BANK LOAN CASE STUDY

Project Description:

The aim of the project is to get an idea of of applying EDA in a real business scenario. In this case study, apart from applying the techniques that we have learnt in the EDA module, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers

In this project ,

1. Clean the raw data provided using various cleaning techniques
2. Perform analysis and visualization to get valuable insights

Approach :

The main approach towards this project is to first understand the dataset provided. Then using various cleaning techniques we can clean the data and get rid of duplicates to prepare data for analysis stage .We perform imputation techniques like mean ,mode and average to fill the null values . For analysis depending upon the insights required we can use various charts , pivot tables ,functions ,etc . We can use "Why's" technique to get to root of the problem and reach the desired solution . At the end we will display the insights extracted using various tables and charts to make it more easy to quickly understand the insights.

Tech used : The main software used during the project is Microsoft Excel

Insights :

A) Approach of analysis:

The banks problem statement is to identify patterns which indicate reasons for client having difficulty in paying the loan back . The insights gained from this will be used for risk assessment by the company. We are provided with the dataset that contains :

1. **application_data.csv`** contains all the information of the client at the time of application.
   The data is about wheather a client has payment difficulties.
2. **`previous_application.csv`** contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
3. **`columns_descrption.csv`** is data dictionary which describes the meaning of the variables.

B) Finding the missing data :

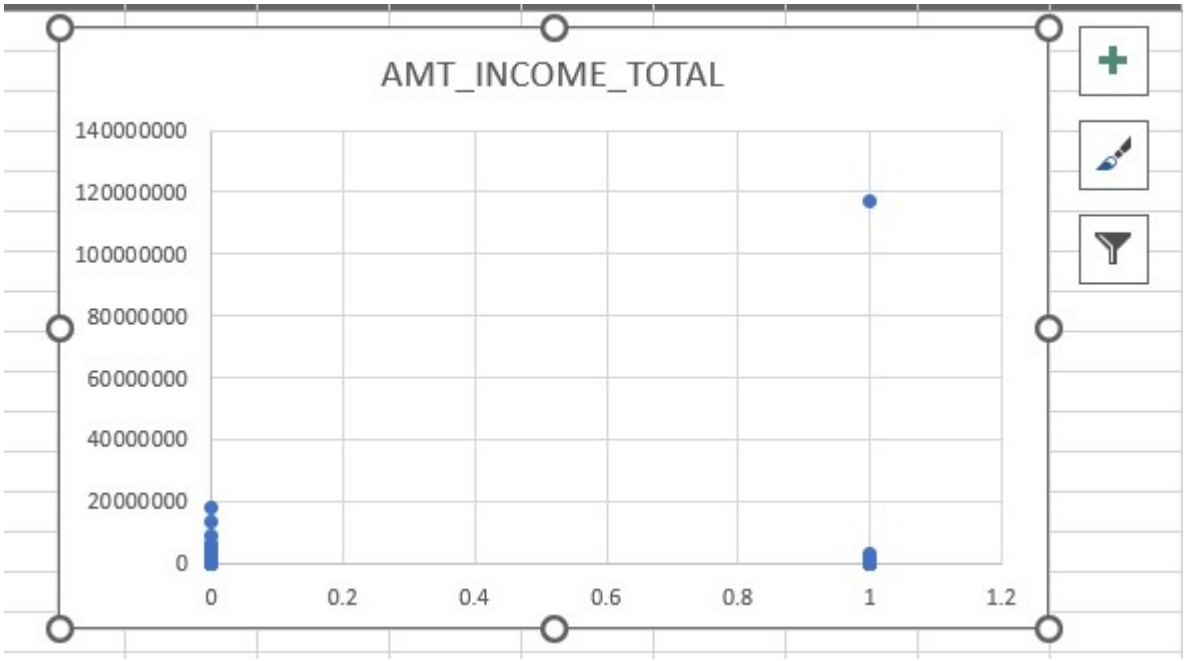
We clean the data from application.csv by deleting duplicates. Then calculate the percentage of empty fields in each column . We delete the columns having >30% of empty fields . Columns which have <30% empty fileds we fill them using mode , mean and average imputation techniques and prepare the data for further analysis

C) Identifying the outliers:

Outlier analysis is based on following values :

- AMT_INCOME_TOTAL
- AMT_GOOS_PRICE
- AMT_CREDIT
- DAYS_EMPLOYED

AMT_INCOME_TOTAL Outlier analysis:

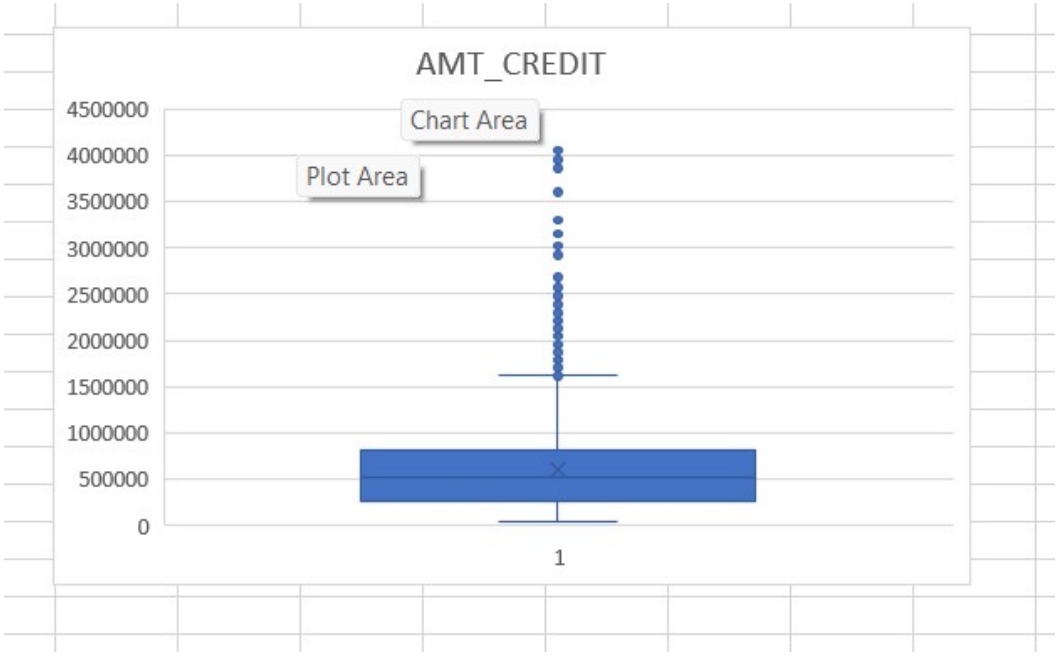


We use XY scatter plot for outlier analysis and we find out that majority of people(Target 1&0) have income less than 20000K . There is one exception who has income >20000K

AMT_GOODS_PRICE Outlier analysis:

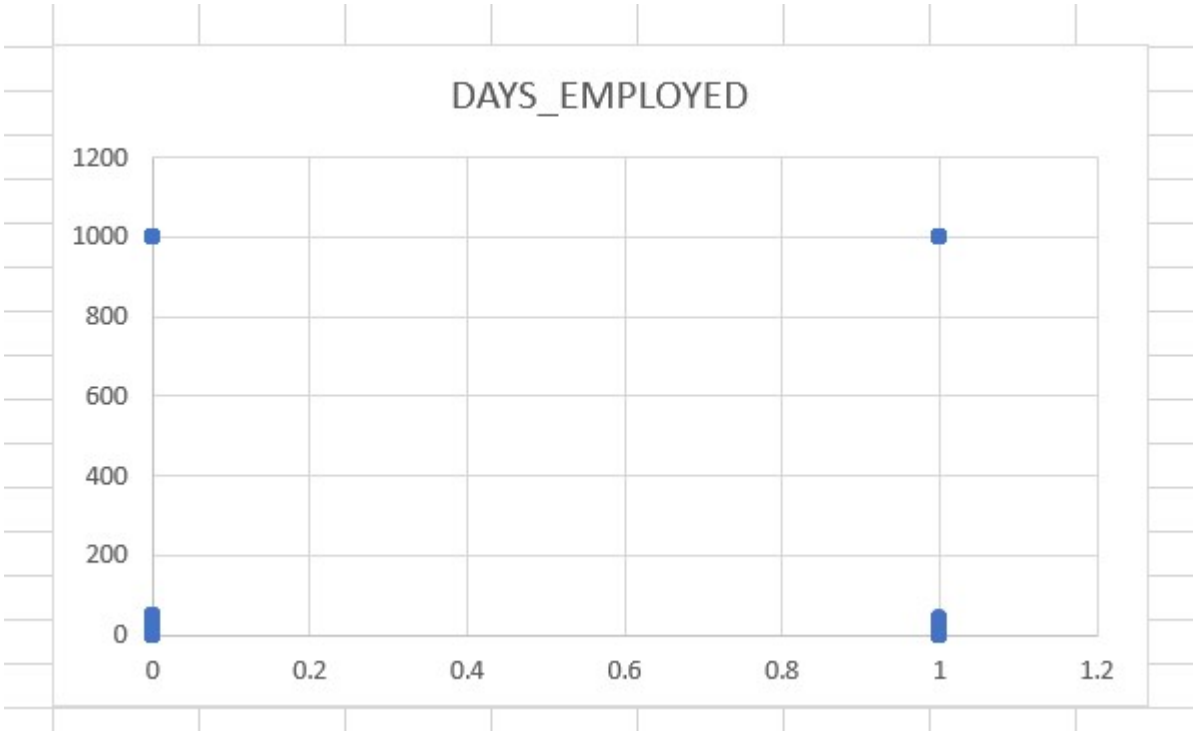
We use boxplot in this to identify outliers and we found a couple of outliers . In box plot points beyond the blue line are considered as outliers which in this data the blue line is just below 1500K . All the values above the blue line are called outliers

AMT_CREDIT



We use boxplot again for outlier analysis . As explained above all the points beyond 75% quartile range are considered as outliers .

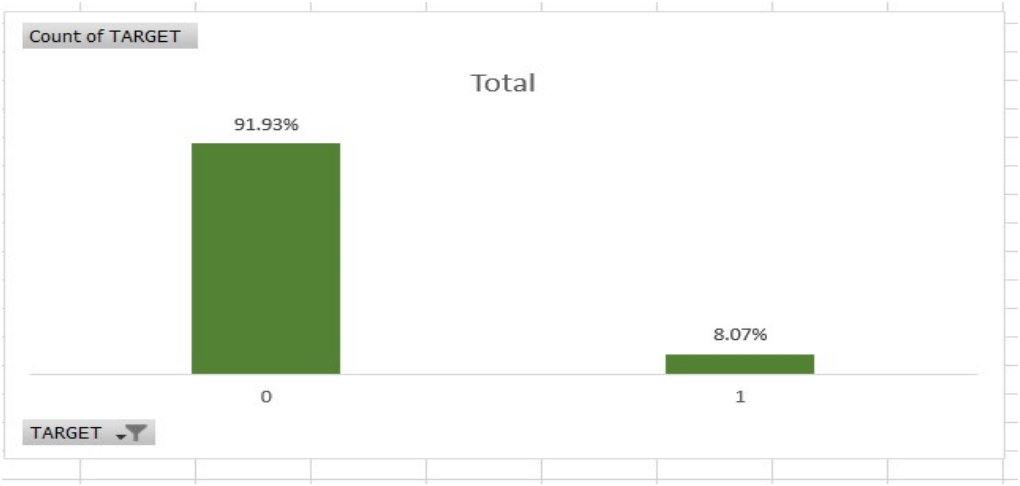
DAYS_EMPLOYED



In this we use XY Scatter plot for outlier analysis and find out that majority of days people(Target 1 & 0) are less than 200 but there were 2 values where days of employment is 1000days which is impossible so we consider those two values to be outliers in the data .

D) Data Imbalance :

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations. We can better understand imbalanced dataset handling with an example.



There is imbalance data in dataset in target column where people marked with Target 0 are very high as compared to people marked with Target 1 . Ratio is 92:8

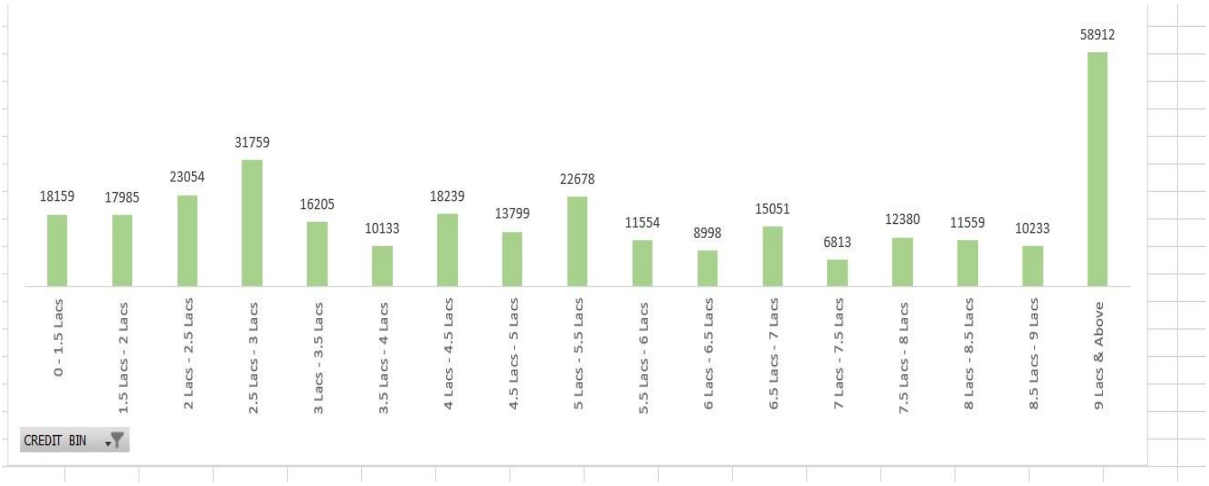
There can be data imbalance in other columns i.e males and females in data but we found out that the difference wasn't much higher so we discarded that set .

E) Types of analysis:

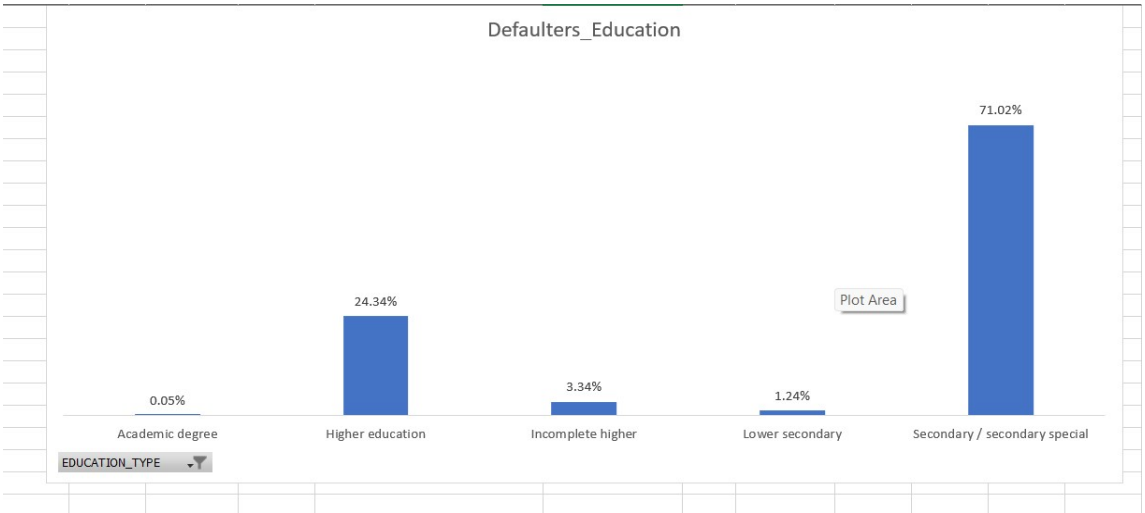
- **Univariate analysis**
- **Segmented univariate analysis**
- **Bivariate analysis**

Univariate analysis :

The term univariate analysis refers to the analysis of one variable. The purpose of univariate analysis is to understand the distribution of values for a single variable.



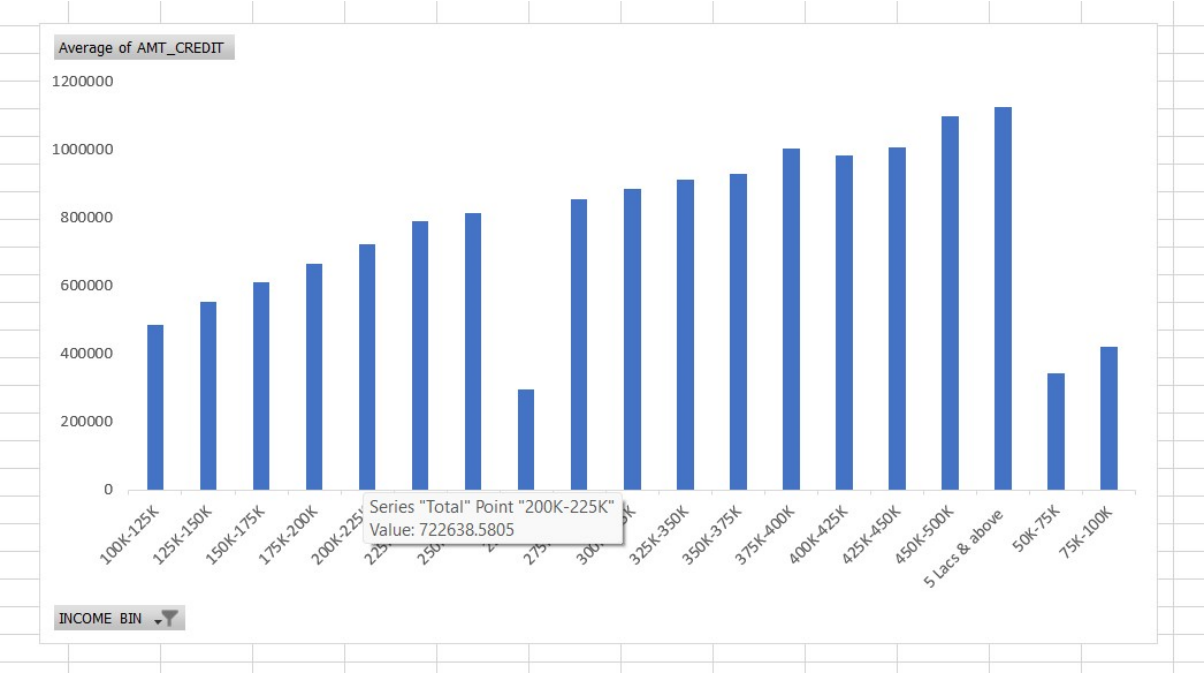
In the above graph univariate analysis is done for credit bin column for both Taregt 1 and 0 people . The conclusion from above analysis is that majority of people whose loan get approved have credit more than 9 Lacs and above .



Above graph is another example of univariate analysis which suggests that majority of people who are not able to pay loans(Defaulters) on time are only educated till secondary .

Bivariate analysis:

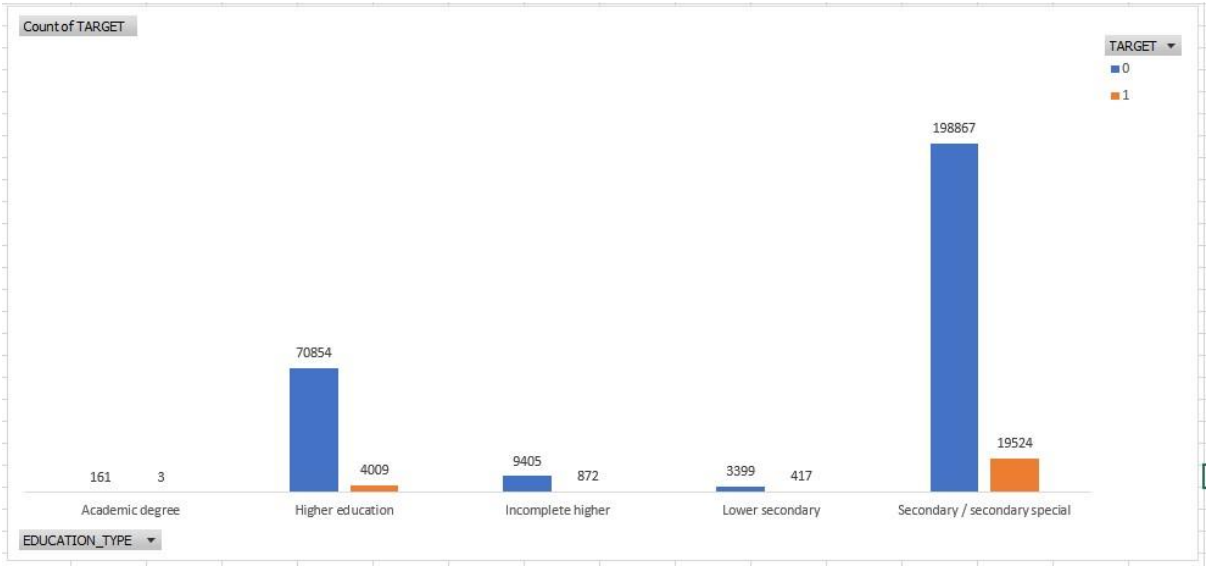
Bivariate analysis is a statistical method examining how two different things are related. The bivariate analysis aims to determine if there is a relationship between the two variables and, if so, how strong and in which direction that link is.



In the above graph bivariate analysis is done with Income Bin and AMT_CREDIT . It shows that both the income bin and AMT_CREDIT are directly proportional to each other . From above bivariate analysis we conclude that as income increases amt_credit also increases .

Segemented Univariate Analysis :

One of the simplest types of data visualisation is segmented univariate analysis. Uni means one, which implies that it only takes into account one data variable when conducting analysis. analysis. Segmented analysis, which refers to the analysis of the data variable in subsets, is particularly beneficial since it may display the pattern of change metric over the many segments of the same variable.



In the graph above we perform segmented univariate analysis that shows education of most of the applicants (Target 0 & 1). We conclude that very few there are very few target 1 applicants which don't pay loans on time despite having an academic degree .

F) Correlation :

Correlation defines the strength of relationship between two variables. Corelation is important in data analysis because it can help discover meaningful realtionships between different variables

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH (Years)	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.027	0.003	-0.024	-0.337	-0.245	0.029	0.023
AMT_INCOME_TOTAL	0.027	1	0.343	0.168	-0.063	-0.140	-0.023	-0.187
AMT_CREDIT	0.003	0.343	1	0.101	0.047	-0.070	0.001	-0.103
REGION_POPULATION_RELATIVE	-0.024	0.168	0.101	1	0.025	-0.007	0.001	-0.539
DAYS_BIRTH (Years)	-0.337	-0.063	0.047	0.025	1	0.626	0.271	-0.002
DAYS_EMPLOYED (Years)	-0.245	-0.140	-0.070	-0.007	0.626	1	0.277	0.038
DAYS_ID_PUBLISH (Years)	0.029	-0.023	0.001	0.001	0.271	0.277	1	0.009
REGION_RATING_CLIENT	0.023	-0.187	-0.103	-0.539	-0.002	0.038	0.01	1

Color scheme used is Green and White . As green color becomes darker co realtion increases. Dark green indicates strongest corealtion and white indicates weakest co realtion. The above picture shows co realtion in target 0 data .

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH (Years)	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.005	-0.002	-0.032	-0.259	-0.193	0.032	0.041
AMT_INCOME_TOTAL	0.005	1	0.038	0.009	-0.003	-0.015	0.004	-0.021
AMT_CREDIT	-0.002	0.038	1	0.069	0.135	0.002	0.052	-0.059
REGION_POPULATION_RELATIVE	-0.032	0.009	0.069	1	0.048	0.016	0.016	-0.443
DAYS_BIRTH (Years)	-0.259	-0.003	0.135	0.048	1	0.582	0.253	-0.034
DAYS_EMPLOYED (Years)	-0.193	-0.015	0.002	0.016	0.582	1	0.229	0.003
DAYS_ID_PUBLISH (Years)	0.032	0.004	0.052	0.016	0.253	0.229	1	-0.001
REGION_RATING_CLIENT	0.041	-0.021	-0.059	-0.443	-0.034	0.003	-0.001	1

Color scheme used is Green and White . Dark green indicates strongest link whereas white color indicates weakest link . The above picture shows co relation in target 1 data .

G) Most Important results :

- Majority of applicants applying for loans have education only till secondary
- There are very few applicants who have higher degree and are still not paying loans on time
- Applicants with higher income have higher credit score
- There are very few people who don’t pay loans on time as compared to people paying loans on time

Dataset attached :

https://drive.google.com/file/d/155b1MezEQsyO0QhYJ_t1C12yXDJ68RrN/view?usp=sharing

Results :

During this project I have learned how a data analyst works in real life . I also got experience to handle vast amounts of data and perform various risk analysis which in turn would help me become a better analyst in future.