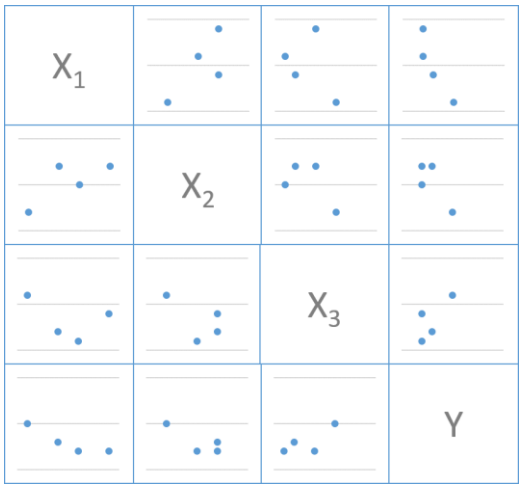
Data Science and Machine Learning Essentials

Module 3 Key Points

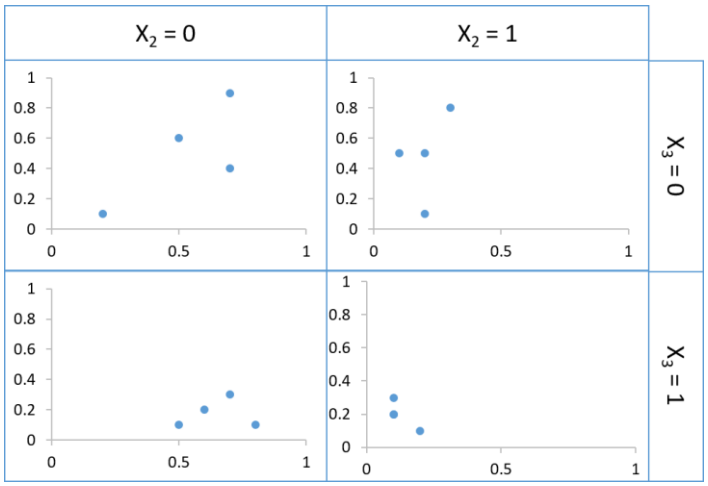
Chapter 13: Data Exploration and Visualization

Key Points

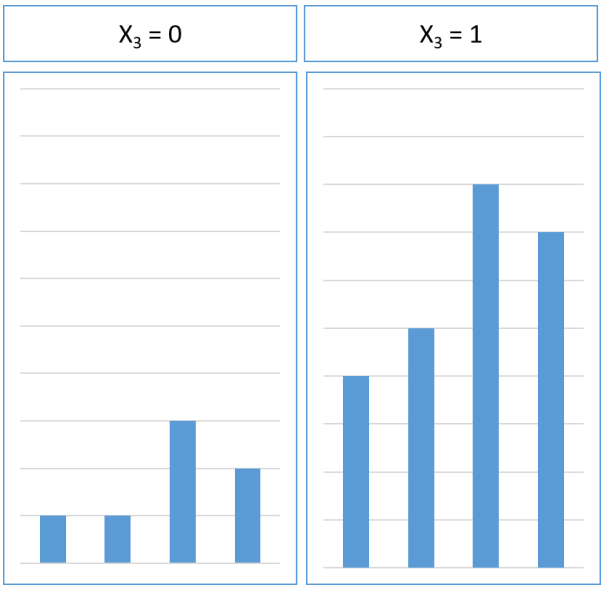
- Data visualization is a highly useful way to explore data, and can help you determine apparent relationships between columns in order to identify candidates for predictive features in a machine learning model.
- A scatter plot matrix, like the one below, shows scatter plots of selected columns in relation 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.



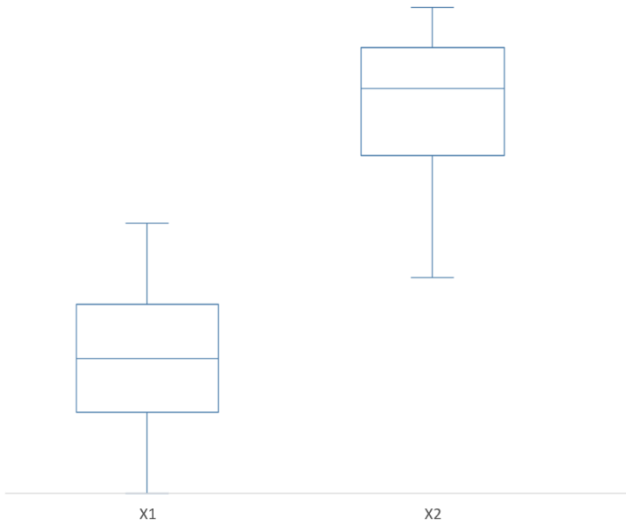
- Scatter plots of individual columns can be useful for detailed exploration of the features in your dataset. A scatter plot enables you to see the intersection of values for two columns as plots in a chart. Additionally, you can condition the visualization on further columns; enabling you to visualize 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.



Further Reading

- *Exploratory Data Analysis*, John Tukey, 1977 (Addison-Wesley)
- 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

Chapter 14: Building Models in Azure ML

Key Points

- Creating models is an iterative process that involves exploring data to select a set of candidate features, splitting the data to create a training set, selecting an appropriate machine learning model and training it, scoring and evaluating the model, and repeating the process until the best model for your predictive scenario and data has been identified.
- When existing columns don't align well with your label, try engineering polynomial columns by squaring or cubing them. This can often create a more useful distribution of values that align to the label better than the base feature values.
- When working on a classification or regression model, use the **Score Model** module to compare the predicted values (or scored *labels*) with the known actual label values in the portion of the dataset that was not used to train the model.
- In addition to the built-in models with the **Train Model** and **Score Model** modules, you create models using custom R or Python code.

Further Reading

- **Train Model** module: https://msdn.microsoft.com/en-us/library/azure/dn906044.aspx
- **Score Model** module: https://msdn.microsoft.com/en-us/library/azure/dn905995.aspx

Chapter 15: Model Evaluation, Comparison, and Selection

Key Points

- Models are evaluated based on statistics about the errors, or *residuals*, in the predicted values.
- You can use the **Evaluate Model** module to calculate commonly used error statistics for residuals in regression and classification models, making this an easy way to compare two scored models.
- In an ideal model, the residuals should be close to zero and conform to a normal probability distribution of values for the model type, which can be shown on a Q-Q plot (a chart that shows *quantiles* for the normal probability distribution against the residual values). For example, in a linear regression model, a Q-Q plot should show a straight line. You can use R or Python to create Q-Q plots and other visualizations of residuals to evaluate models.

Further Reading

- **Evaluate Model** module: https://msdn.microsoft.com/en-us/library/azure/dn905915.aspx