1.0.0.0. **INTRODUCTION**

1.1.0.0. Setting the context

1.1.1.0. This report is to be read in conjunction with the Google Colab notebook for EDA & Feature Analysis on the Home Credit loan defaulter prediction problem hosted on Kaggle here.

1.1.2.0. An understanding of the complete problem context and high level summary of the datasets used can be sought from here.

1.2.0.0. A quick refresher about Home Credit's motivation for predicting potential defaulters

1.2.1.0. Though there are a lot of people seeking loans from banks and lending institutions, only a few of them get approved. This is primarily because of insufficient or non-existent credit histories of the applicant. Such a population is taken advantage of by untrustworthy lenders. In order to make sure that these applicants have a positive loan taking experience, Home Credit uses Data Analytics to predict the applicants' loan repayment abilities, trying to ensure that the clients capable of loan repayment do not have their applications rejected.

1.3.0.0. EDA and Feature Analysis - What is it about?

1.3.1.0. Exploratory Data Analysis [EDA] is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the actual modeling task.

1.3.2.0. EDA is primarily making sense of data at hand, before getting them dirty with it.

1.3.3.0. Occam's razor, which summarizes that of two competing theories, the simpler explanation of an entity is to be preferred; forms the basis for feature analysis.

1.3.4.0. Feature engineering refers to a process of selecting and transforming variables when creating a predictive model using machine learning or statistical modeling. The process involves a combination of data analysis, applying rules of thumb and domain-based knowledge.

1.4.0.0. Objectives of performing EDA and Feature Analysis on the Home Credit dataset

1.4.1.0. Understanding the given data sets individually and interactions among them.

1.4.2.0. Gaining insights regarding the features by variable-level visualizations and their relations with the 'Target' outcome.

1.4.3.0. Listing missing values and devising a strategy for filling these out logically.

1.4.4.0. Identifying outliers and rationally addressing them.

2.0.0.0. **STEPWISE OVERVIEW OF THE ENTIRE EDA PROCESS AND CORRELATION WITH THE COLAB NOTEBOOK**

2.1.0.0. The following are the major phases of the EDA as carried out in the Colab notebook -

2.2.0.0. Section 1.0.0 - A brief summary of the project, dataset and intent of this notebook.

2.3.0.0. Section 2.0.0 - Contains the necessary groundwork for carrying out the EDA.

2.3.1.1. Section 2.1.0 comprises the code for installing dependencies for some of the specialised libraries used in this notebook.

2.3.1.2. Section 2.2.0 has the list of all the libraries loaded for the purpose of EDA in this notebook along with a context for each library.
2.3.1.3. Section 2.3.0 lists the custom functions defined for the utilities which shall be frequently used or a feature which is important for the purpose of EDA.
2.3.1.4. Section 2.4.0 loads the available datasets for EDA and further analysis.

2.4.0.0. Section 3.0.0 - Comprises of dataset-level analysis
2.4.1.1. Section 3.1.0 presents a comprehensive summary for each dataset loaded. This helps form an action plan regarding EDA and downstream feature-engineering strategies.
2.4.1.2. Section 3.2.0 gives an elaborate quantification of interactions between the main datasets for getting context of applicants in the sample population.
2.4.1.3. Section 3.3.0 summarizes the key insights of the dataset-level summary.

2.5.0.0. Section 4.0.0 - consists of feature-level univariate and multivariate analysis.
2.5.1.1. Section 4.1.0 contains visualizations for features which 'appear' to be important for defaulter prediction based on domain knowledge or common sense. Each visualization is followed by a summary for the same and some observations.
2.5.1.2. Section 4.2.0 has visualizations for interactions among the features or bi-variate feature plots. These help in viewing the relation between the features involved coupled with the TARGET variable and derive key insights.
2.5.1.3. Section 4.3.0 has correlation heatmaps among the features of each dataset along with the TARGET variable.
2.5.1.4. Purpose for doing this is to get a fair idea of the degree of correlation among features which will give some basis for selecting features for modelling or assessing features selected in case of model-based feature selection [done in section 9.0.0].
2.5.1.5. Each heatmap set is followed by insights derived from the same.

2.6.0.0. Section 5.0.0 - consists of feature engineering.
2.6.1.1. Section 5.1.0 has some extracted features as well as the description for the same. Creation of these features is based on acquired domain knowledge and literature review.
2.6.1.2. Section 5.2.0 comprises the process of merging the main datasets. Rationale for merging the datasets is explained as is the process followed in the same.
2.6.1.3. Section 5.3.0 deals with encoding the categorical features and imputation of missing values. Strategy for the same is described in brief.
2.6.1.4. Section 5.4.0 involves splitting the processed dataset into train, validation and test dataset.
2.6.1.5. Based on future modeling insights, the validation set may be merged into the train dataset itself, opting for k-fold CV.

2.7.0.0. Section 6.0.0 - is a checkpoint where datasets processed thus far are saved or 'pickled' for further analysis, saving on RAM consumption.
2.7.1.1. This is done as running the notebook on Colab free edition hits the RAM limit, even after using optimised datasets with context to size.

2.8.0.0. Section 7.0.0 - loads the saved files from section 6.0.0 for further processing.
2.8.1.1. This is done in order to save on RAM consumption.

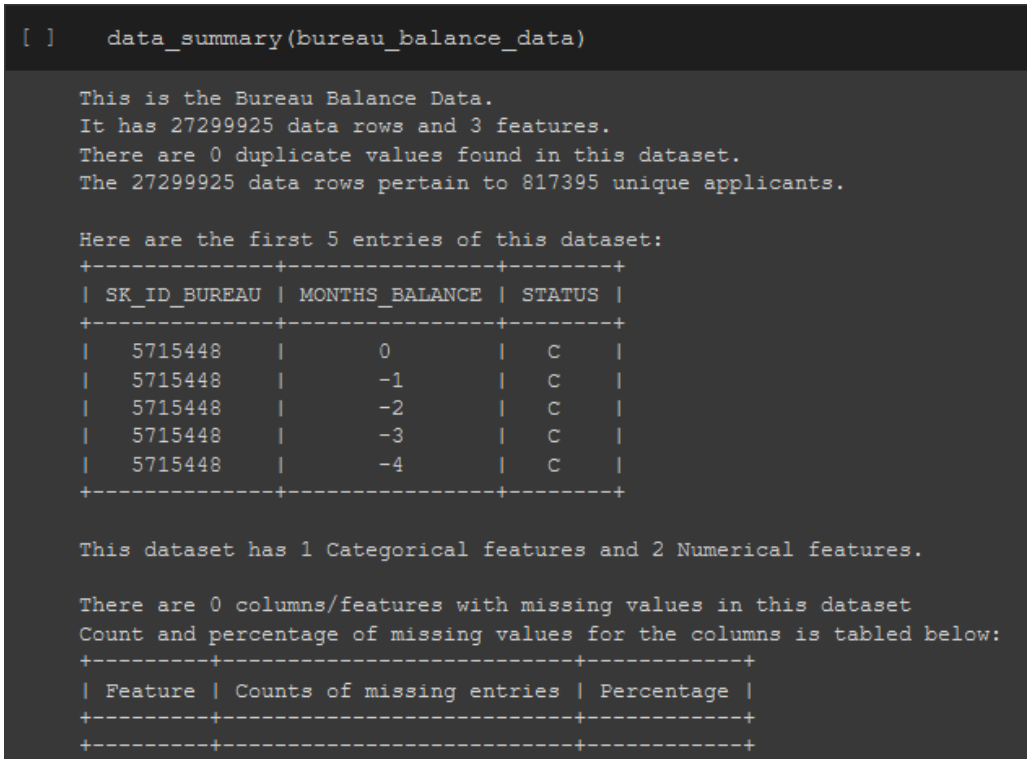
2.9.0.0. Section 8.0.0 - deals with removal of outliers.
2.9.1.1. Presence of outliers was determined during EDA and the same shall be addressed in this section.
2.9.1.2. Section 8.1.0 involves usage of the pyOD library to detect and remove the outliers.
2.9.1.3. Section 8.2.0 involves verifying whether the outlier removal was effective or not.

2.10.0.0. Section 9.0.0 - comprises feature selection.
2.10.1.1. As there are many features in the processed dataset and it is quite likely that some may very negligibly - or not at all, contribute towards defaulter prediction, a sort of selection of useful features needs to be done.
2.10.1.2. In this notebook, presently 2 models for feature selection are run and output of one of them is selected for further engineering.
2.10.1.3. In future phases, more complex combinations of feature selection strategies shall be applied based on modeling and outcomes.
2.10.1.4. Section 9.1.0 is about Extra tree Classifier from SKLearn for feature selection.
2.10.1.5. Output of this is considered for further processing in the notebook.
2.10.1.6. Section 9.2.0 is about Random Forest Regressor, again from SKLearn, for feature selection.
2.10.1.7. Almost all the features thought of as important towards Defaulter prediction during EDA do figure in the top selected features, as do the created or extracted features as listed in section 5.1.0
2.10.1.8. Section 9.3.0 has a correlation matrix heatmap for the top 25 features, just as a sanity check and for visualization.

2.11.0.0. Section 10.0.0 is about High level data visualization.
2.11.1.1. Section 10.1.0 and Section 10.2.0 deal with saving and loading the 'pickled' datasets, due to RAM constraints on Colab.
2.11.1.2. Section 10.3.0 is about the actual tSNE visualization for getting a 'feel' of the data separability.
2.11.1.3. Owing to the significant compute times for tSNE, only 2 combinations of parameters are evaluated and visualized.

2.12.0.0. Section 11.0.0 concludes this phase by summarizing key takeaways of the EDA and Feature Analysis.
2.12.1.1. Section 11.1.0 lists the general highlights of the EDA phase.
2.12.1.2. Section 11.2.0 summarizes the context-specific quirks and insights.

3.0.0.0. **SUMMARY OF EDA AND FEATURE ANALYSIS ON HOME CREDIT DATASET**
3.1.0.0. Dataset level analysis
3.1.1.0. Objective of the dataset-level analysis is understanding each dataset in context of types of data/features it contains, total data points and their uniqueness, proportion of missing values and finally, interaction among the main datasets.
3.1.2.0. All the datasets are loaded as Pandas DataFrame and a high level summary is tabulated for each.
3.1.3.0. A screengrab of the summary tabulation for one of the datasets [bureau balance dataset] is shown for representation.



3.1.4.0. The summary for each dataset includes listing total and unique number of entries, nature of the features, tabulation of features with missing values and a dataslice for visualizing the dataset.

3.1.5.0. Highlights of interactions among the main datasets is summarized below:

3.1.5.1. Of the 307511 unique applicants in application_train dataset, 263491 [85.69% of the total applicants] have at least one recorded instance of past non-Home Credit history in the Bureau dataset. Alternately, of the 305811 unique applicants in Bureau dataset, 42320 [13.84% of the total unique entries] are exclusive to the Bureau and have no presence in the application_train dataset.

3.1.5.2. Of the 48744 unique applicants in application_test dataset, 42320 [86.82% of the total applicants] have at least one recorded instance of past non-Home Credit history in the Bureau dataset. Alternately, of the 305811 unique applicants in Bureau dataset, 263491 [86.16% of the total unique entries] are exclusive to the Bureau and have no presence in the application_test dataset.

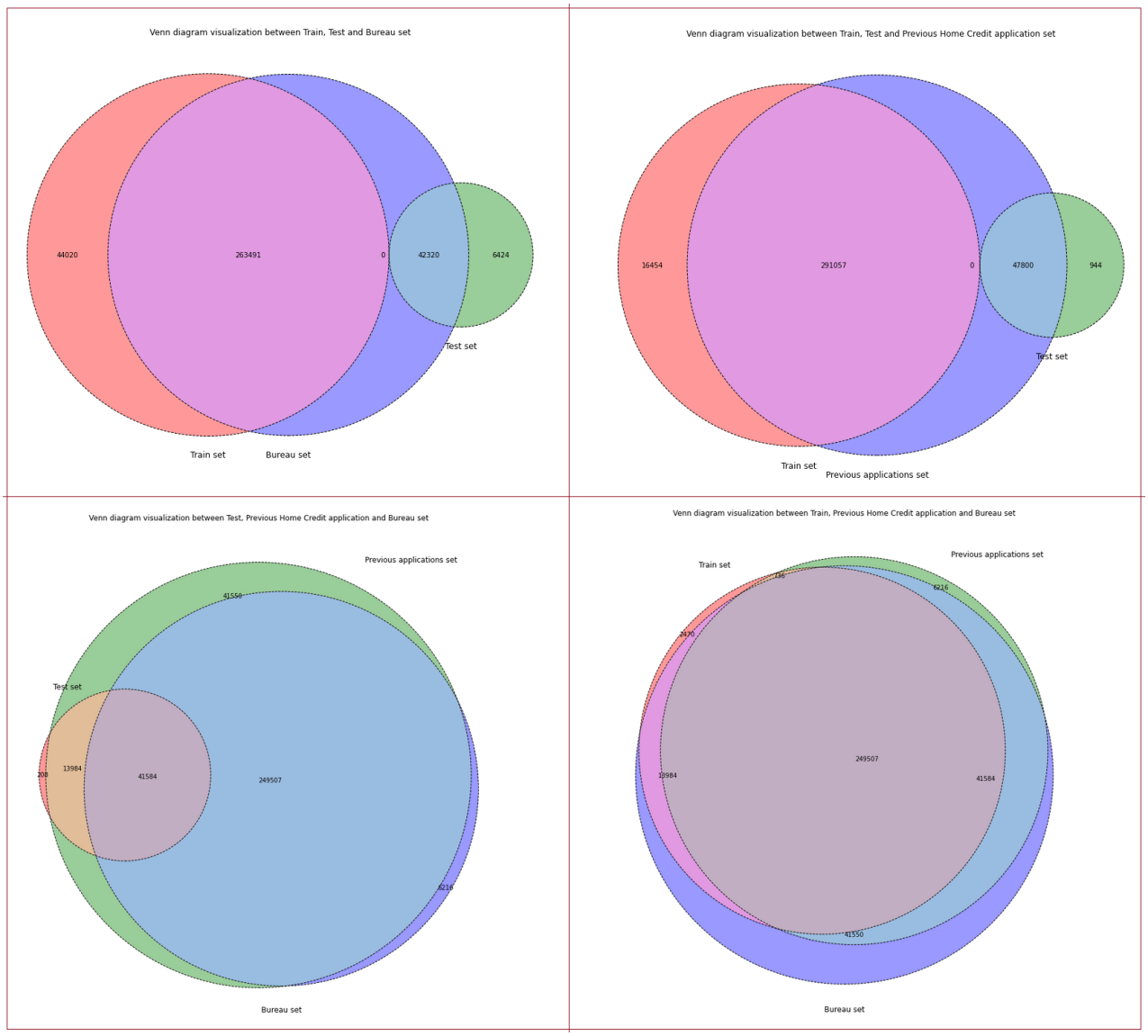
3.1.5.3. Of the 307511 unique applicants in application_train dataset, 291057 [94.65% of the total applicants] have at least one recorded instance of past Home Credit history in the previous applications dataset. Alternately, of the 338857 unique applicants in the previous applications dataset, 47800 [14.11% of the total unique entries] are exclusive to the previous applications and have no presence in the application_train dataset.

3.1.5.4. Of the 48744 unique applicants in application_test dataset, 47800 [98.06% of the total applicants] have at least one recorded instance of past Home Credit history in the previous applications dataset. Alternately, of the 338857 unique applicants in the previous applications dataset, 291057 [85.89% of the total unique entries] are exclusive to the previous applications and have no presence in the application_test dataset.

3.1.5.5. Of the 305811 unique applicants in bureau dataset, 291091 [95.19% of the total applicants] have at least one recorded instance of past Home Credit history in the previous applications dataset. Alternately, of the 338857 unique applicants in previous applications dataset, 47766 [14.1% of the total unique entries] are exclusive to the Home Credit previous applications and have no credit history with other agencies recorded in the bureau dataset.

3.1.5.6. Of the 305811 unique applicants in bureau dataset, 291091 [95.19% of the total applicants] have at least one recorded instance of past Home Credit history in the previous applications dataset. Alternately, of the 338857 unique applicants in previous applications dataset, 47766 [14.1% of the total unique entries] are exclusive to the Home Credit previous applications and have no credit history with other agencies recorded in the bureau dataset.

3.1.6.0. Venn Diagram visualization for the interactions among the main datasets



3.1.7.0. Key insights from dataset-level EDA and analysis

3.1.7.1. Around 86% of the training sample applicants are not first time credit seekers and have some credit history with lending agencies apart from Home Credit as recorded in the Bureau dataset.

3.1.7.2. Around 95% of the training sample applicants already have some credit history with Home Credit recorded in the previous applications dataset. Such a high number of repeat applicants might be indicative of customer's preference to Home Credit's lending processes and products over competition.

3.1.7.3. Barely 1% of the training sample applicants are first-time credit seekers from Home Credit with no recorded financial history in any agency.

3.1.7.4. As there is a very high number of applicants having previous loan records in the Bureau or Previous Home Credit database, these shall be used along with the Application Train dataset for modeling purposes.

3.1.7.5. There are many features/columns with missing values across the datasets with some having over 60% data missing. A suitable imputation strategy shall be employed unless it is evidenced by further analysis that dropping these features is a better strategy.

3.2.0.0. Feature-level univariate & multivariate analysis

3.2.1.0. Objective of the feature-level analysis is understanding each feature in context to its distribution, relation with the output or key variable, the values it takes, and possible anomalies.

3.2.2.0. To this effect, grouped bar charts, pie charts, box plots and histograms are plotted as per suitability with the type of feature under analysis.

3.2.3.0. As the whole objective of this exercise is predicting a potential defaulter, all the features shall mostly be plotted with the ‘Target’ variable as criterion.

3.2.4.0. Secondly, as there are 122 features in the Application Train dataset alone, visualizing each and every feature and deriving meaningful insights can be pretty time-consuming.

3.2.5.0. Hence, based on literature reviews and consequent domain knowledge coupled with practical intuition, features which ‘may have’ significant bearing on the defaulter prediction shall be visualized.

3.2.6.0. Commencing the EDA with the distribution of defaulters in the Application Train dataset [feature - ‘TARGET’] , it is observed that the application_train dataset is heavily imbalanced, as expected for a healthy lending company. This fact shall govern the major decisions such as model evaluation metrics.

3.2.7.0. Visualizing the gender-wise distribution [feature - CODE_GENDER], it can be observed that women secured a greater number of loans as compared to men, almost twice as much.

3.2.7.1. Moreover, the credit default rate is slightly lower for women than for men.

3.2.7.2. These demographic insights can help Home Credit formulate focused products and campaigns catering to females as well as introspect the disparity in genders of applicants.

3.2.7.3. There are 4 entries where Gender='XNA'. Defaulting tendency for this category is 0. Since this is not providing much information to be retained as a separate representative category, these entries may be dropped eventually unless significant insights prove contrary.

3.2.8.0. Plotting the graphs for type of loans availed [feature - NAME_CONTRACT_TYPE], it is evident that a vast majority of the applicant sample population have availed cash loans over revolving loans.

3.2.8.1. The number of defaulters for revolving loan type is a tad little lower than cash loans and may be explored by Home Credit for in-depth assessment. With context to defaulter prediction considering sample the rates are not too different and hence, loan type does not highlight a quirk.

3.2.8.2. The gender-wise split is also expected, given the ratio of female-to-male applicants.

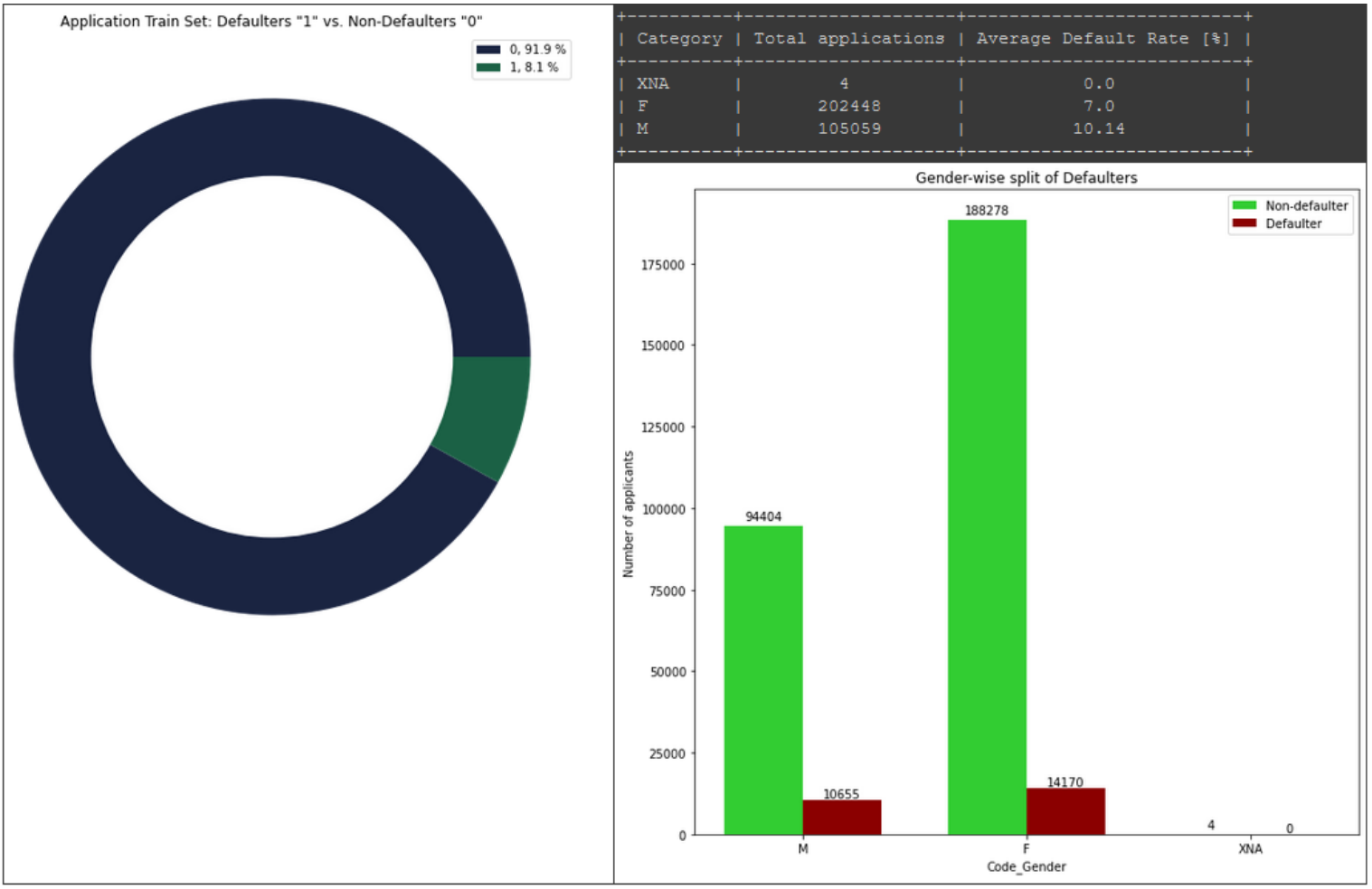


Image - Sample visualizations for context

3.2.9.0. From the 6 familial infographics [features - CNT_CHILDREN & NAME_FAMILY_STATUS], following are the insights -

3.2.9.1. A majority of applicants are married as well as having no offspring, indicative of a young demographic. However, this does not translate to a pattern for defaulter rate.

3.2.9.2. There is a significant variability in defaulter rate among other classes [ex. - High number of offsprings or unknown marital status]. However, the data points are far too few to make any sort of meaningful inference or generalization.

3.2.10.0. Relatively greater number of applicants do not own a car than those that do [feature - FLAG_OWN_CAR].

3.2.10.1. However, the defaulter rate is almost the same for both the cases and does not indicate a unique pattern.

3.2.11.0. A majority of the applicants are owning flats or some form of real estate [feature - FLAG_OWN_REALTY]. This can give an insight into the primary client-base patronising Home Credit.

3.2.11.1. However, there is no defaulter-wise aberration or pattern observed with context to realty ownership.

3.2.12.0. The statistics regarding car and realty ownership can provide Home Credit some insights regarding general wealth levels of their client base though as a feature for defaulter prediction, these statistics may not be too profound.

3.2.13.0. A majority of applicants have not provided their occupation type in the application [approx. 31.3%] [feature - OCCUPATION_TYPE].

3.2.13.1. Low-skill labourers, Drivers & Waiters have a relatively greater defaulter rate than other occupations.

3.2.13.2. Relatively high-skill applicants such as High-skill tech staff, HR staff, Core staff, Accountants and IT staff have a relatively lower defaulter rate than other occupations. Though this can be attributable to a very limited sample population, considering otherwise, this occupation demographic may be offered special incentives to avail products by Home Credit, depending on Home Credit's business goals and values [values is emphasised on as - the primary goal of Home Credit for predicting defaulters is to ensure that first time credit seekers as well as marginalised borrowers are given equal opportunities and occupation-wise promotion may be contrasting to their guiding spirit.]

3.2.13.3. These insights may help in understanding the borrowing patterns and possibly wilful defaulters in conjunction with income data.

3.2.13.4. Since there is a chance of this feature being significant to defaulter prediction, filling in the missing values is an important consideration.

3.2.13.5. Also, it would serve Home Credit well to record this data for future clients with due diligence as it 'may' affect their loan approval status.

3.2.14.0. Majority of applicants have attained secondary education [feature - NAME_EDUCATION_TYPE].

3.2.14.1. Cursory glance at defaulter-rate-per-education level suggests an inverse relation. This is more of a social insight.

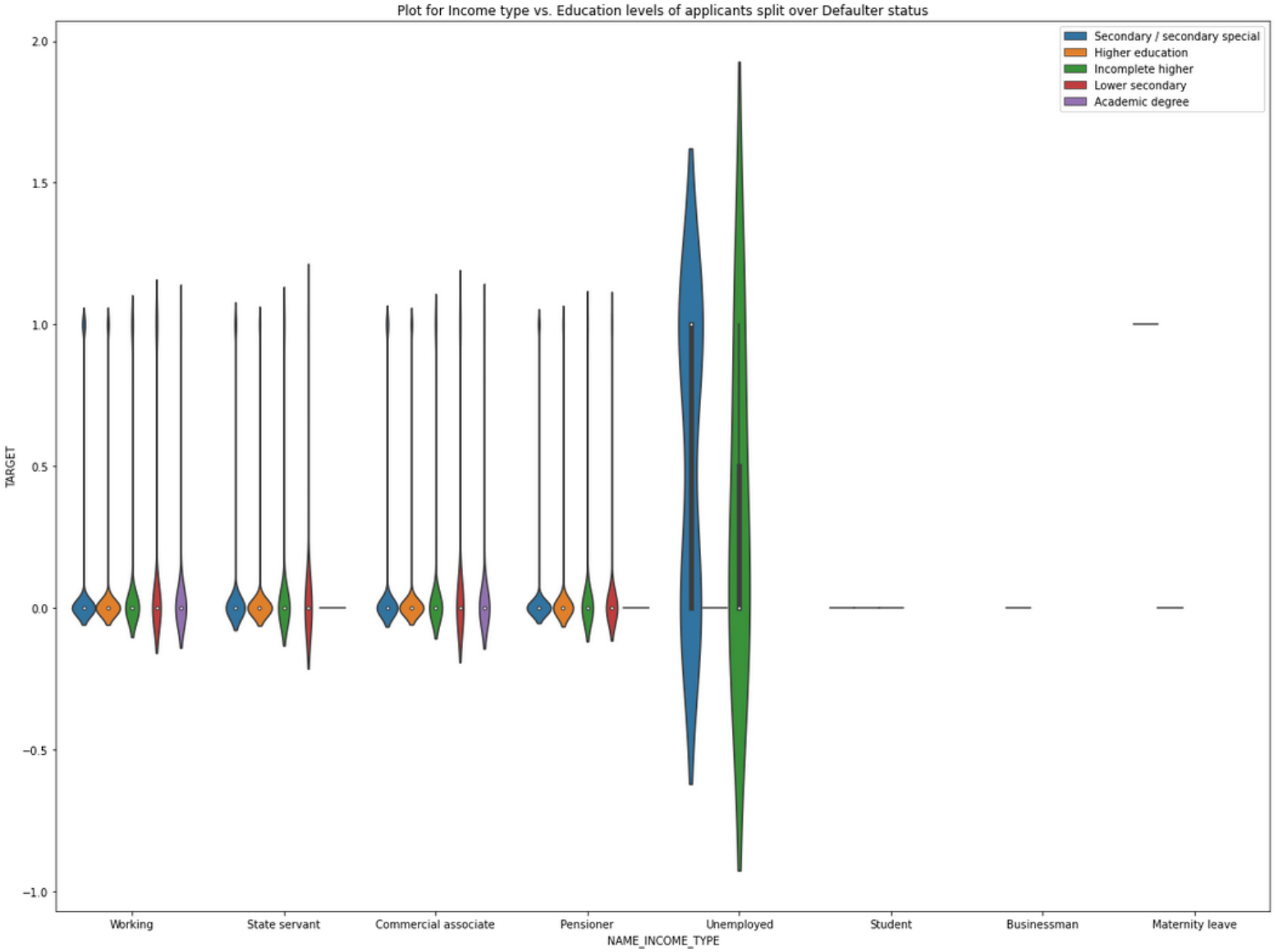
3.2.14.2. To elaborate, defaulter rate is high among applicants with secondary education and this can be attributed to the vast majority in the sample population.

3.2.14.3. Among the other levels, as mentioned earlier, defaulter rate lowers substantially with increasing education.

3.2.15.0. 'Working' income category applicants avail the most number of loans whereas Commercial Associates, Pensioners and State Servants take considerably lesser number of loans [feature - NAME_INCOME_TYPE].

3.2.16.0. Unemployed applicants and those on maternity leave have a very high default rate whereas Students & Businessmen have no defaults. However, considering the available data points' extremely limited representation, there can be no generalization possible.

3.2.17.0. A bivariate violin plot gives a visual insight regarding correlation between education, income source & defaulter status.



3.2.18.0. Other key insights in a nutshell

3.2.18.1. The External Source Normalized scores show different natures for the different defaulter states and may prove to be an important feature.

3.2.18.2. The graph for loan amount split over defaulter status is almost similar for both the defaulter classes, which suggests that defaulter tendency is independent of loan amount.

3.2.18.3. Loan amount and Annuity are directly proportional to each other which is logical. If the loan amount is high, the annuity amount for the same will also be high. However, the defaulters are split almost uniformly over the entire space which makes logistic regression'esque binary classification almost useless.

3.2.18.4. It is observed that the Default tendency for those who do provide work phone numbers is more than those who do not. This can be attributed to the fact that the wilful defaulters might be providing their work phone numbers so that they do not get disturbed on their personal mobile phone.

3.2.18.5. The bivariate graph for employment in years vs. loan amount sanctioned indicates weird values for days/years of employment. Hence this is also a case for outlier detection. Upon plotting with sanitised values, one can see defaulters are somewhat

concentrated towards the lower left side indicating lower employment as well as lower loan amounts.

3.3.0.0. Feature Engineering

3.3.1.0. Based on the domain-specific literature reviews and the features available, a few indicators of financial health or default tendency can be created.

3.3.2.0. The following are the additional features created -

3.3.2.1. Debt-to-Income Ratio - This is the ratio of loan annuity (AMT_ANNUITY) and income (AMT_INCOME_TOTAL) of the applicants.

3.3.2.2. Loan-to-Value Ratio - This is the ratio of loan amount (AMT_CREDIT) and price of the goods for which loan is given (AMT_GOODS_PRICE) to the applicants.

3.3.2.3. Loan-to-Income Ratio - This is the ratio of loan amount (AMT_CREDIT) and income (AMT_INCOME_TOTAL) of the applicants.

3.3.3.0. As was evidenced in the Venn diagram visualization, Bureau and Previous Application datasets also have valuable records worth investigation and same are merged with training dataset for further analysis and modeling.

3.3.4.0. However, an alternative approach may be explored eventually as Home Credit also aspires to cater to first time credit-seekers or marginalised populace and the model should reflect this thought process. Bureau and previous application data shall be virtually non-existent for such applicants.

3.4.0.0. Filling-in missing values and transformation

3.4.1.0. Categorical features are ‘one-hot encoded’ using Pandas’ get dummies operator with the methods for handling NaN as a category.

3.4.2.0. Missing values in numerical features are filled using the median in order to mitigate the effects of outlier values.

3.5.0.0. Outlier detection and handling

3.5.1.0. While performing the feature analysis on AMT_INCOME_TOTAL, the histogram was heavily distorted.

3.5.2.0. Generating the boxenplot, it is observed that there are some extreme income levels which are skewing the distribution.

3.5.3.0. Investigating further, there is a female applicant with a very high income level who is also a defaulter. Analysing dataset further, it is observed that loan amount is almost lying in the mid-levels which 'might' be indicative of an error in recording income levels rather than a wilful defaulter.

3.5.4.0. Considering this logic, there is a case for outlier removal.

3.5.5.0. Outlier detection is performed using the Cluster-Based Local Outlier Factor (CBLOF) scheme of the outlier detection module of pyOD library.

3.5.6.0. After specifying the parameters and carrying out the outlier removal, the dataset is checked for its split with context to class [TARGET] imbalance and it is found to be almost unchanged which is a good thing.

3.5.7.0. Moreover, plotting the boxenplot on cleansed data shows the effectiveness of the removal process as the data is much more legible as seen by the shape.

3.6.0.0. Feature selection

3.6.1.0. After the processing of data upto this point, 444 features are present in the train dataset. As many of the features may not contribute at all towards outcome prediction or even to a varying degree, it serves one well to weed out those superfluous features as is the main idea of Occam's Razor.
3.6.2.0. Currently, 2 standard feature selection models in SKLearn library are used and top 25 features are displayed for visualization.
3.6.3.0. High points of this visualization of the important features are -
3.6.3.1. The engineered features created are figuring in the top 25 features.
3.6.3.2. Many of the features thought as important during EDA do figure in the list.

3.7.0.0. High-dimensional data visualization
3.7.1.0. From bivariate analysis, it is already observed that the data is not linearly separable and hence, PCA may not provide additional insights. Towards high-dimensional visualization of the processed data, t-SNE which considers non-linear relations, is carried out as it gives one a sense or intuition of how the data is arranged in a high-dimensional space.
3.7.2.0. Visualizing the output, there is no immediate separation between defaulters and non-defaulters. Basically, it implies that both are a part of a similar class with overall similar properties.
3.7.3.0. It can be inferred that linear models may not work well and hence, other ML models capable of handling complex non-linear relationships shall be employed.

4.0.0.0. **KEY TAKEAWAYS OF EDA AND FEATURE ANALYSIS PHASE**
4.1.0.0. General Highlights
4.1.1.0. This is a very time-intensive phase involving tinkering with a myriad of features and its combinations, and requires participation from varied sources such as programmers to code effective visualizations and domain experts in order to know what to visualize.
4.1.2.0. Outcomes of this phase are very visually rich and are most useful to convey data to 'non-technical' populace.
4.1.3.0. Insights obtained in the EDA phase may be very valuable as they show patterns not usually discernable; helping translate them into tangible business outcomes.

4.2.0.0. Specific Highlights to Home Credit dataset-context EDA performed on Google Colab [Free edition]
4.2.1.0. The dataset is pretty big and needs some form of size optimization as well as storing of intermediate outputs in order to be performed on Colab's free boxes, owing to RAM usage.
4.2.2.0. There are too many features to visualize owing to time constraint and summarization and hence only a few are actually visualized in the notebook and among them, those with significant insights are listed in this report.
4.2.3.0. Owing to the various processes involved, section and subsection headings used in this report are fairly consistent with the ones used in the Colab notebook for ease of understanding and correlation.
4.2.4.0. There is a very high scope for further feature engineering based on preliminary model outputs and feature selection methods employed.

4.2.5.0. Hence, data processing done in this phase such as the features created, imputation strategies followed may be revisited based on outcome of future phases and modeling.

4.2.6.0. Realling the objectives of this phase listed in 1.4.0.0. above, the work carried out so far accomplishes the intended objectives and is helpful in the formulation of model selection and other activities for upcoming phases.

4.2.7.0. Regarding the t-SNE visualization, which is pretty time-consuming on the colab box, combinations of perplexity and iterations were tried out with help of Saurabha Daa, my diploma batchmate doing this same project, and the results are not all that different.

4.2.8.0. Regarding some visualizations such as correlations, advanced statistics such as the Phi-K statistic is used for categorical as well as mixed features. A complex correlation visualization may be constructed in the upcoming Advanced Modeling and Feature Engineering phase.