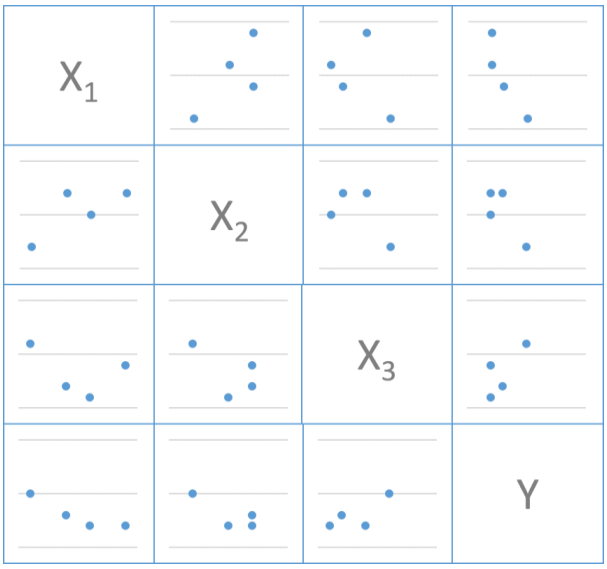
Data Science Essentials

Visualizing Data

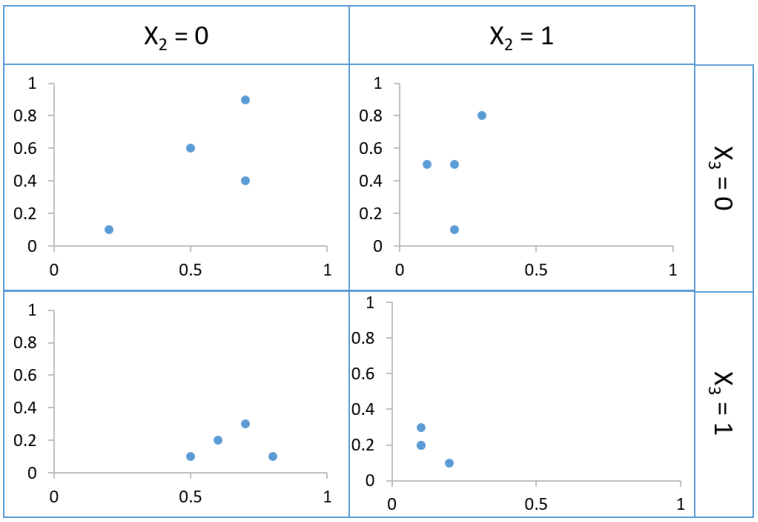
Data visualization is a highly useful way to explore data, and can help you determine apparent relationships between columns in order to identify candidates for predictive features in a machine learning model.

Common Visualizations

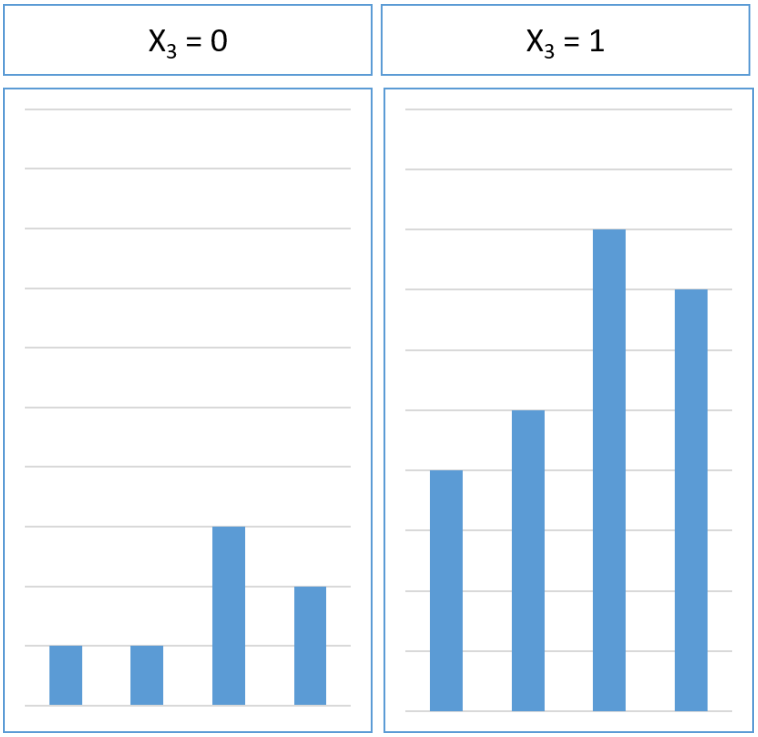
A scatter plot matrix, like the one below, shows scatter plots of selected columns in relation to each other, and is often a good starting point for data exploration. With a scatter plot matrix, you can easily spot variables that are collinear; which often indicates redundant features that should be removed from the model.



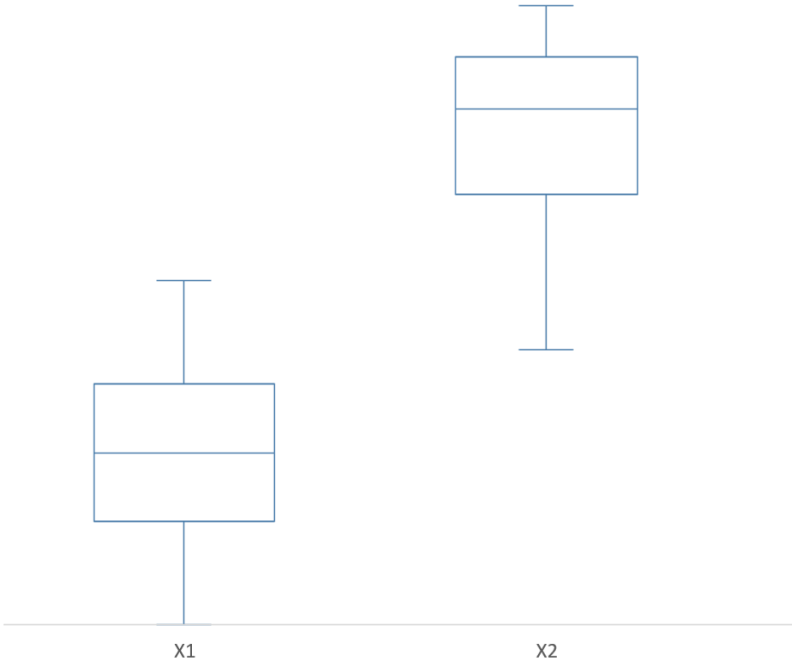
Scatter plots of individual columns can be useful for detailed exploration of the features in your dataset. A scatter plot enables you to see the intersection of values for two columns as plots in a chart. Additionally, you can condition the visualization on further columns; enabling you to visualize multiple dimensions of your data on a two dimensional chart. Scatter plots are particularly useful for spotting linear and non-linear relationships between variables.



Histograms that show the distribution of data density are useful for identifying potential outlier values. When conditioned, they show clearly how values for one variable may be distributed differently against specific values of another variable.



Box plots show the quartiles of numeric variables, with the median value indicated within a box that shows the first and third quartile. Additionally, lines known as *whiskers* can be extended from the box to show values outwith the values in the box. By comparing two values with box plots, you can easily see where the majority of the values for each variable lie, and to what extent the values in the two variables overlap.

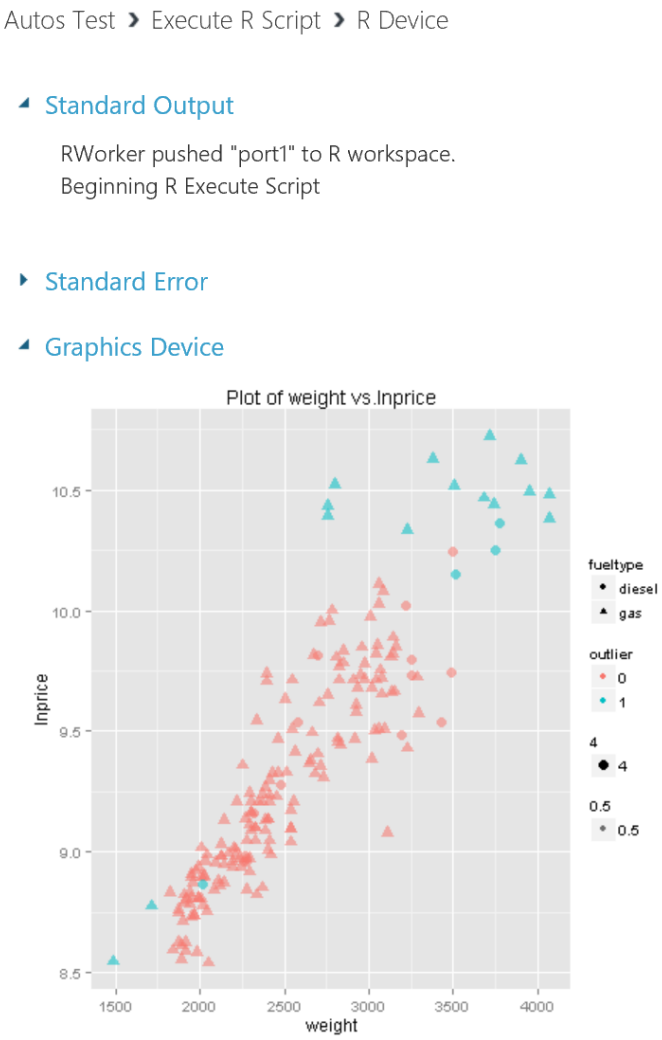


Visualizing Data with R or Python

You can generate visualizations using R or Python. See the following resources for more information about specific libraries and functions to create visualizations:

- R Visualization Resources
    - Documentation for ggplot2: http://docs.ggplot2.org/current/
    - Cheat sheet for ggplot2: https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
- Python Visualization Resources
    - Pandas plotting tutorial: http://pandas.pydata.org/pandas-docs/stable/visualization.html
    - Matplotlib tutorial: http://matplotlib.org/users/pyplot_tutorial.html

When working with Azure Machine Learning, you can include data visualization code in script modules, which can be useful if you need to visualize data that has been manipulated in the data flow of an Azure ML experiment. To view the visualizations, your code should print them to the device port, which you can view in the Azure ML Studio environment as shown here:



Alternatively, you can use Jupyter notebooks in Azure ML to visualize data. You will need to convert the output of the data flow to CSV formal, and then create a notebook based on the CSV data. Visualizations in Jupyter notebooks are displayed beneath the cell containing the code to create them, as shown here: