STATS 382 PROJECT 1

Aditya Kapoor

2023-03-04

```
library(e1071)
```

Above is our library code needed to run some functions in this project.

Task 1

> Convert status and continent to factors in the data frame. Convert thin_youth and thin_child to ordered factors in the data frame. Please include the code in your project, but you do not need to comment on it

```
lifeexp <- read.csv("lifeexp_by_country.csv", header = TRUE)
lifeexp$status <- factor(lifeexp$status)
lifeexp$continent <- factor(lifeexp$continent)

lifeexp$thin_youth <- ordered(lifeexp$thin_youth, levels = c("Low",
    "Medium", "High"))
lifeexp$thin_child <- ordered(lifeexp$thin_child, levels = c("Low",
    "Medium", "High"))
```

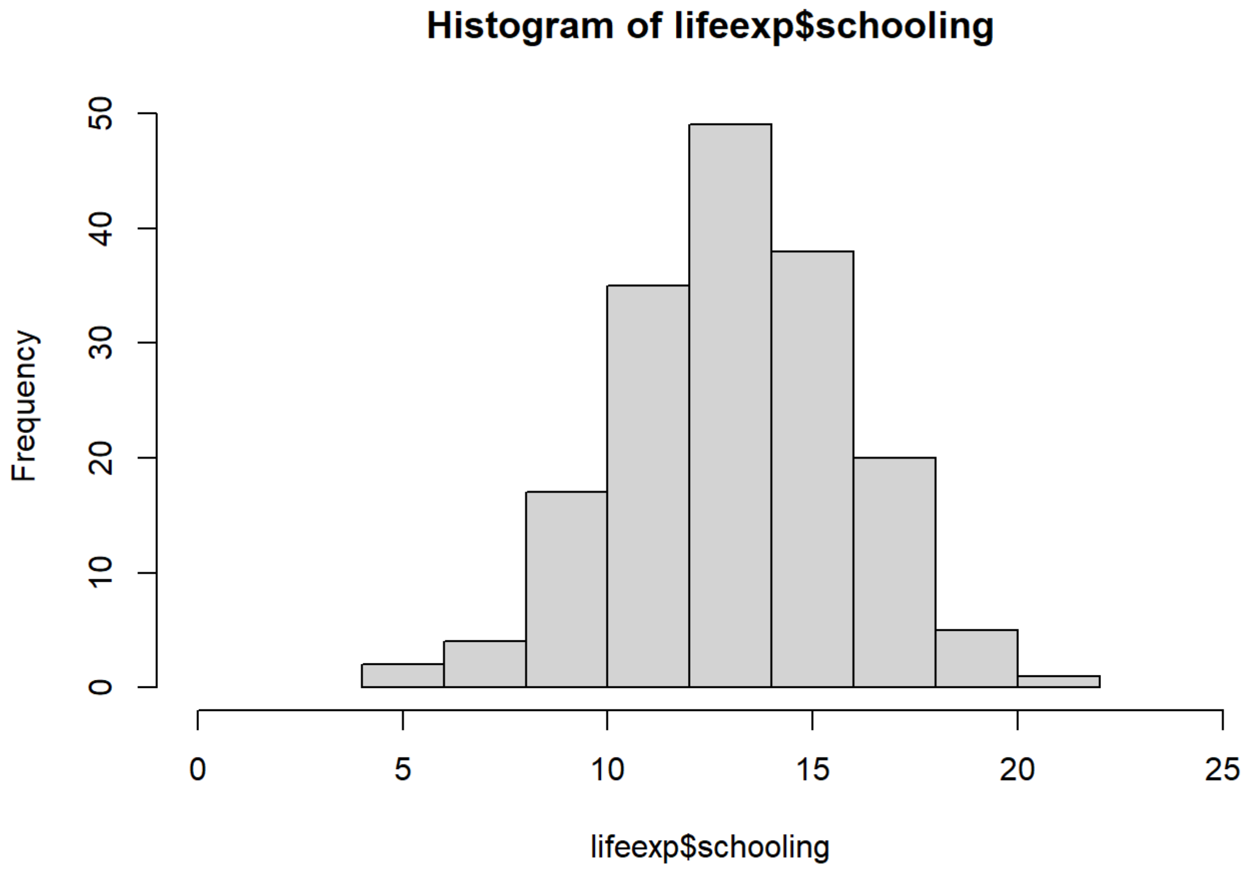
Task 2

> A quantitative variable that is of interest is schooling. Write a paragraph summarizing and describing the variable. The explanations should be such that a person with limited statistical knowledge can understand.
> * You should initially check for NA values.
> * Include relevant graphs (minimally a histogram and a boxplot).
> * Include detailed descriptions of the graphs.
> * Include appropriate descriptive statistics (minimally measuring the center and spread).
> * Explain what those statistics describe about the data.

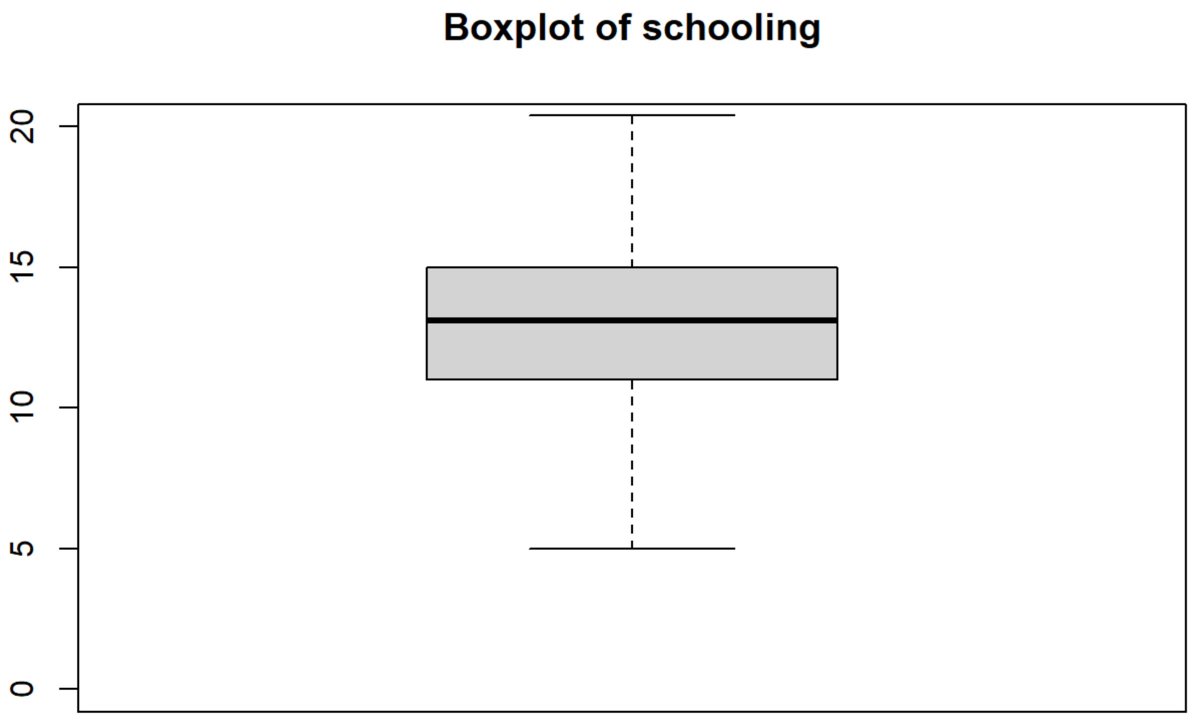
```
sum(is.na(lifeexp$schooling))
```

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

```
hist(lifeexp$schooling, right = FALSE, xlim = c(0, 25))
```



```
boxplot(lifeexp$schooling, ylim = c(0, 20))
title(main = "Boxplot of schooling")
```



```
sd(lifeexp$schooling)
```

```
## [1] 2.828989
```

```
range(lifeexp$schooling)
```

```
## [1]  5.0 20.4
```

```
summary(lifeexp$schooling)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.00   11.00   13.10   13.01   15.00   20.40
```

The histogram displays fairly evenly distributed data with a mean of 13.01 and no skew—all of the data are bell-shaped. The spread is a little bit tighter and centered around the mean, though. The boxplot demonstrates that the black bar, which represents the median of 13.10, is nearly in the center. There are no outliers either. According to our coding, the range (standard deviation) is 2.83 and the mean (center) is 13.01. Also, the range of our results is 5.0 to 20.4, with 50% of them occurring between 11 and 15. Together, we may conclude that the data we have from the collection is fairly evenly distributed. There are no skews or outliers. The data-plots, for instance, show us that nations are neither uniform in value or tilted toward the bottom half of the schooling years. In actuality, the data-plots also inform us that no country in the dataset is accumulating up school hours at an alarming rate or is utterly deficient in them. We can concentrate on the overall image when there are no outliers. With around half of our sample nations having schooling years between 11 and 15 years, the average length of education across the countries in our sample set is around 13 years. We may make more estimates based on the dispersion of our data.We may estimate that around 95% of our nations have schooling years that are within two standard deviations of our average, in this case: 7.34 and 18.66, using our spread, standard deviation, and the 68-95-99.7 rule.

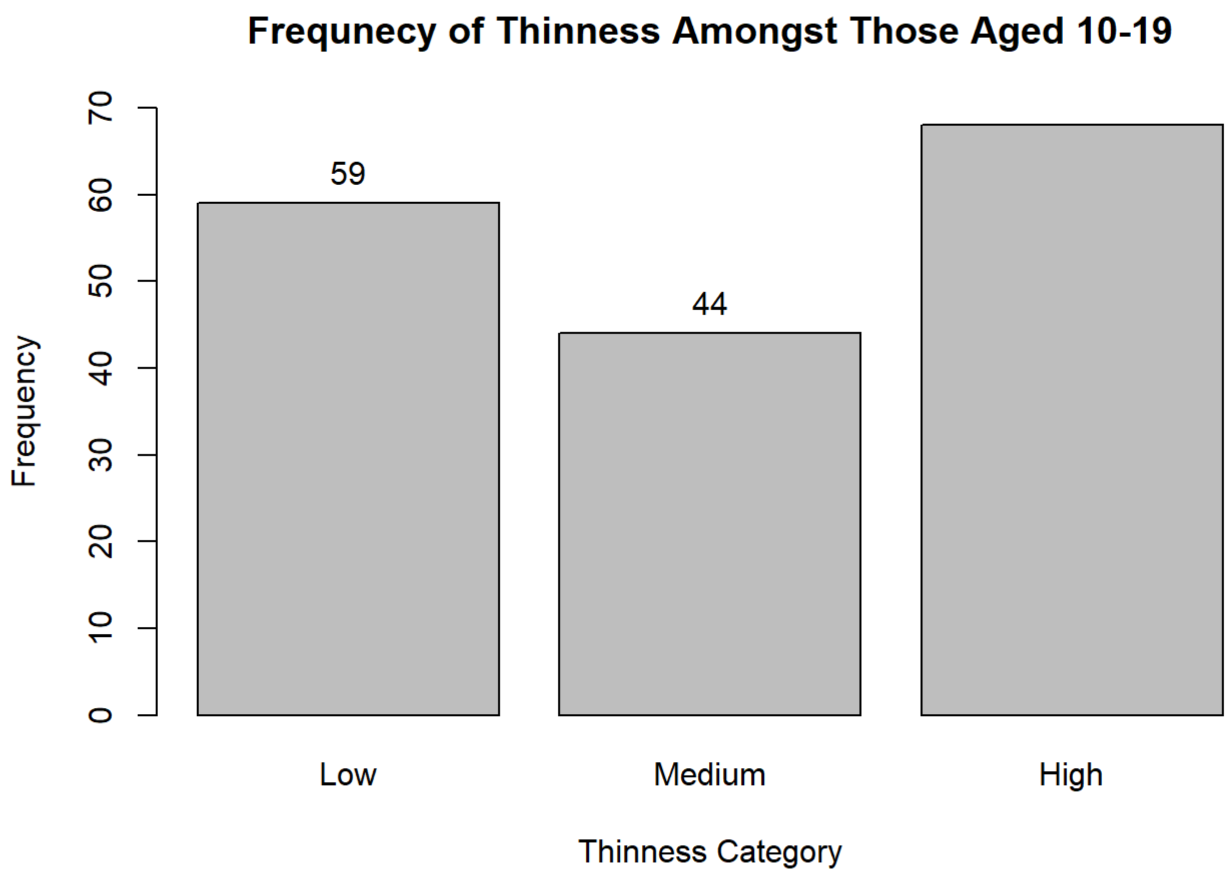
Task 3

> A categorical variable that is of interest is thin_youth. Write a paragraph summarizing and describing the variable. The explanations should be such that a person with limited statistical knowledge can understand.
> * Include at least one relevant graph.
> * Include a detailed description of the graph.
> * Include appropriate descriptive statistics (such as frequency).
> * Explain what those statistics describe about the data.

```
table(lifeexp$thin_youth)
```

```
## 
##    Low Medium   High 
##     59     44     68
```

```
thincounts <- table(lifeexp$thin_youth)
thinplot <- barplot(thincounts, xlab = "Thinness Category", ylab = "Frequency",
    main = "Freqnecy of Thinness Amongst Those Aged 10-19",
    ylim = c(0, 70))
text(x = thinplot, y = thincounts, label = thincounts, pos = 3,
    cex = 1, col = "black")
```



This barplot displays the frequency counts for the various categories of thinness across all nations. Thus far, it appears to have two peaks at both low and high frequencies, with the HIGH category having the highest frequency and the MID category having the lowest.

By examining the data above, we can more clearly identify the frequency. Here, we can see that the frequencies for the following categories—Low, Medium, and High—were 59, 44, and 68, respectively. These frequency figures help to further quantify the earlier justification. Our Low and High frequencies are both above average, however our Middle frequencies are below normal, according to the average, which we computed.

We determine our average as follows:

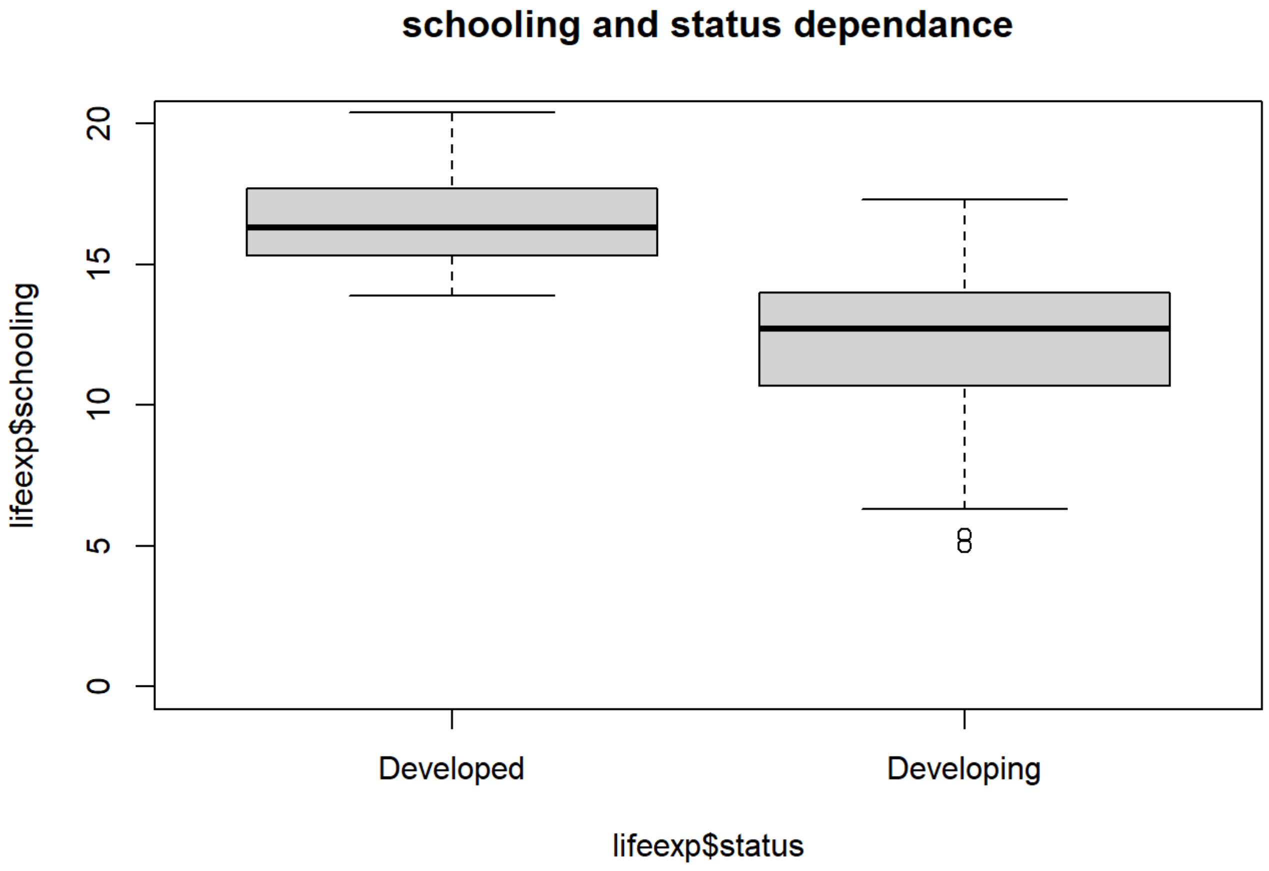
```
mean(table(lifeexp$thin_youth))
```

```
## [1] 57
```

Task 4

> One would like to know if schooling varies by status. Write a paragraph explaining whether or not you think this variable varies by status and detailing how you reached each conclusion.
> • Include graphs (minimally side by side boxplots) for each group that support your conclusions.
> • Include summary statistics (minimally measuring the center and spread) for each group that support your conclusions.

```
boxplot(lifeexp$schooling ~ lifeexp$status, ylim = c(0, 20))
title(main = "schooling and status dependance", )
```



```
tapply(lifeexp$schooling, lifeexp$status, summary)
```

```
## $Developed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   13.90   15.30   16.30   16.54   17.70   20.40 
## 
## $Developing
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.00   10.70   12.70   12.29   13.97   17.30 
```

```
tapply(lifeexp$schooling, lifeexp$status, sd)
```

```
##  Developed Developing 
##   1.647382   2.454269
```

Notwithstanding the differences in our data and charts, we may conclude with confidence that educational attainment differs by status.

As compared to the boxplot for poor nations, the schooling years for rich countries are on the higher end of values with a narrower range. The range is much wider for emerging nations, with notable outliers at the lower end and the median (solid) line leaning more toward the upper end. Not to add that the median in wealthy nations is not comparable to that in poor nations.Not to add that emerging nations have a wider dispersion than industrialized nations.

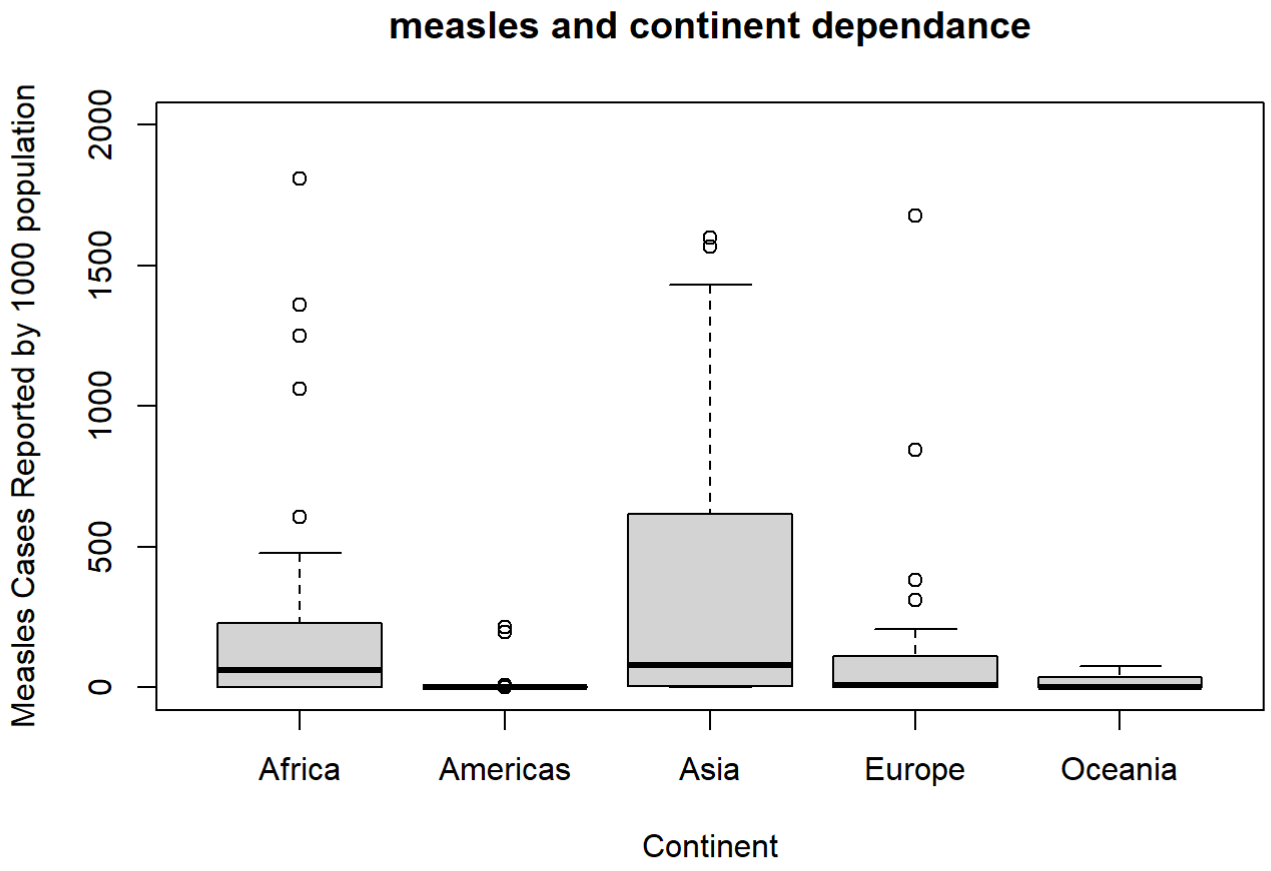
All of this demonstrates that whereas poor nations have a greater variety and wider distribution of schooling years, with an average of 12.29 years, industrialized countries tend to have more comparable years of schooling with an average of 16.54 years.

A glance at the average is another option. The average age in wealthy nations is 16.54 years, compared to 12.29 in developing nations.

Task 5

> One would like to know if measles varies by continent. Write a paragraph explaining whether or not you think this variable varies by continent and detailing how you reached ach conclusion.
> • Include graphs (minimally side by side boxplots) for each group that support your conclusions.
> • Include summary statistics (minimally measuring the center and spread) for each group that support your conclusions.

```
boxplot(lifeexp$measles ~ lifeexp$continent, ylim = c(0, 2000),
    xlab = "Continent", ylab = "Measles Cases Reported by 1000 population",
    main = "measles and continent dependance")
```



```
tapply(lifeexp$measles, lifeexp$continent, summary)
```

```
## $Africa
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0     2.5    60.5  1090.7   222.0 17745.0 
## 
## $Americas
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    0.00    0.00   13.22    0.00  214.00 
## 
## $Asia
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0       6      80    4402     615   90387 
## 
## $Europe
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0     1.0     7.5   184.7   107.5  2464.0 
## 
## $Oceania
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0     0.0     0.0    16.1    31.0    74.0
```

```
tapply(lifeexp$measles, lifeexp$continent, sd)
```

```
##      Africa    Americas        Asia      Europe     Oceania 
##  3293.05568    50.26622 15123.10351   499.49179    25.70970
```

We may reasonably conclude from our data and charts that the number of measles cases differs by continent.

The boxplot shows that each continent has a completely unique range, with multiple outliers (dots) for several continents. This demonstrates that while measles cases are substantially more dispersed on some continents, they are not on the Americas and Oceania. They are a lot smaller.

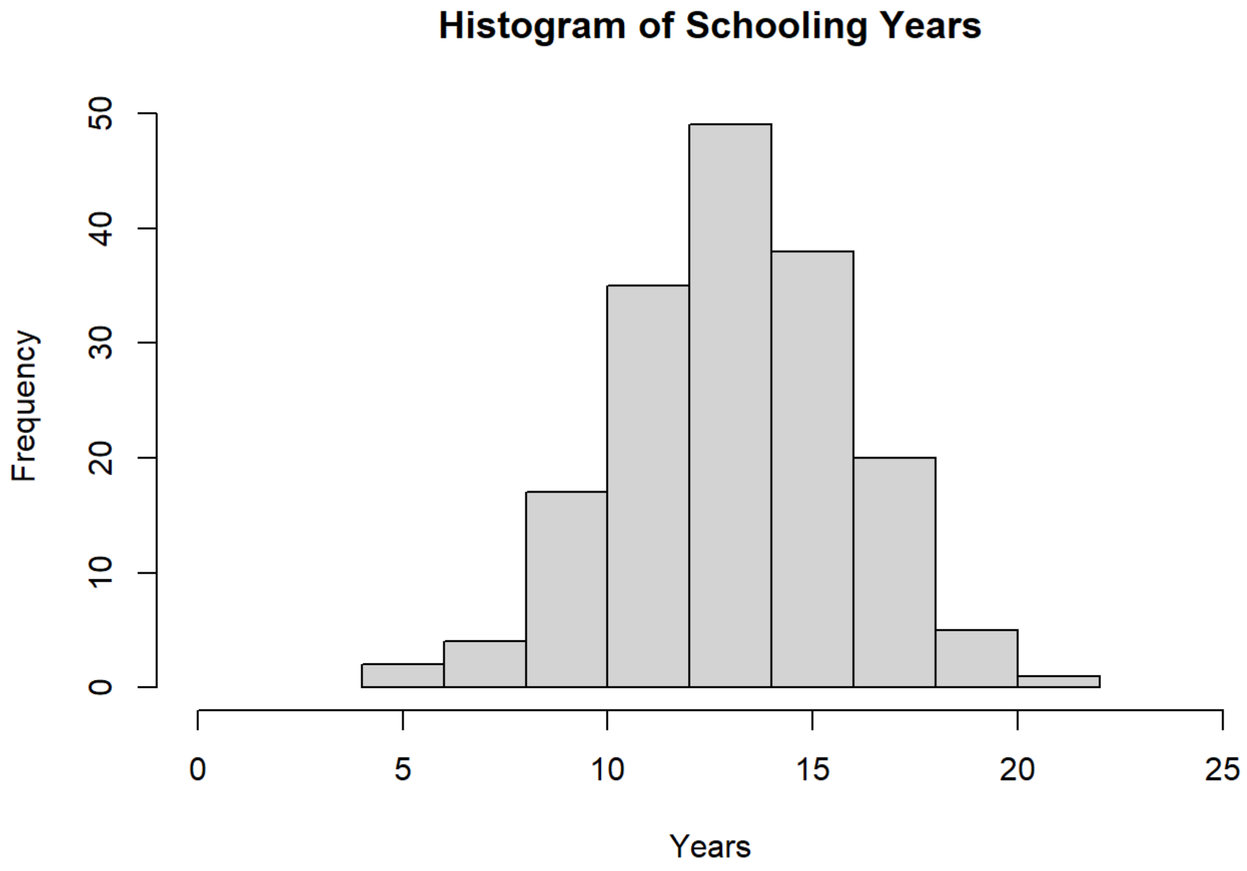
Our standard deviation, or the “sd” in “tapply,” determines our spread. Again, this demonstrates how the distribution of data is substantially more dispersed for some continents than for others, most notably the Americas and Oceania. Compared to Africa, Asia, and Europe, these two continents have spreads that are substantially smaller. The Americas and Oceania’s mean is once again substantially lower than that for Africa, Asia, and Europe when we look at our mean, which is provided by the ‘Mean’ function in our’summary’ function within ‘tapply’.

We can observe that the number of measles cases varies greatly by continent when we combine the spread of data, average of our data, and plot of our data.

Task 6

> Two variables that might be studied more in the future are schooling and measles. It would be helpful to know if these variables are normally distributed. Write a paragraph for each variable explaining why one should or should not assume the variable is normally distributed. Explain in a way that a person with limited statistical knowledge would understand.
> • Include all necessary graphs (at least one of a histogram or a Q-Q Plot with reference line). Explain the implications of the graph(s) in regards to normality of the variable.
> • Calculate the skew and kurtosis. Explain the implications of the values reported in regards to the normality of the variable.
> • Perform the Shapiro-Wilk Test. State your hypotheses, your decision, and conclusion. Use a 1% significance level.

```
hist(lifeexp$schooling, right = FALSE, xlim = c(0, 25), main = "Histogram of Schooling Year
s",
    xlab = "Years")
```



```
shapiro.test(lifeexp$schooling)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lifeexp$schooling
## W = 0.9957, p-value = 0.9072
```

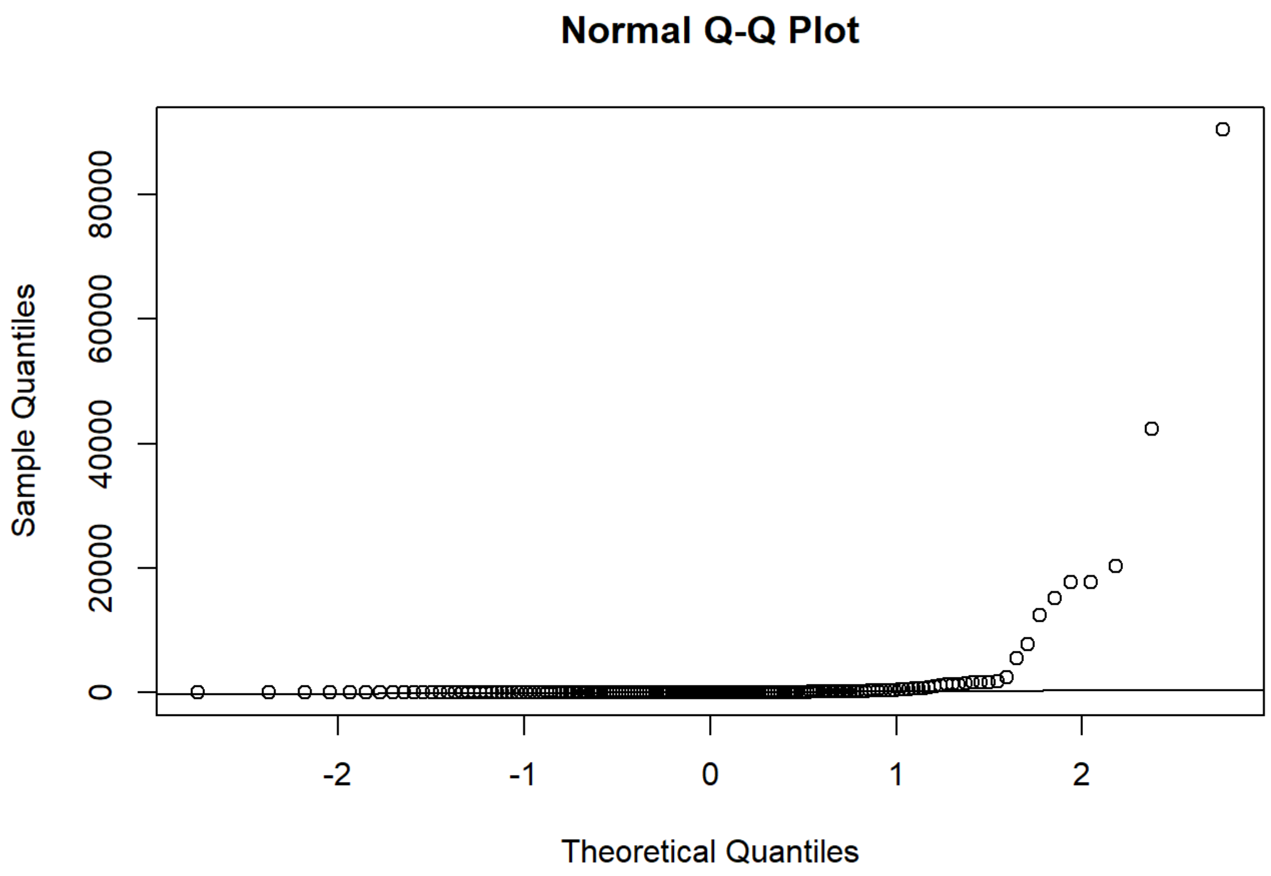
```
skewness(lifeexp$schooling, type = 3)
```

```
## [1] -0.1476361
```

```
kurtosis(lifeexp$schooling, type = 3)
```

```
## [1] -0.08845119
```

```
qqnorm(lifeexp$measles)
qqline(lifeexp$measles)
```



```
shapiro.test(lifeexp$measles)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lifeexp$measles
## W = 0.17902, p-value < 2.2e-16
```

```
skewness(lifeexp$measles, type = 3)
```

```
## [1] 8.720267
```

```
kurtosis(lifeexp$measles, type = 3)
```

```
## [1] 86.47514
```

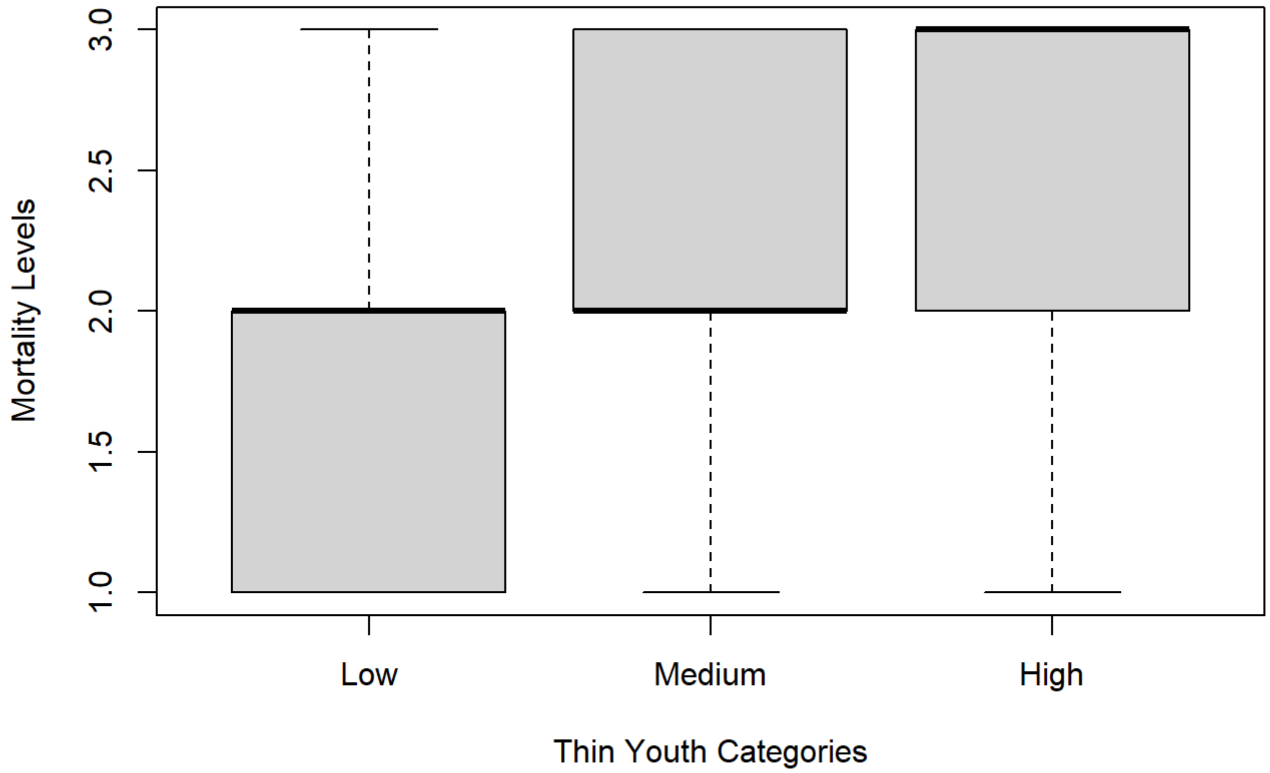
Our schooling data set may appear to be typical for whatever reason. If you look at the histogram plot up top, you won't see any obvious skews to the left or right, and there aren't any large tails at the end. The bell-shaped curve that surrounds all of the data points is centered on the mean. Also, there are no anomalies. Also, the results of our Shapiro-Wilk normality test support this likelihood. Only when the p-value is greater than our significance value is normality taken into account (in this 0.01). We may think of it as usual because it is. Any result near to 0 is seen favorably for normalcy when calculating the skewness, which is a measure of how far to the right or left our data is pushed.Our skewness, which is -0.1476361, supports normalcy as well. Last but not least, kurtosis is a tailedness indicator; a value near to 0 indicates symmetry and supports normalcy. Our kurtosis value, which is -0.08845119, further supports normalcy.

The same cannot be true with our measles data set, though. We utilized a QQ plot in place of a histogram, which compares actual percentiles to those predicted by a normal distribution. We can presume normalcy if a line connects the spots. A big number of points are at stake. The pronounced tail at the extreme right, however, casts doubt on the ability to determine normalcy.There must be more testing done. With regard to our Shapiro-Wilk test, there is sufficient evidence to conclude that our data are NOT normally distributed because our p-value is not greater than our significance of 0.01 and is considerably less than our significance. Last but not least, our values for skewness and kurtosis are far from zero. They exhibit a skew and a tail that contradict normalcy.

Task 7

> Mortality Level.
> • Create a new ordered factor called mortality_level that takes on the value of “Low” if the adult.mortality is less than 80, “Moderate” if the adult.mortality is at least 80 and less than 150, and “High” if the adult.mortality is at least 150. Please provide the code in your project, but you do not need to comment on it.
> • Write a paragraph explaining whether or not you think mortality_level varies by thin_youth and detail how you reached your conclusion.
> – Include graphs (minimally stacked or side-by-side barplots) for each group that support your conclusions.
> – Include summary statistics (such as frequency) by group that support your conclusions.

```
lifeexp$mortality_level[lifeexp$adult.mortality < 80] <- "Low"
lifeexp$mortality_level[lifeexp$adult.mortality >= 80 & lifeexp$adult.mortality <
    150] <- "Moderate"
lifeexp$mortality_level[lifeexp$adult.mortality >= 150] <- "High"
lifeexp$mortality_level <- ordered(lifeexp$mortality_level, levels = c("Low",
    "Moderate", "High"))
boxplot(lifeexp$mortality_level ~ lifeexp$thin_youth, ylab = "Mortality Levels",
    xlab = "Thin Youth Categories", data = lifeexp$mortality_level)
```



```
tapply(lifeexp$mortality_level, lifeexp$thin_youth, summary)
```

```
## $Low
##      Low Moderate     High 
##       28       17       14 
## 
## $Medium
##      Low Moderate     High 
##        6       19       19 
## 
## $High
##      Low Moderate     High 
##       15       12       41
```

Certainly, there is sufficient information to demonstrate that the death rate does differ for the groups of thin kids. The low thin juvenile groups on the plots frequently have lower death rates. Remember that in our graphic, 1.0 represents a Low mortality level, 2.0 a Moderate mortality level, and 3.0 a High mortality level. When greater death rates are associated with higher categories of thinness, everything is mirrored on the other side.

We have a clearer perspective after looking at our summary. Lower levels of thinness were correlated with lower rates/counts of low mortality. When we move up the thinness scale, mortality tends to increase.As we reach a medium level of thinness, moderate and high death rates outnumber low mortality rates. In contrast, consider great thinness. While the thinness group was categorized as having a high death rate, there were more than 40 incidents documented.