CS513Midterm

Taylor Niedzielski

3/28/2021

Question 2

Perform the EDA Analysis for the fictional dataset COVID19_v2.csv I.Summarizing each column (e.g. min, max, mean) II.Identifying missing values III.Displaying the frequency table of "Infected" vs. "MaritalStatus" IV.Displaying the scatter plot of "Age", "MaritalStatus" and "MonthAtHospital", one pair at a time V.Show box plots for columns: "Age", and "MonthAtHospital"VI.Replacing the missing values of "Cases" with the "mean" of "Cases". VI.Replacing the missing values of "Cases" with the "mean" of "Cases".

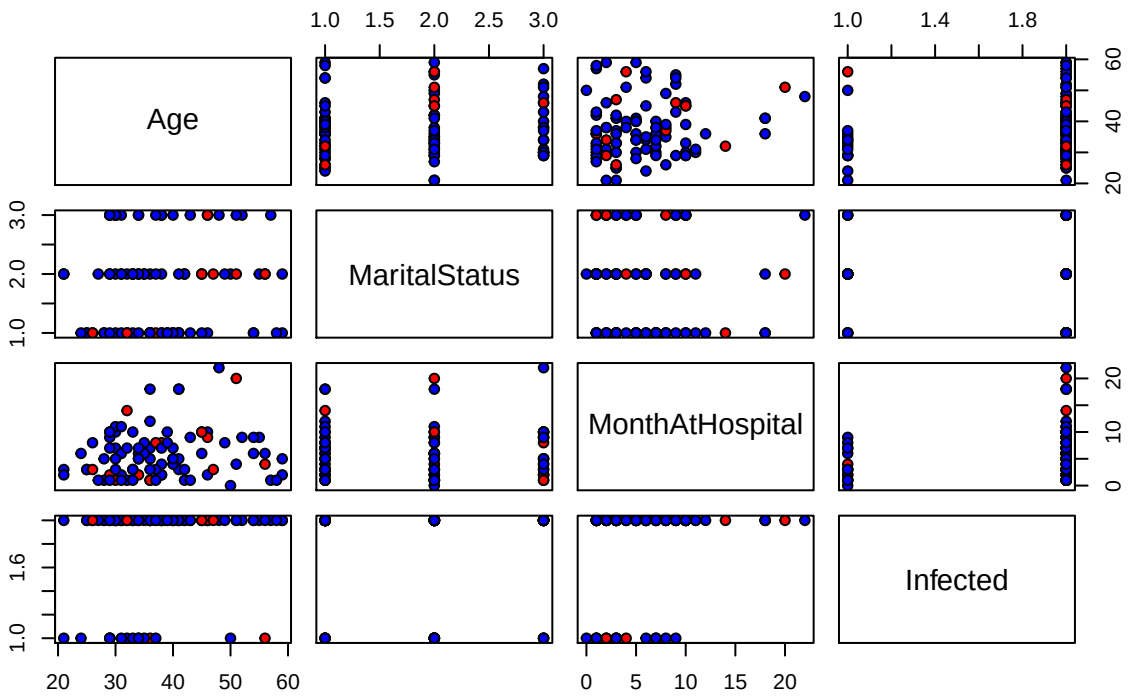
```
covid19 <- read.csv("/Users/taylorniedzielski/Desktop/cs513/COVID19_v2.csv")

# I.    Summarizing each column (e.g. min, max, mean )
summary(covid19[,1:6])
```

```
##        ID            Age           Exposure    MaritalStatus
##  Min.   :1001   Min.   :21.00   Min.   :1.00   Length:100
##  1st Qu.:1026   1st Qu.:31.00   1st Qu.:1.00   Class :character
##  Median :1050   Median :36.50   Median :3.00   Mode  :character
##  Mean   :1050   Mean   :38.32   Mean   :2.66
##  3rd Qu.:1075   3rd Qu.:44.50   3rd Qu.:4.00
##  Max.   :1100   Max.   :60.00   Max.   :4.00
##                 NA's   :2
##      Cases       MonthAtHospital
##  Min.   : 5434   Min.   : 0.000
##  1st Qu.:17782   1st Qu.: 2.500
##  Median :20276   Median : 5.000
##  Mean   :19062   Mean   : 6.116
##  3rd Qu.:22354   3rd Qu.: 9.000
##  Max.   :25000   Max.   :22.000
##  NA's   :2       NA's   :5
```

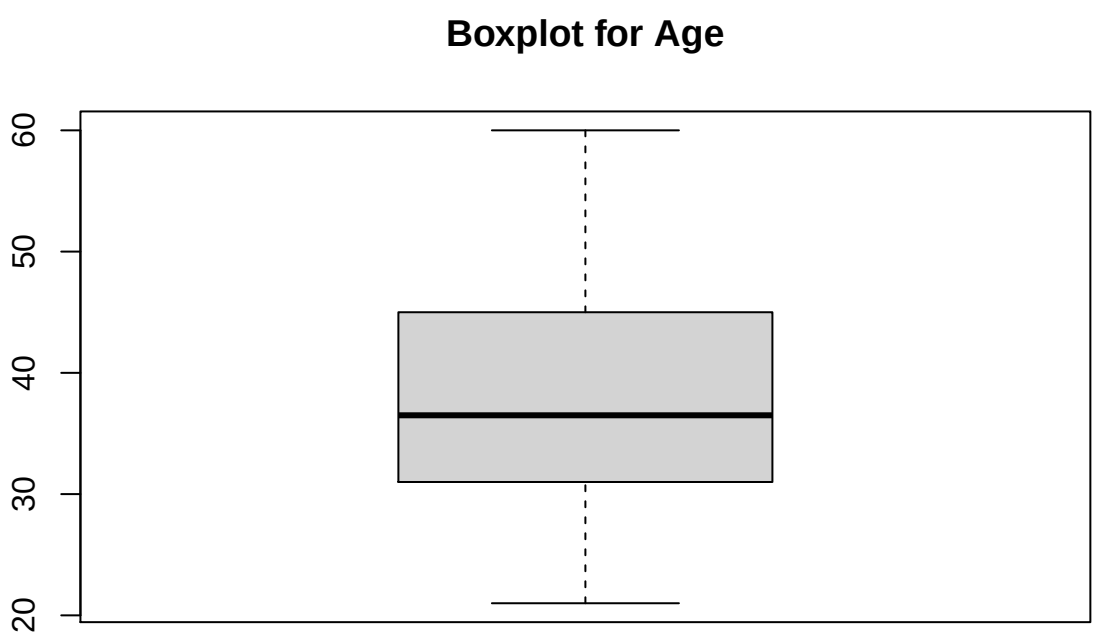
```
# II.   Identifying missing values
missing <- is.na(covid19[,1:6])

# III.  Displaying the frequency table of "Infected" vs. "MaritalStatus"
table(Class = covid19$MaritalStatus, F6 = covid19$Infected)
```

```
##           F6
## Class      No Yes
##   Divorced 20   3
##   Married  40   4
##   Single   25   8
```

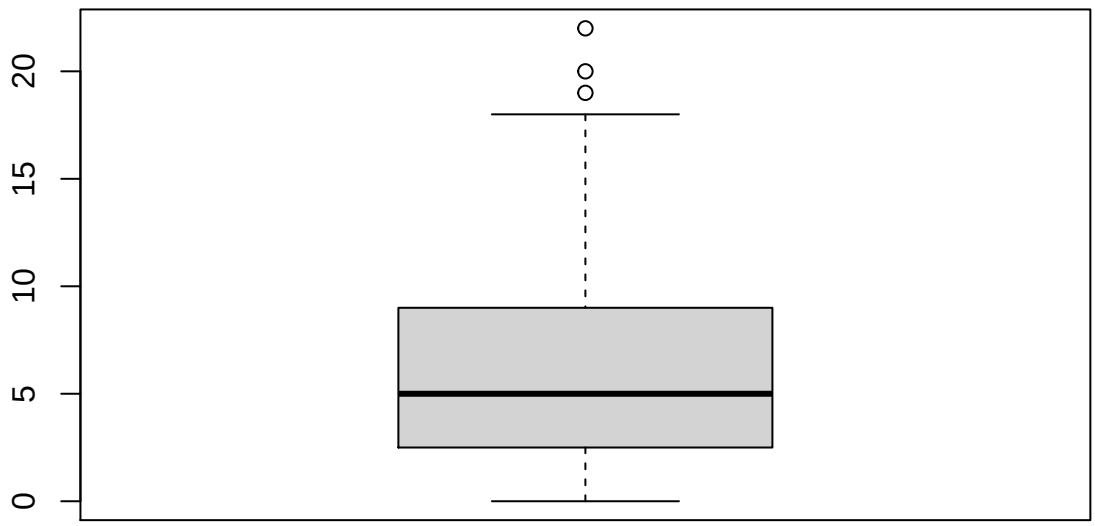
```r
#####################################Clean Data#######################################
##locate all non numeric types
def = c(0,0,0,0,0,0,0)
for(i in 1:ncol(covid19)) {
  def[i] <- typeof(covid19[,i])
}

##convert married/single/divorced status to numeric to perform operations
covidcol4 <- covid19[,4]
for(i in 1:length(covidcol4)) {
  if(covidcol4[i] == "Married") {
    covidcol4[i] = 1;
  }
  if(covidcol4[i] == "Single") {
    covidcol4[i] = 2;
  }
  if(covidcol4[i] == "Divorced") {
    covidcol4[i] = 3;
  }
}
covid19[,4] <- covidcol4
covid19[,4] <- as.numeric(covid19[,4])

##convert yes/no to numeric to perform operations
covidcol7 <- covid19[,7]
for(i in 1:length(covidcol7)) {
  if(covidcol7[i] == "Yes") {
    covidcol7[i] = 1;
  }
  if(covidcol7[i] == "No") {
    covidcol7[i] = 2;
  }
}
covid19[,7] <- covidcol7
covid19[,7] <- as.numeric(covid19[,7])

naOmitCovid19<- na.omit(covid19)
#################################################################################
# IV.Displaying the scatter plot of "Age", "MaritalStatus" and "MonthAtHospital",
# one pair at a time

pairs(naOmitCovid19[c(2,4,6,7)], main = "Scatter Plot of COVID19 Data",
      pch = 21,bg =c("red","blue")[factor(covid19$Infected)] )
```




```
# V.Show box plots for columns:  "Age", and "MonthAtHospital"
boxplot(covid19$Age, main = "Boxplot for Age")
```




```
boxplot(covid19$MonthAtHospital, main = "Boxplot for Months At Hospital")
```




```
# VI.Replacing the missing values #of "Cases" with the "mean" of "Cases".
for(i in 2:ncol(covid19)){
  covid19[is.na(covid19[,i]), i] <- round(mean(covid19[,i], na.rm = TRUE))
}

summary(covid19)
```

```
##        ID            Age          Exposure    MaritalStatus      Cases
##  Min.   :1001   Min.   :21.00   Min.   :1.00   Min.   :1.00   Min.   : 5434
##  1st Qu.:1026   1st Qu.:31.00   1st Qu.:1.00   1st Qu.:1.00   1st Qu.:17864
##  Median :1050   Median :37.00   Median :3.00   Median :2.00   Median :20246
##  Mean   :1050   Mean   :38.31   Mean   :2.66   Mean   :1.79   Mean   :19062
##  3rd Qu.:1075   3rd Qu.:43.50   3rd Qu.:4.00   3rd Qu.:2.00   3rd Qu.:22345
##  Max.   :1100   Max.   :60.00   Max.   :4.00   Max.   :3.00   Max.   :25000
##  MonthAtHospital    Infected
##  Min.   : 0.00   Min.   :1.00
##  1st Qu.: 3.00   1st Qu.:2.00
##  Median : 6.00   Median :2.00
##  Mean   : 6.11   Mean   :1.85
##  3rd Qu.: 8.25   3rd Qu.:2.00
##  Max.   :22.00   Max.   :2.00
```

Question 4

Load the CANVAS fictional "COVID19_v2.CSV" dataset into R/Python. Remove the missing values. Use unweighted knn(k=5) to predict infection rate (infected) for a random sample (30%) of the data (test dataset)

```
rm(list=ls())
covid19 <- read.csv("/Users/taylorniedzielski/Desktop/cs513/COVID19_v2.csv",
                    colClasses=c("Infected"="factor" ))

#get rid of N/A
naOmitCovid19<- na.omit(covid19)

# create training and test data sets
index<-sort(sample(nrow( naOmitCovid19),round(.30*nrow(naOmitCovid19 ))))
training<- naOmitCovid19[-index,]
test<- naOmitCovid19[index,]

library(kknn)
predict_k1 <- kknn(formula= Infected~., training[,c(-1)] , test[,c(-1)], k=5,
                   kernel ="rectangular"  )

 fit <- fitted(predict_k1)
 table(test$Infected,fit)
```

```
##      fit
##       No Yes
##   No  23   0
##   Yes  4   0
```

```
 # Measure the performance of knn

wrong<- ( test$Infected!=fit)
rate<-sum(wrong)/length(wrong)
rate
```

```
## [1] 0.1481481
```

Question 5

Load the CANVAS "COVID19_v2.CSV" dataset into R/Python. Remove the missing values. Discretize the "MonthAtHospital" into "less than 6 months" and "6 or more months". Also discretize the age into "less than 35", "35 to 50" and "51 or over". Construct a Naïve Bayes model to classify infection ("infected') based on the other variables. Measure the accuracy of the model.Do not use the original MonthAtHospital and age variables as predictors

```
rm(list=ls())
covid19 <- read.csv("/Users/taylorniedzielski/Desktop/cs513/COVID19_v2.csv",
                    colClasses=c("Infected"="factor" ))

natesting <-na.omit(covid19)
naOmitCovid19<- na.omit(covid19)
```

```r
months <- naOmitCovid19[,6]
for(i in 1:length(months)) {
  if(months[i] < 6) {
    months[i] = 1;
  }
  if(months[i] >= 6) {
    months[i] = 2;
  }
}
naOmitCovid19[,6] <- months

age <- naOmitCovid19[,2]
for(i in 1:length(age)) {
  if(age[i] < 35) {
    age[i] = 1;
  }
  if(age[i] >= 35  || age[i] <= 50) {
    age[i] = 2;
  }
  if(age[i] > 50) {
    age[i] = 3;
  }
}
naOmitCovid19[,2] <- age

library(e1071)

nBayes_all <- naiveBayes(factor(Infected)~., data = naOmitCovid19)
category_all<-predict(nBayes_all,natesting )

table(nBayes_all=category_all, Samples=natesting$Infected)
```

```
##           Samples
## nBayes_all No Yes
##        No  76  14
##        Yes  0   1
```

```r
NB_wrong<-sum(category_all!=natesting$Infected)

NB_error_rate<-NB_wrong/length(category_all)
NB_error_rate
```

```
## [1] 0.1538462
```