A Study of UCI Credit Approval

Lawrence Reed

MSDS 696 Regis University

Professor Doug Hart

**Introduction**

In beginning to work with the UCI Credit Approval dataset. The first thing that comes to mind is that the data appears to be cryptic in nature. Indeed, it is because there are no column names and all the data has been masked to make it difficult to discern the people whose actions are represented in the dataset. At first, and for some time working with the dataset was quite difficult due to the amount of work that went into masking the original data. Upon reading the white paper that accompanies it, I was expecting to be given, at least, the column names, but no such luck for me.

**Commencement**

In starting, I examined the dataset as it was.

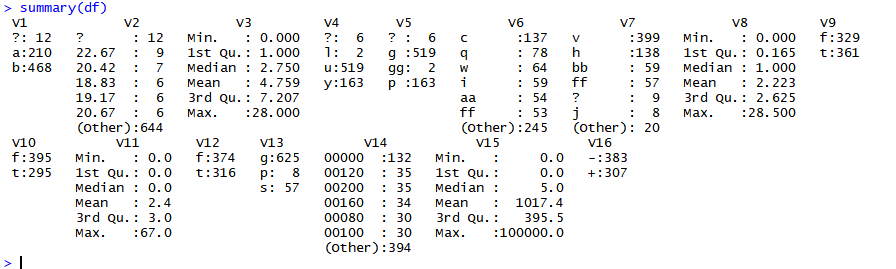


Figure 1

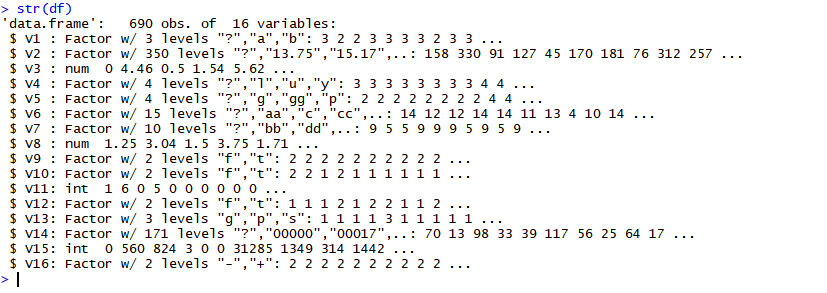


Figure 2

In examining the summary and the structure, it is noticeable that there are several different columns that have missing data. But the surprise waiting here is that there are factor levels labeled as question marks. This is my first encounter with data that is intentionally made difficult to read. The fake factor level tells that the confusion is absolutely by design. It might be easy to become over whelmed with this, but the best thing to do is to remain calm and recall the strategies taught to handle data that is not at all succinct.

The cases with missing values represented about 5% of the total of all records. This is small enough to consider simply removing the incomplete records. I decided to remove the records, for the sake of brevity. Below, I used “mice” and “VIM” to visualize the missing data and to quantify it. It is a set of two pictures that demonstrate a before and an after.

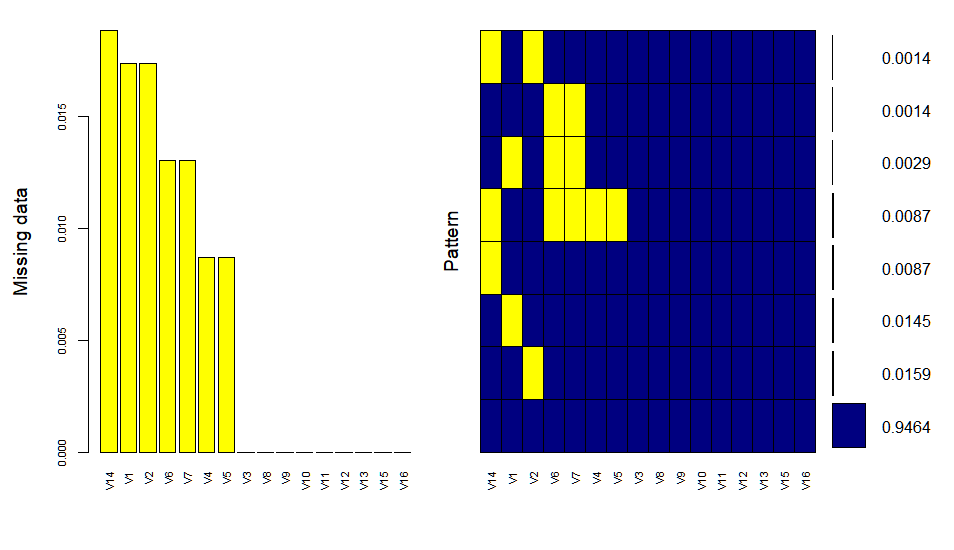


Figure 3

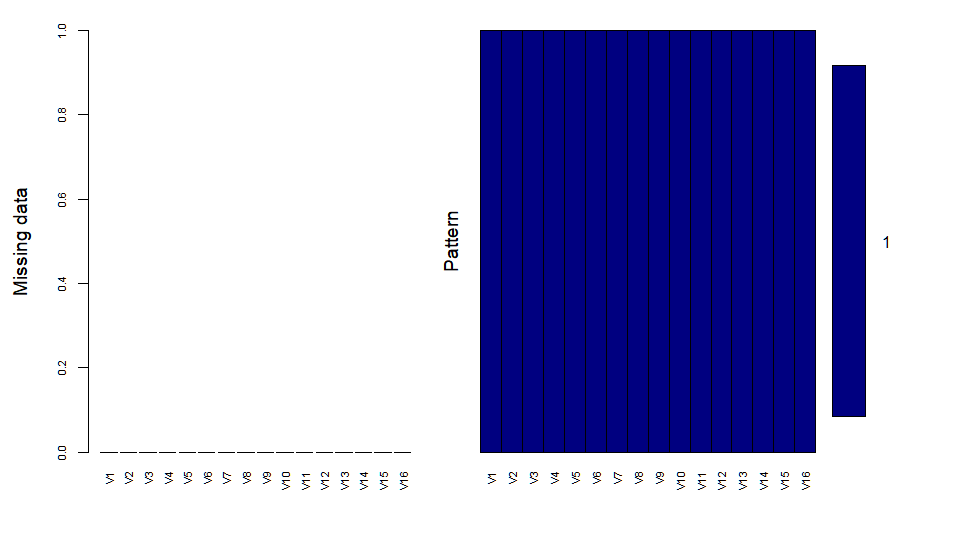


Figure 4

To handle the problem with missing data, I updated the question marks to NA, then I performed and NA omit, which resulted in Figure 3. The data frame was reduced from 690 obsevations to 653.

Column 1 had a dummy factor level, but the data only had a two different characters, I determined that the extra factor needed to be removed, so I decided to remove all factor leveling. Afterwards, I rebuilt the factor, using the rules that were given me, by past professors. I repeated that process for all other columns that followed suit to ensure that the factoring is proper and concise for each column.

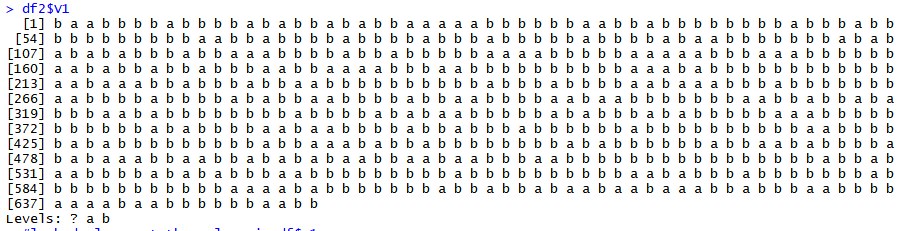


Figure 5

Column 2 began as floats with 2 decimal points. To add to the confusion there were too many factor levels programmed for this column. After clearing it out, in checking its summary, the minimum was about 13 and the maximum was about 77. This appeared to be age, but it was confusing due to the decimals. So, I rounded the column to the ones. The ages seemed to be a bit to young at thirteen, so I researched Japan and found that there are no set laws that define the boundary between child and adult. While the official age is 18, there are credit grantors who will allow people as young as 13 to acquire credit.

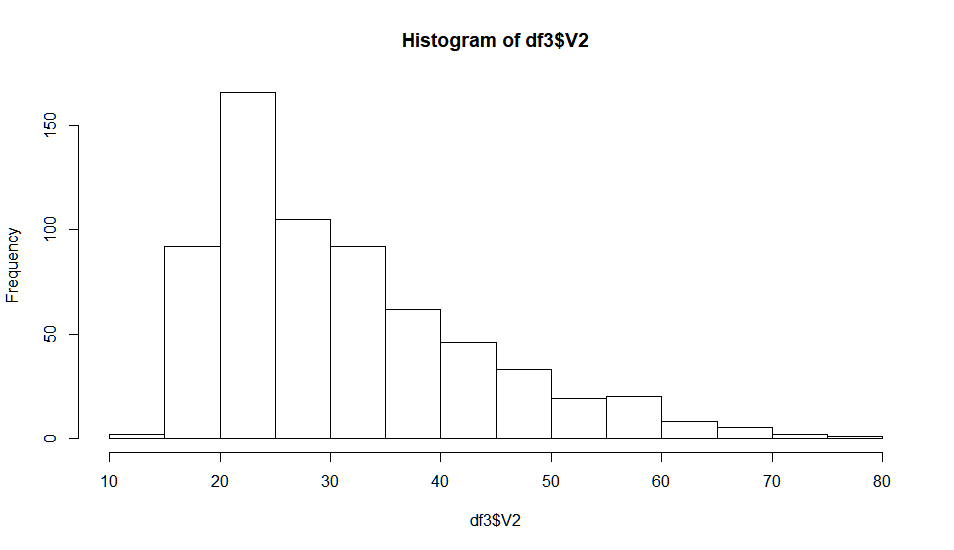


Figure 6

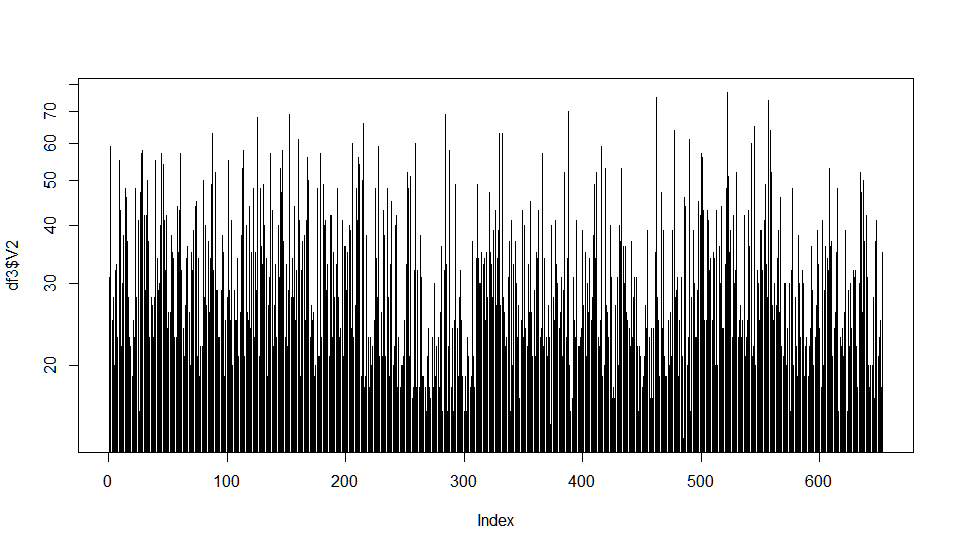


Figure 7

In examination of the histogram of the ages, we can observe that the distribution is slightly skewed to the right, but the log of the column gives a sense that the column is a bit more balanced than the histogram shows. Giving more attention to the age column, a histogram and the log view of the column shows that of the cases with minors the peak is with 15.5 to 16 years olds.

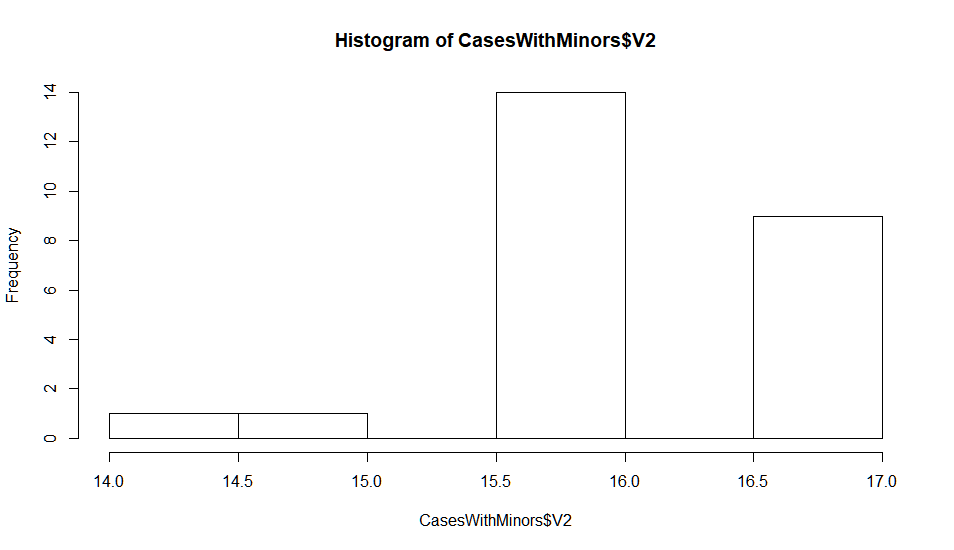


Figure 8

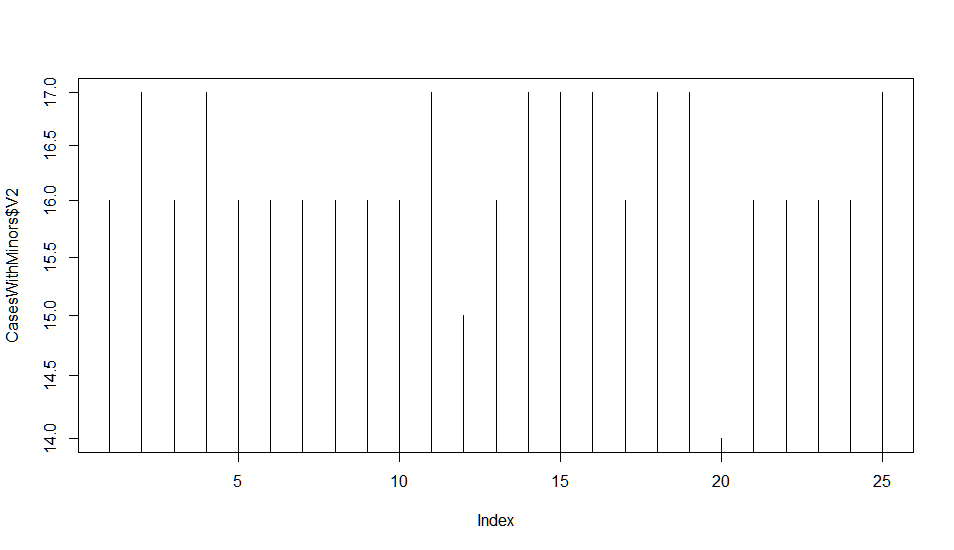


Figure 9

Reaching variable 3, it is almost second nature to recognize that the column is coded to resemble thousands, but with a period where the comma should be. This makes it apparent that one only needs to multiply by 1000 to get the column in proper form. This is even more apparent when it is done, creating a column that one could not imagine in any other state, as shown below in Figure 10 and 11.

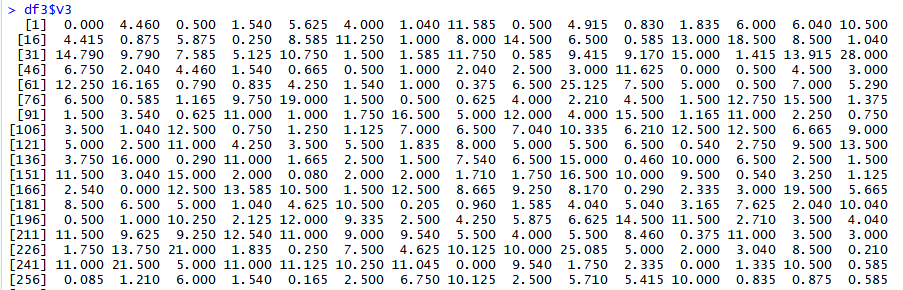


Figure 10

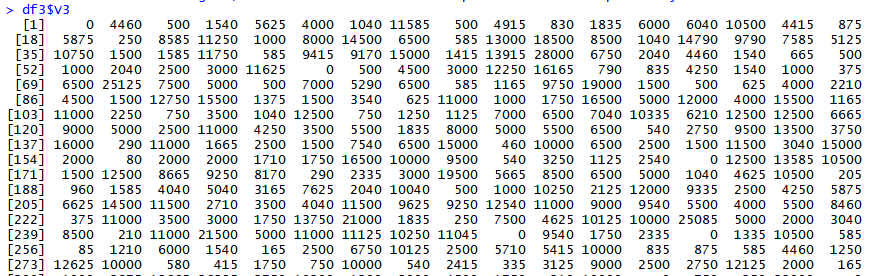


Figure 11

A histogram of the new values reveals quite a few zero values and a skew to the right, but values are evenly balanced highs, mids and lows, as demonstrated in Figures 12 and 13.

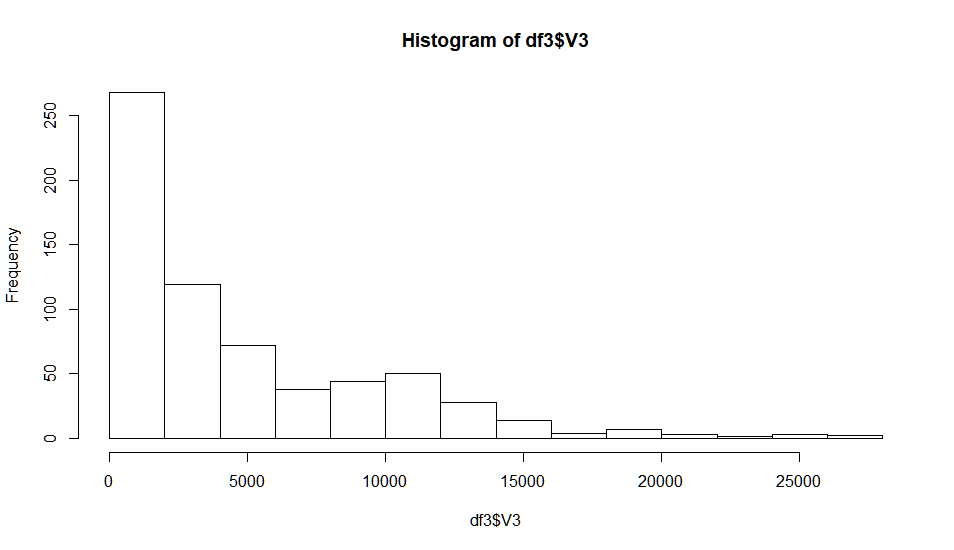


Figure 12

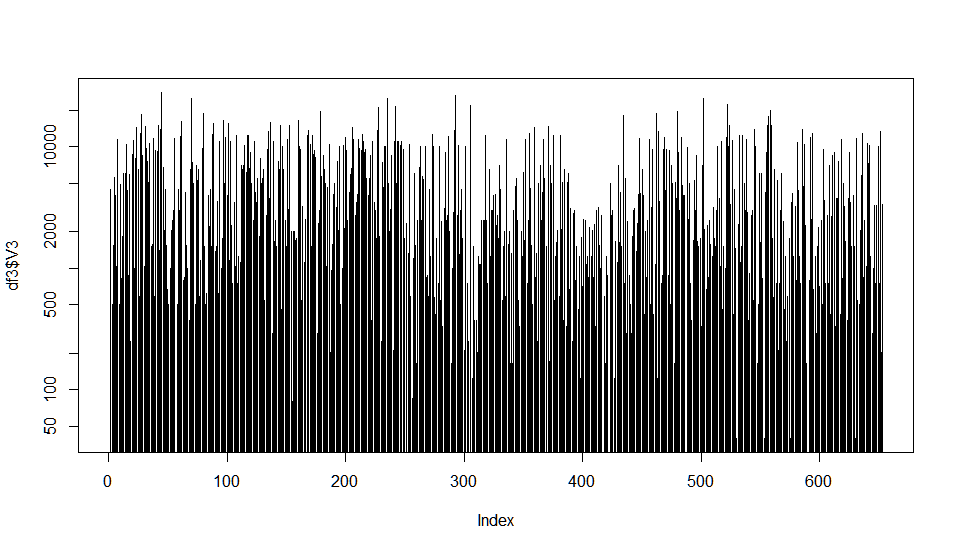


Figure 13

For the handling of variable 4, I used the same principles that were used for variable 1. That made it easy to refactor the column correctly. These principles were further expanded upon with variables 5 through 7, due to the number of unique factor levels presented increasing by quite a bit. Variable 8 was a brief return to the principles that were used to decipher variable 3. The histogram of variable 8 shows a larger representation and presence of 0. Further research reveals that most people in Japan have not been introduced to the credit system, despite the age of introduction being much earlier than 18 years of age and Japan being a known industrialized nation.

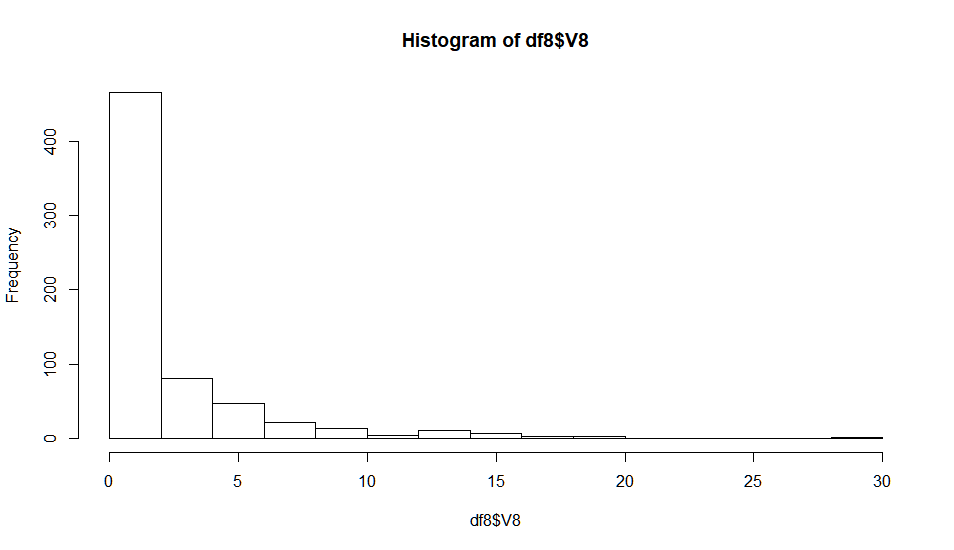


Figure 14

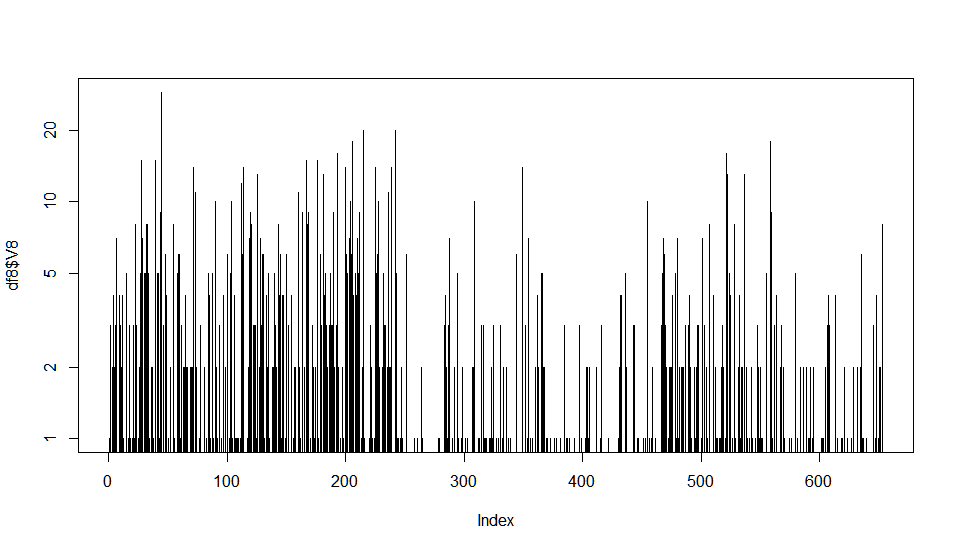
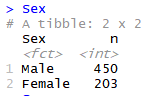


Figure 15

Variables 9 and 10 were properly created Boolean columns, with no issues. Variable 11 marks a return of the prevalent 0, which cannot be ruled out because of most people not being introduced to the official industrialization of Japan, to included credit granting. Variable 12 is again Boolean and properly formatted. Variable 13 is a factor with three levels. Variable 14 is a factor with several levels, but it should not be a factor. I made it numeric, then turned it to character, because it should not be used for calculations. Variable 15 shows the prevalence of 0, which is now to be expected. Variable 16 was coded as plus and minus for approved and denied. I fixed this by dropping the factor levels then recreating them with the proper wording.

**Information Extrication**

Examination of the data presented reveals that men are more than twice as likely to apply for credit in Japan, with male applications at 450 and female applications at 203.



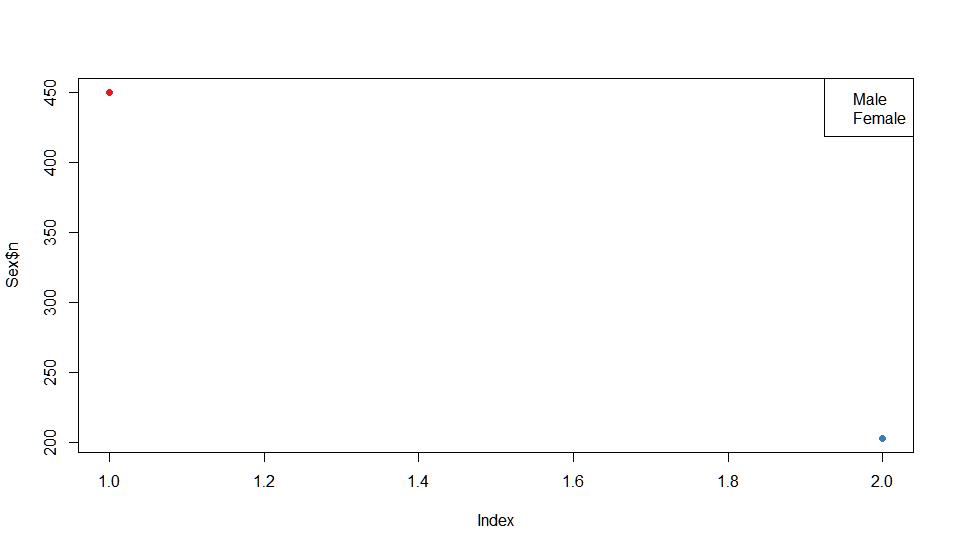
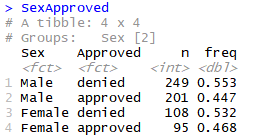
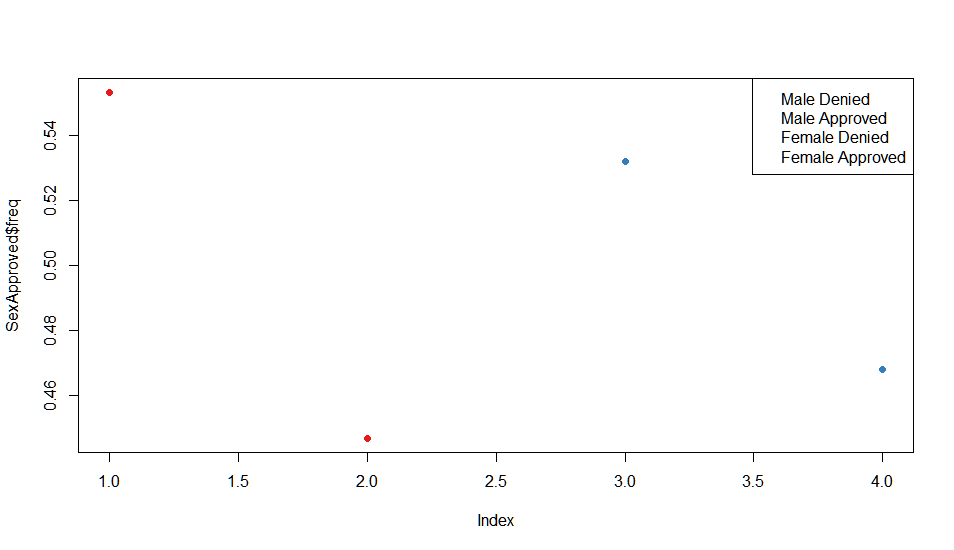


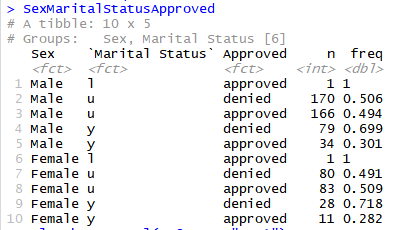
Figure 16

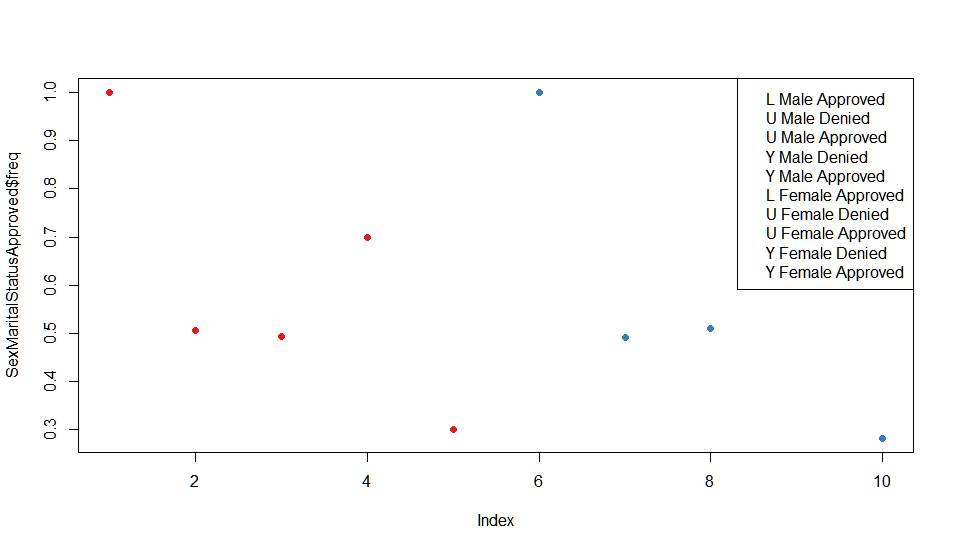
It is also demonstrated that when women do apply for credit, they are 2% more likely than men to be approved.



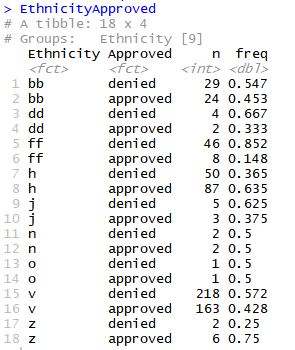


In examination of sex and marital status, there is a large difference in approval rates for both men and women. A married man is 19% more likely to be approved than an unmarried man, while a married woman is 22% more likely to be approved than if she were not married.



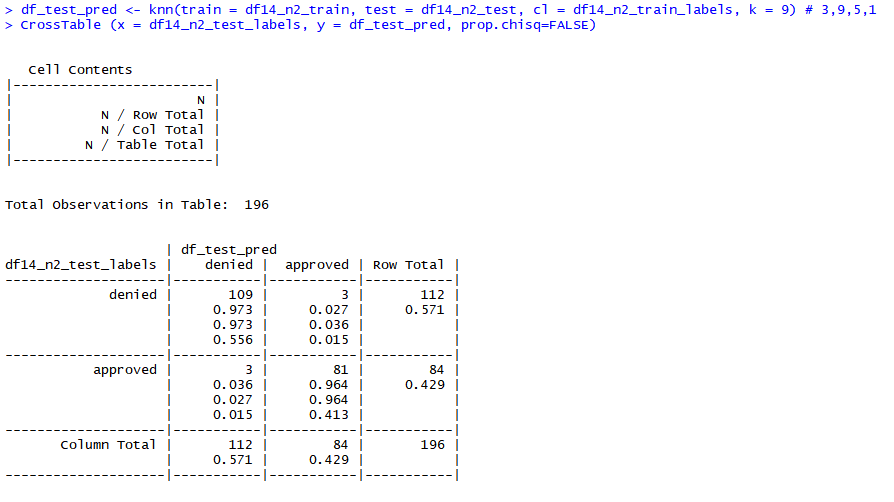


It is also shown that ethnic group “h” enjoys a credit approval rate of 63%, while ethnic group “ff ” only enjoys a 15% approval rate. Highly unusual, ethnic groups “n” and “o” have a perfectly balanced approval and denial rate.

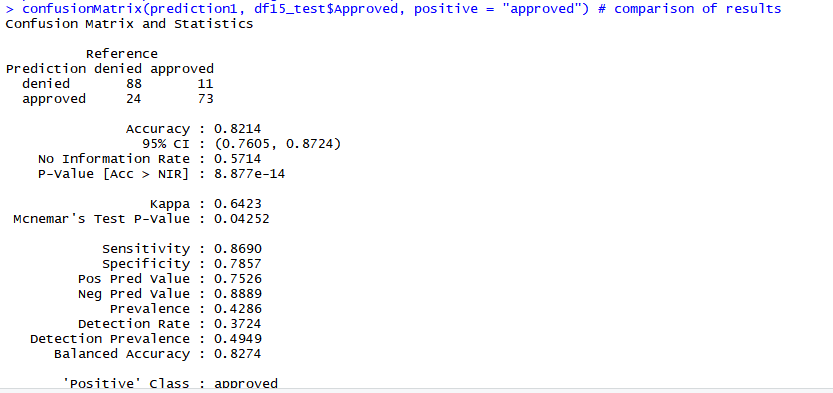


**Model Performance**

To model the dataset, I first chose knn. I was able to achieve a 96% success rate on classifying approval and denial, as shown below.



My second choice was to use bagging to model the dataset, however I was only able to achieve an 82% correct decision using this method.



**Conclusion**

There will be times when the data that we are given will appear very difficult to work with, not to mention trying to extract insights from it or model it. If we keep the principles that were given us and let those principles guide our actions, then we can and will succeed in handling even the most difficult datasets. Here we have managed to get data from and model a most complicated dataset, which would not have been possible without proper technique.

References

https://www.huffingtonpost.com/creditcom/how-do-credit-scores-work\_b\_1723362.html