**Case Study: Analyze data for Home loan approval prediction**

**Part 1 Analysis - Graph Analysis**

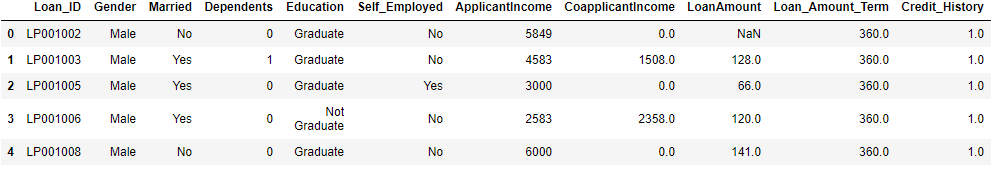
**Introduction**

Dream Housing Finance company is a growing company that provides Home loans in all over US Areas. They first check the eligibility of the customers before customers apply for the application. The company is interested in automating the pre-approval eligibility process by evaluating the customer's data provided by them on the application. This is important, as this provides the customer their likelihood of getting approved before they apply for the loan. The data is provided by the company to determine the eligibility of the customer so they can target these customers. There are many factors that can determine the eligibility of the customer, such as education level, Income, Credit History, etc. For this project, we will attempt to analyze what factors are significant in approving the loan.

**Data**

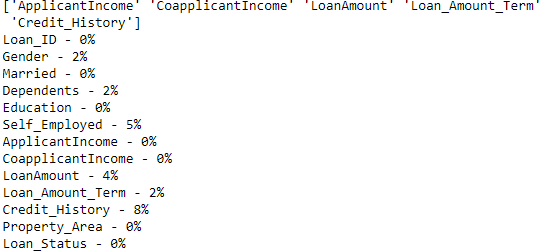
1. Data Source: Loan Eligibility Data from kaggle.com.
2. Data variables in the file are: Key Name Description

* Gender - Male/ Female
* Married - Applicant married (Y/N)
* Dependents - Number of dependents Education - Applicant Education (Graduate/ Under Graduate)
* Self\_Employed - Self-employed (Y/N)
* ApplicantIncome - Applicant income
* CoapplicantIncome - Coapplicant income
* LoanAmount - Loan amount in thousands
* Loan\_Amount\_Term - Term of a loan in months
* Credit\_History - credit history meets guidelines
* Property\_Area - Urban/ Semi-Urban/ Rural
* Loan\_Status - Loan approved (Y/N)



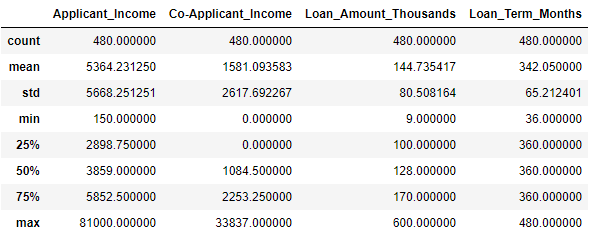
**Data Cleaning Steps**

1. Checked for duplicates using Loan\_ID, identified unique records.
2. Checked for missing values and dropped missing values. This will be revisited after dimensionality reduction:

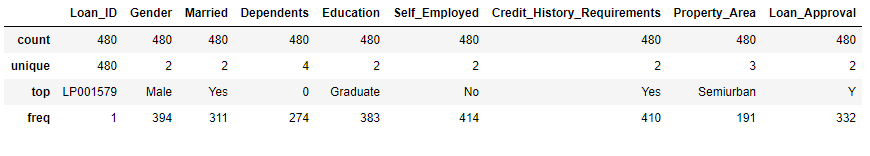


1. Stripped white space from the text data
2. Renamed Headers

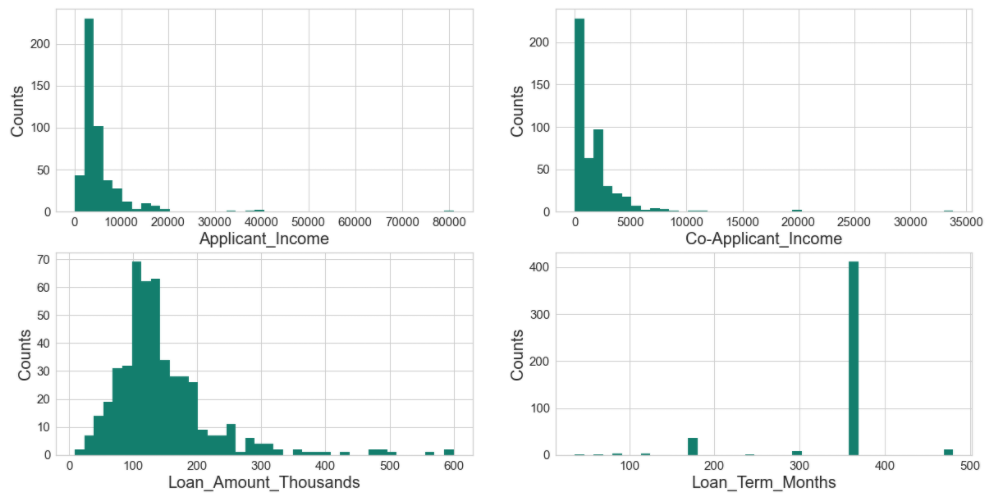
**Summary Statistics**

**Numerical**

**Categorical**

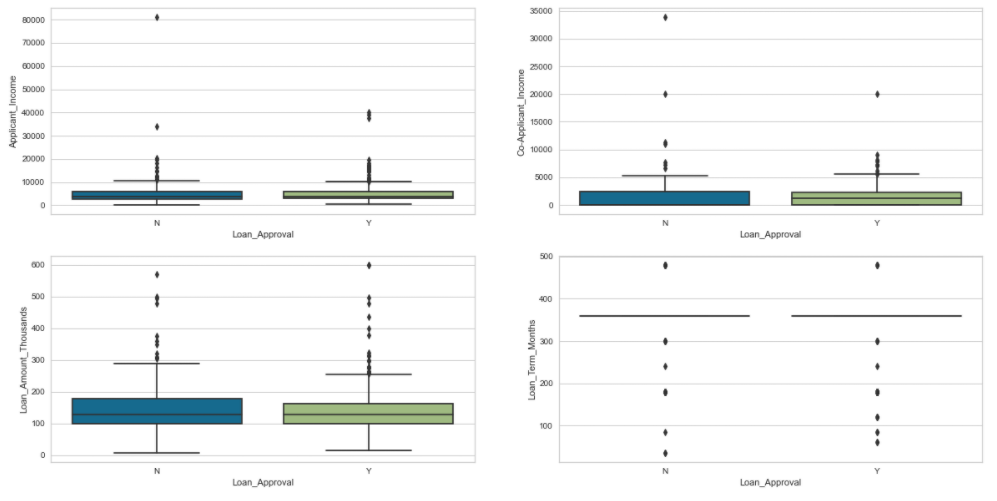
The minimum and maximum of applicant income is 150 and 81,000. So, Income can be an important feature. Co-applicant income can also be an important variable. Credit history requirements has more frequent Yes, while self\_employed has more frequent No.

**Plots**

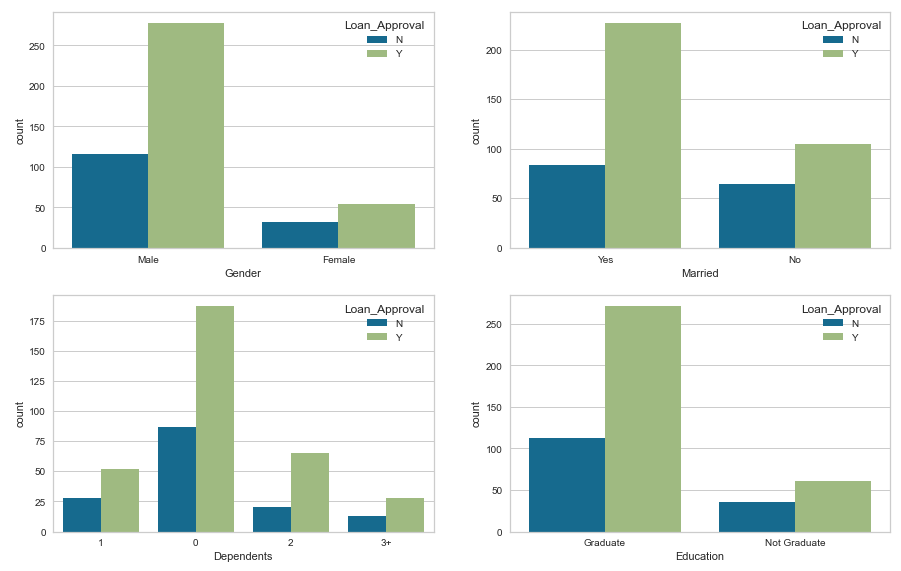
**Histograms for Numerical Data**

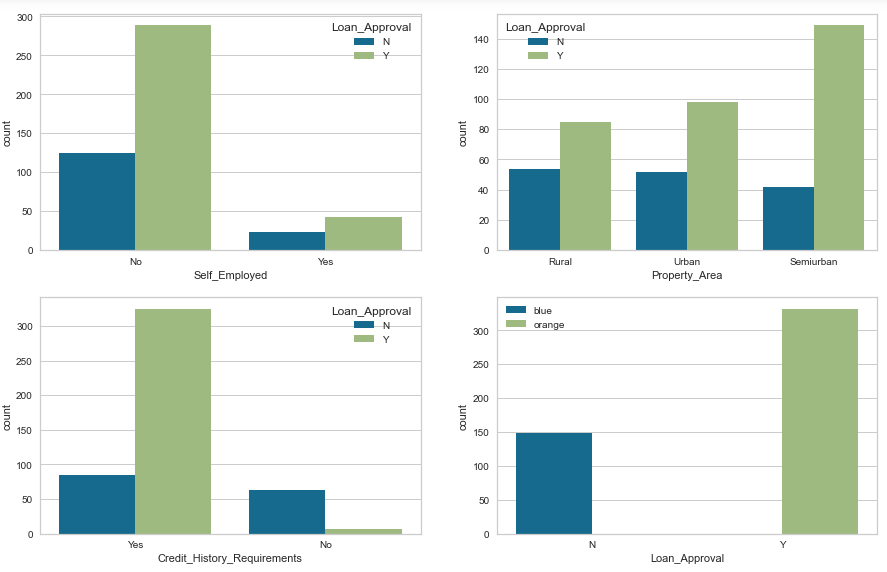
**Observations**

1. Applicant Income has a median ~5000 with extreme value of 81,000.
2. Most Applicants had no Co-applicants with mean co-applicant income is ~25000.
3. Median loan amount is ~140,000.
4. Majority of loan term is around 360 months.

**Boxplot for Numerical Data**

**Observation:** Numerical Data has no significant relevance to the Loan Approval

**Bar Charts for Categorical Data**

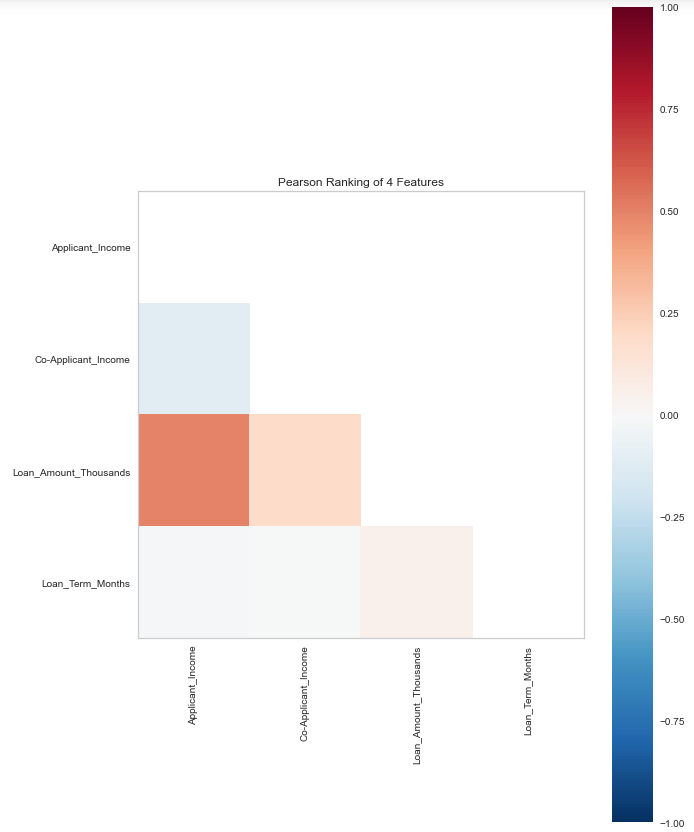


**Observations**

1. More males applied for the loan.
2. There were more married people applied to the loan and has higher rates for approval.
3. There are higher approvals for applicants with 0 Dependents.
4. Applicants with Graduate degree has higher approval rates.
5. Self-Employed applicants have low approval rate.
6. Applicants from Semi-urban areas are likely to be approved.
7. Applicants who meet the credit history requirements are more likely to be approved for the loan.
8. About 69% of the Loans are approved.

**Correlation Plot for Numerical Data**

**Observation:** There seem to have some positive correlation (~.50) between loan amount and applicant salary. There is a weak correlation between loan amount and co-applicant salary.



**Summary Observation**

1. Married couple with zero dependents is more likely to get approved.
2. It is possible for higher approval in males (need more analysis to confirm).
3. Applicants with graduate degree and applicants who are not self-employed has more chances for approval.
4. Applicant's residence area matter.

**Part 2 Steps**

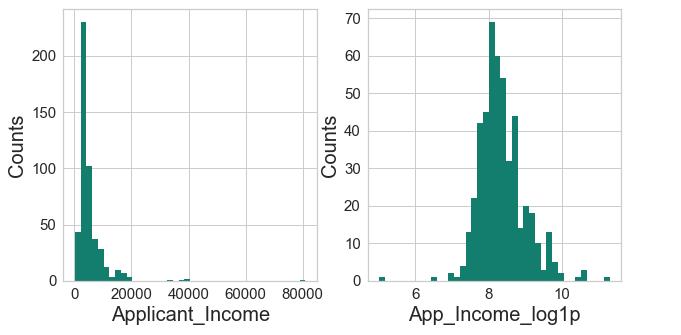
1. Analyze more based on individual categories.
2. Feature selection if needed.
3. Deal with Anomalies and see whether to keep or drop it.
4. Select classifier model to build automated system to approve loans.

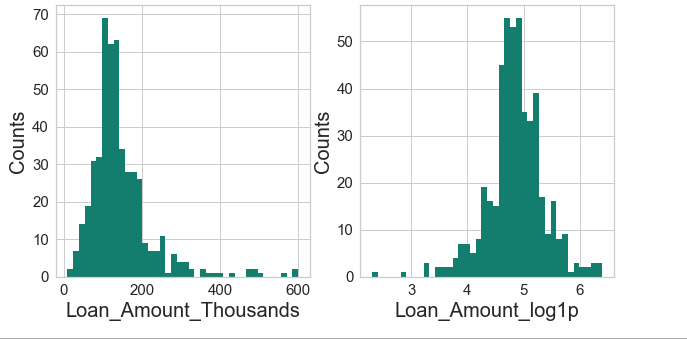
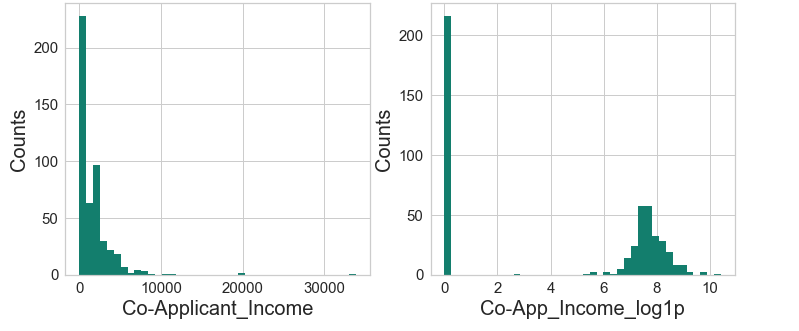
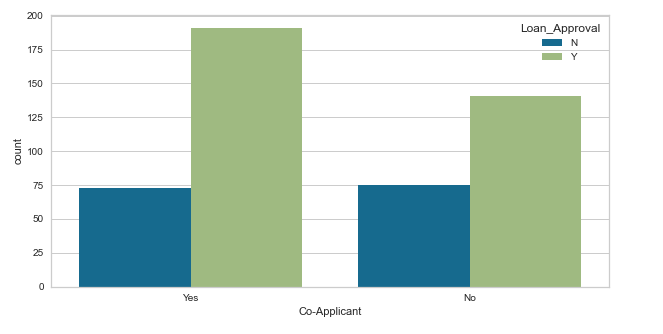
**Part 2 Analysis – Dimensionality Reduction**

**Variable Transformation**

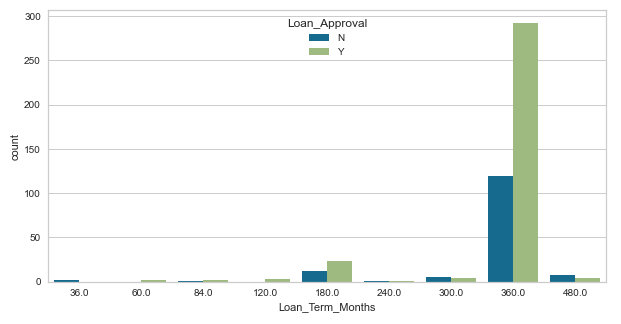
Log transforming variables (applicant's income, co-applicant's income, and loan amount because it is highly skewed data. Adding a Boolean variable of co-applicant and further extend the graph analysis for these three variables.

**Plots for new variables**





**Replotting Loan term as categorical data**



Applicants with Co-applicant are more likely to get approved for the loan compared to applicants with no co-applicant. All the log transformed variables are approximately normal except co-applicant’s income, it has lot of 0’s indicating no applicant. Loan\_Term is indeed categorical data type with numerical values.

**Dimensionality Reduction**

**Feature Selection: Removing variables by filtering columns**

I dropped columns Loan\_ID variable because it is unique and adds no value to the analysis.



For Dimensionality reduction I will be creating 2 main data frames from original data frame: df\_cat with numerical variables turned into categorical and df\_num with all Boolean, Nominal numeric and continuous variables.

* **Converting Numerical variables into categorical**

I converted numerical variables ‘App\_Income\_log1p’, ‘Co-App\_Income\_log1p’, ‘Co-App\_Income\_log1p’ by grouping them into their percentile bins (below 25th, 26-50th, 51-75th and 75th above). I did this so I can use chi-square test for feature selection. Created data frame **df\_cat** with 12 variables and data frame **subset\_cat** with 9 variables (exclusing ‘**App\_Log\_Income\_Group’**, ‘**Co-App\_Log\_Income\_Group**’ and ‘**Loan\_Log\_Amount\_Group**’).

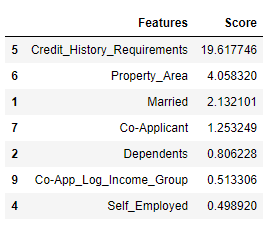
1. Gender: Male = 1, Female = 0
2. Married: Yes = 1, No = 0
3. Education: Graduate = 1, Not Graduate = 0,
4. Self\_Employed: Yes = 1, No = 0,
5. Credit\_History\_Requirements: Yes = 1, No = 0,
6. Co-Applicant: Yes = 1, No = 0
7. Dependents: 1 = 1, 0 = 0, 2 = 2 and 3+ = 3
8. Property\_Area: Semiurban = 1, Urban = 2, Rural = 3
9. App\_Log\_Income\_Group: below\_25th\_percentile = 1, 26th-50th\_percentile = 2, 51th-75th\_percentile = 3, above\_75th\_percentile = 4
10. Co-App\_Log\_Income\_Group: no\_applicants = 0, below\_25th\_percentile = 1, 26th-50th\_percentile = 2, 51th-75th\_percentile = 3, above\_75th\_percentile = 4
11. Loan\_Log\_Amount\_Group: below\_25th\_percentile = 1, 26th-50th\_percentile = 2, 51th-75th\_percentile = 3, above\_75th\_percentile = 4
12. Loan\_Term\_Months: 36, 60, 84, 120, 180, 240, 300, 360, and 480 are the categories

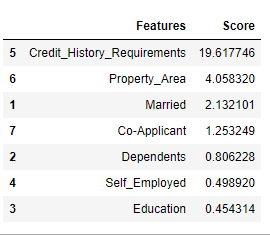
* **Converting categorical variables into Boolean variables, adding numerical variables and identifying nominal variable**

I created **subset\_num** data frame with only continuous variables 'App\_Income\_log1p','Co-App\_Income\_log1p', 'Loan\_Amount\_log1p' and created **df\_num** with all continuous variables and discrete variables.

* **Feature Selection: Chi-Square Method**

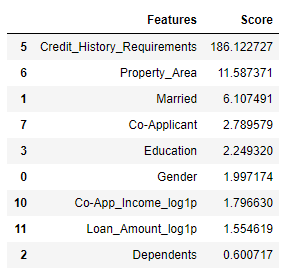
Top 7 with App\_Log\_Income\_Group, App\_Log\_Income\_Group, and Loan\_Amount\_Group variables



Top 7 without App\_Log\_Income\_Group, App\_Log\_Income\_Group, and Loan\_Amount\_Group variables

Both have Top 5 variables:

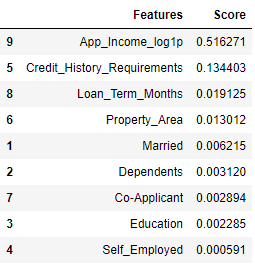
* + Credit\_History\_Requirements
  + Property\_Area
  + Married
  + Co-Applicant
  + Dependents
* **Feature Selection: ANOVA F-value statistic**



Almost same as Chi-Square results, top 4 matches.

* **Feature Selection: Mutual information (MI) selector**

This method is for set of variables, that have continuous and discrete variables such as Boolean or Nominal. It uses score\_func=mutual\_info\_classif



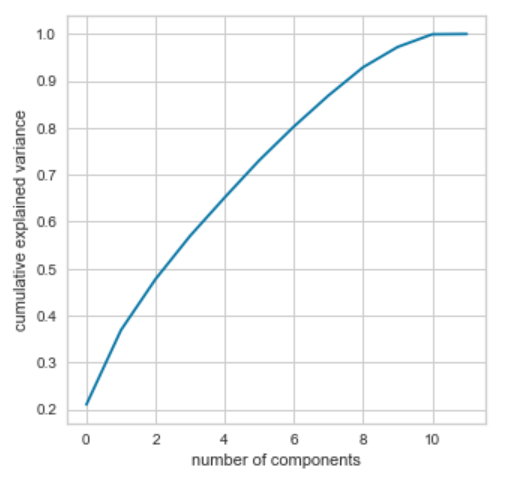
Top 7 variables are:

* App\_Income\_log1p
* Credit\_History\_Requirements
* Loan\_Term\_Months
* Property\_Area
* Married
* Dependents
* Co-Applicant
* **Correlation Check**

1. App\_Income\_log1p and Self\_Employed variables has some relationship with r = 0.23
2. Married and Dependents has little positive relationship with r = .39
3. Married and Co-Applicants has little positive relationship with r = .33

* **Feature Extraction: Principal Component Analysis**

Variance Plot for Components



I had 12 features to start with, and with PCA, it seems like 10 components is a good choice.

**Summary**

All Feature extraction or Selection methods has some limitations. Chi-Square works best when all features are categorical, and ANOVA works best when all features are numerical. I had mixture of both data types. For this reason, I attempted to convert continuous variables into category bins. Although, I had the option to do separate analysis for both data types, I do not believe separate analysis would give me any insights about patterns among all features. Mutual Information method works for both data types, but I must identify which ones are discrete. I believe this method gave me better insights on features. Further correlation analysis of these features showed some of the top features had some correlation. Later I will be eliminating one of these variables.

Lastly, to predict Loan\_Approval for customer, there are couple of models that will go best with this type of data, since most data type is binary, I may choose from Decision tree or Logistic Regression.

**References**

1. Brownlee, J. (2020, August 20). How to Choose a Feature Selection Method For Machine Learning. Retrieved February 02, 2021, from https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/
2. Wijaya, C. Y. (2020, July 07). 5 Must-Know Dimensionality Reduction Techniques via Prince. Retrieved February 02, 2021, from https://towardsdatascience.com/5-must-know-dimensionality-reduction-techniques-via-prince-e6ffb27e55d1