Edinburgh Napier UNIVERSITY

Coursework Report

WEIGUANG RAN
40412989@napier.ac.uk
Edinburgh Napier University - Data Analytic (SET09120)

1 Background

Rstudio:
RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio was founded by JJ Allaire,[5] creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio
ggplot2:
ggplot2 is a data visualization package for the statistical programming language R. Created by Hadley Wickham in 2005, ggplot2 is an implementation of Leland Wilkinson's Grammar of GraphicsâĂŤa general scheme for data visualization which breaks up graphs into semantic components such as scales and layers. And this report's most data are using ggplot2 to analyze.
R language:
R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. R language is good for analyzing data, and have many packages to help people update their software.

2 Introduction

Data: These data record 10,000 results from different ages and genders in different regions. And the regions are ss, sj, st, sd, so. Here are the definition to these regions.
*Gender (can be M or F)
*Location (can be A, B, C, D or E)
*Age. The age is in four bands, Teenager, Under 25 (but over 20), Late 20s, and Over 30.
*Sprints. Score out of 100.
*Jumps. Score out of 100.
*Throws. Score out of 100.
*Distance. Score out of 100.
*Overall. Score out of 100.

TEST: In this test data there are 7 relationships, so we need to use some plots to find these relationships. We should also write a report, which included 3 pages of words and maximum 3 pages for pictures.

Preparation: After I got this data, the first thing I did was import this file to Rstudio, and explore each title's meaning. After I got the meaning of each title, I began to clean up the data and summary the data. And then, conduct a specific analysis for each label and find some deep rules. Finally, modify and fill some graphics to improve the rules of their own discovery.

3 Relationships:7

1.The relationship between gender and individual scores:Females only have higher average scores on SO than males, and others males are higher than females.

2.Relationship between region and total score: The average score of students in D area is the highest, the average score of students in E area is the lowest, there are some outliers in the B and E areas, and the difference in the scores of the "middle part"(numbers between Q1 and Q3) in the E area is the smallest.

3.Relationship between age and various sports: The "middle part" of the "O" age group is higher than all other age groups in all sports. And the long jump score is the most outstanding in 'O' Area. 'SO' has outliers in all ages.

4.Distribution of grades in each category: The density trends of the four sports are not much different, and most of last item are between 40 and 60, also the final total score is not necessarily related to the performance of the first four sports.

5.Linear correlation analysis: ss and st are negatively correlated, sj is positively related to so, and st has nothing to do with sd

6.The distribution of performance of each sport is parabolic, that means the number of people in the middle is the most.

7.The average score of so is the highest, and the average score of sj is the lowest. And the aver. in sj for female is the lowest, the aver. in so for female is the largest.

4 Analyzing

Relationship1:
By calculating the average scores of male and female in each sport category and comparing them, we can get the results.
Coding aver1 ¡- tapply(sj, Gender, mean)View(aver1)
aver2 ¡- tapply(st, Gender, mean)View(aver2)

aver3 ¡- tapply(sd, Gender, mean)View(aver3)
aver4 ¡- tapply(so, Gender, mean)

Relationship2:
According to 'plot1.1.0' and the total grades for each location(Ps:all plots can be found at Appendix). The aver. of all locations:A:246.1910 B:244.8515 C:250.0295 D:274.3750 E:244.6475

Relationship3:
According to 'plot1.1.1' and 'plot1.1.2', plot1.1.1 is the plot for five boxplots, which for location and each sport,and for these five boxplots we can summarize that location o is the place where people have the best grades. And plot2.2.2 is the boxplot for sj and other personal informational(Gender, age, location), so we can find the top boxes are all come from location o, location o is the best place for jumping.

Relationship4:
According to 'plot1.1.3' and the aver. for five grades' items. 'plot1.1.3' is the trending for five grades' items, we can find some useful information for five sports, density can be thought for numbers, the higher the density, the more students there are. Also by calculating the average number of each exercise, we can find the aver. of overall grades is not the aver. for other four sports except 4, which means aver.so =/ aver.(ss+sj+st+sd)/4.

Relationship5:
Based on 'plot1.1.4', plot1.1.4 is a plot describe the linear correlation between variables. From this figure we can see that if there is a horizontal line between the two variables, there is no linear correlation between them. If it is a rising line, it means that it is positively linear, and the line that falls is negative linear correlation.

Relationship6:
Based on 'plot1.1.3', from 1.1.3 we can see all functions change trend roughly the same—look like parabola, that means many people in the middle grade. Also we can find that the trend of so the greatest change line in the chart.

Relationship7:
According to 'plot1.1.5' and the excel 'aver', the plot 1.1.5 shows the relationship between long jump and personal informationïijĹGender, location, Age), In the boxplot we can clearly see that a set of box is significantly lower than others, and that box is old women in T location. The excel 'aver' is the aver. about each sports in male and female, here is the excel contents:

Location	Female	male
ss	48.6922	50.0212
sj	45.8322	49.4168
st	49.5884	50.0798
sd	50.0074	50.1392
so	52.2340	55.0256

So in this excel we can find which is the highest and lowest grade.

4.1 Maths

How to Calculate the average:
A+B+C+..+N/N

How to calculate weighted average:
In mathematical terminology, you express this type of average this way:
âĹŚ(m1...mn) n
where the symbol âĹŚ means "sum all the measurements from 1 to n."

How to calculate Q1 and Q3(quartiles):

Quartiles are the values that divide a list of numbers into quarters:
*Put the list of numbers in order
*Then cut the list into four equal parts
*The Quartiles are at the "cuts"

4.2 Coding

boxplot using ggplot2:
ggplot(dataset, aes(x=interaction(Gender,Age,Location),y=st))+ geom-boxplot()+scale-color-manual(values=c("999999", "E69F00", "56B4E9"))+ +theme(legend.position="right") +labs(title="compare with st and others",x="location,ages,sex", y = "numbers")

Density analysis:
plot(density(so), col='blue', lwd=2)
lines(density(ss), col='red', lwd=2)
lines(density(st), col='black', lwd=2)
lines(density(sd), col='green', lwd=2)
lines(density(sj), col='orange', lwd=2)

Linear distributio:
pairs.panels(dataset[c("ss","sj","st","sd","so")])

5 Conclusion

5.1 What I learned form this cw:

1.I learned that R as a programming language has its own unique charm.
*R is not just a statistics package, it is a language.
*R is designed to operate the way that problems are thought about.
*R is both flexible and powerful.

2.Gain problem solving skills
At heart, analytic is all about solving problems. The problems just happen to be on a much larger scale than what many of us are used to - effecting entire businesses, along with the staff and customers that they serve. The ability to think analytically and approach problems in the right way is a skill that's always useful, not just in the professional world, but in everyday life as well.

3.learn something about the frame and package in R language.

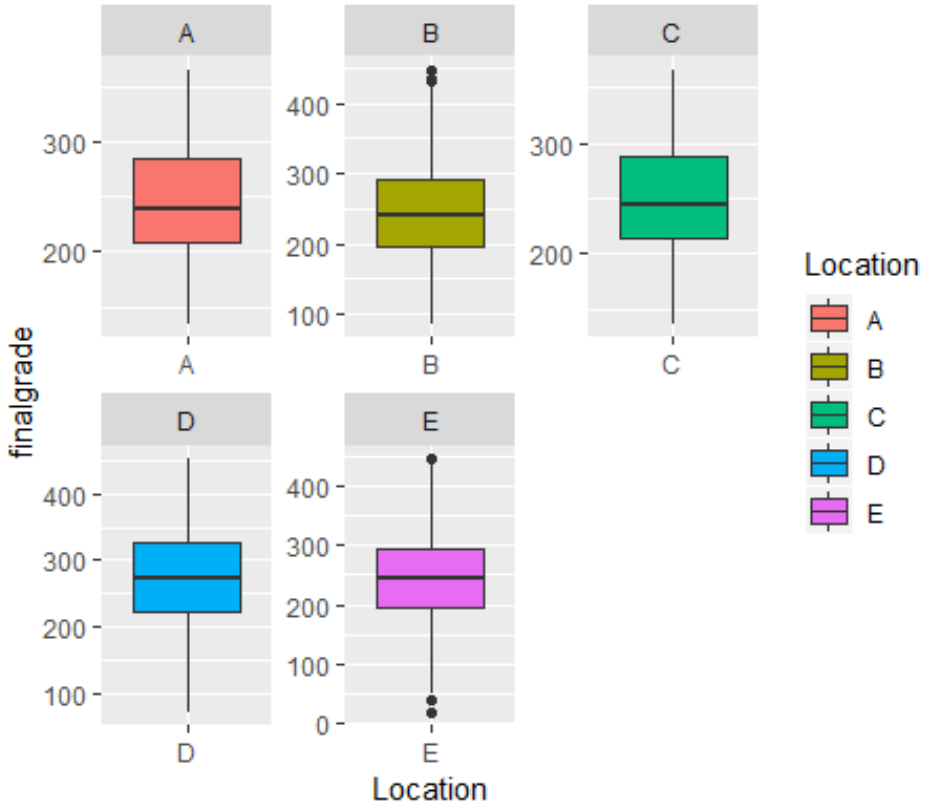
In this coursework I learned how to use ggplot2 package and rstudio, they are quite useful to help me analytic data in the future, and they help me open the door for database and data analytic.

5.2 To sum up:

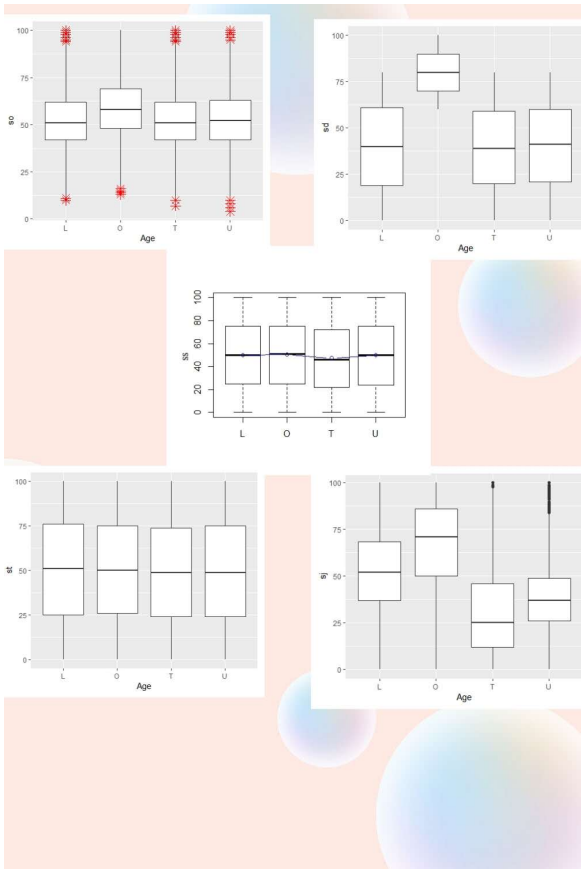
Ultimately, there really isn't any doubt that analytic is going to be a huge element of enterprises in the future. Getting ahead of the curve by learning analytic now provides a pathway to success, as well as transferrable skills that can help in every facet of life. So do not give up, continue to learn data analytic.

6 Appendix:

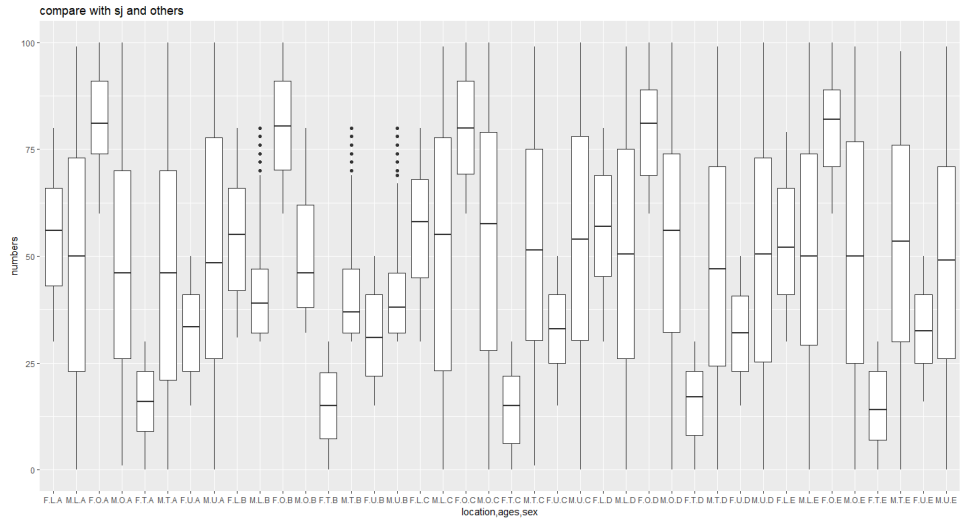
6.0.1 plot1.1.0:



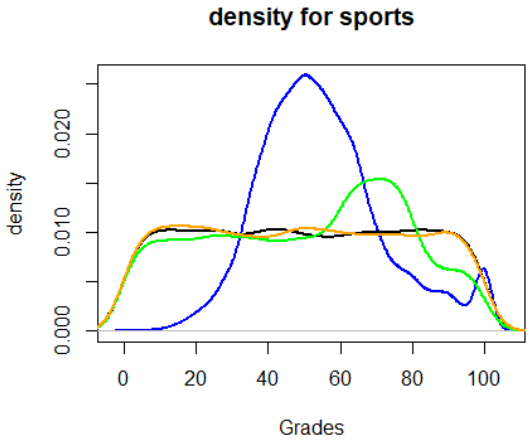
6.0.2 plot1.1.1:



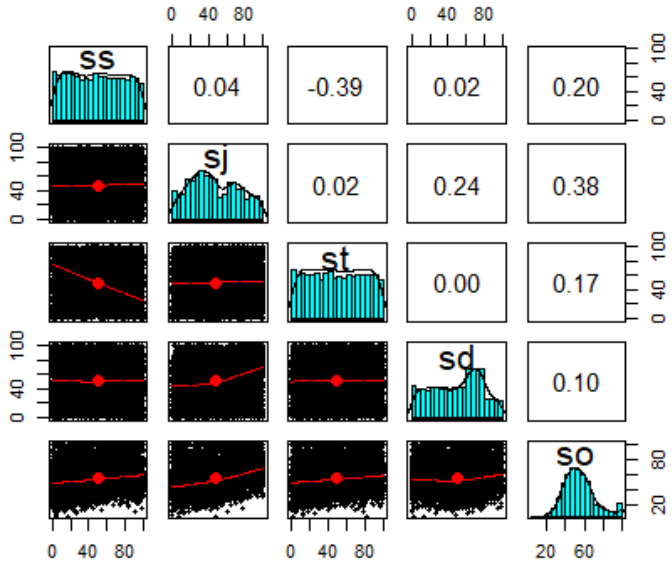
6.0.3 plot1.1.2:



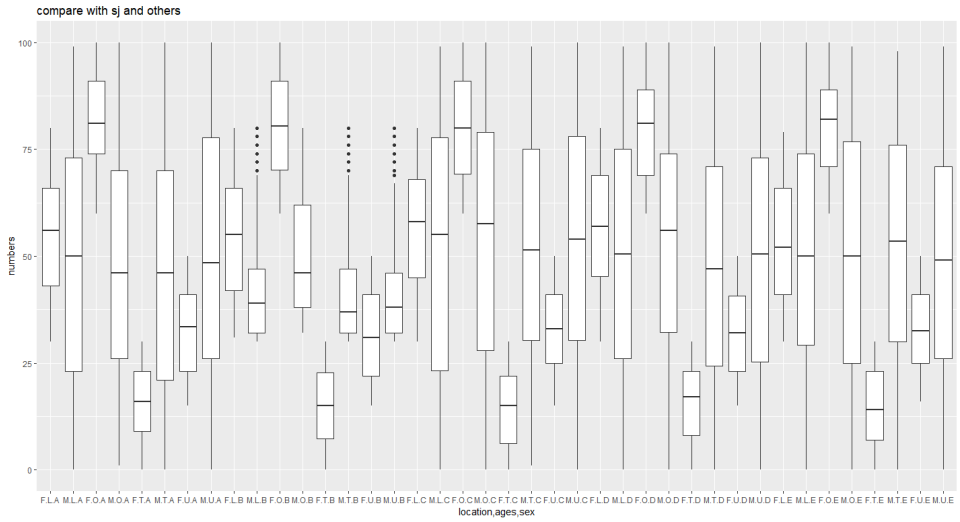
6.0.4 plot1.1.3:



6.0.5 plot1.1.4:



6.0.6 plot1.1.5



6.0.7 References:

1.https://www.kancloud.cn/thinkphp/latex/41808

2.https://en.wikipedia.org/wiki/R-(programming-language)

3. "Syntax Highlighting". Kate Development Team. Archived from the original on 2008-07-07. Retrieved 2008-07-09.

4. "R: What is R?". R-Project. Retrieved 2018-08-07