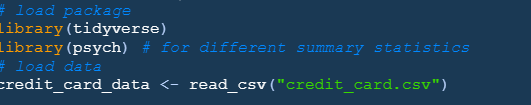
descriptive and exploratory analysis for credit card dataset

2022-12-21

**INTRODUCTION**

Our second set of data analysis will use the credit card data stored in the variable called credit\_card.csv. Its in comma separated value format meaning it can be loaded using the readr package which has the read\_csv() function. Before any analysis we always require a data, therefore loading it is paramount. We also would like to load the required packages since they will help us manipulate well dataset eg visualizations will require the ggplot2 package. The best way of loading a package is loading the meta package called tidyverse which carries a lot of packages useful for most statistical analysis and data manipulation.

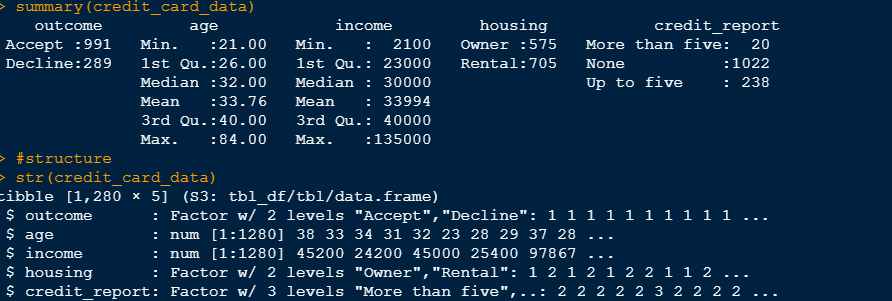
Figure 1: loading data and packages



**Summary Statistic of credit data set.**

We will require to know what types of data is in our dataset. Using the str() function we can understand well what type of data is in our dataset. Summary() function can also give us a clear understanding of numerical variables in our dataset eg the mean, median, 1st quartile and the 3rd quartile. We can also use the describeBy() function in the psych package also enables the reporting of a number of summary statistics (including the quantity of valid case scenarios, mean, standard deviation, median, trimmed mean, MAD: median absolute deviation, lower limit, maximum, range, skewness, and kurtosis) by a grouping variable. This can be accomplished as follows.

Figure 2: summary statistics of data



Some value are missing in our dataset therefore they will be dealt with below. We also see that some data type are not correctly assigned therefore that will be delt with when cleaning the dataset belows. We also need to relevels some levels since there is a lot of repetition.

**Cleaning of Credit Card Dataset**

We will start by cleaning our credit card dataset. Data cleaning is the process of converting unreliable data into reliable data that is simple to evaluate. Its goal is to filter the substance of statistical claims based on the data and the validity of those claims. Additionally, it affects the statistical conclusions drawn from the data and raises your total productivity and data quality. The numerous goals of data cleaning include the following:

* Remove errors
* Reduce Redundancy
* Upgrading Data Reliability
* Delivery Precision
* Ensure Uniformity
* Ensure thoroughness
* standardize your techniques

**Handling missing values denoted with NA**

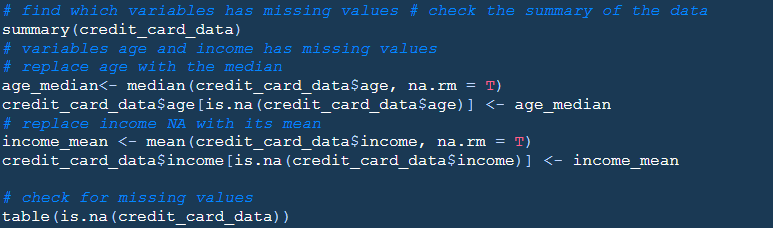
We will start by checking for missing values in our dataset. Using table() and is.na() function we can see the total number of true missing values on the whole dataset. This is shown below with codes done using R.

Figure 3: checking for missing values.



There are 7 missing values in our data set. We can decide to use the mean and median of the corresponding values to replace this values. Codes shown below will perform this operation.

Figure 4: replacing NA values

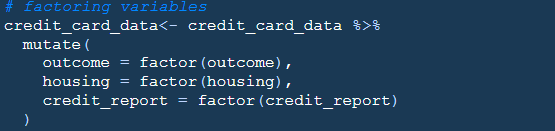


Above codes eliminate the missing value by replacing these value in age variable with its median and the income variable with its mean. Finally we check and see there’s no missing values present after performing the above process.

**Refactoring the variables present in the data**

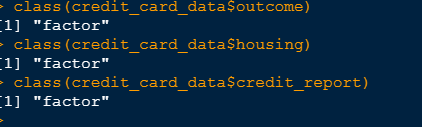
Some variables are supposed to be converted to factors. In an experiment, factors are the variables that the investigator controls to see how they affect the response variable. Only a tiny subset of values, referred to as factor levels, are possible for a factor. Factors may be based on a continuous variable or a categorical variable, but they can only employ a small number of values that the experimenters select (Frost 2022). Factoring of variable will be done with factor() function and mutate() function that will change this class completely.

Figure 5: Factoring variables



The required data variables were converted into factors and this can be confirmed using class() function.

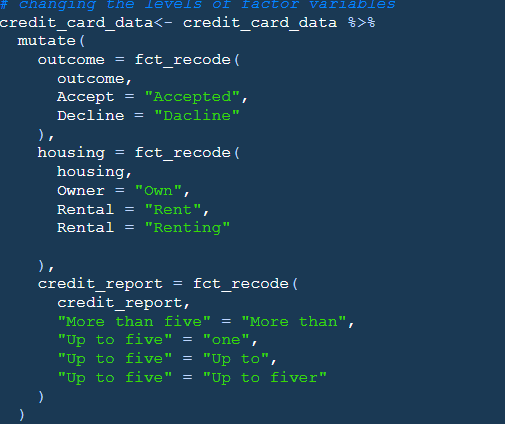
Figure 6: Checking the class of the variables



**Changing the Levels of Different Factor Variables.**

Some of the levels present in the factor variables have some typing error and have some common meaning. For instance, in credit\_report factor variable, we can create three levels such that levels like “one” and “Up to fiver” may fall under the same levels. This will be accomplished using the forcats package.

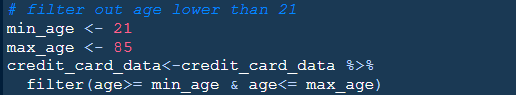
Figure 7: releveling levels of factor variables.



**filtering out age variable according to minimum age for applying credit card and filtering out the outliers.**

Since the minimum age for applying for the credit card is 21, we need to filter out ages lower than 21. In the US, individuals over the age of 21 can independently apply and be authorized for a credit card without providing evidence of a suitable income (May 2017). Also the age variable seems to have outliers, ie, high ages that does not make sense.US women live 5.7 years longer on average, until they are 80.2 years old. At 70.6 years for males and 75.1 years for women, the average death age around the world is a few years lower. These are 77.8 and 83.3 years in the European Union, respectively (WorldData.info 2022). Therefore, we need also to filter out this outliers with a mortality rate of 85. This can be accomplished using the filter function as shown below.

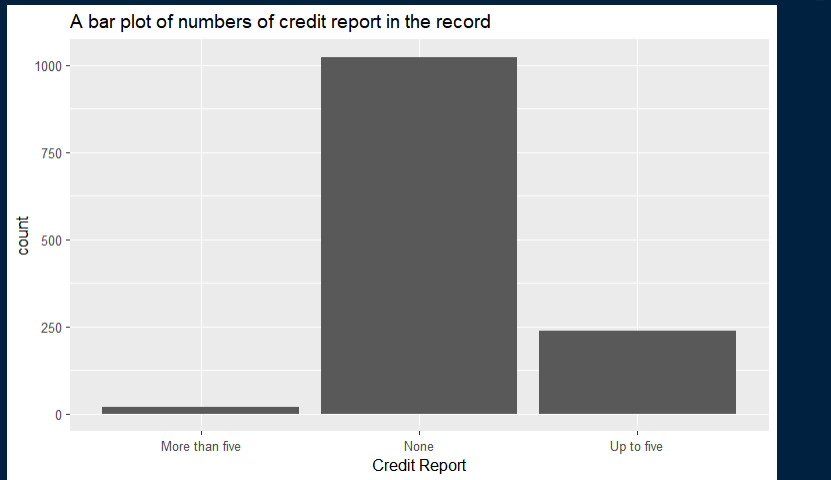
Figure 8: Filtering age according to minimum age for credit card application and outliers



**DESCRIPTIVE AND EXPLORATORY ANALYSIS ON CREDIT CARD DATASET**

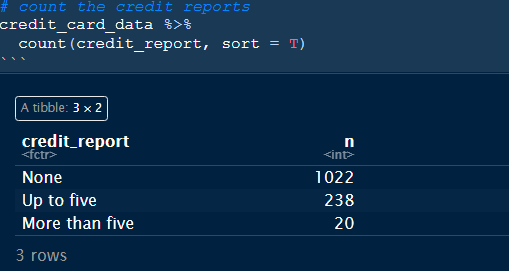
We are going to analyse now the credit card data since its a clean dataset. We first start by looking at the credit\_report variable , that is, how the number of credit reports in the applicants record. A barplot will show us the relationship between the levels of this variable.

Figure 9: Barplot for credit reports



From the above plot its clear that the number of bad record for credit card applications for these applicants is displayed above with bars. The level with the highest count is “None” with a count of more than 1000. This means that more than 1000 applicants have no bad records on the credit report. We can show a count of this levels of credit reports on the applicant record.

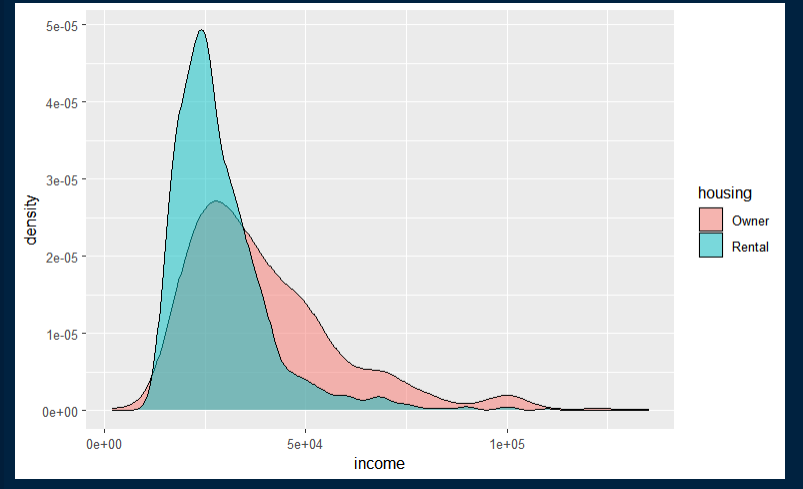
Figure 10: count for the most frequent credit report



We therefore see that the most report recorded on the credit card applications is none, ie, no record was showing for most applicant with a count of 1022 followed by up to five with a count of 238 and more than five with a count of 20.

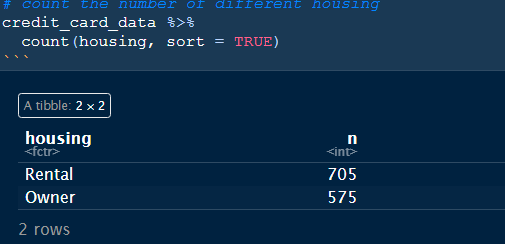
We can further this analysis by looking at the housing method for the applicants together with the income earned. This is shown below with a density plot.

Figure 11: A density plot for income filled with housing method



Most applicants who have applied for credit cards lives on rental housing and an income of less 50000 as shown in the density plot above. Although those who owns houses are few compared to those who rent, it is seen that they earn a high income spreading to about 500000 per annum. Lets count the number of different housing as per the given data.

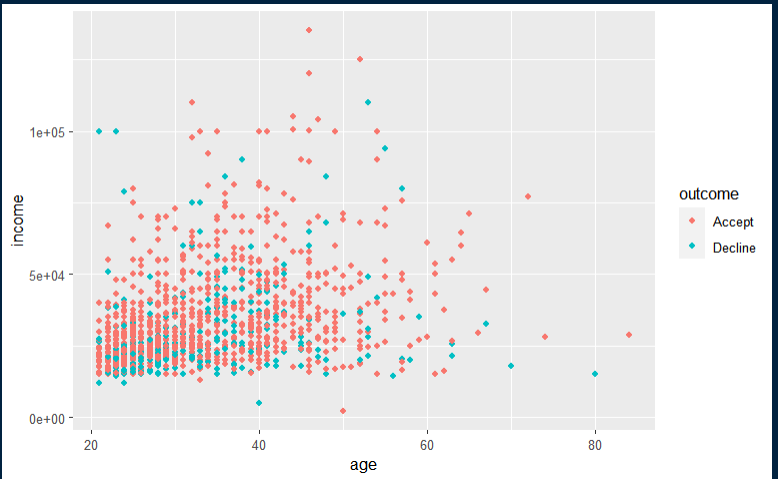
Figure 12: count for housing method



Most credit card applicants lives in rentals with a highest number of 705 while those who owns houses and applied for credit card are 575.

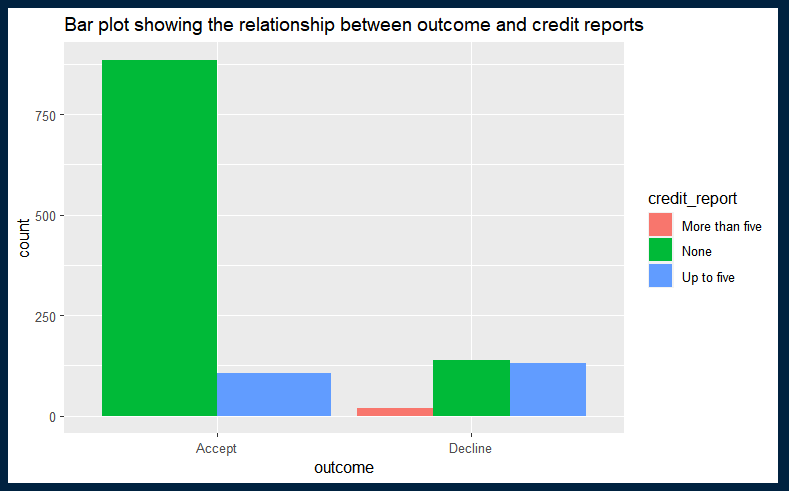
We can also check the outcome of credit card application ie, whether it was accepted or declined comparing it to the income earned and the different age group.

Figure 13: Scatter plot for income versus age



For the above scatter plot, we see that income is the dependent variable depending on age and coloring with different outcomes, we clearly sees that the most outcome provided to credit card applicant is “Accept”. We can say this is due to most of the applicants have attained the required age and they are also earning. The income is highly concentrated between the ages 20 and 60 since most of this people are still working but above that most have retired thats why we see a lower rate above 60. We can also make different bar plots to show how the outcome correlate with the credit\_report.

Figure 14: barplot for outcome and credit\_report



Its clearly evident that most applicants who applied for credit cards and were accepted had no bad records and also those whose outcome was accepted, some had an up to five bad record. Looking at the declined bars we also see that none applicants had any bad records, a lower rate of up to five bad credit record cases were also declined and a few rate that was declined had more than five cases of bad credit record.

**CONCLUSION**

we can conclude that after the analysis, most applicants have no bad credit reports in the applicant’s record. Also, we can conclude and say most applicants who applied for credit cards lives in rental houses. Most applicant’s age lies between 21 and 85 but most applicants who earns an income are in between age 21 and 60. This is due to after 60 years its highly likely most people have retired.

**Reference**

# Staff G.S. 2017, What is the Minimum Age for a Credit Card? June 15, 2017.

# Available from: < [*https://growingsavings.com* )](%20https://growingsavings.com%20)) > [15 June, 2017].

WorldData.info 2020, Life Expectancy for Men and Women. Available from: [*https://www.worlddata.info/life-expectancy.php#:~:text=The%20world%20average%20age%20of,data%20from%20the%20year%202020.*](https://www.worlddata.info/life-expectancy.php#:~:text=The%20world%20average%20age%20of,data%20from%20the%20year%202020.)