**Bank Marketing: Portfolio Project**

Kelsey Thompson

Colorado State University Global

MIS 450: Data Mining

February 9, 2024

Dr. Mamdouh Babi (turned in)

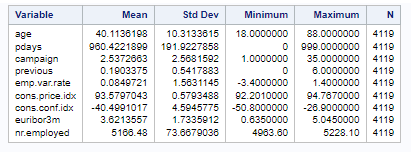
Examining provided marketing data to gain insights and create predictions into potential future customers. I created a **Binary Logistic Regression Model** to classify outcomes and generate actionable insights for future marketing strategies. I focused on whether potential clients would subscribe to a term deposit for a lender. I examined phone-based marketing campaigns that promote financial products.

**Bank Marketing**

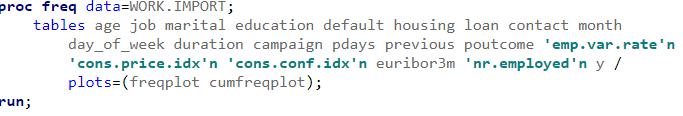
For this task, I utilized the bank.csv dataset. Given the common practice of banks promoting their financial products via phone calls, my role as a data mining (advertising) specialist involves examining the marketing data to gain insights into potential future customers. For detailed information about the variables, I referred to the summary provided in the description-bank.docx file.

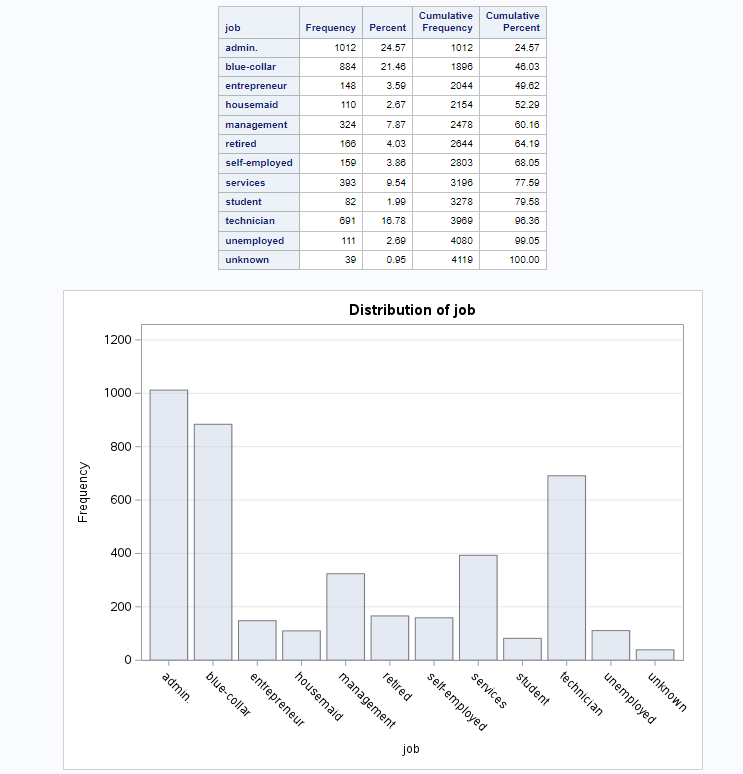
The end goal was to build an appropriate model (or tool) to successfully predict whether a potential client will subscribe a term deposit or not. While exploring classification techniques for this dataset, I opted for a Binary Logistic Regression Model.

* 1. I Explored the dataset by providing summary statistics and graphical summaries of ALL the variables.
     1. Here are the Summary Stats for ALL of the quantitative variables in the dataset (SAS® Help Center, 2012). I can view that there are no missing values for these variables.

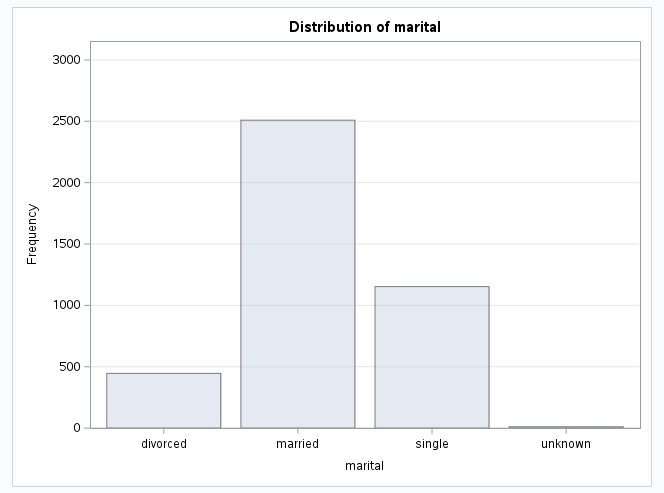


* + 1. Following, are the One-Way Frequencies for all non-numeric variables. I have also integrated step two after each graph (an explanation per some of the key aspects of the data):

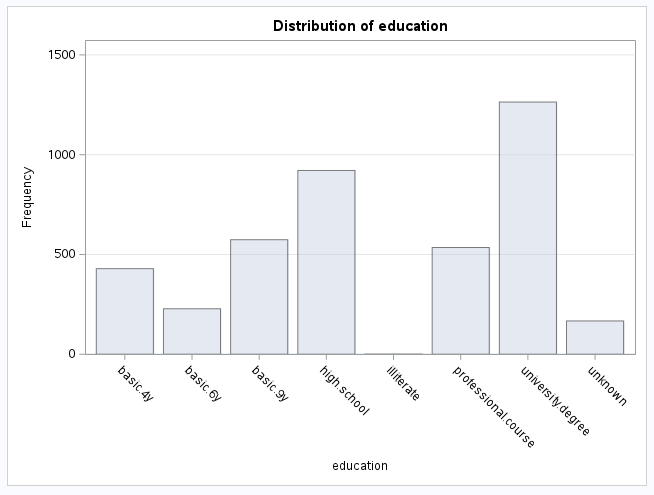




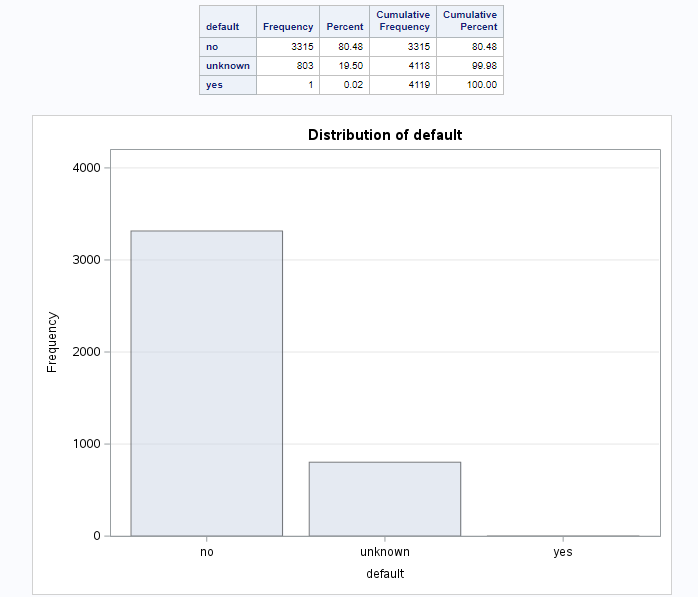
Above, it's apparent that the highest bar corresponds to 1000 under *admin*, followed by *blue-collar* as the second largest. The lowest bar represents *unknown*.



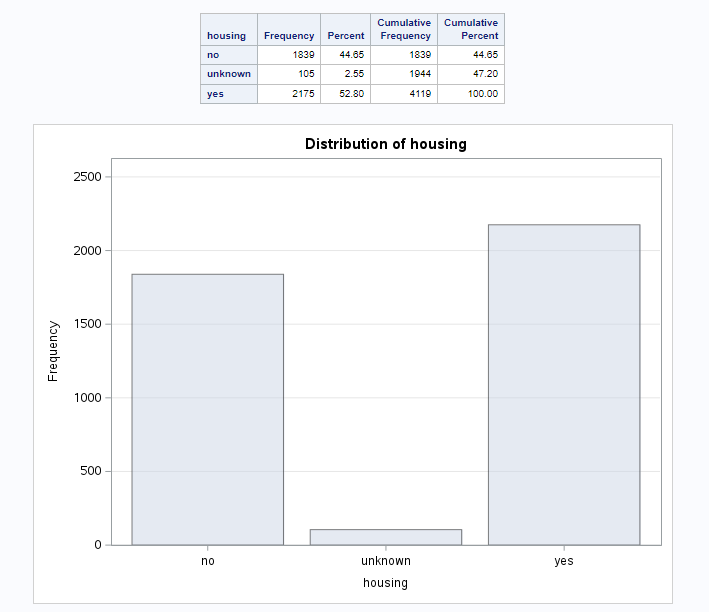
The table above clearly indicates that the majority of customers are married.



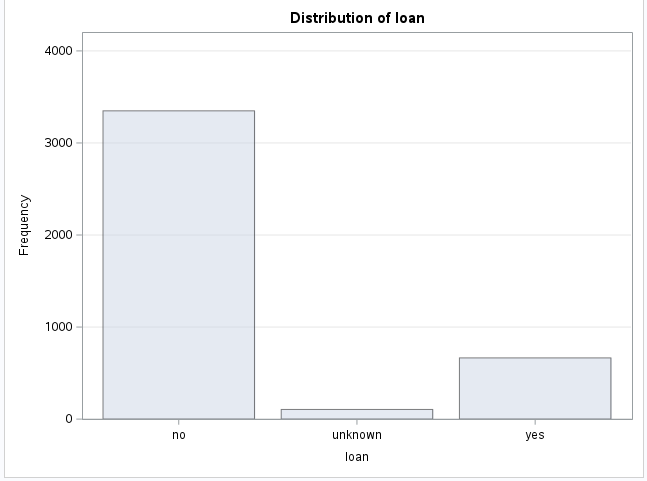
Above, for the distribution of *education*, it's notable that the majority of customers attended university, while the lowest proportion is represented by those who are illiterate.

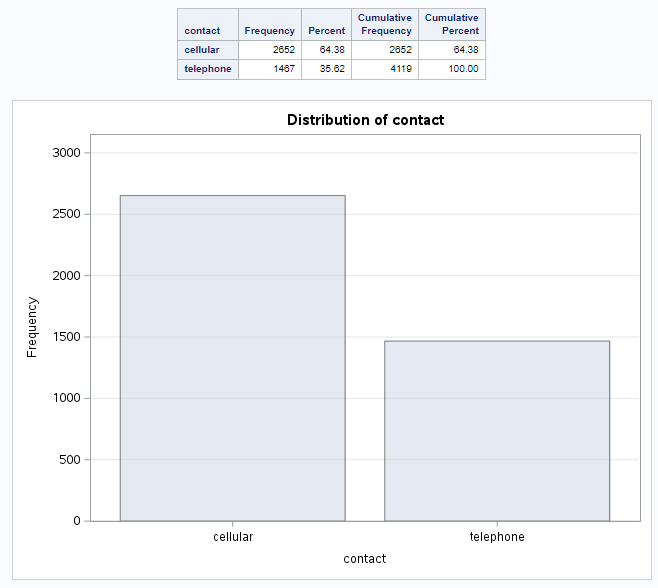


Looking at the distribution of *default*, the majority (80.48%) of customers have not defaulted, while 19.50% are unknown, and a mere 0.02% have defaulted.

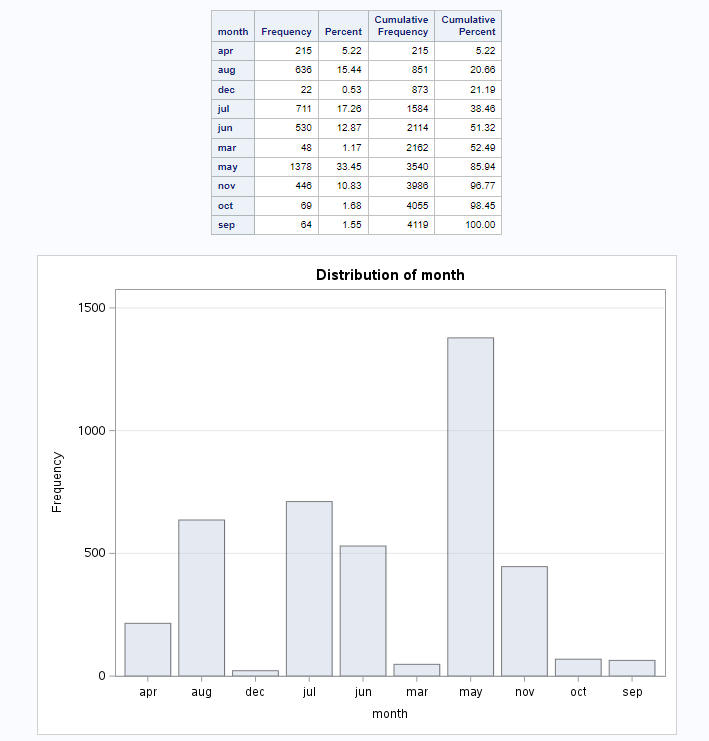


Above, we observe the distribution of *housing* status. The majority fall under 'yes', with 'no' being a close second and 'unknown housing' having a significantly lower count in comparison.

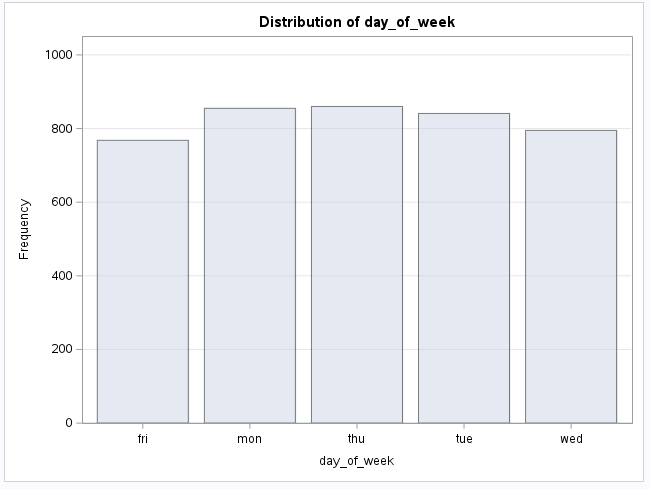
  
The distribution of *loan* indicates that 'no' significantly outweighs both 'yes' and 'unknown loan’.



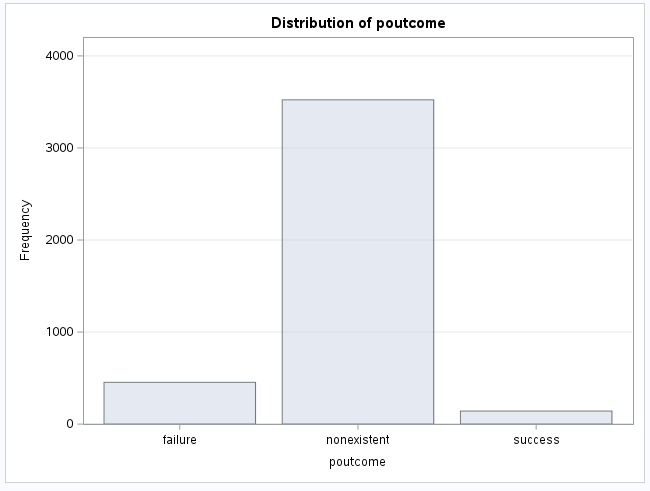
Above, we observe the contact variable. The primary method of communication is ‘cellular’, accounting for 64.38% of the recorded outreach, while 'telephone' represents the second means, comprising of 35.62% of the recorded interactions.



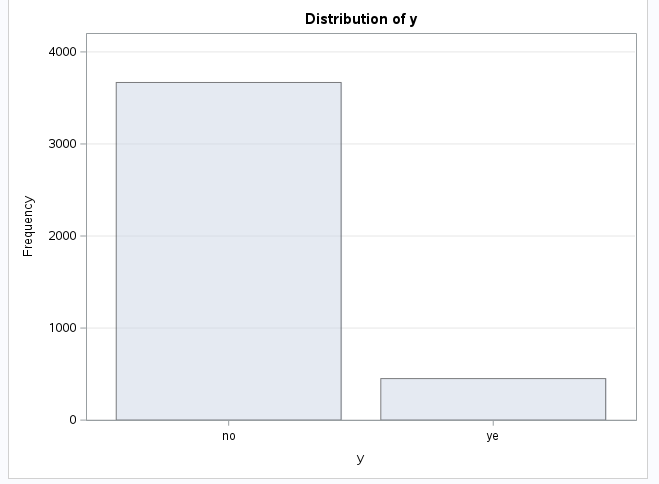
This represents the values by *month*, indicating the last month that a client was contacted. The majority were last contacted in May, while the fewest were reached in December.



The above, bar graph represents the last contact *day of the week* for the customer. There is minimal variation among the weekdays.



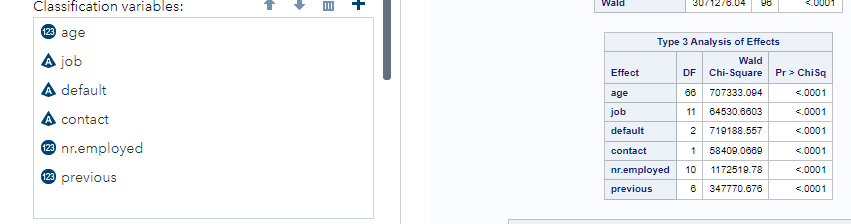
The *poutcome* of the marketing campaign (categorical: "failure," "nonexistent," "success") is depicted here. The majority of outcomes were ‘nonexistent.’



(Above) This represents the presence or absence of a sale. The majority were recorded as non-sales (‘no’).

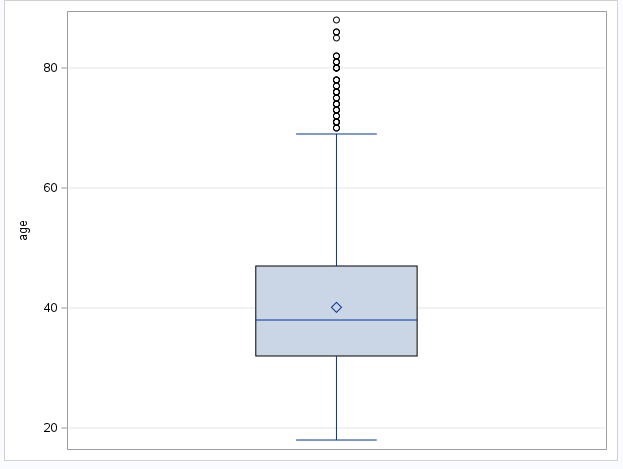
1. This step was integrated into step 1, providing an explanation for some of the key aspects of the data outlined in part 1.
2. I tested various combinations of logically selected variables (as outlined in Milestone 1) to observe changes in significance. For the model, I selected *age, job, default, contact, nr.employed* and *previous*. Then, I reviewed the dataset (chosen variables) for anomalies and have detailed here the methods I employed, along with the results.

Here are my selected variables:

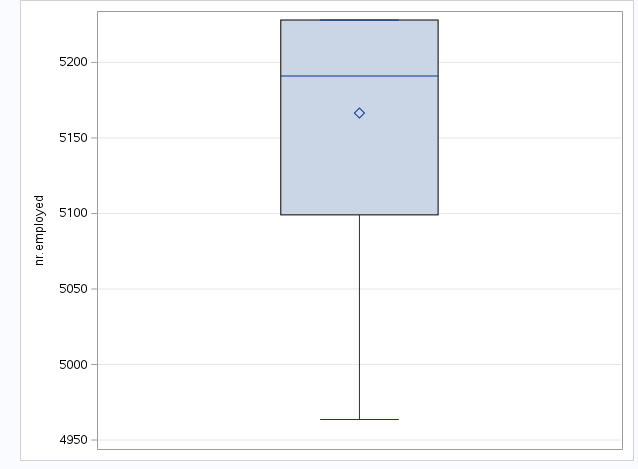


I decided to utilize a box plot to detect any anomalies within the quantitative data, specifically focusing on *age* and *nr.employed*:

1. *Age* (many outliers were detected):

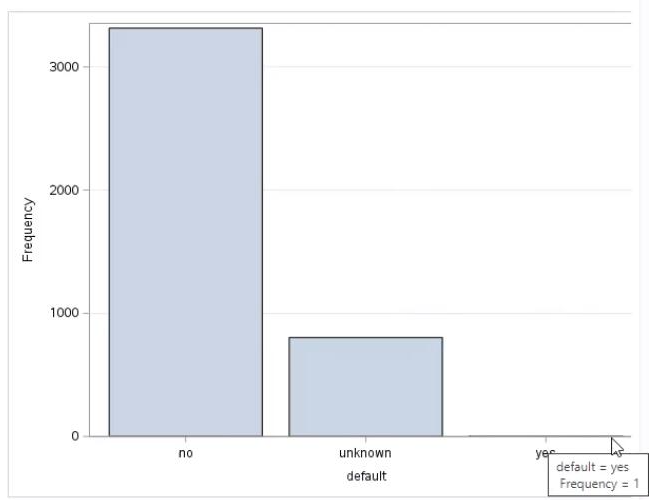


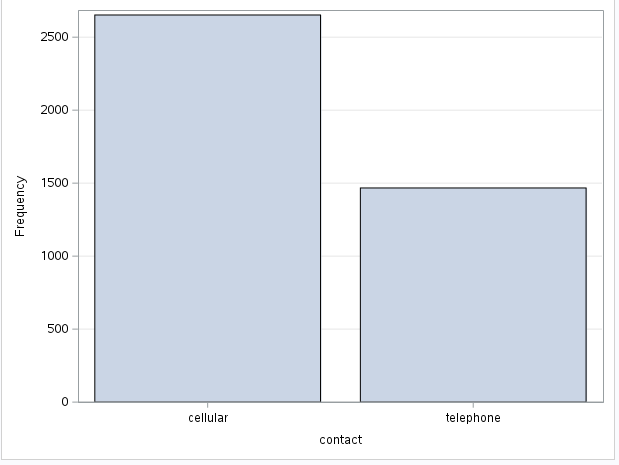
1. *Nr.employed* (no outliers/anomalies were founded):

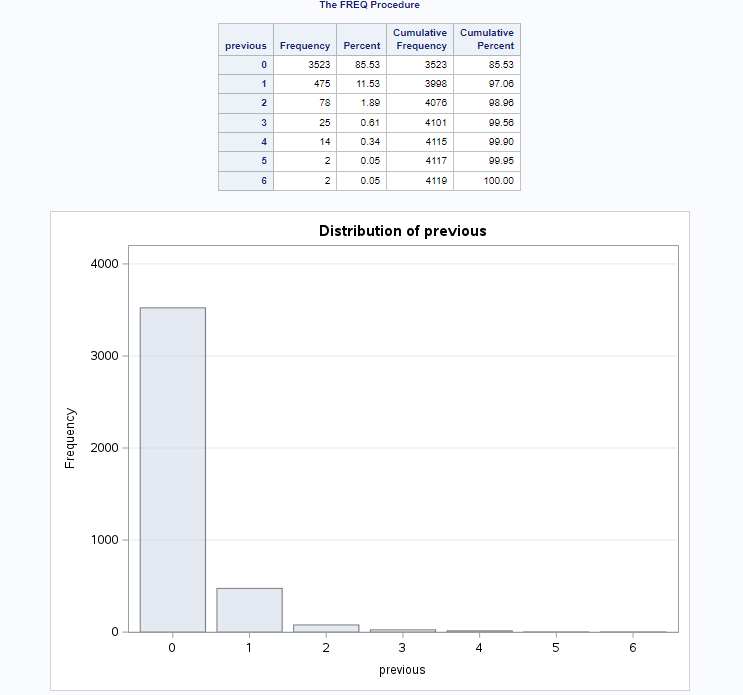


1. For the categorical data, I used bar charts:

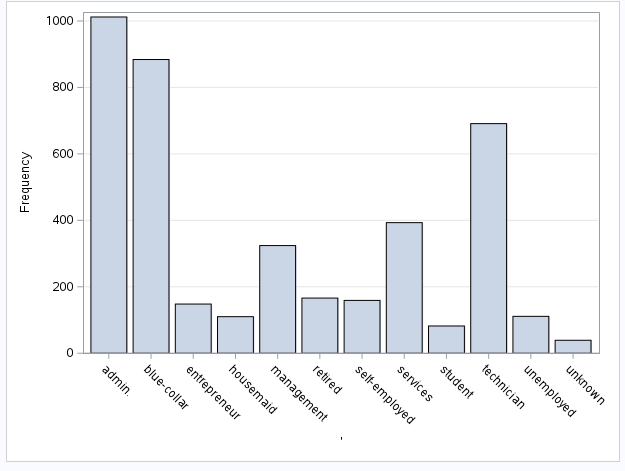
*Default* (no outliers)



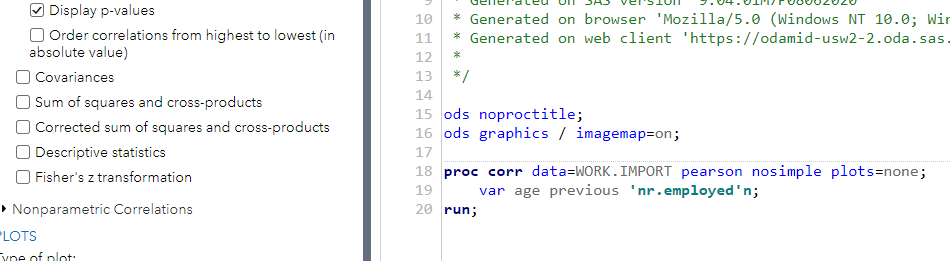
1. *Contact* (no anomalies were present)
2. For *previous*, I used a one-way frequency (3 and 4 are sparse and 5 and 6 are very low).

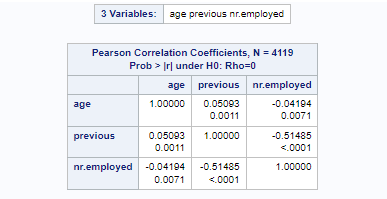


1. For *job*, I utilized a bar chart (no anomalies were found)

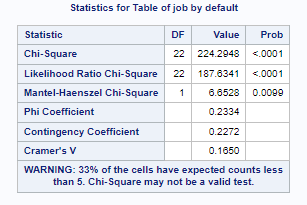


1. In step four, I explored potential correlations/associations among the variables. I will outline the approaches I used and discuss the results obtained.
2. For the quantitative variables (*nr.employed, age,* and *previous*), I conducted a Correlation Analysis. Below is the code and table. These variables exhibit significant correlation, with low P-values. However, there is negligible association between most of the variables, except for *previous* and *nr.employed*.

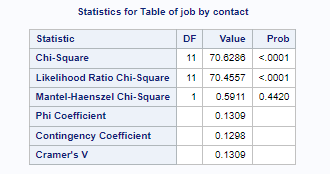


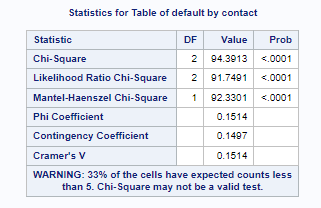


1. I utilized the Chi-squared test for the categorical variables. Below are the results for *job* and *default* (contact was added in step c):

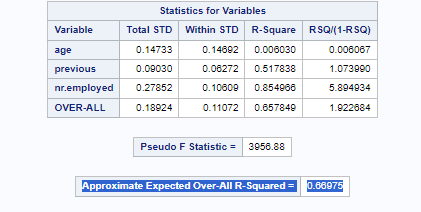


Now, when pairing the categorical variable *contact* with *job* and *default*, I observed statistically significant associations. Both *job* and *contact*, as well as *default* and *contact*, exhibit significant associations.



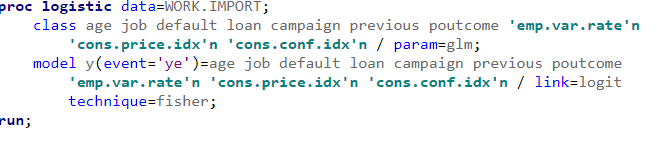


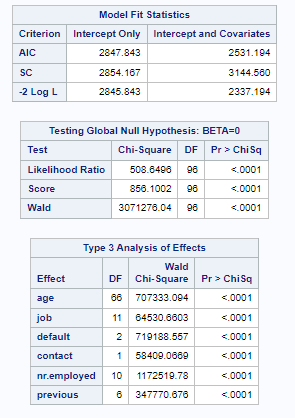
1. Applying a clustering technique, I analyzed all of the quantitative variables, focusing on *age, previous*, and *nr.employed*. The results indicate an overall R-squared of 67%, suggesting that the model fits the data moderately well.



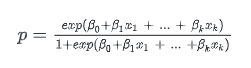
1. I was directed to utilize one of the classification techniques from the course and construct a model that predicts whether the client will subscribe. Further explanation on why I consider my chosen model to be the most suitable for this dataset is further discussed in step 7. I selected this model because I found it to have high accuracy, as evidenced by reviewing the totals of the Confusion Matrix Table. The model fits well.
2. After exploring classification techniques for this dataset, I chose a Binary Logistic Regression Model over a Decision Tree due to its concise nature and the fact that the outcome will is binary (zyBooks, 2022). The model demonstrated a good fit and proved to be an appropriate choice.

Below is the code and the resulting model:



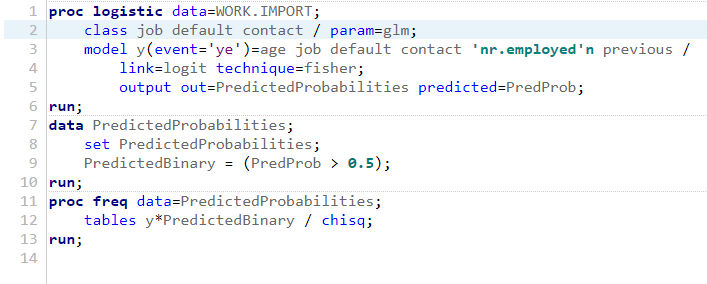


1. Additionally, I would need to apply the following formula to accurately make predictions, regarding client subscriptions (although it is not a requirement of this project).

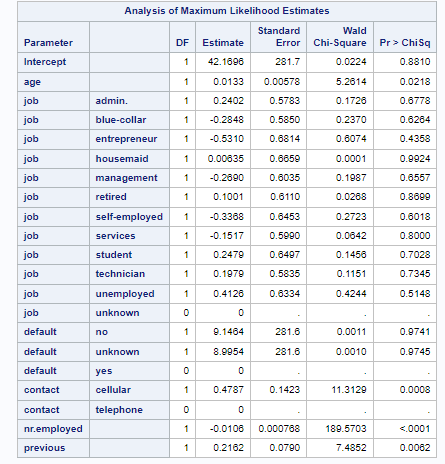


1. To assess the model's fit, I employed a Logistic Progression Model, demonstrating its satisfactory fit (see below). The only potential improvement would be to include additional significant variables.

Here is my Logistic Regression model code:

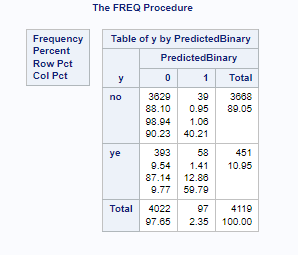


1. Below is the Logistic Regression Model:



7) Upon analysis, I would conclude that the model fits well, as indicated by the confusion matrix table.

a) I utilized a confusion matrix to evaluate the model's fit. When 0 represents no and 1 represents yes. The model has a high accuracy, proven by reviewing the totals. There is a 98.94% accuracy for ‘no.’ There is an 87.14% accuracy for ‘yes’. Therefore, it is an effective model. (SAS® Help Center, 2012)



1. To improve the model, I could add more variables to aid the fit.

References

Elliot, A.C. & Woodward, W.A. (2023). SAS essentials: Mastering SAS for analytics (3rd ed.). John Wiley and Sons. ISBN-13: 978-1-119-90161-7

Kantardzic, M. (2020). Data mining (3rd ed.). John Wiley and Sons, Inc. ISBN 13: 978-1-119- 51604-0

SAS® Help Center (2012). <https://documentation.sas.com/doc/en/helpcenterwlcm/> 1.0/home.htm#/?softwareId=STUDIOMID&softwareVersion=3.7&requestor=inapp&loc ale=en\_US

zyBooks. (2022). Statistics for data analytics.