SCMA 648 Business Data Analytics

Assignment 2 - Data Preparation and Visualization

The quarter 3 data had to be preprocessed. The features were assigned their respective datatypes. The features containing dates were given a standard format. This was saved as an "rda" file. The working directory was set according to the source file location. The rda file that was created was loaded to the R environment along with the required package dplyr.

1. Consider the sample dataset: 2_hw.csv. Each row corresponds to a customer transaction. The first column contains the date and time of the transaction, the second column contains the product ID, and the third column contains the number of items purchased for a retailer. Suppose that you are planning to aggregate and summarize such data collected over the period of a year by product ID. Describe 10 data fields that you could generate based on these three columns of data that would contribute to understanding customer behavior. Each field should provide a value for each product ID. In other words, there will be one row for each product ID in the resulting dataset.

   The dataset has three attributes: Timestamp of the purchase, Product ID and the number of products sold. Different features can be extracted based on these attributes. They are as follows:

   - Total number of each product sold in a year.
   - Total number of each product sold on different days of the week (Monday – Sunday). There will be 7 separate features according to this.
   - This can also be converted to 2 separate features - Total number of each product sold on weekdays (Monday – Friday) and Total number of each product sold on weekends (Saturday – Sunday).
   - Total number of each product sold in different months (January – December). There can be 12 separate features corresponding to each month.
   - This can also be converted to 4 separate features considering quarters in a year - Total number of each product sold in Quarter 1 (January – March), Quarter 2 (April – May), Quarter 3 (June – September) and Quarter 4 (October – December).
   - Total number of each product sold in a day.

- Total number of each product sold at different hours during the day (24 hours). The sale of one product might be higher than that of other during one particular hour of the day.
- Average time (in weeks or days) between two consecutive purchases of each product.
- Frequency of buying each product in the same week, month or quarter.

2. Determine the mean and median of the borrowers' months since last public record (mths_since_last_record). What is the number of observations for which this value is missing? Create a new variable where the value is imputed with the median. What are the mean and median after imputation?

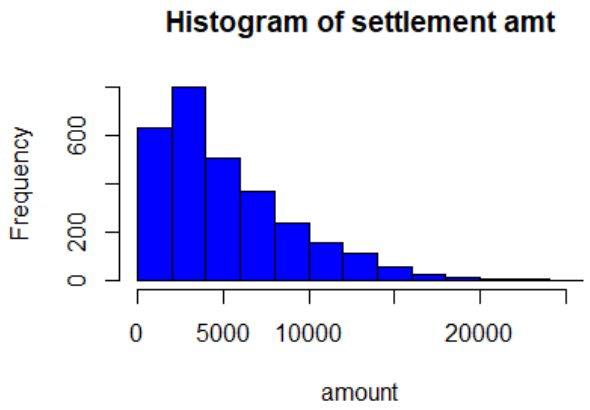
   The mean and median values were determined using summary.

   - The mean is 75.8.
   - The median is 79.
   - The total number of observations for which this value is missing is 103863.

   After imputation,

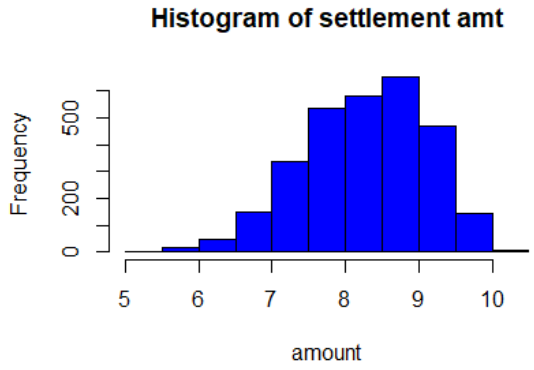
   - The mean is 78.51.
   - The median is 79.

3. Plot a histogram of the settlement amount for loans (for those with a settlement) and determine if the distribution is skewed. If so, create a transformation of the settlement amount data and create a histogram of the result. Does the new data appear to be normally distributed?

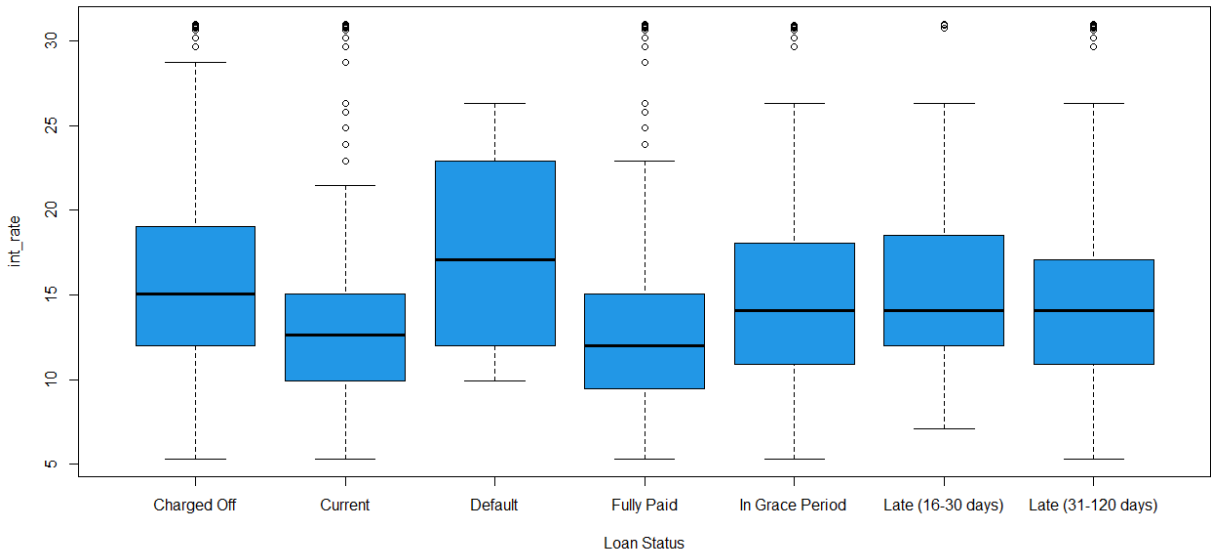


To determine the skewness of distribution, histogram has been used for the feature settlement_amount. As seen, the distribution is right-skewed.

The distribution is skewed. This can be changed using log transformation. After log transformation, the distribution appears to be normal.



4. Create a boxplot of interest rate (int_rate) for each loan status (loan_status). Which status has loans with the highest median interest rate? Do the interest rates for fully paid loans tend to be higher or lower than those that are charged off?



A boxplot of interest rate and each loan status has been created as seen. The default status has the highest interest rate with respect to median. The interest rates for fully paid loans tend to be lower than those that are charged off.