Monish Kumar V – CE18B118

CE4720: Assignment 2

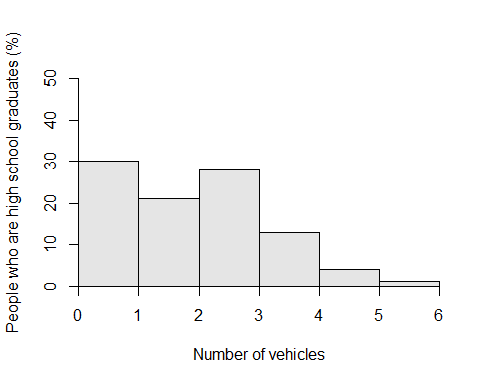
Question No.: 3

vehicle.count = c(0,1,2,3,4,5)  
highschool = c(30.2,21.8,28.4,13.7,4.4,1.5)  
college = c(47.9,19.4,22.7,8.0,1.5,0.5)  
  
# https://www.rforge.net/doc/packages/FSA/histFromSum.html  
require(FSA)

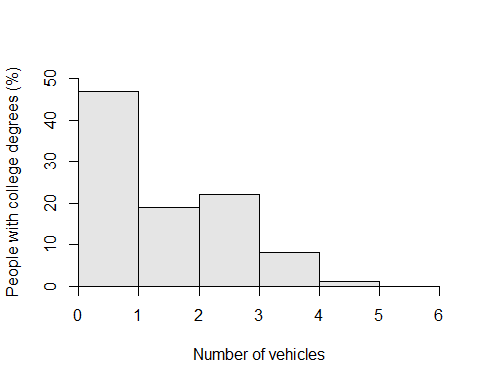
## Loading required package: FSA

## ## FSA v0.8.30. See citation('FSA') if used in publication.  
## ## Run fishR() for related website and fishR('IFAR') for related book.

histFromSum(highschool~vehicle.count, breaks = seq(0,6,1),  
 xlab='Number of vehicles',  
 ylab = 'People who are high school graduates (%)',  
 ylim = c(0,50))

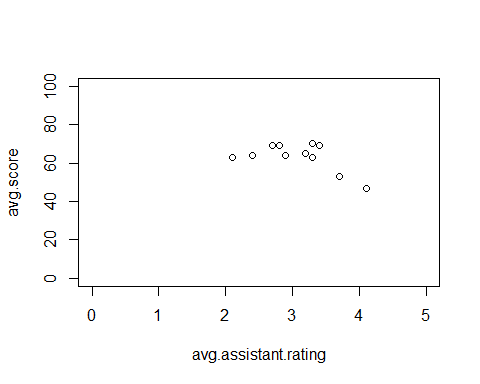


histFromSum(college~vehicle.count, breaks = seq(0,6,1),  
 xlab='Number of vehicles',  
 ylab = 'People with college degrees (%)',  
 ylim = c(0,50))

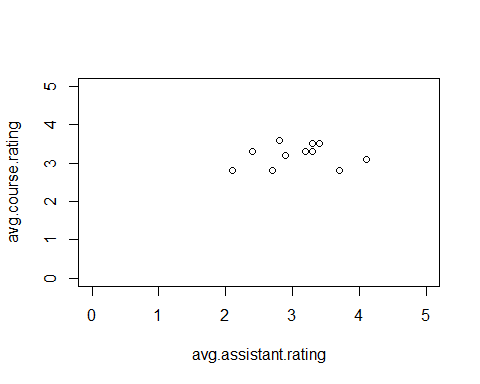


Question No.: 4

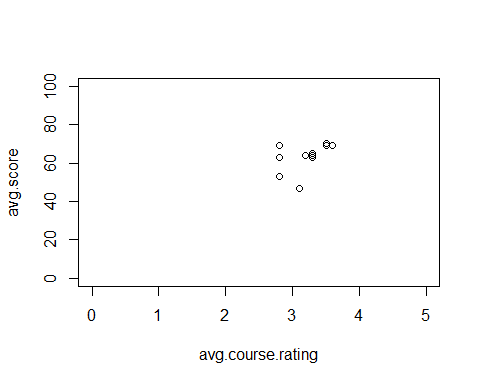
avg.assistant.rating = c(3.3,2.9,4.1,3.3,2.7,3.4,2.8,2.1,3.7,3.2,2.4)  
avg.course.rating = c(3.5,3.2,3.1,3.3,2.8,3.5,3.6,2.8,2.8,3.3,3.3)  
avg.score = c(70,64,47,63,69,69,69,63,53,65,64)  
  
# a)  
plot(avg.assistant.rating,avg.score, xlim = c(0,5), ylim = c(0,100))



# b)  
plot(avg.assistant.rating,avg.course.rating, xlim = c(0,5), ylim = c(0,5))



# c)  
plot(avg.course.rating,avg.score, xlim = c(0,5), ylim = c(0,100))

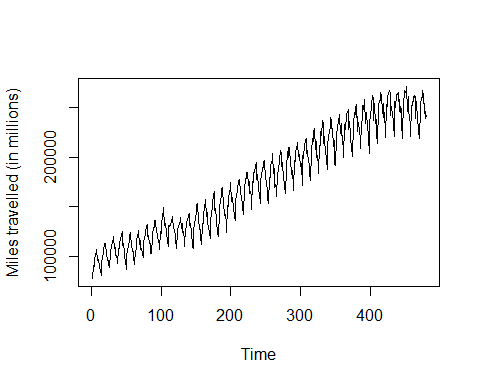


Question No.: 5

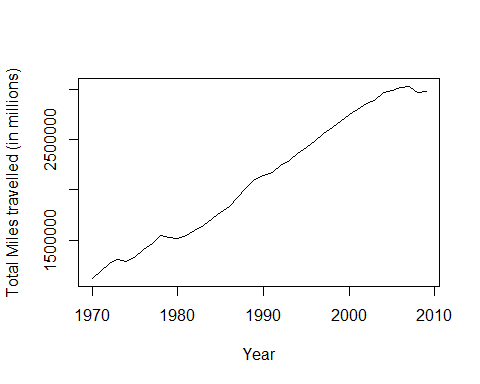
# a)  
getwd()

## [1] "C:/Users/Monish Kumar/R projects/L slot"

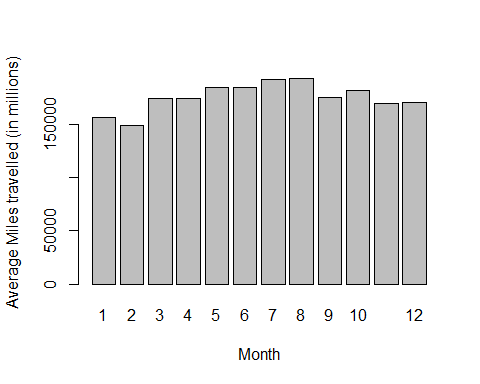
historic = read.csv('historicvmt.csv', header = TRUE)  
  
# b)  
plot.ts(historic$Total, ylab = 'Miles travelled (in millions)')



# c)  
yearly = aggregate(historic$Total, by = list(historic$Year),FUN='sum')  
plot(yearly, xlab = 'Year', ylab = 'Total Miles travelled (in millions)', type='l')  
lines(yearly, xlab = 'Year', ylab = 'Total Miles travelled (in millions)')



monthly = aggregate(historic$Total, by = list(historic$Month),FUN='mean')  
barplot(monthly$x,names.arg = monthly$Group.1,  
 xlab = 'Month', ylab = 'Average Miles travelled (in millions)')



# d)  
gasoline.price = read.table(file="clipboard")  
  
# e)  
m = c(1,1,1,1,1,1,1,1,1,1,1,1,0)  
gasoline.price$month = m  
  
na = 10000000  
  
a = gasoline.price$V1[(gasoline.price$V1 != na) & (gasoline.price$month == 1)]  
b = historic$Total[historic$Year >= 1978] # because 1976 and 1977 are null values  
cor(a,b)

## [1] 0.6747087

Question No.: 6

speed = c(38.4, 34.5, 25.6, 27.9, 30.2, 43, 42.3, 29.9, 41.5, 49.1, 30.2, 39.6, 28.5, 39.1, 40.3, 36.6, 34.9, 21.3, 42.5, 49.7, 37.9, 42.1, 19.5, 45.6, 54.2, 43.92, 43.35, 39.89, 40.86, 34.41, 47.22, 37.41, 44.26, 41.88, 40.57, 41.13, 32.47, 46.50, 41.51, 40.59, 39.40, 39.39, 36.25, 34.42, 41.68, 36.98, 41.06, 40.38)  
  
# custom function for 'mode'  
Mode <- function(x) {  
 ux <- unique(x)  
 ux[which.max(tabulate(match(x, ux)))]  
}  
  
summary(speed)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.50 34.80 40.09 38.54 42.15 54.20

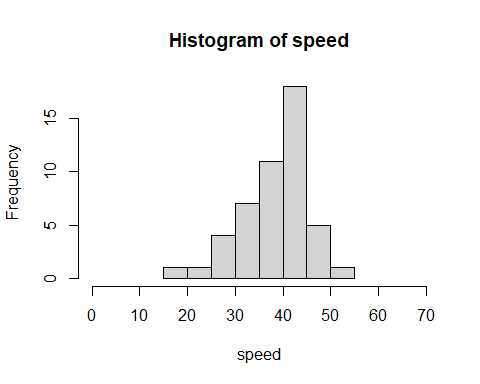
Mode(speed)

## [1] 30.2

var(speed) \*\* 0.5 # std deviation

## [1] 6.937628

hist(speed, xlim = c(0,70))

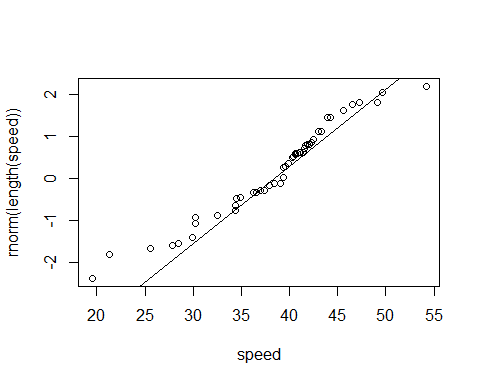


quantile(speed,0.85)

## 85%   
## 43.8915

set.seed(12345)  
qqplot(speed, rnorm(length(speed)))  
qqline(speed, rnorm(length(speed)))

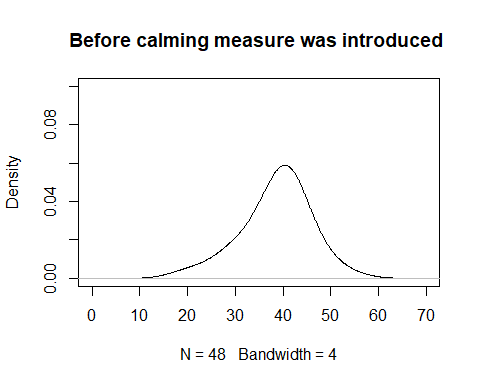
## Warning in if (datax) {: the condition has length > 1 and only the first element  
## will be used



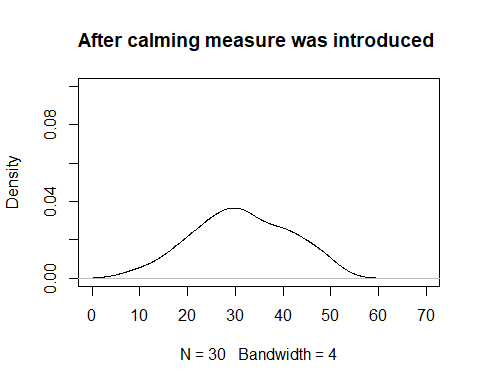
new.speed = c(40.97, 30.38, 26.82, 47.27, 33.78, 22.05, 25.41, 47.08, 17.75, 15.21, 27.46, 32.67, 9.72, 32.31, 33.84, 40.50, 19.38, 24.90, 49.12, 20.01, 38.61, 40.55, 24.09, 40.41, 30.30, 44.48, 31.21, 37.06, 26.33, 30.36)  
  
quantile(new.speed,0.85)

## 85%   
## 40.823

plot(density(speed, bw=4),main='Before calming measure was introduced ',  
 xlim = c(0,70), ylim = c(0,0.1))



plot(density(new.speed, bw=4),main='After calming measure was introduced ',  
 xlim = c(0,70), ylim = c(0,0.1))



t.test(x=speed, y=new.speed, alternative = 'greater')

##   
## Welch Two Sample t-test  
##   
## data: speed and new.speed  
## t = 3.4617, df = 46.492, p-value = 0.0005809  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 3.712385 Inf  
## sample estimates:  
## mean of x mean of y   
## 38.54021 31.33433

Question No.: 7

walking.speeds = read.table('Assignments/2/HW2/Speed.txt', sep = '\t')  
colnames(walking.speeds) = c("speed","age","luggage","gender")  
  
summary(aov(speed ~ age, data = walking.speeds))

## Df Sum Sq Mean Sq F value Pr(>F)   
## age 3 2.306 0.7686 21.1 3.74e-12 \*\*\*  
## Residuals 238 8.671 0.0364   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(aov(speed ~ luggage, data = walking.speeds))

## Df Sum Sq Mean Sq F value Pr(>F)   
## luggage 2 0.726 0.3629 8.462 0.000281 \*\*\*  
## Residuals 239 10.251 0.0429   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(aov(speed ~ gender, data = walking.speeds))

## Df Sum Sq Mean Sq F value Pr(>F)   
## gender 1 0.335 0.3351 7.558 0.00643 \*\*  
## Residuals 240 10.642 0.0443   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(aov(speed ~ age \* luggage, data = walking.speeds)) # no interaction effect

## Df Sum Sq Mean Sq F value Pr(>F)   
## age 3 2.306 0.7686 23.306 3.19e-13 \*\*\*  
## luggage 2 0.843 0.4217 12.786 5.40e-06 \*\*\*  
## age:luggage 4 0.176 0.0440 1.335 0.258   
## Residuals 232 7.651 0.0330   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(aov(speed ~ age \* gender, data = walking.speeds))

## Df Sum Sq Mean Sq F value Pr(>F)   
## age 3 2.306 0.7686 24.166 1.17e-13 \*\*\*  
## gender 1 0.377 0.3766 11.839 0.000687 \*\*\*  
## age:gender 3 0.852 0.2839 8.926 1.27e-05 \*\*\*  
## Residuals 234 7.443 0.0318   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(aov(speed ~ luggage \* gender, data = walking.speeds)) # no interaction effect

## Df Sum Sq Mean Sq F value Pr(>F)   
## luggage 2 0.726 0.3629 8.745 0.000217 \*\*\*  
## gender 1 0.359 0.3586 8.640 0.003614 \*\*   
## luggage:gender 2 0.097 0.0486 1.172 0.311615   
## Residuals 236 9.795 0.0415   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

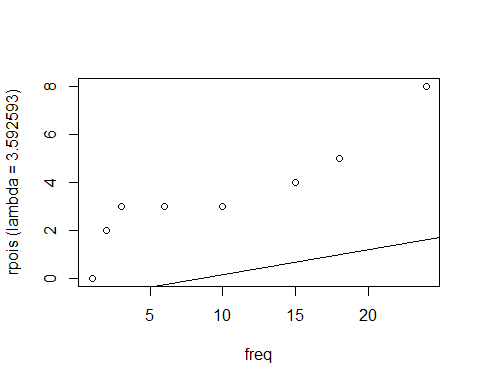
summary(aov(speed ~ age \*luggage \* gender, data = walking.speeds)) # no interaction effect

## Df Sum Sq Mean Sq F value Pr(>F)   
## age 3 2.306 0.7686 27.532 3.47e-15 \*\*\*  
## luggage 2 0.843 0.4217 15.104 7.05e-07 \*\*\*  
## gender 1 0.394 0.3940 14.113 0.00022 \*\*\*  
## age:luggage 4 0.195 0.0487 1.745 0.14106   
## age:gender 3 0.861 0.2869 10.276 2.31e-06 \*\*\*  
## luggage:gender 2 0.044 0.0219 0.785 0.45744   
## age:luggage:gender 3 0.108 0.0362 1.295 0.27682   
## Residuals 223 6.226 0.0279   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Question No.: 8

count = c(0,1,2,3,4,5,6,7,8)  
freq = c(2,3,15,18,24,10,6,2,1)  
  
pois.lambda = sum(count\*freq) / sum(freq)  
  
qqplot(freq,rpois(length(count),lambda = pois.lambda), ylab = 'rpois (lambda = 3.592593)')  
qqline(freq,rpois(length(count),lambda = pois.lambda))

## Warning in if (datax) {: the condition has length > 1 and only the first element  
## will be used



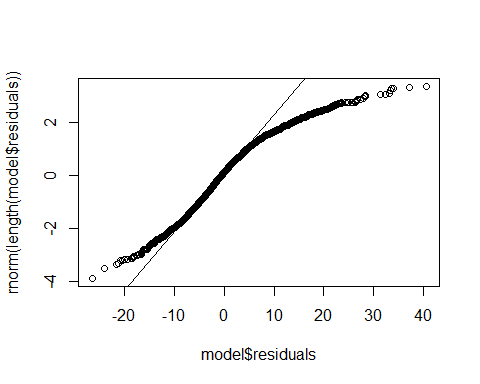
Question No.: 9

df = read.csv('Assignments/2/HW2/hhcomp.csv', header = TRUE)  
attach(df)  
  
# dummy variable  
RATIO.DRVEH = DRVRCNT/HHVEHCNT  
RATIO.DRVEH[is.na(RATIO.DRVEH)] = 0  
RATIO.DRVEH[is.infinite(RATIO.DRVEH)] = 0  
  
  
model = lm(CNTTDHH ~ RATIO.DRVEH + HHSIZE + DRVRCNT + TRAVDAY + HHFAMINC + HHNUMBIK)  
summary(model)

##   
## Call:  
## lm(formula = CNTTDHH ~ RATIO.DRVEH + HHSIZE + DRVRCNT + TRAVDAY +   
## HHFAMINC + HHNUMBIK)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.419 -3.309 -0.480 2.780 40.487   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.940711 0.202263 -4.651 3.35e-06 \*\*\*  
## RATIO.DRVEH 0.361156 0.145697 2.479 0.0132 \*   
## HHSIZE 2.855930 0.055181 51.755 < 2e-16 \*\*\*  
## DRVRCNT 0.488661 0.100326 4.871 1.13e-06 \*\*\*  
## TRAVDAY 0.145400 0.030412 4.781 1.77e-06 \*\*\*  
## HHFAMINC 0.075236 0.008038 9.360 < 2e-16 \*\*\*  
## HHNUMBIK 0.514434 0.056122 9.166 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.642 on 9776 degrees of freedom  
## Multiple R-squared: 0.3748, Adjusted R-squared: 0.3744   
## F-statistic: 976.6 on 6 and 9776 DF, p-value: < 2.2e-16

qqplot(model$residuals, rnorm(length(model$residuals)))  
qqline(model$residuals, rnorm(length(model$residuals)))

## Warning in if (datax) {: the condition has length > 1 and only the first element  
## will be used



#shapiro.test(model$residuals)  
ks.test(model$residuals, "pnorm", mean(model$residuals), sd(model$residuals))

## Warning in ks.test(model$residuals, "pnorm", mean(model$residuals),  
## sd(model$residuals)): ties should not be present for the Kolmogorov-Smirnov test

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: model$residuals  
## D = 0.065937, p-value < 2.2e-16  
## alternative hypothesis: two-sided