**Note: Data Analysis with Python**

**Week 1:**

1. Defining Problem Statement
   1. Setting goals and objectives
   2. Analytical Approach
2. Data Requirements & Data Collection

**Week 2:**

**Data Wrangling:**

Data wrangling is the process of converting data from the initial format to a format that may be better for analysis.

1. Identify and handle missing values
2. Data Formatting
3. **Data Normalization**: (centering/scaling):
   1. Data normalization can be defined as a process designed to facilitate a more cohesive form of data entry, essentially ‘cleaning’ the data. When you normalize a data set, you are reorganizing it to remove any unstructured or redundant data to enable a superior, more logical means of storing that data.
   2. The main goal of data normalization is to achieve a standardized data format across your entire system. This allows the data to be queried and analyzed more easily, which can lead to better business decisions.
   3. Examples
      1. Miss ANNA will be written Ms. Anna
      2. 4158488400 will be written 415-848-8400
      3. 37 buttercup AVE will be written 37 Buttercup Avenue
      4. Amazon will be written Amazon.com, Inc.
      5. VP product will be written Vice President of Product
      6. Price range should be between 5000 – 10000. Over/under these values, column won’t store the values.
4. **Data Binning :** Data binning, also called data discrete binning or data bucketing, is a data pre-processing technique used to reduce the effects of minor observation errors. The original data values which fall into a given small interval, a bin, are replaced by a value representative of that interval, often a central value.
5. **Turning Categorical values into numeric variables**

**Data Normalization Methods/Techniques:**

This technique is very important to understand the data in data pre-processing process.

When it comes to build a model that predict the outcome of variables, if needed, the data normalization is very important and it can be achieved through several techniques such as;

1. **Simple Feature Scaling**

In this method, simply divide each value by maximum value of the feature and formula would be:

**Xnew = Xold / Xmax**

1. **Min-Max Method**

In this method, we simply divide the difference of Xold and Xmin to the difference of Xmax and Xmin.

**Xnew = (Xold – Xmin) / (Xmax – Xmin)**

1. **Z-score or Standard Score**

In this method, we divide the difference of Xold and Xavg with the standard deviation of the feature variable.

**Xnew = (Xold – Xavg) / Xstd**

**Or** ( **Note**: The resulting values hover around zero, and typically range between -3 to +3 but can be higher or lower.)

**Xnew = (Xold – Xavg) / standard deviation**

**Note:** All these methods result in the providing the Xnew value between 0 to 1.

**Data Binning:**

**Data Standardization**

Data is usually collected from different agencies in different formats. (Data standardization is also a term for a particular type of data normalization where we subtract the mean and divide by the standard deviation.)

**What is standardization?**

Standardization is the process of transforming data into a common format, allowing the researcher to make the meaningful comparison.

**Example**

Transform mpg to L/100km:

In our dataset, the fuel consumption columns "city-mpg" and "highway-mpg" are represented by mpg (miles per gallon) unit. Assume we are developing an application in a country that accepts the fuel consumption with L/100km standard.

We will need to apply **data transformation** to transform mpg into L/100km.

## Data Normalization

**Why normalization?**

Normalization is the process of transforming values of several variables into a similar range. Typical normalizations include scaling the variable so the variable average is 0, scaling the variable so the variance is 1, or scaling the variable so the variable values range from 0 to 1.

**Example**

To demonstrate normalization, let's say we want to scale the columns "length", "width" and "height".

**Target:** would like to normalize those variables so their value ranges from 0 to 1

**Approach:** replace original value by (original value)/(maximum value)

**Binning**

**Why binning?**

Binning is a process of transforming continuous numerical variables into discrete categorical 'bins' for grouped analysis.

**Example:**

In our dataset, "horsepower" is a real valued variable ranging from 48 to 288 and it has 59 unique values. What if we only care about the price difference between cars with high horsepower, medium horsepower, and little horsepower (3 types)? Can we rearrange them into three ‘bins' to simplify analysis?

We will use the pandas method 'cut' to segment the 'horsepower' column into 3 bins.

**Indicator Variable (or Dummy Variable)**

**What is an indicator variable?**

An indicator variable (or dummy variable) is a numerical variable used to label categories. They are called 'dummies' because the numbers themselves don't have inherent meaning.

**Why we use indicator variables?**

We use indicator variables so we can use categorical variables for regression analysis in the later modules.

**Example**

We see the column "fuel-type" has two unique values: "gas" or "diesel". Regression doesn't understand words, only numbers. To use this attribute in regression analysis, we convert "fuel-type" to indicator variables.

We will use pandas' method 'get\_dummies' to assign numerical values to different categories of fuel type.

**Week 3:**

**Chi-Square Test:**

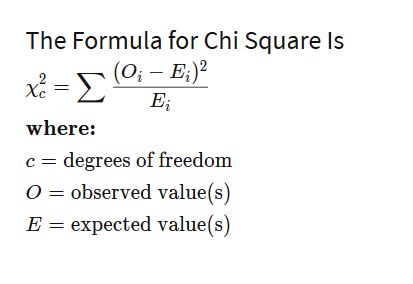
The Chi-square test is intended to test how likely it is that an observed distribution is due to chance. It measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

The Chi-square tests a null hypothesis that the variables are independent. The test compares the observed data to the values that the model expects if the data was distributed in different categories by chance.

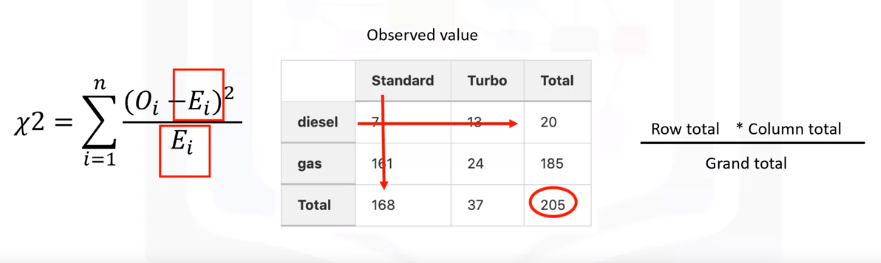
Anytime the observed data doesn't fit within the model of the expected values, the probability that the variables are dependent becomes stronger, thus proving the null hypothesis incorrect.

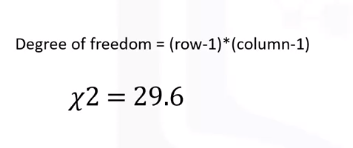
The Chi-square does not tell you the type of relationship that exists between both variables, but only that a relationship exists.

A crosstab is a table showing the relationship between two or more variables. When the table only shows the relationship between two categorical variables, the crosstab is also known as a contingency table.



How to calculate expected values for Chi-square test:





**If p <0.05, then we reject the NULL Hypothesis, meaning that two variables are independent to each other and conclude that there is evidence of association between these two variables.**

## Continuous Numerical Variables:

Continuous numerical variables are variables that may contain any value within some range. They can be of type "int64" or "float64". A great way to visualize these variables is by using scatterplots with fitted lines.

In order to start understanding the (linear) relationship between an individual variable and the price, we can use "regplot" which plots the scatterplot plus the fitted regression line for the data.

### **ANOVA: Analysis of Variance**

The Analysis of Variance (ANOVA) is a statistical method used to test whether there are significant differences between the means of two or more groups. ANOVA returns two parameters:

**F-test score**: ANOVA assumes the means of all groups are the same, calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A larger score means there is a larger difference between the means.

**P-value**: P-value tells how statistically significant our calculated score value is.

If our price variable is strongly correlated with the variable we are analyzing, we expect ANOVA to return a sizeable F-test score and a small p-value.

**Pearson Correlation Co-efficient and P-Value:**

**P-value**

What is this P-value? The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant.

By convention, when the

**p-value is <0.001:** we say there is strong evidence that the correlation is significant

**the p-value is < 0.05:** there is moderate evidence that the correlation is significant.

**the p-value is < 0.1:** there is weak evidence that the correlation is significant.

**the p-value is > 0.1:** there is no evidence that the correlation is significant.

We can obtain this information using "stats" module in the "scipy" library.

**Week 4:**

**Model Development:**

A model or estimator can be thought of as a mathematical equation used to predict the value given one or more other values. Relating one or more independent variables or features to dependent variables.

Usually, the more relevant data you have, the more accurate your model is.

Simplest form of regression models are :   
  
Simple Linear Regression Model

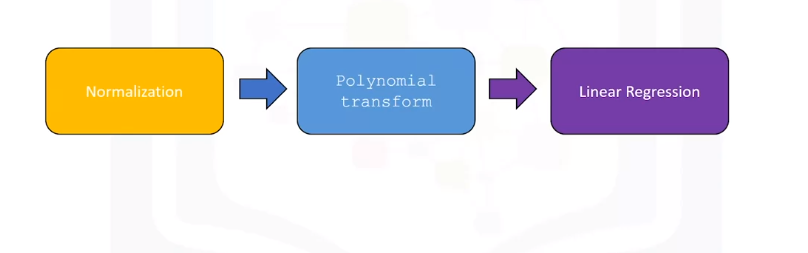
Multiple Linear Regression Model

Polynomial Regressions Model

**Polynomial Regression:**

Polynomial regression is a special case of general linear regression. For this type of regression, we transform our data into a polynomial, then use linear regression to fit the parameters.

There are many steps involved when building an accurate machine learning model such as ;



**Determining a Good Model Fit:**

How can we determine if the predictive model is correct?

When building a best-fit predictive model, do consider the following points;

1. Make sure the predicted values make sense
2. Do the visualization of results such as visualization of residual errors, normal distribution curve or bell-curved visualization, visualization of actual values, predicted values, and mean along with residuals and such.
3. Make sure to create a scatter plot along with best-fitted line which will help in determining the best fit model.
4. Comparing between different models

**SLOPE inf Regression Model:**

In a regression context, the slope is the heart and soul of the equation because it tells you how much you can expect Y to change as X increases. ***In general, the units for slope are the units of the Y variable per units of the X variable. It's a ratio of change in Y per change in X.***

* a refers to the **intercept** of the regression line, in other words: the value of Y when X is 0
* b refers to the **slope** of the regression line, in other words: the value with which Y changes when X increases by 1 unit

For example: Predicting the price(dependent variable/target variable/response variable or predicated values) of a car based on highway-mpg(independent variable/predictor variable/explanatory variable).

Price = 38423.31 – 821.75 \* highway-mpg (Simple linear model equation)

How to interpret the above equation:

This value (821.75) corresponds to the multiple of the highway miles per gallon feature. As such, an increase of one unit in highway miles per gallon, the value of the car decreases approximately 821 dollars. This value also seems reasonable.

***Simply: one unit of increase in X (explanatory variable) responds to unit increase or decrease in Y.***

**Note:** The mean square error is perhaps the most intuitive numerical measure for determining if a model is good or not.

**R-Squared:**

It tells you how well your line fits into the model. R - squared values range from 0 - 1. R - squared tells us what percent of the variability in the dependent variable is accounted for by the regression on the independent variable.

An R - squared of 1 means that all movements of dependent variable are completely explained by movements in the independent variables.

# Lesson Summary

In this lesson, you have learned how to:

**Define the explanatory variable and the response variable:** Define the response variable (y) as the focus of the experiment and the explanatory variable (x) as a variable used to explain the change of the response variable. Understand the differences between Simple Linear Regression because it concerns the study of only one explanatory variable and Multiple Linear Regression because it concerns the study of two or more explanatory variables.

**Evaluate the model using Visualization:** By visually representing the errors of a variable using scatterplots and interpreting the results of the model.

**Identify alternative regression approaches:** Use a Polynomial Regression when the Linear regression does not capture the curvilinear relationship between variables and how to pick the optimal order to use in a model.

**Interpret the R-square and the Mean Square Error:** Interpret R-square (x 100) as the percentage of the variation in the response variable y  that is explained by the variation in explanatory variable(s) x. The Mean Squared Error tells you how close a regression line is to a set of points. It does this by taking the average distances from the actual points to the predicted points and squaring them.