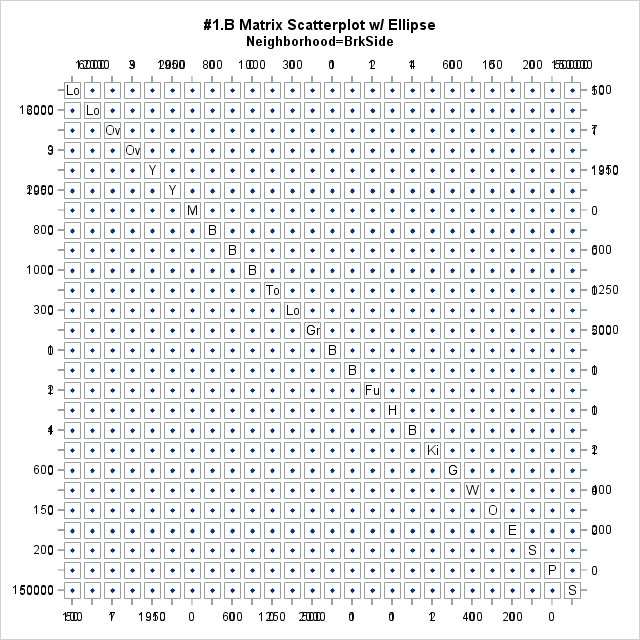
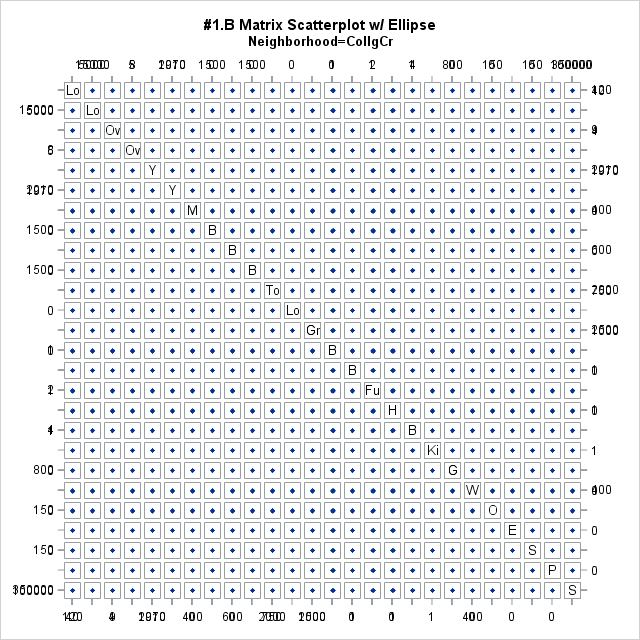
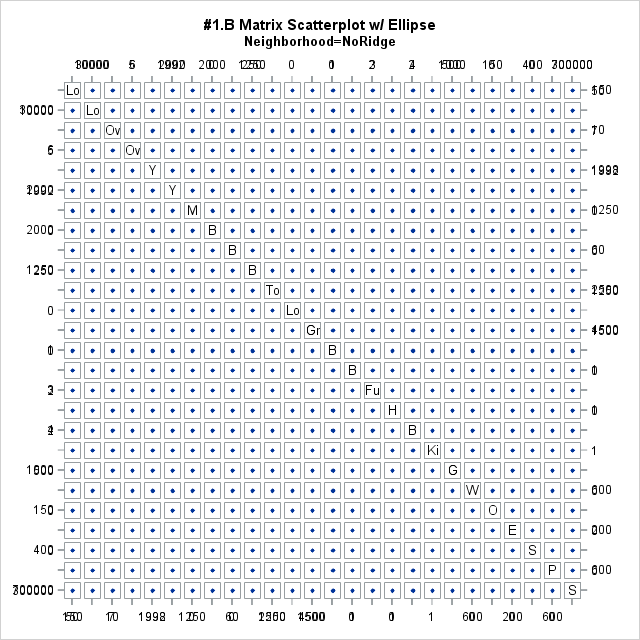
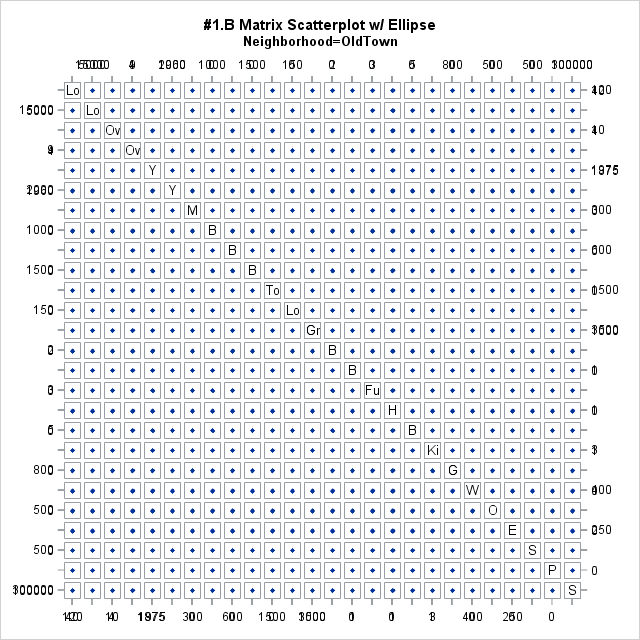
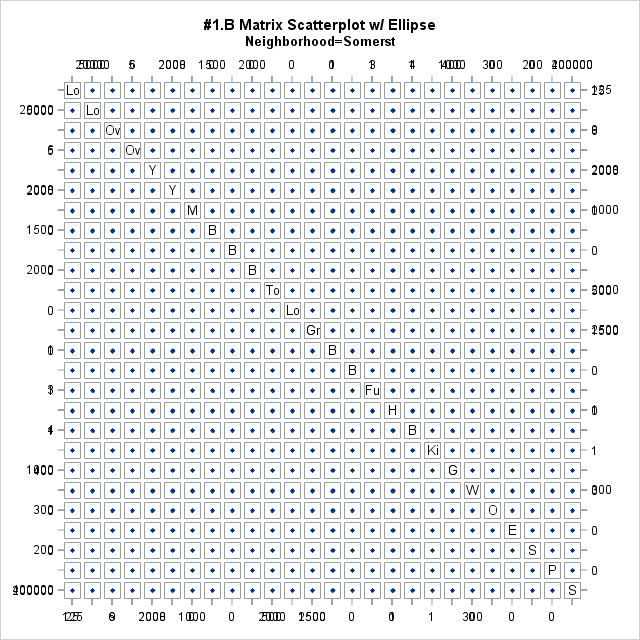
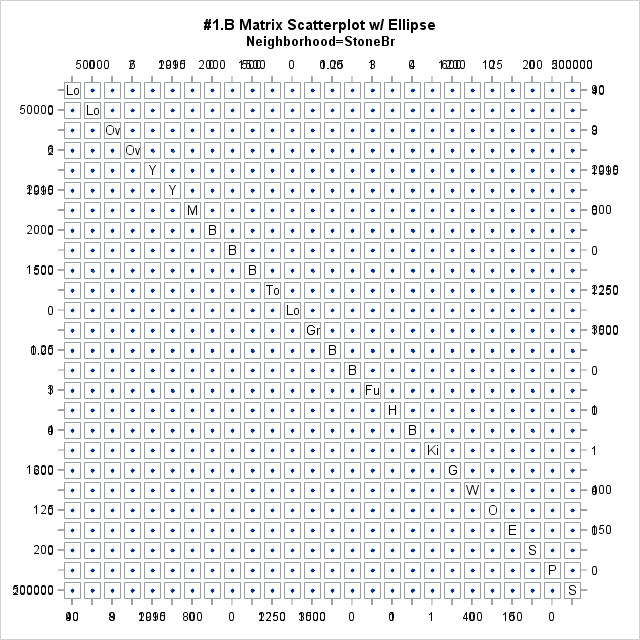
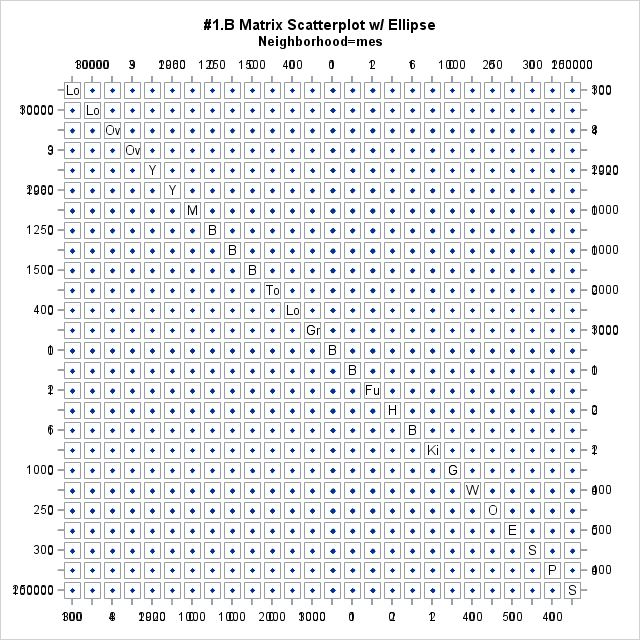
UNIT 8 HW: MANOVA and LDA

Consider the Ames housing data from a Kaggle project. I have adapted the data set (HW8AmesData.xlsx) so that there are only 7 neighborhoods represented. Using this data set we would like to perform the following analysis:

1. We would like to perform a MANOVA to test if there is evidence that any of the means of the of the response variables are different between any of the neighborhoods. In order to do this, we must first check the assumptions of the MANOVA:
   1. Note that for larger sample sizes, there is a multivariate normal central limit theorem and thus the MANOVA is robust to deviance from multivariate normality. We will assume that is the case here.
   2. Generate, copy and paste a matrix of scatter plots of the continuous variables for each of the neighborhoods. Make sure there is a fit ellipse on each. Make a judgement about the equality of the variance and covariances based on these plots. Please include your code as well. There is nothing to do or perform for this part.

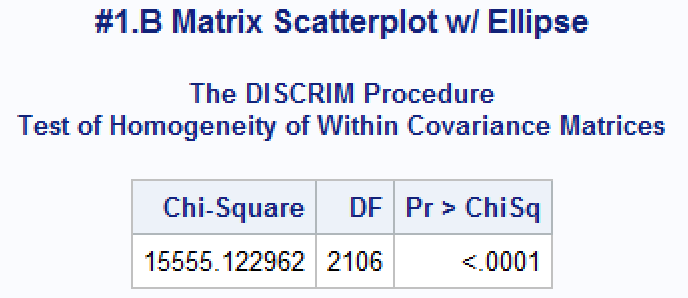




\*Considering there is a large number of continuous variables included, it is nearly impossible to see t he ellipses and the scatterplots as intended.

* 1. It turns out that Bartlett’s test is not robust to departures from the normality assumption and does not perform well in terms of type 1 error for tests with different sample sizes. Just for practice, run Bartlett’s test and write a formal conclusion (one sentence) for Bartlett’s test. Note, for Barlett’s test you will need to use proc discrim with pool=test option**.** Did it agree with your visual assessment above? Copy and paste your code and the table that shows the results of Bartlett’s test.

Based on the result from the Bartlett test, we would proceed with a quadratic model.



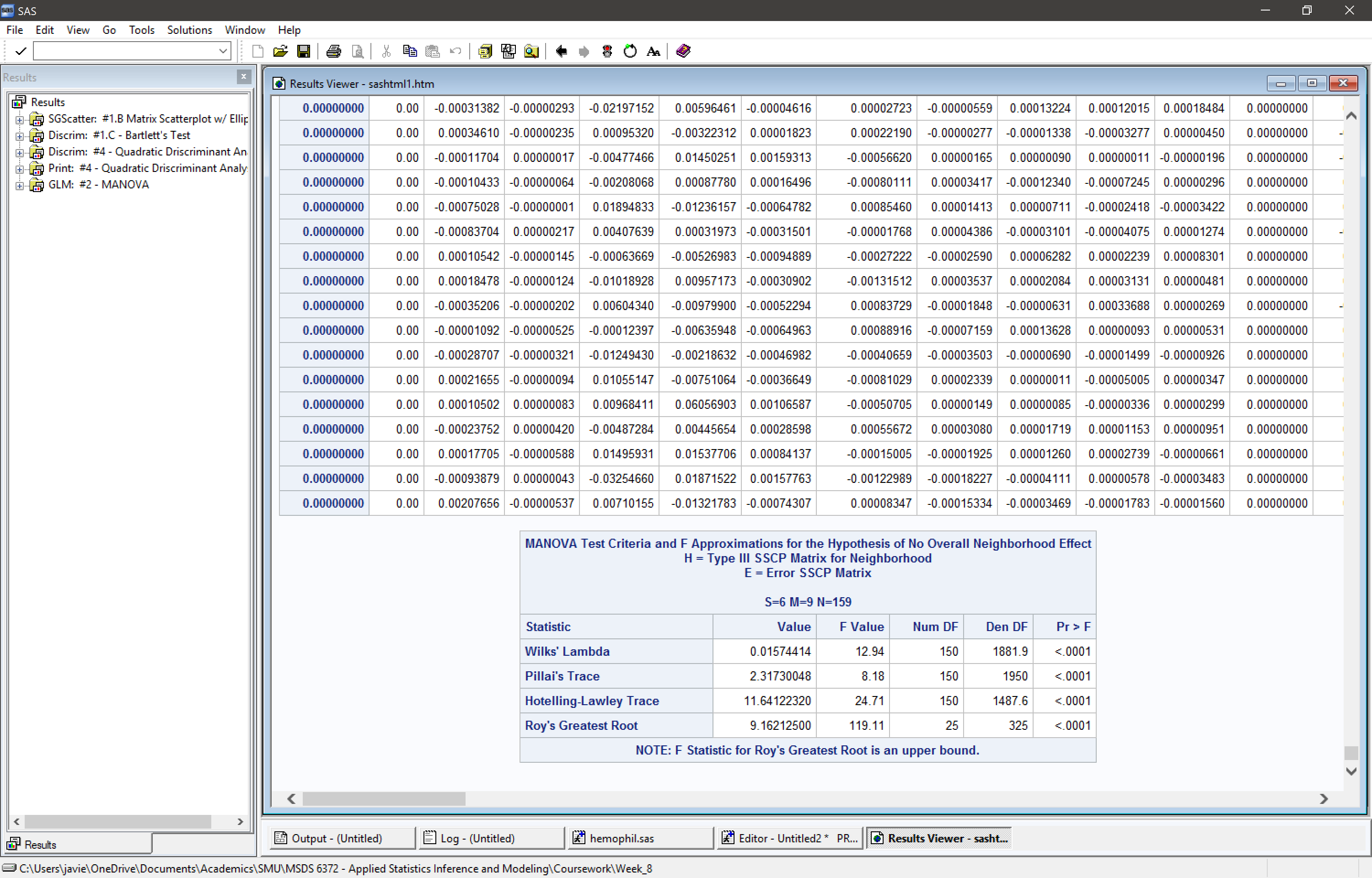


* 1. From here you may proceed with the MANOVA assuming that the test is robust to any departures from multivariate normality and that the variance covariance matrices are equal (despite Bartlett’s test). Simply proceed with caution with respect to the homoscedasticity assumption. You may also assume that observations for the same variables are independent both within and between neighborhoods. There is nothing to do or perform for this part.

1. Perform the MANOVA and copy and paste the MANOVA test result table below. This is the table with Wilk’s Lambda, Hotelling-Lawley Trace, etc. Write a conclusion (one or two sentences) with respect to the test. Is there evidence that at least one mean between at least one pair of neighborhoods is different?

Based on the statistical output from the Wilks Lambda and the other tests, there is strong evidence to suggest that at least one mean between one pair of neighborhoods is different at the 95% confidence level (p-values all <0.0001).

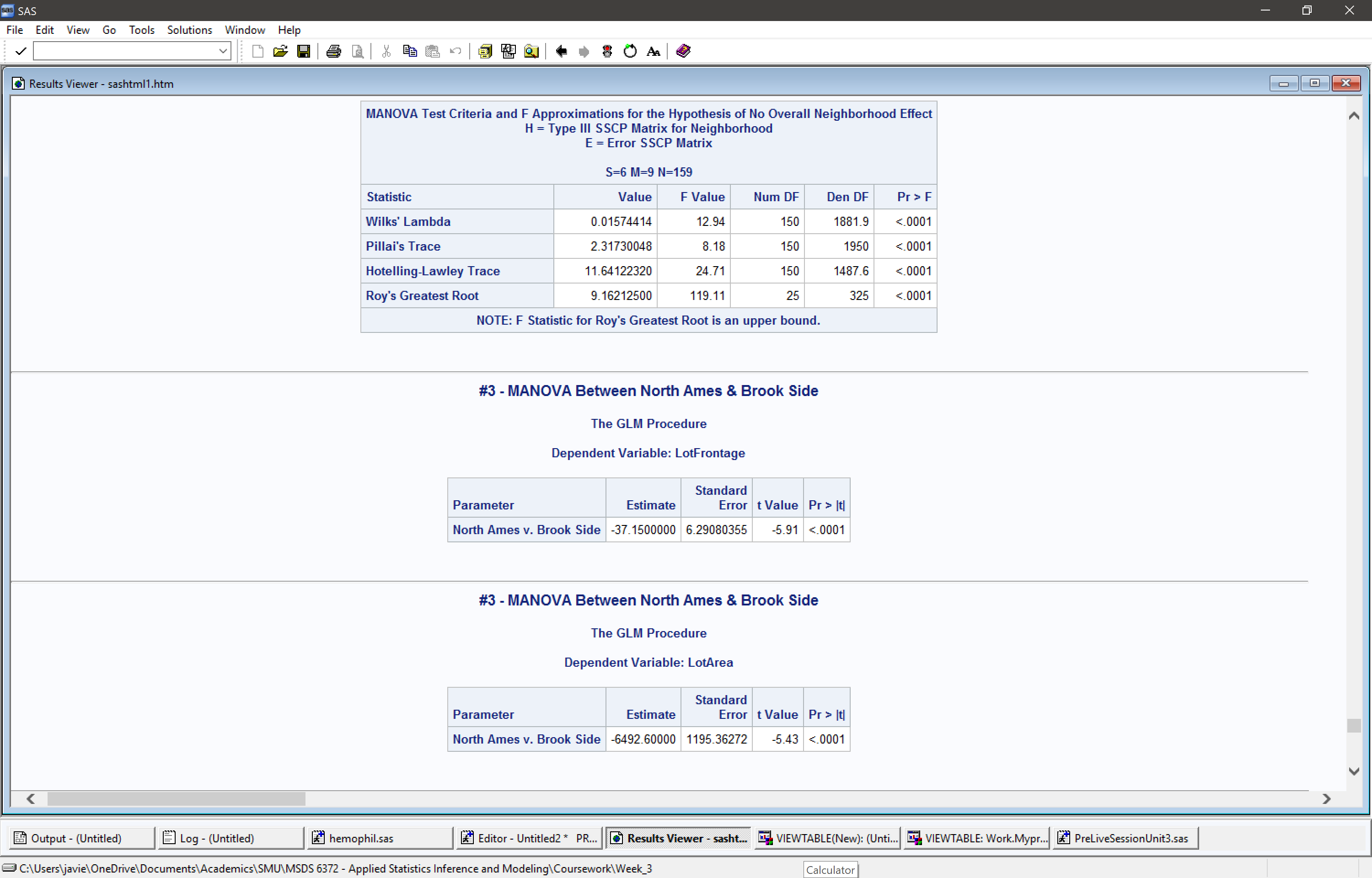




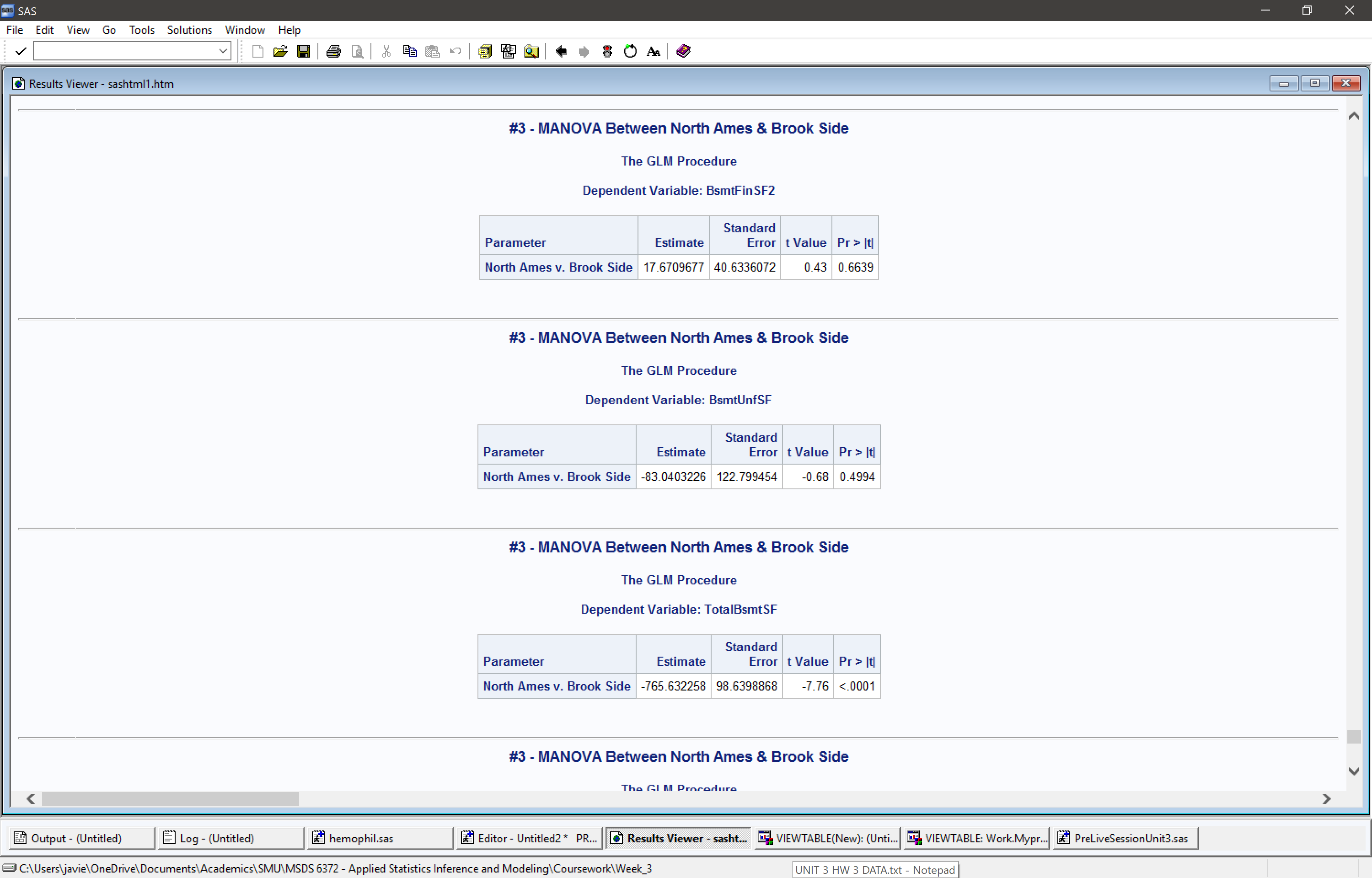
1. You should find that there is evidence of a difference in means for at least one variable between at least one pair of neighborhoods! Now we need to investigate which variables and which pairs of neighborhoods. There are lots of combination to look at here; therefore, simply use a contrast to test for which variables have evidence of different means between the North Ames and Brook Side neighborhoods. Copy and paste the relevant tables as well as the code and be sure and provide at least a one or two sentence conclusion / summary (similar to what is found in the powerpoints).

Based on the test statistics, there is at least one mean that is different between North Ames and Brook Side. The following tables indicate which exact variables are different and which are not by looking at the significance (p-value <0.0001).

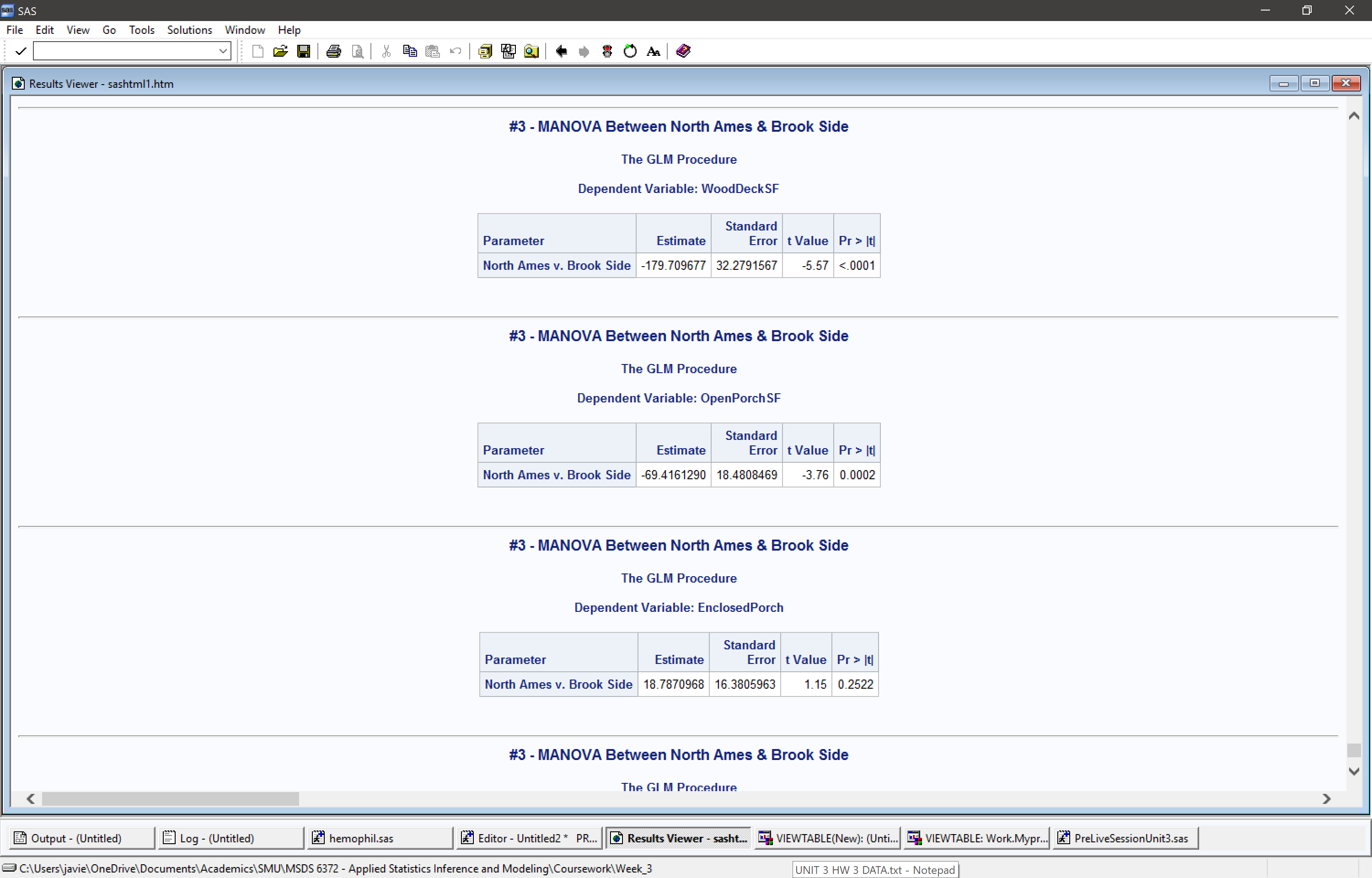
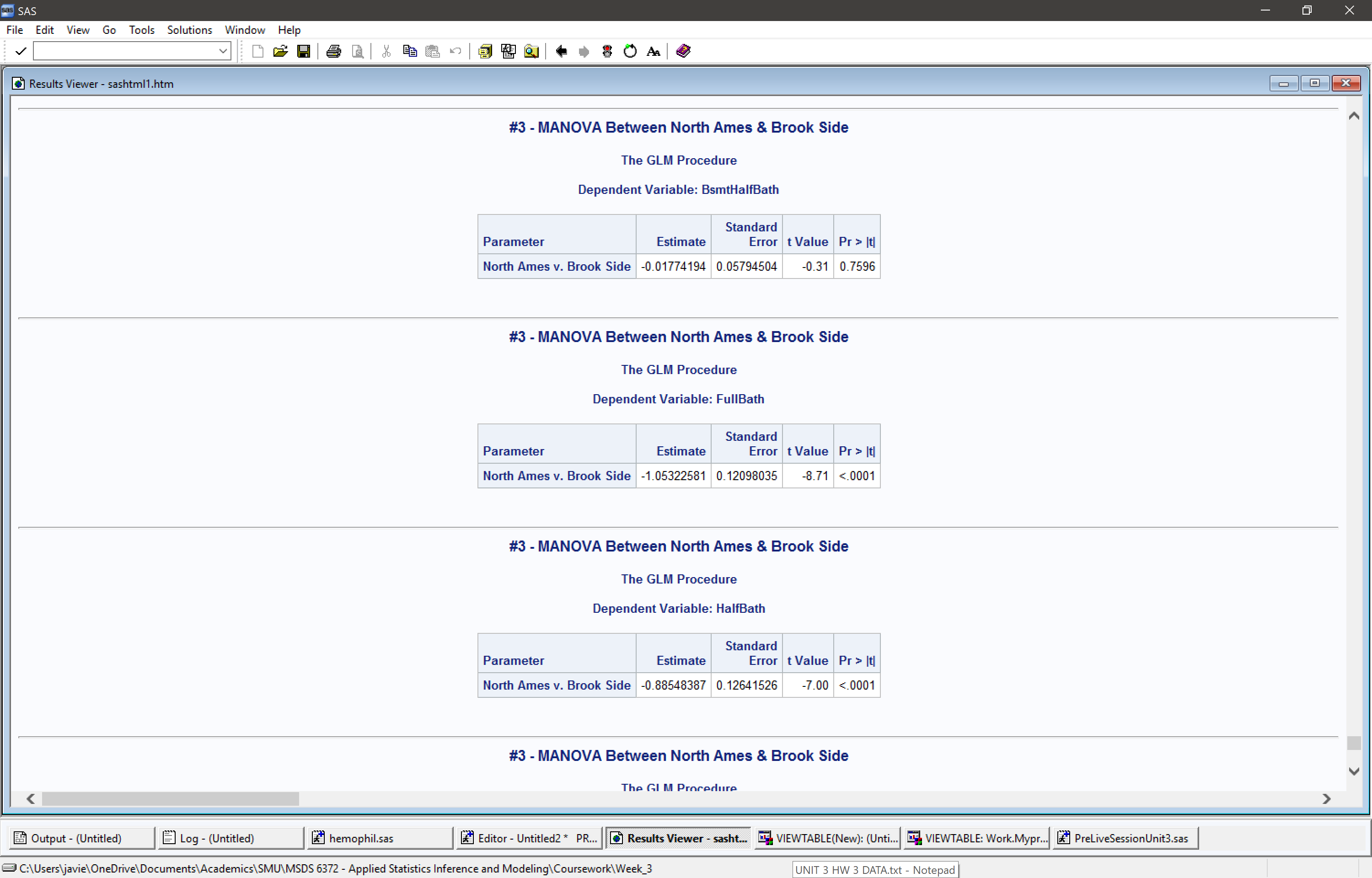
















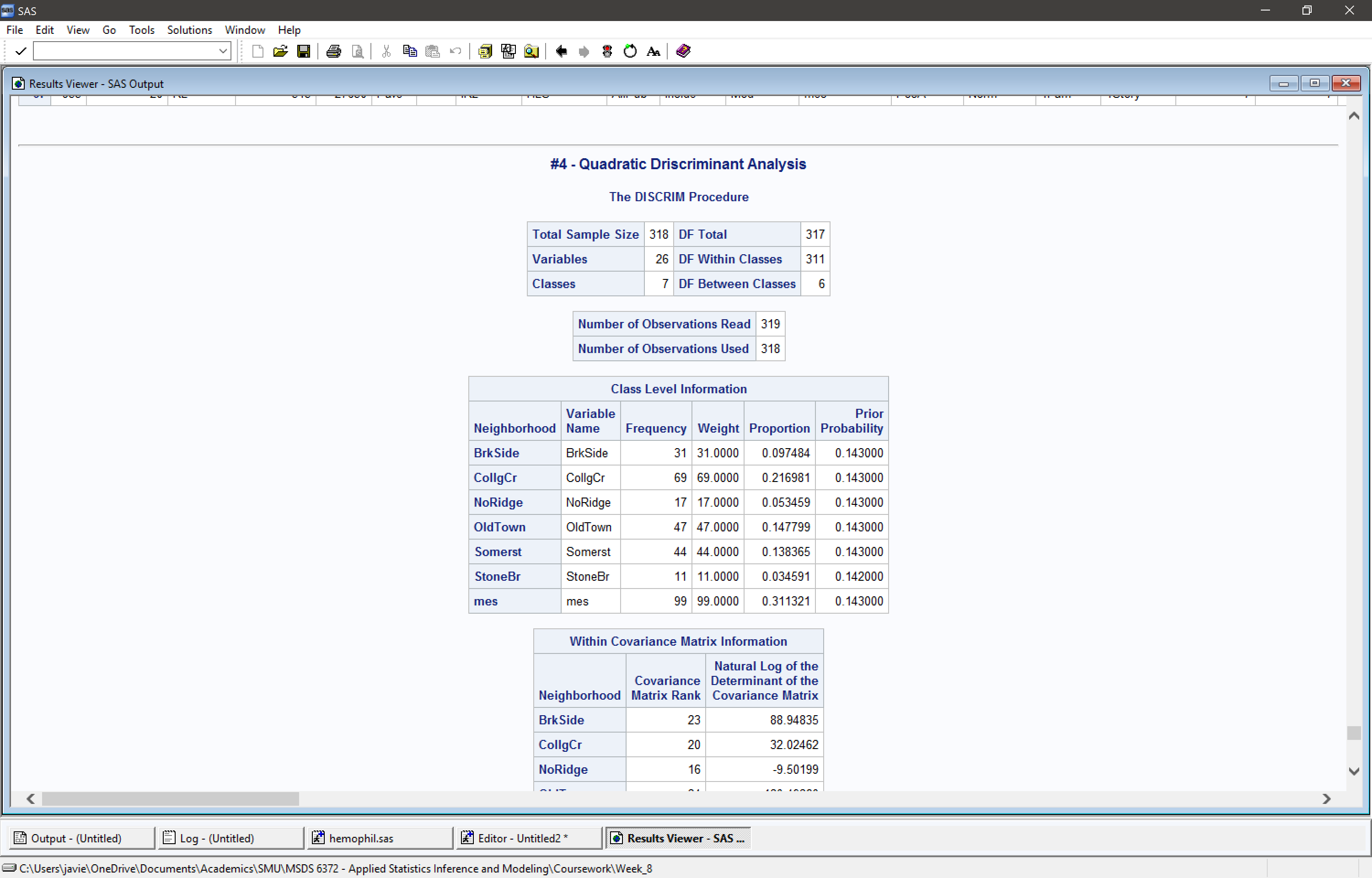


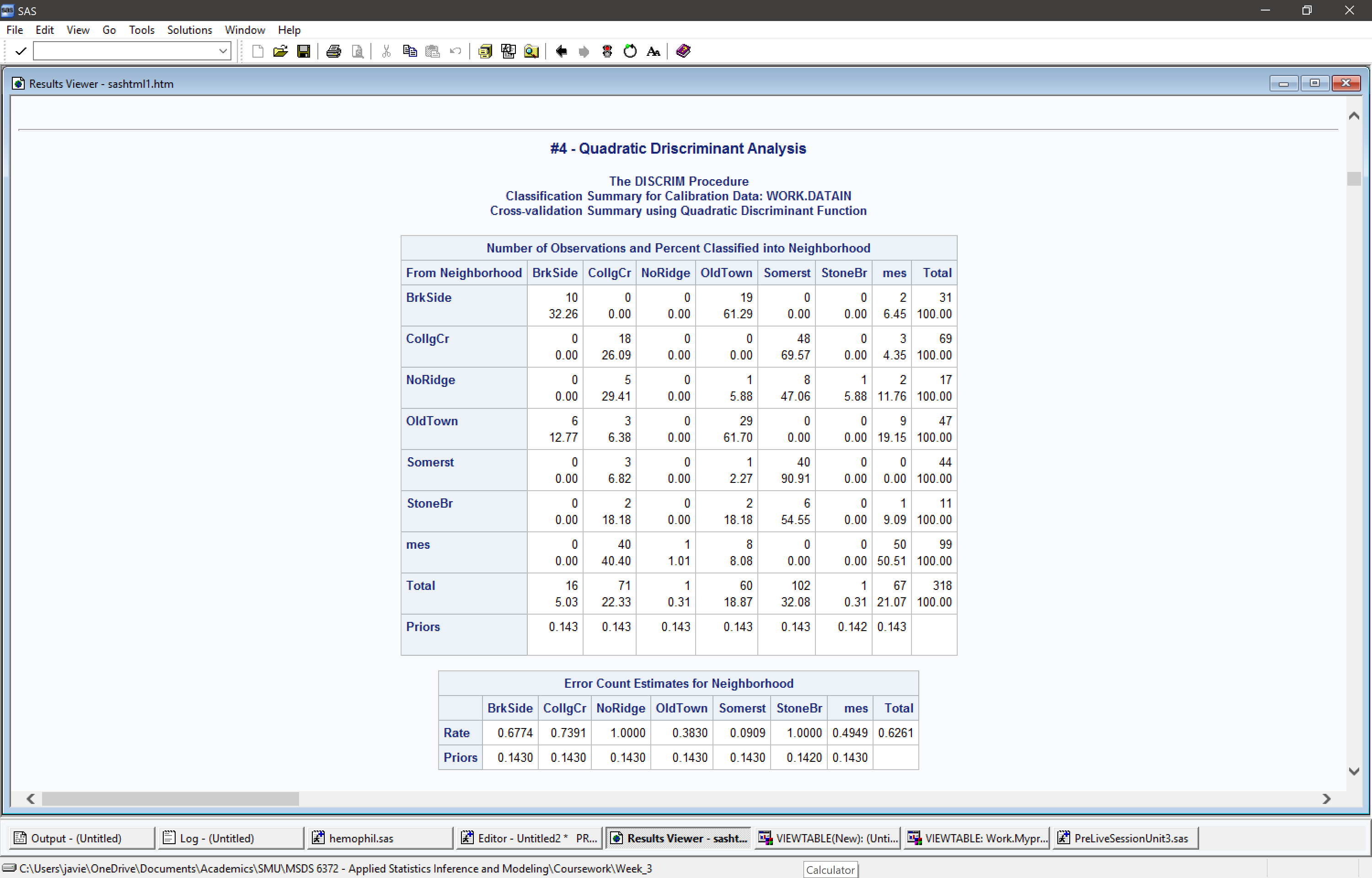
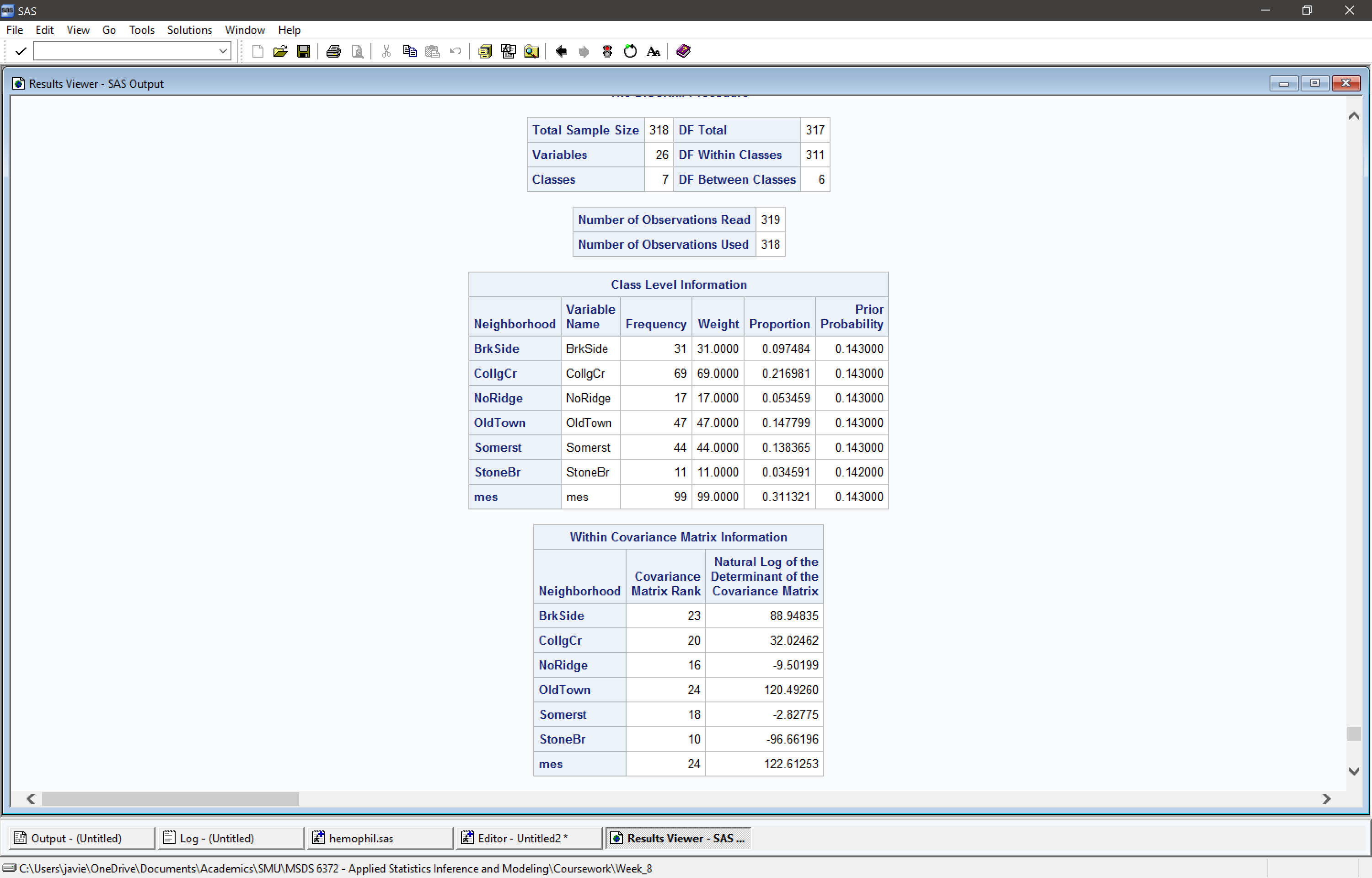
1. Next let’s perform a discriminant analysis (LDA or QDA) to help classify (discriminate between) neighborhoods using the variables above as explanatory variables (features).

Evaluate your model (LDA or QDA) with respect to mis-classification rate from a cross validation rather than a resubstitution. Copy and paste the cross validated confusion matrix and and mis-classification table as well as your code. Again, summarize your findings in at least one or two sentences. (Assume equal priors for the neighborhoods.)

Based on the output, it appears the model had a 100% error rate for StoneBr and NoRidge, and moderate error rate for the other neighborhoods. However, Somerst appears to have had the highest success rate with a 9% error rate. Overall, the model exhibited a 62% error rate, in which it misclassified the neighborhoods.







1. Finally, we would like to predict / classify / impute a house that we do not know what neighborhood in which it belongs. This house has a lot frontage of 52 ft, a lot area of 6000 sqft, an above ground living area of 1,400 sqft and a sale price of $110,000. Copy and paste the relevant table, your code and please include a short summary of your findings.

Based on the findings, there was a greater probability the neighborhood of the given dataset would be Somerst (49%). As a result, the model predicted the neighborhood was Somerst using QDA.



