SWINBURNE UNIVERSITY OF TECHNOLOGY

Team Members:

**Predicting customer behaviour in banking**

**INF30030 – Business Analysis**

* Phan Trong Hieu – 103426489
* Ngo Cong Thanh - 103433609

# Defining Business Objective

The objective of the project is to use a random forest model to analyze data from direct marketing campaigns of a Portuguese banking institution, with the aim of predicting whether a client will subscribe to a term deposit. The project's focus is on utilizing this predictive model to achieve this goal. The information on customers and their past campaign data, which was originally presented by Moro et al. in 2014 (refer to Moro, Cortez, and Rita), is accessible through the UCI repository at <https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip>. Ultimately, the business objective is to improve the effectiveness of marketing campaigns and ultimately increase the conversion rate of potential clients to actual clients.To anticipate whether a customer will accept (“yes”) or decline (“no”) a term deposit offer, we employ various classification models and algorithms. Afterward, we analyze the outcomes in alignment with our business goals, which are to identify potential clients who are more inclined to sign up for the deposit.

This prediction can be based on various factors such as the client's age, income, employment status, education level, previous transaction history, and other demographic and behavioral data. By analyzing these data points and using machine learning algorithms, businesses can build predictive models that can accurately identify which clients are most likely to subscribe to a term deposit.

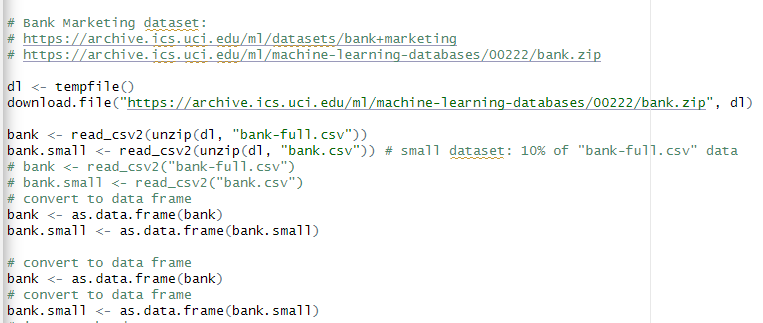
This can help businesses optimize their marketing campaigns by targeting the right clients with the right message at the right time. It can also lead to increased customer satisfaction and loyalty by providing personalized recommendations and offers to clients who are most likely to be interested in term deposits. Ultimately, the business objective of predicting the client's likelihood of subscribing to a term deposit is to increase revenue and profitability by converting more potential clients into actual clients.

There are several benefits to using a random forest model to predict customer behavior in banking:

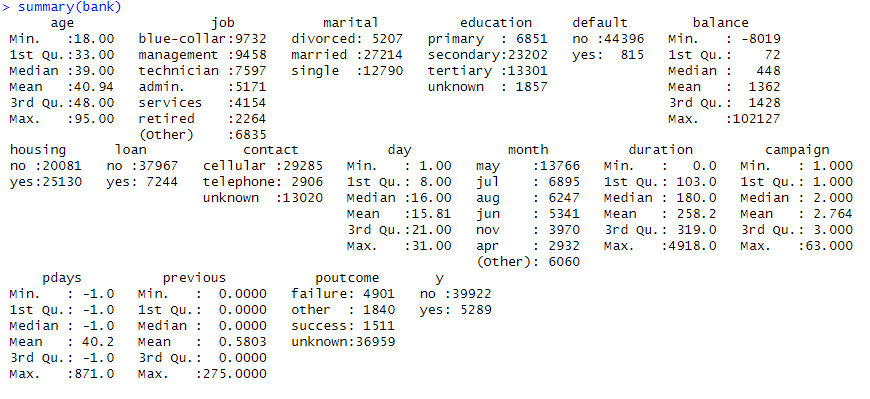
* Accuracy: Random forests have been shown to be highly accurate in predicting customer behavior. They are able to handle large datasets with many variables and can identify complex patterns in the data.
* Interpretability: Random forests provide information on which variables are important in predicting customer behavior. This information can be used to gain insights into customer behavior and inform business decisions.
* Robustness: Random forests are robust to outliers and missing data, which is common in banking datasets.
* Scalability: Random forests can handle large datasets and can be easily parallelized to speed up computation time.
* Versatility: Random forests can be used for both classification and regression problems, making them suitable for a wide range of customer behavior prediction tasks in banking.

# Preparing and exploring data

You can retrieve and extract the bank marketing dataset by utilizing the subsequent code:



There are a total of 45,211 observations with 17 variables that were gathered over the course of five years, comprising the dataset. The factors that are used to make predictions include both numerical values, such as age, balance, and duration, as well as discrete or categorical data, such as job, marital status, education, housing, and poutcome. The result of the prediction is either "yes" or "no," which is a categorical outcome.



***Figure 1.*** *Summary of bank-full.csv*

The dataset authors noted that the distribution of the outcome variable y reveals an imbalance, with 78% of the responses being "no" and only 12% being "yes". Such an uneven distribution is frequently observed in real-life datasets. In the modeling phase, we will employ undersampling to create a dataset with more balance. However, there are other methods that can be helpful as well. Next, we will run models on both the original dataset and the balanced dataset to compare their performance metrics.

To facilitate analysis and modeling, we divided the data into two parts, namely training data and test data. The training data consists of 90% of the original dataset, while the remaining 10% is reserved for testing. If the dataset is small, it might be beneficial to allocate a higher percentage of the data for testing purposes. However, this particular dataset contains sufficient data for training and evaluating on the 10% set aside for testing.

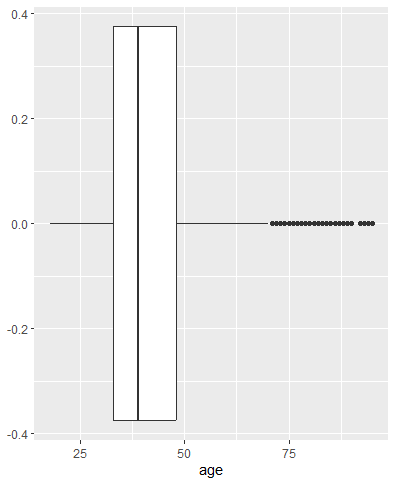
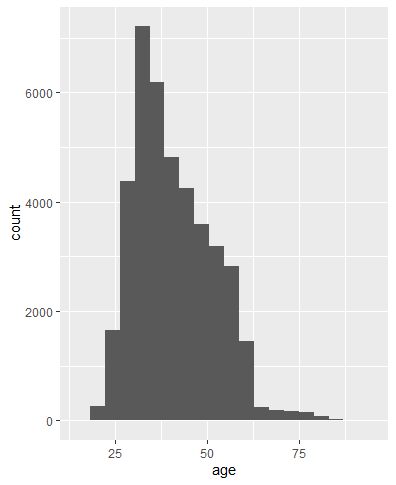
Furthermore, there are variables in the dataset that have unknown values. These variables will be treated as a separate category and no attempts will be made to fill in the missing values. Details about these variables will be provided below.

## The visual exploration of the training data

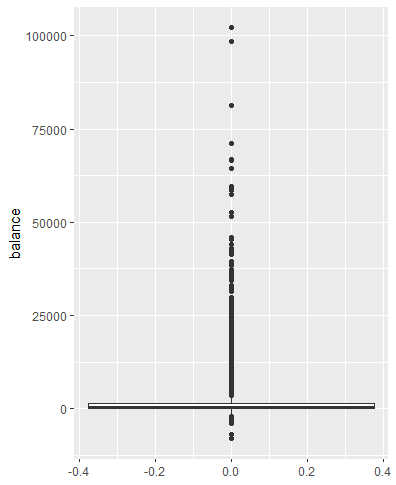
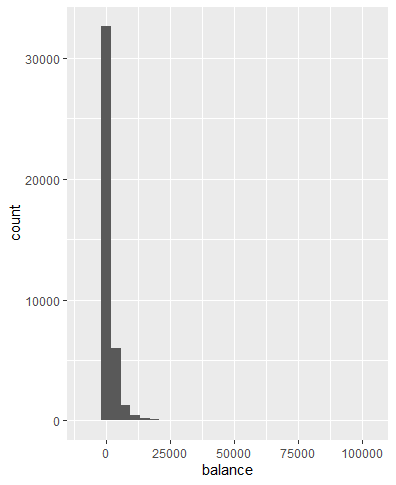
Although feature selection could be performed on the provided features, it is not within the scope of this analysis. However, considering our knowledge of the domain, we anticipate that several features, including balance, age, housing, loan, duration of contact, and previous outcome, would have an impact on the prediction.

**Numerical data**

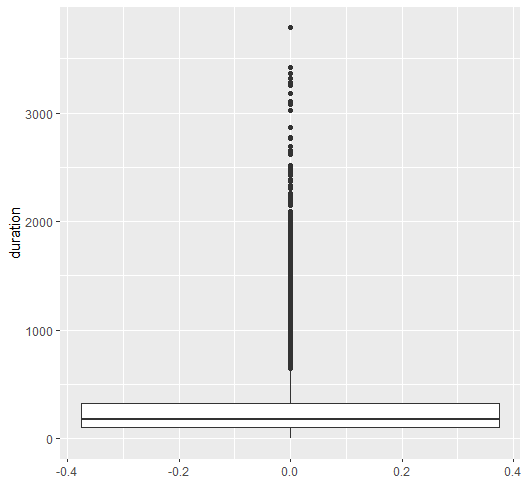
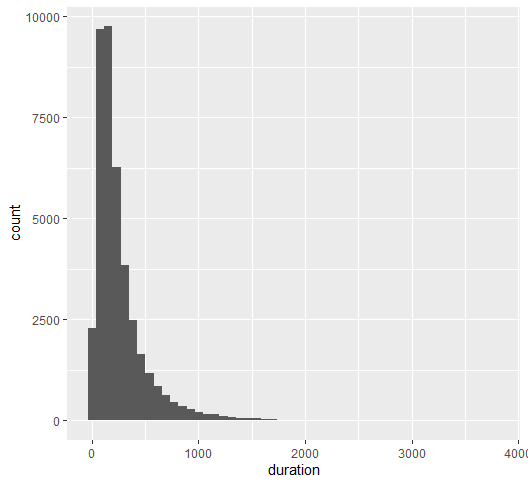
The mean and median are indicated to be close together, the value is around 40 years.



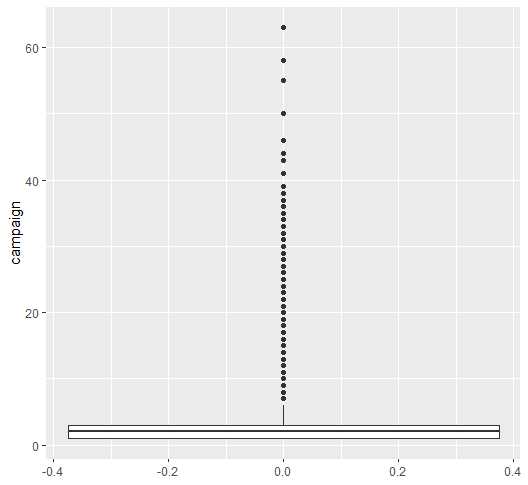
The distribution of balance: right skewed, the mean value is around 1365.923



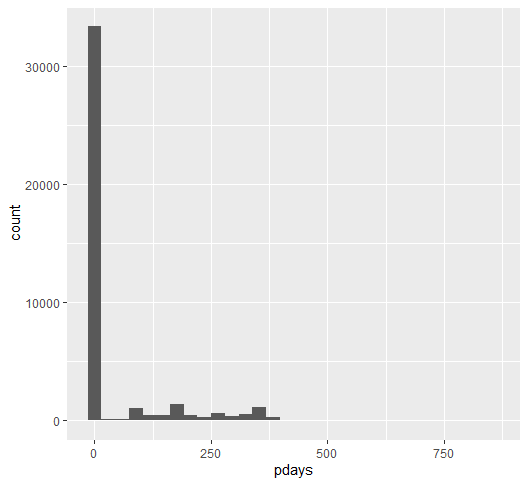
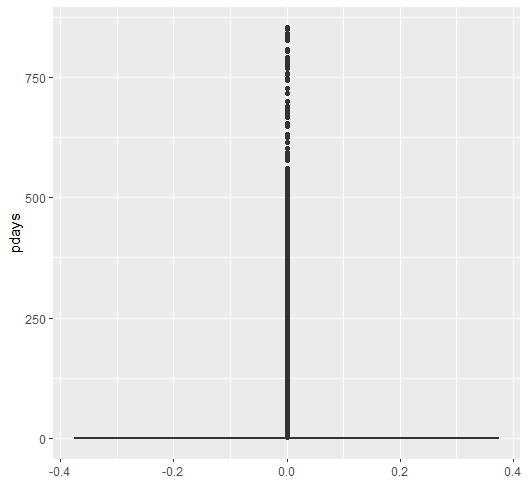
The contact duration: right skewed, remains some outliers.



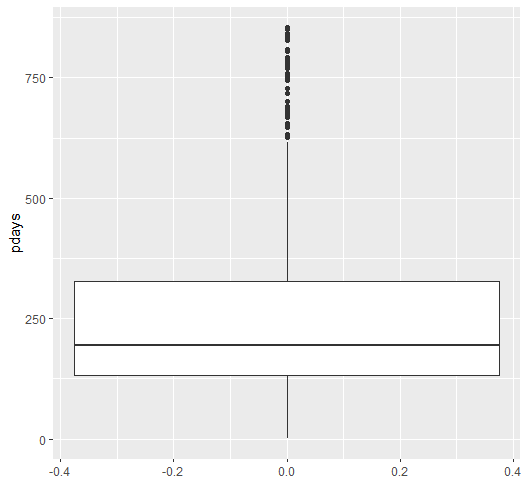
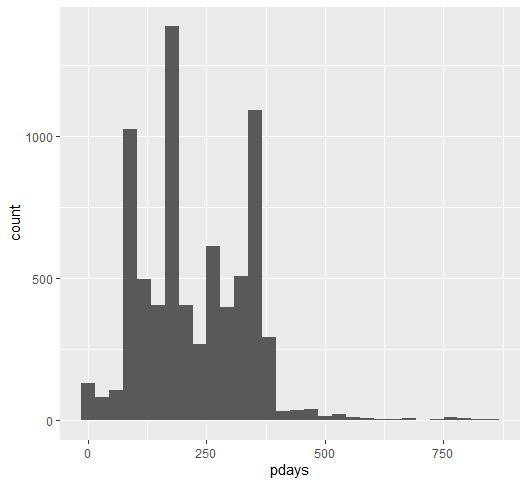
The campaign data: right skewed, this stores the count of the contacts that were made for this client during the campaign.



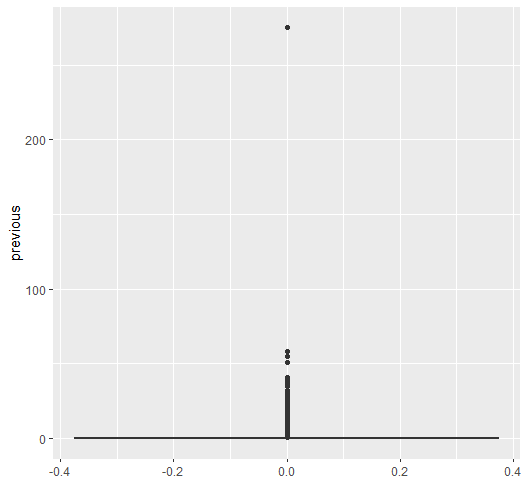
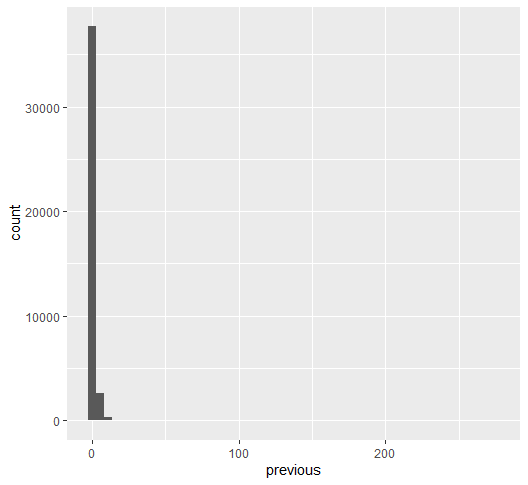
The pdays data: right skewed with outliers, shows a significant number of values as -1(33252 values). It records the duration in days since the client was last contacted, where a value of -1 indicates that the client was not contacted previously.

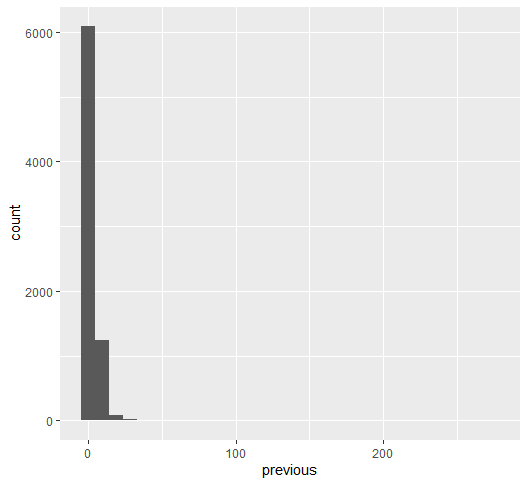
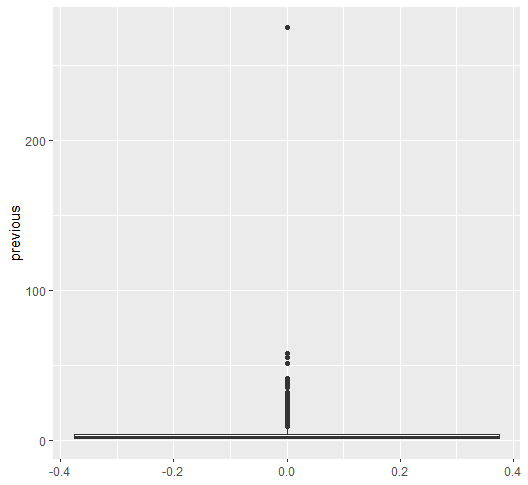
Removing -1 values: still remains some outliers



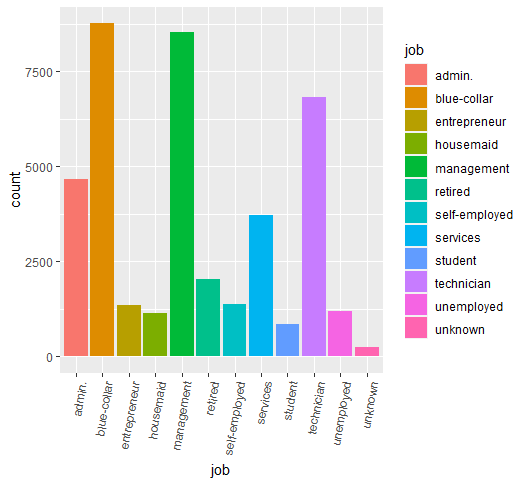
The variable called "previous" is comparable to the pdays variable mentioned earlier, but instead of -1, it uses 0 to indicate that no previous contact was made. This variable keeps track of the number of contacts made for this client in earlier campaigns.



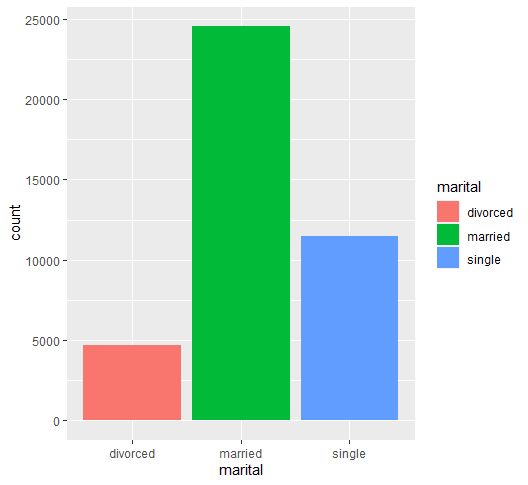
Removing 0 entries

 **Categorical data**:

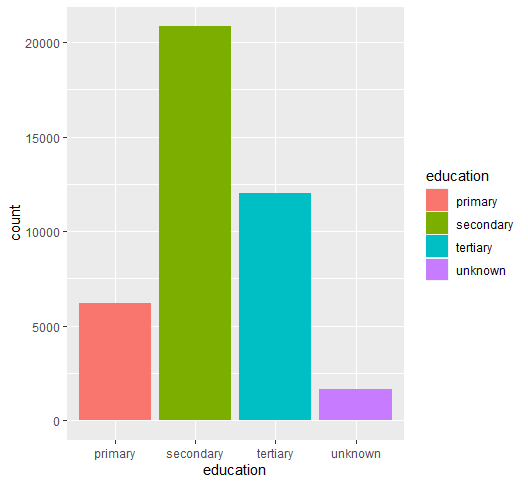
blue-collar jobs, management, technician, admin and services are 5 sectors that have most employees.



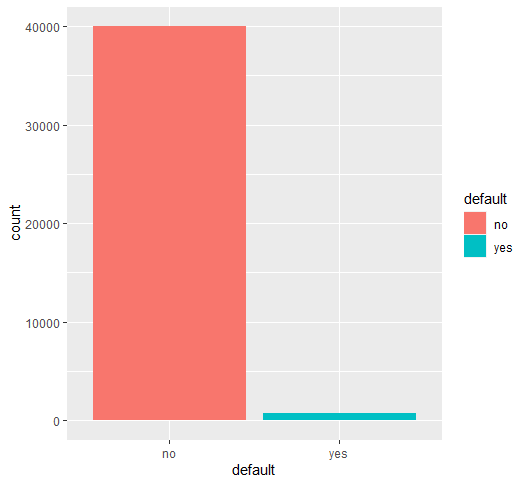
The majority of individuals in the dataset are in a married relationship (24520 people)



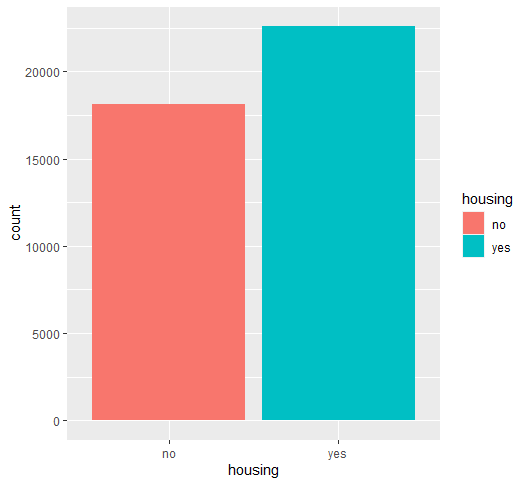
A majority of individuals have finished their secondary education at least



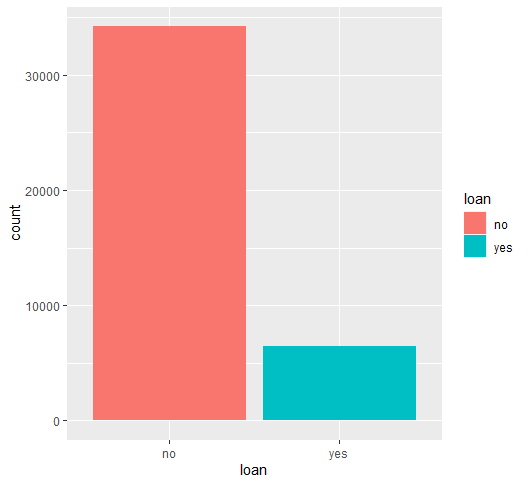
Most people were able to fulfill their financial responsibilities and did not default.



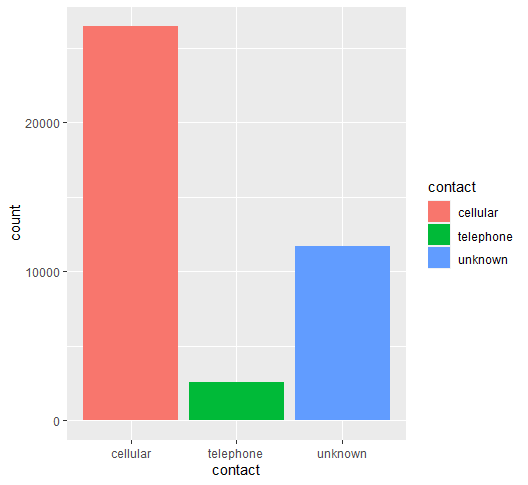
Over 50% of individuals possess mortgages, however, the differences in proportions are less pronounced compared to other aspects.(55% yes vs 45% no)



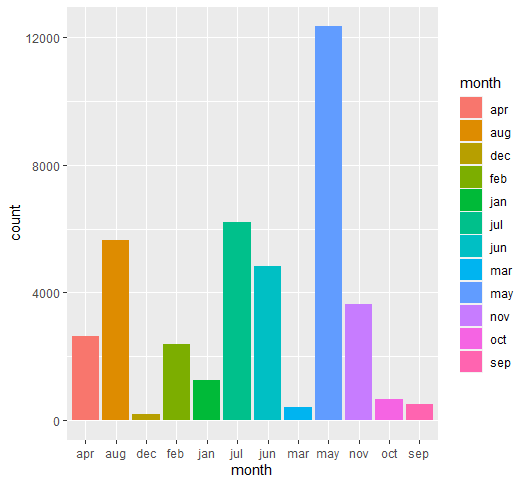
The majority of individuals do not hold personal loans.



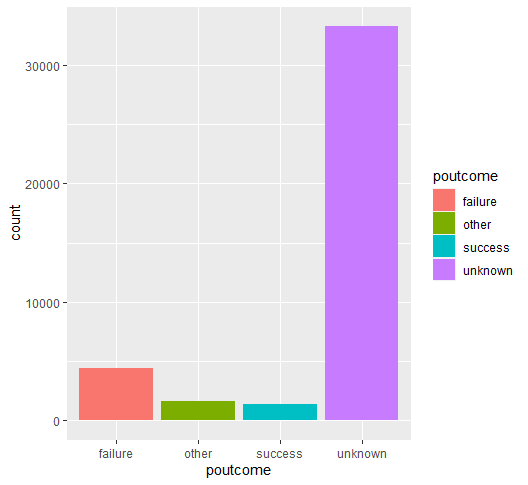
A significant number of individuals were reached through mobile phones, however, the count of unknown values is relatively elevated.



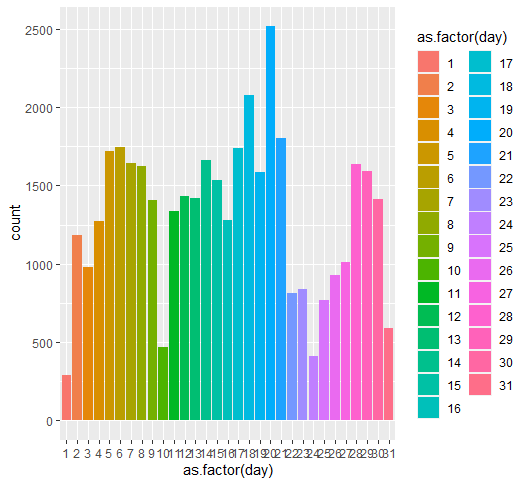
The highest number of individuals were contacted between May to August, whereas the least amount of people were contacted during December and March.



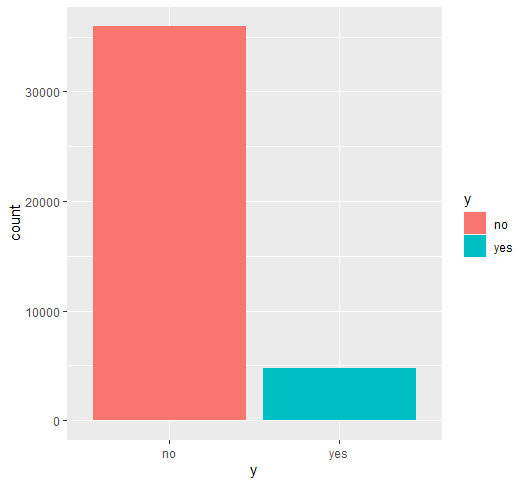
For the available data, the previous result is largely unsuccessful, however, a majority of the entries are unknown.



The highest number of individuals were contacted on the 18th, 19th and 20th days of the month, whereas the lowest number of people were contacted on the 1st, 10th and 24th days.

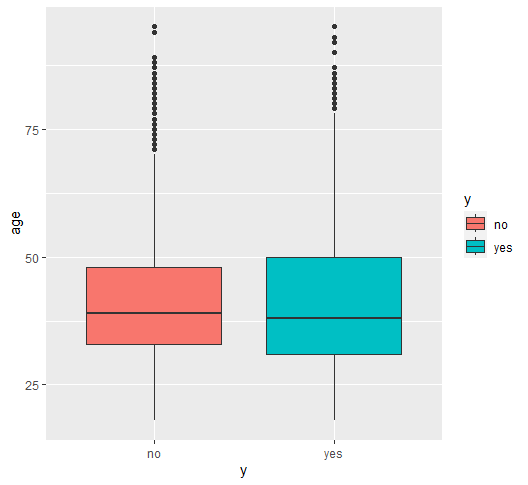


The dependent variable, outcome y, has a greater number of "no" values compared to "yes" values. Data is not balanced, might use under-/upsampling or weights

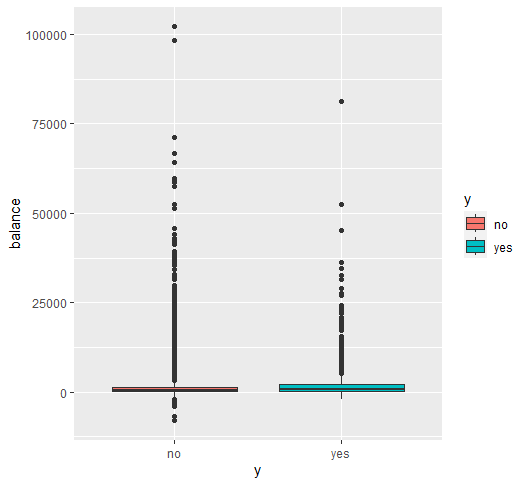


Upon examining the connections between the characteristics and the outcome y, the following factors are discovered:

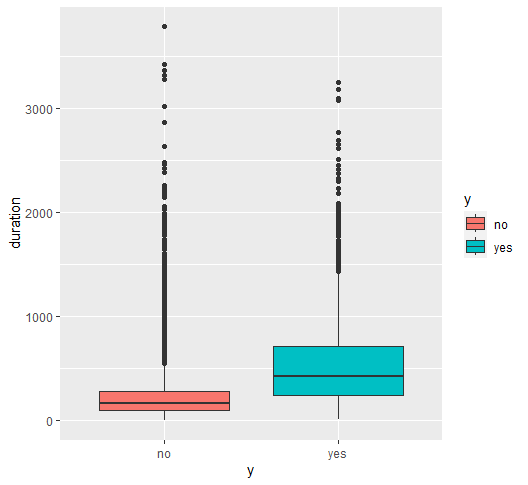
The median age for "yes" responses is somewhat lower, however, the data is more broadly distributed.



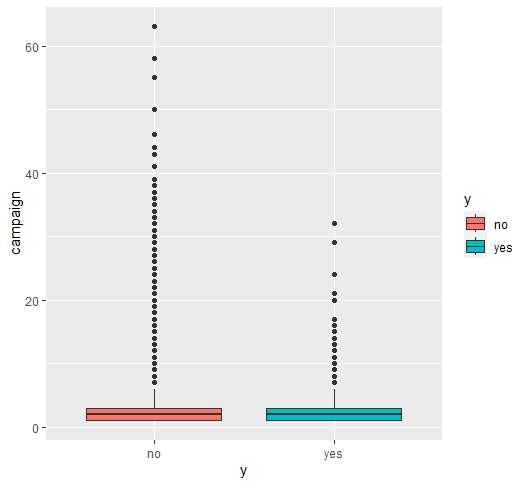
The distribution of balance is comparable for both "yes" and "no," except for the presence of more significant outliers in the "no" category.



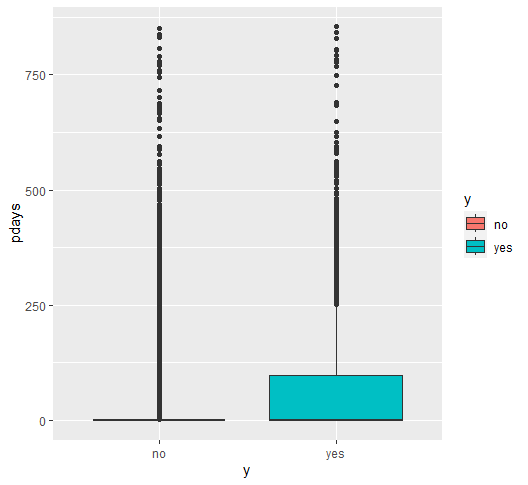
Despite the wider spread of data, the median duration for "yes" is greater than that of "no." This implies that the duration of contact is a crucial factor in predicting the outcome.



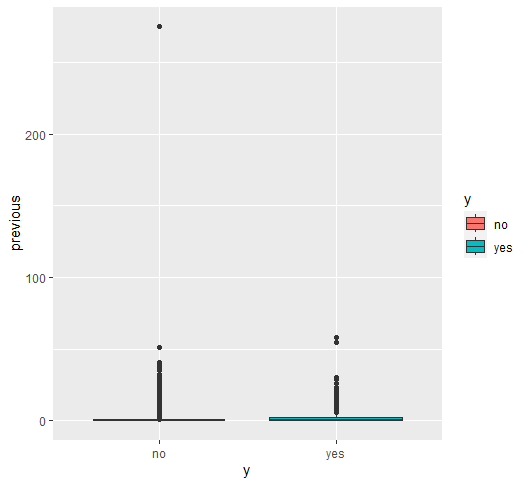
The distribution of the campaign is comparable for both "yes" and "no," except for the presence of more significant outliers in the "no" category.



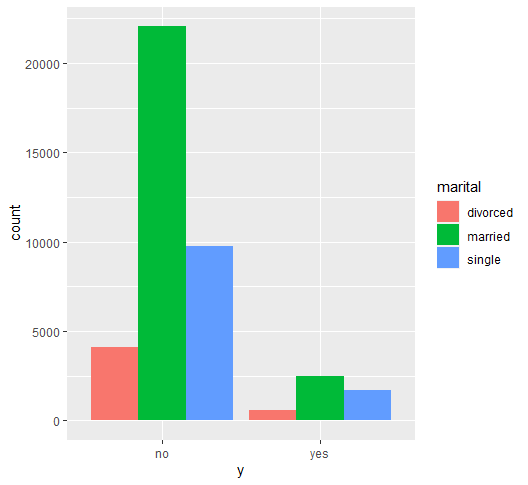
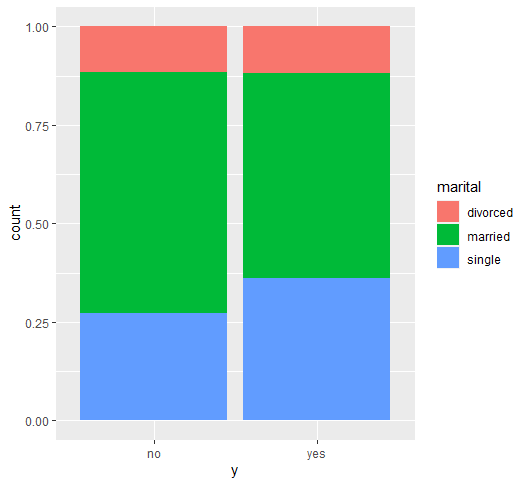
The “yes” values for pdays have wider spread.



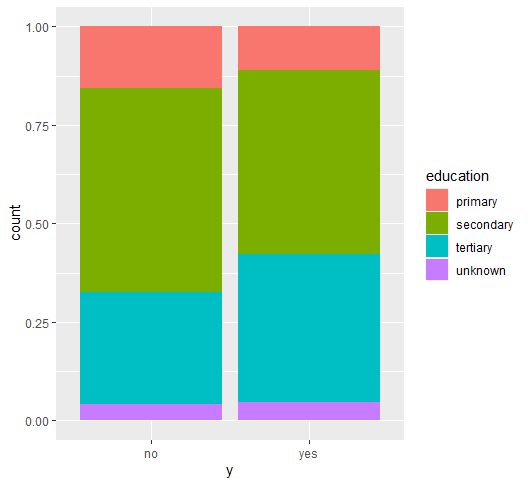
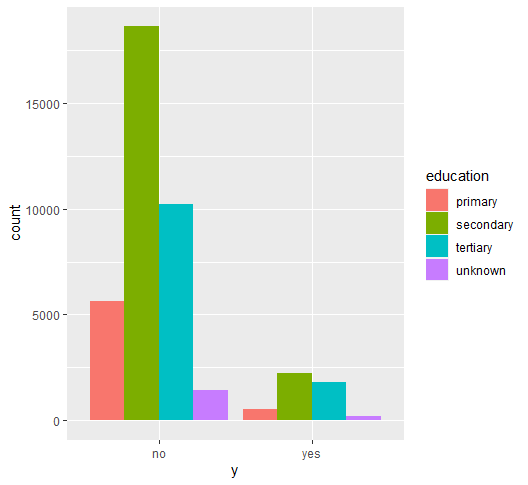
The distribution of the "previous" variable is alike for both "yes" and "no," except for a solitary high outlier that is specifically linked to the "no" category.

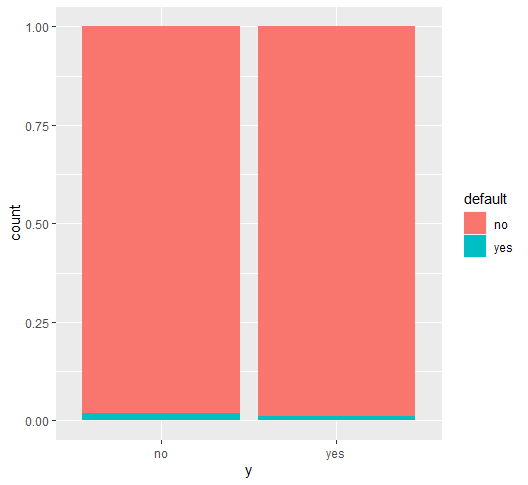
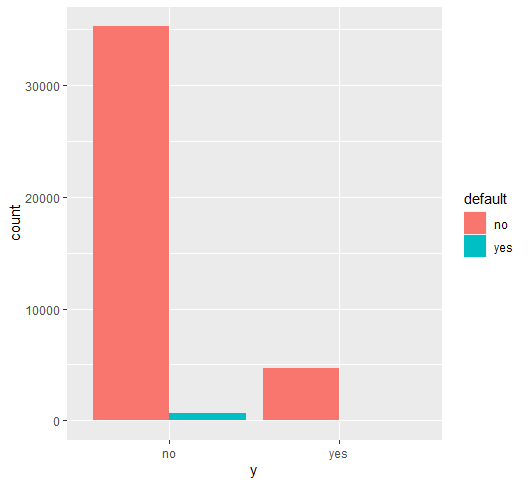


Although the majority of individuals in both groups are married, single people have a higher tendency to respond "yes."

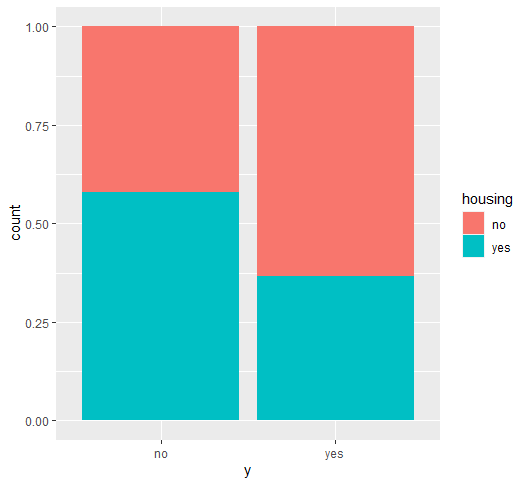
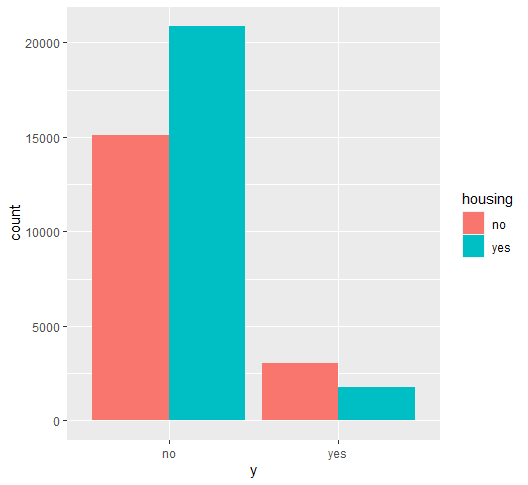
 

Individuals with a secondary level of education constitute the majority in both groups. However, those with tertiary education are more likely to respond "yes" when compared proportionally, whereas those with a primary level of education are more likely to respond "no."

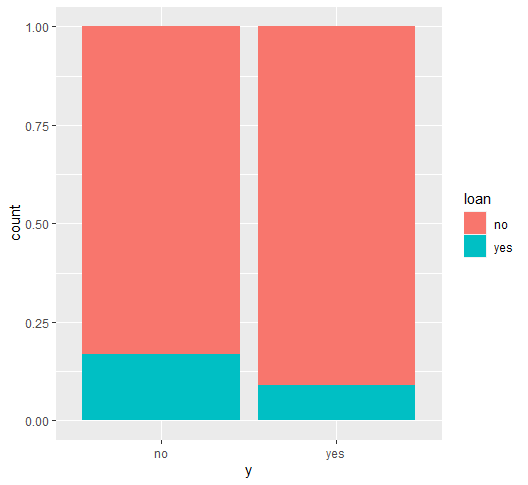
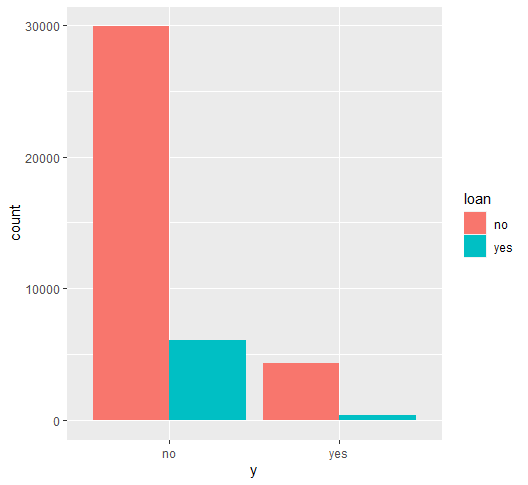


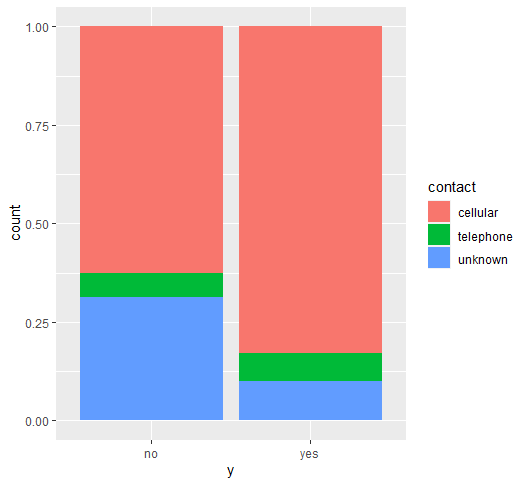
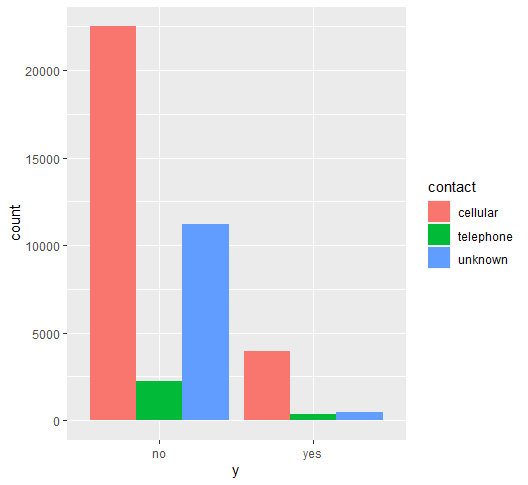
It is unclear what is being referred to as "unknown." However, the sentence "The majority of "yes" responses come from individuals who did not default, but the situation is similar for "no" answers" can be paraphrased as: Most of the individuals who responded with a "yes" did not default, but the same pattern is observed among those who responded with a "no."

Individuals who do not have housing loans are responsible for the majority of "yes" responses, while those with housing loans constitute the majority of "no" answers. This observation is consistent with the goal of acquiring new customers.

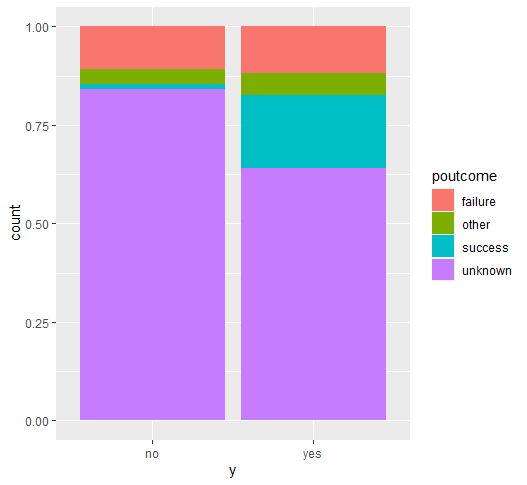
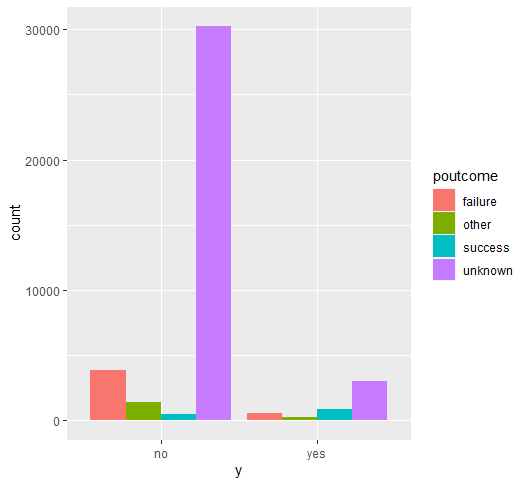


Most of the individuals who responded with a "yes" do not have a loan, which is in line with the objective of attracting new customers.

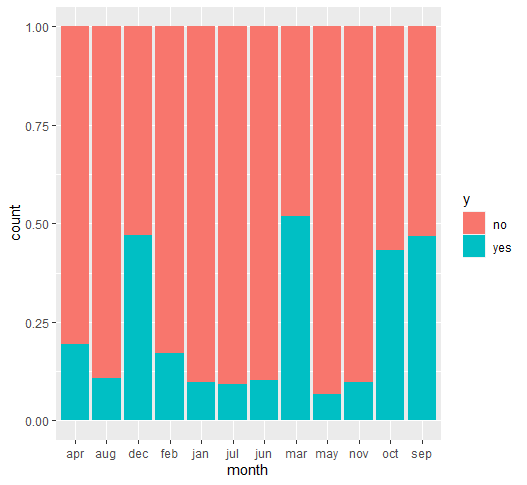
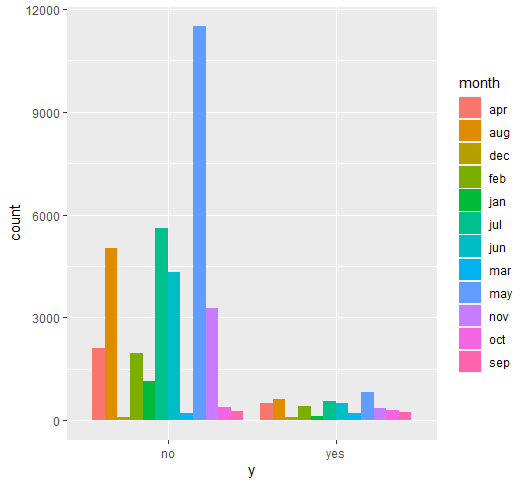
The majority of "yes" responses were obtained from individuals who were contacted through their mobile phones.



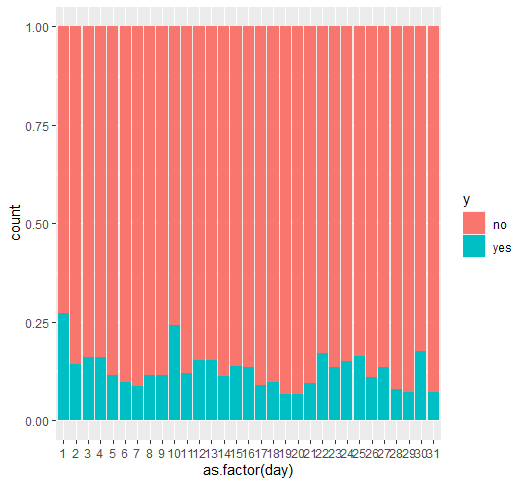
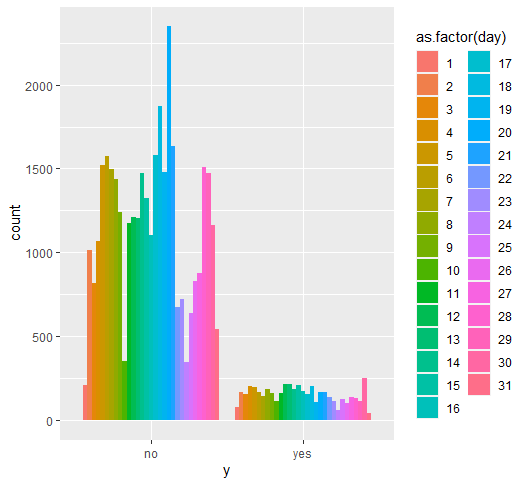
A significant number of "yes" responses were provided by individuals who had experienced successful outcomes in the past. Although this could be linked to a greater level of trust, it is not a topic that falls within the scope of the analysis.



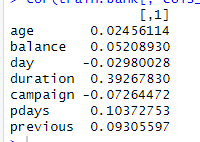
The months of March, December, October, and September have the highest proportion of "yes" responses, and these are also the months during which fewer calls are made. Conversely, during the months with the highest number of calls, the proportion of "yes" responses remains relatively stable.



The days with the greatest proportion of "yes" responses are the 1st, 10th, and 30th of each month. Although the 1st and 10th are days when fewer calls are made, the 30th is a day when a larger number of calls are made.



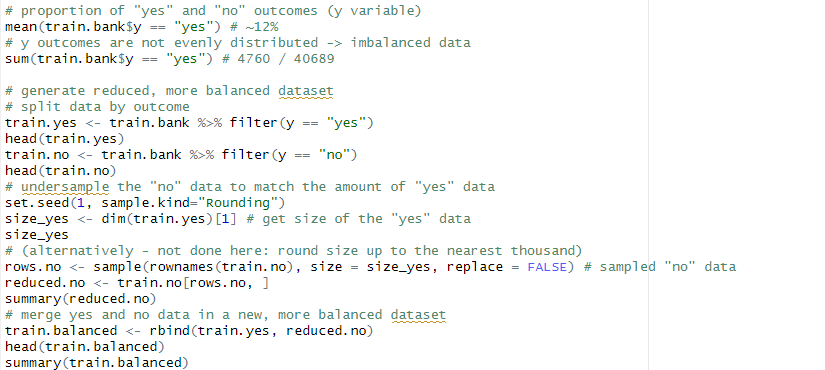
Correlation

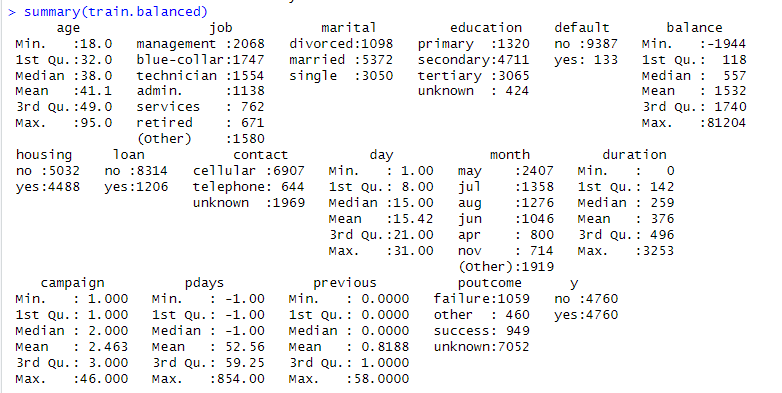


# Sampling data

As previously mentioned, the dataset is imbalanced and contains only about 12% "yes" outcomes. However, these outcomes are the ones that are relevant from a business standpoint, as the objective of sales/marketing campaigns is to acquire new customers or expand the product range for existing ones. In datasets with a limited number of observations for a particular class, identifying the features that describe that class is more challenging than in datasets where each class has a comparable number of observations. This is because comparable features will be present in the other dominant classes, making it more difficult to distinguish the relevant features.

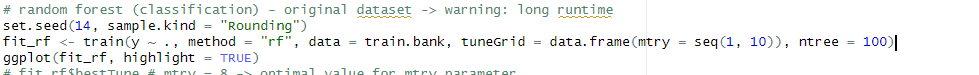
To address this issue, we generate a more balanced dataset by undersampling and use it for modeling in conjunction with the original dataset. The new dataset will include the "yes" outcomes and a similar number of "no" outcomes sampled from the original dataset.

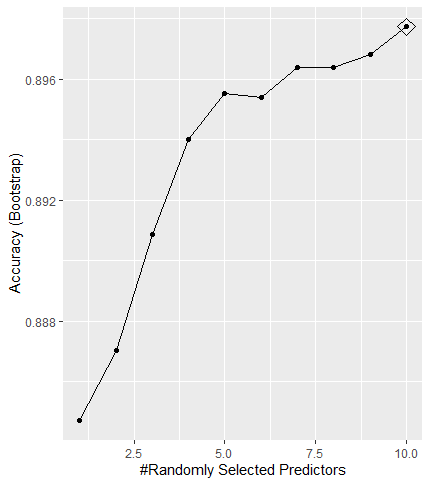


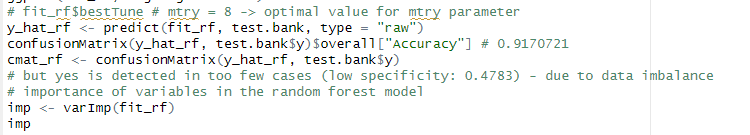


# Building the model using training dataset

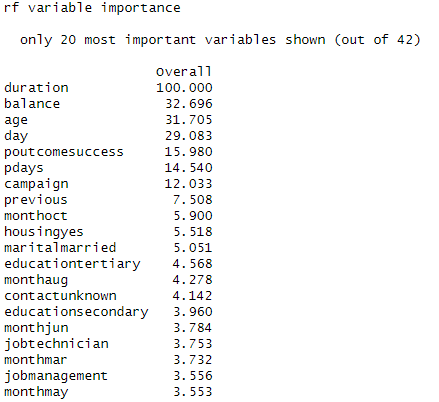
When it comes to this type of problem, random forests are a popular option. Random forests average predictions over multiple trees and can produce a more dependable prediction. Additionally, the random forest model generates a ranking of feature importance for the selected model, which can aid in comprehending the classification process. This model type can identify non-linear relationships in the data.

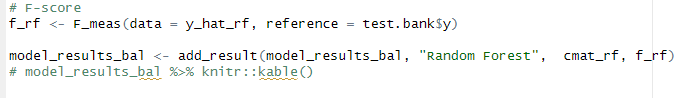




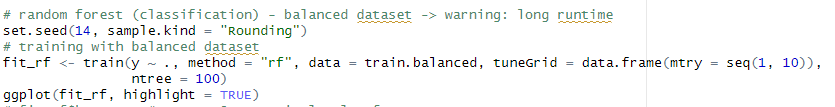


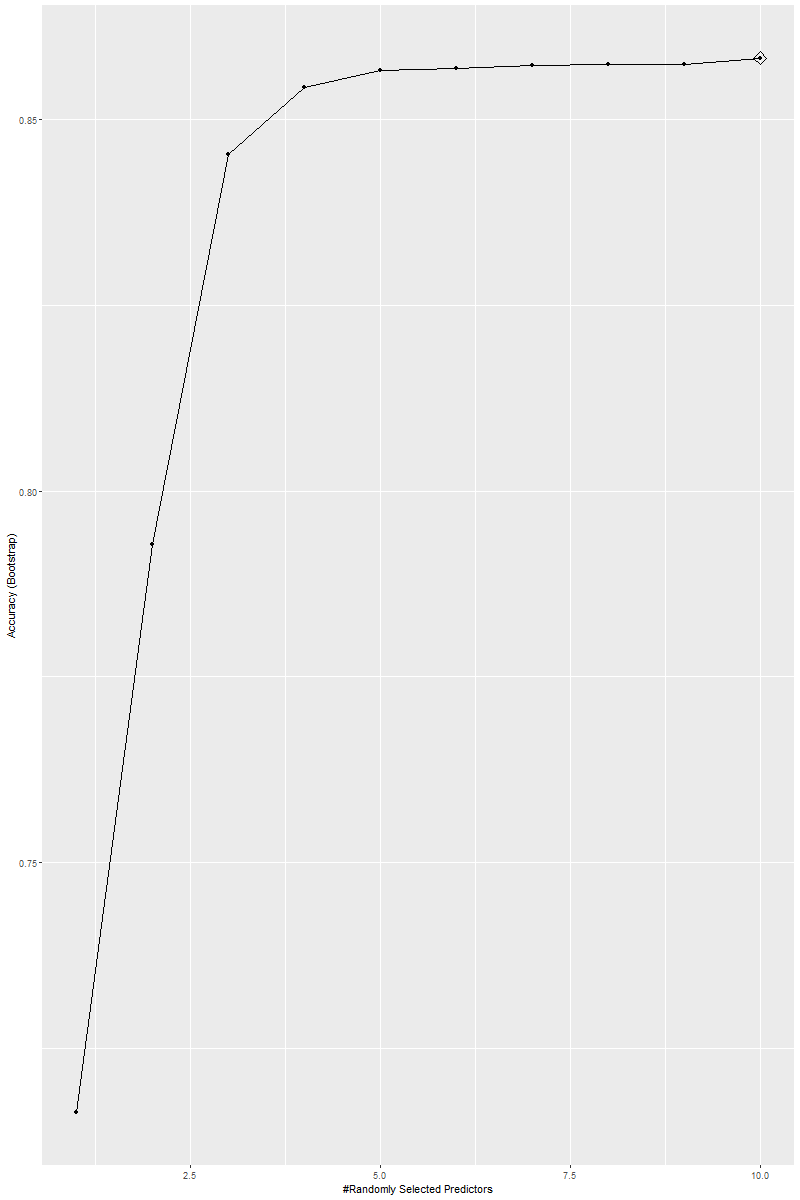




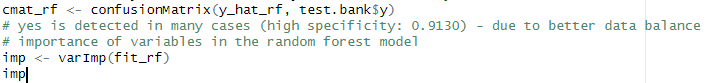


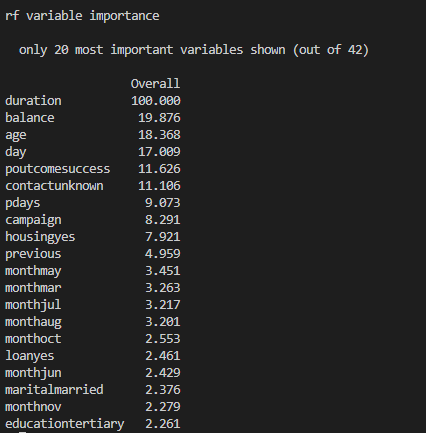
The dataset that is more balanced requires fewer resources and is smaller in size.

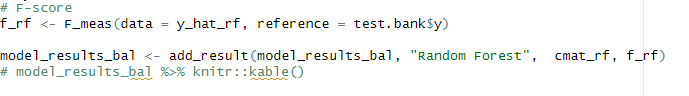












# Conclusion

We analyzed the data from the Bank's marketing campaign and experimented with various techniques to classify customer responses. After evaluating the results, we found that the Random Forest model was the most effective and provided the best balanced accuracy. However, this model can be computationally expensive when dealing with large amounts of data. Although there are other promising alternatives like neural networks or time-series methods that can capture more complex relationships between features and time-related effects, selecting the best model for deployment requires considering factors such as available data, its structure, domain knowledge, and business goals.