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

Question A1 - 3 pts

Consider the following incomplete R output:

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	?	?	3.6122	?	0.004
Error	?	9.415	?		
TOTAL	?	?			

Fill in the missing values in the analysis of the variance table.

```
y <- c(0.53, 0.36, 0.2, -0.37, -0.6, -0.64, -0.68, -1.27, 0.73, 0.31, 0.03, -0.29,
    -0.56, -0.96, -1.61, -0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83)
group <- factor(c("Control", "Control", "Control", "Control", "Control", "Control",
    "Control", "Control", "Knees", "Knees", "Knees", "Knees", "Knees", "Knees", "Knees",
    "Eyes", "Eyes", "Eyes", "Eyes", "Eyes", "Eyes", "Eyes"))
```

```r
df <- data.frame(group, y)
```

```r
model <- aov(y ~ group, data = df)
summary(model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## group        2  7.224   3.612   7.289 0.00447 **
## Residuals   19  9.415   0.496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Completed Table:

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	2	7.224	3.6122	7.289	0.004
Error	19	9.415	0.496		
TOTAL	21	16.639			

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.

```r
model.tables(model, type = "means")
```

```
## Tables of means
## Grand mean
##
## -0.7127273
##
##  group
##     Control   Eyes   Knees
##     -0.3087 -1.551 -0.3357
## rep  8.0000  7.000  7.0000
```

μ_1 = -0.3087 = mean phase shift in hours of Control group (did not have any body part exposed to light)
μ_2 = -1.551 = mean phase shift in hours of group that had eyes exposed to light
μ_3 = -0.3357 = mean phase shift in hours of group that had knees exposed to light

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
H_0: $\mu_1 = \mu_2 = \mu_3$

b. **1 pts** Write the alternative hypothesis of the ANOVA F-test, H_A
H_A: All or some of μ_1, μ_2, μ_3 are different

c. **1 pts** Fill in the blanks for the degrees of freedom of the ANOVA F-test statistic: F(_____, _____)
F(k-1,N-k) = F(3-1,22-3) = F(2,19)

d. **1 pts** What is the p-value of the ANOVA F-test?
p-value = 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.
Since the p-value is small and less than 0.05, we reject the null hypothesis of equal means. This means that in this experiment, the mean of the control, eyes and knees groups are different from each other.

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()`.

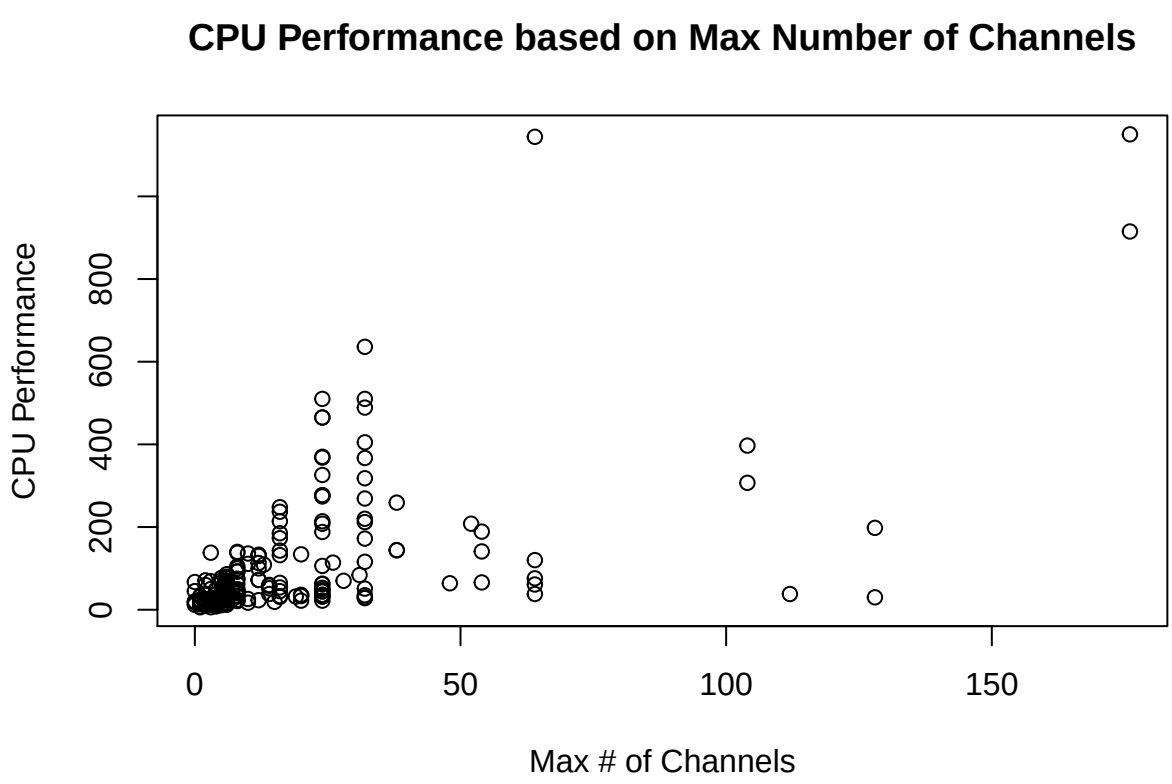
```
# Read in the data
data = read.csv("machine.csv", head = TRUE, sep = ",")
# Show the first few rows of data
head(data, 3)
```

```
##    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(x = data$chmax, y = data$performance, main = "CPU Performance based on Max Number of Channels",
    xlab = "Max # of Channels", ylab = "CPU Performance")
```



The general trend of this relationship is linear and positively increasing. The more channels that are available, the higher the CPU performance.

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)
```

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

The correlation coefficient is 0.61, a moderate positive correlation.

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

Using the plot observed in part a, I would not recommend a simple linear regression model. There is a dense cluster of data points to the bottom left of the graph that needs further exploration.

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

```
model1 = lm(performance ~ chmax, data)
summary(model1)
```

```
##
## 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?
Model parameters are:
$\beta_0 = 37.2252$ = the intercept, the CPU performance when there are no channels in the CPU
$\beta_1 = 3.7441$ = each additional channel increases CPU performance by 3.7 performance points

b. **2 pts** Write down the estimated simple linear regression equation.
performance = 37.2252 + 3.7441*chmax

c. **2 pts** Interpret the estimated value of the β_1 parameter in the context of the problem.
β_1 is positive so there exists a direct relationship with between the number of maximum channels and the CPU performance (ie - the more channels are, the higher we can expect the CPU performance to be). As stated above, each additional channel increases CPU performance by ~3.7 performance points.

d. **2 pts** Find a 95% confidence interval for the β_1 parameter. Is β_1 statistically significant at this level?

```
confint(model1, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 15.817392 58.633048
## chmax        3.069251  4.418926
```

The 95% confidence interval for β_1 is (3.069251, 4.418926). Based on the summary results above, the p-value for chmax is less than 0.05 so we can conclude that β_1 is statistically significant at this level.

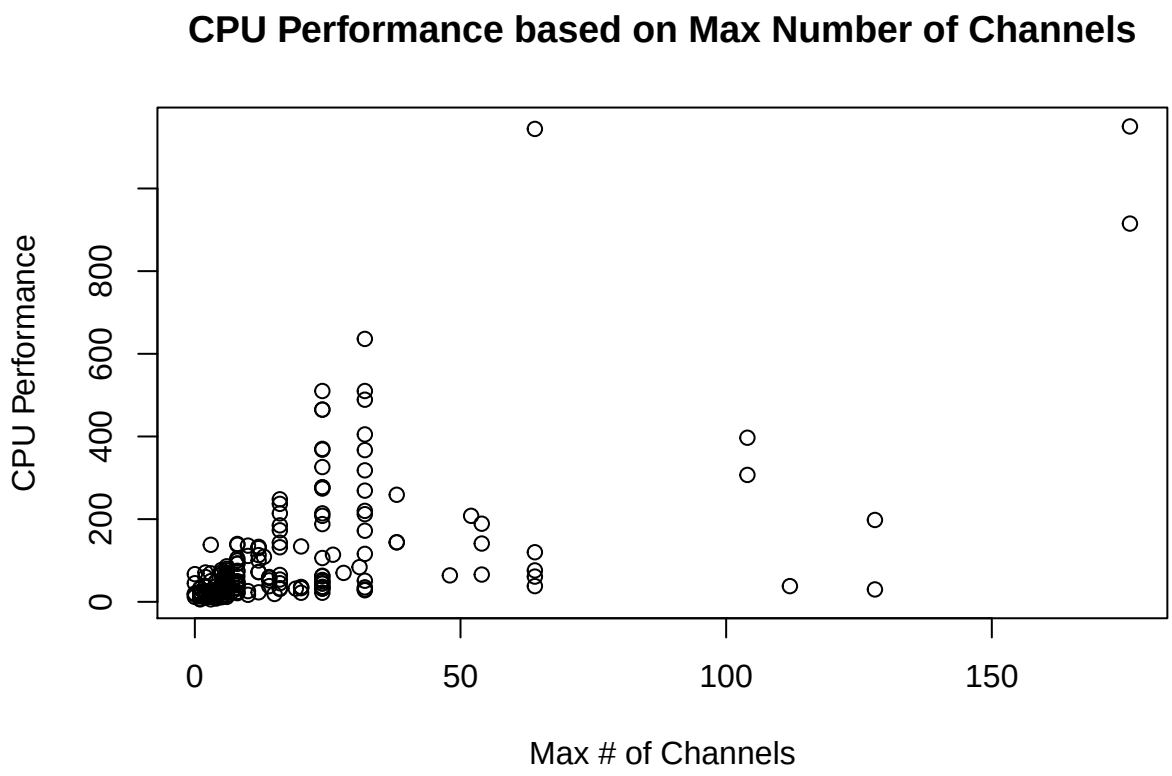
e. **2 pts** Is β_1 statistically significantly positive at an α-level of 0.01? What is the approximate p-value of this test?
Similar to above, the p-value for chmax is very small and less than 0.01 so we can conclude that β_1 is statistically significant.

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(x = data$chmax, y = data$performance, main = "CPU Performance based on Max Number of Channels",
    xlab = "Max # of Channels", ylab = "CPU Performance")
```

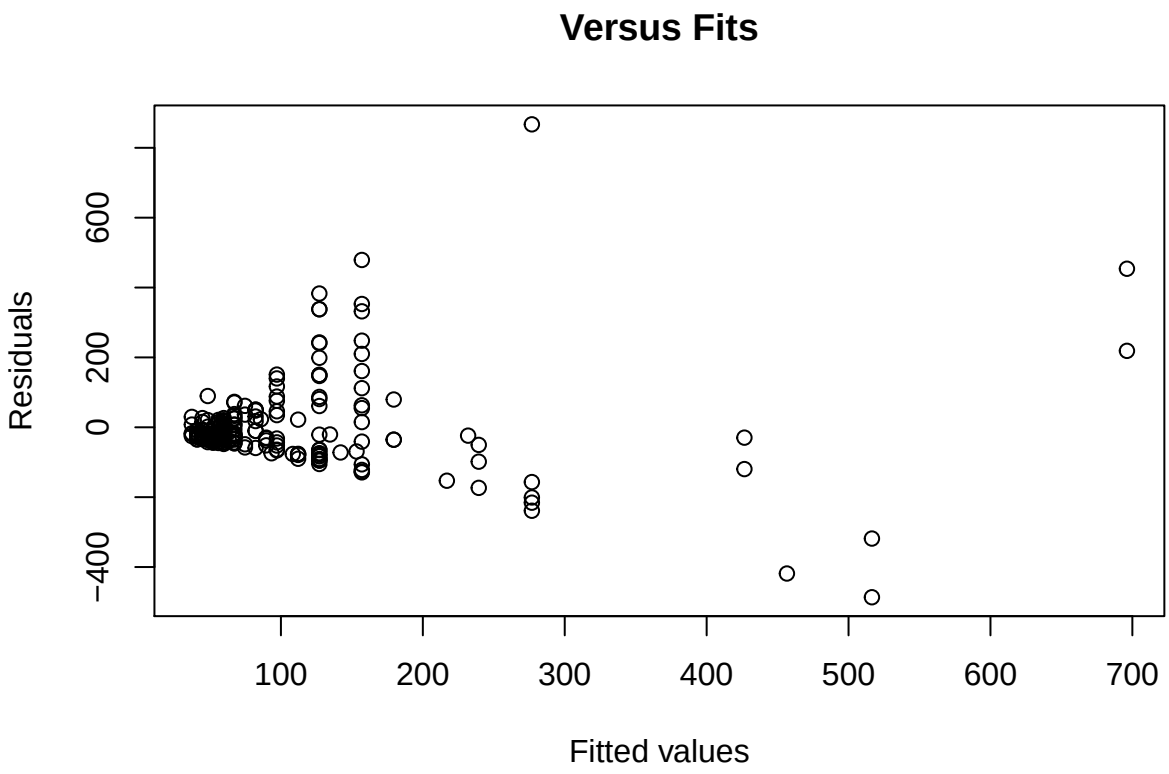


Model Assumption(s) it checks: Linearity (if there is a linear relationship between response and predictive variable)

Interpretation: There is a moderately positive linear relationship between the response and predictive variable. The relationship is not strong, so we need to check the other assumptions to determine if we need to transform the data.

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

```
plot(fitted(model1), residuals(model1), xlab = "Fitted values", ylab = "Residuals",
    main = "Versus Fits")
```

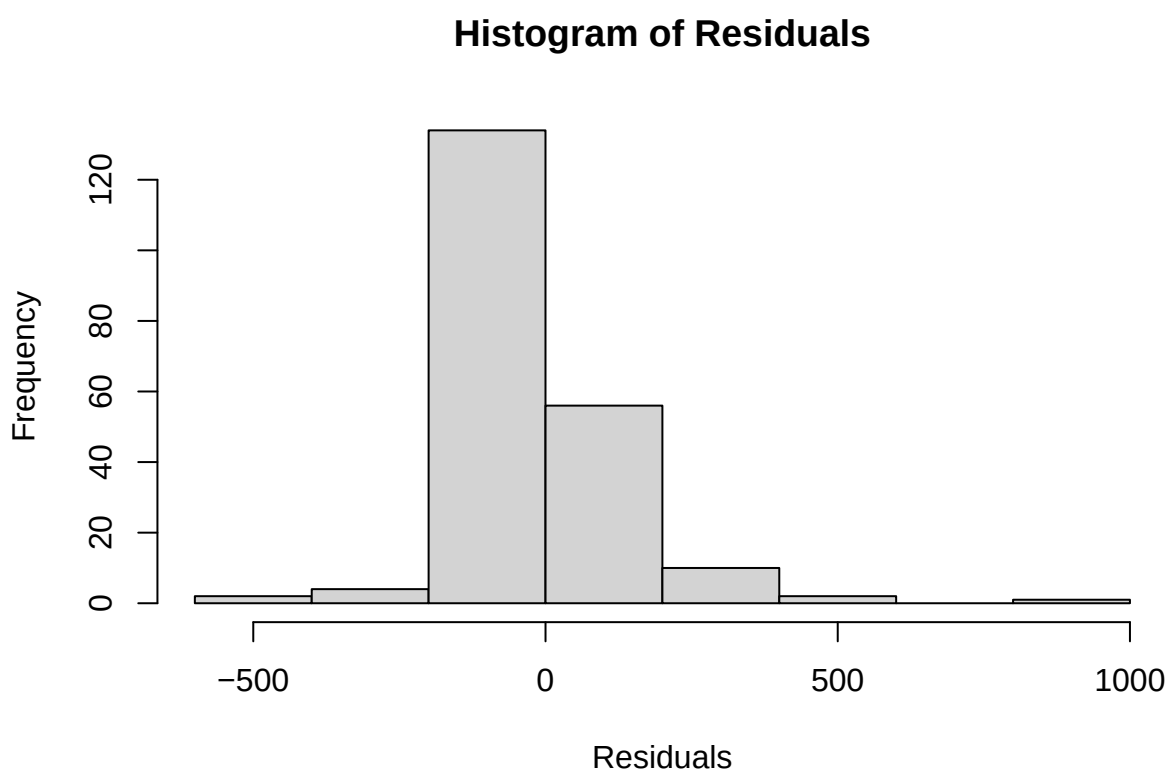


Model Assumption(s) it checks: Constant Variance

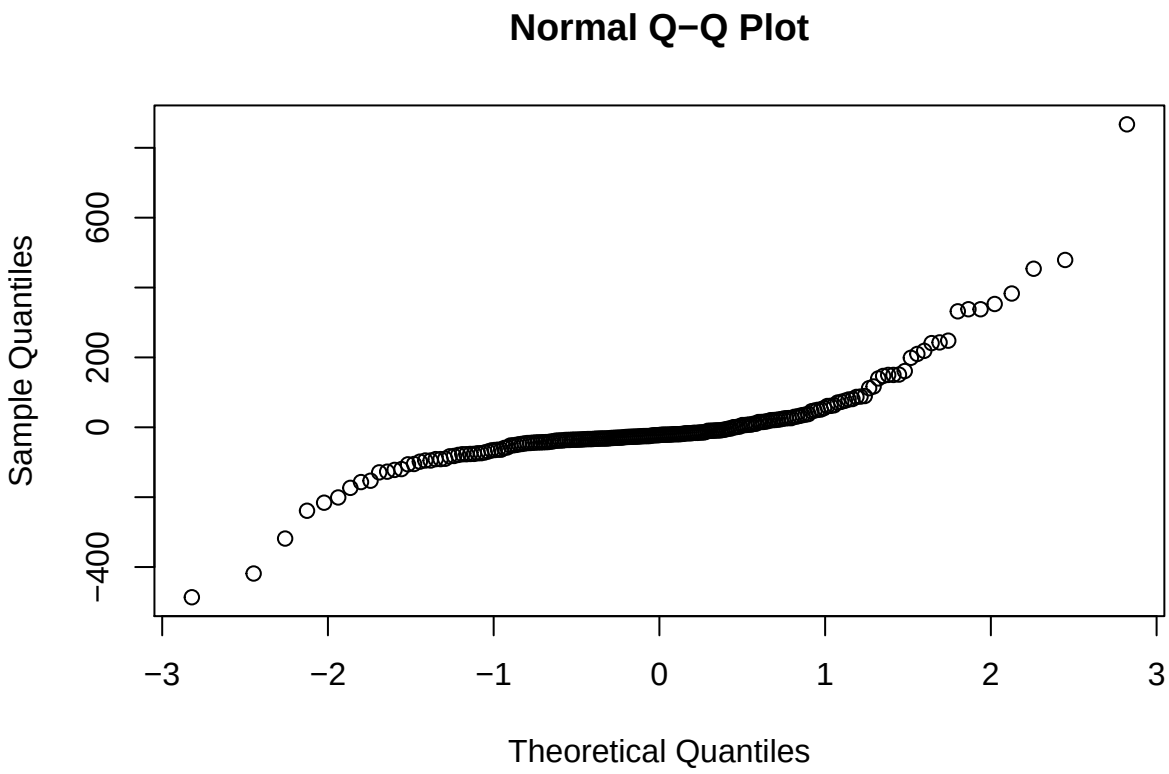
Interpretation: Residuals show a larger variance as the fitted values increase.

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

```
hist(residuals(model1), main = "Histogram of Residuals", xlab = "Residuals")
```




```
qqnorm(residuals(model1))
```



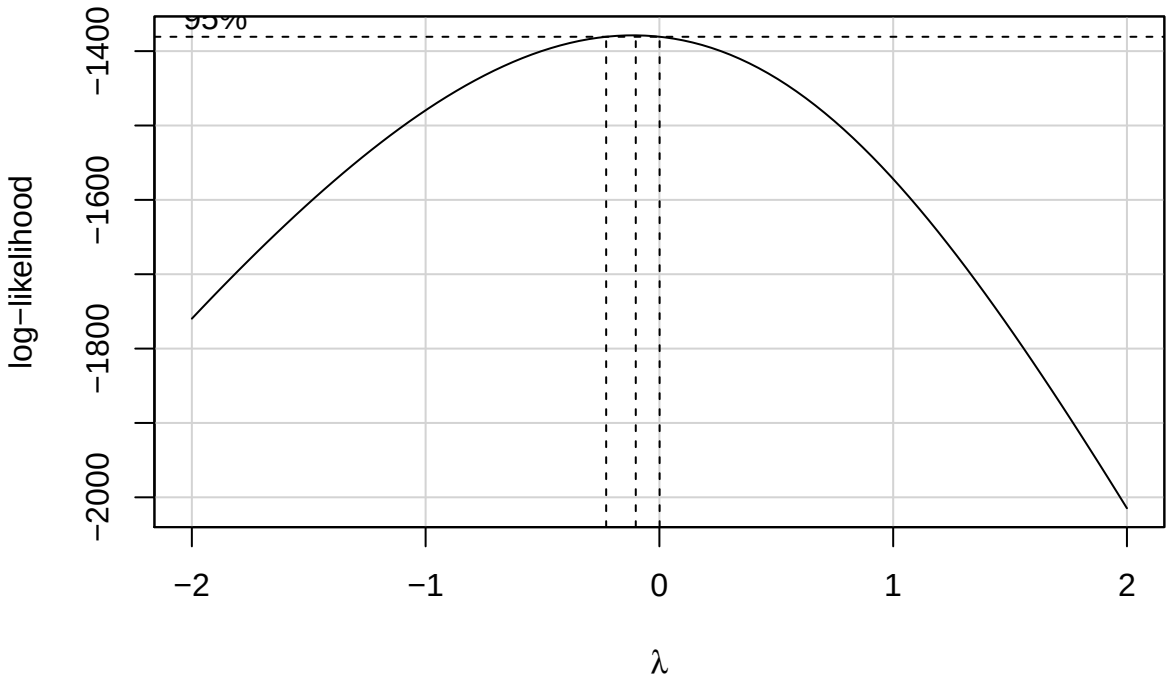
Model Assumption(s) it checks: Normality

Interpretation: Histogram shows residuals are slightly skewed to the left. Curvature at the ends of the QQ plot suggest that non-normality exists in the data.

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?

```
library("car")
model1_bC <- boxCox(model1)
```




```
lambda <- model1_bC$x[which.max(model1_bC$y)]
lambda_round <- round(lambda/0.5) * 0.5
lambda_round
```

```
## [1] 0
```

Since $\lambda = 0$, it is suggested that a log transformation is performed.

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(log1p(performance) ~ chmax, data)
summary(model2)
```

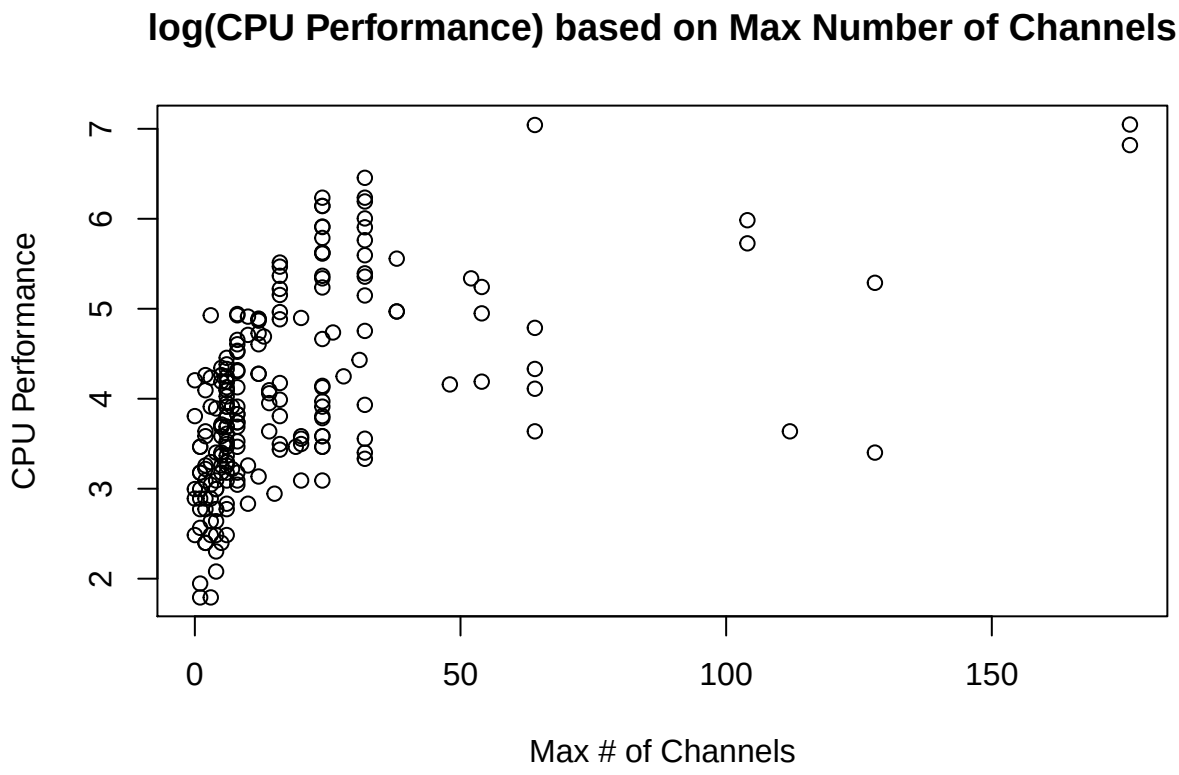
```
##
## Call:
## lm(formula = log1p(performance) ~ chmax, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.83056 -0.60271 -0.08479  0.52552  2.11705
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.698106   0.075053  49.273  < 2e-16 ***
## chmax       0.020050   0.002366   8.475  4.4e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8871 on 207 degrees of freedom
## Multiple R-squared:  0.2576, Adjusted R-squared:  0.254
## F-statistic: 71.82 on 1 and 207 DF,  p-value: 4.404e-15
```

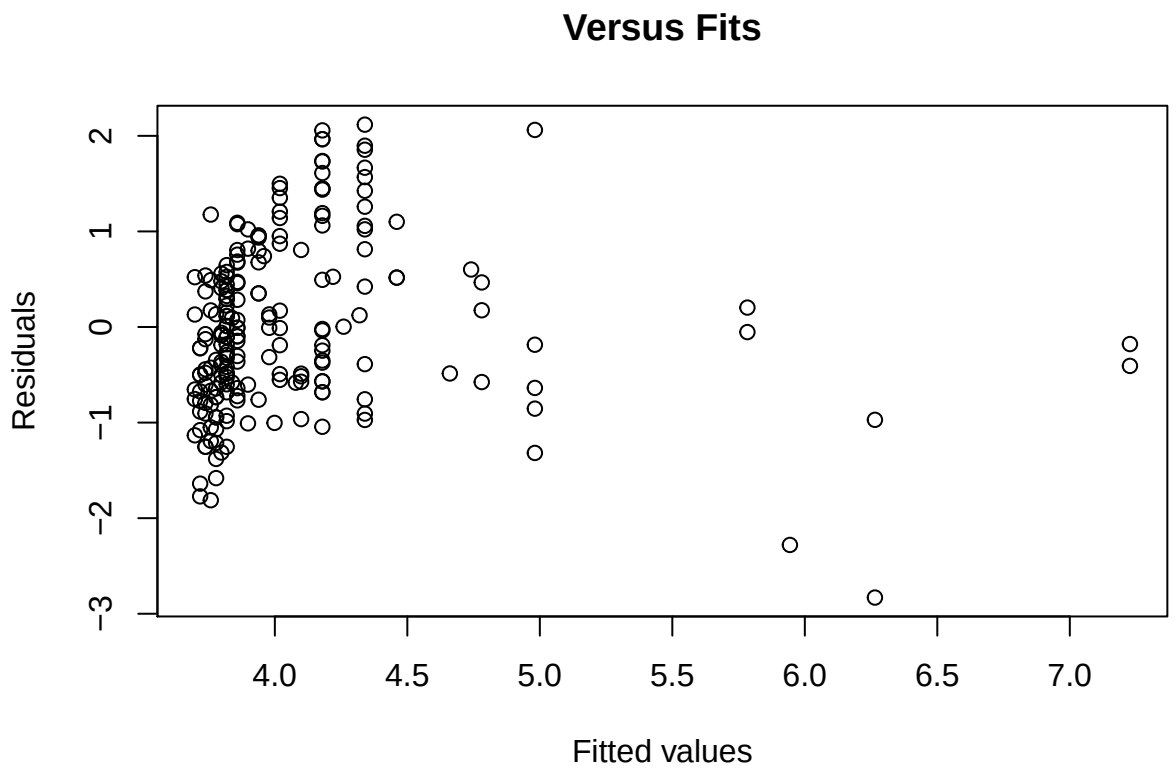
e. **2 pts** Compare the R-squared values of *model1* and *model2*. Did the transformation improve the explanatory power of the model?
Model 1 R-Squared = 0.3663
Model 2 R-Squared = 0.8871
The transformation improved the explanatory power of the model.

f. **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?

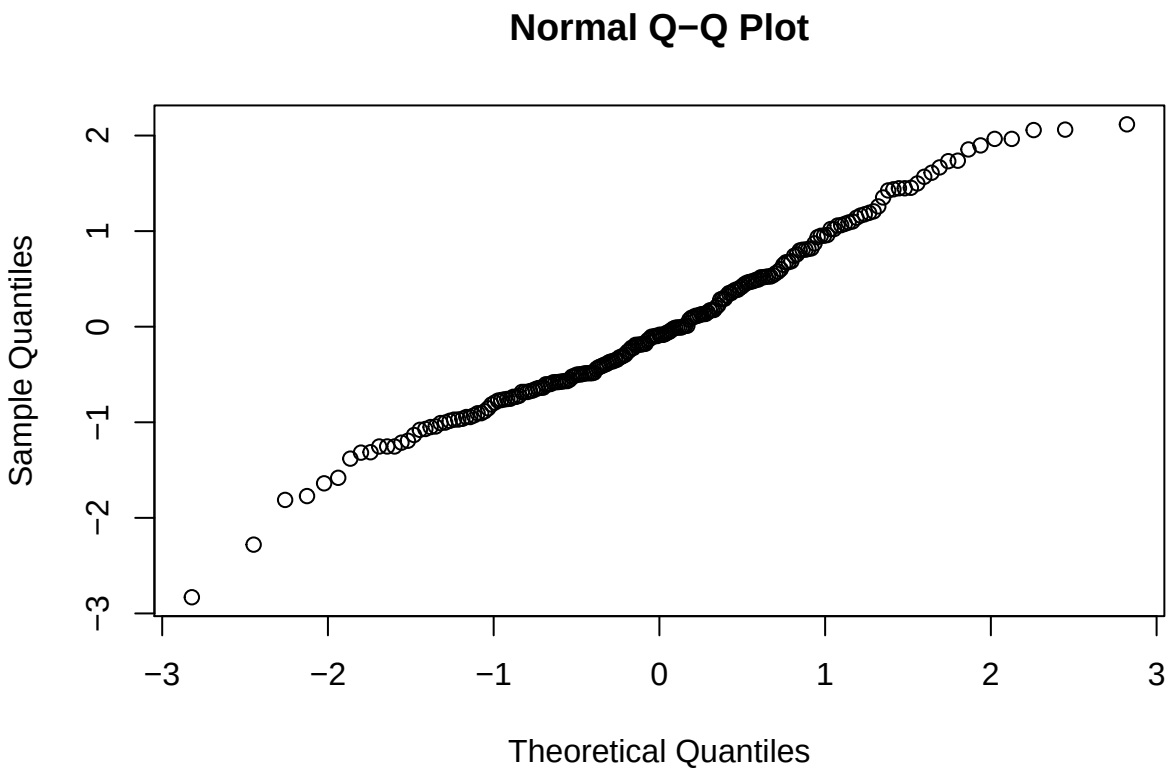
```
# Linearity
plot(x = data$chmax, y = log(data$performance), main = "log(CPU Performance) based on Max Number of Cha
    xlab = "Max # of Channels", ylab = "CPU Performance")
```



```r
# Constant Variance
plot(fitted(model2), residuals(model2), xlab = "Fitted values", ylab = "Residuals",
    main = "Versus Fits")
```




```r
# Normality
qqnorm(residuals(model2))
```



While the plots that check for linearity and constant variance still show some clustering to the left side of the graphs, the results are much improved from the first model. The normality assumption also holds much better in this model than the first model. I wouldn't say model 2 is a good fit, but it is definitely a better fit than model 1.

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 <- data.frame(chmax = 128)

model1_results <- predict.lm(model1, new_data, interval = "predict", level = 0.95)
model2_results <- predict.lm(model2, new_data, interval = "predict", level = 0.95)

# Changing the scale of model2 results
fit = exp(model2_results[1, 1])
lwr = exp(model2_results[1, 2])
upr = exp(model2_results[1, 3])

model1_results
```

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

```
fit
```

```
## [1] 525.6027
```

```
lwr
```

```
## [1] 84.63453
```

```
upr
```

```
## [1] 3264.131
```

Based on these results, model 2 gives a much wider prediction interval when transformed back to the original scale. However, the fitted values for each model's predictions are not far off from each other (516 for model 1 and 526 for model 2).

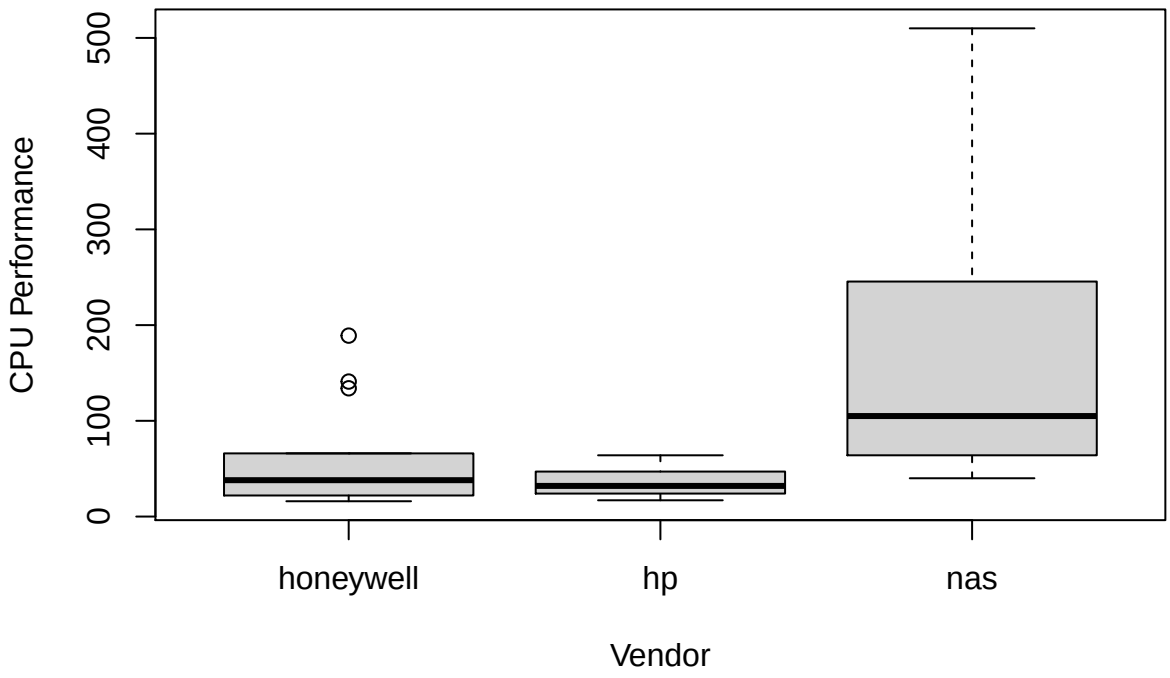
Part C. ANOVA - 8 pts

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, xlab = "Vendor", ylab = "CPU Performance")
```



Based on these plots, nas has higher within-variability compared to honeywell and hp. There is high between-variability between nas and the other 2 vendors.

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.

```
model_vendors <- aov(data2$performance ~ data2$vendor)
summary(model_vendors)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## data2$vendor  2 154494   77247   6.027 0.00553 **
## Residuals    36 461443   12818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.tables(model_vendors, type = "means")
```

```
## Tables of means
## Grand mean
##
## 112.8718
##
##  data2$vendor
##     honeywell    hp   nas
##         60.46 36.43 176.9
## rep     13.00  7.00  19.0
```

Since the p-value is small and less than 0.05, we can reject the null hypothesis and conclude that the means of the groups are not equal to each other.

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(model_vendors)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data2$performance ~ data2$vendor)
##
## $`data2$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
```

Based on an α-level of 0.05, nas-honeywell and nas-hp are statistically significant and we can conclude that the mean for nas is statistically significantly different from hp and honeywell.