**Problem 1.** *Load the Snodgrass dataset from the ArchData package*

**Problem 2.** *How many observations are there in this dataset? How many variables?*

91 observations, 15 variables

**Problem 3.** *Use the help function for the Snodgrass dataset to look up the meanings of each of the different variables in the dataset. Describe the variables and the data types represented.*

| Variable Name | Description | Data Type |
|---|---|---|
| East | East grid location of house in ft (according to the excavation grid system) | num |
| South | East grid location of house in ft (according to the excavation grid system) | num |
| Length | House length in ft | num |
| Width | House width in ft | num |
| Segment | Three areas within the site (1, 2, 3) | Factor w/ 3 levels "1", "2", "3" |
| Inside | Location within or outside the "white wall" | Factor w/ 2 levels "Inside", "Outside" |
| Area | Area in square ft | num |
| Points | Number of projectile points | int |
| Abraders | Number of abraders | int |
| Discs | Number of discs | int |
| Earplugs | Number of earplugs | int |
| Effigies | Number of effigies | int |
| Ceramics | Number of ceramics | int |
| Total | Total number of artifacts listed above | int |
| Types | Number of kinds of artifacts listed above | int |

**Problem 4.** *Discuss the dataset's data quality in terms of the characteristics we discussed in lectures and assess the documentation/description of the dataset in the Help document. Are there ways the documentation or data quality could be improved?*

From a numerical standpoint, the dataset seems to be of high quality. The East and South values were all measured CONSISTENTLY to the nearest 0.01 ft; all measurements under the Length and Width values are also consistent, since they were all measured to the nearest 0.1 ft. There is also a clear UNIFORMITY in the distance units used to conduct measurements—each distance unit (East, South, Length, Width) is measured in feet. These factors, when combined with the tabular format of the data, make the data easily REPRO-DUCIBLE (whether digitally or through printing) for future archaeologists to consult.

Despite these positives, there are still areas where the quality of the dataset could be improved. For instance, there appears to be a typo in the description of the South variable, as the R documentation seems to give it the same description as that of the East variable—this could lead to confusion for outside observers, and ought to be fixed. Also, the name of the *Points* variable is not very descriptive, and could perhaps be replaced with something like "ProjectilePoints" to clarify the true purpose of the variable.

**Problem 5.** *Compare the sizes of the houses located in different areas (Segments) of the site. What variable(s) will you use to measure house size? What function(s) will you use to summarize the house sizes for the different areas of the site, and why? Do there seem to be any differences between the sizes of houses in different areas? Is there any difference in size between houses located inside the White Wall vs. those located outside?*

Using the *Area* variable of Snodgrass, as well as the *group_by*, *summarise*, *mean*, and *median* functions in R, I was able to compare the sizes of the houses in each segment. I chose the *group_by* and *summarise* functions in order to first separate all the observations into their respective segments and then obtain the key values that I wanted to find: the mean and median area for each segment. It was mentioned in lecture that these values were particularly pertinent for gaining insight on the tendencies of any dataset as a whole, so I felt that calculating both of them would yield a comprehensive analysis for this question.

It appears that the houses in Segment 1 were generally the largest, with a mean house area of 317.37 ft$^2$ and median of 319.50 ft$^2$. This was followed by Segment 3, which had a mean area of 192.79 ft$^2$ and median of 195.75 ft$^2$, and then Segment 2, which had the smallest mean and median area of the three segments at 166.79 ft$^2$ and 156.00 ft$^2$ respectively.

Using the *group_by* and *summarise* functions again, I found the mean and median areas of houses inside and outside the White Wall: those inside had a mean area of 317.37 ft$^2$ and median of 319.50 ft$^2$, while those outside had a mean of 179.06 ft$^2$ and median of 175.50 ft$^2$.

**Problem 6.** *Compare the numbers of artifacts found in the houses in the different areas of the site. What variable(s) and function(s) will you use to summarize the total numbers of artifacts in the houses for the different areas of the site, and why? Do there seem to be any differences in the total number of artifacts per house between different areas? Is there any difference in the total artifact numbers between houses located inside the White Wall vs. those located outside?*

Using the *Total* variable of Snodgrass, as well as the *group_by*, *summarise*, *sum*, and *mean* functions, I was able to compare the number of artifacts in each segment. I chose the

*group_by* and *summarise* functions in order to first separate all the observations into their respective segments and then obtain the key values that I wanted to find: the total number of artifacts in each segment, and the mean number of artifacts per house. This is because, when comparing the number of artifacts across different areas, it makes sense to find the total number of artifacts in each area, in order to consider each area as a whole; as well as the average number of artifacts per house, in order to gain information about the average material circumstances of owners of houses in one area relative to those of owners in another area.

Segment 1 had the most artifacts in total at 503, followed by Segment 2 with 56 and then Segment 3 with 50. In terms of the average number of artifacts per house, Segment 1 also had the highest at 13.24, while Segments 2 and 3 both had an average of 2.

Using the *group_by* and *summarise* functions again, I also found the total number of houses inside and outside the White Wall, as well as the mean number of artifacts per house for these regions: there were 503 artifacts in total and 13.24 per house inside the White Wall, and 106 artifacts in total and 2.00 per house outside.

**Problem 7.** *Compare the diversity of artifacts found in the houses in different areas of the site. What variable(s) and function(s) will you use to summarize the diversity of artifacts found in the houses in different areas of the site, and why? Do there seem to be any differences in the variety of artifacts occurring in the houses found in different areas? Is there any difference in artifact diversity between houses located inside the White Wall vs. those located outside?*

Using the *Types* variable of Snodgrass, as well as the *group_by*, *summarise*, and *mean* functions, I was able to compare the diversity of the artifacts in each segment. I chose the *group_by* and *summarise* functions in order to first separate all the observations into their respective segments and then obtain the key value that I wanted to find: the mean number of artifact types per house in each segment. I did not use the *sum* function here like I did for #6, because there might be different houses containing the same type of artifact; calculating the total number of types in the dataset using the *sum* function would count these types separately instead of as one single type, thereby leading to a misrepresentation of the data.

Segment 1 had the most types of artifacts per house at an average of 7.68, followed by Segment 2 with 1.82 and Segment 3 with 1.68.

Using the *group_by* and *summarise* functions again, I found the mean number of types of artifacts inside and outside the White Wall: there were 7.68 types per house inside, and 1.75 per house outside.

**Problem 8.** *Create a table that summarizes the frequencies of different artifact types in the different areas of the site. What function(s) will you use to summarize these frequencies, and why? Do you notice any patterns in the frequencies of particular artifact types in different areas?*

The table (named *ArtifactFrequency*) was created using R code and can be viewed using the R file I've attached to this email. I used the *group_by*, *summarise*, *mean*, and *sum* functions because I wanted to obtain the average number of each artifact type per house in the different areas, as well as the total number of each artifact type in each area.

I noticed that Segment 1 (i.e. the White Wall) had the highest number in total of *every* type of artifact, as well as the highest average number of each type of artifact per house.

**Problem 9.** *What kinds of possible hypotheses might you suggest to explain the differences in house sizes and artifact distributions in the different areas of the site that you have observed in your answers to the previous questions?*

From the disparities in house size and artifact distributions between houses in the White Wall and houses outside, I would infer that people living in the White Wall were more affluent, while those who lived outside were generally less so. This is because having a bigger house and more possessions is generally seen today as a sign of wealth and/or social status, and I believe there is no reason to think that things would be otherwise for people living in the 1300-1400s.

**Problem 10.** *How would you test your hypotheses from your answers to #9 above? What kinds of other data/variables might you gather to test them?*

I would create a box plot to gain a better understanding of the number of outliers in the data: if more outliers are present, then they could have a significant impact on the mean and total values of the data, and would thus call my hypothesis from #9 into question.

If I were an archaeologist and wanted to test my hypothesis, I would try to gather written records such as historical documents or journals, which could yield valuable insight on the way of life of the population and potentially the reasoning for the distribution of artifacts in different areas. Next, I would analyze the material composition of each artifact; if those made of valuable materials such as gold or diamonds were more common within the White Wall, this would help to confirm my hypothesis. Finally, I would search for more artifacts which are commonly found among affluent people: jewelry, larger/more elaborate tombstones, and perfume containers are a few examples.