How are education resources distributed in Australia (city vs rural and public vs private) and what effect does this have on mathematics literacy?

Abstract

In this research, I want to explore whether there is a discrepancy among scores in mathematic literacy within Australia, and how education resource(say, shortage of math teacher) will impact the assessment result. For instance, comparing the education resources between remote and city areas and the discrepancies between private or public schools. Based on the PISA2012 mathematics literacy survey data using statistical analysis like linear regression model and ANOVA test. Also, I will try to break down the relationship between school location and how it affects the scores in math literacy, the relationship between private or public schools and how it relates to the students' scores in math literacy.

Introduction

- PISA2012: the term PISA stands for Programme for International Student Assessment is a survey that randomly select 15 year-old students(students who are attending secondary schools) as samples for assessment. In 2012, a total of 65 OECD countries and economics and about half a million 15 year-old students participated in the PISA assessment(*PISA 2012 Results in Focus* 2014). Generally, the assessed reaults lies in 5 categories: Level 5 are high performers, whereas students who lies below the international standard baseline level 2 are cosidered low performers. In 2012, 775 Australian schools and 14,481 students participated in this assessment(Thomson and Buckley 2013). To ensure the authenticity, an amount of indigenous students were also sampled.
- Students are categorized in 6 different levels in accordance to their proficiency in the assessment(Thomson and Buckley 2013):
- Level 6: Students who score higher than 669.3 scores belongs to level6.
- Level 5: Students who score higher than 607.0 scores belongs to level5.
- Level 4: Students who score higher than 544.7 scores belongs to level4.
- Level 3: Students who score higher than 482.4 scores belongs to level3.
- Level 2: Students who score higher than 420.1 scores belongs to level2.
- Level 1: Students who score higher than 357.8 scores belongs to level1.
- Below 1: not demonstrate even the most basic types of mathematical literacy that PISA measures. These students are likely to be seriously disadvantaged in their lives beyond school.
- Mathematic Literacy: in the mathematic literacy domain, the assessment focused on students' ability to solve mathematic problems described in a real-life situation. In PISA2012 framework, it defined mathematic literacy as follows: "Mathematic literacy is an individual's capacity to formulate, employ and interpret mathematics in a variety of contexts. It includes reasoning mathematically nd using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to recognise the role that Mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens."

The assessment in math literacy is designed according three main components:

1. the context of a challenge or problem that arises in the real world
2. the nature of mathematical thought and action that can be used to solve the problem
3. the processes that the problem solver can use to construct a solution.

Overview of Australian Education System:

–Terms and definition

It is defined in the handbook from PISA(*SCHOOL Questionnaire for Pisa 2012* 2011) that the size of school is defined as : Village School or Rural Area: less than 3,000 people Small Town School: 3,000 to about 15,000 people Town School: 15,000 to about 100,000 people City School: 100,000 to about 1,000,000 people, for example, Hobart, Tasmania Large City School: ith over 1,000,000 people, for example, Sydney and Melbourne

–Facts of Australian Education(Halsey 2017), ("The Australian Education System | Australian Education Technology" 2017)

1. In 2016, there are over 9,400 schools in which 1400 of them are secondary schools across Australia. Which means more than half of the amount of secondary schools were participants of PISA2012 assessment.
2. In 2016, there are over 6,000 government(public) schools, more than 1700 Catholic schools(private) and more than 1,000 independent schools(private)
3. In 2017, around 65% of students attended government school, 19% students attended Catholic schools and 16% students attended independent schools.
4. In 2014, the proportion of residents aged between 25~34 years old who has a degree was: Major City 42.2 %, Inner Regional 21.8%, Outer Regional 19.5%, and Remote and Very Remote 17.8%.

Methods Used

– Libraries used

1. ggplot2: for plot generation (Wickham 2009)
2. PISA2012lite: provides original data set (Biecek, n.d.)
3. lme4: for linear regression analysis (Bates et al. 2015)
4. magrittr: for specific functions(Bache and Wickham 2014)
5. data.table: enables faster and efficient data handling (Dowle and Srinivasan 2017)
6. dplyr: to enable specific function for example, "select" function (Wickham et al. 2017)

– PISA2012 data

In this analysis, the original dataset is provided by library (Biecek, n.d.) The PISA2012lite dataset contains 10 data tables, including the survey result of school questionnaire and parent questionnaire, plus the student's questionnaire and assesment result. Including 775 Australian schools and 14,481 students participated. Indegi I used the 5 plausible values(PV1Math~PV5MATH) for as the measure of a student's mathematical literacy. The minimum score in this assessment is 0 and maximum is 1,000. For reliability, smaller states and indigenous students were oversampled in this assessment.

– Statistical Test Used:

Since the the data to be analyzed contained samples more than 30, so we will be running a F-test instead of a t-test. Before we start with the statistical testing, we have to make sure the data is normally distributed to allow further F-test.

Linear Regression Analysis: linear regression is one of the most popular statistical test, it's an approach for modeling the relationship between a dependent variable(Y) and one or multiple variables(X).

ANOVA Test: the term ANOVA test originally stands for Analysis of Variance which is used to analyze the differences between means among two test groups.

Tukey Test: Tukey test is often used with ANOVA test, it allows researchers to compare the means of all possible pairs.

– Hypothesis:

H0(null hypothesis): the variables mentioned is not going to affect the students' test result on PISA 2012 mathematical literacy.

H1(alternative hypothesis): school location, education resource and whether or not the school is privately owned or publicly owned will affect the students' performance on mathematical literacy.

Results

```
library(data.table)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

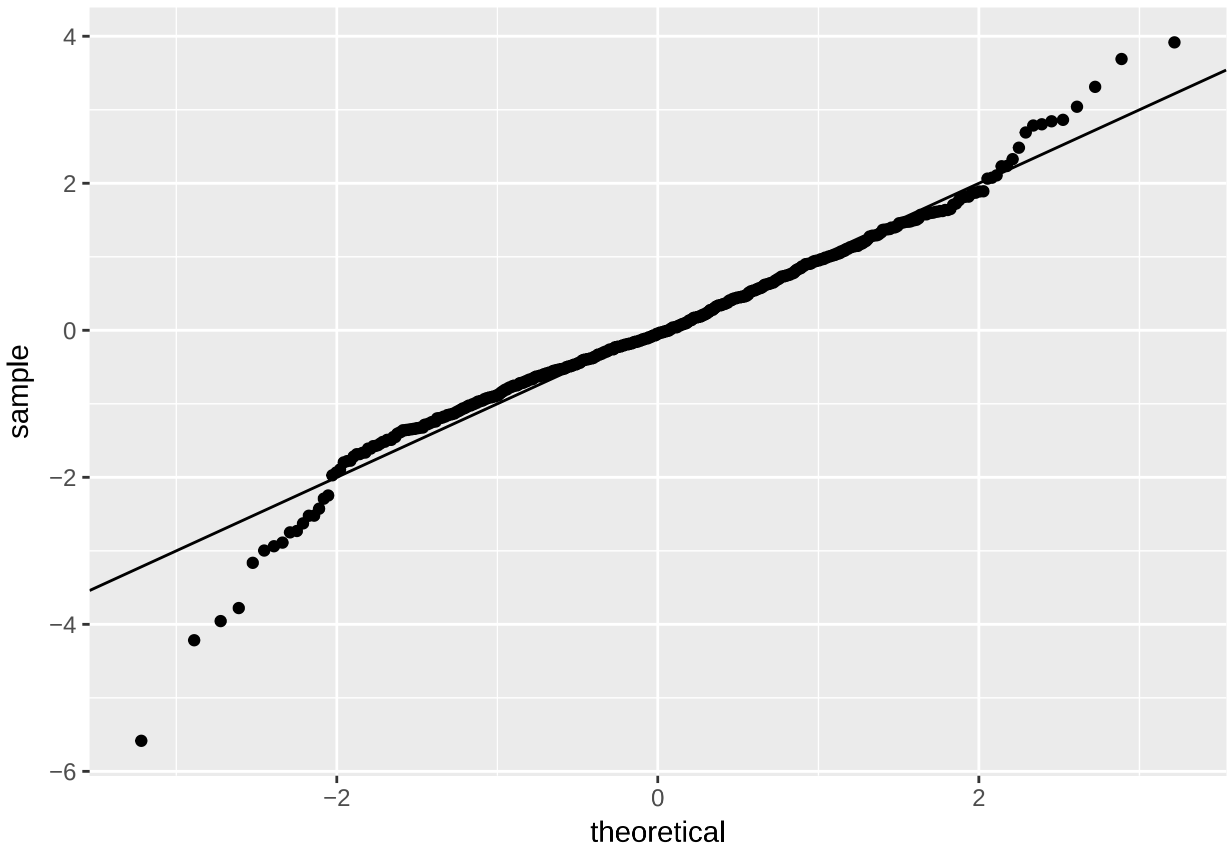
```
library(ggplot2)
library(PISA2012lite)
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(magrittr)
data("school2012")
setDT(school2012)
data("computerStudent2012")
setDT(computerStudent2012)

#calculate the weighted average of the Math Literacy performance using PV1MATH to PV5MATH
pv_cols <- paste0("PV", 1:5, "MATH")
student_data <- computerStudent2012[NC == "Australia",
                                    lapply(.SD, weighted.mean, w = W_FSTUWT / sum(W_FSTUWT)),
                                    by = SCHOOLID, .SDcols = c(pv_cols, "ESCS") ][, .(SCHOOLID, ESCS,
                                                                       Mean_PVMATH = Reduce(`+`, .SD) /
                                                                       .SDcols = pv_cols]
#a summary table showing the overall mean math score
student_data[,  .(mu = mean(Mean_PVMATH, na.rm = T), sigma = sd(Mean_PVMATH, na.rm = T))]
```

```
##          mu    sigma
## 1: 495.5452 59.07824
```

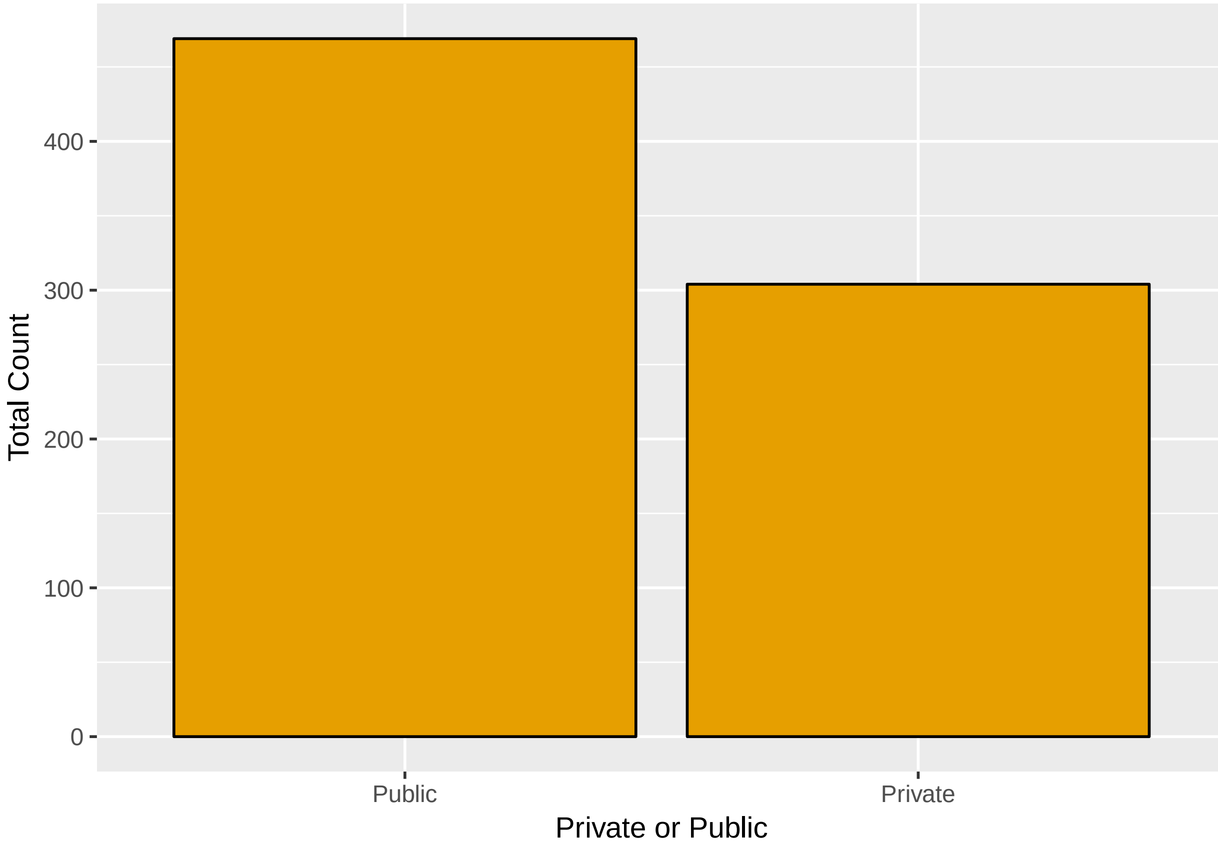
```r
#test whether the data is normally distributed
#' Put data onto a standard normal scale
#' @param x A numeric vector
#' @return A numeric vector of same length as `x`. This vector has mean 0 and sd 1.
standardise <- function(x) {
  (x - mean(x)) / sd(x)
}

#a plot test for normality
ggplot(student_data) + stat_qq(aes(sample = Mean_PVMATH %>% standardise)) + geom_abline()
```



As we can see from the plot above, the column Mean_PVMATH is normally distributed. Thereby, we can proceed to the statistical test we are going to perform in the following paragraph.

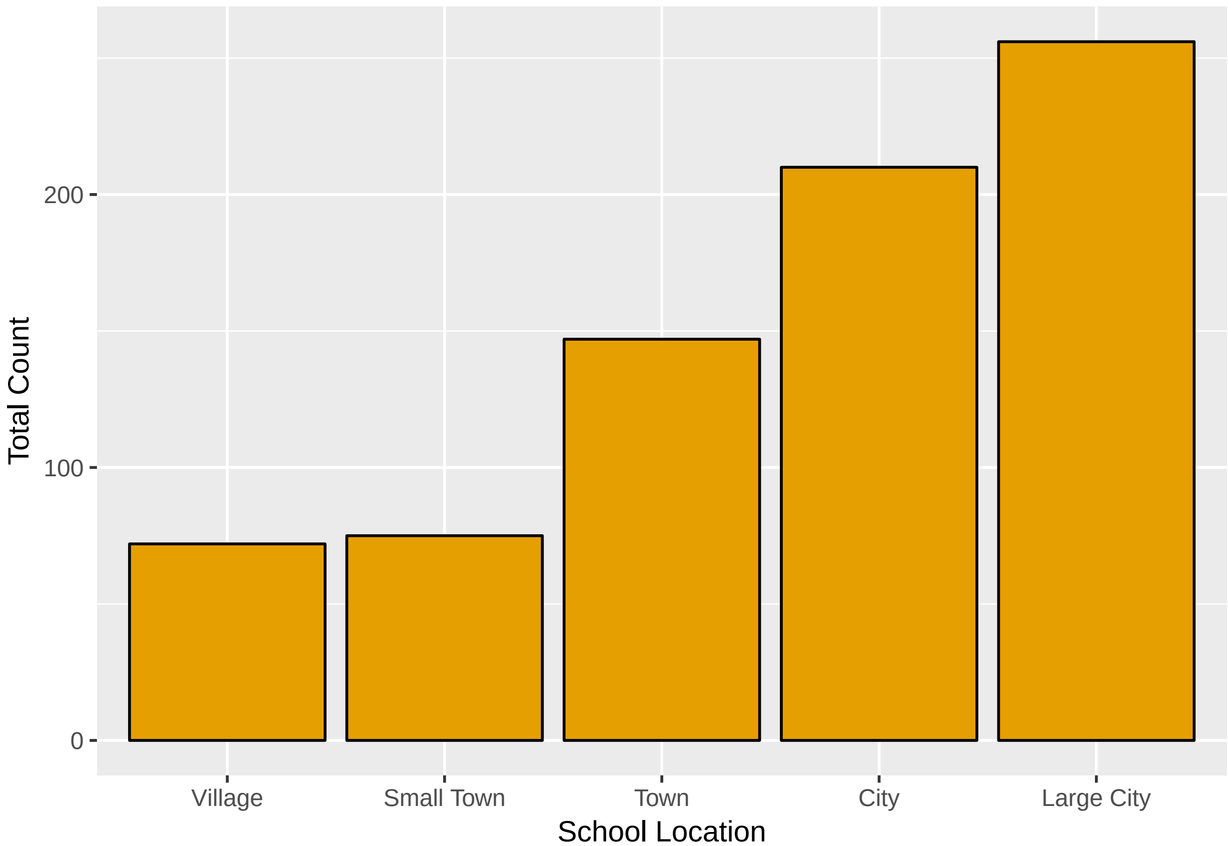
The following graph shows the amount of private and public schools in Australia that participated:

```r
#plot a bar chart showing the count of private and public schools participated
ggplot(school2012[NC == "Australia" & !is.na(SC01Q01)], aes(x=SC01Q01))+geom_bar(fill="#E69F00", colour=
```



The following is a graph about the distribution of schools in Australia that participated:

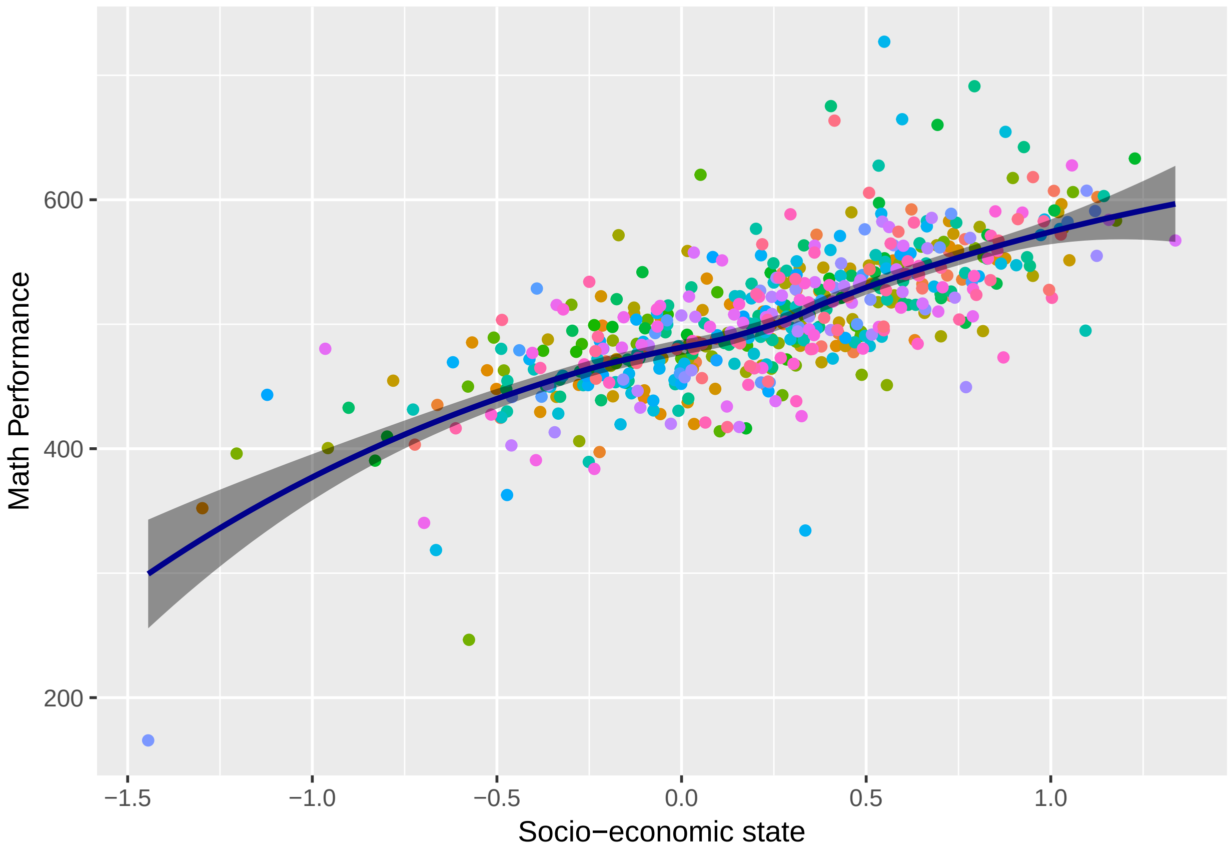
```
#plot bar chart showing the school location for schools participated:
ggplot(school2012[NC == "Australia" & !is.na(SC03Q01)], aes(SC03Q01))+geom_bar(fill="#E69F00", colour="
```



The following cosidering the socio-economic state of the student using Mean_PVMATH:

```
#a point plot showing the relationship between socio-economic state and math performance
ggplot(data = student_data, aes(x = ESCS, y = Mean_PVMATH)) +
  geom_point(aes(colour = SCHOOLID)) + geom_smooth(fill="black", colour="darkblue", size=1)+ theme(legen
```

```
## `geom_smooth()` using method = 'loess'

## Warning: Removed 234 rows containing non-finite values (stat_smooth).

## Warning: Removed 234 rows containing missing values (geom_point).
```

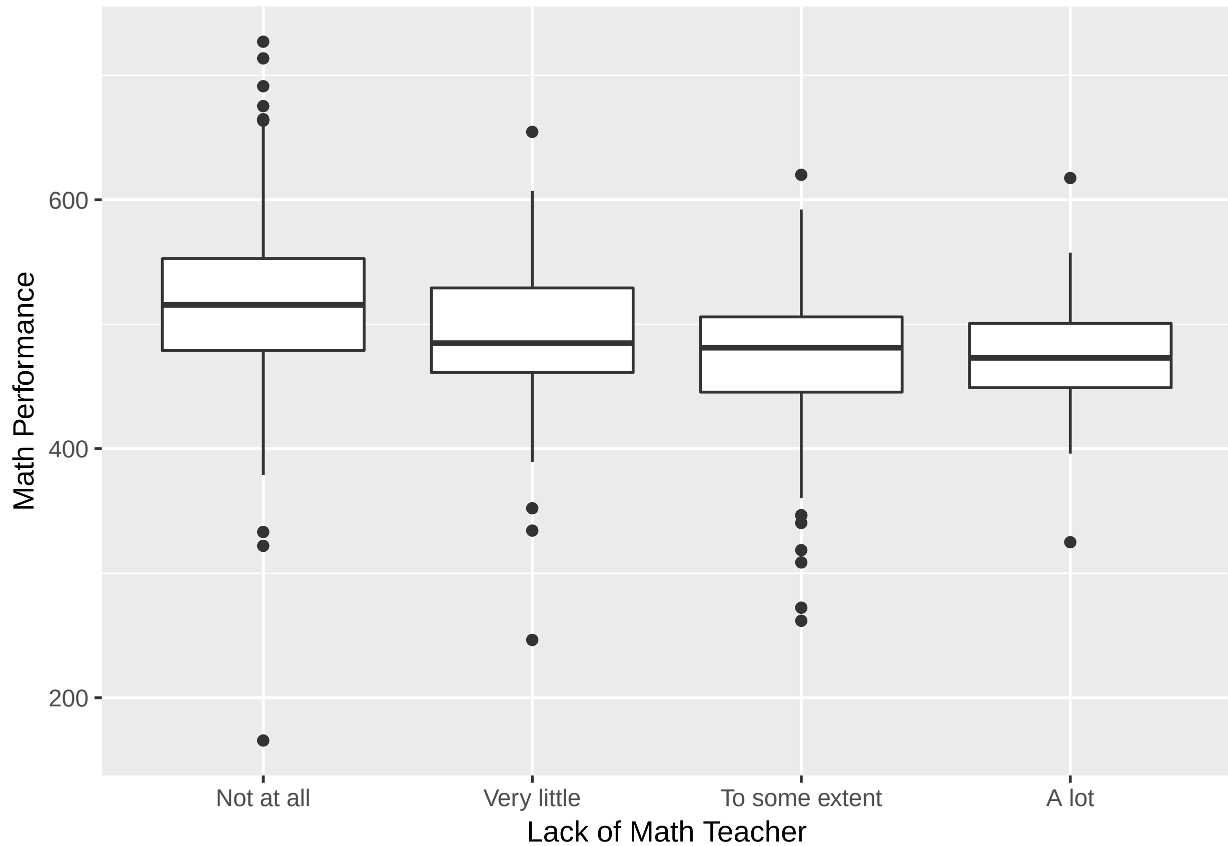


The graph shows a positive relationship between socio-economic state of a student and their math performance.

```
#create a view selecting country Australia and column SC14Q02 (Lack of math teacher)
school_view <- school2012[NC =="Australia" & !is.na(SC14Q02)]
school_plot_data <- select(school_view, SC14Q02, SCHOOLID)

#merge two tables into one table for plotting use
plot_data <- merge(student_data, school_plot_data, by="SCHOOLID")

#a simple summary table
plot_data[, .(mu = mean(Mean_PVMATH, na.rm = T),
              sigma = sd(Mean_PVMATH, na.rm = T)),
              by = SC14Q02]
```

```
##           SC14Q02       mu    sigma
## 1: To some extent 474.3121 51.97706
## 2:    Very little 491.2367 55.49166
## 3:     Not at all 514.8029 60.78728
## 4:          A lot 475.7340 47.41558
```

```
#create a box plot
ggplot(plot_data, aes(x=factor(SC14Q02), y=Mean_PVMATH))+geom_boxplot()+labs(x="Lack of Math Teacher", y
```




```r
#run linear regression test
fit <- lm(Mean_PVMATH~SC14Q02, data=plot_data)
summary(fit)
```

```
##
## Call:
## lm(formula = Mean_PVMATH ~ SC14Q02, data = plot_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -349.17  -31.84    0.88   36.44  212.15
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             514.803      3.072 167.606  < 2e-16 ***
## SC14Q02Very little      -23.566      5.424  -4.345 1.59e-05 ***
## SC14Q02To some extent   -40.491      4.917  -8.236 7.82e-16 ***
## SC14Q02A lot            -39.069      8.972  -4.354 1.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.55 on 757 degrees of freedom
## Multiple R-squared:  0.09157,    Adjusted R-squared:  0.08797
## F-statistic: 25.43 on 3 and 757 DF,  p-value: 1.101e-15
```

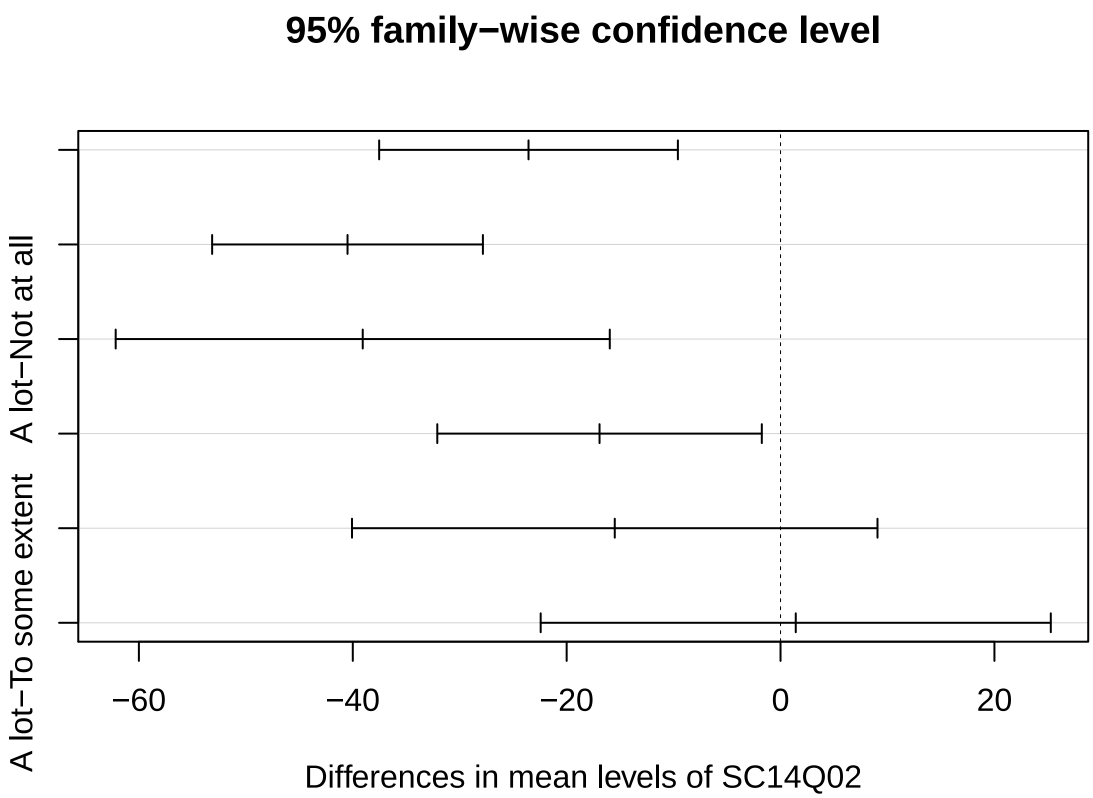
```r
#run ANOVA test
plot_anova <- aov(Mean_PVMATH~SC14Q02, data = plot_data)
summary(plot_anova)
```

```
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## SC14Q02       3  244035   81345   25.43 1.1e-15 ***
## Residuals   757 2421025    3198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#run Tukey test
tuk <- TukeyHSD(plot_anova)
tuk
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Mean_PVMATH ~ SC14Q02, data = plot_data)
##
## $SC14Q02
##                                  diff       lwr        upr     p adj
## Very little-Not at all     -23.566233 -37.53238  -9.600085 0.0000933
## To some extent-Not at all  -40.490861 -53.14970 -27.832021 0.0000000
## A lot-Not at all           -39.068966 -62.17076 -15.967172 0.0000894
## To some extent-Very little -16.924628 -32.09747  -1.751783 0.0217653
## A lot-Very little          -15.502733 -40.07227   9.066799 0.3653172
## A lot-To some extent         1.421895 -22.42878  25.272572 0.9987112
```

```
plot(tuk)
```



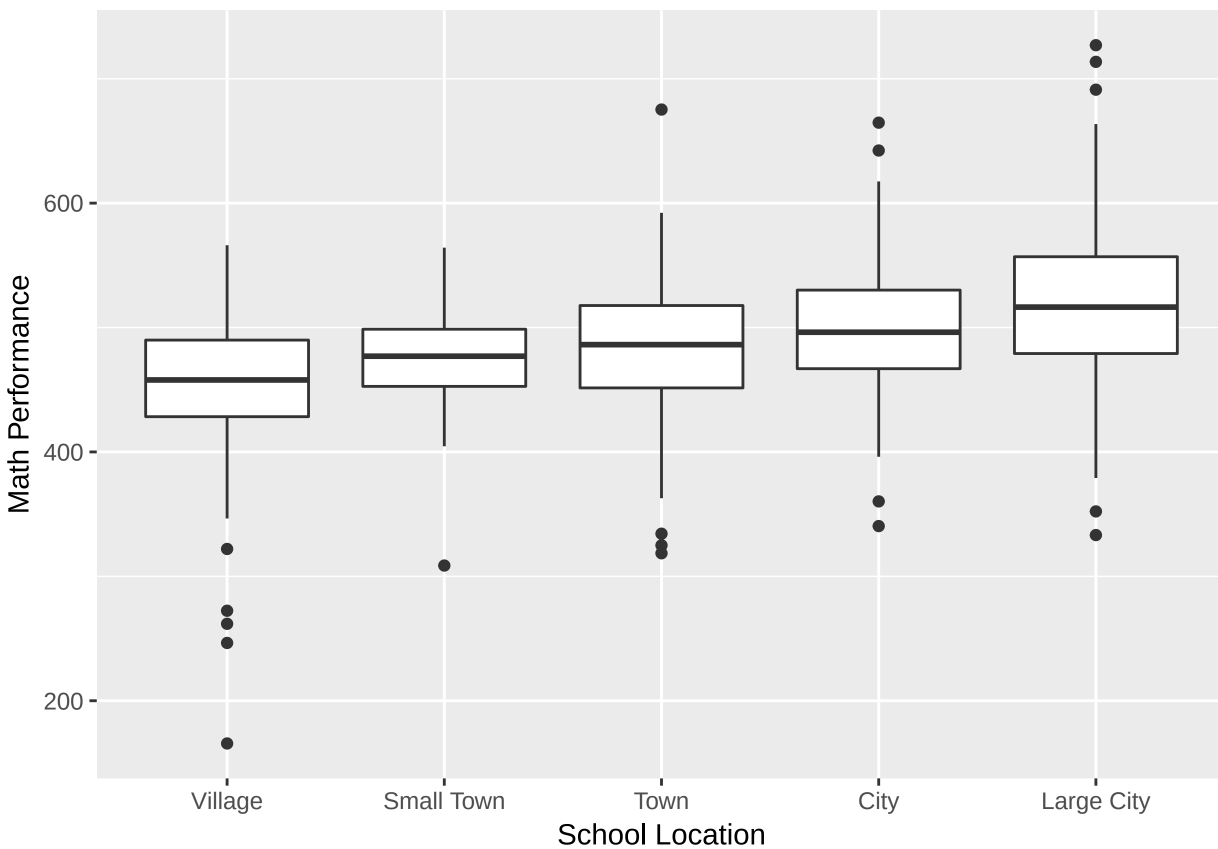
The boxplot infers that there is a difference in the mean score of students from different shortage situation of math teachers. And it impacts the students' math assessment result greatly.

Now we want to look into the relationship between schools and weighted math performances from PV1MATH to PV5MATH:

```r
#filter out the data from Australia and select column SC03Q01(School location) and SchoolID
school_view <- school2012[NC =="Australia" & !is.na(SC03Q01)]
school_plot_data <- select(school_view, SC03Q01, SCHOOLID)

#merge two tables into one table for plotting use
plot_data <- merge(student_data, school_plot_data, by="SCHOOLID")

#a simple summary table
plot_data[, .(mu = mean(Mean_PVMATH, na.rm = T),
              sigma = sd(Mean_PVMATH, na.rm = T)),
              by = SC03Q01]
```

```
##       SC03Q01       mu    sigma
## 1:       Town 484.0591 52.41388
## 2: Large City 518.5862 60.02301
## 3:       City 500.7306 49.46786
## 4:    Village 448.4095 69.54368
## 5: Small Town 472.2424 40.09360
```

```r
#merge two tables: student_data and school_plot_data and create a boxplot
ggplot(plot_data, aes(x=factor(SC03Q01), y=Mean_PVMATH))+geom_boxplot()+labs(x="School Location", y="Ma
```




```r
#run linear regression test
fit <- lm(Mean_PVMATH~SC03Q01, data=plot_data)
summary(fit)
```

```
##
```

```
## Call:
## lm(formula = Mean_PVMATH ~ SC03Q01, data = plot_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -282.773  -34.668    0.743   35.401  208.365
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         448.410      6.496  69.033  < 2e-16 ***
## SC03Q01Small Town    23.833      9.094   2.621  0.00895 **
## SC03Q01Town          35.650      7.928   4.497 7.99e-06 ***
## SC03Q01City          52.321      7.527   6.951 7.85e-12 ***
## SC03Q01Large City    70.177      7.352   9.545  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.12 on 755 degrees of freedom
## Multiple R-squared:  0.1362, Adjusted R-squared:  0.1316
## F-statistic: 29.76 on 4 and 755 DF,  p-value: < 2.2e-16
```

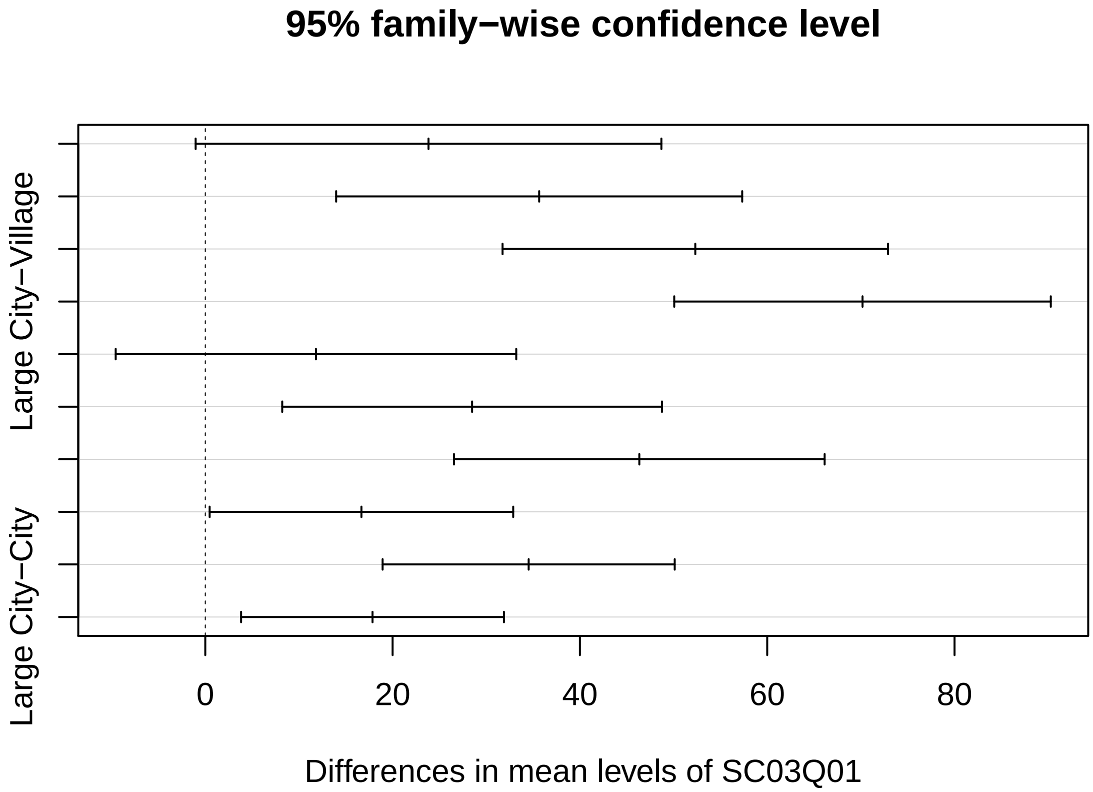
```
#run ANOVA test
plot_anova <- aov(Mean_PVMATH~SC03Q01, data = plot_data)
summary(plot_anova)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## SC03Q01       4  361608   90402   29.76 <2e-16 ***
## Residuals   755 2293569    3038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#run Tukey test
tuk <- TukeyHSD(plot_anova)
tuk
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Mean_PVMATH ~ SC03Q01, data = plot_data)
##
## $SC03Q01
##                             diff        lwr      upr     p adj
## Small Town-Village    23.83292 -1.0329231 48.69877 0.0676216
## Town-Village          35.64960 13.9706366 57.32857 0.0000781
## City-Village          52.32112 31.7389970 72.90324 0.0000000
## Large City-Village    70.17664 50.0721922 90.28109 0.0000000
## Town-Small Town       11.81668 -9.5692727 33.20264 0.5556866
## City-Small Town       28.48819  8.2149344 48.76145 0.0012437
## Large City-Small Town 46.34372 26.5555839 66.13185 0.0000000
## City-Town             16.67151  0.4643414 32.87868 0.0402473
## Large City-Town       34.52704 18.9309607 50.12311 0.0000000
## Large City-City       17.85552  3.8240010 31.88704 0.0048120
```

```
plot(tuk)
```

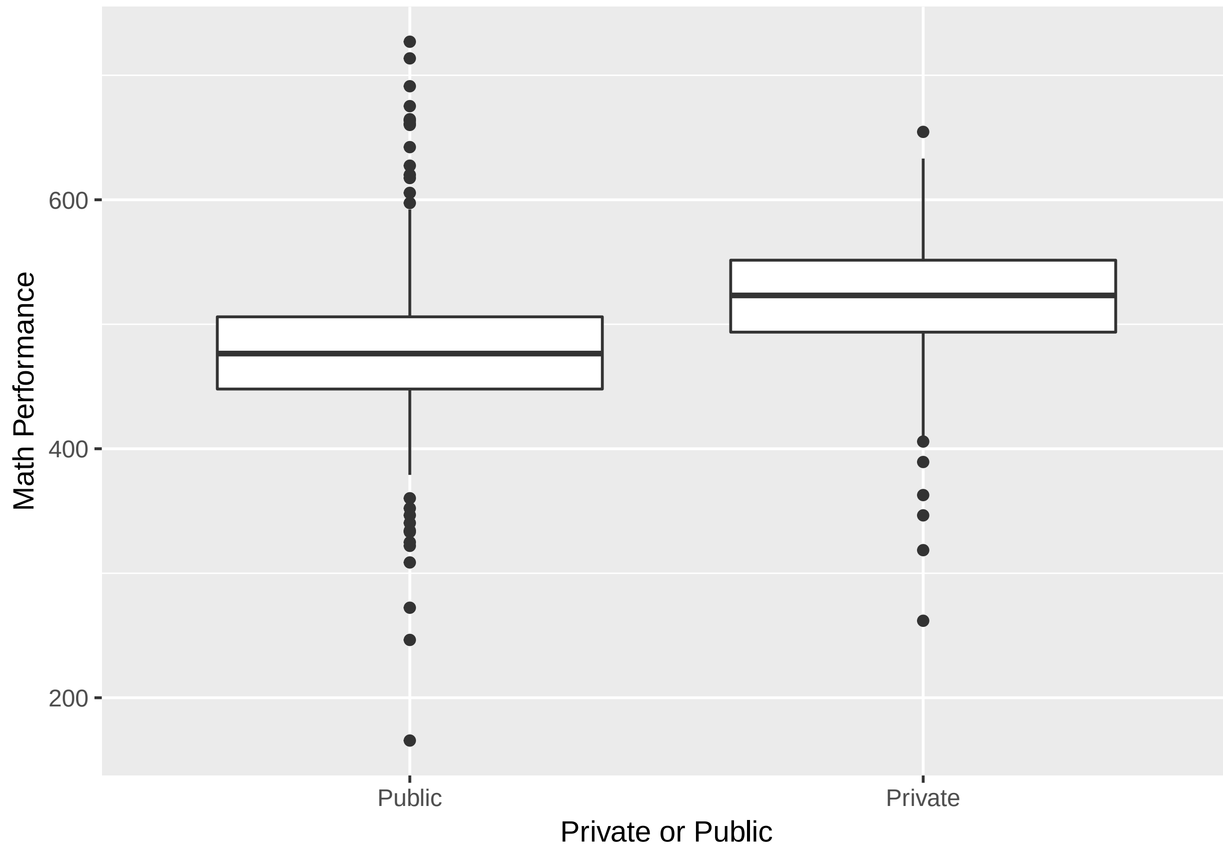


The boxplot shows there is a significant difference in the mean score of students from different region of Australia. The graph implies that students from large city in Australia, for example, Sydney and Melbourne normally scores higher than students from a village. Those pairs that are significantly different according to to result of Tuckey test are those that does not across the 0 value.

```
school_view <- school2012[NC =="Australia" & !is.na(SC01Q01)]
school_plot_data <- select(school_view, SC01Q01, SCHOOLID)

#merge two tables into one table for plotting use
plot_data <- merge(student_data, school_plot_data, by="SCHOOLID")

#a simple summary table
plot_data[, .(mu = mean(Mean_PVMATH, na.rm = T),
              sigma = sd(Mean_PVMATH, na.rm = T)),
              by = SC01Q01]
```

```
##    SC01Q01       mu    sigma
## 1: Private 521.0669 49.04944
## 2:  Public 479.2423 59.25269
```

```
#create a boxplot comparing the mean value
ggplot(plot_data, aes(x=factor(SC01Q01), y=Mean_PVMATH))+geom_boxplot()+labs(x="Private or Public", y="
```




```r
#run linear regression test
fit <- lm(Mean_PVMATH~SC01Q01, data=plot_data)
summary(fit)
```

```
##
## Call:
## lm(formula = Mean_PVMATH ~ SC01Q01, data = plot_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -313.605  -29.561    0.133   28.136  247.709
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     479.242      2.561  187.11   <2e-16 ***
## SC01Q01Private   41.825      4.084   10.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.47 on 771 degrees of freedom
## Multiple R-squared:  0.1197, Adjusted R-squared:  0.1186
## F-statistic: 104.9 on 1 and 771 DF,  p-value: < 2.2e-16
```

```r
#run ANOVA test
plot_anova <- aov(Mean_PVMATH~SC01Q01, data = plot_data)
summary(plot_anova)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
```

```
## SC01Q01       1  322649  322649   104.9 <2e-16 ***
## Residuals   771 2372064    3077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The boxplot shows there is a great difference in the mean score of students from different sectors of Australia. The graph implies that students from private schools in Australia outperform the students from government/public schools.

Discussion

The summary shws that the mean score for all Australian students is 495.5 and there are more public schools than private schools attended.

A stuent's socio-economic state will impact their mathematical literacy assessment.

A statistical look into the shortage of math teachers shows that students from schools that has no shortage of math teachers at all have a hghest mean score of 514.80, whereas students from schools that is lacks a lot of math teachers is 475.73. The linear regression test shows that whether the school is in short of math teachers will impact the student's math assessment result greatly since the p-value is very small(<0.05)

From the Tukey test, we can conclude that:

1. There is no significant discrepancy in performances between schools that is facing A lot-Very little and A lot-To some extent shortage of math teachers.(since the p-value is greater than 0.05)
2. There is a significant difference between schools that is facing To some extent-Not at all, Very little-Not at all and A lot-Not at all shortage of math teachers.(since the p-value is equal to 0 or almost equal to 0)

According to the school location analysis, we can tell that there is a positive relationship between the school location and the student's math literacy score. The mean score for all Australian students is 495.5, whereas the students from a village is 448.41, mean score for students from a large city is 518.59. The difference between mean scores is (518.59-448.41)=70.18 points.

In the linear regression test for school location, the p values are all smaller than 0.05 which implies that the store locations are highly related to a student's math performance. In the ANOVA test, the F value is 29.76 and p-value is very small(less than 0.05). Which also implies that there is a significant relationship between the school location and the scores in students' math literacy. And since there are multiple(more than 2) factors in this dataset, we will run another tukey test to compare the differences between each paired groups.

From the Tukey test above, telling from the column diff and p adj, these facts can be able to conclude:

1. There is no significant discrepancy in performances between a Town school and a Small Town School since the p=0.55>0.05 and Small Town-Village since p=0.07>0.05.
2. There is significant discrepancy in math performace between City-Village, Large City-Village, Large City-Samll Town, Large City-Town, since the p value all equal to 0. These significant difference is shown in the plotted Tukey graph.

According to the private/public analysis, we can tell that there is a positive relationship between whether the student enters a private or public/government school is a factor that is affecting their PISA mathematical literacy assessment. While the mean score for all Australian students is 495.5, the mean score for students from independent/private sector is 521.067, mean score for students who attend in public/government school is 479.2. The difference is (521.067-479.2)= 41.87 points. In the linear regression test from the private/public test, the p values are all smaller than 0.05 which implies that the whether the school is a private on or public one is highly related to a student's math performance. In the ANOVA test, the F value is 104.9 and p-value is very small(less than 0.05). Suggesting that there is a strong relationship between whether a school is private or public can affect the student's score on math literacy. Since there are only two levels of factors in this test, we will not run a Tukey test like above.

Evidence shows that whether the school is privately or publicly owned and also the school location and education resource is going to affect the students' mathematical literacy score on PISA 2012. Thus we will be rejecting my H0 hypothesis, hence we will be accepting my H1 hypothesis that whether a school is privately or publicly owned and the location of schools will impact on students' test reults on PISA2012 mathematical literacy assessment.

Conclusion

A few results can be concluded accroding to the analysis result:

1. Students from private/independent schools significantly outperform public/government students.
2. Students from large city outperform students from other scetors of students.
3. Students with a higher socio-economic state also show a higher level of proficiency in math literacy.
4. Averagely, the result of all Australian students fall in level3 of proficiency.
5. Although the mean score for large city student's is highest among the but the standard deviation is a largest too, implying that the score distribution of large city students is wider than of other sectors.
6. Since students from a higher socio-economic family is more likely to enter a private school, these batch of students contributed to the higher performance in mathematical literacy of private schools.
7. The shortage of math resource will also impact on the student's math performance. However, it's interesting that once the school is in short of math teachers, the difference shortage level of math teachers does not impact greatly on the performance. For example, there is no great difference between the mean score of students from schools who are facing: shortage of math teacher to some extent, very little shortage of math teacher and a lot shortage of math teachers.

To enable deeper investigation in the future, I think it's necessary that we incoporate data source from the Government of Education and try to analyze the discrepancy of education resources and funds distributed across the nation. This report can serve as an evidence to the analysis of equalization in education in the future.

Nevertheless, I think the difference in the students' math performance can not serve as the evidence to tell if the students from a village or remote area is "poorly" educated. Since the assessment is standardized, it's actually unable to evaluate intangible attributes of a student, such as their EQ and their values. As far as I know, math is probably something that is comparetively irrelevant for students who comes from indigenous family and live their life in the outback. An assessment on hunting skill may seem more crucial to them.

Reference

Bache, Stefan Milton, and Hadley Wickham. 2014. *Magrittr: A Forward-Pipe Operator for R.* https://CRAN.R-project.org/package=magrittr.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. doi:10.18637/jss.v067.i01.

Biecek, Przemyslaw. n.d. *PISA2012lite: Set of Datasets from Pisa 2012 Study.* https://CRAN.R-project.org/package=PISA2012lite.

Dowle, Matt, and Arun Srinivasan. 2017. *Data.table: Extension of 'Data.frame'.* https://CRAN.R-project.org/package=data.table.

Halsey, J. 2017. *Title: Independent Review into Regional Rural and Remote Education— Discussion*

Paper. https://docs.education.gov.au/system/files/doc/other/independent_review_into_regional_rural_and_remote_education.pdf.

PISA 2012 Results in Focus. 2014. https://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf.

SCHOOL Questionnaire for Pisa 2012. 2011. http://www.oecd.org/pisa/pisaproducts/PISA12_ScQ_ENG.pdf.

"The Australian Education System | Australian Education Technology." 2017. https://www.austrade.gov.au/edtech/the-australian-education-system/.

Thomson, Bortoli, S., and S. Buckley. 2013. *PISA 2012: How Australia Measures up*. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1015&context=ozpisa.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Müller. 2017. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.