**Assignment Submission Cover Sheet**

|  |  |
| --- | --- |
| **Programme Title:** | **MSc Business Analytics** |
| **Module Code and Title:** | **BU7142 -** **Foundations of Business Analytics** |
| **Assessment Title:** | **Group Assignment** |
| **Group Number:** | **Group 3** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Student Name and Contribution** | **%** |  | **%** |
| **1.** |  | **4.** |  |
| **2.** |  | **5.** |  |
| **3.** |  | **6.** |  |

For group work – individual % contributions need to be stated **only** where they **are not equal.**

Please read the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at:

<http://www.tcd.ie/calendar>

Online Tutorial on avoiding plagiarism ‘Ready, Steady, Write’, is located at

<http://tcd-ie.libguides.com/plagiarism/ready-steady-write>

**Introduction**

Brief of the Report

The dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls.

In response to the data provided, we put forward the following three questions for analysis：

1. Do certain points in the year or month affect a person's decision to subscribe to a deposit? Additionally, are there certain points in the year with more purchases?

2. The factors affecting a client’s balance.

3. What type of person is most likely to subscribe to a term deposit (factors age, marital, education, employment, homeowner)? Which type of person should we target when marketing to subgroups of people**.**

The main purpose of our research on the above three questions is to find out the target group that affects bank deposits based on the data, and the best time for bank deposit marketing, and then formulate a more detailed marketing strategy based on this, and finally achieve the goal of increasing bank deposits.

An Overview of the Report

The contents are as follows:

* **Introduction:** Brief introduction about dataset and questions we mentioned
* **Analysis:**

**- Method of analysis:** Method and model used for every question, based on the analysis of the dataset, identify the most appropriate statistical model.

**- Coding and visualization:** Detailed coding and related explanations for each step of analysis.

**- Insights from graph:** After visualizing the data results, develop several insights from tables or histograms which can contribute to the marketing campaign.

* **Conclusion:** Draw an overall conclusion based on analysis and come up with a feasible marketing strategy for the bank.

**Analysis**

**Q1：Do certain points in the year or month affect a person's decision to subscribe to a deposit?**

**Method of analysis**

* Histograms counts, and proportions on a per month basis based off the outcome variable y (Yes/No). Y was first separated in the data set and then the distribution of the corresponding frequencies in each month was presented in the form of a histogram. However, examining only the frequencies only gives limited information, so the proportion of Yes/No in each month extends the analysis.
* Correlation testing was conducted by creating dummy variables and assigning 1/0 to the categorical variable of each month.
* Logistic regression was run to test the significance of the variables; a chi-square test was used to obtain the results.

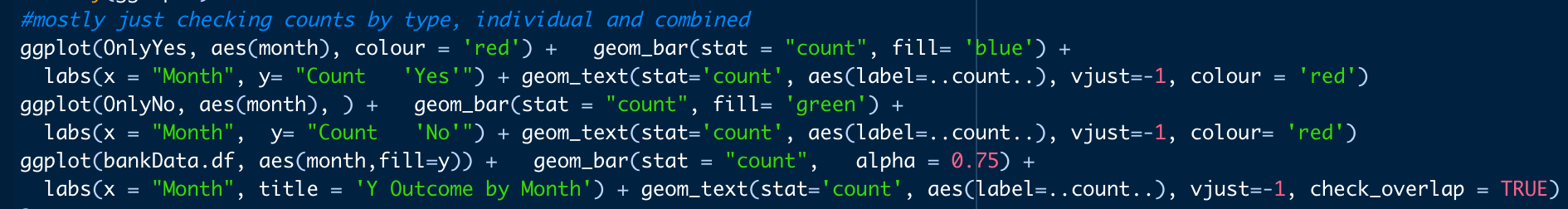
**Analysis of Codes and Visualization**

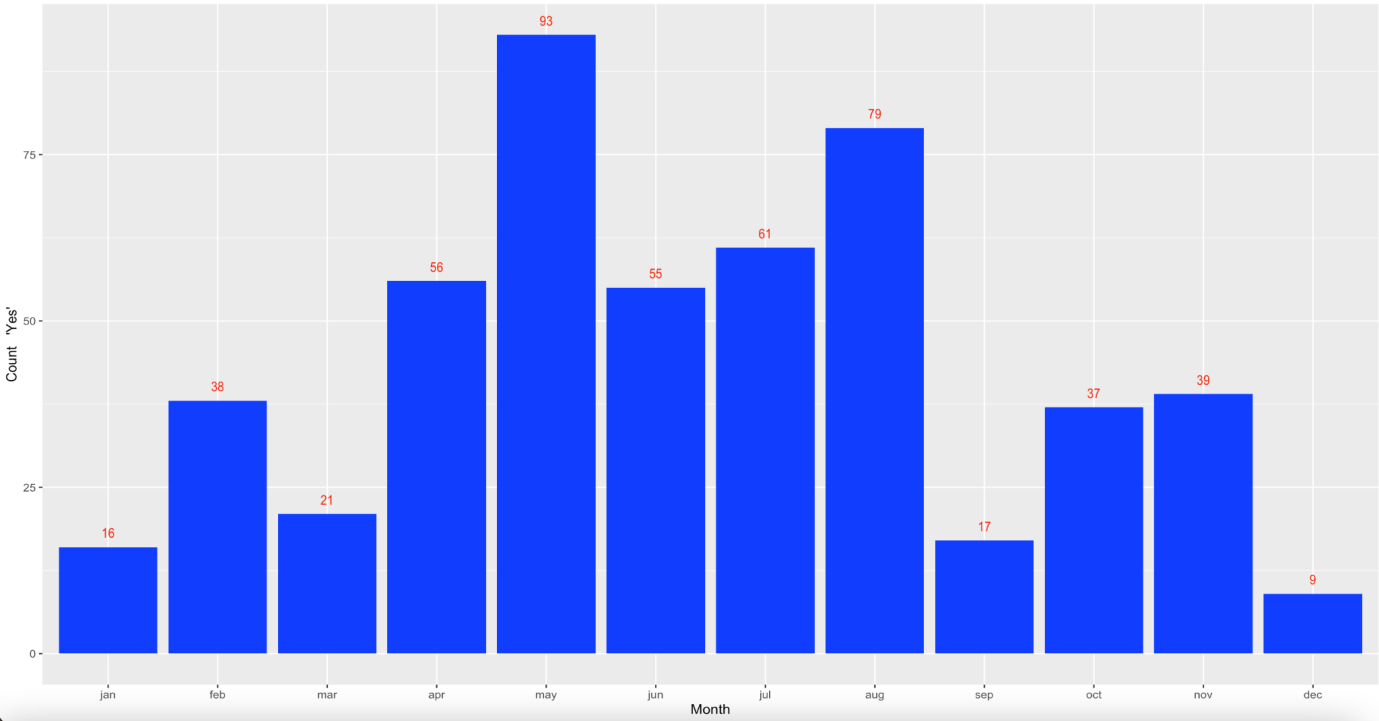
This code looks into the effect of months on sales and distributions etc.

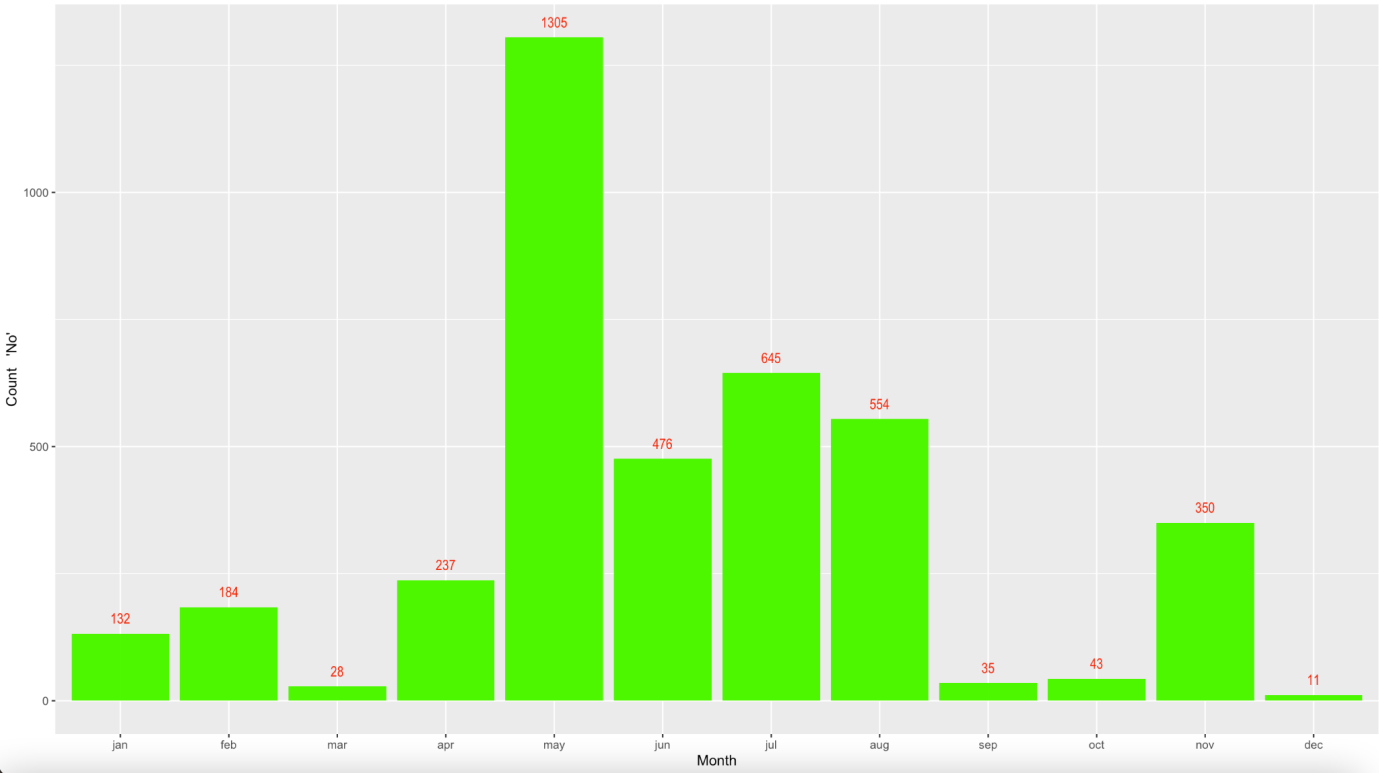
Partition data into ‘yes and no’ y outcomes then we can see frequency distribution:

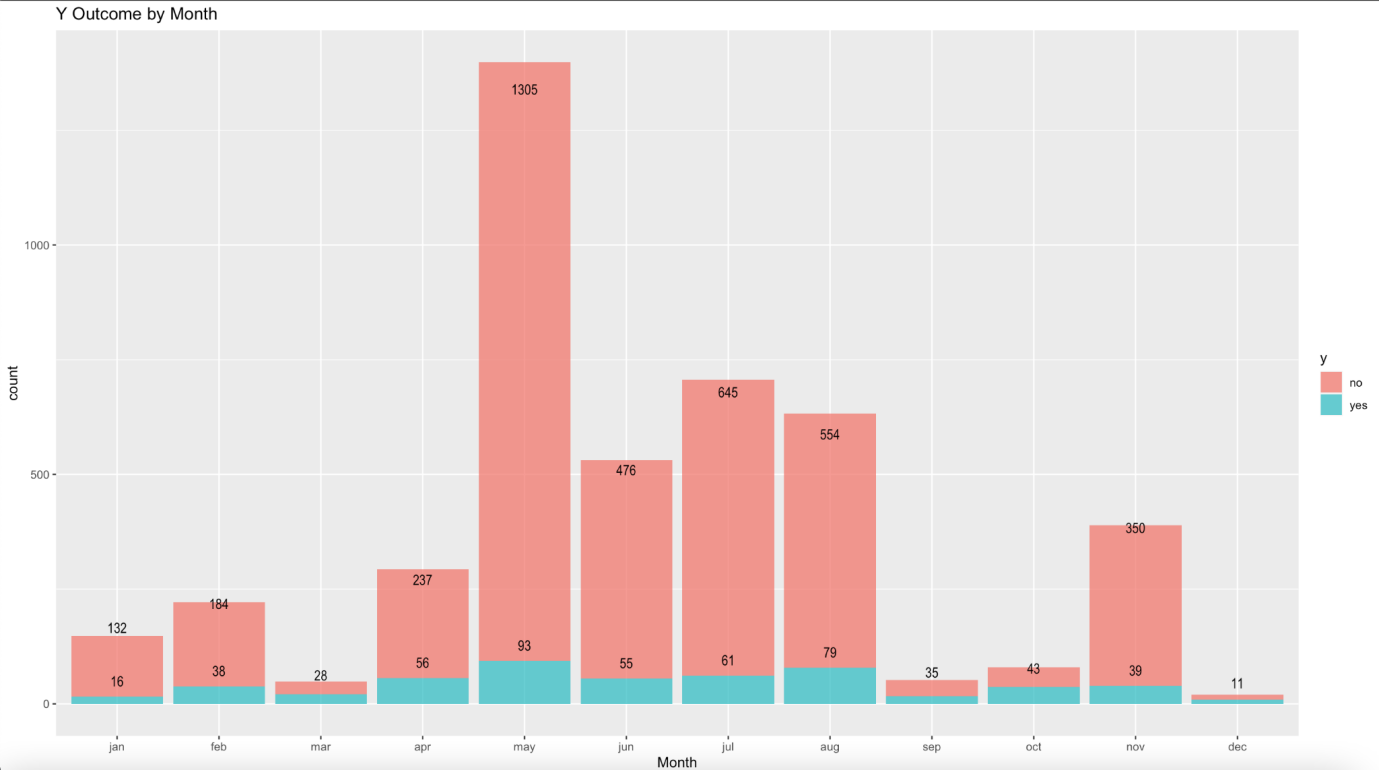


Check frequency yes/no counts by month.



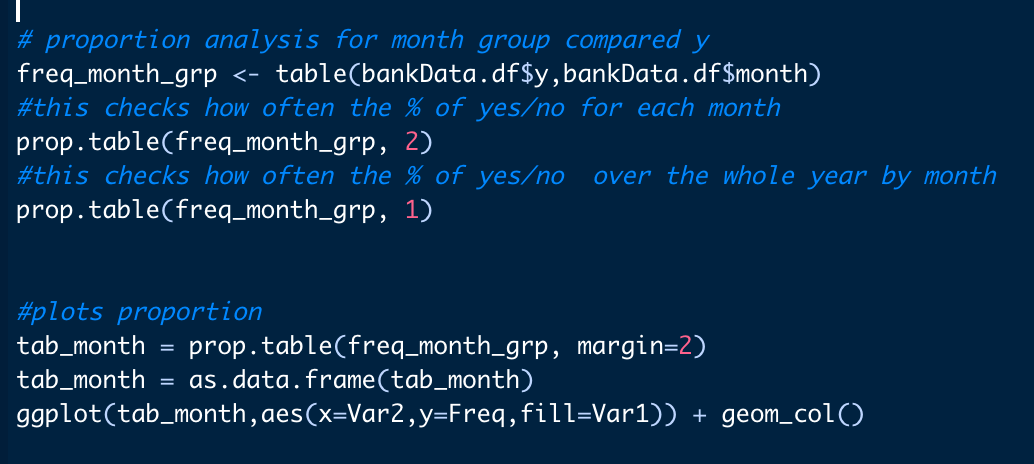




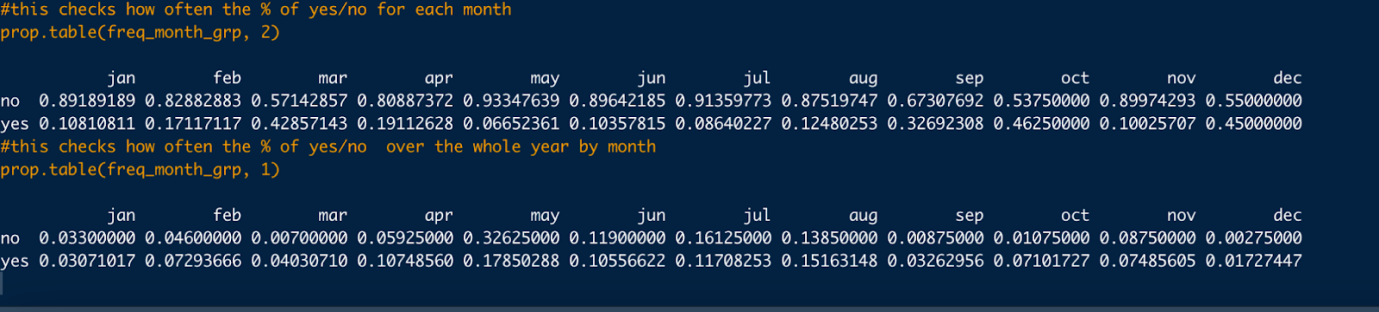


According to the count histogram, May has the highest number of yes's, i.e. the highest number of depositors; but correspondingly, May also has the highest count of no's, i.e. those who did not deposit. It is therefore not enough to focus on the absolute value (i.e., count), we still need to look further at the proportion of yes (i.e., frequency) to determine whether May should be a priority focus month.

Below, analysis is performed on the proportion for month group compared to y, this checks how often the percentage of yes/no occurs for each month over the whole year.



The proportion table is listed below, however is more effective in a bar chart.

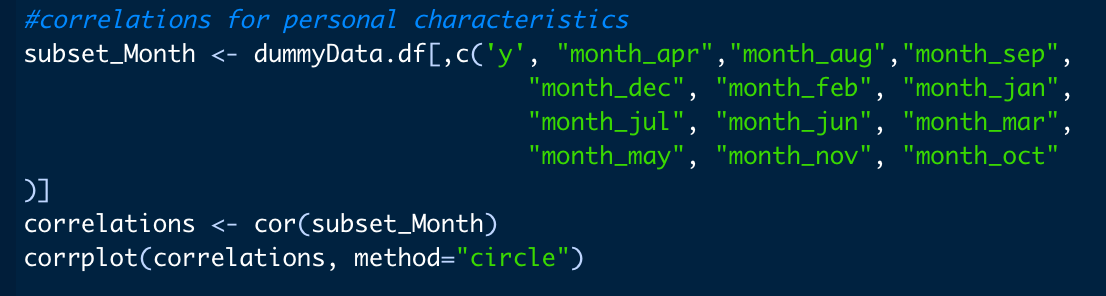


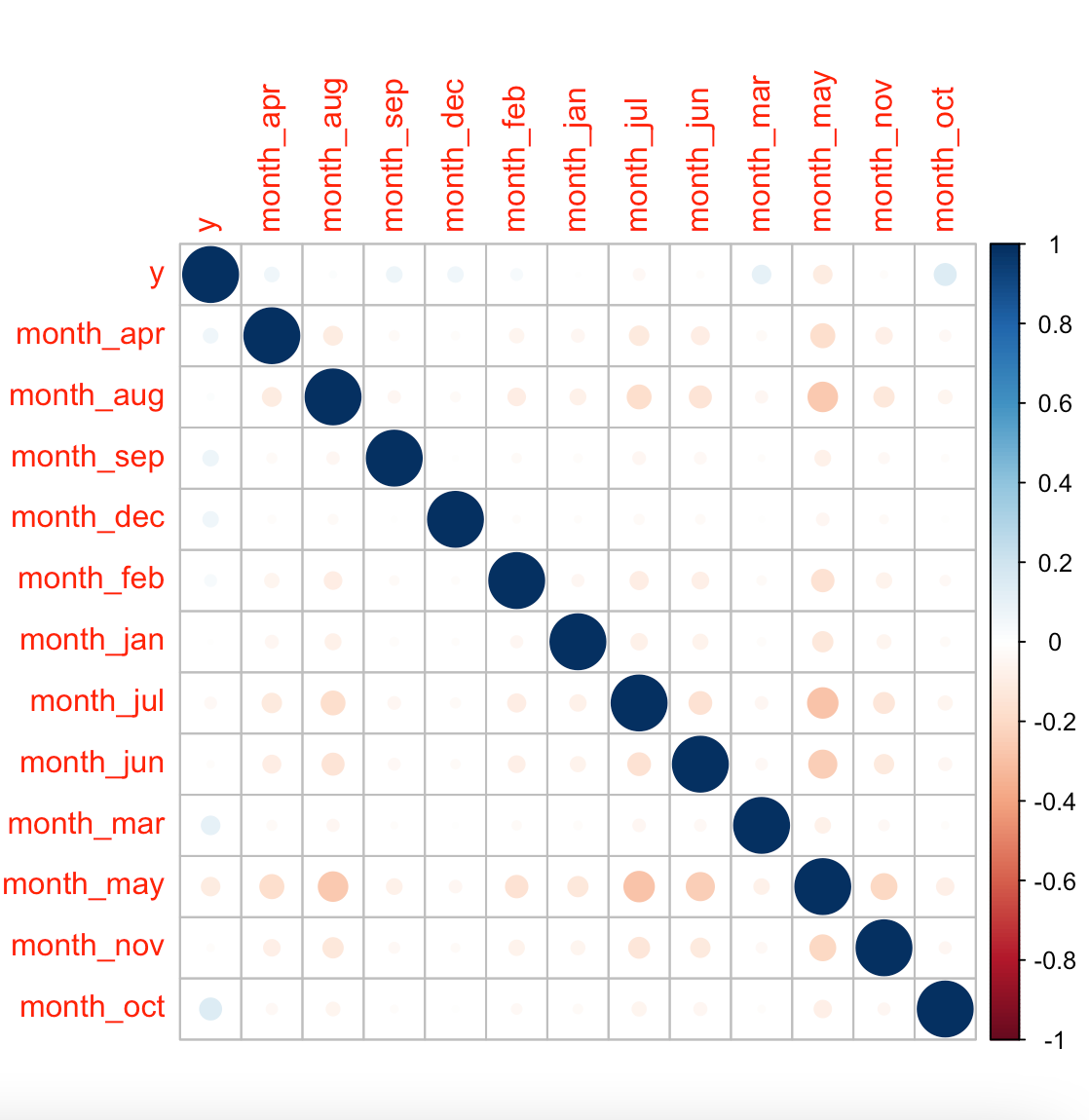


From this we see the highest frequency of yes occurs in the months of December, October,September and Mar. However considering the counts from the histograms above, we need to be cautious about its actual affect on outcome variable y as the data is limited compared to months like May.

Next, run correlation to see how everything is related. Firstly, create dummy variables and compare them to y.

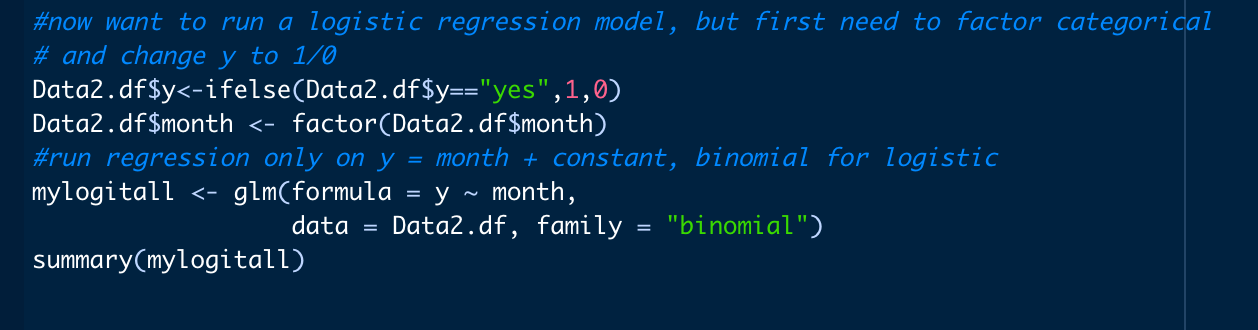
Correlations for personal characteristics:

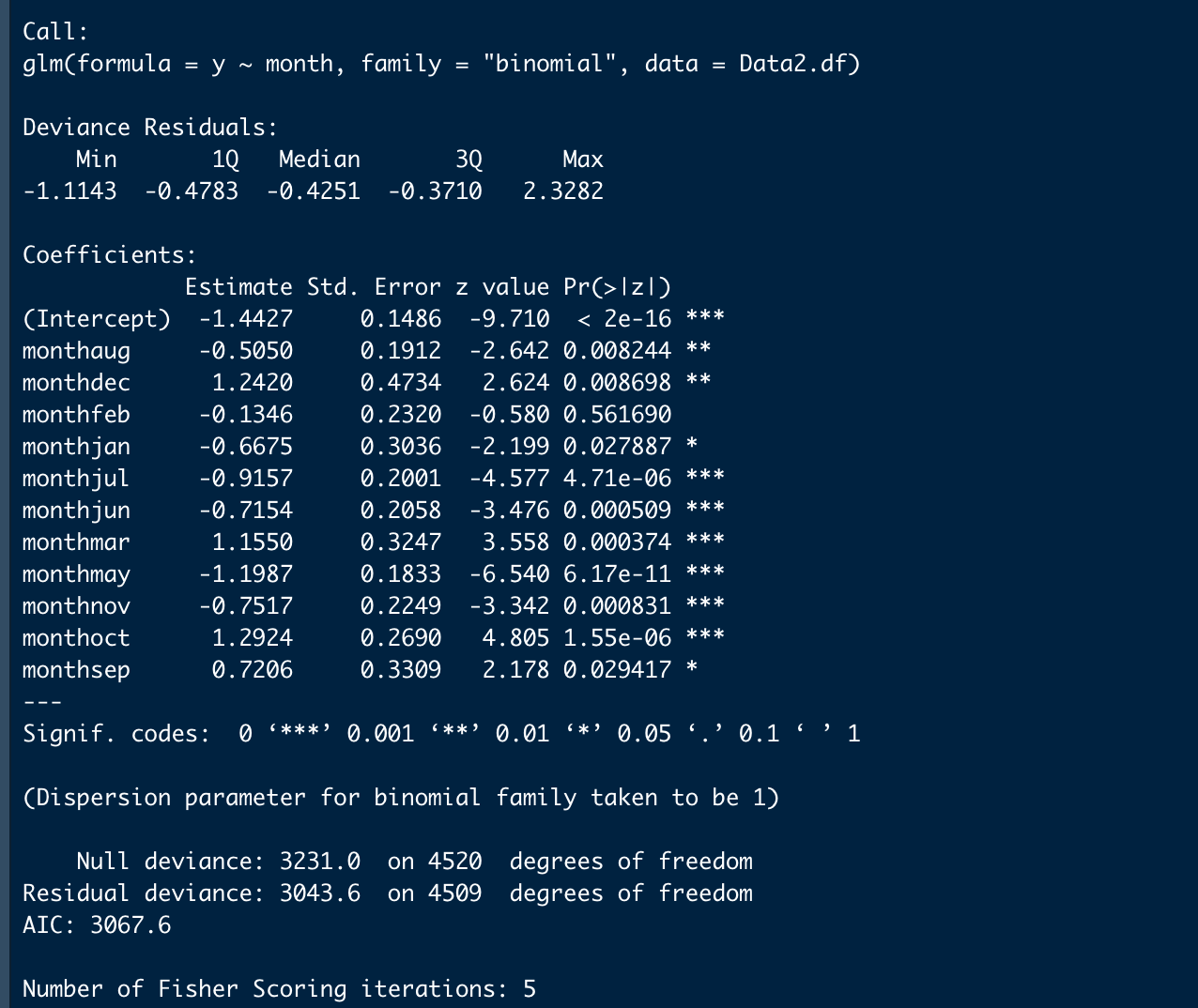


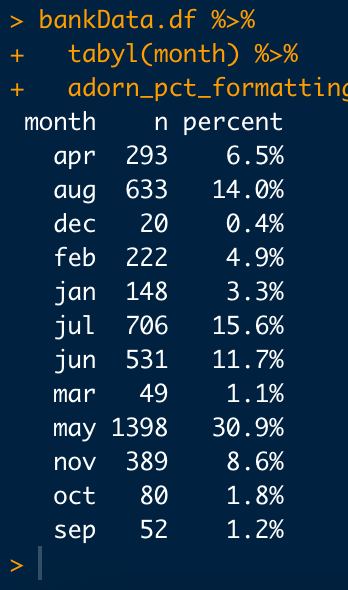
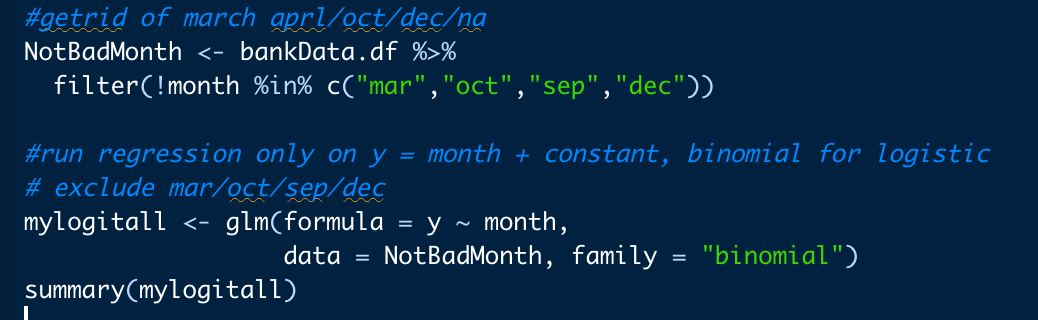


We mostly care about the y column, with months. From this we see that Oct, Mar, Sep, Dec have positive correlation with the outcome variable y, whereas May and July have some negative correlation with y.

Next, running a logistic regression test to see which variables are significant and create a hypothesis test. But first change the classification factor and y to 1&0 and then run the regression for y = month + constant only.

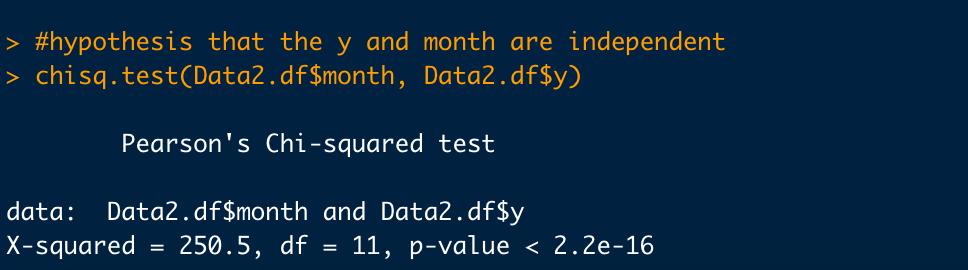




****

Once again, intercept is significant, take hypothesis tests from that or not. Additionally run chi square, take insight as desired.

Hypothesis that the y and month are independent:

****

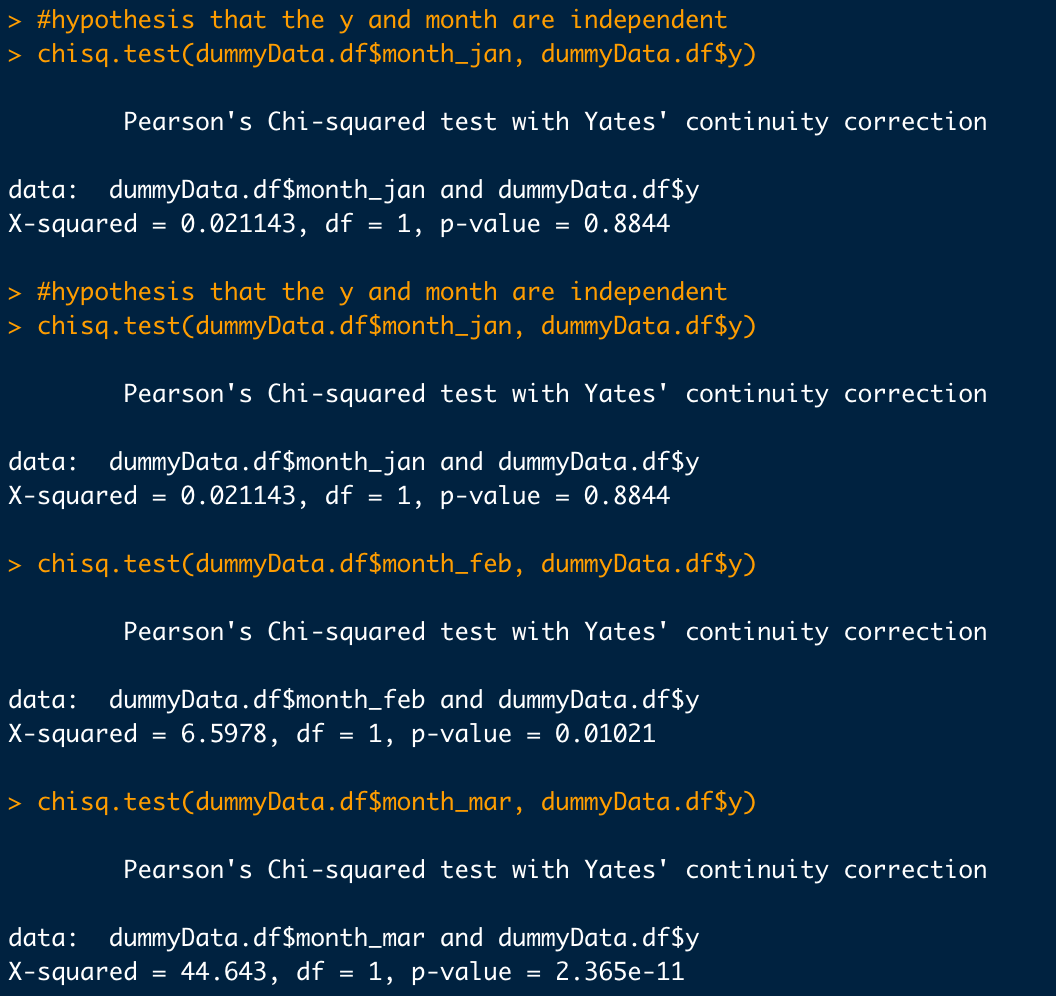
A chi-square test was conducted on each month's data to investigate whether the month was significantly correlated with the Y variable.

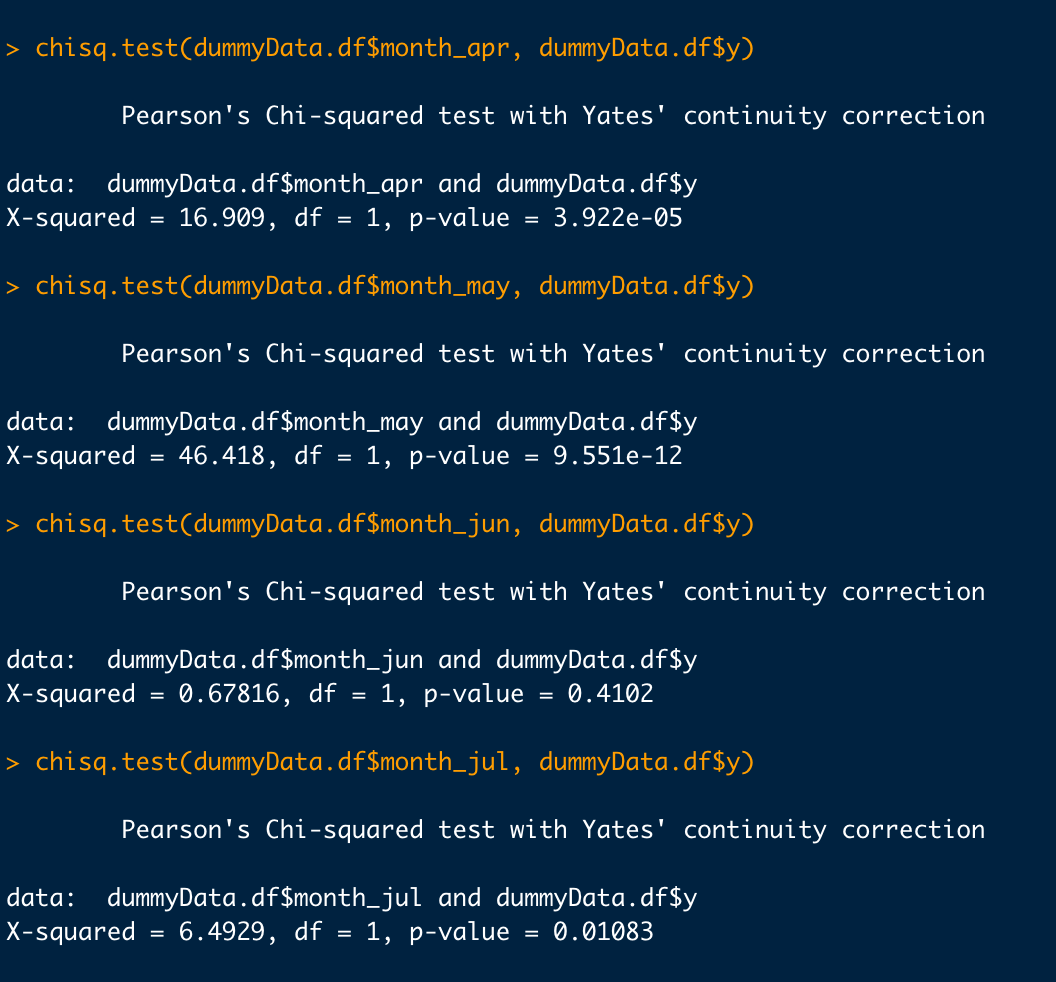
hypo:  
H0：the y and month are independent;

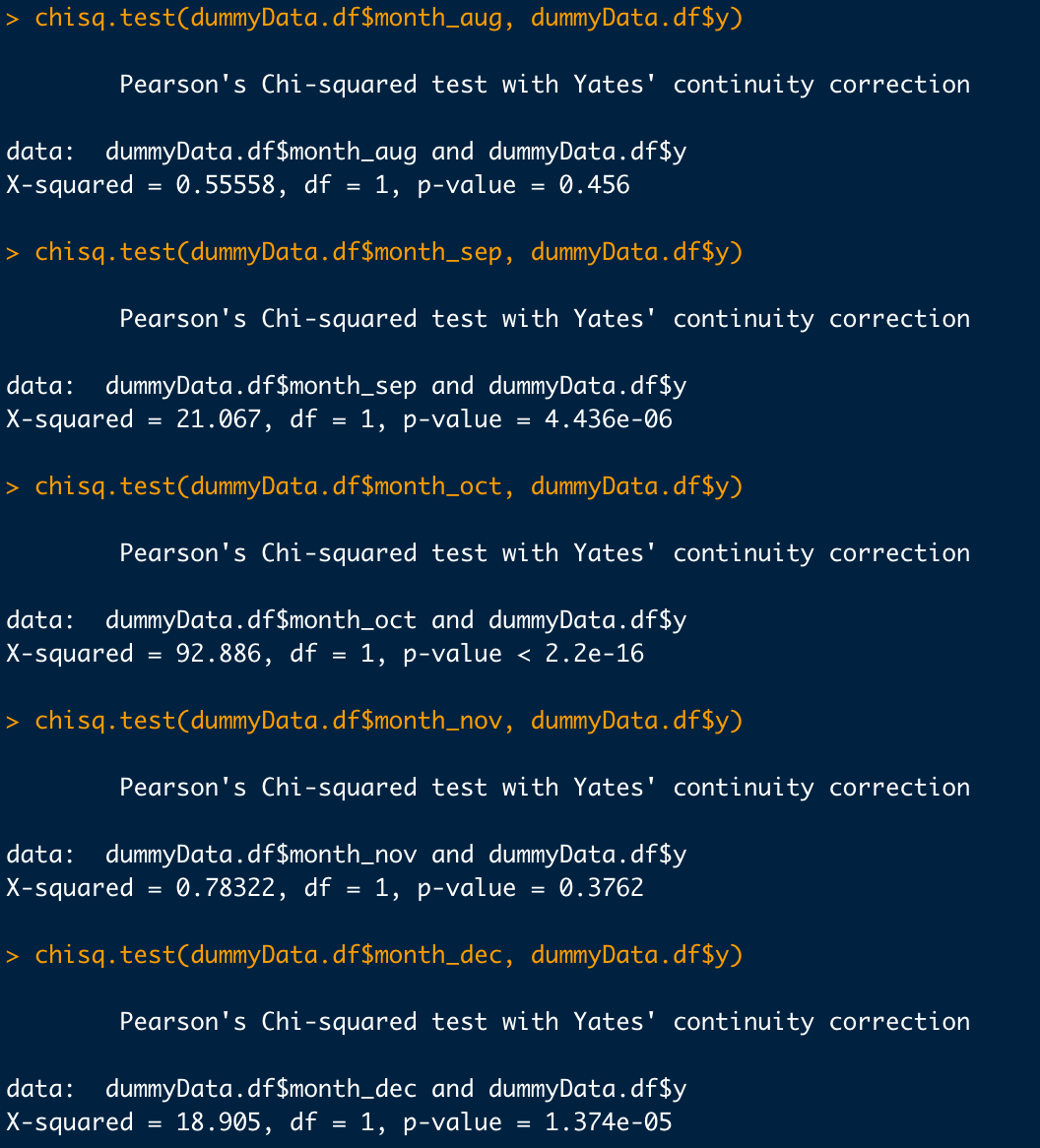
H1：the y and month are correlated (not independent)

The p-values for each month were calculated using R. The months with p-values below 0.05 were selected as having a significant correlation with the Y variable.

****

****

****



From this we conclude that February / March / April / May / July / September / October / December and y are significantly correlated.

**Insights from the graphs for Q1:**

According to the P value, February/March/April/May/July/September/October/December is significantly related to y.

Through data analysis, it is found that October and December are the months with the largest proportion of successful deposits, which means that more people will have a successful deposit in banks in October and December of the year.

For banks, they can consider conducting more promotional activities and customer return visits in October and December to attract customers to make deposits, increase customers' willingness to deposit and deposit amount, and further increase the bank's deposit amount and loan scale.

A limitation to the October and December “sales” months, would be the limited amount of data, where proportionalities may significantly be affected by reduced sample size.

**Q2: Factors affecting a client’s balance**

**Method for analysis**

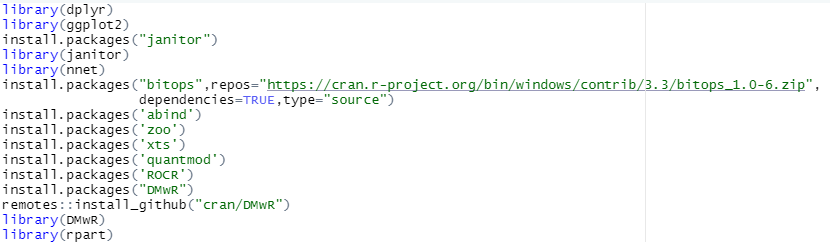
1. Before data cleaning: In the first question, we have carried out simple data observation and data analysis.

2. Data mining: After importing the data, delete values with very few magnitudes, use the classification method in the repeated multiple complement categorical variables to predict, predict the binomial classification method in order, and then sort the categorical variable coding, and disorder individual Dummy variables, variability of individual variables

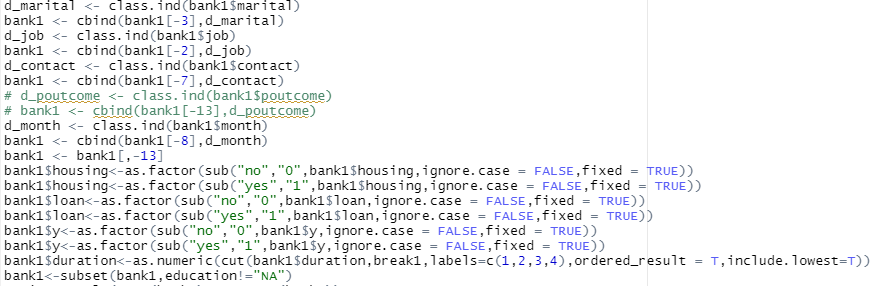
3. After data tracking: use linear regression to model first. Due to variable reasons, use a step wise iterative method, through AIC variable variables, and finally select a suitable variable method

**Codes and Analysis**

The following is the process of data cleaning (refer to the notes for a detailed explanation).

****

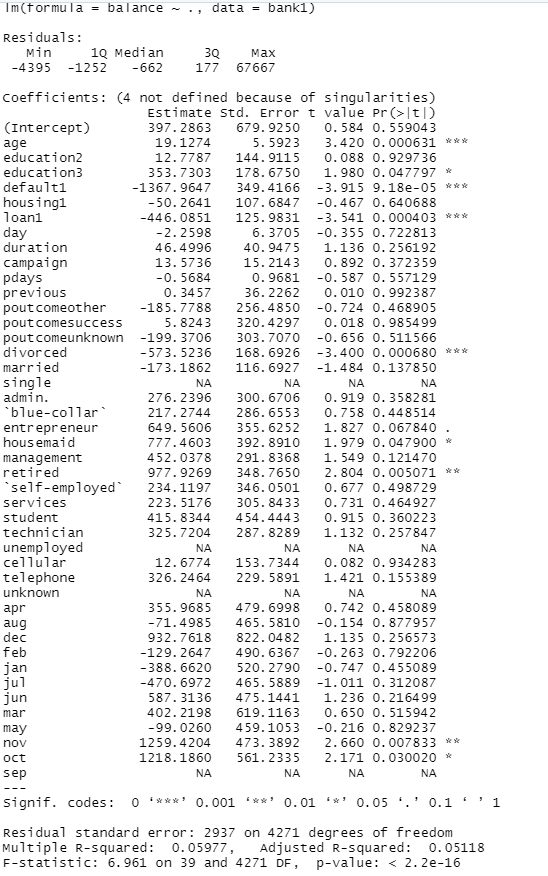
****

****

Next, a linear regression is fitted to the balance column:



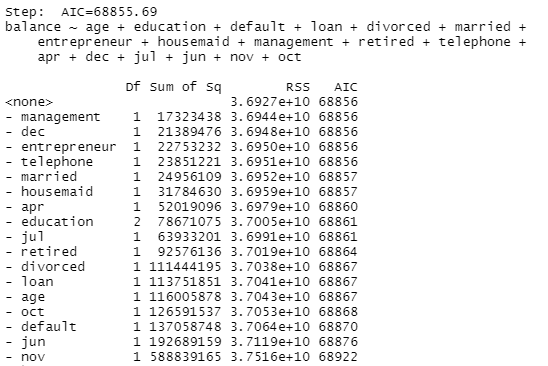
( See the next page)



As there are too many variables, a stepwise regression method is considered here to filter the variables by AIC values to finally obtain the appropriate variables:

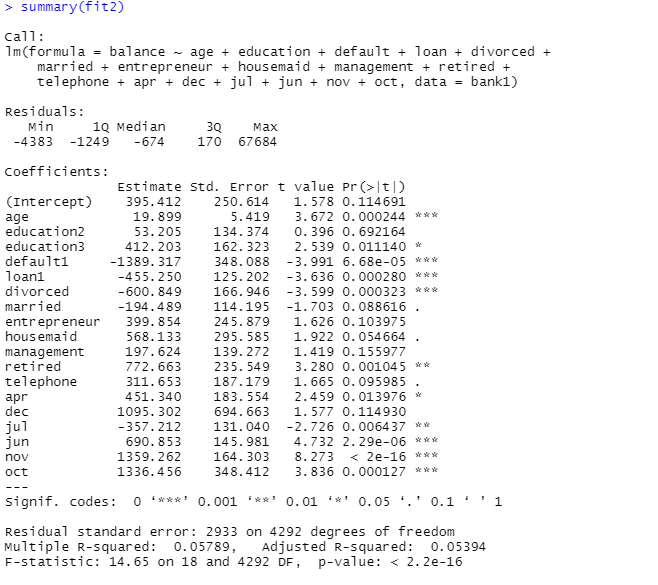


Find the smallest AIC value:

****

Several independent variables obtained according to the stepwise regression method were re-regressed.

****

****

**Insights from the graphs**

According to the P value, the factors that affect the balance are: age, education 3 (tertiary), default is yes (credit in default), loan is yes, divorce, retirement, April, July, June, November, October.

1. According to the data, as people grow older, their deposit balance will continue to accumulate.
2. If a customer has a history of credit default, his/her deposit balance will be lower.
3. If the customer is divorced or divorced many times, the deposit balance will decrease accordingly.
4. If the time of the last contact with the customer is in April, June, November, October, the customer's deposit balance will have a positive correlation change; if the time of the last contact with the customer is July, the customer's deposit balance will have a negative correlation change. This affects of months on balance is more difficult to explain a reason why, but seems to provide statistical insight nonetheless.

**Q3. What type of person is most likely to subscribe to a term deposit (factors age, marital, education, employment)? Which type of person should we target when marketing to subgroups of people.**

**Methods for Analysis:**

1. Combined with the methodology of the previous two questions, the frequency and proportions of each variable were analyzed using categorical histograms.

2. This is followed by hypothesis testing of the data using logistic regression to derive the correlation between the variables and whether a deposit is made, to determine who is most significantly likely to subscribe to a time deposit.

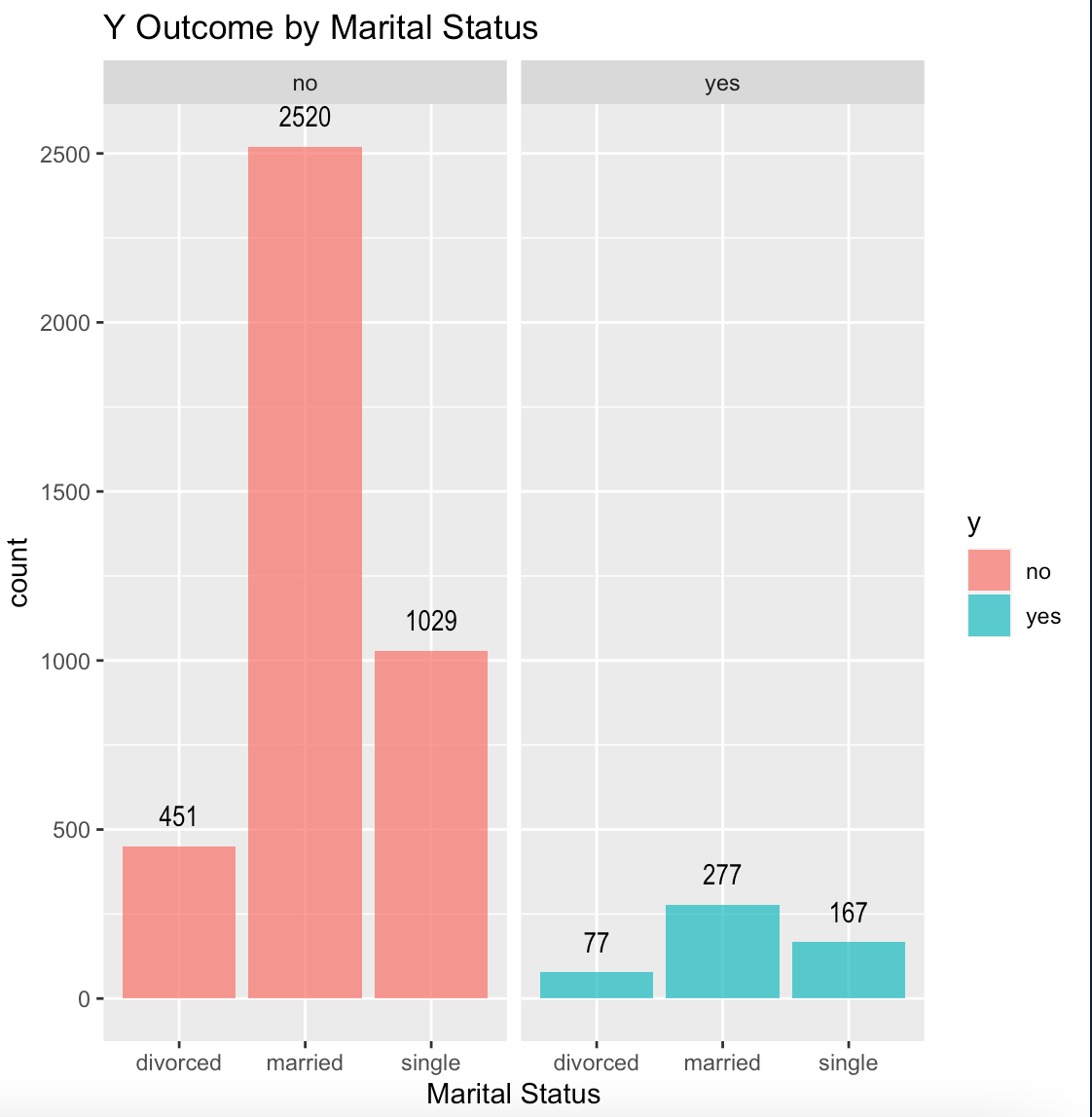
3. Finally, this was interpreted in terms of people's habits and marketing strategies (e.g., if older people are more likely to subscribe to time deposits, then we should spend more effort marketing to older people)

The following personal characteristics are looked into:

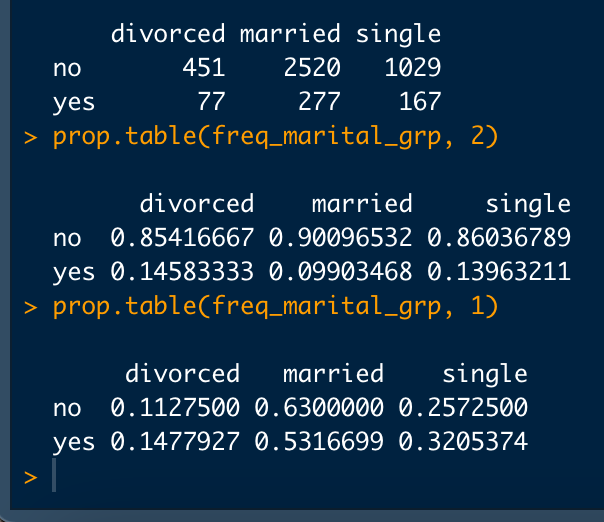
**1. Marital:**

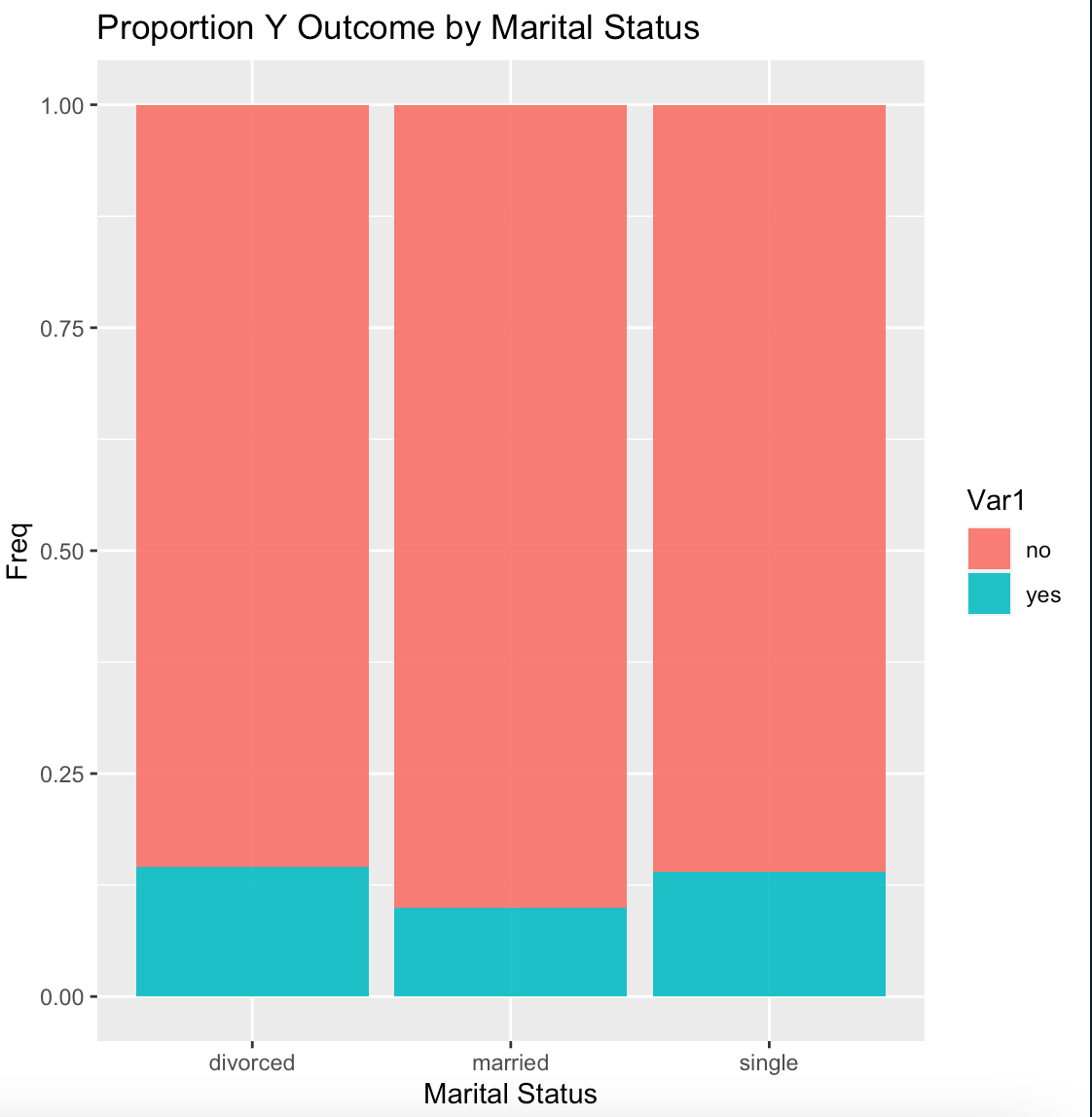
Analysis of the relationships between marital status and Y:





The proportion of people who have been a success for outcome variable y:





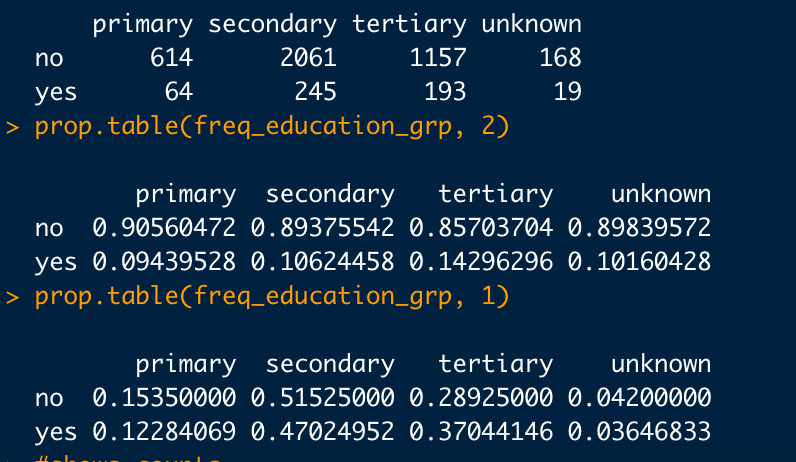
Marital Status Perspective:

From the frequency histogram distribution, married people are the most often to result in a successful deposit, followed by singles. Divorced people have the least deposits, but the frequency is not intuitively expressed, so we need to analyze the proportional histogram to conclude that divorced people are the most likely to have a successful impact on y.

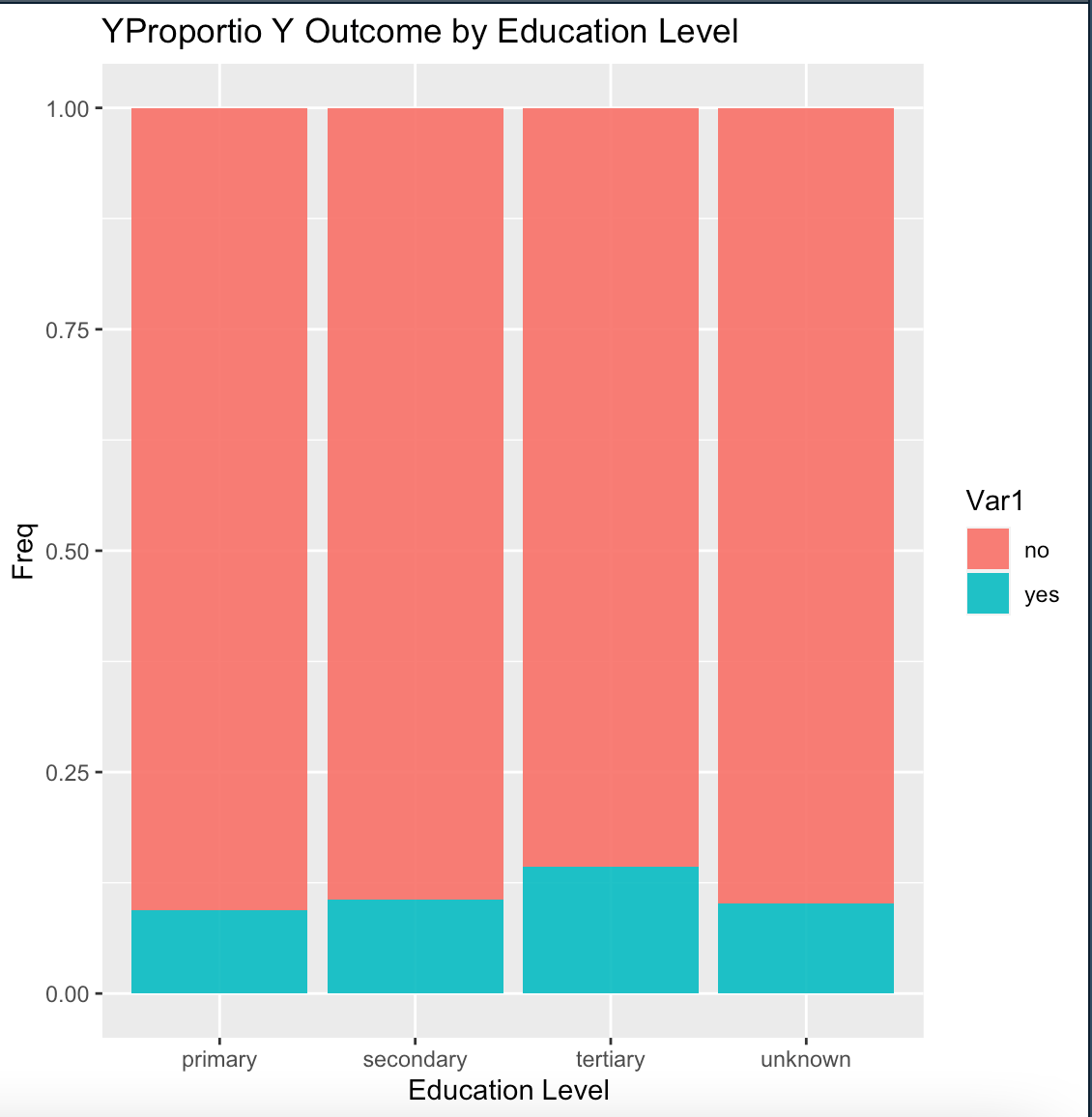
**2.Education:**

Analysis of the relationship between educational status and Y







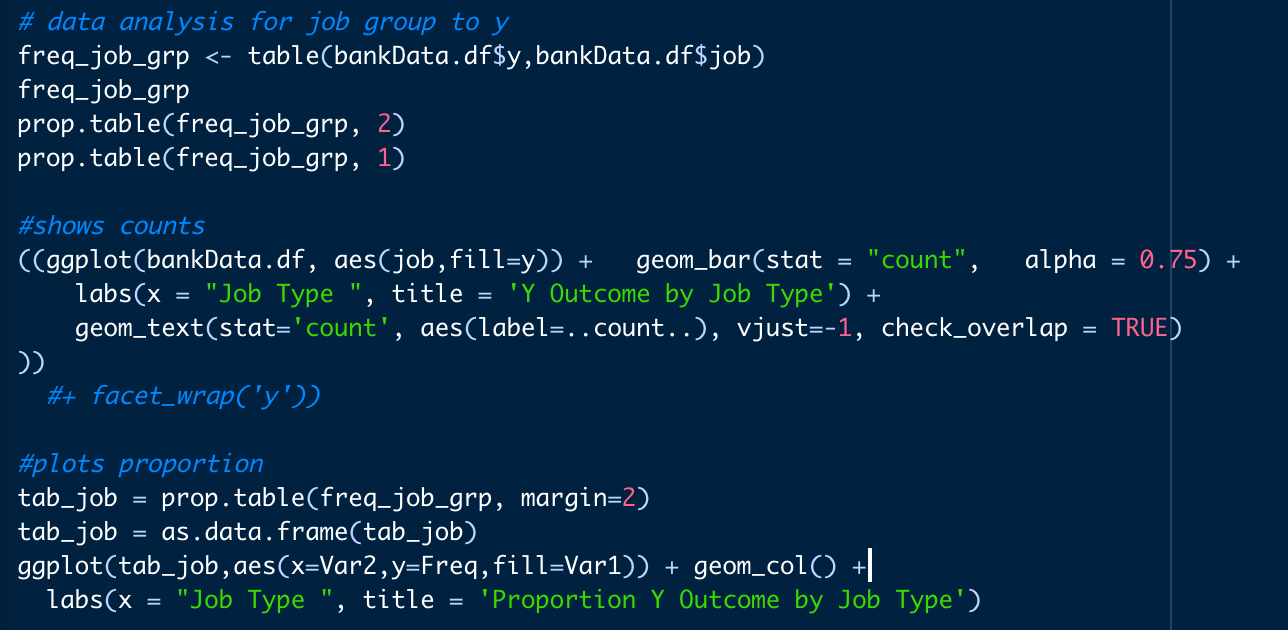


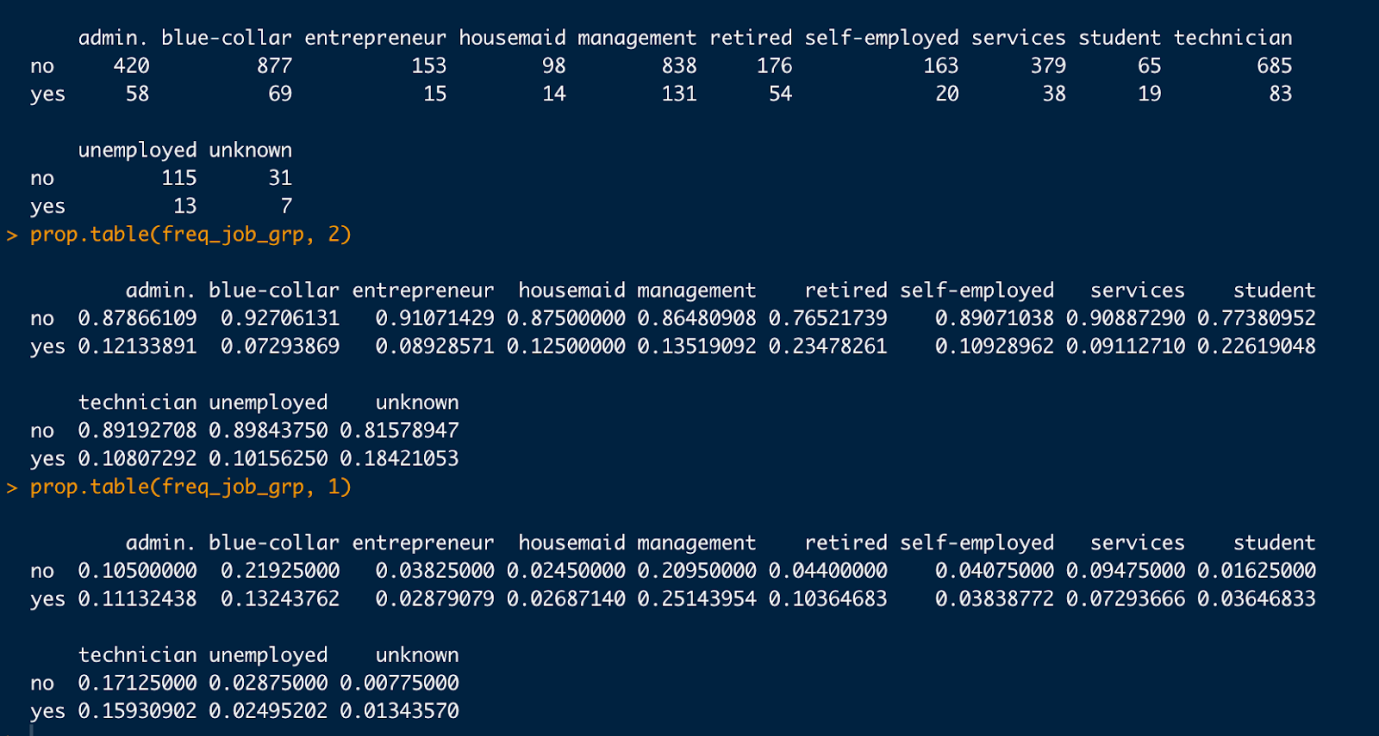
Education Perspective:

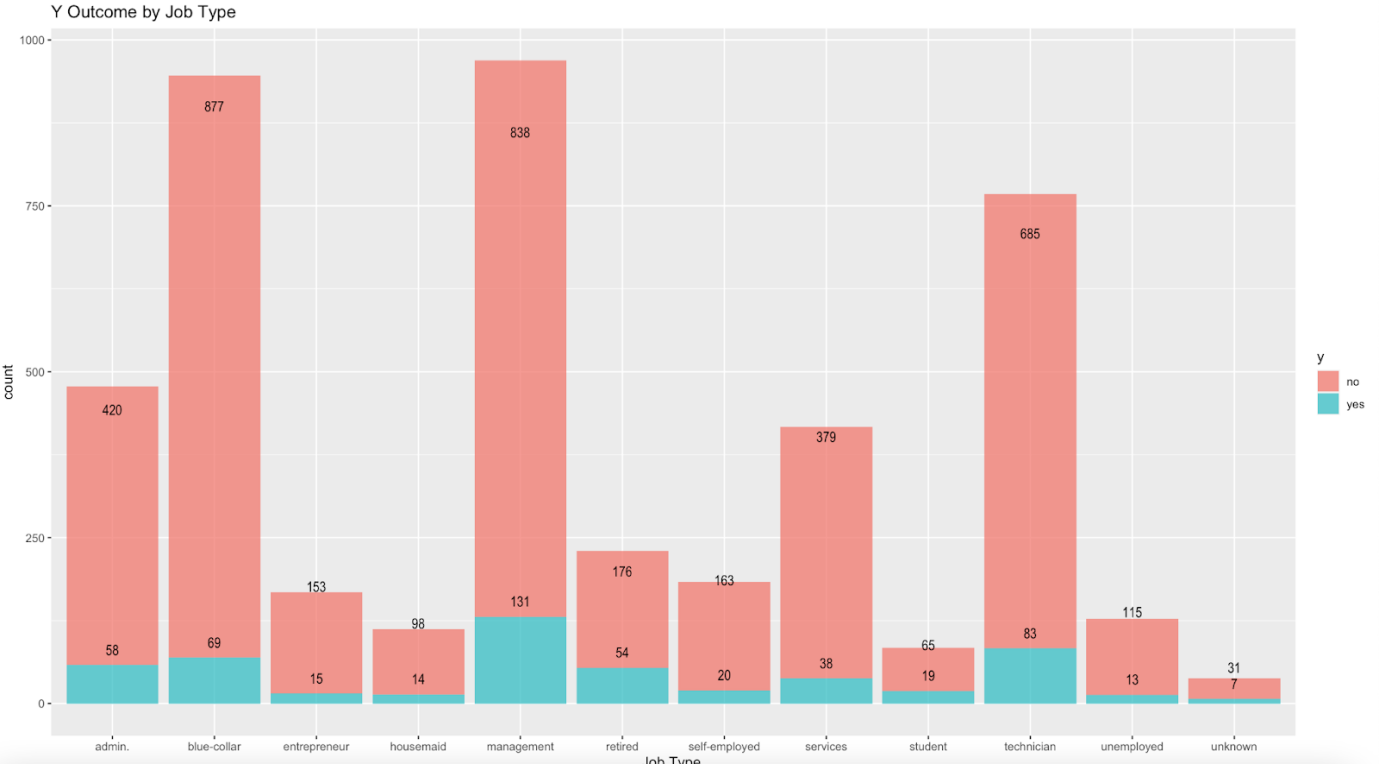
From the frequency histogram distribution, the highest count of successful deposits is a secondary education level, followed by the tertiary education level. The frequency is not intuitively expressed, so we have to analyze the proportion ratio graph, we can get the education level tertiary Yes, it is more likely to make a deposit. As education level increases, the likelihood to have a successful deposit seems to increase.

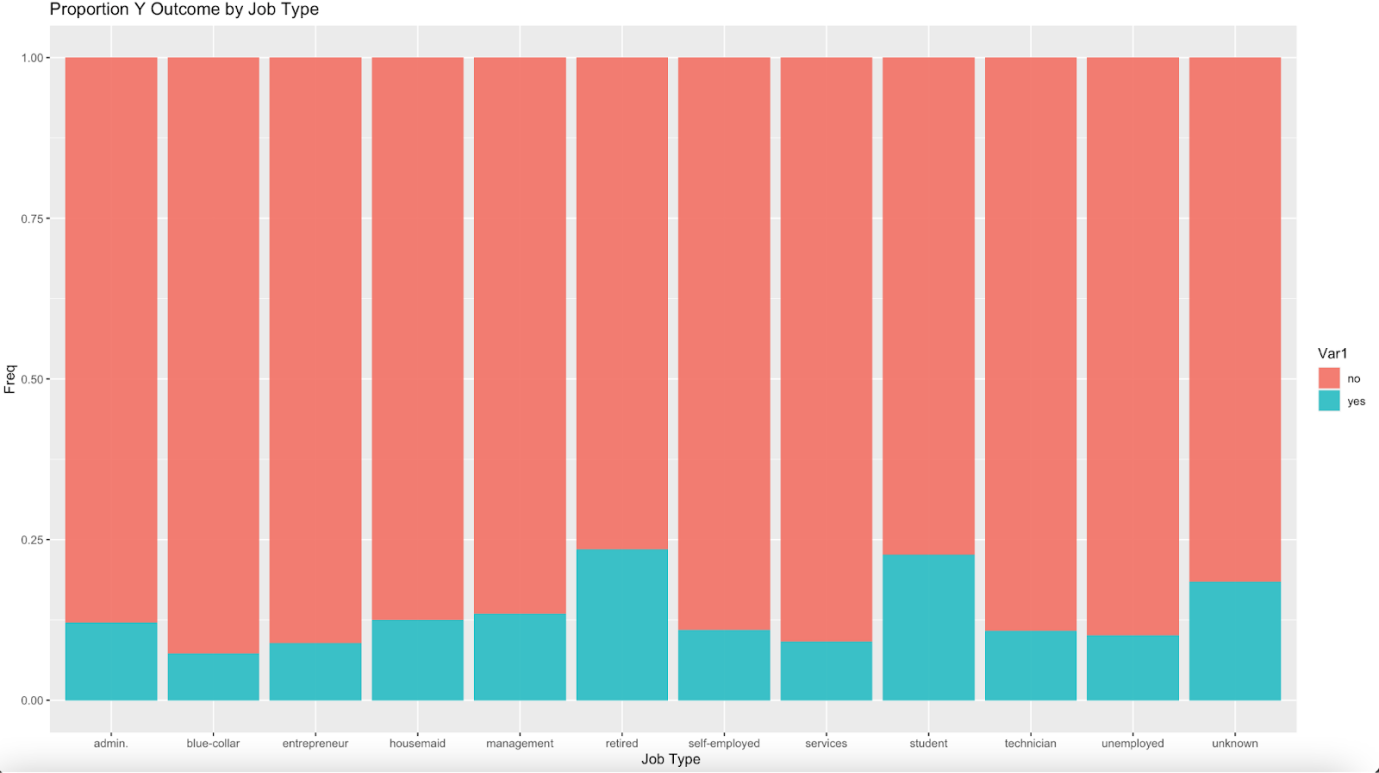
**3. Job:**

Analysis of the relationship between job status and Y









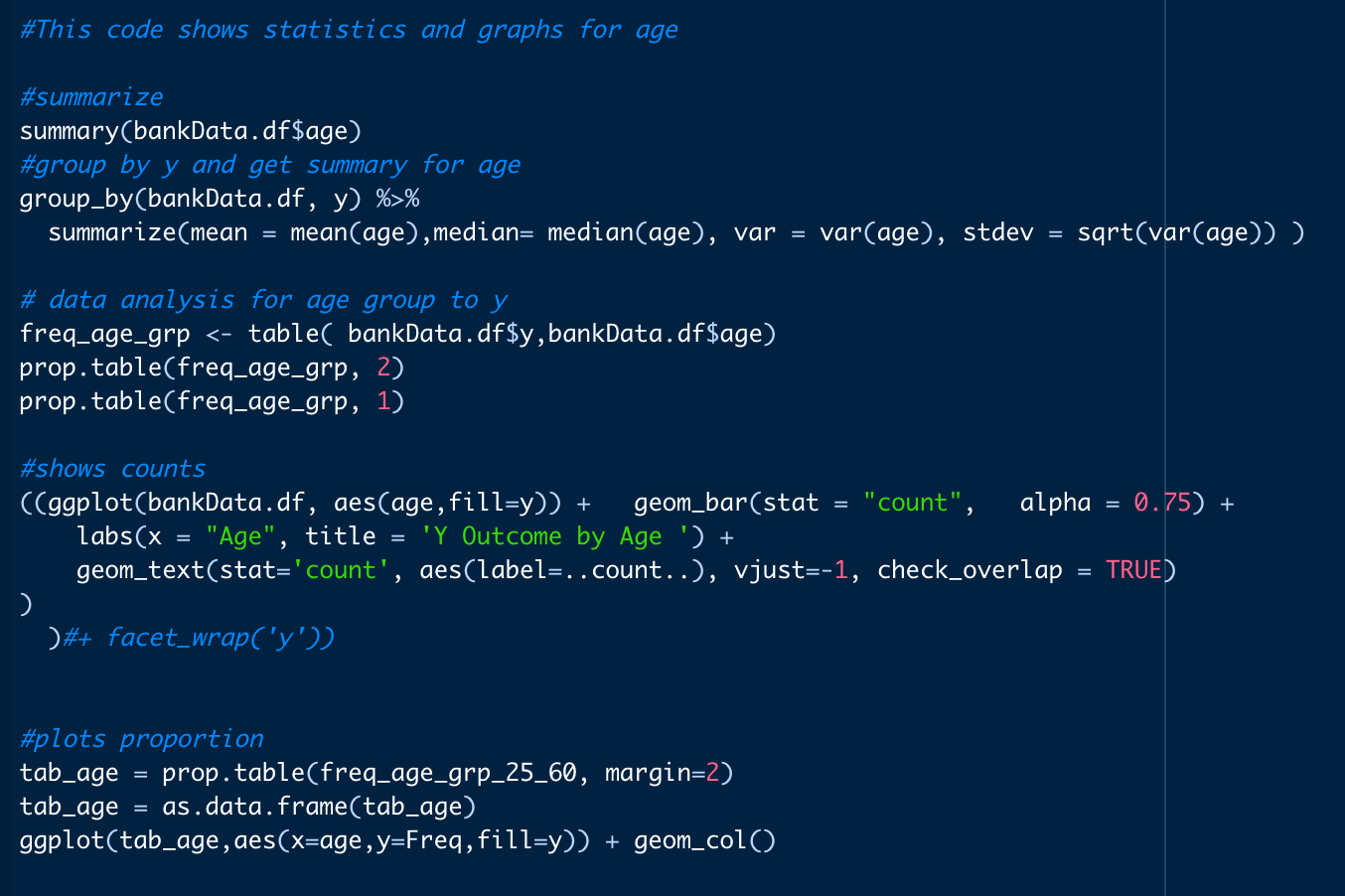
Job perspective:

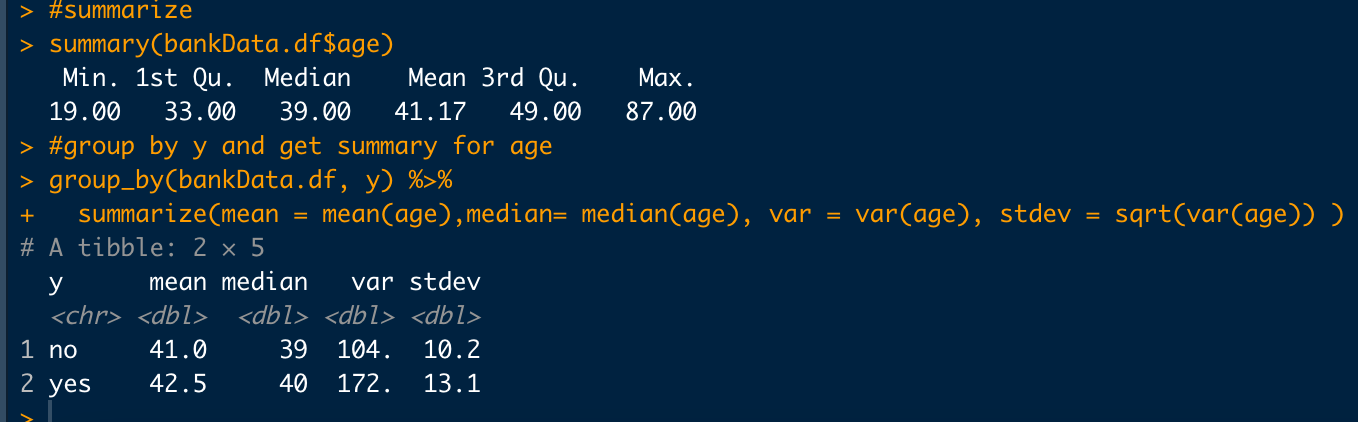
From the frequency distribution histogram, the top three people with the most deposits are management, technician, and blue-collar, and the least three are the unemployed, housemaid, and entrepreneurs (there is not enough “unknown” job type data, so we eliminated them).

From the frequency distribution histogram, the top three people with the highest deposit ratio are retired, student, and management (“unknown” data is less, so we eliminated them). This will be discussed further in analysis below.

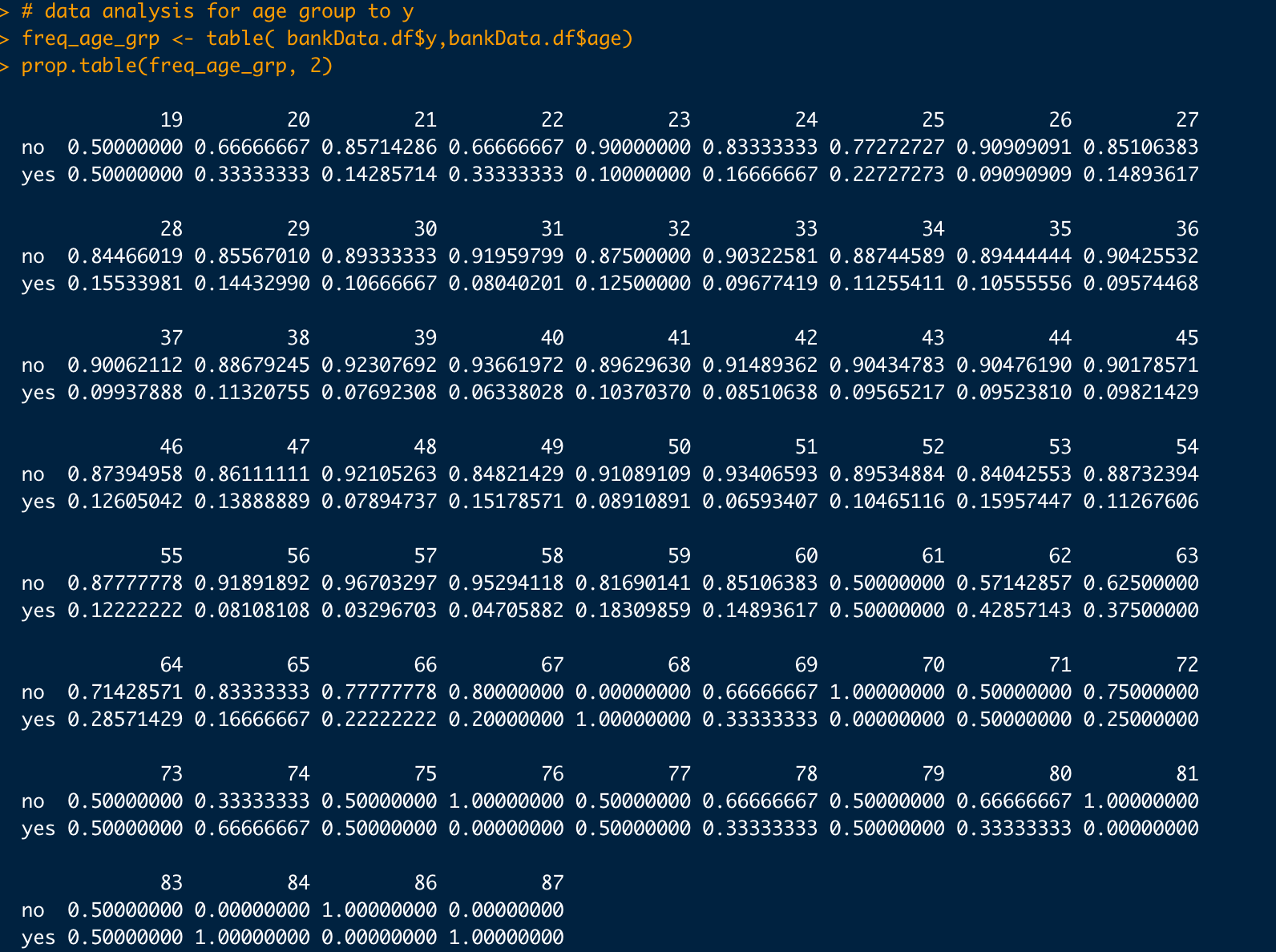
**4. Age:**

Analysis of the relationship between age and Y:

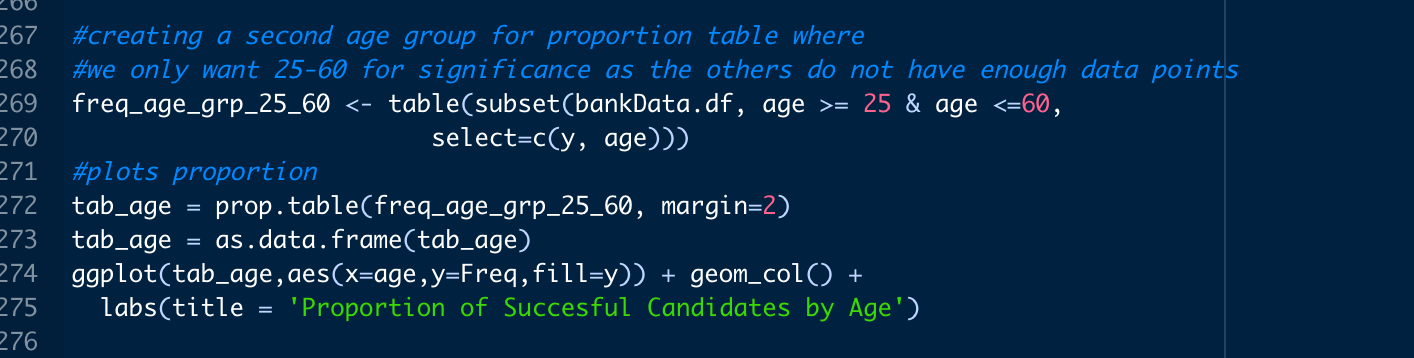


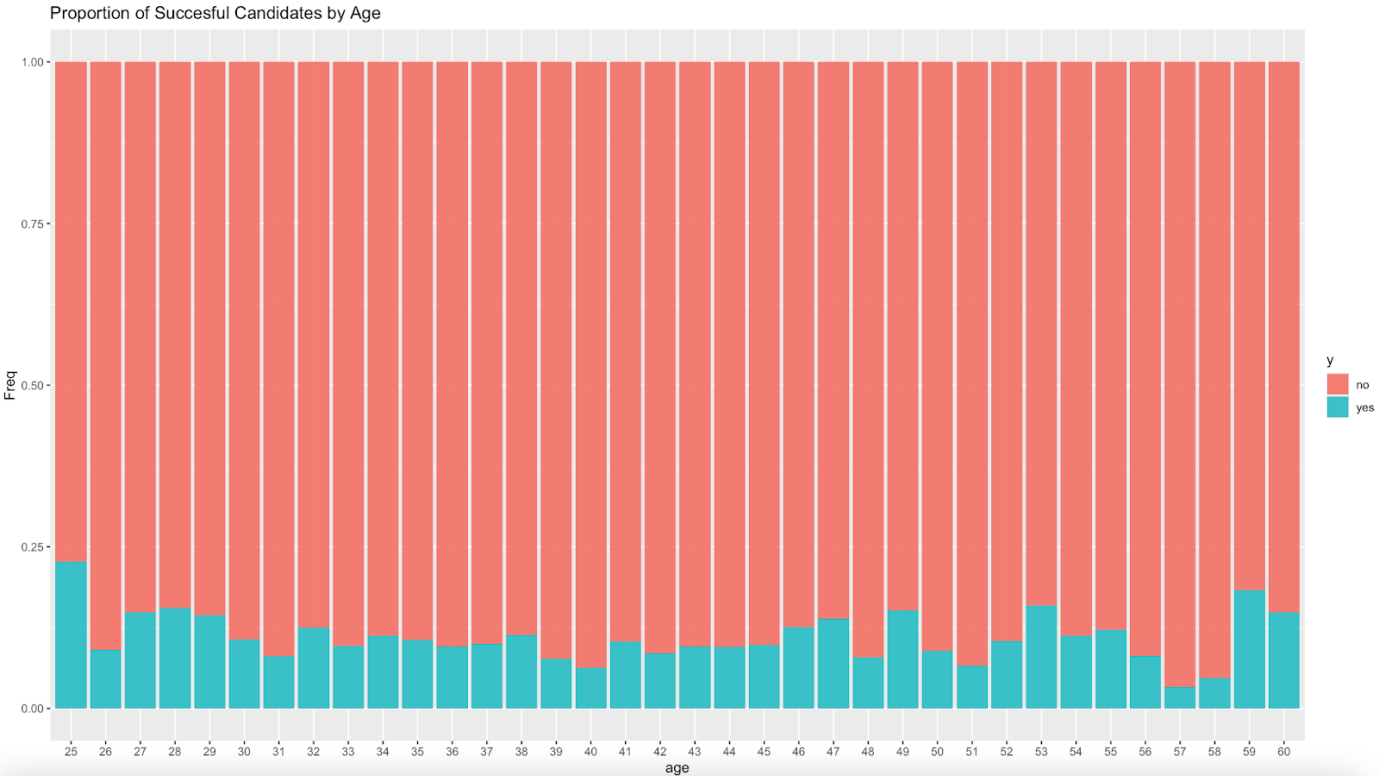
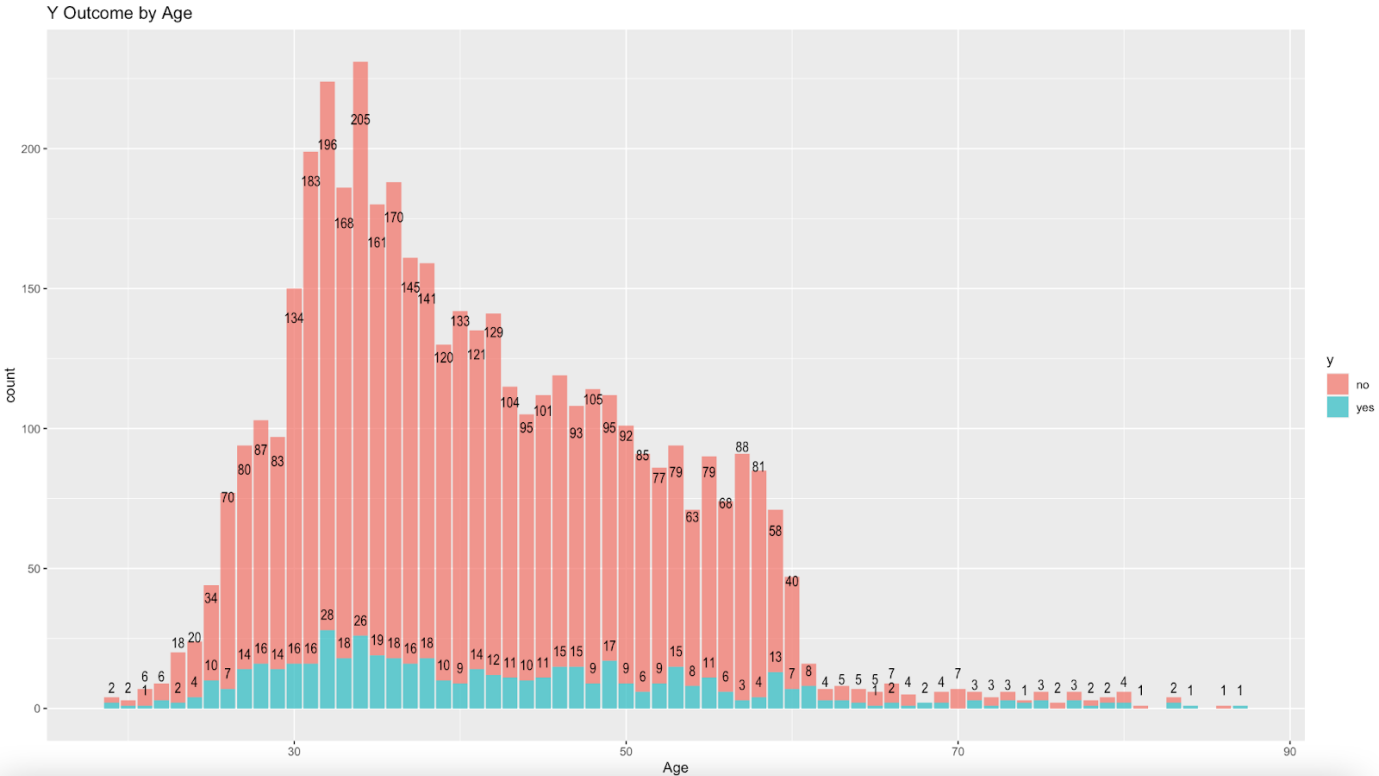


Proportion table for yes/no by age:



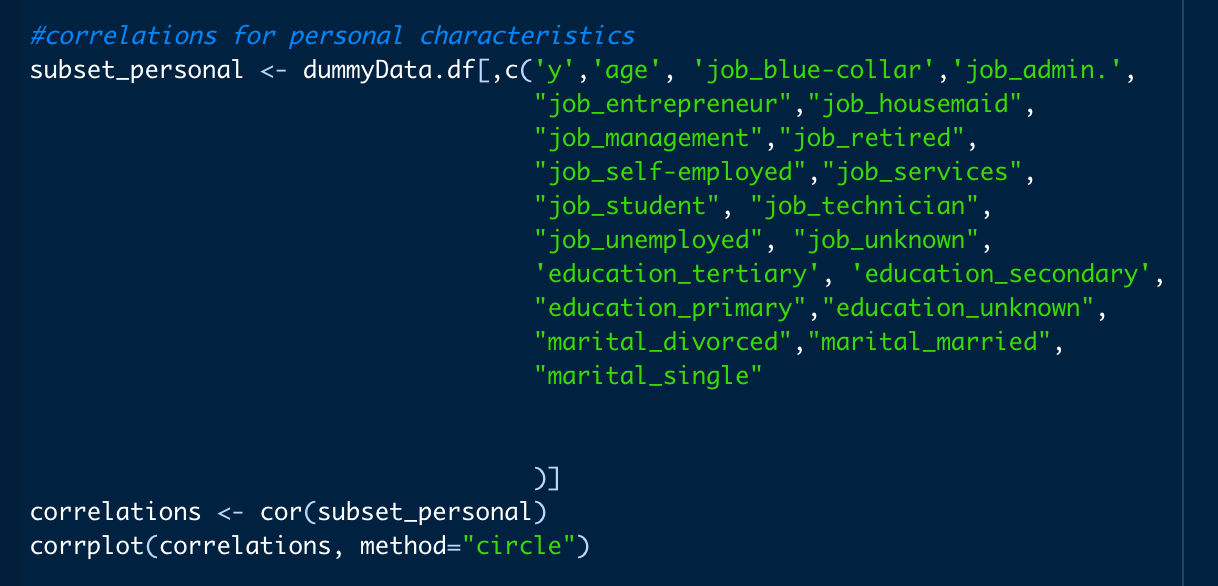
Here we limit out data greater than 60 and less than 25 as not enough data points to make informed decisions.

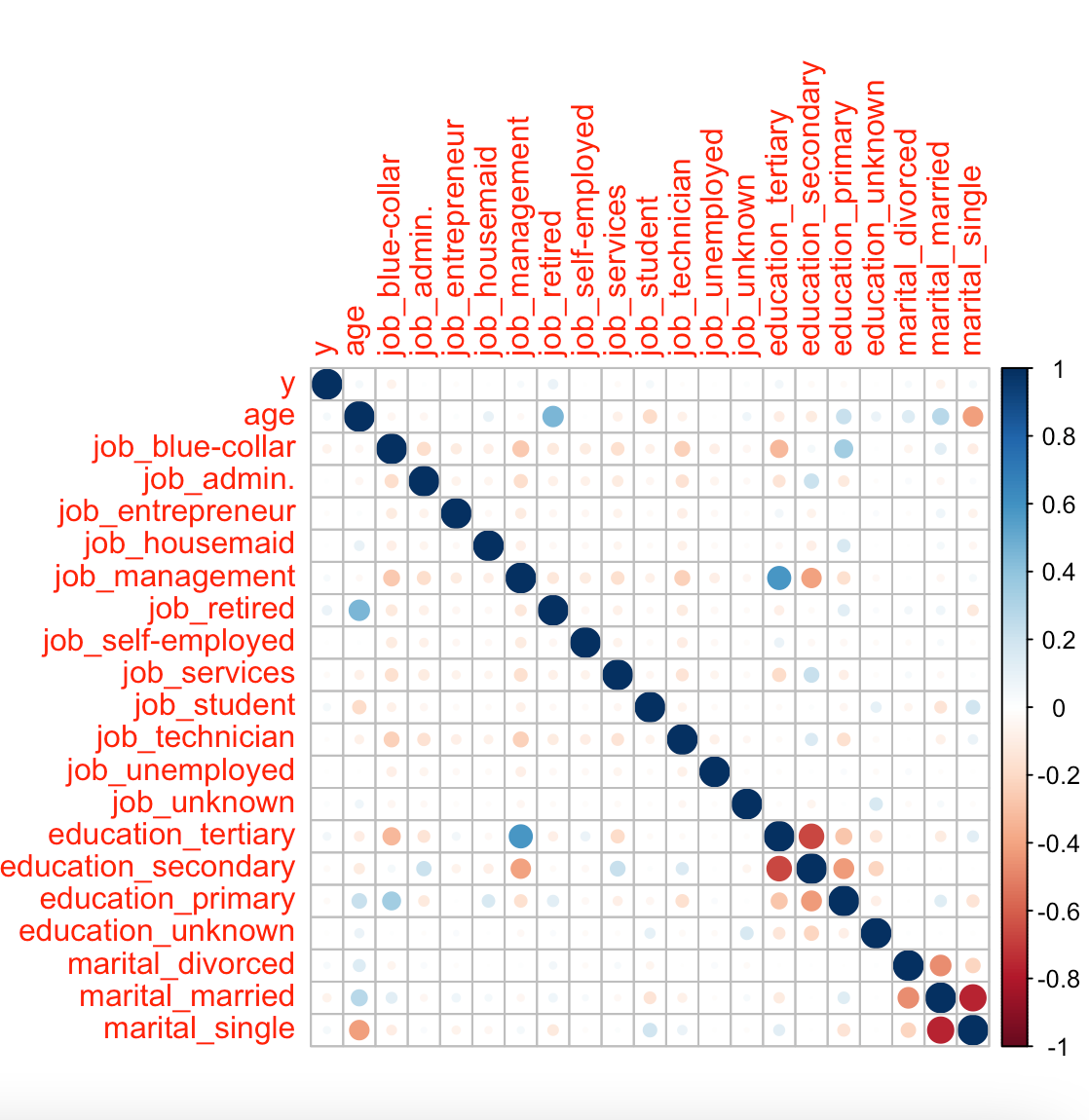




Data is excluded for certain ages from our data and then a proportion table is plotted. As we can see between 25-60 there is no clear pattern of people purchasing at a certain age, just a higher volume of people in general are being contacted to subscribe in 30s-40s.

Next, we ran logistic regressions and correlations for the relationship between personal traits and Y:



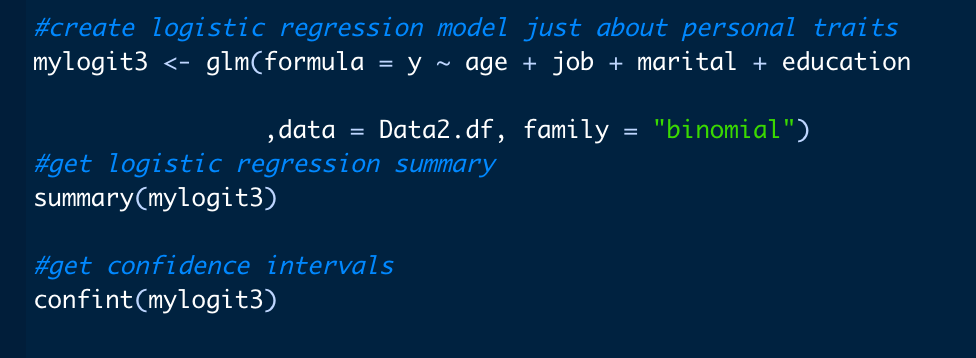


The strongest correlations we see here with y are:

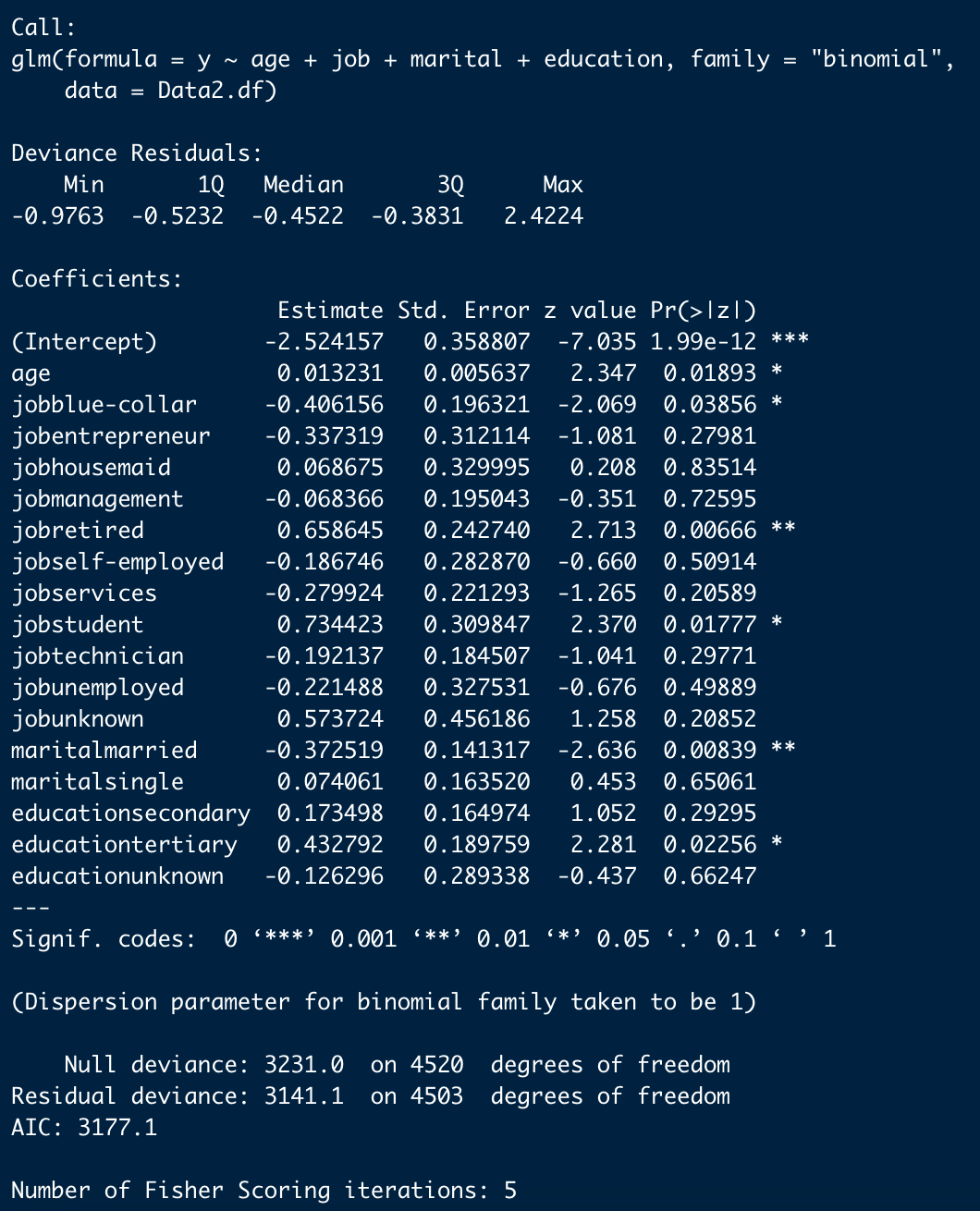
1. Marital status has a negative correlation for married people, positive for single or divorced people.
2. Education tertiary as a positive correlation with y.
3. Retirees and students have a positive correlation, while blue collar workers have a negative.

There are other interesting aspects of the correlation matrix, but go beyond the scope of this question.

Now a logistic regression model is run on the ‘Person Trait’ variables:



(See next page)



The variables marked with \* have a p-value less than 0.05, indicating a strong correlation between these variables and term deposit.

**Insights from the graphs**

1. According to the frequency histogram, divorced persons are more likely to have more deposits and married people result in a negative likelihood in successful outcome variable y, as shown in the correlation matrix and logistic regression outcome.

2. In terms of educational background, people with a higher education background are more likely to have a positive impact on the outcome variable y.

3. Depending on the occupation the client is engaged in, retirees and students are more likely to have more subscribe to the outcome variable.

4. By analyzing the age of customers, as we can see between 25-60 there is no clear pattern of people purchasing at a certain age, just a higher volume of people in general are looking to or subscribe in 30s-40s.

Data analysis shows that: divorced or single persons, higher educated people, retirees and students, are more likely to have more successful y deposits. For banks, in order to obtain more fixed deposits, these four types of people will be the main target groups as the bank should pay more attention to these clients. Specifically, the bank should spend more energy to maintain the relationship with these customers, and carry out more activities and return visits to these customers so as to increase their willingness to subscribe deposits. In addition, for some customer groups with a small absolute number, however, with a high deposit frequency and a large deposit amount, special marketing policies should be implemented.

**Conclusion**

The success of marketing activities is inseparable from the choice of time and target. Therefore, we will summarize from the two dimensions of time and object:

**Time Point**

1. As one of the most important sources of income for banks [1], the interest income created by credit business depends to a large extent on customers' bank deposits. From the results of our data analysis, if the bank wants to increase deposits, the first step needs to determine the appropriate promotion time. It is recommended to conduct more deposit business promotion and customer interaction visits in October and December.

2. In addition to increasing customer deposits, expanding the volume of credit business is one of the most direct and effective measures. From our data, people have a stronger willingness to spend in April, May, June, July and August of the year, so they are considering launching more credit card/loan businesses during this period and carry out corresponding marketing activities accordingly.

**The Target Groups**

1. The main target groups can be divorced persons, highly educated groups, students, managers, and retirees.

2. Judging from the current development situation, banks should attract more new target groups based on maintaining the main customer groups, and continuously expand the scale of customer groups. For key customers, it is necessary to enhance customer stickiness, call key customers on a regular basis, and launch targeted wealth management products suitable for key customer groups.

-For example, for retired in key customer groups, banks can develop pension deposit products with lower risks and stable returns, because pensions are the first and most urgent need for the elderly.

And starting from their emotional needs, make full use of the advantages of the safety and reliability of savings deposits, and continuously amplify this advantage through deposit design and channel experience, so as to increase the sense of security of the elderly in bank deposits.

For example, for young customers in the new customer group, because young people's income surplus level is relatively low, it is possible to launch targeted small wealth management products. Promote and promote for a long time, cultivate customers' awareness of savings, and effectively develop them into long-tail customers.

**Reference**

1. Xing Zheng Finance 2020, Finance Sina Home Page, 15 July 2020, Finance Sina. Available from <https://finance.sina.com.cn/stock/stockzmt/2020-07-15/doc-iivhvpwx5570474.shtml> [ 26 October 2021 ]