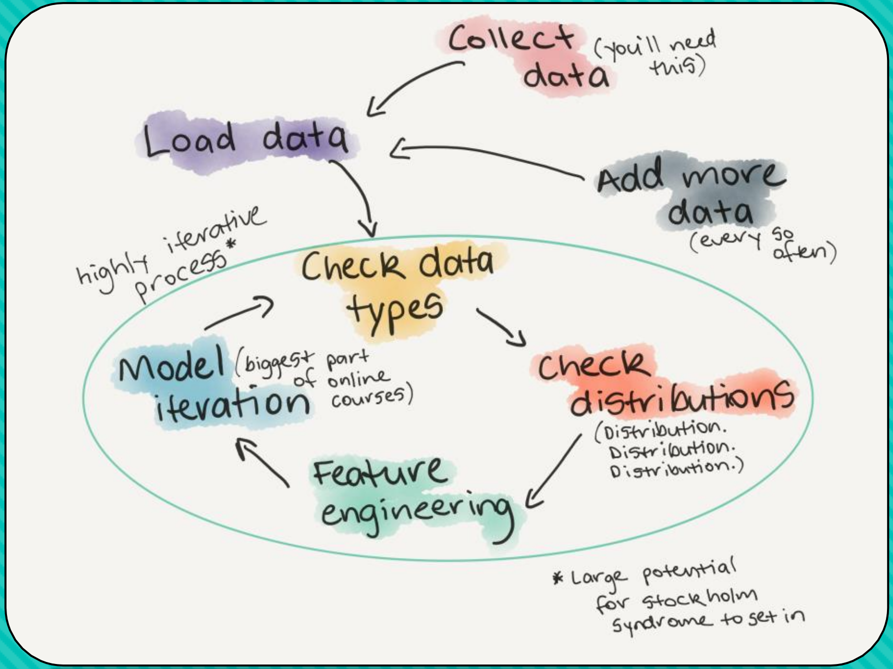
ASSIGNMENT

Credit EDA Presentation.



By→Joyita Sadhukhan

INTRODUCTION:

1. What is EDA?
2. Importance of EDA.
3. Process of EDA.

Check Next Slide

What is EDA?

- Full form of EDA is Exploratory Data Analysis.
- As per IBM EDA is a process used by Data Scientists to analyse and investigate the data sets. It helps to summarize the insights of dataset as a visual representation which is easy to understand by any common mon.
- This process was first developed by American Mathematician Mr. John Tukey in the 1970s.

John Tukey

Importance of EDA:

- EDA helps to determine the errors in Dataset.
- Helps to observe a trend for a particular dataset.
- Detect Data anomalies in EDA.
- Helps in Fraud Detection for Financial Businesses.
- It also helps us to understand the relationships among multiple Variables.
- EDA can also predict future of a business just by analysis the Data.

Process Followed for EDA:

1. Understanding Business Problem
2. Data Collection
3. Data Understanding
4. Data Cleaning
5. Missing Value Treatment
6. Outlier Checks
7. Data Analysis
8. Correlation
9. Insights

Metadata for Both the Datasets.

Check Next Slide

Application_Data (df1):

This data set is having 307511 Rows and 122 Columns.

Most of the columns are either Integer, Float or String.
65 float columns, 41 integer columns, and 16 string or object columns.

```
In [5]: #checking shape o
        df1.shape

Out[5]: (307511, 122)
```

```
df1.info(verbose=True)
 105  FLAG_DOCUMENT_11             int64
 106  FLAG_DOCUMENT_12             int64
 107  FLAG_DOCUMENT_13             int64
 108  FLAG_DOCUMENT_14             int64
 109  FLAG_DOCUMENT_15             int64
 110  FLAG_DOCUMENT_16             int64
 111  FLAG_DOCUMENT_17             int64
 112  FLAG_DOCUMENT_18             int64
 113  FLAG_DOCUMENT_19             int64
 114  FLAG_DOCUMENT_20             int64
 115  FLAG_DOCUMENT_21             int64
 116  AMT_REQ_CREDIT_BUREAU_HOUR   float64
 117  AMT_REQ_CREDIT_BUREAU_DAY    float64
 118  AMT_REQ_CREDIT_BUREAU_WEEK   float64
 119  AMT_REQ_CREDIT_BUREAU_MON    float64
 120  AMT_REQ_CREDIT_BUREAU_QRT    float64
 121  AMT_REQ_CREDIT_BUREAU_YEAR   float64
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

Previous_Application (df2):

This data set is having 1670214 Rows and 37 Columns.

```
df2.shape

Out[9]: (1670214, 37)
```

Most of the columns are either Integer, Float or String.
15 float columns, 6 integer columns, and 16 string or object columns.

```
 21   NAME_CLIENT_TYPE             1670214 non-null  object
 22   NAME_GOODS_CATEGORY          1670214 non-null  object
 23   NAME_PORTFOLIO               1670214 non-null  object
 24   NAME_PRODUCT_TYPE            1670214 non-null  object
 25   CHANNEL_TYPE                 1670214 non-null  object
 26   SELLERPLACE_AREA             1670214 non-null  int64
 27   NAME_SELLER_INDUSTRY         1670214 non-null  object
 28   CNT_PAYMENT                  1297984 non-null  float64
 29   NAME_YIELD_GROUP             1670214 non-null  object
 30   PRODUCT_COMBINATION          1669868 non-null  object
 31   DAYS_FIRST_DRAWING           997149 non-null   float64
 32   DAYS_FIRST_DUE               997149 non-null   float64
 33   DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34   DAYS_LAST_DUE                997149 non-null   float64
 35   DAYS_TERMINATION             997149 non-null   float64
 36   NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB
```

Missing Value Treatment.

Check Next Slide

Missing Value Treatment for df1

- We checked the percentage of missing value in each column.
- If there are more than 50% Missing Values we deleted those columns.
- And imputed median or mode value for the columns which are having less than 50% missing values.
- Imputed median values for numeric columns and mode value for object type columns.
- In df1 almost 41 columns were having more than 50% missing values.

```
In [16]: null_val[null_val>50.00].count()

Out[16]: 41
```

This is showing 41 columns which are having missing values more than 50%

Missing Value Treatment for df2

- We dealt the df2 data set same like df1. more than 50% missing value columns are deleted and others are imputed with median or mode.
- Imputed median values for numeric columns and mode value for object type columns.
- In df2 almost 4 columns were having more than 50% missing values.

```
In [19]: null_val_1[null_val_1>50.00]

Out[19]: AMT_DOWN_PAYMENT              53.636480
         RATE_DOWN_PAYMENT             53.636480
         RATE_INTEREST_PRIMARY         99.643698
         RATE_INTEREST_PRIVILEGED      99.643698
         dtype: float64
```

This is showing 4 columns which are having missing values more than 50%

Data Binning

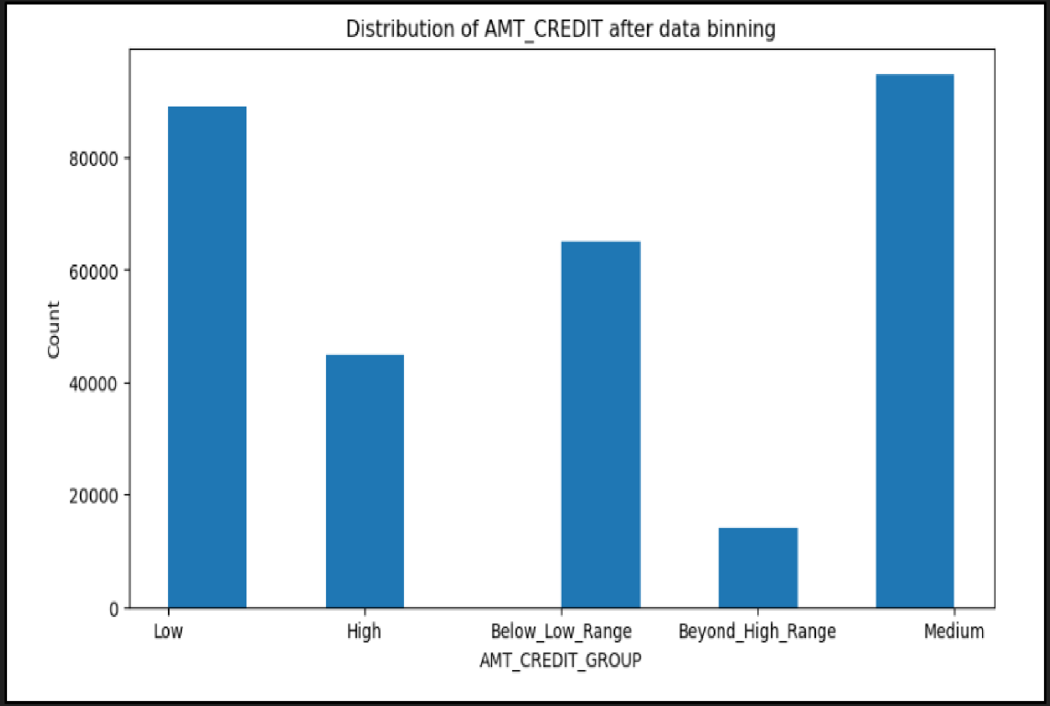
Check Next Slide

Data Binning for df1

- Data Binning is categorizing or grouping numeric values in a range.
- Let’s say we are having a large number of data about people weight. Here instead of analysing all individuals we can group them in a range let’s say 35kgs to 45kgs as Under weight, 46kgs to 55kgs as good weight, 56kgs to 65kgs as moderate weight, 66kgs to 80kgs as over weight.
- This helps us to understand the distribution of numeric values in a range.
- We used binning in 4 variables for df1.
  1. AMT_CREDIT
  2. AMT_ANNUITY
  3. AMT_INCOME_TOTAL
  4. DAYS_BIRTH

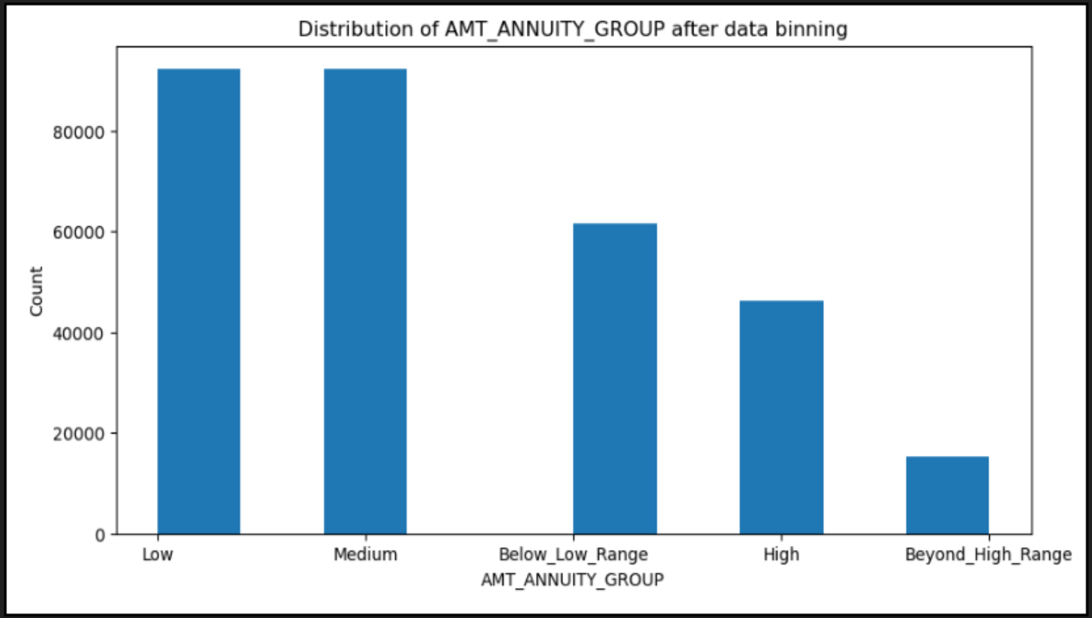
AMT_CREDIT :

- We binned data here with respective of quartiles (0, 0.2, 0.5, 0.8, 0.95, 1) 0.8 quartile values are comes into medium range.
- And we can also observe that highest AMT_CREDIT value are from middle amount range.



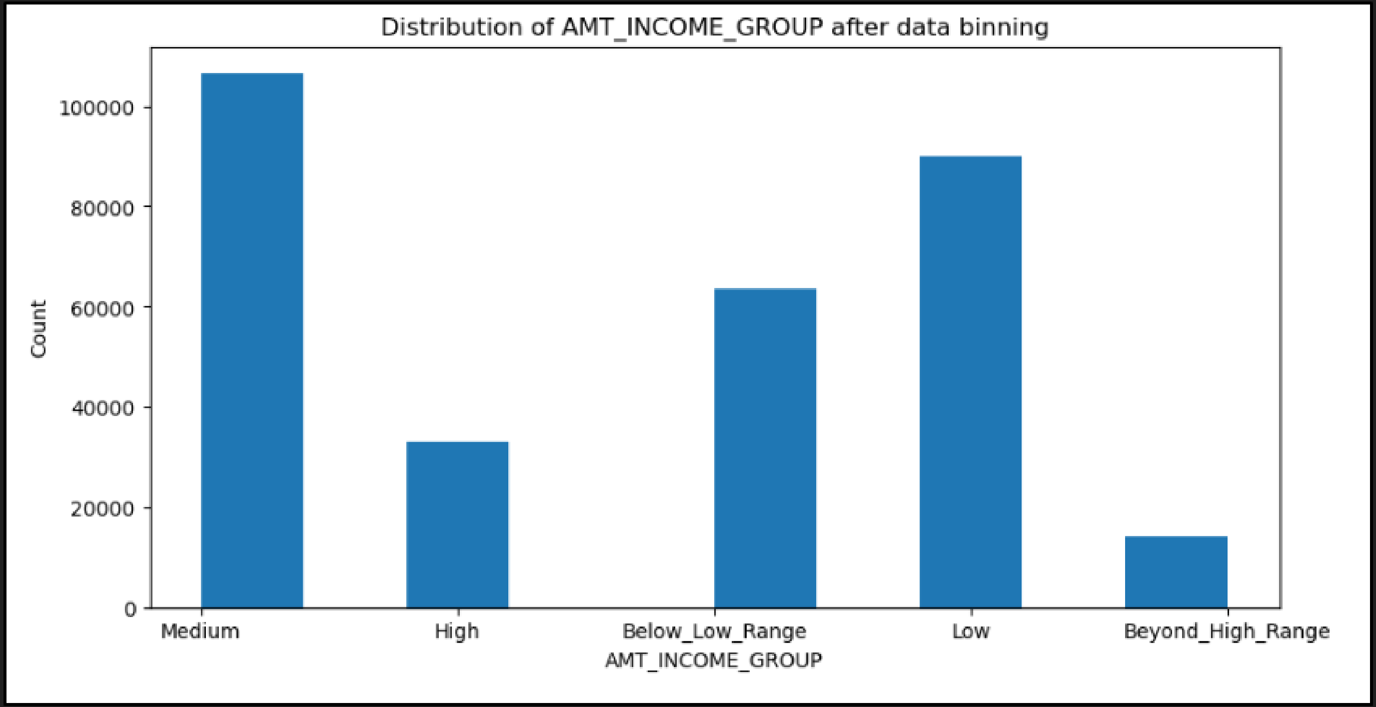
AMT_ANNUITY :

- Same method applied here as well to binning the data like AMT_CREDIT.
- We can observe that highest range of EMIs are belongs to Low and Medium range.



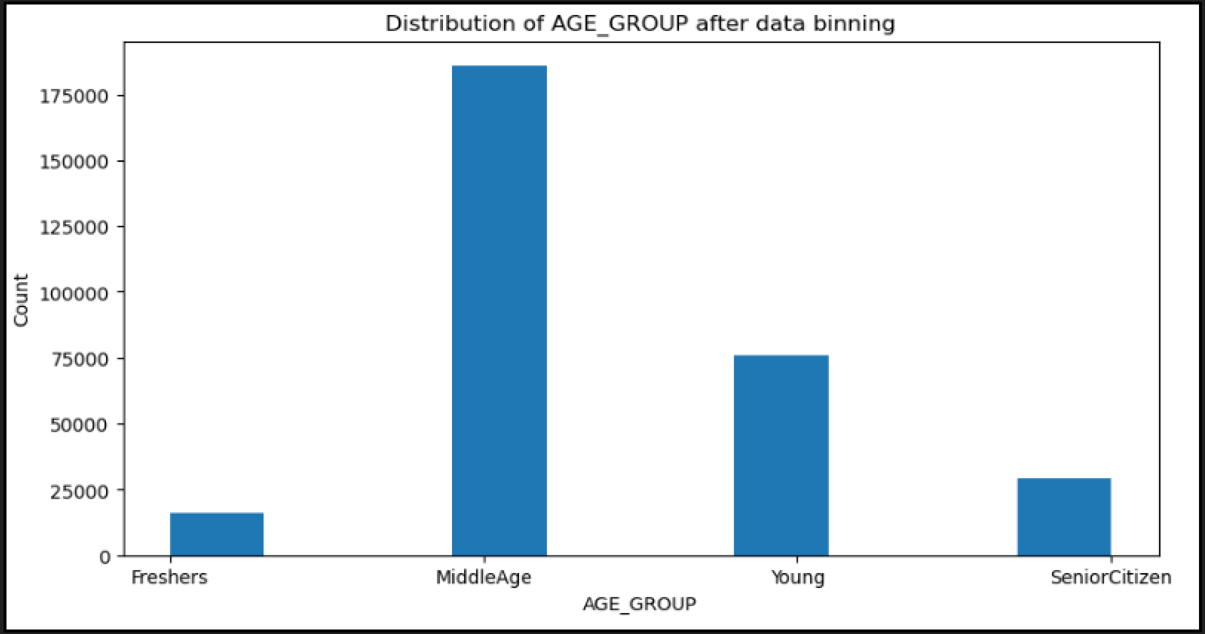
AMT_INCOME_TOTAL :

- Here we can observe that most of the Loan takers are belongs to Medium income range count is more than 100000



DAYS_BIRTH:

- Here data binning is different than others.
- Here is the range:
  1. 19 – 24 = 'Freshers'
  2. 25 – 34 = 'Young'
  3. 35 – 59 = 'Middle Age'
  4. 60 – 100 = 'Senior Citizen'
- Most of the loan takers are belong to middle age groups.



Outlier Checks

Check Next Slide

What is Outlier and Methods:

Values which are beyond the range or some values which are not normal for a column are identified as Outliers.
There are 2 ways to check outliers which are discussed below:

Z-score

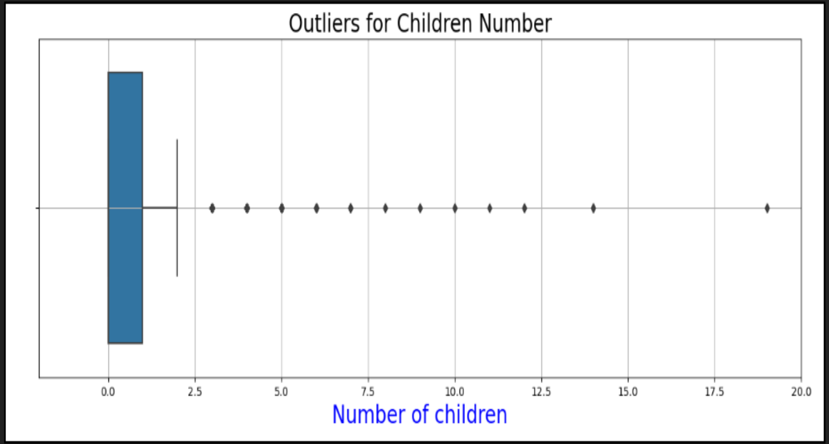
- it tells the distance from a point to the mean of the dataset in the units of standard deviation. if z-score is greater than 3 or less than -3 (-4 / -3.5 is outlier) considered as outliers.
- Limitations:
  1. It gives output according to normal distribution.
  2. Also mean and standard deviation can be easily distorted by extremely high data point.
  3. To avoid this we need to use median and then need to modify the z-score which is again a long process and error prone.

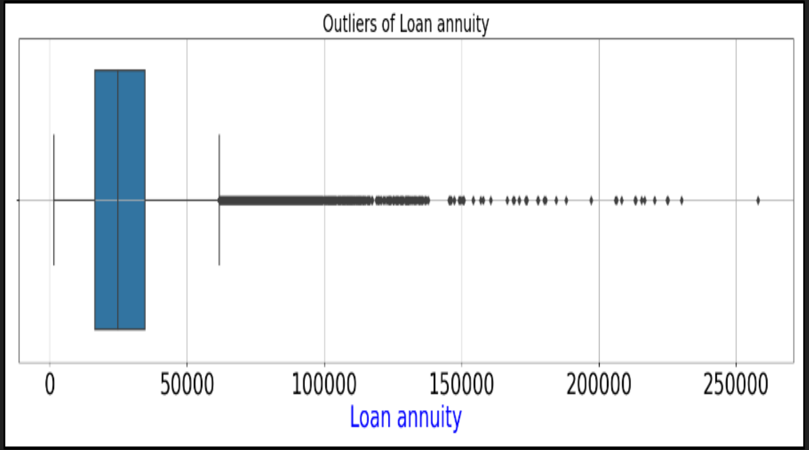
Boxplot

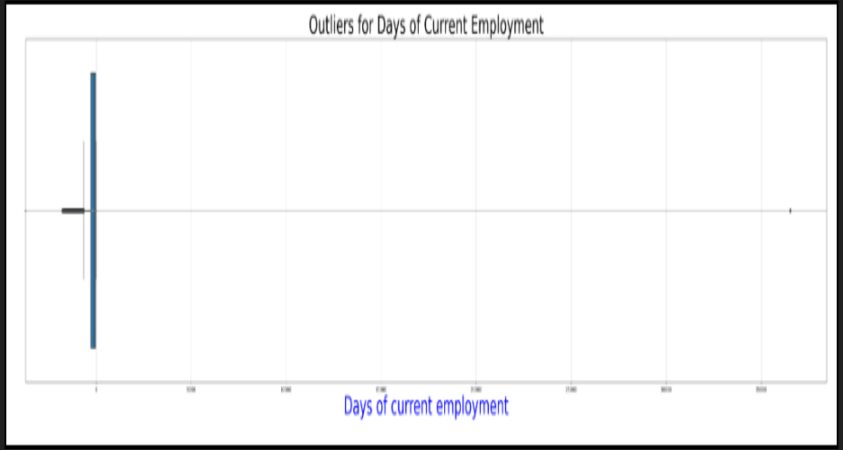
- It gives value of outliers on the basis of IQR or Interquartile Range. it always take median as standard so no other process need to be followed.
- Here we just need to calculate IQR which can be done by the formula:
  IQR = Q3 - Q1 where Q3 = 75th percentile and Q1 = 25th percentile.
- We also needs to found lower bound and upper bound in order to find the outliers.
  Lower Bound = Q1-(1.5*IQR)*
  *Upper Bound = Q3+(1.5*IQR)
- If any value lies beyond Lower and upper bound those are considered as outliers.
- We will proceed with boxplot method here as it is more efficient and time saving.

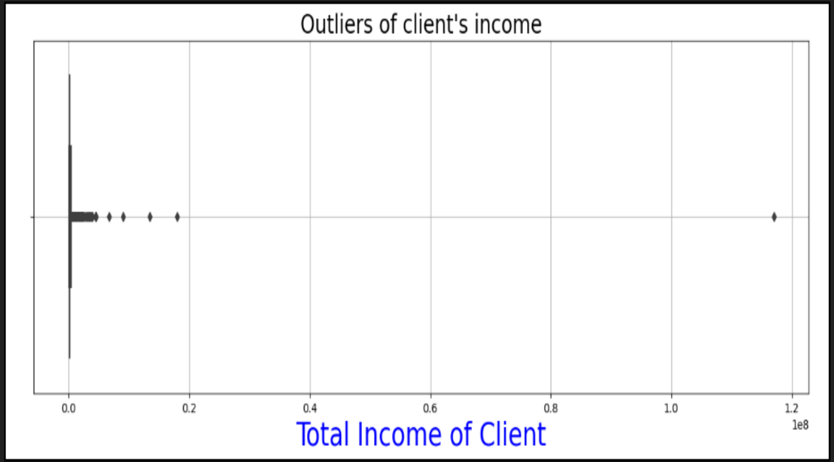
Outliers Checked For Variables:

- We have checked outliers for almost 9 variables.
- Observations are noted below:
  1. In case of **children numbers** from 4 to 20 numbers of children can be counted as Outlier.
  2. In case of **income** IQR is very slim and a large number of outliers are present.
  3. For **Loan Credit** large number of outliers are present but IQR is richer.
  4. **Loan Annuity** is also having a good number of IQR unlike AMT_INCOME_TOTAL. range of outliers is also big and the highest outlier is greater then 250000.
  5. In **Goods Price** 3rd quartile is larger than 1st Quartile and outliers are also having a greater range.
  6. For **Days Birth** there is no outliers.
  7. In **Days of Employment** most of the outliers are on lower bound side and 1 outlier is present after 350000.
  8. In **Count of Family Members** IQR is having a good range and outliers started from 10 to 20. We can also conclude most of the clients are having 4 family members or they are Nuclear family.
  9. **External Sources** is another example of 0 outliers.
- In the next slide we will put all the Boxplots for each Variables.

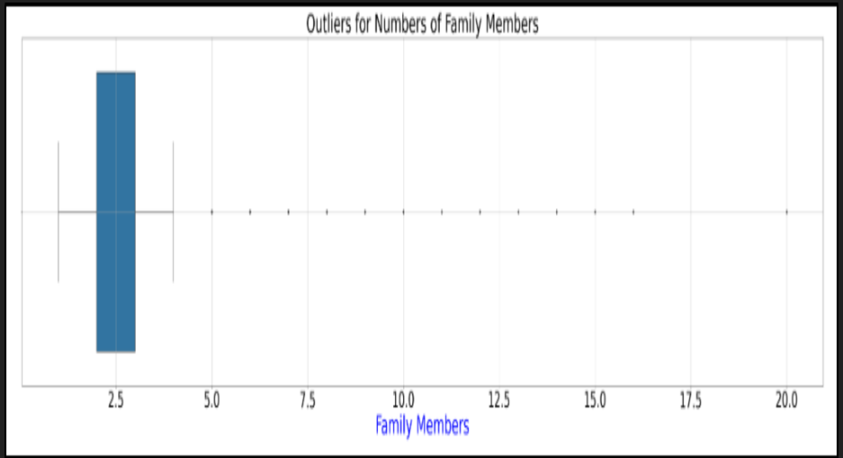

Outliers for Children Number
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of children

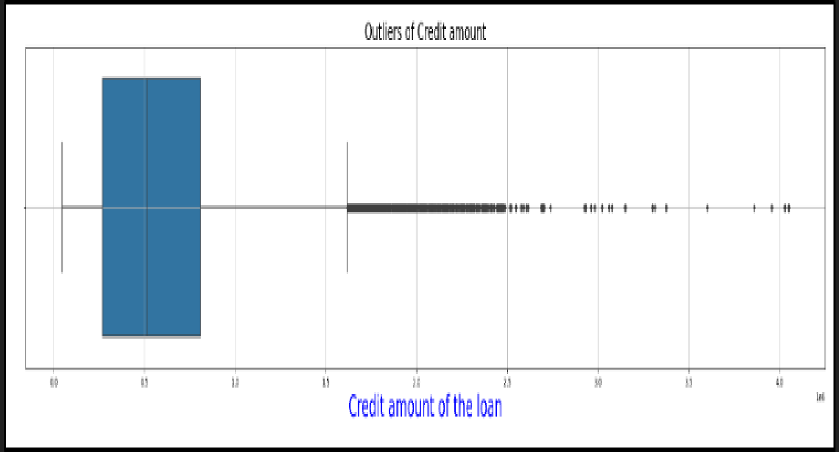

Outliers of Loan annuity
0
50000
100000
150000
200000
250000
Loan annuity

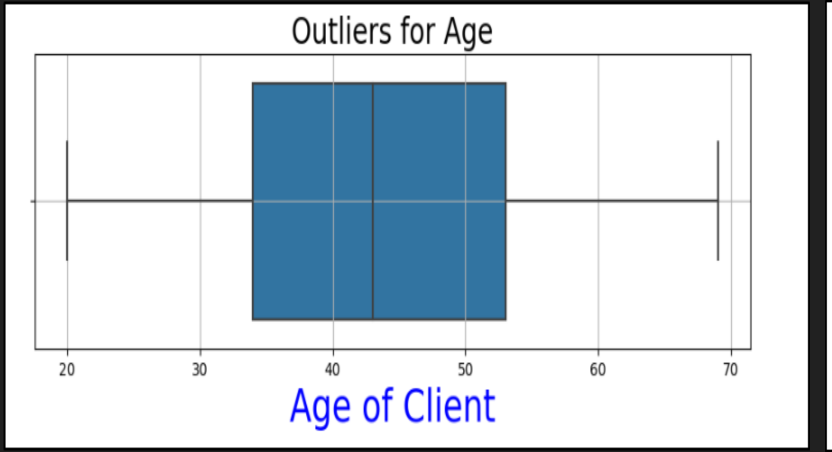

Outliers for Days of Current Employment
Days of current employment

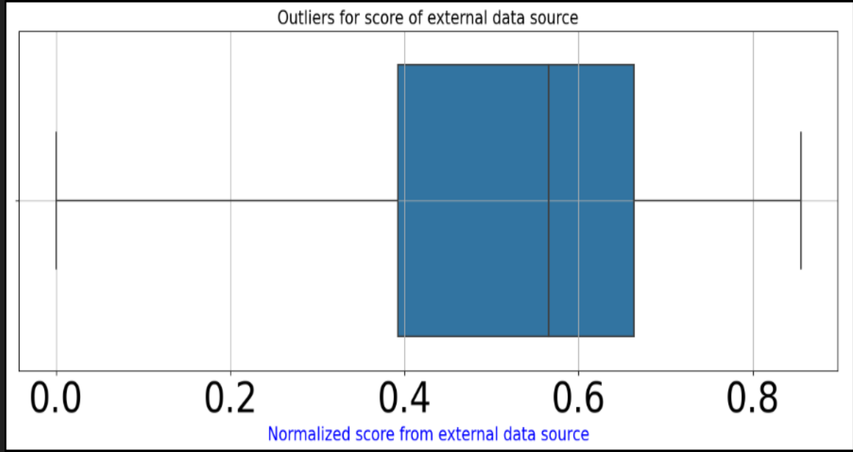

Outliers of client's income
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e8
Total Income of Client


Outliers of Good price
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
Goods price


Outliers for Numbers of Family Members
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Family Members


Outliers of Credit amount
Credit amount of the loan


Outliers for Age
20
30
40
50
60
70
Age of Client


Outliers for score of external data source
0.0
0.2
0.4
0.6
0.8
Normalized score from external data source

Data Imbalance

Check Next Slide

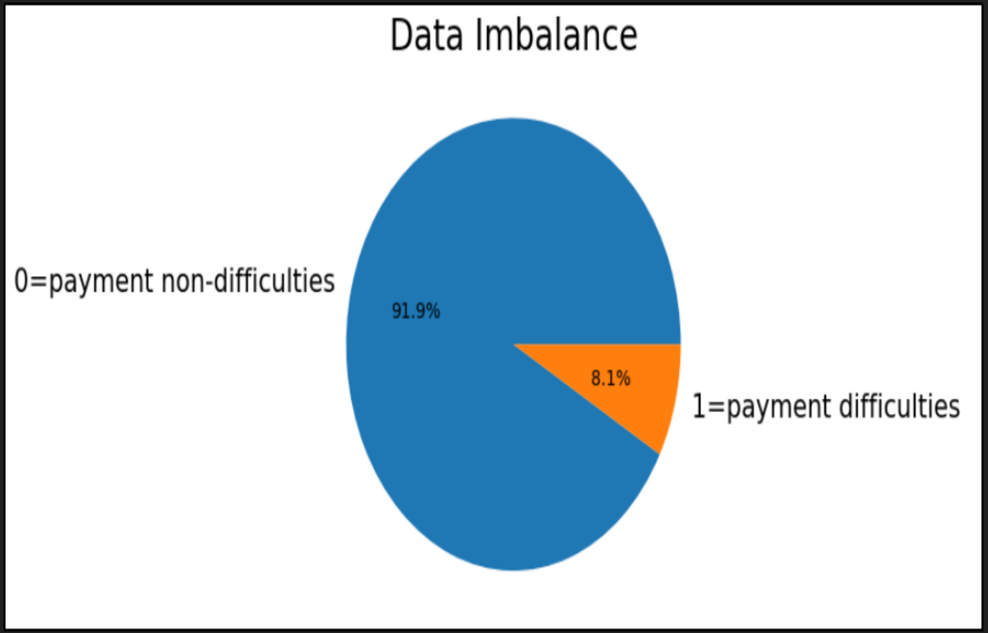
What is Data Imbalance?

- When in a data set only one Target class shows major observation. Then it is called Imbalanced Data. This issue happens in real world data.
- In case of imbalanced date we need to focus on statistic choices more.
- Uses of Data Imbalance:
  1. Fraud Detection
  2. Disease Diagnosis
  3. Data Anomaly

Difference between balanced and imbalance data

Data imbalance in df1.

- In case of df1 the data is highly imbalanced.
- There are 91.9% or ~92% population is non-defaulters and remaining 8.1% are defaulters.
- On basis of this defaulters and non-defaulters we will proceed with data analysis in next step.
- We have represented this data imbalance with a pie chart which will give us a clear understanding.



Univariate Data Analysis

What is univariate analysis?

univariate analysis describe the trend or data graphically for a single variable. Here variable means column. there are multiple graphs available for univariate analysis. like:

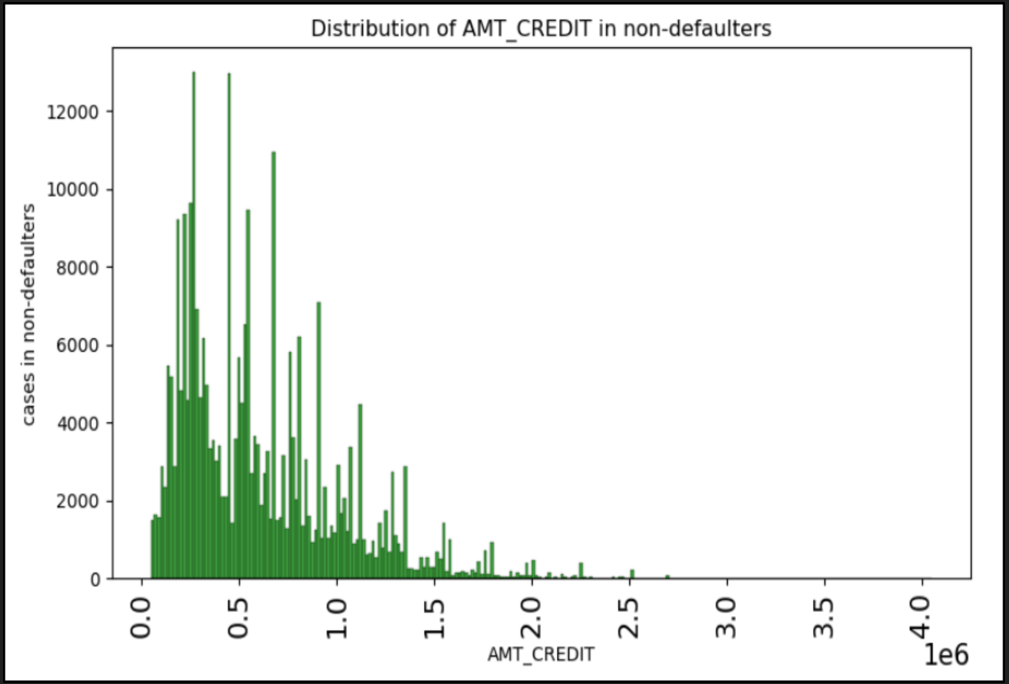
1. Histogram
2. Pie chart
3. Line chart
4. boxplot

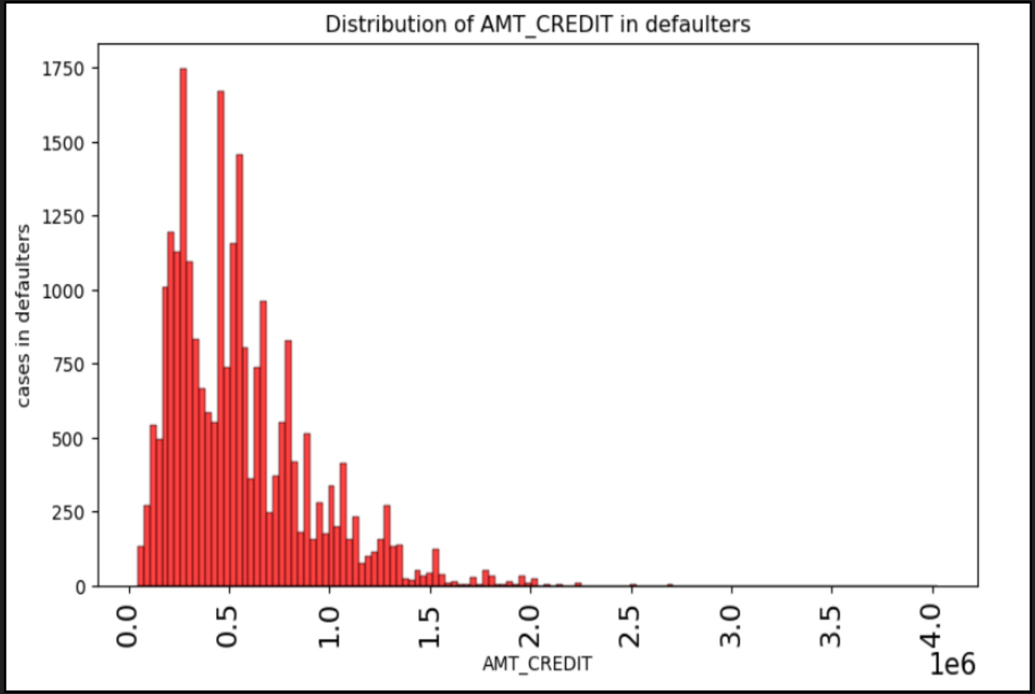
but **histogram** is the most efficient in terms of visualization and for time trend **line chart** is the one.

Check Next Slide

AMT_CREDIT :

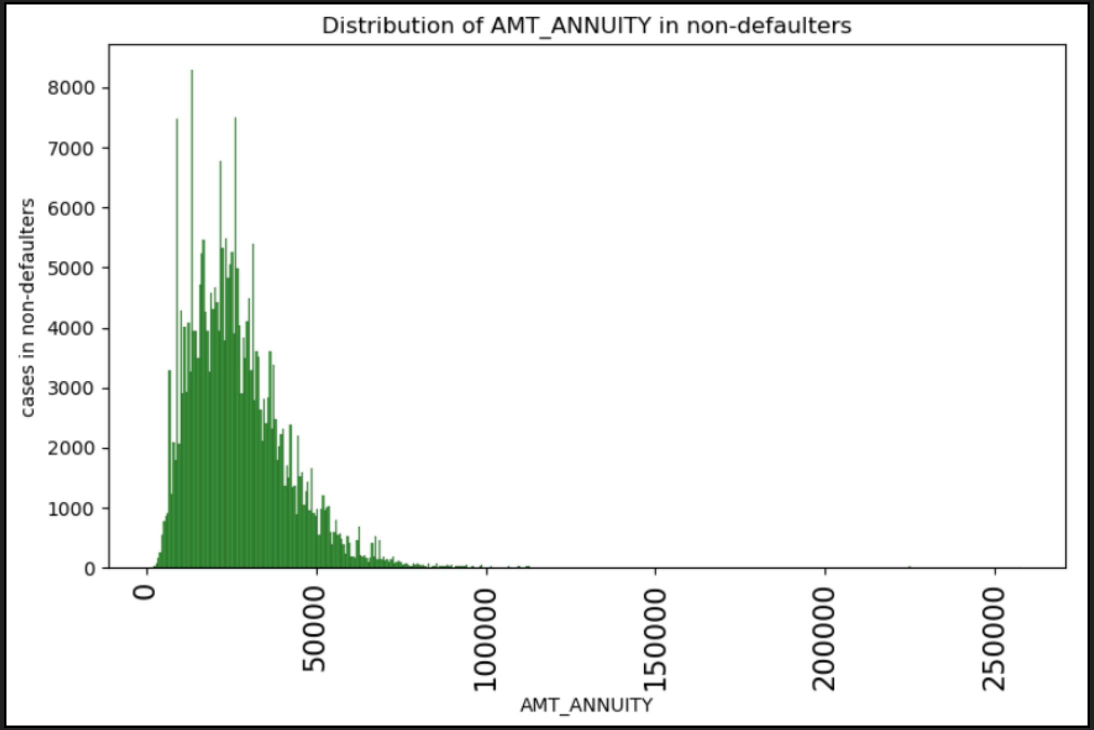
- Most of the non defaulters are taken loan in range of 0.3 to 0.5 quartiles.

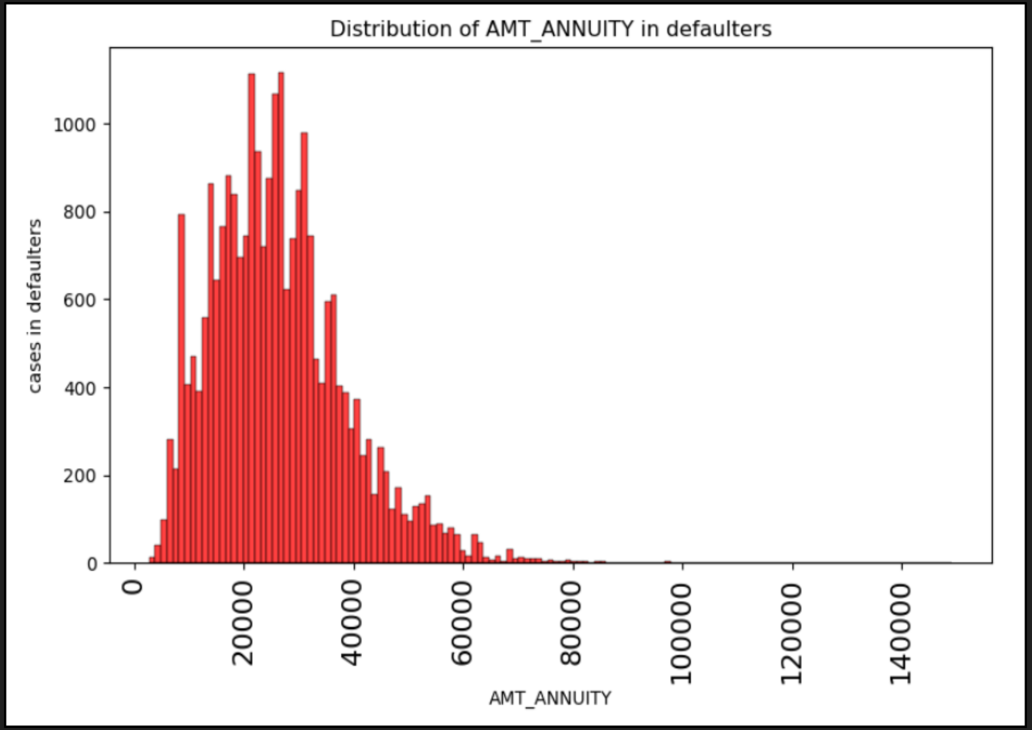




AMT_ANNUITY :

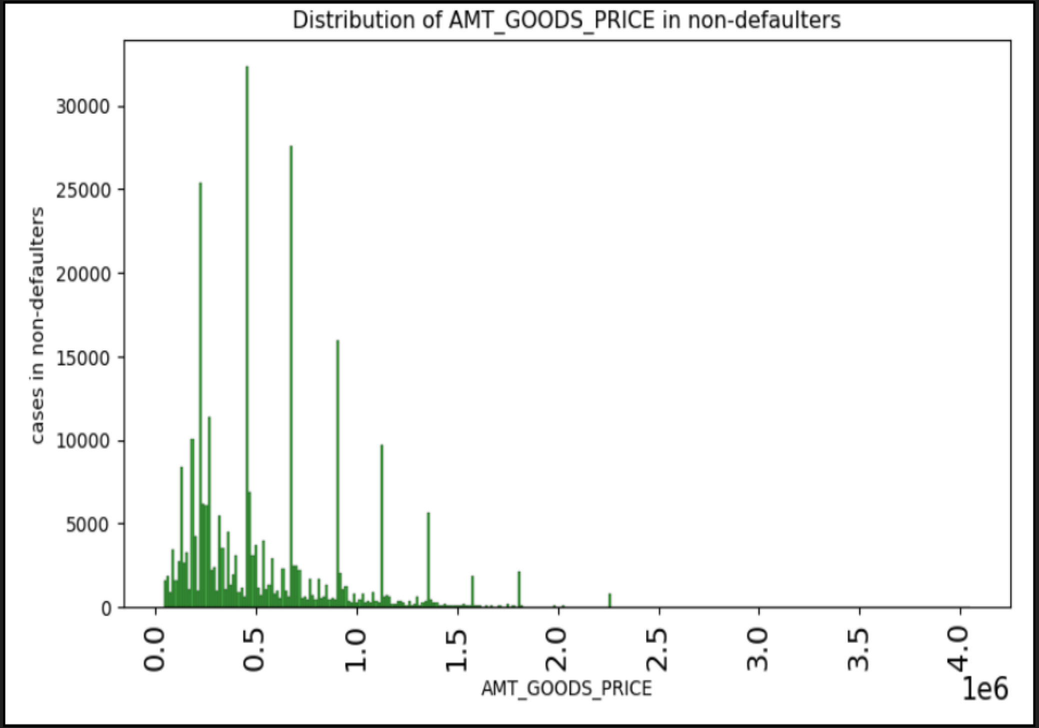
- People who have **loan annuity** of 20k to 50k are more likely to be a non-defaulters

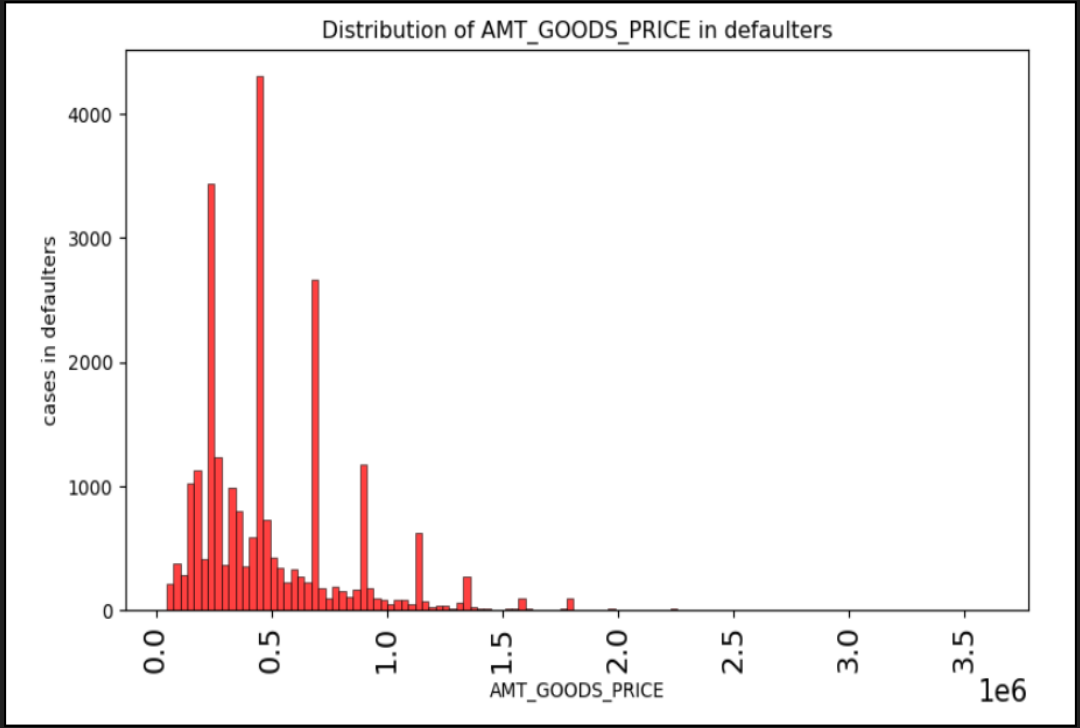




AMT_GOODS_PRICE :

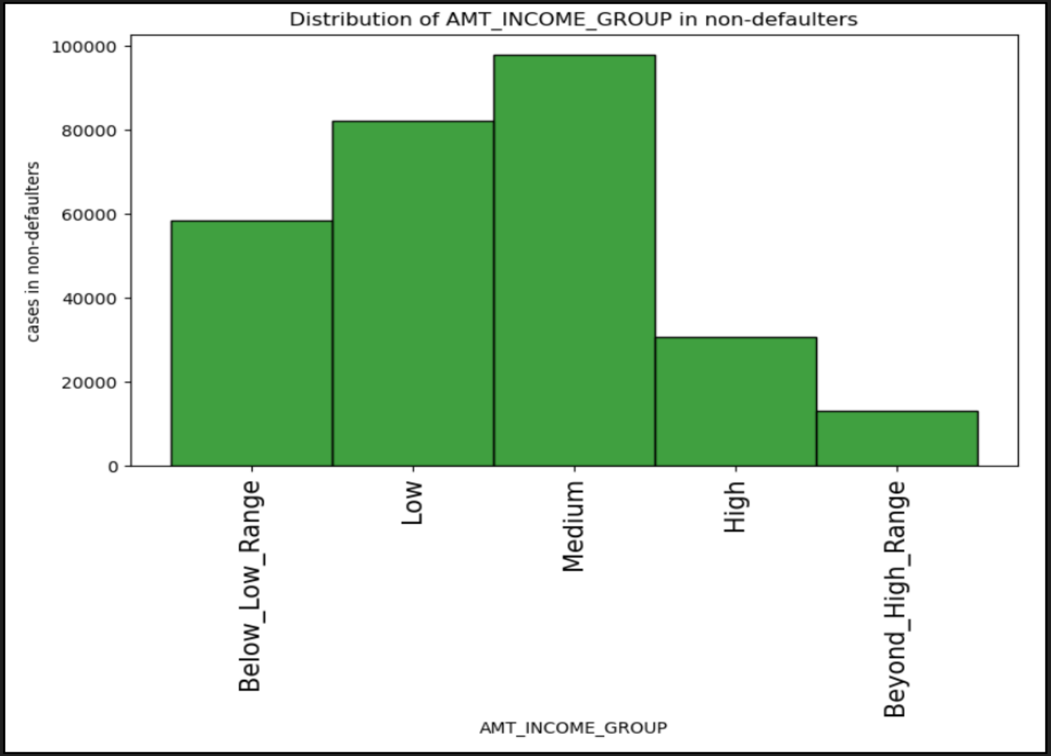
- People who have goods in range of 0.25 to 0.55 quartile are mostly a non-defaulters

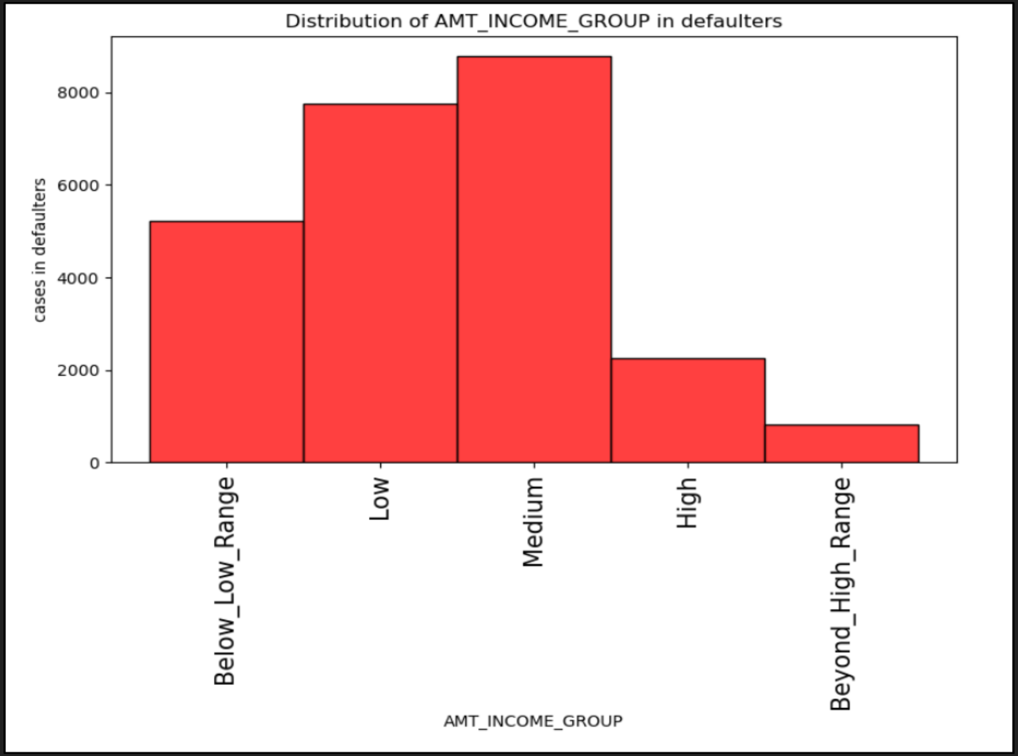




AMT_INCOME_GROUP :

- People who are having **income range of medium** they are more likely to be a non-defaulters.





Bivariate Data Analysis

What is Bivariate Analysis?

bivariate analysis always gives us a relationship between 2 variables. for bivariate analysis we can use multiple graphs like:

1. Bar Graph
2. Scatterplot and
3. boxplot

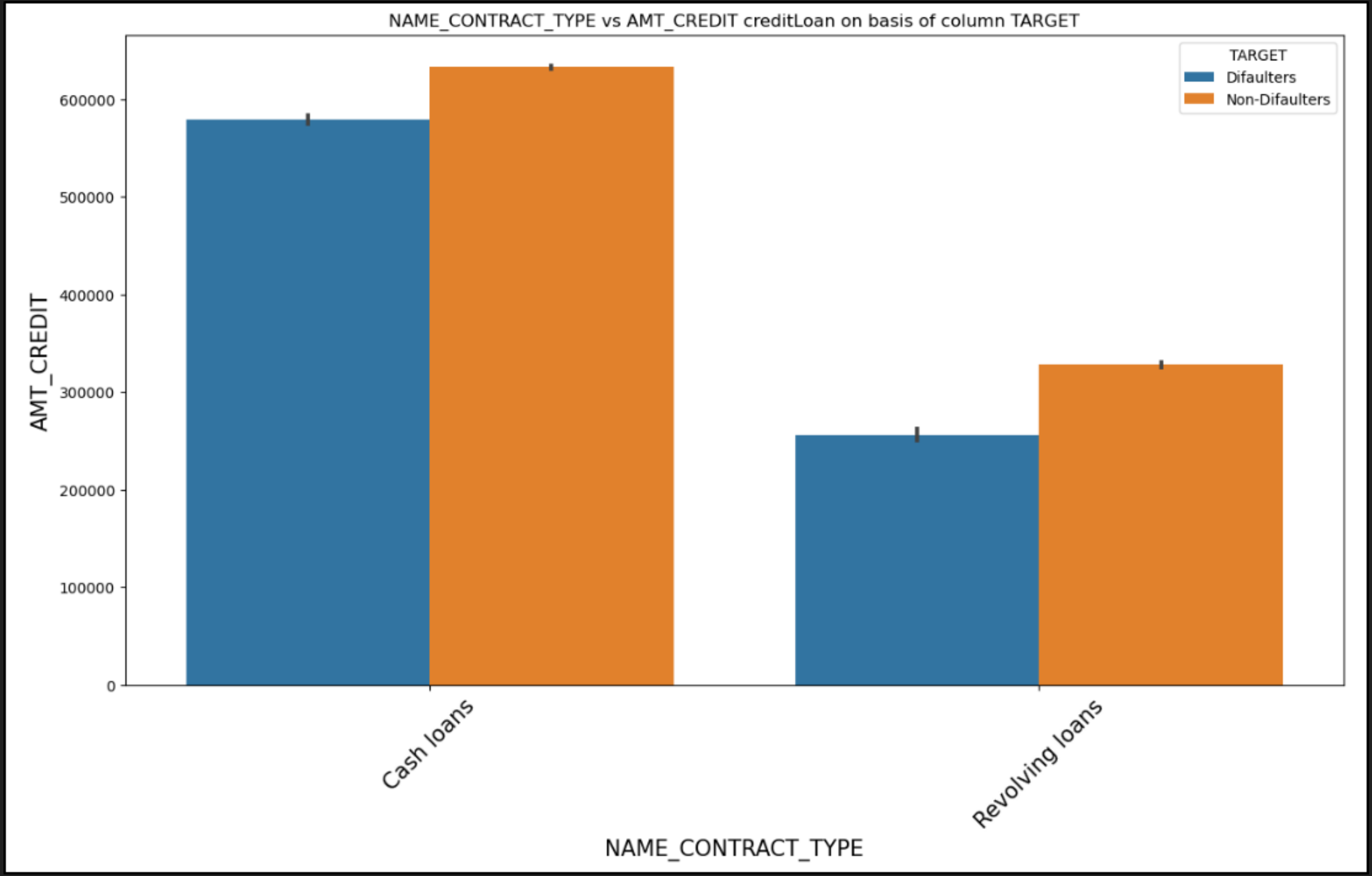
But we will mostly use here scatterplot and Bar graphs.

- Scatterplot = this use for showing relation between 2 numerical variables
- Bar Graph = this use for showing relationship between categorical and numerical varible

Check Next Slide

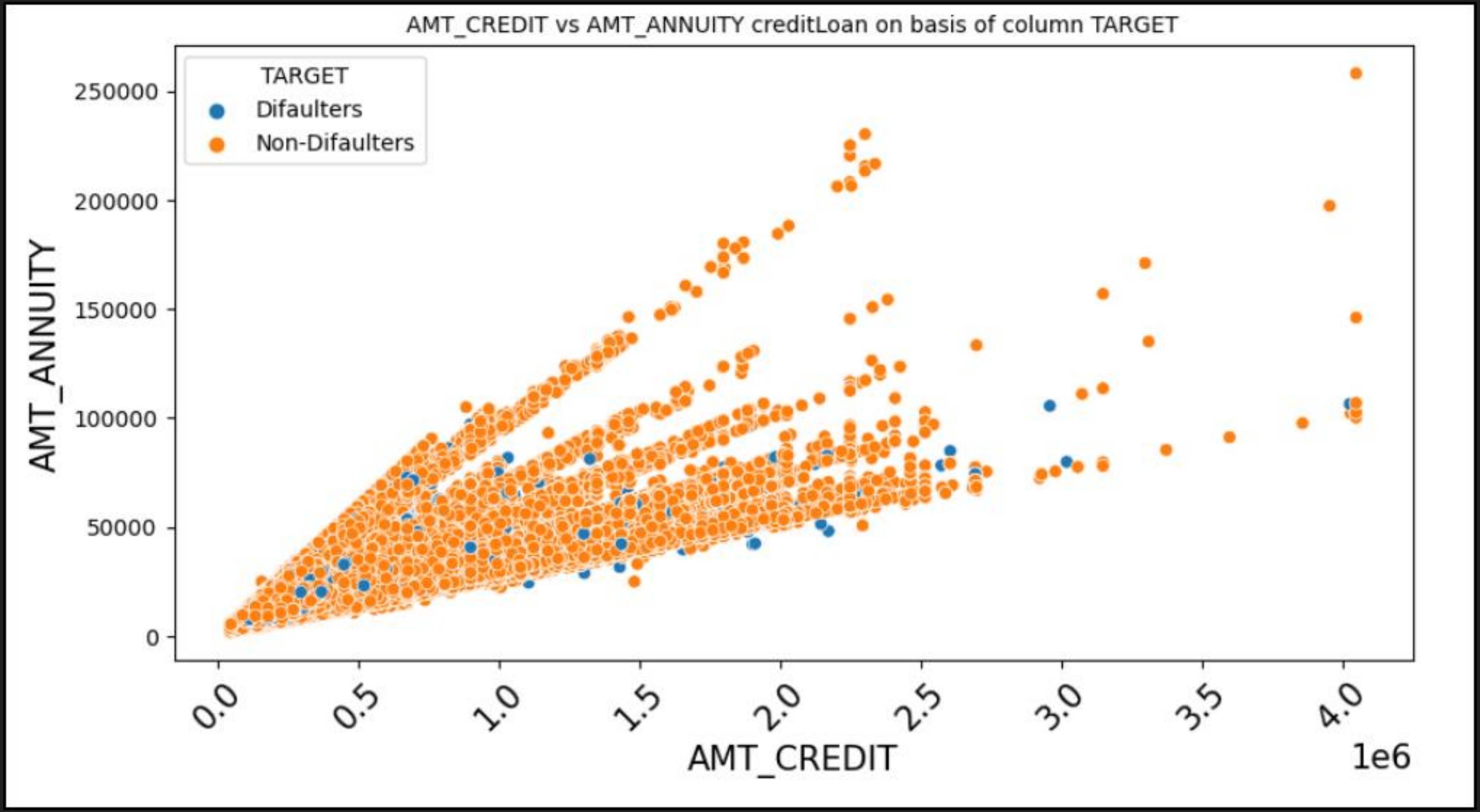
AMT_CREDIT vs NAME_CONTRACT_TYPE :

- Most of the loans are taken as Cash and count of non-defaulters are higher in number.
- Amount of highest cash loan is more than 6 Lakhs.
- Also defaulters are from cash loans.



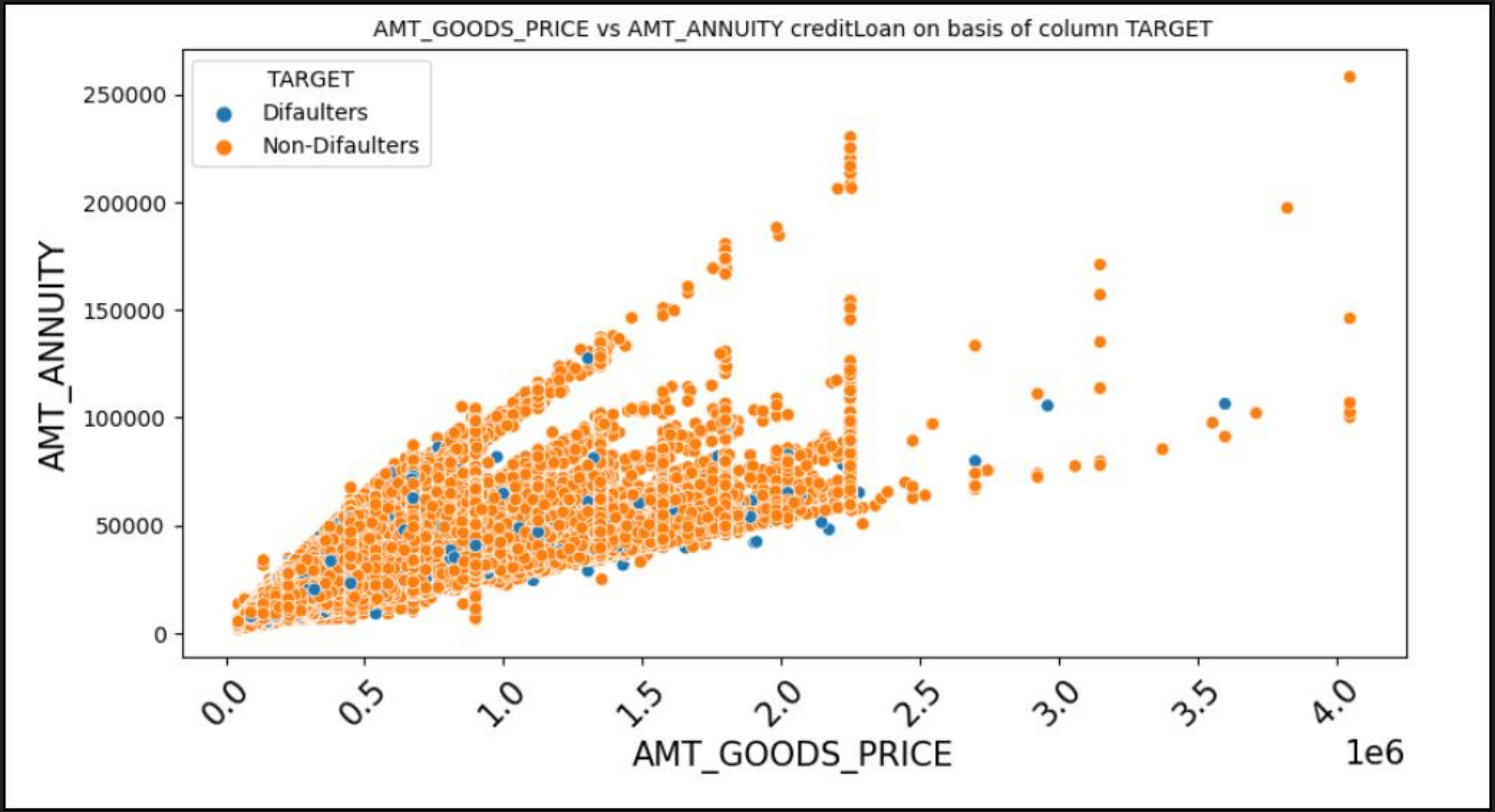
AMT_CREDIT vs AMT_ANNUITY :

- From below graph we can say that both the variable are highly correlated with each other. Also showing us a positive correlation.



AMT_ANNUITY vs AMT_GOODS_PRICE :

- From below graph we can say that both the variable are highly correlated with each other. Also showing us a positive correlation.



Multivariate Data Analysis

What is Multivariate Analysis?

It gives us the result of relationship between all the variables present in a data frame. We perform it with heatmap mostly.

What is correlation?

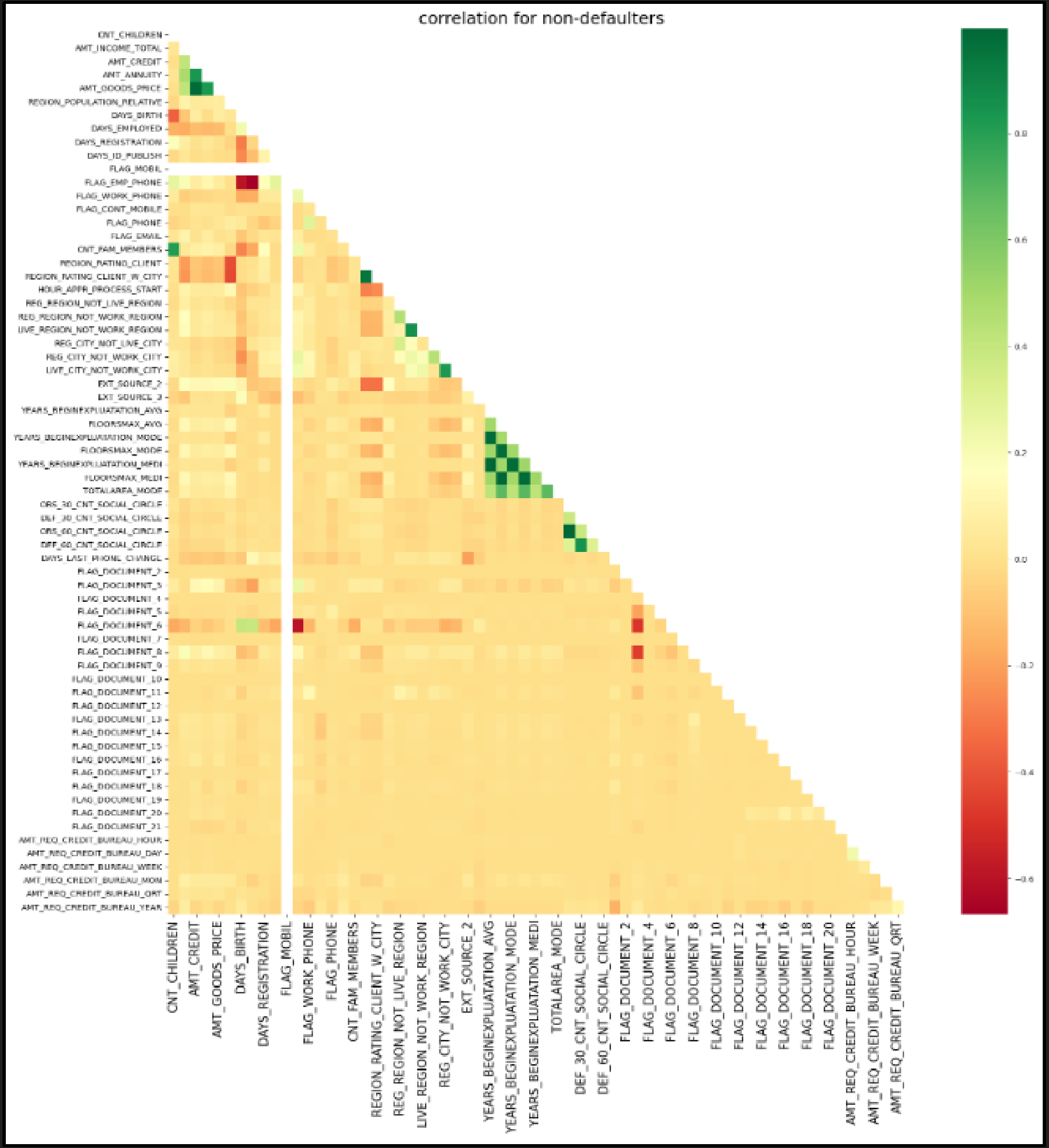
correlation defines us the relationship between minimum 2 or more than 2 variables. this is 3 types:

1. positive correlation = it shows us a positive relationship between 2 variants let's say if variable A increases then variable B will also increase.
2. Negative correlation = it shows us a negative relationship between 2 variants let's say if variable A increases then variable B will also decrease.
3. No correlation = where there is no relationship between 2 diagrams and points are present in the graph without any trend.

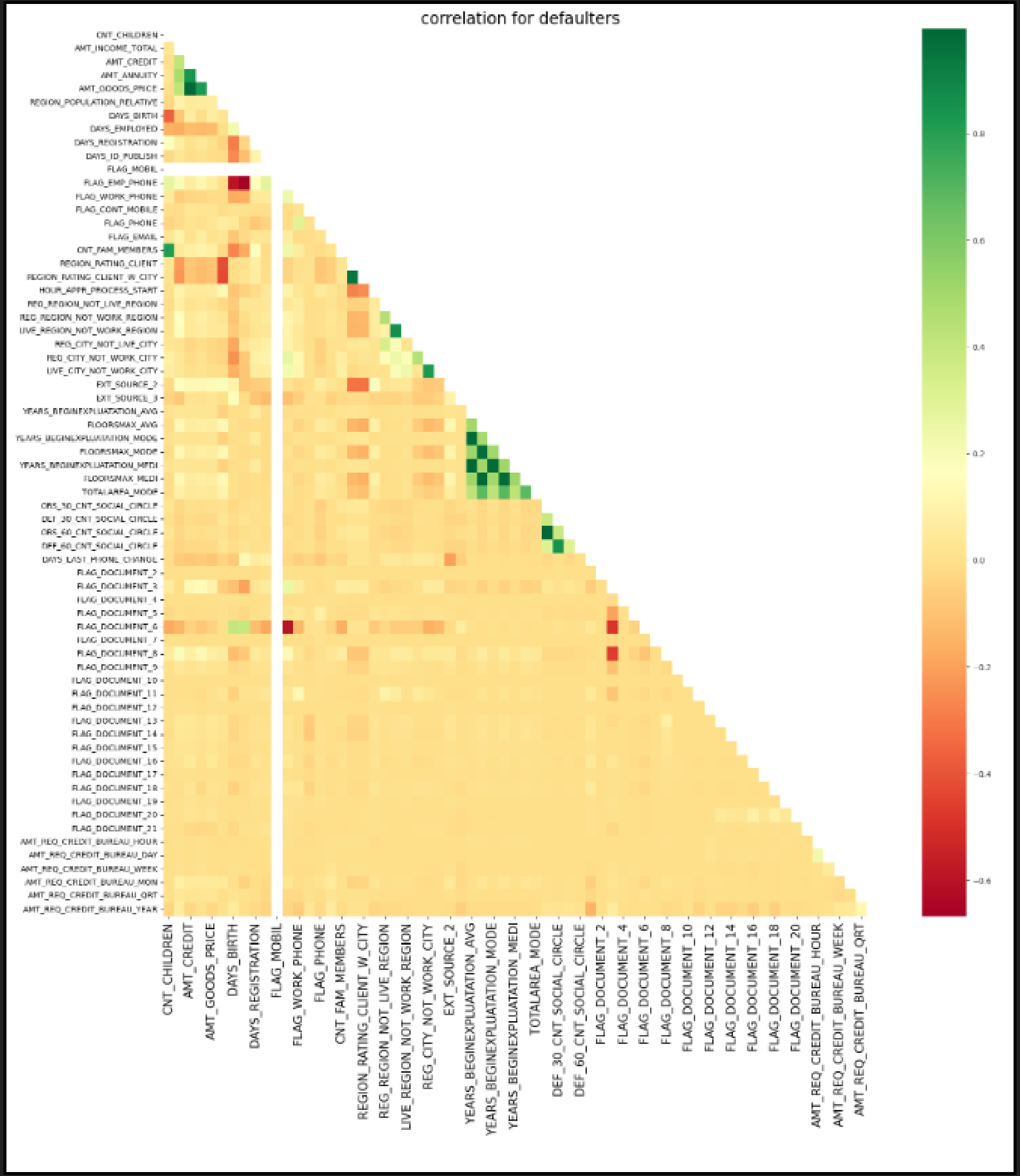
Check Next Slide ⟶

Correlation For Both Defaulters and Non-Defaulters :

- This gives us a structure of negative correlation for both Defaulters and Non-Defaulters.
- It showed negative correlation between AMT_CREDIT and DAYS_BIRTH.
- Also it is cleared that income total is inversely proportional to each other.
- Graphs are given in next slide.



Heatmap for Non-Defaulters



Heatmap for Defaulters

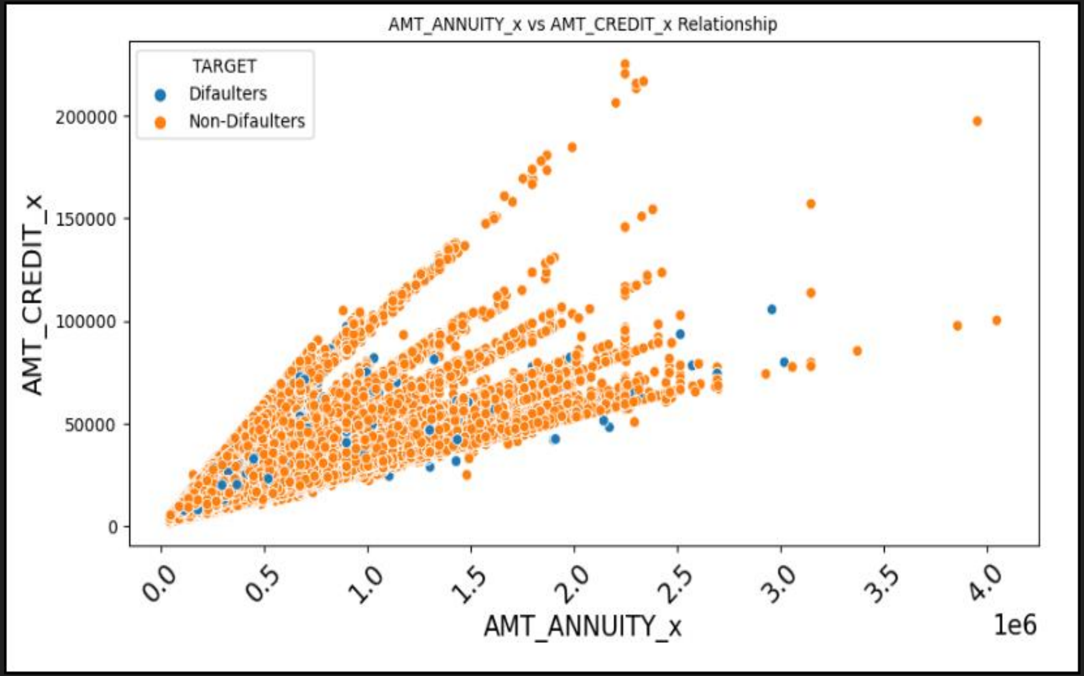
Merged Data Analysis (df1 and df2) :

Check Next Slide

We didn't notice much difference between merged data and individual data. Only 2 main Variables are analysed here.

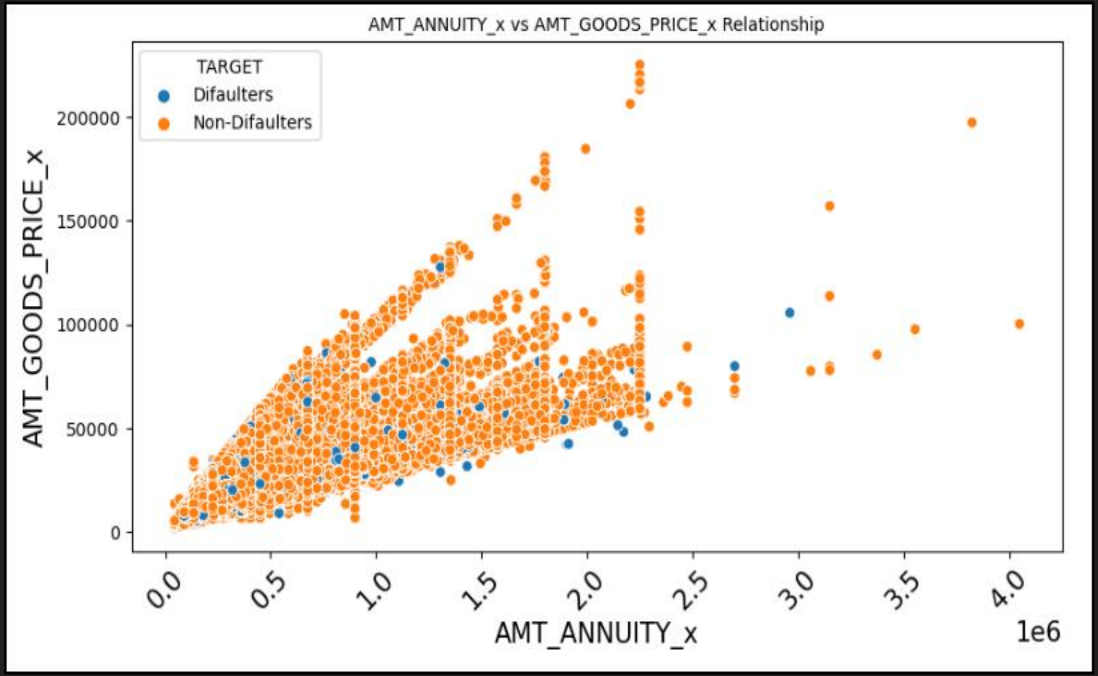
AMT_ANNUITY_x vs AMT_CREDIT_x

- These 2 variables are highly correlated to each other just like Application_Data



AMT_ANNUITY_x vs AMT_GOODS_PRICE_x

- Same for these also. Both are highly correlated.

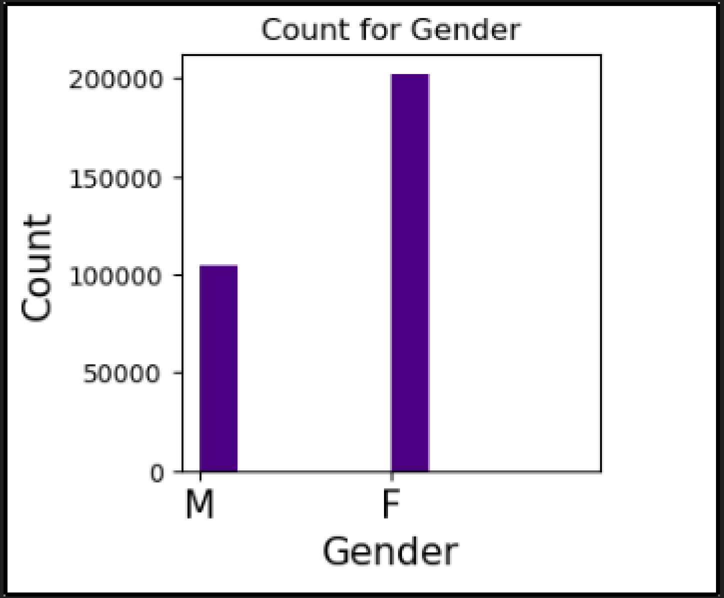


CONCLUSION / INSIGHTS :

Check Next Slide ⟶

- Data is highly imbalanced here because 91.9% population are payment non-defaulters and 8.1% are payment defaulters. So, we can say that even though majority of population are non-defaulters in the data but if we look for accuracy then the model is performing poorly. In this case we need to focus on statistics choices even more.

- Most of the loan appliers are female which is almost 2 Lakhs.



- Also observed that more number f payment non-defaulters are either not having any children or having 1 child.

- Most of the loan defaulters are having loan annuity in range of 20,000 to 50,000 because distribution for non-defaulters is very high there.

- People who are having income range in medium INCOME_TOTAL_GROUP are most likely to be a non-defaulters and population for taking loan is higher for this particular income group.

- Most of the defaulters as well as non-defaulters are taken cash loans.
- People who are older than 40 years are most likely to be a non-defaulters.
- AMT_CREDIT & AMT_ANNUITY are highly correlated to each other which is acceptable. Because if loan amount increases then EMI will also increase.
- AMT_GOODS_PRICE & AMT_ANNUITY are also highly correlated to each other. If, asset's price if high then loan amount will be high so automatically EMI will also be high.
- From correlation we can see it is giving us a negative relationship with most of the variables for both defaulters and non defaulters. Like young people are taken loan in high amount of money which is an example of negative correlation. Because, one variable is decreasing while another one is increasing.
- Merged data analysis doesn't show much difference from application_data.csv, both are almost similar.

Thank You.