# CLASSIFICATION IN DEPOSIT ACCOUNTS

Marc Sandoval, Yahui Peng, Sabrina Fautheree
STA 6933 Advanced Topics in Statistical Learning
Department of Management Science and Statistics
UTSA
May 2, 2022

# Agenda

- Business Problem
- Data Understanding
- Proposed Modeling Approaches

# Business Problem

**Business Problem:** after an investigation, a bank has determined their decline in revenues is due to their existing customers not subscribing to long term deposits. The bank wants to focus their attention on customers that have a higher propensity to obtain these accounts

**Business Objective:**

- Predict which banking customers that are likely to obtain long-term deposit accounts based on customer information and demographics

*https://www.kaggle.com/datasets/rashmiranu/banking-dataset-classification

# Data Understanding

**Number of Observations:** 32,950

**Number of Predictors:** 15

**Binary Response Variable**

| Response | Data Type | Description |
|---|---|---|
| y | binary | has the client subscribed a term deposit? ('yes','no') |

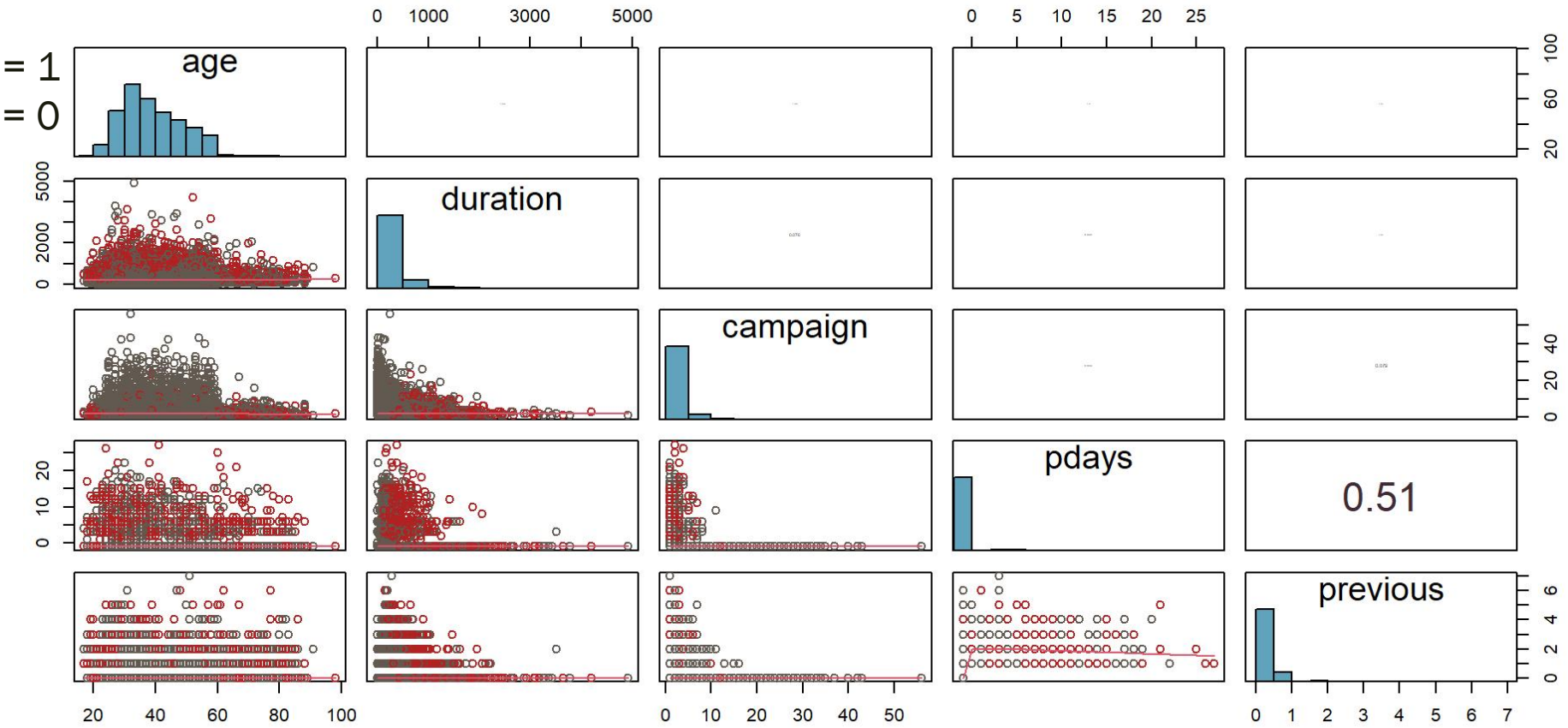
| Predictor | Data Type | Description | Missing Values |
|---|---|---|---|
| age | numeric | age of person | no |
| job | nominal | type of job | yes |
| marital | nominal | marital status | yes |
| education | nominal | level | yes |
| default | nominal | has credit in default? | yes |
| housing | nominal | has housing loan? | yes |
| loan | nominal | has personal loan? | yes |
| contact | nominal | contact communication type | no |
| dayofweek | nominal | last contact day of the week | no |
| duration | numeric | last contact duration, in seconds . | no |
| campaign | numeric | number of contacts performed during this campaign and for this client | no |
| pdays | numeric | number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted) | no |
| previous | numeric | number of contacts performed before this campaign and for this client | no |
| poutcome | nominal | outcome of the previous marketing campaign | no |

# Pairwise Scatterplots, Correlation and Histograms

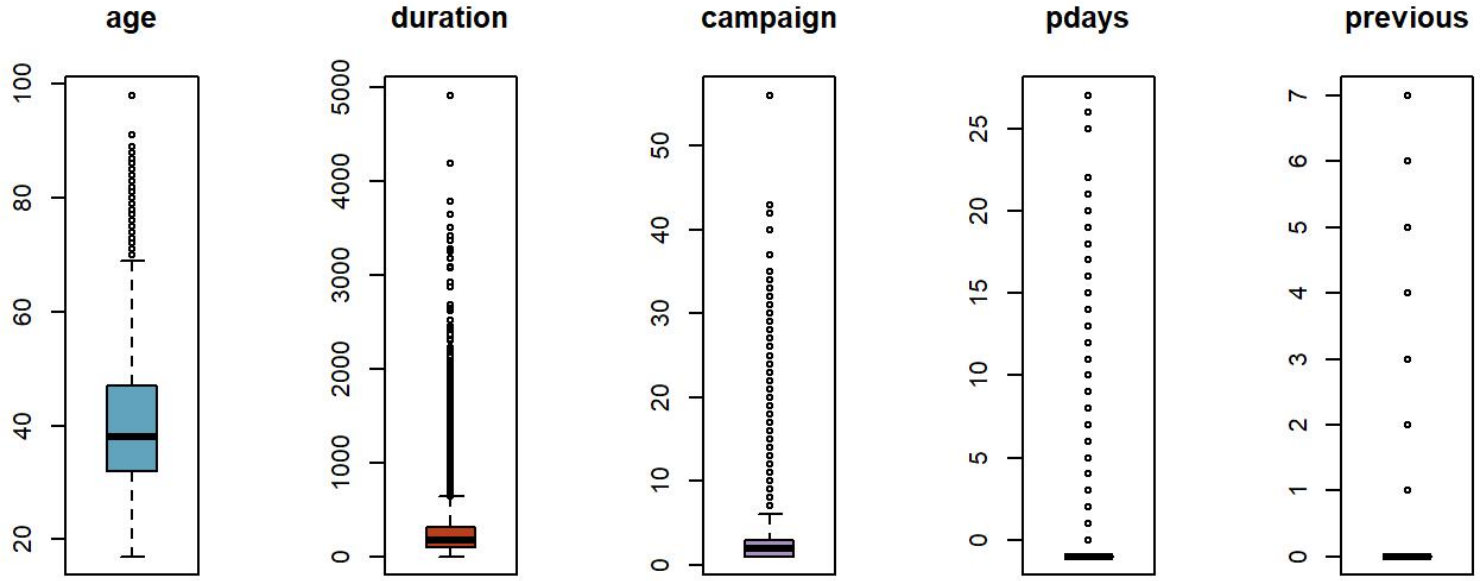
31724 out of 32950 obs have a *pdays* value of 999 (client was not previously contacted)

Encode *pdays* value 999 to -1 for better visualization purpose.

Red dots: y = 1
Grey dots: y = 0

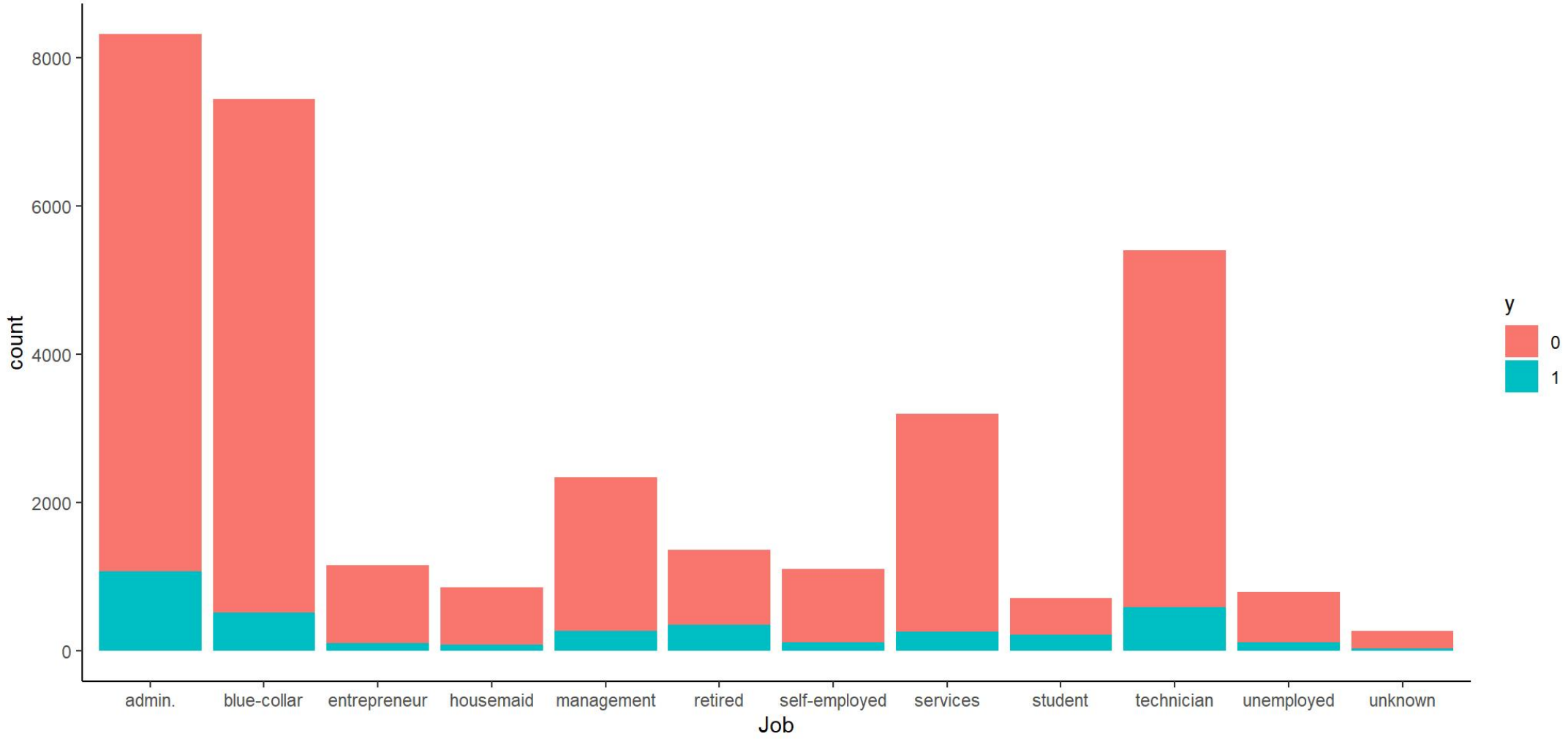


# Boxplots and Summary Statistics

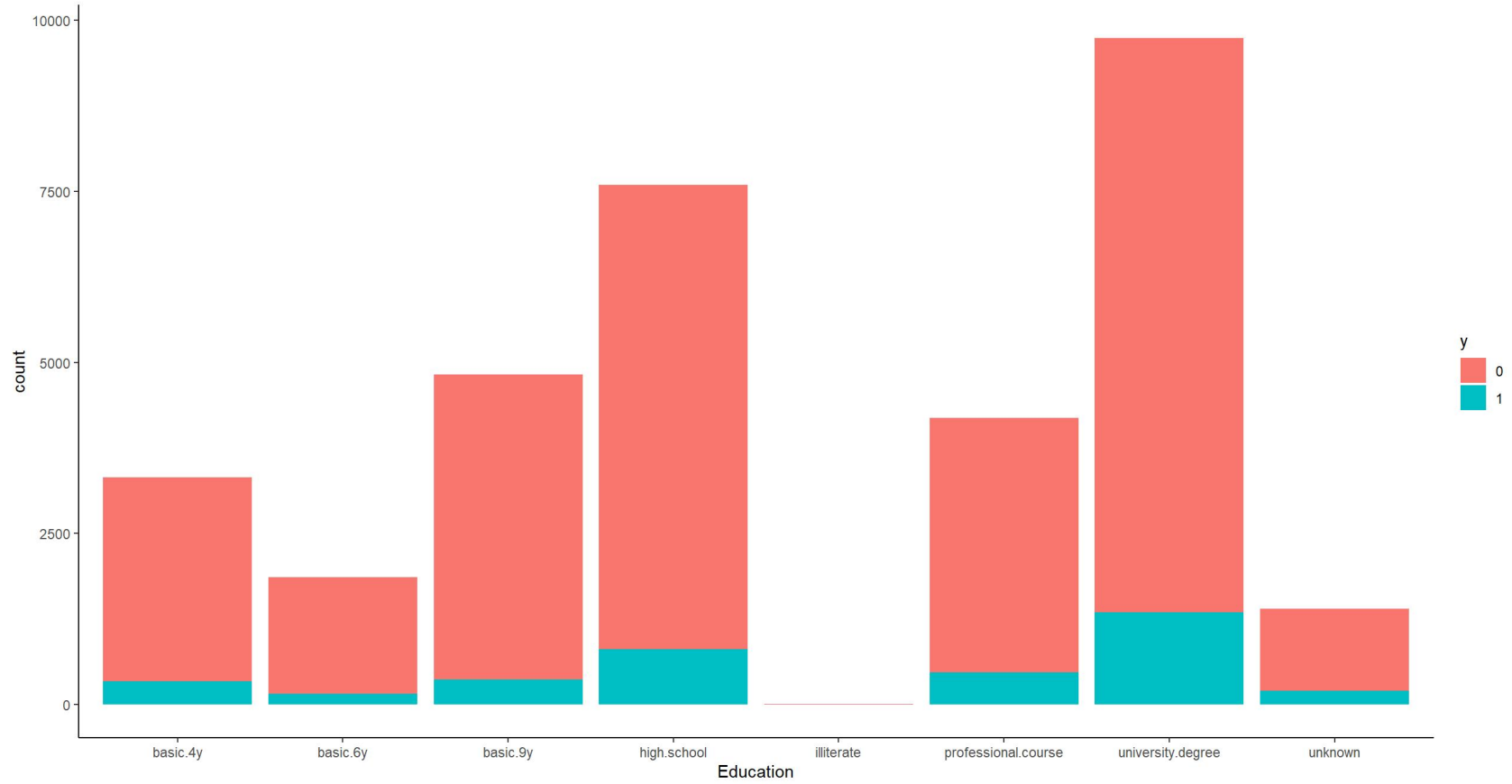



| age | duration | campaign | pdays | previous |
|---|---|---|---|---|
| Min. :17.00 | Min. : 0.0 | Min. : 1.000 | Min. :-1.0000 | Min. :0.0000 |
| 1st Qu.:32.00 | 1st Qu.: 103.0 | 1st Qu.: 1.000 | 1st Qu.:-1.0000 | 1st Qu.:0.0000 |
| Median :38.00 | Median : 180.0 | Median : 2.000 | Median :-1.0000 | Median :0.0000 |
| Mean :40.01 | Mean : 258.1 | Mean : 2.561 | Mean :-0.7397 | Mean :0.1747 |
| 3rd Qu.:47.00 | 3rd Qu.: 319.0 | 3rd Qu.: 3.000 | 3rd Qu.:-1.0000 | 3rd Qu.:0.0000 |
| Max. :98.00 | Max. :4918.0 | Max. :56.000 | Max. :27.0000 | Max. :7.0000 |

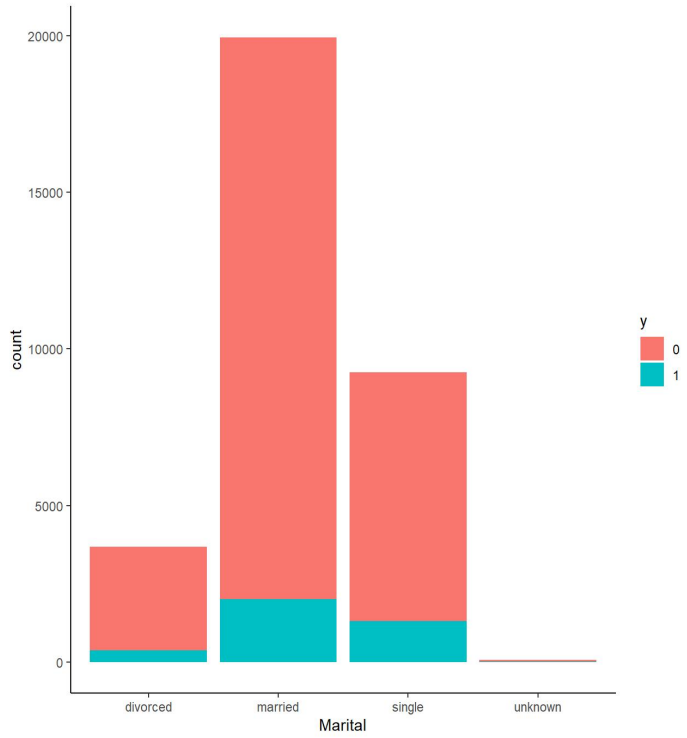
# Bivariate Analysis of Categorical Features

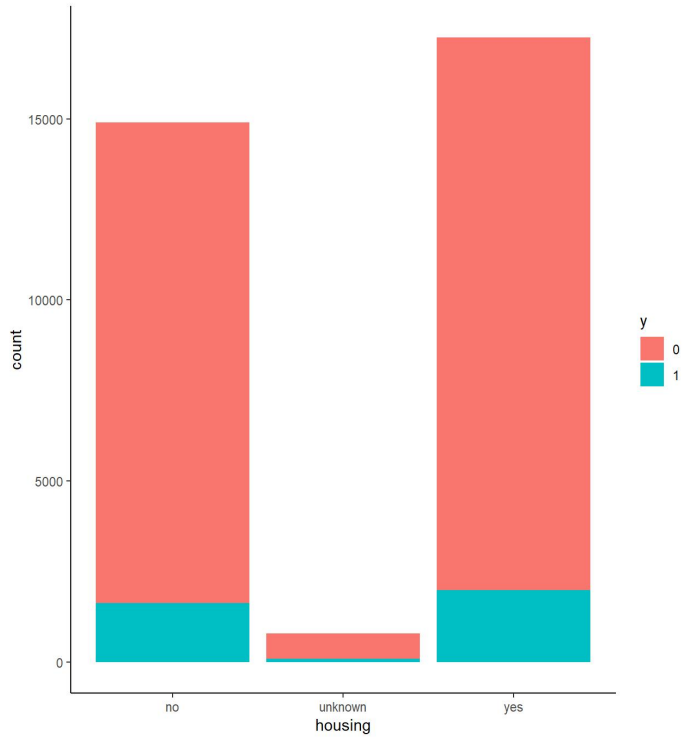


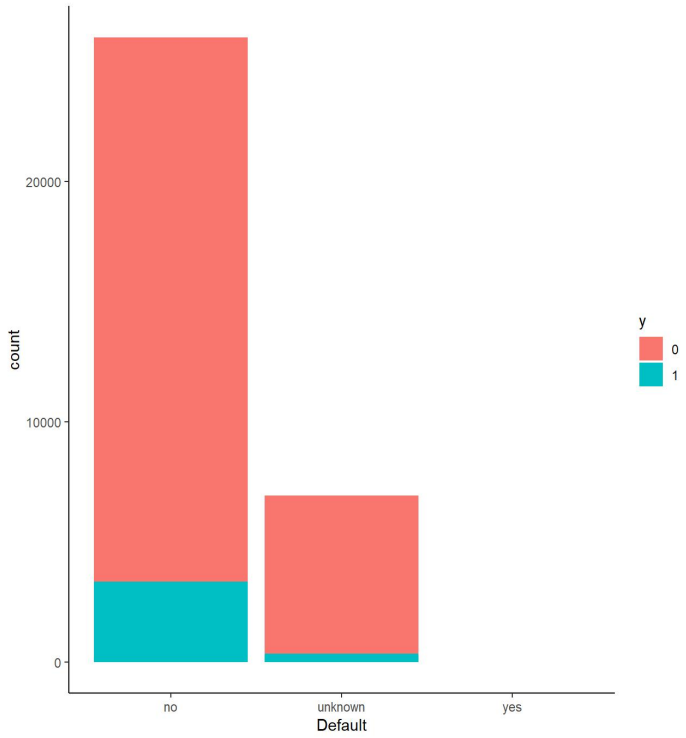
# Bivariate Analysis of Categorical Features



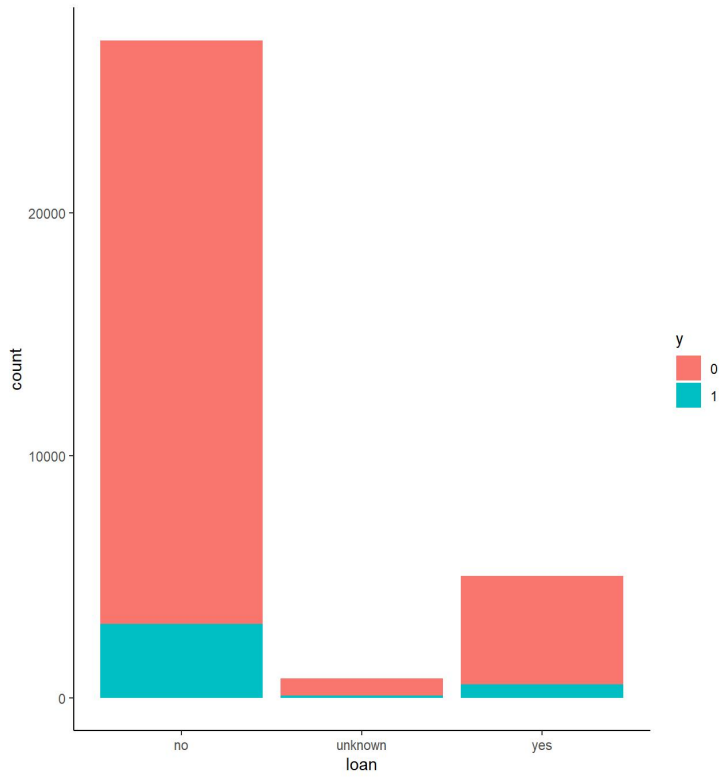
# Bivariate Analysis of Categorical Features

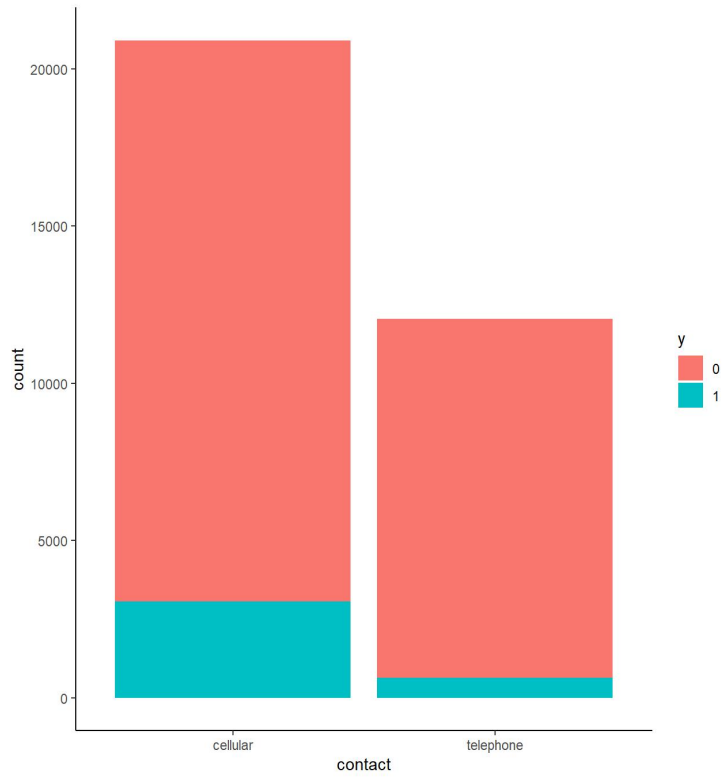


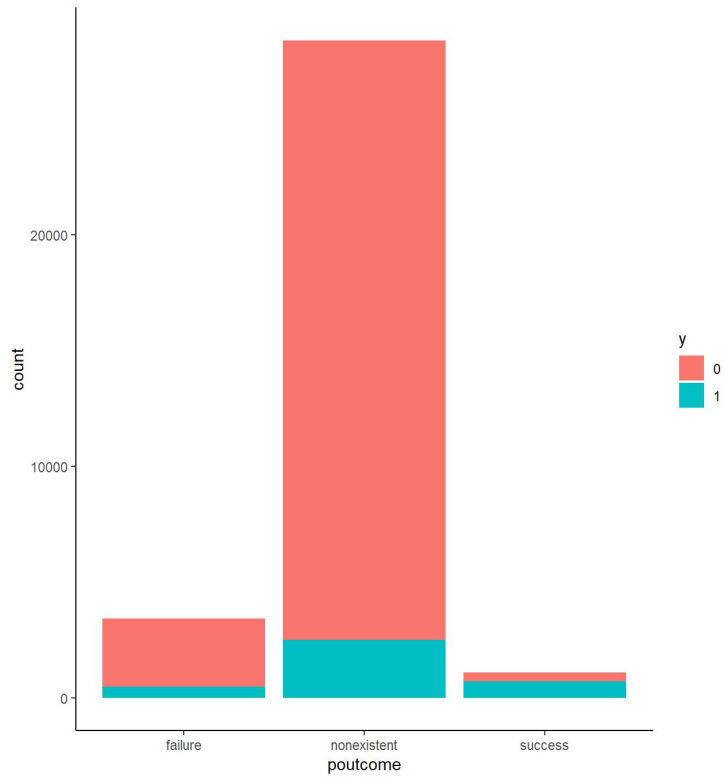



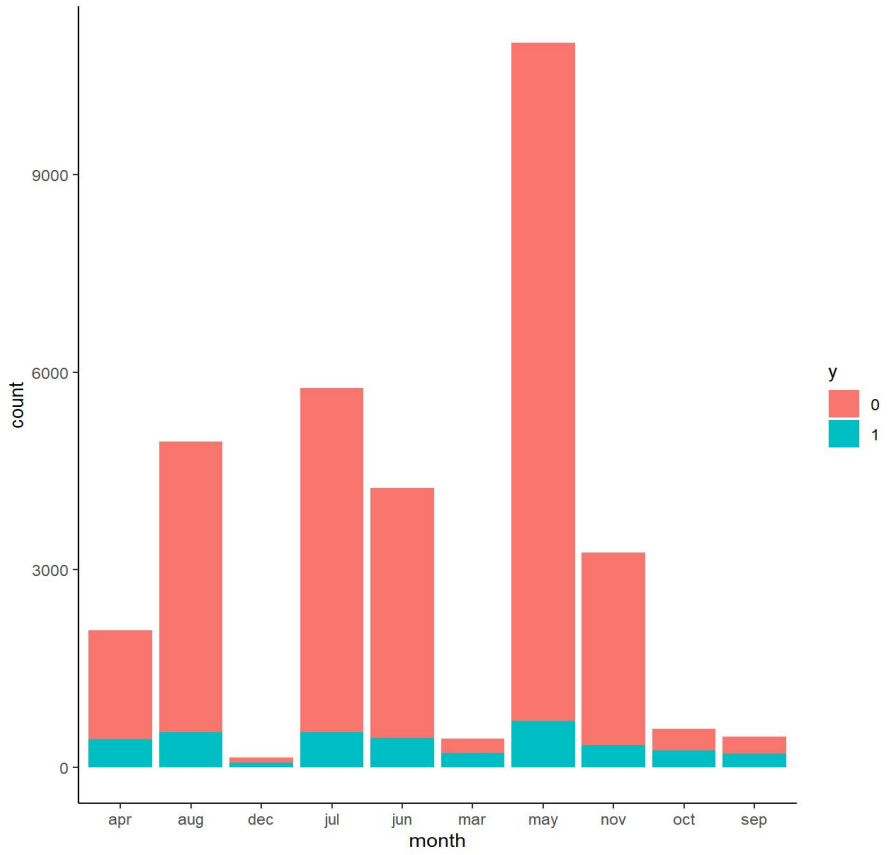
# Bivariate Analysis of Categorical Features

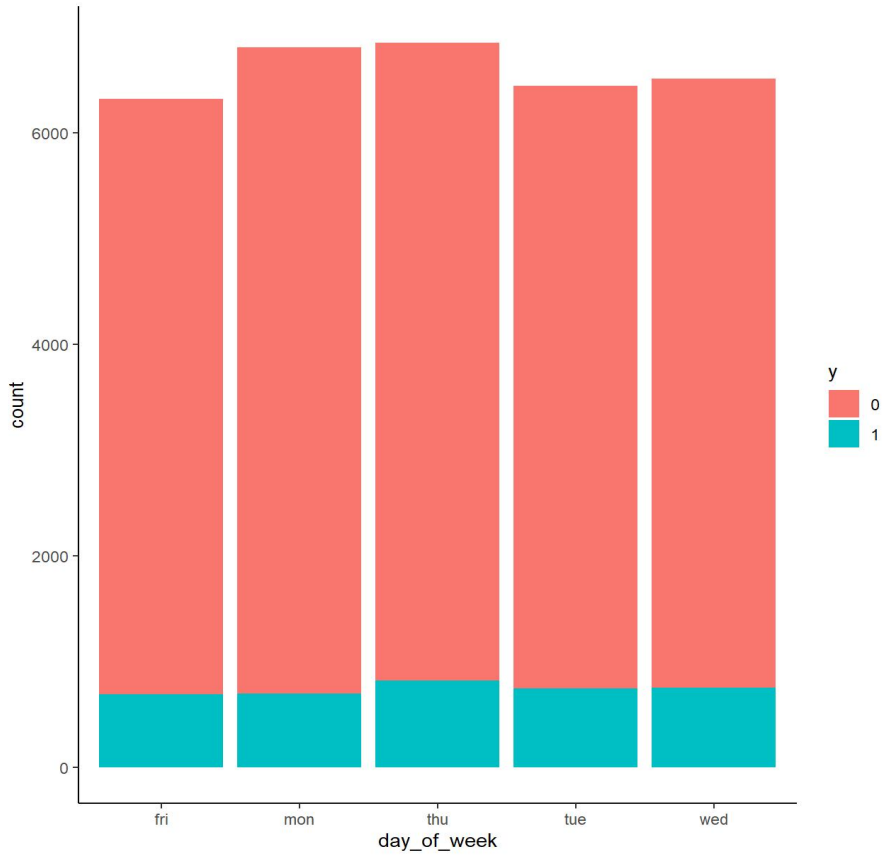


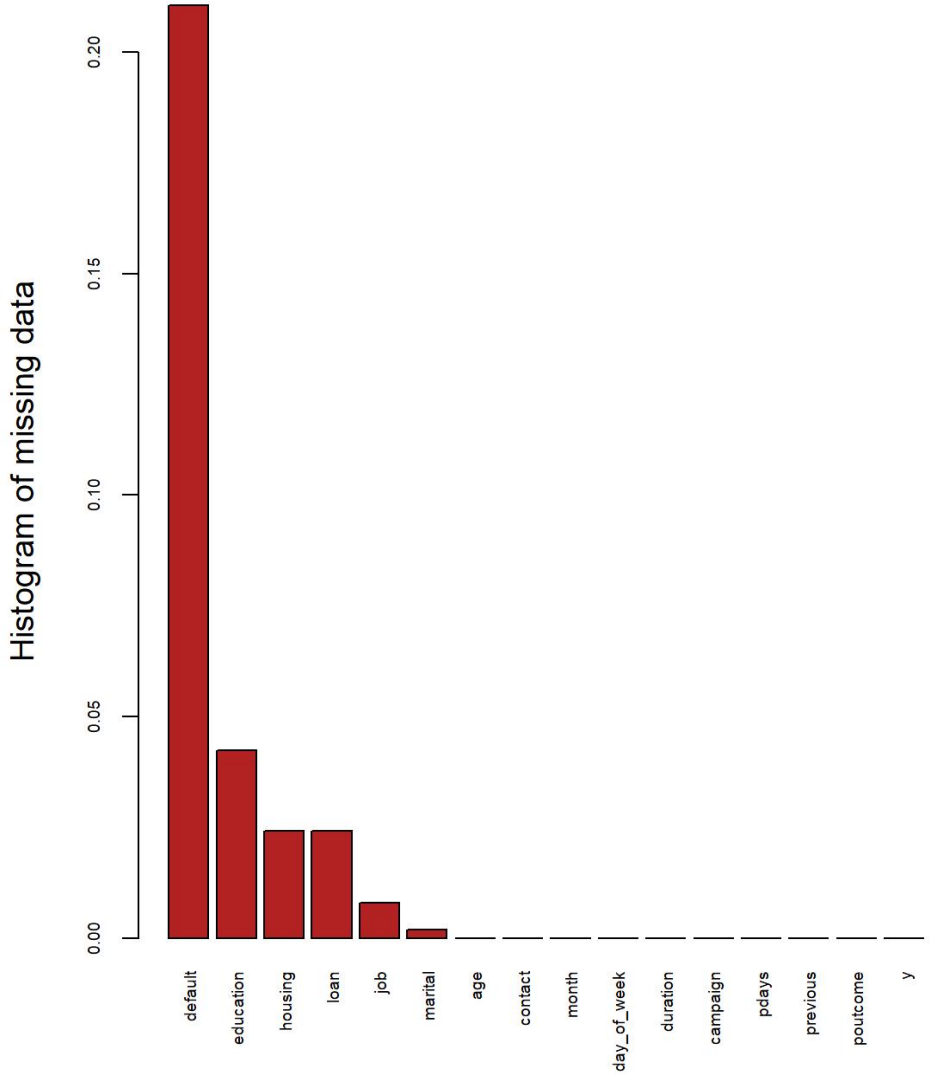



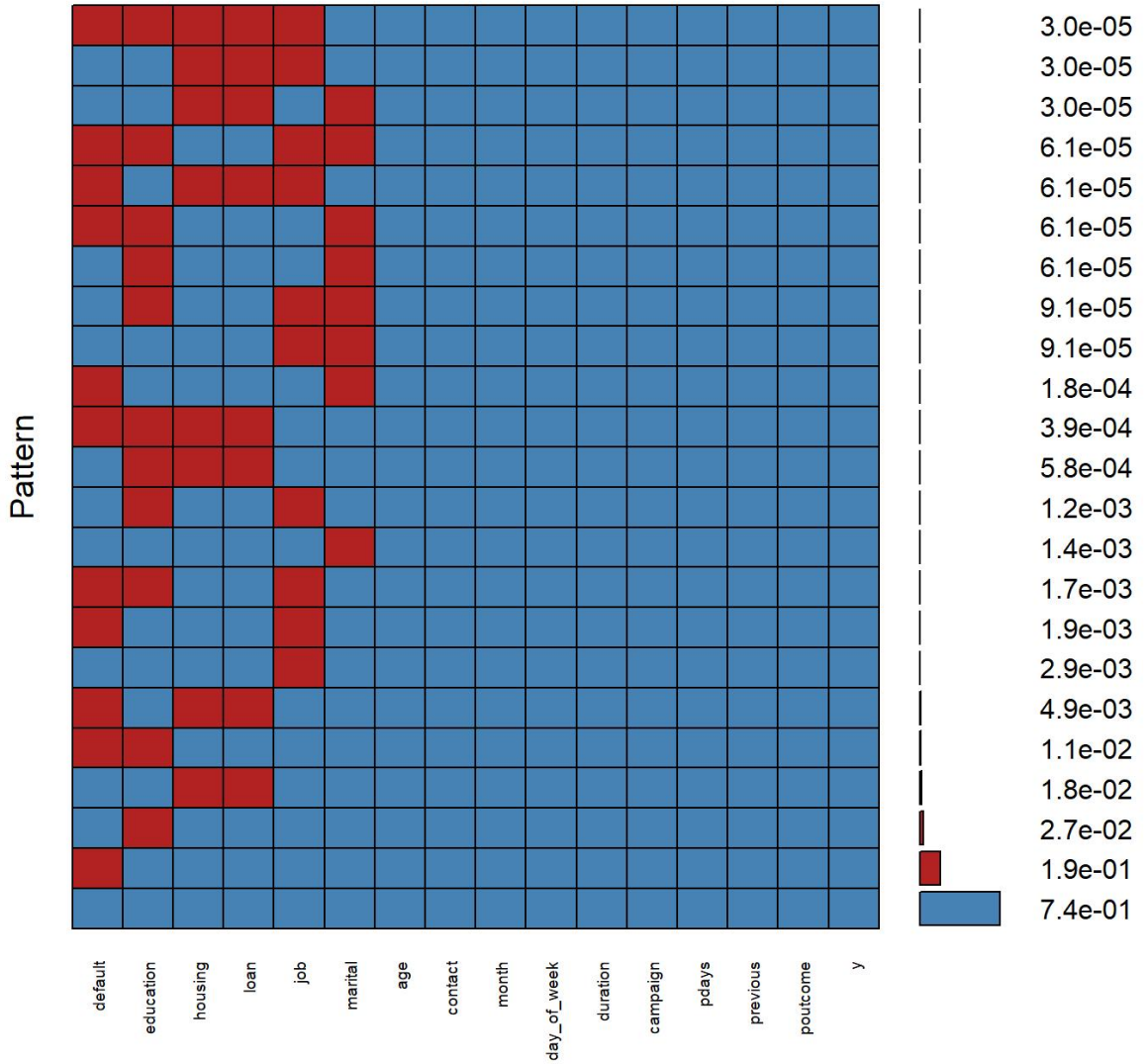
# Bivariate Analysis of Categorical Features



# Missing Patterns



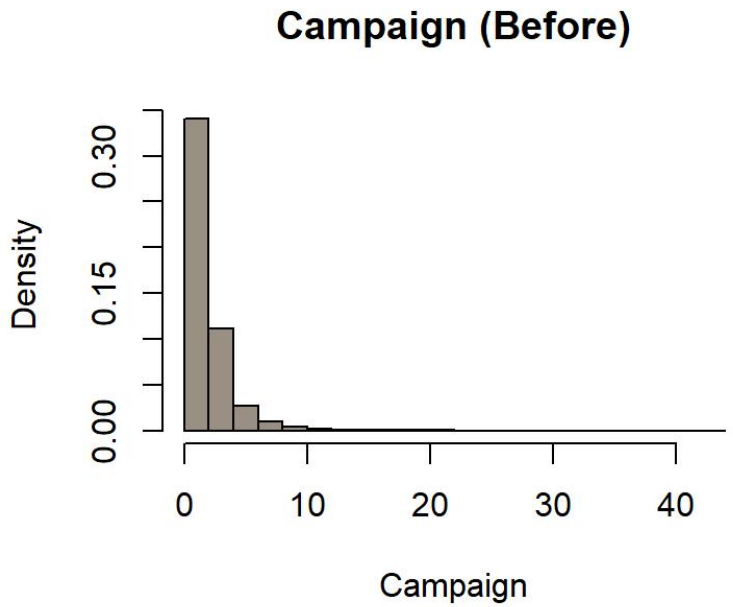
# Data Imputation

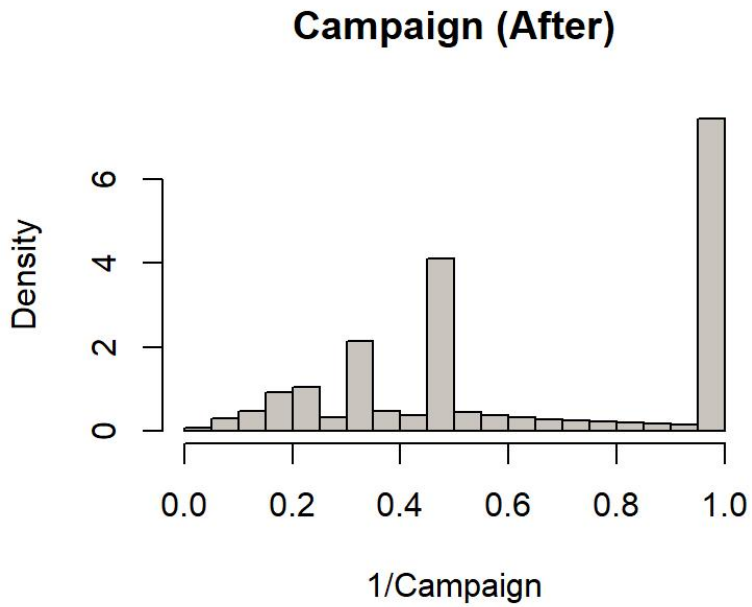
| Predictor | % Missing |
|---|---|
| job | 0.8% |
| marital | 0.2% |
| education | 4.24% |
| default | 21.06% |
| housing | 2.42% |
| loan | 2.42% |

# Skewness Resolving for Training Data

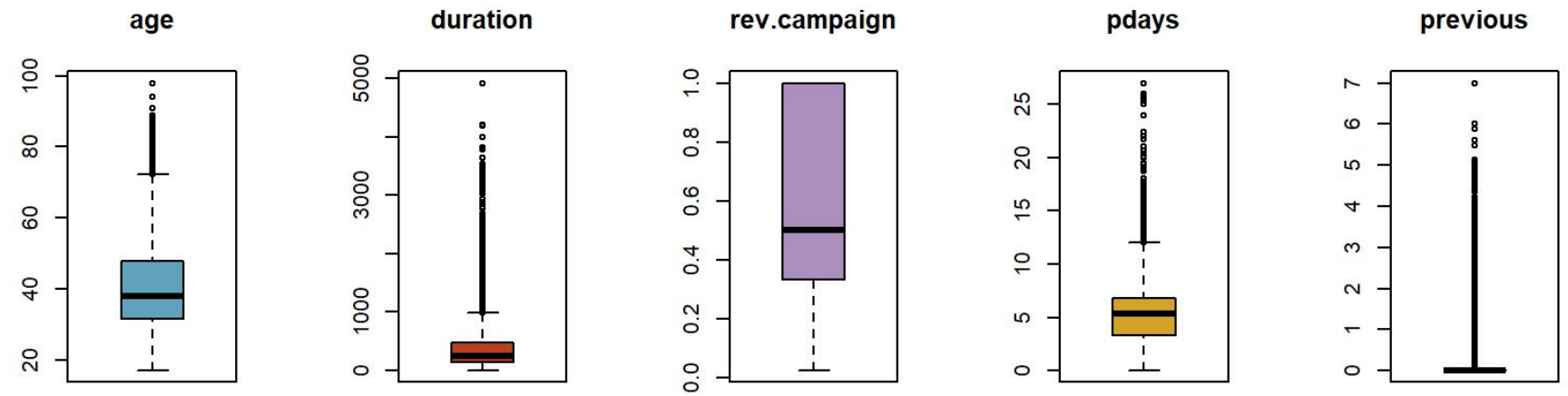
Encode *pdays* of 999 to NA for preprocessing purposes.

```
      age  duration  campaign     pdays  previous
0.5661764 1.1564039 5.1403350 1.0219250 1.2362670
```

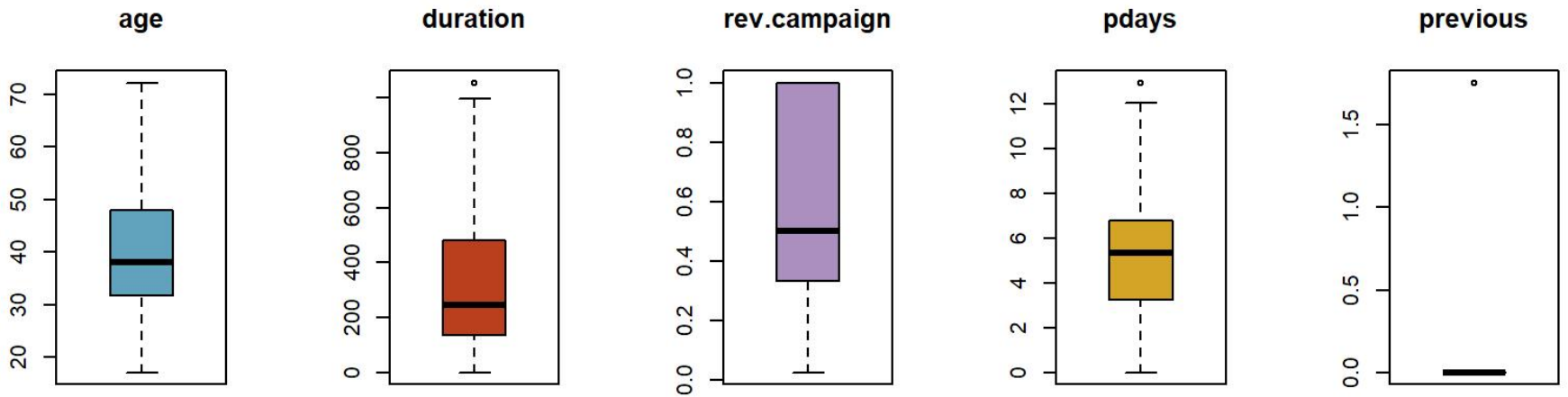


# Outliers Resolving for Training Data




After Capping Low and High Outliers with 5% and 95% Quantile Values, Respectively



# Near-zero Variance Features

```
> names(data_train)[nearZeroVar(data_train)]
 [1] "default"    "job3"       "job4"       "job6"       "job7"       "job9"
 [7] "job11"      "education2" "education5" "month3"     "month9"     "month10"
[13] "month12"
```

# Imbalanced Data

We observe that our response variable is imbalanced, which may cause poor performance in the minority class:

| Class | Count | % |
|---|---|---|
| Y=0 | 29,238 | 0.887 |
| Y=1 | 3,712 | 0.113 |

Apply a “Synthetic Minority Oversampling Technique” (SMOTE) to balance the data in the training partition. The algorithm duplicates records in the minority class using KNN:

| Class | Count | % |
|---|---|---|
| Y=0 | 23,391 | 0.568 |
| Y=1 | 17,820 | 0.432 |

# Proposed Modeling Approaches

- **We will focus on tree-based approaches**
  - Nonparametric methods for classification would be appropriate for our data set, since it consists of categorical data
  - Interpretable methods are appropriate for our business problem: provide our client with demographics they can focus on
  - Models with better predictive power, such as neural nets, might be computationally expensive given the size of our data set

- **Methods**: random forest, bagging, boosting, and XGBoost. Explore other ensemble approaches if time allows

Questions?