**Description**

# Credit score classification

<https://www.kaggle.com/datasets/parisrohan/credit-score-classification>

**Problem Statement**

You are working as a data scientist in a global finance company. Over the years, the company has collected basic bank details and gathered a lot of credit-related information. The management wants to build an intelligent system to segregate the people into credit score brackets to reduce the manual efforts.

**Task**

Given a person’s credit-related information, build a machine learning model that can classify the credit score.

Possible uses:

With the Credit Score, banks can predict whether a customer could pay by the due date. Then decide whether to give the loan and also the limit to give.

**Attributes**

1. ID : A unique identification of an entry
2. Customer\_ID : A unique identification of a person
3. Month : Represents the month of the year
4. Name : Represent the name of a person.
5. Age : Represents the age of the person
6. SSN : Represents the social security number of a person
7. Occupation : Represents the occupation of the person
8. Annual\_Income : Represents the annual income of the person
9. Monthly\_Inhand\_Salary : Represents the monthly base salary of a person
10. Num\_Bank\_Accounts : Represents the number of bank accounts a person
11. Num\_Credit\_Card : Represents the number of other credit cards held by a person
12. Interest\_Rate : Represents the interest rate on credit card
13. Num\_of\_Loan : Represents the number of loans taken from the bank
14. Type\_of\_Loan : Represents the type of loan taken by a person
15. Delay\_from\_due\_date : Represents the average number of days delayed from the payment date.
16. Num\_of\_Delayed\_Payment : Represents the average number of payments delayed by a person
17. Changed\_Credit\_Limit : Represents the percentage change in credit card limit
18. Num\_Credit\_Inquiries : Represents th number of credit card inquiries
19. Credit\_Mix : Represents the classification of the mix of credits
20. Outstanding\_Debt : Represents the remaining debt to be paid (in USD)
21. Credit\_Utilization\_Ratio : Represents the utilization ratio of credit card
22. Credit\_History\_Age : Represents the age of credit history of the person
23. Payment\_of\_Min\_Amount : Represents whether only the minimum amount was paid by the person
24. Total\_EMI\_per\_month : Represents the monthly EMI payments (in USD)
25. Amount\_invested\_monthly : Represents the monthly amount invested by the customer (in USD)
26. Payment\_Behaviour : Represents the payment behavior of the customer (in USD)
27. Monthly\_Balance : Represents the monthly balance amount of the customer (in USD)
28. Credit\_Score : Represents the bracket of credit score (Poor, Standard, Good)

I am using the credit\_score for classification in this dataset.

Preparation

Modifying:

1. Handling the unuseful column in this dataset.

| Column ID | Removed Column ID | I have removed the ID because we cannot predict the class from it since it is a unique value |
| --- | --- | --- |
| Column Customer\_ID | Removed Column Customer\_ID | I have removed the Customer ID because we cannot imply class from it since it is a unique value |
| Column Month | Removed Column Month | I have removed the Month because we cannot imply class from it |
| Column Name | Removed Column Name | I have removed this column because we cannot imply class from it |
| Column SSN | Removed Column SSN | I have removed this column because we cannot imply classes from it since it is a unique value |
| Column Num\_Bank\_Accounts | Removed Column Num\_Bank\_Accounts | I have removed this column because we cannot imply classes from it |
| Column Type\_of\_Loan | Removed Column Type\_of\_Loan | I have removed this column because we cannot imply classes from it |
| Column delay\_from\_due\_date | Removed Column Delay\_from\_due\_Date | I have removed this column because I think Number of delayed due date is a better attribute to predict class |
| Column Num\_Credit\_Inquiries | Removed Column Num\_Credit\_Inquiries | I have removed this column because we cannot imply classes from it |
| Column  Credit\_mix | Removed column  Credit\_Mix | I have removed this column because I am trying to imply credit score, which is a more detailed version of credit mix |
| Column Total\_EMI\_per\_month | Removed Column Total\_EMI\_per\_month | I have removed this column because we cannot imply classes from it |
| Column Payment\_Behaviour | Removed Column Payment\_Behaviour | I have removed this column because we cannot imply classes from it |
| Column Credit\_Limit | Removed Column Credit\_limit | I have removed this column because we cannot imply classes from it |
| Column Interest rate | Removed Column Interest rate | I have removed this column because we cannot imply classes from it |
| Column Monthly\_Invest | Removed Column Monthly\_Invest | I have removed this column because we cannot imply classes from it |

Handling Missing values and Outliers.

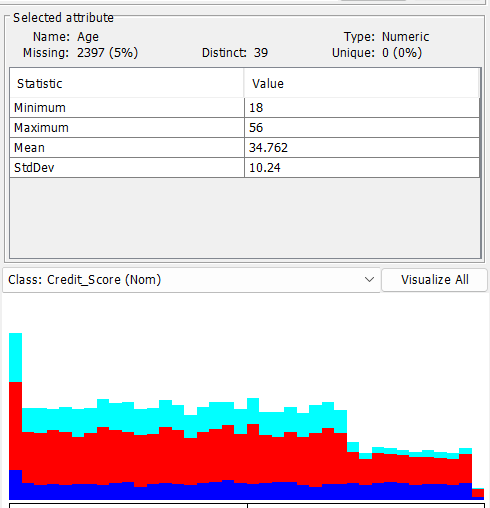
| Column age | Delete blank and wrong format ages(7635 rows) | Customers should be with sufficient age |
| --- | --- | --- |
| Occupation | Delete blank occupations(6547 rows) | We are not giving loans to people without occupations since they may not be able to pay the loan |
| Annula\_income | Delete blank and wrong format() | We should not give loans to someone whose income is not known |
| Monthly\_inhand\_salary | Delete blank and wrong format | We should not give loans to someone whose salary is not known |
| Num\_of\_credit\_card | Removing Num\_of\_credit\_card > 10 and Num\_of\_credit\_card < 0 | 96% of people have No more than 10 credit cards in this dataset. And after people with 11 credit cards is a huge gap. So rows with too many credit cards are not useful in my opinion. |
| outStanding\_debt | Removing rows with wrong format | We are not using the rows with insufficient data |
| credit\_history\_age | Removing rows with wrong format | We are not using the rows with insufficient data |
| Monthly\_balance | Removing insufficient Monthly\_balance | We should not give loans to someone whose Monthly balance is insufficient |

| Column age | Deleting rows with age>56 or age<0 | After age >56 is aged >95, which are not reliable datas here, and age <0 should not be of consideration |
| --- | --- | --- |
| Column occupation | Deleting rows without Occupation | We are not considering people without occupation |
| Column Num\_of\_Loan | Deleting rows with Num\_of\_loan >9 and Num\_of\_Loan < 0 | After Num\_of\_Loan =9, there was a big gap, so those people with Num\_of\_Loan> 9 are not reliable. |
| Column Num\_of\_delayed\_payment | Deleting rows with Num\_of\_delayed\_payment > 20 | After Num\_of\_delayed \_payment = 20, there is a huge gap to the next value, and people from 0 to 20 covers 99% of people in this dataset. |

1. After Modifying the data I converted the file from .csv to .arff
2. I have chosen the credit\_score as the class attribute
3. Discretising the class attribute

I do not need to do it for this dataset

1. Statistics



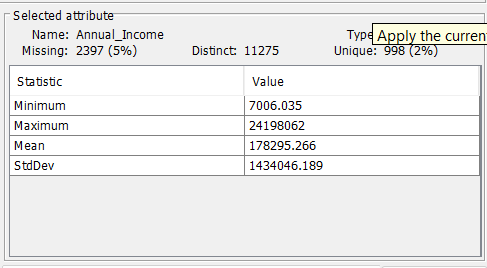
Age:

Min 18

Max 56

Mean 34.762

StdDev 10.24



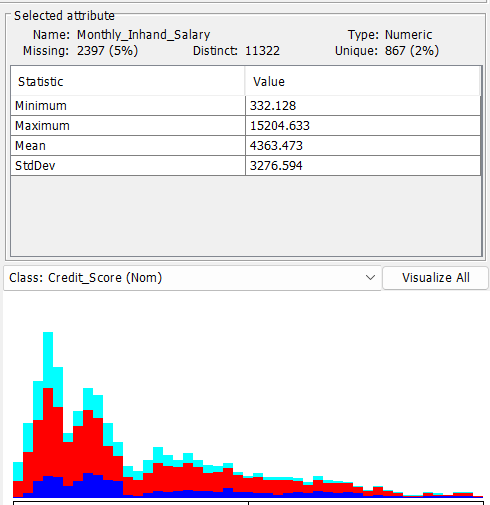
Annual\_Income:

Min 7006.035

Max 24198062

Mean 178295.266

StdDev 1434046.189



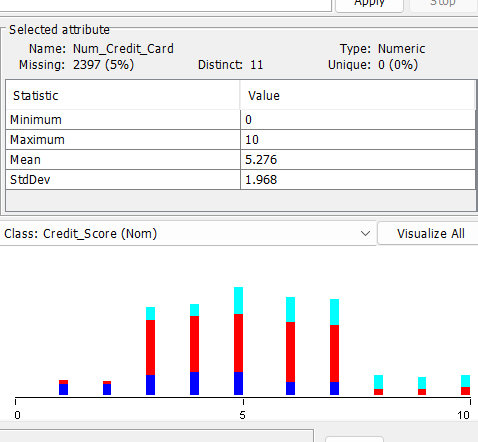
Monthly\_Inhand\_Salary:

Min 332.128

Max 15204.633

Mean 4363.473

StdDev 3276.594



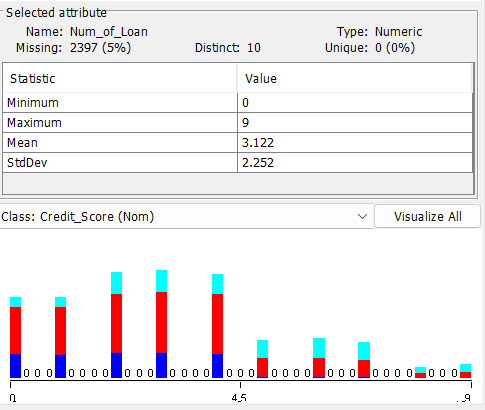
Num\_Credit\_Card:

Min 0

Max 10

Mean 5.276

StdDev 1.968



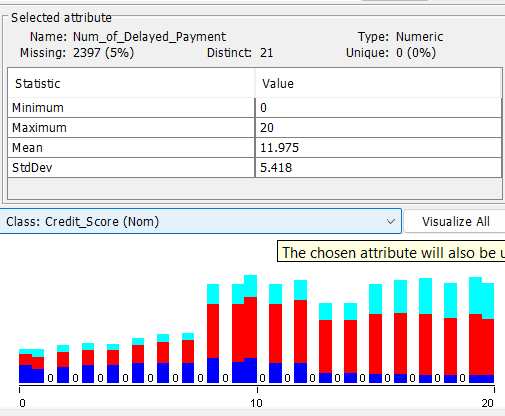
Num\_of\_Loan:

Min 0

Max 9

Mean 3.122

StdDev 2.252



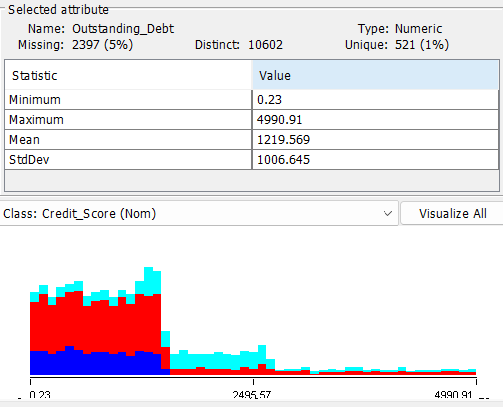
Num\_of\_Delayed\_Payment:

Min 0

Max 20

Mean 11.975

StdDev 5.418



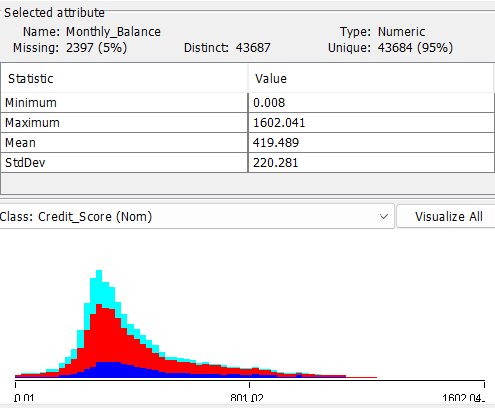
Outstanding\_Debt:

Min 0.23

Max 4990.91

Mean 1219.569

StdDev 1006.645



Monthly\_Balance:

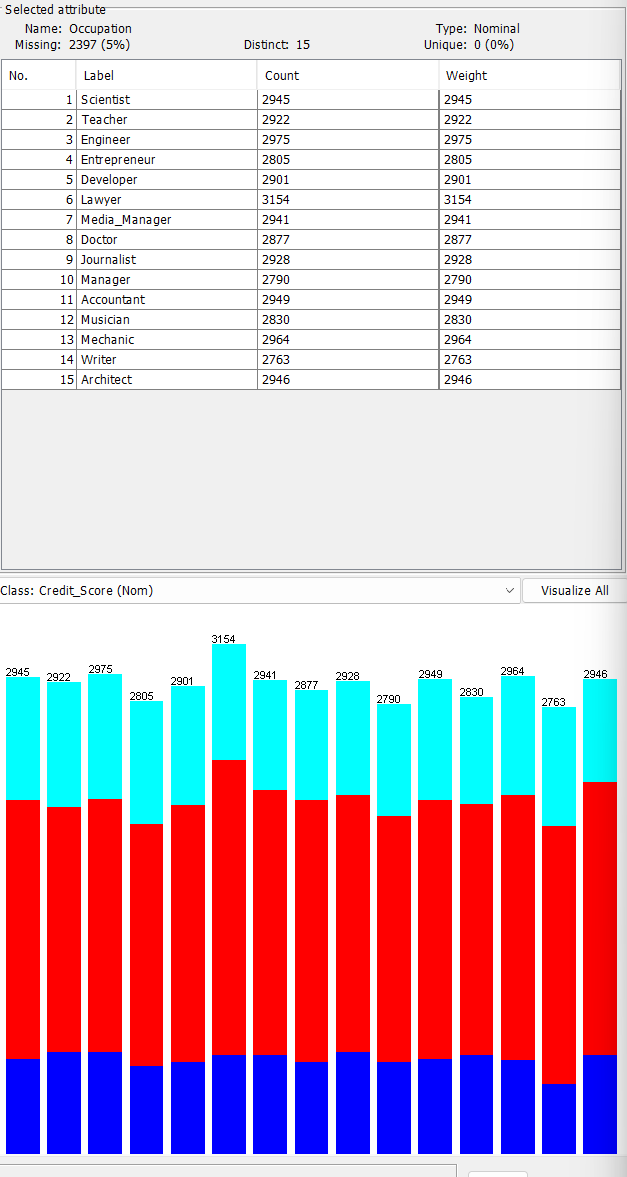
Min 0.008

Max 1602.041

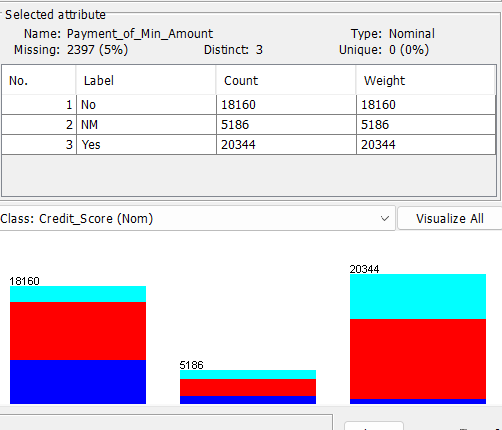
Mean 419.489

StdDev 220.281

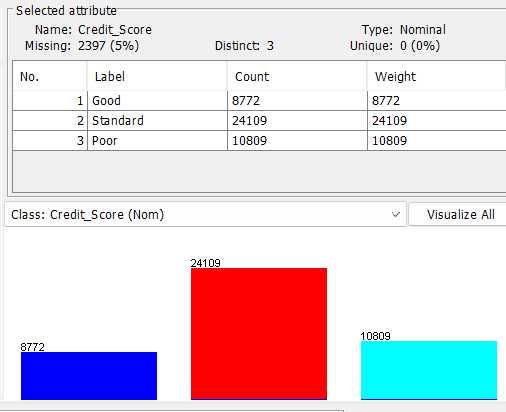
Occupation:

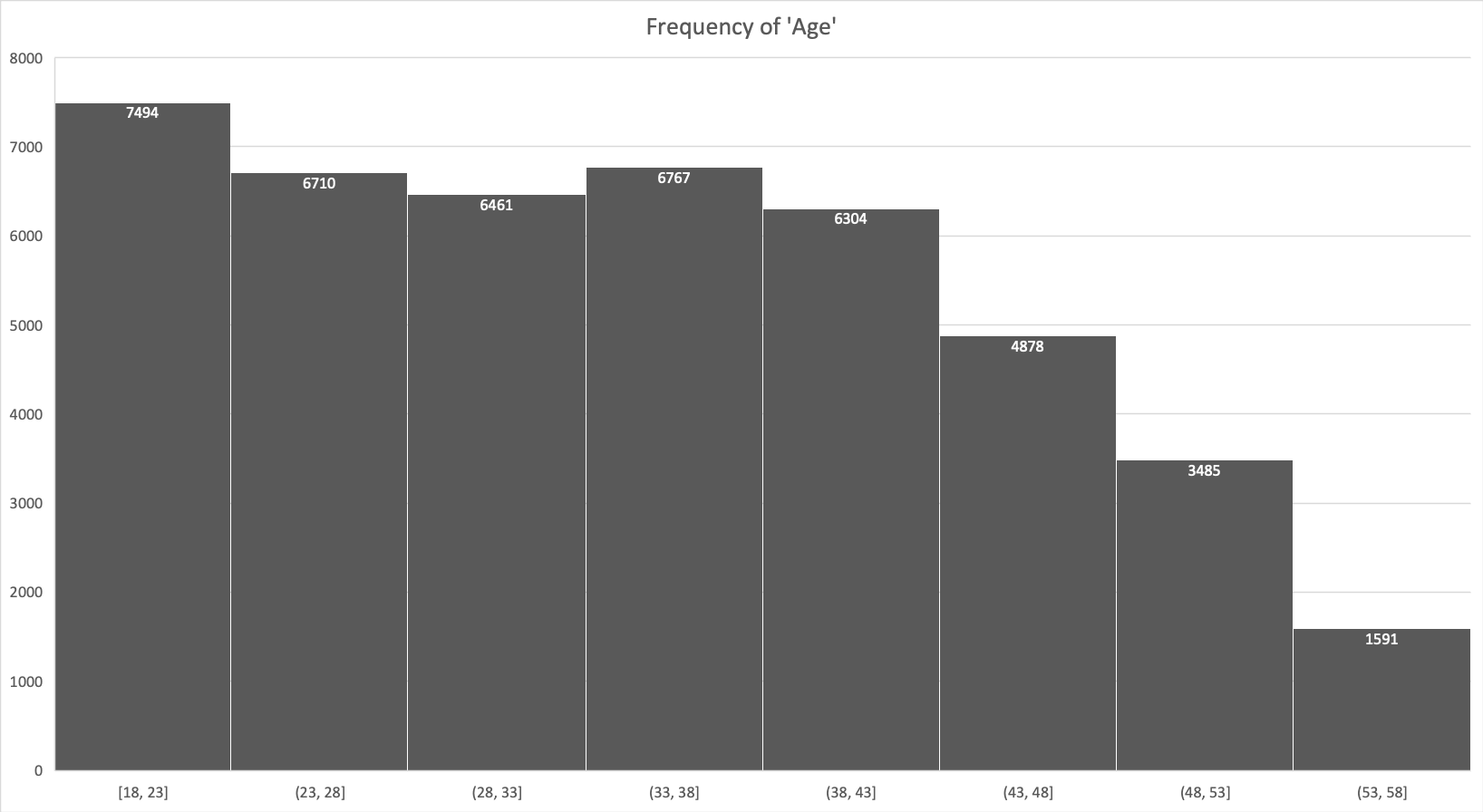


Payment\_of\_Min\_Amount:

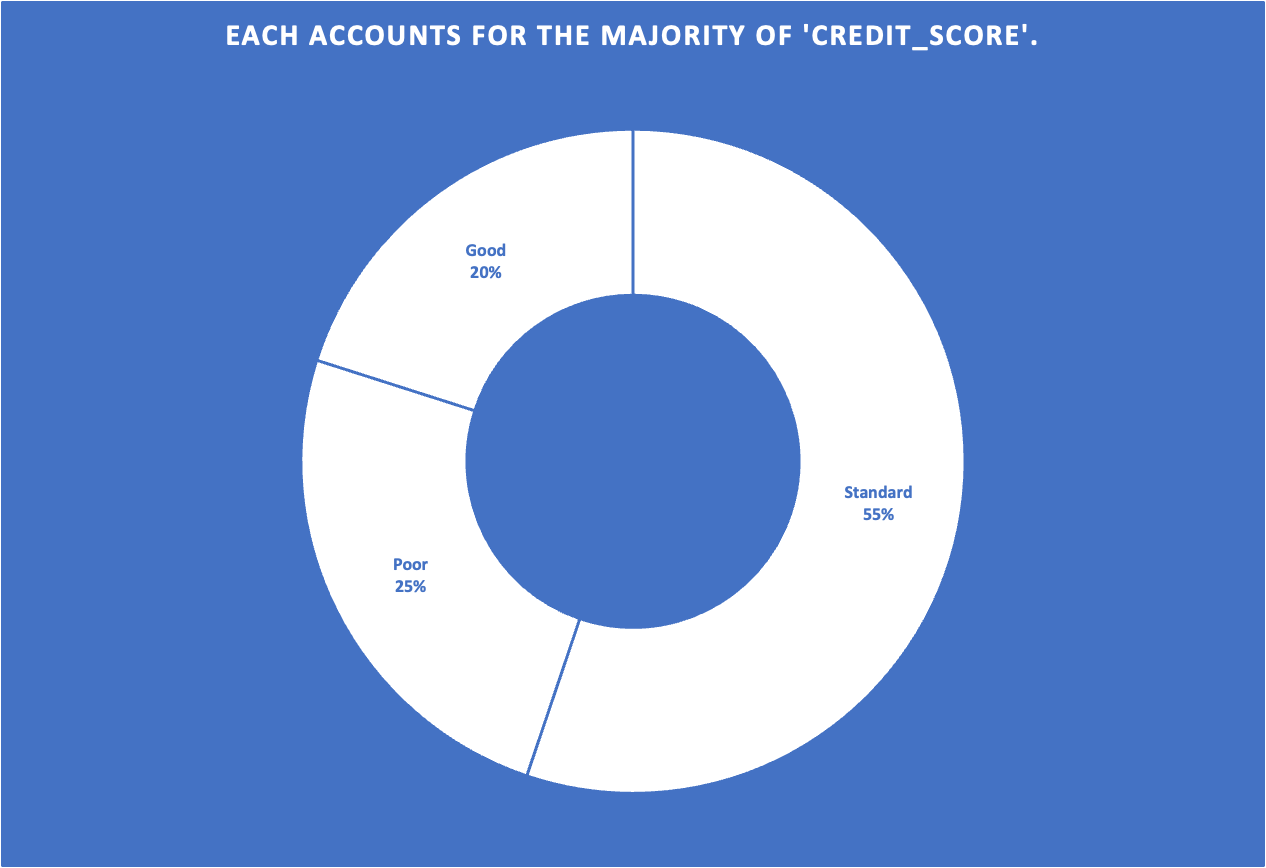


Credit\_Score:

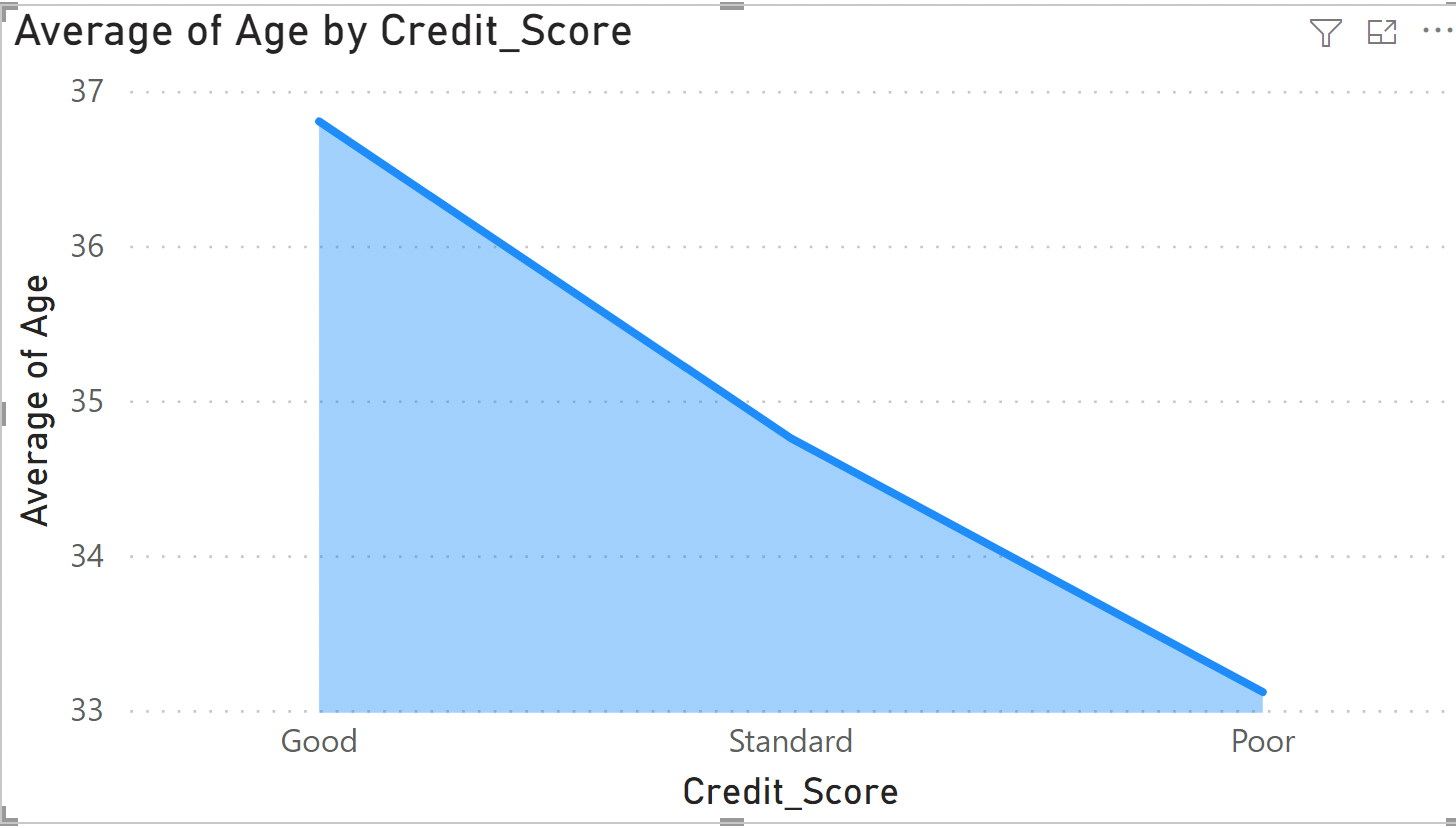




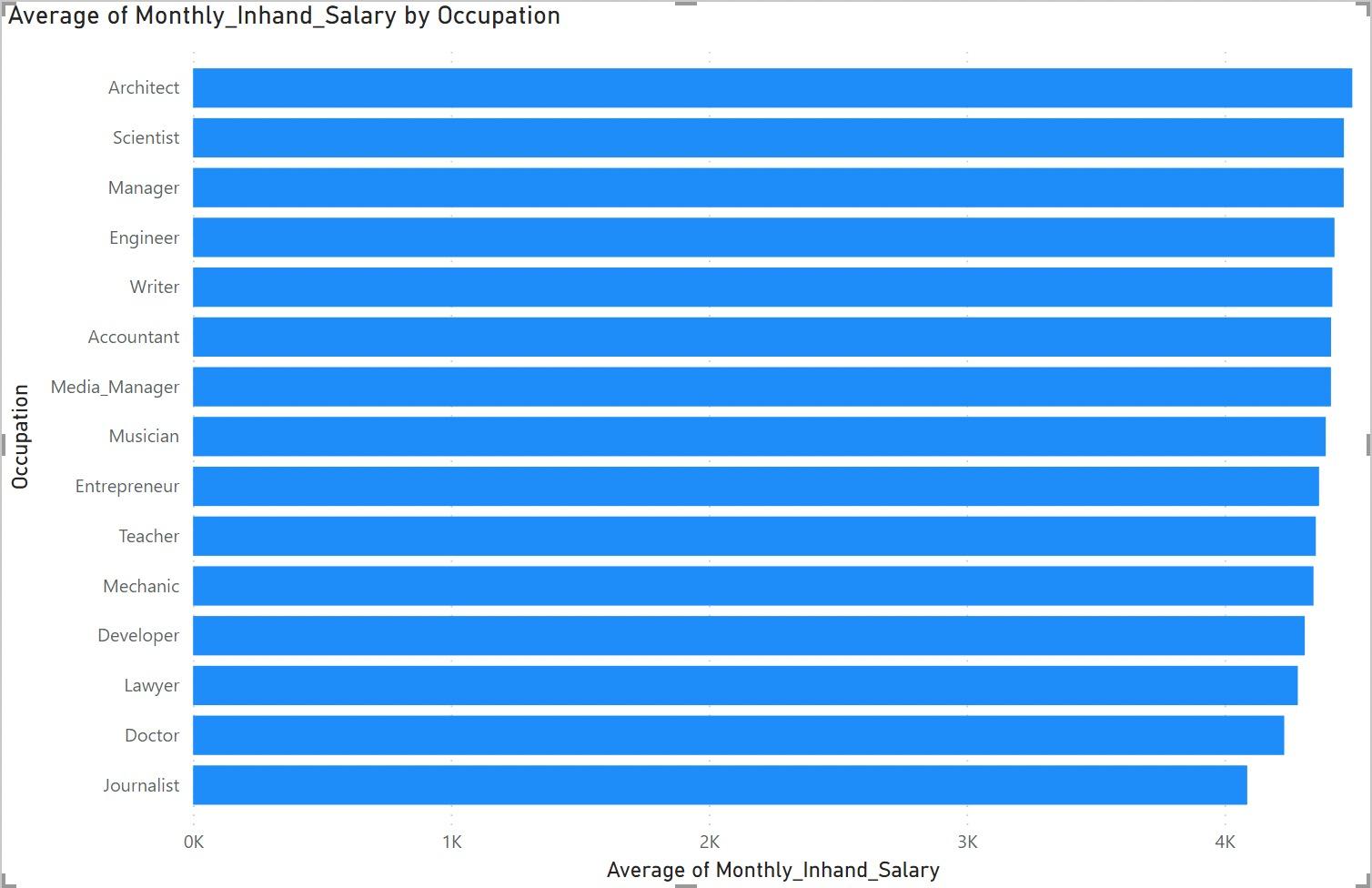
We can see 18-23 years customers are most in this database. 53-58 years old customers take the least part. Number of (23,28], (28,33],(33,38],(38,43] people are nearly the same. Numbers of customers dropped from group (43, 48] to (53, 58].



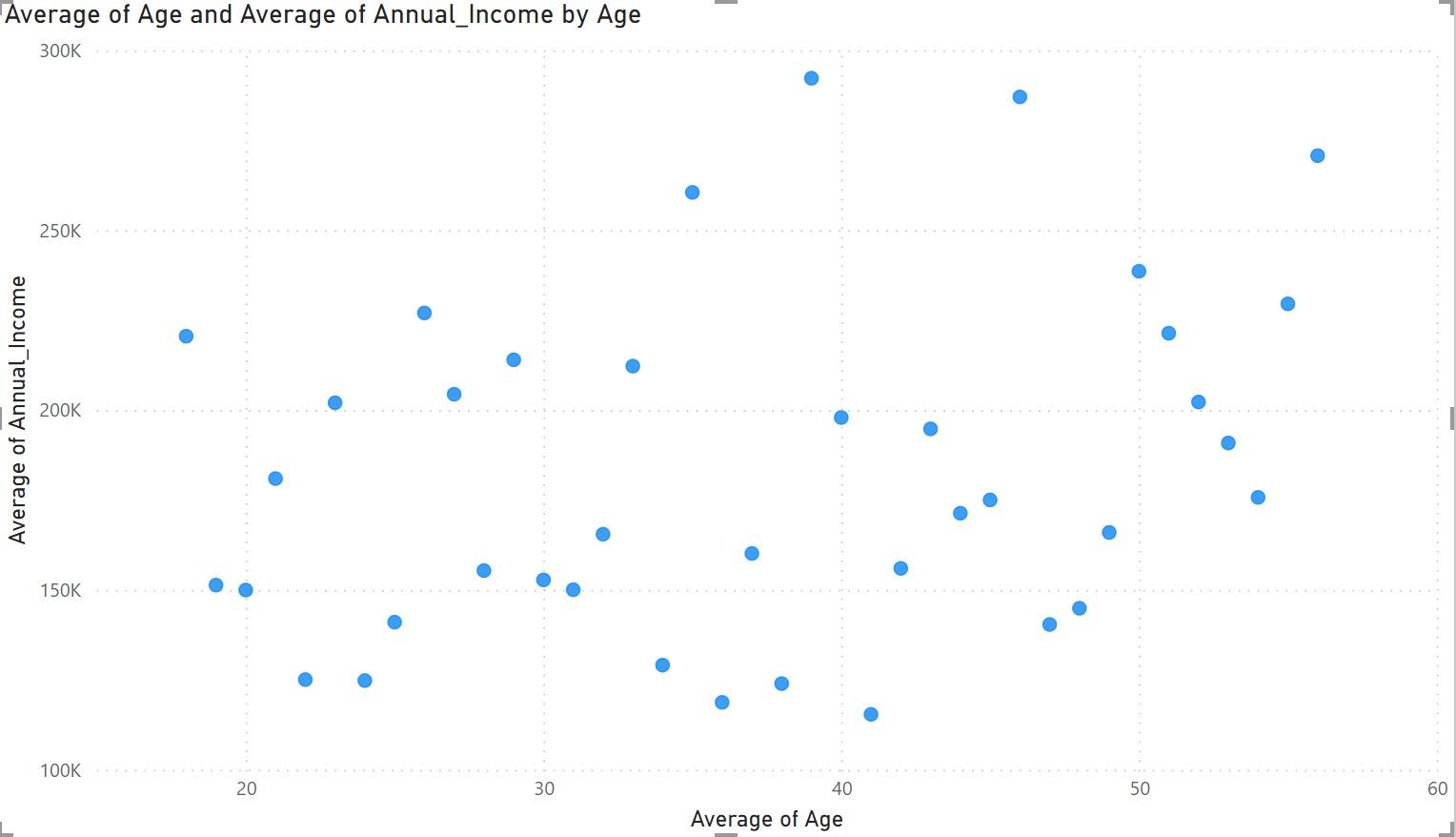
We can see standard Credit\_Score customers take more than half of this dataset(55%), whereas good credit scored ones are the least in this dataset(20%). The Poor credit scored ones takes ¼.



We can see the poor credit scored people’s average age is approximately 33 years old, standard credit scored people average age is approximately 35, good credit scored people’s average age is approximately 37. This shows the younger, the worth credit\_score people may have.



From this figure, we can see architects get highest average monthly inhand salary, whereas journalists get lowest monthly inhand salary. And average monthly inhand salary of all occupations from this dataset are from 4k to 5k.



We can see the average of annual\_income are from 100k to 300k for the people from 18 to 56 in this dataset.

4.

Classification

|  | Training set size | TP Rate | FP Rate | Accuracy | Reason |
| --- | --- | --- | --- | --- | --- |
| J48 | 60% | 0.646 | 0.318 | 64.6058% | The J48 has many useful features including accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc |
| 80% | 0.639 | 0.325 | 63.8759 % |
| Naive Bayes | 60% | 0.588 | 0.210 | 58.7829 % | Naive Bayes is good on multi-class prediction problems. It needs less data to be more accurate. |
| 80% | 0.579 | 0.210 | 57.8761 % |
| Random Sub Space | 60% | 0.658 | 0.344 | 65.759 % | The main advantage of the random subspace method is to **randomly select feature subsets, resulting in low-correlated multiple weak learners**. |
| 80% | 0.645 | 0.368 | 64.5147 % |
| Attribute Selected Classifier | 60% | 0.653 | 0.359 | 65.2966 % | It contains 2 steps(1) dimensionality reduction through attribute selection, and (2) classification, which you could check respectively |
| 80% | 0.649 | 0.348 | 64.8911 % |

1. When I tried to use Simple Cart classification I found it never stops.
2. I can not use AD Tree as there are no 2 class attributes in my datasets.

Reason

80% Comparison : 64.8911 %(Attribute Selected Classifier)>64.5147 %(Random Sub Space)>63.8759 %(J48)> 57.8761 %(Naive Bayes)

60% Comparison: 65.759 %(Random Sub Space)>65.2966 %(Attribute Selected Classifier)>64.6058%(J48)>58.7829 %(Naive Bayes)

I prefer to use RandomSubSpace or Attribute Selected Classifier to classify the dataset since the TP Rate from these 2 method are better(Attribute Selected Classifier is a little better in 80% and RandomSubSpace is a little better in 60%).

But I think Random SubSpace is the best one to predict since it gets the highest accuracy for 80%.