Exploratory Data Analysis and Model Selection

Presented by: Derek Kane

Overview of Topics

- Introduction to Exploratory Data Analysis
- Dataset Features
- Data Munging
- Descriptive Statistics
- Data Transformations
- Variable Selection Procedures
- Model Selection Procedures

Introduction to Exploratory Data Analysis

The "law of the instrument" developed by Abraham Kaplan in 1964 and Abraham Maslow's hammer in 1966 state: "If all you have is a hammer, everything looks like a nail."

Introduction to Exploratory Data Analysis

There is no panacea for advanced analytics; data scientists need to know what the correct tool is required for a particular task to be most effective and offer the proper solution.

Introduction to Exploratory Data Analysis

- Whenever we engage in a predictive modeling activity, we need to first understand the data for which we are working from. This is called Exploratory Data Analysis or EDA.
- The primary purpose for the EDA is to better understand the data we are using, how to transform the data, if necessary, and how to assess limitations and underlying assumptions inherent in the data structure.
- Data scientists need to know how the various pieces of data fit together and nuances in the underlying structures in order to decide what the best approach to the modeling task.
- Any type of method to look at data that does not include formal statistical modeling and inference generally falls under the EDA.

Exploratory Data Analysis

Here are some of the main reasons why we utilize EDA:

- Detection of mistakes.
- Checking of assumptions.
- Preliminary selection of appropriate models and tools.
- Determining relationships of the explanatory variables (independent).
- Detecting the direction and size of relationships between variables.

Features of a Data Set

Variables

Observations

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
$42,000.00	5850	3	1	2	yes	no	yes
$38,500.00	4000	2	1	1	yes	no	no
$49,500.00	3060	3	1	1	yes	no	no
$60,500.00	6650	3	1	2	yes	yes	no
$61,000.00	6360	2	1	1	yes	no	no
$66,000.00	4160	3	1	1	yes	yes	yes
$66,000.00	3880	3	2	2	yes	no	yes
$69,000.00	4160	3	1	3	yes	no	no
$83,800.00	4800	3	1	1	yes	yes	yes
$88,500.00	5500	3	2	4	yes	yes	no
$90,000.00	7200	3	2	1	yes	no	yes
$30,500.00	3000	2	1	1	no	no	no
$27,000.00	1700	3	1	2	yes	no	no

Quantitative Variables

- Numeric

Qualitative Variables

- Categorical

Features of a Data Set

Dependent Variable

Independent Variables

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
$42,000.00	5850	3	1	2	yes	no	yes
$38,500.00	4000	2	1	1	yes	no	no
$49,500.00	3060	3	1	1	yes	no	no
$60,500.00	6650	3	1	2	yes	yes	no
$61,000.00	6360	2	1	1	yes	no	no
$66,000.00	4160	3	1	1	yes	yes	yes
$66,000.00	3880	3	2	2	yes	no	yes
$69,000.00	4160	3	1	3	yes	no	no
$83,800.00	4800	3	1	1	yes	yes	yes
$88,500.00	5500	3	2	4	yes	yes	no
$90,000.00	7200	3	2	1	yes	no	yes
$30,500.00	3000	2	1	1	no	no	no
$27,000.00	1700	3	1	2	yes	no	no

The dependent variable is the variable that we are interested in predicting and the independent variables are the variables which may or may not help to predict the dependent variable.

Data Munging

- Data Munging is the transformation of raw data to a useable format.
- Many datasets are not readily available for analysis.
- Data needs to be transformed or cleaned first.
- This process is often the most difficult and the most time consuming.

Data Munging Tasks

Data Munging tasks include:

- Renaming Variables
- Data Type Conversion
- Encoding, Decoding, recoding data.
- Merging Datasets
- Transforming Data
- Handling Missing Data (Imputation)
- Handling Anomalous values

- These data munging tasks are an iterate process and can occur at any stage throughout the overall EDA procedure.

Understanding the Data

Non-Numeric

Missing Values

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
$42,000.00	5850	3	1	2	yes	no	yes
$38,500.00	4000	2	1	1	yes	no	no
$49,500.00	3060	3	1	1	yes	no	no
$60,500.00	6650	3	1	2	yes	yes	no
	6360	2	1	1	yes	no	no
	4160	3	1	1	yes	yes	yes
$66,000.00	3880	3	2	2	yes	no	yes
$69,000.00	4160	3	1	3	yes	no	no
$83,800.00	4800	3	1	1	yes	yes	yes
$88,500.00	5500	3	2	4	yes	yes	no
$90,000.00	7200	3	2	1	yes	no	yes
$30,500.00	3000	2	1	1	no	no	no
$27,000.00	1700	15	1	2	yes	7	no

Outlier

Error

Observation: The first row should contain variable names and all of the data should be completely filled after the data munging process is complete.

Data Munging Tasks

Renaming Variables

- The names of variables should make intuitive sense to non-practitioners and does not have to conform to IT protocols and standards.

T1K5X
$42,000.00
$38,500.00
$49,500.00
$60,500.00

Price
$42,000.00
$38,500.00
$49,500.00
$60,500.00

Data Type Conversion

- Depending upon the modeling task at hand and the software, the data may need to be expressed in a specific format in order to process correctly.

Date
January 1st, 2014

Date
1/1/2014

SQL: Text String
Varchar (max)

SQL: Date Value
Datetime

Data Munging Tasks

Encoding Data

- There are times when we need to change the underlying contents in a variable to prepare them for analytics. Ex. Qualitative to Quantitative.

driveway
yes
no
yes

driveway
1
0
1

- If we are using categorical variables, we need to clean them to get rid of non response categories like "I don't know", "no answer", "n/a", etc… We also need to order the encoding of categories (potentially reverse valence) to ensure that models are built and interpreted correctly.

Response
Strongly Agree
Strongly Disagree
Agree
Disagree
No Preference

Response
Strongly Agree
Agree
No Preference
Disagree
Strongly Disagree

Response
Strongly Agree
Agree
Disagree
Strongly Disagree

Response
4
3
2
1

- Usually non response categories is coded with values like 999. If this was a value in the variable "Age", this will skew the results and should be turned to NULL and reviewed further.

Data Munging Tasks

Merging Datasets

- It is quite rare that you will have a dataset readily constructed for analysis. This may require some data manipulation and merging in order to get the data in the correct form.

ID	price	lotsize
A1234	$42,000.00	5850
A1235	$38,500.00	4000
A1236	$49,500.00	3060
A1237	$60,500.00	6650
A1238	$61,000.00	6360
A1239	$66,000.00	4160
A1240	$66,000.00	3880
A1241	$69,000.00	4160
A1242	$83,800.00	4800
A1243	$88,500.00	5500
A1244	$90,000.00	7200
A1245	$30,500.00	3000
A1246	$27,000.00	1700

ID	bedrooms	bathrms	stories	garagepl
A1234	3	1	2	1
A1235	2	1	1	0
A1236	3	1	1	0
A1237	3	1	2	0
A1238	2	1	1	0
A1239	3	1	1	0
A1240	3	2	2	2
A1241	3	1	3	0
A1242	3	1	1	0
A1243	3	2	4	1
A1244	3	2	1	3
A1245	2	1	1	0
A1246	3	1	2	0

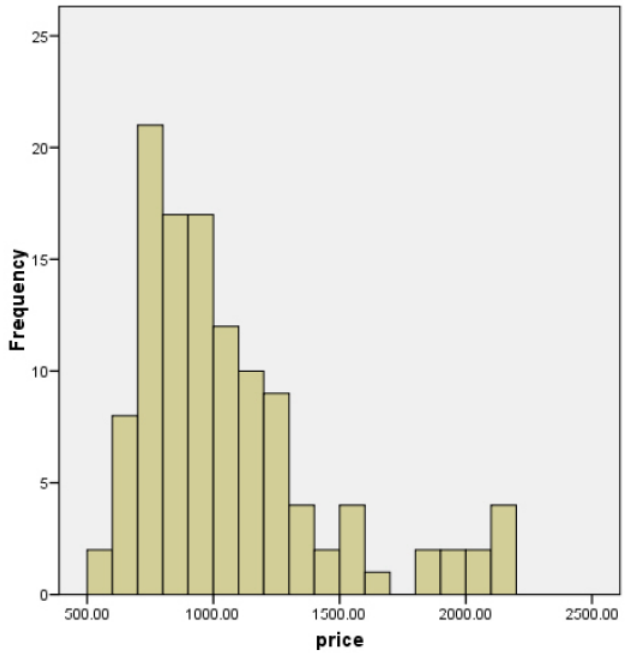
ID	price	lotsize	bedrooms	bathrms	stories	garagepl
A1234	$42,000.00	5850	3	1	2	1
A1235	$38,500.00	4000	2	1	1	0
A1236	$49,500.00	3060	3	1	1	0
A1237	$60,500.00	6650	3	1	2	0
A1238	$61,000.00	6360	2	1	1	0
A1239	$66,000.00	4160	3	1	1	0
A1240	$66,000.00	3880	3	2	2	2
A1241	$69,000.00	4160	3	1	3	0
A1242	$83,800.00	4800	3	1	1	0
A1243	$88,500.00	5500	3	2	4	1
A1244	$90,000.00	7200	3	2	1	3
A1245	$30,500.00	3000	2	1	1	0
A1246	$27,000.00	1700	3	1	2	0

Observation: The datasets will need to have a common ID as the link to join the data. After the data has been merged, the ID may not be necessary to retain for model building.

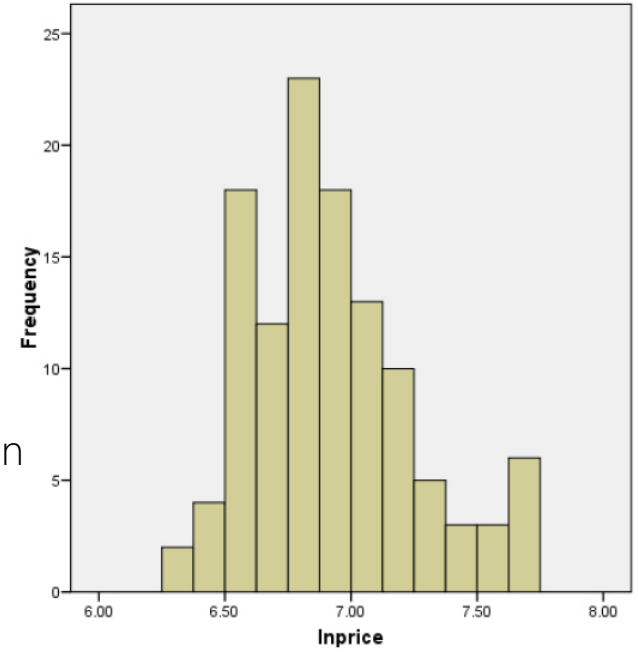
Data Munging Tasks

Transforming Variables

- There may be times where a variable will need to be transformed in order to achieve linearity. This will aid in strengthening the results of parametric based methods.



Natural Log Transformation



Data Munging Tasks

Imputation

- If there are missing values in a column, these cannot be left unattended. We must decided if we want to:
  - Remove the observation from the dataset
  - Calculate a value for the null (impute). This usually is determined with the mean or median, however, a more advanced version can use a multiple linear regression formula.

price
$ 42,000.00
$ 38,500.00
$ 49,500.00
$ 60,500.00
$ 66,000.00
$ 69,000.00
$ 83,800.00
$ 88,500.00
$ 90,000.00
$ 30,500.00
$ 27,000.00

price
$42,000.00
$38,500.00
$49,500.00
$60,500.00
$58,660.00
$58,660.00
$66,000.00
$69,000.00
$83,800.00
$88,500.00
$90,000.00
$30,500.00
$27,000.00

Mean = 58,660

Median = 60,500

Data Munging Tasks

Handling Anomalous Values

- Depending upon the analytic task, we need to assess points which exhibit a great deal of influence on the model.
- Outliers are data points that deviate significantly from the spread or distribution of other similar data points. These can typically be detected through the use of scatterplots.
- Many times we will delete the entry with an outlier to achieve normality in the dataset.
- In some instances, an outlier can be imputed but this must be approached with caution.

- Important: The drivers of outlying data points need to first be understood prior to devising an approach to dealing with them. They can hold the clues to new insights.

Descriptive Statistics

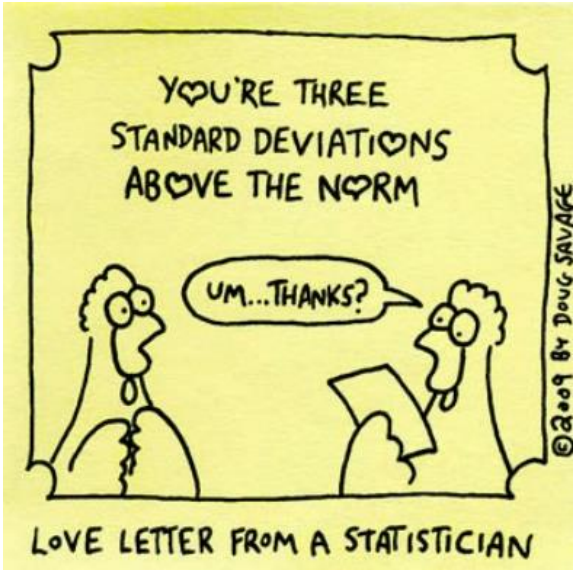
"Data don't make any sense,
we will have to resort to statistics."

- Describes the data in a qualitative or quantitative manner.
- Provides a summary of the shape of the data.
- These statistics help us to understand when transformations, imputations, and removal of outliers are necessary prior to model building.

Descriptive Statistics

- Descriptive Statistics are a collection of measurements of two things: Location and variability.
- *Location* tells you of the central value of your variable.
- *Variability* refers to the spread of the data from the center value.
- Statistics is essentially the study of what causes variability in the data.

Location	Variability
Mean	Variance
Median	Standard Deviation
Mode	Range



Descriptive Statistics - Location

Mean

- The sum of the observations divided by the total number of observations. It is the most common indicator of central tendency of a variable.

$$\overline{X} = \frac{\sum X_i}{n}$$

Median

- To get the median, we need to sort the data from lowest to highest. The median is the number in the middle of the data

2 2 5 6 7 8 9

Mode

- Refers to the most frequent of commonly occurring number within the variable.

2 2 5 6 7 8 9

Descriptive Statistics - Variability

Variance

- Measures the dispersion of the data from the mean.

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{(n-1)}$$

Standard Deviation

- The Standard Deviation is the Squared Root of the Variance. This indicates how close the data is to the mean.

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{(n-1)}}$$

Range

- Range is a measure of dispersion. It is the simple difference between the largest and smallest values.

2 2 5 6 7 8 9

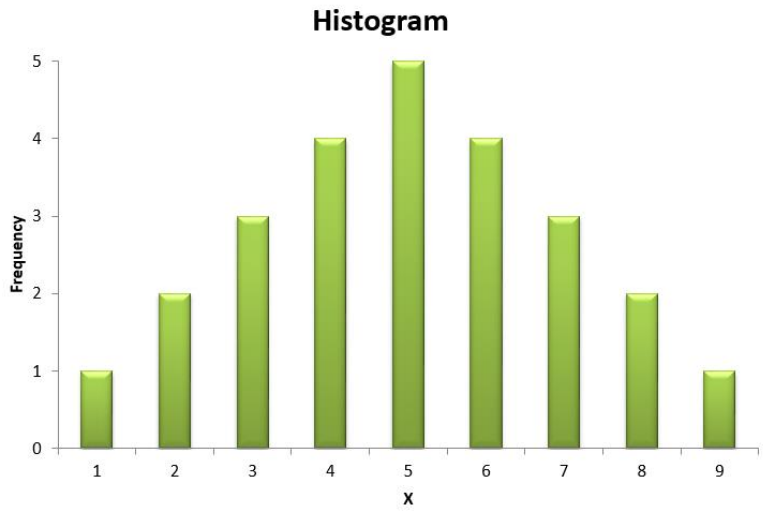
2 to 9

Histograms

- Histograms are a graphical display of data using bars of different heights. This allows us to evaluate the shape of the underlying distribution. Essentially, a histogram is a bar chart that groups numbers into ranges or bins.

X
1
2
2
3
3
3
4
etc…

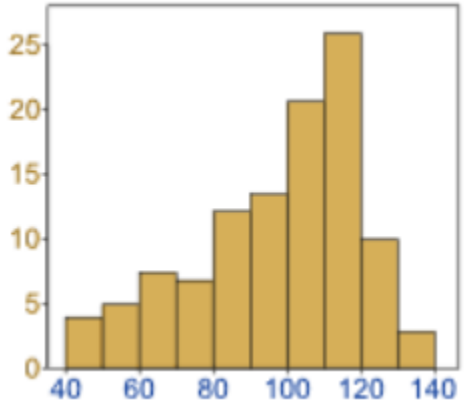
X	Frequency
1	1
2	2
3	3
4	4
5	5
6	4
7	3
8	2
9	1



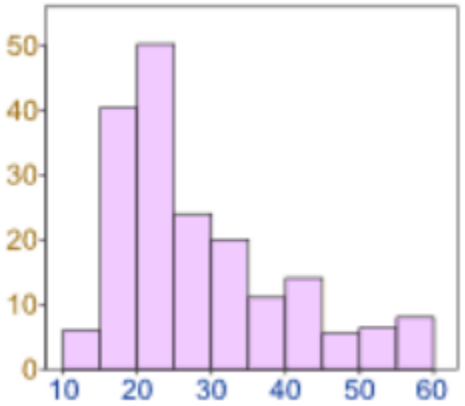
Histograms

- Data can be distributed (spread out) in many different ways.

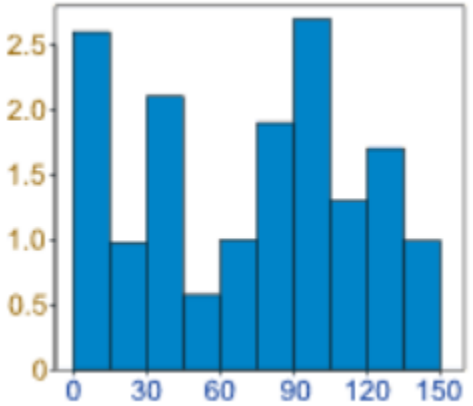
It can be spread out more on the left



Or more on the right

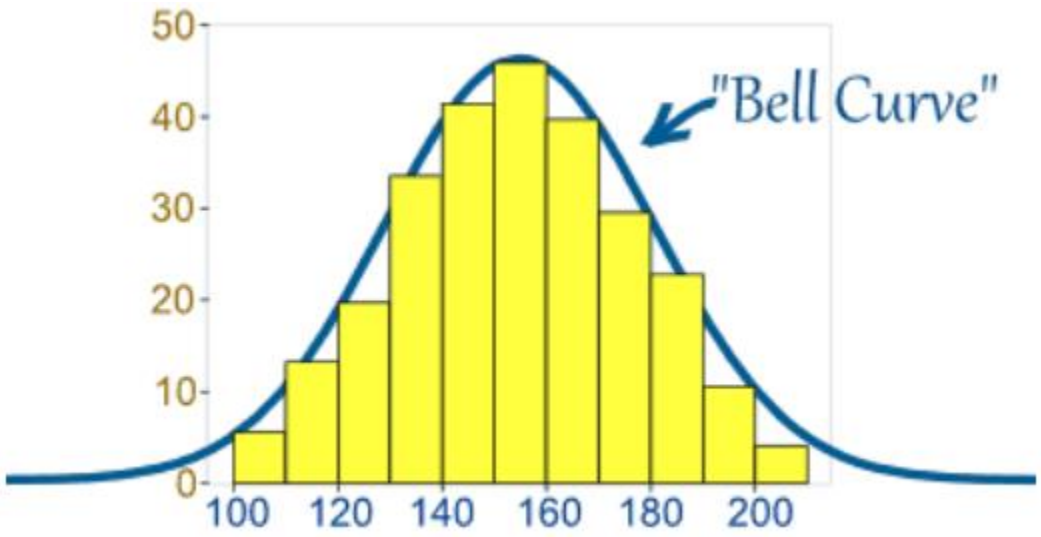


Or all jumbled up.



Normal Distribution

- But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



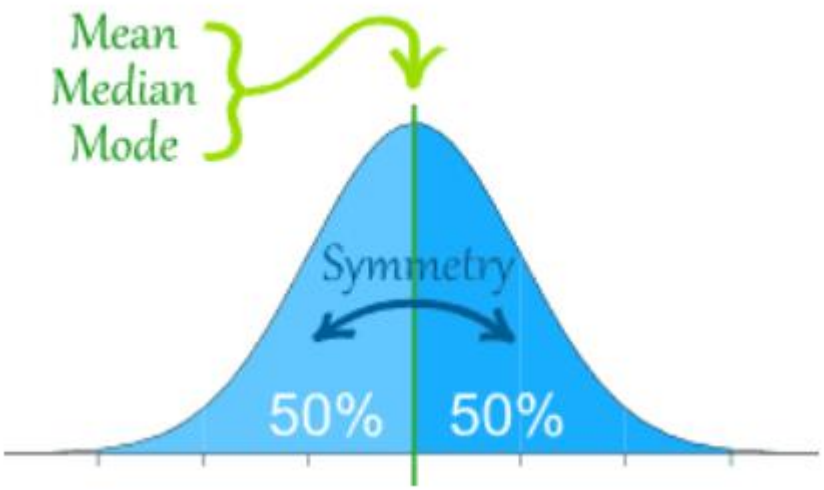
The "Bell Curve" is a Normal Distribution. The yellow histogram shows some data that follows it closely, but not perfectly (which is usual).

Normal Distribution

Many things follow a normal distribution:

- Height of People
- Size of things produced by machines
- Errors in measurements
- Blood Pressure
- Academic Test Scores
- Qunicunx (Plinko)

Normal Distribution

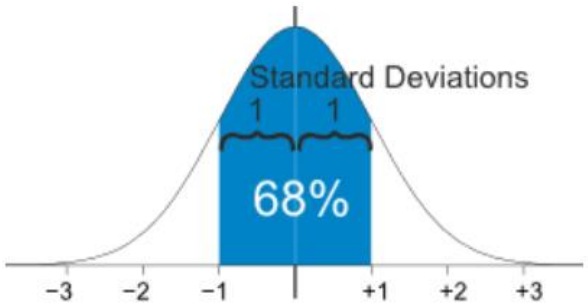


We say that data is normally distributed when the histogram has:

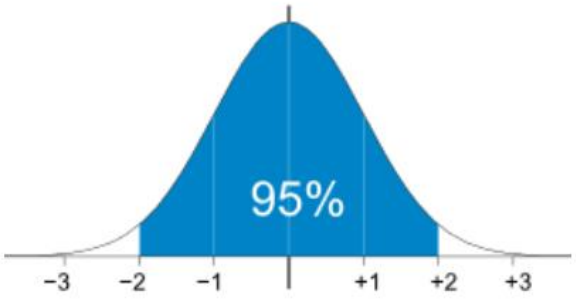
- Mean = Median = Mode
- Symmetry around the center
- 50% of the values less than the mean and 50% greater than the mean.

Normal Distribution

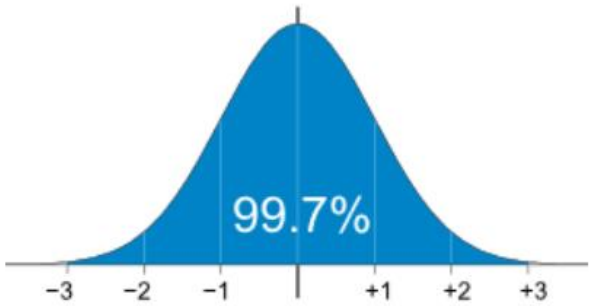
When we calculate the standard deviation of the mean (σ) we find that:



68% of values are within 1 deviation of the mean



95% of values are within 2 deviations of the mean



99.7% of values are within 2 deviations of the mean

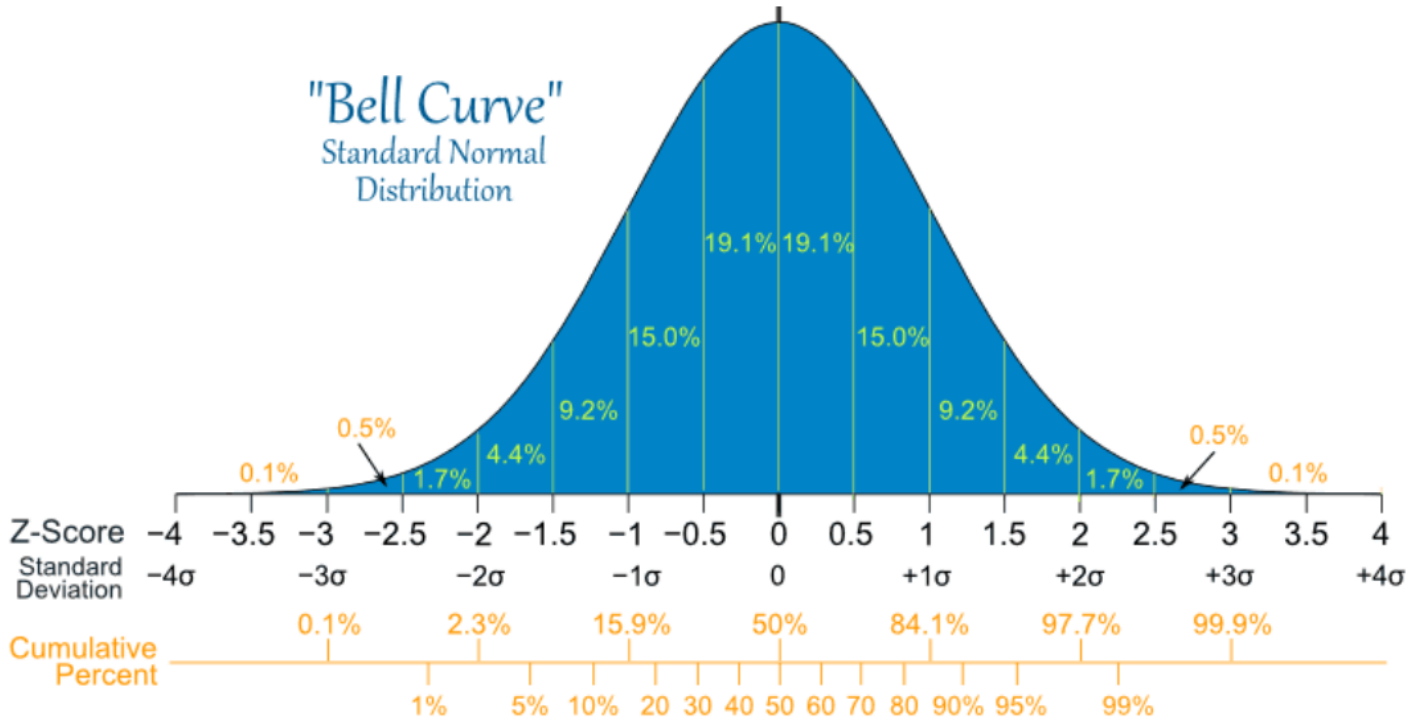
Normal Distribution

$$z = \frac{x - \mu}{\sigma}$$

- The number of standard deviations from the mean is also called the "standard score" or "z-score".
- We can take any normal distribution and convert to the standard normal distribution.
- To convert a value to the Z-score:
  - First subtract the mean.
  - Then divide by the standard deviation.
- Standardizing helps us make decisions about the data and makes our life easier by only having to refer to a single table for statistical testing.

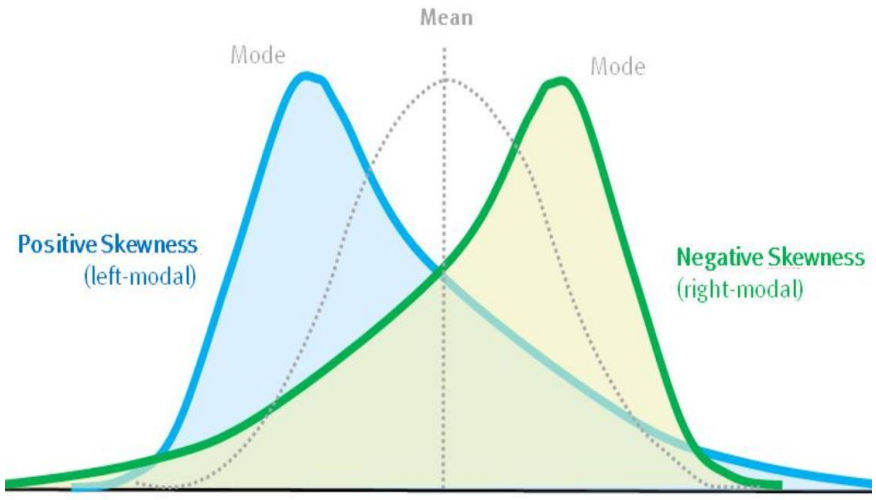
Normal Distribution

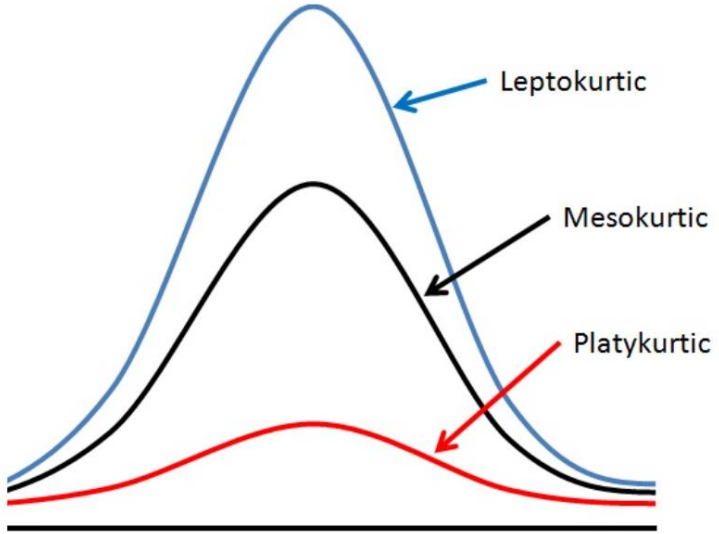
- Here is the Standard Normal Distribution with percentages for every half of a standard deviation, and cumulative percentages:



Normal Distribution

- There are other moments about the mean, skewness and kurtosis, that can used to understand what distribution should be used when there are 3 or more parameters to estimate.

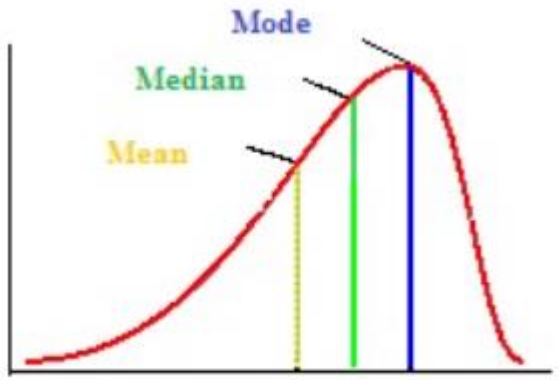




Normal Distribution

- The Skewness statistic refers to the lopsidedness of the distribution. If a distribution has a negative Skewness (sometimes described as left skewed) it has a longer tail to the left than to the right. A positively skewed distribution (right skewed) has a longer tail to the right, and zero skewed distributions are usually symmetric.

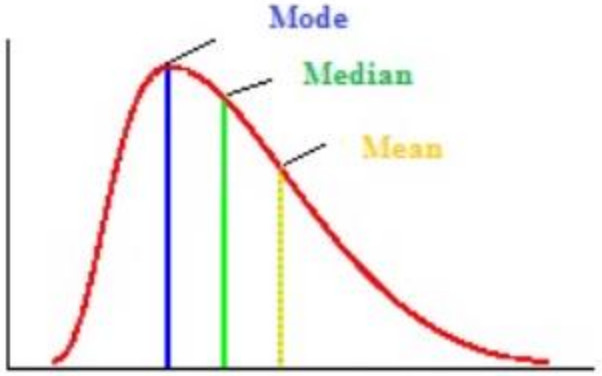
$$Skewness = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{SD(x)}\right)^3$$



Left-Skewed (Negative Skewness)

Interpretation:

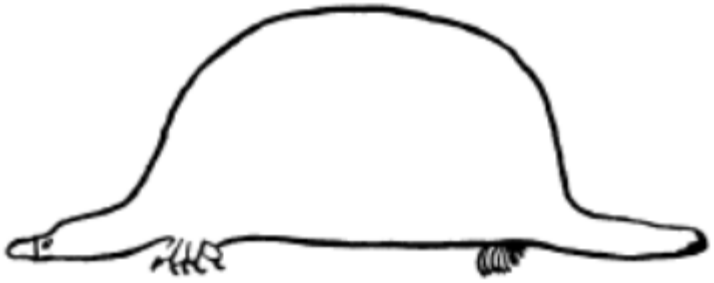
- Skewness > 0 - Right skewed distribution - most values are concentrated on left of the mean, with extreme values to the right.
- Skewness < 0 - Left skewed distribution - most values are concentrated on the right of the mean, with extreme values to the left.
- Skewness = 0 - mean = median, the distribution is symmetrical around the mean.



Right-Skewed (Positive Skewness)

Normal Distribution

Leptokurtic



Platykurtic

- Kurtosis is an indicator used in distribution analysis as a sign of flattening or "peakedness" of a distribution.
- The Kurtosis is calculated from the following formula:

$$Kurtosis = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{SD(x)}\right)^4$$

Interpretation:

- Kurtosis > 3 - Leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and longer tails. This means high probability for extreme values.
- Kurtosis < 3 - Platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution, and the values are wider spread around the mean.
- Kurtosis = 3 - Mesokurtic distribution - normal distribution for example.

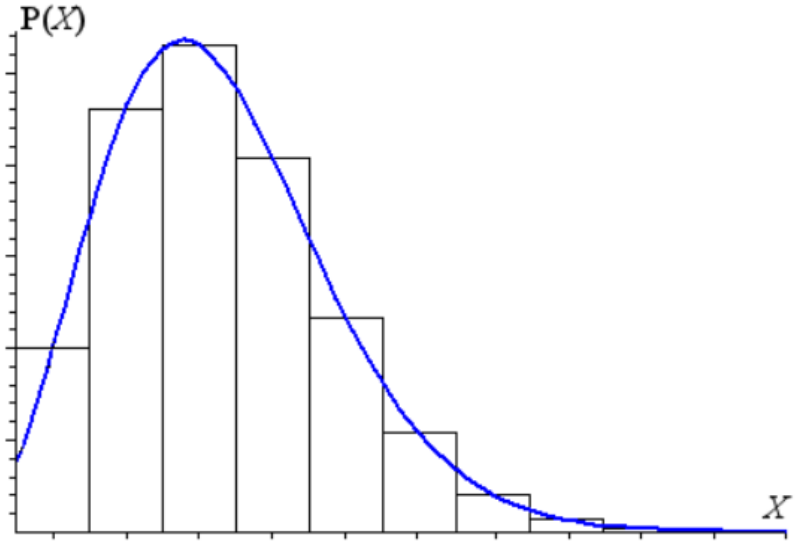
Normal Distribution

- The following table gives some examples of skewness and kurtosis for common distributions.

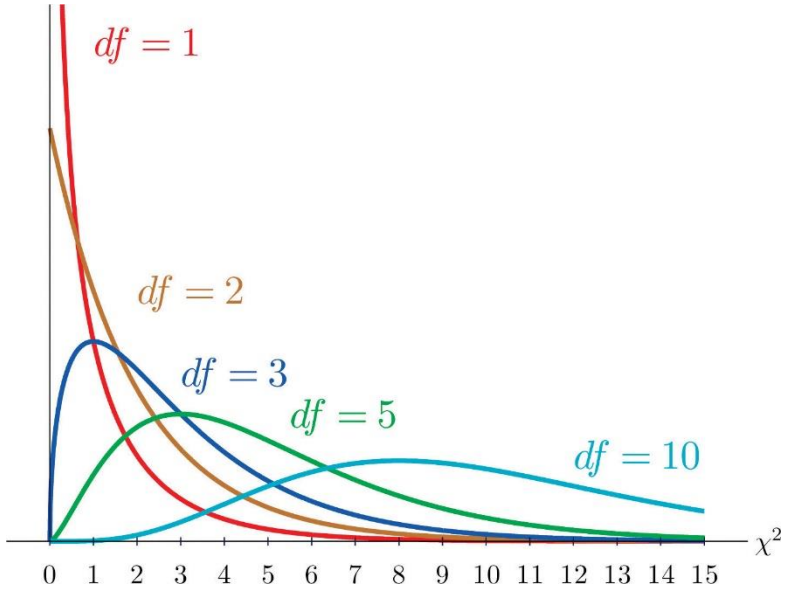
Distribution	*Skewness*	*Kurtosis*
Binomial	-∞ to +∞	1 to +∞
Chisq	0 to 2.828	3 to 15
Exponential	2	9
Lognormal	0 to +∞	3 to +∞
Normal	0	3
Poisson	0 to +∞	3 to +∞
Triangle	-0.562 to 0.562	2.388
Uniform	0	1.8

Other Distributions

Here are some examples of other common distributions you may encounter:



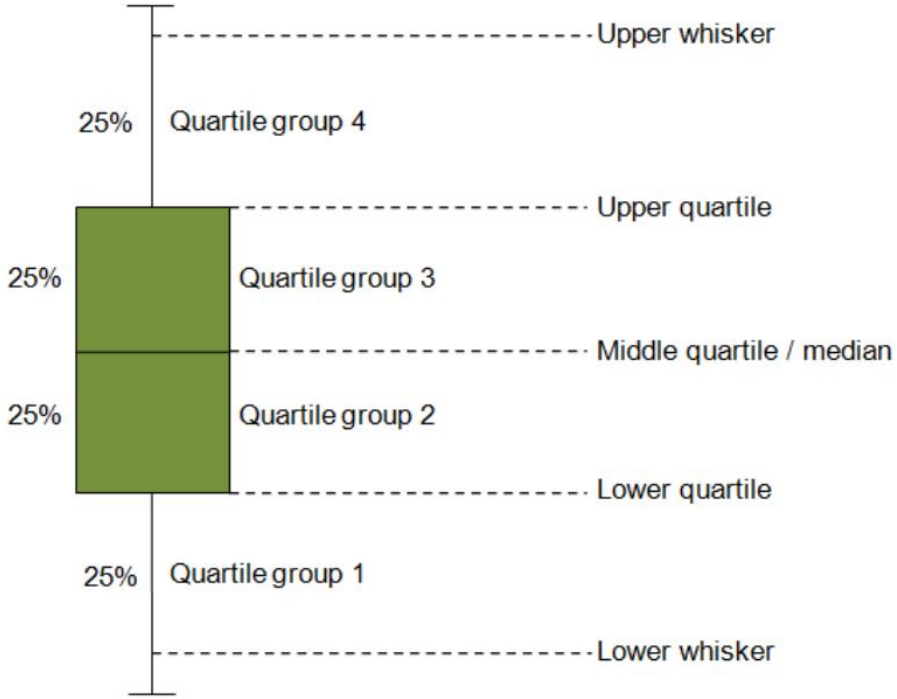
Poisson Distribution



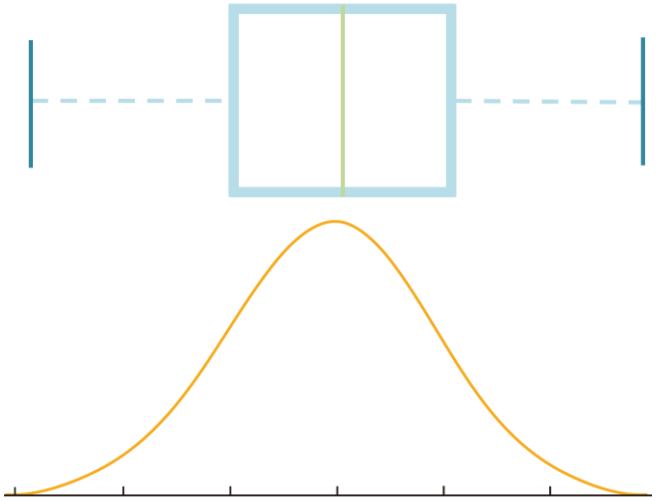
Chi-Square Distributions

Box Plots

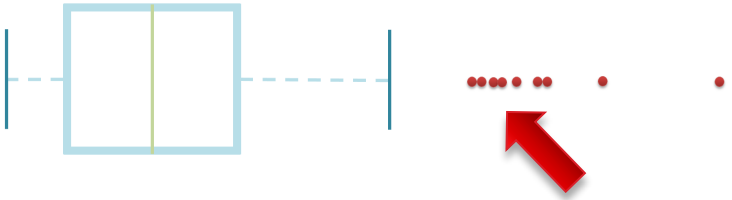
- A Boxplot is a nice way to graphically represent the data in order to communicate the data through their quartiles.



Normal Distribution



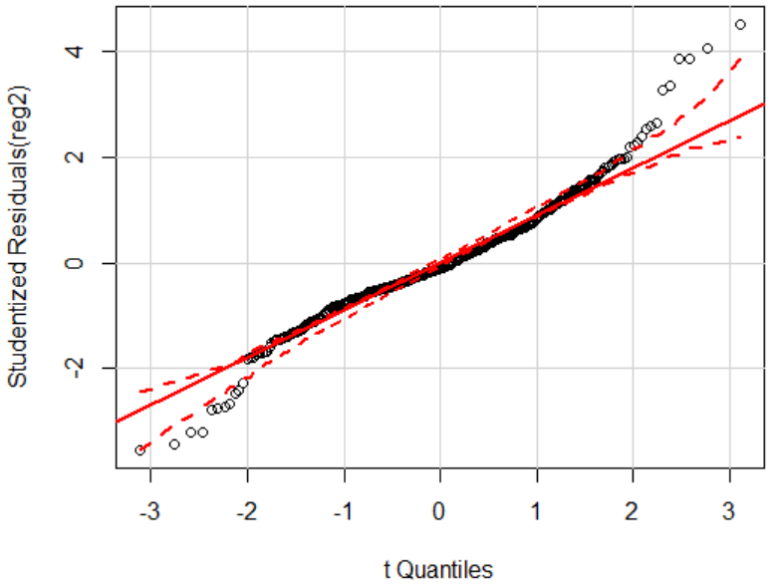
Right Skewed Distribution



Outlier

QQ Plots

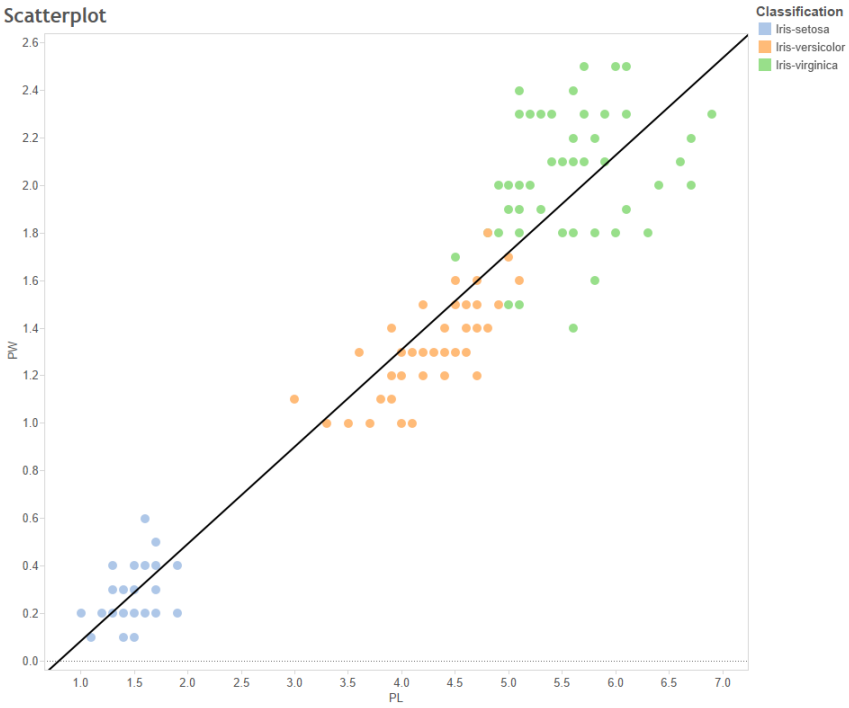
- In statistics, a QQ plot ("Q" stands for quantile) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- A QQ plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
- If the quantiles of the theoretical and data distributions agree, the plotted points fall on or near the line $y = x$
- This can provide an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary.



Quantile-Quantile Plot Diagnostics	
Description of Point Pattern	**Possible Interpretation**
all but a few points fall on a line	outliers in the data
left end of pattern is below the line; right end of pattern is above the line	long tails at both ends of the data distribution
left end of pattern is above the line; right end of pattern is below the line	short tails at both ends of the data distribution
curved pattern with slope increasing from left to right	data distribution is skewed to the right
curved pattern with slope decreasing from left to right	data distribution is skewed to the left
staircase pattern (plateaus and gaps)	data have been rounded or are discrete

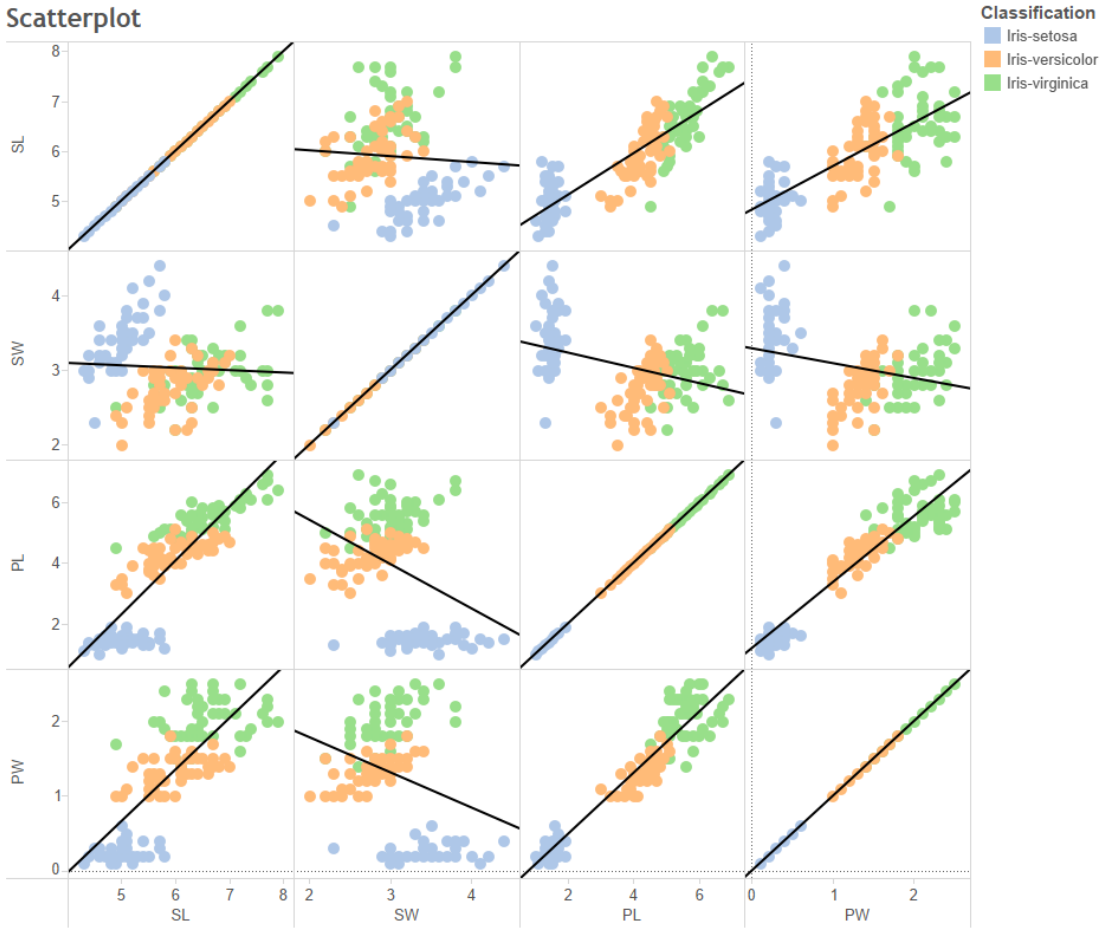
Scatterplots

- A point plot between two variables used to understand the spread of the data.
- The spread of the data allows for us to understand if the data has a non-linear or linear relationship and the relative degree of the correlation in the data.
- This technique can allow be used to detect outliers in the data.
- A scatterplot becomes more powerful when you incorporate categorical data as an additional dimension.



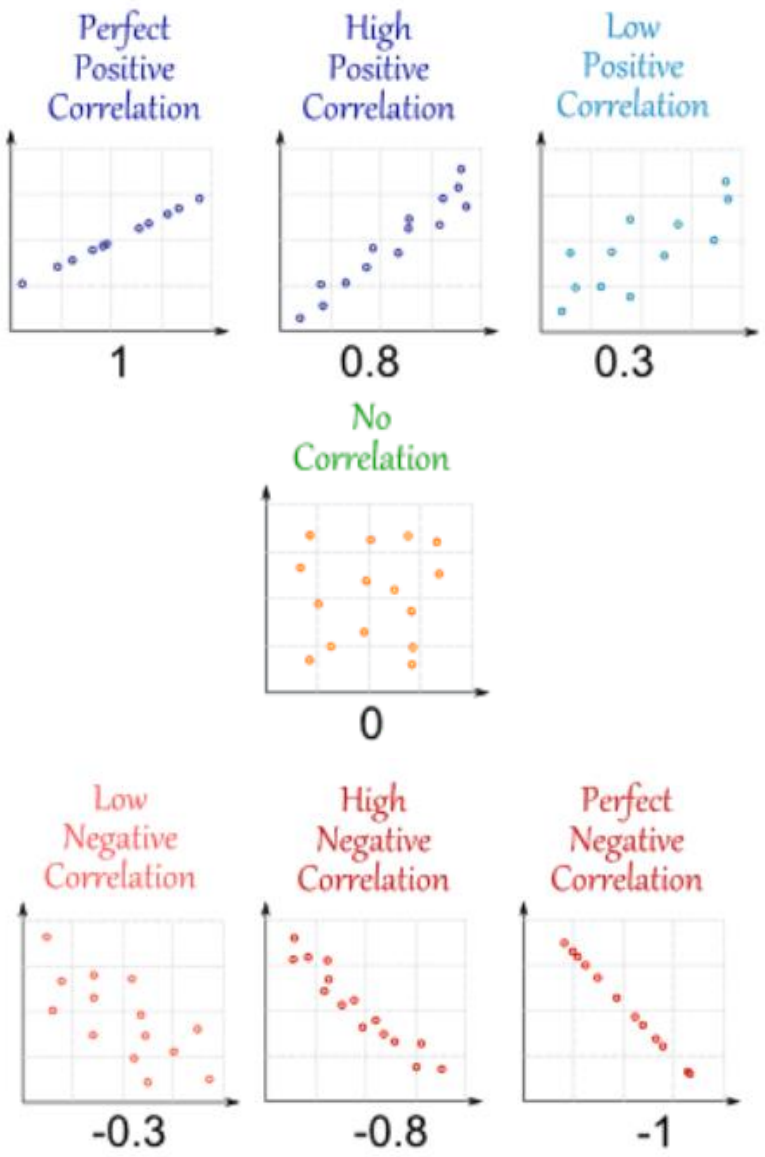
Scatterplots

- Scatterplot Matrices can offer a 10,000 feet view of the patterns within the data.



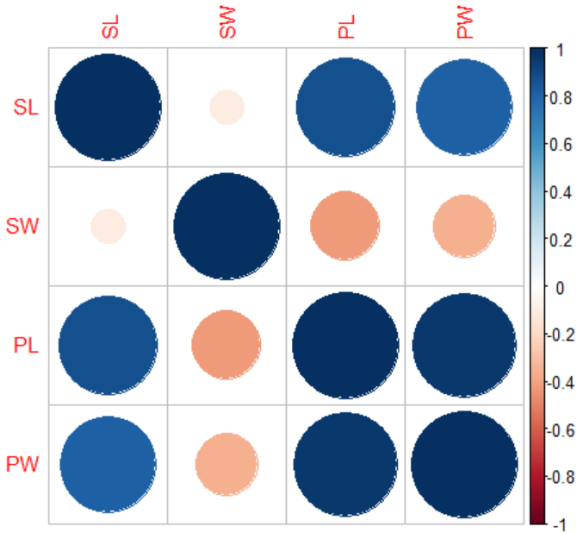
Correlation

- When two sets of data are strongly linked together we say they have a high correlation.
  - Correlation is Positive when the values increase together, and
  - Correlation is Negative when one value decreases as the other increases.
- Correlation can have a value ranging from:
  - 1 is perfect positive correlation
  - 0 implies that there is no correlation
  - -1 is perfect negative correlation
- Correlation Does Not Imply Causation !!!!



Correlation

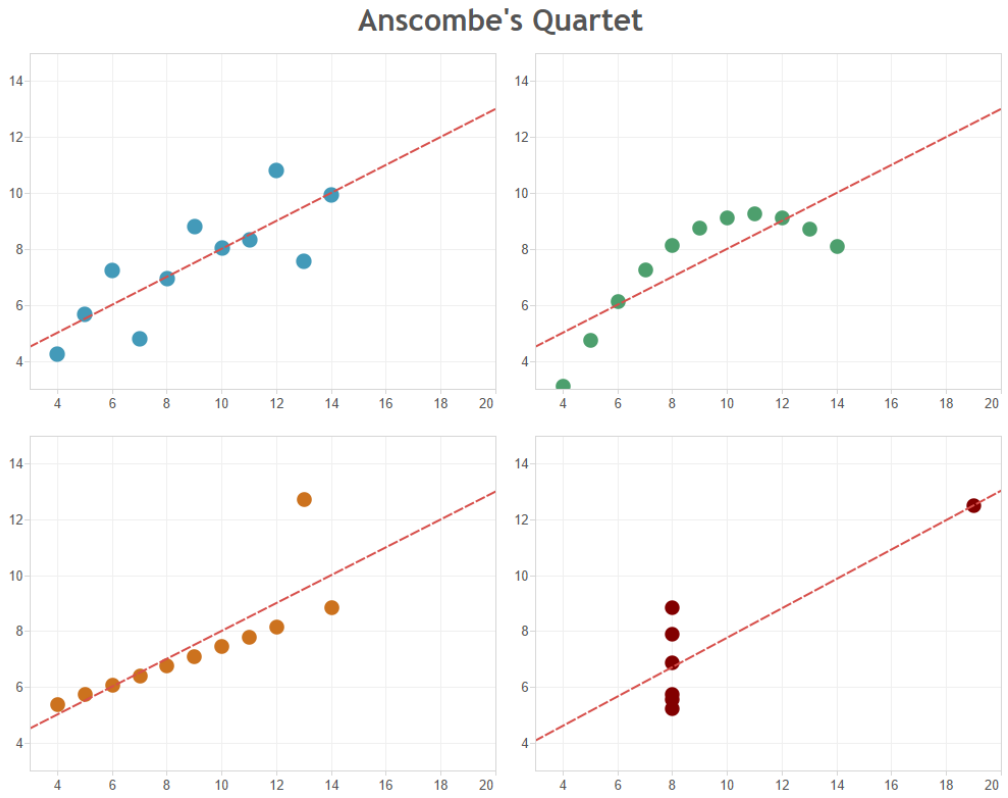
	SL	SW	PL	PW
SL	1.00	-0.11	0.87	0.82
SW	-0.11	1.00	0.42	-0.36
PL	0.87	-0.42	1.00	0.96
PW	0.82	-0.36	0.96	1.00



- The creation of a correlation matrix and correlation heat maps make it easier to have a top level view of the data.
- Remember that data that is closer to perfect negative or positive correlations will be stronger for linear model building.
- We need to be careful about the assuming that lower correlation intervals are potentially poor variables for model building.
- Evaluation of correlation should be handled on a case by case basis.
- Example: In behavioral sciences, it is not uncommon to have correlations between 0.3 and 0.4 for significant variables.

Understanding the Data

- This is a powerful example on why we need to understand the data in order to model the information correctly.
- Anscombe's quartet shows how different data spreads can produce identical regression models and coefficients.
- These are the same models with identical mean, variance, and correlation values.
- Some of these models requires data transformations to achieve linearity, others have outliers that strongly influence the models performance.



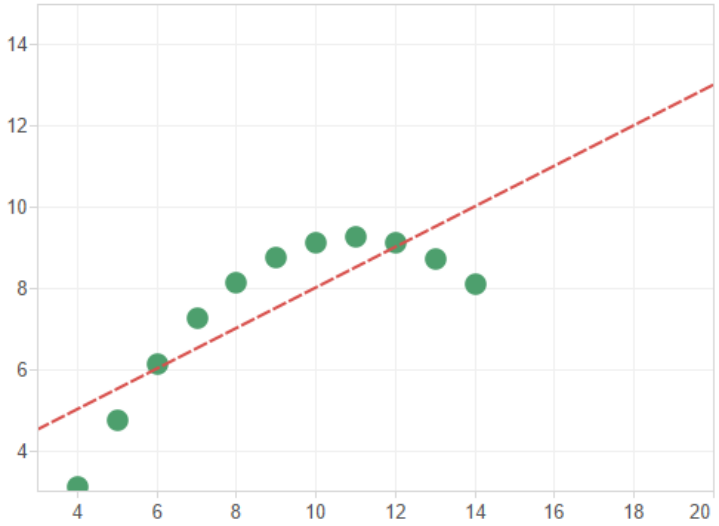
Fundamentals of Transformations

When data in a scatterplot (box plot or bell curve) indicates that a non-linear relationship exists between the dependent and independent variable, we can use a data transformation to achieve linearity.

This is a vital task for models which require linearity (regression, time series, etc…) in the variables.

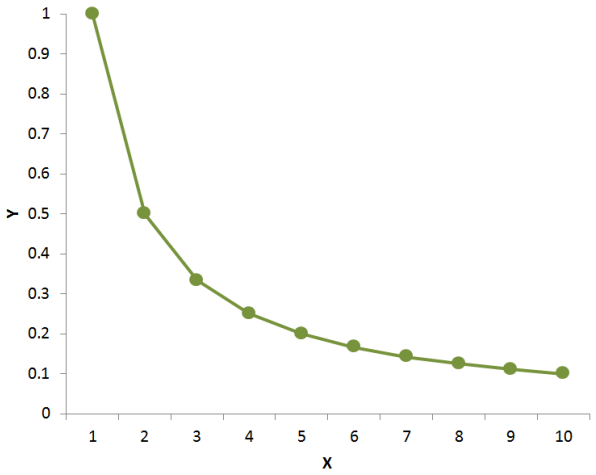
Here are some types of transformations we will go over:

- Power Function
- Exponential Function
- Polynomial Function
- Logarithmic Function
- Square Root Function

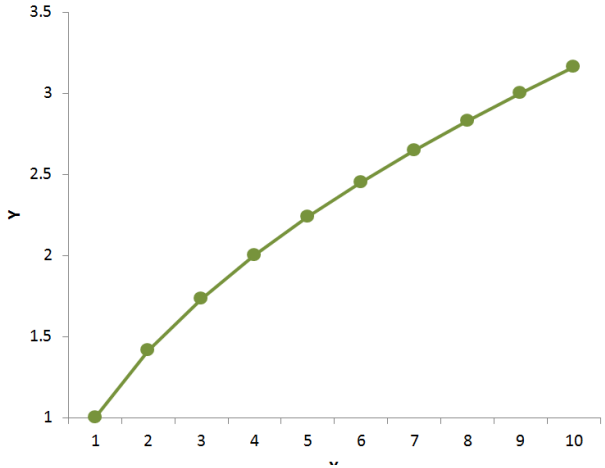


Fundamentals of Transformations

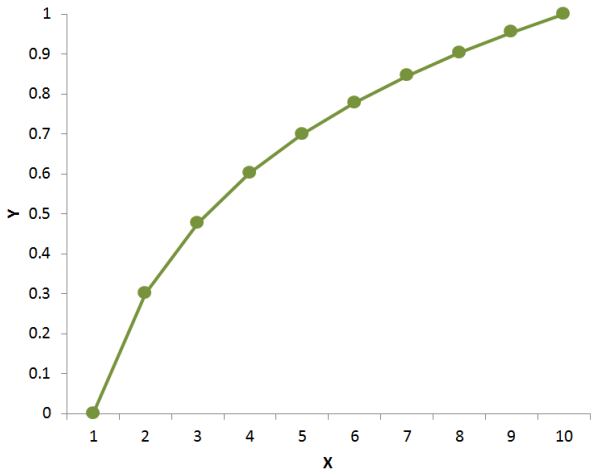
Reciprocal Function: Y = 1 / X



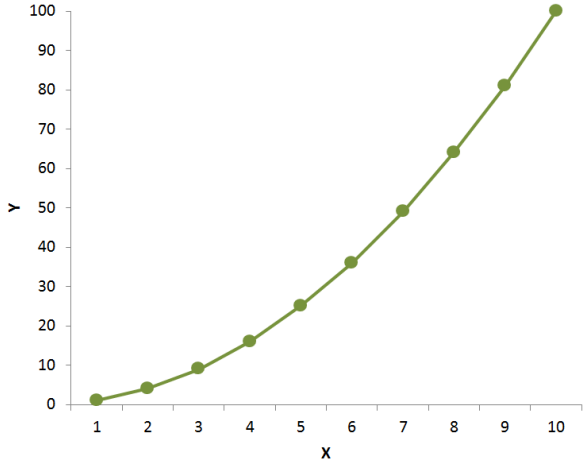
Square Root Function: Y = √ X



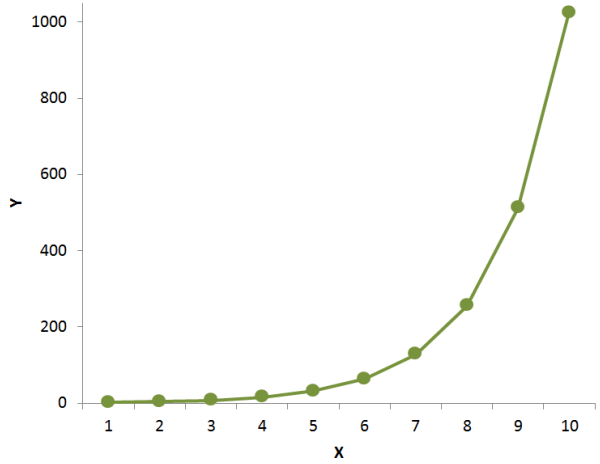
Log Function: Y = log X



Power Function: Y = X^2



Exponential Function: Y = 2^x



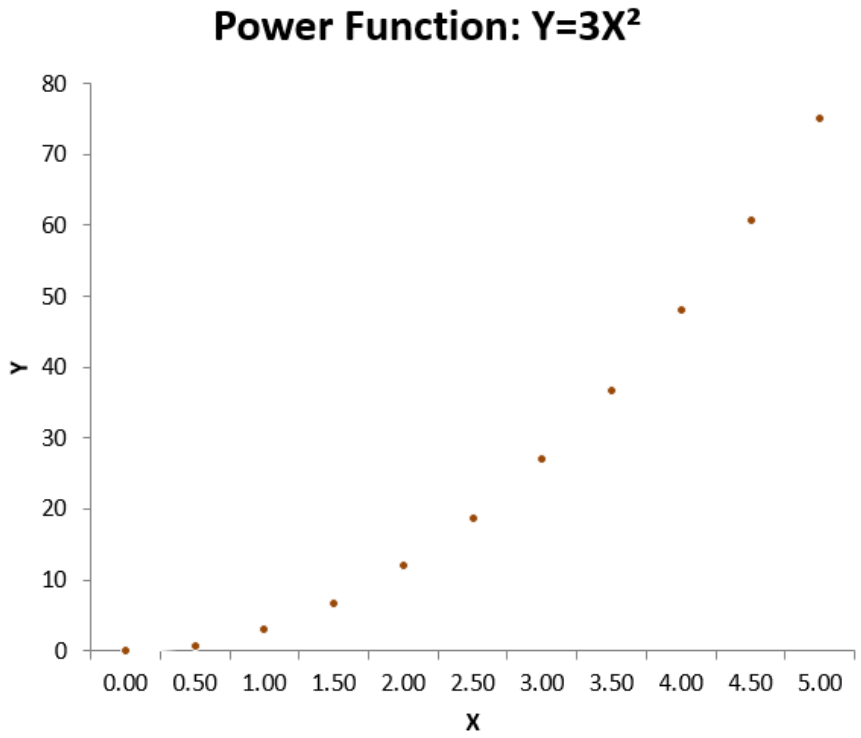
Power Function

A power function takes the following form:

$$y = a * x^b$$

- It's important to note the location of the variable x in the power function. Later in this activity, we will contrast the power function with the exponential function, where the variable x is an exponent, rather than the base as in the equation.

Power Function



- It could be argued that the data is somewhat linear, showing a general upward trend. However, it is more likely that the function is nonlinear, due to the general bend.

Power Function

- Let's take the logarithm of both sides of the power function:

$$y = 3x^2$$

- The base of the logarithm is irrelevant; however, log x is understood to be the natural logarithm (base e) in R. That is, log x = $\log_e$ x in R.

$$\log y = \log 3x^2$$

- The log of a product is the sum of the logs, so we can write the following.

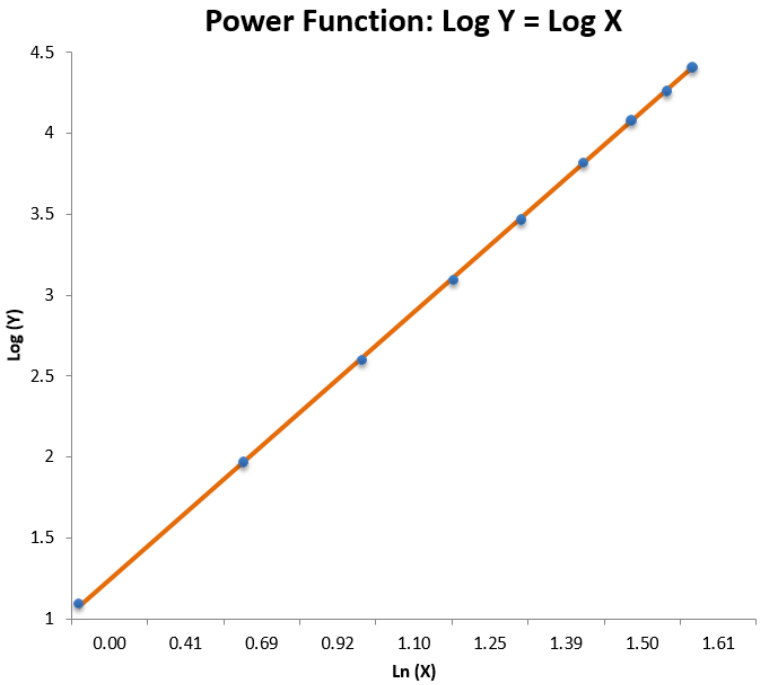
$$\log y = \log 3 + \log x^2$$

- Another property of logs allows us to move the exponent down.

$$\log y = \log 3 + 2 \log x$$

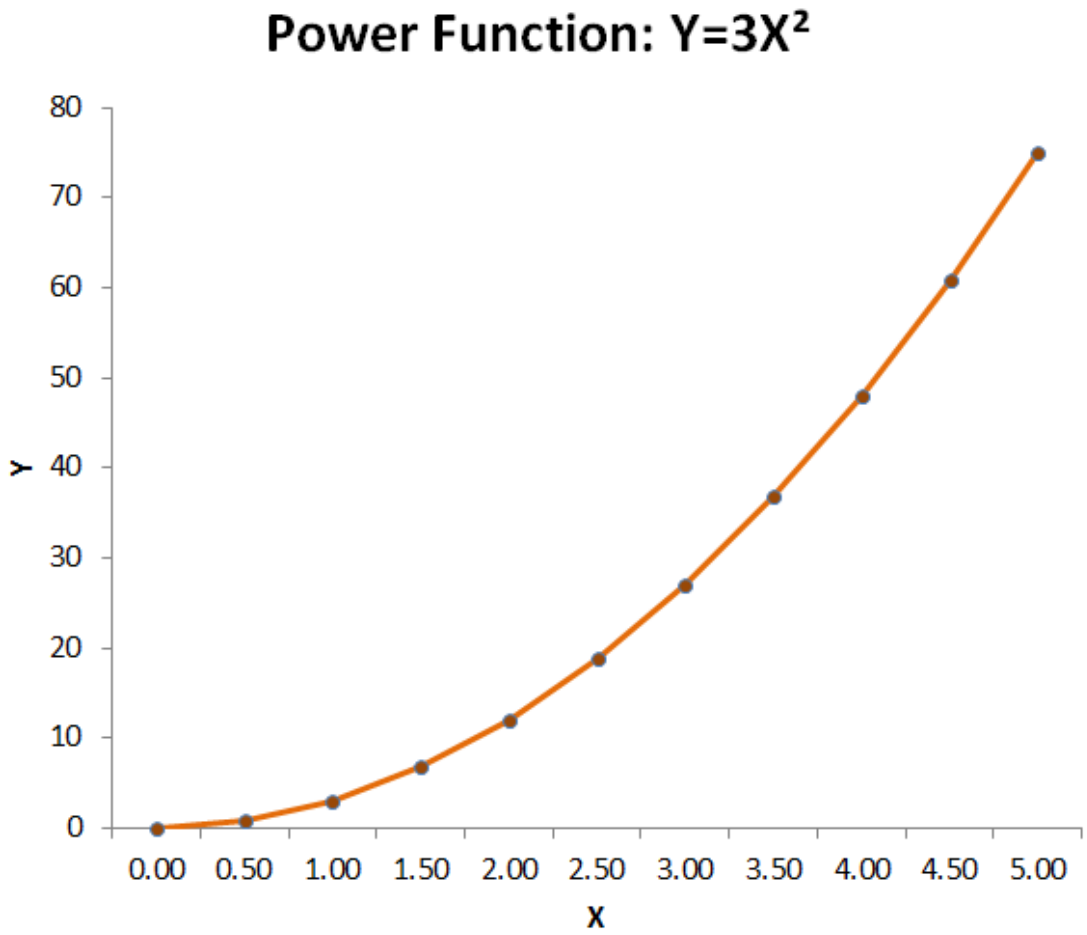
Power Function

```
Call:
lm(formula = log(y) ~ log(x))

Coefficients:
(Intercept)       log(x)
      1.099        2.000

>
> log(3)
[1] 1.098612
```



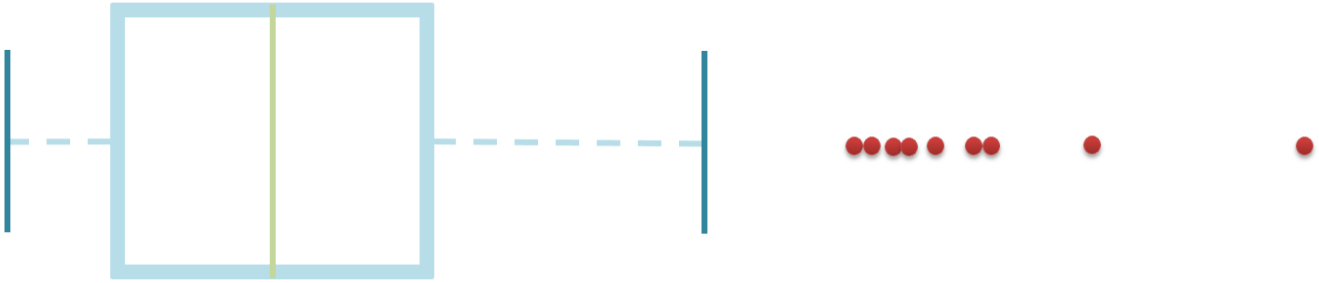
- The slope is 2, which is the slope indicated in log y = log 3 + 2 log x. Supposedly, the intercept should be log 3.
- Important: If the graph of the logarithm of the response variable versus the logarithm of the independent variable is a line, then we should suspect that the relationship between the original variables is that of a power function.

Power Function

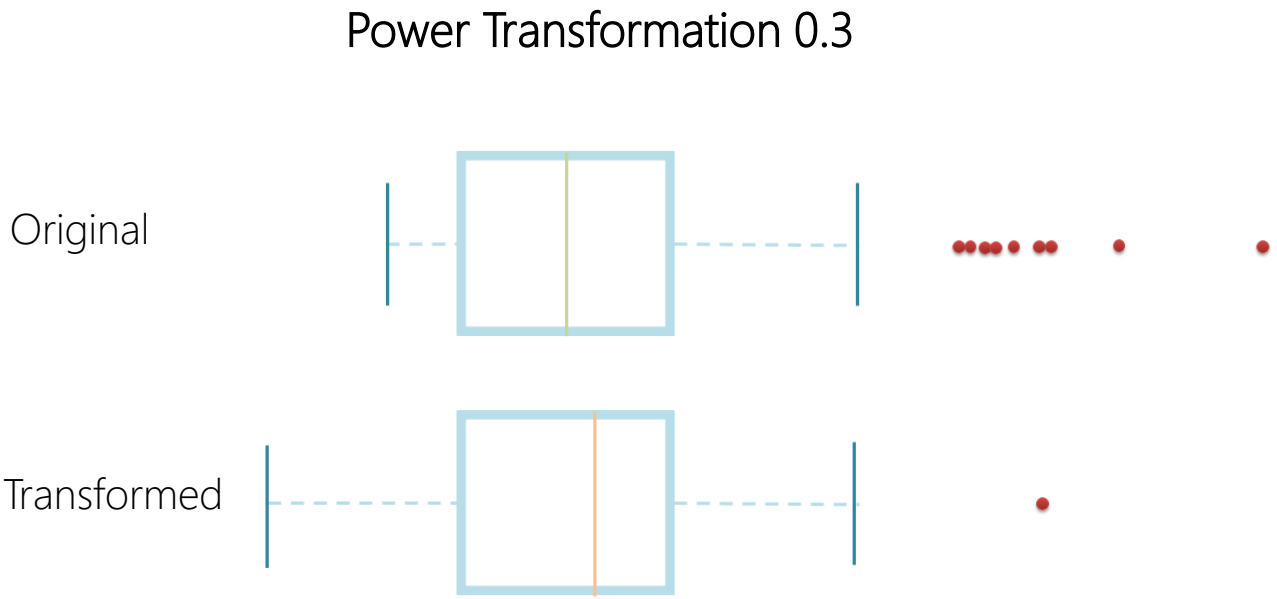


Power Transformation

- This box-plot is a representation of salaries for 500 employees at a University.
- What is the shape of the dataset? It depends what you are looking at…
- If you look at the middle 50% of the salaries, they look roughly symmetric.
- On the other hand, if you look at the tails (the portion of the data beyond the quartiles), then there is clear right-skewness, and a number of high outliers. One would expect to find a few large salaries at a university.



Power Transformation



Observations:

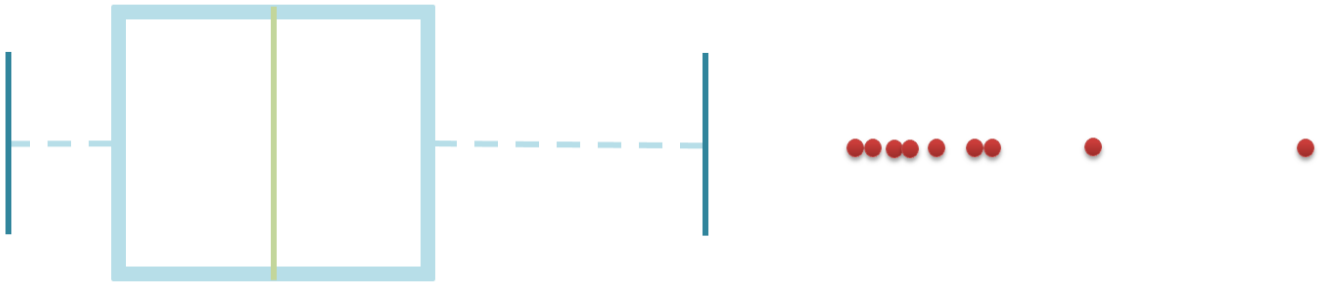
- If we look at the middle 50%, we don't have symmetry (data is left skewed)
- If we look at the tails, the re-expressed data is roughly symmetric

Power Transformation

What have we learned from this exercise?

- For some data, it can be difficult to achieve symmetry of the whole batch. Here the power transformation was helpful in making the tails symmetric (and control the outliers), but it does not symmetrize the middle salaries.

- In practice, one has to decide on the objective. If we are focusing on the middle part of the salaries, then no re-expression is necessary. But if we want to look at the whole dataset and control the outliers, then perhaps the choice of power = 0.3 is a better choice.



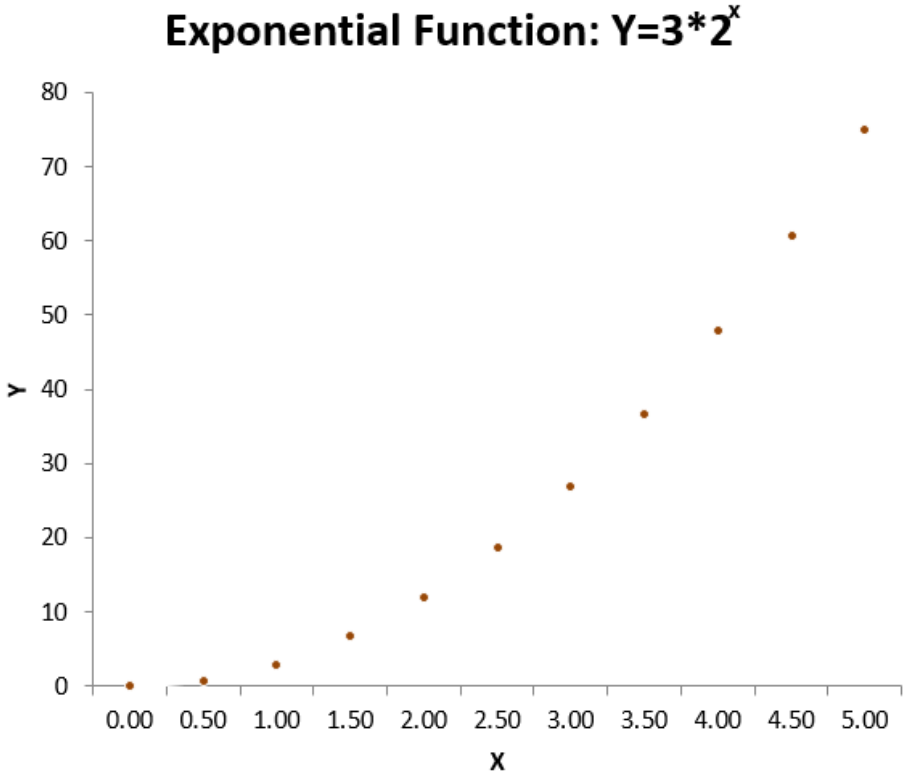
Exponential Function

A exponential function takes the following form:

$$y = a * b^x$$

- In the power function, the independent variable was the base, as in $y = a\ x^b$. In the exponential function, the independent variable is now an exponent, as in $y = a\ b^x$. This is a subtle but important difference.

Exponential Function



- It could be argued that the data is somewhat linear, showing a general upward trend. However, it is more likely that the function is nonlinear, due to the general bend.

Exponential Function

- Let's take the logarithm of both sides of the exponential function:

$$y = 3*2^x$$

- The base of the logarithm is irrelevant; however, log x is understood to be the natural logarithm (base e) in R. That is, log x = $\log_e$ x in R.

$$\log y = \log 3(2^x)$$

- The log of a product is the sum of the logs, so we can write the following.

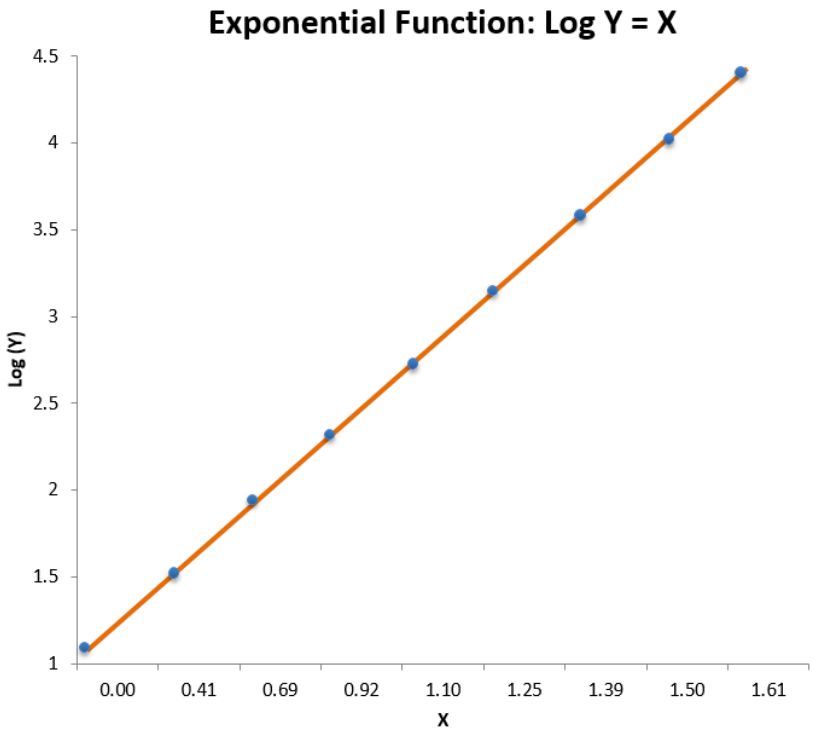
$$\log y = \log 3 + \log 2^x$$

- Another property of logs allows us to move the exponent down.

$$\log y = \log 3 + x \log 2$$

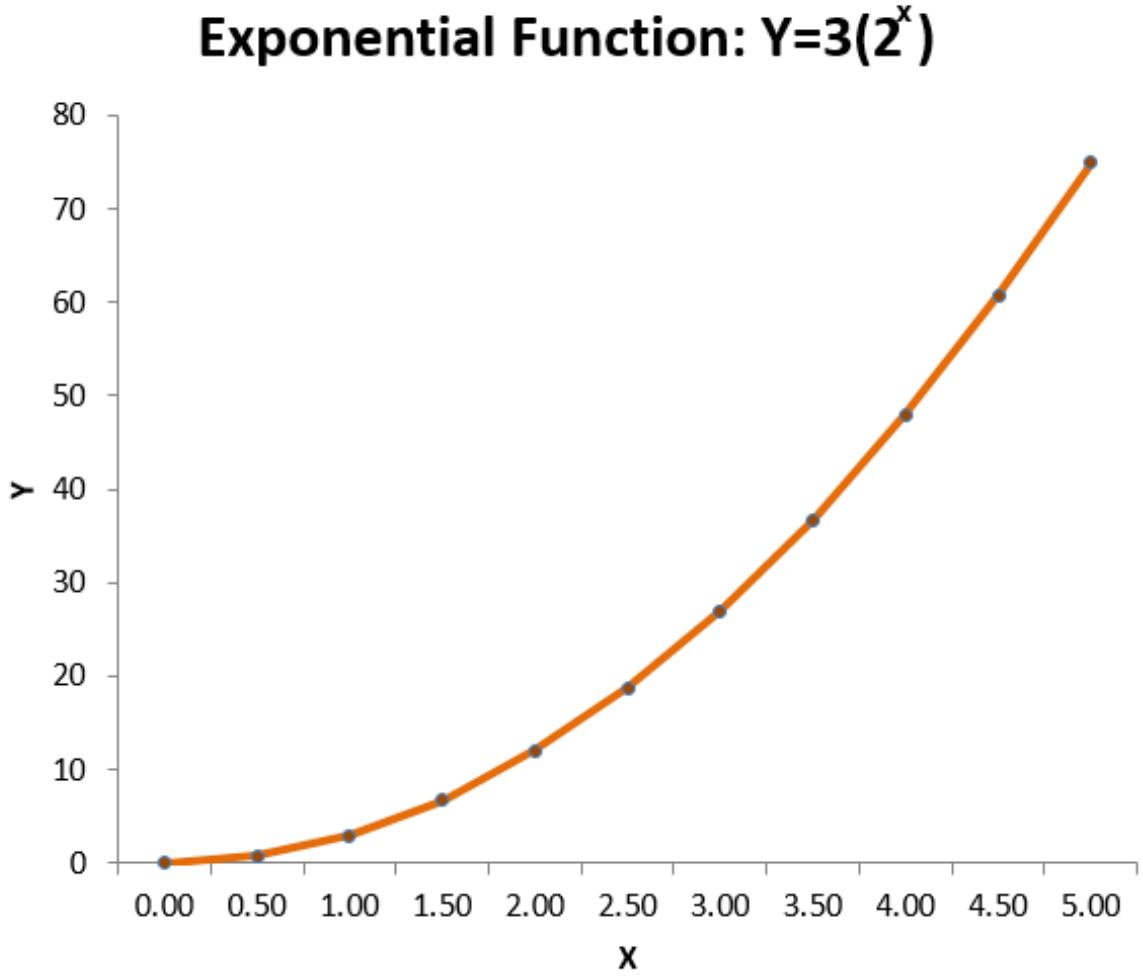
Exponential Function

```
Call:
lm(formula = log(y) ~ log(x))

Coefficients:
(Intercept)       log(x)
      1.099        2.000

>
> log(3)
[1] 1.098612
```

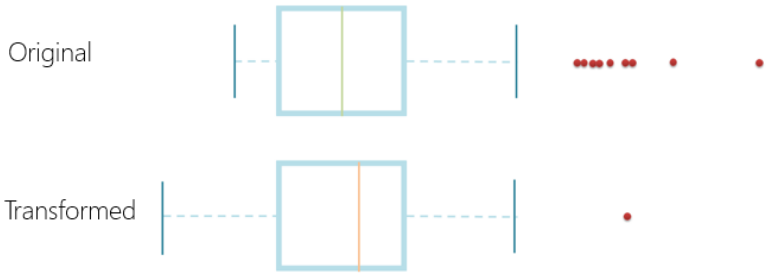


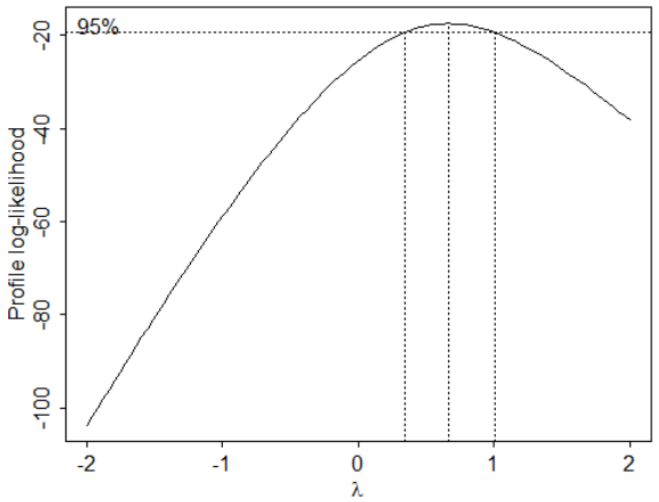
- The slope is 2, which is the slope indicated in log y = log 3 + 2 log x. Supposedly, the intercept should be log 3.
- Important Result: If the graph of the logarithm of the response variable versus the logarithm of the independent variable is a line, then we should suspect that the relationship between the original variables is that of a power function.

Exponential Function



Box-Cox Method





- How did we know to use a power function of 0.3 in the power transformation example?
- The statisticians George Box and David Cox developed a procedure to identify an appropriate exponent (Lambda = l) to use to transform data into a "normal shape."
- The Lambda value indicates the power to which all data should be raised.
- In order to do this, the Box-Cox power transformation searches from Lambda = -5 to Lamba = +5 until the best value is found

Box-Cox Method

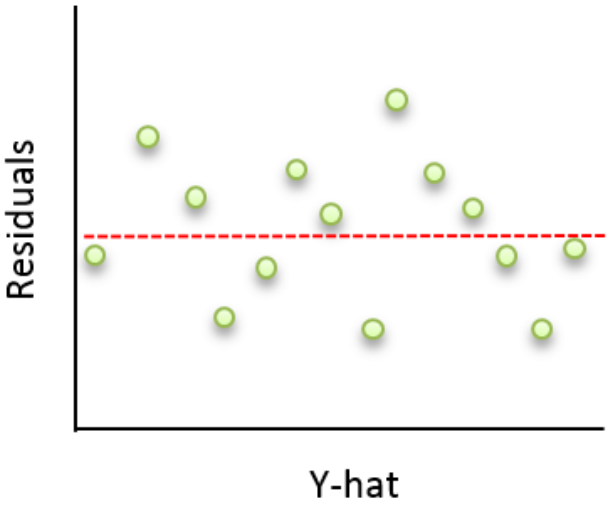
- Here is a table which indicates which linear transformation to apply for various lambda values:

Best λ	Equation	Name
-2.5 to -1.5	$1/y^2$	inverse square
-1.5 to -0.75	$1/y$	reciprocal
-0.75 to -0.25	$1/\sqrt{y}$	inverse square root
-0.25 to 0.25	$\ln(y)$	natural log
0.25 to 0.75	$\sqrt{y}$	square root
0.75 to 1.5	y	none
1.5 to 2.5	y^2	square

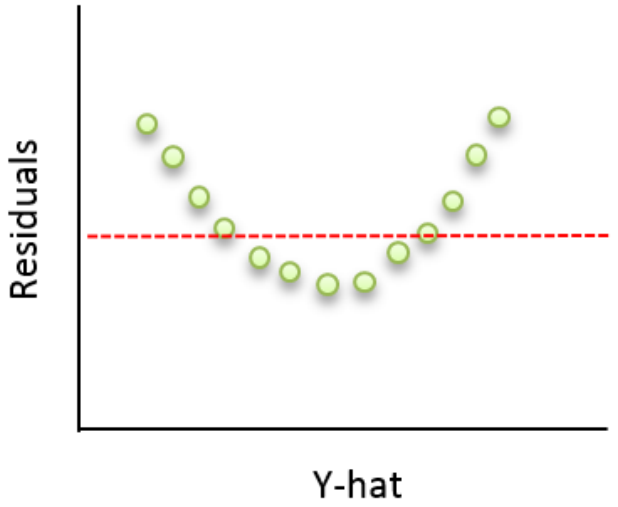
Residual Plots

- A residual plot is a scatterplot of the residuals (difference between the actual and predicted value) against the predicted value.
- A proper model will exhibit a random pattern for the spread of the residuals with no discernable shape.
- Residual plots are used extensively in linear regression analysis for diagnostics and assumption testing.
- For our purposes in the EDA, they can help to identify when a transformation is appropriate.
- If the residuals form a curvature like shape, then we know that a transformation will be necessary and can explore some methods like the Box-Cox.

Random Residuals



Curved Residuals



Transformation Guide

- Transforming a data set to enhance linearity is a multi-step, trial-and-error process.
- Step 1: Conduct a standard regression analysis on the raw data.
- Step 2: Construct a residual plot.
  - If the plot pattern is random, do not transform data.
  - If the plot pattern is not random, continue.
- Step 3: Compute the coefficient of determination (R2).
- Step 4: Choose a transformation method.
- Step 5: Transform the independent variable, dependent variable, or both.
- Step 6: Conduct a regression analysis, using the transformed variables.
- Step 7: Compute the coefficient of determination (R2), based on the transformed variables.
  - If the transformed R2 is greater than the raw-score R2, the transformation was successful. Congratulations!
  - If not, try a different transformation method.

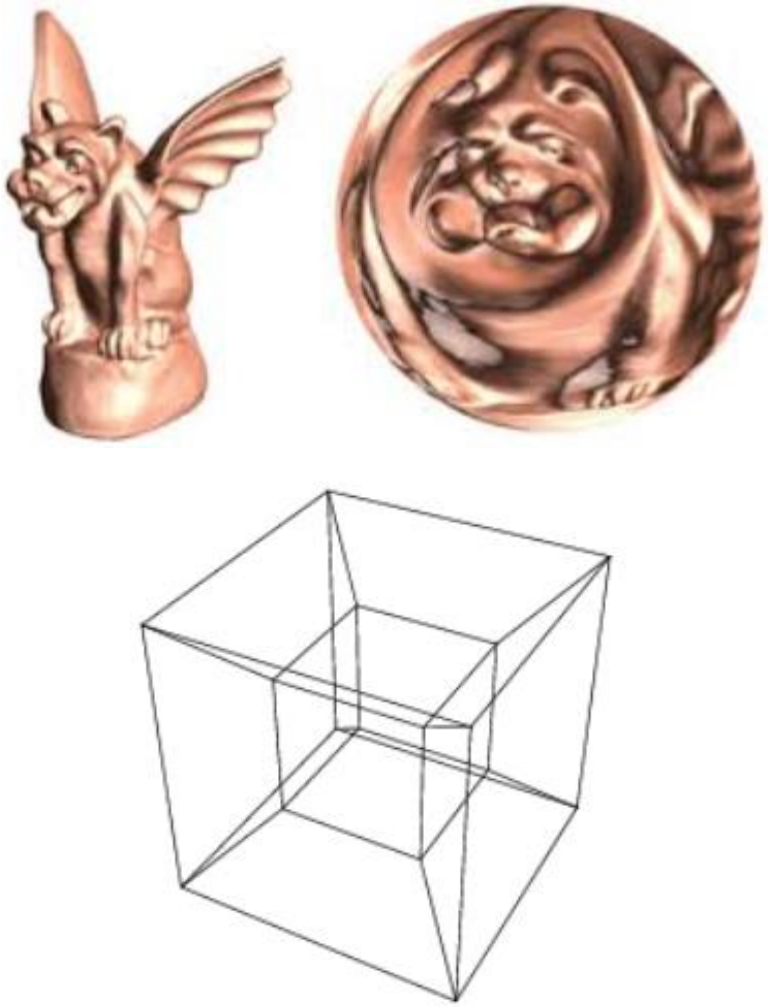
Variable Selection Methods

- How do we approach a modeling task with 100's or 1000's of independent variables? What if there are a significant number of variables with low correlations?

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	etc...	etc...	Variable 100
0	0	0	1	0	etc...	etc...	0
1	0	1	0	1	etc...	etc...	1
0	1	0	1	0	etc...	etc...	0
1	1	0	0	1	etc...	etc...	1
0	0	1	1	0	etc...	etc...	0
0	1	0	0	0	etc...	etc...	0
1	0	1	0	1	etc...	etc...	1
0	0	0	1	1	etc...	etc...	0
1	1	1	0	0	etc...	etc...	1
1	0	0	1	0	etc...	etc...	0
0	1	0	0	1	etc...	etc...	0
0	0	1	0	1	etc...	etc...	1

Dimensionality Reduction

- Dimensionality reduction is the process of reducing the number of variables under consideration and can be represented in feature selection and feature extraction approaches.
- Feature Selection – Methodology that attempts to find a subset of the original variables using some form of selection criteria. These approaches are primarily through filtering (Ex. information gain) and wrappers (search guided by accuracy).
- Feature Extraction – Transforms the data in a higher dimensional space to a space of lower dimensions. A linear approach is called a Principal Component Analysis (PCA).



Feature Selection Methods

- When attempting to describe a model containing dependent and independent variables, it is desirable to be as accurate as possible, but at the same time to use as few variables as possible (principal of parsimony).
  - Forward Selection Procedure - Starts with no variables in the model. The term that adds the most is identified and added to the model. The remaining variables are analyzed and the next most "helpful" variable is added. This process continues until none of the remaining variables will improve the model.
  - Backward Selection Procedure - Starts with the full model, and proceed by sequentially removing variables that improve the model by being deleted, until no additional deleting will improve the model.
  - Stepwise Selection Procedure – Starts like a forward selection procedure. After adding a variable to the model, run a backward selection procedure on the existing variable pool within the model to try and eliminate any variables. Continue until every remaining variable is significant at the cutoff level and every excluded variable is insignificant.

Feature Selection Methods

Analysis of Deviance Table					
Initial Model:					
Mileage ~ Price + Make + Model + Trim + Type + Cyl + Liter + Doors + Cruise + Sound + Leather					
Final Model:					
Mileage ~ Price + Model + Trim + Sound + Leather					
Step	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**AIC**
1			730	12854358402	13484.23
2 - Doors	0	0	730	12854358402	13484.23
3 - Liter	0	0	730	12854358402	13484.23
4 - Cyl	0	0	730	12854358402	13484.23
5 - Type	0	0	730	12854358402	13484.23
6 - Make	0	0	730	12854358402	13484.23
7 - Cruise	1	4733781	731	12859092183	13484.23

- This is an example of stepwise selection procedure which contains a large number of independent variables.
- After the routine has completed, note the variable list has been reduced.
- The variables which were removed did not meet a certain statistical confidence level (P-Value) which was pre-defined.
- Unless we have specific subject matter expertise and expert level knowledge of the dataset at hand, we should set the P-Value as 0.05 or smaller.

Principal Component Analysis

- The Principal Component Analysis procedure will take a number of variables and reduce them down to a smaller number of variables called components.

- This example took the dependent variables Price, Software, Aesthetics, & Brand and created 4 Components.

Price	Software	Aesthetics	Brand
6	5	3	4
7	3	2	2
6	4	4	5
5	7	1	3
7	7	5	5
6	4	2	3
5	7	2	1

Importance of components:				
	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**
Standard deviation	1.558939	0.980409	0.681667	0.379258
Proportion of Variance	0.607573	0.240301	0.116198	0.035959
Cumulative Proportion	0.607573	0.847873	0.964041	1

Loadings:				
	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**
Price	0.523		0.848	
Software	0.177	0.977	-0.12	
Aesthetics	-0.597	0.134	0.295	-0.734
Brand	-0.583	0.167	0.423	0.674

- The cumulative proportion figure of 0.847 for Comp. 2 tells us that the first two components account for 84.7% of the total variability of the data. Generally speaking, 80% accounts for the data rather well.

Principal Component Analysis

- The loadings can be thought of as new variables which have been created for our analysis.
- The way to interpret and calculate the first component variable is: Comp.1 = 0.523 * Price + 0.177 * Software - 0.597 * Aesthetics - 0.583 * Brand

Importance of components:				
	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**
Standard deviation	1.558939	0.980409	0.681667	0.379258
Proportion of Variance	0.607573	0.240301	0.116198	0.035959
Cumulative Proportion	0.607573	0.847873	0.964041	1

Loadings:				
	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**
Price	0.523		0.848	
Software	0.177	0.977	-0.12	
Aesthetics	-0.597	0.134	0.295	-0.734
Brand	-0.583	0.167	0.423	0.674

```
##########################################################
# PCA and logistic Regression
##########################################################

# Fit the model with the first two principal components.

model <- glm(OS ~ pca$scores[,1] + pca$scores[,2],
             data=mydata2, family=binomial)

summary(model)
```

Before You Start Model Building

Once you have the data in the proper format, before you perform any analysis you need to explore the data and prepare it first.

- All of the variables are in columns and observations in rows.
- We have all of the variables that you need.
- There is a at least one variable with a unique ID.
- Have a backup of the original dataset handy.
- We have properly addressed any data munging tasks.

Splitting the Dataset

- Another important step regarding predictive model building involves splitting the dataset into a training and testing dataset. This is performed to ensure that our models are validated against data that was not part of the models construction.

- To accomplish this, we will randomly select observations for training dataset (70%) and the remaining into the test dataset (30%).

Training Dataset

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
$42,000.00	5850	3	1	2	yes	no	yes
$38,500.00	4000	2	1	1	yes	no	no
$49,500.00	3060	3	1	1	yes	no	no
$60,500.00	6650	3	1	2	yes	yes	no
$61,000.00	6360	2	1	1	yes	no	no
$66,000.00	4160	3	1	1	yes	yes	yes
$66,000.00	3880	3	2	2	yes	no	yes
$69,000.00	4160	3	1	3	yes	no	no
$83,800.00	4800	3	1	1	yes	yes	yes
$88,500.00	5500	3	2	4	yes	yes	no
$90,000.00	7200	3	2	1	yes	no	yes
$30,500.00	3000	2	1	1	no	no	no
$27,000.00	1700	3	1	2	yes	no	no

Testing Dataset

Spending Our Data

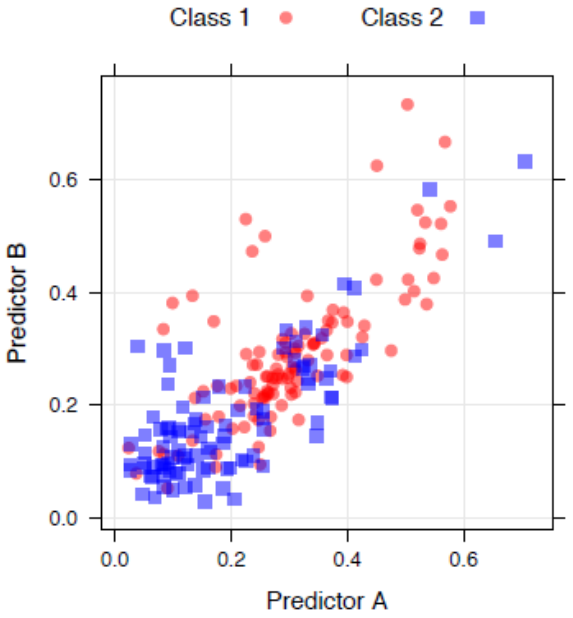
- In our previous example, we had decided to use a split of 70% for the model building dataset or training set and 30% for our validation dataset.
- The more data that we spend on the training set, the better estimates that we will get. (Provided the data is accurate.)
- Given a fixed amount of data:
  - Too much spent in training wont allow us to get a good assessment of predictive performance. We may find a model that fits the training data very well, but is not generalizable (over-fitting).
  - Too much spent in testing wont allow us to get a good assessment of model parameters.
- Many business consumers of these models emphasize the need for an untouched set of samples to evaluate performance.

Over-Fitting and Resampling

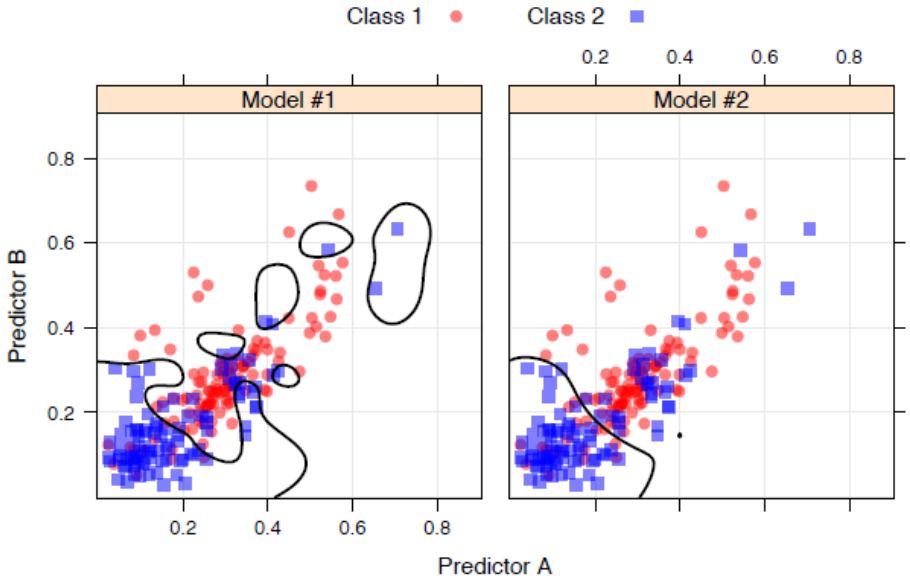
- Over-fitting occurs when a model inappropriately picks up on trends in the training set that do not generalize to new samples.
- When this occurs, assessments of the model based on the training set can show good performance that does not reproduce in future samples.
- Some predictive models have dedicated "knobs" to control for over-fitting.
  - Neighborhood size in nearest neighbor models for example.
  - The number of splits in a decision tree model.
- Often, poor choices for these parameters can result in over-fitting. When in doubt, consult with statisticians and data scientists on how to best approach the tuning.

Over-Fitting and Resampling

Data Set



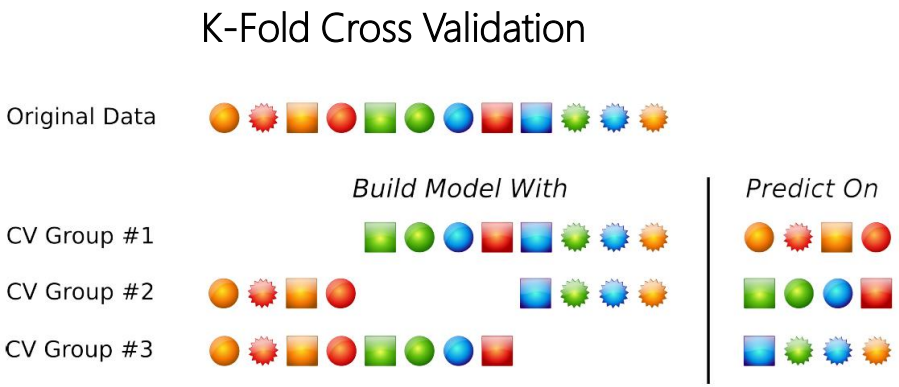
Two Model Fits



Over-Fitting and Resampling

- One obvious way to detect over-fitting is the use of a test set. The predictive performance measures that we use to evaluate the performance should be consistent across the training and test-set.
- However, repeated "looks" at the test-set can also lead to over-fitting.
- Resampling the training samples allows us to know when we are making poor choices for the values of these parameters (the test set is not used).
- Resampling methods try to "inject variation" into the system in order to approximate the models performance on future samples.
- We will walk through several types of resampling methods for training set samples.

Resampling – K Folds Cross Validation

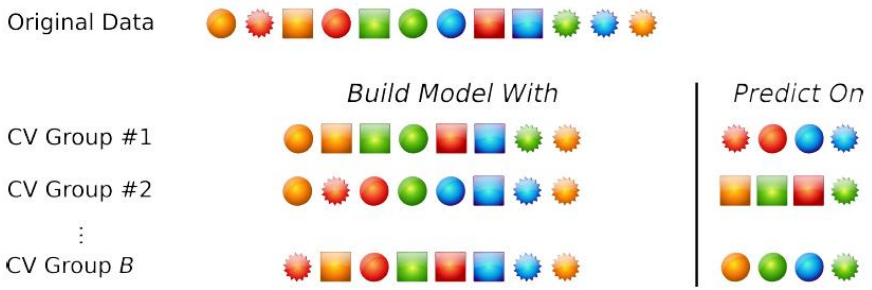


- Here, we randomly split the data into K-distinct blocks of roughly equal size.
  - We leave out the first block of data and fit a model.
  - This model is used to predict the held out block.
  - We continue this process until we've predicted all K held out blocks.
- The final performance is based upon the hold-out predictions.
- K is usually taken to be 5 or 10 and leave one out cross validation as each sample as a block.

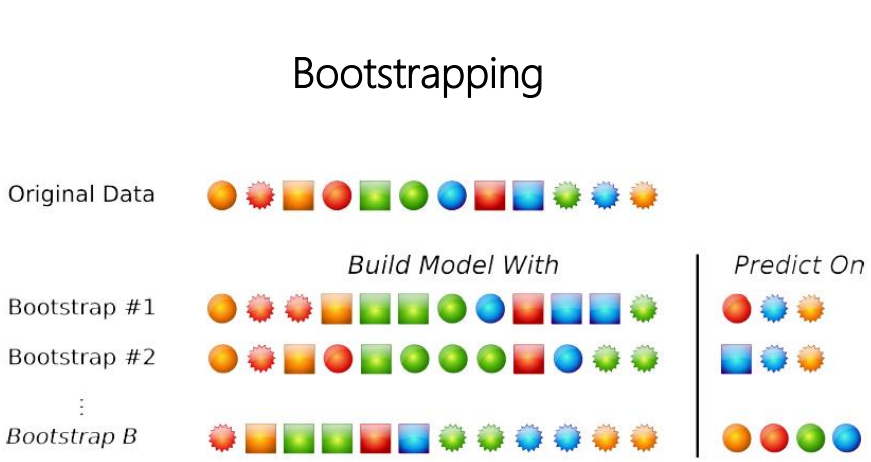
Resampling – Repeated Splits

- A random proportion of data (say 70%) are used to train a model while the remainder is used for prediction.
- The process is repeated many times and the average performance is used.
- These splits can also be generated using stratified sampling.
- With many iterations (20 to 100), this procedure has smaller variance than K-Fold CV, but is likely to be biased.

Repeated Training/Testing Splits



Resampling – Bootstrapping



- Bootstrapping takes a random sample of the data with replacement. The random sample is the same size as the original dataset.
- Samples may be selected more than once and each sample has approx. 63.2% chance of showing up at least once.
- Some samples wont be selected and these samples will be used to predict performance.
- The process is repeated multiple times (say 30-100).
- The procedure also has low variance but non zero bias when compared to K-fold CV.

The Big Picture

- We think that resampling will give us honest estimates of future performance, but there is still the issue of which model to select.

- One algorithm to select predictive models:


```
Define sets of model parameter values to evaluate;
for each parameter set do
    for each resampling iteration do
        Hold–out specific samples ;
        Fit the model on the remainder;
        Predict the hold–out samples;
    end
    Calculate the average performance across hold–out predictions
end
Determine the optimal parameter set;
```

Selecting the Right Algorithm

- There are 3 broad categories of activities that we can perform and we need to understand how each modeling activity we are performing corresponds to each bucket.
  - Predictive Models - Predictive models are models of the relation between the specific performance of a unit in a dataset and one or more known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different dataset will exhibit the specific performance.
  - Descriptive Models - Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.
  - Decision Models - Decision models describe the relationship between all the elements of a decision — the known data (including results of predictive models), the decision, and the forecast results of the decision — in order to predict the results of decisions involving many variables. These models can be used in optimization, maximizing certain outcomes while minimizing others.

Predictive Models

- If the goal of the activity is to predict a value in a dataset, we first need to understand more about the dataset to determine the correct "tool" or approach.
- Lets examine the following dataset:
- Our goal is to predict the price of a home based upon the lot size and number of bedrooms. Notice that there is no time element in the dataset.
- The price variable is called the dependent variable and the lotsize/ bedrooms are the independent variables.
- For now, we will focus our attention on understanding the characteristics of the dependent variable.

price	lotsize	bedrooms
$ 42,000.00	5850	3
$ 38,500.00	4000	2
$ 49,500.00	3060	3
$ 60,500.00	6650	3
$ 61,000.00	6360	2
$ 66,000.00	4160	3
$ 66,000.00	3880	3
$ 69,000.00	4160	3
$ 83,800.00	4800	3
$ 88,500.00	5500	3
$ 90,000.00	7200	3
$ 30,500.00	3000	2
$ 27,000.00	1700	3

Dependent Variable

Independent Variables

Predictive Models

- The house price variable seems to fall into a range of values which implies that it is continuous in nature.

Dependent Variable Type	Value	Example
Continuous	-∞ to +∞	House Price: $45,000
Binary	0/1	Will they purchase a house?
Qualitative (Classification)	X, Y, Z, etc...	What type of house is it? Large, Medium, Small

Dependent Variable Type	Model Algorithm	Notes
Continuous	Linear Regression	Used to predict a continous numeric value based upon 1 or more independent variables. This approach does not consider time as a factor.
Continuous	Time Series	Used to predict a time dependent continous numeric value based upon 1 or more independent variables.
Continuous	CART	A decision tree model that can be used to determine continuous numeric values
Continuous	Neural Network	A machine learning algorithm that can be applied to a number of tasks.

- Observation: Since there is no time component for this house price in the dataset, we can effectively consider the use of linear regression, CART, or neural network models.

Predictive Models

- If the dependent variable was a binary response (yes/no), there are a number of different methods we would consider using:

Dependent Variable Type	Model Algorithm	Notes
Binary	Logistic Regression	Regression technique used to describe a binary response variable using a logit or probit distribution.
Binary	Survival Analysis	A logistic regression model that also incorporates a time dependent variable.
Binary	Neural Network	A machine learning algorithm that can be applied to a number of tasks.
Binary	Naïve Bayes	Based on Bayes conditional probability rule is used for performing classification tasks.
Binary	Decision Tree	A decision tree is a type of classification algorithm that works by selecting a series of carefully selected questions about the attributes of the test record.
Binary	Support Vector Machine	A machine learning algorithm that can be applied to a number of tasks.
Binary	kNN	A non-parametric method for classifying objects based on closest training examples in the feature space.

Descriptive Models

- If we need to classify/cluster variables together beyond a dichotomous variable, we can use some of the following descriptive techniques:

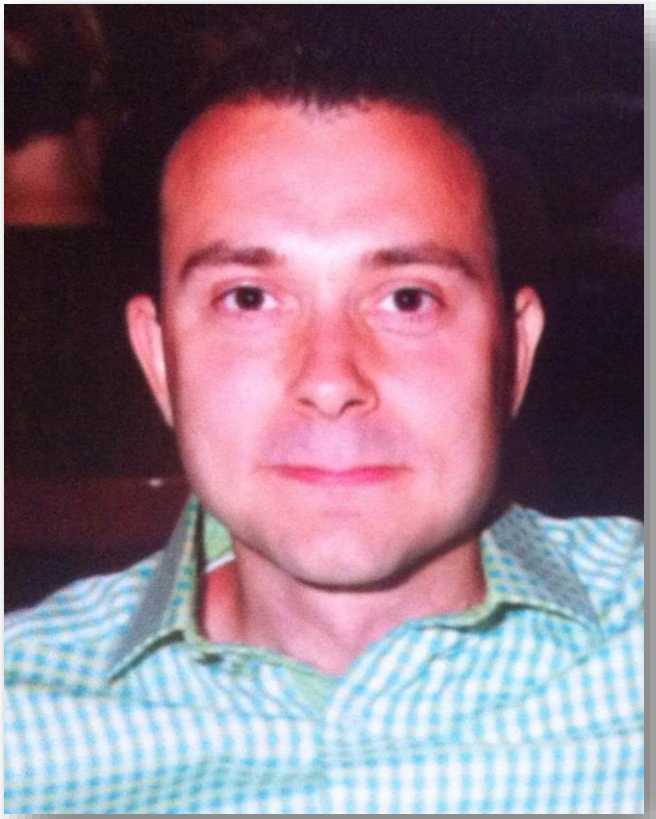
Dependent Variable Type	Model Algorithm	Notes
Qualitative (Classification)	Association Rule	Technique used to find assocaitions between variables in a dataset.
Qualitative (Classification)	Neural Network	A machine learning algorithm that can be applied to a number of tasks.
Qualitative (Classification)	Naïve Bayes	Based on Bayes conditional probability rule is used for performing classification tasks.
Qualitative (Classification)	Decision Tree	A decision tree is a type of classification algorithm that works by selecting a series of carefully selected questions about the attributes of the test record.
Qualitative (Classification)	Support Vector Machine	A machine learning algorithm that can be applied to a number of tasks.
Qualitative (Classification)	kNN	A non-parametric method for classifying objects based on closest training examples in the feature space.
Qualitative (Classification)	Random Forest	Ensemble method using multiple decision trees to create classifications

General Modeling Strategy

- There is usually an inverse relationship between model flexibility / power and interpretability.
- In the best case, we would like a parsimonious and interpretable model that has excellent performance.
- Unfortunately, that is not realistic.
- A recommended strategy:
  - Start with the most powerful black-box type models.
  - Get a sense of the best possible performance.
  - Then fit more simplistic/ understandable models.
  - Evaluate the performance cost of using a simpler model.

About Me

- Reside in Wayne, Illinois
- Active Semi-Professional Classical Musician (Bassoon).
- Married my wife on 10/10/10 and been together for 10 years.
- Pet Yorkshire Terrier / Toy Poodle named Brunzie.
- Pet Maine Coons' named Maximus Power and Nemesis Gul du Cat.
- Enjoy Cooking, Hiking, Cycling, Kayaking, and Astronomy.
- Self proclaimed Data Nerd and Technology Lover.



Fine

Acknowledgements

- http://en.wikipedia.org/wiki/Regression_analysis
- http://www.ftpress.com/articles/article.aspx?p=2133374
- http://www.matthewrenze.com/presentations/exploratory-data-analysis-with-r.pdf
- http://msenux.redwoods.edu/math/R/TransformingData.php
- https://exploredata.wordpress.com/2014/10/01/reexpressing-salaries/
- http://www.princeton.edu/~otorres/DataPrep101.pdf
- http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_univariate_sect040.htm
- http://www.mathsisfun.com/data/correlation.html
- http://www.mathsisfun.com/data/standard-normal-distribution.html
- http://www.edii.uclm.es/~useR-2013/Tutorials/kuhn/user_caret_2up.pdf