Assignment 1 Q2

Nimit Bhanshali

2022-07-17

Part 1

1 (i)

```
data <- read_csv( # reading the .csv file
  file = "C:/Users/nbhan/OneDrive - University of Toronto/Year 2/STA238/R/Assignment 1/toronto-apartmen
  col_names = TRUE
)
```

```
## Rows: 3446 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): EVALUATION_COMPLETED_ON, PROPERTY_TYPE, RESULTS_OF_SCORE, SITE_ADD...
## dbl (27): _id, BALCONY_GUARDS, CONFIRMED_STOREYS, CONFIRMED_UNITS, ELEVATORS...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df <- data[c("WARD", "SCORE", "CONFIRMED_STOREYS")]
colnames(df) <- c("ward", "score", "storeys") # renaming the columns

ward11_obsv1 <- sum(df["ward"] == "11") # number of Ward 11 observations
ward13_obsv1 <- sum(df["ward"] == "13") # number of ward 13 observations

print(
  paste("There are", ward11_obsv1, "observations in Ward 11.")
  )
```

```
## [1] "There are 132 observations in Ward 11."
```

```
print(
  paste("There are", ward13_obsv1, "observations in Ward 13.")
  )
```

```
## [1] "There are 208 observations in Ward 13."
```

1 (ii)

```
df1 <- na.omit(df) # removing observations without a score
ward11_obsv2 <- sum(df1["ward"] == "11") # updated number of Ward 11 observations
ward13_obsv2 <- sum(df1["ward"] == "13") # updated number of Ward 13 observations

print(paste("There were", ward11_obsv1 - ward11_obsv2,
            "observations removed from Ward 11."))
```

```
## [1] "There were 0 observations removed from Ward 11."
```

```
print(paste("There were", ward13_obsv1 - ward13_obsv2,
            "observationss removed from Ward 13.") )
```

```
## [1] "There were 1 observationss removed from Ward 13."
```

By removing properties with missing scores, we are assuming that these properties are no longer considered as apartment options in the data we will proceed to analyze.

Such an assumption is made because if were to keep the properties with missing scores, we would unable to compare the quality of such apartments to others as there is no variable to compare them with. Hence, keeping properties with missing scores would prevent us from achieving the goal of this analysis, which is to compare the quality of apartment options.

Part 2

2 (i)

```
# Summary statistics for apartments scores: mean, median, standard deviation
mean(df1$score)
```

```
## [1] 72.28048
```

```
median(df1$score)
```

```
## [1] 72
```

```
sd(df1$score)
```

```
## [1] 7.117172
```

2(ii)

```
ward11 <- filter(df1, ward %in% c(11))
ward13 <- filter(df1, ward %in% c(13))

# Statistics for Ward 11 apartments scores: mean, median, standard deviation
print("Statistics for aparments in Ward 11")
```

```
## [1] "Statistics for aparments in Ward 11"
```

```
mean(ward11$score)
```

```
## [1] 70.41667
```

```
median(ward11$score)
```

```
## [1] 71
```

```
sd(ward11$score)
```

```
## [1] 7.666763
```

```
# Statistics for Ward 13 apartments scores: mean, median, standard deviation
print("Statistics for aparments in Ward 13")
```

```
## [1] "Statistics for aparments in Ward 13"
```

```
mean(ward13$score)
```

```
## [1] 72.36232
```

```
median(ward13$score)
```

```
## [1] 73
```

```
sd(ward13$score)
```

```
## [1] 6.965179
```

2(iii)

Overall, we see that the quality of apartments in Ward 13 is better than the quality of apartments in Ward 11. This is evident when comparing the statistics of the apartments score in both wards. We see that the mean apartment score in Ward 11, 70.417, is lower by approximately 2 units than the mean apartment score in Ward 13, 72.362. This implies that on average, quality of apartments in Ward 13 is better than the quality of apartments in Ward 11. Additionally, given that the median apartment score for both wards are rather close to their mean, we can conclude that the apartment scores in both wards are about evenly distributed and that there are not many outliers that have influenced the mean apartment scores of either ward. Lastly, the standard deviation of the apartment scores for both wards, tell us that the data doesn't deviate too much from the mean, even less so for the data from Ward 13. From this we can conclude, that the mean apartment scores are a rather strong representative of the apartment scores for both wards, 11 and 13. Thus, allowing us to claim that the quality of apartments in Ward 13 are better than the quality of apartments in Ward 11.

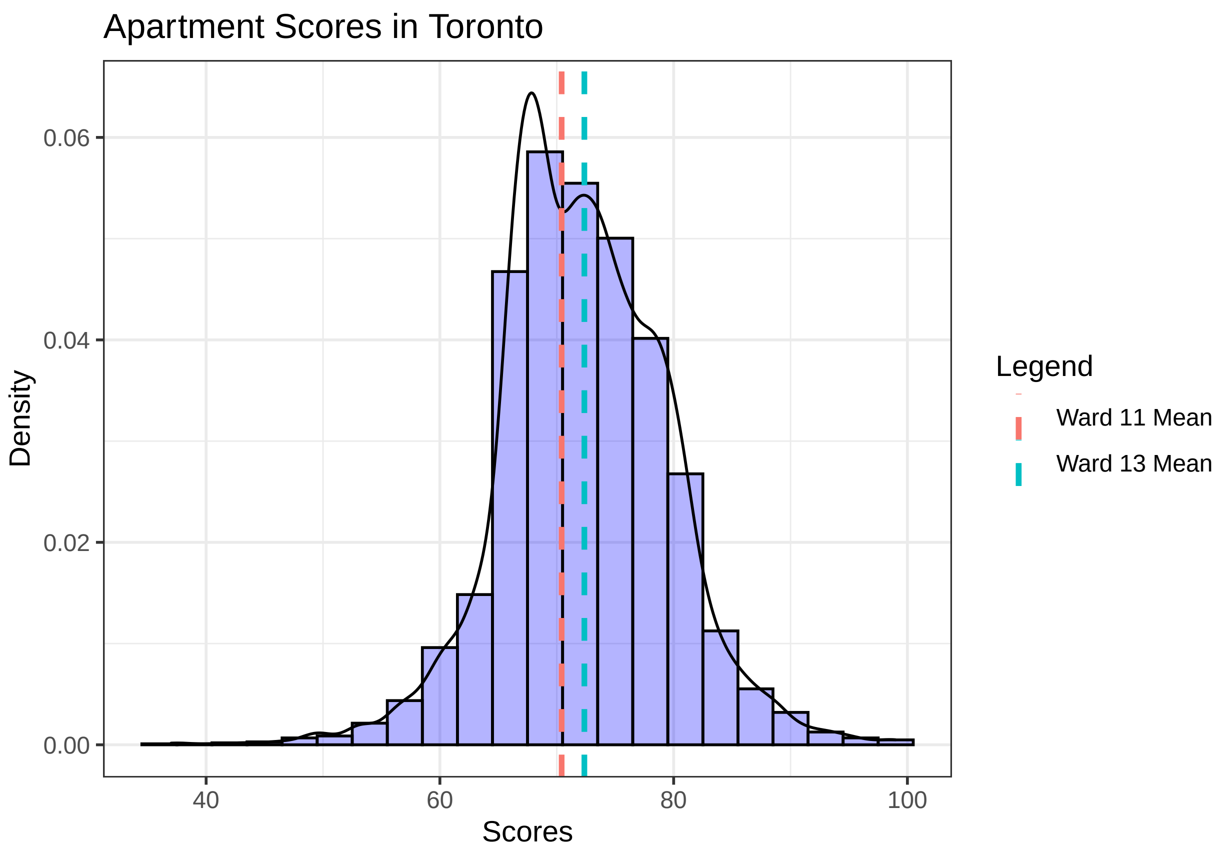
The quality of apartments in Ward 13 is better than the quality of apartments overall in Toronto for similar reasons. Firstly, we see that the mean apartment score in Ward 13, 72.362, is greater than the mean apartment score overall in Toronto, 72.280, implying that on average the apartment scores in Ward 13 are higher than the apartment scores overall in Toronto. Furthermore, medians that are close to their respective means and similar standard deviations inform us of a roughly even spread of data with few outliers, making the mean apartment scores a strong representatives of the apartments scores in Ward 13 as well as the apartment scores in Toronto overall. Hence, we can conclude that the quality of apartments in Ward 13 are better than the quality of apartments overall in Toronto.

On the other hand, quality of apartments in Ward 11 are worse than the quality of apartments overall in Toronto. First indication of this can be seen in the mean apartments scores, where the mean apartments scores in Ward 11, 70.417, are lower than the mean apartments scores overall in Toronto, 72.280. Like mentioned above, due to median values being close to the mean and similar standard deviations, we can conclude that the data is evenly spread with very few outliers. All of this contributes towards making a strong case for the mean apartment scores being a good representative for the apartments scores in Ward 11 as well as the apartment scores overall in Toronto. Hence, we can claim that the quality of apartments in Ward 11 are worse than the quality of apartments overall in Toronto.

Part 3

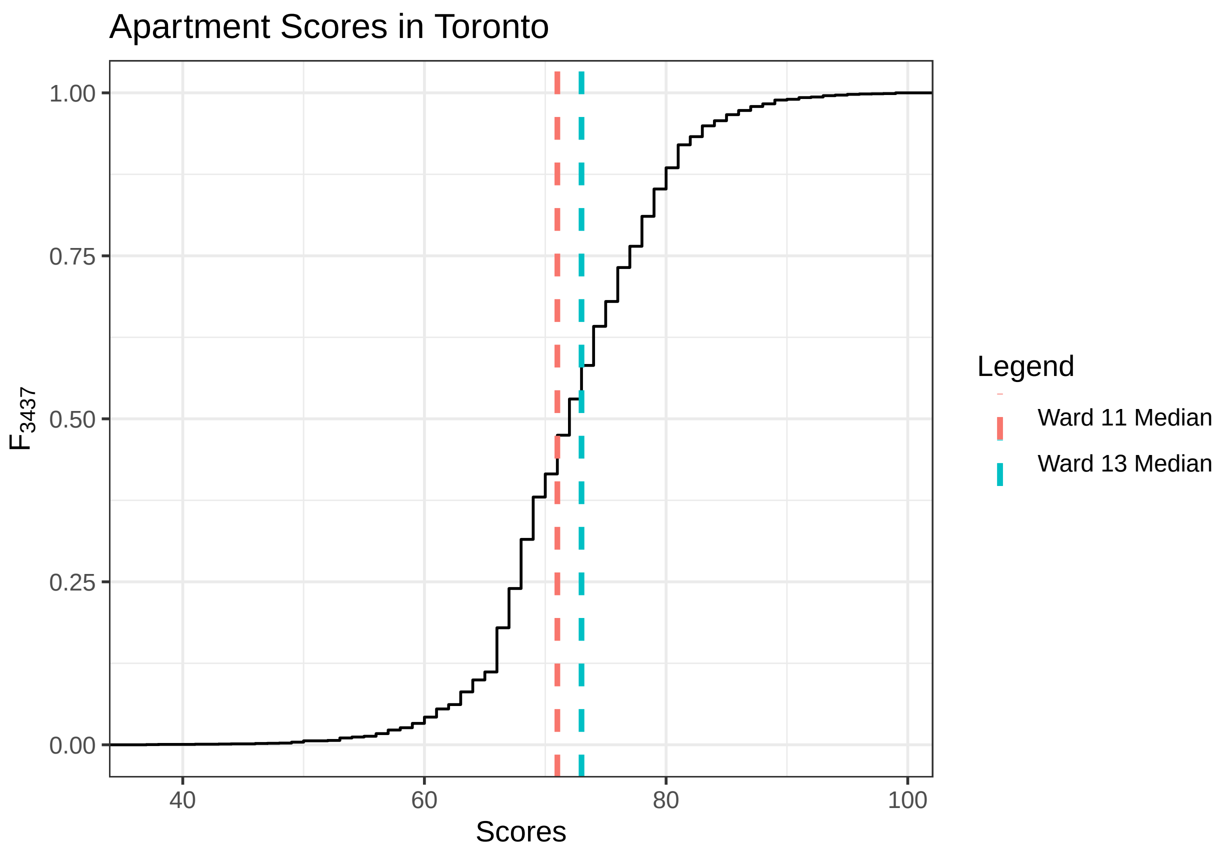
3(i)

```
ward11mean <- mean(ward11$score)
ward13mean <- mean(ward13$score)
df1 %>%
  ggplot(aes(x = score)) +
  theme_bw() +
  # Histogram of apartment scores in Toronto
  geom_histogram(aes(y = ..density..), binwidth = 3, colour = "black",
                 fill = "blue", alpha = .3) +
  # KDE of apartment scores in Toronto
  geom_density() +
  # Vertical line intercepting at the Ward 11 Mean
  geom_vline(aes(xintercept=ward11mean, color= "Ward 11 Mean"),
             linetype="dashed", size=1) +
  # Vertical line intercepting at the Ward 13 Mean
  geom_vline(aes(xintercept=ward13mean, color= "Ward 13 Mean"),
             linetype="dashed", size=1) +
  labs(title = "Apartment Scores in Toronto",
       x = "Scores",
       y = "Density",
       color = "Legend")
```



3(ii)

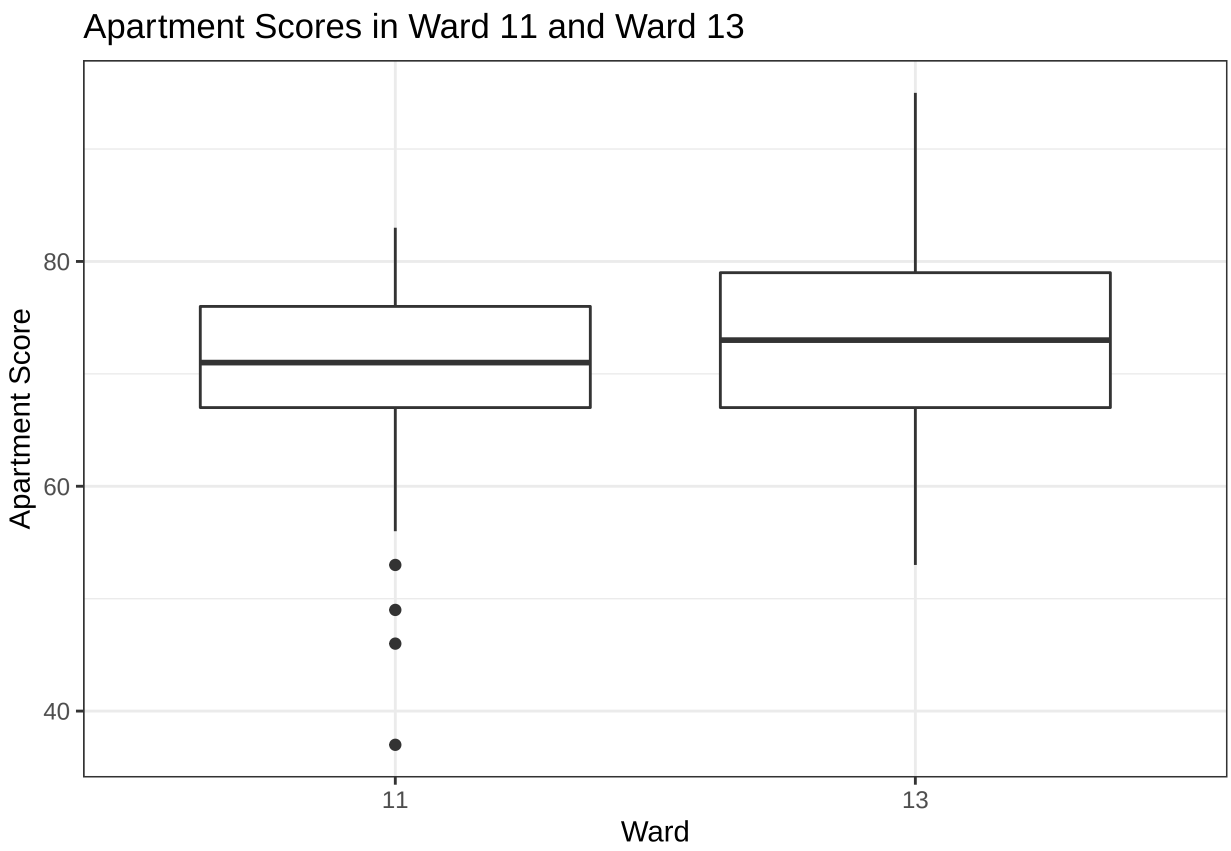
```
ward11med <- median(ward11$score)
ward13med <- median(ward13$score)

ggplot() +
  theme_bw() +
  # eCDF of Apartment Scores in Toronto
  stat_ecdf(aes(x = df1$score)) +
   # Vertical line intercepting at the Ward 11 Median
  geom_vline(aes(xintercept=ward11med, color= "Ward 11 Median"),
             linetype="dashed", size=1) +
   # Vertical line intercepting at the Ward 13 Median
  geom_vline(aes(xintercept=ward13med, color= "Ward 13 Median"),
             linetype="dashed", size=1) +
  labs(title = "Apartment Scores in Toronto",
       x = "Scores",
       y = expression(paste("F"["3437"])),
       color = "Legend")
```



You would add the median to the eCDF because a median is easily visible on an eCDF. By drawing a horizontal line at 0.5 on the y-axis, the corresponding value on the x-axis gives you the median of the dataset. Hence, median is added to the eCDF as it a measure of the center of a dataset that can be best determined visually. Whereas, a mean is added to the histogram and KDE because these visualizations don't work with quantiles and percentiles like eCDF. Hence, the more appropriate measure of the center of a dataset is the mean of the dataset instead. Therefore, we see the mean being added to the KDE and the histogram, while the median being added to the eCDF.

3(iii)

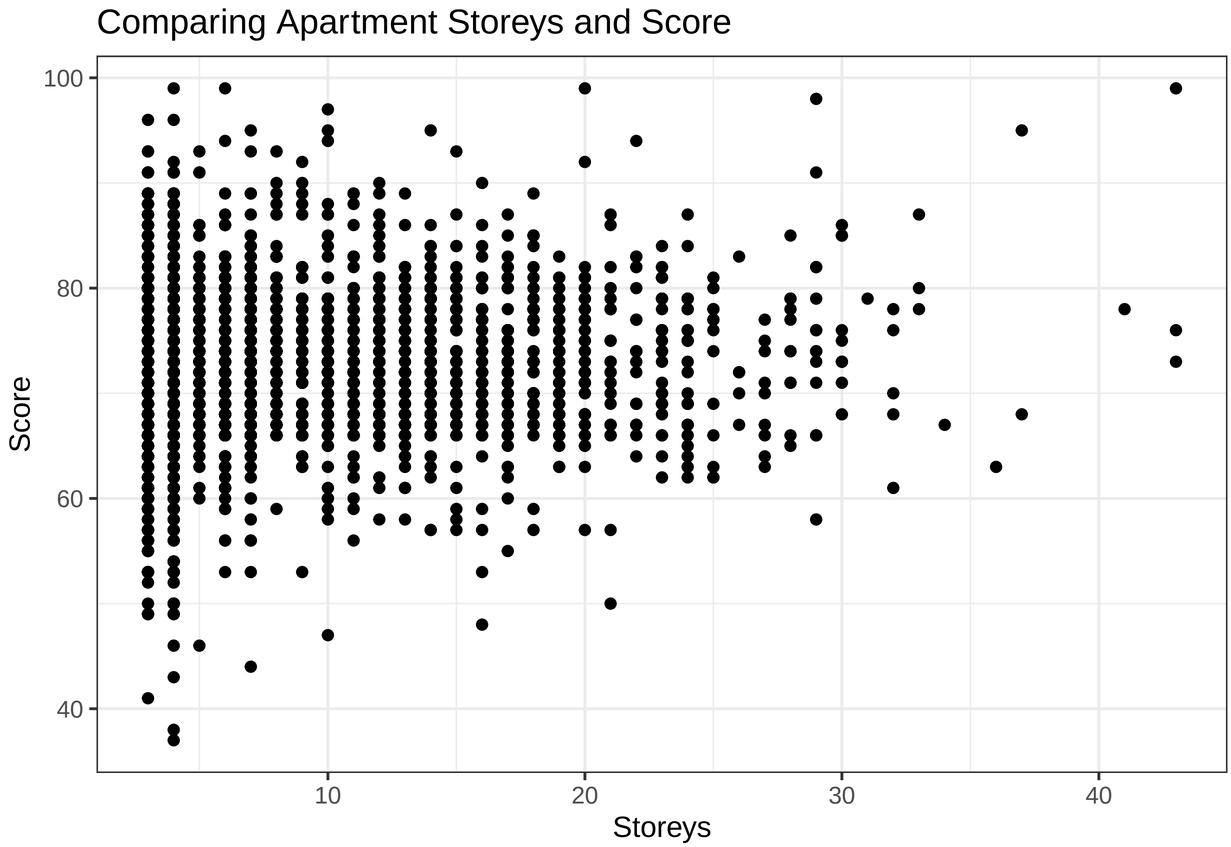
```
ward <- filter(df1, ward %in% c(11, 13))
ggplot() +
  # Boxplots to compare Ward 11 and Ward 13
  geom_boxplot(aes(x = ward$ward, y = ward$score)) +
  theme_bw() +
  labs(
    title = "Apartment Scores in Ward 11 and Ward 13",
    x = "Ward",
    y = "Apartment Score"
  )
```



Part 4

4(i)

```
df1 %>%
ggplot(aes(x = storeys, y = score)) +
  # Scatterplot comparing storeys and score
  geom_point() +
  theme_bw() +
  labs(
    title = "Comparing Apartment Storeys and Score",
    x = "Storeys",
    y = "Score"
  )
```



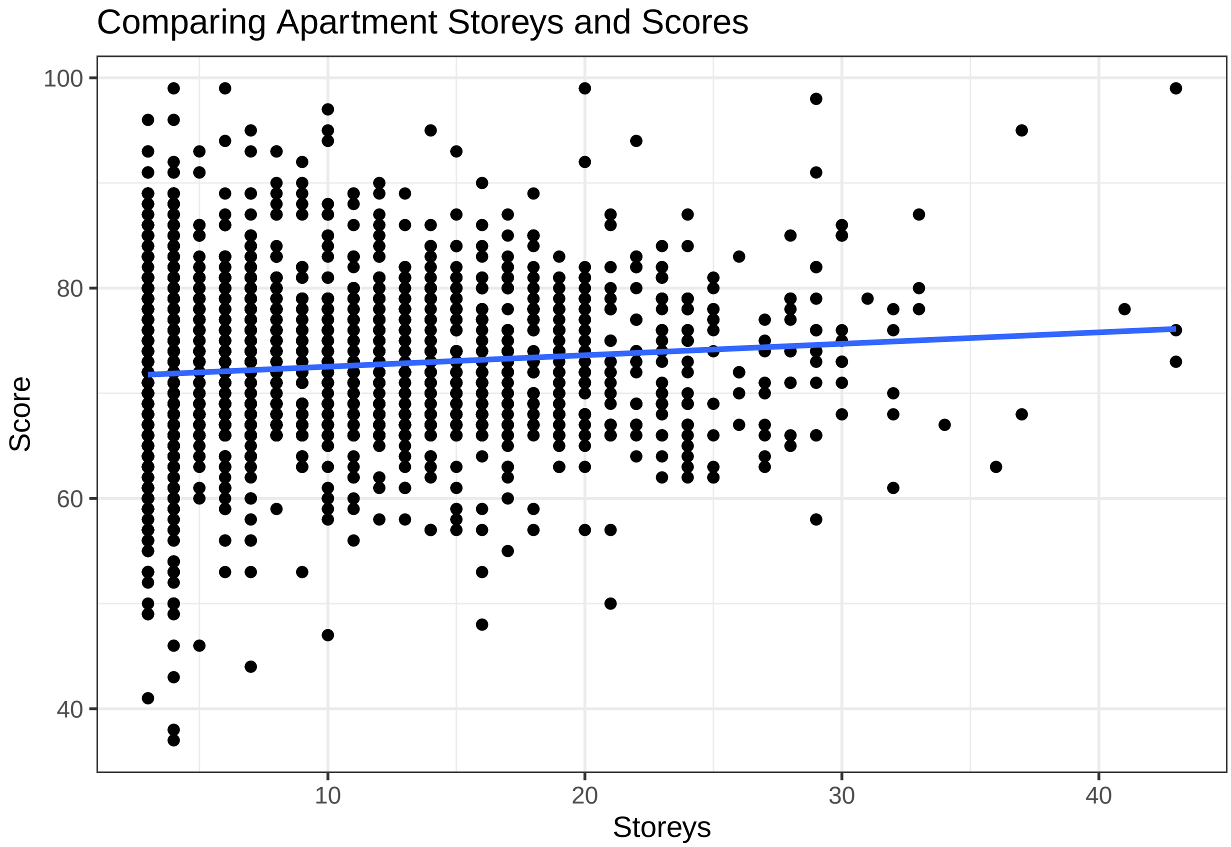
4(ii)

The parameters of this model are α and β, which represent the intercepts and slope, respectively. The U_i is considered to be the random component of this statistical model to account for the unknown. It is assumed that U_i independent and has an expectation of 0, implying that values of $score_i$'s are equally far away from the model are scattered around the model in a random manner. Another assumption is that U_i has a variance of σ^2, implying the same amount of variability about the model everywhere.

4(iii)

```
df1 %>%
ggplot(aes(x = storeys, y = score)) +
  geom_point() +
  theme_bw() +
  labs(
    title = "Comparing Apartment Storeys and Scores",
    x = "Storeys",
    y = "Score"
  ) +
  # Linear model to the scatterplot above
  geom_smooth(method = "lm", se=FALSE)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Part 5

Ward 11 overall has lower quality apartments compared to the rest of Toronto. This can be seen in both their mean and median values, both measures displaying higher values for the apartment scores. Whereas, Ward 13 is seen to have slight higher quality apartments compared to the rest of Toronto. Again, this is visible in the mean and median values for the apartment scores, which are both higher for Ward 13 than the rest of Toronto.

As for Ward 11 and Ward 13, it is evident that on average Ward 13 does have higher quality apartments in comparison to Ward 11 as seen in the mean and median values. However, it might be worth noticing the boxplots for the two wards that show Ward 11 having a few outliers. This may indicate that a few low quality apartments in Ward 11 may be influencing the mean apartment score in Ward 11. Additionally, although we see that the median and maximum value of Ward 13 apartment scores are higher than Ward 11 apartment scores, we also see that the lower quartile and the minimum value excluding outliers are higher for Ward 11 compared to Ward 13. The boxplots also shows the apartments scores to be spread over a greater range for Ward 13 compared to Ward 11. Hence, this may indicate that although on average you may get a lower quality apartment in Ward 11 in comparison to Ward 13, but the chance of getting apartments of a very low quality may be lower in Ward 11 than in Ward 13.

According to the scatterplot above, I would prefer to live in a high rise. This is because although the trend and the simple linear regression model is not too strong, it is indicative that apartments with higher storeys do tend to have a better apartment score. Hence, in the search for a higher quality apartment, I would prefer to live in a high rise.

However, there are some limitations of this analysis. Firstly, the number of Ward 11 apartment samples are significantly less then the number of Ward 13 apartment samples. Hence, we do not know whether these samples are appropriate representatives of all the apartments in Ward 11 and Ward 13. Additionally, in this analysis, we have chosen to specify on very few factors that help in deciding between rental housing. Hence, the results of this analysis do not help in making the decision of which ward have better quality of apartment options. Additionally, this analysis didn't really focus on the gaps and modes in the apartments scores of Ward 11 and Ward 13. This analysis focused on these attributes of apartment scores in Toronto overall but not specifically for Ward 11 and 13 which are the two wards we are comparing. Being able to visualize the shape of the data will have allowed us to make a more informed decision on which ward has better quality of apartment options. These are some limitation in our analysis that prevent us from making an informed decision regarding which ward has better quality of apartment options.