Abstract

Data is meaningless unless its conversion into valuable information. The critical thing of Data Science is to know the value of data, find the data is insight, and apply it to the real world. Nowadays, Python is one of the most potent computer languages in the Computer Science world. Not only effective in integration systems and easy to code in terms of data analysis and construct distribution data, but also there are a lot of libraries for choosing to use, such as matplotlib, Pandas, and Seaborn. matplotlib is a library for data visualisation and graphical plotting, which is very helpful for Data Science Pandas library is easy, flexible, and fast for data analysis, and Seaborn is a library to makes statistical and visualization data. Overall, Data scientists use Python and R for data preparation and statistical analysis. Python used for general purposes is more readable, more straightforward, and offers more flexibility while learning.

Introduction

In this report, Python is the primary language to explore insight and generate graphs. The given data set is the record time interval on one day when a fisherman has fished by using three types of Bait in each fishing rod by labelling it with A, B, and C. the data set consists of three columns which is the time that fish caught into the first column of the data set(X), recorded the weight of fishes in the second column of the data set(Y), and for the third column of the data set(Z) is recording the Bait that fishes caught. The description of this data set is a sample from a larger population and gives mean values with 95 percent confidence intervals.

Methods and Result

The first step of methods is to explore the given data by using describe and generate a plot illustrates of mode function in pandas library of Python that shows preliminary characterise and measures of centrality, spread, and suitable additional measures of data set in the form of numeric for X (Time(Hr)) Y (Weight(Kg)) and Z (Bait) Values. The numeric of describing, Mode, and correlation of data set is in Fig.1 Characterise of data, Fig.2 represent Mode of the dataset by mode1 is order the numbers highest and see which number appears the most often,mode2 and mode3 are the highest number as mode1 in different values and Fig.3 Correlation of the dataset

	X	Y
count	400.000000	400.00000
mean	9.370525	1.66740
std	5.796400	1.10816
min	0.010000	0.01000
Q1(25%)	4.325000	0.70750
median(50%)	9.020000	1.61500
Q3(75%)	13.747500	2.40000
max	22.270000	4.88000

Figure 1: Characterise of data set by describing function

	X	Y	Z
mode1	0.93	0.11	C
mode2	3.11	0.70	NaN
mode3	4.99	0.87	NaN

Figure 2: Mode of data set by mode function

	Times(hr.)	Weight(kg.)
Times(hr.)	1.000000	-0.120593
Weight(kg.)	-0.120593	1.000000

Figure 3: Correlation of the dataset by the correlation function

In the second step, the data set has shown more insight by using a library of Pandas, Seaborn, NumPy, Scipy, and Matplotlib through various types of the histogram. First is a set histogram of times the fisherman has caught fish within 24 hours by different bins. Fig.4 is a histogram by using containers 133 from the Freeman-Diaconis rule [1]

$$Bin\ width = 2\frac{IQR(x)}{\sqrt[3]{n}}$$

This histogram shows detail of data more appropriate and range of time 01:00 to 4:00 and 11:00 to 12:00 fish catch is raised very high. Its peaks at 11:00 to 12:00 and sparse of fish catch from 21:00 to 00:00.

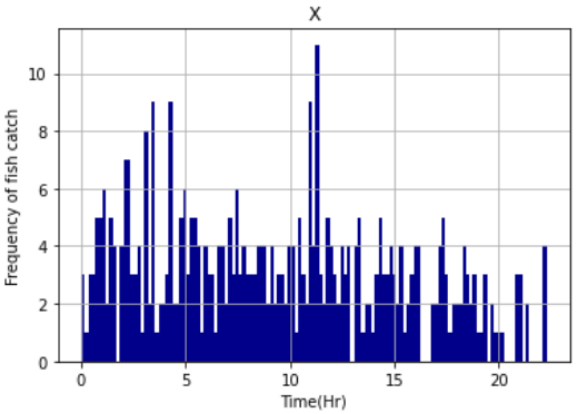

Figure 4: The histogram of time that Fishman made a caught fish with 133 bins

Next is the histogram of the weight of fishes that the fisherman has made a catch, as shown in Fig 5. This histogram uses bins = 400 because of the Freeman-Diaconis rule, same as Fig4, and data is a high spread data rate. From histogram show that fish size 2 kg is a very peak size and interval of 3 kg to 4.8 kg is very rare

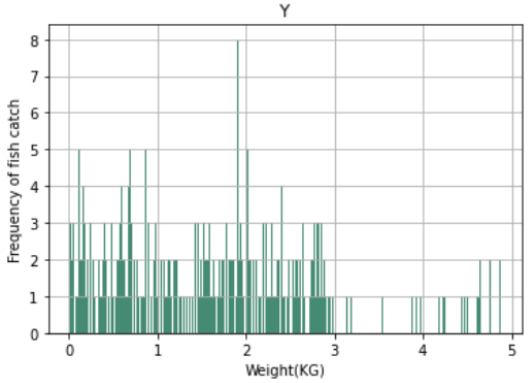

Figure 5: The histogram of the weight of fish that fisherman made caught with 400 bins

The 95 percent Confidence Intervals [2] for time and weight distribution and Population Mean are presented in Figures 6, 7, 8, and 9, respectively. First, the 95% confidence interval for times distribution is -1.9904 to 20.7314. and confidence interval for a population mean of time is 8.8025 to 9.9385. On the other hand, the 95% confidence interval for weight distribution is -0.5045 to 3.8394. and confidence interval for a population mean of weight is 1.5584 to 1.7764. a use equation computes the 95 percent, confident intervals $\bar{x} \pm 1.96$. The 95% confidence interval is a range of values that can be 95% confident contains the population's actual mean.

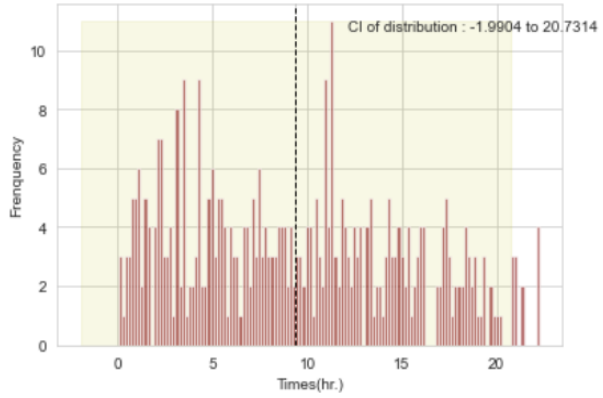

Figure 6: The 95 percent Confidence Intervals for Time(X) Distribution

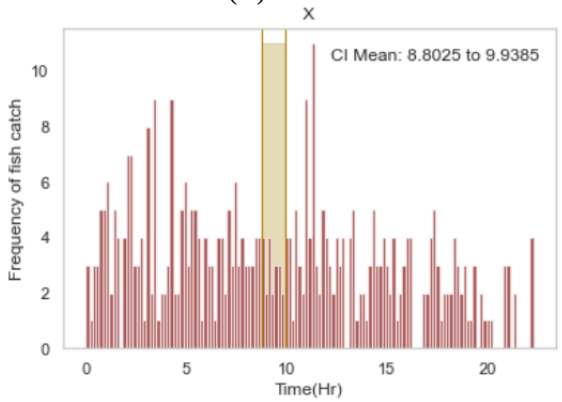

Figure 7: The 95 percent Confidence Intervals for Time(X) Population Mean

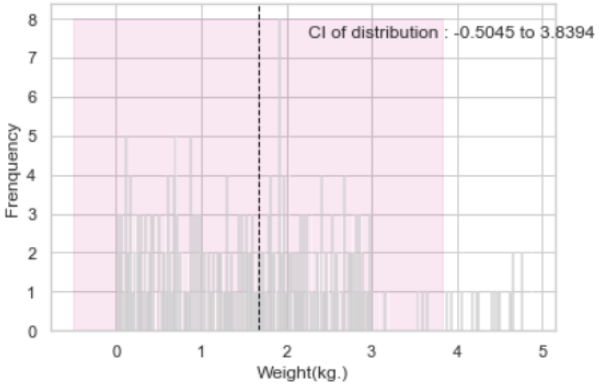

Figure 8: The 95 percent Confidence Intervals for Weight(Y) Distribution

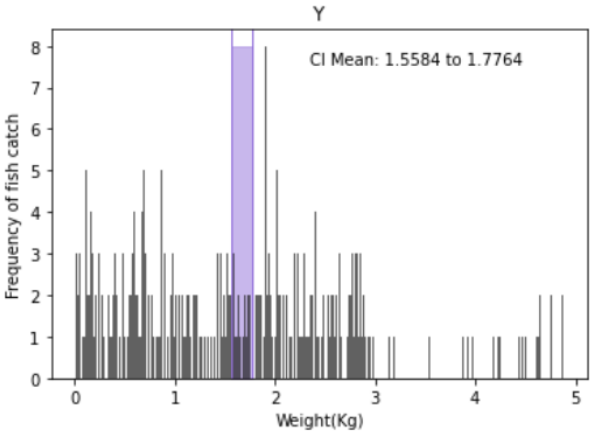

Figure 9: The 95 percent Confidence Intervals for Weight(Y) Population Mean

And the next is a histogram for types of Bait (Z) shown in Fig 10 in terms of frequency, what kind of Bait is most effective, which is categories C used to fish 257 fish and the lowest is category B 64 fish, and category A is 79 in the middle

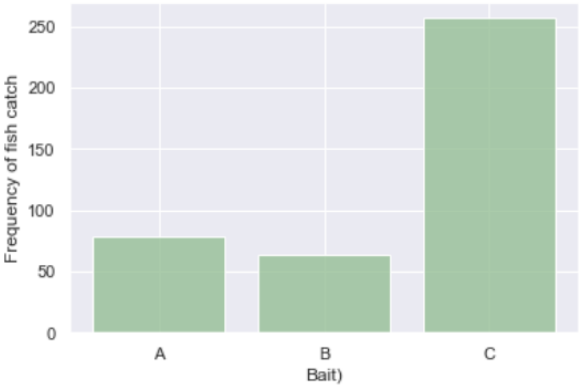

Figure 10: the histogram of categories of Bait (Z)

Then, apply the time and weight of fish caught to Kernel density estimation (KED) [3] shown in Fig 11, Fig. 12, respectively It illustrates ti see the smooth curve of the density distribution of data and visualise continuous histogram

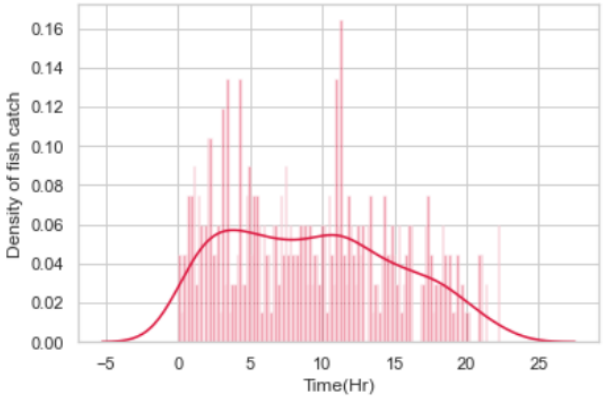

Figure 11: Kernal density estimate Time(Hr) of fish caught

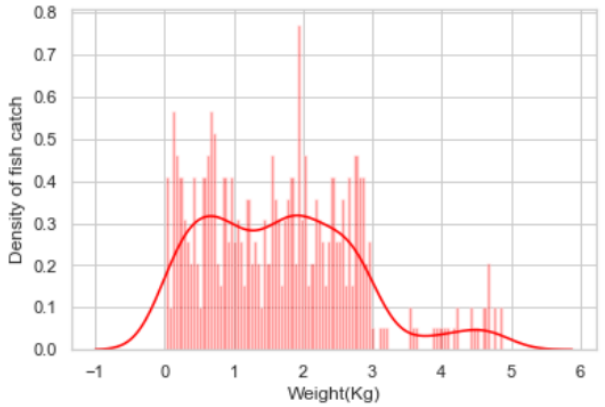

Figure 12: Kernal density estimate Weight(Kg) of fish caught

The next part is plot the histogram by Seaborn for made charts in terms of Scatter plot of the dataset that has Kernel Density estimate of time, weight in each Bait by on the x-axis represents Time(Hr), and the y-axis represents Weight(Kg) for each type of Bait shown in Fig 13 and Fig 14, From the graph is show that Bait type A has the best performance in the morning, type B is dispersed, and type C has the best performance in the noon

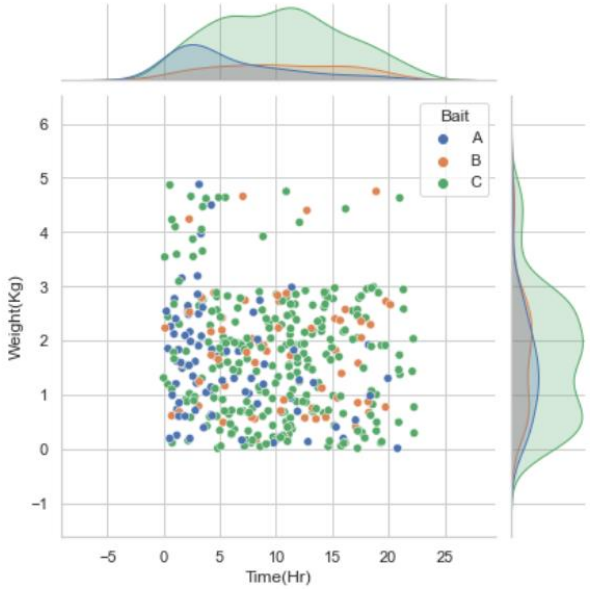

Figure 13: KED and Scatter plot of data set in terms of Bait (Z)

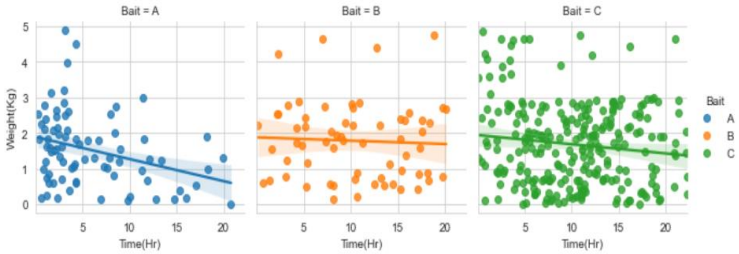

Figure 14: KED and Scatter plot of data set in terms of each Bait (Z) separately with Liner regression

Discussion

The discussion of the main point and answer the question in the coursework.

1. What is the Correlation between X and Y?

Fig 15. Show the correlation between X and Y and represent the relationship amount of the dataset by building to check Pearson's coefficient [4] that $r = -0.12$ and visualise linear regression of correlation between X and Y. This means the variable changes in the opposite direction. So terms of the graph show that X and Y are common inversions. It can cause by variables in a dataset that decrease the correlation of information

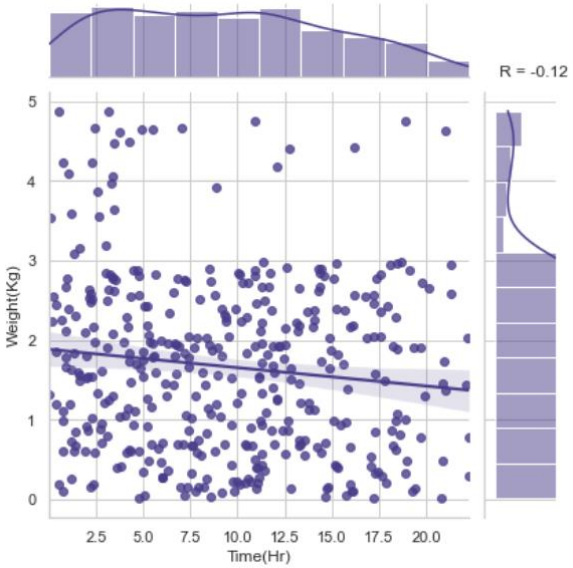

Figure 15: The Scatter and histogram of Linear regression and correlation of weight and time

2. What is the best time to go fishing at this lake?

For this question, in my opinion, in terms of the best time of fishing at this lake. It is not only focused on frequency but the weight (size) of fish also. So, according to Fig 16. The statistic of the data set shows that the maximum recorded time is 22.27, which means there is no recorded data at an hour of 23:00. The correct number of bins to represent the exact fraction of hours should be 23. It shows that the best fishing time in terms of frequency is 11.00-12.00 and, if needed, it is more especially a time of fishing shown in Fig 17. that When looking at the 11:00 box plot, it can see Q1 is at 0.26 (11:16), and Q3 is at 0.74 (11:45) that at 11.16 to 11:45 have 50 percent of frequency at 11:00 to 12:00. So, 11:16 to 11:45 is the best time to go fishing and refer to Fig 18. Illustrate that in terms of weight, the best time for fishing the most

significant fish is 3.00-4.00, but If judging by weight and frequency together, 11.00-12.00 is the best time for fishing in this lake

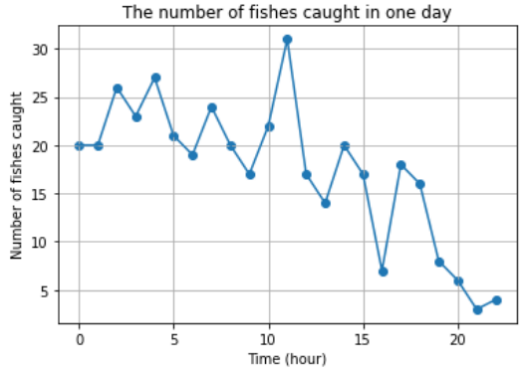

Figure 16: Frequency of fish caught in each hour

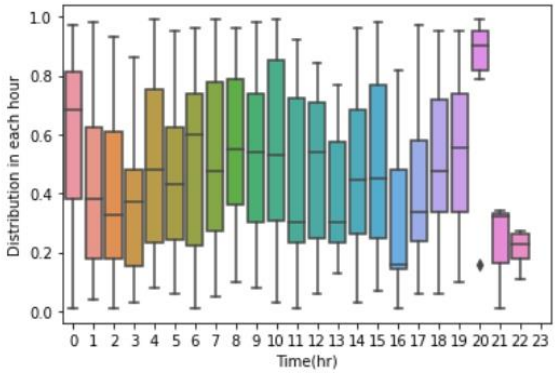

Figure 17: Distribution of time of fish caught in each hour by box plot

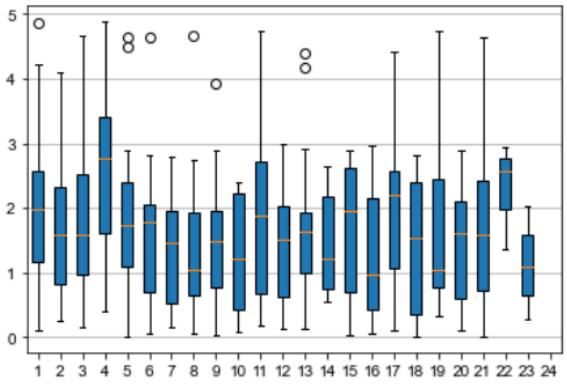

Figure 18: Distribution of weight of fish caught in each hour by box plot

3. Which Bait is most effective

To find the most effective Bait, it is seen from Fig 14. And to analyse the effectiveness of each type, it is illustrated by Fig 19 in terms of times and Fig 20 in terms of weight. In Fig 19, one can see the significance of times in each Bait. Bait type A can many caught fishes from 01:00 to 05:00, but it wasn't performing too much to catch fish at other times. Bait type B can be used to catch fish every time and stable, but the performance doesn't too much as bait type C that the most performance almost every time. Next, looking at Fig. 20 shows effectiveness in each Bait in the aspect of weight. Bait type A and B look similar to the efficacy of consequences, and bait type C looks to have the best performance. So bait type C can catch more fish than others, making this bait special. Thus, after analyses, the effectiveness of each type of Bait can see that bait type C is the most effective.

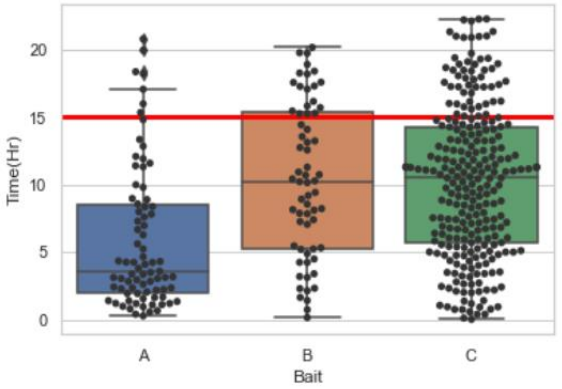

Figure 19: The Box plot and Swarm plot of times in each Bait with 3 pm. Line indicator

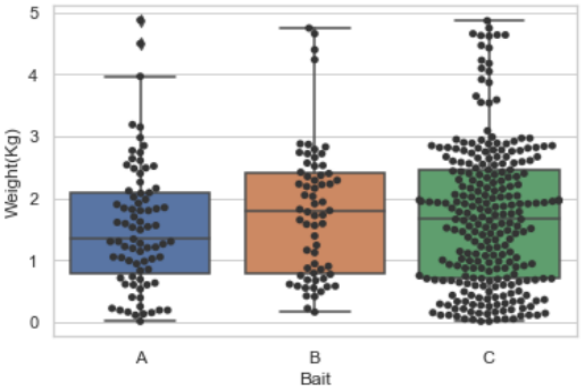

Figure 20: The Box plot and Swarm plot of weight in each Bait

4. What is the best type of Bait to use at 3 pm in the afternoon

From Fig 19, the Box plot illustrates which Bait is the most effective by showing the red line of 3 pm (15.00) of mark the Time(Hr), and it indicates that Bait category C gains the highest performance and effectiveness of all Bait

Conclusion

This data set is an excellent material to practice and explore the data. The dataset requires several libraries of Python to do data analytics and practical statistical knowledge to do the process. It makes this dataset is an excellent example for a data analyses

References

[1] Determining Histogram Bin Width using the Freedman-Diaconis Rule Available: http://www.jtrive.com/determining-histogram-bin-width-using-the-freedman-diaconis-rule.html [Accessed November, 20 2021].

[2] The 95% Confidence Intervals estimate – SPH Available: https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_confidence_intervals/bs704_confidence_intervals_print.html [Accessed November, 20 2021].

[3] Kernel density estimators. Available:https://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0405/MISHRA/kde.html[Accessed November 20 2021]

[4] Pearson correlation coefficient Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient [Accessed November 20 2021].