SESSION 12: Generalized Linear Models

Assignment 1

Problem Statement

1. Use the given link below:

https://archive.ics.uci.edu/ml/machine-learning-databases/communities/

Perform the below operations:

a. Find out top 5 attributes having highest correlation (select

only Numeric features).

```
#splitting
library(caTools)
set.seed(123)
split = sample.split(dataset$Profit,SplitRatio = 0.75)
traindata = subset(dataset,split == T)
testdata = subset(dataset,split == F)
#model building
regressor = lm(formula = Profit ~ ., data = traindata)
summary(regressor)
Profit_predicted = predict(regressor,newdata = testdata)
Profit_predicted


dtset <- read.csv(file.choose())
View(dtset)
library(caTools)
set.seed(123)
split=sample.split(dtset$Unit.Price, Split=0.75)
traindt=subset(dtset,split==T)
tstdata=subset(dtset,split==F)
##model building
```

```
regressor=lm(formula = Unit.Price ~ units,data = traindt)
# Multi Linear Regression
housing<- read.csv(file.choose())
View(housing)
#datadictionary
#price respresents the sale price of the house
#area gives total size of the property in square feet
#bedrooms gives the total no of bedrooms
#bathrooms gives the total no of bathrooms
#stories gives the no of stories the building has
#mainroad=1,whether the house is facing the maidroad

str(housing)
#main road
str(housing$mainroad)
summary(factor(housing$mainroad))
#simple way is to convert yes and no into 0 and 1's
levels(housing$mainroad)<- c(0,1)
housing$mainroad<- as.numeric(levels(housing$mainroad))[housing$mainroad]

levels(housing$guestroom)<- c(0,1)
housing$guestroom<-as.numeric(levels(housing$guestroom))[housing$guestroom]

levels(housing$basement)<- c(0,1)
housing$basement<-as.numeric(levels(housing$basement))[housing$basement]

levels(housing$hotwaterheating)<- c(0,1)
housing$hotwaterheating<-as.numeric(levels(housing$hotwaterheating))
[housing$hotwaterheating]
```

```
levels(housing$airconditioning)<- c(0,1)
housing$airconditioning<-as.numeric(levels(housing$airconditioning))[housing$airconditioning]

levels(housing$prefarea)<- c(0,1)
housing$prefarea<-as.numeric(levels(housing$prefarea))[housing$prefarea]

#adding new columns
dummy_1<-data.frame(model.matrix(~furnishingstatus,data = housing))
View(dummy_1)
#removing a column
dummy_1<-dummy_1[,-1]
View(dummy_1)
# removing existing column and adding new column
housing_1<-cbind(housing[,-13],dummy_1)
View(housing_1)

#####Derived metrics

housing_1$areaperbedroom<- housing_1$area/housing_1$bedrooms

housing_1$bbratio <- housing_1$bathrooms/housing_1$bedrooms

#####Model Building

set.seed(100)
# to creat train data & test data
trainindices <- sample(1:nrow(housing_1),0.7*nrow(housing_1))
train <- housing_1[trainindices,]
View(train)
#taking the data into test
```

```
test <- housing_1[-trainindices,]

View(test)


model_1 <- lm(price~.,data=train)

summary(model_1)


library(car)

vif(model_1)

summary(model_1)

#drop Bbratio

model_2 <- lm(formula = price ~area + bedrooms + bathrooms + stories + mainroad +
guestroom +

                basement + hotwaterheating + airconditioning + areaperbedroom +
furnishingstatussemi.furnished + furnishingstatusunfurnished +

                parking + prefarea , data = train)

vif(model_2)

summary(model_2)
```

b. Find out top 3 reasons for having more crime in a city.

```
ggplot()+ geom_point(aes(x=testdata$TV,y = testdata$Sales),colour = 'red')+
  geom_line(aes(x = trainingdata$TV,y =predict(regressor, newdata = trainingdata)),
          colour ='blue')+
          ggtitle('TV vs Sales')+
          xlab("TV Spend")+
          ylab('Sales')
```

c. Which all attributes have high correlation with crime rate?

```
model_8 <- lm(formula = price ~area+ bathrooms + stories + guestroom +
              airconditioning + furnishingstatusunfurnished +
               parking + prefarea , data = train)
```