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\_assignment2//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

# 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

# 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

# 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

# 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

# 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

# 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

write.csv(df, file='D:/diHub/Assessment2\_RandPython\_Marked/aliejaz1749\_khi\_r\_assignment2/updated\_hospitaldata.csv')