HBAT Data Examination (July 2019)

Jason Huggy, Student, Lewis University

I. INTRODUCTION

The purpose of this study was to examine data from the HBAT dataset utilized in *Multivariate Data Analysis* [1]. The data set includes 99 lines of unlabeled data, and 23 variables. A variety of methods are utilized to visualize and analyze the data, ranging from univariate analysis to multivariate analysis. This study uses SAS Enterprise Guide (EG) to use tools such as distribution analysis, scatter plot matrices, correlation, boxplots, and ANOVA analytics. The hope is to determine any significance in the data presented.

II. Methodology

To conduct this study, SAS Enterprise Guide is used to produce all visualizations and data analytics. To begin, the HBAT data is imported into EG and verified to be free of missing values or errors. A quick examination of the data shows that the first five columns of data, as well as the last, are nonmetric features. The rest of the features are all metric.

A. SHAPE OF THE *DISTRIBUTION*

To begin the first part of this study, a distribution analysis is conducted on variable X_6. After loading the HBAT data into EG, this is done by choosing *Tasks, Describe* and then *Distribution Analysis.* The distribution analysis is specified to create a histogram to display a normal distribution based on the data in column X_6. The resulting report displays a table and graph displaying the distribution of the data in X_6.

B. RELATIONSHIP BETWEEN VARIABLES

In the second part of the study, bivariate analysis is conducted using five different variables. This is done first using a scatter plot matrix, which creates a five by five matrix of scatter plots, with the histogram for each variable running diagonally down the middle. The scatter plot matrix is created by selecting *Tasks, Graph,* and then *Scatter Plot Matrix*. X_6, X_7, X_8, X_{12}, and X_{13} are selected as the variables used, and the option to display histograms down the middle of the matrix is selected. In addition, a correlation study is created on these variables, which is done by selecting *Tasks, Multivariate,* and then *Correlations.*

C. GROUP DIFFERENCES

In the last part of this study, two pairs of variables are analyzed using box plots and ANOVA analysis. The creation of the box plots is done by choosing *Tasks, Graph,* and then *Box plot.* Both box plots are created with variable X_1 on the X-axis, and then with X_6 and then X_7 on the Y-axis. Two tables are created to display the ANOVA analysis of the same variables using *Tasks, ANOVA,* and *One-Way ANOVA.*

III. Results and Discussion

A. SHAPE OF THE *DISTRIBUTION*

After plotting the histogram for X_6, see Figure 1 below, it is seen that the graph is not skewed to either side, but the middle of the graph drops below the normal curve. Ideally, we would prefer that the middle of the graph at least mostly matches up with the normal curve to be used in multivariate analysis. To make X_6 more useable, it is advised that a transformation is used to deal with the slump in the middle of the data.

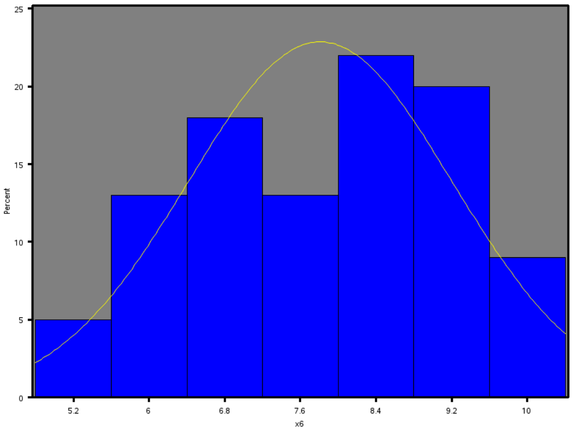


Fig. 1. Distribution Analysis of X_6.

B. RELATIONSHIP BETWEEN VARIABLES

In the next part of the study five variables are plotted in a scatter plot matrix and the correlations between each variable pair is analyzed. Figure 2 shows the scatter plot matrix and Table 1 shows the correlation results. From the scatter plots, it is obvious that there is a positive correlation between X_7 and X_{12}. Looking at Table 1, the correlation is 0.79. Since it is so close to 1, it signifies that the two variables are strongly, positively correlated. Meaning, as one goes higher so does the other. There is also a slight negative correlation between variable X_{13} and X_6, but the correlation is only -0.40 as seen in table 1. This signifies that there is some correlation, but not the strongest. Since it is negative that means as one goes lower the other goes lower. It can be argued that there may be slight correlation between a few of the other variable pairs, but they are very slight.

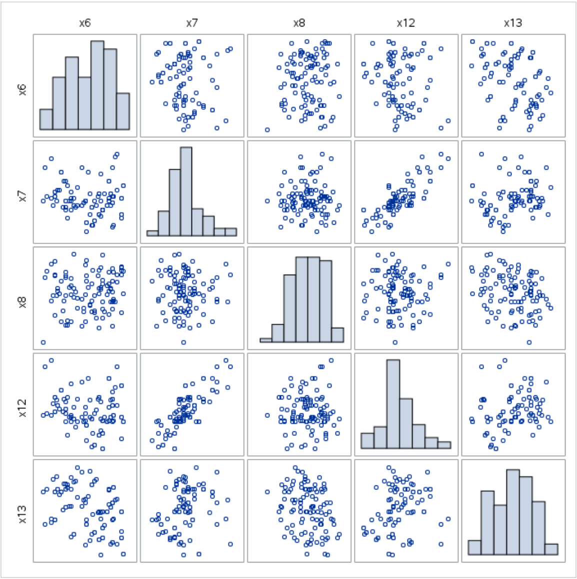
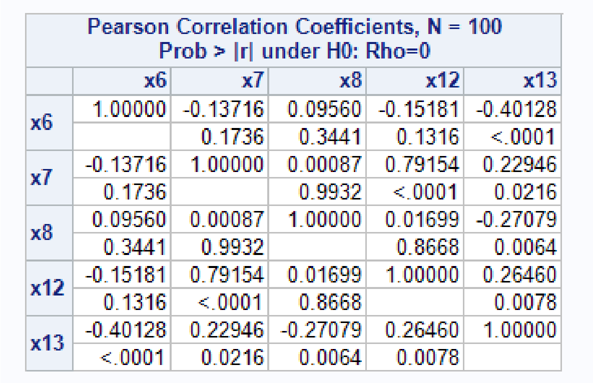


Fig. 2. Scatter plot matrix of variables 6, 7, 8, 12, and 13.



Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	x6	x7	x8	x12	x13
x6	1.00000	-0.13716 0.1736	0.09560 0.3441	-0.15181 0.1316	-0.40128 <.0001
x7	-0.13716 0.1736	1.00000	0.00087 0.9932	0.79154 <.0001	0.22946 0.0216
x8	0.09560 0.3441	0.00087 0.9932	1.00000	0.01699 0.8668	-0.27079 0.0064
x12	-0.15181 0.1316	0.79154 <.0001	0.01699 0.8668	1.00000	0.26460 0.0078
x13	-0.40128 <.0001	0.22946 0.0216	-0.27079 0.0064	0.26460 0.0078	1.00000

Table. 2. Correlation between variables 6, 7, 8, 12, and 13.

C. GROUP DIFFERENCES

Finally, two pairs of variables are analyzed using box plots and ANOVA analysis. Figure 3 and 4 show the box plots for each pair. Looking at the two box plots there are obvious differences. Figure 3 shows significant variance between the three categories, while figure 4 hardly shows and significant variance between the three categories. This is supported by the ANOVA analysis data. The first pair, that shows the higher variance in the box plots, has an ANOVA of 83, while the second pair has an ANOVA of only 0.86. This is a significant difference and shows why there is so much variance in the box plots between X_6 and X_1. This variance between categories can help differentiate between the categories when it comes down to modeling the data.

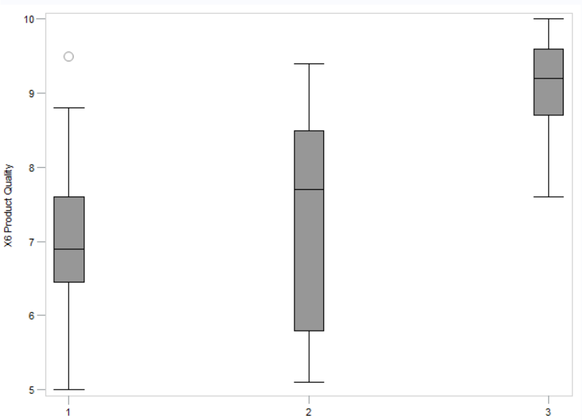


Fig. 3. Box plot between variable 1 and 6.

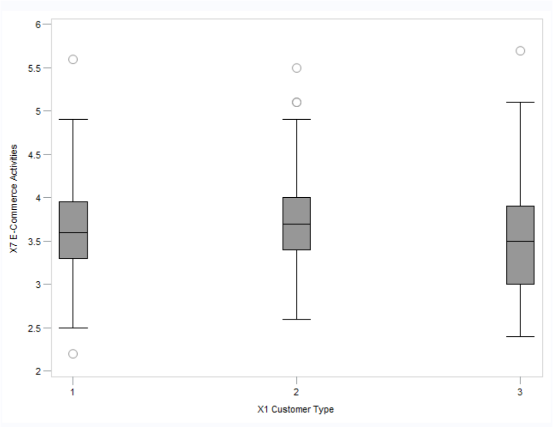


Fig. 4. Box plot between variable 1 and 7.

IV. CONCLUSION

This study serves the purpose of showing the benefits from examining data before trying to do anything to it. It is important to ensure that your data meets any requirements needed to use the specified multivariate analysis technique, and performing simple analysis on your data can help ensure those criteria are met. Simply visualizing your data can have a significant effect on the results of your study. It is not something that should ever be skipped.

REFERENCES

[1] J. Hair, W. Black, B. Babin, R. Anderson, *Multivariate Data Analysis: Pearson New International Edition*. 2013.