ADAPT Program Day 4

# Loading packages

library(readxl)  
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(tidyr)  
library(ggplot2)  
library(reshape2)

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

library(stringr)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(data.table) #This is to use fread instead of read.csv (faster for big files)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':  
##   
## hour, isoweek, mday, minute, month, quarter, second, wday,  
## week, yday, year

## The following objects are masked from 'package:reshape2':  
##   
## dcast, melt

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

### Transforming variables

mtcars2 <- within(mtcars, {  
 vs <- factor(vs, labels = c("V", "S"))  
 am <- factor(am, labels = c("automatic", "manual"))  
 cyl <- ordered(cyl)  
 gear <- ordered(gear)  
 carb <- ordered(carb)  
 })  
  
str(mtcars2)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : Ord.factor w/ 3 levels "4"<"6"<"8": 2 2 1 2 3 2 3 1 1 2 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : Factor w/ 2 levels "V","S": 1 1 2 2 1 2 1 2 2 2 ...  
## $ am : Factor w/ 2 levels "automatic","manual": 2 2 2 1 1 1 1 1 1 1 ...  
## $ gear: Ord.factor w/ 3 levels "3"<"4"<"5": 2 2 2 1 1 1 1 2 2 2 ...  
## $ carb: Ord.factor w/ 6 levels "1"<"2"<"3"<"4"<..: 4 4 1 1 2 1 4 2 2 4 ...

# Hypothesis testing

## Testing normality

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

shapiro.test(mtcars2$mpg)

##   
## Shapiro-Wilk normality test  
##   
## data: mtcars2$mpg  
## W = 0.94756, p-value = 0.1229

#H0 is accepted: mpg is Normal

## Testing if variance are equal

leveneTest(mtcars2$mpg ~ mtcars2$am)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 1 4.1876 0.04957 \*  
## 30   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#H0 is rejected: Variances are not equal

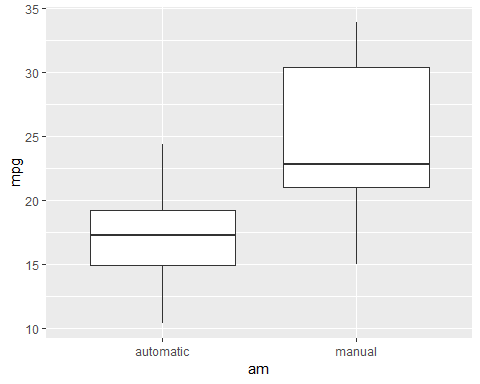
## T-Student test

#Null hypothesis: Mean MPG of automatic cars is the same of manual cars  
t.test(filter(mtcars2,am=="automatic") %>% select(mpg),  
 filter(mtcars2,am=="manual") %>% select(mpg),  
 alternative = "two.sided", var.equal = F)

##   
## Welch Two Sample t-test  
##   
## data: filter(mtcars2, am == "automatic") %>% select(mpg) and filter(mtcars2, am == "manual") %>% select(mpg)  
## t = -3.7671, df = 18.332, p-value = 0.001374  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -11.280194 -3.209684  
## sample estimates:  
## mean of x mean of y   
## 17.14737 24.39231

## A graphic to compare the distribution

ggplot(mtcars2,aes(am,mpg)) + geom\_boxplot()



# Making a regression

## Linear regression using gear type

lmodel <- lm(mtcars2$mpg ~ mtcars2$am)  
summary(lmodel)

##   
## Call:  
## lm(formula = mtcars2$mpg ~ mtcars2$am)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3923 -3.0923 -0.2974 3.2439 9.5077   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.147 1.125 15.247 1.13e-15 \*\*\*  
## mtcars2$ammanual 7.245 1.764 4.106 0.000285 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.902 on 30 degrees of freedom  
## Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385   
## F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

## Linear regression using weight

lmodel <- lm(mpg ~ wt, data=mtcars2)  
summary(lmodel)

##   
## Call:  
## lm(formula = mpg ~ wt, data = mtcars2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.5432 -2.3647 -0.1252 1.4096 6.8727   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.2851 1.8776 19.858 < 2e-16 \*\*\*  
## wt -5.3445 0.5591 -9.559 1.29e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.046 on 30 degrees of freedom  
## Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446   
## F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10

## Prediction for weight = 3.000lb

predict(lmodel, newdata = data.frame(wt=3.000))

## 1   
## 21.25171