MSDS Bridge Final Project-Week 3

Emmanuel Hayble-Gomes

7/29/2019

The Goal of the project is to analyze the impact of the portuguese banking campaign conducted by the Marketing Department to promote term deposit.

The sample size used for this project is 10% of the full bank data corresponding to 4521 randomly selected observations from the full dataset.

My goal for this project is focused on understanding the features of the data and attempt to predict term deposit using a categorical outcome variable.

Specifically, I will like to know what is the relationship between the age and the average yearly balance and find answers to questions about the population in terms of the age distribution, employment level, marital status, banking relationship,education and home ownership. This project will also attempt to identify the significant variables for predicting term deposit.

Task 1. Data Exploration (summary statistics, means, medians, quartiles, or any other relevant information about the dataset.
Please include some conclusions in the R Markdown text.)

```
getwd()
```

```
## [1] "C:/Users/Emahayz_Pro/Desktop/CUNY_Bridge/R-Class/Week3"
```

```
setwd("C:/Users/Emahayz_Pro/Desktop/CUNY_Bridge/R-Class/Week3")
Port_Bank <- read.csv("bank.csv", sep = ",")

head(Port_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
```

```r
summary(Port_Bank) # See Task 4 for answers
```

```
##       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                            
## 
```

```r
str(Port_Bank)    # See Task 4 for answers
```

```
## '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 ...
```

Task 2. Data wrangling (Perform some basic transformations.
to include column renaming,creating a subset of the data, replacing values, or creating new columns with derived data (for example summing two columns together)).

```
names(Port_Bank)[names(Port_Bank)== 'y'] <- 'term_deposit' # I renamed y
categorical variable as term_deposit.
head(Port_Bank) #View the new name
```

```
##   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 term_deposit
## 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
```

```
Port_Bank$term_deposit <- ifelse(Port_Bank$term_deposit=="yes",1,0)
str(Port_Bank)  #View the new number
```

```
## '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 ...
##  $ term_deposit: num  0 0 0 0 0 0 0 0 0 0 ...

# I renamed y categorical variable as term_deposit for the purpose of
analysis.
# This is a categorical variable with two factors “Yes” or “No”,
# I also replaced or converted the term_deposit factor values to numeric
using binary “1” and “0” with Yes = 1 and No = 0.

# Creating a subset of the data:
set.seed(101)

train.size <- 0.7 # I created a subset/sample with 70% of the data known as
train.
Port_train <- runif(nrow(Port_Bank))< train.size
Bank_train <- Port_Bank[Port_train, ]
Bank_test <- Port_Bank[!Port_train, ]

head(Bank_train) #Viewing the new dataframe for Bank_train

##   age           job marital education default balance housing loan
## 1  30    unemployed married   primary      no    1787      no   no
## 2  33      services married secondary      no    4789     yes  yes
## 4  30    management married  tertiary      no    1476     yes  yes
## 5  59   blue-collar married secondary      no       0     yes   no
## 6  35    management  single  tertiary      no     747      no   no
## 7  36 self-employed married  tertiary      no     307     yes   no
##    contact day month duration campaign pdays previous poutcome
## 1 cellular  19   oct       79        1    -1        0  unknown
## 2 cellular  11   may      220        1   339        4  failure
## 4  unknown   3   jun      199        4    -1        0  unknown
## 5  unknown   5   may      226        1    -1        0  unknown
## 6 cellular  23   feb      141        2   176        3  failure
## 7 cellular  14   may      341        1   330        2    other
##   term_deposit
## 1            0
## 2            0
## 4            0
## 5            0
## 6            0
## 7            0
```

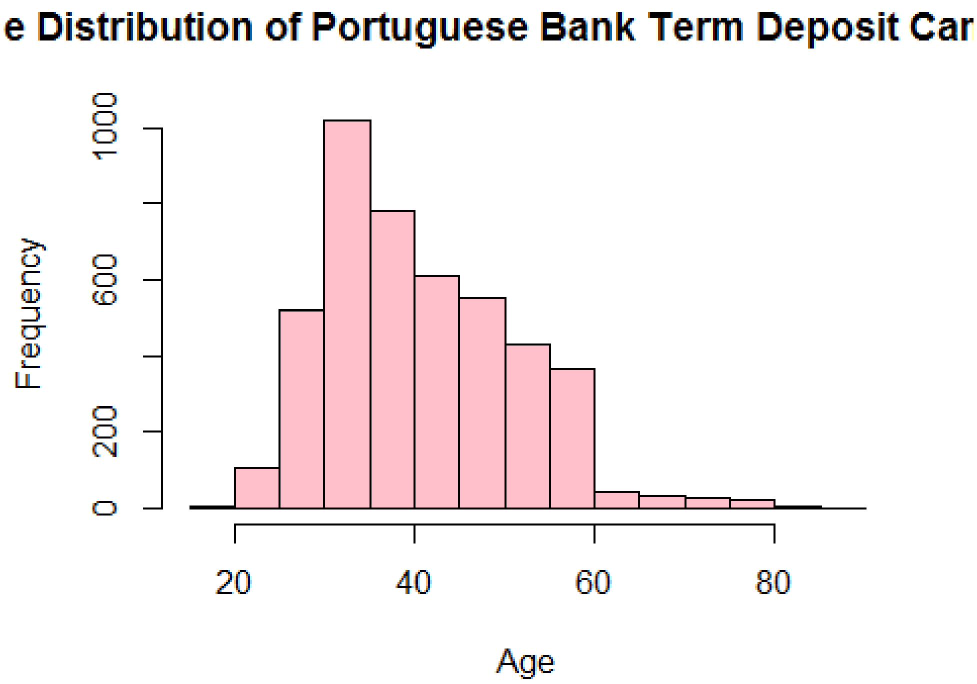
```r
head(Bank_test)  #Viewing the new dataframe for Bank_test
```

```
##    age        job marital education default balance housing loan  contact
## 3   35 management  single  tertiary      no    1350     yes   no cellular
## 11  39   services married secondary      no    9374     yes   no  unknown
## 12  43     admin. married secondary      no     264     yes   no cellular
## 13  36 technician married  tertiary      no    1109      no   no cellular
## 14  20    student  single secondary      no     502      no   no cellular
## 17  56 technician married secondary      no    4073      no   no cellular
##    day month duration campaign pdays previous poutcome term_deposit
## 3   16   apr      185        1   330        1  failure            0
## 11  20   may      273        1    -1        0  unknown            0
## 12  17   apr      113        2    -1        0  unknown            0
## 13  13   aug      328        2    -1        0  unknown            0
## 14  30   apr      261        1    -1        0  unknown            1
## 17  27   aug      239        5    -1        0  unknown            0
```

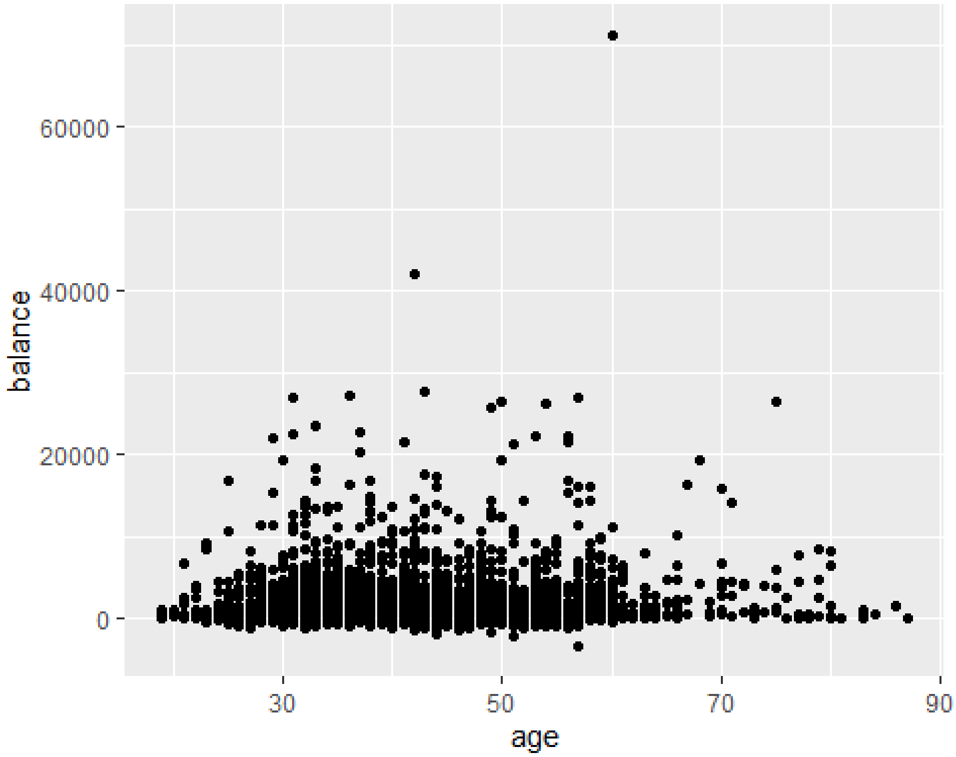
Task 3. Graphics (Please make sure to display at least one scatter plot, box plot and histogram.
Don’t be limited to this.
Please explore the many other options in R packages such as ggplot2).

Visualization:

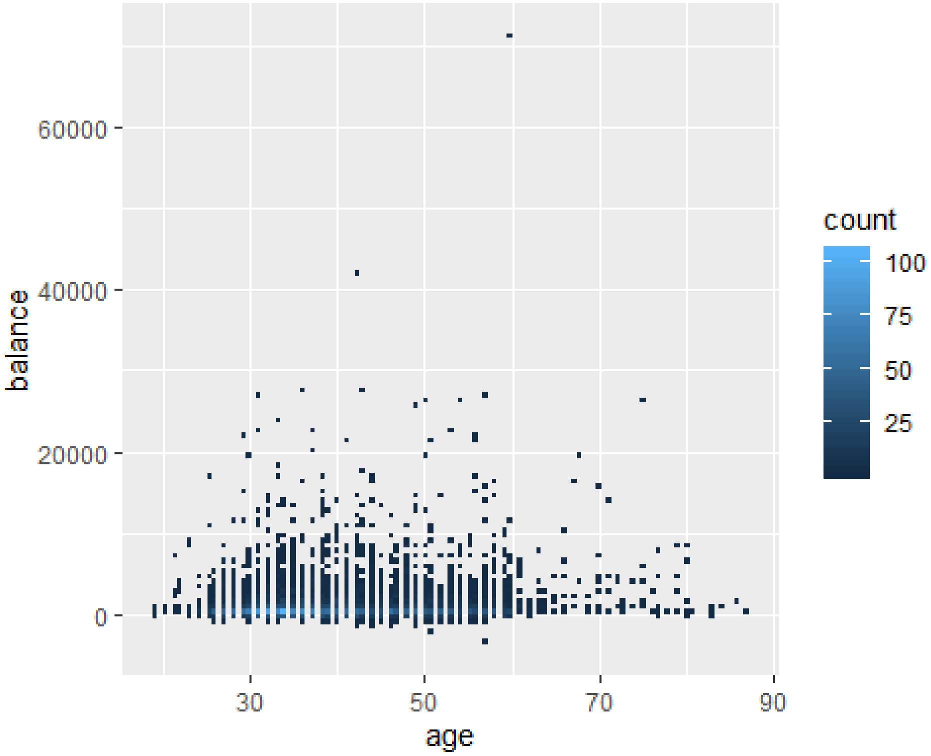
```r
library(ggplot2)

# Histogram using age variable
hist(Port_Bank$age, main = "Age Distribution of Portuguese Bank Term Deposit Campaign", xlab = "Age", ylab = "Frequency", col = "pink") #The histogram shows that majority of the population is between the age of 30 to 35.
```



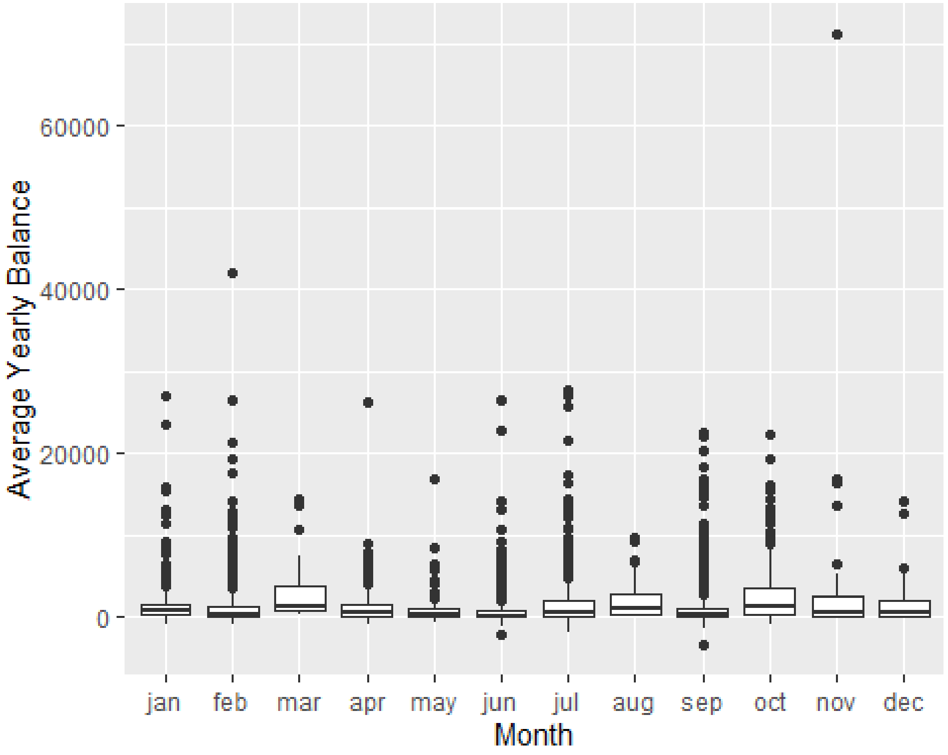
```
# Scatter plot using age and account balance variables
ggplot(Port_Bank, aes(x = age, y = balance))+
    geom_point() # Scatter Plot
```



```
ggplot(Port_Bank,aes(age,balance))+ geom_bin2d(bins = 120) # Improved scatter
plot with 2d bins
```



```
# Box plot
# Just Exploring here with ploting a box plot using factor variable (Month)
Port_Bank$month <- factor(Port_Bank$month,
                          labels =
c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","dec"))

ggplot(Port_Bank,aes(x = month, y = balance))+
    geom_boxplot()+ scale_x_discrete(name = "Month")+
  scale_y_continuous(name = "Average Yearly Balance")
```



Task 4. Meaningful question for analysis (Please state at the beginning a meaningful question for analysis.
Use the first three steps and anything else that would be helpful to answer the question you are posing from the data ##set you chose.
Please write a brief conclusion paragraph in R markdown at the end.).

Data Exploration: There are 10 factor variables and 7 integer variables in this dataset. Please see the Conclusion below for details.

Data Wrangling: I renamed y categorical variable as term_deposit for the purpose of analysis. This is a categorical variable with two factors “Yes” or “No”, I also replaced or converted the term_deposit factor ##values to number using binary “1” and “0” with Yes = 1 and No = 0.

Building a Logistic Regression Model to Predict term deposit. I will use the train dataset which is the 70% sample ##created earlier.

Bank_train #70% of the Port_Bank dataframe Bank_test #30% of the Port_Bank dataframe

```
Port_Bank_logit <- glm(term_deposit ~., data = Bank_train, family =
binomial(), maxit = 100)
summary(Port_Bank_logit) # Note AIC shows 1524.9 is this a good fit? Good for
comparing models, the smaller AIC score is better.
```

```
##
## Call:
## glm(formula = term_deposit ~ ., family = binomial(), data = Bank_train,
##     maxit = 100)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.6957  -0.3742  -0.2395  -0.1407   3.1441  
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        -2.304e+00  7.448e-01  -3.093  0.00198 ** 
## age                -7.800e-03  8.869e-03  -0.879  0.37914    
## jobblue-collar     -6.222e-01  3.068e-01  -2.028  0.04254 *  
## jobentrepreneur    -2.095e-01  4.914e-01  -0.426  0.66989    
## jobhousemaid        6.294e-02  4.675e-01   0.135  0.89290    
## jobmanagement       6.870e-04  3.009e-01   0.002  0.99818    
## jobretired          8.641e-01  3.800e-01   2.274  0.02296 *  
## jobself-employed   -3.514e-01  4.455e-01  -0.789  0.43031    
## jobservices        -2.854e-01  3.448e-01  -0.828  0.40788    
## jobstudent          8.543e-02  5.159e-01   0.166  0.86847    
## jobtechnician      -3.427e-01  2.930e-01  -1.170  0.24213    
## jobunemployed      -7.978e-01  5.047e-01  -1.581  0.11393    
## jobunknown          8.720e-01  7.208e-01   1.210  0.22636    
## maritalmarried     -2.887e-01  2.162e-01  -1.336  0.18166    
## maritalsingle      -2.162e-01  2.533e-01  -0.854  0.39332    
## educationsecondary -8.938e-03  2.467e-01  -0.036  0.97110    
## educationtertiary   2.459e-01  2.867e-01   0.858  0.39099    
## educationunknown   -6.552e-01  4.763e-01  -1.376  0.16895    
## defaultyes          6.288e-01  4.679e-01   1.344  0.17900    
## balance            -6.330e-06  1.991e-05  -0.318  0.75055    
## housingyes         -1.902e-01  1.711e-01  -1.111  0.26641    
## loanyes            -4.789e-01  2.383e-01  -2.010  0.04448 *  
## contacttelephone    5.375e-02  2.723e-01   0.197  0.84350    
## contactunknown     -1.558e+00  2.913e-01  -5.350 8.81e-08 ***
## day                 7.922e-03  1.006e-02   0.788  0.43089    
## monthaug           -2.754e-01  3.003e-01  -0.917  0.35919    
## monthdec           -1.459e+00  1.131e+00  -1.290  0.19704    
## monthfeb            2.811e-01  3.413e-01   0.824  0.41018    
## monthjan           -1.407e+00  5.260e-01  -2.676  0.00746 ** 
## monthjul           -7.890e-01  3.062e-01  -2.577  0.00997 ** 
## monthjun            5.619e-01  3.619e-01   1.552  0.12056    
## monthmar            1.462e+00  4.538e-01   3.222  0.00127 ** 
## monthmay           -6.204e-01  2.890e-01  -2.147  0.03183 *  
## monthnov           -1.156e+00  3.600e-01  -3.210  0.00133 ** 
## monthoct            1.272e+00  4.373e-01   2.908  0.00364 ** 
## monthsep            8.196e-01  4.893e-01   1.675  0.09390 .  
## duration            4.170e-03  2.421e-04  17.226  < 2e-16 ***
## campaign           -8.963e-02  3.611e-02  -2.482  0.01306 *  
## pdays              -3.485e-04  1.316e-03  -0.265  0.79117    
## previous           -2.118e-02  4.679e-02  -0.453  0.65081    
## poutcomeother       7.728e-01  3.324e-01   2.325  0.02006 *  
## poutcomesuccess     2.385e+00  3.402e-01   7.011 2.36e-12 ***
## poutcomeunknown    -1.950e-03  4.075e-01  -0.005  0.99618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2163.2  on 3145  degrees of freedom
## Residual deviance: 1438.9  on 3103  degrees of freedom
## AIC: 1524.9
##
## Number of Fisher Scoring iterations: 6
```

```
# Test of Varaible Significance using Chi Square
anova(Port_Bank_logit, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: term_deposit
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       3145     2163.2
## age        1    10.46      3144     2152.8  0.001217 **
## job       11    48.36      3133     2104.4 1.231e-06 ***
## marital    2     8.61      3131     2095.8  0.013514 *
## education  3     7.96      3128     2087.8  0.046802 *
## default    1     0.29      3127     2087.6  0.592701
## balance    1     0.00      3126     2087.6  0.952796
## housing    1    16.76      3125     2070.8 4.245e-05 ***
## loan       1     8.86      3124     2061.9  0.002916 **
## contact    2    47.10      3122     2014.8 5.909e-11 ***
## day        1     5.02      3121     2009.8  0.025063 *
## month     11    87.72      3110     1922.1 4.657e-14 ***
## duration   1   400.56      3109     1521.5 < 2.2e-16 ***
## campaign   1     7.98      3108     1513.6  0.004721 **
## pdays      1     7.56      3107     1506.0  0.005959 **
## previous   1     0.65      3106     1505.3  0.421488
## poutcome   3    66.48      3103     1438.9 2.414e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Predicting the term deposit
Predict_term <- predict(Port_Bank_logit,type = "response")
table(Bank_train$term_deposit,Predict_term > 0.5)
```

```
##
##     FALSE TRUE
##   0  2744   60
##   1   226  116

# Using the term_deposit variable from the train dataset to generate a
confusion matrix.
# The model accurately predicted 2,744 as True Negative (TN)and 116 as True
Positive (TP)
```

I want to validate the Model using the test data

```
Port_Bank_Val <- glm(term_deposit ~., data = Bank_test, family = binomial(),
maxit = 100)
summary(Port_Bank_Val)
```

```
## 
## Call:
## glm(formula = term_deposit ~ ., family = binomial(), data = Bank_test, 
##     maxit = 100)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.3762  -0.3931  -0.2570  -0.1526   2.8618  
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        -2.722e+00  1.104e+00  -2.465 0.013690 *  
## age                -2.914e-03  1.289e-02  -0.226 0.821077    
## jobblue-collar      1.668e-01  4.181e-01   0.399 0.689953    
## jobentrepreneur    -2.461e-01  6.469e-01  -0.380 0.703605    
## jobhousemaid       -2.009e+00  1.208e+00  -1.663 0.096405 .  
## jobmanagement      -1.502e-01  4.275e-01  -0.351 0.725350    
## jobretired          1.226e-01  6.239e-01   0.196 0.844264    
## jobself-employed    2.098e-01  6.022e-01   0.348 0.727619    
## jobservices         3.186e-01  4.792e-01   0.665 0.506160    
## jobstudent          7.205e-01  6.193e-01   1.163 0.244689    
## jobtechnician       1.658e-01  3.979e-01   0.417 0.676838    
## jobunemployed      -8.070e-02  7.851e-01  -0.103 0.918127    
## jobunknown         -1.883e-01  1.065e+00  -0.177 0.859704    
## maritalmarried     -8.800e-01  3.151e-01  -2.793 0.005222 ** 
## maritalsingle      -4.868e-01  3.683e-01  -1.322 0.186221    
## educationsecondary  2.056e-01  3.729e-01   0.551 0.581314    
## educationtertiary   4.295e-01  4.298e-01   0.999 0.317613    
## educationunknown   -1.642e-01  5.984e-01  -0.274 0.783820    
## defaultyes          9.417e-01  1.244e+00   0.757 0.449229    
## balance             2.715e-05  3.851e-05   0.705 0.480867    
## housingyes         -4.705e-01  2.544e-01  -1.849 0.064389 .  
## loanyes            -9.185e-01  3.838e-01  -2.393 0.016708 *  
## contacttelephone   -1.820e-01  4.630e-01  -0.393 0.694223    
## contactunknown     -1.302e+00  3.816e-01  -3.412 0.000646 ***
## day                 3.765e-02  1.498e-02   2.513 0.011957 *  
## monthaug           -5.603e-01  4.720e-01  -1.187 0.235215    
## monthdec            1.365e+00  1.156e+00   1.181 0.237719
## monthfeb           -3.142e-01  6.572e-01  -0.478 0.632549
## monthjan           -1.153e+00  6.125e-01  -1.883 0.059713 .
## monthjul           -7.383e-01  4.482e-01  -1.647 0.099518 .
## monthjun            6.291e-01  5.488e-01   1.146 0.251691
## monthmar            1.651e+00  8.210e-01   2.010 0.044394 *
## monthmay           -2.663e-01  4.189e-01  -0.636 0.524894
## monthnov           -5.131e-01  4.540e-01  -1.130 0.258418
## monthoct            1.405e+00  5.495e-01   2.556 0.010576 *
## monthsep            1.475e-01  8.253e-01   0.179 0.858184
## duration            4.592e-03  3.925e-04  11.700  < 2e-16 ***
## campaign           -3.839e-02  4.752e-02  -0.808 0.419151
## pdays               3.228e-04  1.624e-03   0.199 0.842475
## previous            3.608e-02  8.040e-02   0.449 0.653595
## poutcomeother       5.338e-02  5.200e-01   0.103 0.918230
## poutcomesuccess     3.109e+00  6.017e-01   5.168 2.37e-07 ***
## poutcomeunknown    -3.609e-01  5.593e-01  -0.645 0.518768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1063.51  on 1374  degrees of freedom
## Residual deviance:  680.35  on 1332  degrees of freedom
## AIC: 766.35
##
## Number of Fisher Scoring iterations: 6
```

```
# Predicting the term deposit in the test data
Predict_term_Val <- predict(Port_Bank_Val,type = "response")
table(Bank_test$term_deposit,Predict_term_Val > 0.5)
```

```
##
##     FALSE TRUE
##   0  1168   28
##   1   104   75
```

```
# Using the term_deposit variable from the test dataset to generate a
confusion matrix.
# The model accurately predicted 1,168 as True Negative (TN)and 75 as True
Positive (TP)
```

BONUS –place the original .csv in a github file and have R read from the link.
This will be a very useful skill as you progress in your data science education and career.

```
library(RCurl) # Loading the RCurl package will enable me to read the csv
file using the link from my Github
```

```
## Loading required package: bitops
```

```
Port_Bank <- read.csv(text =
getURL("https://raw.githubusercontent.com/Emahayz/MSDS_R_Class/master/bank.cs
v"), header = T, sep = ",")
head(Port_Bank) # The original Salaries csv file is successfully read.
```

```
##   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
```

Conclusion

Data Exploration:

There are 10 factor variables and 7 integer variables in this dataset. From the summary statistics of the data, the average age of this population is about 41 years old and the median age is 39, the lower and upper quartiles are 33 and 49 respectively.

The histogram shows that majority of the population is between the age of 30 to 35. Most of the people have jobs in Management representing 969 of the population while 230 people are retired. There are 2,797 married couples and 1,196 unmarried people while 528 are divorced.

A significant portion of the population has at least secondary education (2,306) while 1,350 has college degree. Only 691 people have existing loan with the bank and 76 of those people have defaulted on a loan. A significant number of the population (3,830) do not have any existing loan with the bank.

About 2,559 of the population are home owners and 521 people already have existing term deposit account. The scatter plots show that account average yearly balance does not increase with age, a significant portion of the population with age greater than 30 had negative average yearly balance. However, there are outliers at age about 42 and 60 years with over €40,000 and €70,000 average yearly balance. The Boxplot shows that the outliers occurred in the month of February and November.

Data Wrangling:

I renamed y categorical variable as term_deposit for the purpose of analysis. This is a categorical variable with two factors “Yes” or “No”, I also replaced or converted the term_deposit factor values to numeric using binary “1” and “0” with Yes = 1 and No = 0.

Test of variable Significance:

The Chi Square shows that the following variables are strongly significant for predicting term deposit: Job situation, housing condition, type of contact used (cell phone, landline etc), the month of the year for the campaign, duration-time since last contact and poutcome- outcome of the previous marketing campaign.