F20DL  
Group 13  
Lab 1

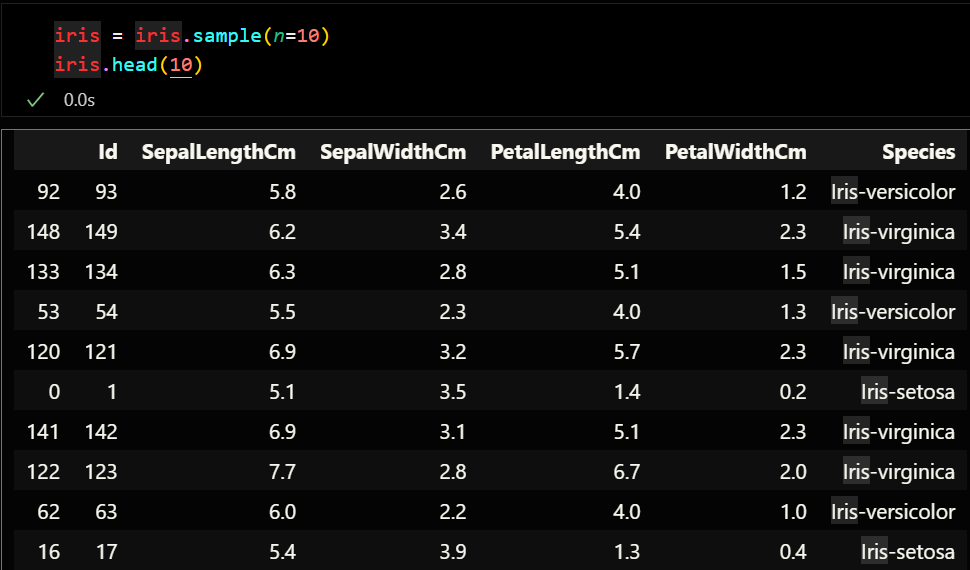
Dataset we explored:

1. Iris Species - <https://www.kaggle.com/datasets/uciml/iris>
2. Titanic - Machine Learning from Disaster - <https://www.kaggle.com/c/titanic/data>
3. Ford Car Price Prediction - <https://www.kaggle.com/datasets/adhurimquku/ford-car-price-prediction>

Iris Species Dataset

Features

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column/features** | **Explanation** | **Data type** | **Instances** | **Missing values** | **Mean value** |
| ID | unique identifier for each flower | integer | 150 | 0 | - |
| SepalLenghtCm | sepal length in cm | float | 150 | 0 | 5.84 |
| SepalWidthCm | sepal width in cm | float | 150 | 0 | 3.05 |
| PetalLengthCm | petal length in cm | float | 150 | 0 | 3.75 |
| PetalWidthCm | petal width in cm | float | 150 | 0 | 1.19 |
| Species | label defining species | object | 150 | 0 | - |

6 features with no null/missing values.  
The output variable “Species “makes this a classification problem.  


Titanic dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column/features** | **Explanation** | **Data type** | **Instances** | **Missing values** | **Mean value** |
| PassengerId | unique identifier for each passenger | integer | 891 | 0 | - |
| Survived | if the passenger survived | integer | 891 | 0 | - |
| Pclass | passenger class (1st, 2nd or 3rd class) | integer | 891 | 0 | - |
| Name | Name of the passenger | object | 891 | 0 | - |
| Sex | Gender of the passenger | object | 891 | 0 | - |
| Age | Age of the passenger | float | 714 | 177 | 29.69 |
| SibSp | siblings or spouse of the passenger | integer | 891 | 0 | - |
| Parch | parent or children of the passenger | integer | 891 | 0 | - |
| Ticket | ticket number of the passenger | object | 891 | 0 | - |
| Fare | how much the passenger pay in dollars | float | 891 | 0 | 32.2 |
| Cabin | which cabin the passenger was located in if any | object | 891 | 0 | - |
| Embarked | country passenger embarked from (S, C or Q) | object | 891 | 0 | - |

12 features with “Age” having 177 null values which could be fixed by applying the mean of all ages to these missing values.  
The output variable “Survived” makes this a classification problem.   




Ford Car Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| column/features | explanation | data type | instances | missing values | mean value |
| model | Model name | object | 17966 | 0 | - |
| year | year of production | integer | 17966 | 0 | 2016.86 |
| price | price of the car | integer | 17966 | 0 | 12279.53 |
| transmission | type of transmission(automatic, manual, semi-auto) | object | 17966 | 0 | - |
| mileage | miles travelled | integer | 17966 | 0 | 23362.6 |
| fuelType | Type of fuel (Petrol, Diesel, Hybrid, Electric, Other) | object | 17966 | 0 | 29.69 |
| tax | annual tax | integer | 17966 | 0 | 113.32 |
| mpg | efficiency/miles per gallon | float | 17966 | 0 | 57.9 |
| engine size | engine Displacement | float | 17966 | 0 | 1.35 |

9 features with no null/missing values.

The output variable “price” makes this a regression problem.

