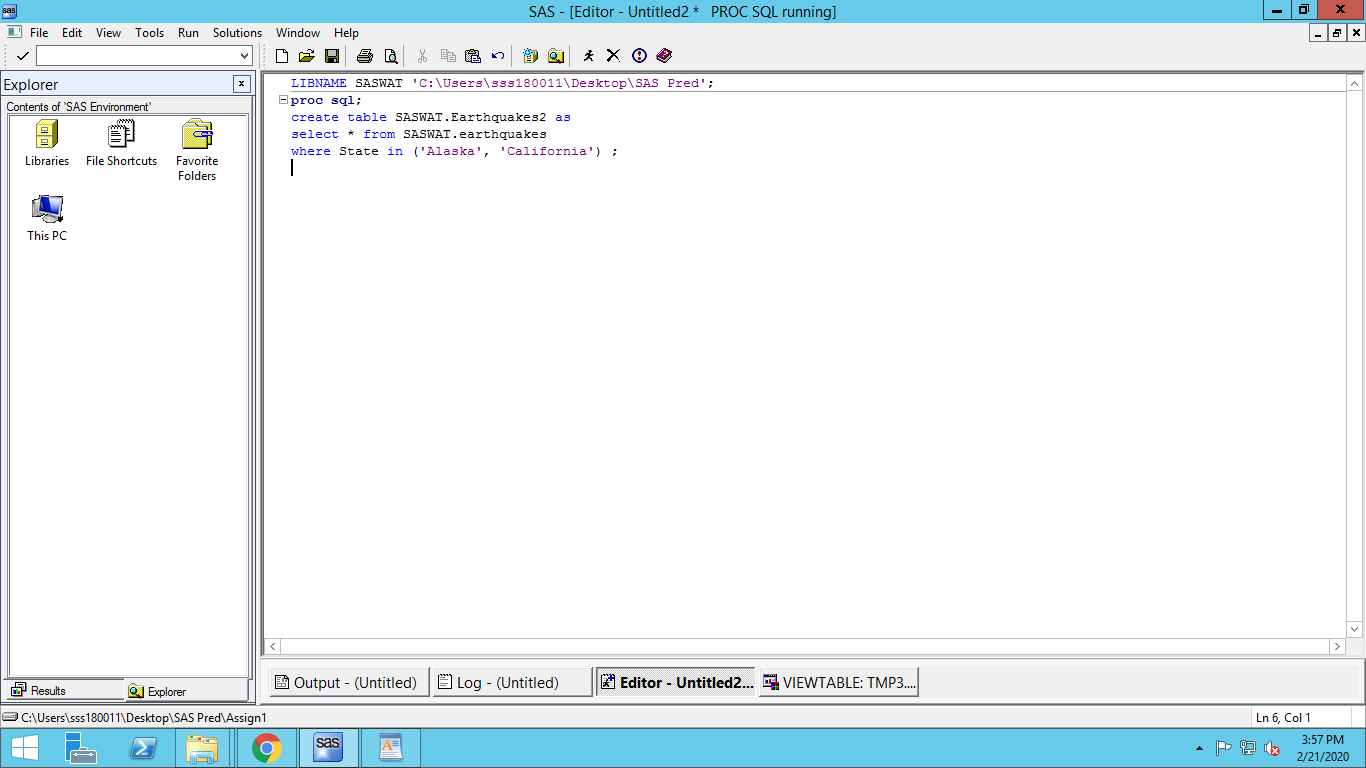
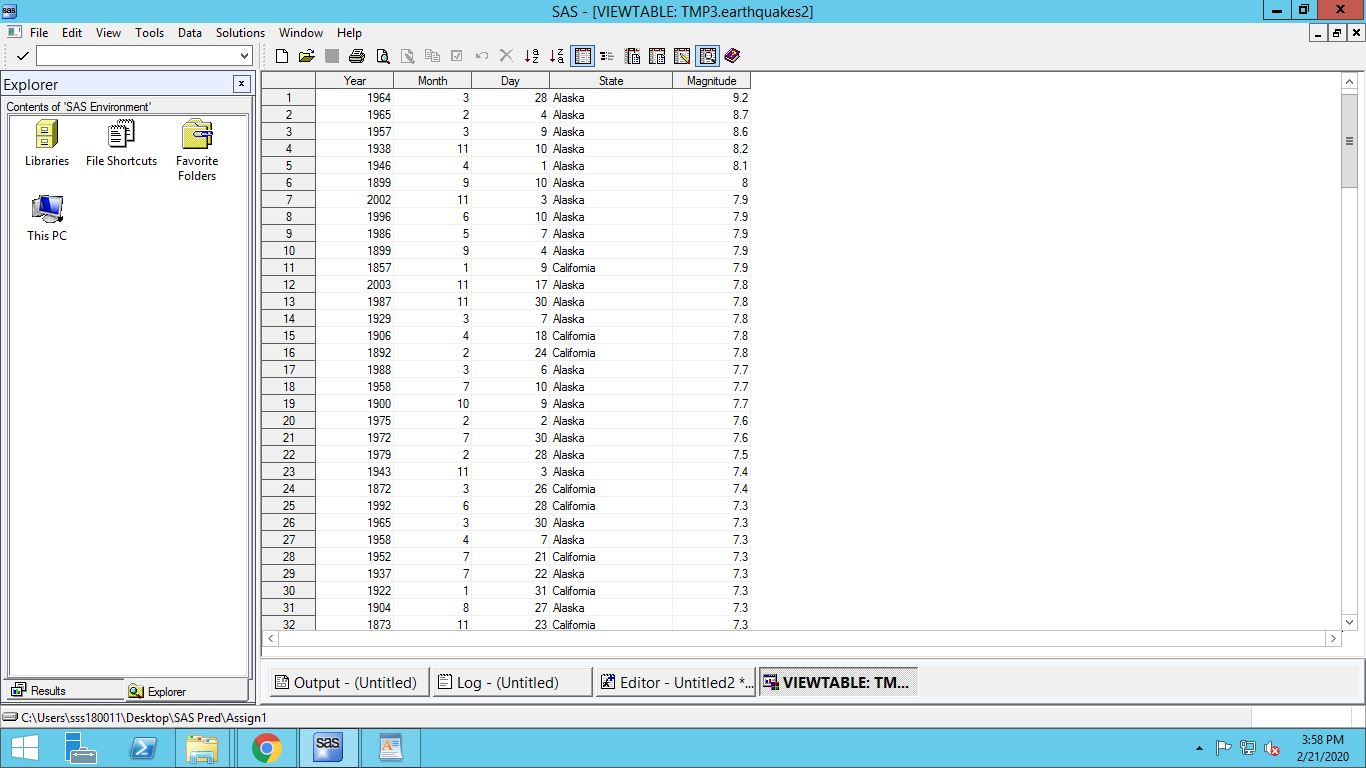
**Important Note1: For all questions below assume the test level is**

**Important Note2: For each question which asks you to do a hypothesis test, you need to (1) Write down your null and alternative hypothesis clearly and mention which test is appropriate in your word report, (2) Write a code in SAS to execute the test and (3) give the results of the test in your word report**

1. The United States Geological Survey provides data on earthquakes of historical interest. The SAS data set called EARTHQUAKES contains data about earthquakes with a magnitude greater than 2.5 in the United States and its territories. The variables are year, month, day, state, and magnitude.
2. California and Alaska are the two states with the highest number of earthquakes in the country. Create a new data set that includes only these two states and use this data set to answer the following questions.

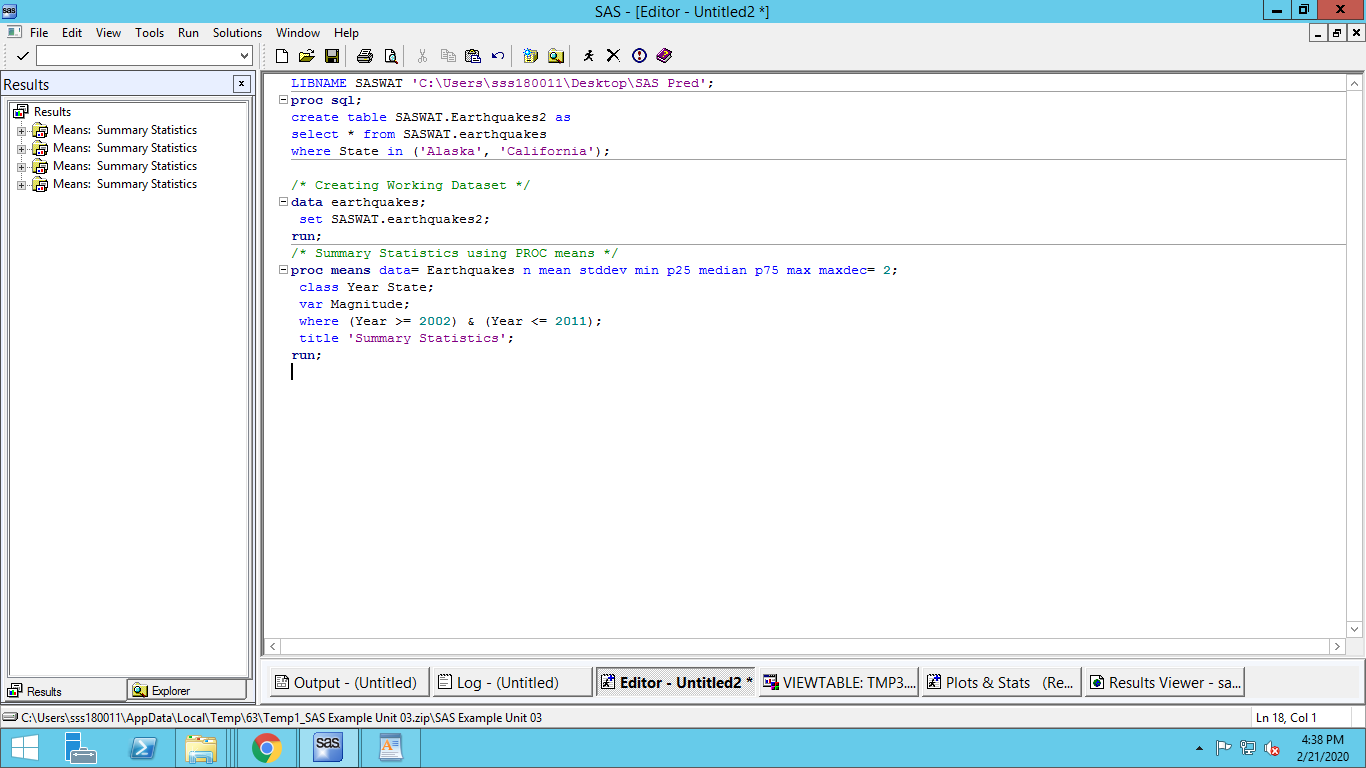


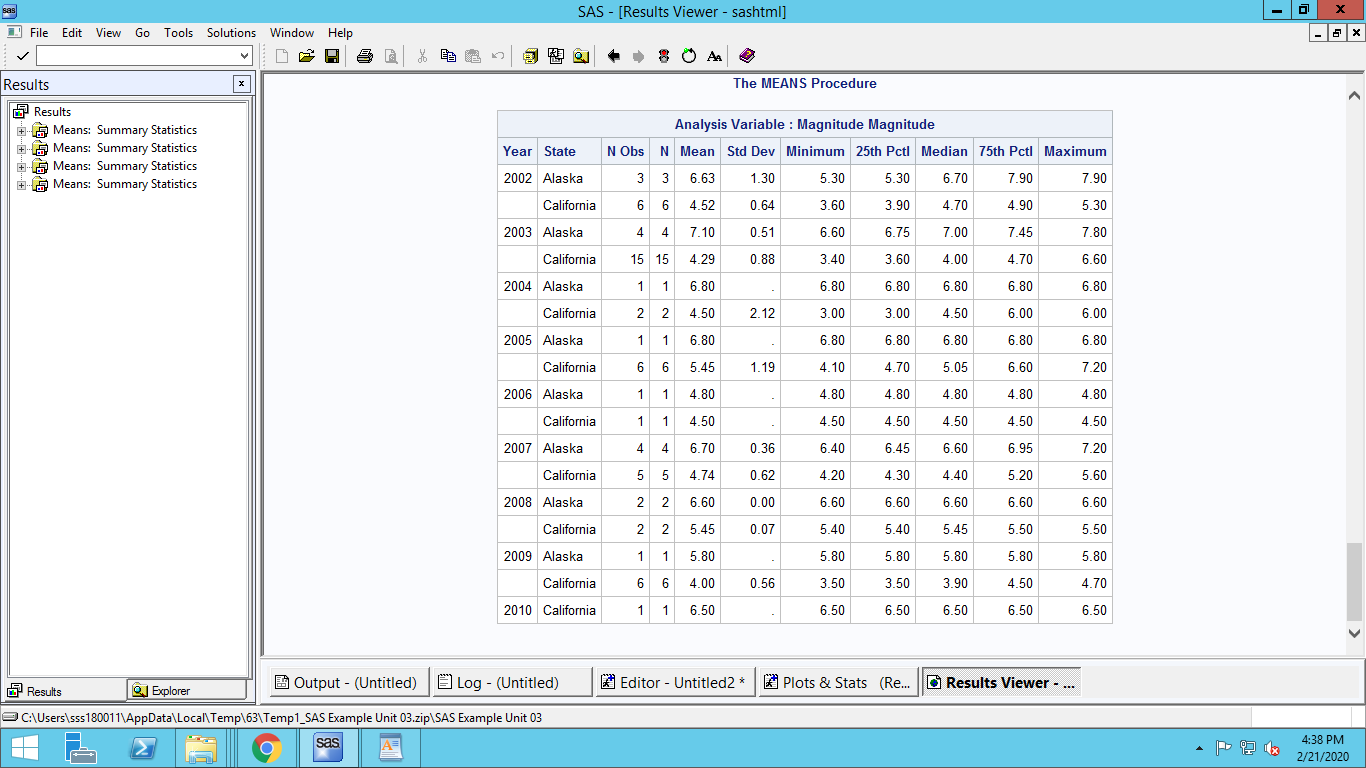


1. You are interested in the following statistics for the magnitude of earthquake:

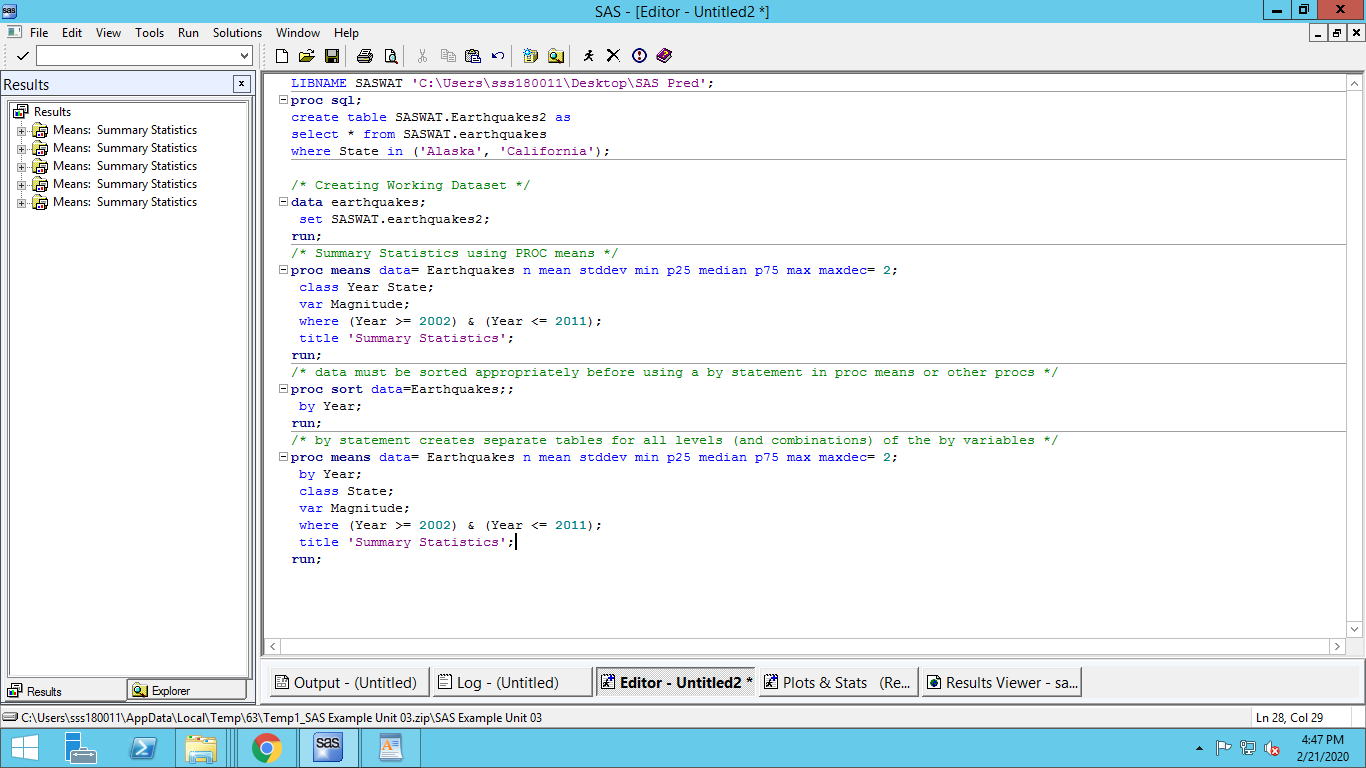
* Mean
* Median
* Standard deviation
* Minimum and maximum
* 25th and 75th percentiles

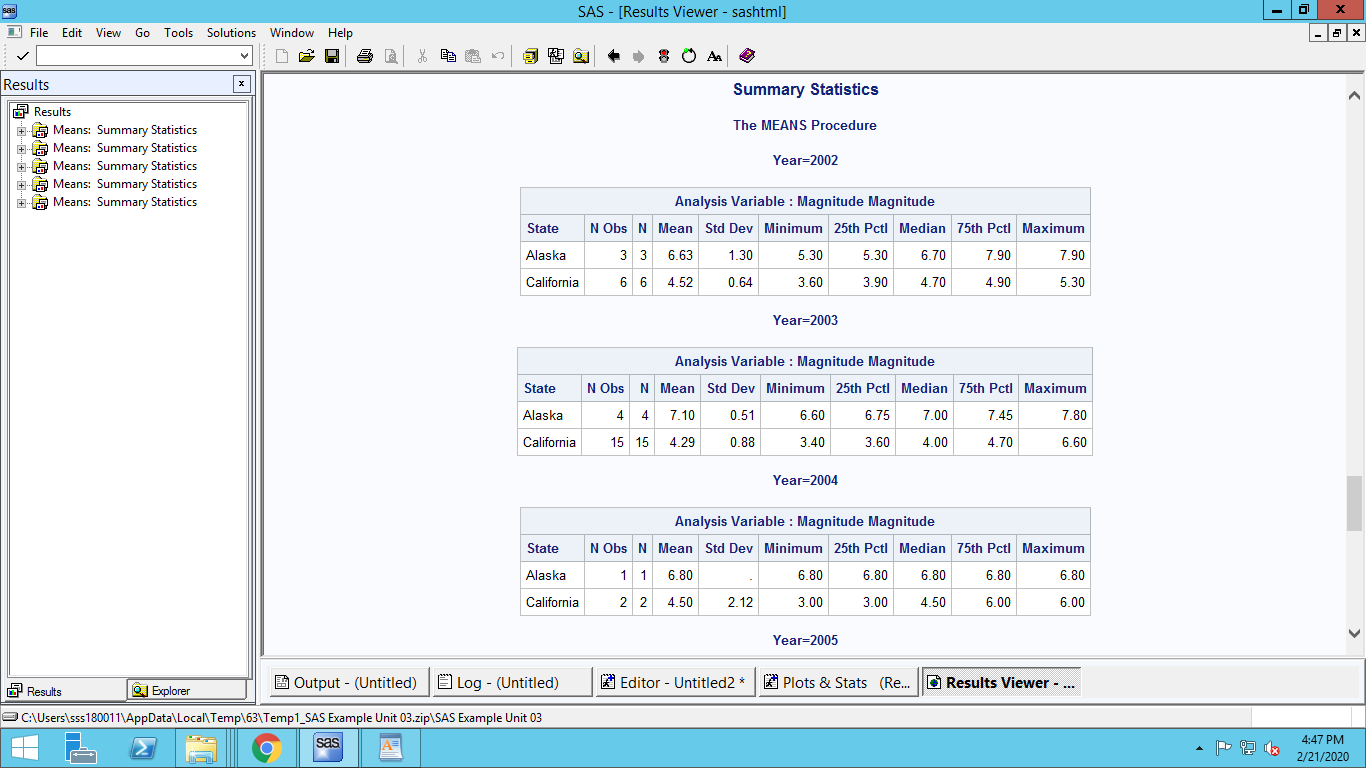
Create a table that shows the above statistics across different states within each year. In particular, your table must have years at the first column and it must break down the results across different states in the second column. In order to make the table short, further assume you are interested only in recent years and want to create a table that shows the desired statistics from 2002 to 2011.

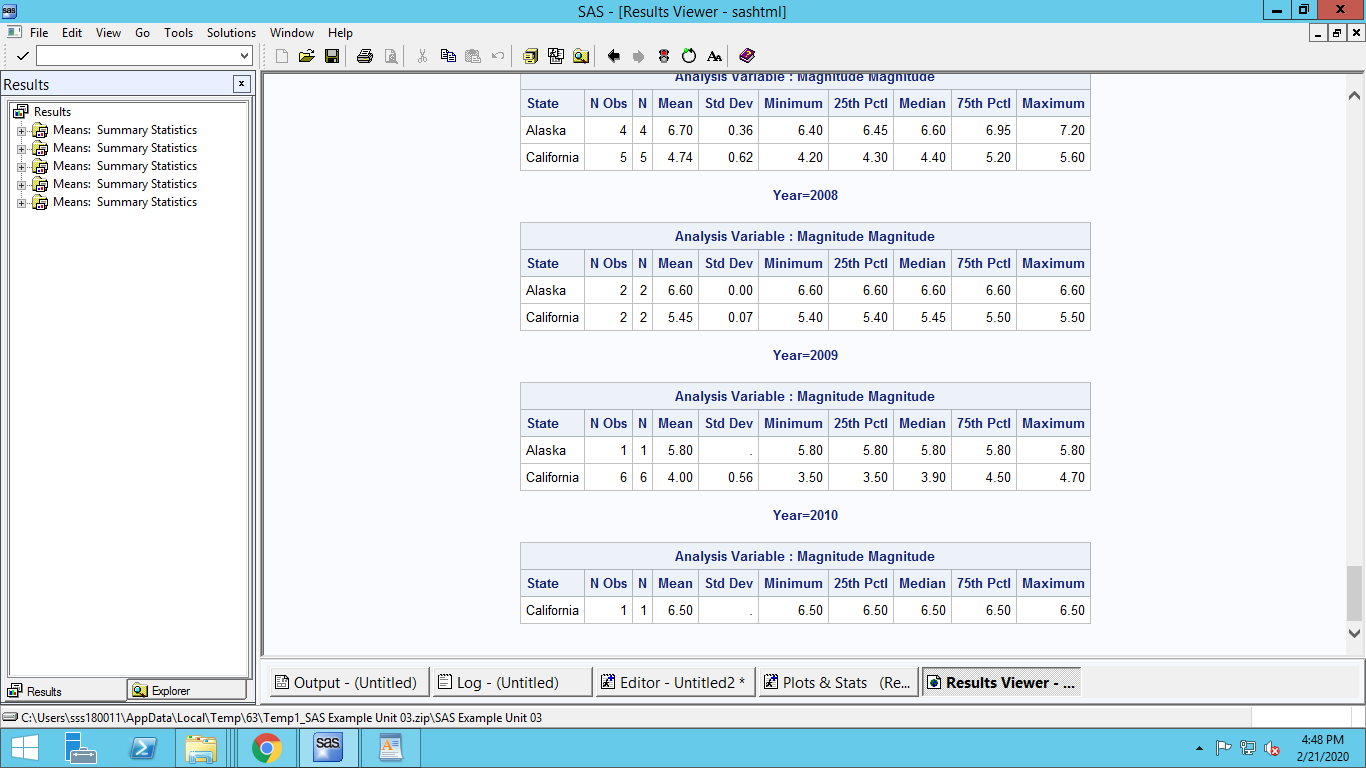




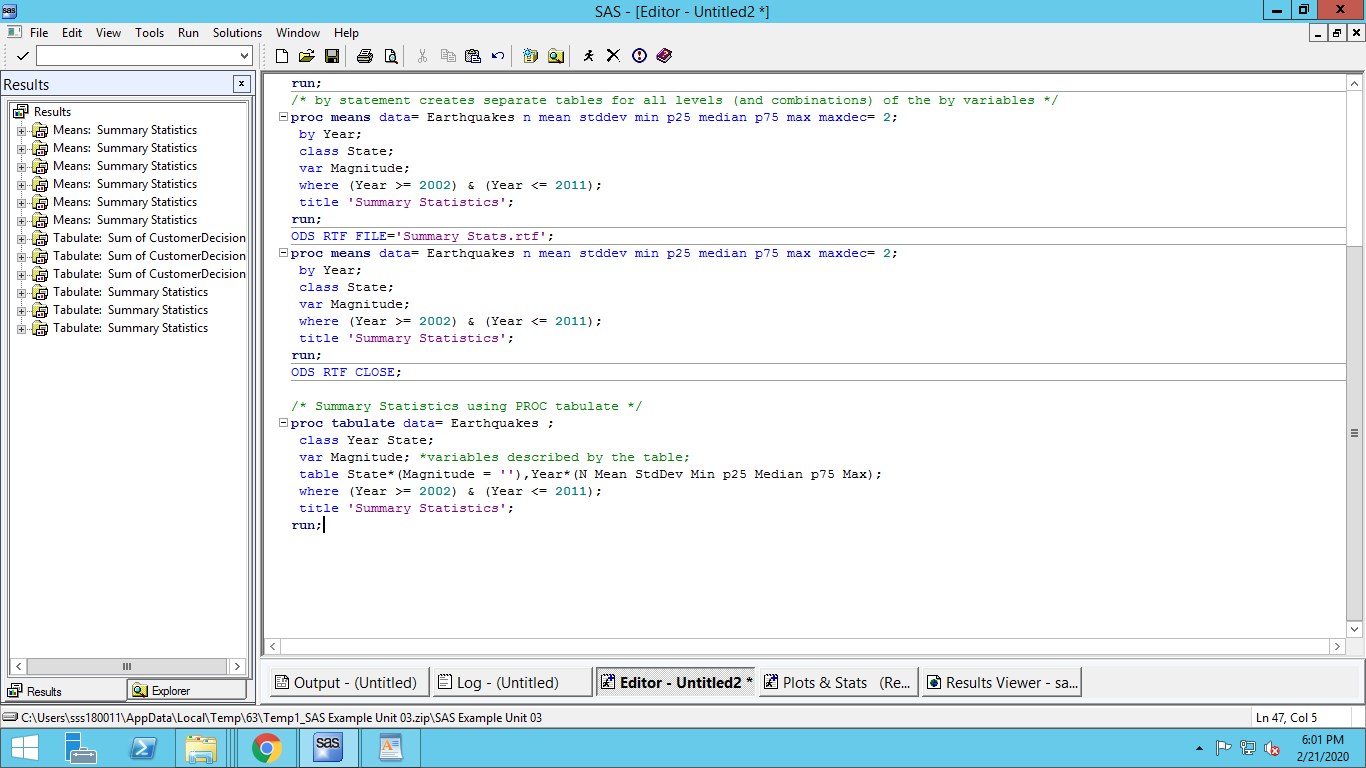
1. Modify you SAS code in (b) such that the results for each year is shown in a separate table.

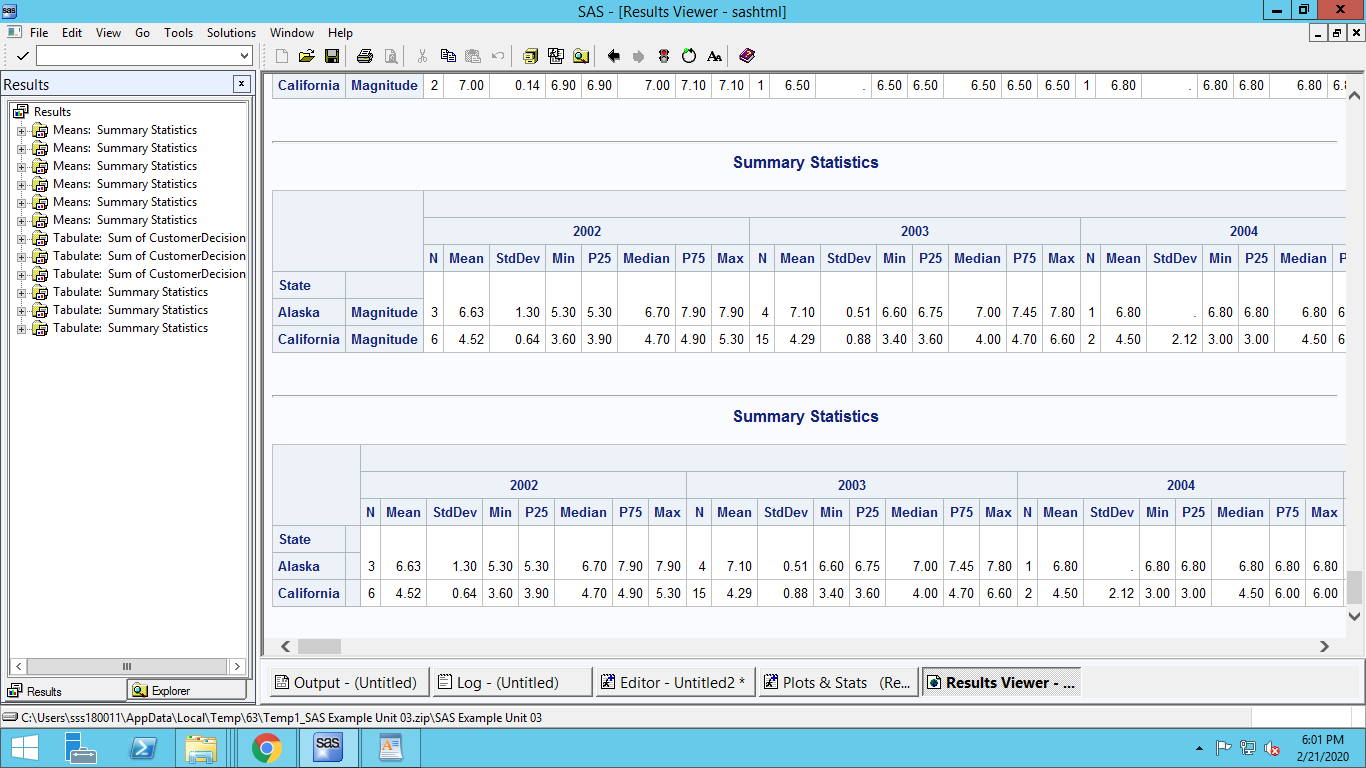


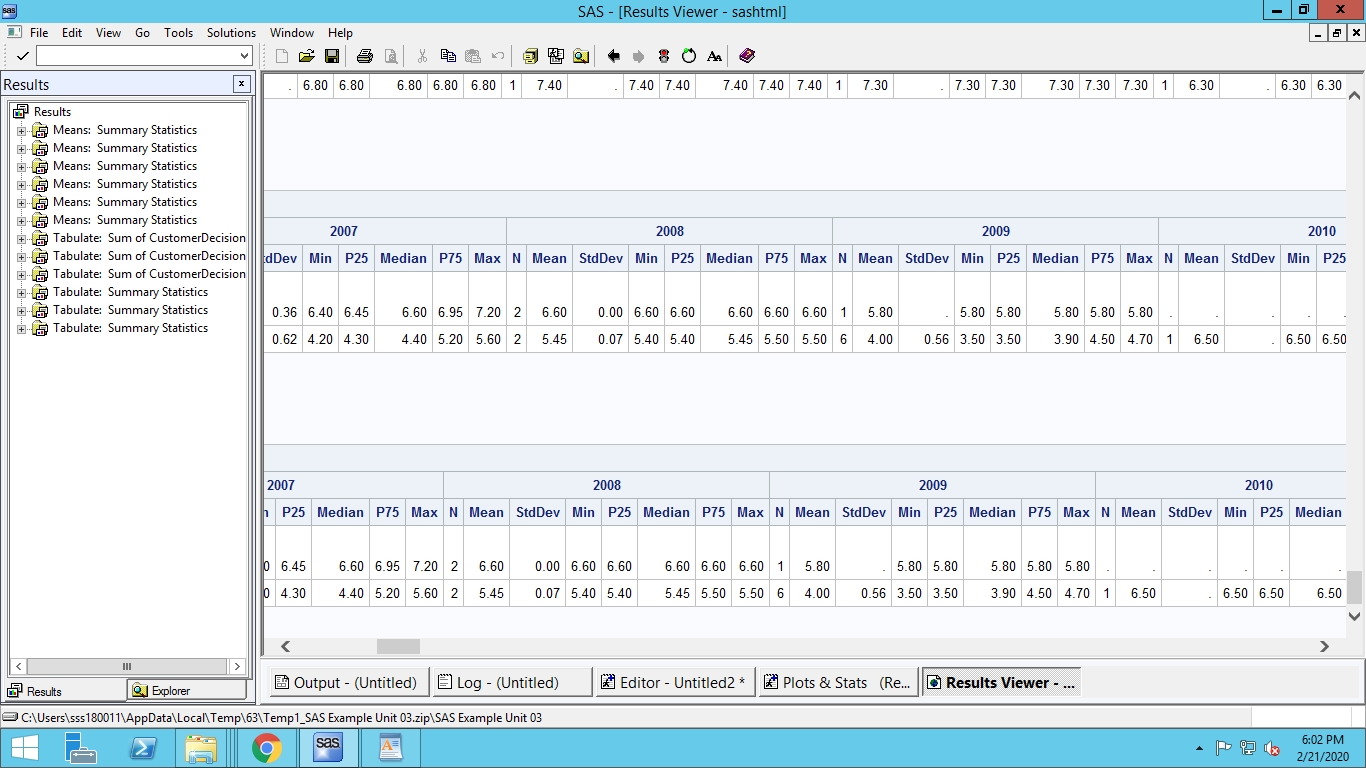




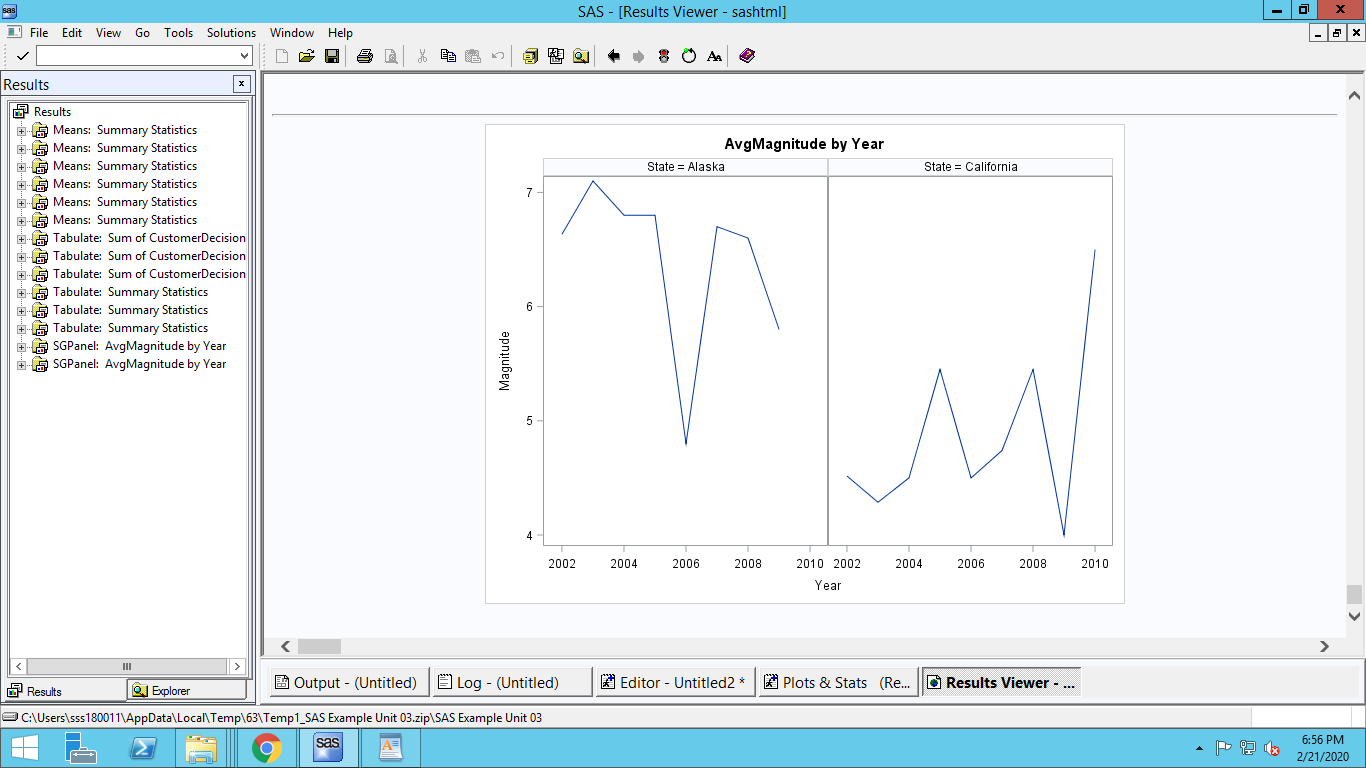
1. Now, assume you want to show the same results in part (b) but with the difference that years are shown is the first column and the states are shown in the top row.

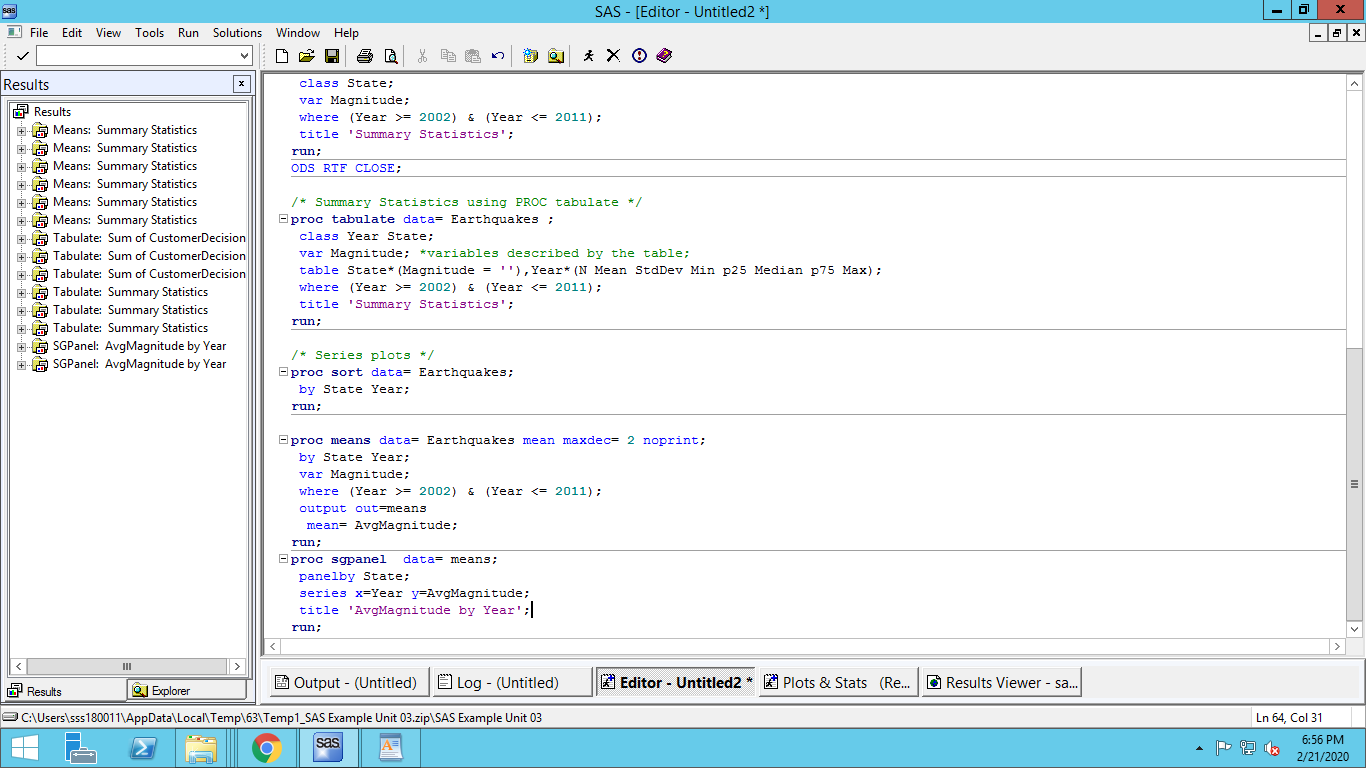






1. You are interested in how the magnitude of earthquakes is trending over time for each state. In one graph, plot two time series plots, side by side, which shows the trend of average magnitude of earthquakes over time for the two states.



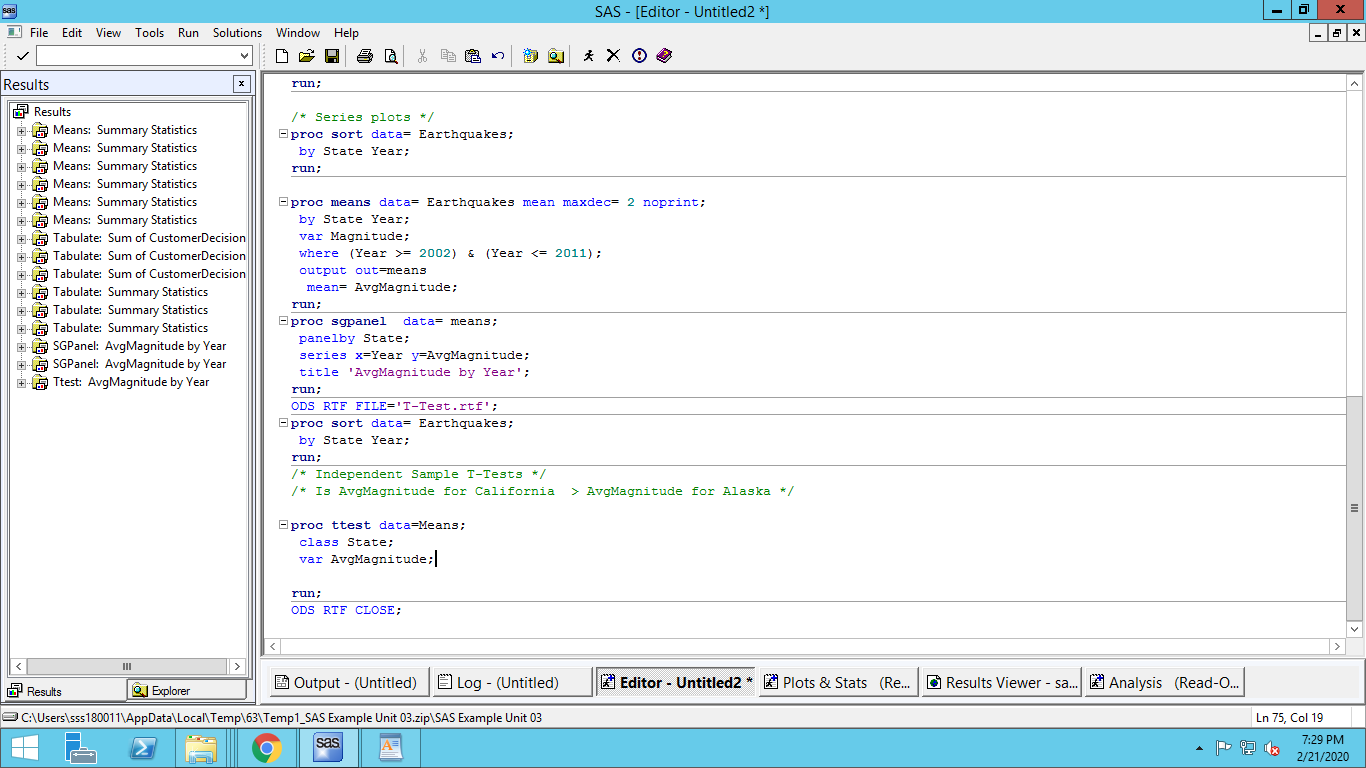


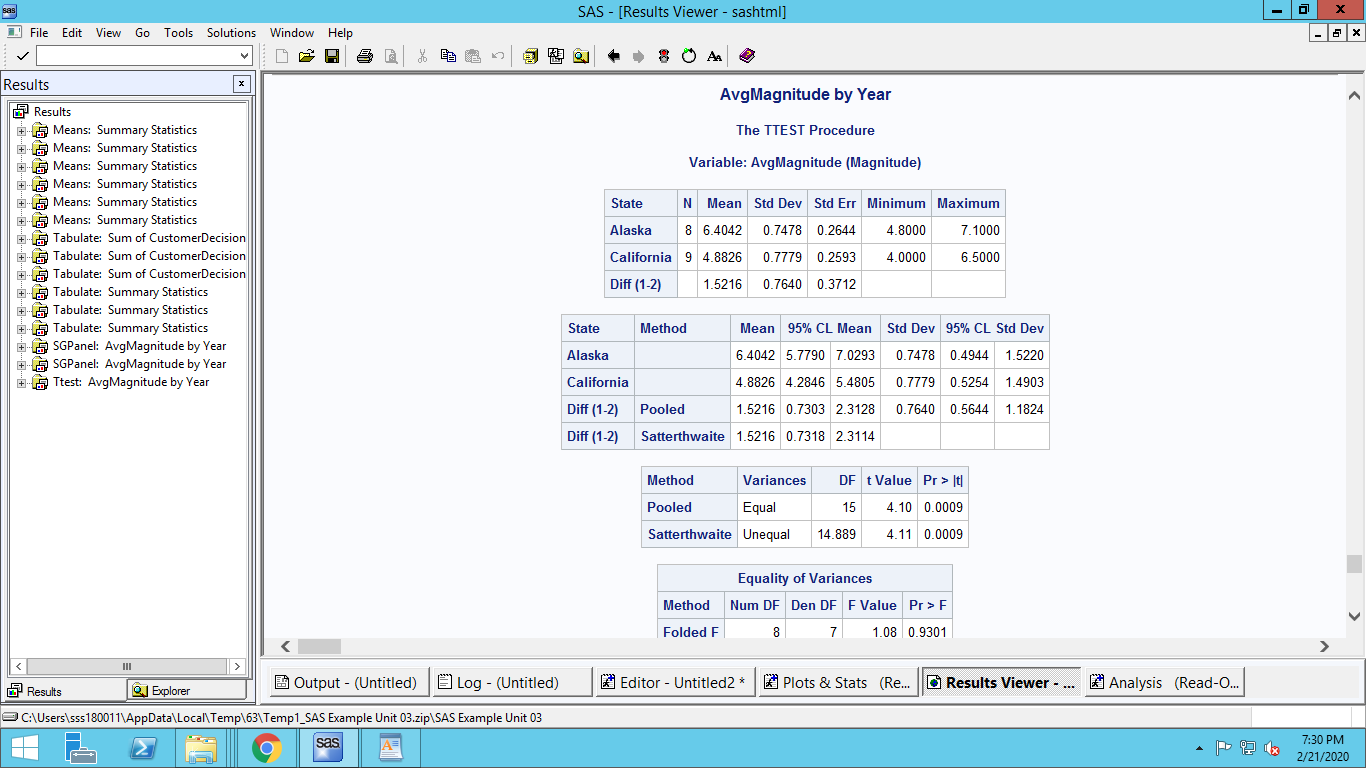
1. Test the following hypothesis: “the average magnitude of earthquakes in California is significantly higher than that in Alaska”.

**We need a Two-Sample T-Test for this hypothesis. This is since the samples are taken from two different population.**

**Null Hypothesis (H0): Avg Magnitude of earthquake in California less than or equal to Alaska**

**Alternate Hypothesis (Ha): Avg Magnitude of earthquake in California is more than Alaska**

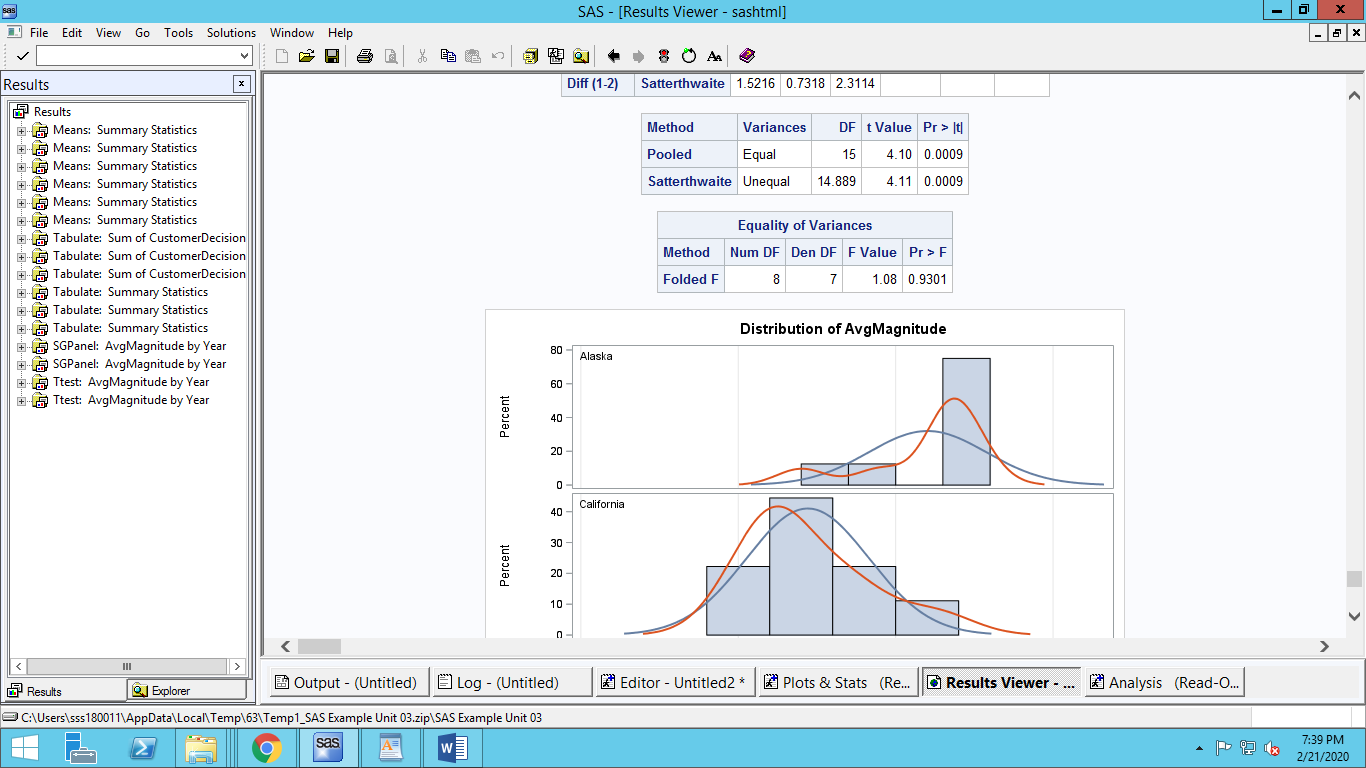




**Since there are 2 underlying populations from which Sample 1 and Sample 2 is drawn, we need to figure out the variances of these 2 populations are same or different**

**We need to do a hypothesis test of the population variance. This is done by a F-Test. The null hypothesis of this F-Test is ‘The Variances are equal’.**

**The table ‘Equality of Variances’ shows this:**

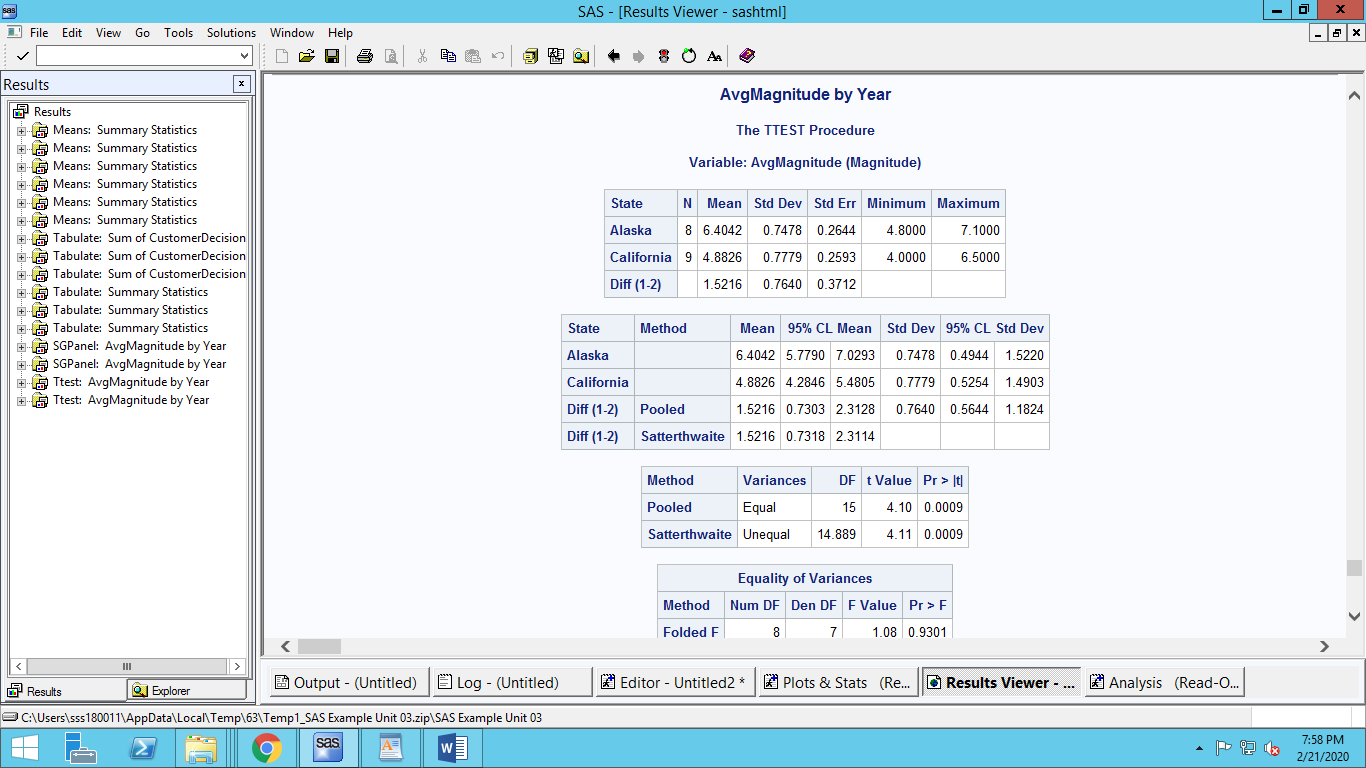


**Since the P value is not < 0.01 we fail to reject the null hypothesis (Variances are equal).**

**Hence use the equal variances t-test**

**In the pooled (Equal) method P value is less than 0.01, hence we reject the null Hypothesis.**

**This means that Average Magnitude of California and Alaska are statistically different.**

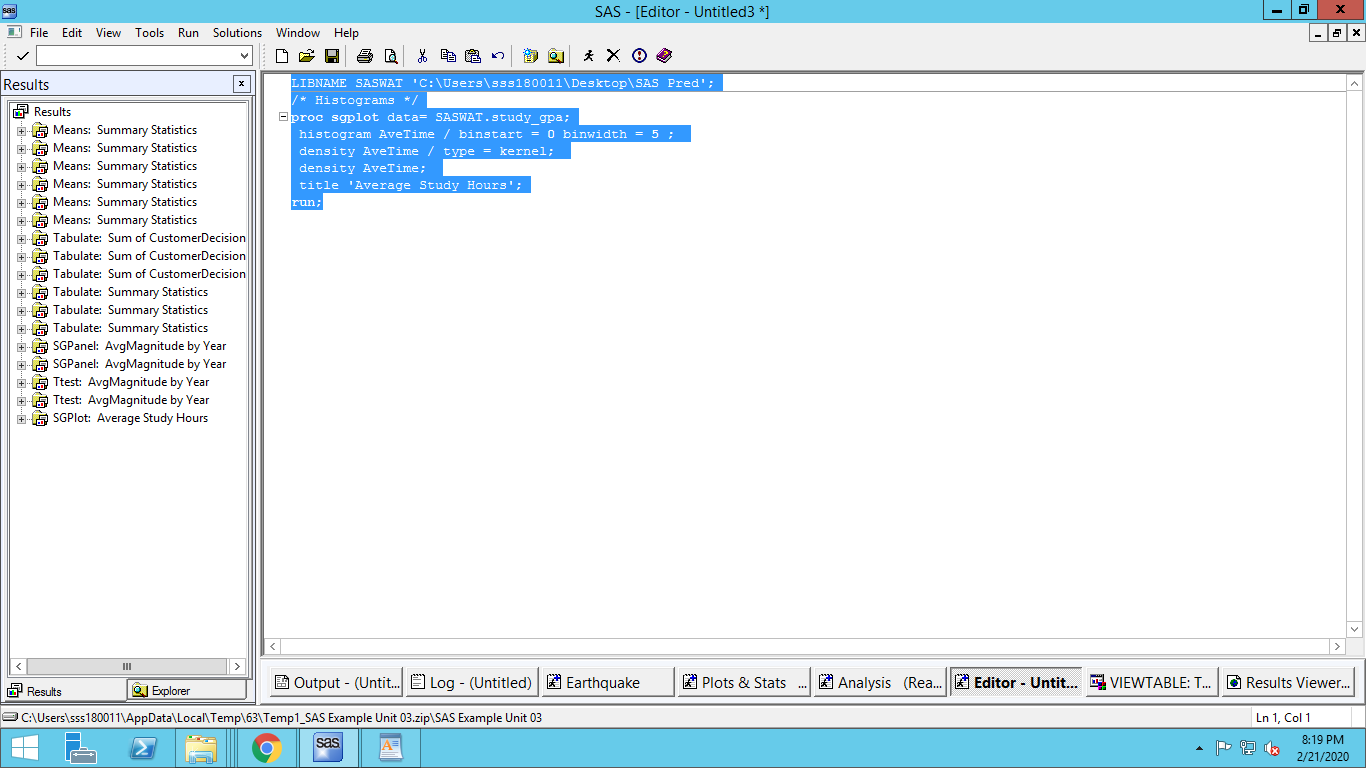


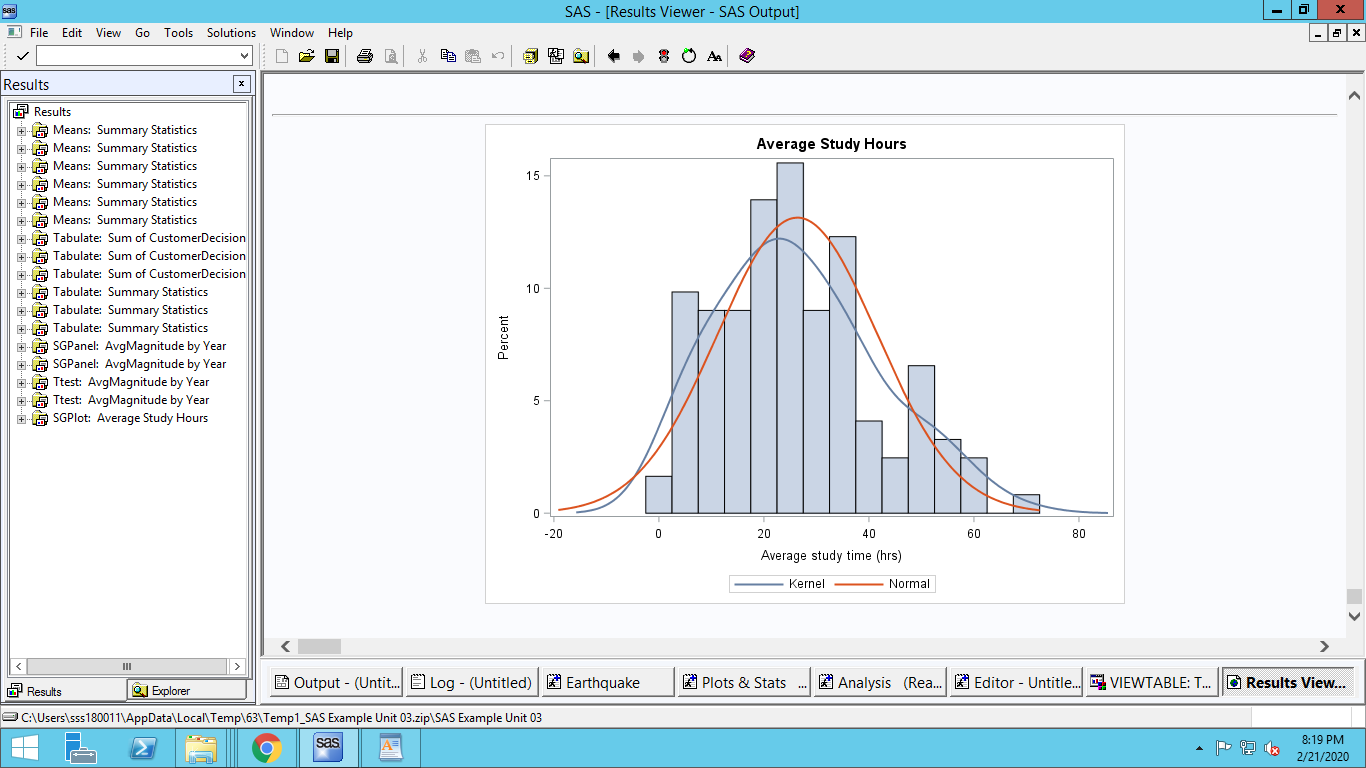
**We can see here mean of Alaska is higher than of California.**

**So the stated hypothesis that “**the average magnitude of earthquakes in California is significantly higher than that in Alaska” **is wrong.**

------------------------------------------------------------------------------------------------------------------------------------------

1. Suppose that at a local university the study guidelines for the College of Science and Math are to study two to three hours per unit per week. The instructor of the class, Orientation to the Statistics Major, takes these guidelines very seriously. He asks students to record their study time each week, and at the end of the term he compares their average study time per week to their term GPA. The SAS data set called STUDY\_GPA contains student identification information, orientation course-section number, number of units enrolled, average time studied, and term GPA.
2. Graph the histogram for hours of study. Use the start point=0 and bandwidth=5. Also, overlaid to this graph, display the plots for the kernel density and the best fitting normal curve. Using an eyeballing approach, can we say the hours of study follows a normal distribution?





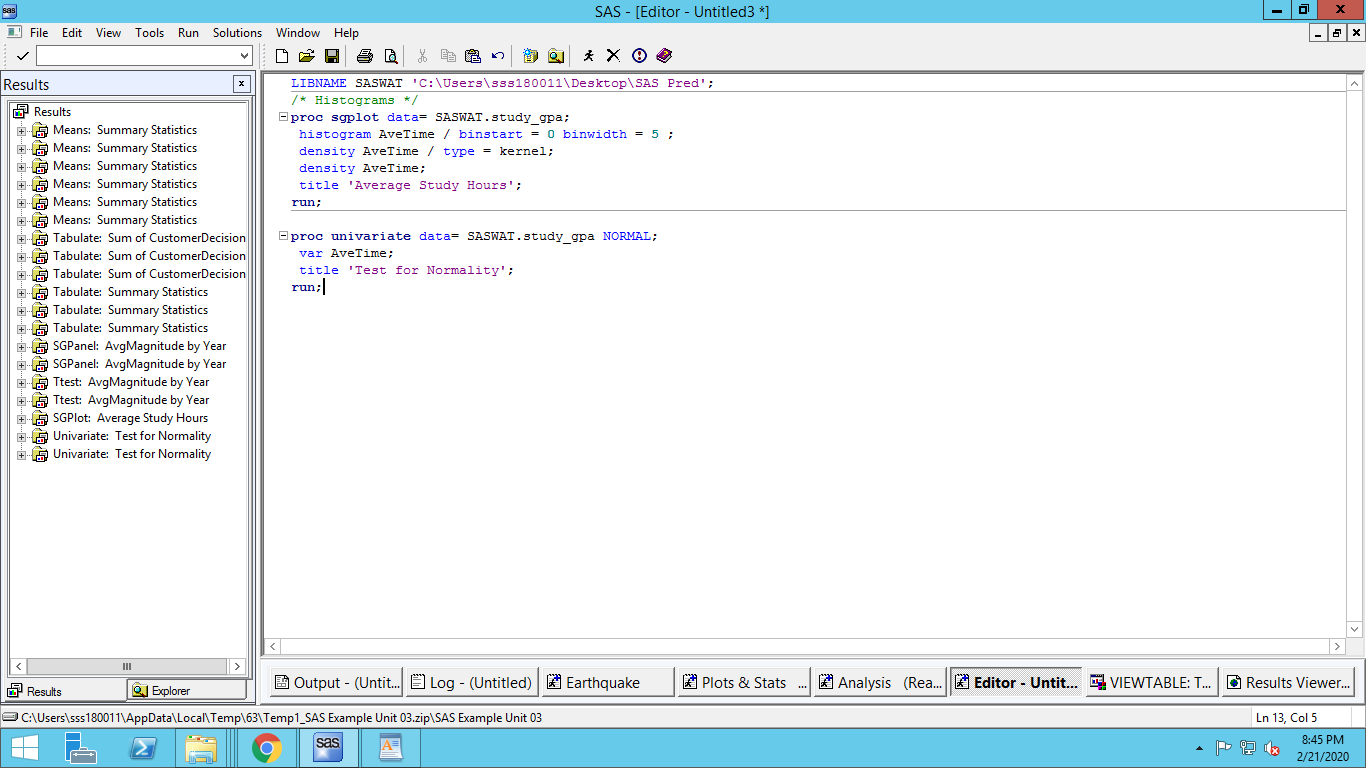
**Just by eyeballing we can say hours of study follows a normal distribution**

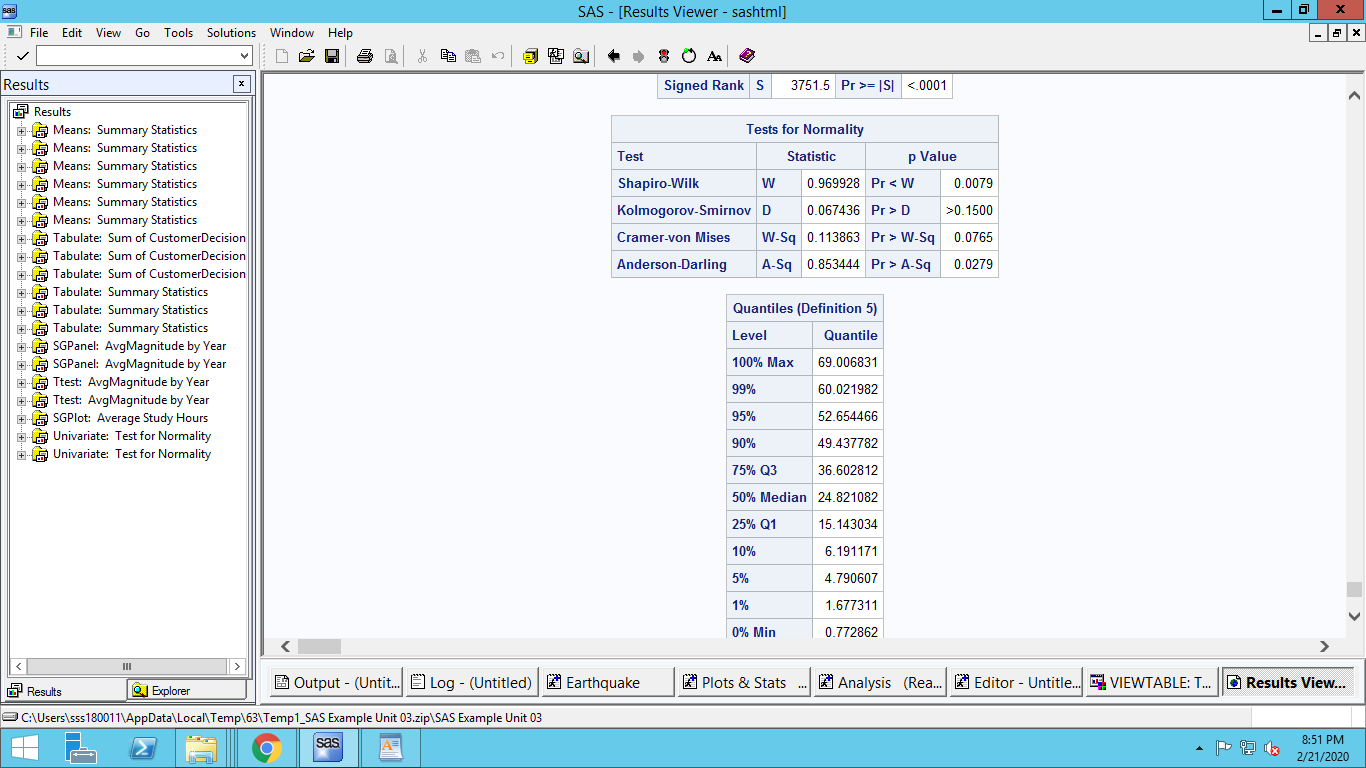
1. Now, suppose you want to test the normality not just by eyeballing. Conduct a statistical test to check whether the hours of study follows a normal distribution.

**For a Statistical test we use PROC UNIVARIATE.**

**Null Hypothesis(H0): The distribution is normal**

**Alternate Hypothesis(Ha): Distribution is not normal**





**Kolmogorov-Smirnov, Cramer-von Mises or Anderson-Darling fail to reject the null hypothesis that the distribution is normal on the alpha=0.01 level.**

**Shapiro-Wilk rejects the null hypothesis that the distribution is normal on the alpha=0.01 level.**

**Since majority of the tests fail to reject the null hypothesis, we can conclude that the distribution is normal.**

The next part focuses on the difference between causality and correlation. In most managerial situations, we are interested in causality rather than just correlation. Think carefully when you answer the next part. There may not be a definite correct answer when it comes to the causality argument. You will be given the credit as long as your reasoning is plausible.

1. Conduct a hypothesis test to check whether there exists a significance correlation between units enrolled, hours of study and GPA for section 1. What is your conclusion? What variable you think may cause the other?

**Pearson:**

**Between Units Taken and GPA, correlation is statistically significant. But practically speaking the correlation value of -0.15 is not that good.**

**Between Average time and Units taken, correlation is statistically significant. But practically speaking the correlation value of 0.425 is not that good.**

**Between Average time and GPA, correlation is statistically significant. But practically speaking the correlation value of -0.34 is not that good.**

**Spearman:**

**Between Units Taken and GPA, P value is more than 0.01. So, correlation is not statistically significant.**

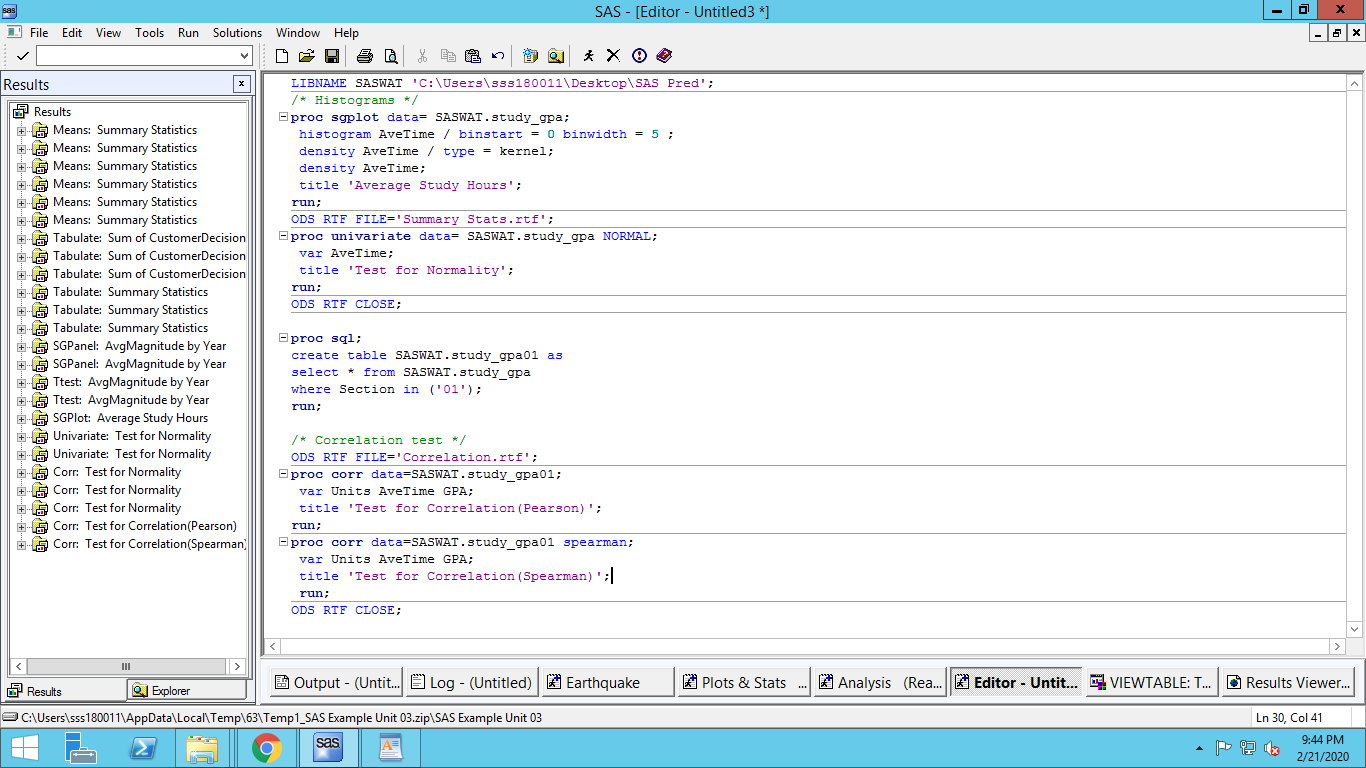
**Between Average time and Units taken, correlation is statistically significant. But practically speaking the correlation value of 0.412 is not that good.**

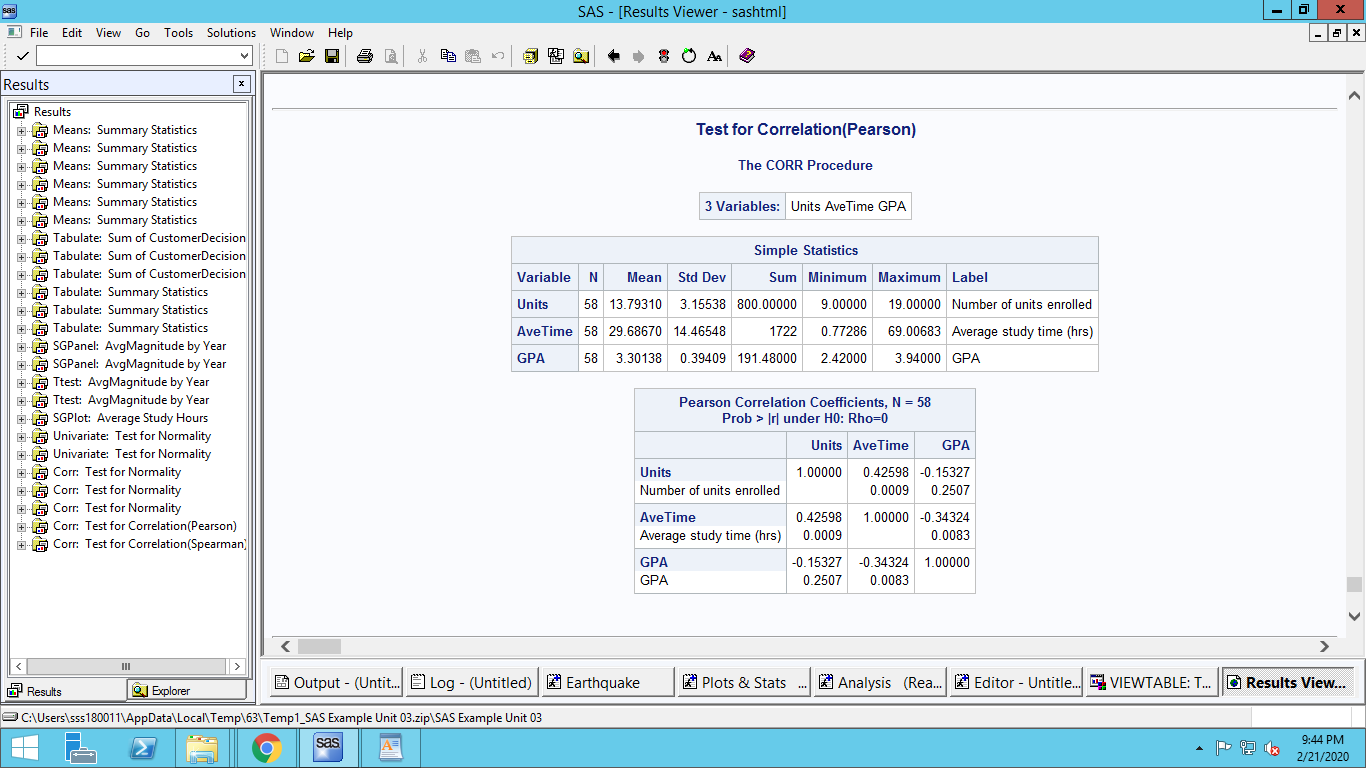
**Between Average time and GPA, correlation is statistically significant. But practically speaking the correlation value of -0.348 is not that good.**

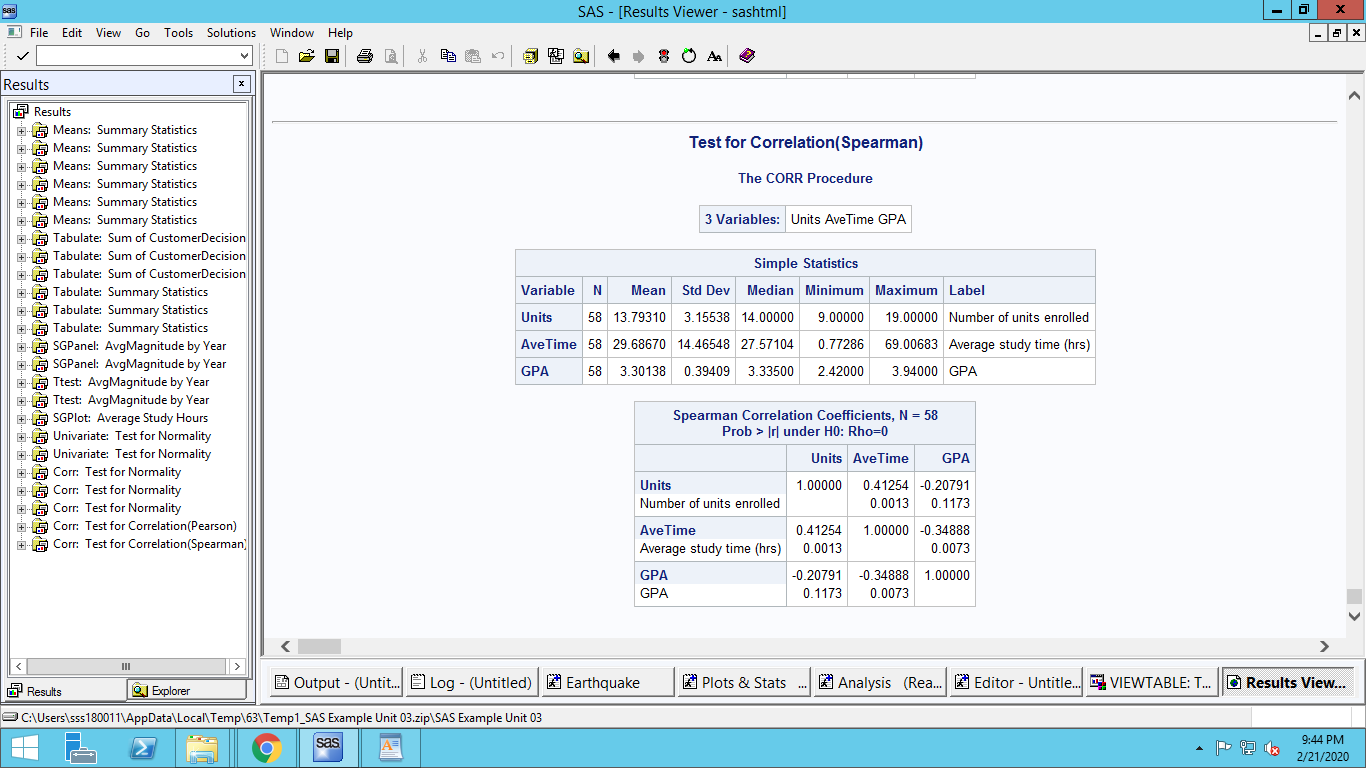
**Out of all correlations, Average time and Units taken is higher than the others. This is obvious, since more units require more time to be invested.**

**Between Average time and GPA correlation is less in absolute terms than Average time and Units taken. But it is negative. This maybe because brighter students require less time to study. Or, students with less units invest less time to score the same GPA as compared to someone enrolled in more credits.**

**Between Units taken and GPA, Pearson correlation test shows statistical significance but not Spearman test. The value is negative and very less in absolute terms. This is reasonable since more units can lead to lower grades if proper time is not invested accordingly.**







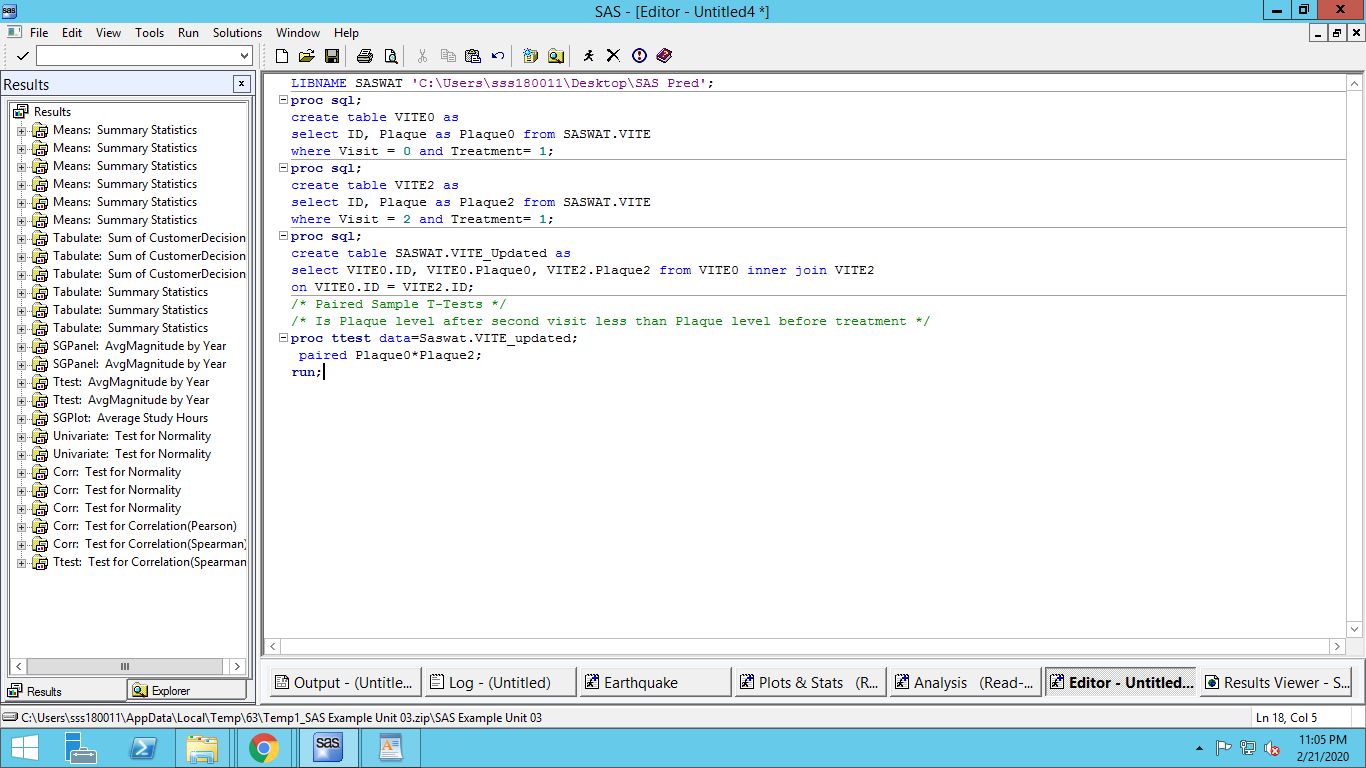
------------------------------------------------------------------------------------------------------------------------------------------

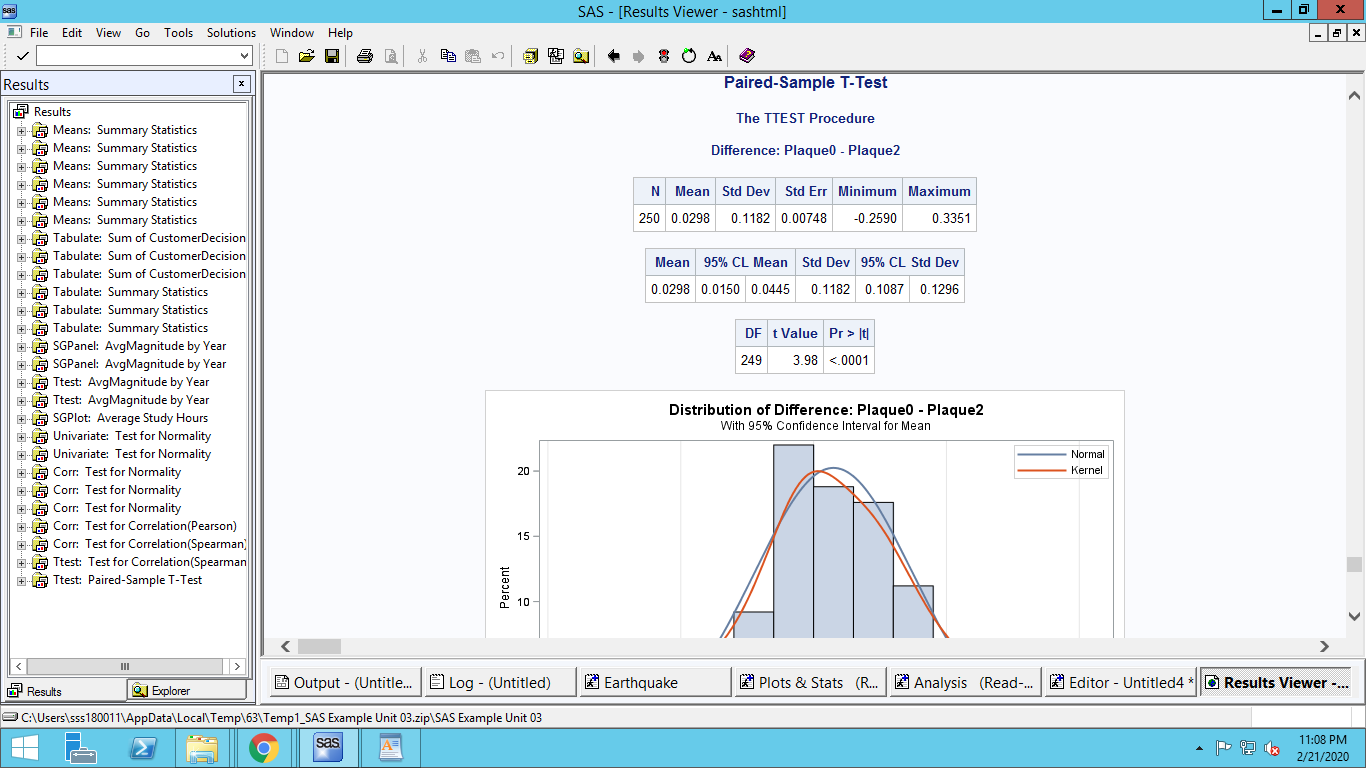
1. A study was conducted to see whether taking vitamin E daily would reduce the levels of atherosclerotic disease in a random sample of 500 individuals. Clinical measurements, including thickness of plaque of the carotid artery (taken via ultrasound), were recorded at baseline and at two subsequent visits in a SAS data set called VITE. Patients were divided into two strata according to their baseline plaque measurement.
2. Assumer there were no placebo (i.e., control) group in your data set. Conduct a test to see whether there is a difference in plaque level before treatment and after the second visit?

**Since samples are not independent (Same individuals are sampled twice), One-tailed paired sample test is needed for this.**

**Null Hypothesis(H0): Plaque level after second visit >= Plaque level before treatment**

**Alternate Hypothesis(Ha): Plaque level after second visit < Plaque level before treatment**





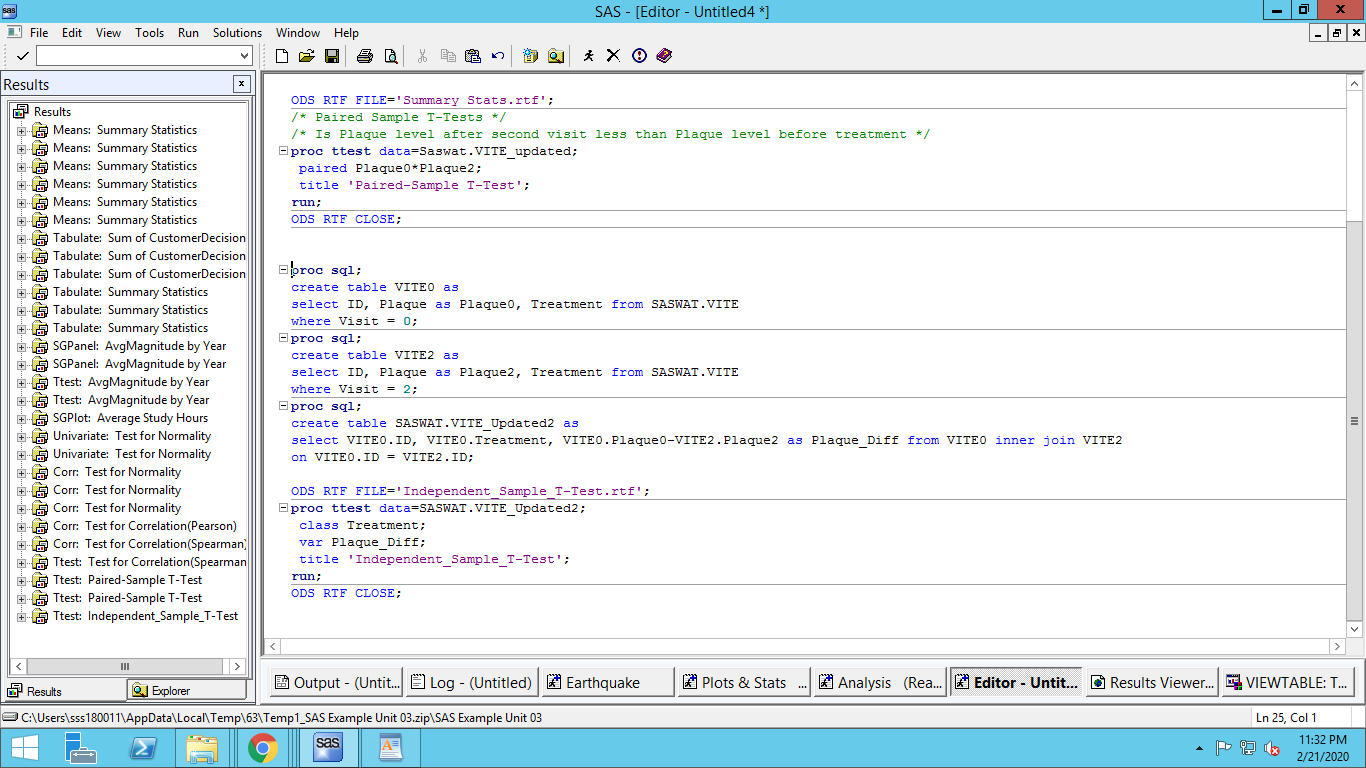
**0 is not in the 95% Confidence interval mean. And, P < 0.01. Therefore, we reject the null hypothesis that Plaque level after second visit remains same as baseline plaque level value.**

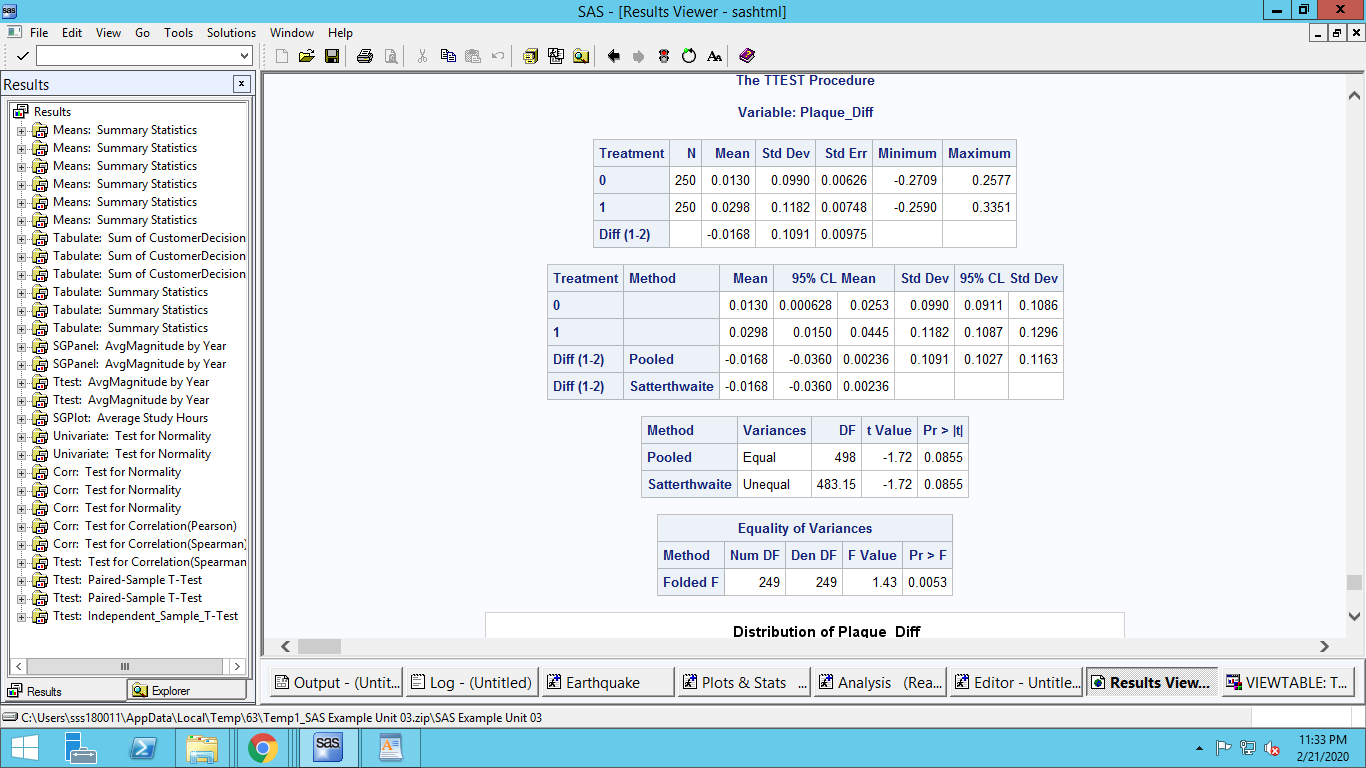
1. Now, considering the fact that there is indeed a control group in your dataset, conduct a new test to check whether there is a difference in plaque level before treatment and after the second visit.

**We need a Two-Sample T-Test for this hypothesis. This is since the samples are taken from two different population.**

**Null Hypothesis (H0): Diff in Plaque level in Placebo >= Diff in Plaque level in non-Placebo group**

**Alternate Hypothesis (Ha): Diff in Plaque level in Placebo < Diff in Plaque level in non-Placebo group**





**Since there are 2 underlying populations from which Sample 1 and Sample 2 is drawn, we need to figure out the variances of these 2 populations are same or different**

**We need to do a hypothesis test of the population variance. This is done by a F-Test. The null hypothesis of this F-Test is ‘The Variances are equal’.**

**Since the P value is < 0.01 we reject the null hypothesis (Variances are equal).**

**Hence use the unequal variances t-test**

**In the Scatterwaithe (UnEqual) method P value is more than 0.01, hence we fail to reject the null Hypothesis.**

**This means that difference in Plaque level in Placebo is not statistically different than difference in Plaque level in non-Placebo group**

1. Which of the tests in part (a) and (b) is more reliable? Explain.

**The second test (b) is more reliable since in the first (a) test we did not take into account the control group. Even though in (a) we found there is a statistically significant difference in Plaque level before and after 2nd visit, the difference between this difference among control and non-control groups is not statistically significant.**

1. One of the critical factors in randomizing the subjects in control and treatment groups is to make sure that the subject are perfectly randomized in all aspects. Using the last two columns (i.e., alcohol and cigarette usage), conduct two tests to check whether subjects are randomized perfectly.

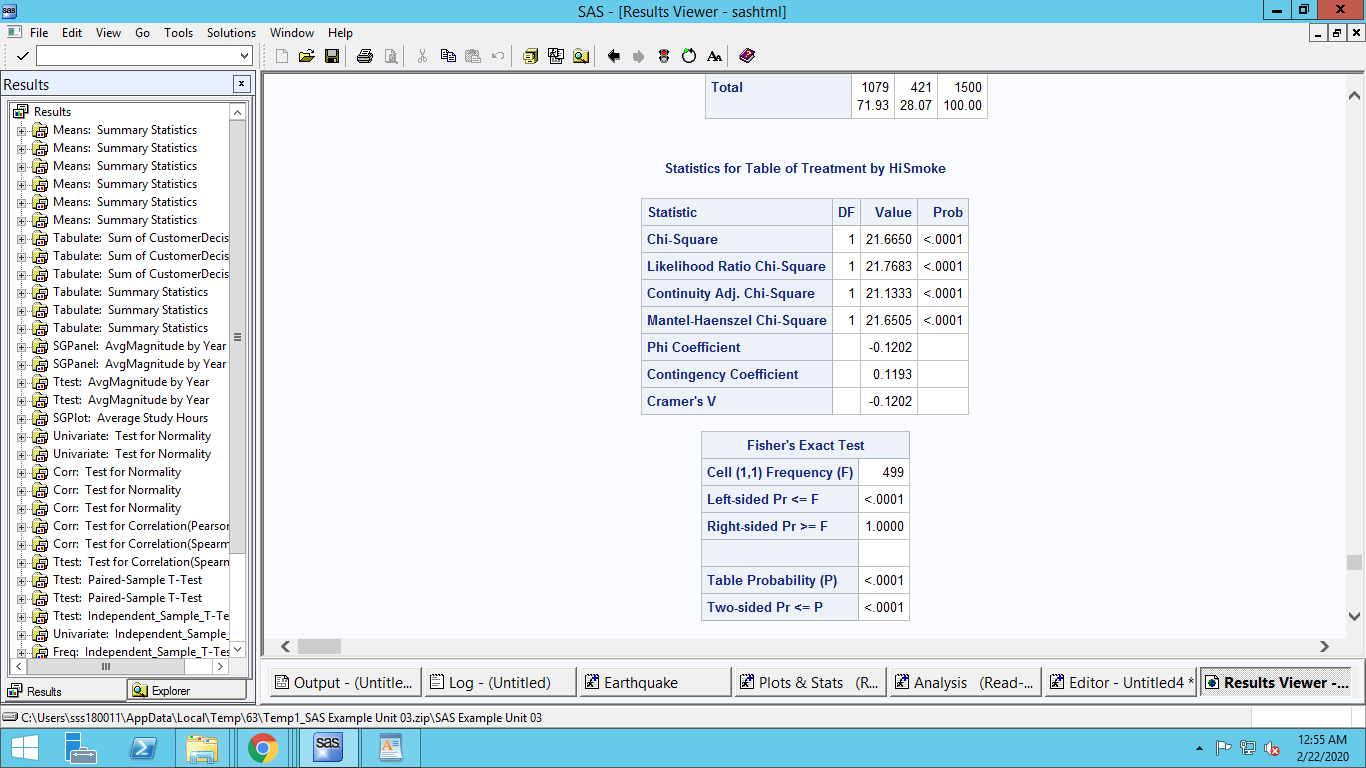
**Categorical Variables are created to determine high smokers and high drinkers**

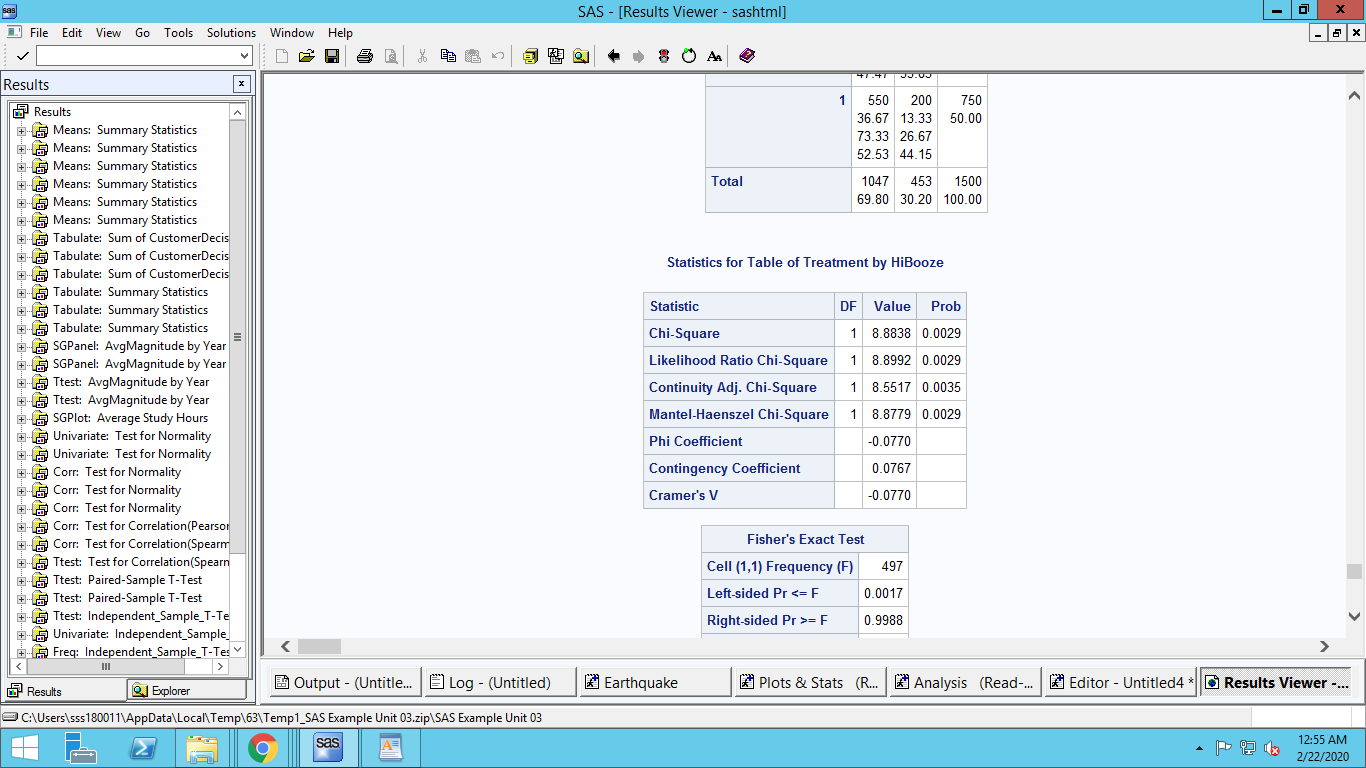
**Step1:**

**Chi-Square test of independence to check whether High smokers/drinkers related to Treatment conditions**

**Null Hypothesis(H0): High Smoker/Drinker and treatment conditions are independent**

**Alternate Hypothesis (Ha): High Smoker/Drinker and treatment conditions are related**





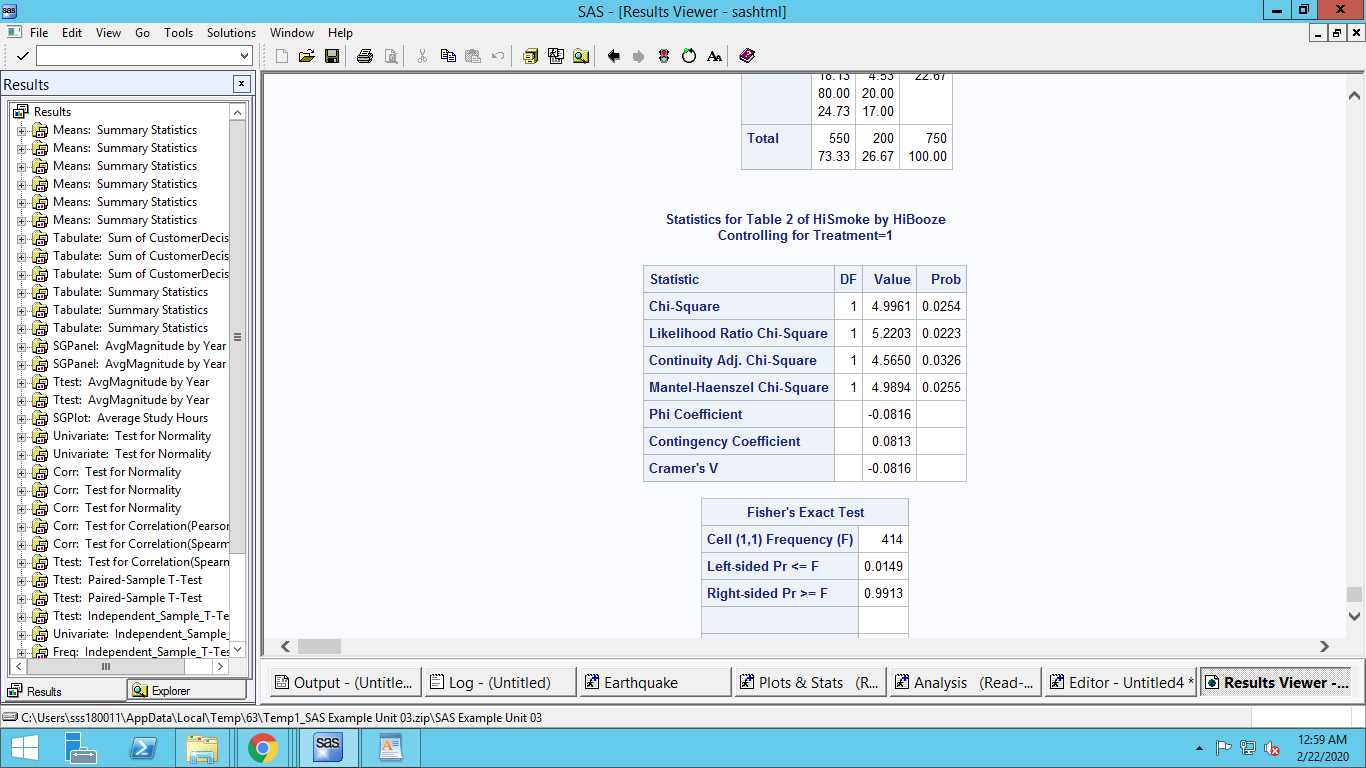
**Since P<0.01, In both the cases we reject the null hypothesis that High Smoke/Drink is independent to Treatment condition**

**Step2:**

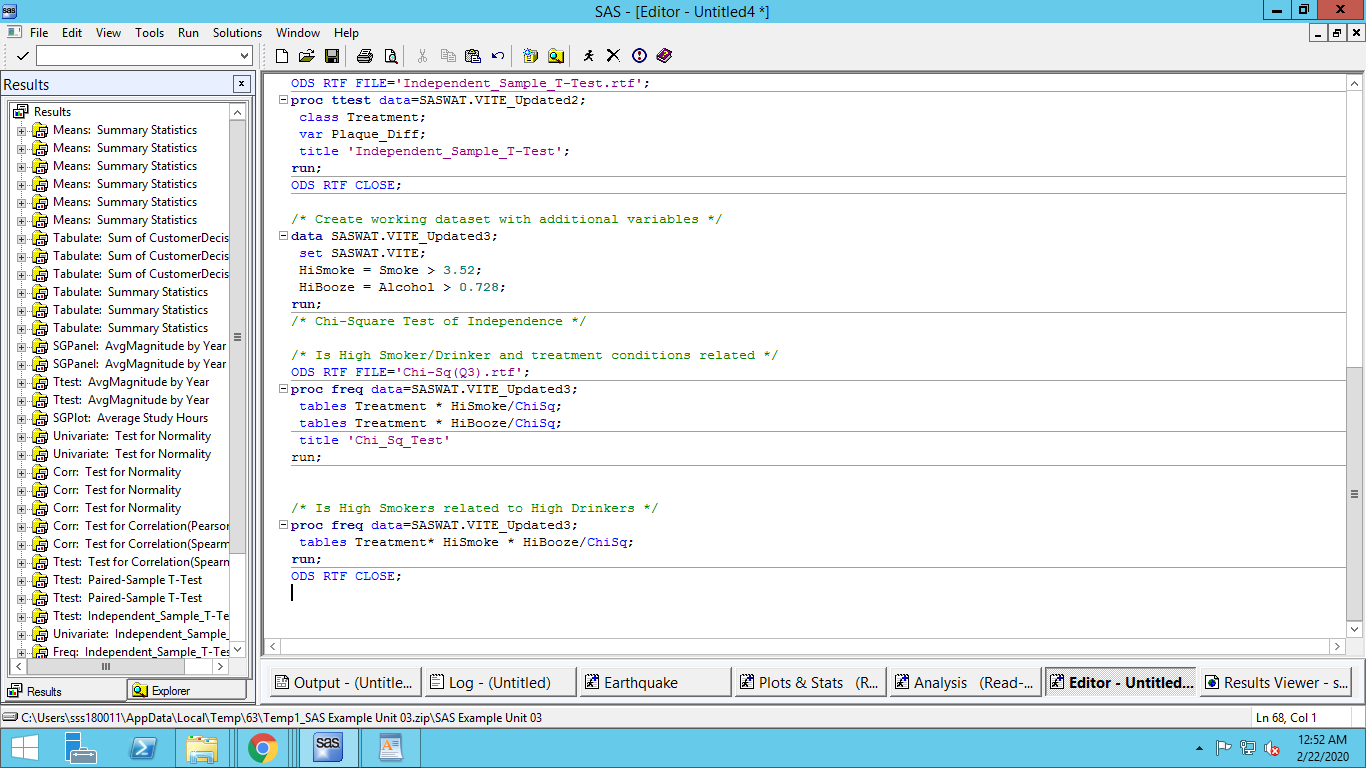
**Chi-square test of independence to check whether High smokers and high drinkers are related for each treatment condition**

**Null Hypothesis(H0): High Smoker and High Drinkers are independent**

**Alternate Hypothesis (Ha): High Smoker and High Drinkers related**



**Since P>0.01, We fail to reject the null hypothesis that High Smoke is independent of High Drinkers**



**So It seems that maybe subjects are not randomized properly.**

------------------------------------------------------------------------------------------------------------------------------------------