**Project 2 - Bank Marketing**

**Introduction:**

asdf

**Data Description:**

The dataset we will be evaluating includes the results of a Portuguese banking institution’s marketing campaign data. The data was collected through phone calls, often more than once to the same client for clarity in the response. The output variable for analysis includes a “yes/no” response category based on if each client has subscribed a term deposit or not. The purpose of this project is to use this data and see how accurate we can predict whether the client subscribes or not.

**Variables:**

In order to better understand the dataset we were working with we begin by separating the data attributes into its data type groups. Considering the data splits 50:50 into a numeric and categorical grouping, we will first look into only the non-categorical data:

Numeric Variables:

* **Age** = age of client
* **Duration** = contact duration (in seconds
* **Campaign** = # of days contacts performed during campaign on the client
* **Pdays** = # of days passed by after client was last contacted from previous campaign)
* **Previous** = prior # of contacts performed before the campaign and for the client)
* **Emp.var.rate** (employment variation rate) = measure of the extent to which
* **Cons.price.idx** = (consumer price index)
* **Cons.confg.idx** = (consumer confidence index)
* **Eurobor3m** = (Euribor interest rate) interest rate at which European banks lend one another funds denominated in euros
* **Nr.employed** = (number of employees)

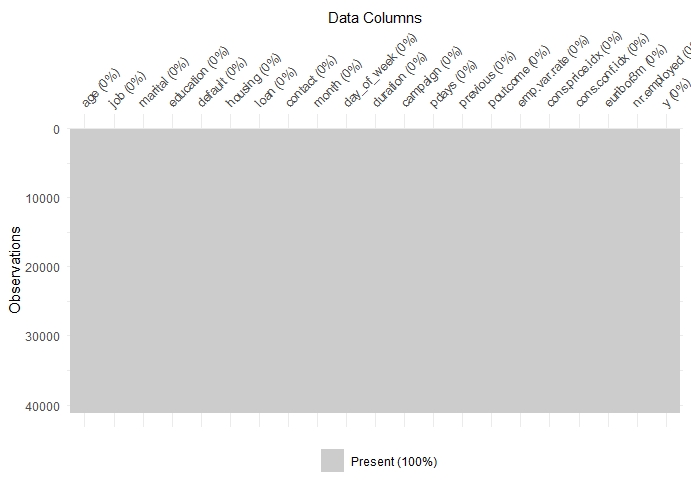
Categorical Variables (with factor ID):

* **Job** = job type
  + [1] 'admin.'
  + [2] 'blue-collar'
  + [3] 'entrepreneur'
  + [4] 'housemaid'
  + [5] 'management'
  + [6] 'retired',
  + [7] 'self-employed'
  + [8] 'services',
  + [9] 'student'
  + [10] 'technician'
  + [11] 'unemployed'
  + [12] 'unknown'
* **Marital** = marital status
  + [1] 'divorced' (or widowed)
  + [2] 'married',
  + [3] 'single',
  + [4] 'unknown'
* **Education** = level of education
  + [1] 'basic.4y'
  + [2] 'basic.6y'
  + [3] 'basic.9y'
  + [4] 'high.school'
  + [5] 'illiterate'
  + [6]'professional.course'
  + [7]'university.degree
  + [8] 'unknown'
* **Default** = credit in default or not
  + [1] no, [2] yes
* **Housing** = has housing loan or not
  + [1] no, [2] yes
* **Loan** = has personal loan or not
  + [1] no, [2] yes
* **Contact** = communication type
  + [1] 'cellular' [2] ,'telephone'
* **Month** = last contact month of the year
* **Day\_of\_Week** = last contact day of the year
* **Poutcome** = outcome of previous marketing campaign

**Data Tidying:**

As part of the data tidying process, we first want to verify if the dataset contains missing data for any of the attributes. As shown in Figure A.1 below, we can confirm that all data is being populated. However, upon closer inspection of the data, we found that instead of populating missing data as NULL, it was instead replaced with “unknown” as shown in Figure A.2. We then imputed these “unknown” variables with non-missing value means as shown in Figure A.2.

**Figure A.1 - Raw Data Figure A.3 - After Replacing ‘unknown’ with NAs**



We next wanted to determine if there was any multicollinearity present between any of the numeric variables.

**Exploratory Data Analysis:**

We first wanted to explore if there was any relationships between the # of campaigns and the months it was run in (Figure B.1). We can see that interestingly, most of the campaigns were done during the summer months with less and less contacts made to clients closer to winter months. Additionally, we can rule out the Day\_of\_Week variable as a potential explanatory variable as it appears there is a somewhat even distribution of calls being made throughout the week.

**Figure B.1**

