**Assignment No. 13.1**

> COBRA.YTD2017 <- read.csv("E:/kamagyana/Computing/DARET/Assignments/COBRA-YTD2017.csv", stringsAsFactors=FALSE)

> View(COBRA.YTD2017)

> str(COBRA.YTD2017)

'data.frame': 26759 obs. of 23 variables:

$ MI\_PRINX : int 8924155 8924156 8924157 8924158 8924159 8924160 8924161 8924162 8924163 8924164 ...

$ offense\_id : num 1.74e+08 1.74e+08 1.74e+08 1.74e+08 1.74e+08 ...

$ rpt\_date : chr "12/31/2017" "12/31/2017" "12/31/2017" "12/31/2017" ...

$ occur\_date : chr "12/30/2017" "12/18/2017" "12/30/2017" "12/30/2017" ...

$ occur\_time : chr "23:15:00" "13:00:00" "22:01:00" "20:00:00" ...

$ poss\_date : chr "12/31/2017" "12/30/2017" "12/31/2017" "12/31/2017" ...

$ poss\_time : chr "00:30:00" "22:00:00" "01:00:00" "01:06:00" ...

$ beat : int 510 501 303 507 409 612 605 603 605 304 ...

$ apt\_office\_prefix: chr "" "" "" "" ...

$ apt\_office\_num : chr "" "" "" "" ...

$ location : chr "43 JESSE HILL JR DR NE" "1169 ATLANTIC DR NW" "633 PRYOR ST SW" "333 NELSON ST SW" ...

$ MinOfucr : int 640 640 640 640 640 650 311 640 640 531 ...

$ MinOfibr\_code : chr "2305" "2305" "2305" "2305" ...

$ dispo\_code : chr "" "" "" "" ...

$ MaxOfnum\_victims : int 2 1 1 1 2 1 1 1 1 1 ...

$ Shift : chr "Morn" "Unk" "Morn" "Eve" ...

$ Avg.Day : chr "Sat" "Unk" "Sat" "Sat" ...

$ loc\_type : int 13 13 18 18 18 18 26 18 13 26 ...

$ UC2.Literal : chr "LARCENY-FROM VEHICLE" "LARCENY-FROM VEHICLE" "LARCENY-FROM VEHICLE" "LARCENY-FROM VEHICLE" ...

$ neighborhood : chr "Downtown" "Home Park" "Mechanicsville" "Castleberry Hill" ...

$ npu : chr "M" "E" "V" "M" ...

$ x : num -84.4 -84.4 -84.4 -84.4 -84.5 ...

$ y : num 33.8 33.8 33.7 33.8 33.7 ...

View(COBRA.YTD2017)

str(COBRA.YTD2017)

sum(is.na(COBRA.YTD2017))

crimedata <- COBRA.YTD2017

colnames(crimedata)[22] <- "longitude"

colnames(crimedata)[23] <- "lattitude"

colnames(crimedata)[19] <- "CrimeType"

colnames(crimedata)[21] <- "NeiPlgUnit"

library(lubridate)

crimedata$rpt\_date <- mdy(crimedata$rpt\_date)

crimedata$occur\_date <- mdy(crimedata$occur\_date)

crimedata$poss\_date <- mdy(crimedata$poss\_date)

crimedata$occur\_time <- hms(crimedata$occur\_time)

crimedata$poss\_time <- hms(crimedata$poss\_time)

**QUESTION: 1**

Find out top 5 attributes having highest correlation (select only Numeric features).

numcrime <- cbind(crimedata[,c(2,22,23)])

str(numcrime)

cor(numcrime, method = "pearson", use = "complete.obs")

> str(numcrime)

'data.frame': 26759 obs. of 3 variables:

$ offense\_id: num 1.74e+08 1.74e+08 1.74e+08 1.74e+08 1.74e+08 ...

$ longitude : num -84.4 -84.4 -84.4 -84.4 -84.5 ...

$ lattitude : num 33.8 33.8 33.7 33.8 33.7 ...

> cor(numcrime, method = "pearson", use = "complete.obs")

offense\_id longitude lattitude

offense\_id 1.000000000 -0.004996624 0.004719228

longitude -0.004996624 1.000000000 -0.999835504

lattitude 0.004719228 -0.999835504 1.000000000

NOTE: If we consider whatever data is given as it is. The numeric data types are only three of them, which are the offence\_id, latitude, and longitude. Prima Facie the correlation of these three data does not make any sense, but for testing whether the offence id is related to the latitude and longitude. The above results show clearly that there is no correlation between the three, and logically the latitude and longitude are strongly correlated.

Further if we continue to extract all the integer variables also and then convert them as numeric and re-test with their correlations, even then there seems to be no major change in the correlation between the variables. So identifying the top 5 variables which are numerical type (as required by the question) is not possible. The following is the analysis of the same.

numcrime <- cbind(crimedata[,c(1,2,8,12,13,15,18,22,23)])

numcrime <- numcrime[,-5]

numcrime$MI\_PRINX <- as.numeric(numcrime$MI\_PRINX)

numcrime$beat <- as.numeric(numcrime$beat)

numcrime$MinOfucr <- as.numeric(numcrime$MinOfucr)

numcrime$dispo\_code <- as.numeric(numcrime$dispo\_code)

numcrime$loc\_type <- as.numeric(numcrime$loc\_type)

str(numcrime)

cor(numcrime, method = "pearson", use = "complete.obs")

MI\_PRINX ffense\_id beat MinOfucr dispo\_code

MI\_PRINX 1.000000000 0.992113449 0.036305597 0.08993369 -0.04108935

offense\_id 0.992113449 1.000000000 0.039430562 0.08267135 -0.03844196

beat 0.036305597 0.039430562 1.000000000 0.07164977 -0.09110722

MinOfucr 0.089933693 0.082671351 0.071649772 1.00000000 -0.03041830

dispo\_code -0.041089349 -0.038441965 -0.091107218 -0.03041830 1.00000000

loc\_type -0.003274369 -0.002780581 0.003423581 -0.11426129 0.07980919

longitude -0.054418151 -0.053894906 0.008780950 -0.36516082 0.05245724

lattitude 0.054623864 0.053987132 -0.010633099 0.36789643 -0.05408415

loc\_type longitude lattitude

MI\_PRINX -0.003274369 -0.05441815 0.05462386

offense\_id -0.002780581 -0.05389491 0.05398713

beat 0.003423581 0.00878095 -0.01063310

MinOfucr -0.114261286 -0.36516082 0.36789643

dispo\_code 0.079809192 0.05245724 -0.05408415

loc\_type 1.000000000 0.03681847 -0.03848476

longitude 0.036818465 1.00000000 -0.99991694

lattitude -0.038484762 -0.99991694 1.00000000

**Question No. 2:**

Top 3 reasons for having crime in the city

> crime <-as.data.frame(table(crimedata$CrimeType))

> colnames(crime)[1] <- "Crime"

> crime

Crime Freq

1 AGG ASSAULT 2024

2 AUTO THEFT 3197

3 BURGLARY-NONRES 758

4 BURGLARY-RESIDENCE 2635

5 HOMICIDE 75

6 LARCENY-FROM VEHICLE 9840

7 LARCENY-NON VEHICLE 6589

8 RAPE 226

9 ROBBERY-COMMERCIAL 157

10 ROBBERY-PEDESTRIAN 1126

11 ROBBERY-RESIDENCE 132

> crime[order(-crime$Freq),]

Crime Freq

6 LARCENY-FROM VEHICLE 9840

7 LARCENY-NON VEHICLE 6589

2 AUTO THEFT 3197

4 BURGLARY-RESIDENCE 2635

1 AGG ASSAULT 2024

10 ROBBERY-PEDESTRIAN 1126

3 BURGLARY-NONRES 758

8 RAPE 226

9 ROBBERY-COMMERCIAL 157

11 ROBBERY-RESIDENCE 132

5 HOMICIDE 75