Table of Contents

[Introduction 1](#_Toc104995215)

[Problem Definition 1](#_Toc104995216)

[Loading Essential Python Libraries 1](#_Toc104995217)

[Data Pre-processing 3](#_Toc104995218)

[Detection and handling of missing values 4](#_Toc104995219)

[Detecting and handling outliers 5](#_Toc104995220)

[Descriptive Analysis 7](#_Toc104995221)

[Hypothesis Generation 9](#_Toc104995222)

[Algorithm Definition 12](#_Toc104995223)

[System Features 13](#_Toc104995224)

[System Architecture 13](#_Toc104995225)

[Methodology 14](#_Toc104995226)

[Model selection 14](#_Toc104995227)

[Use Case Diagram 18](#_Toc104995228)

[CONCLUSION 20](#_Toc104995229)

[REFERENCES 20](#_Toc104995230)

# 

**Introduction**

This report is about Machine learning algorithms that automate decision making by generalizing from known examples called supervised learning. Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this report is to provide quick, immediate and easy way to choose the deserving applicants. Customer first applies for loan after that the bank validates the customer eligibility for loan. The bank wants to automate the loan eligibility process based on customer details provided while filling application form.

**Problem Definition**

The bank wants to automate the loan eligibility process based on customer detail provided while filling an application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments; those are eligible for loan amount so that they can specifically target these customers.

[1]Loan prediction is a very common real-life problem that every bank faces in their lending operations. If the loan approval process is automated, it can save a lot of hours and improve the speed of service to the customers. The increase in customer satisfaction and savings in operational costs are significant.

**Loading Essential Python Libraries**

We imported pandas to load data and also explore the data both with descriptive statistics and data visualization. We chose pandas because it provides support for working with tabular data such as CSV, Excel and many more, it is used for data analysis and associated manipulation of tabular data in data frames.

Why import pandas

* Data representation: It provides extremely streamlined forms of data representation, this helps in analysing and understanding data better.
* Efficiently handles large data: It helps to save a lot of time by importing large amounts of data very fast.
* Supports common visualization graphs and plots
* Extensive file format compatibility
* Makes data flexible and customizable: Pandas provides a huge feature set to apply on data one has so that one can customize it, edit and pivot it according to one’s will or desire.

Numpy is referred to as a python package used for performing various logical operations and mathematical operations on arrays. Numpy module works with numerical data and it is said it consumes less memory as compared to Pandas

Why import Numpy

* Numpy is very good for data analysis
* It has a better performance for 50K rows or less
* Numpy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them efficiently.
* It aims to provide an array object that is up 50x faster than the traditional python lists.

|  |  |  |
| --- | --- | --- |
| Point of Comparison | Pandas | Numpy |
| Type of data supported | For working with tabular data such as CSV, Excel and others | Data in the form of arrays and matrix |
| Performance | Pandas performs best with more amounts of data, around 500K or more rows | Numpy performs best with lesser amounts of data, around 50000 or less rows |
| Popular Industry Usage | Is popularly used for data analysis and visualization | Is popularly used for numerical calculations |

Matplotlib is referred to as a plotting library for the python programming language and its numerical mathematics extension Numpy. It provides an objected-oriented API for embedding plots into applications using general purpose. Matplotlib allows a user to add markers and change colours and for some types of visualizations.

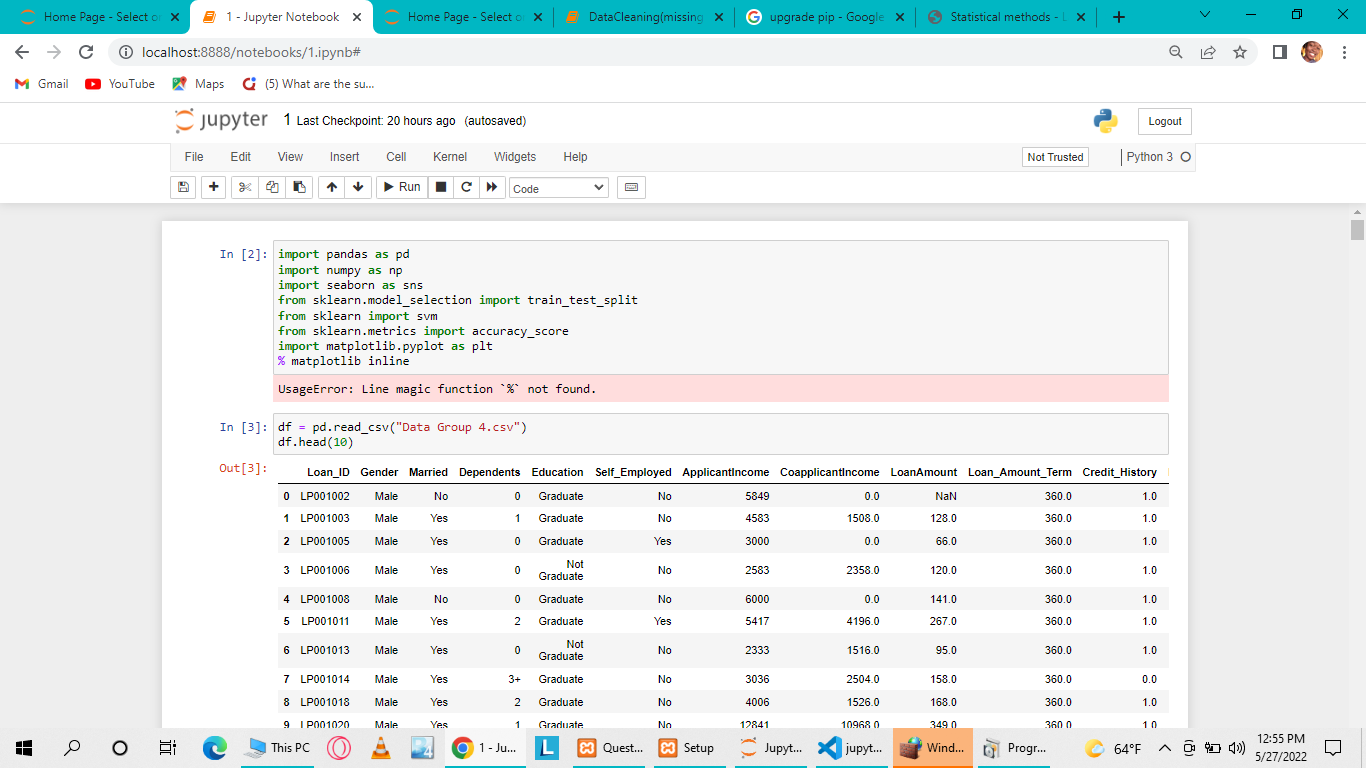
Advantages of using Matplotlib:

* Versatile and accessible
* Customizable
* Good documentation

Seaborn is a library that uses Matplotlib underneath to plot graphs, it will be used to visualize random distributions.

Advantages of Seaborn

* Seaborn works well with pandas data structure compared to Matplotlib
* Can easily customize plots
* It generates a dynamic and informative plot to represent data



**Data Pre-processing**

Is defined as the steps needed to follow in order to transform or encode data so that the machine may easily parse it. Importance of data pre-processing is to improve the overall data quality.[2]

4 Steps in Data Preprocessing

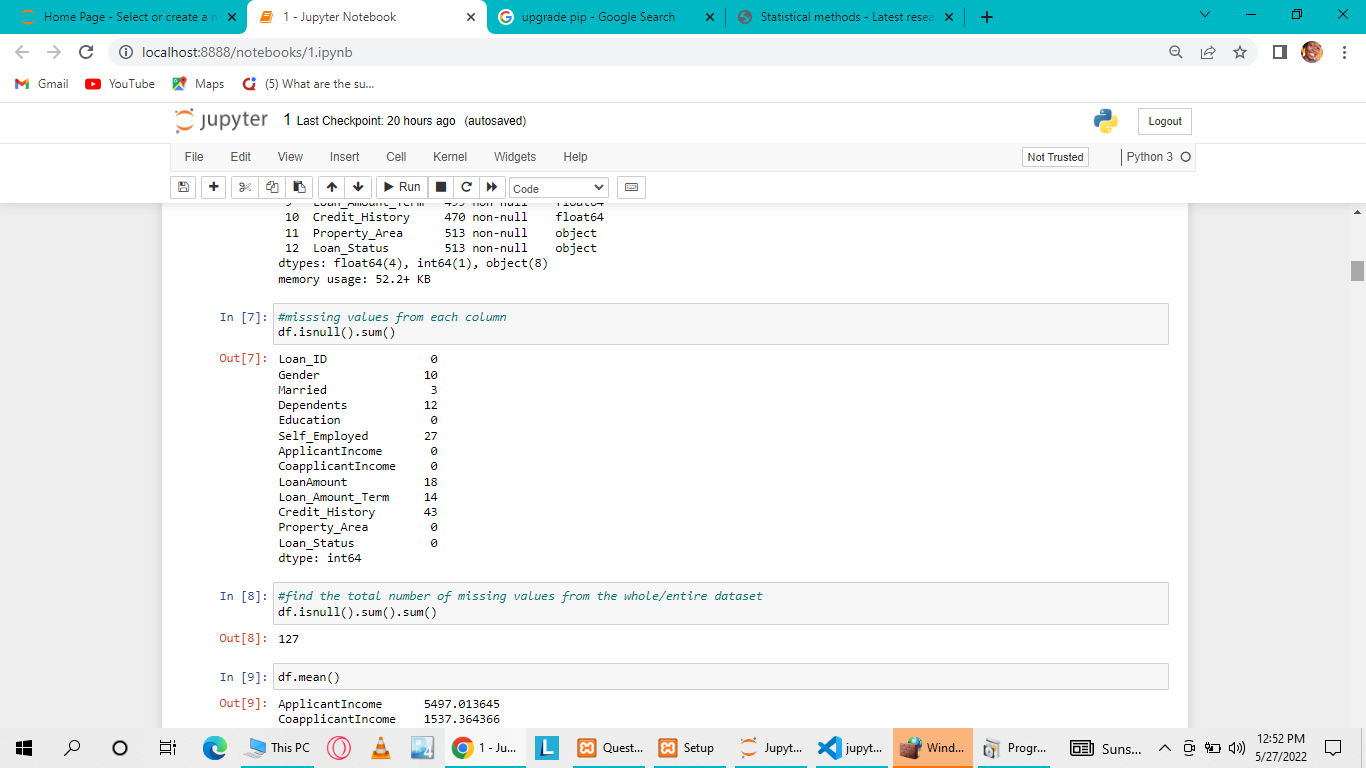
* Data Cleaning: Cleaning data by filling the missing values, resolving the inconsistency and also removing outliers.
* Data Integration: it is used to merge data present in multiple sources into a single large data store.
* Data Transformation: Once the data clearing has been done, we need to consolidate the quality data into alternate forms by changing the value, structure or format of data using generalization and normalization.
* Data Reduction or Dimension Reduction: Obtaining a reduced representation of the dataset that is much smaller in volume but produces the same quality of analytical results [2].

**Detection and handling of missing values**

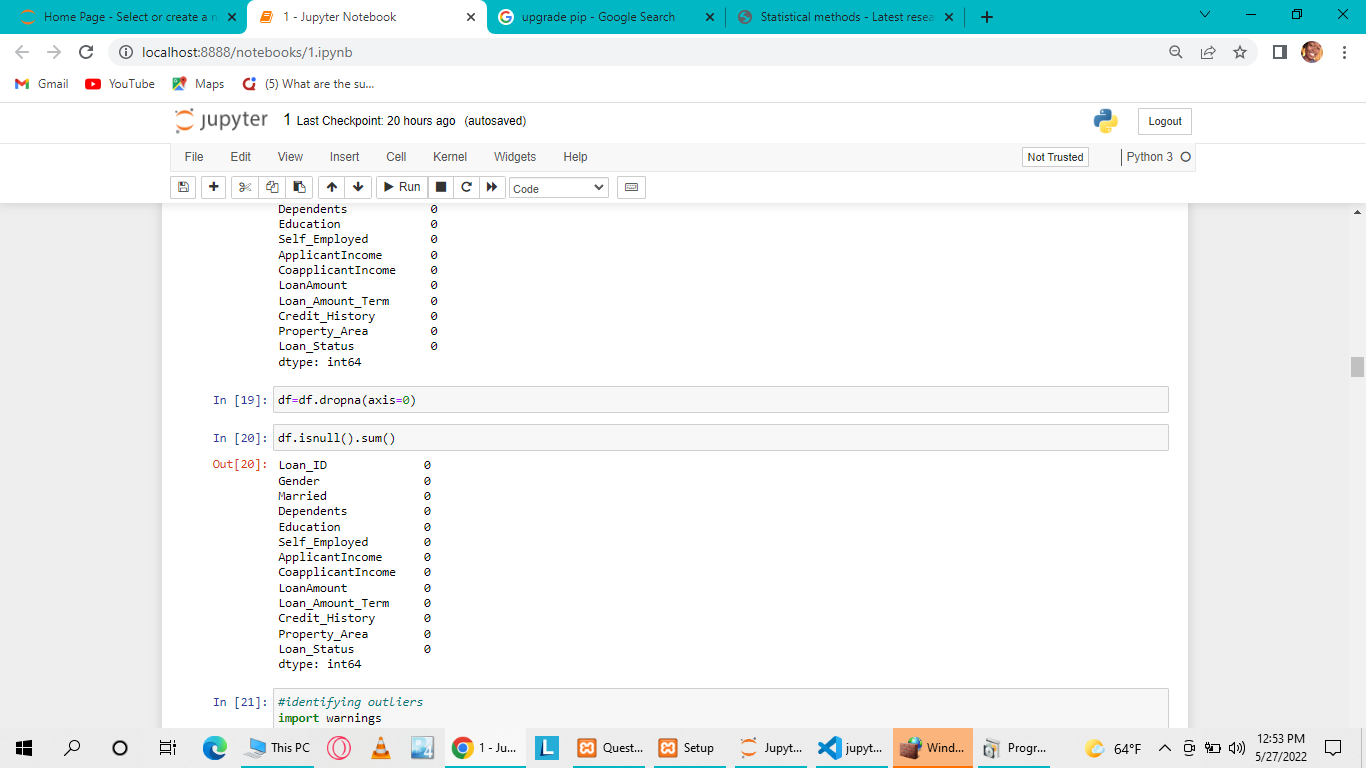
Missing values may give incorrect view of the overall statistics of data so we have to identify them from the entire dataset and drop unwanted columns [3].

Handling-missing values: Since we are assigned to create a model that will decide whether to grant a loan or not, we must have all the information required. If there is a missing value, we either find it or delete a row with a missing value.

* Replacing with mode: Mode is used in the categorical features. We used the columns of Married and Self-Employed
* Replacing with median: Median is used when data is skewed. Loan Amount Term, Loan Amount
* Replacing with mean: filled the missing values from Loan\_Amount using mean because it is the most common method of imputing missing values of numeric columns.
* Forward fill: Credit\_History
* Replacing using arbitrary value: Dependants
* Deleted rows with missing values on Gender. We decided to delete rows where gender is missing because one cannot guess whether a person is a male ore a female.



The figure above shows the missing values for each column

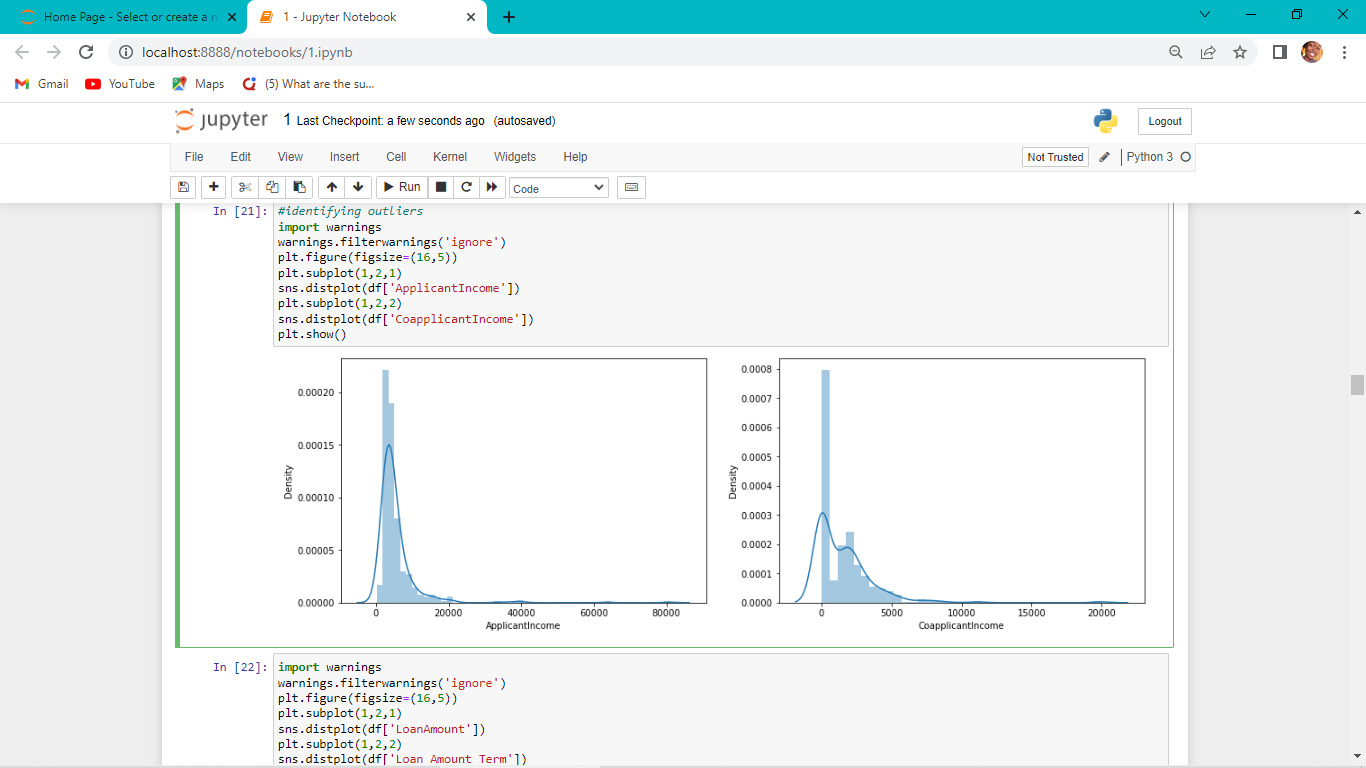


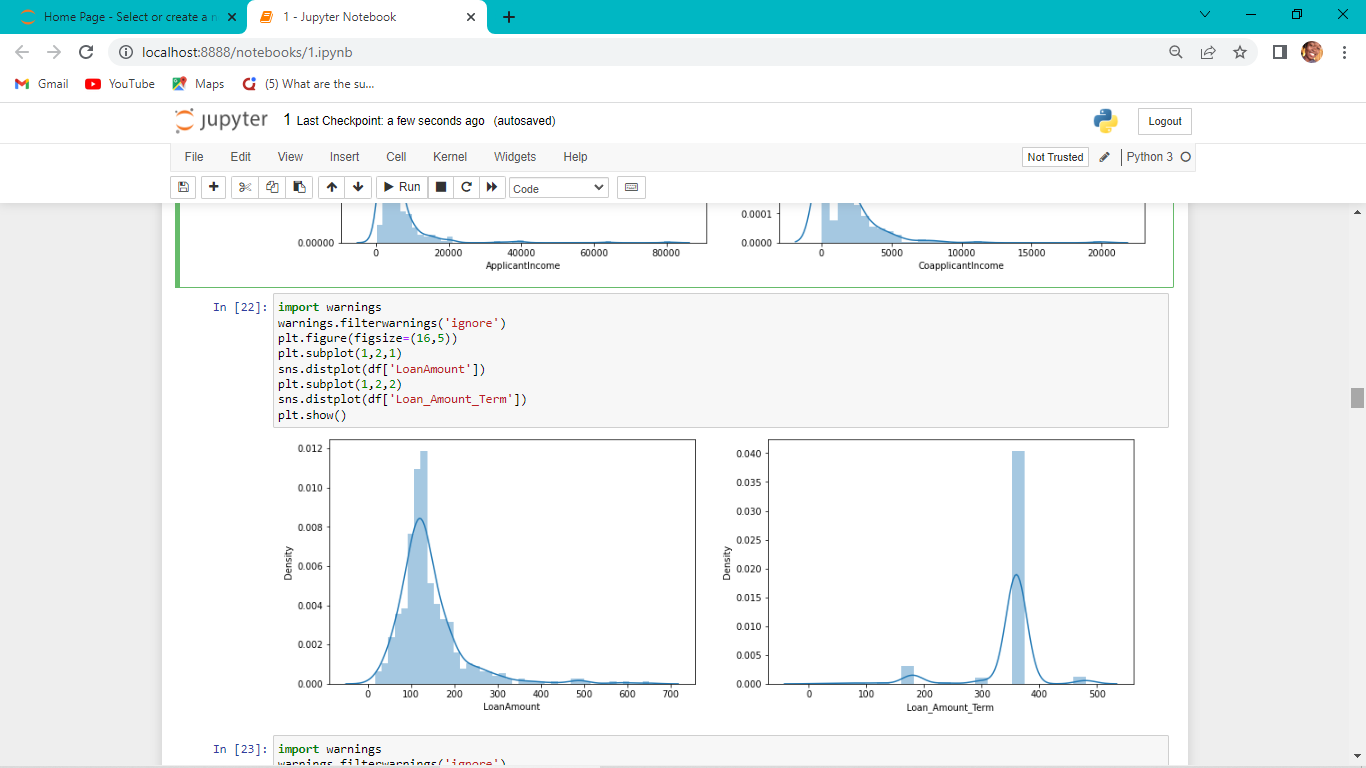
The figure above shows each column after filling the missing values

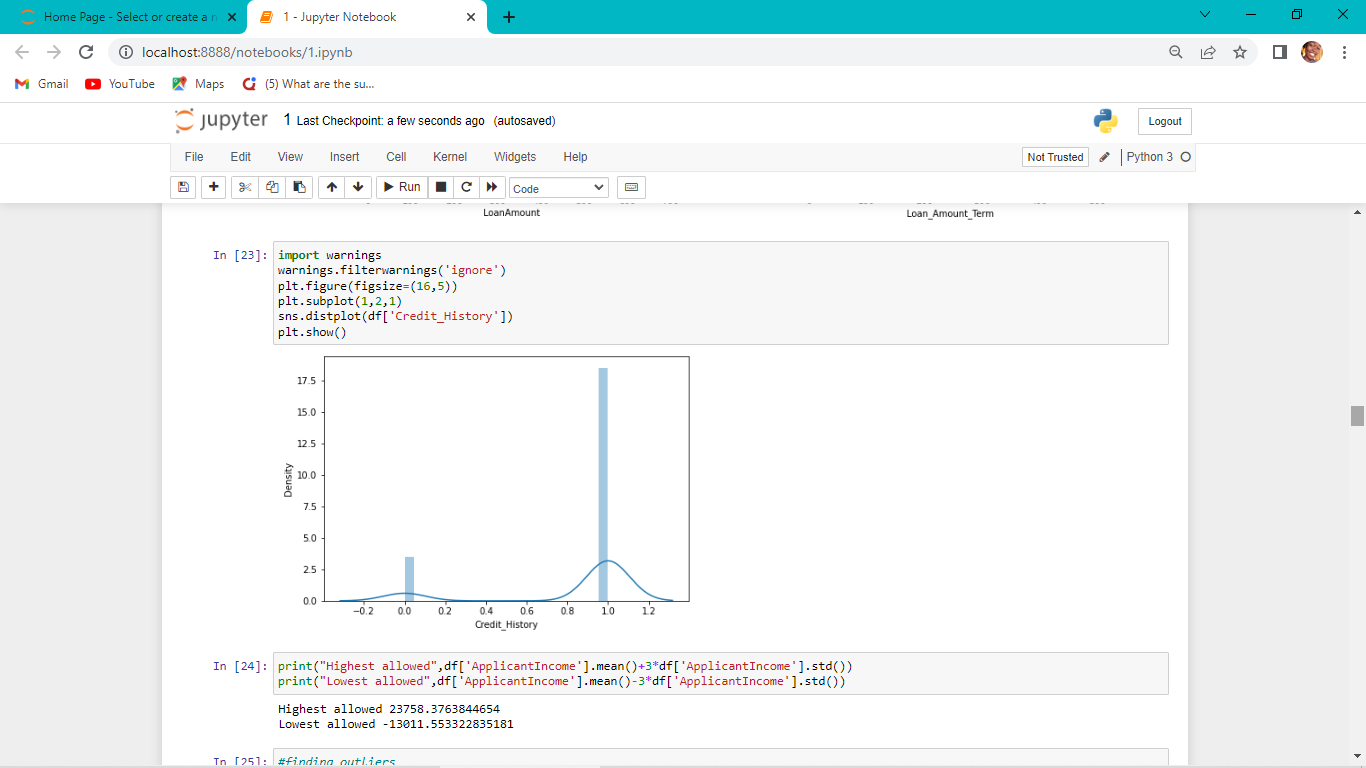
**Detecting and handling outliers**

Outliers are defined as extremely high or low data point relative to the nearest data point and the rest neighbouring co-existing values in a dataset given They are values that stand out greatly from the overall pattern of values in a dataset [4]. To identify the outliers in each column, we used the distribution plot, which makes charts and plots look engaging and enables some of the common data visualization needs. Basically, it makes the data visualization and exploration easy to conquer.

The below figure shows identifying outliers using a graph



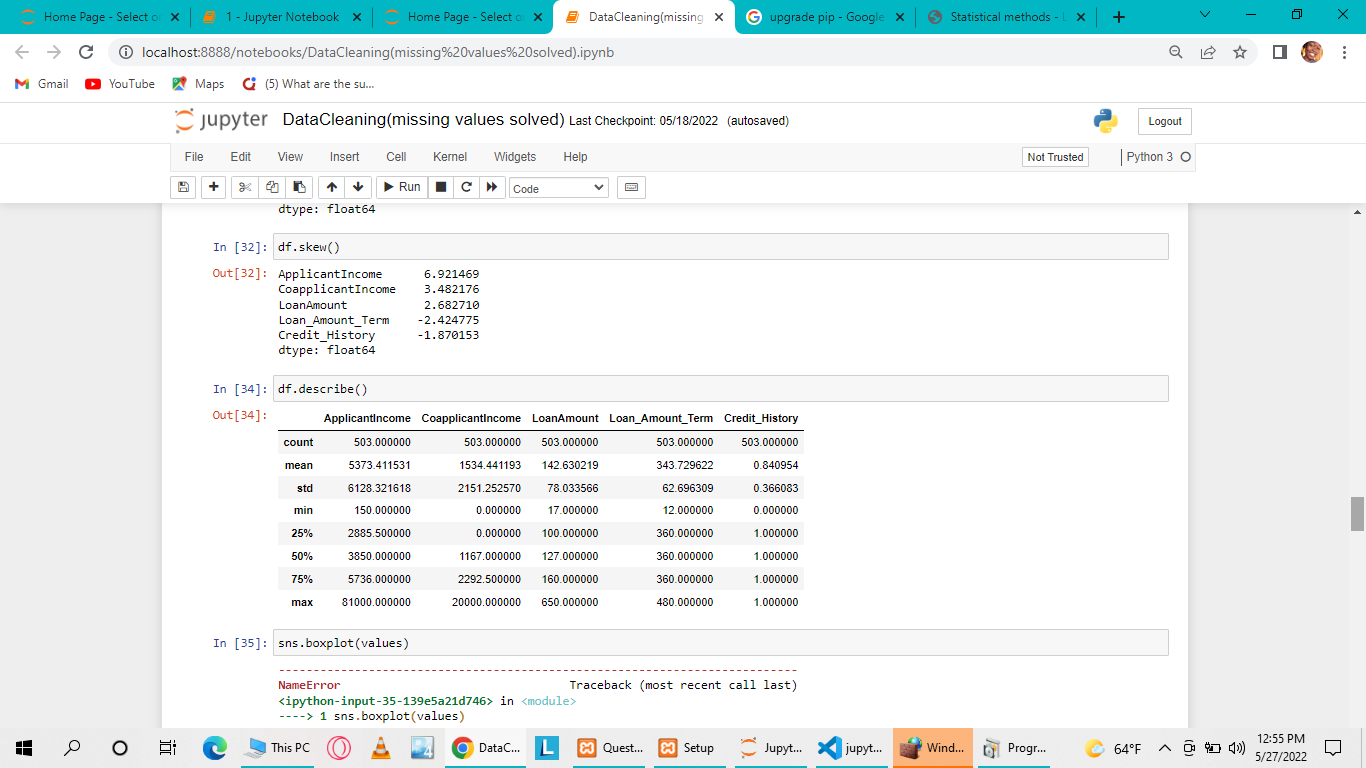




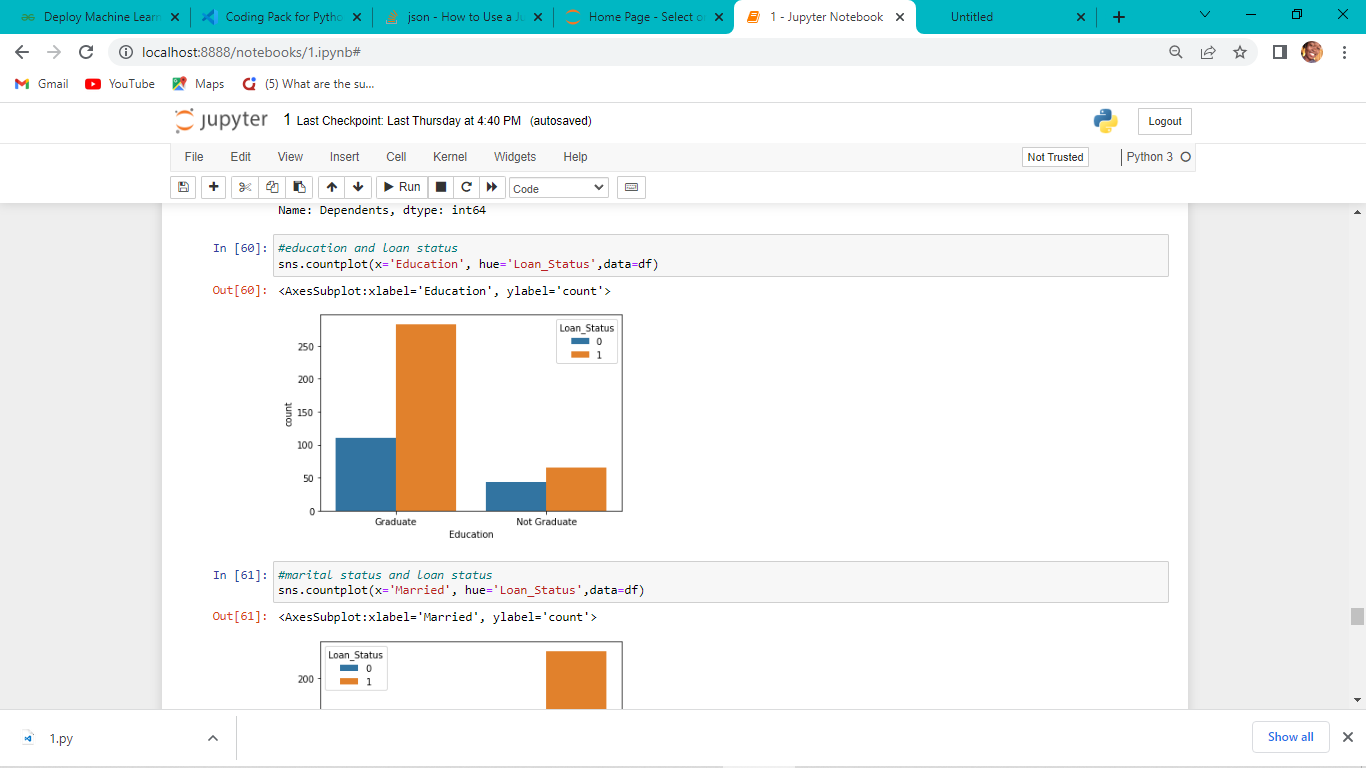
* Capping – We cap our outliers data and make the limit, that is above a particular value or less than that value, all the values are considered as outliers and the number of outliers in the dataset gives that capping number. Its advantage is its fastest nature
* Trimming – Data excludes the outlier values from our analysis

**Descriptive Analysis**

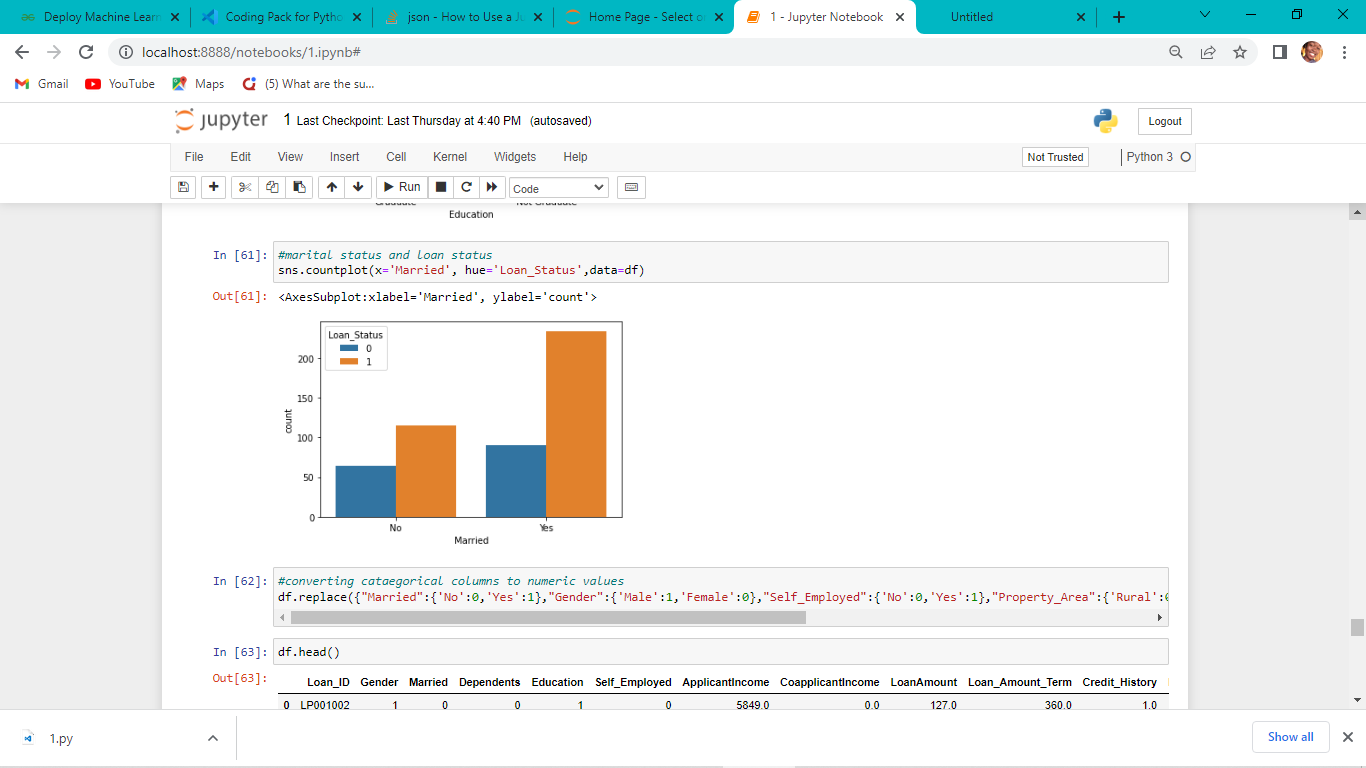
Used describe function to show the description of the dataset. This gives the summary of the count, mean, standard deviation, minimum, maximum and the quintile values.



Exploring data



The above figure shows that being a graduate increases ones chance of getting a loan.

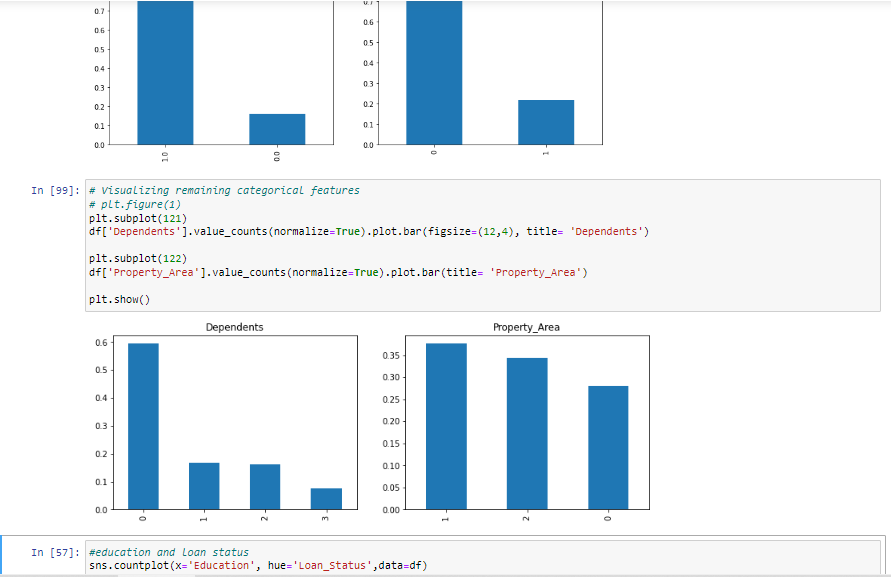


The above figure shows that being married increases ones chance of getting the loan based on the data.

**Hypothesis Generation**

Hypothesis Generation is the process of listing out all the possible factors that can affect the outcome i.e. which of the features will have an impact on whether a loan will be approved or not [5]. Some of the hypothesis are:

* Education: Applicants with higher education level i.e. graduate level should have higher chances of being granted a loan. Meaning the proportion of loans getting approved for graduates is higher compared to non-graduates.
* Income: Applicants with higher income should have more chances of loan approval
* Loan amount: If the loan amount is less, the chances of loan approval should be high
* Loan term: Loans with a shorter time period should have higher chances of approval
* Previous credit history: Applicants who have repayed their previous debts should have higher chances of loan approval. Applicants with credit history as 1 are more likely to get their loans approved
* The proportion of loans getting approved in semi-urban area is higher as compared to that in rural or urban areas.
* The proportion of male and female applicants is more or less same for both approved and unapproved loans
* Proportion of married applicants is higher for the approved loans
* The distribution of applicants with 1 or 3+ dependents is similar across both the categories of Loan\_Status.





It can be inferred from the above bar plots that:

80% applicants in the dataset are male.

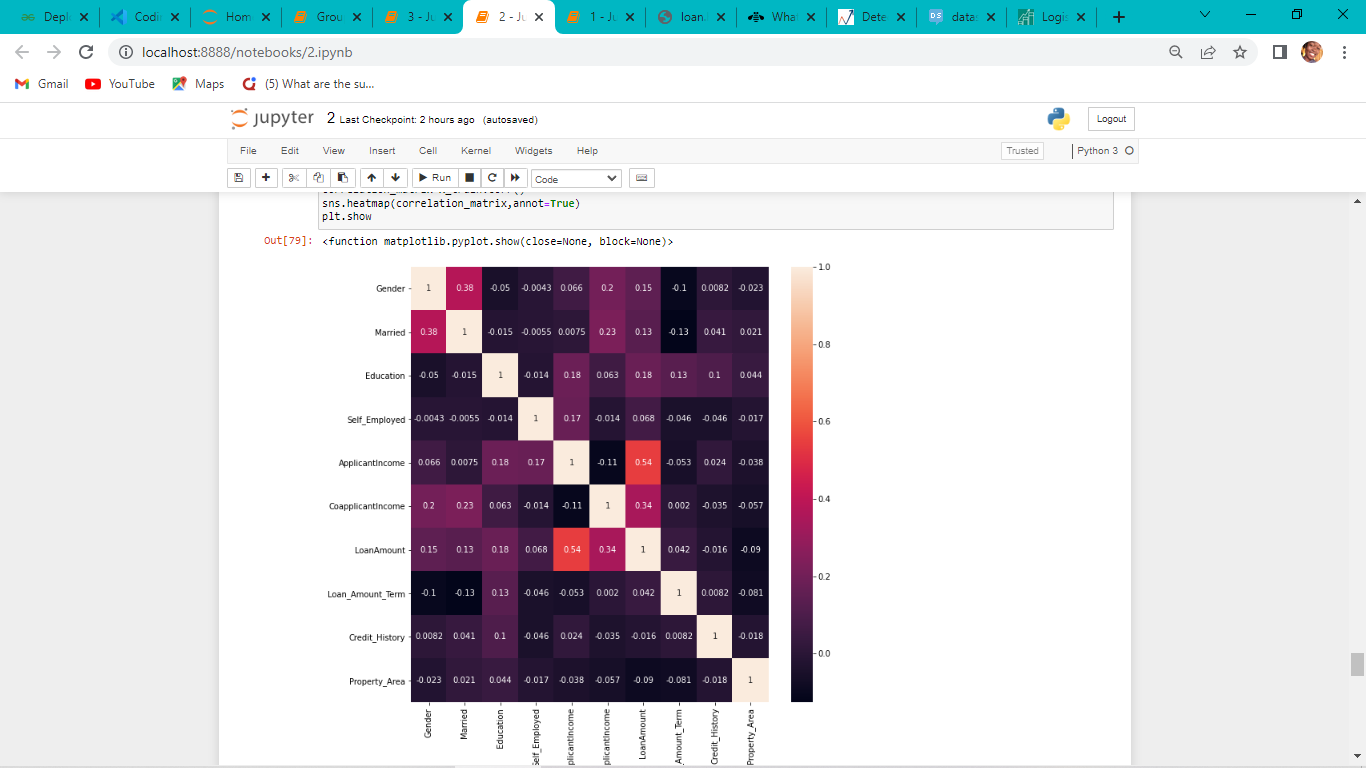
Around 65% of the applicants in the dataset are married.

Around 15% applicants in the dataset are self employed.

Around 85% applicants have credit history (repaid their debts).

Around 80% of the applicants are Graduate.

The below map

**Algorithm Definition**

The algorithm that we have decided to use for this purpose is Logistic Regression. It falls under the supervised machine learning algorithms category. Logistic regression is a classification algorithm used to find the probability off event success and event failure. It is a statistical analysis method used to predict a binary outcome such as yes or no based on observations of a data set. It is used to predict the output of a categorical dependent variable therefore the outcome must be a categorical or discrete value, being 0 and 1, No or Yes [6].

Pseudocode

The following shows the pseudocode for the proposed loan prediction method

1. Import libraries
2. Load the data
3. Determine the training and testing data
4. Data cleaning and pre-processing
5. Fill the missing values with median, mean, mode and forward fill
6. Apply modelling for prediction
7. Determine the accuracy for the model applied

**System Features**

• Data collection.

• Data cleaning and preprocessing

• Model selection

• Data verification

• Classification.

• Report deliver.

**System Architecture**

This is a loan prediction architecture



Implementation Details (Modules):

Loan Dataset: Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset the system will automatically predict which costumer’s loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.

Determine the training and testing data: Typically , Here the system separate a dataset into a training set and testing set ,most of the data use for training ,and a smaller portions of data is used for testing. After a system has been processed by using the training set, it makes the prediction against the test set.

Data cleaning and processing: In Data cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative.

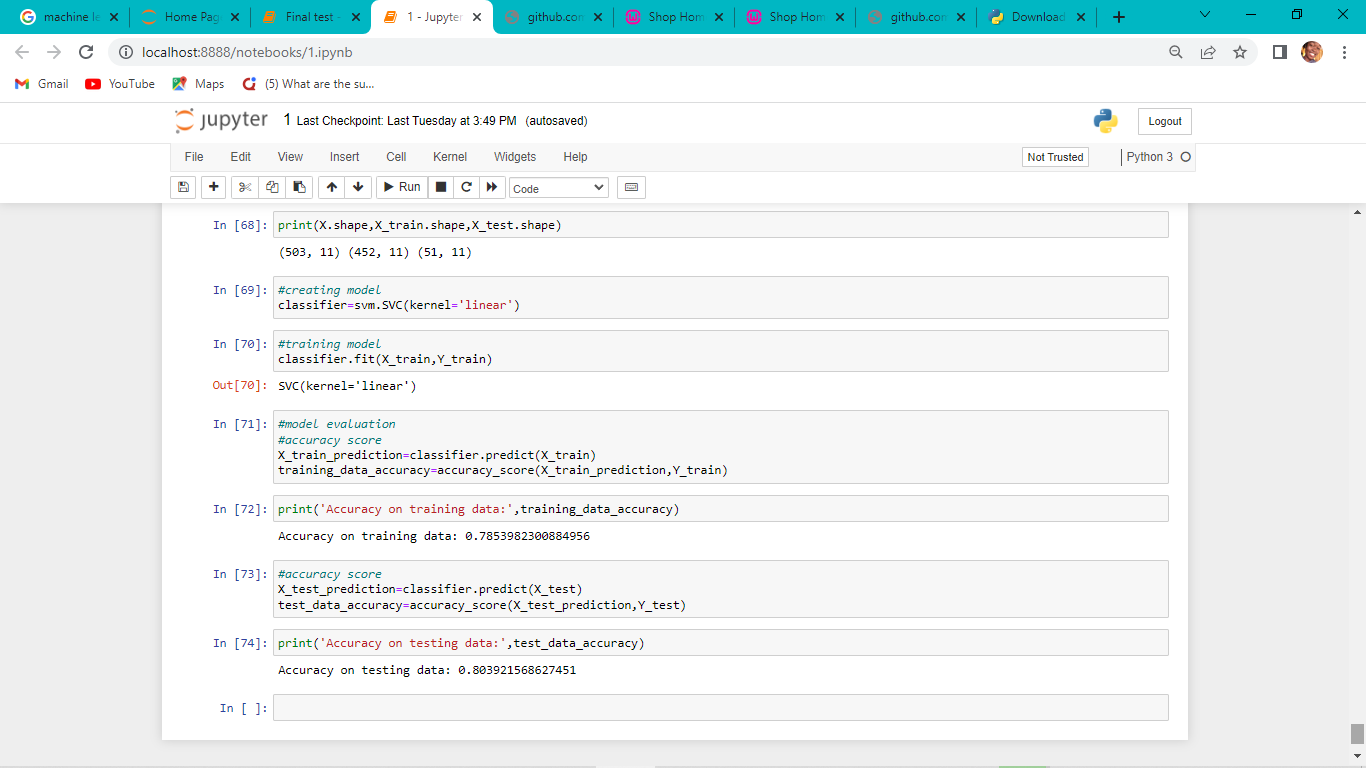
**Methodology**

The proposed model focuses on predicting the credibility of customers for loan application by analysing their information. The input to the model is the customer information collected. On the output from the classifier, decision on whether to approve or reject the customer request can be made. In this process it is required to train the data using different algorithms and then compare user data with train data to predict the nature of the loan. Every new applicant details filled at the time of application form acts as a test. After the operation of testing, model predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it conclude on the basis of the training data sets to extract important information predicts if a customer would be able to repay his loan or not.

**Model selection**

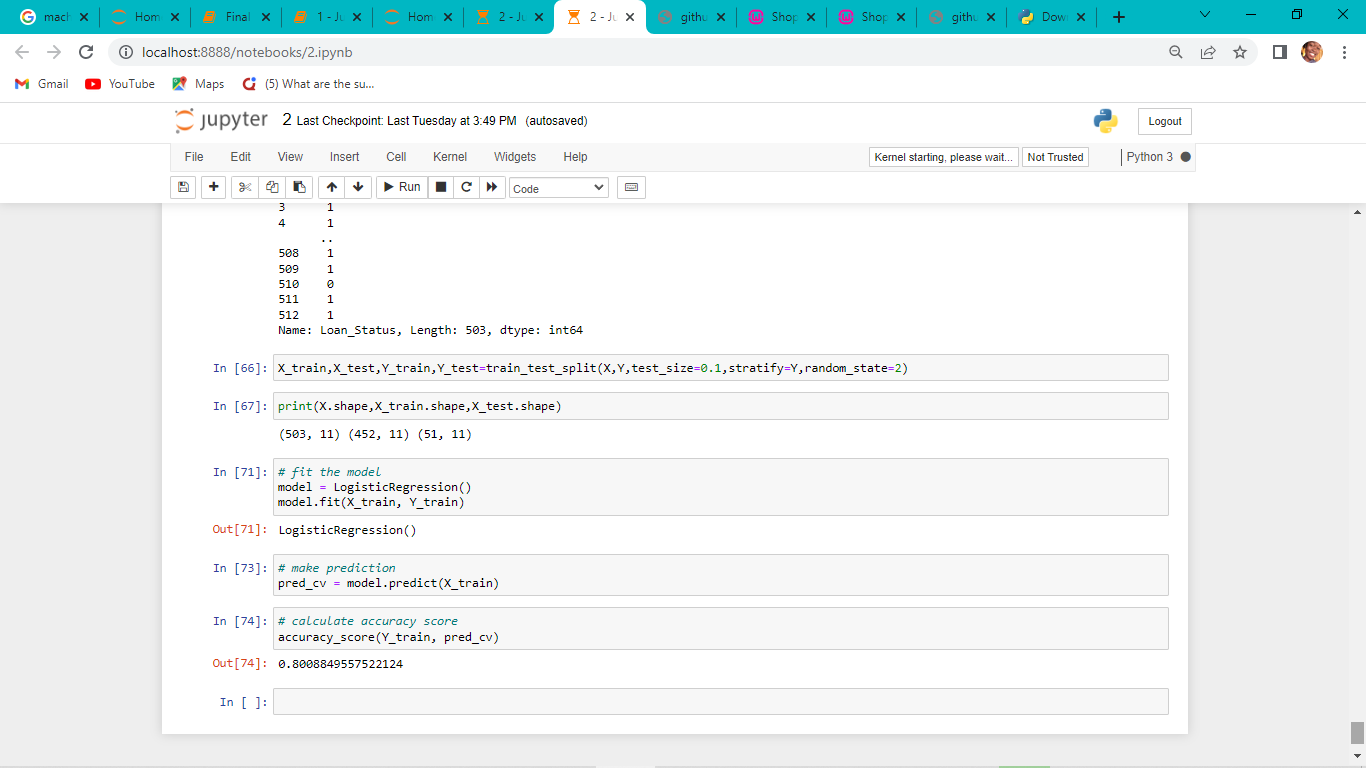
SVM: In this approach, each data item is plotted in a n-dimensional space, where n represents the number of features with each feature represented in a corresponding co-ordinates [7].

Fig.1 The below image indicates the accuracy score while using Support Vector Model (SVM)



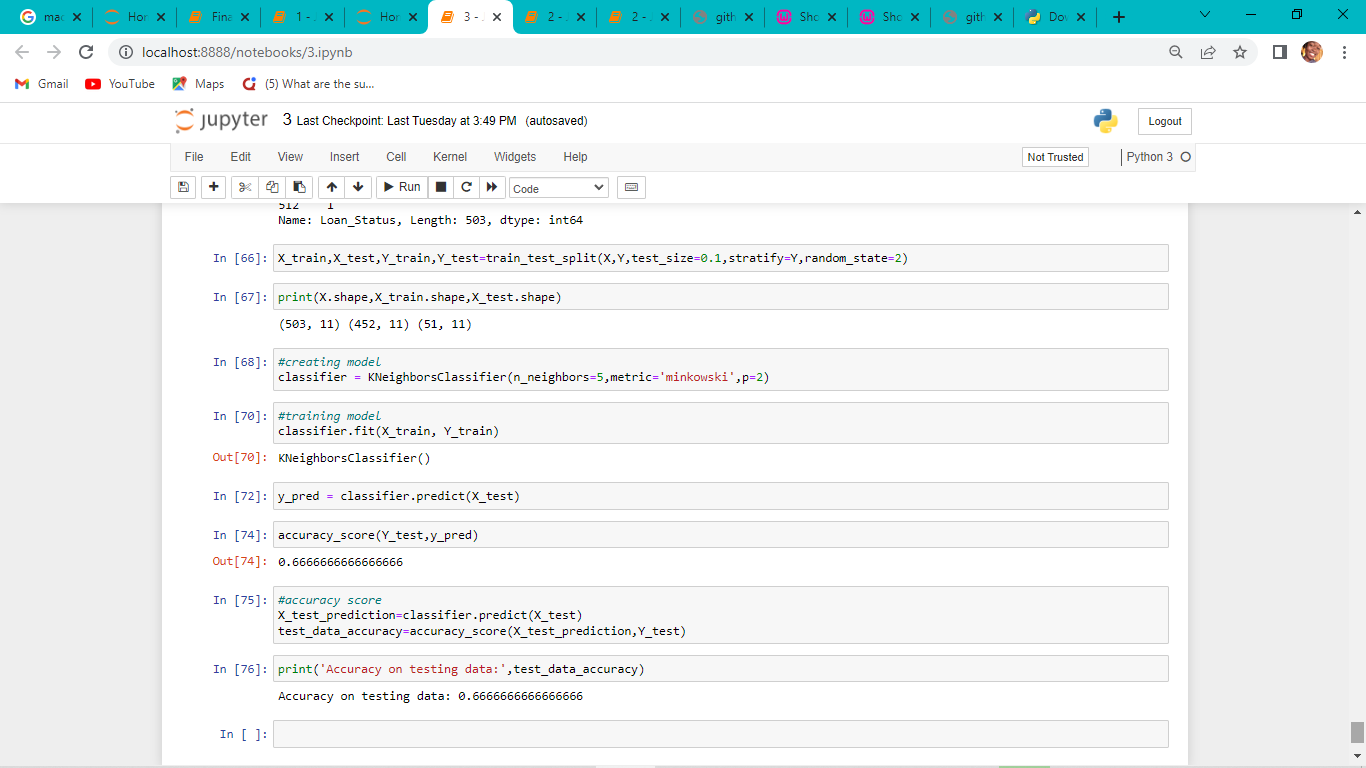
Logistic regression is supervised learning classification algorithm such that we are able to anticipate the yield values for new information based on those connections, which it learned from the previous information sets [6].

Fig.2 The below image indicates the accuracy score while using Logistic Regression



The KNN algorithm is a simple supervised machine learning algorithm that can be utilized to unravel both classification and replace issues. It is easy to implement and understand but significantly slows as the size of that data on use grows [8].

Fig.3 The below image indicates the accuracy score while using KNN.



Results (Discussion)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model name | Accuracy | Precision | Recall score | F1-score |
| SVM | 78% | 79.1% | 97.1% | 87.2% |
| Logistic Regression | 80% | 82.9% | 97.1% | 89.5% |
| KNN | 66% | 71.4% | 85.7% | 77.9% |

Accuracy: It’s worn to find the portion of correctly classified values. It is tell us how often our classifier is right. Sum of all true values divided by total values. Number of classified samples =TP+TN

Formula for calculating the accuracy

(TP + TN)

(TP + FP + TN + FN)

Precision: It is used to calculate the models ability to classify positive values correctly. Number of classified values = TP Number of actual values = TP+FP

Precision

TP

TP + FP

Recall: To calculate the models ability to predict positive values

TP

TP + FN

Discussion

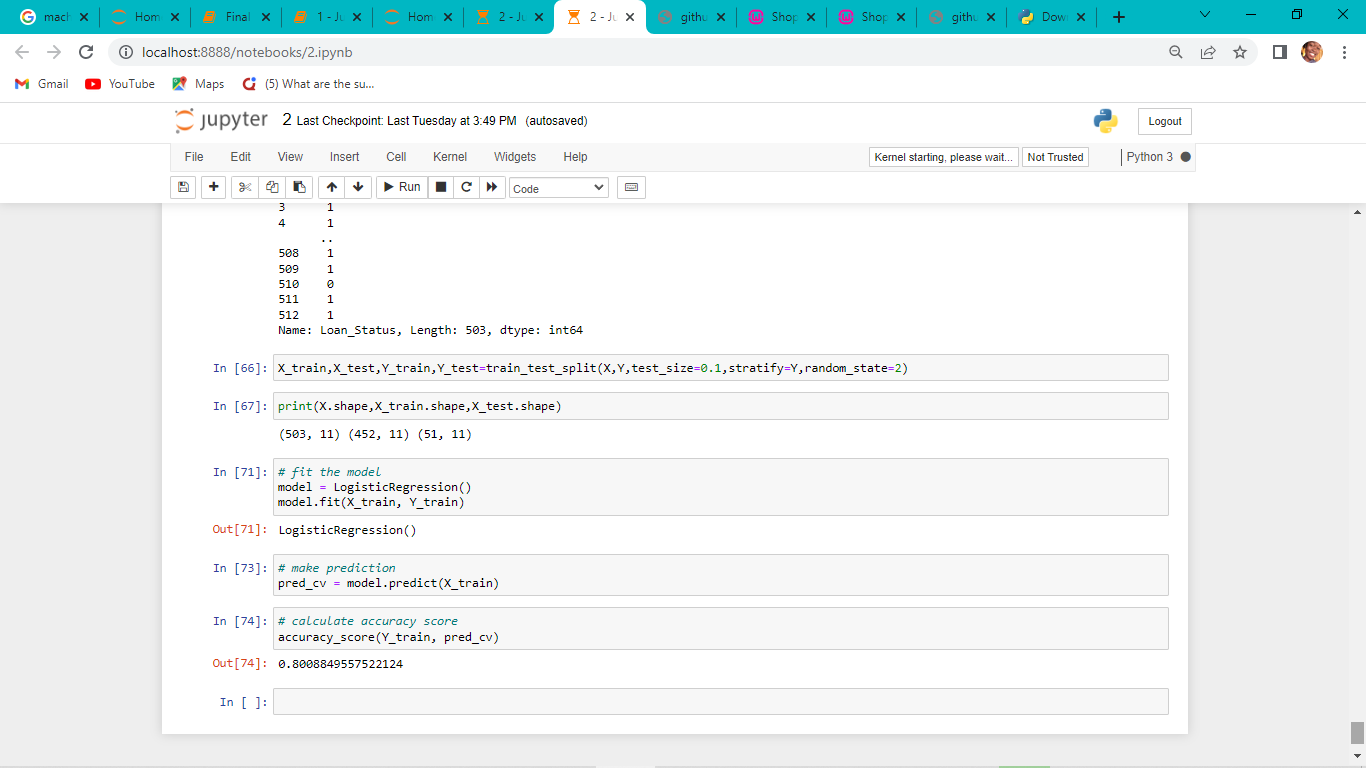


Fig.4

Strengths

* Easier to implement, interpret and very efficient to train
* It not only provides a measure of how appropriate a predictor is, but also its direction of association.
* Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
* It is very fast at classifying unknown records.

Weaknesses

* The major limitation of the logistic regression is the assumption of linearity between the dependent variable and the independent variable
* It can only be used to predict discrete functions, hence the dependent variable of the logistic regression is bound to discrete number set

So here, it can be concluded with confidence that the logistic regression model is extremely efficient and gives a better result when compared to other models. It works correctly and fulfills all requirements of bankers. This system properly and accurately calculate the result. It predicts the loan is approve or reject to loan applicant or customer very accurately.

**Use Case Diagram**

The diagram below depicts the relationship between the use cases, actors and the system.



**CONCLUSION**

In this report, we have proposed customer loan prediction using supervised learning techniques for loan candidate as a valid or fail to pay customer. Various algorithms were implemented to predict customer loan. Optimum results were obtained using Logistic Regression, KNN, and SVM. Compared those three algorithms and logistic regression has the high accuracy. From a correct analysis of positive points and constraints on the part, it can be safely ended that the merchandise could be an extremely efficient part. In close to future this module of prediction can be integrated with the module of machine-driven processing system.

# References

|  |  |
| --- | --- |
| [1] | T. Massaoudi, “towardsdatascience.com,” 06 June 2019. [Online]. Available: https://towardsdatascience.com/ml-basics-loan-prediction-d69ba7f31f6. [Accessed 25 May 2022]. |
| [2] | P. Baheti, “v7labs.com,” 26 May 2022. [Online]. Available: https://www.v7labs.com/blog/data-preprocessing-guide. [Accessed 28 May 2022]. |
| [3] | N. Tamboli, “analyticsvidhya.com,” 29 October 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/. [Accessed 26 May 2022]. |
| [4] | D. Lemonaki, “freecodecamp.org,” 24 August 2021. [Online]. Available: https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics. [Accessed 23 May 2022]. |
| [5] | K. Shetty, “analyticsvidhya.com,” 30 September 2020. [Online]. Available: https://analyticsvidhya.com/blog/2020/09/hypothesis-generation-data-science-projects/#:~:text=Trip%20Duration%Prediction-,What%20is%20Hypothesis%20Generation%3F,generated%20based%20on%20any%20evidence. [Accessed 29 May 2022]. |
| [6] | J. Brownlee, “machinelearningmastery.com,” 01 April 2016. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning. [Accessed 01 June 2022]. |
| [7] | M. McGregor, “freecodecamp.org,” 01 July 2020. [Online]. Available: https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explauned-with-code-examples/. [Accessed 30 May 2022]. |
| [8] | O. Harrison, “towardsdatascience.com,” 10 September 2018. [Online]. Available: https://www.towardsdatascience.com/machine--basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761. [Accessed 01 June 2022]. |

**GROUP MEMBERS**

LIEKETSENG MOTSOHI 901010922

MPHO SESOANE 901011059

MANTS’E MAKATENG 901010847

MOTS’ELISI MOKHOEMA 901010825