# Name:

# Part 1: Exploration

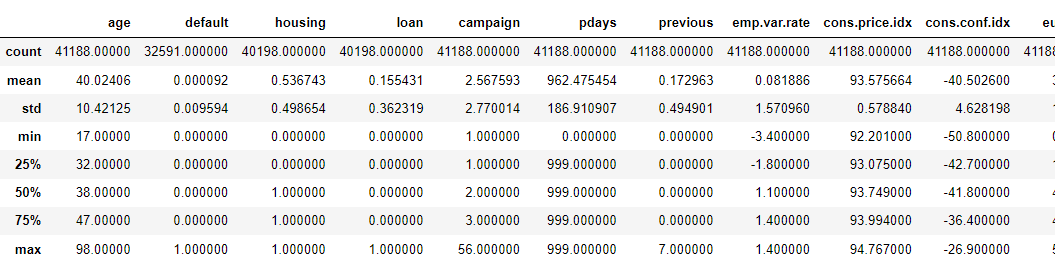
In this section, you should describe:

* Results of **descriptive statistics** about your columns.
* Your investigation into **missing values** and how you dealt with them. (Remember: leaving them alone is a valid option if it’s justified!)
* Your investigation into **outliers** and how you dealt with them. (Remember: leaving them alone is a valid option if it’s justified!)
* The exploration of the relationship between your potential features and the target, e.g. answering questions like, “**How did the percentage of people who bought the product vary with the age of customers**?” and “**Are older or younger customers more likely to buy?**” Based on these answers, which **features** did you choose for modeling?

Provide supporting visuals where appropriate.

**Analysis**

1. Find the count, mean, standard deviation, min, and max values of all numerical columns.



1. There were missing values in default, housing, and loan columns. To handle these, first, I have checked the relation between the target variable and variables having missing values. The results showed that the high percentage of missing values related to the clients who didn’t subscribe to a term deposit.
2. There are outliers in age, Campaign, pdays, and previous columns. To handle the outliers, check the IQR of some columns and drop the values above and below the IQR range; other columns, such as pdays (999 means the client did not previously contact), have a considerable number of outliers and prefer to drop.
3. The analysis shows that the percentage of people who bought the product varies with the age of customers, and customers having the age group between 25 and 40 are more likely to buy.
4. Based upon the EDA, all features are used for modeling except pdays, default because of huge outliers and missing values.

# Part 2: Modeling

After completing the Jupyter notebook and training 2 different machine learning models, you should:

* Describe the features you chose for each model.
* Describe the model you used for each model.
* Detail the results of both models.
  + What was their accuracy score?
  + What did the confusion matrix reveal? Include some discussion about false positives and false negatives.
* Decide which model performed better overall and justify your decision. Is it because one has a higher accuracy, or is it the makeup of the confusion matrices?

**Analysis**

1. ['age', 'housing', 'loan', 'month', 'day\_of\_week', 'campaign', ‘previous', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx','euribor3m', 'nr.employed', 'output','job\_blue-collar', 'job\_entrepreneur', 'job\_housemaid','job\_management', 'job\_retired', 'job\_self-employed', 'job\_services','job\_student', 'job\_technician', 'job\_unemployed', 'job\_unknown', ‘marital\_married', 'marital\_single', 'marital\_unknown','education\_basic.6y', 'education\_basic.9y', 'education\_high.school','education\_illiterate','education\_professional.course','education\_university.degree','education\_unknown','contact\_telephone']
2. 1. Logistic Regression, K nearest neighbours (neighbours=5,7)
3. **For logistic regression:**

The accuracy of the model Logistic Regression is **89.32%**

**Confusion matrix:**

array([[6412, 54],

[ 725, 100]]

**For K-nearest neighbours:**

The accuracy score for the knn2 model is **88.79%**

**Confusion matrix:**

array([[6383, 83],

[ 734, 91]], dtype=int64)

1. Both models have a similar **accuracy score of ~ 89%.** So, both are good, but we can use different cross-validation techniques using gridsreachCV to find the beat model.