WEEK 2 ASSIGNMENT

Concepts of Statistics 2 – DATA-51200 | Spring 2 2020

Christina Morgenstern

---

1. Missing Data

Before diving into the analysis of a data set, we must get an understanding of the data as well as the putative relationships between individual variables. Furthermore, we need to make sure that the data is valid and accurate and fulfills the requirements for multivariate analysis. One of the first tasks involved in this data examination endeavor is checking for missing data in the data set which are values of variables that are not available to the analysis. Apart from determining the extend of missing data, we must also investigate if there is an underlying pattern discernable i.e. if there are relationships between individual missing values. Diagnosing this missing data process is crucial for subsequent analysis because missing data can have unintended effects on the results such as inaccurate or biased outcomes.

Hair *et. al.,* describe a four-step procedure for examining the data set for missing values. In the first step the type of missing data is determined which can be of *ignorable* or *non-ignorable* nature. Ignorable missing data confers to values that are expected to be missing and are part of the design process. Examples are censored data or missing data that resulted from the data collection process as well as sample data that doesn´t take into account the full population. Non-ignorable missing data is further structured into either known or unknown sources. Whereas the latter results from responses that were not given by the survey participants, the former is due to a procedural error. In step two of the missing data diagnosing process, the extend of missing data is defined. Tabulating data helps with assessing the level of missingness, whereby the percentage of missing values for each variable is determined. A general rule of thumb states that 10% or less of missing data in a variable is acceptable without adopting any measures. If the level of missing data exceeds this threshold, actions must be taken to diagnose the level of randomness in the missing data. This is step three in the missing data diagnosis sequence and deals with the feature of randomness and the degree of association between individual variables. According to these two characteristics, three types of data sets can be defined: *Missing data at random (MAR), Missing completely at random (MCAR)* and *Not missing at random (MNAR).* Whereas in MAR the relationship to other variables somehow can be made out, MCAR data shows a higher level of randomness and in MNAR data, distinct non-random patterns are discernable. Several diagnostic tests, such as the *t-Test of missingness* or the *Little´s MCAR* test aid in finding these patterns. Selecting an imputation method, which estimates the missing value based on other variables values, is performed in step four. It should be noted that nonmetric variables are not subject to imputation. Depending on whether the data has been identified as being MCAR or MAR in the previous step, the researcher can try different imputation methods. Data with MCAR missing data patterns can choose from a variety of approaches such as imputation using valid data, imputation using known replacement values or imputation through calculating replacement values. Very often mean imputation is used whereby the mean value, which was derived from all other values in the sample, is used. If the missing data process has been assigned a MAR pattern, only the modeling approach can be applied. Maximum Likelihood or Multiple Imputation models are popular choices.

In general, there is no single best remedy to deal with missing data in a data set. The researcher needs to apply several imputation methods and quantify the results.

References:

[1] Multivariate Data Analysis by Joseph F. Hair Jr, William C. Black, Barry J. Babin and Rolphe E. Anderson, Pearson, 8th edition, 2019

For the following programming exercises, I am using SAS Studio from SAS University Edition running on my VirtualBox VM. Please note, that the SAS Studio software is running as a German version as my default browser settings are German. Since I am not familiar with SAS code, I will be using the SAS Visual Programming Interface.

2. Working with the HBAT dataset.

Importing the HBAT(3).xlsx data set into SAS Studio. For the data to be visible to the SAS software, it needs to be moved to the folder that was specified in the setup. The following code was generated by SAS upon import of the HBAT data set.

```
OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
 72
 73         /* Generierter Code (IMPORT) */
 74         /* Quelldatei: HBAT(3).xls */
 75         /* Quellpfad: /folders/myfolders */
 76         /* Code generiert am: 26.03.20 22:15 */
 77
 78         %web_drop_table(WORK.IMPORT);
 79
 80
 81         FILENAME REFFILE '/folders/myfolders/HBAT(3).xls';
 82
 83         PROC IMPORT DATAFILE=REFFILE
 84         DBMS=XLS
 85         OUT=WORK.IMPORT;
 86         GETNAMES=YES;
 87         RUN;

 NOTE: The import data set has 100 observations and 24 variables.
 NOTE: WORK.IMPORT data set was successfully created.
 NOTE:  Verwendet wurde: PROZEDUR IMPORT - (Gesamtverarbeitungszeit):
       real time           0.09 seconds
       cpu time            0.02 seconds
```

The data set comprises of 24 variables and 100 observations (Table 1).

Table 1. Summary of data set.

Dateiname	WORK.IMPORT	**Beobachtungen**	100
Membertyp	DATA	**Variablen**	24
Engine	V9	**Indizes**	0
Erstellt	27.03.2020 18:18:13	**Beobachtungslänge**	192
Zuletzt geändert	27.03.2020 18:18:13	**Gelöschte Beobachtungen**	0
Schutz		**Komprimiert**	NEIN
Dateityp		**Sortiert**	NEIN
Etikett			
Datendarstellung	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Codierung	utf-8 Unicode (UTF-8)		

The first step in the data exploration is to check for missing values. In the navigation pane of SAS Studio there is button with which this can be achieved. I used this feature and selected all variables, to see if there are any missing data in the HBAT data set. As summarized in Table 2, there are no missing values in the data set.

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend

x13	x14	x15	x16	x17	x18	x19	x20	x21	x22	x23	Häufigkeit	Prozent
Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	Nicht fehlend	100	100

Table 2. Summary of missing data evaluation result. None of the 24 variables contains missing values as denoted by the statement „Nicht fehlend“, which means „nothing missing“.

The code that was generated to search for missing data is as follows:

```
ods noproctitle;

proc format;
      value _nmissprint low-high="Nicht fehlend";
run;

proc freq data=WORK.IMPORT;
      title3 "Häufigkeit für fehlende Daten";
      title4 h=2 "Legende: ., A, B, etc = Fehlend";
      format id x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19
            x20 x21 x22 x23 _nmissprint.;
      tables id x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19
            x20 x21 x22 x23 / missing nocum;
run;

proc freq data=WORK.IMPORT noprint;
      table id * x1 * x2 * x3 * x4 * x5 * x6 * x7 * x8 * x9 * x10 * x11 * x12 * x13
            * x14 * x15 * x16 * x17 * x18 * x19 * x20 * x21 * x22 * x23 / missing
            out=Work._MissingData_;
      format id x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19
            x20 x21 x22 x23 _nmissprint.;
run;

proc print data=Work._MissingData_ noobs label;
      title3 "Variablenübergreifende Muster für fehlende Daten";
      title4 h=2 "Legende: ., A, B, etc = Fehlend";
      format id x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19
            x20 x21 x22 x23 _nmissprint.;
      label count="Häufigkeit" percent="Prozent";
run;

title3;

/* Bereinigung */
proc delete data=Work._MissingData_;
run
```

a) Histogram of X6, X7, X16 and X17 and associated normal curve.

In the Navigation Pane on the left-hand side, go to Tasks & Utilities and click Diagrams, select Histogram. Add data to the program through selecting the previously imported HBAT file. I started with generating the histogram for the Product Quality variable X6. The Product Quality variable is a metric variable that captures the perceived level of quality of HBAT´s paper products and can have values from 0 to 10. The histogram displays the count frequency for each bin with the number of bins set to default. A normal curve was added to the diagram.

The following SAS code was generated when creating the histogram for X6.

```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
	title height=14pt "Histogram of Product Quality Variable X6";
	histogram x6 / scale=count fillattrs=(color=CXcad5e5);
	density x6;
	xaxis label="X6 Product Quality";
	yaxis grid label="Count";
run;

ods graphics / reset;
title;
```

Through exchanging the variable in the data set up, histograms for the other variables X7, X16 and X17 were generated. The metric variable X7 summarizes the customer satisfaction of E-Commerce Activities, where the HBAT´s website and its user-friendliness are rated from 0 to 10. The metric variable X16 confers the customer perception of correct and efficient Ordering and Billing activities and takes values between 0 and 10. Price Flexibility is denoted by the metric variable X17 which represents the perceived willingness to negotiate prices on purchases of paper products from 0 to 10. In all cases, 0 is labelled as poor and 10 as excellent.
The resulting histograms for the four variables can be seen in Figure 1.

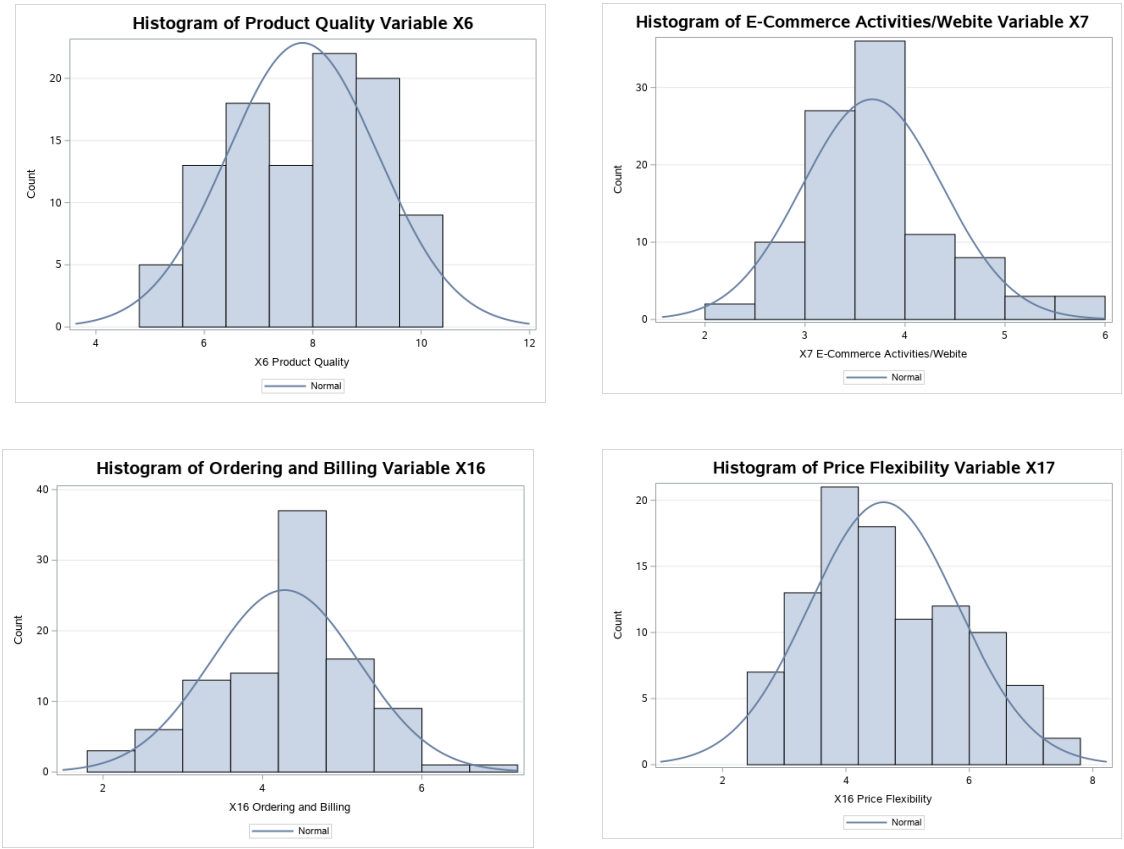

Figure 1. Histograms of variables X6, X7, X16 and X17 with superimposed normal curves.

b) For part a, which variables satisfy the normality test. Examine this visually and using skewness and kurtosis tests.

Skewness and Kurtosis tests can be performed using the Distribution Analysis feature from the Statistics collection within Tasks. Within the Options section, Add insert statistics and select Number of observations, Skewness and Kurtosis. A check for Normality is performed through selecting the Histogram and Goodness-of-fit-tests as well as the display of T-Test Statistic and the P-Value.
Again, I run these options on variable X6 and reuse the code to generate the results for the other variables (Figure 2).

```
ds noproctitle;
ods graphics / imagemap=on;

/* Daten untersuchen */
proc univariate data=WORK.IMPORT;
	ods select Histogram;
	var x6;
	histogram x6 / normal;
	inset n skewness kurtosis / position=ne;
run;

proc univariate data=WORK.IMPORT normal;
	ods select Histogram GoodnessOfFit;
	var x6;

	/* Auf Normalität prüfen */
	histogram x6 / normal(mu=est sigma=est);
	inset normaltest pnormal skewness kurtosis n / position=ne;
run;
```

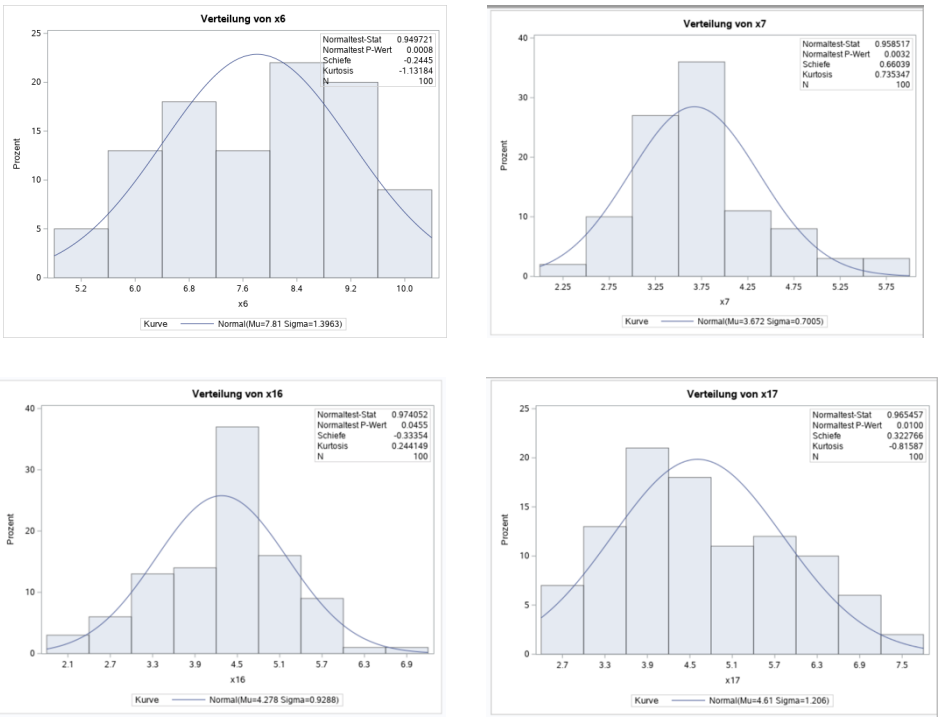

Figure 2. Distribution analysis of variables X6, X7, X16 and X17 with superimposed normal curves and test statistics displayed.

A visual examination of the plots leads to the assumption that only variable X16 has a somewhat normal distribution which is characterized by following the bell-shaped curve around the mean. The other three

variables, X6, X7 and X17 do not show a normal distribution as jugged visually and are skewed to the left for X6 and right for X7 and 17, respectively.

Normality can also be assessed by carrying out statistical tests using the Skewness and Kurtosis values. If the z-values for Skewness and Kurtosis lie between $\pm$ 2.58 for a significance level of 0.01 or $\pm$ 1.96 for a significance level of 0.05 then the distribution can be assumed to be normal.

In the following, I am calculating the z-values for Skewness for each variable. The values for Skewness are taken from the histograms in Figure 2 (termed "Schiefe").

$$Z_{Skewness\ X6} = \frac{Skewness\ value}{\sqrt{\frac{6}{N}}} = \frac{-0.2445}{\sqrt{\frac{6}{100}}} \cong -0.998$$

$$Z_{Skewness\ X7} = \frac{Skewness\ value}{\sqrt{\frac{6}{N}}} = \frac{0.66039}{\sqrt{\frac{6}{100}}} \cong 2.696$$

$$Z_{Skewness\ X16} = \frac{Skewness\ value}{\sqrt{\frac{6}{N}}} = \frac{-0.33354}{\sqrt{\frac{6}{100}}} \cong -1.362$$

$$Z_{Skewness\ X17} = \frac{Skewness\ value}{\sqrt{\frac{6}{N}}} = \frac{0.322766}{\sqrt{\frac{6}{100}}} \cong 1.312$$

From this analysis we can see that X7 is not normal at both significance levels 0.01 and 0.05 because the value of 2.696 is larger than 2.58 and 1.96.
The other three variables, X6, X16 and X17 fulfill the criteria of smaller values and are normal at both significance levels because the values are smaller than 2.58 and 1.96.
Thus, it is not always evident from visual examining the graph if the distribution follows a normal distribution.

In the following, I am calculating the z-values for Kurtosis for each variable. The values for Kurtosis are taken from the histograms in Figure 2.

$$Z_{Kurtosis\ X6} = \frac{Kurtosis\ value}{\sqrt{\frac{24}{N}}} = \frac{-1.13184}{\sqrt{\frac{24}{100}}} \cong -2.31$$

$$Z_{Kurtosis\ X7} = \frac{Kurtosis\ value}{\sqrt{\frac{24}{N}}} = \frac{0.735347}{\sqrt{\frac{24}{100}}} \cong 1.501$$

$$Z_{Kurtosis\ X16} = \frac{Kurtosis\ value}{\sqrt{\frac{24}{N}}} = \frac{0.244149}{\sqrt{\frac{24}{100}}} \cong 0.498$$

$$Z_{Kurtosis\ X17} = \frac{Kurtosis\ value}{\sqrt{\frac{24}{N}}} = \frac{-0.81587}{\sqrt{\frac{24}{100}}} \cong -1.665$$

Concerning the peakedness of the distributions, the z-values for all variables lie within the 0.01 significance level with values smaller than 2.58. Even variable X7 seems to follow a normal distribution now, albeit a skewed one.

c) The Normal Probability Plots for X6, X7, x16 and X17. Which variables are normal?

Normal probability plots are used to graphically visualize if data is distributed approximately normal. It plots the actual data values against data from a normal distribution. If the data is following a normal distribution, then the points should form an approximate straight line.

The following SAS code was generated when creating Normal Probability Plots for variable X17.

```
ods noproctitle;
ods graphics / imagemap=on;

proc univariate data=WORK.IMPORT normal;
    ods select ProbPlot;
    var x17;

    /* Auf Normalität prüfen */
    probplot x17 / normal(mu=est sigma=est);
    inset normaltest pnormal skewness kurtosis n / position=nw;
run
```

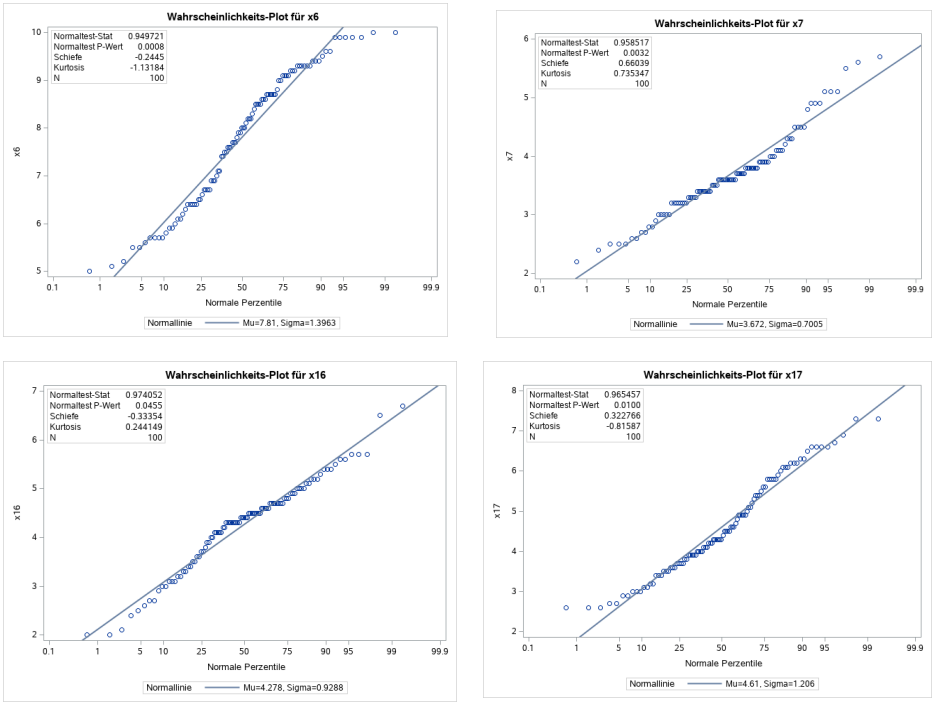

Figure 3. Normal Probability Plots for variables X6, X7, X16 and X17.

From Figure 3, we can see that the distributions for all four variables follow an approximate straight line. However, there are some deviations from the line discernable with the distribution falling both below and above the diagonal denoting a flatter or more peaked normal curve, respectively.

d) Plot the scatterplots for (X7 versus X19) and (X6 versus X19). Are there any outliers and how many? Use an ellipse representing the 95% confidence interval.

Bivariate relationships can be visualized using the scatterplot, a graph of data points based on two metric variables. Adding an ellipse representing the 95 percent confidence interval of a bivariate normal distribution can aid in visual outlier detection.

The scatterplots for X7 vs. X19 and X6 vs. X19 were generated using the following SAS code:

```
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=WORK.IMPORT pearson nosimple noprob
      plots=scatter(ellipse=prediction alpha=0.05);
    var x19;
    with x6;
run;
```

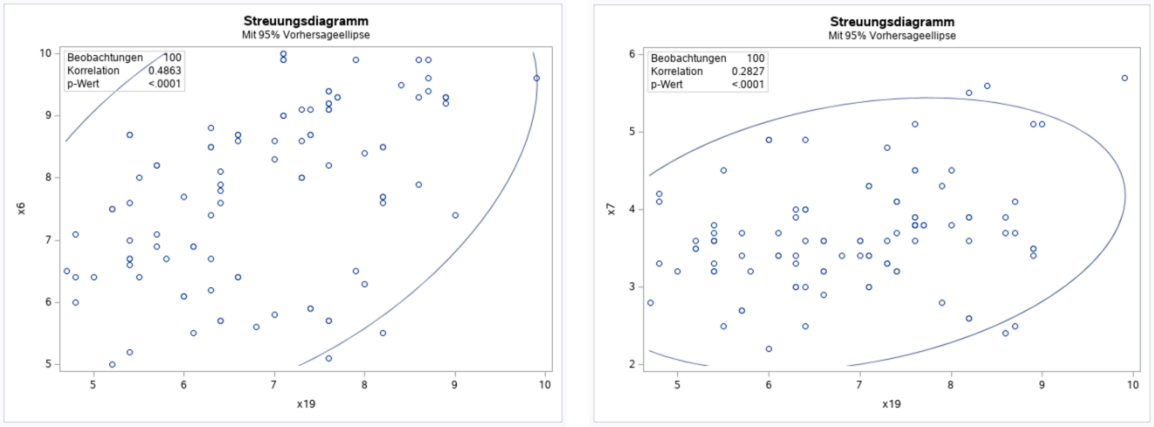

Figure 4. Scatterplots of variables X6 and X7 with X19.

From the scatterplot of X6 with X19 (Figure 4, left) we can see that two values fall outside the ellipse whereas in scatterplot of X7 with X19 (Figure 4, right) five values lie outside the range.

e) The bivariate profiling (correlation) of relationships between X6, X7, X8, X12. Which variables are correlated?

A scatterplot matrix allows for visual inspection of pair-wise relationship between variables.

The following SAS code was used to generate a scatterplot matrix of variables X6, X7, X8 and X12.

```
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=WORK.IMPORT pearson cov nosimple plots=matrix;
    var x6 x7 x8 x12;
run;
```

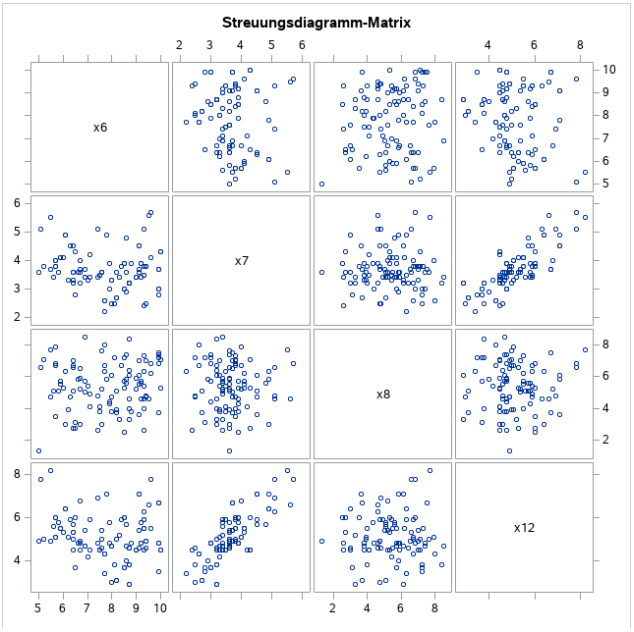

Figure 5. Scatterplot matrix of variables X6, X7, X8 and X12 for analysis of correlation.

From Figure 5 we can see that the highest correlation occurs between variables X7 and X12 because when plotted together pairwise, the dots follow an approximate straight line. The other variables don´t show any linear correlation with the selected variables as denoted by the widely dispersed pattern of points.

f) Boxplots of (X6 with X1) and (X7 with X1). Can we use Boxplots to detect outliers? Explain.

A boxplot is a diagram that allows for graphical representation of the data distribution of a metric variable for a category of non-metric variable. In SAS studio the Box-Plot feature can be found in the navigation pane under Tasks and Diagram. In Figure 6 the box plots from the variable X6 Product quality and variable X7 E-Commerce Activities/Website are displayed for each of three Customer Types (X1): less than one year (1), 1 to 5 years (2) and over five years (3).

SAS code to generate the box plots.

```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
	vbox x7 / category=x1;
	yaxis grid;
run;

ods graphics / reset;
```

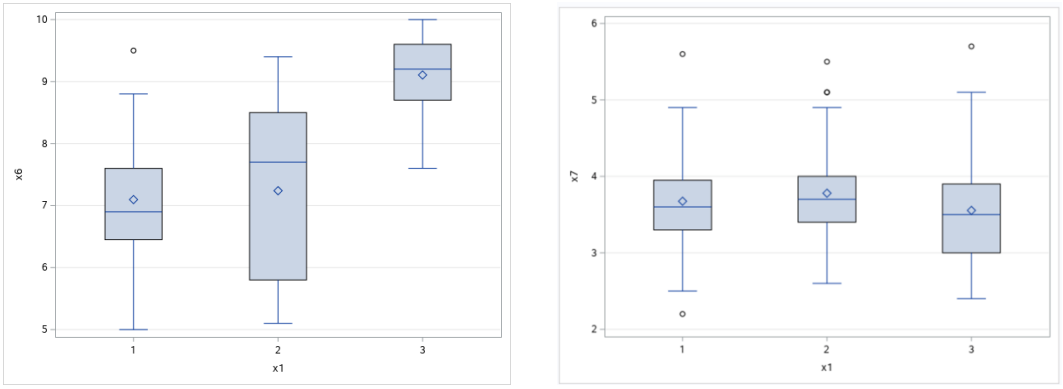

Figure 6. Boxplots of X6 (left) and X7 (right) with the three categories of X1. (1) less than one year, (2) 1-5 years, (3) over 5 years.

Boxplots allow for visual outlier detection. Outliers are represented as circle symbols outside the whiskers. There is one outlier in the boxplot of X6 with X1 and five outliers in the boxplot of X7 with X1. This confirms the outlier detection seen in the scatterplots in Figure 4.

Whereas there is no difference in the rating of the website and other E-commerce activities by customers that have been buying at HBAT for different time periods (Figure 6, right), there are differences in the perceived quality of HBAT´s products as noted by short- and long-term customers (Figure 6, left). While customers of categories 1 and 2 who are buying HBAT products for less than five years seem to rate the product quality equally, customers that have been clients for more than five years rate products to be of a higher-level quality.

3. For the following data, draw (by hand) the histograms for X1 and X2 and the scatterplot for X1 versus X2. Which variable(s) do think is(are) normal? Explain.

For the histograms, the values of the variables were arranged in ascending order. Minimum and maximum values were determined as well as the frequency of each value. For such a small data set, distributing the values into five bins, seemed appropriate. The bin width for each histogram was calculated using the difference of max and min values divided by the number of bins.

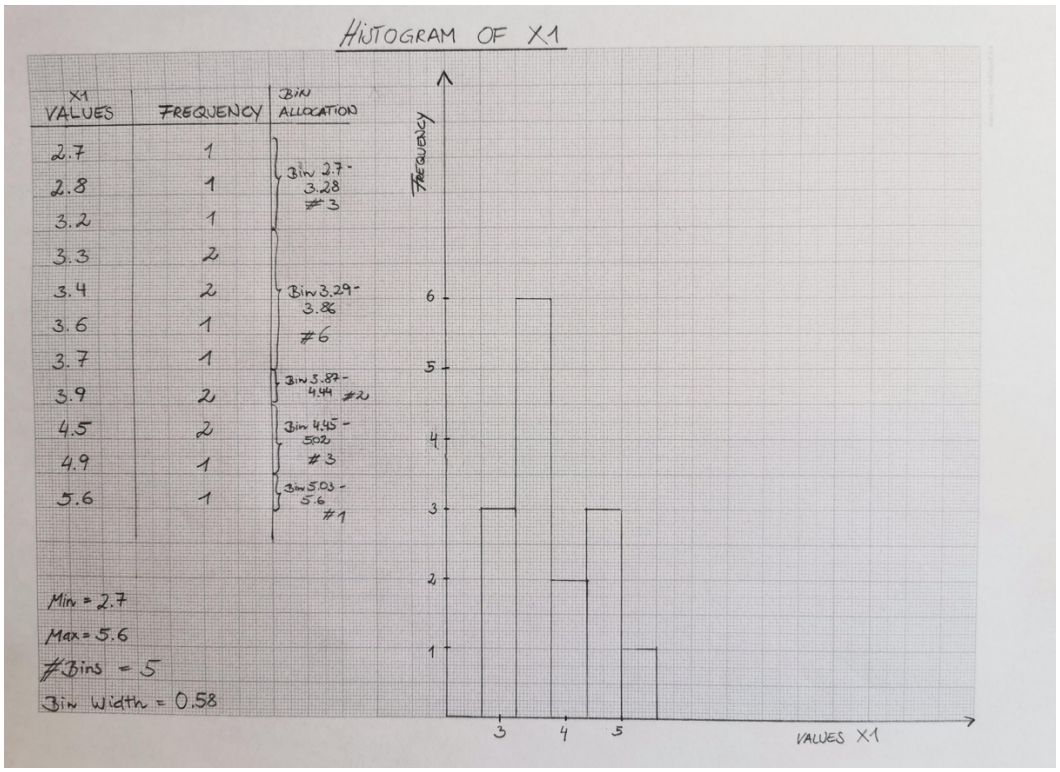

Figure 7. Histogram of variable X1.

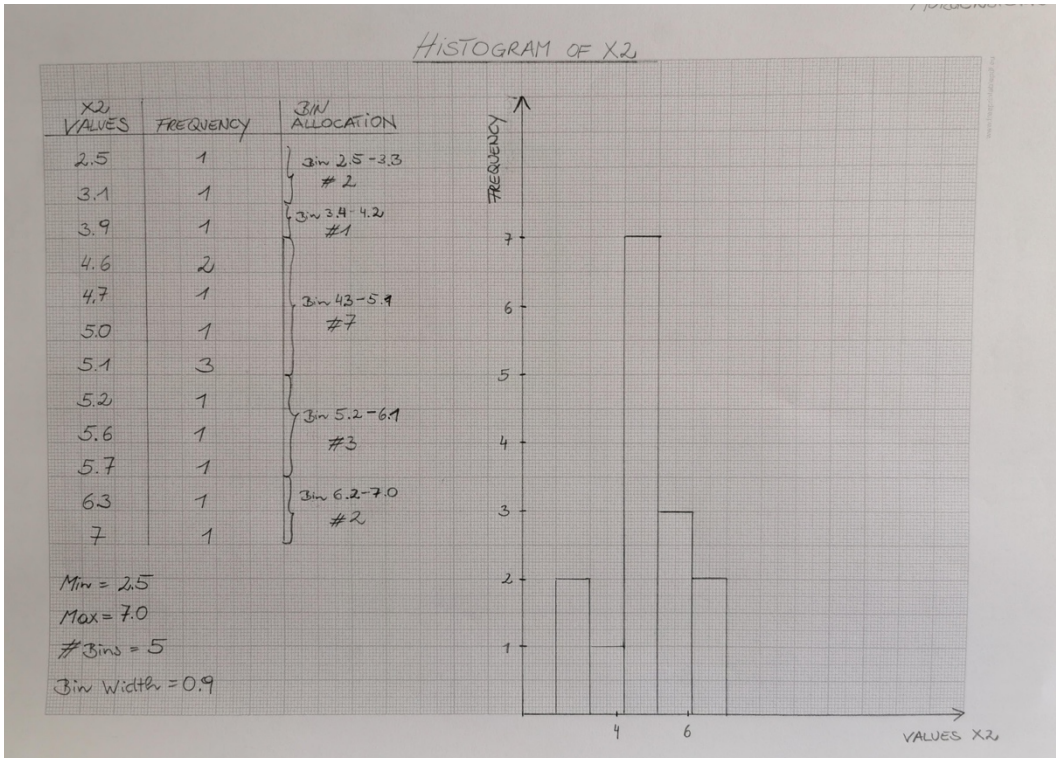

Figure 8. Histogram of variable X2.

We can assume normality of a distribution, if the graph is approximately bell-shaped and symmetric around the mean. Looking at the two graphs in Figure 7 and Figure 8, I would assume that variable X2 is more likely to be normal distributed because I can visually make out the bell-shaped distribution around the mean of 4.9. The histogram of variable X1 is more skewed to the right.

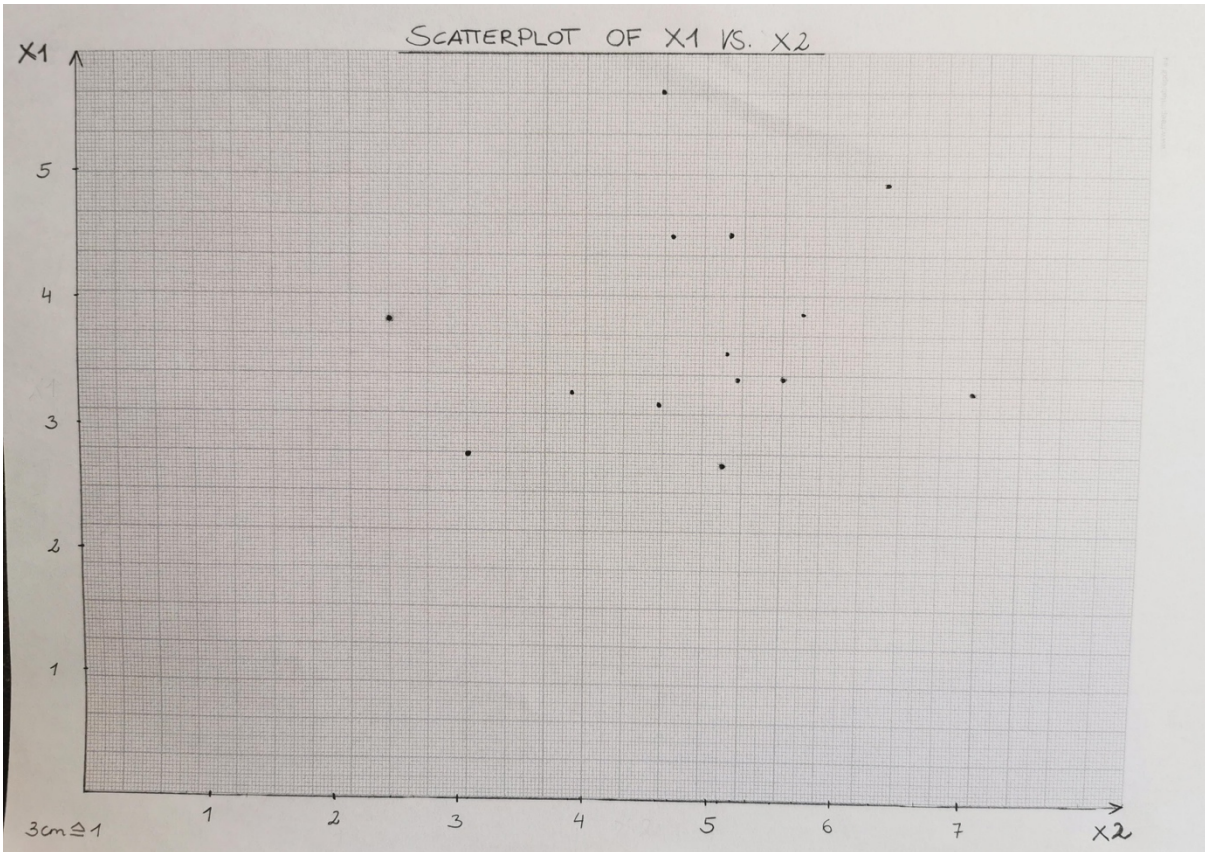

Figure 9. Scatterplot of variable X1 plotted against X2.

From the scatterplot, I can see that the variables X1 and X2 are not correlated (Figure 9), because I cannot make out a relationship such as a straight line neatly passing through. The points seem to be scattered randomly assuming no relationship between X1 and X2.