R Notebook

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

### Dependency

##library(tidyverse)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readxl)  
library(ggplot2)  
library(ggpubr)

### Read xls file

file\_name = "Hydrocarbon(2).xlsx"  
df = read\_excel(file\_name)

### Preview tibble

df

## # A tibble: 32 x 6  
## Index X1 X2 X3 X4 Y  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 33 53 3.32 3.42 29  
## 2 2 31 36 3.1 3.26 24  
## 3 3 33 51 3.18 3.18 26  
## 4 4 37 51 3.39 3.08 22  
## 5 5 36 54 3.2 3.41 27  
## 6 6 35 35 3.03 3.03 21  
## 7 7 59 56 4.78 4.57 33  
## 8 8 60 60 4.72 4.72 34  
## 9 9 59 60 4.6 4.41 32  
## 10 10 60 60 4.53 4.53 34  
## # ... with 22 more rows

# Q-1

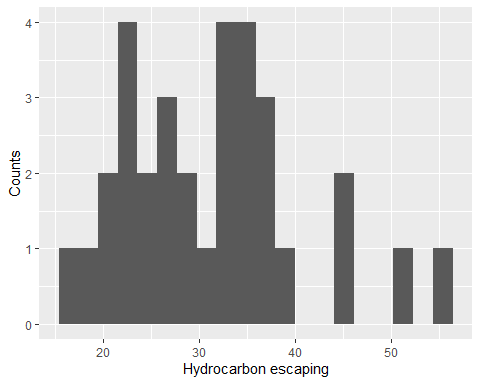
## SUmmary

summary(df)

## Index X1 X2 X3   
## Min. : 1.00 Min. :31.00 Min. :35.00 Min. :2.590   
## 1st Qu.: 8.75 1st Qu.:37.00 1st Qu.:41.00 1st Qu.:3.290   
## Median :16.50 Median :60.00 Median :60.00 Median :4.285   
## Mean :16.50 Mean :57.91 Mean :55.91 Mean :4.422   
## 3rd Qu.:24.25 3rd Qu.:62.00 3rd Qu.:62.00 3rd Qu.:4.630   
## Max. :32.00 Max. :92.00 Max. :92.00 Max. :7.450   
## X4 Y   
## Min. :2.590 Min. :16.00   
## 1st Qu.:3.373 1st Qu.:23.75   
## Median :4.090 Median :31.50   
## Mean :4.324 Mean :31.12   
## 3rd Qu.:4.540 3rd Qu.:34.50   
## Max. :7.450 Max. :55.00

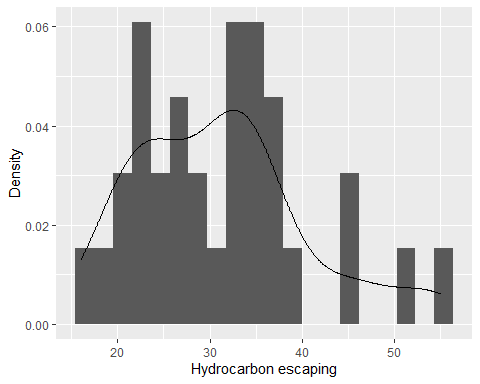
## Histogram

ggplot(df, aes(x=Y)) + geom\_histogram(bins=20) + labs( y="Counts", x="Hydrocarbon escaping")



## Histogram with density curve

ggplot(df, aes(x=Y)) + geom\_histogram(aes(y=..density..),bins=20) + geom\_density(alpha=.2)+ labs( y="Density", x="Hydrocarbon escaping")



*So looking at the graphs above, we can see that the curve is right skewed. This means more observations were found with hydrocarbon escaping values less than mean*

# Q-2

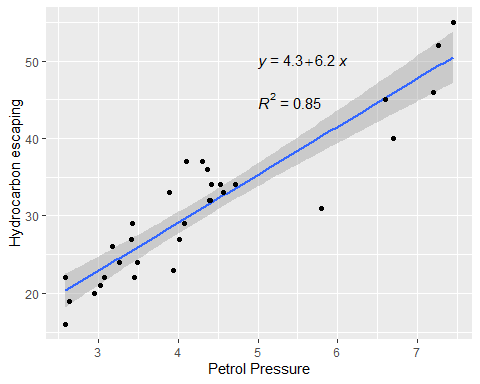
fit<- lm(df$Y~df$X4)  
summary(fit)

##   
## Call:  
## lm(formula = df$Y ~ df$X4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.2820 -2.1353 0.3438 2.0466 7.2629   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.3054 2.1674 1.986 0.0562 .   
## df$X4 6.2029 0.4779 12.980 7.65e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.703 on 30 degrees of freedom  
## Multiple R-squared: 0.8489, Adjusted R-squared: 0.8438   
## F-statistic: 168.5 on 1 and 30 DF, p-value: 7.654e-14

## Scatter plot

#create plot with regression line, regression equation, and R-squared  
ggplot(data=df, aes(x=X4, y=Y)) +  
 geom\_smooth(method="lm") +  
 geom\_point() +  
 stat\_regline\_equation(label.x=5, label.y=50) +  
 stat\_cor(aes(label=..rr.label..), label.x=5, label.y=45) + labs( x="Petrol Pressure", y="Hydrocarbon escaping")

## `geom\_smooth()` using formula 'y ~ x'



cor(x= df$X4, y= df$Y)

## [1] 0.9213333

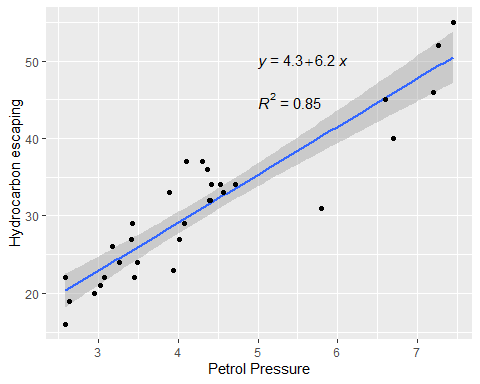
*So based on the graph and the correlation value we can say that the relation is linear, pretty strong and positive. There seems to be one outlier for X4 = 5.8*

# Q-3

##Least Square Regression Line

#create plot with regression line, regression equation, and R-squared  
ggplot(data=df, aes(x=X4, y=Y)) +  
 geom\_smooth(method="lm") +  
 geom\_point() +  
 stat\_regline\_equation(label.x=5, label.y=50) +  
 stat\_cor(aes(label=..rr.label..), label.x=5, label.y=45) + labs( x="Petrol Pressure", y="Hydrocarbon escaping")

## `geom\_smooth()` using formula 'y ~ x'

 ## testing df

t\_df = data.frame(X4 = c(3.4,4,4.5))  
t\_df

## X4  
## 1 3.4  
## 2 4.0  
## 3 4.5

intercept <- fit$coefficients[1]  
slope <- fit$coefficients[2]  
intercept

## (Intercept)   
## 4.305422

slope

## df$X4   
## 6.202851

### predict\_sales <- intercept + slope\*(x)

### Y = 4.305422 + 6.202851(X)

t\_df$pred\_Y = intercept + slope\*(t\_df$X4)  
t\_df

## X4 pred\_Y  
## 1 3.4 25.39512  
## 2 4.0 29.11683  
## 3 4.5 32.21825

# Q-4

df.lm = lm(Y ~ X4, data=df)   
df.res = resid(df.lm)  
df

## # A tibble: 32 x 6  
## Index X1 X2 X3 X4 Y  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 33 53 3.32 3.42 29  
## 2 2 31 36 3.1 3.26 24  
## 3 3 33 51 3.18 3.18 26  
## 4 4 37 51 3.39 3.08 22  
## 5 5 36 54 3.2 3.41 27  
## 6 6 35 35 3.03 3.03 21  
## 7 7 59 56 4.78 4.57 33  
## 8 8 60 60 4.72 4.72 34  
## 9 9 59 60 4.6 4.41 32  
## 10 10 60 60 4.53 4.53 34  
## # ... with 22 more rows

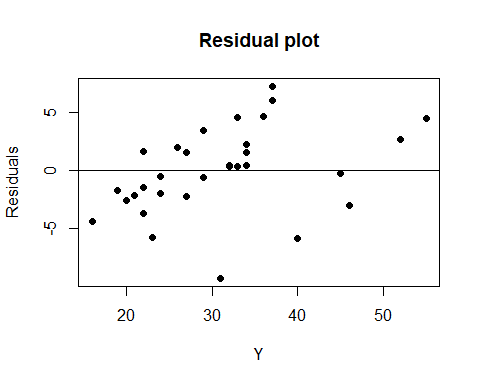
## Residual values

df.res

## 1 2 3 4 5 6 7   
## 3.4808267 -0.5267171 1.9695110 -1.4102039 1.5428552 -2.1000613 0.3475479   
## 8 9 10 11 12 13 14   
## 0.4171202 0.3400041 1.5956620 -2.6038332 4.6501466 2.2779756 -5.7446559   
## 15 16 17 18 19 20 21   
## -1.9533729 0.4640611 -5.8645250 -2.9659506 4.4833366 2.6618783 -0.6130550   
## 22 23 24 25 26 27 28   
## -3.7052588 -9.2819590 -0.2442399 6.0223177 7.2628879 4.5654867 -2.2408840   
## 29 30 31 32   
## 1.5956620 -1.6809494 -4.3708068 1.6291932

## Residual plot

plot(df$Y, df.res, ylab="Residuals", xlab="Y", main="Residual plot", pch=19)   
abline(0, 0)



mean(df.res)

## [1] -9.714451e-17

*So the mean is pretty close to zero. The data seems a little over fitted though.*

t\_df = filter(df[,5:6], X4 %in% c(3.45,4.02,5.8))  
t\_df

## # A tibble: 3 x 2  
## X4 Y  
## <dbl> <dbl>  
## 1 3.45 22  
## 2 5.8 31  
## 3 4.02 27

t\_df$pred\_Y = intercept + slope\*(t\_df$X4)  
t\_df

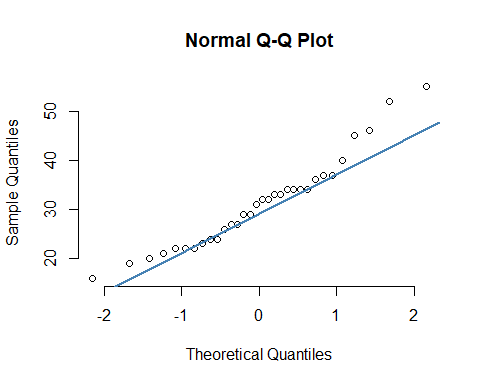
## # A tibble: 3 x 3  
## X4 Y pred\_Y  
## <dbl> <dbl> <dbl>  
## 1 3.45 22 25.7  
## 2 5.8 31 40.3  
## 3 4.02 27 29.2

t\_df$res = t\_df$Y-t\_df$pred\_Y  
t\_df

## # A tibble: 3 x 4  
## X4 Y pred\_Y res  
## <dbl> <dbl> <dbl> <dbl>  
## 1 3.45 22 25.7 -3.71  
## 2 5.8 31 40.3 -9.28  
## 3 4.02 27 29.2 -2.24

# Q-5

qqnorm(df$Y, pch = 1, frame = FALSE)  
qqline(df$Y, col = "steelblue", lwd = 2)



*It seems like a good fit to as it is close to a straight line. The distribution of Hydrocarbon escaping values is close to Normal*