ETC 3250
FINAL PROJECT

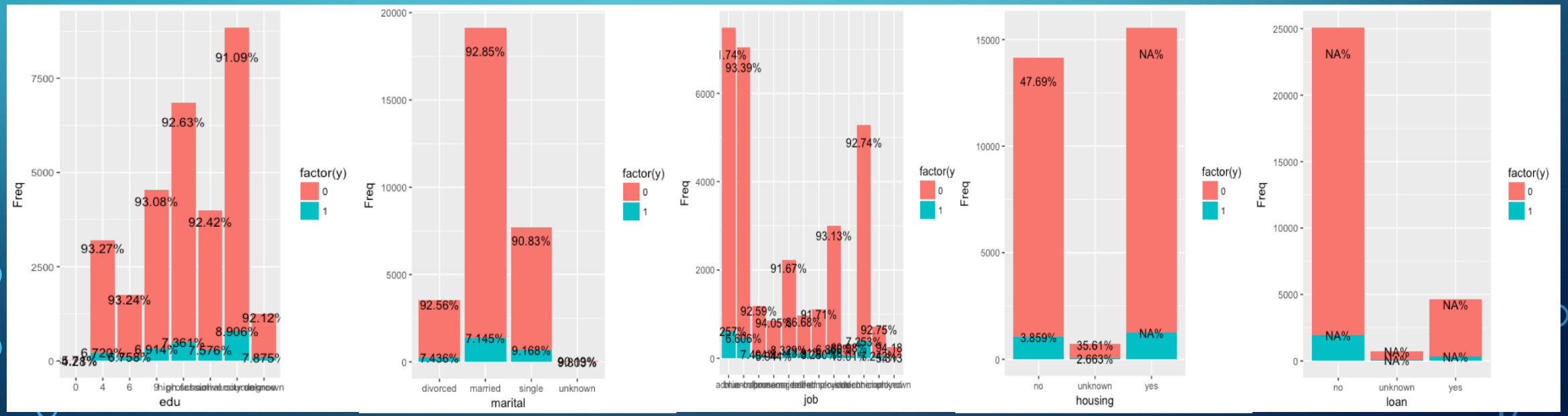
GROUP MEMBER: YOUCHENG CAO， JINGYING WANG, YI YU

INTRODUCTION

The purpose of the project is to build a model to predict the probability that a client will subscribe to a bank term deposit on the basis of these predictors. The main process we are going through includes accessing data, dealing with data and fitting them into different models, comparing models and choosing the best one.

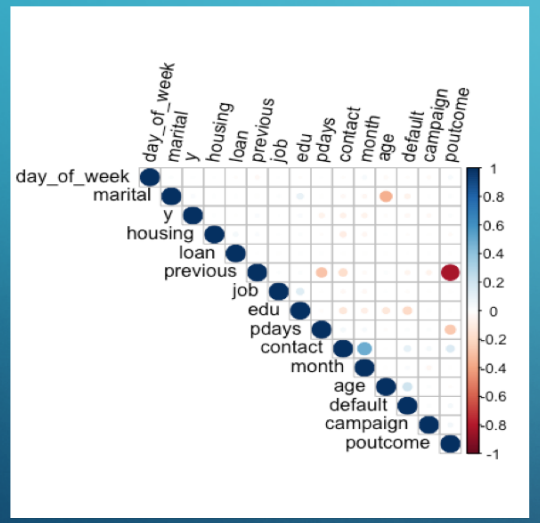
ACCESSING DATA AND VISUALISING THEM

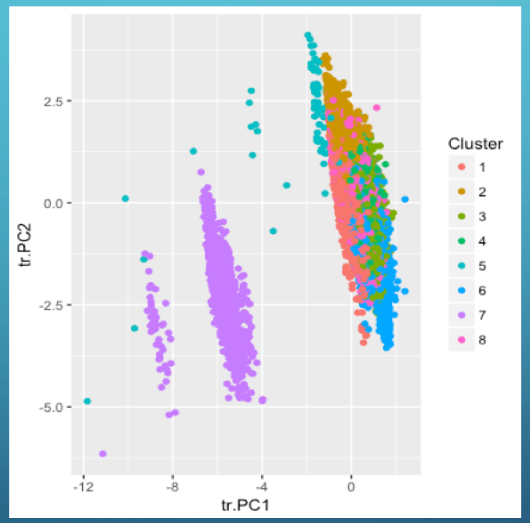
1. Drawing graphs to find the relationship between factor and each variables.



DEALING WITH DATA

1. Drawing graph to find the relationship between variables
2. Select outliers using PCA and k means method.

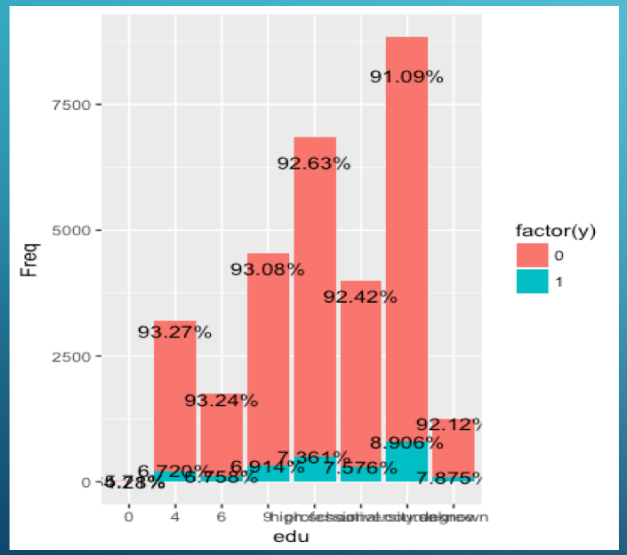




DEALING WITH DATA

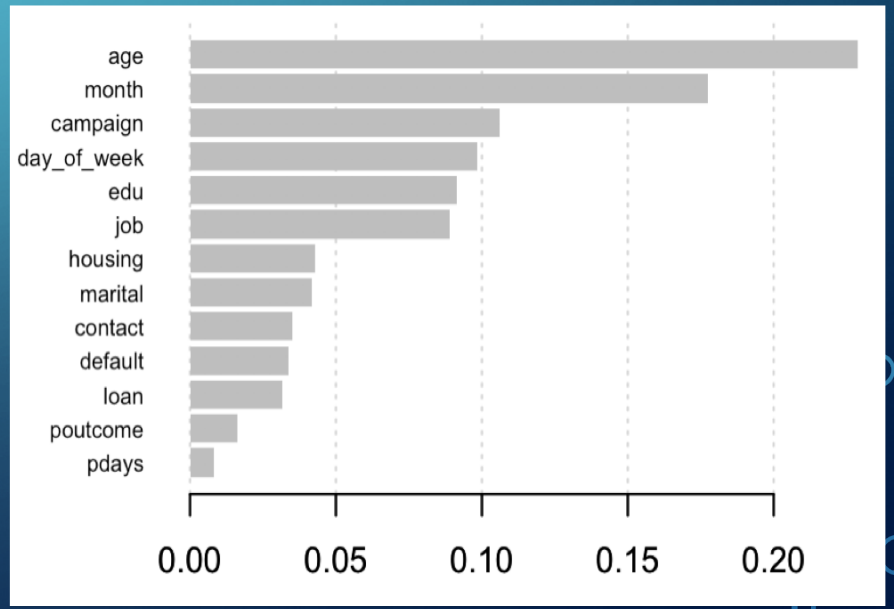
3. Deleting the outliers and variables with abnormal attribute(unknown)

4. Discretization



MODEL FITTING

1. Dividing data into train set and test set
2. Use train set to build model, then use test to check the error rate
   (1) using stepwise regression in logistic model
   (2) setting depth in decision tree method



BUILDING MODELS

- Advantages and Disadvantages for models:
- Logistic: More robust, able to handle nonlinear effect. Hard to predict continuous outcomes and cannot ensure accuracy.
- Tree: Easy to understand and explain, able to deal with qualitative predictors without creating dummy variables. Can be very non-robust.

SUMMARY

Error rate logistic:0.57646

Error rate tree:0.57755

Choose lower error rate: logistic model

```
> summary(glm.fit.full.dev)

Call:
glm(formula = y ~ age + campaign + pdays + job_retired_numeric +
    job_student_numeric + default_no_numeric + contact_numeric +
    month_mar_numeric + month_may_numeric + month_jun_numeric +
    month_jul_numeric + month_aug_numeric + month_oct_numeric +
    month_nov_numeric + day_mon_numeric + day_wed_numeric + day_thu_numeric +
    poutcome_failure_numeric + edu_basic__numeric + edu_high__numeric +
    edu_prof__numeric, family = binomial, data = tr_full[, -5:-6])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6751  -0.3973  -0.3478  -0.3048   2.7383

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.0252366  0.2505797  -0.101 0.919778
age                     -0.0048942  0.0025601  -1.912 0.055913 .
campaign                -0.0324302  0.0090234  -3.594 0.000326 ***
pdays                   -0.0011674  0.0002143  -5.449 5.07e-08 ***
job_retired_numeric      0.6051690  0.1148049   5.271 1.35e-07 ***
job_student_numeric      0.4764695  0.1628024   2.927 0.003426 **
default_no_numeric       0.1296002  0.0581927   2.227 0.025941 *
contact_numeric         -0.2527286  0.0988894  -2.556 0.010598 *
month_mar_numeric        1.1396784  0.1336264   8.529  < 2e-16 ***
month_may_numeric       -1.3392728  0.1218229 -10.994  < 2e-16 ***
month_jun_numeric       -1.0657715  0.1268662  -8.401  < 2e-16 ***
month_jul_numeric       -1.0044602  0.0747890 -13.431  < 2e-16 ***
month_aug_numeric       -1.2899039  0.0827139 -15.595  < 2e-16 ***
month_oct_numeric        2.1465330  0.2758826   7.781 7.22e-15 ***
month_nov_numeric       -1.1180502  0.0849012 -13.169  < 2e-16 ***
day_mon_numeric         -0.1808049  0.0639439  -2.828 0.004690 **
day_wed_numeric          0.1740754  0.0605964   2.873 0.004070 **
day_thu_numeric          0.1221188  0.0585116   2.087 0.036880 *
poutcome_failure_numeric -0.5811122  0.1109817  -5.236 1.64e-07 ***
edu_basic__numeric       -0.1814289  0.0574727  -3.157 0.001595 **
edu_high__numeric        -0.1642478  0.0606119  -2.710 0.006732 **
edu_prof__numeric        -0.1272991  0.0716117  -1.778 0.075465 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16512  on 30435  degrees of freedom
Residual deviance: 15445  on 30414  degrees of freedom
AIC: 15489

Number of Fisher Scoring iterations: 5

>
```