Exploratory data analysis

Montserrat Guillen

2017

Contents

Introduction

This is a very short introduction to the exploration of data using RStudio and Rmarkdown.

This document sequentially applies a set of Data Science techniques to gain insights from the Direct Marketing campaign of a Portuguese Banking Institution.

There are two public data sets that are linked to the article by S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014. The two datasets ontain similar information, but not exactly the same.. Here we will analyse the smaller data set (called **bank.csv**). The file can be downloaded from: https://archive.ics.uci.edu/ml/datasets/bank+marketing

or (for this course)

http://www.ub.edu/rfa/docs/DATA/bank.csv

First we need to read the data from the file “bank.csv”.

```
#setwd("..")
### CHUNK 1

bank<-read.table(file="bank.csv",header=T,sep=";")
```

The dataset contains information on `4521` clients and `17`variables.

Note that the input variables are not the same in this file than in the “additional” data set, that has different attributes. There are many recent analysis of all these data but one has to check which exact data file is used in each case.

Input variables:

# bank client data:

1 - age (numeric)

2 - job : type of job (categorical: “admin.”,“unknown”,“unemployed”,“management”,“housemaid”,“entrepreneur”,“student”, “blue-collar”,“self-employed”,“retired”,“technician”,“services”)

3 - marital : marital status (categorical: “married”,“divorced”,“single”; note: “divorced” means divorced or widowed)

4 - education (categorical: “unknown”,“secondary”,“primary”,“tertiary”)

5 - default: has credit in default? (binary: “yes”,“no”)

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: “yes”,“no”)

8 - loan: has personal loan? (binary: “yes”,“no”)

# related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: “unknown”,“telephone”,“cellular”)

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: “jan”, “feb”, “mar”, . . . , “nov”, “dec”)

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: “unknown”,“other”,“failure”,“success”)

Output variable (desired target):

17 - **y - has the client subscribed a term deposit? (binary: “yes”,“no”)**

Names of the variables

We print de names of the variables:

```
### CHUNK 2

colnames(bank)
```

```
##  [1] "age"       "job"       "marital"   "education" "default"  
##  [6] "balance"   "housing"   "loan"      "contact"   "day"      
## [11] "month"     "duration"  "campaign"  "pdays"     "previous" 
## [16] "poutcome"  "y"
```

We will use function **attach** so that we can call variables just by their name instead of **bank$name**.

```
### CHUNK 3

attach(bank)

# search() tells you the search order for objects:
search()
```

```
##  [1] ".GlobalEnv"        "bank"              "package:stats"    
##  [4] "package:graphics"  "package:grDevices" "package:utils"    
##  [7] "package:datasets"  "package:methods"   "Autoloads"        
## [10] "package:base"
```

An overview of basic Data Analysis for this dataset

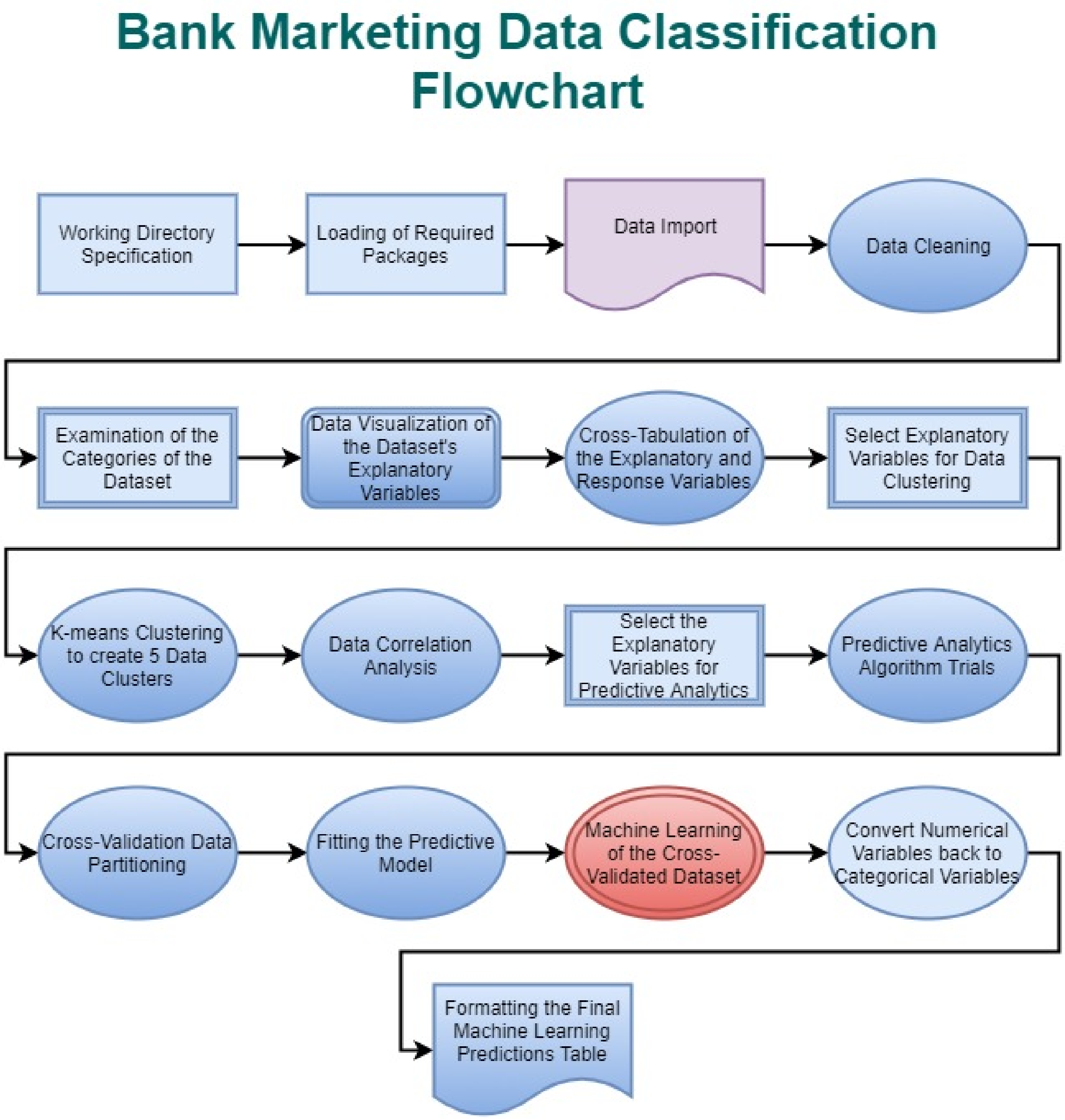

Bank Marketing Data Classification Flowchart
Working Directory Specification
Loading of Required Packages
Data Import
Data Cleaning
Examination of the Categories of the Dataset
Data Visualization of the Dataset's Explanatory Variables
Cross-Tabulation of the Explanatory and Response Variables
Select Explanatory Variables for Data Clustering
K-means Clustering to create 5 Data Clusters
Data Correlation Analysis
Select the Explanatory Variables for Predictive Analytics
Predictive Analytics Algorithm Trials
Cross-Validation Data Partitioning
Fitting the Predictive Model
Machine Learning of the Cross-Validated Dataset
Convert Numerical Variables back to Categorical Variables
Formatting the Final Machine Learning Predictions Table

image:

Required packages

The function, "install.packages()", downloads and installs R programming language packages from CRAN-like repositories or from local files. If these packages are not installed, they shoule be installed before running the code.

```
### CHUNK 4

# I've set warnings=FALSE to avoid warnings on packages name collisions.
# Required Packages
# install.packages("ggplot2")   # plotting
# install.packages("dplyr")     # data management
# install.packages("cluster")   # kmeans clustering
# install.packages("HSAUR")     # silhouette plotting
# install.packages("fpc")       # numbers cluster plot
# install.packages("lattice")   # cluster plotting
# install.packages("rpart")     # Decision Tress data classification
# install.packages("kernlab")   # Support Vector Machines machine learning
# install.packages("randomForest") # Random Forest machine learning

library(ggplot2)
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(cluster)
#library(HSAUR)
#library(fpc)
#library(lattice)
#library(rpart)
#library(kernlab)
#library(randomForest)
```

Session information

This is information on the R version used in this example:

```
### CHUNK 5

sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 16299)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
## [3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
## [5] LC_TIME=Spanish_Spain.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
## [1] dplyr_0.7.4   ggplot2_2.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.14     bindr_0.1        knitr_1.17       magrittr_1.5
##  [5] munsell_0.4.3    colorspace_1.3-2 R6_2.2.2         rlang_0.1.4
##  [9] stringr_1.2.0    plyr_1.8.4       tools_3.4.3      grid_3.4.3
## [13] gtable_0.2.0     htmltools_0.3.6  yaml_2.1.15      lazyeval_0.2.1
## [17] rprojroot_1.2    digest_0.6.12    assertthat_0.2.0 tibble_1.3.4
## [21] bindrcpp_0.2     glue_1.2.0       evaluate_0.10.1  rmarkdown_1.8
## [25] stringi_1.1.6    compiler_3.4.3   scales_0.5.0     backports_1.1.1
## [29] pkgconfig_2.0.1
```

Print of first six records (all variables)

```
### CHUNK 6

head(bank)
```

```
##   age         job marital education default balance housing loan  contact
## 1  30  unemployed married   primary      no    1787      no   no cellular
## 2  33    services married secondary      no    4789     yes  yes cellular
## 3  35  management  single  tertiary      no    1350     yes   no cellular
## 4  30  management married  tertiary      no    1476     yes  yes  unknown
## 5  59 blue-collar married secondary      no       0     yes   no  unknown
## 6  35  management  single  tertiary      no     747      no   no cellular
##   day month duration campaign pdays previous poutcome  y
## 1  19   oct       79        1    -1        0  unknown no
## 2  11   may      220        1   339        4  failure no
## 3  16   apr      185        1   330        1  failure no
## 4   3   jun      199        4    -1        0  unknown no
## 5   5   may      226        1    -1        0  unknown no
## 6  23   feb      141        2   176        3  failure no
```

We can also use:

```
### CHUNK 7

glimpse(bank)
```

```
## Observations: 4,521
## Variables: 17
## $ age       <int> 30, 33, 35, 30, 59, 35, 36, 39, 41, 43, 39, 43, 36, ...
## $ job       <fctr> unemployed, services, management, management, blue-...
## $ marital   <fctr> married, married, single, married, married, single,...
## $ education <fctr> primary, secondary, tertiary, tertiary, secondary, ...
## $ default   <fctr> no, no, no, no, no, no, no, no, no, no, no, no, no,...
## $ balance   <int> 1787, 4789, 1350, 1476, 0, 747, 307, 147, 221, -88, ...
## $ housing   <fctr> no, yes, yes, yes, yes, no, yes, yes, yes, yes, yes...
## $ loan      <fctr> no, yes, no, yes, no, no, no, no, no, yes, no, no, ...
## $ contact   <fctr> cellular, cellular, cellular, unknown, unknown, cel...
## $ day       <int> 19, 11, 16, 3, 5, 23, 14, 6, 14, 17, 20, 17, 13, 30,...
## $ month     <fctr> oct, may, apr, jun, may, feb, may, may, may, apr, m...
## $ duration  <int> 79, 220, 185, 199, 226, 141, 341, 151, 57, 313, 273,...
## $ campaign  <int> 1, 1, 1, 4, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 5, 1...
## $ pdays     <int> -1, 339, 330, -1, -1, 176, 330, -1, -1, 147, -1, -1,...
## $ previous  <int> 0, 4, 1, 0, 0, 3, 2, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 2...
## $ poutcome  <fctr> unknown, failure, failure, unknown, unknown, failur...
## $ y         <fctr> no, no, no, no, no, no, no, no, no, no, no, no, no,...
```

Data visualization

Data Summary of the Bank Dataset

We check all variables and conclude a few on interesting things about our data.

```
### CHUNK 8

dim(bank)
```

```
## [1] 4521   17
```

```
summary(bank)
```

```
##       age                 job          marital          education   
##  Min.   :19.00   management :969   divorced: 528   primary  : 678  
##  1st Qu.:33.00   blue-collar:946   married :2797   secondary:2306  
##  Median :39.00   technician :768   single  :1196   tertiary :1350  
##  Mean   :41.17   admin.     :478                   unknown  : 187  
##  3rd Qu.:49.00   services   :417                                   
##  Max.   :87.00   retired    :230                                   
##                  (Other)    :713                                   
##  default       balance      housing     loan           contact    
##  no :4445   Min.   :-3313   no :1962   no :3830   cellular :2896  
##  yes:  76   1st Qu.:   69   yes:2559   yes: 691   telephone: 301  
##             Median :  444                         unknown  :1324  
##             Mean   : 1423                                         
##             3rd Qu.: 1480                                         
##             Max.   :71188                                         
##                                                                   
##       day            month         duration       campaign     
##  Min.   : 1.00   may    :1398   Min.   :   4   Min.   : 1.000  
##  1st Qu.: 9.00   jul    : 706   1st Qu.: 104   1st Qu.: 1.000  
##  Median :16.00   aug    : 633   Median : 185   Median : 2.000  
##  Mean   :15.92   jun    : 531   Mean   : 264   Mean   : 2.794  
##  3rd Qu.:21.00   nov    : 389   3rd Qu.: 329   3rd Qu.: 3.000  
##  Max.   :31.00   apr    : 293   Max.   :3025   Max.   :50.000  
##                  (Other): 571                                  
##      pdays           previous          poutcome      y       
##  Min.   : -1.00   Min.   : 0.0000   failure: 490   no :4000  
##  1st Qu.: -1.00   1st Qu.: 0.0000   other  : 197   yes: 521  
##  Median : -1.00   Median : 0.0000   success: 129             
##  Mean   : 39.77   Mean   : 0.5426   unknown:3705             
##  3rd Qu.: -1.00   3rd Qu.: 0.0000                            
##  Max.   :871.00   Max.   :25.0000                            
## 
```

What about term diposit and default? Is it possible?

```r
### CHUNK 9

table(y,default)
```

```
##      default
## y       no  yes
##   no  3933   67
##   yes  512    9
```

Who are these 9 people?

```r
### CHUNK 10

default_termdip=subset(bank, default=='yes' & y=='yes')
glimpse(default_termdip)
```

```
## Observations: 9
## Variables: 17
## $ age       <int> 49, 41, 56, 39, 41, 55, 30, 36, 32
## $ job       <fctr> entrepreneur, blue-collar, housemaid, technician, b...
## $ marital   <fctr> divorced, married, divorced, divorced, single, marr...
## $ education <fctr> unknown, secondary, primary, tertiary, secondary, s...
## $ default   <fctr> yes, yes, yes, yes, yes, yes, yes, yes, yes
## $ balance   <int> -701, 720, 1238, 3, -386, -308, 239, 12, -53
## $ housing   <fctr> yes, no, no, no, no, no, yes, no, yes
## $ loan      <fctr> no, yes, no, no, yes, no, no, no, no
## $ contact   <fctr> cellular, cellular, unknown, cellular, cellular, ce...
## $ day       <int> 30, 24, 5, 6, 20, 2, 21, 12, 16
## $ month     <fctr> jul, jul, jun, may, nov, feb, may, aug, apr
## $ duration  <int> 988, 651, 1558, 488, 477, 781, 412, 587, 648
## $ campaign  <int> 2, 1, 1, 1, 1, 1, 1, 2, 1
## $ pdays     <int> -1, -1, -1, -1, -1, -1, -1, -1, 272
## $ previous  <int> 0, 0, 0, 0, 0, 0, 0, 0, 1
## $ poutcome  <fctr> unknown, unknown, unknown, unknown, unknown, unknow...
## $ y         <fctr> yes, yes, yes, yes, yes, yes, yes, yes, yes
```

Specific statistical measures

```r
### CHUNK 11

sapply(bank[c("age", "duration")], median, 1)
```

```
##      age duration
##       39      185
```

```r
### CHUNK 12

tapply(age, y, median)
```

```
##  no yes
##  39  40
```

```r
tapply(duration, y, median)
```

```
##  no yes
## 167 442
```

Tables and proportions

```
### CHUNK 13

table(y)
```

```
## y
##   no  yes
## 4000  521
```

```
prop.table(table(y))
```

```
## y
##      no     yes
## 0.88476 0.11524
```

```
round(prop.table(table(y))*100, 2)
```

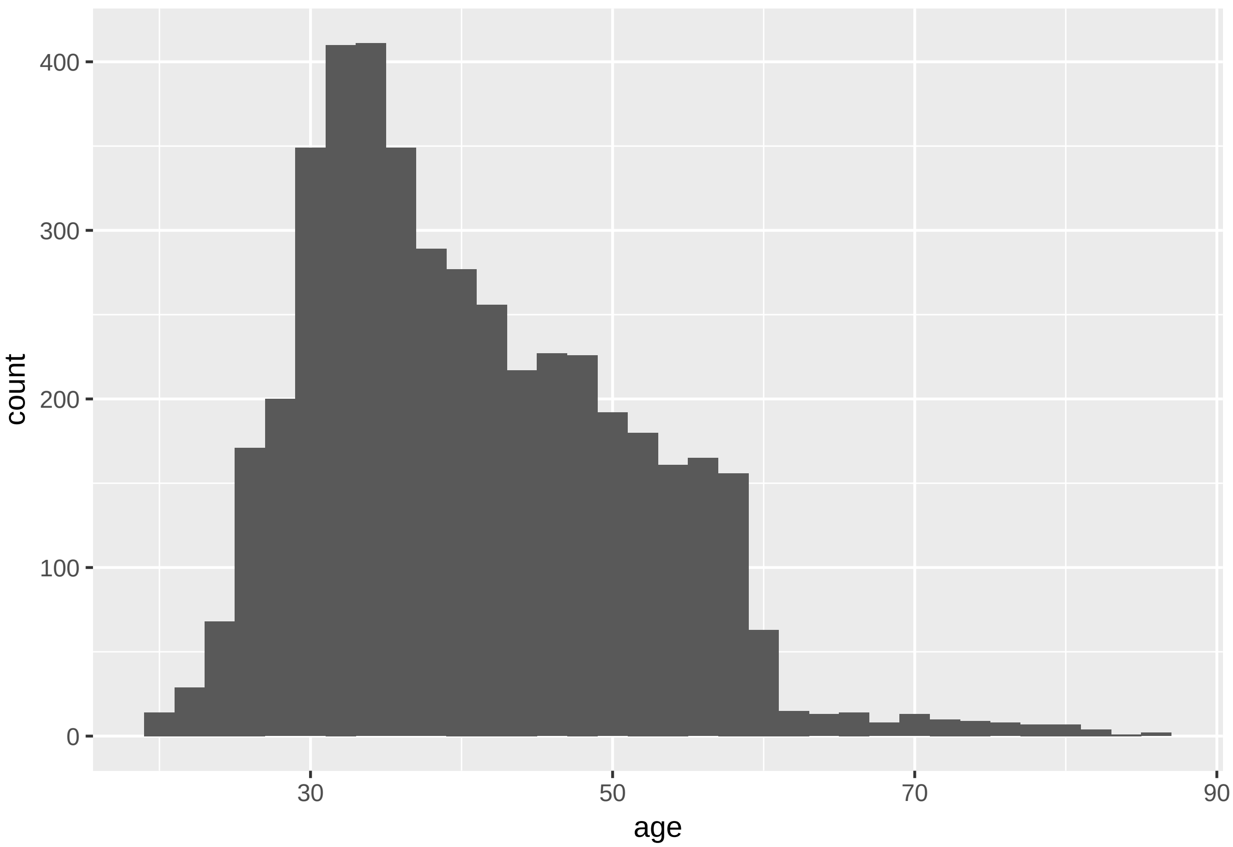
```
## y
##    no   yes
## 88.48 11.52
```

Histograms of age and duration

Regular histograms

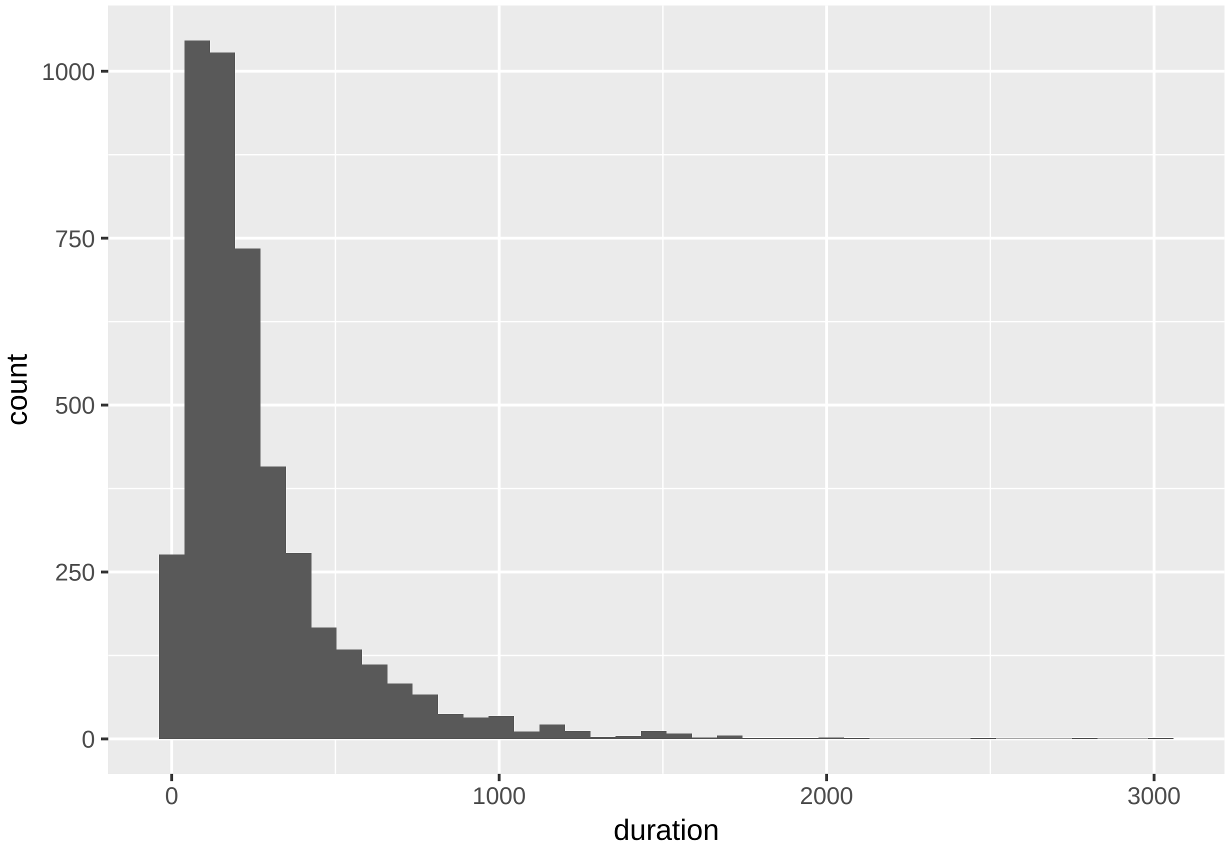
Easy histogram with 35 bins and label.

```
### CHUNK 14

ggplot(data=bank, aes(age)) + geom_histogram(bins=35)+xlab("age")
```



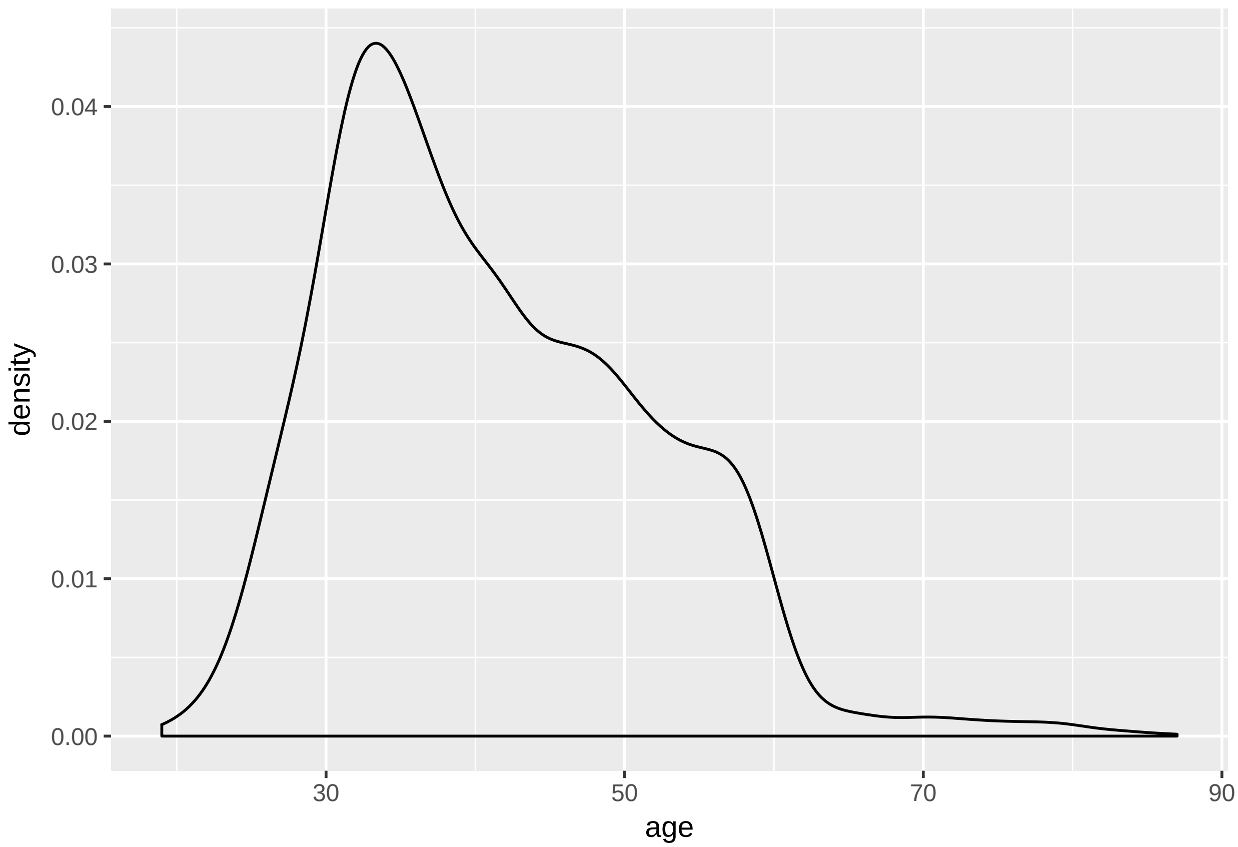
Easy histogram with 40 bins and lanel

```
### CHUNK 15

ggplot(data=bank, aes(duration)) + geom_histogram(bins=40)+xlab("duration")
```



With density plots

```
### CHUNK 16

ggplot(data=bank, aes(age)) + geom_density()
```



Plot colors

Palettes:

Diverging BrBG, PiYG, PRGn, PuOr, RdBu, RdGy, RdYlBu, RdYlGn, Spectral

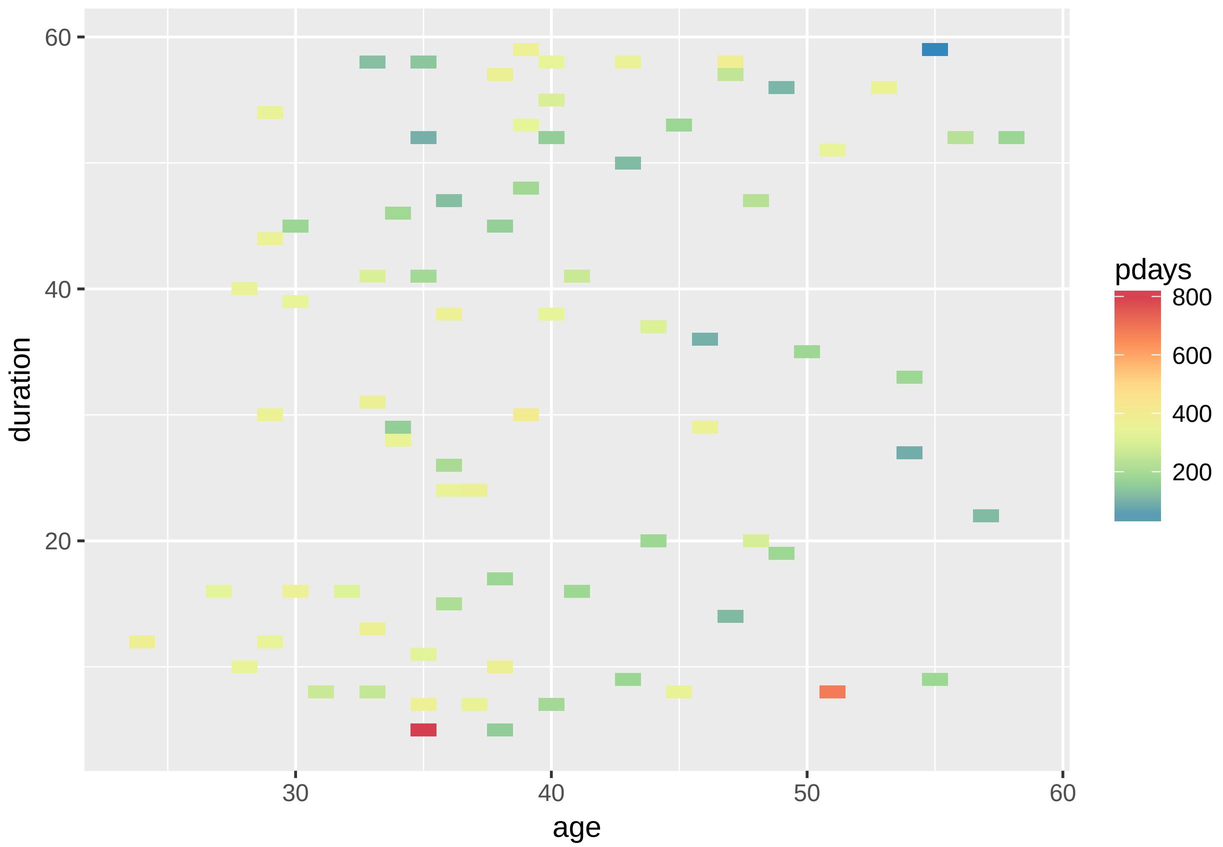
Qualitative Accent, Dark2, Paired, Pastel1, Pastel2, Set1, Set2, Set3

Sequential Blues, BuGn, BuPu, GnBu, Greens, Greys, Oranges, OrRd, PuBu, PuBuGn, PuRd, Purples, RdPu, Reds, YlGn, YlGnBu, YlOrBr, YlOrRd

Mutiple dimension graphics

We have a problem with the variable **pdays**, because when the customer was not contacted, the value is -1.

```
### CHUNK 17

  t2=subset(bank, pdays>=0 & duration<60)
 ggplot(t2) +
  geom_tile(aes(age, duration,fill = pdays))+
   scale_fill_distiller(palette = "Spectral")
```

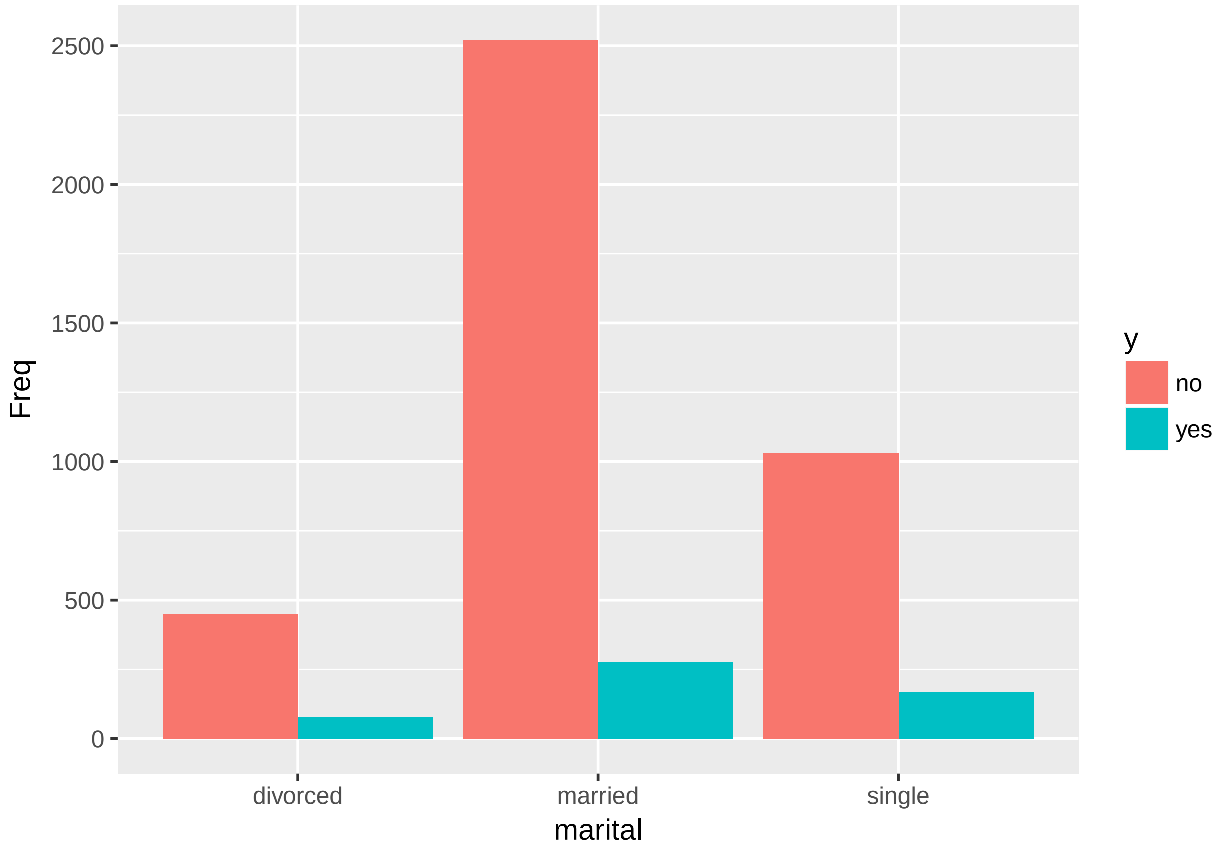


Plots

Vertical bars:

Grouping and showing frequencies

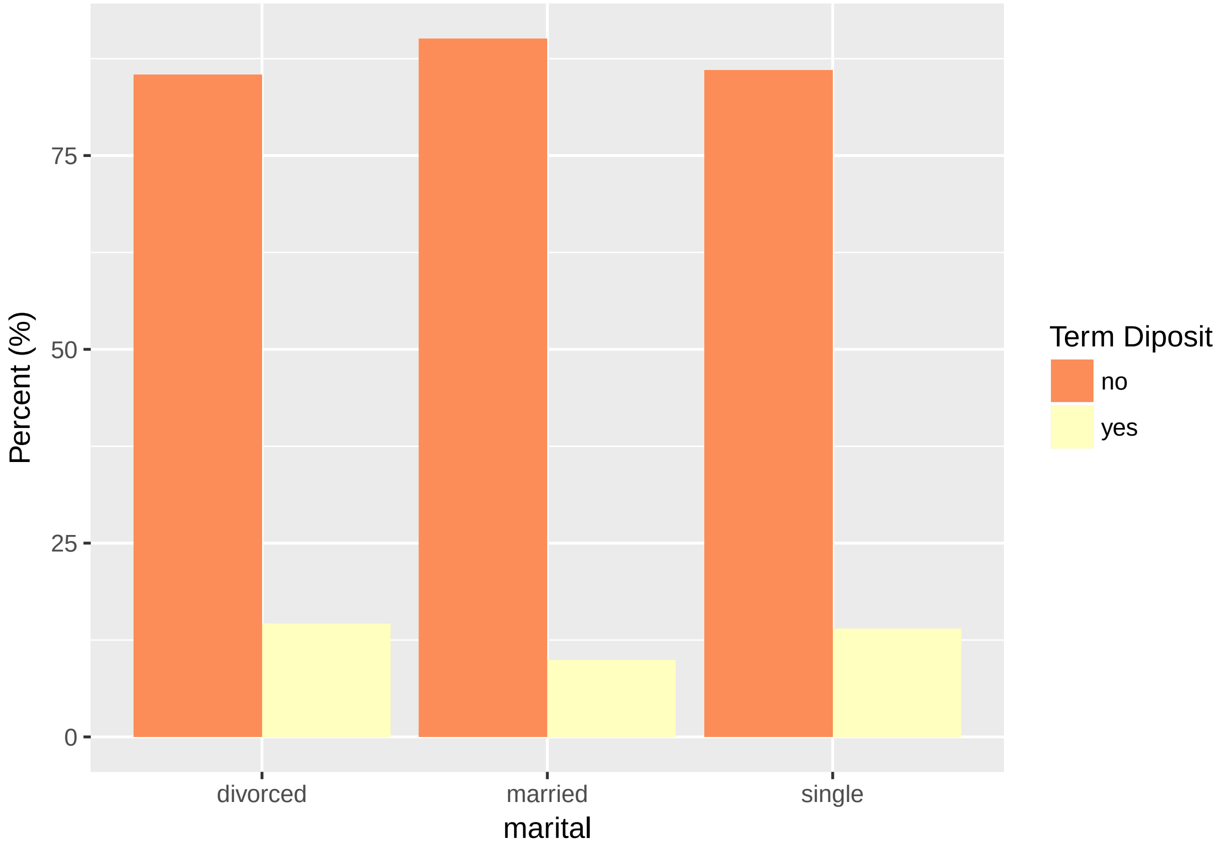
```
### CHUNK 18

t=data.frame(table(y, marital))
ggplot(t, aes(x=marital, y=Freq, fill=y)) +
  geom_bar(position='dodge', stat='identity')
```



Grouping bars and showing percent

You can try a number of different palettes “Greens”, “Set1”, “Set2”, etc…

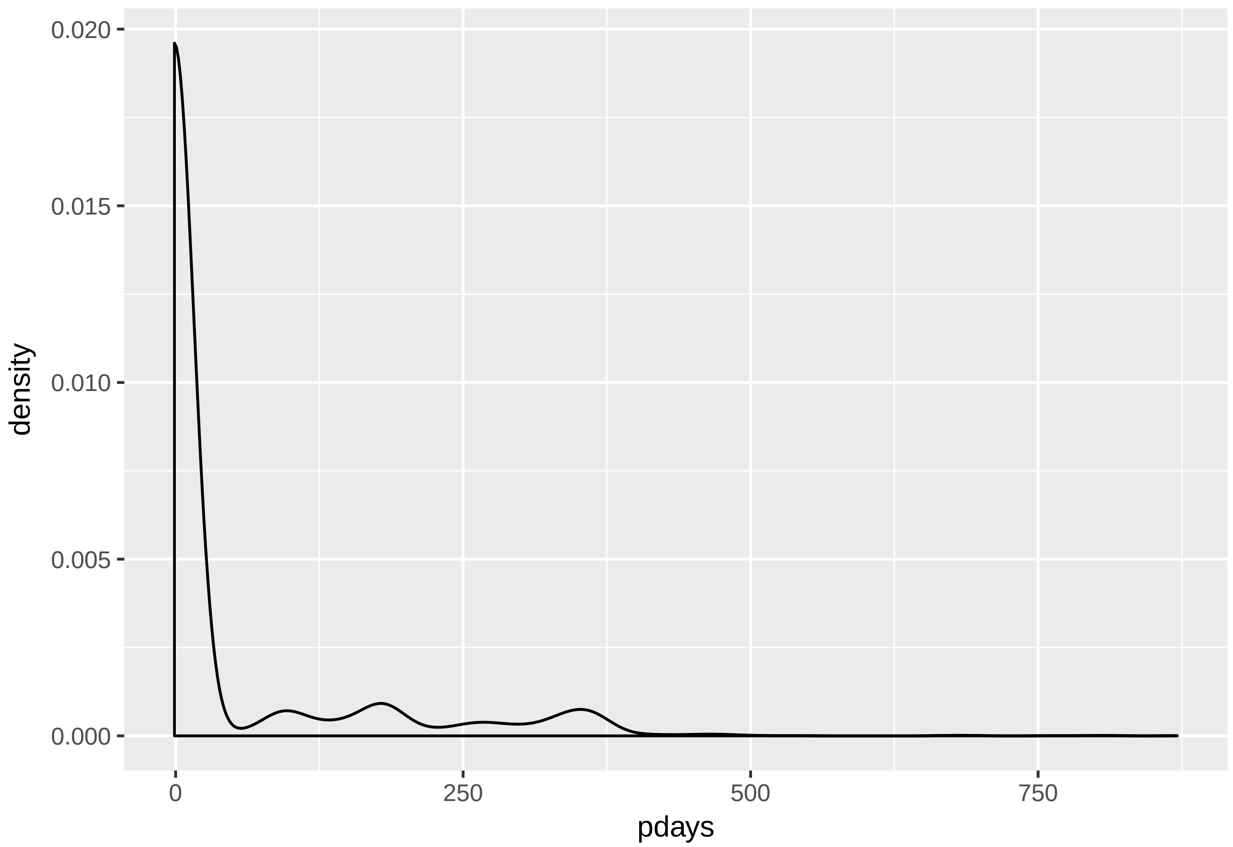
```
### CHUNK 19

t=data.frame(prop.table(table(y ,marital), 2))
ggplot(t, aes(x=marital, y=Freq*100, fill=y)) +
  geom_bar(position='dodge', stat='identity')+
  ylab("Percent (%)")+ scale_fill_brewer("Term Diposit", palette="Spectral")
```



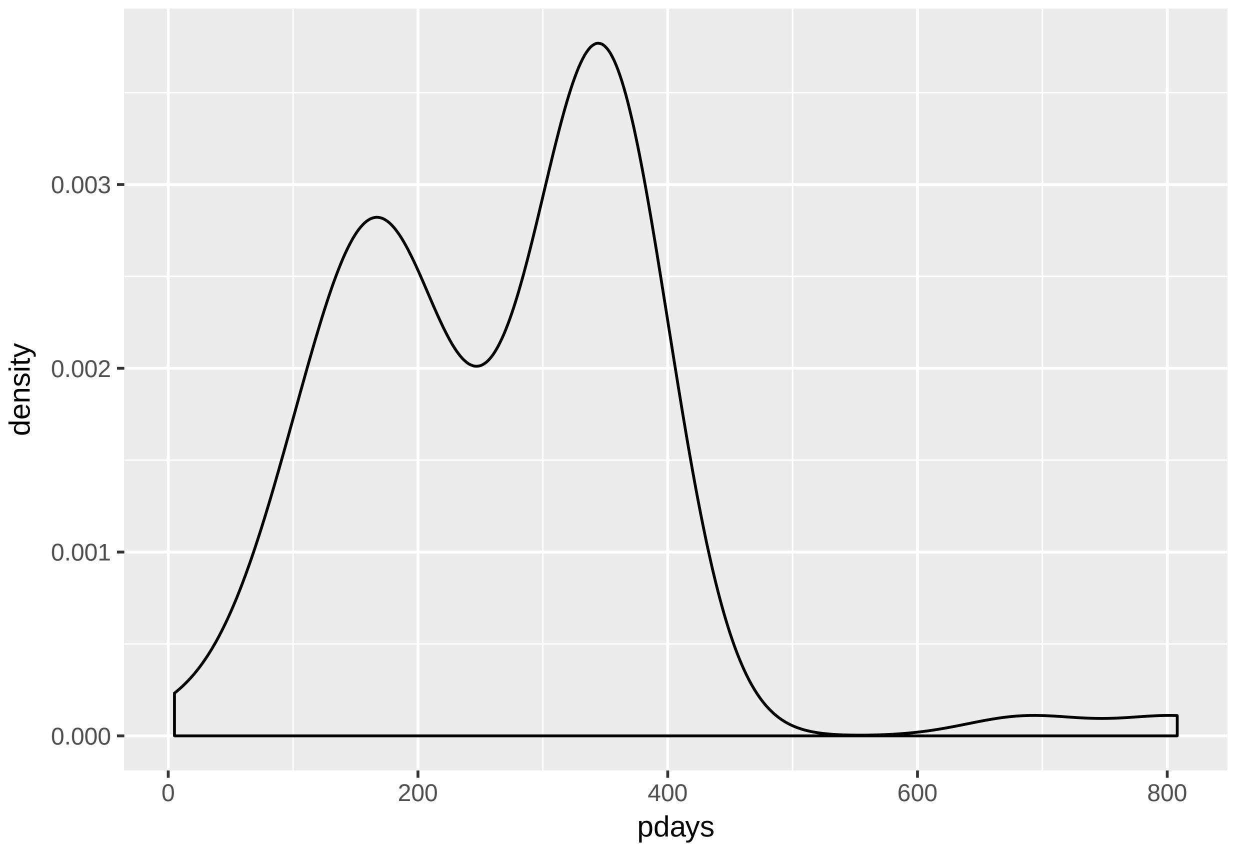
Miscelaneous plots and boxplots

The simplest thing to do is a box plot.

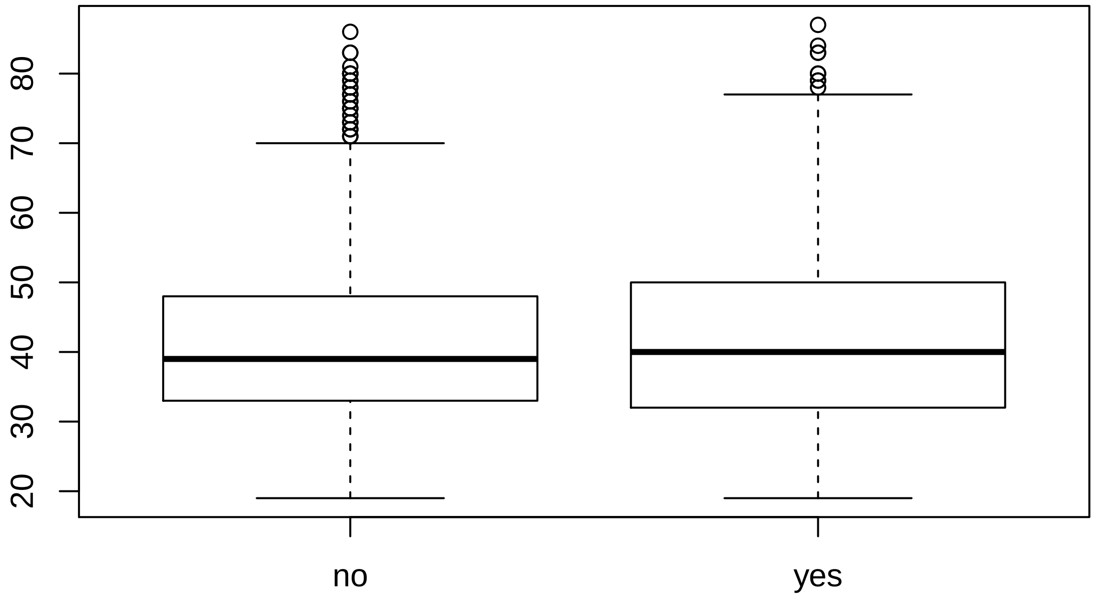
```
### CHUNK 20

ggplot(data=bank, aes(pdays))+ geom_density()+scale_fill_brewer()
```



```r
ggplot(data=t2, aes(pdays))+ geom_density()+scale_fill_brewer()
```

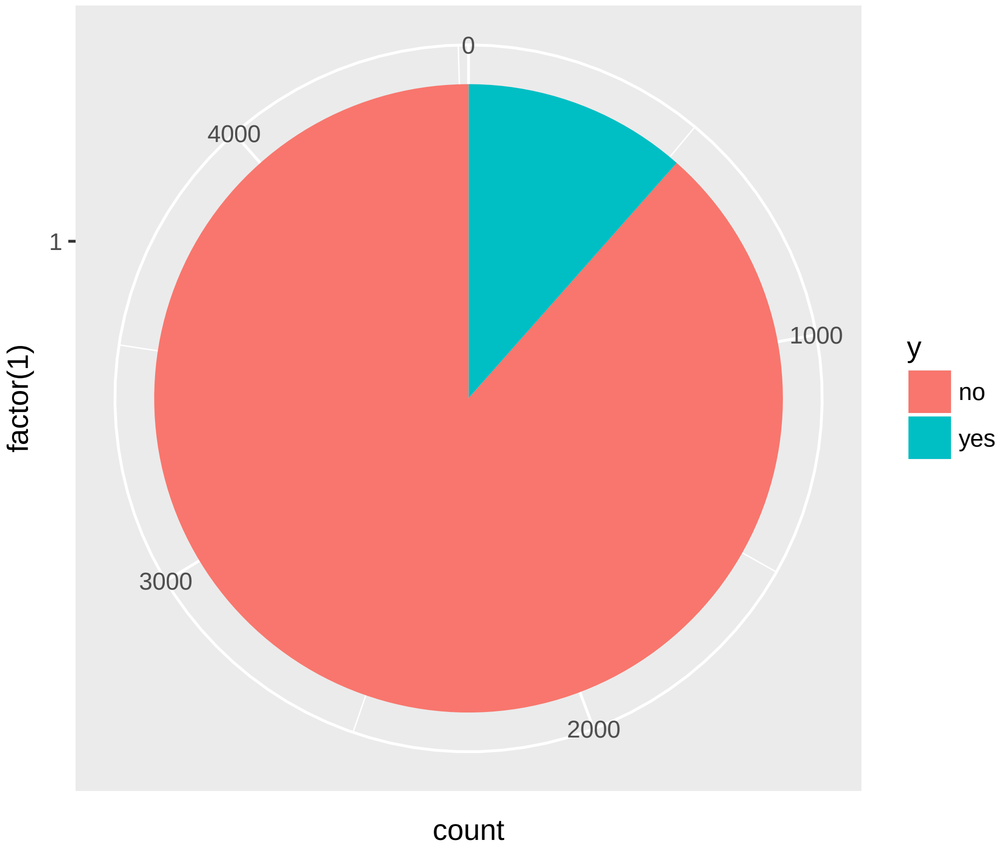


```r
plot(y, age)
```



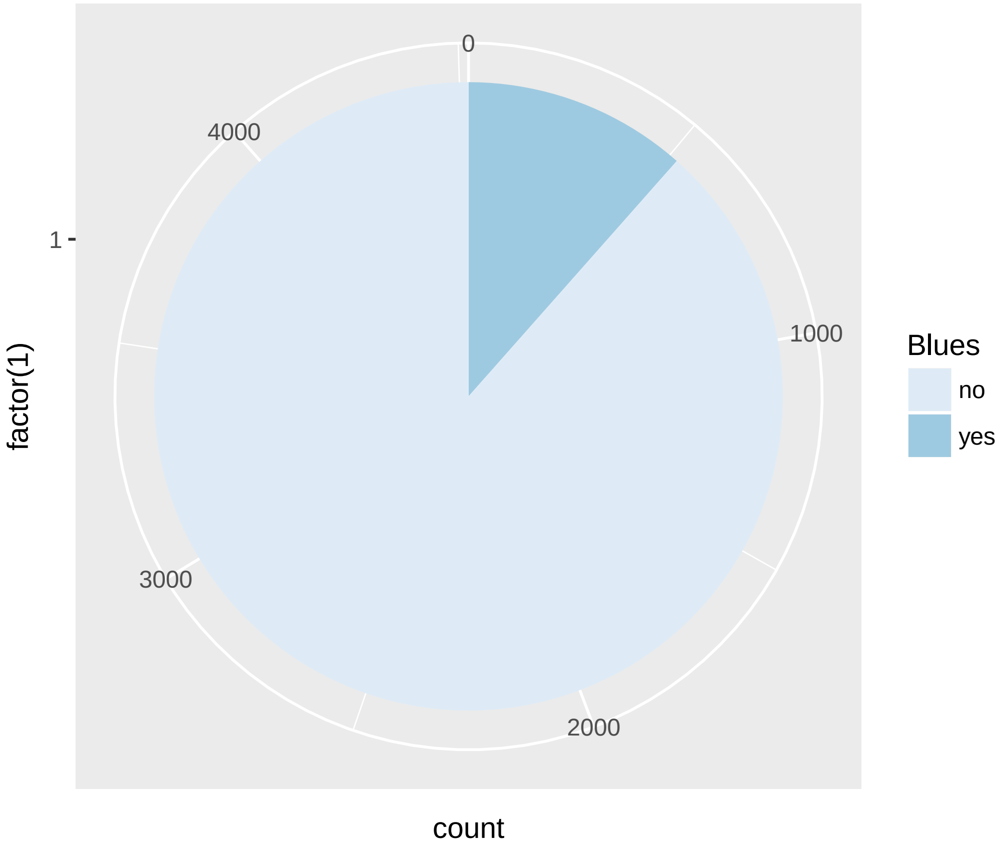
Simple Pie charts

```
### CHUNK 21

ggplot(bank, aes(x=factor(1), fill=y))+
  geom_bar(width = 1)+
  coord_polar("y")
```

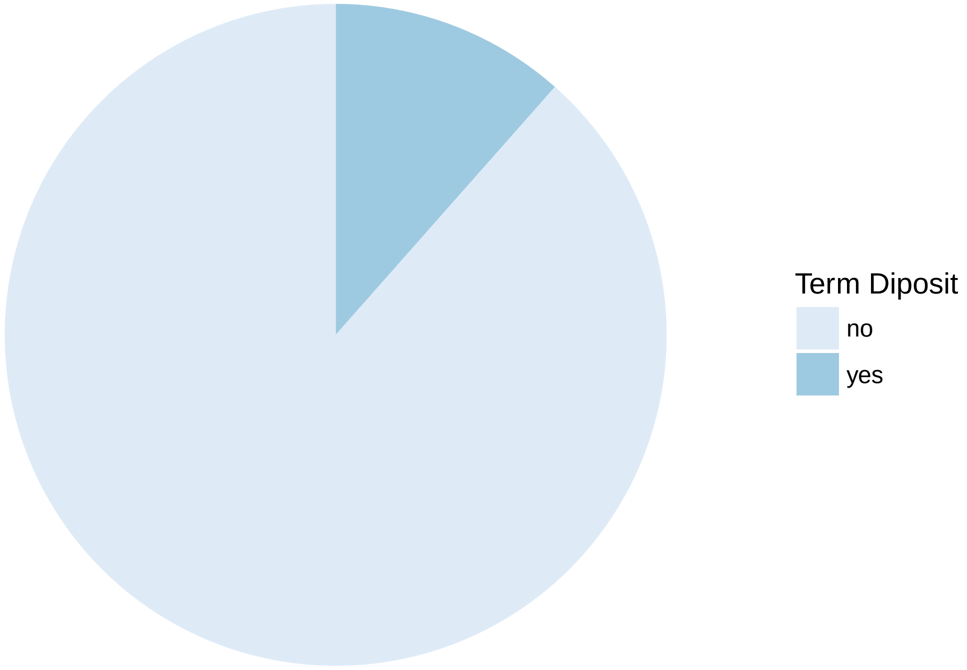


We now use another palette and labeling.

```
### CHUNK 22

ggplot(bank, aes(x=factor(1), fill=y))+
  geom_bar(width = 1)+
  coord_polar("y")+ scale_fill_brewer("Blues")
```



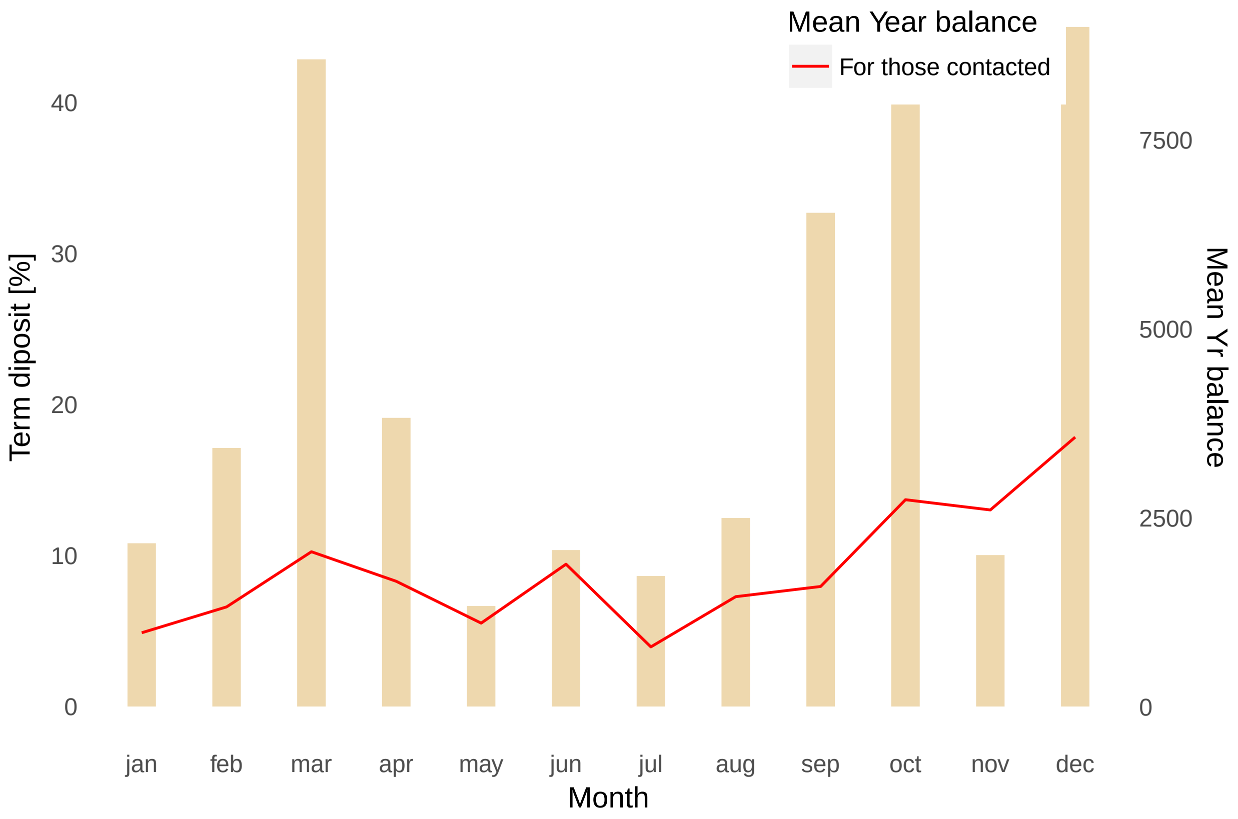
```
blank_theme <- theme_minimal()+
  theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.border = element_blank(),
  panel.grid=element_blank(),
  axis.ticks = element_blank(),
  plot.title=element_text(size=14, face="bold")
  )

ggplot(bank, aes(x=factor(1), fill=y))+
  geom_bar(width = 1)+
  coord_polar("y")+ scale_fill_brewer("Term Diposit")+ blank_theme +
  theme(axis.text.x=element_blank())+
  theme(axis.text.y=element_blank())
```



Evolution over time

Example of a double scale graphic

```
### CHUNK 23

t3=data.frame(prop.table(table(y, month),2))
t3$month_order=factor(as.character(t3$month), levels = c("jan","feb","mar", "apr", "may", "jun","jul",

t4 = group_by(bank, month) %>% summarise(Yearlybalance=mean(balance)) %>% ungroup()
t4$month_order=factor(as.character(t4$month), levels = c("jan","feb","mar", "apr", "may", "jun","jul",


ggplot(subset(t3, y=='yes'), aes(x=month_order, y=Freq*100)) +
  theme(plot.background = element_blank(),
              # panel.grid.minor = element_blank(),
              #  panel.grid.major = element_blank(),
                panel.border = element_blank(),
                panel.background = element_blank(),
                axis.ticks = element_blank() ) +
    geom_linerange(subset(t3, y=='yes'), mapping=aes(x=month_order, ymin=0, ymax=Freq*100), colour = "w
   geom_line(t4, mapping=aes(x=month_order, y=Yearlybalance/200, group=1, colour= "For those contacted"
scale_y_continuous(sec.axis = sec_axis(~.*200, name = "Mean Yr balance")) +
scale_colour_manual(values = c("red")) +
 labs(y = "Term diposit [%]",
               x = " Month ",
               colour = "Mean Year balance") +
 theme(legend.position = c(0.8, 0.9))
```



A function that produces a series of graphics

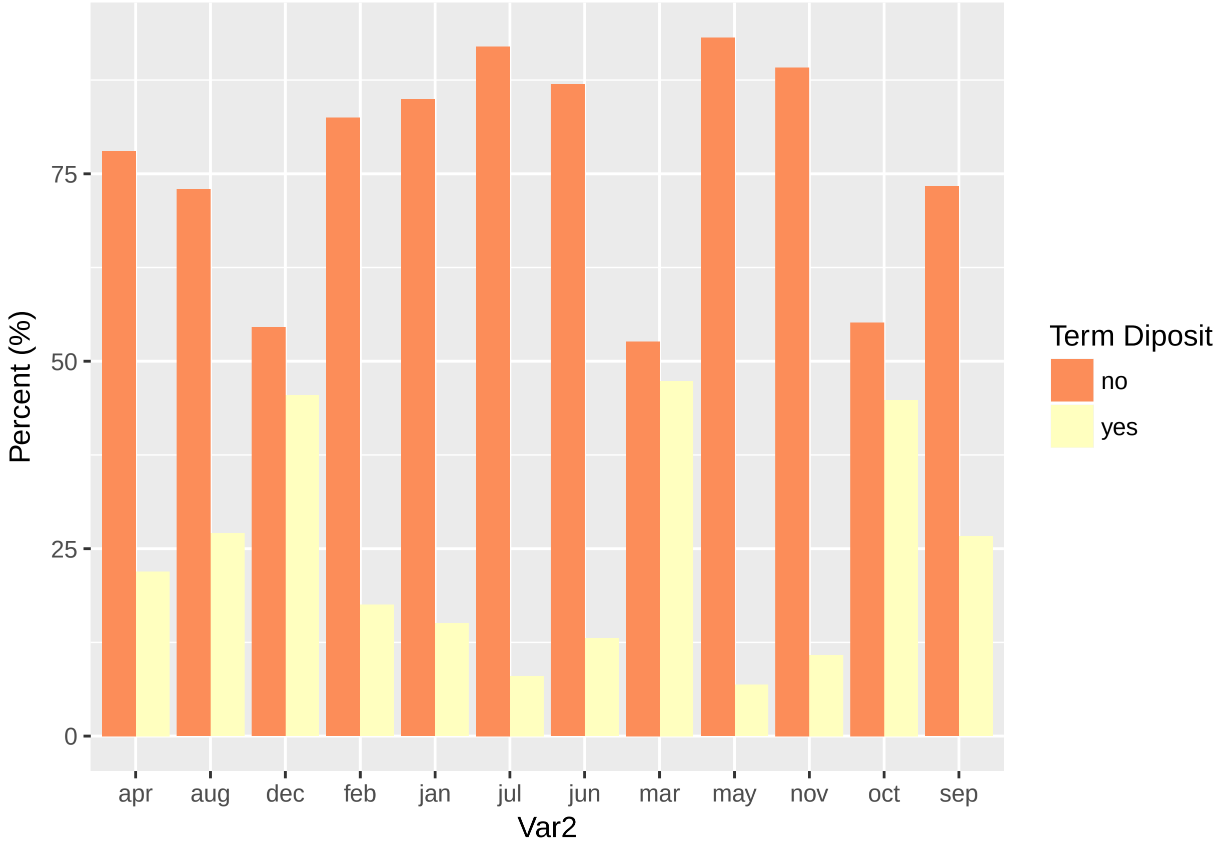
```
### CHUNK 24

table(bank$campaign)
```

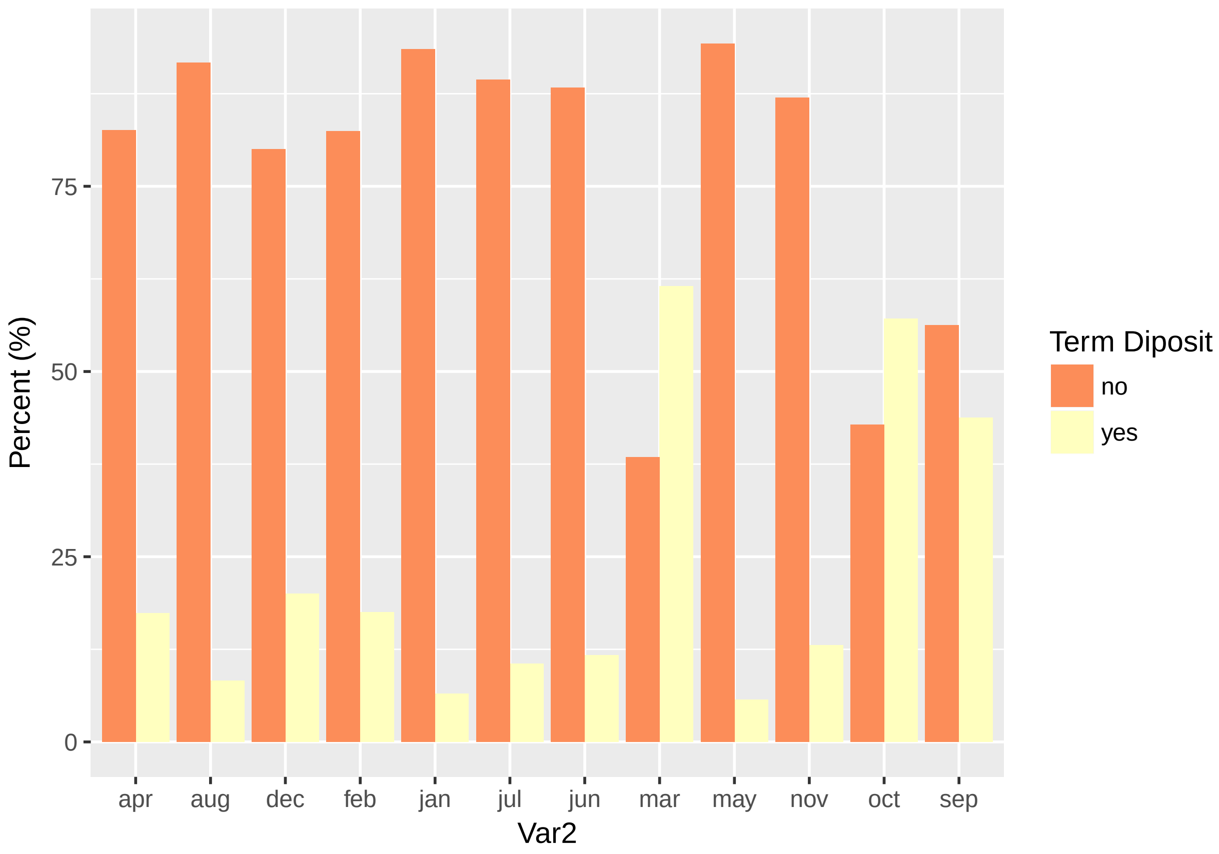
```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 1734 1264  558  325  167  155   75   56   30   27   22   21   17   10    9
##   16   17   18   19   20   21   22   23   24   25   28   29   30   31   32
##    8    7    7    3    3    2    2    2    3    4    3    1    1    1    2
##   44   50
##    1    1
```

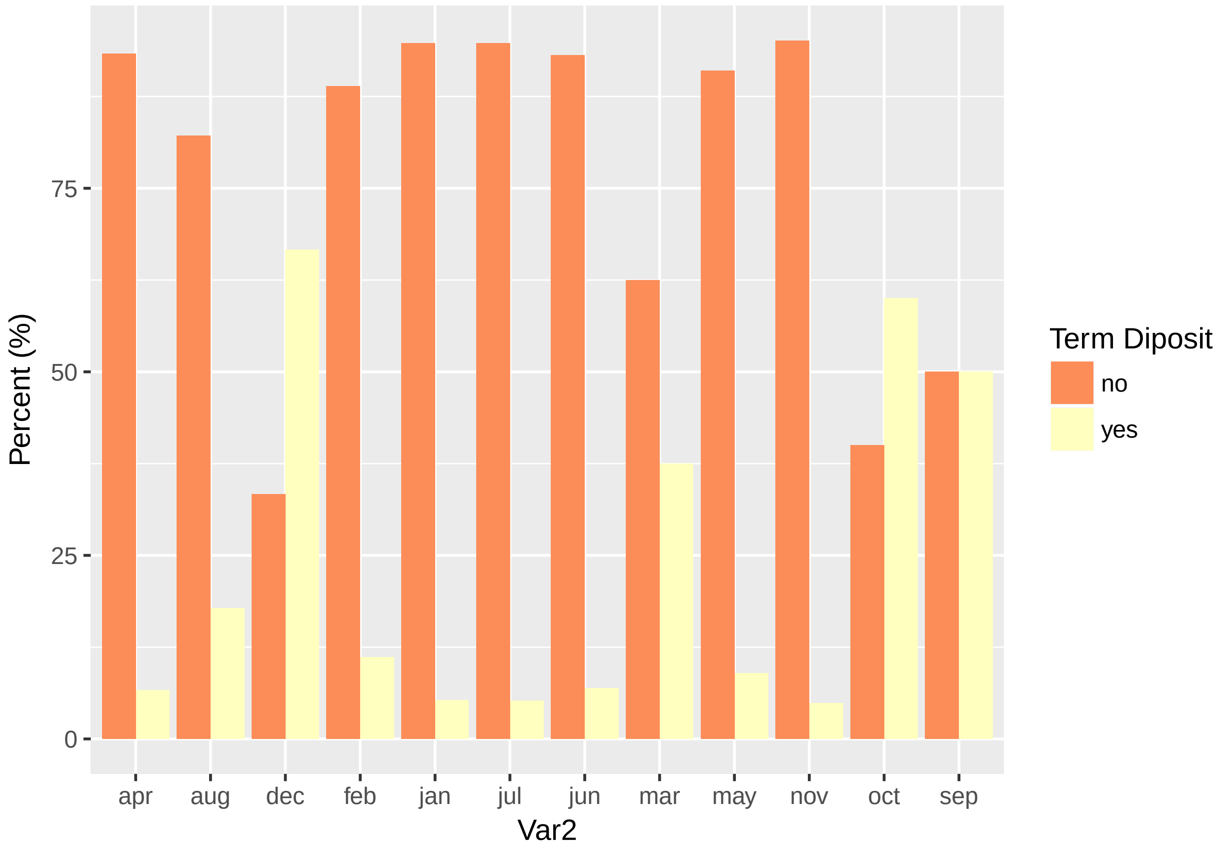
```
bank$campaign2=ifelse(bank$campaign>=10, 10, bank$campaign)
table(bank$campaign2)
```

```
##
##    1    2    3    4    5    6    7    8    9   10
## 1734 1264   558   325   167   155    75    56    30   157
```

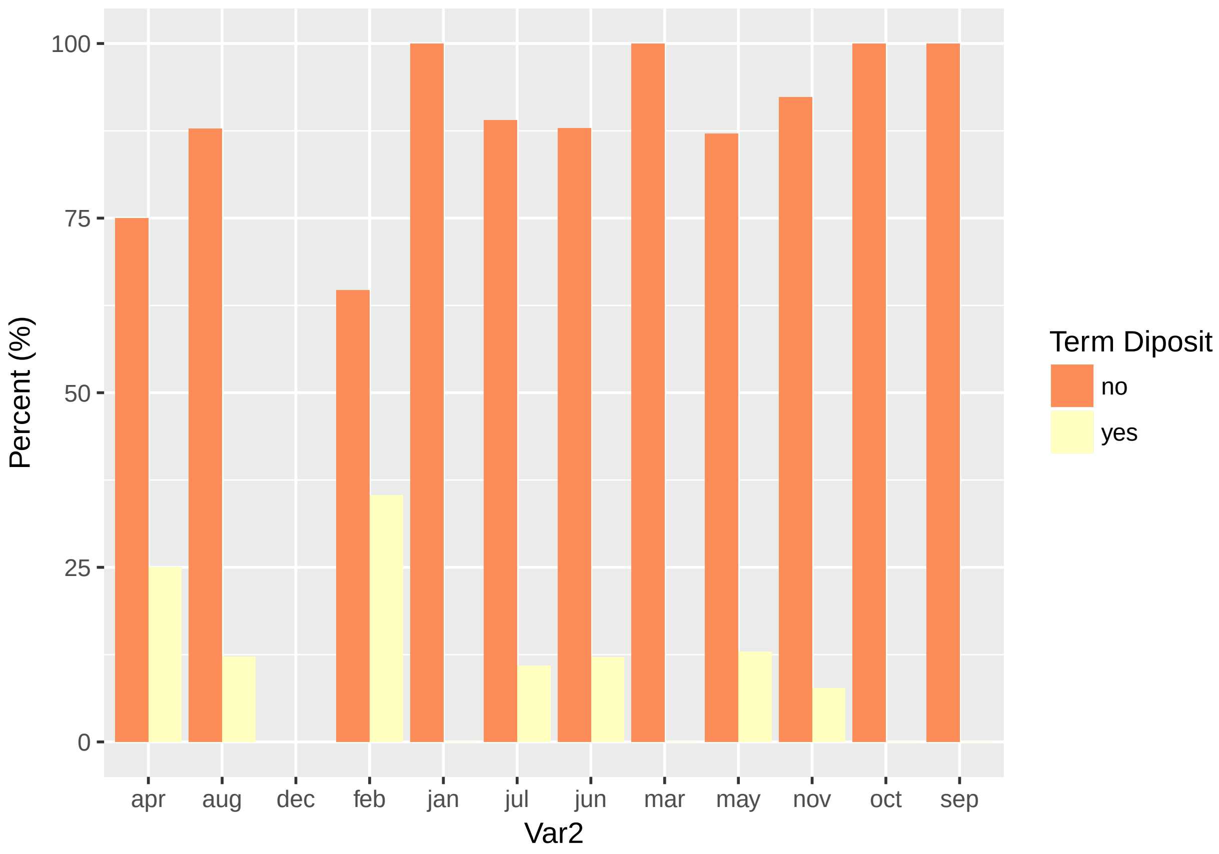
```
CoolPlot<-function(icampaign){

df=subset(bank, bank$campaign2==icampaign)
t3=data.frame(prop.table(table(df$y, df$month),2))

ggplot(t3, aes(x=Var2, y=Freq*100, fill=Var1)) +
  geom_bar(position='dodge', stat='identity')+
  ylab("Percent (%)")+ scale_fill_brewer("Term Diposit", palette="Spectral")
}



par(mfrow=c(5,2))

for (i in 1:10){
  aa<-CoolPlot(i)
  print(aa)
}
```



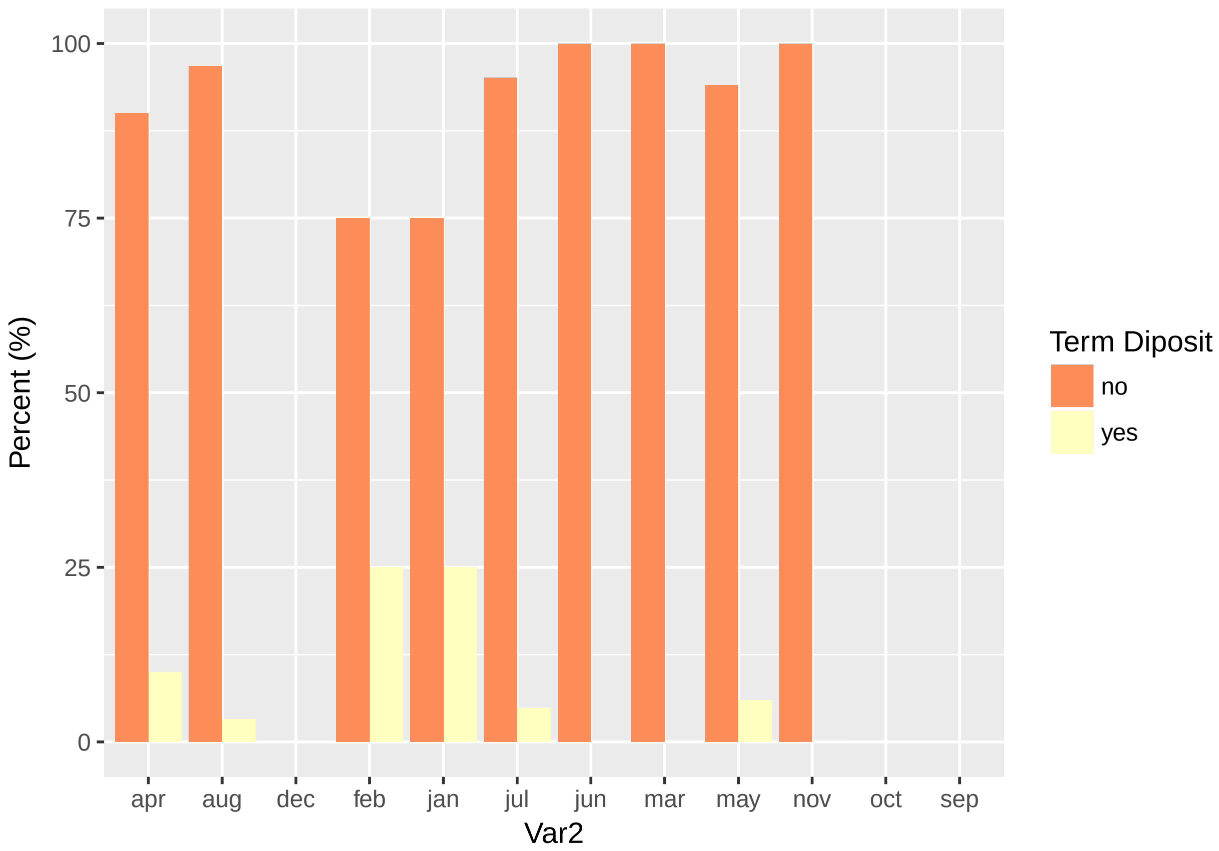

Percent (%)
0
25
50
75
apr
aug
dec
feb
jan
jul
jun
mar
may
nov
oct
sep
Var2
Term Diposit
no
yes



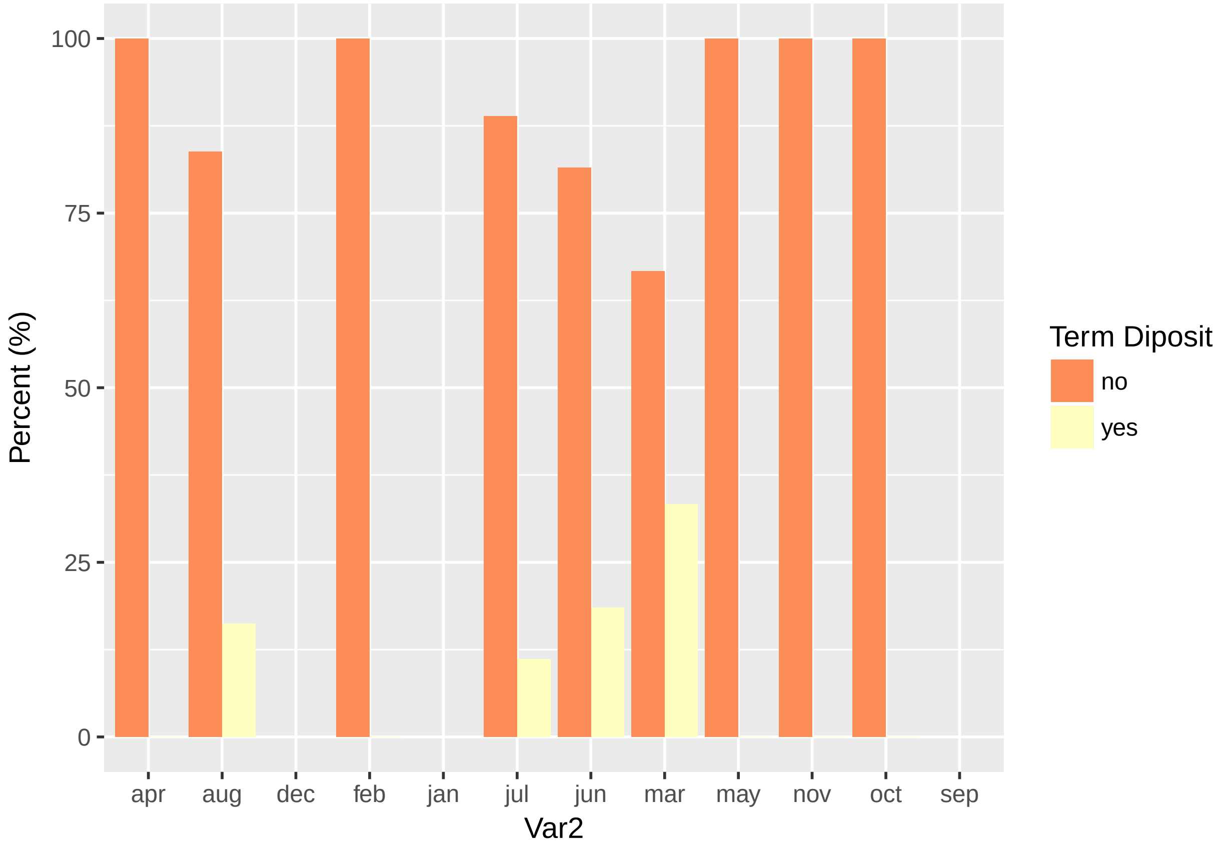
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



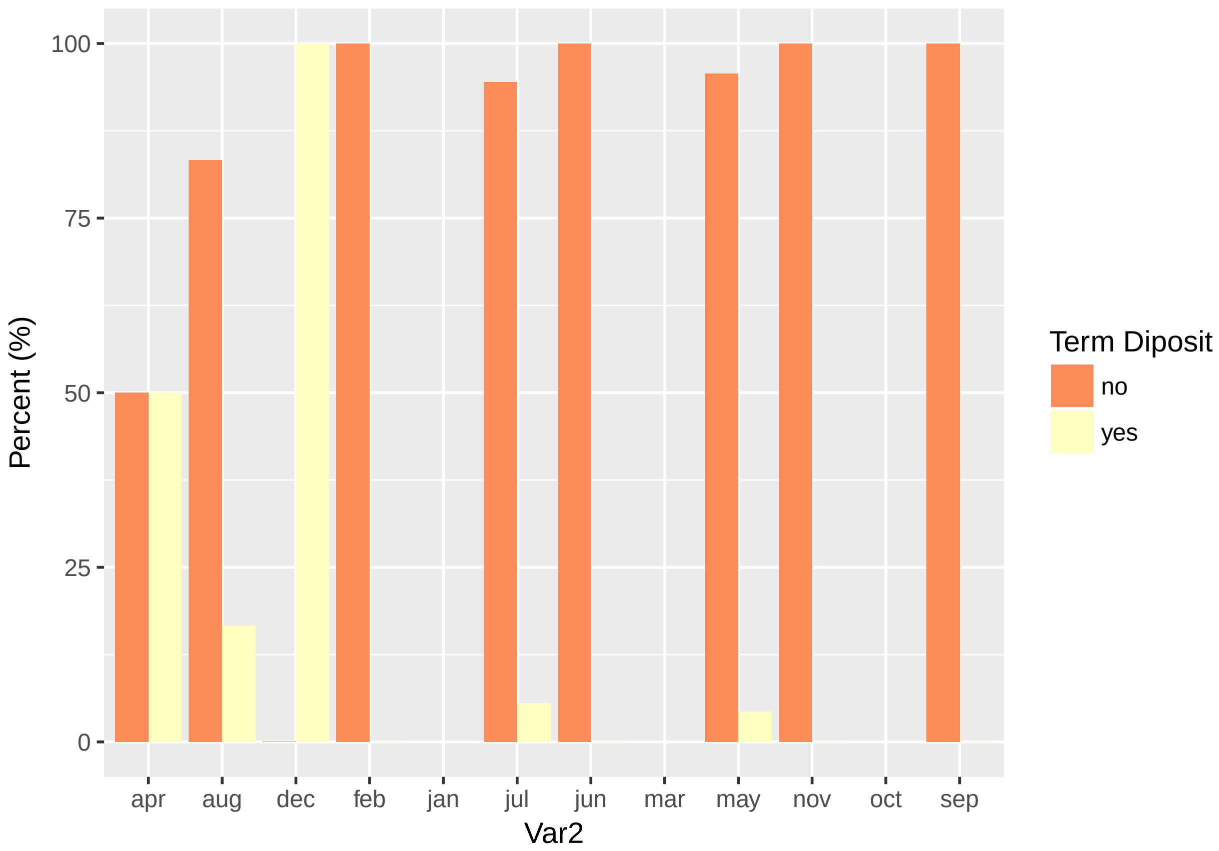
```
## Warning: Removed 6 rows containing missing values (geom_bar).
```



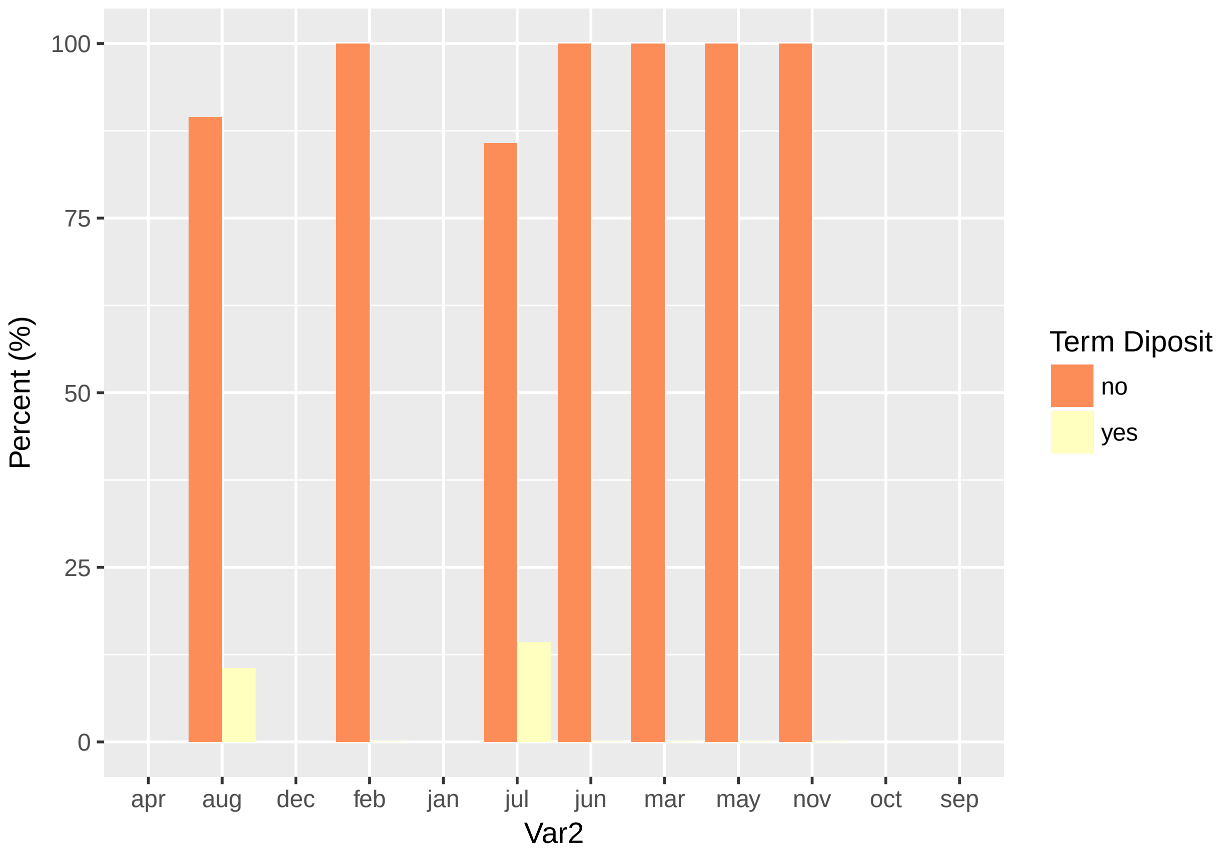
```
## Warning: Removed 6 rows containing missing values (geom_bar).
```



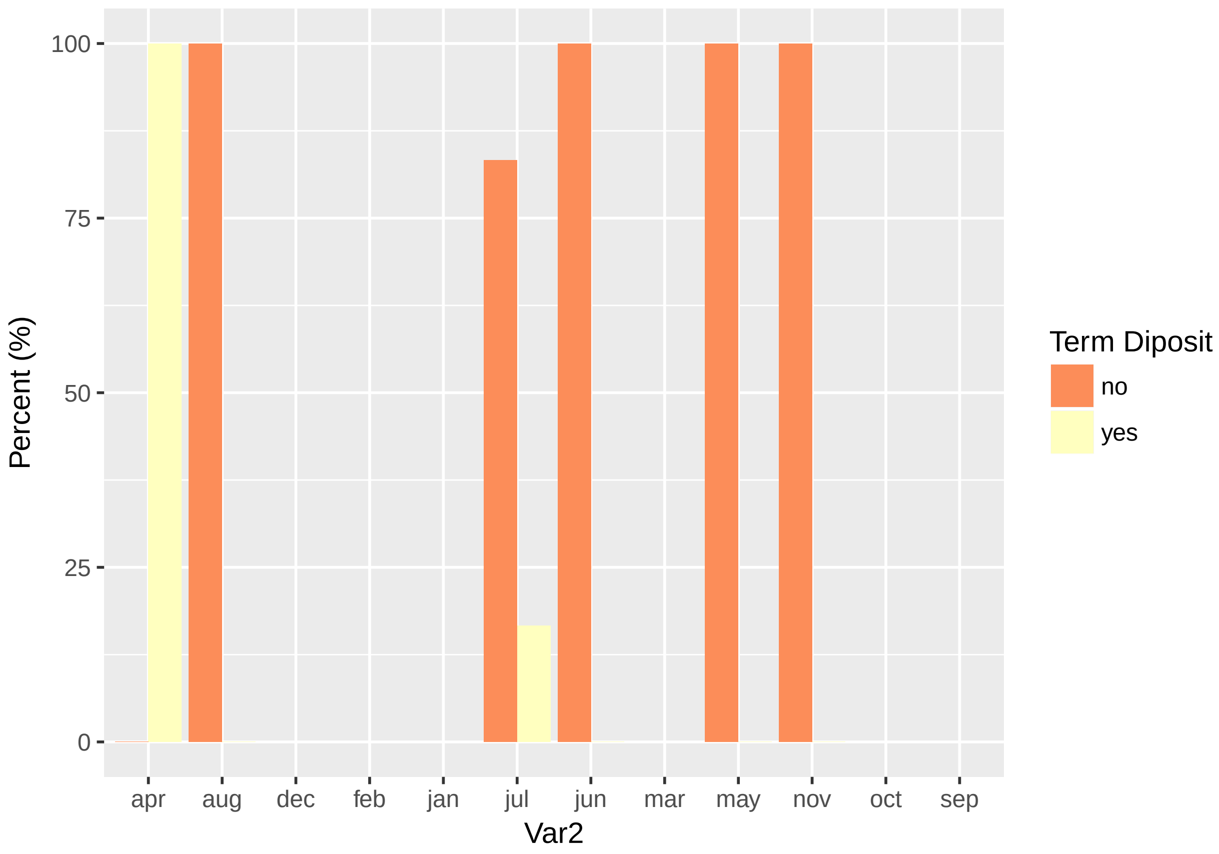
```
## Warning: Removed 6 rows containing missing values (geom_bar).
```



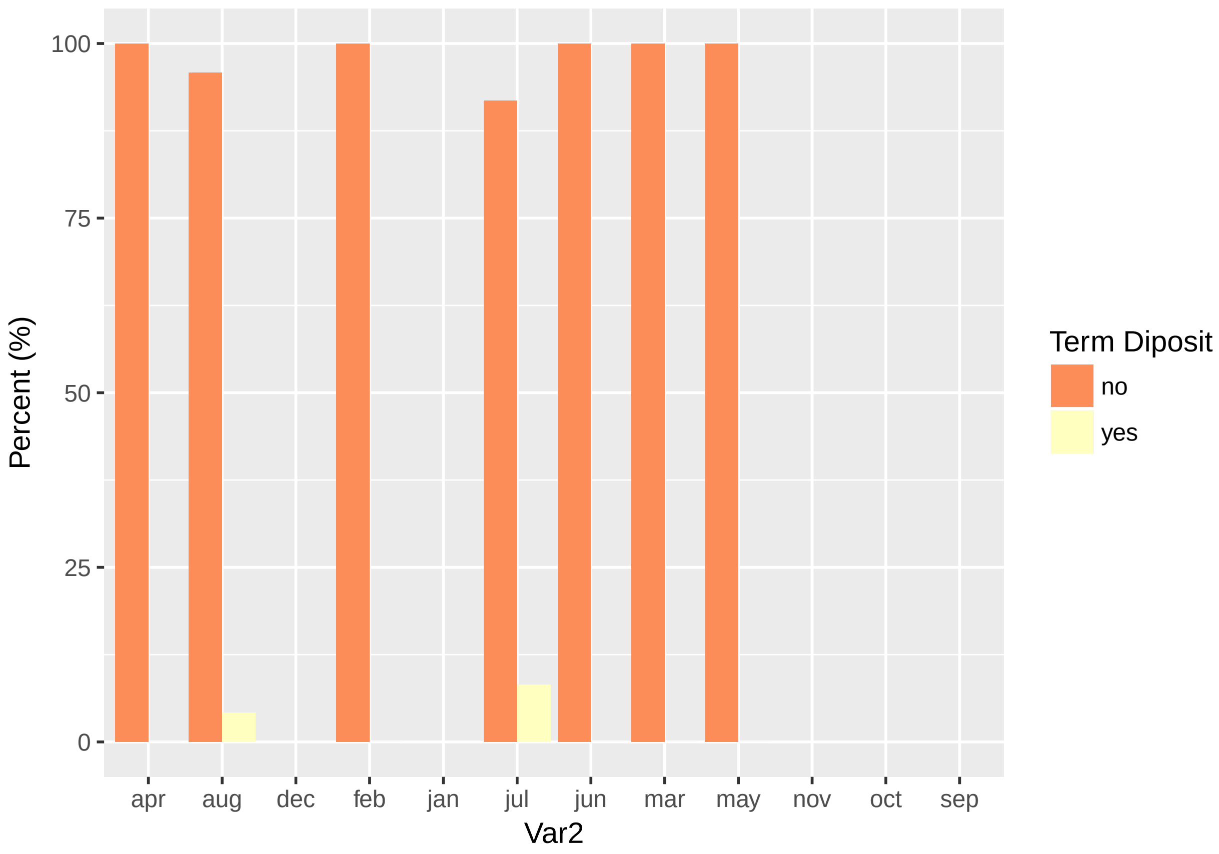
```
## Warning: Removed 10 rows containing missing values (geom_bar).
```



```
## Warning: Removed 12 rows containing missing values (geom_bar).
```



```
## Warning: Removed 10 rows containing missing values (geom_bar).
```



Reference

More information on graphics with R (ggplot2)

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html