aliejaz1749_khi_r_assignment2.R

Ali.Ejaz

Mon Mar 27 17:40:19 2017

```
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(readr)

# List all objects in the workspace
#ls()

# Or remove all files from your workspace
#rm(list = ls())

# Load CSV file
hospitaldata <-
read.csv('D://diHub//Assessment2_RandPython_Marked//aliejaz1749_khi_r_assignm
ent2//hospitaldata.csv', header = TRUE, stringsAsFactors = F)
str(hospitaldata)
```

```
## 'data.frame':    222 obs. of  15 variables:
##  $ Date              : chr  "Sunday, January 01, 2017" "Monday, January
02, 2017" "Monday, January 02, 2017" "Monday, January 02, 2017" ...
##  $ id                : int  101 150 58 75 97 101 26 149 20 72 ...
##  $ Time              : chr  "11:00" "10:45AM" "12:38PM" "1:00PM" ...
##  $ Age               : chr  "40" "26" "30" "40" ...
##  $ Sex               : chr  "F" "M" "F" "M" ...
```

```
##  $ Consulting..Doctor: chr  "Dr Kinza Alam" "Nursing Staff" "Dr Riffat
Naheed" "Dr Riffat Naheed" ...
##  $ Specialty         : chr  "Gynae" "" "Psychotherapist" "Psychotherapist"
...
##  $ Procedure         : chr  "C Section" "Dressing" "Consultation"
"Consultation" ...
##  $ Total..Charges    : chr  "30000" "1500" "1000" "1500" ...
##  $ Amount..Received. : int  30000 1500 1000 1500 2000 35000 2000 500 NA
500 ...
##  $ Amount..Balance   : chr  " -   " " -   " " -   " " -   " ...
##  $ Amount.Received.By: chr  " Mrs Shamsa " " Dr Saniya " " Mrs Shamsa " "
Mrs Shamsa " ...
##  $ Amount.in.Hospital: int  NA NA 300 450 600 NA NA 500 NA 500 ...
##  $ Receptionist..Name: chr  "Hamza" "Haris" "Fiza" "Zaheer" ...
##  $ Next.Apt          : chr  "" "" "" "" ...

#create copy of dataframe
df <- tbl_df(hospitaldata)
glimpse(df)

## Observations: 222
## Variables: 15
## $ Date               <chr> "Sunday, January 01, 2017", "Monday, Januar...
## $ id                 <int> 101, 150, 58, 75, 97, 101, 26, 149, 20, 72,...
## $ Time               <chr> "11:00", "10:45AM", "12:38PM", "1:00PM", "2...
## $ Age                <chr> "40", "26", "30", "40", "27", "40", "43", "...
## $ Sex                <chr> "F", "M", "F", "M", "M", "F", "M", "F", "F"...
## $ Consulting..Doctor <chr> "Dr Kinza Alam", "Nursing Staff", "Dr Riffa...
## $ Specialty          <chr> "Gynae", "", "Psychotherapist", "Psychother...
## $ Procedure          <chr> "C Section", "Dressing", "Consultation", "C...
## $ Total..Charges     <chr> "30000", "1500", "1000", "1500", "2000", "3...
## $ Amount..Received.  <int> 30000, 1500, 1000, 1500, 2000, 35000, 2000,...
## $ Amount..Balance    <chr> " -   ", " -   ", " -   ", " -   ", " -   "...
## $ Amount.Received.By <chr> " Mrs Shamsa ", " Dr Saniya ", " Mrs Shamsa...
## $ Amount.in.Hospital <int> NA, NA, 300, 450, 600, NA, NA, 500, NA, 500...
## $ Receptionist..Name <chr> "Hamza", "Haris", "Fiza", "Zaheer", "Haris"...
## $ Next.Apt           <chr> "", "", "", "", "", "", "", "", "", "", "",...

# Qus1. Please remove the dots in the names, so it may become easier for you
to work through it.
names(df) <- gsub("\\.", "", names(df))

# Qus2. Which day of the week is expected to have most visits?
NameOfDays <- wday(mdy(df$Date), label = TRUE)
which.max(summary(NameOfDays))

## Mon
##   2

# Qus3. What is the average age of patients?
unique(df$Age)
```

```
##  [1] "40"  "26"  "30"  "27"  "43"  "28"  "2"   "32"  "76"  "75"  "36"
## [12] "42"  "23"  "48"  "25"  "50"  "60"  "-"   "57"  ""    "39"  "6"
## [23] "5"   "9"   "29"  "45"  "34"  "52"  "21"  "58"  "33"  "10"  "19"
## [34] "53"  "55"  "28M" "47"  "49"  "31"  "7"   "8"   "17"  "54"  "35"
## [45] "80"  "70"  "20"  "13"  "18"  "14"  "6M"  "24"  "3"   "38"  "22"
## [56] "65"  "64"  "78"  "56"
```

```
class(df$Age)
```

```
## [1] "character"
```

```
p_age_var <- df$Age
p_age <- as.numeric(p_age_var)
```

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

```
mean(p_age, na.rm = TRUE)
```

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

```
# Qus4. How many children were entertained? (Make a Bracket of Age from 1-12)
p_child_age <- as.numeric(df$Age)
```

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

```
p_child_age[is.na(p_child_age)] <- 0
sum(p_child_age >= 12)
```

```
## [1] 169
```

```
# Qus5. Which gender type had what kind of procedure in abundance? i.e.
Female visit mostly because of Gynae Problem
gender_type<-df%>%
filter(!is.na(Sex))%>%
group_by(Procedure,Sex)%>%
summarize(total_gender= n())%>%
filter(total_gender==max(total_gender))
gender_type
```

```
## Source: local data frame [51 x 3]
## Groups: Procedure [48]
##
##                  Procedure   Sex total_gender
##                      <chr> <chr>        <int>
## 1                              F            3
## 2           22 Unit Bridge     F            2
## 3            4 Unit Bridge     F            2
## 4    8 Unit Bridge+2 R.C.T     M            1
## 5                      BSR     M            1
## 6               C Section     F            2
## 7               Cancelled                  1
## 8             Consultation     F           45
```

```
## 9   Consultation +   X Ray     F            1
## 10 Consultation + Dressing     M            1
## # ... with 41 more rows

# Qus6. Which Doctor is earning highest?
d_high_ern <- select(df, ConsultingDoctor , AmountReceived)
d_high_ern <- filter(df , ConsultingDoctor!='Nursing Staff' ,
!is.na(AmountReceived))
grp_d_high_ern <- group_by(df, ConsultingDoctor)
summriz_doc_ern <- summarize(grp_d_high_ern, sum(AmountReceived), na.rm =
TRUE)
summriz_doc_ern[is.na(summriz_doc_ern)] <- 0
as.numeric(summriz_doc_ern$`sum(AmountReceived)`)

##  [1]  2800  3750  1000     0 26100  1400  1500     0 11000 76700  1000
## [12]  1000  1000 20900 18800 52000  5700  1500     0  3200  6000  1700
## [23]     0

max(summriz_doc_ern$`sum(AmountReceived)`)

## [1] 76700

# Qus7. Which procedure type earns more money?
p_proc_typ_high <- select(df, Procedure , AmountReceived)
x <- p_proc_typ_high%>%
 filter(!is.na(AmountReceived))%>%
  group_by(Procedure)%>%
  summarize(Procedure_1 = sum(AmountReceived))%>%
  filter(Procedure_1 == max(Procedure_1))
  x

## # A tibble: 1 × 2
##      Procedure Procedure_1
##          <chr>       <int>
## 1 Consultation       83450

# Qus8. Which time of the day has highest frequency of visits by hour?
x <- df %>%
  filter(!is.na(Time), Time != '-') %>%
  group_by(Time) %>%
  summarize(time_wise_cnt = n()) %>%
  filter(Time != '') %>%
  filter(time_wise_cnt == max(time_wise_cnt))
x

## # A tibble: 4 × 2
##      Time time_wise_cnt
##     <chr>         <int>
## 1  1:00PM             9
## 2  1:30PM             9
## 3 12:00PM             9
## 4  3:00PM             9
```

```r
# Qus9. Create a bracket of time by Morning, Afternoon, Evening, Night (6am - 12pm - Morning, 12 pm- 4 pm, Afternoon, 4 pm- 7pm, Evening, 7pm - 6 am, Night).


# Qus10. How many patients are repeated visitors?
p_rep_patient_visit <- df %>%
  group_by(id)%>%
  summarize(p_count = n()) %>%
  filter(p_count > 1) %>%
  summarize(tot_rep_vis = n())
p_rep_patient_visit
```

```
## # A tibble: 1 × 1
##   tot_rep_vis
##         <int>
## 1          37
```

```r
# Qus11. Give us the id of repeated visitors.
p_rep_vistors <- df %>%
  group_by(id)%>%
  summarize(p_rep_Vist = n()) %>%
  filter(p_rep_Vist > 1) %>%
  arrange(desc(p_rep_Vist))
p_rep_vistors
```

```
## # A tibble: 37 × 2
##       id p_rep_Vist
##    <int>      <int>
## 1      1         12
## 2     46          5
## 3    122          5
## 4     17          4
## 5     94          4
## 6    140          4
## 7     45          3
## 8     63          3
## 9    101          3
## 10   107          3
## # ... with 27 more rows
```

```r
# Qus12. Which patients visited again for the same problem?
p_p_prob_Vist <- df %>%
  group_by(Procedure, id)%>%
  summarize(p_prob_Vist = n()) %>%
  filter(p_prob_Vist > 1) %>%
  arrange(desc(p_prob_Vist))
p_p_prob_Vist
```

```
## Source: local data frame [24 x 3]
## Groups: Procedure [15]
```

```
##
##         Procedure    id p_prob_Vist
##             <chr> <int>       <int>
## 1        Pharmacy     1          10
## 2       Injection   122           5
## 3        Dressing    46           4
## 4    Consultation   114           3
## 5           Crown   145           3
## 6       Injection    94           3
## 7  22 Unit Bridge    12           2
## 8   4 Unit Bridge   140           2
## 9       C Section   101           2
## 10   Consultation    13           2
## # ... with 14 more rows
```

```
# Qus13. What is the median age for Females and Males?
p_medi_gender <- df %>%
  group_by(Sex)%>%
  summarize(p_Sex = n()) %>%
  filter(p_Sex > 1) %>%
  arrange(desc(p_Sex))
p_medi_gender
```

```
## # A tibble: 4 × 2
##     Sex p_Sex
##   <chr> <int>
## 1     M   105
## 2     F   102
## 3          12
## 4     -     2
```

```
# Qus14. What is the total amount in balance?
p_am_blc <- df$AmountBalance
p_am_blc <- as.numeric(parse_number(p_am_blc))
```

```
## Warning: 211 parsing failures.
## row col expected actual
##   1  -- a number      -
##   2  -- a number      -
##   3  -- a number      -
##   4  -- a number      -
##   5  -- a number      -
## ... ... ........ ......
## See problems(...) for more details.
```

```
p_am_blc <- as.numeric(p_am_blc)
p_am_blc[which(is.na(as.numeric(as.character(p_am_blc))))]<-0
p_am_blc = sum(p_am_blc)
p_am_blc
```

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

```r
# Qus15. How much money was made by Procedure Type "Consultation"?
p_consultation_max <- df %>%
  filter( Procedure == 'Consultation', !is.na(AmountReceived),
AmountReceived!= '-') %>%
  group_by(Procedure) %>%
  summarize(p_consultation_max = sum(AmountReceived))
p_consultation_max
```

```
## # A tibble: 1 × 2
##      Procedure p_consultation_max
##          <chr>              <int>
## 1 Consultation              83450
```

```r
# Qus16. Is there a relation between Age and Total Charges paid?


# Qus17. Which Age group had highest number of visits?
p_max_visit <- df %>%
  filter(Age!= '-', Age!= '', !is.na(Age)) %>%
  group_by(Age) %>%
  summarize(p_max_visit = n()) %>%
  filter(p_max_visit == max(p_max_visit))
p_max_visit
```

```
## # A tibble: 1 × 2
##     Age p_max_visit
##   <chr>       <int>
## 1    30          20
```

```r
# Qus18. What is the total cost earned by Procedure Type X Ray and Scalling
together?
p_tot_cost <- df %>%
  filter(Procedure == 'X Ray' | Procedure == 'Scalling' , Procedure!= '-',
Procedure!= '', !is.na(Procedure)) %>%
  group_by(Procedure) %>%
  summarize(p_tot_cost = sum(AmountReceived)) %>%
  filter(p_tot_cost == max(p_tot_cost))
p_tot_cost
```

```
## # A tibble: 1 × 2
##   Procedure p_tot_cost
##       <chr>      <int>
## 1  Scalling      16500
```

```r
write.csv(df,
file='D:/diHub/Assessment2_RandPython_Marked/aliejaz1749_khi_r_assignment2/up
dated_hospitaldata.csv')
```