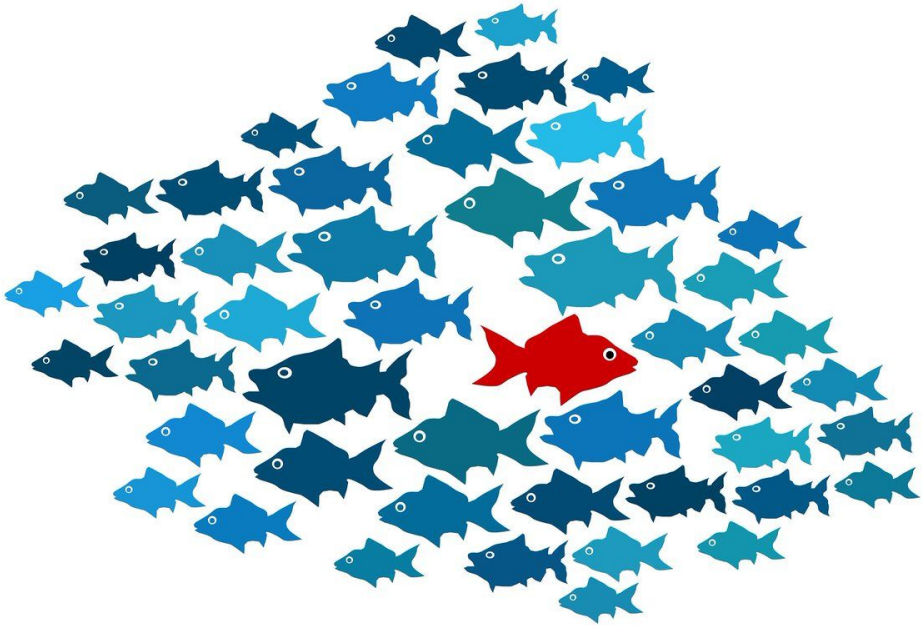
Data Preprocessing

Part-2

Table of Contents

Outlier Detection



a person or thing situated away or detached from the main body or system

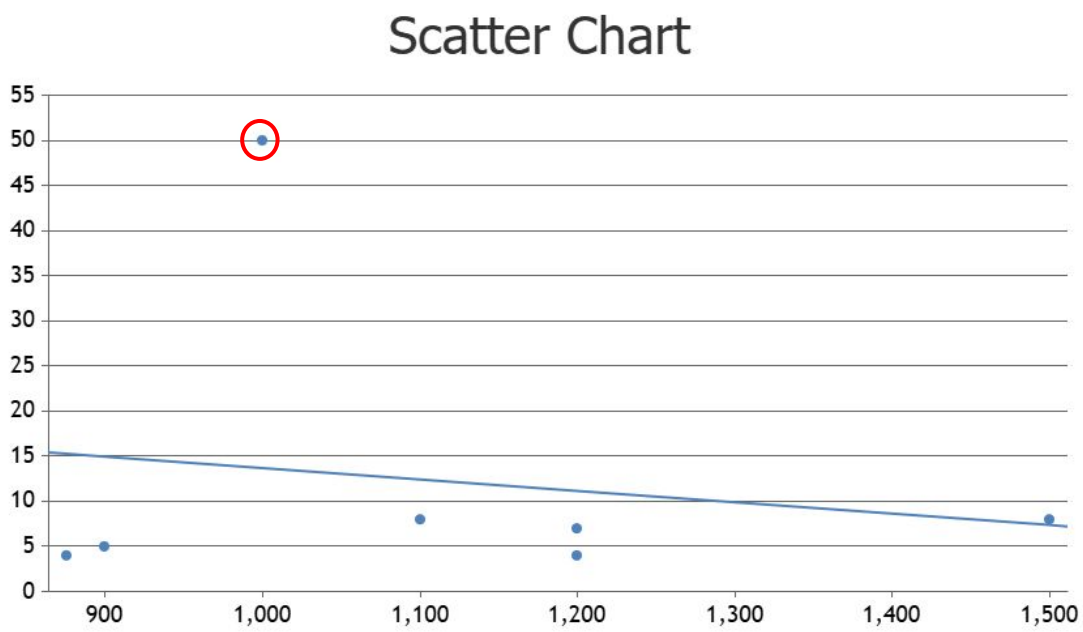
- Oxford

Outlier Detection

Unit	No. of rooms	Price
900	5	7000
1200	4	10,000
1500	8	12,000
876	4	6,000
1000	50	7,500
1200	7	12,000
1100	8	13

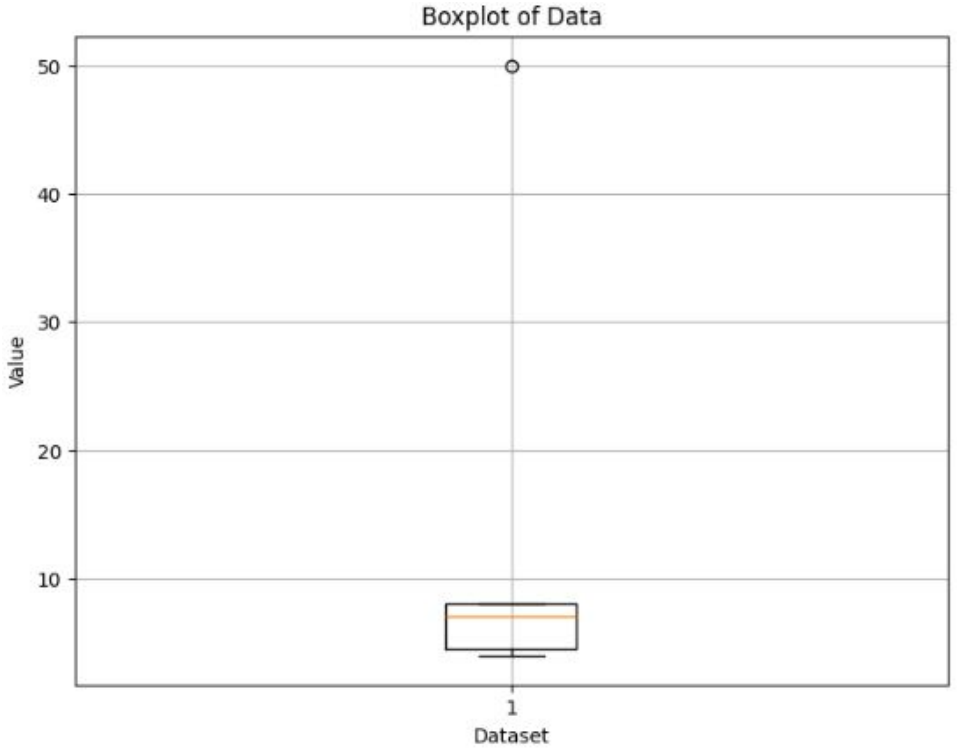
How to detect Outlier?

Visualize Data



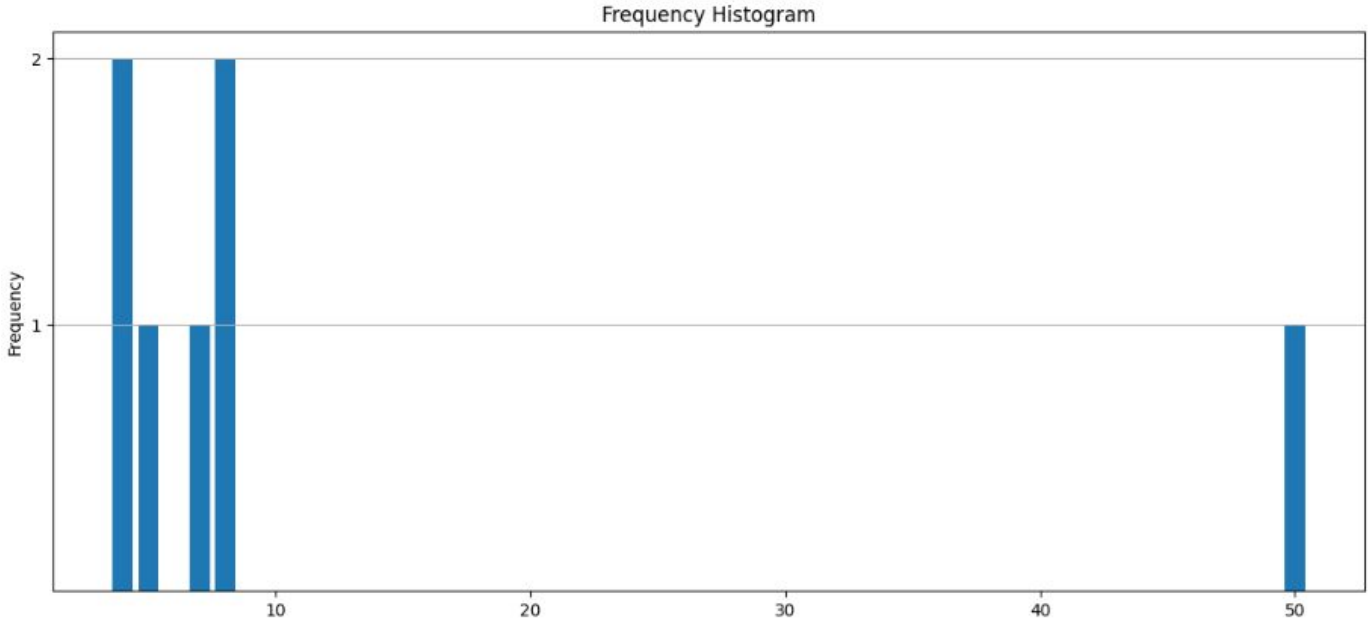
How to detect Outlier?

Visualize Data



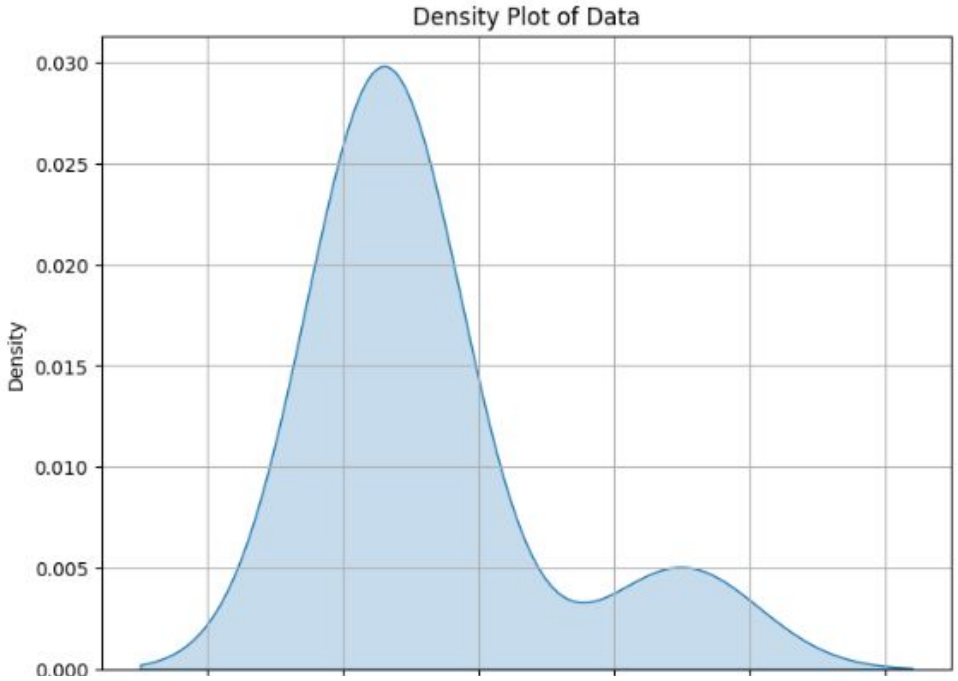
How to detect Outlier?

Visualize Data



How to detect Outlier?

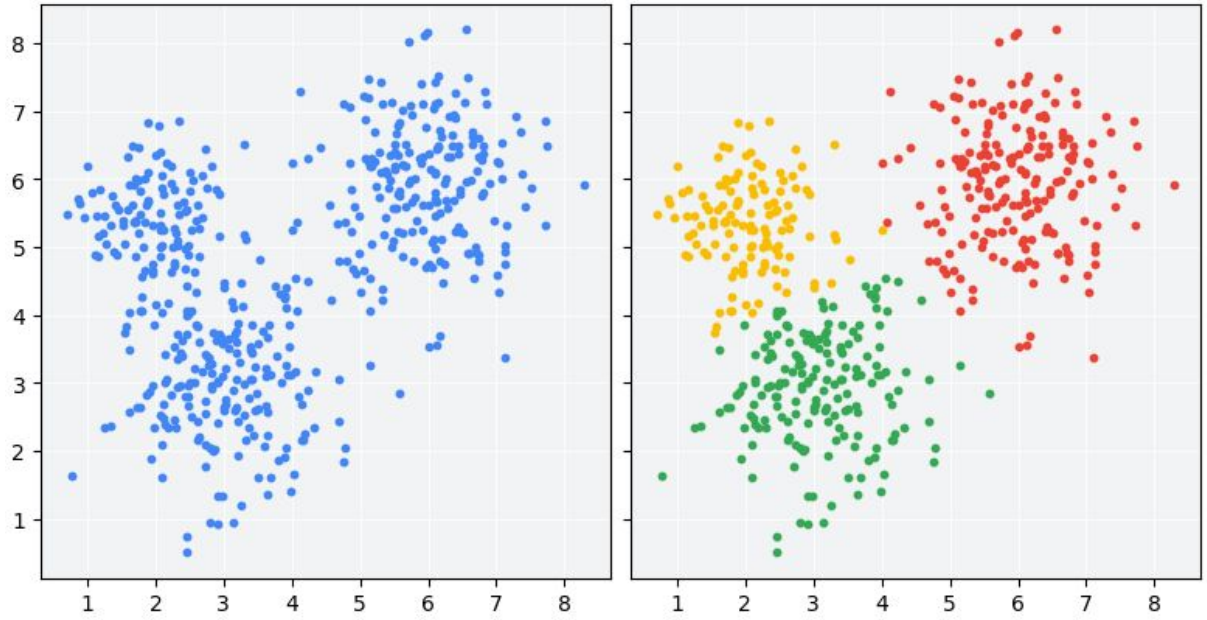
Visualize Data



How to detect Outlier?

Machine Learning Methods:

- Clustering
- Isolation Forest
- One-Class SVM



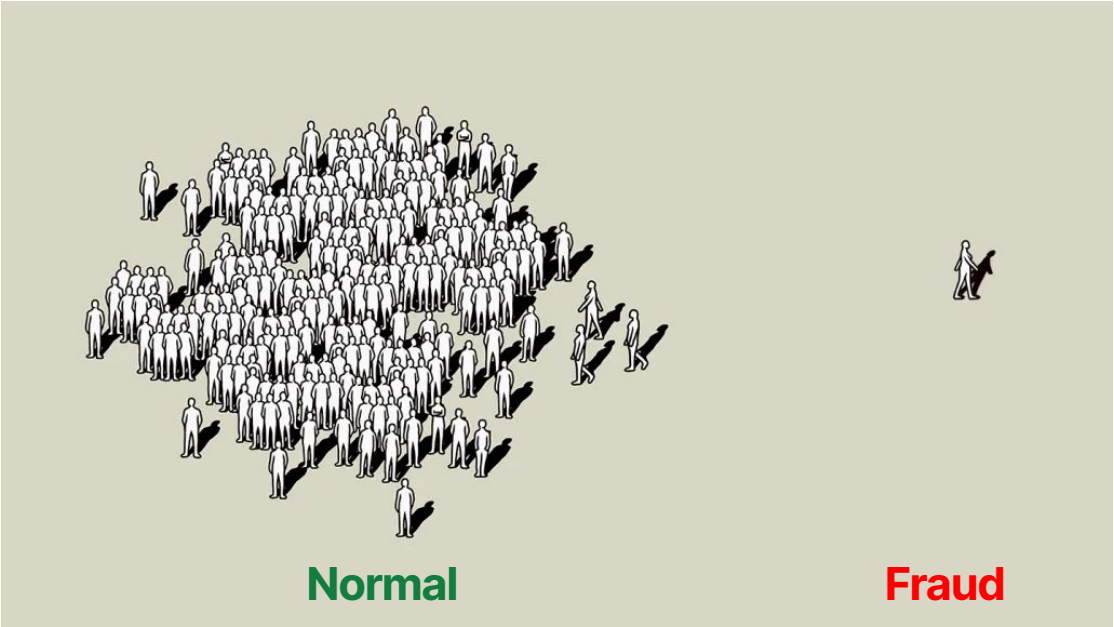
Removing Outliers

Exclude outlier data points if they are errors or irrelevant to the analysis.

Unit	No. of rooms	Price
900	5	7000
1200	4	10,000
1500	8	12,000
876	4	6,000
~~1000~~	~~50~~	~~7,500~~
1200	7	12,000
~~1100~~	~~8~~	~~13~~

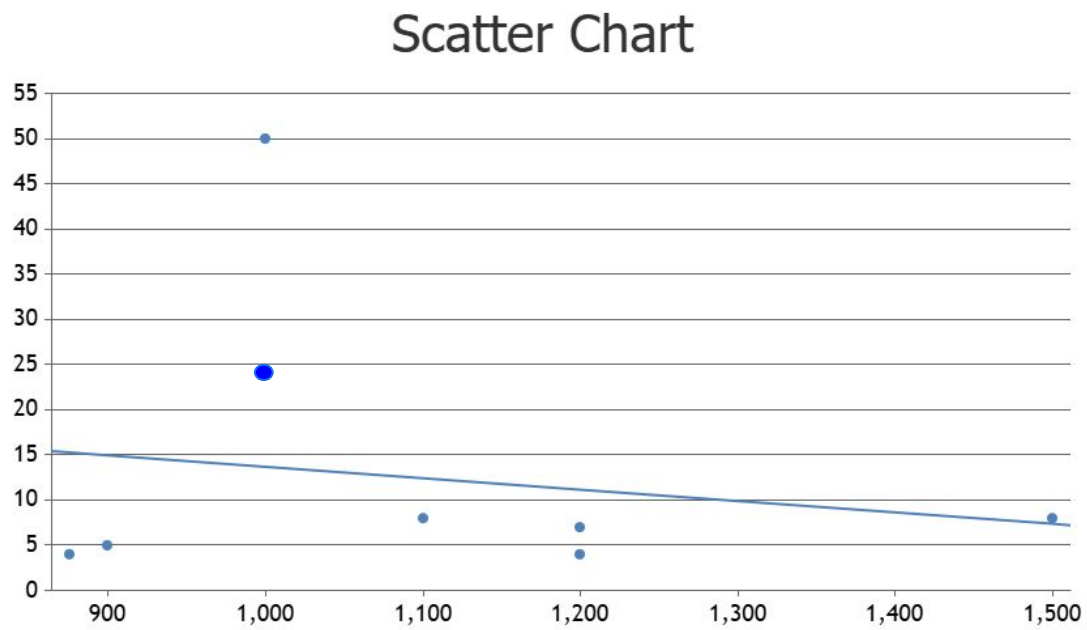
Removing Outliers

Sometimes outlier is so important in dataset that you can't exclude them



Removing Outliers

Winsorization: extreme values are replaced with the nearest value



Removing Outliers

Imputation

Unit	No. of rooms	Price
900	5	7000
1200	4	10,000
1500	8	12,000
876	4	6,000
1000	50	7,500
1200	7	12,000
1100	8	13

6

13,000

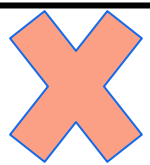
Feature Engineering

- Removing irrelevant features
- Creating new features

Feature Engineering

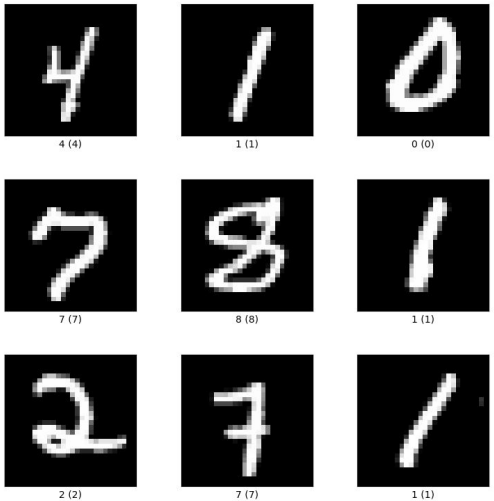
Removing irrelevant features

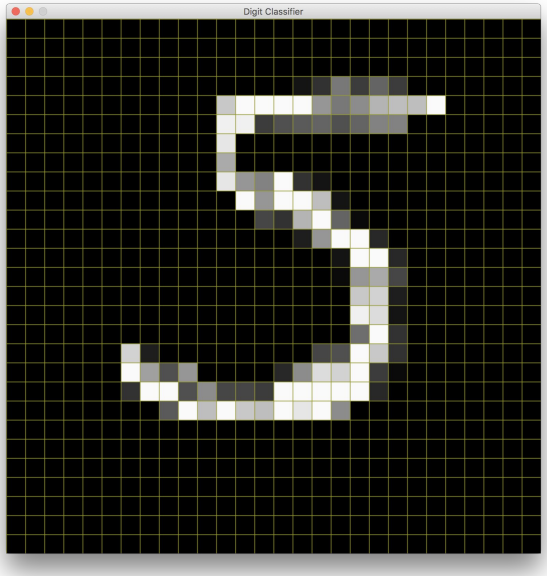
Unit	No. of rooms	Price	Colour	Lift_Availability



Feature Engineering

Removing irrelevant features





Feature Engineering

Creating New Features

Unit	No. of rooms	Price	Lift_Availability

Unit	No. of rooms	Price	Lift_Availability	**Per Unit Price**

Per Unit Price = Price / Unit

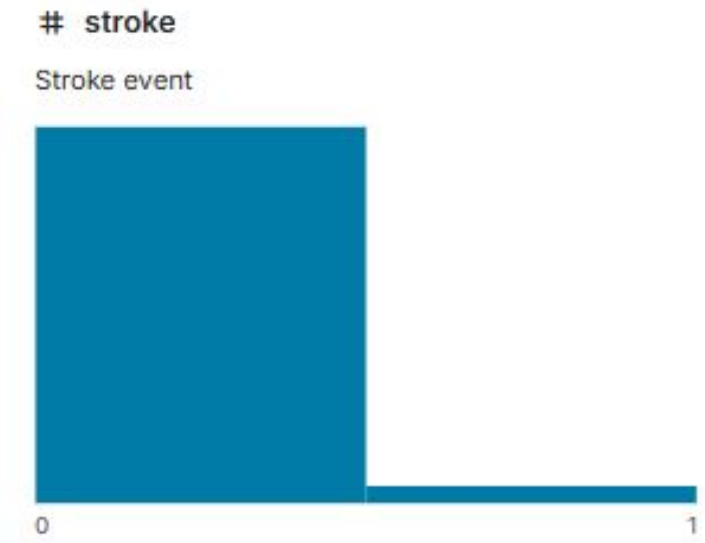
Data Cleaning

Removing Duplicate Data

Correcting inconsistent or invalid data entries.

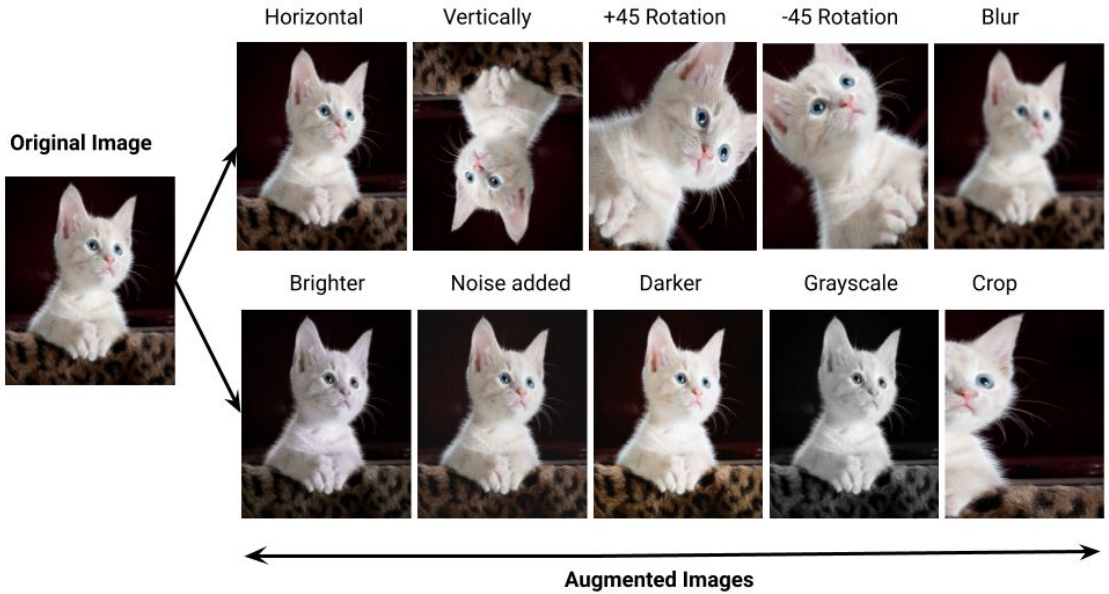
Data Augmentation

artificially increasing the size and diversity of a dataset



Data Augmentation

Image Augmentation



Data Augmentation

Text Augmentation

The movie was very interesting and enjoyable.

The movie was very fascinating and enjoyable.

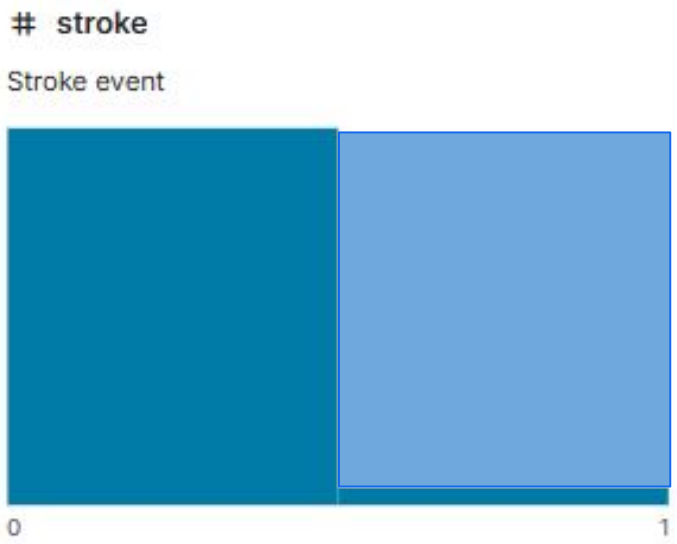
The movie was very fascinating and delightful.

The film was very fascinating and delightful.

Oversampling

used to address class imbalance in a dataset by increasing the number of instances in the minority class.

SMOTE (Synthetic Minority Oversampling Technique)



Undersampling

address class imbalance in a dataset by reducing the number of instances in the majority class.