CMSC320 Final Project

Authors:

I Cheng Chan (writing and coding)

Xinyao Zhang (writing and coding)

Data Science

Data science is extremely important nowadays in our life. Artificial intelligence and machine learning depends largely on the big data analysis. Besides computers, human decision makers also make detailed analyses based on data about specific field, in order to make better decisions.

This course would involve data analyzing from cvs file and sometimes data from a website. The data could have over thousands of entities (row) and multiples of attributes (column). The users should pay close attention on useful analysis such as standard distribution, mean, center and other types of analysis which would introduce later in this document.

Before analyzing data, manipulating data is often needed so that unnecessary data could be eliminated from analyzing progress.

There are multiple ways to manipulate data before analysis such as handling missing data, tidying data, etc.

Visualizing data is also an important feature of data analyzing. Different kinds of graph could provide different information to the read, as well as avoiding unimportant information regarding what reader is looking for.

DATA RESOURCES:

https://www.kaggle.com/uciml/student-alcohol-consumption#student-mat.csv

This is a data set that is related to a survey conducted between 2 school's mathematics course students. The entity of the data would contain some personal background information. Our aim is to answer the question regarding whether a student's alcohol consumption and guardian would affect their performance in their studies.

This first step is to create a Rmd file in order to analysis the dataset in R language.

Before the analysis, we have to download the csv from the website above and store it at the same location of where the Rmd file is created, so that we could use the dataset within the Rmd file.

library(tidyverse)

- Load the installed package name tidyverse, before using this library, user might need to install the package in console by 'install.packages("tidyverse")' command.

read\_csv(file)

- read\_csv command is part of the readr R package and allows you to read a dataset stored in a csv file. In this function, there parameter aside for file is not required, unless user wants to change anyone of them, there is no need to type that out.

<-

- this symbol means assigning the result from the right variable to the left variable

typing the variable name alone can show the content contained by this variable

table(…)

- show the corresponding data set in summarized factor form

Here we chose tab$sex, mean information in the column named sex, below showed that there are 2 types of entity in the column, F and M, this is categorical attribute.

Categorical attributes

- something that can be categorize like true or false, disagree or agree, race, etc. Data maybe order or unorder, sex is unordered.

Numerical attributes

- can be discrete or continuous. Discrete meaning that the number is no continuous, for example, something that can be counted (able to be answered by How many). As for continuous, is the length or width etc. of something (able to be answered by How much)

In our chosen column G1,G2,G3 would be discrete numerical attributes, because from 0 to 20, we are counting how many 0, 1.in total.

%>%

- pipe operator, this could create a pipeline, so that functions that are piped can be perform within one step before output the result. User can use multiple pipe operator, the function will be run from top to bottom accordingly.

Here, we are adding (mutate) an extra column name G to the dataset, the column G is average calculated from the column G1,G2 and G3, the type of this column is double (dbl).

By now, we are ready to analyze the data. First let's take a look at the grade represented in graph.

We used arrange (G) to arrange the order of the dataset in ascending order by default according to the column G. Next, we assign the row number for each entity (rowed\_to\_column()).

ggplot() is a function we used to draw graph, inside the function we assigne the row number to x-axis, the average grade column G to the y-axis. The + sign is use to say that what type of graph we would like to draw, there are many other kinds of graph like geom\_boxplot() for boxplot graph which would be used latter, grom\_point() is drawing the graph by point for each value.

Now we draw the histogram for the same input.

From we graph, we could have some basic idea through observation, which is the following:

Exploratory data analysis:

central trends (mean)

spread (variance)

skew

outliers

In this density plot which shows the distribution of numerical variable, we could see that that data is slightly skewed, most of the variable is around grade that is 10. This graph is similar to the histogram above.

Boxplot is a graph that should show the median (the line inside the box), max (the top line's ending point), min (bottom line's the ending point), 3rd quartile (box's horizontal top line), 1st quartile (box's bottom line). Also, according to the boxplot, there is not outliner (extra dots).

Generate min, max, mean, median from the dataset. The summarize() function allow the result only shows the expected result from calculation.

Here we are adding extra information to the histogram, "bins=20" inside grom\_histogram() represent that the number of bins is set to 20 (default is 30), see y-axis range. Then are add a red vertical line to represent the median of G column, which is the same value as the calculation above.

Centering and scaling

To make the analysis easier, we could try to standardize the data through standard deviation. This is useful when there are huge number of entity in the dataset.

standardization transformation formula

0 would represent mean

This graph show the distribution of the standardized grade. Extra function inside histogram binwidth=0.5, this group the data for x-axsis by 0.5, this can simplify the data because some would have decimal place that is close to each other.

Here, we are converting categorical attribute to numerical ones. Since the attributes have 3 options: father, mother, and other, we can represent the guardian by adding 2 columns.

Guardian father mother

father 1 0

mother 0 1

other 0 0

We also want to see how different the result is for different guardian. We created dummy numerical attributes (one for mother, one for father). In this way, the option "other" will be considered as default value since both of its "father" and "mother" attributes are 0.

First we draw the graph representing the relationship between alcohol consumption (x-axsis) and the average grade (y-axsis), the line would represent the trend of the information. We can see that as consumption increase the average grade decrease.

For our null hypothesis, we are saying that there is no relationship between alcohol consumption and grade.

To confirm the hypothesis, we use the lm function. According to the calculation, average grade would decrease by -0.25~ for each consumption rate. By looking at the p value (0.0805~) >0.05, meaning we are accepting the null hypothesis, there is no relationship between alcohol consumption and grade.

Using the same concept, we now calculate the relationship for different guardian together with the student's alcohol consumption compare to average grade. We can see that if the guardian is mother the impact would be smaller than those who guardian is father (0.93~ < 1.25~), also, through observing the p value for both (0.18~ and 0.1~), we could also conclude that there is no relationship between guardian, alcohol consumption and grade.

Through observing the graph, the score of student is generally less with guardian is other, and the slop is steeper than guardian is father or mother, this could mean that the impact is even greater than the two.

Conclusion

Generally, we expect that alcohol consumption would affect student's grade, however, through this experiment, we accept the null hypothesis. Thus, we obtain an opposite conclusion - The alcohol consumption won't affect student's grade.

This should not be the final conclusion regarding to student's grade. There are lots of different aspects that could affect a student's grade, for example, the understand of material, and effectiveness of learning. Different guardian might have different ways to raise their children, but not all of them will be using the same way. On the other hand, we are just using sample will less than 400 coming from 2 schools. The sample size is not large enough to conclude anything.

Through the use of data analysis, people can do lots of analysis related to different areas and situation, some of them might not have the result of what people usually expect, like this case. Therefore, data science is not only a useful technique, but also a tool for people to understand the essense of the topic, thus making wise decsions.

Resources

Here is a website provided by R where user could search for other functionality for their analysis.

https://www.rdocumentation.org/

Linear regression for machine learning, linear regression is not only a useful topic in data science, but also in machine learning.

https://medium.com/datadriveninvestor/understanding-linear-regression-in-machine-learning-643f577eba84

Data science has a big market. There are lots of job opportunities related to data science and the salary is high.

https://www.northeastern.edu/graduate/blog/data-science-careers-shaping-our-future/

Data science help solving real world problem in business and many other areas.

https://www.viget.com/articles/solving-problems-with-data-science/

Real topic related to data science

https://www.kdnuggets.com/2019/02/odsc-10-trending-data-science-topics.html