R_assg_2.R

Syed.Adeel

Mon Mar 27 15:32:31 2017

```
df <-read.csv("E://DIH//Assignment2//hospitaldata.csv", strip.white = T, na.strings = c("-","","  ","\t"
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
```

```
library("lubridate")
```

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

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

```
library("ggplot2")
View(df)
head(df)
```

```
##                       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            <NA>         Mrs Shamsa                 NA              Hamza
## 2            <NA>          Dr Saniya                 NA              Haris
## 3            <NA>         Mrs Shamsa                300               Fiza
## 4            <NA>         Mrs Shamsa                450             Zaheer
## 5            <NA>         Mrs Shamsa                600              Haris
## 6            <NA>          Dr Saniya                 NA              Haris
##   Next.Apt
## 1     <NA>
```

```
## 2     <NA>
## 3     <NA>
## 4     <NA>
## 5     <NA>
## 6     <NA>
```

```
#renaming columns
df<-rename(df,
       Consulting_Doctor=`Consulting..Doctor`,
       Total_Charges=`Total..Charges`,
       Amount_Received=`Amount..Received.`,
       Amount_Balance=`Amount..Balance`,
       Amount_Received_By=`Amount.Received.By`,
       Amount_in_Hospital=`Amount.in.Hospital`,
       Receptionist_Name=`Receptionist..Name`,
       Next_Apt=`Next.Apt`)

df$Date <- as.Date(strptime(df$Date, "%a, %B %d, %Y"))

df$Time <- format(strptime(df$Time, format='%I:%M %p'), '%I:%M %p')

health_data<-tbl_df(df)

health_data<-mutate(health_data,
                    Age= as.integer(Age),
                    Total_Charges=as.integer(Total_Charges),
                    Amount_in_Hospital=as.integer(Amount_in_Hospital),
                    Time=as.character(Time),
                    Date=as.character(Date))
```

```
## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion

## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion
```

```
health_data%>%
select(Date)%>%
transmute(Date=substring(Date, regexpr(' ', Date)+1, nchar(Date)))-> Dates
Dates
```

```
## # A tibble: 222 × 1
##          Date
##         <chr>
## 1  2017-01-01
## 2  2017-01-02
## 3  2017-01-02
## 4  2017-01-02
## 5  2017-01-02
## 6  2017-01-02
## 7  2017-01-02
## 8  2017-01-02
## 9  2017-01-02
## 10 2017-01-02
## # ... with 212 more rows
```

```
health_data%>%
select(Date)%>%
transmute(day= substring(Date,1,regexpr(',', Date)-1))-> Days
Days
```

```
## # A tibble: 222 × 1
##      day
##    <chr>
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## # ... with 212 more rows
```

```
#Q2
#Dates<-select(health_data, Date)
#Dates<-mutate(Dates, substring(Date,",")[[1]][1])
df['Date']<-lapply(df['Date'], as.character)

#Q3
summarize(health_data, average=mean(Age))
```

```
## # A tibble: 1 × 1
##   average
##     <dbl>
## 1      NA
```

```
#Q4
health_data%>%
filter(Age<12)%>%
select(Age)%>%
summarize(total.count=n())
```

```
## # A tibble: 1 × 1
##   total.count
##         <int>
## 1          23
```

```
#Q5
health_data%>%
filter(Sex=='M')%>%
group_by(Procedure)%>%
tally(sort=T)
```

```
## # A tibble: 29 × 2
##             Procedure     n
##                 <chr> <int>
## 1       Consultation    37
## 2              X Ray    11
## 3          Injection     9
```

```
## 4               Crown     6
## 5            Dressing     5
## 6  Consultation+X Ray     4
## 7     Laboratory Test     3
## 8               R.C.T     3
## 9            Scalling     3
## 10         Extraction     2
## # ... with 19 more rows
```

```
health_data%>%
  filter(Sex=='F' | Sex=='f')%>%
  group_by(Procedure)%>%
  tally(sort=T)
```

```
## # A tibble: 30 × 2
##         Procedure     n
##             <chr> <int>
## 1    Consultation    46
## 2      Extraction     8
## 3           X Ray     4
## 4       Injection     3
## 5    Orthodontics     3
## 6        Scalling     3
## 7             USG     3
## 8            <NA>     3
## 9  22 Unit Bridge     2
## 10  4 Unit Bridge     2
## # ... with 20 more rows
```

```
#Q6
health_data%>%
  group_by(Consulting_Doctor)%>%
  summarize(sum=sum(Total_Charges))%>%
  arrange(desc(sum))
```

```
## # A tibble: 23 × 2
##    Consulting_Doctor   sum
##                <chr> <int>
## 1      Dr Kinza Alam 76700
## 2            Dr Saad 52000
## 3             Dr Ali 26100
## 4    Dr Qurat ul Ain 20900
## 5   Dr Riffat Naheed 18800
## 6           Dr Irfan 11000
## 7     Dr Waqar Azeem  6000
## 8       Dr Saad Riaz  5700
## 9       Brig Farrukh  3750
## 10        Dr Shireen  3200
## # ... with 13 more rows
```

```
#Q7
health_data%>%
  group_by(Procedure)%>%
  summarize(sum=sum(Total_Charges))%>%
  arrange(desc(sum))
```

```
## # A tibble: 48 × 2
##                          Procedure    sum
##                              <chr>  <int>
## 1                     Orthodontics 240000
## 2                   22 Unit Bridge  69500
## 3                        C Section  65000
## 4                        Operation  50000
## 5  RCT (4 teeth) Bridge (9 teeth)  48000
## 6           8 Unit Bridge+2 R.C.T  30000
## 7                            Crown  20000
## 8                         Scalling  16500
## 9                            R.C.T  15500
## 10                      Extraction  14600
## # ... with 38 more rows
```

```
#Q8
time<-hour(strptime(health_data$Time, "%I:%M %p"))
which.max(table(time))
```

```
## 13
##  5
```

```
#Q9
health_data%>%
  mutate(time_bracket = ifelse(time >= 6 & time <= 12, "Morning",
                        ifelse(time > 12 & time <= 16, "Afternoon",
                        ifelse(time > 16 & time <= 19, "Evening",
                        ifelse(time > 19, "Night", NA)))))
```

```
## # A tibble: 222 × 16
##          Date    id     Time   Age   Sex Consulting_Doctor       Specialty
##         <chr> <int>    <chr> <int> <chr>              <chr>           <chr>
## 1  2017-01-01   101     <NA>    40     F     Dr Kinza Alam           Gynae
## 2  2017-01-02   150 10:45 AM    26     M     Nursing Staff            <NA>
## 3  2017-01-02    58 12:38 PM    30     F  Dr Riffat Naheed Psychotherapist
## 4  2017-01-02    75 01:00 PM    40     M  Dr Riffat Naheed Psychotherapist
## 5  2017-01-02    97 02:45 PM    27     M  Dr Riffat Naheed Psychotherapist
## 6  2017-01-02   101 03:00 PM    40     F     Dr Kinza Alam           Gynae
## 7  2017-01-02    26 03:28 PM    43     M         Dr Saniya             M/o
## 8  2017-01-02   149 03:45 PM    28     F         Dr Fakiha         Dentist
## 9  2017-01-02    20 03:45 PM     2     F         Dr Fakiha         Dentist
## 10 2017-01-02    72 05:00 PM    40     M         Dr Fakiha         Dentist
## # ... with 212 more rows, and 9 more variables: Procedure <chr>,
## #   Total_Charges <int>, Amount_Received <int>, Amount_Balance <chr>,
## #   Amount_Received_By <chr>, Amount_in_Hospital <int>,
## #   Receptionist_Name <chr>, Next_Apt <chr>, time_bracket <chr>
```

```
#Q10
health_data%>%
  group_by(id)%>%
  summarize(length=n())%>%
  filter(length>1)%>%
  nrow()
```

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

```
#Q11
health_data%>%
  group_by(id,Procedure)%>%
  summarize(length=n())%>%
  filter(length>1)%>%
  print()
```

```
## Source: local data frame [24 x 3]
## Groups: id [23]
##
##       id                       Procedure length
##    <int>                           <chr>  <int>
## 1      1                        Pharmacy     10
## 2     12                  22 Unit Bridge      2
## 3     13                    Consultation      2
## 4     17                    Consultation      2
## 5     17 RCT (4 teeth) Bridge (9 teeth)      2
## 6     20                    Consultation      2
## 7     25                    Consultation      2
## 8     45                   R.P.D + Crown      2
## 9     46                        Dressing      4
## 10    63                    Consultation      2
## # ... with 14 more rows
```

```
#Q12
health_data%>%
  group_by(id)%>%
  summarize(length=n())%>%
  filter(length>1)%>%
  print()
```

```
## # A tibble: 37 × 2
##       id length
##    <int>  <int>
## 1      1     12
## 2      4      2
## 3     12      2
## 4     13      2
## 5     17      4
## 6     20      2
## 7     25      2
## 8     40      2
## 9     45      3
## 10    46      5
## # ... with 27 more rows
```

```
#Q13
health_data%>%
  filter(Sex=='M')%>%
  select(Age)%>%
  summarize(median=median(Age,na.rm=TRUE))%>%
  print()
```

```
## # A tibble: 1 × 1
##   median
```

```
##    <int>
## 1     29
```

```
health_data%>%
  filter(Sex=='F'| Sex=='f')%>%
  select(Age)%>%
  summarize(median=median(Age,na.rm=TRUE))%>%
  print()
```

```
## # A tibble: 1 × 1
##   median
##    <int>
## 1     30
```

```
#Q14
health_data$Amount_Balance<-gsub(",", "" ,health_data$Amount_Balance, fixed=TRUE)
health_data$Amount_Balance<-gsub(".00", "" ,health_data$Amount_Balance, fixed=TRUE)

health_data%>%
  mutate(Amount_Balance=as.integer(Amount_Balance))%>%
  select(Amount_Balance)%>%
  summarize(sum=sum(Amount_Balance,na.rm=TRUE))%>%
  print()
```

```
## # A tibble: 1 × 1
##      sum
##    <int>
## 1 222500
```

```
#Q15
health_data%>%
  filter(Procedure=="Consultation")%>%
  select(Total_Charges)%>%
  summarize(sum=sum(Total_Charges,na.rm=TRUE))%>%
  print()
```

```
## # A tibble: 1 × 1
##     sum
##   <int>
## 1 83950
```

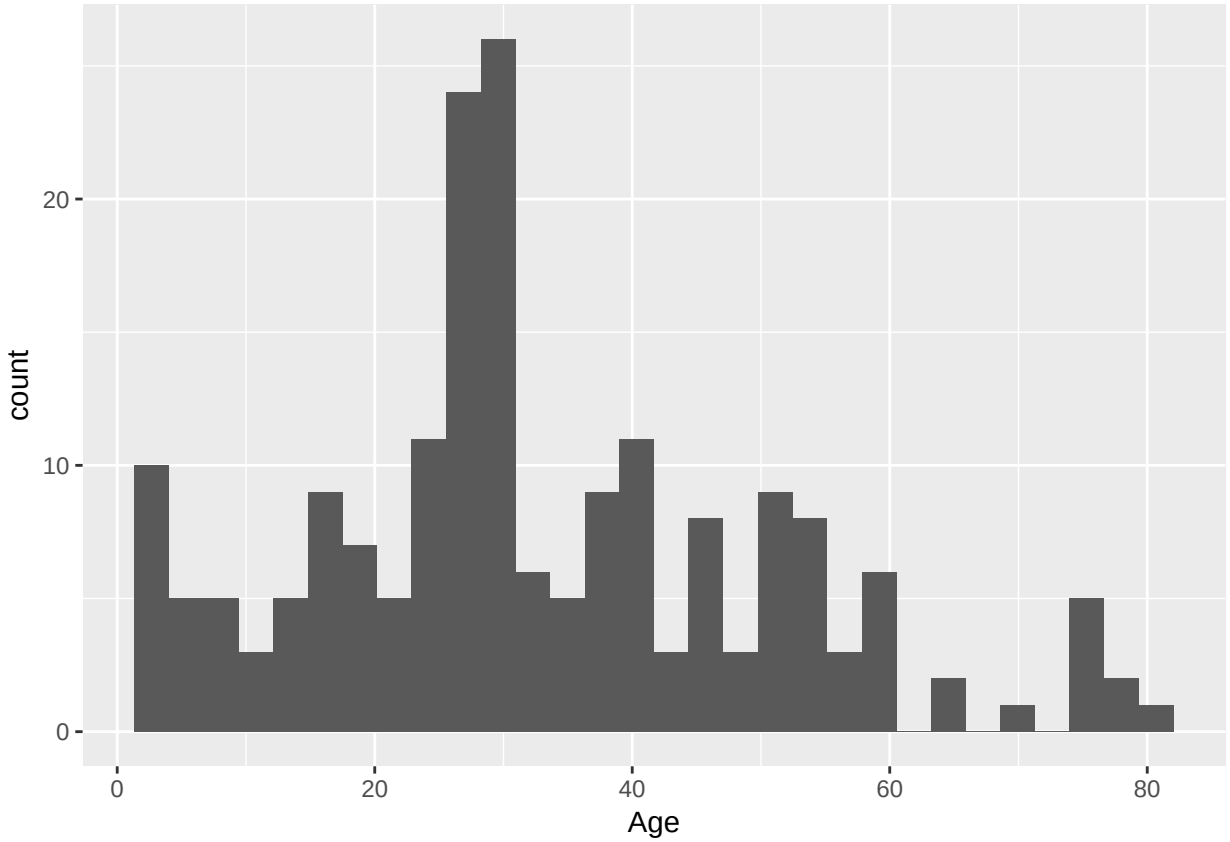
```
#Q16
health_data%>%
  select(Age,Total_Charges) %>%
  filter(!is.na(Age),!is.na(Total_Charges)) ->corr

cor(corr$Age,corr$Total_Charges)
```

```
## [1] 0.02809046
```

```
#Q17
qplot(health_data$Age, xlab="Age")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 30 rows containing non-finite values (stat_bin).
```




```
#Q18
health_data%>%
  filter(Procedure== 'X Ray' | Procedure== 'Scalling' ) %>%
  select(Total_Charges)%>%
  summarize(sum=sum(Total_Charges,na.rm=TRUE))%>%
  print()
```

```
## # A tibble: 1 × 1
##     sum
##   <int>
## 1 22300
```

```
write.csv(health_data, "E://DIH//Assignment2//clean_hospital_data_R.csv")
```