HW2\_6337\_MXB220061\_HXD220000\_2023

2023-02-20

Authors : Mankirat Singh Bharma MXB220061 - Harikrishna Dev HXD220000

Loading required libraries and cleaning environment

rm(list = ls())  
demo = T  
require(psych)

## Loading required package: psych

require(data.table)

## Loading required package: data.table

require(dplyr)

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

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

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

require(ggplot2)

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

if(demo) {setwd("~/Library/Mobile Documents/com~apple~CloudDocs/School Work/Sem 2/BUAN 6337/HW/pset2")}

Section - 1

1. The United States Geological Survey provides data on earthquakes of historical interest. Earthquakes.csv contains data about earthquakes with a magnitude greater than 2.5 in the United States and its territories. The variables are year, month, day, state, and magnitude.
2. California and Alaska are the two states with the highest number of earthquakes in the country. Read the data and create a new data set that includes only these two by filtering only the relevant rows.

ertqks <- fread("earthquakes.csv",header = T)[State %in% c("Alaska","California"),]  
if(demo) {str(ertqks)  
summary(ertqks)  
}

## Classes 'data.table' and 'data.frame': 169 obs. of 5 variables:  
## $ Year : int 1964 1965 1957 1938 1946 1899 2002 1996 1986 1899 ...  
## $ Month : int 3 2 3 11 4 9 11 6 5 9 ...  
## $ Day : int 28 4 9 10 1 10 3 10 7 4 ...  
## $ State : chr "Alaska" "Alaska" "Alaska" "Alaska" ...  
## $ Magnitude: num 9.2 8.7 8.6 8.2 8.1 8 7.9 7.9 7.9 7.9 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

## Year Month Day State   
## Min. :1812 Min. : 1.000 Min. : 1.00 Length:169   
## 1st Qu.:1933 1st Qu.: 4.000 1st Qu.: 8.00 Class :character   
## Median :1984 Median : 6.000 Median :17.00 Mode :character   
## Mean :1965 Mean : 6.343 Mean :16.42   
## 3rd Qu.:2003 3rd Qu.: 9.000 3rd Qu.:24.00   
## Max. :2010 Max. :12.000 Max. :31.00   
## NA's :1   
## Magnitude   
## Min. :3.000   
## 1st Qu.:5.400   
## Median :6.500   
## Mean :6.228   
## 3rd Qu.:7.100   
## Max. :9.200   
##

1. You are interested in the following statistics for the magnitude of earthquake:- Mean-Median-Standard Deviation-Minimum and maximum-25thand 75thpercentiles Create a table that shows the above statistics across different states within each year. In particular, your table must have years at the first column and it must break down the results across different states in the second column. In order to make the table short, further assume you are interested only in recent years and want to create a table that shows the desired statistics from 2002 to 2011

summary\_ertqks <- ertqks[,':='(mean=mean(Magnitude), median = median(Magnitude),std = sd(Magnitude),min = min(Magnitude),max=max(Magnitude),percentile.25 = quantile(Magnitude,0.25),percentile.75 = quantile(Magnitude,0.75)),by=c("Year","State")]  
  
if(demo) {summary(summary\_ertqks)}

## Year Month Day State   
## Min. :1812 Min. : 1.000 Min. : 1.00 Length:169   
## 1st Qu.:1933 1st Qu.: 4.000 1st Qu.: 8.00 Class :character   
## Median :1984 Median : 6.000 Median :17.00 Mode :character   
## Mean :1965 Mean : 6.343 Mean :16.42   
## 3rd Qu.:2003 3rd Qu.: 9.000 3rd Qu.:24.00   
## Max. :2010 Max. :12.000 Max. :31.00   
## NA's :1   
## Magnitude mean median std   
## Min. :3.000 Min. :4.000 Min. :3.900 Min. :0.0000   
## 1st Qu.:5.400 1st Qu.:5.450 1st Qu.:5.050 1st Qu.:0.4950   
## Median :6.500 Median :6.500 Median :6.500 Median :0.6229   
## Mean :6.228 Mean :6.228 Mean :6.163 Mean :0.6623   
## 3rd Qu.:7.100 3rd Qu.:7.100 3rd Qu.:7.000 3rd Qu.:0.8758   
## Max. :9.200 Max. :9.200 Max. :9.200 Max. :2.1213   
## NA's :60   
## min max percentile.25 percentile.75   
## Min. :3.000 Min. :4.50 Min. :3.50 Min. :4.450   
## 1st Qu.:5.000 1st Qu.:6.40 1st Qu.:5.00 1st Qu.:5.800   
## Median :6.200 Median :7.00 Median :6.40 Median :6.600   
## Mean :5.841 Mean :6.82 Mean :5.99 Mean :6.412   
## 3rd Qu.:7.000 3rd Qu.:7.30 3rd Qu.:7.00 3rd Qu.:7.100   
## Max. :9.200 Max. :9.20 Max. :9.20 Max. :9.200   
##

summary\_ertqks\_flt <- summary\_ertqks[Year>=2002 & Year<=2011,]  
summary\_ertqks\_flt <- summary\_ertqks\_flt[order(Year,State)]  
  
if(demo) {str(summary\_ertqks\_flt)  
 summary(summary\_ertqks\_flt)}

## Classes 'data.table' and 'data.frame': 61 obs. of 12 variables:  
## $ Year : int 2002 2002 2002 2002 2002 2002 2002 2002 2002 2003 ...  
## $ Month : int 11 10 2 6 5 9 3 11 12 11 ...  
## $ Day : int 3 23 6 17 14 3 16 24 24 17 ...  
## $ State : chr "Alaska" "Alaska" "Alaska" "California" ...  
## $ Magnitude : num 7.9 6.7 5.3 5.3 4.9 4.8 4.6 3.9 3.6 7.8 ...  
## $ mean : num 6.63 6.63 6.63 4.52 4.52 ...  
## $ median : num 6.7 6.7 6.7 4.7 4.7 4.7 4.7 4.7 4.7 7 ...  
## $ std : num 1.301 1.301 1.301 0.643 0.643 ...  
## $ min : num 5.3 5.3 5.3 3.6 3.6 3.6 3.6 3.6 3.6 6.6 ...  
## $ max : num 7.9 7.9 7.9 5.3 5.3 5.3 5.3 5.3 5.3 7.8 ...  
## $ percentile.25: num 6 6 6 4.08 4.08 ...  
## $ percentile.75: num 7.3 7.3 7.3 4.88 4.88 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

## Year Month Day State   
## Min. :2002 Min. : 1.000 Min. : 2.00 Length:61   
## 1st Qu.:2003 1st Qu.: 4.000 1st Qu.:10.00 Class :character   
## Median :2004 Median : 6.000 Median :17.00 Mode :character   
## Mean :2005 Mean : 6.344 Mean :17.33   
## 3rd Qu.:2007 3rd Qu.: 9.000 3rd Qu.:24.00   
## Max. :2010 Max. :12.000 Max. :31.00   
##   
## Magnitude mean median std   
## Min. :3.000 Min. :4.000 Min. :3.900 Min. :0.0000   
## 1st Qu.:4.100 1st Qu.:4.287 1st Qu.:4.000 1st Qu.:0.5621   
## Median :4.900 Median :4.740 Median :4.700 Median :0.6432   
## Mean :5.167 Mean :5.167 Mean :5.028 Mean :0.7709   
## 3rd Qu.:6.500 3rd Qu.:6.500 3rd Qu.:6.500 3rd Qu.:0.8758   
## Max. :7.900 Max. :7.100 Max. :7.000 Max. :2.1213   
## NA's :6   
## min max percentile.25 percentile.75   
## Min. :3.000 Min. :4.500 Min. :3.500 Min. :4.45   
## 1st Qu.:3.400 1st Qu.:5.500 1st Qu.:3.700 1st Qu.:4.65   
## Median :4.100 Median :6.600 Median :4.300 Median :5.20   
## Mean :4.464 Mean :6.316 Mean :4.736 Mean :5.53   
## 3rd Qu.:5.400 3rd Qu.:7.200 3rd Qu.:6.000 3rd Qu.:6.50   
## Max. :6.800 Max. :7.900 Max. :6.825 Max. :7.30   
##

1. Modify your R code in (b) such that the results for each year is shown in a separate table.

year <- unique(summary\_ertqks\_flt$Year)  
  
for(i in year) {  
 assign(paste0("ertqk",i),as.data.table(summary\_ertqks\_flt[Year ==i,]))  
 str(get(paste0("ertqk",i)))  
}

## Classes 'data.table' and 'data.frame': 9 obs. of 12 variables:  
## $ Year : int 2002 2002 2002 2002 2002 2002 2002 2002 2002  
## $ Month : int 11 10 2 6 5 9 3 11 12  
## $ Day : int 3 23 6 17 14 3 16 24 24  
## $ State : chr "Alaska" "Alaska" "Alaska" "California" ...  
## $ Magnitude : num 7.9 6.7 5.3 5.3 4.9 4.8 4.6 3.9 3.6  
## $ mean : num 6.63 6.63 6.63 4.52 4.52 ...  
## $ median : num 6.7 6.7 6.7 4.7 4.7 4.7 4.7 4.7 4.7  
## $ std : num 1.301 1.301 1.301 0.643 0.643 ...  
## $ min : num 5.3 5.3 5.3 3.6 3.6 3.6 3.6 3.6 3.6  
## $ max : num 7.9 7.9 7.9 5.3 5.3 5.3 5.3 5.3 5.3  
## $ percentile.25: num 6 6 6 4.08 4.08 ...  
## $ percentile.75: num 7.3 7.3 7.3 4.88 4.88 ...  
## - attr(\*, ".internal.selfref")=<externalptr>   
## Classes 'data.table' and 'data.frame': 19 obs. of 12 variables:  
## $ Year : int 2003 2003 2003 2003 2003 2003 2003 2003 2003 2003 ...  
## $ Month : int 11 3 6 2 12 8 2 1 3 5 ...  
## $ Day : int 17 17 23 19 22 15 22 25 11 25 ...  
## $ State : chr "Alaska" "Alaska" "Alaska" "Alaska" ...  
## $ Magnitude : num 7.8 7.1 6.9 6.6 6.6 5.3 5.2 4.7 4.6 4.2 ...  
## $ mean : num 7.1 7.1 7.1 7.1 4.29 ...  
## $ median : num 7 7 7 7 4 4 4 4 4 4 ...  
## $ std : num 0.51 0.51 0.51 0.51 0.876 ...  
## $ min : num 6.6 6.6 6.6 6.6 3.4 3.4 3.4 3.4 3.4 3.4 ...  
## $ max : num 7.8 7.8 7.8 7.8 6.6 6.6 6.6 6.6 6.6 6.6 ...  
## $ percentile.25: num 6.83 6.83 6.83 6.83 3.7 ...  
## $ percentile.75: num 7.28 7.28 7.28 7.28 4.65 ...  
## - attr(\*, ".internal.selfref")=<externalptr>   
## Classes 'data.table' and 'data.frame': 3 obs. of 12 variables:  
## $ Year : int 2004 2004 2004  
## $ Month : int 6 9 5  
## $ Day : int 28 28 30  
## $ State : chr "Alaska" "California" "California"  
## $ Magnitude : num 6.8 6 3  
## $ mean : num 6.8 4.5 4.5  
## $ median : num 6.8 4.5 4.5  
## $ std : num NA 2.12 2.12  
## $ min : num 6.8 3 3  
## $ max : num 6.8 6 6  
## $ percentile.25: num 6.8 3.75 3.75  
## $ percentile.75: num 6.8 5.25 5.25  
## - attr(\*, ".internal.selfref")=<externalptr>   
## Classes 'data.table' and 'data.frame': 7 obs. of 12 variables:  
## $ Year : int 2005 2005 2005 2005 2005 2005 2005  
## $ Month : int 6 6 6 6 6 9 5  
## $ Day : int 14 15 17 12 16 22 6  
## $ State : chr "Alaska" "California" "California" "California" ...  
## $ Magnitude : num 6.8 7.2 6.6 5.2 4.9 4.7 4.1  
## $ mean : num 6.8 5.45 5.45 5.45 5.45 5.45 5.45  
## $ median : num 6.8 5.05 5.05 5.05 5.05 5.05 5.05  
## $ std : num NA 1.19 1.19 1.19 1.19 ...  
## $ min : num 6.8 4.1 4.1 4.1 4.1 4.1 4.1  
## $ max : num 6.8 7.2 7.2 7.2 7.2 7.2 7.2  
## $ percentile.25: num 6.8 4.75 4.75 4.75 4.75 4.75 4.75  
## $ percentile.75: num 6.8 6.25 6.25 6.25 6.25 6.25 6.25  
## - attr(\*, ".internal.selfref")=<externalptr>   
## Classes 'data.table' and 'data.frame': 2 obs. of 12 variables:  
## $ Year : int 2006 2006  
## $ Month : int 7 10  
## $ Day : int 27 20  
## $ State : chr "Alaska" "California"  
## $ Magnitude : num 4.8 4.5  
## $ mean : num 4.8 4.5  
## $ median : num 4.8 4.5  
## $ std : num NA NA  
## $ min : num 4.8 4.5  
## $ max : num 4.8 4.5  
## $ percentile.25: num 4.8 4.5  
## $ percentile.75: num 4.8 4.5  
## - attr(\*, ".internal.selfref")=<externalptr>   
## Classes 'data.table' and 'data.frame': 9 obs. of 12 variables:  
## $ Year : int 2007 2007 2007 2007 2007 2007 2007 2007 2007  
## $ Month : int 12 8 8 12 10 5 8 7 7  
## $ Day : int 19 2 15 26 31 9 9 2 20  
## $ State : chr "Alaska" "Alaska" "Alaska" "Alaska" ...  
## $ Magnitude : num 7.2 6.7 6.5 6.4 5.6 5.2 4.4 4.3 4.2  
## $ mean : num 6.7 6.7 6.7 6.7 4.74 4.74 4.74 4.74 4.74  
## $ median : num 6.6 6.6 6.6 6.6 4.4 4.4 4.4 4.4 4.4  
## $ std : num 0.356 0.356 0.356 0.356 0.623 ...  
## $ min : num 6.4 6.4 6.4 6.4 4.2 4.2 4.2 4.2 4.2  
## $ max : num 7.2 7.2 7.2 7.2 5.6 5.6 5.6 5.6 5.6  
## $ percentile.25: num 6.47 6.47 6.47 6.47 4.3 ...  
## $ percentile.75: num 6.83 6.83 6.83 6.83 5.2 ...  
## - attr(\*, ".internal.selfref")=<externalptr>   
## Classes 'data.table' and 'data.frame': 4 obs. of 12 variables:  
## $ Year : int 2008 2008 2008 2008  
## $ Month : int 5 4 7 4  
## $ Day : int 2 16 29 30  
## $ State : chr "Alaska" "Alaska" "California" "California"  
## $ Magnitude : num 6.6 6.6 5.5 5.4  
## $ mean : num 6.6 6.6 5.45 5.45  
## $ median : num 6.6 6.6 5.45 5.45  
## $ std : num 0 0 0.0707 0.0707  
## $ min : num 6.6 6.6 5.4 5.4  
## $ max : num 6.6 6.6 5.5 5.5  
## $ percentile.25: num 6.6 6.6 5.43 5.43  
## $ percentile.75: num 6.6 6.6 5.47 5.47  
## - attr(\*, ".internal.selfref")=<externalptr>   
## Classes 'data.table' and 'data.frame': 7 obs. of 12 variables:  
## $ Year : int 2009 2009 2009 2009 2009 2009 2009  
## $ Month : int 1 5 1 3 4 6 3  
## $ Day : int 24 18 9 30 30 8 8  
## $ State : chr "Alaska" "California" "California" "California" ...  
## $ Magnitude : num 5.8 4.7 4.5 4.3 3.5 3.5 3.5  
## $ mean : num 5.8 4 4 4 4 4 4  
## $ median : num 5.8 3.9 3.9 3.9 3.9 3.9 3.9  
## $ std : num NA 0.562 0.562 0.562 0.562 ...  
## $ min : num 5.8 3.5 3.5 3.5 3.5 3.5 3.5  
## $ max : num 5.8 4.7 4.7 4.7 4.7 4.7 4.7  
## $ percentile.25: num 5.8 3.5 3.5 3.5 3.5 3.5 3.5  
## $ percentile.75: num 5.8 4.45 4.45 4.45 4.45 4.45 4.45  
## - attr(\*, ".internal.selfref")=<externalptr>   
## Classes 'data.table' and 'data.frame': 1 obs. of 12 variables:  
## $ Year : int 2010  
## $ Month : int 1  
## $ Day : int 10  
## $ State : chr "California"  
## $ Magnitude : num 6.5  
## $ mean : num 6.5  
## $ median : num 6.5  
## $ std : num NA  
## $ min : num 6.5  
## $ max : num 6.5  
## $ percentile.25: num 6.5  
## $ percentile.75: num 6.5  
## - attr(\*, ".internal.selfref")=<externalptr>

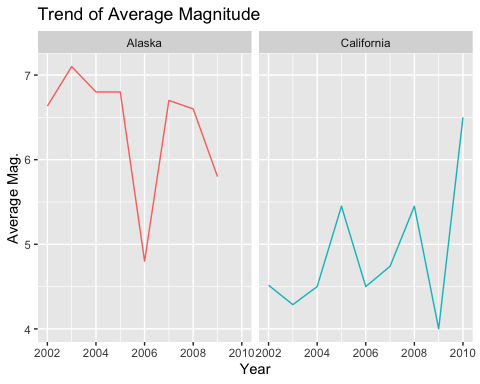
1. Now, assume you want to show the same results in part (b)but with the difference that years are shown is the first column and the states are shown in the top row.

ertqk\_summary <- dcast(summary\_ertqks\_flt,Year ~ State,fun = mean,value.var = c("mean", "median", "std", "min", "max","percentile.25","percentile.75"))  
if(demo) {str(ertqk\_summary)}

## Classes 'data.table' and 'data.frame': 9 obs. of 15 variables:  
## $ Year : int 2002 2003 2004 2005 2006 2007 2008 2009 2010  
## $ mean\_Alaska : num 6.63 7.1 6.8 6.8 4.8 ...  
## $ mean\_California : num 4.52 4.29 4.5 5.45 4.5 ...  
## $ median\_Alaska : num 6.7 7 6.8 6.8 4.8 6.6 6.6 5.8 NaN  
## $ median\_California : num 4.7 4 4.5 5.05 4.5 4.4 5.45 3.9 6.5  
## $ std\_Alaska : num 1.3 0.51 NA NA NA ...  
## $ std\_California : num 0.643 0.876 2.121 1.195 NA ...  
## $ min\_Alaska : num 5.3 6.6 6.8 6.8 4.8 6.4 6.6 5.8 NaN  
## $ min\_California : num 3.6 3.4 3 4.1 4.5 4.2 5.4 3.5 6.5  
## $ max\_Alaska : num 7.9 7.8 6.8 6.8 4.8 7.2 6.6 5.8 NaN  
## $ max\_California : num 5.3 6.6 6 7.2 4.5 5.6 5.5 4.7 6.5  
## $ percentile.25\_Alaska : num 6 6.83 6.8 6.8 4.8 ...  
## $ percentile.25\_California: num 4.08 3.7 3.75 4.75 4.5 ...  
## $ percentile.75\_Alaska : num 7.3 7.28 6.8 6.8 4.8 ...  
## $ percentile.75\_California: num 4.88 4.65 5.25 6.25 4.5 ...  
## - attr(\*, ".internal.selfref")=<externalptr>   
## - attr(\*, "sorted")= chr "Year"

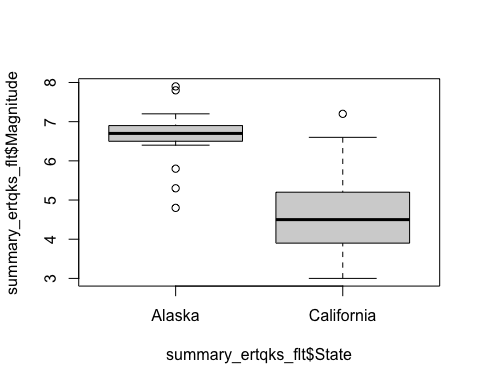
1. You are interested in how the magnitude of earthquakes is trending over time for each state. In one graph, plot two time series plots, side by side,which shows the trend of average magnitude of earthquakes over time for the two states

summary\_ertqks\_flt %>%  
ggplot(aes(Year,mean))+  
geom\_line(aes(colour = summary\_ertqks\_flt$State))+ labs(x = "Year" , y ="Average Mag." )+ facet\_wrap(~summary\_ertqks\_flt$State, nrow = 1)+ theme(legend.position = "none")+  
ggtitle("Trend of Average Magnitude")



1. Test the following null hypothesis: “the average magnitude of earthquakes in California is equal to that of Alaska”

boxplot(summary\_ertqks\_flt$Magnitude ~ summary\_ertqks\_flt$State)



H0: Avg Magnitude (California) = Avg Magnitude (Alaska)

H1: Avg Magnitude (California) <> Avg Magnitude (Alaska)

t.test(summary\_ertqks\_flt$Magnitude ~ summary\_ertqks\_flt$State)

##   
## Welch Two Sample t-test  
##   
## data: summary\_ertqks\_flt$Magnitude by summary\_ertqks\_flt$State  
## t = 8.4493, df = 36.554, p-value = 4.043e-10  
## alternative hypothesis: true difference in means between group Alaska and group California is not equal to 0  
## 95 percent confidence interval:  
## 1.528420 2.493237  
## sample estimates:  
## mean in group Alaska mean in group California   
## 6.617647 4.606818

Since p- value was less than 0.05, we reject the null hypothesis

Q. 2 Suppose that at a local university the study guidelines for the College of Science and Math are to study two to three hours per unit per week. The instructor of the class, Orientation to the Statistics Major, takes these guidelines very seriously. He asks students to record their study time each week, and at the end of the term he compares their average study time per week to their term GPA. “study\_gpa.csv” contains student identification information, orientation course-section number, number of units enrolled, average time studied, and term GPA.

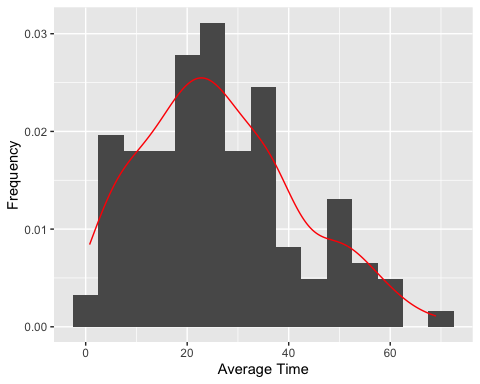
1. Graph the histogram for hours of study. Use the startpoint=0 and bandwidth=5.Also, overlaid to this graph, display the plots for the kernel density and the best fitting normal curve. Using an eyeballing approach, can we say the hours of study follows a normal distribution?(Hint: usegeom\_histogram() and geom\_density() in ggplot2)

gpa <- fread("study\_gpa.csv",header = T)  
  
if(demo) {str(gpa)}

## Classes 'data.table' and 'data.frame': 122 obs. of 7 variables:  
## $ ID : int 1005 1026 1045 1063 1071 1082 1096 1108 1120 1181 ...  
## $ FInitial: chr "J" "E" "R" "T" ...  
## $ LastName: chr "Bryant" "Fisher" "Turner" "Howard" ...  
## $ Section : int 2 2 2 1 1 2 2 2 2 2 ...  
## $ Units : int 10 18 19 9 14 12 19 11 16 12 ...  
## $ AveTime : num 21.4 10.4 48.4 18.3 49.7 ...  
## $ GPA : num 1.93 2.19 2.23 3.3 2.42 2.42 2.45 2.48 2.5 2.5 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

hist\_plot <- ggplot(gpa, aes(AveTime)) +  
geom\_histogram(aes(y=..density..), binwidth = 5) + # scale histogram y  
geom\_density(col = "red")  
print(hist\_plot + labs(x="Average Time",y = "Frequency"))

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(density)` instead.



Upon looking at the histogram, we can conclude that the distribution is positively(right) skewed distribution.

1. Check statistics of the average hours of study.

summary(gpa$AveTime)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.7729 15.1945 24.8211 26.3651 36.5434 69.0068

1. Conduct a hypothesis test to check whether there exists a significance correlation between units enrolled, hours of study and GPA for section 2. What is your conclusion? Doe correlation mean that one variable causes the other?

# H0: GPA   
  
var = names(gpa)[5:7]  
  
forms <- lapply(1:length(var),function(i) formula(paste(var[i], "~", paste(var[-i], collapse = "+"))))  
  
models <- lapply(forms,aov,data = gpa[Section == 2])  
  
for(i in 1:length(forms)) {  
 print(paste("Model:",forms[i]))  
 print(summary(models[[i]]))  
 }

## [1] "Model: Units ~ AveTime + GPA"  
## Df Sum Sq Mean Sq F value Pr(>F)   
## AveTime 1 71.0 71.03 8.261 0.00557 \*\*  
## GPA 1 1.5 1.49 0.173 0.67914   
## Residuals 61 524.5 8.60   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## [1] "Model: AveTime ~ Units + GPA"  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Units 1 1752 1751.7 8.361 0.00531 \*\*  
## GPA 1 191 190.8 0.911 0.34365   
## Residuals 61 12781 209.5   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## [1] "Model: GPA ~ Units + AveTime"  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Units 1 0.002 0.00174 0.008 0.927  
## AveTime 1 0.187 0.18718 0.911 0.344  
## Residuals 61 12.535 0.20549

We can see that GPA is highly correlated with Units enrolled and hours of study

Q.3 A study was conducted to see whether taking vitamin E daily would reduce the levels of atherosclerotic disease in a random sample of 500 individuals. Clinical measurements, including thickness of plaque of the carotid artery (taken via ultrasound), were recorded at baseline and at two subsequent visits in a data set “vite.csv”. Patients were divided into two strata according to their baseline plaque measurement.

1. First,read the data.The variable descriptions are as follows:

ID: individual identifier

Strata: 1=baseline plaque above 0.60mm+, 2=baseline plaque below 0.60mm Treatment: 0=placebo group, 1=vitamin E treatment

Plaque: Plaque measurement (mm)

HDL: HDL cholesterol (mg/DL)

LDL: LDL cholesterol (mg/DL)

Visit: 0=baseline, 1=first year, 2=second year

Trig: Triglycerides mg/DL

SBP: Systolic blood pressure (mm/Mg)

DBP: Diastolic blood pressure (mm/Mg)

Alcohol: # alcoholic drinks per day

Smoke: # cigarettes smoked per day

vite <- fread("vite.csv",header = T,na.strings = c("NA",""),sep = "auto",stringsAsFactors = F)  
  
if(demo) {str(vite)  
 summary(vite)}

## Classes 'data.table' and 'data.frame': 1500 obs. of 12 variables:  
## $ ID : int 1 1 1 2 2 2 3 3 3 4 ...  
## $ Visit : int 0 1 2 0 1 2 0 1 2 0 ...  
## $ Strata : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Treatment: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Plaque : num 0.807 0.758 0.81 0.758 0.687 ...  
## $ HDL : int 42 44 40 39 46 53 53 36 39 46 ...  
## $ LDL : int 127 143 158 138 147 161 133 146 119 139 ...  
## $ Trig : int 149 49 98 211 29 177 163 198 140 247 ...  
## $ SBP : int 106 131 136 157 154 65 169 172 140 142 ...  
## $ DBP : int 70 109 87 100 108 70 106 91 90 103 ...  
## $ Alcohol : int 1 2 2 1 2 3 1 2 3 0 ...  
## $ Smoke : int 0 0 0 6 6 6 0 0 0 0 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

## ID Visit Strata Treatment Plaque   
## Min. : 1.0 Min. :0 Min. :1.0 Min. :0.0 Min. :0.2209   
## 1st Qu.:125.8 1st Qu.:0 1st Qu.:1.0 1st Qu.:0.0 1st Qu.:0.4812   
## Median :250.5 Median :1 Median :1.5 Median :0.5 Median :0.5998   
## Mean :250.5 Mean :1 Mean :1.5 Mean :0.5 Mean :0.6329   
## 3rd Qu.:375.2 3rd Qu.:2 3rd Qu.:2.0 3rd Qu.:1.0 3rd Qu.:0.7830   
## Max. :500.0 Max. :2 Max. :2.0 Max. :1.0 Max. :1.0808   
## HDL LDL Trig SBP   
## Min. :22.00 Min. : 83.0 Min. : 25.0 Min. : 65.0   
## 1st Qu.:41.00 1st Qu.:126.0 1st Qu.:106.0 1st Qu.:123.0   
## Median :46.00 Median :136.0 Median :167.0 Median :142.0   
## Mean :45.87 Mean :135.5 Mean :173.6 Mean :141.9   
## 3rd Qu.:50.00 3rd Qu.:145.0 3rd Qu.:229.0 3rd Qu.:161.2   
## Max. :71.00 Max. :185.0 Max. :503.0 Max. :234.0   
## DBP Alcohol Smoke   
## Min. : 38.00 Min. :0.0000 Min. : 0.000   
## 1st Qu.: 84.00 1st Qu.:0.0000 1st Qu.: 0.000   
## Median : 92.00 Median :0.0000 Median : 0.000   
## Mean : 92.03 Mean :0.7287 Mean : 3.523   
## 3rd Qu.:101.00 3rd Qu.:1.0000 3rd Qu.: 5.000   
## Max. :138.00 Max. :7.0000 Max. :34.000

1. Note that the current data is in long format. We first want to transform the data to wide format so that we can conduct certain statistical analyses. Basically, long formats have repeated observations for a given person, whereas wide formats record those observations column-wise. Create the data so that plague values for each visit (0, 1, 2) are recorded in 3 separate columns as opposed to 3 rows, by ID and treatment. (cf. Reference: <https://data.library.virginia.edu/reshaping-data-from-wide-to-long/> although you may want to stick with the data.table() syntax and commands)

vite\_fnl <- dcast(vite,formula = ID+Treatment~Visit,value.var = "Plaque")  
  
colNames <- c("baseline","first\_year","second\_year")  
colnames(vite\_fnl)[3:5] <- colNames  
  
##vite\_0 <- vite\_fnl[vite\_fnl$Treatment == 0]  
  
if(demo) {str(vite\_fnl)}

## Classes 'data.table' and 'data.frame': 500 obs. of 5 variables:  
## $ ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Treatment : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ baseline : num 0.807 0.758 0.752 0.816 0.798 ...  
## $ first\_year : num 0.758 0.687 0.786 0.6 0.957 ...  
## $ second\_year: num 0.81 0.823 0.803 0.969 0.797 ...  
## - attr(\*, ".internal.selfref")=<externalptr>   
## - attr(\*, "sorted")= chr [1:2] "ID" "Treatment"

1. Assume there were no placebo group (i.e., treatment = 0) in your data set. Conduct a test to see whether there is a difference in plaque level before treatment and after the second visit? Interpret your results.

H0: Baseline = Second year H1: Baseline <> Second year

t.test(vite\_fnl$baseline[vite\_fnl$Treatment == 1],vite\_fnl$second\_year[vite\_fnl$Treatment == 1],paired = T)

##   
## Paired t-test  
##   
## data: vite\_fnl$baseline[vite\_fnl$Treatment == 1] and vite\_fnl$second\_year[vite\_fnl$Treatment == 1]  
## t = 3.9816, df = 249, p-value = 8.987e-05  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## 0.01504613 0.04450187  
## sample estimates:  
## mean difference   
## 0.029774

As p\_value < 0.05, we can rejected the null hypothesis.

1. Now, considering the fact that there is indeed a control group in your dataset, conduct a new test to check whether there is a difference in plaque level before treatment and after the second visit. Interpret your results

H0: treatment hasn’t effected the levels of plaque H1: treatment has effected the levels of plaque

vite\_fnl$diff <-vite\_fnl$second\_year - vite\_fnl$baseline  
t.test(vite\_fnl$diff[vite\_fnl$Treatment == 1],vite\_fnl$diff[vite\_fnl$Treatment == 0], paired = T)

##   
## Paired t-test  
##   
## data: vite\_fnl$diff[vite\_fnl$Treatment == 1] and vite\_fnl$diff[vite\_fnl$Treatment == 0]  
## t = -1.6986, df = 249, p-value = 0.09065  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## -0.036300591 0.002681391  
## sample estimates:  
## mean difference   
## -0.0168096

As p\_value > 0.05, H0 cannot be rejected.

1. Which of the tests in part (c) and (d) is more reliable? Explain.

Test - 2 helps us understand the effect of Vitamin E between the control and test group. This would give us a better idea on actual impact of the experiment.

1. One of the critical factors in randomizing the subjects in control and treatment groups is to make sure that the subjects are perfectly randomized in all aspects. Using the last two columns (i.e., alcohol and cigarette usage) of the original (long format) data, conduct two tests to check whether subjects are randomized perfectly. If they are perfectly randomized, then we should not expect much difference in alcohol (or cigarette) consumption for control vs. treatment groups.

H0: Distribution of alcohol consumers is randomized (Mean of treatment is same between both groups) H1: Distribution of alcohol consumers is not randomized (Mean of treatment is not same between both groups)

t.test(data = vite, Alcohol~Treatment)

##   
## Welch Two Sample t-test  
##   
## data: Alcohol by Treatment  
## t = 2.5104, df = 1488.5, p-value = 0.01216  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## 0.03702277 0.30164389  
## sample estimates:  
## mean in group 0 mean in group 1   
## 0.8133333 0.6440000

As p\_value < 0.05, we can reject null hypothesis.

H0: Distribution of smokers is randomized (Mean of treatment is same between both groups) H1: Distribution of smokers is not randomized (Mean of treatment is not same between both groups)

t.test(data = vite, Smoke~Treatment)

##   
## Welch Two Sample t-test  
##   
## data: Smoke by Treatment  
## t = 5.4701, df = 1365.5, p-value = 5.344e-08  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## 1.150199 2.436468  
## sample estimates:  
## mean in group 0 mean in group 1   
## 4.420000 2.626667

As p\_value < 0.05, we can reject null hypothesis.