CAPSTONE PROJECT 1:
HOME LOAN PREDICTION

An analytical approach

https://github.com/adjakka/Springboard_Capstone_Projects/tree/master/Springboard_Capstone_Project_1

Aditya Jakka
October 2018

Abstract

Dream Housing is a financial institution that grants loans to people for the purpose of acquiring property. There are a number of variables that are taken into account when granting a loan. Some of them are: the type of property being acquired (urban, rural, semi-urban), the applicant's income, the loan term, marital status, number of dependents, level of education, their credit history and various other variables.

I hope we can all agree that there is a certain degree of risk involved when granting loans to applicants. For example, it would definitely be a lot riskier to grant loans to people with no credit history. The company is also taking into account if the applicants are self-employed. Will they be able to continue making payments if they suffer a loss in business? Should applicants with incomes on the lower scale be granted large loan amounts? A 'risk analysis' is performed before a final decision is made.

This is clearly a complex and time consuming process which is also prone to human error (if done manually). Hence, with this project, a prediction model will be built which will make decisions on behalf of the company. That is, human intervention will not be necessary.

TABLE OF CONTENTS

INTRODUCTION

Financial institutions in recent years are trying to automate the process of loan approvals. One of the major motivating factors is that it eliminates the scope for human error. The loan approvals could be for a car, education, house or any other property. As discussed in the abstract, there are certain "**variables**" like **gender, income, loan amount, loan term, type of property** that are taken into consideration. These are all **independent variables.** These **independent variables** are taken into consideration when determining the outcome, which is our **dependent variable**, the **Loan_status.**

The key goal of this project is to design a prediction model that will make decisions (in this case predict the value of the **dependent variable**) in real time. For the design of this model, historical data will be used. This ***historical data*** is available in the csv file, '***train_data.csv'.*** The historical data contains information about the applicants (all the **independent variables)** and whether they were approved for a loan or not (i.e the **dependent variable**, the **Loan_status)**. The information of applicants for whom the loan approval decision is yet to be determined (**Loan_status)** is present in the ***test*** data file, '***test_data.csv***'.

The next part of the question is – "Who can benefit from this model ?". A model of this nature could benefit any financial institutions that is in the business of granting loans. Depending on the type of loan granted, the dependent variables to be considered could vary. However, the crux of the concept and approach remains the same. This model is by no means a "perfect prediction model" and this project serves as a model upon which better prediction models can be built.

Proposed Approach

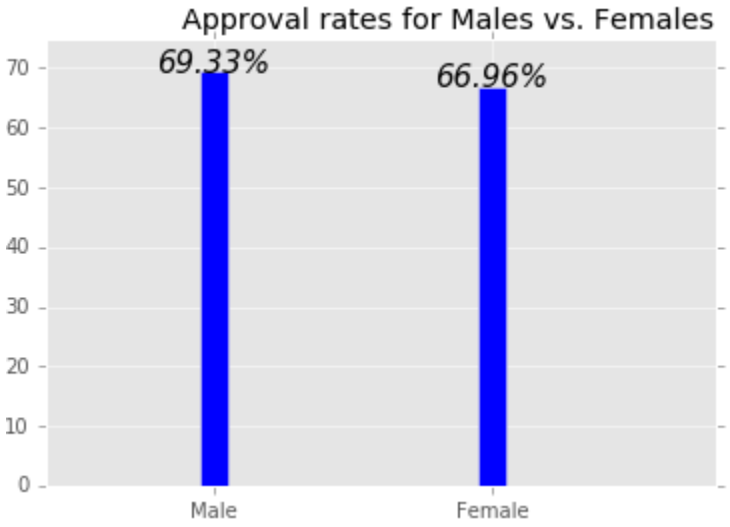
The dataset consists of a total of thirteen columns. The last column in the dataset, the **Loan_Status** is the **dependent variable**. Initially, the assumption is that all the other **independent variables** are critical in determining the loan approval. However, it is clear that the '**Loan_ID'** column plays no role in determining a decision. Similarly, through a statistical analysis, we need to determine which other dependent variables play little or no part in determining the loan approval. For example, it's possible that the "**CoApplicantIncome" variable** really doesn't affect the decision process. We may also find similar results with "**Gender**".

Once we have determined all such variables that play no role in the decision making process, we omit these variables from our analysis.

Initial Hypotheses

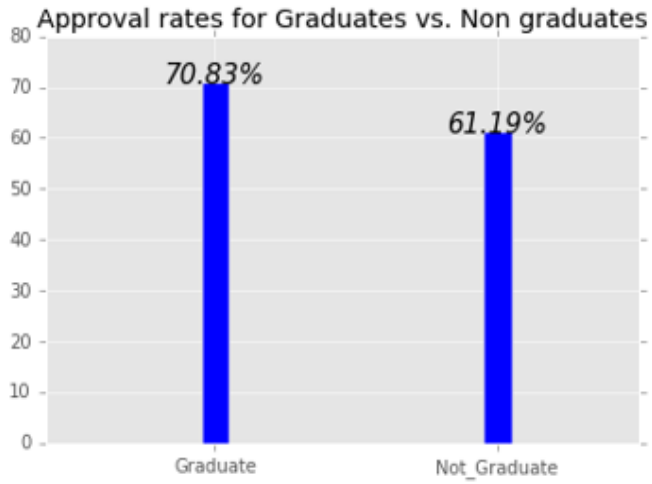
Before we clean the data and get into the nuts and bolts of predictive modeling, let us try and make a hypotheses and observe if the data echoes our assumptions.

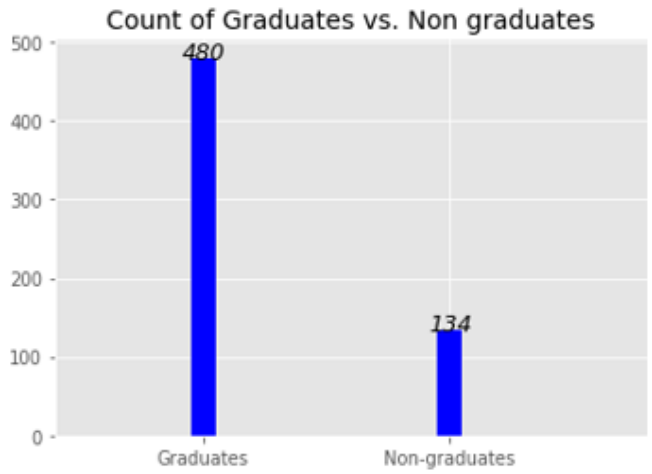
- Let us explore the data and see if males have a greater chance of approval. I computed the percentages of males vs. females approved for the loan.



From the bar plot above, we observe that indeed males do have a higher approval rating than females. However, the difference in percentage isn't significant enough to definitively confirm our hypothesis.

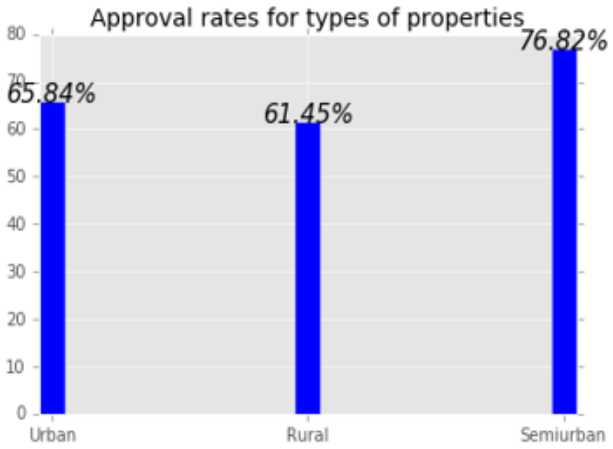
- Next, let us explore to see if Graduates have a higher chance of approval in comparison to non-graduates. From the bar graphs below, we see that graduates indeed have a ~10% higher approval chance. However, let's also see how many graduates and non-graduates were in the list. ***If there are an insignificant number of non-graduates in comparison to graduates or vice-versa, then we might have to consider rejecting the hypothesis as there aren't enough samples to support our hypothesis. [Note: This process will be followed for evaluating the other variables as well.]***

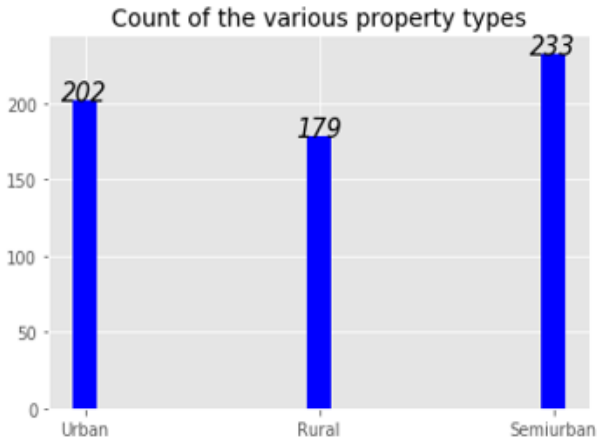




For a total of 614 applicants, 134 were non-graduates. Hence, we can argue that our hypotheses could be true.

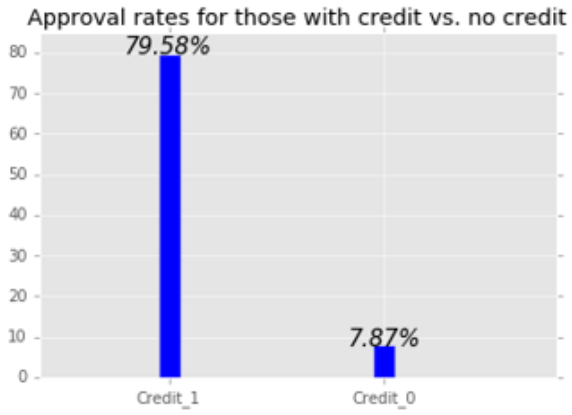
- Next, let us consider the property type and the percentage approvals. We see that semi-urban areas have a higher approval rate. They have a 15% higher approval rate than rural properties and more than 10% approval rate than urban areas. But let us also consider the count of the number of Urban, rural and semiurban properties in question. Two bar plots has been illustrated below.

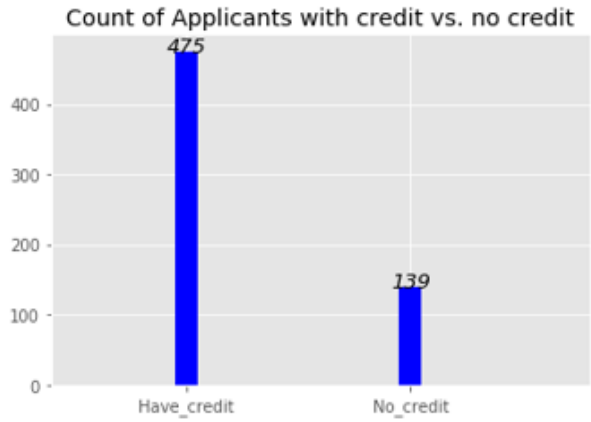




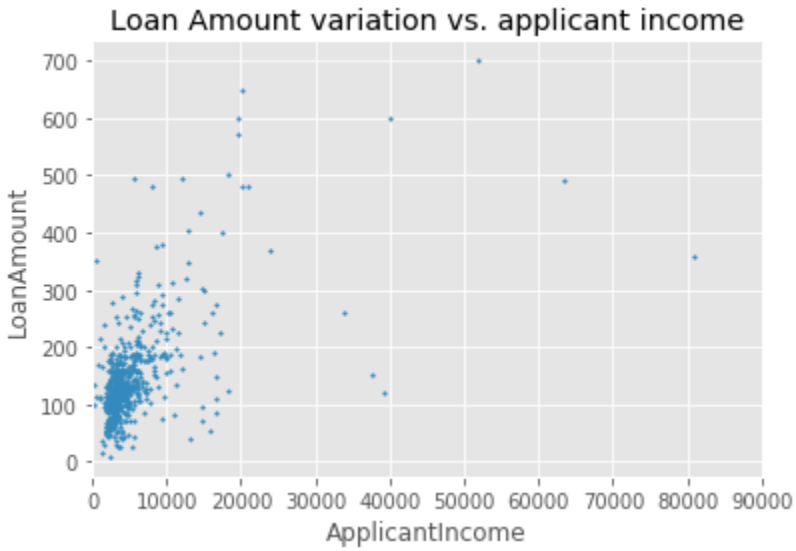
We see a significant number of loan applications for each of the property types. Hence, we could again argue that perhaps, our hypotheses holds true.

- Next, we will try to determine if people with a credit history has a higher chance of loan approval as opposed to those who don't. Let us explore the bar graph below. It would seem that applicants with a credit history are 10 times more likely to get approved. But before we confirm our hypothesis, let's investigate the number of people with credit vs no credit. Again, there are a significant number of samples for people with no credit. Hence, we could again argue that for now, our hypothesis holds good.

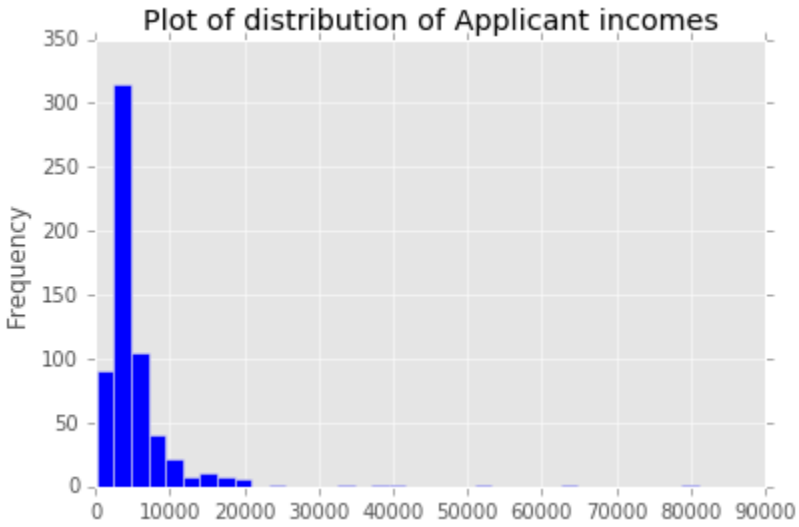




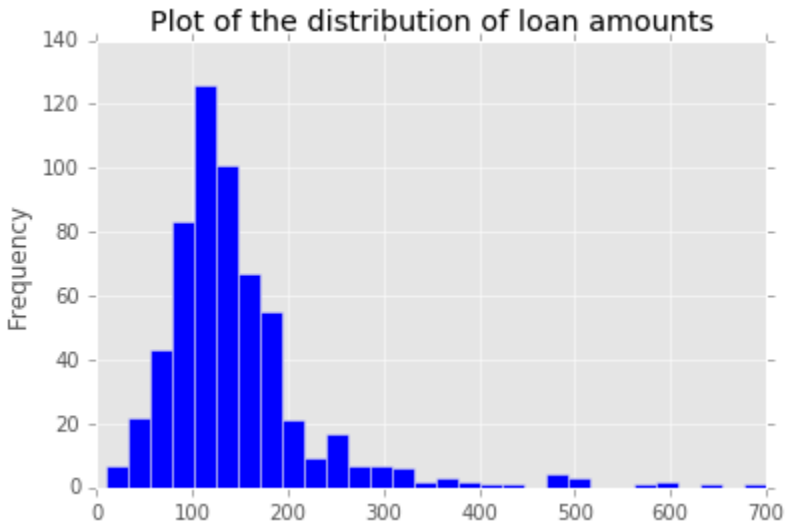
- Next, let us try to see if there is a relationship between the applicants' incomes and their loan amounts requested. While there is some indication that applicants with higher incomes requester larger loan amounts, the data points are diversely spread through the graph and we cannot confidently conclude that there is a solid pattern that supports our hypothesis.



- Next, let us consider the distributions for applicant income. This has been illustrated with a histogram below. From the histogram plot below, it's evident that most of our applicants make $5000 a month or less. In fact, more than 50% of the applicants make between $2500 and $5000.



- Let us also consider the distribution for loan amounts requested. We see that most of the applicants request a loan amount of $200K or less.



Cleaning Steps performed

Before I get started, I will attach a snapshot of all the columns in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
Loan_ID              614 non-null object
Gender               601 non-null object
Married              611 non-null object
Dependents           599 non-null object
Education            614 non-null object
Self_Employed        582 non-null object
ApplicantIncome      614 non-null int64
CoapplicantIncome    614 non-null float64
LoanAmount           592 non-null float64
Loan_Amount_Term     600 non-null float64
Credit_History       564 non-null float64
Property_Area        614 non-null object
Loan_Status          614 non-null object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.4+ KB
```

There are 614 observation in total. Some of the columns have missing values. These values had 'NaN' assigned for missing values.

Before we can an in-depth analysis of the cleaning steps performed, I will explain two concepts: Mode, median and Outliers

Mode - The **mode** of a set of data values is the value that appears most often. Let us consider a list, x = [1,2,2,3]. The mode of the list x is 2 since it is the most repeated value.

Median - The median is the value separating the higher half from the lower half of a data sample. Let's consider a list x = [3,4,5,6,7]. The median of x in this case is 5. For an even number of observations, the average of the two middle observations is computed as the median. This is represented by the ***50th percentile*** line in the box plot below.

Gender – To fill the missing values of this column, I computed the mode value (most repeated value) of the Gender column. The 'NaN' or missing values were then substituted with this value. In this case, it was '*Male*'.

Married – Just as the gender column, missing values for the married column were filled with the mode value. The 'NaN' or missing values were then substituted with this value. It was '*Married*' for this column.

Dependents – Again, the mode value was used. It was '0'.

Self Employed – The mode value was computed once again.

LoanAmount – The mean of Loan amounts was computed and assigned to the NaN values.

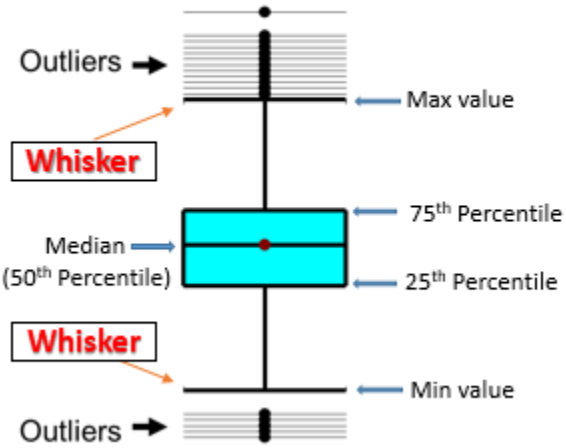
Loan_Amount_Term – The mode was computed and assigned to the missing values. It was 360.

Credit History – The mode was computed once again. It was 1.0

Dealing with Outliers

Outliers - An **outlier** is an observation point that is distant from other observations. These data points could be a result of errors or wrongly recorded observations. These are often excluded from analysis as it could have an adverse effect on the reliability of our model.

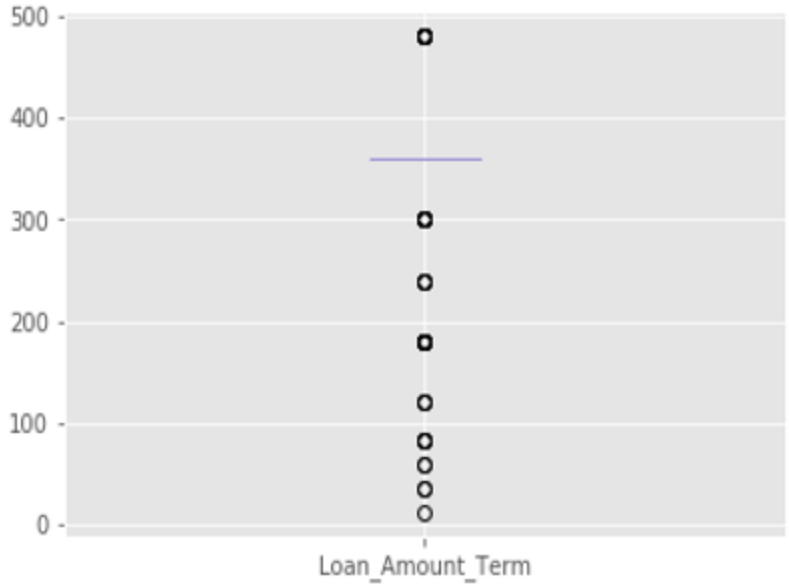
We will identify outliers with the help of box plots. A typical box plot with outliers is illustrated below:



The data points above and below the **whiskers** are our **outliers**. The min and max values in this plot have been plotted by excluding the outliers from the calculation.

I will now outline the steps I took to clean the data:

Loan_Amount_Term – From the box plot, there is one data point > 360. On pulling values from the dataframe for Loan_Amount_term > 360, I observed that there quite a few applicants for loan_amount term = 480 and considering the relatively small size of this dataset, these data points will be taken into consideration for analysis. However, there are a few data points below 100.

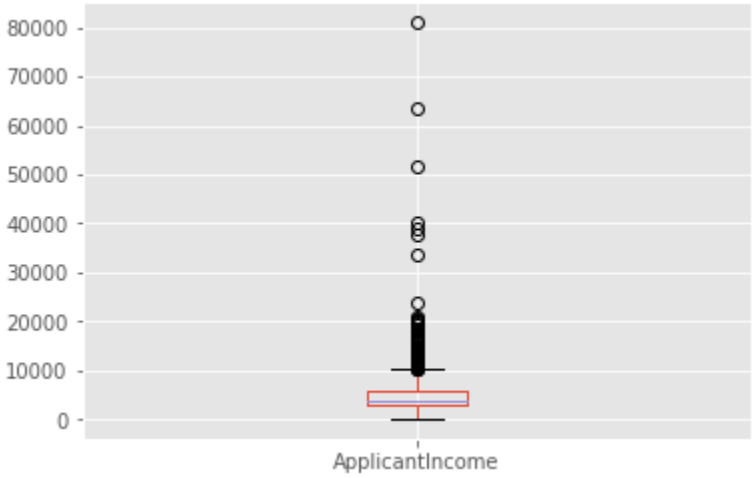


	Loan_Amount_Term
Loan Amount terms < 100	60.0
	60.0
	36.0
	84.0
	84.0
	12.0
	36.0
	84.0
	84.0

So when I take a peek at that data, here's what I get: A total of 9 data points below 100. However, we can see that there are just 2 values below 50: 12 and 24. Therefore, it seems highly unlikely that applicants would apply for a 12 or 36 term housing loan.

Proposed course of action: Omit data points below 50.

ApplicantIncome - By observing the box plot for applicant incomes, we again see quite a few outliers.

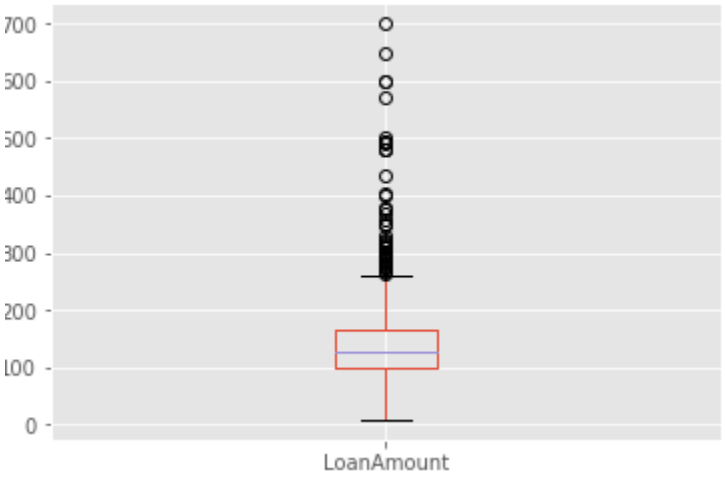


ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
51763	0.0	700.0	300.0	1.0	Urban	Y
63337	0.0	490.0	180.0	1.0	Urban	Y
81000	0.0	360.0	360.0	0.0	Rural	N

All the values in the boxplot from 50000 and above seem pretty far away and we see that there are indeed just three data points.

Proposed course of action: Even if we do assume these rows to be legitimate, I think perhaps we could drop these rows for analysis? Simply because the % of people with an income > 50000 probably won't apply to Dream Housing.

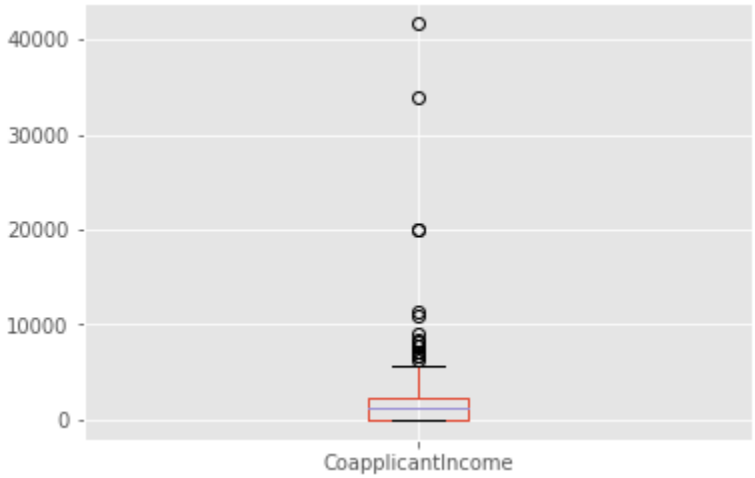
LoanAmount - The box plot shows quite a few outliers. We see that some outliers over 500 (illustrated beside the box plot) that are pretty isolated. We find 3 data points in the 600s and one at 700.



LoanAmount
650.0
600.0
700.0
570.0
600.0

Proposed course of action: Omit the data point at 700.

CoapplicantIncome - On observing the box plot, we again see quite a few outliers but most of them are located close to the box plot. However, there are three data points over the 20000 mark that appear to be pretty isolated.



Proposed course of action: There are only 4 data points in the data set. Unsure what to do. Need to discuss during our call.