codebasics / py

Code Issues 30 Pull requests 32 Actions Projects Wiki Sec

master

py / ML / FeatureEngineering / 3_outlier_IQR / 3_outliers_iqr.ipynb

dhavalsays added utsav resume

History

1 contributor

626 lines (626 sloc) 15.1 KB

Outlier Detection and Removal Using IQR

In [3]:
```
import pandas as pd
df = pd.read_csv("heights.csv")
df
```

Out[3]:

	name	height
0	mohan	1.2
1	maria	2.3
2	sakib	4.9
3	tao	5.1
4	virat	5.2
5	khusbu	5.4
6	dmitry	5.5
7	selena	5.5
8	john	5.6
9	imran	5.6
10	jose	5.8
11	deepika	5.9
12	yoseph	6.0
13	binod	6.1
14	gulshan	6.2
15	johnson	6.5
16	donald	7.1
17	aamir	14.5
18	ken	23.2
19	Liu	40.2

In [4]:
```
df.describe()
```

Out[4]:

	height
count	20.000000
mean	8.390000
std	8.782812
min	1.200000

25%	5.350000
50%	5.700000
75%	6.275000
max	40.200000

Detect outliers using IQR

```
In [5]: Q1 = df.height.quantile(0.25)
        Q3 = df.height.quantile(0.75)
        Q1, Q3

Out[5]: (5.3500000000000005, 6.275)
```

```
In [6]: IQR = Q3 - Q1
        IQR

Out[6]: 0.9249999999999998
```

```
In [7]: lower_limit = Q1 - 1.5*IQR
        upper_limit = Q3 + 1.5*IQR
        lower_limit, upper_limit

Out[7]: (3.962500000000001, 7.6625)
```

Here are the outliers

```
In [8]: df[(df.height<lower_limit)|(df.height>upper_limit)]

Out[8]:
```

	name	height
0	mohan	1.2
1	maria	2.3
17	aamir	14.5
18	ken	23.2
19	Liu	40.2

Remove outliers

```
In [9]: df_no_outlier = df[(df.height>lower_limit)&(df.height<upper_limit)]
        df_no_outlier
```

Out[9]:

	name	height
2	sakib	4.9
3	tao	5.1
4	virat	5.2
5	khusbu	5.4
6	dmitry	5.5
7	selena	5.5
8	john	5.6
9	imran	5.6
10	jose	5.8
11	deepika	5.9
12	yoseph	6.0
13	binod	6.1
14	gulshan	6.2
15	johnson	6.5
16	donald	7.1

Exercise

You are given height_weight.csv file which contains heights and weights of 1000 people. Dataset is taken from here, https://www.kaggle.com/mustafaali96/weight-height (https://www.kaggle.com/mustafaali96/weight-height)

You need to do this,

(1) Load this csv in pandas dataframe and first plot histograms for height and weight parameters