How to do Effective and Sucessful Bank Telemarketing

Arindam Barman[1], Mohamed Elmoudni[1], Shazia Khan[1], Kishore Prasad[1]

[1] City University of New York (CUNY)

Author note

Abstract

the Title page, the Abstract page, and the references page do not count when using APA Style . . . as such, you need to have 12 pages besides those three pages use 250 words or less to summarize your problem, methodology, and major outcomes. Even though direct marketing is a standard method for banks to utilize in the face of competition and financial unstability, it has, however, been shown to exhibit poor performance. The telemarketing calls are simply not answered or answered and immediately disconnected. It is however welcomed by the right person who is in need of financial relief. The aim of this exercise is to target clients more effectively and efficiently based on the data from a Portuguese bank telemarketing effort. We first used logistic regression to predict the binary response variable. The outcomes. . . .

Keywords: select a few key words (up to five) related to your work. . . .logistic regression model, linear discriminant analysis (LDA), predictive modeling, bank telemarketing, direct marketing, Data Mining

How to do Effective and Sucessful Bank Telemarketing

Introduction

describe the background and motivation of your problem–

After looking at various options, we settled for this project for our final since it met all the requirements.

"Regression analysis is one of the most commonly used statistical techniques in social and behavioral sciences as well as in physical sciences. Its main objective is to explore the relationship between a dependent variable and one or more independent variables (which are also called predictor or explanatory variables)." This is the definition provided by www.unesco.org for Regression Analysis

The most successful direct marketing is to predict the customers that have a higher probability to do business. Data exploration technique, is crucial to understand customer behavior. Many banks and services are moving to adopt the predictive technique based on the data mining to predict the customer profile before targeting them. The prediction or classification is the most important task in the data exploration and model building that is usually applied to classify the group of data. In classification, the outcome is a categorical variable and several combinations of input variable are used to build a model and the model that gives a better prediction with the best accuracy is chosen to target the prospective customers.

The data set contains approximately 41188 obs. of 21 variables.

This dataset is based on "Bank Marketing" UCI dataset (please check the description at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The data is enriched by the addition of five new social and economic features/attributes (national wide indicators from a ~10M population country), published by the Banco de Portugal and publicly available at: https://www.bportugal.pt/estatisticasweb./

The binary classification goal is to predict if the client will subscribe a bank term deposit (variable y).

This dependent variable tells whether the client will subscribe a bank term deposit or not. This is a binary variable and as such we will be using a Logistic Regression Model.

Literature Review

discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are. Explain how your investigation is similar or different to the state-of-the-art. Please do not discuss paper one at a time, instead, identify key characteristics of your topic, and discuss them in a whole. Please cite the relevant papers where appropriate.

We will be reviewing three papers addressing the same problem of bank telemarketing.

1. http://bru-unide.iscte.pt/RePEc/pdfs/13-06.pdf
2. http://www.ijmbs.com/Vol6/1/4-vaidehi-r.pdf
3. http://www.columbia.edu/~jc4133/ADA-Project.pdf

Methodology

discuss the key aspects of your problem, data set and regression model(s). Given that you are working on real-world data, explain at a high-level your exploratory data analysis, how you prepared the data for regression modeling, your process for building regression models, and your model selection. .

The data is available on website for UC Irvine Machine Learning Repository. There are two different data sets available. The “bank” data has 45,211 records with 16 attributes and 1 response variable. The “bank-additional” data has 41,188 records with additional attributes added to “bank” data, it has 20 attributes and 1 response variable. We chose to use the data with additional attributes.

The data consists of four groups of information.

- Client’s personal infomation

- Client's bank information
- Bank's telemarketing campaign information
- Social and economic information

The main problem with the dataset is that it consists of many missing values which are labeled "Unknown". The missing data consists of 26% of the data. We decided to retain the missing data to help with our regression modeling. The other problem with the data is that only 12% of the data shows the response variable to be "y".

We looked at each varable and the unique values contained in each variable and what they represented. We can divide the variables in the following three categories:

1 - Binary values of "yes" and "no" wit null values given as "unknown".

2 - Categorical values with "unknown" as missing values. The categorical variable require dummy variables to be created for each unique value. We included "unknown" as one of the dummy variable.

3 - numeric values with "999" as indication of null value. We created a variable to indicate if the data was missing or present.

Experimentation and Results

describe the specifics of what you did (data exploration, data preparation, model building, selection, evaluation) and what you found out (statistical analysis, inter presentation and discussion of the results)

Data Exploration

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:

-Variable identification

-Missing values and Unique Values

-Variables relationship to y

We notice that the variables are numerical, categorical and binary. The response variable y is binary.

Based on the original dataset, our predictor input has 21 variables. And our response variable is 1 variable called y.

Binomial Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more metric (interval or ratio scale) independent variables.

5.1.3 Preliminary Data Analysis

5.1.4 Analysis of Predictor variable

Table 1

Variable Analysis

Variable	Data.Type	Analysis
age	Numeric	No significant trend with responses variable, better response with age grp<
job	Catagorical	12 levels, proportion of responses from admin and blue collar job profiles ar
marital	Catagorical	4 levels, % response from marital status from single is greater compare to o
education	Catagorical	8 levels, responses from education with university degree are higher
default	Binary	3 levels, response is from no default group is dominant and some responses
housing	Binary	3 levels, no significant difference in association for three different groups
loan	Binary	4 levels, no significant difference in association for three different groups
contact	Catagorical	2 levels, responses from cellular contact is higher

Variable	Data.Type	Analysis
day_of_week	Catagorical	5 levels, response from customer is better on Wed,Thu, Tue
month	Catagorical	10 levels, there is significant variations of responses from Customers
duration	Numeric	closely associated with response variable with threshold for positive respons
campaign	Numeric	Number of campaign has impact on positive response of the campaign
pdays	Numeric	This variable does not seem to have strong relationship with response varial
previous	Numeric	previous contacts seems to have influence on the positive response of the ca
poutcome	Catagorical	have relationship with campaign outcome, earlier success has better respons
emp.var.rate	Numeric	lower the variation rates higher the number of positive outcome
cons.price.idx	Numeric	lower consumer price index seems to have higher positive response rate
cons.conf.idx	Numeric	lower confidence index brings more success to the campaign as people tend
euribor3m	Numeric	lower rate has association with more number of positive cases
nr.employed	Numeric	lower the number of employee higher the number of positive responses

5.1.4 Missing values

We see that there are no missing values in our dataset as shown in table 2 and graph format. The unique values are given in the table

5.1.5 Proportion of Response Variables

5.2 Data Preparation

-Convert Binary to 0 and 1

-Create dummy variables

-Data Summary Analysis

-Correlation of Variables with y

5.2.1 Convert Binary yes and no to 0 and 1

Now in order to prepare the data for modeling, we need to update Yes = 1 and No = 0.

5.2.2 Create dummy variables

Now we need to create dummy variables to find out the relationship between y variables and dependent variables, for all categorical variables.

Prepare test data. We will treat the test data the same way as the train data, and then apply models created using the treated train data.

5.2.3 Data Summary with Dummy variables

5.2.4 Correlation between Response and Predictor of Variables

Now we will produce the correlation table between the independent variables and the dependent variable

2.5 Outliers Handling

5.2.6 Analysis the link function for given variables

In this section, we will investigate how our initial data aligns with a typical logistic model plot.

Recall the Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). The fundamental equation of generalized linear model is:

g(E(y)) = a+ Bx1+B2x2+ B3x_3+...

where, g() is the link function, E(y) is the expectation of target variable and B0 + B1x1 + B2x2+B3x3 is the linear predictor B0,B1,B2, B3 to be predicted. The role of link function is to "link" the expectation of y to linear predictor.

In logistic regression, we are only concerned about the probability of outcome dependent variable success or failure. As described above, g() is the link function. This function is established using two things: Probability of Success as p and Probability of Failure as 1-p. p should meet following criteria: It must always be positive (since p >= 0) It must always be less than equals to 1 (since p <= 1).

Now let's investigate how our initial data model aligns with the above criteria. In other words, we will plot regression model plots for each variable and compare it to a typical logistic model plot:

The main objective in the transformations is to achieve linear relationships with the dependent variable or, really, with its logit.

Methodology

CRISP-DM Methodology has been used for this assignment .....Need material???? image/process flow

Business Understanding :

Data Exploration :

5.1 Data Exploration

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:

-Variable identification

-Understanding predictor variables relationship with response variable -Missing values and Unique Values

Data Preparation :

Methodology

Business Understanding :

Data Exploration :

5.1 Data Exploration

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:

-Variable identification

-Understanding predictor variables relationship with response variable -Missing values and Unique Values

Data Preparation :

Modeling:

Logistics Regression:. Logistic Regression is a probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor.Logistics model doesn't suffer a lot from severe class imbalance. Logistic Regression creates log odds of the response as a linear function of predictor variables. Many of the categorical predictors in the data set for this project have sparse and unbalanced distributions. Using logistics model with the given set of data would need adjustment of variables to fine tune the model.

Classification Tree. Classification Tree is used to predict the outcome of a categorical response variable.The purpose of the analyses via tree-building algorithms is to determine a set of logical conditional split that permit accurate classification of cases and accurate prediction.Effectiveness of classification tree model with binary variable is one of the reason for selection for this analysis study. This model though has problem with over fitting. We will also create RandomForest model to overcome that.'

RandomForest Model. Random Forests grows many classification trees for given set of response and predictor variables. Each tree gives a classification, and all the outputs from different trees are "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Over fitting problem with the classification tree can be overcome by this approach with weighted average of more number of trees. This method is good for prediction but a little bit difficult to interpret. Since we are facing the binary category,Random Forest is a good classification method to try.

Evaluation

There are number of ways to evaluate the regression and classification models based on the purpose like prediction, classification, variable selection etc. In the given business

scenario objective is to classification of the response variable by building a model that can predict likelihood of response from Customer. Following evaluation criteria we have used for model evaluation-

(1) The Hosmer-Lemeshow test assesses the model calibration and how predicted values tend to match the predicted frequency when split by risk decides. This test will be used for Logistics regression model validation.

(2) AUC along with Model Accuracy will be used for model evaluation. Accuracy is calculated based on certain threshold where as AUC is overall performance evaluation of model as various points.AUC criteria will be given more weight age for model evaluation in this case.

Experimentations:

In this section experimentation will be carried out with the data by formulating three different types of models with three different approaches. Following are the three different approaches that will be used here-

-Model 1- This model will be created by using logit function of Generalized Logistics Model(GLM).

-Model 2: This model will be created by using Classification tree function.

-Model 3- This model will be created by using classification technique RandomForests model.

There are two data set given with the business case training and test set. Training set will be used to train the model and the test set will be used to evaluate the model performance.

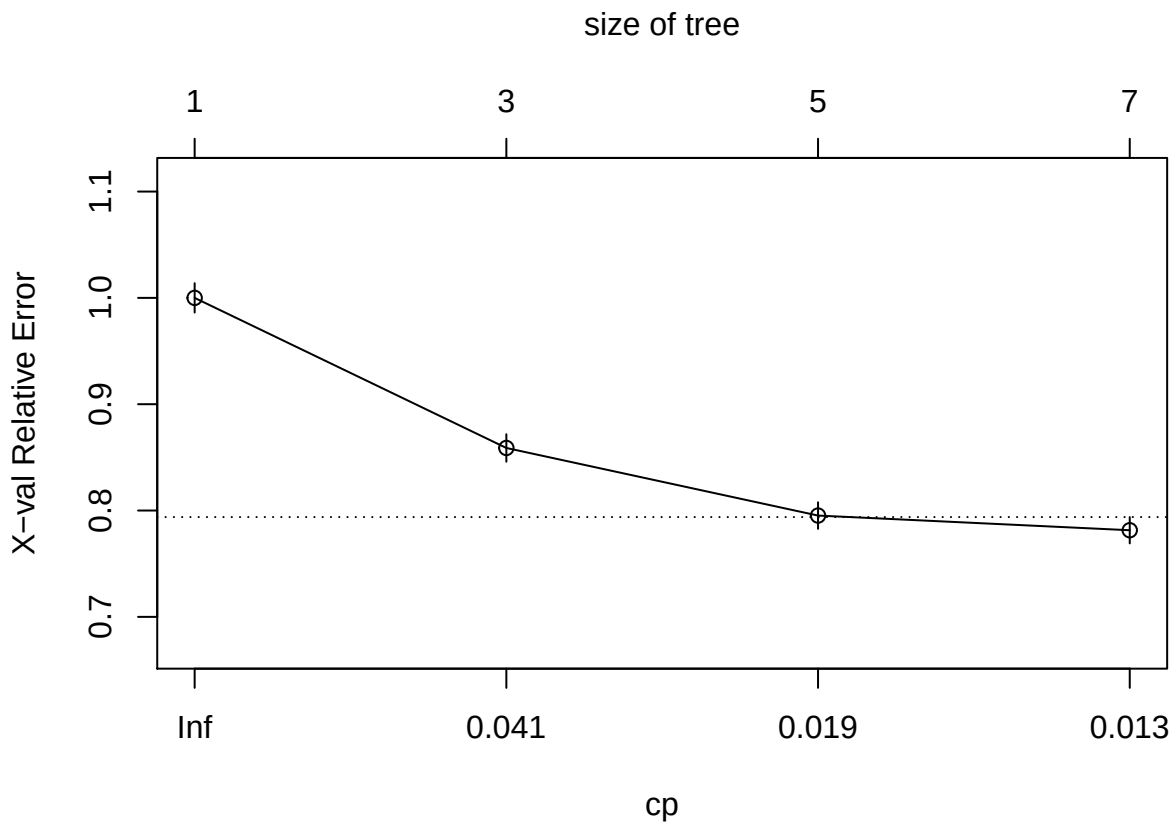
Logistics regression- Model 1:. Logistics regression function GLM has been used to classify the campaign response variable. Basic model generated by using GLM function has been enhanced by making necessary adjustments to non associated predictor variables shown as “NA” in basic model output. Next the model has been validated by using k=5 fold

cross validation press to do necessary adjustment to the model.

Interpretation from Logistics Regression model. There were total 10 iterations been performed before final selection of variables were made. AIC value from model 1 and model1_update(enhanced) model were same 13776. Hence removing variables from basic model does not help performance wise but reduced complexity with less degrees of freedom.By using k=5 cross validation, ($delta) error value came out to be low 0.06289177.

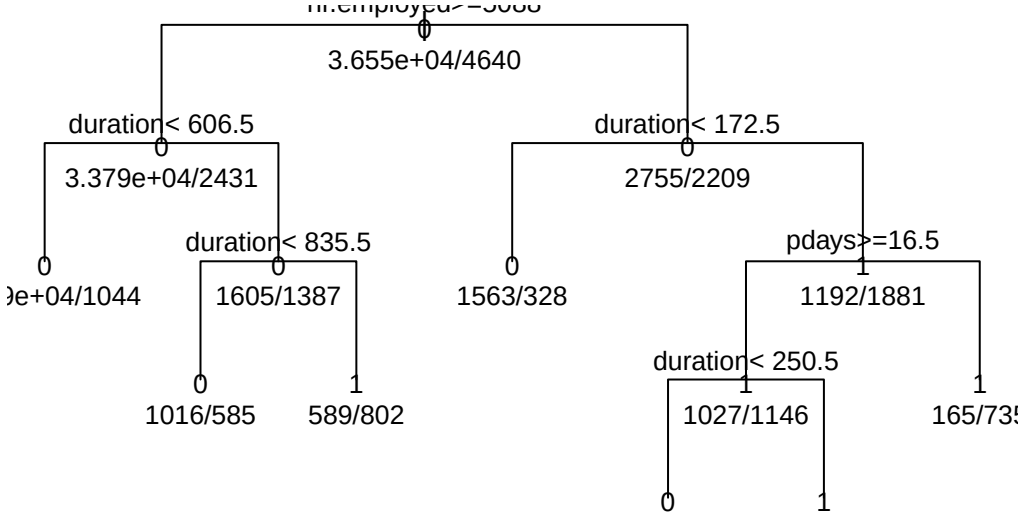
Table below provides details on significance of the variables and its odd ratio.

Classification Tree- Model 2. The basic idea of classification tree model is to predict a response variable y for the campaign from predictor variables. Model does this by growing a binary tree. At each node in the tree, a test is applied to one of the inputs. Depending on the outcome of the test two routes to be followed left or right. Eventually a leaf node is reached where a prediction is made about the binary outcome of campaign response. Model 2 has been rated using the Classification function from ROCR package.Basic model has been optimized using prune function.

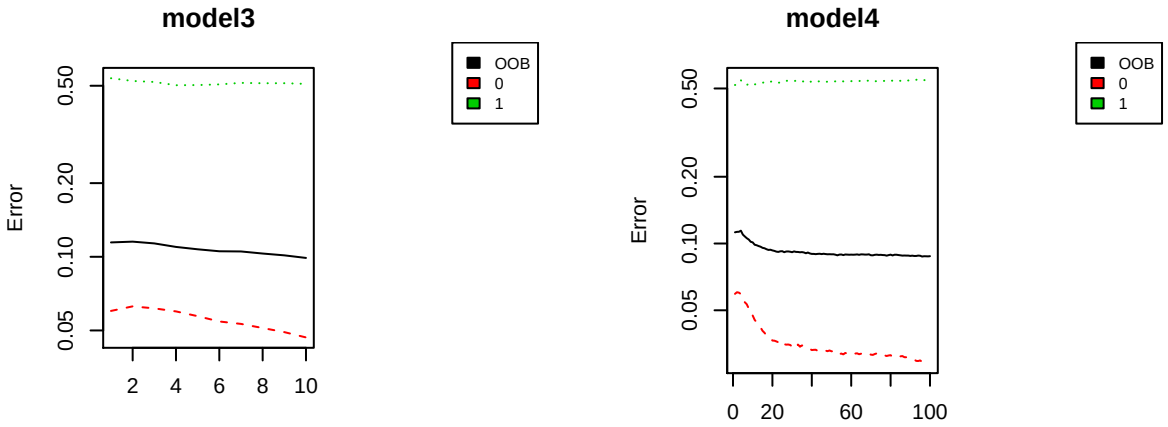


232

233 **Interpretation of Classification Tree Model.** Following are the most important
234 variables from this model-duration ,nr.employed ,euribor3m ,emp.var.rate, cons.conf.idx ,
235 cons.price.idx.Total 6 leafs(decision points) have been formed from this model. Complete
236 Classification tree is given below in the diagram.



RandomForest- Model 3. In Random Forests many classification trees are formed to classify campaign response variable y. Each tree creates separate set of classification, each tree is voted for performance for that classification. The forest chooses the classification having the most votes (over all the trees in the forest). One model will be created using this method with tree size 50. Then this model will be evaluated with a model of tree size 100.



Interpretation of RandomForest. From the chart above it can be seen that classification error rate to classify negative responses reduces with the increase in number of trees but there is no significant change in error rate for positive response.There is only slight reduction in error rate for negative responses when tree size is increased to 100 from 50.Number of variables tried at each split are 7 with negative classification rate of 0.03 and positive classification error rate of 0.51.

Below chart provides importance of various variables used in the model.

Results

Results from Regression Model

Table 2

Model 1 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	0.9142996	0.0857004	0.4323725	0.6678082	0.9331069	0.3607211	0.7029638

Results from Regression Model

Table 3

Model 1 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	0.9142996	0.0857004	0.4323725	0.6678082	0.9331069	0.3607211	0.7029638

From above table it can be seen Logistics Regression model has a very high accuracy rate of 91.42% when model was evaluated using the validation data set. Though the AUC value for this model was comparatively lower 0.702 which indicates not good fitment of the model.

`Hosmer and Lemeshow goodness of fit (GOF) test`

data: model1_update$y, fitted(m) X-squared = 14.926, df = 8, p-value = 0.0606

By using Hosmer-Lemeshow goodness-of-fit (GOF) tests when model was evaluated p value came to be greater than 0.05. With this test if the p value is lower than 0.05 model is rejected and if it's high, then the model passes the test. Regression model passed this test.

Results from Classification Tree- Model 2

Table 4

Model 1 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
2	0.918184	0.081816	0.5343681	0.6548913	0.9440149	0.4377405	0.8650875

It can be seen from the table above, this model 2 has also very high accuracy rate of 91.81% which is very good.This model has AUC value of 0.865 which seem to be inline with given high accuracy.

Results from RandomForest Model

Table 5

Model 1 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
3	0.97815	0.02185	0.8359202	0.9592875	0.9801396	0.8269327	0.9157791

The model created using Randomforest has accuracy of 98.64% which is extraordinary results and give rise to suspicion model is able to separate out the classification based on certain variable. When we looked at the importance of variable "duration" it becomes apparent that this variable is being used in a big way to classify response accurately. It can be seen that this model also shows the similar kind of trend in classification of data in earlier stages with very stiff line till true positive rate of 0.4 and then sharp increase in false positive rate.

5 Discussion and Conclusions:

Table 6

Comparison of 3 Model3

	Model_No	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	GLM_model1	0.9142996	0.0857004	0.4323725	0.6678082	0.9331069	0.3607211	0.7029638
2	CRT_model2	0.9181840	0.0818160	0.5343681	0.6548913	0.9440149	0.4377405	0.8650875
3	RF_model3	0.9781500	0.0218500	0.8359202	0.9592875	0.9801396	0.8269327	0.9157791

$ Final model selection:

Based on the Accuracy of the model, model 1 and model 2 are very close around 91% accuracy with probability threshold of 0.5. Model 3 has much higher value of 98%. But Accuracy is not always the key criteria for a model as Accuracy is calculated based on a

defined threshold. Also due to imbalance of data o 10% to 90% distribution of response variable forced to choose the model based on other criteria. Model Based on AUC value is model 3 having AUC value of 0.9398 which is a very good score. Model 3 stands out among the three models.

$ Key predictor variables:

For all three models it is found variables “duration” is most important variables by far. This variable has positive impact in campaign outcome. This could be due to the fact that longer the Customer stays on phone more productive conversation is taking place to get the Customer start their term deposit Account. “euribor3m” is most important variable which denotes inter bank interest rate in Eurozone. Term deposit interest rates are generally interlinked and tends to go up together. This variable has positive impact on response variable. Predictor “nr.employed” denotes number of employees for the bank. This variable also has positive impact on campaign response. More the number of employees more visible the bank is and in turn more customers it gets through the campaign.

Among the negative variables “emp.var.rate” has negative impact on response. As negative rate of this variable indicates issues with economy and lower economic activities. That in turn could impact the savings rate and people tend to use their savings that time.

$ Shortcomings

Imbalance of response variable only 10% of population was the main shortcomings that we have in the model creation. This issue has been addressed partially by using Area Under Curve as the criteria for model selection.

$ Final Recommendation

In conclusion it can be suggested to the bank management that focus should be given in hiring more people, doing more quality phone calls. Also to time the campaign in a stable macroeconomic environment to get better return on investment from this campaign.

References

be sure to cite all references used in the report (APA format). We used R (3.2.2, R Core Team, 2016) and the R-packages *papaja* (0.1.0.9054, Aust & Barth, 2015), *papaja* (0.1.0.9054, Aust & Barth, 2015), *Amelia* (1.7.4, Honaker, King, & Blackwell, 2011), *aod* (1.3, Lesnoff, M., Lancelot, & R., 2012), *AUC* (0.3.0, Ballings & Poel, 2013), *dplyr* (0.4.3, H. Wickham & Francois, 2015), *faraway* (1.0.7, Faraway, 2016), *gdata* (2.17.0, Warnes et al., 2015), *ggplot2* (2.0.0, H. Wickham, 2009), *gplots* (3.0.1, Warnes et al., 2016), *gridExtra* (2.2.1, Auguie, 2016), *ISLR* (1.0, James, Witten, Hastie, & Tibshirani, 2013), *knitr* (1.12.3, Xie, 2015), *lattice* (0.20.33, Sarkar, 2008), *leaps* (2.9, Fortran code by Alan Miller, 2009), *MASS* (7.3.43, W. N. Venables & Ripley, 2002), *popbio* (2.4.3, Stubben & Milligan, 2007), *pscl* (1.4.9, Zeileis, Kleiber, & Jackman, 2008), *psych* (1.5.8, Revelle, 2015), *Rcpp* (0.12.5, Eddelbuettel & François, 2011), *reshape* (0.8.5, Wickham & Hadley, 2007), *ROCR* (1.0.7, Sing, Sander, Beerenwinkel, & Lengauer, 2005), *stringr* (1.0.0, H. Wickham, 2015), and *xtable* (1.8.2, Dahl, 2016) for all our analyses.

Appendix

Supplemental tables and/or figures. R statistical programming code.

```
{r
code=readLines(knitr::purl('https://raw.githubusercontent.com/kishkp/data621-ctg5/mast
                documentation = 0)), eval = FALSE} #
```

6.1 Data Analysis details

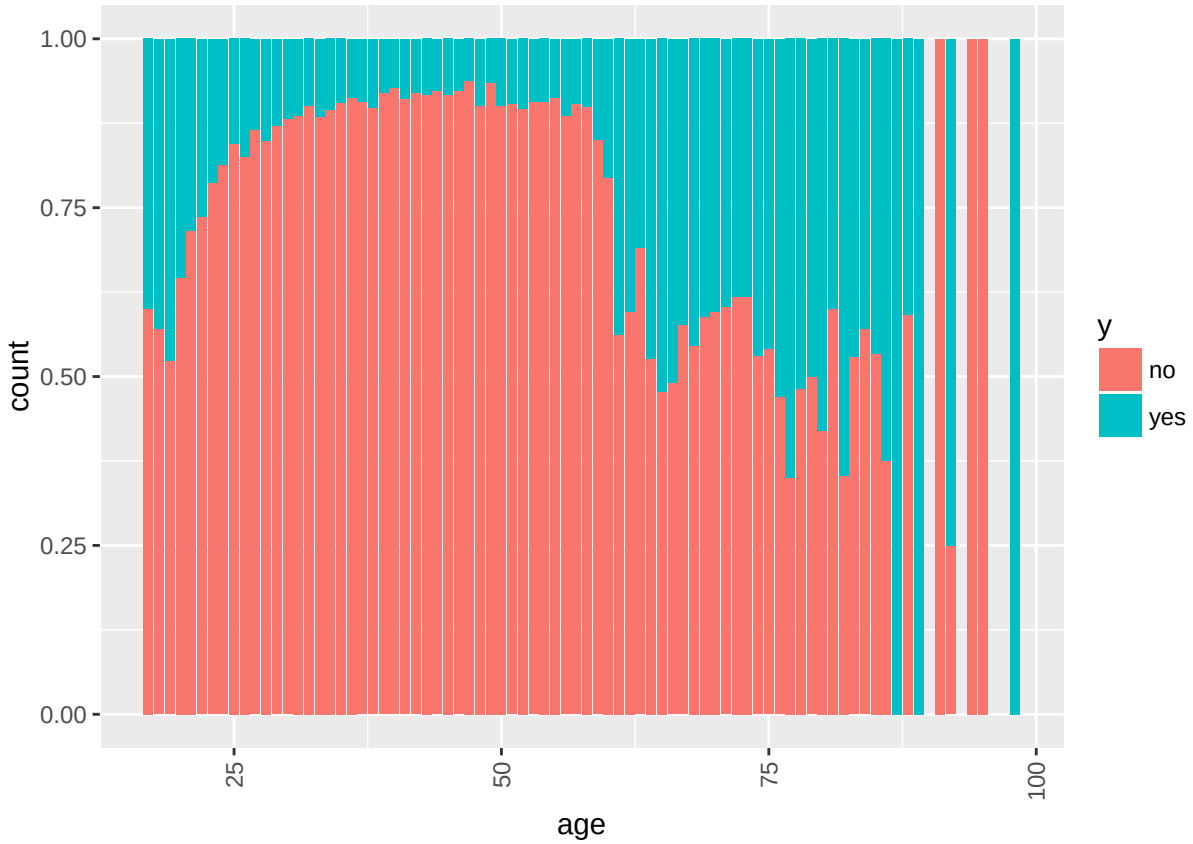
6.1.1 Variable Description

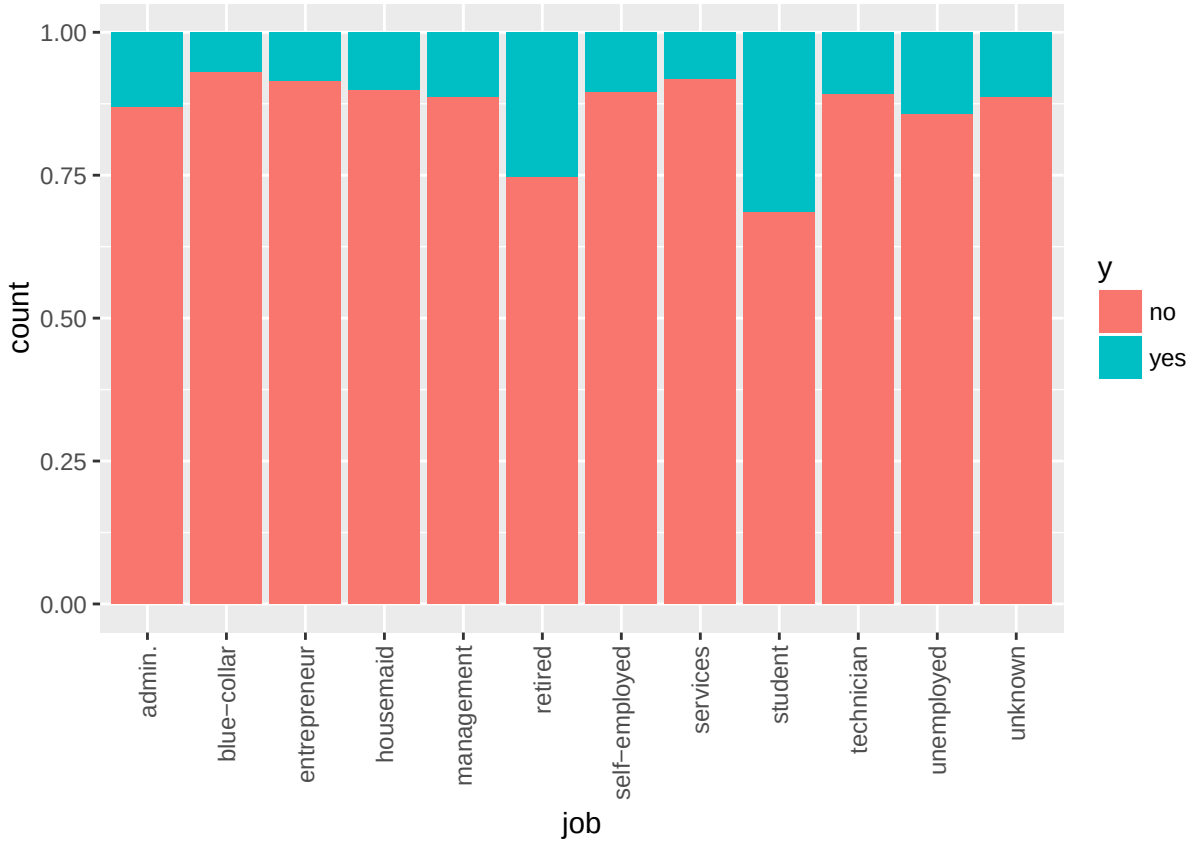
Table 7

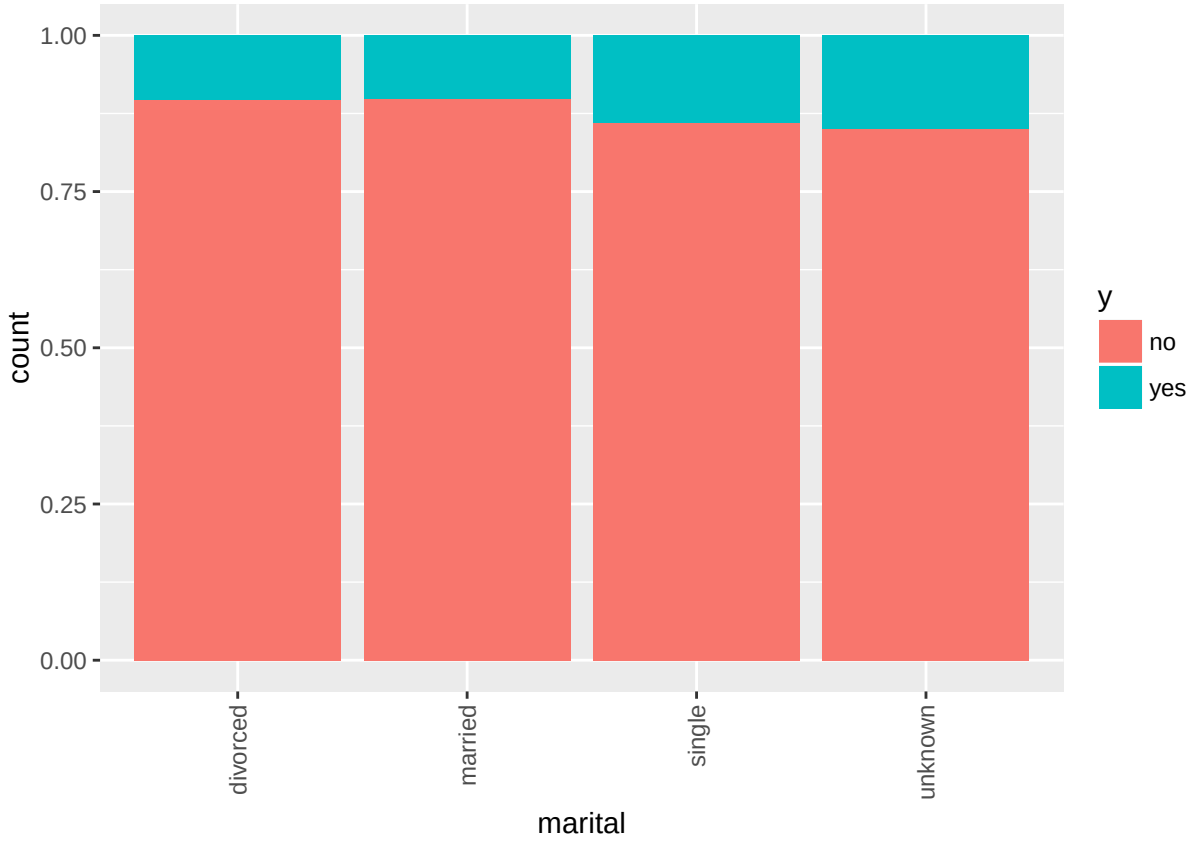
Variable Description

Variable	Data.Type	Type	Description
age	Numeric	Predictor	Client's age
job	Catagorical	Predictor	Client's job
marital	Catagorical	Predictor	Client's marital status
education	Catagorical	Predictor	Client's education level
default	Binary	Predictor	Credit in default?
balance	Numeric	Predictor	Client's average yearly balance, in euros
housing	Binary	Predictor	Client has housing loan?
loan	Binary	Predictor	Client has personal loan?
contact	Catagorical	Predictor	Client's contact communication type
day	Catagorical	Predictor	Client last contact day of the month
month	Catagorical	Predictor	Client last contact month of year
duration	Numeric	Predictor	Client last contact duration, in seconds
campaign	Numeric	Predictor	Client number of contacts performed during this campaign
pdays	Numeric	Predictor	Client days that passed after first contact
previous	Numeric	Predictor	Number of contacts performed before this campaign
poutcome	Catagorical	Predictor	Outcome of the previous marketing campaign
emp.var.rate	Numeric	Predictor	Quarterly employment variation rate
cons.price.idx	Numeric	Predictor	Monthly consumer price index
cons.conf.idx	Numeric	Predictor	Monthly consumer confidence index
euribor3m	Numeric	Predictor	Daily euribor 3 month rate
nr.employed	Numeric	Predictor	Quarterly number of employees
y	Binary	Response	Has the client subscribed a term deposit?

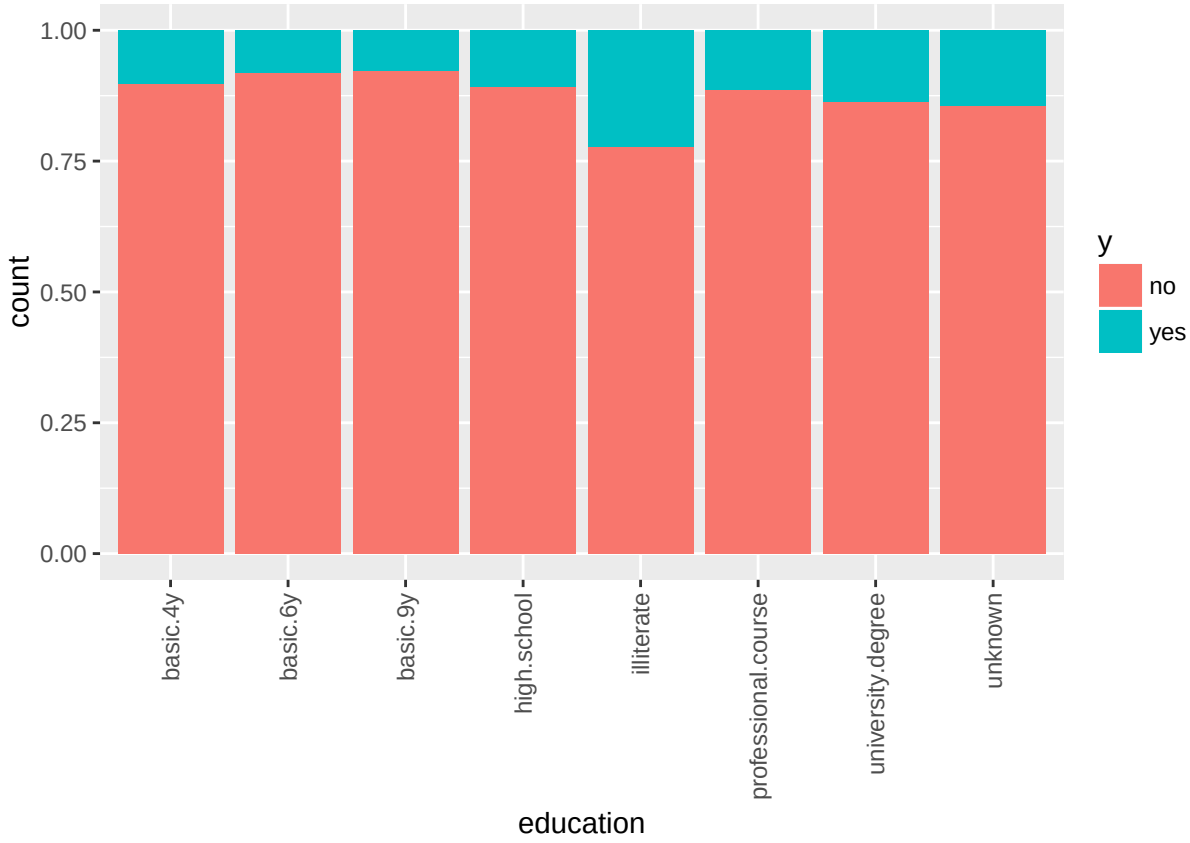
6.1.2 Predictor and Response variable Association



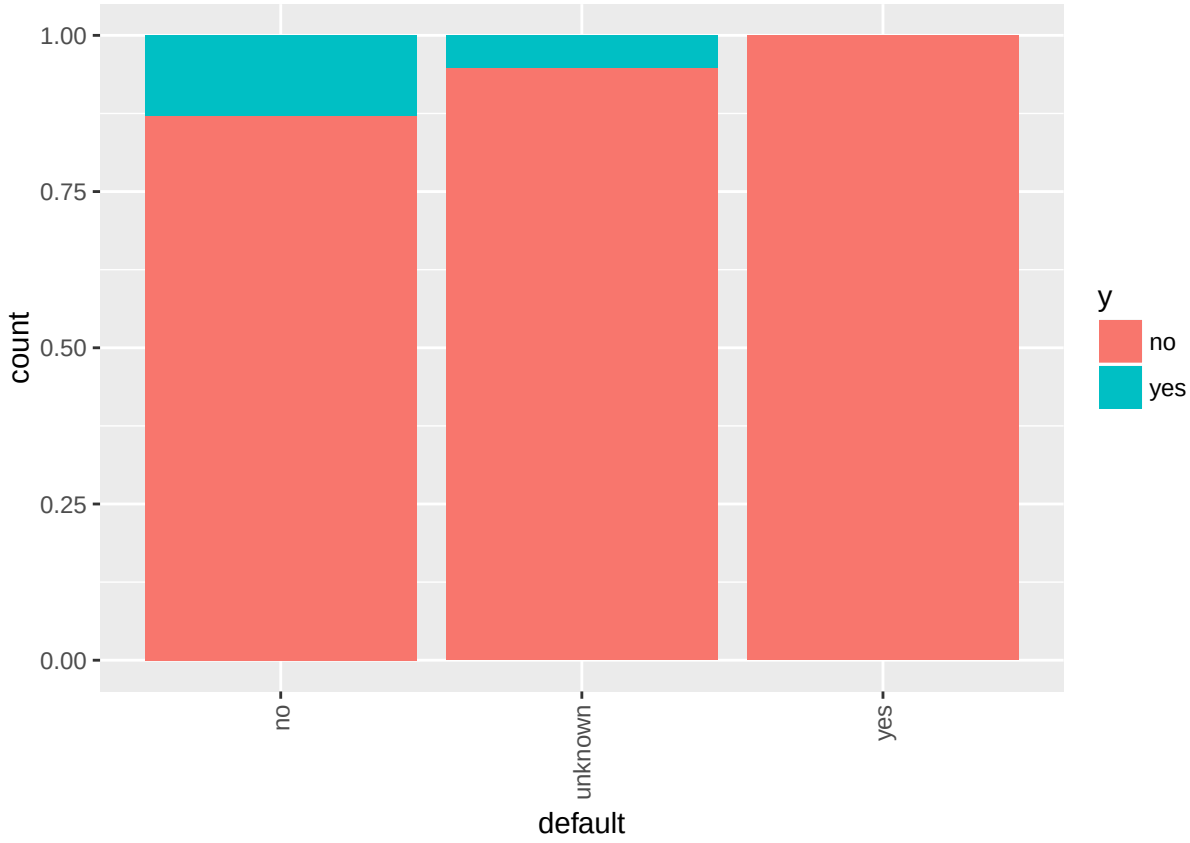



1.00
0.75
0.50
0.25
0.00
count
divorced
married
single
unknown
marital
y
no
yes

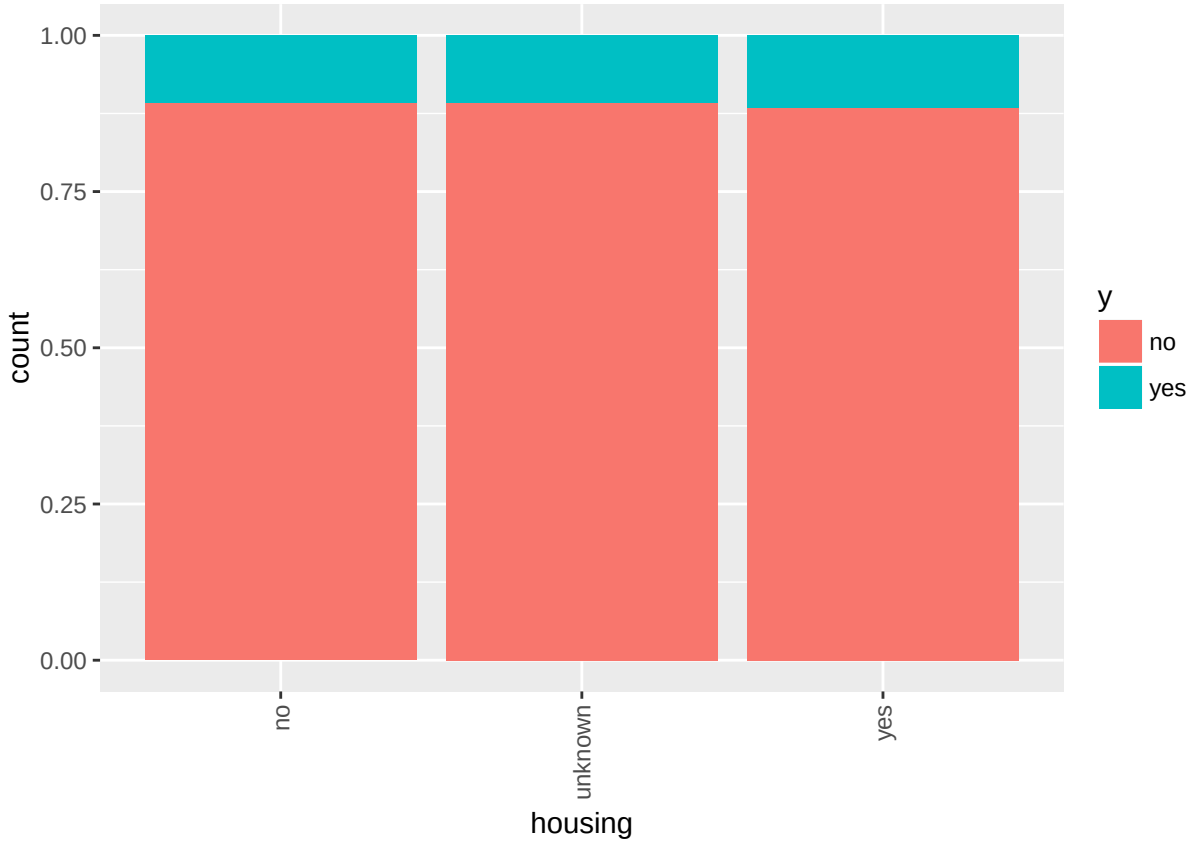
330


1.00
0.75
0.50
0.25
0.00
count
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
education
y
no
yes

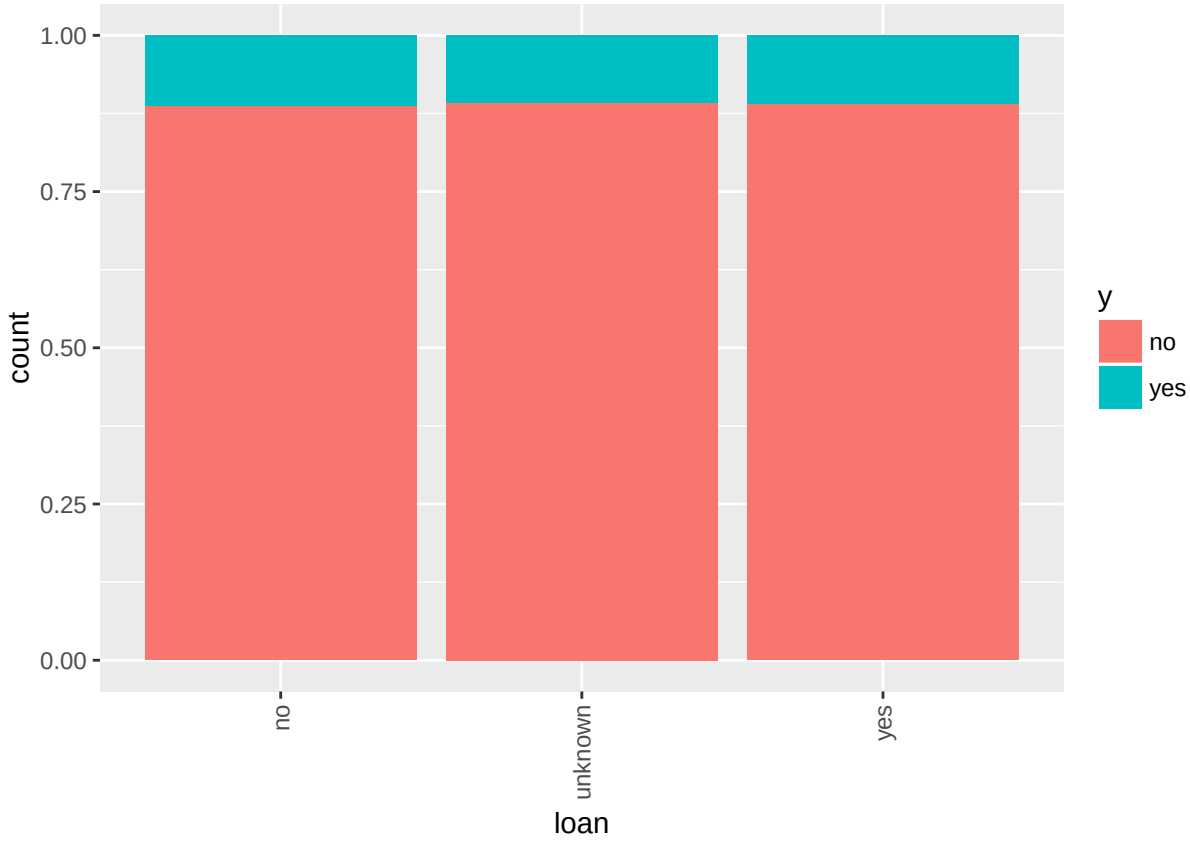
331



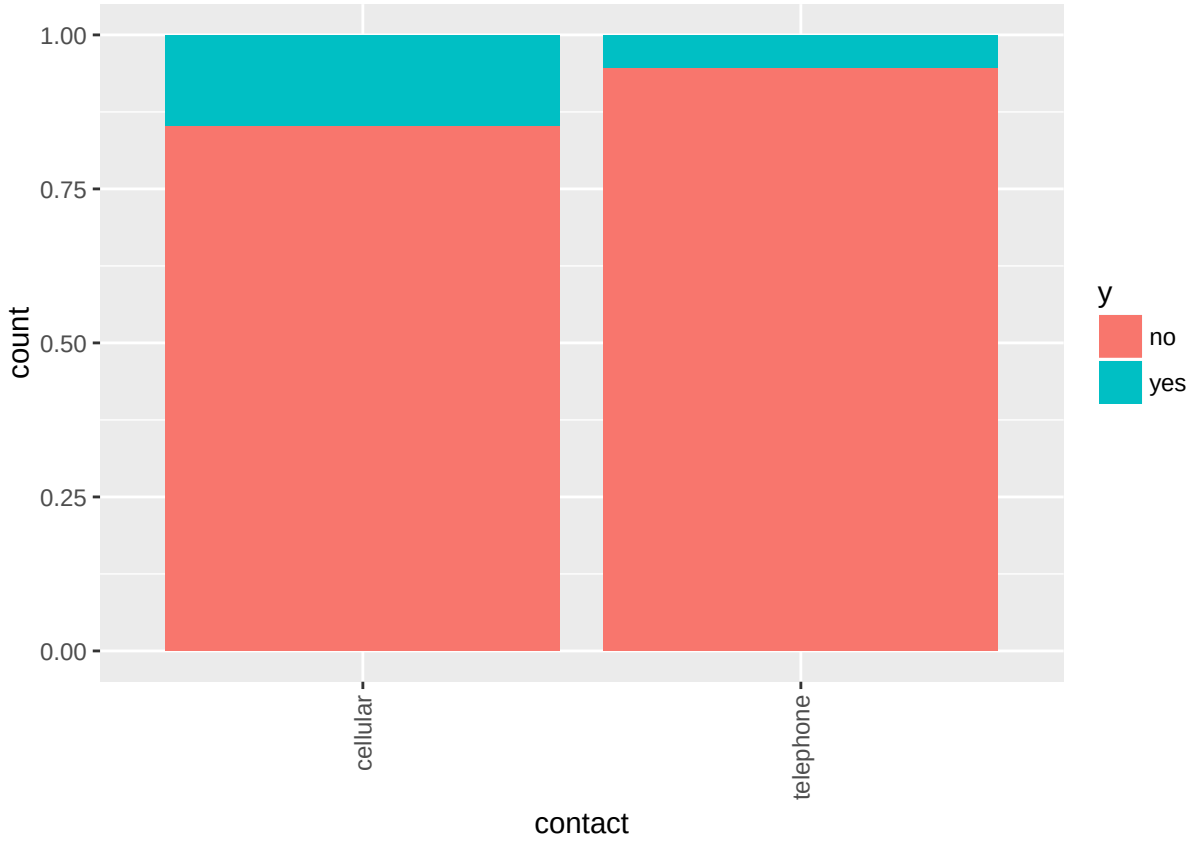
332



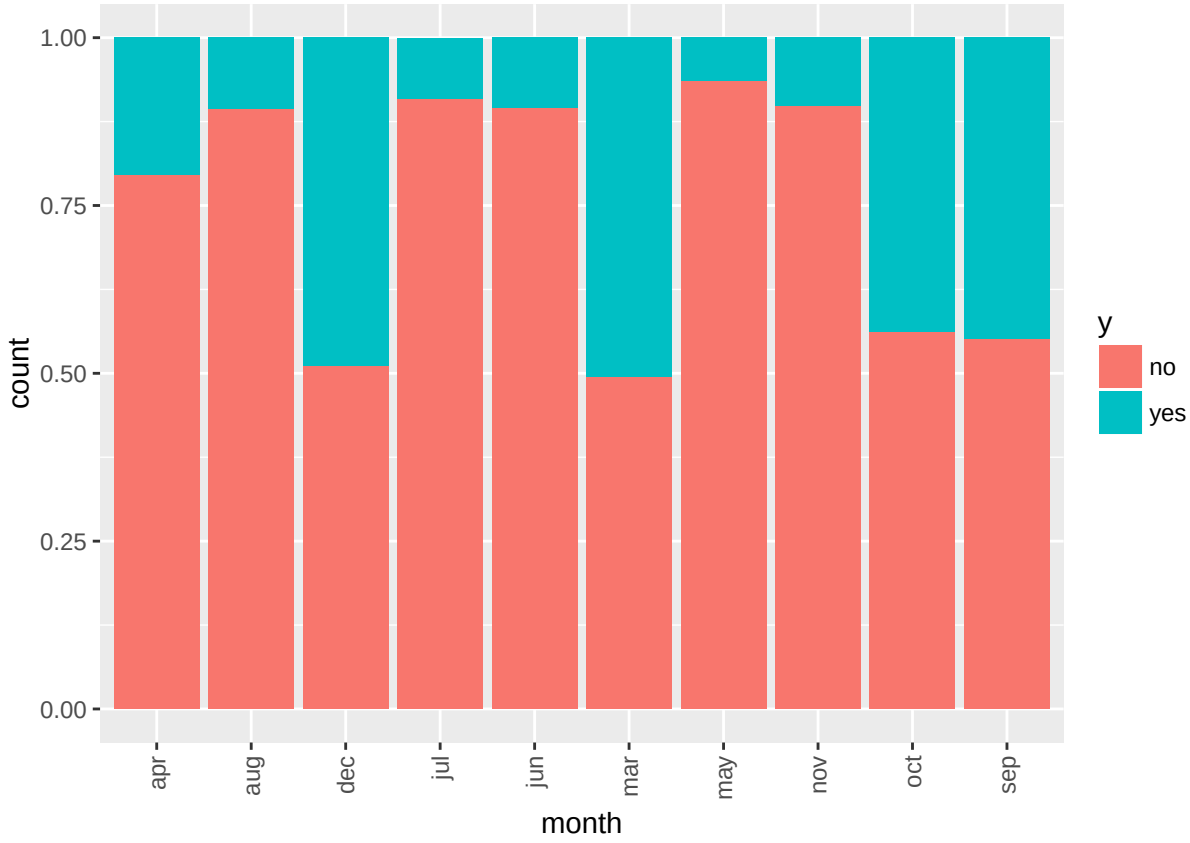
333



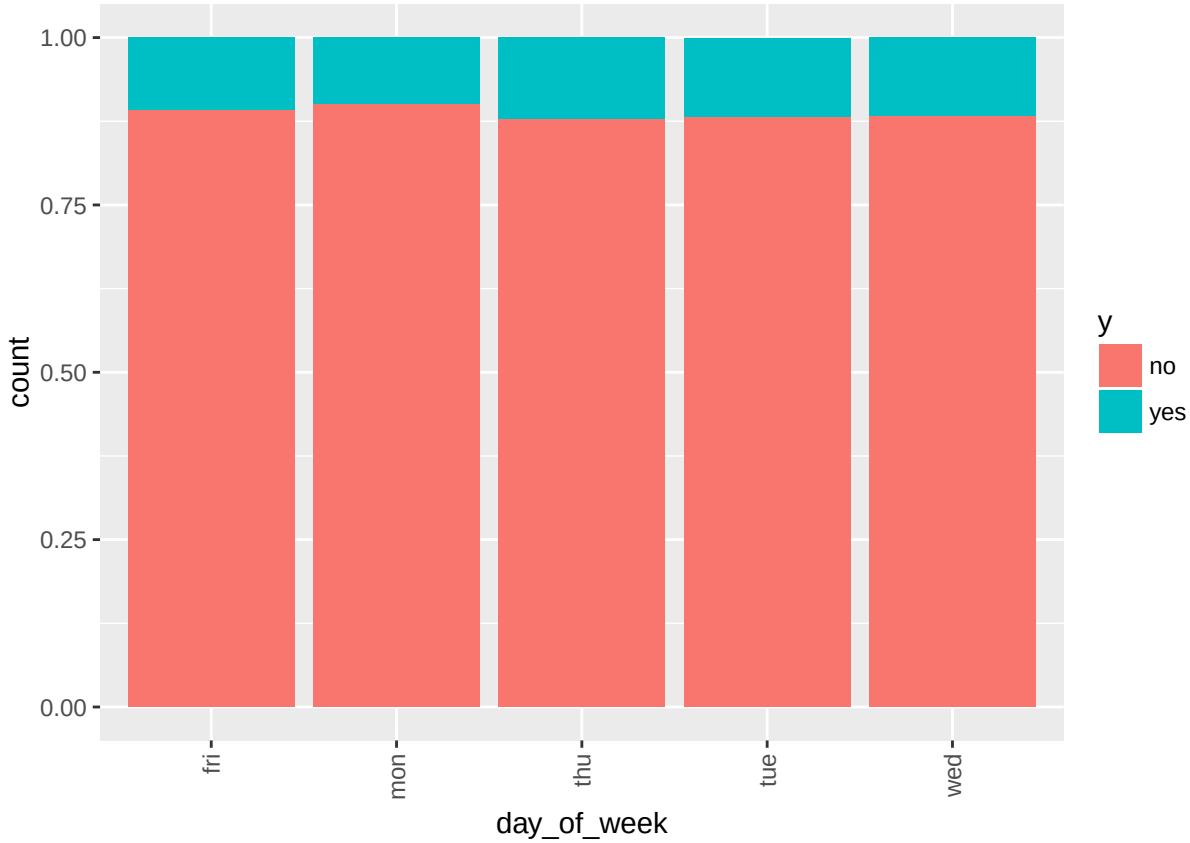
334


1.00
0.75
0.50
0.25
0.00
count
cellular
telephone
contact
y
no
yes

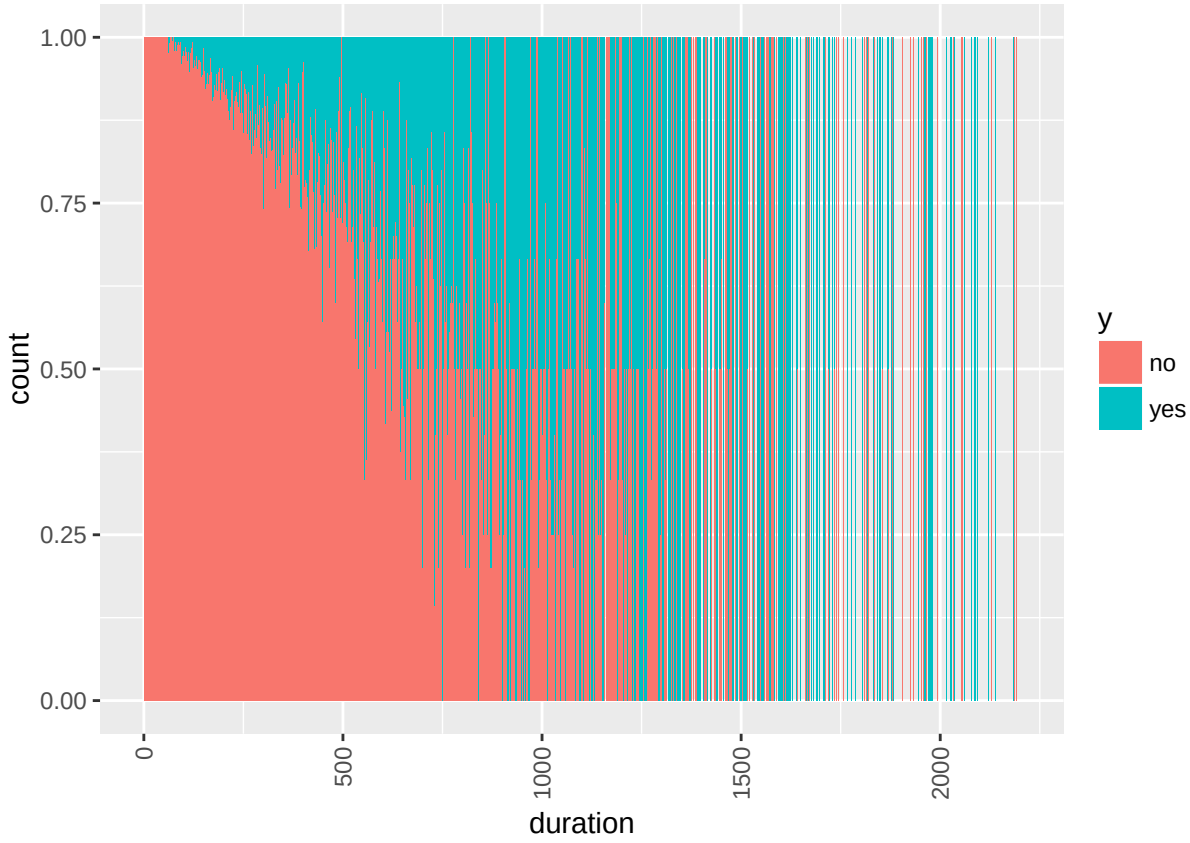
335


1.00
0.75
0.50
0.25
0.00
count
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
month
y
no
yes

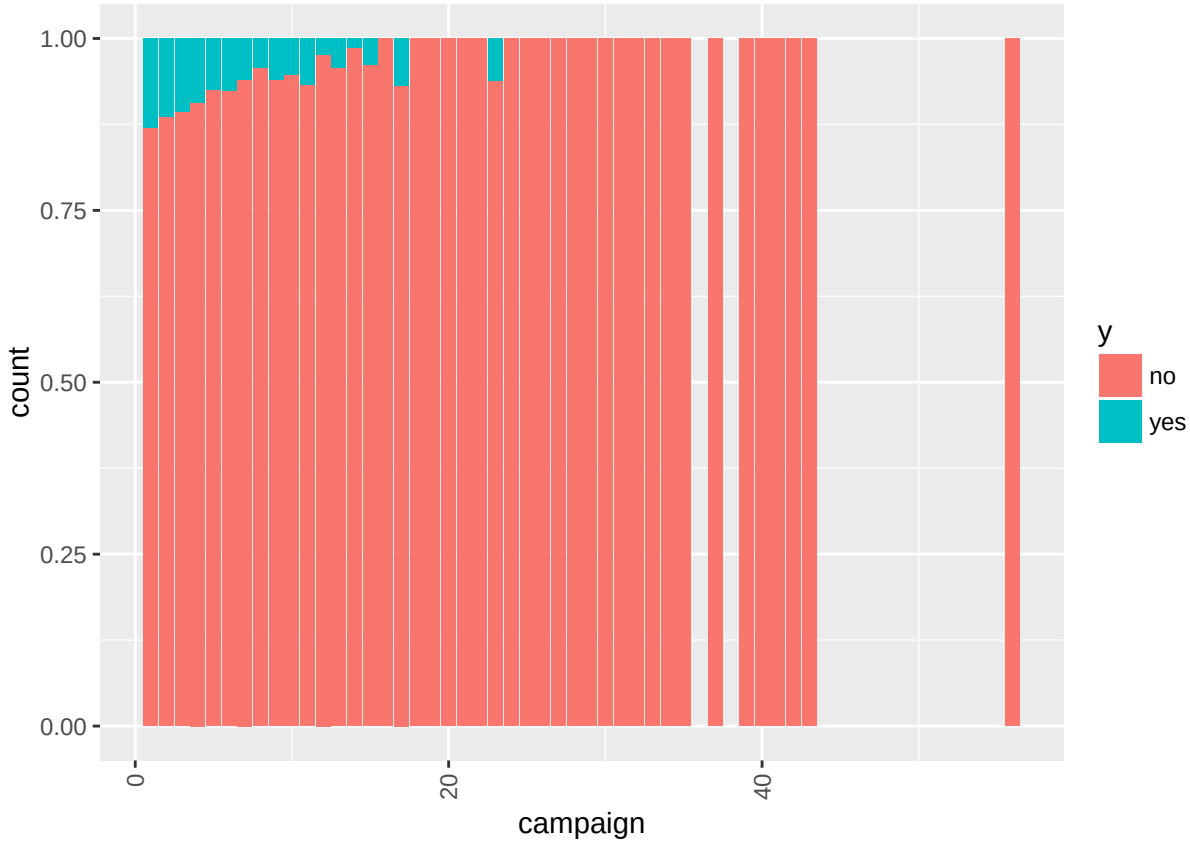
336


1.00
0.75
0.50
0.25
0.00
count
fri
mon
thu
tue
wed
day_of_week
y
no
yes

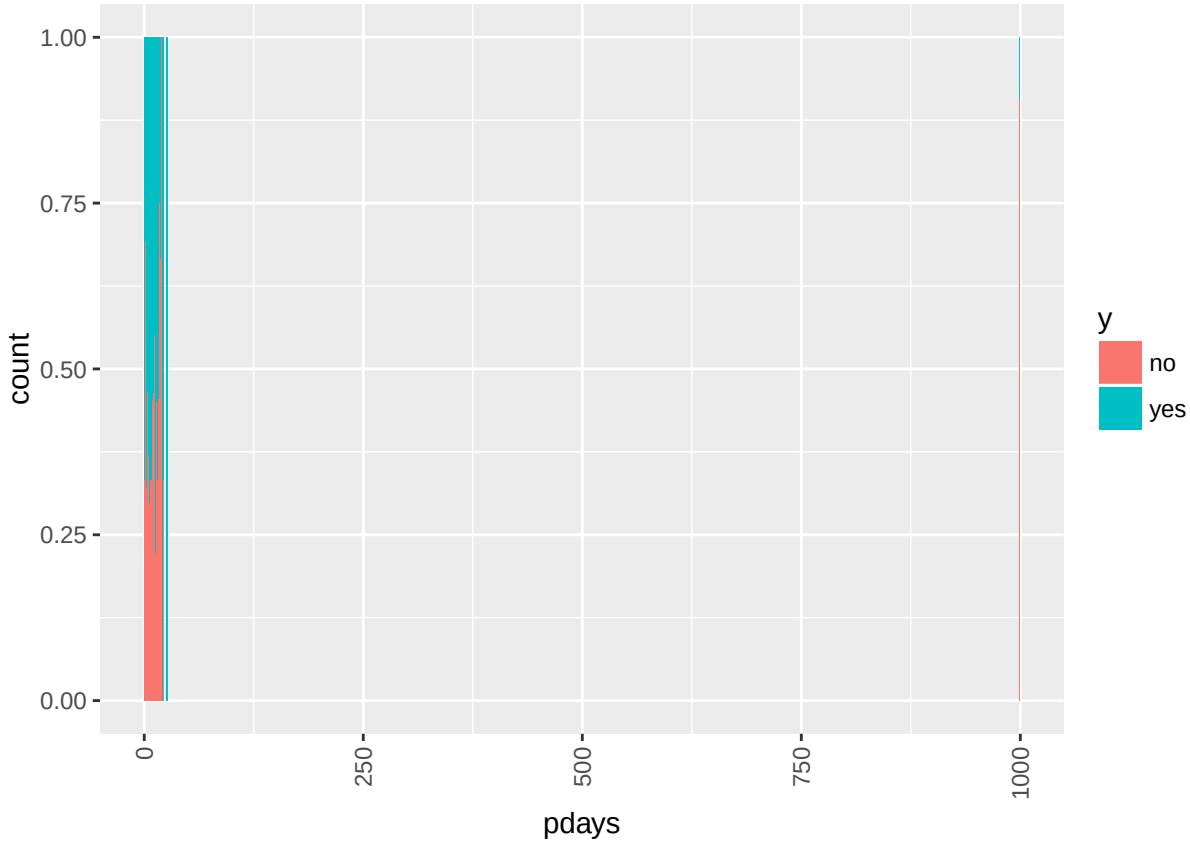
337


1.00
0.75
0.50
0.25
0.00
count
0
500
1000
1500
2000
duration
y
no
yes

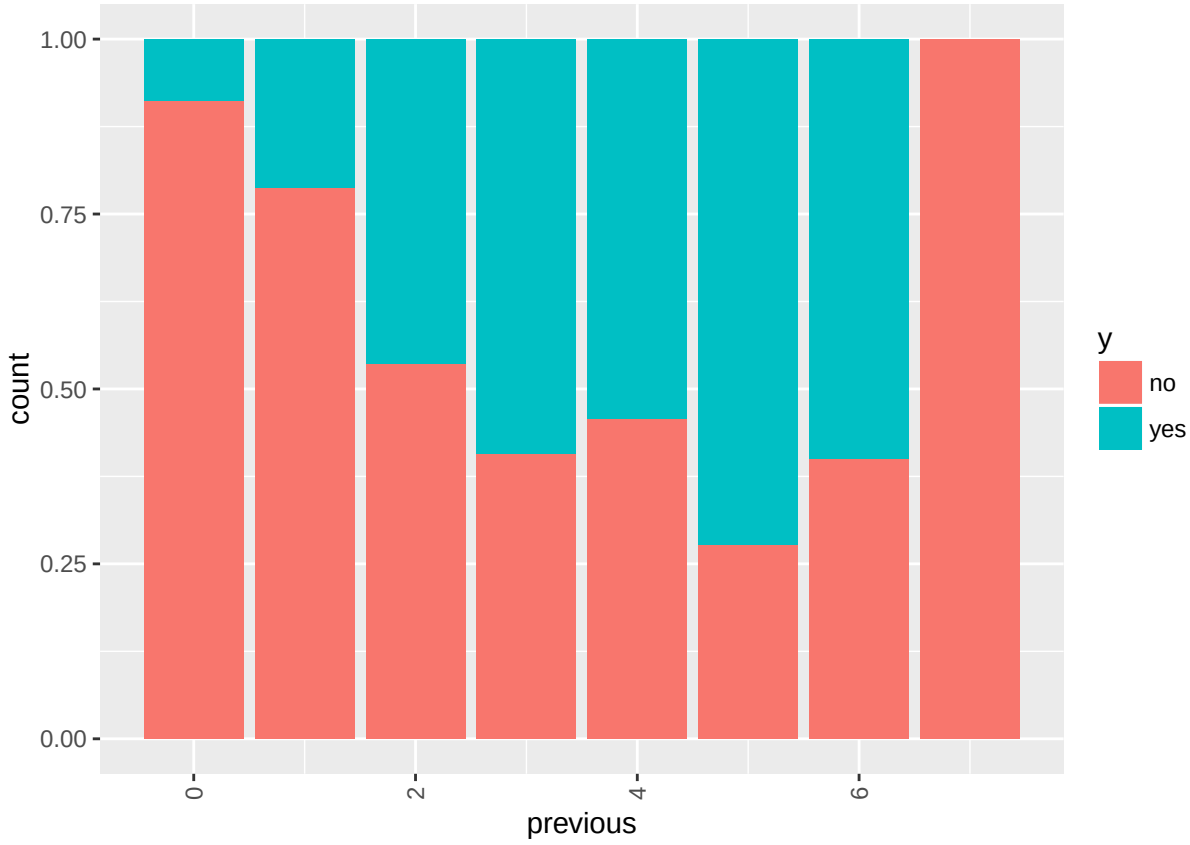
338



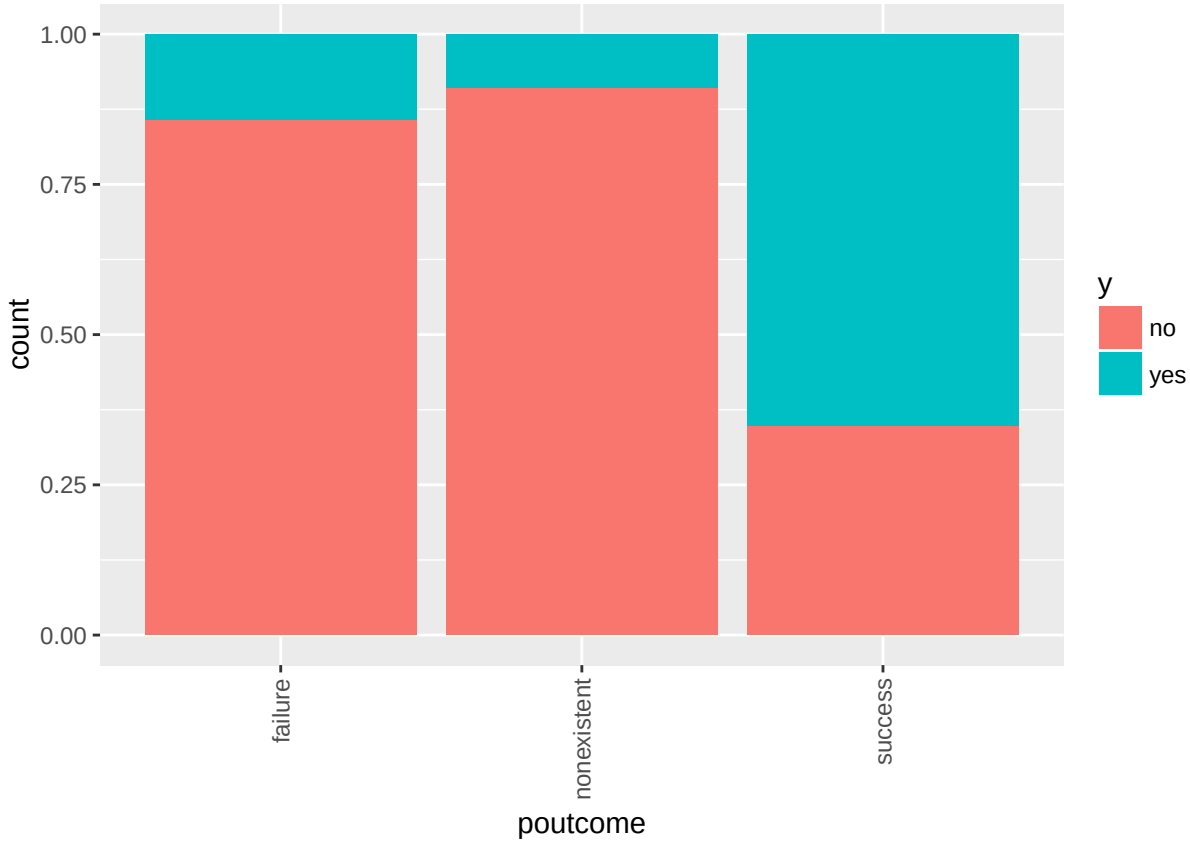
339



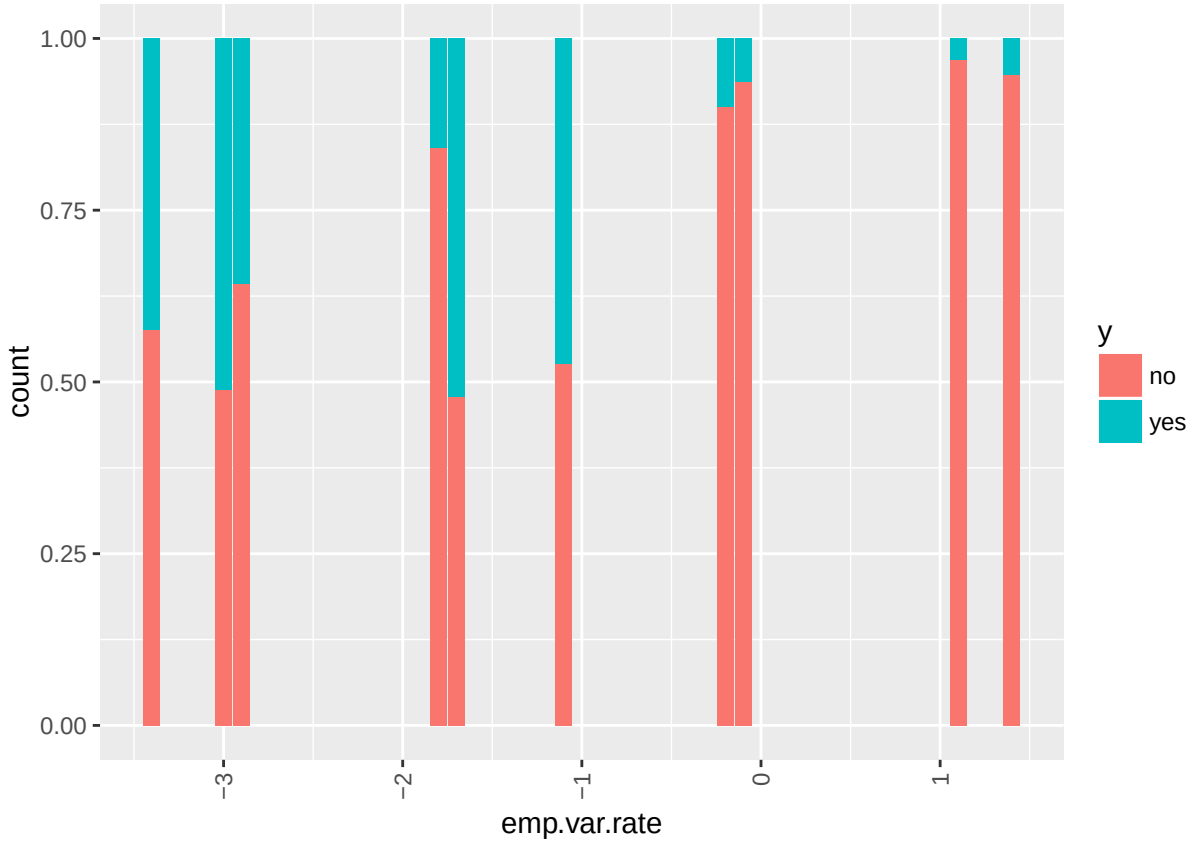
340



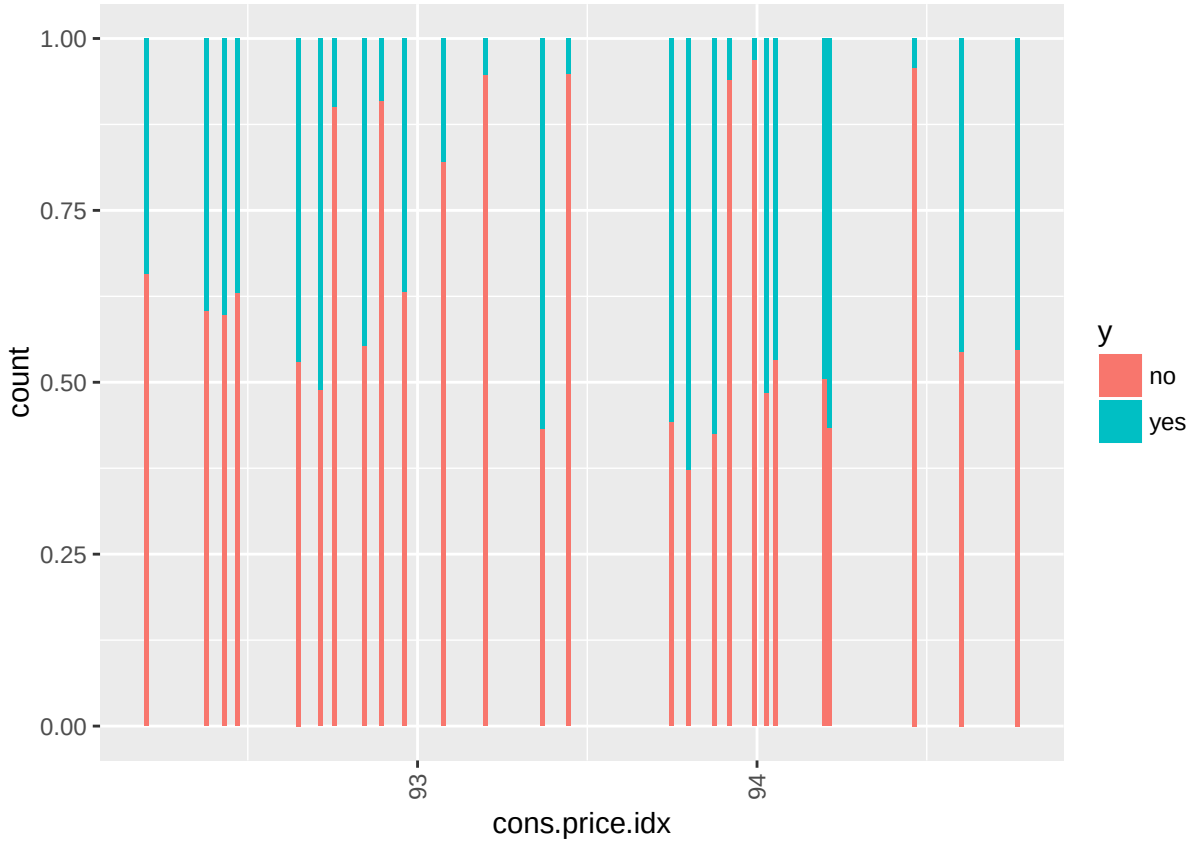
341


1.00
0.75
0.50
0.25
0.00
count
failure
nonexistent
success
poutcome
y
no
yes

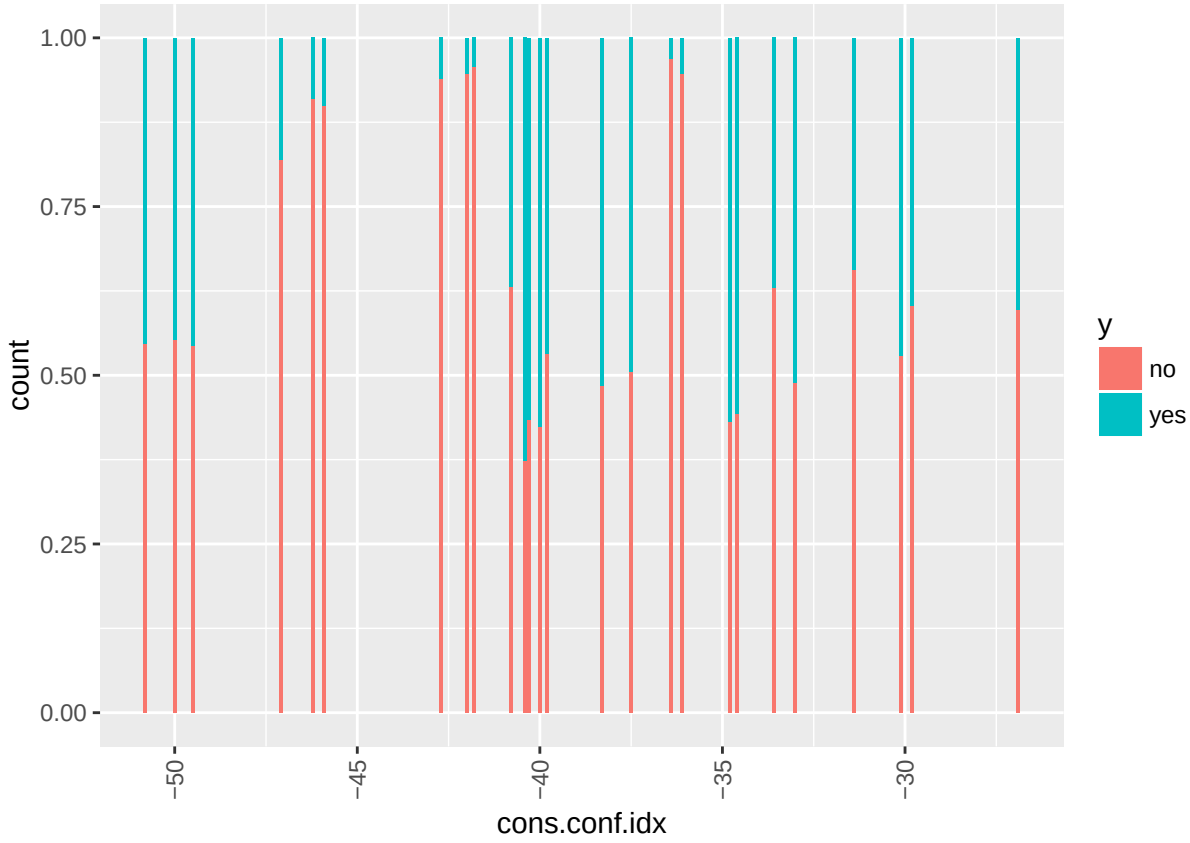
342



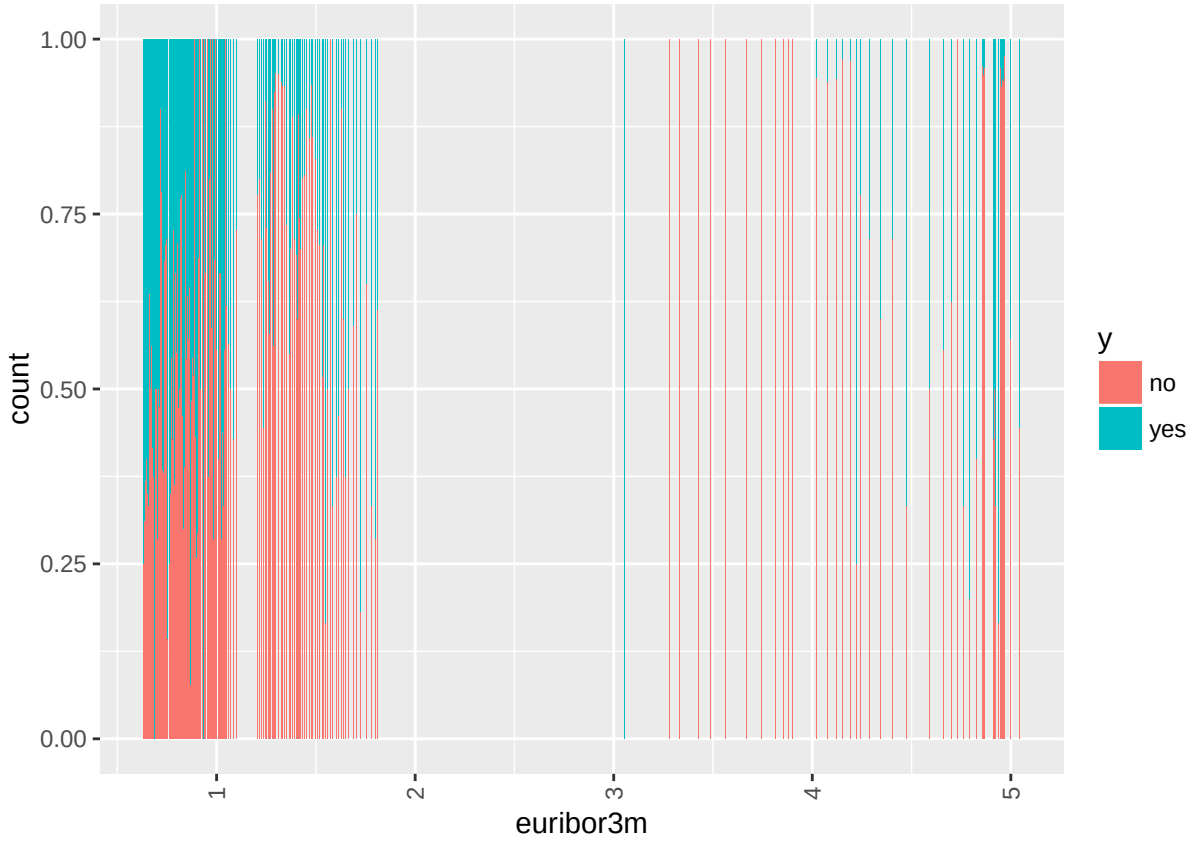
343



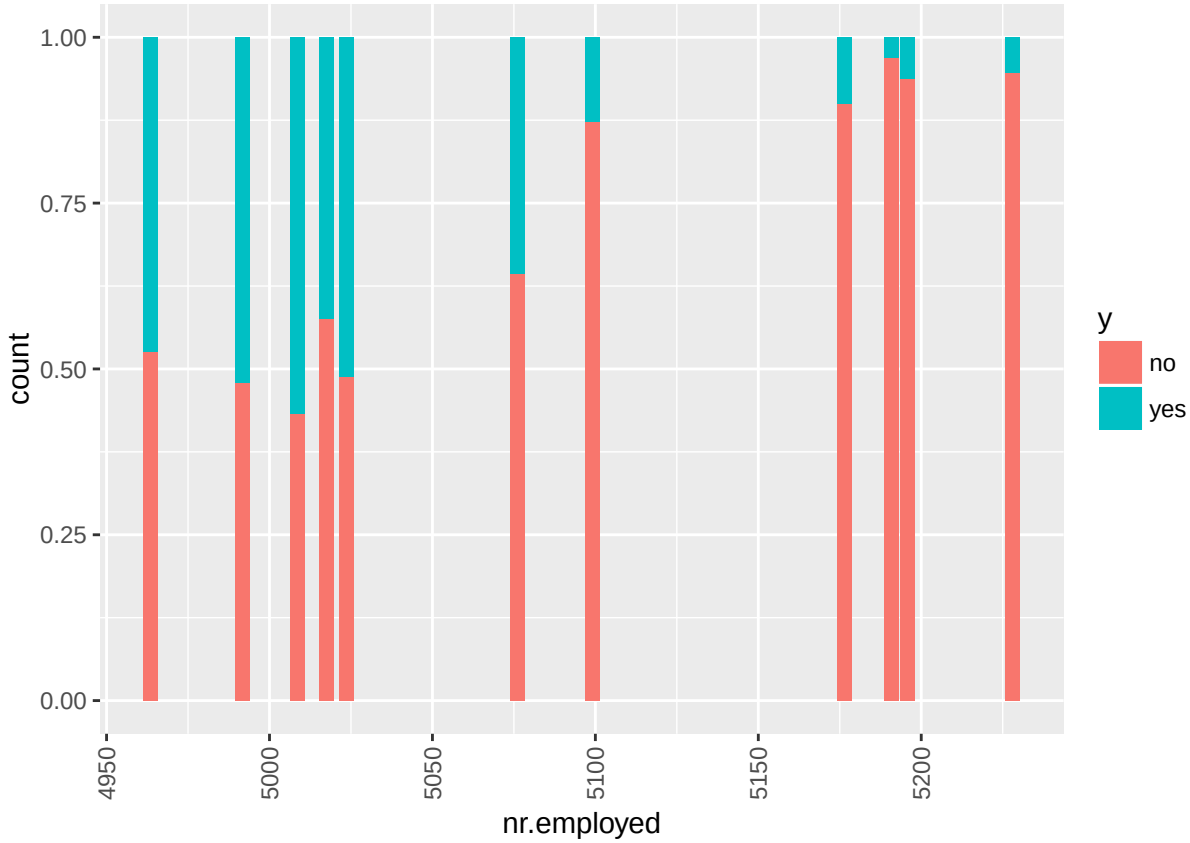
344



345



346



6.1.3 Unique Value & Missing value

We see that there are no missing values in our dataset as shown in table 2 and graph format. The unique values are given in the table

Table 8

Missing Values

	Missing Values
age	0
job	0
marital	0
education	0
default	0
housing	0

	Missing Values
loan	0
contact	0
month	0
day_of_week	0
duration	0
campaign	0
pdays	0
previous	0
poutcome	0
emp.var.rate	0
cons.price.idx	0
cons.conf.idx	0
euribor3m	0
nr.employed	0
y	0

Table 9

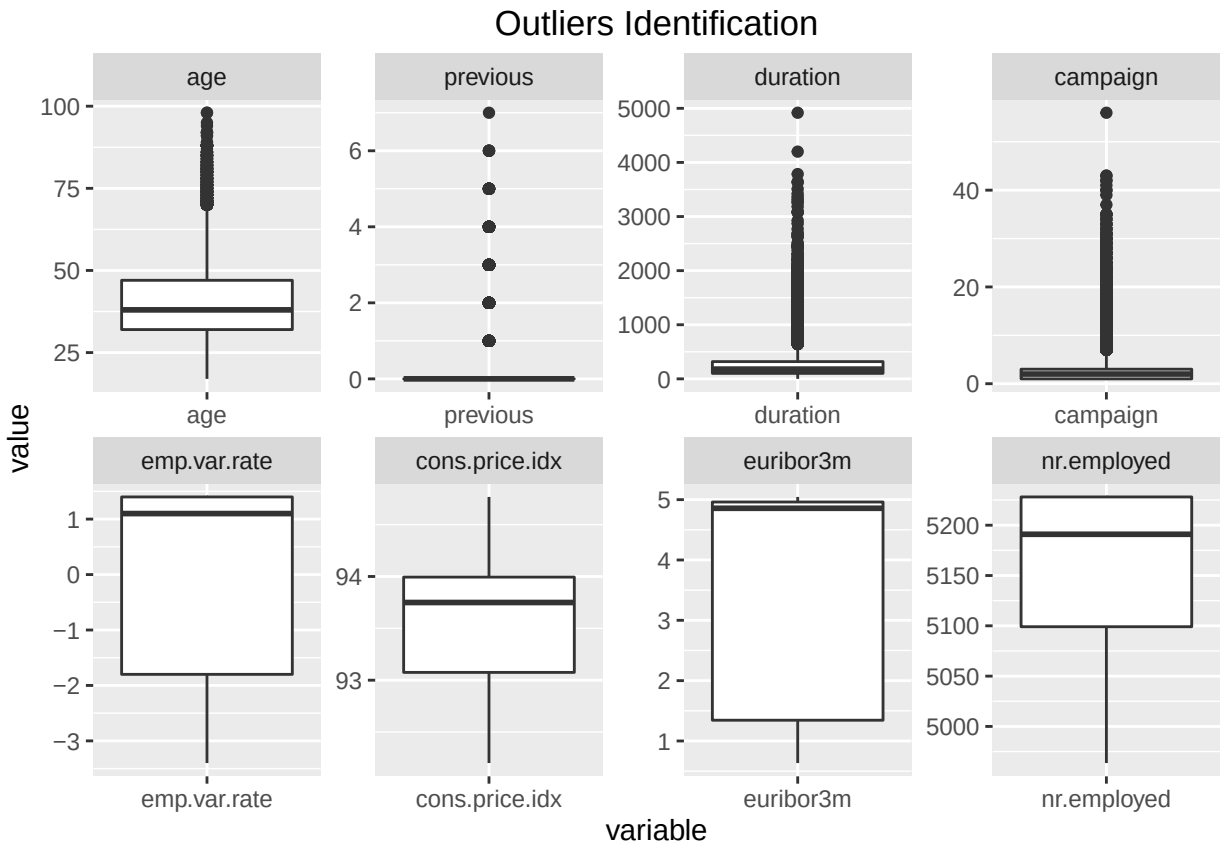
Unique Values

	Unique Values
age	78
job	12
marital	4
education	8
default	3
housing	3

	Unique Values
loan	3
contact	2
month	10
day_of_week	5
duration	1544
campaign	42
pdays	27
previous	8
poutcome	3
emp.var.rate	10
cons.price.idx	26
cons.conf.idx	26
euribor3m	316
nr.employed	11
y	2

6.1.4 Data Summary post conversion

6.1.5 Outliers Analysis



6.1.6 Analysis of link functions for given variables

Auguie, B. (2016). *GridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from http://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2015). *Papaja: Create aPA manuscripts with rMarkdown*. Retrieved from https://github.com/crsh/papaja

Ballings, M., & Poel, D. V. den. (2013). *AUC: Threshold independent performance measures for probabilistic classifiers*. Retrieved from

http://CRAN.R-project.org/package=AUC

Dahl, D. B. (2016). *Xtable: Export tables to laTeX or hTML*. Retrieved from http://CRAN.R-project.org/package=xtable

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. Retrieved from http://www.jstatsoft.org/v40/i08/

Faraway, J. (2016). *Faraway: Functions and datasets for books by julian faraway*. Retrieved from http://CRAN.R-project.org/package=faraway

Fortran code by Alan Miller, T. L. using. (2009). *Leaps: Regression subset selection*. Retrieved from http://CRAN.R-project.org/package=leaps

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47. Retrieved from http://www.jstatsoft.org/v45/i07/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *ISLR: Data for an introduction to statistical learning with applications in r*. Retrieved from http://CRAN.R-project.org/package=ISLR

Lesnoff, M., Lancelot, & R. (2012). *Aod: Analysis of overdispersed data*. Retrieved from http://cran.r-project.org/package=aod

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2015). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from

http://CRAN.R-project.org/package=psych

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in r. *Bioinformatics*, *21*(20), 7881. Retrieved from http://rocr.bioinf.mpi-sb.mpg.de

Stubben, C. J., & Milligan, B. G. (2007). Estimating and analyzing demographic models using the popbio package in r. *Journal of Statistical Software*, *22*(11).

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., ... Venables, B. (2016). *Gplots: Various r programming tools for plotting data*. Retrieved from http://CRAN.R-project.org/package=gplots

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., ... others. (2015). *Gdata: Various r programming tools for data manipulation*. Retrieved from http://CRAN.R-project.org/package=gdata

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://had.co.nz/ggplot2/book

Wickham, H. (2015). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from http://CRAN.R-project.org/package=stringr

Wickham, H., & Francois, R. (2015). *Dplyr: A grammar of data manipulation*. Retrieved from http://CRAN.R-project.org/package=dplyr

Wickham, & Hadley. (2007). Reshaping data with the reshape package. *Journal of*

406 *Statistical Software*, *21*(12). Retrieved from http://www.jstatsoft.org/v21/i12/paper

407 Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida:
408 Chapman; Hall/CRC. Retrieved from http://yihui.name/knitr/

409 Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R.
410 *Journal of Statistical Software*, *27*(8). Retrieved from
411 http://www.jstatsoft.org/v27/i08/

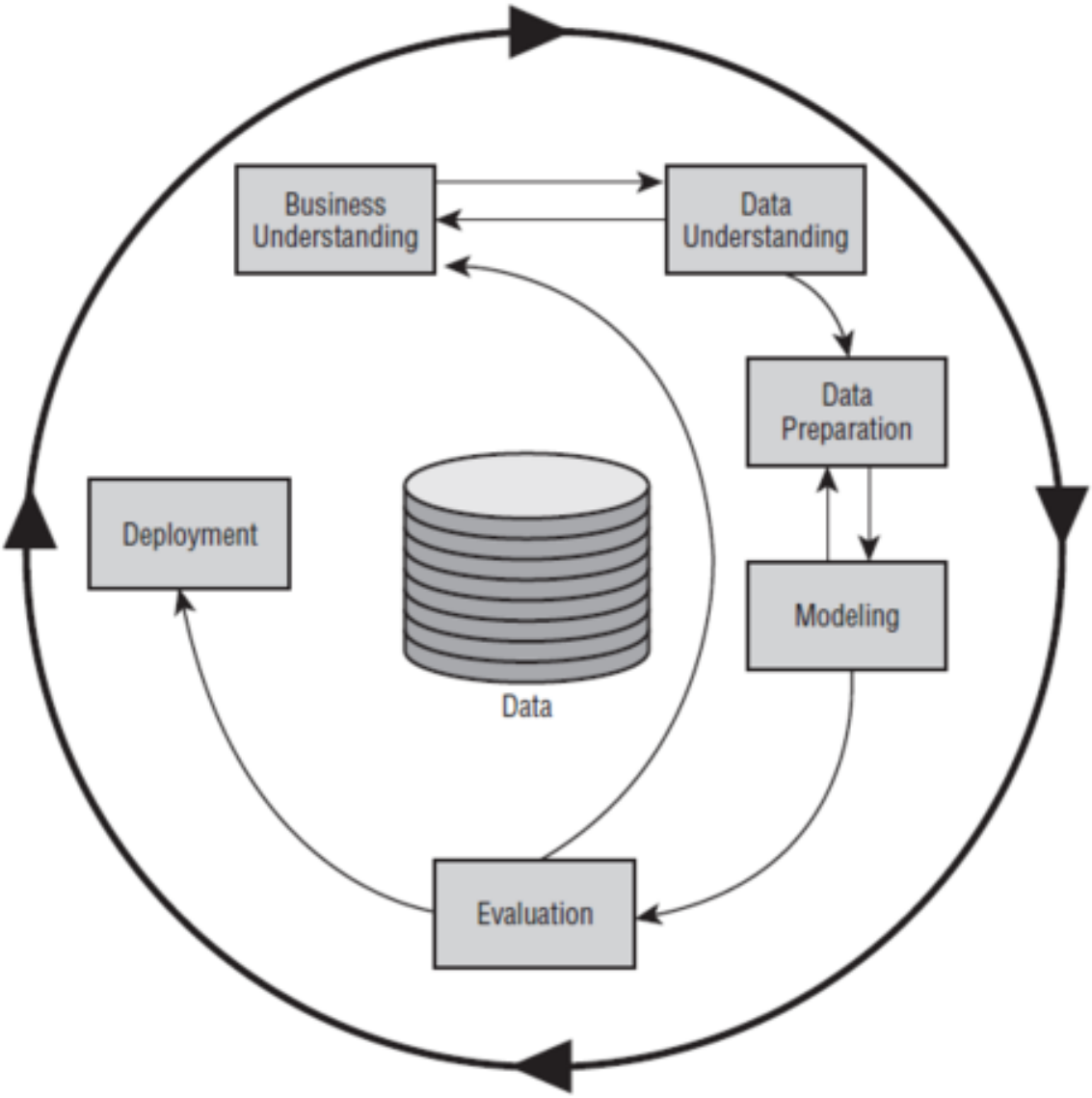


Figure 1. CRISP-DM