Unit 3 HW Solutions

Question 1 (30 points total)

In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, let's suppose that a random sample of the ages of fired and not fired people in the American Samoa Government are listed below:

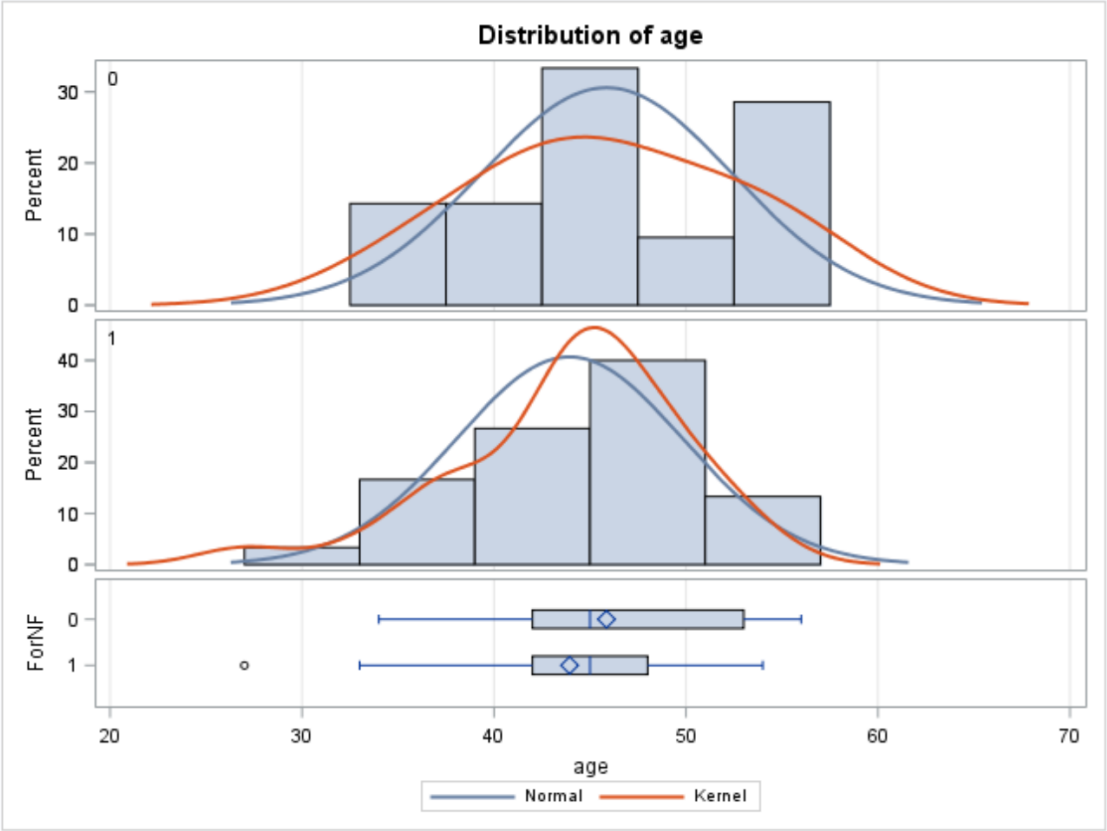
Fired
> 34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56
Not Fired
> 27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54

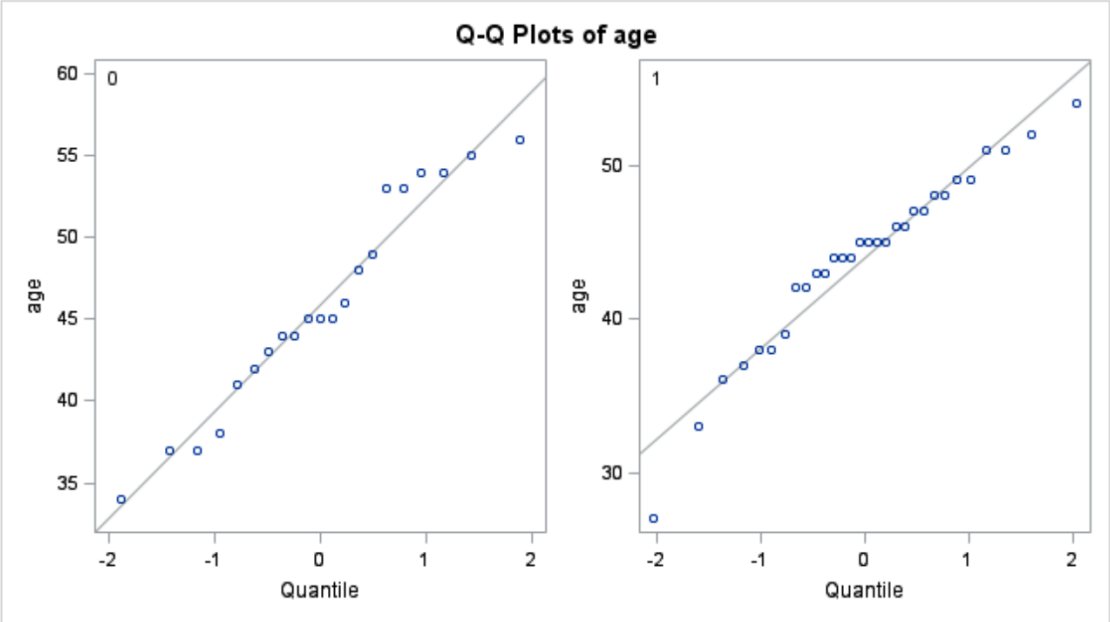
Part A (10 points total)

Check the assumptions (with SAS) of the two-sample t-test with respect to this data. Address each assumption individually as we did in the videos and live session and make sure and copy and paste the histograms, q-q plots or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate?

SAS code for histograms, q-q plots, and box plots as well as the t-test below

```
*/Assumes a data file named samoa with variables ForNF and age has been uploaded,
where 1 is Fired and 0 is Not Fired;
proc ttest data = samoa;
class ForNF;
var age;
run;
```





(3 points) Normality: There is little if any evidence from the histograms and QQ plots of any departures from normality from these data. We will assume they do come from normal distributions.

(3 points) Equal standard deviations: There is little evidence that the samples are pulled from distributions that have different standard deviations, thus we will assume that the standard deviations are equal.

(3 points) Independence: We will assume that the observations are independent both between and within groups.

(1 point) Decision: the two-sample t-test and confidence intervals are appropriate to use for these data.

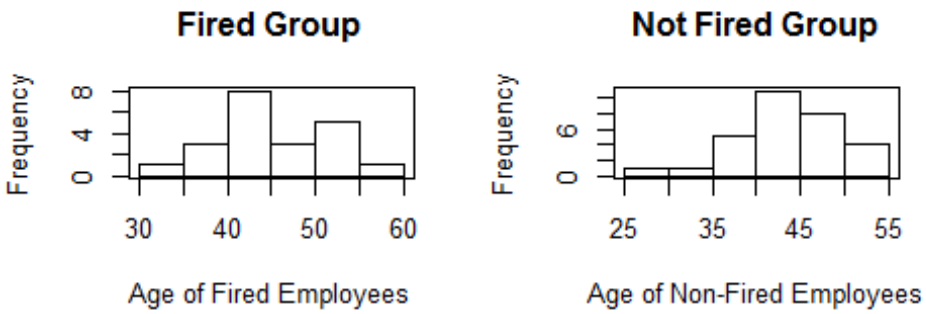
Note: Remember, your graphs do not have to look exactly the same as the answer key. There are multiple ways to create histograms in SAS - just ensure the same information is being displayed.

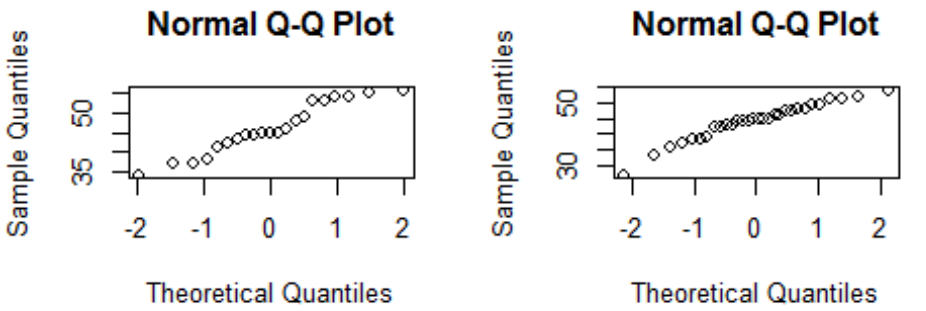
Part B (10 points)

Check the assumptions with R and compare them with the plots from SAS.

```
fired <- c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53,
53, 54, 54, 55, 56)
not.fired <- c(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45,
45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54)
label1 <- rep('fired', 21)
label2 <- rep('not.fired', 30)
label.all <- as.factor(c(label1, label2))
samoa <- data.frame(status=label.all, age=c(fired, not.fired))

par(mfrow=c(2,2))
hist(fired, xlab='Age of Fired Employees', main='Fired Group')
box()
hist(not.fired, xlab='Age of Non-Fired Employees', main='Not Fired Group')
box()
qqnorm(fired)
qqnorm(not.fired)
```





You should observe exactly the same results, aside from slight formatting differences between SAS and R. No real discussion is needed aside from reaching the same conclusion as before!

Part C (10 points)

Now perform a complete analysis of the data. You may use either the permutation test from HW 1 or the t-test from HW 2 (copy and paste) depending on your answer to part a. In your analysis, be sure and cover all the steps of a complete analysis:

1. State the problem.

2. Address the assumptions of t-test (from part A)

3. Perform the t-test if it is appropriate and a permutation test if it is not. (Judging from your analysis of the assumptions.

4. Provide a conclusion including the p-value and a confidence interval.

5. Provide the scope of inference.

(Steps 3-5 are from your previous HW; you are just putting everything together.)

NOTE: THIS QUESTION SHOULD BE EASY AS YOU ARE SIMPLY FORMATTING YOUR RESULTS FROM EARLIER IN THE ABOVE FORM. (Steps 3-5 are from your previous HW; you are just putting everything together.) IT REALLY JUST EQUATES TO ADDING A STATEMENT OF THE PROBLEM AND ADDRESSING THE ASSUMPTIONS (1 and 2 above). You can basically copy and paste the rest. We are simply putting everything together to make a complete report.

Problem (1 point): We wish to test the claim that the mean age of the fired group is different than the mean age of the not fired group for this population of American Samoa working citizens. In other words, we are testing the claim that age discrimination exists for this population. (Note that a one-sided test addresses the problem more directly, as age discrimination typically refers to preferential treatment of younger workers. However, both a one-sided test and two-sided test are acceptable for this assignment.)

The assumptions (1 point) are as stated in parts A and B. Because the assumptions for the t-tools are met, we will proceed with a t-test.

What follows is a repeat from HW #2, but you can see how everything flows together in a complete analysis. If you chose the permutation test (with an approximate p-value or exact p-value), then see the homework solutions from the permutation test.

Step 1 - Hypotheses (1 point):

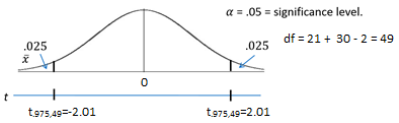
$H_o: \mu_{Fired} = \mu_{NotFired}$
$H_a: \mu_{Fired} \neq \mu_{NotFired}$

or

$H_o: \mu_{Fired} \leq \mu_{NotFired}$
$H_a: \mu_{Fired} > \mu_{NotFired}$

Note that the latter set of hypotheses more closely aligns with our question of interest, as age discrimination typically focuses on older workers being treated less favorably.

Step 2 - Identification of Critical Value (1 point): ± 2.01 **(2-sided) or** 1.677 **(1-sided)**



This graph is for a two-sided test.

Step 3 - Value of Test Statistic (1 point): $t = 1.10$

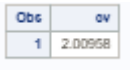
Step 4 - Give p-value (1 point): $p = 0.2771$ **(2-sided) or** $p = 0.1385$ **(1-sided)**

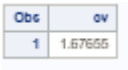
Step 5 - Decision (1 point): Fail to Reject H_o at significance level alpha = 0.05.

Step 6 - Conclusion (1 point for the statistical conclusion, 1 point for the confidence interval, 1 point for discussing the scope): On the basis of this test, there is not enough evidence to suggest that the mean ages of the fired and not fired groups are different. In other words, there is not enough evidence to suggest that there is discrimination based on age ($p = 0.2771$ **from a two-sample t-test). A 95% confidence interval for this difference is** $[-1.60, 5.44]$ **years. Since the subjects in this sample were randomly sampled, inference can be generalized to the population of all employees in the American Samoa Government.**

Note: if you did a 1-sided test and provided a 2-sided 90% CI, this interval should be [-1.01, 4.86]. The 1-sided 95% CI is [1.01, inf], although when we consider that a confidence interval should be the set of all plausible values for the difference in means, the interpretive value of this CI is not as strong as a 2-sided 90% CI.

SAS Code:	
*/For a two-sided test at alpha = 0.05; proc ttest data = samoa; class ForNF; var age; run;	ForNF / Method / Mean / 95% CL Mean / Std Dev / 95% CL Std Dev 0 / 45.8571 / 42.8886 / 48.8256 / 6.5214 / 4.9893 / 9.4173 1 / 43.9333 / 41.7364 / 46.1303 / 5.8835 / 4.6857 / 7.9093 Diff (1-2) / Pooled / 1.9238 / -1.5936 / 5.4413 / 6.1519 / 5.1389 / 7.6661 Diff (1-2) / Satterthwaite / 1.9238 / -1.6790 / 5.5266 Method / Variances / DF / t Value / Pr > |t| Pooled / Equal / 49 / 1.10 / 0.2771 Satterthwaite / Unequal / 40.268 / 1.08 / 0.2870
*/For a one-sided test at alpha = 0.05; proc ttest data = samoa sides = u; class ForNF; var age; run;	ForNF / Method / Mean / 95% CL Mean / Std Dev / 95% CL Std Dev 0 / 45.8571 / 42.8886 / 48.8256 / 6.5214 / 4.9893 / 9.4173 1 / 43.9333 / 41.7364 / 46.1303 / 5.8835 / 4.6857 / 7.9093 Diff (1-2) / Pooled / 1.9238 / -1.0107 / Infty / 6.1519 / 5.1389 / 7.6661 Diff (1-2) / Satterthwaite / 1.9238 / -1.0780 / Infty Method / Variances / DF / t Value / Pr > t Pooled / Equal / 49 / 1.10 / 0.1385 Satterthwaite / Unequal / 40.268 / 1.08 / 0.1435
*/For a 90% two-sided confidence interval to match the 0.05 one-sided hypothesis test; proc ttest data = samoa sides = u alpha = 0.1; class ForNF; var age; run;	ForNF / Method / Mean / 90% CL Mean / Std Dev / 90% CL Std Dev 0 / 45.8571 / 43.4027 / 48.3116 / 6.5214 / 5.2038 / 8.8537 1 / 43.9333 / 42.1082 / 45.7585 / 5.8835 / 4.8568 / 7.5292 Diff (1-2) / Pooled / 1.9238 / -1.0107 / 4.8583 / 6.1519 / 5.2872 / 7.3929 Diff (1-2) / Satterthwaite / 1.9238 / -1.0780 / 4.9256

*/Critical value for a two-sided test at alpha = 0.05;
data critval;
cv = quantile("T", .975, 49);
 	
proc print data = critval;
run;
	
	Obs cv
	1 2.00958

*/Critical value for a one-sided test at alpha = 0.05;
data critval;
cv = quantile("T", .95, 49);
 	
proc print data = critval;
run;
	
	Obs cv
	1 1.67655

Code for the permutation test can be found in prior assignments.	

Note: Perhaps you might be wondering at this point in the HW, "Why are we always testing the assumptions of the t-test? Is it the best test? Should we always run the t-test when we can?" These are very good questions and open questions that are up for debate! The one thing that is mathematically proven and not up for debate is that if the assumptions are met, the two-sample t-test is the most powerful (in terms of Power = 1 - beta) test in the universe at testing the claim of the difference of means. Two questions may arise here. 1. Do we every really have the assumptions fully met in the real world and just how much power do we give up at varying degrees of violation of the assumptions? 2. Do we always want inference on the equality/difference of means? We will continue to answer these questions in Chapter 4. (Also note that we started to answer number two with a t-test of log transformed data. The inference there is on the equality (ratio) of medians, which may be a better measure of center when dealing with right or left skewed data!

Question 2 (30 points total)

In the last homework, it was mentioned that a Business Stats class here at SMU was polled and students were asked how much money (cash) they had in their pockets at that very moment. The idea was to see if there was evidence that those in charge of the vending machines should include the expensive bill / coin acceptor or if they should just have the credit card reader. However, a professor from Seattle University polled her class with the same question. Below are the results of the polls.

SMU
> 34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0
Seattle U
> 20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0

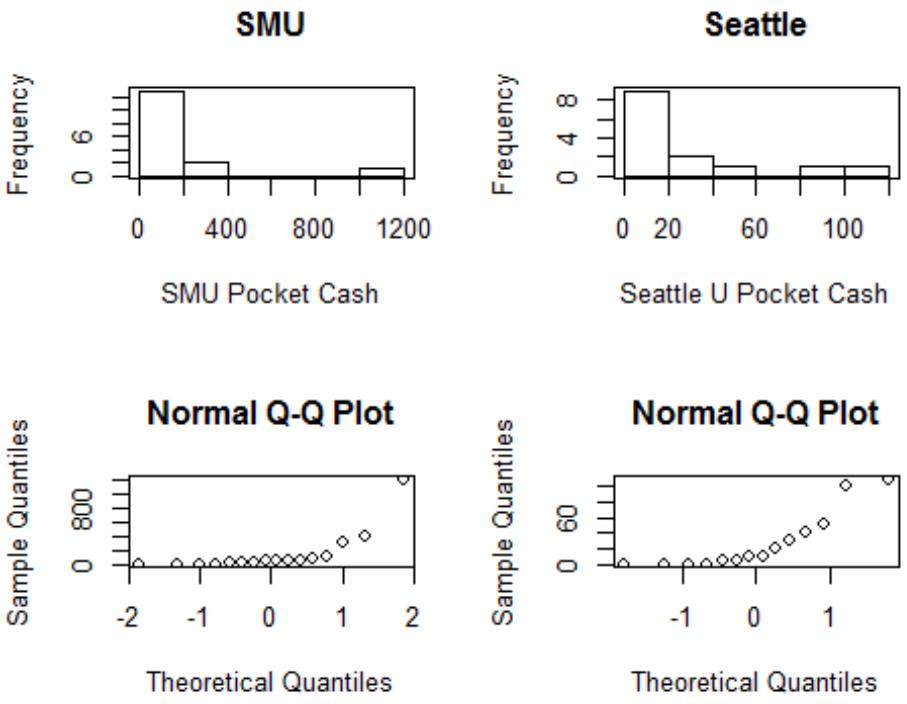
Part A (9 points total)

Check the assumptions (with SAS or R) of the two-sample t-test with respect to this data. Address each assumption individually as we did in the videos and live session and make

sure to copy and paste the histograms, q-q plots, or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate?

```
SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)
Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)
school1 <- rep('SMU', 16)
school2 <- rep('Seattle', 14)
school <- as.factor(c(school1, school2))
all.money <- data.frame(name=school, money=c(SMU, Seattle))

par(mfrow=c(2,2))
hist(SMU, xlab='SMU Pocket Cash', main='SMU')
box()
hist(Seattle, xlab='Seattle U Pocket Cash', main='Seattle')
box()
qqnorm(SMU)
qqnorm(Seattle)
```



(3 points) Normality: There is significant evidence from the histograms and q-q plots of severe departures from normality for these data. We will assume they do not come from normal distributions. In addition, the sample size does not look adequate to make the t-test robust to this assumption.

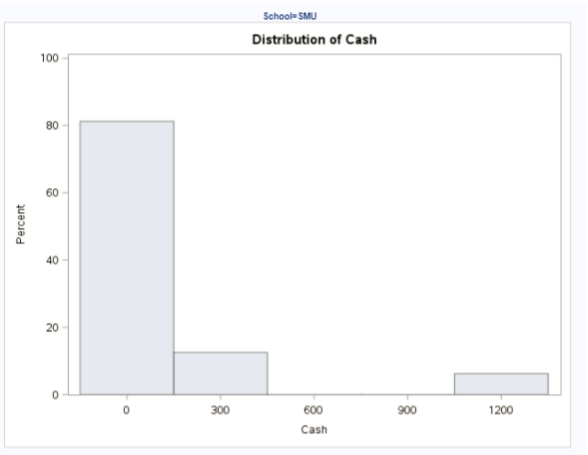
(3 points) Equal standard deviations: There is significant evidence to suggest that the standard deviations of these distributions are different.

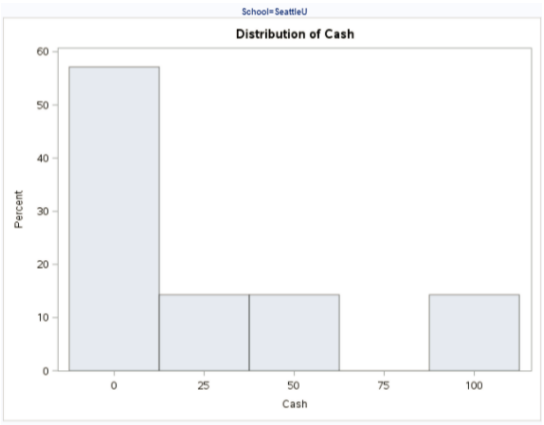
(3 points) Independence: We will assume that the observations are independent both between and within groups.

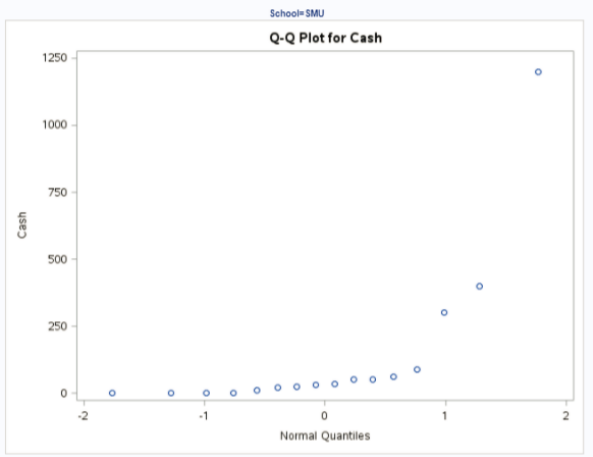
The assumptions do not appear to be met; we will NOT proceed with the t-tools.

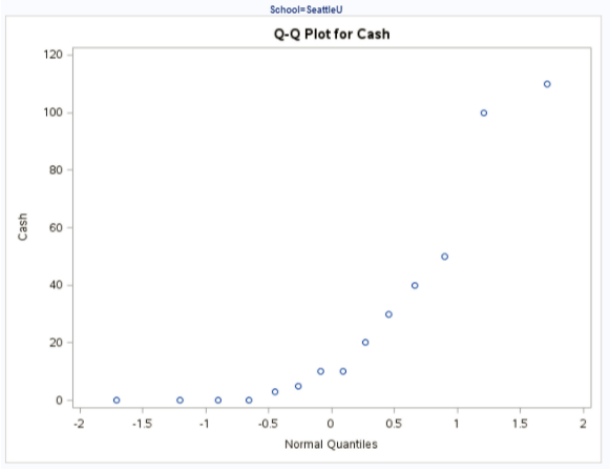
Assumptions check in SAS:

```
*To address assumptions of the t-test with histograms and q-q plots;
proc univariate data = schoolcash;
by school;
histogram cash;
qqplot cash;
run;
```

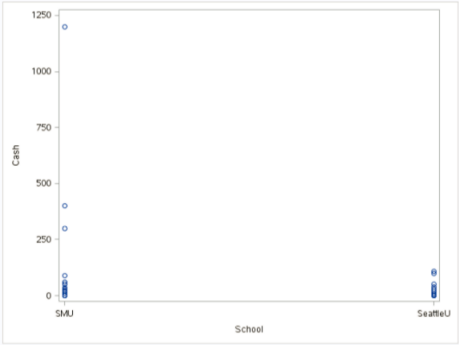









```
*To address t-test assumptions with scatter plot;
proc sgplot data = schoolcash;
scatter x= school y = cash;
run;
```



Part B (15 points total)

Now perform a complete analysis of the data. You may use either the permutation test from HW 1 or the t-test from HW 2 (copy and paste) depending on your answer to part a. In your analysis, be sure to cover all the steps of a complete analysis.

1. State the problem.

2. Address the assumptions of the t-test (from part a)

3. Perform the t-test if it is appropriate and a permutation test if it is not (judging from your analysis of the assumptions).

4. Provide a conclusion, including the p-value and a confidence interval.

5. Provide the scope of inference.

NOTE: AGAIN, THIS QUESTION SHOULD BE EASY, AS YOU ARE SIMPLY FORMATTING YOUR RESULTS FROM EARLIER IN THE ABOVE FORM. IT REALLY JUST EQUATES TO ADDING A STATEMENT OF THE PROBLEM AND ADDRESSING THE ASSUMPTIONS (1 or 2 above.) Steps 3-5 are from your previous HW; you are just putting everything together. You can basically copy and paste the rest. We are simply putting everything together to make a complete report.

Problem (2 points): Test the claim that the mean amount of pocket cash in SMU's and Seattle U's students' pockets is different.

Assumptions (3 points): The assumptions are as stated in parts A and B (you do not necessarily need to re-state them). Since the assumptions of the t-test are not met, we will instead conduct a permutation test for the difference in sample means.

What follows is an abridged version of the problem from HW #1, but you can see how everything flows together in a complete analysis. The code, output, and steps have been omitted for parsimony, but they are available in the HW #1 solutions.

Note: Remember, your p-values may be slightly different but will in all likelihood be within 0.05 of this answer key.

(10 points for the test and conclusion) To test for a difference of population means between the SMU and Seattle groups, a permutation test was conducted on 1,000 random permutations of the data. Note that a full permutation test is ideal, but SAS will likely crash due to the computing work required for this data. Hence, we perform our analysis on 1,000 random permutations of the data. A histogram of the 1,000 differences of sample

means from the 1,000 permutations can be viewed (in prior HW). The observed difference was $114, where 135 of the 1000 permutations yielded a difference in sample means that was as extreme or more extreme than this observed difference (p-value = 135/1000 = 0.135). This does not provide sufficient evidence against the null hypothesis that the mean pocket cash of SMU students is equal to that of Seattle University students. These 30 students were not from a random sample; therefore, inference cannot be extended beyond the 30 subjects in the sample. Since we failed to reject Ho, there is no need to write up whether causal inference can be drawn.

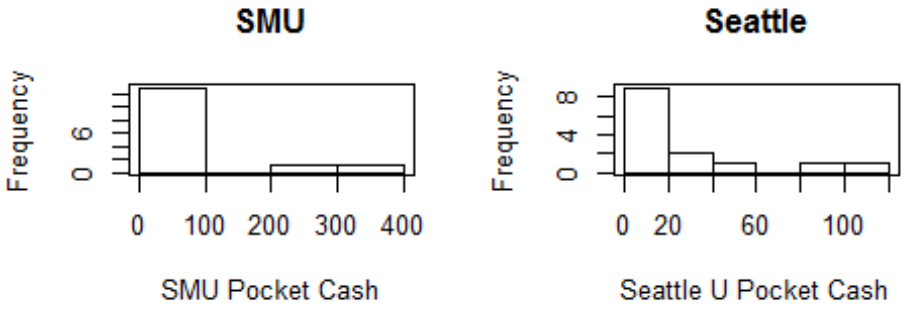
Part C (6 points total)

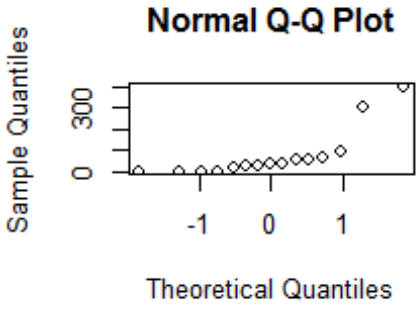
Note the potential outlier in the SMU data set. Re-check the assumptions in SAS or R without the outlier. Does this change your decision about the appropriateness of the t-tools? Compare the p-value from t-test with and without the outlier. Based on your analysis so far, what should we do with this outlier? Consult the outlier flowchart in Section 3.4.

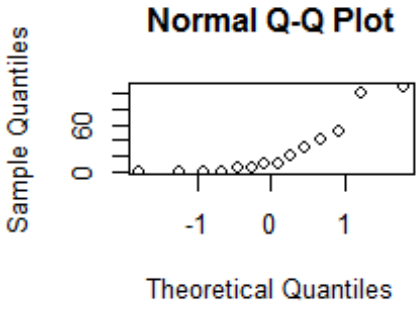
Note: 3 points for running the t-test and 3 points for reporting and discussing the results.

```
##Remove the $1,200 from SMU
SMU = c(34, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)
Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)
school1 <- rep('SMU', 15)
school2 <- rep('Seattle', 14)
school <- as.factor(c(school1, school2))
all.money <- data.frame(name=school, money=c(SMU, Seattle))

par(mfrow=c(2,2))
hist(SMU, xlab='SMU Pocket Cash', main='SMU')
box()
hist(Seattle, xlab='Seattle U Pocket Cash', main='Seattle')
box()
qqnorm(SMU)
qqnorm(Seattle)
```








```
t.test(money ~ school, data=all.money, var.equal=T)

##
##  Two Sample t-test
##
## data:  money by school
## t = -1.3402, df = 27, p-value = 0.1913
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -111.53155   23.39821
## sample estimates:
## mean in group Seattle     mean in group SMU
##              27.00000              71.06667
```

Even without the outlier, there is still considerable evidence against normality and equal standard deviations; while the sample size may be big enough to make the test robust to the normality assumption and the similar sample size may imply that the test is also robust to the equal standard deviation assumption, it is more conservative to go with a permutation test here. The p-value from the t-test with the outlier was 0.1732 and without the outlier it was 0.1913. We note that it does not make a large difference (and no difference in the decision not to reject Ho at alpha = 0.05), although we also did not think the t-test was appropriate for these data. A more appropriate analysis would be to run the permutation test with and without the outlier. This was done and the p-value with the outlier was 0.135 (above) and the p-value without the outlier was 0.261. Again, the decision is the same, thus we will keep the analysis above and report the result with the outlier.

```
*SAS Code without outlier;
*To remove outlier;
data schoolcashNoOutlier;
set schoolcash;
```

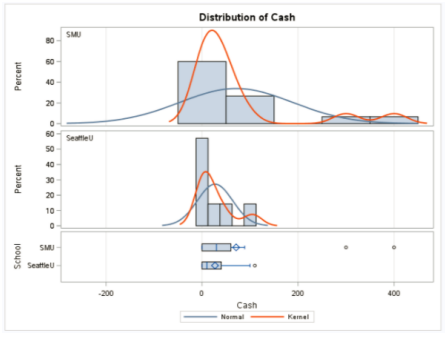
where cash < 1200;
run;

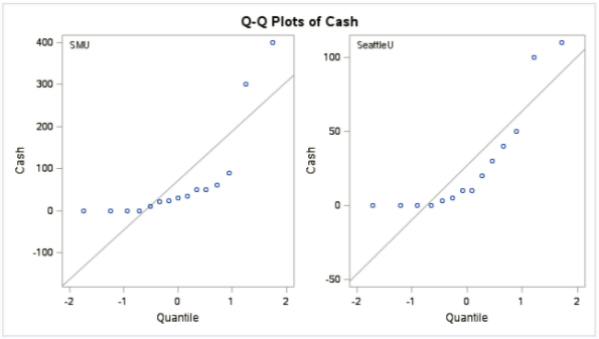
*To address t-test assumptions with scatter plot without outlier;
proc sgplot data = schoolcashNoOutlier;
scatter x= school y = cash;
run;

*To run a t-test and check assumptions of t-test with histograms and q-q plots without outlier;
proc ttest data = schoolcashNoOutlier;
class school;
var cash;
run;

School	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
SMU		71.0667	5.9029	136.2	117.7	88.1497	185.6
SeattleU		27.0000	5.7989	48.2011	36.7193	26.6198	59.1564
Diff (1-2)	Pooled	44.0667	-23.3982	111.5	88.4804	69.9543	120.4
Diff (1-2)	Satterthwaite	44.0667	-23.3333	111.5			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	27	1.34	0.1913
Satterthwaite	Unequal	16.876	1.38	0.1855





Question 3 (40 points total)

Find the "Education Data" data in the course materials. This data set includes annual incomes in 2005 of the subset of National Longitudinal Survey of youth (NLSY79) subjects who had paying jobs in 2005 and who had completed either 12 or 16 years of education by the time of their interview in 2006. All the subjects in this sample were between 41 and 49 years of age in 2006. Test the claim that the distribution of incomes for those with 16 years of education exceeds the distribution for those with 12 years of education. (Hint: pay careful attention to the ratio between the largest and smallest incomes in each group… also… is the bigger mean associated with the bigger standard deviation? Transformation?)

Note: There is some SAS code in the course materials to help you download the data into SAS. It is a very large dataset… "datalines" is not a good idea here! You could also use the File/Import option.

Finally, make sure you present your findings as you would to a client:

1. State the Problem.
2. Address the Assumptions (graphically and using words).
3. Perform the Most Appropriate (Powerful) Test (in reality, this may be a pooled t-test on the original data, a t-test on the log transformed data, or a permutation test on the original data, since these are the ones we have studied so far. For now, assume you must choose between the pooled t-test on the original data or on the log transformed data.)
4. Provide a conclusion including a p-value and a confidence interval.
5. Provide a scope of inference.

(5 points) Problem: Test the claim that the distribution of incomes for those with 16 years of education exceeds the distribution for those with 12 years of education.

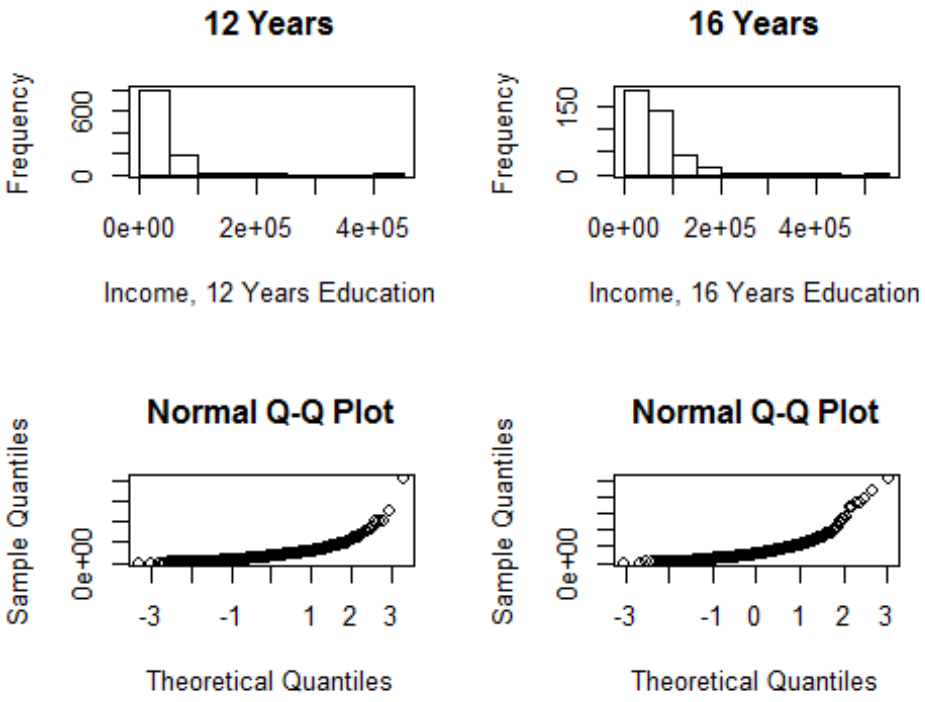
(10 points) Assumptions: In order to test the claim that the "distribution" of incomes of those with 16 years of education exceed the "distribution" of incomes with 12 years of education we will focus on the location parameters: the median. We will test if there is sufficient evidence to suggest that the median income of those with 16 years of education exceeds the median income of those with only 12 years of education. The reason to look at medians is that the t-test is not an appropriate test for the raw data. The histogram and box plot below indicate strong evidence of inequality of variance between the two populations. A quick look at sample size indicates that the smaller sample size is associated with the larger standard deviation, which is when the t-test is least robust.

```
##Read in the data, note your directory will be different
##You could've used SAS as well!

edu <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 -
Statistical Foundations for Data Science/UNIT 3/HW/EducationData.csv')

par(mfrow=c(2,2))
hist(subset(edu, Educ==12)$Income2005, xlab='Income, 12 Years Education',
main='12 Years')
box()
hist(subset(edu, Educ==16)$Income2005, xlab='Income, 16 Years Education',
main='16 Years')
box()
```

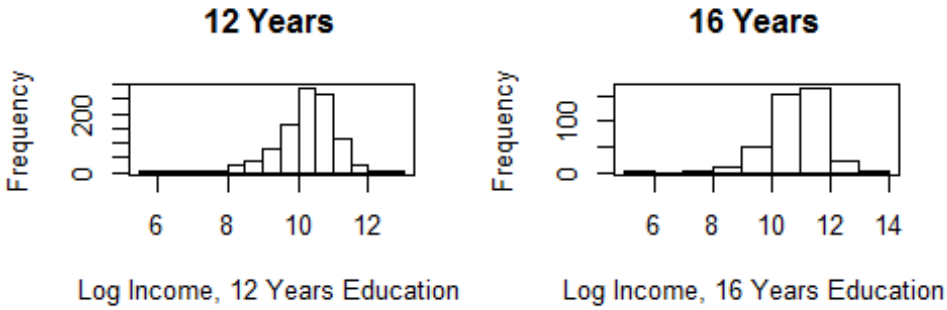
```
qqnorm(subset(edu, Educ==12)$Income2005)
qqnorm(subset(edu, Educ==16)$Income2005)
```

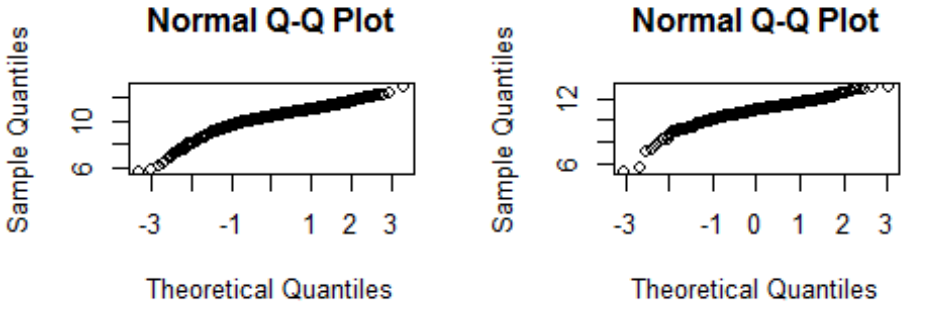


However, the larger mean is associated with the larger standard deviation, which indicates that a log transformation of the data may be appropriate. The histogram and box plot below are for the log transformed data and indicate that there is evidence that the standard deviations may be equivalent for the log transformed data. While the q-q plots looks more normal, the normality of the original data is of little concern, given that the large sample size should ensure the sampling distribution of the means is normal (from the Central Limit Theorem CLT). Since the visual check provides strong evidence for the equality of standard deviations, we will proceed with a t-test to test the difference of means of the log transformed data. Note: this is actually a test of the ratio of medians of the original data.

```
edu$log.income <- log(edu$Income2005)

par(mfrow=c(2,2))
hist(subset(edu, Educ==12)$log.income, xlab='Log Income, 12 Years Education',
main='12 Years')
box()
hist(subset(edu, Educ==16)$log.income, xlab='Log Income, 16 Years Education',
main='16 Years')
box()
qqnorm(subset(edu, Educ==12)$log.income)
qqnorm(subset(edu, Educ==16)$log.income)
```

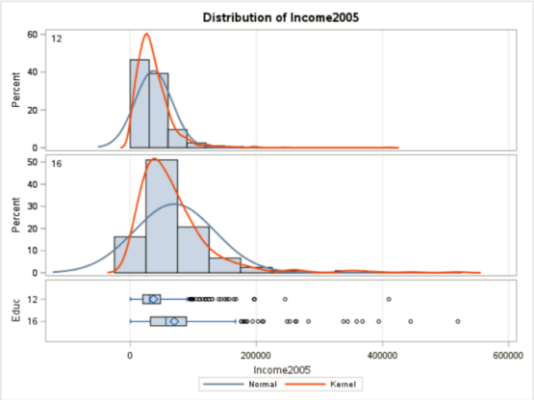


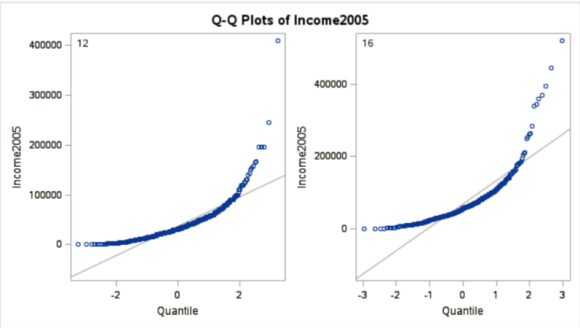



```
*To import education data;
FILENAME REFFILE '/home/sadiet0/my_courses/bsadler0/MSDS 6371/UNIT
3/EducationData.csv';
PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=education;
        GETNAMES=YES;
RUN;
```

Original Data:

```
*To check t-test assumptions on original data;
proc ttest data = education sides = l;
class educ;
var income2005;
run;
```





Logged data:

***To create a logged variable;**
data education;
set education;
logincome2005 = log(income2005);
run;

***To perform the t-test and check assumptions on log transformed data;**
proc ttest data = education sides = l;
class educ;
var logincome2005;
run;

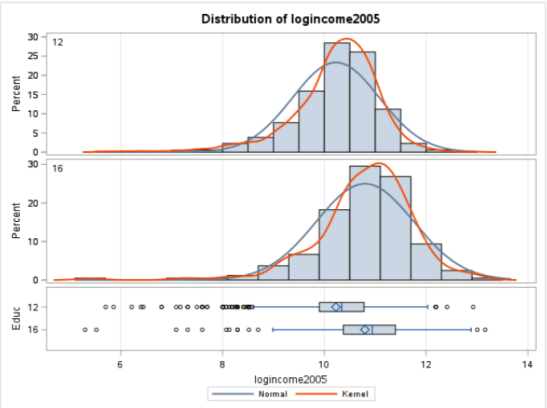
*Note that the test statistic is negative because of the way the data is sorted.

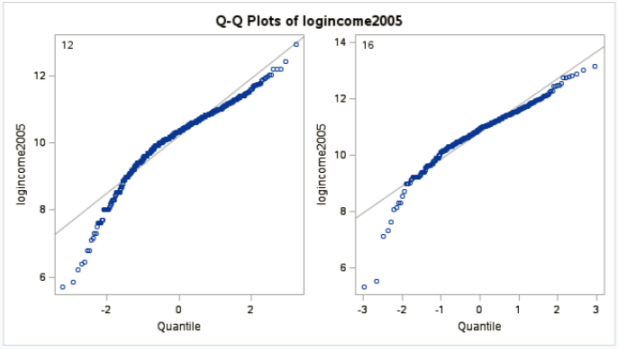
Variable: logincome2005

Educ	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
12		1020	10.2272	0.8540	0.0267	5.7038	12.9239
16		406	10.7971	0.9581	0.0475	5.2983	13.1603
Diff (1-2)	Pooled		-0.5699	0.8848	0.0519		
Diff (1-2)	Satterthwaite		-0.5699		0.0546		

Educ	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
12		10.2272	10.1747	10.2797	0.8540	0.8185	0.8927
16		10.7971	10.7036	10.8906	0.9581	0.8964	1.0290
Diff (1-2)	Pooled	-0.5699	-Infty	-0.4844	0.8848	0.8535	0.9186
Diff (1-2)	Satterthwaite	-0.5699	-Infty	-0.4800			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	1424	-10.98	<.0001
Satterthwaite	Unequal	674.82	-10.45	<.0001





***To get critical value for one-sided t-test at level alpha = 0.05;**
data critval;
cv = quantile("t", .05, 1424);
run;
proc print critval;
run;

cv
-1.64592439

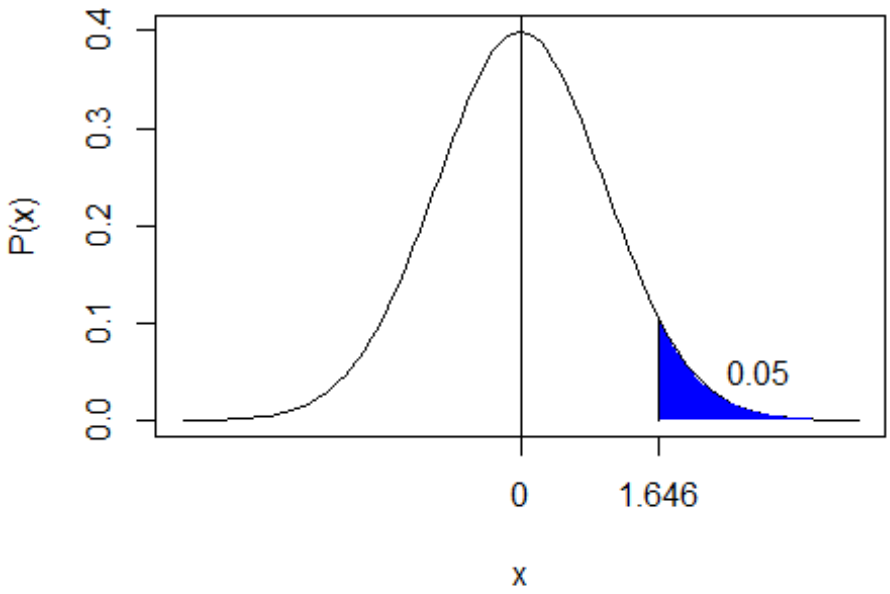
***To find a 90% confidence interval to align with an alpha = 0.05 one-sided hypothesis test;**
proc ttest data = education sides = 2 alpha = 0.1;
class educ;
var logincome2005;
run;

Educ	Method	Mean	90% CL Mean		Std Dev	90% CL Std Dev	
12		10.2272	10.1832	10.2712	0.8540	0.8241	0.8864
16		10.7971	10.7187	10.8755	0.9581	0.9060	1.0171
Diff (1-2)	Pooled	-0.5699	-0.6553	-0.4844	0.8848	0.8585	0.9131
Diff (1-2)	Satterthwaite	-0.5699	-0.6597	-0.4800			

Step 1 - Hypotheses (2 points):
$H_o: \mu_{log(16years)} = \mu_{log(12years)}$
$H_a: \mu_{log(16years)} > \mu_{log(12years)}$

Note: the null hypothesis could also be less than or equal to. If you did a 2-sided test, 2 points should be deducted here and the remaining work should be evaluated as if a 2-sided test were acceptable. (In other words, we don't want you to miss the entire problem because you tested the wrong hypothesis.)

Step 2 - Identification of Critical Value (1 point for drawing, 1 point for value): 1.646



Step 3 - Value of Test Statistic (2 points): $t = 10.98$

Step 4 - Give p-value (2 points): $p < 0.0001$

Step 5 - Decision (2 points): Reject H_o

Step 6 - Conclusion (5 points for the statistical conclusion, 5 points for the confidence interval, 5 points for discussing the scope): There is overwhelming evidence at the alpha = 0.05 level of significance ($p < 0.0001$) that the median income in 2005 for people with 16 years of education is 1.77 times as large as the median income for those in the study that had only 12 years of education. A 90% confidence interval for this factor is [e-.6553, e-

.4844]= [1.62, 1.93]. This was an observational study, and thus we cannot confirm that the years of education caused the increase in income, only that they are associated with each other. There is little detail about the randomness of the sample, although it is doubtful that it was a random sample. We must limit the inference gained from this study to only the subjects of this sample.

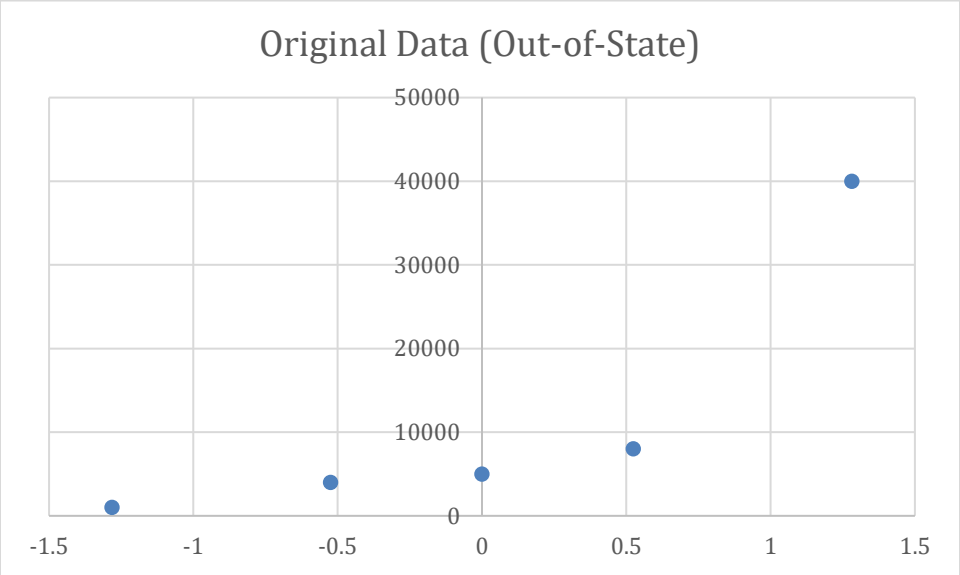
Note: an alternative way to state the conclusion would be to say the median income for those with 16 years of education was 77% larger than the median income for those that had only 12 years of education, with the confidence interval being 62% to 93%.

Note: if the order of your data was different, you would have gotten that the median income for those with 12 years of education was 0.57 times the median income of those with 16 years of education. The confidence interval would have been [$e^{-.6553}$, $e^{-.4844}$] =[0.519, 0.616]. This is completely fine as long as you interpret the results properly.

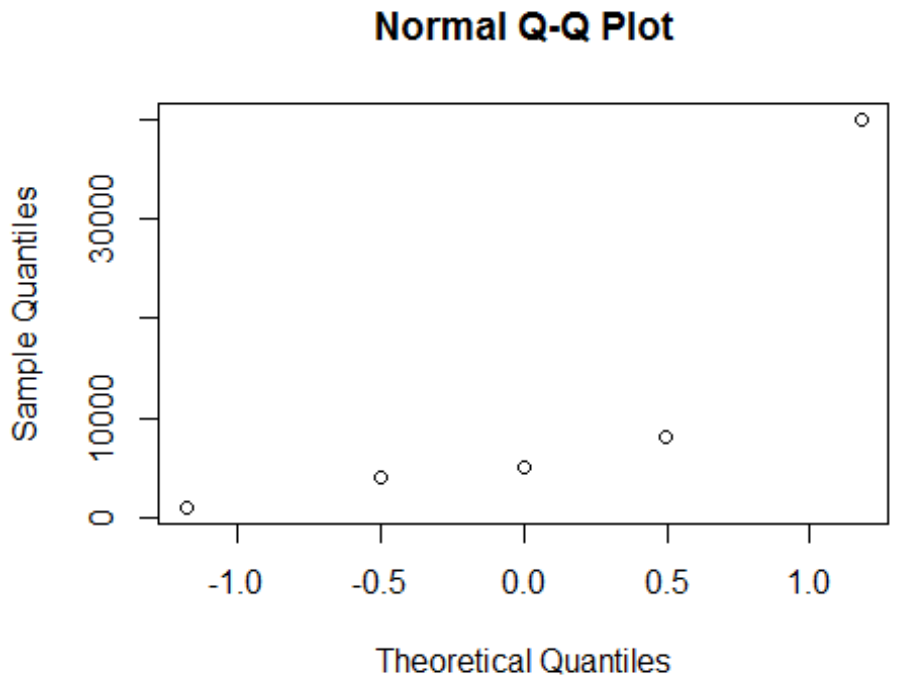
Bonus (+5 points total)

Create 2 q-q plots (by hand) for the original data in Chapter 3, question 20 of the text book. A q-q plot for the In-State and a q-q plot for the Out-Of-State data. Show all work by filling in a table like the one below (one for In-State and one for Out-of-State):

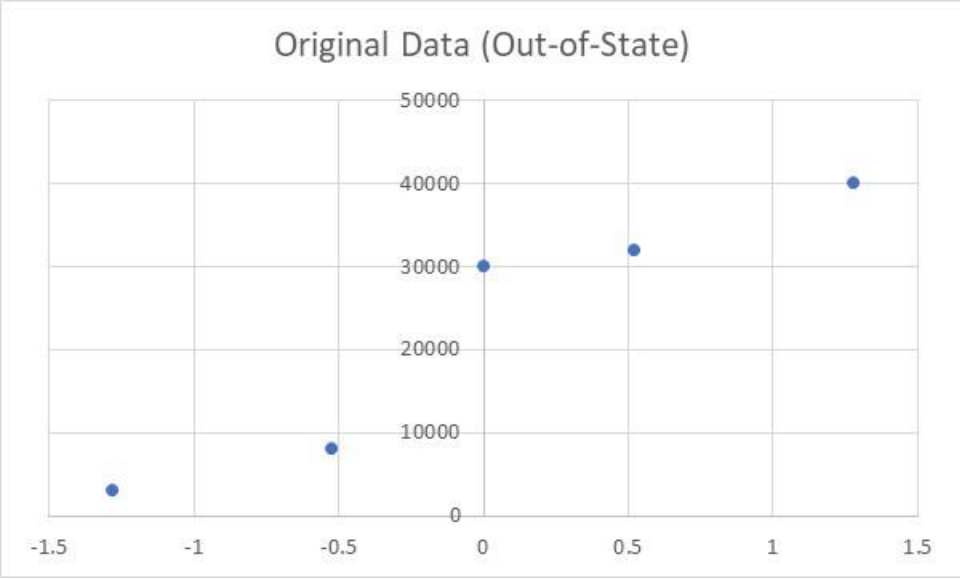
Original Data (In-State)	Percentage for Percentiles	Z-Score of Original Data	Z-Scores for Percentiles
1000	0.1	-0.65954	-1.28155
4000	0.3	-0.47288	-0.5244
5000	0.5	-0.41066	0
8000	0.7	-0.22399	0.524401
40000	0.9	1.76708	1.281552



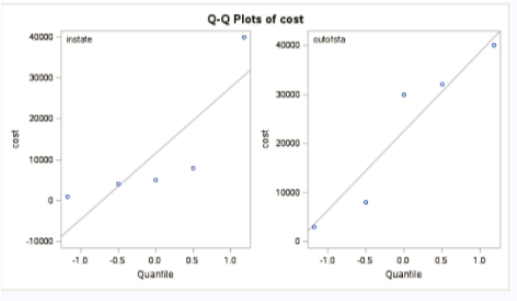

```
in.state <- c(1000,4000,5000,8000,40000)
qqnorm(in.state)
```



Original Data (Out-of-State)	Percentage for Percentiles	Z-Score of Original Data	Z-Scores for Percentiles
3000	0.1	-1.21367	-1.28155
8000	0.3	-0.90406	-0.5244
30000	0.5	0.458224	0
32000	0.7	0.582069	0.524401
40000	0.9	1.077446	1.281552



Check your q-q plots by comparing them with the ones from proc ttest. (Run proc ttest but just for the q-q plots. You do not need to run a full hypothesis test.) What would you conclude about the normality of the distributions these data came from?

*To get q-q plots; proc ttest data = tuition; class location; var cost; run;	

The out-of-state q-q plot provides more evidence of normality than the in-state. However, with a sample size of only 5, neither q-q plot provides evidence of extreme departures from normality.