A) What is an outlier?

- An outlier is a data point in a dataset that is distant from all other observations.
- A data point that lies outside the overall distribution of the dataset.
- An outlier is an extremely high or extremely low value in the dataset.

The reason for an outlier to exists in a dataset are:

- Variability in the data
- An experimental measurement error

The impact of an outlier:

Outliers makes it difficult to analyze data and build a predictive model/algorithm, therefore there is a need to identify and remove them.

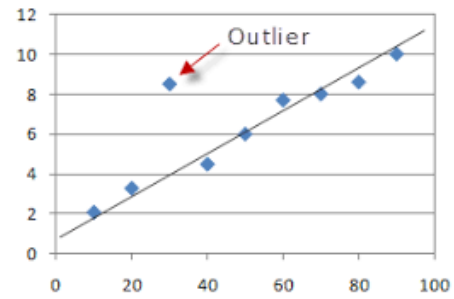
.

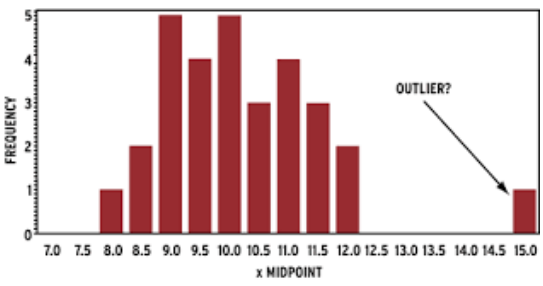
How to identify an outlier in any dataset?

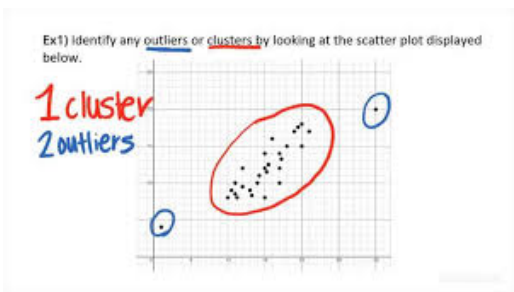
There is no clear criteria for determining whether or not a data point is an outlier. As a result, there are a variety of standard approaches that are used to identify outliers.

The most common used technique using data visualization.

Example of Outlier in graph are:







B) Data Visualization / (Using Plotting to find outliers)

- Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations.
- These visual displays of information communicate complex data relationships and data-driven insights in a way that is

C) Plotting in Pandas

- Pandas is a data analysis tool, but it also provides great options for data visualization.
- Pandas uses the plot() method to create graphs.

The Pandas plot() Method

- Pandas plot() make plots of Series or DataFrame.

Syntax:

> **df.plot()**

Parameters:

- **data** : Series or DataFrame
- **x and y parameters** specify the values that you want on the x and y column.
- **figsize** specifies the size of the figure object.
- **Title** to be used for the plot.
- **legend** to be placed on-axis subplots.
- **X and y label:** name to use for the label on the x-axis and y-axis.
- **Subplots:** make separate subplots for each column.
- **Kind:** the kind of plot to produce.

Different Plotting methods

In df.plot(), the kind parameter has the following options to implement plotting. These are:

1. Line Plot

> kind = 'line'

> By default, Line is used for plot()

2. Bar plot

> kind = 'bar' or king = 'barh'

> A bar plot is a plot that presents categorical data with rectangular bars. The lengths of the bars are proportional to the values that they represent.

> barh gives horizontal plotting.

3. Scatter plot or Point plot

> kind = 'scatter'

> A scatter plot is used to plot correlations between two variables. These correlations are plotted in the form of markers of varying colors and sizes.

4. 'hist' for histogram

> > kind = 'hist'

> A histogram is the most commonly used graph to show frequency distributions.

> A histogram is a graphical display of data with bars of different heights, where each bar groups numbers into ranges.

5. 'kde' or 'density' for density plots

> kind = 'density' or king = 'kde'

> Pandas can generate a **Kernel Density Estimate (KDE)** plot using Gaussian kernels. A kernel density estimate plot shows the distribution of a single variable and can be thought of as a smoothed histogram.

6. 'pie' for pie plots

> > kind = 'pie'

Let's create a DataFrame to learn to apply plot() in it.

```
import pandas as pd

df = pd.DataFrame({
    'name':['john','mary','peter','jeff','bill','lisa','jose'],
    'age':[23,78,22,19,45,33,20],
    'gender':['M','F','M','M','M','F','M'],
    'state':['california','dc','california','dc','california','texas','texas'],
    'num_children':[2,10,3,3,20,15,4],
    'num_pets':[5,1,0,5,2,2,3]
})

df
```

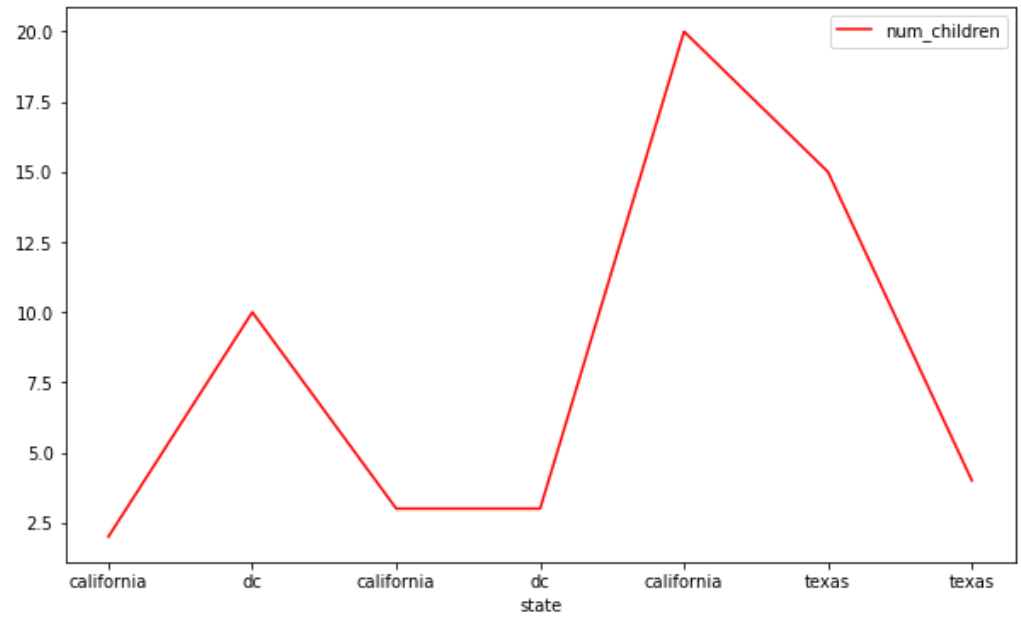
	name	age	gender	state	num_children	num_pets
0	john	23	M	california	2	5
1	mary	78	F	dc	10	1
2	peter	22	M	california	3	0
3	jeff	19	M	dc	3	5
4	bill	45	M	california	20	2

1) Line Plot

In df.plot(), we give kind = 'line', x axis column name, y axis column name, color of line, size of output image.

```
df.plot(kind='line', x='state', y='num_children', color='red', figsize=(10,6))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe5a23fbf90>
```



2) Bar Plot

In df.plot(), we give kind = 'bar', x axis column name, y axis column name, color of line, size of output image.

```
df.plot(kind='bar',x='name',y='num_children',color='green', figsize=(10,6))
```

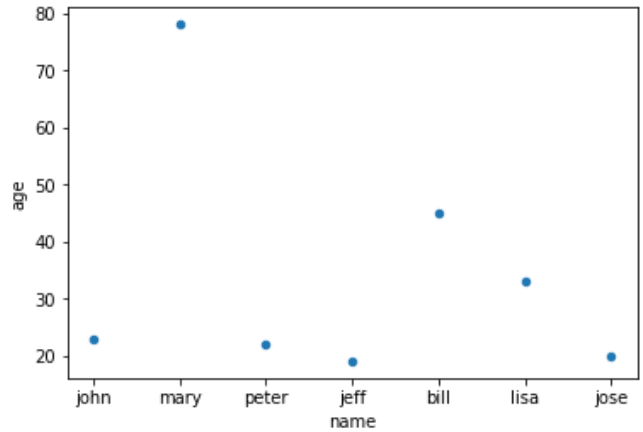
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe5af556dd0>
```

3)Scatter

In df.plot(), we give kind = 'scatter', x axis column name, y axis column name, color of line, size of output image.

```
df.plot(kind='scatter',x='name',y='age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe5a2261a90>
```

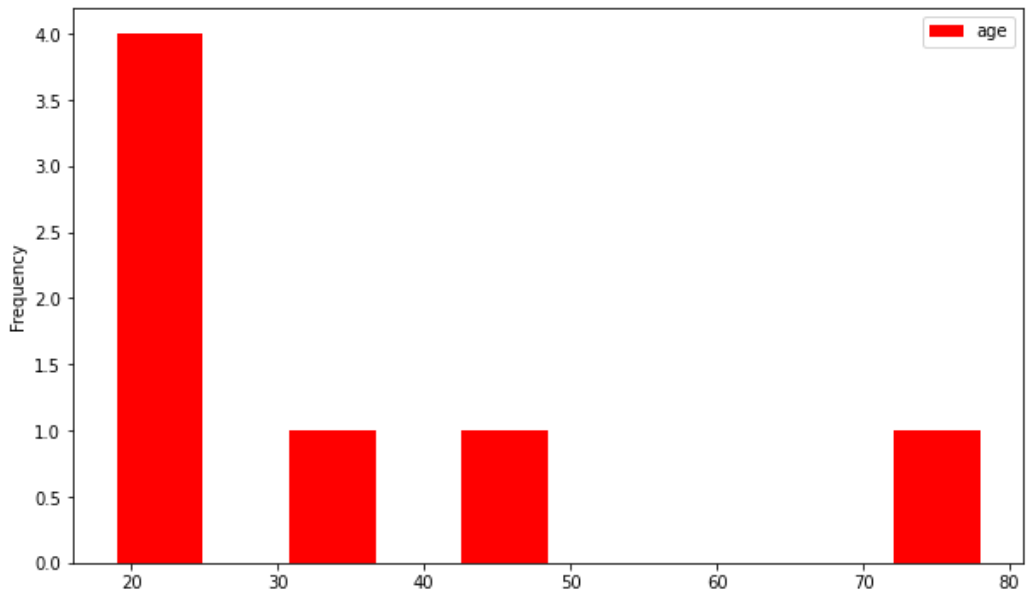


4) Histogram

In df.plot(), we give kind = 'hist', x axis column name, y axis column name, color of line, size of output image.

```
df.plot(kind='hist',y='age',x='num_children',color='red', figsize=(10,6))
```

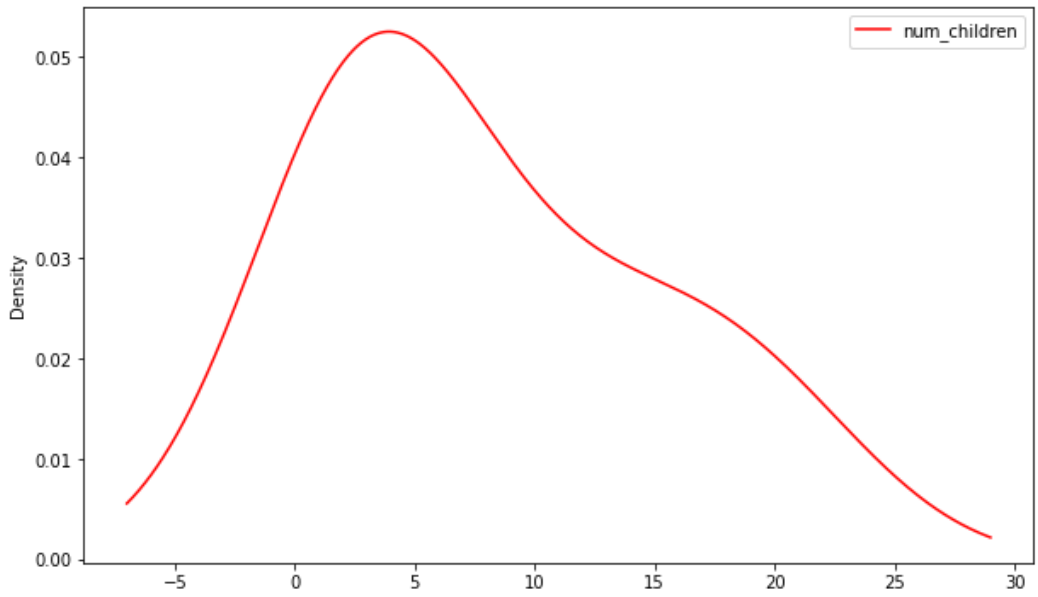
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe5ae06ebd0>
```



5) Density

```
df.plot(kind='kde',x='name',y='num_children',color='red', figsize=(10,6))
```

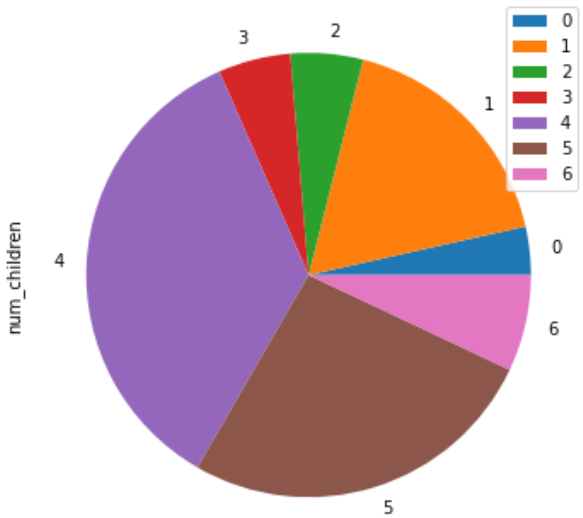
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe5adf08fd0>
```



6) Pie Plot

```
df.plot(kind='pie',x='name',y='num_children', figsize=(10,6))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe5a24ea690>
```



Additional Points

The Complete parameters of plot() are:

DataFrame.plot(x=None, y=None, kind='line', ax=None, subplots=False, sharex=None, sharey=False, layout=None, figsize=None, use_index=True, title=None, grid=None, legend=True, style=None, logx=False, logy=False, loglog=False, xticks=None, yticks=None, xlim=None, ylim=None, rot=None, fontsize=None, colormap=None, table=False, yerr=None, xerr=None, secondary_y=False, sort_columns=False, **kwds)

Parameter Description:

- x The columns argument mentions the set of columns to be considered as the x axis in the plotting process. The default value of the argument is None.
- y The columns argument mentions the set of columns to be considered as the y axis in the plotting process. The default value of the argument is None.
- kind The kind argument is used to mention the type of graph to be used as a process of this plotting process. ‘line’: line plot (This is the default plotting kind value). ‘bar’: vertical bar plot. ‘barh’: horizontal bar plot. ‘hist’: histogram. ‘box’: boxplot. ‘kde’: Kernel Density Estimation plot. ‘density’: same as ‘kde’. ‘area’: area plot. ‘pie’: pie plot. ‘scatter’: scatter plot. ‘hexbin’: hexbin plot.
- ax The Index argument mentions the axes object, the default value of this argument is none.
- subplots This is another boolean argument which is used to apply separate subplots for each and every column in place.
- sharex This is another boolean argument when the subplots=True, some among the x axis labels are set as invisible and x axis is shared, defaulting to True if ax is None otherwise False when an ax is passed in.
- sharey This is another boolean argument when the subplots=True, some among the y axis labels are set as invisible and y axis is shared, defaulting to True if ay is None otherwise False when an ay is passed in.
- layout This is a tuple type based on which the subplots are laid upon.
- figsize This mentions the size value in inches. use_index This uses the index as ticks for the x axis.
- title This is used for setting the title value for the graph. The value mentioned in the title will be in the top of the graph.
- grid A boolean argument with the default value as None, this is used to maintain the axis grid lines.
- legend Place the legend on the subplots of the axis.
- style The style argument holds values in list or dict. This argument is used to represent the style of line for each column.
- logx X axis level log scaling , this is a boolean argument and the default value is false.
- logy Y axis level log scaling, this is a boolean argument and the default value is false.
- loglog The loglog is used to maintain the log scaling in both x axis and y axis levels. This is a boolean argument and the default value is false.
- xticks The Xticks are associated with a value using this xticks argument. This argument takes input in the form of sequence.
- yticks The yticks are associated with a value using this yticks argument. This argument takes input in the form of sequence.
- xlim 2-tuple/list.
- ylim 2-tuple/list.
- rot The default value is None and this argument holds value in integer format. This mentions the rotation value for the ticks.
- fontsize The xticks and yticks are associated with a fontsize using this argument. The default value is None and this argument holds value in integer format.
- colormap The colormap is used for selecting colors from. This argument is ofstring type.
- colorbar This is of boolean type, when set to true it is helpfull in plotting the colorbar.
- position The bar and plot layout specific alignments. table When set to true it maps a table using the data associated to the dataframe. So the matplotlib’s default layout will be used to meet this data. If a Series or DataFrame is passed, use passed data to draw a table.
- yerr The error bars will be displayed for more detail, corresponds to y axis.
- xerr The error bars will be displayed for more detail. corresponds to x axis. stacked Bar plots.
- sort_columns The plot ordering is determined based on the column names used.

- secondary_y Determines which column to plot.
- mark_right On using the secondary_y axis automatically markings are placed on the column labels.
- kwds keywords.

Note:

A histogram represents the frequency distribution of continuous variables Conversely, a bar graph is a diagrammatic comparison of discrete variables.

Histogram presents numerical data whereas bar graph shows categorical data.