NYS Math Test Scores for Charter and Public Schools

Nathan Williamson and Camilo Salazar

May 13, 2017

```
setwd("C:/Users/Nate/Desktop/Statistics_R")
```

Introduction Our data is from the NYC Department of Education website (http://schools.nyc.gov/Accountability/data/TestResults/ELAandMathTestResults). The data are the NYS common core mathematics test scores for all NYC public schools from 2013-2016 in grades 3-8. The data are the average scores for each school in NYC, separated by charter school scores and public school scores. The public school data consisted of 4,429 scores and the charter school data consisted of 576 scores.

The purpose of the project is investigate the distribution of the test scores. Specifically, does the frequency distribution of the test scores resemble some kind of distribution. According to standard practice we would be looking for the test scores to resemble a normal distribtution. We will test for skewness and normality.

We will also investigate the relationship between public and chater schools and perform tests to determine whether there is a statistically significant difference between charter school and public schools test scores.

First we take out all of the Charter school math scores for 2013-2016.

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.3.3
```

```
chartermath <- read_excel("CharterSchoolResults20132016Public.xlsx",col_names=TRUE,skip=6,sheet="Math")
attach(chartermath)
head(chartermath)
```

```
## # A tibble: 6 × 17
##      DBN                                `School Name`      Grade  Year
##    <chr>                                        <chr>      <chr> <dbl>
## 1 84K037   BEGINNING WITH CHILDREN CHARTER SCHOOL II          3  2015
## 2 84K037   BEGINNING WITH CHILDREN CHARTER SCHOOL II          3  2016
## 3 84K037   BEGINNING WITH CHILDREN CHARTER SCHOOL II          4  2016
## 4 84K037   BEGINNING WITH CHILDREN CHARTER SCHOOL II All Grades  2015
## 5 84K037   BEGINNING WITH CHILDREN CHARTER SCHOOL II All Grades  2016
## 6 84K125 SUCCESS ACADEMY CHARTER SCHOOL - BED-STUY 2          3  2015
## # ... with 13 more variables: Category <chr>, `Number Tested` <dbl>,
## #   Mean.Scale.Score <chr>, `#` <chr>, `%` <chr>, `#__1` <chr>,
## #   `%__1` <chr>, `#__2` <chr>, `%__2` <chr>, `#__3` <chr>, `%__3` <chr>,
## #   `#__4` <chr>, `%__4` <chr>
```

Now we take out just the total scores for all students in the schools in all four years

```
charter.math.scores <- subset(chartermath, subset=Grade=="All Grades",select="Mean.Scale.Score")
charter.math.scores<-as.numeric(unlist(charter.math.scores))
head(charter.math.scores)
```

```
## [1] 320.8636 332.1087 347.3036 360.4480 358.9726 361.5256
```

```
length(t(charter.math.scores))
```

```
## [1] 576
```

Next we take out all of the math scores for public schools

```
library(readxl)
publicmath <- read_excel("SchoolMathResults20132016Public.xlsx",col_names=TRUE,skip=6,sheet="All Studen
attach(publicmath)
```

```
## The following objects are masked from chartermath:
##
##     #, #__1, #__2, #__3, #__4, %, %__1, %__2, %__3, %__4,
##     Category, DBN, Grade, Mean.Scale.Score, Number Tested, School
##     Name, Year
```

```
head(publicmath)
```

```
## # A tibble: 6 × 17
##      DBN             `School Name` Grade  Year     Category
##    <chr>                     <chr> <chr> <dbl>        <chr>
## 1 01M015 P.S. 015 ROBERTO CLEMENTE     3  2013 All Students
## 2 01M015 P.S. 015 ROBERTO CLEMENTE     3  2014 All Students
## 3 01M015 P.S. 015 ROBERTO CLEMENTE     3  2015 All Students
## 4 01M015 P.S. 015 ROBERTO CLEMENTE     3  2016 All Students
## 5 01M015 P.S. 015 ROBERTO CLEMENTE     4  2013 All Students
## 6 01M015 P.S. 015 ROBERTO CLEMENTE     4  2014 All Students
## # ... with 12 more variables: `Number Tested` <dbl>,
## #   Mean.Scale.Score <chr>, `#` <chr>, `%` <chr>, `#__1` <chr>,
## #   `%__1` <chr>, `#__2` <chr>, `%__2` <chr>, `#__3` <chr>, `%__3` <chr>,
## #   `#__4` <chr>, `%__4` <chr>
```

Now we take out the scores for all grades for all years.

```
public.math.scores <- subset(publicmath, subset=Grade=="All Grades",select="Mean.Scale.Score")
public.math.scores<-as.numeric(unlist(public.math.scores))
head(public.math.scores)
```

```
## [1] 276.3944 278.0317 278.2931 285.6731 300.5000 307.8558
```

```
length(t(public.math.scores))
```

```
## [1] 4430
```

We see there is an NA value so we remove it from the data.

```
which(is.na(public.math.scores))
```

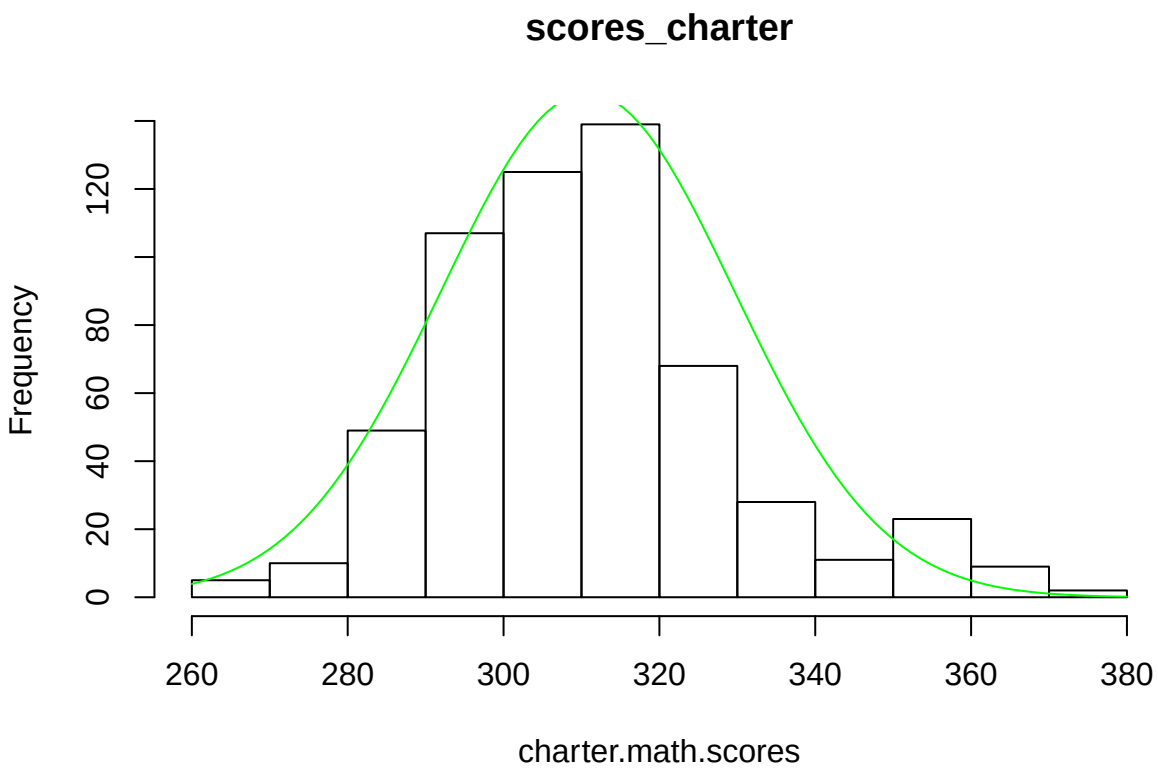
```
## [1] 764
```

```
public.math.scores <- public.math.scores[-764]
which(is.na(public.math.scores))
```

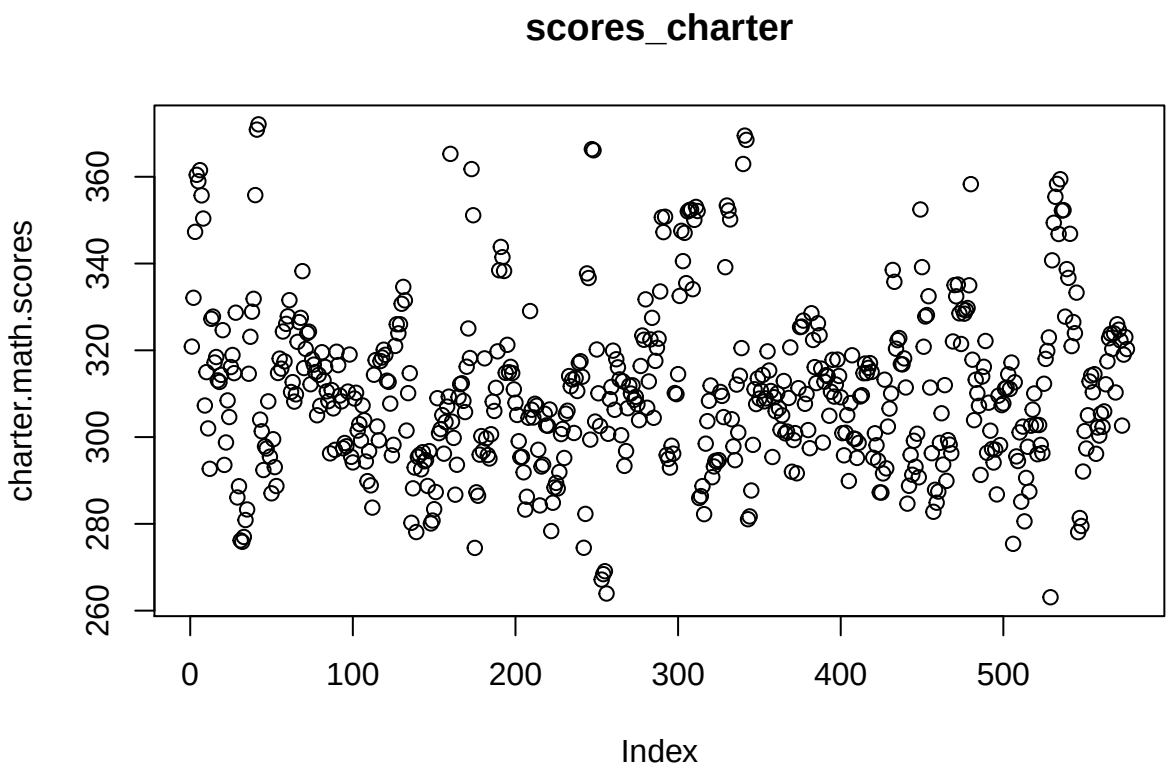
```
## integer(0)
```

Now we perform an EDA on each data set. For the charter school math scores we have:

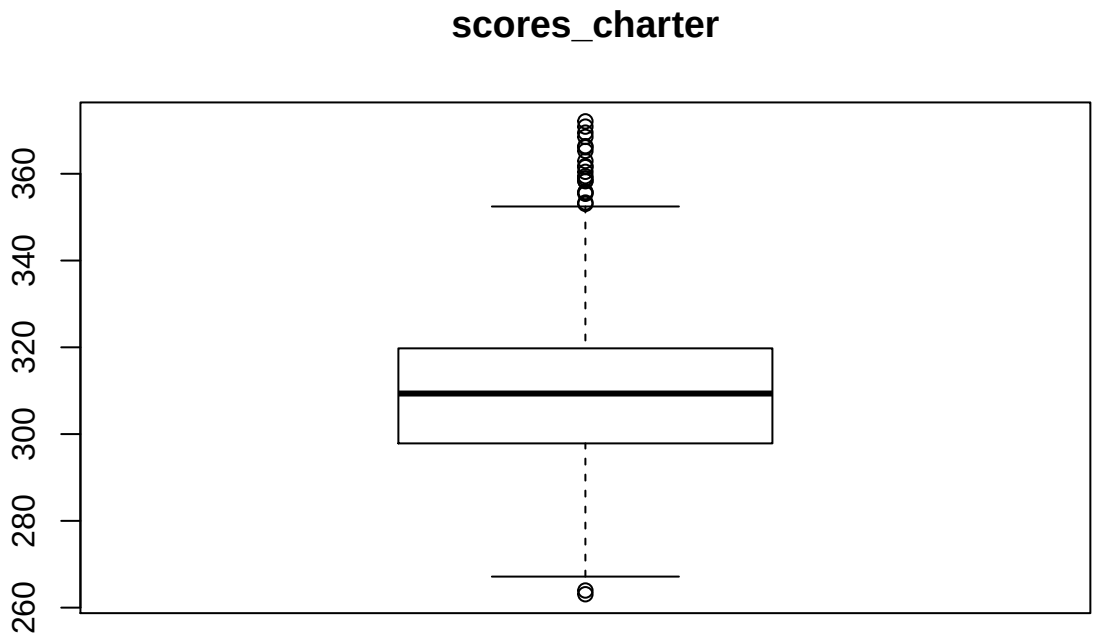
```
#EDA charter
hist(charter.math.scores,main="scores_charter")
curve(7000*dnorm(x,mean=310.8,sd=sqrt(355.024)), col='green', add=TRUE)
```



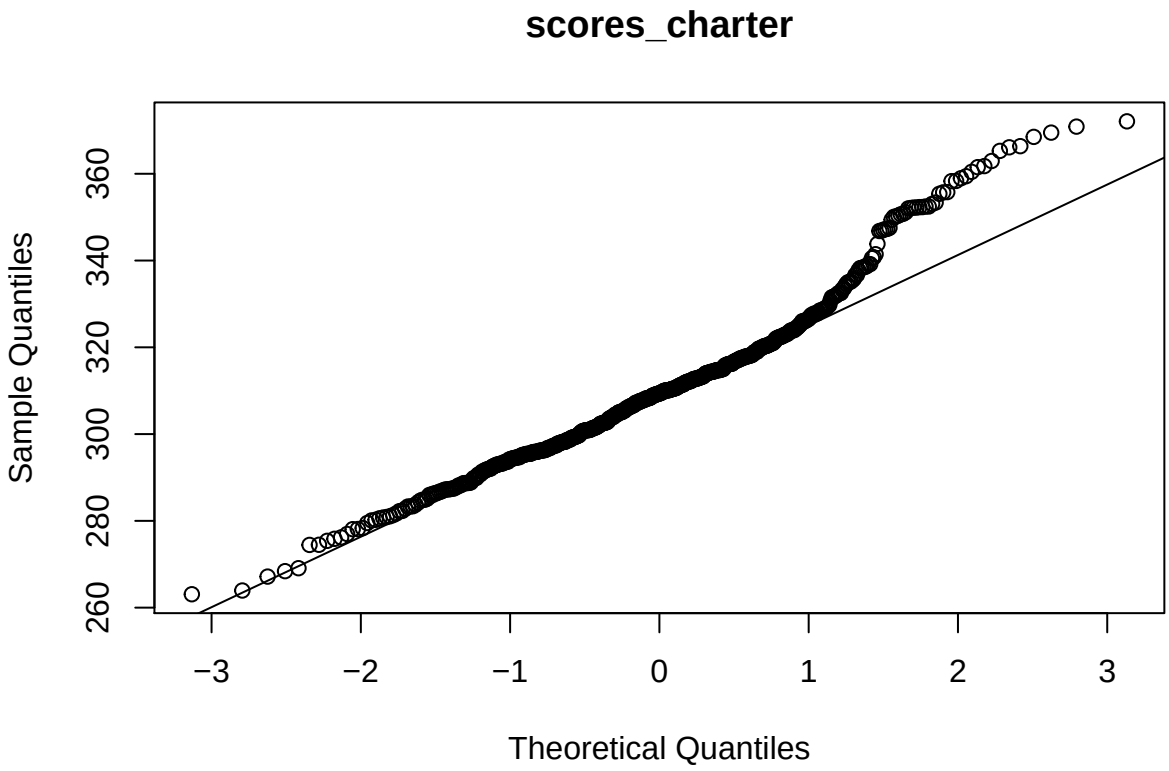

```
plot(charter.math.scores,main="scores_charter")
```




```
boxplot(charter.math.scores,main="scores_charter")
```




```
qqnorm(charter.math.scores,main="scores_charter")
qqline(charter.math.scores)
```




```
summary(charter.math.scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   263.1   297.9   309.4   310.8   319.8   372.1
```

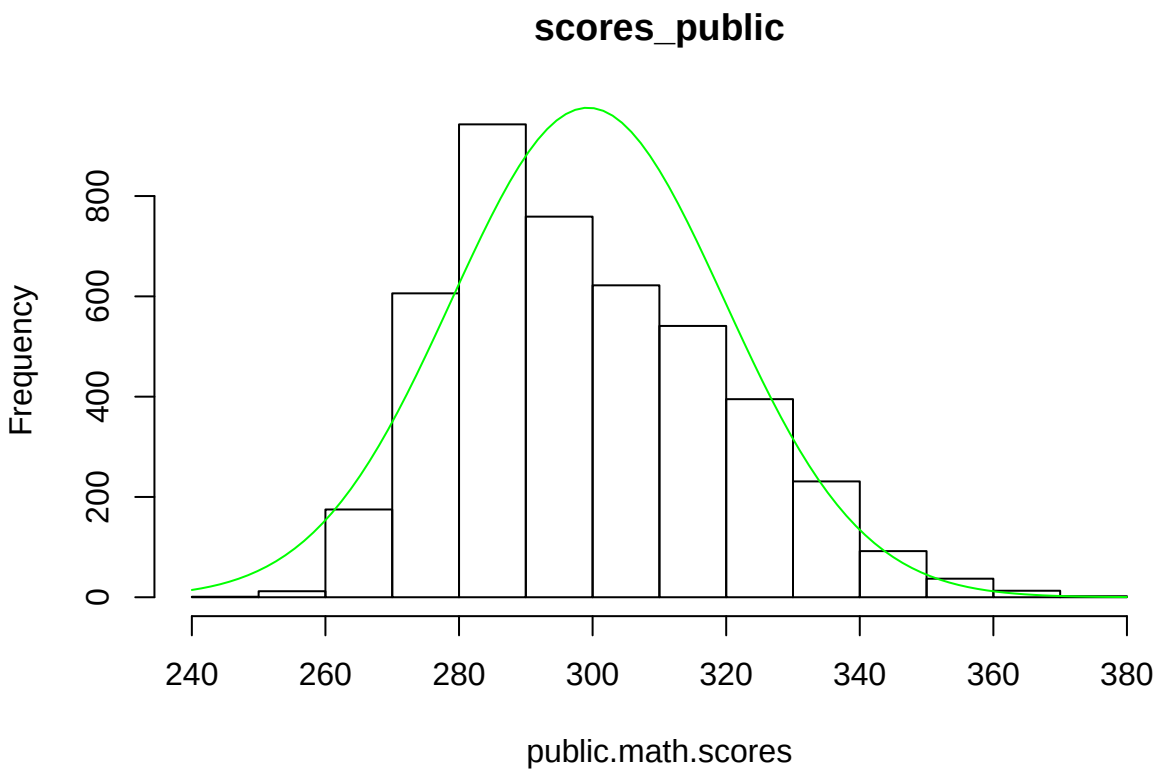
```
var(charter.math.scores)
```

```
## [1] 355.024
```

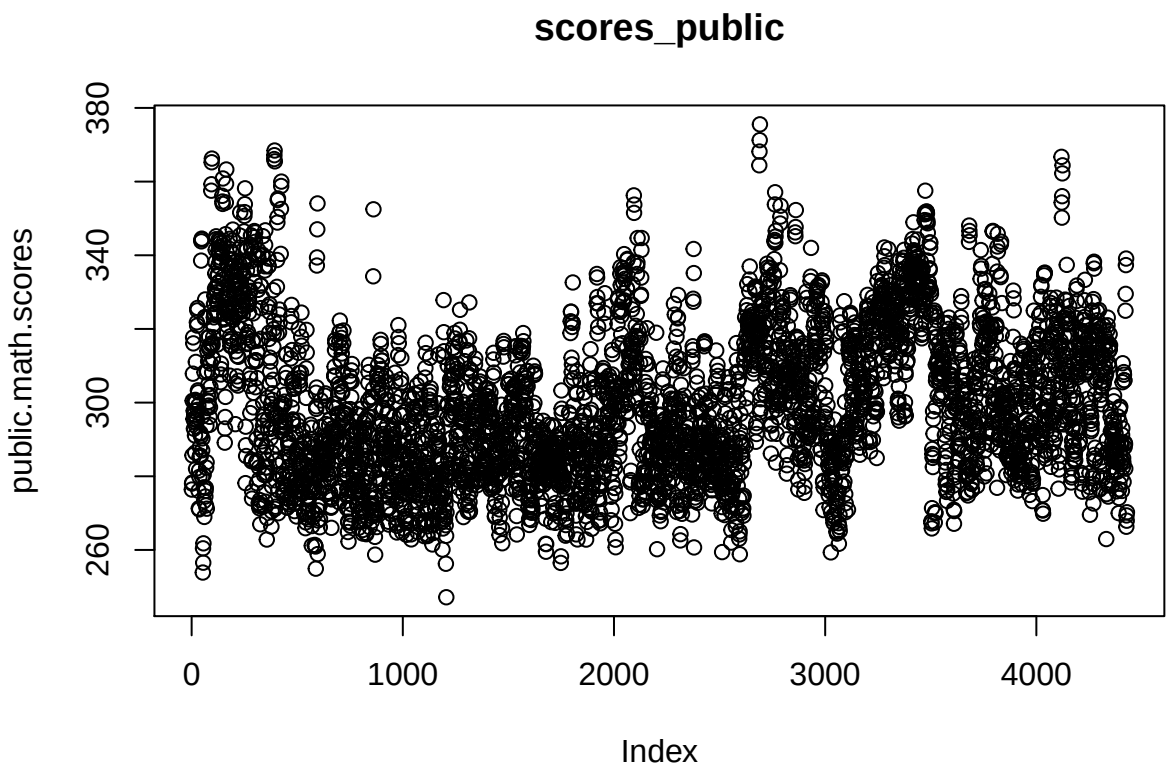
The histogram appears to be bell shaped and skewed to the right. There is no apparent correlation from the plot of the scores as we would expect because the x "index" axis is categorical, and we would not expect a correlation. We can see there are several outliers in the high range of the data from the box plot. From the qqplot we can see there is significant non-normality on the high end of the data.

Now we perform an EDA for the public school math scores

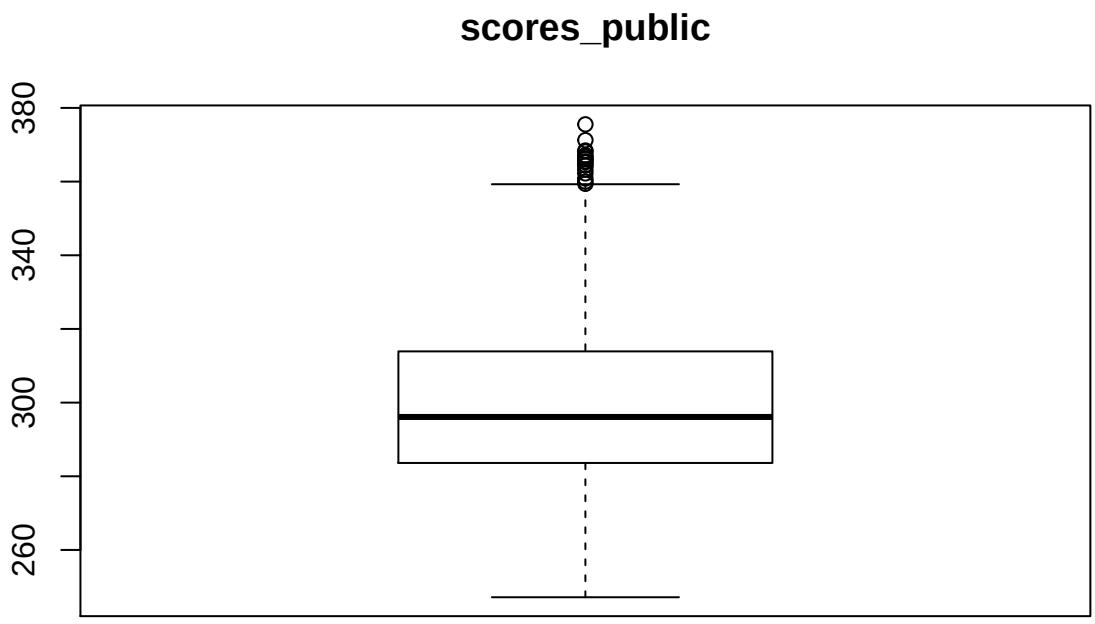
```
#EDA public
hist(public.math.scores,main="scores_public")
curve(50000*dnorm(x,mean=299.3,sd=sqrt(417.5698)), col='green', add=TRUE)
```



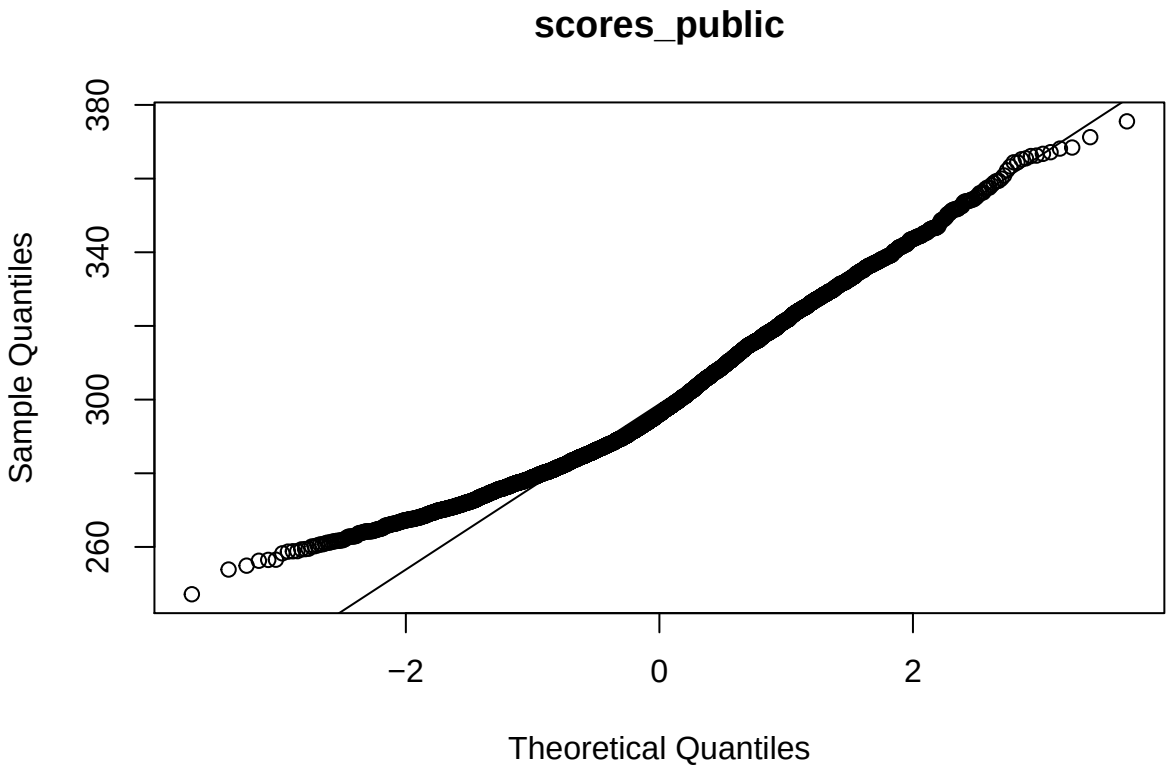

```
plot(public.math.scores,main="scores_public")
```




```
boxplot(public.math.scores,main="scores_public")
```




```
qqnorm(public.math.scores,main="scores_public")
qqline(public.math.scores)
```




```
summary(public.math.scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   247.2   283.6   296.1   299.3   313.9   375.5
```

```
var(public.math.scores)
```

```
## [1] 417.5698
```

The histogram is bell-shaped and skewed to the right. There is no apparent correlation in the plot of the scores as we would expect. There are outliers in the high end of the box plot. There is non-normality apparent in the s-shape of the low end of the distribution in the qq-plot.

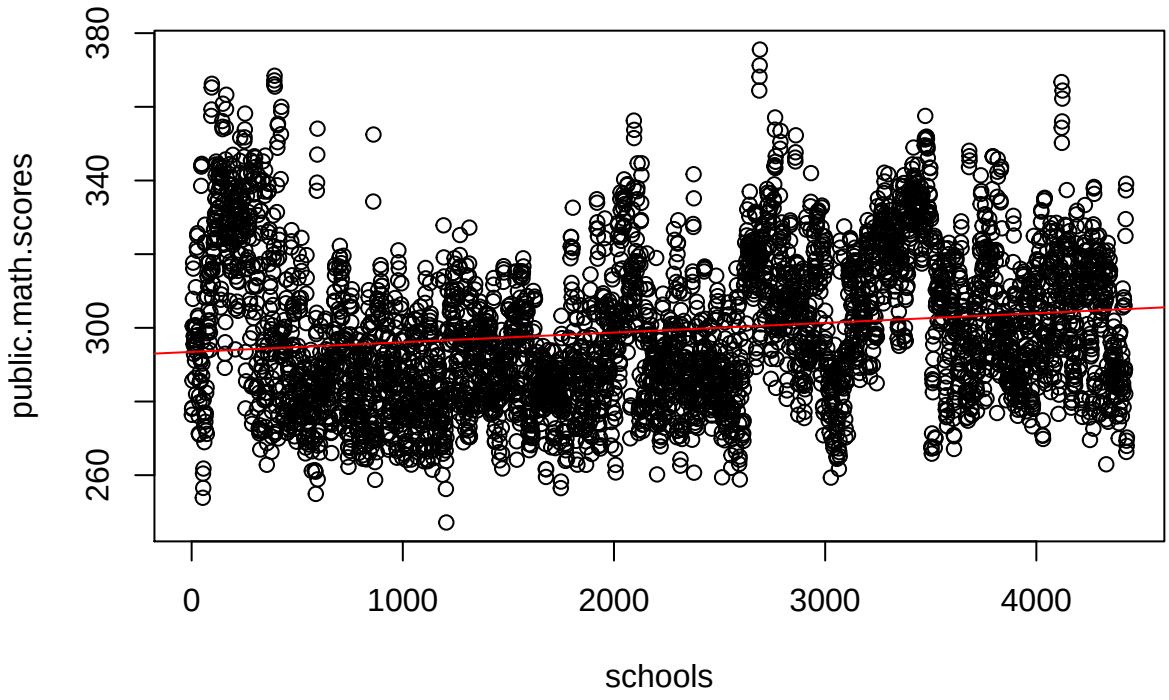
Linear Regression

```
schools <- seq(1,4429,1)
linear.model <- lm(public.math.scores ~ schools)
summary(linear.model)
```

```
##
## Call:
## lm(formula = public.math.scores ~ schools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.475 -15.325  -3.215  13.414  75.007
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 2.935e+02  6.059e-01  484.32   <2e-16 ***
## schools     2.624e-03  2.369e-04   11.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.16 on 4427 degrees of freedom
## Multiple R-squared:  0.02696,    Adjusted R-squared:  0.02674
## F-statistic: 122.7 on 1 and 4427 DF,  p-value: < 2.2e-16
```

```
plot(schools,public.math.scores)
abline(linear.model,col="red")
```



It is obvious that a linear regression is not a good fit for our data.

Quadratic Regression

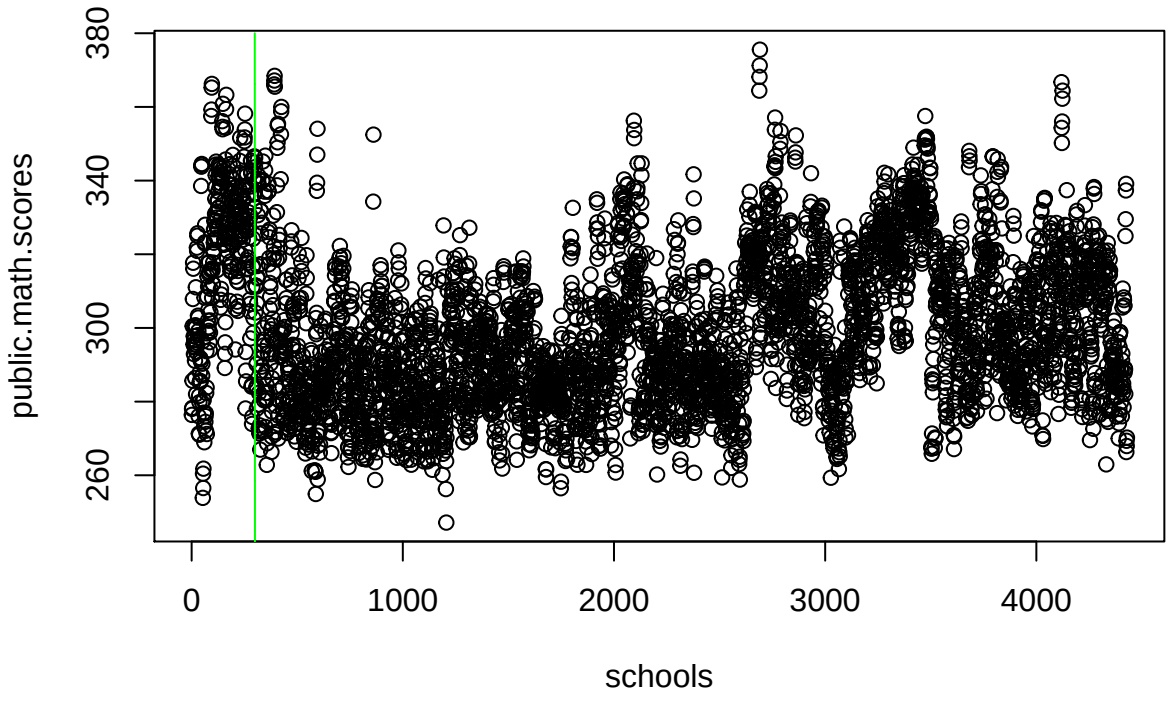
```
#quadratic model
schools <- seq(1,4429,1)
schools2 <- schools^2
quadratic.model <-lm(public.math.scores ~ schools + schools2)
summary(quadratic.model)
```

```
##
## Call:
## lm(formula = public.math.scores ~ schools + schools2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.099 -14.820  -3.036  13.215  78.581
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.018e+02  8.937e-01  337.66   <2e-16 ***
## schools     -8.620e-03  9.318e-04   -9.25   <2e-16 ***
## schools2     2.538e-06  2.037e-07   12.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.82 on 4426 degrees of freedom
## Multiple R-squared:  0.05995,    Adjusted R-squared:  0.05952
## F-statistic: 141.1 on 2 and 4426 DF,  p-value: < 2.2e-16
```

```
plot(schools,public.math.scores)
x <- seq(240,380,0.001)
fit <- 301.8-x*0.008619+0.000002538*x^2
lines(fit,x, col='green', add=TRUE)
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "add" is not a
## graphical parameter
```



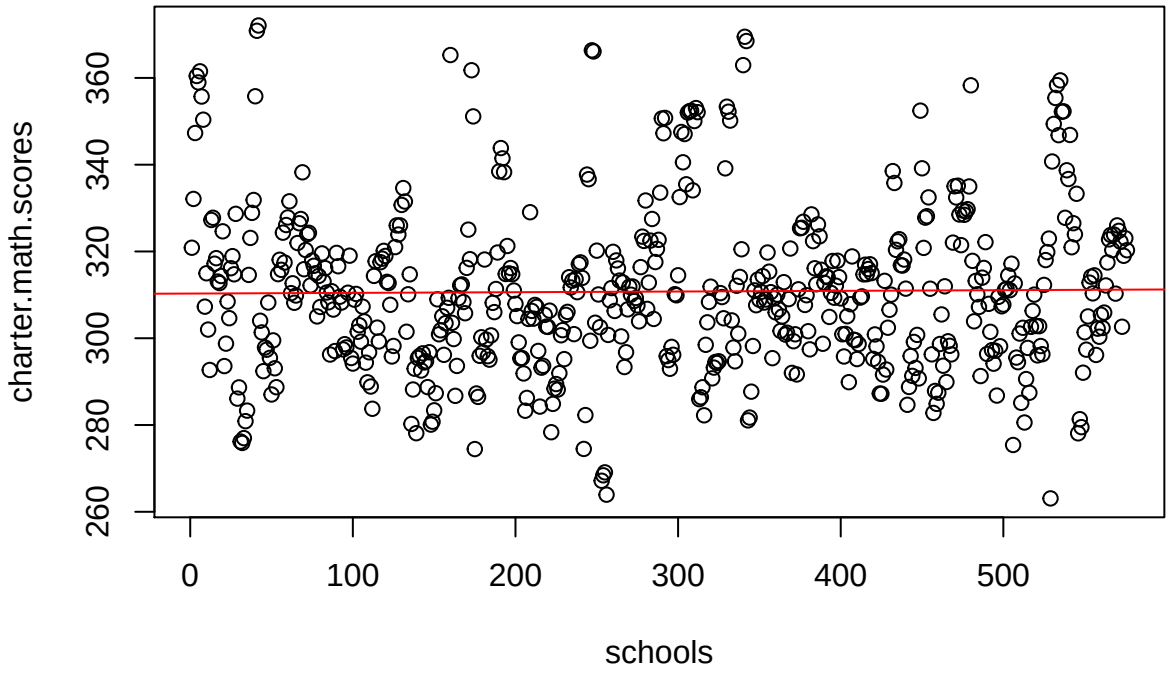
We can see that a quadratic model does not provide a good fit for our data.

```
schools <- seq(1,576,1)
linear.model <- lm(charter.math.scores ~ schools)
summary(linear.model)
```

```
##
```

```
## Call:
## lm(formula = charter.math.scores ~ schools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.038 -12.934  -1.391   9.101  61.719
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.103e+02  1.573e+00 197.211   <2e-16 ***
## schools     1.565e-03  4.725e-03   0.331    0.741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.86 on 574 degrees of freedom
## Multiple R-squared:  0.0001912,  Adjusted R-squared:  -0.001551
## F-statistic: 0.1098 on 1 and 574 DF,  p-value: 0.7405
```

```
plot(schools,charter.math.scores)
abline(linear.model,col="red")
```



A linear model is not a good fit for the charter school data.

Quadratic model for Charter Schools

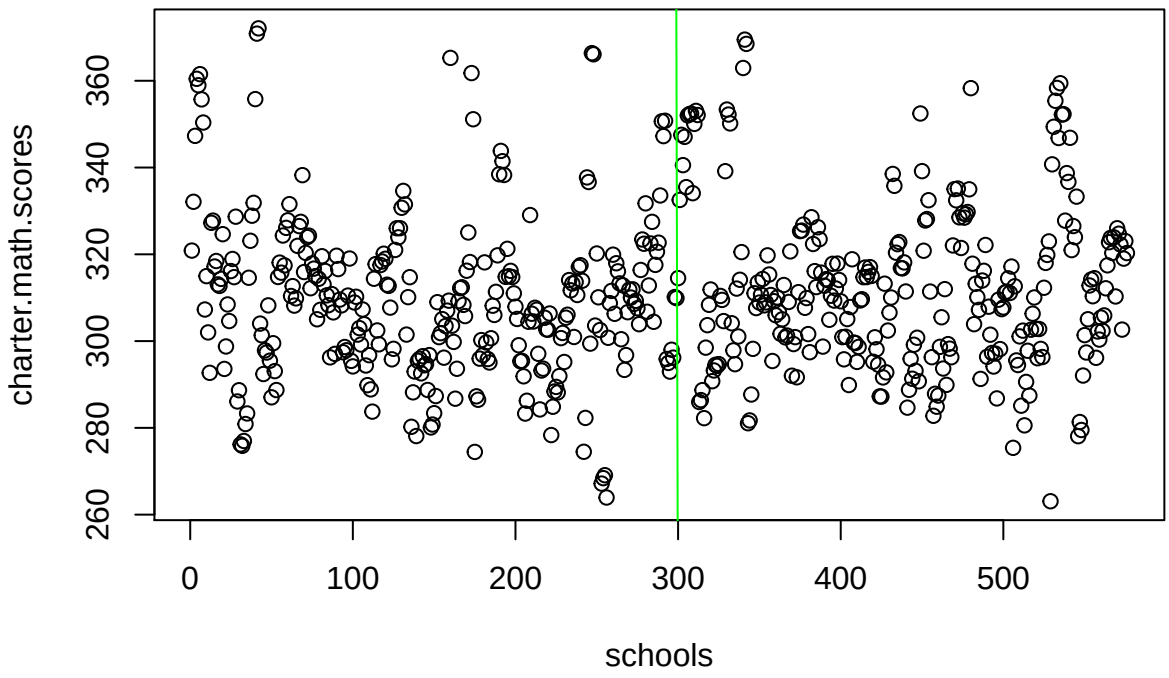
```
#quadratic model
schools <- seq(1,576,1)
schools2 <- schools^2
```

```r
quadratic.model <-lm(charter.math.scores ~ schools + schools2)
summary(quadratic.model)
```

```
##
## Call:
## lm(formula = charter.math.scores ~ schools + schools2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.220 -12.982  -0.918   8.396  60.455
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.143e+02  2.357e+00 133.374   <2e-16 ***
## schools     -4.015e-02  1.886e-02  -2.128   0.0337 *
## schools2     7.229e-05  3.166e-05   2.284   0.0228 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.79 on 573 degrees of freedom
## Multiple R-squared:  0.009208,   Adjusted R-squared:  0.00575
## F-statistic: 2.663 on 2 and 573 DF,  p-value: 0.07062
```

```r
plot(schools,charter.math.scores)
x <- seq(240,380,0.01)
fit <- 301.8-x*0.008619+0.000002538*x^2
lines(fit,x, col='green', add=TRUE)
```

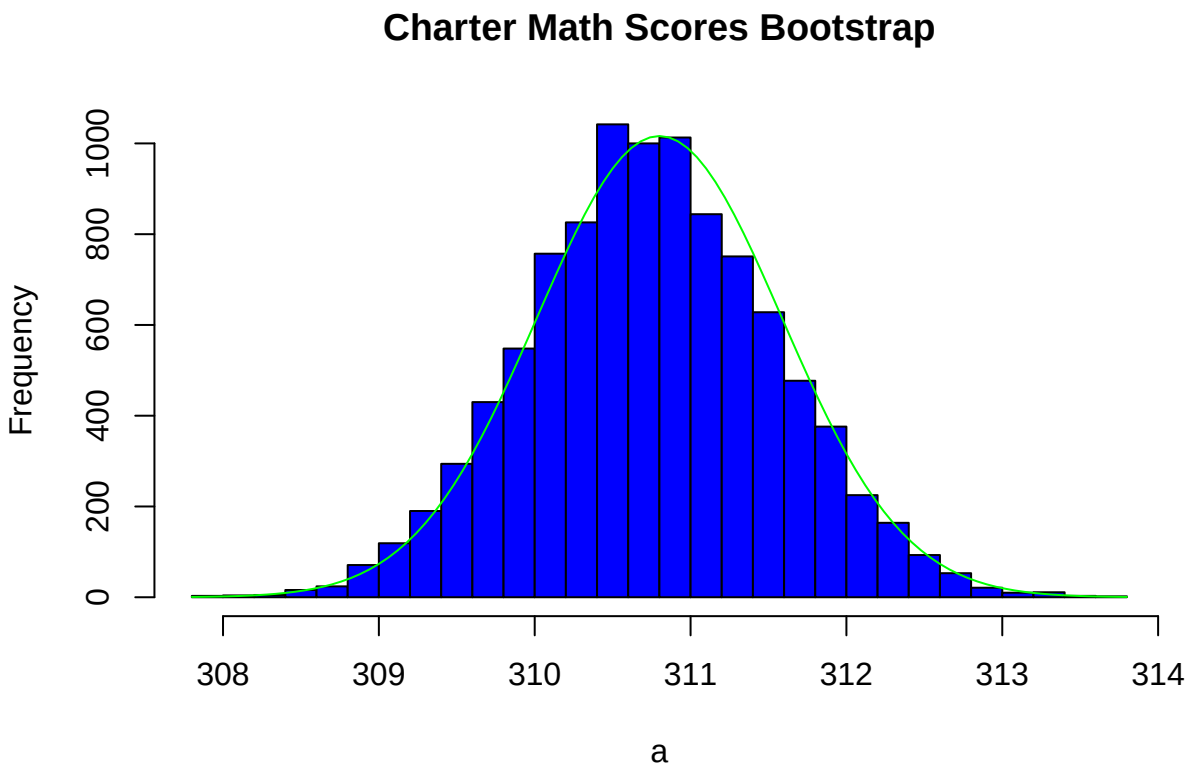
```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "add" is not a
## graphical parameter
```



A quadratic regression is clearly a bad fit for the data. Because of the lack of correlation between the variables these regressions are not useful to our analysis.

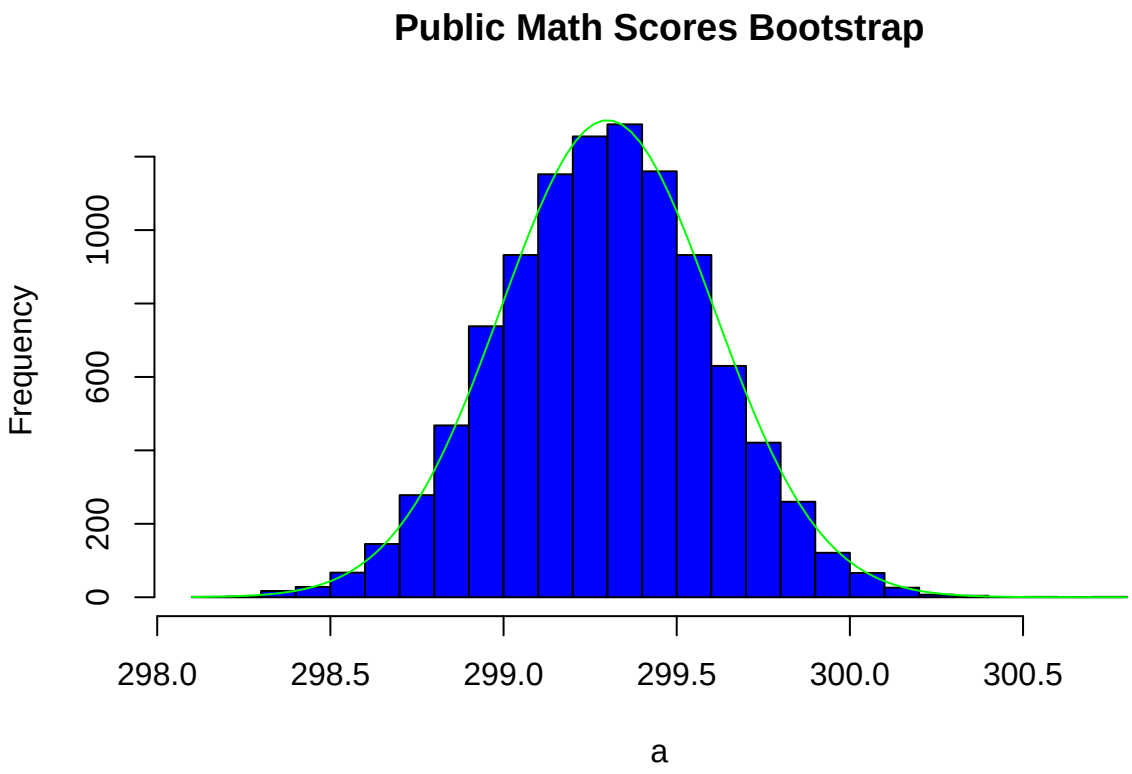
Now we will perform the bootstrap on each distribution. For the charter school data:

```
a <- numeric(10000)
for(i in 1:10000) a[i] <- mean(sample(charter.math.scores,replace=T))
hist(a,main="Charter Math Scores Bootstrap",col="blue",breaks=25)
x <- seq(307,314,0.01)
curve(2000*dnorm(x,mean=310.8,sd=sqrt(355.024)/sqrt(576)), col='green', add=TRUE)
```



Bootstrap for the public school math scores:

```
a <- numeric(10000)
for(i in 1:10000) a[i] <- mean(sample(public.math.scores,replace=T))
hist(a,main="Public Math Scores Bootstrap",col="blue",breaks=25)
x <- seq(298,301,0.01)
curve(1000*dnorm(x,mean=299.3,sd=sqrt(417.5698)/sqrt(4430)), col='green', add=TRUE)
```



The bootstrap sampling distribution of the mean for each data set appear to be normal and we have overlayed the theoretical normal curve for the data. The bootstrap has the same mean as the original data and var = var/sqrt(n). This data appears to be normally distributed as the theoretical curves are a close fit to the histogram of the bootstrap.

Now we separate the data for Asian, Black, Hispanic and White students.

```
library(readxl)
publicmathethnic <- read_excel("SchoolMathResults20132016Public.xlsx",sheet="Ethnicity",col_names=TRUE,s
attach(publicmathethnic)
```

```
## The following objects are masked from publicmath:
##
##     #, #__1, #__2, #__3, #__4, %, %__1, %__2, %__3, %__4,
##     Category, DBN, Grade, Mean.Scale.Score, Number Tested, School
##     Name, Year
```

```
## The following objects are masked from chartermath:
##
##     #, #__1, #__2, #__3, #__4, %, %__1, %__2, %__3, %__4,
##     Category, DBN, Grade, Mean.Scale.Score, Number Tested, School
##     Name, Year
```

```
asian <- subset(publicmathethnic, select="Mean.Scale.Score",subset=(Grade=="All Grades" & Category=="As
```

```
## Warning: drop ignored
```

```
black <- subset(publicmathethnic, select="Mean.Scale.Score",subset=(Grade=="All Grades" & Category=="Bl
```

```
## Warning: drop ignored
```

```
white <- subset(publicmathethnic, select="Mean.Scale.Score",subset=(Grade=="All Grades" & Category=="Wh
```

```
## Warning: drop ignored
```

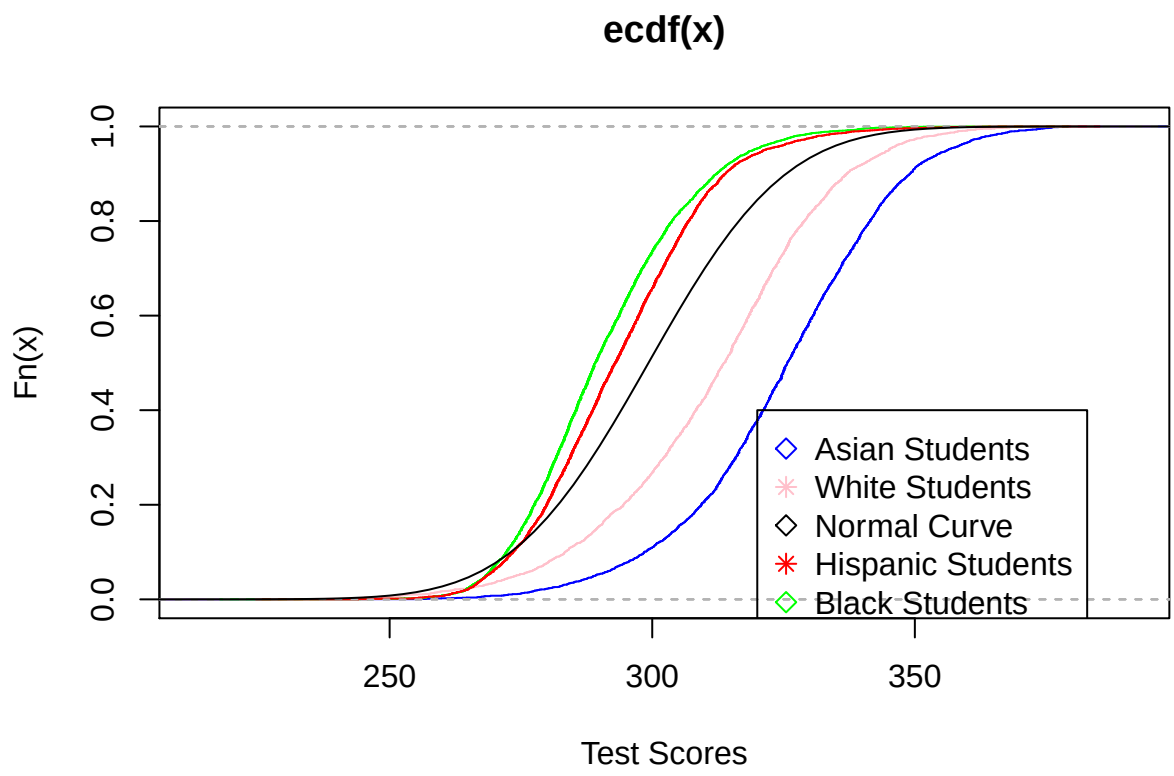
```
hispanic <- subset(publicmathethnic, select="Mean.Scale.Score",subset=(Grade=="All Grades" & Category==
```

```
## Warning: drop ignored
```

Now we clean the data

```
int.asian <- as.numeric(unlist(asian))
int.asian <- int.asian[!is.na(int.asian)]
int.black <- as.numeric(unlist(black))
int.black <- int.black[!is.na(int.black)]
int.white <- as.numeric(unlist(white))
int.white <- int.white[!is.na(int.white)]
int.hispanic <- as.numeric(unlist(hispanic))
int.hispanic <- int.hispanic[!is.na(int.hispanic)]
```

```
#Plot the ecdf's of the different ethnicities
plot.ecdf(int.asian,xlab="Test Scores", col="blue", pch=5)
plot.ecdf(int.white,col="pink", pch=8, add=TRUE)
plot.ecdf(int.black, col="green", pch=5, add=TRUE)
plot.ecdf(int.hispanic,col="red", pch=8, add=TRUE)
curve(pnorm(x,mean(public.math.scores),sd(public.math.scores)),150,400, col="black",add=TRUE)
legend(320,0.4,c("Asian Students","White Students","Normal Curve","Hispanic Students","Black Students")
```



From the ECDF we can see clear differences between the test scores based on ethnicity. We can see that the scores

Now we performs Student's t-test.

```
int.public.math <- as.numeric(unlist(public.math.scores))
int.charter.math <- as.numeric(unlist(charter.math.scores))
t.test(int.charter.math,int.public.math)
```

```
##
##  Welch Two Sample t-test
##
## data:  int.charter.math and int.public.math
## t = 13.602, df = 762.05, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.811948 13.121688
## sample estimates:
## mean of x mean of y
##  310.7503  299.2835
```

The Welch Two-Sample t-test has a small p-value, indicating that there is a close to 0 probabililty that the difference in means is 0. The 95% confidence interval for the difference in means is (9.8,13.1), indicating that the charter school test scores are between 9.8 and 13.1 higher than the public school scores.

F-test to compare two variances

```
var.test(int.public.math,int.charter.math)
```

```
##
##  F test to compare two variances
##
## data:  int.public.math and int.charter.math
## F = 1.1762, num df = 4428, denom df = 575, p-value = 0.01178
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.037019 1.326155
## sample estimates:
## ratio of variances
##           1.176173
```

According to the p=value, 0.01178, which indicates the variances are probably not the sample but still not significantly different between the charter and public school data. The ratio of the variances is 1.176 which is not significantly different. The test indicates that the public school data has a higher variance than the charter school data.

```
ks.test(charter.math.scores,public.math.scores)
```

```
## Warning in ks.test(charter.math.scores, public.math.scores): p-value will
## be approximate in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  charter.math.scores and public.math.scores
## D = 0.31112, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

According to the Kolmogorov-Smirnov test our two samples are likely not from the same distribution.

Conclusion

Our models for public and charter school scores appear to be bell shaped but exhibit skewness and non-normality. High achieving students caused skewness in the charter school scores, whereas low achieving students caused skewness in the public school scores. Charter schools performed significantly better than public schools on the state common core tests. The 95% confidence interval for the difference in means is (9.8,13.1), indicating that the charter school test scores are between 9.8 and 13.1 higher than the public school scores.