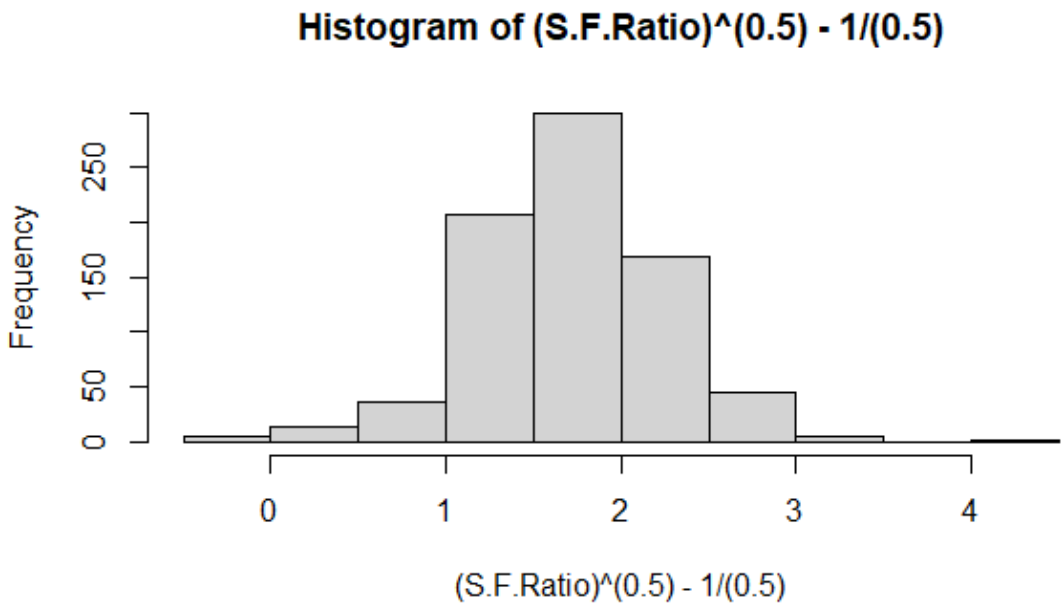
P1 - Analise Exploratória de Dados

Nome: Davi dos Santos Mattos - 119133049 ; Lucas Jesus Lapa - 117086199

Questão 1.

1.a)



Acima temos um boxplot da variável “S.F.Ratio”, onde podemos ver sua distribuição em torno de 5 à 25 estudantes por professor em cada faculdade, e casos extremos, onde podemos ter poucos estudantes por professor ou muitos estudantes. Tendo sua dispersão, ou seja, a maioria de casos, ao redor de estudantes por professor.

Códigos Utilizados:

```
library(ISLR)
attach(College)

#Criando a imagendo do BoxPlot
jpeg("S.F.Ratio_Tr.jpeg")
hist((S.F.Ratio)^(0.5)-1/(0.5))
dev.off()
```

1.b)

Calculando a diferença interquartil, posteriormente as cercas superior e inferior da variável, e comparando com os dados da variável, nós chegamos aos seguintes outliers:

Nome da Universidade	S.F.Ratio	Tipo de Outlier
Case Western Reserve University	2.9	Anormalmente Baixo
Johns Hopkins University	3.3	Anormalmente Baixo
University of Charleston	2.5	Anormalmente Baixo
Goldey Beacom College	27.6	Anormalmente Alto
Indiana Wesleyan University	39.8	Anormalmente Alto
Lesley College	27.8	Anormalmente Alto
Mesa State College	28.8	Anormalmente Alto
Saint Joseph's College	27.2	Anormalmente Alto
University of Texas at San Antonio	25.3	Anormalmente Alto
Western Michigan University	24.7	Anormalmente Alto

Tendo tais dados de forma explícita, é recomendado que as faculdade cujo o nome consta no tipo de outlier, anormalmente alto, contratassem mais professores. No caso das faculdade onde o tipo de outlier estar categorizado como “Anormalmente Baixo”, é recomendado que eles alocassem mais mais estudantes por professor

Códigos utilizados:

```
library(ISLR)
attach(College)

# Analisando os Dados e Calculando os Cercas e Diferença Interquartil
summary(S.F.Ratio)
quantile(S.F.Ratio)
DIQ <- 16.5-11.5
inf <- 11.5- 1.5 * DIQ
sup <- 15.5 + 1.5 * DIQ

#Filtrando os dados
filtro1 <- subset(College, round(S.F.Ratio) < inf, select=c(S.F.Ratio)) # Outlines Inferiores
filtro2 <- subset(College, round(S.F.Ratio) > sup, select=c(S.F.Ratio)) # Outlines Superiores

View(filtro1)
View(filtro2)
```

1.c)

Gráfico sem transformação de variável:

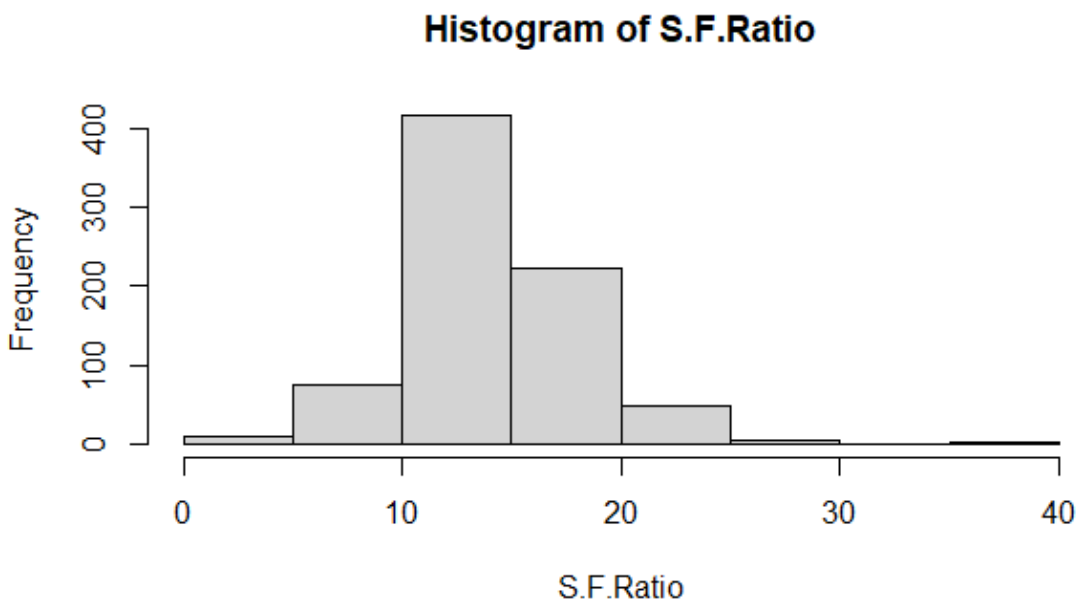
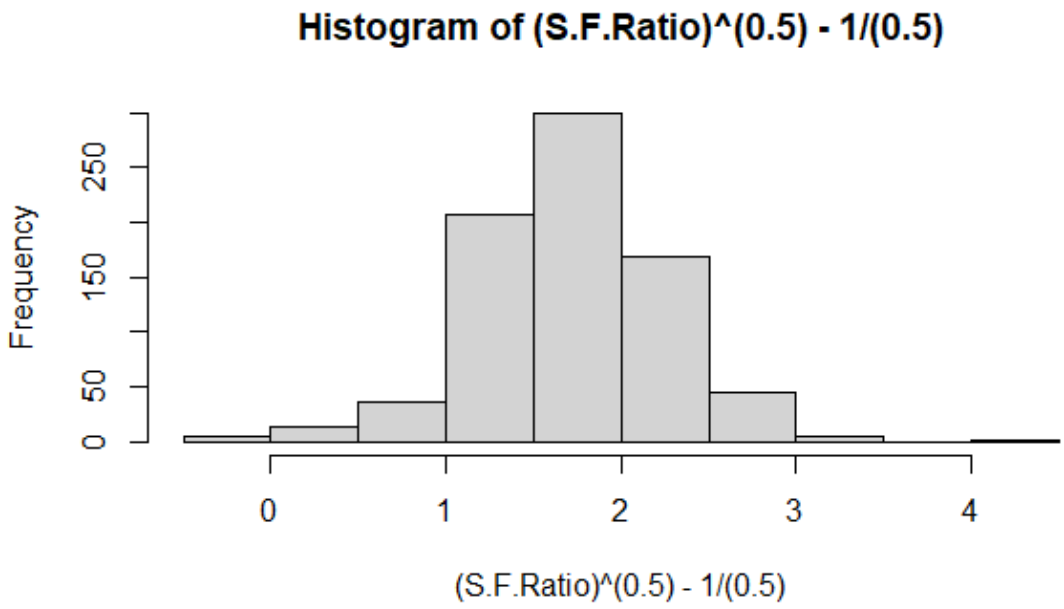


Gráfico com transformação de variável:



Como podemos ver acima, no gráfico sem a transformação, podemos ver que a variável tem uma distribuição assimétrica, e assim portanto sendo necessária uma transformação. A transformação utilizada foi arredondar os valores da variável para ter uma visão absoluta da mesma. Pois como a variável se trata de “Número de estudantes por professor”, uma variável cujo o valor seja, por exemplo, 3.9 estudantes por professor

e sua cerca inferior fosse 4, poderia acabar sendo tratada como outline sendo que na prática não há 3,9 estudantes, mas 3 ou 4 estudantes.

(row)	S.F.Ratio
Case Western Reserve University	2.9
Johns Hopkins University	3.3
University of Charleston	2.5

(row)	S.F.Ratio
Case Western Reserve University	2.9
Johns Hopkins University	3.3
University of Charleston	2.5
Washington University	3.9

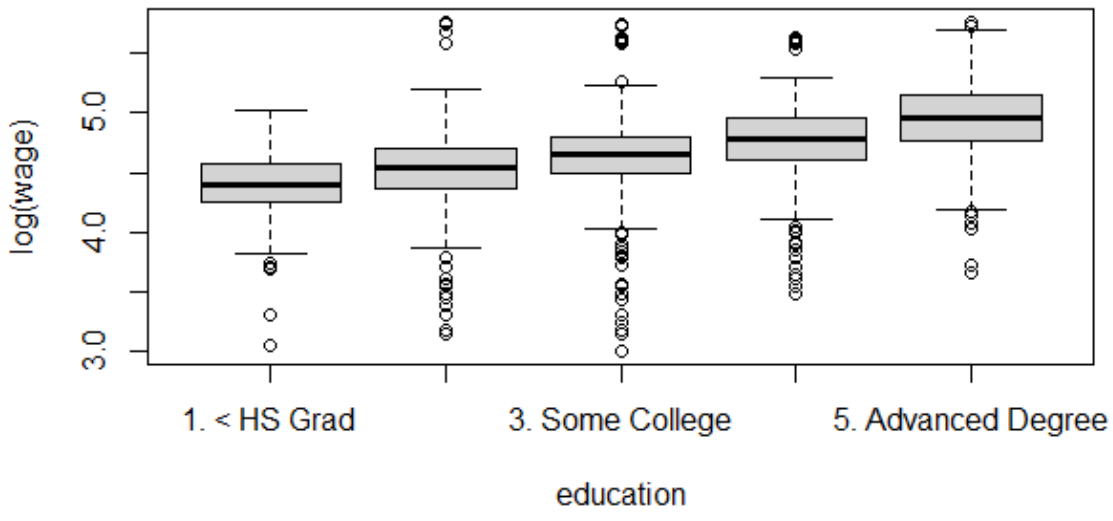
O mesmo se aplica caso peguemos a Cerca superior:

(row)	S.F.Ratio
Goldey Beacom College	27.6
Indiana Wesleyan University	39.8
Lesley College	27.8
Mesa State College	28.8
Saint Joseph's College	27.2
University of Texas at San Antonio	25.3
Western Michigan University	24.7

(row)	S.F.Ratio
Goldey Beacom College	27.6
Indiana Wesleyan University	39.8
Lesley College	27.8
Lindenwood College	24.1
Mesa State College	28.8
Saint Joseph's College	27.2
University of Texas at San Antonio	25.3
Western Michigan University	24.7

Questão 2.

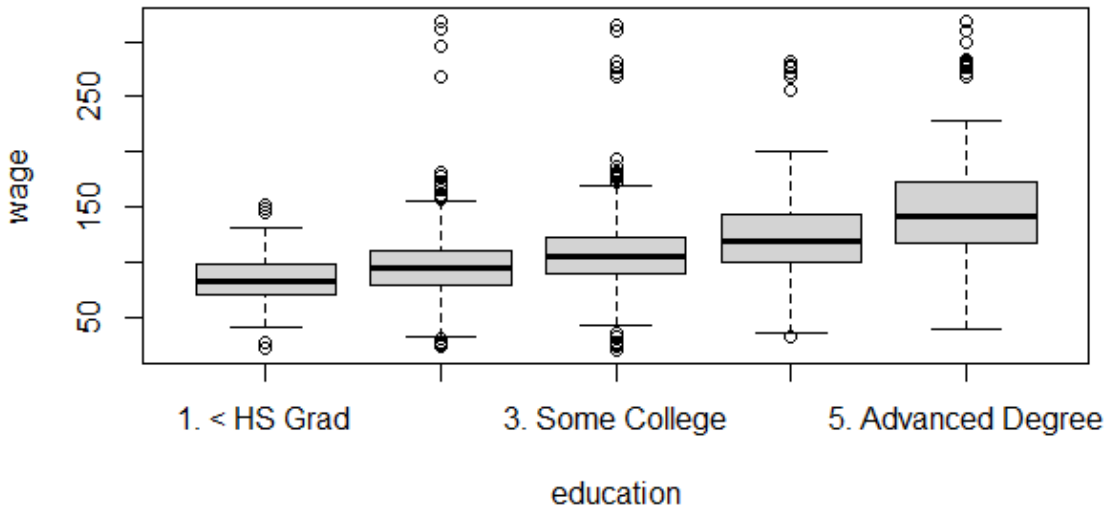
2.a)



Através do gráfico acima nós podemos concluir que, conforme maior for o grau de educação melhor será o salário, porém com alguns casos discrepantes onde o salário pode se menor ou maior que a média devido ao grau educacional. Porém com a tendência de que o salário seja menor ou anormalmente menor, caso o grau de educação seja menor que Ensino Médio.

2.b)

Gráfico sem transformação:



Através do gráfico acima da variável sem transformação, nós podemos concluir que não há homogeneidade entre as dispersões, apenas a terceira e a segunda são homogêneas entre si, mas em relação as outras dispersões, não há homogeneidade. Portanto foi necessário utilizar a transformação de variável $log(wage), \lambda = 0$.

Comandos Utilizados:

```
simetr <- function(lambda){
    if(lambda == 0){tr <- log(wage)} else{tr <- ((wage^lambda)-1)/lambda}
    q1 <- quantile(tr,0.25)
    q2 <- quantile(tr,0.5)
    q3 <- quantile(tr,0.75)
    return(abs(q2-0.5*(q1+q3))/q2)
}

lambda <- c(-2,-1,-0.5,0,0.5,1,2)

valorTr <- 999
lbd <- 0
```

```
for(num in lambda){
    print(num)
    print(simetr(num))
}

lbd


boxplot(log(wage)~education)
```

2.c)

Olhando para a educação equivalente ao Ensino Médio completo, nós temos alguns outliers, como por exemplo:

Salário	Tipo de Outlier
318.3424	Anormalmente Alto
311.9346	Anormalmente Alto
295.9912	Anormalmente Alto
267.9011	Anormalmente Alto
267.9011	Anormalmente Alto
23.95294	Anormalmente Baixo
23.27470	Anormalmente Baixo

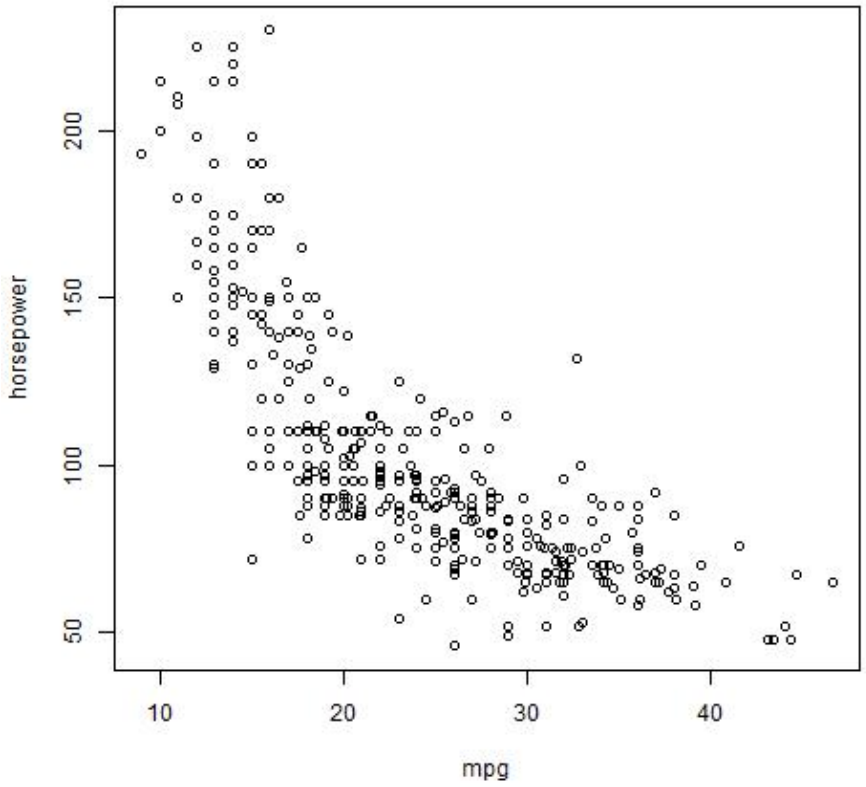
Analisando os salário discrepantes e os comparando com o resto do conjunto de dados, nós conseguimos ver que os salários mais altos para a médio do grau de escolaridade, geralmente possuem plano de saúde, diferentemente dos que recebem salários extremamente baixos, onde nenhum possui plano de saúde.

Comandos utilizados:

```
df_Wage <- data.frame(education=rep(education), wage=rep(wage))
filtro1 <- subset(df_Wage, education == "2. HS Grad")
summary(filtro1)

DIQ <- 109.83 - 77.95
inf <- 77.95 - 1.5 * DIQ
sup <- 109.83 + 1.5 * DIQ

filtro3 <- subset(Wage, wage < inf & education == "2. HS Grad")
filtro4 <- subset(Wage, wage > sup & education == "2. HS Grad")

View(filtro4)
View(filtro3)
```

Questão 3.

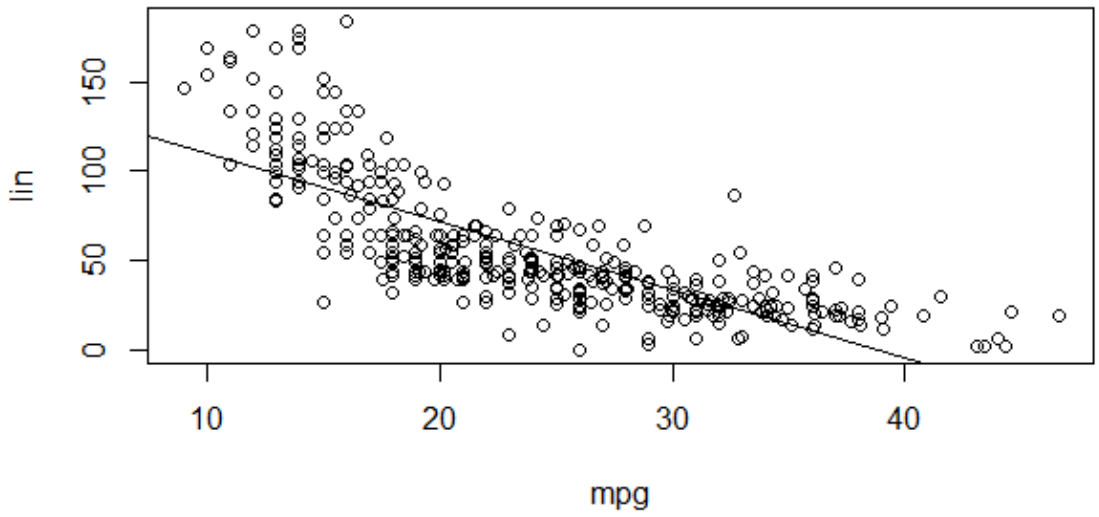
3.a)



Pelo gráfico acima podemos concluir que quanto maior a potência dos carros, menor será seu desempenho em relação km/l (mpg).

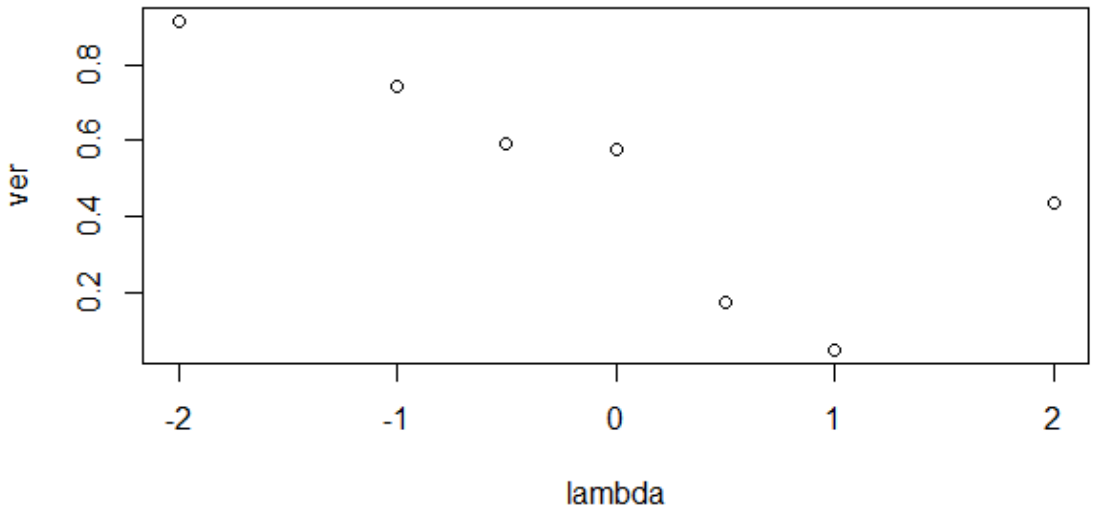
3.b) e c)

$variável\_transformada = 148.476 + (-3.839) * horsepower$



De acordo com o gráfico acima, onde nós transformamos a variável $hoserpower$ e o relacionamos novamente com a variável mpg, podemos ver que na maioria das vezes a potência do carro influencia em seu desempenho, ou seja, a equação apresentada é considerada forte.

Tendo em vista que originalmente os dados orginalmente indicam que o desempenho do carro não é uma função linear, tornou-se necessário utilizar a transformação de variável $\frac{y^{\lambda}-1}{\lambda}, \lambda \neq 0$. Porém, tendo em vista que apenas utilizar $\lambda = 1$ não é o suficiente, tivemos que subtrair 45 do valor de $horspower$, para reduzirmos para aproxima-lo a origem, ou seja, pegar o menor numero da variável e subtrair uma constate, no caso $46(valor\_variável) - 45(constante)$, e também escolher 3 pontos do diagrama para chegarmos numa melhor dispersão dos dados, sendo esses ponto um a extrema esquerda, um perto do centro e outro a extrema direita.

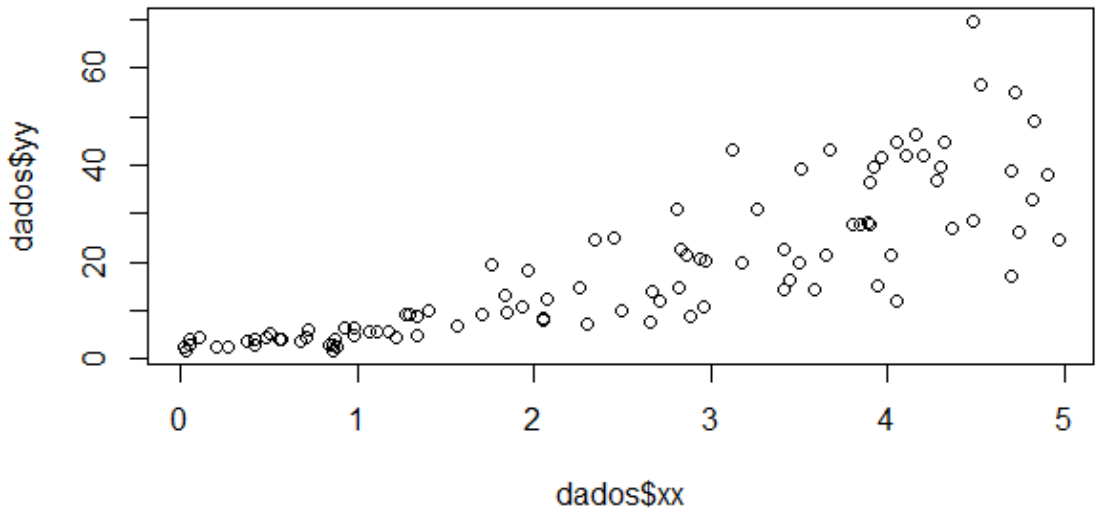



```
# Escolha dos pontos
yA<-195
yB<-125
yC<-65
xA<-9
xB<-28
xC<-46
sAB<- (yB-yA)/(xB-xA)
sBC<- (yC-yB)/(xC-xB)
sAB
sBC

simetr <- function(lambda){
  if(lambda == 0){tr <- log(horsepower-45)} else{tr <- (((horsepower-45)^lambda)-1)/lambda}
  q1 <- quantile(tr,0.25)
  q2 <- quantile(tr,0.5)
  q3 <- quantile(tr,0.75)
  return(abs(q2-0.5*(q1+q3))/q2)
}

lambda <- c(-2,-1,-0.5,0,0.5,1,2)

eq.fn<-function(lambda){
  if(lambda == 0){zA=log(yA-60)} else {zA=((yA-45)^lambda-1)/lambda}
  if(lambda == 0){zB=log(yB-60)} else {zB=((yB-45)^lambda-1)/lambda}
  if(lambda == 0){zC=log(yC-60)} else {zC=((yC-45)^lambda-1)/lambda}
  nslAB=(zB-zA)/(xB-xA)
  nslBC=(zC-zB)/(xC-xB)
  return(abs((nslBC-nslAB)/(nslBC+nslAB)))
}

ver<-c(eq.fn(-2),eq.fn(-1),eq.fn(-0.5),eq.fn(0),eq.fn(0.5),eq.fn(1),eq.fn(2))
plot(lambda,ver)

lin <- ((horsepower-45)^1-1)/1
```

```
plot(mpg,lin)
teste=lm(lin~mpg)
teste
abline(a=148.476,b=-3.839)
```

Questão 4.

4.a)

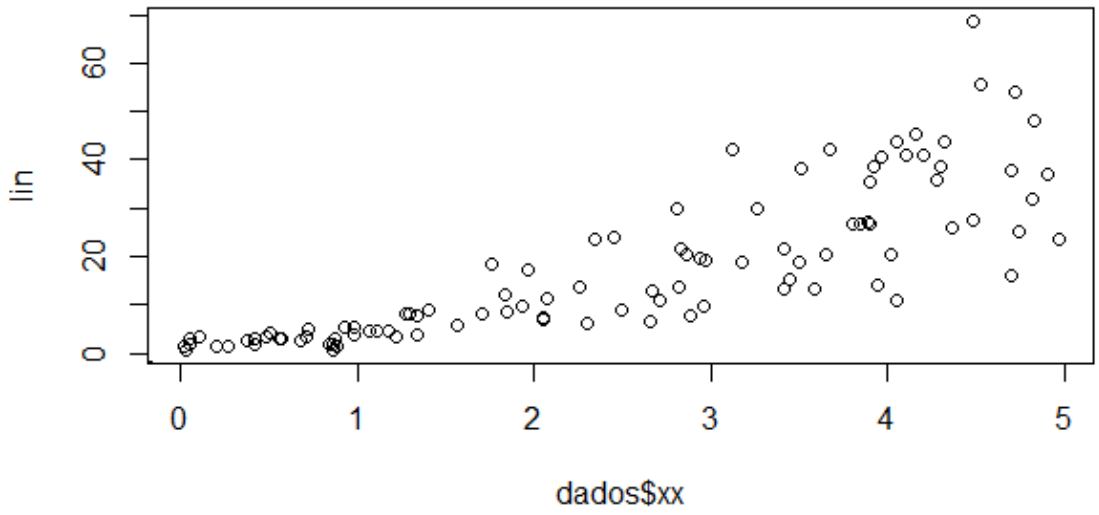


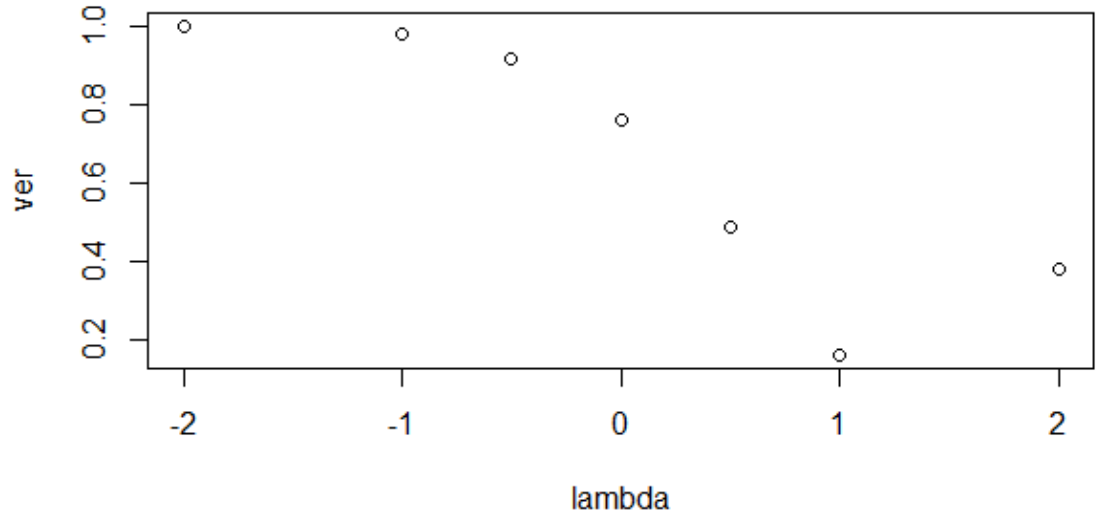
4.b)

Analisando o gráfico acima, parece existir uma relação linear entre as variáveis.

4.c)

Gráfico após linearização:





4.d)

Após usar o critério de linearização o valor de lambda escolhido é igual 1 como pode ser visto acima após escolher 3 pontos do gráfico referentes ao extrema esquerda, médio e extrema direita para achar sua transformação usual, chegando assim na expressão matemática abaixo.

$$variável\_transformada = 3.224 + (8.373) * dados\$yy$$

```
plot(dados$xx,dados$yy)

yA<-1
yB<-30
yC<-65
xA<-1
xB<-2.5
xC<-5
sAB<- (yB-yA)/(xB-xA)
sBC<- (yC-yB)/(xC-xB)
sAB
sBC

simetr <- function(lambda){
  if(lambda == 0){tr <- log(dados$yy)} else{tr <- (((dados$yy)^lambda)-1)/lambda}
  q1 <- quantile(tr,0.25)
  q2 <- quantile(tr,0.5)
  q3 <- quantile(tr,0.75)
  return(abs(q2-0.5*(q1+q3))/q2)
}

lambda <- c(-2,-1,-0.5,0,0.5,1,2)

eq.fn<-function(lambda){
  if(lambda == 0){zA=log(yA)} else {zA=((yA)^lambda-1)/lambda}
  if(lambda == 0){zB=log(yB)} else {zB=((yB)^lambda-1)/lambda}
  if(lambda == 0){zC=log(yC)} else {zC=((yC)^lambda-1)/lambda}
  nslAB=(zB-zA)/(xB-xA)
  nslBC=(zC-zB)/(xC-xB)
  return(abs((nslBC-nslAB)/(nslBC+nslAB)))
}

ver<-c(eq.fn(-2),eq.fn(-1),eq.fn(-0.5),eq.fn(0),eq.fn(0.5),eq.fn(1),eq.fn(2))
plot(lambda,ver)

lin <- ((dados$yy)^1-1)/1

plot(dados$xx,lin)
teste=lm(lin~dados$xx)
teste
abline(a=-3.224,b=-8.373)
```