NMC268: Data Science Applications to Archaeology Assignment #3

125crossave

Part 1. ***Data Cleaning***

(a) *How many records are in the dataset?*

124 records

(b) *What would be an appropriate primary key for this dataset? Are any corrections necessary in order to use this field as the primary key?*

The ***Accession Number*** column would be an appropriate primary key, since it is a unique identifier for each artifact and thus corresponds with the typical purpose served by a primary key, which is to serve as a unique identifier for each row in a dataset. However, it is necessary to first clean entries with duplicate and/or conflicting data—for example, there were two accession numbers in one of the cells in this column, and one of these accession numbers (UPM 60-20-365a.2) was present in another column, which in turn contained duplicate values of artifacts UPM 60-20-365b.1 and UPM 60-20-365b.2, so I had to clean the data in order to separate these accession numbers into unique cells. I also trimmed all leading and trailing whitespace to make the data more readable.

(c) *Clean the provided dataset so that you are able to answer the questions below. Provide a full description of the data cleaning operations you have undertaken, and why you chose to perform them the way you did. What fields needed cleaning? What did you do to format them in a way that allows for the analysis below? What decisions did you make when cleaning the dataset?*

As previously mentioned, I edited the *Accession Number* column by using the Trim leading and trailing whitespace tool. I also edited the *Dimensions* column by using the Split into several columns feature to split the dimensions of each artifact into columns storing the *Length*, *Width*, and *Height*, with each value being measured in cm and stored as a number instead of as text; this was done to facilitate statistical analysis for the second part of the assignment, since it is much easier to create histograms if the required values are stored as numbers and not as text. To store the values as numbers, I used the Replace tool to remove all text within these columns, and the To Number transform to change the text to numbers. The same methods were applied to the *Warp diam.* and *Warp count per cm* columns, and to the *Building Location* columns—for the latter, I separated entries into *Building Location* and *Room* columns, and also made use of the Cluster and Edit tool in order to merge similar entries in these columns (e.g. different buildings were labelled as VIIE and VII East, so I had to merge them into the same category), as well as those in the *Description* column, together in order to improve the consistency of the dataset. For clustering, I used the *NearestNeighbour* method because doing so helped cluster the data points I wanted to merge the quickest.

For the *Weft count per cm* column, two entries were found with values of 0.7 mm and 0.1 mm respectively, so I assumed that these numbers were supposed to represent the average distance between wefts. Thus, I divided 1 cm by these numbers to get values of 14 and 100 wefts per cm.

Regarding descriptions of the artifacts, I edited two entries in the *Condition* column by using the Rename feature to alter their descriptions, which had the value of "Charred and not charred". Since this description seemed to suggest that the artifacts were "Partially charred", I changed the descriptions to this value in order to maintain consistency across the entire column. I also used cluster and edit to change the descriptions of artifacts in the *Weave Structure* column, because the questions in the second part of the assignment require statistical analysis that focuses on the primary kinds of weave structures. As such, I decided to group the artifacts into two primary groups: weft-faced plain weaves and balanced plain weaves. However, other types of weave structures (i.e. interlooping and twine) were not merged into these groups, as well as objects whose weave structures could not be identified.

For all columns, I made sure to trim leading and trailing whitespace so that all entries containing the same descriptions would be grouped together and no entries would be left out accidentally.

(d) *When you are satisfied with your dataset, export it to an appropriate file format.*

The file was exported to .csv format.

Part 2. ***Data Analysis***

(a) *What are the most common types of textile objects in the dataset?*

Using the *sort* and *table* functions in R, I found that the three most common types of textile objects were textile fragments (92), pseudomorphs (8), and rope fragments (5). With respect to condition, more objects were charred (91) than not charred (23) or partially charred (2), and with respect to colour, most were black (88).

(b) *Which of the buildings produced the greatest number of the textile-related items included in the dataset?*

Using the *sort* and *table* functions in R, I found that ***Building IV-V*** contained the most textile-related items (43), followed by Building II (38).

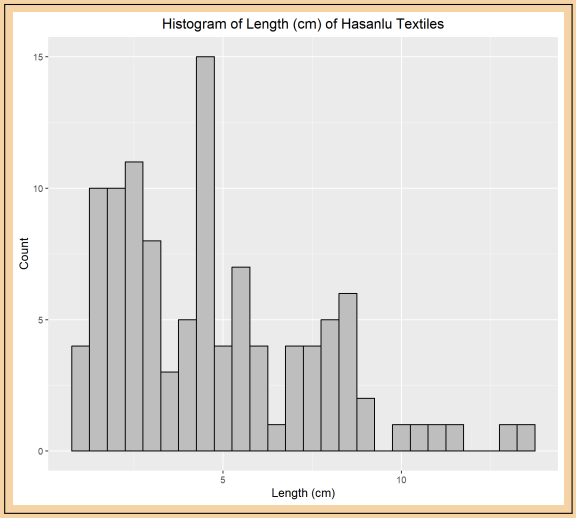
(c) *For the two buildings with the most textile-related items, explore which rooms in those buildings the textile items originate from. Locate the most common findspots for textile-related items on the plan above. Do you notice any patterns in the spatial distribution of these textiles?*

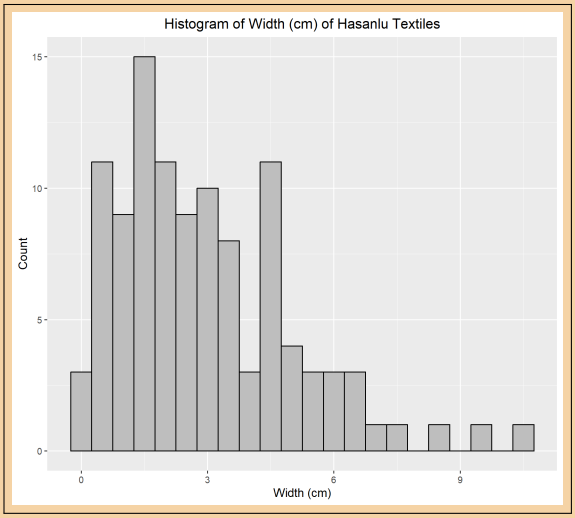
Using the *sort*, *table*, and *filter* functions in R, I found that textiles within Building IV-V were found in Rooms 1, 2, and 4, with the most common finds being in Room 4 (22). For Building II, textiles were found in Rooms 5 and 8, as well as in various other undefined areas, with the most common finds being in Room 5 (32). Regarding patterns, it seems that for Building II the vast majority of the textiles were located in Room 5—looking at the map, this appears to make sense as Room 5 is one of the largest rooms on the entire map. Overall, one can see that the most textiles are generally concentrated around the lower court, in rooms that are only accessible by entering other rooms first (if one starts in the Lower Court, then they must go through Room 2 to get to Room 5 in BBII, and must go through Room 1 to get to Room 4 in BBIV-V).

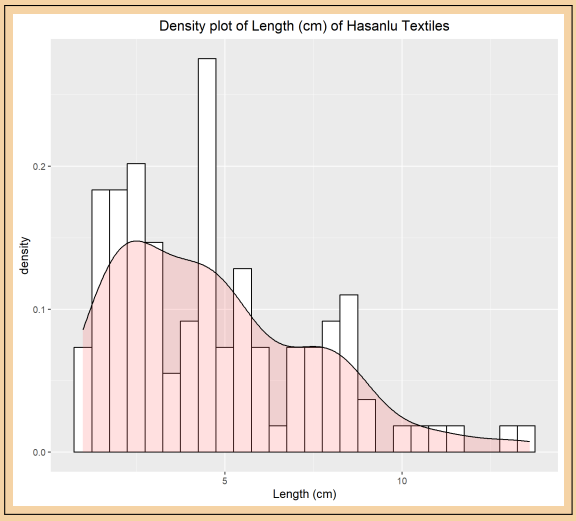
(d) *What are the most common kinds of weave structures among the textile fragments?*

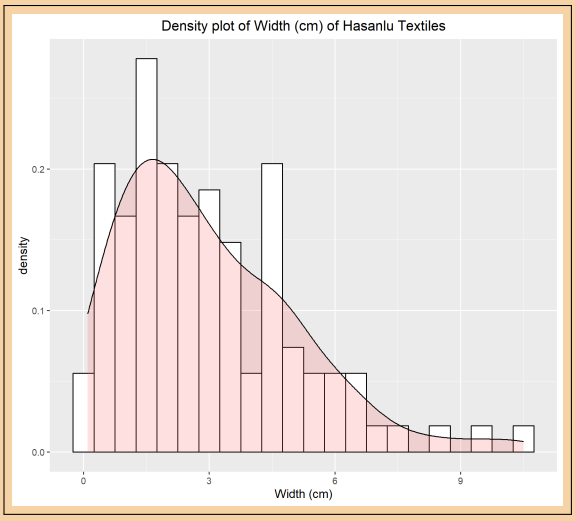
Using the *sort* and *table* functions in R, I found that the most common type of weave structure was the ***weft-faced plain weave*** (69), followed by the balanced plain weave (19). Note that within these two categories, there are sub-categories (e.g. weft face plain weave with supplementary pile and/or loops) that I have not considered, since the criteria used by the original creators of the dataset for categorizing textiles by weave structure seems to be inconsistent—some have more detailed descriptions of the weave structure, while others do not.

(e) *Make histograms that plot the distributions of the length and the width measurements of the textile fragments. Do these distributions appear to be normal distributions? Conduct a test to determine whether in fact each of these distributions is normal, and report the results. Replot the histograms to compare the length and width distributions based on weave structure, and compare the results by eye. Are there any obvious differences in these measurements for different weave structures?*

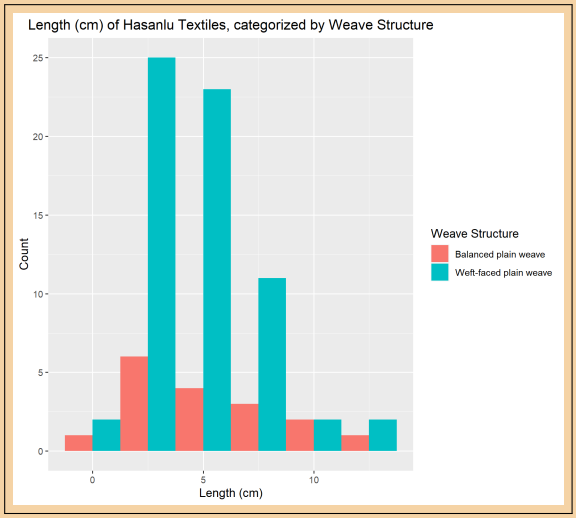


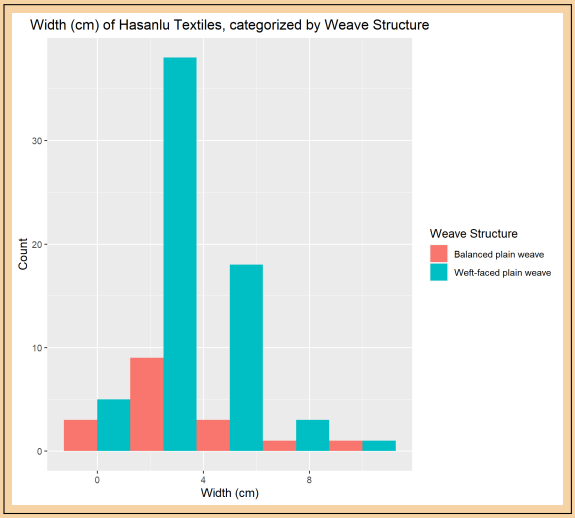






Both distributions appear to be asymmetric and left-skewed, and hence not normal. Using a *Shapiro-Wilke* test via the shapiro.test function in R, it was found that neither of the distributions were in fact normal: the length distributions yielded a p-value of 1.244×10^{-5}, while the width distributions yielded a $p-$value of 7.679×10^{-6}: these values are both below the required $p \geq 0.05$ to accept the null hypothesis of the distributions being normal.



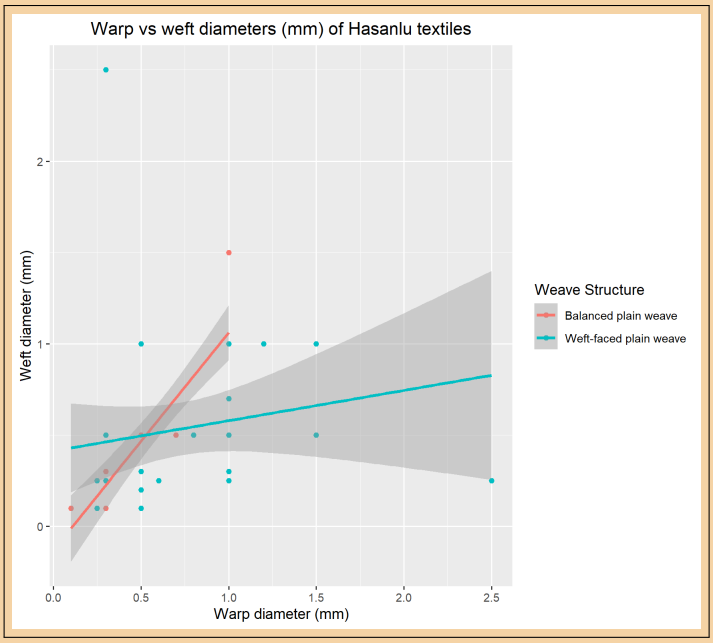


For length and width distributions based on weave structure, it appears that the mean and median lengths are greater for weft-faced plain weaves than for balanced plain weaves. Despite this, the length and width density distributions for both weave structures appear to be *somewhat* similarly shaped (asymmetrical, left-skewed curve), although the weft-faced distributions are shifted to the right of the balanced distributions since the values within the former are generally greater.

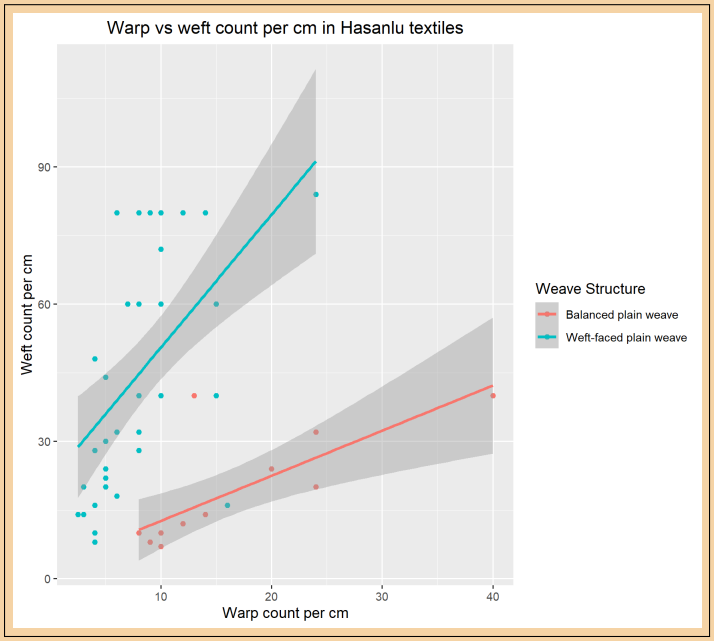
There is not enough data in the histogram to make any conclusive judgements about the interlooping and twined distributions.

(f) *Make a scatter plot of the warp diameter vs. the weft diameter of the textile fragments, grouped by weave structure. Then make a scatter plot of the Warp count per cm vs. the Weft count per cm for the textile fragments, again grouped by weave structure. For both, use the primary kind of weave structure identified to categorize the fragments. Are there any noticeable patterns in the size of the yarn diameters or the warp/weft densities based on weave structure?*

For this question, I excluded the interlooping and twine textiles from my analysis, because they do not contribute any meaningful information which can help me discern the necessary patterns for answering this question.



For the warp vs. weft diameter scatterplot, it appears that weft-faced plain weaves generally have greater warp and weft diameters than balanced plain weaves. There are also more outliers in the former group than the latter, but this may be explicable by the presence of more data points in the former group. Both groups appear to experience an increase in weft diameter as warp diameter increases, and vice versa (i.e. an increase in one will generally lead to an increase in the other). However, the rate of increase for the two groups seems to be substantially different; the rate at which weft diameter increases when warp diameter increases is considerably higher for balanced plain weaves, while the rate at which warp diameter increases when weft diameter increases is considerably higher for weft-faced plain weaves.



For the warp vs. weft count per cm scatterplot, it appears that the mean and median weft count per cm are higher for the weft-faced plain weaves, while the mean and median warp count per cm are higher for the balanced plain weaves. The rate of increase appears to be somewhat linear for both groups, but also differs despite this similarity: the rate at which weft count per cm increases when warp count per cm increases is higher for weft-faced plain weaves, while the rate at which warp count per cm increases when weft count per cm increases is much higher for balanced plain weaves.

Generally, it seems that the distribution of weft-faced plain weaves is more varied (i.e. residuals are higher on average) than the distribution of balanced plain weaves for both scatterplots.