TEECE 1/1L – ECE Elective 1/1L

M2: Exploratory Data Analysis

Lectured by:

Engr. Timothy M. Amado
Faculty, Electronics Engineering Department

Copyright Notice

Outline

Exploratory Data Analysis

- Introduction
- Example 1: Titanic Survival Dataset
- Example 2: Iris Flower Dataset

Exploratory Data Analysis

Exploratory Data Analysis refers to an approach used to analyze data sets to summarize their main characteristics, often using visual methods. Its primary objective is to discover patterns, detect anomalies, test hypotheses, and check assumptions using statistics and graphical representations.[1]

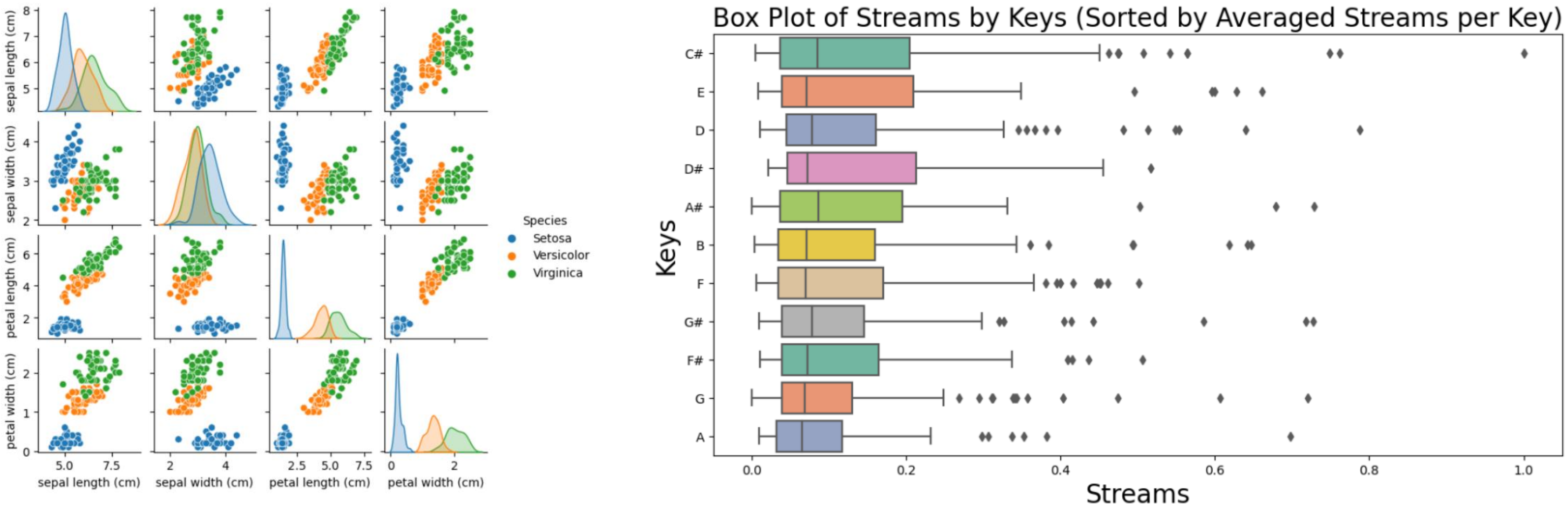


Fig.1: Correlogram visualization of the Iris Flower Data Set and box plot diagram of streams by keys of Most Streamed Spotify Songs of 2023

[1] *https://www.ibm.com/topics/exploratory-data-analysis*

Different Modalities of Data

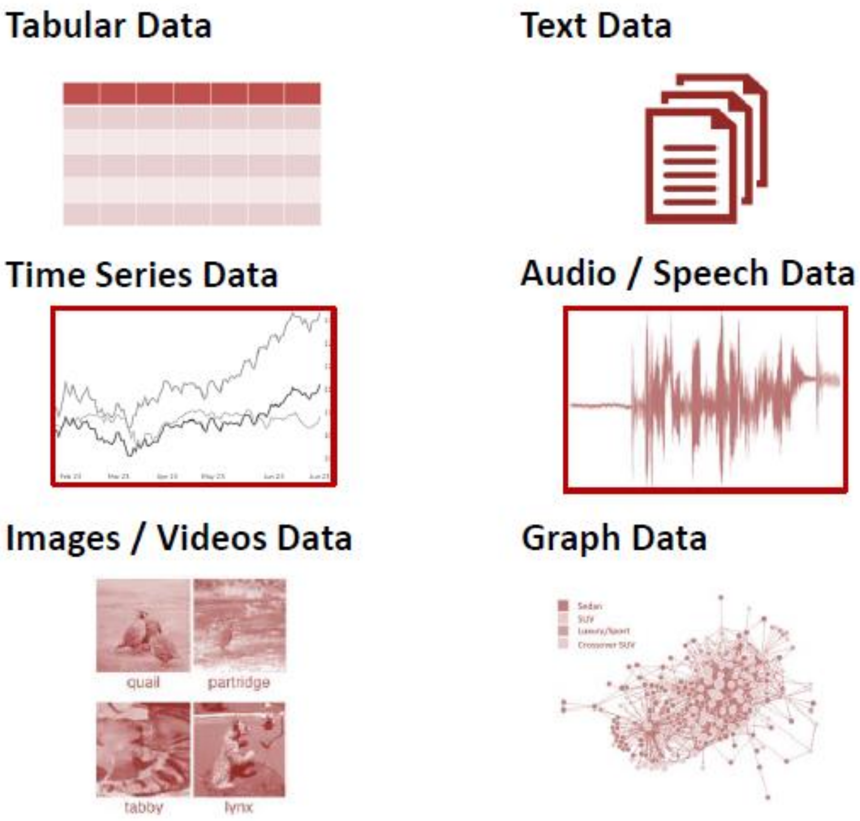


Fig.2: Different modalities of data that can be found in most datasets. Data sets can be heterogeneous, i.e., different types of data are contained in one set.

Different Modalities of Data

- Tabular Data - *Structured data arranged in rows and columns.*
- Text Data - *Unstructured text that can be analyzed for patterns and insights.*
- Time Series Data - *Sequential data points indexed in time order, often used in forecasting.*
- Audio/Speech Data - *Acoustic signals that can be processed for speech recognition and other audio analysis applications.*
- Images/Videos Data - *Visual data used in tasks like image recognition and video processing.*
- Graph Data - *Data represented in nodes and edges, useful in networking, social media analysis, and more.*

EDA Examples

Example: Titanic Survival Dataset

Contains information on 1309 passengers aboard the Titanic and whether they survived or not.

Goal: To predict the survival of passengers based on their attributes.

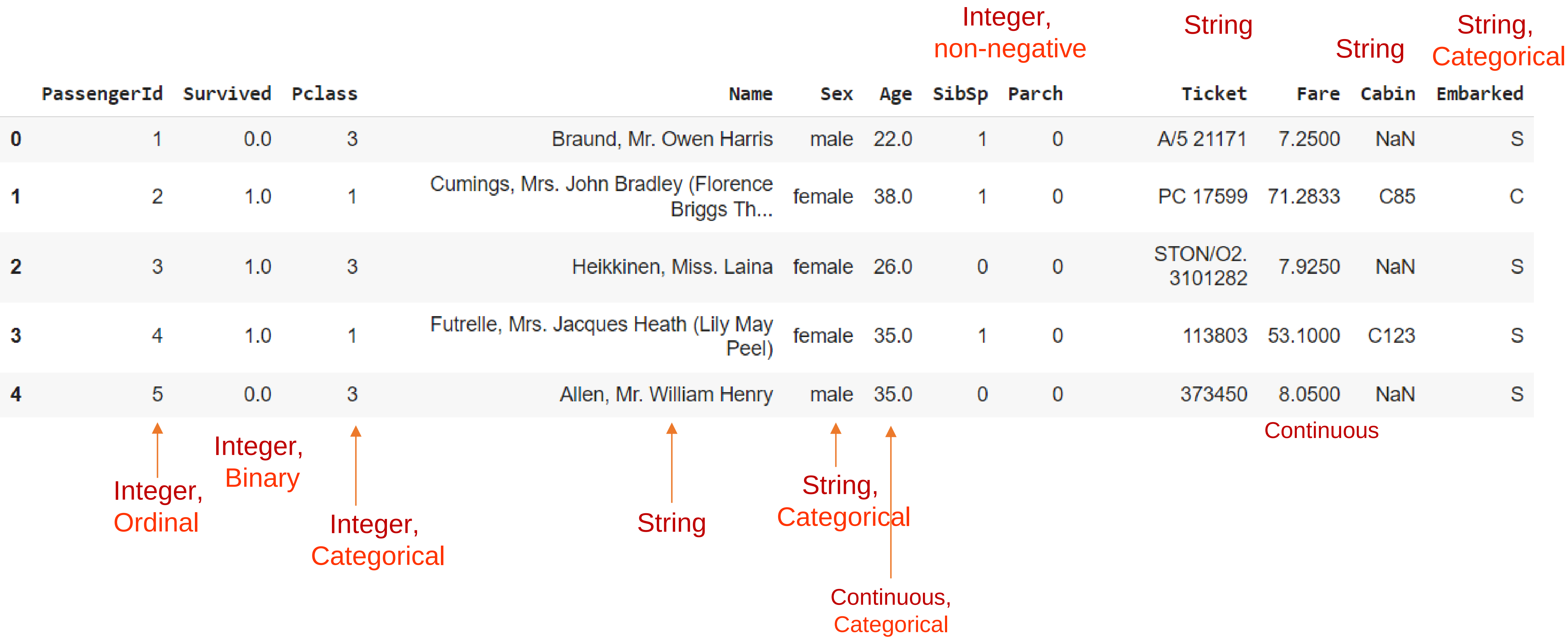
kaggle

PassengerID	An identifier unique to a passenger
Name	Passenger’s name
survival	Survival (0 = No; 1 = Yes)
pclass	Ticket class (1 = 1st, 2= 2nd, 3 = 3rd)
sex	Sex (Male/Female)
Age	Age in years
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation *(C = Cherbourg, Q = Queenstown, S = Southampton)*

Source: *https://www.kaggle.com/competitions/titanic/data*

EDA Examples

Example: Titanic Survival Dataset



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Integer, non-negative

String

String

String, Categorical

Integer, Ordinal

Integer, Binary

Integer, Categorical

String

String, Categorical

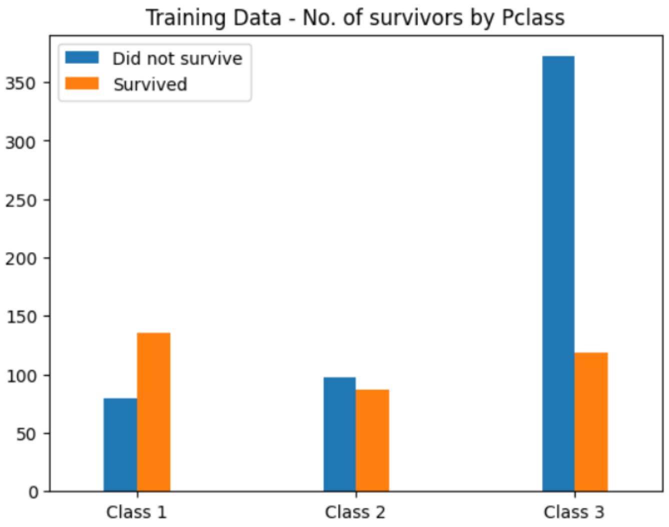
Continuous, Categorical

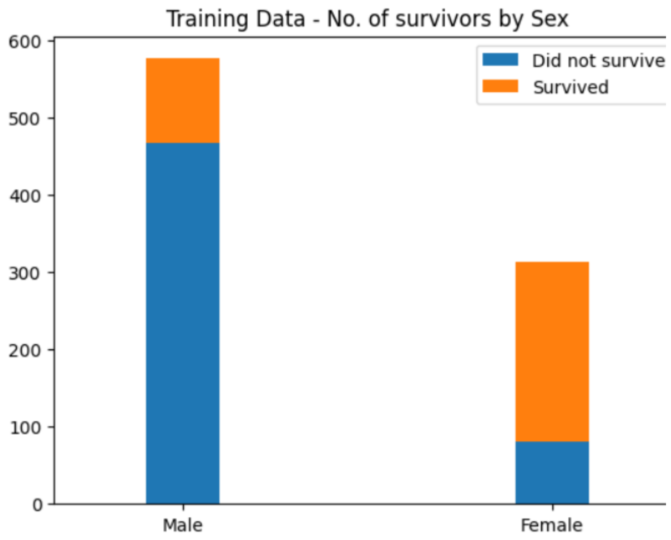
Continuous

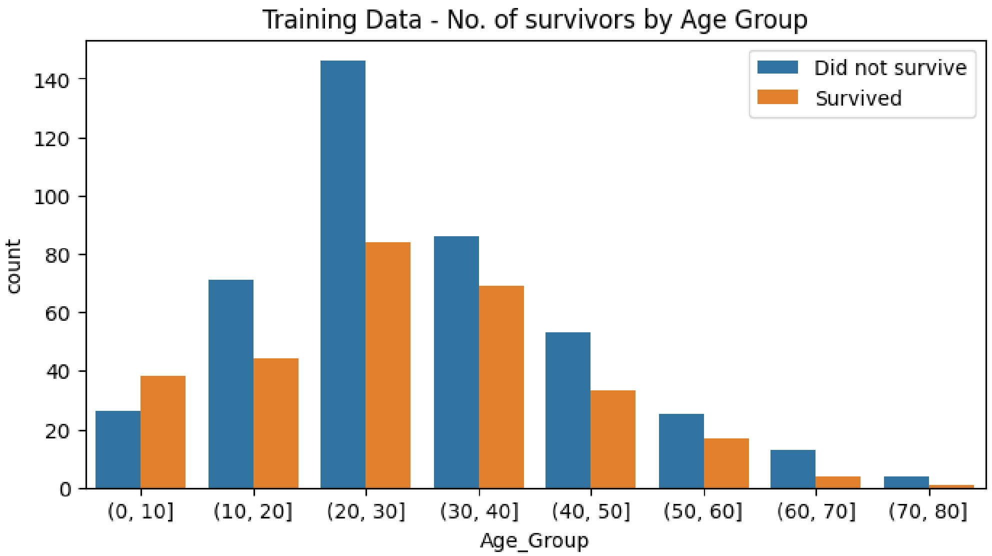
EDA Examples

Example: Titanic Survival Dataset

Bar plots and **histograms** are useful for visualizing the “count” of values in the data set.



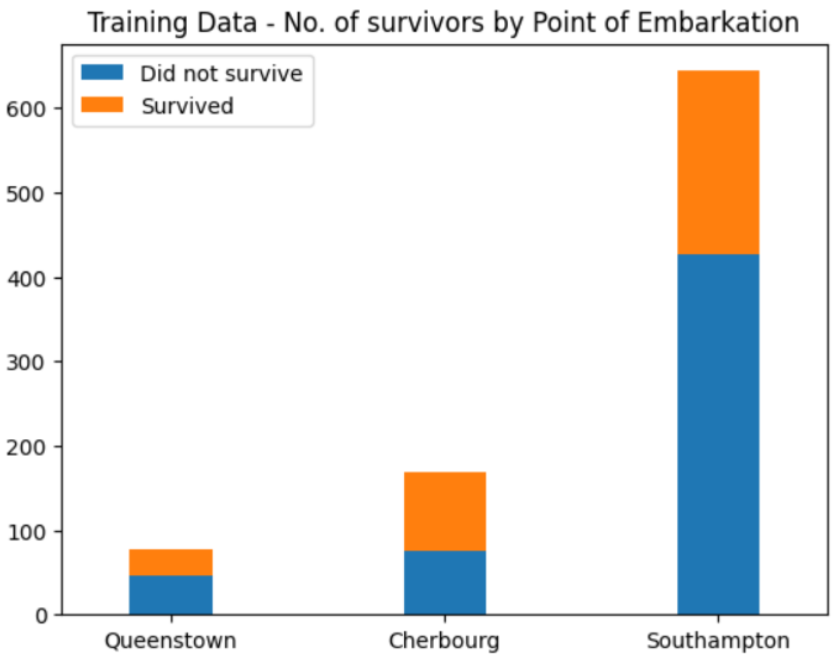


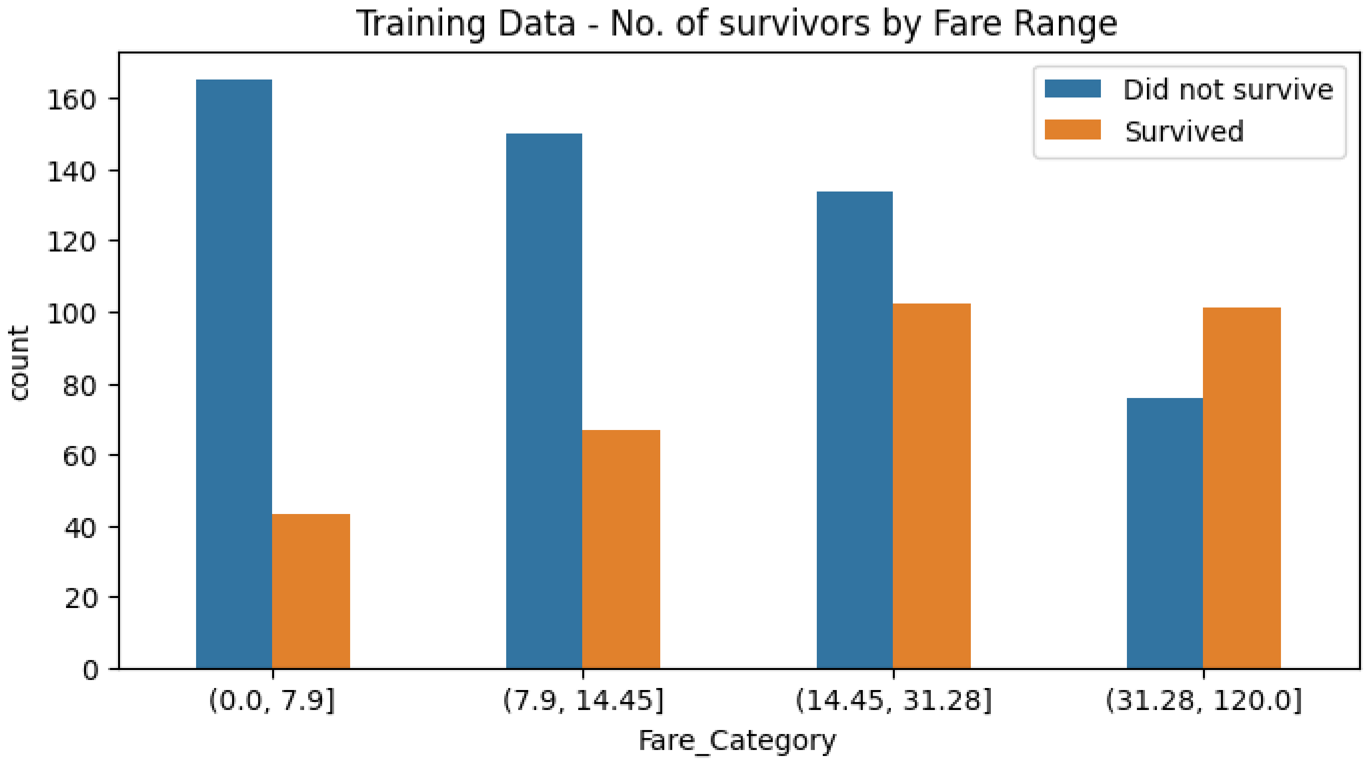


EDA Examples

Example: Titanic Survival Dataset

Bar plots and **histograms** are useful for visualizing the “count” of values in the data set.

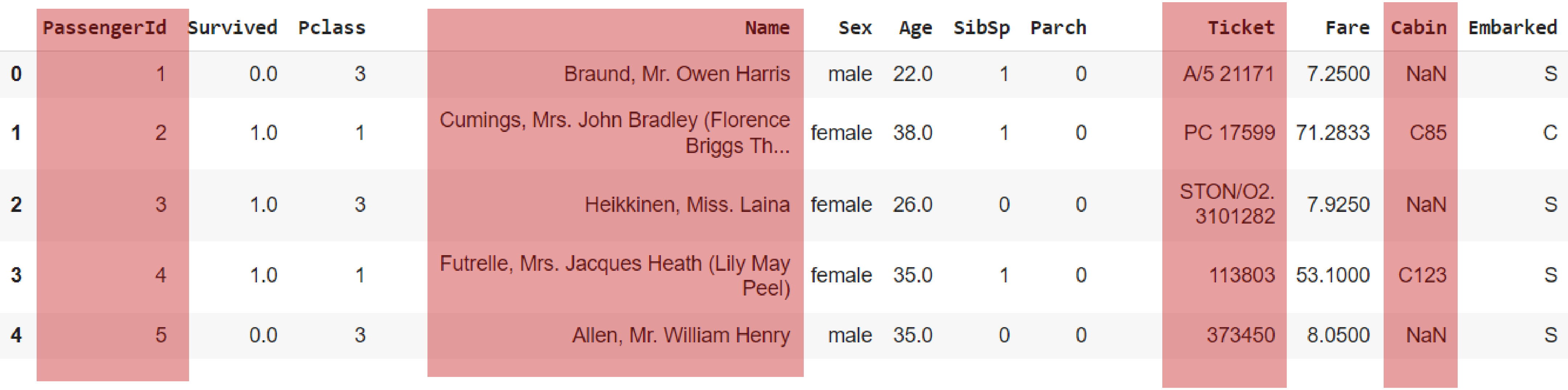




EDA Examples

Example: Titanic Survival Dataset

Before training a classifier, it is essential to keep only the relevant **numerical** and **categorical** columns, while all other columns should be dropped.



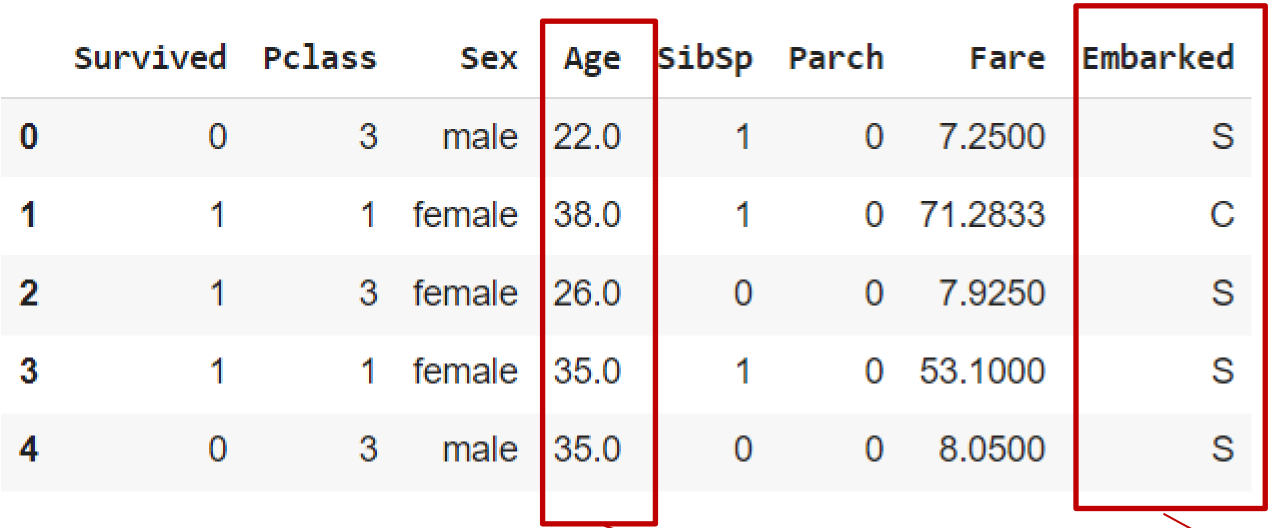
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

This is called **feature selection.**

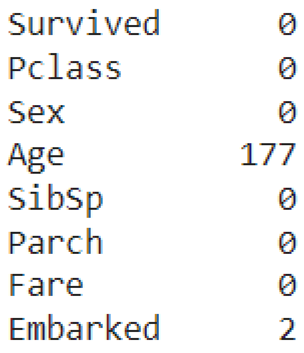
EDA Examples

Example: Titanic Survival Dataset

How to deal with missing data?



	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

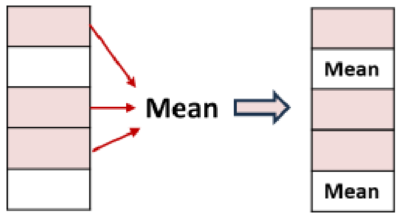


```
Survived      0
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Fare          0
Embarked      2
```

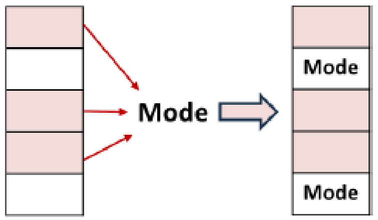
To deal with missing data we can:

- Remove rows containing missing values
- Perform **data imputation**

Mean Imputation



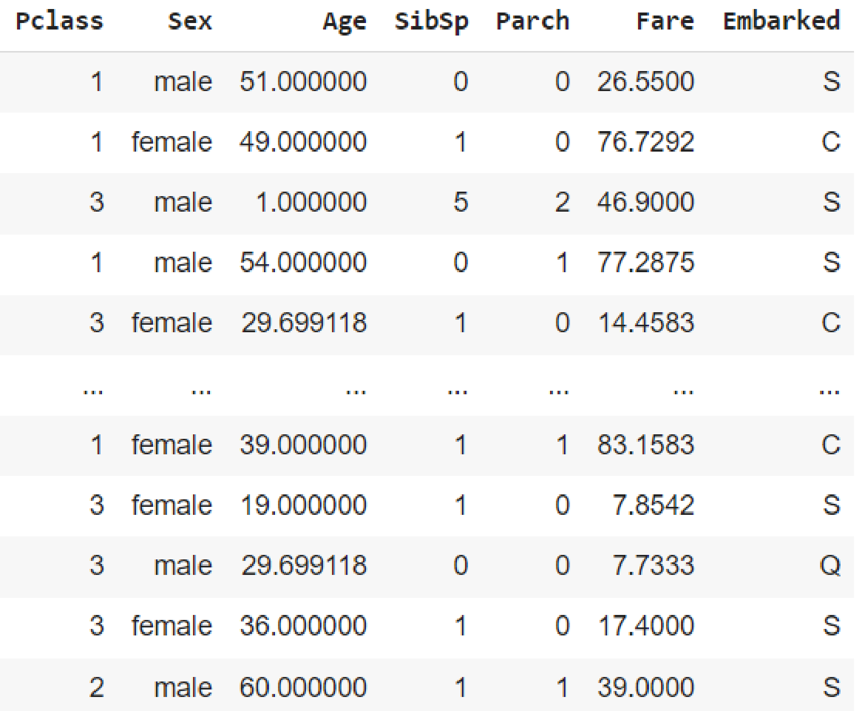
Most-frequent Imputation



EDA Examples

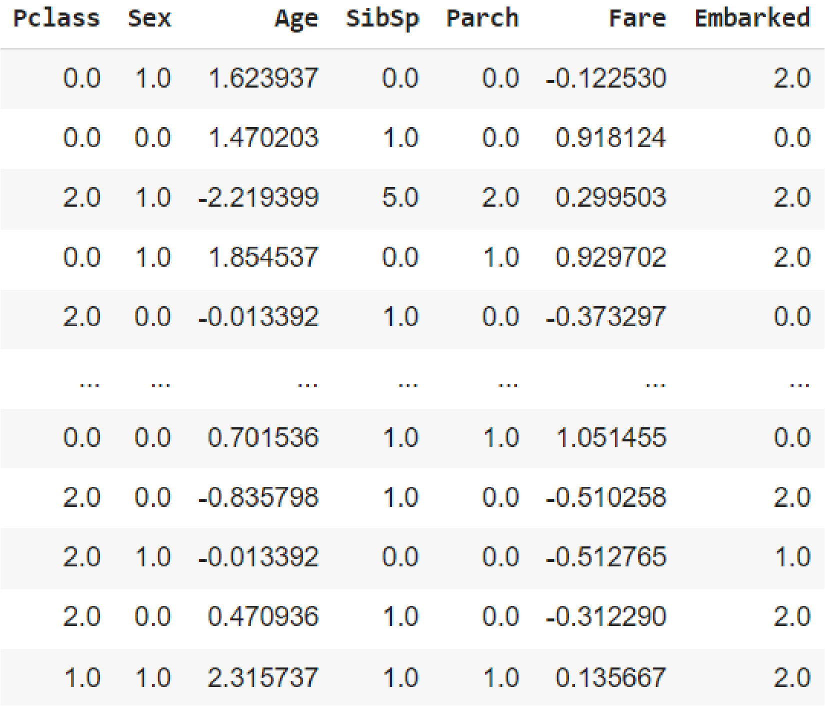
Example: Titanic Survival Dataset

Another important preprocessing task is the column transformation.



Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
1	male	51.000000	0	0	26.5500	S
1	female	49.000000	1	0	76.7292	C
3	male	1.000000	5	2	46.9000	S
1	male	54.000000	0	1	77.2875	S
3	female	29.699118	1	0	14.4583	C
...	...	...	...	...	...	...
1	female	39.000000	1	1	83.1583	C
3	female	19.000000	1	0	7.8542	S
3	male	29.699118	0	0	7.7333	Q
3	female	36.000000	1	0	17.4000	S
2	male	60.000000	1	1	39.0000	S

BEFORE

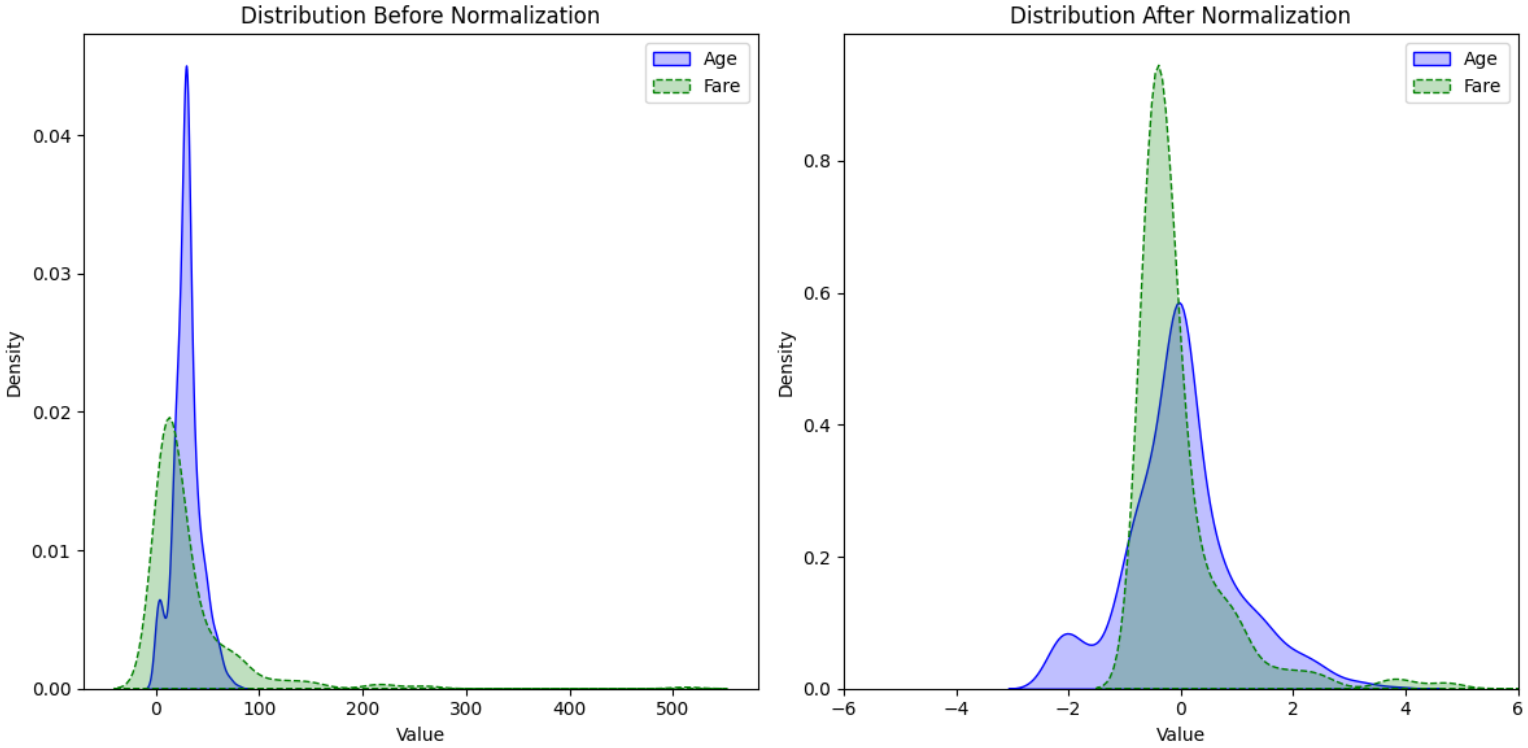


Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0.0	1.0	1.623937	0.0	0.0	-0.122530	2.0
0.0	0.0	1.470203	1.0	0.0	0.918124	0.0
2.0	1.0	-2.219399	5.0	2.0	0.299503	2.0
0.0	1.0	1.854537	0.0	1.0	0.929702	2.0
2.0	0.0	-0.013392	1.0	0.0	-0.373297	0.0
...	...	...	...	...	...	...
0.0	0.0	0.701536	1.0	1.0	1.051455	0.0
2.0	0.0	-0.835798	1.0	0.0	-0.510258	2.0
2.0	1.0	-0.013392	0.0	0.0	-0.512765	1.0
2.0	0.0	0.470936	1.0	0.0	-0.312290	2.0
1.0	1.0	2.315737	1.0	1.0	0.135667	2.0

AFTER

Data Normalization

- Normalization removes the effect of differing scales and biases.
- All data are centered to zero-mean and scaled to unit-variance



Data Standardization
(Standard scaler)

$$x_i' = \frac{x_i - \mu}{\sigma_P}$$

Min-max scaler

$$x_i' = \frac{x_i - \min x_i}{\max x_i - \min x_i}$$

Max-abs scaler

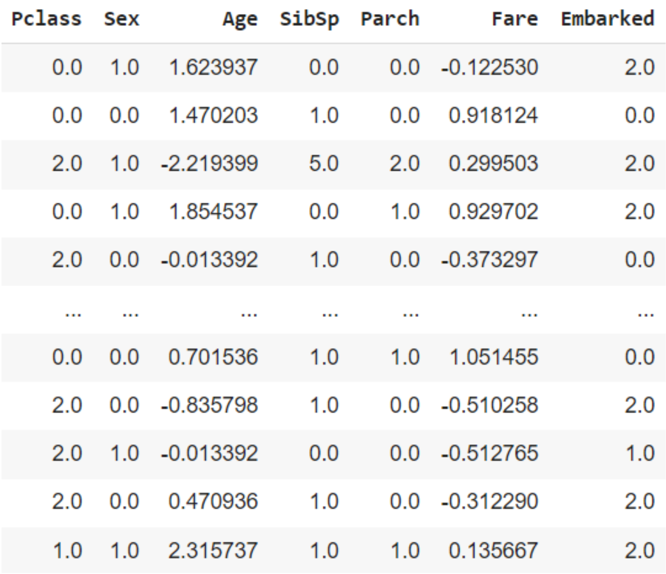
$$x_i' = \frac{x_i}{\max |x_i|}$$

- The scatter of data is preserved.
- Normalization improves machine learning by treating all features equally.

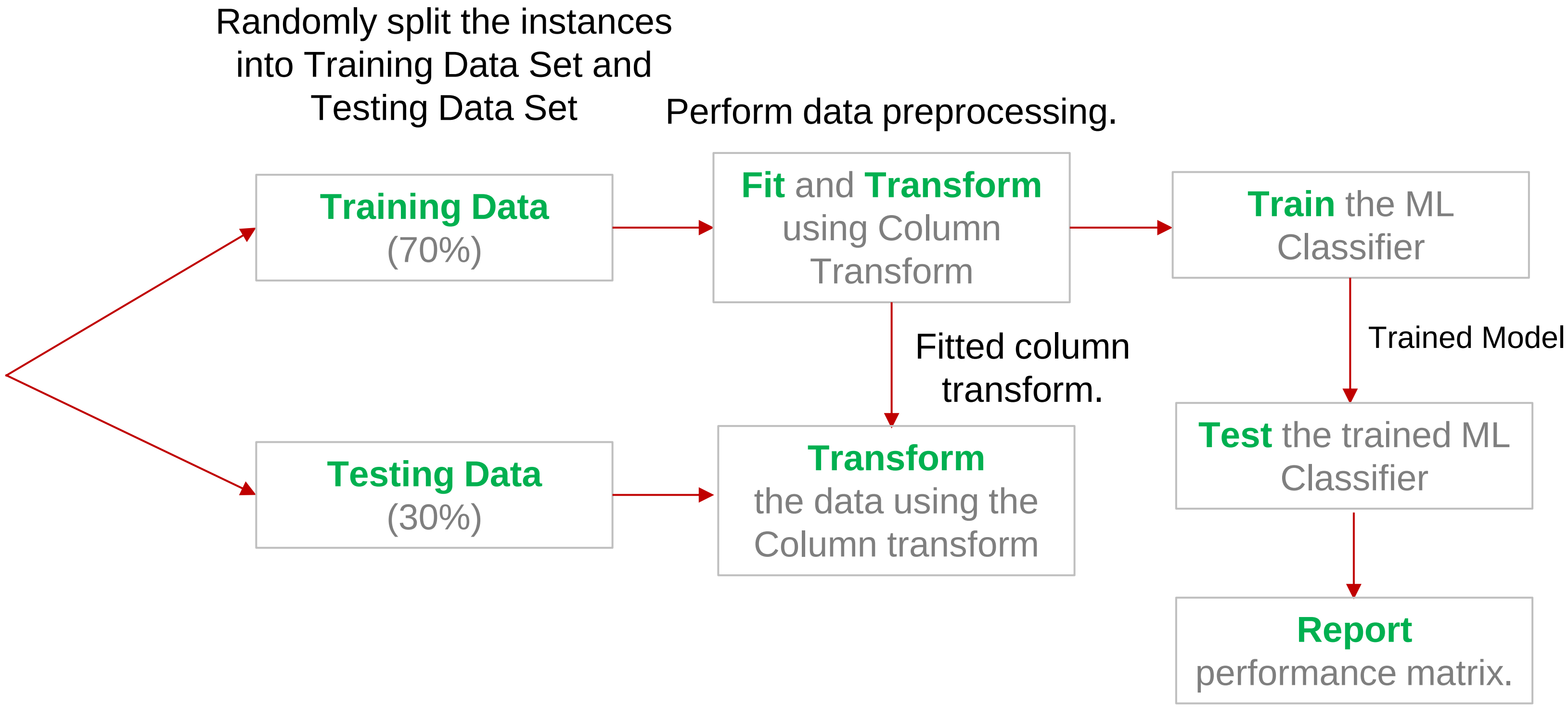
EDA Examples

Example: Titanic Survival Dataset

We can now set the ML pipeline



Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0.0	1.0	1.623937	0.0	0.0	-0.122530	2.0
0.0	0.0	1.470203	1.0	0.0	0.918124	0.0
2.0	1.0	-2.219399	5.0	2.0	0.299503	2.0
0.0	1.0	1.854537	0.0	1.0	0.929702	2.0
2.0	0.0	-0.013392	1.0	0.0	-0.373297	0.0
...	...	...	...	...	...	...
0.0	0.0	0.701536	1.0	1.0	1.051455	0.0
2.0	0.0	-0.835798	1.0	0.0	-0.510258	2.0
2.0	1.0	-0.013392	0.0	0.0	-0.512765	1.0
2.0	0.0	0.470936	1.0	0.0	-0.312290	2.0
1.0	1.0	2.315737	1.0	1.0	0.135667	2.0



Randomly split the instances into Training Data Set and Testing Data Set

Perform data preprocessing.

Training Data (70%)

Fit and **Transform** using Column Transform

Train the ML Classifier

Fitted column transform.

Trained Model

Testing Data (30%)

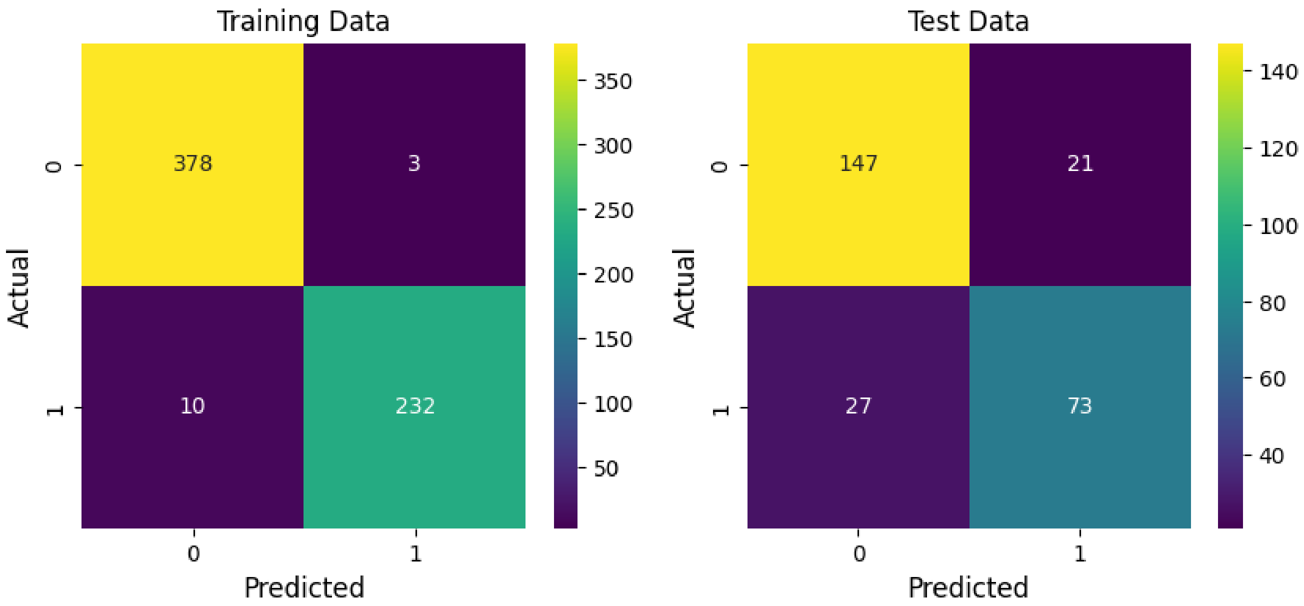
Transform the data using the Column transform

Test the trained ML Classifier

Report performance matrix.

EDA Examples

Example: Titanic Survival Dataset

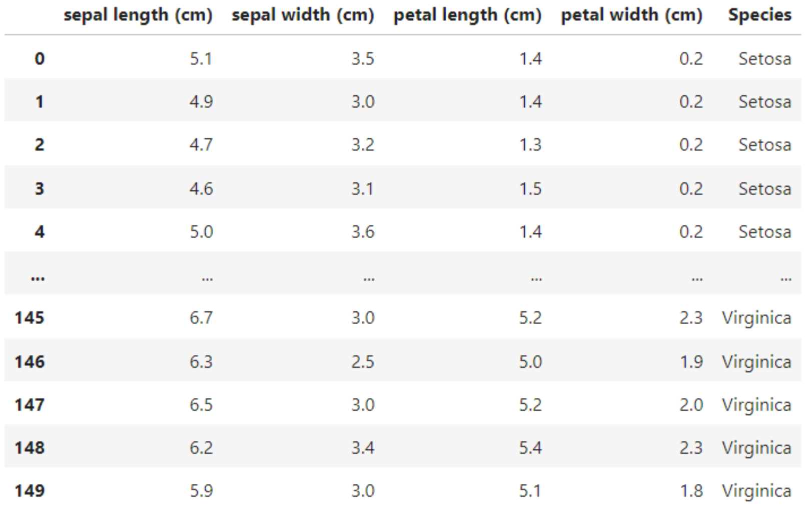


```
RF train accuracy: 0.979
RF test accuracy: 0.821
```

EDA Examples

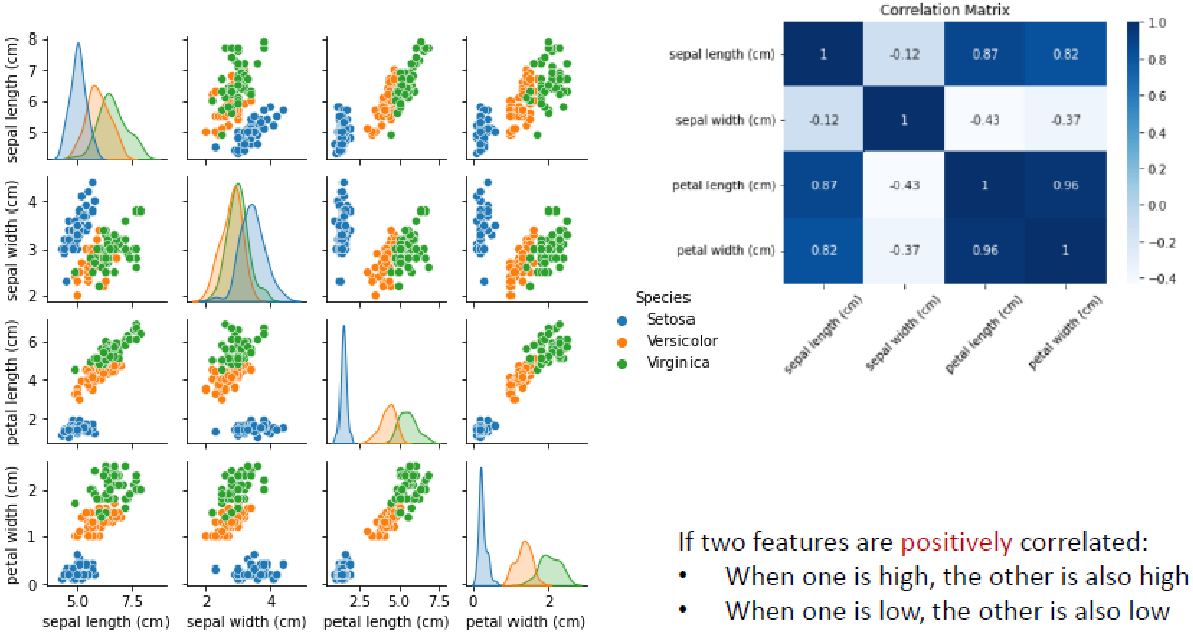
Example: Iris Flower Data Set

The data set contains measurements of 150 iris flowers in terms of their sepal length, sepal width, petal length, and petal width. There are 3 species of flowers, Setosa, Versicolor, and Virginica, with 50 samples each.



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	Species
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Virginica
146	6.3	2.5	5.0	1.9	Virginica
147	6.5	3.0	5.2	2.0	Virginica
148	6.2	3.4	5.4	2.3	Virginica
149	5.9	3.0	5.1	1.8	Virginica

Pair plots (or **correlograms / correlation matrices**) are useful for finding correlated features.



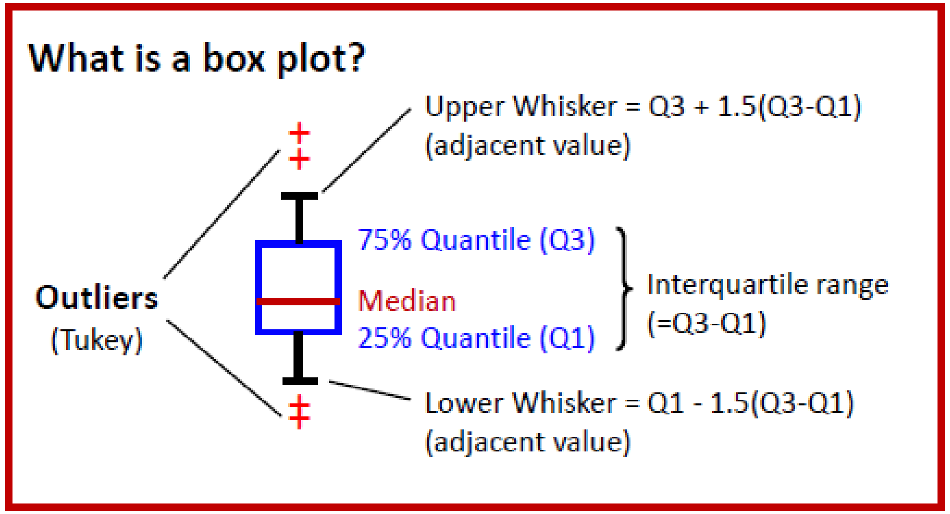
If two features are positively correlated:

- When one is high, the other is also high
- When one is low, the other is also low

EDA Examples

Example: Iris Flower Data Set

Box-and-whisker plots (or simply box plots) and violin plots are useful for visualizing the distribution of values.


What is a box plot?

Upper Whisker = Q3 + 1.5(Q3-Q1) (adjacent value)

75% Quantile (Q3)

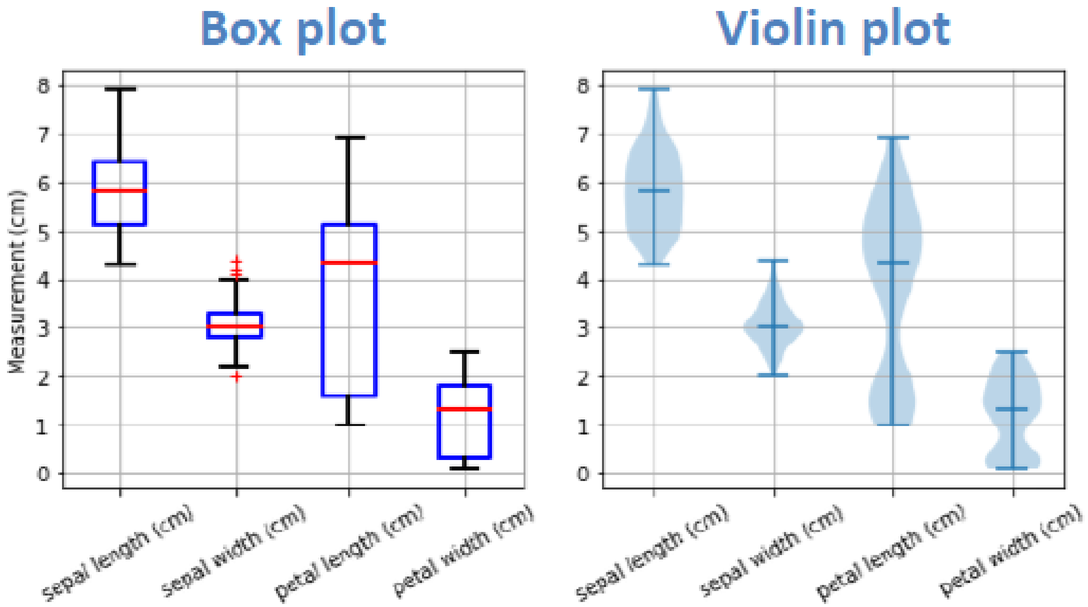
Median

25% Quantile (Q1)

Interquartile range (=Q3-Q1)

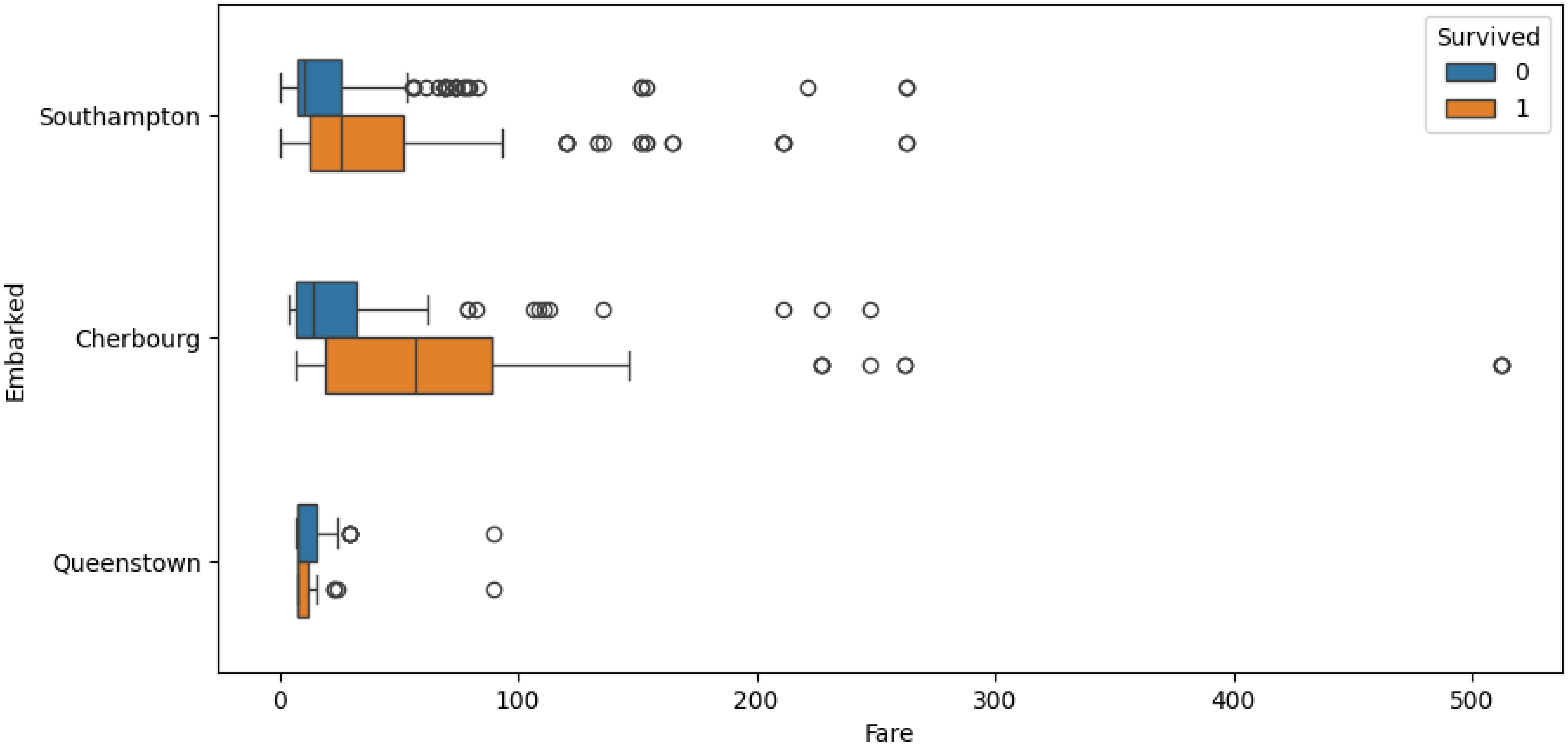
Outliers (Tukey)

Lower Whisker = Q1 - 1.5(Q3-Q1) (adjacent value)



EDA Examples

Example: Titanic Survival Dataset



Thank You!