**Exploratory Data Analysis**

This is another preliminary step for Machine Learning problems. Exploratory Data Analysis (EDA) is not only used in Machine Learning, but also in Data Analysis business cases.

**Why Exploratory Data Analysis (EDA)?**

👉Summarizes main characteristics of the data

👉Gains better understanding of the data

👉Uncovers relationships between variables

👉Extracts important variables

Let’s say we want to answer question given below.

*“What are the characteristics that have the most impact on car price?”*

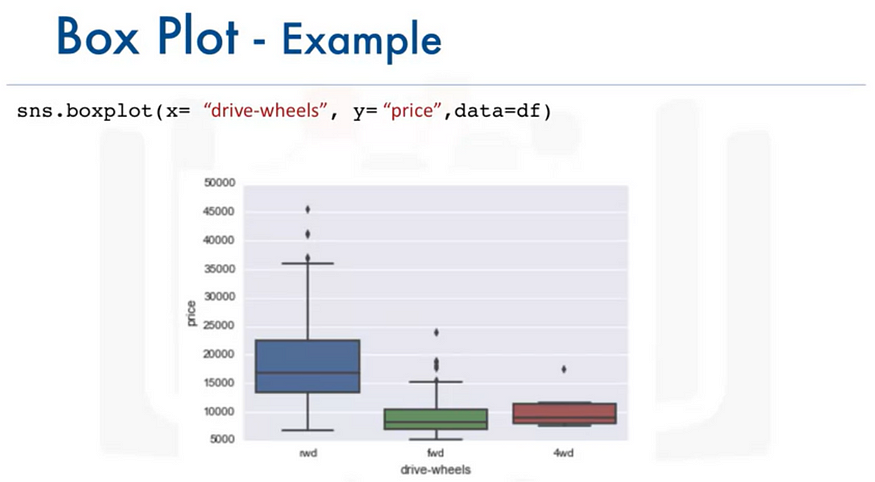
EDA will help answering those questions.

**Descriptive Statistics**

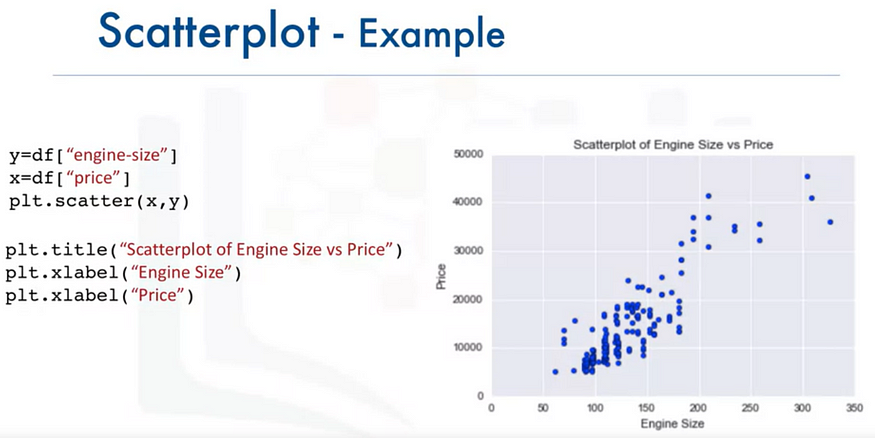
It describes basic features of a data-set and obtains short summary of the sample and measures of the data.

**Summarize Categorical features**— using value\_counts()

**Visualize and Compare Continuous variable feature with Categorical feature** — Use Boxplots. With Boxplots, you can easily spot outliers and also see the distribution and skewness of the data. In below image, we see that price for the rear wheel drive (rwd) is quite different from other categories. While price for front wheel drive (fwd) and 4 wheel drive (4wd) are almost same.



**Visualize and Compare Continuous variable feature with another Continuous variable feature**— Use Scatterplot. Each observation in scatterplot is a point. It shows relationship between two variables, generally between target/dependent variable and predictor/independent variable. In below image, we see the “Price” variable on y-axis is a target variable and “Engine Size” on x-axis is a predictor variable. As the engine size increases, price also go higher which indicates there is a positive relationship between these two variables.

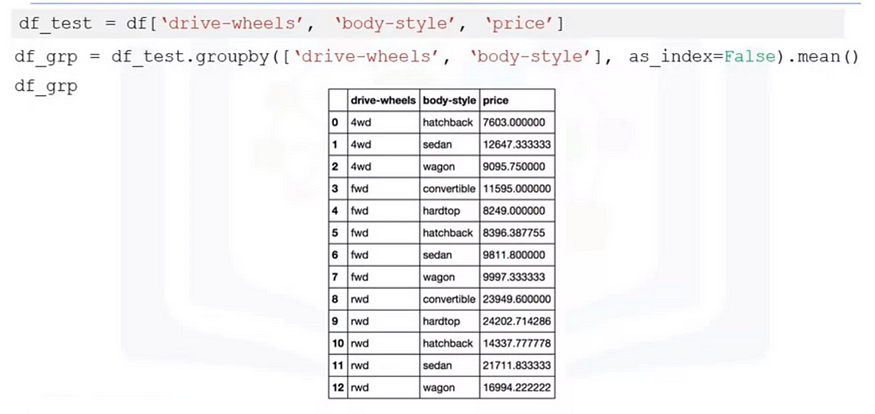


**GroupBy**

It is used to group data by different features and hence, it discovers the important categorical features based on their relationship with a numeric feature.

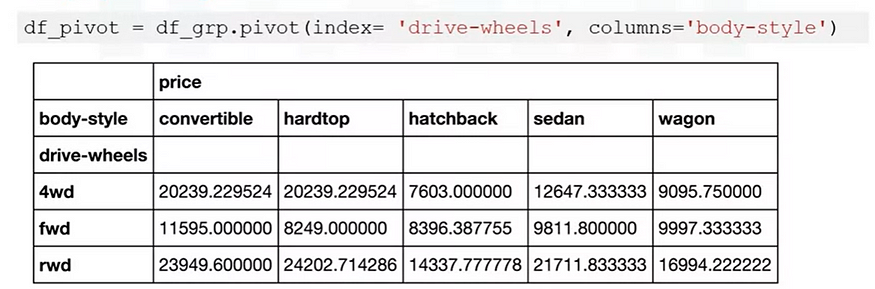
1. Can be applied on categorical variables
2. Groups data into categories
3. Can be applied on single or multiple variables

In below image, cars are grouped by their drive-wheels and body-style. We see that “rear wheel drive, convertibles” and “rear wheel drive, hardtops” have the highest price while “4 wheel drive, hatchbacks” have the lowest price.



**Pivot tables**

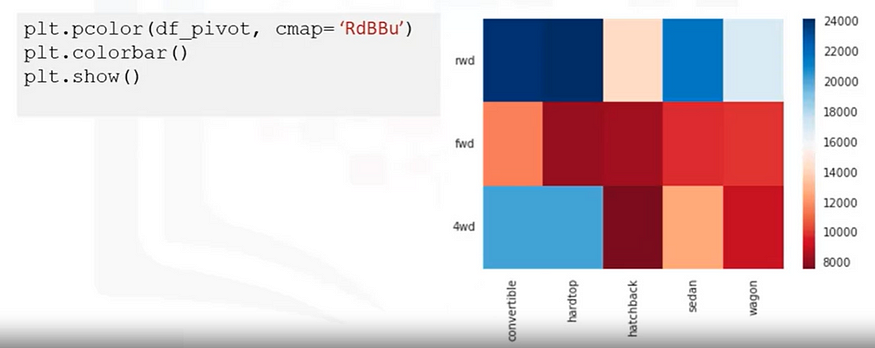
Data can be arranged in a pivot table for better readability. It does this by arranging one variable on column and another variable along the rows as shown below.



**Heatmap**

It is very important visualization technique for showing relationship between multiple variables.

Below image shows the previous pivot table in Red-Blue color scheme. We can clearly see “rear wheel drive, convertibles” and “rear wheel drive, hardtops” have the highest price and are shown in dark blue color while “4 wheel drive, hatchbacks” have the lowest price and is shown dark red color.



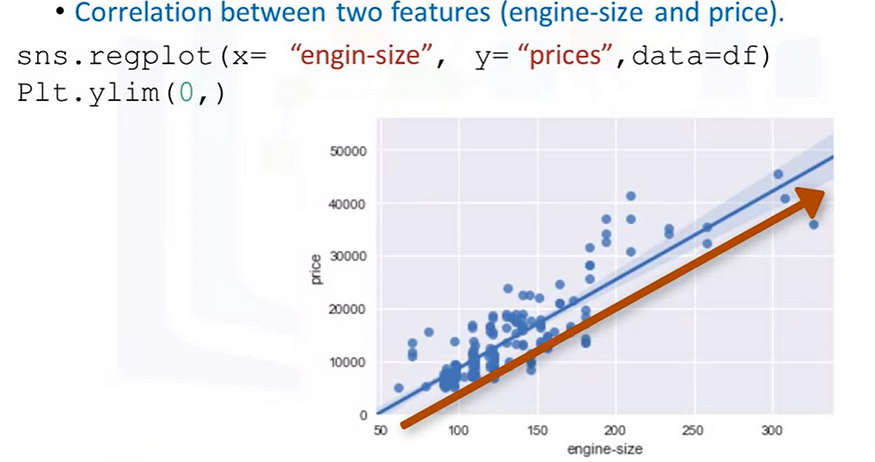
**Correlation**

It is a statistical metric to measure what extent different variables are interdependent. In short, if one variable changes, how does it affect other variable.

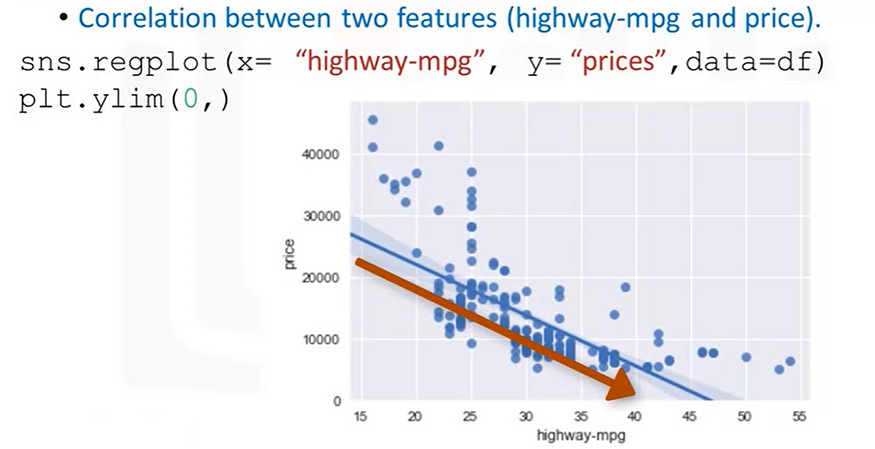
Examples:

1. There is correlation between smoking and lung cancer. More you smoke, higher are the chances of getting lung cancer.
2. There is correlation between rain and umbrella. More it rains, higher are the chances of people using umbrella.

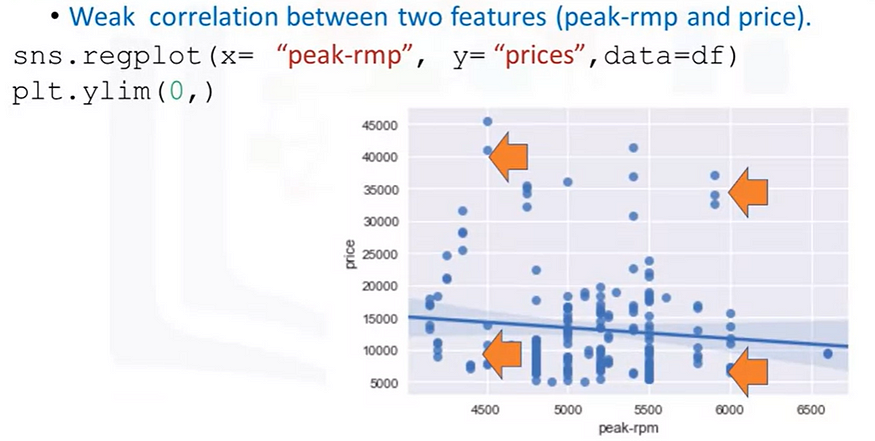
**Positive Correlation**: In below image, we see a straight regression line along data points which shows positive relationship or correlation between two variables engine-size and price. As engine-size increases, price goes up. We can say engine-size is a good predictor of price.



**Negative Correlation**: In below image, we see a straight regression line along data points which shows negative relationship or correlation between two variables highway-miles-per-gas and price. As highway-mpg increases, price goes down. We can say highway-mpg is a good predictor of price.



**Weak Correlation**: Sometimes, there is neither positive nor negative correlation. In below image, we see that both low and high peak-rpm values have high prices and they have low prices as well. We can say peak-rpm is NOT a good predictor of price.



👉**Correlation does not imply Causation**

Even though smoking and lung cancer are correlated, we cannot say that lung cancer is only caused by smoking. Or umbrellas are only being used because of rain.

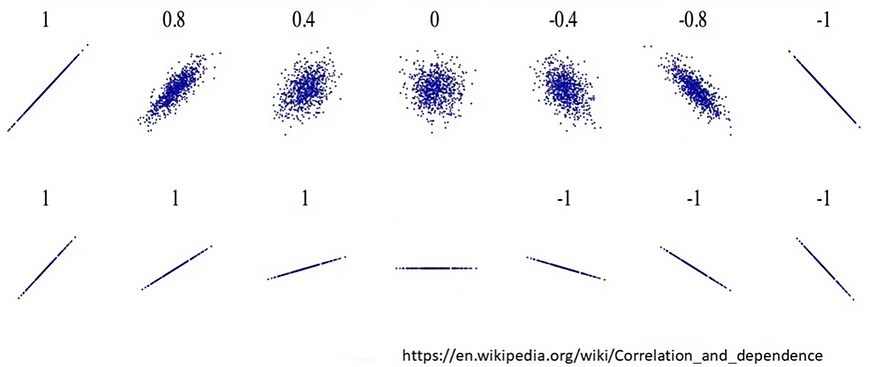
**Correlation Coefficient and P-Value**

The two most commonly used statistical tests for establishing relationship between variables are correlation coefficient and p-value. Correlation coefficient is a way to test, how strong is the relationship between two variables, whereas p-value tells us if the result of an experiment is statistically significant.

1. **Correlation Coefficient**: There are several types of correlation coefficients (e.g. Pearson, Kendall, Spearman), but the most commonly used is the Pearson’s correlation coefficient. Assigns value -1 to +1. Value close to or equal to +1 represents strong positive relation between two features. Value close to or equal to -1 represents strong negative relation between two features. While value close to or equal to 0 represents very weak or no relationship at all.
2. **P-value**: It tells how certain we are about the correlation that we calculated. P-value evaluates how well your data rejects the [null hypothesis](http://www.statisticshowto.com/probability-and-statistics/null-hypothesis/), which states that there is no relationship between two compared groups. Successfully rejecting this hypothesis tells you that your results may be statistically significant. This is done by specifying a significance cutoff, known as the [alpha value](http://blog.minitab.com/blog/michelle-paret/alphas-p-values-confidence-intervals-oh-my). Alpha is usually set to 0.05, meaning the probability of achieving the same or more extreme results assuming the null hypothesis is 5%. If the p-value is less than the specified alpha value, then we reject the null hypothesis.

* value < 0.001 implies **very strong** certainty in the result
* value < 0.05 implies **strong** certainty in the result
* value < 0.1 implies **weak**certainty in the result
* value > 0.1 implies **no** certainty at all in the result

Below image shows the data points for different Correlation coefficients.

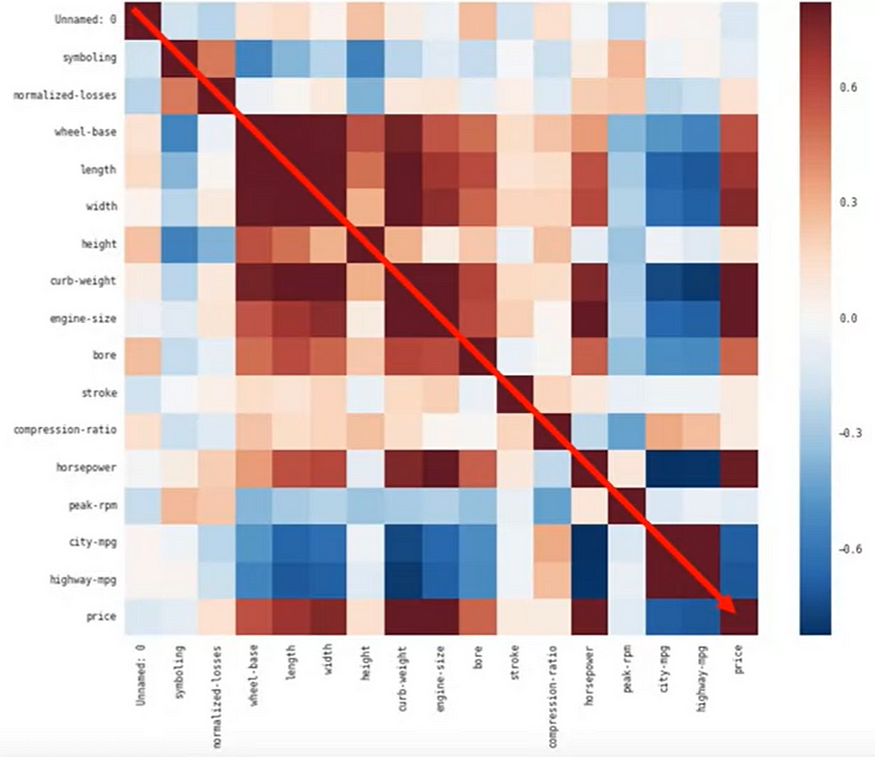


In below example, we see that pearson correlation coefficient is 0.8, that is closer to 1 and p-value is much smaller than 0.001. So, we can say that horsepower and price of car are strongly correlated and the results are statistically significant.



**Correlation heat-maps**

We can visualize the correlation between all features of a dataframe using correlation heatmap. The color scheme indicates the Pearson Correlation Coefficient indicating the strength of relationships between two features. We can ignore the diagonal blocks as it represents relationship of a feature with itself and by default, it will be 1.

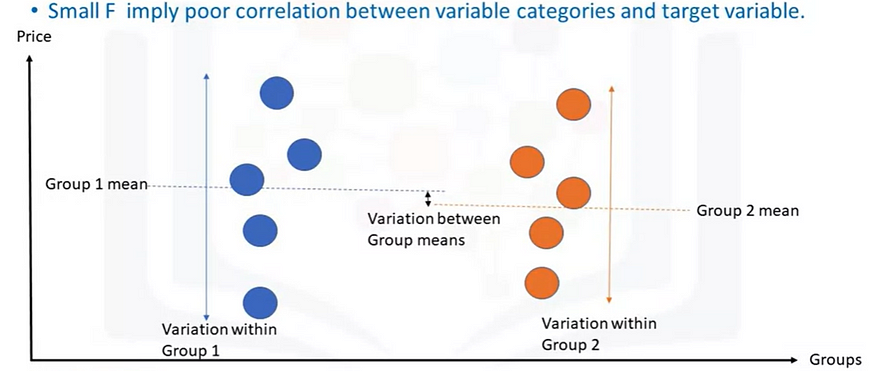


**ANOVA — The Analysis of Variance**

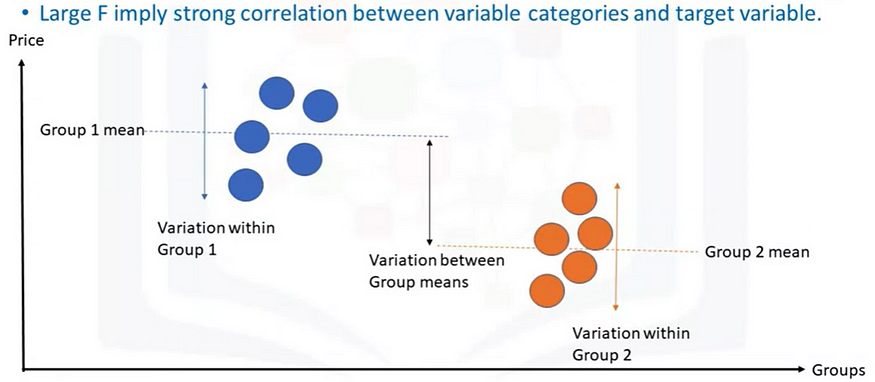
A statistical method in which the variation in a set of observations is divided into distinct components

**F-test score**: Variation between sample group means/averages divided by variation within each of the sample group.

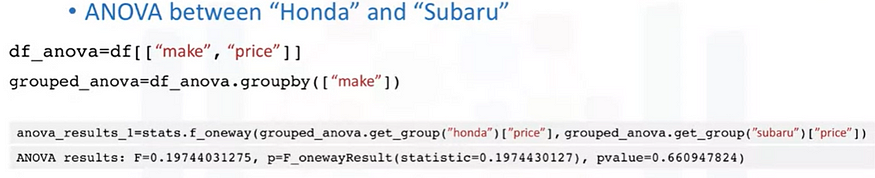
Below image shows that variation within Group 1 and Group 2 are much larger than variation between group means. So, F-score will be very low and imply that there is a poor correlation between variable categories (Group 1 and Group 2) and target variable.



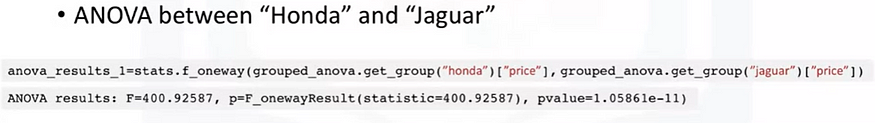
Below image shows that variation within Group 1 and Group 2 are smaller than variation between group means. So, F-score will be very high and imply that there is a strong correlation between variable categories (Group 1 and Group 2) and target variable.



Let’s see an example below where Group 1 is “Honda” car maker and Group 2 is “Subaru”. Here, F-score is 0.19 which is quite low and p-value is also greater than 0.5.



However, in below example where Group 1 is “Honda” car maker and Group 2 is “Jaguar”. Here, F-score is 0.19 which is quite high and p-value is also less than 0.001.



So, ANOVA test predicts that there will be huge difference in price if car model changes from “Honda” to “Jaguar” and vice-versa. However, there will not be enough change in price if car model changes from “Honda” to “Subaru” and vice-versa.