```
---
editor_options:
  markdown:
    wrap: 72
---

## Project 1

## Yelizaveta Semikina

# Task 1

# Graphs

hist(mydata\$Rating, right = FALSE, ylim = c(0,35), main = "Histogram of
Rating", xlab = "Rating")

![](images/Task1_Histogram%20of%20Rating-01.png)

hist(mydata\$Reviews, right = FALSE, ylim = c(0, 200), xlim = c(0,
9000000), main = "Histogram of Reviews", xlab = "Reviews")

![](images/Task1_Histogram%20of%20Reviews.png)

boxplot(mydata\$Rating, horizontal = TRUE, main = "Boxplot of Rating",
xlab = "Rating")

![](images/Task1_Boxplot%20of%20Rating.png)

boxplot(mydata\$Reviews, horizontal = TRUE, breaks = c( 4, 516, 7306,
204111, 82166, 8118609), main = "Boxplot of Reviews", xlab = "Reviews")

![](images/Task1_Boxplot%20of%20Reviews.png)

# Summary

Reviews is a numeric, discrete variable and Rating is a numeric,
continuous variable. The average total for reviews is 4.3 and the
average total ratings is 204110. The min for rating is 3.8 and for
reviews is 4. Also, the first and third quartiles for rating is between
4.3 and 4.4 and for reviews between 516 and 82166. It means that data
points for first quartile are 25% of data are found under 4.2 for rating
and 516 for reviews and for the third quartile that are 4.4 for rating
and 82166 for reviews, found 75% of data. The max for rating is 4.8 and
for reviews is 8118609. In rating column we can see which app has the
best rating and we can see which app is better, as for reviews column,
based on this column we can tell how many people reviewed the app and we
can conclude which app is the most popular. As for graphs, the histogram
for rating looks pretty normal, it is bell shaped but a little bit
skewed to the right. The histogram for Reviews is skewed right and it
does not look like normal distribution. The boxplot for rating is
normal, and for reviews it is not normally distributed. The reviews
boxplot is positively skewed. On box boxplots we can see the Min, max,
and first and third quartiles as well as IQR and median.

# Task 2

# Histogram

hist(category1, right = FALSE, ylim = c(0, 100), main = "Histogram of
Category", xlab = "Category")

![](images/Task2_Histogram%20of%20Category.png)
```

The Category is a categorical nominal variable and Content.Rating is a categorical ordinal variable. The Content.Rating variable contains people ages such as teen or Mature 17+ and etc. and can be sorted. When the Category variable contains three types such as family, game and tools and can not be sorted. The Category and Content.Rating variables have a length of 164 and class of character. In content rating, we have 28 teens, 8 mature 17+, 10 everyone 10+ and 118 everyone, thus, we know that the amount of people who use the specific app and categorized by their aged. Also, based on category column we know under what category each app. For example, Crazy Doctor app is for Family.\
In category we have 88 for family values, 41 for game and 35 for tools. Based on the histogram for Category, it is clear that it is not normally distributed.

Task 3

Graphs

```
boxplot(Rating \~ Category, data = mydata)
```



```
boxplot(Reviews \~ Category, data = mydata)
```


Summary Statistics

We can see that the code above shows us the means and standard deviations of the variables Rating and Reviews sorted by Category. Based on the code above we also see that Game Category by Reviews has the highest mean 480220.93 and standard deviation 1381986.3 and Family categorized by Reviews has the lowest mean 83313.22 and standard deviation 216613.4. As for Rating variable, the category Family has the highest mean 4.314773 and standard deviation 0.2152463 and Tools has the lowest mean 4.271429 and the Game has the lowest standard deviation 0.1737674. The box plot for Rating by Category Game has pretty normal distribution, the box plot for Family right skewed and Tools is left skewed. I assume that the box plot for Reviews by Category does not looks like normally distributed at all also, the means in Reviews by Category are very different, the Family mean is 83313.22, the Game mean is 480220.93 and the Tools mean is 184386.14. The same difference is in standard deviations. Thus, we can say that variable Rating vary by Category due to small differences and variable Reviews does not vary by Category.

Task 4

Graph

```
boxplot(Content.Rating \~ Category, data = mydata)
```


We have a variable Content.Rating by Category, the highest mean has category Tools 4.000000 and the lowest mean 2.365854 has category Game.The highest standard deviation 1.337088 has Game category and the lowest standard deviation 0.000000 has Tools category. We can notice that the there is a small difference in means and standard deviations. As for the box plot, the Game category is skewed right and Family and Tools is very squeezed. Based on the means and standard deviations little differences I think that Content.Rating vary by Category.

Task 5

```
# App Size Hypothesis

Ho: mydata\$App_Size = Normal

H1: mydata\$App_Size != Normal

# App Size Graph

hist(mydata\$App_Size, right = FALSE, main = "Histogram of App Size",
xlab = "App Size")
```



```
# Tests

For App Size, we can see that the the histogram is not normally
distributed, it is skewed to the right, so it is positively skewed. We
can notice that the frequencies on the left side of the graph are higher
than on the right side. Skewness of the App Size is 0.8831487, it means
it is pretty close to 0 and could be normally distributed because the
closer the skewness value to 0 the more likely it to be normally
distributed. Kurtosis for App Size is -0.1156519, in this case, we say
that the more it closer to zero, it means that it is normally
distributed. Based on Shapiro-Wilk Test we see that the p-value is
2.633e-09, it is small compare to α. Thus, do not reject Ho. There is
not enough evidence that the data is not normal.

# Rating Hypothesis

Ho: mydata\$Rating = Normal

H1: mydata\$Rating != Normal

# Rating Graph

hist(mydata\$Rating, right = FALSE, ylim = c(0,35), main = "Histogram of
Rating", xlab = "Rating")
```



```
# Tests

For Rating histogram, we see that it is bell shaped, slightly skewed to
the left. Based on it we can say that it is pretty normally distributed.
Skewness for App Size is 1.296361e-15 which is not close to 0. Thus, we
can say that it is not normally distributed. As for kurtosis, it is
-0.3117621, it is small and pretty close to 0 too. Shapiro-Wilk Test
shows that the p-value is a little bigger compare to α.Thus, do not
reject Ho. There is not enough evidence that the data is not normal.

# Task 6

# Graph

hist(genre1, right = FALSE, ylim = c(0, 25), main = "Histogram of
Genre", xlab = "Genre")
```



```
We can see that the variable before converting it to factor and numeric
was a character. The variable Genre contains 22 games of Action and
Adventure genre, 11 games of Arcade genre, 6 games of Board and Card
genres and 2 games of Puzzle genres. The genre variable has mean of
1.707317 and standard deviation of 0.9012187. The histogram of Genre is
right skewed, which means it is positively skewed and the highest
```

```
frequencies are more to the right and low frequencies are on the left
side of the histogram. Thus, we can conclude that it is not normally
distributed.

# Task 7

# Graph

boxplot(mydata2$Size_Tier ~ mydata2$Genre, data = mydata2, xlab =
"Genre", ylab = "Size Tier")

![](images/Task7_Boxplot%20of%20Size_Tier~Genre.png)

After doing some summary statistics and graph on Size_Tier variable by
Genre, we see that Size Tier has 10 Small App Sizes, 20 moderate size
and 11 large size. Also, we can notice that the highest mean is 2.500000
and the lowest mean is 1.909091. The highest standard deviation is
0.8312094 and the lowest is 0.6900656. The mean of Size Tier is 2.02439
and the mean of Genre is 1.707317. The SD of Size Tier is 0.7241479 and
the SD of Genre is 0.9012187. We see that the differences between means
and standard deviations are not that significant, there are some small
differences. As for the box plot of Size Tier and Genre, we can notice
that it is does not look like normally distributed. Also, we see that
Arcade, Board and Card and Puzzle Genres are mostly skewed and Action
and Adventure genre is squeezed. Thus,since the difference between means
and standard deviation are small, we can assume that Size Tier vary by
Genre.
```