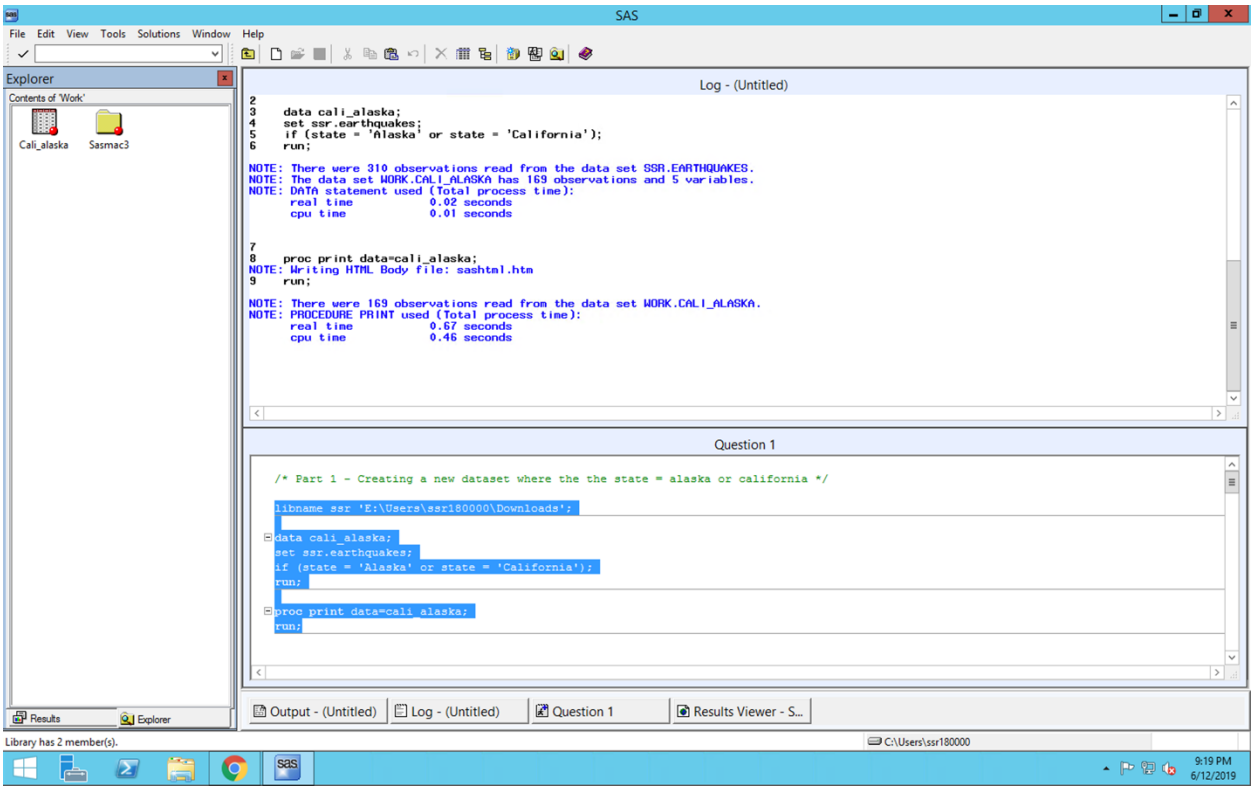
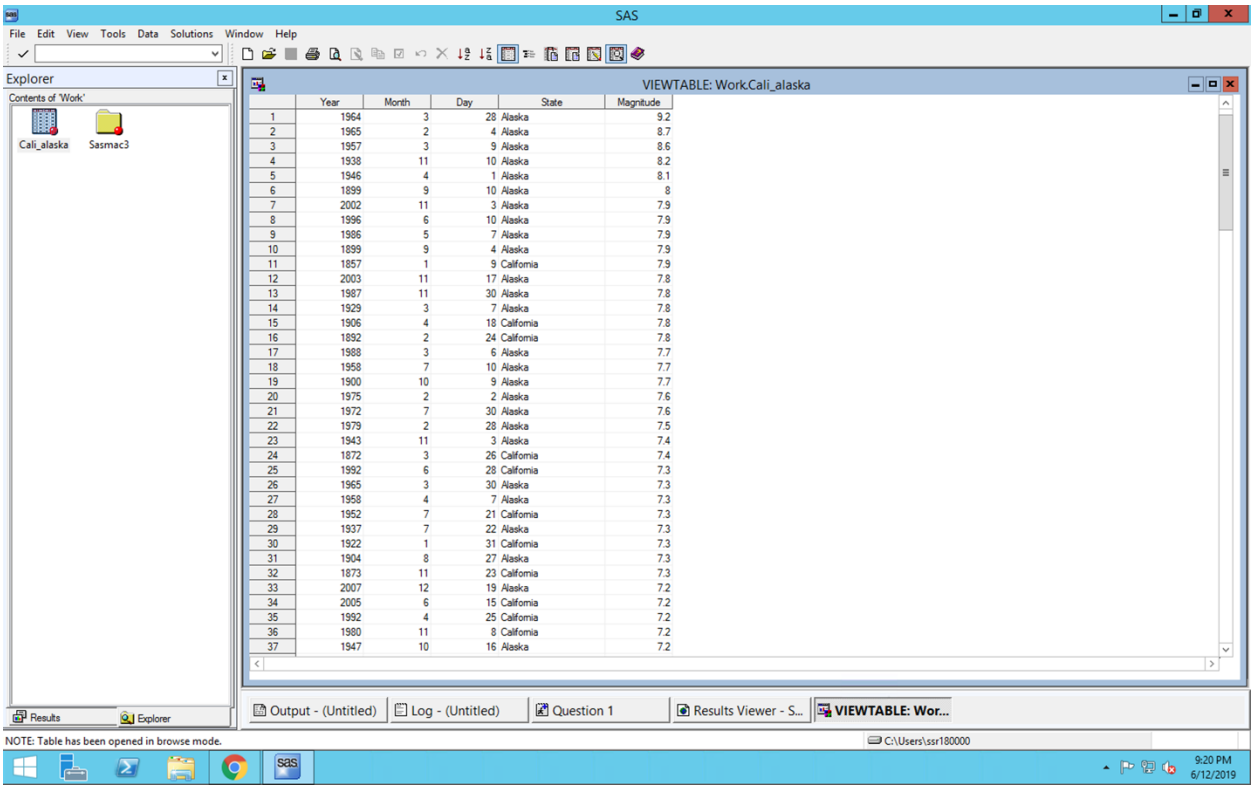
REPORT

TOPIC – HYPOTHESIS TESTING

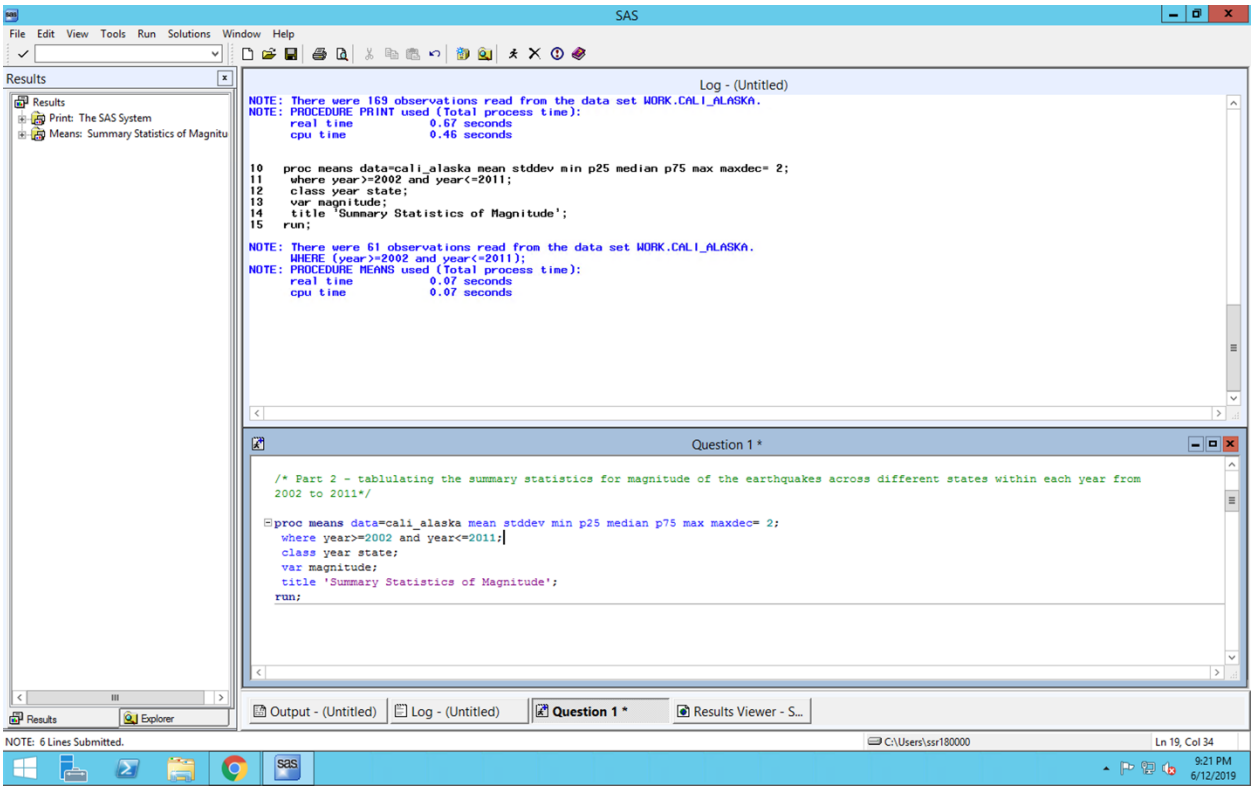
Problem 1:

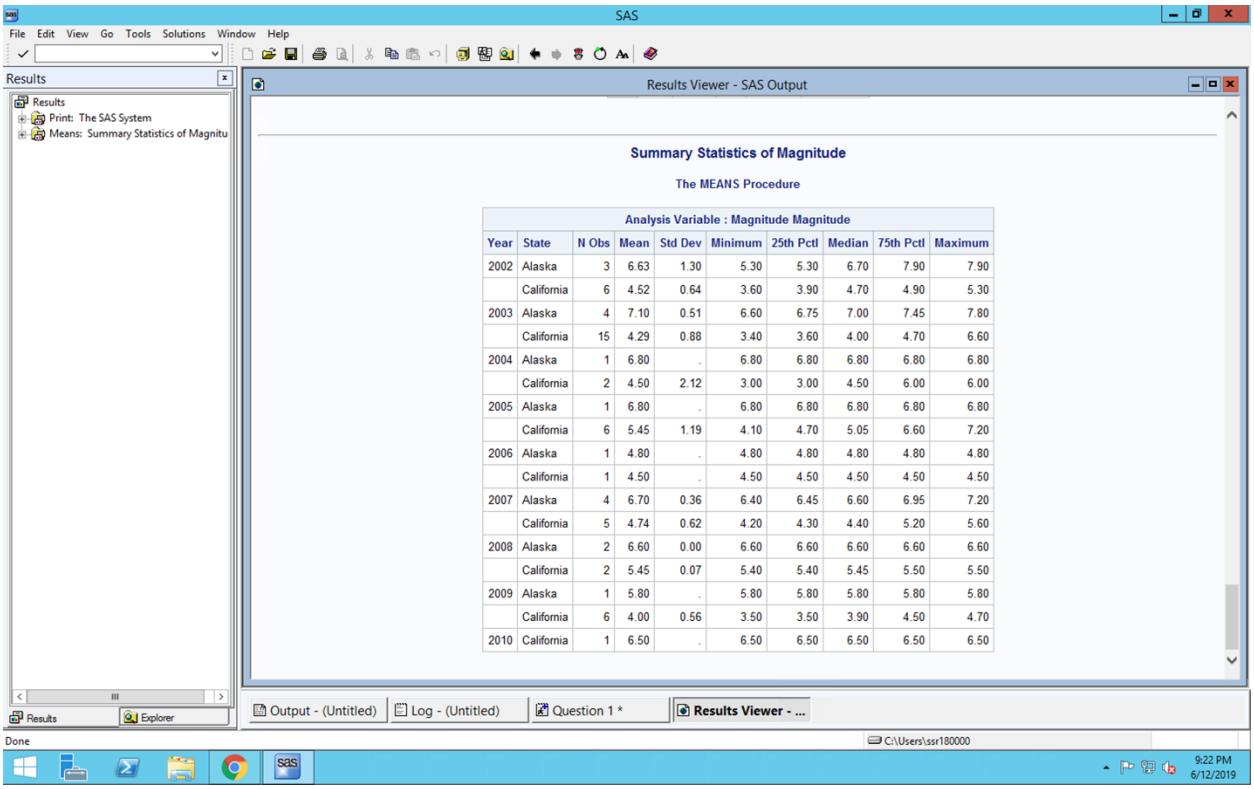
The United States Geological Survey provides data on earthquakes of historical interest. The SAS data set called EARTHQUAKES contains data about earthquakes with a magnitude greater than 2.5 in the United States and its territories. The variables are year, month, day, state, and magnitude.

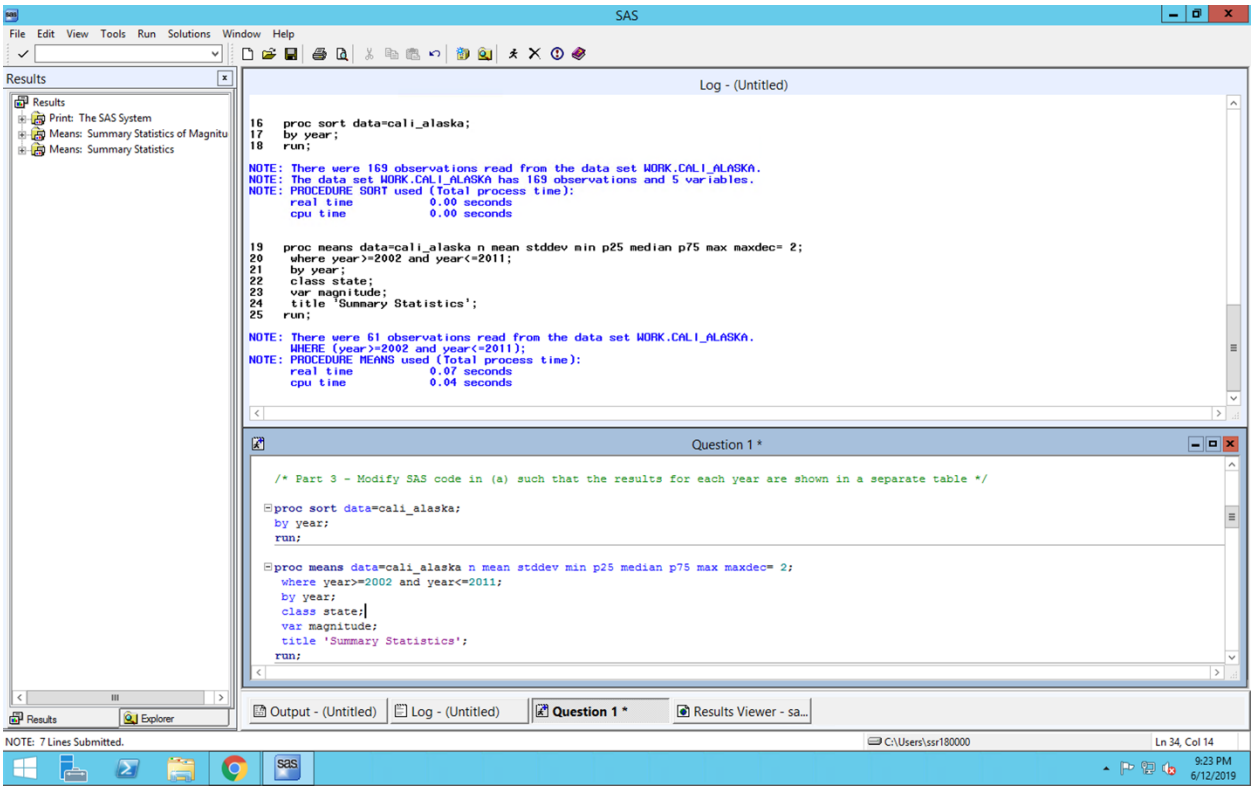


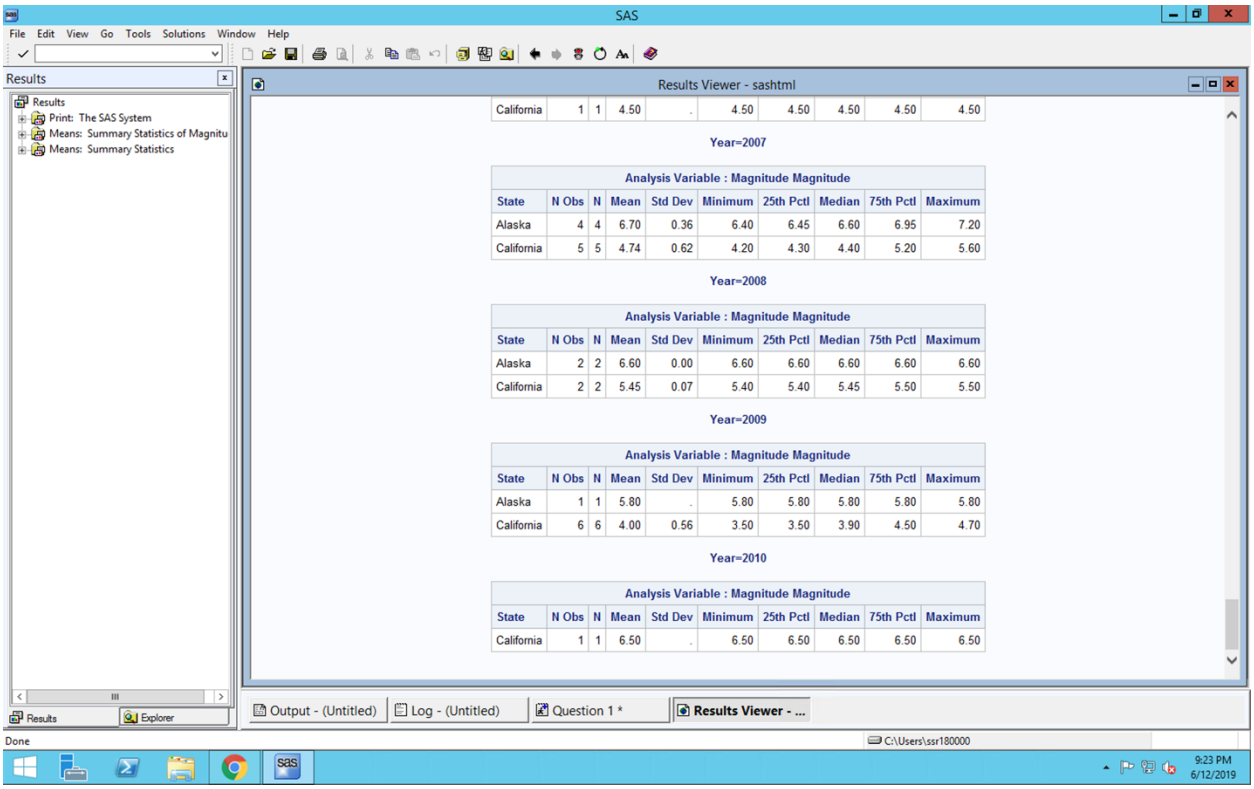


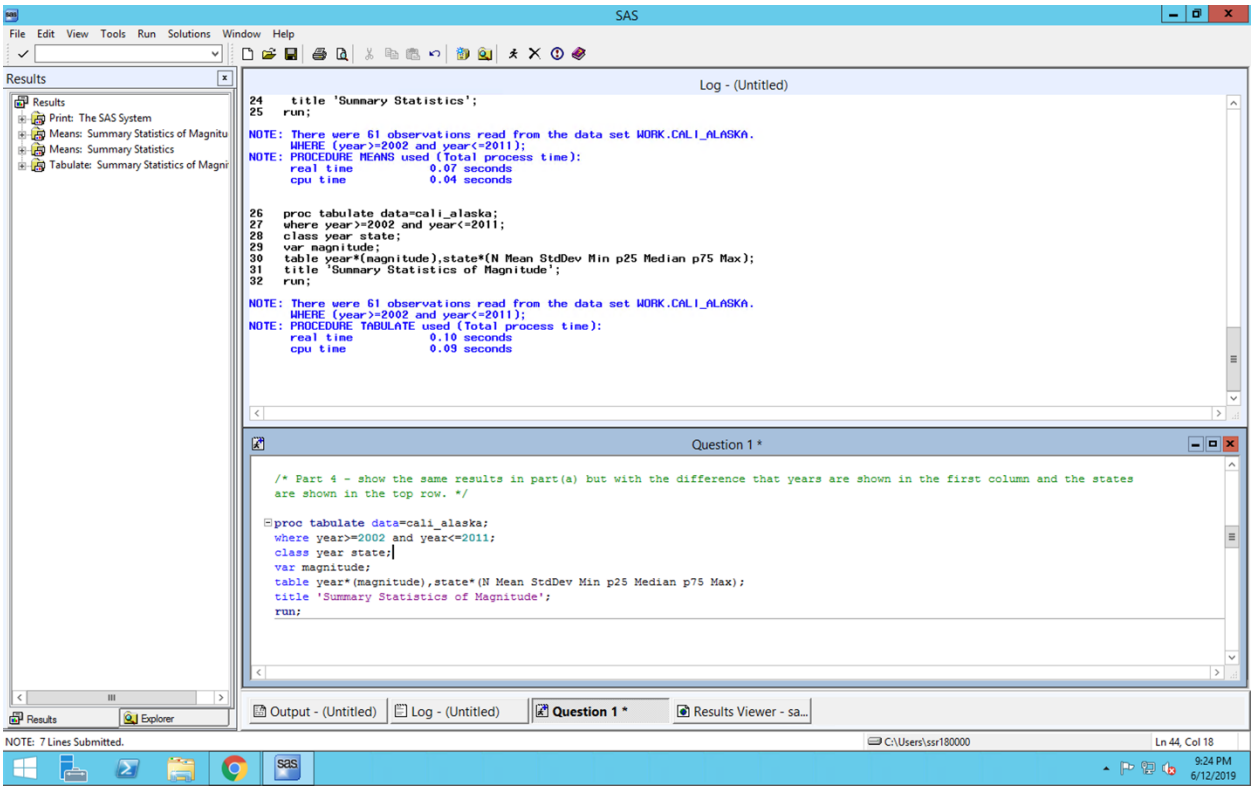
	Year	Month	Day	State	Magnitude
1	1964	3	28	Alaska	9.2
2	1965	2	4	Alaska	8.7
3	1957	3	9	Alaska	8.6
4	1938	11	10	Alaska	8.2
5	1946	4	1	Alaska	8.1
6	1899	9	10	Alaska	8
7	2002	11	3	Alaska	7.9
8	1996	6	10	Alaska	7.9
9	1986	5	7	Alaska	7.9
10	1899	9	4	Alaska	7.9
11	1857	1	9	California	7.9
12	2003	11	17	Alaska	7.8
13	1987	11	30	Alaska	7.8
14	1929	3	7	Alaska	7.8
15	1906	4	18	California	7.8
16	1892	2	24	California	7.8
17	1988	3	6	Alaska	7.7
18	1958	7	10	Alaska	7.7
19	1900	10	9	Alaska	7.7
20	1975	2	2	Alaska	7.6
21	1972	7	30	Alaska	7.6
22	1979	2	28	Alaska	7.5
23	1943	11	3	Alaska	7.4
24	1872	3	26	California	7.4
25	1992	6	28	California	7.3
26	1965	3	30	Alaska	7.3
27	1958	4	7	Alaska	7.3
28	1952	7	21	California	7.3
29	1937	7	22	Alaska	7.3
30	1922	1	31	California	7.3
31	1904	8	27	Alaska	7.3
32	1873	11	23	California	7.3
33	2007	12	19	Alaska	7.2
34	2005	6	15	California	7.2
35	1992	4	25	California	7.2
36	1980	11	8	California	7.2
37	1947	10	16	Alaska	7.2

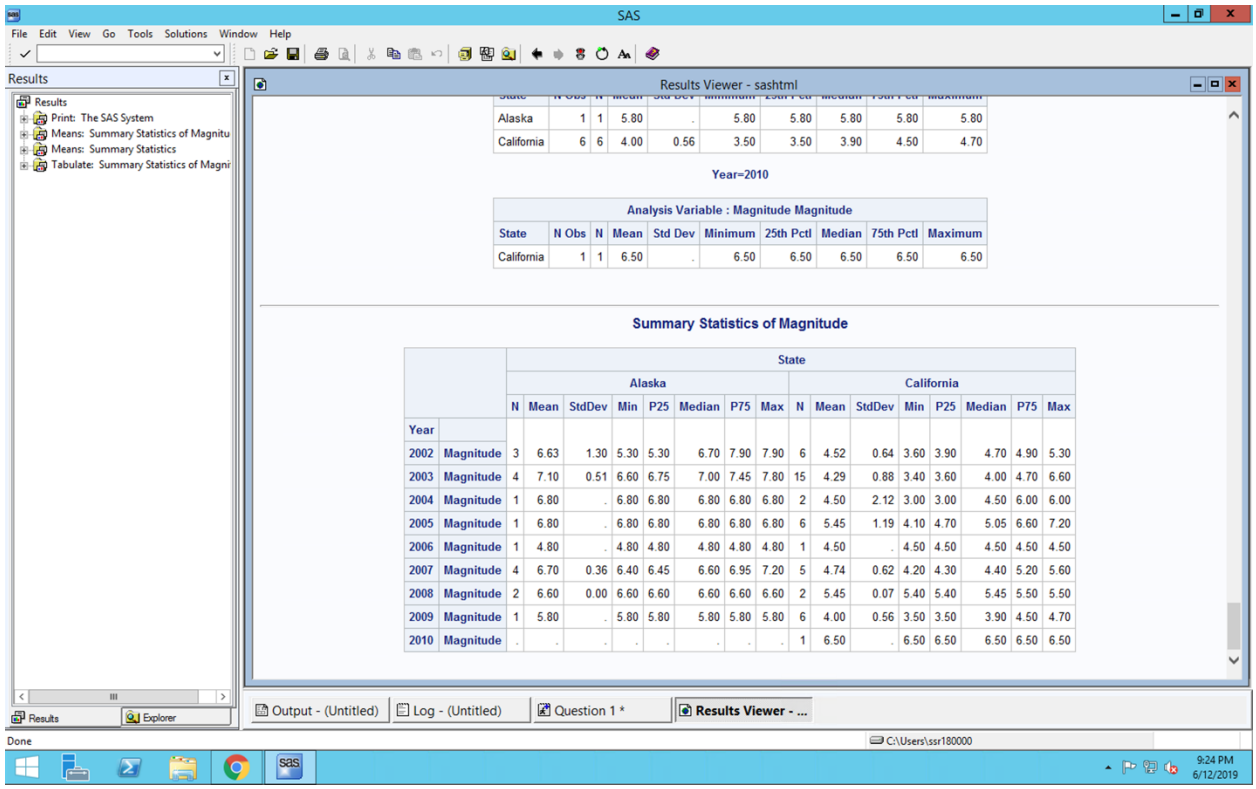


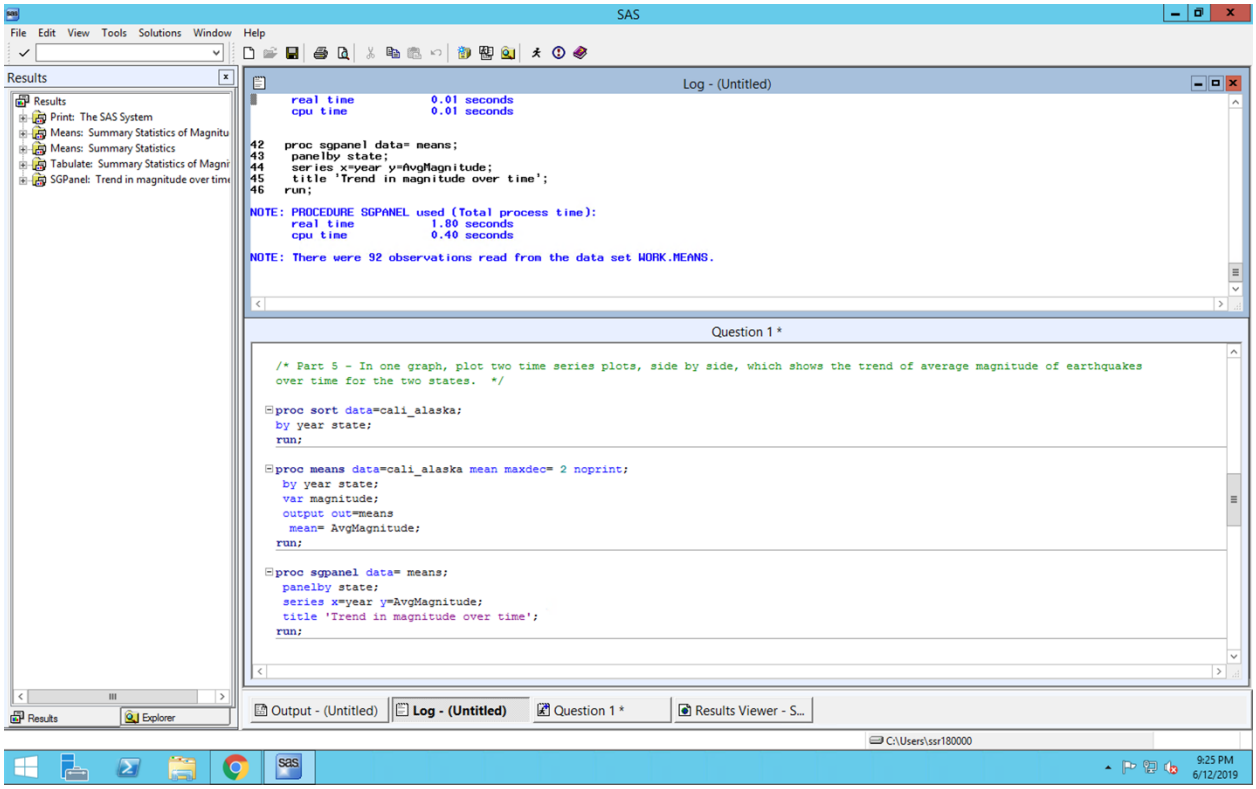

SAS
File Edit View Go Tools Solutions Window Help
Results
Results
Print: The SAS System
Means: Summary Statistics of Magnitu
Results Viewer - SAS Output
Summary Statistics of Magnitude
The MEANS Procedure
Analysis Variable : Magnitude Magnitude
Year State N Obs Mean Std Dev Minimum 25th Pctl Median 75th Pctl Maximum
2002 Alaska 3 6.63 1.30 5.30 5.30 6.70 7.90 7.90
California 6 4.52 0.64 3.60 3.90 4.70 4.90 5.30
2003 Alaska 4 7.10 0.51 6.60 6.75 7.00 7.45 7.80
California 15 4.29 0.88 3.40 3.60 4.00 4.70 6.60
2004 Alaska 1 6.80 . 6.80 6.80 6.80 6.80 6.80
California 2 4.50 2.12 3.00 3.00 4.50 6.00 6.00
2005 Alaska 1 6.80 . 6.80 6.80 6.80 6.80 6.80
California 6 5.45 1.19 4.10 4.70 5.05 6.60 7.20
2006 Alaska 1 4.80 . 4.80 4.80 4.80 4.80 4.80
California 1 4.50 . 4.50 4.50 4.50 4.50 4.50
2007 Alaska 4 6.70 0.36 6.40 6.45 6.60 6.95 7.20
California 5 4.74 0.62 4.20 4.30 4.40 5.20 5.60
2008 Alaska 2 6.60 0.00 6.60 6.60 6.60 6.60 6.60
California 2 5.45 0.07 5.40 5.40 5.45 5.50 5.50
2009 Alaska 1 5.80 . 5.80 5.80 5.80 5.80 5.80
California 6 4.00 0.56 3.50 3.50 3.90 4.50 4.70
2010 California 1 6.50 . 6.50 6.50 6.50 6.50 6.50
Output - (Untitled)
Log - (Untitled)
Question 1 *
Results Viewer - ...
Done
C:\Users\ssr180000
9:22 PM
6/12/2019

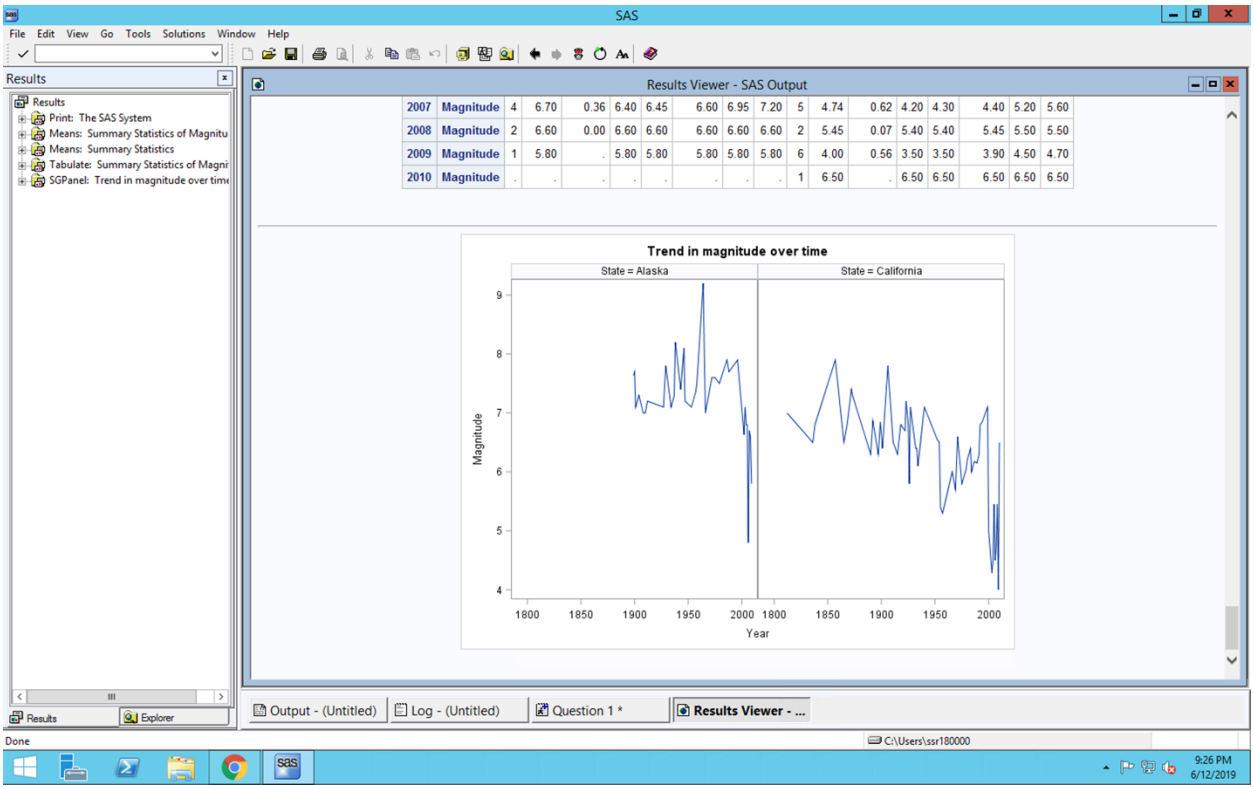

SAS
File Edit View Tools Run Solutions Window Help
Results
Results
Print: The SAS System
Means: Summary Statistics of Magnitu
Means: Summary Statistics
Log - (Untitled)
16 proc sort data=cali_alaska;
17 by year;
18 run;
NOTE: There were 169 observations read from the data set WORK.CALI_ALASKA.
NOTE: The data set WORK.CALI_ALASKA has 169 observations and 5 variables.
NOTE: PROCEDURE SORT used (Total process time):
real time 0.00 seconds
cpu time 0.00 seconds
19 proc means data=cali_alaska n mean stddev min p25 median p75 max maxdec= 2;
20 where year>=2002 and year<=2011;
21 by year;
22 class state;
23 var magnitude;
24 title 'Summary Statistics';
25 run;
NOTE: There were 61 observations read from the data set WORK.CALI_ALASKA.
WHERE (year>=2002 and year<=2011);
NOTE: PROCEDURE MEANS used (Total process time):
real time 0.07 seconds
cpu time 0.04 seconds
Question 1 *
/* Part 3 - Modify SAS code in (a) such that the results for each year are shown in a separate table */
proc sort data=cali_alaska;
by year;
run;
proc means data=cali_alaska n mean stddev min p25 median p75 max maxdec= 2;
where year>=2002 and year<=2011;
by year;
class state;
var magnitude;
title 'Summary Statistics';
run;
Output - (Untitled)
Log - (Untitled)
Question 1 *
Results Viewer - sa...
NOTE: 7 Lines Submitted.
C:\Users\ssr180000
Ln 34, Col 14
9:23 PM
6/12/2019


SAS
File Edit View Go Tools Solutions Window Help
Results
Results
Print: The SAS System
Means: Summary Statistics of Magnitu
Means: Summary Statistics
Results Viewer - sashtml
California 1 1 4.50 . 4.50 4.50 4.50 4.50 4.50
Year=2007
Analysis Variable : Magnitude Magnitude
State N Obs N Mean Std Dev Minimum 25th Pctl Median 75th Pctl Maximum
Alaska 4 4 6.70 0.36 6.40 6.45 6.60 6.95 7.20
California 5 5 4.74 0.62 4.20 4.30 4.40 5.20 5.60
Year=2008
Analysis Variable : Magnitude Magnitude
State N Obs N Mean Std Dev Minimum 25th Pctl Median 75th Pctl Maximum
Alaska 2 2 6.60 0.00 6.60 6.60 6.60 6.60 6.60
California 2 2 5.45 0.07 5.40 5.40 5.45 5.50 5.50
Year=2009
Analysis Variable : Magnitude Magnitude
State N Obs N Mean Std Dev Minimum 25th Pctl Median 75th Pctl Maximum
Alaska 1 1 5.80 . 5.80 5.80 5.80 5.80 5.80
California 6 6 4.00 0.56 3.50 3.50 3.90 4.50 4.70
Year=2010
Analysis Variable : Magnitude Magnitude
State N Obs N Mean Std Dev Minimum 25th Pctl Median 75th Pctl Maximum
California 1 1 6.50 . 6.50 6.50 6.50 6.50 6.50
Output - (Untitled)
Log - (Untitled)
Question 1 *
Results Viewer - ...
Done
C:\Users\ssr180000
9:23 PM
6/12/2019

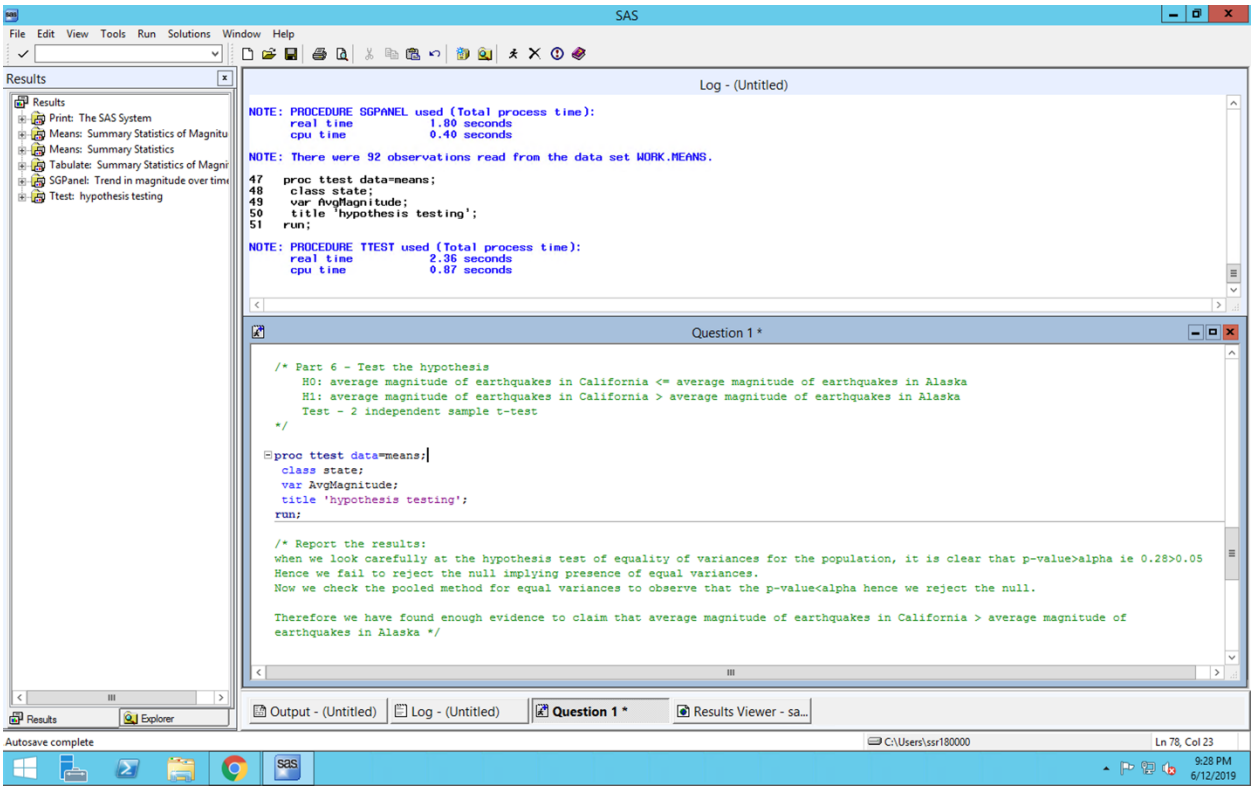

SAS
File Edit View Tools Run Solutions Window Help
Results
Results
Print: The SAS System
Means: Summary Statistics of Magnitu
Means: Summary Statistics
Tabulate: Summary Statistics of Magni
Log - (Untitled)
24 title 'Summary Statistics';
25 run;
NOTE: There were 61 observations read from the data set WORK.CALI_ALASKA.
WHERE (year>=2002 and year<=2011);
NOTE: PROCEDURE MEANS used (Total process time):
real time 0.07 seconds
cpu time 0.04 seconds
26 proc tabulate data=cali_alaska;
27 where year>=2002 and year<=2011;
28 class year state;
29 var magnitude;
30 table year*(magnitude),state*(N Mean StdDev Min p25 Median p75 Max);
31 title 'Summary Statistics of Magnitude';
32 run;
NOTE: There were 61 observations read from the data set WORK.CALI_ALASKA.
WHERE (year>=2002 and year<=2011);
NOTE: PROCEDURE TABULATE used (Total process time):
real time 0.10 seconds
cpu time 0.09 seconds
Question 1 *
/* Part 4 - show the same results in part(a) but with the difference that years are shown in the first column and the states
are shown in the top row. */
proc tabulate data=cali_alaska;
where year>=2002 and year<=2011;
class year state;
var magnitude;
table year*(magnitude),state*(N Mean StdDev Min p25 Median p75 Max);
title 'Summary Statistics of Magnitude';
run;
Output - (Untitled)
Log - (Untitled)
Question 1 *
Results Viewer - sa...
NOTE: 7 Lines Submitted.
C:\Users\ssr180000
Ln 44, Col 18
9:24 PM
6/12/2019


SAS
File Edit View Go Tools Solutions Window Help
Results
Results
Print: The SAS System
Means: Summary Statistics of Magnitu
Means: Summary Statistics
Tabulate: Summary Statistics of Magni
Results Viewer - sashtml
Alaska 1 1 5.80 . 5.80 5.80 5.80 5.80 5.80
California 6 6 4.00 0.56 3.50 3.50 3.90 4.50 4.70
Year=2010
Analysis Variable : Magnitude Magnitude
State N Obs N Mean Std Dev Minimum 25th Pctl Median 75th Pctl Maximum
California 1 1 6.50 . 6.50 6.50 6.50 6.50 6.50
Summary Statistics of Magnitude
State
Alaska
California
N Mean StdDev Min P25 Median P75 Max N Mean StdDev Min P25 Median P75 Max
Year
2002 Magnitude 3 6.63 1.30 5.30 5.30 6.70 7.90 7.90 6 4.52 0.64 3.60 3.90 4.70 4.90 5.30
2003 Magnitude 4 7.10 0.51 6.60 6.75 7.00 7.45 7.80 15 4.29 0.88 3.40 3.60 4.00 4.70 6.60
2004 Magnitude 1 6.80 . 6.80 6.80 6.80 6.80 6.80 2 4.50 2.12 3.00 3.00 4.50 6.00 6.00
2005 Magnitude 1 6.80 . 6.80 6.80 6.80 6.80 6.80 6 5.45 1.19 4.10 4.70 5.05 6.60 7.20
2006 Magnitude 1 4.80 . 4.80 4.80 4.80 4.80 4.80 1 4.50 . 4.50 4.50 4.50 4.50 4.50
2007 Magnitude 4 6.70 0.36 6.40 6.45 6.60 6.95 7.20 5 4.74 0.62 4.20 4.30 4.40 5.20 5.60
2008 Magnitude 2 6.60 0.00 6.60 6.60 6.60 6.60 6.60 2 5.45 0.07 5.40 5.40 5.45 5.50 5.50
2009 Magnitude 1 5.80 . 5.80 5.80 5.80 5.80 5.80 6 4.00 0.56 3.50 3.50 3.90 4.50 4.70
2010 Magnitude . . . . 1 6.50 . 6.50 6.50 6.50 6.50 6.50
Output - (Untitled)
Log - (Untitled)
Question 1 *
Results Viewer - ...
Done
C:\Users\ssr180000
9:24 PM
6/12/2019


SAS
File Edit View Tools Solutions Window Help
Results
Results
Print: The SAS System
Means: Summary Statistics of Magnitu
Means: Summary Statistics
Tabulate: Summary Statistics of Magni
SGPanel: Trend in magnitude over time
Log - (Untitled)
real time 0.01 seconds
cpu time 0.01 seconds
42 proc sgpanel data= means;
43 panelby state;
44 series x=year y=AvgMagnitude;
45 title 'Trend in magnitude over time';
46 run;
NOTE: PROCEDURE SGPANEL used (Total process time):
real time 1.80 seconds
cpu time 0.40 seconds
NOTE: There were 92 observations read from the data set WORK.MEANS.
Question 1 *
/* Part 5 - 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. */
proc sort data=cali_alaska;
by year state;
run;
proc means data=cali_alaska mean maxdec= 2 noprint;
by year state;
var magnitude;
output out=means
mean= AvgMagnitude;
run;
proc sgpanel data= means;
panelby state;
series x=year y=AvgMagnitude;
title 'Trend in magnitude over time';
run;
Output - (Untitled)
Log - (Untitled)
Question 1 *
Results Viewer - S...
C:\Users\ssr180000
9:25 PM
6/12/2019



Is the average magnitude of earthquakes in California is significantly higher than that in Alaska?

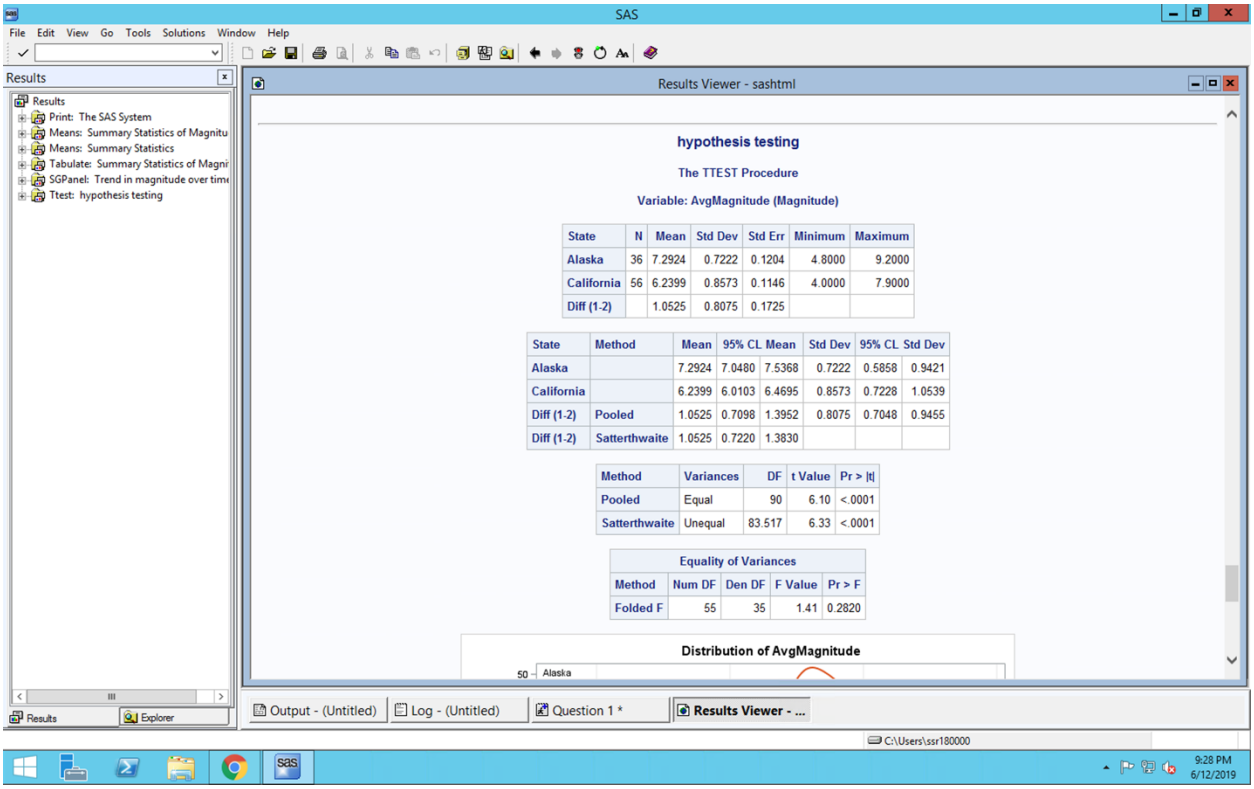


The null hypothesis is:

$$H_0: \mu_{California} \leq \mu_{Alaska}$$

Or equivalently

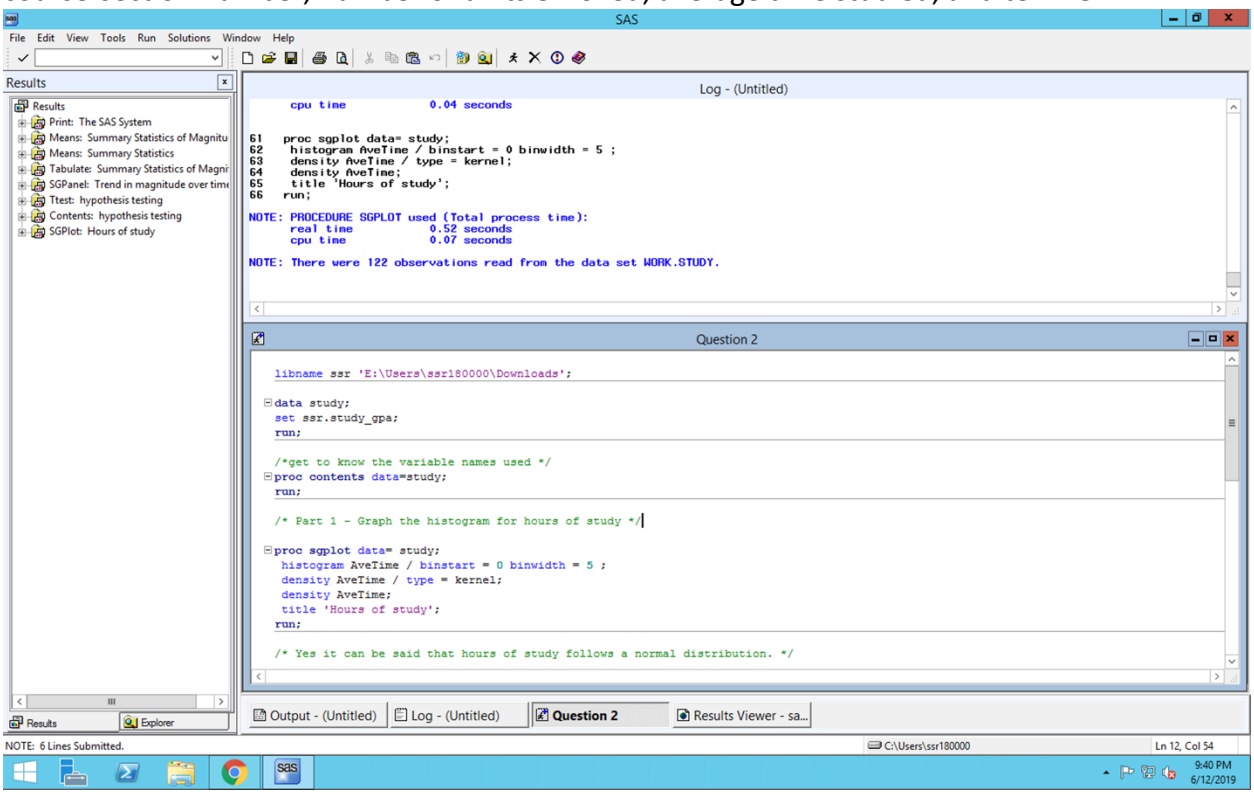
$$H_0: \mu_{Alaska} - \mu_{California} \geq 0$$



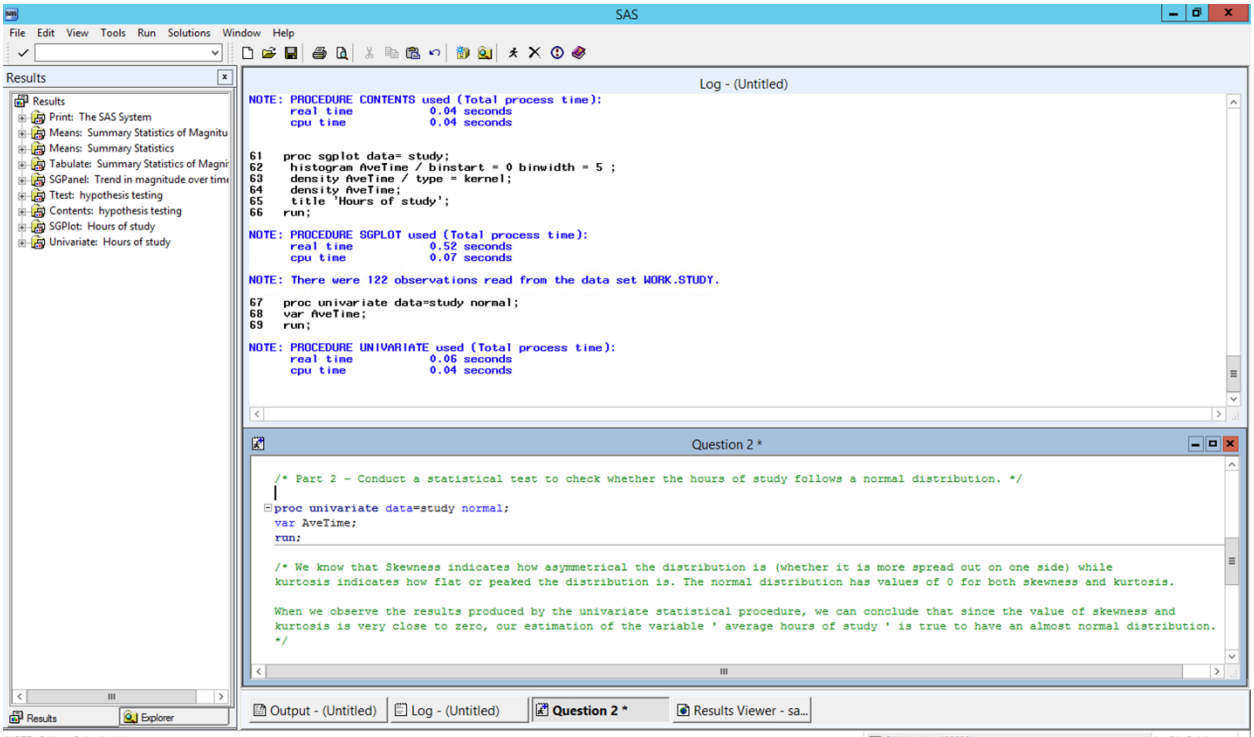
The p-value is very large, implying that we cannot reject the null. Hence, statistically, there is no significance evidence to believe that the average magnitude or earthquakes in California is larger.

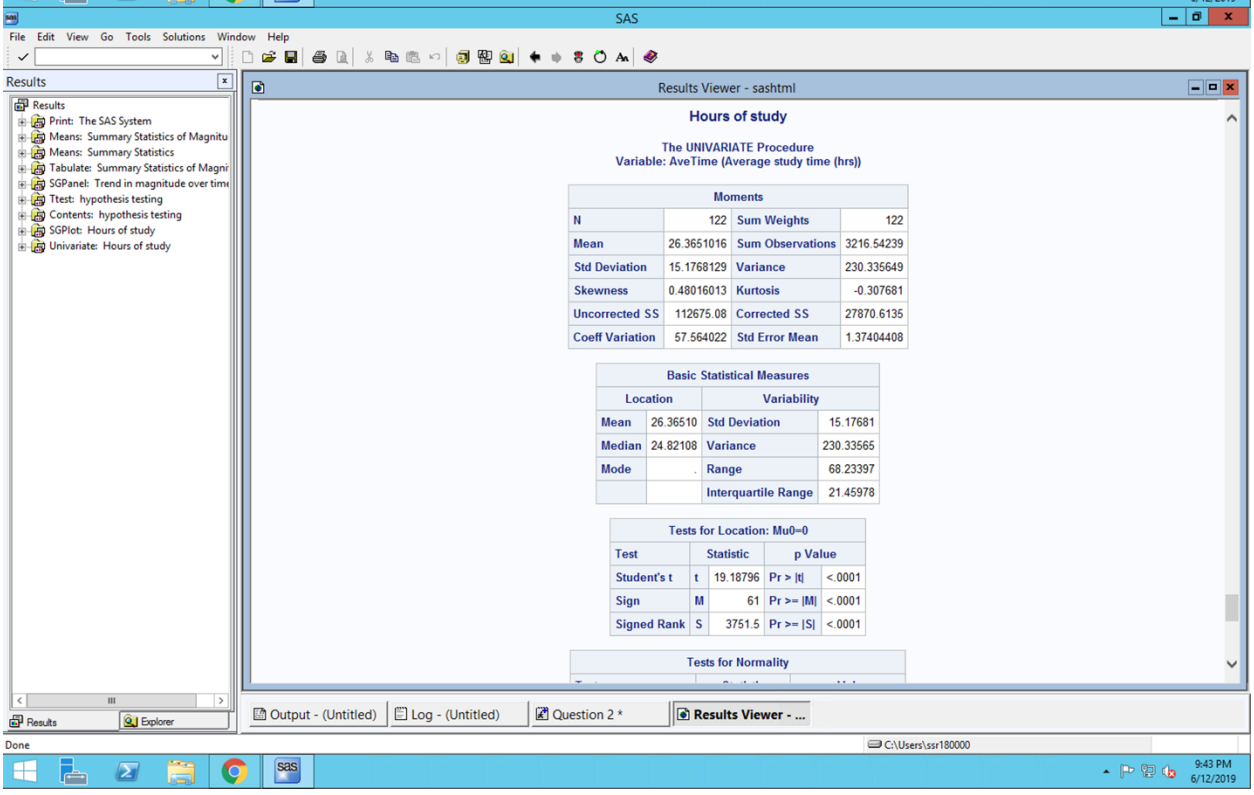
Problem 2:

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



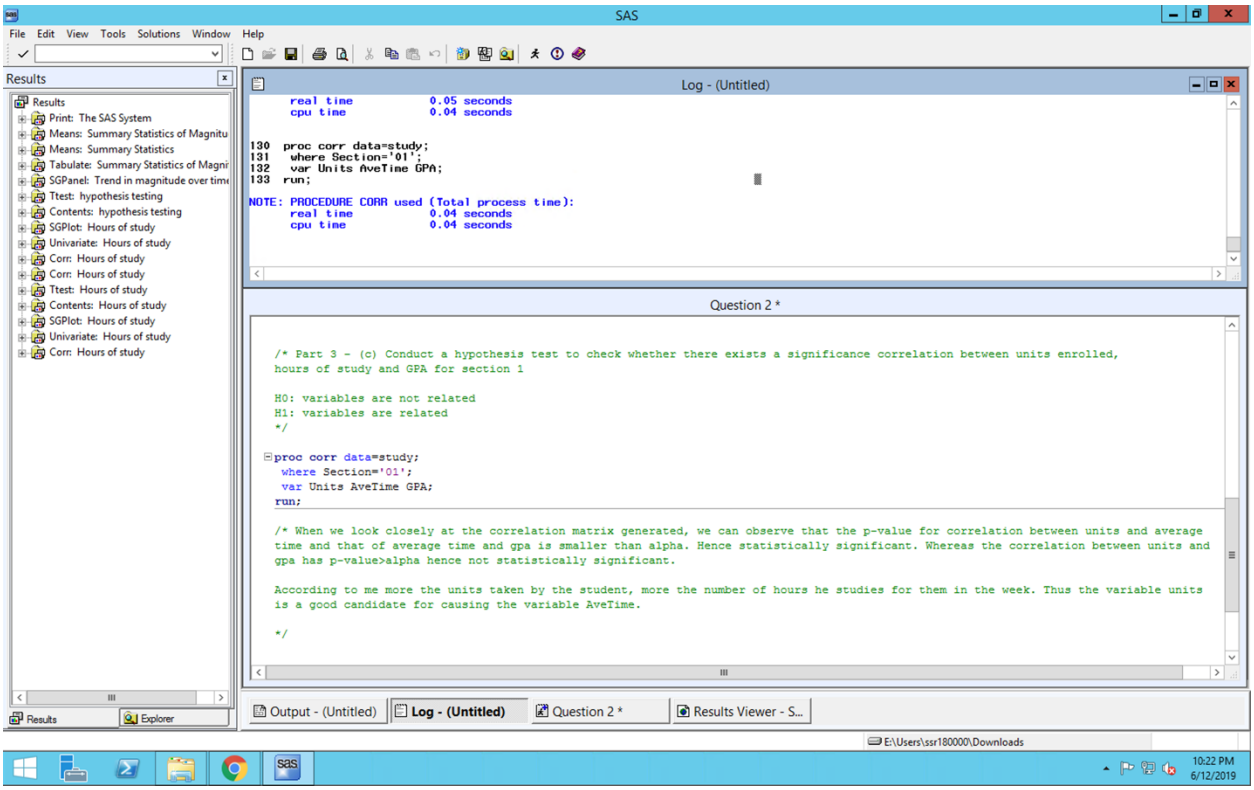
Eyeballing: The kernel density is a bit skewed to the left, but not too much. We can't strongly argue that it looks like a normal distribution, however, normality assumption may not be a big assumption since the skewedness is not significant.

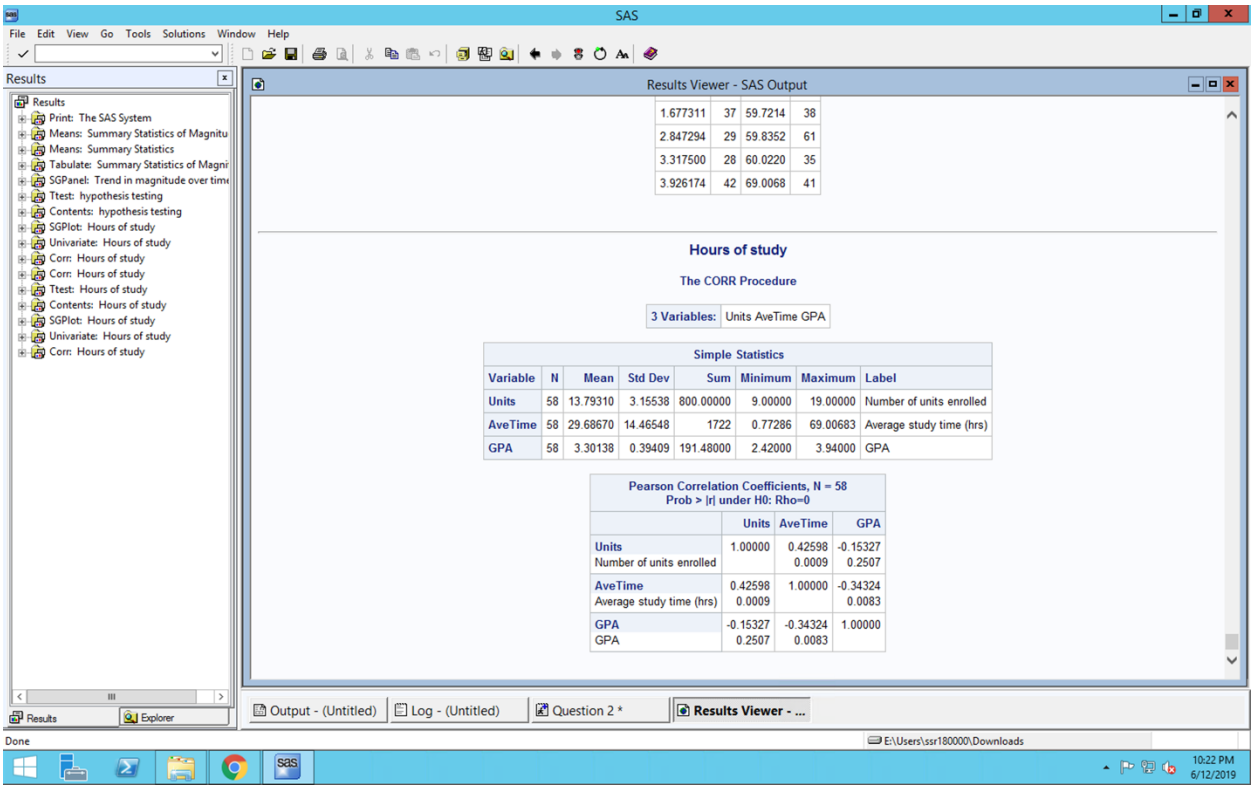




Causality and Correlation.

Check whether there exists a significance correlation between units enrolled, hours of study and GPA for section 1.





From this result, it can be seen that average study time and number of units registered are positively correlated. There is no other significant correlation. This correlation is very sensible: the more units you register, the more time you spend on studying.

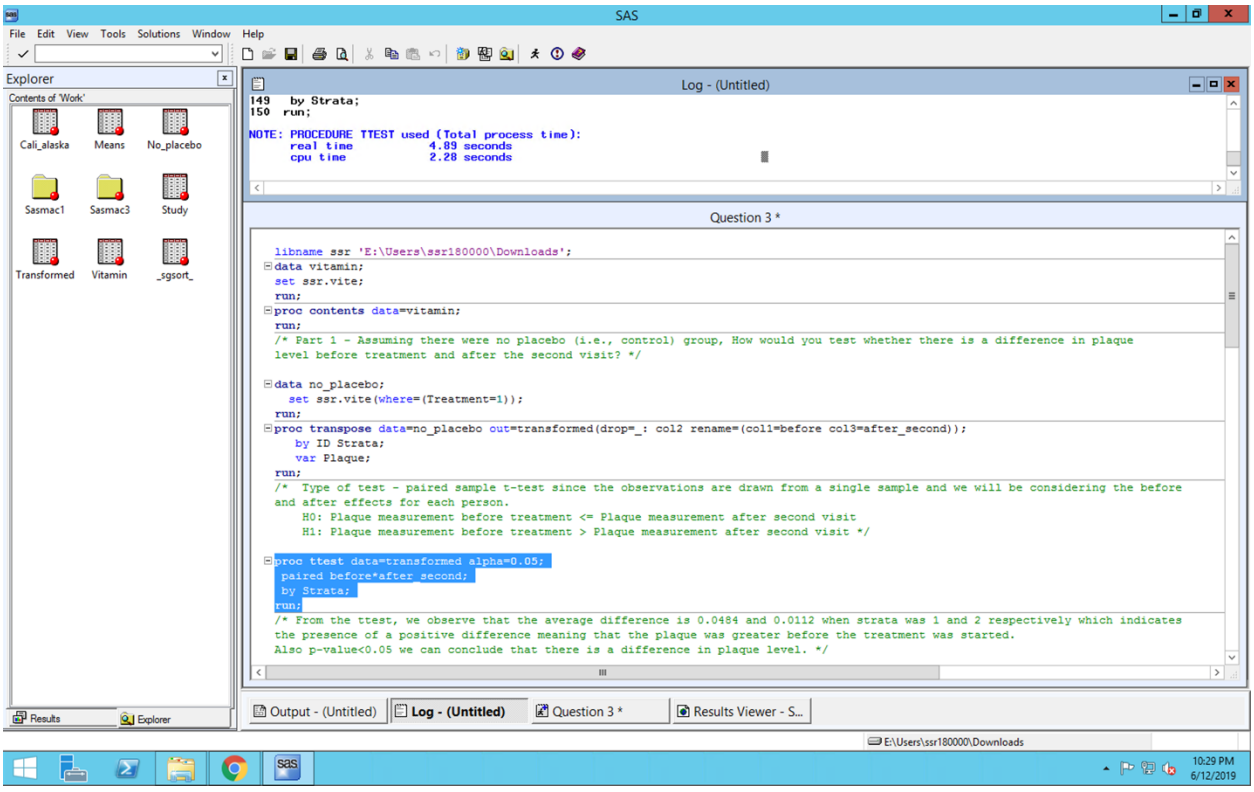
Problem 3:

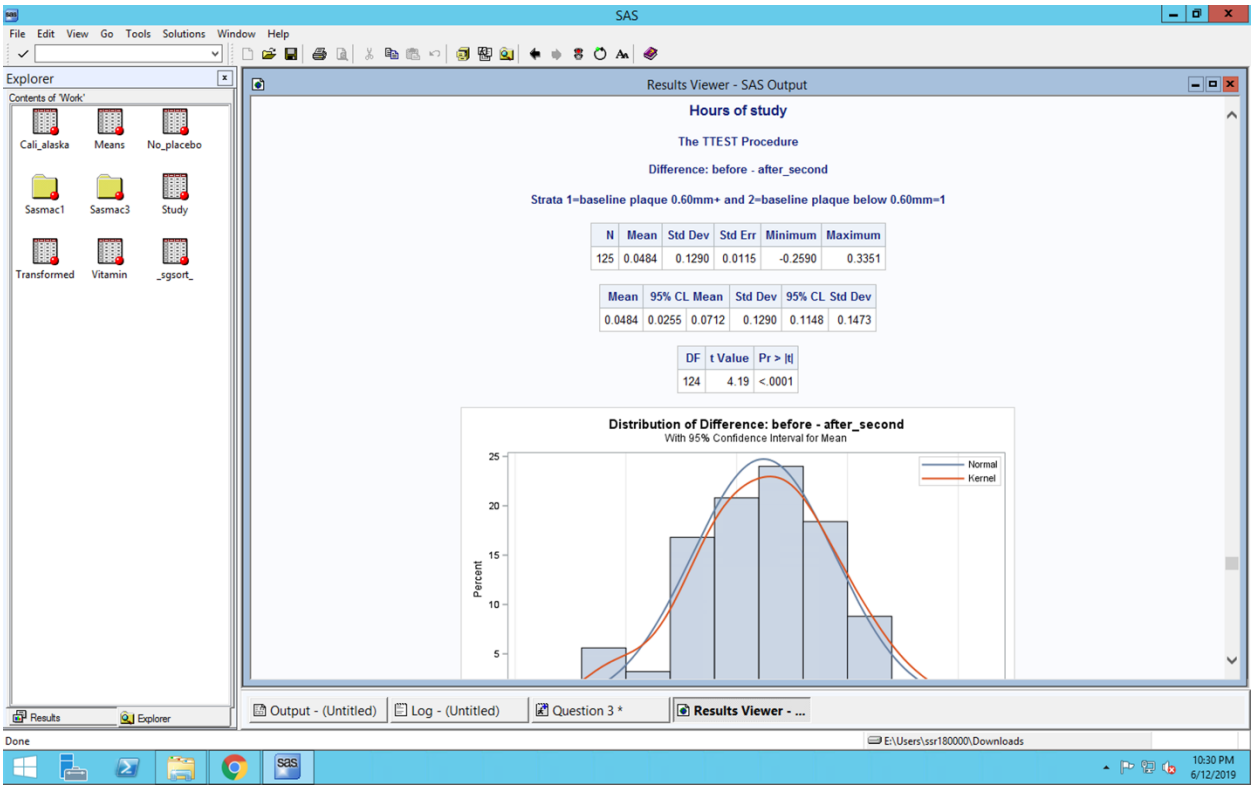
A study was conducted to see whether taking vitamin E daily would reduce the levels of atherosclerotic disease in a random sample of 500 individuals. Clinical measurements, including thickness of plaque of the carotid artery (taken via ultrasound), were recorded at baseline and at two subsequent visits in a SAS data set called VITE. Patients were divided into two strata according to their baseline plaque measurement.

Difference in plaque level before treatment and after the second visit?

$$H_0 : \textit{For treatment group, plaque level after the second visis is greater than plaque level at the beginning}$$

For this part, we need to run a paired t test for treatment group to check whether taking vitamin E changes the plaque level. If we reject the null, we can conclude that taking vitamin E is effective in reducing the plaque level for treatment group.
Note that, in general, rejecting a one sided test is easier. However, when using a one sided test we should have a reasonable prior assumption. In this case, for example, it is believed that taking vitamin E should reduce the plaque level, hence, a one sided test is reasonable. If there is no such reasoning, we have to use a two sided test.





The null is rejected. So, taking vitamin E has a significant effect on reducing the plaque level for treatment group after the second visit.

Now, considering the fact that there is indeed a control group in your dataset, change in the hypothesis test will be:

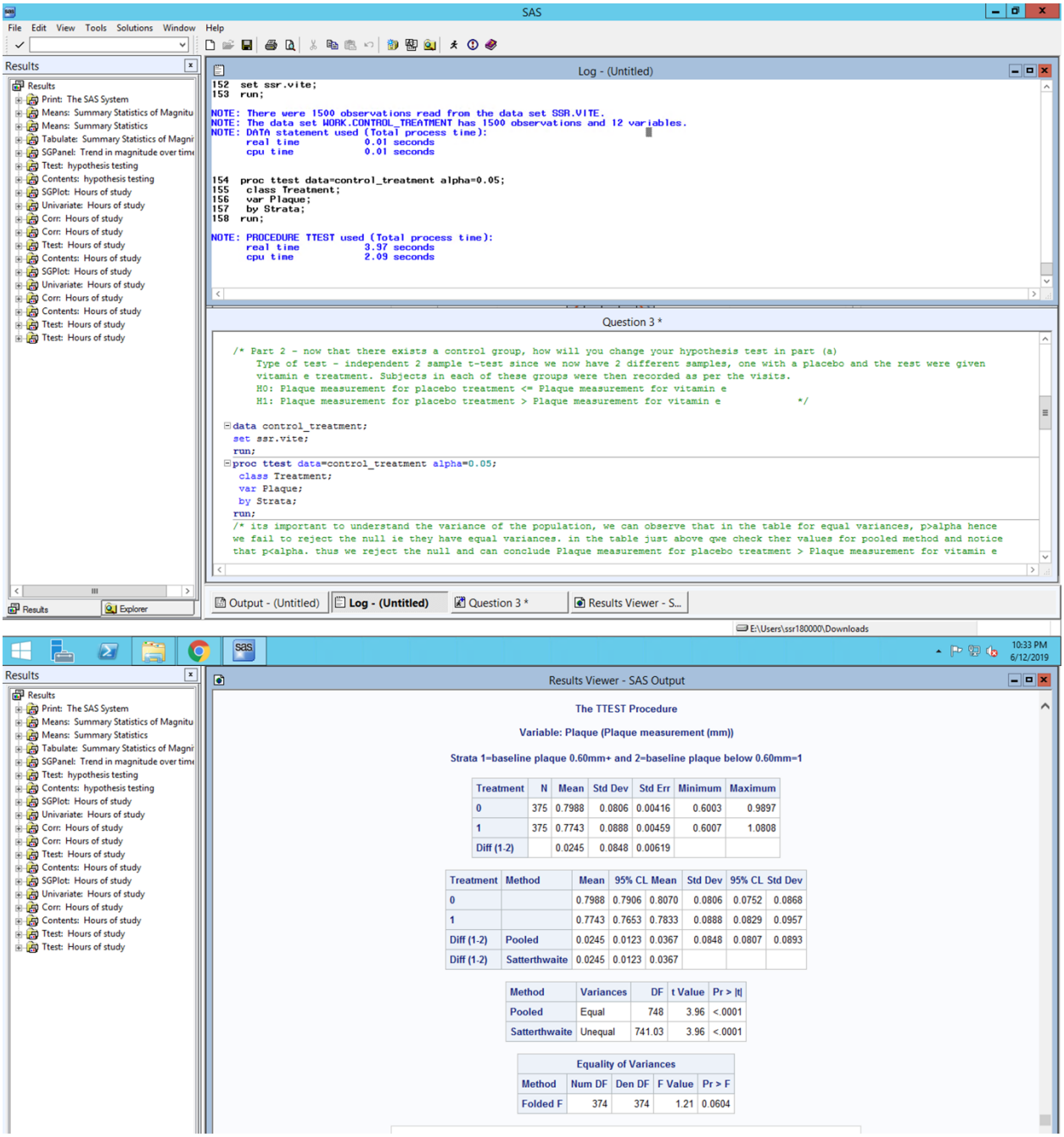
In many application, researchers divide individuals in control and treatment groups. Having the control group will insure that the effect is due to the factors we are interested in. From the results in part (a), one may conclude that reduction in plaque level is due to taking vitamin E. However, the reduction in plaque level may be due to some other factors which are not observed by researcher. Here is how the control group will help. We conduct a hypothesis test for control group as well, then two things might happen:

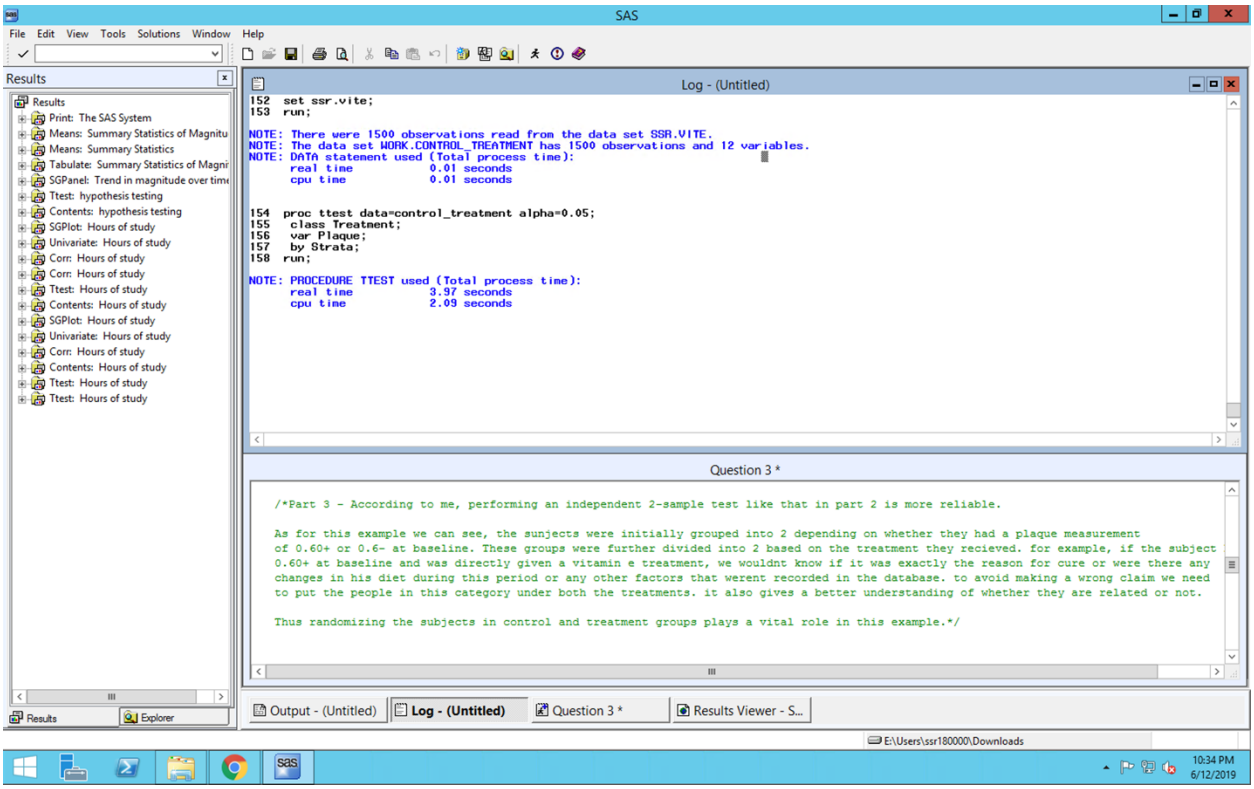
1- The plaque level for control group is the same before and after treatment. This implies that the reduction in plaque level for treatment group is indeed due to taking vitamin E. The reason is that if this reduction was due to some factors other than vitamin E, the plaque level for control group must have changed to.
2- The plaque level for control group is also changing. Here we cannot make an immediate conclusion. We need to know the amount of change. If the reduction in plaque level for treatment group is much more than that for control group, we may still conclude that taking vitamin E is effective. This method is known as Difference in Difference approach (Diff in Diff).

In summary, the null hypothesis that we need to conduct is:

H_0 : *The reduction in plaque level for treatment group is less than that for control group*

If we reject the hypothesis above, we can conclude that the reduction in plaque level in significantly greater for treatment group compared with control group, implying that taking vitamin E is significantly effective in reducing the plaque level.



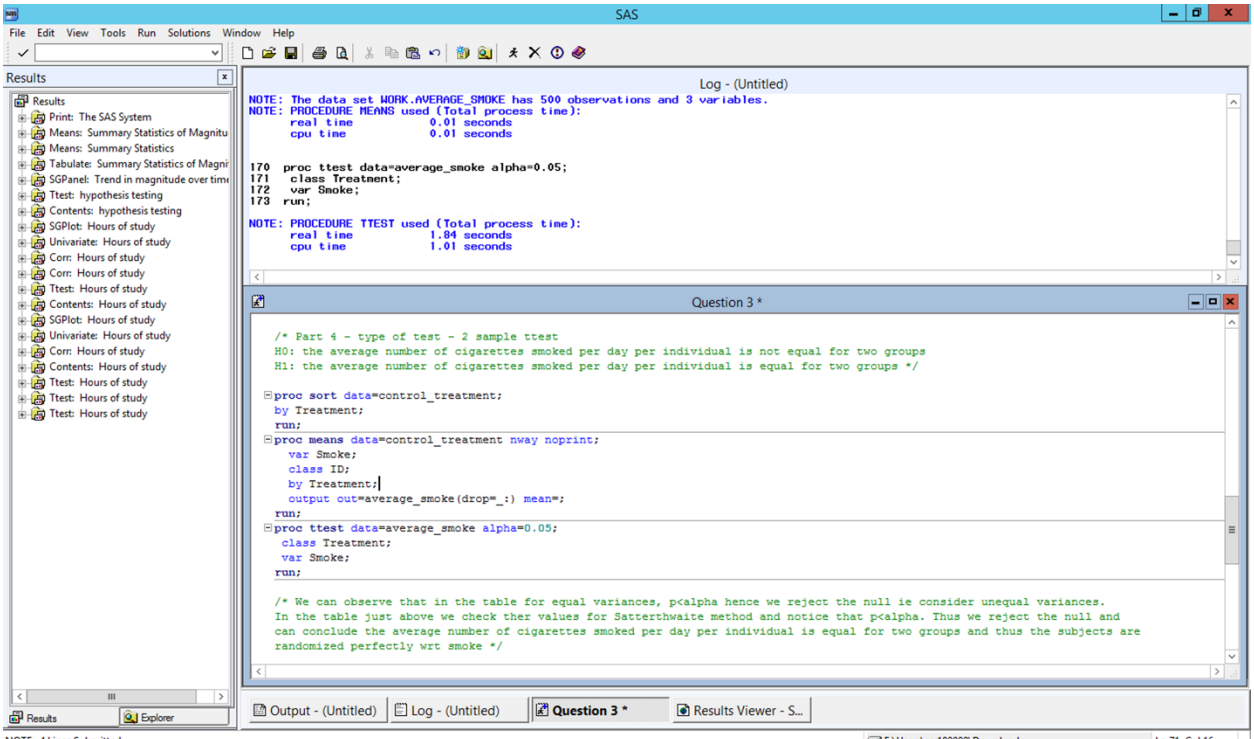


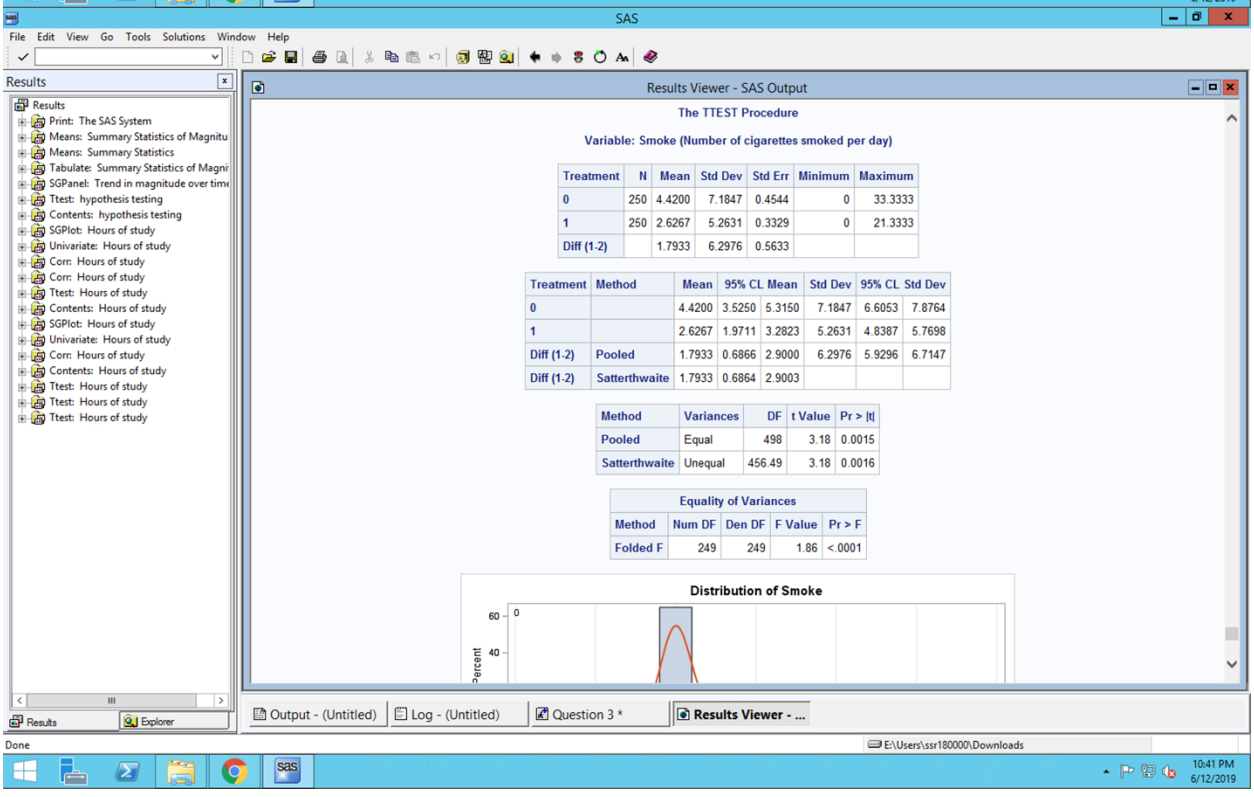
One of the critical factors in randomizing the subjects in control and treatment groups is to make sure that the subject are perfectly randomized in all aspects. Using the last two columns (i.e., alcohol and cigarette usage) conduct two hypothesis tests to check whether subjects are randomized perfectly.

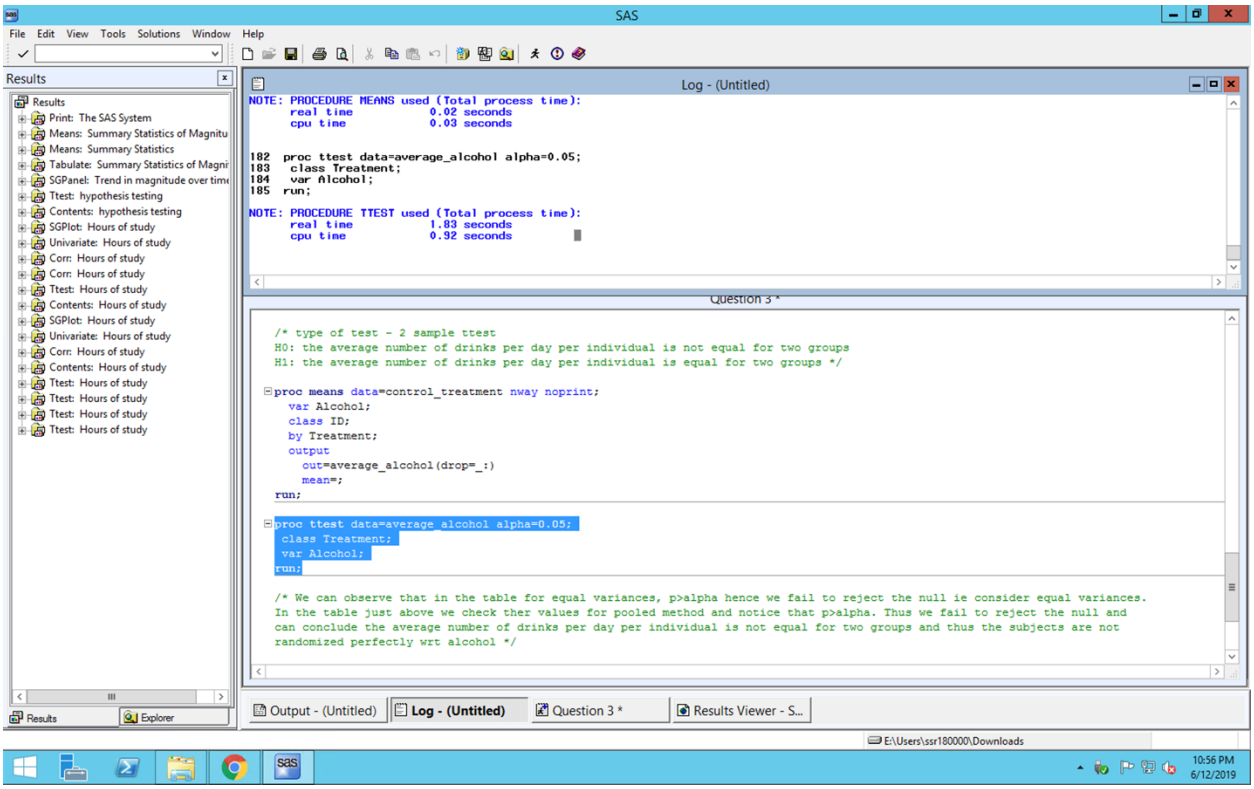
Following the arguments above, we need to make sure that the only difference between treatment and control group is taking/not taking vitamin E and they are similar in every other aspect. This implies that randomizing people in two groups has to be done very carefully. A bad randomizing, for example, is a case where a high portion of individuals in treatment group drink alcohol regularly, whereas, no one in control group drinks alcohol. This will ruin the results. In our example, we need to run the following two tests.

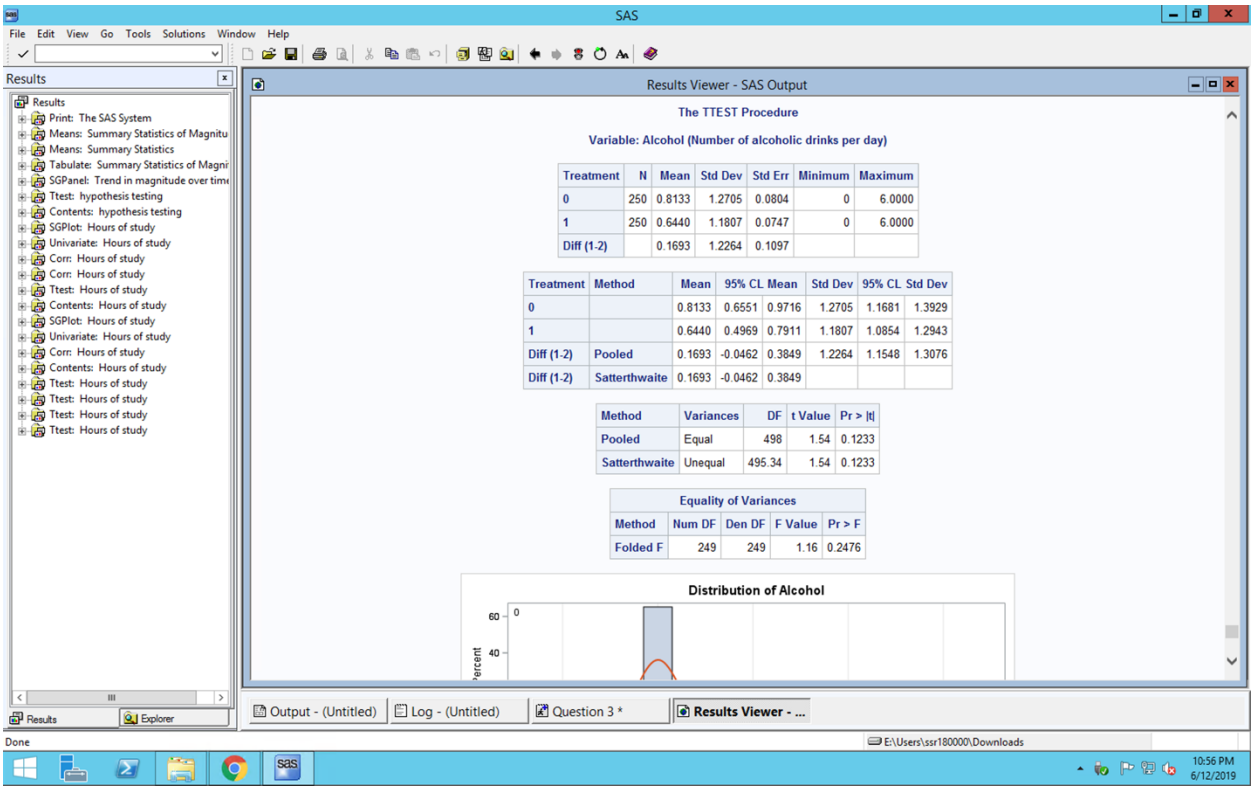
H_0: The average number of cigarettes smoked per day per individual is the same for both groups.
H_0: The average amount of alcohol drunk per day per individual is the same for both groups.

If we reject any of these tests, it means the randomization is not perfect. I don't put the results here. These two tests should be very straightforward to do. You can find the SAS codes posted. So, I'll leave it as your own exercise.
Remark: Ideally we want also the portions to be the same. In other words, we need to also make sure that: **The portion of people who smoke cigarettes/drink alcohol is the same for both groups.**
For the purpose of this assignment, doing only the first test is accepted.


SAS
File Edit View Tools Run Solutions Window Help
Results
Log - (Untitled)
NOTE: The data set WORK.AVERAGE_SMOKE has 500 observations and 3 variables.
NOTE: PROCEDURE MEANS used (Total process time):
real time 0.01 seconds
cpu time 0.01 seconds
170 proc ttest data=average_smoke alpha=0.05;
171 class Treatment;
172 var Smoke;
173 run;
NOTE: PROCEDURE TTEST used (Total process time):
real time 1.84 seconds
cpu time 1.01 seconds
Question 3 *
/* Part 4 - type of test - 2 sample ttest
H0: the average number of cigarettes smoked per day per individual is not equal for two groups
H1: the average number of cigarettes smoked per day per individual is equal for two groups */
proc sort data=control_treatment;
by Treatment;
run;
proc means data=control_treatment nway noprint;
var Smoke;
class ID;
by Treatment;
output out=average_smoke(drop=_:) mean=;
run;
proc ttest data=average_smoke alpha=0.05;
class Treatment;
var Smoke;
run;
/* We can observe that in the table for equal variances, p<alpha hence we reject the null ie consider unequal variances.
In the table just above we check ther values for Satterthwaite method and notice that p<alpha. Thus we reject the null and
can conclude the average number of cigarettes smoked per day per individual is equal for two groups and thus the subjects are
randomized perfectly wrt smoke */
Output - (Untitled)
Log - (Untitled)
Question 3 *
Results Viewer - S...
NOTE: 4 Lines Submitted.
E:\Users\ssr180000\Downloads
Ln 71, Col 16
10:41 PM
6/12/2019


SAS
File Edit View Go Tools Solutions Window Help
Results
Results Viewer - SAS Output
The TTEST Procedure
Variable: Smoke (Number of cigarettes smoked per day)
Treatment N Mean Std Dev Std Err Minimum Maximum
0 250 4.4200 7.1847 0.4544 0 33.3333
1 250 2.6267 5.2631 0.3329 0 21.3333
Diff (1-2) 1.7933 6.2976 0.5633
Treatment Method Mean 95% CL Mean Std Dev 95% CL Std Dev
0 4.4200 3.5250 5.3150 7.1847 6.6053 7.8764
1 2.6267 1.9711 3.2823 5.2631 4.8387 5.7698
Diff (1-2) Pooled 1.7933 0.6866 2.9000 6.2976 5.9296 6.7147
Diff (1-2) Satterthwaite 1.7933 0.6864 2.9003
Method Variances DF t Value Pr > |t|
Pooled Equal 498 3.18 0.0015
Satterthwaite Unequal 456.49 3.18 0.0016
Equality of Variances
Method Num DF Den DF F Value Pr > F
Folded F 249 249 1.86 <.0001
Distribution of Smoke
Percent
60
40
0
Output - (Untitled)
Log - (Untitled)
Question 3 *
Results Viewer - ...
Done
E:\Users\ssr180000\Downloads
10:41 PM
6/12/2019


SAS
File Edit View Tools Solutions Window Help
Results
Log - (Untitled)
NOTE: PROCEDURE MEANS used (Total process time):
real time 0.02 seconds
cpu time 0.03 seconds
182 proc ttest data=average_alcohol alpha=0.05;
183 class Treatment;
184 var Alcohol;
185 run;
NOTE: PROCEDURE TTEST used (Total process time):
real time 1.83 seconds
cpu time 0.92 seconds
Question 3 *
/* type of test - 2 sample ttest
H0: the average number of drinks per day per individual is not equal for two groups
H1: the average number of drinks per day per individual is equal for two groups */
proc means data=control_treatment nway noprint;
var Alcohol;
class ID;
by Treatment;
output
out=average_alcohol(drop=_:)
mean=;
run;
proc ttest data=average_alcohol alpha=0.05;
class Treatment;
var Alcohol;
run;
/* We can observe that in the table for equal variances, p>alpha hence we fail to reject the null ie consider equal variances.
In the table just above we check ther values for pooled method and notice that p>alpha. Thus we fail to reject the null and
can conclude the average number of drinks per day per individual is not equal for two groups and thus the subjects are not
randomized perfectly wrt alcohol */
Output - (Untitled)
Log - (Untitled)
Question 3 *
Results Viewer - S...
E:\Users\ssr180000\Downloads
10:56 PM
6/12/2019


SAS
File Edit View Go Tools Solutions Window Help
Results
Results Viewer - SAS Output
The TTEST Procedure
Variable: Alcohol (Number of alcoholic drinks per day)
Treatment N Mean Std Dev Std Err Minimum Maximum
0 250 0.8133 1.2705 0.0804 0 6.0000
1 250 0.6440 1.1807 0.0747 0 6.0000
Diff (1-2) 0.1693 1.2264 0.1097
Treatment Method Mean 95% CL Mean Std Dev 95% CL Std Dev
0 0.8133 0.6551 0.9716 1.2705 1.1681 1.3929
1 0.6440 0.4969 0.7911 1.1807 1.0854 1.2943
Diff (1-2) Pooled 0.1693 -0.0462 0.3849 1.2264 1.1548 1.3076
Diff (1-2) Satterthwaite 0.1693 -0.0462 0.3849
Method Variances DF t Value Pr > |t|
Pooled Equal 498 1.54 0.1233
Satterthwaite Unequal 495.34 1.54 0.1233
Equality of Variances
Method Num DF Den DF F Value Pr > F
Folded F 249 249 1.16 0.2476
Distribution of Alcohol
Percent
60
40
0
Output - (Untitled)
Log - (Untitled)
Question 3 *
Results Viewer - ...
Done
E:\Users\ssr180000\Downloads
10:56 PM
6/12/2019