Course	STA258H5S: Statistics with Applied Probability
Course Coordinator	Asal Aslemand
Group Members	Mohammad Fahim Uddin Alvi, Srabon Ratan, Chunyuan Wang, Keying Rao, Wenyi Wang, Yubin Yao
Topic	School Diversity
Rationale	School diversity is always an eye-attracting topic. In this research, what we are interested in is, if there is a real significant difference in school diversity and which factors lead to these differences. Firstly, we choose the Asian, Black and white variables to figure out the difference in proportion for different schools. In this case, we may do ANOVA analysis to see if there is a significant difference. Thereafter, we will choose the variable ST and d_locate_Txt as it includes the characteristics about the schools such as state and (school district, town, rural, city, etc). We believe that both school size and type of school district will have a positive relationship to the school diversity. Hopefully, we can get some meaningful insights through this research.

TidyTuesday link	https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-09-24
Data Files used	school_diversity.csv

School Diversity Descriptive Analysis

Mohammad Fahim Uddin Alvi, Srabon Ratan, Chunyuan Wang, Keying Rao, Wenyi Wang, Yubin Yao

Load the data set from Github repository

```
data <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   LEAID = col_character(),
##   LEA_NAME = col_character(),
##   ST = col_character(),
##   d_Locale_Txt = col_character(),
##   SCHOOL_YEAR = col_character(),
##   AIAN = col_double(),
##   Asian = col_double(),
##   Black = col_double(),
##   Hispanic = col_double(),
##   White = col_double(),
##   Multi = col_double(),
##   Total = col_double(),
##   diverse = col_character(),
##   variance = col_double(),
##   int_group = col_character()
## )
```

```
attach(data)
```

Number of respondents in School year 1994-1995 and 2016-2017

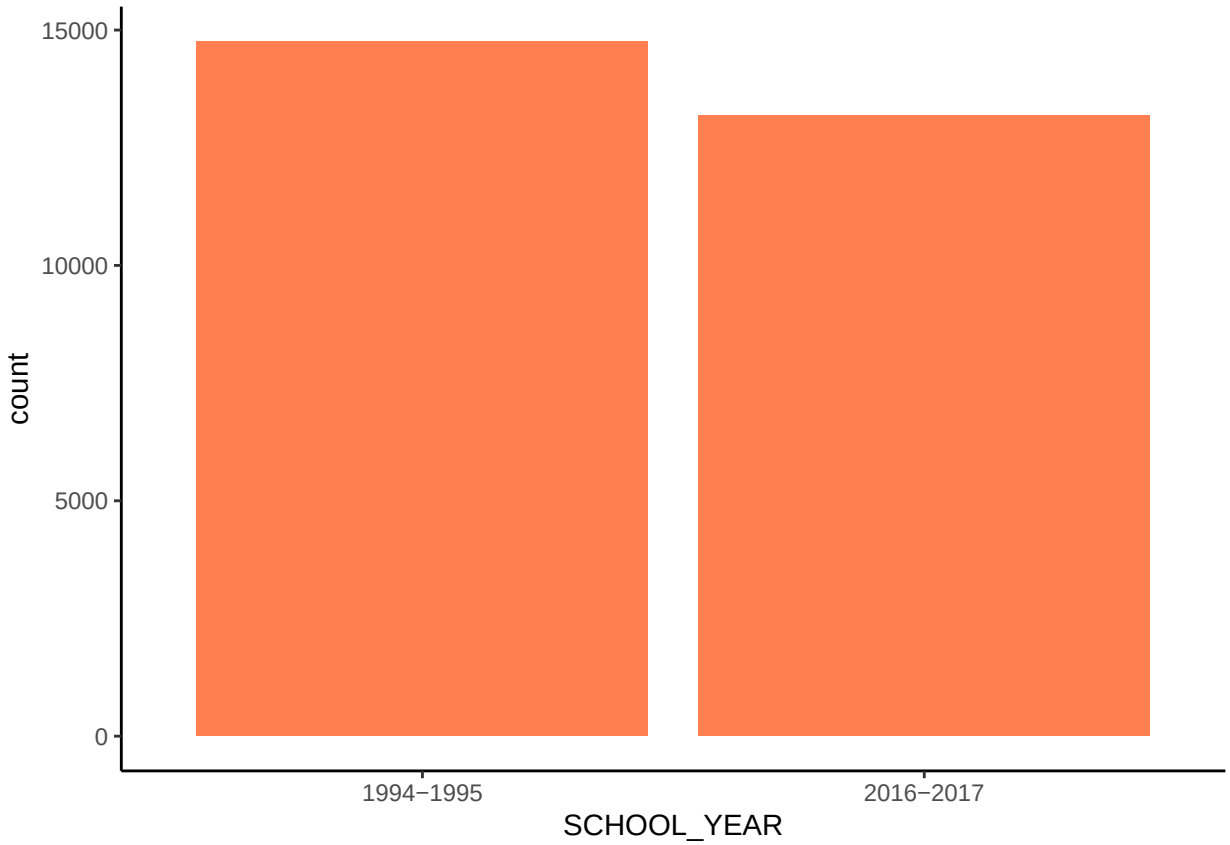


Figure 1

Number of respondents in each diversity group

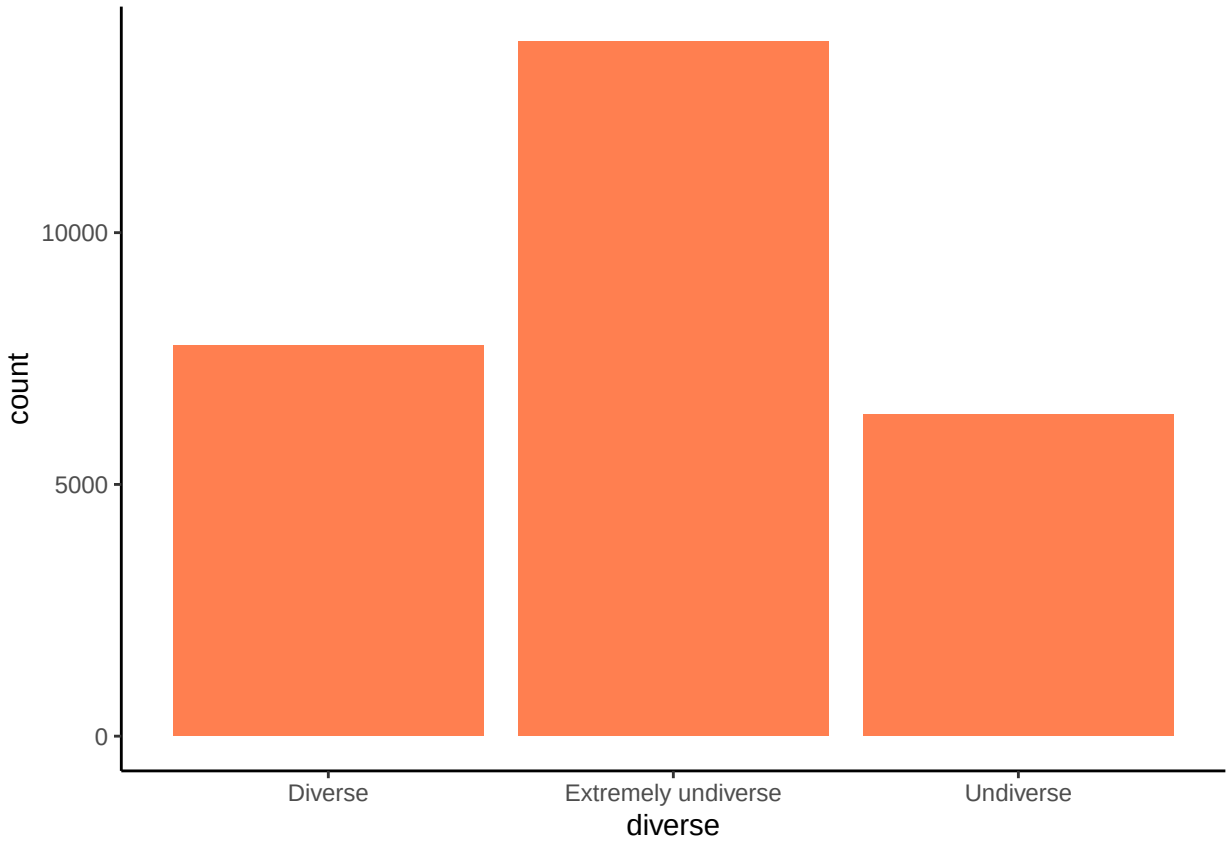


Figure 2

Diversity among two school year groups

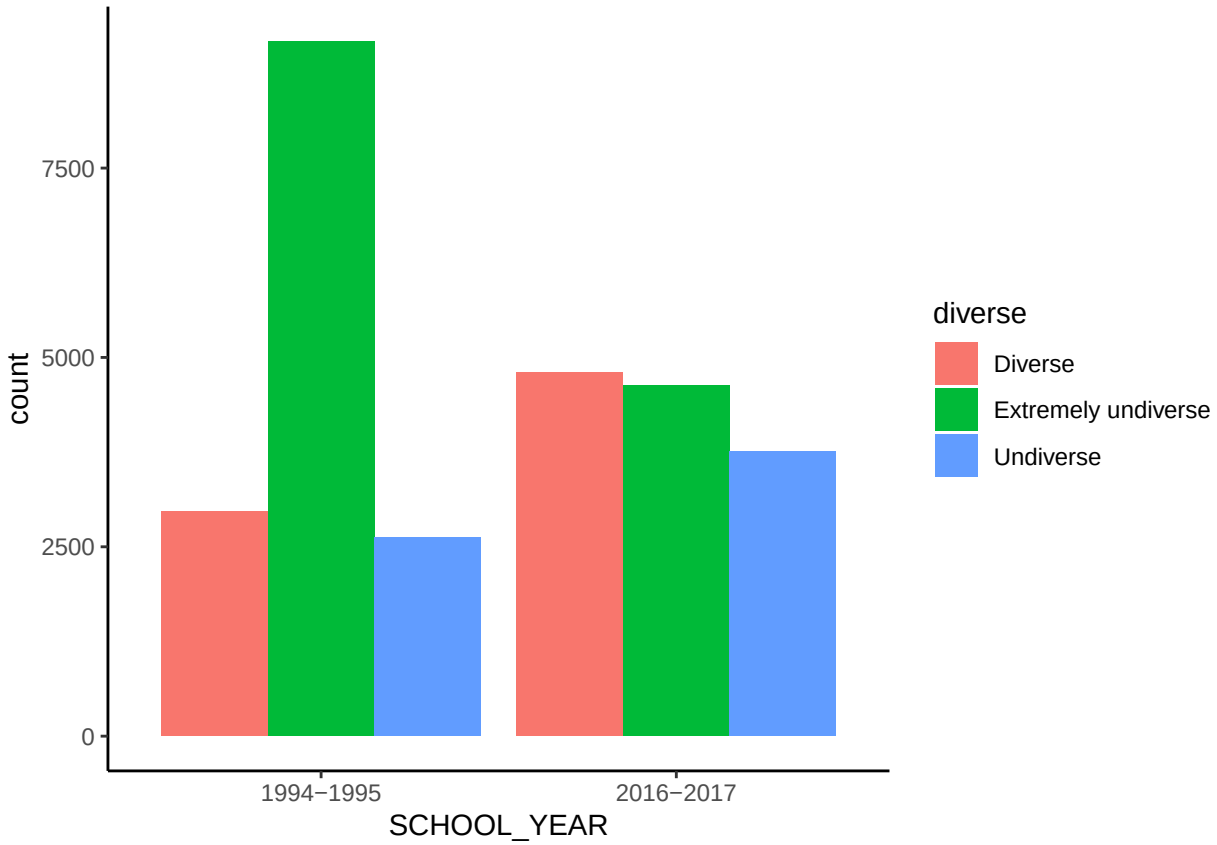


Figure 3

Type of school district among two school year groups

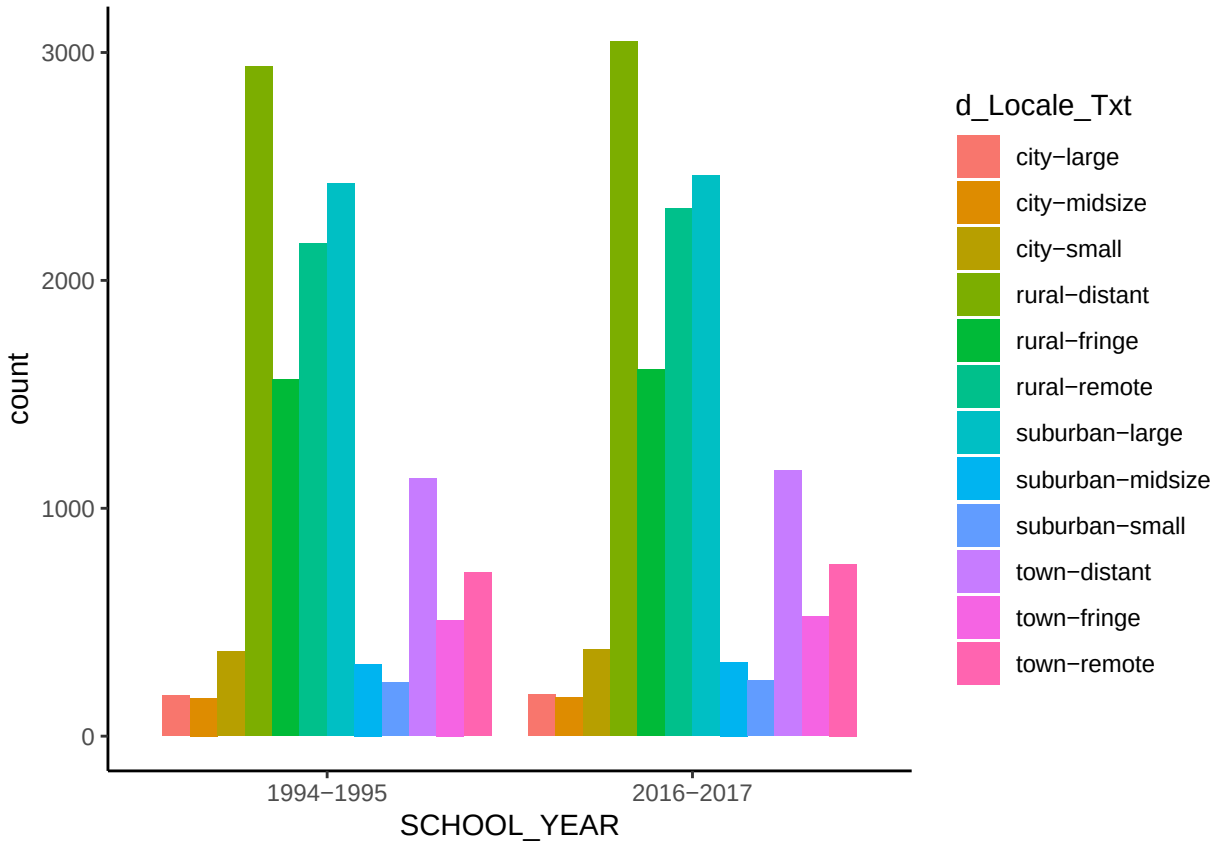


Figure 4

Type of school district among diversity groups

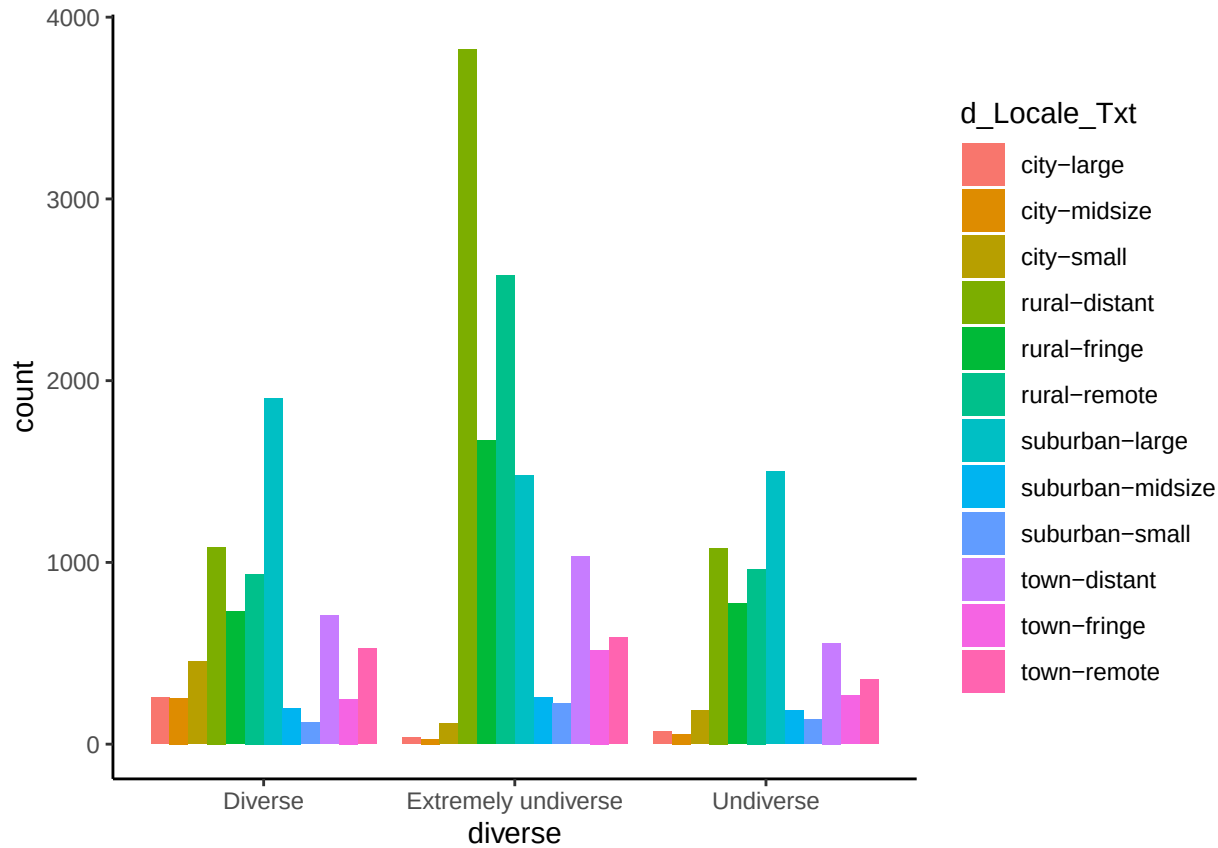


Figure 5

Group Summary Statistics of American Indian and Alaskan native proportion of student population

By School year:

```
## $`1994-1995`
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00000   0.00000   0.08416   2.73855   0.54054 100.00000
##
## $`2016-2017`
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00000   0.03113   0.22443   2.84303   0.68048 100.00000
```

Table 1

By School district:

```
## $`city-large`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1739  0.4194  1.4578  1.0524 67.8899
```

```
##
## $`city-midsize`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.006471 0.176131 0.364027 0.694799 0.738110 7.158314
##
## $`city-small`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1234  0.2740  0.9025  0.5928 73.0435
##
## $`rural-distant`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0000   0.1031   2.9849   0.6322 100.0000
##
## $`rural-fringe`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.1356  1.6648  0.4511 98.1132
##
## $`rural-remote`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.232   6.963   1.931 100.000
##
## $`suburban-large`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0306  0.1354  0.4063  0.3121 42.8482
##
## $`suburban-midsize`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00000  0.08283  0.23706  0.73284  0.65326 13.42657
##
## $`suburban-small`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00000  0.04129  0.18018  0.99201  0.48928 89.66942
##
## $`town-distant`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00000  0.04063  0.17655  1.80991  0.52673 88.62559
##
## $`town-fringe`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00000  0.03164  0.19015  1.11976  0.54391 37.50000
##
## $`town-remote`
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.1011   0.4337   4.8324   2.2478 100.0000
```

Table 2

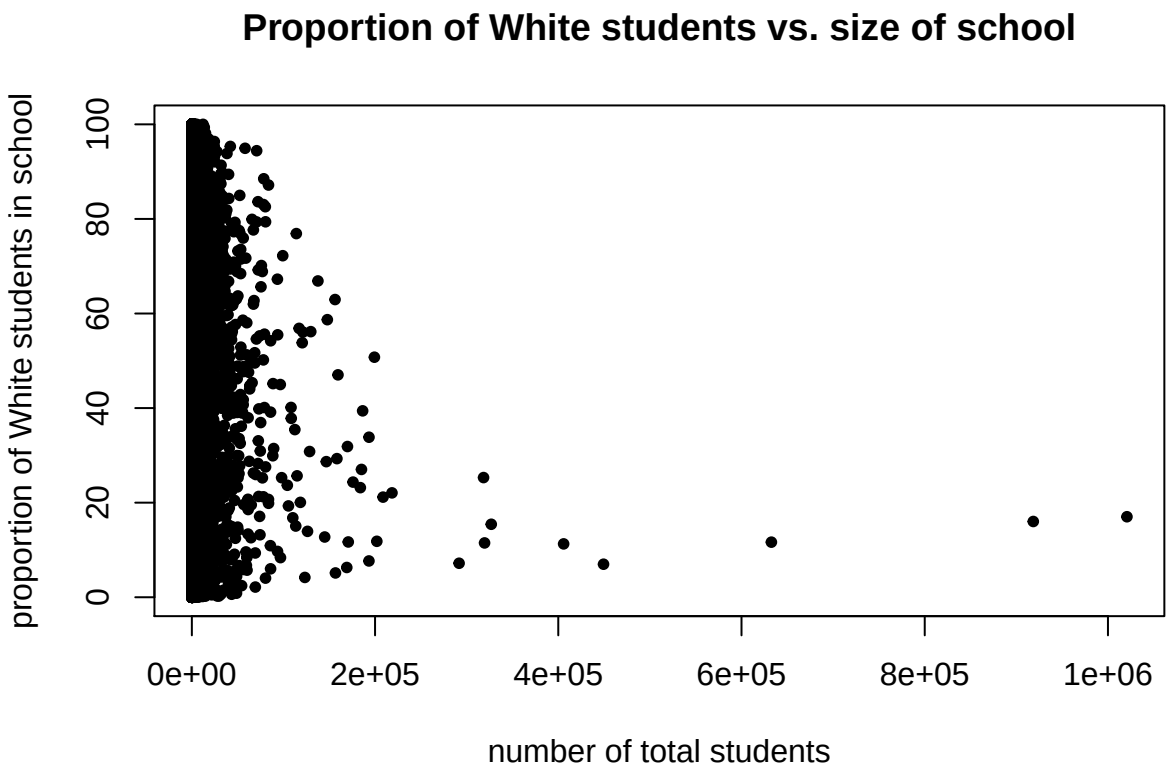
As per Figure 1 we can see that there were approximately equal number of students taken into the study for school years 1994-1995 and 2016-2017. But in figure 2 it is visible that the extremely un-diverse group is much higher than the other diversity groups. Further exploration was carried out to see this overall difference reflect in both school years. However,

as per figure 3, in school year 1994-1995, extremely un-diverse group is much higher than the other two diversity groups. Whereas in schools year 2016-2017, there is no significant difference among these diversity groups. Further exploration was performed to see if there is any difference in school district mix between two school years. As per figure 4 there isn't any significant difference of school district mix. Finally, a graph was plotted to identify any differences in school district mix by diversity groups. As per figure 5, there is a significant spike in rural districts within extremely un-diverse group than other two diversity groups. Apart from above graphical data exploration, further analysis was carried out for the variable: American Indian and Alaskan native proportion of student population. In particular to see if there is any difference in this population among groups such as School year and school district. As per Table 1, there seems to be no difference in this group proportion between two School years. However as per Table 2, there is a significant difference in this proportion among school district mix. In fact, the proportion gets larger moving from rural to urban districts.

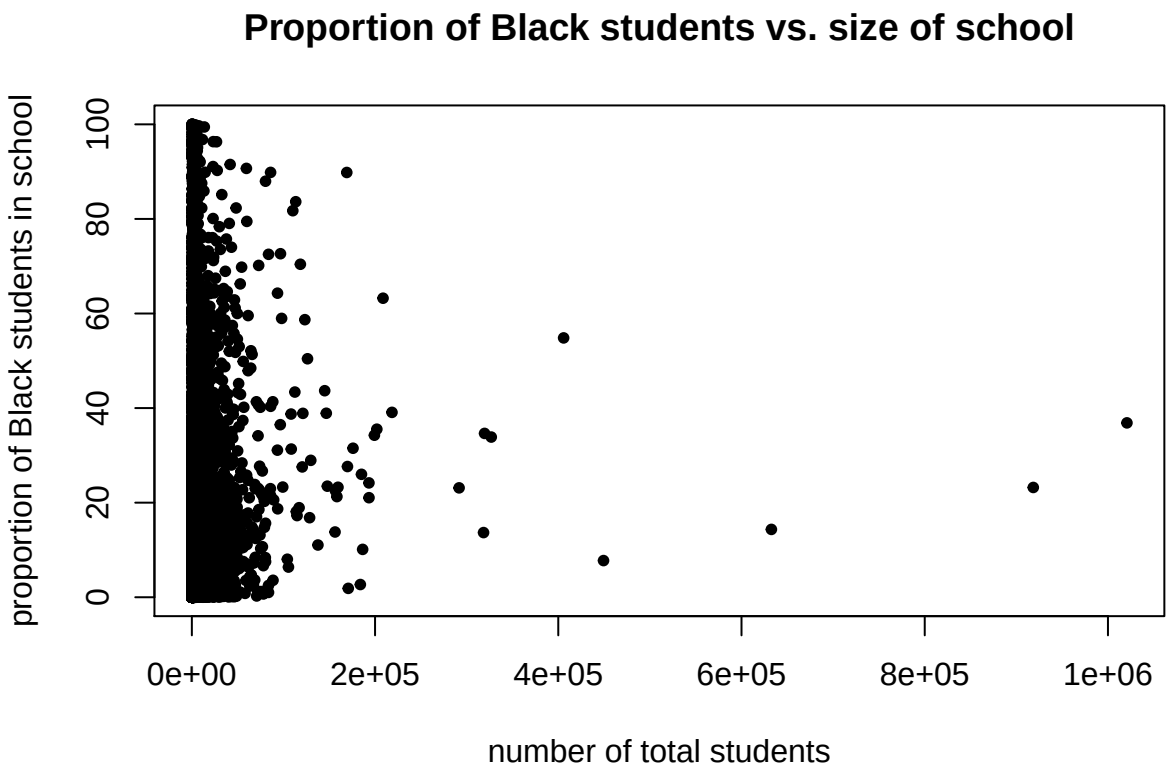
Relationship among vairables

In this part, we are trying to figure out the relationships between variables.First of all, we are trying to figure out if the school diversity is relation to the total amount of school students. Since the number of students of each school is numerical, we can directly draw the scatter plots. Here we draw the scatter plots between white and total, black and total and Asian and total:

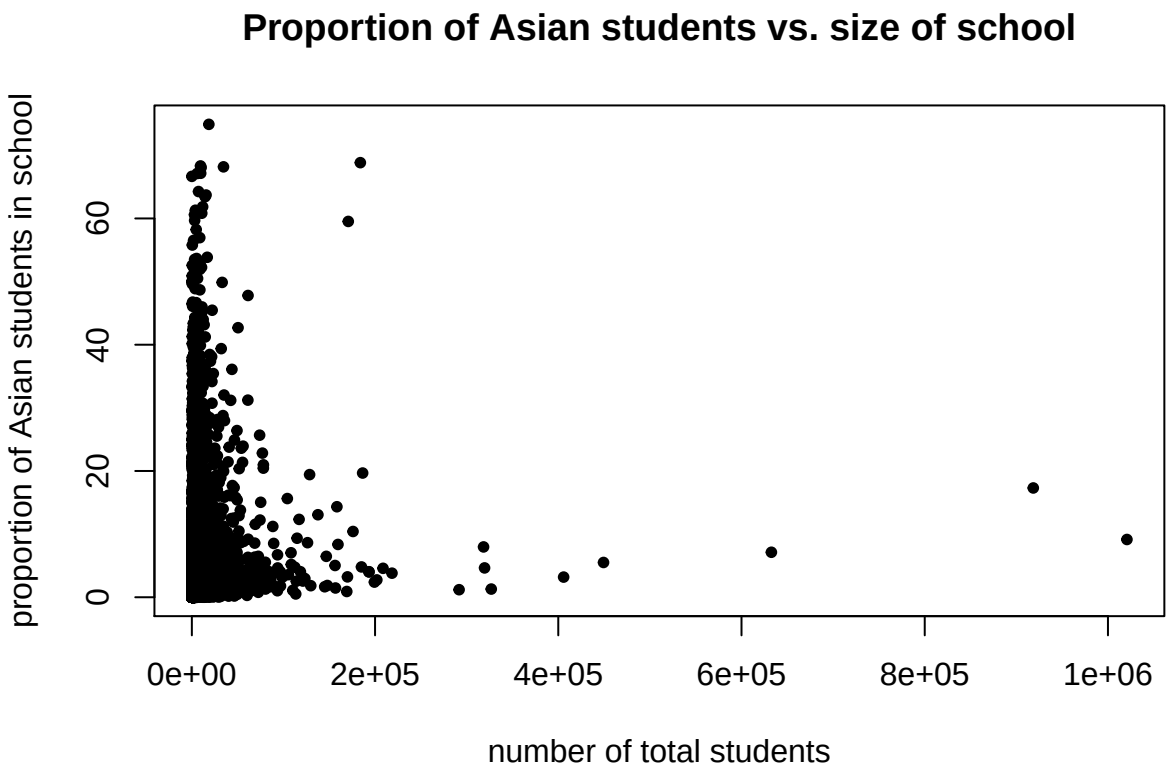
```
plot(Total,White,main="Proportion of White students vs. size of school",
     xlab="number of total students",
     ylab="proportion of White students in school",col="black",pch=20)
```




```
plot(Total,Black,main="Proportion of Black students vs. size of school",
     xlab="number of total students",
     ylab="proportion of Black students in school",col="black",pch=20)
```




```
plot(Total,Asian,main="Proportion of Asian students vs. size of school",
     xlab="number of total students",
     ylab="proportion of Asian students in school",col="black",pch=20)
```



As shown from the three diagrams, we can tell that there is no relationship in either White proportion and school size nor Black proportion and school size. So we can conclude that school diversity has no relationship with school size.

Furthermore, we want to know the relationship between time and school diversity. In this case, we try to split dataset within different subsets by the school year 1994-1995 and school year 2016-2017. Then, we use the difference between the two school years to say if there is a relationship.

```
sd9495 = subset(data ,SCHOOL_YEAR == "1994-1995")
mean(White)
```

```
## [1] 76.9282
```

```
mean(Black)
```

```
## [1] 6.551511
```

```
mean(Asian)
```

```
## [1] 1.769562
```

Here we get the average proportion for white, black and Asian students in 1994-1995 school year.

```
sd9495 = subset(data, SCHOOL_YEAR == "1994-1995")
mean(White)
```

```
## [1] 76.9282
```

```
mean(Black)
```

```
## [1] 6.551511
```

```
mean(Asian)
```

```
## [1] 1.769562
```

Compare the two sets of mean, we can see that the average proportion of white students is getting smaller while both black and Asian students are increasing. Hence, we can guess that the school diversity has a positive relation with time. We can try to prove it in next part.

Now we are finding the relationship between the size of city and the diversity of the univeristies.

Firstly, we find the proportion of white students in different city-types and find the biggest proportion.

```
sdcity_large = subset(data,d_Locale_Txt == "city-large")
mean(White)
```

```
## [1] 76.9282
```

```
mean(Black)
```

```
## [1] 6.551511
```

```
mean(Asian)
```

```
## [1] 1.769562
```

```
sdcity_midsize = subset(data,d_Locale_Txt == "city-midsize")
mean(White)
```

```
## [1] 76.9282
```

```
mean(Black)
```

```
## [1] 6.551511
```

```
mean(Asian)
```

```
## [1] 1.769562
```

```
sdcity_small = subset(data,d_Locale_Txt == "city-small")
mean(White)
```

```
## [1] 76.9282
```

```r
mean(Black)
```

```
## [1] 6.551511
```

```r
mean(Asian)
```

```
## [1] 1.769562
```

In cities, "city-small" has the largest proportion of white student. In this case, we can conclude that the school in the bigger city have more diversiform which means that the bigger proportion of the black students and the Asian students.

And then we are going to find the proportion of white students in the different rural-types, then we can find which type of rural has the biggest proportion.

```r
sdrural_distant = subset(data,d_Locale_Txt == "rural-distant")
mean(White)
```

```
## [1] 76.9282
```

```r
mean(Black)
```

```
## [1] 6.551511
```

```r
mean(Asian)
```

```
## [1] 1.769562
```

```r
sdrural_fringe = subset(data,d_Locale_Txt == "rural-fringe")
mean(White)
```

```
## [1] 76.9282
```

```r
mean(Black)
```

```
## [1] 6.551511
```

```r
mean(Asian)
```

```
## [1] 1.769562
```

```r
sdrural_remote = subset(data,d_Locale_Txt == "rural-remote")
mean(White)
```

```
## [1] 76.9282
```

```r
mean(Black)
```

```
## [1] 6.551511
```

```r
mean(Asian)
```

```
## [1] 1.769562
```

In rurals, "rural-fringe" has the biggest proportion of white students. But the difference between "rural-fringe" and "rural-remote" are almost similar (only 0.5%), so we can infer that rural-remote/fringe has the bigger proportion of white student. On the another hand, we can conclude that the bigger type of rural the more diversiform.

Afterwards, we are finding the biggest proportoin of white students in the different types of suburban. The biggest proportion of white student means the less diversiform in the suburban.

```
sdsuburban_large = subset(data,d_Locale_Txt == "suburban-large")
mean(White)
```

```
## [1] 76.9282
```

```
mean(Black)
```

```
## [1] 6.551511
```

```
mean(Asian)
```

```
## [1] 1.769562
```

```
sdsuburban_midsize = subset(data,d_Locale_Txt == "suburban-midsize")
mean(White)
```

```
## [1] 76.9282
```

```
mean(Black)
```

```
## [1] 6.551511
```

```
mean(Asian)
```

```
## [1] 1.769562
```

```
sdsuburban_small = subset(data,d_Locale_Txt == "suburban-small")
mean(White)
```

```
## [1] 76.9282
```

```
mean(Black)
```

```
## [1] 6.551511
```

```
mean(Asian)
```

```
## [1] 1.769562
```

In suburban, the biggest proportion of white students in the different types of suburban. This means the bigger suburban the less diversiform.

Then, we are going to find the biggest proportion of the white student in the different of towns. The biggest proportion in the type of town means it has the lowest diversiform.

```r
sdtown_distant = subset(data,d_Locale_Txt == "town-distant")
mean(White)
```

```
## [1] 76.9282
```

```r
mean(Black)
```

```
## [1] 6.551511
```

```r
mean(Asian)
```

```
## [1] 1.769562
```

```r
sdtown_fringe = subset(data,d_Locale_Txt == "town_fringe")
mean(White, na.rm=TRUE)
```

```
## [1] 76.9282
```

```r
mean(Black, na.rm=TRUE)
```

```
## [1] 6.551511
```

```r
mean(Asian, na.rm=TRUE)
```

```
## [1] 1.769562
```

```r
sdtown_remote = subset(data,d_Locale_Txt == "town-remote")
mean(White)
```

```
## [1] 76.9282
```

```r
mean(Black)
```

```
## [1] 6.551511
```

```r
mean(Asian)
```

```
## [1] 1.769562
```

The biggest proportion of white students is "town-distance".

Overall, we know that the bigger proportion of the non-white students, more diversiform the unviersities are. After we serched all the types of cities, we can conclude that more diversiform of the universities' students in the bigger cities rather than the smaller one.

Construction and Interpretation of Confidence Intervals

Confidence Interval for the population mean of American Indian and Alaskan native proportion of student population.

```
##
##  One Sample t-test
##
## data:  data$AIAN
```

```
## t = 41.069, df = 27943, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.654793 2.920899
## sample estimates:
## mean of x
##  2.787846
```

Output 1

Confidence Intervals for population means of difference student group proportions

```
##
##  One Sample t-test
##
## data:  data$Asian
## t = 66.997, df = 27943, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.717792 1.821332
## sample estimates:
## mean of x
##  1.769562
```

Output 2

```r
t.test(data$Black, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  data$Black
## t = 72.052, df = 27943, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.373288 6.729734
## sample estimates:
## mean of x
##  6.551511
```

Output3

```r
t.test(data$Hispanic, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  data$Hispanic
## t = 94.211, df = 27943, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  10.23453 10.66944
## sample estimates:
## mean of x
##  10.45199
```

Output4

```r
t.test(data$White, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  data$White
## t = 486.99, df = 27943, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  76.61858 77.23782
## sample estimates:
## mean of x
##   76.9282
```

Output5

```r
t.test(data$Multi, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  data$Multi
## t = 101.7, df = 13183, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3.140688 3.264127
## sample estimates:
## mean of x
##  3.202408
```

Output6

In Output 1 we are 95% confident that the population mean of American Indian and Alaskan native proportion of student population lies between 2.65 and 2.92. Likewise, as per the above outputs from Output 2 to Output 6, below confidence for their proportion of student populations were drawn with a 95% confidence,

Asian:

[1.71, 1.82]

Black:

[**6.37, 6.72**]

Hispanic:

[**10.23, 10.66**]

White:

[**76.61, 77.23**]

Multi-ethnic:

[**3.14, 3.26**]

As per the above intervals we can clearly see that there is significant difference in White proportion with compared to other groups.

As per above Outputs we can see that the groups: has the lowest average proportions. Even though their confidence interval does not overlap, further analysis was carried out to see if there is a difference between Multi-ethnic and Asian groups.

```
##
##  One Sample t-test
##
## data:  data$d1
## t = 18.069, df = 13183, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8526861 1.0601921
## sample estimates:
## mean of x
## 0.9564391
```

Output 7

As per above Output 7 we can see that the difference in mean proportion between Asian and Multi-ethnic groups lies between 0.85 and 1.06 with 95% confidence. Hence, we can conclude that there is a significant difference between two groups since the interval does not contain 0 in it.

School Diversity Inference Report

TUT9101, Group 6: Mohammad Fahim Uddin Alvi, Srabon Ratan, Chunyuan Wang, Keying Rao, Wenyi Wang, Yubin Yao

Load the data set from Github repository

```
#load required libraries
library(tidyr)
library(dplyr)
library(ggplot2)
library(gridExtra)

#laod data set
school_diversity <- read.csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/
```

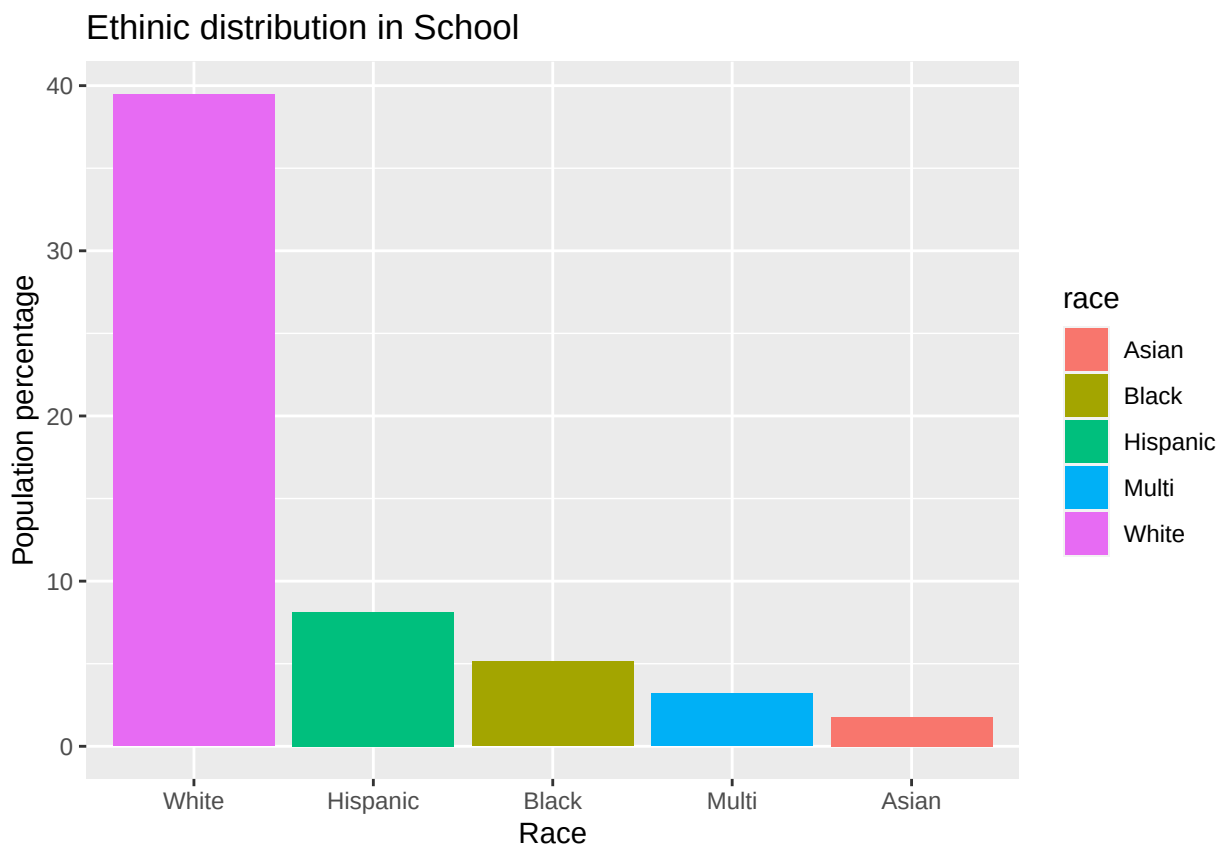
Prepare data

```
#Drop, which populations are less than 100
data <- filter(school_diversity, Total >=500)

#gather different races in one column
data <- gather(school_diversity, "race", "rate", 7:11)

#drop the NA values
data <- filter(data,rate !="NA" )

#remove outliers from their populaton rate
Q <- quantile(data$rate , probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(data$rate )
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range

data <- subset(data , data$rate > (Q[1] - 1.5*iqr) & data$rate  < (Q[2]+1.5*iqr))
```

Distyribution of different Ethinic students in School

```
data %>% group_by(race) %>%
  summarise(mean = mean(rate)) %>%
  arrange(desc(mean))  %>%
  ggplot(aes(reorder(race,-mean),mean, fill= race)) +geom_histogram(stat = "identity")+
  labs(x= "Race",title = "Ethinic distribution in School", y= "Population percentage" )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

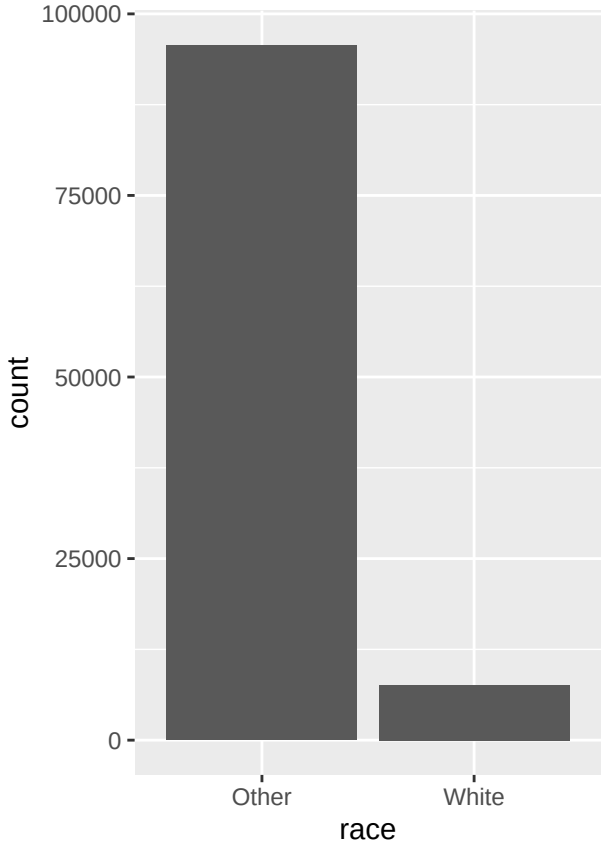


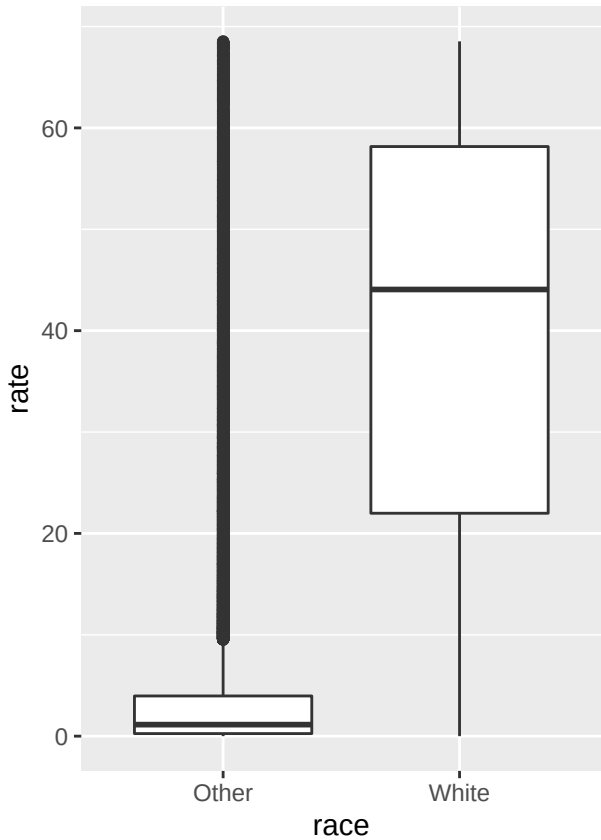
Adding other races to compare population of White students with all others

```
data$race[data$race != "White"] <- "Other"


p1 <- ggplot(data,aes(race)) +geom_histogram(stat = "count" )
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
p2 <- ggplot(data,aes(race,rate)) +geom_boxplot( )
grid.arrange(p1,p2,nrow=1)
```





Statistical Hypothesis

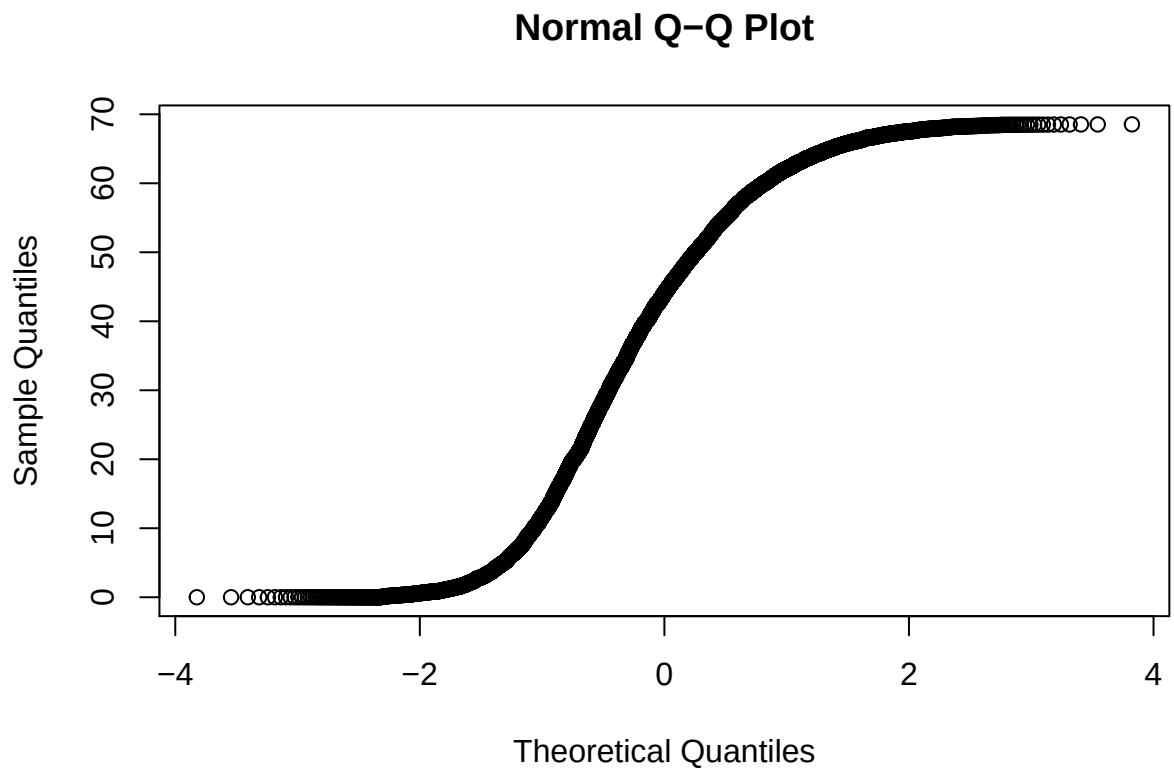
Null Hypothesis H0: Population of white students is greater than population of all other races.

Alternative Hypothesis Ha: Population of white students is not greater than population of all other races.

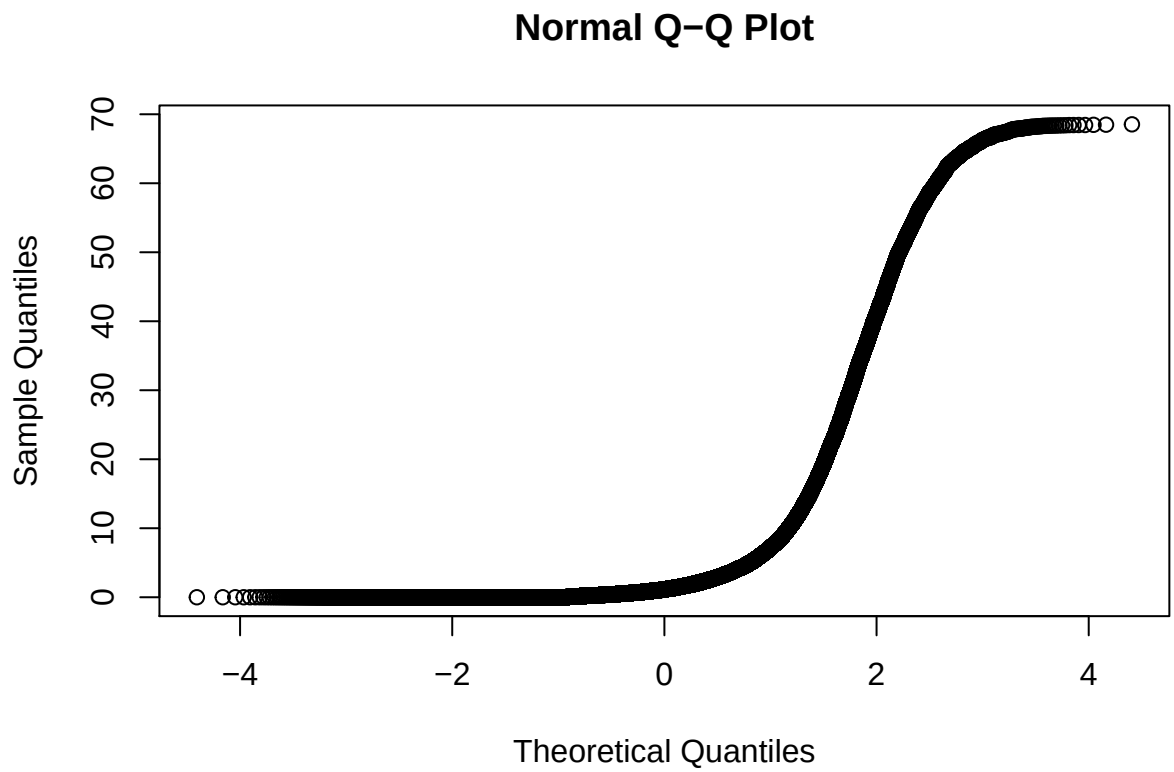
checking for assumptions

The assumption for the test is that both groups are sampled from normal distributions with equal variances.

```
#sample 1
qqnorm(data$rate[data$race == "White"])
```




```r
#sample 2
qqnorm(data$rate[data$race == "Other"])
```



t-Test

```
t.test(rate ~ race,data, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  rate by race
## t = -142.39, df = 7870.1, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -35.14366       Inf
## sample estimates:
## mean in group Other mean in group White
##            4.731364           39.473646
```

T statistic value

T statistic value of the test is -142.39

p value

P value of the test is 1

Decision

As p value is far greater than 0.05 so we are unable to reject the null hypothesis.

Conclusion

Population of the white students is greater than the population of the student which belong to the other ethnic groups.