R report

Foundation of Data Science

Name: Zhihan Wang
ID:30005124

The data set comes from one fisherman who fished in a lake. It records data including the weight of fishes, time that he made every caught and different types of fishing rods he used. In this assignment, I am going to report what I found from the data set, such as relationship between different factors, the effectiveness of each type of bait and so on.

Firstly, I plotted all points in the data set to have a preview. At *Figure.1*, fishing rod C probably catches the most fishes, as the number of grey points are biggest; besides, it seems that with time goes by, the weight of fishes caught in the lake is decreasing. Assuming there is a linear relationship between time and weight, I calculated the coefficients of the linear model and did a hypothesis test. To be more clearly, I am going to show much more exhaustive analysis below.

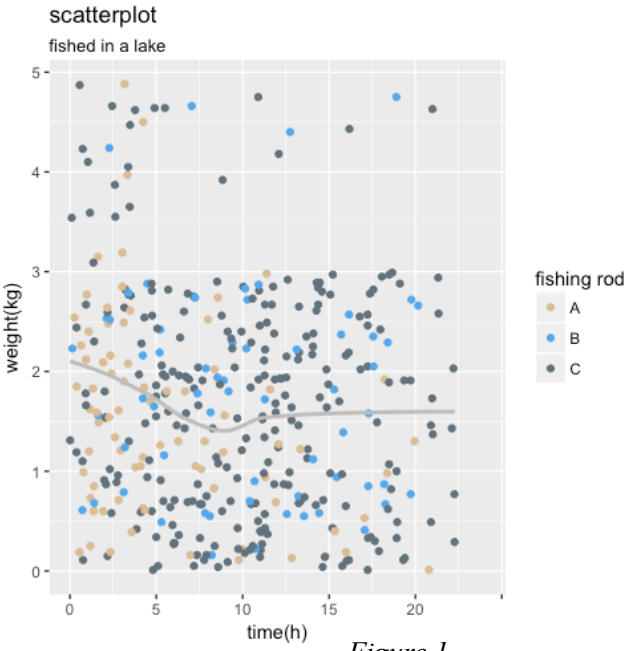

Figure.1

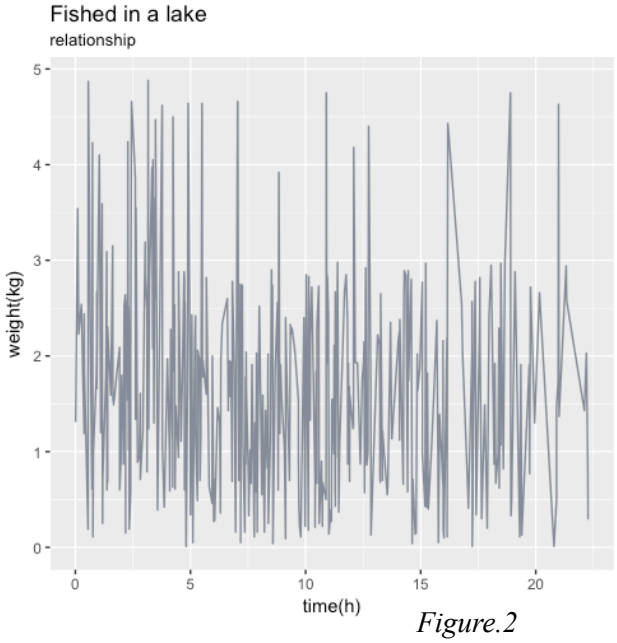

Figure.2

The relationship between time and weight is slightly negative, with a not very small p value 0.014.The figure on the left illustrates the relationship between size of fish(weight) and time the fisherman made every caught. In this plot, there are more bigger size fishes (weights above 3kg) caught between 00:00 and 05:00 than after 05:00. And in one day, the weight of most fishes ranges from around 0.5kg to 2.5kg .

Next, I am going to show some detailed information about time and weight. The number of fishes caught changes in one day. *Figure.3*

tells us only a small proportion of fishes are caught at the beginning and a much smaller at the end of one day. Some messages about weight can be acquired from *Figure.4.* We can see that only few fishes' weights are larger than 3kg and most fishes are in a middle size(0.5~2.5kg), or much smaller.

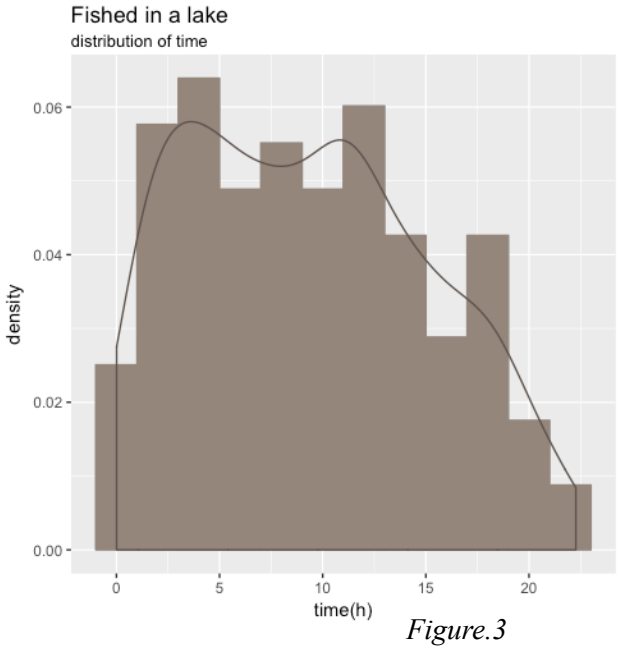

Figure.3

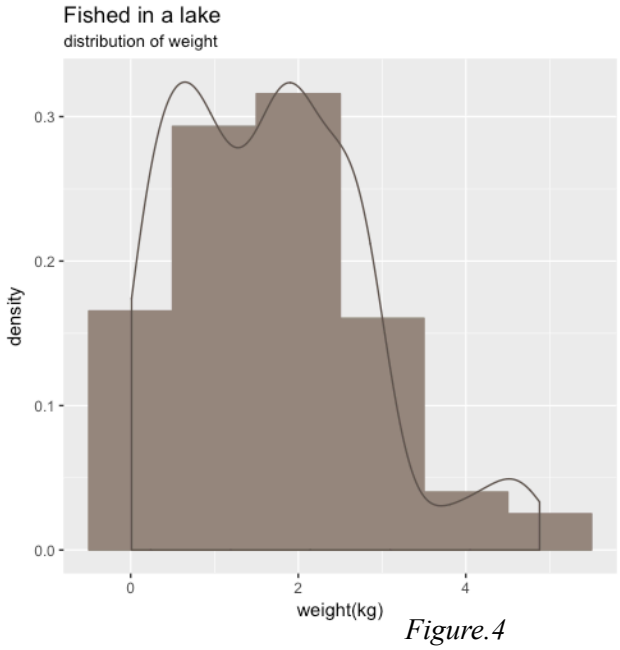

Figure.4

To be more precisely, I calculate some features about "time" and "weight" in R. The results are on the right in the Table.1. The weights of fishes are primarily between 0.705kg and 2.4kg, nearly same as the result got from *Figure.2(4)*. The time of most fishes caught is from 04:19 to 13:43.

Table.1

time		weight	
Min.	: 0.010	Min.	:0.010
1st Qu.	: 4.320	1st Qu.	:0.705
Median	: 8.960	Median	:1.610
Mean	: 9.356	Mean	:1.666
3rd Qu.	:13.715	3rd Qu.	:2.400
Max.	:22.270	Max.	:4.880

In the next step, assuming that the data set is just a sample of a larger population, to acquire a confidence interval about the mean value of the weight of fishes and time the fisherman made every caught in the larger population, I operated all of these in R. A 95% confidence interval of mean value of ' weight of fish' is from 1.556422 to 1.774706 with the mean value 1.665564. And the mean value of 'time of every caught' is 9.355789, with a 95% confidence interval between 8.785329 and 9.926250.

As said at the beginning, from the scatterplot, there are more points labeled C. To acquire more accurate information, I counted the number of points labeled A,B,C respectively. There are 79 points labeled A, 63 points labeled B and 257 points labeled C. The difference of effectiveness in 3 types of rods are really obvious. Bait C is preferred by most fishes. So in the next part, I am going to analyse 3 types of baits, such as their effectiveness in different time period and so on.

The figure on the right illustrates the effectiveness of 3 rods in one day. Rod A mainly works at the very beginning of one day, but after around 08:00 it is less effective. After about 05:00, rod B and rod C begin to perform much effective.

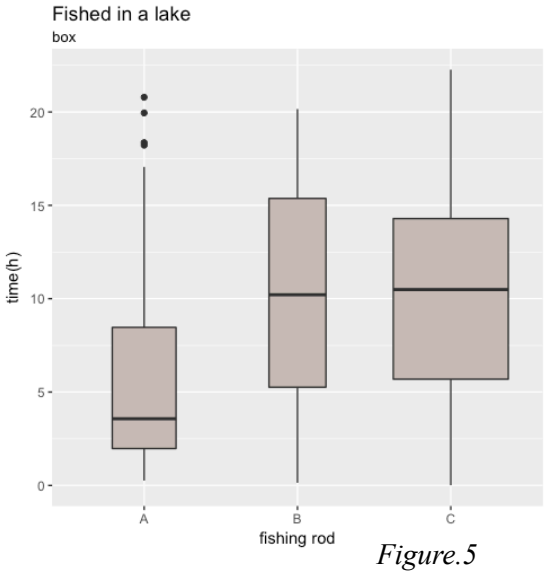

Figure.5

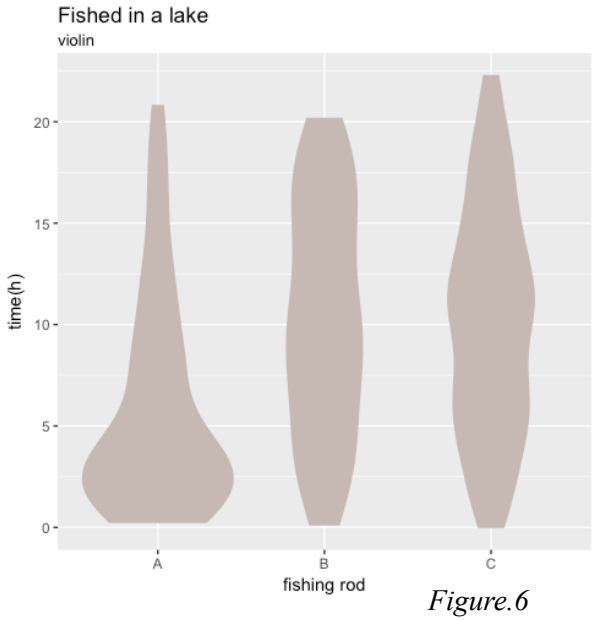

Figure.6

A more vivid image is on the left. In the *Figure. 6*, we can clearly see that rod A is better at the beginning, and the effectiveness of rod B is relatively balanced in one day. Rod C works from the beginning to the end and most effective in the midday. After comparing 3 rods in different time period, I am going to see the size of fishes caught by 3 rods.

The median value of weight in 3 types of rods A, B and C is 1.34, 1.78 and 1.66 respectively. The mean value of these three are 1.529, 1.774 and 1.681. According to these figures, rod A always has the smallest value. Rod B has a higher value but not too much distinguished from rod C. To be more easily perceived, *Figure.7* generates a density figure about weights. The weight of fishes caught by A is mainly around 1.5kg. And B and C perform better on catching

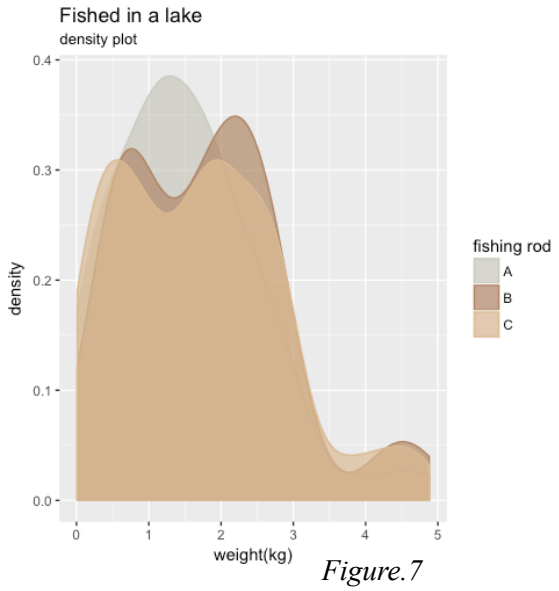

Figure.7

bigger size fishes. Then I calculated the variance of 3 rods, with 1.025, 1.176, and 1.306. The variance value of C is the biggest. So the weight of fishes caught by B tends to be much closer to the mean value. I infer that the reason why rod C has a slightly smaller value in median and mean than B is that the number of fishes caught by C is far larger than A and B, which also caused a bigger variance value. To be certain about what I thought, I generated two plots below. From *Figure.8*, at most time, the number of fishes caught by rod C is bigger than other 2 rods. And from *Figure.9* rod C always caught most fishes no matter they are bigger or smaller.

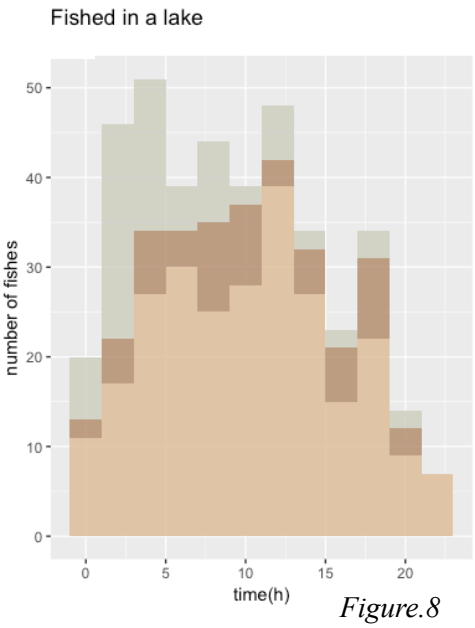

Figure.8

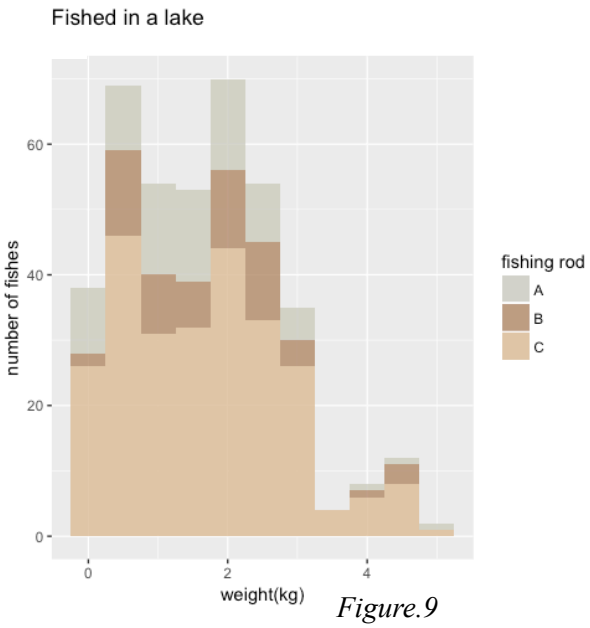

Figure.9

In conclusion, if we want to catch more fishes, bait C is relatively more effective. However, if we want to catch bigger size fishes without considering quantity, bait B would be a good choice as it has a bigger mean and medium value, and a smaller variance. And about good fishing time, if you want to catch bigger ones, you can go fishing before 05:00. And since most fishes were caught between 04:19 and 13:43, you can get more fishes at this period of time. At 3pm, though only bait B includes 3pm in the box at *Figure.5*; however, from *Figure.8*, we can clearly see at 15:00, rod C is able to catch most fishes. So if you want to fish at 3pm, I think bait C would be the best.