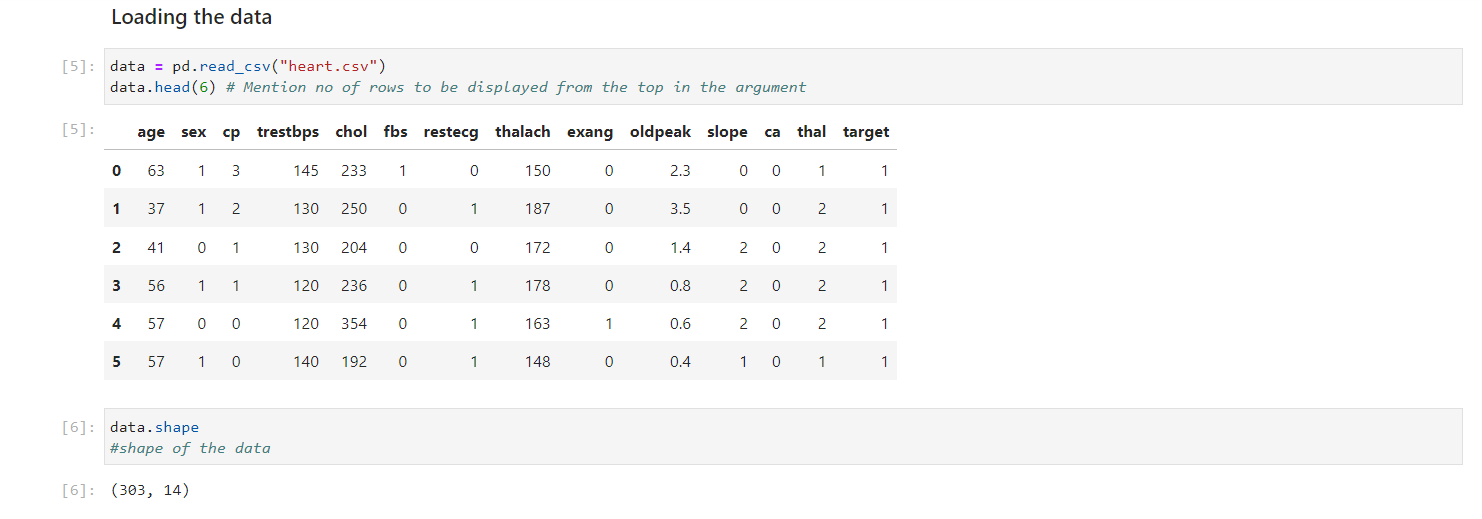
**Machine Learning**

**Code: Rushikesh Shinde, Data cr: Simplilearn**

1. Preliminary analysis:



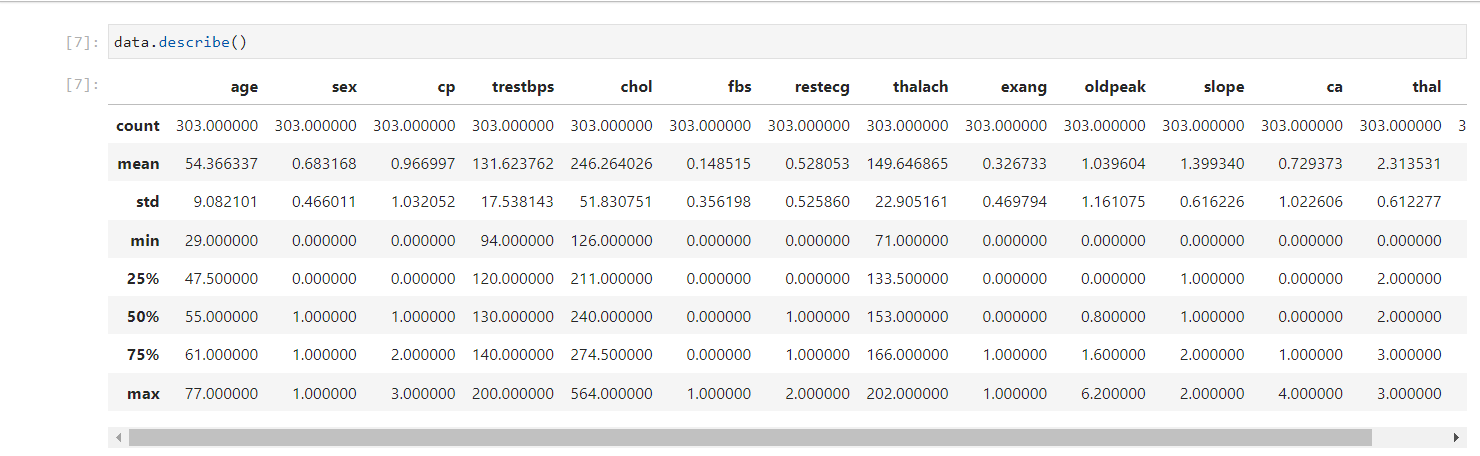


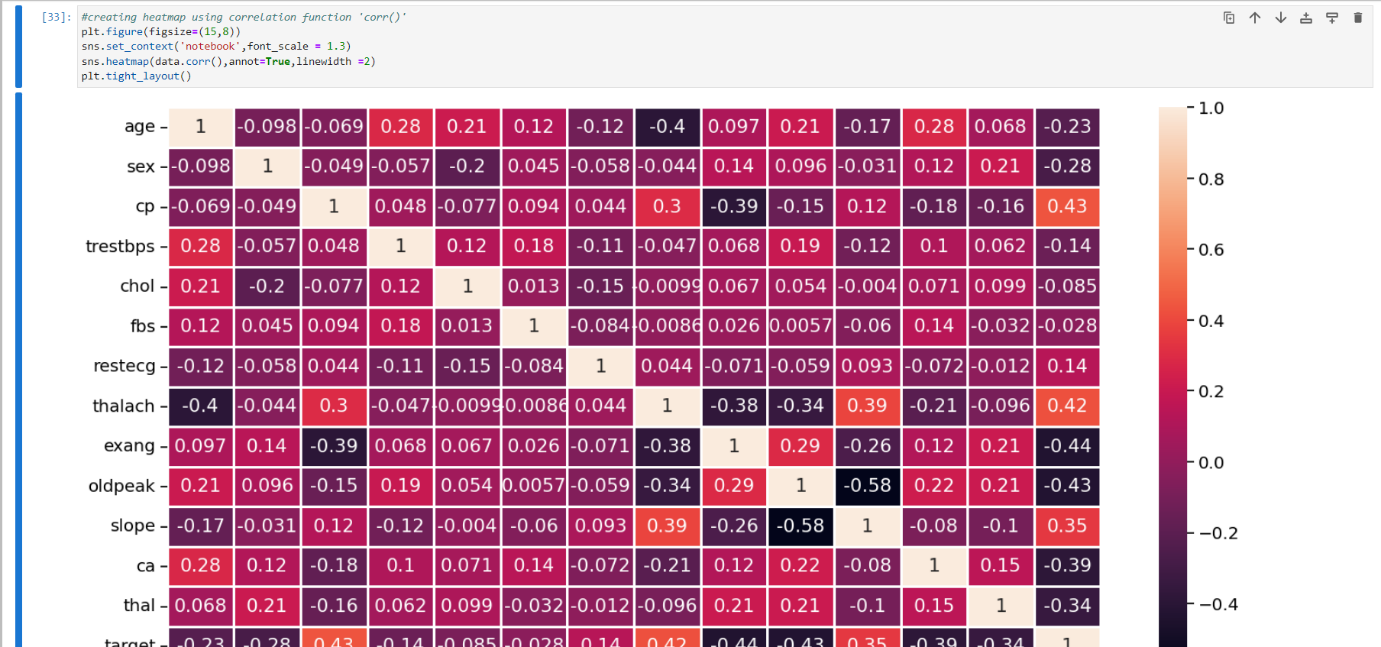
2. Prepared a report about the data explaining the distribution of the disease and the related factors using the steps listed below:

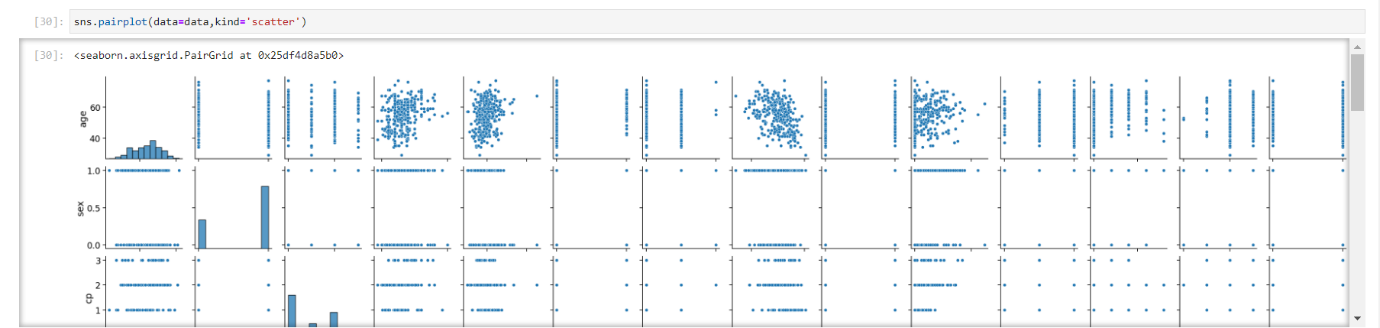
* 1. Preliminary statistical summary of the data i.e. shape, dtypes, null

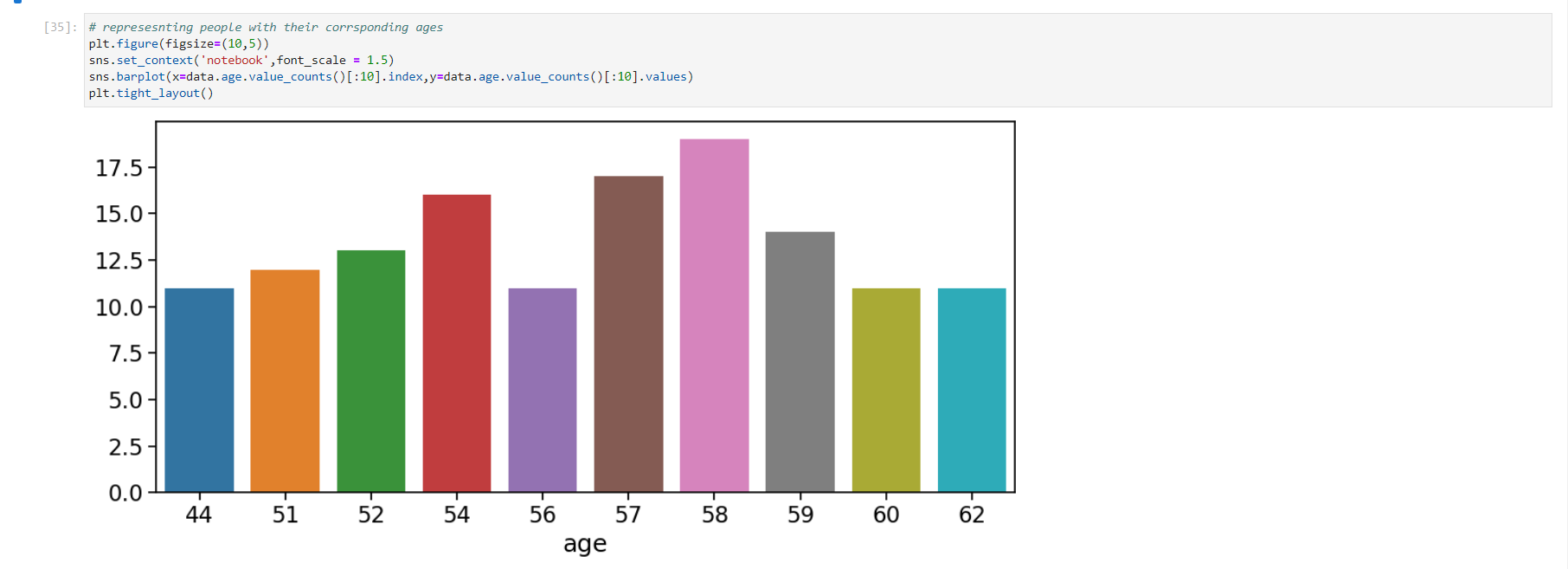
Value counts.

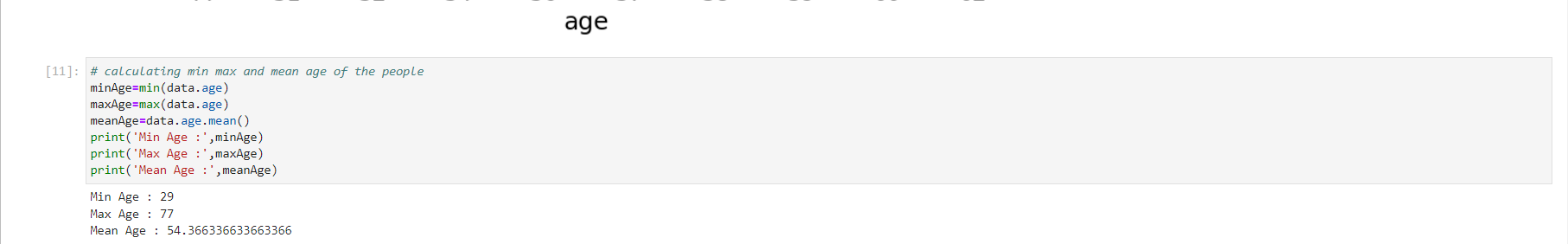
* 1. Identifying the data variables which are categorical and described and explored these variables.
  2. Study of the occurrence of CVD across the Age category
  3. Study of the composition of all patients with respect to the Sex category
  4. Studied if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient
  5. Described the relationship between cholesterol levels and a target variable
  6. Stated what relationship exists between peak exercising and the occurrence of a heart attack
  7. If thalassemia is a major cause of CVD
  8. List of how the other factors determine the occurrence of CVD
  9. Used a pair plot to understand the relationship between all the given variables

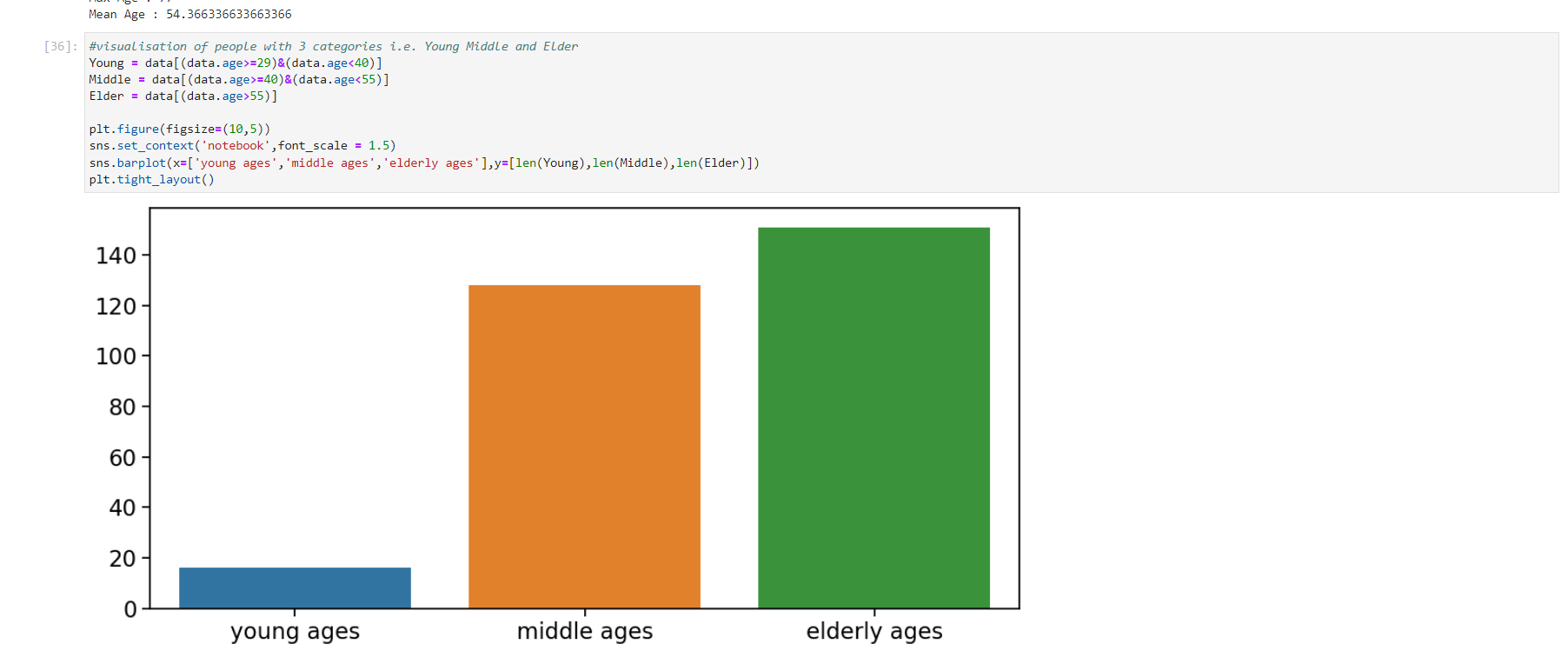


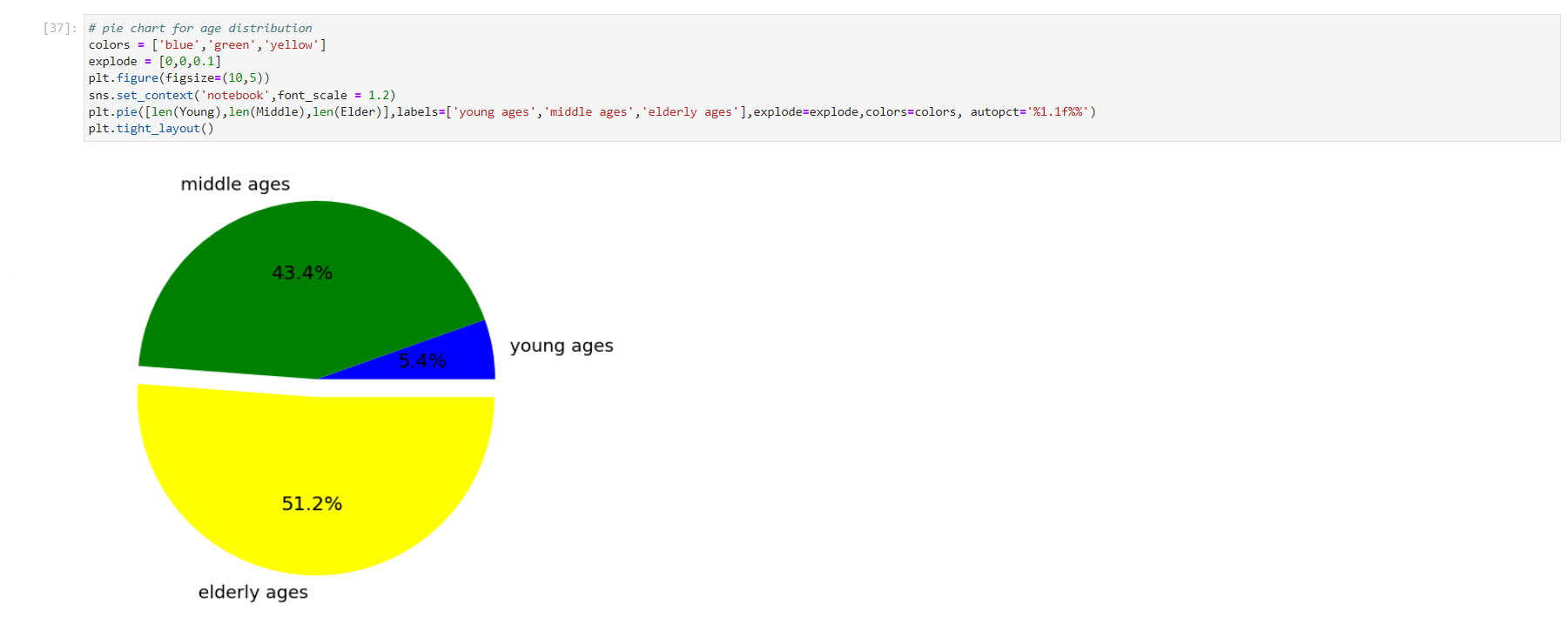


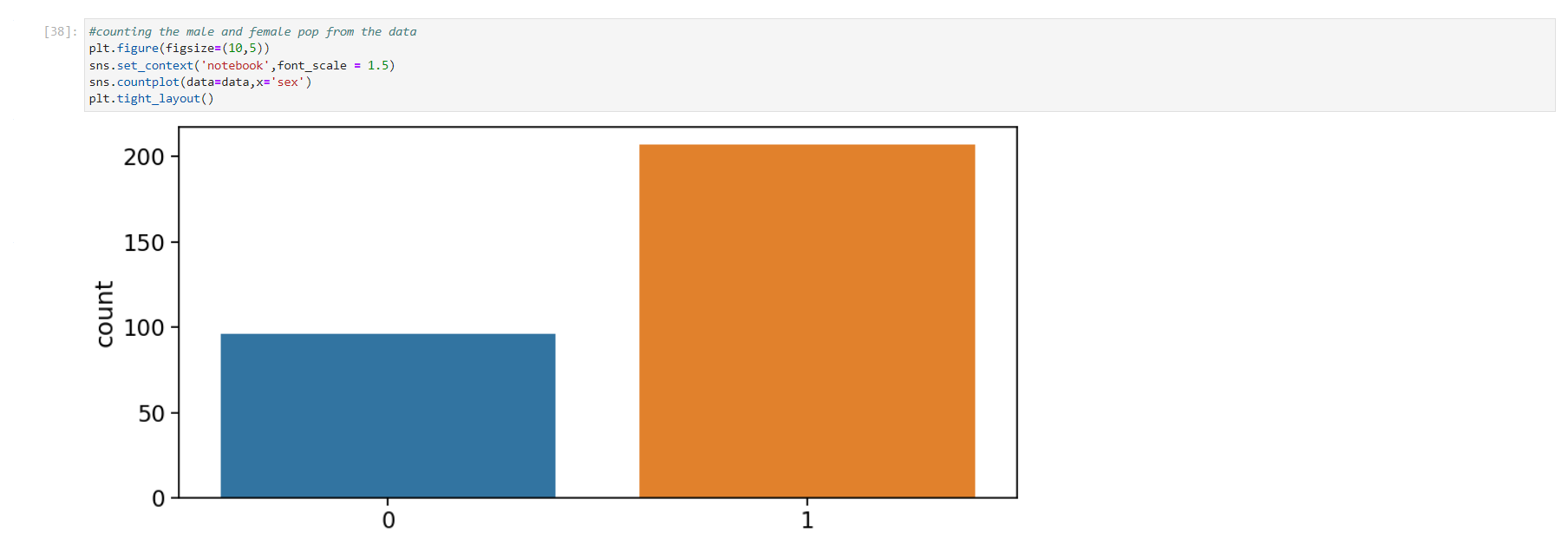


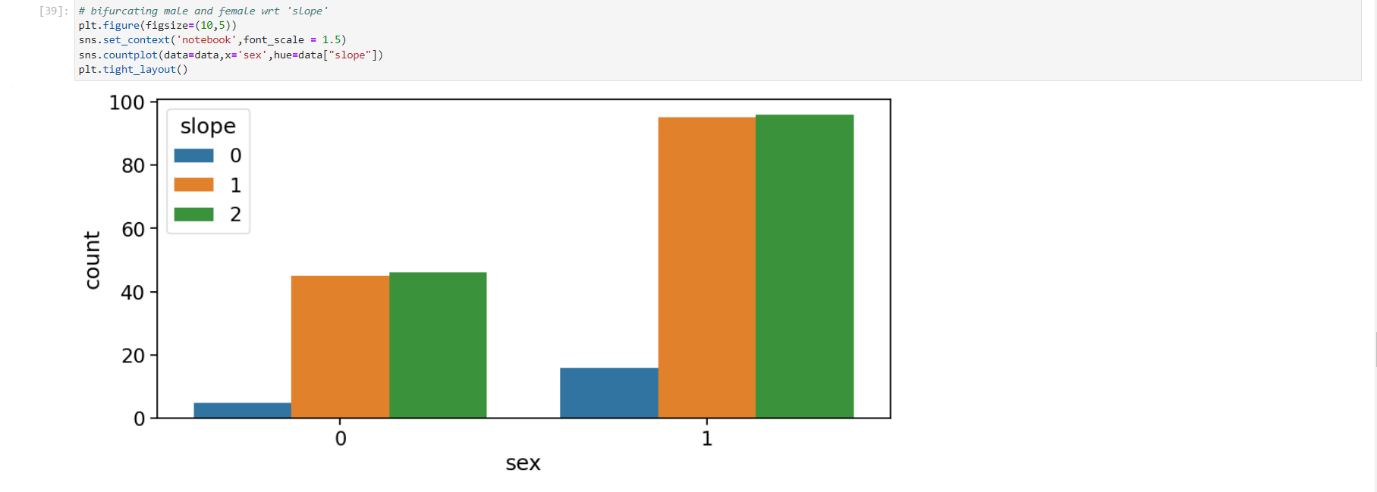


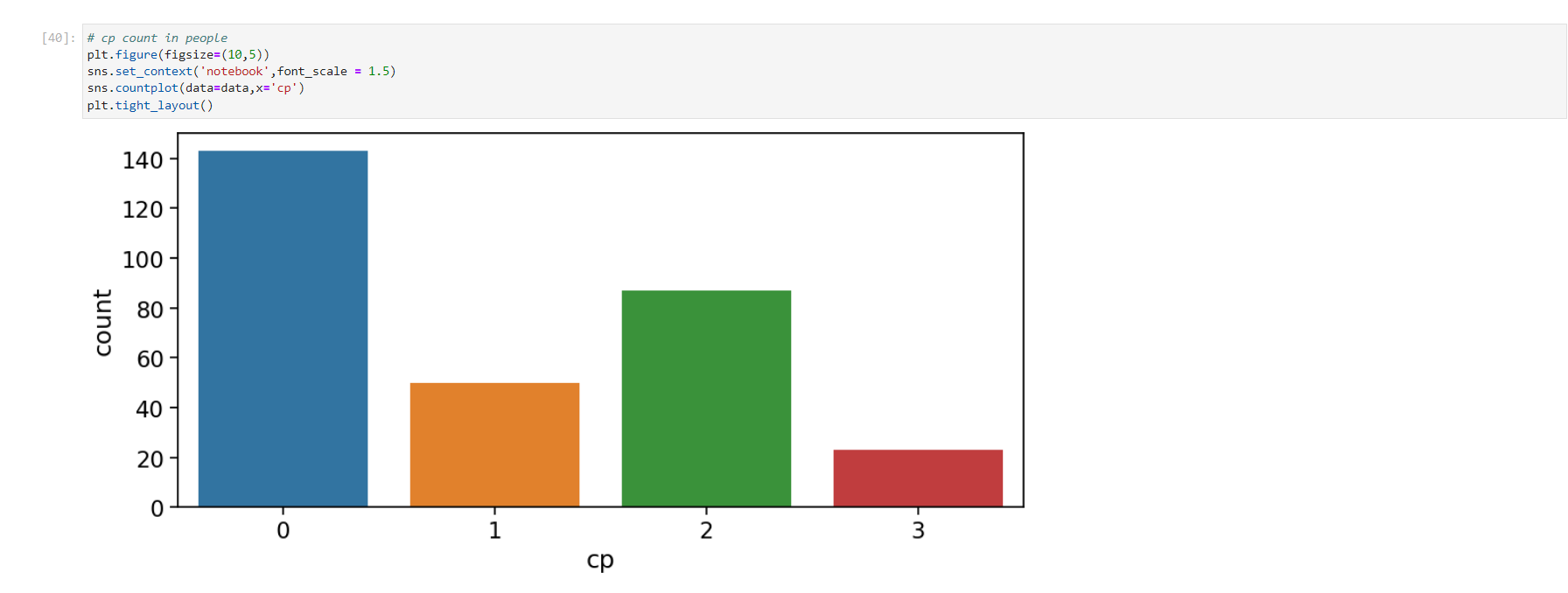


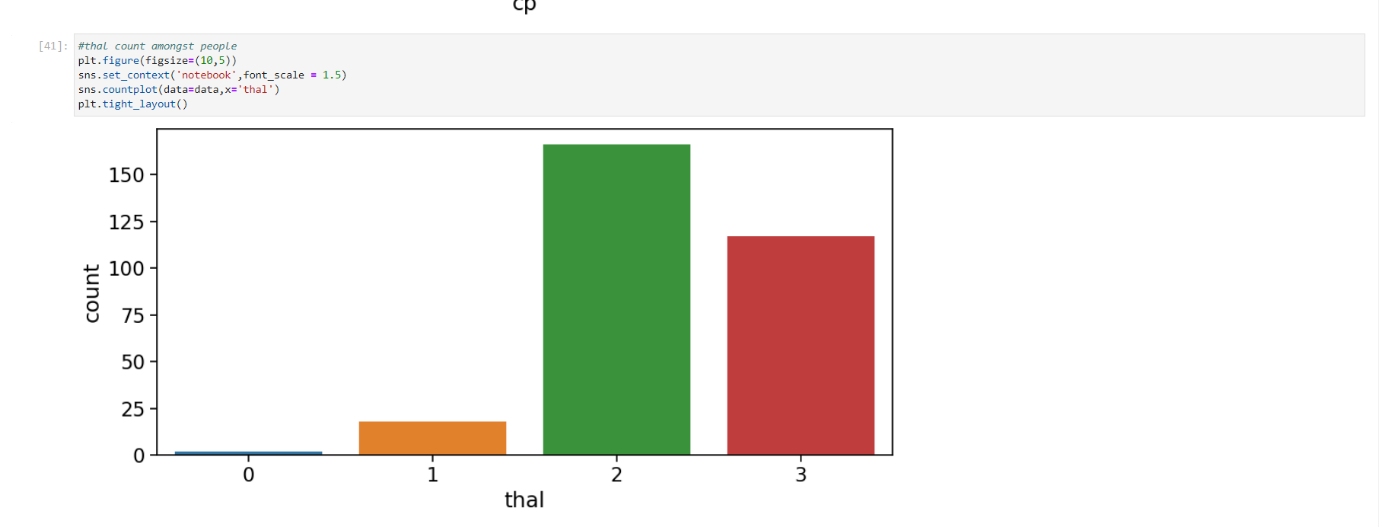


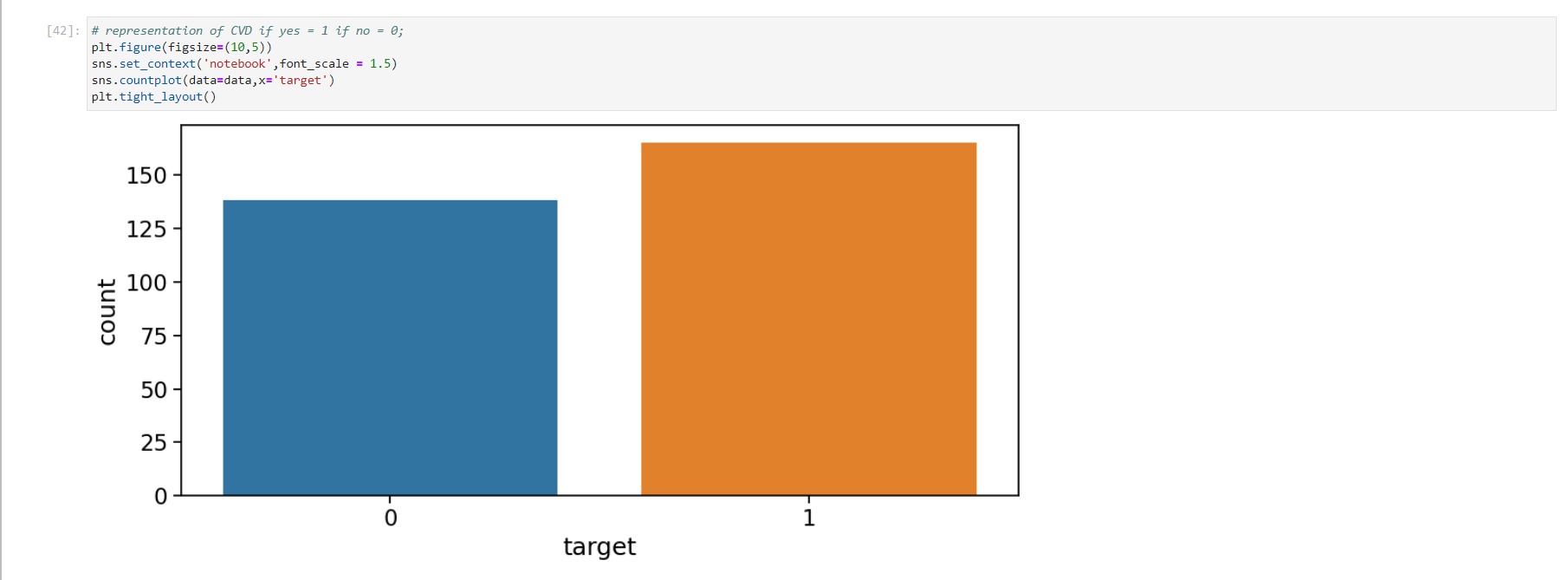


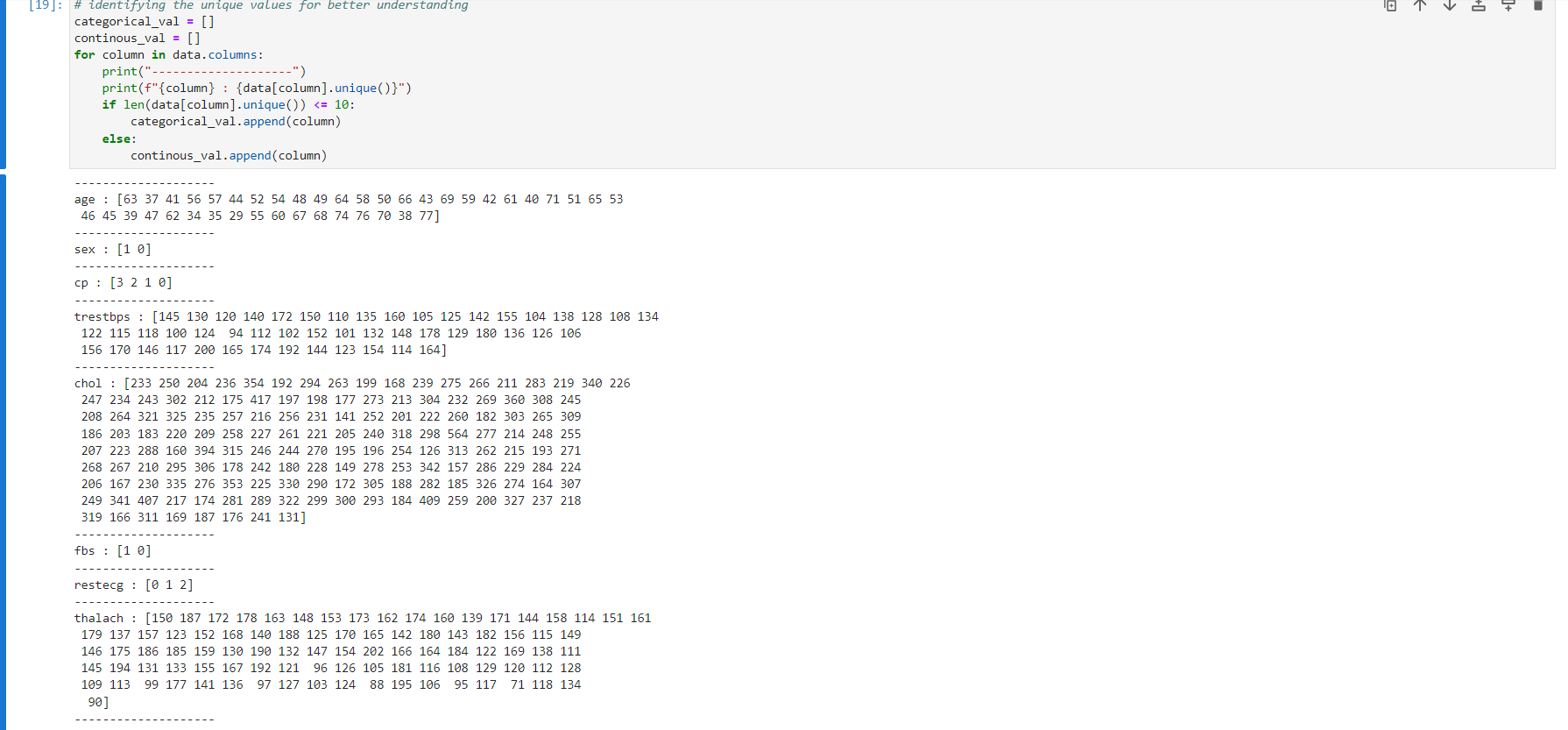




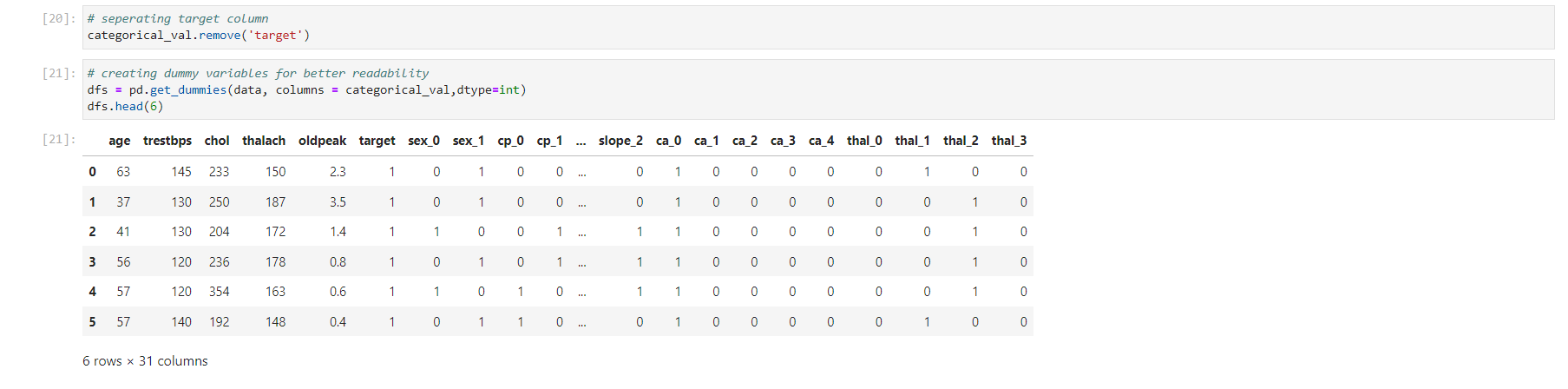


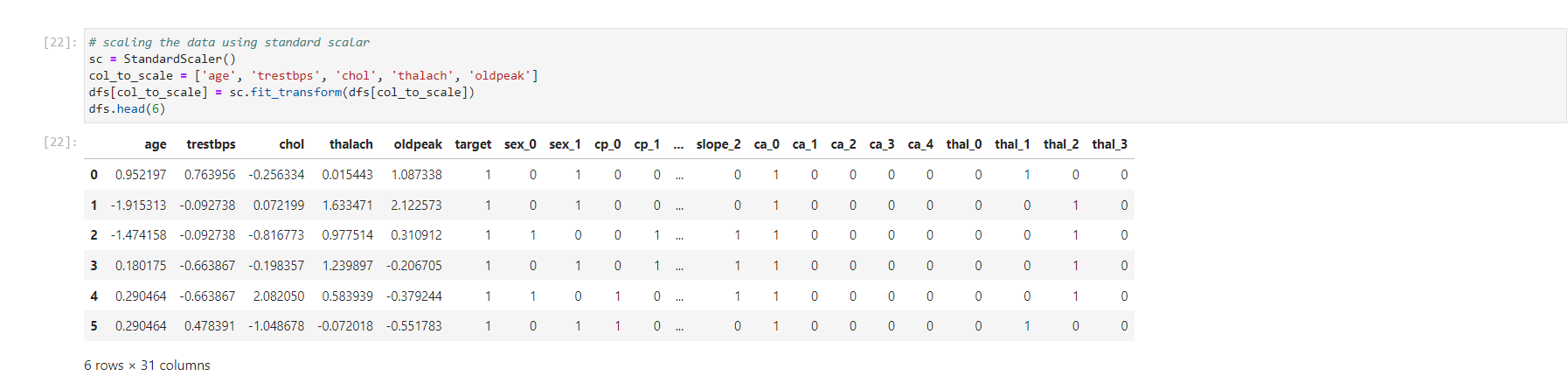






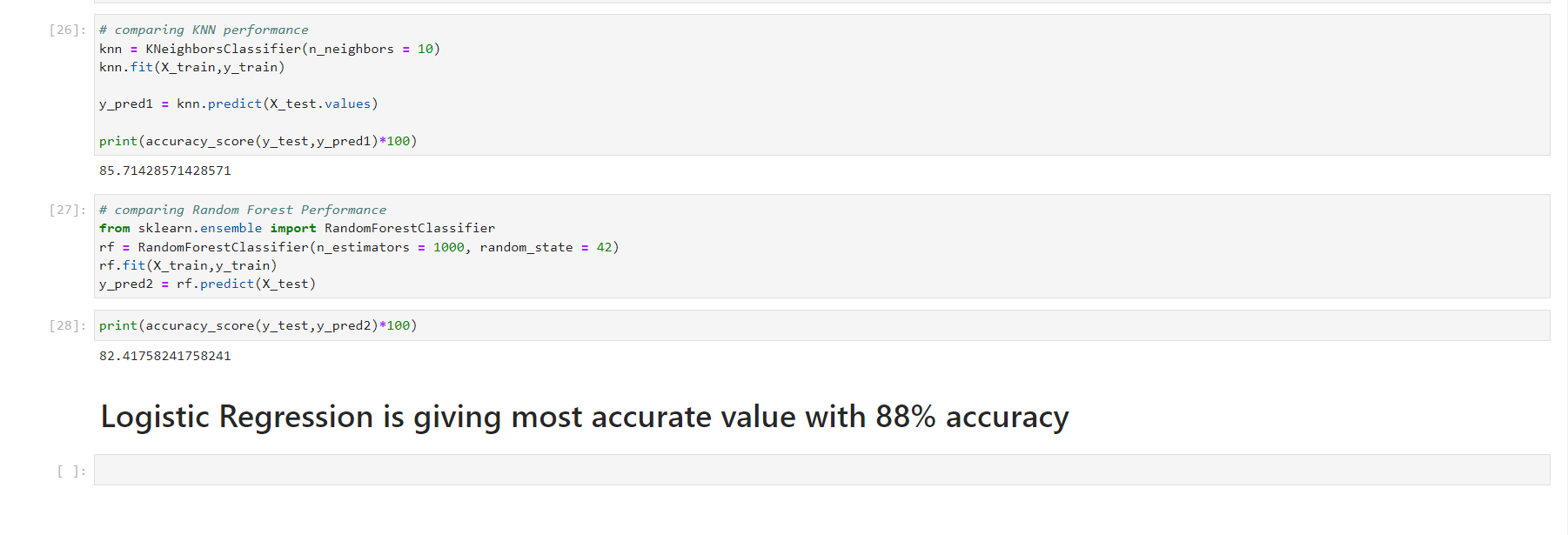






1. Built a Random Forest model as well as Logistic Regression model and KNN for better performance comparison hence we can conclude that Logistic Regression is giving highest ‘‘88%’’ accuracy and best model for this data.





Code File:-

Thank You!