**Northeastern University**

Logo

Description automatically generated

**Final Project Complete Research Report**

**Intermediate Analytics**

**ALY6015 CRN20419**

College of Professional Studies, Northeastern University

**Submitted To:**

Prof. Valeriy Shevchenko

**Submitted by: Group Delta**

Anukriti Baijal ([baijal.a@northeastern.edu](mailto:baijal.a@northeastern.edu))

Adhira Sahni ([sahni.ad@northeastern.edu](mailto:sahni.ad@northeastern.edu))

Kateryna Dzhyndzhyrysta ([dzhyndzhyrysta.k@northeastern.edu](mailto:dzhyndzhyrysta.k@northeastern.edu))

Raghav Gupta ([gupta.ragha@northeastern.edu](mailto:gupta.ragha@northeastern.edu))

Pranav Marla ([marla.p@northeastern.edu](mailto:marla.p@northeastern.edu))

**Date:** 16 February 2023

**Assignment Summary, Goal, Plan and Dataset**

The purpose of this project was to work within our groups to identify a data set and perform various statistical analysis on it being taught in this course. The data set identified was cleaned and prepared before doing exploratory data analysis using descriptive statistics. As part of this assignment, we aimed to simulate the working of a data analytics team in a corporate structure for achieving a common project goal.

We already know that analytics has become a part of every field today, it has also become one of the most useful tools for any bank or company to get an idea about their customers. Data helps them track, and eventually predict customer behavior. This allows banks/ companies to create policies and products in anticipation of a person’s future requirements, thereby keeping a strong hold on their numbers.

The dataset consists of data related to the direct marketing campaigns of a Portuguese bank. It consists of 45211 rows and 17 columns. The columns in the dataset are:

*age* - Age of the client (numeric)

*job* - Job category of the client

*marital* - Marital status of the client

*education* - Education level of the client

*default* - Does the client have credit in default?

*balance* - Average yearly balance of the client, in euros

*housing* - Does the client have a housing loan?

*loan* - Does the client have a personal loan?

*contact* - Contact communication type of the client

*day* - Last contact day of the month

*month* - Last contact month of year

*duration* - Last contact duration, in seconds

*campaign* - Number of contacts performed during this campaign and for this client.

*pdays* – No. of days passed since the client was last contacted from a previous campaign.

*previous* - Number of contacts performed before this campaign and for this client.

*poutcome* - Outcome of the previous marketing campaign

*target* *-* Has the client subscribed to a term deposit?

In this assignment, we will deploy several statistical analytic techniques to answer various questions that we gathered initially. In the process, we also came across some inconsequential results which helped us mend our further research. We answered questions relating to relation between various parameters, dependency/ impact of variables on one another, etc.

The following Analysis section will go in depth with regards to the tests conducted and findings obtained by the team members.

**Analysis**

We start the project by performing an **Exploratory Data Analysis** to get insight about various variables and data structure as a whole. EDA helps us design our research as it can highlight relations and pinpoint irregularities at a high level.

Importing the dataset from the .csv file



File.choose() ensures portability of code.

Displaying the dimensions of the dataset. The banking data contains 45211 rows &17 columns.

A picture containing graphical user interface

Description automatically generated

Displaying the structure of the dataset

Text, letter

Description automatically generated

The structure of a dataset helps us understand what the *values* of a column look like. By analyzing the structure of the dataset, we can notice trends which can help us ask questions to be investigated further. Some of these trends are:

* + Incorrectly assigned data types
  + Repeating values, especially for quantitative data types
  + Missing data

From the structure, we can see that the datatypes of all the columns seem to be correctly assigned.

* We can notice some repeating values. The character “unknown” seems to be repeated in *contact* and *poutcome* columns and is also visible in the *education* column. How often is the “unknown” repeating in the dataset?
* The *campaign*, *pdays* and *previous* columns show repeating values of 1,-1 and 0 respectively. Do these columns only contain these values?
* The default and Target column show repeating values of “no”. The *housing* column has repeating “yes” values. How often are these values repeating in these columns?

We will investigate the above observations further by looking at the 5-point summary of the numerical columns.

Displaying the 5-point summary of the column in the dataset

A screenshot of a computer

Description automatically generated with medium confidence

This goes to show us that some columns are non-numeric, and thus a 5-point summary showing minimum, maximum, median and quartiles cannot be displayed.

We can see that the values for *campaign*, *pdays* and *previous* don’t seem to be only 1, -1 and 0 respectively. However, the median of the *previous* column is 0 and its mean is 0.5803 so it might be something to investigate further if required.

The minimum value for the *balance* column is -8019, which means there exist negative values for the average yearly bank balance. This information can be crucial if we decide to investigate the *balance* column further.

Next, we made a bar plot of number of clients by call success and observed some information.

Chart, bar chart

Description automatically generated

* We observed that the vast proportion (~87%) of the calls did not result in the customer subscriber to the term deposit.
* This information can be useful when we are analyzing the *Target* column further.

Plotting a jitter Plot between the bank balance and marital status of the clients.

Diagram

Description automatically generated

* As visible from the above plot, the jitter pattern of all three categories are similar, with most values being concentrated towards the bottom.
* However, for the instances where the client is married, the points are also concentrated slightly higher than the other two categories.

Plotting a jitter plot between bank balance and education level of the clients.

A picture containing diagram

Description automatically generated

* In the above plot we see that the jitter patterns are again similar for most categories, but we see that the clients with secondary and tertiary education seem to have slightly higher balances.
* The clients with “unknown” education seem to have the lowest average yearly bank balances.

Plotting a stacked bar plot of clients by profession and success of call.

Chart, bar chart

Description automatically generated

* We observed that 3 categories of professions; “technician”, “management” and “blue-collar” have a disproportionate number of calls.
* We also observed that the ratio of whether the client subscribed to the term deposit was approximately the same.

Stacked Bar plot of no. of clients by Marital status and Call success.

Chart, bar chart

Description automatically generated

* We observed that most of the clients were married and that the ratio of whether the client subscribed to the term deposit was approximately the same, similar to the previous plot.
* Similar to the previous plot, we also observed that the ratio of whether the client subscribed to the term deposit was approximately the same across marital status

Plotting a histogram of the age of the clients

Chart, histogram

Description automatically generated

* We observed that the age was asymmetrically distributed with most of the calls being made to clients above the age of 30.
* We can see that after the age of 60, there was a sharp decline in the calls being made. We can assume that lesser calls had been made to clients who were older than 60 years.

**Question: At alpha = 0.05, is there a relation between the education level and bank balance of the customers? (One way ANOVA)**

Checking tally of negative balance, grouped by education level.

Graphical user interface

Description automatically generated with low confidence

We observe that secondary education level has highest number of people with a negative balance.

For this part of the analysis, we are trying to determine if a relation exists between the education level and the bank balance. But before that we will check the class of these two variables and convert any categorical variables into factors.

Graphical user interface, text

Description automatically generated

We created a boxplot of balance grouped by education level to see if we can visually observe a pattern.

Chart, box and whisker chart

Description automatically generated

Since the median values of balance are rather close for all education levels, it is difficult to draw a conclusion just by looking at the plot. We do notice that there are plenty of outliers in all education levels, mostly on the upper side. There are few outliers below the minimum value.

Next, we created 2 new data frames. First one to show the tally of entries for each education level. We see that the secondary education level has highest number of people. The second data frame showing the sum of balance for each education level. (Note: This is algebraic sum, and some observations also have negative balance or debt as observed earlier).

Graphical user interface, text, email

Description automatically generated

Next, we see a bar plot for a more visual graphic about sum of balances under each education level.

Chart, bar chart

Description automatically generated

It is interesting to note that secondary education level has maximum number of people with debt, but the sum of bank balances still adds up to be the highest. This reflects that the number of people with secondary education level is highest in Portugal, this is in accordance with what we observed earlier.

We also note that number of people with tertiary level of education is almost half that of secondary, however in terms of sum of bank balance they are only slightly behind. This shows that mean balance must be higher for a person with higher level of education.

We confirm this by looking at mean balance for each education level.

For this, we start by first creating different subsets for each education level, and then calculating mean bank balance. We use rbind() to combine these data frames together which we will eventually use to run our hypothesis test.

Text

Description automatically generated

We note that mean bank balance is highest for third level of education, followed by primary, and lastly secondary level of education. So even though sum of bank balance was highest for secondary education, but it is a misrepresentation of the true situation.

Similarly, the median bank balance also follows the same trend with tertiary education level at the top, followed by primary and secondary respectively. We note here that since the values are close to each other, that’s why the boxplot was unable to generate insights.

Next, we performed an ANOVA test to check if there is a significant relation between the education level and bank balance. We ran the test for a **significance level of alpha = 0.05.**

**Step 1. State Hypotheses**

**Null Hyp, H0:** There is no statistically significant relation between education level and bank balance.

**Alternate Hyp, Ha**: There is statistically significant relation between education level and bank balance.

**Step 2. Find Critical Value**

Graphical user interface, text, application

Description automatically generated

We ran the test, as shown above. From the test result summary, we observe the degrees of freedom, the F test value, and the p value that shows the level of significance of our hypothesis.

We can find the critical value using F tables available, or using the R functions as follows.

Graphical user interface, text

Description automatically generated

We note that the **critical value is 2.605.**

**Step 3. Find the Test Value**

Text

Description automatically generated

**Test value is found to be 116.682**, and **p value is found to be less than 2x10-16.**

**Step 4. Making Decision**

Comparing the p value with alpha, we draw the conclusion that the Null Hypothesis can be rejected.

Text

Description automatically generated with medium confidence

**Step 5. State Conclusion**

Since we reject the null hypothesis, we conclude that at 95% confidence level there is a statistically significant relation between the different education levels and the bank balance.

Next, we want to understand **if there is a difference between the number of calls made to the customers last time and the number of calls made to the customers during the campaign**. To answer this, we perform **Wilcoxon Rank Test.**

**Step 1:** H0: There is no difference in the number of times a customer is called during the current and the previous campaigns.

H1: There is a difference in the number of times a customer is called during the current and the previous campaigns.

**Step 2:** Since α=0.05, the critical value=±1.96

**Step 3:** The p-value is less than 2.2e-16.

Graphical user interface, text, application

Description automatically generated

**Step 4:** Since the p-value < α, we reject the null hypothesis.

**Step 5:** There is enough evidence to reject the claim that there is no difference in the number of times a customer is called during the current and the previous campaigns with a confidence of 95%.

Further, after the above analysis, we want to understand **if the number of phone calls made to the customers during the current campaign leads to success, failure, or other outcomes with respect to each customer.** For this, we perform an **ANOVA test.**

**Step 1:** H0: There is no difference in mean outcome being failure, success, other or unknown.

H1: There is a difference in mean outcome being failure, success, others or unknown.

**Step 2:** Since α=0.05, the critical value= 2.61

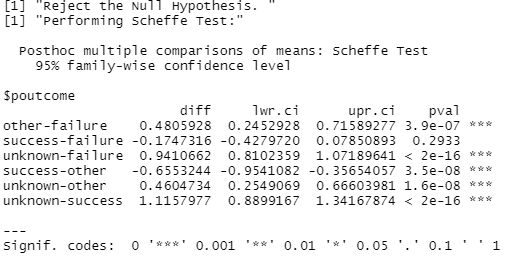


**Step 3:** From the test, we gather that the F-value=192.8 and the p-value < 2e-16

A picture containing text

Description automatically generated

We further conduct a Scheffe’s Test:



**Step 4:** We notice that all the p-values are less than α. Since the p-value is less than α, we reject the null hypothesis.

**Step 5:** Based on the results of the test, we reject the null hypothesis. There is enough evidence to reject the claim that there is no difference in mean outcome being failure, success, other or unknown based on the number of phone calls made during the campaign with 95% confidence.

**Question: Which factor has the highest correlation with bank balance? (Correlation Matrix)**

Now, we want to see if which factors have the most effect on the bank balance of an individual. We start by looking at the structure of the dataset to confirm our character and integer variables.

Text, letter

Description automatically generated

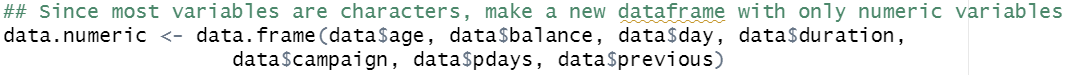
We can see that most of the variables are categorical, so to get a precise correlation, we must consider all variables. For this we will use the ggcorplot() library.

Since there are a lot of variables, it hard to draw conclusions from this correlation matrix. A closer look of the matrix tells us that the correlation coefficient of majority of variables, with respect to balance lie in the range of **-0.07 to 0.2.** These indicate a very weak correlation as closer the values to -1 and 1, stronger the correlation is.

Chart, scatter chart

Description automatically generated

It can be confirmed that only a few variables are integer type, such as age, duration etc. So we will make a new data frame called ‘data.numeric’ to plot a correlation matrix between the numeric variables.



We use the cor() function to plot a matrix, it’s result can be seen as below. It can be observed that the correlation coefficients with respect to balance are not very high. In fact, the highest value is that of 0.1 for the variable ‘age’. This implies that the factors age, day, duration, campaign, pay days and previous don’t have a significant impact on the bank balance of an individual. We can also confirm this by plotting a correlation plot using the corplot() function.

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing chart

Description automatically generated

A picture containing scatter chart

Description automatically generated

Our initial analysis is confirmed, and we can see that none of the above factors have a significant impact on the bank balance. If we run a linear regression using Balance as our predictor and duration and age as our response variables, we get the following results.

Text

Description automatically generated

We would expect **age** to have a more significant impact on someone’s bank balance but observing the correlation matrix, we can see that it only has a value of **0.1.** Age is still the highest indicator of the balance out of all numeric variables. We can conclude that, there are certain other factors such as education level, marital status, and job status etc. which impact the bank balance. We will need to run further analysis and tests to know the relationship between these variables.

**Question: Significance of variables with respect to their housing status. (G.L.M.)**

Next, we want to check the significance of the variables, with ‘Housing’ as our response variable. For this we will use a generalized linear model.

We start by creating a training and test data set, using an **80/20 split**. For this we will make use of the caret library. Set.seed() ensures the reproducibility of results.

Graphical user interface, text, application

Description automatically generated

The caret\_train data frame contains 80% of the observations from the original data set, banking and caret\_train contains the remaining 20%.

Since housing is a categorical variable with 2 responses ‘Yes’ and ‘No’, we will use binomial, logit family type.

Table

Description automatically generated

Most variables, except ‘marital married’, ‘job services’ ‘job technician’ show significant relationship with the ‘Housing Status’.

Table

Description automatically generated

It can be said that the housing status of an individual is dependent on various factors such as age, job status etc.

**What is the main cause of loans? (Correlation and Chi-Square Independence Test)**

To answer this question, we first started with isolating the needed variables and storing them separately. Therefore, we created a new data frame that contained our dataset without the variables about the campaign.

Text

Description automatically generated

Next, we have created a correlation matrix. Since most of the variables are categorical, we had to find another way to create the matrix, other than shown in previous assignments. Through some research, we were able to build a function that made the correlation matrix with the categorical variables.

Text

Description automatically generated with low confidence

Looking at the matrix, we can see that the highest correlation with housing loans is the job type – blue collar (0.18). The highest correlation with personal loans is credit card in default (0.08), however the biggest cause is the secondary education (0.07).

Chart, scatter chart

Description automatically generated

Usually, the next step would be to build a scatterplot, however given the nature of the data, the scatterplot may not be useful. Below is the code we used to build such scatterplot and the graph itself. As seen on the graph, the data points are too close to each other, and we found it hard to interpret.

Graphical user interface, text

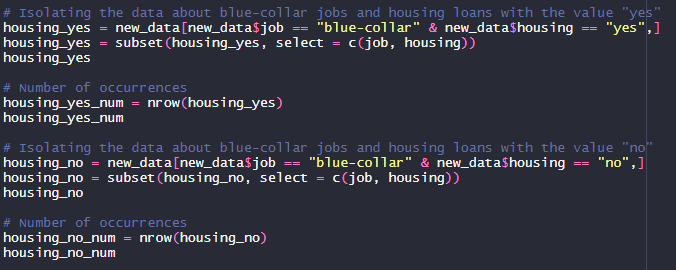
Description automatically generated

Table

Description automatically generated

Since we know from the correlation matrix the highest correlations, we decided to run Chi-Square Independence Test on those variables. As established before, the highest correlation with the housing loan was blue-collar job type. On the other hand, the highest correlation with the personal loan was the credit card in default variable. However, since we are trying to find the cause of loans, we will be working with second highest correlation – secondary education.

To perform the test, we first built a matrix containing the blue-collar job type and secondary education, and their respective amounts of “yes” and “no” related to the loans variables. Since personal and housing loans variables have the same type of response and we are trying to find the cause of any type of loan, we did not have to separate those two variables and build additional matrices. To create the matrix, we isolated the variables into new data frames with their respective responses.



A screenshot of a computer

Description automatically generated with medium confidence

Next, having the data frames ready, we were able to build the matrix. Now we are ready to perform the Chi-Square Independence Test.

A screenshot of a computer

Description automatically generated with medium confidence

**Null hypothesis**: The presence of loans is independent on the job type or education. The **Alternative hypothesis**: The presence of loans is dependent on the job type or education. Our alpha level is 0.05. Finally, we ran the Chi-Square Independence Test.

Text

Description automatically generated

Since the p-value is < 2.2e-16 which is a number very close to zero and it is less than our alpha level of 0.05, we reject the null hypothesis. Therefore, we can confirm that the presence of loans is dependent on the job type or education. Having performed this test, this further solidified our assumptions of the strong correlations between these variables that we saw from the correlation matrix. As a result, we can conclude that the main reasons for any type of loan are the blue-collar job type or secondary education.

Text

Description automatically generated

**Is there a difference between the balance of different job types? (ANOVA)**

To explore this question, we first isolated the job type and balance variables and stored them in a separate data frame.

Graphical user interface, text

Description automatically generated

Next, we built a correlation matrix to have a visual representation on how the variables correlate between each other. We can see that the balance is highly correlated with management job type and retirement.

Chart, histogram, scatter chart

Description automatically generated

**Step 1. State Hypotheses**

H0: there is no effect of the job type on the balance.

H1: job type affects the balance.

**Step 2. Find Critical Value**

After running ANOVA test, we could see that our critical value is 1.7886.

A picture containing text

Description automatically generated

**Step 3. Find the Test Value**

From displaying the summary of the ANOVA test, we can see that our F value is 43.01, with the p-value being <2e-16.

Graphical user interface, text

Description automatically generated

**Step 4. Making Decision**

Since our p-value is <2e-16 which is less than our alpha level of 0.05, we reject our null hypothesis. Which means that job type does affect the balance.



**Step 5. State Conclusion**

Based on our results, we can conclude that the job type has an effect on the balance. As discovered previously, the highest correlations were with management job type and retirement. The lowest correlation was with blue collar job type. We also built a scatter plot, in order to further analyze the balances per job type. From the graph, we can see that management has the highest balances amongst any other job types. We can also see that blue collar has a point with the lowest balance.

A picture containing text, light

Description automatically generated

**Conclusion**

To conclude, we made use of multiple statistical tools for working with our data, this includes performing exploratory data analytics using descriptive statistics, regression modeling for multiple factors, various types of hypotheses testing including non-parametric tests. The dataset has been briefly described containing 45211 rows and 17 columns. Each parameter has been briefly explained in the introduction section. Some characteristics are categorical in nature while others are numeric. In our first week of analysis, we identified some questions that we desired to solve as part of this project. After going deeper into our research, we expanded the scope of work progressively.

Exploratory data analysis allowed us to gauge the information about the data set at a high level and point out certain anomalies and discrepancies. Further plotting of variables against one another in our EDA began to show us relations between some of them.

A number of questions were answered as part of analysis; however, we did come across some roadblocks in the process as well.

With the help of one-way Anova testing a relation was established and authenticated between education levels and bank balance.

Using Wilcoxon Rank Sum Test, we verified that there is a difference in the number of times that clients were contacted as part of the previous campaigns and the current campaign.

Using Anova test we also conclude that the mean number of failures/ successes/ and unknowns vary significantly for the present campaign.

Last but not the least, correlation matrices were illustrated to show collinearity between variables. It was concluded that the type of job also has a significant impact on the bank balance of customers.

**References**

James Chen (March 20, 2022). *Time Deposit: Definition, How It’s Used, Rates, and How to Invest*. <https://www.investopedia.com/terms/t/termdeposit.asp>

Kranti Walke (2020). *Bank-full*. <https://www.kaggle.com/datasets/krantiswalke/bankfullcsv>