HireArt Case Study

Chuyi Guo

3/14/2019

Data Pre-processing.

```
# read in the data into system.
library(readxl)
library(dplyr)
library(ggplot2)
data = read_excel('/Users/ChuyiGuo/Desktop/HireArt/HireArt - Data Analyst Exercise 10.12.17.xlsx')
data = as.data.frame(data)
# extract month, year for each line of data
data$month = as.numeric(format(as.Date(data$'Date of Contact'), "%m"))
data$year = as.numeric(format(as.Date(data$'Date of Contact'), "%Y"))
# extract some of the columns
data = data[,c(colnames(data) %in% c('Client Name', "month", "year"))]
```

According to the spreadsheet, a client may be contacted multiply times within one month. For this situation, the client should be counted as one, based on a monthly basis. Those duplicated records need to be taken care.

```
data2 = data[!duplicated(data),]
```

Get a client list for the team. That is the clients that had been contacted within those 4 years. Below shows some of the clients' name.

```
client_list = unique(data2$`Client Name`)
head(client_list)
```

```
## [1] "Wyman, Farrell and Haag"       "Veum, McClure and Schuster"
## [3] "Armstrong Group"               "Lueilwitz, Moore and Hahn"
## [5] "Abbott Group"                  "Oga, Gottlieb and Cruickshank"
```

There are totally ‘r length(client_list)’ clients for this team.

```
length(client_list)
```

```
## [1] 35
```

The numbers of clients for each month of each year are calculated (named n). I.e. the team contacted 9 clients in January 2014.

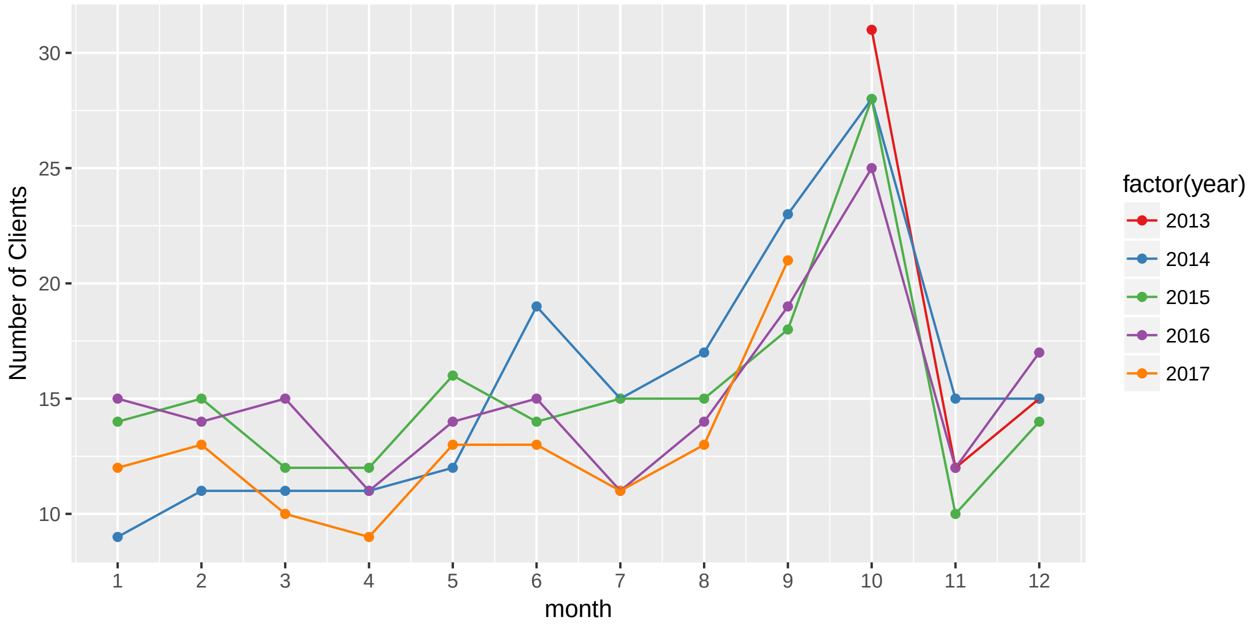
Also, the percentage of its clients for that month (named prop) are calculated. I.e. the team contacted 25% of its clients in January 2014.

```
total = as.data.frame(
  data2 %>%
    group_by(month, year) %>%
    summarise(n = n()) %>%
    mutate(prop = n/length(client_list))
)
total
```

```
##    month year  n      prop
## 1      1 2014  9 0.2571429
## 2       1 2015 14 0.4000000
## 3       1 2016 15 0.4285714
## 4       1 2017 12 0.3428571
## 5       2 2014 11 0.3142857
## 6       2 2015 15 0.4285714
## 7       2 2016 14 0.4000000
## 8       2 2017 13 0.3714286
## 9       3 2014 11 0.3142857
## 10      3 2015 12 0.3428571
## 11      3 2016 15 0.4285714
## 12      3 2017 10 0.2857143
## 13      4 2014 11 0.3142857
## 14      4 2015 12 0.3428571
## 15      4 2016 11 0.3142857
## 16      4 2017  9 0.2571429
## 17      5 2014 12 0.3428571
## 18      5 2015 16 0.4571429
## 19      5 2016 14 0.4000000
## 20      5 2017 13 0.3714286
## 21      6 2014 19 0.5428571
## 22      6 2015 14 0.4000000
## 23      6 2016 15 0.4285714
## 24      6 2017 13 0.3714286
## 25      7 2014 15 0.4285714
## 26      7 2015 15 0.4285714
## 27      7 2016 11 0.3142857
## 28      7 2017 11 0.3142857
## 29      8 2014 17 0.4857143
## 30      8 2015 15 0.4285714
## 31      8 2016 14 0.4000000
## 32      8 2017 13 0.3714286
## 33      9 2014 23 0.6571429
## 34      9 2015 18 0.5142857
## 35      9 2016 19 0.5428571
## 36      9 2017 21 0.6000000
## 37     10 2013 31 0.8857143
## 38     10 2014 28 0.8000000
## 39     10 2015 28 0.8000000
## 40     10 2016 25 0.7142857
## 41     11 2013 12 0.3428571
## 42     11 2014 15 0.4285714
## 43     11 2015 10 0.2857143
## 44     11 2016 12 0.3428571
## 45     12 2013 15 0.4285714
## 46     12 2014 15 0.4285714
## 47     12 2015 14 0.4000000
## 48     12 2016 17 0.4857143
```

Plot out the number of clients per month to get a intuitive overviews for this data set. It shows most of the clients are contacted in October.

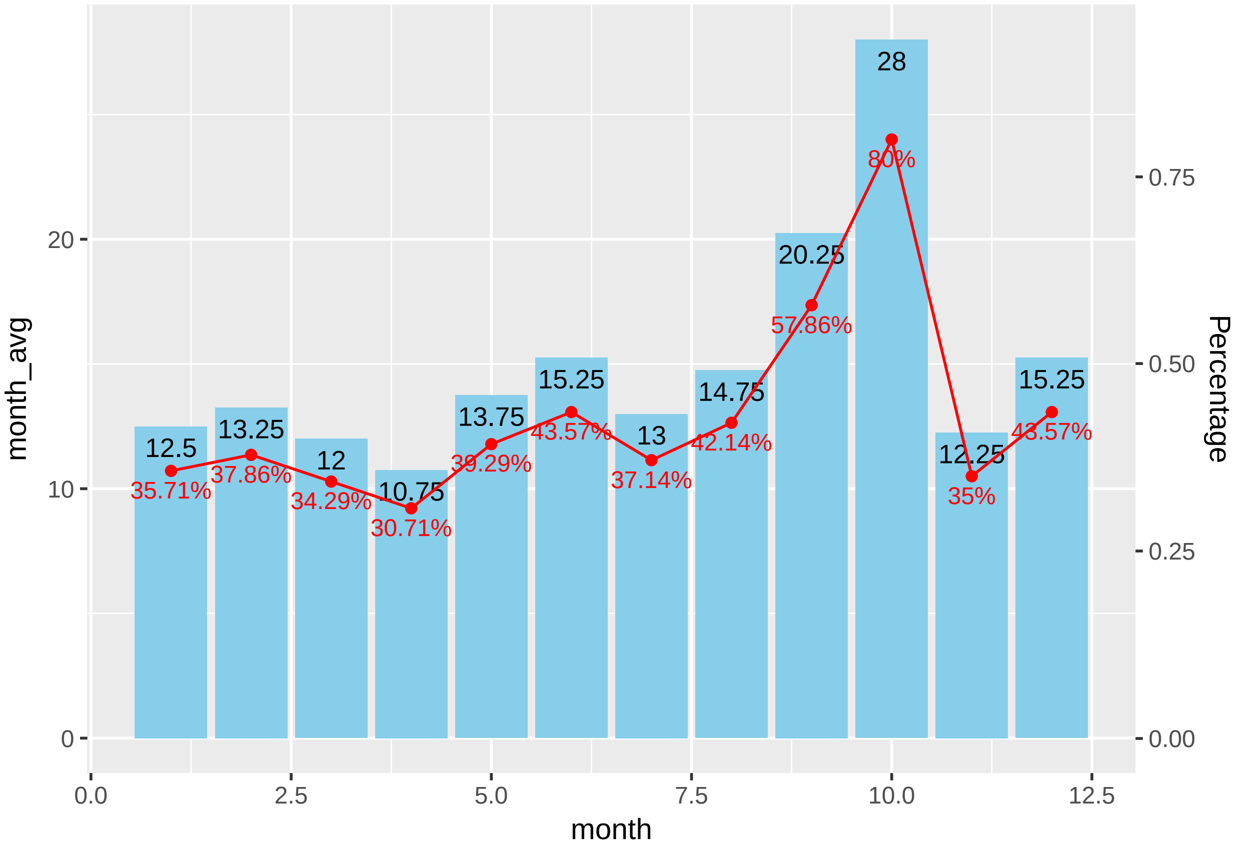
```
ggplot(total,aes(x = month, y = n )) +
  geom_point(aes(color = factor(year))) +
  geom_line(aes(color = factor(year))) +
  labs(y="Number of Clients") +
  scale_x_continuous(breaks = seq(1,12,1)) +
  scale_color_brewer(palette = "Set1")
```



Below shows the average number of clients and its percentage for each month during the four-year period. October has the highest value, which indicates on average, 80% of its clients had been contacted in October.

```
month_avg = as.data.frame(
  total %>%
    group_by(month) %>%
    summarise(month_avg = mean(n), month_prop_avg = mean(prop))
)

ggplot(month_avg, aes(x = month, y = month_avg)) +
  geom_bar(stat="identity", position=position_dodge(),fill='skyblue') +
  geom_text(aes(label=month_avg), vjust=1.6,
            color="black",position = position_dodge(0.9), size=3.5) +
  scale_y_continuous(sec.axis = sec_axis(~./30, name="Percentage")) +
  geom_line(aes(x = month, y = month_prop_avg*30, group=1),col='red') +
  geom_point(aes(x = month, y = month_prop_avg*30, group=1),col='red') +
  geom_text(aes(x=month, y=month_prop_avg*30,
                label=paste(round(month_prop_avg*100,digits=2),"%",sep="")),
            vjust=1.6, size=3,col='red') +
  scale_colour_brewer(palette="Set1")
```



Therefore, it can be concluded that October is the month that the team is likely to contact the most clients.