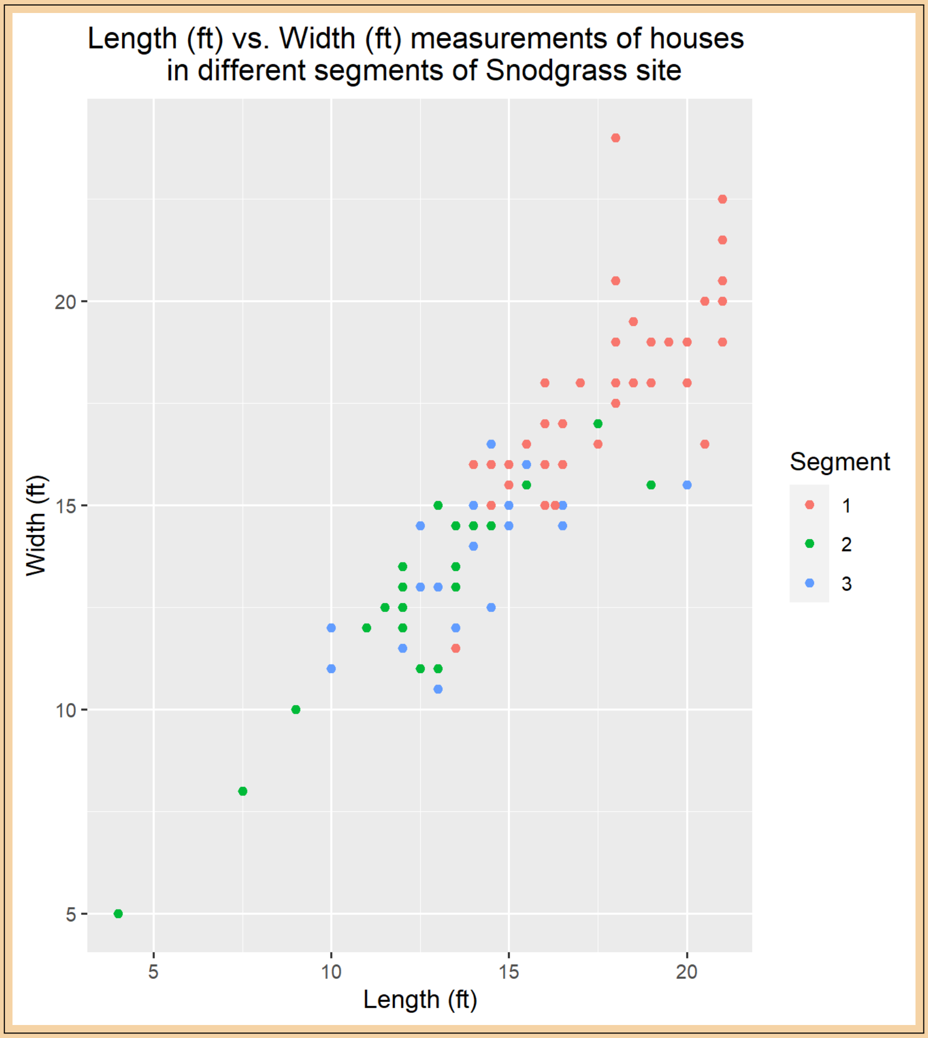
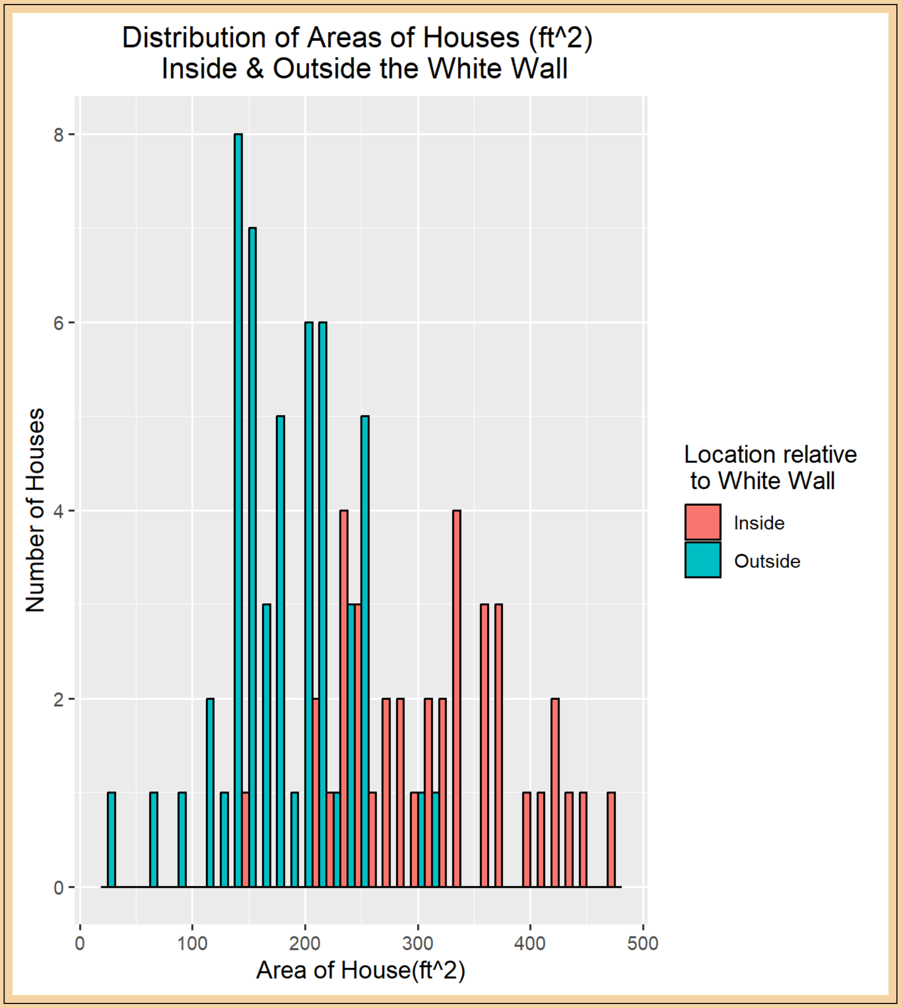
Part 1. ***Examining House Size***

(a) *Create a scatterplot that compares Length and Width measurements for the houses in the different segments of the site. What patterns do you notice in the scatterplot, and how do you interpret them?*



From the scatterplot, it can be seen that length and width of houses are generally correlated through a linear relationship similar to the function $y = x$: as the length increases, the width seems to increase by the same amount (and vice versa). Also, Segment 1 clearly tends to have the longest and widest houses, followed by Segment 3 and then Segment 2, barring a few outliers. As such, I would infer that the houses in Segment 1 were the largest, which suggests that people living in this area of Snodgrass might have had some social and/or economic advantages over residents of the other two regions. There is also a slight difference in overall length vs. width measurements of the houses in Segments 2 and 3, which reveals itself most notably through the three outliers belonging to Segment 2 which all have lengths and widths less than or equal to 10 ft. I would interpret this pattern as suggesting that overall, the house sizes in Segments 2 and 3 were largely similar, but there might still be a possibility that those living in Segment 2 were slightly smaller and possibly less affluent as a result.

(b) *Create a histogram that compares the distribution of house area inside vs. outside the White Wall. What patterns do you notice in the histogram, and how do you interpret them?*



From the histogram, we can see that both distributions are somewhat, though not perfectly, symmetric, and that the area of houses outside the White Wall is generally much lower than the area of houses inside, with the exception of a few outliers. Again, I would interpret this data as suggesting some sort of economic disparity between residents within the White Wall and residents outside, since a bigger house area can typically be equated to a higher income. In addition, it is interesting to note that the people living inside the White Wall were clearly outnumbered by people living outside—perhaps those living inside possessed some sort of power which prevented outsiders from organizing against them?

(c) *What kind of probability distribution does the histogram you created for part b appear to represent? Test to see whether this distribution is appropriate for describing the distributions of house area for both inside and outside the wall, and report the results of both tests. How do you interpret these results?*

At first glance, it appears that the data for house areas both inside and outside the White Wall follow a **normal distribution**, since the shapes of both data are somewhat symmetrical and bell-shaped.

To test the normality of the distributions, I conducted a *Shapiro-Wilke* test using the shapiro.test function in R, as shown in the attached R file. The results of this test suggest that the distributions are indeed normal, since the p-values were found to be 0.7753 for house areas inside the White Wall and 0.6066 outside—both these values agree with the conventional standard of $p >= 0.05$ as a prerequisite for accepting the null hypothesis, which in this case is that the distribution is normal. Thus, I interpret these tests as confirming my initial impression that the distributions of house areas inside and outside the White Wall are both normal—the p-values yielded by the test are *much* greater than 0.05, which means that the probability of the distribution not being normal is very low.

(d) *Based on the results of the previous step, conduct a test to determine whether there are any significant differences in house area between houses located inside vs. outside the White Wall? Report the results of this test and describe how you interpret them.*

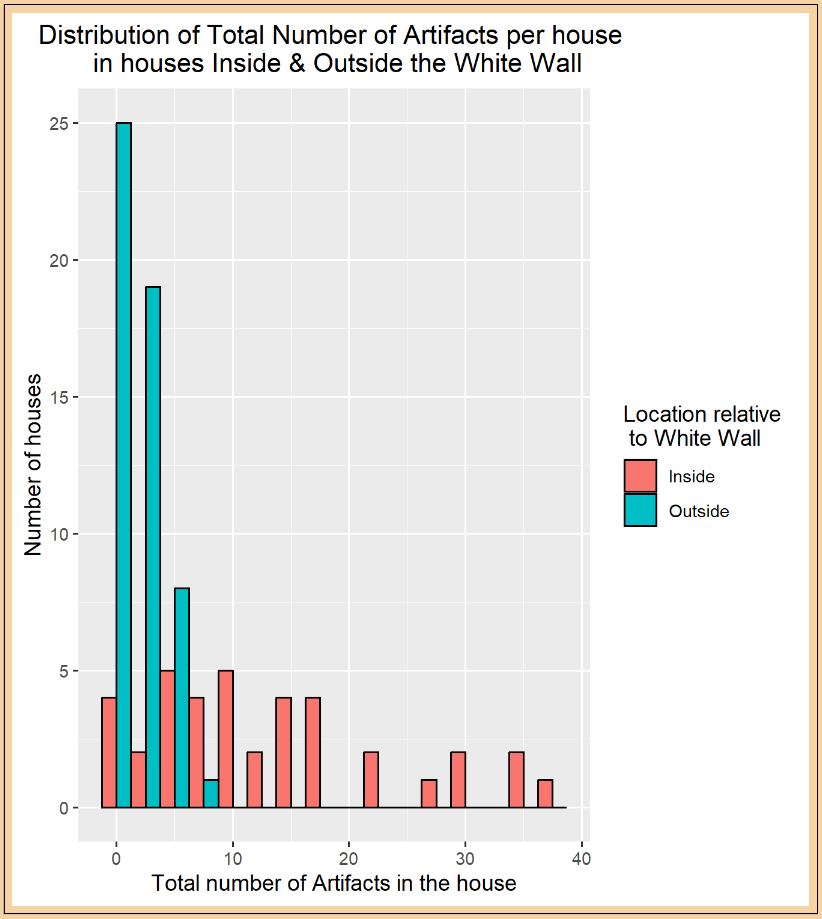
t-tests require an assumption of normality, which was confirmed in part (c) to be appropriate. As such, I used the t.test function in R to conduct a t-test on the data, with the assumption that variances of house area for the two samples (in vs. out of the White Wall) were unequal—this latter assumption was supported by a *Levene* test which yielded a p-value of $p = 0.02175 < 0.05$, thereby showing that the null hypothesis of the variances being equal for both samples should be rejected.

As a result of the t-test, I obtained estimates of 317.3711 and 179.0566 ft^2 respectively for the mean area of houses inside and outside the White Wall, which confirms that the difference in house area between the two regions is statistically significant, given the 95% confidence interval of $[109.2929, 167.3360]$ as the difference between the two means, as well as the p-value of 6.811×10^{-14} which is far lower than the minimum of 0.05 needed to accept the null hypothesis of the difference in mean between the two groups being 0.

The result of this test solidifies and adds further credible evidence to support a location-related disparity of mean house area between the two regions—the extremely low p value suggests that it is quite unlikely for the mean values of the two distributions to be equal. As such, this test provides clear statistical evidence to support my interpretation of the populace within the White Wall as having a higher income and better societal standing, since said interpretation was predicated on the existence of a statistically significant difference in mean house area between the two regions.

Part 2. *Examining Artifact Numbers*

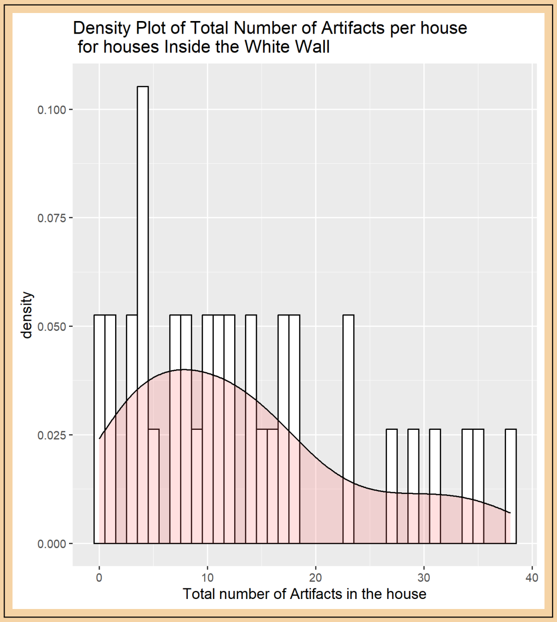
(a) *Create a histogram that compares the total number of artifacts per house in houses inside vs. outside the White Wall. What patterns do you notice in the histogram, and how do you interpret them?*

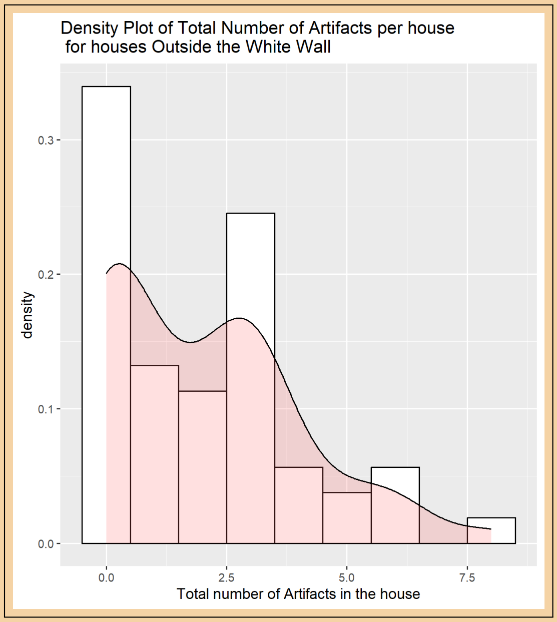


From the histogram, we can see that the range of total artifact number per house is much lower for houses outside the White Wall, and that most houses inside the White Wall have more artifacts than the house with the most artifacts outside the White Wall. From this info, I infer that people living inside the White Wall must generally be more affluent, since having more material possessions betokens greater wealth and material prosperity. Moreover, I would infer that the people living outside the White Wall had to deal with largely similar material circumstances, due to the high frequency of houses possessing a relatively small range (0-8) of total number of artifacts compared to those living inside.

(b) *What kinds of probability distributions might be appropriate for this data, based on the histogram you created for part a? Test to see whether a normal distribution is appropriate for describing the distributions of total artifact counts per house for both houses inside and outside the wall, and report the results of both tests. How do you interpret these results?*

Using code to create density distributions for artifacts inside and outside the White Wall, I obtained the following results:


Density Plot of Total Number of Artifacts per house
for houses Inside the White Wall


Density Plot of Total Number of Artifacts per house
for houses Outside the White Wall

The asymmetrical, left-skewed distribution of samples inside the White Wall, and the bimodal distribution of samples outside, suggest that it is inappropriate to characterize these samples as being normally distributed.

To test for normality, I used the shapiro.test function again to determine the p-values of the data—the resulting values of 0.007752 (inside) and 3.057×10^{-5} (outside) returned by this test show that normality cannot be assumed for either distribution, since both values are less than the conventional 0.05 required to make this assumption. Thus, I interpret this test as confirming my initial impression that the distributions of total number of artifacts per house both inside and outside the White Wall cannot be normal—the extremely low p-values for both distributions suggest that the possibility of their being normal is very unlikely. This also means that a t-test cannot be conducted on the data, since a t-test only works on data which are normally distributed.

(c) *If you determined in part (b) that a normal distribution* ***was*** *appropriate for describing these distributions, conduct a test to determine whether there are any significant differences in the total number of objects per house between houses inside vs. outside the White Wall. Report the results of this test and describe how you interpret them.*

It was found in (b) that the distributions were not normal; thus, no test was made for this question. However, if the distributions in (b) *were* normal, I would have followed the same steps as I did in Part 1(d).

(d) *Create a contingency table that summarizes the sum total for each artifact type for all houses inside the White Wall vs. all houses outside the White Wall. Does this table fulfill the requirements for conducting a chi-square test?*

Using the summarise function in R, I created a contingency table containing the sum total of each artifact type in Snodgrass for houses both inside and outside the White Wall—this table can be viewed in the attached R file. The two requirements mentioned in lecture for a chi-squared test are that at least 80% of the entries in the contingency table must be greater than 5, and that none of the entries can be equal to zero. Since the table I produced meets both of these conditions (83.33% of the cells are greater than 5; none of the entries are equal to 0), it fulfills the requirements for conducting a chi-squared test.

(e) *If your answer to part c above is* ***yes****, conduct a chi-square test on this contingency table, and present the results of this test. What would the null hypothesis for this test be? What is the alternative hypothesis? Which of these hypotheses would you accept, based on the results of your chi-square test? Which cells of the table are contributing most to any difference in the distribution of different artifact types between the two areas of the site?*

Using the chisq.test function in R, I conducted a chi-squared test which yielded $\chi^2 = 9.5857$, df $=$ 5, and $p = 0.08786$. Given the null hypothesis, which is that no relationship exists between the location (inside or outside the White Wall) and the sum total for each artifact type, and the alternative, which is that a relationship between these attributes *does* exist, the p-value of 0.08786 produced by the test suggests that I should accept the former, since it is greater than the conventional prerequisite of 0.05.

To determine which entries have the greatest impact on the difference in distribution between the two areas, I used the residuals function in R—entries with the highest absolute value of residual are those altering the difference the most. Since the 2nd entry of the Outside row (11 abraders, residual of 2.301) and the 6th entry of the Outside row (30 ceramics, residual of -1.188) have the greatest absolute residuals out of all entries in the contingency table, it follows that they have the biggest effect on the difference in distribution of different artifact types between the two areas.

Part 3. *Interpretation*

(a) *Based on the results you have obtained from Parts 1 and 2, revisit the hypothesis you proposed in your answer to Question 9 from the first assignment. How would you modify this hypothesis based on the visualizations and tests you have conducted?*

Overall, my hypothesis in Assignment #1 was corroborated by the tests—the results from Part 1, which revealed a statistically significant difference in house areas inside and outside of the White Wall, as well as the histogram produced in Part 2(a), serve to support the notion that residents inside the White Wall were more affluent than those outside. However, the chi-squared test conducted in Part 2(e) seems to introduce an element of uncertainty that I hadn't previously been aware of, since it found that no relationship was present between the location and the sum total for each artifact type. As such, I propose to modify my hypothesis slightly: The disparity in house size and number of each artifact type is *likely* due to the relative affluence of residents within the White Wall compared to residents outside, but there might be other valid explanations for the distributions of house areas and artifacts which I have not accounted for in my analysis thus far.

For example, it might be the case that certain houses within the White Wall were used as communal spaces or to safeguard various items, and that no-one (or very few people) actually lived inside those houses; it might also be possible that the original community consisted only of houses inside the White Wall, and that the residents built houses around this wall for their children after they married and produced offspring. The uncertainty introduced by my findings in Part 2(e) suggest that it may be worthwhile to investigate these theories further, both through collecting new data and through more extensive analysis of data that already exists.