Tidy Tuesday Data Project Step 2

Saakshi Shah

2/27/2021

1) Exploring Single Variables

Total

```
# Summary for Total
favstats(Total)
```

```
##  min      Q1 median       Q3    max     mean       sd   n missing
##  124 4549.75  15104 38909.75 393735 39370.08 63483.49 172       1
```

```
# Number of outliers
Q1 = 4549.75
Q3 = 38909.75
IQR = Q3 - Q1
upper_whisker = Q3 + 1.5 * IQR
Total_outliers = recent_grades %>% filter(Total > upper_whisker)
count(Total_outliers)
```

```
##    n
## 1 21
```

```
# Histogram of Total
ggplot(data = recent_grades,
       mapping = aes(x = Total,
                     fill = Major_category)) +
    geom_histogram(alpha = 0.8, bins = 20) +
    scale_x_continuous(name = "Number of Studnets") +
    scale_y_continuous(name = "Frequency") +
    scale_fill_discrete(name = "Major Category") +
    ggtitle("US Graduate Students Enrolled in Major")
```

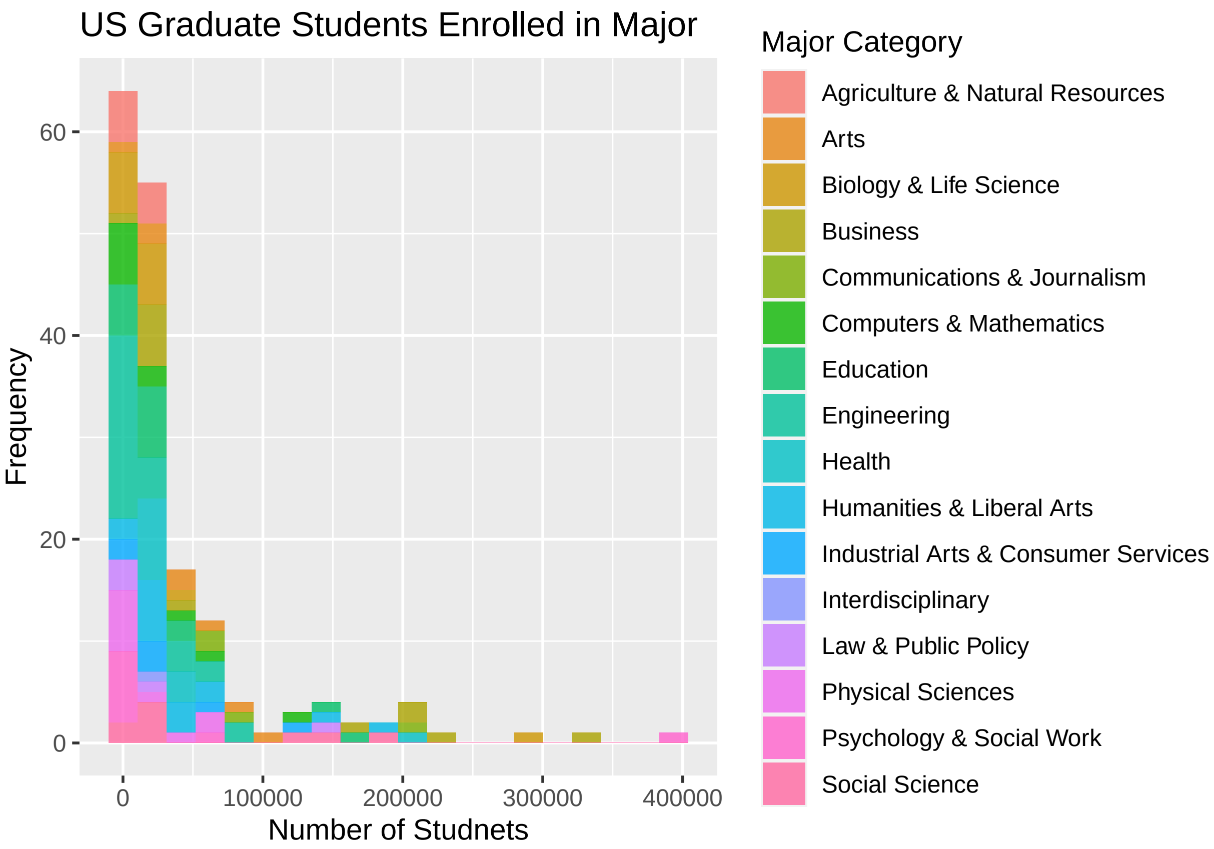
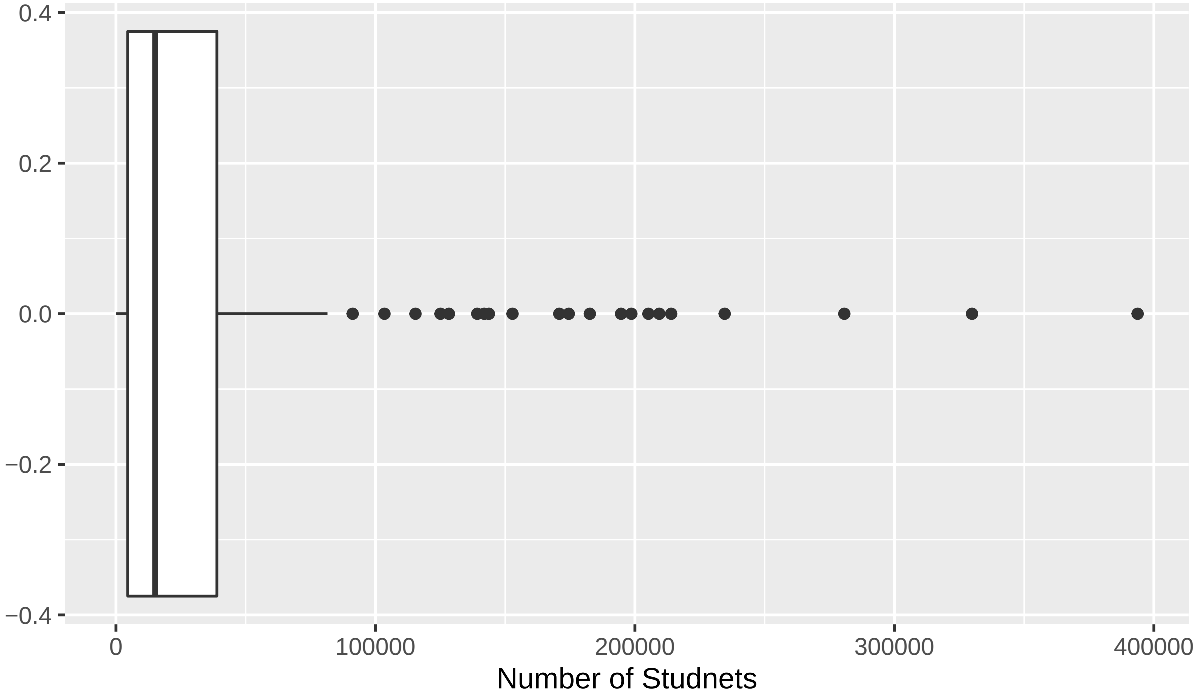


Figure 1: Histogram of US Graduate Students Enrolled in Major

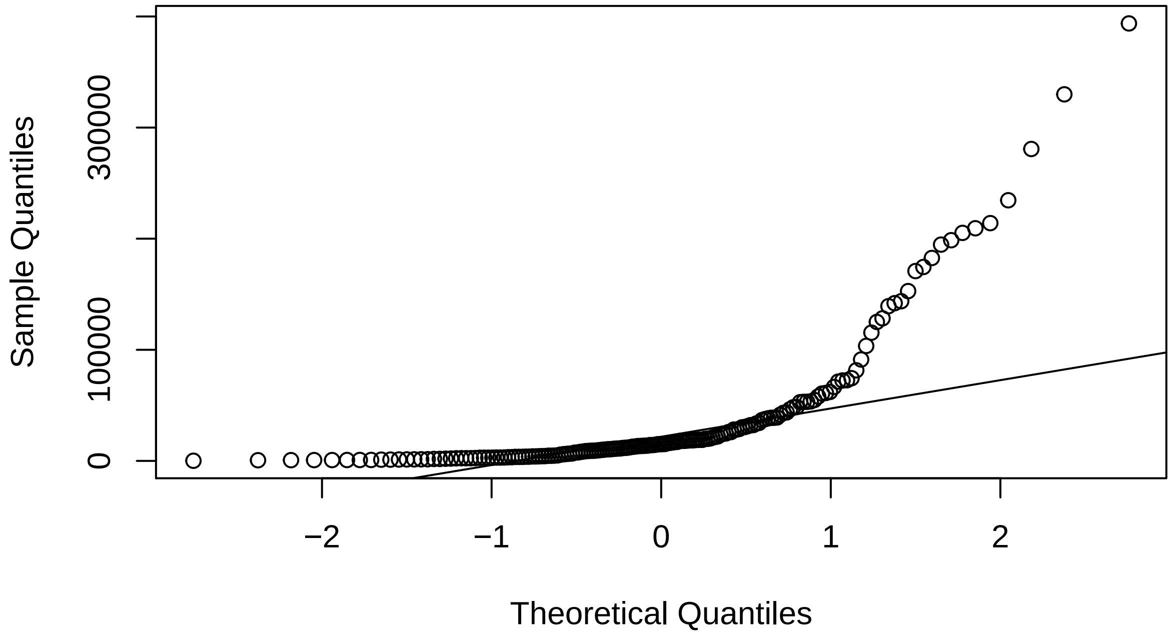
```
# Boxplot of Total
ggplot(data = recent_grades,
       mapping = aes(x = Total)) +
    geom_boxplot() +
    scale_x_continuous(name = "Number of Studnets") +
    scale_y_continuous(name = "") +
    ggtitle("Figure 2: Boxplot of US Graduate Students Enrolled in Major")
```

Figure 2: Boxplot of US Graduate Students Enrolled in Major



```r
# normal QQ Plot of Total
qqnorm(Total, main = "Figure 3: Normal QQ Plot of US Graduate Students Enrolled in Major")
qqline(Total)
```

Figure 3: Normal QQ Plot of US Graduate Students Enrolled in Majo



The variable “Total” represents the total number of graduate students enrolled with a major in the United States. Through the visualization from figure 1, it is clear that the number of graduates enrolled in a major is right-skewed. The median number of students enrolled in a major is 15104, and the average is 39370. Consequently, this states that 50% of the majors have at most 15104 students enrolled, and the other 50% of the majors have more than 15104 students enrolled. The middle half of the distribution extends across 4550 to 38910, with an interquartile range of 34360. The calculation shows that there exist 21 extreme outliers in this set of data. For example, the number of students enrolled in Psychology major, Business Administration Management, and Biology/Life Science, with total enrollment over 250,000 students each. We can conclude that they are the most popular majors in US graduate schools, and the data does not comes from a normal distribution.

Unemployment Rate

```
# Summary of Unemployment Rate
favstats(Unemployment_rate)
```

```
##  min         Q1     median         Q3       max       mean         sd   n
##    0 0.05030643 0.06796077 0.08755711 0.1772264 0.06819083 0.03033094 173
##  missing
##        0
```

```
# Histogram of Unemployment Rate
ggplot(data = recent_grades,
       mapping = aes(x = Unemployment_rate,
                     fill = Major_category)) +
    geom_histogram(alpha = 0.8, bins = 20) +
    scale_x_continuous(name = "Unemplyment Rate",
                       labels = scales::percent) +
    scale_y_continuous(name = "Frequency") +
    scale_fill_discrete(name = "Major Category") +
    ggtitle("US Graduate Students' Unemplyment Rate")
```

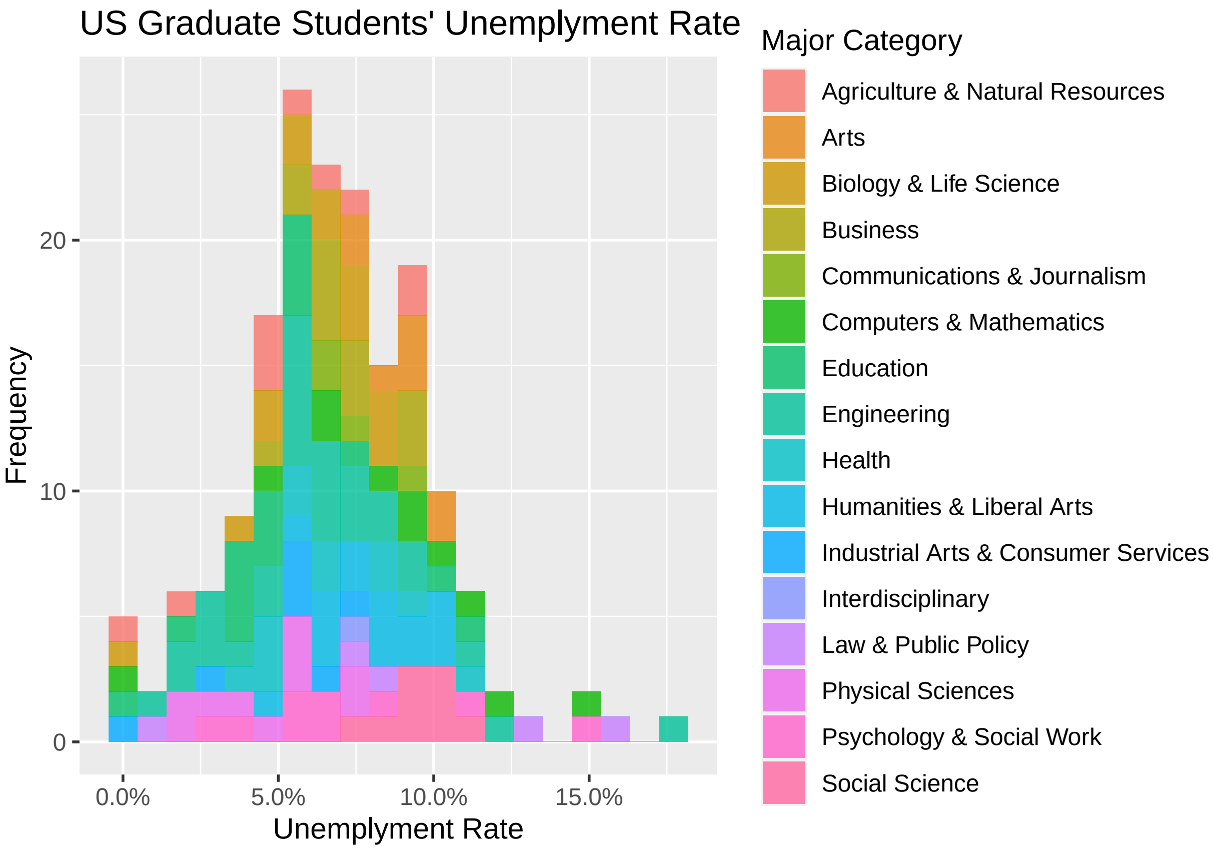


Figure 4: Histogram of US Graduate Students' Unemployment Rate

```
# Boxplot of Unemployment Rate
ggplot(data = recent_grades,
       mapping = aes(x = Unemployment_rate)) +
    geom_boxplot() +
    scale_x_continuous(name = "Unemplyment Rate",
                       labels = scales::percent) +
    ggtitle("Boxplot of US Graduate Students' Unemplyment Rate")
```

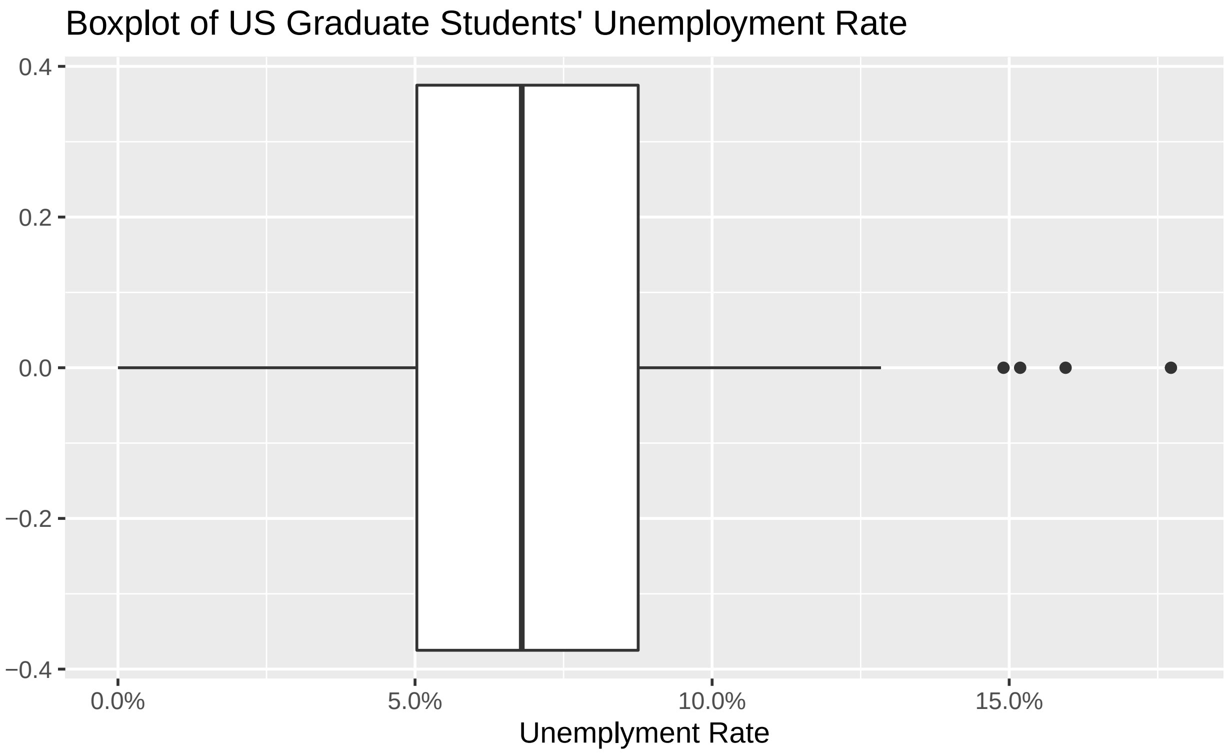


Figure 5: Boxplot of US Graduate Students' Unemployment Rate

```
# normal QQ Plot of Unemployment Rate
qqnorm(Unemployment_rate, main = "Normal QQ Plot of US Graduate Students' Unemployment Rate")
qqline(Unemployment_rate)
```

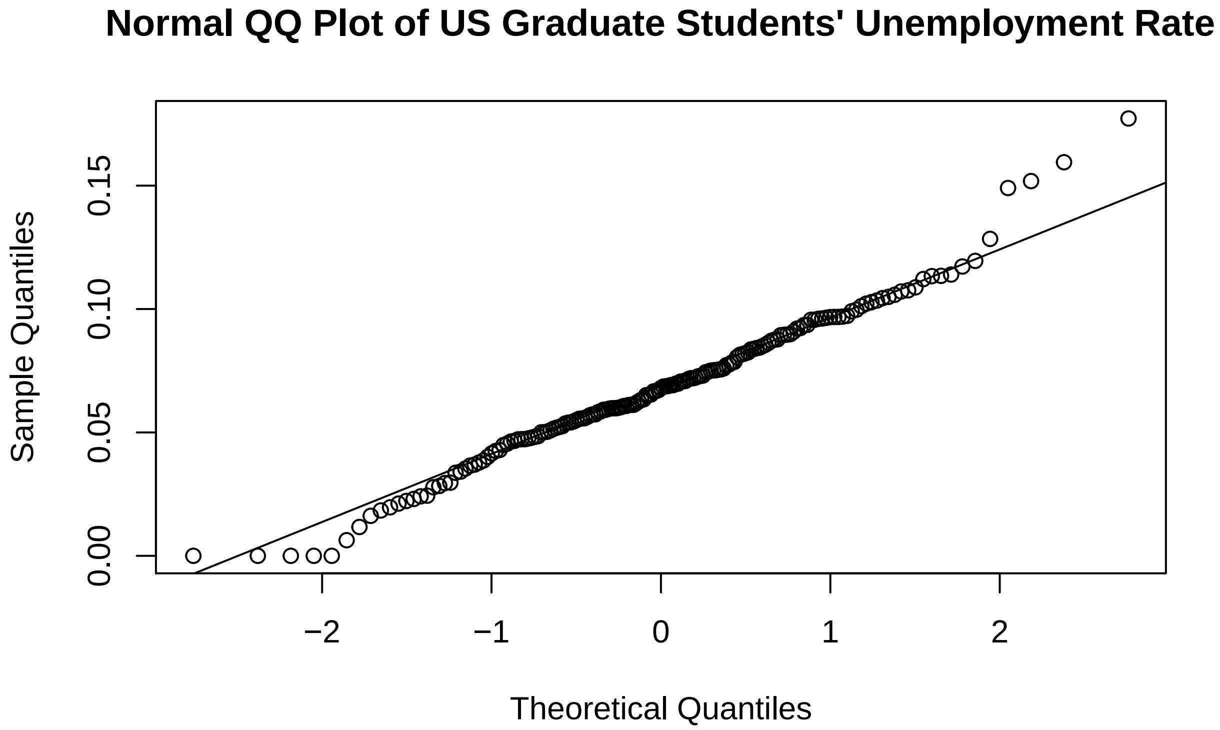


Figure 6: Normal QQ Plot of US Graduate Students' Unemployment Rate

The variable "Unemployment_rate" represents the proportion of unemployed students vs. employed graduate students. Examining figure four, the unemployment rate is unimodal, slightly right-skewed. It ranges from 0.00% to 17.7%. 50% of the students after graduation have an unemployment rate of at most 6.80%, and the other 50% are higher than 6.80%. The middle half of the distribution extends across a range of 5.03% to 6.82%, with an Interquartile Range of 1.79%. We can expect that percentage of unemployment differ from the mean by about 3.00%, on average. Figure 3 shows that there exist four extreme data points, these are Nuclear Engineering, Public Administration, Computer Networking/Telecommunications, and Clinical Psychology. With an unemployment percentages of 17.7, 16.0, 15.1, and 14.9 respectively. Yet, we do not have enough evidence to conclude that the set of data was not normally distributed from figures four to six.

Female Graduates Population Percentage

```r
# Summary of Sharewomen
favstats(ShareWomen)
```

```
##  min        Q1   median        Q3       max      mean       sd   n missing
##    0 0.3360262 0.534024 0.7032992 0.9689537 0.5222234 0.231205 172       1
```

```r
# Histogram of Sharewomen
ggplot(data = recent_grades,
       mapping = aes(x = ShareWomen,
                     fill = Major_category)) +
    geom_histogram(alpha = 0.8, bins = 15) +
    scale_x_continuous(name = "Population Percentage",
                       labels = scales::percent) +
    scale_y_continuous(name = "Frequency") +
    scale_fill_discrete(name = "Major Category") +
    ggtitle("Population Percentage of Female Graduates")
```

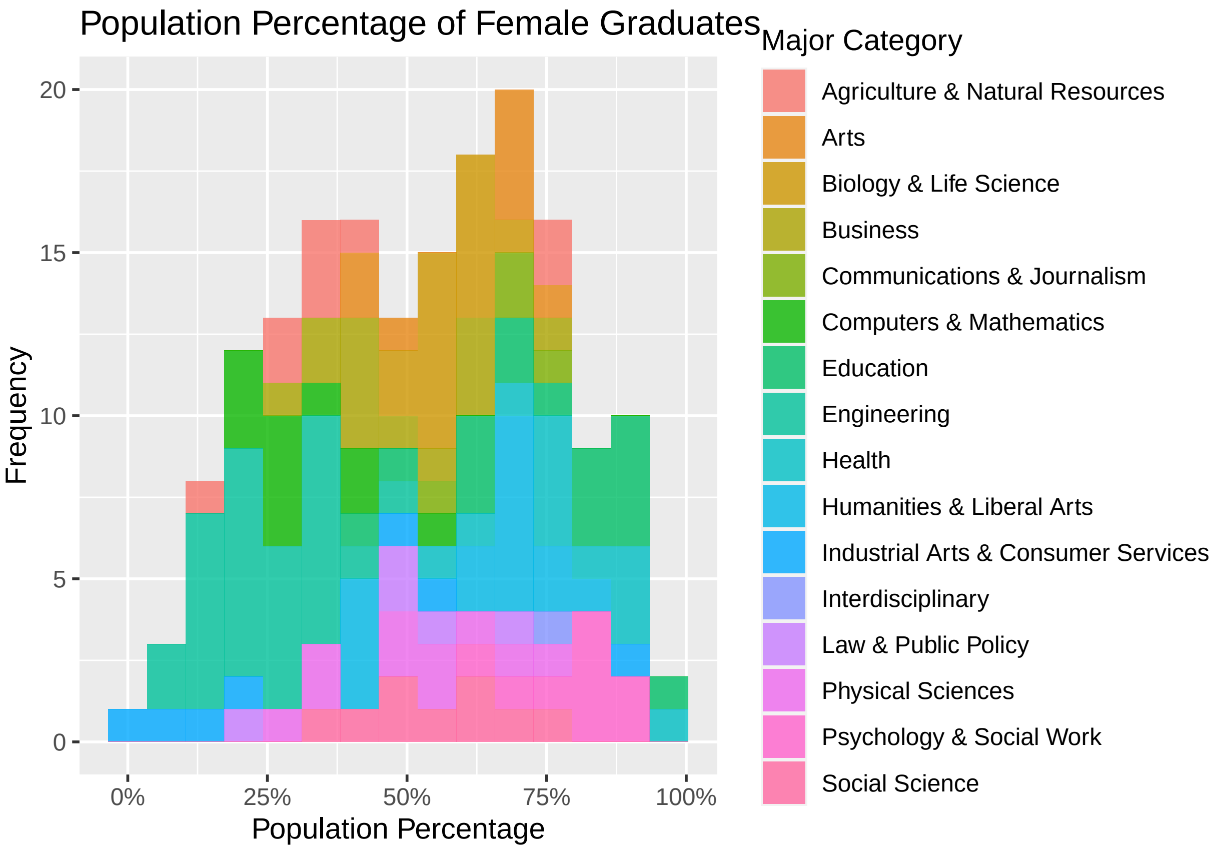


Figure 7: Population Percentage Histogram of US Female Graduates

```
# Boxplot of Unemployment Rate
ggplot(data = recent_grades,
       mapping = aes(x = ShareWomen)) +
    geom_boxplot() +
    scale_x_continuous(name = "Population Percentage",
                       labels = scales::percent) +
    ggtitle("Boxplot of US Female Graduates' Population Percentage")
```

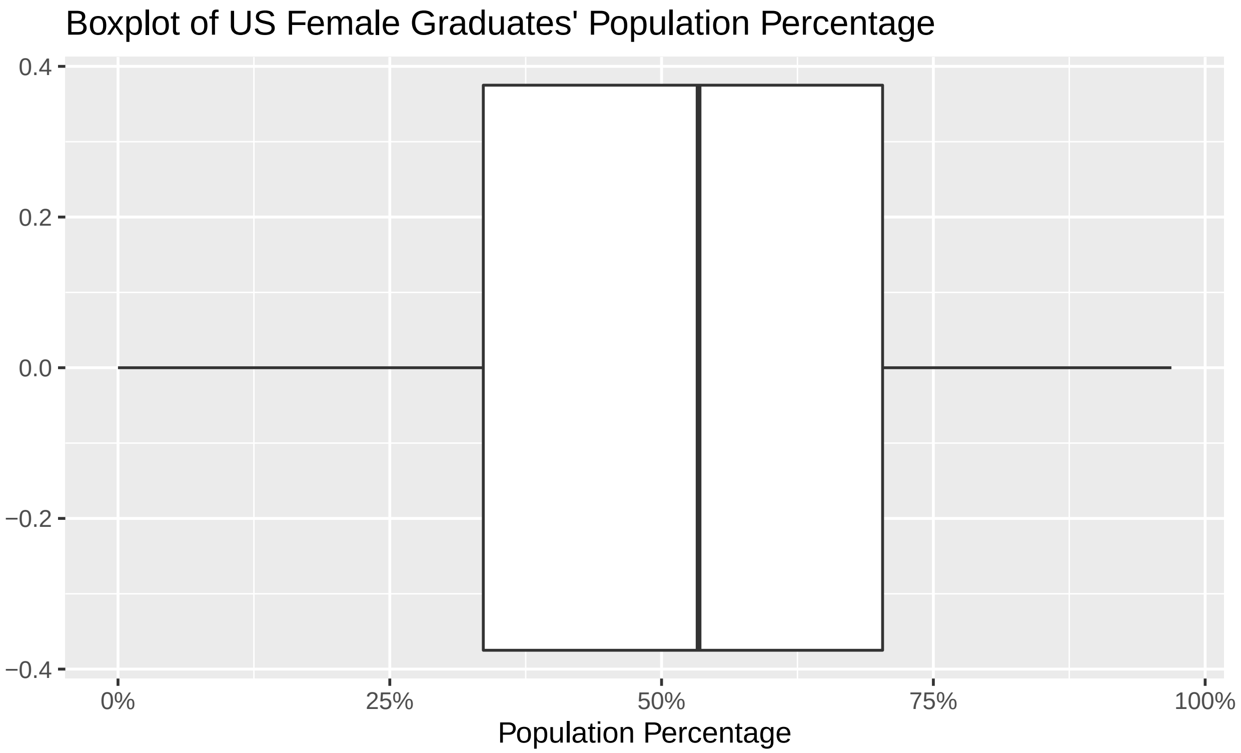


Figure 8: Boxplot of US Female Graduate Students' Population Percentage

```
# normal QQ Plot of Unemployment Rate
qqnorm(ShareWomen, main = "Normal QQ Plot of Female Graduates' Population Percentage")
qqline(ShareWomen)
```

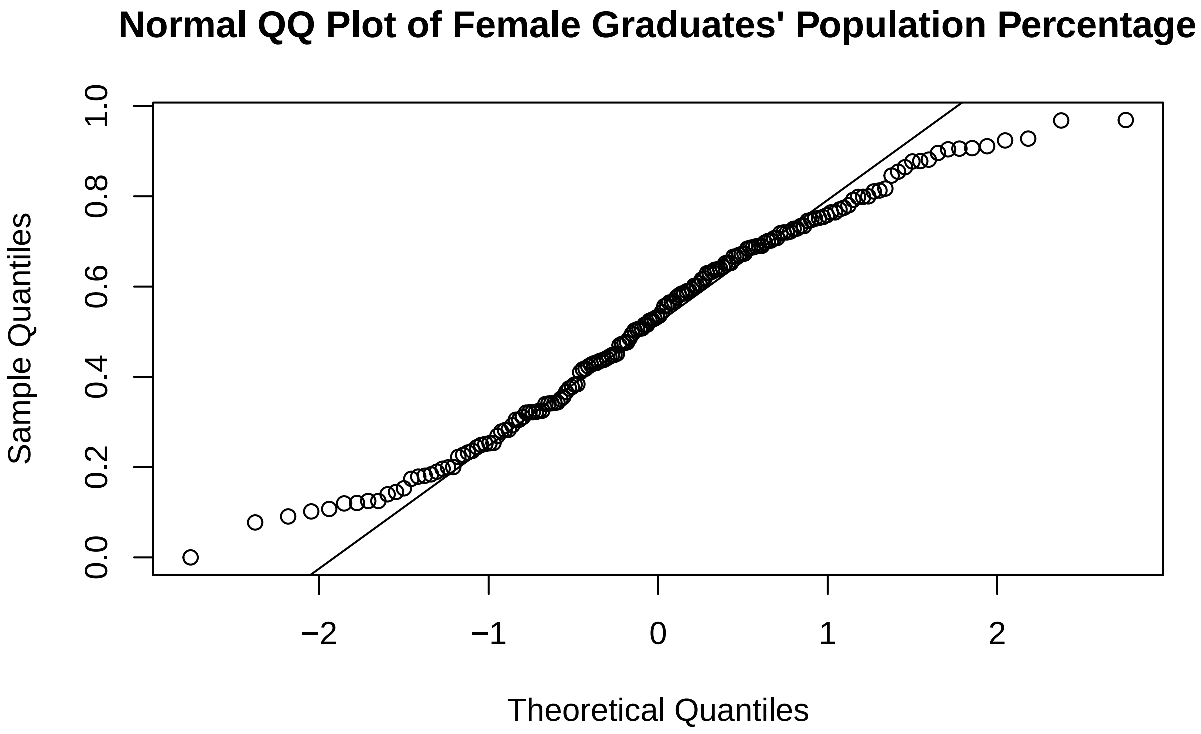


Figure 9: Normal QQ Plot of US Female Graduate Students' Population Percentage

The variable “ShareWomen” shows the female graduate student’s population percentage. In the 137 different majors, the female graduate population range from 0.00% to 96.9%, with an average of around 52.2%. The median population percentage is 53.4%. That is, 50% of the students have a female population with at most 53.4%, whereas 50% of the students have more than 53.4% of the female population. The middle half of the distribution ranges from 33.6% to 70.3%, with an Interquartile Range of 36.7%. We can expect that percentage of the female population differs from the mean by about 23.1%, on average. Education and Psychology have the highest female population enrolled, while Industrial Arts & Engineering have the least female population enrolled. Figure eight shows that there exist no extreme data points. We do not have enough evidence to conclude that the data was not normally distributed with figures seven to nine

Annual Median Salary

```r
# Summary of Median
favstats(Median)
```

```
##    min    Q1 median    Q3    max     mean       sd   n missing
##  22000 33000  36000 45000 110000 40151.45 11470.18 173       0
```

```r
# Histogram of Median
ggplot(data = recent_grades,
       mapping = aes(x = Median,
                     fill = Major_category)) +
    geom_histogram(alpha = 0.8, bins = 20,
                   position = "identity") +
    scale_x_continuous(name = "Annual Salary",
                       labels = scales::dollar) +
    scale_y_continuous(name = "Frequency") +
    scale_fill_discrete(name = "Major Category") +
    ggtitle("US Graduates' Annual Median Salary")
```

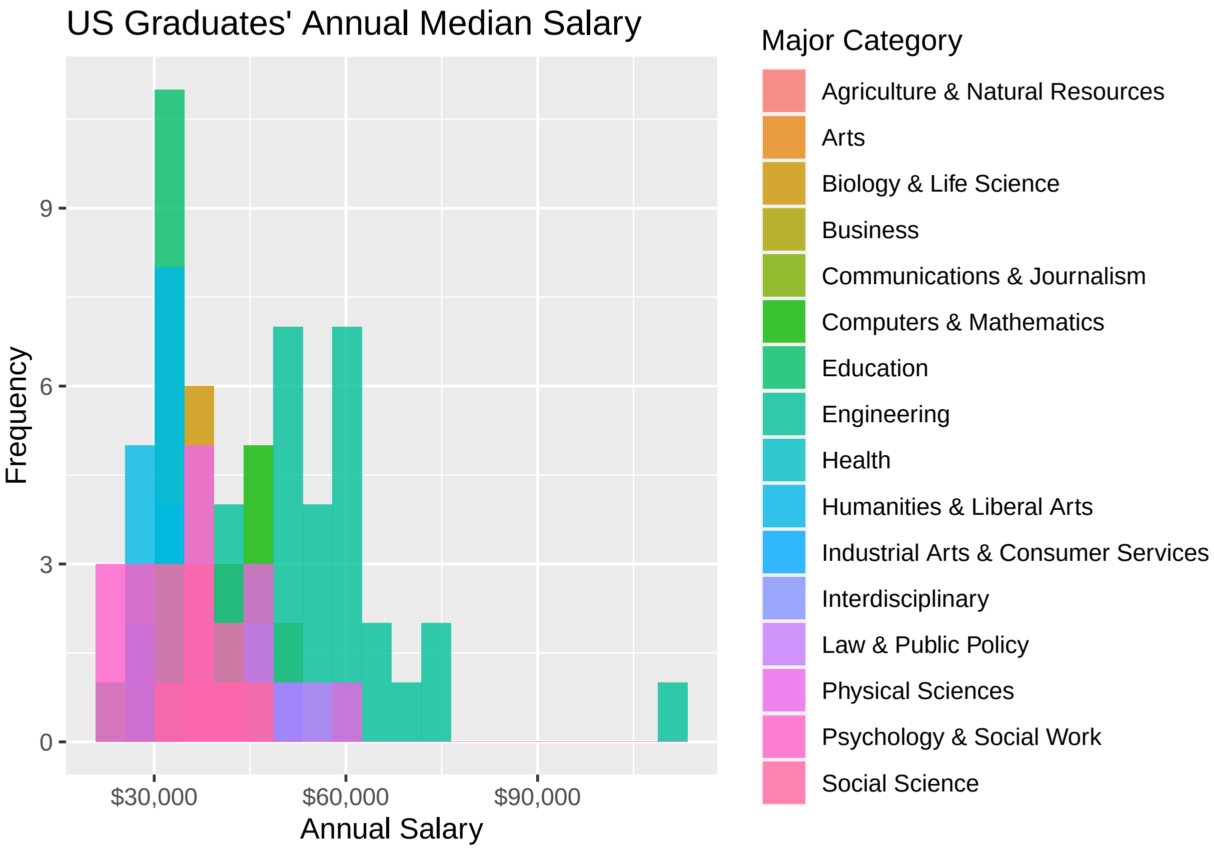


Figure 10: Histogram of US Graduate Students' Annual Median Salary

```
# Boxplot of Median
ggplot(data = recent_grades,
       mapping = aes(x = Median)) +
    geom_boxplot() +
    scale_x_continuous(name = "Annual Slary",
                       labels = scales::dollar) +
    ggtitle("Boxplot of US Graduate Students' Annual Median Salary")
```

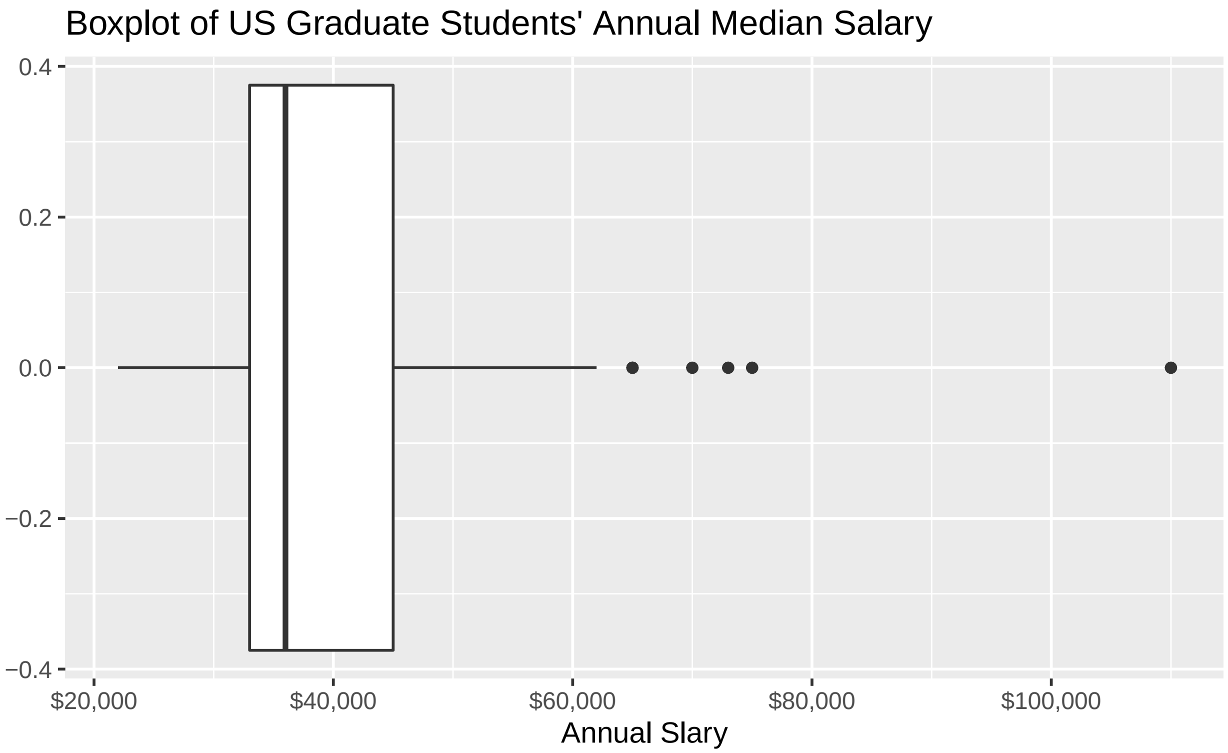


Figure 11: Boxplot of US Graduate Students' Annual Median Salary

```r
# normal QQ Plot of Total
qqnorm(Total, main = "Figure 3: Normal QQ Plot of US Graduate Students Enrolled in Major")
qqline(Total)
```

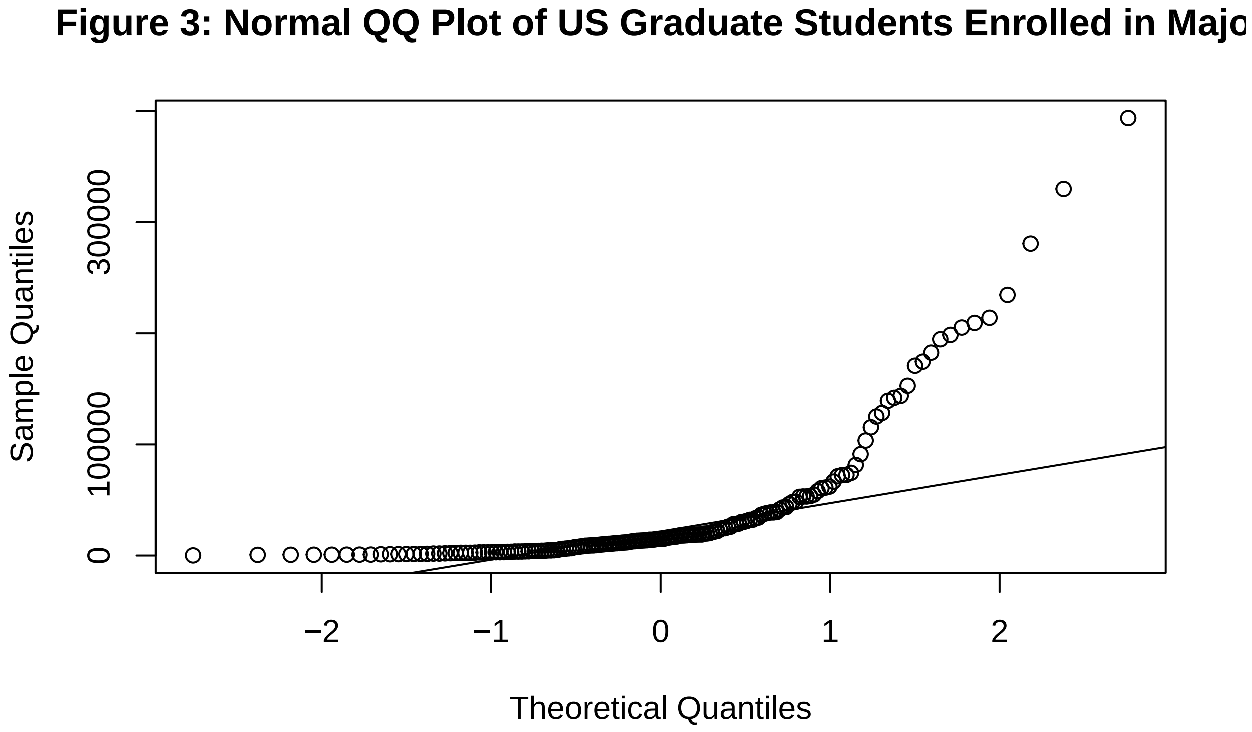


Figure 12: Boxplot of US Graduate Students' Annual Median Salary

The variable “Median” represents the annual median salary of a US graduate student. In the 137 different majors, the annual median salary range from $22,000 to $110,000, with an average of around $40,151. The median annual salary is in the left of the box of the boxplot, it is positioned at value of $36,000. The middle half of the salary ranges from $33,000 to $45,000, with an Interquartile Range of $12,000. We can expect that the annual median salary on average differs from the mean by about $11,470. Five data point are plotted individually with salaries of $65,000, $70,000, $73,000, $75,000 and $110,000. Figure 10 to 12 illustrates that the data is right skewed with extreme outliers. Thus, it does not appear that this data could have come from a normal distribution.