

Natasha Patel, Benya Chongolnee, Jerry Hsu

Dr. Chao

STAT 350-030

Lab: Wed: 11:30-12:20, Ryan Murphy

3 December 2017

Final Project

A. **Introduction and Questions:** We drew three inference questions to determine a general conclusion pertaining to median income. The three inference situations are: median income vs. region, median income vs. urban indicator, and median income vs. percent college graduate. We will perform an ANOVA test, a 2-sample test, and a regression test, respectively, to collect data on these three inference situations. Based off of this, we will answer the following question: what general effects does median income have on an individual? This question is important to know because it will allow the government and society to cater to the needs of each income bracket.

B. **Data:**

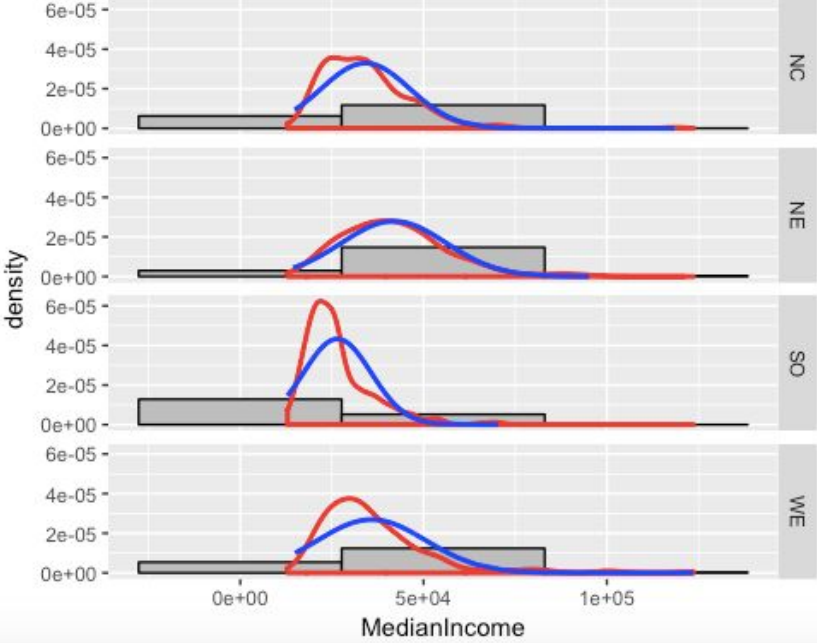
	Brief Description	Numeric or Categorical
Median Income	It is the median household income of a county.	Numeric continuous
Region	It is one of the following four regions: Northeast, North Central, South, or West.	Categorical
Urban Indicator	It is whether a county is an urban county or not.	Categorical
Percent College Graduate	It is the percent of people aged 25 and over that graduated from college.	Numeric continuous

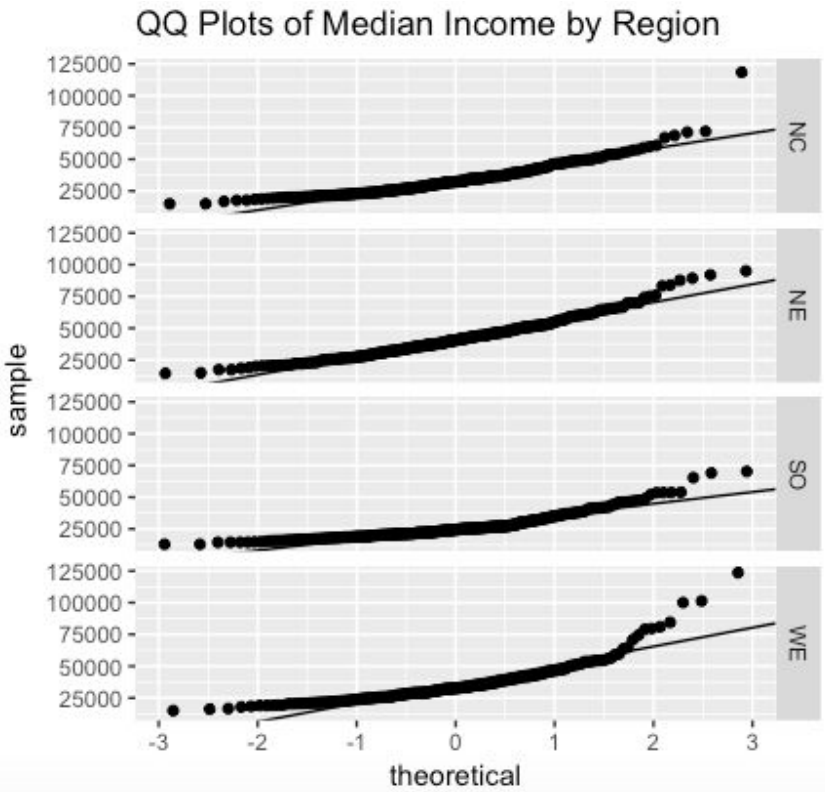
C. **Inference 1:** ANOVA of Median Income vs Region

Completed by Natasha

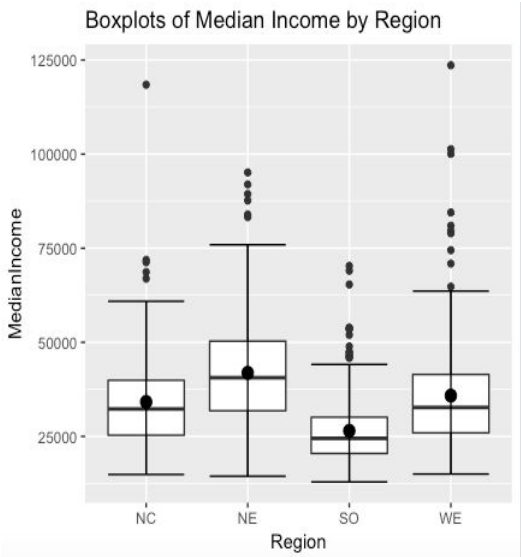
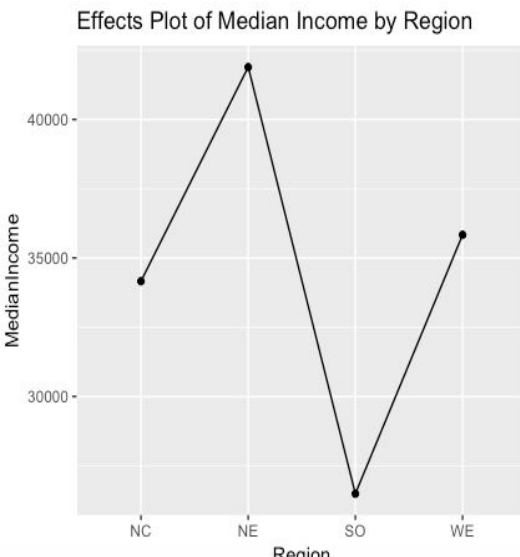
a. Code displayed in Appendix

- b. The statistical procedure being used to analyze the relationship between median income and region is a one-sided ANOVA test. An ANOVA test is used to determine if several populations have the same means by comparing how far apart the sample means are and with how much variance. In this case, we are comparing different regions (populations) to see if they have the same mean median incomes and what the variance is between them. This is a two-sided inference because the alternate hypothesis is that at least two μ_i 's are different.
- c. There are three assumptions for an ANOVA test: each population comes from an independent SRS, the populations have a normal distribution, and all the populations have the same unknown variance. I have outlined each assumption in the table below to prove that all of the assumptions are satisfied:

Independent SRS	<i>Assumed to be true</i> SATISFIED
Normal populations	<p style="text-align: center;">Histograms of Median Income versus Regions</p> 

	<p style="text-align: center;">QQ Plots of Median Income by Region</p>  <p>Based off of the histogram and the QQ plots, we can say that the populations are normal. The red and blue curves on the histogram are similar, meaning that the populations match the normal distribution. The plots on the GG plots fall approximately in a straight line, meaning that the populations match the normal line.</p> <p>SATISFIED</p>
<p>Variance</p>	<pre> > tapply(data_cleaned.subset\$MedianIncome, data_cleaned.subset\$Region, length) NC NE SO WE 260 299 307 232 > tapply(data_cleaned.subset\$MedianIncome, data_cleaned.subset\$Region, mean) NC NE SO WE 34161.07 41890.14 26489.23 35835.56 > tapply(data_cleaned.subset\$MedianIncome, data_cleaned.subset\$Region, sd) NC NE SO WE 12127.624 14278.279 9218.282 14844.251 </pre> <p>We can test that all the populations have the same variance by using the following equation:</p> $S_{max}/S_{min} < 2$ $14844.251/9218.282 < 2$ $1.610 < 2$ <p>SATISFIED</p>

- d. I chose to display the data using a boxplot and an effects plot to see if there is evidence to suggest that the median incomes vary by region.

Boxplot	Effects Plot
	
<p>Based off of these boxplots, it looks like the median incomes for each region are very similar to one another. We can use an ANOVA test to determine if there is evidence to suggest that the median income levels are actually the same or if they are different.</p>	<p>Based off this effects plot we can see that there is some variation in the median incomes for each region. The NE has the highest and the South has the lowest with more than a \$10,000 difference between the two. We can use an ANOVA test to determine if there actually is a difference in the median incomes for each region.</p>

- e. ANOVA Hypothesis Test:

Step 1: Definition of the terms

μ_{NE} is the population mean median income for the Northeast region

μ_{NC} is the population mean median income for the North Central region

μ_{SO} is the population mean median income for the South region

μ_W is the population mean median income for the West region

Step 2: State the hypotheses

Ho: $\mu_{NE} = \mu_{NC} = \mu_{SO} = \mu_W$

Ha: at least two μ_i 's are different

Step 3: Find the test statistic, p value, report DF

```
> summary(fit)
              Df    Sum Sq   Mean Sq F value Pr(>F)
Region          3 3.647e+10 1.216e+10   75.68 <2e-16 ***
Residuals    1094 1.758e+11 1.606e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fts= 75.68

P value= $2e^{-16}$

DF= 3 (1) and 1094 (2)

Step 4: Conclusion

$\alpha = 0.05$

Since $2e^{-16}$ is smaller than the significance level of 0.05, we should reject the null hypothesis (Ho). The data provides evidence that the population mean median incomes of at least one of the regions is different from the rest.

Multiple Comparison Test (Tukey):

I chose the Tukey method because I want to compare the population mean for each region to each other, not to a control group.

```
Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = MedianIncome ~ Region, data = data_cleaned.subset)

$Region
      diff      lwr      upr    p adj
NE-NC  7729.071  4963.654 10494.488 0.0000000
SO-NC -7671.838 -10420.445 -4923.231 0.0000000
WE-NC  1674.487 -1270.809  4619.782 0.4604920
SO-NE -15400.909 -18050.681 -12751.137 0.0000000
WE-NE  -6054.584 -8907.866 -3201.303 0.0000004
WE-SO   9346.325   6509.332 12183.317 0.0000000
```

Zero is in the interval for one pairing: West-North Central. This means that we have evidence to suggest that these two regions could have the same population

mean median income. However, since zero does not lie in the interval for any other pair, we can say that we have evidence that the other population means are different from one another.

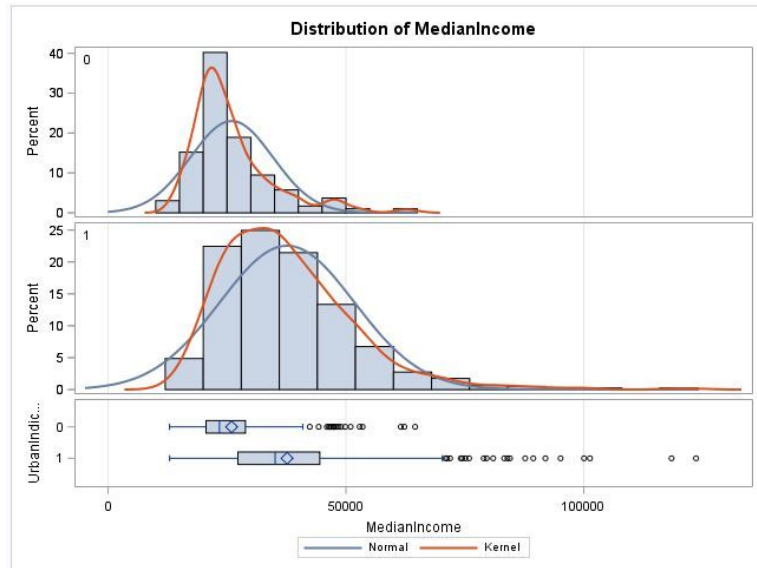
- f. Based off of the ANOVA test (hypothesis and multiple comparison test), we can conclude that the sample mean median incomes for the four different regions are not the same. This means that each region has different median incomes, which suggests that region plays a factor into this. Even though the results from the Tukey test suggest that the West and North Central regions could have the same median incomes, it is unlikely because the difference between the two is a positive value. This is only possible if the two regions have different median incomes. Referring back to our conclusion, we can say that the region a person lives in affects what their median income will be. More specifically, we can say that the Northeast tends to have higher median incomes and the South tends to have lower median incomes. This is important for the government and businesses to recognize because it means that adjustments will have to be made to prices depending on the location of their target market.

D. Inference 2: 2-sample test Urban Indicator vs Median Income

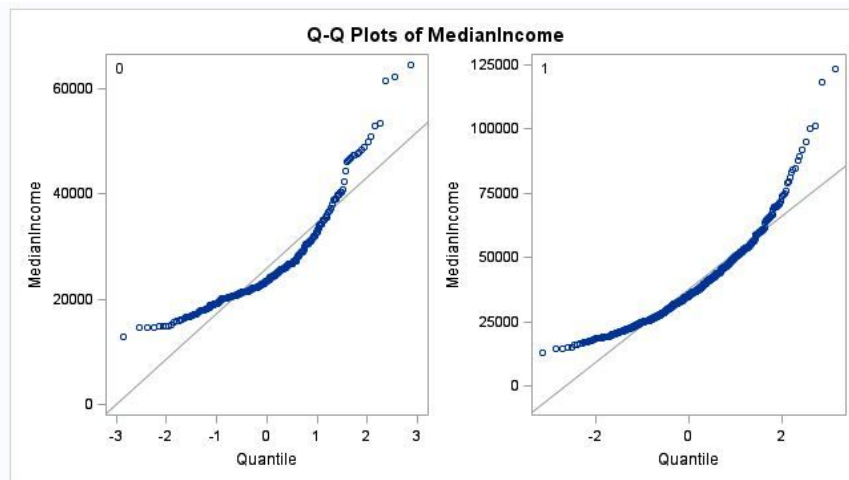
Completed by Benya

- a. Code displayed in Appendix
- b. In order to compare the median income and whether or not they are in the urban or rural area, the two-sample test must be performed. Two-sample test is performed when the data are not dependent on each other. The Urban Indicator are 2 completely different different category. The Urban Indicator in the US Data set is 1 and 0, meaning they are either in the urban area or rural area. We should use a two-sided alternative for this analysis. This is because the null hypothesis is that each urban indicator has the same/similar median income, while the alternative should be that they are significantly difference, making one urban indicator (rural or urban area) less than OR greater than another region.
- c. Assumptions for 2 sample t-test for independent variables are as follow:

- i. SRS: assumed correct
- ii. Urban indicator 0 and 1 are assumed to be independent
- iii. Normal:



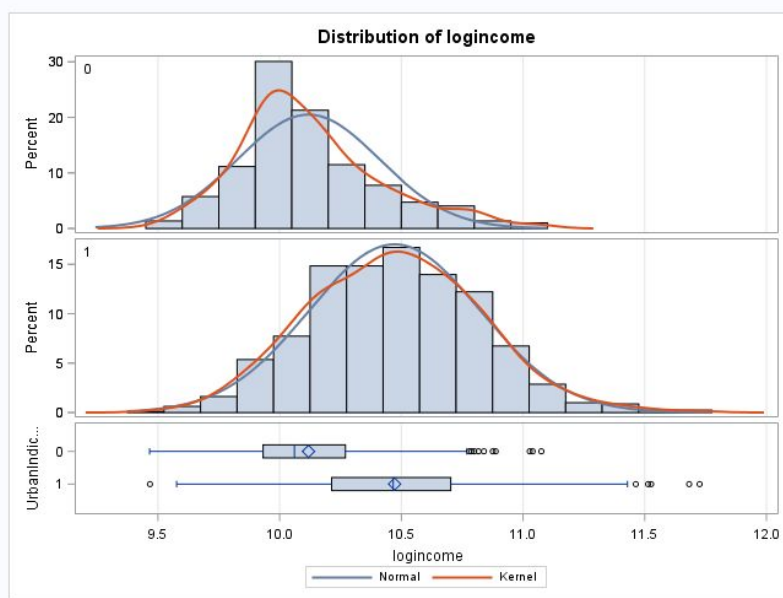
Histograms and boxplots



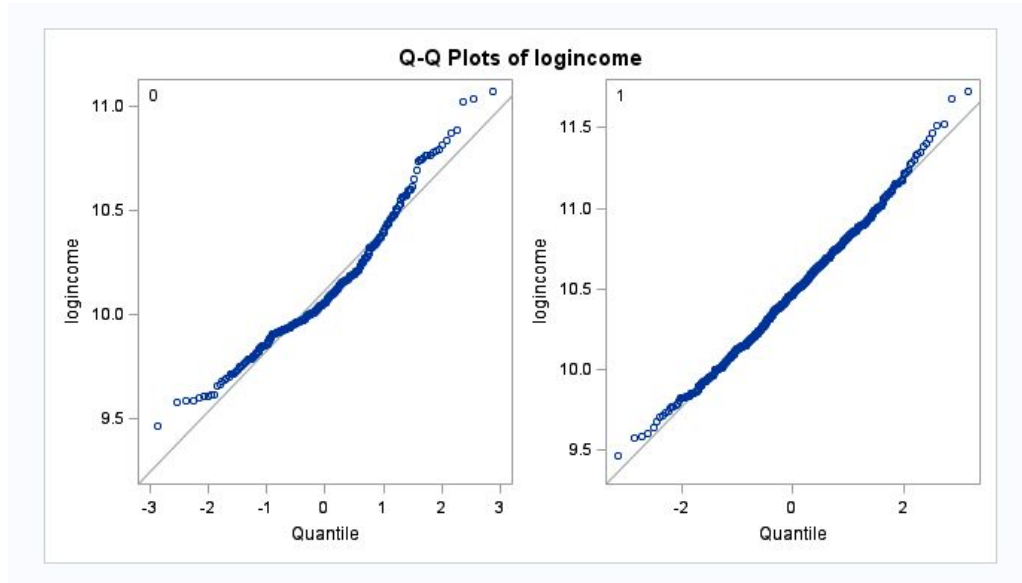
Q-Q plots

Though the data is a simple random sample, the data is not normally distributed from the graphs above. First, the histogram are not symmetrical, showing that that the data is not normal. Next, the box plot shows a few outliers and that the median is not approximately close to the mean. These attributes shows even more reasons why the data

is not normally distributed. Lastly, in the Q-Q plots above, the points are skewed right. Because of these reasonings, it can be concluded that the data are not normally distributed. Because it is not normally distributed; therefore, we are unable to perform a t-test, I will perform a log transformation in order to make the data normally distributed. Performing a log transformation will be able to better distribute the data points. Log transformation is useful when the data is highly skewed, which is the case with the median income data. The following graphs is when the data have gone through a log transformation.



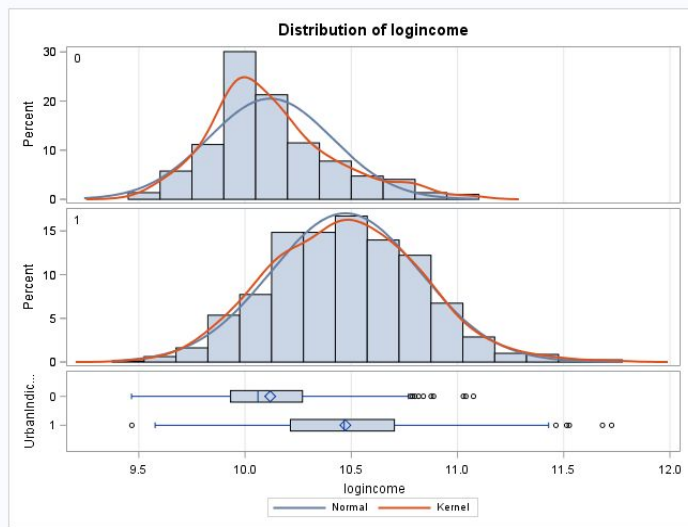
Histograms and boxplots after a log transformation



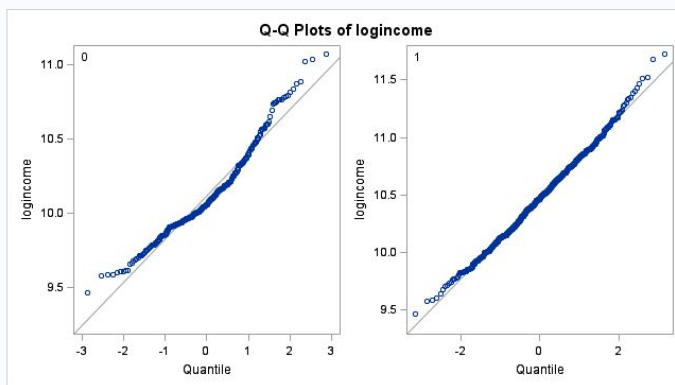
Q-Q Plots after a log transformation

Because it is assumed that the data is SRS, we must next see if the log data is normally distributed in order to ensure that all of the assumptions are met in order to perform a 2-sample t-test. From the 2 graphs above, it can be stated that the data is now normalized after the log transformation. The histogram are now approximately symmetric and the mean and median are now approximately equal. The Q-Q plots shows that the graph are not longer skewed and that the points are closely align to the normal line. Because of these reasonings, the data is now normally distributed. Since the data is normal and that it is from a simple random sample, it is safe to continue with the t-test.

d. *The following graphs are the same graphs as the above 2 graphs.*



The picture on the right has two parts. The histogram and the boxplot. The histogram shows that the median income in the urban indicator of 1 and 0 are very different. They are differently distributed. For example, about 30% of the people in the urban indicator of 0 has a log income of ~10. At the same time, only ~5% of the population in the urban indicator has a log income of ~10. Though both histograms are normally distributed, this shows that the data in urban indicator 0 and 1 are not similar. Similar conclusions can be made with the box plot. The mean/media of the log income of urban indicator 0 is lower than the urban indicator 1. This means that on average, the income of the people that lives in the urban indicator 0 has less income than the people in urban indicator 1.



The Q-Q plot shows that the log income data is normally distributed. It also shows the outliers in each urban indicator.

e. Confidence interval and hypothesis test

Confidence Interval:

UrbanIndicator	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		10.1185	10.0851	10.1519	0.2920	0.2702	0.3176
1		10.4721	10.4478	10.4965	0.3515	0.3351	0.3696
Diff (1-2)	Pooled	-0.3536	-0.3985	-0.3087	0.3366	0.3230	0.3513
Diff (1-2)	Satterthwaite	-0.3536	-0.3949	-0.3123			

From the data in the table above, we are 95% confidence that the population mean of the log median income for Urban Indicator 0 lies in between 10.0851 and 10.1519. We are also 95% confidence that the population mean of the log median income for Urban Indicator 1 lies in between 10.4478 and 10.4965. With that being said, we are 95% confidence that the difference in population mean of the log median income for Urban Indicator 0 and 1 lies in between -0.3949 and -0.3123. Since the confidence interval of the difference between Urban Indicator 0 and 1 does not include 0, we can reject the null hypothesis which states that the two urban indicator's median income are equal.

Hypothesis Test:

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	1096	-15.45	<.0001
Satterthwaite	Unequal	628.73	-16.82	<.0001

Step 1:

μ_{zero} = population mean of log median income of the urban indicator 0.

μ_{one} = population mean of log median income of the urban indicator 1.

Step 2: Null Hypothesis = $\mu_{zero} - \mu_{one} = 0$

Alternative Hypothesis = $\mu_{zero} - \mu_{one} \neq 0$

Step 3:

Test Statistics: -16.82 (given in the table above)

Degree of Freedom: 628.73 (given in the table above)

P-Value: <0.0001 (given in the table above)

Step 4:

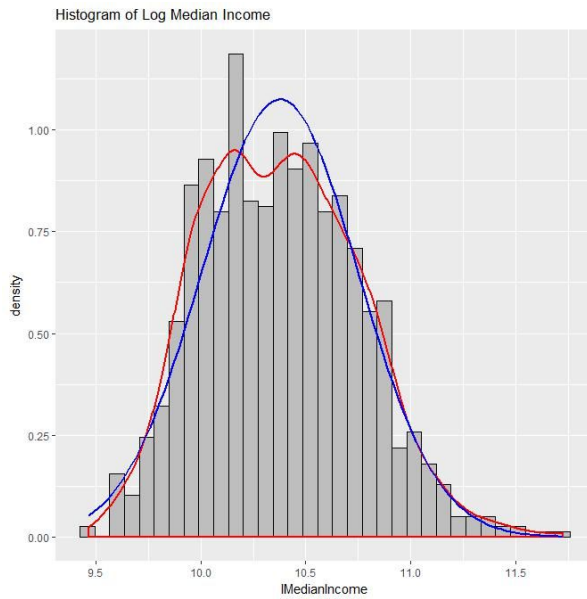
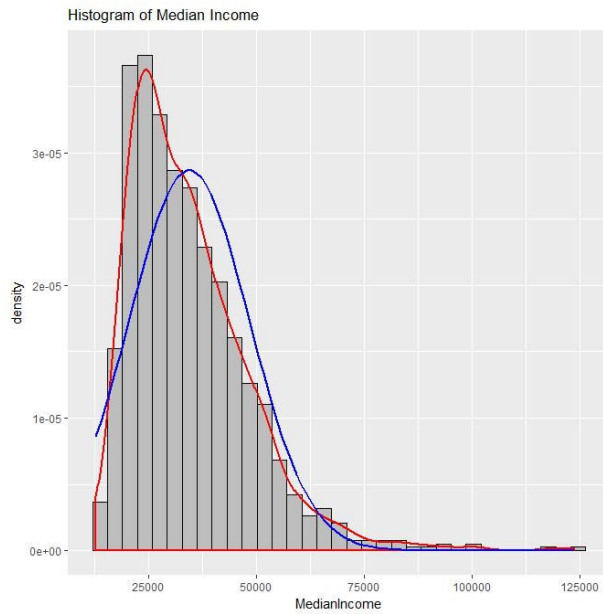
Since 0.0001 (the P-Value) is less than 0.05 (alpha), we will reject the null hypothesis which states that the median income in Urban Indicator 0 and 1 are equal. Because of this conclusion, there is enough evidence to prove that the two urban indicator's median income are not equal.

f. By doing the 2-sample independent t-test, it can be safe to state that the median income of the Urban Indicator 0 and 1 are not equal. This does not prove that the urban indicator determines the median income. There could be many other variable that could contribute to median income. What this test does is prove that a person that lives in Urban Indicator 0 on average, do not have the same income as a person that lives in Urban Indicator 1. By knowing this information, the government may want to look as to why the median income in both Urban Indicator are not equal. They would want to close that gap down in order to create a better place for people to live in.

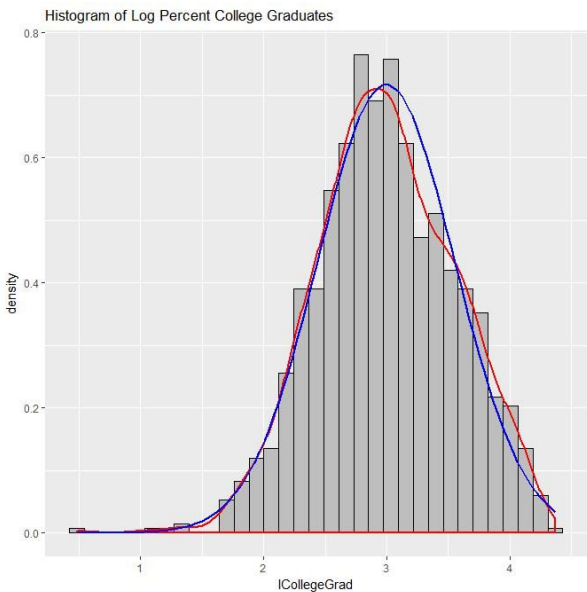
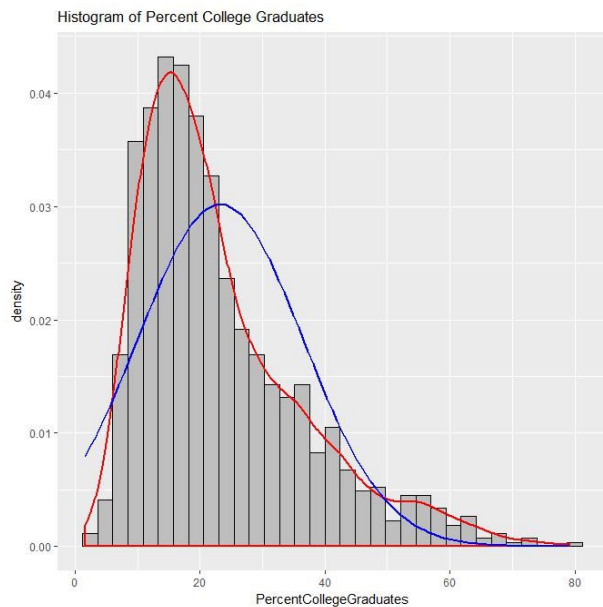
E. **Inference 3:** Regression of Median Income vs Percent College Graduate

- a. Code displayed in Appendix
- b. A regression should be used because because we want to know the relationship between the median income and percent college graduate with the former as the explanatory variable and latter as the response variable. As education costs start to be a non issue in the higher incomes, we will be restricting our inference to people with an income between \$20420 and \$41320, the federal poverty line for 3 and 8 member families.
- c. Assumptions and Transformations
 - i. Transformations
 1. Median Income will need a logarithmic transformation. Looking at the histogram of the median income before the transformation, the distribution appears to be unimodal but heavily skewed to the

right. Centered with a logarithmic transformation, the transformed data histogram is unimodal, and symmetric.

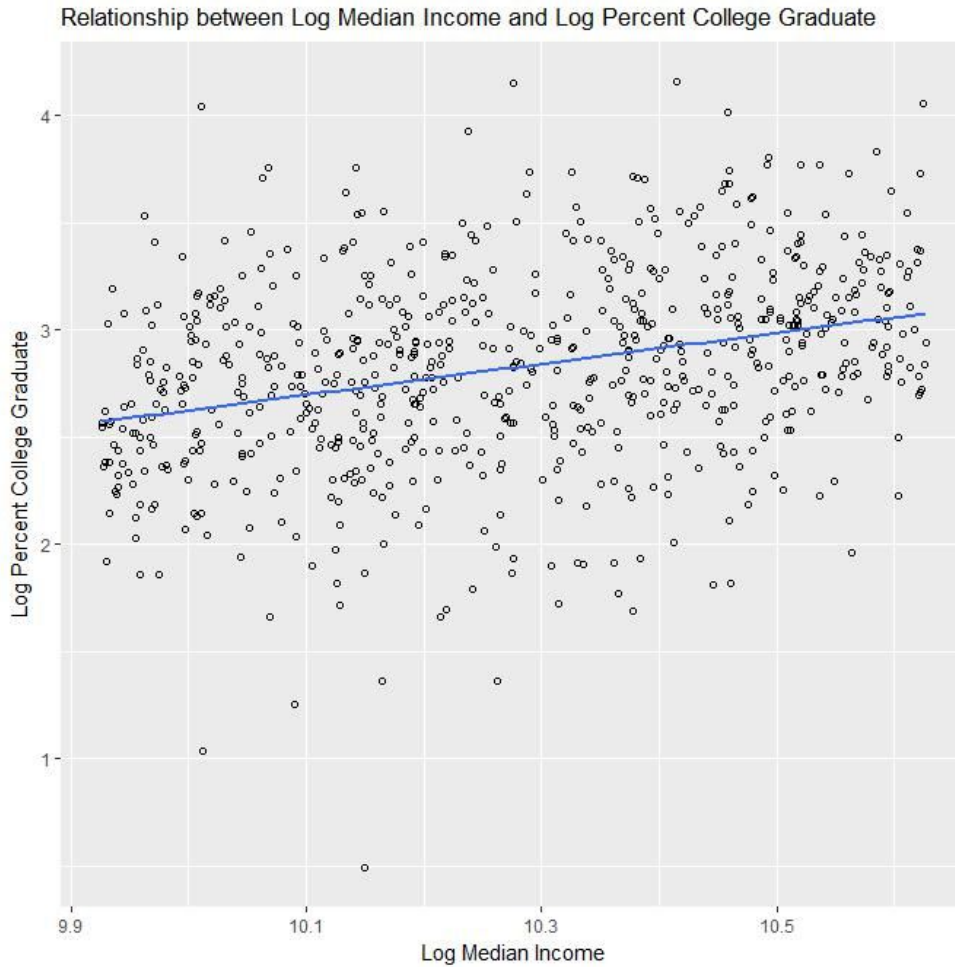


- Percent Graduate over 21 will need a logarithmic transformation. Looking at the histogram of the percent graduates before the transformation, the distribution appears to be unimodal but heavily skewed to the right. Centered with a logarithmic transformation, the transformed data histogram is unimodal, and symmetric.

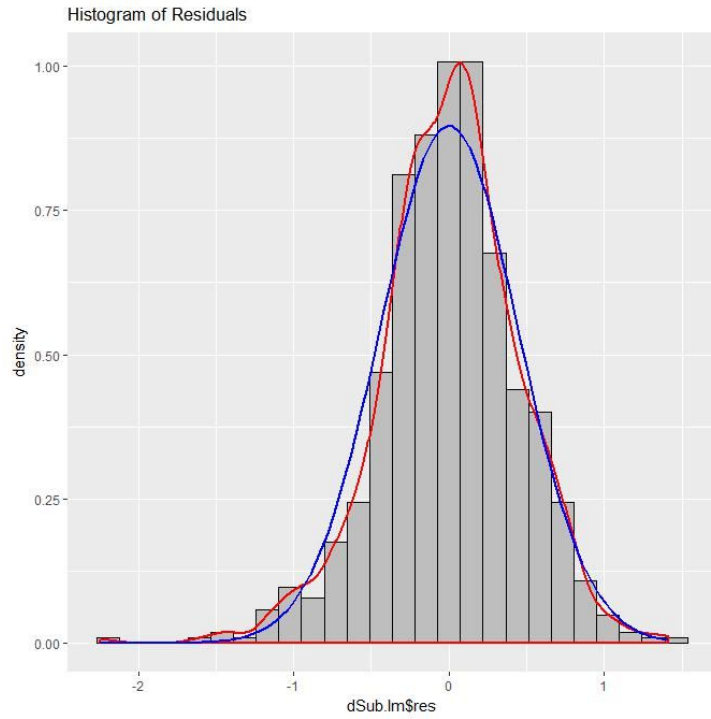


ii. Assumptions

1. A SRS is assumed to be true. Each county is also assumed to be independent of one another.
2. Looking at the scatterplot, the relationship appears to be linear, but, because of its near horizontal nature, it does not have a high coefficient of correlation at 0.3101.

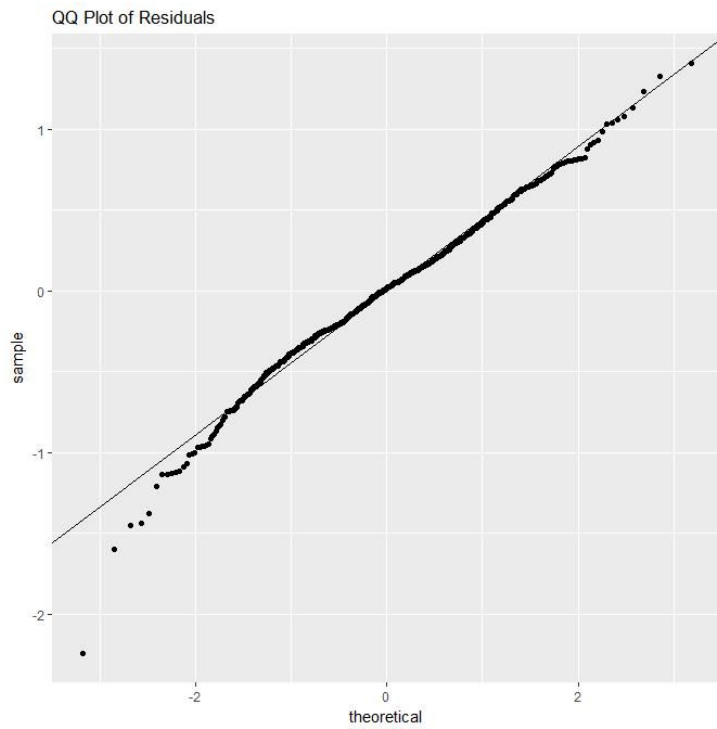


3. Looking at both the histogram and the QQ plot, the residual appears to be normally distributed in general. This shows that the response is normally distributed around the regression line.



a.

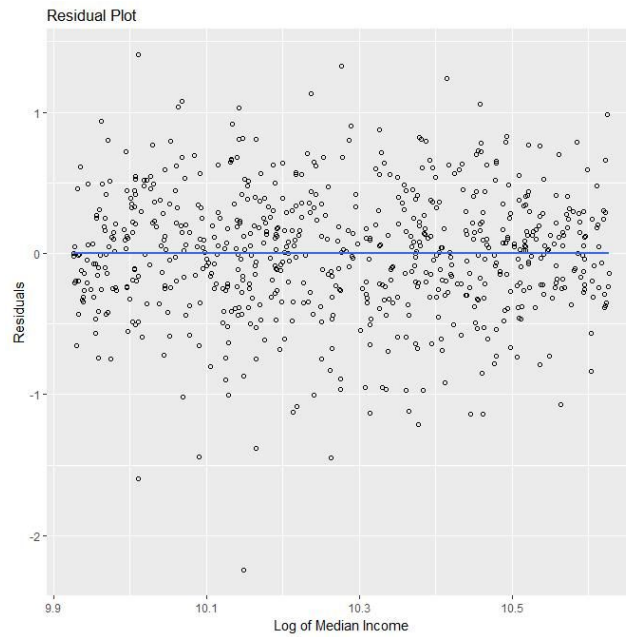
The histogram of the residual appears to be unimodal, and symmetric with no obvious outliers.



b.

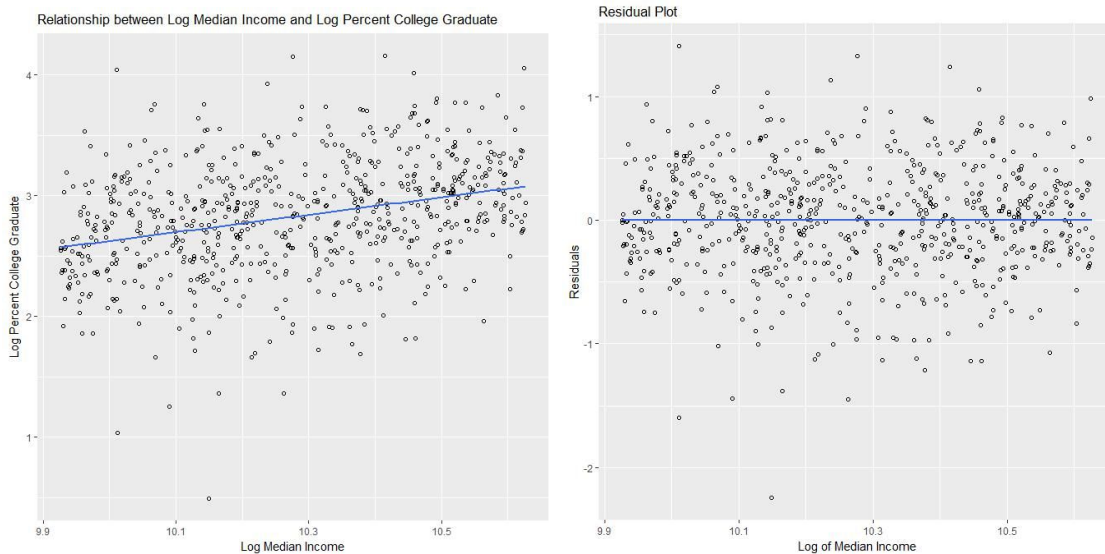
The qq appears to be normally distributed with most data lying close to or on the diagonal.

4. The standard deviation of the residual appears to be constant.



The residual plot is of equal variance throughout.

d. For a regression, it is appropriate to use a scatter plot to visualize the data and residues.



There does not appear to be any discernable pattern in the residual plot.

e. Test to see if the slope is zero. This indicates if the two are independent or not.

Step 1: Definition of the terms

Let μ_1 be the population slope.

Step 2: State the hypotheses

$H_0: \mu_1 = 0$

$H_A: \mu_1 \neq 0$

Step 3: Find the Test Statistic, p-value, report DF

```
call:
lm(formula = lCollegeGrad ~ lMedianIncome, data = dSub)

Residuals:
    Min       1Q   Median       3Q      Max
-2.24426 -0.25186  0.01741  0.27768  1.41009

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.61031    0.86361  -5.338 1.27e-07 ***
lMedianIncome  0.72353    0.08403   8.610 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4452 on 697 degrees of freedom
Multiple R-squared:  0.09614,    Adjusted R-squared:  0.09484
F-statistic: 74.13 on 1 and 697 DF,  p-value: < 2.2e-16
```

df1=1

df2=697

$t_s = 8.610$

Equation of line: Predicted percent graduate = $-4.61031 + 0.72353 * (\text{median income})$

P-value ≈ 0

Step 4: Conclusion

$\alpha = 0.01$

$0 < 0.01$

With a p-value of close to 0, there is strong evidence for rejecting the null hypothesis the population slope is 0 in favor of the alternative that the slope is not 0. This suggests that there is an association between the log median income and log percent college graduates over 21.

- f. In conclusion, the tests suggest that there is an association between log median income and log percent college graduates over 21 in the domain of our restrictions. More specifically, for every increase of one in log median income, the log of percent college graduates is predicted to increase by 0.72353. This, however, does not imply that median income directly causes in a higher percent of college graduates.

F. Final conclusion

The three tests done (ANOVA test, 2-sample test, and regression test) effectively explain the general effects that median income have on an individual. The conclusion that comes from the tests varies. First, the ANOVA test concluded that the sample mean median incomes for four different regions are not the same. This means that a person's income may be estimated using the region they live in. The 2-sample independent t-test concluded that the median income in Urban Indicator 0 and Urban Indicator 1 are also not equal. This also means that a person's income may be estimated using their urban indicator as well (whether they live in an urban area or rural area). Lastly, the regression test concluded that there is an association between median income and college graduates over 21. Overall, we can say that the median income has a general effect on these three variables. The reasoning behind why they have an effect is inconclusive due to not having enough data given; however, it can be concluded that the median income definitely has an effect on each region, urban indicator, and college graduates over 21. All of these conclusions are very important to know especially for the government because it will allow the government to cater to the needs of each income bracket in order to create a better environment for the people that live in the population. The data found from these tests will be able to be utilized in order to understand each income brackets better.

G. Appendix:

Codes:

a. **Inference 1:** ANOVA of Median Income vs Region

Code in R

```
data_cleaned$Region <- as.factor(data_cleaned$Region)

data_cleaned_subset <- subset(data_cleaned,
  Region == "NE" | Region == "NC" | Region == "SO" | Region == "WE",
  select = c("Region", "MedianIncome"))

library(ggplot2)
xbar <- tapply(data_cleaned_subset$MedianIncome, data_cleaned_subset$Region, mean)
s <- tapply(data_cleaned_subset$MedianIncome, data_cleaned_subset$Region, sd)

ggplot(data_cleaned_subset, aes(x = Region, y = MedianIncome)) +
  geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  stat_summary(fun.y=mean, col="black", geom="point", size = 3) +
  ggtitle("Boxplots of Median Income by Region")

ggplot(data = data_cleaned_subset, aes(x = Region, y = MedianIncome)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line", aes(group = 1)) +
  ggtitle("Effects Plot of Median Income by Region")

tapply(data_cleaned_subset$MedianIncome, data_cleaned_subset$Region, length)
tapply(data_cleaned_subset$MedianIncome, data_cleaned_subset$Region, mean)
tapply(data_cleaned_subset$MedianIncome, data_cleaned_subset$Region, sd)

xbar <- tapply(data_cleaned_subset$MedianIncome, data_cleaned_subset$Region, mean)
s <- tapply(data_cleaned_subset$MedianIncome, data_cleaned_subset$Region, sd)
data_cleaned_subset$theoretical.density <- apply(data_cleaned_subset, 1, function(x){
  dnorm(as.numeric(x["MedianIncome"]), xbar[x["Region"]], s[x["Region"]])})
ggplot(data_cleaned_subset, aes(x = MedianIncome)) +
  geom_histogram(aes(y=..density..),
    bins = sqrt(length(data_cleaned_subset))+2, fill="grey", col="black") +
  facet_grid(Region ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y=theoretical.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of Median Income versus Regions")

data_cleaned_subset$intercept <- apply(data_cleaned_subset, 1, function(x){xbar[x["Region"]])}
data_cleaned_subset$slope <- apply(data_cleaned_subset, 1, function(x){s[x["Region"]])}
ggplot(data_cleaned_subset, aes(sample=MedianIncome)) +
  stat_qq() +
  facet_grid(Region ~ .) +
  geom_abline(data=data_cleaned_subset, aes(intercept = intercept, slope = slope)) +
  ggtitle("QQ Plots of Median Income by Region")

fit <- aov(MedianIncome ~ Region, data=data_cleaned_subset)
summary(fit)
test.Tukey<-TukeyHSD(fit,conf.level=0.95)
test.Tukey
```

b. **Inference 2: 2-sample test Urban Indicator vs Median Income**

Code in SAS:

```
data USDataSubsetnew;  infile "W:\fall 2017\stats350\us-data-cleaned.txt"
delimiter = '09'x firstobs = 2 ;
length IncomeCategory $11.;
input IncomeCategory      State $ Region $ CountyIndex $
UrbanIndicator           $ Population LandArea      PopulationDensity
PercentMaleDivorce PercentFemaleDivorce MedianIncome
PercentCollegeGraduates MedianHouseAge RobberiesPerPopulation
AssaultsPerPopulation BurglariesPerPopulation LarceniesPerPopulation
EducationSpending EducationSpendingP2 TestScore;

Run;

data USDataSubsetFinal;
set USDataSubsetnew;
if UrbanIndicator = '1' or UrbanIndicator = '0';
Run;

proc ttest data = USDataSubsetFinal H0 = 0 sides = 2 alpha = 0.05;
class UrbanIndicator;
var MedianIncome;
Run;

data logged;
set USDataSubsetFinal;
logincome = log(MedianIncome);

proc ttest data = logged H0 = 0 sides = 2 alpha = 0.05;
class UrbanIndicator;
var logincome;
```

run;

c. **Inference 3:** Regression of Median Income vs Percent College Graduate

Code in R

```
library(ggplot2)
dSub <- subset(USDataCleaned, 20420 < USDataCleaned$MedianIncome &
USDataCleaned$MedianIncome < 41320)

attach(dSub)

#Histogram of Median Income
windows()
ggplot(dSub, aes(MedianIncome)) +
  geom_histogram(aes(y=..density..),
                bins = sqrt(nrow(dSub)), fill="grey", col="black") +
  geom_density(col="red", lwd=1) +
  stat_function(fun=dnorm, args=list(mean=mean(MedianIncome),
                                     sd=sd(MedianIncome)), col="blue", lwd=1) +
  ggtitle("Histogram of Median Income")

#Tranform the data
dSub$lMedianIncome <- log(dSub$MedianIncome)

#Transformed Median Income
windows()
ggplot(dSub, aes(lMedianIncome)) +
  geom_histogram(aes(y=..density..),
                bins = sqrt(nrow(dSub)), fill="grey", col="black") +
  geom_density(col="red", lwd=1) +
  stat_function(fun=dnorm, args=list(mean=mean(lMedianIncome),
                                     sd=sd(lMedianIncome)), col="blue", lwd=1) +
  ggtitle("Histogram of Log Median Income")

#Percent cd
windows()
ggplot(dSub, aes(PercentCollegeGraduates)) +
  geom_histogram(aes(y=..density..),
                bins = sqrt(nrow(dSub)), fill="grey", col="black") +
  geom_density(col="red", lwd=1) +
  stat_function(fun=dnorm, args=list(mean=mean(PercentCollegeGraduates),
                                     sd=sd(PercentCollegeGraduates)),
                col="blue", lwd=1) +
  ggtitle("Histogram of Percent College Graduates")

#Tranform data
dSub$lCollegeGrad <- log(dSub$PercentCollegeGraduates)

#Percent cd tranformed
windows()
ggplot(dSub, aes(lCollegeGrad)) +
  geom_histogram(aes(y=..density..),
                bins = sqrt(nrow(dSub)), fill="grey", col="black") +
```

```

geom_density(col="red",lwd=1) +
stat_function(fun=dnorm,args=list(mean=mean(lCollegeGrad),
                                sd=sd(lCollegeGrad)), col="blue",lwd=1) +
ggtitle("Histogram of Log Percent College Graduates")

#Scatterplot of log median income vs log percent grad
windows()
ggplot(dSub, aes(x=lMedianIncome, y=lCollegeGrad))+
  geom_point(shape = 1) +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Relationship between Log Median Income and Log Percent College Graduate")
+
  xlab("Log Median Income") +
  ylab("Log Percent College Graduate")

rs<-cor(dSub$lMedianIncome, dSub$lCollegeGrad)

dSub.lm <- lm(lCollegeGrad ~ lMedianIncome, data = dSub)
summary(dSub.lm)
confint(dSub.lm, level = 0.95)

#Res plot
windows()
ggplot(data.frame(residuals=dSub.lm$res, lMedianIncome=dSub$lMedianIncome),
aes(x=lMedianIncome, y=residuals))+
  geom_point(shape = 1) +
  geom_smooth(method = lm, se = FALSE) +
  ggtitle("Residual Plot") +
  xlab("Log of Median Income") +
  ylab("Residuals")

#Res dist
windows()
ggplot(dSub.lm, aes(dSub.lm$res)) +
  geom_histogram(aes(y=..density..),
                bins = sqrt(nrow(dSub)), fill="grey",col="black") +
  geom_density(col="red",lwd=1) +
  stat_function(fun=dnorm,args=list(mean=mean(dSub.lm$res),
                                sd=sd(dSub.lm$res)), col="blue",lwd=1) +
  ggtitle("Histogram of Residuals")

#Res QQ
xbar <- mean(dSub.lm$res)
s <- sd(dSub.lm$res)
windows()
ggplot(dSub.lm, aes(sample=dSub.lm$res)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle("QQ Plot of Residuals")

```