Banking_2.R

Ram

Wed Nov 09 20:51:19 2016

```
#               Visual Analysis using ggplot2                      |
#        Exploratory analysis of Bank customers data (Stay/Exit)   |
#  (keep in mind that ususally these kind of problems comes with   |
#    imbalanced data ==> while modeling, care has to be taken )    |
```

```
setwd('G:/DATASCIENCE/DS-PROJECTS/15_Visual_Analytics/Banking/')
rm(list=ls())
```

```
library('dplyr')
```

```
## Warning: package 'dplyr' was built under R version 3.2.5
```

```
##
## 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('ggplot2')
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
library('gridExtra')
library('grid')
library('scales')
library('corrplot')
```

```
## Warning: package 'corrplot' was built under R version 3.2.5
```

```
library('mlr')
```

```
## Warning: package 'mlr' was built under R version 3.2.5
```

```
## Loading required package: BBmisc
```

```
## Warning: package 'BBmisc' was built under R version 3.2.5
```

```
##
## Attaching package: 'BBmisc'

## The following object is masked from 'package:grid':
##
##     explode

## The following objects are masked from 'package:dplyr':
##
##     coalesce, collapse

## Loading required package: ParamHelpers

## Warning: package 'ParamHelpers' was built under R version 3.2.5

## Loading required package: stringi
```

```
data <- read.csv('Churn-Modelling.csv', na.strings = c('',' ', '?', 'NA'), stringsAsFactors = T)
# As I am not intended to do predictive modeling here, read only the input data

str(data)
```

```
## 'data.frame':    10000 obs. of  14 variables:
##  $ RowNumber      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CustomerId     : int  15634602 15647311 15619304 15701354 15737888 15574012 15592531 15656148 1579
##  $ Surname        : Factor w/ 2932 levels "Abazu","Abbie",..: 1116 1178 2041 290 1823 538 178 2001 1
##  $ CreditScore    : int  619 608 502 699 850 645 822 376 501 684 ...
##  $ Geography      : Factor w/ 3 levels "France","Germany",..: 1 3 1 1 3 3 1 2 1 1 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 2 2 1 2 2 ...
##  $ Age            : int  42 41 42 39 43 44 50 29 44 27 ...
##  $ Tenure         : int  2 1 8 1 2 8 7 4 4 2 ...
##  $ Balance        : num  0 83808 159661 0 125511 ...
##  $ NumOfProducts  : int  1 1 3 2 1 2 2 4 2 1 ...
##  $ HasCrCard      : int  1 0 1 0 1 1 1 1 0 1 ...
##  $ IsActiveMember : int  1 1 0 0 1 0 1 0 1 1 ...
##  $ EstimatedSalary: num  101349 112543 113932 93827 79084 ...
##  $ Exited         : int  1 0 1 0 0 1 0 1 0 0 ...
```

```
# Target : Exited

data$HasCrCard      <- ifelse(data$HasCrCard == 0, 'NO',"YES")
data$IsActiveMember <- ifelse(data$IsActiveMember == 0, 'NO','YES')
data$Exited         <- ifelse(data$Exited == 0, 'Stayed', 'Exited')

data$Tenure <- as.factor(data$Tenure)
data$NumOfProducts <- as.factor(data$NumOfProducts)
data$HasCrCard <- as.factor(data$HasCrCard)
data$IsActiveMember <- as.factor(data$IsActiveMember)
data$Exited <- as.factor(data$Exited)
```

```r
# Data Summarizations and Tabulations
```

```r
# Though I prefer 'dplyr' , base R's prop.table() is providing the
# cleaner output --> I have sticked to prop.table() in the analysis
```

```r
# data %>% group_by(Geography) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
# data %>% group_by(Gender) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
# data %>% group_by(Geography,Gender) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
# data %>% group_by(Exited) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
# data %>% group_by(Geography,Exited) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
# data %>% group_by(HasCrCard,Exited) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
# data %>% group_by(Gender,Exited) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
# data %>% group_by(Geography,Gender,Exited) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
```

```r
prop.table(table(data$Geography))
```

```
##
##  France Germany   Spain
##  0.5014  0.2509  0.2477
```

```r
prop.table(table(data$Gender))
```

```
##
## Female   Male
## 0.4543 0.5457
```

```r
prop.table(table(data$Tenure))
```

```
##
##      0      1      2      3      4      5      6      7      8      9
## 0.0413 0.1035 0.1048 0.1009 0.0989 0.1012 0.0967 0.1028 0.1025 0.0984
##     10
## 0.0490
```

```r
prop.table(table(data$NumOfProducts))
```

```
##
##      1      2      3      4
## 0.5084 0.4590 0.0266 0.0060
```

```r
prop.table(table(data$HasCrCard))
```

```
##
##     NO    YES
## 0.2945 0.7055
```

```r
prop.table(table(data$IsActiveMember))
```

```
## 
##     NO    YES 
## 0.4849 0.5151
```

```r
prop.table(table(data$Exited))
```

```
## 
## Exited Stayed 
## 0.2037 0.7963
```

```r
prop.table(table(data$Geography,data$Exited),1)
```

```
##          
##              Exited    Stayed
##   France  0.1615477 0.8384523
##   Germany 0.3244320 0.6755680
##   Spain   0.1667340 0.8332660
```

```r
prop.table(table(data$Gender,data$Exited),1)
```

```
##         
##             Exited    Stayed
##   Female 0.2507154 0.7492846
##   Male   0.1645593 0.8354407
```

```r
prop.table(table(data$Tenure,data$Exited),1)
```

```
##     
##         Exited    Stayed
##   0  0.2300242 0.7699758
##   1  0.2241546 0.7758454
##   2  0.1917939 0.8082061
##   3  0.2111001 0.7888999
##   4  0.2052578 0.7947422
##   5  0.2065217 0.7934783
##   6  0.2026887 0.7973113
##   7  0.1721790 0.8278210
##   8  0.1921951 0.8078049
##   9  0.2164634 0.7835366
##   10 0.2061224 0.7938776
```

```r
prop.table(table(data$NumOfProducts,data$Exited),1)
```

```
##    
##         Exited     Stayed
##   1 0.27714398 0.72285602
##   2 0.07581699 0.92418301
##   3 0.82706767 0.17293233
##   4 1.00000000 0.00000000
```

```r
prop.table(table(data$HasCrCard,data$Exited),1)
```

```
##
##          Exited    Stayed
##   NO  0.2081494 0.7918506
##   YES 0.2018427 0.7981573
```

```r
prop.table(table(data$IsActiveMember,data$Exited),1)
```

```
##
##          Exited    Stayed
##   NO  0.2685090 0.7314910
##   YES 0.1426907 0.8573093
```

```r
summary(data)
```

```
##    RowNumber       CustomerId           Surname      CreditScore
##  Min.   :    1   Min.   :15565701   Smith   :  32   Min.   :350.0
##  1st Qu.: 2501   1st Qu.:15628528   Martin  :  29   1st Qu.:584.0
##  Median : 5000   Median :15690738   Scott   :  29   Median :652.0
##  Mean   : 5000   Mean   :15690941   Walker  :  28   Mean   :650.5
##  3rd Qu.: 7500   3rd Qu.:15753234   Brown   :  26   3rd Qu.:718.0
##  Max.   :10000   Max.   :15815690   Genovese:  25   Max.   :850.0
##                                     (Other) :9831
##    Geography       Gender          Age            Tenure
##  France :5014   Female:4543   Min.   :18.00   2      :1048
##  Germany:2509   Male  :5457   1st Qu.:32.00   1      :1035
##  Spain  :2477                 Median :37.00   7      :1028
##                               Mean   :38.92   8      :1025
##                               3rd Qu.:44.00   5      :1012
##                               Max.   :92.00   3      :1009
##                                               (Other):3843
##     Balance       NumOfProducts HasCrCard  IsActiveMember
##  Min.   :     0   1:5084        NO :2945   NO :4849
##  1st Qu.:     0   2:4590        YES:7055   YES:5151
##  Median : 97199   3: 266
##  Mean   : 76486   4:  60
##  3rd Qu.:127644
##  Max.   :250898
##
##  EstimatedSalary        Exited
##  Min.   :    11.58   Exited:2037
##  1st Qu.: 51002.11   Stayed:7963
##  Median :100193.91
##  Mean   :100090.24
##  3rd Qu.:149388.25
##  Max.   :199992.48
##
```

```r
cred.sco_hist <- ggplot(data,aes(CreditScore)) + geom_histogram(binwidth=1.0,fill=alpha('blue',0.4)) +
  scale_x_continuous(limits=c(300,820)) + labs(title='CreditScore_histogram')
```

```r
a2 <- ggplot(data,aes(CreditScore)) + geom_histogram(binwidth=0.1,fill=alpha('blue',0.4)) +
  scale_x_sqrt() # this is also better
a3 <- ggplot(data,aes(CreditScore)) + geom_histogram(binwidth=0.002) + scale_x_log10()
```

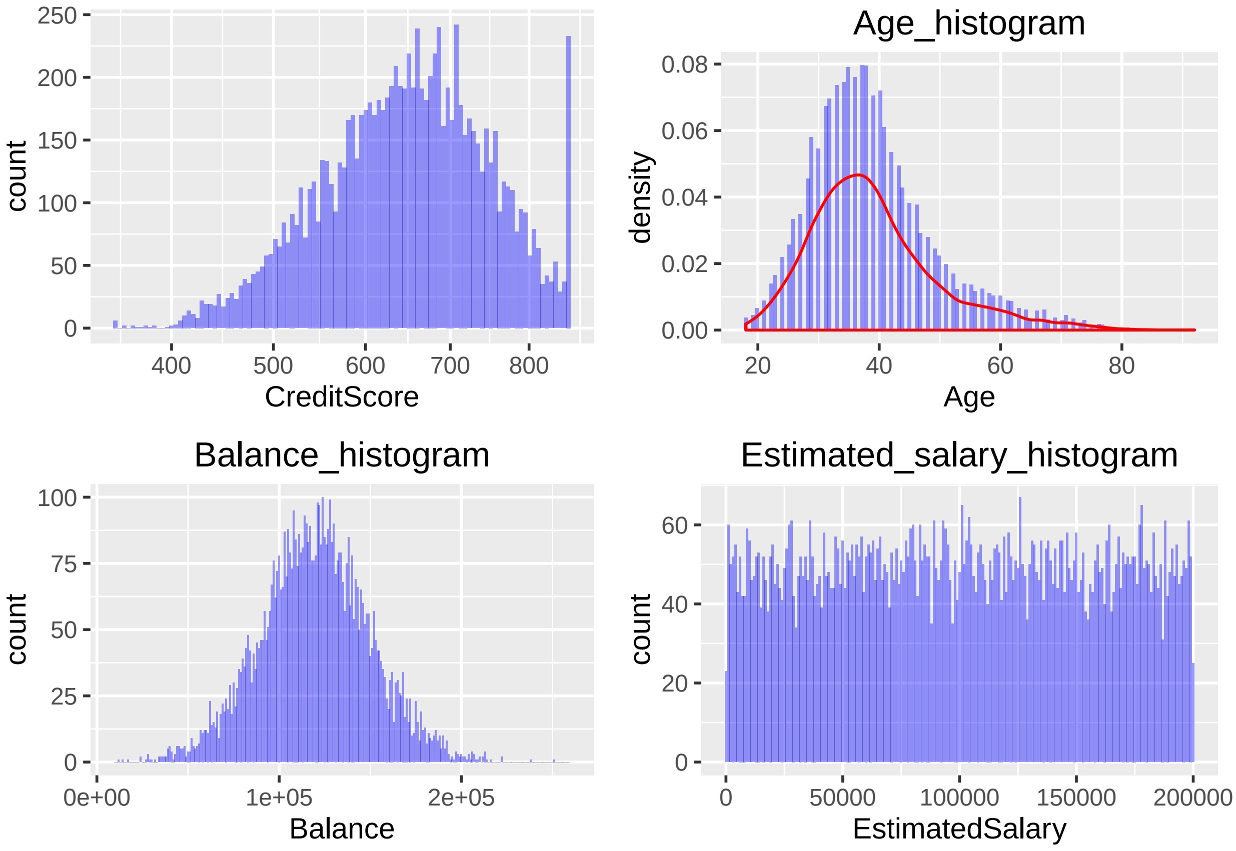
```r
age_hist <- ggplot(data,aes(Age, ..density..)) + geom_histogram(binwidth=0.6,fill=alpha('blue',0.4)) +
  geom_density(color='red') + labs(title='Age_histogram')

#ggplot(data,aes(Age)) + geom_freqpoly(binwidth=1.0)
```

```r
# ggplot(data,aes(Balance)) + geom_histogram() + scale_x_sqrt()
b <- ggplot(data,aes(Balance)) + geom_histogram(binwidth = 0.01) + scale_x_log10()
# Though log10 helped, instead cap the small values
balancee_hist <- ggplot(data,aes(Balance)) + geom_histogram(binwidth = 1000,fill=alpha('blue',0.4)) +
  scale_x_continuous(limits=c(9000,260000)) + labs(title='Balance_histogram')
```

```r
Est_sal <- ggplot(data,aes(EstimatedSalary)) + geom_histogram(binwidth = 1000, fill=alpha('blue',0.4))
  labs(title='Estimated_salary_histogram')
```

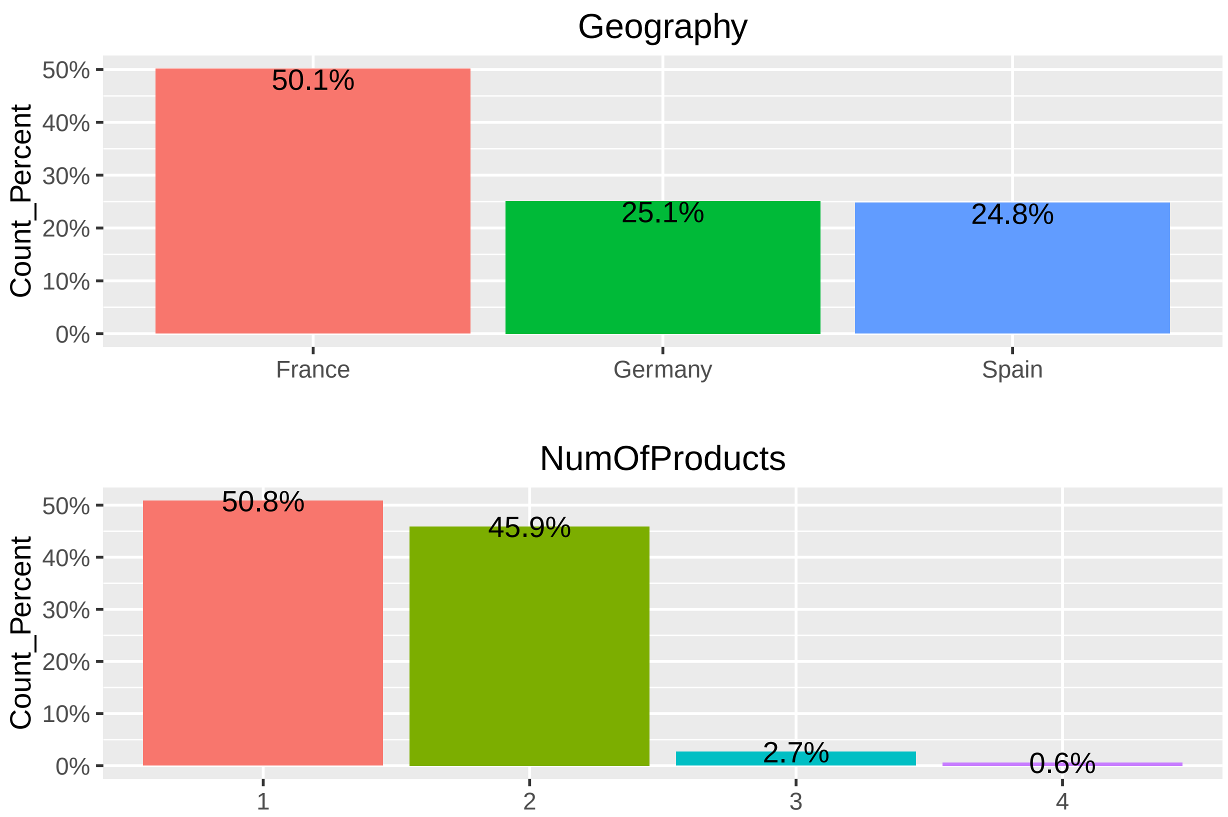
```r
grid.arrange(a2,age_hist,balancee_hist,Est_sal,nrow=2)
```

```
## Warning: Removed 3618 rows containing non-finite values (stat_bin).
```

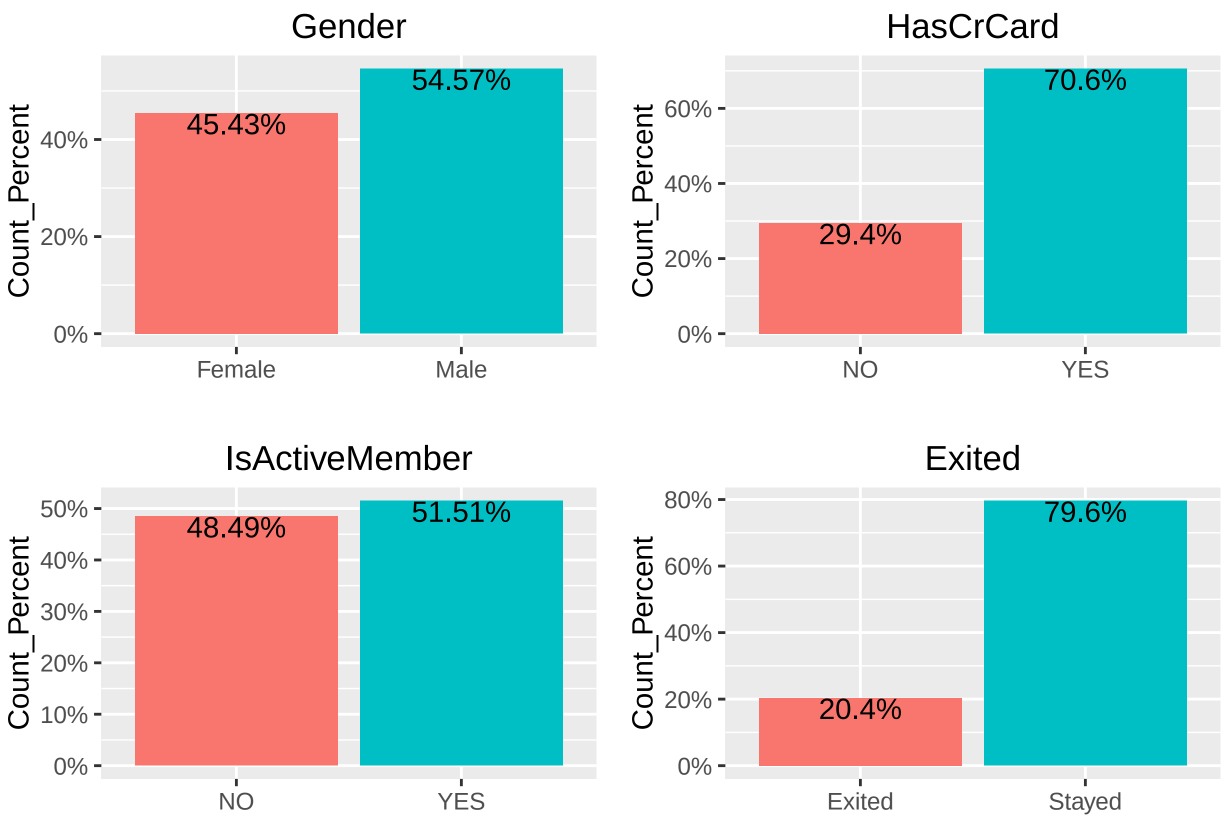


```
# -- 'Credit Score'     : did applied a 'square root' transformation
# -- 'Age'              : is somewhat rigth skewed
# -- 'Balance'          : applied some scale restictions
# -- 'Estimated Salary' : (not well distributed)
```

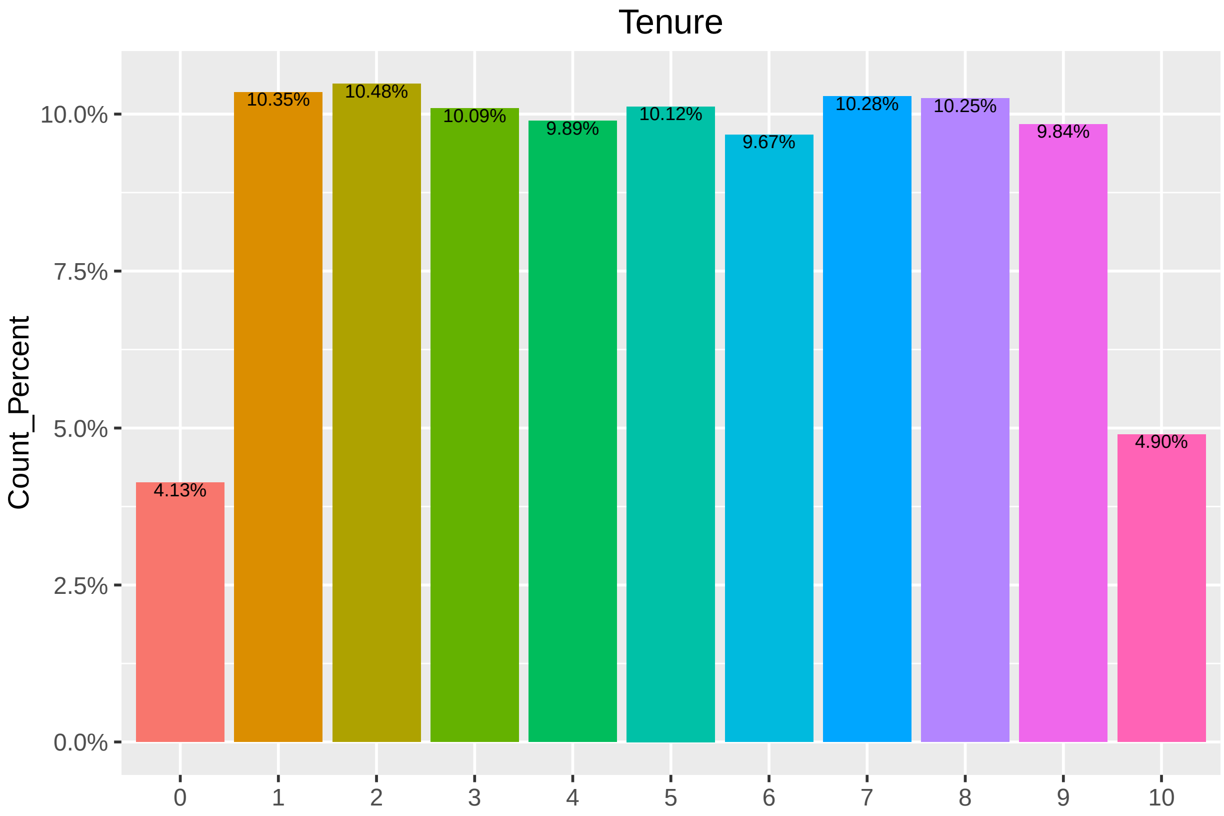
```
# Geograpy
Geography <- ggplot(data, aes(Geography,fill=Geography)) + geom_bar(aes(y=(..count../sum(..count..)))) +
  geom_text(aes(y=(..count../sum(..count..)),label=scales::percent((..count..)/sum(..count..))),
  stat='count', vjust=1.0) + scale_y_continuous(labels=percent_format()) +
  labs(title='Geography',y='Count_Percent', x='') +  theme(legend.position='none')
# Gender
Gender <- ggplot(data, aes(Gender,fill=Gender)) + geom_bar(aes(y=(..count../sum(..count..)))) +
  geom_text(aes(y=(..count../sum(..count..)),label=scales::percent((..count..)/sum(..count..))),
            stat='count', vjust=1.0) + scale_y_continuous(labels=percent_format()) +
  labs(title='Gender',y='Count_Percent',x='') +  theme(legend.position='none')
# Tenure
Tenure <- ggplot(data, aes(Tenure,fill=Tenure)) + geom_bar(aes(y=(..count../sum(..count..)))) +
  geom_text(size=2.5,aes(y=(..count../sum(..count..)),label=scales::percent((..count..)/sum(..count..))
            stat='count', vjust=1.0) + scale_y_continuous(labels=percent_format()) +
  labs(title='Tenure',y='Count_Percent',x='') +  theme(legend.position='none')
# NumOfProducts
No.Prods <- ggplot(data, aes(NumOfProducts,fill=NumOfProducts)) + geom_bar(aes(y=(..count../sum(..count
  geom_text(aes(y=(..count../sum(..count..)),label=scales::percent((..count..)/sum(..count..))),
            stat='count', vjust=0.5) + scale_y_continuous(labels=percent_format()) +
  labs(title='NumOfProducts',y='Count_Percent',x='') +  theme(legend.position='none')
# HasCrCard
Crcard <- ggplot(data, aes(HasCrCard,fill=HasCrCard)) + geom_bar(aes(y=(..count../sum(..count..)))) +
  geom_text(aes(y=(..count../sum(..count..)),label=scales::percent((..count..)/sum(..count..))),
            stat='count', vjust=1.0) + scale_y_continuous(labels=percent_format()) +
  labs(title='HasCrCard',y='Count_Percent',x='') +  theme(legend.position='none')
# IsActiveMember
Active <- ggplot(data, aes(IsActiveMember,fill=IsActiveMember)) + geom_bar(aes(y=(..count../sum(..count
  geom_text(aes(y=(..count../sum(..count..)),label=scales::percent((..count..)/sum(..count..))),
            stat='count', vjust=1.0) + scale_y_continuous(labels=percent_format()) +
  labs(title='IsActiveMember',y='Count_Percent',x='') +  theme(legend.position='none')
# Exited
Ex_stay <- ggplot(data, aes(Exited,fill=Exited)) + geom_bar(aes(y=(..count../sum(..count..)))) +
  geom_text(aes(y=(..count../sum(..count..)),label=scales::percent((..count..)/sum(..count..))),
            stat='count', vjust=1.0) + scale_y_continuous(labels=percent_format()) +
  labs(title='Exited',y='Count_Percent',x='') +  theme(legend.position='none')

# Tenure is not included -- probably needs to regroup them

grid.arrange(Geography,No.Prods,nrow=2)
```



```r
grid.arrange(Gender,Crcard,Active,Ex_stay, nrow=2)
```



```
grid.arrange(Tenure,nrow=1)
```



```
# 'Geography'   -- Overall, 50% of the bank custormer's are from France
#                          approx. 25% each are from Germany and Spain
# 'NumOfProds'  -- approx. 51% of customer's are having 1 product and
#                         46% are having 2 products with the bank
#             (Generally, the more the products, they are less likely to exit)
# 'Gender'      -- The male customers are dominant in the bank (55%)
# 'HasCrCard'   -- nearly 71% of the customers have credit cards
#             (Generally, customers with credit cards are less likely to exit)
# 'IsActive'    -- they are almost equal in proportions (approx. 50%)
# 'Exited'      -- 20% of the customers are exited
```

```
# Role of Geograpy on Exit
prop.table(table(data$Geography,data$Exited),1)
```

```
##
##               Exited    Stayed
##    France  0.1615477 0.8384523
##    Germany 0.3244320 0.6755680
##    Spain   0.1667340 0.8332660
```

```
geo_ext <- ggplot(data, aes(x=Geography,fill=Exited)) + geom_bar(position='fill') +
  scale_y_continuous(labels=percent_format()) +
  geom_hline(yintercept=0.2) +
  annotate('text',x=1,y=c(0.10,0.60),label=c('16%','84%'),size=5) +
  annotate('text',x=2,y=c(0.10,0.60),label=c('32%','68%'),size=5) +
  annotate('text',x=3,y=c(0.10,0.60),label=c('17%','83%'),size=5) +
  scale_fill_manual(values=alpha(c('red','blue'),.5)) +
  labs(title = "Geography", y = "% of Total Number of Records", x='') +
  theme_classic()

# Exit rate is quite high in Germany (32%, infact, its double compared to
# France and Spain)
```

```
# Role of Gender on Exit
prop.table(table(data$Gender,data$Exited),1)
```

```
##
##             Exited    Stayed
##   Female 0.2507154 0.7492846
##   Male   0.1645593 0.8354407
```

```
gender_ext <- ggplot(data, aes(x=Gender,fill=Exited)) + geom_bar(position='fill') +
  scale_y_continuous(labels=percent_format()) +
  geom_hline(yintercept=0.2) +
  annotate('text',x=1,y=c(0.13,0.60),label=c('25%','75%'),size=5) +
  annotate('text',x=2,y=c(0.13,0.60),label=c('16%','84%'),size=5) +
  scale_fill_manual(values=alpha(c('red','blue'),.5)) +
  labs(title = "Gender", y = "% of Total Number of Records", x='') +
  theme_classic()

# Percentage of male customer's leaving the bank (16%) is less compared to the
# female customers(25%)  ==> female customer's are more likely to exit (when
# all other things are held constant) and infact, their exit rate is higher than
# the overall exit rate (20% -- shown with a hotizontal line)of the bank

# This is a kind of statitstical A/B test
```

```
# Tenure
prop.table(table(data$Tenure,data$Exited),1)
```

```
##
##         Exited    Stayed
##   0  0.2300242 0.7699758
##   1  0.2241546 0.7758454
##   2  0.1917939 0.8082061
##   3  0.2111001 0.7888999
##   4  0.2052578 0.7947422
##   5  0.2065217 0.7934783
##   6  0.2026887 0.7973113
##   7  0.1721790 0.8278210
##   8  0.1921951 0.8078049
##   9  0.2164634 0.7835366
##   10 0.2061224 0.7938776
```

```
tenure_ext <- ggplot(data, aes(x=Tenure,fill=Exited)) + geom_bar(position='fill') +
  scale_y_continuous(labels=percent_format()) +
  geom_hline(yintercept = 0.2) +
  annotate('text',x=1,y=c(0.13,0.60),label=c('23%','77%'),size=3.5) +
  annotate('text',x=2,y=c(0.13,0.60),label=c('22%','78%'),size=3.5) +
  annotate('text',x=3,y=c(0.13,0.60),label=c('19%','81%'),size=3.5) +
  annotate('text',x=4,y=c(0.13,0.60),label=c('21%','79%'),size=3.5) +
  annotate('text',x=5,y=c(0.13,0.60),label=c('20%','80%'),size=3.5) +
  annotate('text',x=6,y=c(0.13,0.60),label=c('20%','80%'),size=3.5) +
  annotate('text',x=7,y=c(0.13,0.60),label=c('20%','80%'),size=3.5) +
  annotate('text',x=8,y=c(0.13,0.60),label=c('17%','83%'),size=3.5) +
  annotate('text',x=9,y=c(0.13,0.60),label=c('19%','81%'),size=3.5) +
  annotate('text',x=10,y=c(0.13,0.60),label=c('22%','78%'),size=3.5) +
  annotate('text',x=11,y=c(0.13,0.60),label=c('21%','79%'),size=3.5) +
  scale_fill_manual(values=alpha(c('red','blue'),.5)) +
  labs(title = "Tenure", y = "% of Total Number of Records",x='') +
  theme_classic()

# Tenure might not provide much insight as the exit percentages are
# almost the same across the Tenure years
```

```
# NumOfProducts
prop.table(table(data$NumOfProducts,data$Exited),1)
```

```
##
##         Exited     Stayed
##   1 0.27714398 0.72285602
##   2 0.07581699 0.92418301
##   3 0.82706767 0.17293233
##   4 1.00000000 0.00000000
```

```
prods_ext <- ggplot(data, aes(x=NumOfProducts,fill=Exited)) + geom_bar(position='fill') +
  scale_y_continuous(labels=percent_format()) +
  geom_hline(yintercept = 0.2) +
  annotate('text',x=1,y=c(0.05,0.90),label=c('27%','73%'),size=4) +
  annotate('text',x=2,y=c(0.05,0.90),label=c('8%','92%'),size=4) +
  annotate('text',x=3,y=c(0.05,0.90),label=c('82%','18%'),size=4) +
  annotate('text',x=4,y=c(0.05),label=c('100%'),size=4) +
  scale_fill_manual(values=alpha(c('red','blue'),.5)) +
  labs(title = "NumOfProducts", y = "% of Total Number of Records",x='') +
  theme_classic()

# Looks like there are some anamolies -- generally, we expect that if there
# are more number of products, they are less likely to leave. This is true
# with customers having 1 and 2 products (exit percentages: 27% and 8%),
# however, the exit percentages of customers with 3 and 4 products are quite
# high (82% and 100% respectively). Though it is unusal, the total no. of
# customers in those categories are also very less (266 and 60).
```

```
# HasCrCard
prop.table(table(data$HasCrCard,data$Exited),1)
```

```
##
##          Exited    Stayed
##   NO  0.2081494 0.7918506
##   YES 0.2018427 0.7981573
```

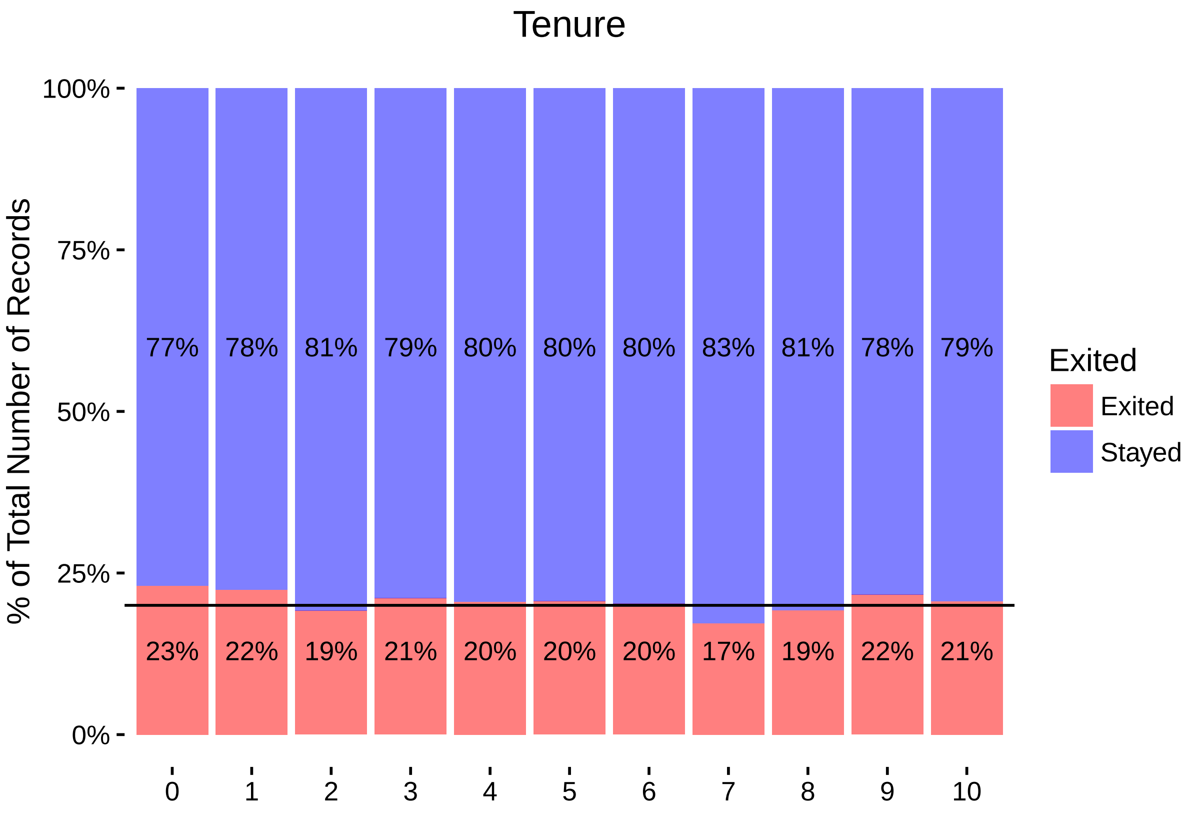
```
card_ext <- ggplot(data, aes(x=HasCrCard,fill=Exited)) + geom_bar(position='fill') +
  scale_y_continuous(labels=percent_format()) +
  geom_hline(yintercept=0.2) +
  annotate('text',x=1,y=c(0.15,0.60),label=c('21%','79%'),size=4) +
  annotate('text',x=2,y=c(0.15,0.60),label=c('20%','80%'),size=4) +
  scale_fill_manual(values=alpha(c('red','blue'),.5)) +
  labs(title = "HasCrCard", y = "% of Total Number of Records", x='') +
  theme_classic()

# Looks like the Credit card has less (or no) impact on the exit rates
# May be not an important feature in this particular case
```

```
# IsActiveMember
prop.table(table(data$IsActiveMember,data$Exited),1)
```

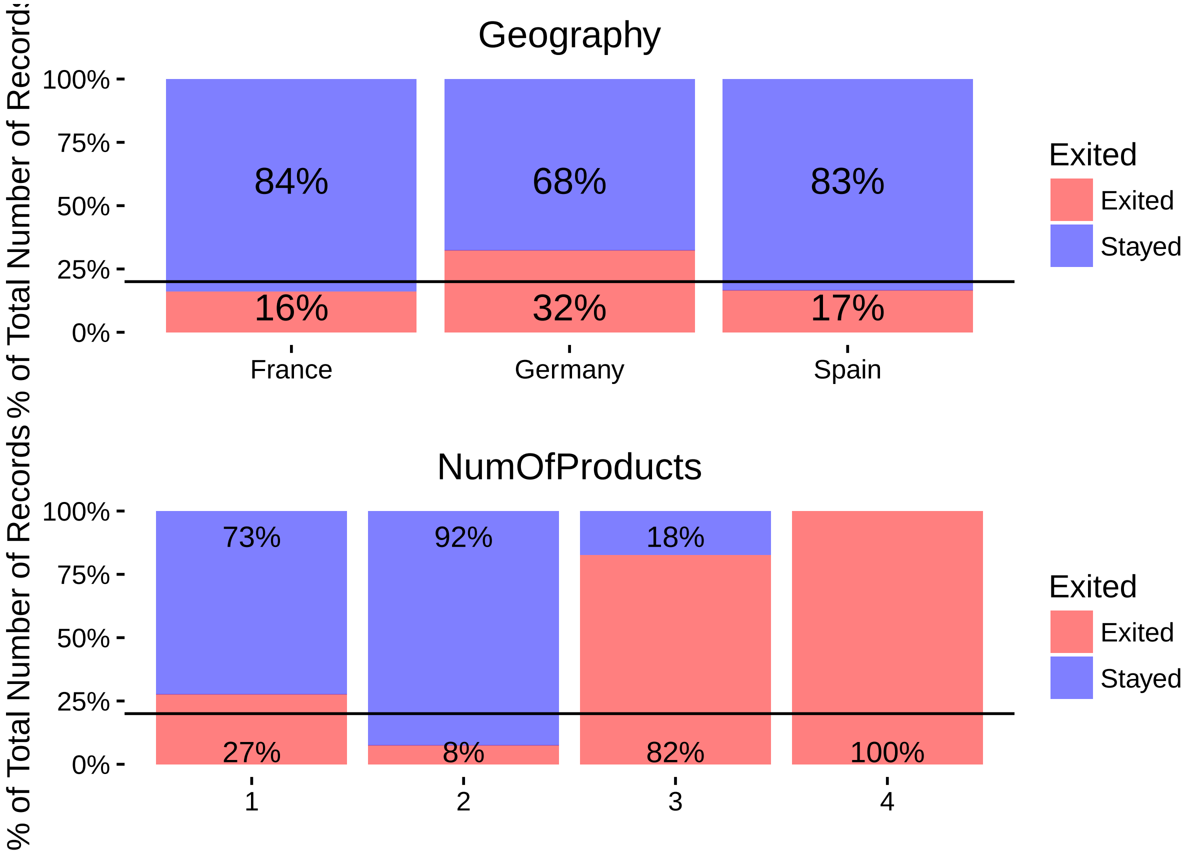
```
##
##          Exited    Stayed
##   NO  0.2685090 0.7314910
##   YES 0.1426907 0.8573093
```

```
active_ext <- ggplot(data, aes(x=IsActiveMember,fill=Exited)) + geom_bar(position='fill') +
  scale_y_continuous(labels=percent_format()) +
  geom_hline(yintercept = 0.2) +
  annotate('text',x=1,y=c(0.12,0.60),label=c('27%','73%'),size=4) +
  annotate('text',x=2,y=c(0.12,0.60),label=c('14%','86%'),size=4) +
  scale_fill_manual(values=alpha(c('red','blue'),.5)) +
  labs(title = "IsActiveMember", y = "% of Total Number of Records", x='') +
  theme_classic()

# Customer's who are not acitve - of them, 27% are left during the period of observation
# ==> bank needs to make their customer's to be active to keep them stay
```

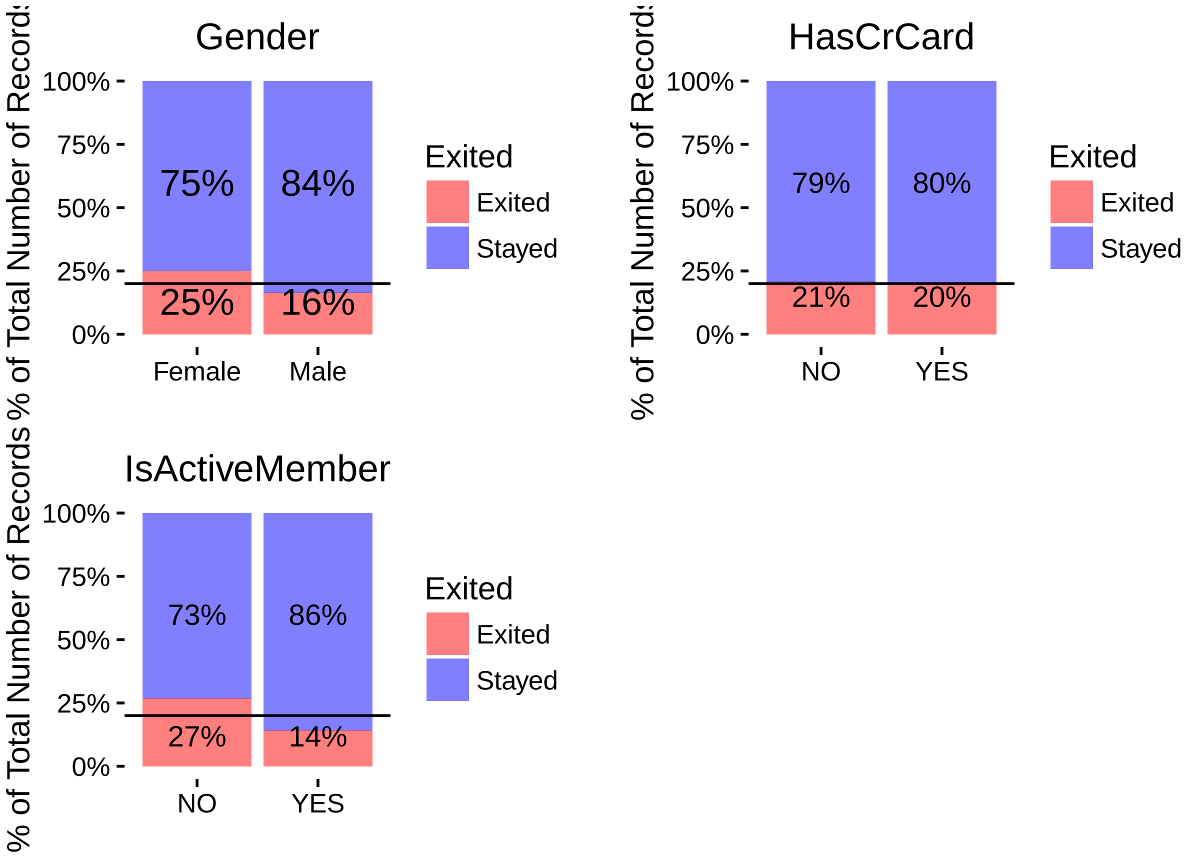
```
tenure_ext
```



```
grid.arrange(geo_ext,prods_ext,nrow=2)
```



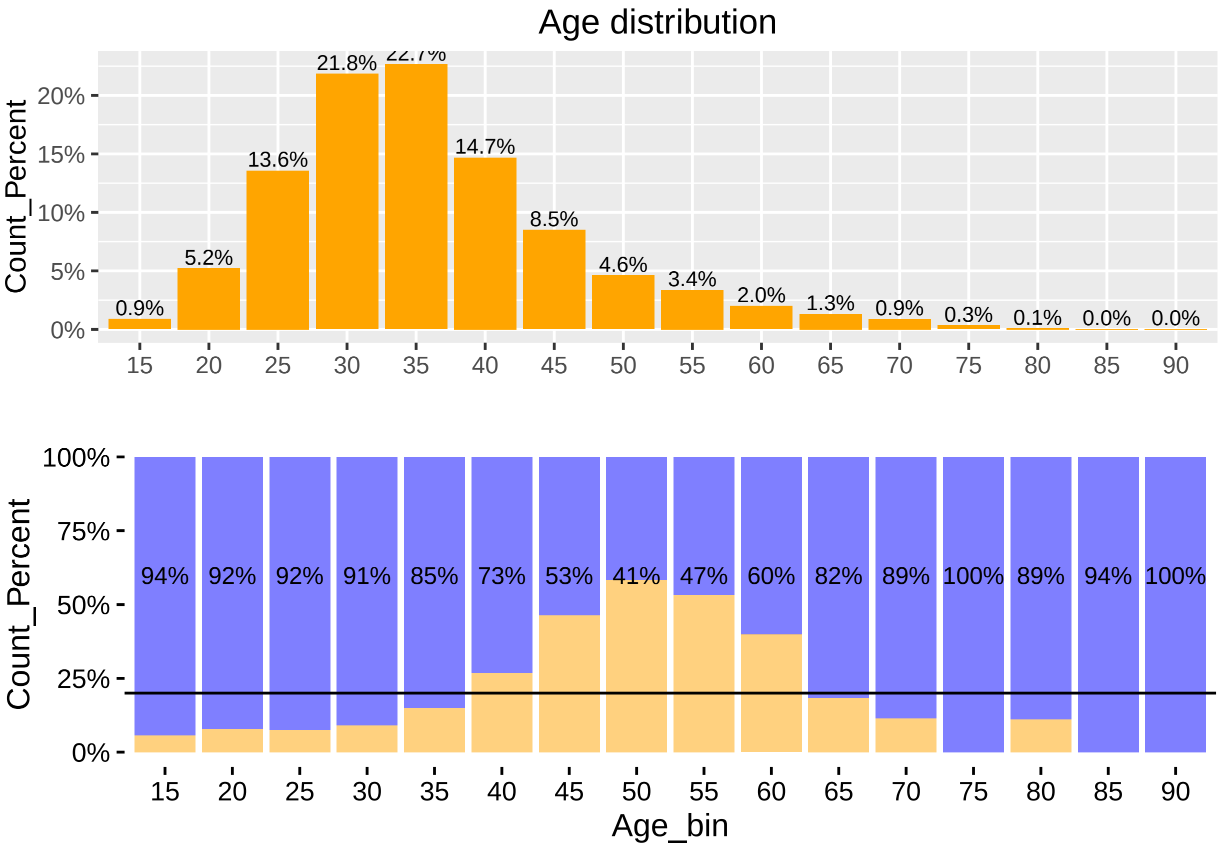
```r
grid.arrange(gender_ext,card_ext,active_ext,nrow=2)
```



```r
data$age_cat <- cut(data$Age, breaks=seq(15,90, by=5),include.lowest = TRUE)

age_hist <- ggplot(data,aes(age_cat)) + geom_bar(aes(y=(..count..)/sum(..count..)),fill='orange') +
  geom_text(size=2.9,aes(y=((..count..)/sum(..count..)),
  label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  scale_x_discrete(labels=seq(15,90,by=5)) +
  scale_y_continuous(labels=percent_format()) + labs(y='Count_Percent', title='Age distribution',x='')
  # +
  # theme(axis.text=element_text(size=10,face='bold'),
  #       axis.title=element_text(size=14,face='bold'))

# Almost 73% of the customers are in the age group of 25-40 ==> the bank has
# good number of younger customers (perhaps the reason why there are more people
# with less number of products)

prop.table(table(data$age_cat,data$Exited),1)
```

```
##
##               Exited     Stayed
##    [15,20] 0.05617978 0.94382022
##    (20,25] 0.07854406 0.92145594
##    (25,30] 0.07516581 0.92483419
##    (30,35] 0.09107551 0.90892449
##    (35,40] 0.14960282 0.85039718
##    (40,45] 0.26802721 0.73197279
##   (45,50] 0.46352941 0.53647059
##   (50,55] 0.58351410 0.41648590
##   (55,60] 0.53273810 0.46726190
##   (60,65] 0.40000000 0.60000000
##   (65,70] 0.18320611 0.81679389
##   (70,75] 0.11363636 0.88636364
##   (75,80] 0.00000000 1.00000000
##   (80,85] 0.11111111 0.88888889
##   (85,90] 0.00000000 1.00000000
```

```r
age_hist3 <- ggplot(data,aes(age_cat,fill=Exited)) + geom_bar(aes(y=(..count..)/sum(..count..)),positio
  #geom_text(size=2.9,aes(y=((..count..)/sum(..count..)),
  #label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  geom_hline(yintercept=0.2) +
  scale_x_discrete(labels=seq(15,90,by=5)) +
  scale_y_continuous(labels=percent_format()) + labs(y='Count_Percent',x='Age_bin')+
  scale_fill_manual(values=alpha(c('orange','blue'),.5)) +
  annotate('text',x=1,y=c(0.60),label=c('94%'),size=3) +
  annotate('text',x=2,y=c(0.60),label=c('92%'),size=3) +
  annotate('text',x=3,y=c(0.60),label=c('92%'),size=3) +
  annotate('text',x=4,y=c(0.60),label=c('91%'),size=3) +
  annotate('text',x=5,y=c(0.60),label=c('85%'),size=3) +
  annotate('text',x=6,y=c(0.60),label=c('73%'),size=3) +
  annotate('text',x=7,y=c(0.60),label=c('53%'),size=3) +
  annotate('text',x=8,y=c(0.60),label=c('41%'),size=3) +
  annotate('text',x=9,y=c(0.60),label=c('47%'),size=3) +
  annotate('text',x=10,y=c(0.60),label=c('60%'),size=3) +
  annotate('text',x=11,y=c(0.60),label=c('82%'),size=3) +
  annotate('text',x=12,y=c(0.60),label=c('89%'),size=3) +
  annotate('text',x=13,y=c(0.60),label=c('100%'),size=3) +
  annotate('text',x=14,y=c(0.60),label=c('89%'),size=3) +
  annotate('text',x=15,y=c(0.60),label=c('94%'),size=3) +
  annotate('text',x=16,y=c(0.60),label=c('100%'),size=3) +
  theme(axis.text=element_text(size=10,face='bold'),axis.title=element_text(size=14,face='bold')) +
  theme_classic() + theme(legend.position='none')

# people in the age group of 40 - 60 (middle age group) are more prone
# to exit the bank (might be having better offers from other banks or
# might be a serious financial crisis or else the programs at the bank
# are more benefitial to the age groups between 15-35) ==> The bank
# might need to focus on these age groups


#age_hist
#age_hist2
grid.arrange(age_hist,age_hist3,nrow=2)
```

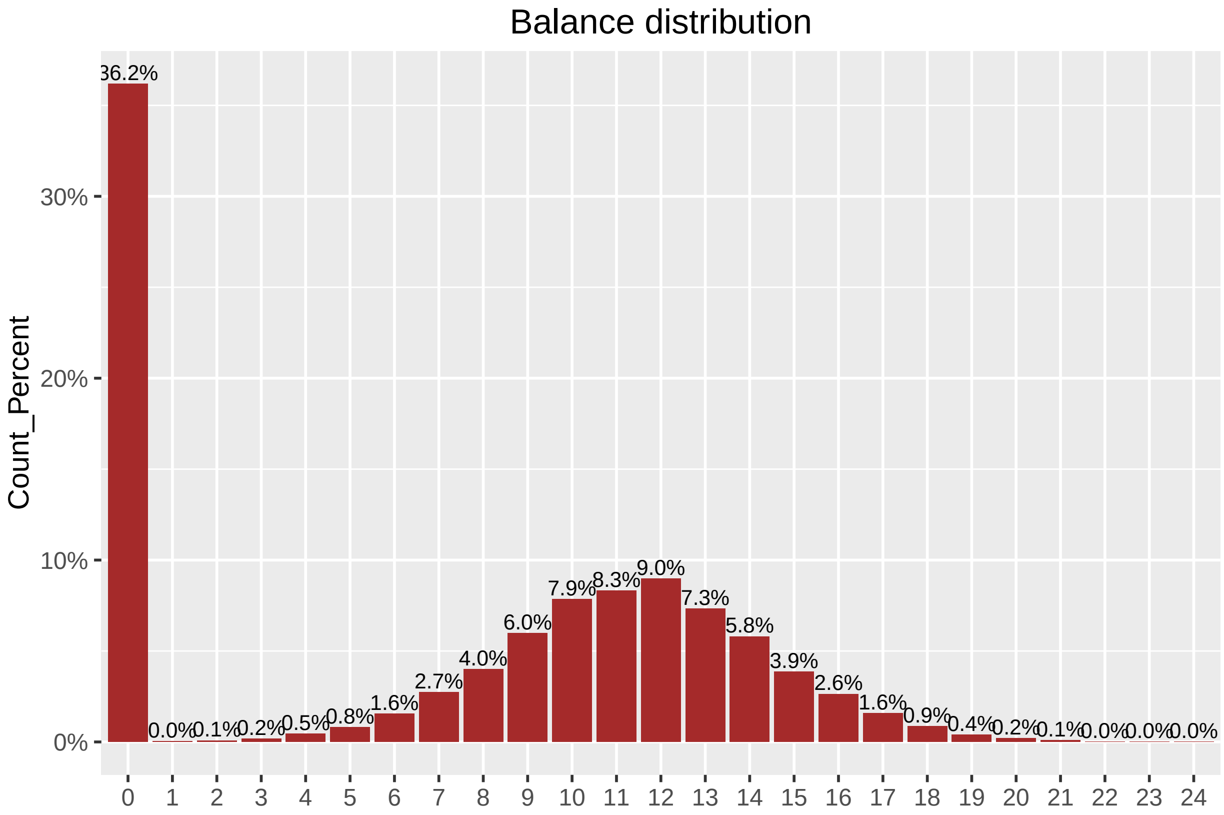


```
# the dual plot provides quite interesting insight:
# we have more number of customers from 25-40 age group (73%), but the customers
# from 40-60 are more intended to exit. Also there are very less number of customers
# in the age groups of 75-90 ==> the observed anamolies
```

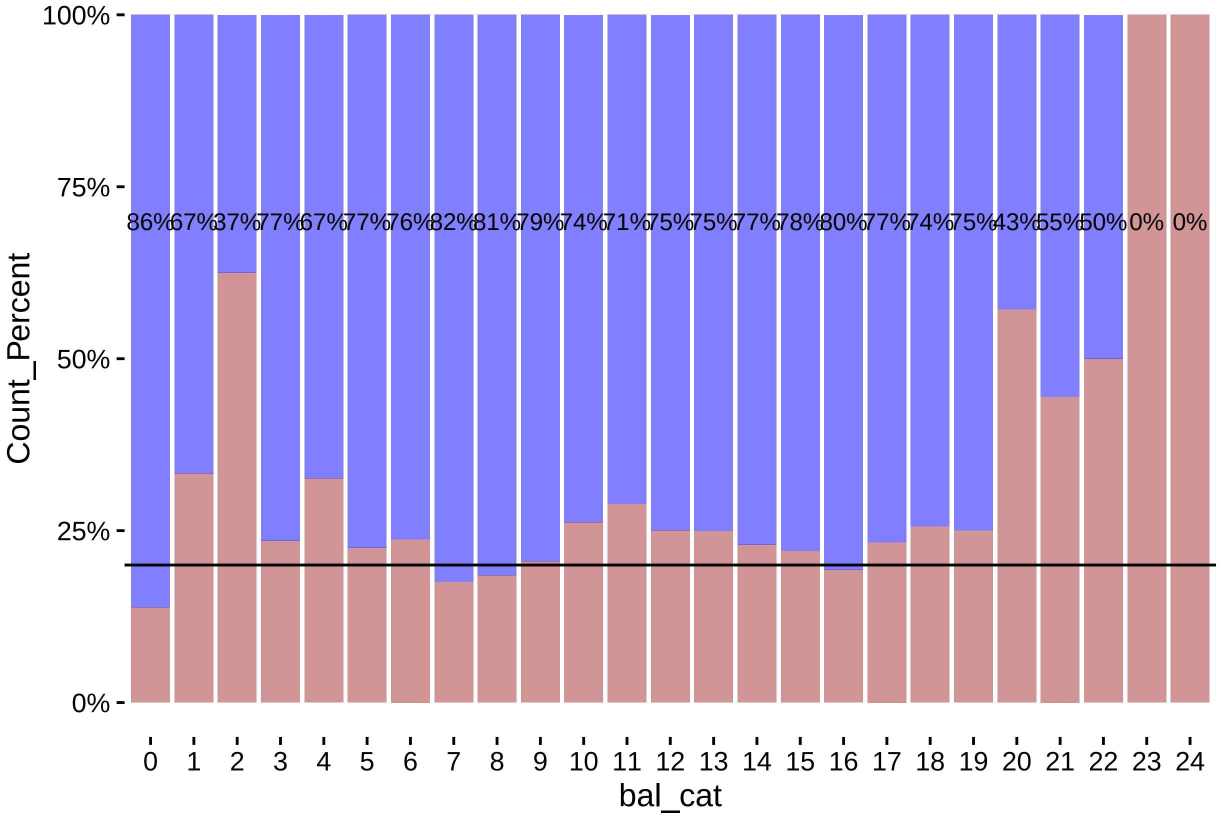
```
data$bal_cat <- cut(data$Balance, breaks=seq(0,260000, by=10000),include.lowest = TRUE)

bal_hist <- ggplot(data,aes(bal_cat)) + geom_bar(aes(y=(..count..)/sum(..count..)),fill='brown') +
  geom_text(size=2.9,aes(y=((..count..)/sum(..count..)),
  label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  scale_x_discrete(labels=seq(0,26,by=1)) +
  scale_y_continuous(labels=percent_format()) + labs(y='Count_Percent',x='',title='Balance distribution
# +
#   theme(axis.text=element_text(size=10,face='bold'),
#         axis.title=element_text(size=14,face='bold'))

# Its a nice distribution except at 0 position (we can cap/ransform it though !!!). It
# clearly indicates there are more people with zero balances

prop.table(table(data$bal_cat,data$Exited),1)
```

```
##
##                          Exited    Stayed
##   [0,1e+04]           0.1384743 0.8615257
##   (1e+04,2e+04]       0.3333333 0.6666667
##   (2e+04,3e+04]     0.6250000 0.3750000
##   (3e+04,4e+04]     0.2352941 0.7647059
##   (4e+04,5e+04]     0.3260870 0.6739130
##   (5e+04,6e+04]     0.2250000 0.7750000
##   (6e+04,7e+04]     0.2371795 0.7628205
##   (7e+04,8e+04]     0.1751825 0.8248175
##   (8e+04,9e+04]     0.1850000 0.8150000
##   (9e+04,1e+05]     0.2053422 0.7946578
##   (1e+05,1.1e+05]   0.2620865 0.7379135
##   (1.1e+05,1.2e+05] 0.2884615 0.7115385
##   (1.2e+05,1.3e+05] 0.2505568 0.7494432
##   (1.3e+05,1.4e+05] 0.2493188 0.7506812
##   (1.4e+05,1.5e+05] 0.2293103 0.7706897
##   (1.5e+05,1.6e+05] 0.2202073 0.7797927
##   (1.6e+05,1.7e+05] 0.1931818 0.8068182
##   (1.7e+05,1.8e+05] 0.2327044 0.7672956
##   (1.8e+05,1.9e+05] 0.2558140 0.7441860
##   (1.9e+05,2e+05]   0.2500000 0.7500000
##   (2e+05,2.1e+05]   0.5714286 0.4285714
##   (2.1e+05,2.2e+05] 0.4444444 0.5555556
##   (2.2e+05,2.3e+05] 0.5000000 0.5000000
##   (2.3e+05,2.4e+05] 1.0000000 0.0000000
##   (2.4e+05,2.5e+05]
##   (2.5e+05,2.6e+05] 1.0000000 0.0000000
```

```r
bal_hist2 <- ggplot(data,aes(bal_cat,fill=Exited)) + geom_bar(aes(y=(..count..)/sum(..count..)),positio
  #geom_text(size=2.9,aes(y=((..count..)/sum(..count..)),
  #label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  geom_hline(yintercept = 0.2) +
  scale_x_discrete(labels=seq(0,26,by=1)) +
  scale_y_continuous(labels=percent_format()) + labs(y='Count_Percent')+
  scale_fill_manual(values=alpha(c('brown','blue'),.5)) +
  annotate('text',x=1,y=c(0.70),label=c('86%'),size=3) +
  annotate('text',x=2,y=c(0.70),label=c('67%'),size=3) +
  annotate('text',x=3,y=c(0.70),label=c('37%'),size=3) +
  annotate('text',x=4,y=c(0.70),label=c('77%'),size=3) +
  annotate('text',x=5,y=c(0.70),label=c('67%'),size=3) +
  annotate('text',x=6,y=c(0.70),label=c('77%'),size=3) +
  annotate('text',x=7,y=c(0.70),label=c('76%'),size=3) +
  annotate('text',x=8,y=c(0.70),label=c('82%'),size=3) +
  annotate('text',x=9,y=c(0.70),label=c('81%'),size=3) +
  annotate('text',x=10,y=c(0.70),label=c('79%'),size=3) +
  annotate('text',x=11,y=c(0.70),label=c('74%'),size=3) +
  annotate('text',x=12,y=c(0.70),label=c('71%'),size=3) +
  annotate('text',x=13,y=c(0.70),label=c('75%'),size=3) +
  annotate('text',x=14,y=c(0.70),label=c('75%'),size=3) +
  annotate('text',x=15,y=c(0.70),label=c('77%'),size=3) +
  annotate('text',x=16,y=c(0.70),label=c('78%'),size=3) +
  annotate('text',x=17,y=c(0.70),label=c('80%'),size=3) +
  annotate('text',x=18,y=c(0.70),label=c('77%'),size=3) +
  annotate('text',x=19,y=c(0.70),label=c('74%'),size=3) +
  annotate('text',x=20,y=c(0.70),label=c('75%'),size=3) +
  annotate('text',x=21,y=c(0.70),label=c('43%'),size=3) +
  annotate('text',x=22,y=c(0.70),label=c('55%'),size=3) +
  annotate('text',x=23,y=c(0.70),label=c('50%'),size=3) +
  annotate('text',x=24,y=c(0.70),label=c('0%'),size=3) +
  annotate('text',x=25,y=c(0.70),label=c('0%'),size=3) +
 # annotate('text',x=26,y=c(0.60),label=c('0%'),size=3) +
  theme(axis.text=element_text(size=10,face='bold'),axis.title=element_text(size=14,face='bold')) +
  theme_classic() + theme(legend.position='none')

bal_hist
```



```
bal_hist2
```



```
#grid.arrange(bal_hist,bal_hist2,norw=2)

# Customers who have low and high balances are leaving more
# compared to the customers with medium range balances
```