

University of Connecticut



Team 6

Group Project: Direct Bank Marketing Data

Latisha Douglas

Madhuri Koyilakonda

Ayush Kumar

Xiang Yu

OPIM 5512 - Python

Professor Ramesh Shankar

1. Introduction

This project is a data model on marketing data set from a Portuguese Bank which is available in the University of California, Irvine (UCI) Machine Learning Repository located at <https://archive.ics.uci.edu/ml/datasets/bank+marketing>.

The objective is to predict what factors enable a client to subscribe to a term deposit. A term deposit or Certificate of Deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution.

The data is from marketing calls made between May 2008 and November 2011 to potential customers.

2. Literature Data

Financial organizations often struggle with differentiating themselves from competitors, who often use similar tactics, market to the same demographics, and offer similar or even the same rates and services. Because of increased competition, it becomes vital for the bank to determine those clients who would subscribe to its term account deposit.

The project is about predicting potential customers who would subscribe to a term account deposit from a Portuguese banking institution by analyzing the data related to the bank's direct marketing campaign.

3. Data Description

The data consists of 41,188 rows and 21 variables. 10 variables are categorical, 10 are numerical, and the target variable 'y', is a binary response indicating '0' for a client not subscribing for a term account, and '1' for a client subscribing for the term account.

Input variables:

1. Age (numeric)
2. Job: type of job (categorical)
3. Marital: marital status (categorical)
4. Education (categorical)
5. Default: has credit in default? (categorical)
6. Housing: has a housing loan? (categorical)
7. Loan: has personal loan? (categorical)

Related with the last contact of the current campaign:

8. Contact: contact communication type (categorical)
9. Month: last contact month of year (categorical)
10. Day_of_week: last contact day of the week (categorical)
11. Duration: last contact duration, in seconds (numeric)

Other attributes:

12. Campaign: number of contacts performed during this campaign and for this client (numeric)
13. Pdays: number of days that passed after client was last contacted from a previous campaign (numeric)
14. Previous: number of contacts performed before this campaign and for this client (numeric)
15. Poutcome: outcome of the previous marketing campaign (categorical)

Social and economic context attributes:

16. Emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. Cons.price.idx: consumer price index - monthly indicator (numeric)
18. Cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. Euribor3m: euribor 3-month rate - daily indicator (numeric)
20. Nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

1. 21 - y - has the client subscribed to a term deposit? (binary)

4. Basic statistics

The following visualizations were completed on the original dataset. The visualizations were performed in Tableau. Images are located in the Appendix.

- **Customers Contacted - by Job** - From the graph of the number of contacts made by job, it can see that admin and blue-collar customers account for half of all customers.
- **Customers Contacted - by Month** - The most customers were contacted during the month of May.
- **Customers Contacted - by Day of the Week** - For days of the week variable, there are no significant differences between different days of week. But Mondays and Thursdays the marketing teams made slightly more calls to their customers.

- **Personal Loan and Housing Loan: Influence Factors** - This plot shows the trend of the number of records for months broken down by personal loan vs housing loan. It can be seen that the trend is almost the same for all the customers despite their housing loan and personal loan status.
- **Marital Distribution** - The chart is a consumer portrayal by marital status. Married customers are more likely to be contacted by the marketing teams.
- **Education Distribution** - Education distribution of the customers is plotted. Over 50% of the customers contacted have degrees above high school.
- **Age Distribution** - The population pyramid is drawn to find patterns of customers' age distribution. For customers who accept a deposit, 37% of them are around their 20s. When the age gets larger, fewer customers would accept the deposit. Customers who accept the deposit are focused between 10 to 40, however, customers who reject the deposit are focused on between 20-50.

5. Data Exploration

- The default column is checked, there are only 3 yes, these rows will not be significant in determining the target variable, hence dropped the rows having default 'yes'
- Checked the correlation matrix and decided to remove columns having correlation greater than 0.7. Dropped 3 columns cons.price.index, eurobinr1, nr.employed
- Changed categorical to continuous, created dummy variables for the following columns, default, housing, loan, contact, poutcome, and target variable y
- Converting job, marital, education, month, day of week categorical variables into dummies will give 35 additional variables, converted those variables into corresponding numbers.
- In pdays column, converted values less than 999 to 0 and values equal to 999 to 1.
- There are no missing values and outliers.

6. Predictive Modeling

This section covers data modeling and model comparison. The final list of variables is selected by employing Random forest and Variation inflation factor techniques. The important variables are then fed to the model as inputs and then model output parameters are analyzed and compared to find the best model.

Variables Selection

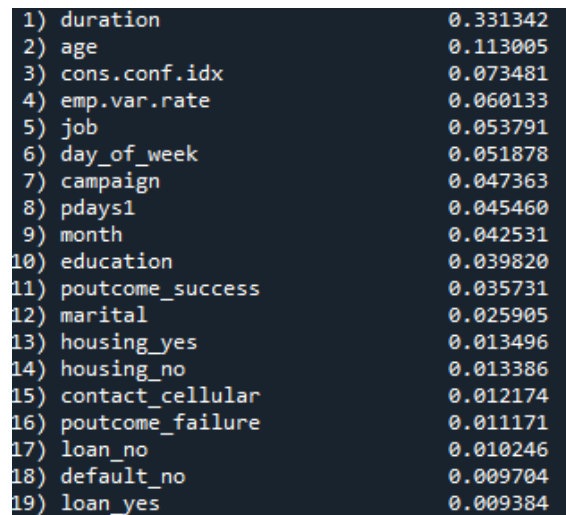
Before modeling the data, it is important to reduce the number of variables as more variables increase the complexity of the model.

Two methods are employed to select the final list of variables: -

- Random forest technique for column contribution
- Variation inflation factor method to remove the variables

Random forest is an ensemble model, which builds multiple decision trees and merges them together to get a more accurate and stable prediction. Great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. *Sklearn* provides a great tool for this task, that measures a feature's importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results, so the sum of all importance is equal to one.

The random forest model is run with all the independent variables and a dependent variable. The number of trees in the forest is set to 10000. The important features can be extracted from this model. The screenshot shows the variable importance by its column contribution.



1)	duration	0.331342
2)	age	0.113005
3)	cons.conf.idx	0.073481
4)	emp.var.rate	0.060133
5)	job	0.053791
6)	day_of_week	0.051878
7)	campaign	0.047363
8)	pdays1	0.045460
9)	month	0.042531
10)	education	0.039820
11)	poutcome_success	0.035731
12)	marital	0.025905
13)	housing_yes	0.013496
14)	housing_no	0.013386
15)	contact_cellular	0.012174
16)	poutcome_failure	0.011171
17)	loan_no	0.010246
18)	default_no	0.009704
19)	loan_yes	0.009384

It is inferred from the results that *duration*, *age*, *cons.conf.inx*, *emp.var. rate* and *job* are top 5 important variables with *duration* and *age* as most important variables.

In the second technique, criteria of *variation Inflation factor* is used to eliminate the variables.

A variance inflation factor (VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors in a model.

- If $VIF = 1$ then predictors are not correlated.

- If $1 < \text{VIF} < 6$ then predictors are moderately correlated.
- If $\text{VIF} > 6$ then predictors are highly correlated.

Data is fitted into OLS (ordinary least square) model and then VIF of the variables is computed. The variable with $\text{VIF} > 6$ is eliminated. Before elimination, the variable contribution from the random forest analysis is observed for its significance. The screenshot shows the VIF values of all the variables after first iteration.

Iteration1

	variables	VIF
0	age	15.202481
1	job	2.754036
2	marital	14.140788
3	education	4.109274
4	month	7.532570
5	day_of_week	5.356899
6	duration	2.002501
7	campaign	1.917850
8	emp.var.rate	1.586814
9	cons.conf.idx	46.602530
10	pdays1	11.096083
11	default_no	5.185261
12	housing_no	inf
13	housing_yes	inf
14	loan_no	inf
15	loan_yes	inf
16	contact_cellular	4.098564
17	poutcome_failure	1.397807
18	poutcome_success	11.121035

The variable *housing_no* is dropped because VIF is equal to Infinity.

A single variable is eliminated each time after running the model. Total 6 iterations were performed until the VIF of the variables was less than 6. The variable *age* is not eliminated, despite having a VIF of 8 because *age* is an important variable from random forest analysis. The screenshot shows the VIF values of all the variables after sixth iteration. These variables are final set of important variables that is fed to different models.

Sixth Iteration:

	variables	VIF
0	age	8.775595
1	job	2.689704
2	education	3.958121
3	month	7.162831
4	day_of_week	4.767630
5	duration	1.947183
6	campaign	1.863206
7	emp.var.rate	1.565836
8	pdays1	1.154504
9	default_no	4.583262
10	housing_yes	2.074348
11	loan_yes	1.178915
12	contact_cellular	3.790476
13	poutcome_failure	1.329977

After selecting the variables, data is split into training and test partitions in 70:30 ratio. The data is then standardized using StandardScaler () function. The dataset is imbalanced dataset. The baseline accuracy is 11.35% as we are interested in the customers who subscribe to a term deposit.

6.2 Modeling

Total 6 models are used to analyze the output. These models are Logistic Regression, Perceptron, SVM, Naïve Bayes, Random forest and Neural Networks.

6.2.1 Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The screenshot shows the model's confusion matrix and other output parameters.

			Predication for Test		Total Accuracy:	90.85%
			Yes	No	Precision	65.93%
			1	0	Baseline Accuracy of 'yes'	11.35%
Actual	Yes	1	563	839	Lift:	5.81
			(True Positive)	(False Negative)	Sensitivity:	40.16%
	No	0	291	10663	Specificity:	97.34%
			(False Positive)	(True Negative)		

6.2.2 Perceptron

This model takes an input, aggregates it and returns 1 only if the aggregated sum is more than threshold. The screenshot shows the model's confusion matrix and other output parameters.

			Predication for Test		Total Accuracy:	89.86%
			Yes	No		
			1	0		
Actual	Yes	1	453	949	Precision	59.84%
			(True Positive)	(False Negative)	Baseline Accuracy of 'yes'	11.35%
	No	0	304	10650	Lift:	5.27
			(False Positive)	(True Negative)	Sensitivity:	32.31%
					Specificity:	97.22%

6.2.3 SVM

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. Refer to the screen for model's output parameters.

			Predication for Test		Total Accuracy:	91.02%		
			Yes	No			Precision	69.16%
			1	0				
Actual	Yes	1	527	875	Lift:	6.09		
			(True Positive)	(False Negative)	Sensitivity:	37.59%		
	No	0	235	10719	Specificity:	97.85%		
			(False Positive)	(True Negative)				

6.2.4 Naïve Bayes

This model is based on the Bayes Theorem for calculating probabilities and conditional probabilities. Find the confusion matrix and output parameters of this model below.

			Predication for Test		Total Accuracy:	90.00%		
			Yes	No			Precision	56.78%
			1	0				
Actual	Yes	1	695	707	Lift:	5.00		
			(True Positive)	(False Negative)			Sensitivity:	49.57%
	No	0	529	10425	Specificity:	95.17%		
			(False Positive)	(True Negative)				

6.2.5 Random Forest

			Predication for Test		Total Accuracy:	91.26%
			Yes	No		
			1	0		
Actual	Yes	1	670	732	Precision	65.82%
			(True Positive)	(False Negative)		
	No	0	348	10606	Lift:	5.80
			(False Positive)	(True Negative)		
					Specificity:	96.82%

Random forest model was run with 10000 trees in the forest. Find below the confusion matrix and the model's output parameters.

6.2.6 Neural Networks

In this model, three hidden layers are used each with size of 20. Find below the confusion matrix and the model's output parameters.

			Predication for Test		Total Accuracy:	90.75%		
			Yes	No			Precision	62.61%
			1	0				
Actual	Yes	1	643	759	Baseline Accuracy of 'yes'	11.35%		
			(True Positive)	(False Negative)			Lift:	5.52
	No	0	384	10570				
			(False Positive)	(True Negative)			Specificity:	96.49%

7. Model Comparison

The purpose of the model comparison is to select the best model. The different output parameters are calculated for these models. For the use case, the model should misclassify the 'yes' for minimum number of times i.e. the model should have the minimum number of false negatives. If the model predicts a greater number of 'yes' as 'no', then there is a cost associated with each misclassified prediction. It's not feasible for a banking institution to deploy a model which increases its cost.

The screenshot shows the model comparison metric.

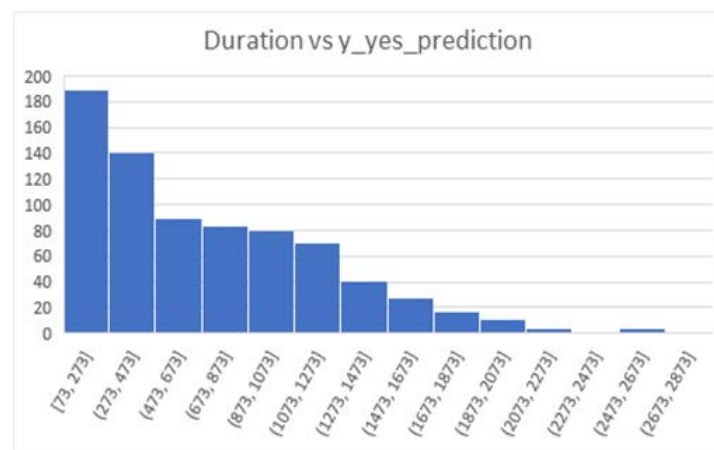
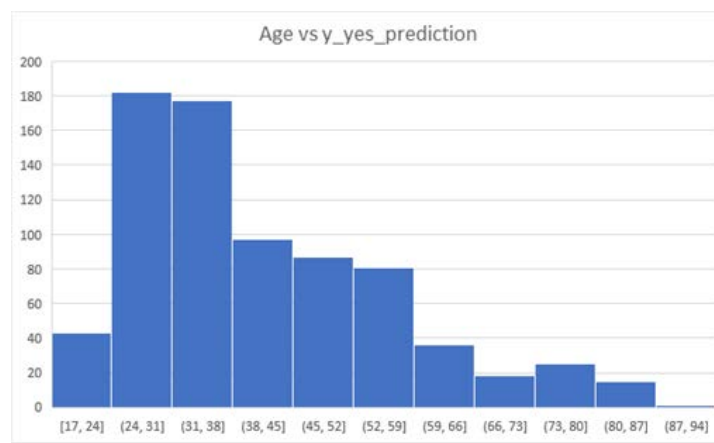
Model /Output	Total Accuracy	precision	Lift	Recall	Specificity	F1 Score	AUC
Logistic	0.91	0.66	5.81	0.40	0.97	0.90	0.69
Perceptron	0.90	0.60	5.27	0.32	0.97	0.88	0.65
SVM	0.91	0.69	6.09	0.38	0.98	0.90	0.68
Naïve Bayes	0.90	0.57	5.00	0.50	0.95	0.90	0.72
Random forest	0.91	0.66	5.80	0.48	0.97	0.91	0.72
Neural	0.91	0.63	5.52	0.46	0.96	0.90	0.71

As compared to other models, Naïve Bayes model has minimum number of false negatives (707) and highest Recall or sensitivity of 50%.

So, Naïve Bayes classifier is the best model.

8. Findings

- When the call duration is between 73 to 473 seconds customers are more likely to subscribe for term deposit.
- Age group 24 to 38 are more likely to subscribe for term deposit



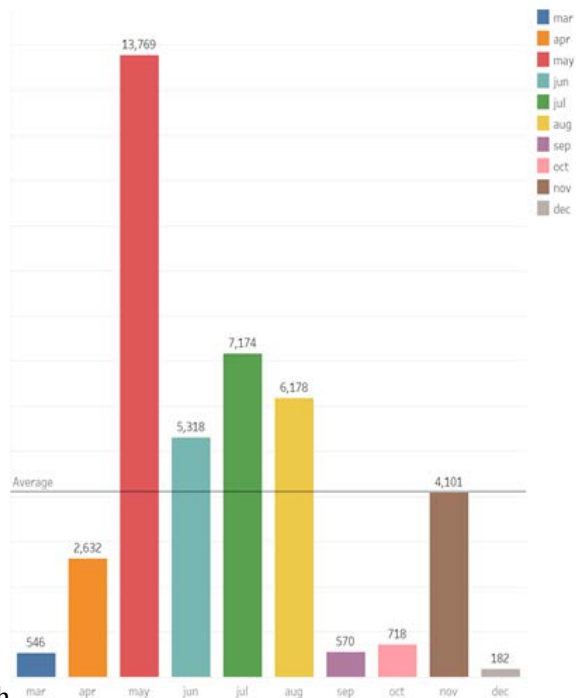
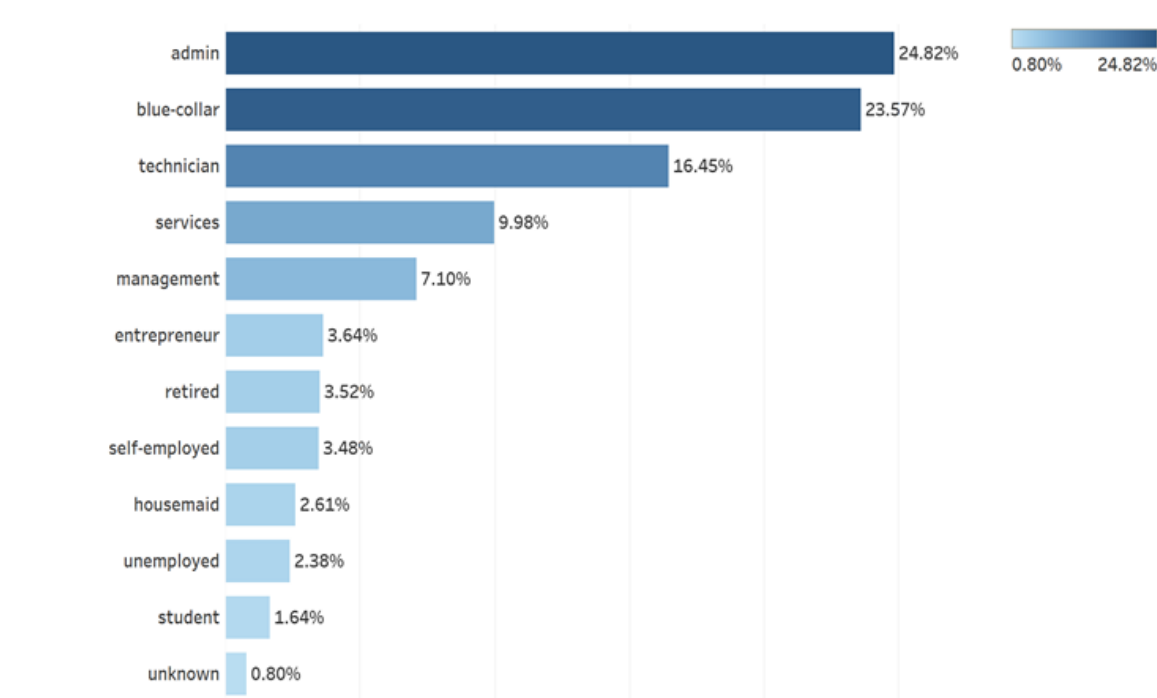
9. Implications/recommendations for business

- Make precise marketing strategy to focus on blue-collar and technician job holders
- Marketing calls that end within 8 minutes, increase the chance for the customers to subscribe to term deposit.
- Marketing Campaigns should be more targeted to Age group 24 to 38

Appendix

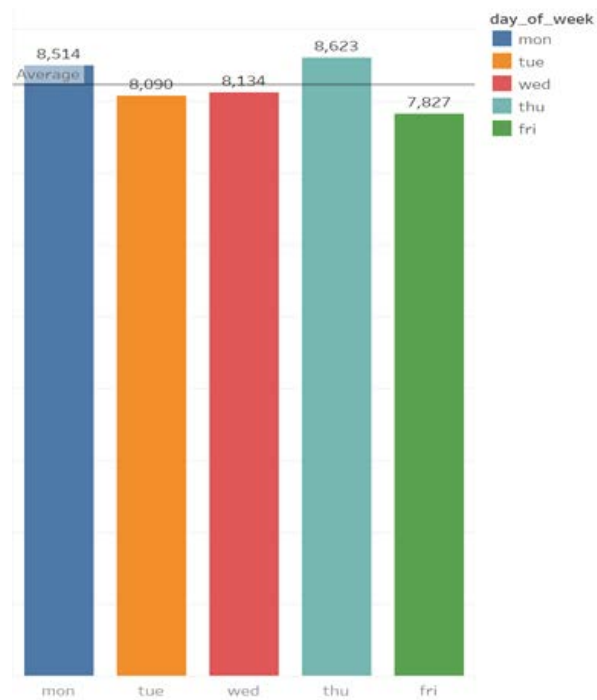
I. Tableau Visualization

Customers Contacted - by Job

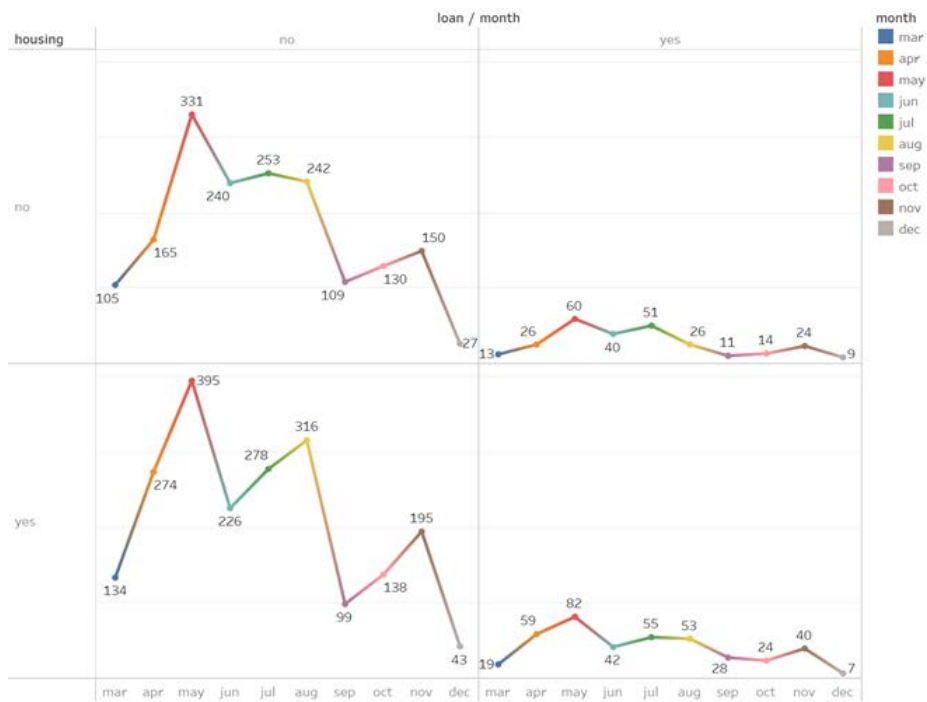


Customers Contacted - by Month

Customers Contacted - by Day of the Week

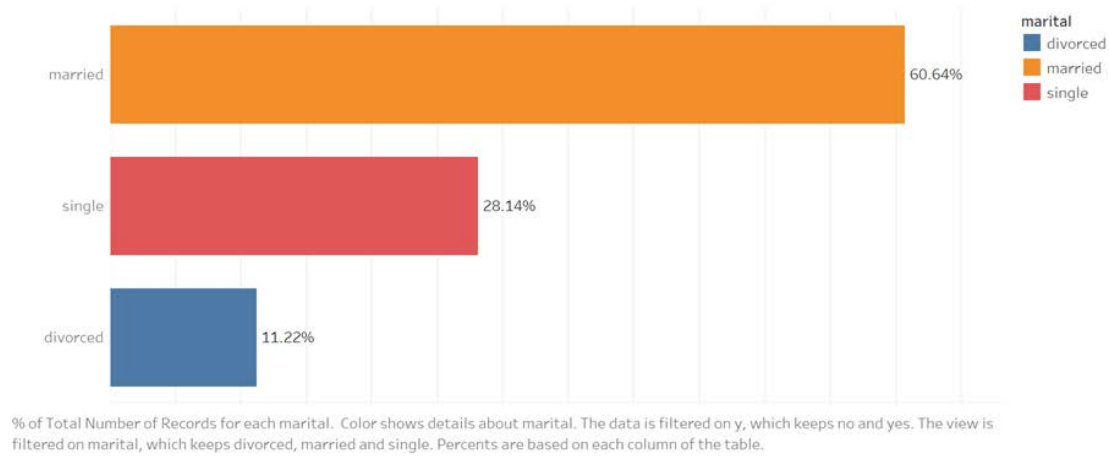


Personal Loan and Housing Loan: Influence Factors

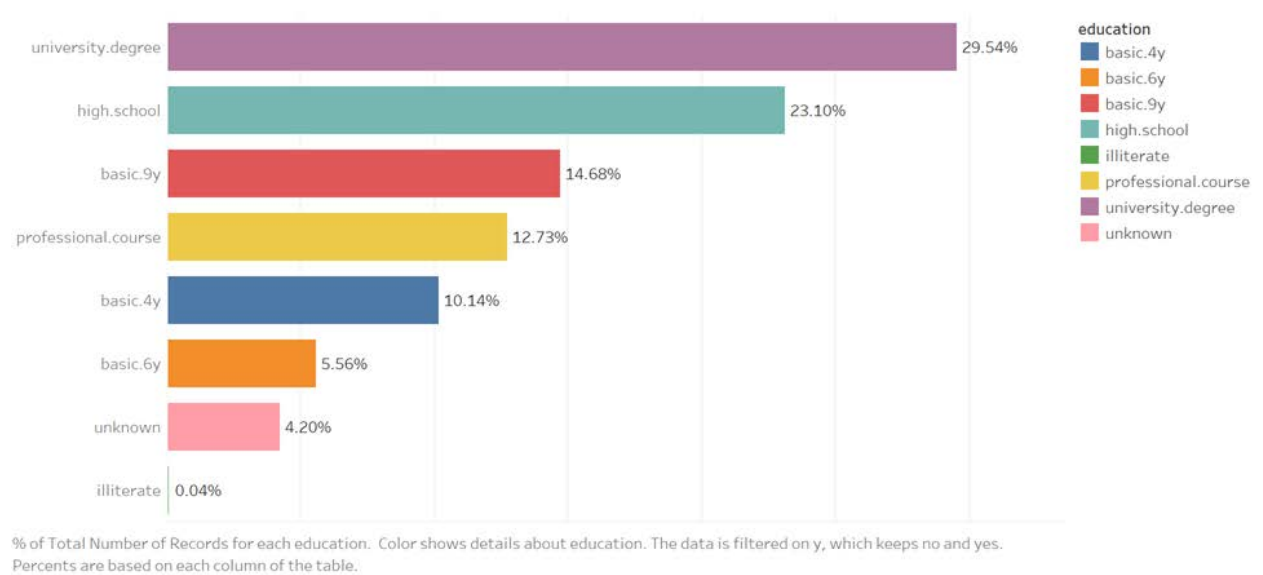


The trend of sum of Number of Records for month broken down by loan vs. housing. Color shows details about month. The data is filtered on y, which keeps yes. The view is filtered on housing, which keeps no and yes.

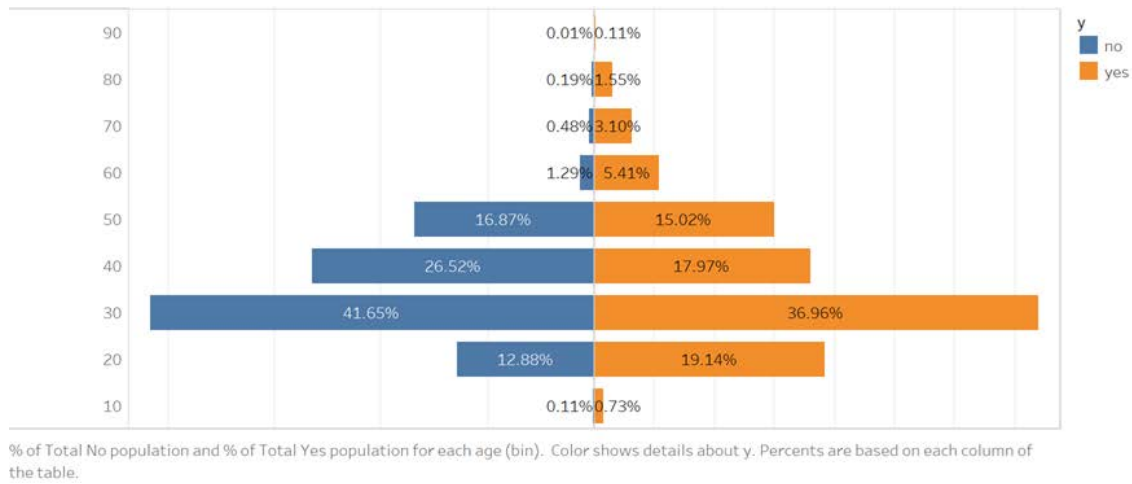
Marital Distribution



Education Distribution



Age Distribution



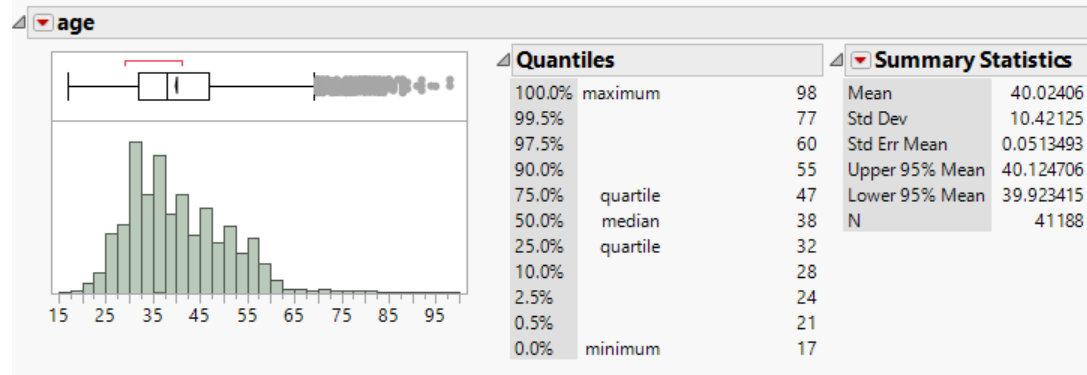
II. Summary Statistics

Variable Descriptions

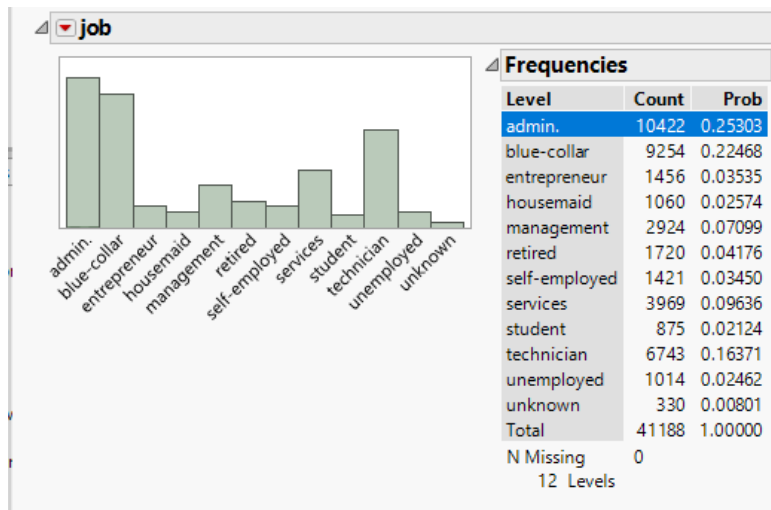
Variable Description	Type
Age	Numeric Continuous
Job	Character Nominal
Marital	Character Nominal
Education	Character Nominal
Default	Character Nominal
Housing	Character Nominal
Loan	Character Nominal
Contract	Character Nominal
Month	Character Nominal
day_of_week	Character Nominal
duration	Numeric Continuous
campaign	Numeric Continuous
pdays	Numeric Continuous
Previous	Numeric Continuous
Poutcome	Character Nominal
emp.var.rate	Numeric Continuous
coms.price.idx	Numeric Continuous
cons.conf.idx	Numeric Continuous
euribor3m	Numeric Continuous
nr.employed	Numeric Continuous
y (Target Variable)	Character Nominal

III. Histograms

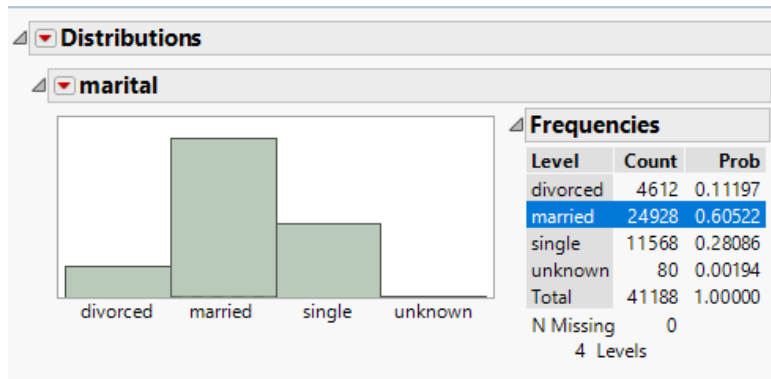
Age Histogram



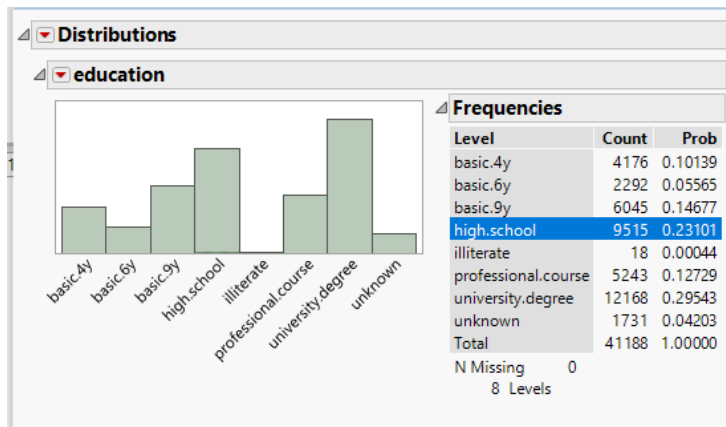
Job Histogram



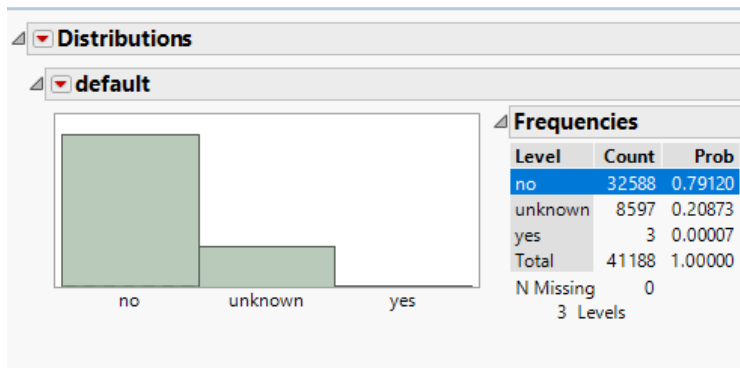
Marital Histogram



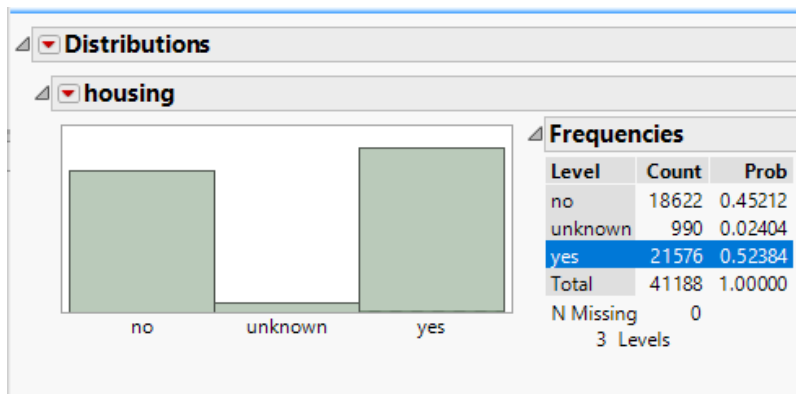
Education Histogram



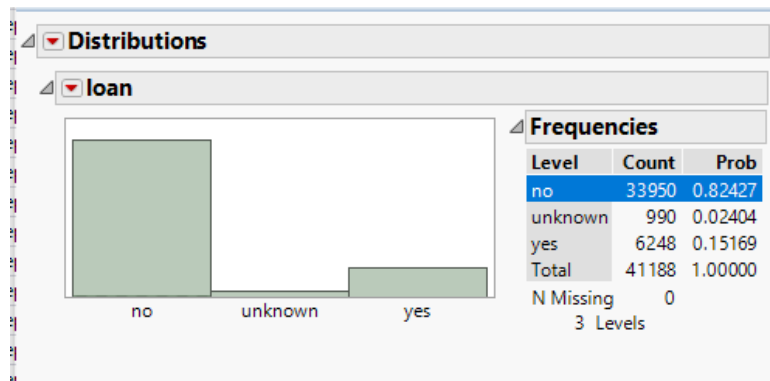
Default Histogram



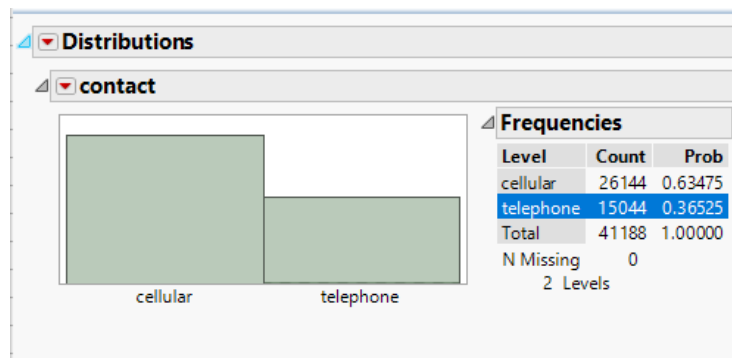
Housing: Has housing loan? Histogram:



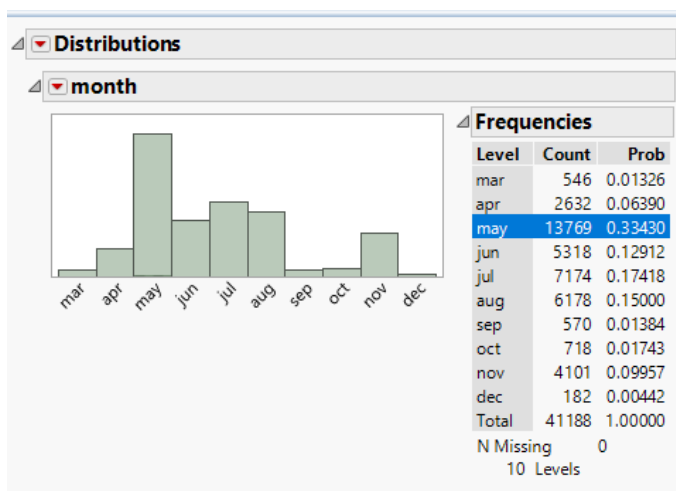
Loan: Has personal loan? Histogram:



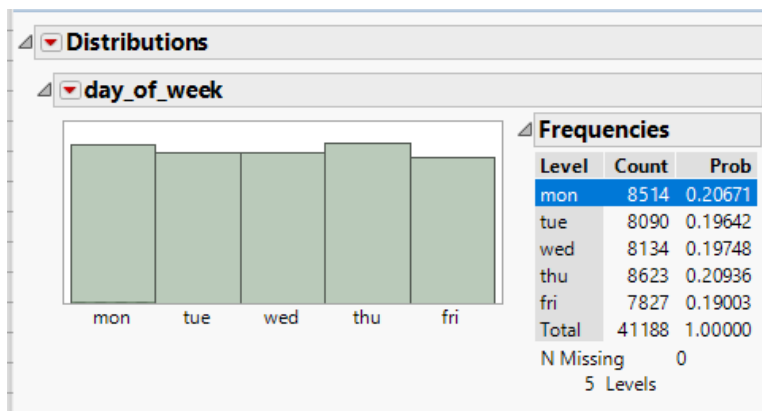
Contract: Contact Communication Type: 'Cellular', Telephone' Histogram:



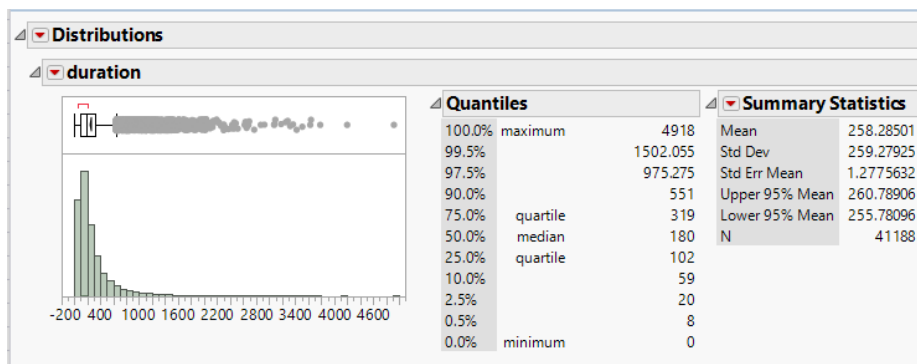
Month: Last Contact month of the year



Day_of_week: Last contact day of the week

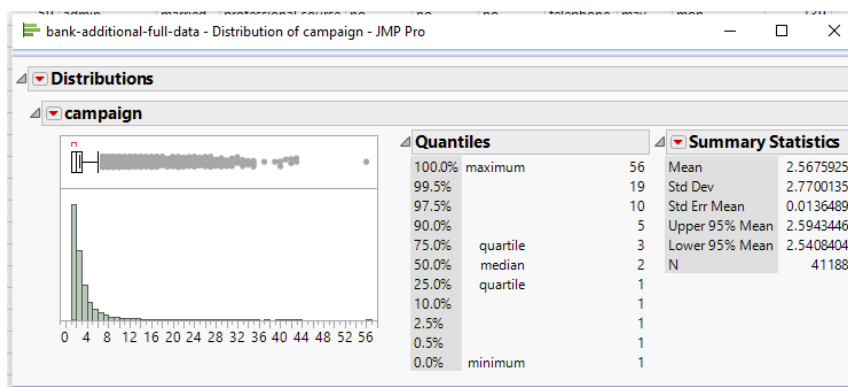


Duration: Last contact duration in seconds

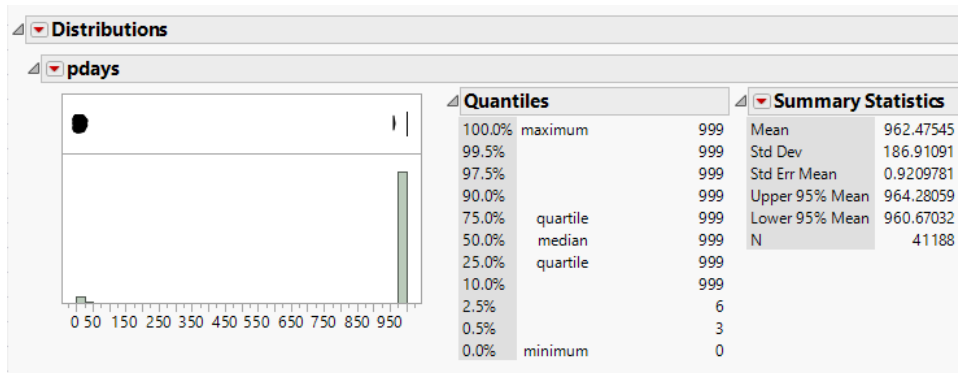


Would binning be a good idea?

Campaign: Number of contacts performed during this campaign and for this client

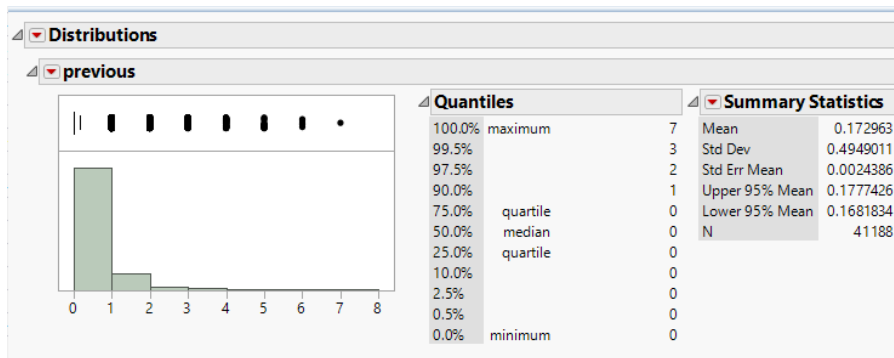


Pdays

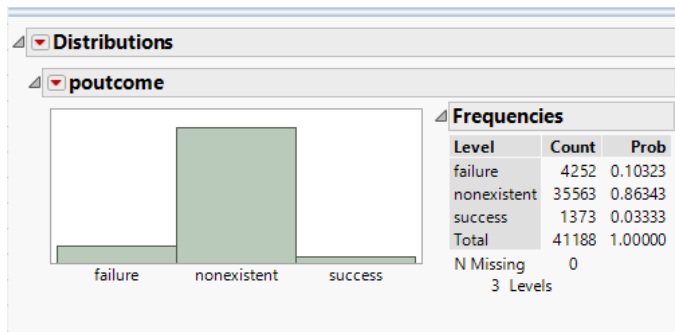


Number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)

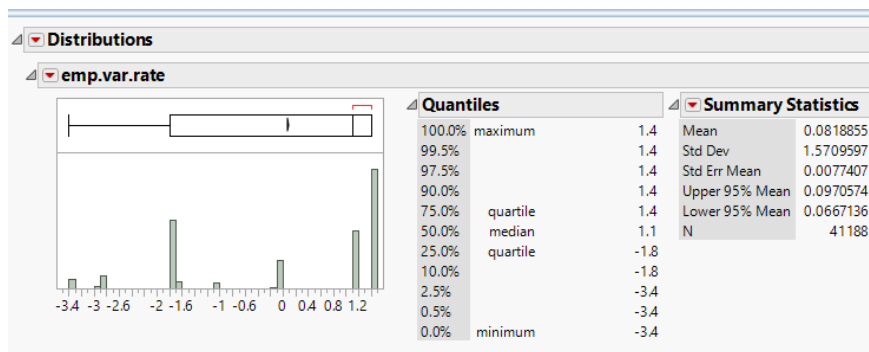
Previous Number of contacts performed before this campaign and for this.



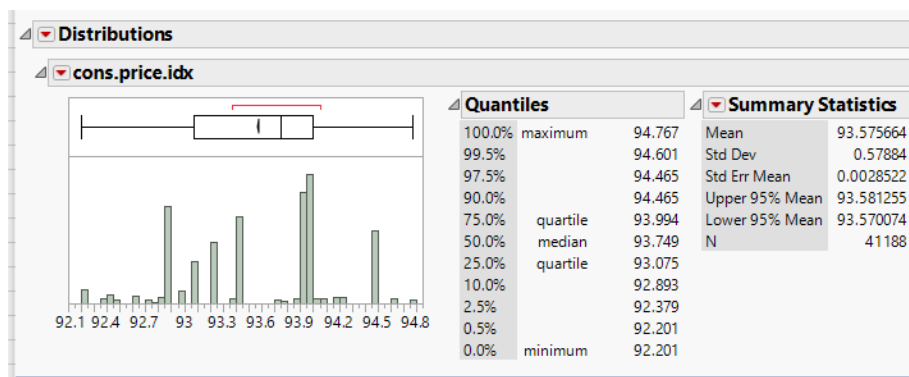
Poutcome: outcome of the previous marketing campaign



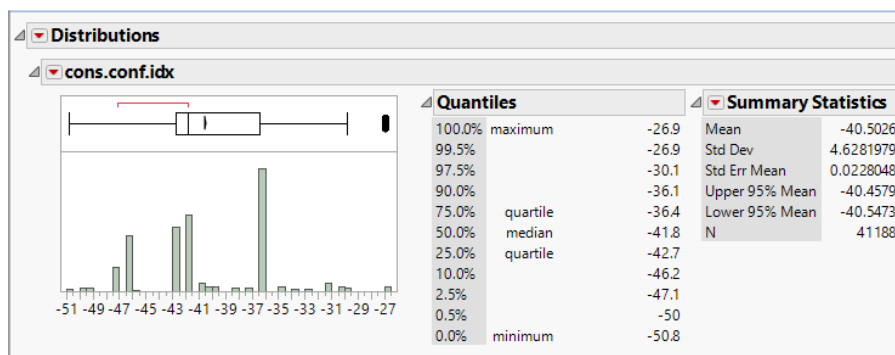
Emp.var.rate: employment variation rate - quarterly indicator



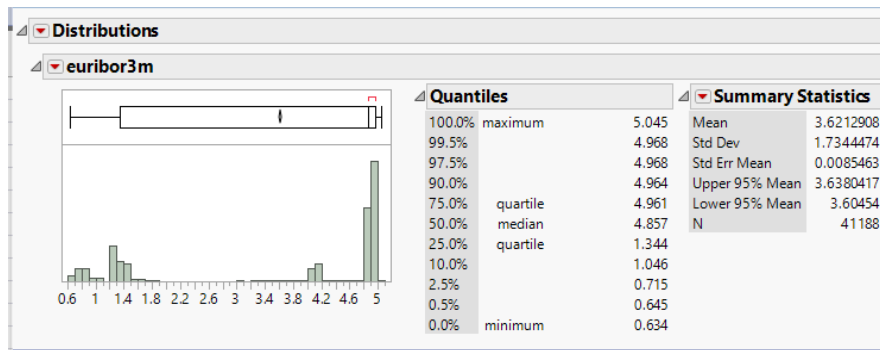
Cons.price.idx: consumer price index - monthly indicator



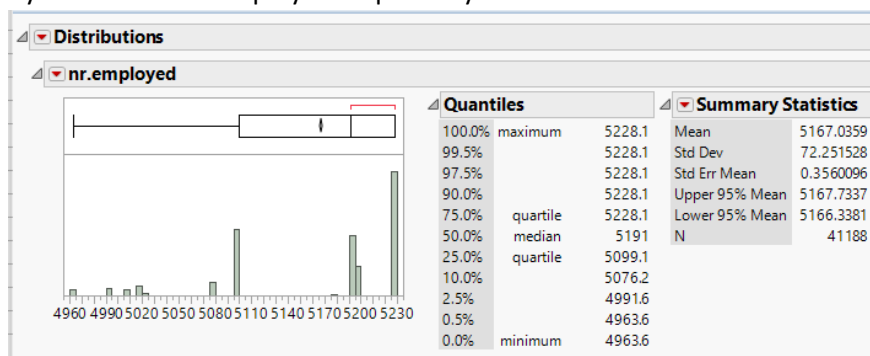
Cons.conf.idx: consumer confidence index - monthly indicator



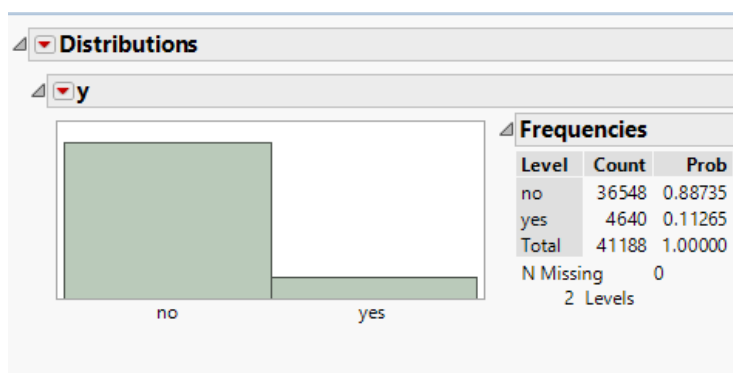
Euribor3m: euribor 3 month rate - daily indicator



Nr.employed: number of employees - quarterly indicator



Target variable: Y - has the client subscribed a term deposit? (binary: 'yes','no')



Summary Statistics

