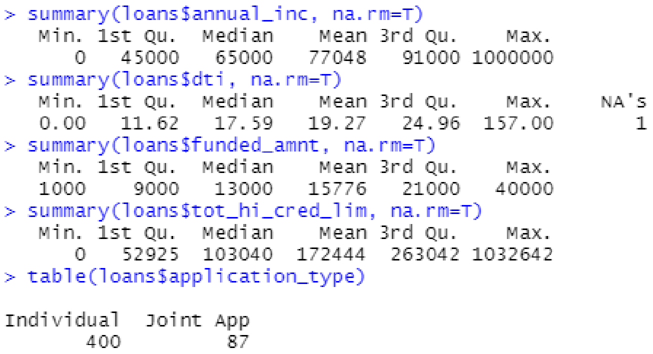
Mini-Project 1

1. Data Description

This data set represents a sample of 487 approved loans from LendingClub during the third quarter of 2018. For this exploratory data analysis, I have elected to focus on five specific variables: four numerical and one categorical. They include *annual_inc*, *dti*, *application_type*, *funded_amnt*, and *tot_hi_cred_lim*. *annual_inc* is the annual income reported by the borrower in their application and has a median of $65,000 and a range of $1,000,000. Debt-to-Income (*dti*) has a median of 17.59, along with a range of 157. Funded Loan Amount (*funded_amnt*) has a median of $13,000 and a range of $39,000. Finally, Credit Limit (*tot_hi_cred_lim*) has a median of $103,040 and a range of $1,032,642. These approved loans are divided into two application types: Individual applications (400 observations) and Joint applications (87 observations).



```
> summary(loans$annual_inc, na.rm=T)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0   45000   65000   77048   91000 1000000 
> summary(loans$dti, na.rm=T)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   11.62   17.59   19.27   24.96  157.00       1 
> summary(loans$funded_amnt, na.rm=T)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1000    9000   13000   15776   21000   40000 
> summary(loans$tot_hi_cred_lim, na.rm=T)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0   52925  103040  172444  263042 1032642 
> table(loans$application_type)

Individual  Joint App 
       400         87 
```

Figure 1: Summary Statistics

2. Variable Distributions

Figures 2 & 3 depict the distributions of the four numeric variables of interest. These distributions are similar in that all are right-skewed, asymmetric, and unimodal. For example, when looking at the boxplot and histogram for *annual_income*, it is evident that this variable is right-skewed because of its many outliers. The right skewness becomes even more noticeable when comparing the variable's mean and median. In general, when the mean is greater than the median, this indicates the distribution is right-skewed. This is exemplified in Figure 1, where *annual_income* has a mean of $77,048 but a median of only $65,000.

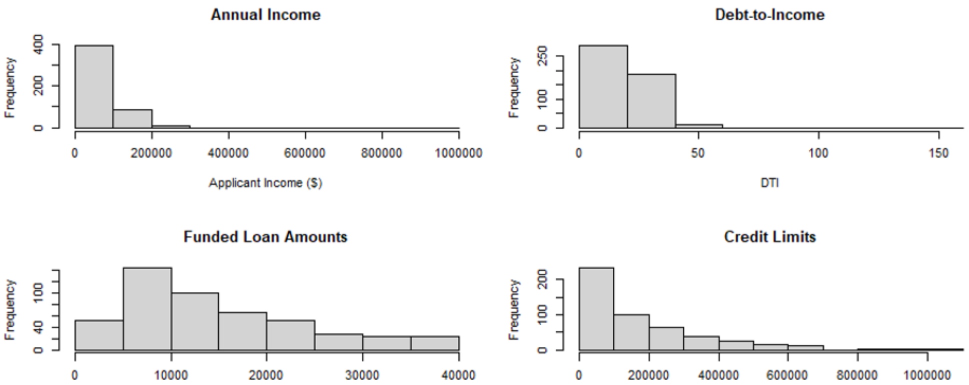


Figure 2: Histograms of Variable Distributions

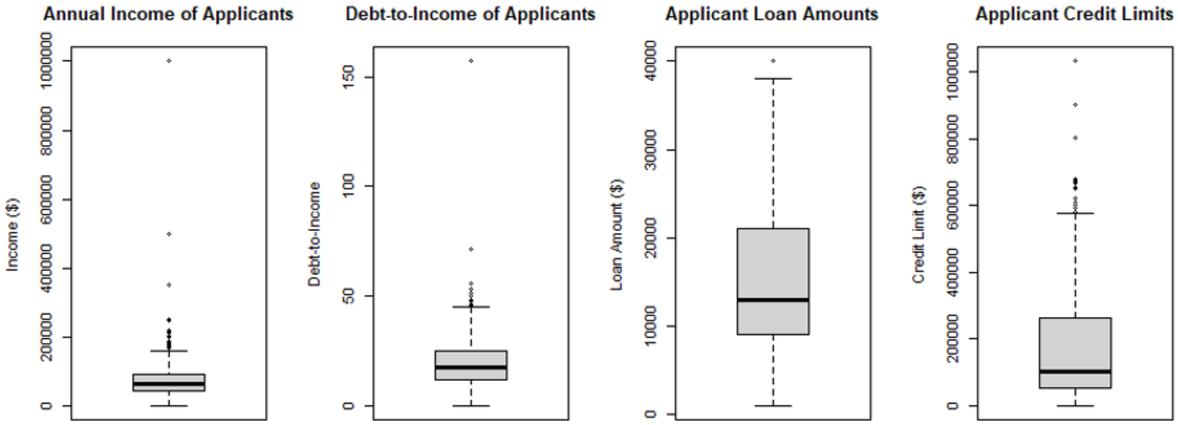


Figure 3: Boxplots of Variable Distributions

3. Patterns and Associations

Figure 4 provides an additional level of detail of the boxplots seen in the "Variable Distributions" section, this time breaking each boxplot down by *application_type*. When looking at "Annual Income of Applicants," one can see that the median income for individual applicants is slightly greater than that of joint applicants. However, this occurrence is not the same, with regard to the latter three boxplots, as the medians of "Debt-to-Income," "Applicant Loan Amounts," and "Applicant Credit Limits" are all greater at the joint applicant level.

The scatterplot in Figure 5 shows all sampled applicants' incomes compared to their credit limits, grouped by *applicant_type*. Without subsetting, there is a weak, positive association between income and credit limit, as is evidenced by a Pearson correlation of 0.4504 in Figure 6. Interestingly enough, when these two variables are subset by *applicant_type*, the Pearson correlation rises to 0.5258 for individual applicants and decreases to 0.3557 for joint applicants, suggesting that there is a stronger association between income and credit limit for individual applications.

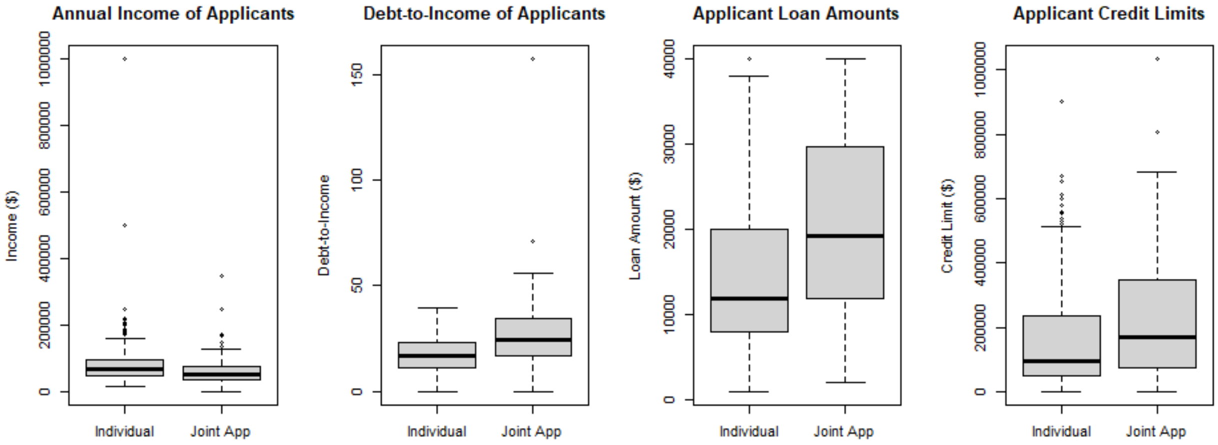


Figure 4: Boxplots of Variables Grouped by Application Type

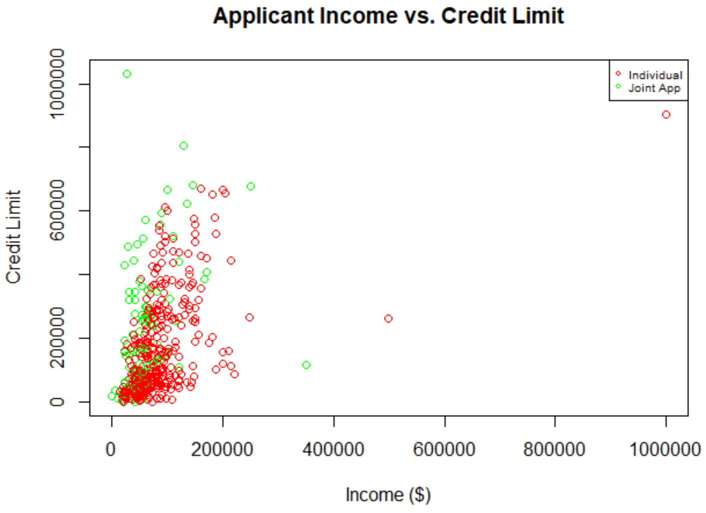
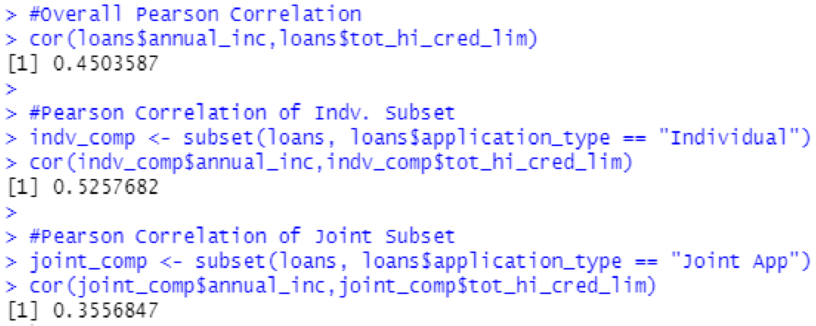


Figure 5: Scatterplot of Applicant Income and Credit Limit Grouped by Application Type



```
> #Overall Pearson Correlation
> cor(loans$annual_inc,loans$tot_hi_cred_lim)
[1] 0.4503587
>
> #Pearson Correlation of Indv. Subset
> indv_comp <- subset(loans, loans$application_type == "Individual")
> cor(indv_comp$annual_inc,indv_comp$tot_hi_cred_lim)
[1] 0.5257682
>
> #Pearson Correlation of Joint Subset
> joint_comp <- subset(loans, loans$application_type == "Joint App")
> cor(joint_comp$annual_inc,joint_comp$tot_hi_cred_lim)
[1] 0.3556847
```

Figure 6: Pearson Correlations

4. Analysis of Missing Values, Outliers, and Other Unusual Features

After an analysis of the data, there are indeed some instances of missing values, outliers, and other unusual features. With respect to missing values, the only variable of interest containing missing values is *dti*. Fortunately, there is only one instance of a missing value for this variable. This lone missing value poses little threat to the integrity of the data and therefore should have no true impact on an analysis. Unfortunately, this same conclusion cannot be made regarding outliers in the data. Looking at the boxplots in Figure 3, it is clear that outliers exist in all four variables. Each of the outliers is 1.5x greater than the third quartile, with none being less than 1.5x the first quartile. These outliers inflate the means for all four variables, leading to a right-skewed distribution. Although not designated as outliers, values of zero were encountered in several variables. For example, *dti*, *annual_income*, and *tot_hi_credit_lim* all had minimums of zero. An applicant having a *dti* of zero could be possible, as such a value would merely indicate that the individual has no outstanding debts. However, an individual having either an annual income or credit limit of zero is highly unusual, especially considering they are applying for a loan. Income and credit worthiness are two of the most important factors in deciding whether someone gets approved for a loan. The fact there were two instances of this represents possible irregularities. One can expect such values to increase measures of dispersion, like range, variance, and standard deviation.

6. Conclusion

In conclusion, we can presume that the data suffers from extreme right skewness, as was demonstrated in the various boxplots, histograms, and comparisons between measures of central tendency. Based on these findings, it is better to consider median as the best measure of central tendency, on account of the fact that it is not as heavily influenced by outliers. While outliers are widespread in the data, there do not appear to be any major problems with missing or unusual values. Although there were significantly fewer observations of it in the sample, joint applicants appeared to be better loan candidates, given their higher median debt-to-income, median accepted loan amount, and median credit limit.