1. This question uses the data HW01pb1data.csv. Download it to your computer.
2. Read in the data in R using

myData<-read.csv("HW01pb1data.csv",header=FALSE). Note, you first need to specify your working directory using the setwd() command. Determine whether each of the attributes (columns) are treated as qualitative (categorical) or quantitative (numeric) using R. Explain how you can tell using R.

> setwd(C:/Users/Som/ITU/RClass")

> myData <- read.csv(“HW01pb1data.csv", header=FALSE)

> View(MyData)

> attributes\_type<- c()

> for(i in 1:5){

+ attributes\_type[i]<-class(myData[,i])

+ }

> attributes\_type

[1] "integer" "integer" "integer" "factor" "factor"

Analyst: The first three attributes are quantitative value because its attribute type is integer. The rest is qualitative value because its attribute type is factor.

1. What is the specific problem that causes two of these attributes to be read in as qualitative (categorical) when it seems it should be quantitative (numeric)?

> for (i in 1:5){

+ if(!is.null(levels(myData[,i])))

+ sprintf("%d th column:/n",i)

+ print(levels(myData[,i]))

+ }

NULL

NULL

NULL

[1] "0" "10" "100" "110"

[5] "120" "140" "15" "150"

[9] "160" "20" "200" "25"

[13] "30" "35" "40" "5"

[17] "50" "55" "60" "65"

[21] "70" "80" "85" "90"

[25] "thirty five"

[1] "0" "10" "120" "140"

[5] "15" "20" "25" "255"

[9] "30" "35" "40" "45"

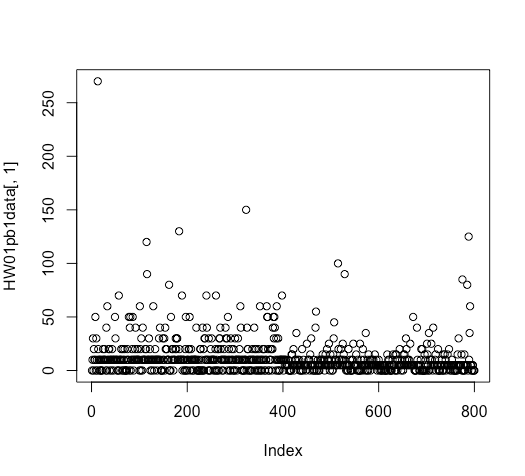
[13] "5" "50" "55" "60"

[17] "70" "80" "twenty five"

Analyst: The specific problem that column 4th and 5th have the factors.

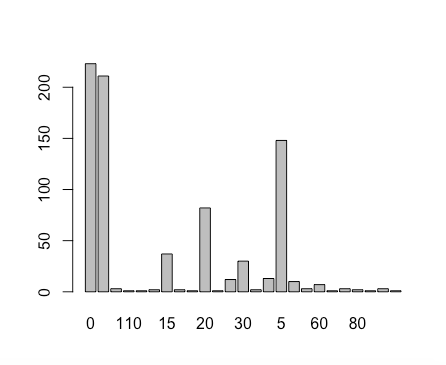
1. Use the command plot() in R to make a plot for column 1 by entering plot(myData[,1]). Use a similar command to plot column 4 (that is plot(myData[,4])). Because one variable is read in as quantitative (numeric) and the other as qualitative (categorical) these two plots are showing completely different things by default. Explain exactly what is being plotted in each case. Include these plots in your homework.

> plot(myData[,1])



Analyst: The command to plot column 1 scatters data on 1 x-y axis. It uses index number as a x-y values.

> plot(myData[,4])



Analyst: The command to plot column 4 a histogram graph. It uses factors to count how many elements are in each factor.

1. (optional) Read the data into Excel. Excel should have no problem opening the file directly since it is .csv. Create a new column that is equal to the forth column plus 10. What is the result for the problem observations (rows) you identified in part b? What specific outcome does Excel display?

A new sell in column 6th will get new value when a sell has a number as its value. If a sell has a string as its value such as “Fifty six” a new sell is not able to receive any value, and returns error “The operation “+” expected a number, date or duration, but cell D405 contain a string.”

1. This question uses the data in the file HW01pb2data.csv. Download it to your computer.

> getwd()

[1] "/Users/Som"

1. Read the data into R using

data<-read.csv("HW01pb2data.csv",header=FALSE). Note, you first need to specify your working directory using the setwd() command. Extract a simple random sample with replacement of 10,000 observations (rows). (Hint: R has a function called sample) Show your R commands for doing this.

> setwd(C:/Users/Som/ITU/RClass")

> data <- read.csv("HW01pb2data.csv",header=FALSE)

> sample\_dataset<-sample(data[,1],10000,replace=TRUE)

1. For your sample, use the functions mean(), max(), var() and quantile(,.25) to compute the mean, maximum, variance and 1st quartile respectively. Show your R code and the resulting values.

> mean(sample\_dataset)

[1] 9.462442

> max(sample\_dataset)

[1] 17.08133

> var(sample\_dataset)

[1] 3.966999

> quantile(sample\_dataset,.25)

25%

8.123561

c) Compute the same quantities in part b on the entire data set and show your answers. How much do they differ from your answers in part b?

> mean(data$V1)

[1] 9.451468

> max(data$V1)

[1] 18.96657

> var(data$V1)

[1] 4.001822

> quantile(data$V1,.25)

25%

8.10388

1. (Optional Part) Save your sample from R to a csv file using the command write.csv(). Then open this file with Excel and compute the mean, maximum, variance and 1st quartile. Provide the values and name the Excel functions you used to compute these.

> write.csv(sample\_dataset, file = "sample\_dataset.csv")

e) (Optional Part) Exactly what happens if you try to open the full data set with Excel?

Analyst: When I tried to open the data set, 65,565 objects out of 2,000,000 data are shown. Then, it might be because of version’s Excel limitation.

1. This question uses a sample of 2000 Ocean View house prices in the file HW01pb3OceanViewdata.csv and a sample of 5000 Desert house prices in the file HW01pb3Desertdata.csv. Download both data sets to your computer. Note that the house prices are in thousands of dollars. (Hint: look at the file MyFirstRLesson.r)

> setwd(C:/Users/Som/ITU/RClass")

> oceanViewData <- read.csv(“HW01pb3OceanViewdata.csv", header=FALSE)

> View(OceanViewData)

> desertData <- read.csv(“HW01pb3Desertdata.csv ", header=FALSE)

> View(desertData)

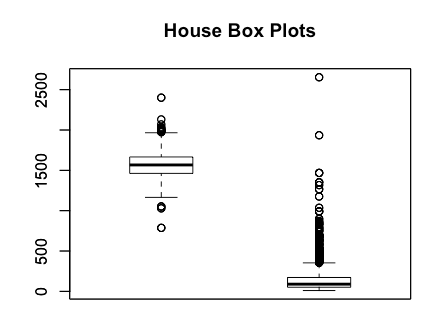
1. Use R to produce a single graph displaying a box plot for each set. Include the R commands and the plot. Put a name in the title of the plot (for example, main="House Box Plots"). Explain the box plot.

> class(oceanViewData)

[1] "data.frame"

> boxplot(oceanViewData, at = 1, xlim = c(0.5, 2.5), ylim = range(c(oceanViewData, desertData)), main = "House Box Plots")

> boxplot(desertData, at = 2, add = TRUE)



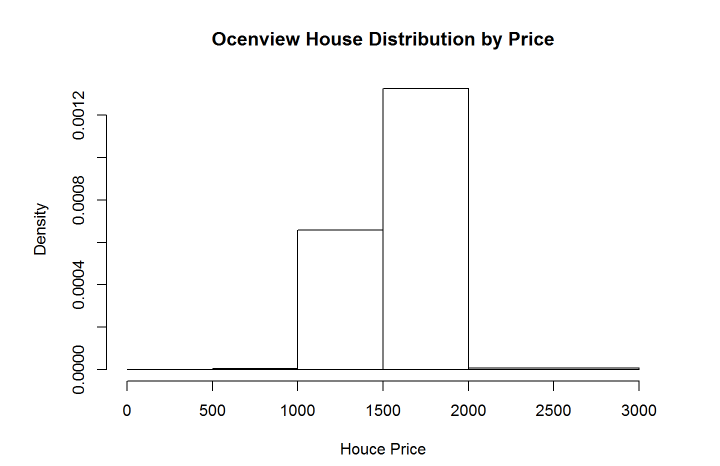
Analyst: The average of ocean view houses’ price are much higher than dessert houses’ price. The graph also presents that the ocean view houses data is almost symmetrically distributed. While, the desert houses data is skewed. It is more outliers in the desert than the ocean data set.

1. Use R to produce a frequency histogram for only the Ocean View house prices. Use intervals of width $250,000 beginning at 0 and ending at $3 million. Include the R commands and the plot. Create an appropriate title for the plot. (Hint: Use the hist R command)

> names(oceanViewData)[1] <- "HousePrice"

> breaks <- c(0, 500, 1000, 1500, 2000, 3000)

> hist(oceanViewData$HousePrice, breaks, main = "Ocenview House Distribution by Price",xlab = "Houce Price")



c) The empirical cumulative distribution function is described in the web site: http://en.wikipedia.org/wiki/ECDF Use R to plot the ECDF of the Ocean View houses. Make another ECDF plot of the Desert houses. Place both plots on the same page. Include legends. Include the R commands and the plot. Create a title for each plot.

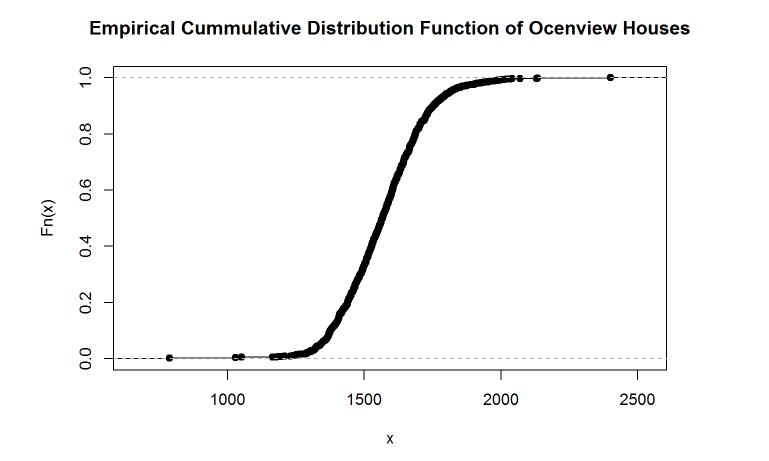
> ecdf(oceanViewData$HousePrice)

Empirical CDF

Call: ecdf(oceanViewData$HousePrice)

x[1:567] = 787, 1029, 1052, ..., 2133, 2401

> plot(ecdf(oceanViewData $HousePrice), main = "Empirical Cummulative Distribution Function of Ocenview Houses")

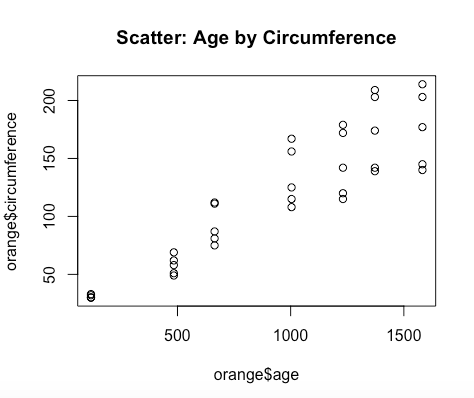


4) This question uses the Orange data set which is included in the R download. Type in the r command: orange <- as.data.frame(Orange). The data frame, orange, consists of three columns: Tree, age, and circumference.

1. Use plot() in R to make a scatter plot for this data with age on the x-axis and circumference on the y-axis. What range should be given for the x-axis? What about the y-axis range? Create an appropriate title for the plot. Include the R commands and the plot.

> orange <- as.data.frame(Orange)

> plot(orange$age,orange$circumference,main = "Scatter: Age by Circumference",xlim=c(min(orange$age),max(orange$age)),ylim=c(min(orange$circumference),max(orange$circumference)))



b) Compute the correlation between the age and circumference of the first tree in R using the function cor().

> cor(orange$age, orange$circumference)

[1] 0.9135189

1. For this problem you may want to use some the following R functions: names, merge, cov, and cor. Create a covariance - correlation chart which has the covariance and correlation of the age and circumference for each tree. Have your code print out the following chart with the same titles and the values filled in.

TREE COVARIANCE CORRELATION

1 1

2 2

3 3

4 4

5 5

Other functions that may be of interest are ddply, tapply, and subset. The ddply() function is in the ply package. Check out <http://had.co.nz/plyr>

> library(plyr)

> ddply(orange, .(Tree), summarize, cor = cor(age, circumference), cov = cov(age, circumference))

Tree cor cov

1 3 0.9881766 22239.83

2 1 0.9854675 22340.07

3 5 0.9877376 30442.81

4 2 0.9873624 34290.45

5 4 0.9844610 37062.62

> covarience<-data.frame(Tree=c(1:max(orange$Tree)),covarience=c(1))

> correlation<-data.frame(Tree=c(1:max(orange$Tree)),correlation=c(1))

>

> for(i in 1:nrow(correlation)){

+ covarience[i,2]<-cov(orange$age[orange$Tree==i],orange$circumference[orange$Tree==i])

+ correlation[i,2]<-cor(orange$age[orange$Tree==i],orange$circumference[orange$Tree==i])

+ }

> covacorr<-merge(covarience,correlation,"Tree")

> print(covacorr)

Tree covarience correlation

1 1 22340.07 0.9854675

2 2 34290.45 0.9873624

3 3 22239.83 0.9881766

4 4 37062.62 0.9844610

5 5 30442.81 0.9877376

5) This question uses the sample of 5,000 Desert Houses from problem three.

a) What is the median value? Is it larger or smaller than the mean?

> setwd(C:/Users/Som/ITU/RClass")

> desertData <- read.csv(“HW01pb3Desertdata.csv ", header=FALSE)

> View(desertData)

> names(desert)[1] <- "HousePrice"

> mean(desertData$HousePrice)

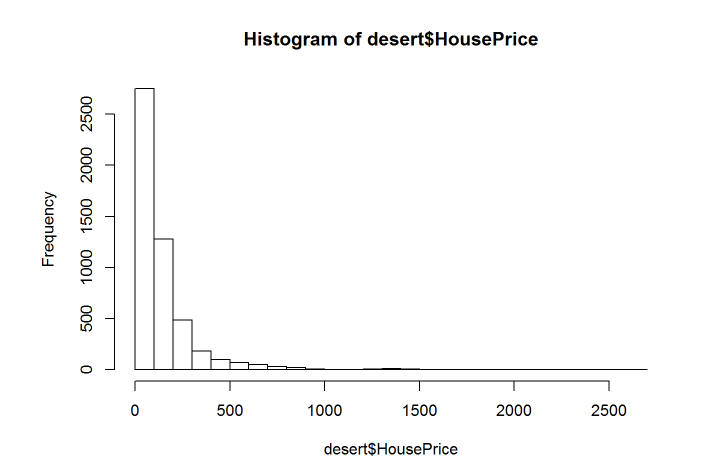
[1] 144.0348

> median(desertData$HousePrice)

[1] 89

1. What does your answer to part a) suggest about the shape of the distribution (rightskewed or left-skewed)? Does the distribution have more weight at one end? Is there a longer tail at the other? The distribution is skewed to the right if there is a long tail to the right. That is if the mean is greater than the median, the distribution is skewed to the right. A few high numbers will pull the mean above the median.

> hist(desertData$HousePrice, breaks = 20)



c) How does the median change if you add 10 (thousand dollars) to all the values?

> desertData <- read.csv("HW01pb3Desertdata.csv", header = FALSE)

> names(Desertdata)[1] <- "HousePrice"

> mean\_data<-mean(desertData $HousePrice)

> medium\_data<-median(desertData $HousePrice)

> mean\_dataplus<-mean\_data+10000

> medium\_dataplus<-median\_data+10000

1. How does the median change if you multiply all the values by 2?

> mean\_dataMulti <-mean\_data \* 2

> multi\_dataMulti <-medium\_data \* 2