Part B. Household income in different regions

1. Code

```
#####
# Jordan Mayer
# STAT 350
# Lab 07
# March 29, 2018
#####

# setup
setwd("W:/Courses Spring 2018/STAT 350/STAT 350 Labs/Lab 07")
  # set working directory
library(ggplot2)  # set up ggplot2 for plotting
graphics.off()  # close any open figures
USData <- read.table("US_Data.txt", header=TRUE, sep="\t")  # get US Data
US_clean <- USData[complete.cases(USData),]  # clean US Data
US_NE <- subset(US_clean, Region == "NE") # subset for Northeast region only
US_NC <- subset(US_clean, Region == "NC") # subset for North Central region
only
US_ESDiff <- US_clean$EducationSpendingP2 - US_clean$EducationSpending
  # differences in education spending between period 2 and period 1

### PART B ###
# create plots for Northeast median income and North Central median income
for (reg in c("NE", "NC")) {
  # attach dataset
  if (reg == "NE") {
    attach(US_NE)
    title = "Median Household Income in Northeast US"
  }
  else {
    attach(US_NC)
    title = "Median Household Income in North Central US"
  }
  # boxplot
  box <-
ggplot(data.frame(MedianIncome=MedianIncome),aes(x="",y=MedianIncome))+
    stat_boxplot(geom="errorbar")+
    geom_boxplot()+
    ggtitle("Test Scores")+
    stat_summary(fun.y=mean,col="black",geom="point",size=3)+
    ggtitle(title)
  ggsave(filename=paste("box",reg,".png"),box,height=6,width=6)

  # histogram
  hist <- ggplot(data.frame(MedianIncome=MedianIncome),aes(MedianIncome))+
```

```
    geom_histogram(aes(y=..density..),
                   bins=sqrt(length(MedianIncome))+2,
                   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(title)+
    xlab("Data")+
    ylab("Proportion")
  ggsave(filename=paste("hist",reg,".png"),hist,height=6,width=6)
  # normal probability plot
  #windows()
  qq <- ggplot(data.frame(MedianIncome),aes(sample=MedianIncome))+
    stat_qq()+
    geom_abline(slope=sd(MedianIncome),intercept=mean(MedianIncome))+
    ggtitle(title)+
    xlab("Theoretical")+
    ylab("Sample")
  ggsave(filename=paste("qq",reg,".png"),qq,height=6,width=6)
  # detach dataset
  if (reg == "NE") {
    detach(US_NE)
  }
  else {
    detach(US_NC)
  }
}

# data are not normally distributed -> transform using log
US_NE_MIlog <- log(US_NE$MedianIncome)  # transformed NE median income
US_NC_MIlog <- log(US_NC$MedianIncome)  # transformed NC median income
# create plots for transformed data
### PART B ###
# create plots for Northeast median income and North Central median income
for (reg in c("NElog", "NClog")) {
  # attach dataset
  if (reg == "NElog") {
    var = US_NC_MIlog
    title = "Median Household Income in Northeast US (transformed)"
  }
  else {
    var = US_NC_MIlog
    title = "Median Household Income in North Central US (transformed)"
  }
  # boxplot
  box <- ggplot(data.frame(var),aes(x="",y=var),height=6,width=6)+
    stat_boxplot(geom="errorbar")+
    geom_boxplot()+
```

```
    ggtitle("Test Scores")+
    stat_summary(fun.y=mean,col="black",geom="point",size=3)+
    ggtitle(title)
  ggsave(filename=paste("box",reg,".png"),box,height=6,width=6)

  # histogram
  hist <- ggplot(data.frame(var),aes(var),height=6,width=6)+
    geom_histogram(aes(y=..density..),
                   bins=sqrt(length(var))+2,
                   fill="grey",col="black")+
    geom_density(col="red",lwd=1)+
    stat_function(fun=dnorm,args=list(mean=mean(var),
                                      sd=sd(var)),
                  col="blue",lwd=1)+
    ggtitle(title)+
    xlab("Data")+
    ylab("Proportion")
  ggsave(filename=paste("hist",reg,".png"),hist,height=6,width=6)
  # normal probability plot
  #windows()
  qq <- ggplot(data.frame(var),aes(sample=var),height=6,width=6)+
    stat_qq()+
    geom_abline(slope=sd(var),intercept=mean(var))+
    ggtitle(title)+
    xlab("Theoretical")+
    ylab("Sample")
  ggsave(filename=paste("qq",reg,".png"),qq,height=6,width=6)
}

# conduct two-sample independent hypothesis test, two-sided
t.test(US_NE_MIlog, US_NC_MIlog, mu=0, conf.level=0.95,
       alternative="two.sided", paired=FALSE)
```

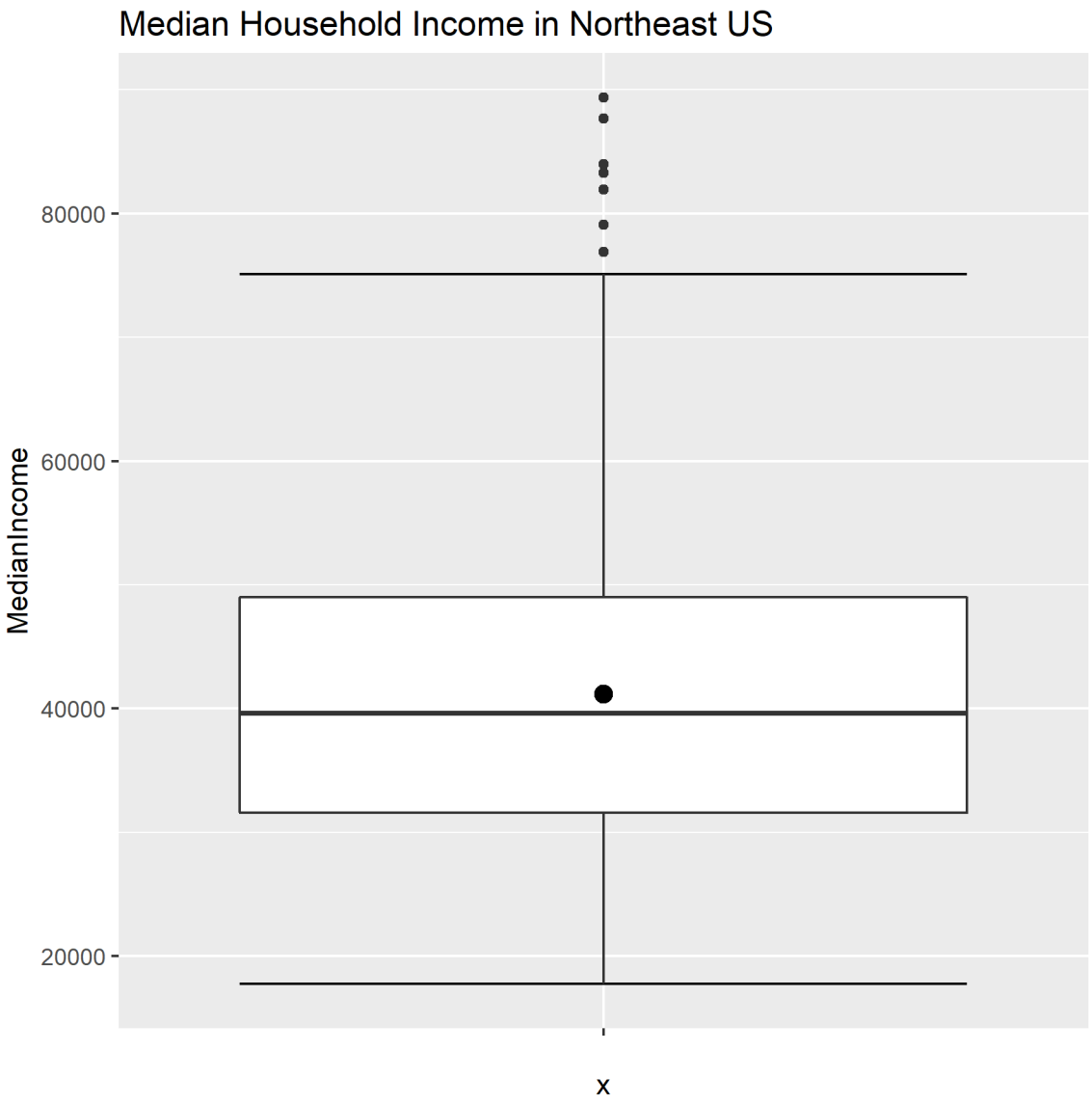
2. Independent or Paired

This data should be analyzed using a two-sample independent procedure. This is because we are analyzing two *independent* samples: the median income in the Northeast region and the median income in the North Central region. A two-sample paired procedure would be more fitting if we were analyzing two observations of the same individuals.

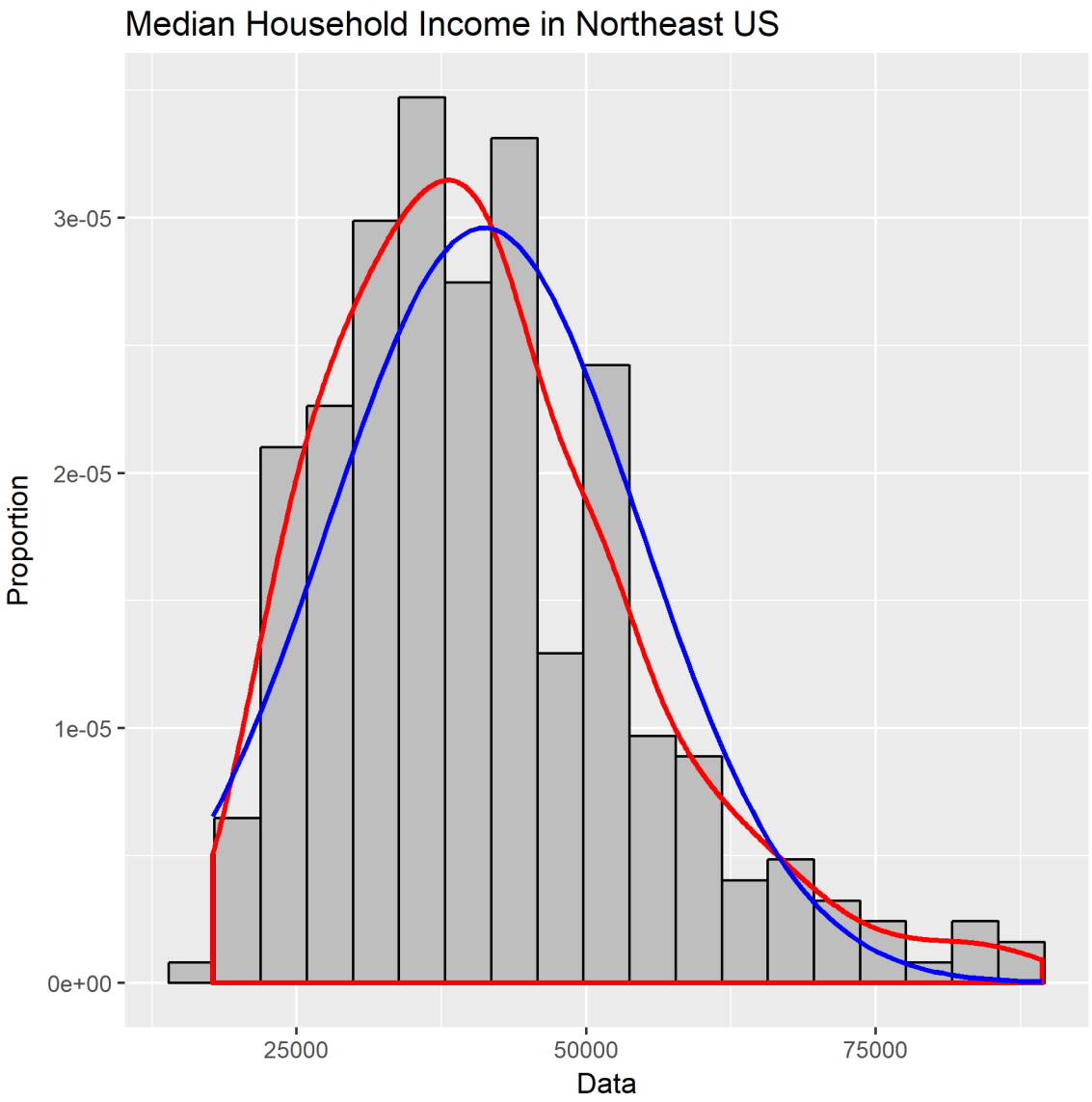
3. One-sided or Two-sided

This data should be analyzed using a two-sided alternative. This is because we are only trying to test whether the means of median household income in the two regions are *significantly different* (e.g. $\Delta_a \neq \Delta_0$), not whether the difference is below or above a specific value.

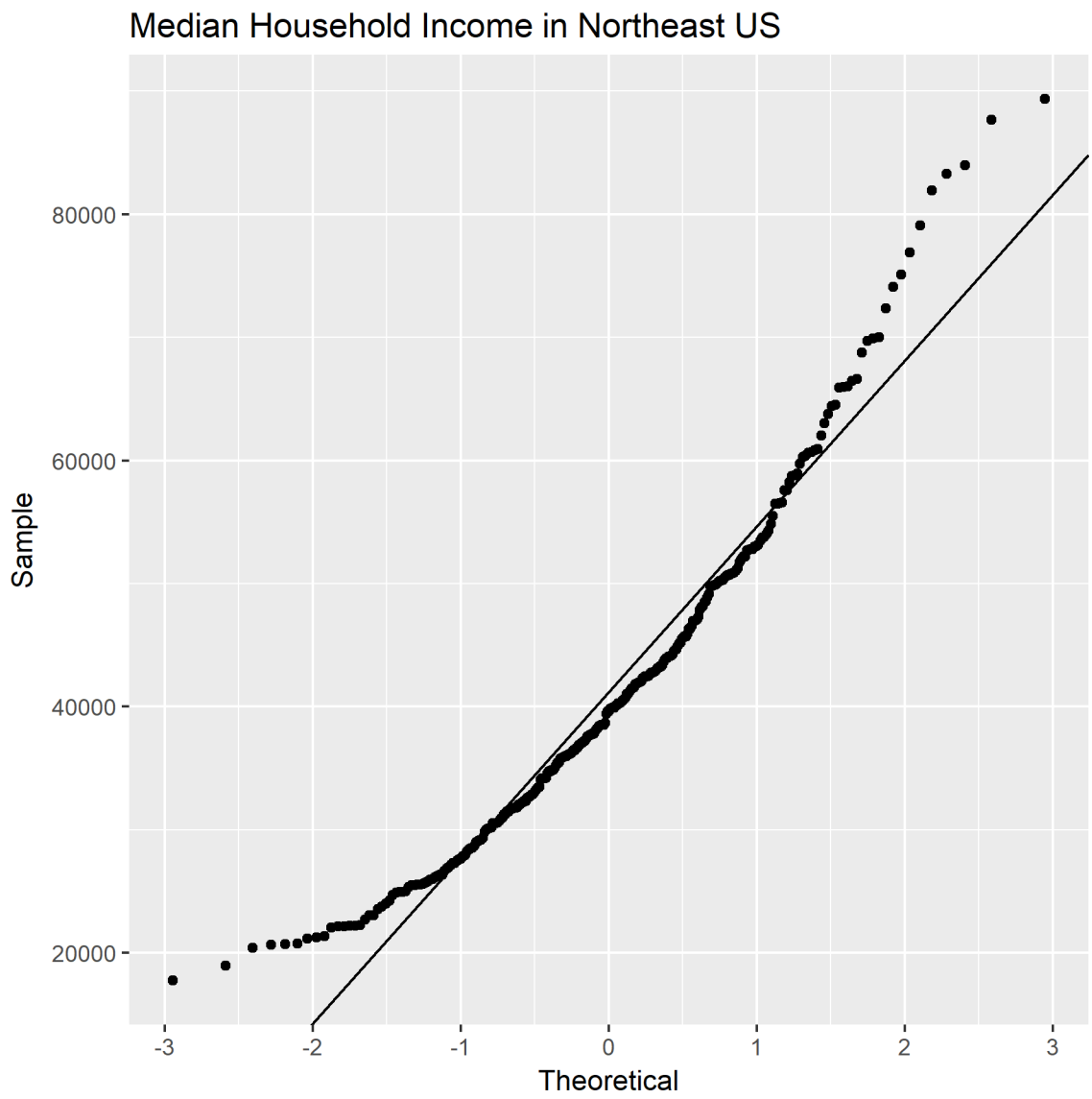
4. Plots



This boxplot demonstrates a significant positive skew.

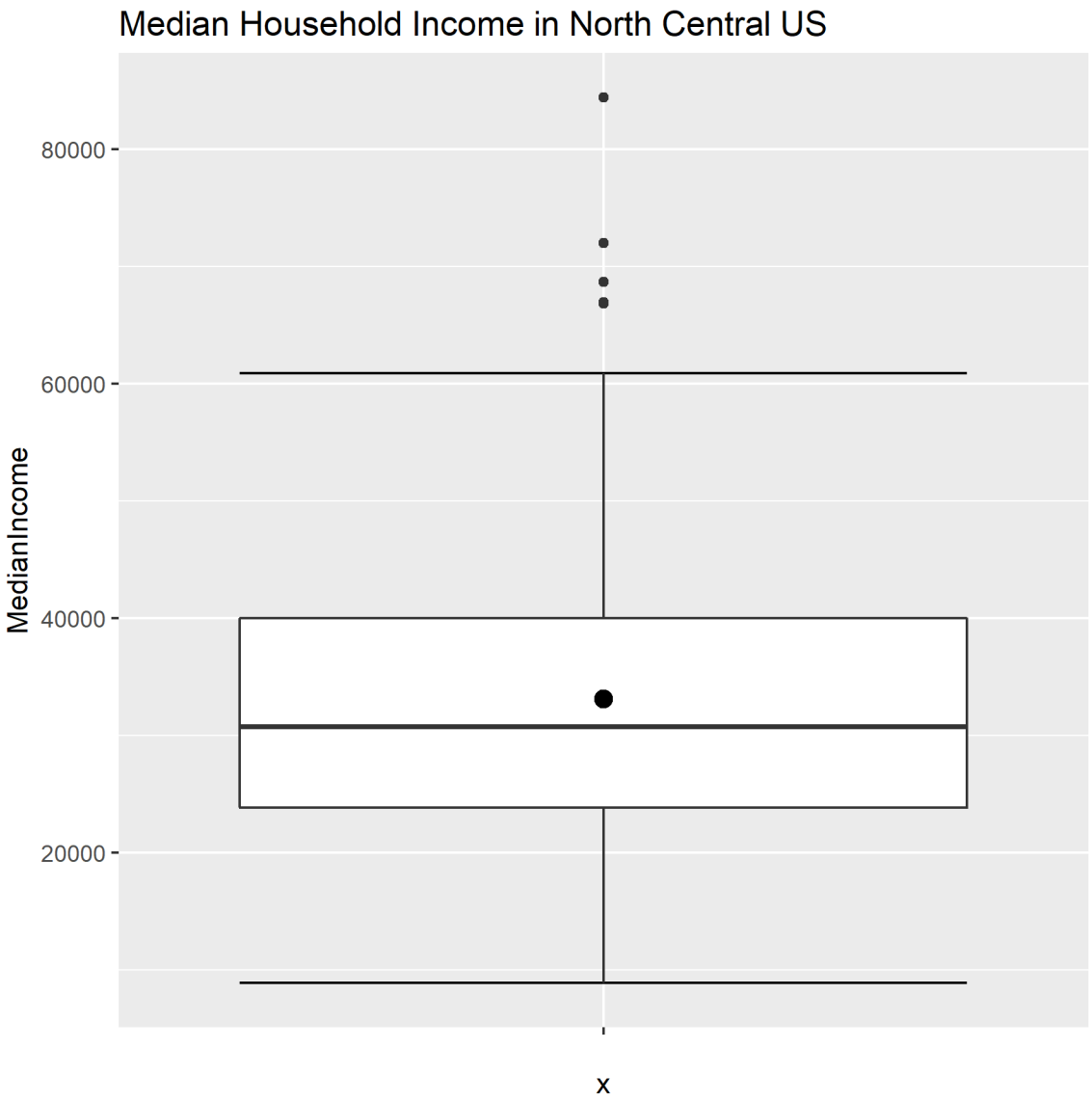


Similarly, this histogram demonstrates a positive skew.

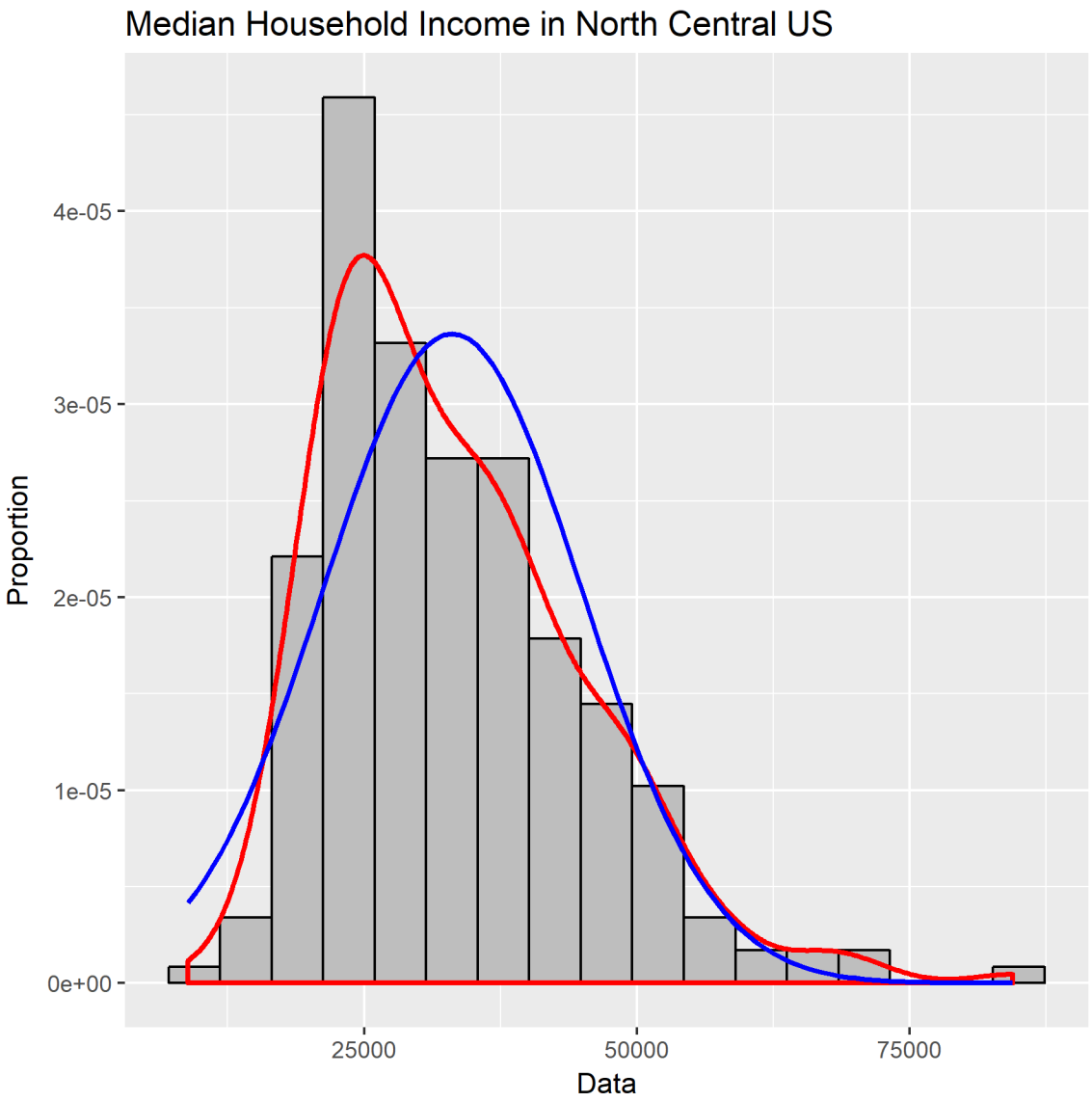


Indeed, the same positive skew is present in this normal probability plot.

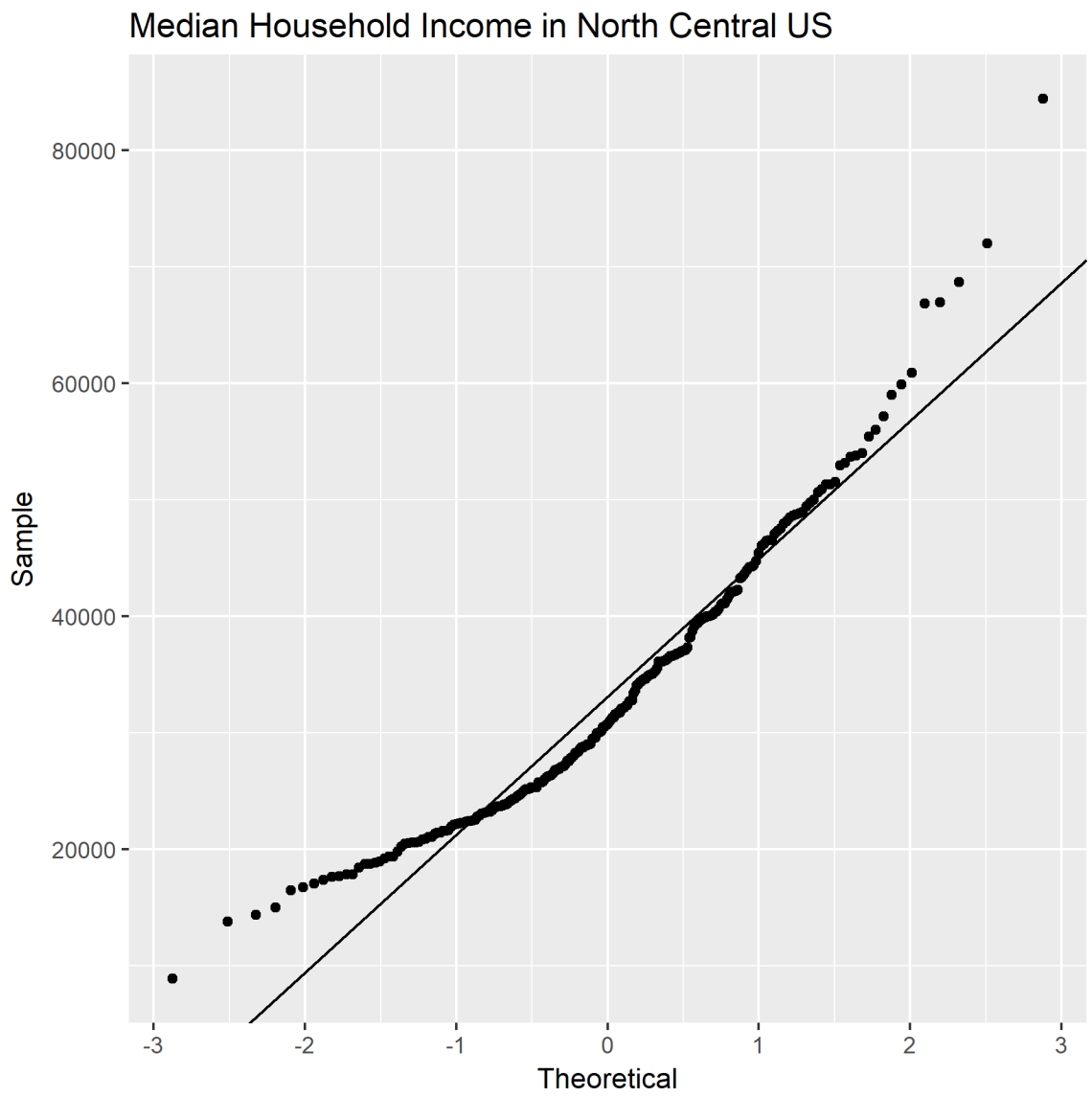
Clearly, the distribution of Median Household Income for the Northeast region is not normally distributed. This data will have to be transformed.



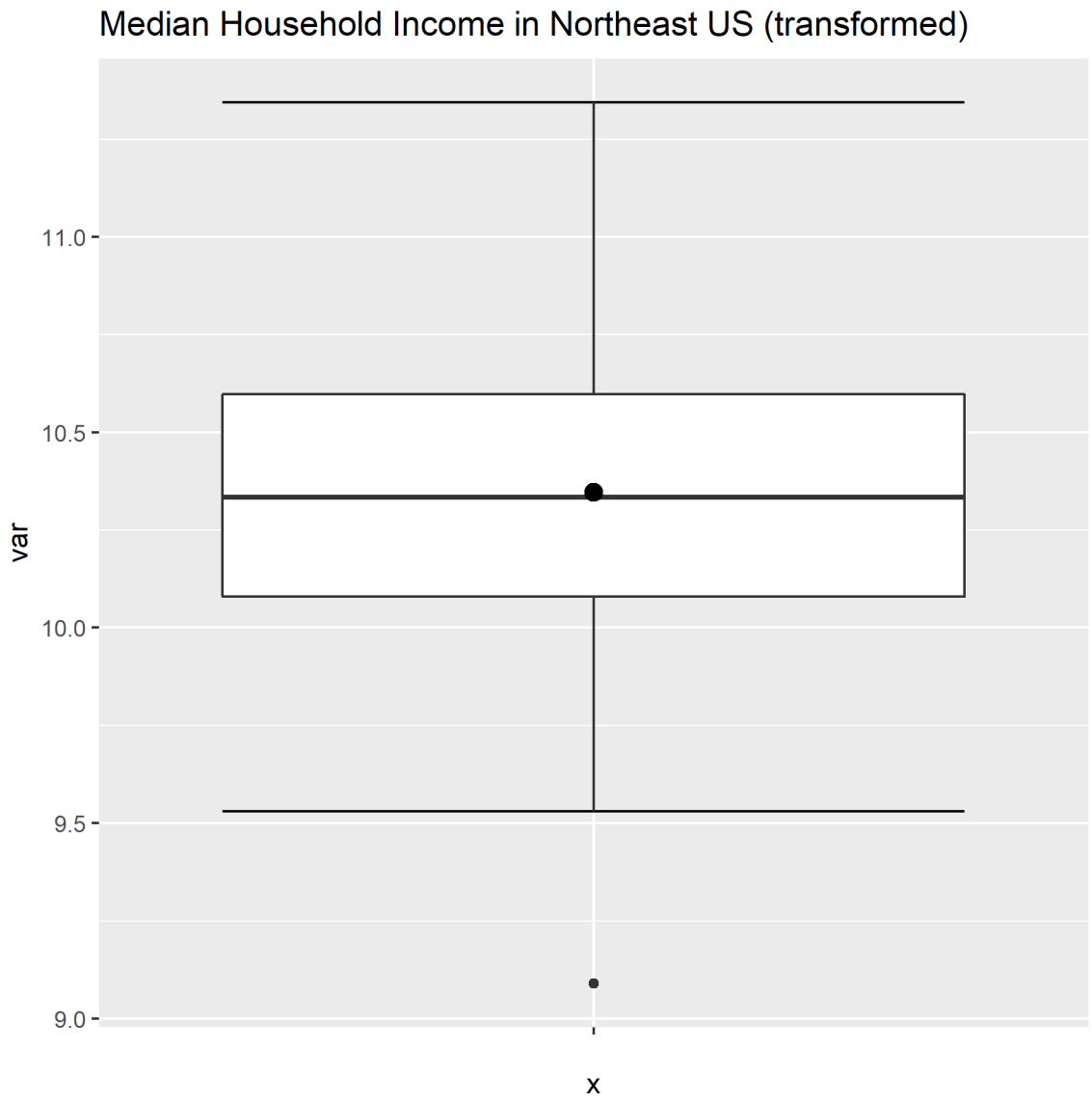
This boxplot, as well, shows a positive skew – though perhaps not as great.



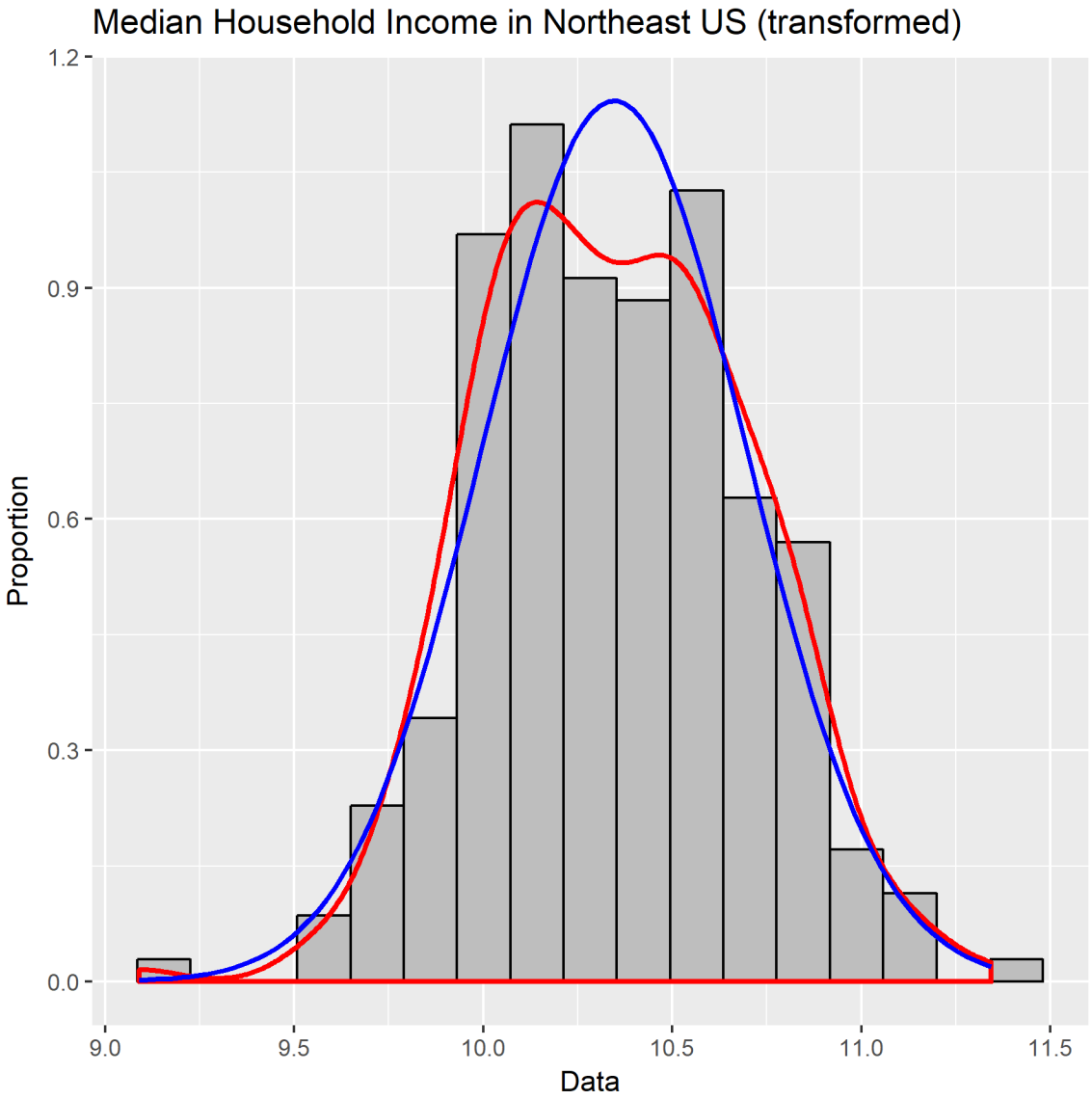
This histogram shows a positive skew for this data as well.



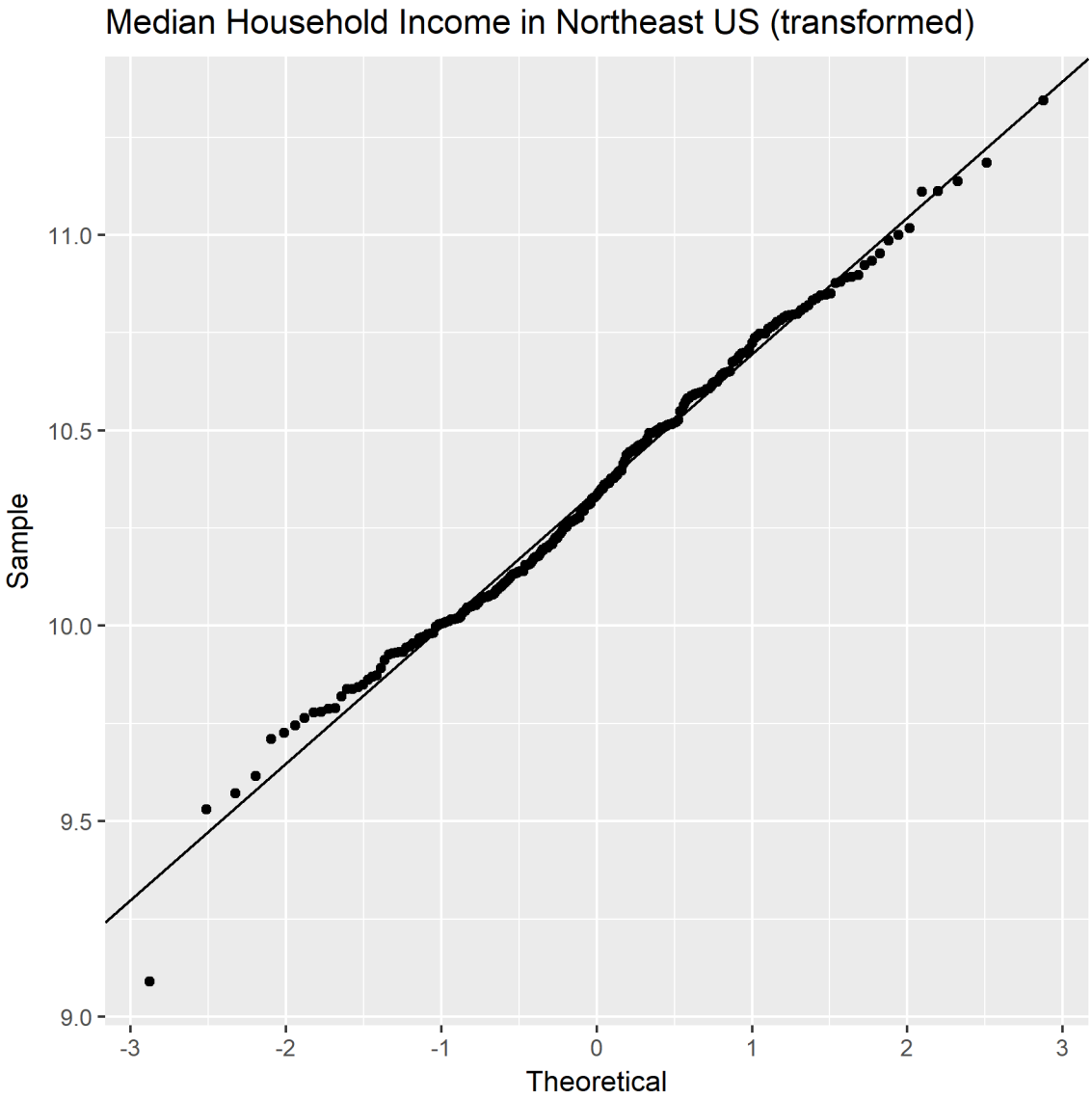
This normal probability plot confirms: the data for Median Household Income in the North Central region is not normally distributed. This data, too, must be transformed.



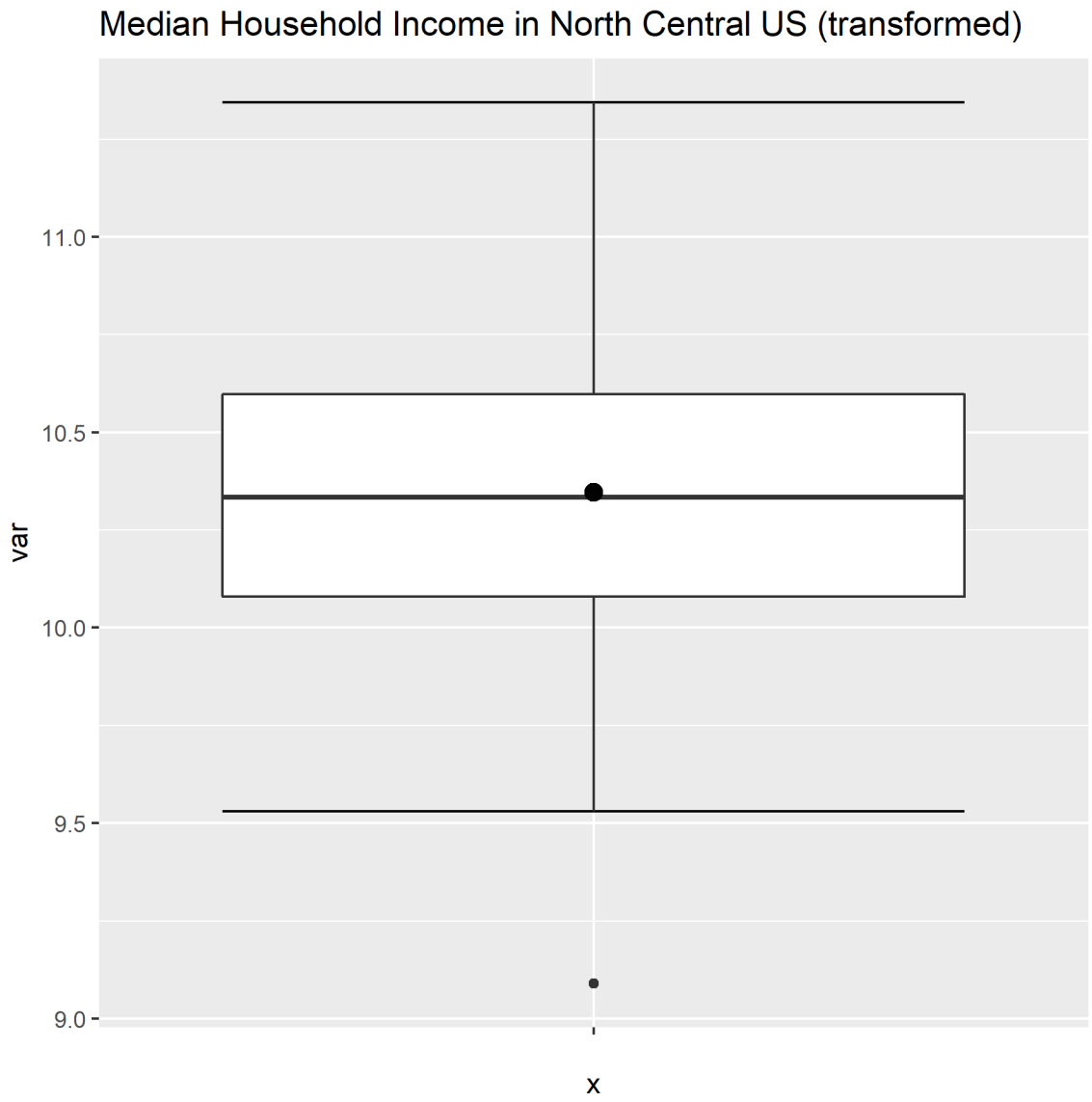
After transformation, this boxplot demonstrates a much more normal distribution, with one outlier.



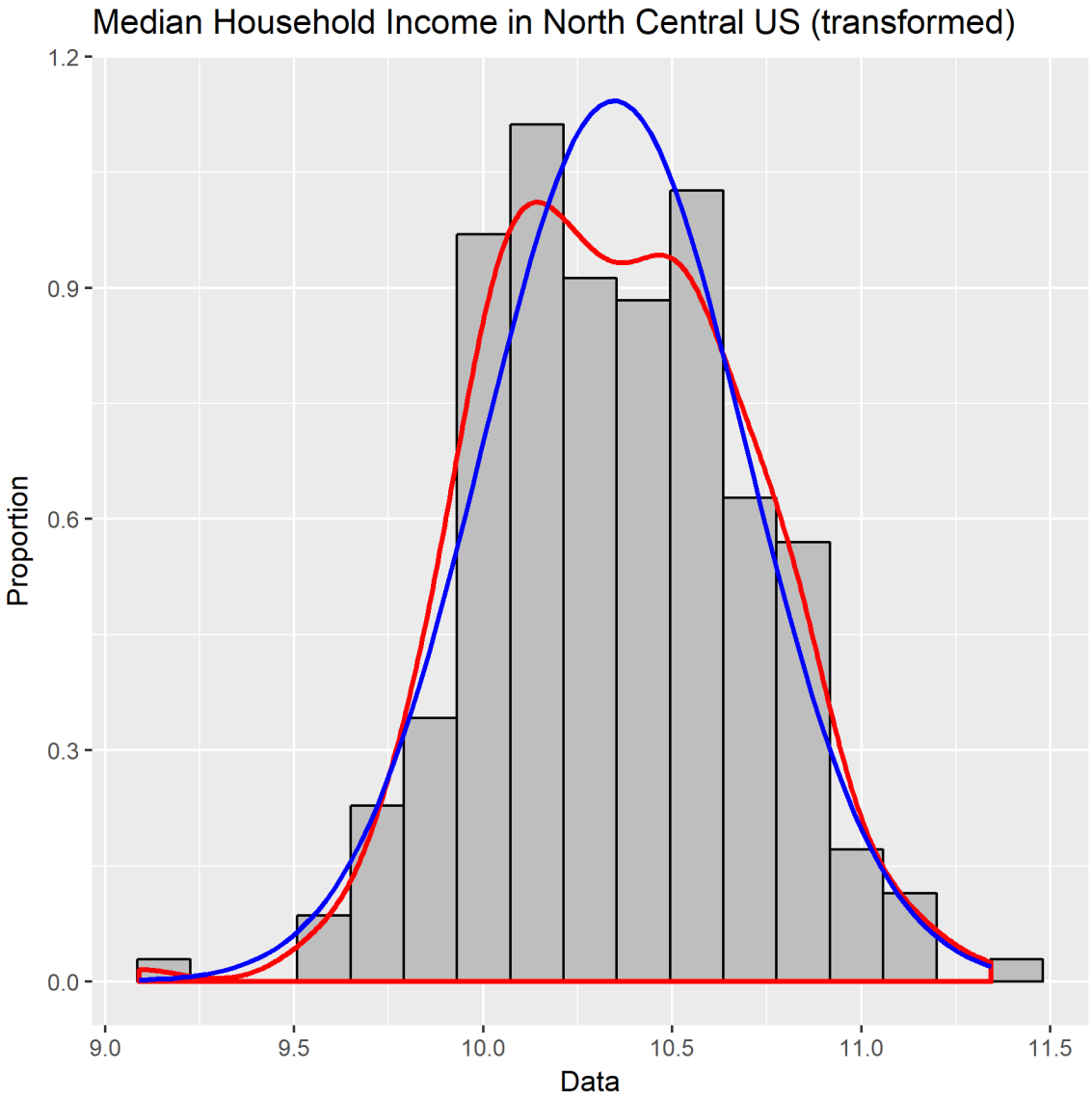
This histogram, as well, shows a nearly normal distribution in the transformed data.



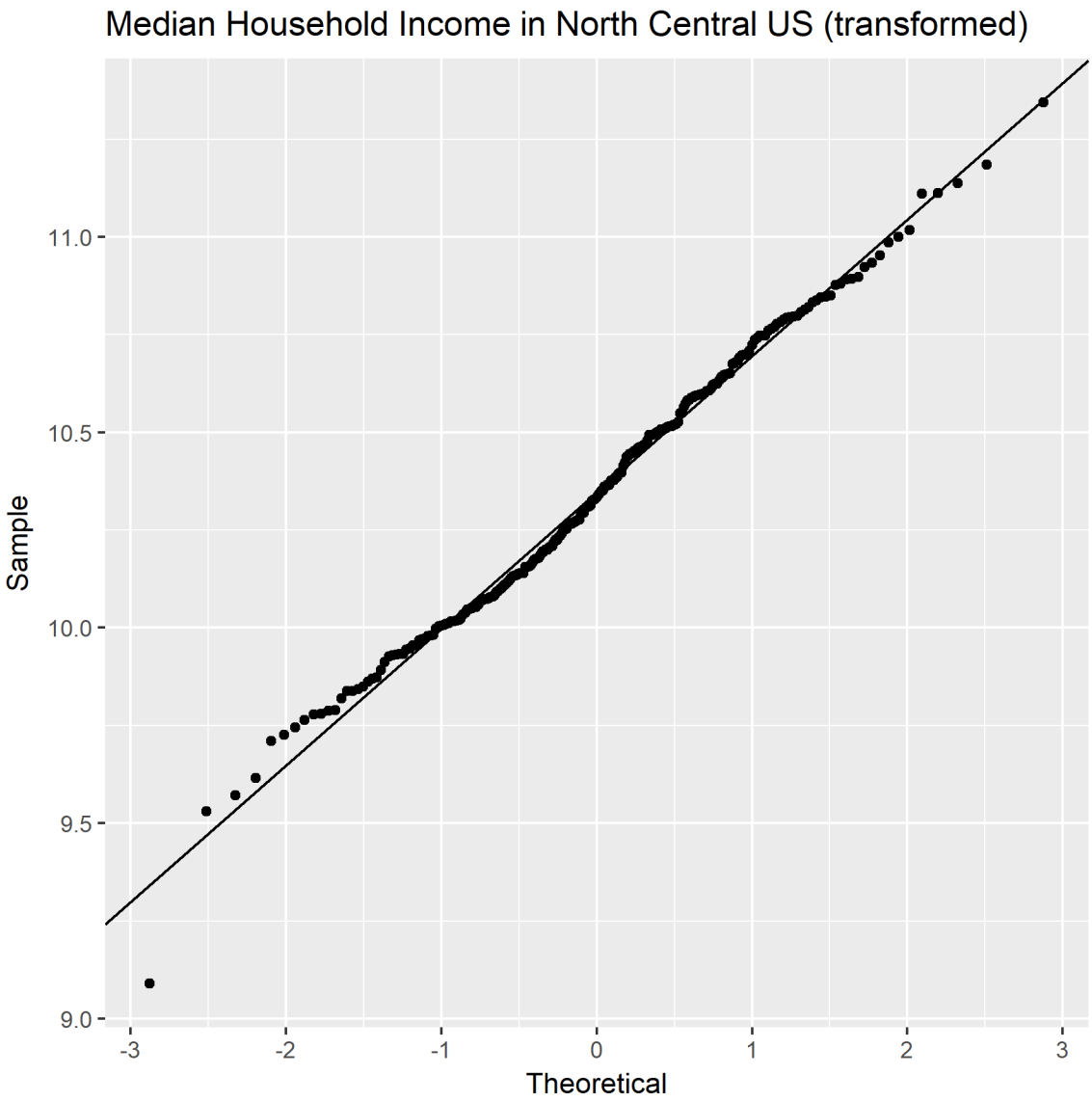
This normal probability plot confirms that the transformed data is normally distributed – notice how all the data points in the plot follow the line of normality quite closely.



This boxplot shows that the transformed North Central data is nearly normally distributed.



This histogram also shows a normal distribution for the transformed NC data.



Once more, the normal probability plot confirms: the transformed data for Median Household Income in the North Central region is normally distributed.

5. 95% Confidence Interval

```
95 percent confidence interval:
 0.1719640 0.2841656
```

6. Hypothesis Test

1. Parameter of interest: $\mu_2 - \mu_1 = \Delta$, the difference in mean income between the Northeast region and the North Central region
2. Hypotheses:
   $H_0: \Delta = \Delta_0 = \$0$
   $H_a: \Delta_a \neq \Delta_0 \rightarrow \Delta \neq \$0$
3. Test statistic (t), degrees of freedom (df), and p-value:

```
 data:  US_NE_MIlog and US_NC_MIlog
 t = 7.9868, df = 508.25, p-value = 9.325e-15
 alternative hypothesis: true difference in means is not equal to 0
```

4. Conclusion:
   $p = 9.325 * 10^{-15} < \alpha = 0.05$
   There is strong evidence (p = 9.325e-15) that the difference in mean income between the Northeast region and the North Central region is not $0.

7. CI/HT Consistency

In part 5, we learn that the 95% confidence interval of the difference in mean household income between the Northeast region and the North Central region ranges from approximately $0.17 to approximately $0.28. That is, we can be 95% confident that this interval captures the true difference. In part 6, we learn that we have very strong evidence (p = 9.325e-15) to reject the null hypotheses that the mean incomes are equal (the difference is 0). These results are consistent because the null hypothesis of 0 is outside our 95% confidence interval.

Part C. Education spending in different time periods

1. Code

```
### PART C ###
# create plots of difference in mean
title="Difference in Education Spending between Period 2 and Period 1"
# boxplot
box <- ggplot(data.frame(US_ESDiff),aes(x="",y=US_ESDiff))+
  stat_boxplot(geom="errorbar")+
  geom_boxplot()+
  ggtitle(title)+
  stat_summary(fun.y=mean,col="black",geom="point",size=3)
ggsave(filename="box ED.png",box,height=6,width=6)
# histogram
hist <- ggplot(data.frame(US_ESDiff),aes(US_ESDiff))+
  geom_histogram(aes(y=..density..),
                 bins=sqrt(length(US_ESDiff))+2,
                 fill="grey",col="black")+
  geom_density(col="red",lwd=1)+
  stat_function(fun=dnorm,args=list(mean=mean(US_ESDiff),
                                    sd=sd(US_ESDiff)),
                col="blue",lwd=1)+
  ggtitle(title)+
  xlab("Data")+
  ylab("Proportion")
ggsave(filename="hist ED.png",hist,height=6,width=6)
# normal probability plot
qq <- ggplot(data.frame(US_ESDiff),aes(sample=US_ESDiff))+
  stat_qq()+
  geom_abline(slope=sd(US_ESDiff),intercept=mean(US_ESDiff))+
  ggtitle(title)+
  xlab("Theoretical")+
  ylab("Sample")
ggsave(filename="qq ED.png",qq,height=6,width=6)

# conduct one-sided hypothesis test with difference, greater
t.test(US_ESDiff, mu=70, conf.level=0.95,alternative="greater")
```

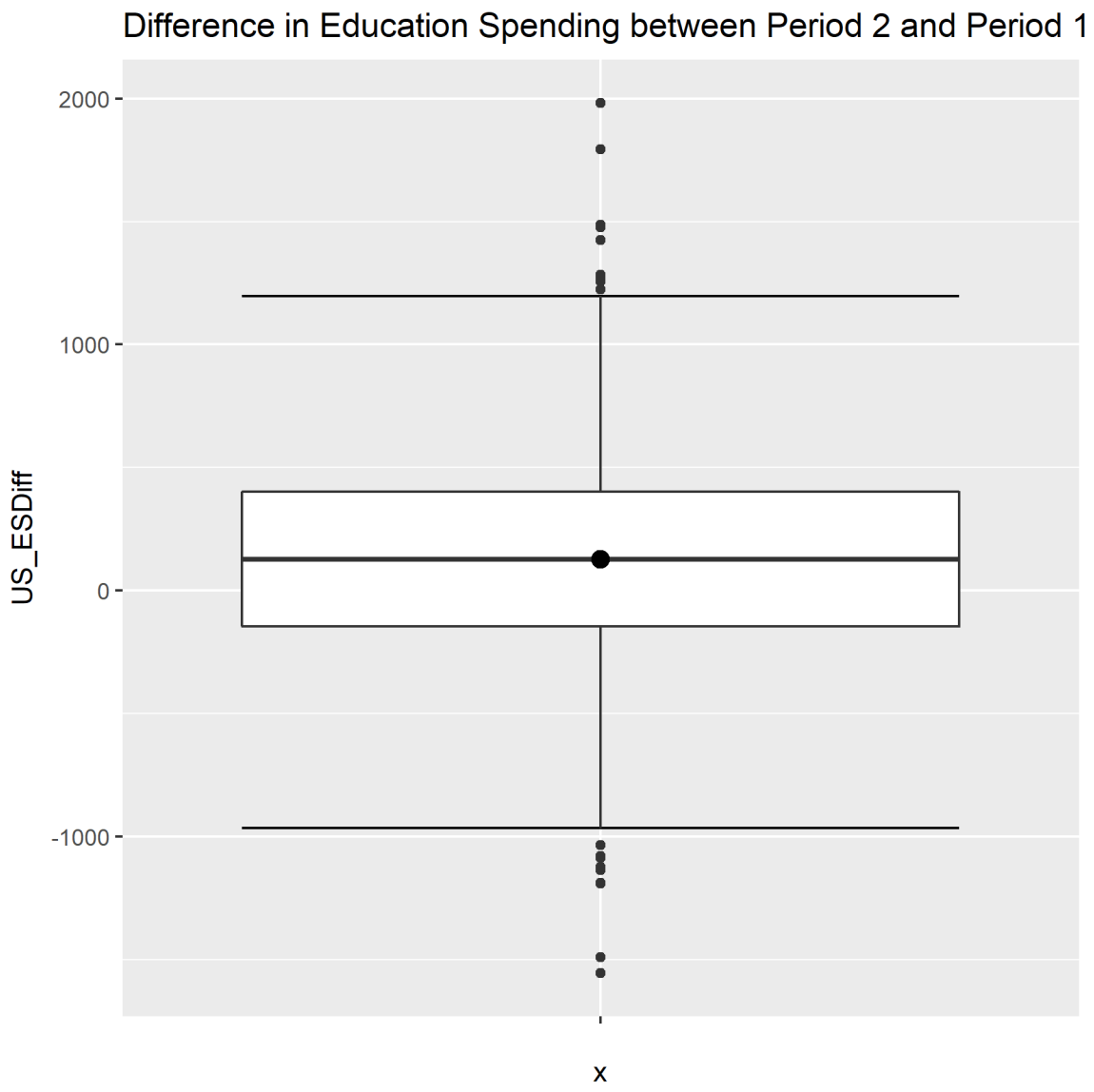
2. Independent or Paired

This data should be analyzed using a two-sample paired procedure. This is because we are analyzing two observations of *the same individuals* (the individuals being the education spending on each pupil and the two observations being period 1 and period 2). A two-sample independent procedure would be more fitting if we were analyzing two independent samples.

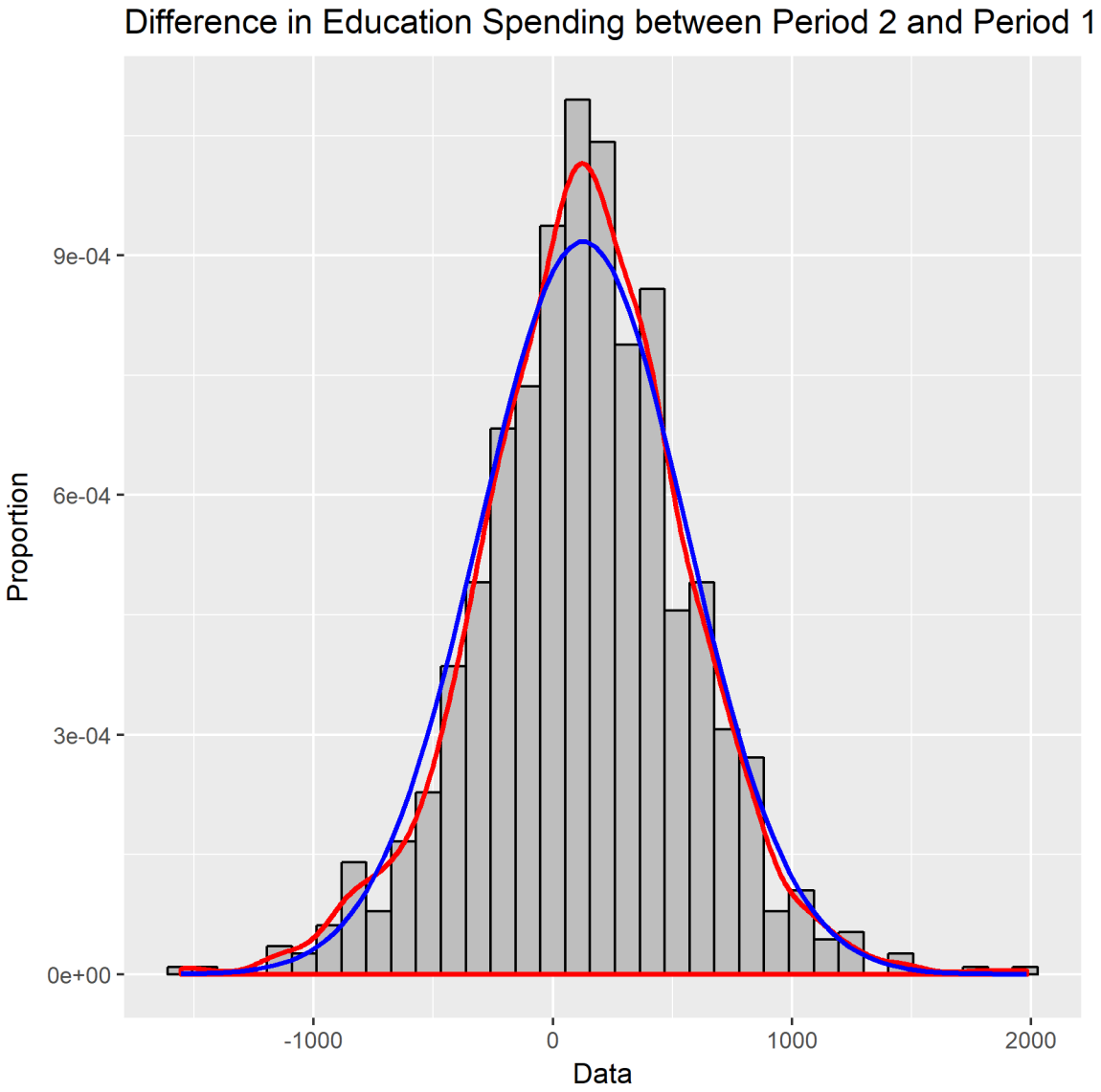
3. One-sided or Two-sided

This data should be analyzed using a one-sided alternative. This is because we are specifically testing whether the spending in period 2 is 70 dollars greater than that in period 1 (e.g. $\Delta_a > \Delta_0$). A two-sided alternative would be more fitting if we were simply testing whether the difference in spending is or is not equal to a specific value (e.g. $\Delta_a \neq \Delta_0$).

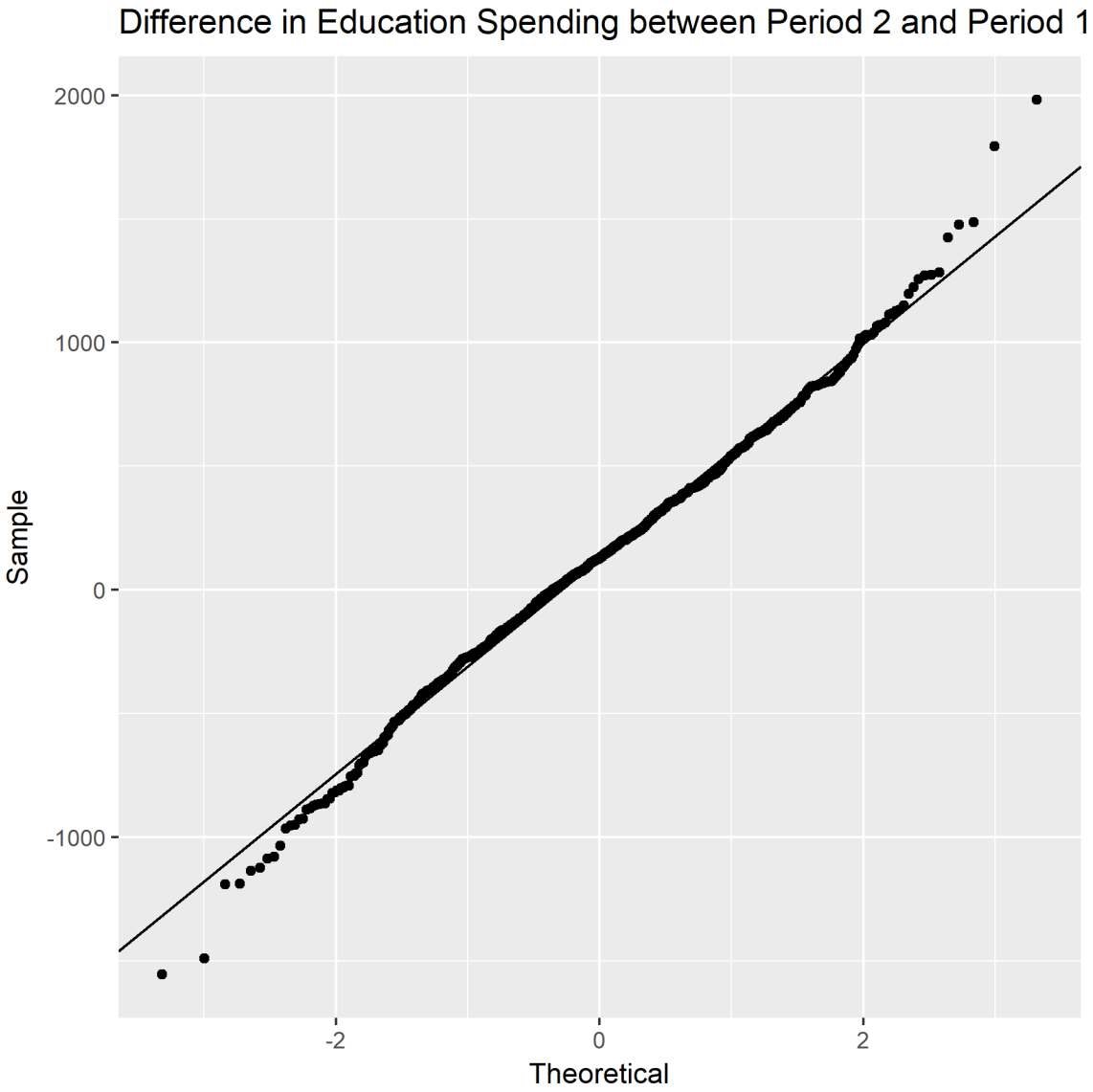
4. Plots



This boxplot suggests that the data for difference in education spending between periods 2 and 1 is normally distributed.



This histogram seems to confirm that the data is, indeed, normally distributed.



Finally, this normal probability plot confirms that the data for Difference in Education Spending between Period 2 and Period 1 is normally distributed; no transformation is necessary.

5. 95% Confidence Interval

```
95 percent confidence interval:
 104.3566      Inf
```

6. Hypothesis Test

1. Parameter of interest: $\mu_2 - \mu_1 = \Delta$, the difference in education spending between periods 2 and 1.
2. Hypotheses:
   $H_0: \Delta = \Delta_0 = \$70$
   $H_a: \Delta_a > \Delta_0 \rightarrow \Delta > \$70$
3. Test statistic (t), degrees of freedom (df), and p-value:

```
data:  US_ESDiff
t = 4.2643, df = 1097, p-value = 1.089e-05
alternative hypothesis: true mean is greater than 70
```

4. Conclusion:
   p = 0.00001089 < α = 0.05
   There is strong evidence (p = 1.089e-05) that the difference in education spending per pupil between period 2 and period 1 is greater than $70.

7. CI/HT Consistency

In part 5, we learn that we can be 95% confident that the interval from approximately $104.35 to infinity captures the true mean difference in education spending per pupil between period 2 and period 1. In part 6, we learn that there is strong evidence (p = 1.089E-05) that this difference is greater than $70. Our results are, therefore, consistent – if we can be 95% confident that the interval from $104.35 and upward captures the true mean difference, then we should also have strong evidence that this true mean is greater than $70 (as this is, obviously, less than $104.35).