STAT 35000 Project
Chakraborty
November 28, 2018

Group Project

Rohan Saxena - MWF 1:30PM
James Kuntz - MWF 1:30PM
Risheek Narayanadevarakere - MWF 1:30PM
Ayush Mehta - MWF 2:30PM

PART A: Introduction and Questions:

Questions:

1. Is the sale price different for houses in poor condition when compared to houses in good condition?
2. Is the sale price different for houses with poor construction grade when compared to houses with good construction grade?
3. Does the sale price vary based upon the era in which the house was built (pre-WWI, WWI-WWII, WWII-present)?

Why these questions are important:

These questions are important because they will give realtors, homeowners, and investors insight into how various factors (condition, construction grade, construction era) influence the value of homes in the area.

One might assume that houses in poor condition are on average less valuable than houses in good condition, but realtors and investors will not want to rely on such assumptions. We will perform a 2-sample independent t-test so that we can know with 95% confidence whether this is the case in Seattle. We will perform the same type of test to compare houses with lower than average building grade to houses with higher than average building grade, so that we can know whether construction grade also affects home value.

To answer our third question, we will perform an ANOVA test on three samples, which are houses built in the pre-WWI era, houses built between WWI and WWII, and houses built in the post-WWII era. The cutoffs for these eras were chosen because of the large societal changes that took place during both World Wars. World War I was significant in that it devastated the world economy (Forbes, 2014), so it's reasonable to think that this may have had an impact on the housing market in the United States. World War II brought many immigrants to the Pacific Northwest, including Seattle, to work in factories and shipyards to produce materials for the war effort ("Modern Society in the Pacific Northwest", n.d.), which likely brought about the need for more housing, thus impacting the housing market as well.

The statistical analysis that we complete in this report is trying to see the effect that different scenarios and conditions have on the sale price of houses in Seattle, Washington. In all three questions, the common dependent variable is the sale price of the houses, log-transformed for normality. In all three questions we are seeing how the sale price changes depending on different independent variables. In the first question, we want to see the relation between the condition of the house and how it affects the sale price. In the second question, we want to see the relation between the construction grade of the house and how it affects the sale price. In the third question we divide our dataset into three subsets based on the time period in which it was built, and we see whether the sale price differs between the pre-WWI era, the era between WWI and WWII, and the post-WWII era.

PART B: Data

a) Variables used:

Variable	Type	Description
priceLog	numeric	This is the log of the price that the house was sold for
grade	categorical	This indicates the construction quality of improvements on the house, on a scale from 1 - 13 (lowest quality to highest quality)
condition	categorical	This indicates the overall condition of the house, on a scale from 1 - 5 (poorest condition to best condition)
yr_built	categorical	This indicates the year in which the house was built

b) Code to clean data set:

```
# Clean data
houseCleaned <- house_dataF18[complete.cases(house_dataF18),]
```

PART C: Inference 1

Is the sale price different for houses in poor condition when compared to houses in good condition?

If I compare two conditions, will there be a statistically significant difference in the sales price for houses?

a) Code:

Provided in the Appendix

b) Analyze Data

We have decided to compare the change in price due to the condition of the houses. As per the data, the building condition varies from 1 to 5 and we have decided anything which comes between 1 and 3 inclusive will be poor condition and anything which comes between 4 and 5 inclusive will be good condition and have decided to see the change in the sale price when a house's condition is poor or good.

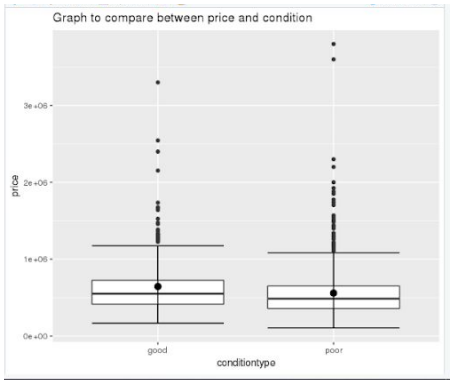
As per the data, we should use the two-sample independent t-test to analyze it because all the houses are different from each other and have no confounding variable. We only are concerned with determining the differences in the price of the houses based on their conditions. Also, we should use a two-sided alternative. We just want to know whether there is a difference between the price and not that one of them has a higher mean than the other.

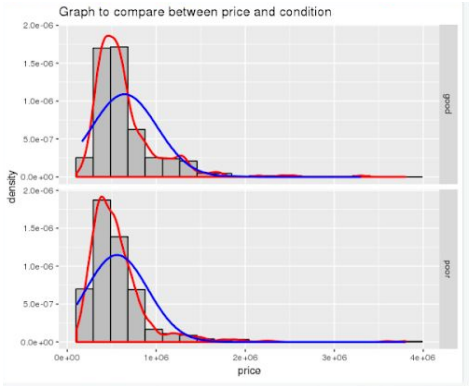
c) Graphical data and Assumptions

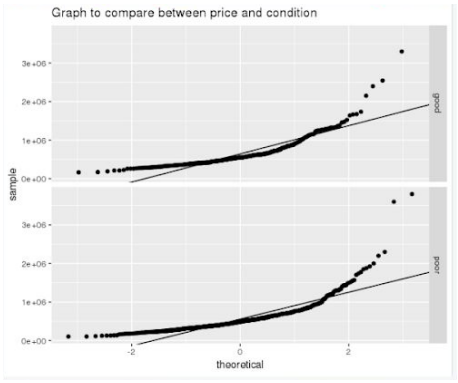
Assumptions:

1. SRS: We have to assume that we have an SRS from each population. We did not collect the data ourselves, so we can only assume that this is true.
2. Independence: We must assume that each population is independent of the others, which makes sense in our case, because each of the houses are distinct, so one house cannot be in two of our samples.
3. Normality: We have to assume that each independent population has a normal distribution, which will be checked from the graphs as shown below:

Graphs for original price data:

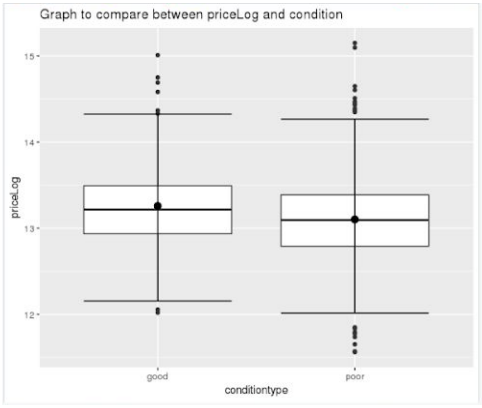


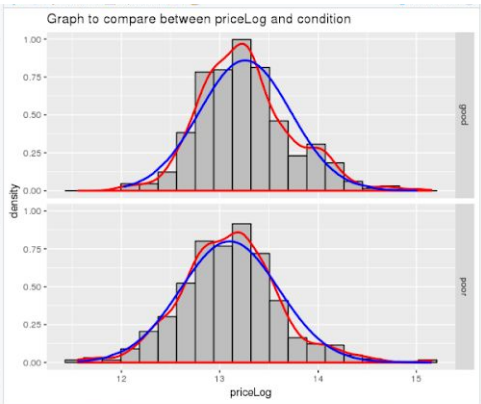


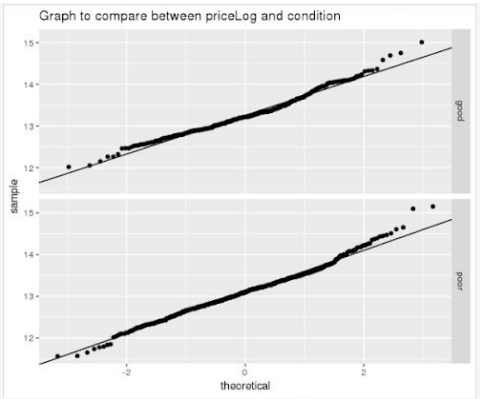


The mean is larger than the median in the box plots and long tails to the right which suggests a right-skewed data. A lot of outliers in the data possibly due to the large data set. The same goes for the Histogram's which suggests data be right skewed because of long tails to the right and therefore tells us that it's not approximated well by a normal distribution.
The data does not seem to appear normal, so we will use a log transformation.

d) Graphs for log-transformed data:







For the boxplots, the distribution appears to have tails of about equal length (right still being a little longer, but acceptable) and the means are close to the medians. However, the poor condition has more outliers than good because the subset for the poor condition is close to twice the size than the subset of the good condition.

The distributions appear to be close to symmetric for the Histogram. The two curves look closer to each other, indicating that the distributions are reasonably approximated by normal distributions.

The normal probability plots show that the data hugs the normal line closely, which means that the data is sufficiently normal.

Therefore, all the assumptions necessary to perform a 2-sample independent t-test have been met.

e) Confidence Interval and Hypothesis Test:

Assume a 0.05 significance level for the test.

```
        Welch Two Sample t-test

data:  dataFrame$priceLog by dataFrame$conditiontype
t = 4.9531, df = 747.27, p-value = 9.037e-07
alternative hypothesis: true difference in means is not equal to 0
5 percent confidence interval:
 0.1553540 0.1593393
sample estimates:
mean in group good mean in group poor
          13.25844           13.10109
```

Hypothesis test:

Step 1: Define parameters

Let μgood be the true mean of the log price for the good condition and μpoor be the true mean of the log price for the poor condition.

Step2: State Hypothesis

$H0$: μgood − μpoor = 0

HA: μgood − μpoor ≠ 0

Step 3: Report Test Statistic, Degrees of Freedom, and P-value

tts = 4.9531

Degrees of freedom: 747.27

p = 9.037e-07

Step 4: State Conclusion

Since 9.037e-07 < 0.05, we reject null hypothesis. The data show strong support to the claim that the population means difference in log price is different between the two building conditions of good and poor.

95% Confidence Interval

```
        Welch Two Sample t-test

data:  dataFrame$priceLog by dataFrame$conditiontype
t = 4.9531, df = 747.27, p-value = 9.037e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0949830 0.2197103
sample estimates:
mean in group good mean in group poor
          13.25844           13.10109
```

The interval is (0.0949830, 0.2197103). We are 95% confident that the difference in population mean of the priceLog between the Poor condition and the Good condition is covered by the interval from 0.0949830 to 0.2197103.

f) Conclusion:

Therefore, we can conclude that the results from the 2 sample t test with 0.05 significance level and the 95% confidence interval test are consistent as we reject the null hypothesis in the t test that the difference in mean is 0 and in the confidence interval test that the confidence interval did not contain a 0.

PART D: Inference 2

Is the sale price different for houses with poor construction grade when compared to houses with good construction grade?
If I compare two grades, will there be a statistically significant difference in the sales price for houses?

a) Code:

Provided in the Appendix

b) Analyze Data

We have decided to compare the change in price due to the construction grade of the houses. As per the data, the construction grade varies from 1 to 13 and we have decided anything which comes between 1 and 7 included will be poor grade and anything which comes between 8 and 13 included will be good grade and have decided to see the change in the sale price when a house's construction is poor or good.

Also we have made a new column in the houseCleaned which is called gradeType which assigns a goodGrade if above 7 and a poorGrade if between 1 and 7. We have decided to use that column to compare against price in the test.

As per the data, we should use the two-sample independent also to analyze it because all the houses are different from each other and have no confounding variable. We only are concerned with determining the differences in the price of the houses based on their construction grade. Also, we should use a two-sided alternative. We just want to know whether there is a difference between the price and not that one of them has a higher mean than the other.

c) Graphical data and Assumptions

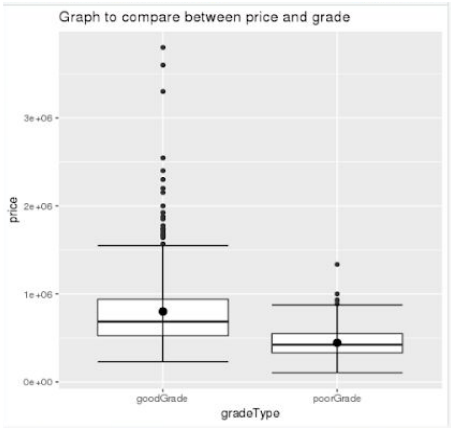
Assumptions:

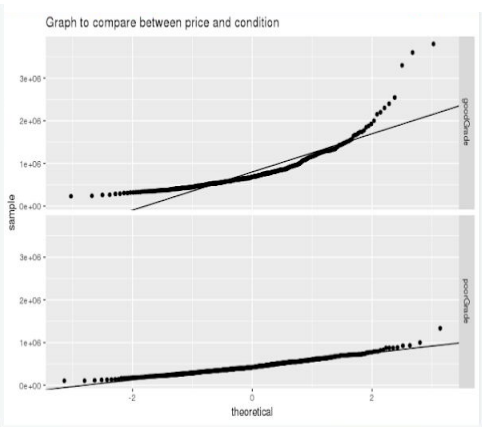
1. SRS: We have to assume that we have an SRS from each population. We did not collect the data ourselves, so we can only assume that this is true.

2. Independence: We must assume that each population is independent of the others, which makes sense in our case, because each of the houses are distinct, so one house cannot be in two of our samples.

3. Normality: We have to assume that each independent population has a normal distribution, which will be checked from the graphs as shown below:

Graphs for original data:



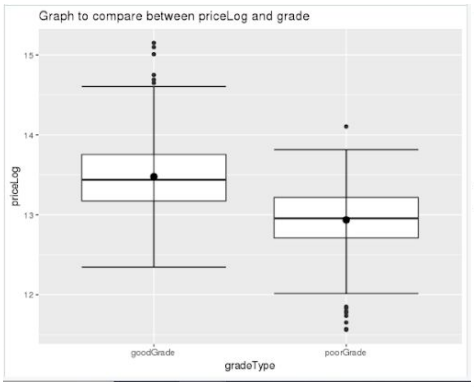


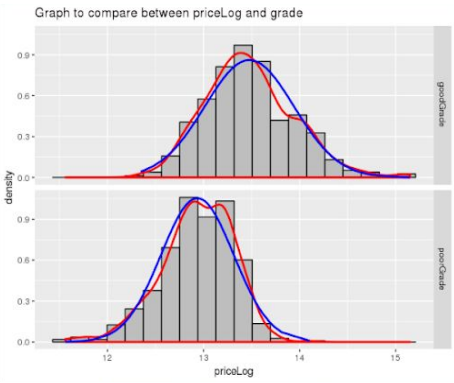
Means seems to be close to the median in the box plot, long tails to the right which suggests a right-skewed data. A lot of outliers in the data for goodGrade possibly due to the large data set. Same goes for the Histogram's which suggests data be right skewed because of long tails to the left and therefore tells us that it's not approximated well by a normal distribution. Also, the QQ-plot suggests that the goodGrade has the convex pattern (sometimes called "concave up") which indicates that the distribution is right skewed. However, the poorGrade lines seems to be followed by the data but the histogram says the opposite in terms of normal.

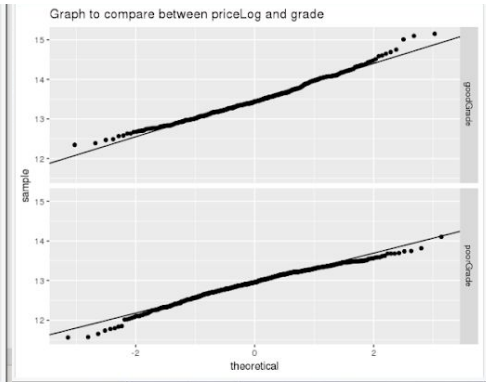
We will continue performing our search by doing a log transformation on this particular dataset because data is not normal.

The graphs obtained indicate that the data given is not normal, therefore we will perform a log transformation to normalize the data.

d) Log Transformed Graphs:







For Boxplot, the distribution appears to have tails of about equal length and the means are close to the medians. The good grade has less outliers now after now. However, the poor grade has more outliers from before however we can ignore them because of our large data set. This happened because even though the box plots distribution looked equal for the poor grade we still had to do log as the data was not normally distributed all together. The distributions appear to be close to symmetric for the Histogram. The two curves look closer to each other, indicating that the distributions are reasonably approximated by normal distributions. The poor grade histogram looks a little right skewed but very minimal which can be ignored. The QQ-plot data seems to follow the line which indicates normal distribution. However, there are minor deviations from both sides of the tails but not a concern for our data cause of the large size.

e) Confidence Interval and Hypothesis Test:

```
        Welch Two Sample t-test

data:  dataFrame$priceLog by dataFrame$gradeType
t = 19.452, df = 748.82, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
5 percent confidence interval:
 0.5395578 0.5430488
sample estimates:
mean in group goodGrade mean in group poorGrade
               13.47683                12.93553
```

Hypothesis test:

Step 1: Define

Let μgoodGrade be the true mean of the log price for the good grade and μpoorGrade be the true mean of the log price for the poor grade.

Step2: State Hypothesis

$H0$: μgoodGrade − μpoorGrade = 0

HA: μgoodGrade − μpoorGrade ≠ 0

Step 3: Report Test Statistic, Degrees of Freedom, and P-value

tts = 19.452

Degrees of freedom: 748.82

p = 2.2e-16

Step 4: State Conclusion

Since 2.22-16 < 0.05, we reject null hypothesis. The data show strong support to the claim that the population means difference in log price is different between the two constructions grade of good and poor.

95% Confidence Interval:

```
        Welch Two Sample t-test

data:  dataFrame$priceLog by dataFrame$gradeType
t = 19.452, df = 748.82, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4866750 0.5959316
sample estimates:
mean in group goodGrade mean in group poorGrade
               13.47683                12.93553
```

The interval is (0.4866750, 0.5959316). We are 95% confident that the difference in population mean of the priceLog between the Poor grade and the Good grade is covered by the interval from 0.4866750 to 0.5959316.

f) Conclusion:

Therefore, we can conclude that the results in 2 sample t test with 0.05 significance level and the 95% confidence interval are consistent as we reject the null hypothesis in 2 sample t test that the difference in mean is 0 and in confidence interval test that the confidence interval did not contain a 0 which means that both the questions are consistent with each other.

PART E: Inference 3

a) Code is in the Appendix

b) We should use the ANOVA (Analysis of Variance) test to analyze this data, because we have 3 independent samples from 3 time periods which we want to compare. We have divided the data into 3 subsets based on these three eras: pre-WWI construction, between WWI and WWII construction, and post-WWII construction. We want to see whether the sale price differs between these different eras.

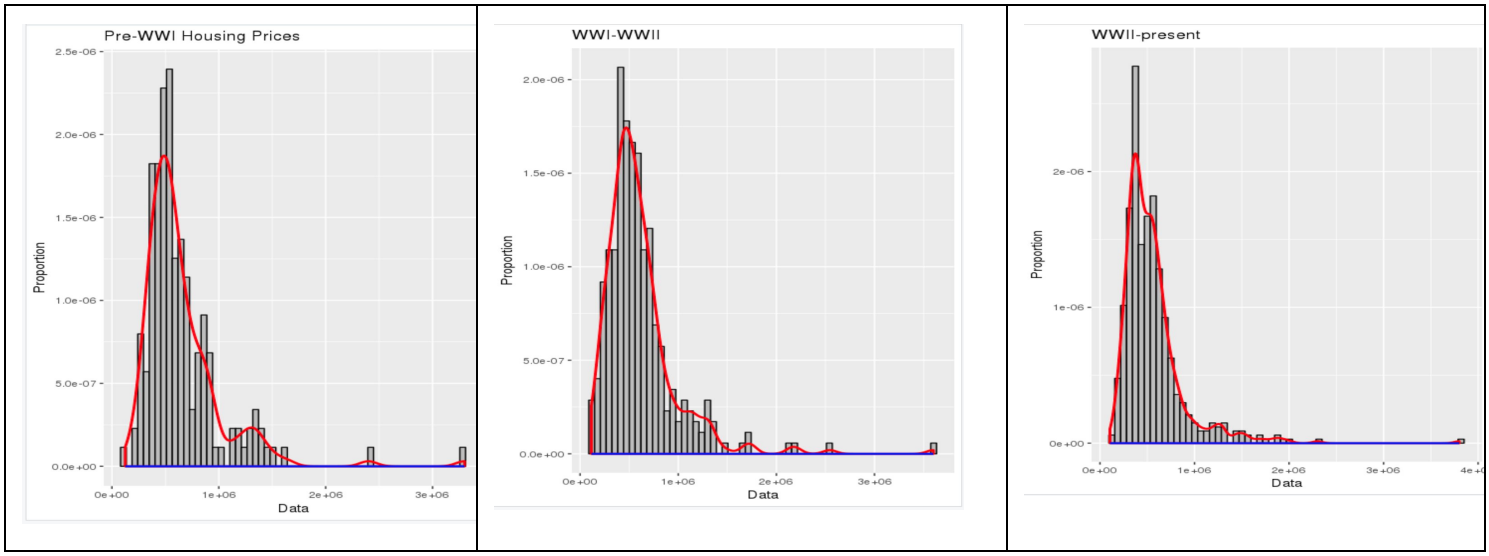
c) Assumptions and Graphs:

1. SRS: We have to assume that we have an SRS from each population. We did not collect the data ourselves, so we can only assume that this is true.

2. Independence: We must assume that each population is independent of the others, which makes sense in our case, because each of the houses are distinct, so one house cannot be in two of our samples.

3. Normality: We have to assume that each independent population has a normal distribution, which will be checked from the graphs as shown below:

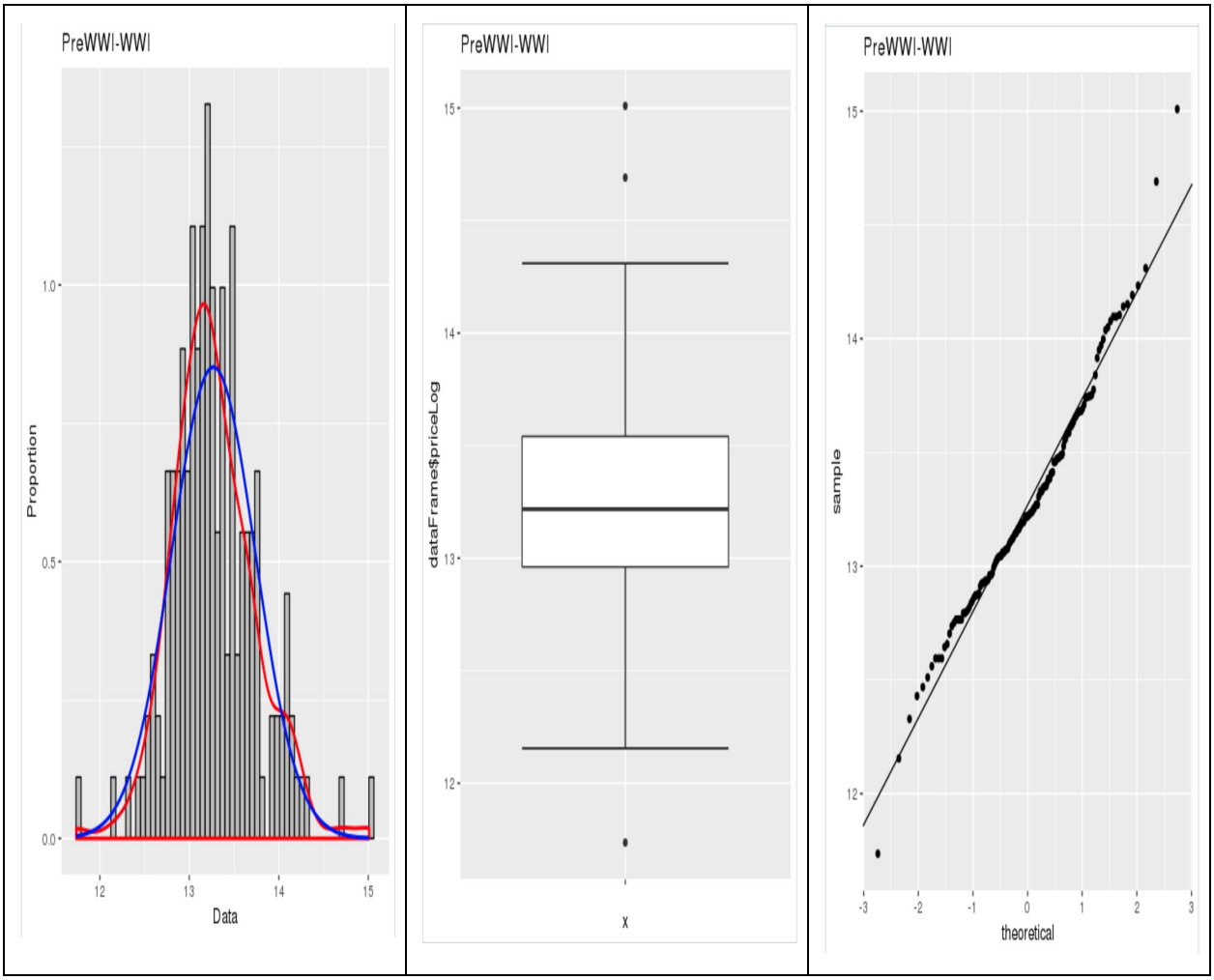
Histograms for original price data for houses from each time period:



It is clear that this data is not normal, as all of our samples have a right skew. So we applied a log transformation in order to satisfy the normality assumption constraint.

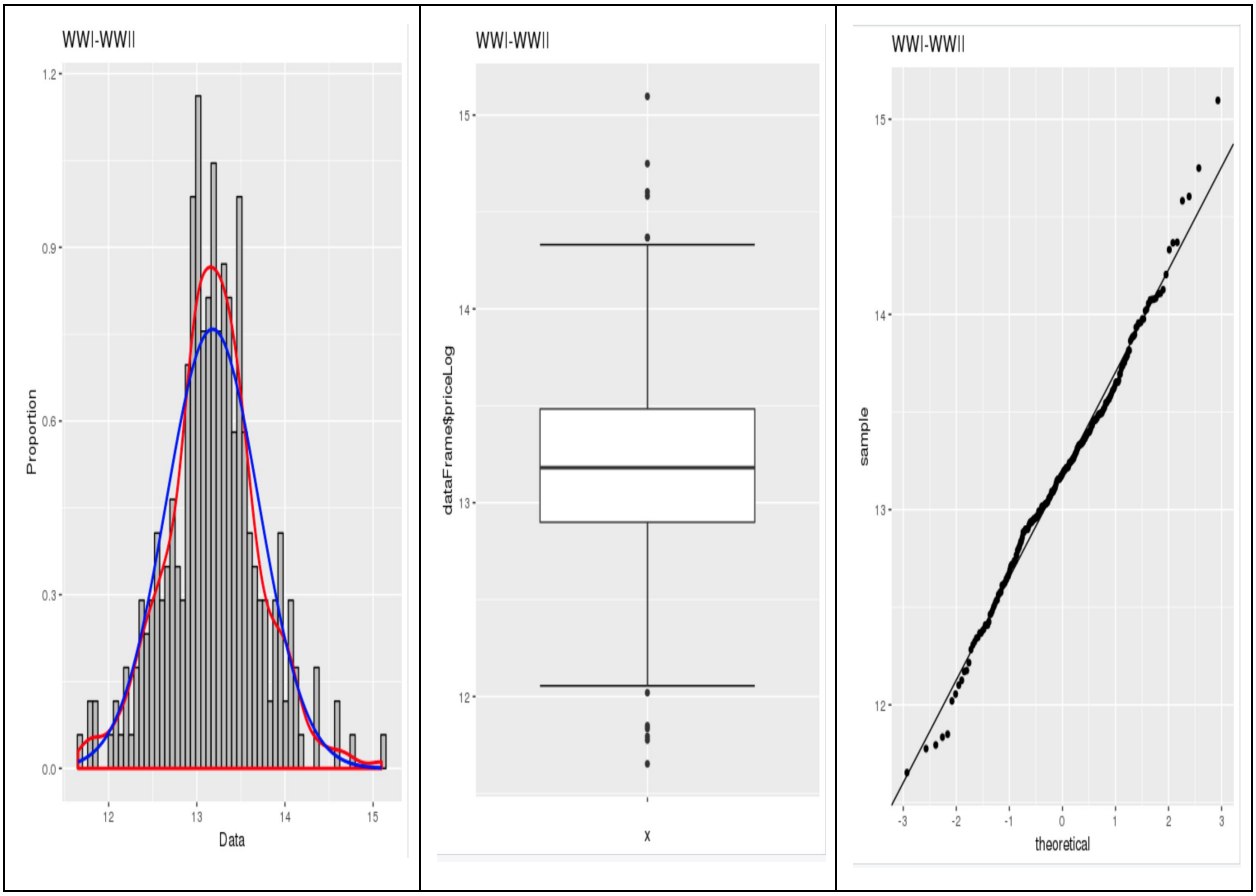
Diagnostic plots for log-transformed data:

Pre-WWI:



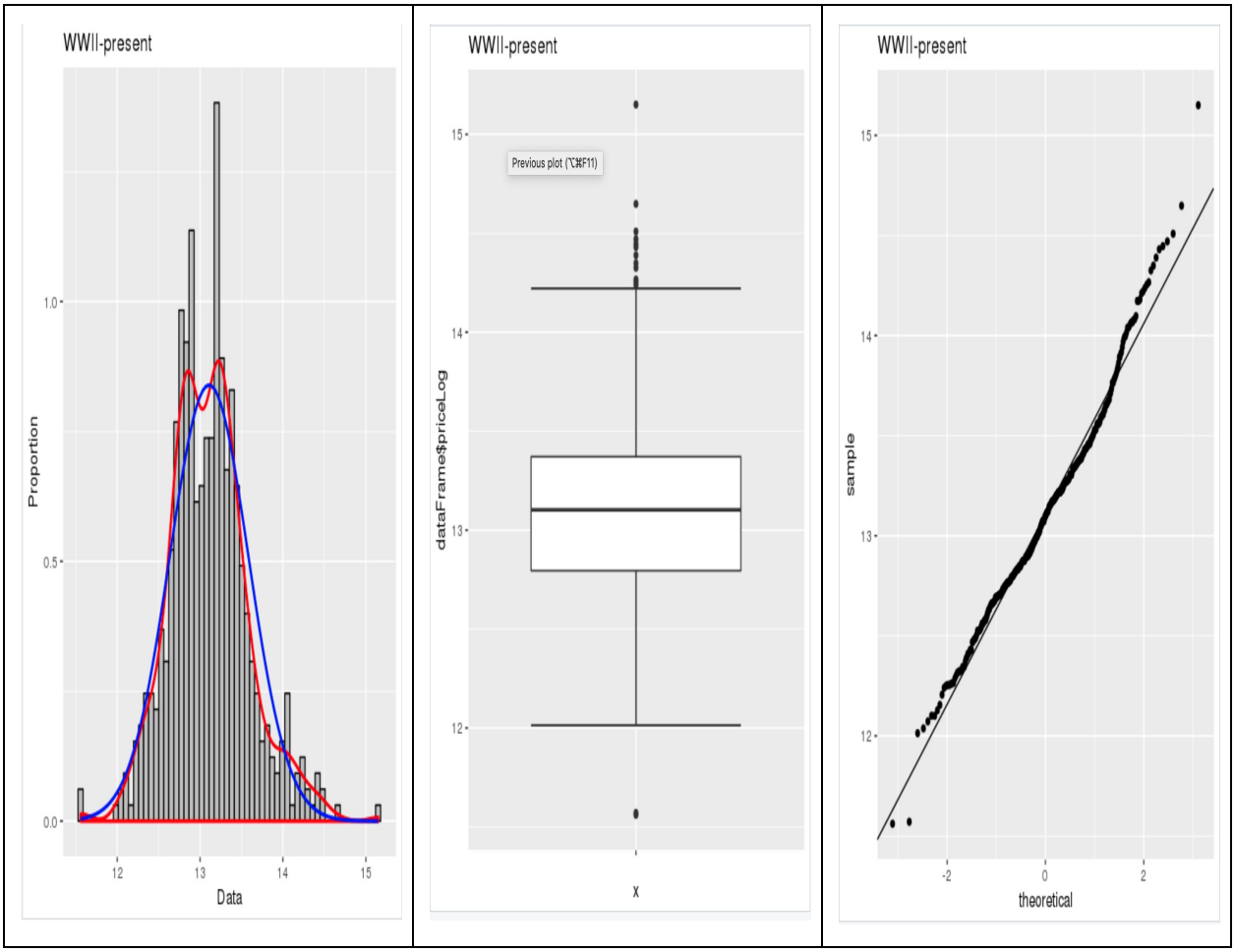
This data looks a lot more normally distributed than the original distribution of the sale price. From the histogram we can see that the distribution is unimodal and has no skewness to either side. The boxplot is evenly spread and there are a few outliers but majority of the data is normally distributed. Again for the qqplot, we can see that there are no major deviations from the normal line, therefore the data is normally distributed for the Pre-WWI subset.

WWI - WWII:



After applying the log transformation, we can see that the distribution is now normally distributed for the WWI-WWII subset. The histogram is unimodal with no skewness to either side. The boxplot looks evenly spread with some outliers but nothing major. Finally for the qqplot, there does not seem to be any serious strong deviations from the normal line.

Post-WWII:



After performing the log transformation, we can see that the Post-WWII subset is now normally distributed. The histogram looks fairly unimodal with no real skewness on either side. The histogram does seem to have some outlying points to the uppertail. The same can be said about the qqplot, however from a holistic perspective, the data seems to be normally distributed.

From these plots we can see that all of our samples are indeed normally distributed. What is more, our sample sizes are all greater than 100, which means that CLT can be used to approximate normality.

4. Constant Variance: We have to be able to assume that all of the populations have the same variance. We will do this using the constant standard deviation assumption below:

Era of Construction:	Mean: (priceLog)	Standard Deviation: (priceLog)
Years: 1900-1918	13.26896	0.468277
Years: 1919-1945	13.17859	0.525763
Years: 1946-2015	13.10868	0.4753999

Constant Standard Deviation Assumption: Smax / Smin < 2

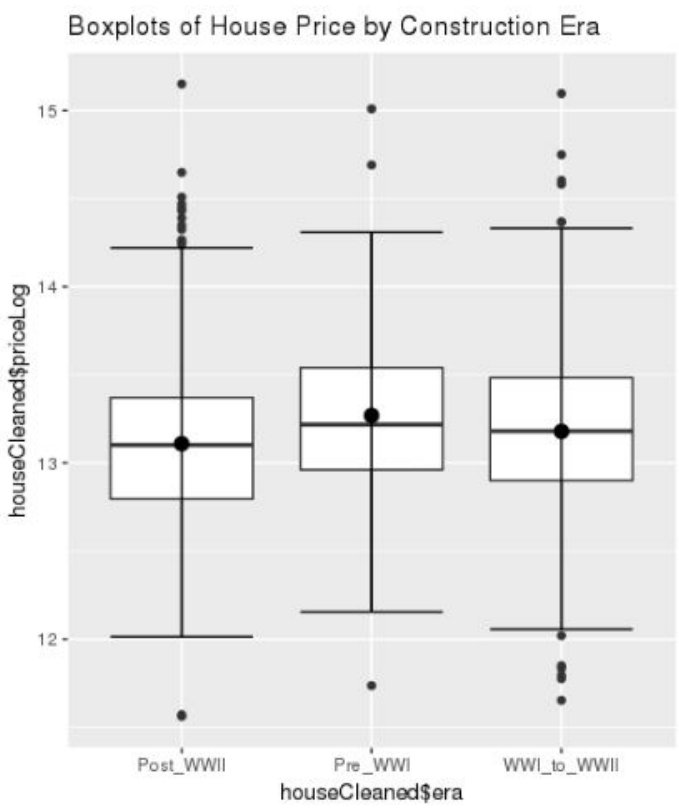
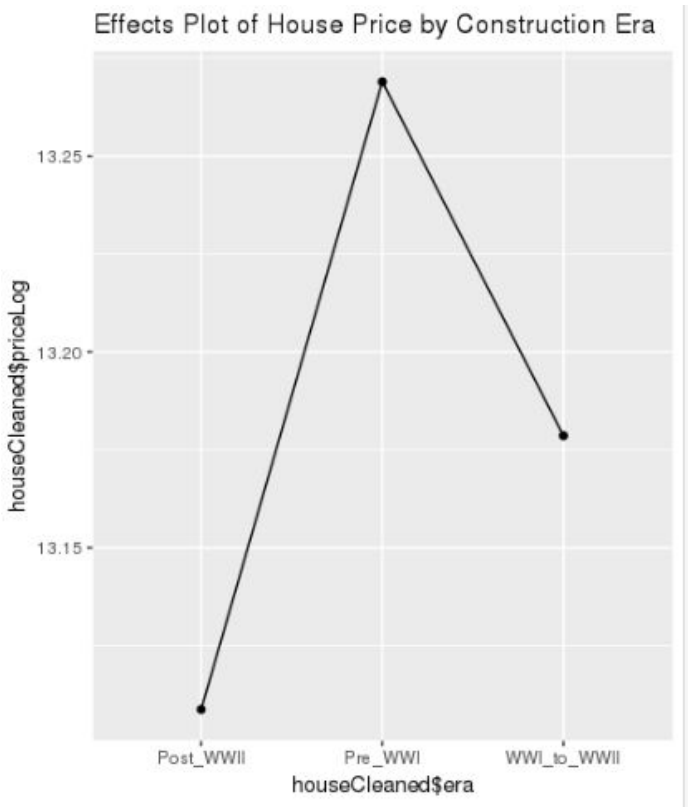
0.525763 / 0.468277 = 1.12276067 < 2

We can safely assume constant variance for our populations.

Therefore, all of our necessary assumptions for ANOVA have been met.

d) Graphs for Data:

Boxplots and Effects Plot of data samples:

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>The boxplot shows that the means for each sample may be equal, because there is a significant amount of overlap. However, we cannot say just by looking.</td>
<td>The effects plot shows us how the means likely compare, if we determine that the means are in fact different.</td>
</tr>
</table>

From these graphs, we can see that the means are likely fairly close. However, if we do determine that they are different, it is likely that older houses will have higher sale prices: we see in the effects plot that the pre-WWI sample has the highest mean, followed by the WWI-WWII sample, followed by the post-WWII sample.

e) Anova Hypothesis Test and Multiple Comparison Test

Program output for ANOVA test with a 95% confidence level:

```
            Df Sum Sq Mean Sq F value  Pr(>F)
era          2   3.43  1.7140   7.145 0.00083 ***
Residuals  990 237.48  0.2399
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypothesis Test:

Step 1: Define parameters

- mu_preWWI is the average sale price for houses constructed from 1900-1918
- mu_WWI-WWII is the average sale price for houses constructed from 1919-1945
- mu_postWWII is the average sale price for houses constructed from 1946-2015

Step 2: State the hypotheses

- Null hypothesis H0: mu_preWWI = mu_WWI-WWII = mu_postWWII
- Alternative hypothesis Ha: at least two mu’s are different

Step 3: Find the test statistic, p-value, and df (from program output, above)

- Test statistic: FTS = 7.145
- P-value: 0.00083
- DF1: 2
- DF2: 990

Step 4: State conclusion

- Because p = 0.00083 < alpha = 0.05, we reject H0 and conclude Ha.
- The data provides strong evidence (p = 0.00083) in support of the claim that at least one average house sale price does vary depending upon which era it was constructed in.

Program output for Tukey multiple comparison test:

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

Fit: aov(formula = priceLog ~ era, data = houseCleaned)

$era
                              diff          lwr         upr     p adj
Pre_WWI-Post_WWII       0.16027399  0.05742263 0.26312535 0.0007823
WWI_to_WWII-Post_WWII   0.06990399 -0.01346493 0.15327291 0.1206941
WWI_to_WWII-Pre_WWI    -0.09037000 -0.20256692 0.02182692 0.1419018
```

From this test, we can conclude the following:

- The average sale price of houses from the pre-WWI era is different from the average sale price of houses from the post-WWII era.

We cannot conclude that any of the other eras are different from one another, because the value 0 is not contained in the interval formed by the lower and upper bounds for these comparisons.

f) Conclusion:

Our question asked, “Does the sale price vary based upon the era in which the house was built (pre-WWI, WWI-WWII, WWII-present)?” Based on the results of our tests, we can conclude with 95% confidence that the average sale price for pre-WWI houses is different from the average sale price for post-WWII houses. We cannot conclude anything about whether the means for either of these eras are different from the mean of the WWI-WWII era.

Part F: Conclusion

Inference 1:

Our statistical analysis indicates that the sale prices differ for houses in different conditions. This conclusion carries value because a potential house buyer would want to balance the house condition and the price based on his or her needs.

Inference 2:

Our statistical analysis indicates that the sale prices differ for houses with different building grades. This conclusion carries value because a potential house buyer would want to balance the house building grade and the price based on his or her needs.

Inference 3:

Our statistical analysis indicates that the sale prices differ for houses built in the pre-WWI era, when compared to houses built in the post-WWII era. This conclusion is valuable because a potential house buyer would want to balance the age of the house with the price based on his or her needs.

We have been able to conclude that the average house sale price is affected by building condition, building grade, and construction era. These findings will be of value to all realtors and investors, as they will need to understand the various factors which influence the value of homes in the area in order to make wise purchases and sales. Homeowners looking to buy or sell will also benefit from understanding the factors which influence home values in the area.

References:

Forbes, S. (2014, August 06). The Horrid Economic Consequences of World War I -- We Still Suffer From Them. Retrieved from https://www.forbes.com/sites/steveforbes/2014/08/02/economic-consequences-of-the-great-war/#1ca68bf2b21b

Modern Society in the Pacific Northwest; The Second World War as Turning Point. (n.d.). Retrieved from http://www.washington.edu/uwired/outreach/cspn/Website/Classroom Materials/Pacific Northwest History/Lessons/Lesson 20/20.html

Appendix:

Starter code (used for all inferences):

```
# Project - STAT 35000

# Import data
library(readr)
house_dataF18 <- read_delim("/depot/statclass/data/stat35000/2018fall/house_dataF18.txt",
                            "\t", escape_double = FALSE, trim_ws = TRUE)

# Clean data
houseCleaned <- house_dataF18[complete.cases(house_dataF18),]

# Log transform data for price
houseCleaned$priceLog <- log(houseCleaned$price)

# Create subsets for each question:

# Q1: Is the sale price lower for houses in poor condition
# when compared to houses in good condition?
poorCond <- subset(houseCleaned, condition <= 3)
goodCond <- subset(houseCleaned, condition > 3)

# Q2: Is the sale price lower for houses with poor construction
# grade  when compared to houses with good construction grade?
poorGrade <- subset(houseCleaned, grade <= 7)
goodGrade <- subset(houseCleaned, grade > 7)

# Q3: Does the sale price vary based upon the era in which the house
# was constructed (pre-WWI, WWI-WWII, WWII-present)?
year19001918 <- subset(houseCleaned, yr_built <= 1918)
year19191945 <- subset(houseCleaned, ((yr_built > 1918) & (yr_built <= 1945)))
year19462015 <- subset(houseCleaned, yr_built > 1945)

```

Code for Inference 1:

```
#conditiontype
houseCleaned$conditiontype[houseCleaned$condition <= 3] <- "poor"
houseCleaned$conditiontype[houseCleaned$condition > 3] <- "good"


dataFrame <- houseCleaned
titleData <- "Graph to compare between price and condition"



#BOXPLOT

ggplot(data = dataFrame, aes(x = conditiontype, y = price)) +

  geom_boxplot() +

  stat_boxplot(geom = "errorbar") +

  stat_summary(fun.y = mean, color = "black", size = 3, geom ="point")+

  ggtitle(titleData)



#HISTOGRAM

tapply(dataFrame$price, dataFrame$conditiontype, length)
tapply(dataFrame$price, dataFrame$conditiontype, mean)
tapply(dataFrame$price, dataFrame$conditiontype, sd)
xbar <- tapply(dataFrame$price, dataFrame$conditiontype, mean)

s <- tapply(dataFrame$price, dataFrame$conditiontype, sd)
dataFrame$normal.density <- apply(dataFrame, 1, function(x){
  dnorm(as.numeric(x["price"]),
        xbar[x["conditiontype"]], s[x["conditiontype"]])})
ggplot(dataFrame, aes(x = price)) +
  geom_histogram(aes(y = ..density..),
                 bins = 20, fill = "grey", col = "black") +
  facet_grid(conditiontype ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle(titleData)


#QQPLOT

houseCleaned$intercept <- ifelse(houseCleaned$conditiontype == "good",
                                 xbar["good"], xbar["poor"])
houseCleaned$slope <- ifelse(houseCleaned$conditiontype == "good",
                             s["good"], s["poor"])

ggplot(houseCleaned, aes(sample = price)) +
```

```
  stat_qq() +
  facet_grid(conditiontype~ .) +
  geom_abline(data= houseCleaned, aes(intercept = intercept,
                                      slope = slope)) +
  ggtitle(titleData)


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


#LOG
houseCleaned$priceLog <- log(houseCleaned$price)



dataFrame <- houseCleaned
titleData <- "Graph to compare between priceLog and condition"



#BOXPLOT

ggplot(data = dataFrame, aes(x = conditiontype, y = priceLog)) +

  geom_boxplot() +

  stat_boxplot(geom = "errorbar") +

  stat_summary(fun.y = mean, color = "black", size = 3, geom ="point"

  ggtitle(titleData)



#HISTOGRAM

tapply(dataFrame$priceLog, dataFrame$conditiontype, length)
tapply(dataFrame$priceLog, dataFrame$conditiontype, mean)
tapply(dataFrame$priceLog, dataFrame$conditiontype, sd)
xbar <- tapply(dataFrame$priceLog, dataFrame$conditiontype, mean)

s <- tapply(dataFrame$priceLog, dataFrame$conditiontype, sd)
dataFrame$normal.density <- apply(dataFrame, 1, function(x){
  dnorm(as.numeric(x["priceLog"]),
        xbar[x["conditiontype"]], s[x["conditiontype"]])})
ggplot(dataFrame, aes(x = priceLog)) +
  geom_histogram(aes(y = ..density..),
                 bins = 20, fill = "grey", col = "black") +
  facet_grid(conditiontype ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle(titleData)
```

```r
#QQPLOT

houseCleaned$intercept <- ifelse(houseCleaned$conditiontype == "good",
                                 xbar["good"], xbar["poor"])
houseCleaned$slope <- ifelse(houseCleaned$conditiontype == "good",
                             s["good"], s["poor"])

ggplot(houseCleaned, aes(sample = priceLog)) +
  stat_qq() +
  facet_grid(conditiontype~ .) +
  geom_abline(data= houseCleaned, aes(intercept = intercept,
                                      slope = slope)) +
  ggtitle(titleData)



#2 SAMPLE T-TEST
titleData <- "Graph to compare between priceLog and condition"


t.test(dataFrame$priceLog ~ dataFrame$conditiontype, mu = 0,
       conf.level = 0.05, alternative = "two.sided",
       paired = FALSE, var.equal = FALSE)


dataFrame <- houseCleaned
titleData <- "Graph to compare between priceLog and condition"

t.test(dataFrame$priceLog ~ dataFrame$conditiontype, mu = 0,
       conf.level = 0.95, alternative = "two.sided",
       paired = FALSE, var.equal = FALSE)
```

Code for Inference 2:

```
#Inference 2

houseCleaned$gradeType[houseCleaned$grade <= 7] <- "poorGrade"
houseCleaned$gradeType[houseCleaned$grade > 7] <- "goodGrade"



dataFrame <- houseCleaned
titleData <- "Graph to compare between price and grade"

#BOXPLOT

ggplot(data = dataFrame, aes(x = gradeType, y = price)) +

  geom_boxplot() +

  stat_boxplot(geom = "errorbar") +

  stat_summary(fun.y = mean, color = "black", size = 3, geom ="point")+

  ggtitle(titleData)

dataFrame <- houseCleaned
titleData <- "Graph to compare between price and grade"

#HISTOGRAM

tapply(dataFrame$price, dataFrame$gradeType, length)
tapply(dataFrame$price, dataFrame$gradeType, mean)
tapply(dataFrame$price, dataFrame$gradeType, sd)
xbar <- tapply(dataFrame$price, dataFrame$gradeType, mean)

s <- tapply(dataFrame$price, dataFrame$gradeType, sd)
dataFrame$normal.density <- apply(dataFrame, 1, function(x){
  dnorm(as.numeric(x["price"]),
        xbar[x["gradeType"]], s[x["gradeType"]])})
ggplot(dataFrame, aes(x = price)) +
  geom_histogram(aes(y = ..density..),
                 bins = 20, fill = "grey", col = "black") +
  facet_grid(gradeType ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle(titleData)


#QQPLOT

houseCleaned$intercept <- ifelse(houseCleaned$gradeType == "goodGrade",
                                 xbar["goodGrade"], xbar["poorGrade"])
houseCleaned$slope <- ifelse(houseCleaned$gradeType == "goodGrade",
                             s["goodGrade"], s["poorGrade"])
ggplot(houseCleaned, aes(sample = price)) +
```

```
ggplot(houseCleaned, aes(sample = price)) +
  stat_qq() +
  facet_grid(gradeType~ .) +
  geom_abline(data= houseCleaned, aes(intercept = intercept,
                                      slope = slope)) +
  ggtitle(titleData)
```

```
#LOG
houseCleaned$priceLog <- log(houseCleaned$price)


dataFrame <- houseCleaned
titleData <- "Graph to compare between priceLog and grade"


#BOXPLOT

ggplot(data = dataFrame, aes(x = gradeType, y = priceLog)) +

  geom_boxplot() +

  stat_boxplot(geom = "errorbar") +

  stat_summary(fun.y = mean, color = "black", size = 3, geom ="point")+

  ggtitle(titleData)


#HISTOGRAM

tapply(dataFrame$priceLog, dataFrame$gradeType, length)
tapply(dataFrame$priceLog, dataFrame$gradeType, mean)
tapply(dataFrame$priceLog, dataFrame$gradeType, sd)
xbar <- tapply(dataFrame$priceLog, dataFrame$gradeType, mean)

s <- tapply(dataFrame$priceLog, dataFrame$gradeType, sd)
dataFrame$normal.density <- apply(dataFrame, 1, function(x){
  dnorm(as.numeric(x["priceLog"]),
        xbar[x["gradeType"]], s[x["gradeType"]])})
ggplot(dataFrame, aes(x = priceLog)) +
  geom_histogram(aes(y = ..density..),
```

```r
#HISTOGRAM

tapply(dataFrame$priceLog, dataFrame$gradeType, length)
tapply(dataFrame$priceLog, dataFrame$gradeType, mean)
tapply(dataFrame$priceLog, dataFrame$gradeType, sd)
xbar <- tapply(dataFrame$priceLog, dataFrame$gradeType, mean)

s <- tapply(dataFrame$priceLog, dataFrame$gradeType, sd)
dataFrame$normal.density <- apply(dataFrame, 1, function(x){
  dnorm(as.numeric(x["priceLog"]),
        xbar[x["gradeType"]], s[x["gradeType"]])})
ggplot(dataFrame, aes(x = priceLog)) +
  geom_histogram(aes(y = ..density..),
                 bins = 20, fill = "grey", col = "black") +
  facet_grid(gradeType ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle(titleData)



houseCleaned$intercept <- ifelse(houseCleaned$gradeType == "goodGrade",
                                 xbar["goodGrade"], xbar["poorGrade"])
houseCleaned$slope <- ifelse(houseCleaned$gradeType == "goodGrade",
                             s["goodGrade"], s["poorGrade"])

ggplot(houseCleaned, aes(sample = priceLog)) +
  stat_qq() +
  facet_grid(gradeType~ .) +
  geom_abline(data= houseCleaned, aes(intercept = intercept,
                                      slope = slope)) +
  ggtitle(titleData)



dataFrame <- houseCleaned
titleData <- "Graph to compare between priceLog and grade"


t.test(dataFrame$priceLog ~ dataFrame$gradeType, mu = 0,
       conf.level = 0.05, alternative = "two.sided",
       paired = FALSE, var.equal = FALSE)


dataFrame <- houseCleaned
titleData <- "Graph to compare between priceLog and grade"

t.test(dataFrame$priceLog ~ dataFrame$gradeType, mu = 0,
       conf.level = 0.95, alternative = "two.sided",
       paired = FALSE, var.equal = FALSE)
```

Code for Inference 3:

```
# Print plots for each data set

# Copy and paste what is below for however many graphs you need
# Just change the two variables in the EDIT section and leave the rest as is

# EDIT - define data set and title
dataFrame <- year19001918
titleData <- "Pre-WWI Housing Prices (1900 - 1918)"

# DO NOT EDIT - create plots

xBar <- mean(dataFrame$priceLog)
SD <- sd(dataFrame$priceLog)
xBar #print
SD   #print
library(ggplot2)

# Create histogram for original data
ggplot(data.frame(dataFrame = dataFrame), aes(x = dataFrame$price)) +
  geom_histogram(aes(y = ..density..), bins = 60,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm, args = list(mean = xBar, sd = SD), col = "blue", lwd = 1) +
  ggtitle(titleData) + xlab("Data") + ylab("Proportion")

# Create plots for log-transformed data
ggplot(data.frame(dataFrame = dataFrame), aes(x = dataFrame$priceLog)) +
  geom_histogram(aes(y = ..density..), bins = 60,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm, args = list(mean = xBar, sd = SD), col = "blue", lwd = 1) +
  ggtitle(titleData) + xlab("Data") + ylab("Proportion")
ggplot(dataFrame, aes(x = "", y = dataFrame$priceLog)) +
  stat_boxplot(geom = "errorbar") +
  geom_boxplot() +
  ggtitle(titleData) +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3)
```

```
ggplot(data.frame(dataFrame = dataFrame), aes(sample = dataFrame$priceLog)) +
  stat_qq() +
  geom_abline(slope = SD, intercept = xBar) +
  ggtitle(titleData)

# EDIT - define data set and title
dataFrame <- year19191945
titleData <- "WWI-WWII Housing Prices (1919 - 1945)"

# DO NOT EDIT - create plots
xBar <- mean(dataFrame$priceLog)
SD <- sd(dataFrame$priceLog)
xBar
SD
library(ggplot2)

# Create histogram for original data
ggplot(data.frame(dataFrame = dataFrame), aes(x = dataFrame$price)) +
  geom_histogram(aes(y = ..density..), bins = 60,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm, args = list(mean = xBar, sd = SD), col = "blue", lwd = 1) +
  ggtitle(titleData) + xlab("Data") + ylab("Proportion")

# Create plots for log-transformed data
ggplot(data.frame(dataFrame = dataFrame), aes(x = dataFrame$priceLog)) +
  geom_histogram(aes(y = ..density..), bins = 60,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm, args = list(mean = xBar, sd = SD), col = "blue", lwd = 1) +
  ggtitle(titleData) + xlab("Data") + ylab("Proportion")
ggplot(dataFrame, aes(x = "", y = dataFrame$priceLog)) +
  stat_boxplot(geom = "errorbar") +
  geom_boxplot() +
  ggtitle(titleData) +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3)
ggplot(data.frame(dataFrame = dataFrame), aes(sample = dataFrame$priceLog)) +
```

```
  stat_qq() +
  geom_abline(slope = SD, intercept = xBar) +
  ggtitle(titleData)

# EDIT - define data set and title
dataFrame <- year19462015
titleData <- "Post-WWII Housing Prices (1946 - 2015)"

# DO NOT EDIT - create plots
xBar <- mean(dataFrame$priceLog)
SD <- sd(dataFrame$priceLog)
xBar
SD
library(ggplot2)

# Create histogram for original data
ggplot(data.frame(dataFrame = dataFrame), aes(x = dataFrame$price)) +
  geom_histogram(aes(y = ..density..), bins = 60,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm, args = list(mean = xBar, sd = SD), col = "blue", lwd = 1) +
  ggtitle(titleData) + xlab("Data") + ylab("Proportion")

# Create plots for log-transformed data
ggplot(data.frame(dataFrame = dataFrame), aes(x = dataFrame$priceLog)) +
  geom_histogram(aes(y = ..density..), bins = 60,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm, args = list(mean = xBar, sd = SD), col = "blue", lwd = 1) +
  ggtitle(titleData) + xlab("Data") + ylab("Proportion")
ggplot(dataFrame, aes(x = "", y = dataFrame$priceLog)) +
  stat_boxplot(geom = "errorbar") +
  geom_boxplot() +
  ggtitle(titleData) +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3)
ggplot(data.frame(dataFrame = dataFrame), aes(sample = dataFrame$priceLog)) +
  stat_qq() +
  geom_abline(slope = SD, intercept = xBar) +
  ggtitle(titleData)


# Create new variable era for diagnostic plots
houseCleaned$era[houseCleaned$yr_built <= 1918] <- "Pre_WWI"
houseCleaned$era[(houseCleaned$yr_built > 1918) & (houseCleaned$yr_built <= 1945)] <- "WWI_to_WWII"
houseCleaned$era[houseCleaned$yr_built > 1945] <- "Post_WWII"

# Create side-by-side box plots for the samples
ggplot(houseCleaned, aes(x = houseCleaned$era, y = houseCleaned$priceLog)) +
  geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3) +
  ggtitle("Boxplots of House Price by Construction Era")

# Create effects plot for the samples
ggplot(data = houseCleaned, aes(x = houseCleaned$era, y = houseCleaned$priceLog)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line", aes(group = 1)) +
  ggtitle("Effects Plot of House Price by Construction Era")

# Perform ANOVA hypothesis test
fit <- aov(priceLog ~ era, data = houseCleaned)
summary(fit)

# Use Tukey method for multiple-comparison test
TukeyHSD(fit, conf.level = 0.95)
```