PLAKSHA UNIVERSITY
DATASCIENCE
USMAN AKINYEMI
ASSIGNMENT – 7

Question 2_1

1. Date int64
This column represents the experimental day, ranging from 1 onwards, indicating the day after stress treatments. It is both numerical and ordinal because it signifies a sequence or order of days.

2. Vial int64
This column contains the individual ID for each fly. It ranges from 1 to 600
It is a numeric type of data and is ordinal as it represents a specific order or sequence.

3. Treat object
This column denotes the stress treatment given to the flies. It contains categorical data with three categories: "Control," "Sham," and "Infection." This is nominal data because it represents different categories without any inherent order.

4. Protein int64
This column indicates the percentage of protein in the diet and ranges from 3% to 65%. It's a numeric column representing the quantity of protein. This data is numeric and continuous.

5. Rep int64
This column signifies the replicate number for treatment and diet groups. It's numeric and ordinal since it represents a specific sequence or order of replicates.

6. Dead float64
This column indicates whether the fly was alive or dead on the experimental day. It's categorical with values "0" for alive, "1" for dead, and "NA" for flies that escaped or were accidentally killed. This is nominal data as it represents categories without an inherent order.

7. Smurf float64
Represents the measure of gut integrity. It's categorical with values "1" for smurf (where dye reached outside the gut) and "0" for no smurf. "NA" is used if the fly escaped on the experimental day. This is nominal data.

8. Eggs float64
Indicates the number of eggs in the vial. It's numeric, but it's not continuous because there might be cases where the count was not done, marked as "NA".

9. Unhatched float64
It refers to the number of unhatched eggs by a vial. It is a numerical variable

10. Hatch float64
It refers to the number of hatched eggs by a vial. It is a numerical variable as well

11. NG float64

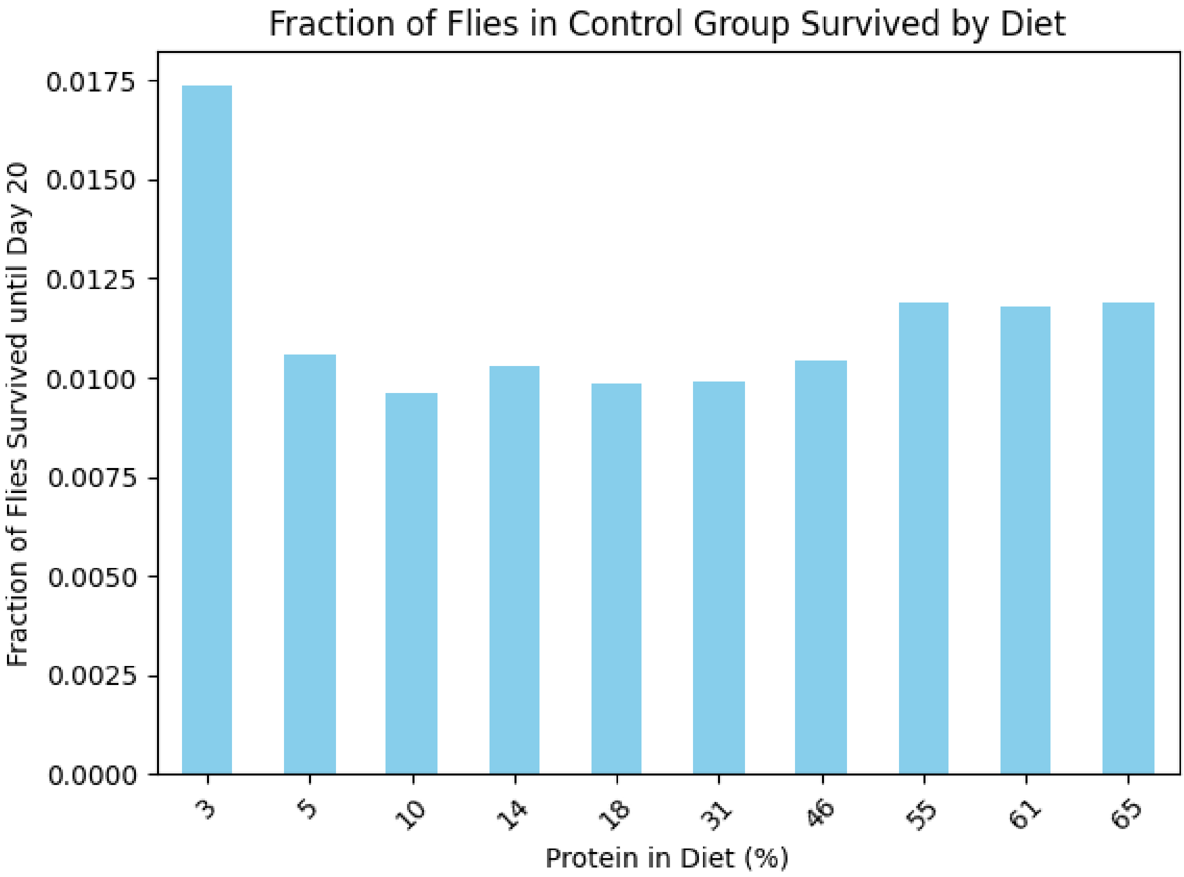
It represents time in seconds for the test, which is numeric and continuous. However, it's ordinal in how it's measured (time taken to climb) but not purely ordinal due to the presence of a capped value (60 seconds) and the "NA" category where no score was recorded.

Question2_2:
Numbers of Null value = 92224
Not all column have it
NG has the highest number of null value - 39626

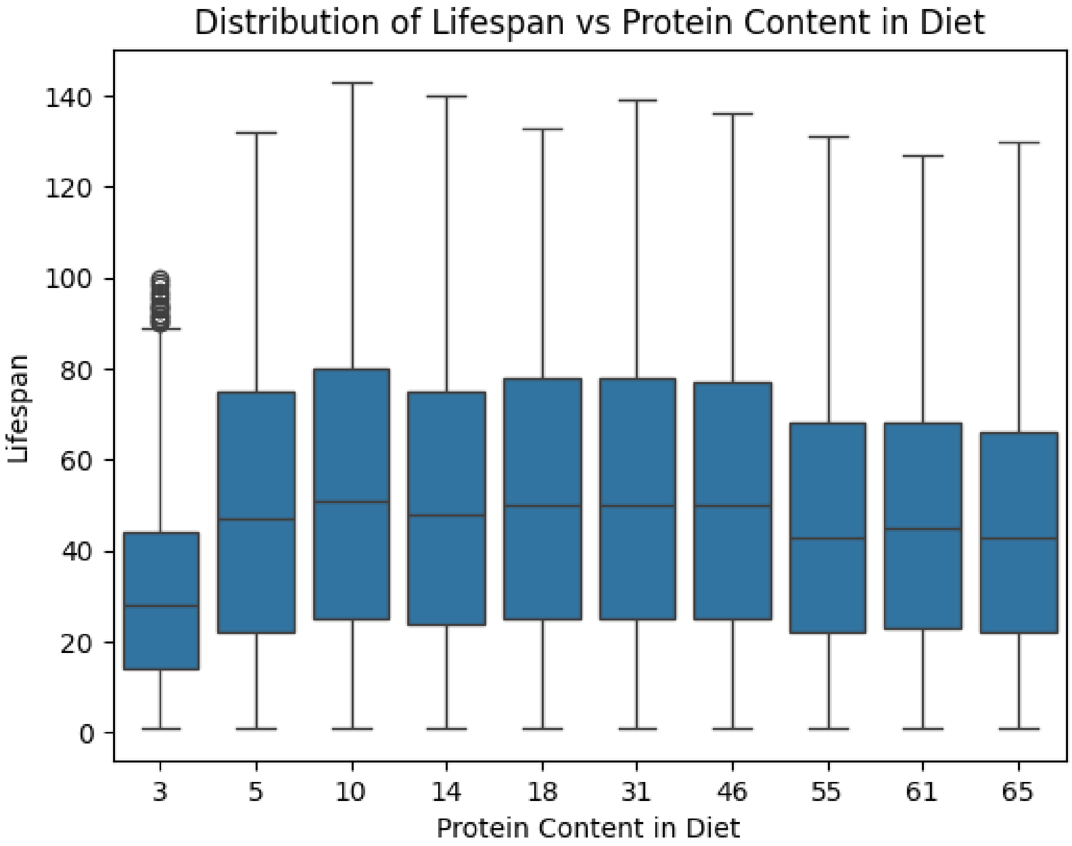
Number of column per column

Date	0
Vial	0
Treat	0
Protein	0
Rep	0
Dead	27
Smurf	27
Eggs	17514
Unhatched	17515
Hatch	17515
NG	39626

Question3.


Fraction of Flies in Control Group Survived by Diet
Fraction of Flies Survived until Day 20
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
3
5
10
14
18
31
46
55
61
65
Protein in Diet (%)

Question4_1:-
Null Hypothesis :- There is no significant difference in average lifespan based on protein content in the diet.
Alternative hypothesis:- There is a significant difference in average lifespan based on protein content in the diet.

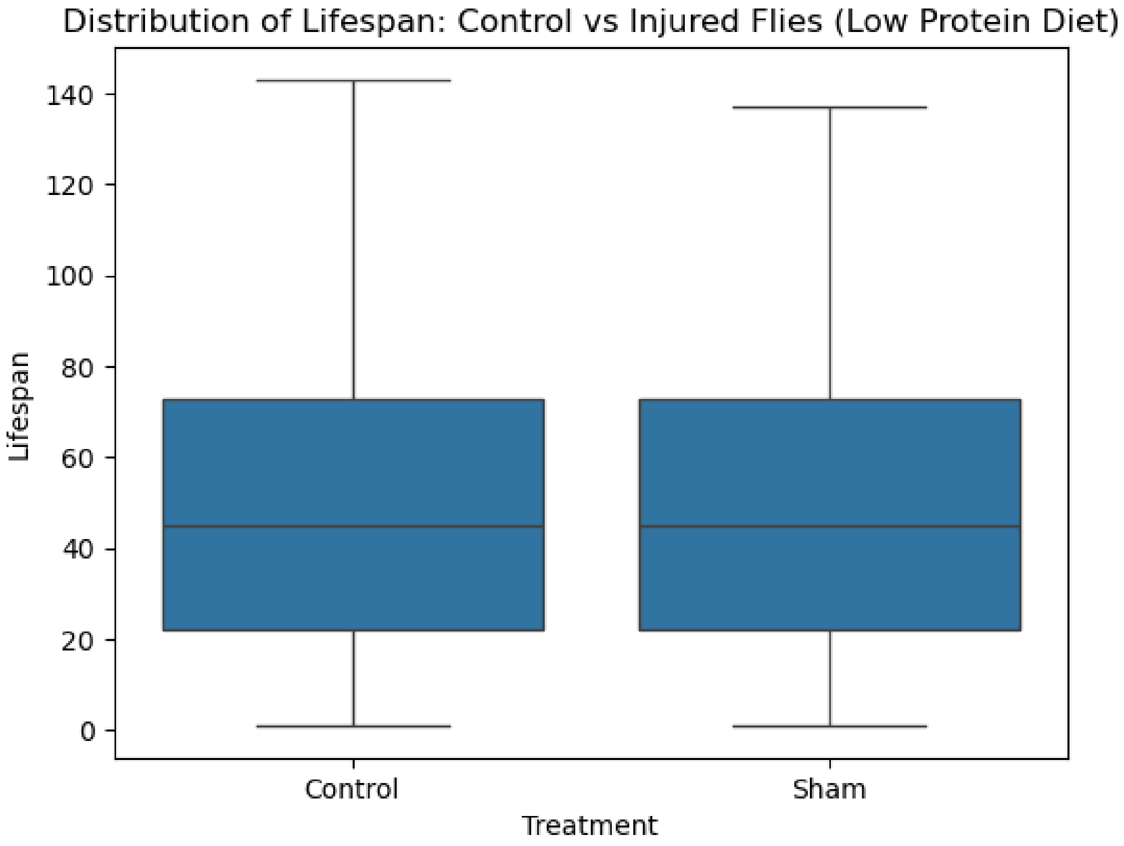


Conclusion:- From the box plot, we can easily see that, the there is a significant difference in average lifespan based on protein content in the diet.'
Therefore
'Reject the null hypothesis.There is a significant difference in average lifespan based on protein content in the diet.'

Question4_2

Null Hypthesis :- There is no significant difference in average lifespan of control vs injured flies under a low protein diet.

Alternative Hypothesis:- There is a significant difference in average lifespan of control vs injured flies under a low protein diet.

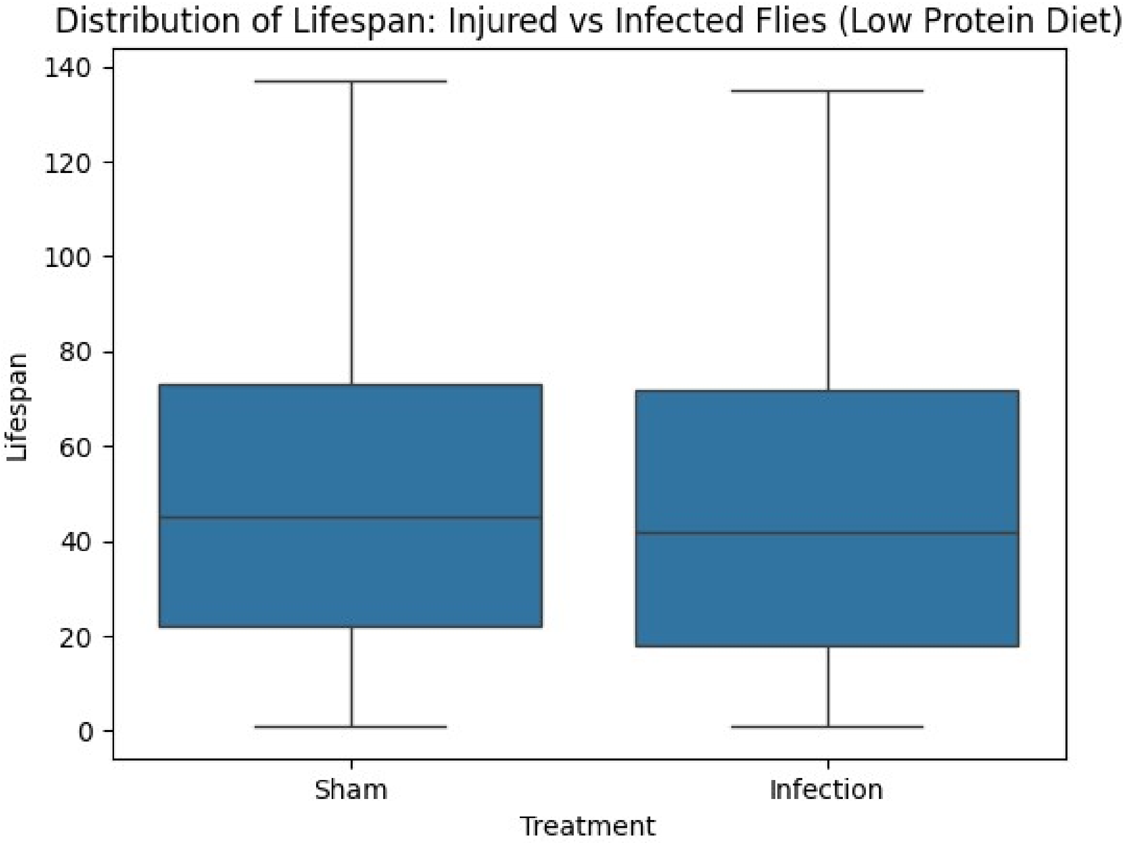


From the box plot, we can clearly see that there is no significant difference in average lifespan of control vs injured flies under a low protein diet.

Therefore
'Fail to reject the null hypothesis. There is no significant difference in average lifespan of control vs injured flies under a low protein diet.'

Question4_3:
Null Hypothesis - There is no significant difference in average lifespan of injured and infected flies under a low protein diet.

Alternative Hypothesis :- There is a significant difference in average lifespan of injured and infected flies under a low protein diet.



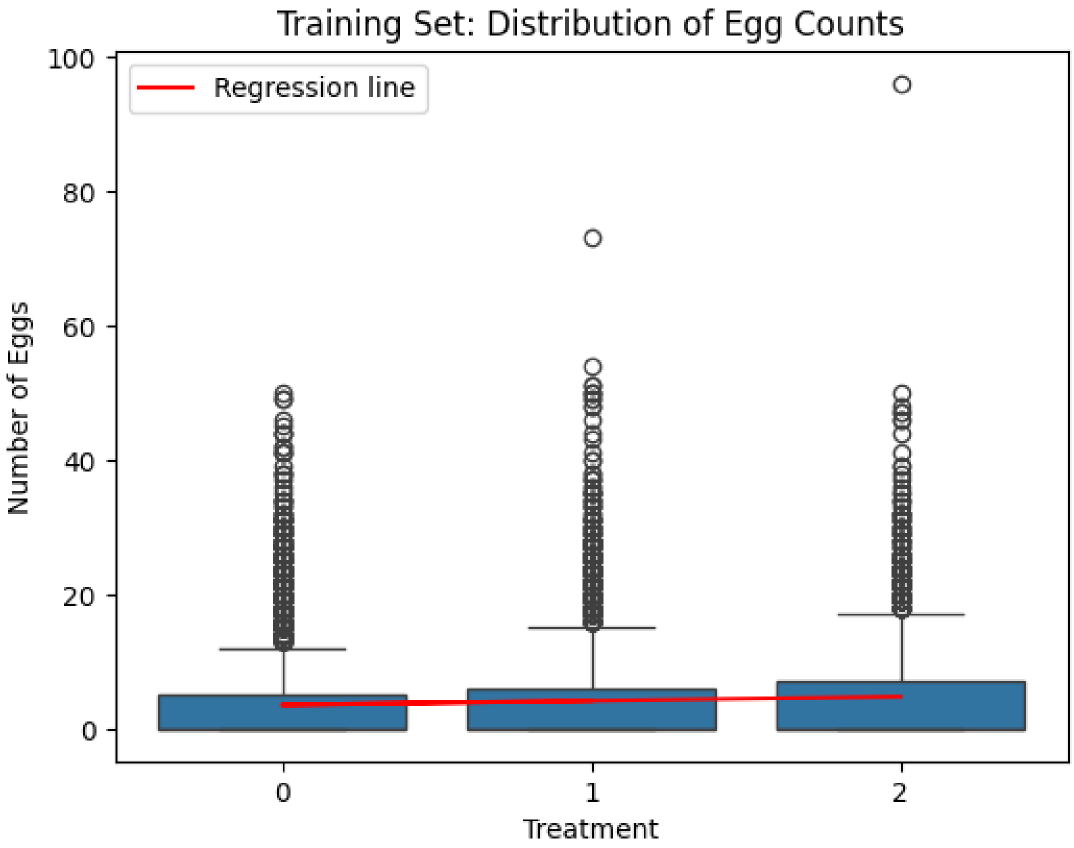
From the box plot, we can clearly see that there is a significant difference in average lifespan of injured and infected flies under a low protein diet.

Therefore:-

'Reject the null hypothesis. There is a significant difference in average lifespan of injured and infected flies under a low protein diet.')

Question4_4:
Null Hypothesis (H0):
There is no significant linear relationship between the severity of treatment (ordinal variable representing control, injury, infection) and the average number of eggs produced under a low protein diet.

Alternative Hypothesis (H1):
There is a significant linear relationship between the severity of treatment (ordinal variable representing control, injury, infection) and the average number of eggs produced under a low protein diet.



Intercept: 3.565603910547476
Slope: 0.6148989042008862
R-squared: 0.007278056980728587

Summary of Results:
Intercept: The estimated average number of eggs produced when the treatment severity is at its lowest level (control) is approximately 3.57.

Slope: For each unit increase in the severity of treatment (moving from control to injury or injury to infection), there is a predicted increase of approximately 0.615 eggs produced on average. This small slope suggests a marginal change in egg production associated with the severity of treatment.

R-squared: The R-squared value of approximately 0.0073 indicates that only about 0.73% of the variability in the number of eggs produced can be explained by the linear relationship with the ordinal variable representing treatment severity. This suggests an extremely weak linear association between treatment severity and the number of eggs produced under a low protein diet.

Conclusion:
Based on the linear regression model:

The severity of treatment, as represented by the ordinal variable, does not appear to be a significant predictor of the number of eggs produced under a low protein diet.
The model, considering treatment severity alone, is inadequate in explaining or predicting variations in egg production.
Overall, the model's poor fit (indicated by the extremely low R-squared value) suggests that the relationship between treatment severity and egg production is very weak.We can say that the treatment does not determine the egg production.