This report presents a data analysis sequence of California Test Score Data. The California Test score (CASchool) data set comes with AER R package. Report of analysis of CASchool data set includes statistical analysis for the research questions, Using the simple to complex methods, computing, making correct inferences and finally a statistical computation and visualization of the data set. The analysis of the data set carried out using the R environment for statistical computing and visualization. R environment is an open-source dialect .The California Test Score data set involves the data from all K-6 to K-8 districts in California with data available for 1998 and 1999. This data set contains data on test performance, school characteristics and student demographic background for school districts in California. The aim of the analysis is to answer two research questions using the statistical analytics techniques. [(package, docs and browser, 2022)](#text1)

The data set contains 420 observations and 15 variables. School characteristics contain enrollment, number of teachers, number of computers per classroom and expenditures per student. The demographic variables contain the percentage of students in the public assistant program CalWorks, the percentage of students that qualify for reduced price lunch and the percentage of students that are English learners.

In order to analysis the data set, first the data set should be loaded to the R environment.



Text

Description automatically generated with medium confidence

Imported California Test Score data set:

A picture containing graphical user interface

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidence

There are 420 observations in the data set. The data set includes 15 variables.

1. District : District code
2. School : School name
3. County : Factor indicating county
4. Grades : Factor indicating grade span of district
5. Students : Total enrollment
6. Teachers : Number of teachers
7. CalWorks : Income assistance
8. Lunch : Percent qualifying for reduced price lunch
9. Computer : Number of computers
10. Expenditure : Expenditures per students
11. Income : Average income of the district
12. English : Percent of English learners
13. Read : Average reading score
14. Math : Average math score

Then str() method is used to examine the basic structure of the data set. It shows number of variables and number of observations in the data set and also the data type of each attribute in the data set. [(Data Analysis and Visualisations using R, 2022)](#text3)

A screenshot of a computer

Description automatically generated with medium confidence

Output of the str() method:

Text

Description automatically generated

The output shows that the school, county and grades variables are char type variables. All the other variables are numeric type variables. All the numeric variables are double variables.

The data set is renamed as cas. It is very useful to type the command easily.

Graphical user interface, text

Description automatically generated

First few rows of the data set can be seen by head() function in R studio.

A picture containing graphical user interface

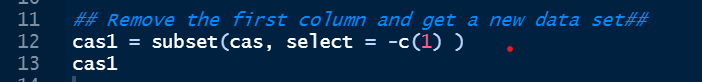
Description automatically generated

The output gives the few rows of the data set with their data types.

A screen shot of a computer

Description automatically generated with low confidence

There are 15 variables in the data set. But the first column is not useful for the analysis. Therefore, the first column is removed by using subset() function. After removing that column, the new data set is named as cas1.

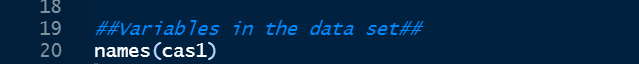


Here are the few rows of new cas1 data set. Now the data set contains only 14 columns.

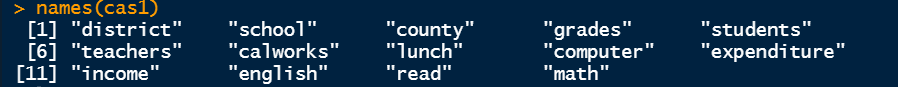
A screen shot of a computer

Description automatically generated with low confidence

The names() function is used to check all the variables in the data set.



The output is as follows.



Now there are only 14 variables in the cas1 data set.

In order to have a quick look at of data, the tail() function can be used. The tail() function shows the last few rows of the data set.

A screenshot of a computer

Description automatically generated with medium confidence

The output are as follows. Table

Description automatically generated with low confidence

The function summary() gives summary statistics about the data set. Summary statistics of a data set is very useful to get an idea about the data set.

Graphical user interface, text

Description automatically generated

The output are as follows.

A screen shot of a computer

Description automatically generated with low confidence

A picture containing graphical user interface

Description automatically generated

The summary() function is most important to summarize the each and every attribute in the data set. This function gives a set of descriptive statistics of a variable. But it depends on the type of the variable. In the cas1 data set, there are three character variables. For character variables, summary() function gives the length and the class of the variable. Length is the number of observations in the data set. In cas1 data set school, county, grades have the length of 420 and the class of those variables is character. All the other variables in the data set are numerical variables. For the numerical variables, summary() function gives mean, median, minimum value, maximum value and the quartiles of the variable. [(RPubs - Final Project - Data110, 2022)](#text2)

Following function is used to take more extensive summary statistics about the data set.



The output of the above function is as follows.

Graphical user interface, text, website

Description automatically generated

Timeline

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface, text

Description automatically generated

By using the above function, rather than the summary() function gives more statistics about each and every variables in the data set. It is very useful for the conclusions.

The average test score and the student-teacher ratio is not included in the caschools data set. But those two variables are important for the future data analysis. Therefore, the average test score and the student-teacher ratio is calculated using following functions and appended it to the cas1 data set.

Text

Description automatically generated

The output is as follows.

A screen shot of a computer

Description automatically generated with low confidence

Now there are 16 variables in the data set. The STR and score variables are the new variables. The data type of these two variables is double.

Following mean(), sd() and quantile() functions are used to compute the mean, standard deviation and the quartiles of the STR and score variables. Then these data is gathered into new data frame called cas1\_summary.

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

The cas1\_summary data frame are as follows.

Text

Description automatically generated

This data frame shows the average, standard deviation and quantile 10, 25, 40, 50, 60, 75, 90 of the quant\_STR and quant\_score variables.

The function plot() can be used to identify the outliers present in the data set. The most accurate plot for this function is scatter plot. Following R commands represent the scatter plot for STR and score variables in the data set. [(Quick-R: Scatterplots, 2022)](#text4)

Text

Description automatically generated

The scatter plot for STR~score is as follows.

Chart, scatter chart

Description automatically generated

Figure 01: Scatter plot of test score and student- teacher ratio.

This plot is the scatter plot of student-teacher ratio and test score variables. The points are strongly scattered, and the variables are negatively correlated.

The following function shows the correlation between STR and score variables.



The output is as follows.

Text

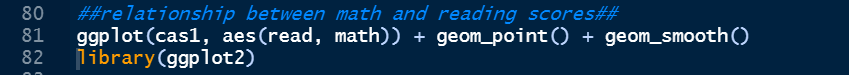
Description automatically generated

According to the above scatter plot, the correlation between str and score variables is negative. It can be confirmed by seeing the above result. The correlation between those variables is -0.2263627. It is also a week correlation. [(California schools analysis demonstrating use of R Markdown, 2022)](#text3)

A statistical analysis is descriptive, simply reporting, summarizing and visualizing a data set. Statistical procedures are used to answer research questions. In such cases the statistical analysis is inferential. The data items in the data set should be motivated by research questions. In this case study two research questions are taken to answer using exploratory data analysis.

1. The relationship between the district level math and reading scores.

In order to find the relationship between the district level math and reading scores, ggplot() function should be used. [(Create Elegant Data Visualisations Using the Grammar of Graphics, 2022)](#text6)



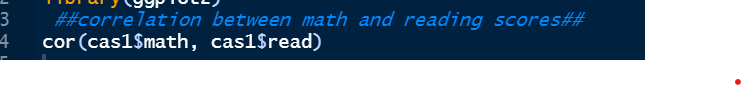
The output is as follows.

Chart, scatter chart

Description automatically generated

Figure 02: Scatter plot for math and read variables

This plot shows that the correlation between math score and reading score is high at the district level. The actual value of the correlation between these two variables is calculated using cor() function.



The output is as follows.

Text

Description automatically generated

The correlation(r value) between the district level math score and reading score is 0.92. The individual level correlations are in the range of 0.3 to 0.6. Therefore the conclusion of the above results is that the group level relationships are much stronger than the individual level relationships.

1. Relationship between the income of schooling districts and the test scores of those districts.

In order to find the relationship between income of schooling districts and the test scores of those districts, the student-teacher ratio(STR)and the average test score(score)should be calculated and appended it to the cas1 data set.

Graphical user interface, text

Description automatically generated

The correlation between income and score variables are calculated using cor() function. [(Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA, 2022)](#text7)



The output is as follows.

Graphical user interface

Description automatically generated with medium confidence

According to the above output, the correlation between income and score is 0.7124308. It is a positive number, which means that the income and the score are correlated positively. It concludes that the schools with high average income achieve high average test scores. In order to model this data, the linear regression model can be used. The model can be built up by plotting the data and then add a linear regression line to the plot. The following code snippet shows how to build the linear regression model for the income and score variables. [(2022)](#text5)



The function lm() is used to build the simple linear model to the score and income variables.

The following code snippet shows the plot for the observations of income and score variables.

Text

Description automatically generated

The output is as follows.

Chart, scatter chart

Description automatically generated

Figure 03: Scatter plot for test score vs. district income

This is the plot for test score vs. district income. Now the regression line is added to the plot by using following R function.

Graphical user interface, text

Description automatically generated

The output is as follows.

Chart, scatter chart

Description automatically generated

Figure 04: scatter plot for test score vs. district income with the regression line

The linear regression line seems to overestimate the actual relationship when income is very high or very low and underestimate it for the middle-income group.

The following equation shows the model test scores as a function of income ad the square of the income.

Score i = β0 + β1 × incomei + β2 × + ui

In the above equation the an additional explanatory variable. Therefore this is a quadratic model. It is a special case of multivariate regression model. The lm() function is used to fit the model for the equation. When fitting this model ^ operator should use with I() to add the quadratic term to the formula as an additional regressor. [(Zach, 2022)](#text)

The following code snippet shows how to do this.

Graphical user interface, text, website

Description automatically generated

In here, the passed regression formula is converted into the object of the class formula. Therefore the objects of this class do not have an arithmetic interpretation with the -, +, \* and ^ operators. The function I() is used to ensure that they are used as an arithmetic operators.

The output of the above coeftest() function is as follows. It shows the summary of the model.

Graphical user interface, text, chat or text message

Description automatically generated

According to the above output, the value of the β0 is 607.3 which is the intercept of the model. The value of β1 is 3.85. The value of β2 is -0.0423. Now the equation is as follows.

Score i = 607.3 + 3.85× incomei  - 0.0423×

The standard errors of β0, β1, β2 are 2.90, 0.27, 0.0048. A hypothesis testing can be done by using this model. Which is the relationship between the district income and test score is linear while the alternative is quadratic. [(2022)](#text8)

The hypothesis is as follows.

H0 : β2 = 0 vs. H1 : β2 ≠ 0.

In here, β2 = 0 implies that the linear regression model while the β2 ≠ 0 implies the quadratic regression model. The t = (β2 – 0)/ SE(β2) = (-8.81). This implies that the null hypothesis is rejected at any level of significance and the conclusion of the above results is the relationship between the test score and the income of the districts is nonlinear.

Below code snippet shows how to add the regression line for the quadratic model. In this case abline() function can not use because of it draws the straight line to the plot. The lines() function is used to avoid this as it draws the non-straight lines to the model.

Graphical user interface, text

Description automatically generated

In this code the scatter plot is drawn for the observations of income and test score variables. Then the regression line is drawn to the plot using abline() function. Line is dText

Description automatically generated

By using the lines() function, a non-straight line is drawn to the plot.

The output of the above functions is as follows.

Chart, scatter chart

Description automatically generated

Figure 05: scatter plot for estimated linear and quadratic regression functions

The quadratic regression line is fitted to the data better than the linear regression line.

Therefore the conclusion of all above analysis is that the relationship between district level income and the test score variables are highly correlated to each other.

REFERENCES:

1. package, F., docs, R. and browser, R., 2022. *CASchools: California Test Score Data in AER: Applied Econometrics with R*. [online] Rdrr.io. Available at: <<https://rdrr.io/cran/AER/man/CASchools.html>> [Accessed 9 May 2022].
2. Rpubs.com. 2022. *RPubs - Final Project - Data110*. [online] Available at: <<https://rpubs.com/dlachaud/844875>> [Accessed 9 May 2022].
3. Medium. 2022. *Data Analysis and Visualisations using R*. [online] Available at: <<https://towardsdatascience.com/data-analysis-and-visualisations-using-r-955a7e90f7dd>> [Accessed 9 May 2022].
4. Gist. 2022. *California schools analysis demonstrating use of R Markdown*. [online] Available at: <<https://gist.github.com/fototo/2732813>> [Accessed 9 May 2022].
5. Zach, V., 2022. *How to Perform Quadratic Regression in R - Statology*. [online] Statology. Available at: <<https://www.statology.org/quadratic-regression-r>> [Accessed 10 May 2022].
6. Statmethods.net. 2022. *Quick-R: Scatterplots*. [online] Available at: <<https://www.statmethods.net/graphs/scatterplot.html>> [Accessed 10 May 2022].
7. 2022. [online] Available at: <<https://www.datacamp.com/tutorial/linear-regression-R>> [Accessed 10 May 2022].
8. Ggplot2.tidyverse.org. 2022. *Create Elegant Data Visualisations Using the Grammar of Graphics*. [online] Available at: <<https://ggplot2.tidyverse.org/>> [Accessed 10 May 2022].
9. Sthda.com. 2022. *Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA*. [online] Available at: <<http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r>> [Accessed 10 May 2022].
10. 2022. [online] Available at: <<https://data-flair.training/blogs/hypothesis-testing-in-r/>> [Accessed 10 May 2022].