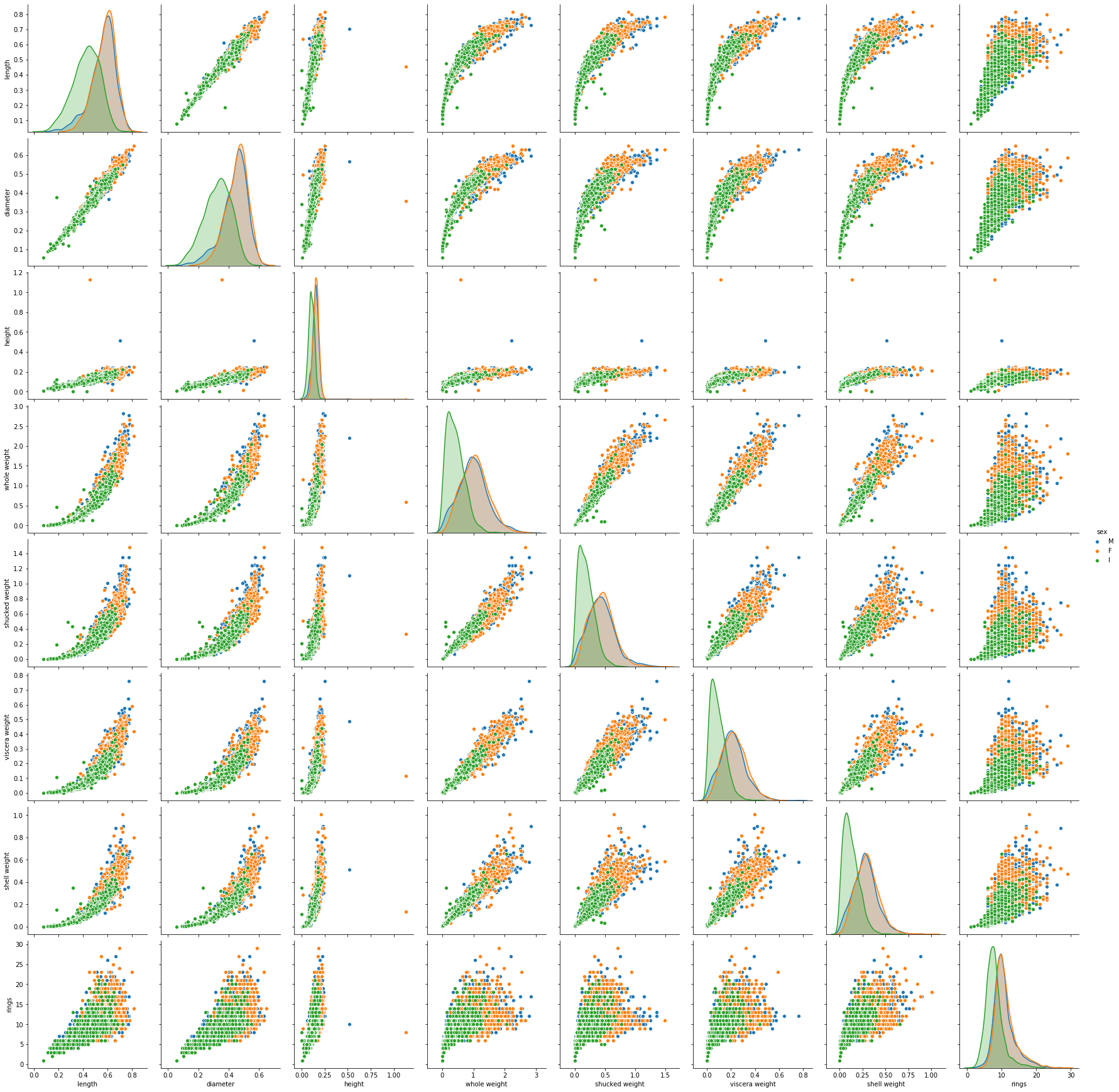
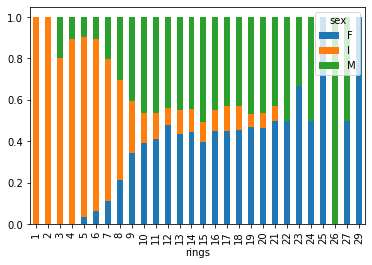
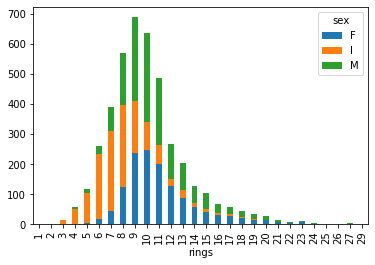
I am using Abalone dataset for this test.

<https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data>

Number of samples: 4177

|  | **sex** | **length** | **diameter** | **height** | **whole weight** | **shucked weight** | **viscera weight** | **shell weight** | **rings** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | M | 0.455 | 0.365 | 0.095 | 0.5140 | 0.2245 | 0.1010 | 0.150 | 15 |
| **1** | M | 0.350 | 0.265 | 0.090 | 0.2255 | 0.0995 | 0.0485 | 0.070 | 7 |
| **2** | F | 0.530 | 0.420 | 0.135 | 0.6770 | 0.2565 | 0.1415 | 0.210 | 9 |
| **3** | M | 0.440 | 0.365 | 0.125 | 0.5160 | 0.2155 | 0.1140 | 0.155 | 10 |
| **4** | I | 0.330 | 0.255 | 0.080 | 0.2050 | 0.0895 | 0.0395 | 0.055 | 7 |





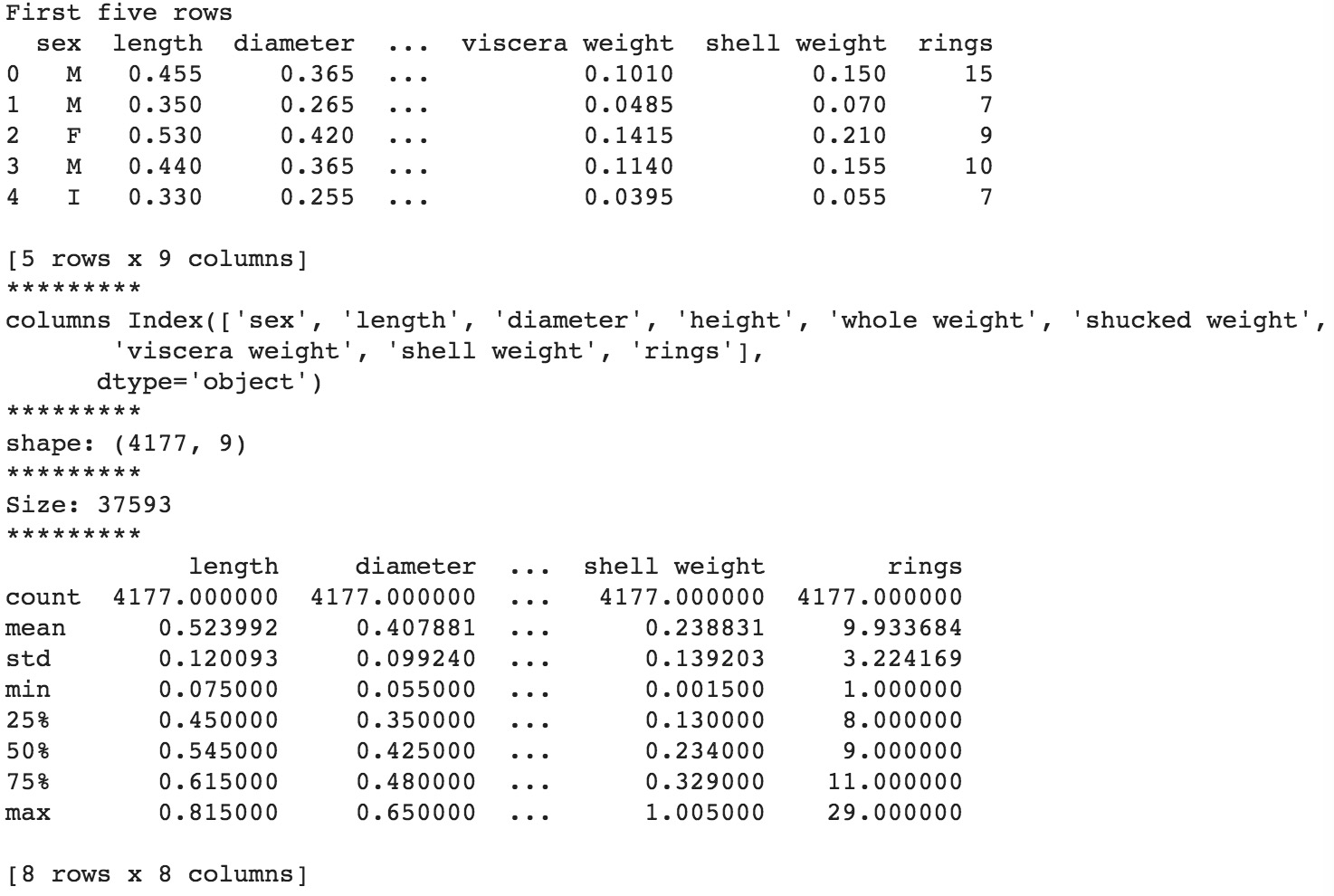
Q1: What is the difference between the last two plots?

Q2: Observation: Does the proportion of number of rings changes based on sex?

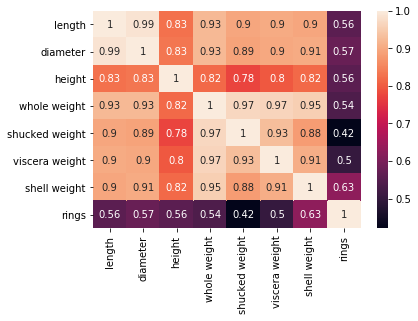
Q3: Write a hypothesis about it and suggest H0 and Ha?

Q4: What is the test conducted in the previous cell? When do we use this kidn of test? Interpret the results.

Q5: What is a t-test? Think of an example of t-test that you can perform on the given dataset (Abalone).



Q6: What is an ANOVA? Think of an example of ANOVA test that you can perform on the given dataset.



Q7: If predictors and target are defined as the previous cell, what kind of algorithm should we use to predict the target?

Q8: Why do we need to split the data in classification algorithms?

Actual Predicted

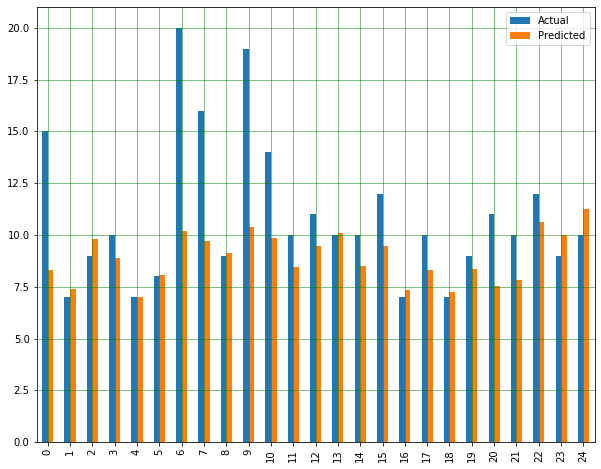
0 15 8.317449

1 7 7.370928

2 9 9.805707

3 10 8.901327

4 7 6.999174



Q9: What can you learn from the previous plot?

Mean Absolute Error: 1.8864971216844468

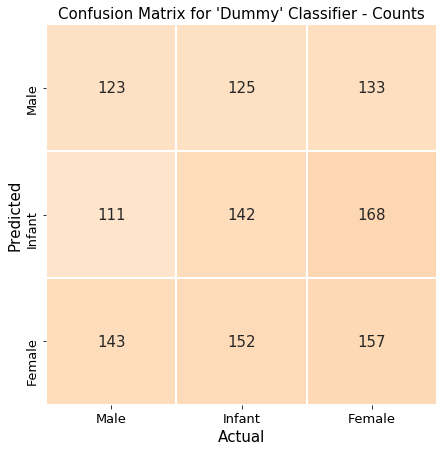
Mean Squared Error: 6.8569439678269966

Root Mean Squared Error: 2.618576706500498

Root Mean Squared Error: 2.618576706500498

R2 score: 0.34022025111581233

Q10: What is R2 and how can we interpret the value of R2 of the previous cell?



Q11: Assume that we are interested in predicting Infants, what is the value of the following parameters (use the Confusion Matrix for Logstic Regression Classifier):

True Postive:

True Negative:

Recall:

Q12: Explain k-means clustering algorithm, name a few other clustering algorithms.

Q13: Why do we need Generalized Linear Models (GLMs)?

Q14: What is overfitting?

Q15: What are the similarities and differences between PCA and LDA?