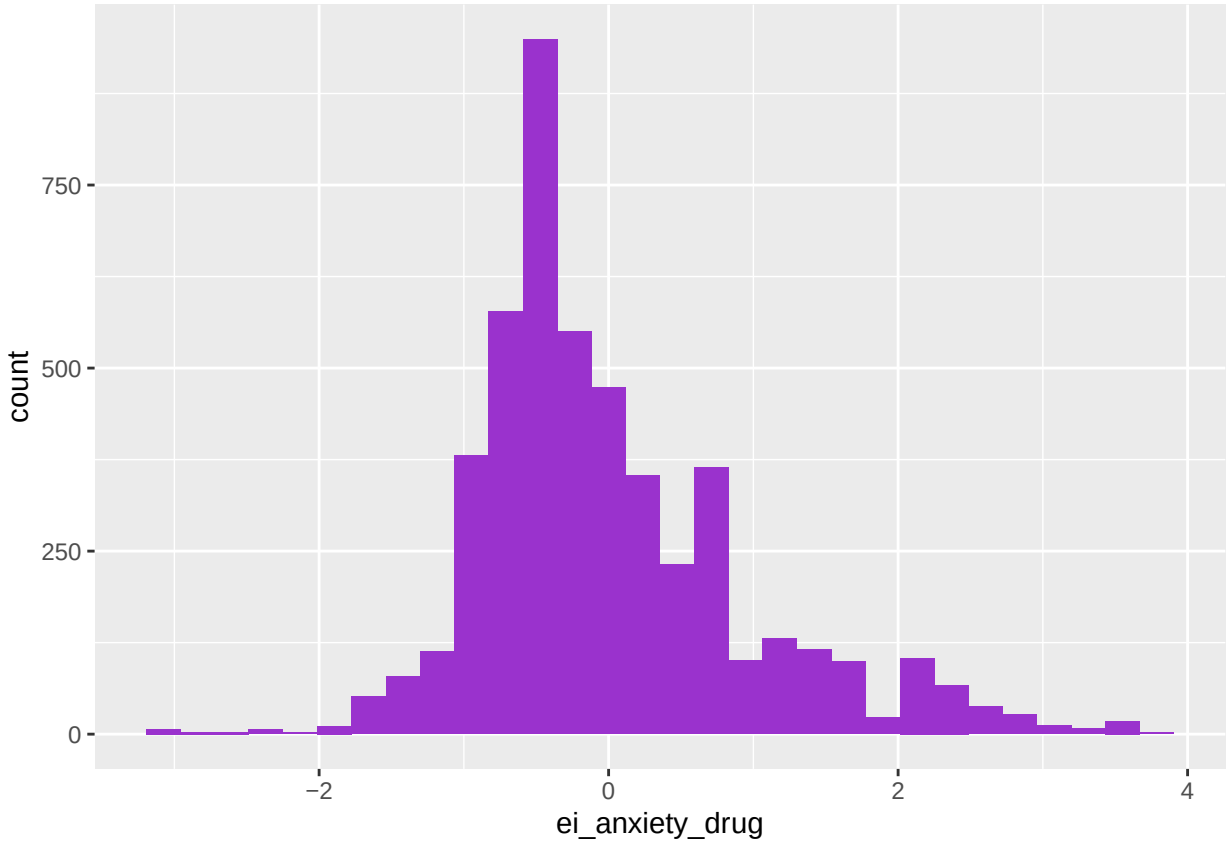
Normality of residuals:

```
ggplot(diagnostic_drug.df, aes(sample = ei_anxiety_drug)) + stat_qq(size=2, col='blue') + stat_qqline(c
```




```
ggplot(diagnostic_drug.df, aes(x=ei_anxiety_drug)) + geom_histogram(fill = "darkorchid3")
```

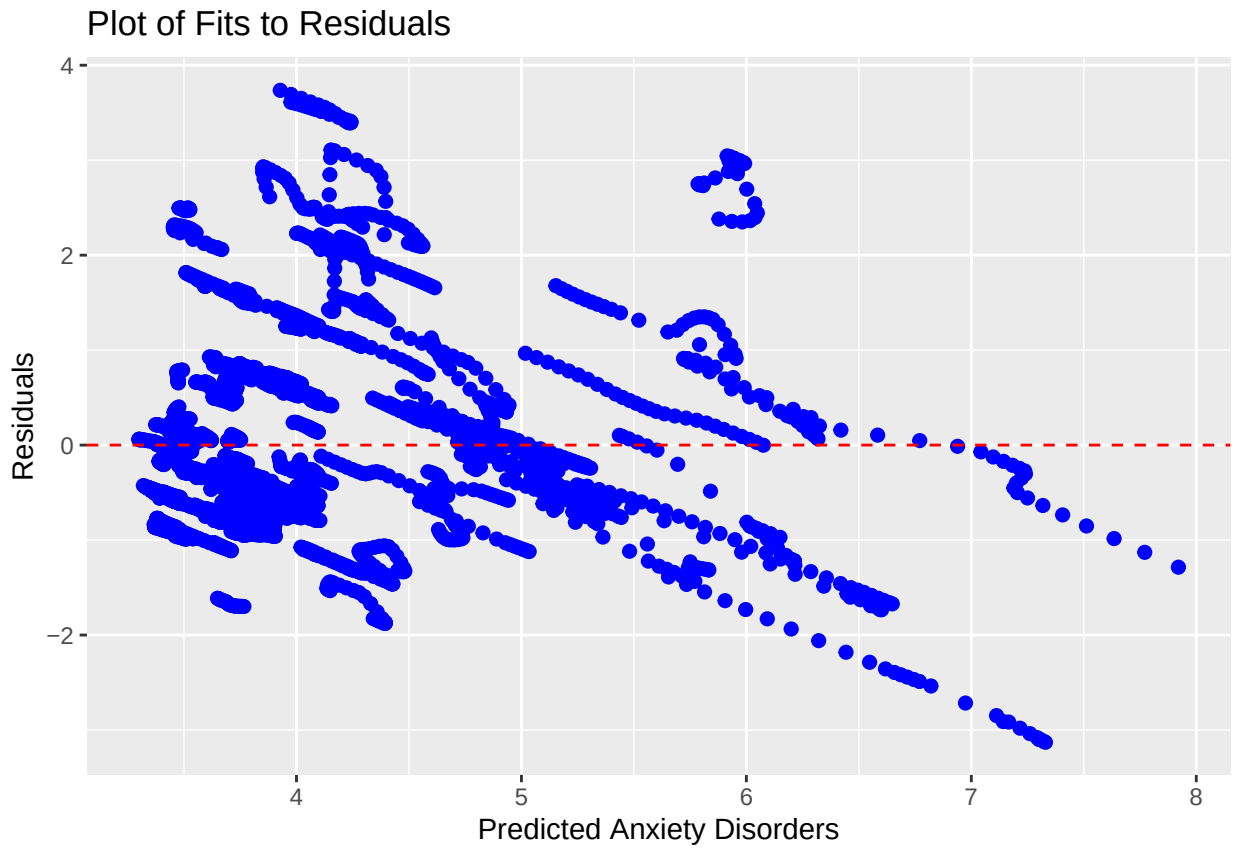
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Looking at our normal probability plot, the residuals seem to be distributed normal enough as the majority of the residuals fall on the line. Therefore we will say that the normality of residuals condition holds.

Homoscedasticity:

```
ggplot(diagnostic_drug.df, aes(x = predicted_anxiety_drug.fit, y = ei_anxiety_drug)) +  geom_point(size=
```



From the above plot, we can see that the large majority of the data points are evenly distributed between -2 and 2. Therefore we can say the condition of homoscedasticity holds.

Since both conditions hold, our linear regression model is valid and we can use it to predict values of Anxiety Disorders based on levels Drug use Disorders.

We will now do a hypothesis test to test whether the regression coefficient of Drug use Disorders is different from 0. Since we are only testing the coefficient of one variable, we will use a t-test.

Null hypothesis: $H_0 : \beta_{Drug} = 0$

Alternative hypothesis: $H_A : \beta_{Drug} \neq 0$

We will set the alpha value to 0.05.

```
coef(summary(predicted_anxiety_drug))
```

```
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        2.721955 0.02840466 95.82773        0
## Drug_use_disorders 1.505773 0.02976999 50.58025        0
```

From our t-test, we get a test statistic of 50.58025 and a p-value of 0. This is smaller than the set alpha value of 0.05, so we reject our null hypothesis that the linear regression coefficient for Bipolar Disorder is 0. And therefore, we can conclude that the Prevalence of Anxiety Disorders can be expressed as a linear function of the prevalence of Bipolar Disorder, and since $\beta_{Drug} > 0$, we can say it is also positive.

In Conclusion for Question 3:

Bipolar disorder and Drug use disorder both have a significant linear relationship to Anxiety disorders in our data. With R-squared values of 0.45 and 0.35 respectively, we can use these linear models to help predict

prevalence of Anxiety disorders. These models are not the greatest as they only explain 45% and 35% of variance in the prevalence of Anxiety disorders respectively, but it is the best that can be done with the data we have.

If we had more time or resources, transforming our data and creating/collecting new variables would be helpful in creating a more powerful simple linear model. Furthermore, creating a multiple linear regression model would be the most helpful thing in improving predictions for Anxiety disorder (when it comes to linear models at least).