Hyperiondev

TASK

Exploratory Data Analysis on the German Credit Dataset

Visit our website

Introduction

The dataset I'm going to work on is the German Credit Dataset which consists of 1,000 credit samples, with each of them having 9 features such as Age, Sex, Job, Housing, Saving/Checking account, Purpose etc., that describes the category and characteristics of each borrower. The original dataset has 1000 rows and 10 columns. It is a widely used sample dataset which can be used to analyse the credit risk, credit classification, and modelling. I'm going to do an exploratory data analysis on this dataset and explain further details in different sections ahead.

Contents of the dataset:

- Age (numeric)
- Sex (text: male, female)
- Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
- Housing (text: own, rent, or free)
- Saving accounts (text - little, moderate, quite rich, rich)
- Checking account (numeric, in DM - Deutsch Mark)
- Credit amount (numeric, in DM)
- Duration (numeric, in month)
- Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)

Dataset source: https://www.kaggle.com/datasets/uciml/german-credit

DATA CLEANING

After having a good understanding of the dataset, I started with cleaning my dataset for effective exploration and better visualization. The major things I did while cleaning the data are presented below:

1. Firstly, I dropped unnamed column which just contains the indexing of the rows, so we don't need that.

2. I've merged the two columns, Saving Accounts and Checking Account to form a single column as Accounts Balance Status. The information on these two columns could be unusable when we use them separately as the large portions of these columns are missing. Even if these two are different types of accounts,

both contains same type of categorical values and reflects the bank balance status of a customer. For that reason, I've merged them into one so that we can have one single complete column rather than two incomplete ones. This will help us understand the financial status of the customers.

3. To merge the columns, I defined a function to get a single category with upper most value from two of the columns Saving accounts and Checking account. With apply() method and lambda function, I used the function to form a new column called 'Accounts Balance Status' and populate that column with the categorical value we get through our function, which gives us the overall accounts status. This simply gives us a brand new column which has the customer's account balance information and it'll help us to have a clearer view towards customers financial status.

4. The missing values in our dataset are represented by '?' symbol. So, before dealing with missing values, the first thing I'm going to do is replace '?' with the NaN value. It'll help us dealing with missing values properly.

5. Dropped the original Saving accounts and Checking account columns after I formed new column 'Accounts Balance Status' merging those two columns.

MISSING DATA

I encountered 99 missing values at the new column 'Accounts Balance Status' that we formed using two columns: Saving accounts and Checking account. The missing values are now represented by NaN value.

I imputed the missing values at "Accounts Balance Status" column with mode values. For that, first thing I did was calculate the mode value and fill NaN values with that using fillna() function. The missing value counts for the columns relevant to us was very little and most of the missing values were missing at random (MAR), so imputing those missing values with mode seem reasonable and efficient.

DATA STORIES AND VISUALISATIONS

After having our dataset cleaned, I wanted to have a descriptive look at our dataset so that we can understand our data better. As we can see below, we have total count, mean, standard deviation, minimum value, maximum value etc., for each column. For example, with this describe function, now we basic idea about few important highlights of the dataset such as, the maximum credit amount is 18424, there are 1000 customers in our dataset who are availing credit for different purposes, the oldest borrower is 75 years old and the youngest one is 19 years old, etc. This information is going to be helpful as we move ahead working with our dataset.

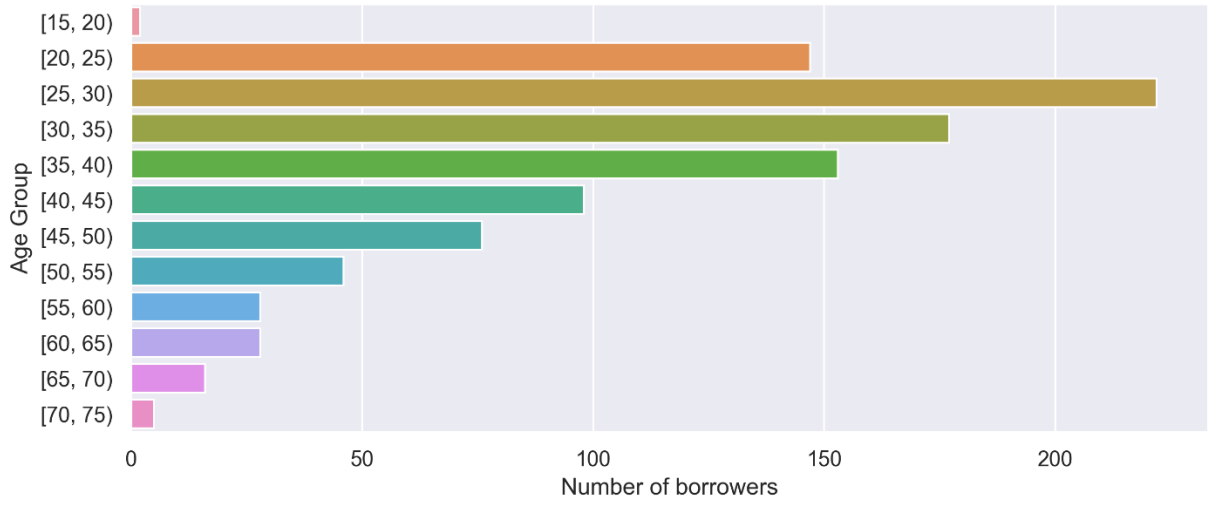
After exploring the German Credit Dataset, I identified many patterns and extracted a few insights. To ensure the clear visualization and neat presentation, the first step was to clean the dataset which I have done thoroughly and now I'm going to start with the visualizations part of report.

I have created few visualizations as a part of an EDA on this dataset, they are presented as under:

Number of borrowers based on different Age Groups:

To have a clear view of number of borrowers based on different age groups, I have created a bar plot which I have presented below.

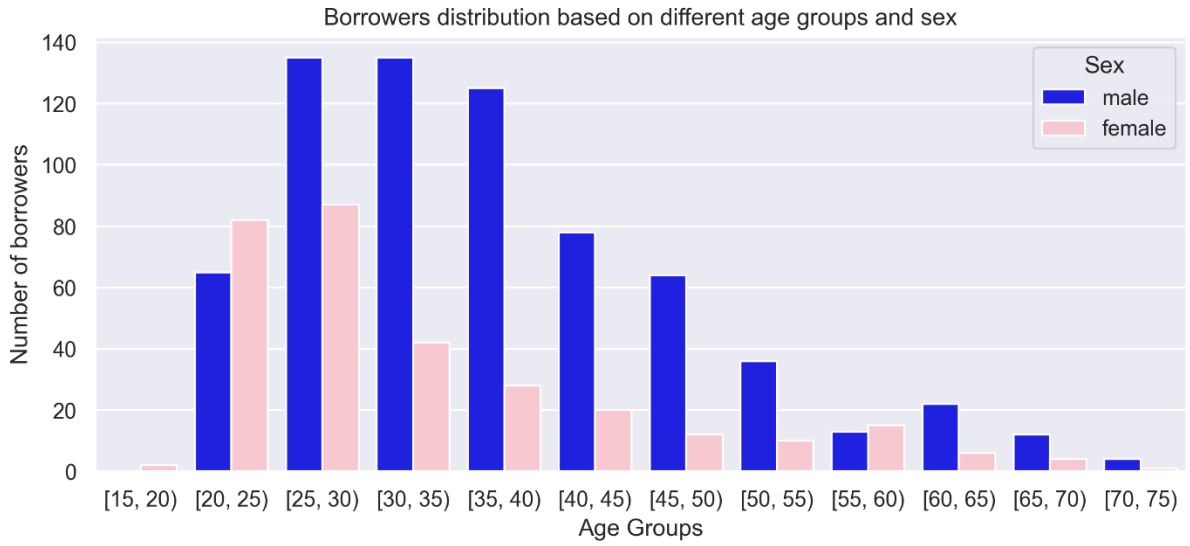
Figure: A



The above graph shows us the category of age group and their distribution. Most of the borrowers in our dataset are aged between 25 to 30 years. There are very few borrowers who are below 20 years of age. There are also borrowers who are above the age of 60. In general, majority of the borrowers are young and middle aged as most of them are 20 to 40 years old.

Borrowers' distribution based on different age groups and gender:

Figure: B



Looking at the above visualization, we can observe that the majority of borrowers are male. In every age group, except at 20-25, there's a majority of male borrowers. It is observed that women who are 20 to 35 years old tends to avail credit more. As we know that most of the borrowers in our dataset are young and middle aged, but there's not equal distribution of male and female borrowers. The number of male borrowers is significantly higher in comparison to the females, male borrower is almost twice the number of females. Hence, we can say that the credit advancement in German is more concentrated in male customers as per this dataset, or this situation may be the result of most women not being interested to avail loans or may be because of the reasons such as employment status, occupation, level of income, gender biasness, etc.

Credit by Purpose

The below three visualizations are the graphical representation of distribution of credit by purpose.

Figure: C

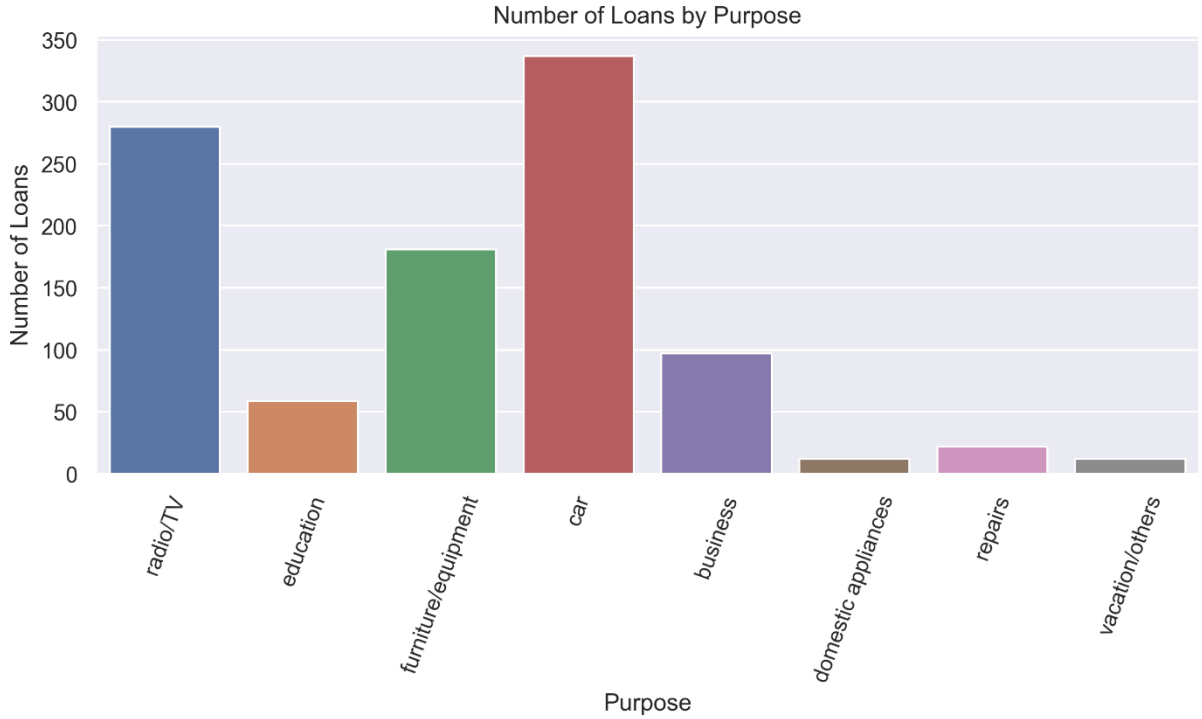


Figure: D

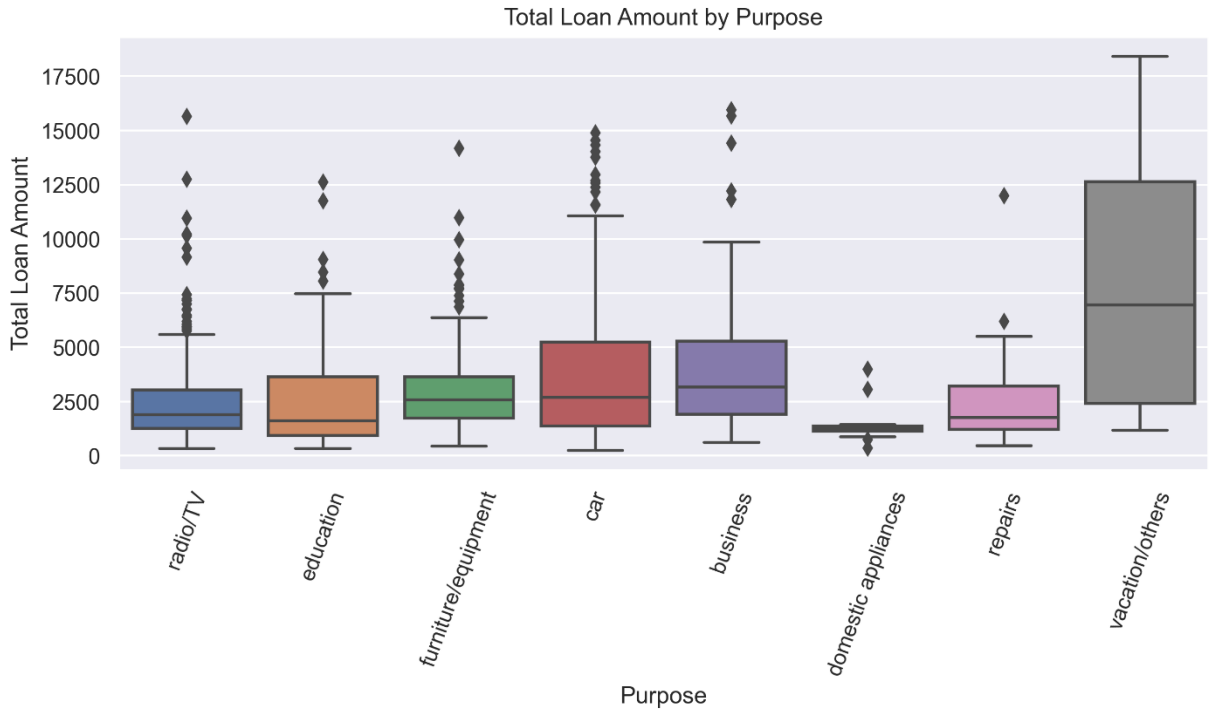
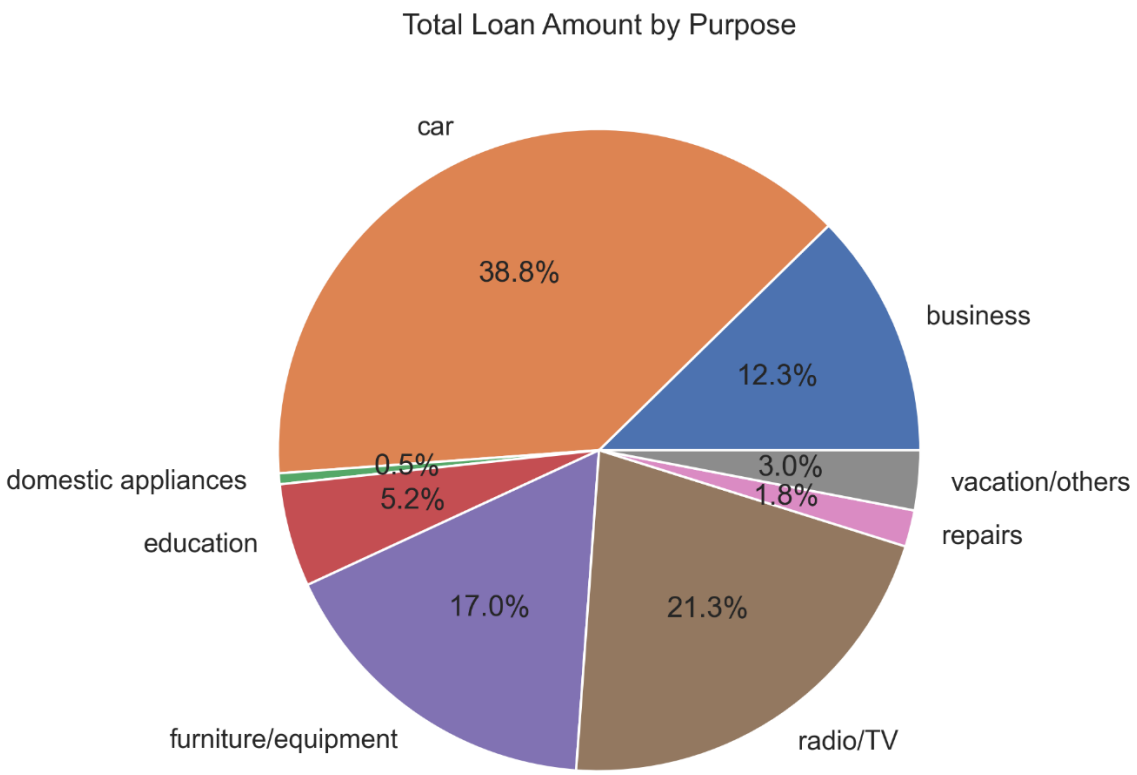


Figure: E



On the first graph, number of loans is represented by y-axis and purpose of the loan is represented by x-axis. It shows that number of loans by purpose. We can observe that majority of the loans are car loans, followed by radio/tv and furniture/equipment loans to be on the second and third place. Business, education and other types of loans are seemed to be a least of their priority while advancing credit. Overall, the credit advancements seem to be concentrated mostly in unproductive sectors when we look at the number of loans.

The second graph shows us the total loan amount based different loan purposes. I created this boxplot mainly to show the spread and central tendency of loan amount for different purposes, it will also help us to identify the extreme values and outliers for each loan purpose. As per this boxplot, the lowest loan amount sanctioned was 250 which was for a car and the highest loan amount ever was 18,424 which was sanctioned for the purpose of vacation/other. We can see a lot of outliers and extreme credit amounts for most of the purposes, so we can assume that loan spread from very little amount to much larger amount for each purpose.

The third visualization is a pie-chart which represents the total loan amount based on each loan purpose. Majority of the credit portfolio is concentrated to cars making it lenders first priority, radio/tv and furniture/equipment loans being on the second and third. Overall, as I said earlier, the credit portfolio does not look very much diversified, rather it is concentrated mostly in unproductive sectors, most of the credit amount being advanced as car loans.

Correlation heatmap: Job, Accounts Balance Status and Credit Amount:
I have created a heatmap below to showcase a correlation between job status, account balance status and credit amount. Let's have a look at it.

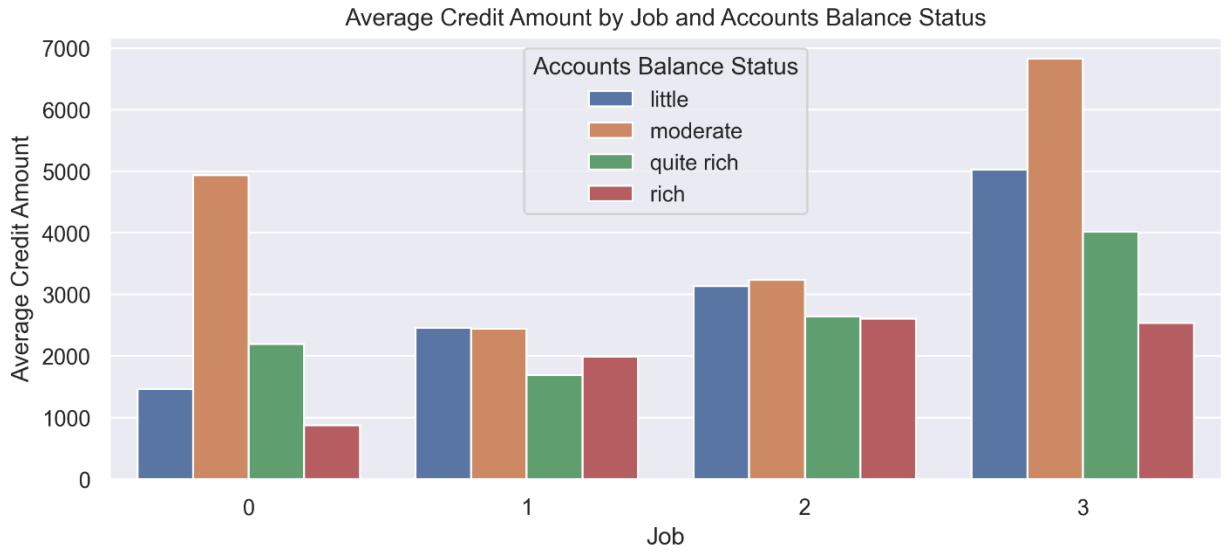
Figure: F

I decided to create a heatmap above just to have a quick look around correlations between Credit amount, Job status and Accounts balance status. Correlations that are close to 0 means there's a little or no relationship between variables. If a correlation coefficient is close to 1, there's a positive correlation and if it is close to -1 there's a negative correlation. One variable increases consistently as another variable increases if there's a positive correlation between variables and vice versa.

In this heatmap, we can observe that job status has a light positive correlation with credit amount whereas the Accounts balance status seem to have little or no correlation at all. It's just an insight to a relationship between few variables, there are other factors as well which influence the sanction of credit facilities. So, there's not much to observe and take aways from this particular graph here.

Average Credit Amount by Job and Accounts Balance Status:
Below is the graphical representation of average credit amount by job and account balance status.

Figure: G



The above graph gives us the overall idea of average credit amount distribution to different categories of borrowers based on their job status and bank account status. firstly, I calculated mean credit amount and then plotted it in the y-axis of the above graph to compare with job status and account balance status of the customers. In general, the customers with 'moderate' accounts balance status and 'highly skilled' job status are availing most of the credit.

The customers with 'unskilled and non-resident' job status are availing significantly less credit compared to customers with other types of job status. We can observe that major portion of the credit amount is sanctioned to the customers with skilled or highly skilled job status, and many of them has moderate balance status, for that reason, we can assume that the credit portfolio within this dataset seems reasonably low-risk based on borrower's accounts balance and job status.

After conducting an exploratory data analysis (EDA) on this dataset, I have identified several interesting and insightful patterns, and I have also extracted few insights which I have presented above.

We could even go further with this dataset to explore other aspects as well, but I decided not to dive into many other aspects this time because of the limited time. This dataset can be very useful for conducting further analysis of credit risk, classification, and modelling since it contains various important information about the credit portfolio.

THIS REPORT WAS WRITTEN BY: AJAY GHIMIRE, ***Data Science / HyperionDev***