Assignment 1 (Practical Data Science)

Tran Hoang Phuc - s3911244

Task 4: Report

Data Preparation

1.1. Data import and basic description

Firstly, I import the required libraries: Pandas and NumPy, to import the bank.csv dataset to the data frame. It is also necessary to copy the data set to a new data frame for modifying and avoid making changes in the original data frame. I use “df.info()” function to check the information, columns and datatypes of the dataset.

1.2. Check column values

Then, I start checking all columns’ values. I created a function “whitespace_removing()” to find all white spaces in the dataset and strip them. Then, duplicated rows are checked by using “df.duplicated()”. I found that there were no duplicated data, so I continued to the next step.

I used the function “value_counts()” to check every columns. I made changes to several rows. First, I looked at the "age" column. According to "https://www.everythingoverseas.com/portugal/banking-in-portugal/," the consumer must be at least 18 years old to open a bank account and use term deposit. Customers under the age of 18 are thus an unattainable value. There are also age 140 and 138, which are both impossible values because the oldest person was 122 years old (https://en.wikipedia.org/wiki/Oldest people). Rows with ages less than 18 and greater than 122 are removed.Then, in the column “job”, there are some typos mistakes: ‘admin.’, ‘bluecollar’, ‘entrepreneurs’ and ‘servicess’ and I fixed them using “replace()” function.

There are also typos in the column "education": 'basic0.4y' and 'basic.6y'. There are additional white spaces in the values of this column, but I removed them at the start. Furthermore, this column holds the value "na" (I determined that this is not a missing value because the datatype of "na" is Object), and I replaced it with "unknown" because it is the closest meaning. In column “housing”, there are typos mistakes and value “na”, which I also solve the problem like above. Some values have capitalized character, which needed to be changed to lower-case.

Moreover, we started to see “NaN” value, which are missing values. I will deal with this problem later. Column “loan” also has value “na” and be replaced by “unknown”. There are typos mistakes in column “day_of_week”: ‘Friday’ and ‘Monday’. They needed to be fixed to ‘fri’ and ‘mon’. In column “duration”, as in the specification, I need to check if the “duration” is 0 then the output “y” should be “no”, and there is no row incorrect. Columns “campaign”, “emp.var.rate”, “cons.price.idx”, “cons.conf.idx” and “euribor3m” have missing values. Besides, the other columns “marital”, “default”, “contact”, “month”, “previous” and “poutcome” are all clean.

1.3. Datatypes

Datatypes are essential in all aspects of programming. However, in order to minimize errors and save

memory, we must utilize appropriate datatypes. It is evident from this dataset that the present memory use is substantial (in my notebook, it is up to 700 KB). Furthermore, Pandas defines a string variable with a limited range of values, and such a string variable can be converted into a category variable to save memory [1]. We can examine the specification and see that all "object" datatype columns have a range of values (such as type of jobs or yes, no answers). As a result, the datatype "category" is preferable to "object."

Moreover, columns “campaign” and “duration” are using “float64”, while all of those values are in integer, so that we should change the datatype to “int”. Integer and Float numbers should be downcast to smaller bit size to save more space. For example, age should be using “int8” (from -127 to 127 numbers) because the highest age currently cannot pass 127 years old.

I used function “astype()” to change the datatypes and downcast numeric columns. Columns “age”, “campaign” and “previous” are downcasted to “Int8”. Columns “duration” and “pdays” are changed to “Int16”. Columns “emp.var.rate”, “cons.price.idx”, “cons.cof.idx”, “euribor3m” and “nr.employed” are downcasted to “Float32” from “Float64”.

After working with datatypes, the memory usage dropped significantly, to approximately 255 KB.

1.4. Missing value

My next step is to address the missing value. Using the function "isna()," I can locate all columns with missing values. As previously stated, there are seven columns with Na values: 'housing', 'duration', 'campaign', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', and 'euribor3m'. For this dataset, I deal with missing values in two ways: replacing them with appropriate values and eliminating unnecessary rows. In this exercise, I filled in the blanks with appropriate data.

There is one nominal value and six numerical values. To deal with missing values in numerical data, I replace them with the mean, median or mode of the column. I needed to utilize the functions "mean()", "median()", and "mode()" to see the results of what data I replaced. If the data distribution is tilted to the left, the mean is often lower than the median, which is frequently lower than the mode. When the data distribution is biased to the right, the mode is typically smaller than the median, which is less than the mean [2]. As a result, if the data distribution is skewed to the left or right, I picked median to substitute missing values because mode and mean can dramatically alter the standard deviation.

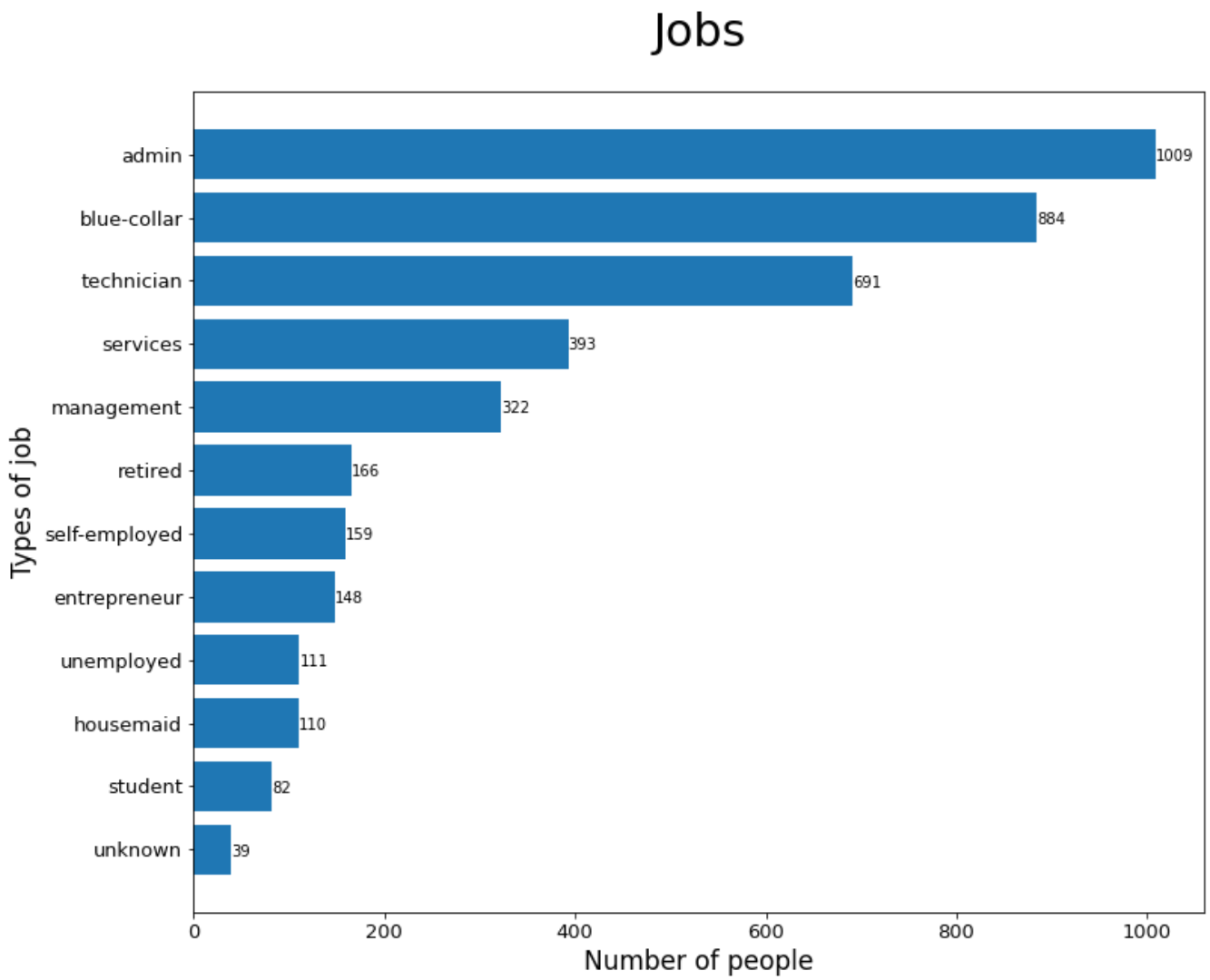
The first column is “housing”, which had two missing values in row 27 and row 140. There are three options: "yes", "no", and "unknown" data. Since there are only two missing values in this column, picking the value that occurs most frequently will be appropriate because it won't significantly change the dataset. I replace “NaN” with the most frequent data is “yes”.

In column “duration”, there are three missing values in row 82, 254 and 299. We can find out the the mode < median < mean, which means that the normal distribution is skewed to the right. I use median to replace the Na data. All other columns used the same technique. “emp.var.rate” column’s missing values are replaced by median value. For column “cons.price.idx”, the mean, median and mode were close to each other, so that the distribution graph was symmetric. Using mean data to replace missing value was suitable. Missing data in column “cons.conf.idx” and “euribor3m” are replaced with median value.

Data Exploration

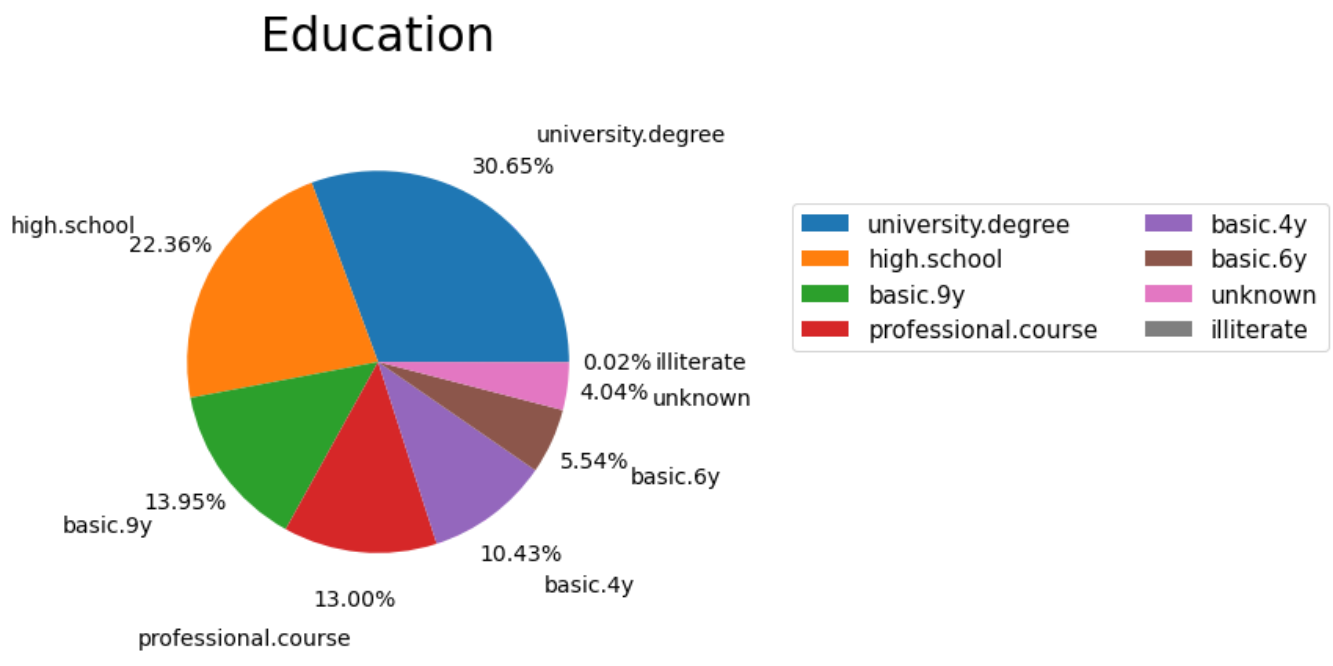
Sub-section 1:

Chart 1:



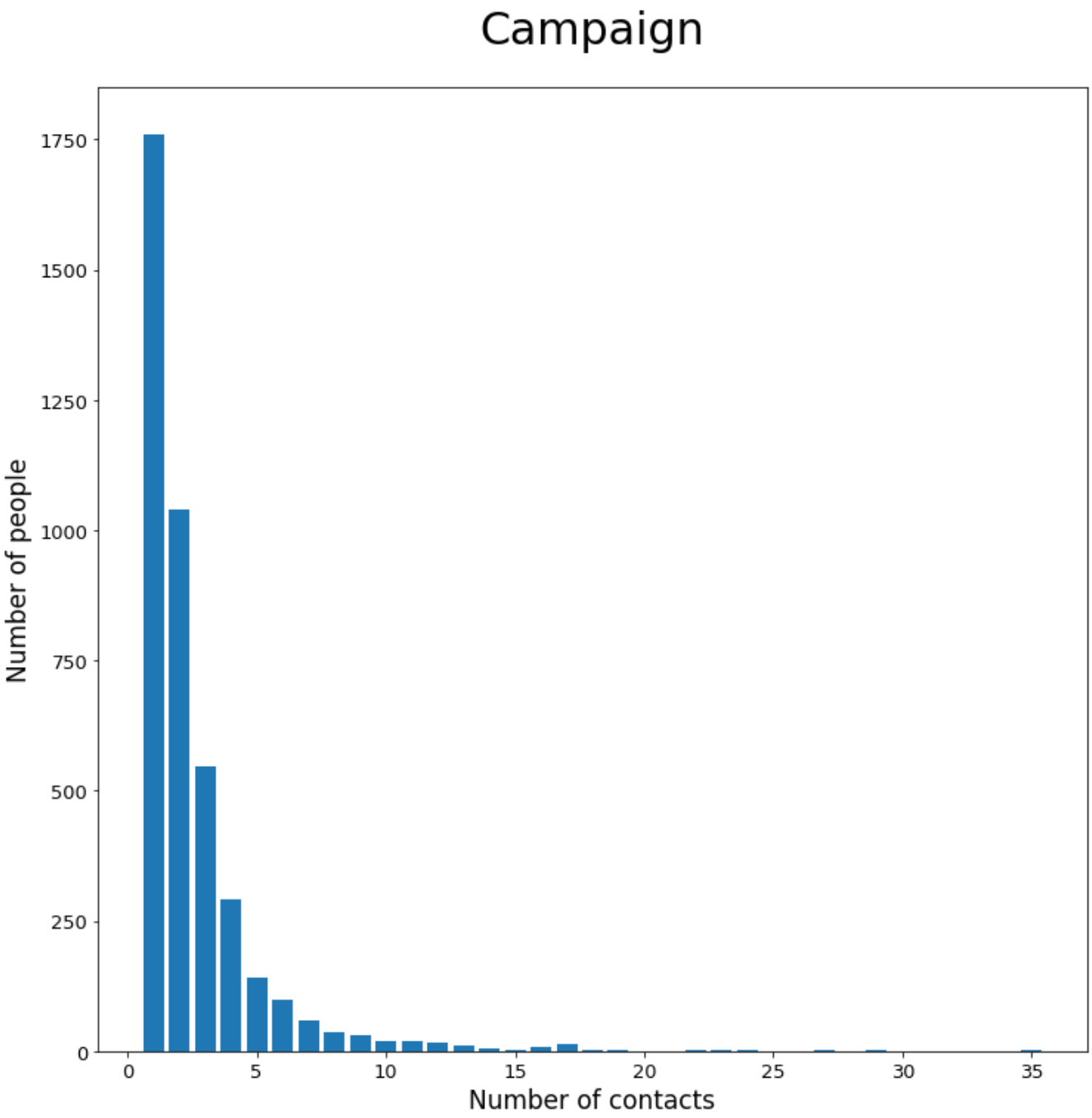
I use horizontal barchart to represent nominal values, the segments of "job" columns' information. The y-axis show the types of job, while x-axis shows the number of people.

Chart 2:



I use piechart to represent ordinal values, "education" columns information. Education is ordinal values type because it has ranking of education level. Each segment shows the percentage of education level. Piechart is suitable for this column because it has only 8 values to represent

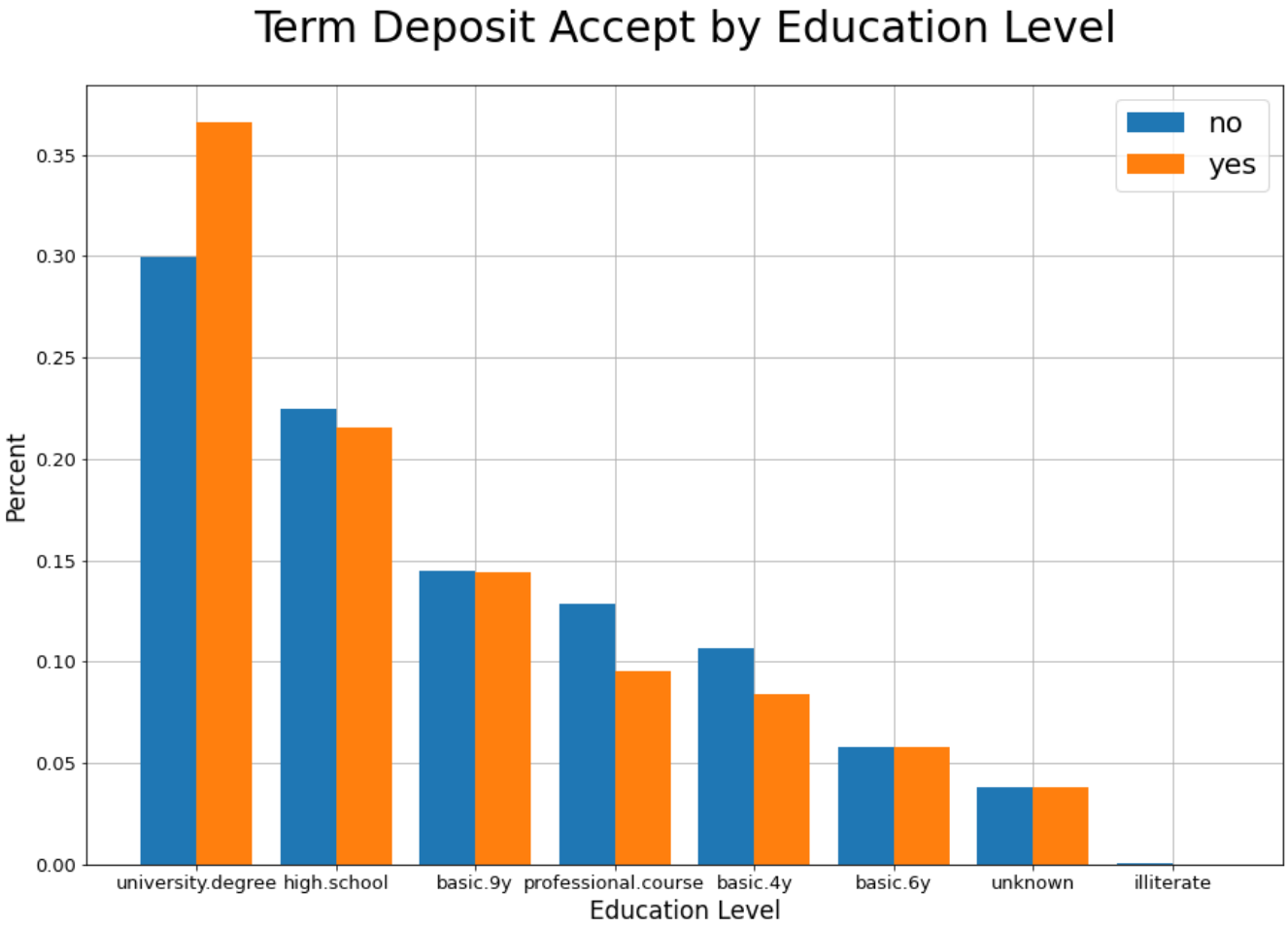
Chart 3:



For columns "campaign", I use barchart to represent the distribution of number of contacts performed during this campaign. y-axis shows the number of people contacted and x-axis shows the number of contacts made.

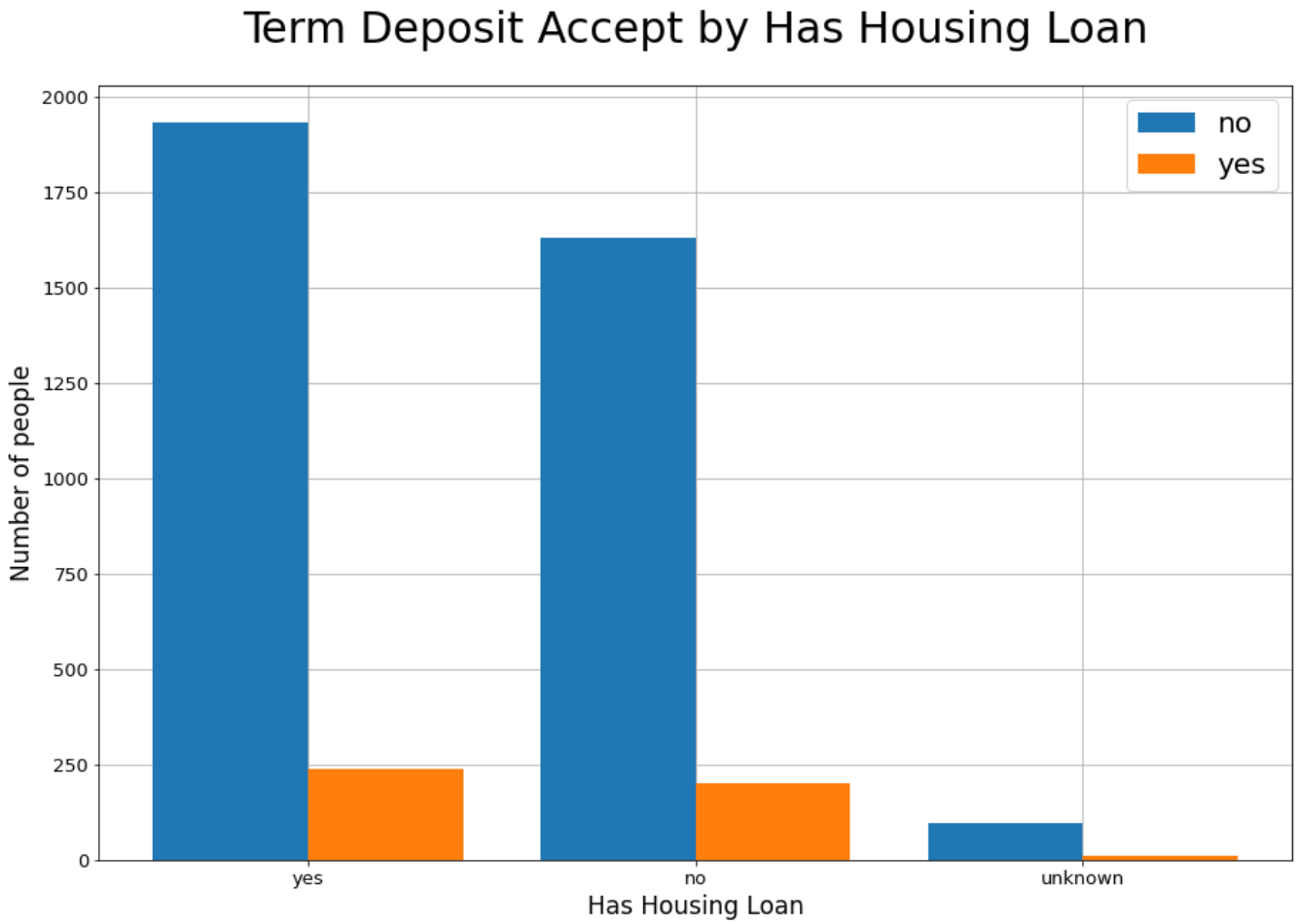
Sub-section 2:

Graph 1:



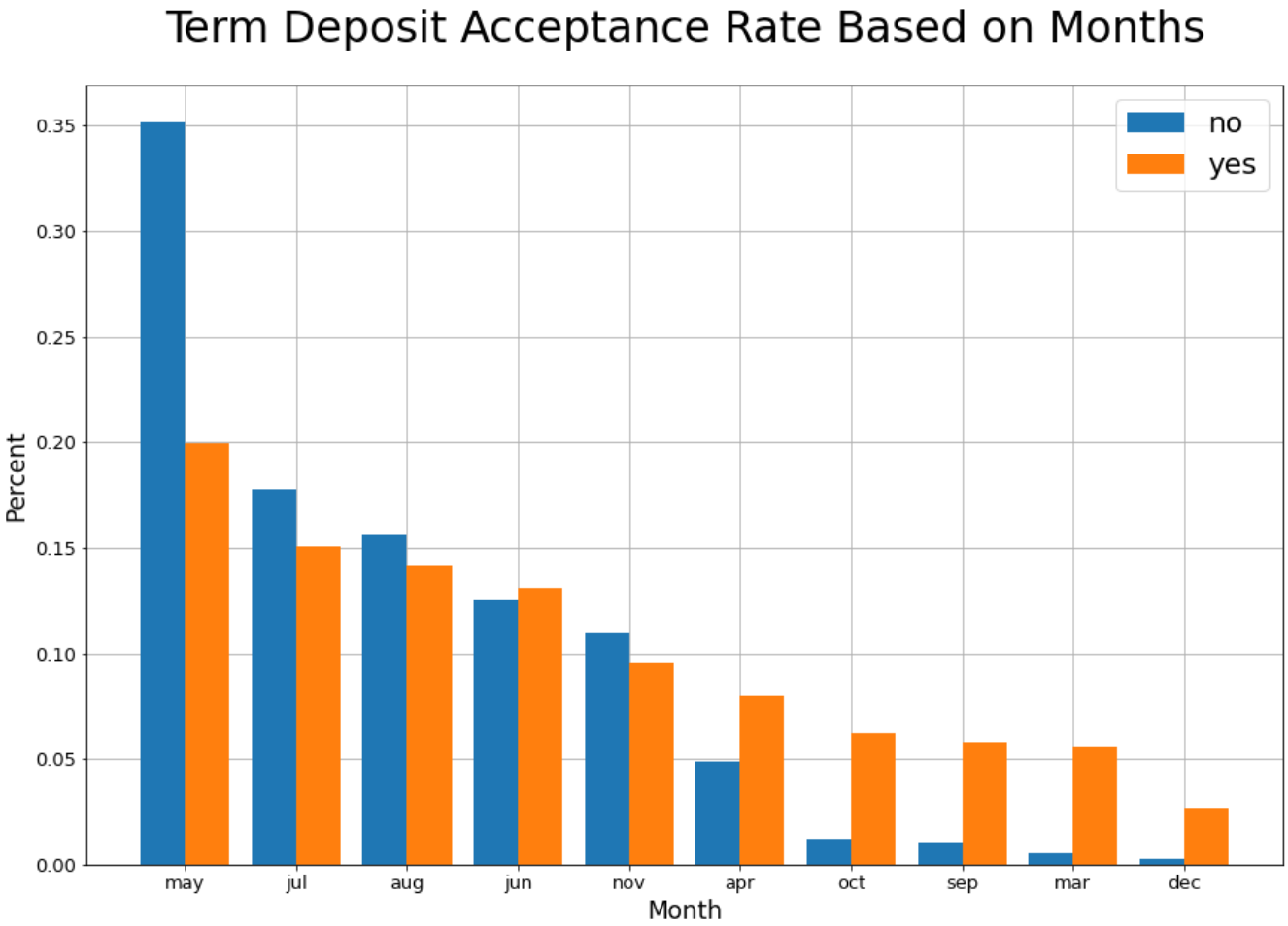
This bar chart shows pair "education" and "y" to see the rate of accepting term deposit by education level. My hypothesis is the higher education level, the higher rate of saying yes to term deposit. Therefore, the bank is likely to contact customers that have higher education level, so that the refuse rate is also high. The relationship of these two columns is increasing based on education level.

Graph 2:



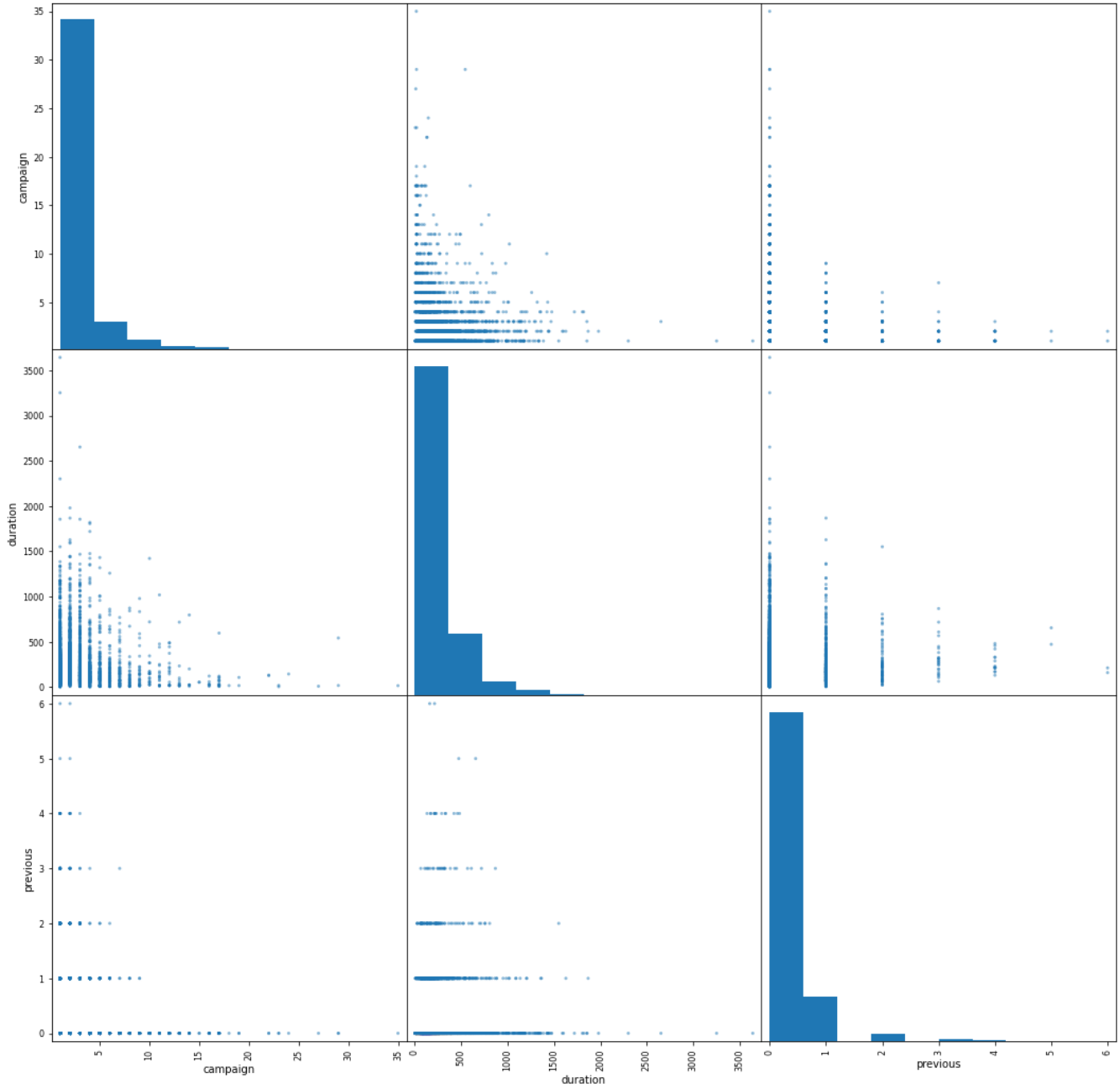
The bar chart represents columns pair "housing" and "y" to see the rate of accepting term deposit for people who has housing loan and not. We can see that people that have housing loan will likely to not accept term deposit.

Graph 3:



The bar chart represents columns pair "month" and "y" to see which part of the year that has the highest term deposit acceptance rate. We can see in the chart, March, April, (end of Quarter 1 and start Quarter 2) and September, October, December (Quarter 4) has the positive rate of saying "yes". My hypothesis is during Quarter 1 and Quarter 4 of the year, the bank increase the euribor rate, so that more people accept term deposit. There is no relationship between these two columns because May has highest rate, while December has lowest rate.

Sub-section 3:

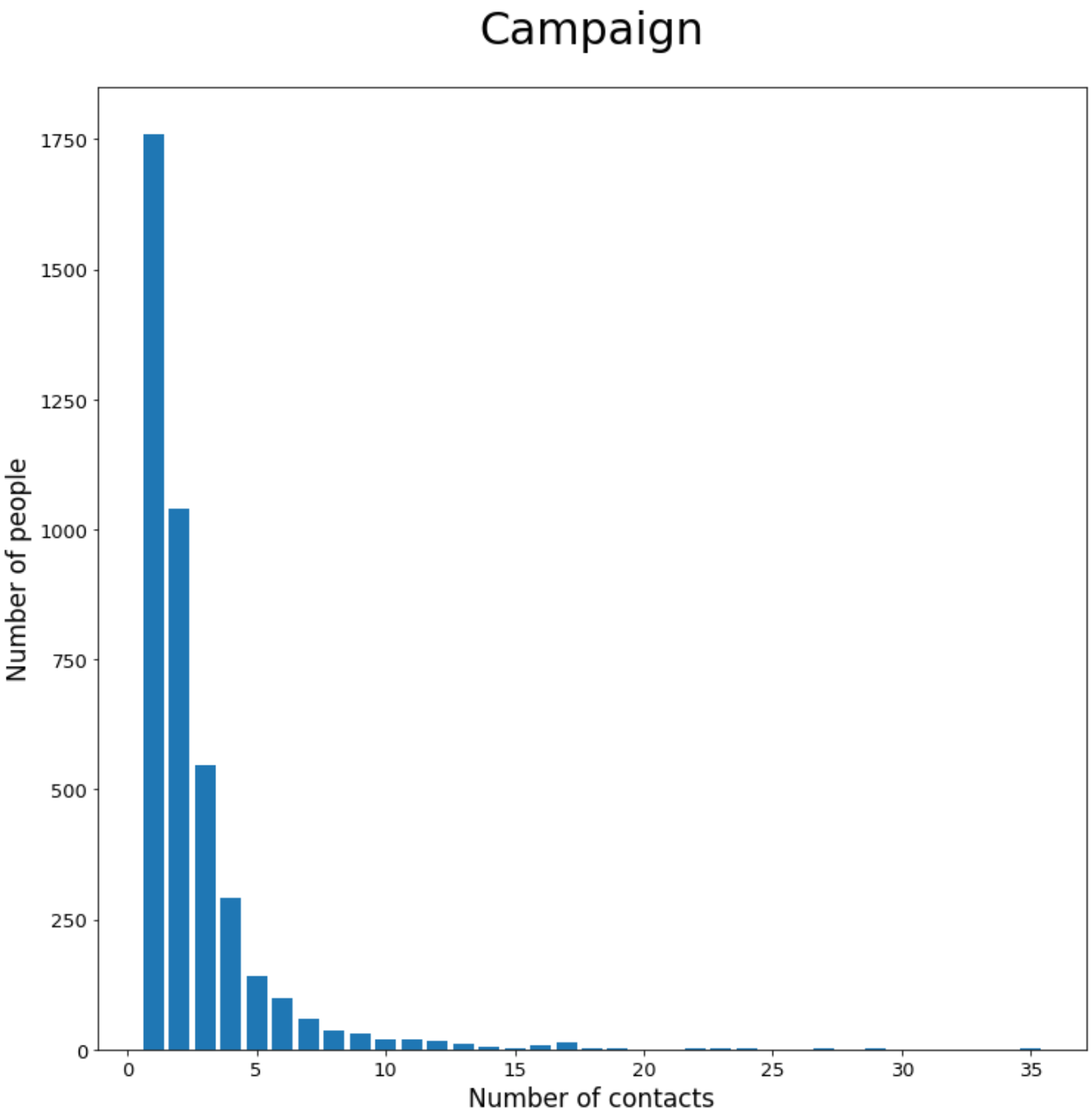


The scatter plot matrix shows three columns: "campaign," "duration," and "previous." We can see that throughout this and earlier campaigns, zero and one customer contacts were created the most. Because the bank is unable to contact customers, call duration is typically zero.

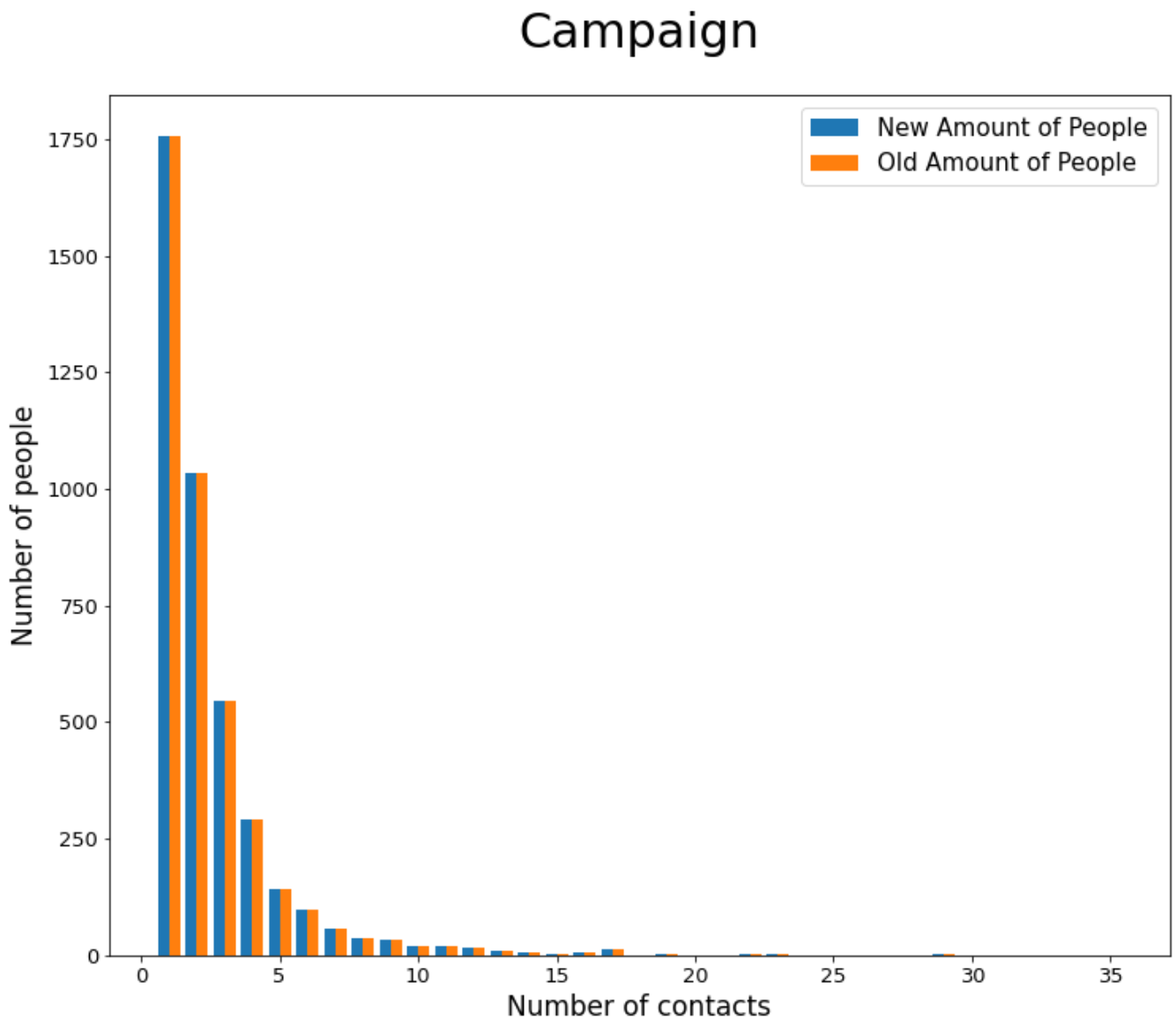
Analysis of Missing Values and Outliers

Sub-section 1:

My first way of dealing with missing value had been mentioned in task 1. In this task, I will remove all missing value rows. Removing rows will lead to the lost of data in other columns too. Therefore, we should only delete rows containing missing value when we cannot do any other replacement.



Above is the graph of deleting missing value. Because of the amount of missing values it too small, I will combine two graphs together to see the different.



If zoom closely, we can see a tiny different. The first and second blue columns are slightly shorter than the orange one. We can see that the impact of dealing missing values with different approaches can be significantly. In this data set, the amount of missing data is small, but for example when we face a data set with lots of missing values, we need to pick a suitable way to deal with them. If we delete all rows containing na value, lots of data will be lost and affect the result.

There are two new csv file:

- "bank_fix1.csv": this csv file contains the data set of the first approach dealing with missing values.
- "bank_fix2.csv": this csv file contains the data set of the second approach dealing with missing values.

Sub-section 2:

a. Do outliers affect standard deviation?

Yes, outliers will affect the standard deviation. Outliers can change the distribution graph from symmetric to skewed. The more outliers, the bigger affect to standard deviation. This is because standard deviation equation needs mean to calculate. And mean will be change significantly if there is outlier, so that standard deviation is affected.

b. When should an outlier not be removed?

We should not remove an outlier when:

When there are too many outliers in the dataframe. Too many outliers means that they have huge affect on the result, and there are no measurement wrong or miss input. If we face this situation, we should make two copies, one that keeps the outliers and one that removes, and compare the results of them. When the outlier does not affect the result alot. For example, when we create a linegraph, if the outlier does not change the regression line, which means that the outlier does not have too many affect to the result. When the data is too important. For example, if the dataset is about car brake testing, it is extremely necessary too keep all of the data.

c. Consider the numeric columns in the bank.csv data, are there outliers in these columns of the data set? How do you detect them?

There are outliers in the numeric columns in bank.csv dataset. To detech the outliers, I will use a statistic equation:

IQR = Q3 - Q1

[Q1 - 1.5 *IQR, Q3 - 1.5* IQR]

Q1 is the median of lower half (or listed as 25% on the describe table above) Q3 is the median of upper half (or listed as 75% on the describe table above) Any values that stands outside of the interquartile range (IQR) is outliers.

Below is the function to detect outliers in this data set

By using the function to find outliers in the notebook, I found that columns "age", "duration", "pdays", "previous", and "cons.conf.idx" have outliers. However, all outliers in this data set are suitable. In "age" column, age larger than approximately 70 years old are outliers, but the data is real so it will not be removed or replaced. Other columns have the same real data, too. This is because the distribution graph of these columns are skewed.

Reference

[1] “Categorical data#,” Categorical data - pandas 1.5.2 documentation. [Online]. Available: https://pandas.pydata.org/docs/user_guide/categorical.html. [Accessed: 24-Nov-2022].

[2] O. S. Statistics and OpenStaxCollege, “Skewness and the mean, median, and mode,” Introductory Statistics, 19-Jul-2013. [Online]. Available: http://pressbooks-dev.oer.hawaii.edu/introductorystatistics/chapter/skewness-and-the-mean-median-and-mode/#:~:text=To%20summarize%2C%20generally%20if%20the,is%20less%20than%20the%20mean. [Accessed: 24-Nov-2022].