**INTERNSHIP – DATA ANALYTICS**

PROJECT 4

**LOAN APPLICATION DATA ANALYSIS**

**SUBMITTED BY**

HRISHIKESH KALITA (EI0123)

COLLEGE

**NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA**

# **GOAL**

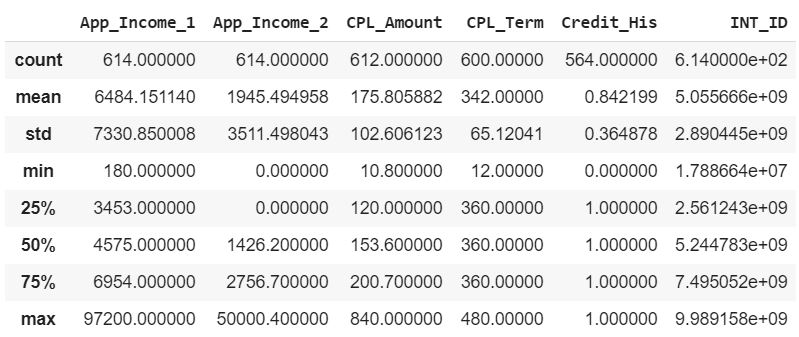
* Checking the Customer’s eligibility to get approval for the loan applied.

# **DATA STATS**

* Shape of the Train Data: (614, 20)
* Columns of the Data:

'Loanapp\_ID', 'Sex', 'Marital\_Status', 'first\_name', 'last\_name', 'email', ‘address', 'Dependents', 'Qual\_var', 'SE', 'App\_Income\_1', 'App\_Income\_2', 'CPL\_Amount', 'CPL\_Term', 'Credit\_His', 'Prop\_Area', 'INT\_ID', 'Prev\_ID', 'AGT\_ID', 'CPL\_Status'

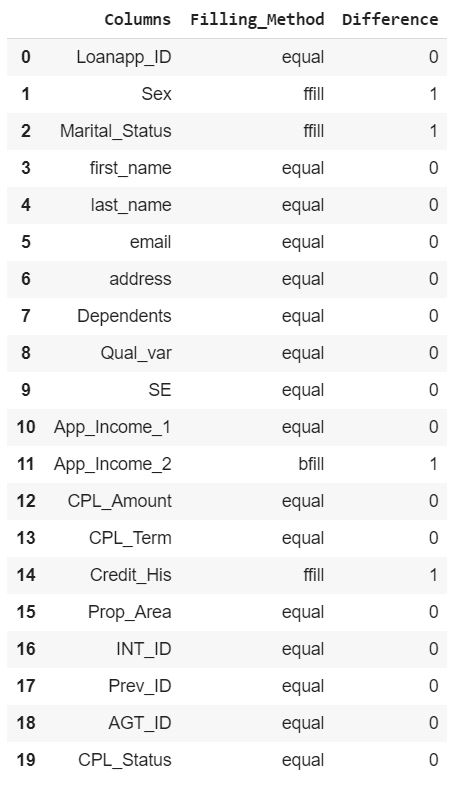
* Data info:



**COMMENTS**

* The discrepancies in the ‘count’ row suggest the presence of null/Nan values in the respective columns.
* The ‘Dependents’ column should have been a numerical column but it is not present in the above table because of several ‘3+’ entries in it which eventually makes it an ‘object’ type column. This was fixed by considering all ‘3+’ as ‘3’ and converting the column into a ‘float’ type.

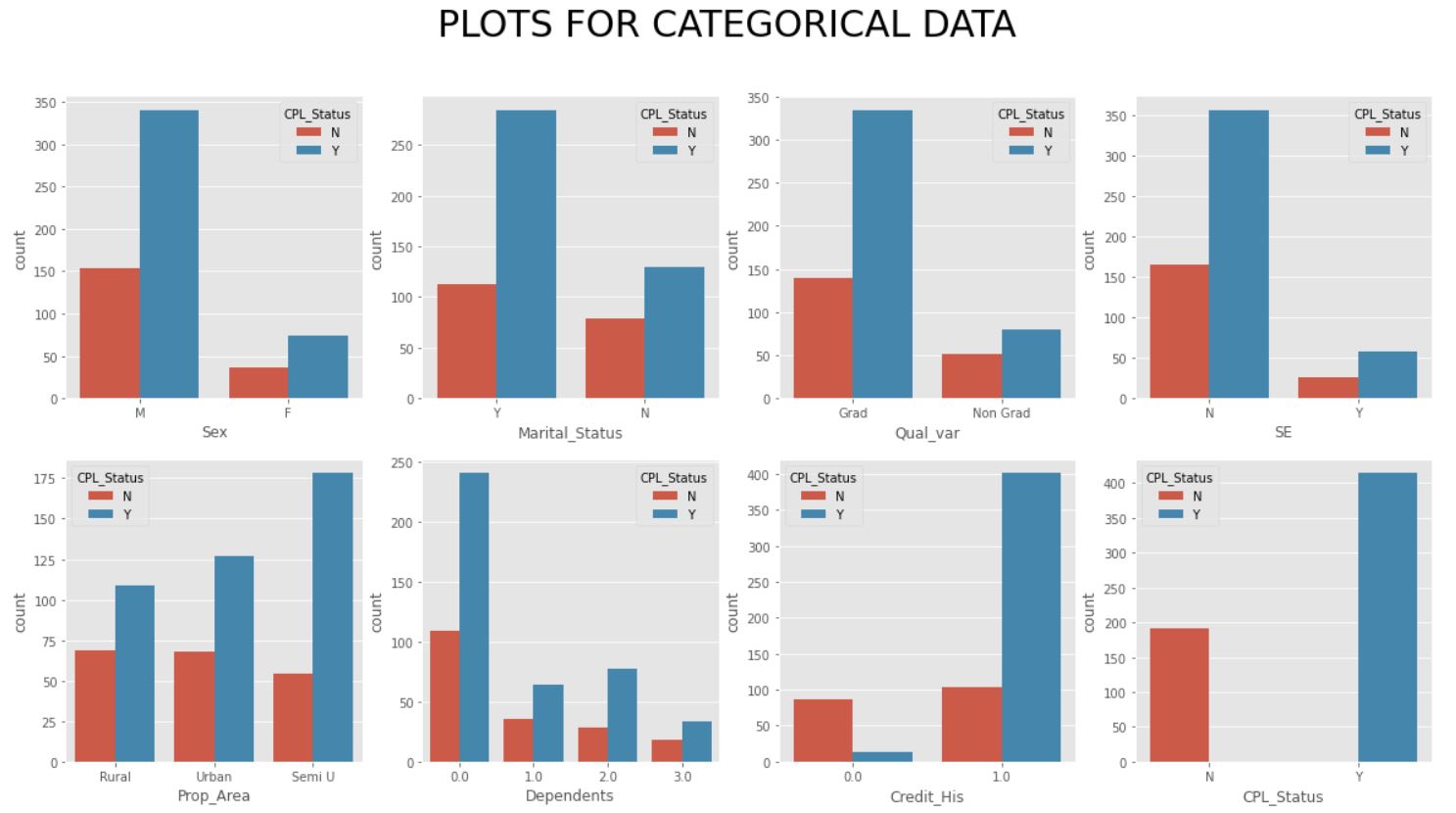
# **DEALING WITH NULL/NAN VALUES**

* A user-based function was used to find the trend in the data regarding the ‘fill’ method. It was coded with the intention of getting an overview of the fact that whether ‘**ffill’** or ‘**bfill’** should be used in Pandas ‘fillna’ function.

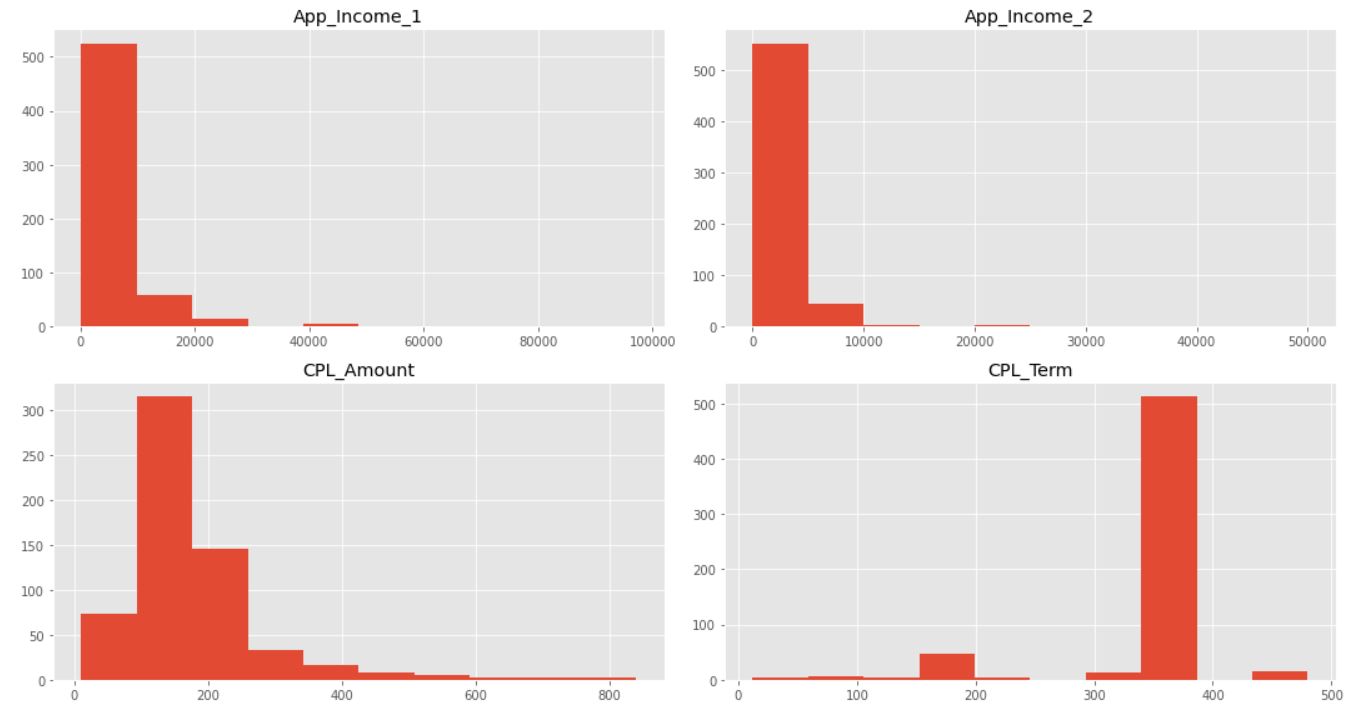
**COMMENTS**

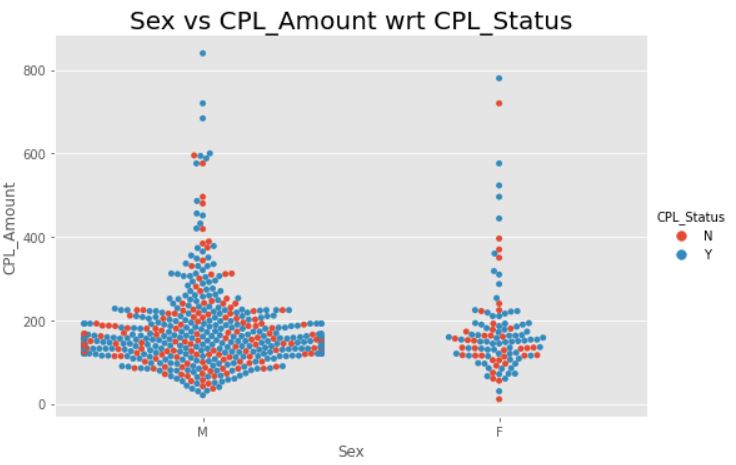
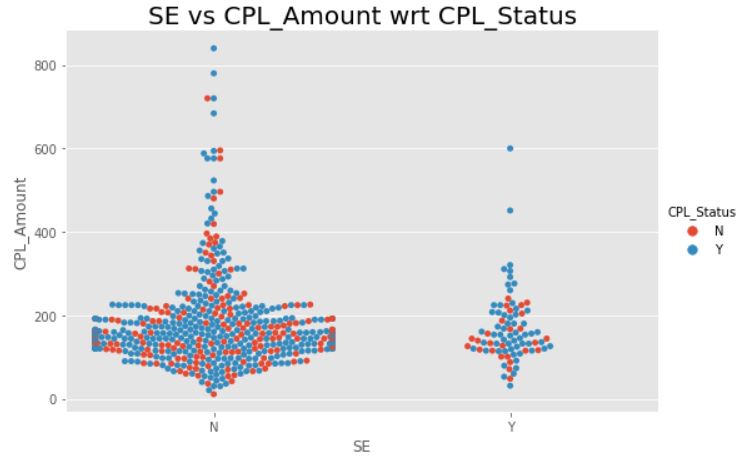
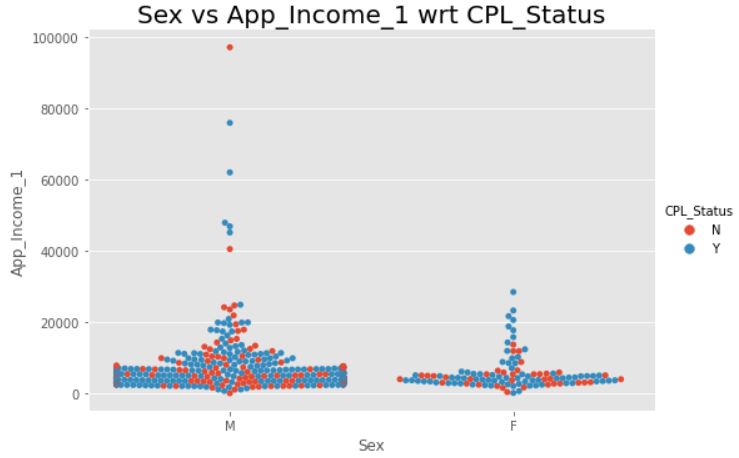
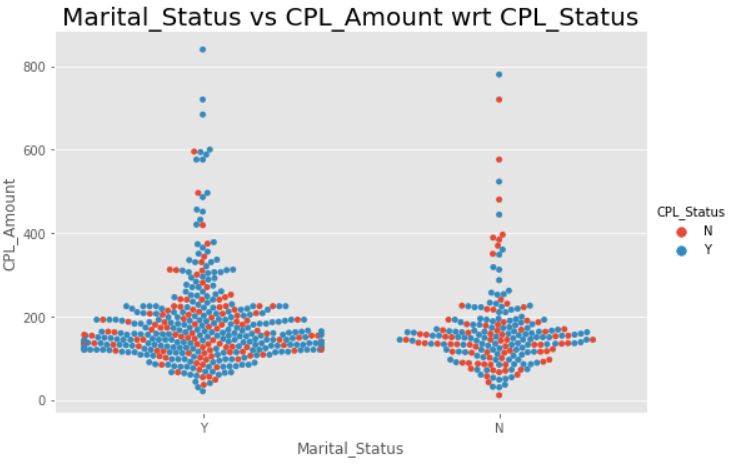
However, the table obtained shows almost similar patterns for all the columns. Only 3 out of all columns had ‘ffill’ majority with a difference of 1, which could be ignored. However, eventually ‘**ffill’** was used to some extent on the training data and the remaining rows with **NaN** values were **removed**. (Train Data Shape: (605, 20) after dealing with missing values.)

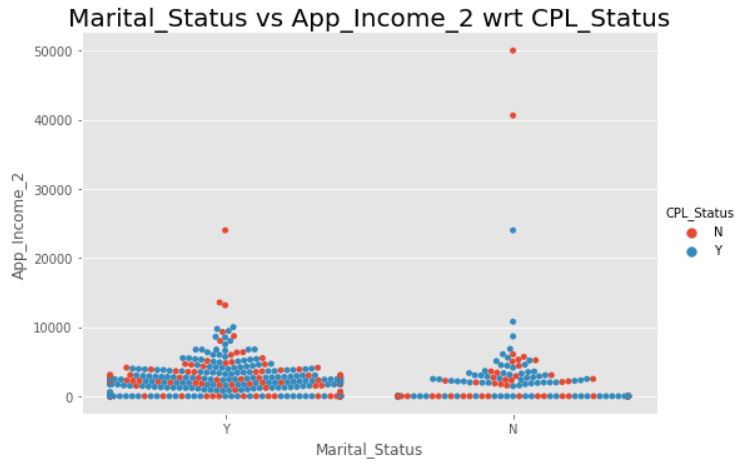
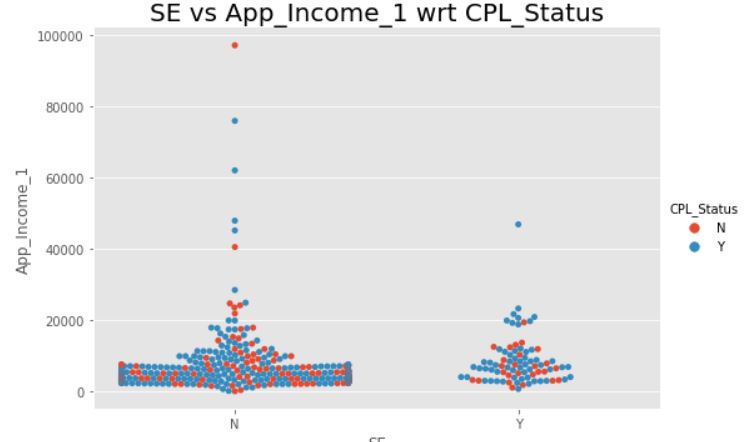
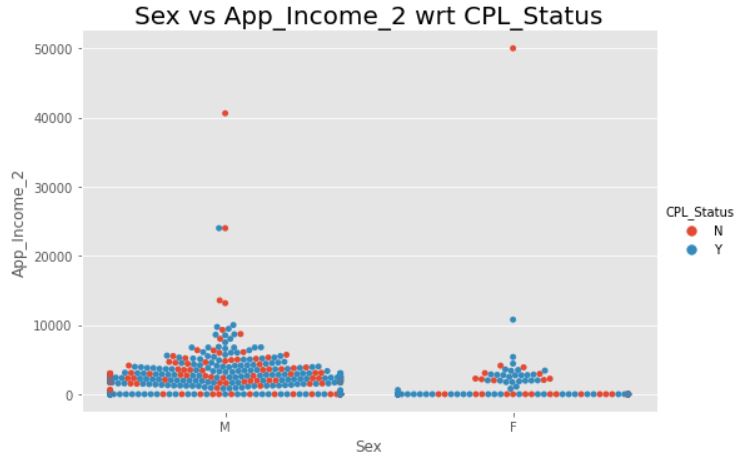
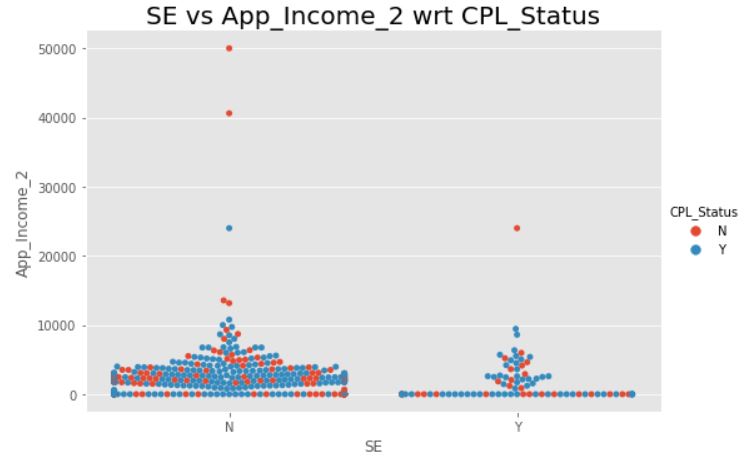
# **EXPLORATORY DATA ANALYSIS (EDA)**

* **CATEGORICAL**

**COMMENTS**

* A general trend of **lesser** ‘**Female’** applicants was observed.
* A **significant** number of **approvals** can be observed in **Semi-Urban** areas.
* In most of the cases the number of **rejected** loan applications is lesser than that of the **approved**. However, in case of applicants **without Credit History**, it is the **opposite**. There is a **greater** number of **rejections**.
* From the data observed above, it may be assumed that a **married, graduated** and **non-self-employed male** with a valid **credit history**, residing in a **semi-urban** area **without any dependents** stand a **higher chance** of being **approved** for a loan.
* **CONTINUOUS**
* **OVERALL**

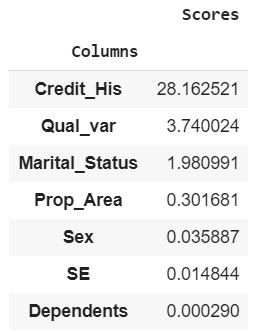
****



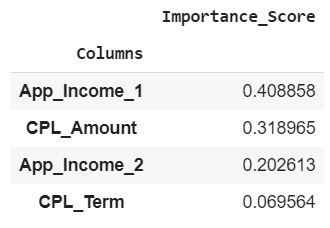
**COMMENTS**

* From the figures above, relationship among various entities can be studied in detail.
* Presence of probable **outliers** can be seen in almost all the diagrams.
* From the visualizations, one can suggest that the dataset **doesn’t contain** **many** people with **very high income** or **high loan amount**.
* Most ‘swarms’ were crowded in the **lower income and lower loan amount** regions.

# **FEATURE IMPORTANCE**



Obtained by using **SelectKBest** feature from **Scikit-Learn**. **Chi2** method was used to draw out the feature importance in deciding the output in the model.



Obtained using **feature importanc**e method in **Random Forest Classifier**.

# **MODELLING**

MODELS USED:

* RANDOM FOREST CLASSIFIER
* DECISION TREE CLASSIFIER
* SUPPORT VECTOR MACHINES
* LOGISTIC REGRESSION
* ADABOOST CLASSIFIER

**CROSS-VALIDATION** was used with **cv = 10** and the **Standard Deviations** among the scores were very low implying the **stability** of the models.

# **PREDICTION OF TEST DATA**

* The missing values were first imputed using Pandas **fillna** function (‘**ffill’**).
* For getting a better judgement, a **user-defined** function was used to **get the predictions from the models** and decide the **final prediction according to the majority rule**.

For example, if 3 of the models used gives the prediction as “**Y**” (Approved), and other two as “**N**” (Rejected), then “**Y**” (Approved) has a 60% majority and hence the respective application is predicted to be ***approved***.

* The final predictions were attached to the test data csv file as a new column and saved as **Test\_Data\_Predictions.csv**