

**Comp6235 – Coursework for the statistics part**

Junming Zhang 29299527 [jz1g17@ecs.soton.ac.uk](mailto:jz1g17@ecs.soton.ac.uk)

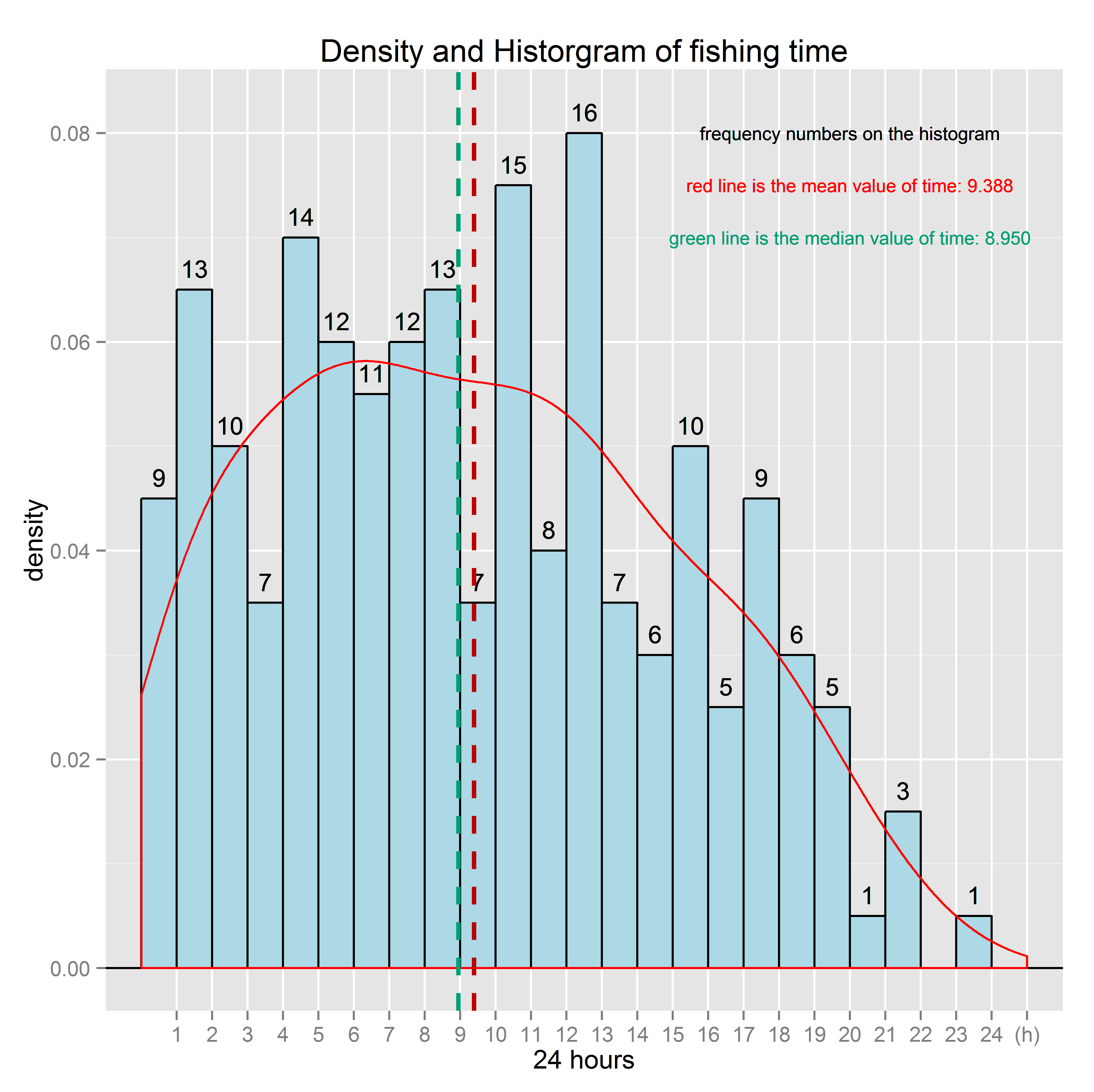
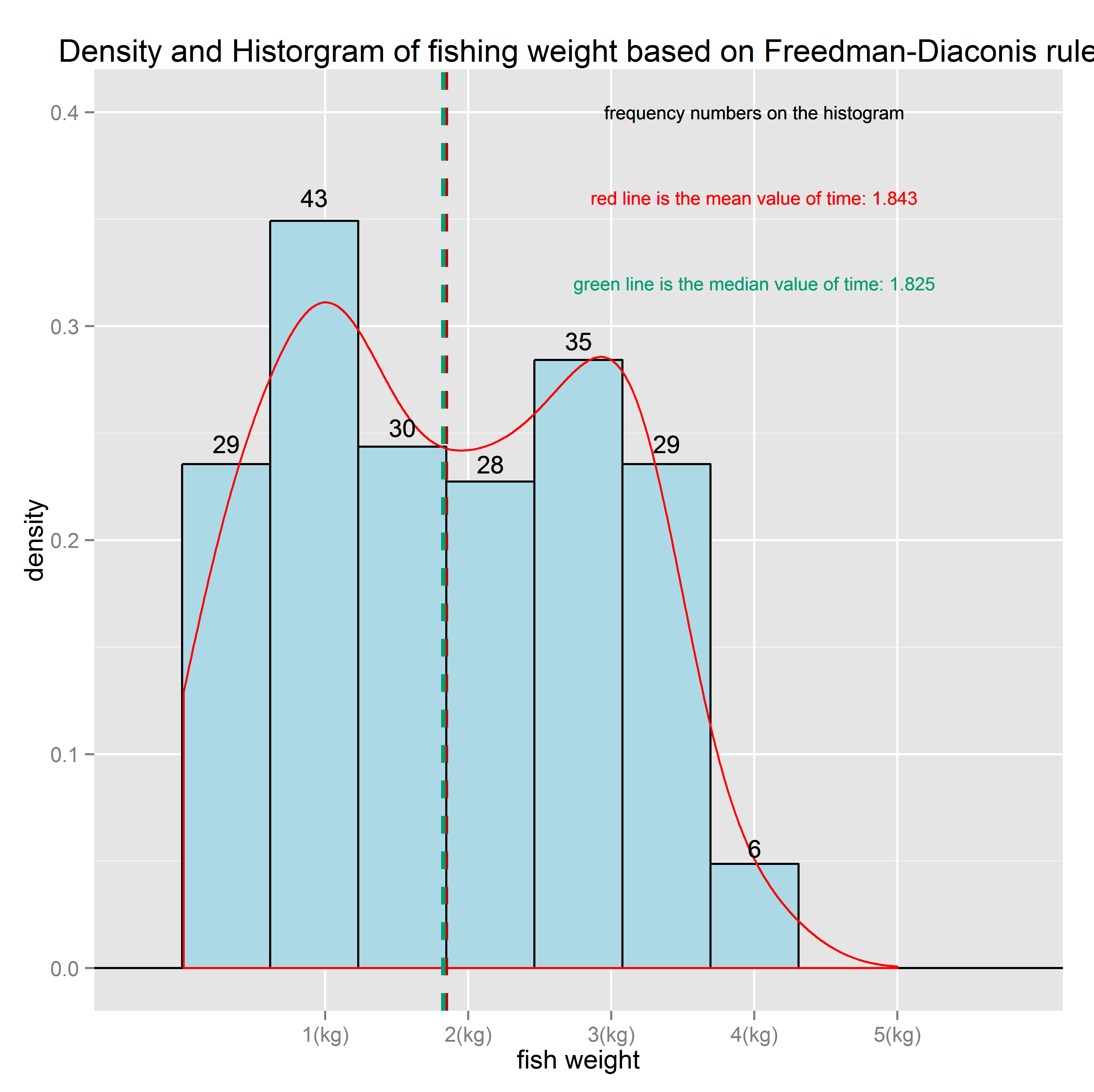
## 1. Method

1. Sort data based on ascending order of time and plot the distribution and linear fitting diagram.
2. Separate data to 24 independent group hour by hour.
3. Calculate the mean, standard deviation and confidence interval of separated data.
4. Plot the line chart between two variables based on processed data.
5. Plotting based on the ggplot2 library.

***Table 1. Processed data from fish.txt***

## 2. Result and Analysis

### 2.1 Character and generate graphs to illustrate the distribution of time and weight

*******Figure 1. Distribution of catching time Figure 2. Distribution of fish weight***

For catching time(unit-h):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Median | Mode | Variance | Geometric mean | Standard deviation |
| 9.388 | 8.950 | 12-13 | 31.998 | 6.786 | 5.657 |
| Skewness | Kurtosis | IQR | Bin size | Kernel density | 95% confidence interval of the mean |
| 0.250 | 2.134 | 8.335 | 1 | Gaussian | 8.600-10.177 |

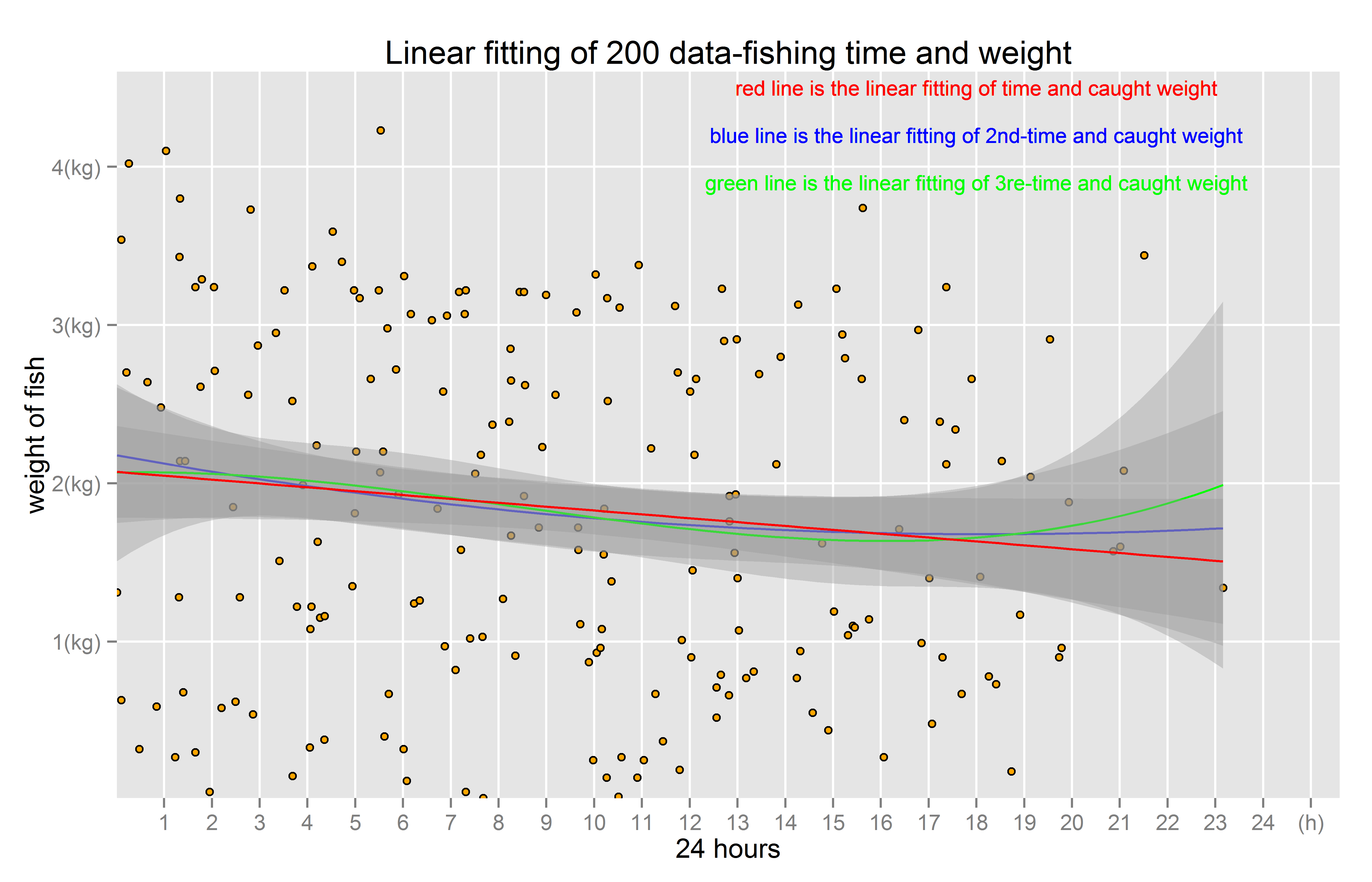
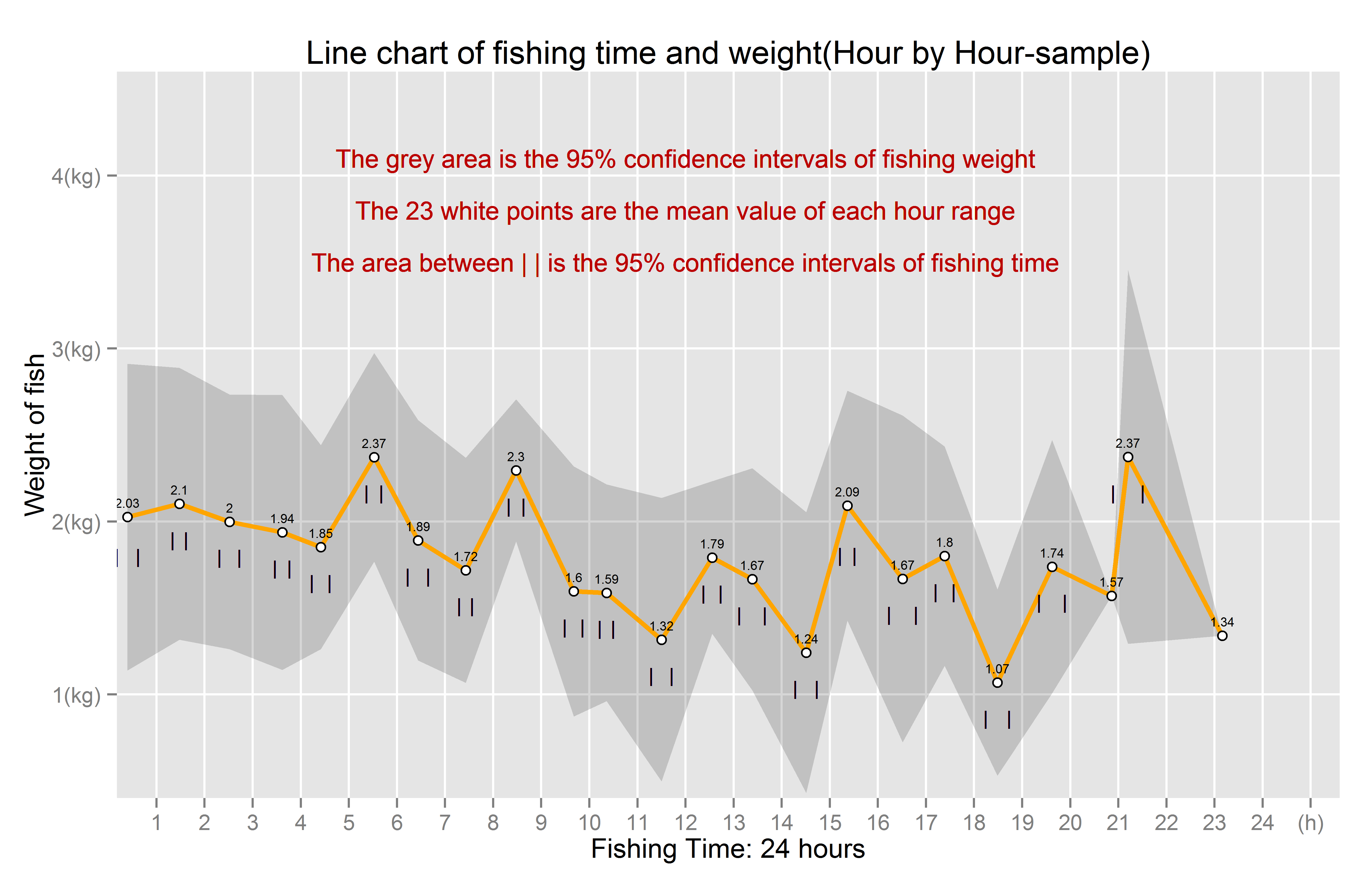
The bin size of catching time is 1 because the time value is within 24 hours.

For catching weight (unit-kg):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Median | Mode | Variance | Geometric mean | Standard deviation |
| 1.843 | 1.825 | 0.616-1.232 | 1.162 | 1.369 | 1.078 |
| Skewness | Kurtosis | IQR | Bin size | Kernel density | 95% confidence interval of the mean |
| 0.090 | 1.849 | 1.800 | 0.616 | Gaussian | 1.693-1.993 |

The bin size of catching weight is based on Freedman-Diaconis rule.

### 2.2 Analyze and generate graphs to illustrate the dependence between time and weight.

***Figure 3. The linear fitting of fishing data Figure 4. The line chart of sampled fishing data***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pearson metric | P value | Kendall metric | P value | Spearman metric | P value |
| -0.128 | 0.070 | -0.088 | 0.063 | -0.126 | 0.073 |

The p-value is not small than 0.05 which means the correction test may not correct. However, in most situation, it could confirm the fishing time and caught weight are independent due to the value of each correction test is close to 0.

Then apply linear fitting to the original data:

|  |  |  |  |
| --- | --- | --- | --- |
| lm1 | lm2 | lm3 | covariance |
| 599.329 | 600.893 | 602.529 | -0.782 |

lm1 has the smallest value which means the red line of Figure 3 is the best regression curve of them. It could find the caught weight decreases with the increasing fishing time.

### 2.3 Estimate the best time to go fishing.

The regression curve of Figure 3 indicates the caught weight decrease with the time. At the meanwhile, for obtaining accurate time to go fishing, Figure 4 gives the average fish caught hour by hour, it could find time between 5:00 and 6:00 has higher average number 2.37 and highest lower band of fish weight., Therefore the estimated best time to go fishing is 5:00~6:00.

## 3. R code, Diagrams and Processed data

<https://github.com/TROUBLEBRO/comp6235_1718_coursework-1>