COMSW4995 Applied Machine Learning

Project Deliverable #2-Data Analysis and Visualization

Team Members: Meilin Guo (mg4578), Chiara Wollner (crw2160), Angela Mu (aym2122), Hussain Doriwala (hd2551), Nagavasavi Jeepalyam (nj2506)

Goal: This project aims to use credit-related information and other banking details to predict and classify an individual's credit score bracket.

Initial data exploration

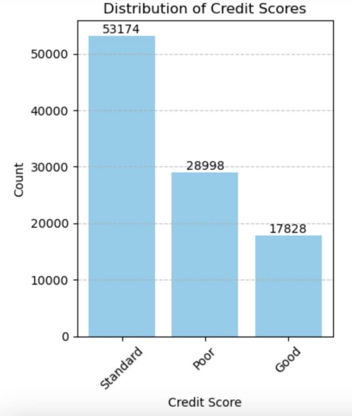
Target: [‘Credit Score’]: **53% Standard; 29% Poor; 18% Good;**

Categorical Features:

[‘Name’,‘Occupation’,’SSN’,’Month’,’Type_of_Loan’,’Credit_Mix’,’Payment_of_Min_Amount’, ’Payment_Behavior’]

Numerical Features:

['Age', 'Annual_Income', 'Monthly_Inhand_Salary', 'Num_Bank_Accounts', 'Num_Credit_Card', 'Interest_Rate', 'Num_of_Loan', 'Delay_from_due_date', 'Num_of_Delayed_Payment', 'Changed_Credit_Limit', 'Num_Credit_Inquiries', 'Outstanding_Debt', 'Credit_Utilization_Ratio', 'Total_EMI_per_month', 'Amount_invested_monthly', 'Monthly_Balance’]



	Monthly_Inhand_Salary	Num_Bank_Accounts	Num_Credit_Card	Interest_Rate	Delay_from_due_date	Num_Credit_Inquiries	Credit_Utilization_Ratio	Total_EMI_per_month
count	84998.000000	100000.000000	100000.00000	100000.000000	100000.000000	98035.000000	100000.000000	100000.000000
mean	4194.170850	17.091280	22.47443	72.466040	21.068780	27.754251	32.285173	1403.118217
std	3183.686167	117.404834	129.05741	466.422621	14.860104	193.177339	5.116875	8306.041270
min	303.645417	-1.000000	0.00000	1.000000	-5.000000	0.000000	20.000000	0.000000
25%	1625.568229	3.000000	4.00000	8.000000	10.000000	3.000000	28.052567	30.306660
50%	3093.745000	6.000000	5.00000	13.000000	18.000000	6.000000	32.305784	69.249473
75%	5957.448333	7.000000	7.00000	20.000000	28.000000	9.000000	36.496663	161.224249
max	15204.633333	1798.000000	1499.00000	5797.000000	67.000000	2597.000000	50.000000	82331.000000

Initial data exploration

Check null values - Before

Customer_ID	0
Month	0
Age	2781
Occupation	7062
Annual_Income	2783
Monthly_Inhand_Salary	15002
Num_Bank_Accounts	1315
Num_Credit_Card	2271
Interest_Rate	2034
Num_of_Loan	4348
Type_of_Loan	11408
Delay_from_due_date	4002
Num_of_Delayed_Payment	7002
Changed_Credit_Limit	2091
Num_Credit_Inquiries	1965
Credit_Mix	20195
Outstanding_Debt	5272
Credit_Utilization_Ratio	4
Credit_History_Age	9030
Payment_of_Min_Amount	0
Total_EMI_per_month	6795
Amount_invested_monthly	4479
Payment_Behaviour	7600
Monthly_Balance	2868
Credit_Score	0
dtype: int64	

Missing data

Column types - Before

ID	object
Customer_ID	object
Month	object
Name	object
Age	object
SSN	object
Occupation	object
Annual_Income	object
Monthly_Inhand_Salary	float64
Num_Bank_Accounts	int64
Num_Credit_Card	int64
Interest_Rate	int64
Num_of_Loan	object
Type_of_Loan	object
Delay_from_due_date	int64
Num_of_Delayed_Payment	object
Changed_Credit_Limit	object
Num_Credit_Inquiries	float64
Credit_Mix	object
Outstanding_Debt	object
Credit_Utilization_Ratio	float64
Credit_History_Age	object
Payment_of_Min_Amount	object
Total_EMI_per_month	float64
Amount_invested_monthly	object
Payment_Behaviour	object
Monthly_Balance	object
Credit_Score	object
dtype: object	

Incorrect data types

Numerical features like Age should be int64or float64

Column types - After Cleaning

Age	float64
Occupation	object
Annual_Income	float64
Monthly_Inhand_Salary	float64
Num_Bank_Accounts	int64
Num_Credit_Card	int64
Interest_Rate	int64
Num_of_Loan	float64
Type_of_Loan	object
Delay_from_due_date	int64
Num_of_Delayed_Payment	float64
Changed_Credit_Limit	float64
Num_Credit_Inquiries	float64
Credit_Mix	object
Outstanding_Debt	float64
Credit_Utilization_Ratio	float64
Credit_History_Age	object
Payment_of_Min_Amount	object
Total_EMI_per_month	float64
Amount_invested_monthly	float64
Payment_Behaviour	object
Monthly_Balance	float64
Credit_Score	object
dtype: object	

Cleaning and sampling

Data Cleaning:

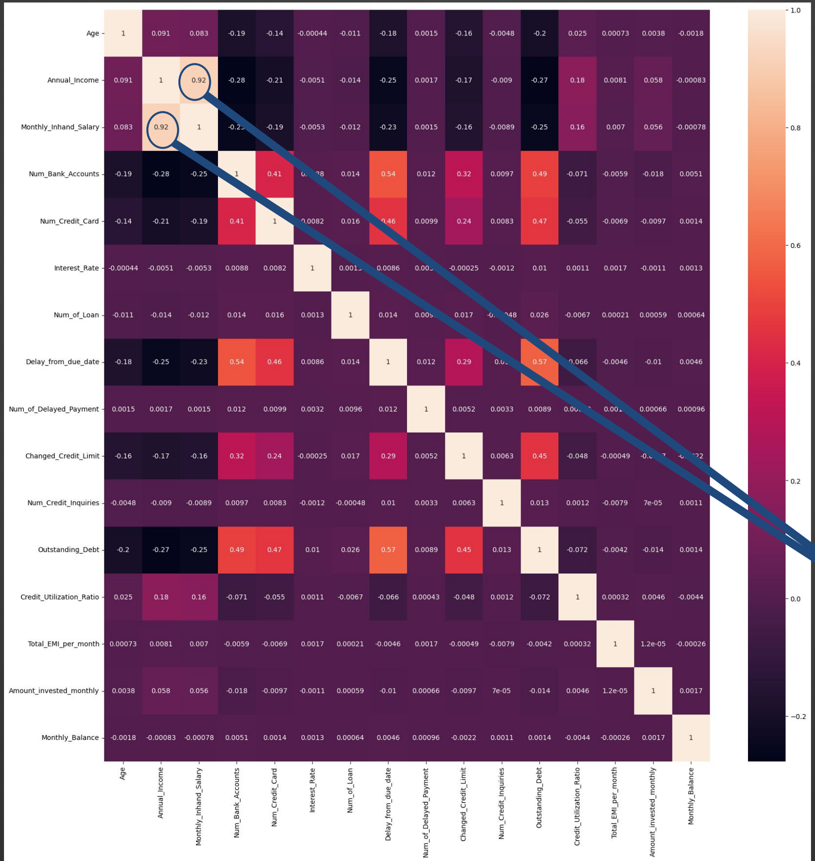
1. Drop irrelevant columns: ID, name, SSN
2. Drop highly correlated numerical columns: Monthly_Inhand_Salary (see next page)
3. Fix poorly-formatted data: numbers w/ underscores, underscores indicating missing data, nonsensical categorical values, etc.
4. Data type: ensure that numerical values are converted to float64 format
5. Fill in negative and missing values w/ the mean of records w/ the same Customer_ID (for numerical) or w/ the mode (for categorical)
6. Parse string version of date and month into numerical format
7. Encode categorical columns: Apply ordinal coding to Month, Credit_Mix, target encoding to Payment_Behaviour, Occupation, Type_of_Loan
8. Scale numerical columns: apply StandardScaler to numerical columns

Sampling:

train/val/test sets splitting: stratified sampling w/ 60, 20, 20 split

Shape of train/val/test sets :((55655, 21), (18553, 21),(18553, 21))

Correlation matrix

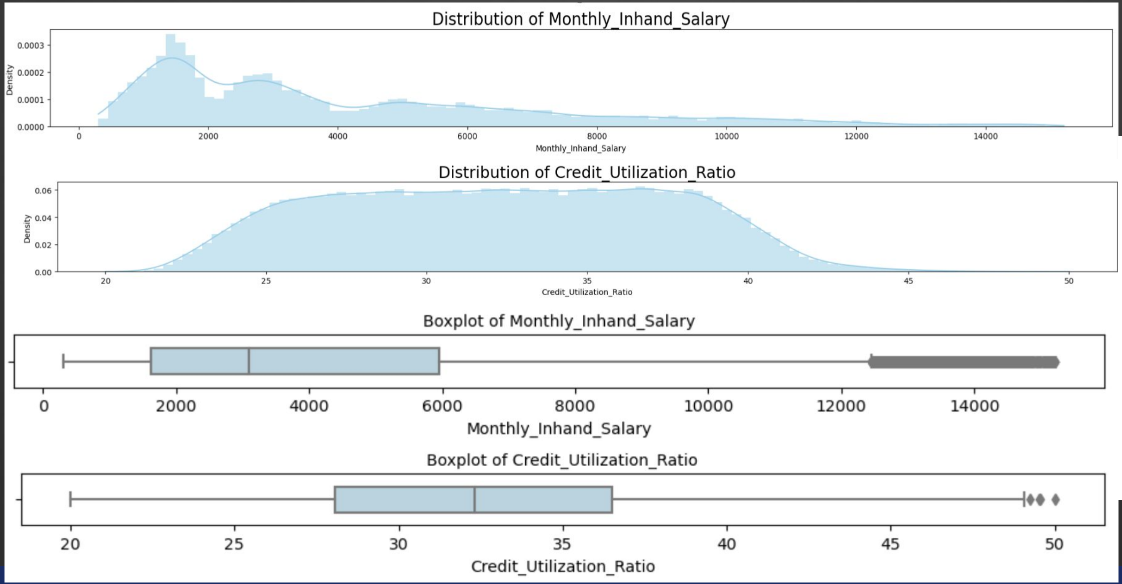


Find and remove redundant columns by identifying features with high correlations (close to 1 or -1).

Drop **Annual_Income** or **Monthly_Inhand_Salary**

Distribution of Numerical Variables

- We generated kernel density plots and boxplots for all numerical variables. The 'Monthly_inhand_salary' and 'Credit_utilization_ratio' are featured below as examples.
- We analyzed means and outlier counts for each numerical variable.
- **Insights:** The distribution of age, annual income, and monthly in-hand salary is right-skewed, suggesting a young population and a general trend towards lower earnings. There is a high average number of bank accounts (17) and credit cards (22) per individual, along with an unusually high mean interest rate of 72.47%, indicating potential outliers or data inaccuracies.

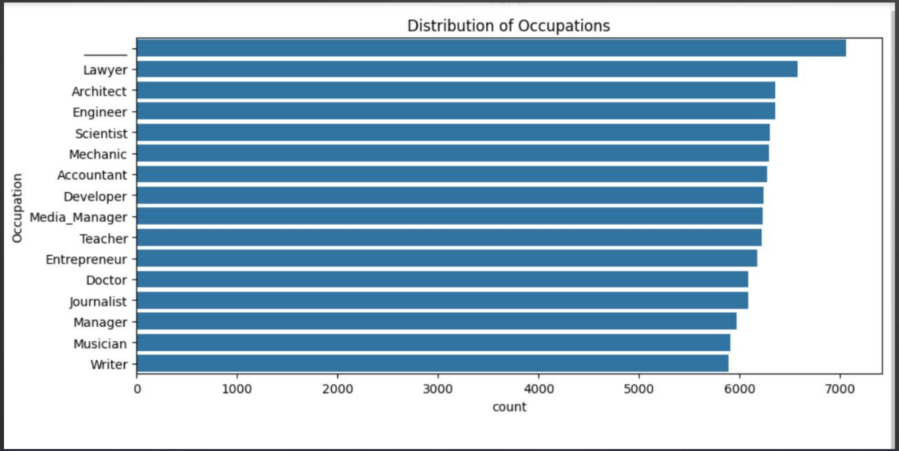


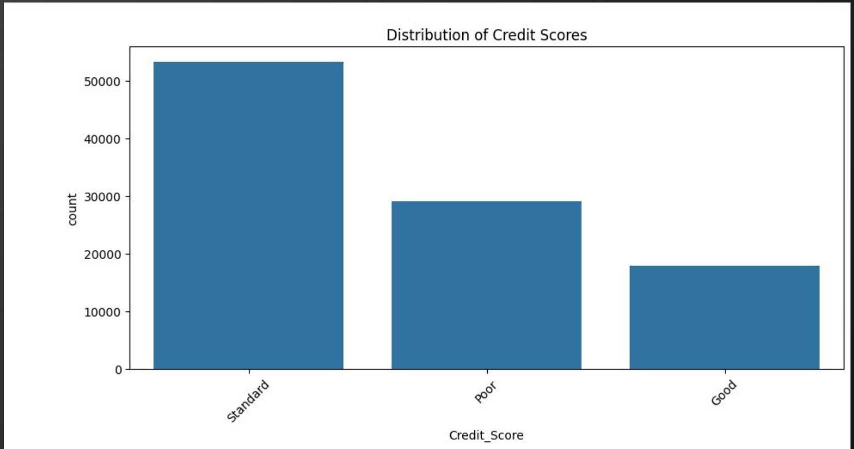

```
Means of the numerical columns:
----------------------------------------
Age: 110.6497
Annual_Income: 176415.70129814997
Monthly_Inhand_Salary: 4194.170849600523
Num_Bank_Accounts: 17.09128
Num_Credit_Card: 22.47443
Interest_Rate: 72.46604
Num_of_Loan: 3.00996
Num_of_Delayed_Payment: 30.923342437471774
Changed_Credit_Limit: 10.389025115157953
Num_Credit_Inquiries: 27.75425103279441
Outstanding_Debt: 1426.220376
Credit_Utilization_Ratio: 32.2851725189436
Credit_History_Age: 221.19540507859733
Total_EMI_per_month: 1403.1182166159933
Amount_invested_monthly: 637.4129984078688
Monthly_Balance: -3.0364372469635625e+22
```

```
Number of outliers in each column:
----------------------------------------
Age: 2781
Annual_Income: 2783
Monthly_Inhand_Salary: 1683
Num_Bank_Accounts: 1315
Num_Credit_Card: 2271
Interest_Rate: 2034
Num_of_Loan: 4348
Num_of_Delayed_Payment: 736
Changed_Credit_Limit: 668
Num_Credit_Inquiries: 1650
Outstanding_Debt: 5272
Credit_Utilization_Ratio: 4
Credit_History_Age: 0
Total_EMI_per_month: 6795
Amount_invested_monthly: 10096
Monthly_Balance: 7636
```

Distribution of categorical variables

- **Month data is evenly spread across the first five months,** showing no preferential or majority entry.
- In **Occupation, "Lawyer" emerges as the predominant profession.**
- **The Credit Mix category reveals "Standard" as the most common type among customers**, overshadowing other classifications and highlighting a potential area for credit improvement.
- In **Payment of Min Amount, the majority of the dataset indicates "Yes,"** suggesting that most customers tend to make their minimum payments.
- **Payment Behaviour is most frequently characterized by "Low spent Small value payments,"** illustrating a cautious spending pattern among a significant portion of customers.

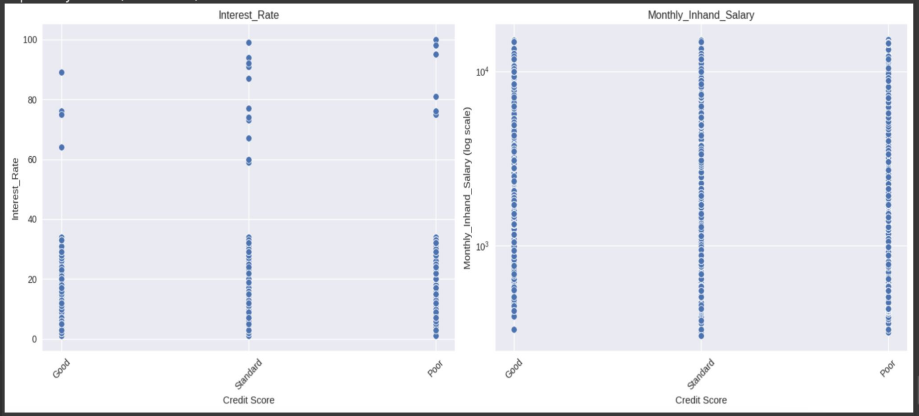




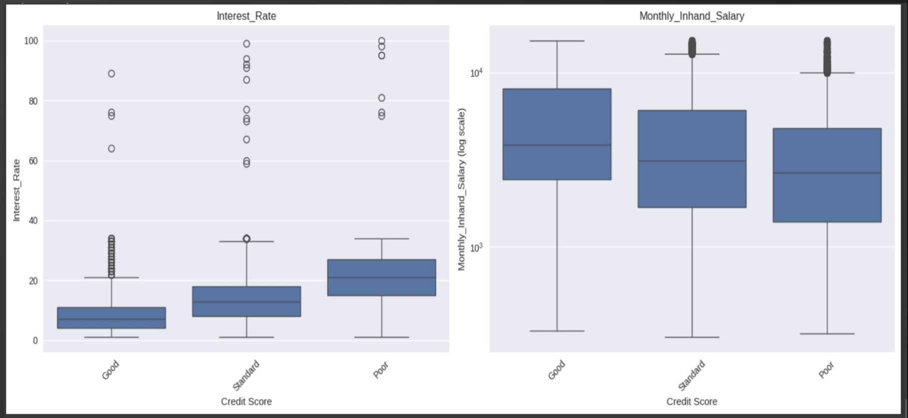
Features and the Target variable

We've chosen 'Interest Rate' and 'Monthly Inhand Salary' to illustrate our analysis in this presentation, though similar analyses were conducted on all features. Borrowers with higher credit scores tend to receive lower interest rates, as reflected in the median scores on the boxplots and the negative correlation in the scatter plot. While the impact of salary is less clear due to the log scale, there is also a suggestion that borrowers with higher incomes may qualify for slightly better rates.

Scatter Plots



Box Plots



Additional Insights from Data Exploration

- Each customer has 8 records in different months, showing how their financial information changes over time
- Total # of unique customers = 12500, Total # of records = 100000
- the Credit Score variable shows that a "Standard" rating is the most prevalent, underscoring the need for credit management and improvement opportunities among the dataset's individuals.
- Together, these insights offer a foundational understanding of the financial habits and statuses of the customers represented in the dataset, revealing areas where financial behavior could be enhanced or further investigated.

Machine Learning Techniques

- **Regression: mapping [Standard,Poor,Good] to numbers**
    - Logistic Regression: Encode categorical labels for binary or multinomial outcome prediction.
- **Classification**
    - Support Vector Machines (SVM): Utilize hyperplanes for classification tasks.
    - Decision Trees: Use a tree-like model for decision-making.
    - Random Forest: Implement an ensemble of Decision Trees, usually for improved accuracy.
    - Bootstrapping: Apply resampling techniques to estimate model accuracy.
- **Clustering: drop target**
    - KNN (K-Nearest Neighbors): Classify based on the closest training examples in the feature space.
    - K-means: Partition n observations into k clusters where each observation belongs to the cluster with the nearest mean.
- **Artificial Neural Networks**
    - Feed-Forward Neural Networks