DSC530
DATA EXPLORATION AND ANALYSIS

Professor : Fadi Alsaleem
Week 12 - Term Project
First name : Saravanan
Last Name : Janarthanan

Introduction

Credit Risk analysis

Lending involves the provision of funds, often in the form of a loan, to individuals, businesses, or other entities, with the anticipation that the borrowed capital will be returned with interest or in adherence to the conditions laid out in a loan agreement. This financial practice is a widespread and pivotal element of the economy.

Loan default takes place when a borrower does not adhere to the specified terms and conditions outlined in the loan agreement. This often involves a failure to make timely payments or a breach of other contractual commitments.

The perpetual chance that a borrower might fail to meet or cease making payments leads to accurately assessing credit risk and it is crucial. Credit risk analytics transforms both historical and projected data into actionable analytical insights, empowering financial institutions to evaluate risk and determine lending and account management strategies. One approach employed by organizations is the integration of credit risk modeling into their decision-making processes.

To analyze the credit risk for awarding loan , various input features are required

- **Person Age** - Age of the individual applying for the loan. Age can be an indicator of stability and reliability in repaying loans.
- **Person Income** - Income level or earnings of the individual. Income is a crucial factor in determining the individual's ability to repay the loan.
- **Previous credit history** - Record of the individual's past borrowing and repayment behavior.
- **Collateral** - Assets offered by the borrower as security for the loan. Collateral serves as a form of protection for lenders in case of default
- **Intent of loan** - Purpose or reason for seeking the loan. The intended use of funds can influence risk assessment.
- **interest rate** - Rate at which interest is charged on the loan amount. The interest rate reflects the cost of borrowing and impacts the affordability of the loan for the borrower

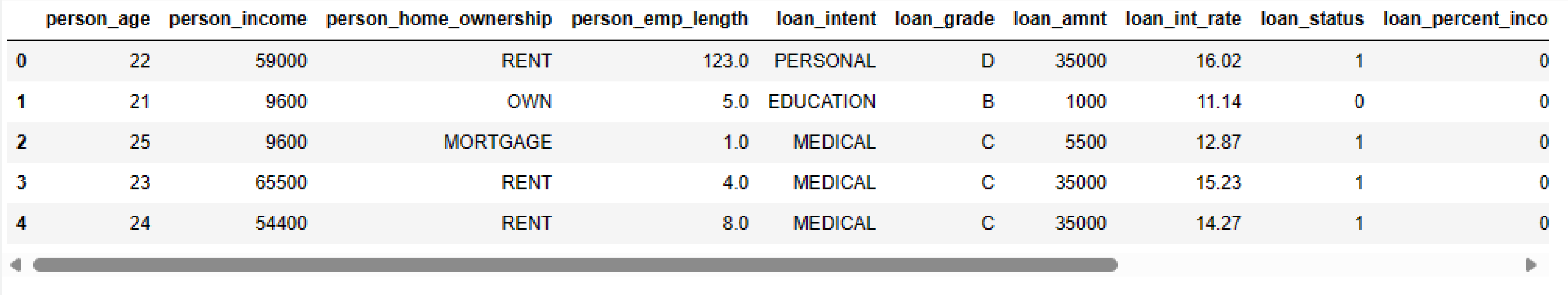
Dataset

The below Credit Risk Analysis dataset will be used

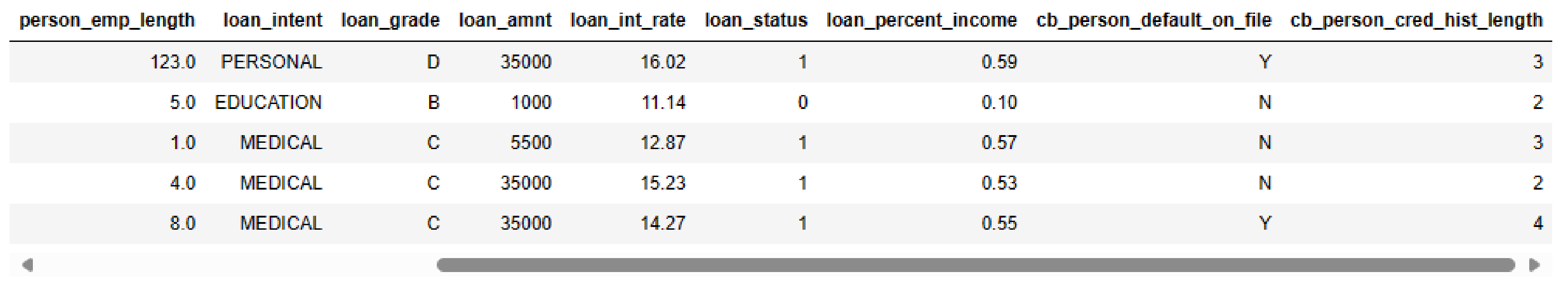
Feature Name	Description
person_age	Age of the person
person_income	Annual Income of the person
person_home_ownership	Home ownership - whether person owns home or mortgage or rented
person_emp_length	Employment length (in years)
loan_intent	Loan intent - intent of the loan
loan_grade	Loan grade - scale based on credit worthiness
loan_amnt	Loan amount
loan_int_rate	Interest rate for the loan
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income - % of income represented by loan
cb_person_default_on_file	Historical default (previous default history from Bureau)
cb_preson_cred_hist_length	Credit history length

Source : https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data

Top 5 rows of the dataset



	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_inco
0	22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1	0
1	21	9600	OWN	5.0	EDUCATION	B	1000	11.14	0	0
2	25	9600	MORTGAGE	1.0	MEDICAL	C	5500	12.87	1	0
3	23	65500	RENT	4.0	MEDICAL	C	35000	15.23	1	0
4	24	54400	RENT	8.0	MEDICAL	C	35000	14.27	1	0



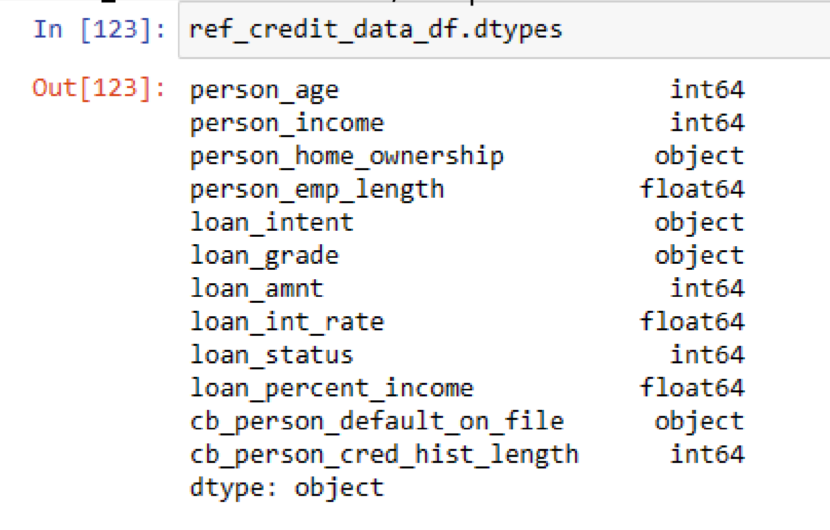
person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
123.0	PERSONAL	D	35000	16.02	1	0.59	Y	3
5.0	EDUCATION	B	1000	11.14	0	0.10	N	2
1.0	MEDICAL	C	5500	12.87	1	0.57	N	3
4.0	MEDICAL	C	35000	15.23	1	0.53	N	2
8.0	MEDICAL	C	35000	14.27	1	0.55	Y	4

The dataset has 32581 rows and 12 columns or features

Following are the categorical variables or features

- person_home_ownership
- loan_intent
- loan_grade
- cb_person_default_on_file

- loan_status - Predictor / response variable



```
In [123]:  ref_credit_data_df.dtypes

Out[123]:  person_age                      int64
           person_income                   int64
           person_home_ownership          object
           person_emp_length             float64
           loan_intent                    object
           loan_grade                     object
           loan_amnt                       int64
           loan_int_rate                 float64
           loan_status                     int64
           loan_percent_income           float64
           cb_person_default_on_file      object
           cb_person_cred_hist_length      int64
           dtype: object
```

Target Hypothesis

Dependent Variable : loan_status this feature has value 0 and 1, 0 indicates loan approved and 1 indicates loan denied.

Statistical Question : What is the statistical impact of independent variables on the dependent variable to determine the loan status.
This will help to determine to predict if a new applicants loan will be approved.

Null Hypothesis is that independent variables does not have statistical impact on dependent variable loan status

The data will be cleaned for incorrect values , outliers, factual errors, missing values etc.

Below five variables will be analyzed and are critical for analyzing the credit worthiness

- person_age - Age of the applicant, Age can be an indicator of stability and reliability in repaying loans. Younger individuals might have fewer financial responsibilities but less established credit histories, while older individuals might have more stability but potentially fixed incomes.

- person_income - Income of the applicant, Income is a crucial factor in determining the individual's ability to repay the loan. Higher income generally indicates a greater capacity to meet financial obligations.

- Home_ownership - Home ownership type, Collateral serves as a form of protection for lenders in case of default. It can mitigate risk by providing recourse to recover losses if the borrower fails to repay the loan.

- loan_amnt - Loan amount sought, is a fundamental aspect of credit risk analysis. It represents the principal balance that the borrower is obligated to repay. Higher loan amounts may pose higher risks for lenders, especially if they exceed the borrower's capacity to repay based on their income and other financial factors.

- loan_percent_income - Loan percent ratio to income, provides insights into the borrower's debt burden relative to their earnings. A high ratio indicates that a significant portion of the borrower's income is allocated to debt repayment, which may increase the risk of default. Lenders often use this ratio to assess the borrower's ability to manage additional debt responsibly.

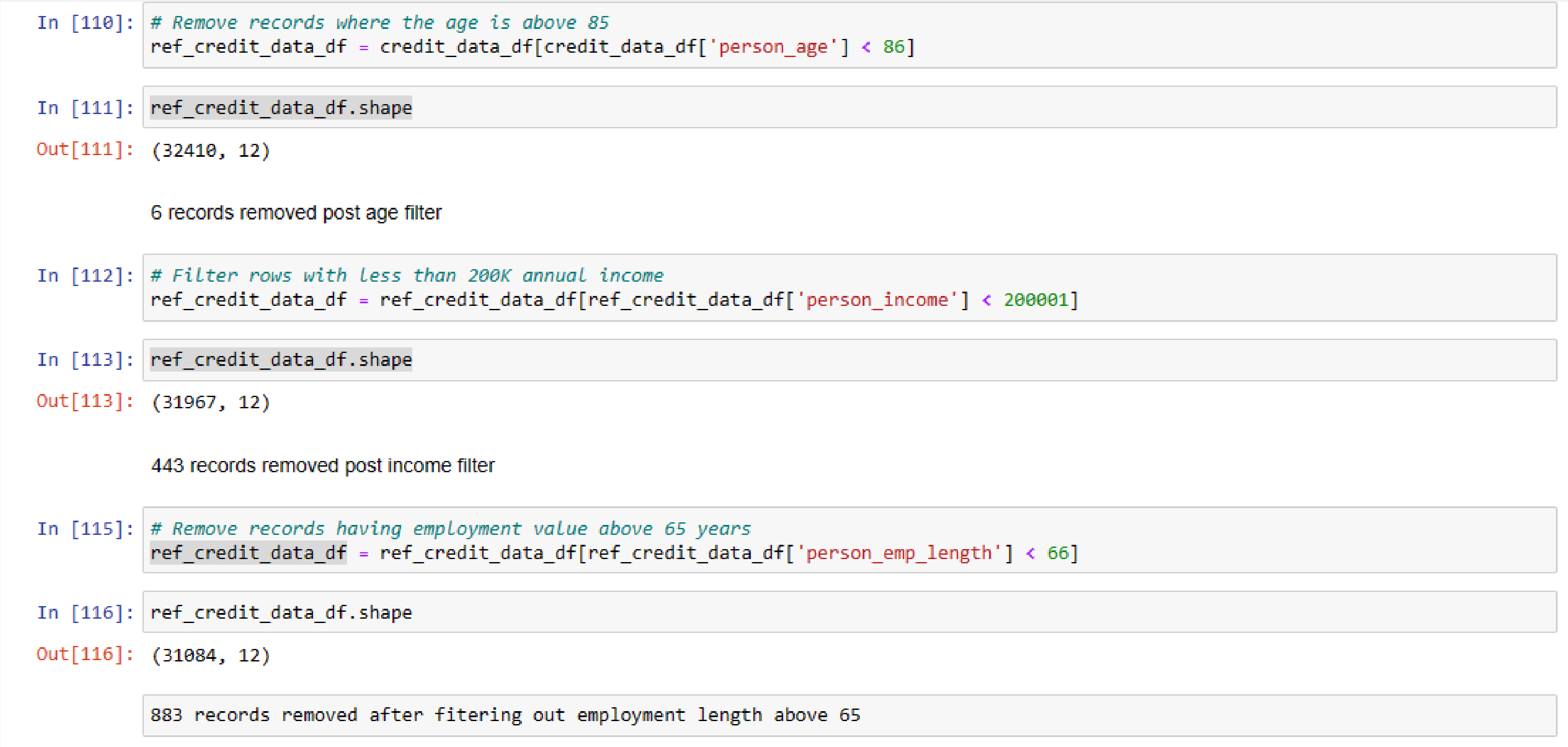
Handling Outliers

From the overall five point summary , There are outliers are data entry errors like

- Age above 100 years
- Loan seeking person having a income up to 600K
- Employment length well above 65 years

Need to cleanup the data for all the above three conditions

- Remove records where the age is above 85 (Probability of Persons above age 85 seeking a loan is very low)
- Retain only rows with less than 200K annual income (Probability of Persons earning above seeking loan for 35K is very low)
- Remove records having employment value above 65 years (with min age at 15 will lead to age at retirement as 80)



```
In [110]: # Remove records where the age is above 85
          ref_credit_data_df = credit_data_df[credit_data_df['person_age'] < 86]

In [111]: ref_credit_data_df.shape

Out[111]: (32410, 12)
```

6 records removed post age filter

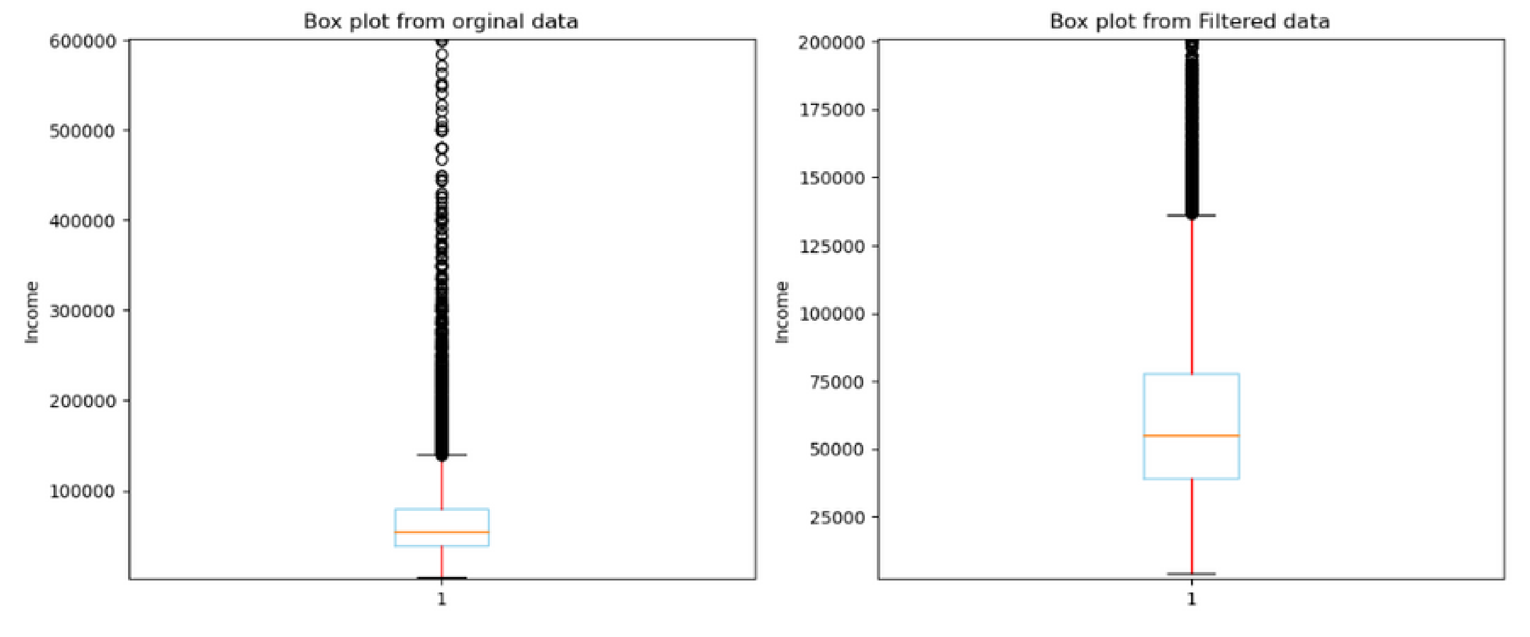
```
In [112]: # Filter rows with less than 200K annual income
          ref_credit_data_df = ref_credit_data_df[ref_credit_data_df['person_income'] < 200001]

In [113]: ref_credit_data_df.shape

Out[113]: (31967, 12)
```

443 records removed post income filter

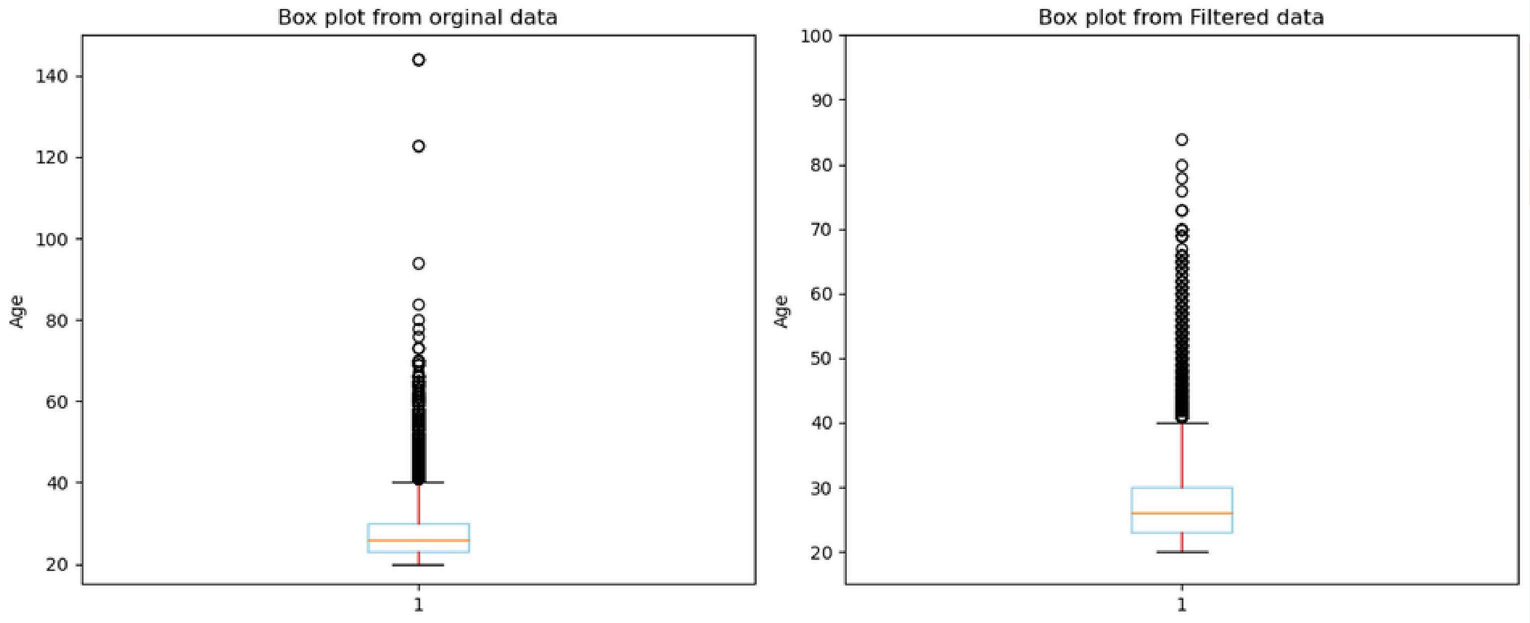
```
In [115]: # Remove records having employment value above 65 years
          ref_credit_data_df = ref_credit_data_df[ref_credit_data_df['person_emp_length'] < 66]

In [116]: ref_credit_data_df.shape

Out[116]: (31084, 12)

          883 records removed after fitering out employment length above 65
```

Outliers displayed in Box plot

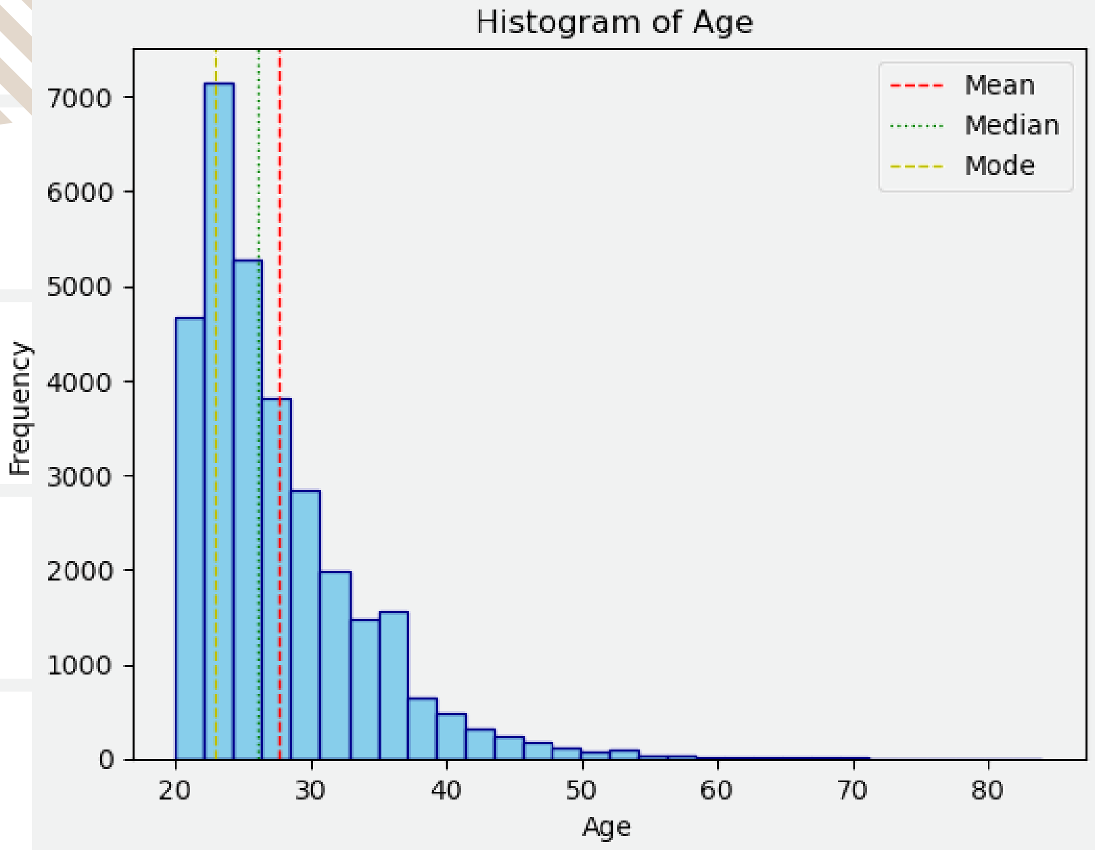
Personal income Outliers before and after cleaning outliers

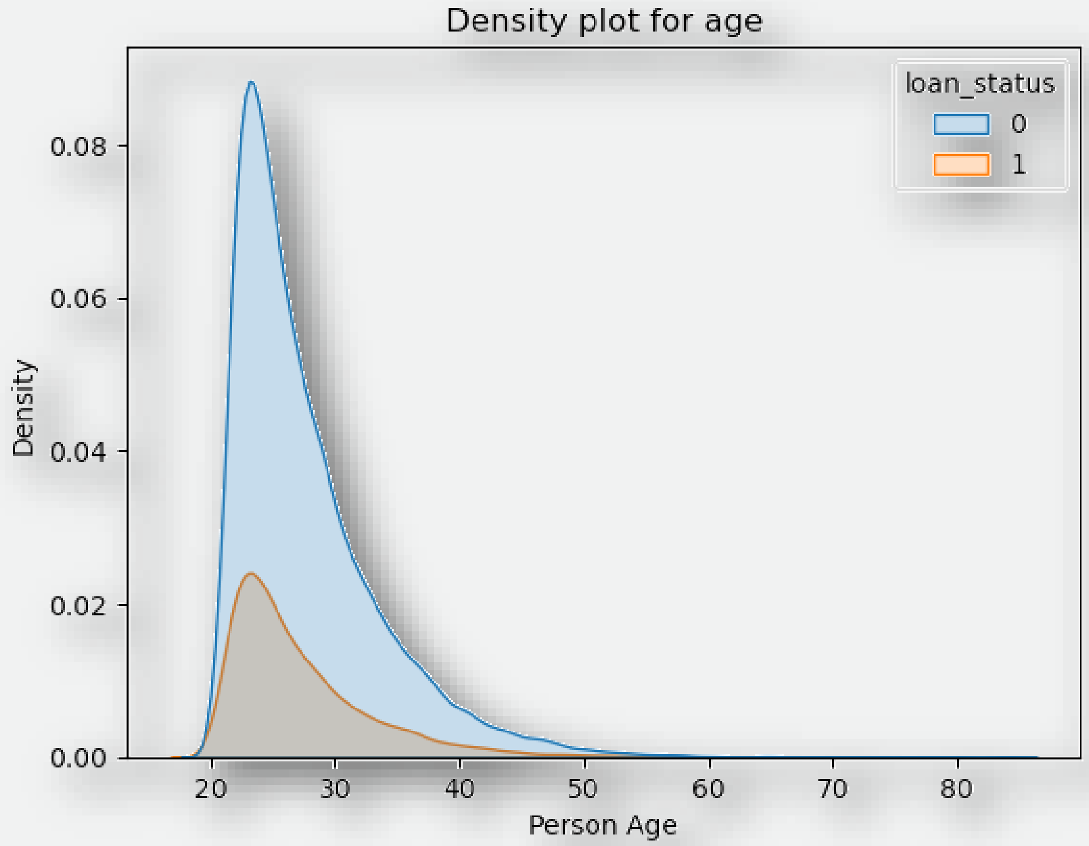


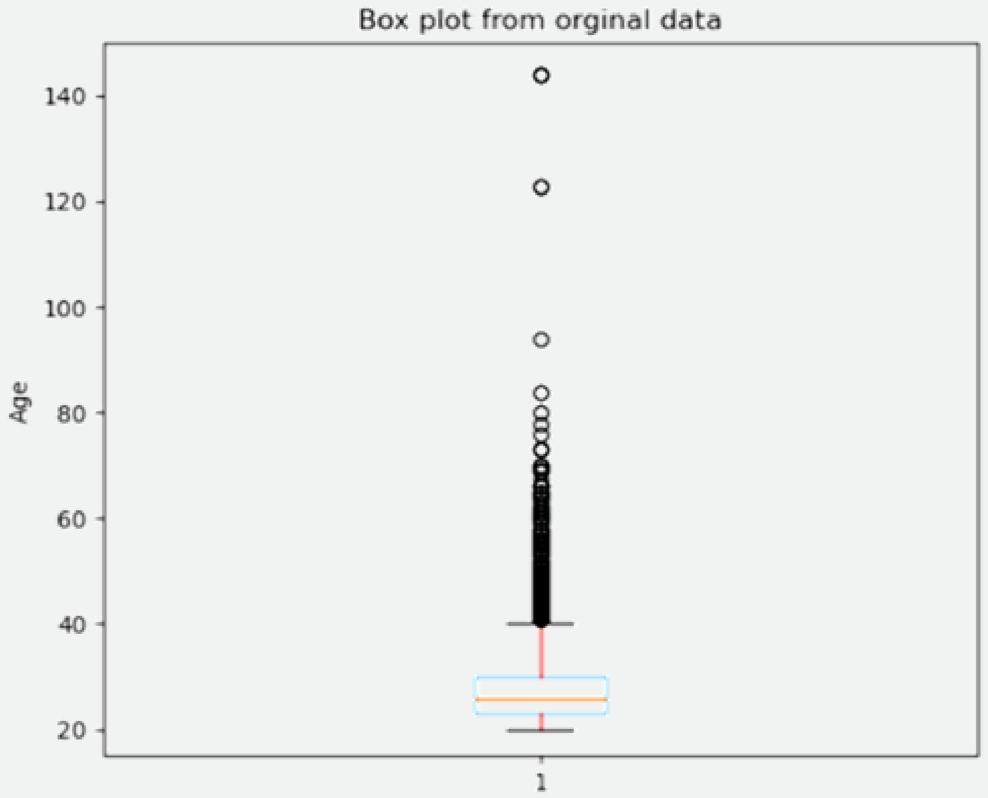
Personal Age Outliers before and after cleaning outliers



Person Age







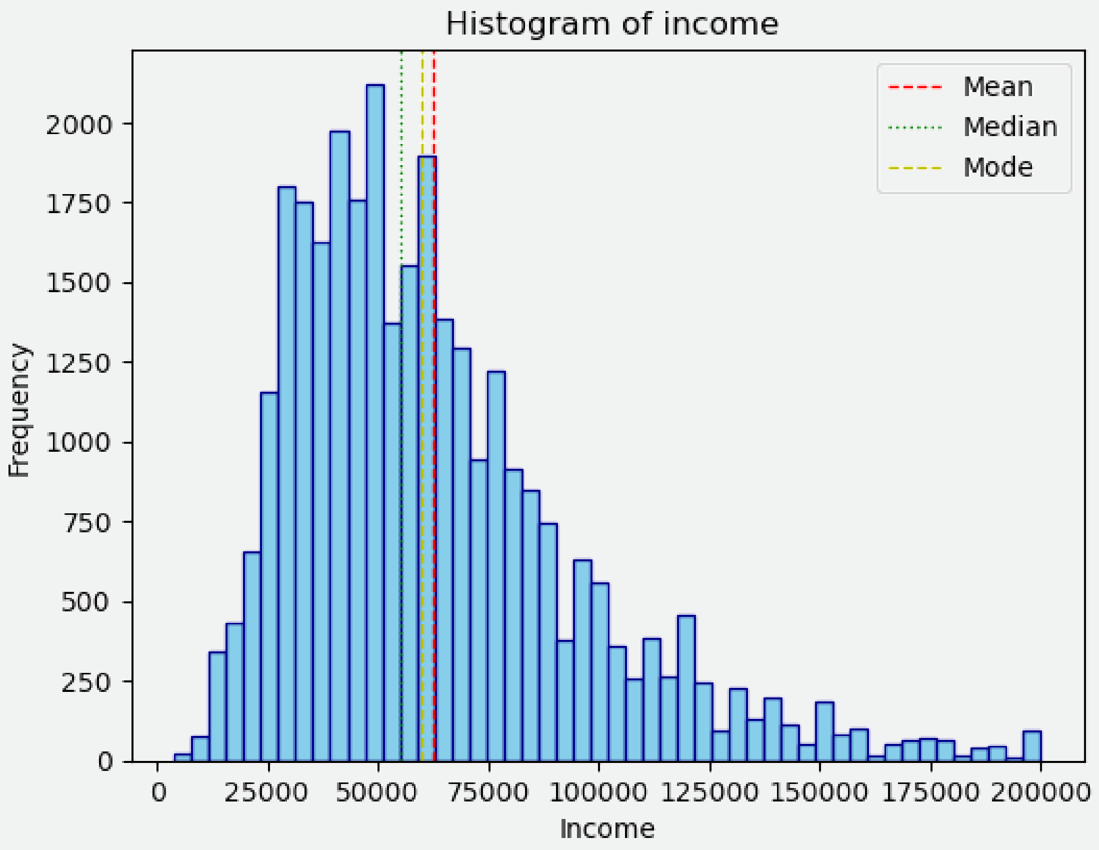
5 Point Summary

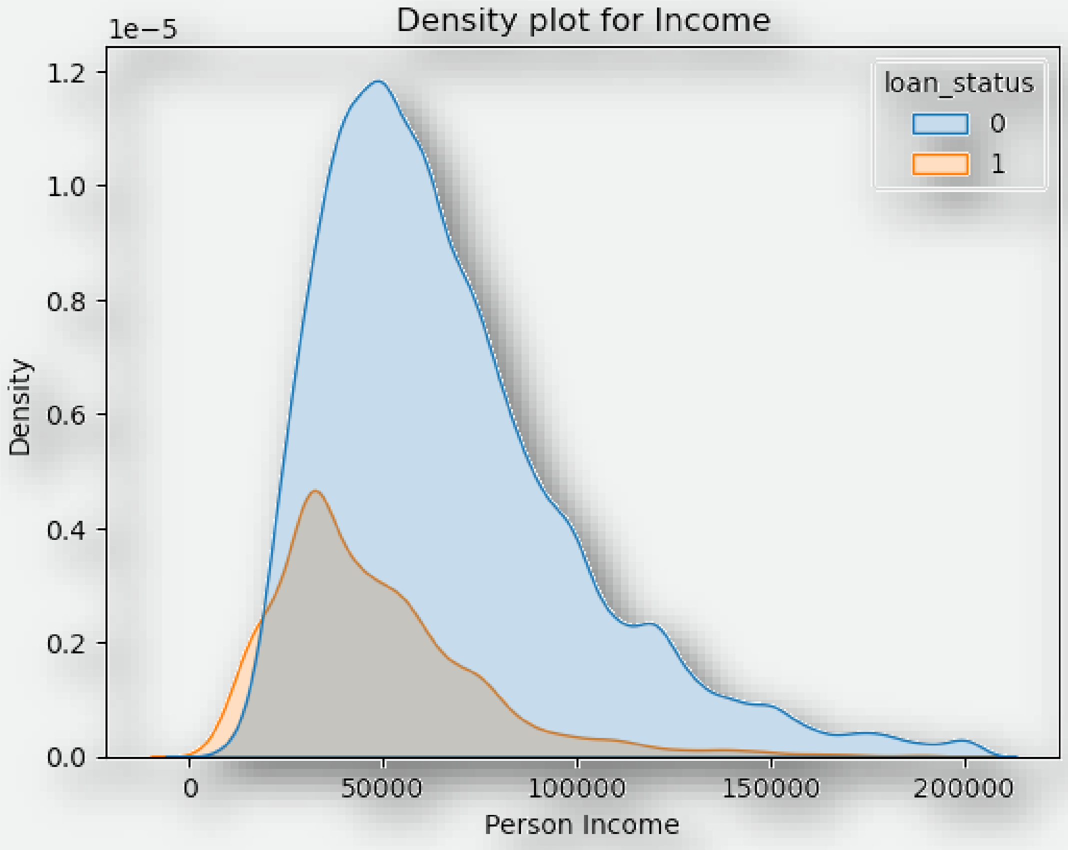
count	31084
mean	27.68
std	6.17
min	20.00
25%	23.00
50%	26.00
75%	30.00
max	84.00
Mode	23.00

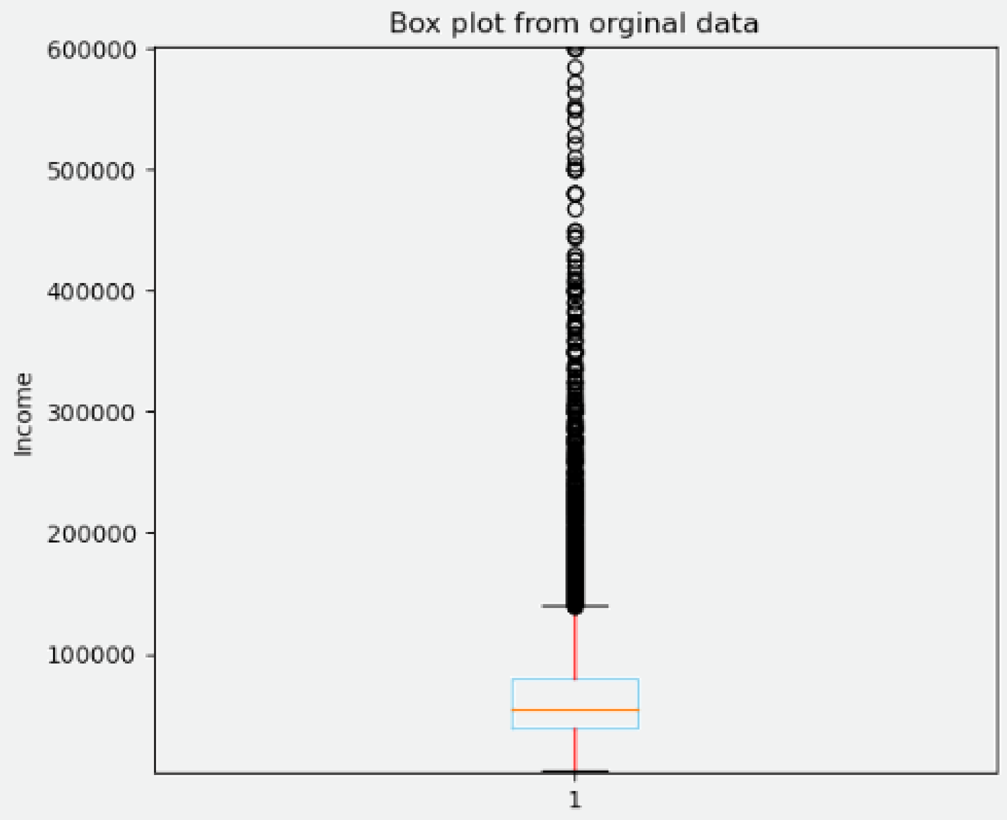
Observations

- Even though most of the applicant age, MODE, is close to 23, the MEDIAN and MEAN values are located away or their values are more than MODE.
- The distribution is again skewed to the right with long tail as visible in the density, box plot and histogram
- From the density plot young people loan applications are getting approved and hence lower age has a high probability of getting their loan approved.
- Outliers: The age had a range from 20 to 144, remove records where the age is above 85 (Probability of Persons above age 85 seeking a loan is very low and age above 100 is more than average life span)

Person Income







5 Point Summary

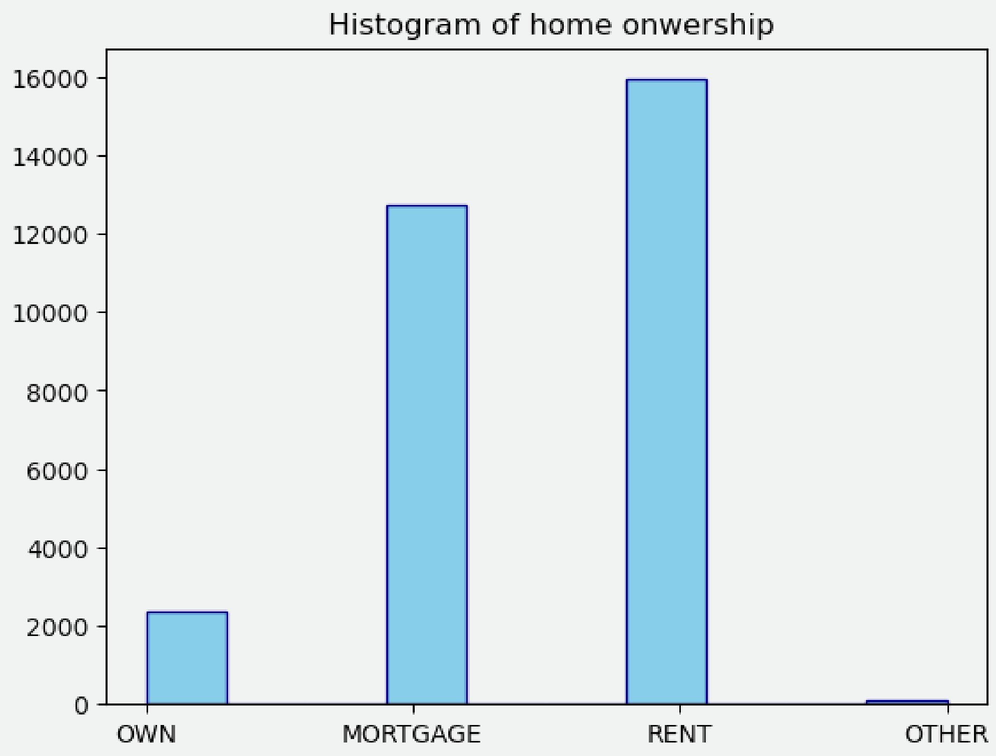
count	31084
mean	62711.41
std	33105.59
min	4000.00
25%	39000.00
50%	55000.00
75%	78000.00
max	200000.00
Mode	60000.00

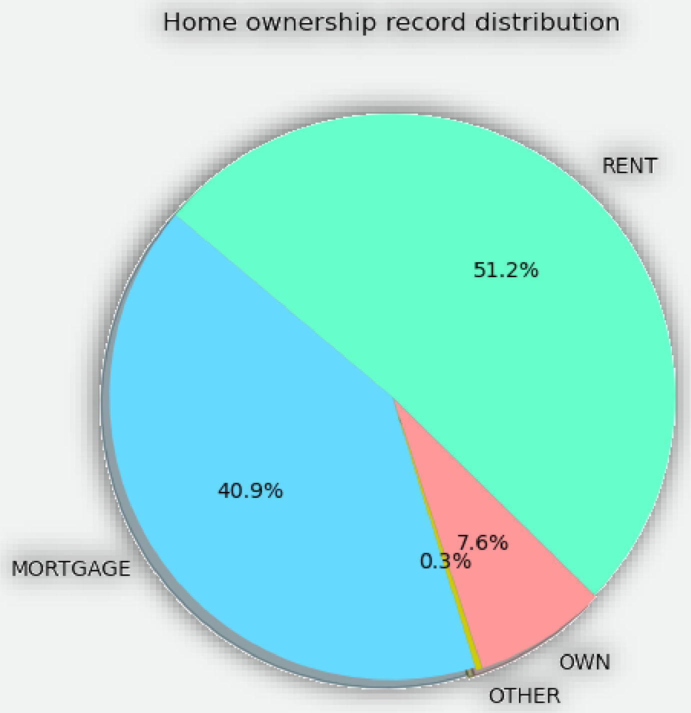
Observations for person income feature

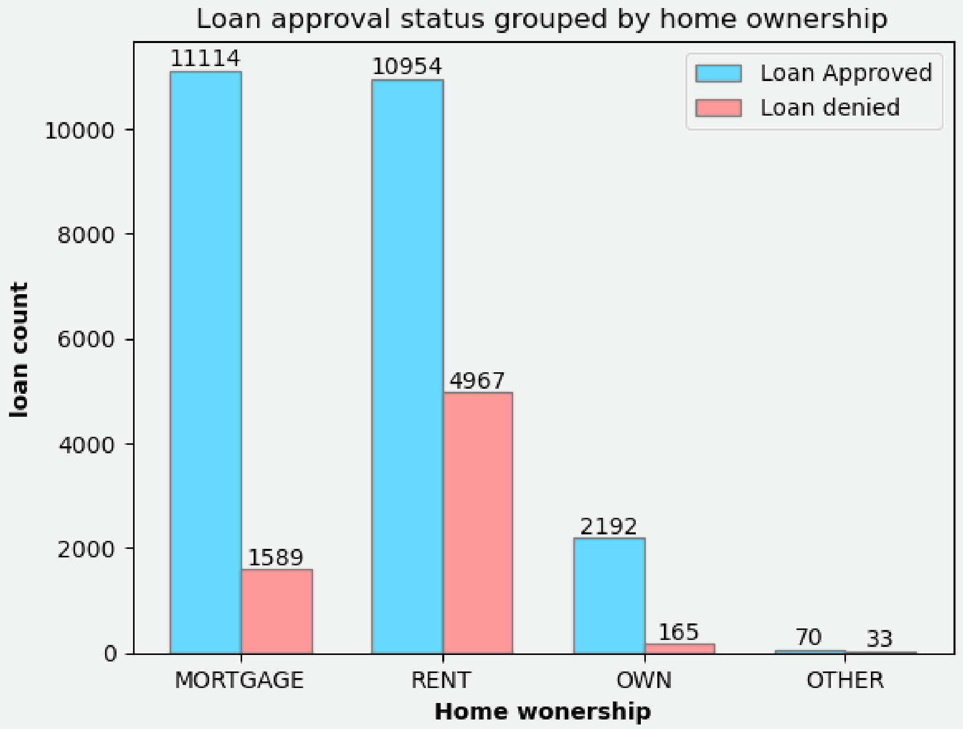
From the initial data summary , it was evident that the data distribution was skewed right with a very long tail. After filtering the dataset with income less than 200K, the distribution is right skewed

- Mean value of 62711.41 is close but located on the right of the median or 2nd quartile confirming the right skew
- Mode value lies between mean and median indicating most of the values are around the mean
- The tails on both side extend beyond 1 standard deviation of 33105.59 on both sides, more than 2 standard deviation on the right side
- Outliers – income ranges from 4K to 600K , Filter rows with less than 200K annual income (Probability of Persons earning above seeking loan for 35K is very low)

Home Ownership







<u>Home Ownership</u>

RENT	15921
MORTGAGE	12703
OWN	2357
OTHER	103

<u>Observations</u>

- Overall home ownership data is unbalanced as mortgage and rented records are available at a higher proportion (91% of total records) than owned and other categories

- Home owners (OWN category) and Mortgaged owners had a higher probability or chance to a get an loan approved.

- 90% of the OWN category had their loan approved. But due to less number of data available, it may be overfitting.

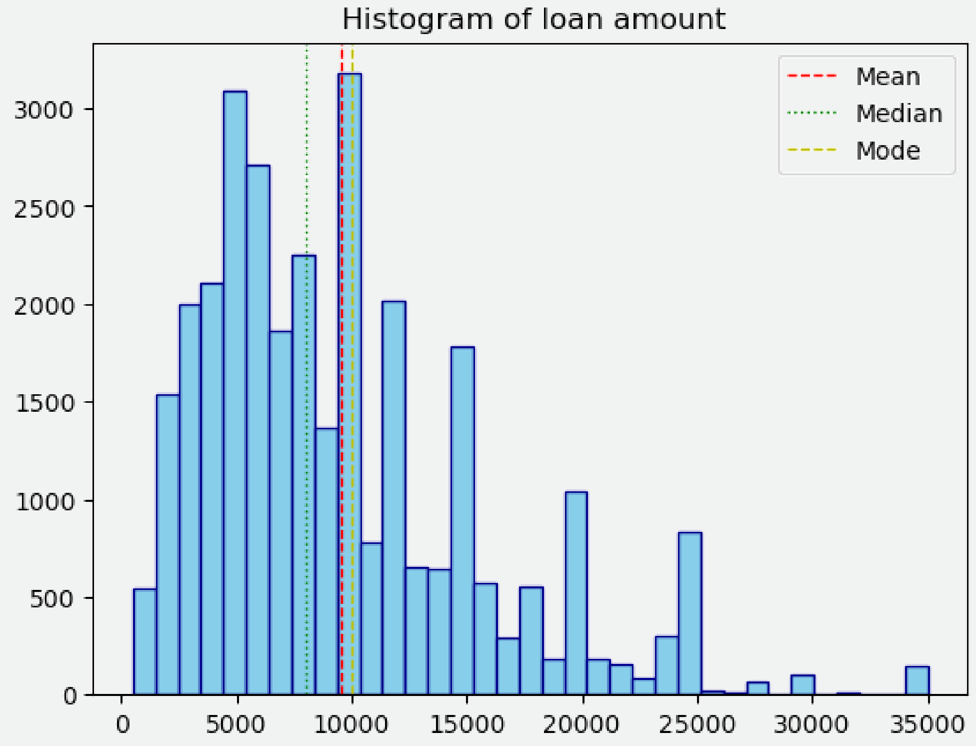
- Mortgaged home owners has a 85% probability of getting loan approved

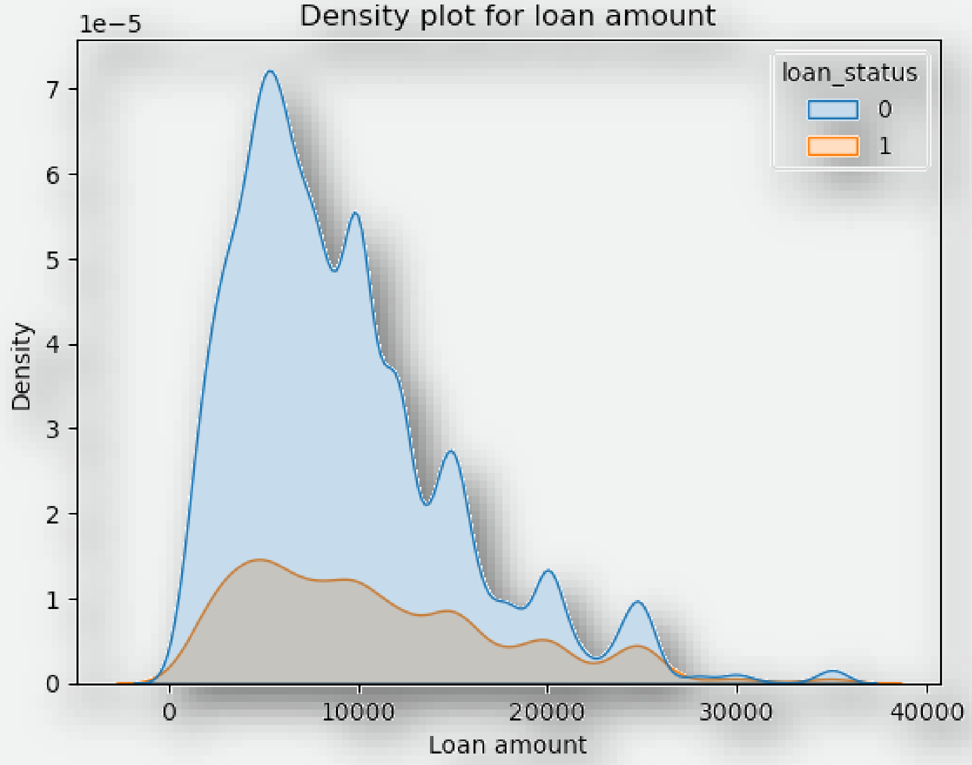
- Rented people have a 69% chance of getting the loan approved.

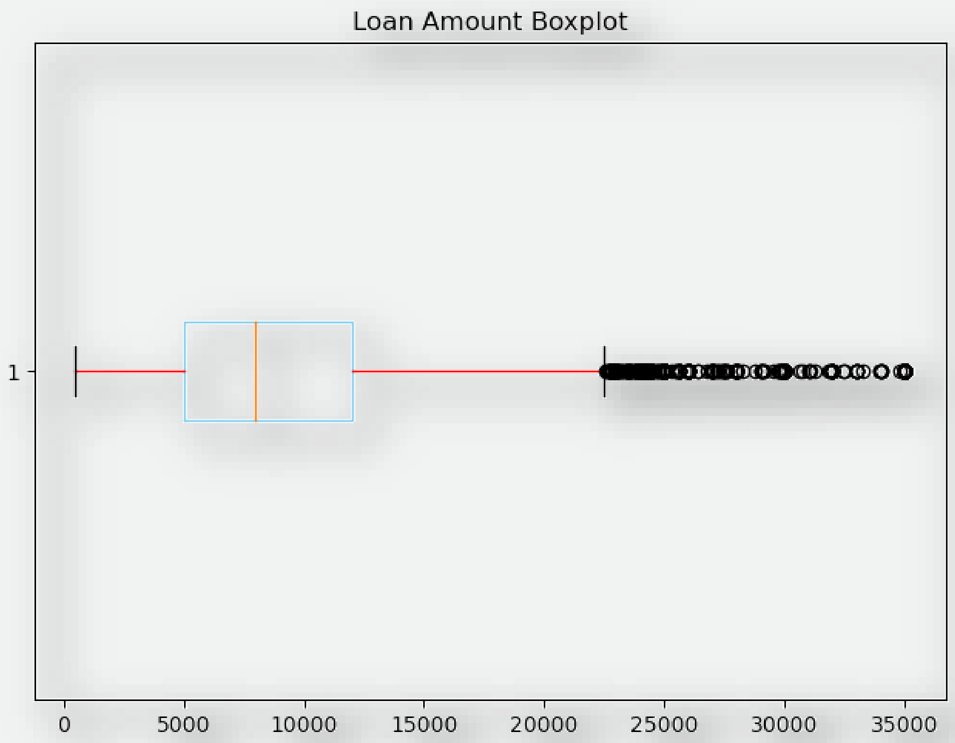
- Loan was approved for 68% of the OTHER category people.

Note: Since Home ownership is a categorical value, providing the value count of unique values

Loan Amount


Histogram of loan amount


Density plot for loan amount


Loan Amount Boxplot

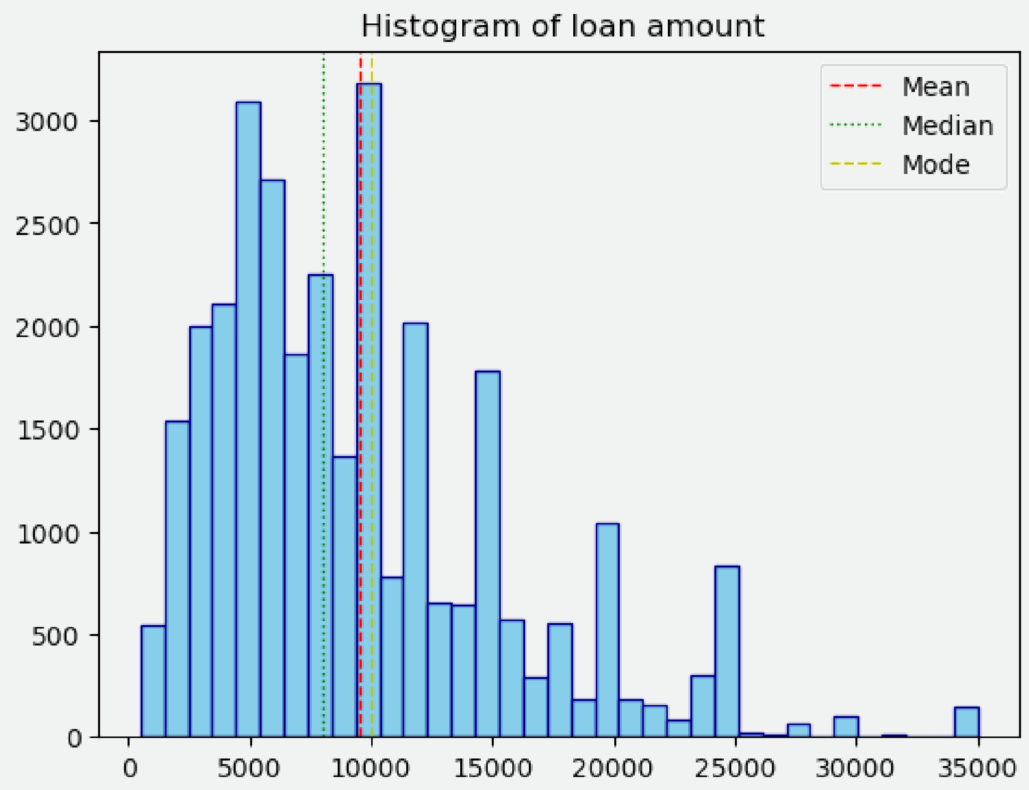
5 Point Summary

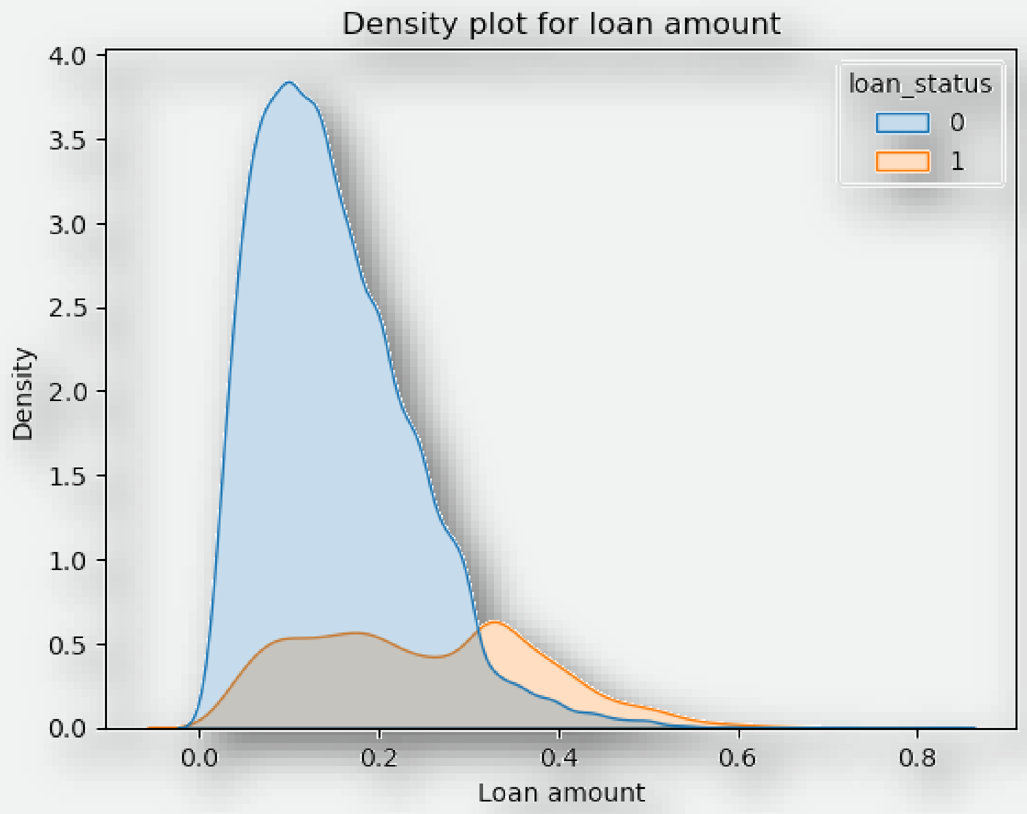
count	31084
mean	9551.81
std	6213.06
min	500.00
25%	5000.00
50%	8000.00
75%	12000.00
max	35000.00
Mode	100000

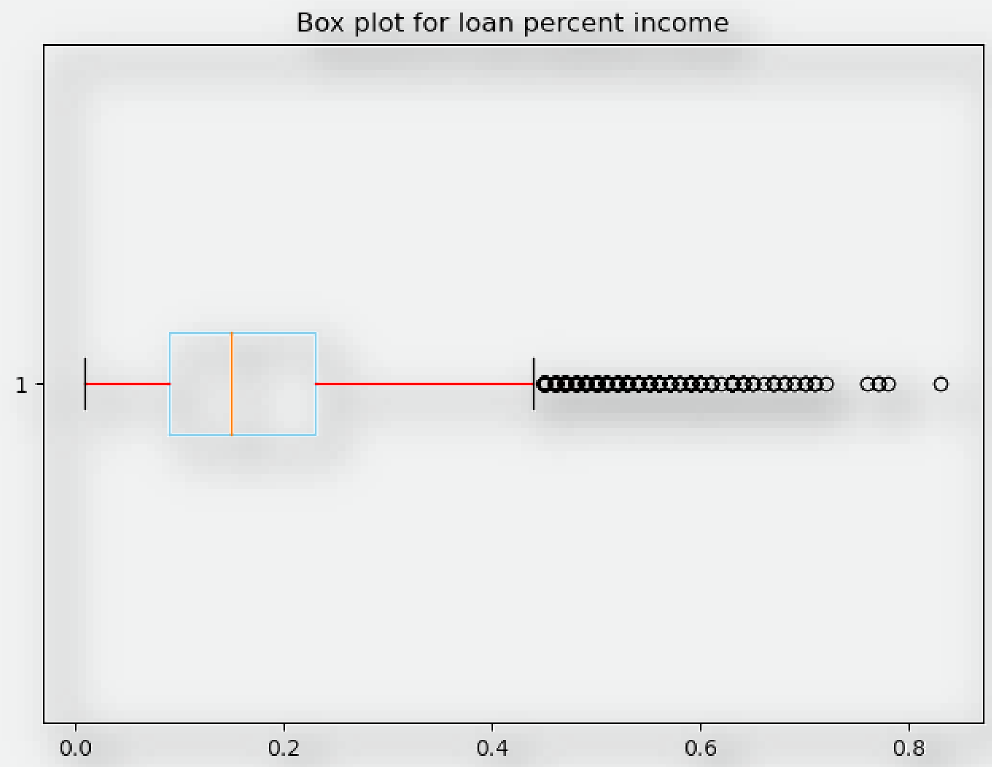
Observation

- Loan amount has a MEAN value of 9551.81 and MEDIAN at 8000.
- Most of the loan amount value, MODE is close to 10000
- Distribution is skewed to the right with along tail as visible in the box plot and histogram.
- Mean is more than the median value also indicates right skewness.
- Outliers : The data ranges from 500 to 35000, but the outliers starts from 23000 onwards, No clean up done

Loan Percent On Income


Histogram of loan amount


Density plot for loan amount


Box plot for loan percent income

5 Point Summary

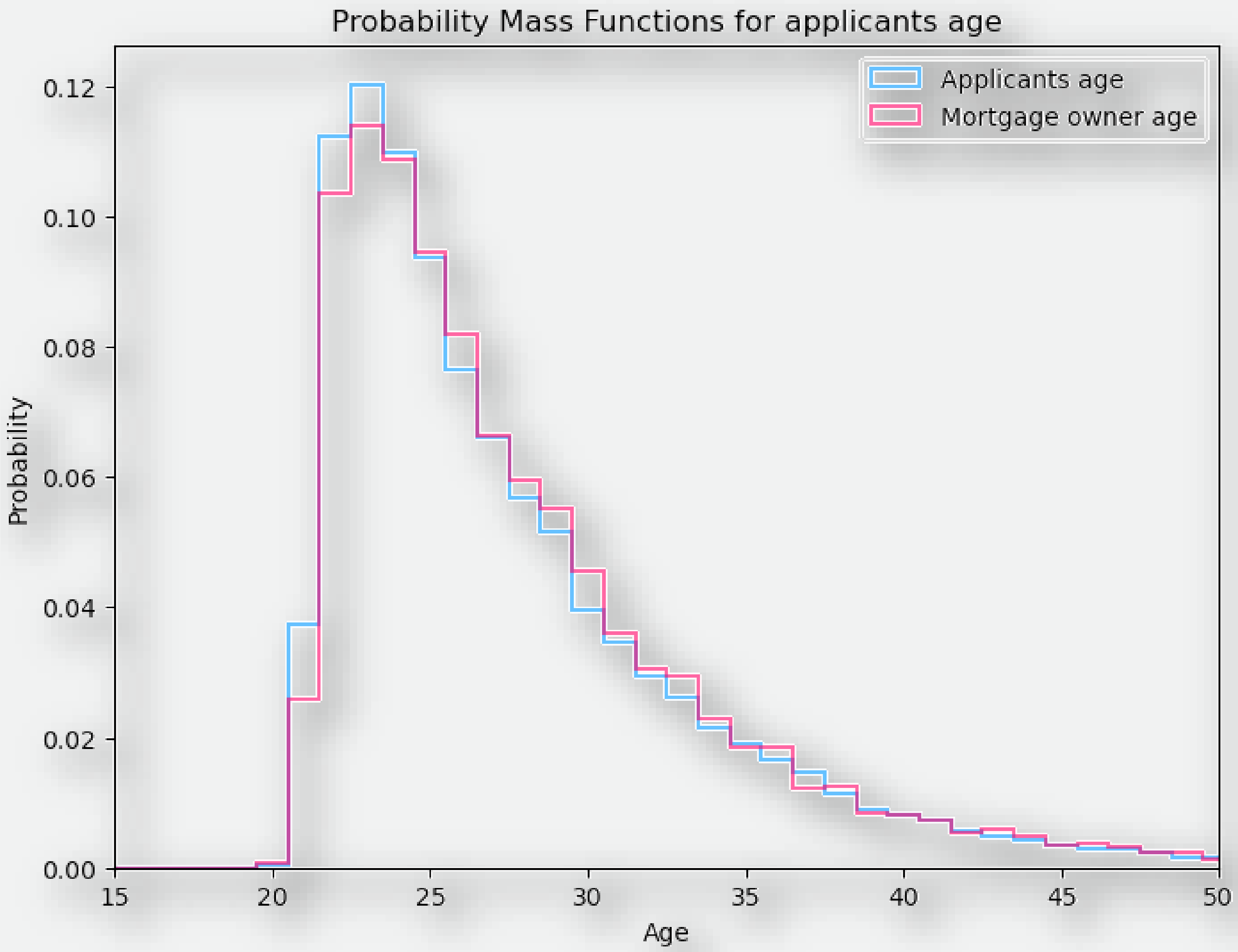
count	31084
mean	0.17
std	0.11
min	0.01
25%	0.09
50%	0.15
75%	0.23
max	0.83
Mode	0.10

Observations

- Loan percent of income has a MEAN value of 0.17 and MEDIAN at 0.15.
- Most of the loan amount value, MODE is 0.10
- Distribution is skewed to the right with along tail as visible in the box plot and histogram.
- Mean is more than the median value also indicates right skewness.
- Outliers : Data ranges from 0 to 0.83, which indicates the percentage of loan amount against income amount

PMF on age

Compare two scenarios in your data using a PMF.



Two sets of data is used to compare the PMF.

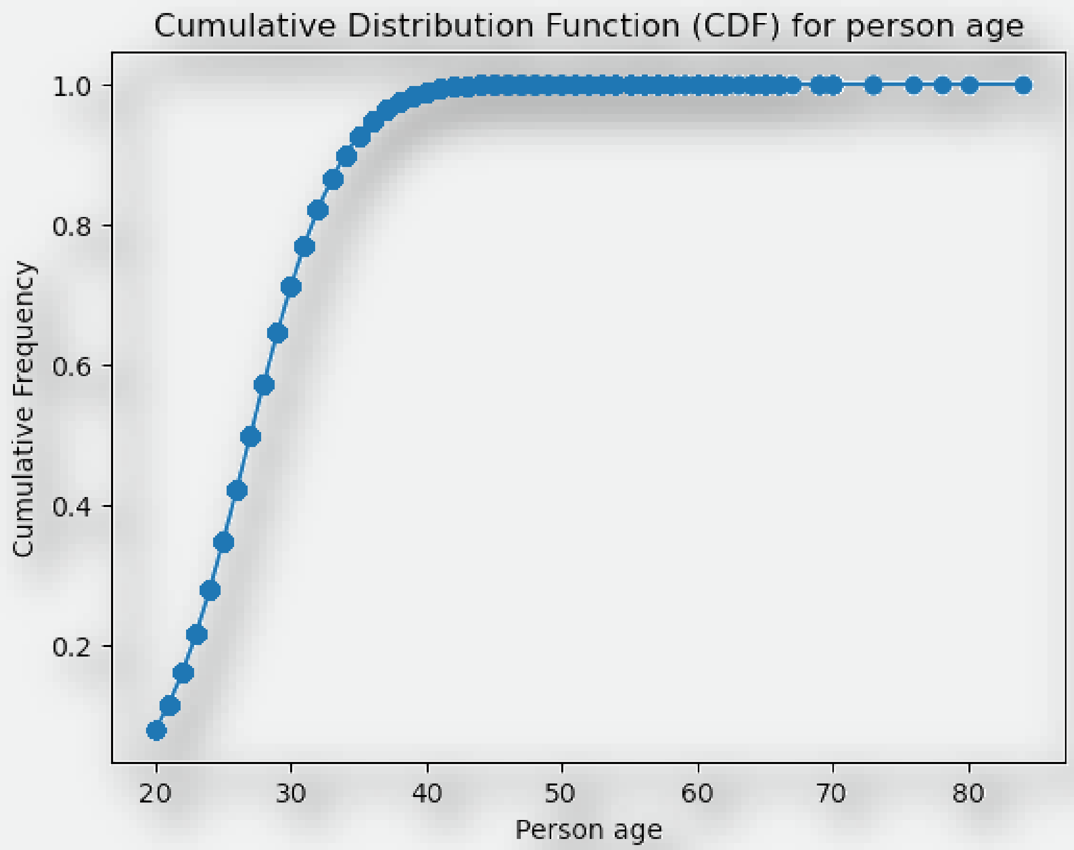
- The first group is age of all the applicants,
- The second group is a subset of the first one, age of applicants which own a house but is on mortgage.

<u>Observations</u>

- The PMF shows that the for any applicants chosen, probability that their age being any value between 23 to 32 is high ,(12% to 4% from 23 to 32)
- The Probability distribution between those two groups is almost identical , except the mortgaged applicants higher probability is 11%

CDF

Create 1 CDF with one of your variables, using page 41-44 as your guide, what does this tell you about your variable and how does it address the question you are trying to answer



From the CDF plot curve on the left , the cumulative % of applicants at a given age value is listed below

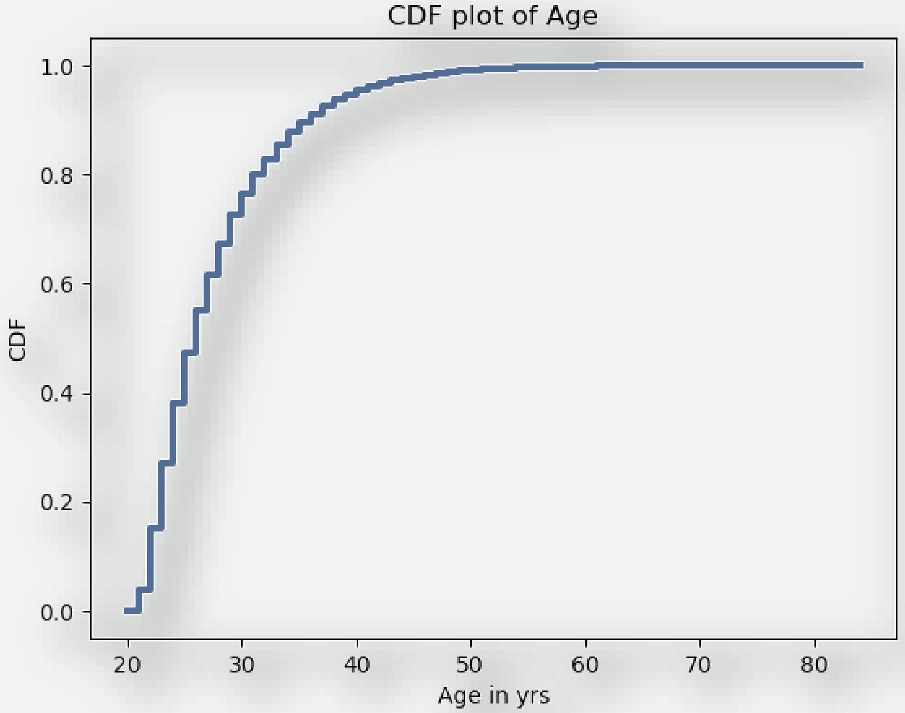
- % of applicant's below age 20 : 0.05%
- % of applicant's below age 21 : 3.79%
- % of applicant's below age 22 : 15.01%
- % of applicant's below age 23 : 27.04%
- % of applicant's below age 24 : 38.02%
- % of applicant's below age 25 : 47.39%
- % of applicant's below age 30 : 76.46%
- % of applicant's below age 35 : 89.53%
- % of applicant's below age 40 : 95.51%
- % of applicant's below age 45 : 98.02%
- % of applicant's below age 50 : 99.14%
- % of applicant's below age 60 : 99.80%
- % of applicant's below age 70 : 99.98%

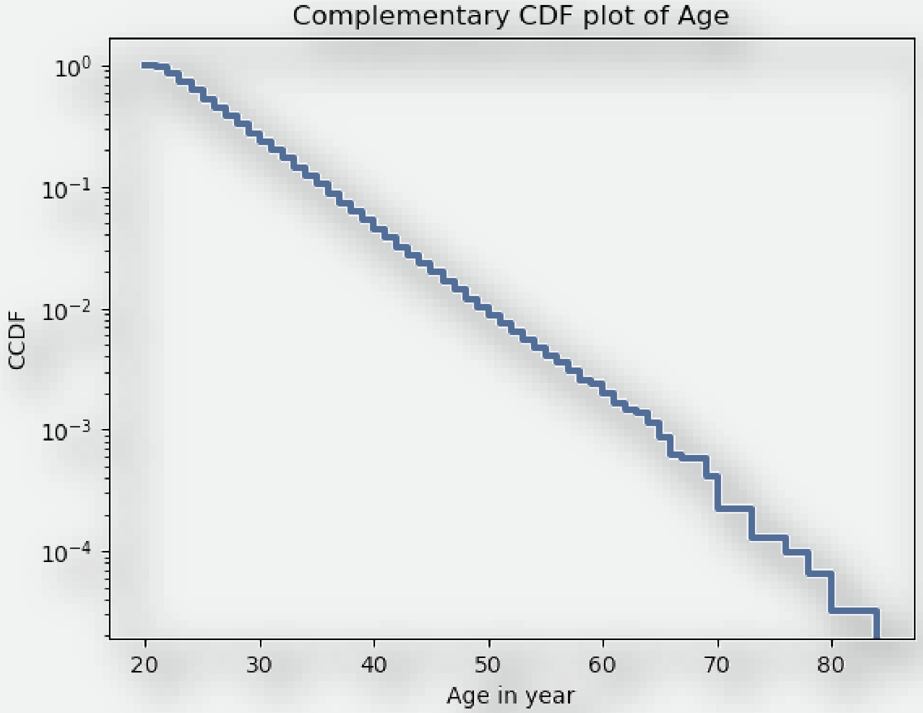
Observations

- From the CDF curve 95% of the applicants are under the 40 (from age 20)
- 75% of the applicants are between 20 to 30 years old

Analytical Distribution

Plot 1 analytical distribution and provide your analysis on how it applies to the dataset you have chosen


CDF plot of Age

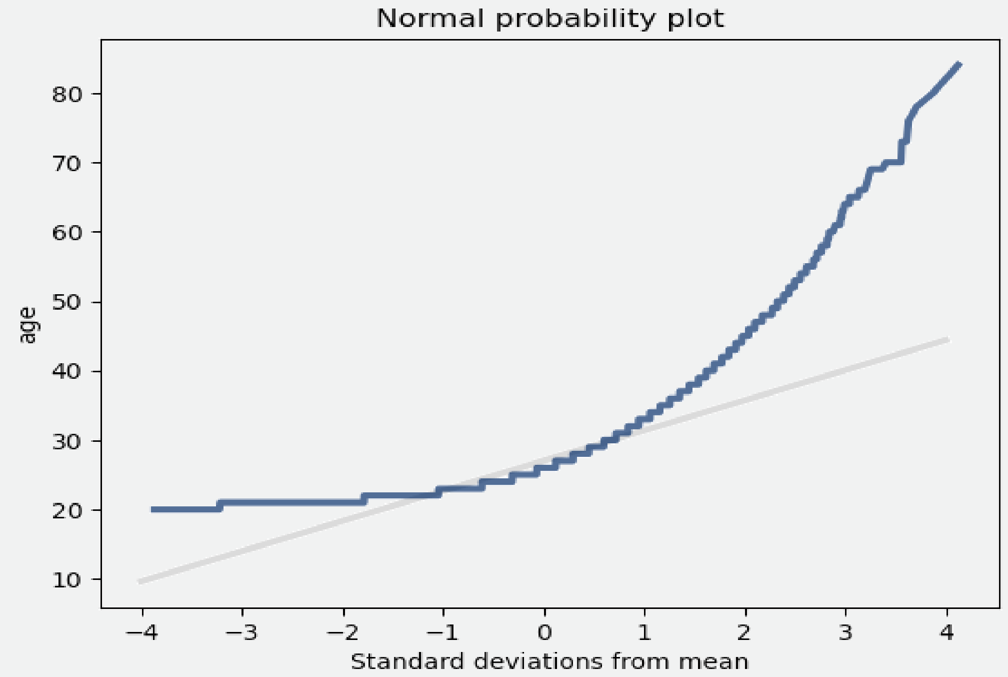

Complementary CDF plot of Age

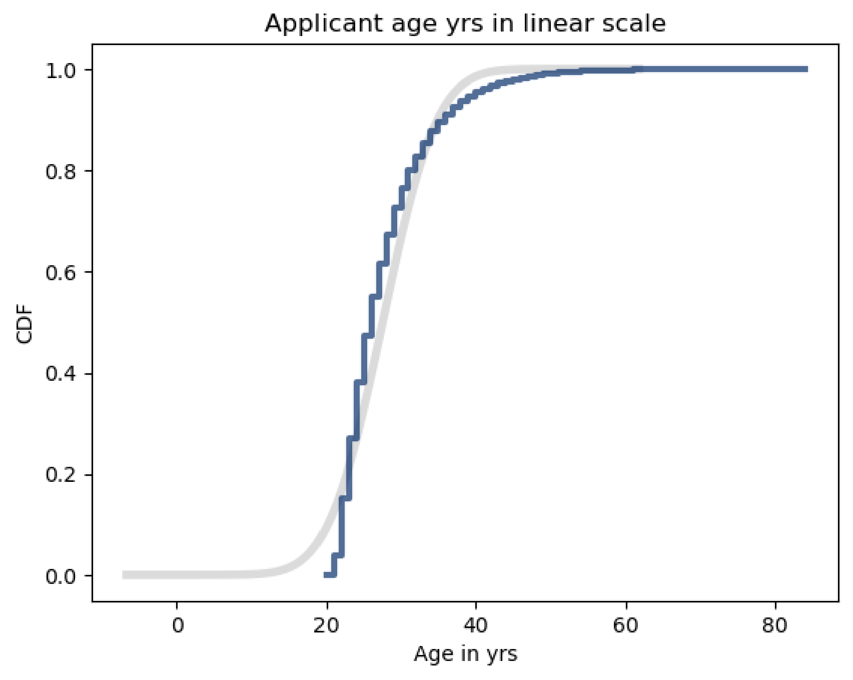
The all applicant age's CDF distribution is plotted, CDF plot of age, to illustrate the **empirical distribution** representing the actual frequencies of probabilities

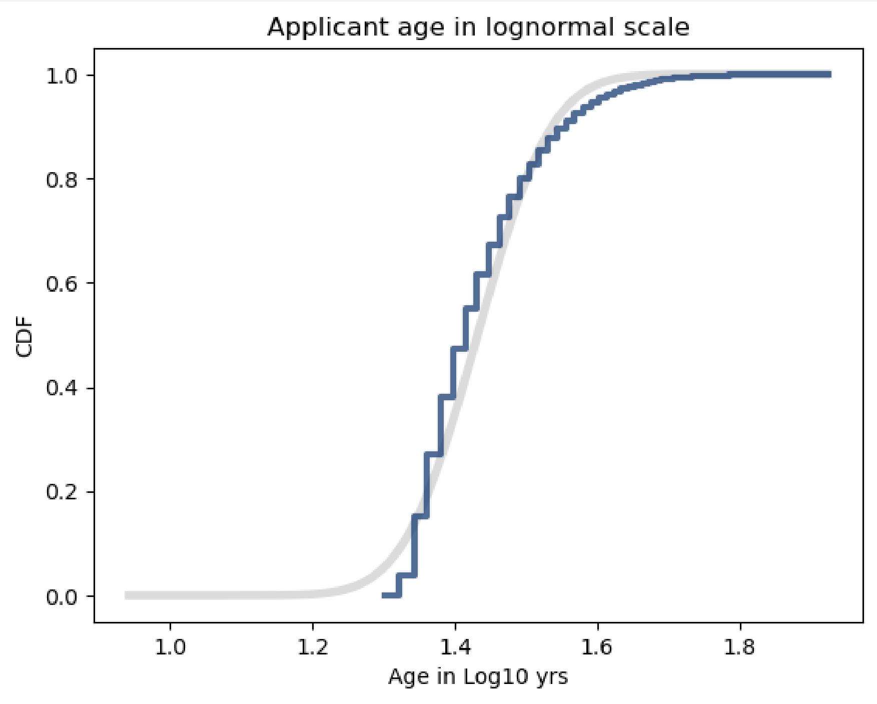
For statistical inference, an analytical distribution is used by using an mathematical function , complement of CDF values i.e. 1 – cdf value, is plotted against a log scale. This helps to understand if the data from an exponential distribution is a straight line. From the Complimentary CDF plot of age , it is almost a straight line breaking after age 65 proving that the data fits for exponential distribution

Analytical Distribution Contd…

Use Normal probability plot to validate the age's data distribution against standard normal distribution







Observation

To validate whether the data distribution is normal by comparing with a standard normal distribution.

Generate a random sample from similar to the applicant's age data using mean value 0 and standard deviation 1

Plot both the standard deviation over mean for both the age data and standard distribution sample.

From the plot both the curves are close or match near the mean and deviate after 1 standard deviation on either side. Meaning most of the data distribution is normal with 1 standard deviation

Observation

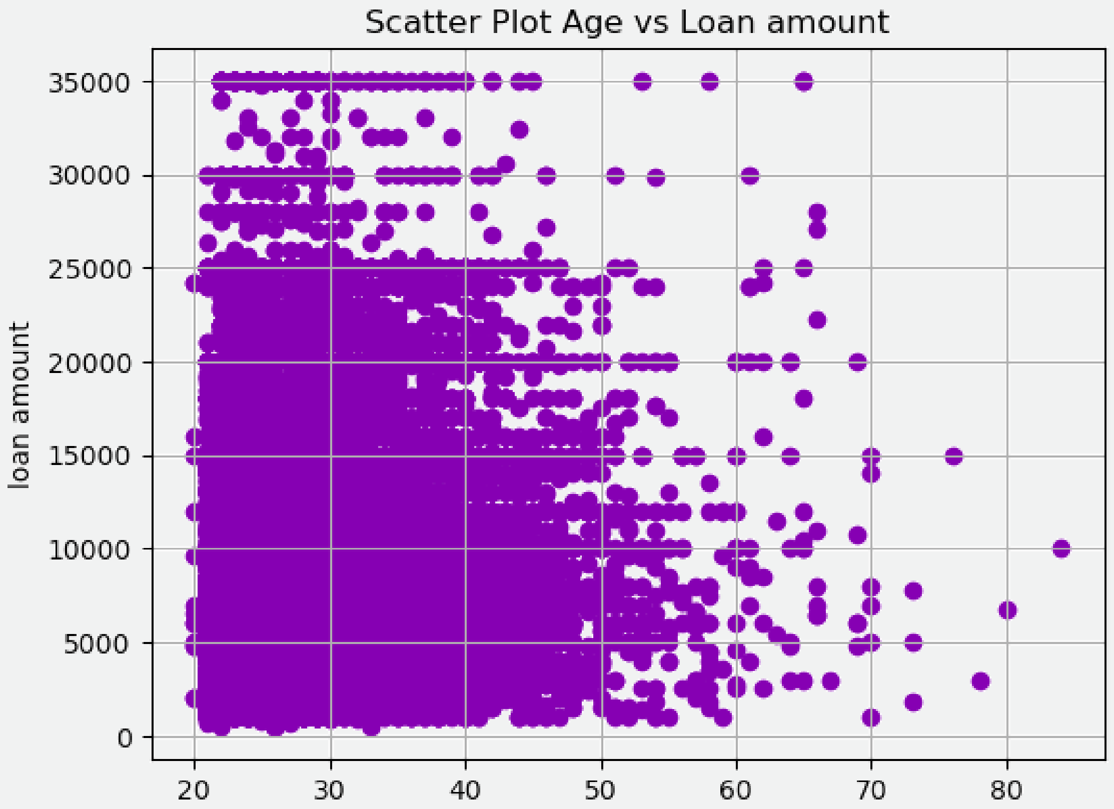
Similar to standard deviation to normal probability plot comparison.

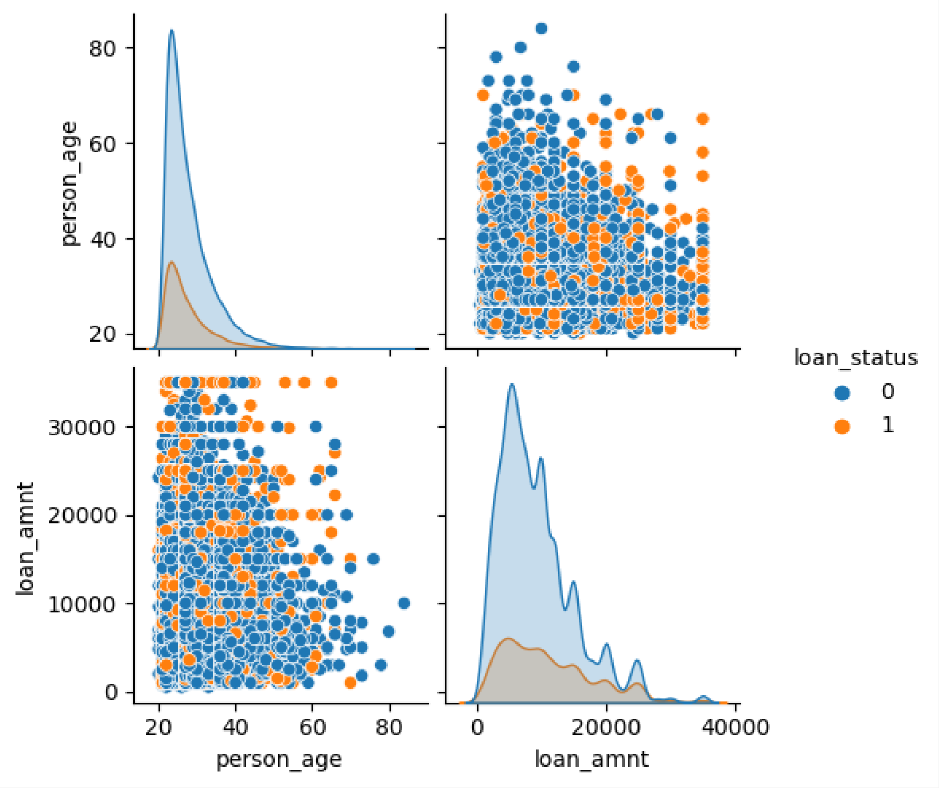
Plot the CDF distribution and compare against standard normal distribution.
To analyze the deviation better compare it against the lognormal exponential distribution of CDF (log values CDF) vs standard normal distribution

From both the plots , it is evident that the applicant age data distribution follows almost a normal distribution. It helps to understand that this data aids in predictions

Scatter Plot and Correlation

Create two scatter plots comparing two variables and provide your analysis on correlation and causation. Remember, covariance, Pearson's correlation, and Non-Linear Relationships should also be considered during your analysis





Observations

A scatter plot comparing age and loan amount is drawn and visually it show a **positive co-relation**, but need the magnitude and statistical significance of the correlation.

Pearson Correlation : 0.0435924739608447
P-value : 1.483073598829262e-14

Pearson correlation coefficient =0.044 suggests a very weak positive linear relationship between the two variables. But extremely small p-value of 1.48×10−141.48×10−^{14} indicates that the observed correlation coefficient is statistically significant.

Spearman's Correlation : 0.057517282050509634
P-value : 3.359184323061208e-24

Spearman correlation coefficient 0.058 indicates a very weak positive monotonic relationship between the two variables
extremely small p-value of 3.36×10−243.36×10−^{24} indicates that the observed Spearman correlation coefficient is statistically significant.

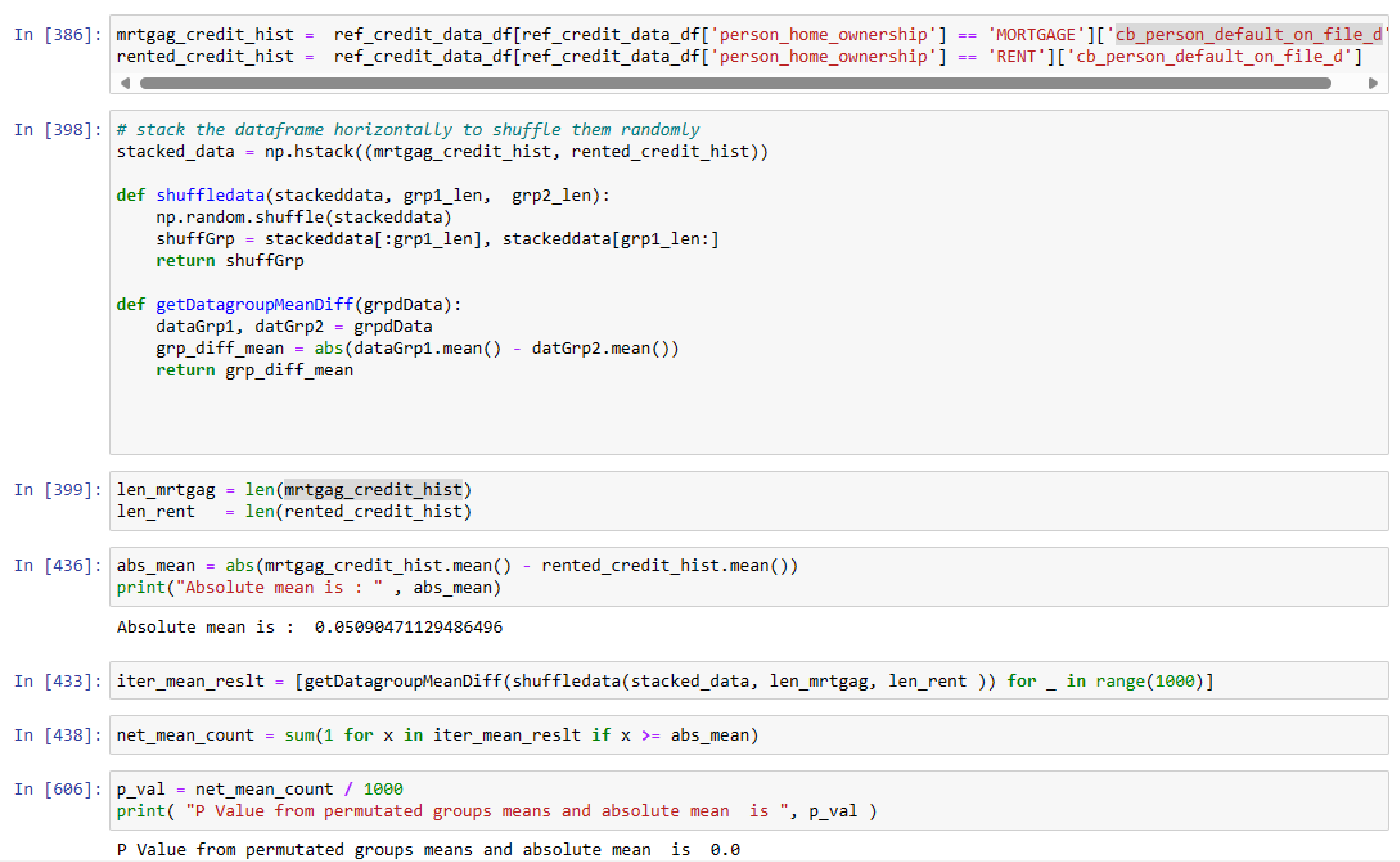
Co=variance : 0.009304957127278272

The magnitude of the covariance (0.0093) indicates that there is some degree of linear association between the two variables

Conclusion
With such a small p-value, the null hypothesis can be rejected and conclude that there is a statistically significant linear relationship between the two variables

HYPOTHESIS

Conduct a test on your hypothesis using one of the methods covered in Chapter 9



```
In [386]: mrtgag_credit_hist =  ref_credit_data_df[ref_credit_data_df['person_home_ownership'] == 'MORTGAGE']['cb_person_default_on_file_d'
          rented_credit_hist =  ref_credit_data_df[ref_credit_data_df['person_home_ownership'] == 'RENT']['cb_person_default_on_file_d']

In [398]: # stack the dataframe horizontally to shuffle them randomly
          stacked_data = np.hstack((mrtgag_credit_hist, rented_credit_hist))

          def shuffledata(stackeddata, grp1_len,  grp2_len):
              np.random.shuffle(stackeddata)
              shuffGrp = stackeddata[:grp1_len], stackeddata[grp1_len:]
              return shuffGrp

          def getDatagroupMeanDiff(grpdData):
              dataGrp1, datGrp2 = grpdData
              grp_diff_mean = abs(dataGrp1.mean() - datGrp2.mean())
              return grp_diff_mean

In [399]: len_mrtgag = len(mrtgag_credit_hist)
          len_rent   = len(rented_credit_hist)

In [436]: abs_mean = abs(mrtgag_credit_hist.mean() - rented_credit_hist.mean())
          print("Absolute mean is : " , abs_mean)

          Absolute mean is :  0.05090471129486496

In [433]: iter_mean_reslt = [getDatagroupMeanDiff(shuffledata(stacked_data, len_mrtgag, len_rent )) for _ in range(1000)]

In [438]: net_mean_count = sum(1 for x in iter_mean_reslt if x >= abs_mean)

In [606]: p_val = net_mean_count / 1000
          print( "P Value from permutated groups means and absolute mean  is ", p_val )

          P Value from permutated groups means and absolute mean  is  0.0
```

Null Hypothesis : One way to model the null hypothesis is by permutation; that is, People those have mortgaged or rented a house have almost similar defaulted rate on file and difference is not statistically different.
Using permutation, combine both the group, shuffle and find the mean using multiple iteration and find the mean difference

Create 2 subsets of 'cb_person_default_on_file_d' , credit default history on file, data by filtering home ownership data using values 'MORTGAGE' and 'RENT'

Calculate the absolute mean difference between these two subset of data (**Absolute mean** is : 0.05090471129486496)

Conduct an iteration (1000 in this case) to permutate the data (by grouping and shuffling two subset of data and then splitting again) and find the mean difference between those shuffled groups.

Find the p-value using the absolute means and the shuffled data means (check if the shuffled mean is above absolute mean) and find ration of count vs number of iterations. (P **Value** from permutated groups mean and absolute mean is 0.0)

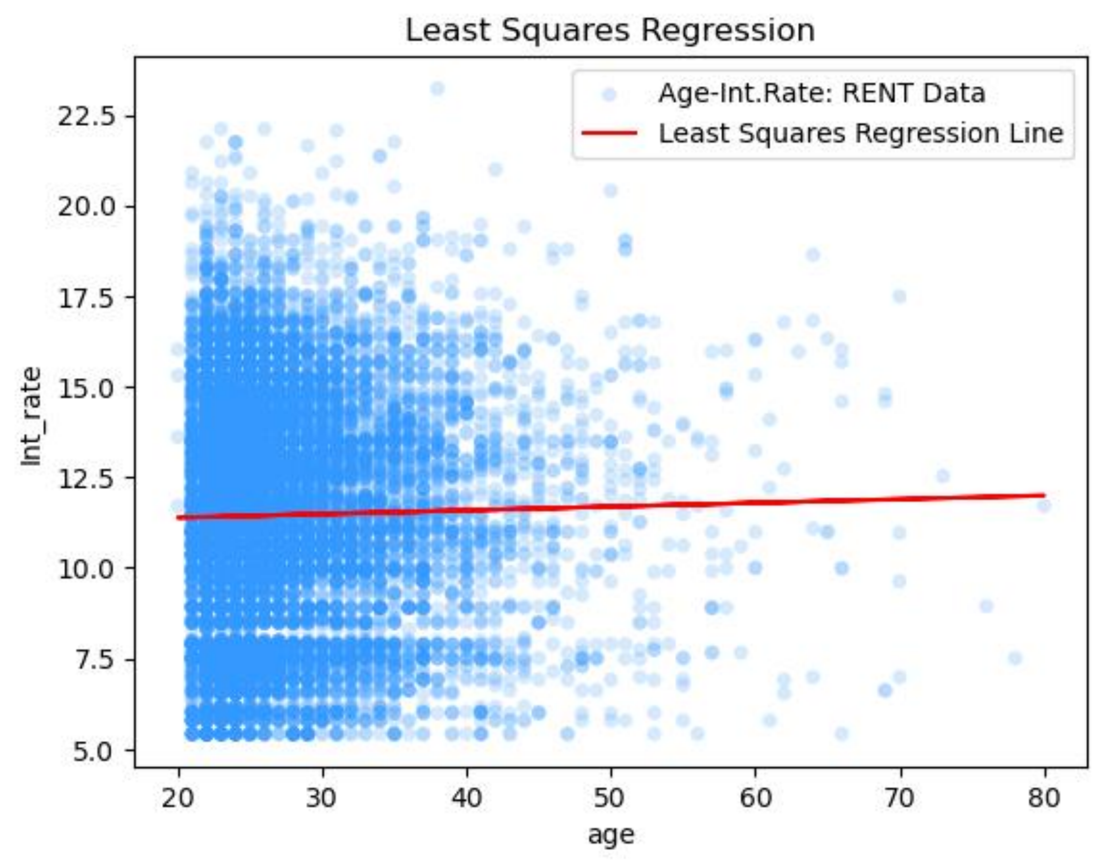
Observations

- The absolute mean difference of 0.05 suggests that there is a substantial difference between the means of the two groups. Since the absolute mean difference is not close to zero, it indicates that the groups are different in terms of the "defaulted rate on file" being measured.

- A permuted p-value of 0 means that in all of the permutations performed, none of the permutations resulted in a difference as extreme as the observed difference between the two groups. It's highly unlikely to observe such a large difference if there were no true difference between the groups. Therefore, the null hypothesis that there is no difference between the groups is rejected.

Linear Regression – Least Squares

Conduct a regression analysis on either one dependent and one explanatory variable, or multiple explanatory variables (Chapter 10 & 11). Your code or screenshots of your code

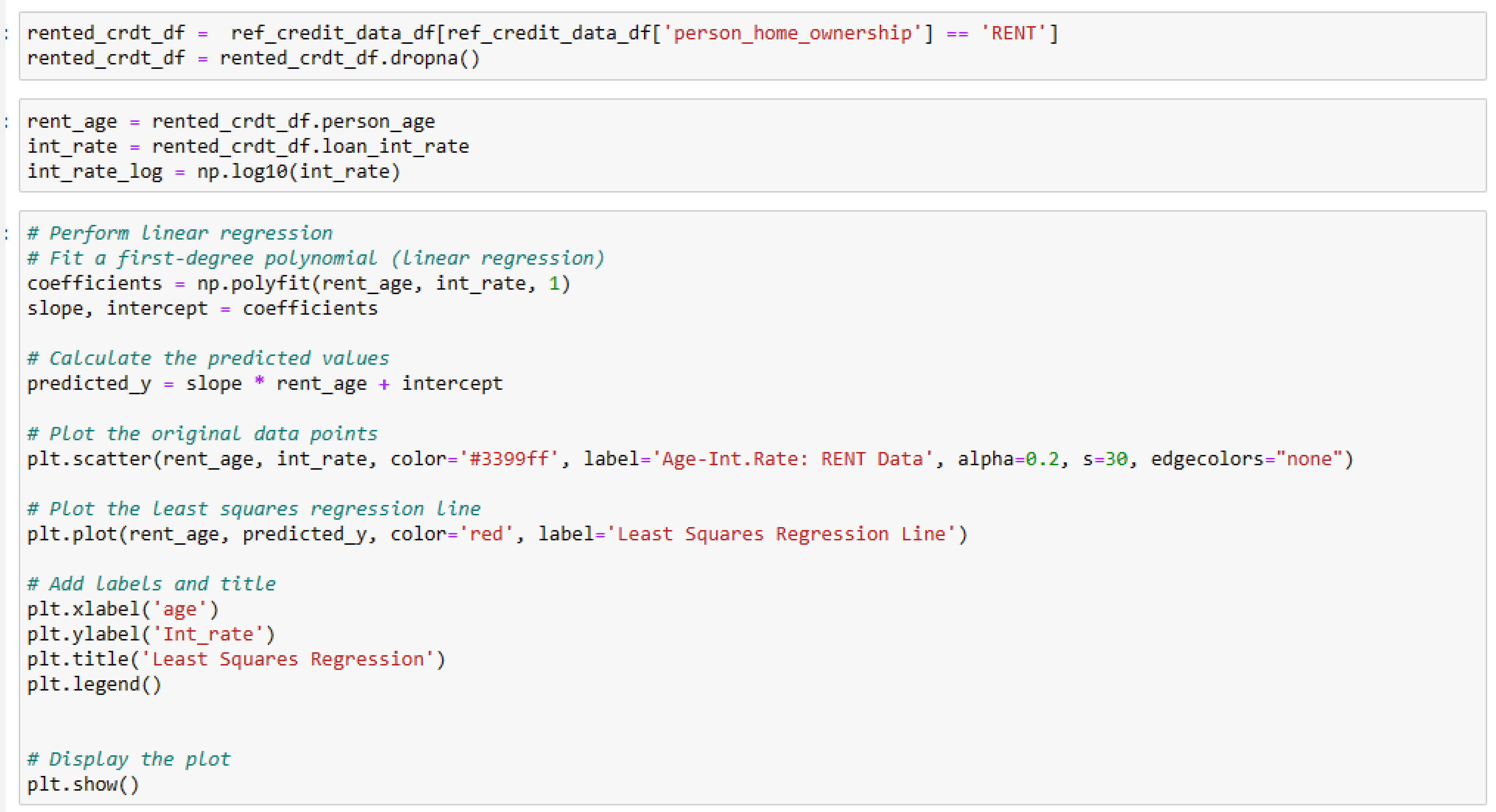


Steps to conduct Least Square regression

To Conduct Least squares we use the age and interest rate for the applicants who has a home rented.

Filter the data frame by rented home owner column.
Create the subset of data of age and interest rate

Find the slope and intercept , first degree polynomial , of age and interest rate
Plot the data and the regression line and understand if the relationship is linear.



```
rented_crdt_df =  ref_credit_data_df[ref_credit_data_df['person_home_ownership'] == 'RENT']
rented_crdt_df = rented_crdt_df.dropna()
```

```
rent_age = rented_crdt_df.person_age
int_rate = rented_crdt_df.loan_int_rate
int_rate_log = np.log10(int_rate)
```

```
# Perform linear regression
# Fit a first-degree polynomial (linear regression)
coefficients = np.polyfit(rent_age, int_rate, 1)
slope, intercept = coefficients

# Calculate the predicted values
predicted_y = slope * rent_age + intercept

# Plot the original data points
plt.scatter(rent_age, int_rate, color='#3399ff', label='Age-Int.Rate: RENT Data', alpha=0.2, s=30, edgecolors="none")

# Plot the least squares regression line
plt.plot(rent_age, predicted_y, color='red', label='Least Squares Regression Line')

# Add labels and title
plt.xlabel('age')
plt.ylabel('Int_rate')
plt.title('Least Squares Regression')
plt.legend()


# Display the plot
plt.show()
```

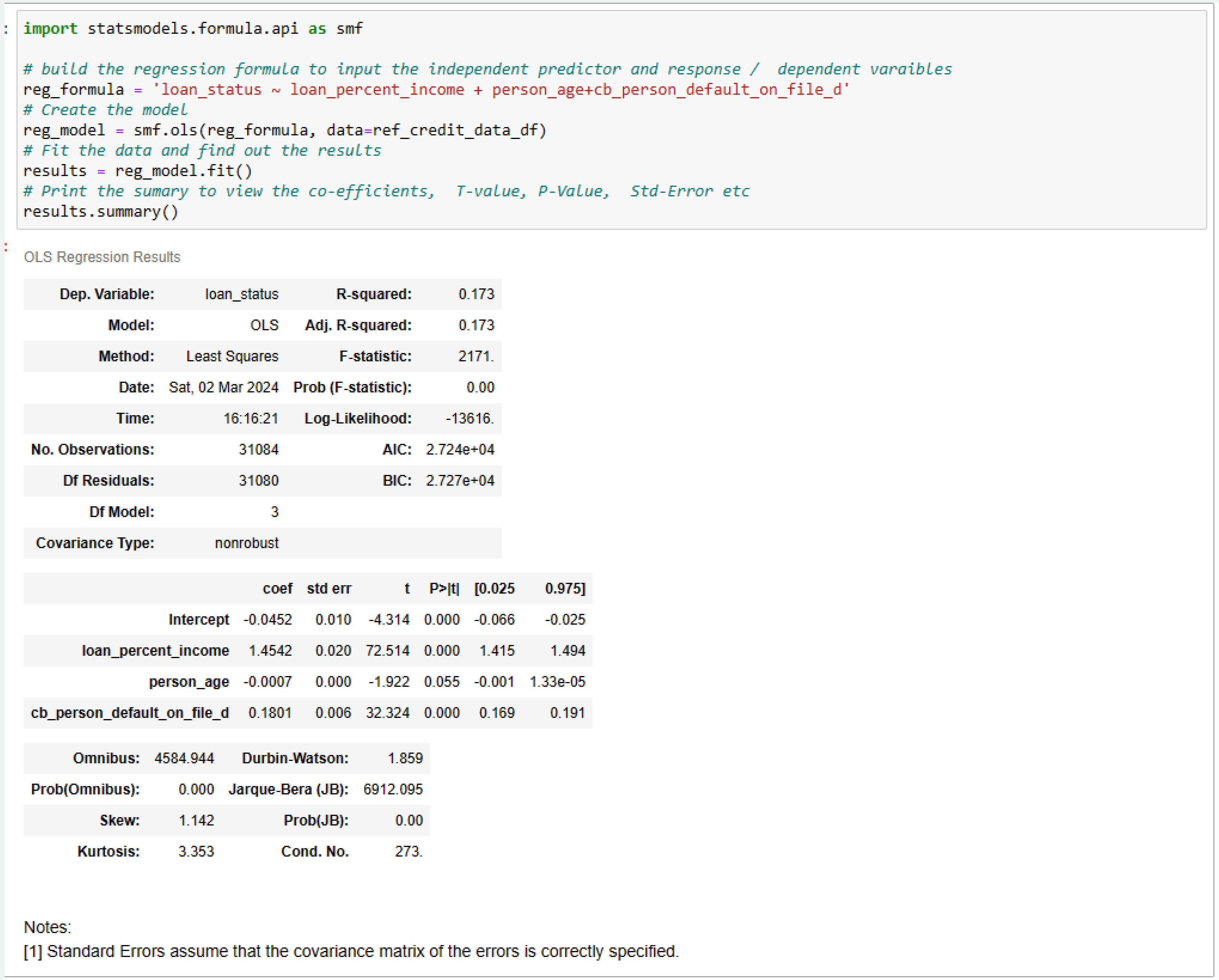
Observations

Slope   : 0.0102
intercept : 11.1806

The Regression line intercept of starts close to interest rate's mean value of 11.46.
The slope indicates a linear relationship of 0.012 unit increase in interest rate for every I unit of age (1 yr.)

Regression Analysis - multiple variables

Perform Regression analysis using multiple variables
Regression analysis on Loan approval status against "Credit default status on file" , "Percentage of loan against income" and “applicant’s age“



```python
import statsmodels.formula.api as smf

# build the regression formula to input the independent predictor and response /  dependent varaibles
reg_formula = 'loan_status ~ loan_percent_income + person_age+cb_person_default_on_file_d'
# Create the model
reg_model = smf.ols(reg_formula, data=ref_credit_data_df)
# Fit the data and find out the results
results = reg_model.fit()
# Print the sumary to view the co-efficients,  T-value, P-Value,  Std-Error etc
results.summary()
```

```
OLS Regression Results

      Dep. Variable:      loan_status        R-squared:      0.173
              Model:              OLS   Adj. R-squared:      0.173
             Method:    Least Squares      F-statistic:      2171.
               Date: Sat, 02 Mar 2024 Prob (F-statistic):     0.00
               Time:         16:16:21   Log-Likelihood:    -13616.
   No. Observations:            31084              AIC:  2.724e+04
       Df Residuals:            31080              BIC:  2.727e+04
           Df Model:                3
    Covariance Type:        nonrobust

                              coef  std err       t  P>|t|  [0.025     0.975]
                 Intercept -0.0452    0.010  -4.314  0.000  -0.066     -0.025
       loan_percent_income  1.4542    0.020  72.514  0.000   1.415      1.494
                person_age -0.0007    0.000  -1.922  0.055  -0.001   1.33e-05
cb_person_default_on_file_d 0.1801    0.006  32.324  0.000   0.169      0.191

       Omnibus: 4584.944    Durbin-Watson:    1.859
 Prob(Omnibus):    0.000 Jarque-Bera (JB): 6912.095
          Skew:    1.142         Prob(JB):     0.00
      Kurtosis:    3.353         Cond. No.     273.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Build the formula similar to ‘R’ to declare the dependent vs independent variables

- “Loan approval status “ , loan_status is declared as dependent / response variable
- "Percentage of loan against income“, loan_percent_income and “applicant’s age”, person_age, and "Credit default status on file" , cb_person_default_on_file_d (dummy variable of category variable cb_person_default_on_file) is used as independent/ **Predictor** variables
- Use the statsmodels library to create and model and fir the data and then print the summary.

Regression Analysis - multiple variables

Observations

"Percentage of loan against income", loan_percent_income

- The predictor has a co-eff of 1.45 suggests a very good positive effect on response variable, 1 unit of loan percent will increase the loan_status by 1.45 units, holding all other variables constant.
- A positive standard error value of 0.02 indicates very less uncertainty with the co-efficient value, suggests co-efficient value is relatively precise
- T-static of 72 indicates co-eff estimate is 72 standard error times away from zero, large absolute value of the t-statistic indicates a strong signal relative to the noise
- The p-value of 0.0 , even though ideal , higher than 0.05 indicates that is predictor may not be statistically significant to reject the null hypothesis that on relationship. It suggests strong evidence against the null hypothesis, indicating that the effect of the predictor variable is statistically significant.

"applicant's age", person_age

- The predictor has a co-eff of -0.0007 suggest that , a one-unit increase in the predictor variable, the response variable is expected to decrease by 0.0007 units.
- A positive standard error value 0 indicates that there is no variability or uncertainty in the coefficient
- T-static of -1.92, a negative t-statistic indicates that the coefficient estimate is negative
- The p-value of 0.055 , suggests weak evidence against the null hypothesis, the value just above the border hence it is statistically significant.

"Credit default status on file" , cb_person_default_on_file_d

- The predictor has a co-eff of 0.18 suggest that the predictor variable has very small effect on response variable.
- A positive small standard error value 0.006 indicates more precise estimation of the coefficient.
- Large T-static value of 32,32 indicates a strong signal relative to the noise
- The p-value of 0.0 , suggests strong evidence against the null hypothesis

Conclusion

- The Exploratory Data Analysis (EDA) results rejected the null hypothesis suggesting no relationship between the response variable, loan_status (determining loan approval), and the predictor variables.
- It was observed that out of 11 predictor variables (comprising 7 continuous and 4 categorical variables), all exhibited statistical significance concerning the response variable.
- Eliminating duplicates ensured that the model remains robust and avoids overfitting.
- Trimming outliers from three variables, namely person_age, person_income, and person_emp_length, aided in reducing distribution tail and skewness, especially due to the presence of non-factual values.
- Examination of the five-point summary, histograms, box plots, and density plots in relation to loan_status provided insights into the predictive variable effectiveness.
- Probability Mass Function (PMF) analysis conducted on two datasets—applicants of all ages and mortgage home owners—indicated a consistent probability distribution within subgroups
- Cumulative Distribution Function (CDF) analysis helped identify the age distribution of applicants, revealing that 75% were aged below 30 years and 95% below 40 years
- Analytical distribution revealed that the selected person_age variable fits both empirical and analytical distributions when compared to the standard normal distribution
- Scatter plots and correlation coefficients (Pearson and Spearman) suggested a weak or small positive correlation between person_age and loan_amount, yet the p-value established this relationship as statistically significant.
- Permutation of data between groups reaffirmed the absence of a difference in credit-on-file data distribution between the two groups, signifying their statistical significance
- Regression analysis provided confirmation of the statistically significant relationship between multiple predictor or independent variables and the response variable, thus rejecting the null hypothesis of no relationship between dependent and independent variables

- **Overall** – The Exploratory Data Analysis (EDA) aided in rejecting the Null Hypothesis, which proposed that independent variables have no statistical impact on the dependent variable, loan status. Instead, it **affirmed that there is a statistically significant relationship between the independent variables and the dependent variable, loan_status.**

Acknowledgment

I am deeply grateful for the unwavering support provided by Professor Fadi Alsaleem throughout this course by providing feedback during the assignment grading. This helped and motivated me to refine my subsequent submissions

Additionally, I extend my heartfelt appreciation to my classmates for fostering a collaborative and enriching learning environment by posting good discussion materials, blogs and links.