**Credit Amount EDA Case Study**



Contents

[**Credit Amount EDA Case Study** 1](#_Toc98211625)

[Introduction ( 10 Minutes) 3](#_Toc98211626)

[Objectives 3](#_Toc98211627)

[Data Dictionary 3](#_Toc98211628)

[EDA - Credit Applications 4](#_Toc98211629)

[1. Data Wrangling ( 50 minutes ) 4](#_Toc98211630)

[1.1 Inspecting data 5](#_Toc98211631)

[1.2 Data Cleaning 7](#_Toc98211632)

[1.3 Imputing Missing Data 8](#_Toc98211633)

[1.4 Fixing erroneous data 11](#_Toc98211634)

[1.5 Adding new columns by Binning Continuous Variables 13](#_Toc98211635)

[1.6 Splitting data based on TARGET 14](#_Toc98211636)

[2. Univariate Analysis ( 60 Minutes ) 15](#_Toc98211637)

[Univariate Analysis of Quantitative Variables 22](#_Toc98211638)

[3. Bivariate & Multivariate Analysis ( 50 minutes ) 24](#_Toc98211639)

[4. Final Insights ( 5 minutes ) 33](#_Toc98211640)

# Introduction ( 10 Minutes)

Notes for the educator –

* The Jupyter notebook is exhaustive with comments, description, topics and introductions. Please use the same to brief students about each section accordingly.
* Explain each code line by line. There are comments for each code line to provide a basic idea about how the code is being executed.
* Data Dictionary – Spend considerable amount of time (5-10 mins) in going through the data dictionary either in Excel sheet or Jupyter notebook. Cover all the fields used in Notebook at least and give students a good understanding of those variables.
* Follow timelines mentioned on each topic. There can be 10-20% deviation as per the skillset level of class but try to adhere to mentioned timelines.
* For all the charts in Univariate/Bivariate analysis, there should be 3 types of discussions in the class –
  + Code/ Function explanation line by line
  + Insights discussion
  + Any possibilities of modifications in current chart for additional insights ( Optional)

We have loan applications data for about 307k applicants. The goal of this case is to perform Risk Analytics with the help of data wrangling and visualisation libraries of Python. The end goal is to derive important insights for the bank to identify the characteristics for bad loan applications. ( Bad loans are loans which are delayed/not paid.)

## Objectives

* Identify what are some common characteristics of bad loan applications
* Identify if there are any patterns related to applicants with loan difficulties
* Identify the driving factors or strong indicators of a bad loan application

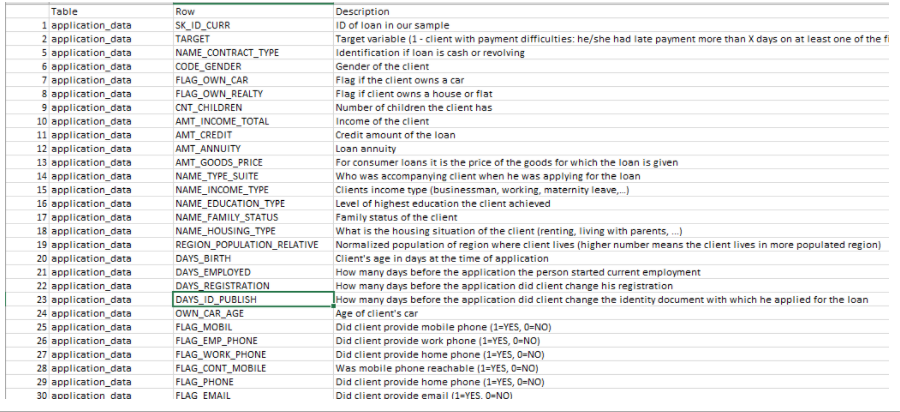
## Data Dictionary

A common starting point in any EDA problem is - Understanding the data.

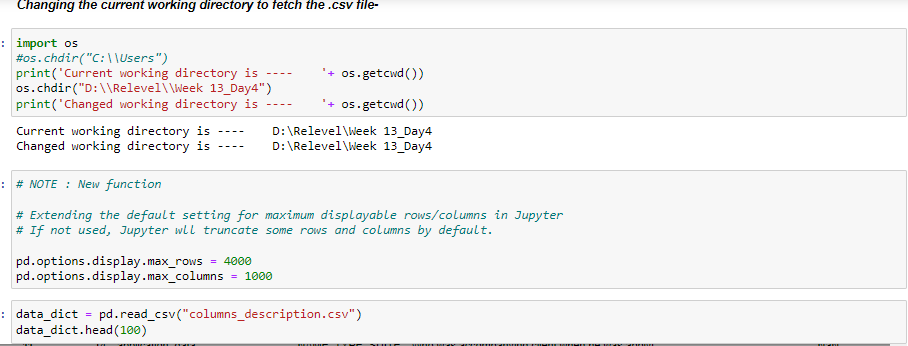
The first step is to check if there is a data dictionary available, and try to get a good understanding of the level of the data and meaning of each of the columns.

The data dictionary document has been provided along with the data. It is advised to go through each of the column in data once before starting with EDA.

A screenshot of the same is given below -



The dictionary can also be imported to Jupyter notebook as below –



# EDA - Credit Applications

Let's begin our EDA now. The flow of the entire case would be as follows -

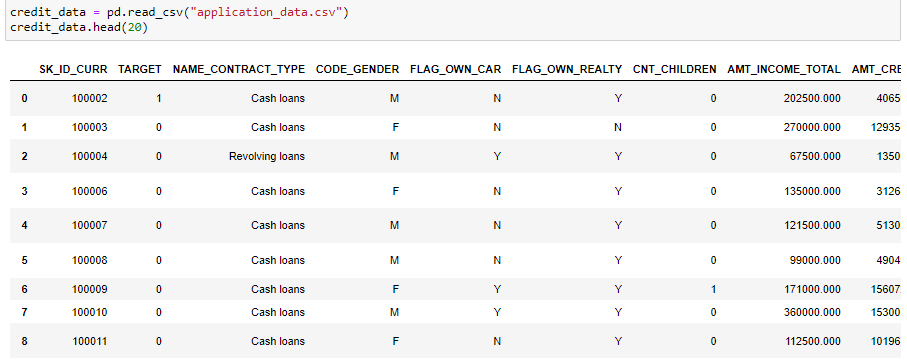
1. **Data Wrangling** – The process of transforming data from a raw form to a more adaptable form for analytics/ML.
2. **Univariate Analysis** – Analysis involving one variable
3. **Bivariate/Multivariate Analysis** – Analysis involving two or more than two multiple variables
4. **Final Insights –** Conclusive inferences of the problem at hand

**Importing libraries**



## 1. Data Wrangling ( 50 minutes )

Loading the data -



We see that the data is at Loan ID level ( SK\_ID\_CURR). There is a mix of quantitative and qualitative variables. There are a lot of Flags as well ( such as *FLAG\_OWN\_CAR* , *FLAG\_OWN\_REALTY* etc. ). There are considerable NAs as well at the first glance of data.

( NAs are missing observations in the data )

### Inspecting data

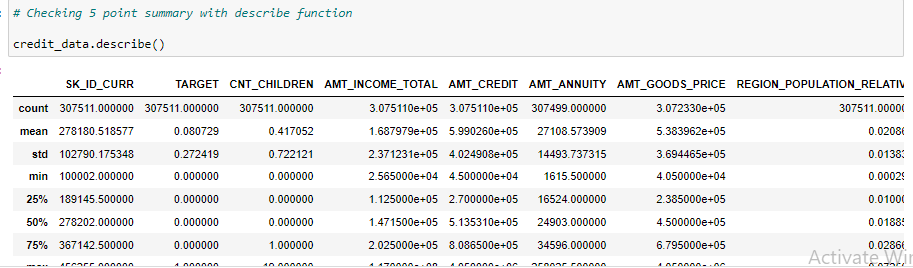


Now we know that there are 307511 loan applications and 122 fields for each application.

##### *5 Point summary*

describe function is used to get the 5 number statistical summary of the quantitative variables of a data. The focus points could be the Range, mean and median for each variable to get a better understanding of the variables.

( Do a quick search about 5 Point summary in case not aware )



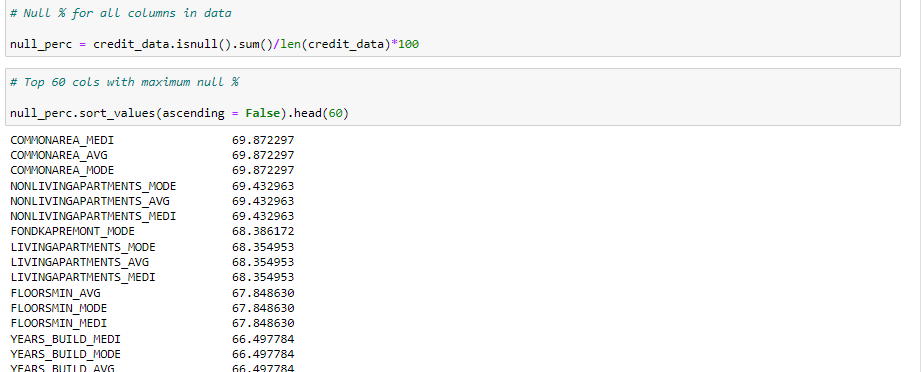
Notes for the educator –

* Discuss 5 point summary of at least 5 important fields. Talk abut how 5 point summary related to a box plot and histogram. Later when histograms are plotted, connect back to this discussion.

##### *Invalid -ve values*

Another important use case of describe function is to check for Invalid -ve values. Here for example, a close look will tell us that all the columns starting with "DAYS\_..." (example - DAYS\_BIRTH ) have -ve values which cannot be valid. We will clean the data by transforming this data appropriately later.

##### *Getting the list of % Nulls in each column*



Notes for the educator –

* Explain how below code work logically – Nulls are summed up, and divided by total row counts to give %
* Similarly, provide logical code explanation of all the codes in notebook as those are being discussed

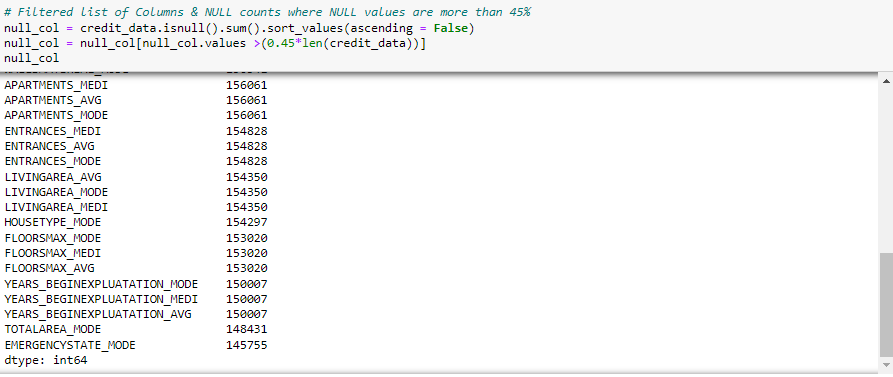
Columns with a lot of NULLs are not useful for us as they would only capture data about a select applications.

There is no standard rule for a good/bad % NULLs for columns to be used or discarded. It should be purely dependent on use case and application of the EDA.

In our case, in order to keep this exercise simpler, we will discard all columns having more than 45% NULLs.

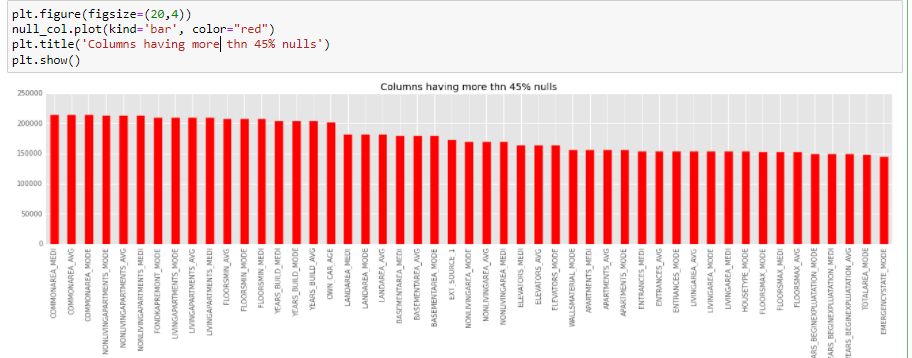
### Data Cleaning

Identifying and removing columns with more than 45% nulls

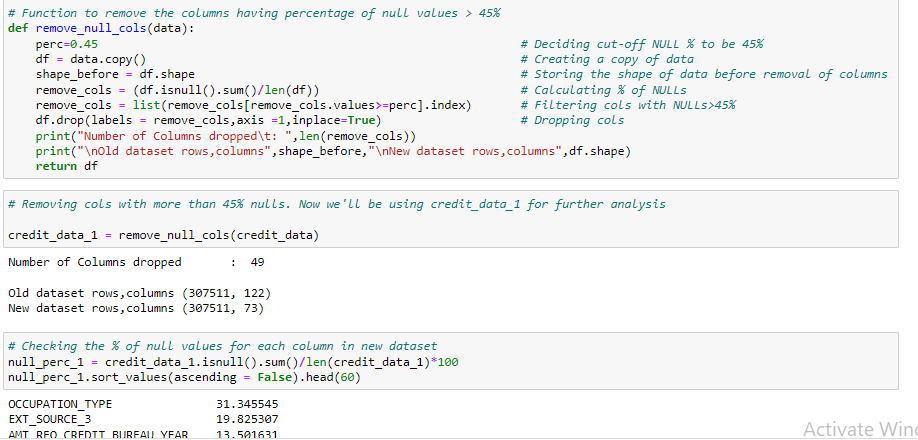




Let's visually look at the columns with NULLs>45% and there NULL counts –



##### *Removal of columns with NULLs>45%*



We have now verified that our modified dataframe "Credit\_data\_1" has no cols with more than 31% NULLs

### Imputing Missing Data

**Why imputation is needed –**

The process of replacing missing fields in the data with a substitute value is known as imputation. It is done for multiple reasons, some of which are listed below –

* To retain most of the data, as truncating rows with missing values leads to reduction of information
* Some Python libraries and functions are not compatible with missing values
* NAs may create a bias in analysis that can lead to faulty analysis
* NAs can affect the accuracy of an ML model

The below listed columns can be categorized into a group of columns with similar significance as they all represent number of queries made to the Credit Bureau.

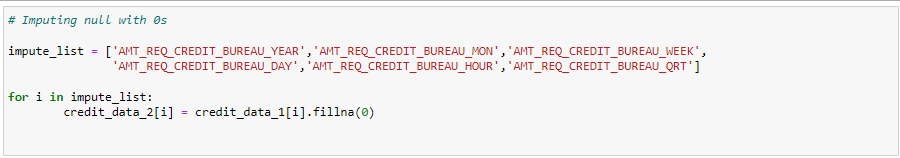
Upon further investigation, we'll see that all have mode as 0. We can impute NULLs for all of these with value 0.

In the end it is also verified that there are 0 NULLs after imputation.

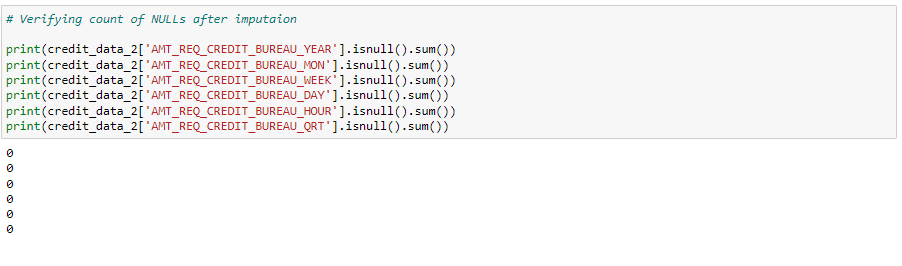
##### 



##### *Imputing NULLs with 0s*



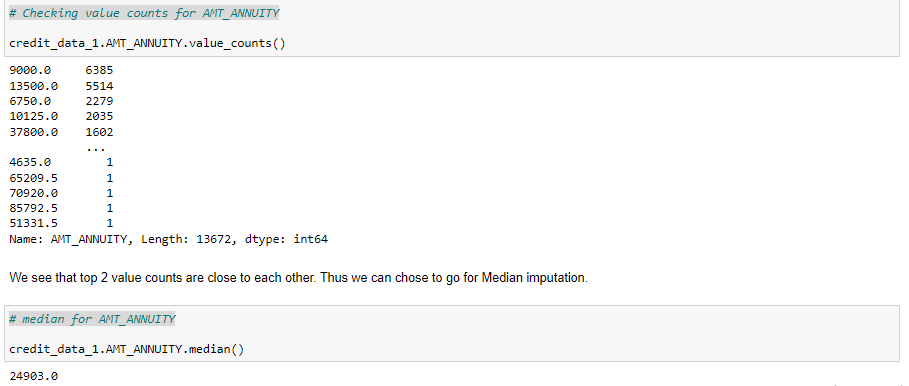
##### *Verifying count of NULLs after imputation*



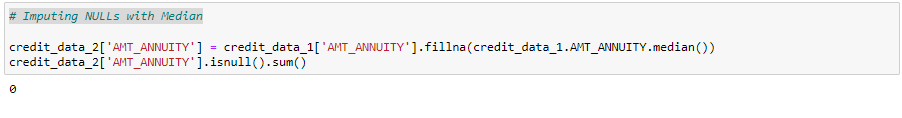
#### AMT\_ANNUITY

Since AMT\_ANNUITY is a continuous variable, unlike AMT\_REQ\_CREDIT\_BUREAU\_YEAR etc ( which could take only integer values), it is better to impute this with the median value.

Another reason for choosing to go for Median instead of Mode is close value counts for top 2 values as we'll see below.

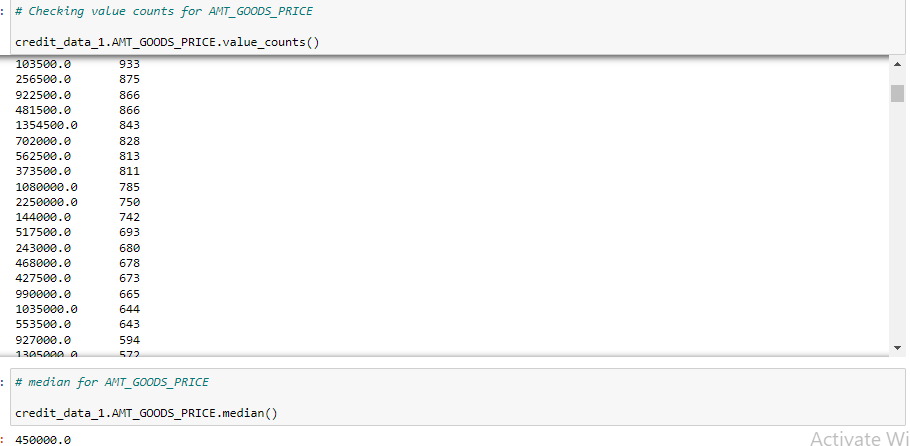


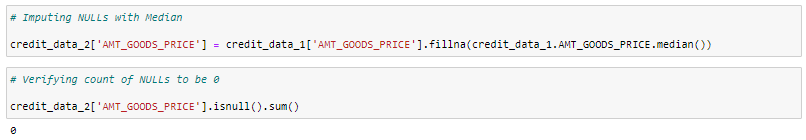
##### *Imputing NULLs with Median*



#### AMT\_GOODS\_PRICE

Similar to AMT\_ANNUITY, imputing NULLs with Median for AMT\_GOODS\_PRICE for similar reasons.

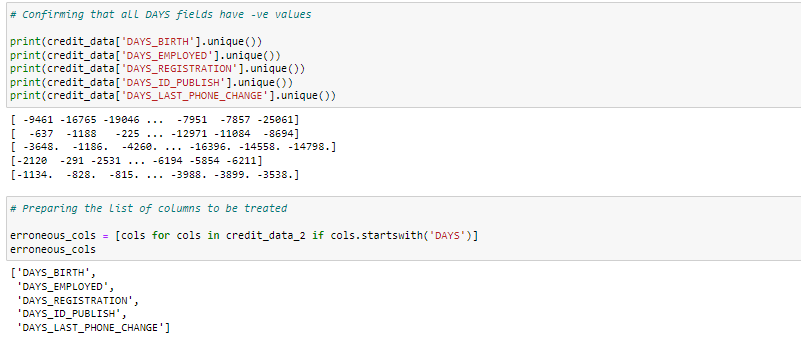


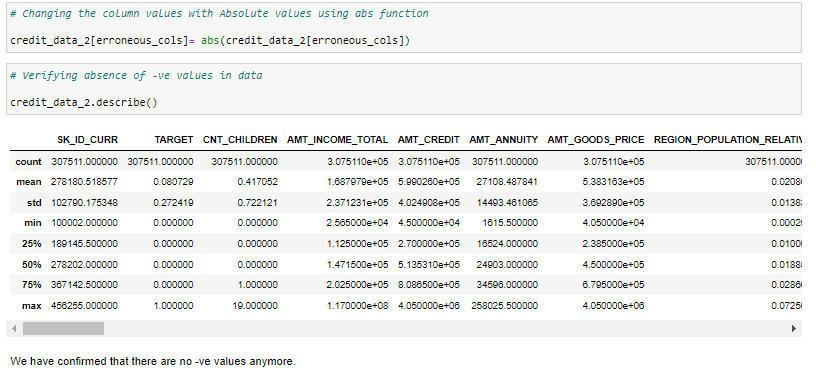


### Fixing erroneous data

As seen already with the help of describe function, we know that we need to treat -ve values in days columns.

We can modify the values to be absolute values, assuming that -ve sign was a technical fault during data feed.





We have confirmed that there are no -ve values anymore.

#### Replacing XNAs for CODE\_GENDER

A quick look on the value counts for CODE\_GENDER shows 4 counts of XNA which is equivalent of a NULL.

Since 4 is a relatively small count, it doesn't matter much on how we impute it. Imputing it with Mode F in our case

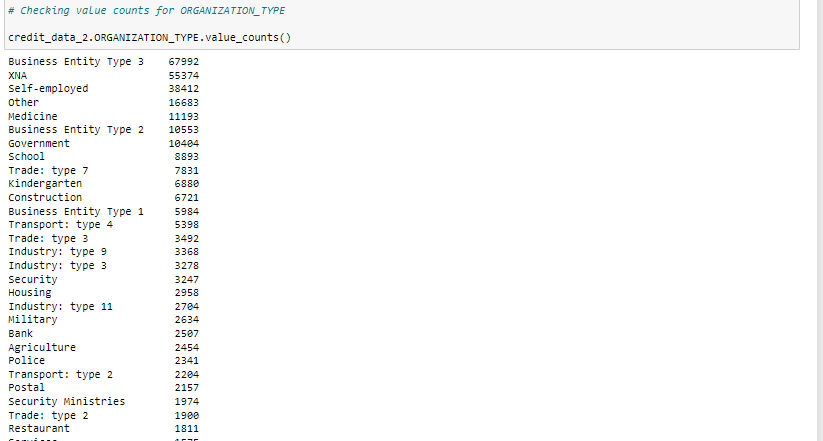


#### Replacing XNAs for ORGANIZATION\_TYPE

XNAs for ORGANIZATION\_TYPE have 2nd highest count in the data. We must be very careful in imputing such a high number of XNAs with any value.

Since it is a categorical variable, and there won't be any aggregrate functions performed on this data, we don't necessarily need whole of the value to be imputed.

Thus, changing all XNAs with NULLs to protect the originality of data.





We have confirmed that there are no more XNAs now in this field.

### Adding new columns by Binning Continuous Variables

It is always a good practice to identify core or highly significant continuous fields in the data and then bin them into specific categories. It allows for an additional categorical analysis for such fields. We'll observe the use case of same later in this EDA exercise.

For now, let's bin some of the continuous variables into 5 bins each as below -

##### *Binning AMT\_INCOME\_TOTAL, AMT\_CREDIT & DAYS\_BIRTH*

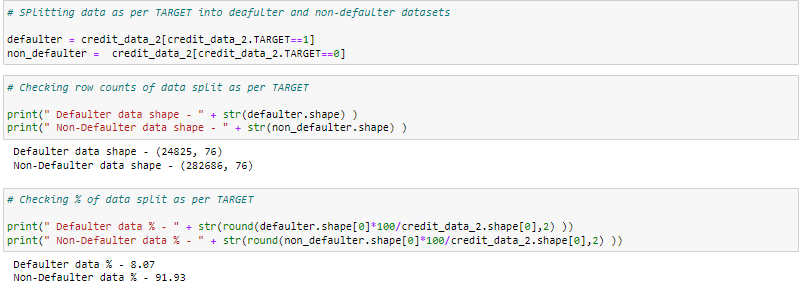




### Splitting data based on TARGET

Splitting data into 2 subsets based on Target Variable- Defaulter Data and Non-Defaulter Data.

This will help us with the comparison among 2 groups later

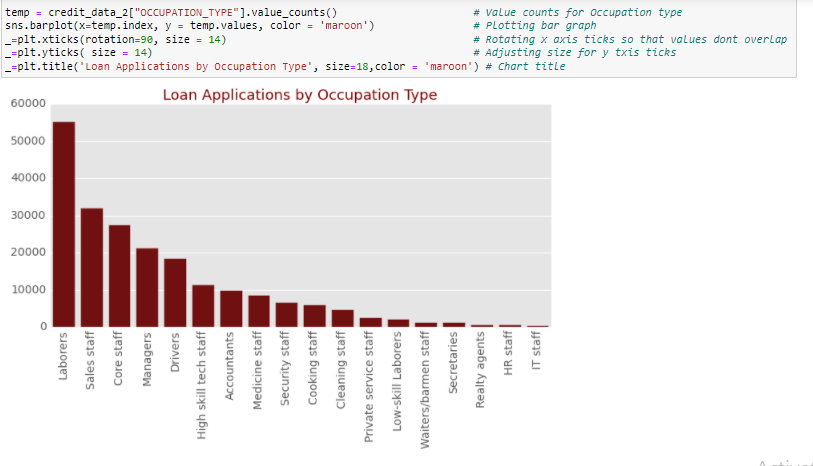


## 2. Univariate Analysis ( 60 Minutes )

Univariate Analysis is simplest form of analysing data. It restricts the analysis to only 1 variable as the name states. ( Uni means One )

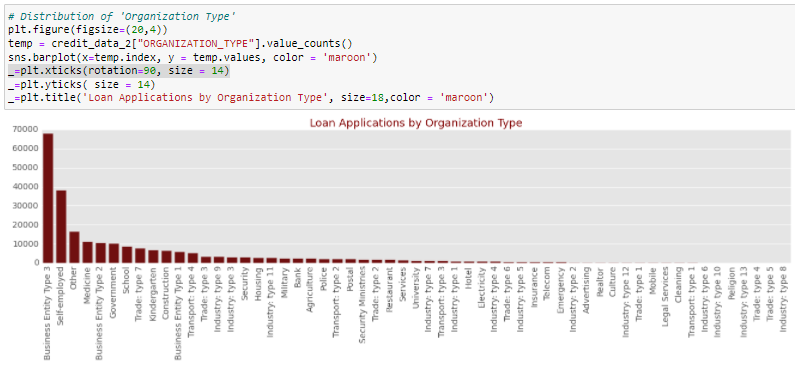
It doesn't take into account the mutual relationships and associations among variables. Rather it focuses on finding patterns through a particular field.

##### *Occupation Type*



We can infer that most of the applications come for Labourers, Sales Staff and Core Staff.

##### *Organization Type*

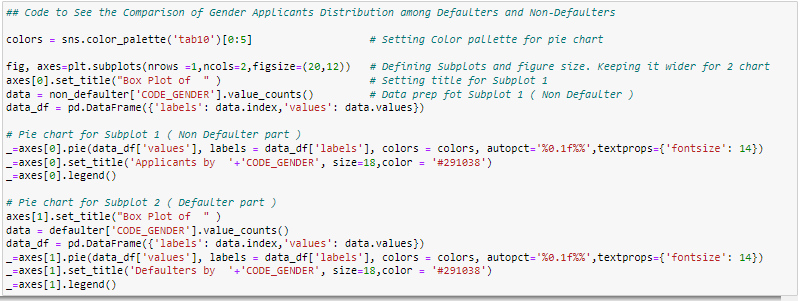


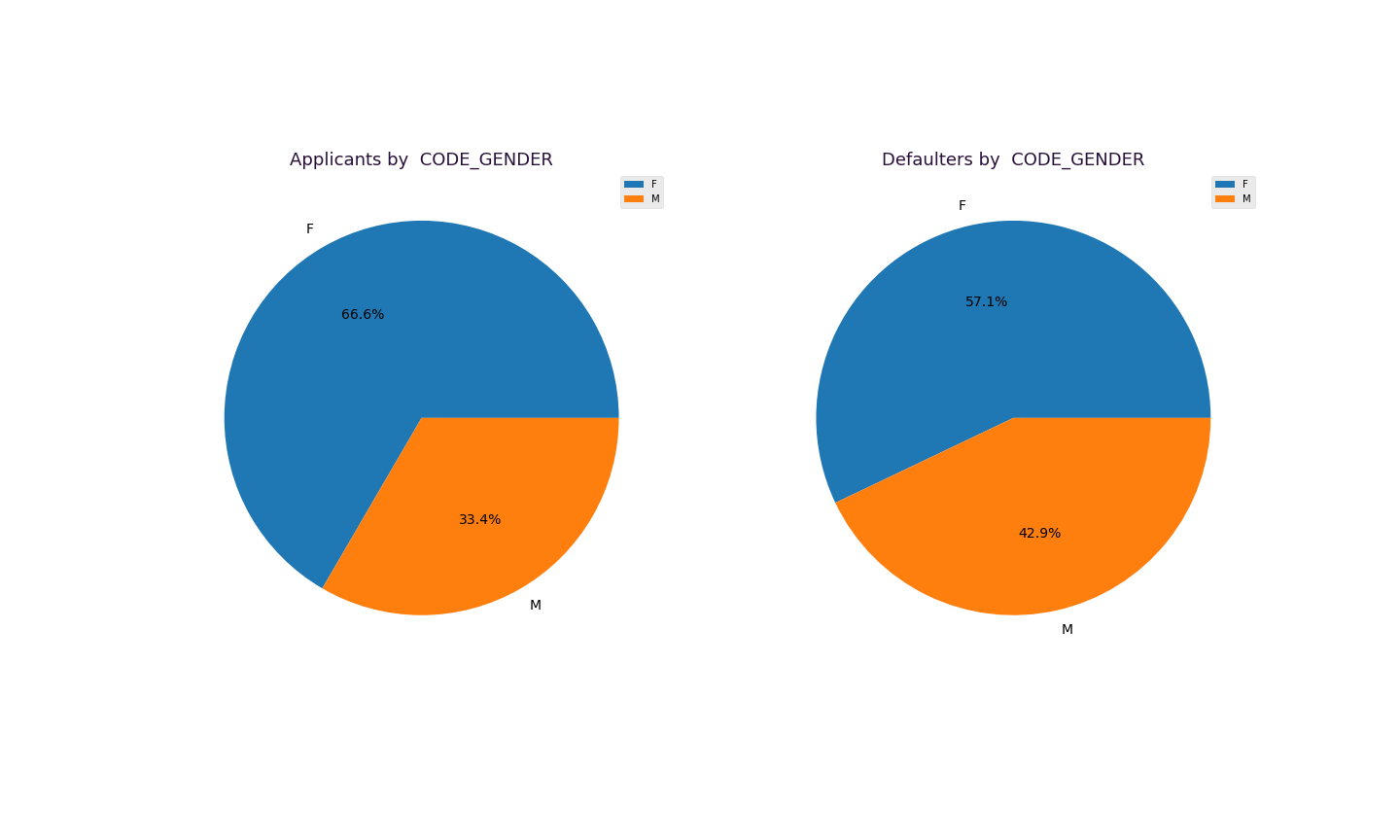
It is observed that majority of the applicants belong to Business Entity Type 3 an Self Employed.

##### *Comparison of Gender Distribution among Defaulters and Non Defaulters*

In this section, we are going to plot pie charts for various categorical variables of interest for Defaulters and Non-defaulters. The purpose is to check for any drastic shift in % distribution for 2 cases. Such comparison may provide us useful insights, as to be done next –

First, we are going to analysis how is the Gender split in 2 cases. Below is the code for plotting the pie charts -





Insights -

* There is majority of Male loan applicants
* More Men default loans as compared to Women, since the % split has increased further for Men in case of Defaulter distribution.

Often we would require to re-use the same code for multiple combination of variables.

It is a common practice to prepare charts by calling functions instead of re-writing the code again and again.

There are following benefits to this approach -

* Code Modularity is improved
* Less code is required to perform same amount of task
* Notebook looks more cleaner

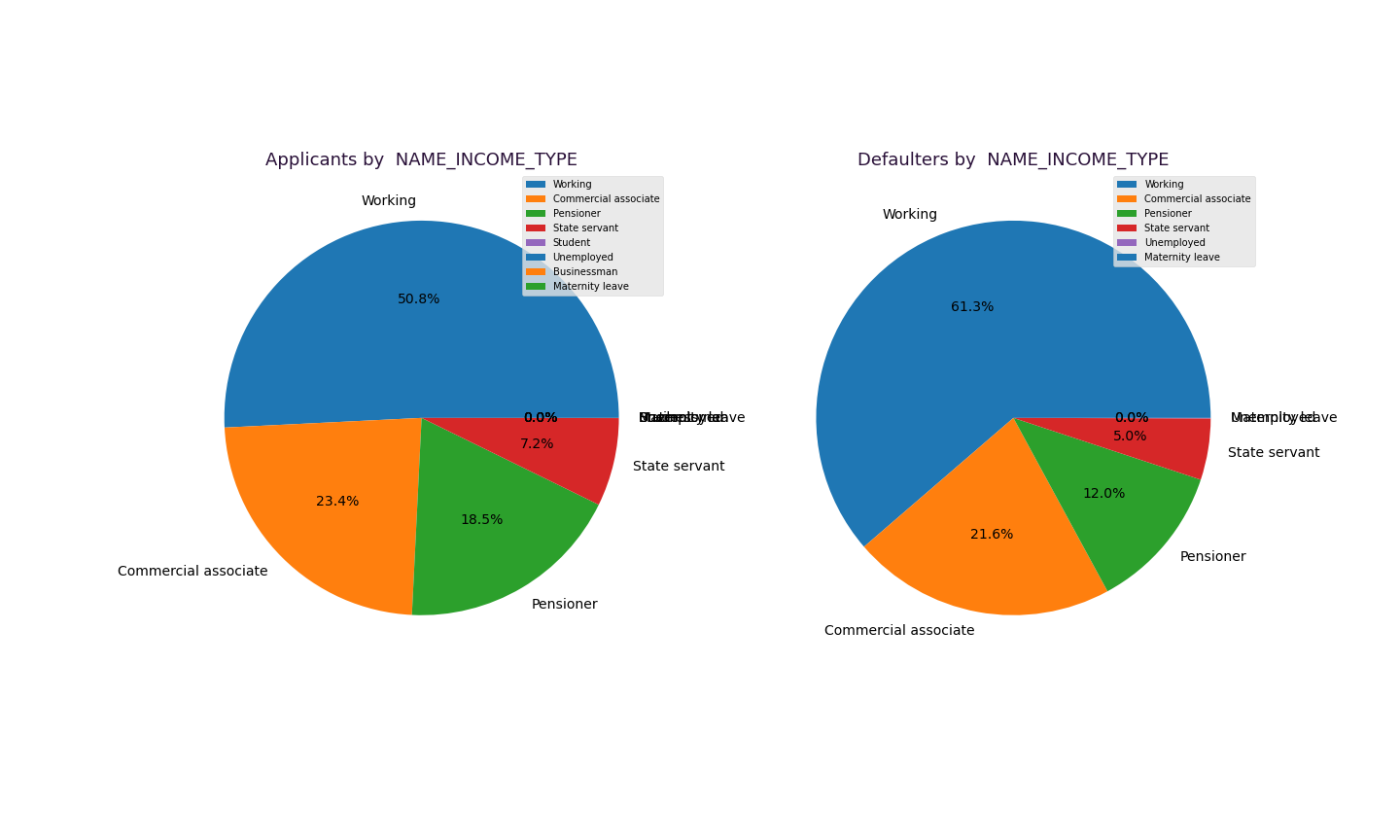
##### *Converting above code to a function -*



##### *Comparison of Income Type Distribution among Defaulters and Non Defaulters*

The next chart will show split of Income Types for Defaulters and Non Defaulters. The purpose is to check if Income type has an effect on loan defaults -



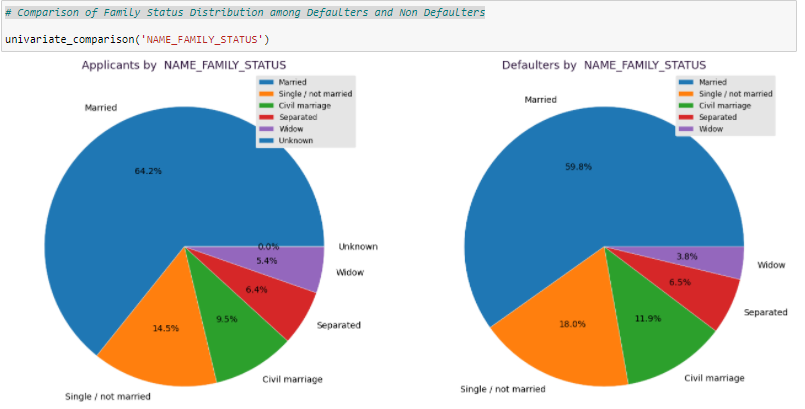


Insights -

* Almost half of the Loan applications come from Working professionals.
* Working professionals contribute more than expected to loan defaults. The % split has increased from 51% to 61%

##### *Comparison of Family Status Distribution among Defaulters and Non Defaulters*

The next chart will show split of Family Status for Defaulters and Non Defaulters. The purpose is to check if type of Family of a loan applicant is a factor in loan defaults -

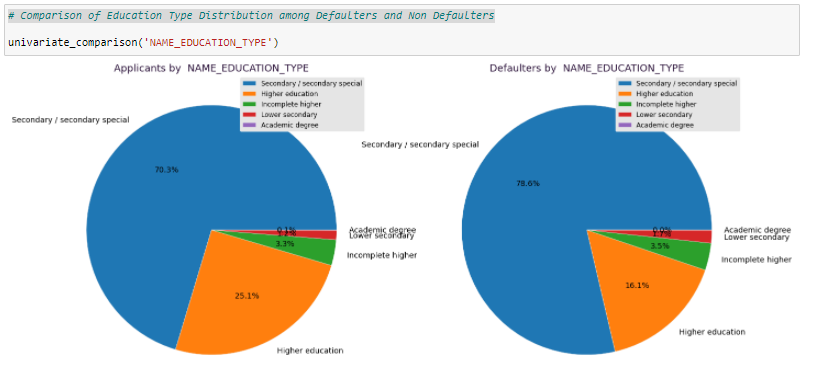


Insights-

* 65 % of the Loan applicants are married.
* Family status doesn't seem to have any major impact on Loan deafults.

##### *Comparison of Education Type Distribution among Defaulters and Non Defaulters*

The next chart will show split of Education Type for Defaulters and Non Defaulters. It will test out the hypothesis that Education has an effect on loan defaults -

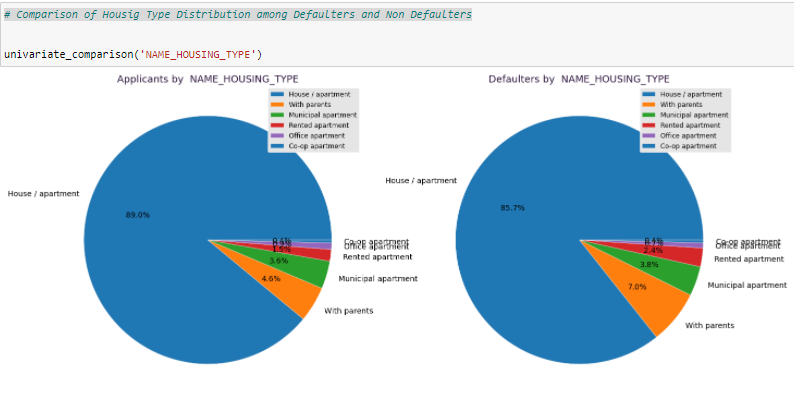


Insights-

* More than 2/3rds of Loan applicants have highest education as Secondary.
* Secondary Education class contribute majorly ( more than expected too) for loan defaults.
* There is a considerable decrease in % split for loan defaults by people with higher education. ( from 25% to 16%)

##### *Comparison of Housing Type Distribution among Defaulters and Non Defaulters*

The next chart will show split of Housing Types for Defaulters and Non Defaulters-

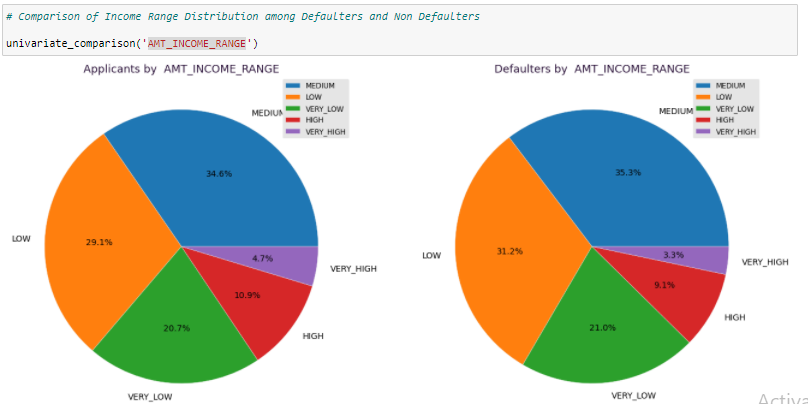


Insights-

* Almost 90% of Loan applicants have their own home.
* Housing type doesn't play a significant role in determining whether there will be a loan defaulter.

##### *Comparison of Income Range Distribution among Defaulters and Non Defaulters*

The next chart will show split of Income Range for Defaulters and Non Defaulters. The purpose is to check if low income range applicants default more than the rest -



Insights-

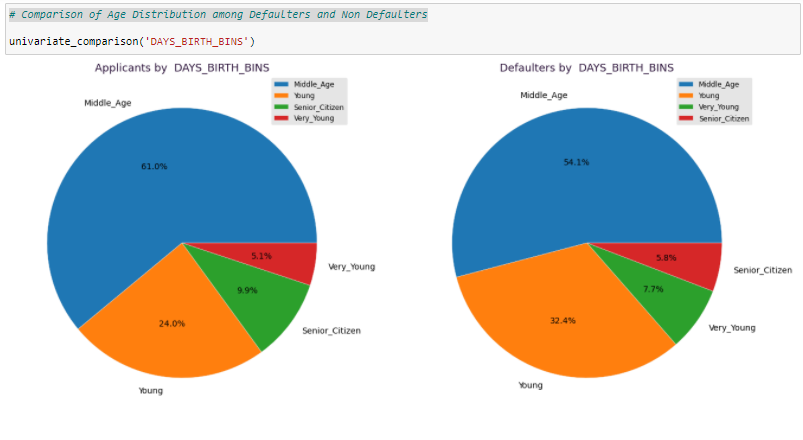
* Here also, the % split is more or less unchanged for Defaulters. It suggests that Income doesn't play a significant role in loan defaults. Although, further drilldown analysis ( later done in this notebook ) would tell us a different story.

It is always good practice to verify our hypotheses by multiple checks and not jump onto conclusions quickly.

NOTE : Let's recall that AMT\_INCOME\_RANGE is a derived variable created by binning earlier. This how binning can be useful in EDA, while this is just one use case, it has many other applications in ML as well.

##### *Comparison of Age Distribution among Defaulters and Non Defaulters*

The next chart will show split of Age for Defaulters and Non Defaulters. The purpose is to check if Age of an applicant has an effect on defaults -

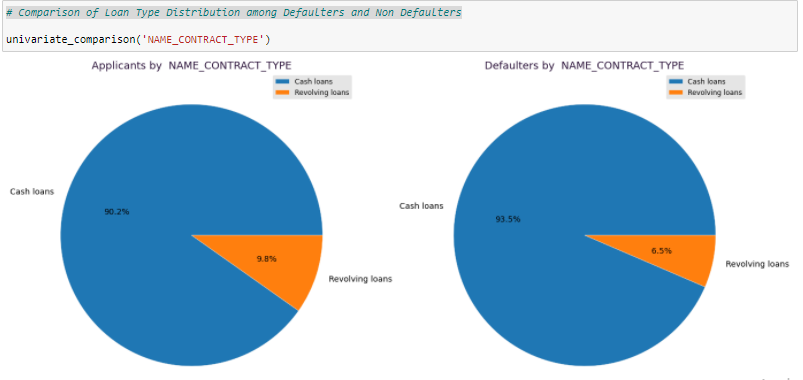


Insights -

* There is a significant shift in % split for Middle Age and Young applicants.
* Middle Aged applicants are contributing lesser to loan defaults
* Young applicants are more expected to default on a loan since there is a change in % split from 24% to 32%

##### *Comparison of Loan Type Distribution among Defaulters and Non Defaulters*

The next chart will show split of Loan Type for Defaulters and Non Defaulters. -

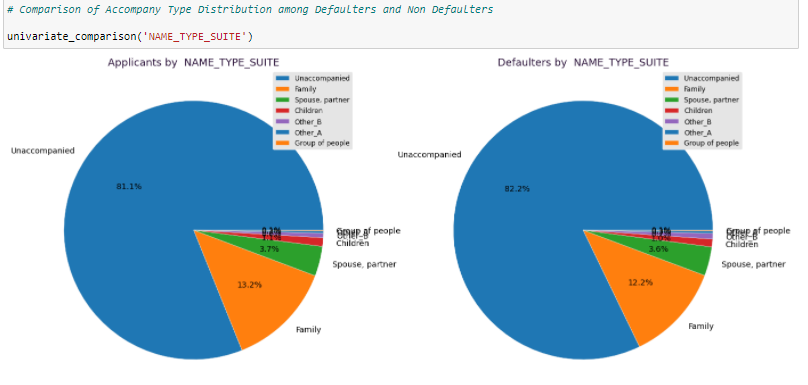


Insights-

* Cash loans are slightly more likely to be defaulted than revolving loans.

##### *Comparison of Accompany Type Distribution among Defaulters and Non Defaulters*

The next chart will show the accompany type for Defaulters and Non Defaulters. The purpose is to see if the number of persons living along can have an impact on loan default -



Insights-

* Majority of loans are applied by single occupants
* This parameter doesn't have any impact on loan defaults as the % split is unchanged in both cases.

*Alternate Chart Ideas –*

The above section of split % analysis can also be represented as following –

* Bar Charts
* Categorical Bar Charts of 2 categories of bars ( Defaulter & Non Defaulters)
* Donut Chart

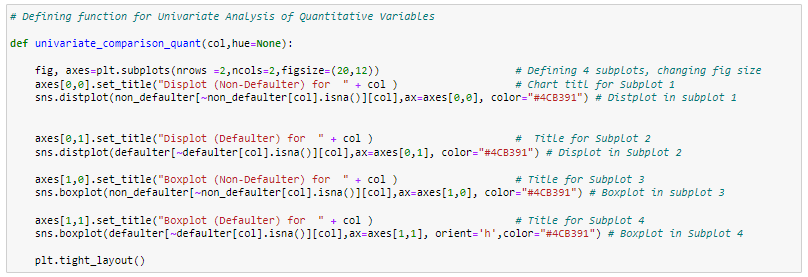
### Univariate Analysis of Quantitative Variables

The Univariate analysis of quantitative variables is fundamentally different from qualitative variables. Since these variables will have more values, and numeric ones, it involves different types of charts which are generally more statistical in nature than basic charts used for categorical variables.

For this part of analysis, we’ll plot Density plots and Box plots of various quantitative variables for Defaulters and Non-defaulters. This will allow us to compare following in both cases –

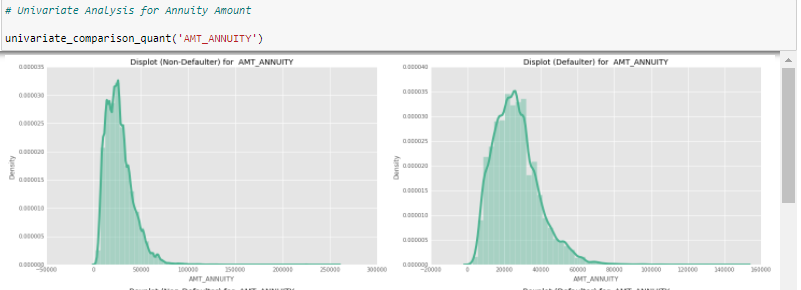
* Comparison of quartile boundaries and Means
* Comparison of distribution
* Outliers information

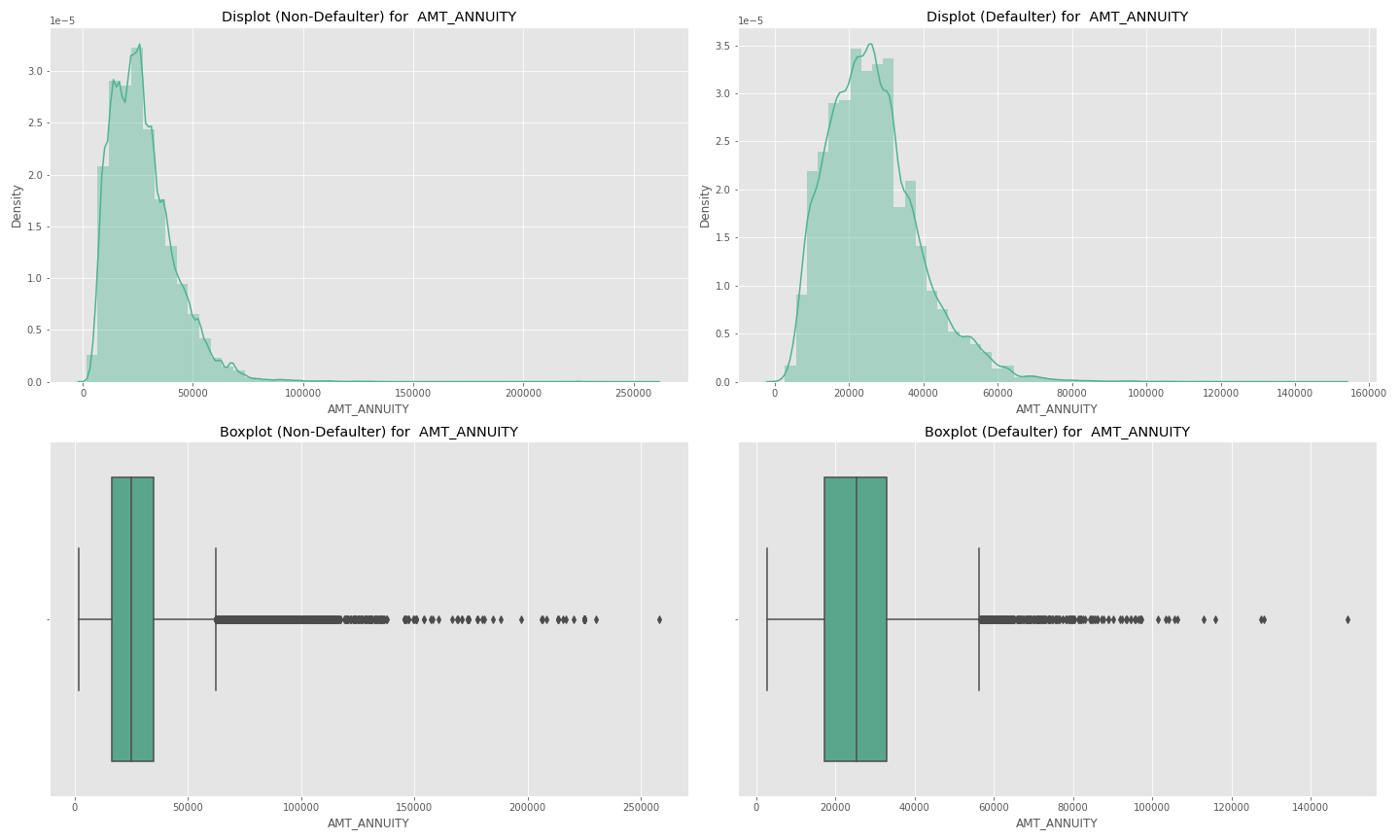
Using this, we can derive meaningful insights, as shown in below section -



*Annuity Amount*

Let’s plot the Density chart and Boxplots for Annuity amount across Defaulters and Non Defaulters. It will help us analysis the impact of Annuity amount on Loan defaults -



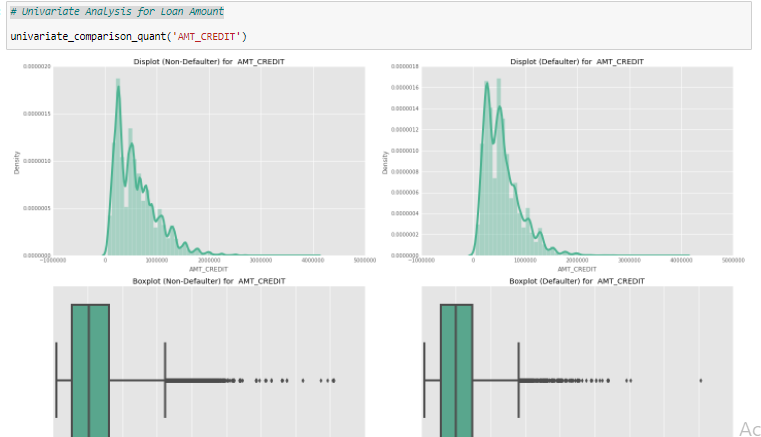


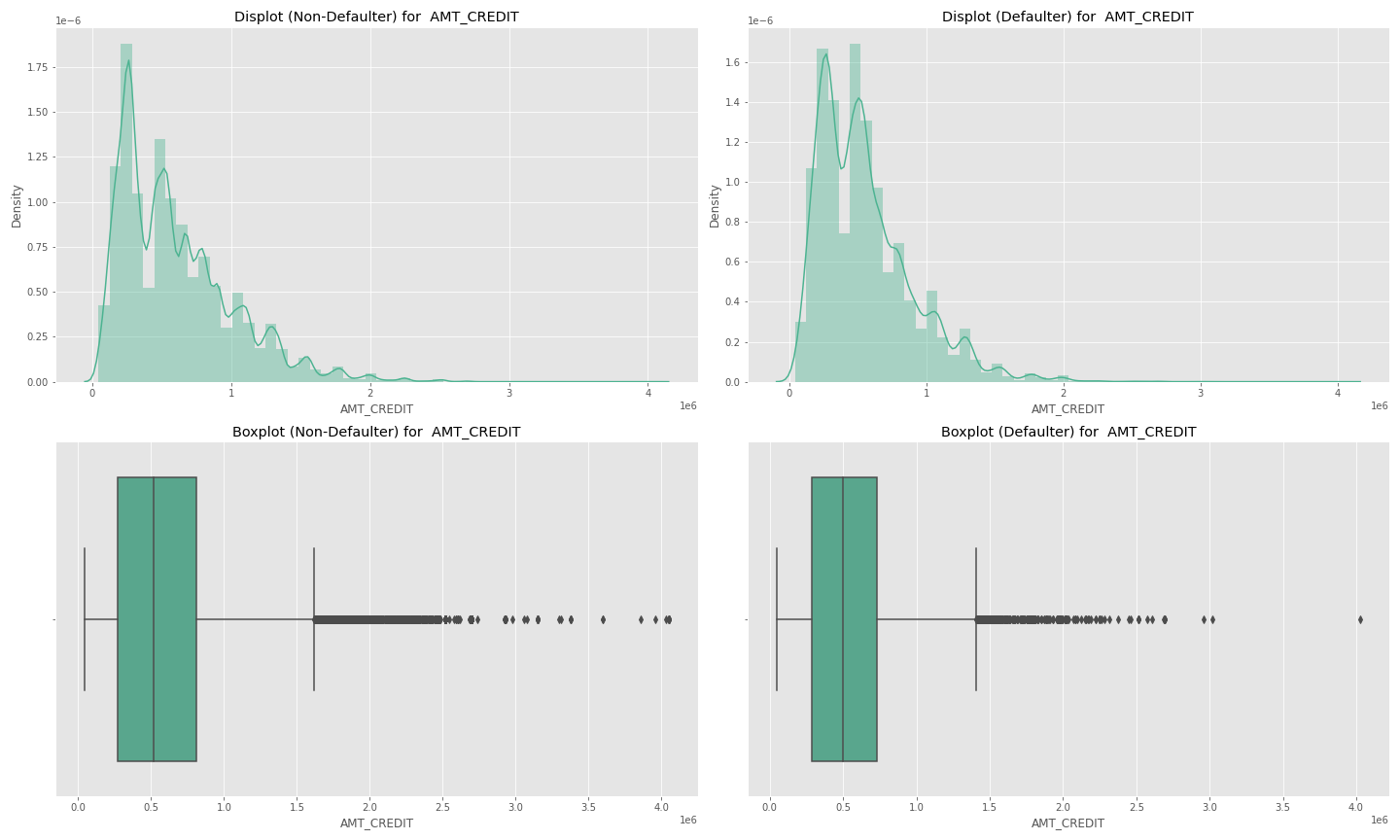
Insights -

* Applicants with lower Annuity Amount are slightly more likely to default on a loan.
* Majority of Loan applicants come from 1st quartile of Annuity data ( Low salary people )

*Loan Amount*

Let’s plot the Density chart and Boxplots for Loan amount across Defaulters and Non Defaulters. It will help us analysis whether lower or higher loans have a correlation with defaults -



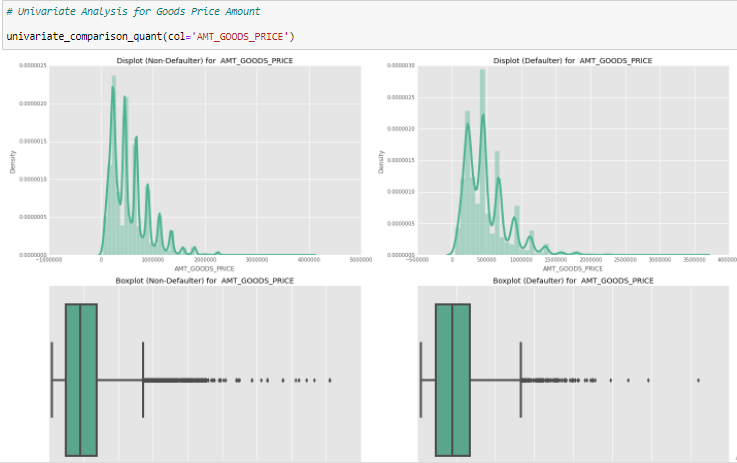


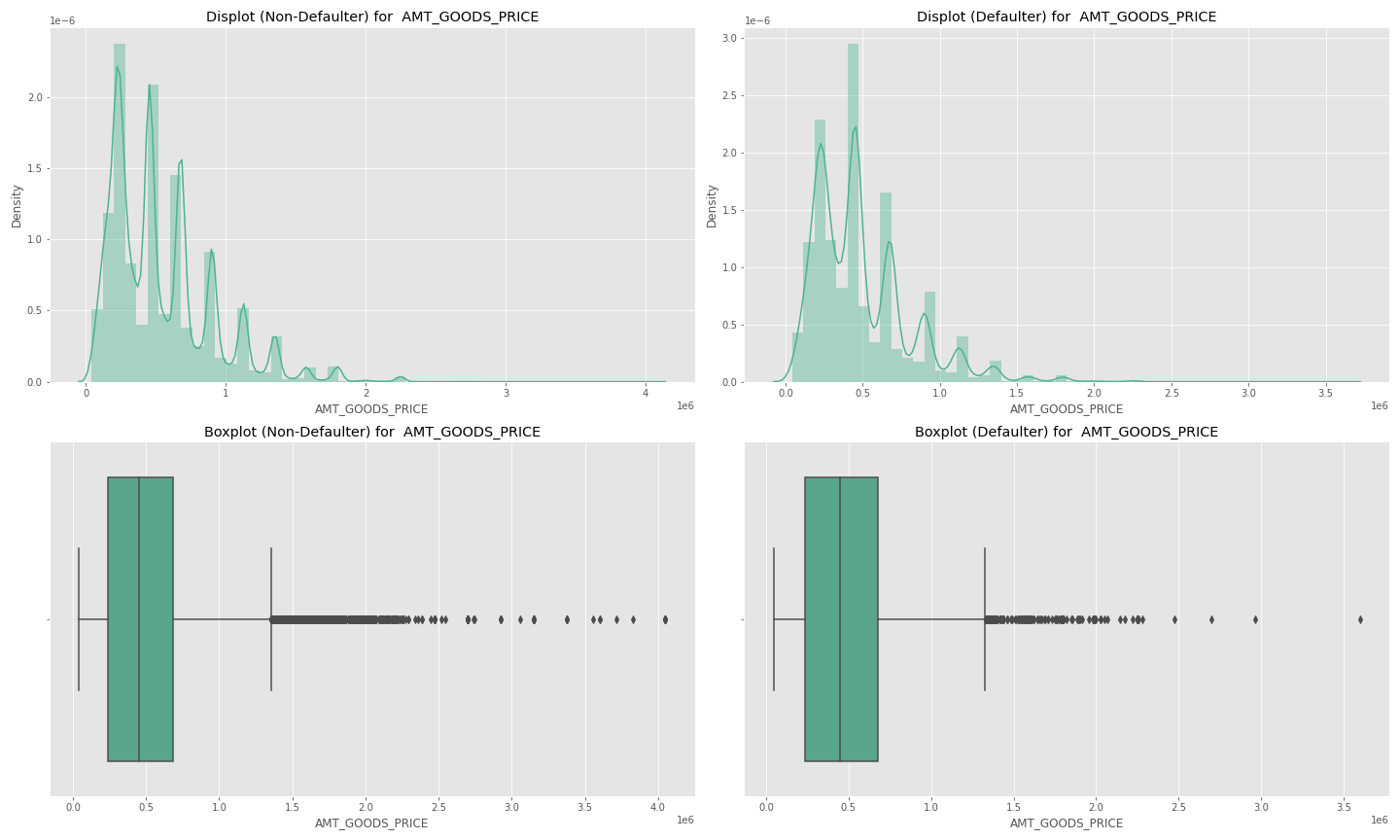
Insights-

* Since the *Loan Amount* have same quartile boundaries in 2 cases, we can infer that *Loan Amount* doesn't have any correlation with *Loan defaults*.

*Goods Price*

Similarly, let’s check whether Good Price amount has an impact on loan defaults -





Insights-

* The distribution are almost unchanged for Defaulters and Non Defaulters, hence we can say that Goods Price doesn't impact the chance of a loan default.

*Alternate Chart Ideas –*

Alternatively, any chart that represents a distribution can be used here for above analysis. Some examples involve Swarmplots and Violinplots.

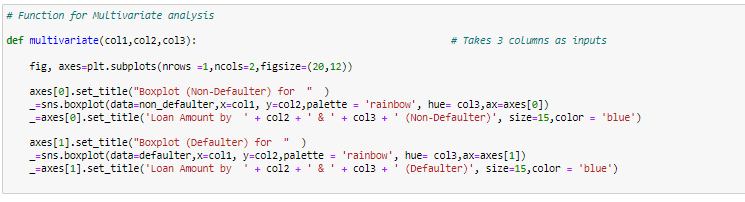
## 3. Bivariate & Multivariate Analysis ( 50 minutes )

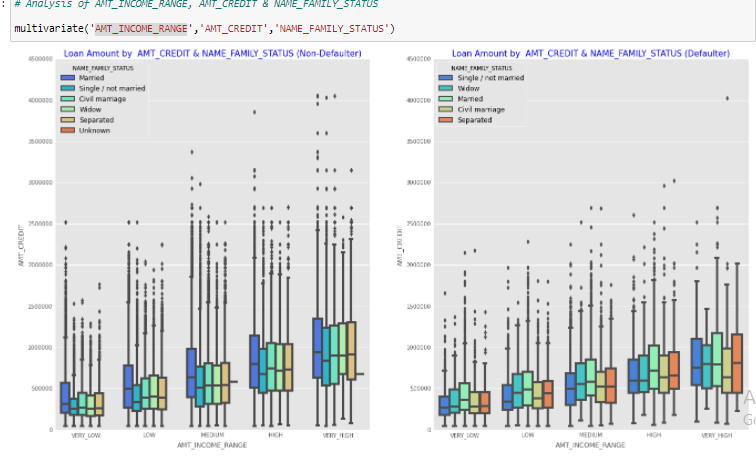
Bivariate Analysis -

It is one of the simplest form of statistical analysis where 2 variables are involved. It looks for relationship among the 2 variables. The applications involve hypothesis validation of association among variables, finding trends, regression etc.

Multivariate Analysis-

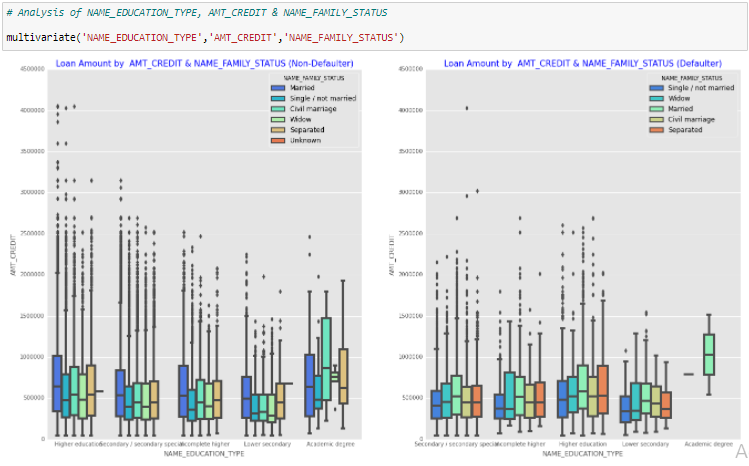
When more than 2 variable are involved in an analysis, it will be a multi-variate analysis. The additional variables may take form of hue color, 3rd axis etc.





Insights-

* With increase in Income range, the loan amount increases proportionally.
* On family status axis, we observe that Married applicants have higher loan amount than others



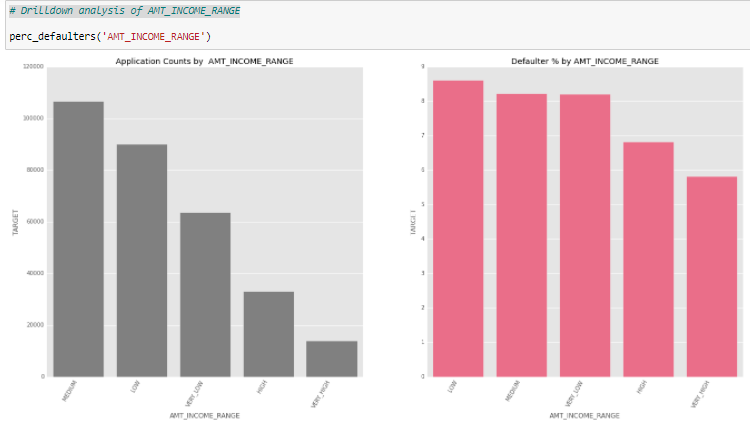
Insights-

* Higher the education, lesser is the likelihood of a loan default
* Among different family status, married ones have the highest likelihood of loan default

#### Drilldown Analysis

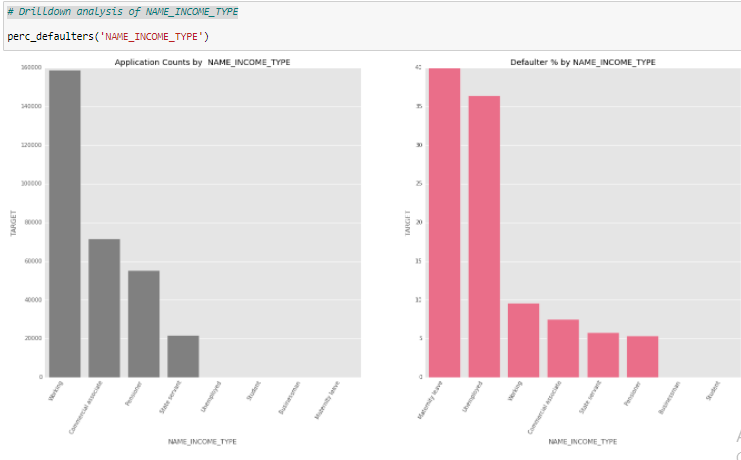
Here we'll look for % defaulters within different classes in a particular variable.





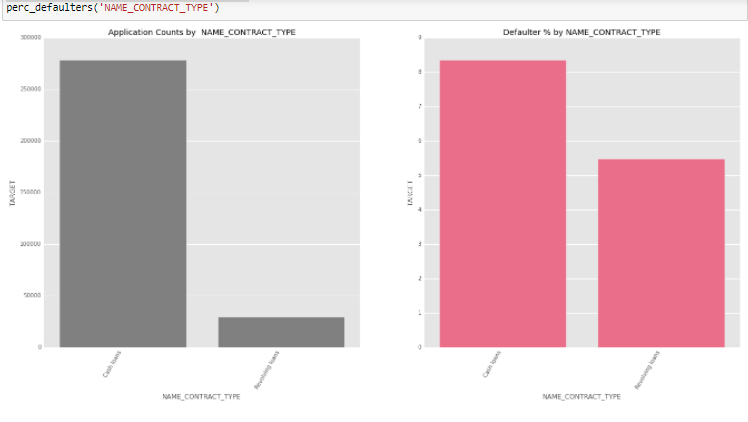
Insights-

* Median income range professionals have maximum applications in the data
* Low Income range have maximum % of loan defaults
* As the Income range increases, loan default probability decreases



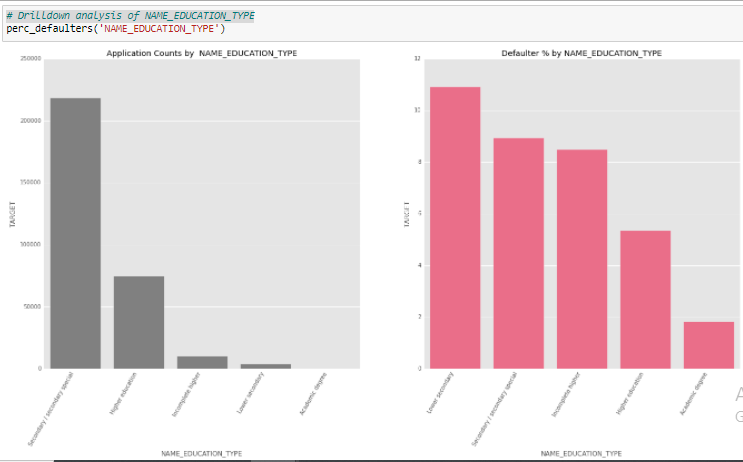
Insights-

* Applicants on Maternity leave have a whopping 40% loan default rate
* The second to the list are Unemployed applicants with 35% loan defaults



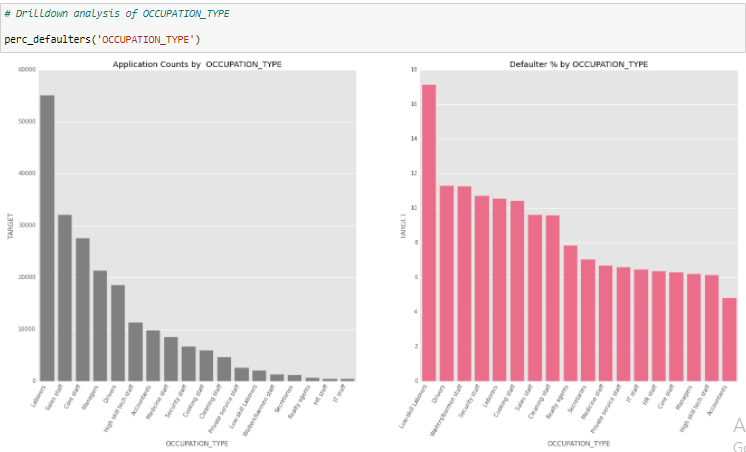
Insights-

* Majority of the loans are cash loans. Cash loans also have almost double probability of a loan default than revolving loans.



Insights-

* Higher the education of an applicant, lesser the chance of loan default
* Lower secondary applicants have a concerning 11% loan default rate, but the count of applicants is low
* The major concern is of Secondary education applicants. They have highest applicants and a significant 9% loan default rate as well.

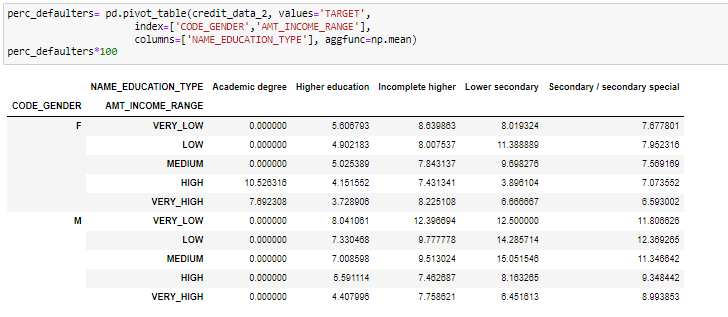


Insights-

* Low skill labourers have an alarming 17% loan default rate. The positive here is that they don't have a high applicant count.
* Labourers & Sales staff will be a major area of concern here, with maximum applicants and a significant loan default rate as well.
* Drivers also have an alarming combination of counts and default %.

##### *Pivot table of all loan default %*

The purpose of deriving this table is have an exhaustive matrix of loan default % for multiple combinations of variables of our interest. We can quickly filter out the higher %s and derive insights as below -



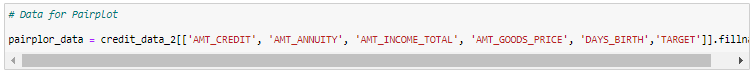
Insights -

Categories with more than 9% default rate -

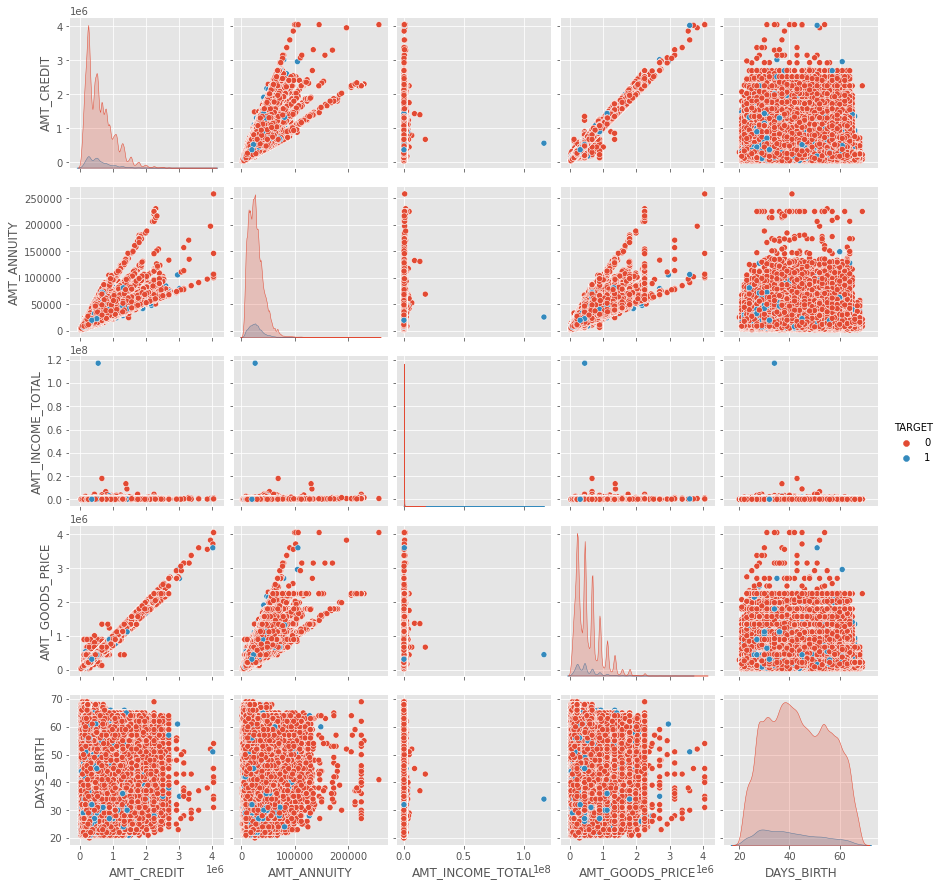
* Females, High Income, Academic degree
* Male, Very Low income , Incomplete higher
* Male, Low Income , Incomplete higher
* Male, Medium Income , Incomplete higher
* Female, Low Income, Lower Secondary
* Female, Medium Income, Lower Secondary
* Male, Very Low Income, Lower Secondary
* Male, Low Income, Lower Secondary
* Male, Medium Income, Lower Secondary
* Male, {ALL INCOME RANGES} , Secondary

##### *Bivariate Analysis using Pairplot*

As discussed in last session, Pairplot is used to check the patterns or trends among multiple combinations of continuous variables in one glance -







Insights-

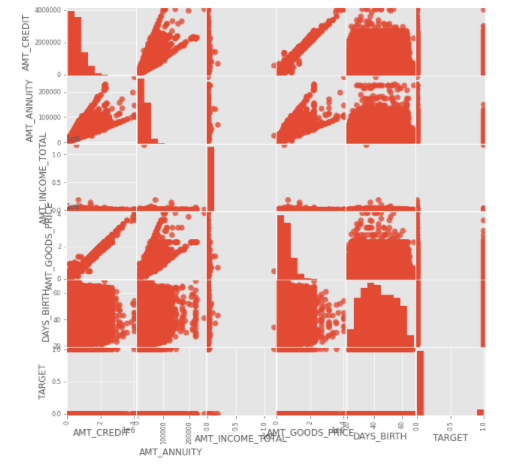
* AMT\_CREDIT & AMT\_GOODS\_PRICE are correlated ( With higher priced goods, loan amount is higher)
* AMT\_ANNUITY & AMT\_GOODS\_PRICE are also correlated ( With higher annuity, expensive goods are purchased)
* AMT\_ANNUITY & AMT\_CREDIT are correlated (Higher the annuity, higher the loan amount)

With respect to TARGET -

* Loan defaulters ( Blue ) are younger in age

*Pairplot using Matplotplib instead of Seaborn –*





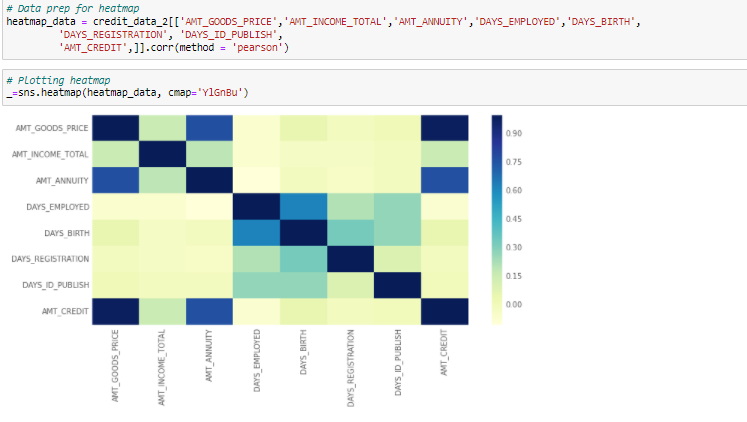
We can see that pairplot is a messy chart with problems such as overlapped labels, overlapped points, split grey areas etc. to name a few. All this is despite being the fact that the code is a more complex one with more variables.

This is a very good example to show the beauty of Seaborn and shows how much under the hood work is done by Seaborn to make such a raw chart visibly aesthetics without much input from the developer/analyst.

Although, as discussed in earlier sessions, there are cases where we still don't have the luxury to use Seaborn - one good example of such case is Pie chart. In seaborn, we don't have a direct function for Pie chart like Pyplot or Matplotplib. Also, in case one wants to experiment with creative charts using subplots and axis objects, there also Matplotlib is the way to go.

##### ***Correlation Check using Heatmap***

The purpose of Heatmap here is to look for any significant correlations among continuous variables in our data. It has similar applications to a pairplot, the heatmap is more useful in scenarios where a pairplot is not sufficient enough to visibly inform regarding correlations in the data.

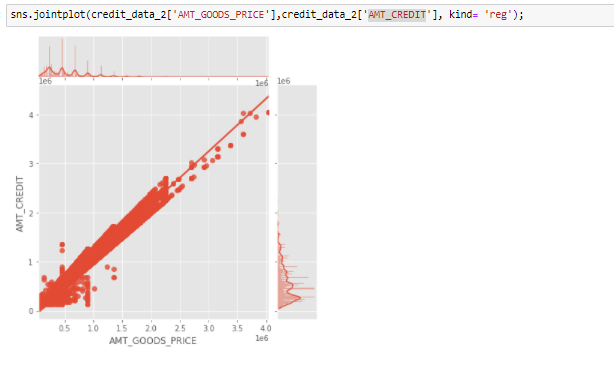


Insights-

* The heatmap confirms our correlation findings from pariplot

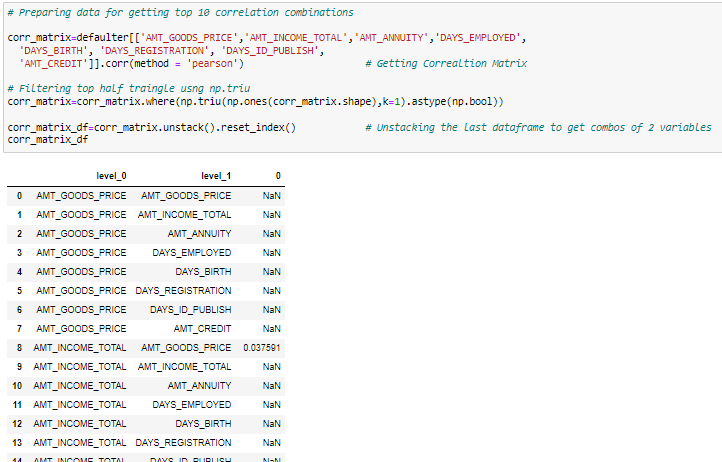
The correlations can be further confirmed and deeply analysed via a *jointplot*.

Let's check the same for AMT\_GOODS\_PRICE & AMT\_CREDIT which have the maximum correlation in our data –



##### *Top 10 correlations in the data*

Although this analysis in not needed here since we have already confirmed 3 insightful correlations in the data, but in other cases with hundreds of variables, this kind of analysis is useful to filter out and focus on the top N correlations as below -



# 4. Final Insights ( 5 minutes )

Following are the driving factors for a loan default -

* Lower the highest education of an applicant, higher the chance of loan default. This is one of the core driving factor in loan defaults.
* Labourers & Sales staff are major area of concern , with maximum applicants and a significant loan default rate. Drivers also have an alarming combination of counts and default %.
* Applicants on Maternity leave have a whopping 40% loan default rate. Unemployed applicants also have 35% loan defaults
* Low Income range have maximum % of loan defaults. As the Income range increases, loan default probability decreases
* Among different family status, married ones have the highest likelihood of loan default
* Applicants with lower Annuity Amount are slightly more likely to default on a loan.
* Young applicants are more expected to default on a loan.
* More Men default loans as compared to Women

Finally, as a conclusion we can recommend to be careful while approving loans for applicants with lower education, as that is constituting majority of the defaults. Alternatively, some policies can be devised to control such loan defaults.

All other factors – Labourers class, Low income range etc. seem to be logically linked with lower education as well. Hence, fixing that alone should have a chain effect, which would significantly improve the loan repayments.

# Home Work

* 1. There is an another file provided which includes Previous Application data. Perform EDA on that data and derive useful insights.
  2. Incorporate Outlier treatment in this EDA.