R-Assignment2.r

Obaid

Sat Mar 25 17:18:37 2017

```r
# Obaid Ur Rehman

#Loading required libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```r
library(ggplot2)
library(tidyr)
library(mosaic)
```

```
## Loading required package: lattice
```

```
## Loading required package: mosaicData
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add additional features.
## The original behavior of these functions should not be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following objects are masked from 'package:dplyr':
```

```
##
##     count, do, tally

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cov, D, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
#Loading data set from csv file named "hospitaldata.csv"

hdf <- read.csv("D:\\Inbox Workplace\\R Workspace\\R Learning Assignment 2\\R-Assignment-2\\Obaid_Islama
dim(hdf)
```

```
## [1] 222  15
```

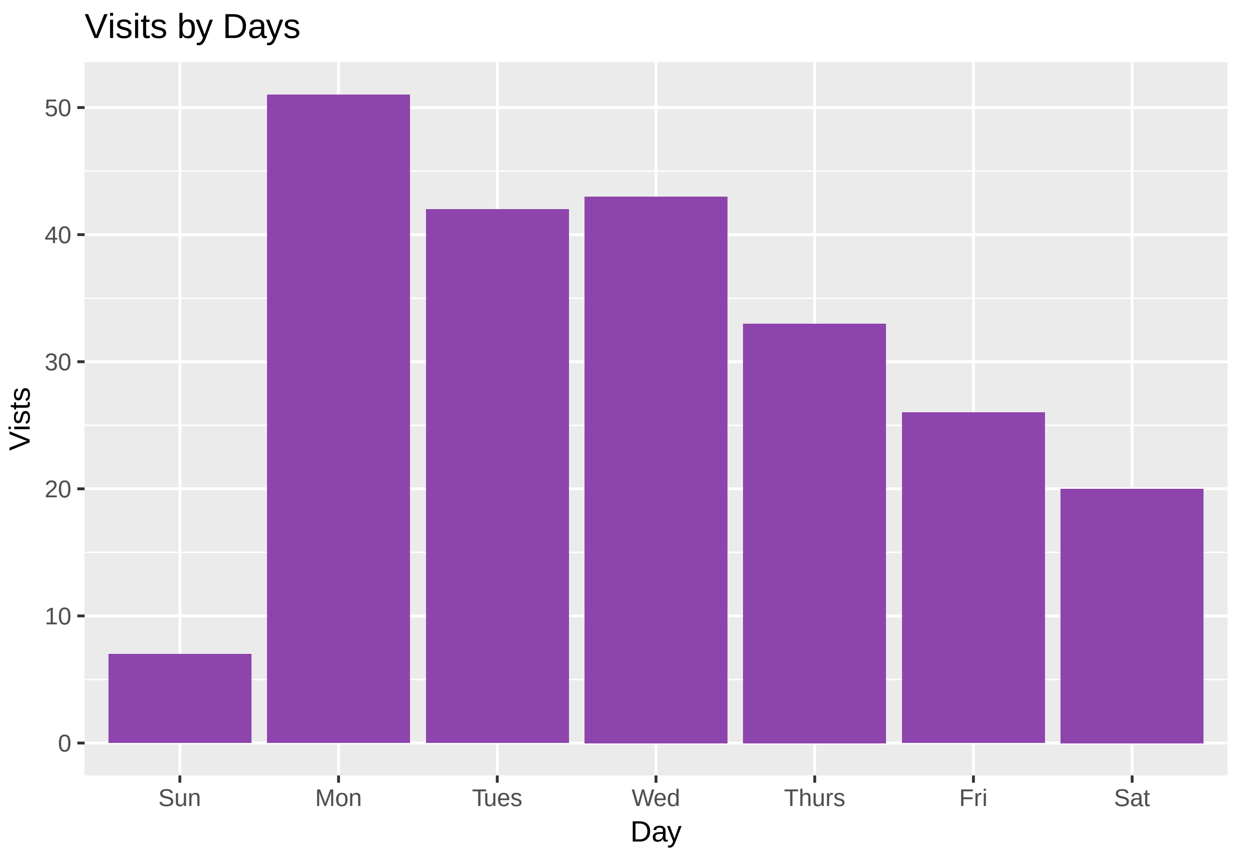
```
# 222 observations and 15 columns

#Printing hdf
head(hdf)
```

```
##                         Date  id    Time Age Sex Consulting..Doctor
## 1 Sunday, January 01, 2017 101   11:00  40   F      Dr Kinza Alam
## 2 Monday, January 02, 2017 150 10:45AM  26   M      Nursing Staff
## 3 Monday, January 02, 2017  58 12:38PM  30   F   Dr Riffat Naheed
## 4 Monday, January 02, 2017  75  1:00PM  40   M   Dr Riffat Naheed
## 5 Monday, January 02, 2017  97  2:45PM  27   M   Dr Riffat Naheed
## 6 Monday, January 02, 2017 101  3:00PM  40   F      Dr Kinza Alam
##         Specialty    Procedure Total..Charges Amount..Received.
## 1           Gynae    C Section          30000             30000
## 2            <NA>     Dressing           1500              1500
## 3 Psychotherapist Consultation           1000              1000
## 4 Psychotherapist Consultation           1500              1500
## 5 Psychotherapist Consultation           2000              2000
## 6           Gynae    C Section          35000             35000
##   Amount..Balance Amount.Received.By Amount.in.Hospital Receptionist..Name
## 1               -         Mrs Shamsa                 NA              Hamza
## 2               -          Dr Saniya                 NA              Haris
## 3               -         Mrs Shamsa                300               Fiza
## 4               -         Mrs Shamsa                450             Zaheer
## 5               -         Mrs Shamsa                600              Haris
## 6               -          Dr Saniya                 NA              Haris
##   Next.Apt
## 1     <NA>
## 2     <NA>
## 3     <NA>
## 4     <NA>
## 5     <NA>
## 6     <NA>
```

```
# Q1. Cleaning the column names
names(hdf)<-gsub("\\.","",names(hdf))
head(hdf) #dots from column names removed
```

```
##                         Date  id    Time Age Sex ConsultingDoctor
## 1 Sunday, January 01, 2017 101   11:00  40   F    Dr Kinza Alam
## 2 Monday, January 02, 2017 150 10:45AM  26   M    Nursing Staff
## 3 Monday, January 02, 2017  58 12:38PM  30   F Dr Riffat Naheed
## 4 Monday, January 02, 2017  75  1:00PM  40   M Dr Riffat Naheed
## 5 Monday, January 02, 2017  97  2:45PM  27   M Dr Riffat Naheed
## 6 Monday, January 02, 2017 101  3:00PM  40   F    Dr Kinza Alam
##         Specialty    Procedure TotalCharges AmountReceived AmountBalance
## 1           Gynae    C Section        30000          30000          -
## 2            <NA>     Dressing         1500           1500          -
## 3 Psychotherapist Consultation         1000           1000          -
## 4 Psychotherapist Consultation         1500           1500          -
## 5 Psychotherapist Consultation         2000           2000          -
## 6           Gynae    C Section        35000          35000          -
##   AmountReceivedBy AmountinHospital ReceptionistName NextApt
## 1      Mrs Shamsa                NA            Hamza    <NA>
## 2       Dr Saniya                NA            Haris    <NA>
## 3      Mrs Shamsa               300             Fiza    <NA>
## 4      Mrs Shamsa               450           Zaheer    <NA>
## 5      Mrs Shamsa               600            Haris    <NA>
## 6       Dr Saniya                NA            Haris    <NA>
```

```r
# Q2. Which day of the week is expected to have most visits?
dayPop <-
  hdf %>%
  mutate(Day=wday(mdy(Date),label=TRUE)) %>%
  group_by(Day) %>%
  summarize(visits=length(Day))

ggplot(dayPop,aes(x=Day,y=visits))+geom_bar(stat="identity",fill="#8E44AD")+ggtitle("Visits by Days")+la
```



```
#The visits on Monday are greater than visits on other days of week, and also the probability of Monday
# therefore, Monday is expected to have most visits

# Q3. What is the average age of patients?
hdfClean<- hdf
hdfClean$Age <-as.numeric(as.character(hdfClean$Age))
```

```
## Warning: NAs introduced by coercion
```

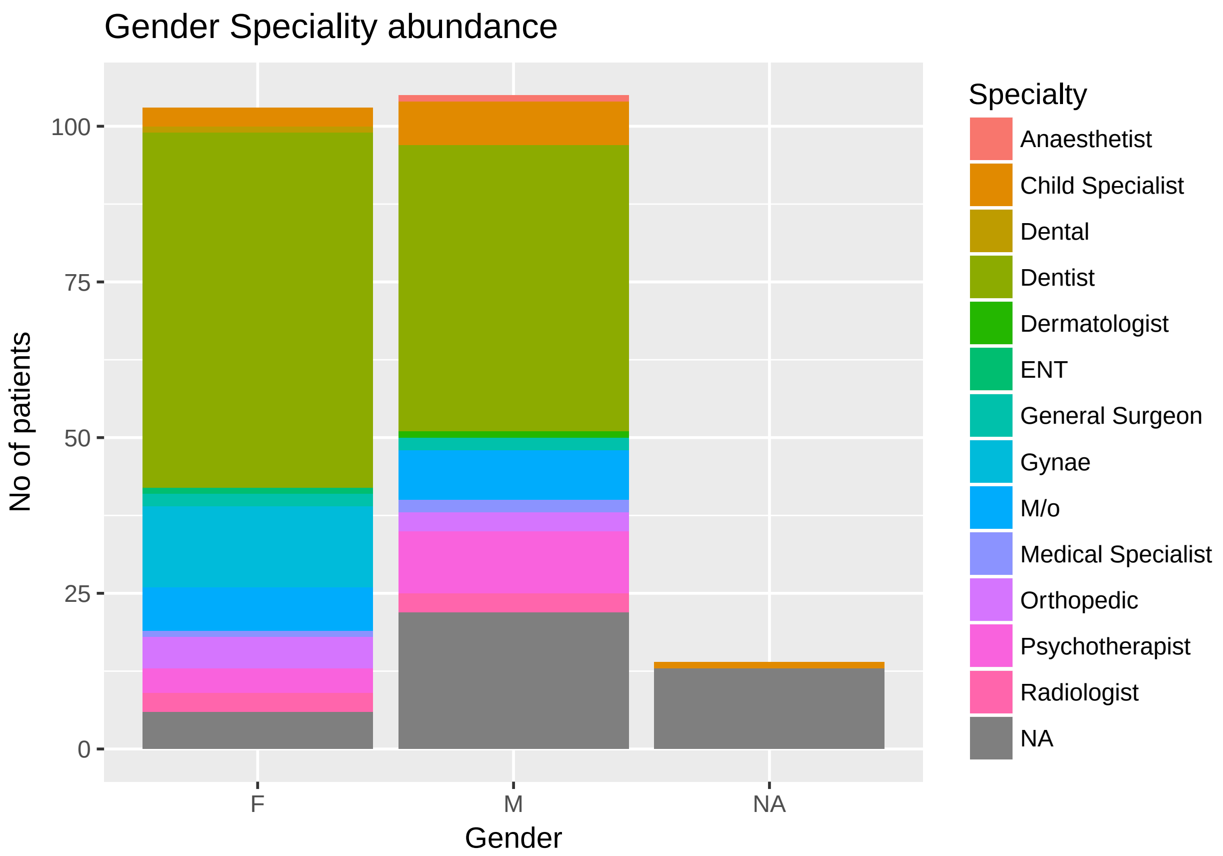
```
mean(hdfClean$Age,na.rm = TRUE) #Average age is 32.7
```

```
## [1] 32.73438
```

```
# Q4. How many childerns were entertained?
count(filter(hdfClean,Age>=1,Age<=12))  #23 childerns were entertained    #Q to ask, if i use length in
```

```
## 1.
## 23
```

```
# Q5. Which gender type had what kind of procedure in abundance?
hdfClean$Sex <- gsub("f","F",hdfClean$Sex)
hdfClean$Sex<-gsub("\\s|-",NA,hdfClean$Sex)
qplot(data=hdfClean,Sex,fill=Specialty)+ggtitle("Gender Speciality abundance")+labs(x='Gender',y='No of
```



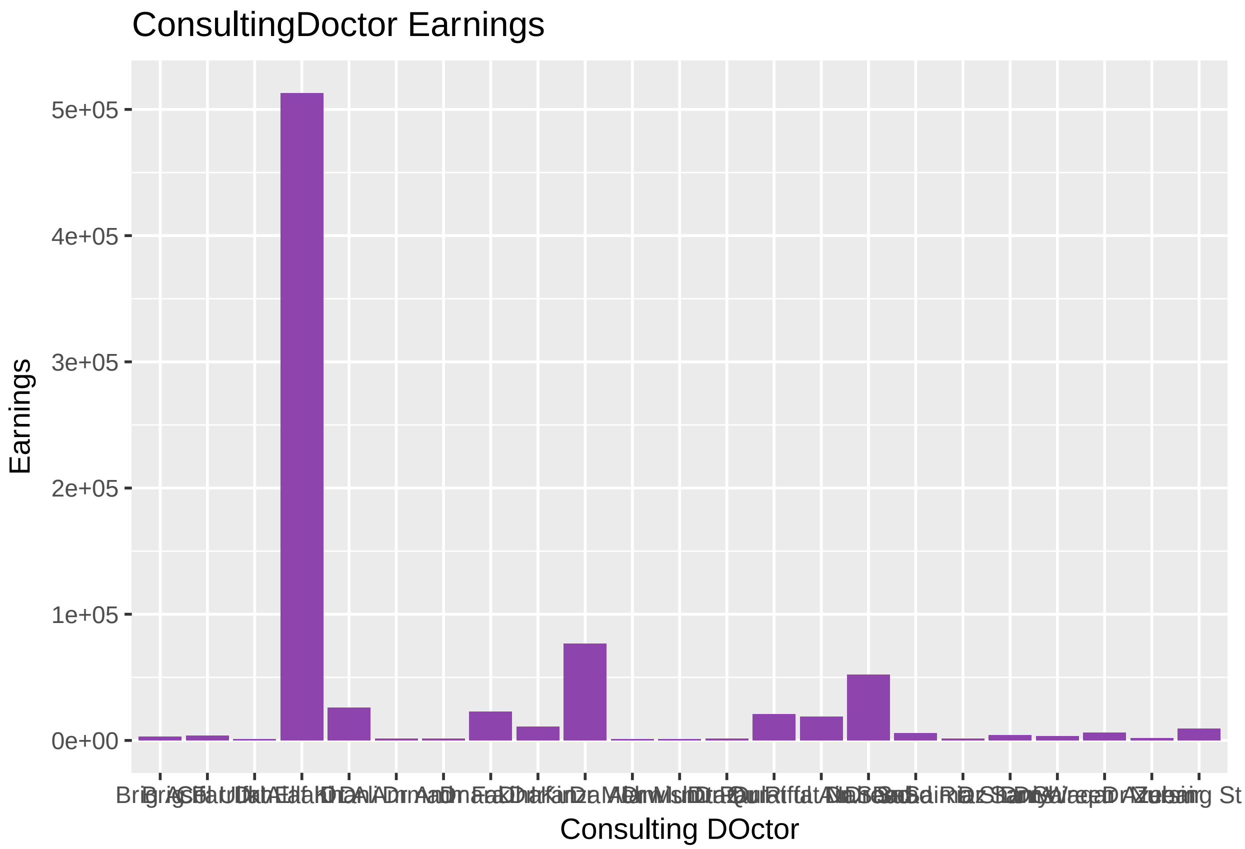
```
# As we can see from plot, both Male and Female have Dentist procedure in abundance



# Q6. Which doctor is earning highest?

#Cleaning totalCharges column (we will need in future to summ charges) by Converting them to numeric an
hdfClean$TotalCharges <- as.numeric(as.character(hdfClean$TotalCharges))
```

```
## Warning: NAs introduced by coercion
```

```
hdfClean[c('TotalCharges')][is.na(hdfClean[c('TotalCharges')])]<-0  #only chnage NA to 0 in TotalCharge
DrEarnings <-
  hdfClean %>%
  group_by(ConsultingDoctor)%>%
  summarize(Earning=sum(TotalCharges)) %>%
  arrange(desc(Earning))

DrEarnings # Dr Alaf Khan has the highest earnings!
```

```
## # A tibble: 23 × 2
##    ConsultingDoctor Earning
##              <fctr>   <dbl>
## 1      Dr Alaf Khan  513050
## 2     Dr Kinza Alam   76700
## 3           Dr Saad   52000
## 4            Dr Ali   26100
## 5          Dr Fakiha   22600
## 6   Dr Qurat ul Ain   20900
## 7  Dr Riffat Naheed   18800
## 8           Dr Irfan   11000
## 9      Nursing Staff    9150
## 10   Dr Waqar Azeem    6000
## # ... with 13 more rows
```

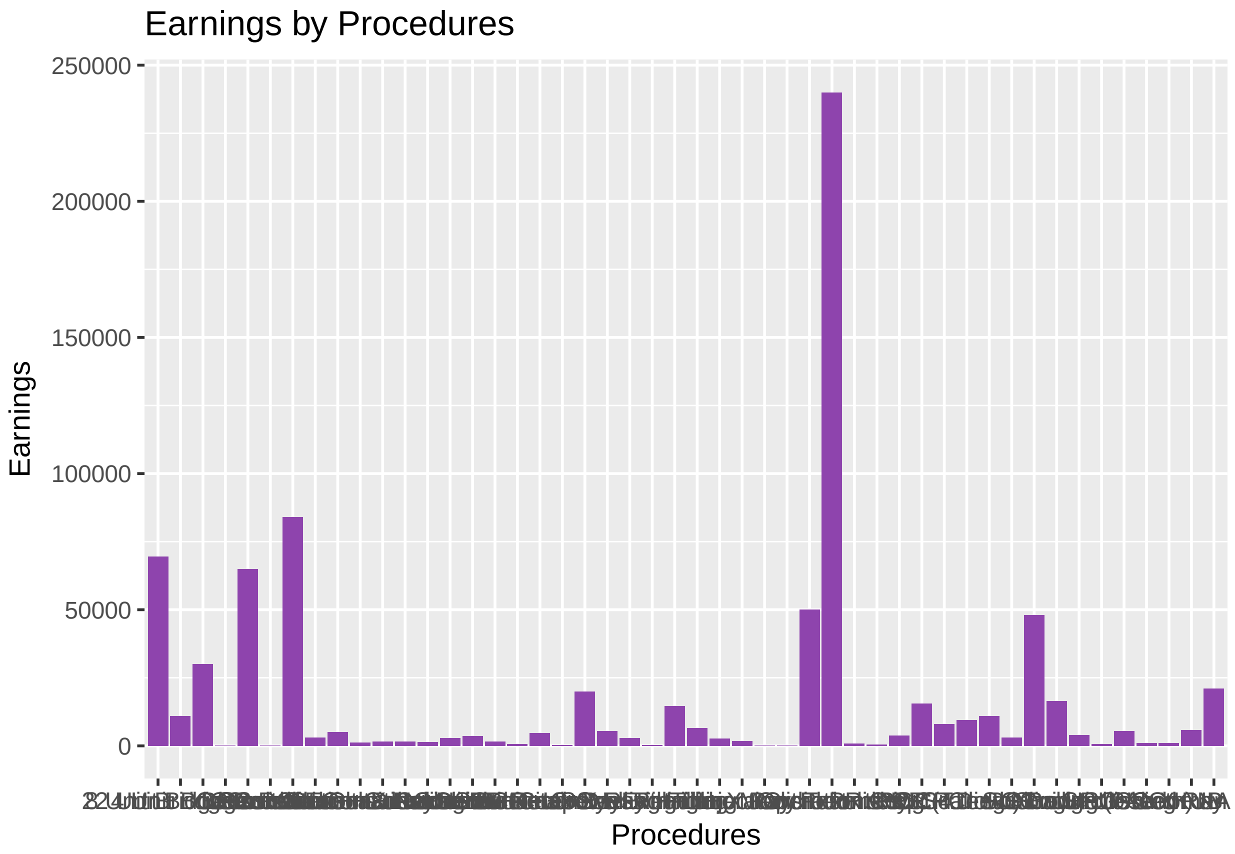
```
#Plottig graph for DoctorEarnings
ggplot(data=DrEarnings,aes(x=ConsultingDoctor,y=Earning))+geom_bar(stat='identity',fill='#8E44AD')+ggti
```



```
# Q7. Which procedure type earns more money?

#its same as above Question, jut need to group_by with Procedur instead of ConsultingDoctor
# We dont need to clean totalcharges column again

ProcedureEarnings <-
  hdfClean %>%
  group_by(Procedure) %>%
  summarize(Earning=sum(TotalCharges)) %>%
  arrange(desc(Earning))
ProcedureEarnings  #Orthodontics earns more money
```

```
## # A tibble: 48 × 2
##                         Procedure Earning
##                            <fctr>   <dbl>
## 1                    Orthodontics  240000
## 2                    Consultation   83950
## 3                  22 Unit Bridge   69500
## 4                       C Section   65000
## 5                       Operation   50000
## 6  RCT (4 teeth) Bridge (9 teeth)   48000
## 7            8 Unit Bridge+2 R.C.T   30000
## 8                              NA   21000
## 9                           Crown   20000
## 10                       Scalling   16500
## # ... with 38 more rows
```

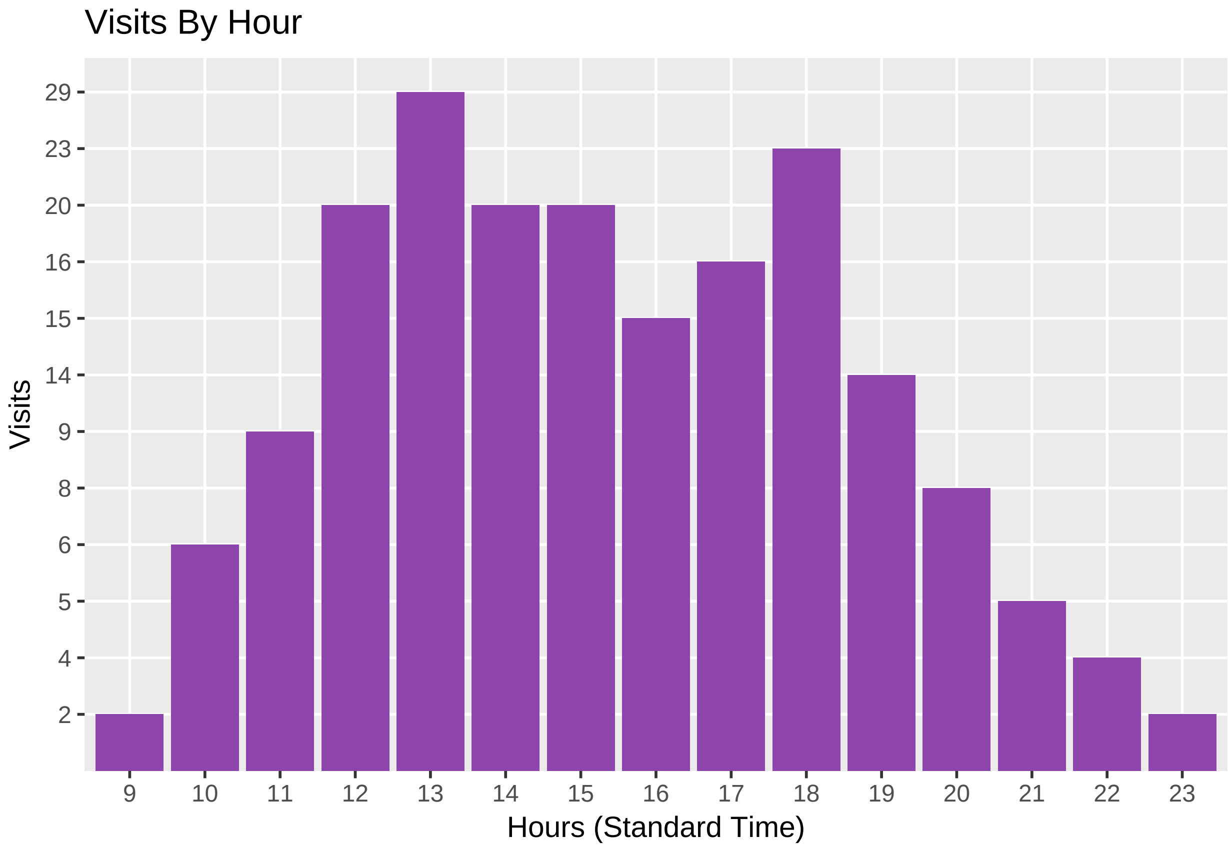
```
#Plotting graph for ProcedureEarnings
ggplot(data=ProcedureEarnings,aes(x=Procedure,y=Earning))+geom_bar(stat='identity',fill='#8E44AD')+ggti
```



```
# Q8. Which time of day has highest frequency of visits by hours

#Creating a column Hour
VisitsByHour <-
  hdfClean %>%
  select(Time) %>%
  mutate(Hour = hour(hm(format(strptime(hdfClean$Time, "%I:%M %p"), "%H:%M")))) %>%
  group_by(Hour) %>%
  summarize(Visits=length(Hour)) %>%
  arrange(desc(Visits))%>%
  filter(!is.na(Hour))
VisitsByHour # it seems at 1:00PM (13:00), the visits are maximum. The Hour for 2nd highest is sadly NA
```

```
## # A tibble: 15 × 2
##     Hour Visits
##    <dbl>  <int>
## 1     13     29
## 2     18     23
## 3     12     20
## 4     14     20
## 5     15     20
## 6     17     16
## 7     16     15
## 8     19     14
## 9     11      9
## 10    20      8
## 11    10      6
## 12    21      5
## 13    22      4
## 14     9      2
## 15    23      2
```

```
#plotting
ggplot(data=VisitsByHour,aes(x=factor(Hour),y=factor(Visits)))+geom_bar(stat='identity',fill='#8E44AD')
```



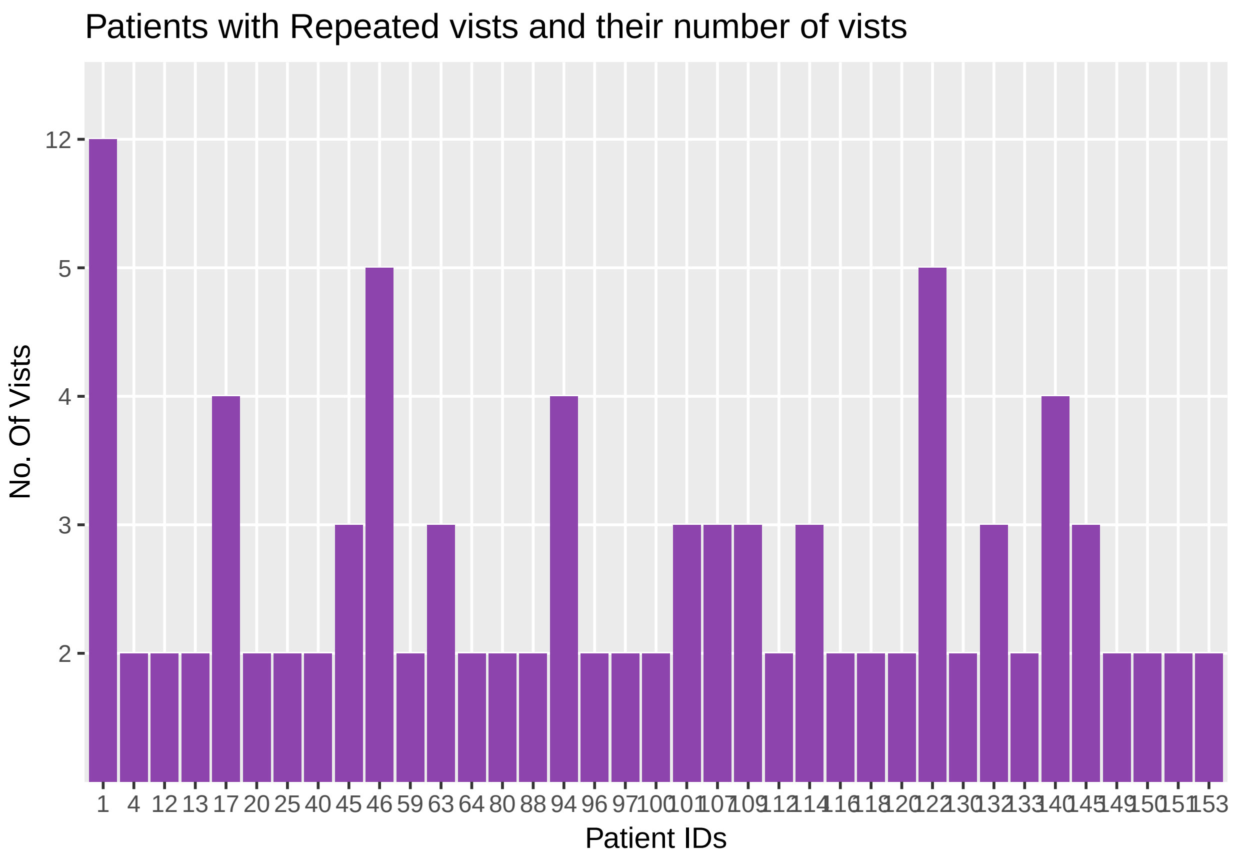
```
# Q9. Create a bracket of time

#Create column hour in hdfClean
hdfClean <-
  hdfClean %>%
  mutate(Hour = hour(hm(format(strptime(Time,"%I:%M %p"),format="%H:%M"))))

hdfClean <-
  hdfClean %>%
  mutate( Bracket = derivedFactor(
    "Morning" = (Hour>=6 & Hour<=12),
    "Afternoon" = (Hour>=12 & Hour<=16),
    "Evening" = (Hour>=14 & Hour<=19),
    "Night" =((Hour>=19 & Hour<=23) | (Hour>=0 & Hour<=6) ),
    .method = "first",
    .default = 0
  ))
select(hdfClean,Time,Hour,Bracket)
```

```
##        Time Hour   Bracket
## 1     11:00   NA      <NA>
## 2   10:45AM   10   Morning
## 3   12:38PM   12   Morning
## 4    1:00PM   13 Afternoon
## 5    2:45PM   14 Afternoon
## 6    3:00PM   15 Afternoon
## 7    3:28PM   15 Afternoon
## 8    3:45PM   15 Afternoon
## 9    3:45PM   15 Afternoon
## 10   5:00PM   17   Evening
## 11   5:00PM   17   Evening
## 12   5:30PM   17   Evening
## 13   1:00PM   13 Afternoon
## 14   3:25PM   15 Afternoon
## 15   6:10PM   18   Evening
## 16  11:45PM   23     Night
## 17  12:40PM   12   Morning
## 18   8:10PM   20     Night
## 19   8:30PM   20     Night
## 20  12:40PM   12   Morning
## 21   2:00PM   14 Afternoon
## 22   2:00PM   14 Afternoon
## 23  12:30PM   12   Morning
## 24   1:00PM   13 Afternoon
## 25   1:30PM   13 Afternoon
## 26        -   NA      <NA>
## 27   8:15PM   20     Night
## 28     <NA>   NA      <NA>
## 29  12:36PM   12   Morning
## 30   1:30PM   13 Afternoon
## 31   2:30PM   14 Afternoon
## 32   3:15PM   15 Afternoon
## 33   5:20PM   17   Evening
## 34   5:30PM   17   Evening
## 35   3:50PM   15 Afternoon
## 36   6:00PM   18   Evening
## 37     <NA>   NA      <NA>
## 38     <NA>   NA      <NA>
## 39   3:00PM   15 Afternoon
## 40   4:30PM   16 Afternoon
## 41   4:30PM   16 Afternoon
## 42  10:45AM   10   Morning
## 43  02:00PM   14 Afternoon
## 44  02:00PM   14 Afternoon
## 45  11:20AM   11   Morning
## 46   3:00PM   15 Afternoon
## 47   8:00PM   20     Night
## 48   4:30PM   16 Afternoon
## 49   6:30PM   18   Evening
## 50   9:00PM   21     Night
## 51     <NA>   NA      <NA>
## 52   1:30PM   13 Afternoon
## 53   6:00PM   18   Evening
## 54   6:20PM   18   Evening
## 55  11:25AM   11   Morning
## 56  11:15AM   11   Morning
## 57   1:10PM   13 Afternoon
## 58   3:30PM   15 Afternoon
## 59   6:15PM   18   Evening
## 60   9:40PM   21     Night
## 61  12:00PM   12   Morning
## 62   2:00PM   14 Afternoon
## 63   5:00PM   17   Evening
## 64     <NA>   NA      <NA>
## 65  11:00AM   11   Morning
## 66     <NA>   NA      <NA>
## 67     <NA>   NA      <NA>
## 68     <NA>   NA      <NA>
## 69  10:15AM   10   Morning
## 70   1:20PM   13 Afternoon
## 71   1:30PM   13 Afternoon
## 72  12:15PM   12   Morning
## 73   1:00PM   13 Afternoon
## 74   1:15PM   13 Afternoon
## 75   4:50PM   16 Afternoon
## 76   1:00PM   13 Afternoon
## 77   1:15PM   13 Afternoon
## 78   2:10PM   14 Afternoon
## 79   1:30PM   13 Afternoon
## 80     <NA>   NA      <NA>
## 81     <NA>   NA      <NA>
## 82  12:50PM   12   Morning
## 83   3:30PM   15 Afternoon
## 84   5:40PM   17   Evening
## 85     <NA>   NA      <NA>
## 86     <NA>   NA      <NA>
## 87     <NA>   NA      <NA>
## 88     <NA>   NA      <NA>
## 89   6:45PM   18   Evening
## 90   9:45PM   21     Night
## 91     <NA>   NA      <NA>
## 92   1:00PM   13 Afternoon
## 93   1:30PM   13 Afternoon
## 94   5:40PM   17   Evening
## 95   5:35PM   17   Evening
## 96   6:00PM   18   Evening
## 97   5:30PM   17   Evening
## 98   6:30PM   18   Evening
## 99   6:50PM   18   Evening
## 100  2:10PM   14 Afternoon
## 101  2:10PM   14 Afternoon
## 102  1:00PM   13 Afternoon
## 103  1:40PM   13 Afternoon
## 104  6:00PM   18   Evening
## 105 12:00PM   12   Morning
## 106  1:00PM   13 Afternoon
## 107  1:25PM   13 Afternoon
## 108  4:45PM   16 Afternoon
## 109  8:00PM   20     Night
## 110  4:00PM   16 Afternoon
## 111  4:00PM   16 Afternoon
## 112  7:30PM   19   Evening
## 113  7:45PM   19   Evening
## 114  1:30PM   13 Afternoon
## 115  1:30PM   13 Afternoon
## 116  4:00PM   16 Afternoon
## 117  6:15PM   18   Evening
## 118 12:00PM   12   Morning
## 119  1:10PM   13 Afternoon
## 120  2:15PM   14 Afternoon
## 121  6:00PM   18   Evening
## 122  8:00PM   20     Night
## 123 10:13AM   10   Morning
## 124 12:00PM   12   Morning
## 125 12:00PM   12   Morning
## 126  2:40PM   14 Afternoon
## 127  2:40PM   14 Afternoon
## 128  2:40PM   14 Afternoon
## 129 10:00AM   10   Morning
## 130  9:30AM    9   Morning
## 131  6:30PM   18   Evening
## 132  7:00PM   19   Evening
## 133 12:00PM   12   Morning
## 134  4:20PM   16 Afternoon
## 135  5:57PM   17   Evening
## 136  6:15PM   18   Evening
## 137  7:15PM   19   Evening
## 138 12:00PM   12   Morning
## 139 11:20AM   11   Morning
## 140  3:40PM   15 Afternoon
## 141  7:00PM   19   Evening
## 142    <NA>   NA      <NA>
## 143  2:30PM   14 Afternoon
## 144  3:00PM   15 Afternoon
## 145  7:02PM   19   Evening
## 146 11:40AM   11   Morning
## 147  4:45PM   16 Afternoon
## 148  6:15PM   18   Evening
## 149  4:10PM   16 Afternoon
## 150  5:30PM   17   Evening
## 151  6:30PM   18   Evening
## 152  6:20PM   18   Evening
## 153  6:10PM   18   Evening
## 154 11:30AM   11   Morning
## 155  2:45PM   14 Afternoon
## 156    <NA>   NA      <NA>
## 157  1:25PM   13 Afternoon
## 158  2:00PM   14 Afternoon
## 159  7:00PM   19   Evening
## 160 10:15PM   22     Night
## 161  1:00PM   13 Afternoon
## 162  6:00PM   18   Evening
## 163  7:11PM   19   Evening
## 164 10:10PM   22     Night
## 165       -   NA      <NA>
## 166  3:00PM   15 Afternoon
## 167  4:30PM   16 Afternoon
## 168  5:00PM   17   Evening
## 169  1:55PM   13 Afternoon
## 170  1:50PM   13 Afternoon
## 171  2:00PM   14 Afternoon
## 172  3:00PM   15 Afternoon
## 173  9:30PM   21     Night
## 174  3:45PM   15 Afternoon
## 175  4:00PM   16 Afternoon
## 176 11:30AM   11   Morning
## 177 12:20PM   12   Morning
## 178       -   NA      <NA>
## 179 10:30PM   22     Night
## 180 12:40PM   12   Morning
## 181    <NA>   NA      <NA>
## 182  3:00PM   15 Afternoon
## 183  8:00PM   20     Night
## 184  5:00PM   17   Evening
## 185  6:00PM   18   Evening
## 186       -   NA      <NA>
## 187  7:00PM   19   Evening
## 188  7:10PM   19   Evening
## 189 12:48PM   12   Morning
## 190  3:00PM   15 Afternoon
## 191  7:05PM   19   Evening
## 192       -   NA      <NA>
## 193 11:20AM   11   Morning
## 194 12:30PM   12   Morning
## 195  1:30PM   13 Afternoon
## 196  4:10PM   16 Afternoon
## 197  5:45PM   17   Evening
## 198  2:40PM   14 Afternoon
## 199       -   NA      <NA>
## 200  1:20PM   13 Afternoon
## 201  5:30PM   17   Evening
## 202  7:00PM   19   Evening
## 203        -   NA      <NA>
## 204  3:00PM   15 Afternoon
## 205        -   NA      <NA>
## 206  7:40PM   19   Evening
## 207  2:00PM   14 Afternoon
## 208  9:35PM   21     Night
## 209  8:30PM   20     Night
## 210 10:00PM   22     Night
## 211  4:45PM   16 Afternoon
## 212  6:55PM   18   Evening
## 213 12:00PM   12   Morning
## 214  7:30PM   19   Evening
## 215 12:00PM   12   Morning
## 216  9:00AM    9   Morning
## 217    <NA>   NA      <NA>
## 218    <NA>   NA      <NA>
## 219  3:30PM   15 Afternoon
## 220  6:00PM   18   Evening
## 221 10:20AM   10   Morning
## 222 11:20PM   23     Night
```

```
# Q10.  How many patients are repeated visitor?
repPat <-
  select(hdfClean,id) %>%
  group_by(id) %>%
  summarize(visits=length(id)) %>%
  arrange(desc(visits)) %>%
  filter(visits >1)
dim(repPat)  #37 Patients have more than one visits. Paient with id= 1 is very unfortunate, with 12 vis
```

```
## [1] 37  2
```

```
#plotting
ggplot(data=repPat,aes(x=factor(id),y=factor(visits)))+geom_bar(stat='identity',fill='#8E44AD')+ggtitle
```



```
# Q11.  Give the id of repeated visitors
ids<-
  repPat %>%
  select(id)
ids #Shows the id(s) of repeated patients
```

```
## # A tibble: 37 × 1
##       id
##    <int>
## 1      1
## 2     46
## 3    122
## 4     17
## 5     94
## 6    140
## 7     45
## 8     63
## 9    101
## 10   107
## # ... with 27 more rows
```

```
# Q12.  Which patients visited again for the same problem?
samep <-
  hdfClean %>%
  select(id,Specialty) %>%
  group_by(id) %>%
  summarize(problems=n_distinct(Specialty), visits=length(Specialty))%>%
  filter(visits>problems)
samep
```

```
## # A tibble: 29 × 3
##       id problems visits
##    <int>    <int>  <int>
## 1      1        1     12
## 2     12        1      2
## 3     13        1      2
## 4     17        3      4
## 5     25        1      2
## 6     40        1      2
## 7     45        1      3
## 8     46        1      5
## 9     63        2      3
## 10    88        1      2
## # ... with 19 more rows
```

```
# The above tabel sgow the id, and no of distinct problems that patient have and no of visits patient m
# so if, no of visits is greater than no of problems patient have this means patient have  come more th
# he got

#Q13. What os median age for female and male?
medianAge<-
  hdfClean %>%
  select(Sex,Age) %>%
  group_by(Sex) %>%
  summarize(MedianAge = median(Age,na.rm=TRUE))
medianAge # Shows the median age for Female(F) and Male(M)
```

```
## # A tibble: 3 × 2
##     Sex MedianAge
##   <chr>     <dbl>
## 1     F        30
## 2     M        29
## 3  <NA>        NA
```

```
# Q14.  What is the total amount in balance?
hdfClean$AmountBalance <-gsub("\\.00|,","",hdfClean$AmountBalance)
hdfClean$AmountBalance <-as.numeric(as.character(hdfClean$AmountBalance))
```

```
## Warning: NAs introduced by coercion
```

```
sum(hdfClean$AmountBalance,na.rm=TRUE) #222500
```

```
## [1] 222500
```

```
# Q15.  How much money was made by Procedure Type "Consultation"?

#cleaning TotalCharges column
hdfClean$TotalCharges <- as.numeric(as.character(hdfClean$TotalCharges))
consult <-
  hdfClean %>%
  select(Procedure,TotalCharges) %>%
  group_by(Procedure) %>%
  filter(Procedure =='Consultation') %>%
  summarize(TotalMoney= sum(TotalCharges,na.rm=TRUE))
```

```
consult #83950
```

```
## # A tibble: 1 × 2
##       Procedure TotalMoney
##          <fctr>      <dbl>
## 1 Consultation      83950
```

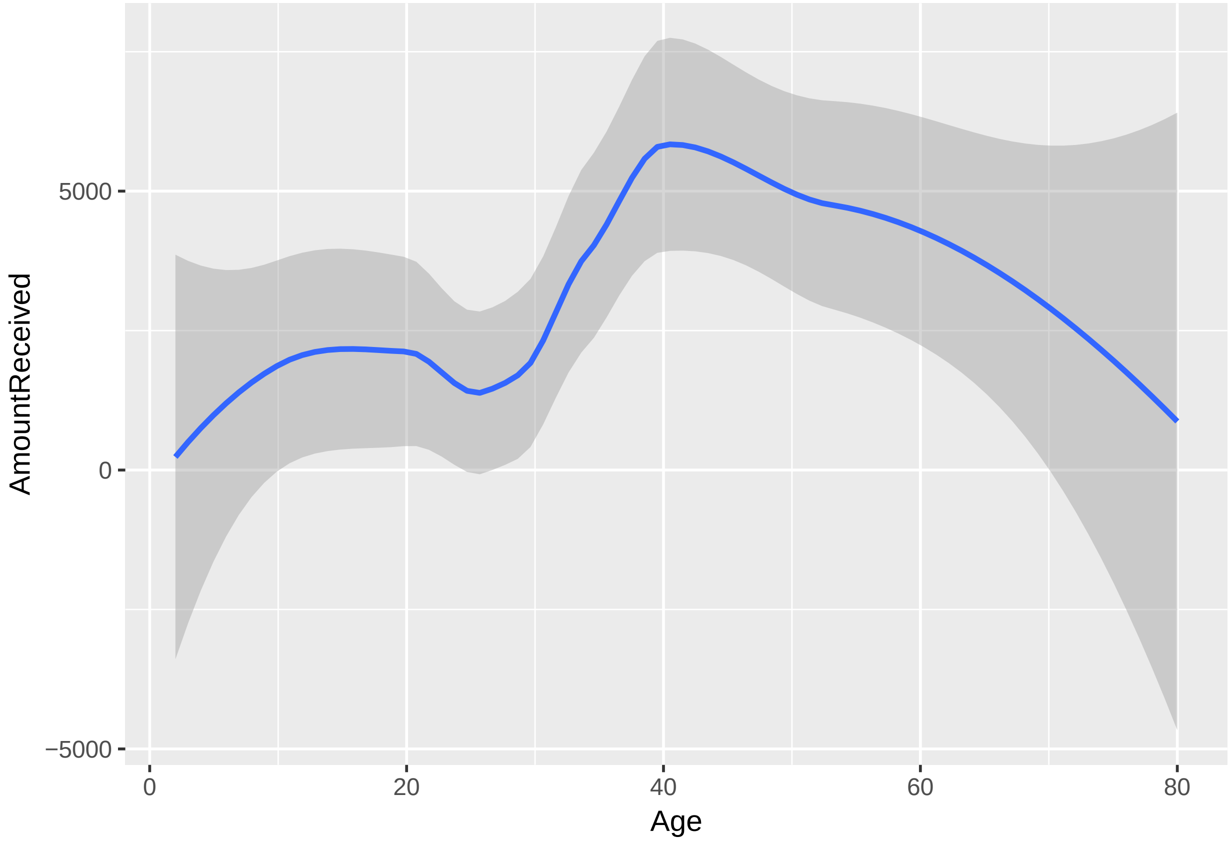
```
# Q16.  Is there any relation between Age and Total charges paid?
cor<-cor(hdfClean$Age,hdfClean$AmountReceived, use='complete.obs')  #use is to ignore NA values
cor
```

```
## [1] 0.1316023
```

```
# The answer is 0.13 which tell us that there is a weak positive (uphill) relation.
ggplot(data=hdfClean,aes(x=Age,y=AmountReceived))+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```

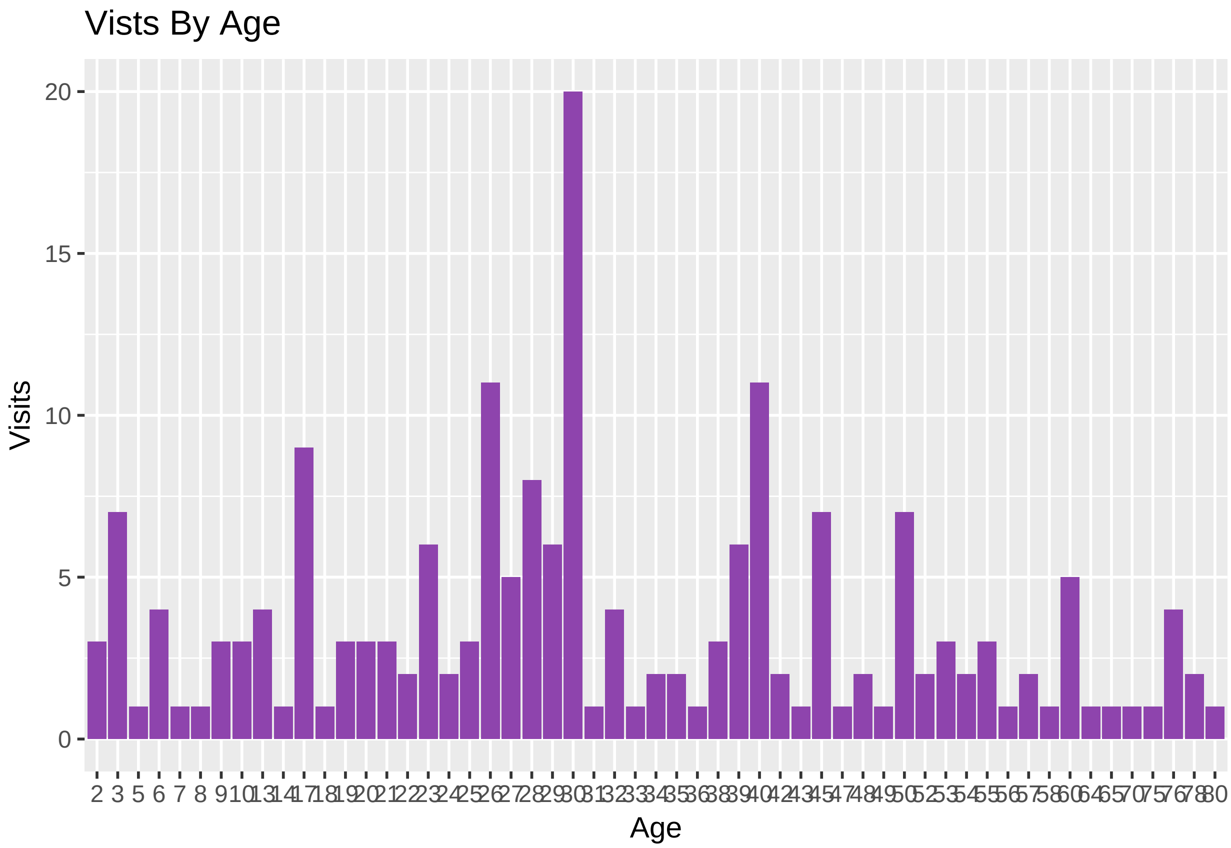
```
## Warning: Removed 38 rows containing non-finite values (stat_smooth).
```



```
# As we can see for plot, There exists a relation but is very weak linear relationship. We usually dnt


# Q17.  Which age group had highest number of visits?
ageVisits<-
  hdfClean %>%
  select(id,Age) %>%
  group_by(Age) %>%
  summarize(Visits=length(Age)) %>%
  arrange(desc(Visits)) %>%
  filter(!is.na(Age))
ageVisits
```

```
## # A tibble: 55 × 2
##       Age Visits
##     <dbl>  <int>
## 1      30     20
## 2      26     11
## 3      40     11
## 4      17      9
## 5      28      8
## 6       3      7
## 7      45      7
## 8      50      7
## 9      23      6
## 10     29      6
## # ... with 45 more rows
```

```
#Plotting
ggplot(data=ageVisits,aes(x=factor(Age),y=Visits))+geom_bar(stat='identity',fill='#8E44AD')+ggtitle("Vi
```



```
#As we can see, Most no of vists are 30 but the Age is NA so we dont include that. After that, patient


# Q18.  What is the total cost earned by Procedure Type X Ray and Scalling together?
earning <-
```

```
  hdfClean %>%
  select(Procedure,TotalCharges) %>%
  filter(Procedure=='X Ray'|Procedure=='Scalling') %>%
  group_by(Procedure) %>%
  summarize(Occurance=n(), Earning=sum(TotalCharges))
earning
```

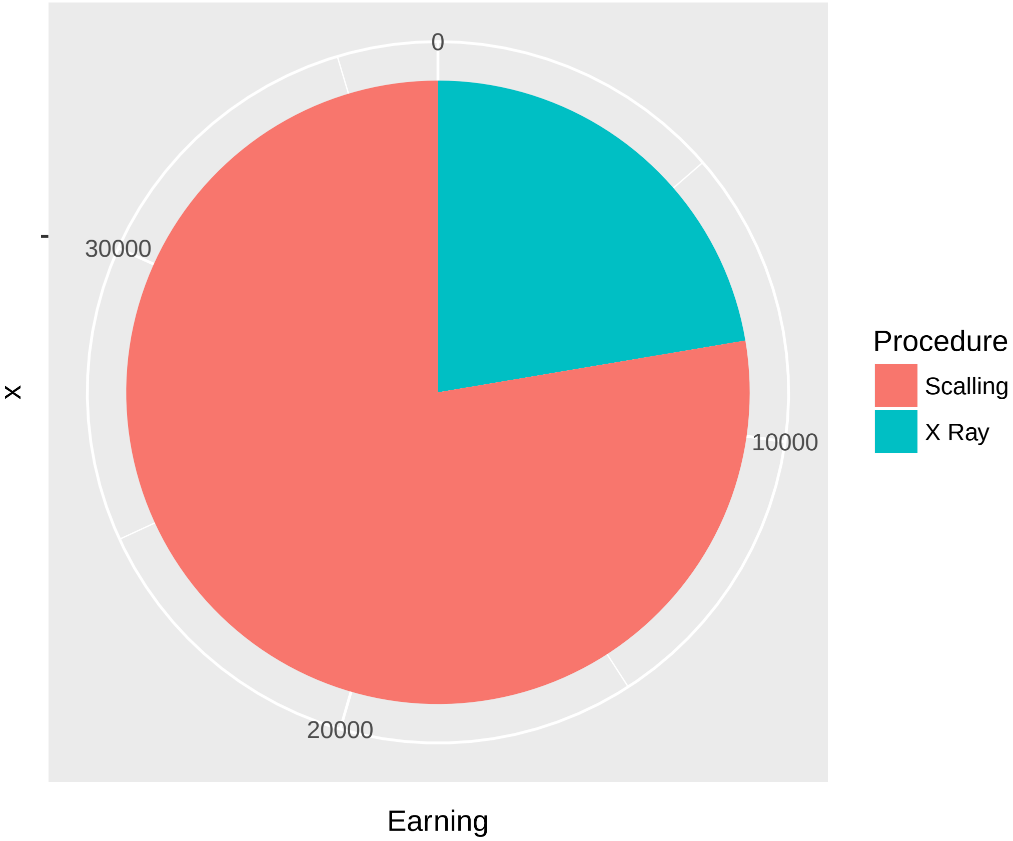
```
## # A tibble: 2 × 3
##   Procedure Occurance Earning
##      <fctr>     <int>   <dbl>
## 1  Scalling         6   16500
## 2     X Ray        15    5800
```

```
# Scalling = 16500, X Ray = 5800.
# As we can see from results, x Ray occured more than Scalling and still earned less than scalling. One
#that the XRay fee is less than Scalling fee.

#BUTTTTTT!!!, there are procedures in which xray was done along with some other procedure :same for sca
# now for better results, we dig deep

earning2 <-
  hdfClean %>%
  select(Procedure,TotalCharges) %>%
  filter(  grepl("X Ray",Procedure) | grepl("Scalling",Procedure),nchar(as.character(Procedure))>8) %>%
  mutate(Procedure= derivedFactor(
    "X Ray" = (grepl("X Ray",Procedure)==TRUE),
    "Scalling" = (grepl("Scalling",Procedure)==TRUE),
    .method = "first",
    .default = 0
  ),
  TotalCharges=derivedFactor(
    "300" = (Procedure =='X Ray'),
    "3000" = (Procedure =='Scalling'),
    .method = "first",
    .default = 0
  ))  %>%
  group_by(Procedure) %>%
  summarize(Occurance=n(),Earning=sum(as.numeric(as.character(TotalCharges))))


totalEarnings <-
  rbind(earning,earning2) %>%
  group_by(Procedure) %>%
  summarize(Occurance=sum(Occurance),Earning=sum(as.numeric(as.character(Earning))))
totalEarnings
```

```
## # A tibble: 2 × 3
##   Procedure Occurance Earning
##      <fctr>     <int>   <dbl>
## 1  Scalling        10   28500
## 2     X Ray        23    8200
```

```
# So that totalEarnings Show the actual earning by X Ray and Scalling , and their occurance in the whol

#lets Plot this
```

```
ggplot(data=totalEarnings,aes(x='',y=Earning,fill=Procedure))+geom_bar(width=1,stat='identity')+coord_po
```



```
#Generating csv file from cleaned data
write.csv(hdfClean, 'D:/Inbox Workplace/R Workspace/R Learning Assignment 2/R-Assignment-2/Obaid_Islamab
```