# Analytics Based Project (Step 2- Group B)

## Abstract

In this document, we outline the advancements executing ETL (Extract, Transform, Load), and conducting Exploratory Data Analysis (EDA) to comprehend and cleanse the data while uncovering any underlying potential issues. We present an overview of our initial discoveries and observations resulting from the exploratory data analysis (EDA) conducted in this preliminary report. For further technical specifics, please refer to the attached python file.

At each juncture, we will proceed thoughtfully, recognizing the assumptions and potential pitfalls we make and considering the ramifications they entail. This analysis, as stated in the prior step, will concentrate on 'expired' loans issued in 2019 (charged-off/fully paid).

## ETL Workflow Breakdown

**1. Data Cohesion and Integration:** We combined quarterly loan data from 2018 and 2019 into cohesive datasets, removing unnecessary headers. We'll focus on analyzing only the 'expired' loans, which make up 69.64% of the dataset.

**2. Data cleaning & Feature Selection:** In this step, our focus was on eliminating any data elements that could distort or hinder our analysis, implemented by:

### Feature Based

**A. Removing Features Based on 50% NaN Threshold:** Features dominated by missing values were omitted as their inclusion could undermine the model's integrity. Relying on imputation for these could introduce bias and overfitting. Although this approach may seem stringent, we carefully reviewed discarded columns to ensure no essential data capable or in need of interpolation was removed.

**B. Removing Features with No Variance:** We removed features that displayed only one unique value across all observations. Such features add unnecessary complexity without contributing predictive power, thus simplifying the model and enhancing performance.

**C. Preventing Data Leakage by Omitting Post-Loan Features:** To avoid data leakage and thus overly optimistic performance estimates and poor generalization, we excluded features that are only known after a loan has been funded. Our approach involved both Business understanding and analyzing the consistency of feature values over the years. By assessing changes in attributes for each loan between 2018 and 2019, we determined whether features reflected conditions at the time of loan application or were influenced by events post-issuance.

**D.Addressing High Cardinality:** we will tackle high cardinality within categorical features to enhance model simplicity and generalizability. This involves converting features with a vast range of values into formats that are more manageable for predictive modeling**.**we'll delve deeper into it during the subsequent stage of data preparation

**E. Business understanding:** Informed by a deep understanding of the business context and strategic project objectives, we selectively omitted features that could compromise the model's effectiveness.

**F.High Correlation Analysis:** This code identifies highly correlated numeric columns in the DataFrame expired_2019 and selects one feature from each group to reduce redundancy and multicollinearity. By keeping only one feature from each correlated group, we maintain data integrity while improving model interpretability and performance.

## Instance Based

**A. Duplicates** - no duplicates where detected

**B. Loans missing critical features -** We opted to disregard a select few features that we deemed essential for comprehending the loan, given their absence.

**C. High Nan Loans** - Removing loans with many NaNs maintains data integrity and quality, preventing inaccurate analyses and biased results

.

**D. Joint applicants** -For simplicity, we have chosen to temporarily exclude joint loan applications, given their marginal representation in our database (1.73%) and the considerable number of secondary columns associated with them, which were removed based on the 50% threshold. (A separate analysis on joint loans may be considered at a later stage).

**E. Outliers (see appendix Ⅲ) -** We removed outliers to ensure analysis/model accuracy and reliability, utilizing a Gaussian kernel approach in the "remove_outliers_gaussian" function. This function calculates mean and standard deviation to establish bounds for outlier detection, creating a mask to identify and filter out these data points from the specified feature in the DataFrame even if not distributed normally (further specifications can be seen in the code).

**3. Log transformation (**see appendix Ⅳ**)-** A logarithmic transformation involves taking the logarithm of each data point. It's useful for skewed data because it compresses large values and spreads out small ones, making the distribution more symmetric and easier to analyze. We're still deciding whether to apply the logarithmic transformation, so it's pending until we decide our next steps, as outlined (see appendix no.9).

**4. Formatting Data**- This section acts as a pre-processing step where we adjust data types and add useful features for upcoming steps, such as "months_active.

## 5. Pitfalls:

**Incomplete Data Understanding**: Lack of thorough data comprehension may lead to misguided decisions.

**Rigid Thresholds:** Strict thresholds for data handling can result in valuable information loss or bias.

**Data Leakage Risk:** Incorrectly excluding post-loan features may bias analysis results. Information Loss in Feature Selection: Statistical feature selection might overlook relevant data relationships.

**Bias in Missing Data Handling:** Removing loans with many NaNs may introduce bias.

**Outlier Treatment Concerns:** Indiscriminate outlier removal can discard valid data points.

**Data Integrity Disregard:** Over-manipulating data may compromise analysis reliability.

**6. Checkpoint of overall changes-**After our first ETL phase, we have 56 features and 302,613 instances, marking a 9.6% decrease from the initial count  (see appendix Ⅱ).

## Insights Extracted From Exploratory Data Analysis (EDA)-

### Key Findings: General EDA-

**1. Loan Amount Distribution (see appendix no.1):**
The distribution of loan amounts indicates a majority of loans below $20,000, with a median around $10,000. While most loans fall within this range, outliers exceeding $35,000 suggest potential financial complexities for certain borrowers. The distribution indicates an asymmetric distribution.

**2. Interest Rate Distribution (see appendix no.2):**
Interest rates exhibit a right-skewed distribution, indicating that the majority of borrowers have lower interest rates. The median interest rate falls around 10%, with most rates clustered between 10% and 15%.

### Bivariate Analysis on Charged-off Loans:

As we progress, we chose to conduct a bivariate analysis on charged-off status loans that provides valuable insights into the factors associated with loan defaults. By examining relationships between independent variables and loan status, we identify predictors of default risk, crucial for risk assessment and lending strategy optimization.

**3. Loan Status Analysis- by loan amount/ interest rate (see appendix no.3):** Examination of loan status by loan amount and interest rate reveals insights into default risk. Charged-off loans tend to have higher amounts and interest rates compared to fully paid ones, emphasizing the significance of these factors in assessing borrower risk.

**4. Borrower Demographics (see appendix no.4 ):**
Analysis based on borrower demographics, such as annual income, state and home ownership, reveals correlations with loan repayment outcomes. Higher-income earners demonstrate a lower likelihood of charged-off loans, suggesting income's role in mitigating default risk. For home ownership, borrowers categorized as "OWN" or "MORTGAGE" exhibit a higher commitment to fulfill loan obligations.

**5. Loan Purpose (see appendix no.5 ):**
Debt consolidation emerges as the most prevalent reason for loan applications, followed by credit card refinancing and home improvement. Interestingly, loans for debt consolidation, house and small business show a higher proportion of charged-off instances compared to other purposes, highlighting the influence of loan purpose on default risk.

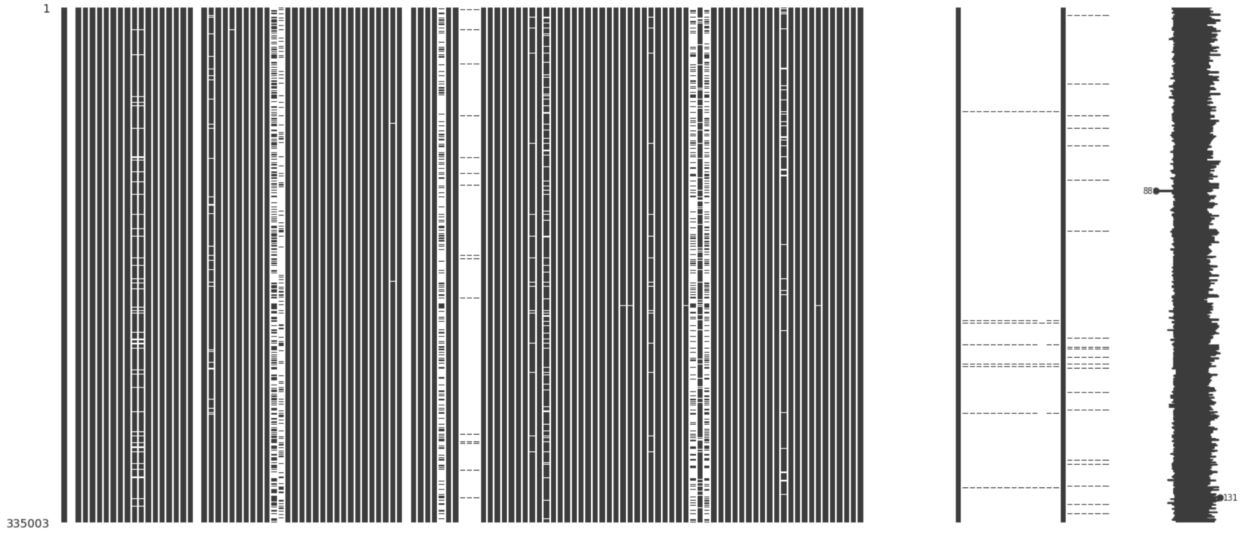
**6. Subgrades by Employee title (see appendix no.6 )**:
The crosstab analysis reveals the distribution of sub-grades among the top 15 employee titles. Notably, distinct patterns emerge, with certain job titles exhibiting a higher prevalence of specific sub-grades. For instance, Directors, Engineers, Presidents, and Vice Presidents are notably prominent categories within Grade A. This analysis offers valuable insights into the interplay between job titles and loan grades, shedding light on potential correlations between employment positions and loan grades.

**7. Grades & SubGrades (see appendix no.7):**
Notably, as we transition from lower grades/sub-grades (e.g., A/A1) to higher grades/sub-grades (e.g., G/G5), there is a noticeable increase in loan amount, interest rate and the proportion of charged-off loans.

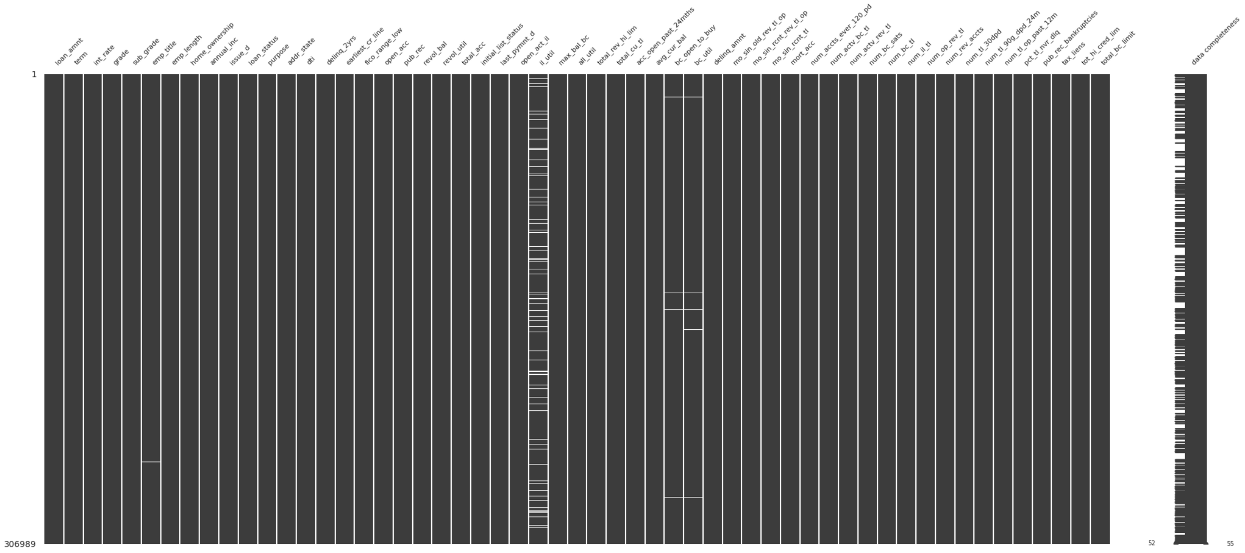
**8. Correlation Analysis (see appendix no.8 ):**
Correlation heatmaps unveil clusters of highly correlated features, such as loan amount, annual income, and total debt-to-income ratio. These insights inform feature selection and model construction for predicting loan outcomes, enhancing risk assessment and lending practices.

# Appendices

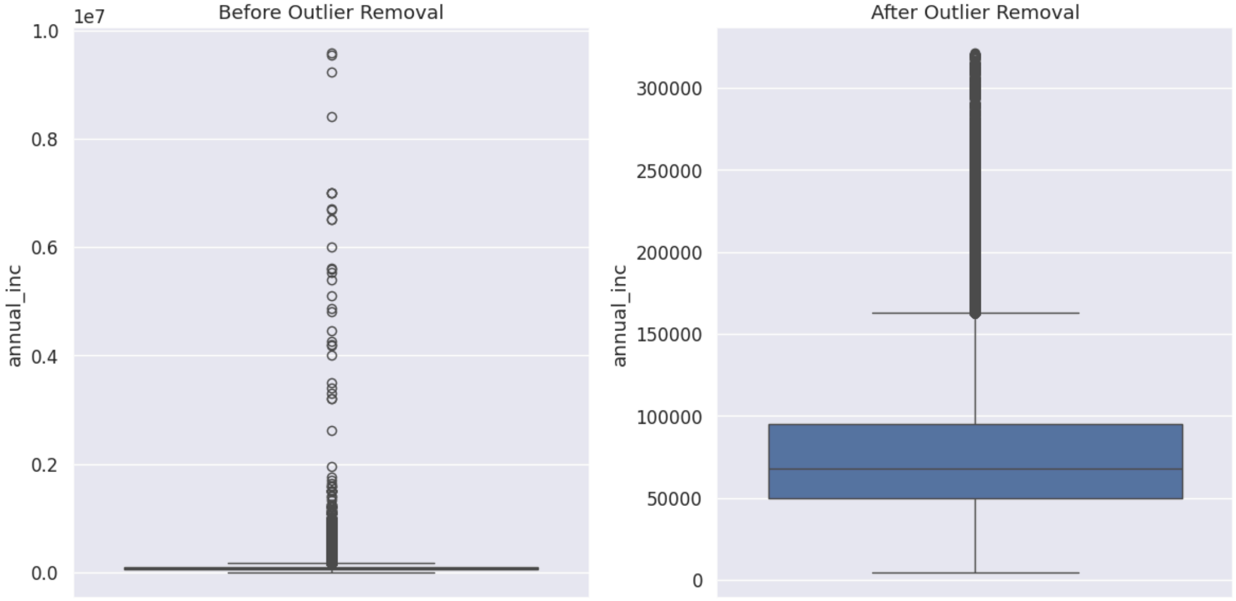
## **Appendix Ⅰ** NAN-Matrix Before (white spaces represent Nans)



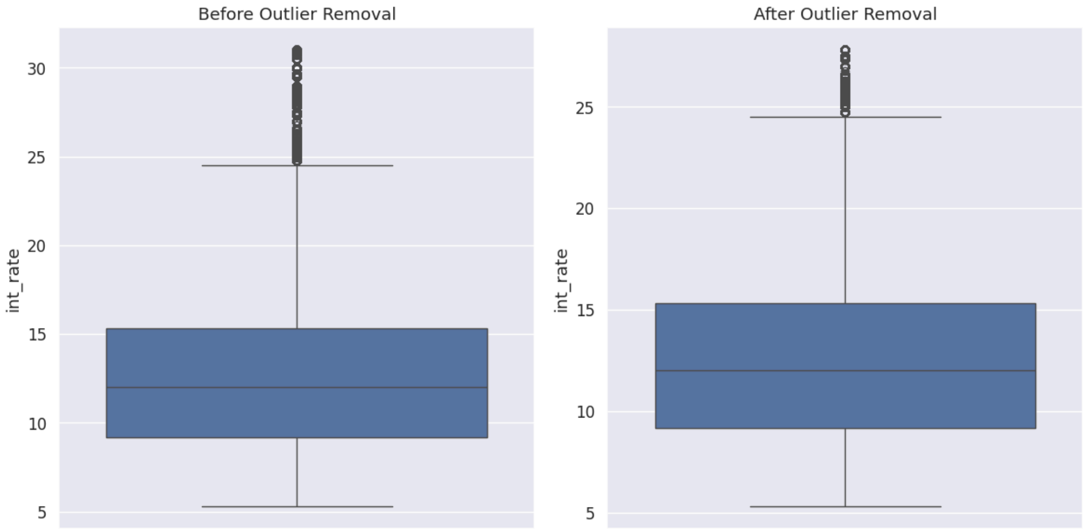
## **Appendix Ⅱ** NAN-Matrix After ETL (white spaces represent Nans)



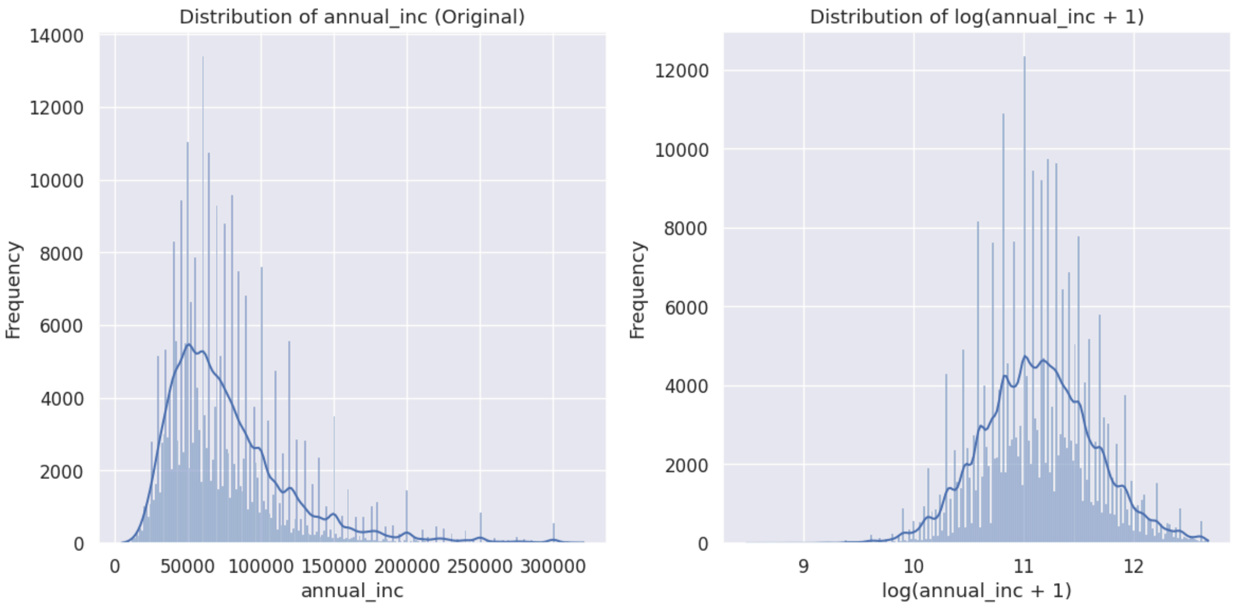
## Appendix Ⅲ Outliers anual_inc



## Outliers int_rate

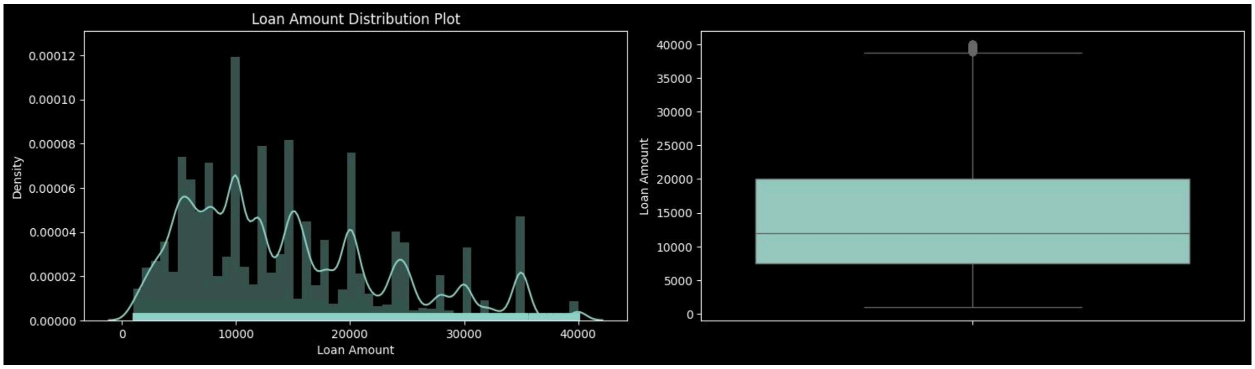


## Appendix Ⅳ Log Transformation

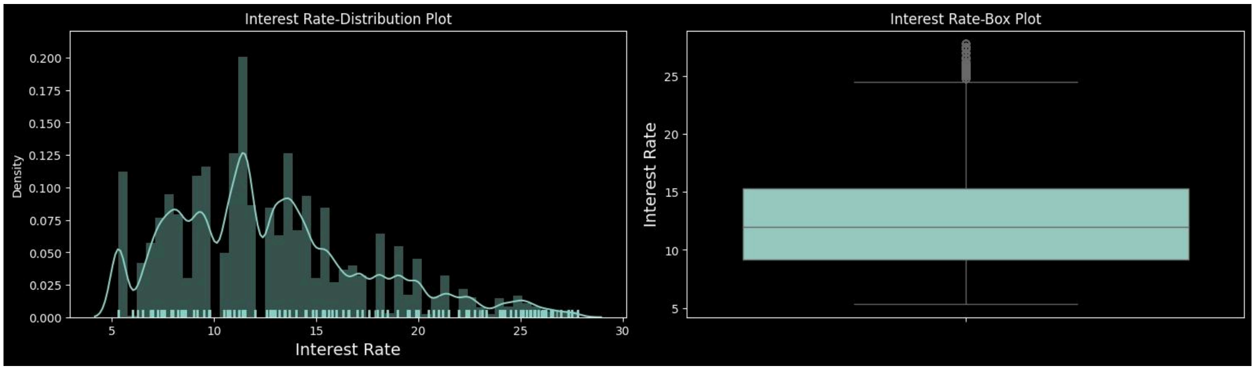


## EDA:

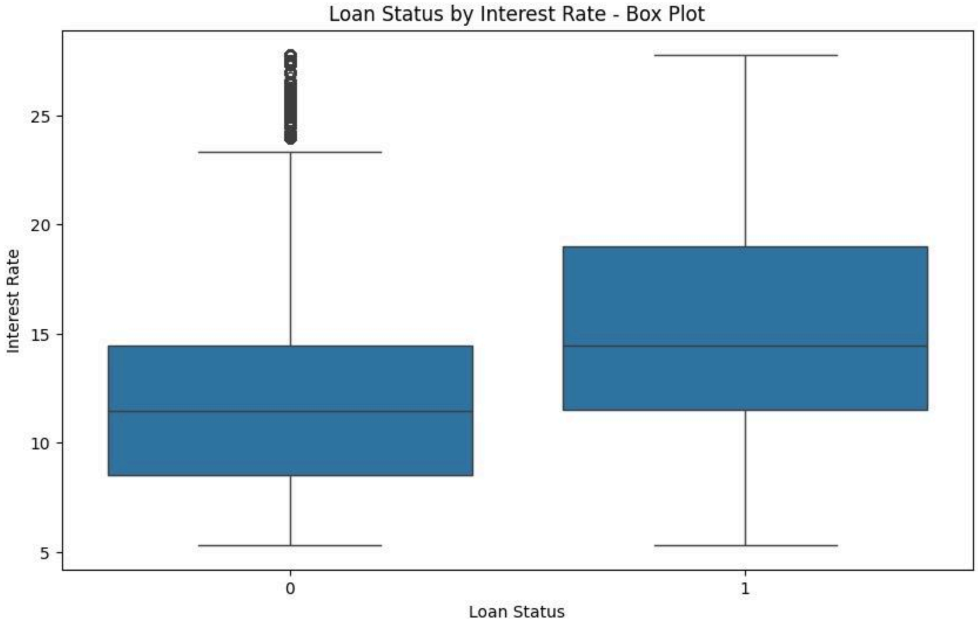
## Appendix no.1:

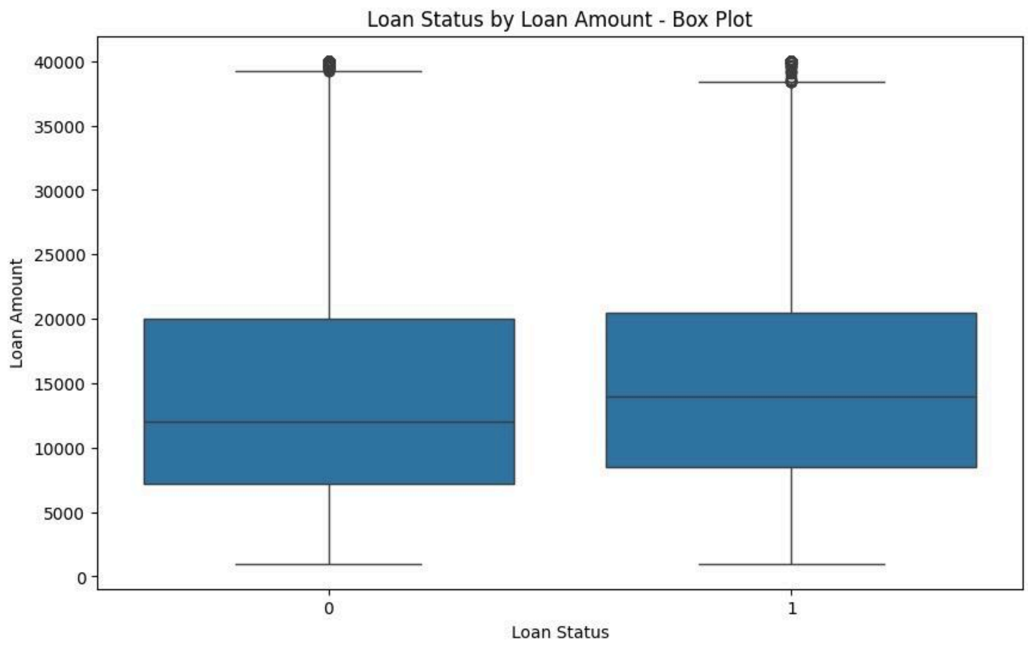


## Appendix no.2:

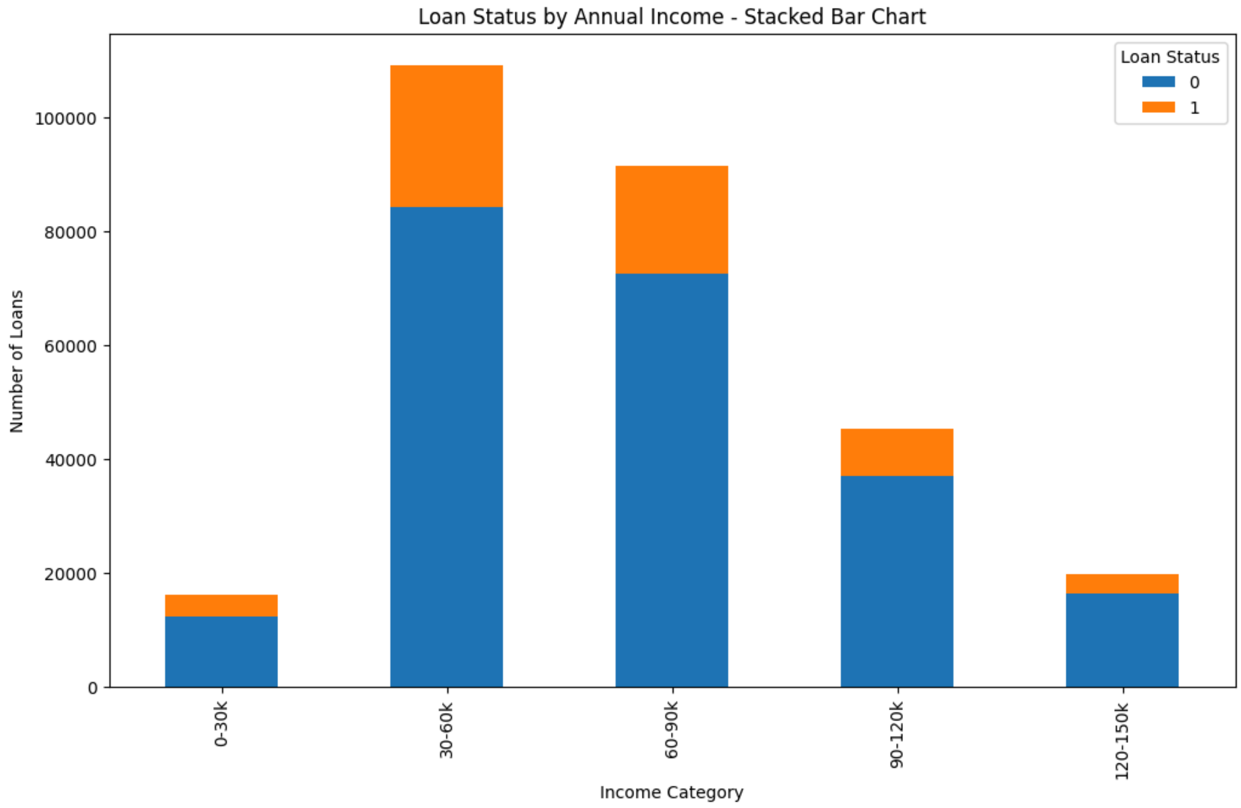


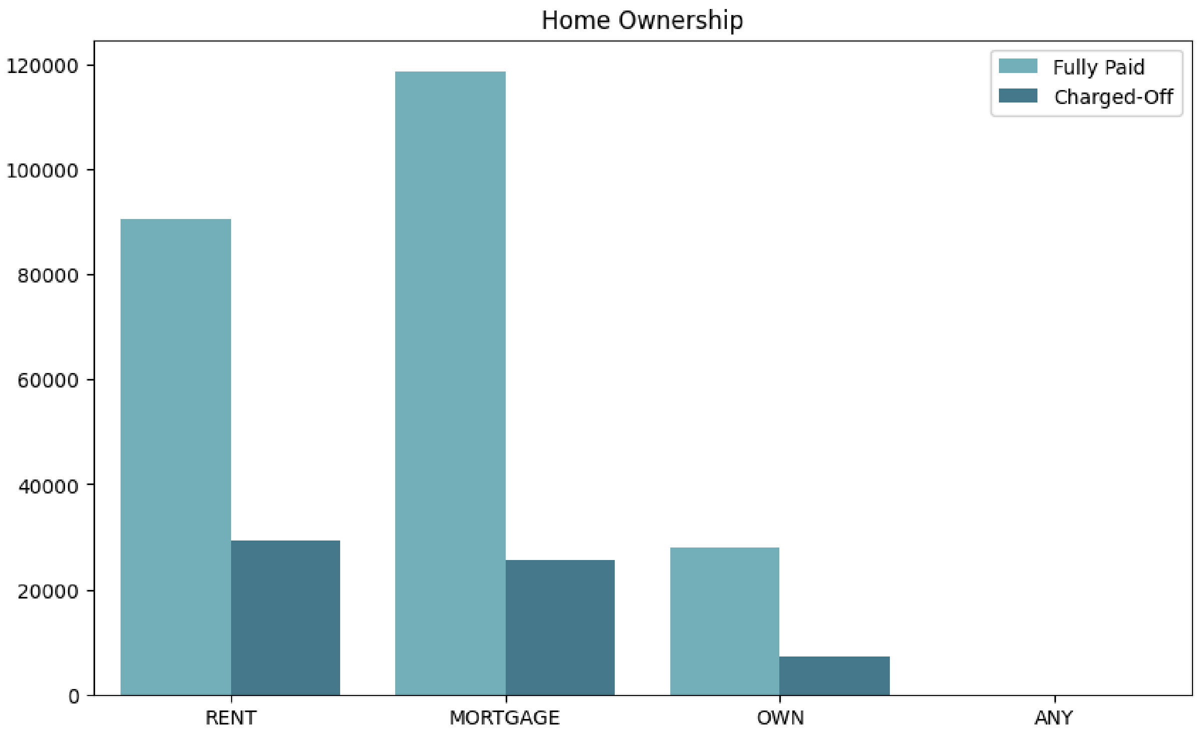
## Appendix No.3:

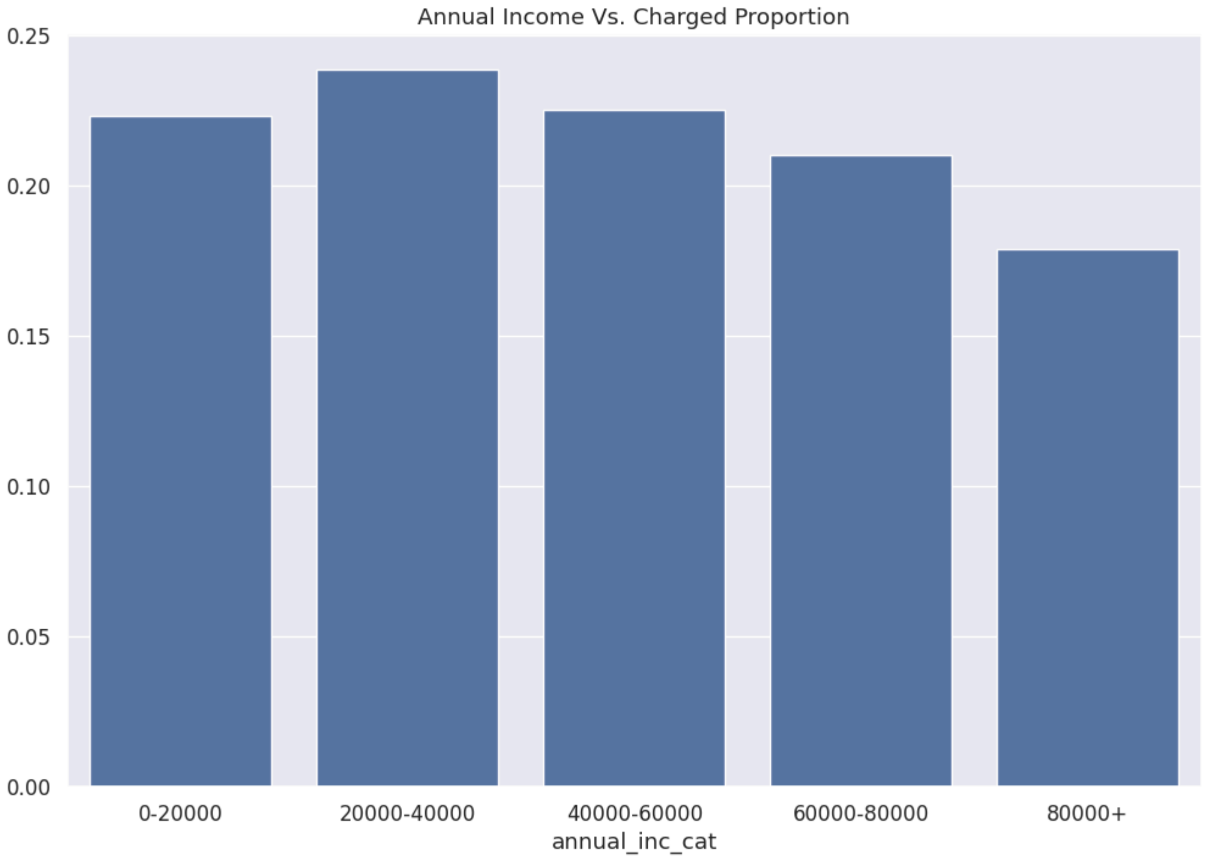


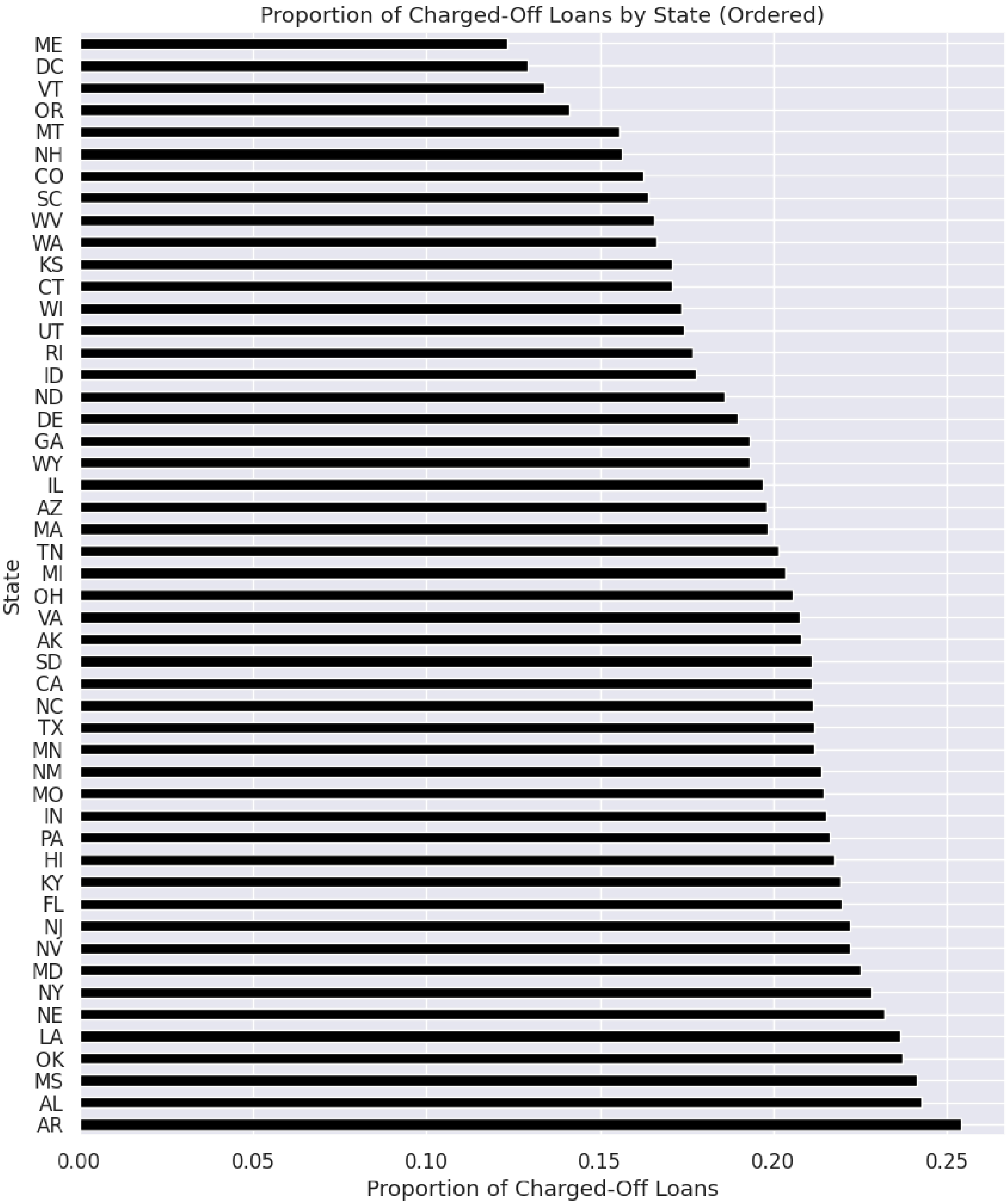


## Appendix no.4:

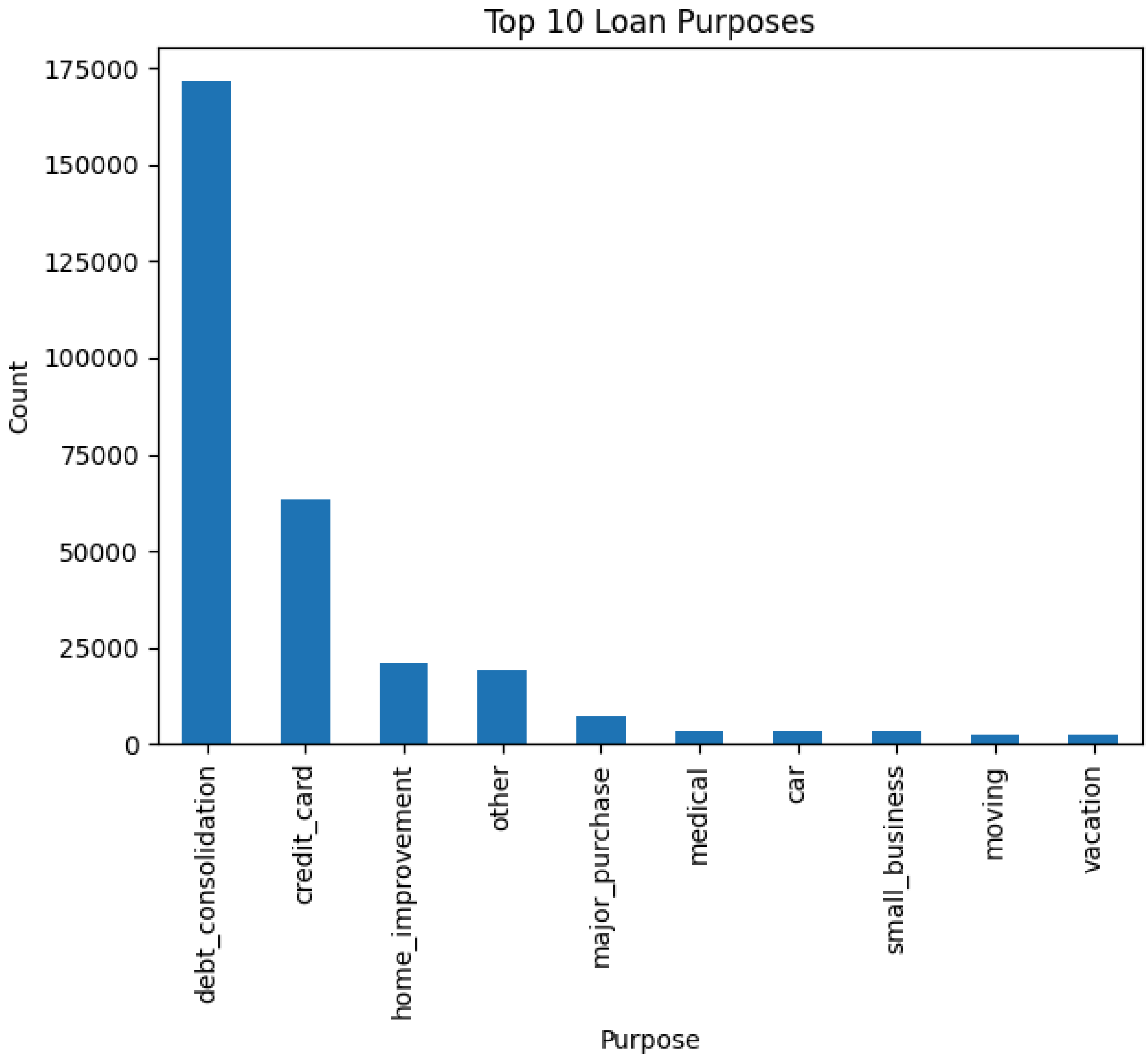


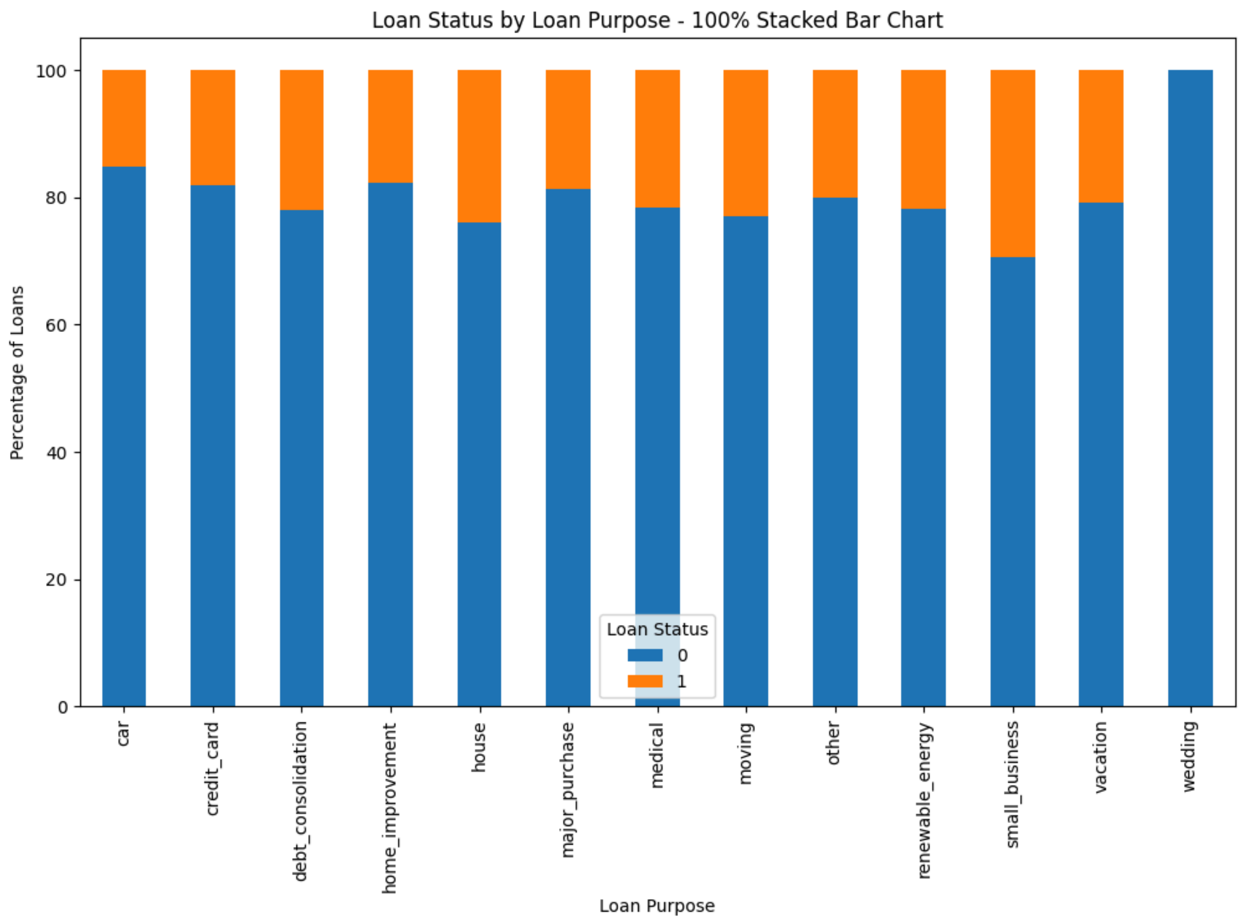

Home Ownership
Fully Paid
Charged-Off
120000
100000
80000
60000
40000
20000
0
RENT
MORTGAGE
OWN
ANY

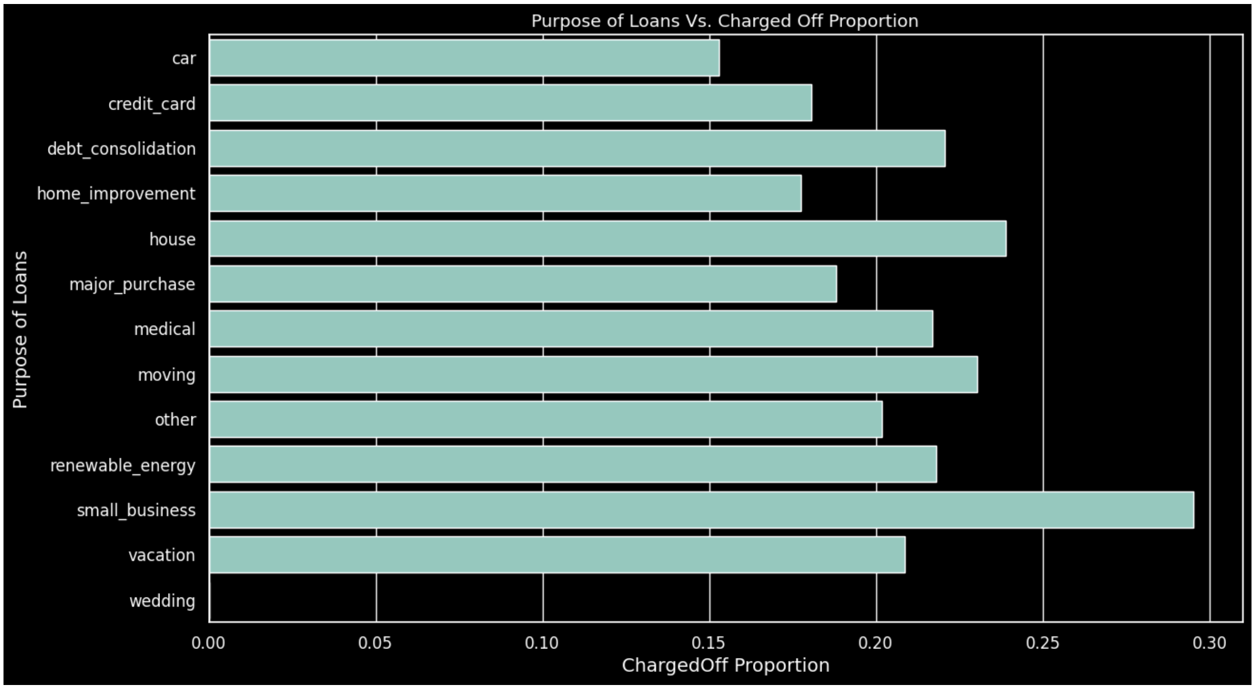

Annual Income Vs. Charged Proportion
0.25
0.20
0.15
0.10
0.05
0.00
0-20000
20000-40000
40000-60000
60000-80000
80000+
annual_inc_cat


Proportion of Charged-Off Loans by State (Ordered)
ME
DC
VT
OR
MT
NH
CO
SC
WV
WA
KS
CT
WI
UT
RI
ID
ND
DE
GA
WY
IL
AZ
MA
TN
MI
OH
VA
AK
SD
CA
NC
TX
MN
NM
MO
IN
PA
HI
KY
FL
NJ
NV
MD
NY
NE
LA
OK
MS
AL
AR
State
0.00
0.05
0.10
0.15
0.20
0.25
Proportion of Charged-Off Loans

**Appendix no.5:**

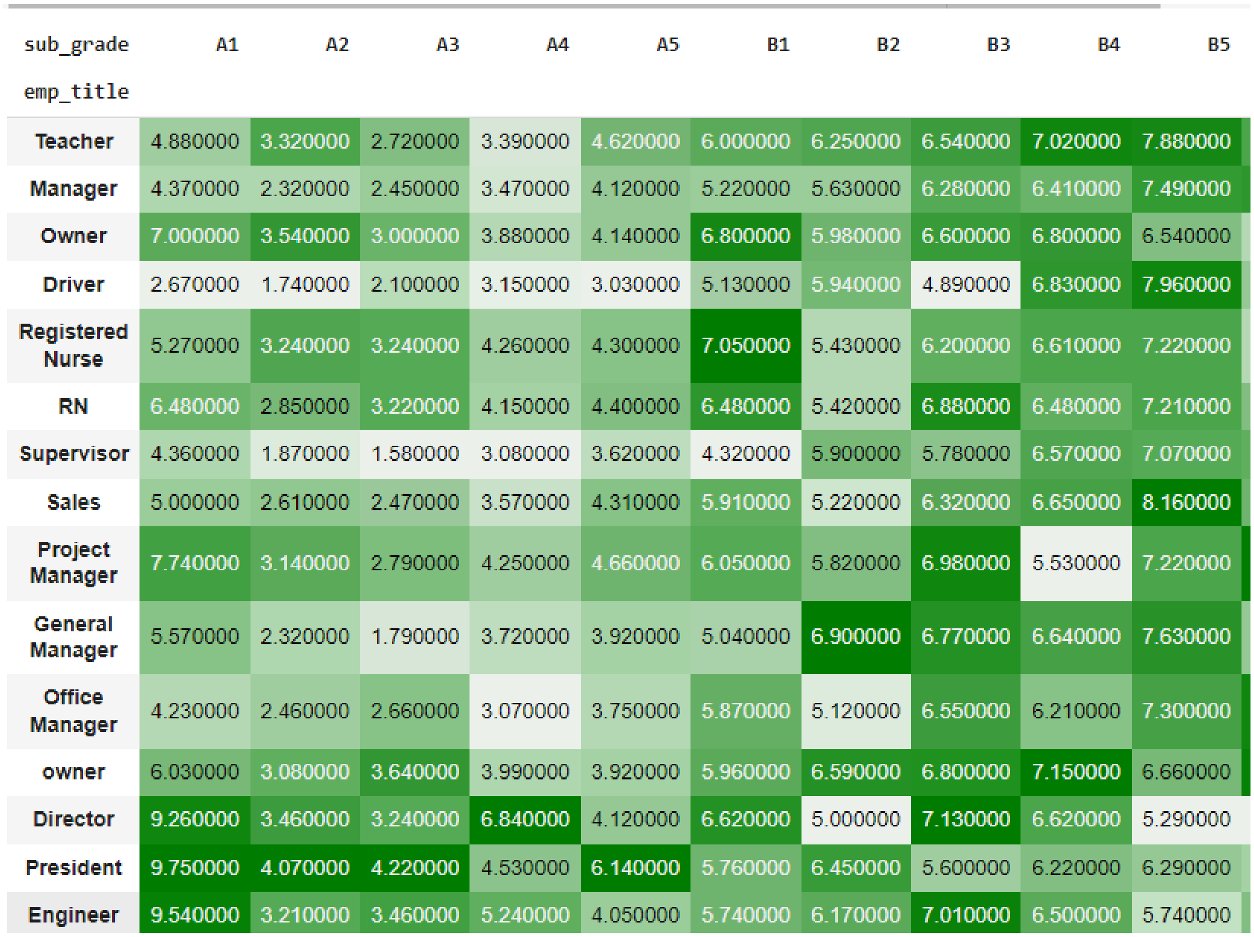



Loan Status by Loan Purpose - 100% Stacked Bar Chart
Percentage of Loans
0
20
40
60
80
100
Loan Status
0
1
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding
Loan Purpose


Purpose of Loans Vs. Charged Off Proportion
Purpose of Loans
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding
0.00
0.05
0.10
0.15
0.20
0.25
0.30
ChargedOff Proportion

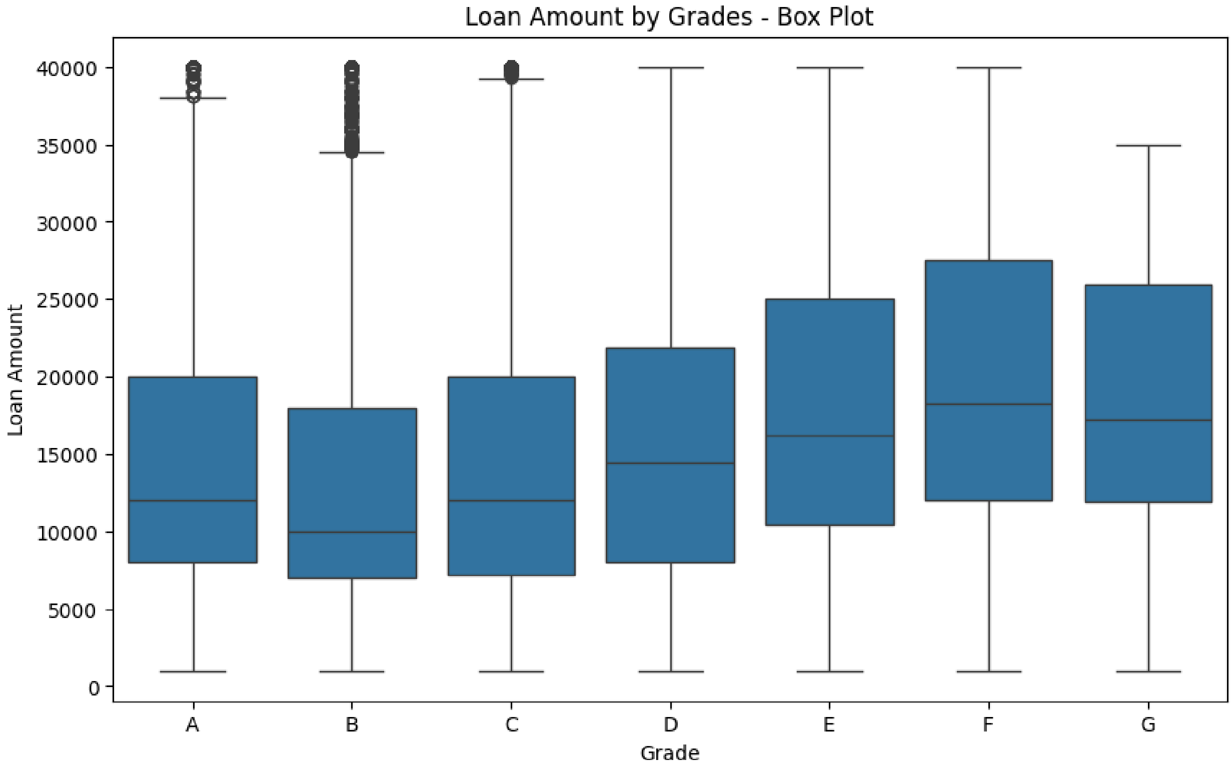
## Appendix no.6:

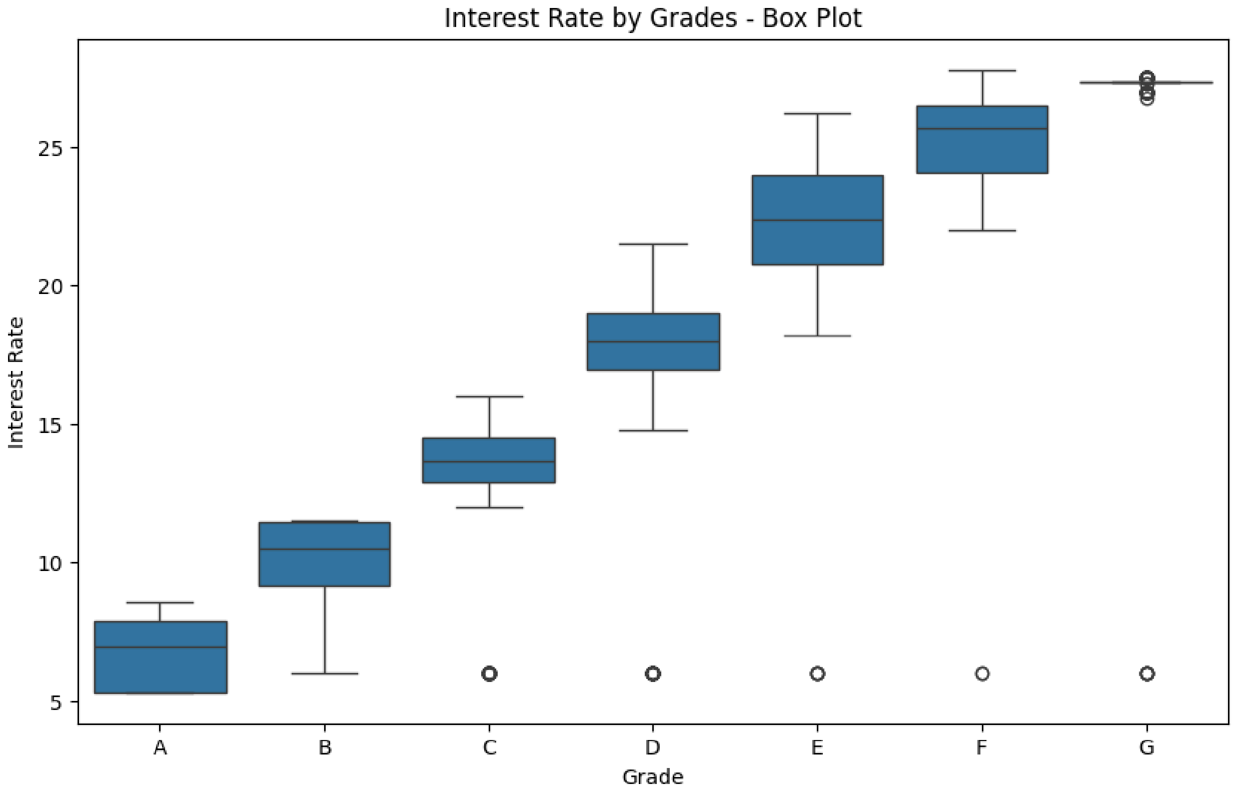
****note that the full crosstab is in the code file.**

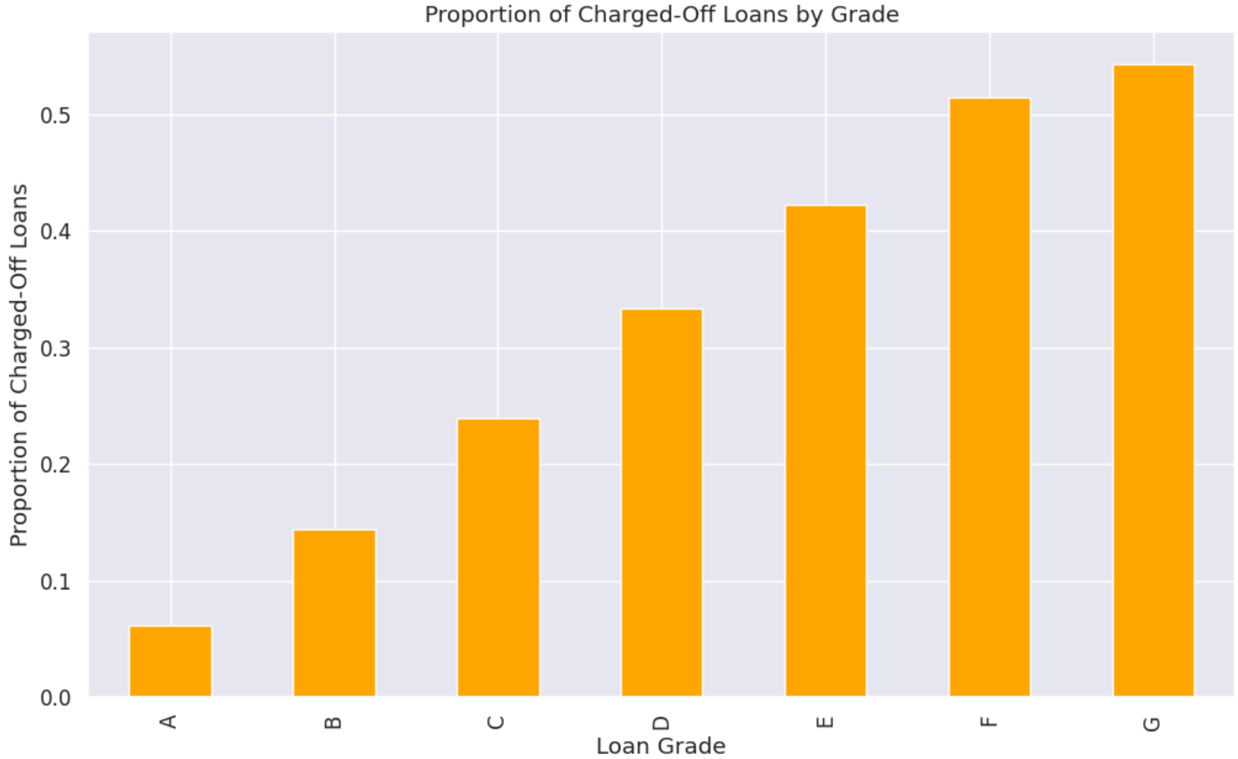


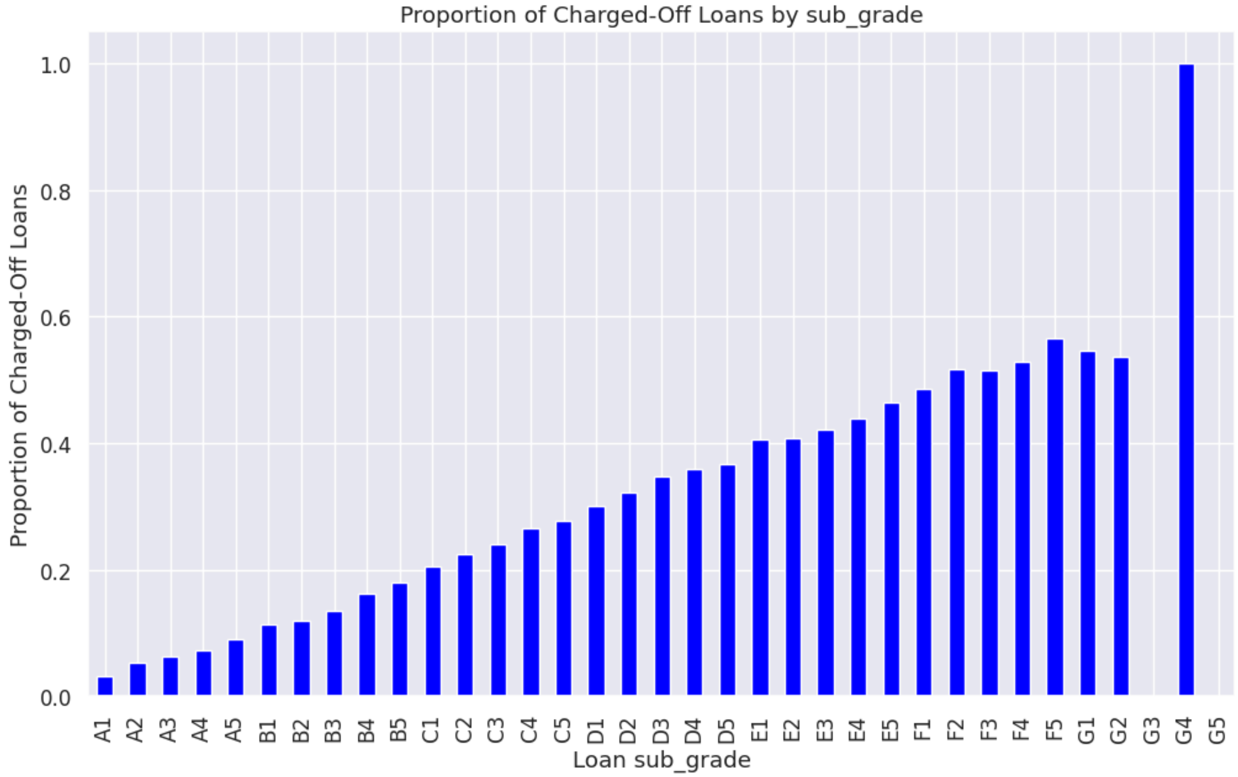
| sub_grade | A1 | A2 | A3 | A4 | A5 | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|---|---|---|---|---|
| emp_title | | | | | | | | | | |
| **Teacher** | 4.880000 | 3.320000 | 2.720000 | 3.390000 | 4.620000 | 6.000000 | 6.250000 | 6.540000 | 7.020000 | 7.880000 |
| **Manager** | 4.370000 | 2.320000 | 2.450000 | 3.470000 | 4.120000 | 5.220000 | 5.630000 | 6.280000 | 6.410000 | 7.490000 |
| **Owner** | 7.000000 | 3.540000 | 3.000000 | 3.880000 | 4.140000 | 6.800000 | 5.980000 | 6.600000 | 6.800000 | 6.540000 |
| **Driver** | 2.670000 | 1.740000 | 2.100000 | 3.150000 | 3.030000 | 5.130000 | 5.940000 | 4.890000 | 6.830000 | 7.960000 |
| **Registered Nurse** | 5.270000 | 3.240000 | 3.240000 | 4.260000 | 4.300000 | 7.050000 | 5.430000 | 6.200000 | 6.610000 | 7.220000 |
| **RN** | 6.480000 | 2.850000 | 3.220000 | 4.150000 | 4.400000 | 6.480000 | 5.420000 | 6.880000 | 6.480000 | 7.210000 |
| **Supervisor** | 4.360000 | 1.870000 | 1.580000 | 3.080000 | 3.620000 | 4.320000 | 5.900000 | 5.780000 | 6.570000 | 7.070000 |
| **Sales** | 5.000000 | 2.610000 | 2.470000 | 3.570000 | 4.310000 | 5.910000 | 5.220000 | 6.320000 | 6.650000 | 8.160000 |
| **Project Manager** | 7.740000 | 3.140000 | 2.790000 | 4.250000 | 4.660000 | 6.050000 | 5.820000 | 6.980000 | 5.530000 | 7.220000 |
| **General Manager** | 5.570000 | 2.320000 | 1.790000 | 3.720000 | 3.920000 | 5.040000 | 6.900000 | 6.770000 | 6.640000 | 7.630000 |
| **Office Manager** | 4.230000 | 2.460000 | 2.660000 | 3.070000 | 3.750000 | 5.870000 | 5.120000 | 6.550000 | 6.210000 | 7.300000 |
| **owner** | 6.030000 | 3.080000 | 3.640000 | 3.990000 | 3.920000 | 5.960000 | 6.590000 | 6.800000 | 7.150000 | 6.660000 |
| **Director** | 9.260000 | 3.460000 | 3.240000 | 6.840000 | 4.120000 | 6.620000 | 5.000000 | 7.130000 | 6.620000 | 5.290000 |
| **President** | 9.750000 | 4.070000 | 4.220000 | 4.530000 | 6.140000 | 5.760000 | 6.450000 | 5.600000 | 6.220000 | 6.290000 |
| **Engineer** | 9.540000 | 3.210000 | 3.460000 | 5.240000 | 4.050000 | 5.740000 | 6.170000 | 7.010000 | 6.500000 | 5.740000 |

**Appendix no.7:**

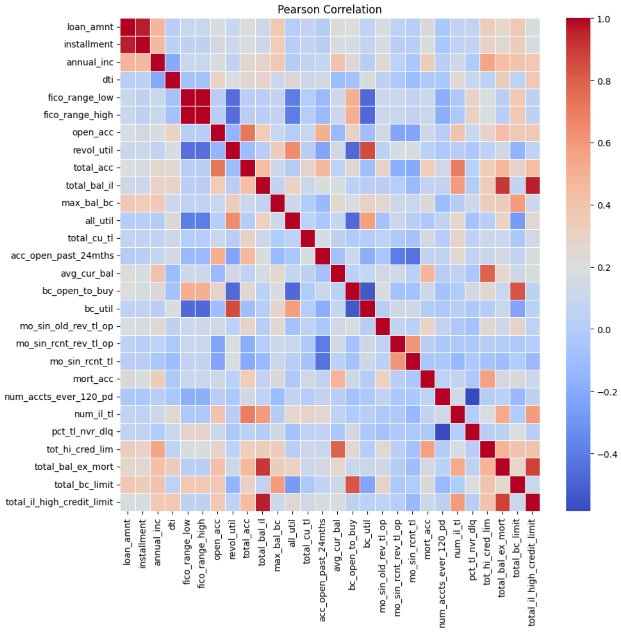


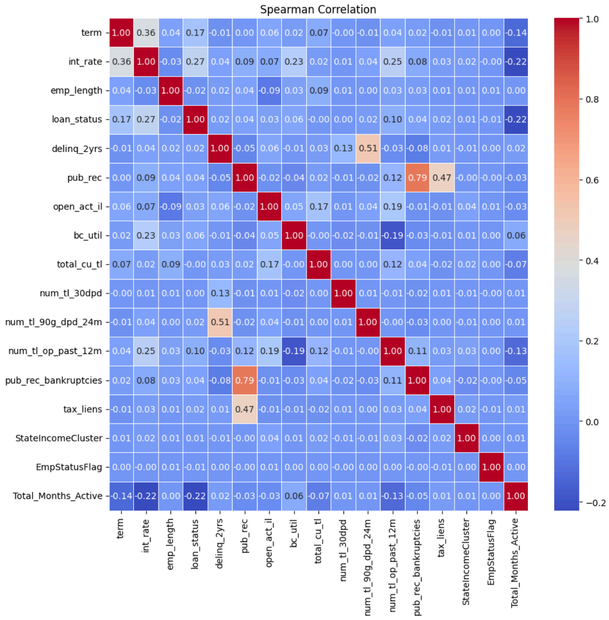


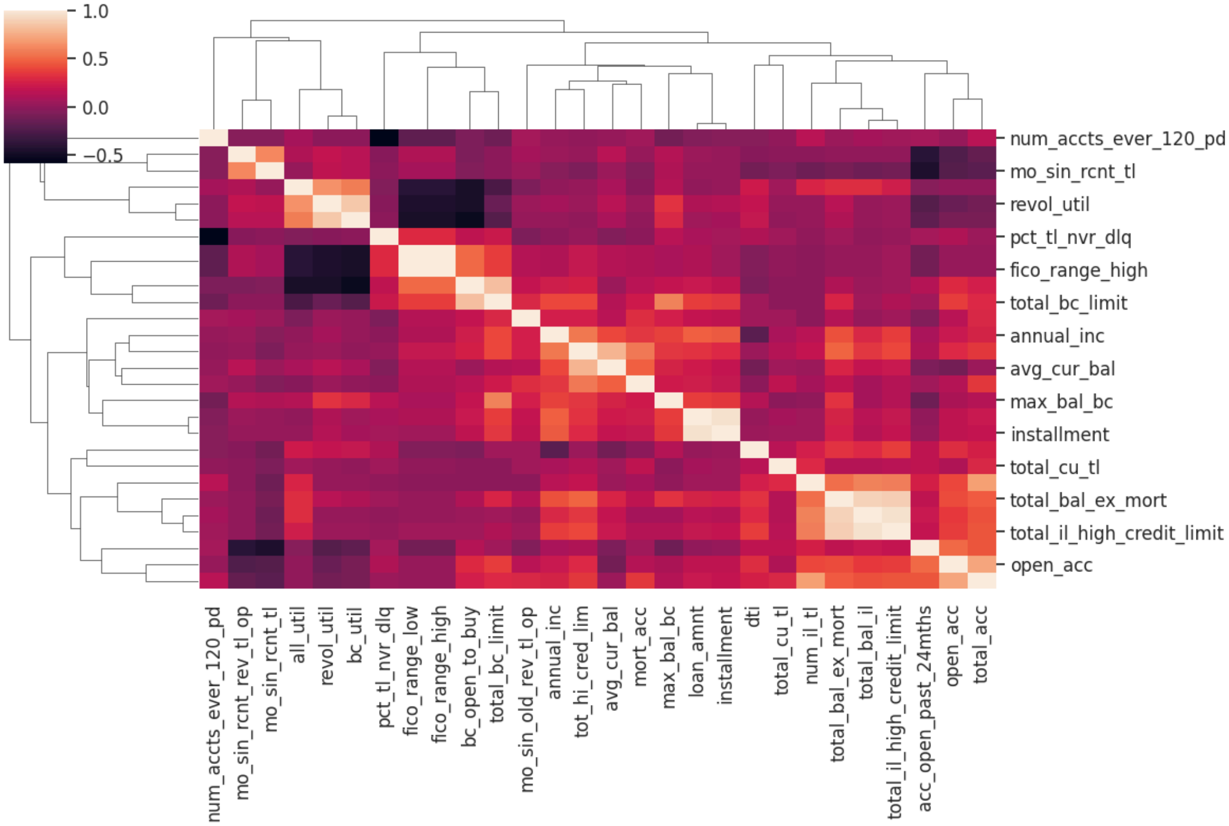

Proportion of Charged-Off Loans by Grade
Proportion of Charged-Off Loans
0.0
0.1
0.2
0.3
0.4
0.5
A
B
C
D
E
F
G
Loan Grade


Proportion of Charged-Off Loans by sub_grade
Proportion of Charged-Off Loans
0.0
0.2
0.4
0.6
0.8
1.0
A1
A2
A3
A4
A5
B1
B2
B3
B4
B5
C1
C2
C3
C4
C5
D1
D2
D3
D4
D5
E1
E2
E3
E4
E5
F1
F2
F3
F4
F5
G1
G2
G3
G4
G5
Loan sub_grade

# Appendix no.8:







## Appendix no.9: Road Map Details

| Task | Status | Notes | Time Estimation |
|---|---|---|---|
| Business understanding | Completed |  | 1-2 weeks |
| Data merge | Completed |  | 2-4 weeks |
| Handling NA | In progress | Decide the specific method for each features with missing data | 2-4 weeks |
| Handling outliers | In progress | Implement our function to other cases | 2 weeks |
| Leakage Data | In progress | Revise our feature selection | 2 weeks |
| EDA | In progress | We need to revise the correlation once the target value is created | 2 weeks |
| Target Variable selection | In progress |  | 2-4 weeks |
| Addressing High Cardinality and Skewness | In progress |  | 4-6 weeks |
| Data normalization | In progress | Decide if we want to implement log-trans or other methods | - |
| Decide what to do with il_util | In progress |  | - |
| One hot implementation (optional) | Not started |  | - |

| Task | Status | Notes | Time Estimation |
|---|---|---|---|
| Go over instances with 0 in 'months active' | In progress | There are some features to be address with values such as ‘other’ or ‘not known’ | - |
| Address Low Freq values | Not started | | - |
| Calculate expected return | Not started | | 4-6 weeks |
| Model selection | Not started | | 6-8 weeks |
| Evaluation of the chosen model | Not started | | 8 weeks |
| Deployment | Not started | | 8 weeks |
| Risk assessment | Not started | | 9 weeks |
| Recommendation for Walter | Not started | | |