How to Find Outliers

Subject: Data mining

Fateme Hasanpour

Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests.

There are four ways to identify outliers:

1. Sorting method
2. Data visualization method
3. Statistical tests (*z* scores)
4. Interquartile range method

True outliers

True outliers should always be retained in your dataset because these just represent natural variations in your sample.

Example: True outlier
You measure 100-meter running times for a representative sample of 560 college students. Your data are normally distributed with a couple of outliers on either end. Most values are centered on the middle, as expected. But these extreme values also represent natural variations because a variable like running time is influenced by many other factors.

Outliers that don't represent true values can come from many possible sources:

- Measurement errors
- Data entry or processing errors
- Unrepresentative sampling

Example: Other outliers

You repeat your running time measurements for a new sample.
For one of the participants, you accidentally start the timer midway through their sprint. You record this timing as their running time.

This data point is a big outlier in your dataset because it's much lower than all of the other times.

Four ways of calculating outliers

Sorting method

You can sort quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.

This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

Example: Sorting method

Your dataset for a pilot experiment consists of 8 values.

180 156 9 176 163 1827 166 171

You sort the values from low to high and scan for extreme values.

9 156 163 166 171 176 180 1872

Using visualizations

You can use software to visualize your data with a box plot .This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for your data.

Statistical outlier detection

Statistical outlier detection involves applying statistical tests or procedures to identify extreme values. (z-score)

Using the interquartile range

The interquartile range (IQR) tells you the range of the middle half of your dataset. You can use the IQR to create “fences” around your data and then define outliers as any values that fall outside those fences.

Interquartile range method

1. Sort your data from low to high
2. Identify the first quartile (Q1), the median, and the third quartile (Q3).
3. Calculate your IQR = Q3 – Q1
4. Calculate your upper fence = Q3 + (1.5 * IQR)
5. Calculate your lower fence = Q1 – (1.5 * IQR)
6. Use your fences to highlight any outliers, all values that fall outside your fences.

Using Z score

We write a function that takes numeric data as an input argument. We find the mean and standard deviation of the all the data points.We find the z-score for each of the data point in the dataset and if the z-score is greater than 3 than we can classify that point as an outlier. Any point outside of 3 standard deviations would be an outlier.

In [4]:
```
import numpy as np
import pandas as pd
outliers=[]
def detect_outlier(data_1):

    threshold=3
    mean_1 = np.mean(data_1)
    std_1 =np.std(data_1)


    for y in data_1:
        z_score= (y - mean_1)/std_1
        if np.abs(z_score) > threshold:
            outliers.append(y)
    return outliers

dataset= [10,12,12,13,12,11,14,13,15,10,10,10,100,12,14,13,12,
          10,10,11,12,15,12,13,12,11,14,13,15,10,15,12,10,14,13,15,10]
outlier_datapoints = detect_outlier(dataset)
print(outlier_datapoints)
```

```
[100]
```

Using IQR

lower_bound is 6.5 and upper bound is 18.5, so anything outside of 6.5 and 18.5 is an outlier.

In [27]:
```
import numpy as np
import pandas as pd

dataset= [10,12,12,13,12,11,14,13,15,10,10,10,100,12,14,13,12,
          10,10,11,12,15,12,13,12,11,14,13,15,10,15,12,10,14,13,15,10]
sorted(dataset)
q1, q3= np.percentile(dataset,[25,75])
iqr = q3 - q1
lower_bound = q1 -(1.5 * iqr)
upper_bound = q3 +(1.5 * iqr)
print('lower_bound:',lower_bound)
print('upper_bound:',upper_bound)
```

```
lower_bound: 6.5
upper_bound: 18.5
```

In [28]:
```
max_num = dataset[0]
for n in dataset:
    max_num = n if n >= max_num else max_num
print("Manual Iteration: ", max_num,'\nindex of max number is: ',dataset.index(max_num))
```

```
Manual Iteration:  100
index of max number is:  12
```

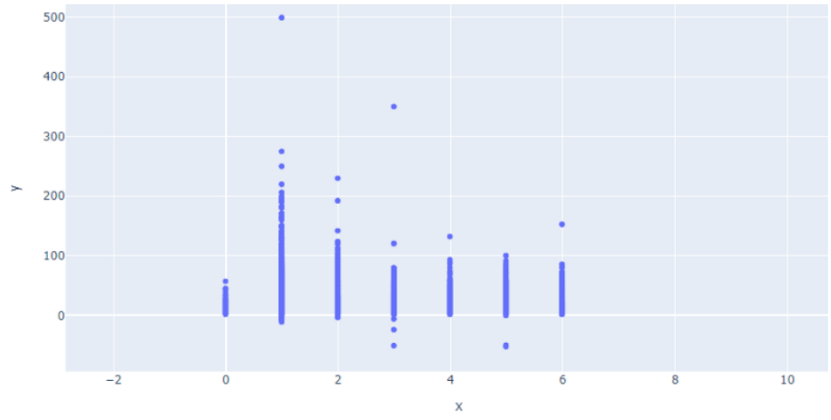
Using pandas describe()

	fare_amount	passenger_count
count	200000.000000	200000.000000
mean	11.359955	1.684535
std	9.901776	1.385997
min	-52.000000	0.000000
25%	6.000000	1.000000
50%	8.500000	1.000000
75%	12.500000	2.000000
max	499.000000	208.000000

As we can see, the fare_amount and passenger_count columns have outliers. For example, the max fare_amount is 499 while its mean is 11.36.

visualize outliers

- Histogram
- Box plot
- Scatter plot



Scatter plot

Drop the outliers

use .dropna()

References:

https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/#:~:text=Using%20pandas%20describe()%20to,not%20the%20dataset%20has%20outliers.

https://www.scribbr.com/statistics/outliers/

End.