Analysis 1_Baldridge

This assignment uses part of data from the Baldrige scoring process collected by the National Institute of Standards and Technology (NIST) for years 1990 to 2006. The Malcolm Baldrige National Quality Award (MBNQA) is considered as one of the most powerful catalysts of quality and organizational performance excellence in the United States, and it has had significant influence throughout the world. This data consists of performance evaluation scores of organizations from different sectors (variable "sector") like (1) manufacturing, (2) services, (3) small business, (4) education, (5) healthcare, and (6) nonprofit from 1990 to 2006. The dataset has scores on 7 categories that include (1) leadership; (2) strategic planning; (3) customer focus; (4) measurement, information and analysis; (5) workforce focus; (6) process management; and (7) results. In turn, these categories may have subcategories and the Criteria has sometimes added, discontinued or modified subcategories over time.

There is a variable "slnoskm17mar11" that simply provides a unique rowname to each observation. The variable "applicant" is applicant number within a particular year. There is another variable called "permanentid" that has an identifier for organizations to identify them uniquely over time if they appear in the data more than once.

Note that the scores are assigned by multiple volunteer examiners for each subcategory. We have individual (median) scores and consensus scores from the examiners. Prefix "i" in the column names tells us that this score is the median value of the all individual scores given by the individual examiners for the specified category and subcategories. Prefix "c" means that the score is a consensus scores received by an organization. In early years, the consensus score were provided only if the organization scored above a certain cut-off value.

1. Download the dataset "baldrige2011.xlsx" posted on Canvas. To answer the questions, please use R markdown to execute the R code and document it with appropriate comments and observations wherever it is required. Please use "stargazer" library for showing all the output tables and "ggplot2" for all graphs.

Pre-processing:

a) What are the number of observations, mean, median, standard deviation, min, maximum, and mode of iirtotal and ccrtotal in this data? Does it make more sense to use mean, median, or mode as a measure of central tendency for these two variables?

```
df$ccrtotal <- as.numeric(df$ccrtotal)

## Warning: NAs introduced by coercion

stargazer(df[c("iirtotal","ccrtotal")],type = "text")

##
## ===============================================================
```

```
## Statistic   N    Mean   St. Dev.   Min   Pctl(25) Pctl(75)   Max
## ------------------------------------------------------------------
## iirtotal  1,098 416.660 152.754  51.000  299.250  538.750  811.000
## ccrtotal   486  520.547 102.186  185.000 446.000  596.000  798.000
## ------------------------------------------------------------------
```

```r
median(df$iirtotal,na.rm=TRUE)
```

```
## [1] 429.5
```

```r
median(df$ccrtotal,na.rm =TRUE)
```

```
## [1] 532
```

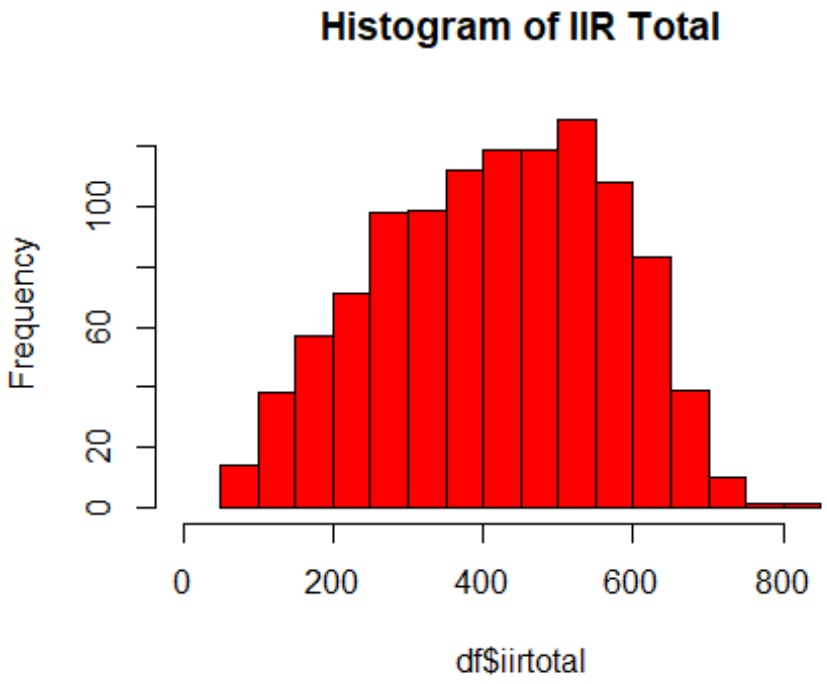
```r
#function for Mode
Mode = function(x){
  ta = table(x)
  tam = max(ta)
  if (all(ta == tam))
    mod = NA
  else
    if(is.numeric(x))
      mod = as.numeric(names(ta)[ta == tam])
  else
    mod = names(ta)[ta == tam]
  return(mod)
}

Mode(df$iirtotal)
```

```
## [1] 437
```

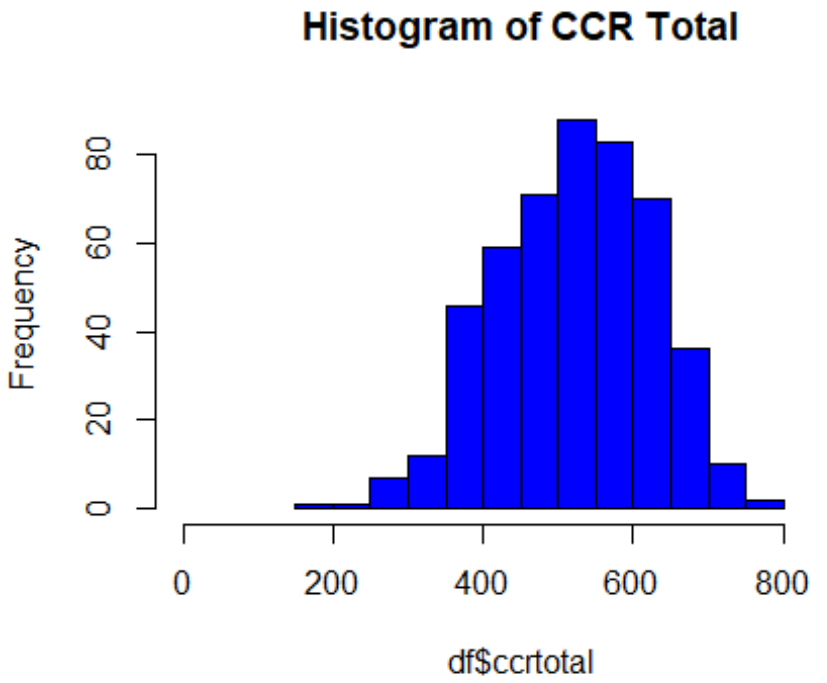
```r
Mode(df$ccrtotal)
```

```
## [1] 535
```

```r
hist(df$iirtotal,xlim=c(0,900),col = 'red',main = "Histogram of IIR Total")
```




```
hist(df$ccrtotal,xlim=c(0,900),col = 'blue',main = "Histogram of CCR Total")
```



We can see from the histograms of these 2 variables that the plots are left skewed due to the presence of

outliers on the lower end. Hence we can conclude median is the best measure of central tendency for these two variables.

b) List the mean, median, standard deviation, min and max of iirtotal and ccrtotal by Sector (make sure that you label sectors such that 1=mfg, 2=service, 3=small, 4=education, 5=health, 6=nonprofit and the output shows sector names and not numerals that denote the sector). Which sector has the highest variation in ccrtotal?

```
stargazer(subset(df[c("ccrtotal","iirtotal")],df$sector==1),
          title="Manufacturing Sector", type = "text",
          digits=2, median = TRUE, omit.summary.stat = c("p25", "p75","N"))

##
## Manufacturing Sector
## ================================================
## Statistic  Mean  St. Dev.  Min   Median  Max
## ------------------------------------------------
## ccrtotal  568.75  92.56   338.00 579.00 798.00
## iirtotal  476.96  152.92  58.00  494.00 811.00
## ------------------------------------------------

stargazer(subset(df[c("ccrtotal","iirtotal")], df$sector==2),
          title="Service Sector", type = "text",
          digits=2, median = TRUE, omit.summary.stat = c("p25", "p75","N"))

##
## Service Sector
## ================================================
## Statistic  Mean  St. Dev.  Min   Median  Max
## ------------------------------------------------
## ccrtotal  543.67  84.77   271.00 552.00 721.00
## iirtotal  466.55  144.06    81   509.5   723
## ------------------------------------------------

stargazer(subset(df[c("ccrtotal","iirtotal")], df$sector==3),
          title="Small Sector", type = "text",
          digits=2, median = TRUE, omit.summary.stat = c("p25", "p75","N"))

##
## Small Sector
## ================================================
## Statistic  Mean  St. Dev.  Min   Median  Max
## ------------------------------------------------
## ccrtotal  536.15  83.32   352.00 535.00 705.00
## iirtotal  353.81  153.45    51    332    716
## ------------------------------------------------

stargazer(subset(df[c("ccrtotal","iirtotal")], df$sector==4),
          title="Education Sector", type = "text",
          digits=2, median = TRUE, omit.summary.stat = c("p25", "p75","N"))
```

```
##
## Education Sector
## ================================================
## Statistic  Mean  St. Dev.  Min   Median  Max
## ------------------------------------------------
## ccrtotal  462.31  83.66   325.00 458.50 652.00
## iirtotal  381.42  127.26   105   392.5   588
## ------------------------------------------------
```

```
stargazer(subset(df[c("ccrtotal","iirtotal")], df$sector==5),
          title="Health Sector", type = "text",
          digits=2, median = TRUE, omit.summary.stat = c("p25", "p75","N"))
```

```
##
## Health Sector
## ================================================
## Statistic  Mean  St. Dev.  Min   Median  Max
## ------------------------------------------------
## ccrtotal  459.44  97.29   185.00 459.50 682.00
## iirtotal  414.20  129.34    89    432    694
## ------------------------------------------------
```

```
stargazer(subset(df[c("ccrtotal","iirtotal")], df$sector==6),
          title="Non-Profit Sector", type = "text",
          digits=2, median = TRUE, omit.summary.stat = c("p25", "p75","N"))
```

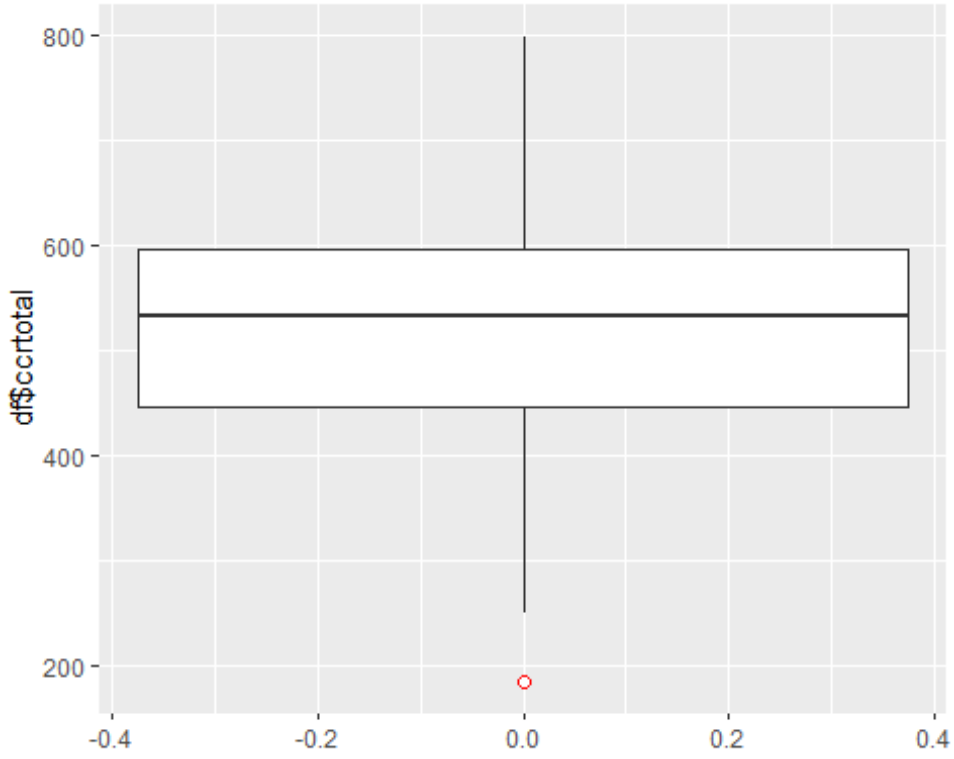
```
##
## Non-Profit Sector
## ========================================
## Statistic  Mean  St. Dev. Min Median Max
## ----------------------------------------
## ccrtotal  411.90  152.06  255 361.5  724
## iirtotal  454.30  117.55  316 438.5  706
## ----------------------------------------
```

We can see from the output that the non-profit sector has the highest variation in the CCRTotal score (sd = 152.06)

c) Identify the outliers in this data set in terms of ccrtotal using a box plot. How does the mean and standard deviation of ccrtotal change if the outliers are included versus excluded from the data set?

```
ggplot(df, aes(y=df$ccrtotal)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 21,
               outlier.size = 2)
```

```
## Warning: Removed 613 rows containing non-finite values (stat_boxplot).
```



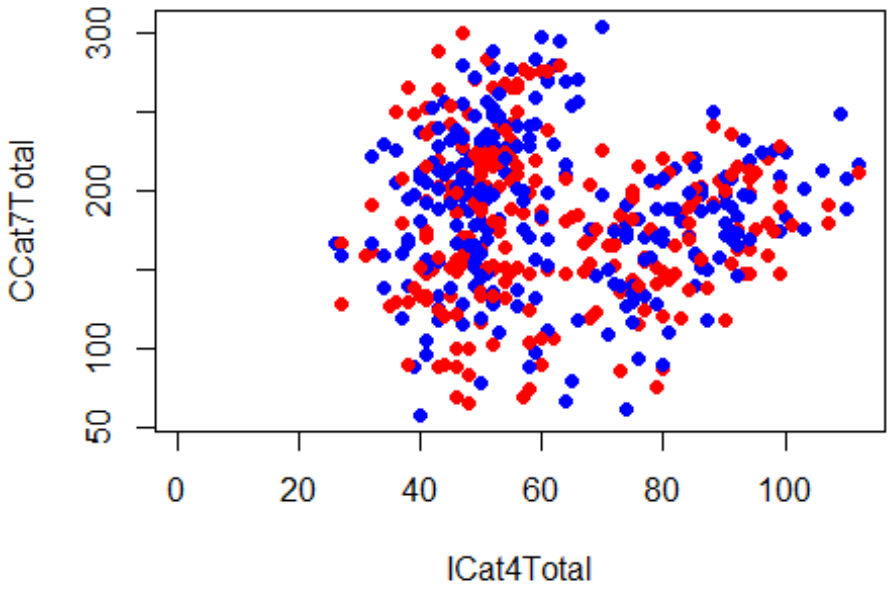

```
min(df$ccrtotal, na.rm = TRUE)

## [1] 185

which(df$ccrtotal == 185)

## [1] 506

df = df[-506,]

stargazer(df[c("iirtotal","ccrtotal")],type = "text")

##
## ===============================================================
## Statistic   N    Mean   St. Dev.   Min   Pctl(25) Pctl(75)   Max
## ---------------------------------------------------------------
## iirtotal  1,097 416.871 152.665  51.000  300.000  539.000  811.000
## ccrtotal   485  521.239 101.146  250.000 446.000  596.000  798.000
## ---------------------------------------------------------------
```

We can infer from the output that the removal of the outlier the mean (by 0.211) and SD (by -0.089) of the score improves slightly.

d) Draw a graph to represent the relationship between icat4total and ccat7total. Make a comment on the graph and how will you interpret it. Also compute the correlation coefficient between these two variables. What does this coefficient tell you about the

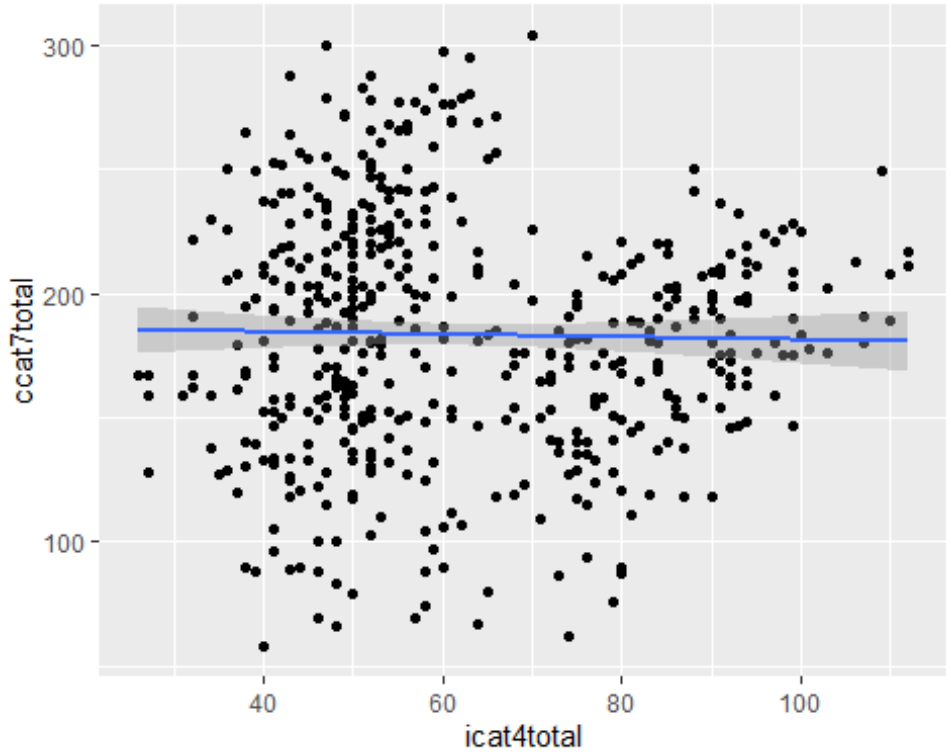
relationship between the variables? (Hint: correlation can't be calculated for 2 variables with different number of observations)

```
plot(df$icat4total,df$ccat7total, xlab = "ICat4Total",ylab =
"CCat7Total",pch=19, col=c("Red","blue"))

## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by
coercion
```




```
df$ccat7total <- as.numeric(df$ccat7total)

## Warning: NAs introduced by coercion

cor(df[c("icat4total","ccat7total")])  #put vars of interest in double quotes

##             icat4total ccat7total
## icat4total          1         NA
## ccat7total         NA          1

cor(x=df$icat4total,y=df$ccat7total,use="complete.obs")

## [1] -0.02093017

# convert ccat7total to numeric value
df$ccat7total <- as.numeric(df$ccat7total)
# creating a new dataframe with both variables of equal # of observations
z = subset(df, (!is.na(df$ccat7total)) & (!is.na(df$icat4total)))
# scatter plot showing the relationship between the variables with regression
line
```

```
ggplot(z, aes(x = icat4total, y = ccat7total)) +
  geom_point() +
  geom_smooth(method = lm)
```




```
cor(z$icat4total, z$ccat7total)

## [1] -0.02093017
```

From the graph, we interpret that there is no relationship between the 2 scores.

e) Provide a count of organizations that have appeared more than once in the data and identify them using their permanentid. Make a table that shows counts of organizations appearing more than once (i.e., twice, thrice, four times, etc).

```
tbl = as.data.frame(sort(table(df$permanentid), decreasing = TRUE))
stargazer(tbl, type = "text", title="Descriptive statistics", digits=2,
summary = FALSE)

##
## Descriptive statistics
## ==============
##     Var1   Freq
## --------------
## 1   B-031   9
## 2   B-030   8
## 3   B-071   7
## 4   B-118   7
## 5   B-015   6
```

```
## 6   B-025  6
## 7   B-135  6
## 8   B-189  6
## 9   B-026  5
## 10  B-037  5
## 11  B-054  5
## 12  B-095  5
## 13  B-097  5
## 14  B-099  5
## 15  B-147  5
## 16  B-173  5
## 17  B-186  5
## 18  B-001  4
## 19  B-062  4
## 20  B-065  4
## 21  B-074  4
## 22  B-085  4
## 23  B-120  4
## 24  B-122  4
## 25  B-149  4
## 26  B-159  4
## 27  B-165  4
## 28  B-170  4
## 29  B-180  4
## 30  B-184  4
## 31  B-198  4
## 32  B-002  3
## 33  B-004  3
## 34  B-005  3
## 35  B-008  3
## 36  B-016  3
## 37  B-017  3
## 38  B-023  3
## 39  B-024  3
## 40  B-034  3
## 41  B-035  3
## 42  B-048  3
## 43  B-055  3
## 44  B-059  3
## 45  B-060  3
## 46  B-064  3
## 47  B-073  3
## 48  B-076  3
## 49  B-081  3
## 50  B-084  3
## 51  B-088  3
## 52  B-107  3
## 53  B-109  3
## 54  B-124  3
## 55  B-125  3
```

```
## 56  B-128  3
## 57  B-130  3
## 58  B-134  3
## 59  B-137  3
## 60  B-141  3
## 61  B-142  3
## 62  B-143  3
## 63  B-146  3
## 64  B-148  3
## 65  B-150  3
## 66  B-153  3
## 67  B-156  3
## 68  B-157  3
## 69  B-161  3
## 70  B-163  3
## 71  B-171  3
## 72  B-179  3
## 73  B-187  3
## 74  B-188  3
## 75  B-190  3
## 76  B-195  3
## 77  B-003  2
## 78  B-006  2
## 79  B-007  2
## 80  B-009  2
## 81  B-010  2
## 82  B-011  2
## 83  B-012  2
## 84  B-013  2
## 85  B-014  2
## 86  B-018  2
## 87  B-019  2
## 88  B-020  2
## 89  B-021  2
## 90  B-022  2
## 91  B-027  2
## 92  B-028  2
## 93  B-029  2
## 94  B-032  2
## 95  B-033  2
## 96  B-036  2
## 97  B-038  2
## 98  B-039  2
## 99  B-040  2
## 100 B-041  2
## 101 B-042  2
## 102 B-043  2
## 103 B-044  2
## 104 B-045  2
## 105 B-046  2
```

```
## 106 B-047  2
## 107 B-049  2
## 108 B-050  2
## 109 B-051  2
## 110 B-052  2
## 111 B-053  2
## 112 B-056  2
## 113 B-057  2
## 114 B-058  2
## 115 B-061  2
## 116 B-063  2
## 117 B-066  2
## 118 B-067  2
## 119 B-068  2
## 120 B-069  2
## 121 B-070  2
## 122 B-072  2
## 123 B-075  2
## 124 B-077  2
## 125 B-078  2
## 126 B-079  2
## 127 B-080  2
## 128 B-082  2
## 129 B-083  2
## 130 B-086  2
## 131 B-087  2
## 132 B-089  2
## 133 B-090  2
## 134 B-091  2
## 135 B-092  2
## 136 B-093  2
## 137 B-094  2
## 138 B-096  2
## 139 B-098  2
## 140 B-100  2
## 141 B-101  2
## 142 B-102  2
## 143 B-103  2
## 144 B-104  2
## 145 B-105  2
## 146 B-106  2
## 147 B-108  2
## 148 B-110  2
## 149 B-111  2
## 150 B-112  2
## 151 B-113  2
## 152 B-114  2
## 153 B-115  2
## 154 B-116  2
## 155 B-117  2
```

```
## 156 B-119  2
## 157 B-121  2
## 158 B-123  2
## 159 B-126  2
## 160 B-127  2
## 161 B-129  2
## 162 B-131  2
## 163 B-132  2
## 164 B-133  2
## 165 B-136  2
## 166 B-138  2
## 167 B-139  2
## 168 B-140  2
## 169 B-144  2
## 170 B-145  2
## 171 B-151  2
## 172 B-152  2
## 173 B-154  2
## 174 B-155  2
## 175 B-158  2
## 176 B-160  2
## 177 B-162  2
## 178 B-164  2
## 179 B-166  2
## 180 B-167  2
## 181 B-168  2
## 182 B-169  2
## 183 B-172  2
## 184 B-174  2
## 185 B-176  2
## 186 B-177  2
## 187 B-178  2
## 188 B-181  2
## 189 B-182  2
## 190 B-183  2
## 191 B-185  2
## 192 B-191  2
## 193 B-192  2
## 194 B-193  2
## 195 B-194  2
## 196 B-196  2
## 197 B-197  2
## 198 B-199  2
## 199 B-200  2
## --------------
```

f) Some companies have been evaluated for 6 or more times in the data. Identify these companies and plot a line graph for any one of these companies that you found most interesting by looking over the trends in icat4total and ccat7total scores during the years they were evaluated. Explain why you found that company interesting and make

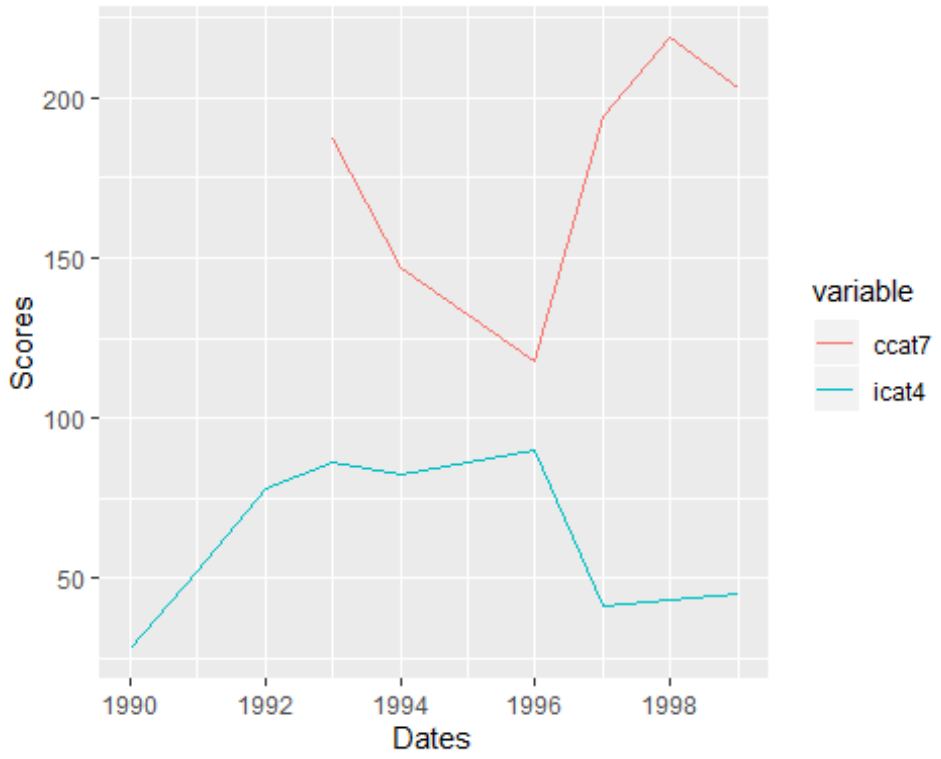
some conjectures about the relationship between icat4total and ccat7total based on what you observe.

```
# identifying companies who had more than 6 evaluations
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

sqldf('SELECT permanentid, count(permanentid)
      FROM df
      GROUP BY permanentid
      Having count(permanentid) >=6
      Order BY count(permanentid) DESC')

##   permanentid count(permanentid)
## 1       B-031                 9
## 2       B-030                 8
## 3       B-118                 7
## 4       B-071                 7
## 5       B-189                 6
## 6       B-135                 6
## 7       B-025                 6
## 8       B-015                 6

#creating a dataframe for companies with more than 6 evaluations
df1<-sqldf('SELECT permanentid, icat4total as score,year,"icat4" as variable
      FROM df
      WHERE permanentid in (SELECT permanentid
      FROM df
      GROUP BY permanentid
      Having count(permanentid) >=6
      Order BY count(permanentid) DESC)
      union
      SELECT permanentid, ccat7total as score, year, "ccat7"
      FROM df
      WHERE permanentid in (SELECT permanentid
      FROM df
      GROUP BY permanentid
      Having count(permanentid) >=6
      Order BY count(permanentid) DESC)')
#convert year to numeric
df1$year<- as.numeric(df1$year)
ggplot() +
  geom_line(data = df1[df1$permanentid =="B-031",], aes(year, score,
group=variable, color=variable))+
  geom_line(data = df1[df1$permanentid =="B-031",], aes(year, score,
group=variable, color=variable)) +
  xlab('Dates') +
```

```
  ylab('Scores') +
  scale_x_continuous(breaks=seq(1990,max(df1$year),2))
```

```
## Warning: Removed 3 rows containing missing values (geom_path).
```

```
## Warning: Removed 3 rows containing missing values (geom_path).
```



We can see from the plot of the company B-031 that there has been a significant difference in the scores from the year 1996 onwards. The consensus scores for results category has significantly increased while the individual score for the category education has decreased. It is interesting to note that as the score for education goes down, the score for results go up.

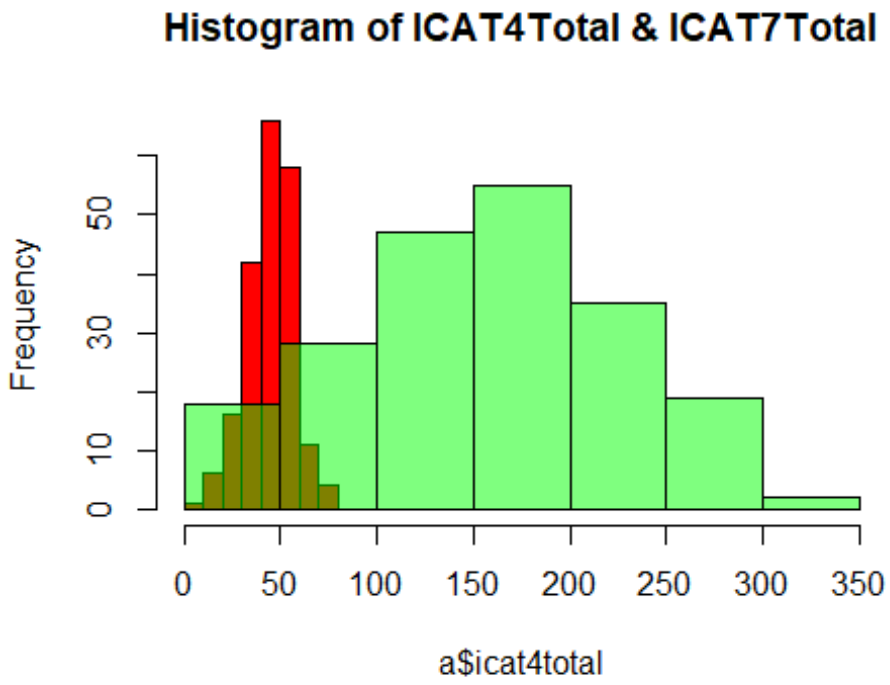
Part 2: Data Visualization

2. Download the dataset "baldrige2011.xlsx" posted on Canvas. To answer the questions, please use R markdown to execute the R code and document it with appropriate comments and observations wherever it is required. Please use "stargazer" library for showing all the output tables and "ggplot2" for all graphs.

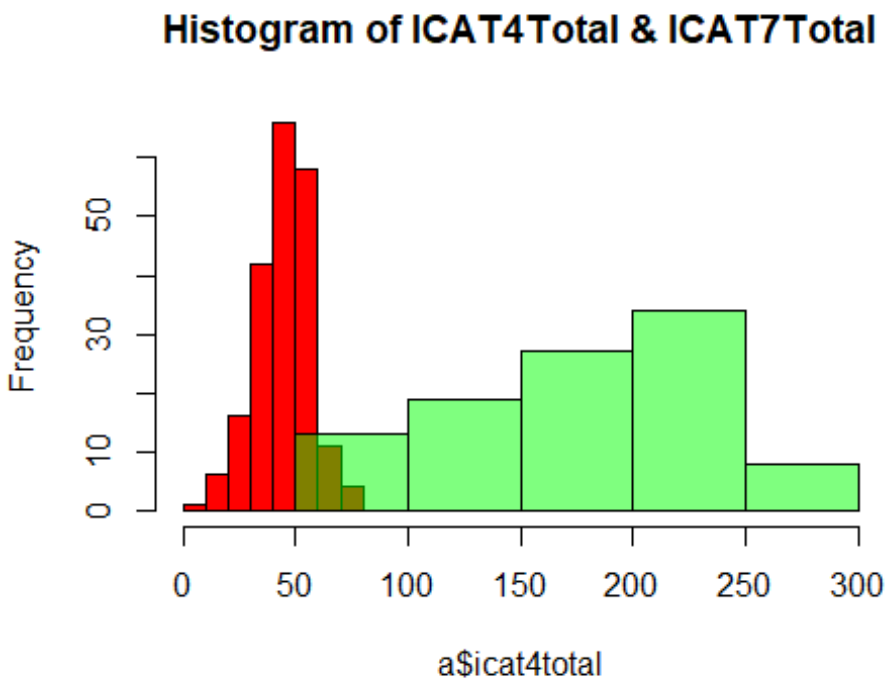
a) Investigate the distribution of icat4total and icat7total scores for healthcare sector with the help of a histogram plot. Attach the resulting graph, copy and paste the accompanying R code for computing this histogram, and list three key observations.

```
a = subset(df[c("icat4total","icat7total")],df$sector==5)

a$icat4total <- as.numeric(a$icat4total)
a$icat7total <- as.numeric(a$icat7total)
```

```r
hist(a$icat4total,xlim=c(0,350),col = 'red',main = "Histogram of ICAT4Total &
ICAT7Total")
hist(a$icat7total,xlim=c(0,350), add=T,col=rgb(0, 1, 0, 0.5))
```



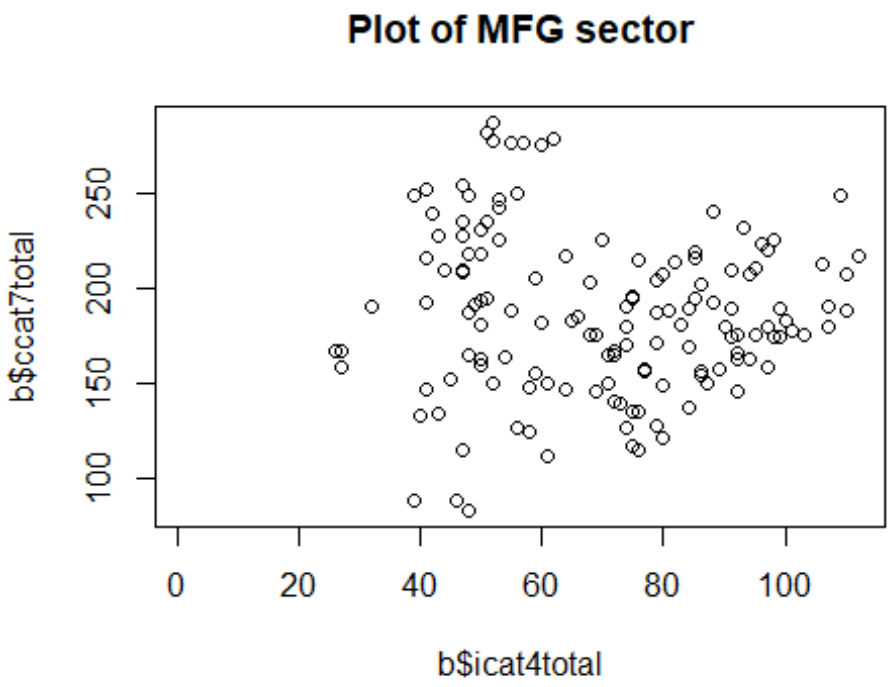

```r
a = subset(df[c("icat4total","ccat7total")],df$sector==5)

a$icat4total <- as.numeric(a$icat4total)
a$ccat7total <- as.numeric(a$ccat7total)

hist(a$icat4total,xlim=c(0,300),col = 'red',main = "Histogram of ICAT4Total &
ICAT7Total")
hist(a$ccat7total, add=T,col=rgb(0, 1, 0, 0.5))
```



The three key observations are: 1. Both the scores are fairly normally distributed. 2. The icat4total (0 to 350) scores are more spread compared to icat7total (0-80) scores. 3. The mean of icat7total is around 50 and that of icat4total is around 150-200.

b) Now, investigate the relationship between icat4total and ccat7total scores for those companies that have both the scores available and belong to manufacturing sector using a scatterplot. Attach the resulting graph, copy and paste the accompanying R code that you used to draw this scatterplot, and list two key observations about the relationship between the scores that you see from this plot

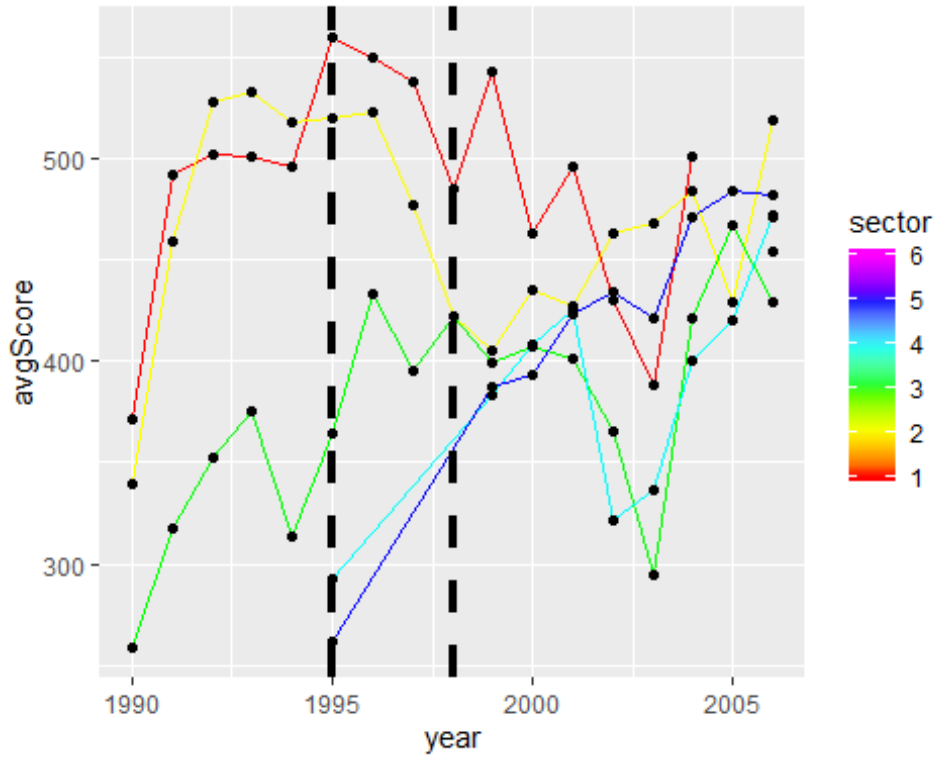
```
b = subset(df[c("icat4total","ccat7total","year")],df$sector==1)

plot(b$icat4total,b$ccat7total, main = "Plot of MFG sector")
```



The two key observations are: 1. There seems to be no correlation between these two scores. 2. We can infer no relationship from these 2 scores.

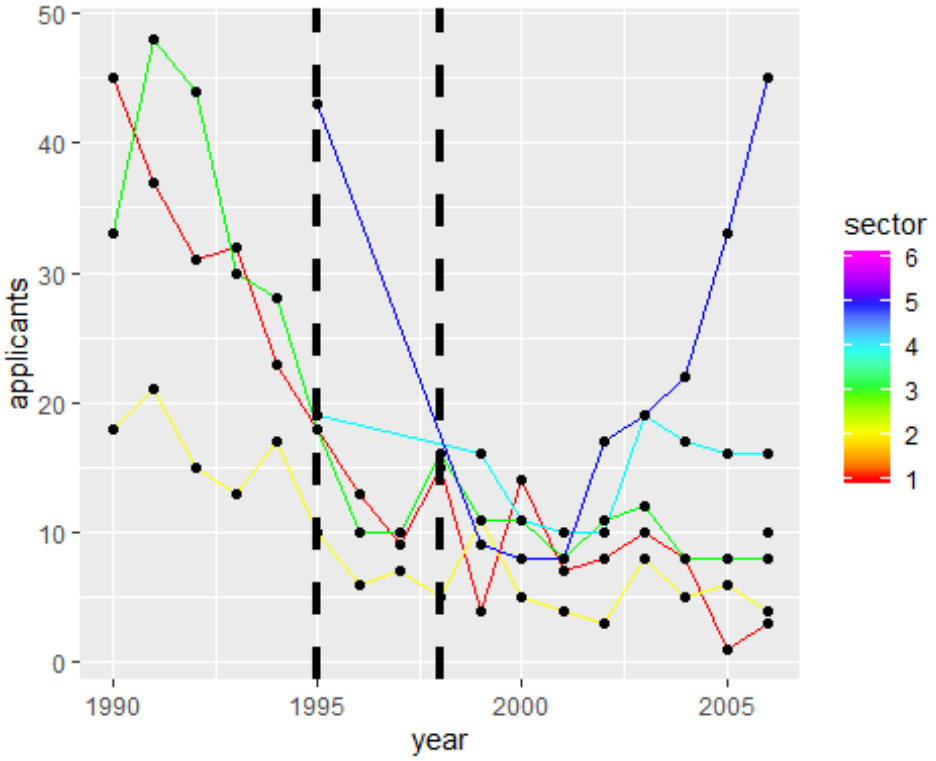
c) Plot a line graph to understand the trends in average iirtotal scores by sector. Also, draw vertical lines at years 1995 and 1998 to separate out and better visualize the trends in 3 different time periods (before 1995, 1995-1998, after 1998).

```
#creating a dataframe to understand the trends in average iirtotal scores by
sector
df4<-sqldf('
      SELECT year, sector, avg(iirtotal) as avgScore
      FROM df
      group by sector,year
      ')
df4$year = as.numeric(df4$year)
# Line plot with multiple groups
ggplot(data=df4, aes(x=year, y=avgScore, group=sector)) +
  geom_line(aes(color=sector))+
  geom_point()+
  scale_color_gradientn(colors =  rainbow(6))+
  geom_vline(xintercept = 1995, color="black", linetype="dashed", size =
1.5)+
  geom_vline(xintercept = 1998, color="black", linetype="dashed", size = 1.5)

## Warning: Removed 1 rows containing missing values (geom_point).
```



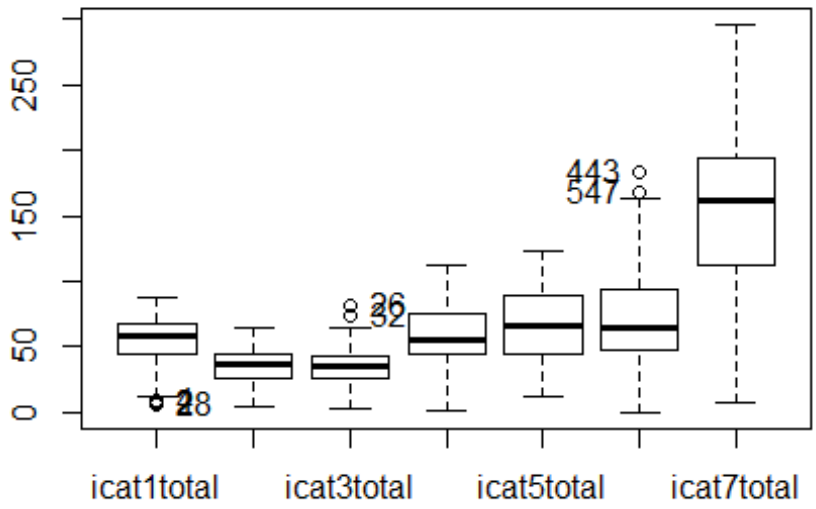
d) Plot a line graph to understand the trends in number of applicants by sector. Also, draw vertical lines at years 1995 and 1998 to separate out and better visualize the trends in 3 different time periods (before 1995, 1995-1998, after 1998).

```
df5<-sqldf('
      SELECT year, sector, count(applicant) as applicants
      FROM df
      group by sector,year
      ')
df5$year = as.numeric(df5$year)
# Line plot with multiple groups
ggplot(data=df5, aes(x=year, y=applicants, group=sector)) +
  geom_line(aes(color=sector))+
  geom_point()+
  scale_color_gradientn(colors =  rainbow(6))+
  geom_vline(xintercept = 1995, color="black", linetype="dashed", size =
1.5)+
  geom_vline(xintercept = 1998, color="black", linetype="dashed", size = 1.5)
```



e) Draw a box plot for all individual category totals (icat1total,icat2total….icat7total) in manufacturing sector and identify the scores that have outliers.

```
library(car)

## Loading required package: carData

mfg =
(subset(df[c("icat1total","icat2total","icat3total","icat4total","icat5total"
,"icat6total", "icat7total")],df$sector==1))

Boxplot(mfg[1:7], data = mfg)
```




```
## [1] "2"   "4"   "28"  "26"  "32"  "443" "547"

sqldf('SELECT count(distinct(permanentid)) as Total
      FROM df ')

##   Total
## 1   199

#table of organizations that appeared more than once
#Top 5
sqldf('SELECT permanentid, count(permanentid) as evaluations
      FROM df
      GROUP BY permanentid
      Having count(permanentid) >0
      ORDER BY count(permanentid) DESC
      LIMIT 5')

##   permanentid evaluations
## 1       B-031           9
## 2       B-030           8
## 3       B-118           7
## 4       B-071           7
## 5       B-189           6

#Bottom 5
sqldf('SELECT permanentid, count(permanentid) as evaluations
      FROM df
      GROUP BY permanentid
```

```
      Having count(permanentid) >0
      ORDER BY count(permanentid) ASC
      LIMIT 5')
```

```
##   permanentid evaluations
## 1       B-003           2
## 2       B-006           2
## 3       B-007           2
## 4       B-009           2
## 5       B-010           2
```

f) Finally, refine your analysis in part 2b by conditioning the scatterplot between icat4total and ccat7total score during 3 different time periods (upto 1994, 1995 to 1998, and 1999 onwards). To do this, install the R library "lattice" by typing install.packages("lattice"), then load it for your current session by typing library(lattice). Since we want to condition the scatterplot for different time periods, create a new variable, say, time_period for which

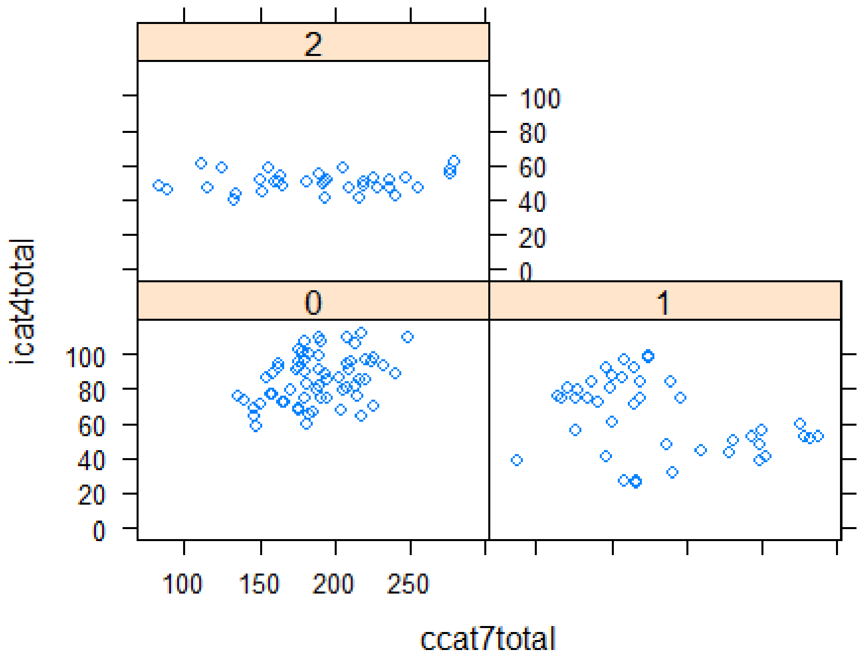
time_period = 0 if year <= 1994 and time_period= 1 if 1995 <= year <= 1998 time_period= 2 if year >= 1999

and compute a scatterplot between icat4total and icat7total score, conditional on time_period. Attach the resulting graph, copy and paste the accompanying R code for computing this scatterplot, and list three key observations about the relationships between advertising, income and sales that you see from this plot.

```
library(lattice)   # Load library into memory

b$time_period = 1

for (i in 1: nrow(b)) {
  if (b[i,'year'] <  1995) {
    b[i,'time_period'] = 0
  } else if (b[i,'year'] > 1998) {
    b[i,'time_period'] = 2
  }
}
b$time_period = factor(b$time_period)

b$time_period
```

```
##   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
1 1 1
## [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
## [223] 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2
## [260] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## Levels: 0 1 2

xyplot(icat4total ~ ccat7total | time_period, data=b)
```



Three key observations: 1. Scores in time_period range is tightly packed around the mean as the central tendency. 2. Scores in time_period 1999 and above, both the scores are tightly dependent on each other. 3. Scores in the time_period between 1995 to 1998, the scores are scattered all around with no apparent relationship.

g) Create a table of summary stats (N, mean, sd, min, max) for all individual category totals (icat1total, icat2total, ..., icat7total). Note: The output should be neatly formatted in a Table and the values should be rounded to 2 decimal places. Please use "stargazer" library for creating the stats table.

```
stargazer(subset(df[c("icat1total","icat2total","icat3total","icat4total","ic
at5total","icat6total", "icat7total")]),
          title="SUMMARY STATISTICS", type = "text",
          digits=2, median = TRUE, omit.summary.stat = c("p25", "p75"))

##
## SUMMARY STATISTICS
## ===================================================
```

```
## Statistic    N    Mean  St. Dev. Min  Median  Max
## ------------------------------------------------
## icat1total 1,097 51.46   18.66   3.00 54.00  91.00
## icat2total 1,097 32.54   13.01   2.00 34.00  64.00
## icat3total 1,097 32.90   13.66   0.00 33.00  82.00
## icat4total 1,097 50.08   20.68   1.00 48.00  112.00
## icat5total 1,097 51.56   22.35   7.00 46.00  123.00
## icat6total 1,097 54.85   30.94   0.00 48.00  183.00
## icat7total 1,097 143.48  66.78   4.00 143.00 308.00
## ------------------------------------------------
```

h) Create a table of pairwise Pearson correlation coefficients for all consensus category totals (ccat1total,ccat2total, …, ccat7total) also showing p-value for correlations. Note: The output should be neatly formatted in a Table and the values should be rounded to 2 decimal places.

```
library(Hmisc)

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

df$ccat1total <- as.numeric(df$ccat1total)

## Warning: NAs introduced by coercion

df$ccat2total <- as.numeric(df$ccat2total)

## Warning: NAs introduced by coercion

df$ccat3total <- as.numeric(df$ccat3total)

## Warning: NAs introduced by coercion

df$ccat4total <- as.numeric(df$ccat4total)

## Warning: NAs introduced by coercion

df$ccat5total <- as.numeric(df$ccat5total)

## Warning: NAs introduced by coercion

df$ccat6total <- as.numeric(df$ccat6total)

## Warning: NAs introduced by coercion
```

```r
df$ccat7total <- as.numeric(df$ccat7total)

c = subset(df,select = c(ccat1total,ccat2total,ccat3total,ccat4total,
                  ccat5total,ccat6total,ccat7total),na.rm=TRUE)

res <- round(cor(c, use = "pairwise", method = "pearson"),2)
res

##            ccat1total ccat2total ccat3total ccat4total ccat5total
ccat6total
## ccat1total       1.00       0.67       0.70       0.11       0.08      -
0.03
## ccat2total       0.67       1.00       0.55       0.26       0.25
0.19
## ccat3total       0.70       0.55       1.00      -0.04      -0.03      -
0.19
## ccat4total       0.11       0.26      -0.04       1.00       0.83
0.79
## ccat5total       0.08       0.25      -0.03       0.83       1.00
0.82
## ccat6total      -0.03       0.19      -0.19       0.79       0.82
1.00
## ccat7total       0.67       0.61       0.63       0.07       0.08      -
0.03
##            ccat7total
## ccat1total       0.67
## ccat2total       0.61
## ccat3total       0.63
## ccat4total       0.07
## ccat5total       0.08
## ccat6total      -0.03
## ccat7total       1.00

res2 <- rcorr(as.matrix(c))

res2

##            ccat1total ccat2total ccat3total ccat4total ccat5total
ccat6total
## ccat1total       1.00       0.67       0.70       0.11       0.08      -
0.03
## ccat2total       0.67       1.00       0.55       0.26       0.25
0.19
## ccat3total       0.70       0.55       1.00      -0.04      -0.03      -
0.19
## ccat4total       0.11       0.26      -0.04       1.00       0.83
0.79
## ccat5total       0.08       0.25      -0.03       0.83       1.00
0.82
## ccat6total      -0.03       0.19      -0.19       0.79       0.82
```

```
1.00
## ccat7total       0.67       0.61       0.63       0.07       0.08      -
0.03
##            ccat7total
## ccat1total       0.67
## ccat2total       0.61
## ccat3total       0.63
## ccat4total       0.07
## ccat5total       0.08
## ccat6total      -0.03
## ccat7total       1.00
##
## n= 485
##
##
## P
##            ccat1total ccat2total ccat3total ccat4total ccat5total
ccat6total
## ccat1total            0.0000     0.0000     0.0122     0.0784     0.5045
## ccat2total 0.0000                0.0000     0.0000     0.0000     0.0000
## ccat3total 0.0000     0.0000                0.3491     0.5471     0.0000
## ccat4total 0.0122     0.0000     0.3491                0.0000     0.0000
## ccat5total 0.0784     0.0000     0.5471     0.0000                0.0000
## ccat6total 0.5045     0.0000     0.0000     0.0000     0.0000
## ccat7total 0.0000     0.0000     0.0000     0.1482     0.0929     0.4547
##            ccat7total
## ccat1total 0.0000
## ccat2total 0.0000
## ccat3total 0.0000
## ccat4total 0.1482
## ccat5total 0.0929
## ccat6total 0.4547
## ccat7total
```

Part 3: Confidence Intervals/ Prediction

3. Download the dataset "baldrige2011.xlsx" posted on Canvas. To answer the questions, please use R markdown to execute the R code and document it with appropriate comments and observations wherever it is required. Please use "stargazer" library for showing all the output tables and "ggplot2" for all graphs.

a. Construct a 95% confidence interval for the average icat7total. Based on this confidence interval, what is the maximum icat7total that a company can score with 95% confidence. Write the R code that you used to arrive at this answer.

```
t.test(df$icat7total, conf.level = 0.95)

##
##  One Sample t-test
##
## data:  df$icat7total
## t = 71.164, df = 1096, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  139.5226 147.4345
## sample estimates:
## mean of x
##  143.4786
```

The limits of the 95% Confidence Interval for average icat7total are 139.52 (lowest) and 147.43 (Highest). The maximum icat7total score the firm can expect with a 95% confidence is 147.43.

b. Is icat7total score statistically different for companies in healthcare and education sector? If so, which set of firms have higher score? To answer these questions, split the data into two subsets for healthcare and education firms and construct a 95% confidence intervals for each subset. Based on these two confidence intervals, is it likely that the average score of healthcare firms is same as that of education firms? Write the R code that you used to derive this inference.

```
edu=subset(df,sector==4,select = c(icat7total))
t.test(edu, conf.level = 0.95)

##
##  One Sample t-test
##
## data:  edu
## t = 24.087, df = 133, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  131.2367 154.7185
## sample estimates:
## mean of x
##  142.9776

health=subset(df,sector==5,select = c(icat7total))
t.test(health, conf.level = 0.95)

##
##  One Sample t-test
##
## data:  health
## t = 31.544, df = 203, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  147.4483 167.1105
## sample estimates:
## mean of x
##  157.2794
```

We can infer from the output above that the health sector companies are scoring higher in the icat7total score. The p-value of both the firms are almost the same which shows they

are not much statistically different. The mean score of health sector is higher than the education sector.