Data exploration Assignment

R Markdown

```
# Setting the working directory

setwd("/Users/myanamandravenkata@unomaha.edu/Downloads")


# Importing the dataset


hfs = read.csv("HFS Service Data.csv")

# Viewing the first few rows of the dataset


head(hfs)
```

```
##   gender  program_name              program_type
## 1   Male Mental Health Counseling and Prevention
## 2 Female Mental Health Counseling and Prevention
## 3 Female Mental Health Counseling and Prevention
## 4 Female Mental Health Counseling and Prevention
## 5 Female Mental Health Counseling and Prevention
## 6 Female Mental Health Counseling and Prevention
##                            facility           job_title       staff_name
## 1 Heartland Family Service - Logan Clinical Supervisor   Poore, Lindsay
## 2               Center Mall Office        THERAPIST II Carlson, Kaitlin
## 3               Center Mall Office        THERAPIST II Carlson, Kaitlin
## 4               Center Mall Office        THERAPIST II Carlson, Kaitlin
## 5               Center Mall Office        THERAPIST II Carlson, Kaitlin
## 6               Center Mall Office        THERAPIST II Carlson, Kaitlin
##   actual_date duration                     event_name activity_type
## 1         961     0:00 Daily Living Assessment DLA 20
## 2         857     0:02                Collateral Note         Phone
## 3         682     0:51             Individual Therapy
## 4         710     0:00             Individual Therapy
## 5         696     0:50             Individual Therapy
## 6         772     1:20           Case Management Note
##   encounter_with is_client_involved is_noshow is_locked is_billed is_paid
## 1                              TRUE     FALSE     FALSE     FALSE   FALSE
## 2         Client               TRUE     FALSE      TRUE     FALSE   FALSE
## 3                              TRUE     FALSE      TRUE      TRUE   FALSE
## 4                              TRUE      TRUE      TRUE     FALSE   FALSE
## 5                              TRUE     FALSE      TRUE      TRUE   FALSE
## 6                              TRUE     FALSE      TRUE     FALSE   FALSE
##   date_entered   user_entered_name approved_date approved_staff_name submi
tted
```

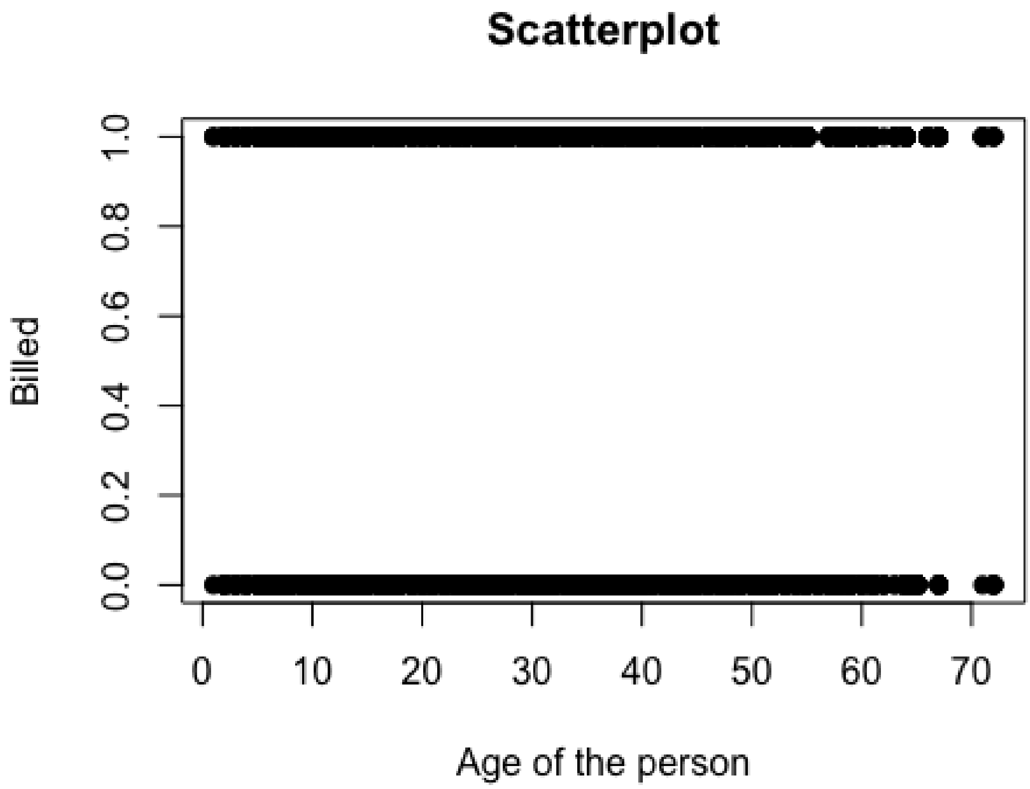
```
## 1          961      Poore, Lindsay            NA                           
## 2          857 Carlson, Kaitlin V.           857 Carlson, Kaitlin V.  Appr
oved
## 3          683 Carlson, Kaitlin V.           689        Stanek, Sean  Appr
oved
## 4          710 Carlson, Kaitlin V.           714        Stanek, Sean  Appr
oved
## 5          696 Carlson, Kaitlin V.           697        Stanek, Sean  Appr
oved
## 6          773 Carlson, Kaitlin V.           773        Stanek, Sean  Appr
oved
##   is_approved is_notapproved is_notapproved_subm
## 1           0              0                   1
## 2           1              0                   0
## 3           1              0                   0
## 4           1              0                   0
## 5           1              0                   0
## 6           1              0                   0
##                program_unit_description sc_code duration_num do_not_bill
## 1 Behavioral Health IA -  Mental Health 1311-16            0       FALSE
## 2  Behavioral Health NE - Mental Health 1311-05            2       FALSE
## 3  Behavioral Health NE - Mental Health 1311-05           51       FALSE
## 4  Behavioral Health NE - Mental Health 1311-05            0       FALSE
## 5  Behavioral Health NE - Mental Health 1311-05           50       FALSE
## 6  Behavioral Health NE - Mental Health 1311-05           80       FALSE
##   do_not_pay   general_location             program_modifier
## 1      FALSE                                No Modifier - IA
## 2      FALSE   Homeless Shelter Heartland Housing Navigation
## 3      FALSE               Home Heartland Housing Navigation
## 4      FALSE Telehealth - Phone Heartland Housing Navigation
## 5      FALSE Telehealth - Phone Heartland Housing Navigation
## 6      FALSE               Home Heartland Housing Navigation
##   program_modifier_code NormalWorkHours duration_other_num duration_other
## 1                 NMODI             Yes                  0           0:00
## 2                   HHN             Yes                  0           0:00
## 3                   HHN             Yes                 10           0:10
## 4                   HHN             Yes                  0           0:00
## 5                   HHN             Yes                 10           0:10
## 6                   HHN             Yes                 10           0:10
##   travel_time_num travel_time planning_time_num planning_time
## 1               0        0:00                 0          0:00
## 2               0        0:00                 0          0:00
## 3               0        0:00                 0          0:00
## 4               0        0:00                 0          0:00
## 5               0        0:00                 0          0:00
## 6               0        0:00                 0          0:00
##   total_duration_num total_duration       reason_for_no_show is_billable z
ip
## 1                  0           0:00                                FALSE  
0
```

```
## 2                  2           0:02                                FALSE 6
81
## 3                 61           1:01                                FALSE 6
81
## 4                  0           0:00 Client No Show - No Call       FALSE 6
81
## 5                 60           1:00                                FALSE 6
81
## 6                 90           1:30                                FALSE 6
81
##   state age recordID simple_race             ethnic_identity gender_identi
ty
## 1    IA  12      298           8 Not Spanish/Hispanic/Latino    Not Obtain
ed
## 2    NE  26      338          16 Not Spanish/Hispanic/Latino            <N
A>
## 3    NE  25      338          16 Not Spanish/Hispanic/Latino            <N
A>
## 4    NE  25      338          16 Not Spanish/Hispanic/Latino            <N
A>
## 5    NE  25      338          16 Not Spanish/Hispanic/Latino            <N
A>
## 6    NE  25      338          16 Not Spanish/Hispanic/Latino            <N
A>
##   sexual_orientation
## 1       Not Obtained
## 2               <NA>
## 3               <NA>
## 4               <NA>
## 5               <NA>
## 6               <NA>
```

```
df <- data.frame(hfs)
df$is_client_involved <- as.numeric(df$is_client_involved)
head(df)
```

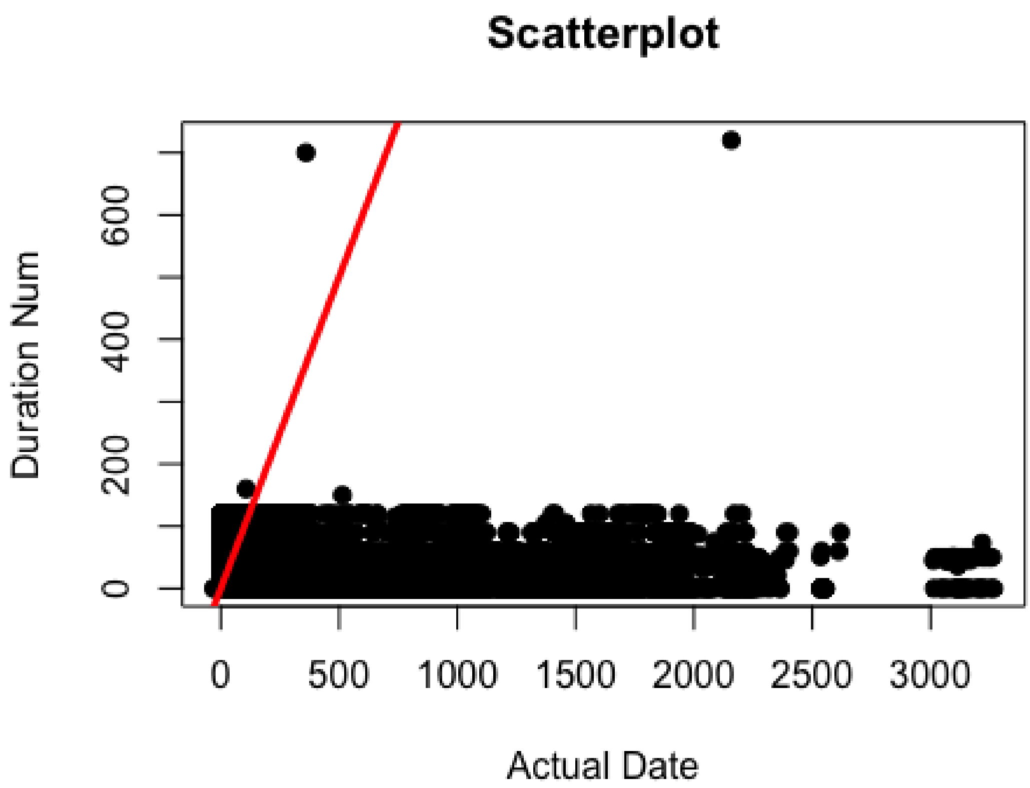
```
##   gender  program_name              program_type
## 1   Male Mental Health Counseling and Prevention
## 2 Female Mental Health Counseling and Prevention
## 3 Female Mental Health Counseling and Prevention
## 4 Female Mental Health Counseling and Prevention
## 5 Female Mental Health Counseling and Prevention
## 6 Female Mental Health Counseling and Prevention
##                           facility           job_title       staff_name
## 1 Heartland Family Service - Logan Clinical Supervisor   Poore, Lindsay
## 2               Center Mall Office        THERAPIST II Carlson, Kaitlin
## 3               Center Mall Office        THERAPIST II Carlson, Kaitlin
## 4               Center Mall Office        THERAPIST II Carlson, Kaitlin
## 5               Center Mall Office        THERAPIST II Carlson, Kaitlin
## 6               Center Mall Office        THERAPIST II Carlson, Kaitlin
##   actual_date duration                     event_name activity_type
## 1         961     0:00 Daily Living Assessment DLA 20              
## 2         857     0:02                Collateral Note         Phone
## 3         682     0:51             Individual Therapy              
## 4         710     0:00             Individual Therapy              
## 5         696     0:50             Individual Therapy              
## 6         772     1:20           Case Management Note              
##   encounter_with is_client_involved is_noshow is_locked is_billed is_paid
## 1                                 1     FALSE     FALSE     FALSE   FALSE
## 2         Client                  1     FALSE      TRUE     FALSE   FALSE
## 3                                 1     FALSE      TRUE      TRUE   FALSE
## 4                                 1      TRUE      TRUE     FALSE   FALSE
## 5                                 1     FALSE      TRUE      TRUE   FALSE
## 6                                 1     FALSE      TRUE     FALSE   FALSE
##   date_entered   user_entered_name approved_date approved_staff_name submi
tted
## 1          961      Poore, Lindsay            NA                          
## 2          857 Carlson, Kaitlin V.           857 Carlson, Kaitlin V.  Appr
oved
## 3          683 Carlson, Kaitlin V.           689        Stanek, Sean  Appr
oved
## 4          710 Carlson, Kaitlin V.           714        Stanek, Sean  Appr
oved
## 5          696 Carlson, Kaitlin V.           697        Stanek, Sean  Appr
oved
## 6          773 Carlson, Kaitlin V.           773        Stanek, Sean  Appr
oved
##   is_approved is_notapproved is_notapproved_subm
## 1           0              0                   1
## 2           1              0                   0
## 3           1              0                   0
## 4           1              0                   0
## 5           1              0                   0
## 6           1              0                   0
##                  program_unit_description sc_code duration_num do_not_bill
## 1 Behavioral Health IA -  Mental Health 1311-16            0       FALSE
## 2  Behavioral Health NE - Mental Health 1311-05            2       FALSE
## 3  Behavioral Health NE - Mental Health 1311-05           51       FALSE
## 4  Behavioral Health NE - Mental Health 1311-05            0       FALSE
## 5  Behavioral Health NE - Mental Health 1311-05           50       FALSE
## 6  Behavioral Health NE - Mental Health 1311-05           80       FALSE
##   do_not_pay   general_location             program_modifier
## 1      FALSE                                No Modifier - IA
## 2      FALSE   Homeless Shelter Heartland Housing Navigation
## 3      FALSE               Home Heartland Housing Navigation
## 4      FALSE Telehealth - Phone Heartland Housing Navigation
## 5      FALSE Telehealth - Phone Heartland Housing Navigation
## 6      FALSE               Home Heartland Housing Navigation
##   program_modifier_code NormalWorkHours duration_other_num duration_other
## 1                 NMODI             Yes                  0           0:00
## 2                   HHN             Yes                  0           0:00
## 3                   HHN             Yes                 10           0:10
## 4                   HHN             Yes                  0           0:00
## 5                   HHN             Yes                 10           0:10
## 6                   HHN             Yes                 10           0:10
##   travel_time_num travel_time planning_time_num planning_time
## 1               0        0:00                 0          0:00
## 2               0        0:00                 0          0:00
## 3               0        0:00                 0          0:00
## 4               0        0:00                 0          0:00
## 5               0        0:00                 0          0:00
## 6               0        0:00                 0          0:00
##   total_duration_num total_duration       reason_for_no_show is_billable z
ip
## 1                  0           0:00                                FALSE
0
## 2                  2           0:02                                FALSE 6
81
## 3                 61           1:01                                FALSE 6
81
## 4                  0           0:00 Client No Show - No Call       FALSE 6
81
## 5                 60           1:00                                FALSE 6
81
## 6                 90           1:30                                FALSE 6
81
##   state age recordID simple_race             ethnic_identity gender_identi
ty
## 1    IA  12      298           8 Not Spanish/Hispanic/Latino    Not Obtain
ed
## 2    NE  26      338          16 Not Spanish/Hispanic/Latino            <N
A>
## 3    NE  25      338          16 Not Spanish/Hispanic/Latino            <N
A>
## 4    NE  25      338          16 Not Spanish/Hispanic/Latino            <N
A>
## 5    NE  25      338          16 Not Spanish/Hispanic/Latino            <N
A>
## 6    NE  25      338          16 Not Spanish/Hispanic/Latino            <N
A>
##   sexual_orientation
## 1       Not Obtained
## 2               <NA>
## 3               <NA>
## 4               <NA>
## 5               <NA>
## 6               <NA>
```

```
# Converting the following fields to numeric because, these had Yes and No or
True and False, therefore, I encoded them to numeric values with is.numeric()
function.

df$is_noshow <- as.numeric(df$is_noshow)
df$is_locked <- as.numeric(df$is_locked)
df$is_billed <- as.numeric(df$is_billed)
df$is_paid <- as.numeric(df$is_paid)

df$do_not_bill <- as.numeric(df$do_not_bill)
df$do_not_pay <- as.numeric(df$do_not_pay)
df$is_billable <- as.numeric(df$is_billable)

attach(df)

# A scatterplot for the variables age and is_billed. This scatter plot shows
the distribution of the age of the person and their billed status.


plot(df$age, is_billed, main = "Scatterplot",xlab = "Age of the person", ylab
= "Billed", pch=19)
```



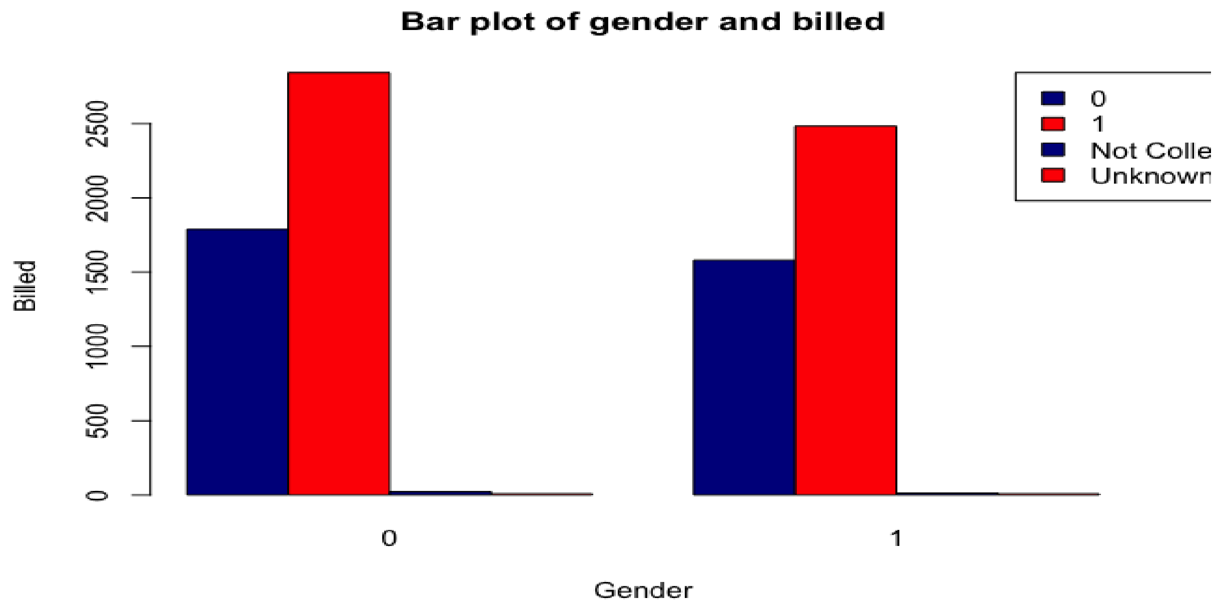
```
# A scatterplot of actual_date and the duration_num field with an abline show
ing the trend whether this data is linear or not. The data is clearly not lin
```

```
ear between actual_date and the duration_num as none of the data points lie o
n the line.

plot(df$actual_date, df$duration_num, main = "Scatterplot", xlab = "Actual Da
te",
     ylab = "Duration Num", pch=19)
abline(0,1,lwd = 3, col = "red")
```

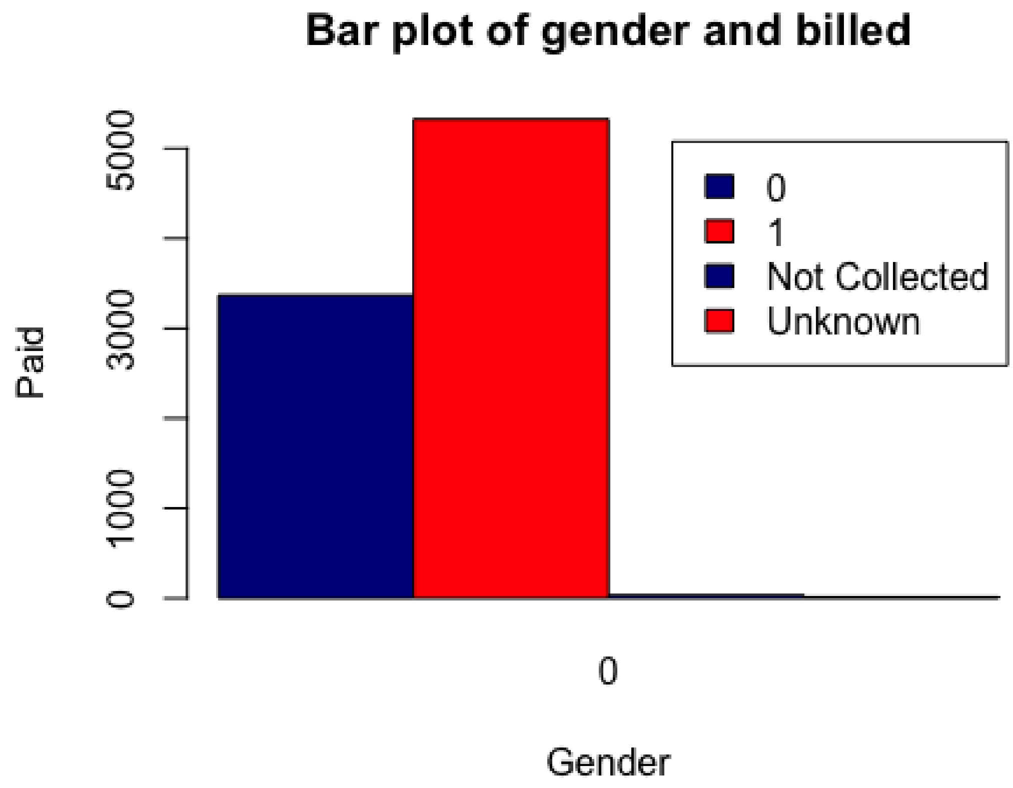


```
# Converting the gender variable for “Male” to 0 and “Female” to 1

df$gender[df$gender =="Male"] <-"0"

df$gender[df$gender=="Female"]<-"1"

# The table value of gender and is_billed is stored in the counts variable wh
ich is further used for the barplot


counts <- table(df$gender, df$is_billed)

# A barplot is plotted between the variables gender and billed status, which
says that Males were billed more when compared with the Females.
```

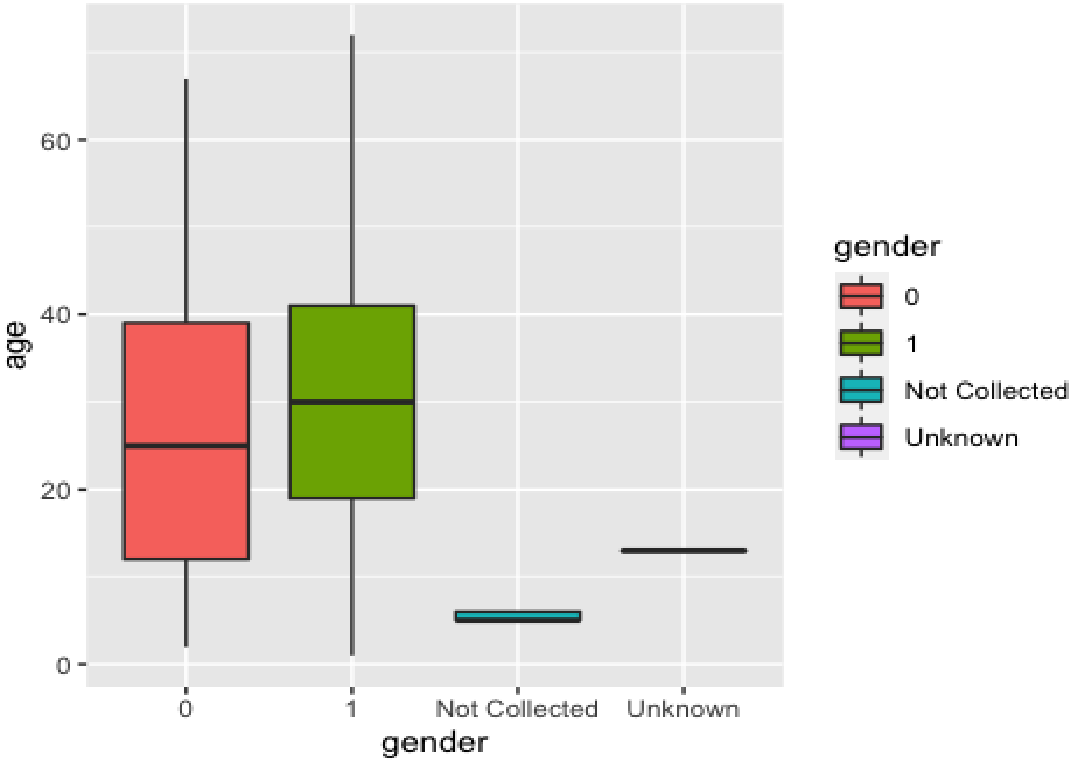
```
barplot(counts, main = "Bar plot of gender and billed",xlab = "Gender", ylab
= "Billed",
        col = c("darkblue","red"),legend = rownames(counts), beside = TRUE)
```



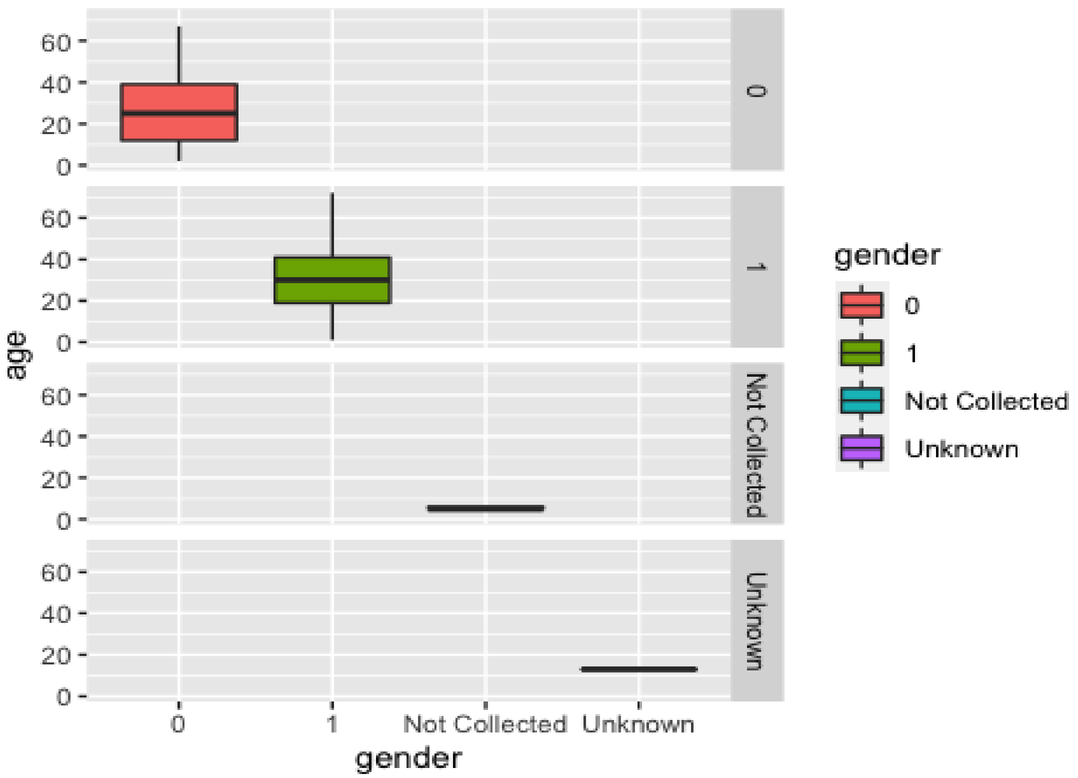
```
#A barplot is plotted between the variables gender and Paid status, which say
s that Males paid more when compared with the Females.

counts_1 <- table(df$gender,df$is_paid)
barplot(counts_1, main = "Bar plot of gender and billed",xlab = "Gender", yla
b = "Paid",
       col = c("darkblue","red"),legend = rownames(counts), beside = TRUE)

library(ggplot2)
```



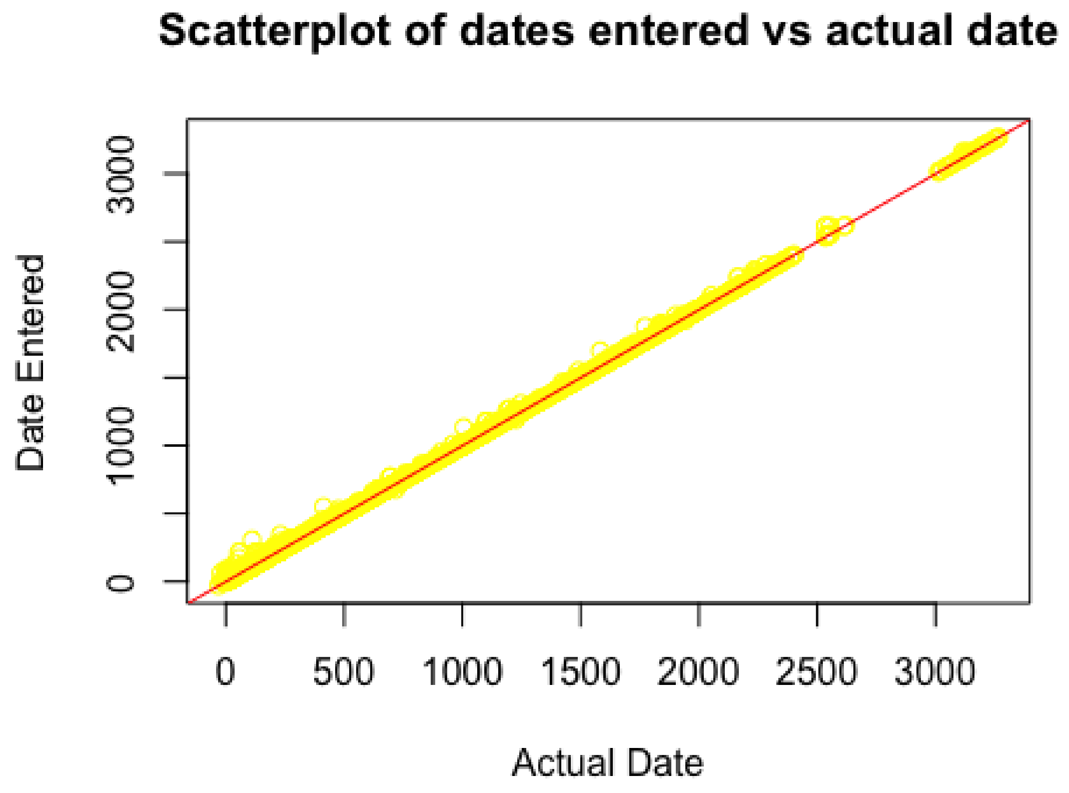
```
#A Box plot is plotted between the variables gender and age, which says that
the median age of Males is 25 whereas the Females is 30. The maximum age of M
ale is ~66 and females is ~66. The first quartile age of Males is ~12, wherea
s Females is 19. The third quartile of Males is ~39 and Females is ~41.


bp <- ggplot(df, aes(x = gender, y = age, group = gender))+
  geom_boxplot(aes(fill = gender))
bp
```



```
#A Facetted plot is plotted between the variables gender and age. FacetGrid t
akes the dataframe as one single input and the name of the variable specified
will be the row of the grid. Facet plot shown below has partitioned the plot
into a matrix of panels. Each panel depicts a different subset of the data.

# With the help of ggplot2 package, we split the gender variable into a facet
grid. Not collected and unknown are the missing values in the data as the dat
a is imbalanced. 0 is Male and 1 is Female as shown in the legend.

bp + facet_grid(gender ~ .)
```



```
# A scatter plot is plotted between the actual_date and the date_entered. The
data is clearly linear between actual_date and the date_entered as the data p
oints lie on the line. This means, the actual date and the date entered on th
e records matches.


plot(df$actual_date, df$date_entered , xlab = "Actual Date", ylab = "Date Ent
ered",
     main = "Scatterplot of dates entered vs actual date", col = "yellow")
abline(lm(actual_date ~ date_entered, data = df), col = "red", pch = 19)
```



```
#The following is some exploratory data analysis I did out of my interest.
Created a table for the gender and is_paid variables. prop.table() function
in R expresses the table entries as proportions for dataframe1 which is
stored in df1.
```

```
dataframe1 <- table(df$gender, df$is_paid)
df1<-prop.table(dataframe1)
df1
```

```
##
##                            0
##   0              0.385363065
##   1              0.609033734
##   Not Collected 0.003773585
##   Unknown        0.001829617
```

```
# The above statistics show that, Males pay 38.53% of the bills, Females pay
60.903% of the bills. The rest of the residual stats are of the missing value
s in the dataset.

ncol(df1)

## [1] 1
```

```
colnames(df1) <- c("The percentage of people who paid")
df1

# Renamed the column of the dataframe to “Percentage of People who paid” to e
nhance the readability for the users.

##
##                  The percentage of people who paid
##   0                                    0.385363065
##   1                                    0.609033734
##   Not Collected                        0.003773585
##   Unknown                              0.001829617
```