HW1 Peer Assessment

Part A. ANOVA

Additional Material: ANOVA tutorial

https://datascienceplus.com/one-way-anova-in-r/

Jet lag is a common problem for people traveling across multiple time zones, but people can gradually adjust to the new time zone since the exposure of the shifted light schedule to their eyes can resets the internal circadian rhythm in a process called "phase shift". Campbell and Murphy (1998) in a highly controversial study reported that the human circadian clock can also be reset by only exposing the back of the knee to light, with some hailing this as a major discovery and others challenging aspects of the experimental design. The table below is taken from a later experiment by Wright and Czeisler (2002) that re-examined the phenomenon. The new experiment measured circadian rhythm through the daily cycle of melatonin production in 22 subjects randomly assigned to one of three light treatments. Subjects were woken from sleep and for three hours were exposed to bright lights applied to the eyes only, to the knees only or to neither (control group). The effects of treatment to the circadian rhythm were measured two days later by the magnitude of phase shift (measured in hours) in each subject's daily cycle of melatonin production. A negative measurement indicates a delay in melatonin production, a predicted effect of light treatment, while a positive number indicates an advance.

Raw data of phase shift, in hours, for the circadian rhythm experiment

Treatment	Phase Shift (hr)
Control	0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -1.27
Knees	0.73, 0.31, 0.03, -0.29, -0.56, -0.96, -1.61
Eyes	-0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83

```
#Create varibales for each group
control <- list(0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -1.27)
knees <- list(0.73, 0.31, 0.03, -0.29, -0.56, -0.96, -1.61)
eyes <- list(-0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83)
labels <- list("control","knees","eyes")

control_df <- do.call(rbind, Map(data.frame,value=control,method=labels[1]))
knees_df <- do.call(rbind, Map(data.frame,value=knees,method=labels[2]))
eyes_df <- do.call(rbind, Map(data.frame,value=eyes,method=labels[3]))

final_df <- rbind(control_df,knees_df,eyes_df)

aov_summary <- do.call(rbind, Map(data.frame,summary(aov(value ~ method, data=final_df))))
model = aov(value ~ method, data=final_df)
model.tables(model,type="means")
```

```
## Tables of means
## Grand mean
```

```
##
## -0.7127273
##
##  method
##     control   eyes   knees
##     -0.3087 -1.551 -0.3357
## rep  8.0000  7.000  7.0000
```

Question A1 - 3 pts

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	2	7.224	3.6122	7.2895	0.004
Error	19	9.415	0.4955		
TOTAL	21	16.639			

Figure 1: Caption for the picture.

Question A2 - 3 pts

Use μ_1, μ_2, and μ_3 as notation for the three mean parameters and define these parameters clearly based on the context of the topic above. Find the estimates of these parameters.

The result is the following:

```
control   eyes   knees
-0.3087 -1.551 -0.3357
```

Question A3 - 5 pts

Use the ANOVA table in Question A1 to answer the following questions:

a. **1 pts** Write the null hypothesis of the ANOVA F-test, H_0 H0: There are equal means for each group representing the method used (i.e. control, knees, eyes)

b. **1 pts** Write the alternative hypothesis of the ANOVA F-test, H_A HA: There are not equal means for each group representing the method used (i.e. control, knees, eyes)

c. **1 pts** Fill in the blanks for the degrees of freedom of the ANOVA F-test statistic: $F(2,19)$

d. **1 pts** What is the p-value of the ANOVA F-test? 0.004

e. **1 pts** According the the results of the ANOVA F-test, does light treatment affect phase shift? Use an α-level of 0.05. Based on our p-value close to 0, we can reject the null hypothesis for an alpha level of 0.05, which means that there is variability across the means for our groups and that certain treatments have a higher impact than others.

Part B. Simple Linear Regression

We are going to use regression analysis to estimate the performance of CPUs based on the maximum number of channels in the CPU. This data set comes from the UCI Machine Learning Repository.

The data file includes the following columns:

- *vendor*: vendor of the CPU
- *chmax*: maximum channels in the CPU
- *performance*: published relative performance of the CPU

The data is in the file "machine.csv". To read the data in `R`, save the file in your working directory (make sure you have changed the directory if different from the R working directory) and read the data using the `R` function `read.csv()`.

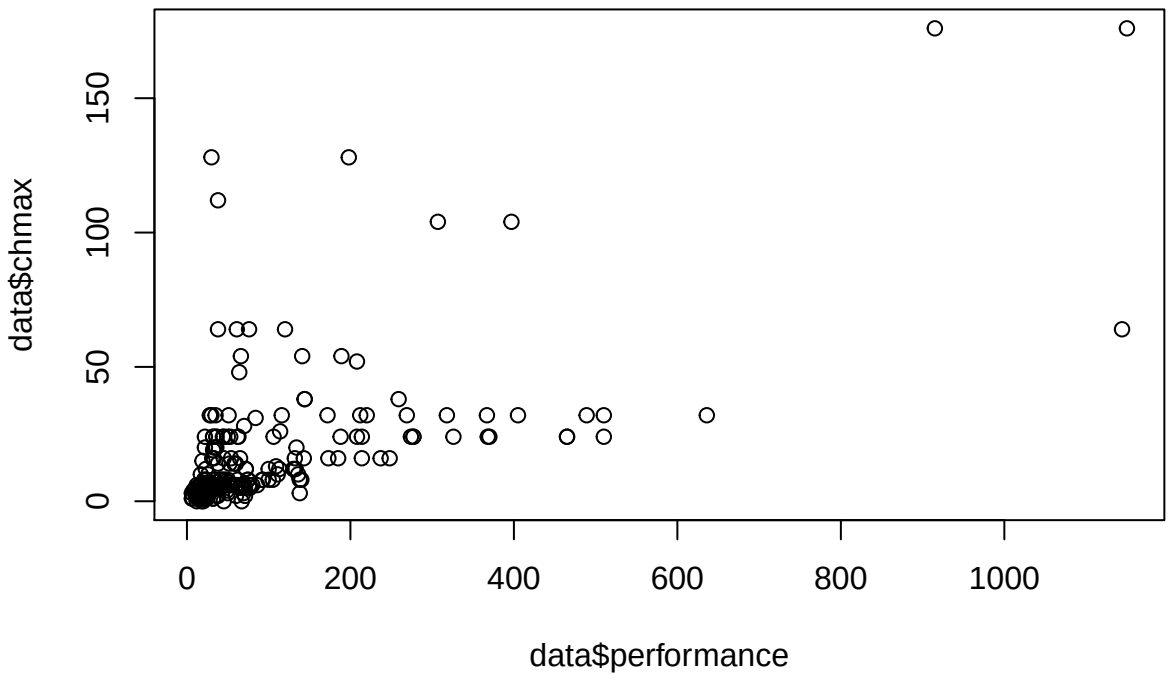
```
# Read in the data
data = read.csv("C:/Users/mjpearl/Desktop/omsa/ISYE-6414-OAN/hw1/machine.csv", head = TRUE, sep = ",")
# Show the first few rows of data
head(data, 3)
```

```
##    vendor chmax performance
## 1 adviser   128         198
## 2  amdahl    32         269
## 3  amdahl    32         220
```

Question B1: Exploratory Data Analysis - 9 pts

a. **3 pts** Use a scatter plot to describe the relationship between CPU performance and the maximum number of channels. Describe the general trend (direction and form). Include plots and R-code used.

```
plot(data$performance,data$chmax)
```



We can see that a majority of the data is clustered in the bottom left quadrant, but we can see that the linearity assumption can still be help as the data start to trend in a straight line from 200 - 600 on the performance side of the axis.

b. **3 pts** What is the value of the correlation coefficient between *performance* and *chmax*? Please interpret the strength of the correlation based on the correlation coefficient.

```
cor(data$performance,data$chmax,method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.6052093
```

Based on our results we can see that there's modest correlation between the two variables but it doesn't signify a strong correlation

c. **2 pts** Based on this exploratory analysis, would you recommend a simple linear regression model for the relationship?

I would still recommend a linear regression just because there's modest correlation between these two variables, and there's additional variables that we could include in order to improve the results. For the purposes of a simple linear regression a score of .6 without any intervention is still worth exploring. There could be additional transformation we could apply to the data in order to improve this score (i.e. log / power transformation, etc.)

```
summary(data)
```

```
##     vendor              chmax         performance
##  Length:209         Min.   :  0.00   Min.   :   6.0
##  Class :character   1st Qu.:  5.00   1st Qu.:  27.0
##  Mode  :character   Median :  8.00   Median :  50.0
##                     Mean   : 18.27   Mean   : 105.6
##                     3rd Qu.: 24.00   3rd Qu.: 113.0
##                     Max.   :176.00   Max.   :1150.0
```

d. **1 pts** Based on the analysis above, would you pursue a transformation of the data? *Do not transform the data.*

Based on our results it seems there are several outliers included in the data between these two variables. We can see this with the influential points of well above the 3rd quartile range for performance. I would conduct perhaps a StandardScaler or QuantileScaler in order to exclude certain observations past a specifc 3rd quantile value. We can also see that the data are on different scales altogether, since the Q1 value for performance is > than the Q3 value for chmax. Therefore, I would recommend removing outliers and scaling our data.

Question B2: Fitting the Simple Linear Regression Model - 11 pts

Fit a linear regression model, named *model1*, to evaluate the relationship between performance and the maximum number of channels. *Do not transform the data.* The function you should use in R is:

```
# Your code here...
model = lm(performance ~ chmax, data)
summary(model)
```

```
##
## Call:
## lm(formula = performance ~ chmax, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -486.47  -42.20  -22.20   20.31  867.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2252    10.8587   3.428 0.000733 ***
## chmax         3.7441     0.3423  10.938  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128.3 on 207 degrees of freedom
## Multiple R-squared:  0.3663, Adjusted R-squared:  0.3632
## F-statistic: 119.6 on 1 and 207 DF,  p-value: < 2.2e-16
```

a. **3 pts** What are the model parameters and what are their estimates?
(Intercept) 37.2252 chmax 3.7441

b. **2 pts** Write down the estimated simple linear regression equation.

performance = 37.2252 + 3.7441*chmax in units

c. **2 pts** Interpret the estimated value of the β_1 parameter in the context of the problem. We can see from the value of b1 that chmax has a positive impact on the performance variable, and 1 unit increase in chmax accounts for a 3.7441 increase in performance.

d. **2 pts** Find a 95% confidence interval for the β_1 parameter. Is β_1 statistically significant at this level? < 2e-16 is measures for Signif. codes: 0 '***' which means this applies for 0.95 significance level and that beta 1 is stastistically significant.

e.

```
confint(model,level=0.99)
```

```
##                0.5 %    99.5 %
## (Intercept) 8.994891 65.455549
## chmax       2.854185  4.633991
```

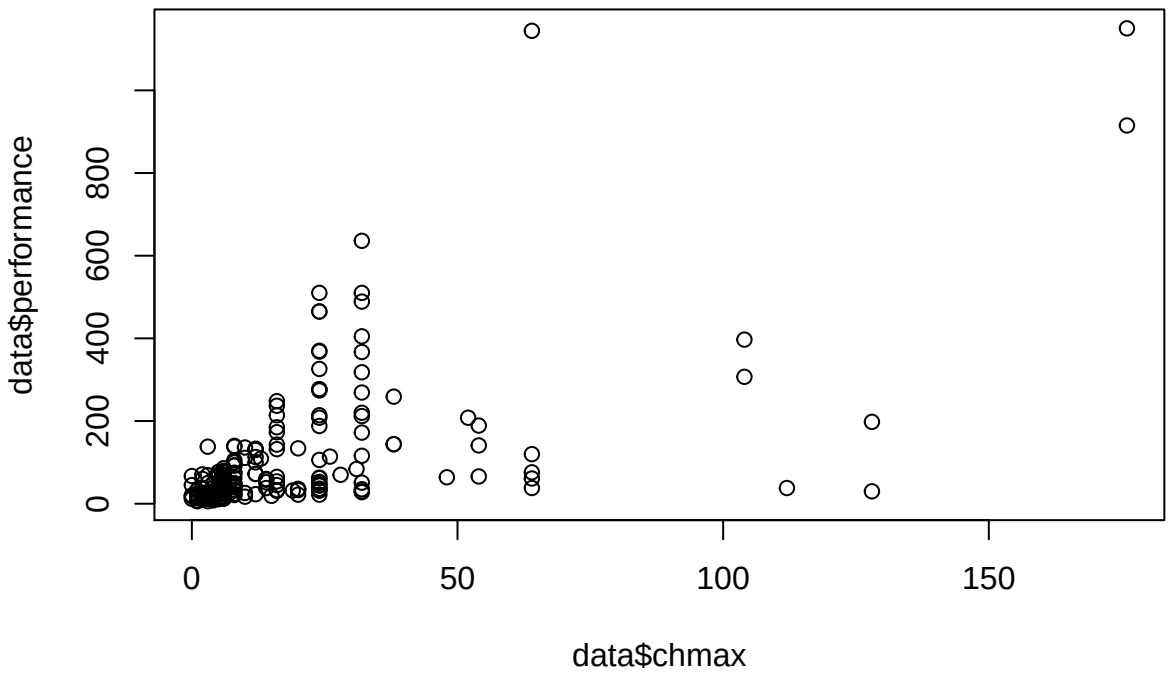
Yes we can conclude that it is statistically positive and has a confidence internal of (2.85, 4.63)

Question B3: Checking the Assumptions of the Model - 8 pts

Create and interpret the following graphs with respect to the assumptions of the linear regression model. In other words, comment on whether there are any apparent departures from the assumptions of the linear regression model. Make sure that you state the model assumptions and assess each one. Each graph may be used to assess one or more model assumptions.

a. **2 pts** Scatterplot of the data with *chmax* on the x-axis and *performance* on the y-axis

```
plot(data$chmax,data$performance)
```

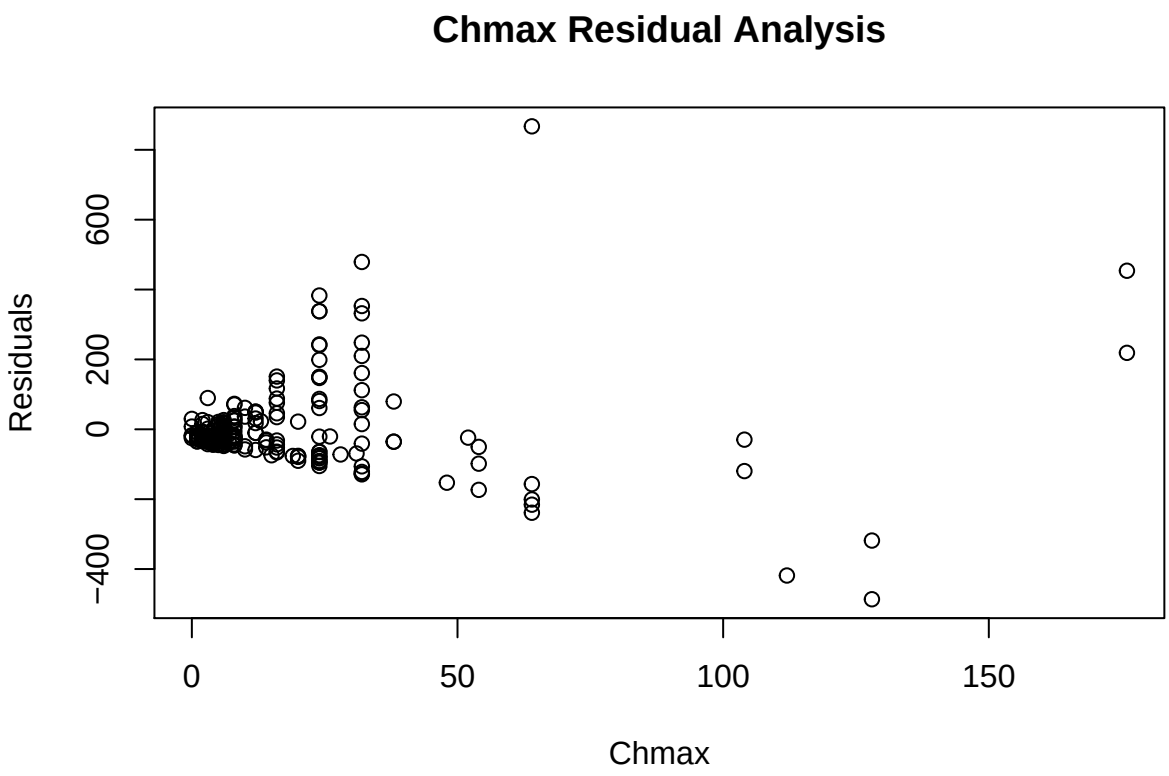


**Model Assumption(s) it checks: For the Linearity Assumption between X and Y

**Interpretation:*Based on our results we can conclude that our model does follow the linearity assumption as we don't have any significant curvatures or non-linear shapes in our scatterplot output. However, there does seem to be a significant amount of clustering around the 0 point, so it would be more ideal if the data was more spread out.*

b. **3 pts** Residual plot - a plot of the residuals, $\hat{\epsilon}_i$, versus the fitted values, $\hat{y}_i$

```
res <- resid(model)
plot(data$chmax, res, ylab="Residuals", xlab="Chmax", main="Chmax Residual Analysis")
```

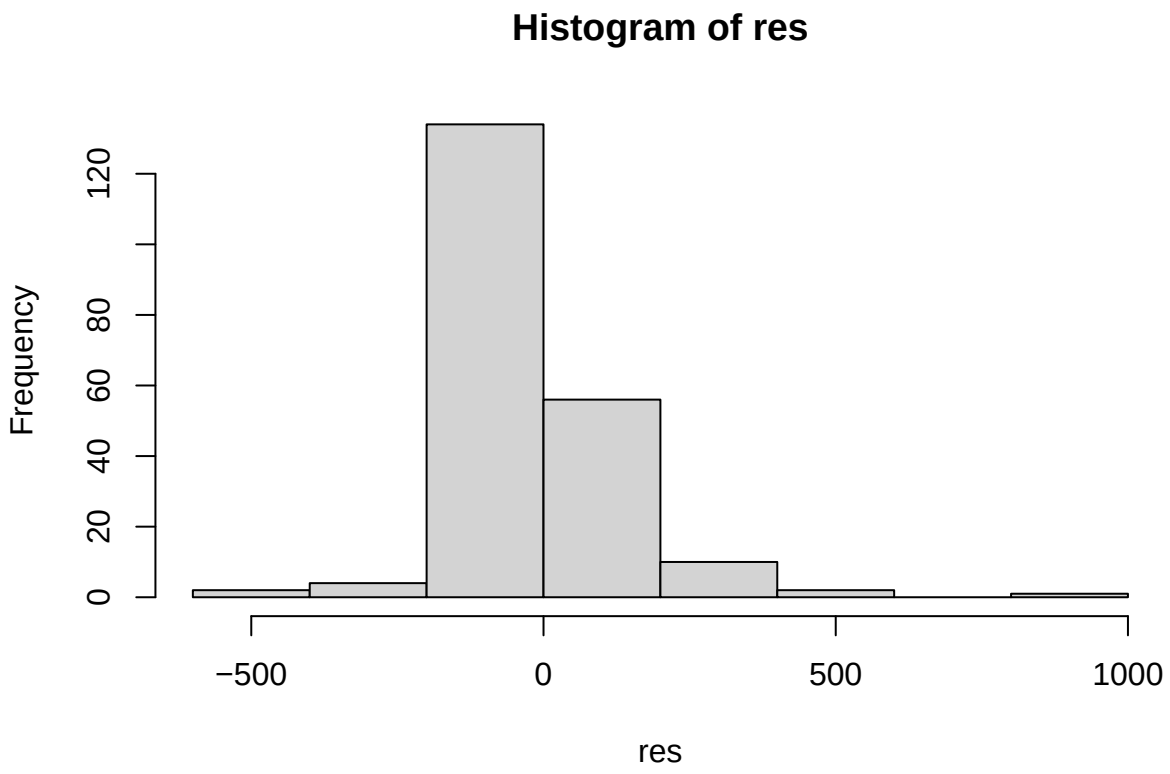


**Model Assumption(s) it checks:*Constant Variance Assumption*

**Interpretation:* From our results we can see that there's no significant cone shape as the values increase, and tends to follow a fairly equal variance throughout the observations when excluding our outliers out of the interpretation.*

c. **3 pts** Histogram and q-q plot of the residuals

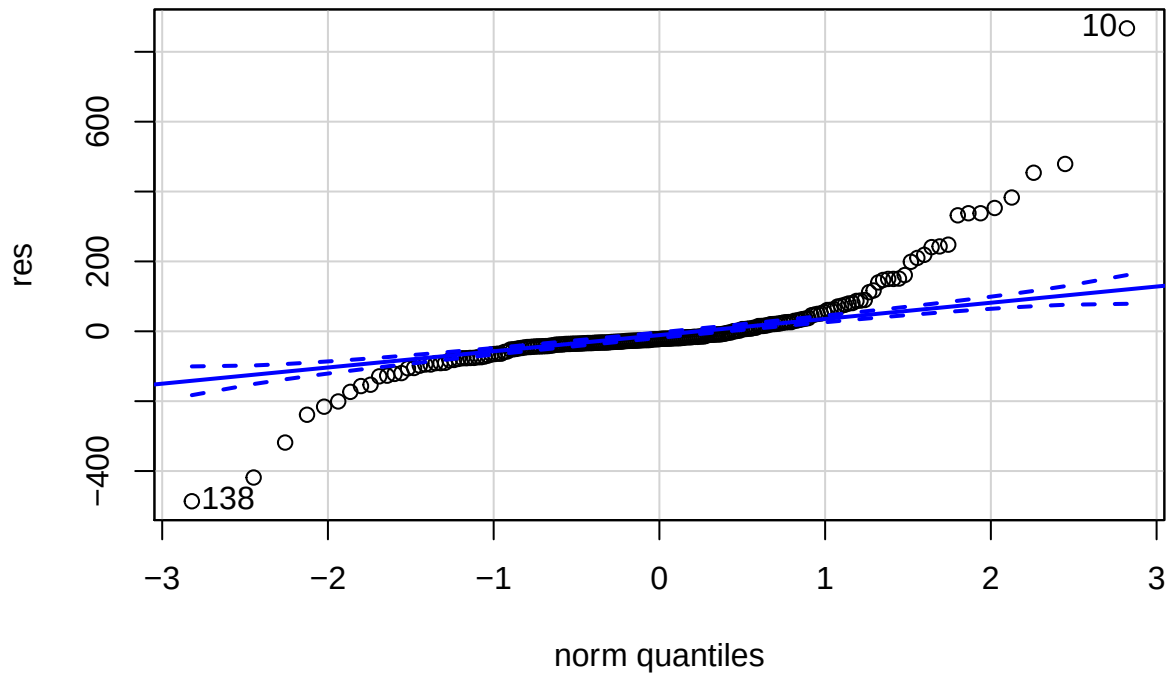
```
hist(res)
```




```
library(car)
```

```
## Loading required package: carData
```

```
qqPlot(res)
```




```
## [1]  10 138
```

**Model Assumption(s) it checks:*Normality Assumption*

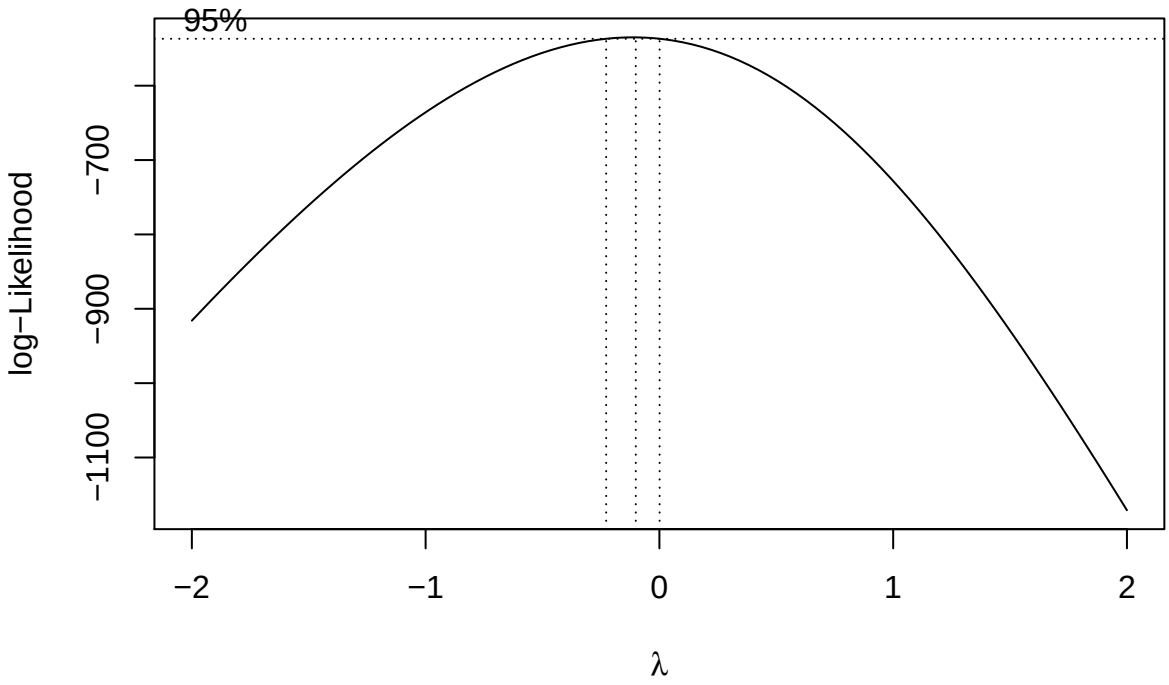
**Interpretation:*Based on our result from the histogram we can see that our data seems to have a skewed distribution. Also based on our qq plot we see a departure from a normal distribution as we have significant tails on either end.*

Question B4: Improving the Fit - 10 pts

a. **2 pts** Use a Box-Cox transformation (`boxCox()`) to find the optimal λ value rounded to the nearest half integer. What transformation of the response, if any, does it suggest to perform?

Based on the results of our BoxCox transformation, we can see that the highest point in the curve is centered around a lambda value of 0. This means that it is suggesting a log transformation to our variable.

```
library(MASS)
bc <- boxcox(performance ~ chmax,data=data)
```



b. **2 pts** Create a linear regression model, named *model2*, that uses the log transformed *performance* as the response, and the log transformed *chmax* as the predictor. Note: The variable *chmax* has a couple of zero values which will cause problems when taking the natural log. Please add one to the predictor before taking the natural log of it

```
model2 <- lm(log(performance) ~ log(chmax+1), data)
summary(model2)
```

```
##
## Call:
## lm(formula = log(performance) ~ log(chmax + 1), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22543 -0.59429  0.01065  0.59287  1.85995
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.47655    0.14152    17.5   <2e-16 ***
## log(chmax + 1)  0.64819    0.05401    12.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.807 on 207 degrees of freedom
## Multiple R-squared:  0.4103, Adjusted R-squared:  0.4074
```

```
## F-statistic:   144 on 1 and 207 DF,  p-value: < 2.2e-16
```

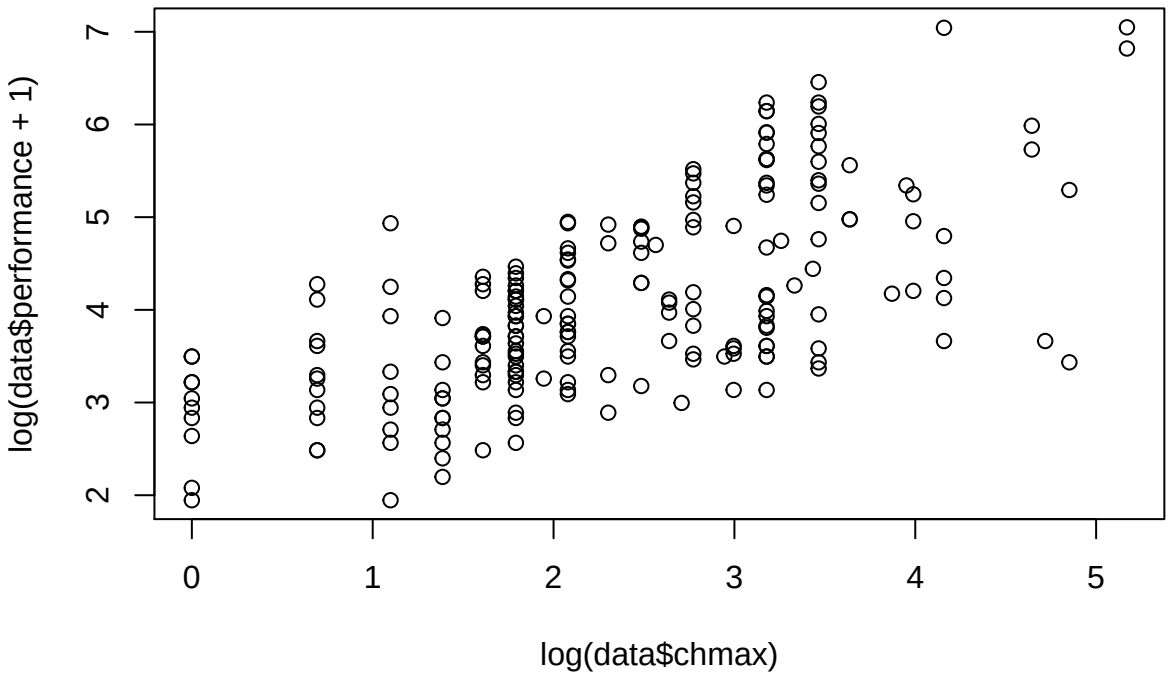
e. **2 pts** Compare the R-squared values of *model1* and *model2*. Did the transformation improve the explanatory power of the model?

Based on the results of the model, we can see that the log transformation did improve the results of the model as the R2 value increased from 0.3663 to 0.4103.

c. **4 pts** Similar to Question B3, assess and interpret all model assumptions of *model2*. A model is considered a good fit if all assumptions hold. Based on your interpretation of the model assumptions, is *model2* a good fit?

Log Transformed Model Assumptions

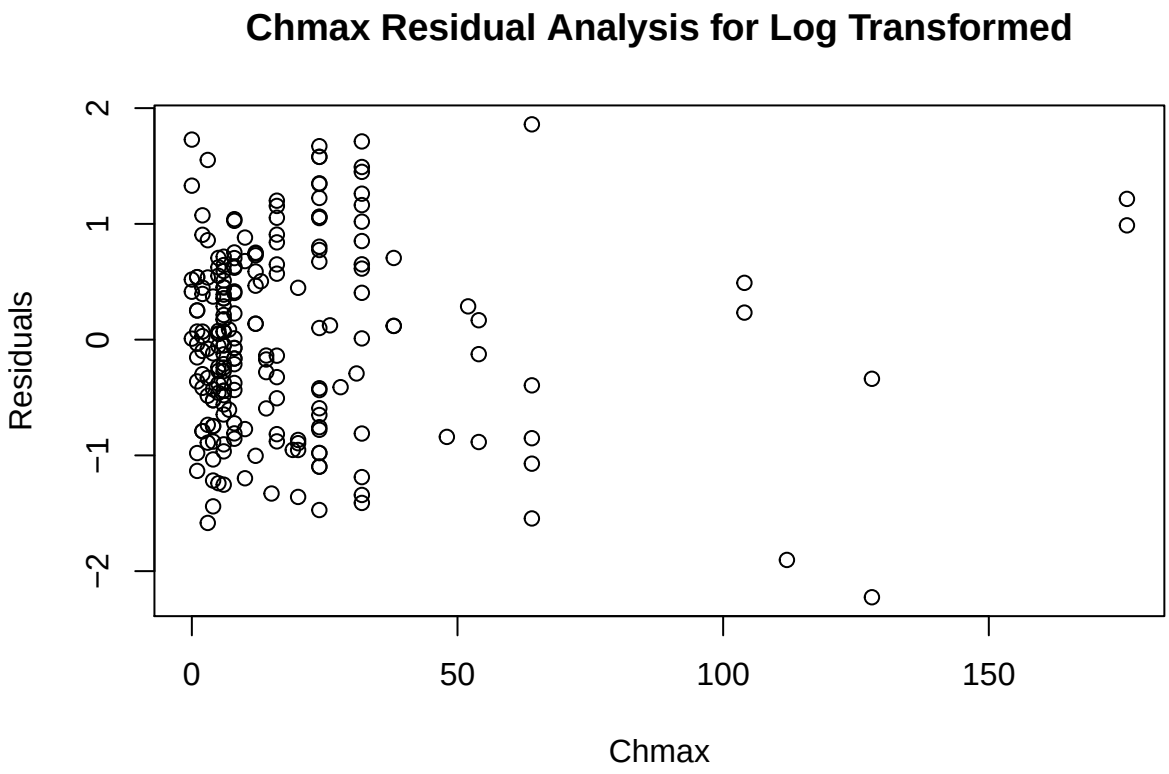
```
plot(log(data$chmax),log(data$performance+1))
```



**Model Assumption(s) it checks: For the Linearity Assumption between X and Y

**Interpretation:*Based on our results we can conclude that our model does follow the linearity assumption as we don't have any significant curvatures or non-linear shapes in our scatterplot output. We can see significant improvement for this assumption when compares to the last plot*

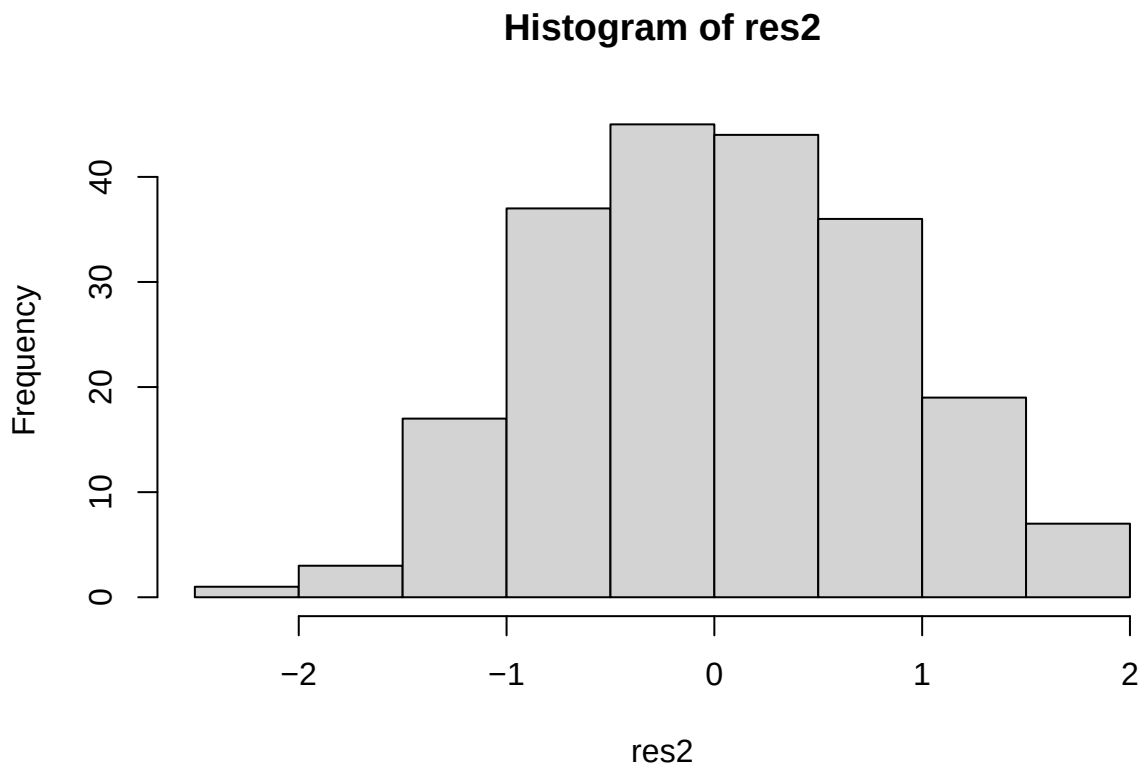
```
res2 <- resid(model2)
plot(data$chmax, res2, ylab="Residuals", xlab="Chmax", main="Chmax Residual Analysis for Log Transforme
```



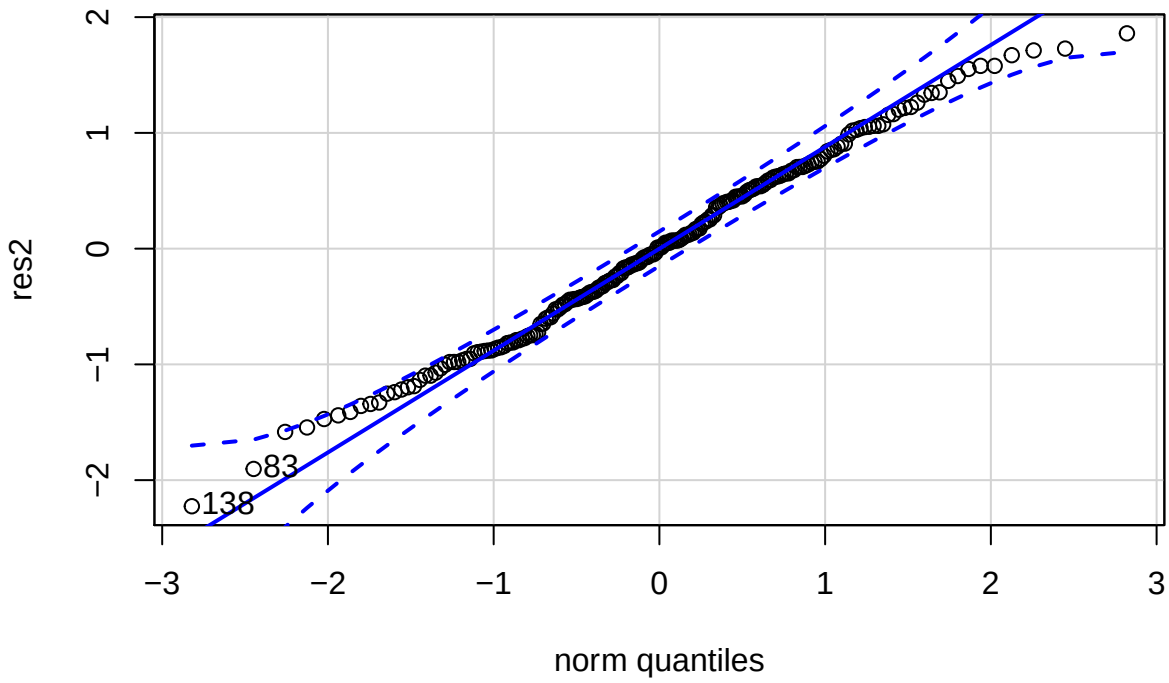
**Model Assumption(s) it checks:*Constant Variance Assumption*

**Interpretation:* From our results we can see that there's no significant cone shape as the values increase, and tends to follow a fairly equal variance throughout the observations when excluding our outliers out of the interpretation.*

```
hist(res2)
```




```
qqPlot(res2)
```




```
## [1] 138  83
```

**Model Assumption(s) it checks:*Normality Assumption*

**Interpretation:*Based on our result from the histogram we can see that the data is much better as it follows a very nice normal distribution now based on the histogram, and the qqplot is no longer exhibiting any tails on the graph*

Question B5: Prediction - 3 pts

Suppose we are interested in predicting CPU performance when `chmax = 128`. Please make a prediction using both *model1* and *model2* and provide the 95% prediction interval of each prediction on the original scale of the response, *performance*. What observations can you make about the result in the context of the problem?

```
new = data.frame(chmax=128)
predict.lm(model,new,interval='predict',level=0.95)
```

```
##        fit      lwr      upr
## 1 516.4685 252.2519 780.6851
```

```
predict.lm(model2,new,interval='predict',level=0.95)
```

```
##        fit      lwr      upr
## 1 5.626624 4.010584 7.242664
```

From our result we can see that the lower and upper limit for the second model is significantly less when compared to the first model. Meaning that there's a much higher likelihood for variability in our prediction when compared to the first model.

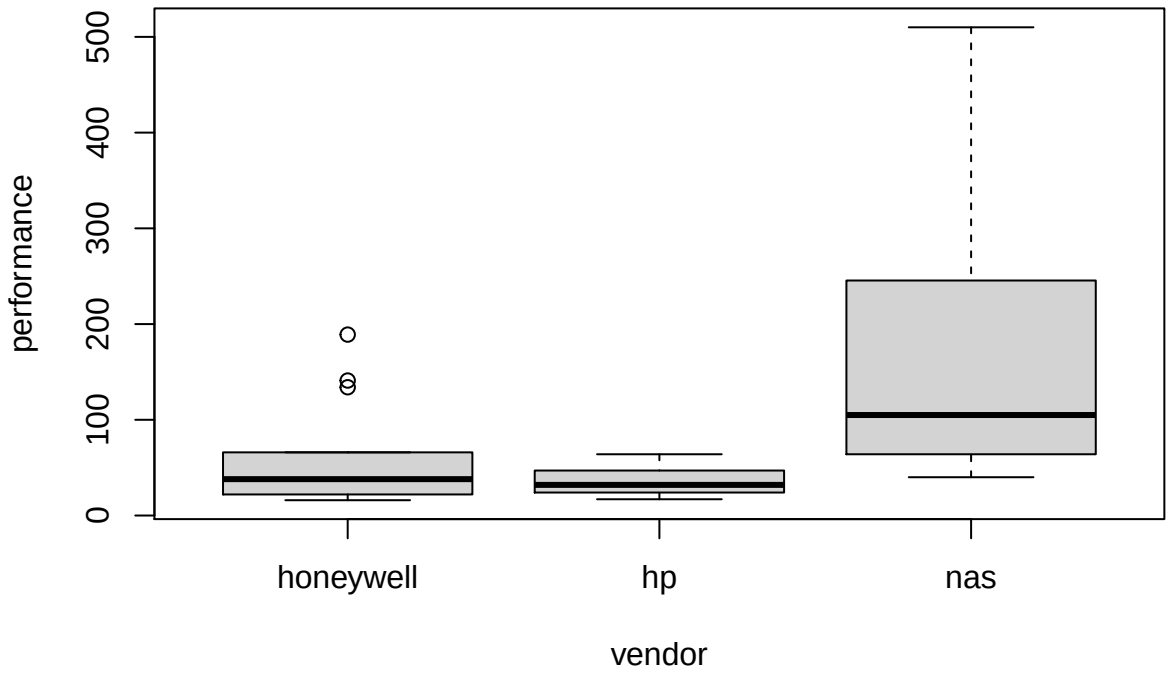
Part C. ANOVA - 8 pts

We are going to continue using the CPU data set to analyse various vendors in the data set. There are over 20 vendors in the data set. To simplify the task, we are going to limit our analysis to three vendors, specifically, honeywell, hp, and nas. The code to filter for those vendors is provided below.

```
# Filter for honeywell, hp, and nas
data2 = data[data$vendor %in% c("honeywell", "hp", "nas"), ]
data2$vendor = factor(data2$vendor)
```

1. **2 pts** Using `data2`, create a boxplot of *performance* and *vendor*, with *performance* on the vertical axis. Interpret the plots.

```
boxplot(performance~vendor,data=data2)
```



Based on our results we can see that nas significantly outperforms the honeywell and hp vendors. In addition, the honeywell provider seems to be experiencing outlier values, but they do not fall significantly out of the quartile range, and would not drastically impact any future modelling results.

Several vendors also have outlier values which may be impacting the results of their performance measurements, as these outliers are influential points falling significantly out of the quartile range.

2. **3 pts** Perform an ANOVA F-test on the means of the three vendors. Using an α-level of 0.05, can we reject the null hypothesis that the means of the three vendors are equal? Please interpret.

```
summary(aov(performance ~ vendor, data=data2))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## vendor       2 154494   77247   6.027 0.00553 **
## Residuals   36 461443   12818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the result

3. **3 pts** Perform a Tukey pairwise comparison between the three vendors. Using an α-level of 0.05, which means are statistically significantly different from each other?

```
TukeyHSD(aov(performance ~ vendor, data=data2))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = performance ~ vendor, data = data2)
##
## $vendor
##                     diff        lwr      upr     p adj
## hp-honeywell  -24.03297 -153.76761 105.7017 0.8934786
## nas-honeywell 116.43320   16.82659 216.0398 0.0188830
## nas-hp        140.46617   18.11095 262.8214 0.0214092
```

From the results of our model we can see that 0 does fall between the range of the lower and upper bounds of -153 and 105, respectively. This indicates that the means between the groups hp and honeywell could be equal, whereas the other pairs are not!