Loan Approval Classification Using Big Data Analytics

Project Report

Loan Approval Classification Using Big Data Analytics

DSCI 5350 – Big Data Analytics

G. Brint Ryan College of Business

University of North Texas

Denton, TX

December 2024

Executive Summary

Objective: "To analyze loan application data using big data tools to identify key factors influencing loan approval."

Key Findings:

- Previous_loan_defaults and loan_percent_income are the strongest predictors of loan approval.Applicants with previous_loan_defaults are less likely to get approved.
- The Random Forest model achieved higher accuracy than linear regression model with 93%.

Impact: "The findings help financial institutions improve risk assessment and streamline loan processing."

Project motivation/background

- Loan approval is vital for financial institutions, affecting profitability and applicants' financial opportunities. With the rise of digital transactions, manual evaluations become inefficient.
- The availability of rich datasets provides an opportunity to leverage machine learning techniques to improve prediction accuracy.
- Minimizing loan defaults by identifying high-risk applicants helps reduce financial losses.
- Benefits:
  - **For Financial Institutions**: Increases efficiency, accuracy, and reduces risks.
  - **For Applicants**: Ensures fair, transparent, and fast decisions, boosting customer satisfaction and inclusion.

Dataset Description

This dataset is retrieved from Kaggle.com. The dataset contains 45000 rows and 14 columns for classifying loan approval status based on the applicant details. The attributes for each applicant are as follows:

Person_age	Age of the applicant
Person_gender	Gender of the applicant
Person_education	Highest level of education of the applicant

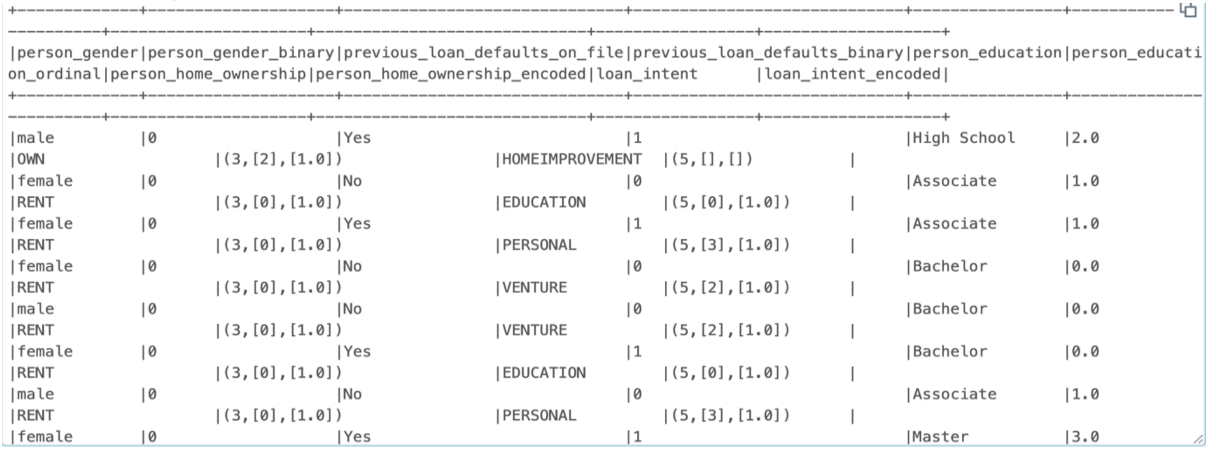
Person_income	Annual income of the applicant
Person_emp_exp	Employment experience in years
Person_home_ownership	Home ownership status of the applicant
Loan_amnt	Requested loan amount
Loan_intent	Purpose of the loan
loan_int_rate	Interest applied to the loan
Loan_percent_income	Loan amount with respect to the percentage of annual income
cb_person_cred_hist_length	Years of credit history
Credit_score	Credit score of the applicant
Previous_loan_defaults_on_file	Categorical indicator of previous loan defaults
Loan_status	Status of the loan (1 = approved, 0 = rejection

Key Variables:

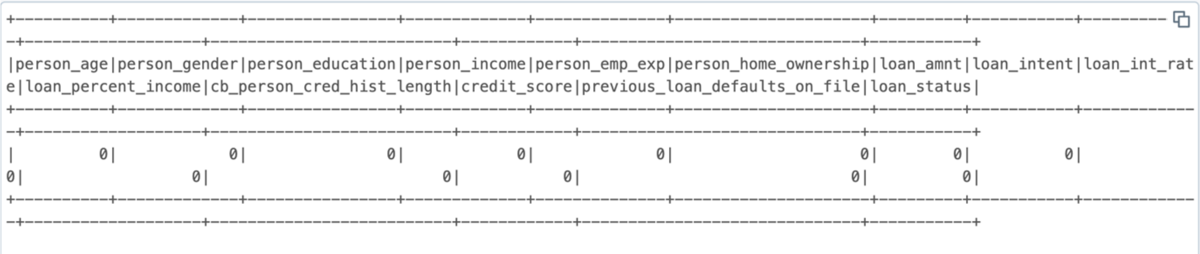
- Demographics: Age, Gender, Education, Employment Experience.
- Financials: Income, Home Ownership, Credit Score, Loan Amount.
- Loan Features: Interest Rate, Loan Purpose, Loan Status.
- Summary Statistics:
  - Average credit score: 632
  - Average loan amount: 9583
  - Approval rate: 22%

Data Preparation

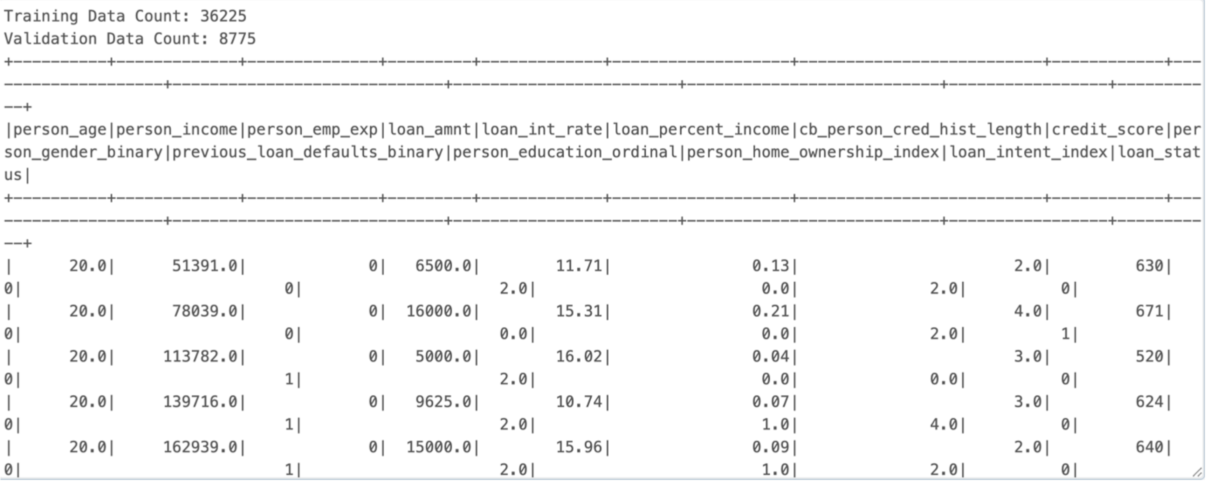
- Checked for missing values: None found.
- Replaced person_age greater than 100 with median values.
- Applied feature encoding (binary, ordinal, one-hot) for categorical variables.
- Scaled numerical features using standard scalar.
- Split data into training and testing sets (e.g., 80/20 split).
- Apache Spark was used to handle data preprocessing and model training efficiently at scale.



```
+-------------+--------------------+------------------------------+-----------------------------+----------------+-----------
----------+---------------------+-----------------------------+-----------------+-------------------+
|person_gender|person_gender_binary|previous_loan_defaults_on_file|previous_loan_defaults_binary|person_education|person_educati
on_ordinal|person_home_ownership|person_home_ownership_encoded|loan_intent      |loan_intent_encoded|
+-------------+--------------------+------------------------------+-----------------------------+----------------+--------------
----------+---------------------+-----------------------------+-----------------+-------------------+
|male         |0                   |Yes                           |1                            |High School     |2.0
|OWN                  |(3,[2],[1.0])                |HOMEIMPROVEMENT  |(5,[],[])          |
|female       |0                   |No                            |0                            |Associate       |1.0
|RENT                 |(3,[0],[1.0])                |EDUCATION        |(5,[0],[1.0])      |
|female       |0                   |Yes                           |1                            |Associate       |1.0
|RENT                 |(3,[0],[1.0])                |PERSONAL         |(5,[3],[1.0])      |
|female       |0                   |No                            |0                            |Bachelor        |0.0
|RENT                 |(3,[0],[1.0])                |VENTURE          |(5,[2],[1.0])      |
|male         |0                   |No                            |0                            |Bachelor        |0.0
|RENT                 |(3,[0],[1.0])                |VENTURE          |(5,[2],[1.0])      |
|female       |0                   |Yes                           |1                            |Bachelor        |0.0
|RENT                 |(3,[0],[1.0])                |EDUCATION        |(5,[0],[1.0])      |
|male         |0                   |No                            |0                            |Associate       |1.0
|RENT                 |(3,[0],[1.0])                |PERSONAL         |(5,[3],[1.0])      |
|female       |0                   |Yes                           |1                            |Master          |3.0
```



```
+----------+-------------+----------------+-------------+--------------+---------------------+---------+-----------+------------
-+-------------------+--------------------------+------------+------------------------------+-----------+
|person_age|person_gender|person_education|person_income|person_emp_exp|person_home_ownership|loan_amnt|loan_intent|loan_int_rat
e|loan_percent_income|cb_person_cred_hist_length|credit_score|previous_loan_defaults_on_file|loan_status|
+----------+-------------+----------------+-------------+--------------+---------------------+---------+-----------+------------
-+-------------------+--------------------------+------------+------------------------------+-----------+
|         0|            0|               0|            0|             0|                    0|        0|          0|
0|                  0|                         0|           0|                             0|          0|
+----------+-------------+----------------+-------------+--------------+---------------------+---------+-----------+------------
-+-------------------+--------------------------+------------+------------------------------+-----------+
```

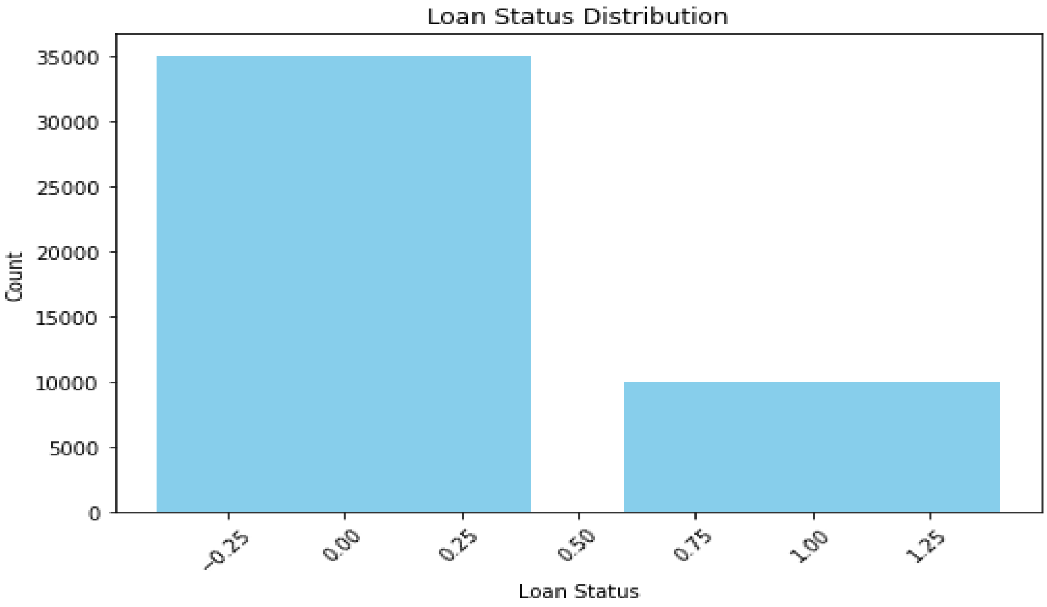


```
Training Data Count: 36225
Validation Data Count: 8775
+----------+-------------+--------------+---------+-------------+-------------------+--------------------------+------------+---
-----------------+-----------------------------+------------------------+---------------------------+-----------------+---------
--+
|person_age|person_income|person_emp_exp|loan_amnt|loan_int_rate|loan_percent_income|cb_person_cred_hist_length|credit_score|per
son_gender_binary|previous_loan_defaults_binary|person_education_ordinal|person_home_ownership_index|loan_intent_index|loan_stat
us|
+----------+-------------+--------------+---------+-------------+-------------------+--------------------------+------------+---
-----------------+-----------------------------+------------------------+---------------------------+-----------------+---------
--+
|      20.0|      51391.0|             0|   6500.0|        11.71|               0.13|                       2.0|         630|
0|                            0|                     2.0|                        0.0|              2.0|          0|
|      20.0|      78039.0|             0|  16000.0|        15.31|               0.21|                       4.0|         671|
0|                            0|                     0.0|                        0.0|              2.0|          1|
|      20.0|     113782.0|             0|   5000.0|        16.02|               0.04|                       3.0|         520|
0|                            1|                     2.0|                        0.0|              0.0|          0|
|      20.0|     139716.0|             0|   9625.0|        10.74|               0.07|                       3.0|         624|
0|                            1|                     2.0|                        1.0|              4.0|          0|
|      20.0|     162939.0|             0|  15000.0|        15.96|               0.09|                       2.0|         640|
0|                            1|                     2.0|                        1.0|              2.0|          0|
```

Exploratory Data Analysis (EDA)

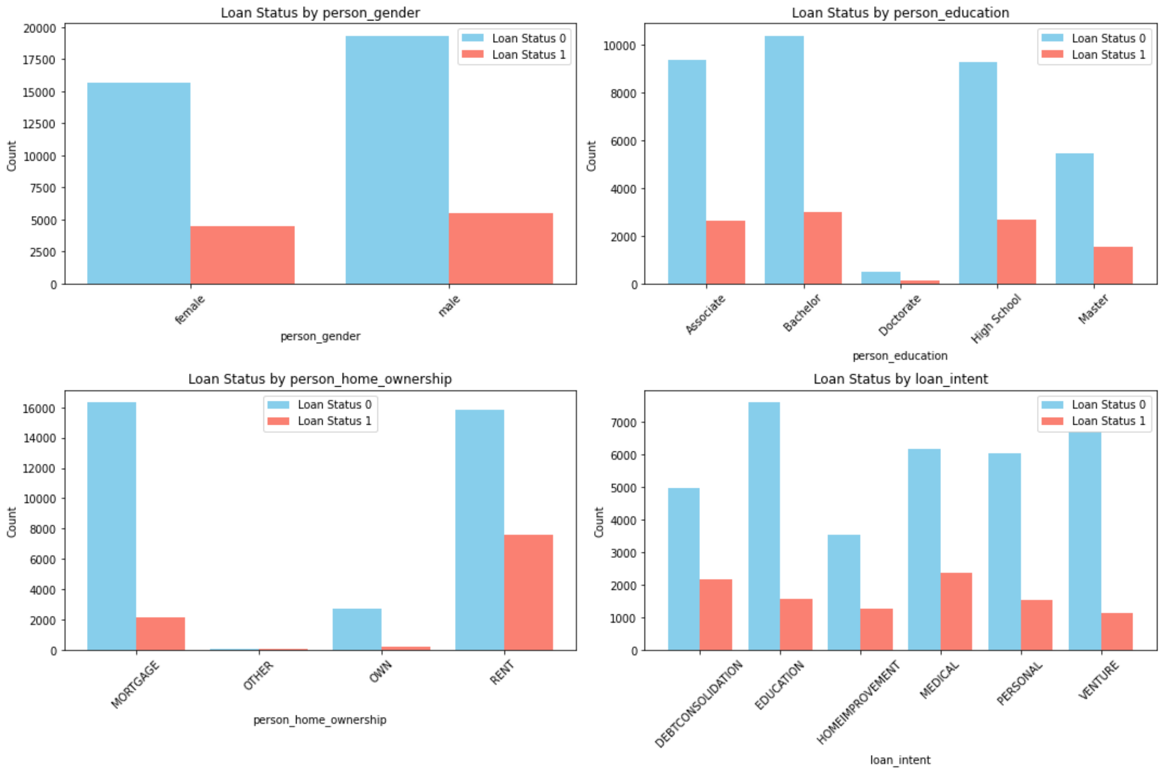
Target Variable Distribution – Key Observations

- The target variable exhibits a significant class imbalance, with rejected loan applications (0) comprising the majority of the dataset.
- Approved loans (1) represent approximately 22% of total observations.
- This imbalance emphasizes the need for evaluation metrics beyond accuracy, as a naive classifier could achieve high accuracy by favoring the majority class.
- To address this challenge, tree-based models such as Random Forest were explored to better capture complex patterns within the minority approval class.



Categorical Variable Analysis:

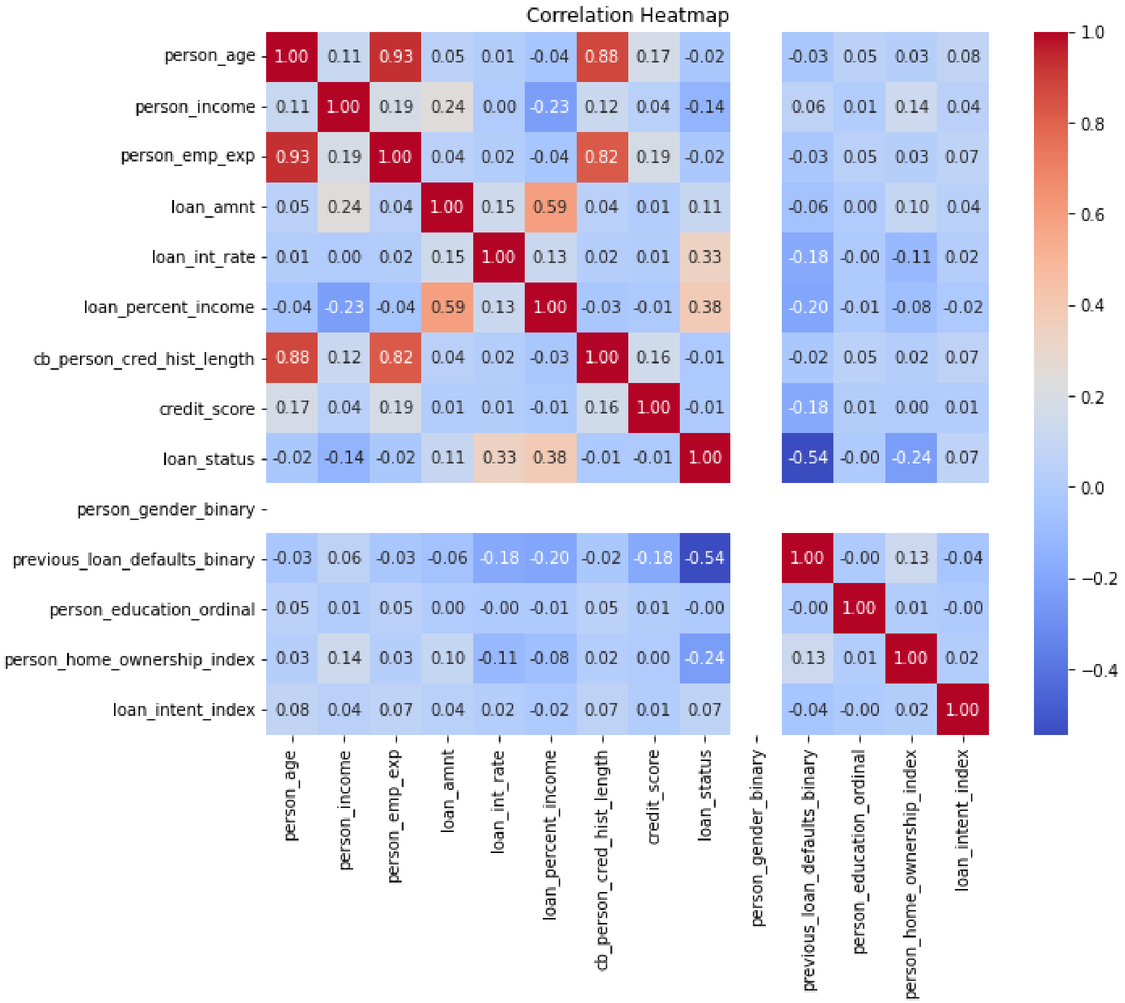
- Gender:
  Loan approval outcomes show minimal variation across genders, suggesting limited bias in approval decisions.
- Education Level:
  Applicants with Bachelor's and Master's degrees exhibit relatively higher approval counts compared to other education categories.
- Home Ownership:
  Renters account for a higher volume of applications and rejections, while mortgage holders display comparatively lower approval rates, indicating home ownership as an important risk indicator.
- Loan Intent:
  Loan purpose influences approval outcomes, with medical and debt consolidation loans showing higher approval counts than venture and education loans.
- Modeling Insight:
  These categorical patterns provide early insights into applicant risk profiles and support informed feature selection for predictive modeling.



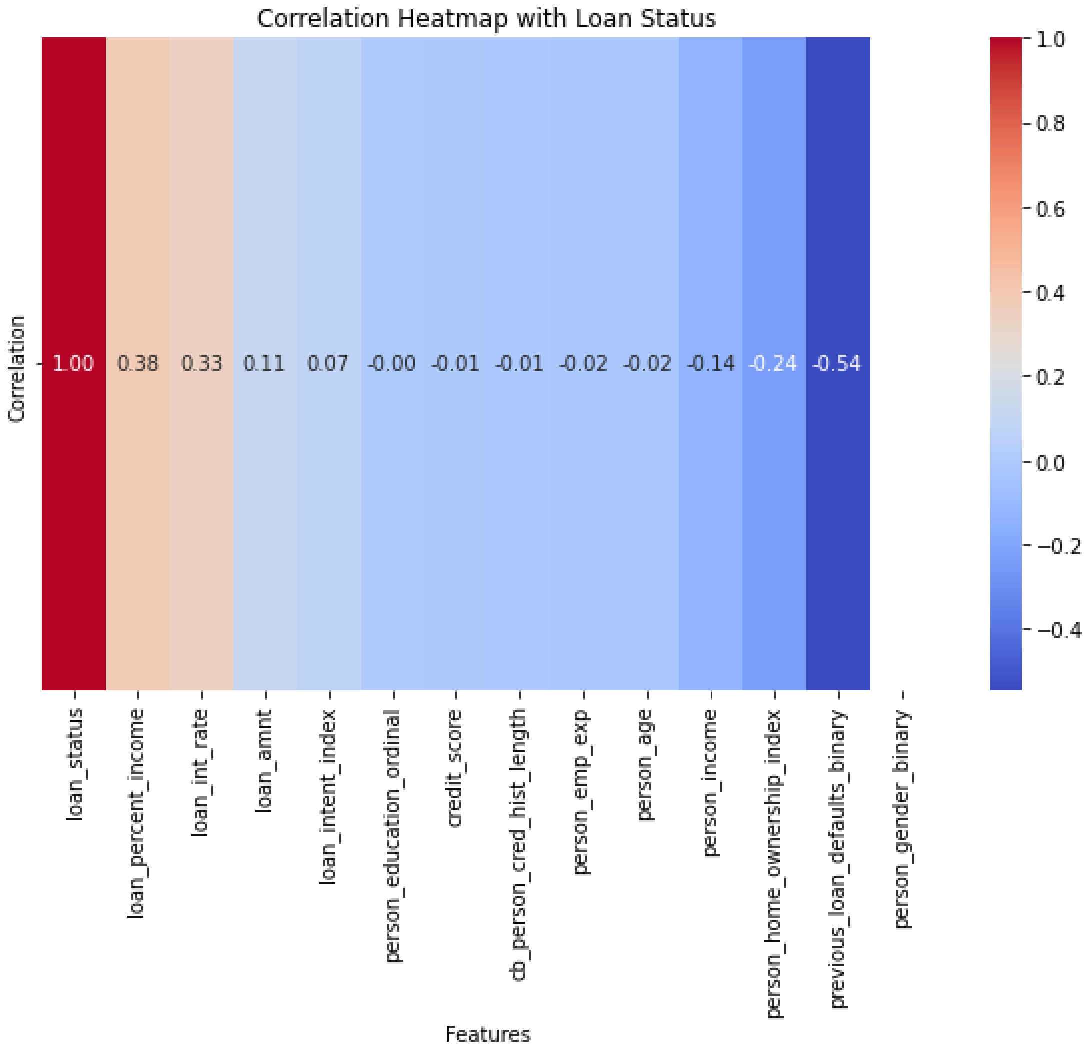
Correlation Heatmap:

- Variables like person_income and loan_amnt show a high correlation (close to 0.59), suggesting that as income increases, the loan amount also tends to increase.

- previous_loan_defaults_binary and loan_status show a negative correlation (-0.54), suggesting that previous defaults negatively impact loan approval.
- person_gender_binary has made no contribution for classification of approval or rejection of loan.

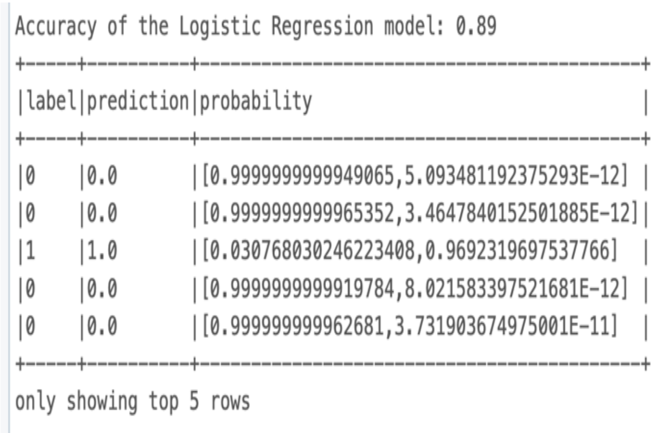


- loan_percent_income (0.38) and loan_int_rate (0.33) show significant positive correlations with loan_status.
- This suggests these features may be influential in determining loan approval.
- previous_loan_defaults_binary (-0.54) has a strong negative correlation, indicating that previous loan defaults significantly reduce the likelihood of loan approval.

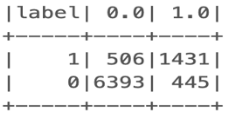


Data Analytics 1 – Logistic Regression

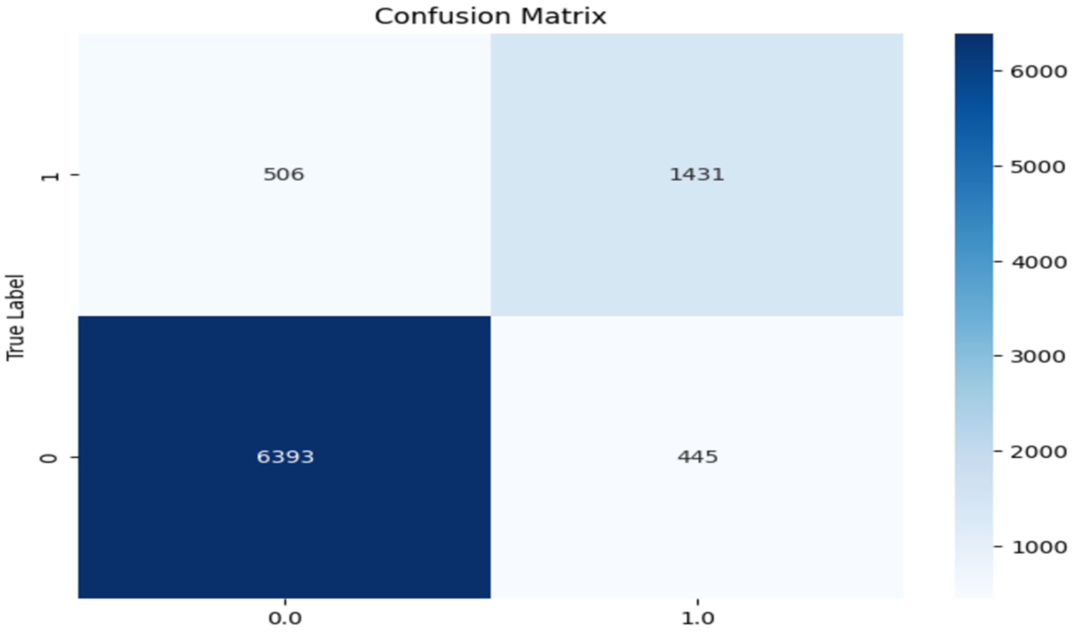
- Logistic Regression was used as a baseline model.
- - Achieved an accuracy of 89%.
- - Key insights: Higher loan-to-income ratios and no previous defaults increase approval chances.
- True Positives (TP):1431 instances were correctly classified as 1 (approved loans).
- False Negatives (FN):506 instances of 1 (approved loans) were misclassified as 0 (rejected loans). This indicates the model is missing a significant portion of approved loans.
- True Negatives (TN):6,393 instances were correctly classified as 0 (rejected loans).
- False Positives (FP):445 instances of 0 (rejected loans) were misclassified as 1 (approved loans).



```
Accuracy of the Logistic Regression model: 0.89
+-----+----------+-------------------------------------------+
|label|prediction|probability                                |
+-----+----------+-------------------------------------------+
|0    |0.0       |[0.9999999999949065,5.093481192375293E-12] |
|0    |0.0       |[0.9999999999965352,3.4647840152501885E-12]|
|1    |1.0       |[0.030768030246223408,0.9692319697537766]  |
|0    |0.0       |[0.9999999999919784,8.021583397521681E-12] |
|0    |0.0       |[0.999999999962681,3.731903674975001E-11]  |
+-----+----------+-------------------------------------------+
only showing top 5 rows
```



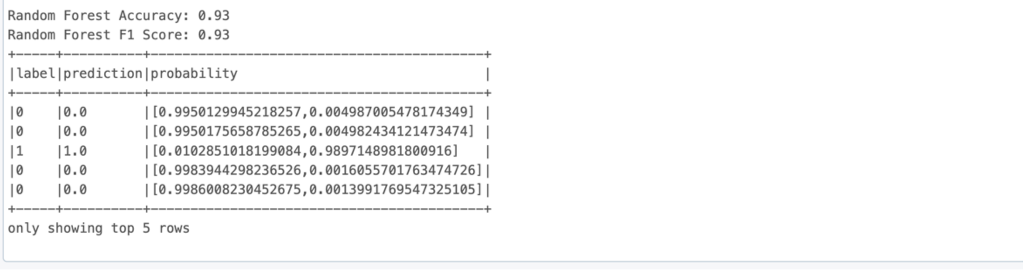
```
|label| 0.0| 1.0|
+-----+----+----+
|    1| 506|1431|
|    0|6393| 445|
+-----+----+----+
```



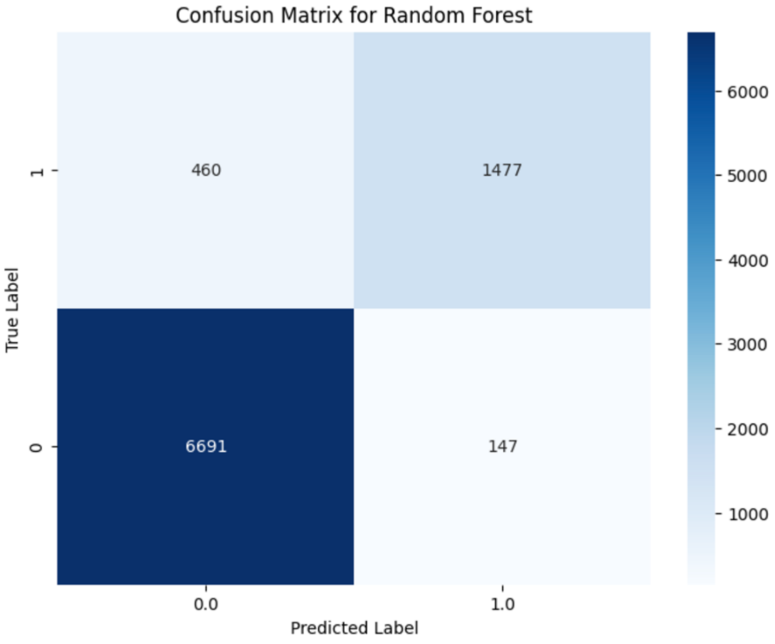
Data Analytics 2 – Random Forest

- Achieved accuracy of 93%
- Feature importance identified previous_loan_defaults,loan_percent_income ,loan_int_rate are most influential.
- True Positives (TP):1477 instances were correctly classified as 1 (approved loans).
- False Negatives (FN):460 instances of 1 (approved loans) were misclassified as 0 (rejected loans). This indicates the model is missing a significant portion of approved loans.
- True Negatives (TN):6,691 instances were correctly classified as 0 (rejected loans).

- False Positives (FP):147 instances of 0 (rejected loans) were misclassified as 1 (approved loans).

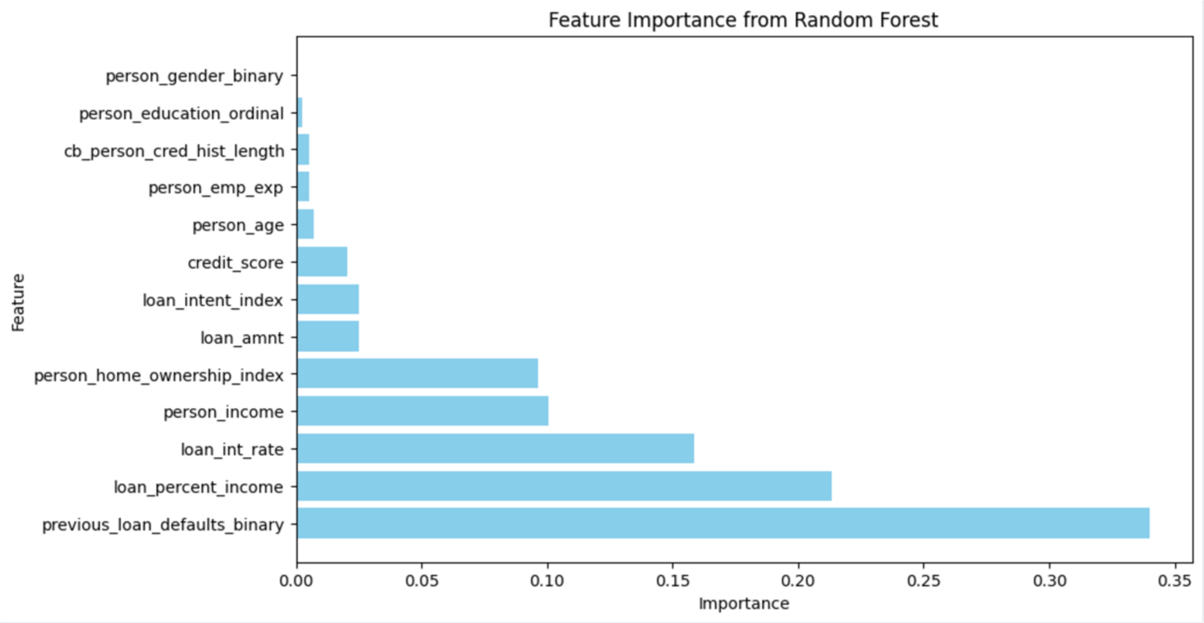


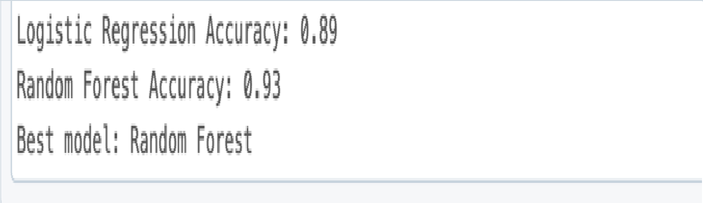
```
Random Forest Accuracy: 0.93
Random Forest F1 Score: 0.93
+-----+----------+------------------------------------------+
|label|prediction|probability                               |
+-----+----------+------------------------------------------+
|0    |0.0       |[0.9950129945218257,0.004987005478174349] |
|0    |0.0       |[0.9950175658785265,0.004982434121473474] |
|1    |1.0       |[0.0102851018199084,0.9897148981800916]   |
|0    |0.0       |[0.9983944298236526,0.0016055701763474726]|
|0    |0.0       |[0.9986008230452675,0.0013991769547325105]|
+-----+----------+------------------------------------------+
only showing top 5 rows
```



Findings

- Key Insights:
  - loan_percent_income, loan_int_rate, and previous_loan_defaults_binary are critical drivers.
  - Random Forest outperformed Logistic Regression with an accuracy of 93%
- Impact:
  - Models provide a reliable basis for predicting loan approvals.
  - Renters and applicants with education loans have slightly higher rejection rates.





```
Logistic Regression Accuracy: 0.89
Random Forest Accuracy: 0.93
Best model: Random Forest
```

Business Implications/Intelligence

Actionable Recommendations:

- Focus on applicants with high loan_percent_income for targeted loan offers.
- Reduce risk by closely evaluating applicants with previous_loan_defaults_binary = 1.
- Optimize loan approval processes using model classifications.

Potential Benefits:

- Increase approval efficiency.
- Income and Credit History are key predictors of loan approval; prioritize these factors in decision-making
- Employment Status and Education Level should be factored into loan decisions for improved applicant profiling
- Implement strategies to assist borrowers with prior defaults in improving creditworthiness.
- Target educational programs to improve financial literacy among potential applicants.

Conclusion:

- The project successfully identified key factors influencing loan approvals.
- Random forest emerged as the best predictive model with 93% accuracy.
- Insights can guide policy adjustments and targeted customer support

Future Work:

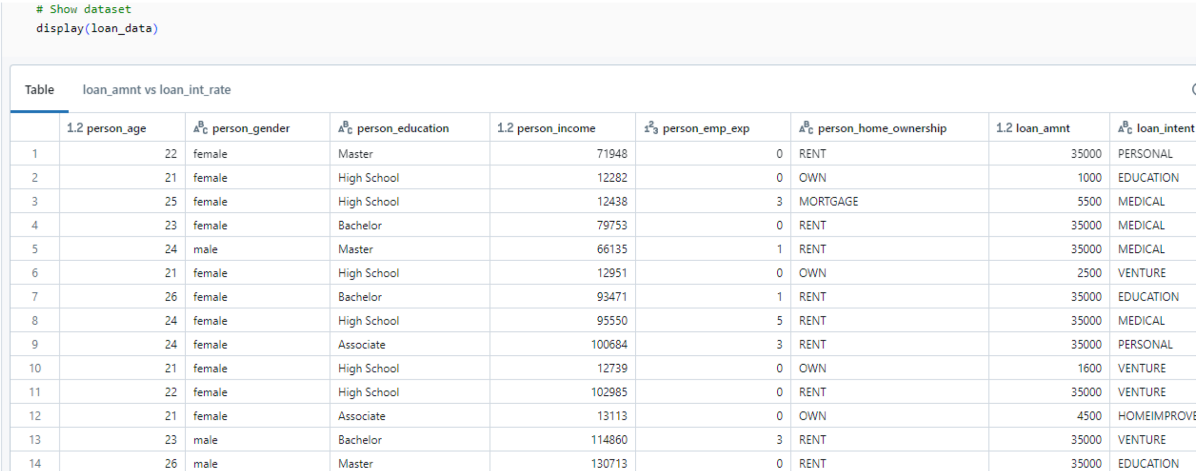
- Experiment with other advanced models (e.g., Gradient Boosting).

- Gather additional data to improve model accuracy further

References

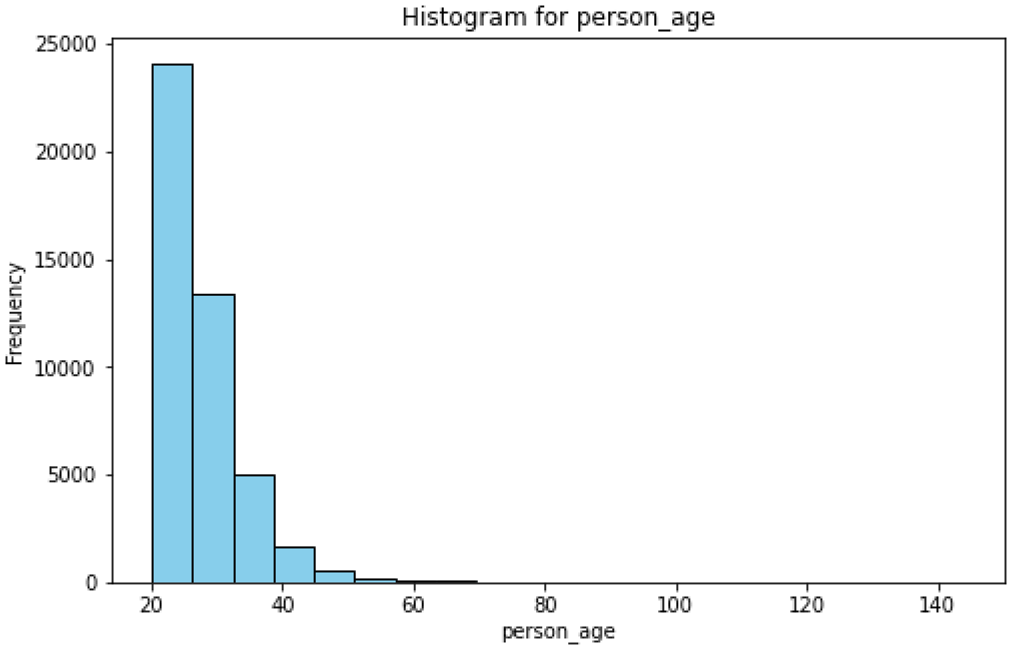
- *https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data*
- *https://www.incharge.org/blog/what-affects-your-ability-to-get-a-home-loan/*
- *Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann*

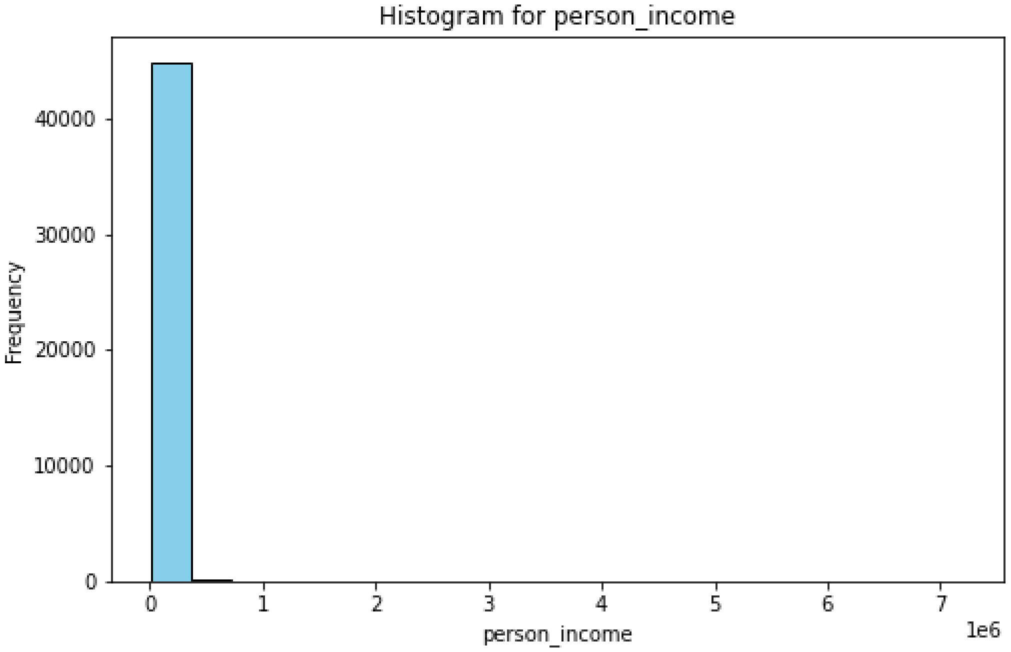
Appendix

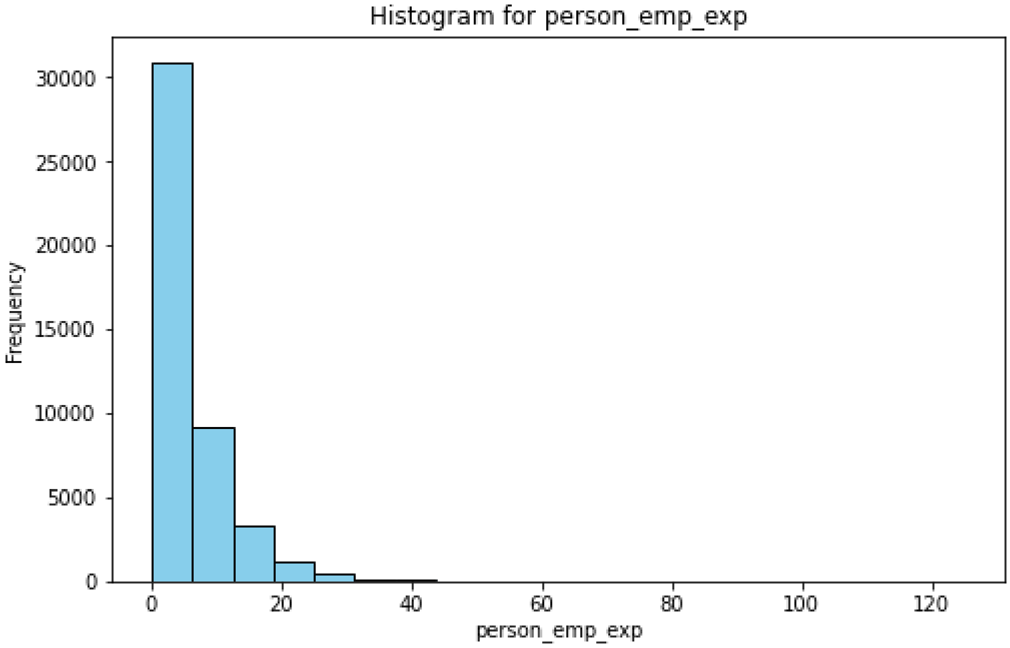


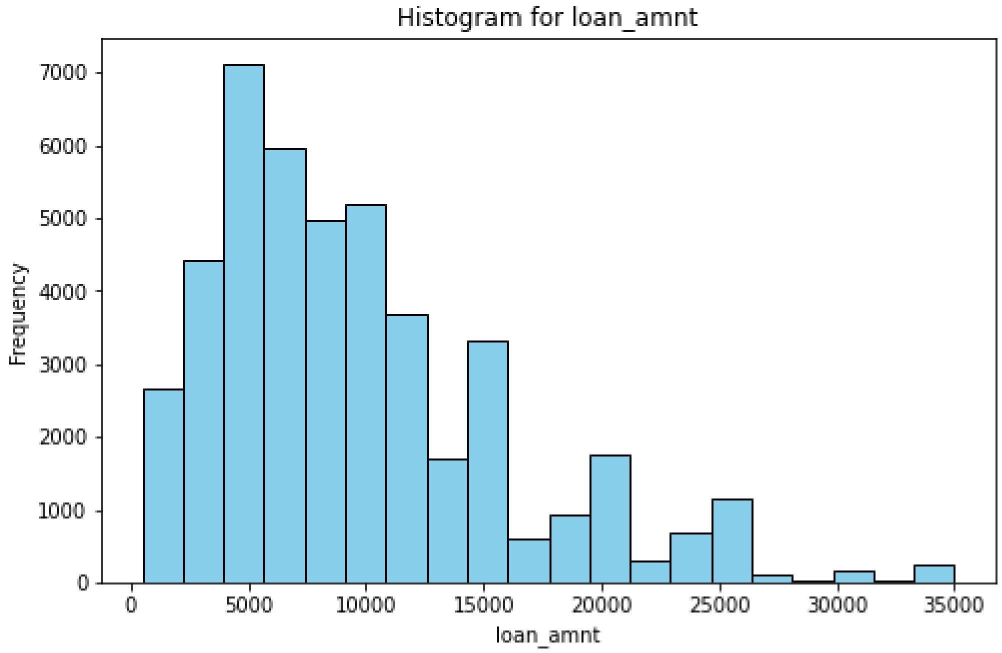
```
# Show dataset
display(loan_data)
```

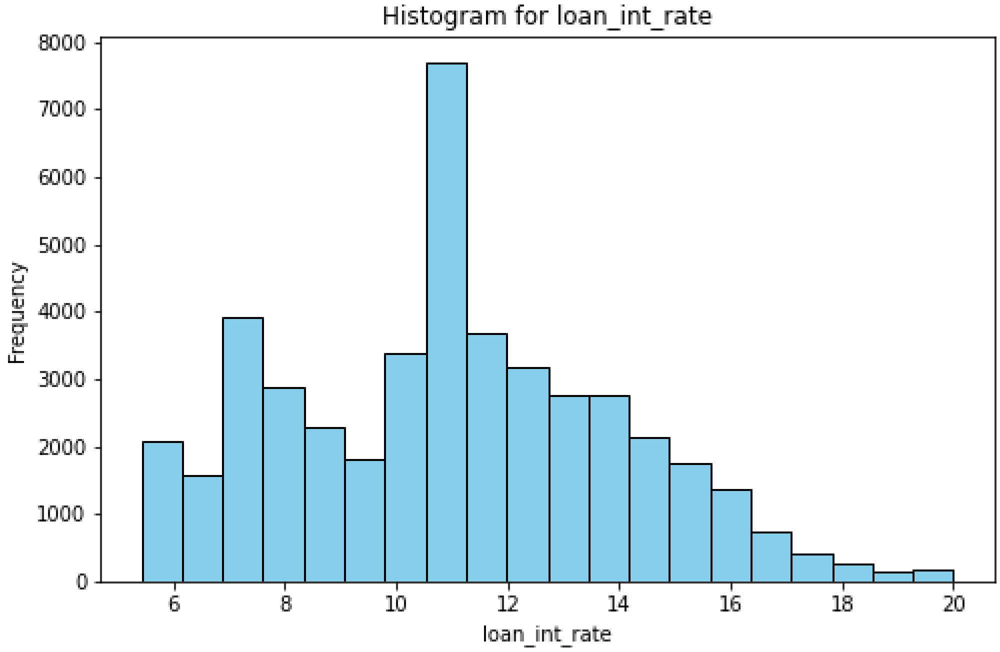
	1.2 person_age	person_gender	person_education	1.2 person_income	person_emp_exp	person_home_ownership	1.2 loan_amnt	loan_intent
1	22	female	Master	71948	0	RENT	35000	PERSONAL
2	21	female	High School	12282	0	OWN	1000	EDUCATION
3	25	female	High School	12438	3	MORTGAGE	5500	MEDICAL
4	23	female	Bachelor	79753	0	RENT	35000	MEDICAL
5	24	male	Master	66135	1	RENT	35000	MEDICAL
6	21	female	High School	12951	0	OWN	2500	VENTURE
7	26	female	Bachelor	93471	1	RENT	35000	EDUCATION
8	24	female	High School	95550	5	RENT	35000	MEDICAL
9	24	female	Associate	100684	3	RENT	35000	PERSONAL
10	21	female	High School	12739	0	OWN	1600	VENTURE
11	22	female	High School	102985	0	RENT	35000	VENTURE
12	21	female	Associate	13113	0	OWN	4500	HOMEIMPROVE
13	23	male	Bachelor	114860	3	RENT	35000	VENTURE
14	26	male	Master	130713	0	RENT	35000	EDUCATION

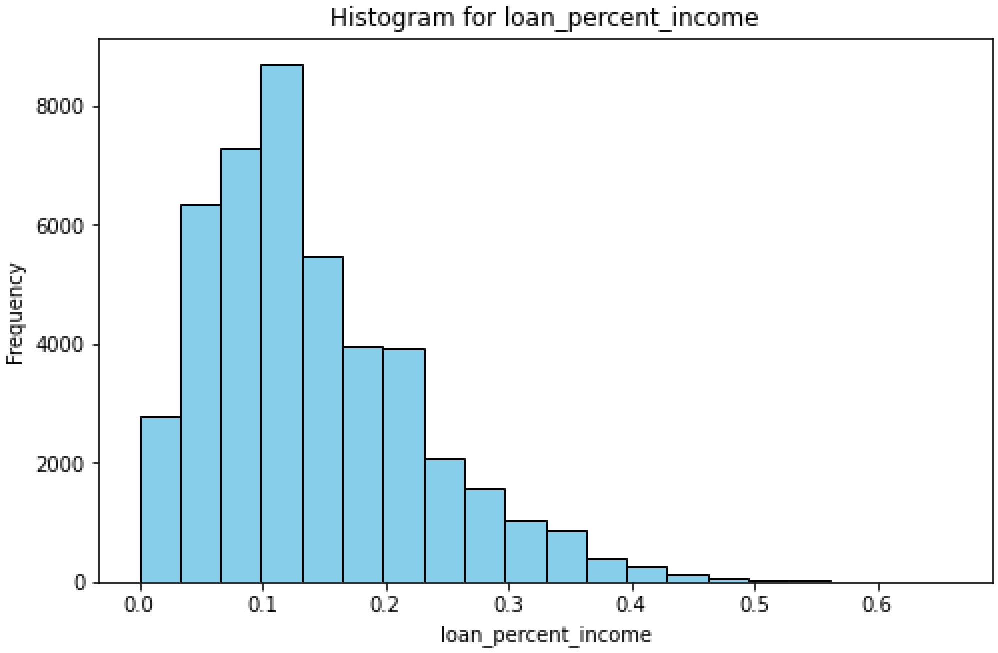


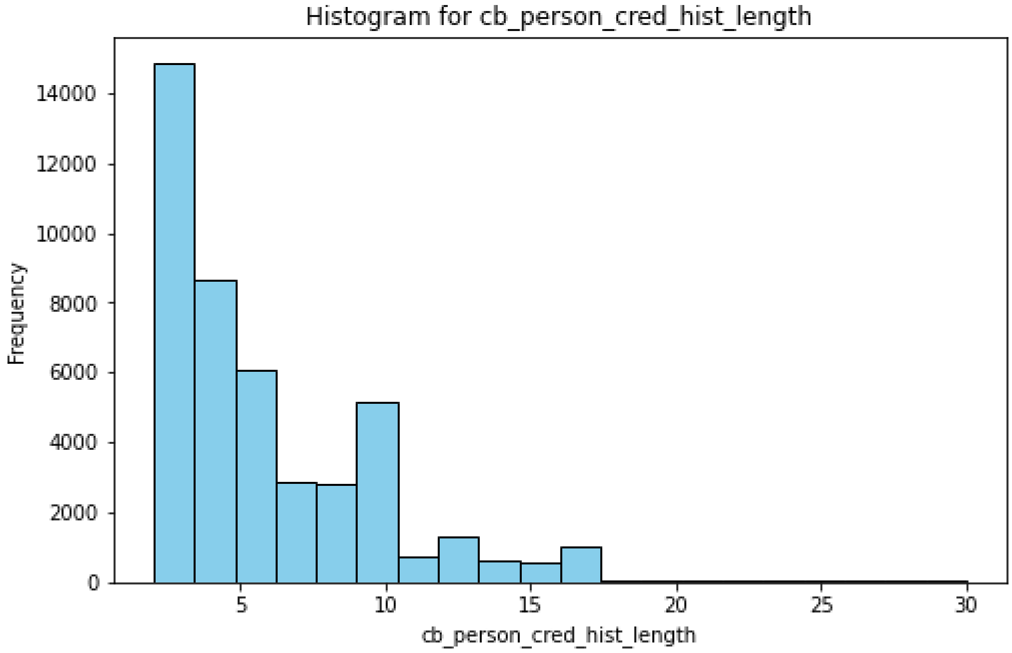

Histogram for person_income
Frequency
0
10000
20000
30000
40000
0
1
2
3
4
5
6
7
1e6
person_income

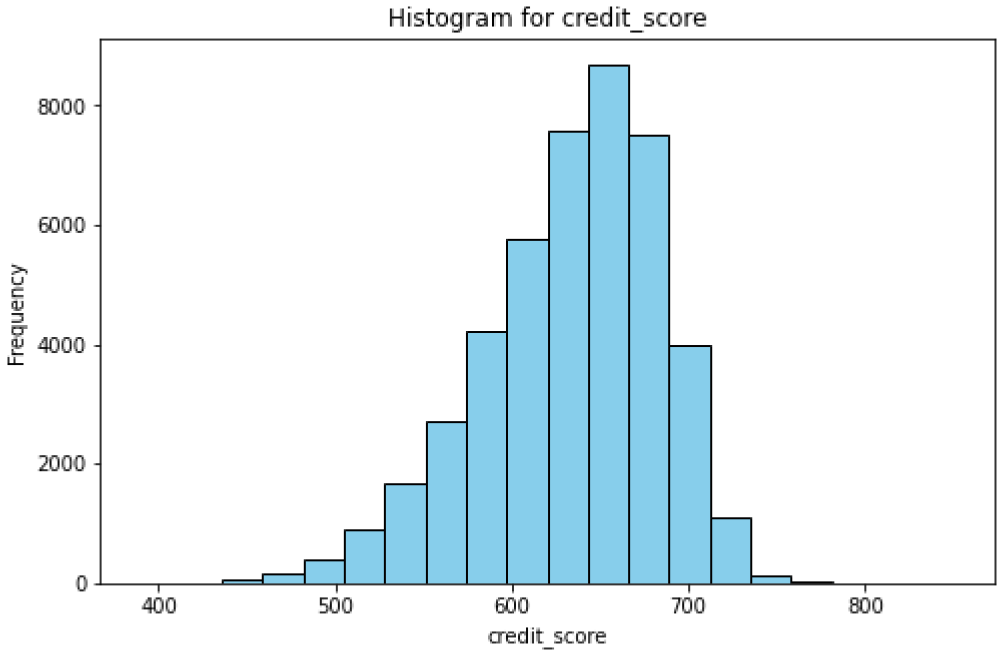

Histogram for person_emp_exp
Frequency
0
5000
10000
15000
20000
25000
30000
0
20
40
60
80
100
120
person_emp_exp

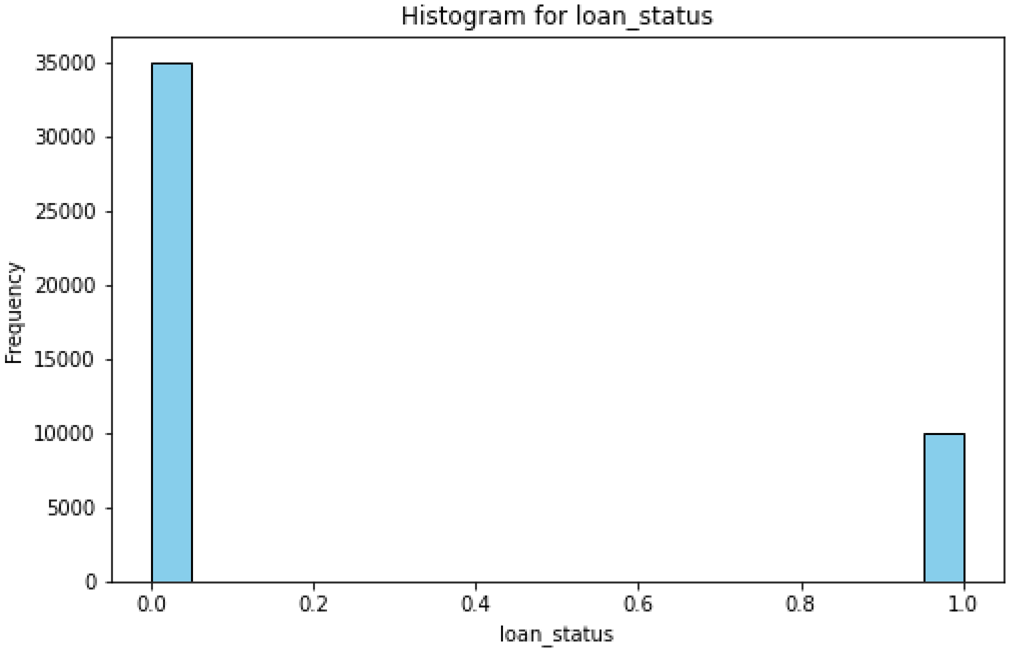

Histogram for loan_amnt
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
5000
10000
15000
20000
25000
30000
35000
loan_amnt

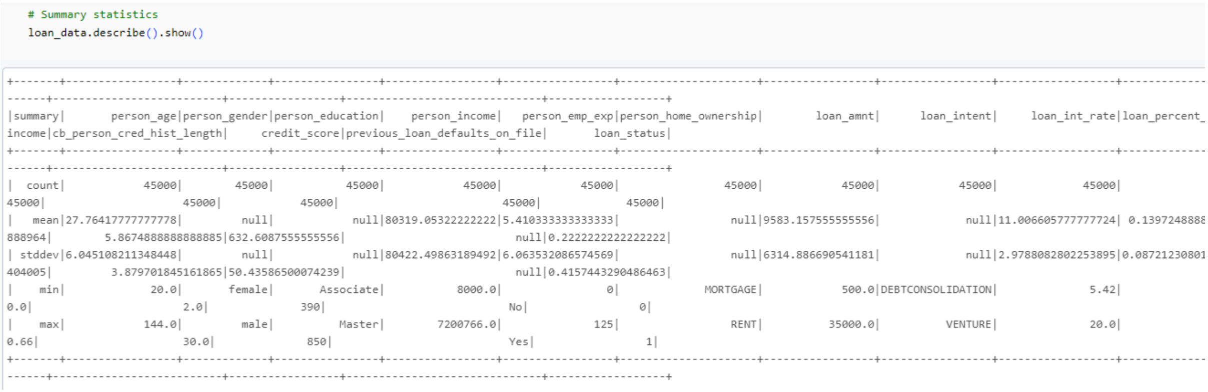

Histogram for loan_int_rate
Frequency
0
1000
2000
3000
4000
5000
6000
7000
8000
6
8
10
12
14
16
18
20
loan_int_rate


Histogram for loan_percent_income
Frequency
0
2000
4000
6000
8000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
loan_percent_income


Histogram for cb_person_cred_hist_length
Frequency
0
2000
4000
6000
8000
10000
12000
14000
5
10
15
20
25
30
cb_person_cred_hist_length

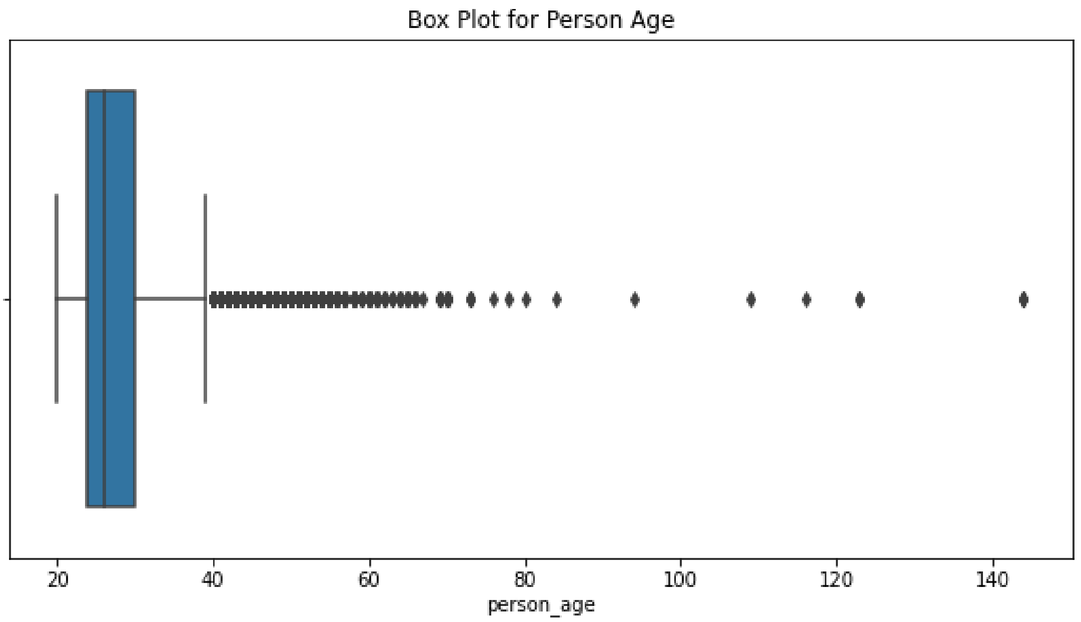

Histogram for credit_score
Frequency
0
2000
4000
6000
8000
400
500
600
700
800
credit_score

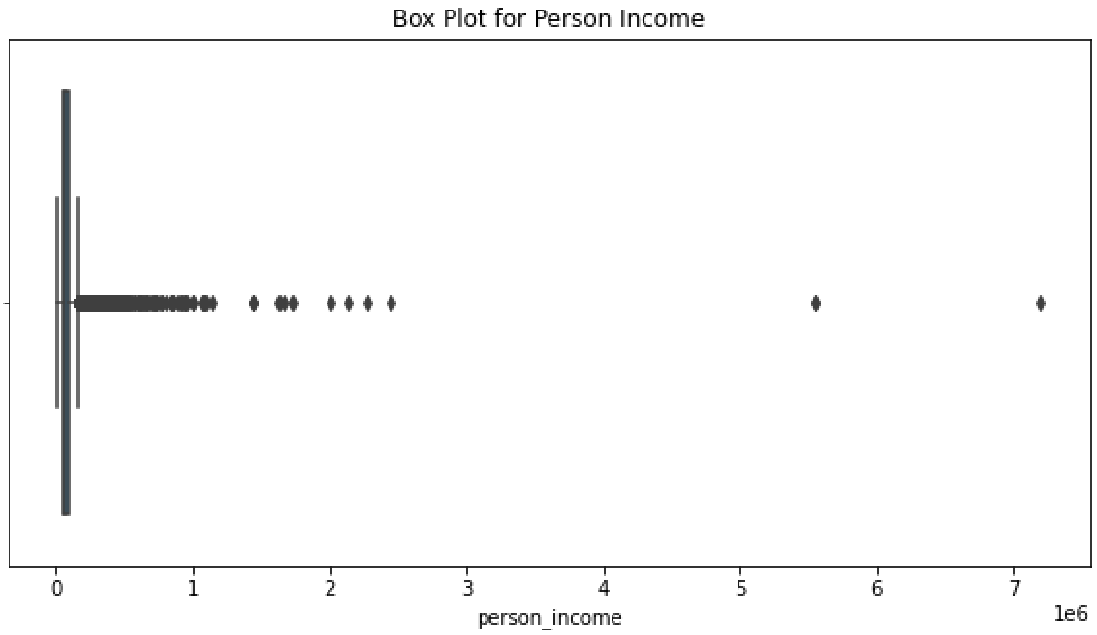

Histogram for loan_status
Frequency
0
5000
10000
15000
20000
25000
30000
35000
0.0
0.2
0.4
0.6
0.8
1.0
loan_status

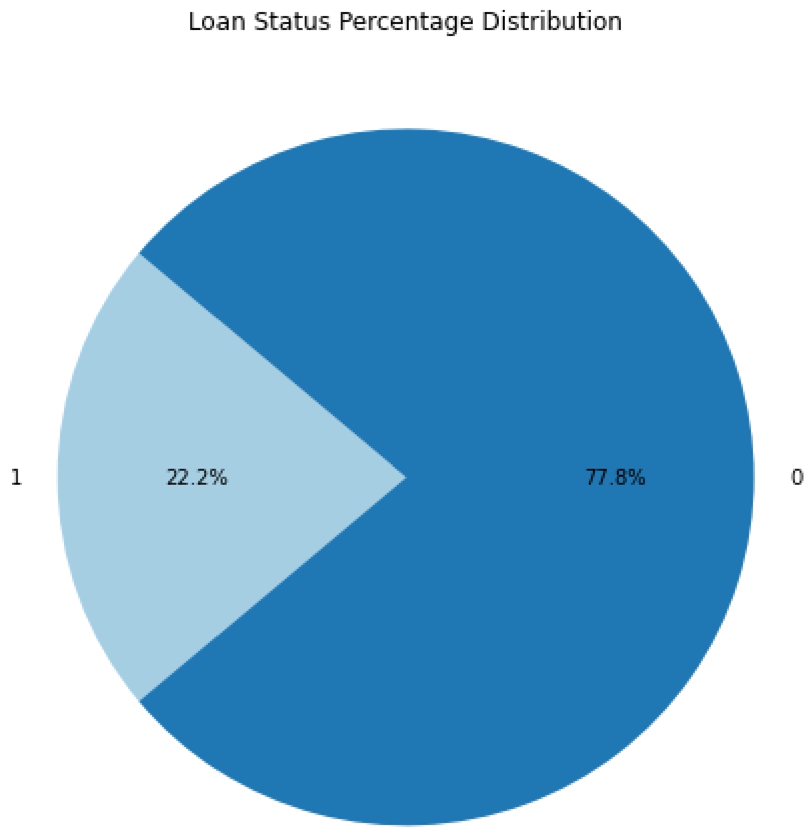


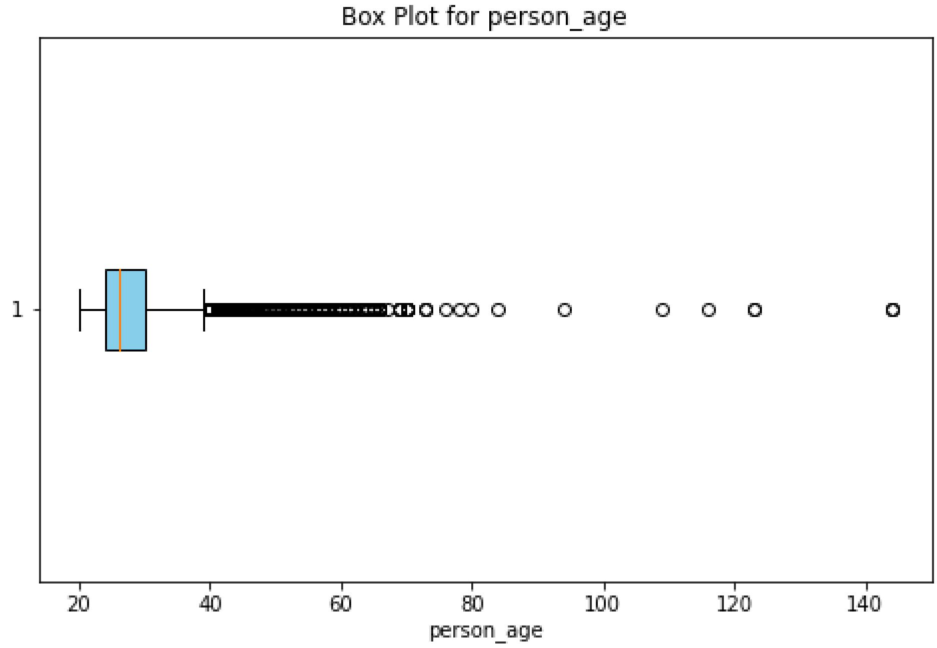
```
# Summary statistics
loan_data.describe().show()
```

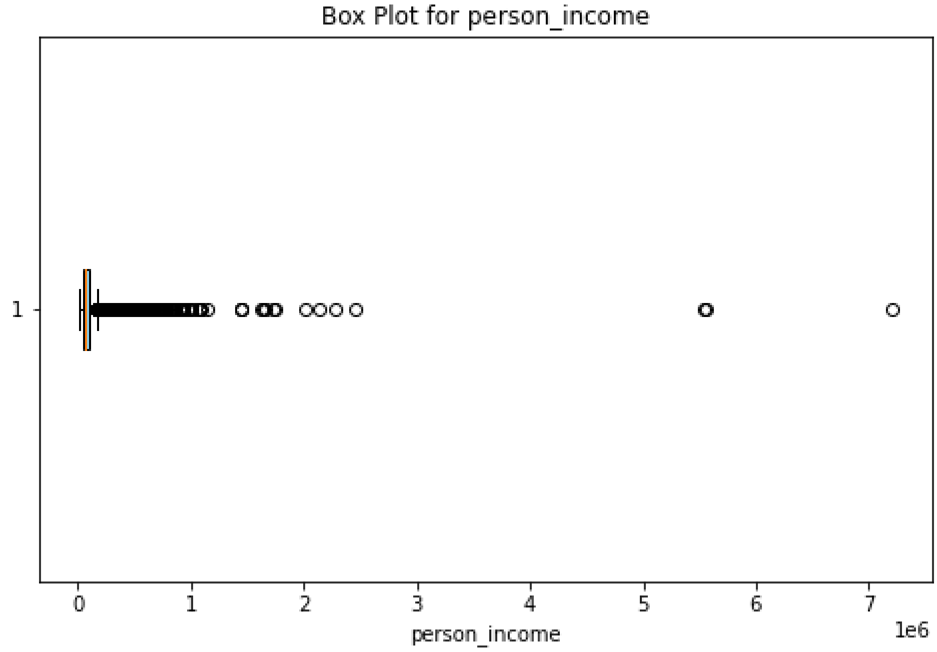
```
+-------+-----------------+------------+---------------+----------------+----------------+--------------------+-----------------+----------------+-----------------+------------
------+-----------------------+----------------+-----------------------------+----------------+
|summary|       person_age|person_gender|person_education|    person_income|   person_emp_exp|person_home_ownership|        loan_amnt|      loan_intent|      loan_int_rate|loan_percent_
income|cb_person_cred_hist_length|     credit_score|previous_loan_defaults_on_file|       loan_status|
+-------+-----------------+------------+---------------+----------------+----------------+--------------------+-----------------+----------------+-----------------+------------
------+-----------------------+----------------+-----------------------------+----------------+
|  count|            45000|        45000|          45000|           45000|           45000|               45000|           45000|           45000|            45000|
45000|                  45000|           45000|                        45000|           45000|
|   mean|27.764177777777778|         null|           null|80319.05322222222|5.410333333333333|                null|9583.157555555556|            null|11.006605777777724| 0.1397248888
888964|        5.867488888888885|632.6087555555556|                         null|0.2222222222222222|
| stddev|6.045108211348448|         null|           null|80422.49863189492|6.063532086574569|                null|6314.886690541181|            null|2.9788082802253895|0.08721230801
404005|        3.879701845161865|50.43586500074239|                         null|0.4157443290486463|
|    min|             20.0|       female|      Associate|          8000.0|               0|            MORTGAGE|           500.0|DEBTCONSOLIDATION|             5.42|
0.0|                    2.0|             390|                           No|               0|
|    max|            144.0|         male|         Master|       7200766.0|             125|                RENT|         35000.0|         VENTURE|             20.0|
0.66|                   30.0|             850|                          Yes|               1|
+-------+-----------------+------------+---------------+----------------+----------------+--------------------+-----------------+----------------+-----------------+------------
------+-----------------------+----------------+-----------------------------+----------------+
```

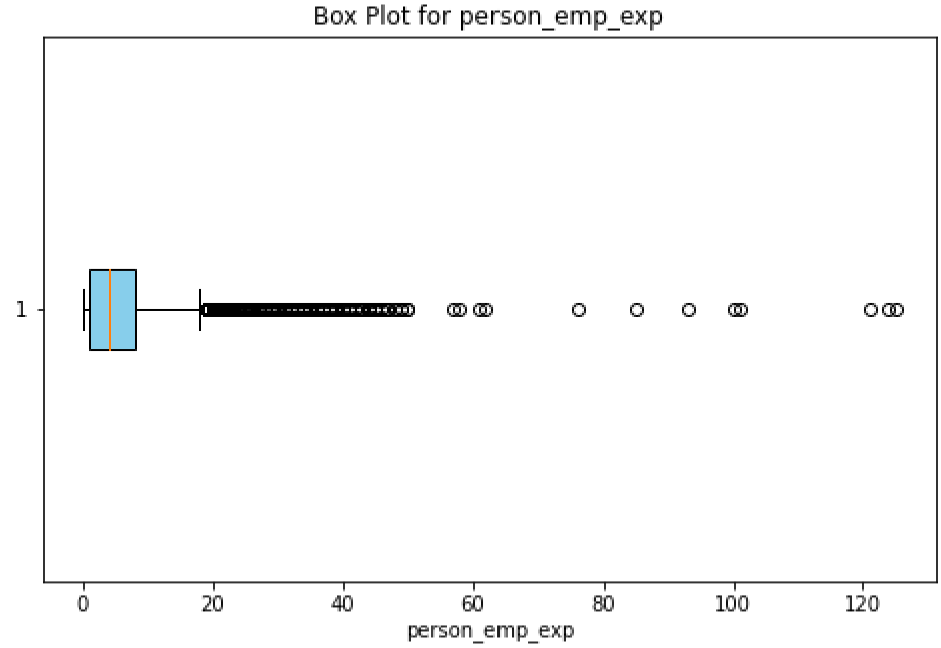


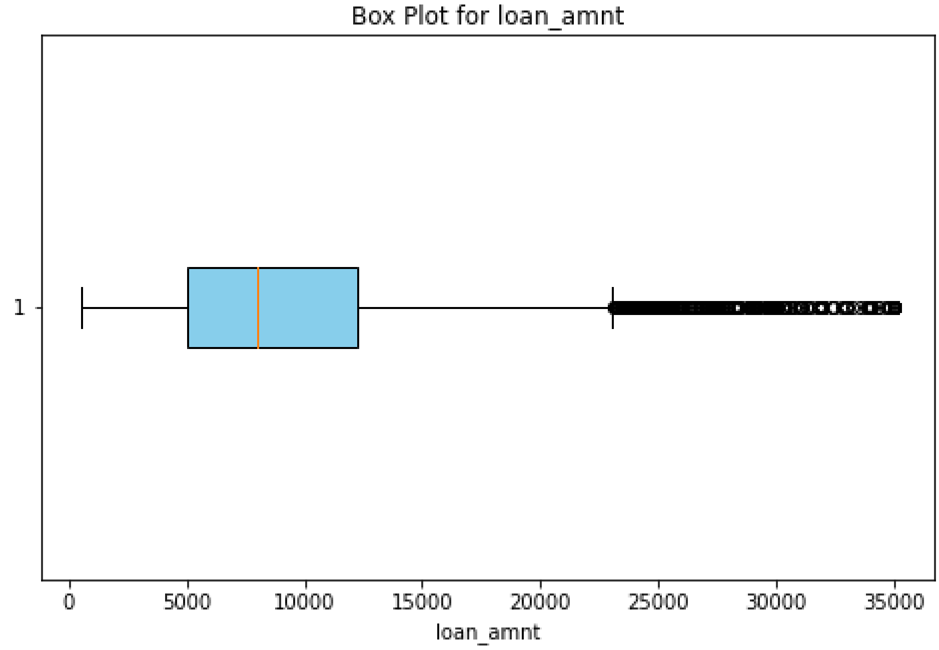


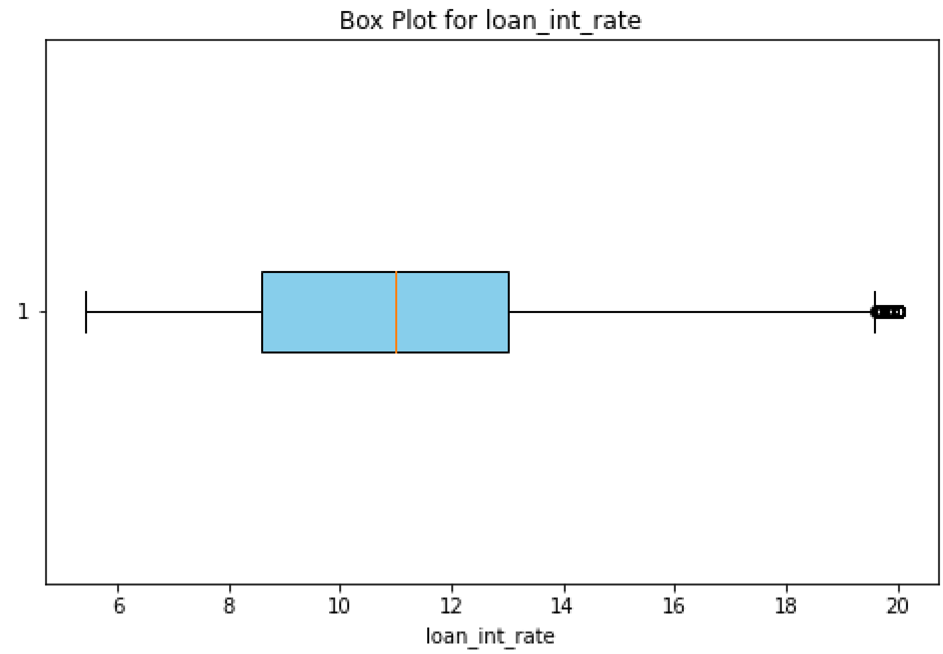

Loan Status Percentage Distribution
1
22.2%
77.8%
0

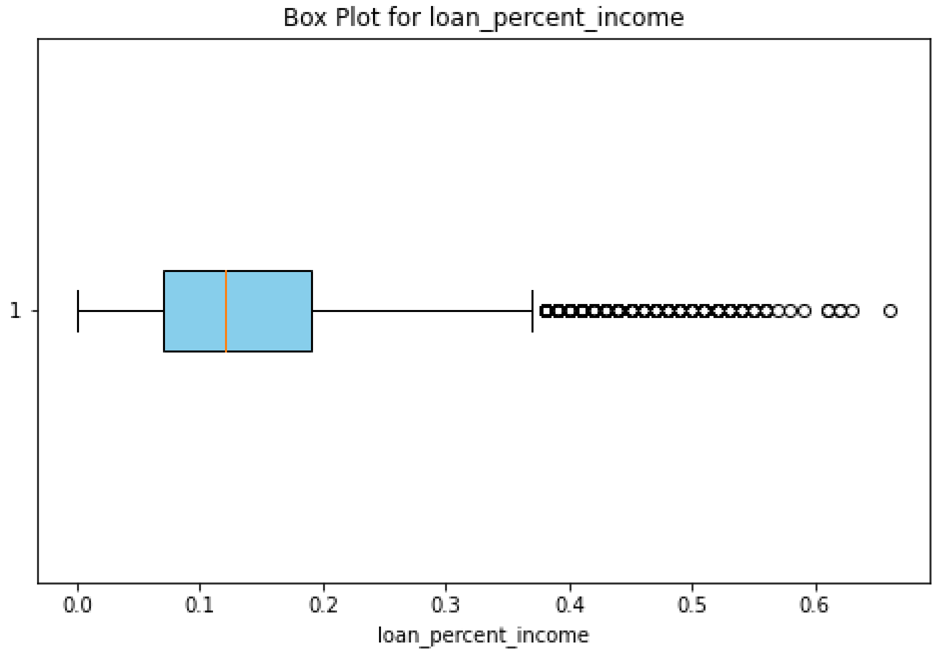

Box Plot for person_age
1
20
40
60
80
100
120
140
person_age

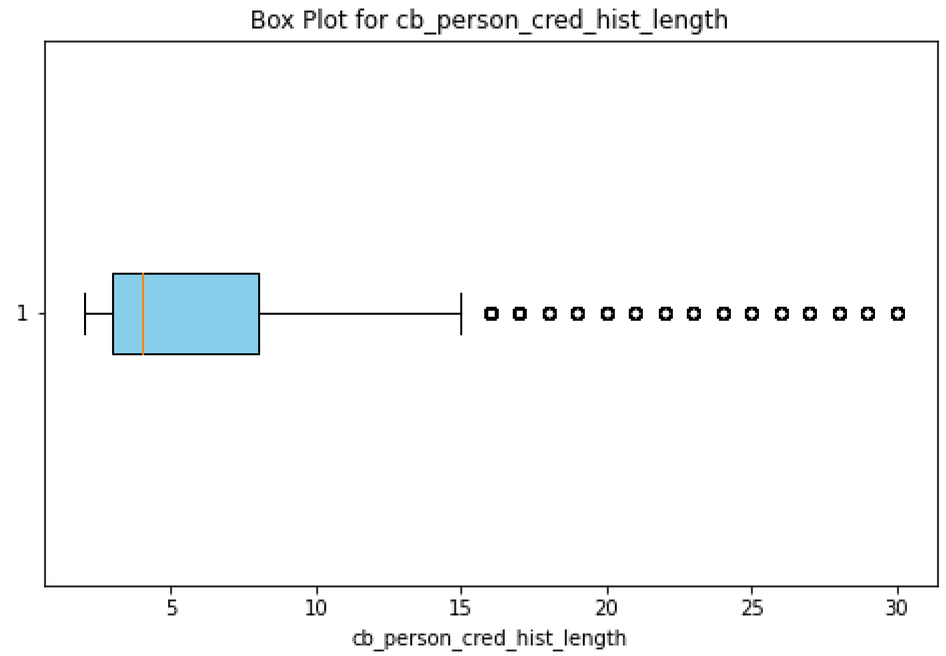

Box Plot for person_income
1
0
1
2
3
4
5
6
7
1e6
person_income

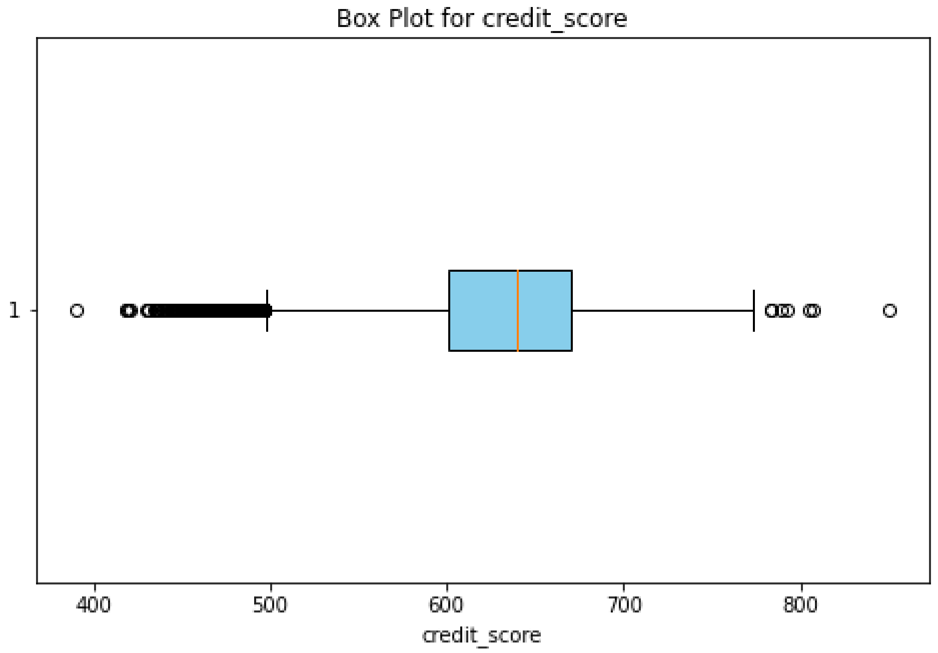

Box Plot for person_emp_exp
1
0
20
40
60
80
100
120
person_emp_exp

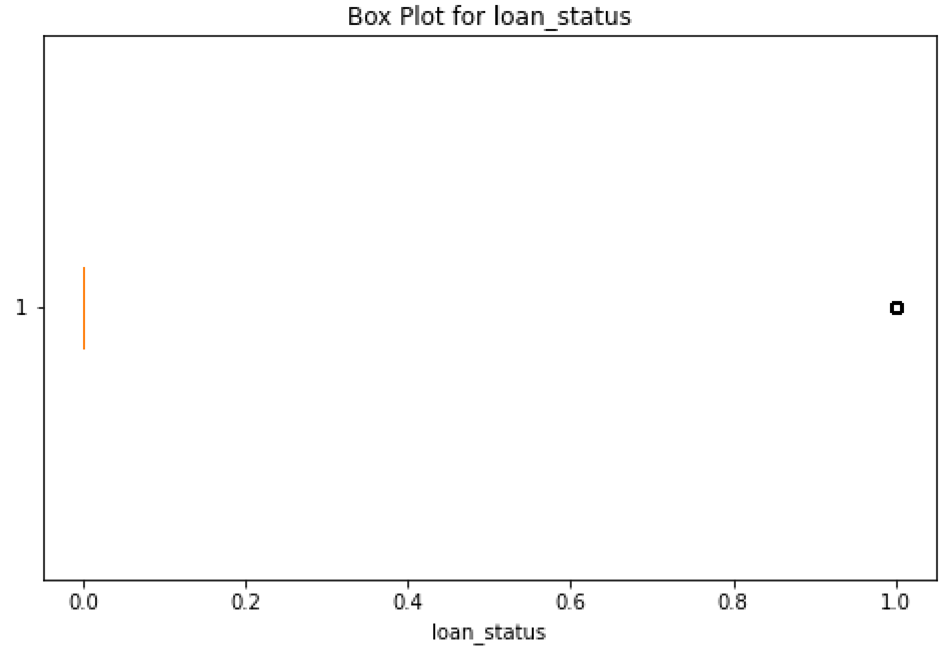

Box Plot for loan_amnt
1
0
5000
10000
15000
20000
25000
30000
35000
loan_amnt

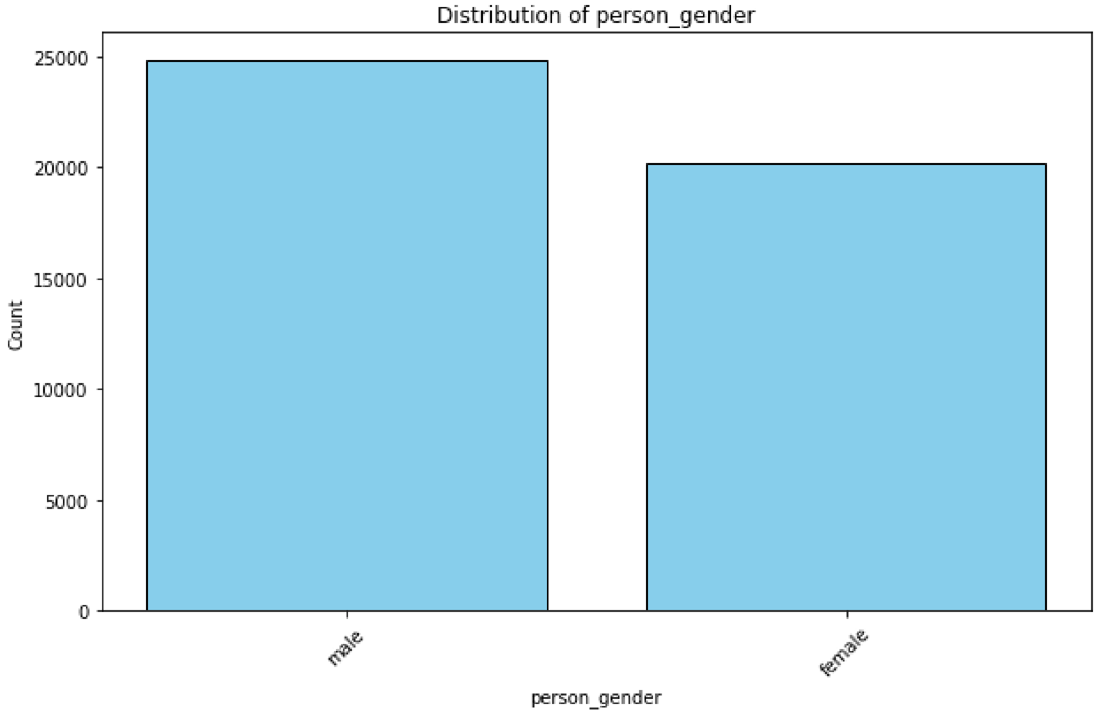

Box Plot for loan_int_rate
1
6
8
10
12
14
16
18
20
loan_int_rate

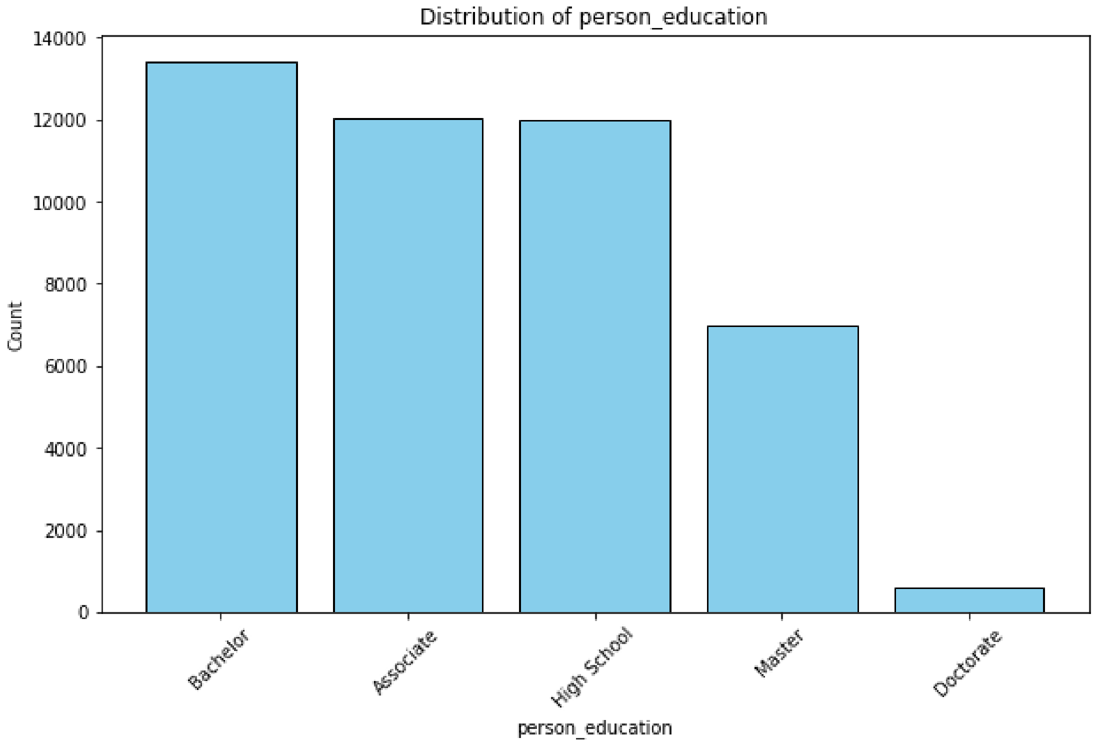

Box Plot for loan_percent_income
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
loan_percent_income

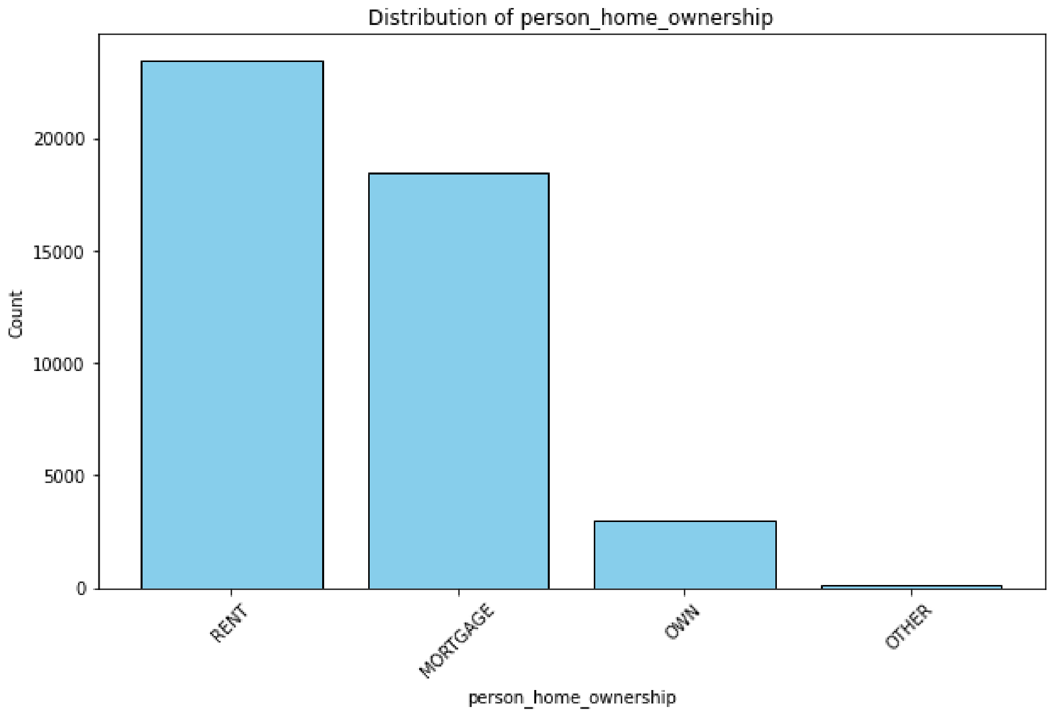

Box Plot for cb_person_cred_hist_length
1
5
10
15
20
25
30
cb_person_cred_hist_length

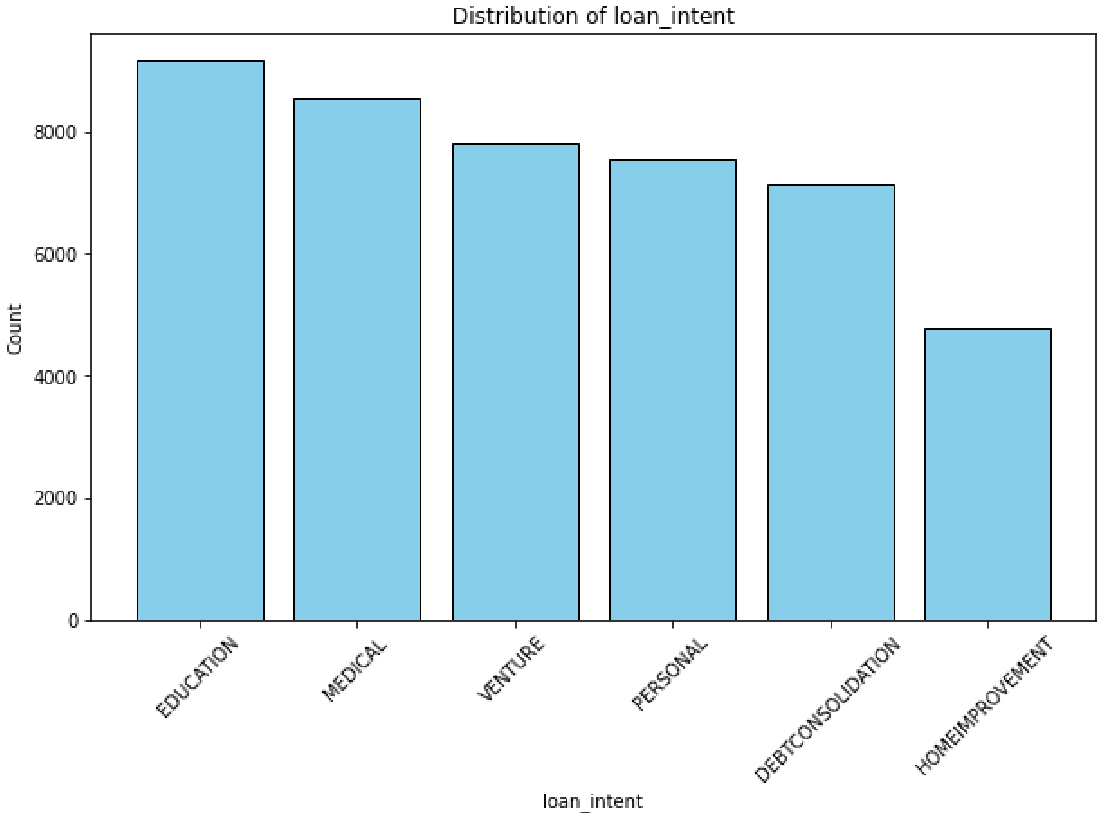

Box Plot for credit_score
1
400
500
600
700
800
credit_score

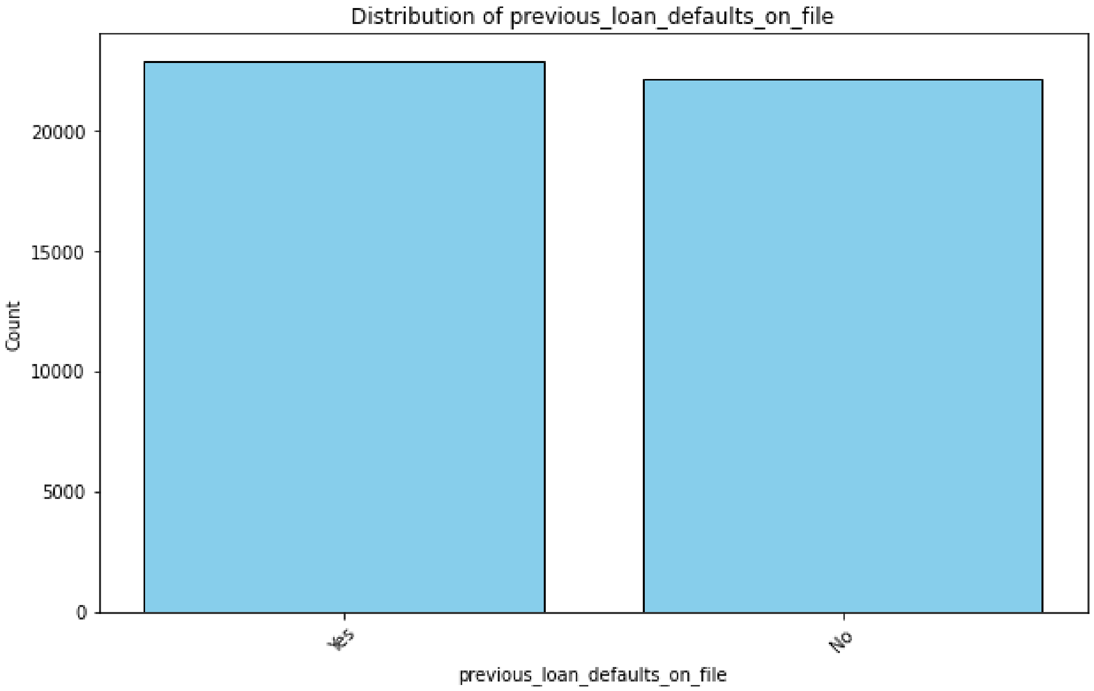

Box Plot for loan_status
1
0.0
0.2
0.4
0.6
0.8
1.0
loan_status


Distribution of person_gender
25000
20000
15000
10000
5000
0
Count
male
female
person_gender


Distribution of person_education
Count
0
2000
4000
6000
8000
10000
12000
14000
Bachelor
Associate
High School
Master
Doctorate
person_education


Distribution of person_home_ownership
Count
0
5000
10000
15000
20000
RENT
MORTGAGE
OWN
OTHER
person_home_ownership


Distribution of loan_intent
Count
0
2000
4000
6000
8000
EDUCATION
MEDICAL
VENTURE
PERSONAL
DEBTCONSOLIDATION
HOMEIMPROVEMENT
loan_intent


Distribution of previous_loan_defaults_on_file
Count
0
5000
10000
15000
20000
Yes
No
previous_loan_defaults_on_file