**Beginners Guide to EDA-Exploratory Data Analysis on a Real Data Set using Numpy & Pandas in Python!**

When working on a new dataset in order to take intelligent action, you need to understand your data. Exploratory data analysis (EDA) allows us to develop the gist of what our data may look like and what kinds of questions can be answered by them. EDA is important because it allows the explorer to make critical decisions about what is interesting to pursue and what probably isn’t worth following up on and thus building a hypothesis using the relationships between variables. EDA helps us to gain insights and help us understand the correlation between the independent variables and target variables.

Topics Covered-

1. Understanding the Data Set.
2. Discovering Patterns Between Categorical and Numerical Columns.
3. Discovering & handling the Missing Values
4. Discovering the trends between the Features w.r.t Target Variable
5. Getting Insights and Conclusions

**Data Set**-

Here I will be using the data set of Bank Marketing Campaign to Predict if the client will subscribe to direct marketing campaign

You can download the complete dataset from-<https://drive.google.com/drive/folders/1urwTQPkUypJ6dGDJgS9Gszb83bfXEG6z?usp=sharing>

Problem-

The data is related to direct marketing campaigns of a Portuguese banking institution. Predict if the client will subscribe to a term deposit based on a marketing campaign

The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. There are four datasets:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]

2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.

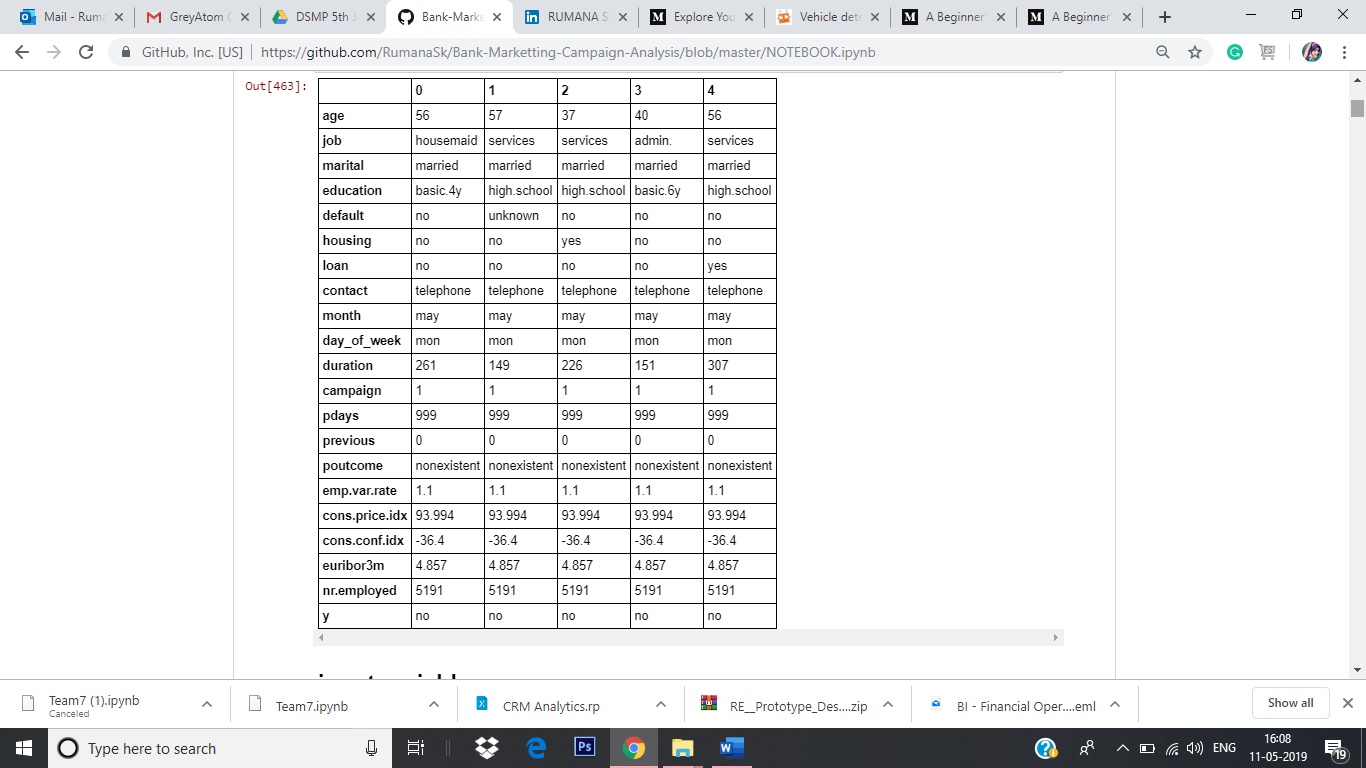
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with fewer inputs).

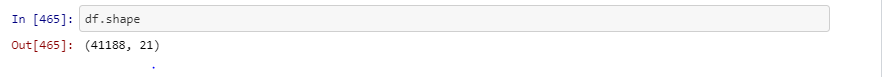
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with fewer inputs). The smallest datasets are provided to test more computationally demanding machine learning algorithms

**Goal:** The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

**Getting to Know about Data**

At the very first of Exploratory Data Analysis, we want to start understanding the data quickly, so here we use df.head(),df.describe() and df.shape().



In this data set, we have 21 columns and almost 41,188 rows.

# **Input variables:**

1. ***age*:** (numeric)  
2. ***job*** **:** type of job (categorical: 'admin.','blue collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3***. marital* :** marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4. ***education*** (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5. ***default:*** has credit in default? (categorical: 'no','yes','unknown')  
6. ***housing:*** has housing loan? (categorical: 'no','yes','unknown')  
7. ***loan:*** has personal loan? (categorical: 'no','yes','unknown')

## related with the last contact of the current campaign:

8. ***contact:*** contact communication type (categorical: 'cellular','telephone')  
9. ***month:*** last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10. ***day\_of\_week:*** last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11. ***duration:*** last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

12. ***campaign:*** number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13. ***pdays:*** number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14. ***previous:*** number of contacts performed before this campaign and for this client (numeric)  
15. ***poutcome:*** outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

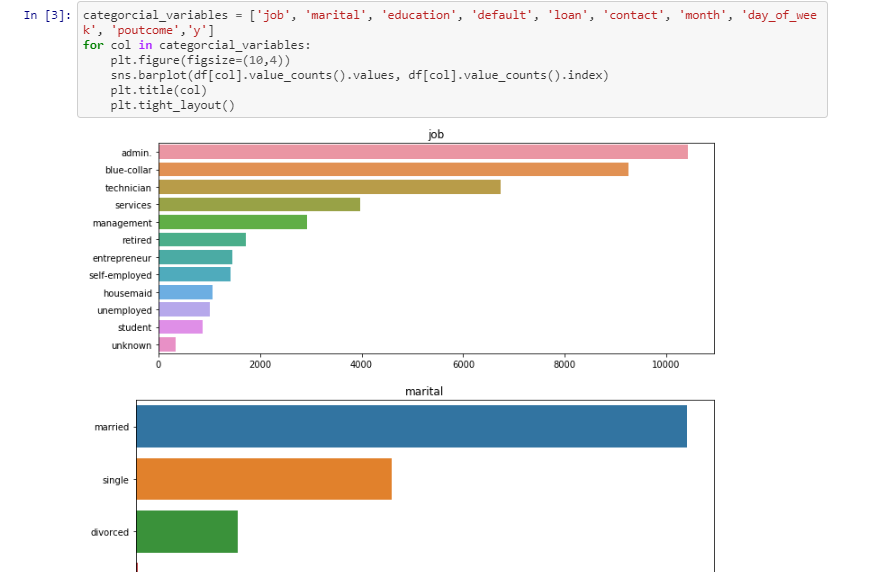
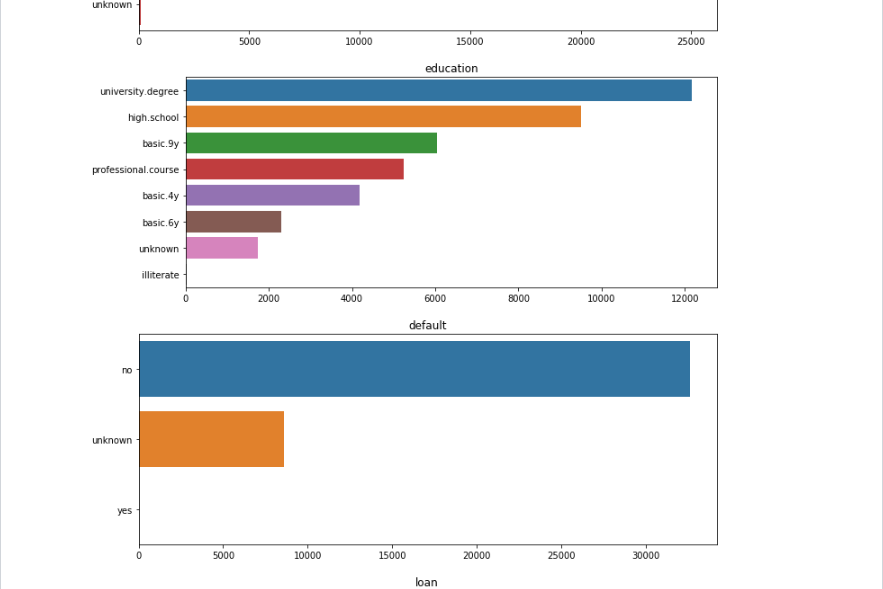
## social and economic context attributes

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):

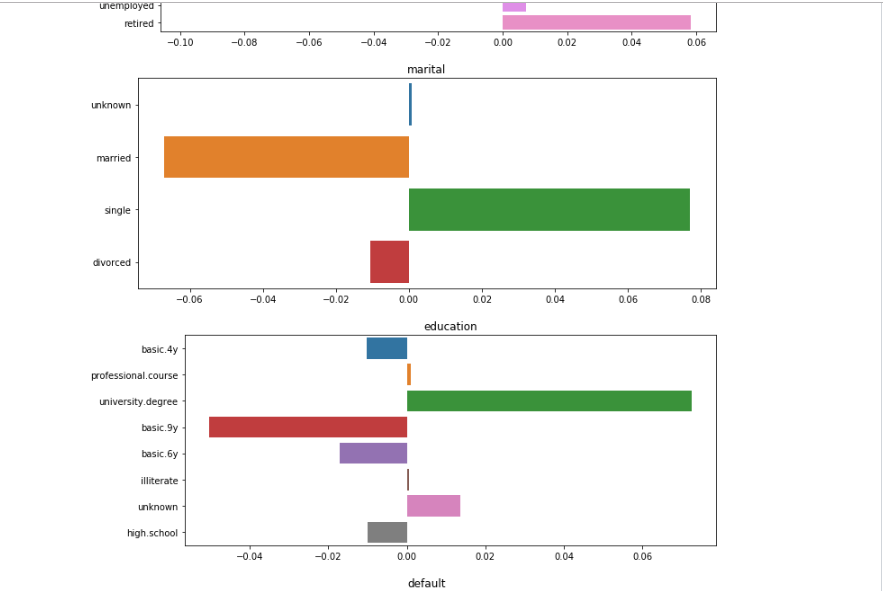
21. **y** - has the client subscribed a term deposit? (binary: 'yes','no')

## Categorical Variables:

We first start the exploratory analysis of the categorical variables and see what are the categories and are there any missing values for these categories. Here, we used the seaborn package to create the bar graphs 

## List of normalized relative frequency of the target class per category.

Normalized distribution of each class per feature and plotted difference between positive and negative frequencies. Positive values imply this category favours clients that will subscribe and negative values categories that favour not buying the product.

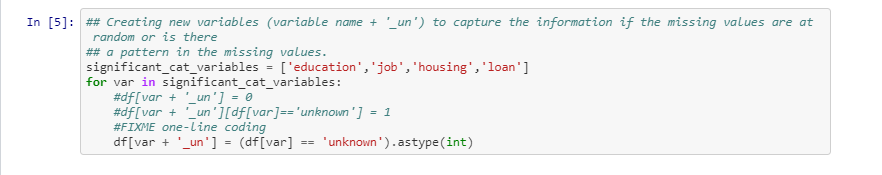
**Handling Missing Values**

There are unknown values for many variables in the Data set. There are many ways to handle missing data. One of the ways is to discard the row but that would lead to reduction of data set and hence would not serve our purpose of building an accurate and realistic prediction model.

Other method is to smartly infer the value of the unknown variable from the other variables. This a way of doing an imputation where we use other independent variables to infer the value of the missing variable. This doesn't guarantee that all missing values will be addressed but majority of them will have a reasonable which can be useful in the prediction.

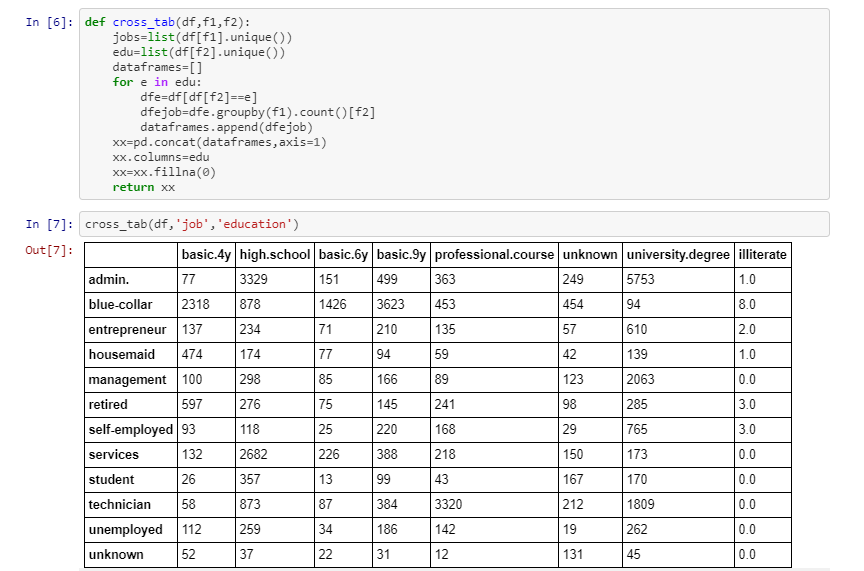
Variables with unknown/missing values are: 'education', 'job', 'housing', 'loan', 'deafult', and 'marital'. But the significant ones are 'education', 'job', 'housing', and 'loan'. The number of unknowns for 'marital' is very low. The unknown for 'default' variable are considered to be recorded as unknown. It may be possible that customer is not willing to disclose this information to the banking representative. Hence the unknown value in 'default' is actually a separate value.

Therefore, we start with creating new variables for the unknown values in 'education', 'job', 'housing' and 'loan'. We do this to see if the values are missing at random or is there a pattern in the missing values



## Imputation:

Now, to infer the missing values in 'job' and 'education', we make use of the cross-tabulation between 'job' and 'education'. Our hypothesis here is that 'job' is influenced by the 'education' of a person. Hence, we can infer 'job' based on the education of the person. Moreover, since we are just filling the missing values, we are not much concerned about the causal inference. We, therefore, can use the job to predict the education.

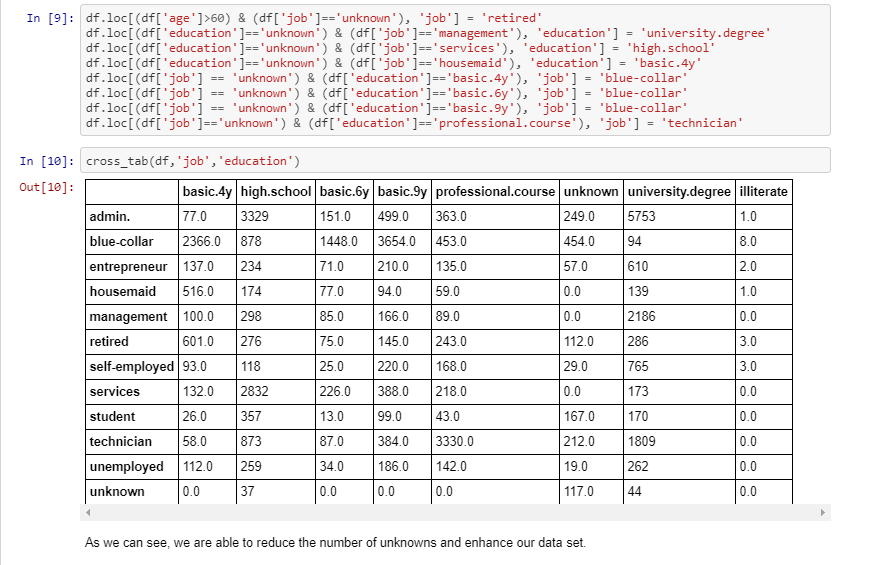


**Inferring education from jobs**: From the cross-tabulation, it can be seen that people with management jobs usually have a university degree. Hence wherever 'job' = management and 'education' = unknown, we can replace 'education' with 'university.degree'. Similarly, 'job' = 'services' --> 'education' = 'high.school' and 'job' = 'housemaid' --> 'education' = 'basic.4y'.

**Inferring jobs from education**: If 'education' = 'basic.4y' or 'basic.6y' or 'basic.9y' then the 'job' is usually 'blue-collar'. If 'education' = 'professional.course', then the 'job' = 'technician'.

**Inferring jobs from age**: As we see, if 'age' > 60, then the 'job' is 'retired,' which makes sense.

While imputing the values for job and education, we were cognizant of the fact that the correlations should make real world sense. If it didn't make real world sense, we didn't replace the missing values.



**Imputations for house and loan**: We are again using cross-tabulation between 'house' and 'job' and between 'loan' and 'job.' Our hypothesis is that housing loan status (Yes or No) should be in the proportion of each job category. Hence using the prior known distribution of the housing loan for each job category, the house loan for unknown people will be predicted such that the prior distribution (% House = Yes's and No's for each job category remains the same). Similarly, we have filled the missing values in the 'loan' variable.

## Numerical variables:

Let see the summary of the data in order to understand the numerical variables

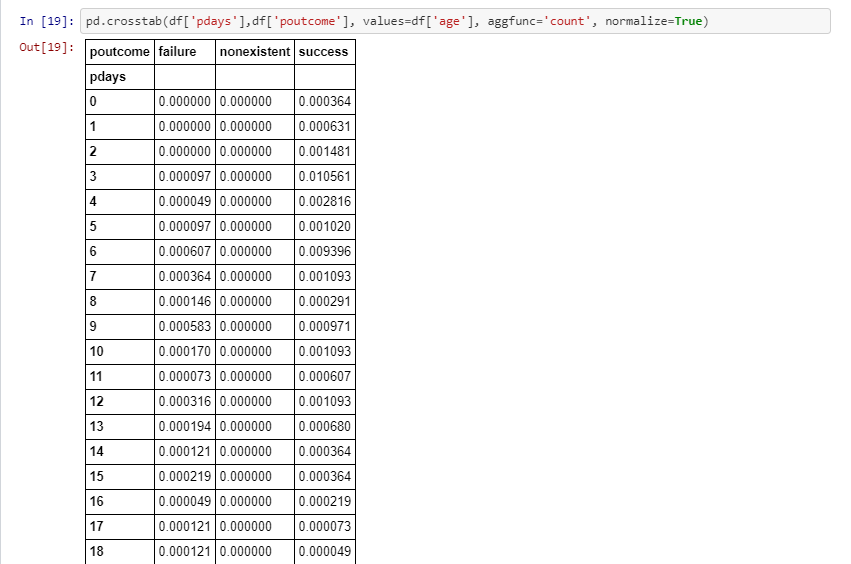
**Missing Values:**From the source of the data (U.C. Irvine ML Repository), we're told that the missing values, or NaNs, are encoded as '999'. From the above table, it is clear that only 'pdays' has missing values. Moreover, a majority of the values for 'pdays' are missing.

**Outliers:** Outliers are defined as 1.5 x Q3 value (75th percentile). From the above table, it can be seen that only 'age' and 'campaign' have outliers as max('age') and max('campaign') > 1.5Q3('age') and >1.5Q3('campaign') respectively.

But we also see that the value of these outliers are not so unrealistic (max('age')=98 and max('campaign')=56). Hence, we need not remove them since the prediction model should represent the real world. This improves the generalizability of the model and makes it robust for real world situations. The outliers, therefore, are not removed.

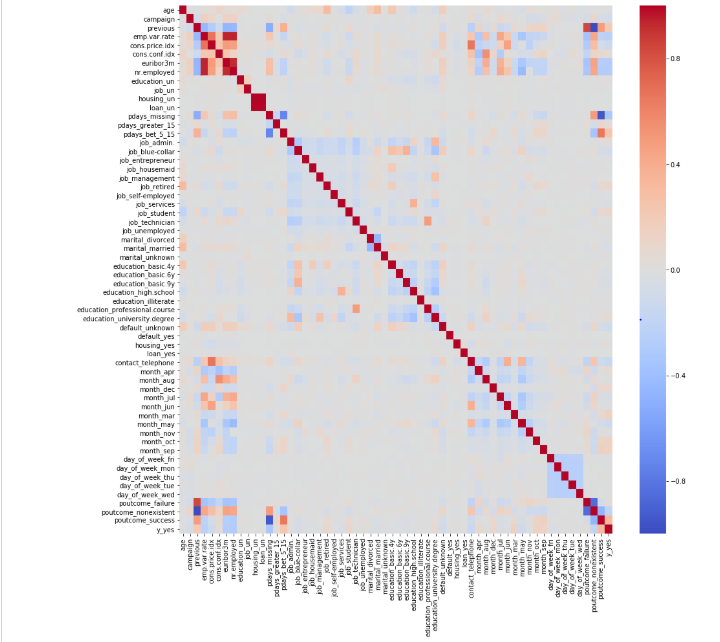
## Missing Values in Numerical Variables

Let's examine the missing values in 'pdays'



As we can see from the above table, the majority of the values for 'pdays' are missing. The majority of these missing values occur when the 'poutcome' is 'non-existent'. This means that the majority of the values in 'pdays' are missing because the customer was never contacted before. To deal with this variable, we removed the numerical variable 'pdays' and replaced it with categorical variables with following categories: p\_days\_missing, pdays\_less\_5, pdays\_bet\_5\_15, and pdays\_greater\_15.





**Inferences**: From the above heat map we can see that 'y\_yes' (our target variable) has good correlation with **'previous', 'emp.var.rate', 'euribor3m', 'nr.employed', 'pdays\_missing', 'poutcome\_success' , 'poutcome\_nonexistent' and 'pdays\_bet\_5\_15'**. We expect to see these independent variables as significant while building the models.

Thus, we can infer the following from our EDA:

1. **Feature -JOB**

Relation with Target variable-The “Blue-Collar are the ones who have said NO to FD while the ones who are “Retired” or working at ”Admin” post comparatively said YES.

1. **Feature -EDUCATION**

Relation with Target variable-While Comparing we realised that the one holding a “University Degree” are most likely to say YES than others.

1. **Feature -Month**

Relation with Target variable-In the month of April we have seen large numbers getting converted while in the May maximum turning up to say no. the only Reason we could figure out was in the month of April they had Financial Year End.

1. **Feature- Day of the Week**

Relation with Target variable- Most of us will definitely agree to the fact that no one likes to get troubled on weekends by telephonic Marketing campaigns. Hence, Observed that on Fridays and Mondays People have Negatively responded as compared to other week days

1. **Feature- Poutcome**

Relation with Target variable- We have received great Response from the people who already had an engagements with the bank previously for some other campaigns

1. **Feature- Contact**

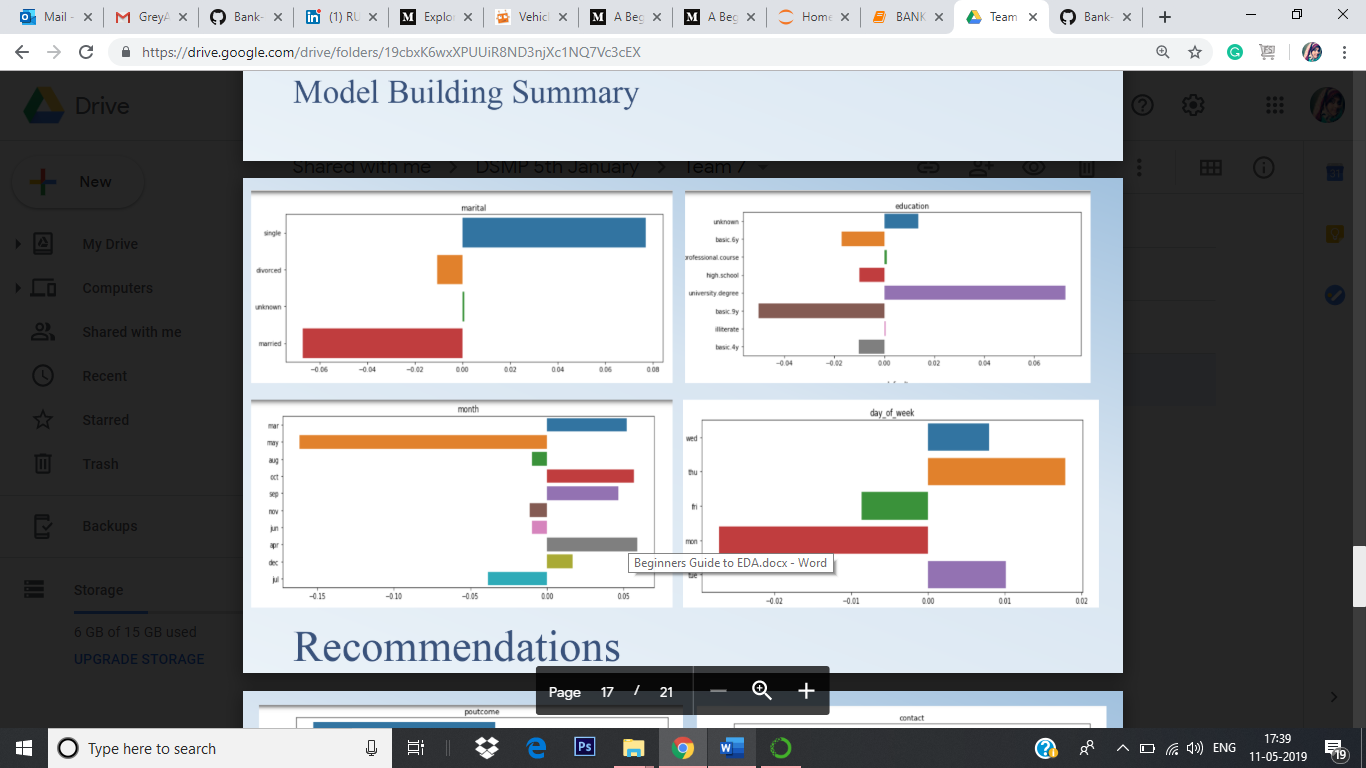
Relation with Target variable- Cellular contact has lead to maximum conversions than the telephonic point of contact.

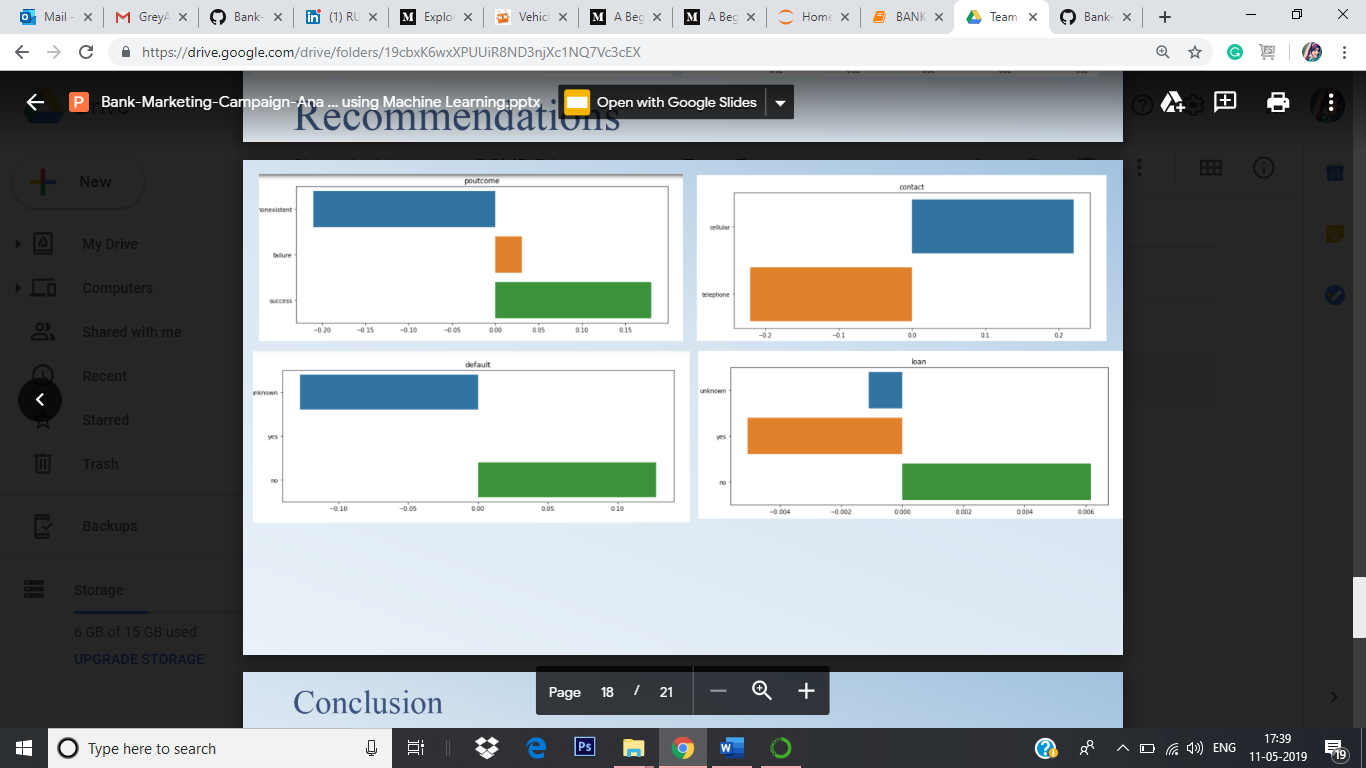
1. **Feature- Default**

Relation with Target variable- Non-Defaulters said YES and Defaulters said NO.

1. **Feature- Loan**

Relation with Target variable- People having personal Loan or Housing Loans didn’t Prefer to join for FD’s.





**Conclusion**

* + Target relatively Old Age people.
  + Prioritize those customers to who were the part of the previous campaigns.
  + According to the plot for both logistic regression and random forest, we can tell that the most influential
  + variables are duration, nr.employed, euribor3m, and emp.var.rate.
  + Based on signs of coefficients of variables in logistic regression, “duration” has positive effect on people
  + saying “yes”. This is because the longer the conversations on the phone, the higher interest the customer
  + will show to the term deposit.
  + “nr.employed”, which is the number of employees in the bank, has positive effect for turning people to
  + subscribe the term deposit. This can be due to the fact that the more employees the bank have, the more
  + influential and prestigious this bank is.
  + “euribor3m” is another important variable, which denotes the euribor 3 month rate.
  + This indicator is based on the average interbank interest rates in Eurozone. It also has positive effect since
  + the higher the interest rate the more willingly customer will spend their money on financial tools.
  + Employment variation rate (emp.var.rate) has negative influence, which means the change of the
  + employment rate will make customers less likely to subscribe a term deposit.
  + This makes sense because the employment rate is an indicator of the macroeconomy. A stable employment
  + rate denotes a stable economic environment in which people are more confident to make their investment.