**ABSTRACT**

The objective of this report is to analyse the data of a Portuguese bank and predict whether a client would subscribe to a term deposit. This data was recorded by the bank through various marketing campaigns that were based on phone calls. This prediction is done to help the bank target prospective clients and launch successful and efficient marketing campaigns for their clients. The objective of this report is achieved by utilizing various data mining and data analysis methods. The target clients and the efficiency of the marketing campaign is decided by classifying the importance of each attribute available in the data.

**INTRODUCTION**

A term deposit is a cash investment that are offered by approved financial institutions like banks and credit unions. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. Term deposits are an extremely safe investment and are therefore very appealing to conservative, low-risk investors. The financial institutions launch marketing campaigns to attract clients to subscribe to a term deposit. The financial institutions can achieve this either by mass marketing campaigns or by target specific marketing. Recently, the financial institutions have started to prefer target specific marketing because mass marketing campaigns exhaust resources and often result in ineffective campaigns. To launch a successful target specific campaign, it is essential for the bank to determine the favourable target. Data analysis techniques are the most popular method to determine the favourable target for such campaigns.

**DATA SET DESCRIPTION**

The data is second-hand and obtained from the machine learning repository of University of California at Irvine. The data contains three types of variable: Numerical (Age, Balance, Duration, Campaign, Pdays, Previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed), Categorical (Day\_of\_week, Marital, Job, Contact, Education, Month, Poutcome, ) and Binary (Housing, Loan, Default, Subscribed).

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | Attribute | Data type | Description |
| 1. | Age | Numeric | The age of the client |
| 2. | Balance | Numeric | The balance in the bank account of the client |
| 3. | Duration | Numeric | The duration of call when last contacted |
| 4. | Campaign | Numeric | The number of times a client was contacted during the campaign |
| 5. | Pdays | Numeric | The number of days that passed by after the last contact |
| 6. | Previous | Numeric | The number of times a client was contacted before the campaign |
| 7. | Emp.var.rate | Numeric | The quarterly employment variation rate |
| 8. | Cons.price.idx | Numeric | The monthly consumer price index |
| 9. | Cons.conf.idx | Numeric | The monthly consumer confidence index |
| 10. | Euribor3m | Numeric | The Euribor three-month rate |
| 11. | Nr.employed | Numeric | The quarterly number of employees |
| 12. | Day\_of\_week | Categorical | The day of the week when last contact was made |
| 13. | Marital | Categorical | The marital status of a client |
| 14. | Job | Categorical | The occupation of a client |
| 15. | Contact | Categorical | The contact communication method of a client |
| 16. | Education | Categorical | The educational qualification of a client |
| 17. | Month | Categorical | The month when last contact was made |
| 18. | Poutcome | Categorical | The outcome of previous marketing campaigns for a client |
| 19. | Housing | Binary | If the client has a housing loan |
| 20. | Loan | Binary | If the client has a personal loan |
| 21. | Default | Binary | If the client has credit in default |
| 22. | Subscribed | Binary | If the client has subscribed to the term deposit |

**Exploratory Data Analysis**

```

*#Importing the dataset*

data <- **read.csv**('bank-additional-full.csv', sep = ';', header = TRUE, string sAsFactors = FALSE)

**head**(data)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | age | job marital | | education | | default housing loan contact month | | | | |
| ## 1 | 56 | housemaid married | | basic.4y | | no |  | no | no telephone | may |
| ## 2 | 57 | services married | | high.school | | unknown |  | no | no telephone | may |
| ## 3 | 37 | services married | | high.school | | no |  | yes | no telephone | may |
| ## 4 | 40 | admin. married | | basic.6y | | no |  | no | no telephone | may |
| ## 5 | 56 | services married | | high.school | | no |  | no | yes telephone | may |
| ## 6 | 45 | services married | | basic.9y unknown | | |  | no | no telephone | may |
| ## | day\_of\_week duration campaign pdays previous | | | | | | | poutcome emp.var.rate | | |
| ## 1 |  | mon | 261 | 1 | 999 | | 0 | nonexistent | | 1.1 |
| ## 2 |  | mon | 149 | 1 | 999 | | 0 | nonexistent | | 1.1 |
| ## 3 |  | mon | 226 | 1 | 999 | | 0 | nonexistent | | 1.1 |
| ## 4 |  | mon | 151 | 1 | 999 | | 0 | nonexistent | | 1.1 |
| ## 5 |  | mon | 307 | 1 | 999 | | 0 | nonexistent | | 1.1 |
| ## 6 |  | mon | 198 | 1 | 999 | | 0 nonexistent | | | 1.1 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | cons.price.idx cons.conf.idx euribor3m nr.employed | | | | y |
| ## 1 | 93.994 | -36.4 | 4.857 | 5191 | no |
| ## 2 | 93.994 | -36.4 | 4.857 | 5191 | no |
| ## 3 | 93.994 | -36.4 | 4.857 | 5191 | no |
| ## 4 | 93.994 | -36.4 | 4.857 | 5191 | no |
| ## 5 | 93.994 | -36.4 | 4.857 | 5191 | no |
| ## 6 | 93.994 | -36.4 | 4.857 | 5191 | no |
|  |  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **str**(data) | |  |  |
| ## 'data.frame': | | 41188 | obs. of 21 variables: |
| ## | $ age | : int | 56 57 37 40 56 45 59 41 24 25 ... |
| ## | $ job | : chr | "housemaid" "services" "services" "admin." ... |
| ## | $ marital | : chr | "married" "married" "married" "married" ... |
| ## | $ education | : chr | "basic.4y" "high.school" "high.school" "basic.6y" |
| ... |  |  |  |
| ## | $ default | : chr | "no" "unknown" "no" "no" ... |
| ## | $ housing | : chr | "no" "no" "yes" "no" ... |
| ## | $ loan | : chr | "no" "no" "no" "no" ... |
| ## | $ contact | : chr | "telephone" "telephone" "telephone" "telephone" .. |
| . |  |  |  |
| ## | $ month | : chr | "may" "may" "may" "may" ... |
| ## | $ day\_of\_week | : chr | "mon" "mon" "mon" "mon" ... |
| ## | $ duration | : int | 261 149 226 151 307 198 139 217 380 50 ... |
| ## | $ campaign | : int | 1 1 1 1 1 1 1 1 1 1 ... |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | $ pdays | : int | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 ... |
| ## | $ previous | : int | 0 0 | 0 0 | 0 0 | 0 0 | 0 0 ... | |  |  |  |  |
| ## | $ poutcome | : chr | "nonexistent" "nonexistent" "nonexistent" "nonexis | | | | | | | | | |
| tent" ... | |  |  |  |  |  |  |  |  |  |  |  |
| ## | $ emp.var.rate | : num | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 ... |
| ## | $ cons.price.idx: num | | 94 94 94 94 | | | 94 ... | |  |  |  |  |  |

* $ cons.conf.idx : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 6.4 | -36.4 ... |  |  |  |  |  |
| ## | $ euribor3m | : num | 4.86 | 4.86 | 4.86 | 4.86 4.86 ... |
| ## | $ nr.employed | : num | 5191 | 5191 | 5191 | 5191 5191 ... |
| ## | $ y | : chr | "no" "no" "no" | | | "no" ... |
|  |  |  |  |  |  |  |

*#Counting the number of nulls*

**sum**(**is.na**(data))

## [1] 0

Looks like the null values have been updated with unknown. Let’s explore the data and get some insights. Let’s start with the target variable.

**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

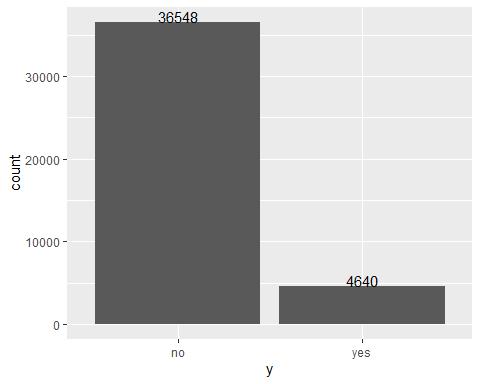
data <- **mutate**(data, **ifelse**(y **==** 'no', 0 ,1))

**names**(data)[**names**(data) **==** 'ifelse(y == "no", 0, 1)'] <-'subscribed' **library**(ggplot2)

**ggplot**(data=data, **aes**(x=y)) **+**

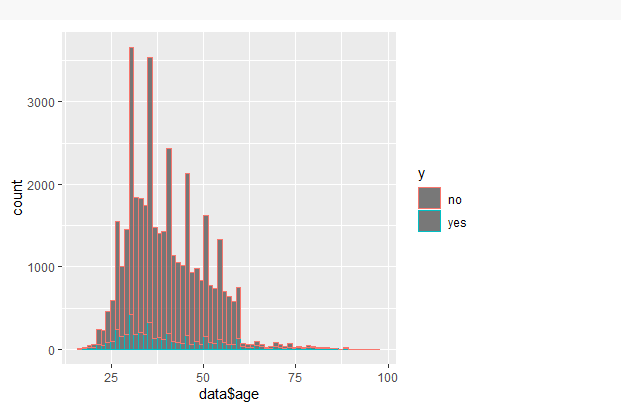
**geom\_bar**() **+**

**geom\_text**(stat='count', **aes**(label=..count..),vjust=0)



The plot shows the number of subscribers v/s non subscribers. In total there are 36548 people who have not subscribed for the term deposit and around 4640 who have subscribed for the term deposit. It is evident that the percent of subscribers is very less as compared to that of non subscribers.

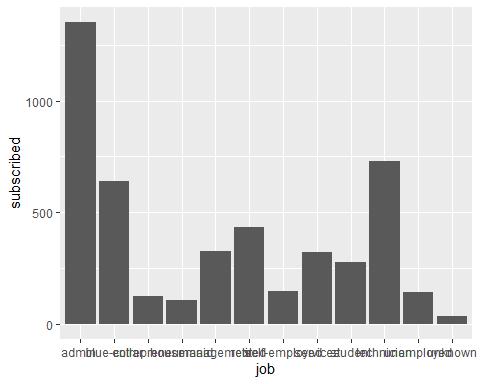
**ggplot**(data=data, **aes**(data**$**age,color =y)) **+ geom\_histogram**(bins =65,alpha=0.8)



The plot shows the histogram for age of all the clients. The subscribers are represented in red and the non subscribers in blue.The data is highly right skewed with more of data centered around the age gap from 25 to 50. The highest peak point is obtained at around 30.

**ggplot**(data, **aes**(job, subscribed)) **+**

**geom\_bar**(stat ='identity')



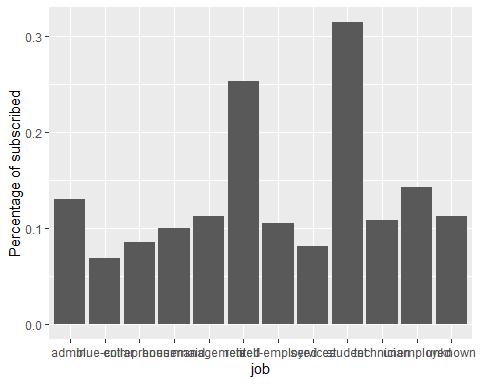
Considering the job variable, the admin group people are most subscribed. The next in line is technician and than blue collar jobs. If we ignore the unknown variable, the least people subscribed belongs to the housemaid job group.

jobs <-

data **%>%**

**group\_by**(job) **%>%**

**summarise**(`Percentage of subscribed`= **sum**(y **==** 'yes')**/length**(y)) **ggplot**(jobs, **aes**(job,`Percentage of subscribed`)) **+ geom\_bar**(stat ='identity')

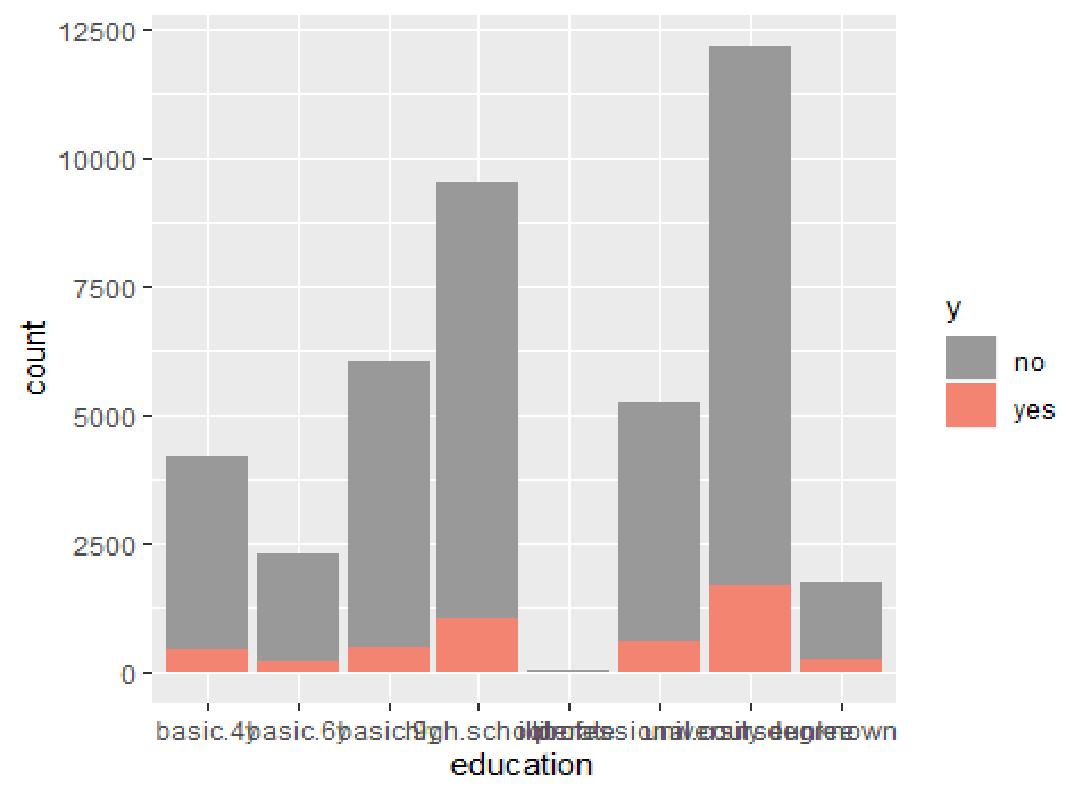


The above plot shows the percentage of subscribed by job title. The number of categories for job plot are admin, blue collar, entrepreneur, housemaid, management, retired, self employed, services, technician and unemployed. This plot portrays a different story. As per this plot the percent of students subscribing is the highest and then the retired one as compared to the count plot displayed above. The services workgroup subscribes the least and admin and technician subscribe moderately.

**ggplot**(data, **aes**(x =marital,fill =y)) **+**

**geom\_bar**() **+**

**scale\_fill\_manual**(values=**c**('#999999','#56B4E9'))



The plot above displays the marital status of the of the people subscribed. As displayed above, married people subscribe the most to the term deposit as compared to divorced which subscribe the least. An interesting feature to note is that the number of married people not subscribing is also the greatest.

**ggplot**(data, **aes**(x =education,fill =y))**+**

**geom\_bar**() **+**

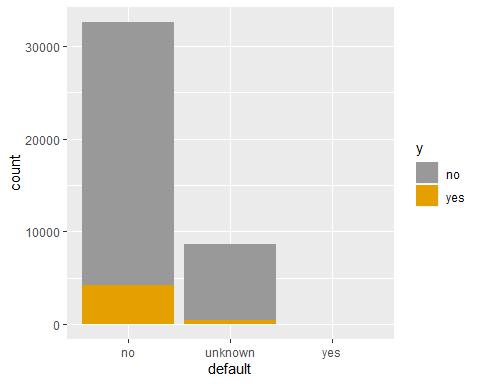
**scale\_fill\_manual**(values=**c**('#999999','#F48472'))

Considering the education plot above, there are various categories of education namely, basic.4y, basic.6y, basic.9y, high school, illiterate, professional course, university degree and unknown. If we ignore the unknown variable, the university degree people subscribe most to the term deposit, next in line being high school and then professional course. The number of non-subscribers follow the same order except the third place here is taken by high school people and then the professional course people.

**ggplot**(data, **aes**(x =default,fill =y)) **+**

**geom\_bar**() **+**

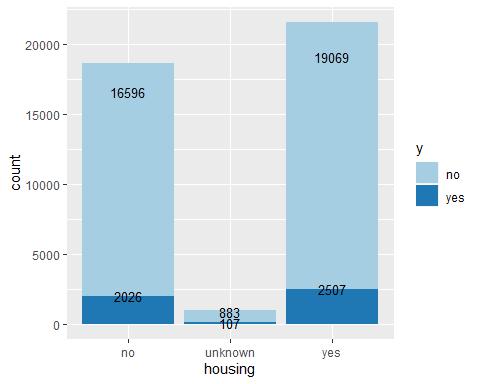
**scale\_fill\_manual**(values=**c**('#999999','#E69F00'))



The default attributes states whether the person has a credit in default or not. Only those people who don’t have credit in default are subscribed to term deposit. The one having credit in default are not subscribed for the term deposit.

**ggplot**(data=data, **aes**(x =housing,fill =y)) **+ geom\_bar**() **+**

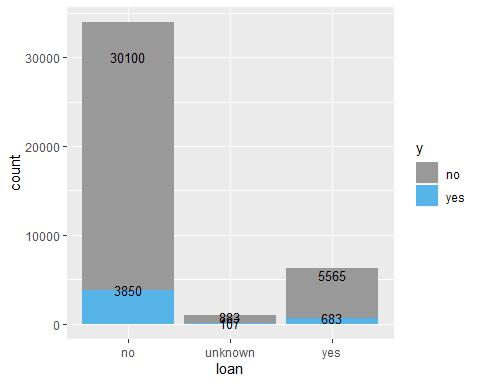
**geom\_text**(stat='count', **aes**(label=..count..),vjust=0.5,size=3.5) **+ scale\_fill\_brewer**(palette="Paired")



Considering the housing plot, the number of people who have taken the housing loan is more than the number of people who haven’t taken the housing loan. But, considering percentage wise, the number of people who haven’t taken the housing loan are more subscribed for the term deposit.

**ggplot**(data=data, **aes**(x =loan,fill =y)) **+ geom\_bar**() **+**

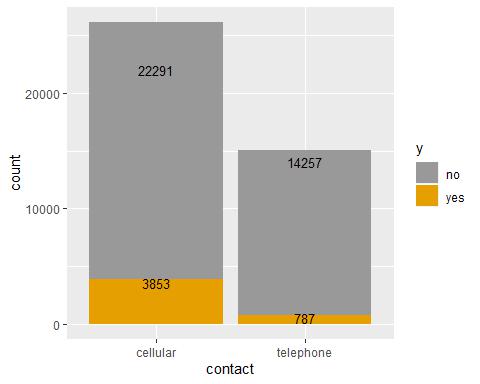
**geom\_text**(stat='count', **aes**(label=..count..),vjust=0.5,size=3.5) **+ scale\_fill\_manual**(values=**c**('#999999','#56B4E9'))



The above plot displays the count of people who have taken loan with respect to the subscriber list. There is very high number of people who haven’t taken the loan as compared to the people who have taken the loan. In addition to that, there are only fewer number of people who haven’t taken the loan and subscribed, as compared to those who have taken the loan.

**ggplot**(data=data, **aes**(x =contact,fill =y)) **+ geom\_bar**() **+**

**geom\_text**(stat='count', **aes**(label=..count..),vjust=.75,size=3.5) **+ scale\_fill\_manual**(values=**c**('#999999','#E69F00'))



The above plot represents the type of communication medium used for contacting the person. There are only two types, the cellular and the telephone. The people who were contacted by cellular medium subscribed more to the term deposit as compared to the people contacted by telephonic medium. There is a higher number of rejections on the cellular side as well.

**library**(Hmisc)

* Loading required package: lattice
* Loading required package: survival
* Loading required package: Formula

##

* Attaching package: 'Hmisc'
* The following objects are masked from 'package:dplyr':
* src, summarize
* The following objects are masked from 'package:base':
* format.pval, units

data**$**month <- **capitalize**(data**$**month)

data**$**month <- **factor**(data**$**month, levels = month.abb)

month <-

data **%>%**

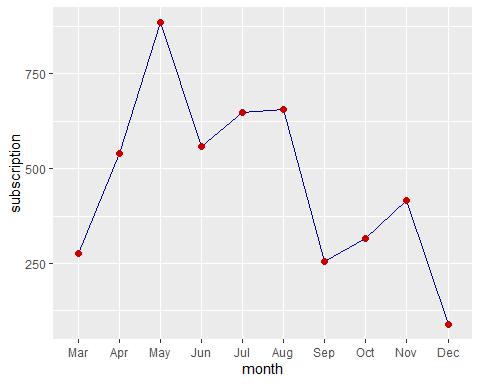
**group\_by**(month) **%>%**

**summarise**(subscription = **sum**(subscribed))

**ggplot**(month, **aes**(month, subscription,group =1)) **+**

**geom\_line**(colour="#000099") **+**

**geom\_point**(size=2,colour="#CC0000")



The above plot shows the average number of subscription for every month. There is peak in the month of may and the number of subscription is the lowest in december. This might be an important variable for predicting subscriptions because the variability among month is higher

data**$**day\_of\_week <- **capitalize**(data**$**day\_of\_week)

data**$**day\_of\_week <- **factor**(data**$**day\_of\_week, levels = **c**('Mon','Tue','Wed','Th u','Fri'))

days <-

data **%>%**

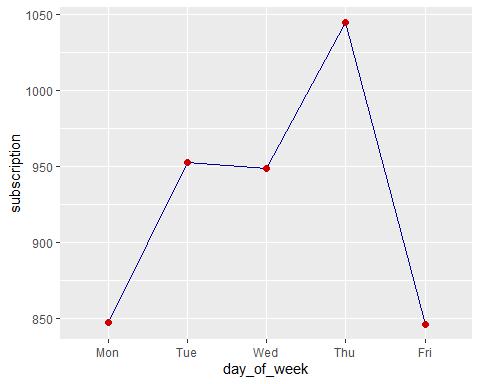
**group\_by**(day\_of\_week) **%>%**

**summarise**(subscription = **sum**(subscribed))

**ggplot**(days, **aes**(day\_of\_week, subscription,group =1)) **+**

**geom\_line**(colour="#000099") **+**

**geom\_point**(size=2,colour="#CC0000")



The above plot shows the average number of subscription for every day of week. The number of subscriptions are higher on thursdays and lowest on Monday and Friday.

month <-

data **%>%**

**group\_by**(month, y) **%>%**

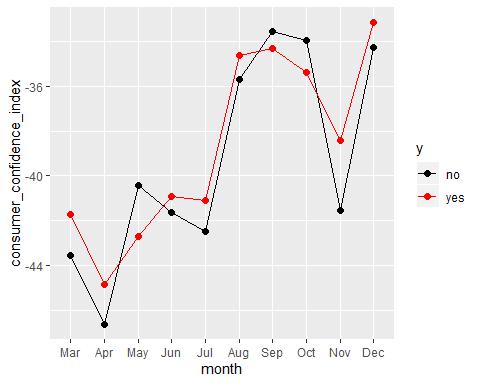
**summarise**(consumer\_confidence\_index = **mean**(cons.conf.idx))

**ggplot**(month, **aes**(month, consumer\_confidence\_index,group =y,color =y)) **+**

**geom\_line**() **+**

**geom\_point**(size=2) **+**

**scale\_color\_manual**(values=**c**("black","red"))



The average trend for consumer confidence index is almost the same for subscribers and non subscribers. The average consumer confidence index is maximum at december and then september, but at the same time the total number of subscription is minimum at december and then september.

month <-

data **%>%**

**group\_by**(month, y) **%>%**

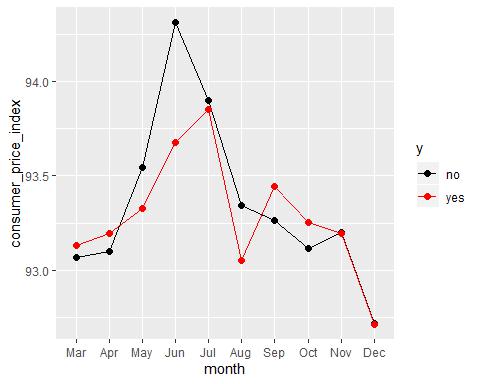
**summarise**(consumer\_price\_index = **mean**(cons.price.idx))

**ggplot**(month, **aes**(month, consumer\_price\_index,group =y,color =y)) **+**

**geom\_line**() **+**

**geom\_point**(size=2) **+**

**scale\_color\_manual**(values=**c**("black","red"))



Considering the average consumer price index by month, the average consumer price index for subscribed clients is greater than that of non-subscribed clients in the month of Jan, Feb, Sep and Oct.

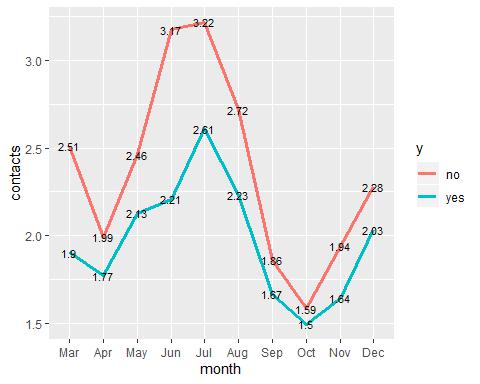
month <-

data **%>%**

**group\_by**(month, y) **%>%**

**summarise**(contacts = **mean**(campaign))

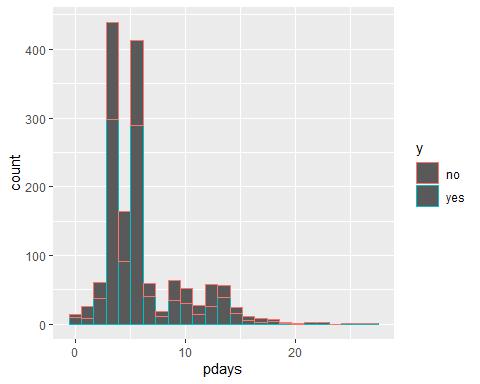
**ggplot**(data =month) **+ geom\_line**(**aes**(x =month,y =contacts,group =y,colour = y),size = 1.25) **+** **geom\_text**(**aes**(x = month, y = contacts, label = **paste**(**r** **ound**(contacts,2))),size =3)



The above graph shows the number of contacts performed during this campaign and the clients. The average number of contacts for subscribed clients are lesser than that of non-subscribed clients in all the months. However the difference between the average number of contacts between subscribed and non-subscribed clients changes from month to month. In addition the average number of contacts varies month to month. For example, in the month of october, on average it only takes 1.5 calls to make a client subscribe but in the month of July it takes 2.61 calls to make a client subscribe.

**ggplot**(data = **filter**(data, pdays **<** 990)) **+**

**geom\_histogram**(**aes**(x=pdays,color =y ),bins =25)



The above plot shows the histogram for pdays. pdays denotes the number of days that passed by after the client was last contacted from a previous campaign.

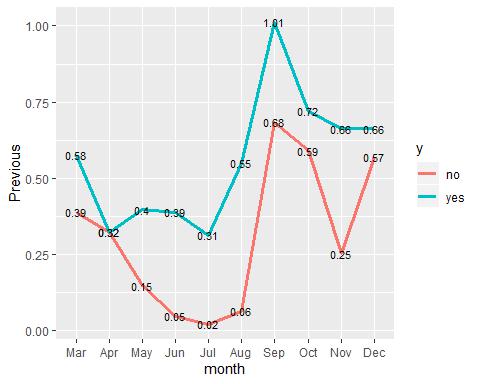
month <-

data **%>%**

**group\_by**(month, y) **%>%**

**summarise**(Previous = **mean**(previous))

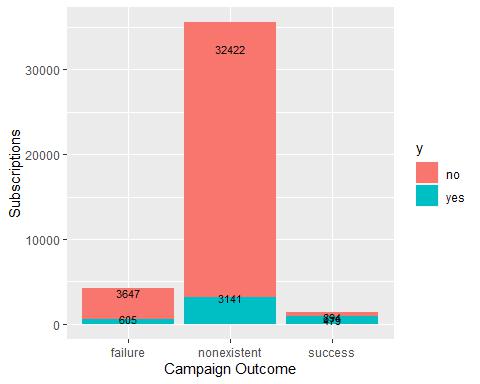
**ggplot**(data =month) **+ geom\_line**(**aes**(x =month,y =Previous,group =y,colour = y),size = 1.25) **+** **geom\_text**(**aes**(x = month, y = Previous, label = **paste**(**r** **ound**(Previous,2))),size =3)



The above plot shows the average number of contacts performed before this campaign and for every client in every month. Unlike the average number of contacts perfomed in this campaign, the average number of contacts for subscribed clients are higher than that of non subscribed clients.

**ggplot**(data,**aes**(x=poutcome,fill=y)) **+ geom\_bar**()**+ geom\_text**(stat ='count'

* **aes**(label=..count..),size =3) **+ labs**(x ="Campaign Outcome",y ="Subscriptions")



The above bar plot shows the number of subscribers and non subscribers by the outcome of the previous marketing campaign. If the previous outcome is a failure then the client is more likely to not subscribe and if the previous outcome is a success then the client is more likely to subscribe.