5-5 EDA Problem Set 3

Curtis ONeal

May 5, 2016

# The following codes sets up libraries, reads the data, and preps some of it.

knitr::opts\_chunk$set(echo = TRUE)  
  
library(ggplot2)  
library(lubridate)

##   
## Attaching package: 'lubridate'

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

library("xlsx", lib.loc="/Library/Frameworks/R.framework/Versions/3.2/Resources/library")

## Loading required package: rJava

## Loading required package: xlsxjars

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':  
##   
## intersect, setdiff, union

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

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

dat <- read.xlsx("~/Dropbox (Personal)/\_Data Science Readings and Class/\_Springboard/Project Data and Files/Birthdays\_Google\_Cal\_Export.xlsx", sheetName="Birthdays\_Google\_Cal\_Export")  
# This is an Excel of My calendar's birthday's exported from Google Calendar   
  
#View(dat)  
names(dat) #check varnames and importation

## [1] "Title" "Start" "End" "Duration" "Description"

#Convert if not already type date using lubridate package  
dat$Start<- ymd(dat$Start)  
dat$month<-months(dat$Start) #pull just the month  
dat$month\_fac<- as.factor(dat$month) #convert months to factor  
dat$day<- as.factor(day(dat$Start))  
dat$Title<- substr(dat$Title, 1, 7) #Try to Protect their Privacy

# Your task is to investigate the distribution of your friends' birth months and days.

Below are questions for you to answer, and we hope you think of others.

# How many people share your birthday? Do you know them?

#set my birthdate  
mine<- as.Date("2016-10-26")   
#sume the logical vector of dates that match mine in the set, remove mine  
count\_my\_bd <- sum(dat$Start == mine)-1  
count\_my\_bd

## [1] 1

#Show the vector to check it  
is\_mine\_vector <- dat$Start == mine  
is\_mine\_vector

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  
## [45] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

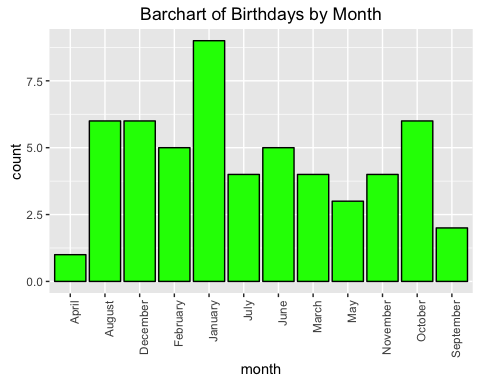
#Use the logical vector to pull the names whose brithday match mine -myslef included.  
dat$Title[is\_mine\_vector]

## [1] "Curtis " "Natalie"

Two people share my birthday. Of course one is me. The other is Natalie and I know her.

# Which month contains the most number of birthdays?

#Plot a barchart of counts -histograms are a bear with factos  
ggplot(dat, aes(month)) +  
 geom\_bar(fill="green", colour="black") +   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
 labs(title= "Barchart of Birthdays by Month")



January has the most birthdays.

# How many birthdays are in each month?

month\_table <- as.data.frame(table(dat$month\_fac))  
colnames(month\_table)<- c("Month", "Freq")  
arrange(month\_table, Freq)

## Month Freq  
## 1 April 1  
## 2 September 2  
## 3 May 3  
## 4 July 4  
## 5 March 4  
## 6 November 4  
## 7 February 5  
## 8 June 5  
## 9 August 6  
## 10 December 6  
## 11 October 6  
## 12 January 9

We also find that January has 9 birthdays, which was not clear from the Bar Chart

# Which day of the year has the most number of birthdays?

DOY\_table <- table(dat$Start)  
DOY\_df <- as.data.frame(DOY\_table)  
colnames(DOY\_df) <- c("DOY", "Freq")  
arrange(DOY\_df, Freq)

## DOY Freq  
## 1 2015-12-31 1  
## 2 2016-01-01 1  
## 3 2016-01-06 1  
## 4 2016-01-07 1  
## 5 2016-01-15 1  
## 6 2016-01-26 1  
## 7 2016-01-27 1  
## 8 2016-01-28 1  
## 9 2016-02-14 1  
## 10 2016-02-19 1  
## 11 2016-02-22 1  
## 12 2016-02-24 1  
## 13 2016-02-25 1  
## 14 2016-03-12 1  
## 15 2016-03-22 1  
## 16 2016-03-29 1  
## 17 2016-03-30 1  
## 18 2016-04-03 1  
## 19 2016-05-02 1  
## 20 2016-05-03 1  
## 21 2016-05-13 1  
## 22 2016-06-20 1  
## 23 2016-06-25 1  
## 24 2016-06-29 1  
## 25 2016-07-05 1  
## 26 2016-07-08 1  
## 27 2016-07-16 1  
## 28 2016-07-26 1  
## 29 2016-08-04 1  
## 30 2016-08-08 1  
## 31 2016-08-27 1  
## 32 2016-08-30 1  
## 33 2016-09-06 1  
## 34 2016-09-23 1  
## 35 2016-10-05 1  
## 36 2016-10-12 1  
## 37 2016-10-19 1  
## 38 2016-10-31 1  
## 39 2016-11-02 1  
## 40 2016-11-08 1  
## 41 2016-11-12 1  
## 42 2016-11-28 1  
## 43 2016-12-03 1  
## 44 2016-12-09 1  
## 45 2016-12-25 1  
## 46 2016-12-28 1  
## 47 2016-12-31 1  
## 48 2016-01-14 2  
## 49 2016-06-01 2  
## 50 2016-08-29 2  
## 51 2016-10-26 2

max\_days <- (select( filter(DOY\_df, Freq == max(DOY\_df$Freq)), DOY))  
#View(max\_days)  
#str(max\_days)  
  
#max\_days[,1] factors  
max\_days <- as.Date(max\_days[,1])  
max\_days

## [1] "2016-01-14" "2016-06-01" "2016-08-29" "2016-10-26"

#These days have the maximum frequencies.

# Who has birthdays on these days?

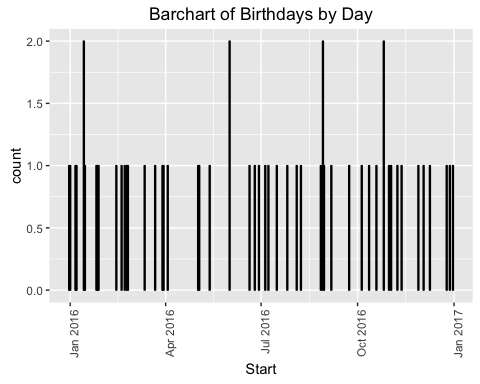
max\_days

## [1] "2016-01-14" "2016-06-01" "2016-08-29" "2016-10-26"

dat$Title[dat$Start %in% max\_days]

## [1] "Seok Jo" "Diwakar" "joanna " "Rev. Be" "christi" "Christi" "Curtis "  
## [8] "Natalie"

#At this scale this is hard to see.  
ggplot(dat, aes(Start)) +  
 geom\_bar(fill="green", colour="black") +   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +  
 labs(title= "Barchart of Birthdays by Day", breaks = 30)



At this scale the barchar of the portion of 365 days is hard to read, but we can tell most days in the set only have one birthday frome the set on it, and four days have 2 birthdays on them.

The most birthdays on any day is 2.

The four days with two birthdays are: 2016-01-14 2016-06-01 2016-08-29 2016-10-26

Four of the eight people with birthdays on those days are:

Seok J birthday Melanie S birthday Mark T birthday  
Amy R birthday

# Do you have at least 365 friends that have birthdays on everyday of the year?

length(dat$Title)-1 #number of friend Observations minus myself

## [1] 54

length(unique(dat$Start)) #Number of unique days of the year with my Frinds birthdays.

## [1] 51

I have 54 friends brithdays in the dataset not including myself.

There are 51 unique days in the dataset.

sort(unique(day(dat$Start))) #Unique Days of the month with Birthdays in the dataset

## [1] 1 2 3 4 5 6 7 8 9 12 13 14 15 16 19 20 22 23 24 25 26 27 28  
## [24] 29 30 31

table(day(dat$Start))

##   
## 1 2 3 4 5 6 7 8 9 12 13 14 15 16 19 20 22 23 24 25 26 27 28 29 30   
## 3 2 3 1 2 2 1 3 1 3 1 3 1 1 2 1 2 1 1 3 4 2 3 4 2   
## 31   
## 3

length(unique(dat$Start)) #Number of unique days of the year with my Frinds birthdays.

## [1] 51

END.