Desafio_Data_Science

Henrique C Vieira

June 4, 2018

Analysis Bank

```
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(ggplot2)
library(tidyr)
library(readr)
```

Load and describe data

```
dataset <- read_delim('bank/bank.csv', delim = ';')
```

```
## Parsed with column specification:
## cols(
##   age = col_integer(),
##   job = col_character(),
##   marital = col_character(),
##   education = col_character(),
##   default = col_character(),
##   balance = col_integer(),
##   housing = col_character(),
##   loan = col_character(),
##   contact = col_character(),
##   day = col_integer(),
##   month = col_character(),
##   duration = col_integer(),
##   campaign = col_integer(),
##   pdays = col_integer(),
##   previous = col_integer(),
##   poutcome = col_character(),
##   y = col_character()
## )
```

```
glimpse(dataset)
```

```
## Observations: 4,521
## Variables: 17
## $ age       <int> 30, 33, 35, 30, 59, 35, 36, 39, 41, 43, 39, 43, 36, ...
## $ job       <chr> "unemployed", "services", "management", "management"...
## $ marital   <chr> "married", "married", "single", "married", "married"...
## $ education <chr> "primary", "secondary", "tertiary", "tertiary", "sec...
## $ default   <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no"...
## $ balance   <int> 1787, 4789, 1350, 1476, 0, 747, 307, 147, 221, -88, ...
## $ housing   <chr> "no", "yes", "yes", "yes", "yes", "no", "yes", "yes"...
## $ loan      <chr> "no", "yes", "no", "yes", "no", "no", "no", "no", "n...
## $ contact   <chr> "cellular", "cellular", "cellular", "unknown", "unkn...
## $ day       <int> 19, 11, 16, 3, 5, 23, 14, 6, 14, 17, 20, 17, 13, 30,...
## $ month     <chr> "oct", "may", "apr", "jun", "may", "feb", "may", "ma...
## $ 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  <chr> "unknown", "failure", "failure", "unknown", "unknown...
## $ y         <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no"...
```

Turn data in correct format

```
dataset$job <- as.factor(dataset$job)
dataset$marital <- as.factor(dataset$marital)
dataset$education <- as.factor(dataset$education)
dataset$default <- ifelse(dataset$default == 'yes', TRUE, FALSE)
dataset$housing <- ifelse(dataset$housing == 'yes', TRUE, FALSE)
dataset$loan <- ifelse(dataset$loan == 'yes', TRUE, FALSE)
dataset$contact <- as.factor(dataset$contact)
dataset$day <- as.factor(dataset$day)
dataset$month <- as.factor(dataset$month)
dataset$campaign <- as.factor(dataset$campaign)
dataset$poutcome <- as.factor(dataset$poutcome)
dataset$y <- ifelse(dataset$y == 'yes', TRUE, FALSE)
dataset$term <- dataset$y
glimpse(dataset)
```

```
## Observations: 4,521
## Variables: 18
## $ 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   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
## $ balance   <int> 1787, 4789, 1350, 1476, 0, 747, 307, 147, 221, -88, ...
## $ housing   <lgl> FALSE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TR...
## $ loan      <lgl> FALSE, TRUE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE...
## $ contact   <fctr> cellular, cellular, cellular, unknown, unknown, cel...
## $ day       <fctr> 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  <fctr> 1, 1, 1, 4, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 5, ...
```

```
## $ 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         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
## $ term      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
```

```
dataset %>%
  summary()
```

```
##       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
##  Mode :logical   Min.   :-3313   Mode :logical   Mode :logical
##  FALSE:4445      1st Qu.:   69   FALSE:1962      FALSE:3830
##  TRUE :76        Median :  444   TRUE :2559      TRUE :691
##                  Mean   : 1423
##                  3rd Qu.: 1480
##                  Max.   :71188
##
##       contact          day           month         duration
##  cellular :2896   20     : 257   may    :1398   Min.   :   4
##  telephone: 301   18     : 226   jul    : 706   1st Qu.: 104
##  unknown  :1324   19     : 201   aug    : 633   Median : 185
##                   21     : 198   jun    : 531   Mean   : 264
##                   14     : 195   nov    : 389   3rd Qu.: 329
##                   17     : 191   apr    : 293   Max.   :3025
##                   (Other):3253   (Other): 571
##     campaign        pdays           previous          poutcome
##  1      :1734   Min.   : -1.00   Min.   : 0.0000   failure: 490
##  2      :1264   1st Qu.: -1.00   1st Qu.: 0.0000   other  : 197
##  3      : 558   Median : -1.00   Median : 0.0000   success: 129
##  4      : 325   Mean   : 39.77   Mean   : 0.5426   unknown:3705
##  5      : 167   3rd Qu.: -1.00   3rd Qu.: 0.0000
##  6      : 155   Max.   :871.00   Max.   :25.0000
##  (Other): 318
##      y              term
##  Mode :logical   Mode :logical
##  FALSE:4000      FALSE:4000
##  TRUE :521       TRUE :521
##
##
##
##
```

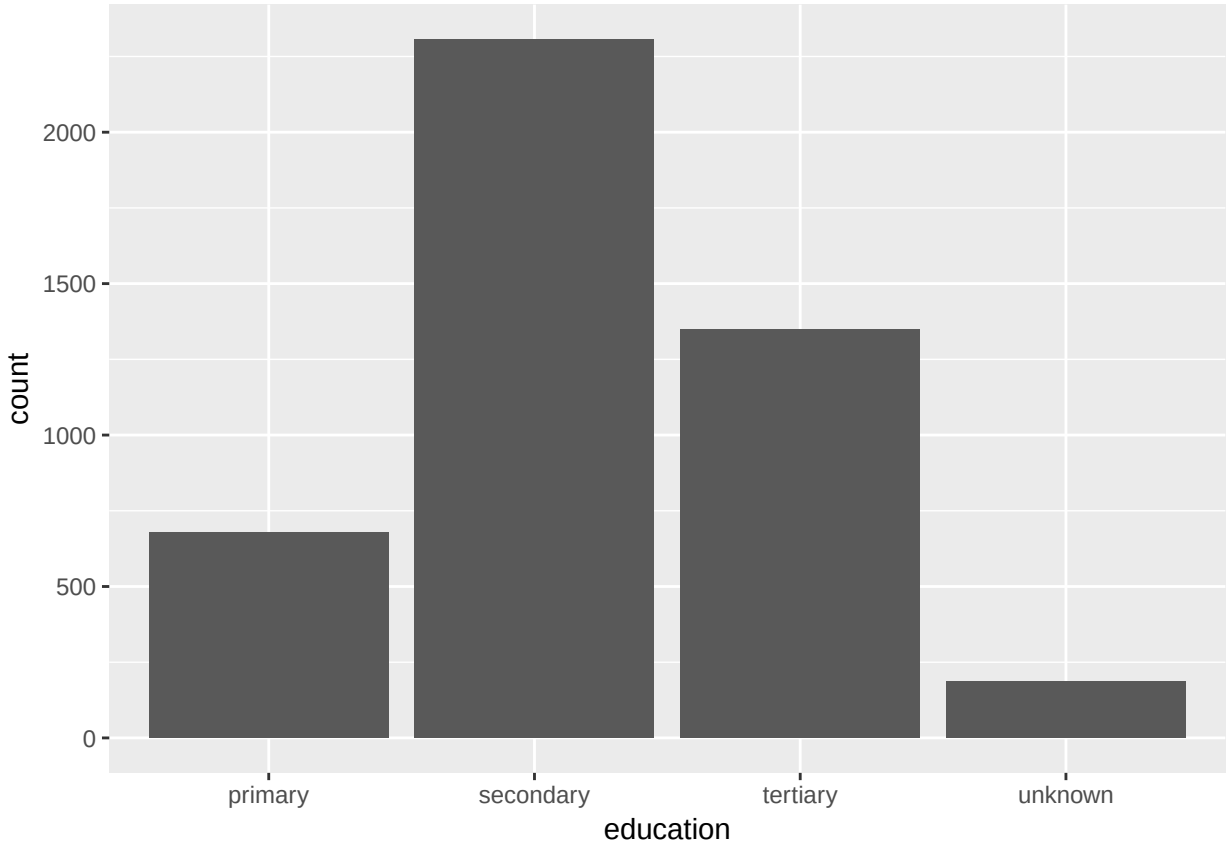
```r
dataset %>%
  group_by(job) %>%
  count(marital, sort = TRUE)
```

```
## # A tibble: 35 x 3
## # Groups:   job [12]
##            job marital     n
##         <fctr>  <fctr> <int>
##  1 blue-collar married   693
##  2  management married   557
##  3  technician married   411
##  4  management  single   293
##  5  technician  single   268
##  6      admin. married   266
##  7    services married   236
##  8     retired married   176
##  9 blue-collar  single   174
## 10      admin.  single   143
## # ... with 25 more rows
```

```r
dataset %>%
  group_by(poutcome) %>%
  filter(poutcome == "success") %>%
  count(campaign, sort = TRUE)
```

```
## # A tibble: 6 x 3
## # Groups:   poutcome [1]
##   poutcome campaign     n
##     <fctr>   <fctr> <int>
## 1  success        1    74
## 2  success        2    30
## 3  success        3    18
## 4  success        6     3
## 5  success        4     2
## 6  success        5     2
```

```r
ggplot(dataset, aes(x=education)) +
  geom_bar()
```

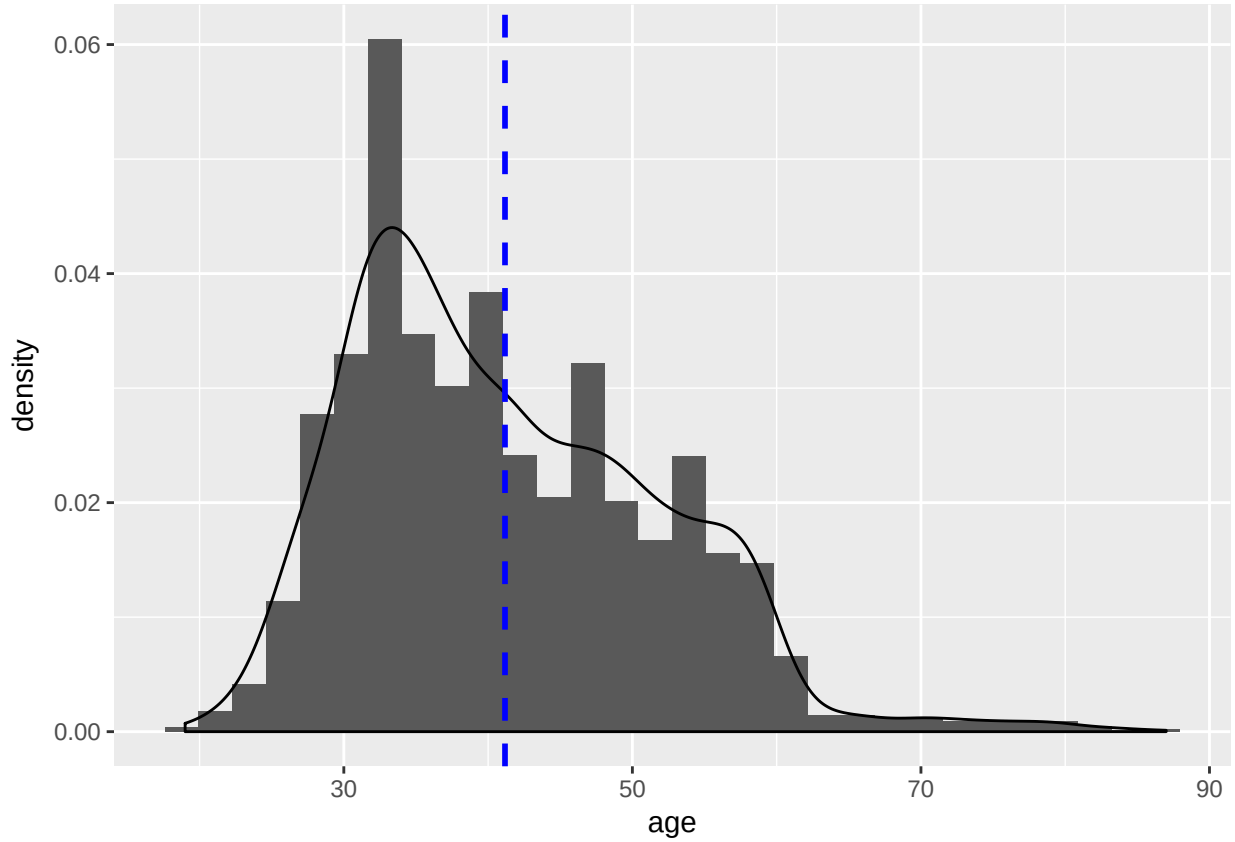



```
dataset %>%
  summarise(avg=mean(age), sd=sd(age))
```

```
## # A tibble: 1 x 2
##       avg       sd
##     <dbl>    <dbl>
## 1 41.1701 10.57621
```

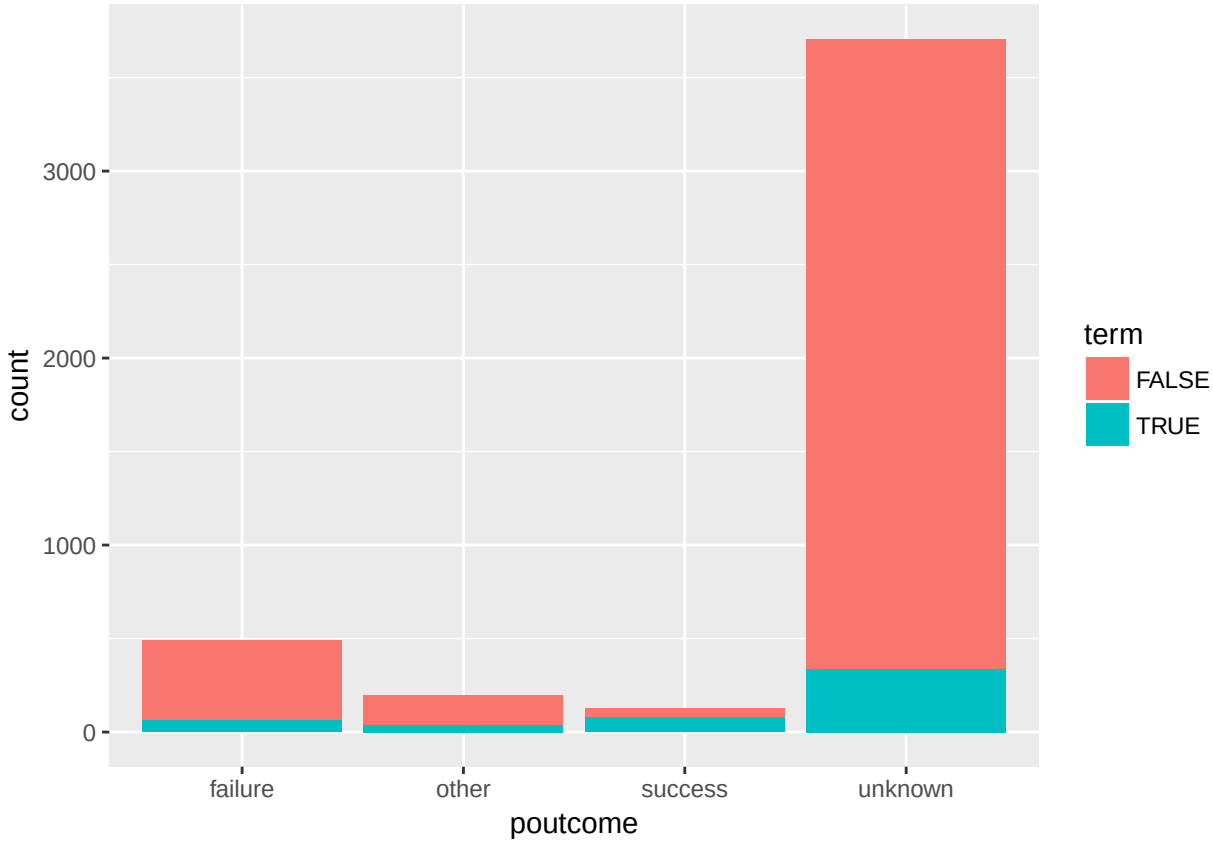
```
ggplot(dataset, aes(x=age, y=..density..)) +
  geom_histogram() +
  geom_density() +
  geom_vline(aes(xintercept=mean(age)),
             color="blue", linetype="dashed", size=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



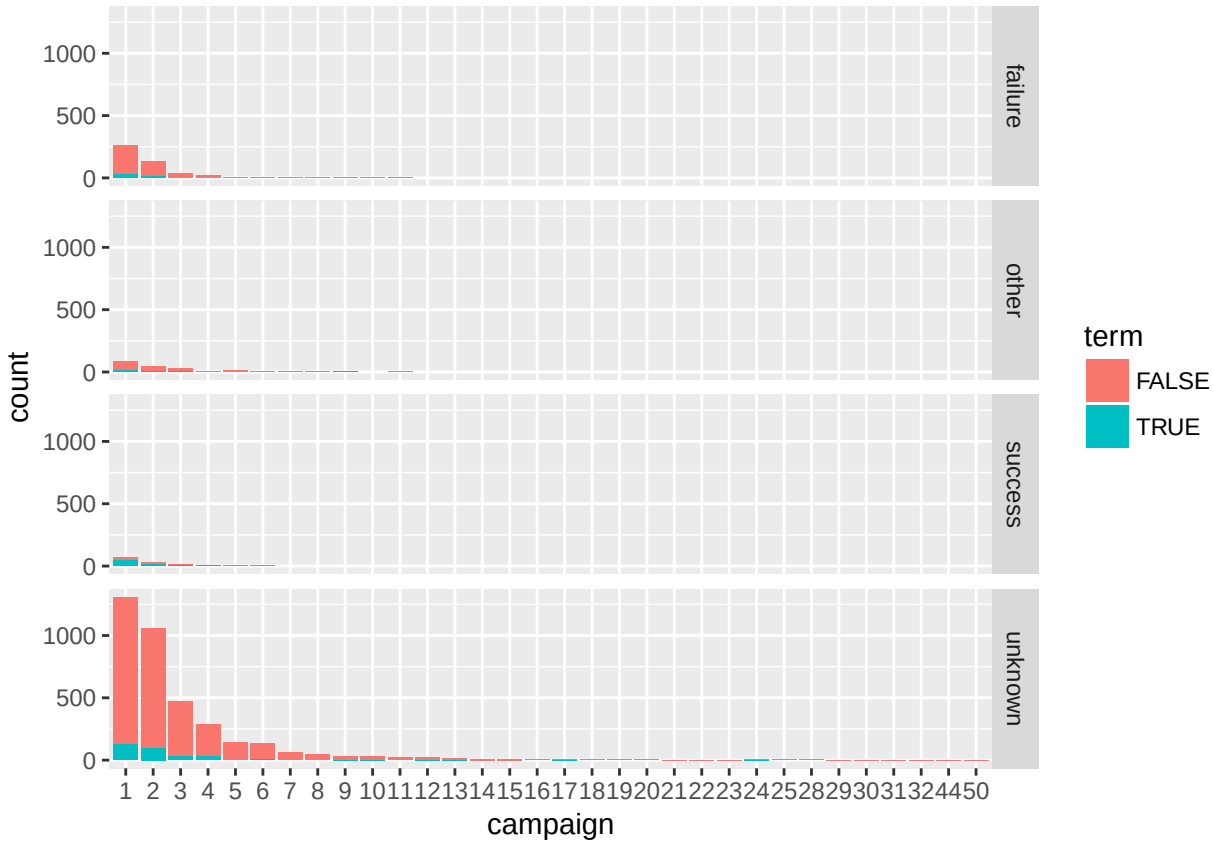
There are many unknown values

```
ggplot(dataset, aes(x=poutcome, fill=term)) +
  geom_bar()
```



View by campaign

```
ggplot(dataset, aes(x=campaign, fill=term)) +
  geom_bar() +
  facet_grid(poutcome ~ .)
```



Predict

```
library(naivebayes)
```

Make two groups: train and control

```
set.seed(100)
pos <- sample(1:nrow(dataset), round(nrow(dataset)*0.1, 0))
dataset_train <- dataset[-pos,]
dataset_test <- dataset[pos,]
class <- dataset_test$term
dataset_test <- dataset_test[,!colnames(dataset_test)=='term']
```

Job - Marital - Education

```
# job
# marital
# education
# default
# housing
# loan
# contact
# day
```

```r
# month
# campaign
# poutcome

m <- naive_bayes(term ~ job +
                        marital +
                        education +
                        default +
                        housing +
                        loan +
                        contact +
                        day +
                        month +
                        campaign +
                        poutcome, dataset_train)
```

```r
pred <- predict(m, dataset_test)
```

```r
confusion_matrix <- table(pred, class)
confusion_matrix
```

```
##        class
## pred    FALSE TRUE
##   FALSE   391   37
##   TRUE     15    9
```

```r
accuracy <- (confusion_matrix[1,2] + confusion_matrix[2,2]) / sum(confusion_matrix)
recall <- confusion_matrix[2,2] / (confusion_matrix[1,1] + confusion_matrix[2,2])
precision <- confusion_matrix[2,2] / (confusion_matrix[2,1] + confusion_matrix[2,2])

accuracy
```

```
## [1] 0.1017699
```

```r
recall
```

```
## [1] 0.0225
```

```r
precision
```

```
## [1] 0.375
```