Data Science | 30 Days of Machine Learning | Day - 8

Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

----Today Topics | Day 08----

- **Panda Profiling**

  How do we use the pandas profiling tool?
- **Feature Engineering**
- Feature Transformation
- Feature Construction
- Feature Selection
- Feature Selection

Dataset Link Kaggle: https://www.kaggle.com/competitions/titanic

GitHub Link: https://github.com/TheiScale/30_Days_Machine_Learning/

Github Link:

#Import Library | pandas

```
import pandas as pd
```

#Import Datasets: Titanic

```
df = pd.read_csv('train.csv')
```

#View Datasets

```
df.head()
```

#Pandas Profiling

```
pip install ydata-profiling
```

#Import Profile Report

```
from ydata_profiling import ProfileReport
prof = ProfileReport(df)
prof.to_file(output_file='output.html')
```

```
-----
-----
```

Generate and Analysis Report <output.html>

```
----
```

#Import Datasets: Hotel

```
hotel = pd.read_csv('hotel.csv')
```

#View Datasets

```
hotel.head()
```

#Pandas Profiling

```
pip install ydata-profiling
```

#Import Profile Report

```
from ydata_profiling import ProfileReport
prof=ProfileReport(hotel)
prof.to_file(output_file='output.html')
```

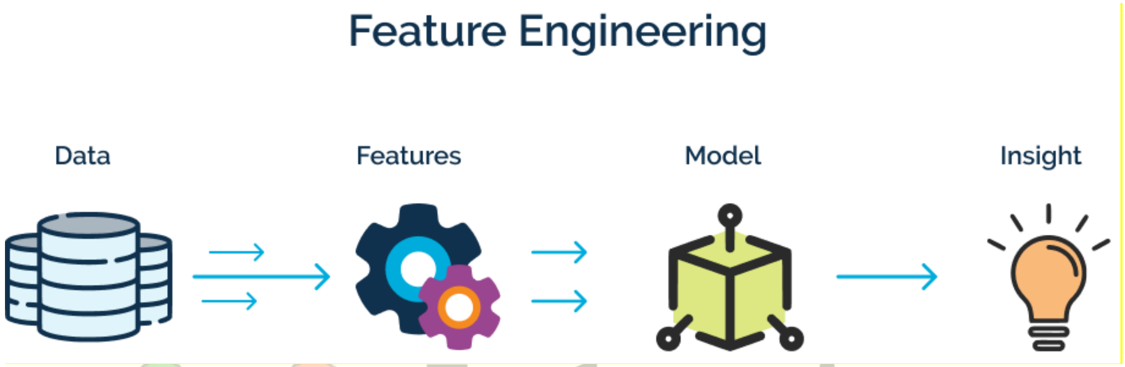
```
-----
-----
```

Generate and Analysis Report <output.html>

```
----
----
```

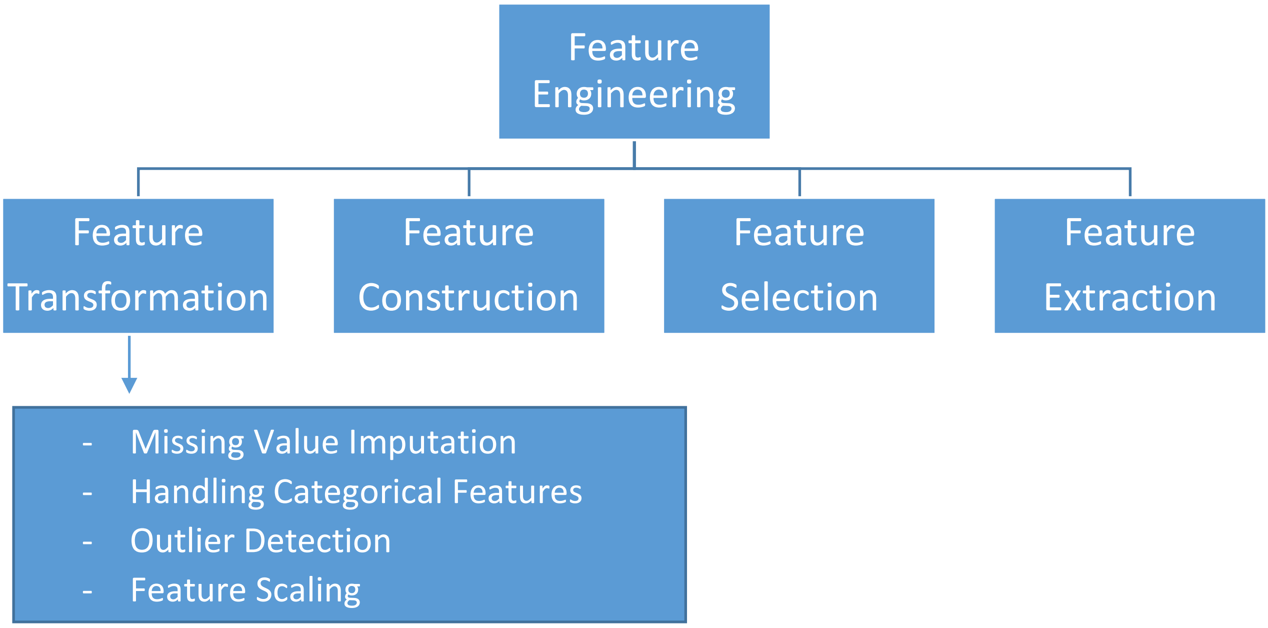
- Feature Engineering

- MLDLC
  - Frame the problem → Gathering data →Data pre-processing → Exploratory data analysis (EDA) → **Feature Engineering**



What is Feature Engineering:

Feature engineering is the process that takes raw data and transforms it into features that can be used to create a predictive model using machine learning.



Feature Engineering

Feature Transformation

Feature Construction

Feature Selection

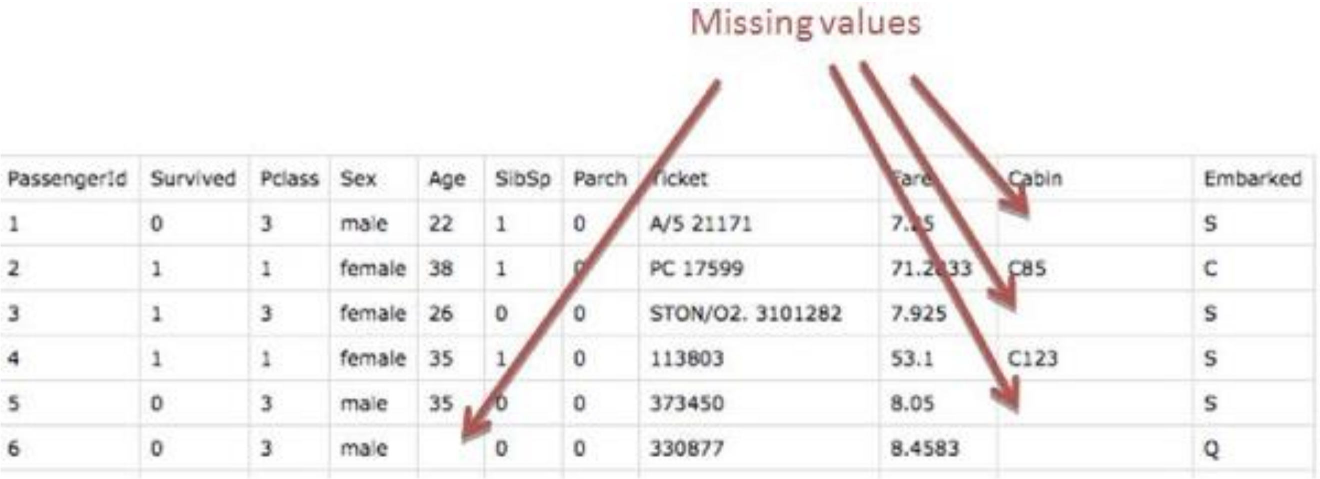
Feature Extraction

- Missing Value Imputation
- Handling Categorical Features
- Outlier Detection
- Feature Scaling

- Missing Value Imputation:

What are imputed values?

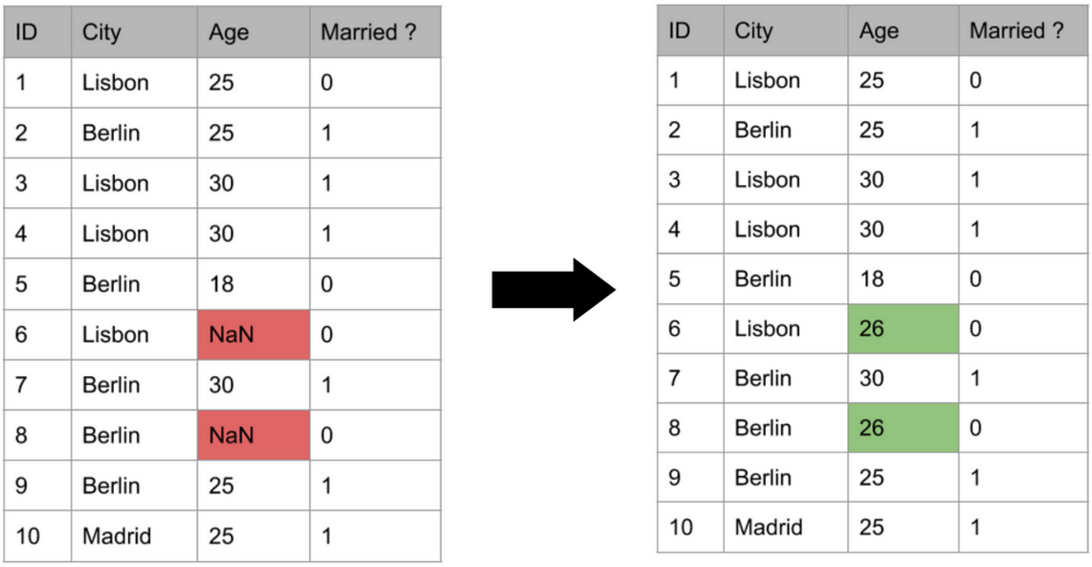
Imputed value, also known as estimated imputation, is an assumed value given to an item when the actual value is not known or available. Imputed values are a logical or implicit value for an item or time set, wherein a "true" value has yet to be ascertained.



Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Average_Age = 26.0

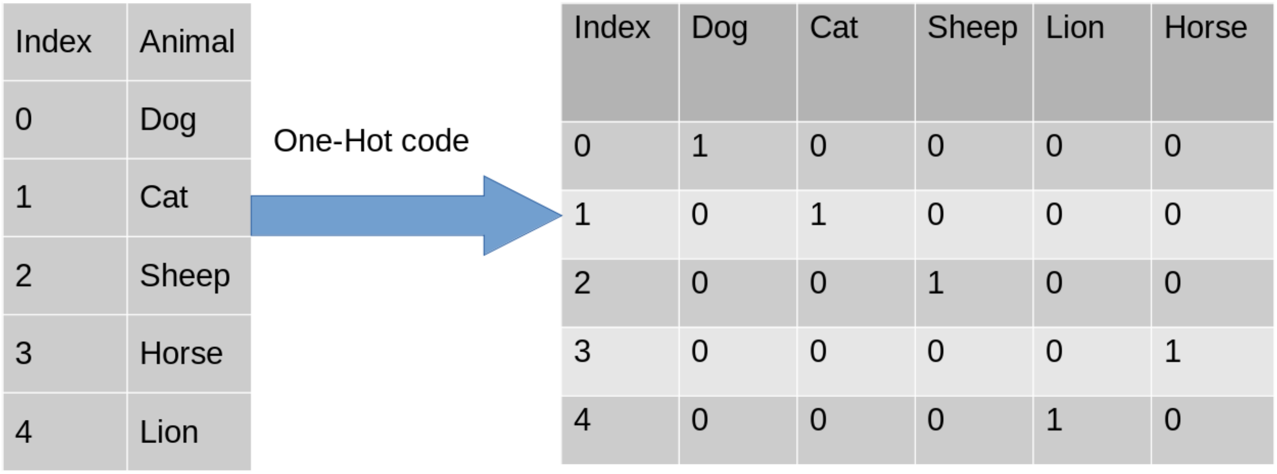


ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	NaN	0
7	Berlin	30	1
8	Berlin	NaN	0
9	Berlin	25	1
10	Madrid	25	1

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	26	0
7	Berlin	30	1
8	Berlin	26	0
9	Berlin	25	1
10	Madrid	25	1

- Handling Categorical Values

How do you handle categorical data?
One of the most common ways to deal with categorical data in machine learning is through a process called one-hot encoding. This technique involves converting categorical data into numerical data by creating a new binary feature for each category.



Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

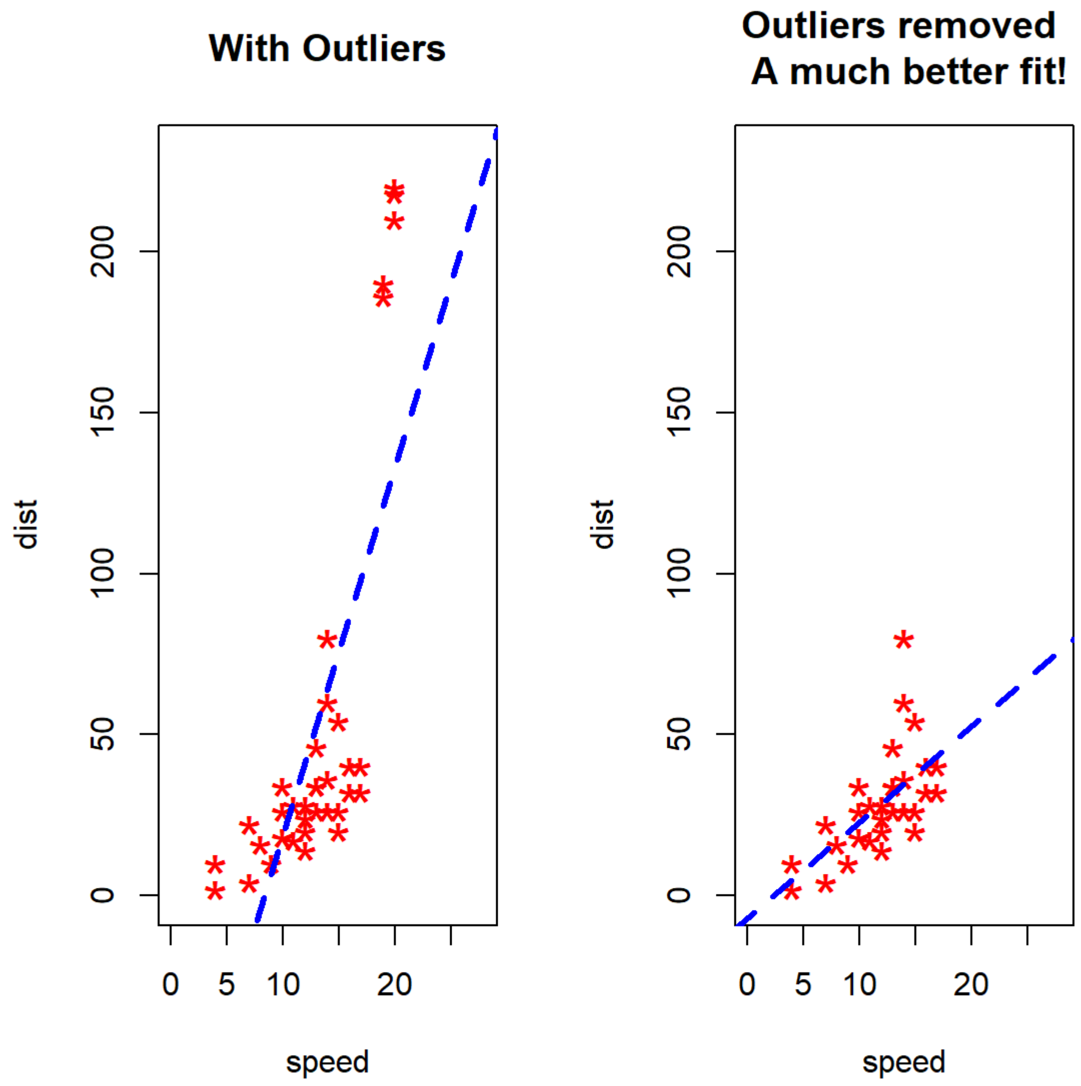
One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

- Outlier Detection

Outlier detection is the process of detecting outliers, or a data point that is far away from the average, and depending on what you are trying to accomplish, potentially removing or resolving them from the analysis to prevent any potential skewing.

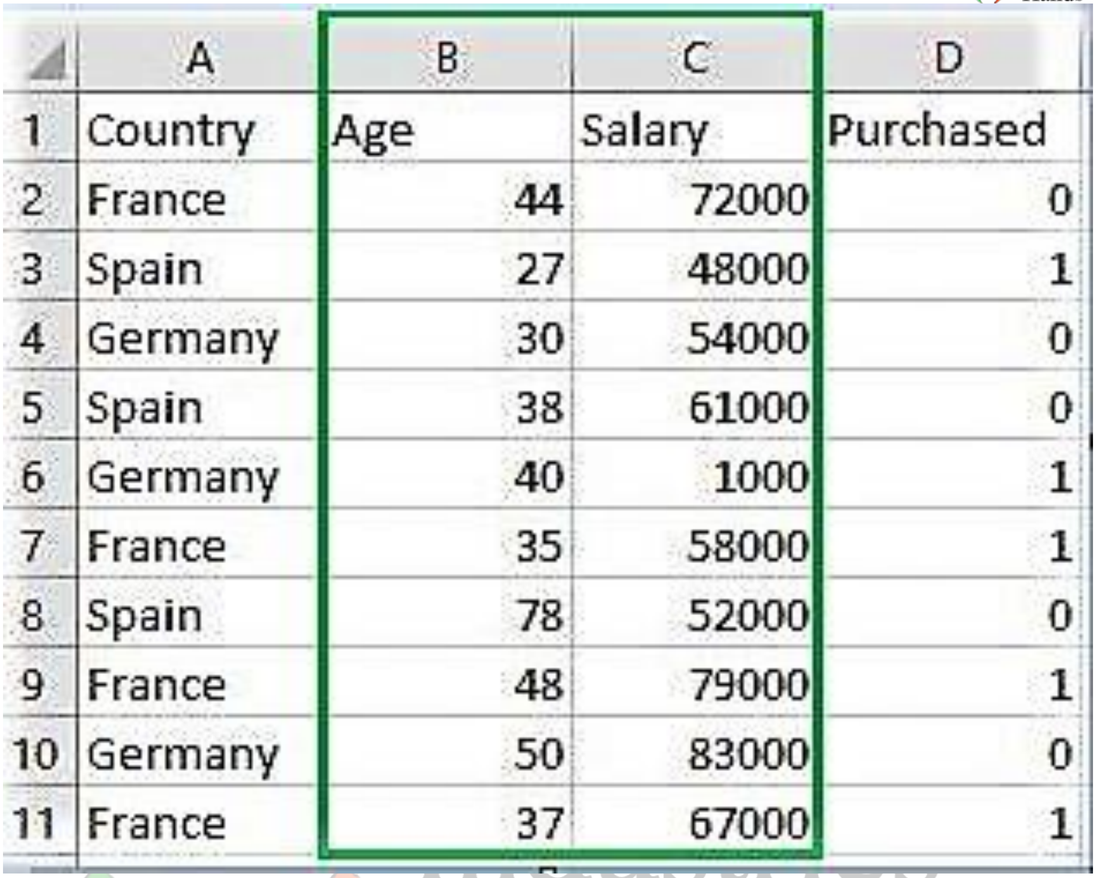
How do you find outliers in linear regression?
We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points outside this extra pair of lines are flagged as potential outliers.



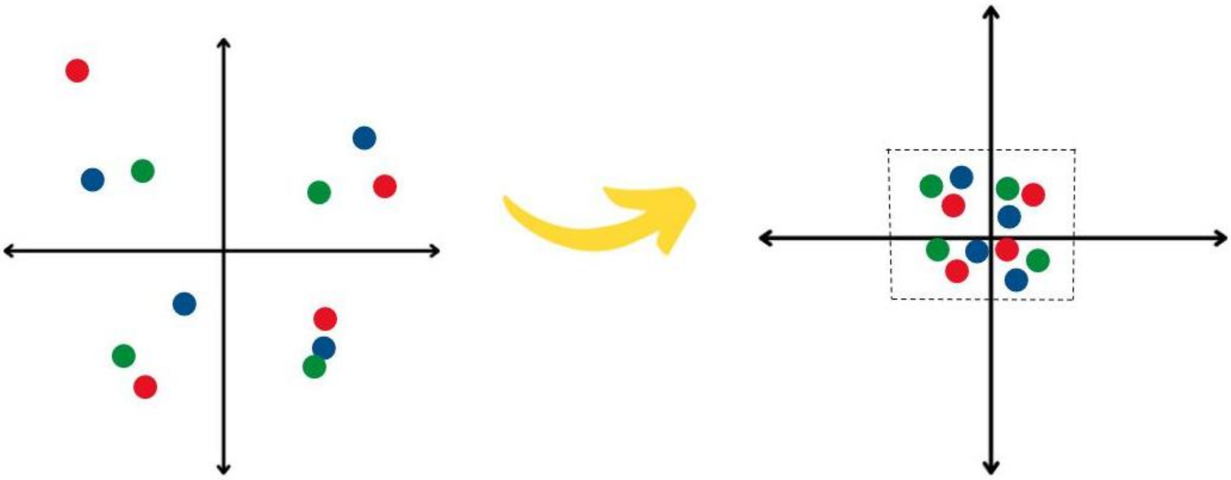
- Feature Scaling

What is Feature Scaling?

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

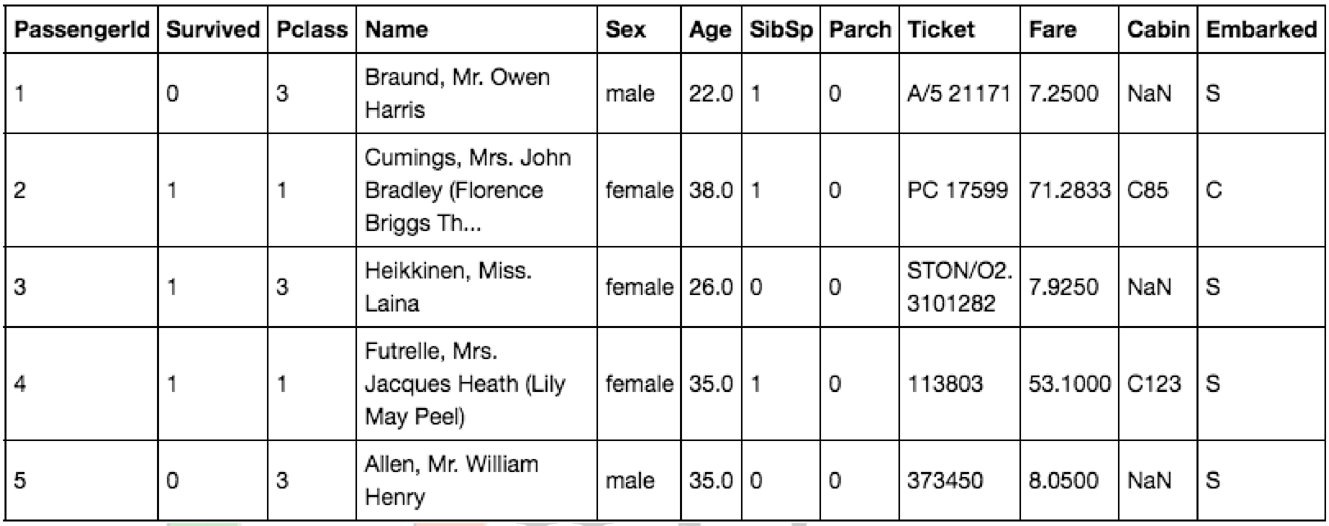


	A	B	C	D
1	Country	Age	Salary	Purchased
2	France	44	72000	0
3	Spain	27	48000	1
4	Germany	30	54000	0
5	Spain	38	61000	0
6	Germany	40	1000	1
7	France	35	58000	1
8	Spain	78	52000	0
9	France	48	79000	1
10	Germany	50	83000	0
11	France	37	67000	1



- **Feature Construction**

In the case of the Titanic dataset, two columns are available: “sibsp” (number of siblings/spouses aboard) and “parent” (number of parents/children aboard). To create a new feature called “family type,” you can combine these columns and assign a specific value to indicate the family size.

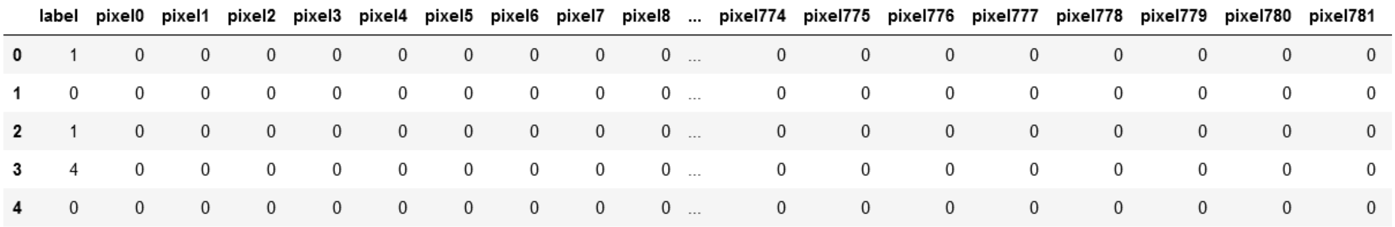


PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282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

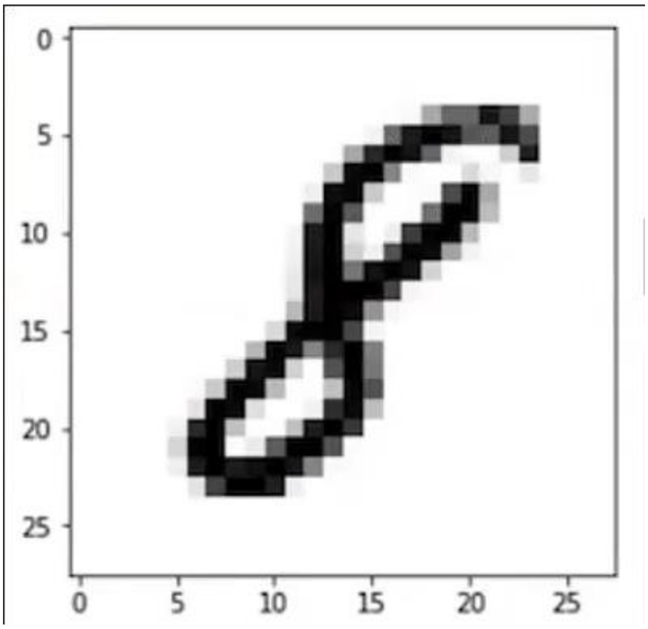
- **Feature selection**

Feature selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

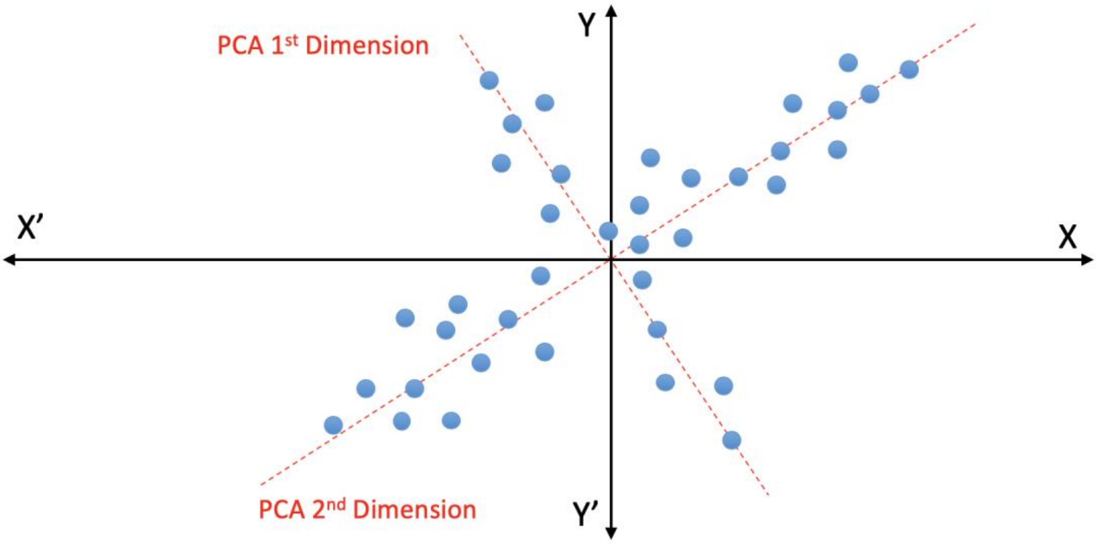


	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	...	pixel774	pixel775	pixel776	pixel777	pixel778	pixel779	pixel780	pixel781
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0



- **Feature Extraction**

What Is Feature Extraction?
Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data.



Data Story Telling (Day 8): Curious Data Minds

How Uber uses data science to reinvent transportation?

Read Blog: https://www.projectpro.io/article/how-uber-uses-data-science-to-reinvent-transportation/290

Read Blog: https://jagan-singhh.medium.com/data-science-at-uber-4380bf8f6aca

Movie Trailer Link (Super Pumped) : https://youtu.be/VMP21LO0Guc