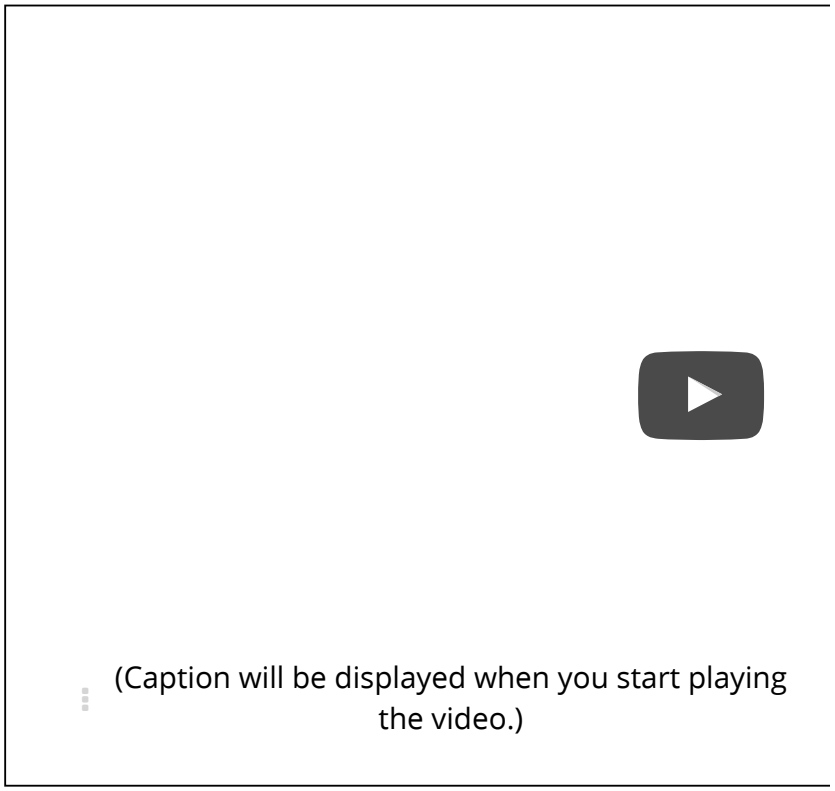
Video

Bookmark this page

2D Scatter Plots



Start of transcript. Skip to the end.

Similar to histograms, 2D scatter plots are also one of

the seven basic tools of quality.

So, again, these are recognized charting techniques, or

practices, that are useful for

troubleshooting your data issues.

The goal of a 2D scatter plot is either to look for

some groupings, or patterns

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Similar to histograms, scatter plots are also one of the Seven Basic Tools of Quality. Let's get them added to your arsenal, starting with the 2D variant.

2D scatter plots are used to visually inspect if a correlation exist between the charted features. Both axes of a 2D scatter plot represent a distinct, numeric feature. They don't have to be continuous, but they must at least be ordinal since each record in your dataset is being plotted as a point with its location along the axes corresponding to its feature values. Without ordering, the position of the plots would have no meaning.

It is possible that either a negative or positive correlation exist between the charted features, or alternatively, none at all. The correlation type can be assessed through the overall diagonal trending of the plotted points.

Positive and negative correlations may further display a linear or non-linear relationship. If a straight line can be drawn through your scatter plot and most of points seem to stick close to it, then it can be said with a certain level of confidence that there is a linear relationship between the plotted features. Similarly, if a curve can be drawn through the points, there is likely a non-linear relationship. If neither a curve nor line adequately seems to fit the overall shape of the plotted points, chances are there is neither a correlation nor relationship between the features, or at least not enough information at present to determine.

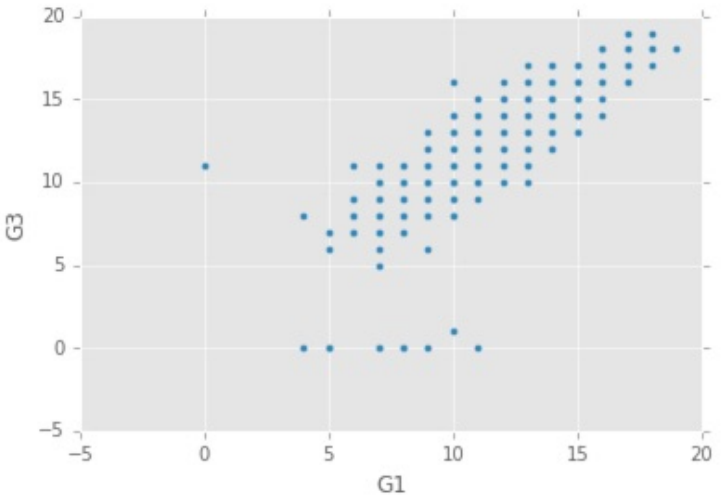
Begin with the code to pull up the students performance dataset, then simply call `.plot.scatter` on your dataset:

```
import pandas as pd
import matplotlib

matplotlib.style.use('ggplot') # Look Pretty
```

```
# If the above line throws an error, use plt.style.use('ggplot') instead
```

```
student_dataset = pd.read_csv("/Datasets/students.data", index_col=0)
student_dataset.plot.scatter(x='G1', y='G3')
```



This is your basic 2D scatter plot. Notice you have to call . `scatter` on a dataframe rather than a series, since two features are needed rather than just one. You also have to specify which features within the dataset you want graphed. You'll be using scatter plots so frequently in your data analysis you should also know how to create them directly from MatPlotLib, in addition to knowing how to graph them from Pandas. This is because many Pandas methods actually return regular NumPy *NDArrays*, rather than fully qualified Pandas *dataframes*.

This plot shows us that there certainly seems to be a positive correlation and linear relationship between a student's first and final exam scores—except for those few students at the bottom of the barrel! None of them did great on their first exam, and they all completely bombed their finals; if only there were a way to chart more than two variables simultaneously, perhaps you could verify if some other variable were at play in causing the students to fail...