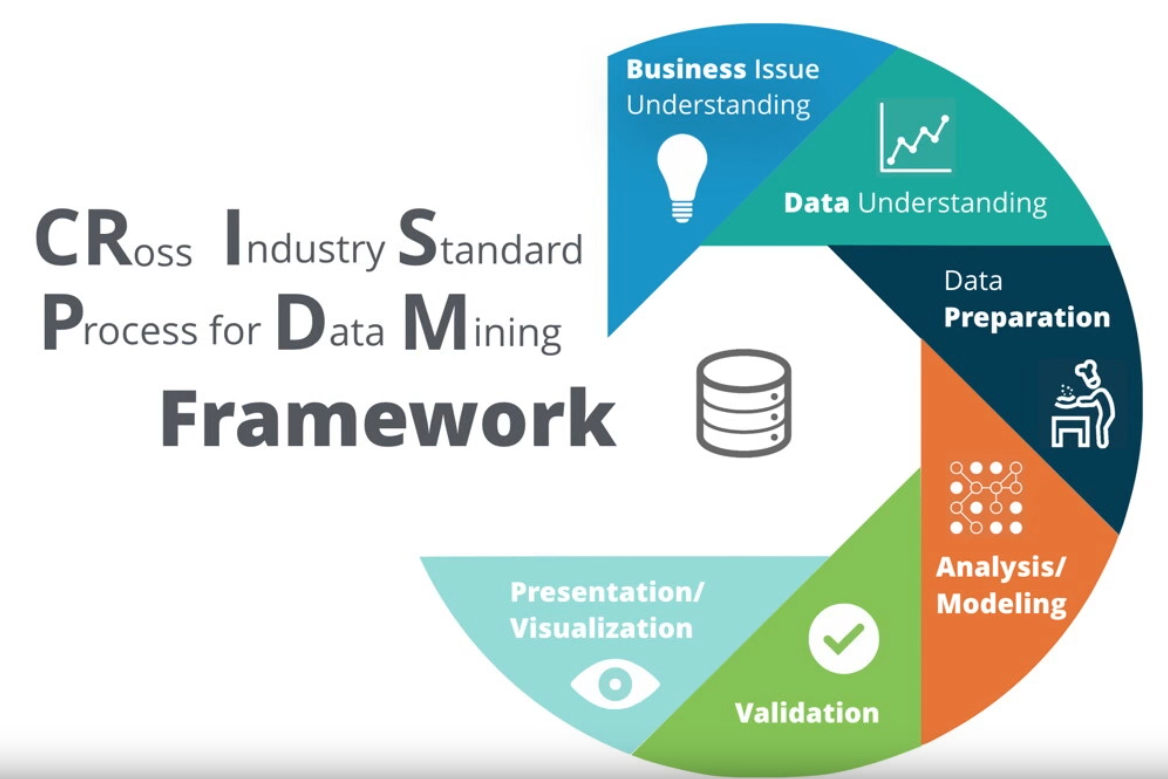
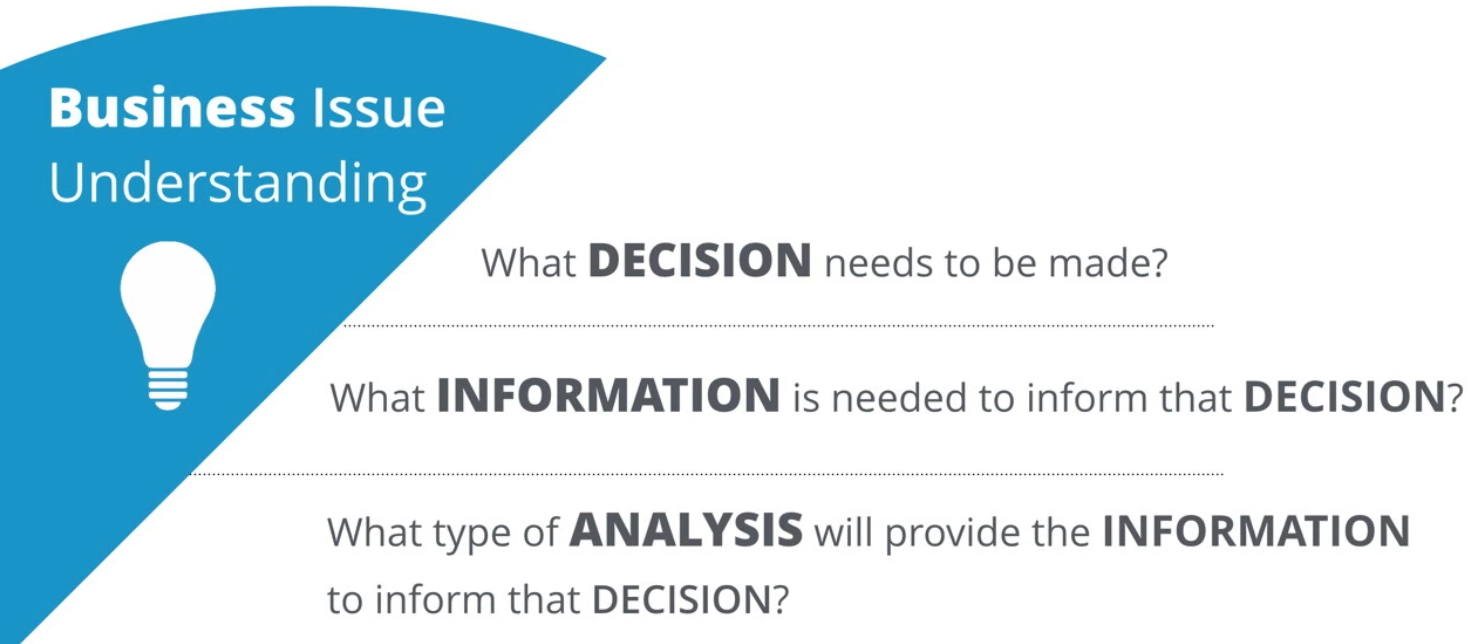
**Problem Solving with Advanced Analytics**

Jocelyn Li

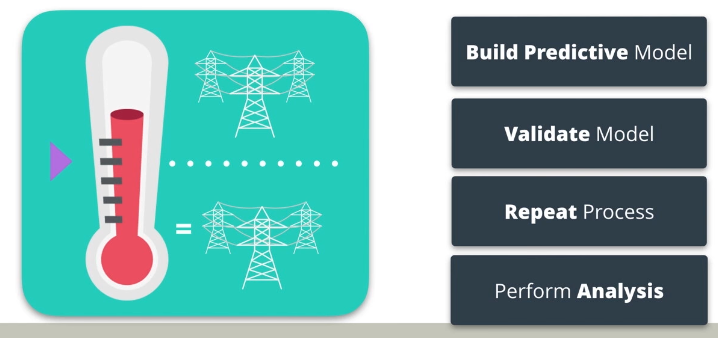
February 2017

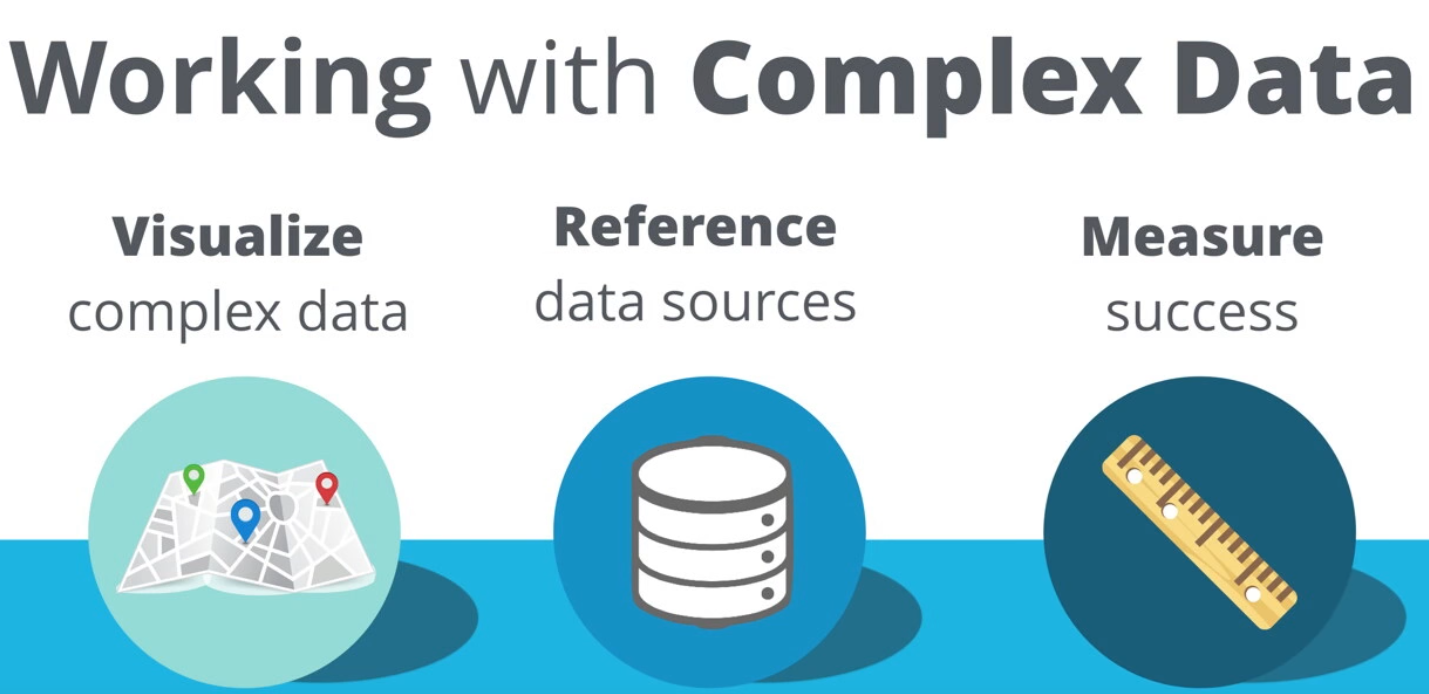
1. **The problem solving framework CRISP-DM**











#### **Data Understanding**

"The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information." – Wikipedia

* What data is needed?
* What data is available?
* What are the important characteristics of the data?

#### **Analysis/Modeling**

"In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed." - Wikipedia

##### Important Steps

* Determine what methodology to use to solve the problem
* Determine the important factors or variables that will help solve the problem
* Build a model to solve the problem
* Run the model and move to the validation phase

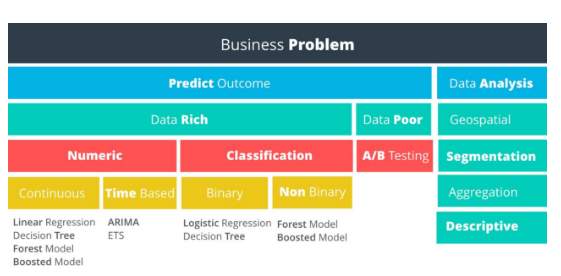
#### **Validation**

"At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached." - Wikipedia

Important Steps

* Observe the key results on the model
* Ensure the results make sense within the content of the business problem
* Determine whether to proceed to the next step or return to a previous phase
* Repeat as many times as necessary

1. **Methodology Map**



1. **Non-Predictive Analysis**

We’ve broken down non-predictive data analysis into four categories:

* Geospatial (**地理空间**)
* Segmentation
* Aggregation
* Descriptive



**Geospatial Analysis**

This type of analysis uses **location based data to help drive your conclusions**. Examples include identifying customers by a geographic region, calculating the distance store locations or creating a trade area based upon customer locations.

**Segmentation Analysis**

Segmentation is the process of **grouping data together**. Groups can be simple, such as customers who have purchased different items, to more complex segmentation techniques where you identify stores that are similar based upon the demographics of their customers.

**Aggregation Analysis**

This methodology simply means calculating a value across a group or dimension and is commonly used in data analysis. For example, you may want to aggregate sales data for a salesperson by month - adding all of the sales closed for each month. Then, you may want to aggregate across dimensions, such as sales by month per sales territory. Aggregation is often done in reporting to be able to "slice and dice" information to help managers make decisions and view performance.

**Descriptive Analysis**

Descriptive statistics provides simple summaries of a data sample. Examples could be calculating average GPA for applicants to a school, or calculating the batting average of a professional baseball player. In our electricity supply scenario, we could use descriptive statistics to calculate the average temperature per hour, per day, or per date. Some of the commonly used descriptive statistics are **Mean, Median, Mode, Standard Deviation, and Interquartile range.**

1. **Predictive Analysis**

**Data Rich vs. Data Poor**

Do you have data on what you are trying to predict? If so, you can proceed down the data rich path, otherwise, the data poor path is your only option. See the following example that demonstrates a data poor scenario.

**Data Poor Business Problems**

**A/B Tests**

If there is not sufficient usable data to solve the problem, then we need to set up an experiment to help us get the data we need. An experiment in a business context is usually referred to as an A/B Test.

**Numeric vs. Non-Numeric Predictive Analysis**

Assuming we have enough data to proceed with the analysis, our next decision is to look at the outcome we’re trying to predict and determine if it’s a numeric outcome or a non-numeric outcome.

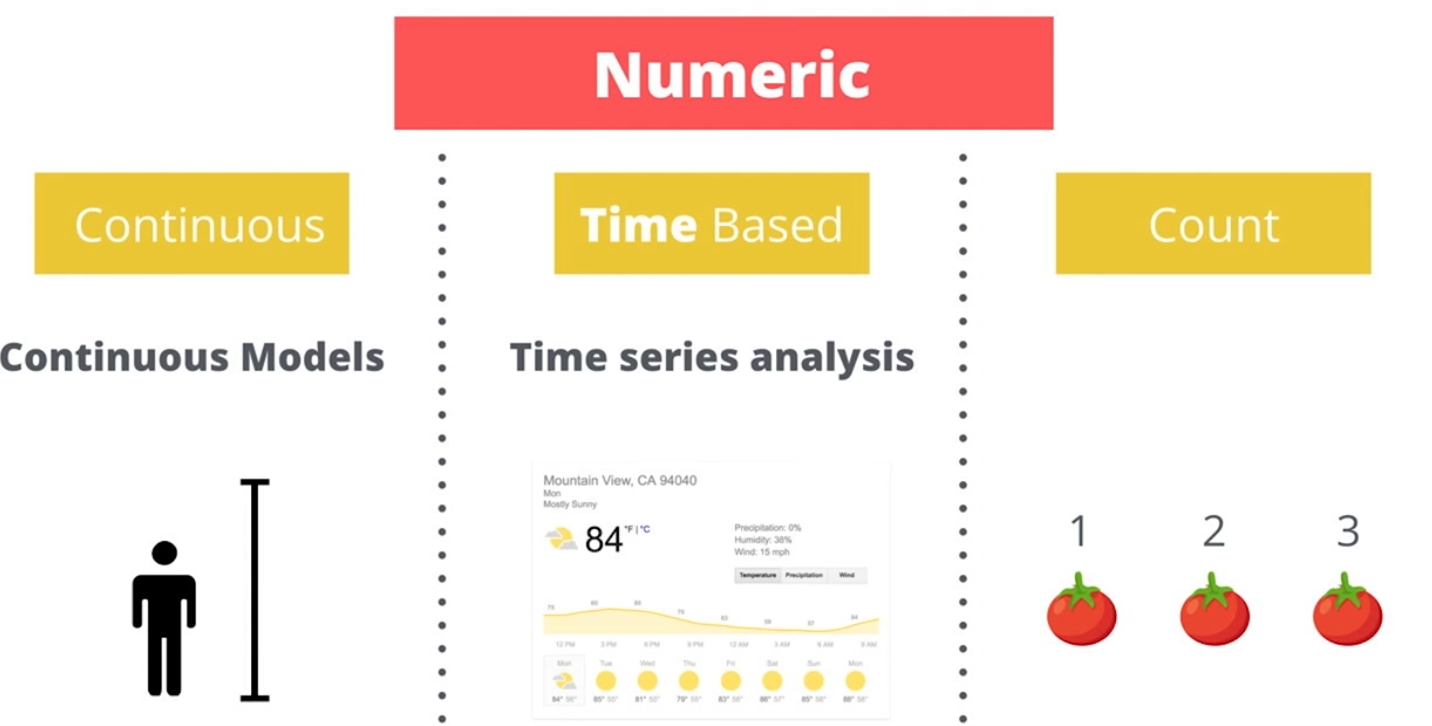
**Regression Models**

Numeric outcomes are those **where the outcome is simply a number**. Predicting the demand for electricity or the hourly temperature are both numeric outcomes. Models predicting numeric data are called regression models.

**Classification Models**

Non-numeric outcomes are those **where we’re trying to predict the category into which a case** (e.g. customer) falls, such as whether a customer will pay on-time, pay late, or default on a payment. Another example is the whether an electronic device will fail before 1000 hours or not. Models predicting non-numeric data are called classification models.

1. **Introduction to Numeric Models**



#### **Target Variables**

Target variables represent the outcome we are trying to predict. In order to select the right predictive model, we first determine whether the target variable is numeric or non-numeric. The type of numeric or non-numeric target variables will then help us select which model is appropriate. Let’s start with numeric variables.

#### **Types of Numeric Variables**

The three most common types of numeric variables are continuous, time-based, and count.

#### **Continuous**

A continuous variable is one that can take on all values in a range. For instance your height can be measured down to many decimal places. We do not grow in even inch intervals. Discrete data is counted, Continuous data is measured

#### **Time-Based**

A time-based numeric variable is one where you are trying to predict what will happen over time. This is often related to forecasting.

#### **Count**

Count variables are numbers that are [**discrete**](https://www.mathsisfun.com/data/data-discrete-continuous.html), positive integers. They’re called count numbers because they’re used to analyze variables that you can count. As modeling these type of variables is not common in business, we won’t be covering this topic in this course.

1. **Introduction to Non-Numeric Models**

#### **Non-Numeric Variables**

A non-numeric variable is often called **categorical**, because the values of the variable take on a discrete number of possible values or categories. Examples include whether an electronic device will fail before 1000 hours or not; whether a customer will pay on-time, pay late, or default on a payment, or whether a store is classified as large, medium or small.

#### **Classification Models: Binary and Non-Binary**

When modeling categorical variables, the number of possible outcomes is an important factor. If there are only two possible categorical outcomes, such as Yes or No, or True or False, then the variable can be described as **Binary**.

If there are more than two possible categorical outcomes, such as small, medium, or large, or pay on-time, pay late, or default on a payment, then the variable can be described as **non-binary**. The important take-away from this lesson is the ability to determine if you should use a classification model, and whether it should be a binary model or a non-binary model.

1. **Approaching the Business Problem**

#### **What decisions need to be made?**

The decision the sales manager needs to make is, “Do we have enough capacity on the support team to handle the support tickets from the new customer?” and “If not, how many people do we need to add to the support team to reach the desired capacity?”

#### **What information do we need to inform this decision?**

We need to calculate the average number of tickets per customer per week. We can then aggregate the average number of tickets for each customer to get a total average number of support tickets that we predict will be submitted per week. Once we have this information, we need to compare the predicted average number of tickets with the current capacity of the support staff, specifically, the average number of tickets each team member can handle.

#### **What type of analysis is needed to get the information needed to make that decision?**

Let’s use our Methodology Map flowchart to help us determine the type of analysis we should use to provide the exact information needed to inform the decision. We want to predict the average number of tickets per week a new customer will submit. Therefore, we’re looking to predict an outcome - that was easy.

1. **Linear Regression**

#### y = mx + b

Y = Target Variable

X = Predictor Variable

m = Slope of the line

b = Y-intercept

##### **Target Variable (dependent variable)**

The target variable is the variable we are trying to understand and predict. It is also referred to as the dependent variable. In our example, we are trying to predict Y, or the average number of tickets.

##### **Predictor Variable (Independent Variable**)

Predictor variables are used to try to predict the target variable and are also known as independent variables. In the example there is just one predictor variable, X, or the number of employees. It is used to predict the number of tickets based.

##### **Validation**

Now that we’ve performed the analysis and run the Linear Regression Model, we need to validate the results of the model. In other words, is there a way to measure how good the model is? Or in this case, is the linear expression we calculated a good fit of our data?

##### **Step 1: Correlation**

Using the correlation function **CORREL(data\_y, data\_x)**, we can calculate the correlation between the target and predictor variable. This value is often referred to as r. The range of r is from -1 to +1. The closer r is to plus or minus 1, the higher the correlation between x and y. In our example, the value of r is 0.987, indicating a strong correlation.

##### **Step 2: Calculate r-squared**

While a strong correlation is good, we really want to know how well the data fits our line. Fortunately, we can get a sense of how good the formula is at approximating the data by calculating the coefficient of determination, or r-squared. R-squared is a coefficient between 0 and 1. **R-squared is interpreted as the percent of variance in observations that is explained by the model, or the explanatory power of the model.** An R-squared value close to 1 would mean that nearly all variance in the target variable is explained by the model. An R-squared value close to 0 would mean that nearly none of the variance in the target variable is explained by the model.

##### **Caution about interpreting R-squared**

How you interpret R-squared depends heavily on the problem you're trying to model and the data you use. For tough problems, a very low R-squared may be acceptable. Also, a high R-squared may result from a poor model. However, in general, the higher the R-squared the better, especially as you add and remove predictor variables to determine the strongest predictive model. To read more about interpreting R-squared, see [**here**](http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit).

**R-squared:** [**http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit**](http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit)

1. **Multiple Linear Regression**

#### **R-Squared vs Adjusted R-Squared**

The adjusted r-squared value should be used with multiple linear regressions due to a phenomenon that occurs when adding additional variables to the model. In a nutshell, the more variables that are included, the higher the r-squared value will be - even if there is no relationship between the additional variables and the target variable. Therefore, we use the Adjusted R-squared value.

You can find more information on Adjusted R-squared [**here**](https://en.wikipedia.org/wiki/Coefficient_of_determination#Adjusted_R2). Note that 'explanatory variables' = 'predictor variables' in the linked article.

1. **Linear Regression using non-numeric predictor variables**

So let’s talk about what happens in linear regression when you add a categorical variable to the mix of predictor variables. Here’s a general regression equation with two predictor variables.

**Y = β 0 + β1X1 + β2 X 2**

Like we discussed, the X's represent the values for each variable. These come directly from the data. The β's come from the linear regression model. β 0 is the intercept. The other β's represent the relationship between the predictor variable X and the target variable Y.

#### **Adding a Categorical Predictor Variable (Transform to dummy variables)**

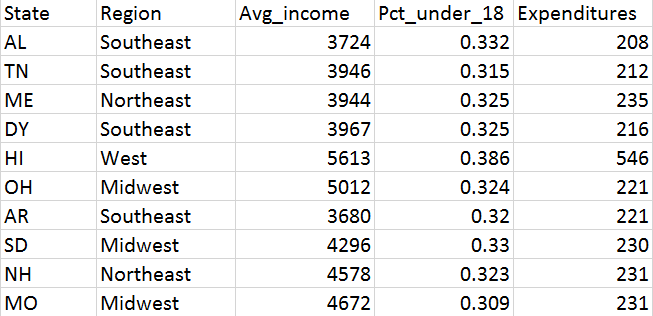
Now let’s say you add a **third variable that is a non-numeric, or categorical variable.** Now putting in the actual value of a category into an equation wouldn’t work because you can't do math string variables, so we have to transform the variable somehow. An inexperienced analyst may simply assign a number to each category and plug it into the model.

**Y = β 0 + β1X1 + β2 X 2 + β3 X 3**

Let's spend a moment exploring the problem with this. In linear regression, the coefficient or slope on each predictor variable represents the relationship between it has with the target variable. So if you transform a category into a numeric variable, you are assuming a linear relationship exists between the target variable and the category number. Since the category number is generally assigned arbitrarily, this doesn't make sense.

#### **Transforming Categorical Variables - Bad Example**

Let’s look at an example. We have a dataset with some information for each of the 50 states in the United States.



We want to predict the per pupil student expenditures in a state. To do so, we'll use three predictor variables: average income of that state, the % of the population under 18, and the region in which the state falls. Below is the regression equation that would result after building the model. Region is a categorical variable with four values: west, midwest, northeast, and southeast. You assign them the numbers 1, 2, 3, and 4 respectively, and run the model. The results give you the following equation:

**Expenditures = -530 + 0.073 Avg\_Income + 1406.36 Pct\_Under\_18 + 6.53 region**

Now let's briefly try and understand the coefficients:

* The coefficient on average income implies that for every one dollar of additional average income, that state spends 7.3 more cents per pupil.
* The coefficient on percent under 18 implies that for every one additional percentage point of the population under 18, the state spends about $14 more per pupil.
* The coefficient on region implies that for every increase in one region, the state spends about $6.52 more per pupil. Now this one doesn't make logical sense. There was no logical order in the numbers, and this format doesn't allow us to take into account enough of the variation among regions.

#### **Transforming Categorical Variables - Good Example**

A much better way of using categorical variables in regression is to use what are called dummy variables. A dummy variable can only take on two values, generally zero or one. You would add one dummy variable for one less than the number of unique values in the categorical variable. So if the variable is binary, you'd add one dummy. If there are four categories, you'd add three dummy variables.

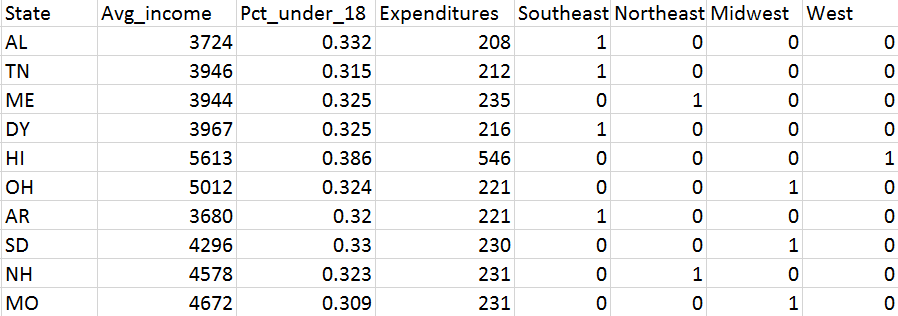
Going back to our example, let's now use dummy variables to represent the categorical variable, region. To represent the four categories west, midwest, northeast, and southeast, you’ll need to add three dummy variables. Let’s create one for southeast, northeast, and west.

**Expenditures = β 0 + β1 Avg\_Income + β2 Pct\_Under\_18 + β3 midwest + β4 southeast + β5 west**

Each of the variables takes on a value of either 1 or 0. If the state is in the southeast, then the value for the southeast variable would be 1 while the other two variables would be zero.

Now we didn't create a variable for northeast. That’s because the equation needs a baseline value that is not coded into a dummy variable. If a state is in the northeast, then the value for all three of the dummy variables would be zero. You always create one less dummy variable *than the number of categories* to make sure that one category is represented by zero values for the dummy variables. That one category, in this case the northeast region, becomes the category that others are compared to.

Note: Many software tools, like Alteryx, will transform categorical variables into dummy variables automatically. If you were to do it manually, the data would now look like this:



1. **Interpreting Regression Results**

Below you see the results of a linear regression. Results are similarly reported by almost any regression tool. Do not be intimidated by the complexity of the presentation of the results; we’ll walk you through the most important values and how to interpret and apply them. For purposes of this exercise, the focus will be on three values that are of particular importance: the coefficient estimates, the p-values, and the R-squared.

Note: [**Here**](https://d17h27t6h515a5.cloudfront.net/topher/2017/February/589b742a_interpreting-regression-results/interpreting-regression-results.pdf) is a PDF version this concept that you can download and use later. You can also find the PDF at the bottom of the page.

* When a P value is less than or equal to the significance level, you reject the null hypothesis. P values to determine statistical significance in a [hypothesis test](http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/what-is-a-hypothesis-test/).
* The **significance level**, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.