**ANSWERS with Questions**

**Lalbabu ray**

**9654943540**

# Question 1:

1. Which of the following statements are correct? (Select all that are correct)
2. Information gain is measured by the reduction in entropy by splitting the data on the basis of a single attribute
3. Decision trees are used for classification but not regression
4. A Gini value of 0 implies a pure node

**ANS : iii**

1. What is the distributional assumption underlying Linear Discriminant Analysis? (Select all that are correct)
2. Classes follow the lognormal distribution
3. Classes follow the bionomial distribution
4. Classes follow a normal distribution

**ANS: iii**

1. Which of the following is a metric to use a split into a decision tree into additional nodes? (Select all that are correct)
2. Median
3. Entropy
4. Gini impurity
5. Standard Deviation

**ANS: ii, iii, iv**

# Question 2:

Import the dataset attached (Invistico\_Airline.csv) in python notebook (Jupyter Notebook/Google Colab) and develop an interesting investigative theme for a data driven story which **should** include the following –

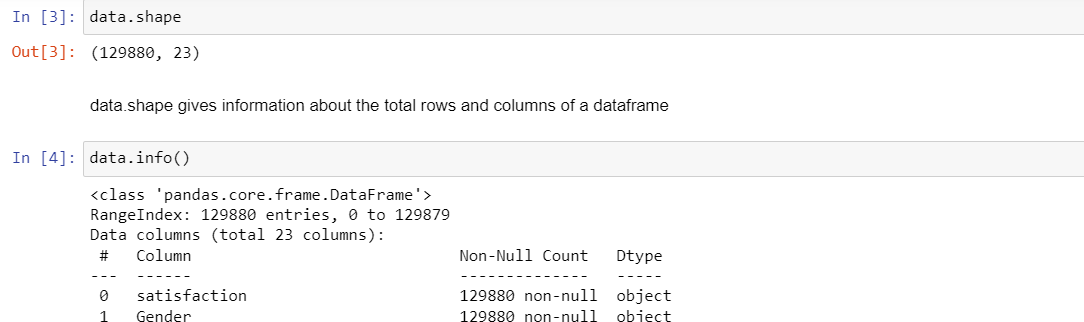
* No. of rows and columns
* Null Analysis
* Duplicated items
* EDA of total count of Satisfied vs Unsatisfied responses
* Correlation amongst variables (this include Customer Satisfaction column)
* EDA of total count of Satisfied vs Unsatisfied responses based on –

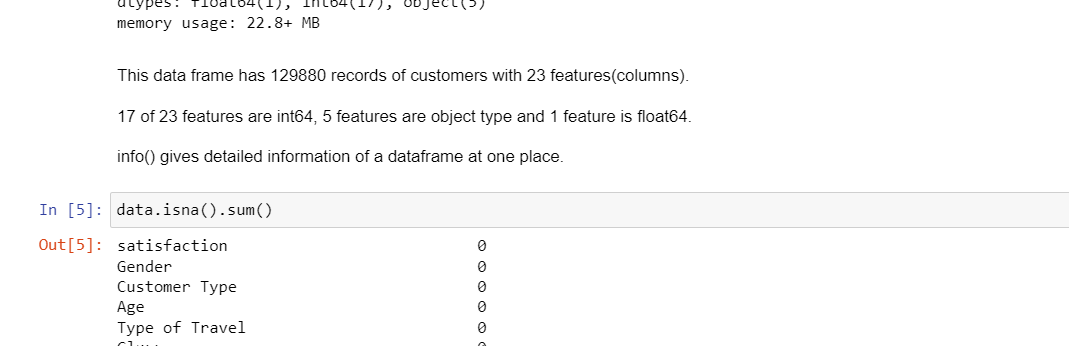
1. Gender
2. Customer Type
3. Age
4. Type of Travel
5. Class
6. Flight Distance
7. Seat Comfort

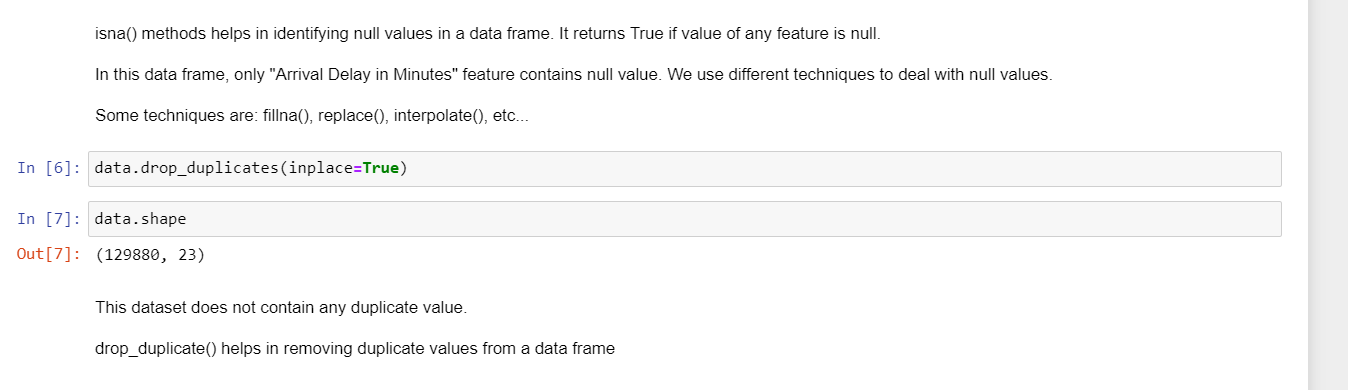
Submit the notebook file (.ipynb) along with a html file of your answer.

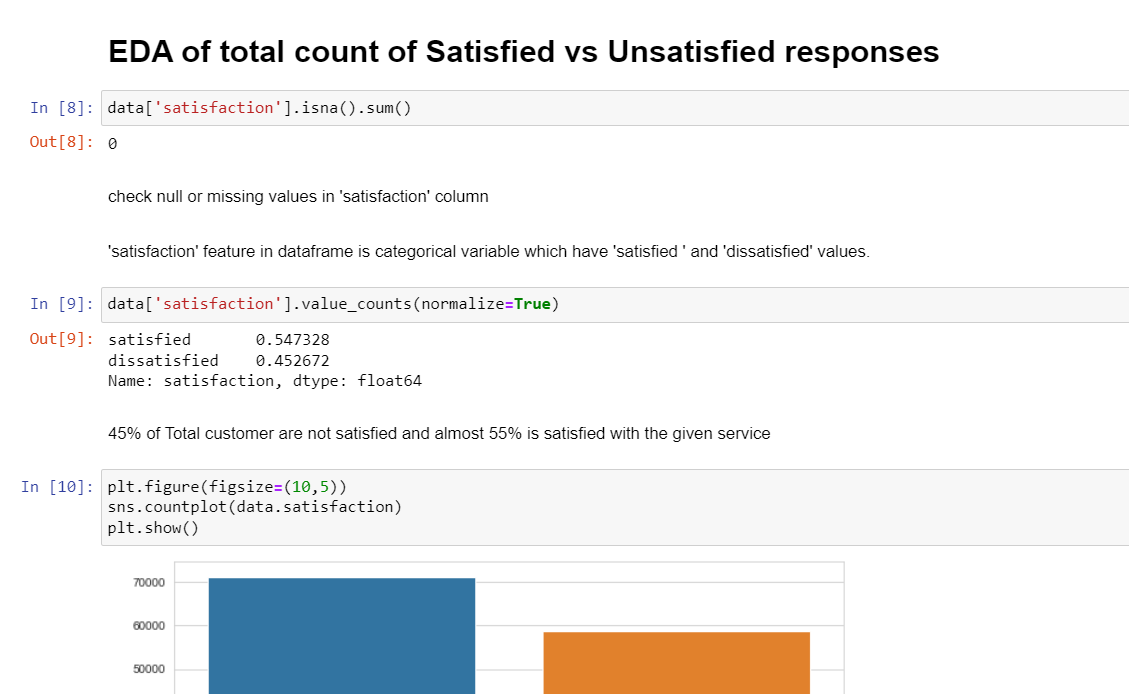
**Note:** Your analysis should not be just python codes and output but also 2-3 sentences describing your analysis in markdown.

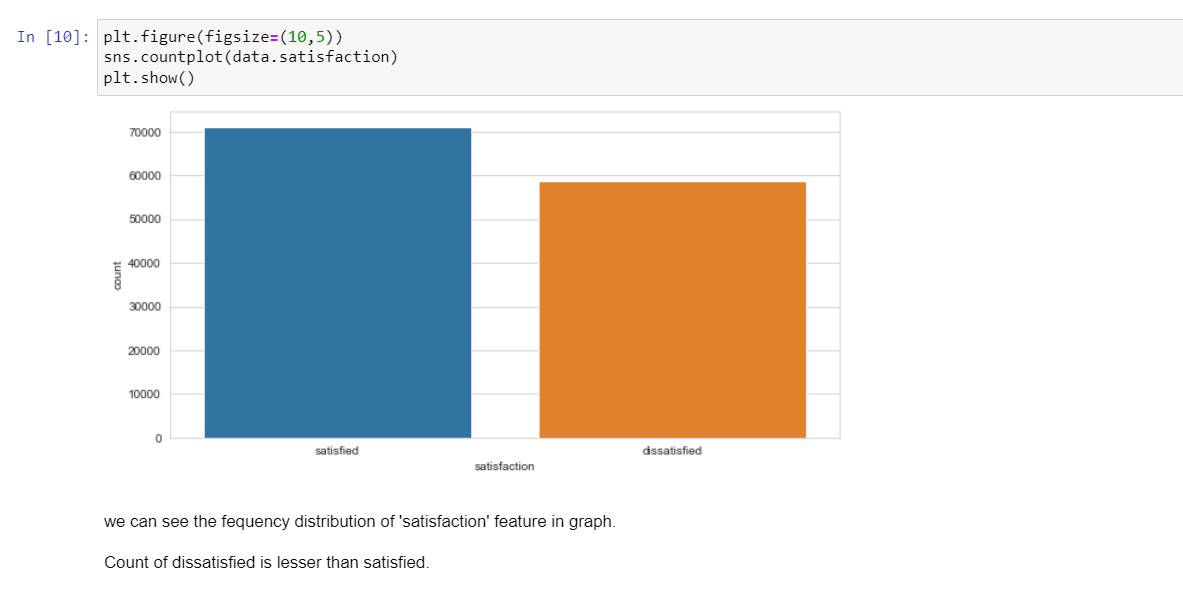


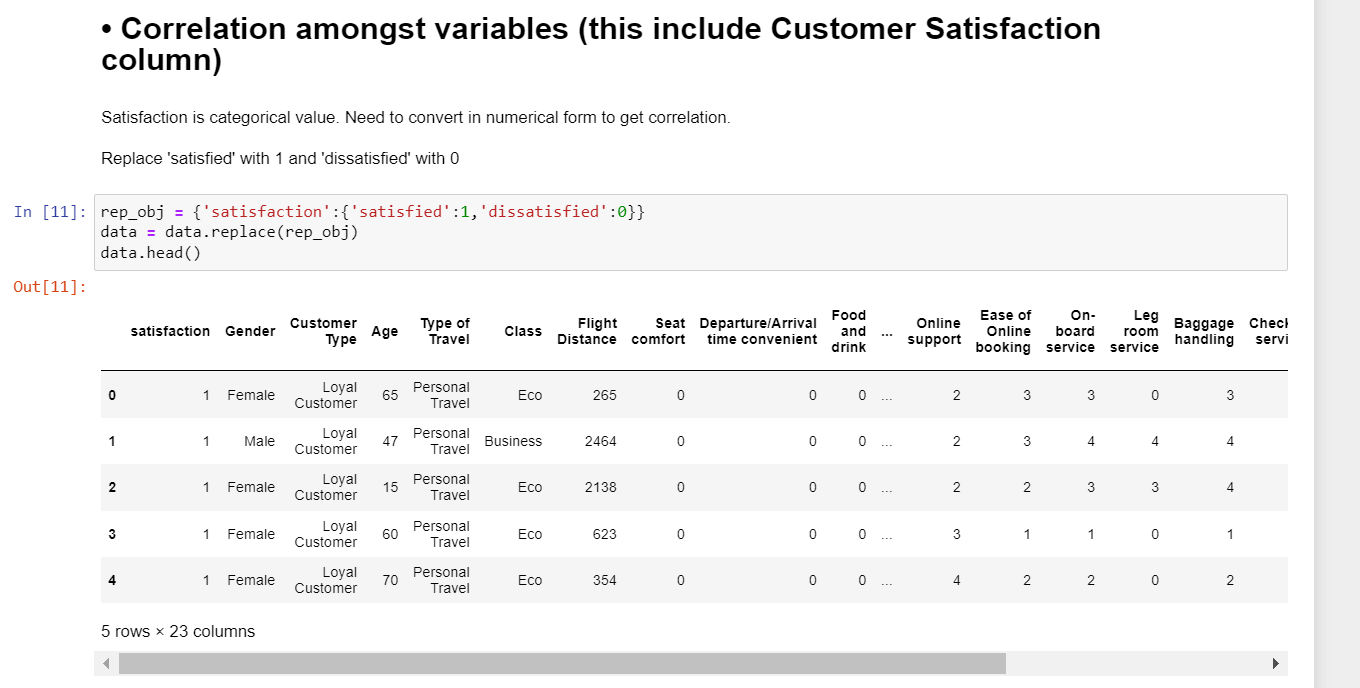


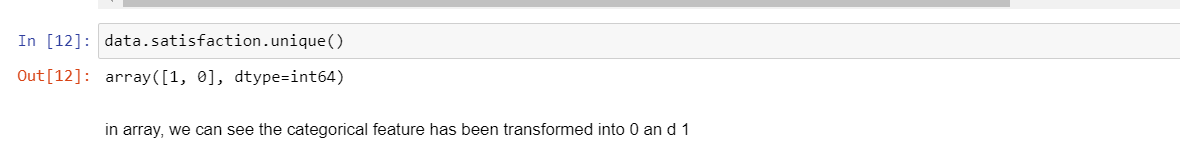






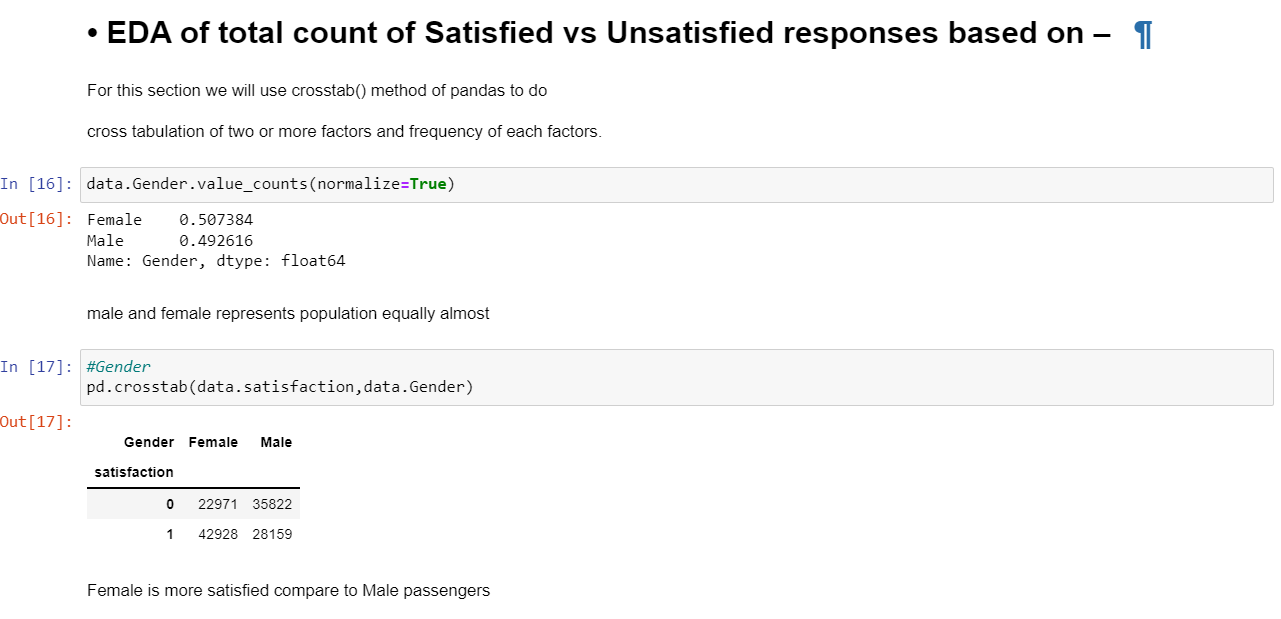


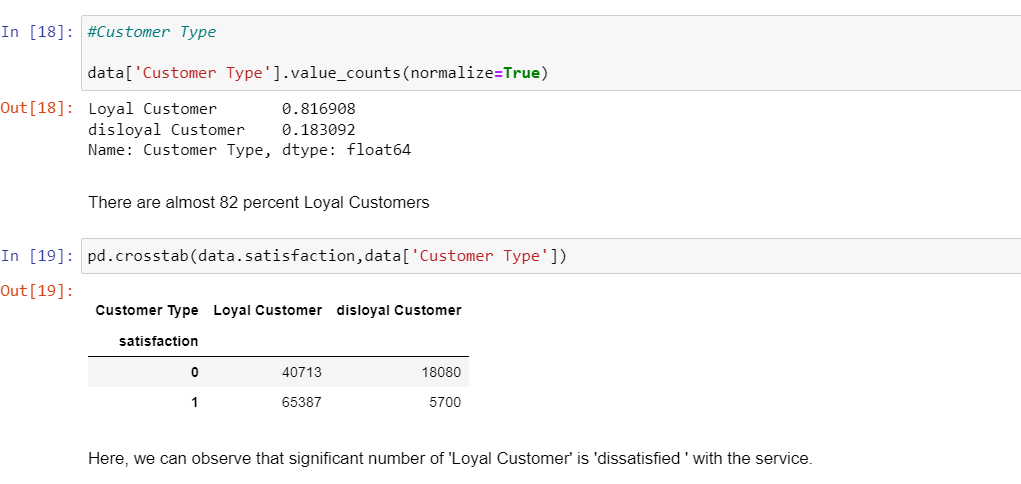




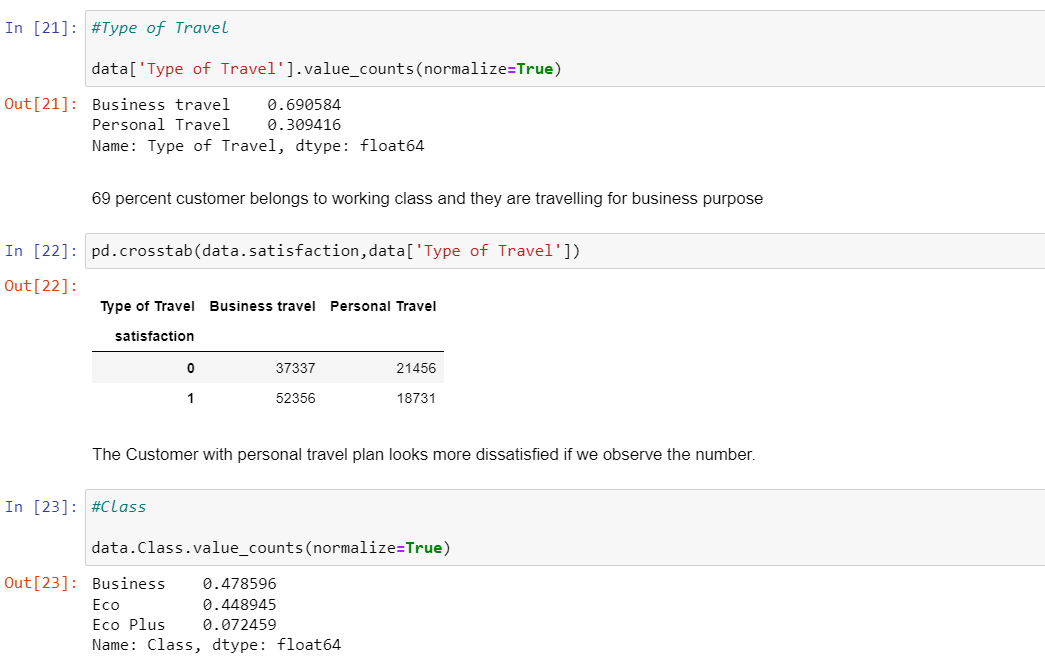
# 

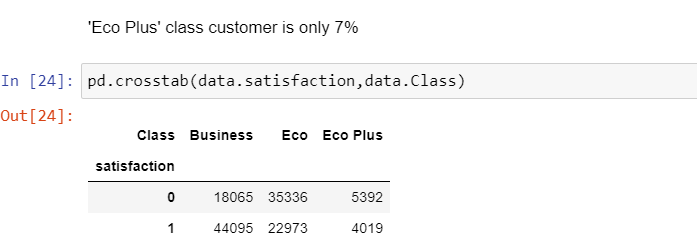
# 

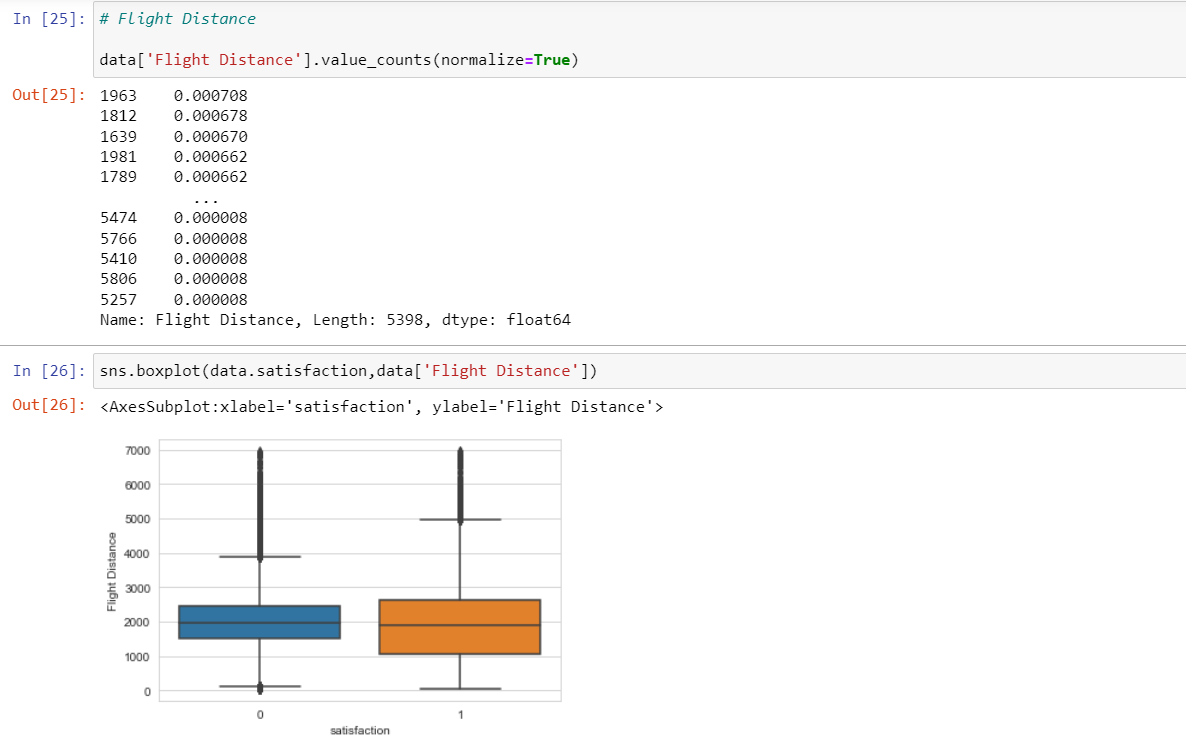


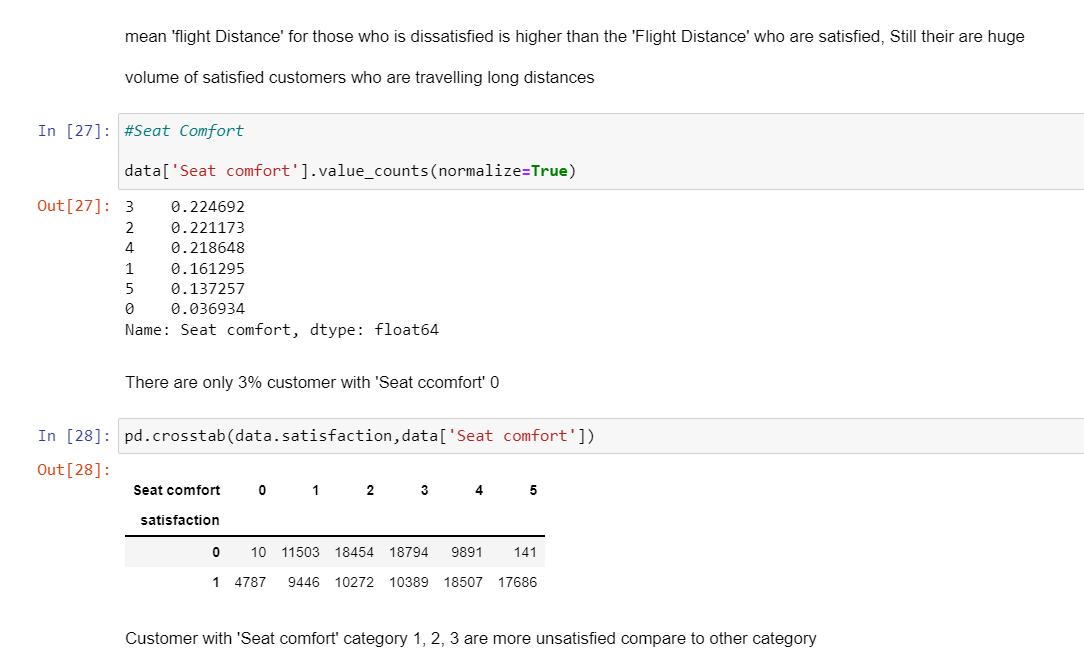










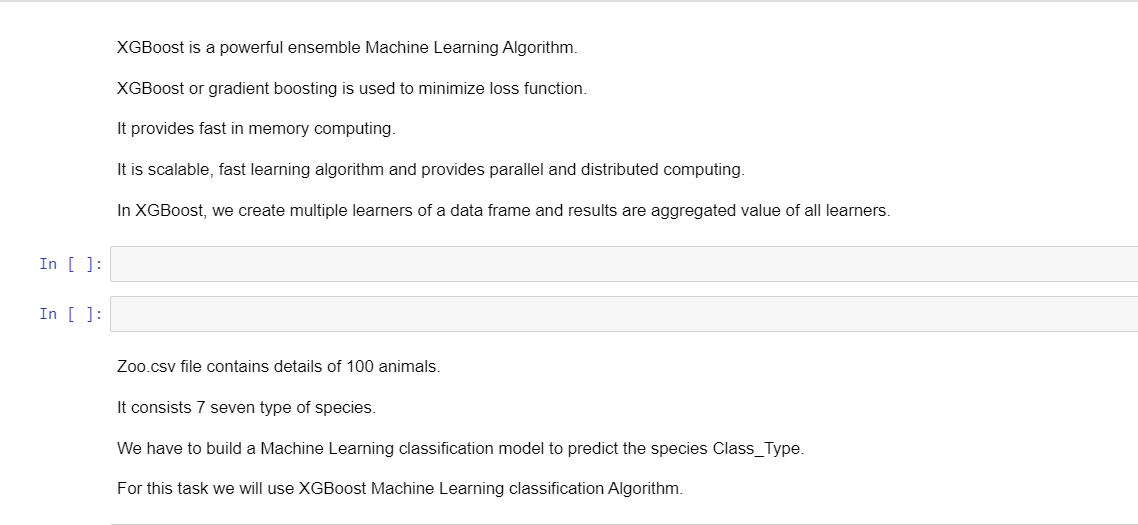


# Question 3:

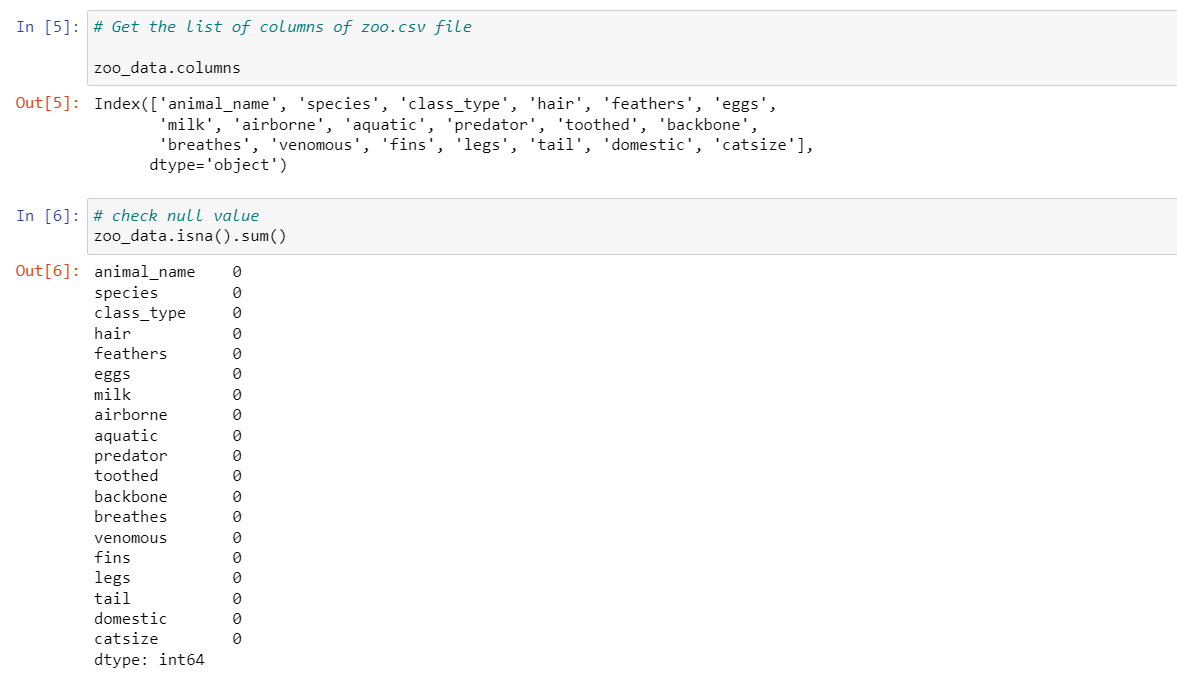
Import the dataset zoo.csv in python notebook (Jupyter Notebook/Google Colab) which contains 100 observations on different animals, with a number of features that describe each animal. The data has a column for species as well as a **class\_type** which is a numerical representation of the species. (the **class\_type** goes from 1 to 7, but the lables may require adjustments).

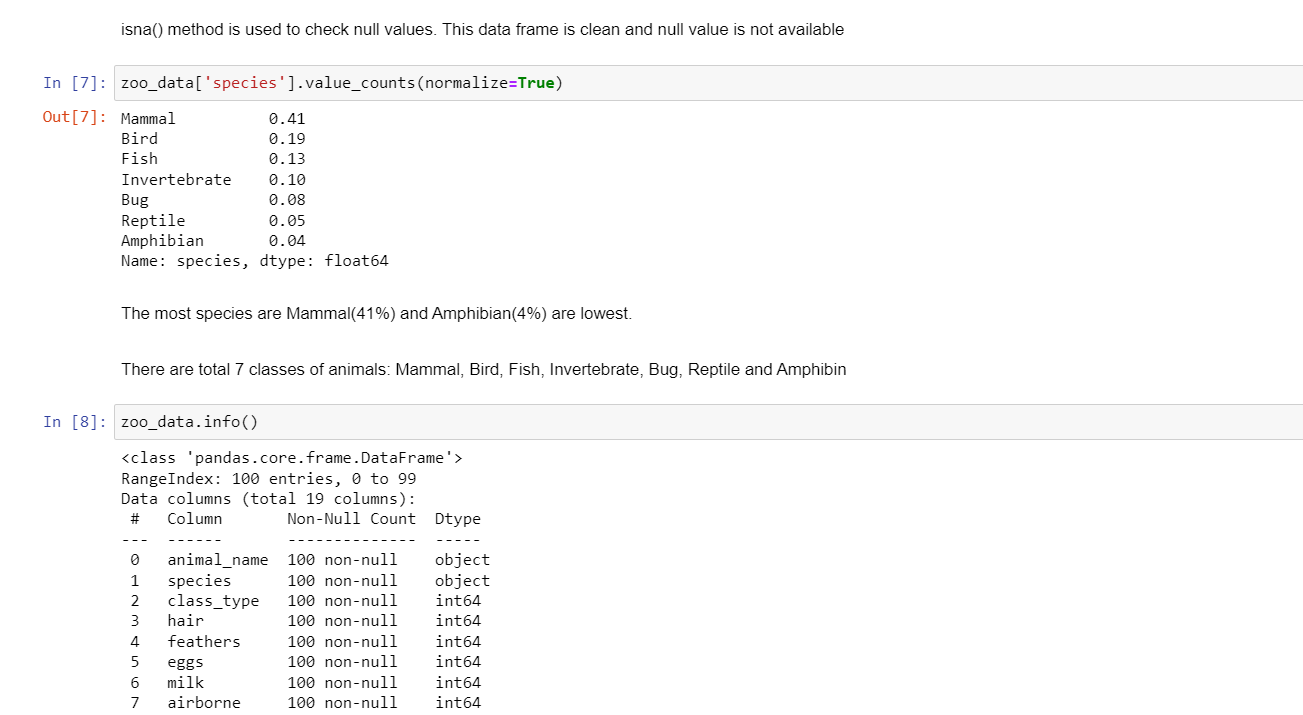
You need to build an XGBoost based classification model to predict the **class\_type** (which corresponds to species) using other features in the dataset. Train-Test split will be 80-20 ratio and **random\_state** is to be set to 1 (**random\_state = 1).**

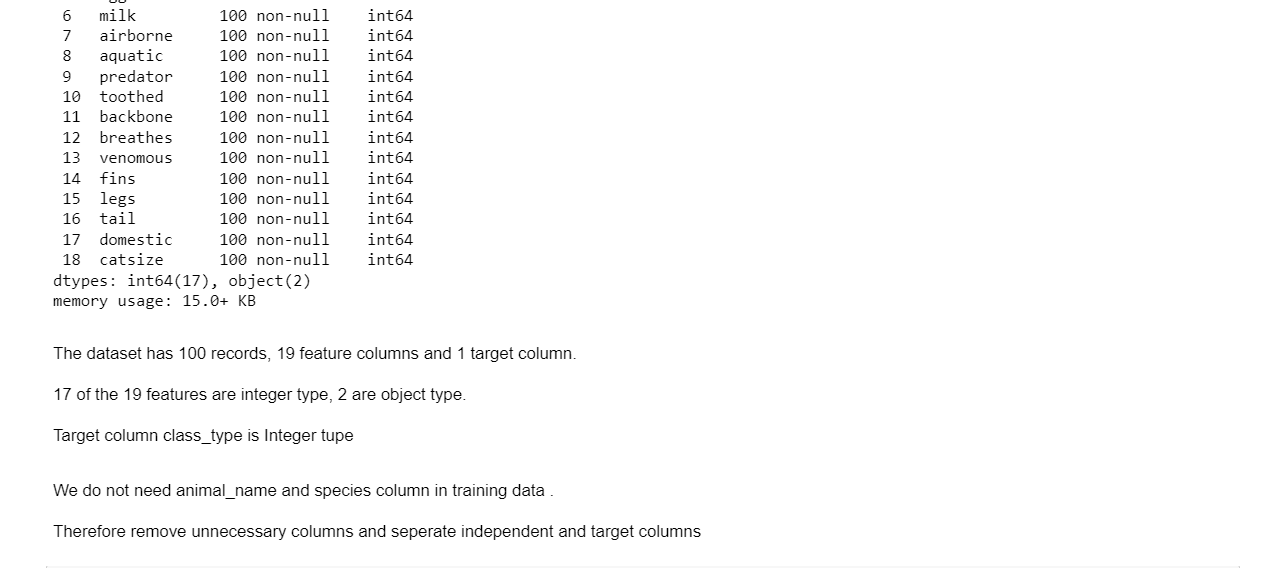
Find the overall accuracy rate of the model.

****

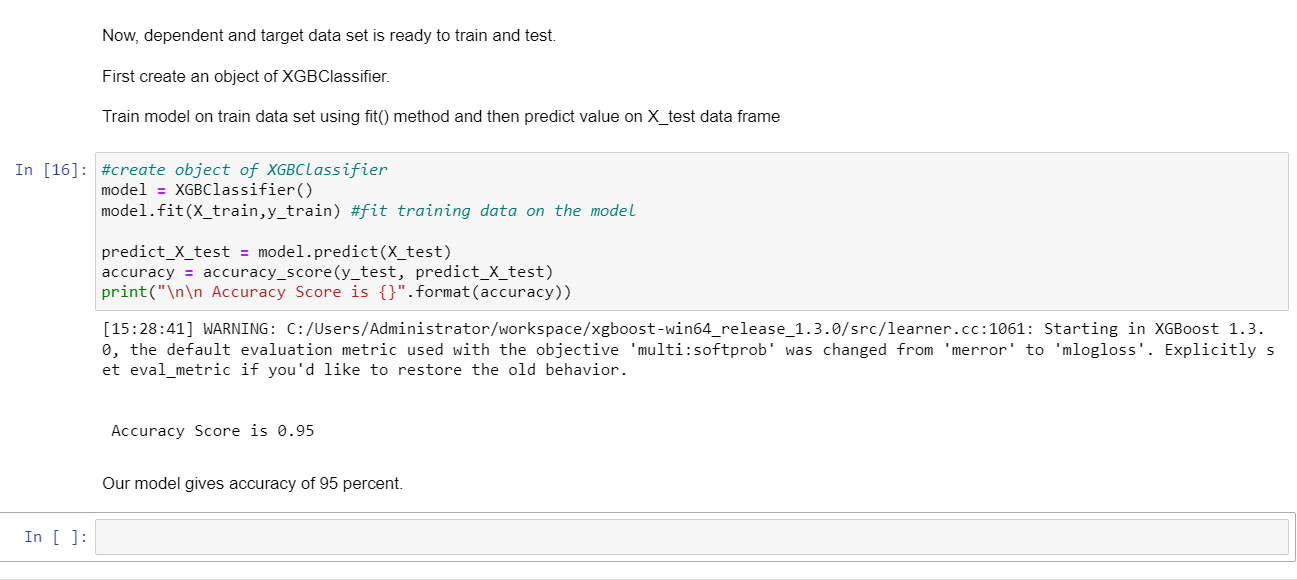
****

****

****

****

****

****