Bank Marketting\_bivariate Analysis

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Importing the dataset

bank <- read.csv("bank.csv",header = TRUE, sep = ",")

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   
##

str(bank)

## 'data.frame': 4521 obs. of 17 variables:  
## $ age : int 30 33 35 30 59 35 36 39 41 43 ...  
## $ job : Factor w/ 12 levels "admin.","blue-collar",..: 11 8 5 5 2 5 7 10 3 8 ...  
## $ marital : Factor w/ 3 levels "divorced","married",..: 2 2 3 2 2 3 2 2 2 2 ...  
## $ education: Factor w/ 4 levels "primary","secondary",..: 1 2 3 3 2 3 3 2 3 1 ...  
## $ default : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ balance : int 1787 4789 1350 1476 0 747 307 147 221 -88 ...  
## $ housing : Factor w/ 2 levels "no","yes": 1 2 2 2 2 1 2 2 2 2 ...  
## $ loan : Factor w/ 2 levels "no","yes": 1 2 1 2 1 1 1 1 1 2 ...  
## $ contact : Factor w/ 3 levels "cellular","telephone",..: 1 1 1 3 3 1 1 1 3 1 ...  
## $ day : int 19 11 16 3 5 23 14 6 14 17 ...  
## $ month : Factor w/ 12 levels "apr","aug","dec",..: 11 9 1 7 9 4 9 9 9 1 ...  
## $ duration : int 79 220 185 199 226 141 341 151 57 313 ...  
## $ campaign : int 1 1 1 4 1 2 1 2 2 1 ...  
## $ pdays : int -1 339 330 -1 -1 176 330 -1 -1 147 ...  
## $ previous : int 0 4 1 0 0 3 2 0 0 2 ...  
## $ poutcome : Factor w/ 4 levels "failure","other",..: 4 1 1 4 4 1 2 4 4 1 ...  
## $ y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

Here we have 4521 observations and 17 variables in the data set.

Let us check for NA’s and blank spaces

colSums(is.na(bank))

## age job marital education default balance housing   
## 0 0 0 0 0 0 0   
## loan contact day month duration campaign pdays   
## 0 0 0 0 0 0 0   
## previous poutcome y   
## 0 0 0

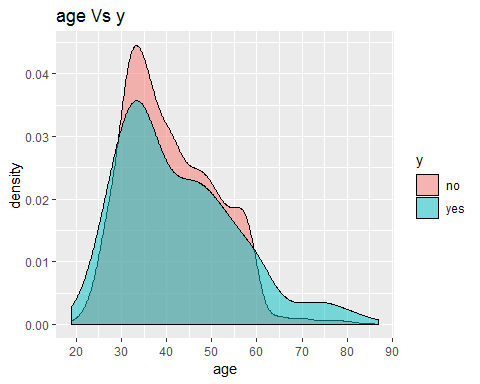
colSums(bank =="")

## age job marital education default balance housing   
## 0 0 0 0 0 0 0   
## loan contact day month duration campaign pdays   
## 0 0 0 0 0 0 0   
## previous poutcome y   
## 0 0 0

There is no missing values in the data set

1. age Vs y Minimum of age is 19 maximum is 87 and mean age is 41.17 Now visualize the data

library(ggplot2)  
ggplot(bank, aes(x = age, fill = y)) +  
geom\_density(alpha=0.5, aes(fill=factor(y))) + labs(title="age Vs y") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

 Most of the customers are from 30 to 40

2.job Vs y

table(bank$job, bank$y)

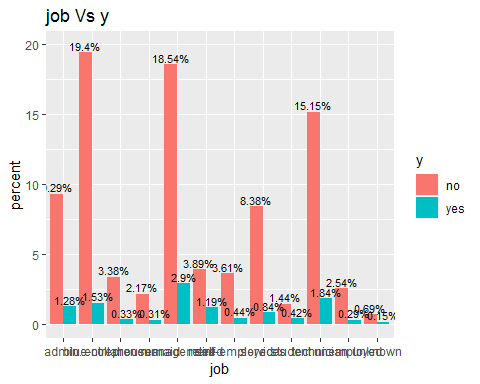
##   
## no yes  
## admin. 420 58  
## blue-collar 877 69  
## entrepreneur 153 15  
## housemaid 98 14  
## management 838 131  
## retired 176 54  
## self-employed 163 20  
## services 379 38  
## student 65 19  
## technician 685 83  
## unemployed 115 13  
## unknown 31 7

prop.table(table(bank$job, bank$y))

##   
## no yes  
## admin. 0.09289980 0.01282902  
## blue-collar 0.19398363 0.01526211  
## entrepreneur 0.03384207 0.00331785  
## housemaid 0.02167662 0.00309666  
## management 0.18535722 0.02897589  
## retired 0.03892944 0.01194426  
## self-employed 0.03605397 0.00442380  
## services 0.08383101 0.00840522  
## student 0.01437735 0.00420261  
## technician 0.15151515 0.01835877  
## unemployed 0.02543685 0.00287547  
## unknown 0.00685689 0.00154833

Now visualize it

ggplot(bank, aes(x = job, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'job', y = 'percent', fill = 'y', title = 'job Vs y')

 Highest number of customers having management job (21.43%) and next is blue-collar job (20.92%)

3.marital Vs y

table(bank$marital, bank$y)

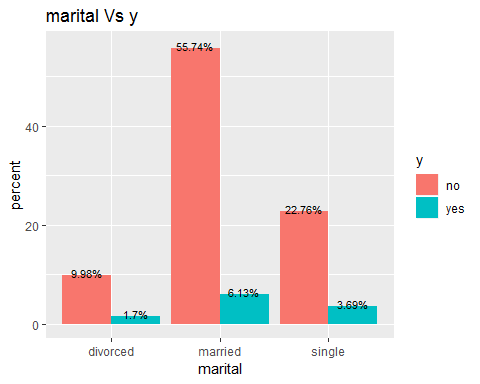
##   
## no yes  
## divorced 451 77  
## married 2520 277  
## single 1029 167

prop.table(table(bank$marital, bank$y))

##   
## no yes  
## divorced 0.09975669 0.01703163  
## married 0.55739881 0.06126963  
## single 0.22760451 0.03693873

Now visualize it

ggplot(bank, aes(x = marital, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'marital', y = 'percent', fill = 'y', title = 'marital Vs y')

 62% customers are married.

4.education Vs y

table(bank$education, bank$y)

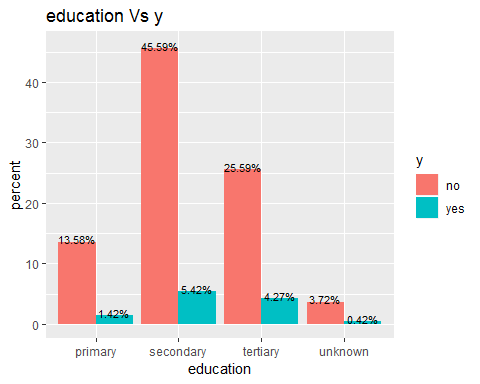
##   
## no yes  
## primary 614 64  
## secondary 2061 245  
## tertiary 1157 193  
## unknown 168 19

prop.table(table(bank$education, bank$y))

##   
## no yes  
## primary 0.13581066 0.01415616  
## secondary 0.45587259 0.05419155  
## tertiary 0.25591683 0.04268967  
## unknown 0.03715992 0.00420261

visualize it

ggplot(bank, aes(x = education, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'education', y = 'percent', fill = 'y', title = 'education Vs y')

 Half of the customers are having secondary education. 5. default Vs y

table(bank$default, bank$y)

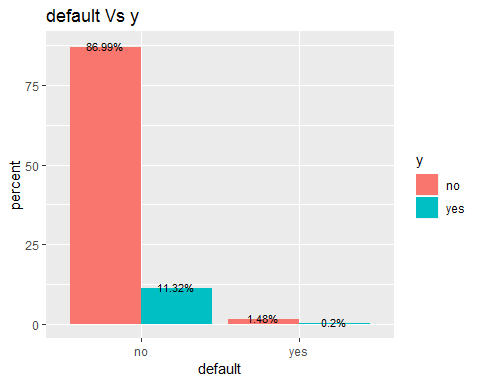
##   
## no yes  
## no 3933 512  
## yes 67 9

prop.table(table(bank$default, bank$y))

##   
## no yes  
## no 0.86994028 0.11324928  
## yes 0.01481973 0.00199071

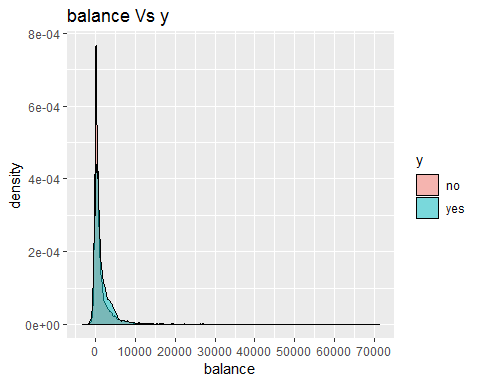
Visualize it

ggplot(bank, aes(x = default, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'default', y = 'percent', fill = 'y', title = 'default Vs y')

 98% customers has no credit in default 6.balance(Average yearly balance in Euros)

minimum balance is -3313, maximum is 71188 and mean baance is 1423. visualize it

ggplot(bank, aes(x = balance, fill = y)) +  
geom\_density(alpha=0.5, aes(fill=factor(y))) + labs(title="balance Vs y") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()



7.housing Vs y

table(bank$housing, bank$y)

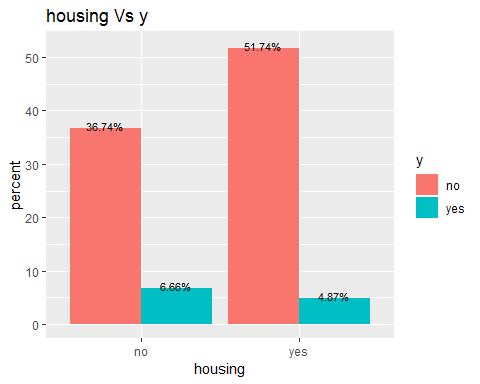
##   
## no yes  
## no 1661 301  
## yes 2339 220

prop.table(table(bank$housing, bank$y))

##   
## no yes  
## no 0.36739659 0.06657819  
## yes 0.51736342 0.04866180

visualize it

ggplot(bank, aes(x = housing, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'housing', y = 'percent', fill = 'y', title = 'housing Vs y')

 56.6% customers has housing loan.

8.loan Vs y

table(bank$loan, bank$y)

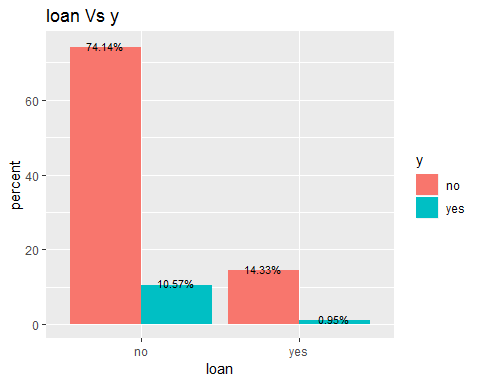
##   
## no yes  
## no 3352 478  
## yes 648 43

prop.table(table(bank$loan, bank$y))

##   
## no yes  
## no 0.74142889 0.10572882  
## yes 0.14333112 0.00951117

visualize it

ggplot(bank, aes(x = loan, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'loan', y = 'percent', fill = 'y', title = 'loan Vs y')

 84.7% has personal loan.

9.contact Vs y

table(bank$contact, bank$y)

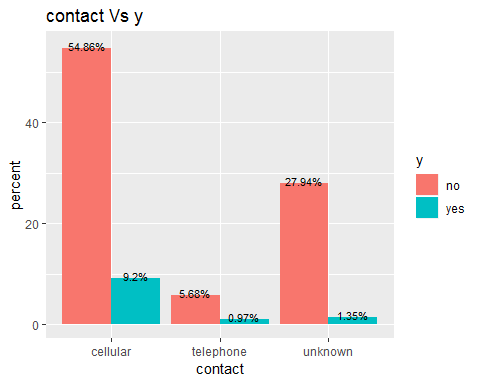
##   
## no yes  
## cellular 2480 416  
## telephone 257 44  
## unknown 1263 61

prop.table(table(bank$contact, bank$y))

##   
## no yes  
## cellular 0.54855121 0.09201504  
## telephone 0.05684583 0.00973236  
## unknown 0.27936297 0.01349259

visualize it

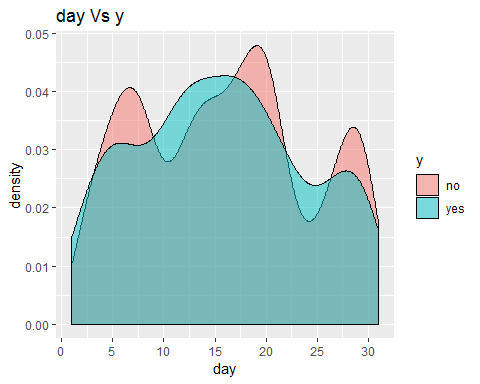
ggplot(bank, aes(x = contact, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'contact', y = 'percent', fill = 'y', title = 'contact Vs y')

 64% customers are contacted by cellular.

10.day(Day of the month last time customer was contacted)

visualize it

ggplot(bank, aes(x = day, fill = y)) +  
geom\_density(alpha=0.5, aes(fill=factor(y))) + labs(title="day Vs y") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

 Most customers contacted on 15 to 20

1. month(month of the year last time customer was contacted)

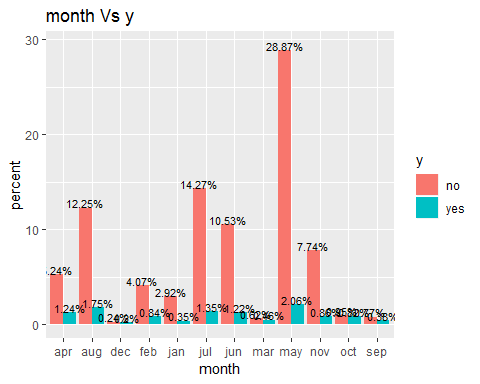
table(bank$month, bank$y)

##   
## no yes  
## apr 237 56  
## aug 554 79  
## dec 11 9  
## feb 184 38  
## jan 132 16  
## jul 645 61  
## jun 476 55  
## mar 28 21  
## may 1305 93  
## nov 350 39  
## oct 43 37  
## sep 35 17

prop.table(table(bank$month, bank$y))

##   
## no yes  
## apr 0.05242203 0.01238664  
## aug 0.12253926 0.01747401  
## dec 0.00243309 0.00199071  
## feb 0.04069896 0.00840522  
## jan 0.02919708 0.00353904  
## jul 0.14266755 0.01349259  
## jun 0.10528644 0.01216545  
## mar 0.00619332 0.00464499  
## may 0.28865295 0.02057067  
## nov 0.07741650 0.00862641  
## oct 0.00951117 0.00818403  
## sep 0.00774165 0.00376023

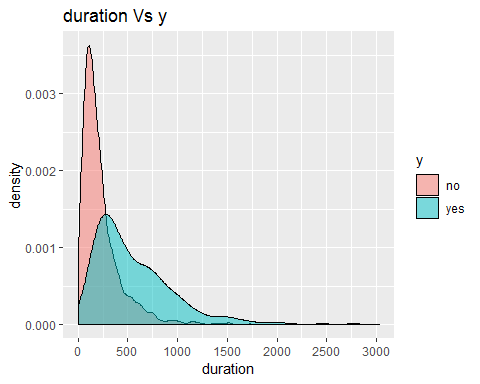
ggplot(bank, aes(x = month, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'month', y = 'percent', fill = 'y', title = 'month Vs y')

 May has more customers contacted

12.duration(last contact duration in seconds)

visualize it

ggplot(bank, aes(x = duration, fill = y)) +  
geom\_density(alpha=0.5, aes(fill=factor(y))) + labs(title="duration Vs y") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

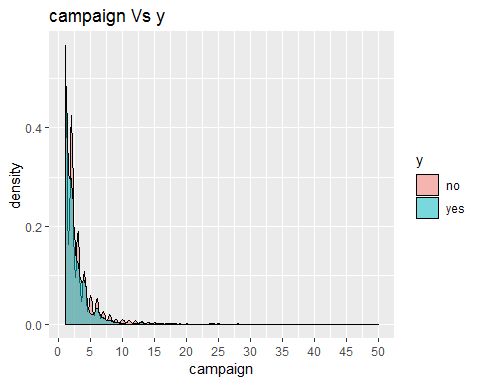


Mean duration for last contact is 264 seconds

13.campaign(Number of contacts performed during the marketing campaign and for this customer)

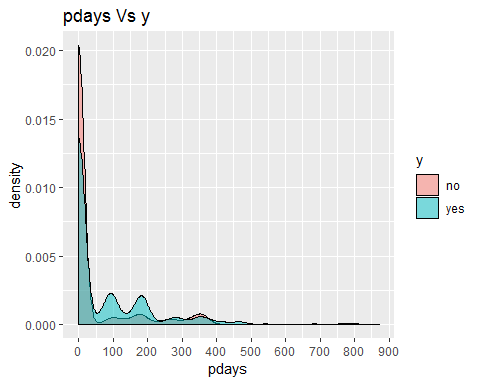
Visualize

ggplot(bank, aes(x = campaign, fill = y)) +  
geom\_density(alpha=0.5, aes(fill=factor(y))) + labs(title="campaign Vs y") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

 Mean number of contacts performed a cuctomer is 2 times

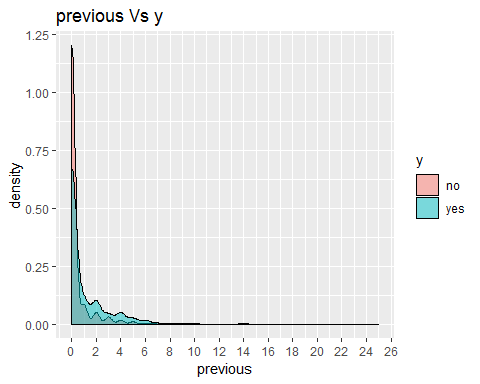
14.pdays(Number of days that passed by after the client was last contacted from a previous campaign) Visualize

ggplot(bank, aes(x = pdays, fill = y)) +  
geom\_density(alpha=0.5, aes(fill=factor(y))) + labs(title="pdays Vs y") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

 Mean number of days that passed by after the client has contacted from previous campaign is 39.

15.previous(Number of contacts performed before this campaign and for this client) Visualize

ggplot(bank, aes(x = previous, fill = y)) +  
geom\_density(alpha=0.5, aes(fill=factor(y))) + labs(title="previous Vs y") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

 Maximum number of contacts performed before this campaign is 25

16.poutcome(Outcome of the previous marketing campaign)

table(bank$poutcome, bank$y)

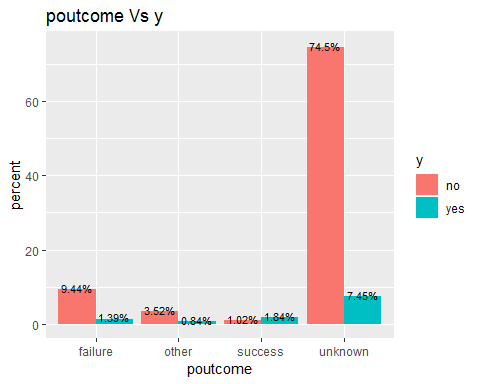
##   
## no yes  
## failure 427 63  
## other 159 38  
## success 46 83  
## unknown 3368 337

prop.table(table(bank$poutcome, bank$y))

##   
## no yes  
## failure 0.09444813 0.01393497  
## other 0.03516921 0.00840522  
## success 0.01017474 0.01835877  
## unknown 0.74496793 0.07454103

Visualize

ggplot(bank, aes(x = poutcome, fill = y)) +  
 geom\_bar(aes(y = prop.table(..count..) \* 100), position='dodge') +  
 geom\_text(aes(y = prop.table(..count..) \* 100 + 0.5,   
 label = paste0(round(prop.table(..count..) \* 100, 2), '%')),   
 stat = 'count',   
 position = position\_dodge(.9),   
 size = 3) +   
 labs(x = 'poutcome', y = 'percent', fill = 'y', title = 'poutcome Vs y')

 Here 10% are failured, 2% success and 81% unknown.