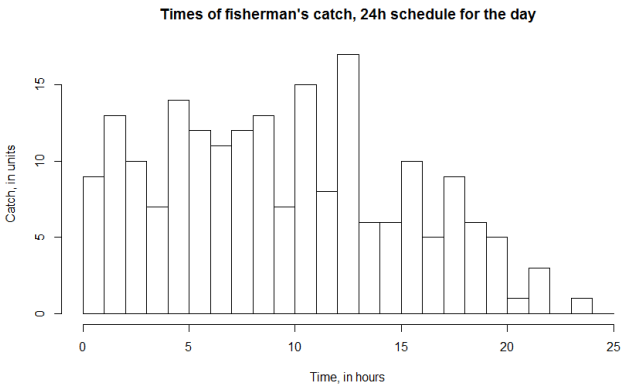
Foundations of Data Science COMP6235

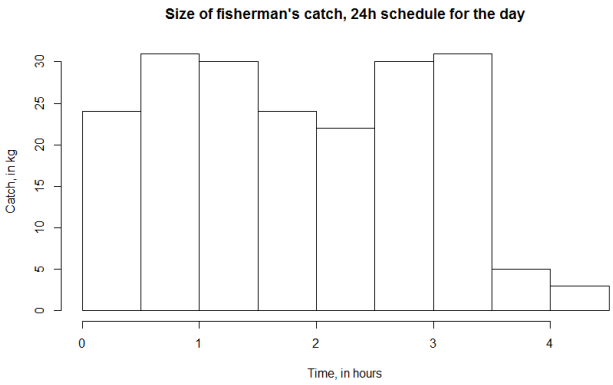
Statistics coursework. R language.

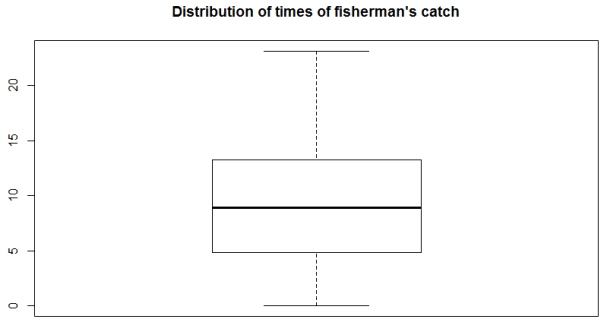
Yulia Efimova ID 28919025

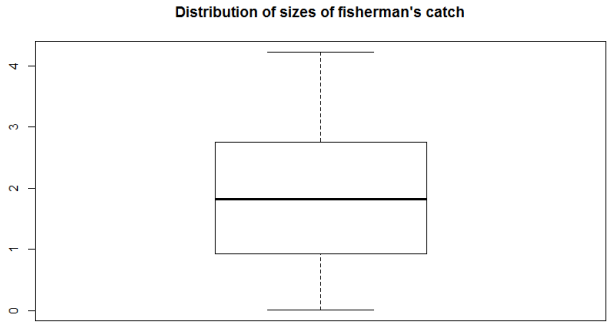
In the case of several datasets, huge and contentiously updated, the best way to analyze and process the data and present it graphically is to use R, software environment and programming language. Being available for Linux, Windows, and Mac OS, R is widely used for such fields as statistics, finance, insurance etc. The aim of this coursework is to demonstrate some features and functions of R language in the following case of the catch of fish during the day that provides the data about the time when the catch was made (X values) and the size of this catch (Y values).

First step









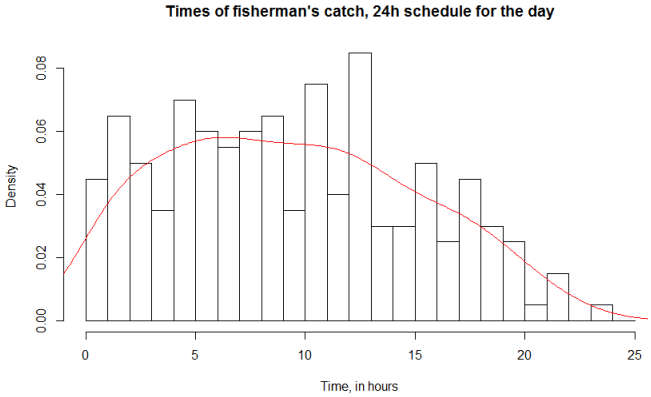


Image 1. X values (generated with R)

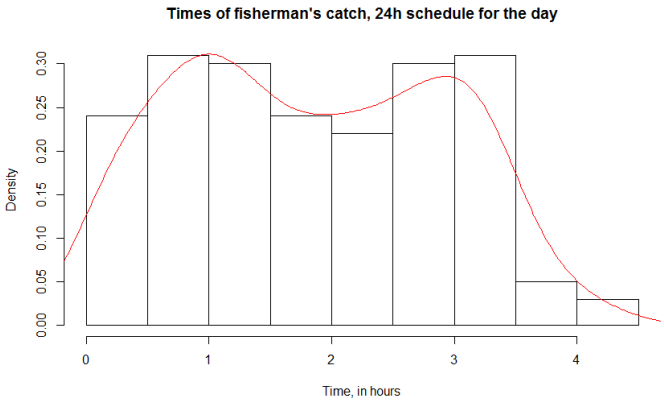


Image 2. Y values (generated with R)

First, to visualize the data histograms of both values were made – Image 1 and Image 2 on the previous page. As it shown on these plots, the biggest number of catches was between 5th and 13th hours with its peak at 13th hour and a lack of tries at 23rd hour. Meanwhile, in Y distribution there is a sharp decline in the end, that means a decrease of performance. However, it has slight slack in the middle.

More precise measures are provided in the table below.

	X	Y
Minimum	0.010	0.0100
Median	8.950	1.8250
Mean	9.388	1.8431
Maximum	23.160	4.2300

According to it, on average, fisherman makes at least 9 catches per hour with almost 2 kilos of fish (also per hour), while in a lucky hour the fisherman makes at least 23 catches with 4.2 kilos in total. The minimum value so small so it can be neglected.

Second step

Scatterplot was considered as the best way to demonstrate the relationship between two variables.

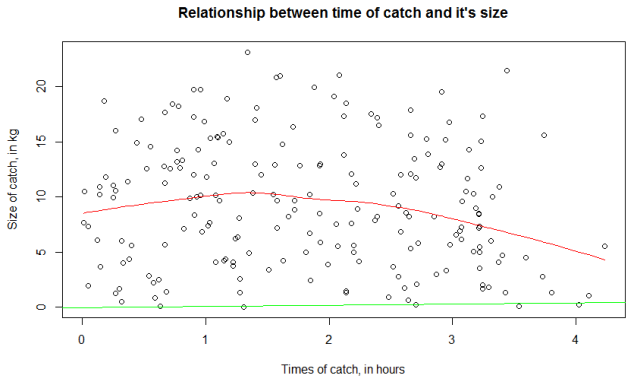


Image 3. The relationship between variables, probability density (red) and locally-weighted regression line (green)

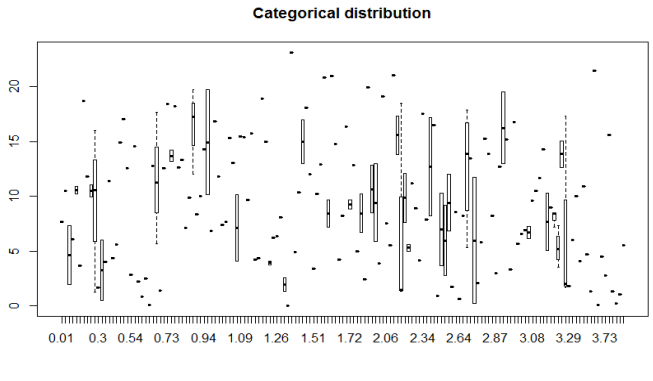


Image 4. Boxplots of distribution of both variables

As it is seen the probability density smoothly declines after approximately 1.4 hours reaching 10 kilos (in size). Moreover, the correlation between both measures is equal -0.1282133 (negative) and the covariance is -0.7818909 (also negative). This means the more tries the fisherman makes the less size of catch he gets. And finally, according to Image 4, it can be said when the fisherman makes his lowest catch and when – his highest.

R language is one of the most efficient tools used to simplify and automatize processes with big data. It allows reducing sources spent on data processing and especially the time.