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.

Answer A1

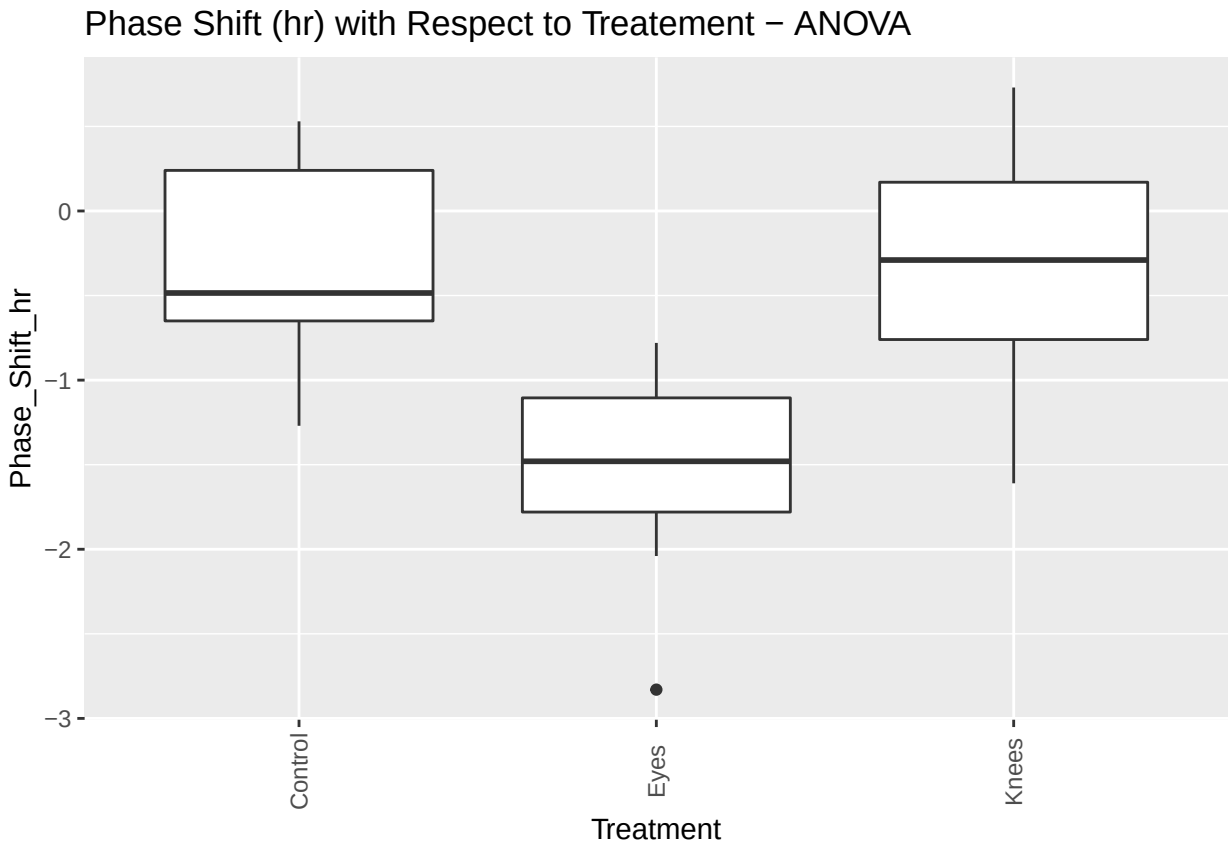
I created a csv file with the data and saved as Treatment_Phase_Shift.csv.

```
treatement_shift_data = read.csv("Treatment_Phase_Shift.csv", header = TRUE)
treatement_shift_data
```

```
##    Treatment Phase_Shift_hr
```

```
## 1    Control           0.53
## 2    Control           0.36
## 3    Control           0.20
## 4    Control          -0.37
## 5    Control          -0.60
## 6    Control          -0.64
## 7    Control          -0.68
## 8    Control          -1.27
## 9      Knees           0.73
## 10     Knees           0.31
## 11     Knees           0.03
## 12     Knees          -0.29
## 13     Knees          -0.56
## 14     Knees          -0.96
## 15     Knees          -1.61
## 16      Eyes          -0.78
## 17      Eyes          -0.86
## 18      Eyes          -1.35
## 19      Eyes          -1.48
## 20      Eyes          -1.52
## 21      Eyes          -2.04
## 22      Eyes          -2.83
```

```r
library(ggplot2)
ggplot(treatement_shift_data, aes(x=Treatment, y=Phase_Shift_hr),
       xlab = "Treatement",
       ylab = "Phase Shift (hr)") + geom_boxplot() +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))+
  ggtitle("Phase Shift (hr) with Respect to Treatement - ANOVA")
```




```
# apply ANOVA
model = aov(Phase_Shift_hr ~ Treatment, data=treatement_shift_data)

summary(model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Treatment    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
```

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

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

Pulling in the information from model summary...

Complete R output:

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.

Answer A2

Calculate Mean using the *aggregate* function

```
aggregate(treatement_shift_data[, 2], list(treatement_shift_data$Treatment), mean)
```

```
##   Group.1          x
## 1 Control -0.3087500
## 2    Eyes -1.5514286
## 3   Knees -0.3357143
```

Variable	Definition	Value
μ_1	Mean of Control Group	-0.3087500
μ_2	Mean of Eyes Group	-1.5514286
μ_3	Mean of Knees Group	-0.3357143

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

Answer : Null hypothesis means all means are equal - $\mu_1 = \mu_2 = \ldots \ \mu_k$. In this case $\mu_{Control} = \mu_{Eyes} = \mu_{Knees}$

b. **1 pts** Write the alternative hypothesis of the ANOVA F-test, H_A

Answer : Alternate Hypothesis means some means are different. $\mu_{Control}$ and μ_{Knee} to be similar but μ_{Eyes} is different. This shows atleast μ_{Eyes} is different from $\mu_{Control}$ & μ_{Knees}

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

```
nrow(treatement_shift_data)
```

```
## [1] 22
```

```
length(unique(treatement_shift_data$Treatment))
```

```
## [1] 3
```

Answer : F(k - 1, N - k) -> in above example it is F**(2, 19)**

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

```
#treatement_shift_data$Treatment
#treatement_shift_data$Phase_Shift_hr
```

```
model = aov(Phase_Shift_hr ~ Treatment, data=treatement_shift_data)
summary(model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Treatment    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
```

Answer - P-Value - **0.00447**

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.

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

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

```
TukeyHSD(model, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Phase_Shift_hr ~ Treatment, data = treatement_shift_data)
##
## $Treatment
##                       diff        lwr        upr     p adj
## Eyes-Control  -1.24267857 -2.1682364 -0.3171207 0.0078656
## Knees-Control -0.02696429 -0.9525222  0.8985936 0.9969851
## Knees-Eyes     1.21571429  0.2598022  2.1716263 0.0116776
```

Answer Since the p-values is less than α-level of 0.05, we reject the null hypothesis that all means are equal. Thus the light treatment does affect Phase shift

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()`.

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

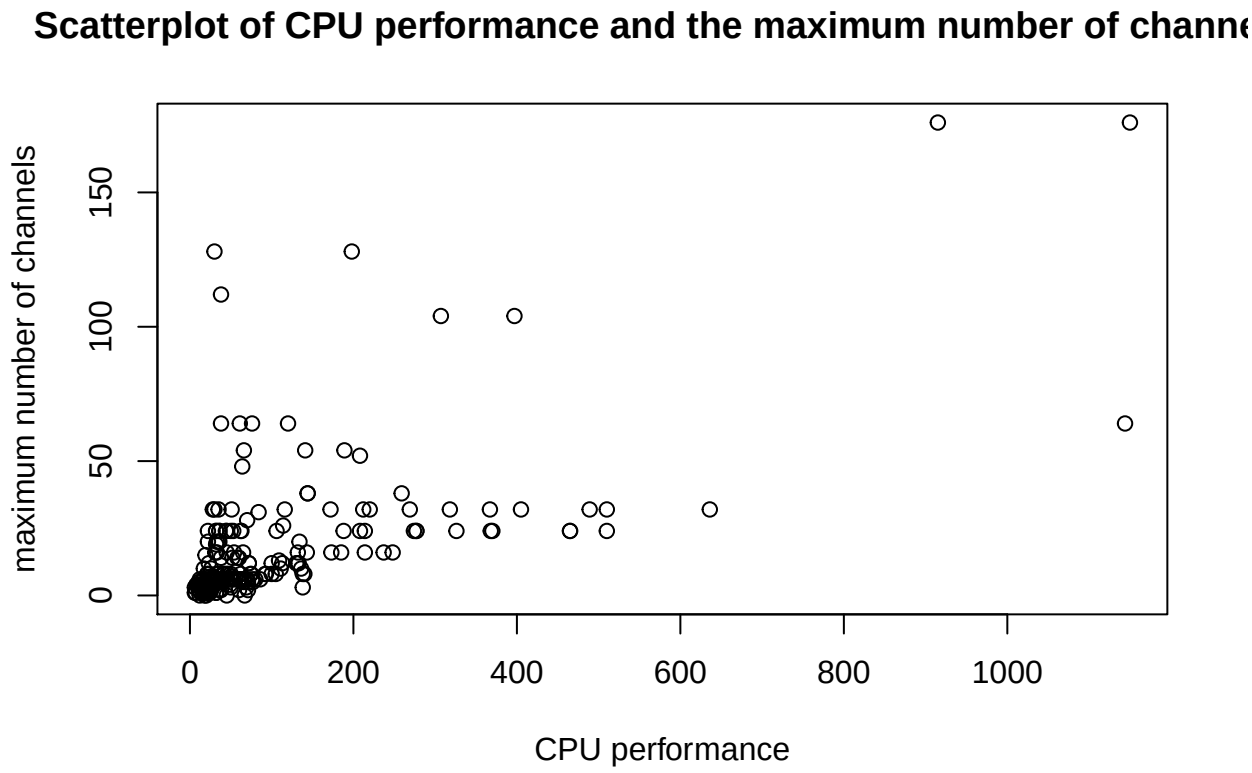
```
##    vendor chmax performance
## 1 adviser   128         198
## 2  amdahl    32         269
## 3  amdahl    32         220
```

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.

Answer:

```r
# Your code here...
plot(data$performance, data$chmax,
     main="Scatterplot of CPU performance and the maximum number of channels",
     xlab="CPU performance",
     ylab="maximum number of channels")
```




```r
#plot(log(data$performance), log(data$chmax))
```

Comments : Scatter plot using original data does show a trend eventhough most of the points are clumped together towards the left bottom corner.

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.

Answer:

```
# Your code here...
cor(data$performance, data$chmax)
```

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

Answer : correlation cofficient of 0.6052093 shows that the correlation is above Moderate and not strong positive relationship

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

Answer: Based on the Exploratory data analysis there exists moderate relationship between the variables. I would still recommend Linear Regression for this... as there are a lot of data-point for lower values of maximum number of channels

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

Answer: Since th data is clumped together in the left bottom corner .. I would pursue data transformation before appling linear Regression to compare the model performace agains the one created without transformation

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...
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?
Answer:
Model parameters

i. $\hat{\beta}_0$ - Intercept - 37.2252
ii. $\hat{\beta}_1$ - Max #Channels - 3.7441
iii. $\hat{\sigma}$ - Std Error- 128.3

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

Answer: $performace = 37.2252 + 3.7441 \cdot chmax$

c. **2 pts** Interpret the estimated value of the β_1 parameter in the context of the problem.

Answer: For every unit of Maximum number of channels increase there is $\sim$ 3.7441 times increase in the CPU performance.

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

Answer:

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

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

95 % confidence Interval for β_1 - 3.069251 - 4.418926

β_1 is statistically significant as evidenced by p-value of 2e-16

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?

Answer:

```
pt(10.938, 207, lower.tail=FALSE)
```

```
## [1] 1.424772e-22
```

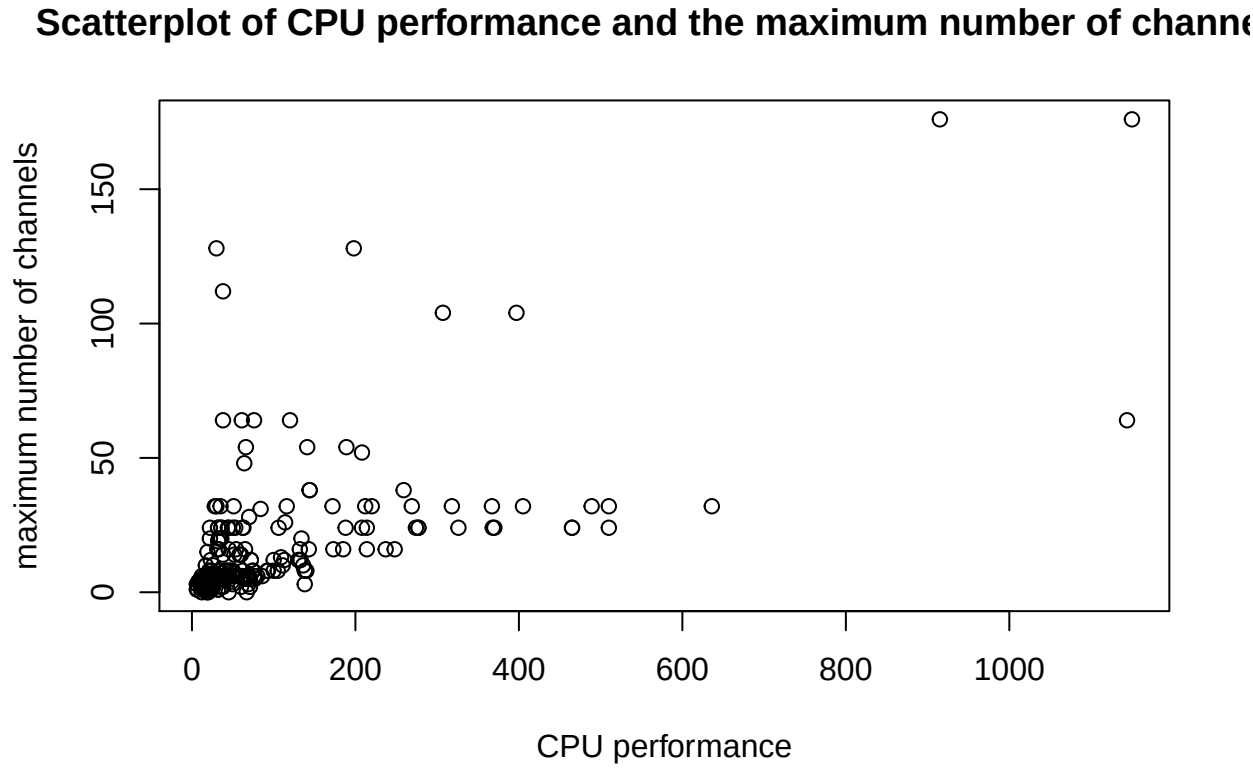
p-value is 1.424772e-22.. This shows that β_1 statistically significantly positive at an α-level of 0.01

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

```
# Your code here...
plot(data$performance, data$chmax,
     main="Scatterplot of CPU performance and the maximum number of channels",
     xlab="CPU performance",
     ylab="maximum number of channels")
```

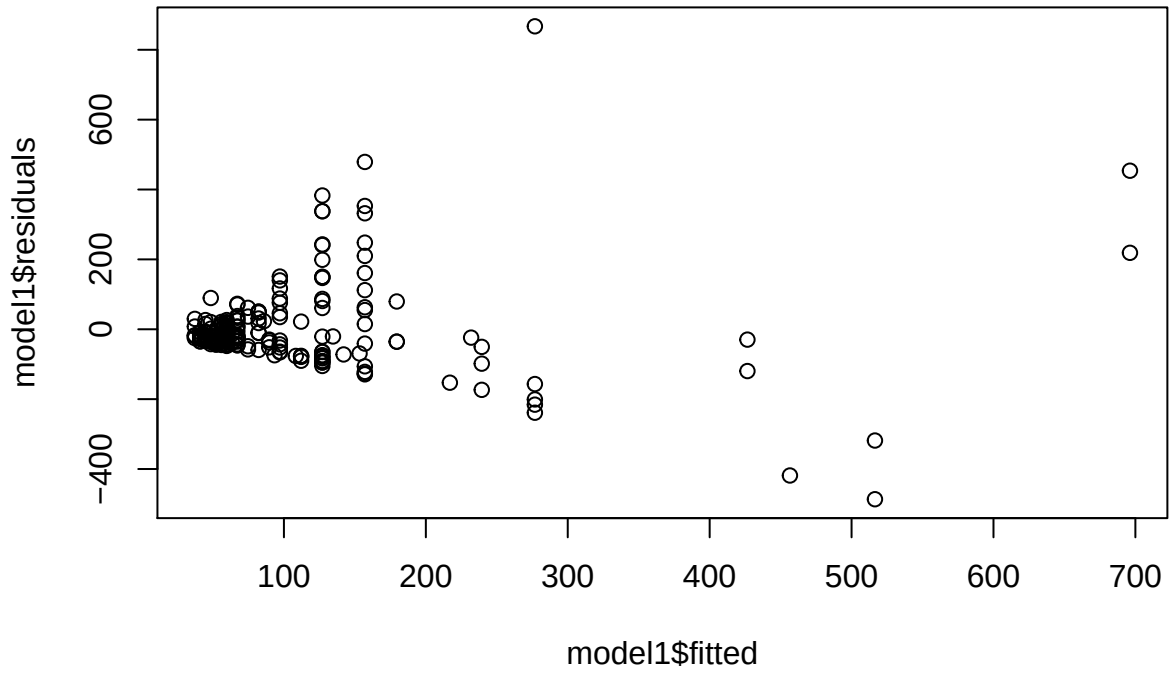


Model Assumption(s) it checks: Answer : Checks for Linearity Assumption.

Interpretation: Answer : The plot looks to have linear relationship.

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

```
# Your code here...
plot(model1$fitted, model1$residuals)
```



Model Assumption(s) it checks: Answer : Checks for Constant variance

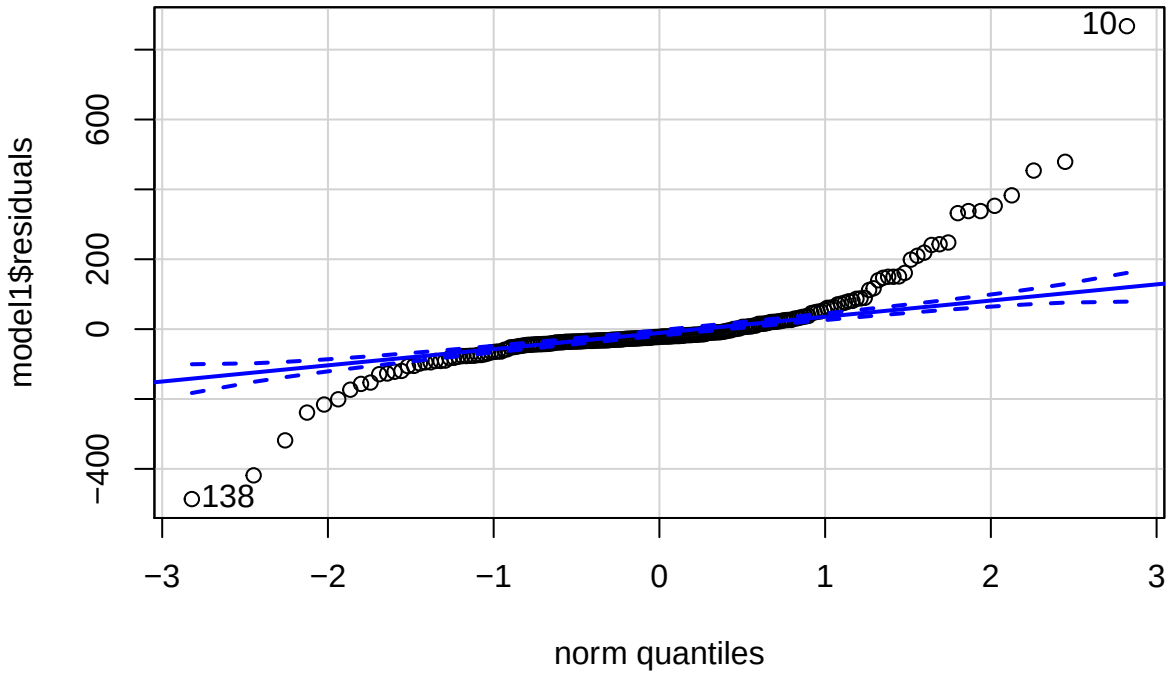
Interpretation: The fitted values vs residuals spread around 0 at random so it looks like the data satisfies Constant variance

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

```
# Your code here...
library(car)
```

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

```
qqPlot(model1$residuals)
```




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

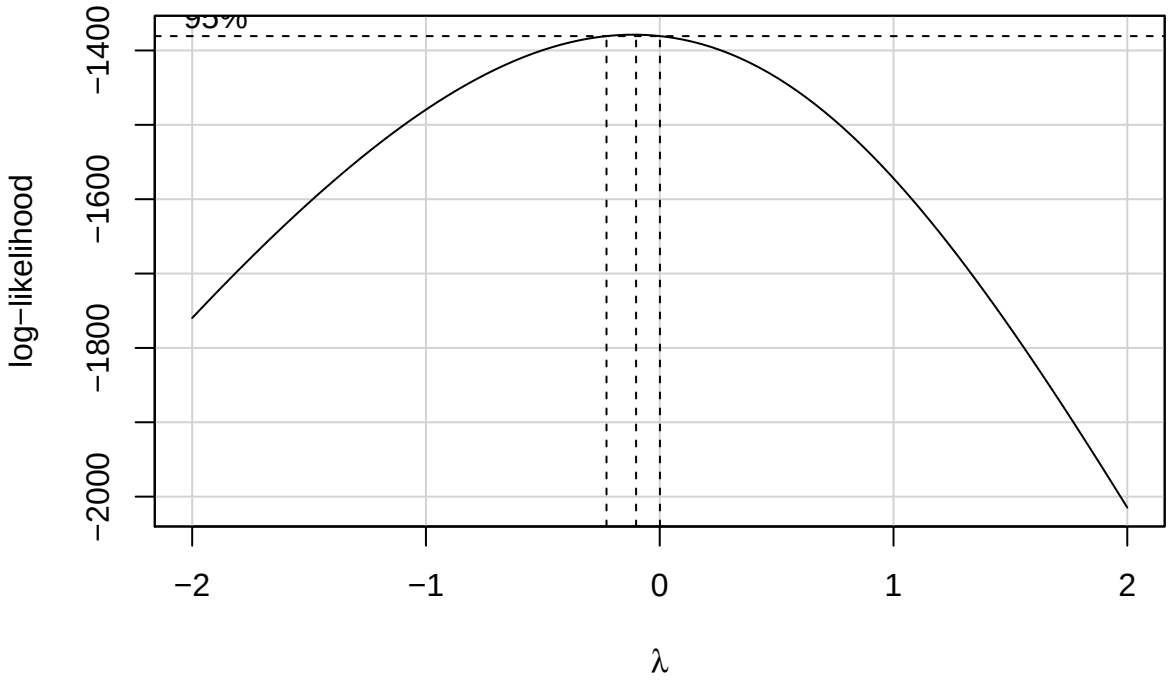
Model Assumption(s) it checks: Answer : Checks for Normality

Interpretation: The data mostly follows normal distribution but is heavy-tailed

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?

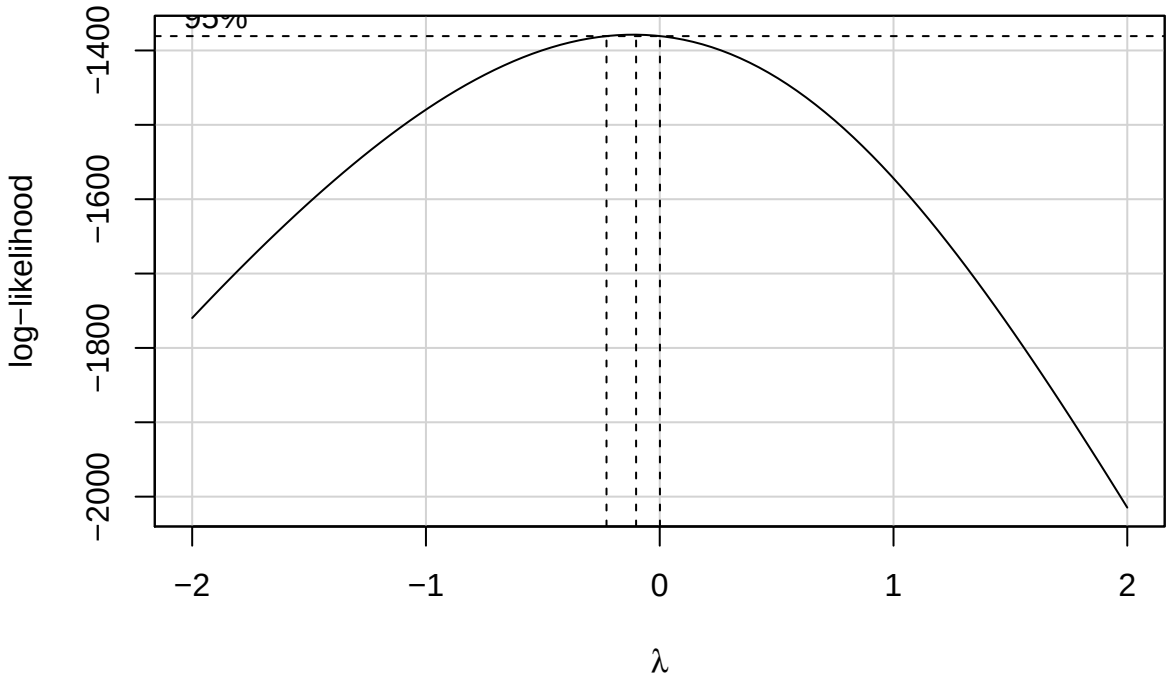
```
b = boxCox(model1)
```




```
lambda <- b$x # lambda values
lik <- b$y # log likelihood values for SSE
bc <- cbind(lambda, lik) # combine lambda and lik
sorted_bc <- bc[order(-lik),] # values are sorted to identify the lambda value for the maximum log like
head(sorted_bc, n = 10)
```

```
##            lambda       lik
##  [1,] -0.10101010 -1378.767
##  [2,] -0.14141414 -1378.874
##  [3,] -0.06060606 -1379.136
##  [4,] -0.18181818 -1379.449
##  [5,] -0.02020202 -1379.986
##  [6,] -0.22222222 -1380.486
##  [7,]  0.02020202 -1381.325
##  [8,] -0.26262626 -1381.977
##  [9,]  0.06060606 -1383.158
## [10,] -0.30303030 -1383.916
```

```
# Your code here...
b = boxCox(data$performance ~ data$chmax)
```




```
lambda <- b$x # lambda values
lik <- b$y # log likelihood values for SSE
bc <- cbind(lambda, lik) # combine lambda and lik
sorted_bc <- bc[order(-lik),] # values are sorted to identify the lambda value for the maximum log like
head(sorted_bc, n = 10)
```

```
##            lambda       lik
##  [1,] -0.10101010 -1378.767
##  [2,] -0.14141414 -1378.874
##  [3,] -0.06060606 -1379.136
##  [4,] -0.18181818 -1379.449
##  [5,] -0.02020202 -1379.986
##  [6,] -0.22222222 -1380.486
##  [7,]  0.02020202 -1381.325
##  [8,] -0.26262626 -1381.977
##  [9,]  0.06060606 -1383.158
## [10,] -0.30303030 -1383.916
```

Answer : the top λ value suggests between -0.06 to -0.01. taking to nearest half integer.. it suggests $\lambda = 0$

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

```
# Your code here...
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
```

```
confint(model2, level = 0.95)
```

```
##                    2.5 %    97.5 %
## (Intercept)    2.1975363 2.7555568
## log(chmax + 1) 0.5417055 0.7546727
```

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

```
summary(model1)$r.squared
```

```
## [1] 0.3662783
```

```
summary(model2)$r.squared
```

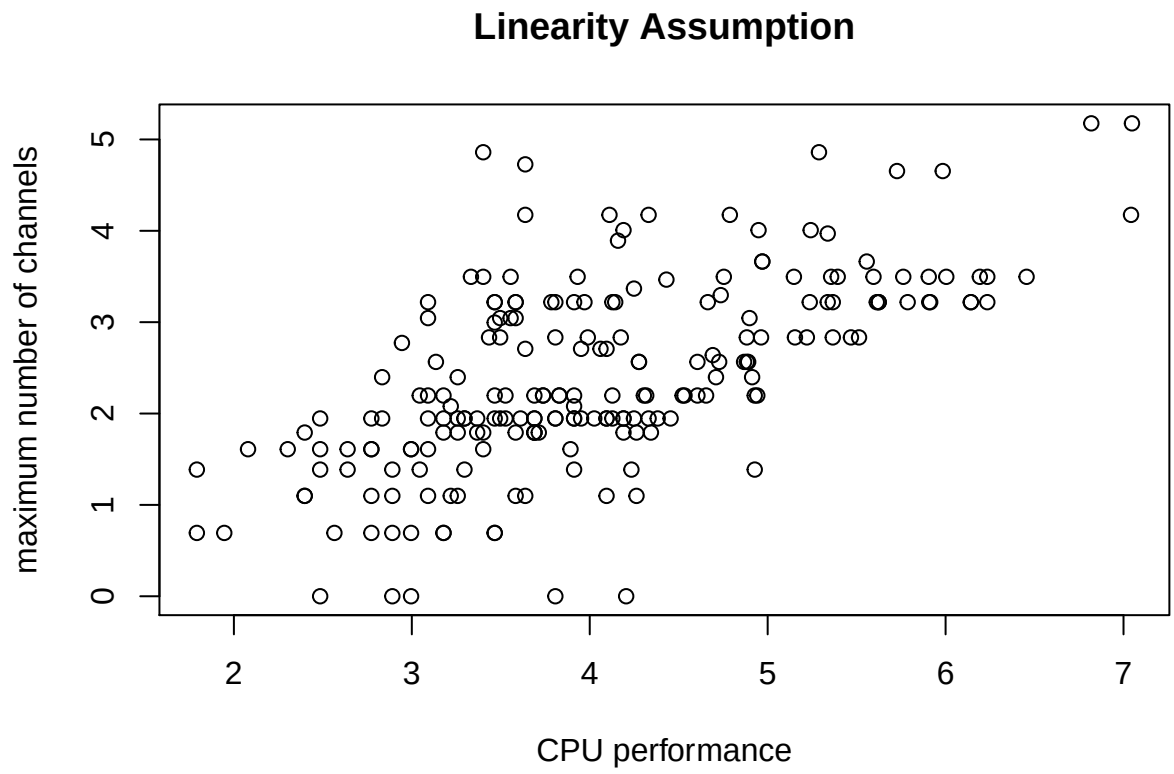
```
## [1] 0.4102926
```

Answer : Model1 -> $R^2 = 0.3662783$. Where as Model2 -> $R^2 = 0.4102926$

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?

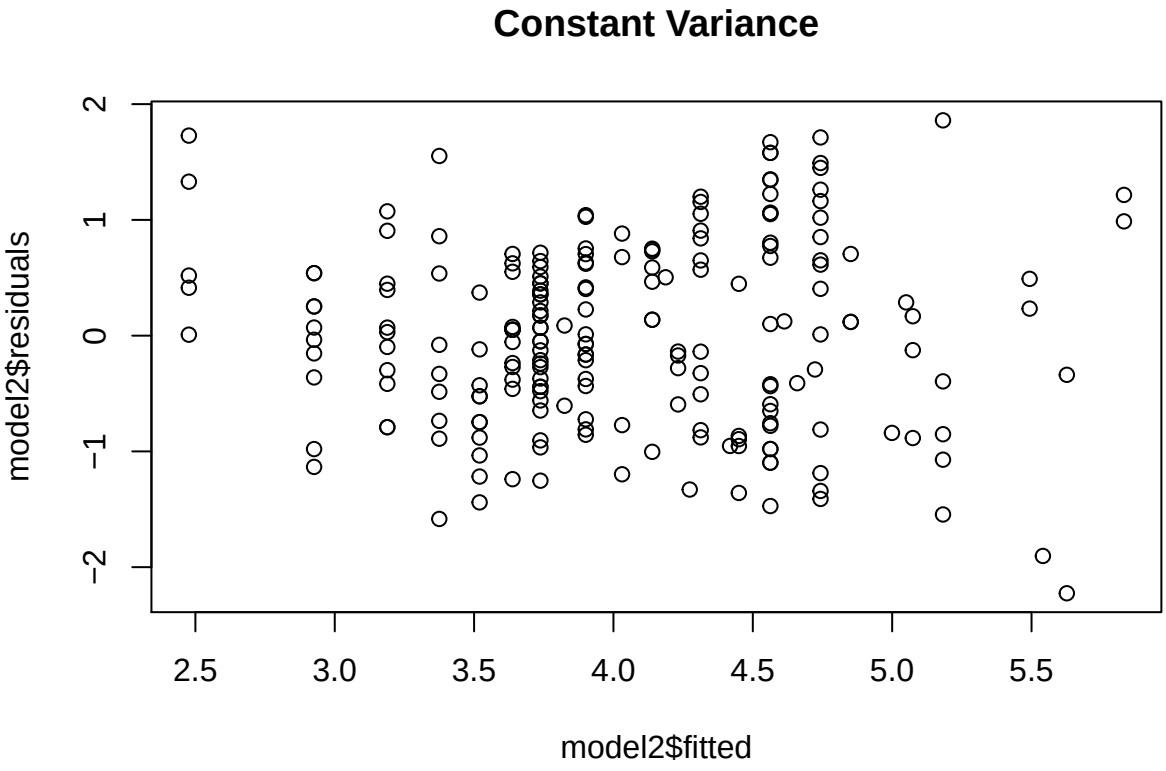
Model Assumption : Linearity Assumption.

```
# Your code here...
plot(log(data$performance), log(data$chmax + 1),
     main="Linearity Assumption",
     xlab="CPU performance",
     ylab="maximum number of channels")
```



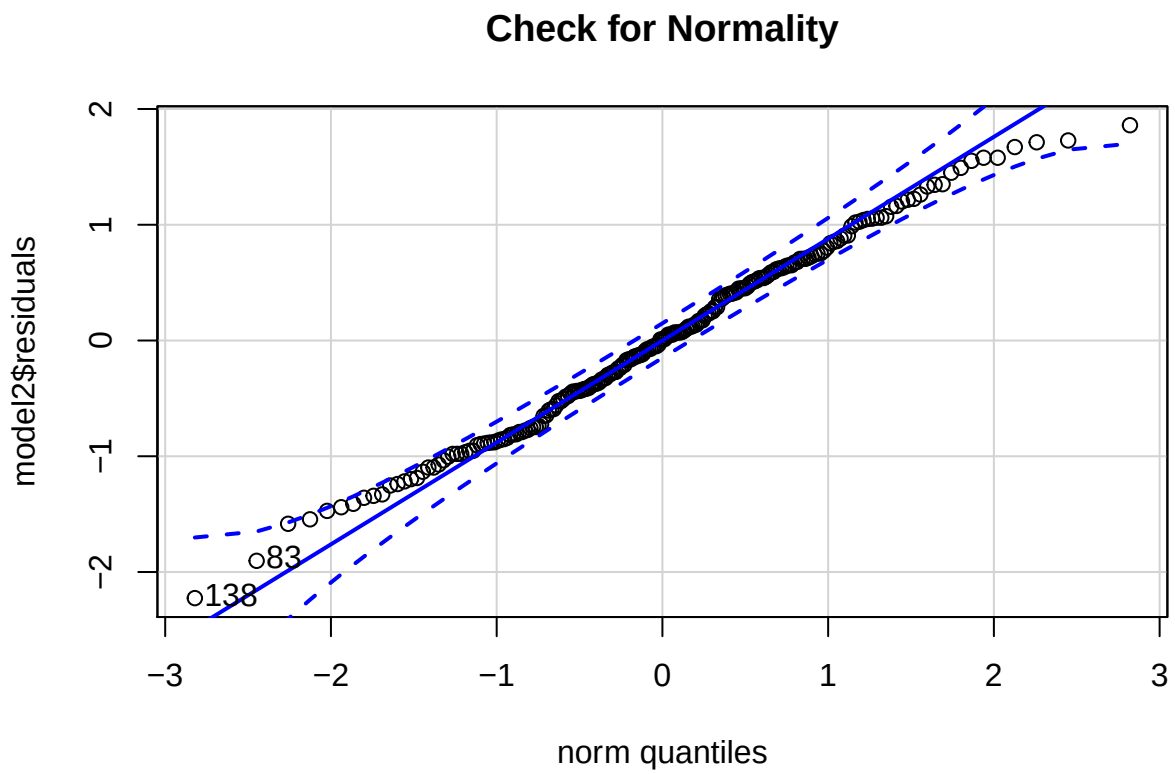
Model Assumption : Constant variance.

```
# Your code here...
plot(model2$fitted, model2$residuals, main="Constant Variance")
```



Model Assumption : Normality

```
# Your code here...
qqPlot(model2$residuals, main="Check for Normality")
```




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

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?

```
# Your code here...
# newppp= data.frame(Inflation.difference= c(-0.68))
newppp= data.frame(chmax= c(128))
predict(model1,newppp,interval=c("prediction"), level = 0.95)
```

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

```
exp(predict(model2,newppp,interval=c("prediction"), level = 0.95))
```

```
##       fit      lwr      upr
## 1 277.723 55.17907 1397.813
```

```
**Answer : **

1. Model 1 - the predicted value - 516.4685 with confidence interval of 252.2519 & 780.6851. The predic
2. Model 2 - the predicted value - 277.723 with confidence interval of 55.17907 & 1397.813. The predicte

Looking at this .. seems like the Model 2 prediction has a tighter confidence interval and cab be deeme
```

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.

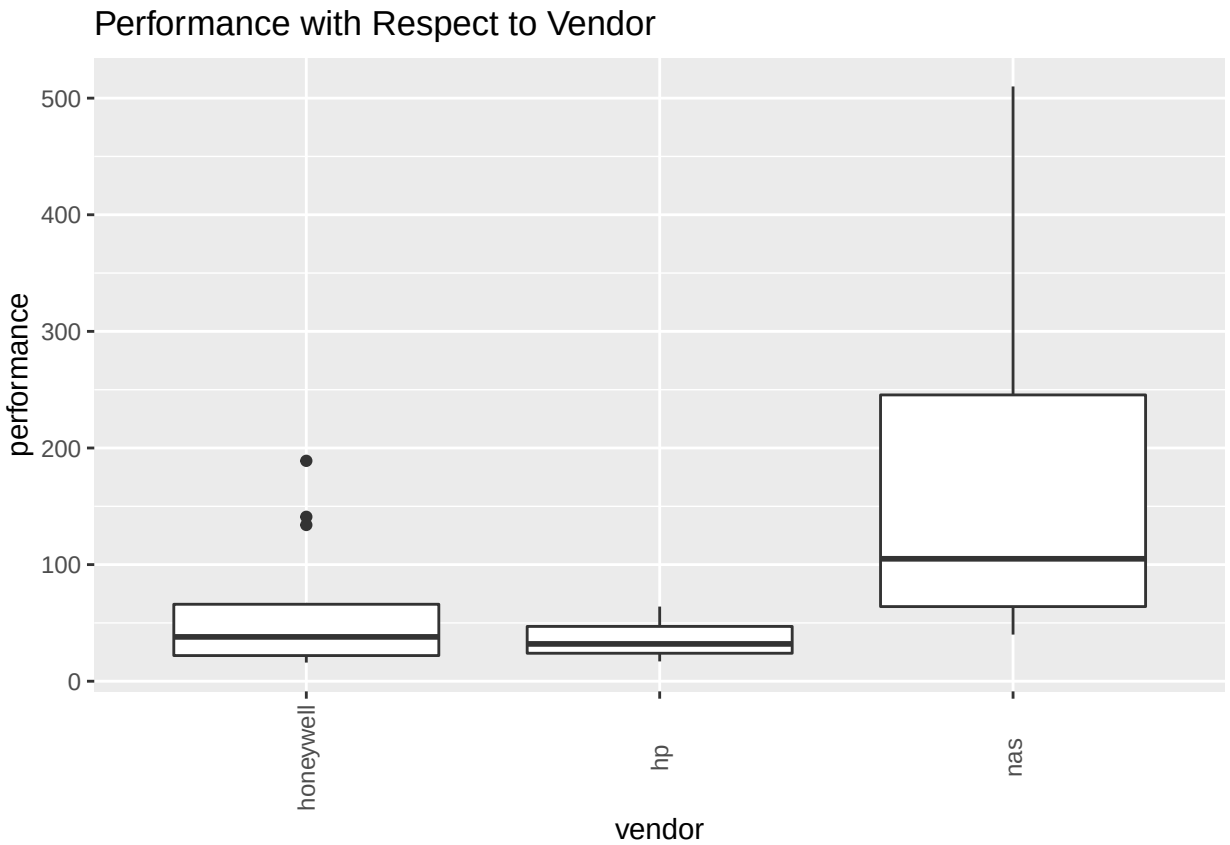
```
# Filter for honeywell, hp, and nas
data2 = data[data$vendor %in% c("honeywell", "hp", "nas"), ]
data2$vendor = factor(data2$vendor)
```

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

```
# Your code here...
aggregate(data2$performance, list(data2$vendor), mean)
```

```
##     Group.1         x
## 1 honeywell  60.46154
## 2        hp  36.42857
## 3       nas 176.89474
```

```
ggplot(data2, aes(x=vendor, y=performance),
       xlab = "Vendor",
       ylab = "Performance") +
  geom_boxplot() +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))+
  ggtitle("Performance with Respect to Vendor")
```



Answer :

* honeywell and hp look to have similar median

* nas seems to have higher medin that honeywell and hp
* honeywell has some outliers

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.

```
# Your code here...
amodel = aov (data2$performance ~ data2$vendor)
summary(amodel)
```

```
##              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
```

Answer : The P-value of 0.00553 suggest that not all means are equal. Thus we reject the Null hypothesis and say that there are some means that are different. The Box Plot also shows the same thing

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?

```
# Your code here...
TukeyHSD(amodel, conf.level = 0.95)
```

```
##   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
```

ANSWER : Based on α-level of 0.05 1. there are 2 pairs of men that are statistically different - nas-honeywell & nas-hp, 2. we cannot say that hp-honeywell have different means. 3. but both honeywell and hp have means different for nas 4. Also both lower and upper thresholds for nas-honeywell and nas-hp are both positive.. so statistically their means are not same