


Do different genre's of songs have significantly different durations? Conduct a one-way ANOVA using data from [the million song database](#)  to answer this question. Draw a conclusion considering the results from your omnibus significance test, posthoc significance tests (i.e. corrected pairwise t-tests), effect size(s) and limitations of these tests. Provide a bar graph which displays mean song durations by song genre (it should include error bars i.e. 95% confidence intervals).

Robert Bruffey

Project 2: ANOVA

R Code for inputting data:

```
library(readr)
#import and view the data set
Songs <- read_csv("~/Downloads/MillionSongsFinal.csv")
View(Songs)
```

Research Question:

Do different genres of songs have significantly different durations?

Genres of the data:

- Classic pop and rock
- Classical
- Dance and electronica
- Folk
- Hip-hop
- Jazz and blues
- Metal
- Pop
- Punk
- Soul and reggae

One-way ANOVA

Step 1:

$H_0: \mu_{\text{classicpopandrock}} = \mu_{\text{classical}} = \mu_{\text{dance}} = \mu_{\text{folk}} = \mu_{\text{hiphop}} = \mu_{\text{jazz}} = \mu_{\text{metal}} = \mu_{\text{pop}} = \mu_{\text{punk}} = \mu_{\text{soul}}$

There is no difference in the population means of the durations of each genre's songs and therefore genre's of songs do not have significantly different durations.

H_1 : At Least one of the population means of the durations of a genre is not equal to the others and therefore, genre's of songs do have significantly different durations.

Step 2: Calculate test statistic and p-value

R code:

```
a = aov(Songs$duration~Songs$genre, data = Songs)
summary(a)
```

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

$SS_{\text{between}} = 581999$

$SS_{\text{within}} = 4960031$

$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}} = 5542030$

Based on the ANOVA test run from the R code above, we see that the p-value is 1.43e-09, which is way less than our significant level of .05. Therefore, we can reject our null hypothesis and conclude that there is at least one difference in the mean durations of a genre's songs.

Corrected Pairwise Comparison

R code:

```
pairwise.t.test(Songs$duration, Songs$genre, p.adj = "bonferroni")
```

```
Pairwise comparisons using t tests with pooled SD

data: Songs$duration and Songs$genre

      classic pop and rock classical dance and electronica folk
classical      2.5e-08          -          -          -
dance and electronica 1.00000      6.2e-06      -          -
folk            1.00000      2.1e-07      1.00000      -
hip-hop         1.00000      1.3e-05      1.00000      1.00000
jazz and blues   0.06805      0.00036      1.00000      0.70095
metal           0.00289      0.00648      0.74653      0.04346
pop             1.00000      5.0e-05      1.00000      1.00000
punk            1.00000      3.2e-05      1.00000      1.00000
soul and reggae  1.00000      2.5e-06      1.00000      1.00000

      hip-hop jazz and blues metal    pop    punk
classical      -      -          -      -      -
dance and electronica -      -          -      -      -
folk           -      -          -      -      -
hip-hop        -      -          -      -      -
jazz and blues 0.90784 -          -      -      -
metal          0.22478 1.00000      -      -      -
pop            1.00000 1.00000      1.00000 -      -
punk           1.00000 1.00000      0.65933 1.00000 -
soul and reggae 1.00000 1.00000      0.46895 1.00000 1.00000

P value adjustment method: bonferroni
```

Based off the Pairwise comparison using t tests with pooled SD, we can take a closer look at each individual genre's durations compared to every other genre. Noticeably, the classical music genre duration is completely different than every other genre's duration as the p-value compared to every genre is less than .05. Another difference I notice is that metal and classic pop and rock have a p-value of .002, therefore they have a low enough p-value to reject the null hypothesis and conclude to have different population mean durations. Finally, we can see that folk and metal also have a p-value of .043, which is less than .05 and so we can reject the null and conclude that folk and metal have different mean population durations of songs.

Effect Sizes

In order to calculate the effect size of the data (R^2) we need to know the SS_{between} , SS_{within} , and SS_{total} .

$$SS_{\text{between}} = 581999$$

$$SS_{\text{within}} = 4960031$$

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}} = 5542030$$

$$R^2 = SS_{\text{between}} / (SS_{\text{between}} + SS_{\text{within}}) = 581999 / 5542030 = .10502$$

The effect size value of .105 tells us that there is a 10.5% of variance in the DV(genre's) explained by IV(durations).

Limitations

The limitations of the previously done tests vary with each test and each is designed to provide a certain piece of information. Beginning with the omnibus ANOVA test, we were only limited to being able to determine if the data as a whole is different. In our test we were able to find that our p-value was less than our significance level of .05, so we could tell that there was at least one mean duration different from the others, but we could not determine which genre's mean it was. Next, for the pairwise comparison we were able to determine the p-values for each individual group to another group. This limited us to only being able to do two sampled t-tests and not allowing us to examine the data at a broader viewpoint such as viewing the data as sections of genres or all of the genres at once. Finally, for the effect size of the data we were only limited to finding the % of variance for all the genres to all the durations. The effect size didn't allow us to see the % of variance between each individual genre to their durations, which would have been useful to determine with genre has the most variance in the data.

Confidence Intervals

95% confidence interval

Means:

- $\mu_{\text{classical pop and rock}} = 222.45$
- $\mu_{\text{classical}} = 441.2338$
- $\mu_{\text{dance and electronic}} = 236.8892$
- $\mu_{\text{folk}} = 232.2343$

- $\mu_{\text{hip-hop}} = 182.4872$
- $\mu_{\text{jazz and blues}} = 273.9929$
- $\mu_{\text{metal}} = 295.6404$
- $\mu_{\text{pop}} = 228.2213$
- $\mu_{\text{punk}} = 206.7587$
- $\mu_{\text{soul and reggae}} = 241.0172$

Confidence intervals

Classical pop and rock

$\text{error} = \text{qt}(0.975, \text{df} = \text{length}(\text{poprock}\$duration) - 1) * \text{sd}(\text{poprock}\$duration) / \sqrt{\text{length}(\text{poprock}\$duration)} = 11.976$

$\text{left} = \text{mean}(\text{poprock}\$duration) - \text{error} = 210.473$

$\text{right} = \text{mean}(\text{poprock}\$duration) + \text{error} = 234.426$

Classical

$\text{error} = \text{qt}(0.975, \text{df} = \text{length}(\text{class}\$duration) - 1) * \text{sd}(\text{class}\$duration) / \sqrt{\text{length}(\text{class}\$duration)} = 221.4285$

$\text{left} = \text{mean}(\text{class}\$duration) - \text{error} = 219.8$

$\text{right} = \text{mean}(\text{class}\$duration) + \text{error} = 662.662$

Dance and Electronica

$\text{error} = \text{qt}(0.975, \text{df} = \text{length}(\text{dance}\$duration) - 1) * \text{sd}(\text{dance}\$duration) / \sqrt{\text{length}(\text{dance}\$duration)} = 50.473$

$\text{left} = \text{mean}(\text{dance}\$duration) - \text{error} = 186.416$

$\text{right} = \text{mean}(\text{dance}\$duration) + \text{error} = 287.362$

Folk

$\text{error} = \text{qt}(0.975, \text{df} = \text{length}(\text{f}\$duration) - 1) * \text{sd}(\text{f}\$duration) / \sqrt{\text{length}(\text{f}\$duration)} = 13.144$

$\text{left} = \text{mean}(\text{f}\$duration) - \text{error} = 219.091$

$\text{right} = \text{mean}(\text{f}\$duration) + \text{error} = 245.378$

Hip-Hop

$\text{error} = \text{qt}(0.975, \text{df} = \text{length}(\text{hiphop}\$duration) - 1) * \text{sd}(\text{hiphop}\$duration) / \sqrt{\text{length}(\text{hiphop}\$duration)} = 85.0287$

$\text{left} = \text{mean}(\text{hiphop}\$duration) - \text{error} = 97.4584$

$\text{right} = \text{mean}(\text{hiphop}\$duration) + \text{error} = 267.516$

Jazz and blues

$\text{error} = \text{qt}(0.975, \text{df} = \text{length}(\text{jazz}\$duration) - 1) * \text{sd}(\text{jazz}\$duration) / \sqrt{\text{length}(\text{jazz}\$duration)} = 38.574$

$\text{left} = \text{mean}(\text{jazz}\$duration) - \text{error} = 235.516$

$\text{right} = \text{mean}(\text{jazz}\$duration) + \text{error} = 312.567$

Metal

$\text{error} = \text{qt}(0.975, \text{df} = \text{length}(\text{m}\$duration) - 1) * \text{sd}(\text{m}\$duration) / \sqrt{\text{length}(\text{m}\$duration)} = 40.399$

$\text{left} = \text{mean}(\text{m}\$duration) - \text{error} = 255.241$

$\text{right} = \text{mean}(\text{m}\$duration) + \text{error} = 336.039$

Pop

```
error=qt(0.975,df=length(p$duration)-1)*sd(p$duration)/sqrt(length(p$duration)) = 36.159
left = mean(p$duration)-error = 192.062
right = mean(p$duration)+error = 264.381
```

Punk

```
error=qt(0.975,df=length(unk$duration)-1)*sd(unk$duration)/sqrt(length(unk$duration)) =
29.092
left = mean(unk$duration)-error = 177.667
right = mean(unk$duration)+error = 235.851
```

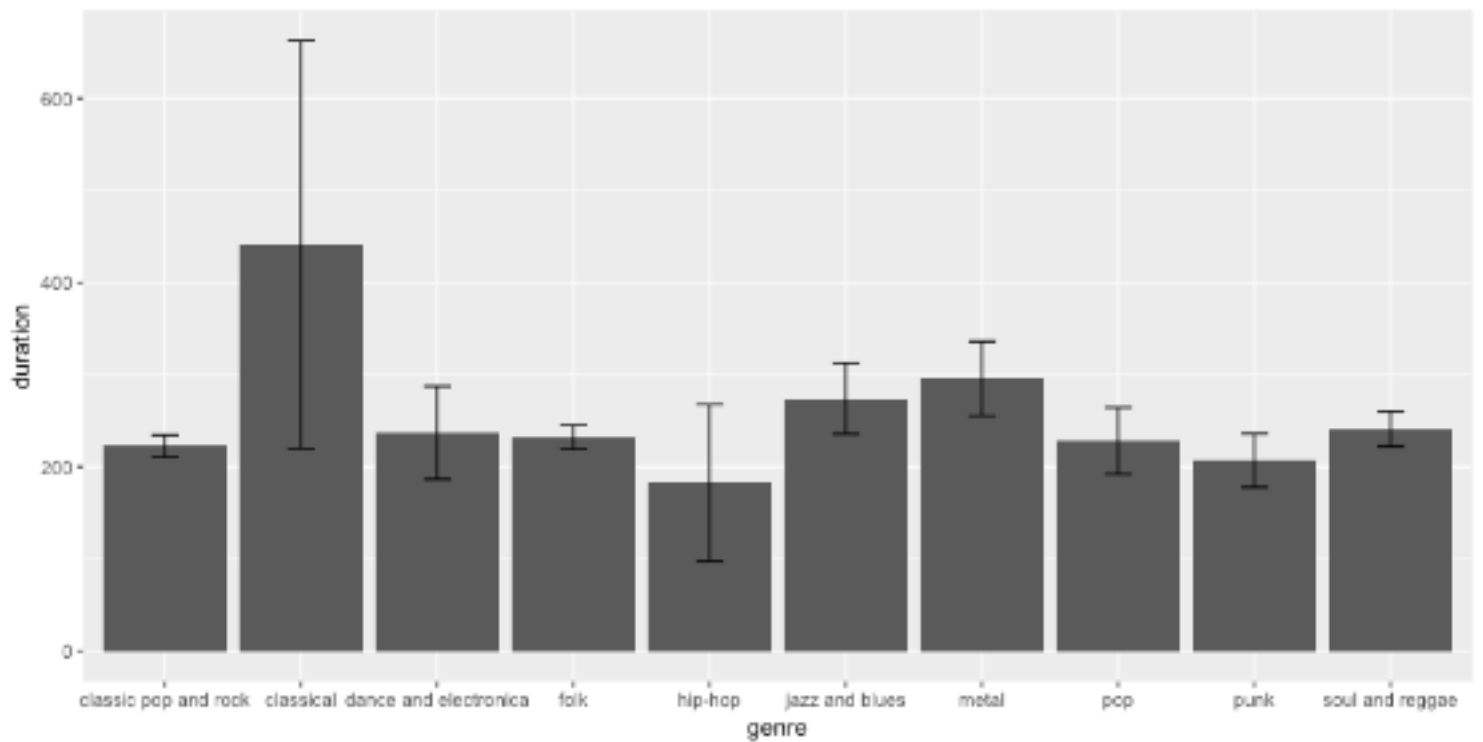
Soul and Reggae

```
error=qt(0.975,df=length(soul$duration)-1)*sd(soul$duration)/sqrt(length(soul$duration))
= 19.012
left = mean(soul$duration)-error = 222.005
right = mean(soul$duration)+error = 260.029
```

Bar Graph

R code:

```
>d.ci = data.frame(genre = c("classic pop and rock","classical","dance and
electronica","folk","hip-hop","jazz and blues","metal","pop","punk","soul and reggae"),duration =
c(mean(poprock$duration),mean(class$duration),mean(dance$duration),mean(f$duration),mea
n(hiphop$duration),mean(jazz$duration),mean(m$duration),mean(p$duration),mean(unk$durati
on),mean(soul$duration)),ci.upper =
c(234.426,662.662,287.362,245.378,267.516,312.567,336.039,264.381,235.851,260.029),ci.lo
wer = c(210.473,219.8,186.416,219.091,97.4584,235.516,255.241,192.062,177.667,222.005))
>library(ggplot2)
>ggplot(d.ci,aes(x=genre,y=duration))+geom_bar(stat="identity")+geom_errorbar
(aes(ymin=ci.lower,ymax=ci.upper),width=.2)
```



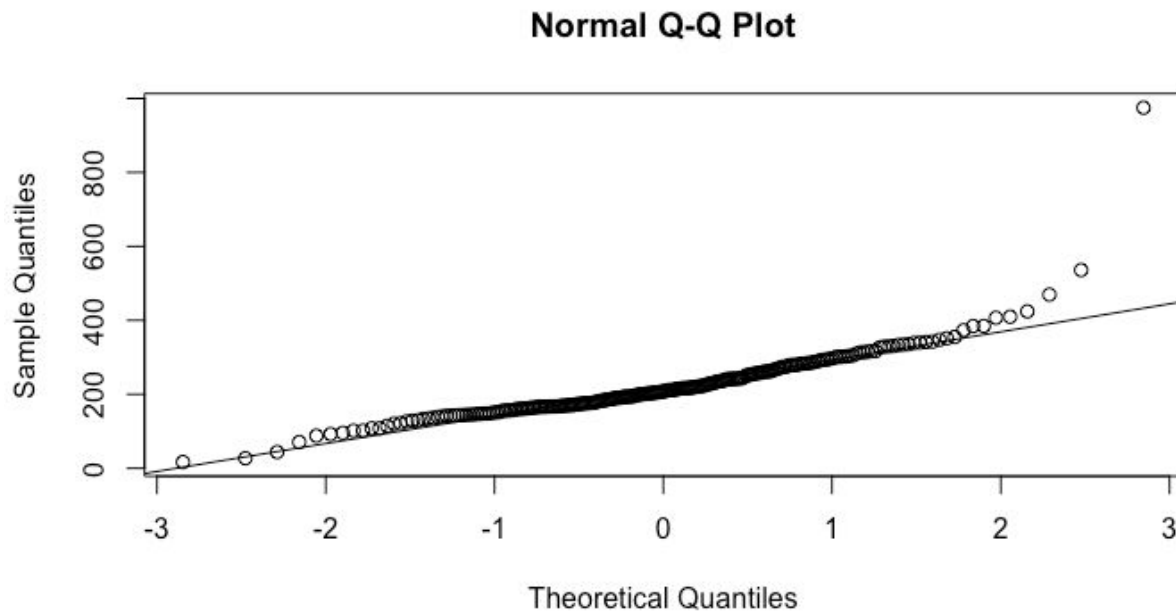
Based off the bar graph classical music tend to be the biggest outlier in the data and has the highest level of variance within its own group. Which would mean that the classical genre is the biggest reason why we reject the null hypothesis when we conduct an ANOVA test.

ANOVA Test Assumptions:

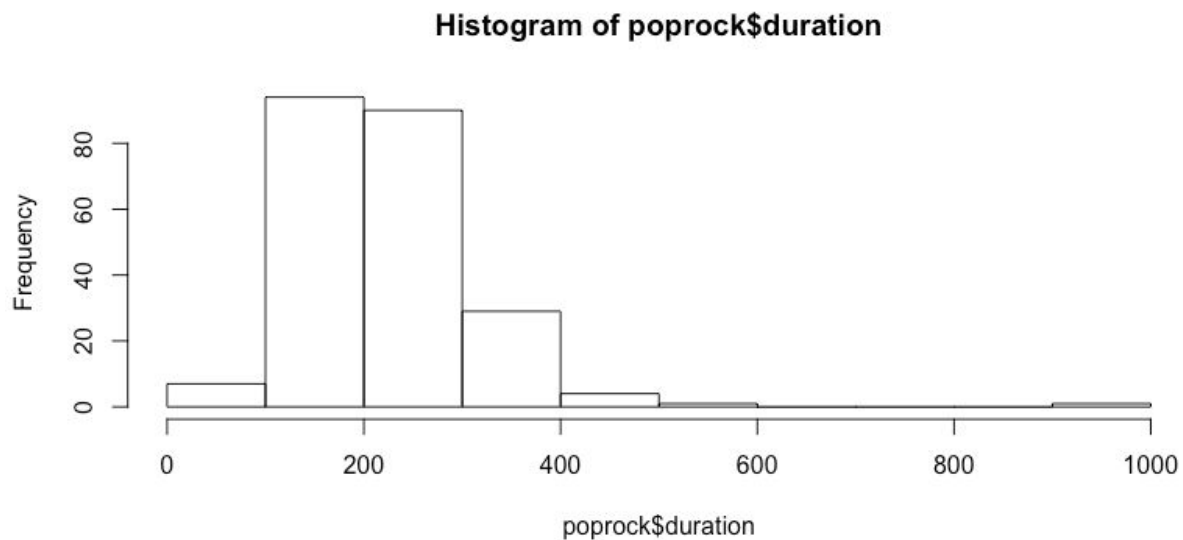
1. Normality

a. Classic pop and rock

- i. `poprock = subset(Songs, genre == "classic pop and rock")`
- ii. `qqnorm(poprock$duration)`
- iii. `qqline(poprock$duration)`



- iv. From the graph above you can see that the data for classical pop and rock is somewhat linear with a few data points outside the line of best fit.

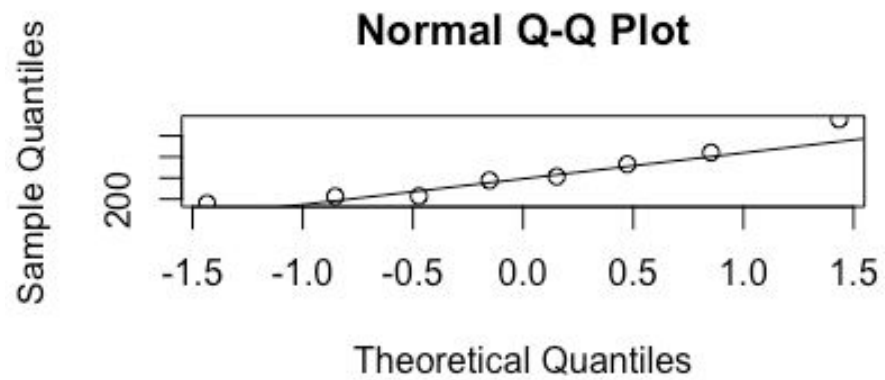


- v. `hist(poprock$duration)`

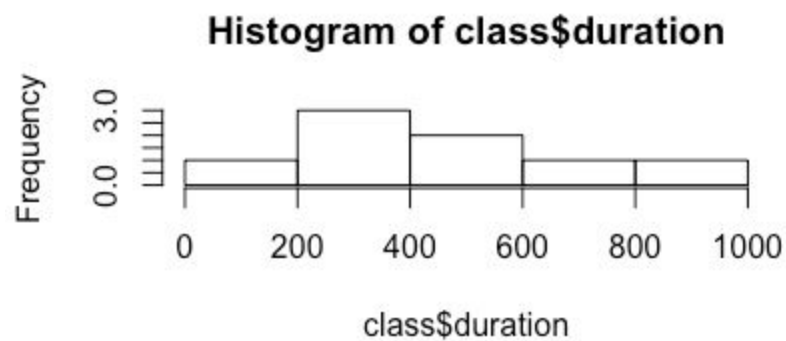
b. Classical

- i. `class = subset(Songs, genre == "classical")`

- ii. `qqnorm(class$duration)`
- iii. `qqline(class$duration)`



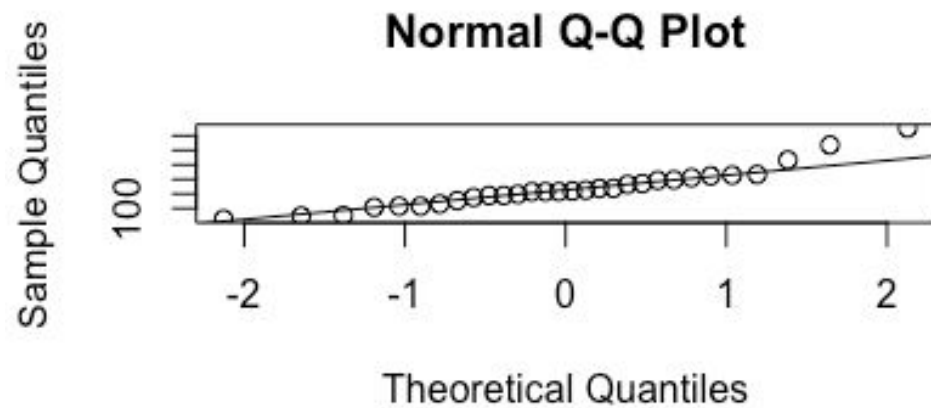
- iv.
- v. `hist(class$duration)`



- vi.

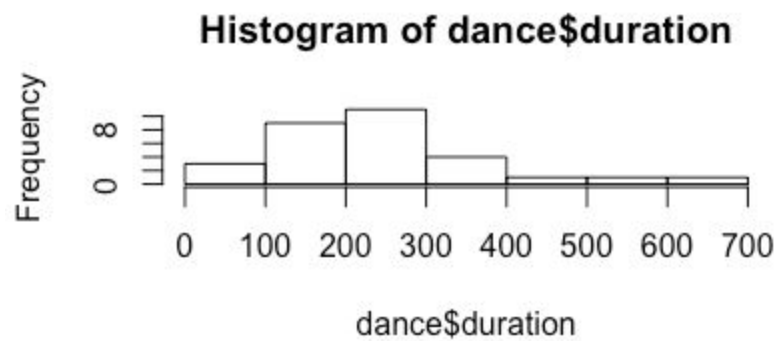
c. Dance and electronica

- i. `dance = subset(Songs, genre == "dance and electronica")`
- ii. `qqnorm(dance$duration)`
- iii. `qqline(dance$duration)`



iv.

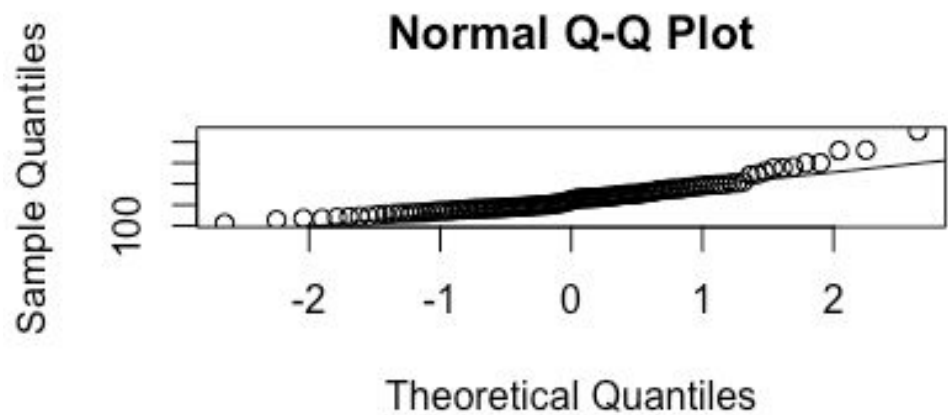
v. `hist(dance$duration)`



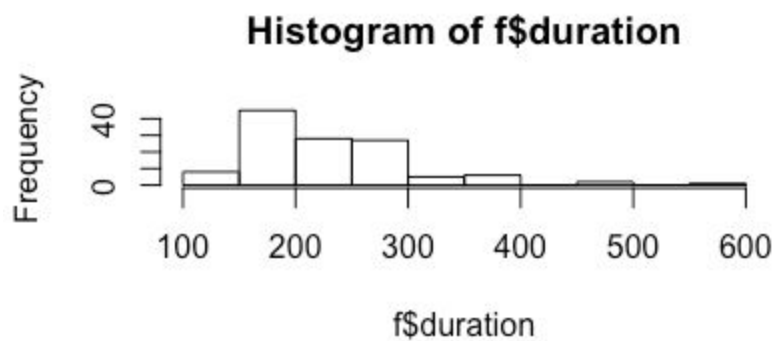
vi.

d. Folk

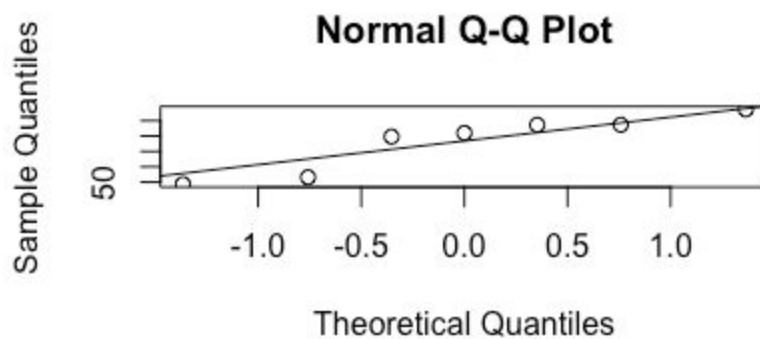
- i. `f = subset(Songs, genre == "folk")`
- ii. `qqnorm(f$duration)`
- iii. `qqline(f$duration)`



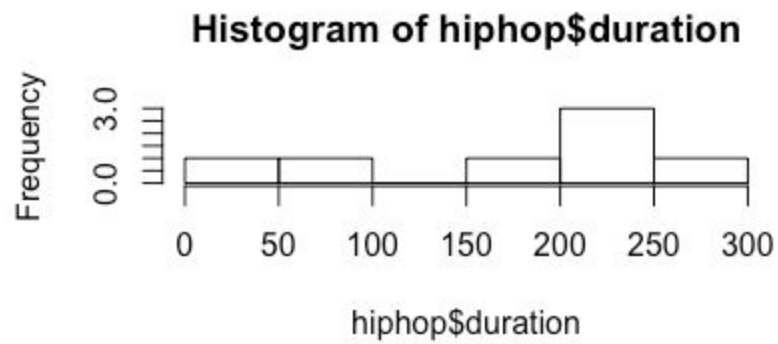
- iv.
- v. `hist(f$duration)`



- vi.
- e. Hip-hop**
- i. `hiphop = subset(Songs, genre == "hip-hop")`
 - ii. `qqnorm(hiphop$duration)`
 - iii. `qqline(hiphop$duration)`



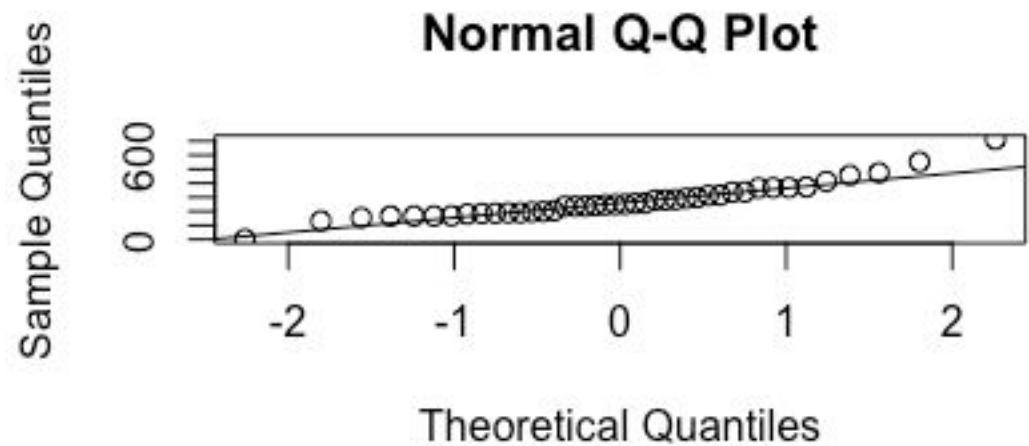
- iv.
- v. `hist(hiphop$duration)`



vi.

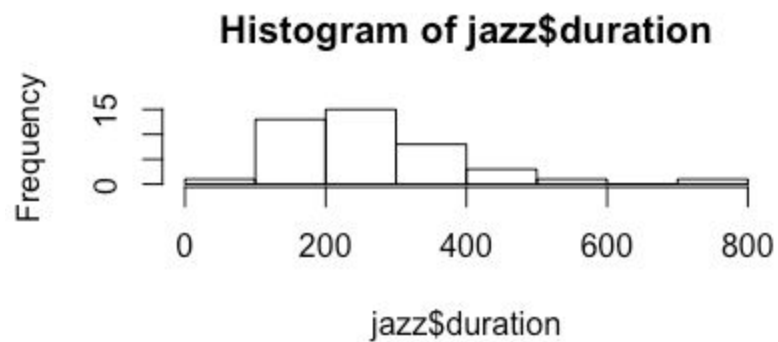
f. Jazz and blues

- i. `jazz = subset(Songs, genre == "jazz and blues")`
- ii. `qqnorm(jazz$duration)`
- iii. `qqline(jazz$duration)`



iv.

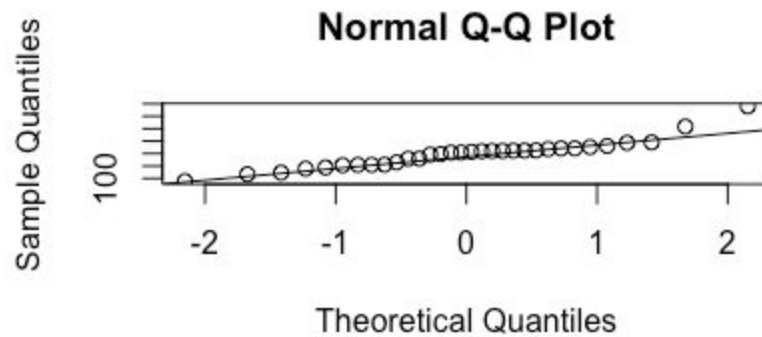
- v. `hist(jazz$duration)`



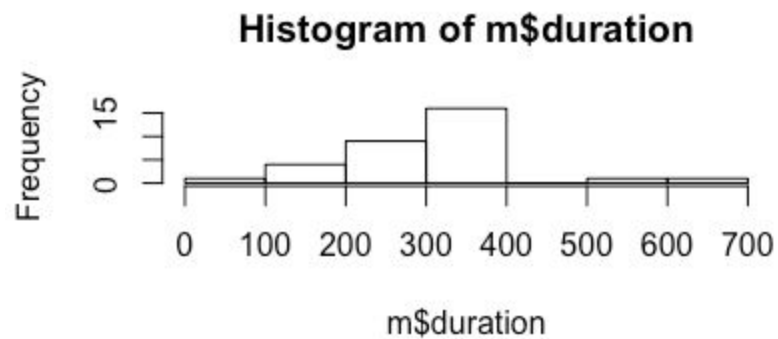
vi.

g. Metal

- i. `m = subset(Songs, genre == "metal")`
- ii. `qqnorm(m$duration)`
- iii. `qqline(m$duration)`



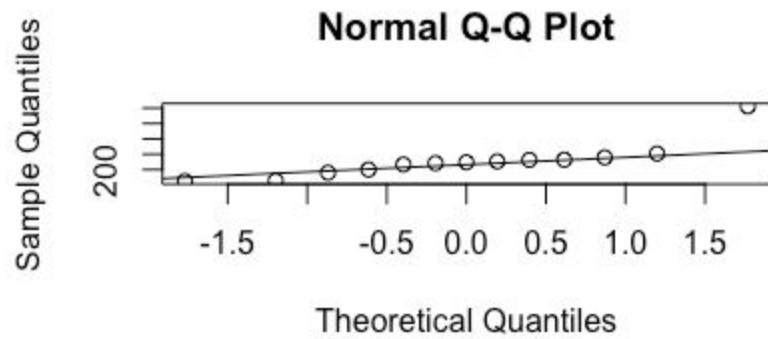
- iv.
- v. `hist(m$duration)`



- vi.

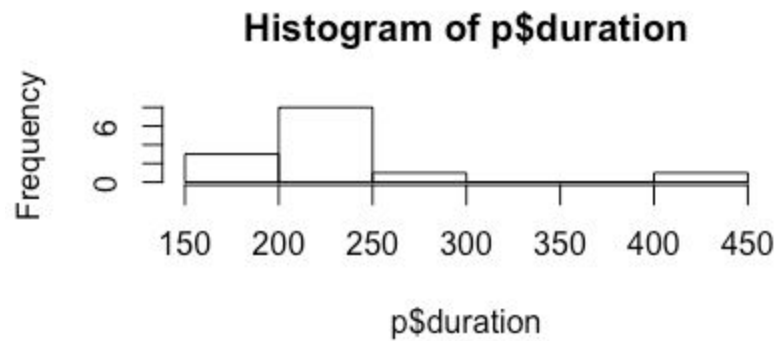
h. Pop

- i. `p = subset(Songs, genre == "pop")`
- ii. `qqnorm(p$duration)`
- iii. `qqline(p$duration)`



iv.

v. `hist(p$duration)`



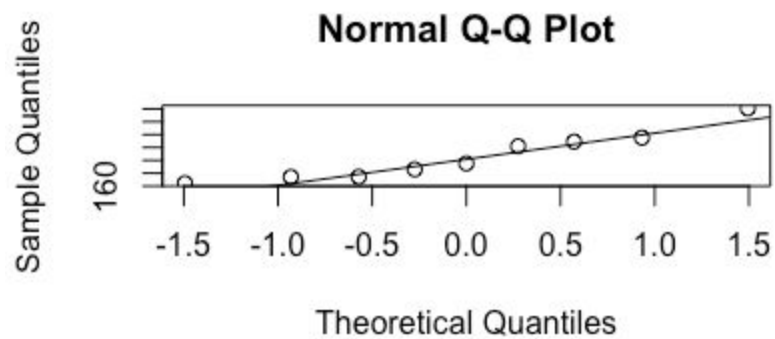
vi.

i. Punk

i. `unk = subset(Songs, genre == "punk")`

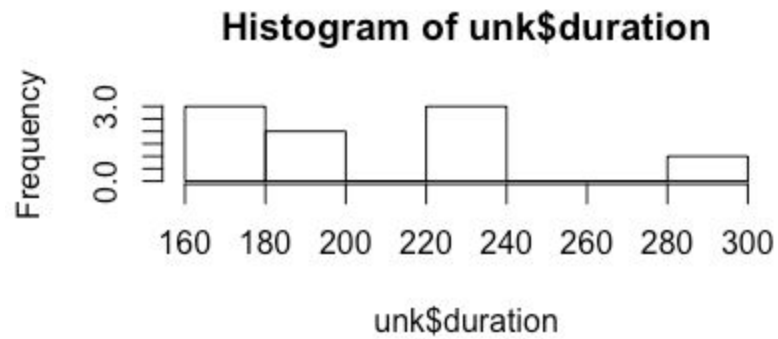
ii. `qqnorm(unk$duration)`

iii. `qqline(unk$duration)`



iv.

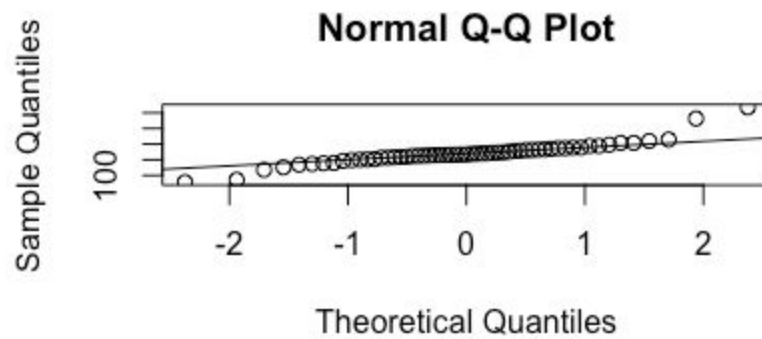
v. `hist(unk$duration)`



vi.

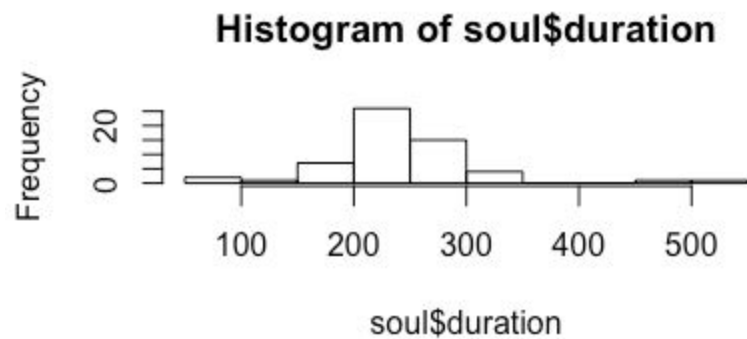
j. Soul and reggae

- i. `soul = subset(Songs, genre == "soul and reggae")`
- ii. `qqnorm(soul$duration)`



iii.

- iv. `qqline(soul$duration)`
- v. `hist(soul$duration)`



vi.

2. From all of the DV of genres we can see that not all of the groups are normally distributed such as punk, hip-hop, and folk music. However, we

can still continue to use ANOVA against some deviation from normality so we move on to the second assumption.

3. Assumption 2: Homogeneity of Variance

- a. Levene test
- b. `> leveneTest(Songs$duration~Songs$genre,Songs)`
- c. Levene's Test for Homogeneity of Variance (center = median)

```
      Df F-value Pr(>F)
group  9   5.578 2.246e-07 ***
      536
---
```

Since the $\text{Pr}(>F)$ is $2.246e-07$ and less than $.05$ we can then reject the null hypothesis and conclude that the assumption of homogeneity of variances is violated.

4. Assumption 3: Independence of observations

Based off the test we know that a song cannot be in two different genres, therefore we know that each song is independent of each other.