Harika Yadavalli

hy3u19@soton.ac.uk

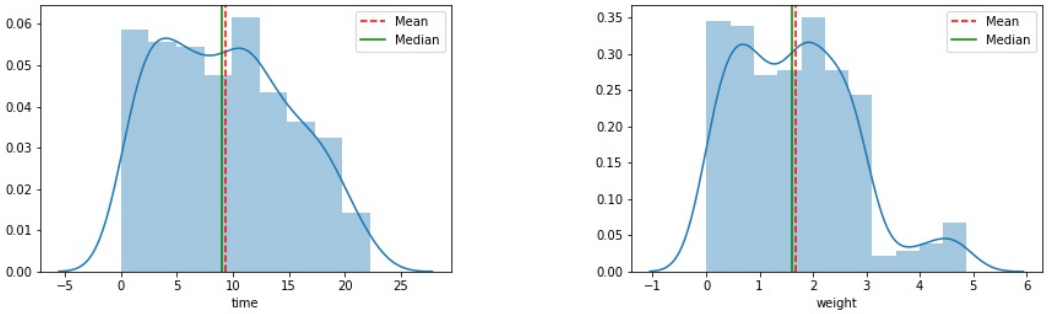
# 1 Introduction to Data set:

**We have been given fishing data set, which consists of 3 pieces(features) of information(data):**
Time of the day, Weight of the Fish and the Bait used

# 2 Explore the data distributions:

## 2.1 Time and Weight data distribution:

Let us plot how the Time and Weight data is distributed and note down the observations in the table for easy reference.




(a) Time distribution (b) Weight distribution

Figure 1: Data Distribution

| **Time distribution** | **Weight distribution** |
|---|---|
| Time data doesn't look like a perfect symmetric normal distribution. | Weight distribution also doesn't look perfect normal distribution. |
| **Mean and Standard deviation:**<br>$\mu$ =9.370525<br>$\sigma$ = 5.789150410412136 | **Mean and Standard deviation:**<br>$\mu$ =1.6674<br>$\sigma$ = 1.10677 |
| **Confidence intervals**= 95%<br>$\mu$= **[8.800759907346904, 9.940290092653097]**<br><br>$\mu = \mu \pm Z * (\sigma/\sqrt{n})$, Use Z-table to get Z-score for 95% confidence interval<br>Used Python package: statsmodels.stats.api | **Confidence intervals**= 95%<br>$\mu$=**[1.558471935919, 1.7763280640809]** |
| **Skewness and Kurtosis:**<br>Skewed to the left, so the Median is less than Mean of the distribution. Kurtosis is not very significant. We can find using:<br><br>`data["time"].skew() => 0.26685`<br>`data["time"].kurtosis() => -0.94659` | **Skewness** > 0.5 ( which is significant) to the left, so the Median is less than Mean of the distribution. Kurtosis is not very significant.<br><br>`data["weight"].skew() => 0.65379`<br>`data["weight"].kurtosis() => 0.16189` |

## 2.2 Effectiveness of Bait:

3 baits are available for fishing: A,B or C

We can say bait is effective if it can catch more fish i.e compare the mean weights for each bait and compare against each other.

```
sns.barplot(x="bait", y="weight", data=time_weight_bait_data, ci=95, order=["A", "B","C"],capsize=.2)
```

**B looks to to be the most effective bait**

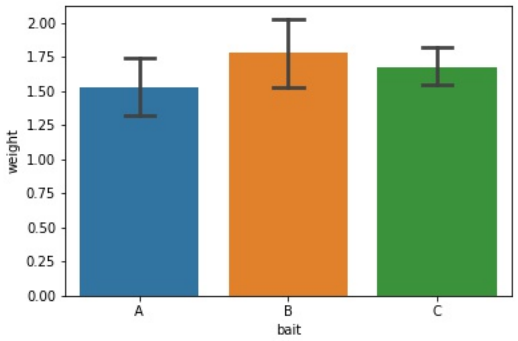


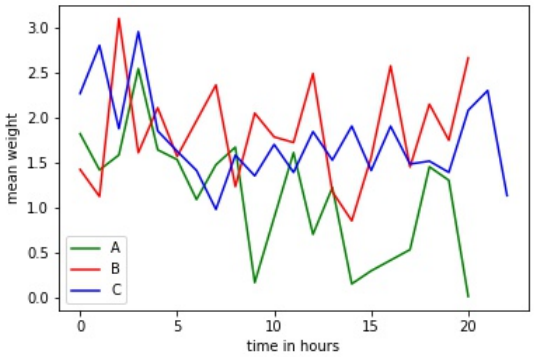

Figure 2: **Bait Vs Mean weights**

# 3 Correlation between the features:

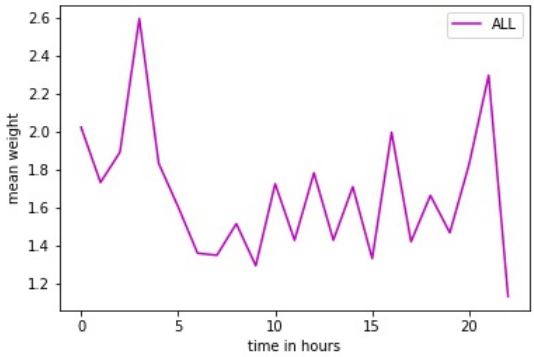
## 3.1 Dependency b/w time, weight and bait:

Let us plot line plot which shows average weight for each hour per each bait and compare against each other.
Sample Code for All baits:

```
temp =time_weight_bait_data.groupby(time_weight_bait_data["time"].astype(int))["weight"].mean()
sns.lineplot(data=temp, color='g',  label="A")
```




(a) Compare for each bait




(b) Combined Result

Figure 3: hour of the day Vs mean weight

**Observations:**

- In the figure(a), we have plotted mean weight during the hour of the day for each bait.
- We clearly see that there is more mean weight for bait A at hour 3am where for B and C at hour 4am
- So it looks like catching more mean fish weight is high b/w 3am to 4am irrespective of the bait chosen.
- **3am to 4am is the best time to go for fishing to catch more fish**

### 3.2 Correlation b/w Time and Weight:

Let us plot seaborn jointplot to see the correlation b/w time and weight

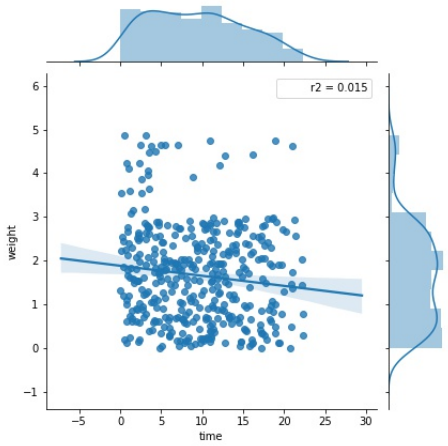



Figure 4: **Correlation: Time Vs Weight**

**Observations:**

- Calculate correlation coefficient r and it's importance i.e P value.

```
stats.pearsonr(time_weight_bait_data["time"],time_weight_bait_data["weight"])
```

- r value is -ve, this infers that with the time of the day the weight of the fish reduces slightly as shown in the figure and the importance is given by p-value which is also not high or considerable.
- We can use pairplot to see the correlation b/w the data distributions

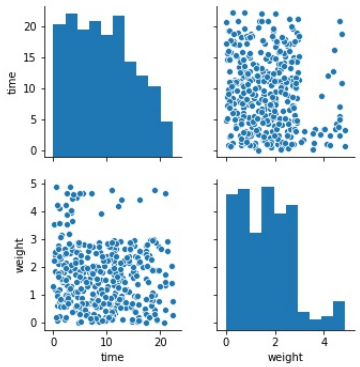



Figure 5: **Pairplot: Time Vs Weight**

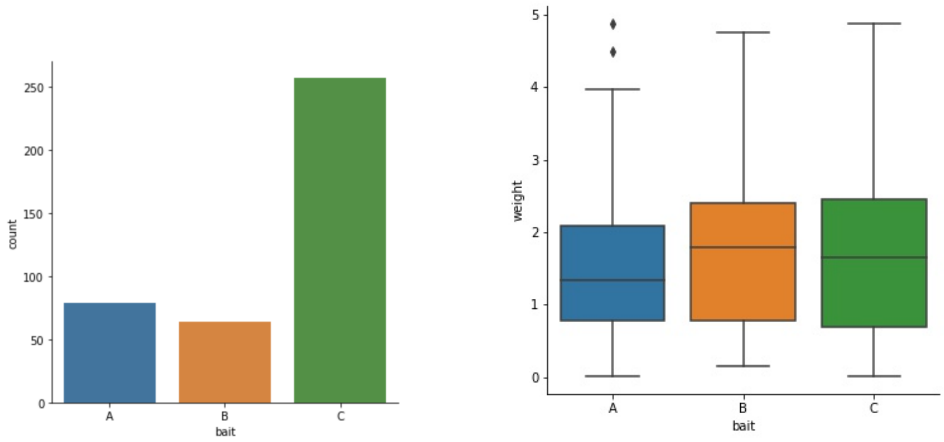
# 4 Questions:

## 4.1 What is the best time to go fishing at this lake?

- **3am to 4am is the best time to go for fishing to catch more fish**, please see the detailed explanation in section 3.1
- Even though this is away from Mean and Median, we have more data available between 3am to 4am from the time distribution on histogram as shown in Figure: 1a
- The correlation factor b/w time and weight certifies this result, that it is more weight during the starting of the day and weight decrease with the time period of the day as shown in Figure : 4

## 4.2 Which bait is most effective?

- Bait B looks most efficient bait when looking at the mean weight values for each bait as shown in figure : 2.
- Let us look at the usage of baits and compare against respective boxplots for weight to find the most efficient one.
- Even though C is most used bait, B is the most efficient one to catch more fish when comparing the mean values for respective weights as shown below in Figure 6b



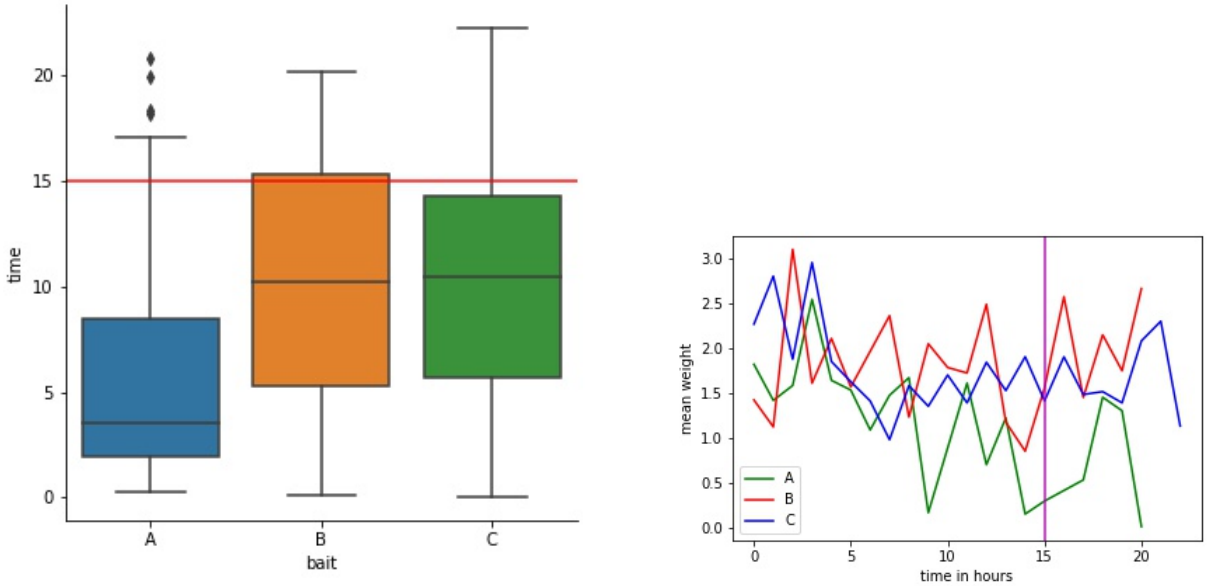

(a) Compare for each bait

(b) Combined Result

Figure 6: hour of the day Vs mean weight

## 4.3 What is the best type of bait to use at 3pm in the afternoon?

- **B is the best bait to use at 3pm** as it has got more confidence levels at as it falls into 50% range of the B boxplot at 3pm as shown below in the Figure : 7a
- Also the mean weight captured at 3pm is more for Bait B and C is the next one to choose after B.




(a) Compare for each bait

(b) Combined Result

Figure 7: Weight boxplot per each Bait