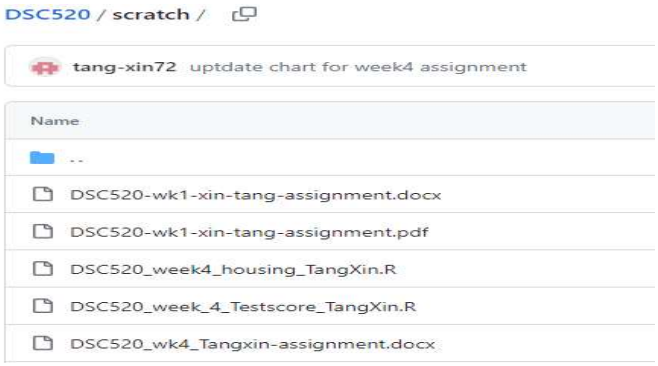
Repo link:

https://github.com/tang-xin72/DSC520/tree/main/scratch



Exercise 1 test score:

```
> data <- read.csv(file = theUrl, header =TRUE, sep =',')
> head(data)
  Count Score Section
1    10   200  Sports
2    10   205  Sports
3    20   235  Sports
4    10   240  Sports
5    10   250  Sports
6    10   265 Regular
```

1. What are the observational units in this study?
   A: There are total 38 units of observations of the scores in the data, grouped by score. The score and count are numeric, and section is a category variable.

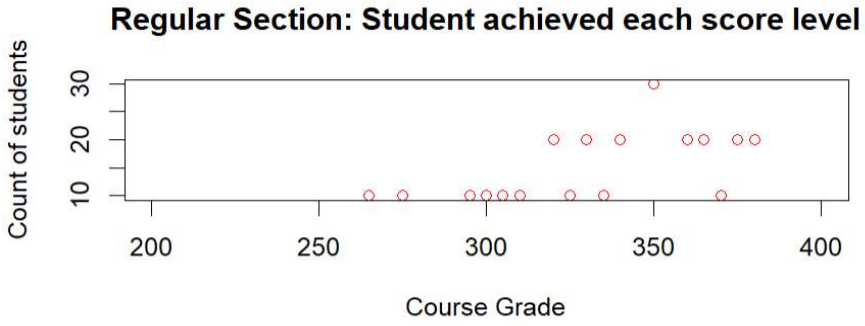
2. Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?
   A: There are 3 variables in the data. The score and count are quantitative variables, and section is a category variable.

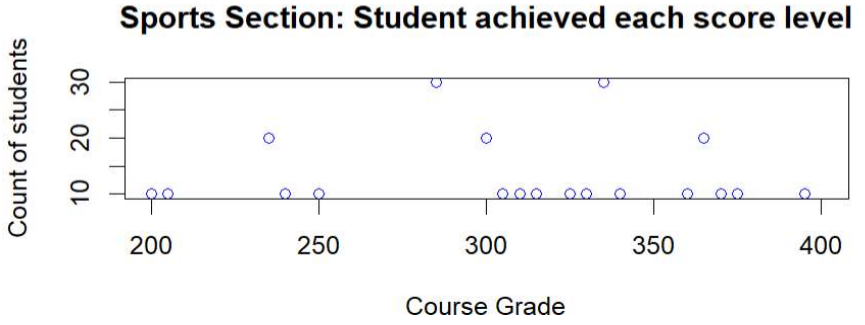
3. Create one variable to hold a subset of your data set that contains only the Regular Section and one variable for the Sports Section.
   A: see R code in program or the screen shot here.

```
regular_data <- subset(data, Section=='Regular')

sports_data <- subset(data, Section=='Sports')
```

4. Use the Plot function to plot each Sections scores and the number of students achieving that score.





4.1 Can you say that one section tended to score more points than the other? Justify and explain your answer.

A: look at the plot, for regular section, more data points of count of 20 and 30 exist in score >300 side. For sports section, more data points of count 10 exist in score > 300 side. Also, there is no data points for score <250 in regular section, while there are some in sports section.

So, more students from regular section scored highs core, especially above 350, also there are no student from regular section scored less than 250. In summary, students from regular section performance better in test.

4.2 Did every student in one section score more points than every student in the other section? If not, explain what a statistical tendency means in this context.

A: look at the plot, it is not true that every student from regular section outperformed all students from sport section. Some students from sport section also get high score, but it is rare. The lowest score the student from regular section gets is 265. While the lowest score in sport section is just 200.

So the mean score for regular section is higher than sport section. Statistically the students in regular section get better test score than those in sport section.

4.3 What could be one additional variable that was not mentioned in the narrative that could be influencing the point distributions between the two sections?

A: the data did not mentioned if the student are taking same test (quiz, mid term or final tec).

Exercise 2 housing data:

See week4_housing_TangXin.R for coding

New variables:

Total_sale_price: created by 'apply' function

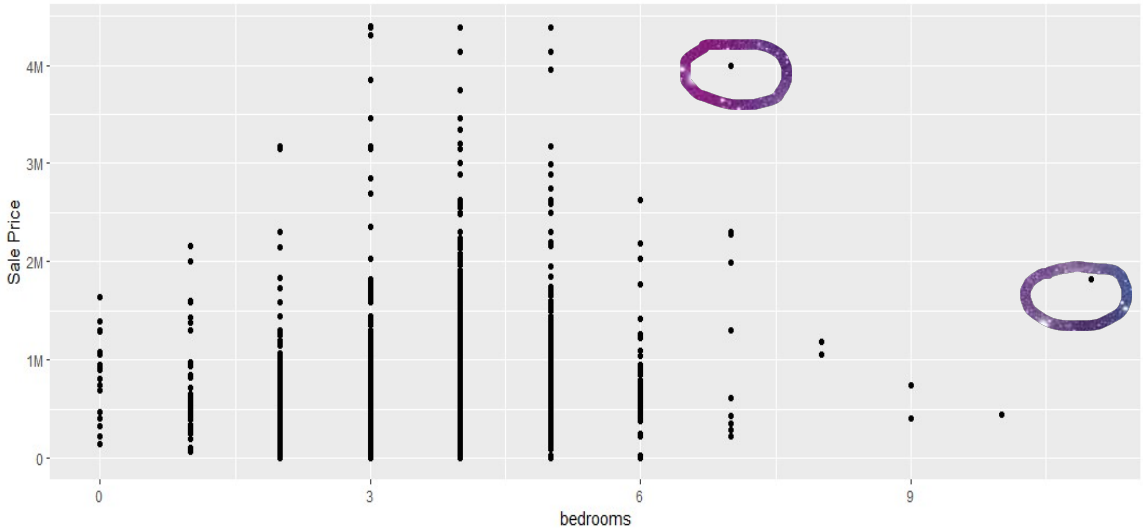
Zipcode_SalePrice_mean: created by 'aggregate' function

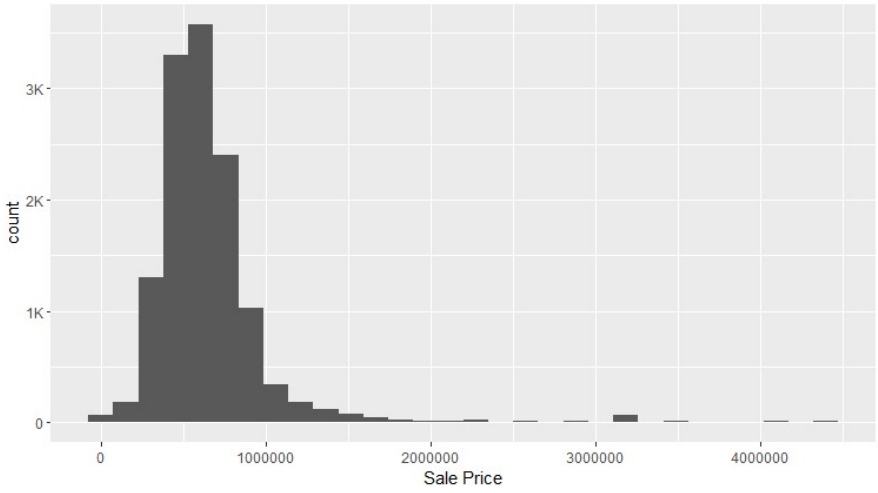
The following variables created using 'plyr' package

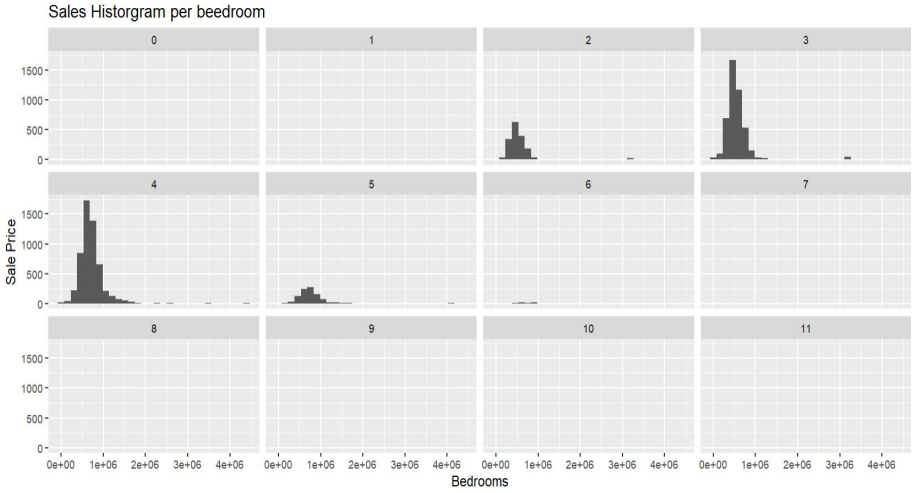
Count_by_sales_room_avg, Count_by_sales_room, Count_by_Sales and Housing_by_zip_bedroom.

Data analysis:

1. Houses with bedroom of 2-4 have most sales.
2. Sales price is not a normal distribution, it is positively skewed, which makes sense since most people cannot afford multimillion mansions.
3. There are a few outliners. Like one house with 7 bedroom sold for 4 million, well above its mean of 2 million. Also the sales of the houses with more than 9 rooms are also outliner in this dataset.




3K
2K
1K
0
count
0
1000000
2000000
3000000
4000000
Sale Price


Sales Historgram per beedroom
0
1
2
3
4
5
6
7
8
9
10
11
1500
1000
500
0
Sale Price
0e+00
1e+06
2e+06
3e+06
4e+06
Bedrooms