Foundations of Data Science: Assignment Report

Adeeba Mahmood
Department of Computer Science
University of Southampton
am8n22@soton.ac.uk

1 Introduction

This report contains the exploratory data analysis of the fishing data, about the catch of a hypothetical fishing fleet. Exploratory data analysis is the initial investigation of the dataset to discover patterns, find the outliers and to check assumptions using statistical methods and Graphical representation.

Fishing dataset consists of 3 features(columns), they are as follows: X represents the time of the catch, Y represents the size of the catch and Z represents the type of the bait. Overall there are a total of 400 data points. You can see the details about the data in Figure 1. The analysis has to be done in python and its respective packages.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   X       400 non-null    float64
 1   Y       400 non-null    float64
 2   Z       400 non-null    object
dtypes: float64(2), object(1)
memory usage: 9.5+ KB
```

Figure 1: Represent the details of the dataset

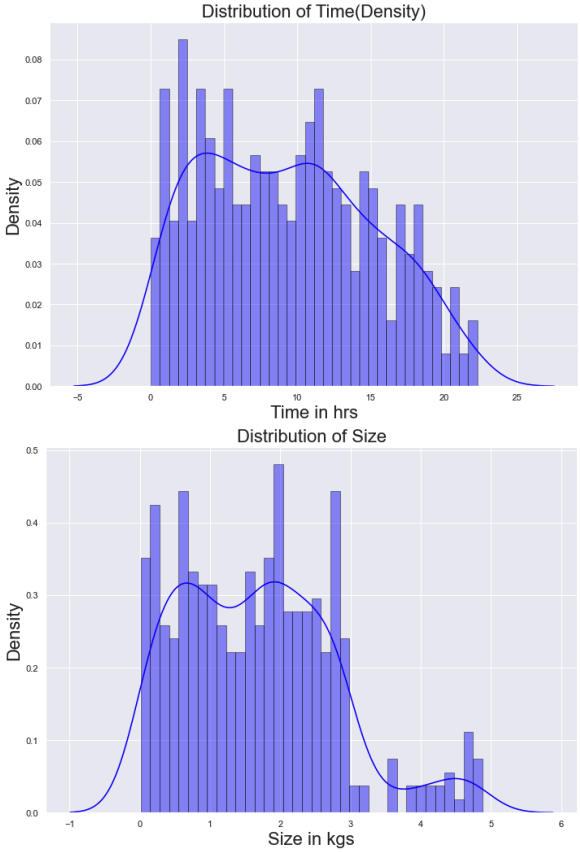

Figure 2: Distribution of Size and Time

2 Methodology

To analyze the dataset, the distribution of the different features need to be plotted. Then for each distribution the centrality and spread needs to be calculated to understand the shape of each distribution.

Centrality is the measure of mean, medium or in some cases mode while the spread is the measure of the variation calculated using range and standard deviation.

Mean of a data is sum of the values/number of a values. Median is the middle value of the data when ordered in from least to greatest. Mode is the value that has the maximum number of occurrences in the dataset.

Also explain the spread and range

To calculate the above value made use of pandas and statistics library in python and for the supporting result, graphs were plotted using matplotlib

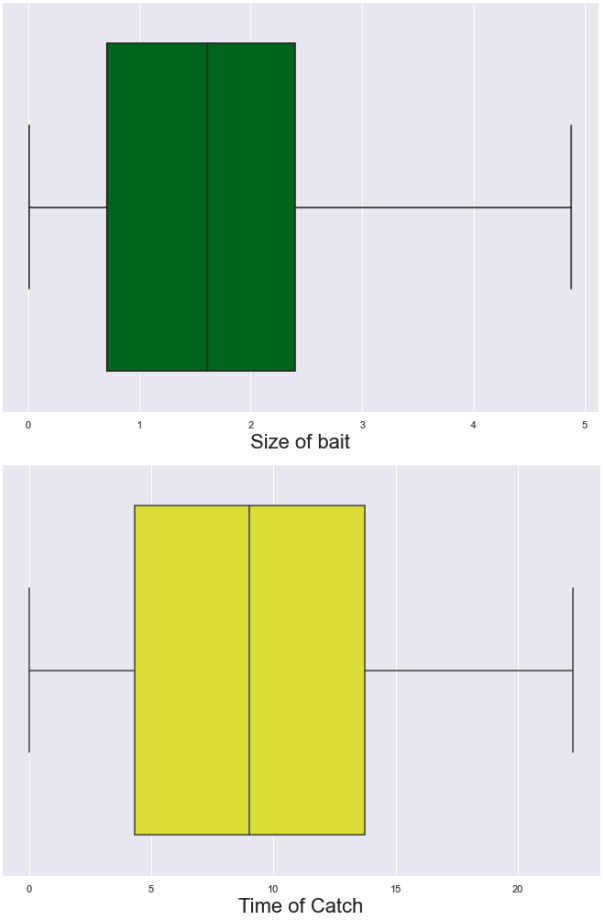

Figure 3: Box-plot to shows the spread of time of catch and bait

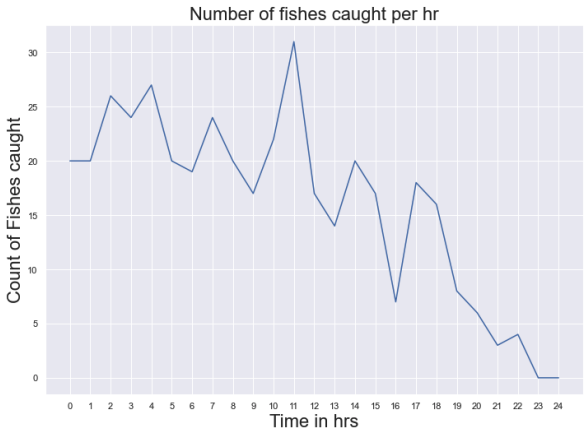

Figure 4: Number of the fishes caught per hr

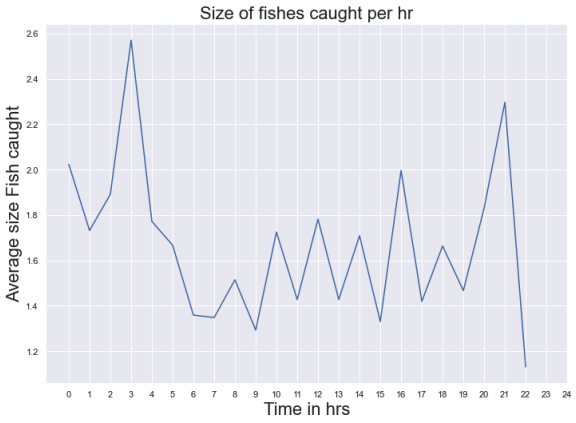

Figure 5: Size of the fishes caught per hr

and seaborn libraries.

3 Discussion and Analysis

Discussion about the centrality spread and shape of the curve

The dataset provided is time sensitive and it depends on the type of bait used. Since the data is time sensitive and it lasts for a whole day, the dataset can be divided into 24 hr periods. By generating a per hr count of the number of fishes caught, one can observe the trend for fishing and can find the best hr for fishing. As you can see in figure 4 the graph moves side-ward peaking at 11 am and then it starts creating lower highs. This shows that the best time for fishing is before 11 am in the morning and after the number of fishes that are being caught goes down.

In figure 5, the highest weighted fishes are caught in the morning at around 3 am. Then it drops around 1.4 to 2.0 kgs for about 9 to 10 hr.

By looking at both the graph, one can say that when there are certain heavy fishes that can be found only in during the day time, also the amount of fishes caught at that time would be also be high. By looking at both the graph, one can say that when there are certain heavy fishes that can be found only in during the day time, also the amount of fishes caught at that time would be also be high.

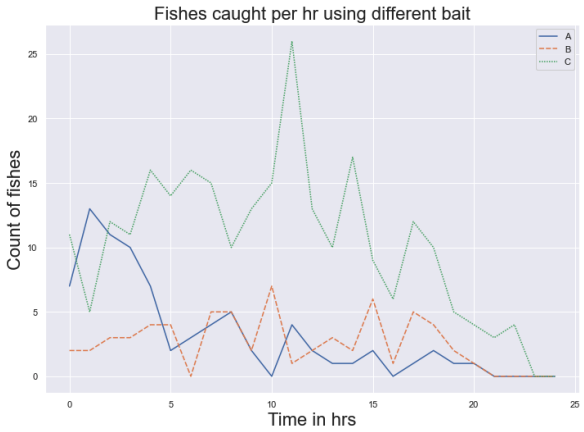

Figure 6: Fishes caught per hr using different bait

Figure 6 and 7, shows the comparison of the type of bait over hr and total respectively. After

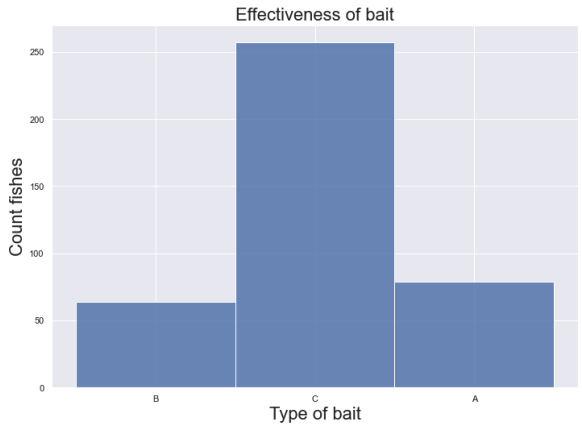

Figure 7: Total number of fishes caught

looking at the graphs, one can conclude that type "C" bait is the best bait. Further looking into figure 6 we observe that during early morning (before 3 am) A type of bait performs better than others. Also looking at Figure 9 we can say that "B" Type bait didn't catch the maximum number of fishes but it caught highest weighted fishes.

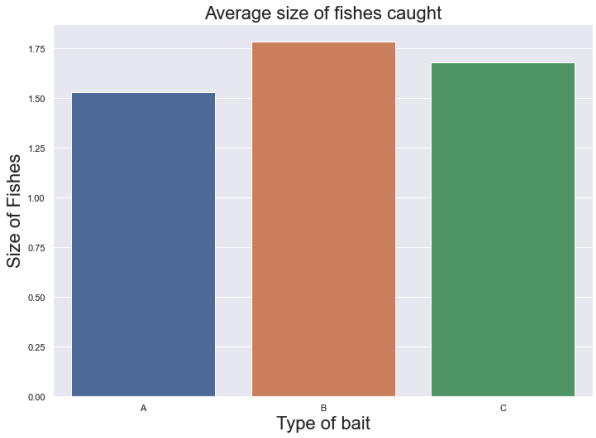

Figure 8: Effectiveness of bait in catching fishes

Correlation between Time and Size is negative(-0.12) which means ¡To be continued¿

4 Conclusions and Future Work

After observing the data and the graph we can conclude by saying as follows:

- Best time to go fishing is around 11 am as we catch the number of fishes at that time.
- "C" Type bait is the best bait.
- Best type of bait after 3 pm afternoon is "C"

Reference

[1] https://pandas.pydata.org/docs/index.html.
[2] https://seaborn.pydata.org/.
[3] https://docs.python.org/3/library/statistics.html.
[4] https://matplotlib.org/stable/index.html.
[5] https://www.hackerearth.com/blog/developers/descriptive-statistics-python-numpy/.