Q1.

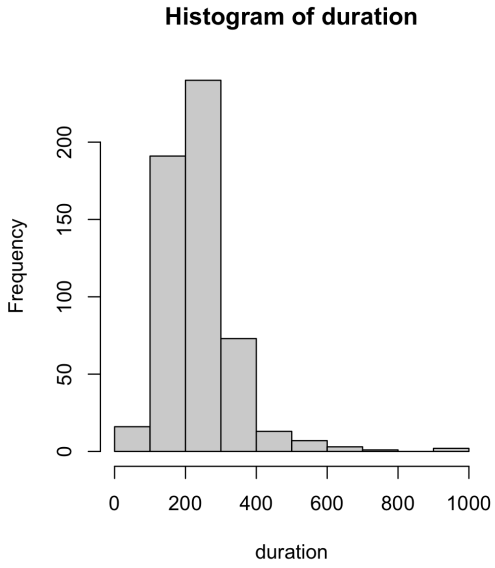
In this assignment you will be using R to analyze data on the million songs dataset (available in Canvas). I want to know whether songs from different genre's have significantly different durations. Please put your answers into words (don't just cut and paste R output).

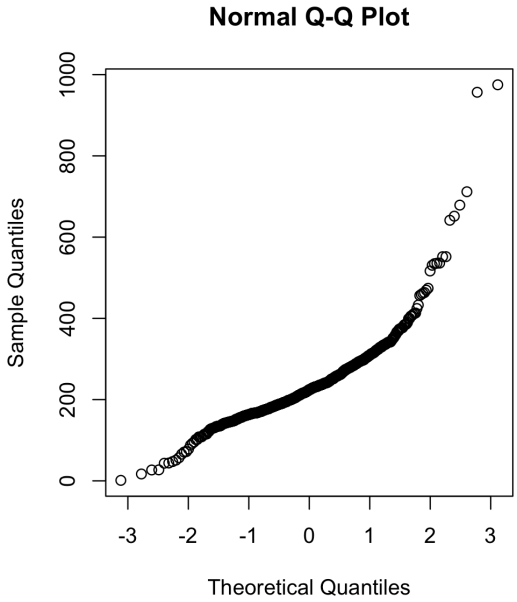
a) Report descriptive statistics for each group (mean song duration and its standard deviation).

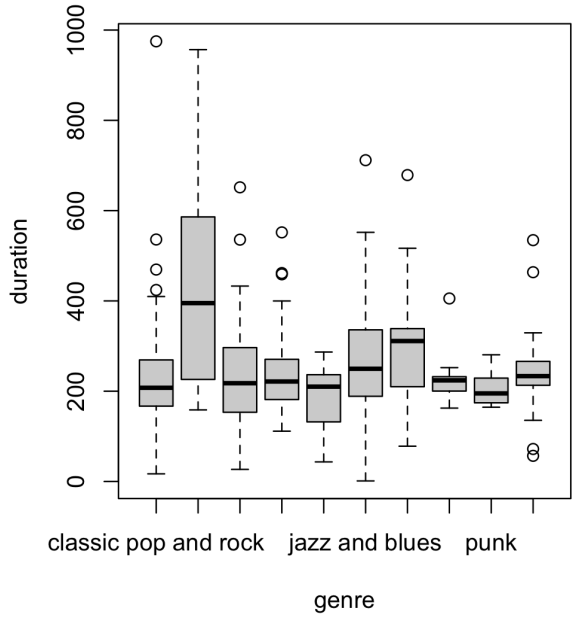
- Here we use the aggregate or tapply function. All of these descriptive statistics pertain to the duration of each genre.
    - classic pop and rock
        - Min = 16.92689
        - Median = 207.5816
        - Max = 974.9938
        - Standard deviation = 91.36827
        - Mean = 222.4497
    - classical
        - Min = 158.58893
        - Median = 395.1277
        - Max = 956.6036
        - Standard deviation = 264.86004
        - Mean = 441.2338
    - dance and electronica
        - Min = 26.80118
        - Median = 217.7040
        - Max = 651.5979
        - Standard deviation = 135.16963
        - Mean = 236.8892
    - folk
        - Min = 111.46404
        - Median = 221.4657
        - Max = 551.8102
        - Standard deviation = 73.33077
        - Mean = 232.2343
    - hip-hop
        - Min = 43.38893
        - Median = 210.0763
        - Max = 286.7718
        - Standard deviation = 91.93833

        - Mean = 182.4872
- jazz and blues
    - Min = 1.22730
    - Median = 249.6649
    - Max = 711.7057
    - Standard deviation = 123.78682
    - Mean = 273.9929
- metal
    - Min = 78.18404
    - Median = 311.1048
    - Max = 678.7653
    - Standard deviation = 112.05150
    - Mean = 295.6404
- pop
    - Min = 162.55955
    - Median = 223.9734
    - Max = 405.4199
    - Standard deviation = 59.83761
    - Mean = 228.2213
- punk
    - Min = 164.54485
    - Median = 195.1081
    - Max = 280.6591
    - Standard deviation = 37.84723
    - Mean = 206.7587
- soul and reggae
    - Min = 56.84200
    - Median = 233.4298
    - Max = 534.6216
    - Standard deviation = 71.65220
    - Mean = 241.0172

b) Evaluate the ANOVA assumptions of normality and equality of variance across groups. *In your write-up, please include the figures that you created.*







- When we look at the histogram and qq plot, the distribution is clearly not normal.
- Levene's Test for Homogeneity of Variance gives a p-value of 2.246e-07, rejecting the null and indicating that the variance among the three groups is not equal.

c) Conduct a one-way ANOVA in R to address the research question. Report all relevant components of a hypothesis test, including test statistic, degrees of freedom, p-value.

- Using the function aov, we get

```
> aov(duration~genre)
Call:
   aov(formula = duration ~ genre)

Terms:
                    genre Residuals
Sum of Squares     581999   4960031
Deg. of Freedom         9       536

Residual standard error: 96.19662
Estimated effects may be unbalanced
```

  - We will use summary() on this.

```
              Df  Sum Sq Mean Sq F value   Pr(>F)
genre          9  581999   64667   6.988 1.43e-09 ***
Residuals    536 4960031    9254
```

  - The degrees of freedom is 9, and the F value is 6.988.
  - Looking at these results, the p-value is very small, indicating that we can reject the null hypothesis and there are significant differences between the groups in the model summary.

d) One-way ANOVA is an omnibus test, which means if you reject the null hypothesis it tells you that at least one mean is different, but it doesn't tell you which ones are different. Using R, get pairwise comparisons with Bonferroni correction. Based on the results, report which, if any, genres are statistically significant from each other (Sig gives p-value for a particular t-test). What would you conclude based on your omnibus test and pairwise comparisons?

- The adjusted p-value for the mean difference in duration between
  - classic pop and rock and classical
  - dance and electronica and classical
  - hip-hop and classical
  - jazz and blues and classical
  - metal and classic pop and rock
  - metal and classical
  - punk and classical
  - soul and reggae and classical
  - metal and folk
- show that there is a significant difference between these genres

e) ANOVA might not be the appropriate analysis approach here? Explain why. (No analysis is needed for this question).

- ANOVA assumes that the responses for each factor level are normally distributed; which they are clearly not in the graphs generated.
- In addition, ANOVA assumes the distributions have the same variance, which Levene's test shows they do not.
- Finally, the data is not independent.

Q2.

Becky Liddle at Auburn University published a study in 1997 on disclosing sexual orientation in class. She taught four sections of the same class, and at the week of the final lecture she disclosed her lesbian identity to two of the sections, and withheld it from the two others. She was concerned with the issue of whether disclosure would influence student evaluations of the course. The means and average variance for the two conditions, further broken down by gender of the students, are presented below. There were 15 students in each cell. Perform a two-way analysis of variance and draw the appropriate conclusions. Please report the F value, degrees of freedom, and p value for every test. Use the fake Liddle Data provided on Canvas.

(Note: The means are the same that Liddle found, but because I could not control for difference in *mid-term* evaluations, as she did, the effect of gender is different from the effect she found. The other effects lead to similar conclusions.)

Please answer the following questions:

a) Is there a significant main effect of disclosure? If so, what do the results mean?

- Running the ANOVA without interaction.
    - The p value of disclose is 0.366, which is not significant, indicating that disclosure does not affect rating.
    - The p value of gender is 2.99e-07, a significant difference, indicating that gender affected rating.
- Running the ANOVA with interaction.
    - The p-value of the interaction is 0.0455, which is significant, indicating that the relationship between disclosure and rating depends on gender.

```
> summary(aov(ratings~disclose + gender))
            Df Sum Sq Mean Sq F value   Pr(>F)
disclose     1   10.7    10.7   0.831    0.366
gender       1  434.0   434.0  33.708 2.99e-07 ***
Residuals   57  733.9    12.9
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(ratings~disclose*gender))
               Df Sum Sq Mean Sq F value   Pr(>F)
disclose        1   10.7    10.7   0.877   0.3530
gender          1  434.0   434.0  35.590 1.73e-07
disclose:gender  1   51.0    51.0   4.184   0.0455
Residuals       56  682.9    12.2

disclose
gender          ***
disclose:gender *
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

b) Is there a significant main effect of gender? If so, what do the results mean?

- Main effect of disclosure: disclosure influenced ratings F = 0.877, p < 0.3530
- Main effect of gender: gender influenced ratings F = 33.708, p < 2.99e-07
- Interaction effect: disclosure had an effect when gender was involved F = 4.184, p < 0.0455

c) Is there a significant interaction between gender and disclosure? If so, produce an interaction graph and explain what the results mean.

- Looking at the interaction between gender and disclosure, there seems to be an effect. Producing an interaction graph shows that disclosing her lesbian identity to male students increased her ratings significantly. Not disclosing her identity to female students had a positive effect on her ratings.