Module 2: Assignment 1

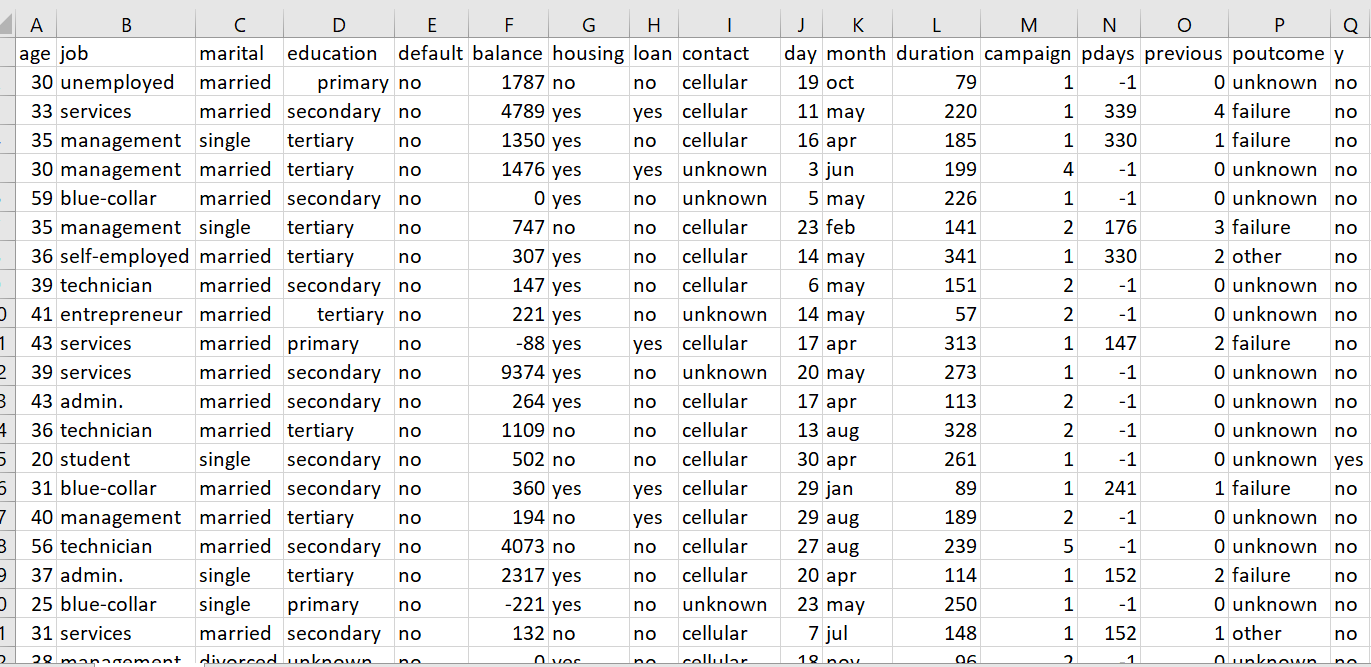
We’re given a dataset named BankTest.csv, the data is related with direct marketing campaigns of a Portugese banking institution. Various information is provided about the person based on which the conclusion is made on whether or not he/she will subscribe to the product(bank term deposit).

Target Variable: Y

This document will take us through the steps used for Data analysis using excel

# Import Data

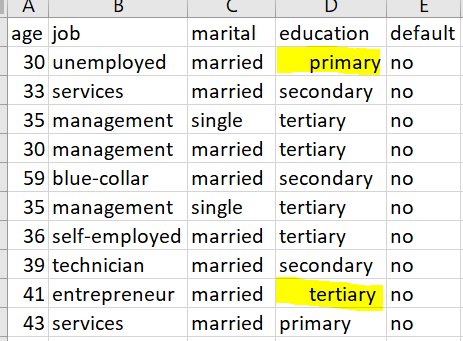
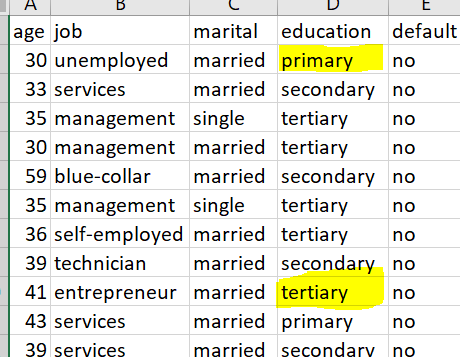
* Downloaded the BankTest.csv file opened it in excel document
* Selected all the data and used format option to autofit the data as per the width of the cell.



# Data Cleaning:

* Clear all the formatting present in the data
* Made sure that there are no blank lines present in the excel
* To ensure the alignment of the cells, if there are any leading or trailing spaces present in the cells, applied trim function.

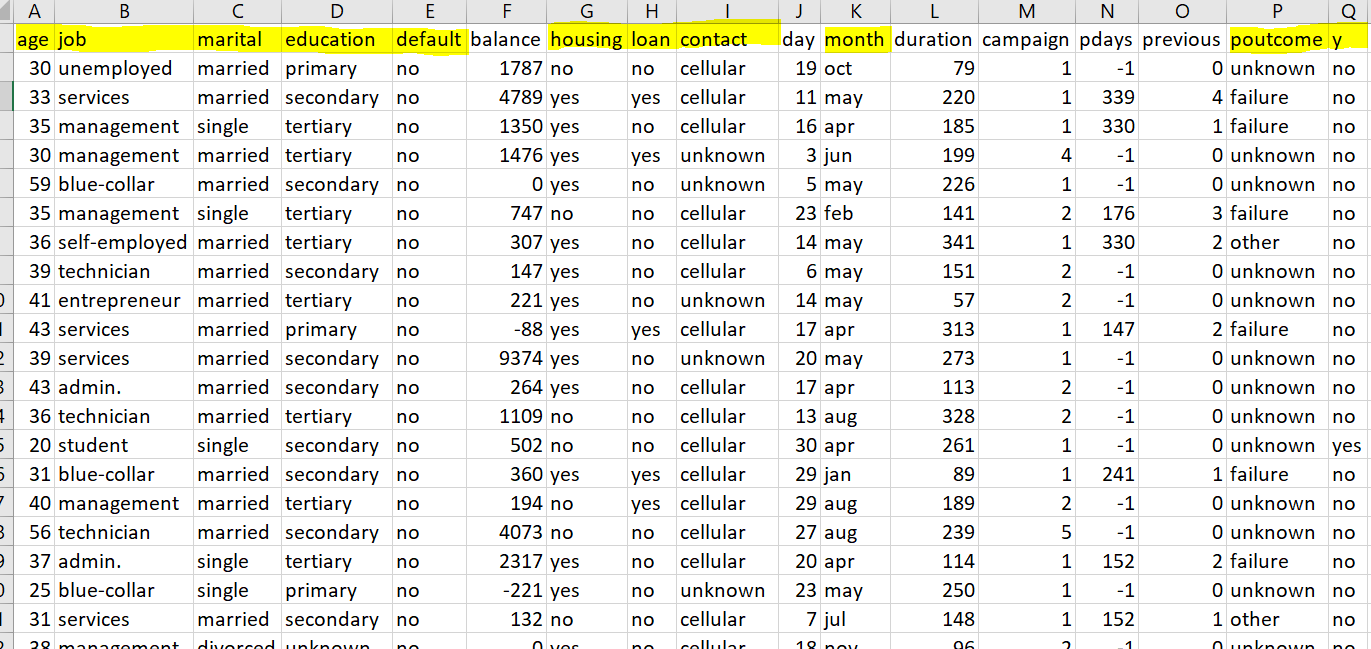
Before: After:

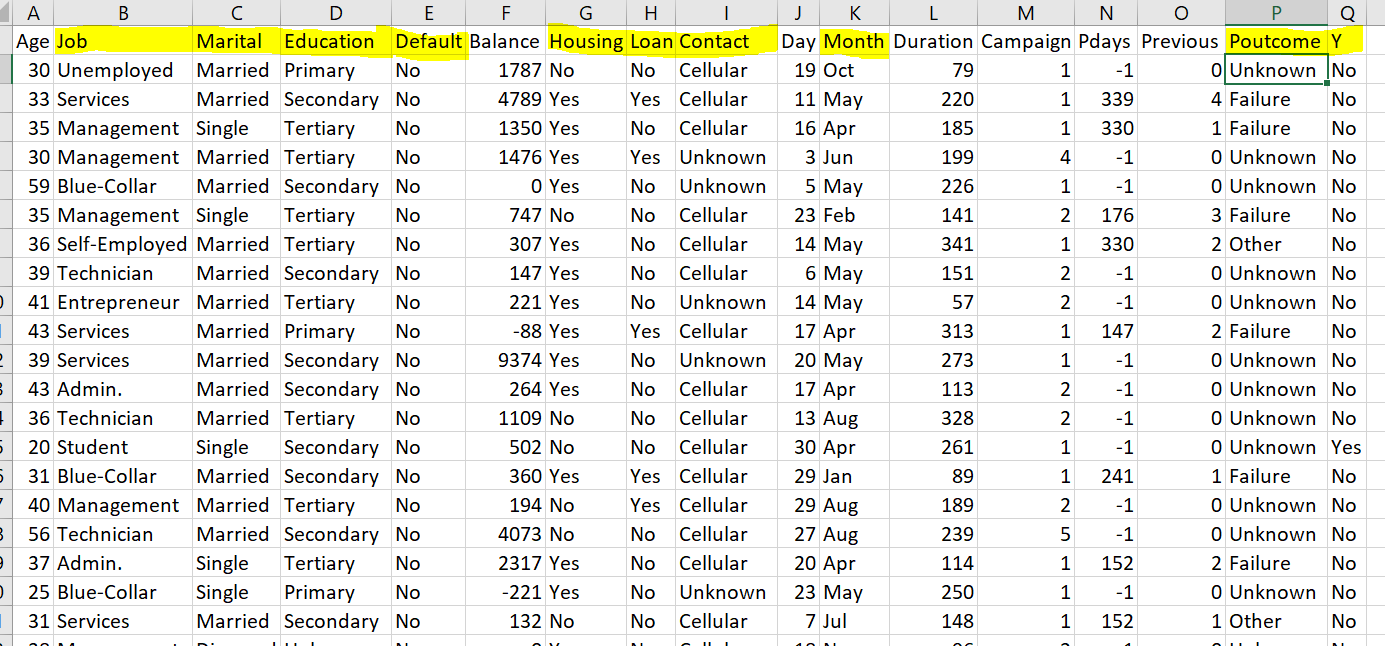
* Used proper function to capitalize the first letter of each word to increase the readability.

Columns affected: Job, Marital,Education,Default,Housing,Loan, Cellular, Poutcome

Before:



After:



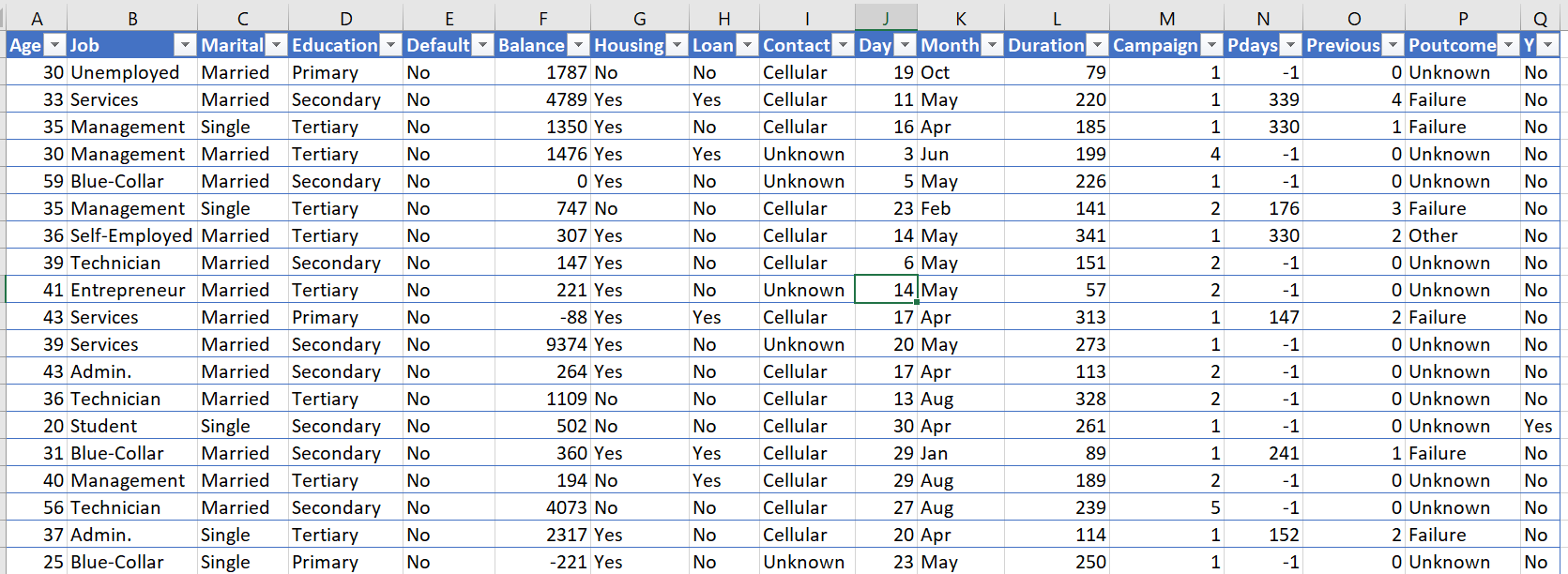
* Removed duplicate data depending on all the columns.

Count before removing redundancy: 4525

Count after removing redundancy: 4522

# Data Filtering:

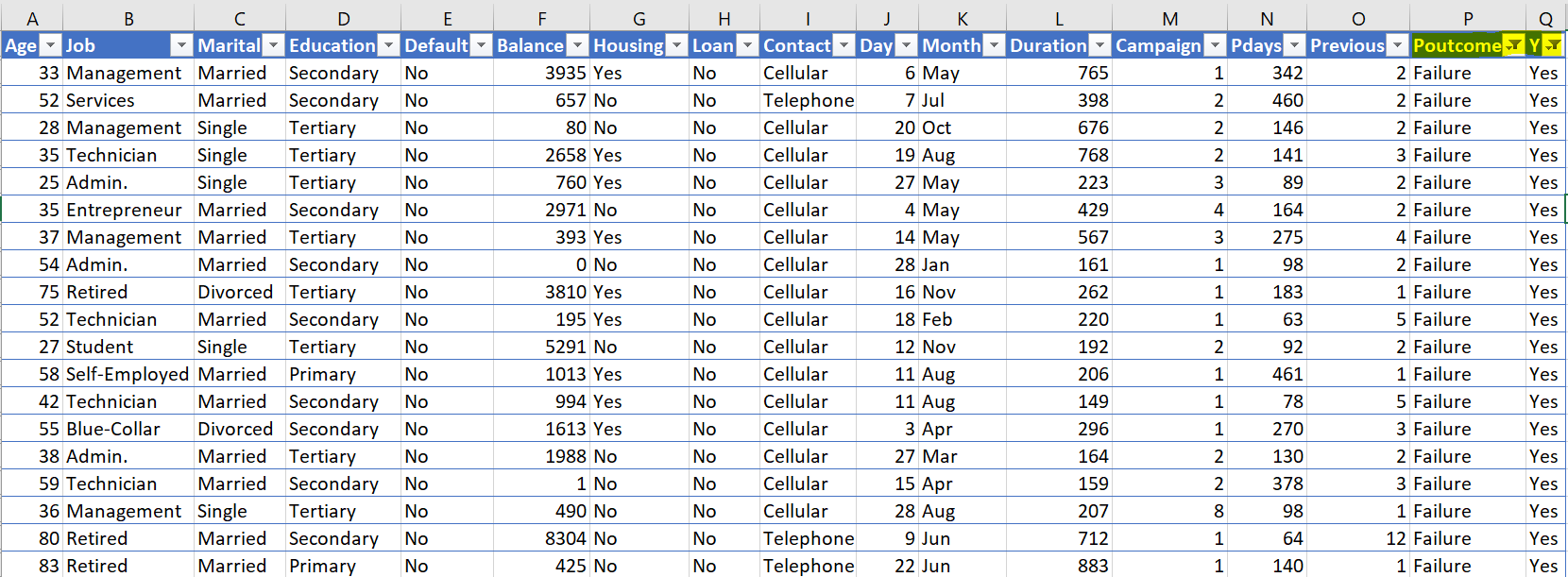
* Used excel feature, Format as table to enhance the visibility of data and applying the filters:



* We can now apply filters to infer patterns from data.
* For instance, we calculated the count(using subtotal function) of yes based on the outcome of previous marketing campaign, Poutcome

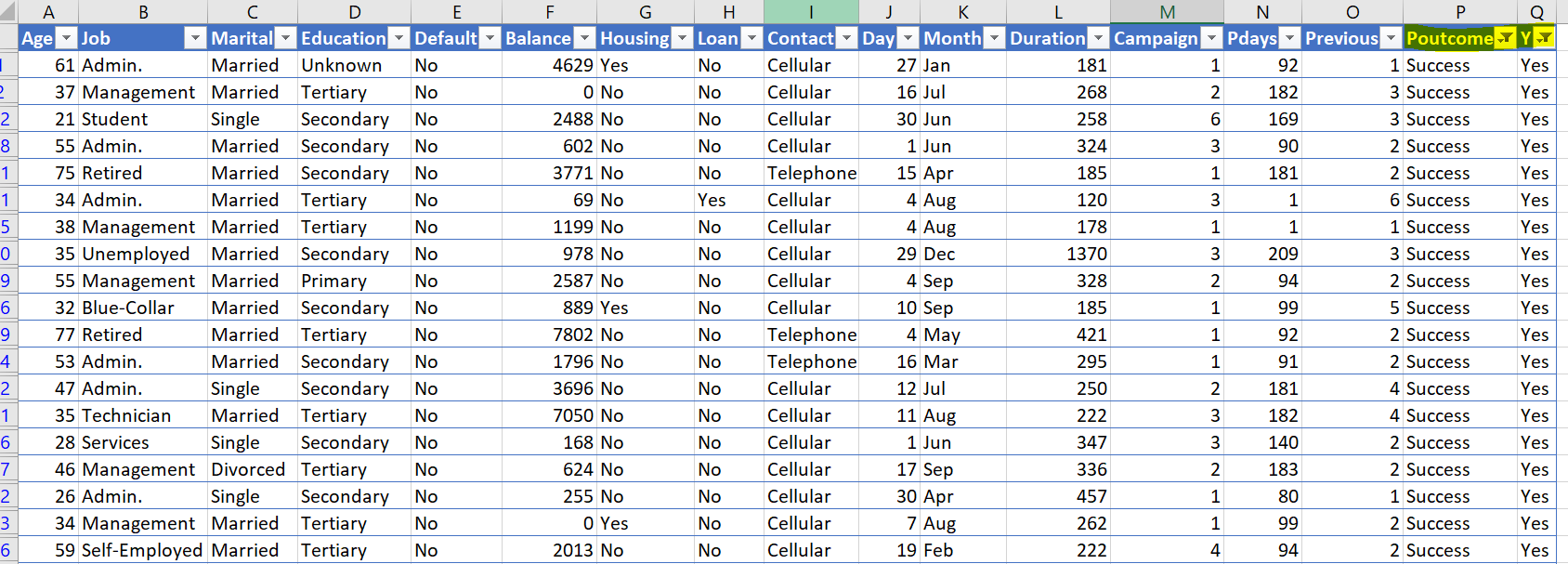
Screenshot for data filtered for Poutcome = “Failure” and Y = “Yes”:

Count = 63



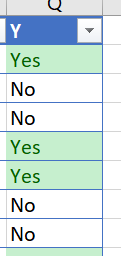
Screenshot for data filtered for Poutcome = “Success” and Y = “Yes”:

Count = 83

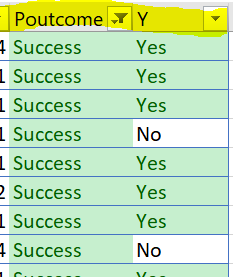


# Data Visualization and Highlighting:

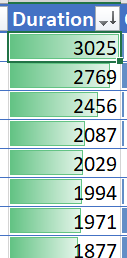
* We can highlight the Target variable having Yes as the outcome with green color:



* We can highlight the Success value in Poutcome column as well to visualize the relationship between the result of previous campaign’s outcome with current campaign’s outcome. And we can see that most of the persons that subscribed in previous campaign, subscribed for the new product as well.



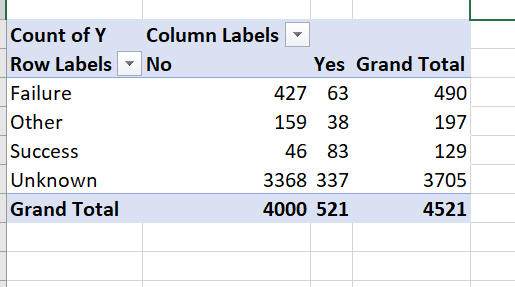
* We can also draw data bars(heat map) for the Duration column which represents the duration of the last call in second



# Pivot Tables:

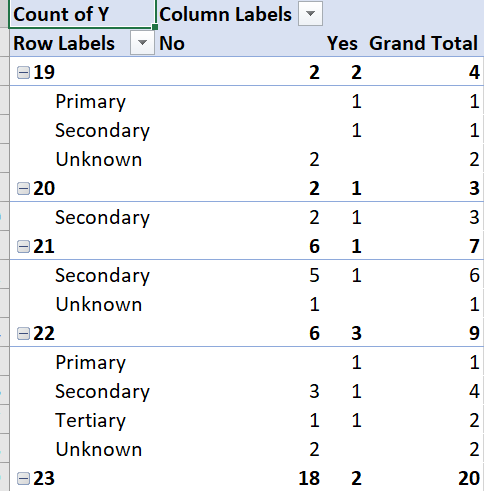
**Pivot 1:**

We took categorical data of Poutcome as row label and data of Y (target variable) as column labels and then used the summation function to show the count for corresponding row and column:



**Pivot 2:**

We took age and education of the person as row label and tried to find correlation of age and education combined with Y :

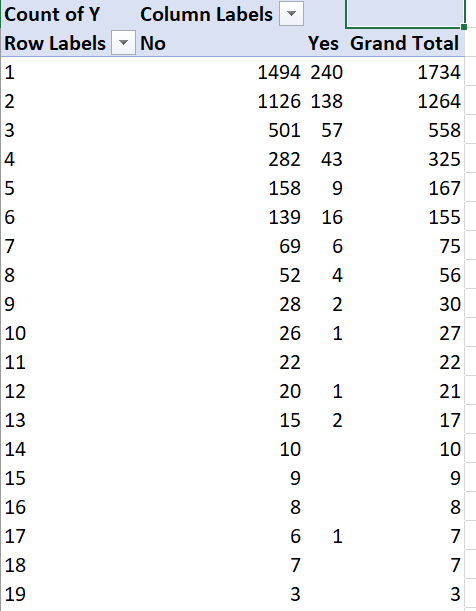


We came to know that maximum Yes is from the person having age 32 and secondary education.

**Pivot 3:**

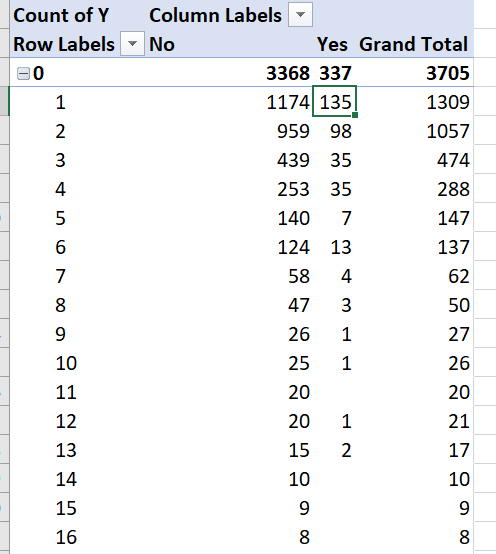
We put campaign that defines the number of times the person was contacted for this campaign as row label and we came to find out that the person that were contacted too much didn’t subscribed for the product

Therefore, our hypothesis is that Campaign has a negative correlation with Y



**Pivot 4:**

We further added Previous column as row label with Campaign and found out that the person who was contacted least number of times for this campaign and has not been contacted ever before this campaign has the most number of subscriptions.



Pivot 5:

We now plotted Housing, Loan and Marital as row labels. Here, Housing column defines whether or not the person has taken housing loan and Loan defines if the person has personal loan or not.

We concluded that the person having no housing loan or personal loan and are married has subscribed the most for the product

