Exploratory Data Analysis or “EDA" is a critical step in a data science life cycle to get the model accuracy and it probably takes good part of project time. Here are the main reasons to use EDA:

* Detection of mistakes
* Checking of assumptions
* Preliminary selection of appropriate models
* Determining relationships among the explanatory features, and
* Assessing the direction and rough size of relationships between explanatory and outcome features.

Any method of looking at data that does not include formal statistical modelling and inference falls under the term exploratory data analysis.

Quality of your inputs decide the quality of your output.

To improve the model accuracy, EDA may be repeated after the model tuning.

The data generally collected into a rectangular array (e.g., spreadsheet or database), most commonly with one row per each scenario of subject and one column for each subject identifier, outcome feature, and explanatory feature. Each column contains the numeric values for a continuous feature or the levels for a categorical feature. (Some more complicated experiments require a more complex data layout.)

Usually we are not very good at looking at a column of numbers or a whole spreadsheet and then determining important characteristics of the data. We find looking at numbers to be tedious or overwhelming. EDA techniques have been devised as an aid in this situation.

Most of these techniques work in part by hiding certain aspects of the data while making other aspects more clear.

EDA is generally cross-classified in two ways:

1. Each method is either non-graphical or graphical.
2. Each method is either univariate or multivariate (usually just bivariate).

Non-graphical methods generally involve calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way.

Univariate methods look at one feature (data column) at a time, while multivariate methods look at two or more features at a time to explore relationships. Usually multivariate EDA will be bivariate (looking at exactly two features), but occasionally it will involve three or more features. It is usually a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

Beyond the categories created by the above cross-classification, each of the categories of EDA have further divisions based on the role (outcome or explanatory) and type (categorical or continuous) of the feature(s).

Below are the techniques involved to understand, clean and prepare your data for building your predictive model:

**Feature Identification**

1. Identify Predictor (Input) and Target (output) features.
2. Identify the data type and category of the features

Data Types

* Numbers: int, float
* Complex: Complex number types, Complex numbers are written in the form x + yj, where x is the real part and y is the imaginary part
* bool: Boolean (true/false) types
* time: Date/time types
* string: string types
* List: List once created can be modified. Ex:[1,2,3]
* Tuple: Tuple once created cannot be modified. Ex:(1,2,3)
* Set: unordered collection of unique items. Ex:{5,1,3,4,6}
* Dictionary: unordered collection of key-value pairs. Ex:{‘key’:2,’value’:1}
* Conversion between data types: we can convert between different data types using int(), float(), str(), set(), tuple(), dict(), list()

Features Category:

* Categorical
* Continuous

Code file reference: DataExploration.py

**Let us discuss the four types of EDA**:

1. Univariate non-graphical
2. Multivariate non-graphical
3. Univariate graphical
4. Multivariate graphical

**Univariate non-graphical**

The goal is to look into the “sample distribution” of a single feature to make some tentative conclusions about how the target column is compatible with that feature. We also find missing and outlier (noise data) values. Outlier data is nothing but data point situated far away from rest of data points. Ex: Sale Price with an abnormal value.

Code file reference: Outlier.py

Initially we find type of feature whether categorical or continuous.

**Categorical data**:

We need to find the range of values and the frequency (or relative frequency) of occurrence for each value.

Only useful univariate non-graphical technique for categorical features is some form of tabulation of the frequencies, usually along with calculation of the percent of data that falls in each category.

For example if we categorize subjects feature by College as Maths, Physics, Chemistry and Social, and there is a large number of students enrolled in the year 2017. If we take a random sample of 20 students for the purpose of performing a memory test, we could list the sample values as Maths, Maths, Physics, Social, Social, Chemistry, Physics, Social, Maths, Physics, Chemistry, Chemistry, Social, Physics, Physics, Maths, Physics, Social, Maths, Chemistry. Our EDA would look like this:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistic** | **Maths** | **Physics** | **Chemistry** | **Social** | **Total** |
| Count | 5 | 6 | 4 | 5 | 20 |
| Proportion | 0.25 | 0.3 | 0.2 | 0.25 | 1 |
| percent | 25% | 30% | 20% | 25% | 100% |

Note that it is useful to have the total count (frequency) to verify that we have an observation for each subject that we recruited. (Losing data is a common mistake, and EDA is very helpful for finding mistakes). In addition, we should expect that the proportions add up to 1.00 (or 100%) if we are calculating them correctly.

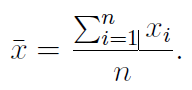
**Continuous data**:

The characteristics of the distribution of a continuous feature are its centre, spread, modality (number of peaks), shape and outliers.

We need to recognize that this would be different each time we get the data, due to selection of a different random sample.

We need to understand the central tendency and spread of the feature. The central tendency or “location" of a distribution has to do with typical or middle values. The common, useful measures of central tendency are the statistics called (arithmetic) mean, median, and mode.

Assuming that we have n data values labelled x1 through xn, the formula for calculating the sample (arithmetic) **mean** is



The arithmetic mean is simply the sum of all of the data values divided by the number of values. It can be thought of as how much each subject gets in a fair re-division of whatever the data are measuring.

The **median** is another measure of central tendency. The sample median is the middle value after all of the values are put in an ordered list. If there are an even number of values, take the average of the two middle values.

For symmetric distributions, the mean and the median coincide. For unsymmetrical distributions, the median is preferred as a measure of central tendency.

The median has a very special property called robustness. A sample statistic is robust if moving some data tends not to change the value of the statistic. The median is highly robust, because you can move nearly the entire upper half and/or lower half of the data values any distance away from the median without changing the median. More practically, a few very high values or very low values usually have no effect on the median.

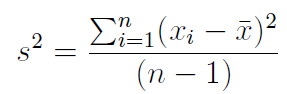
A rarely used measure of central tendency is the **mode**, which is the most likely or frequently occurring value. More commonly, we simply use the term mode when describing whether a distribution has a single peak (unimodal) or two or more peaks (bimodal or multi-modal).

The most common measure of central tendency is the mean. For unsymmetrical distribution or when there is concern about outliers, the median may be preferred.

Code file reference: Imputer.py

Several statistics are commonly used as a measure of the spread of a distribution, including **variance**, **standard deviation**, and **interquartile range**. Spread is an indicator of how far away from the center we are still likely to find data values.

The sample formula for the **variance** of observed data conventionally has n-1 in the denominator instead of n to achieve the property of unbiasedness.

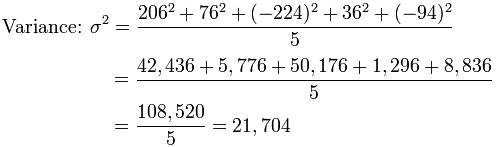


Which is essentially the average of the squared deviations, except for dividing by n-1 instead of n. This is a measure of spread, because the bigger the deviations from the mean, the bigger the variance gets. (In most cases, squaring is better than taking the absolute value because it puts special emphasis on highly deviant values.)

Sample values 600 470 170 430 300

Mean = 600 + 470 + 170 + 430 + 3005  =  19705  =  394

600-394 = 206 similarly rest



So the Variance is 21,704

|  |  |
| --- | --- |
| The **standard deviation** is simply the square root of the variance. Therefore, it has the same units as the original data, which helps make it more interpretable. The sample standard deviation is usually represented by the symbol s or σ. | |
|  |  |
| σ | = √21,704 |
|  | = 147.32... |
|  | = 147 |

A third measure of spread is the **interquartile range**. To define IQR, we first need to define the concepts of quartiles. The quartiles of a sample are the three values, which divide the distribution or observed data into even fourths. So one quarter of the data fall below the first quartile, usually written Q1; one-half fall below the second quartile (Q2); and three fourths fall below the third quartile (Q3). As half of the values fall above Q2, one quarter fall above Q3, and also that Q2 is a synonym for the median. Once the quartiles are defined, it is easy to define the IQR as IQR = Q3 - Q1.

By definition, half of the values (and specifically the middle half) fall within an interval whose width equals the IQR. If the data are more spread out, then the IQR tends to increase, and vice versa.

The IQR is a more robust measure of spread than the variance or standard deviation. Any number of values in the top or bottom quarters of the data can be moved any distance from the median without a affecting the IQR at all. More practically, a few extreme outliers have little or no effect on the IQR.

**Univariate graphical**

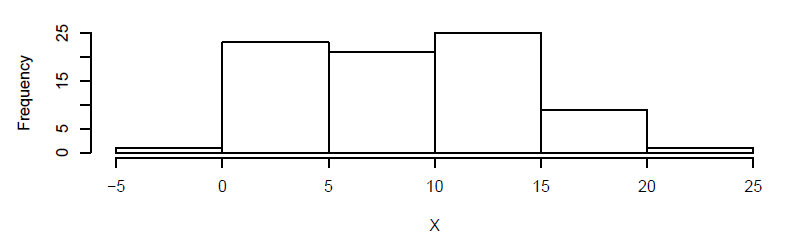
If we are focusing on data from observation of a single feature on n subjects, i.e., a sample of size n, then in addition to looking at the various sample statistics discussed in the previous section, we also need to look graphically at the distribution of the sample. Non-graphical and graphical methods complement each other. While the non-graphical methods are quantitative and objective, they do not give a full picture of the data; therefore, graphical methods, which are qualitative and involve a degree of subjective analysis, are also required.

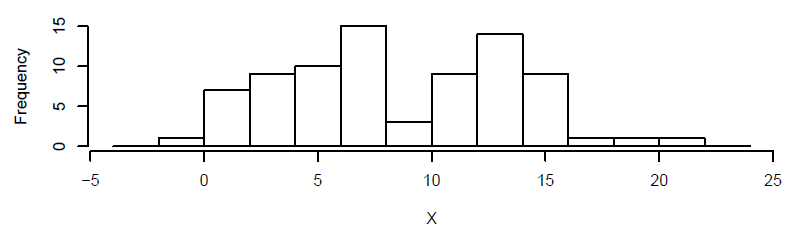
**Histograms**

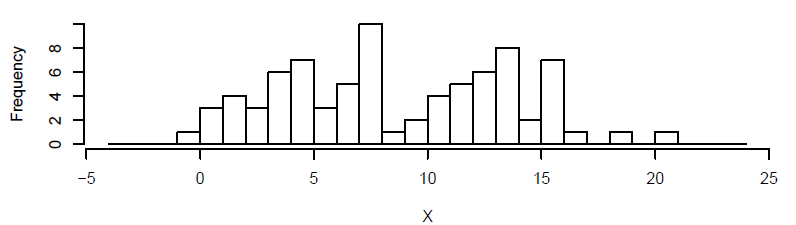
The most basic graph is the histogram, which is a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values. Typically, the bars run vertically with the count (or proportion) axis running vertically. To manually construct a histogram, define the range of data for each bar (called a bin), count how many cases fall in each bin, and draw the bars high enough to indicate the count.

Generally, we will choose between about 5 and 30 bins, depending on the amount of data and the shape of the distribution. Of course, we need to see the histogram to know the shape of the distribution, so this may be an iterative process. It is often worthwhile to try a few different bin sizes/numbers because, especially with small samples, there may sometimes be a different shape to the histogram when the bin size changes. Nevertheless, usually the difference is small.

Below figure shows three histograms of the same sample from a bimodal population using three different bin widths (5, 2 and 1). The top panel appears to show a unimodal distribution. The middle panel correctly shows the bimodality. The bottom panel incorrectly suggests many modes.







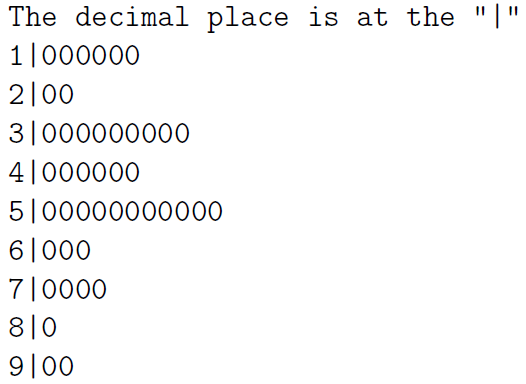
With practice, histograms are one of the best ways to quickly learn a lot about data, including central tendency, spread, modality, shape and outliers.

Code file reference: Histograms.py

**Stem-and-leaf plots**

A simple substitute for a histogram is a stem and leaf plot. A stem and leaf plot is sometimes easier to make by hand than a histogram, and it tends not to hide any information. Nevertheless, a histogram is generally considered better for appreciating the shape of a sample distribution than is the stem and leaf plot.

Here is a stem and leaf plot for a sample data:



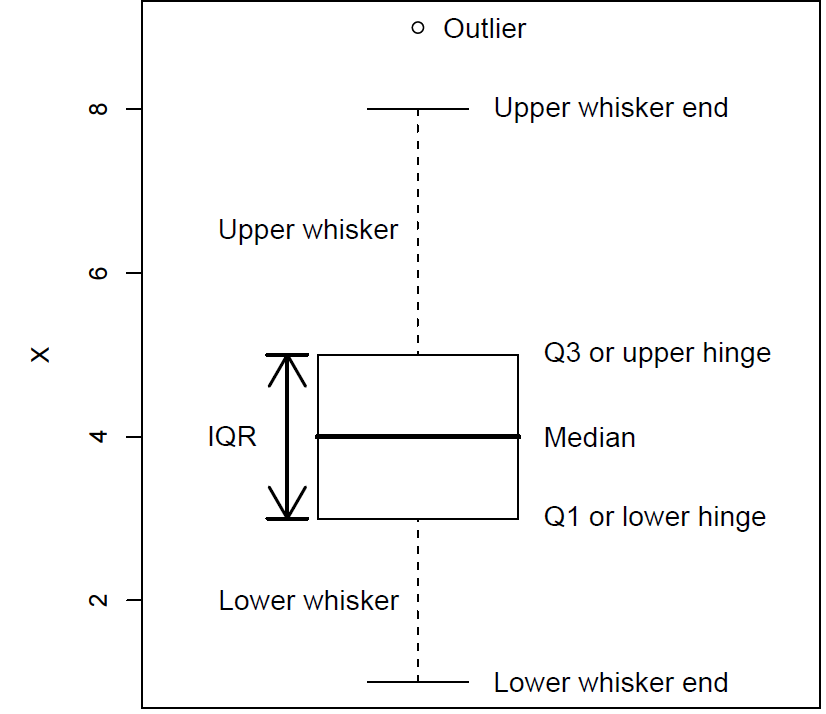
Because this particular stem and leaf plot has the decimal place at the stem, each of the 0's in the first line represent 1.0, and each zero in the second line represents 2.0, etc. So, we can see that there are six 1's, two 2's etc. in our data.

**Boxplots**

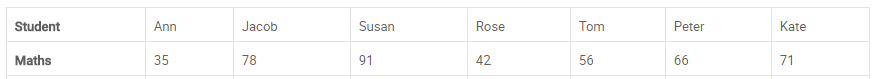
Another very useful univariate graphical technique is the boxplot. Boxplots are very good at presenting information about the central tendency, symmetry and skew, as well as outliers.

Here you can see that the boxplot consists of a rectangular box bounded above and below by hinges that represent the quartiles Q3 and Q1 respectively, and with a horizontal median line through it. You can also see the upper and lower whiskers, and a point marking an outlier. The vertical axis is in the units of the continuous feature.

* Lower whisker end: the minimum number in the data set
* Upper whisker end: the maximum number in the data set
* Median: If data set is arranged in ascending order, it is the middle number
* Lower hinge (Q1): If data set is arranged in ascending order, the 25% of data is below it
* Upper hinge (Q3): If data set is arranged in ascending order, the 75% of data is below it



Example:



First, we shall arrange the marks in ascending order, Scores in Maths: 35, 42, 56, 66, 71, 78, 91

Median – 66

The First (Lower) hinge is the midpoint of the lower half of our data.

Lower half of scores in Maths (in Bold): **35**, **42**, **56**, 66, 71, 78, 91

Q1 – 42

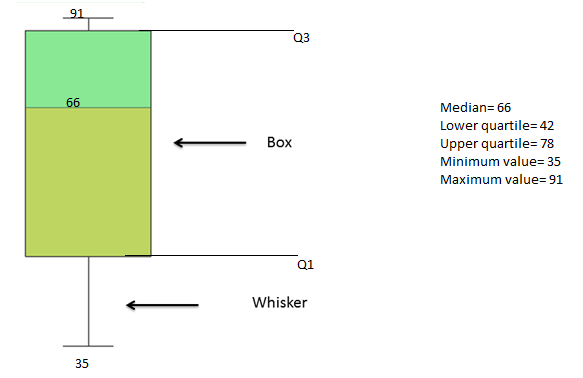
The Third (Upper) hinge is the midpoint of the upper half of our data.

Upper half of scores in Maths (in Bold): 35, 42, 56, 66, **71**, **78**, **91**

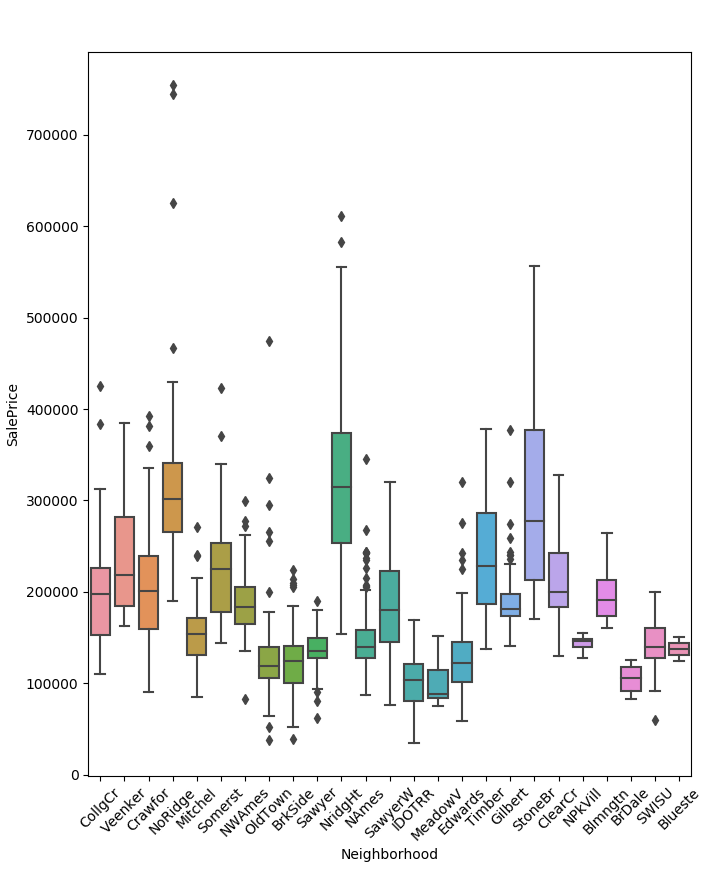
Q3 – 78

Lower whisker end value – 35,

Upper whisker end value – 91



Any data points that lies out of lower and upper whisker end are considered as outliers.



The boxplot is useful because, with practice, all of the above and more can be appreciated at a quick glance. The additional things you should notice on the plot are the symmetry of the distribution. Symmetry is appreciated by noticing if the median is in the center of the box and if the whiskers are the same length as each other. For this purpose, as usual, the smaller the dataset the more variability you will see from sample to sample, particularly for the whiskers. In a skewed distribution, we expect to see the median pushed in the direction of the shorter whisker. If the longer whisker is the top one, then the distribution is positively skewed (or skewed to the right, because higher values are on the right in a histogram). If the lower whisker is longer, the distribution is negatively skewed (or left skewed.) In cases where the median is closer to the longer whisker, it is hard to draw a conclusion.

Boxplots are excellent EDA plots because they rely on robust statistics like median and IQR rather than more sensitive ones such as mean and standard deviation.

Code file reference: Boxplot.py

**Multivariate non-graphical**

Multivariate non-graphical EDA techniques generally show the relationship between two or more features in the form of either cross-tabulation or statistics.

**Categorical data**:

For categorical data (and continuous data with only a few different values), an extension of tabulation called cross-tabulation is very useful.

For two variables, cross-tabulation is performed by making a two-way table with column headings that match the levels of one-variable and row headings that match the levels of the other variable.

Here is an example of a cross-tabulation. For each subject we observe sex and age as categorical variables.

|  |  |  |
| --- | --- | --- |
| **Subject ID** | **Age Group** | **Sex** |
| GW | young | F |
| JA | middle | F |
| TJ | young | M |
| JMA | young | M |
| JMO | middle | F |
| JQA | old | F |
| AJ | old | F |
| MVB | young | M |
| WHH | old | F |
| JT | young | F |
| JKP | middle | M |

|  |  |  |  |
| --- | --- | --- | --- |
| **Age Group/Sex** | **Female** | **Male** | **Total** |
| young | 2 | 3 | 5 |
| middle | 2 | 1 | 3 |
| old | 3 | 0 | 3 |
| Total | 7 | 4 | 11 |

**Chi-Square Test**: The chi-square test is for testing null hypothesis.

A null hypothesis is a hypothesis that says there is no statistical relationship between the two categorical features means they are independent features. It is usually the hypothesis a researcher or experimenter will try to disprove.

Significance level: We can choose significance levels from 0.01 to 0.10; but any value between 0 and 1 can be used.

Before we can proceed, we need to know degrees of freedom.

Degrees of freedom = (number of rows minus one) x (number of columns minus one)

DF = (r - 1) \* (c - 1)

Probability of 0: It indicates that both categorical variables are dependent

Probability of 1: It shows that both categorical variables are independent.

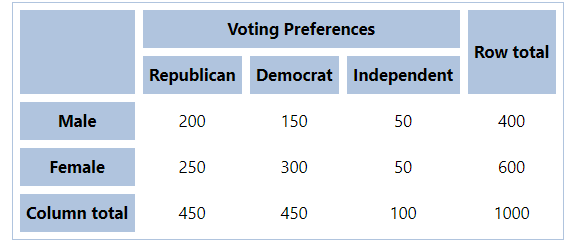
It is found by

\\v-ramprasadm\Shared\CHI SQUARE.bmp

* DF is the degrees of freedom
* r is the number of levels of gender Ex: 2
* c is the number of levels of the voting preference Ex: 3
* nr is the number of observations from level r of gender Ex: 400, 600
* nc is the number of observations from level c of voting preference Ex: 450, 450, 100
* n is the number of observations in the sample Ex: 1000
* Er,c is the expected frequency count where gender is level r and voting preference is level c
* Or,c is the observed frequency count where gender is level r voting preference is level c.

Sample Analysis

A public opinion poll surveyed a simple random sample of 1000 voters. Voters were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table/Two-way table below.



Here we need to know whether men's voting preferences differ significantly from the women's preferences.

Let us use 0.05 as significance level.

The solution to this problem takes four steps:

(1) State the hypotheses

The first step is to state the [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an alternative hypothesis.

A null hypothesis is a hypothesis that says there is no statistical significance between the two variables. It is usually the hypothesis a researcher or experimenter will try to disprove or discredit. An alternative hypothesis is one that states there is a statistically significant relationship between two variables.

(2) Formulate an analysis plan

For this analysis, let’s take significance level as 0.05.

(3) Analyse sample data

Applying the chi-square test for independence to sample data

DF = (r - 1) \* (c - 1) = (2 - 1) \* (3 - 1) = 2

Er,c = (nr \* nc) / n  
E1,1 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,2 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,3 = (400 \* 100) / 1000 = 40000/1000 = 40  
E2,1 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,2 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,3 = (600 \* 100) / 1000 = 60000/1000 = 60

Χ2 = Σ [ (Or,c - Er,c)2 / Er,c ]   
Χ2 = (200 - 180)2/180 + (150 - 180)2/180 + (50 - 40)2/40  
    + (250 - 270)2/270 + (300 - 270)2/270 + (50 - 60)2/60  
Χ2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60  
Χ2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2

(4) Interpret results.

From the Chi-square Distribution Table

For the degree of freedom 2 our p value is less than 0.05. Therefore, we can conclude these two categorical features are dependent features.

**Continuous data**:

For two continuous features, the basic statistics of interest are the sample covariance and/or sample correlation, the sample covariance is a measure of how much two variables co-vary, i.e., how much (and in what direction) should we expect one variable to change when the other changes.

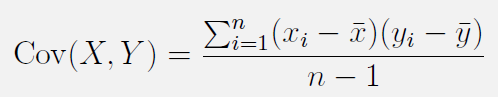
Positive covariance values suggest that when one measurement is above the mean the other will probably also be above the mean, and vice versa.

Negative covariance suggest that when one variable is above its mean, the other is below its mean.

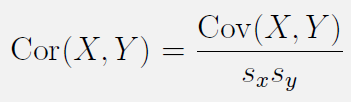
And covariance near zero suggest that the two variables vary independently of each other.

Covariance tend to be hard to interpret, so we often use correlation instead. The correlation has the nice property that it is always between -1 and +1, with -1 being a perfect negative linear correlation, +1 being a perfect positive linear correlation and 0 indicating that X and Y are uncorrelated.

Sample covariance:



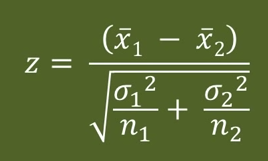
Sample correlation:



Where Sx is the standard deviation of X and Sy is the standard deviation of Y.

**Categorical and Continuous data**:

**Z-Test: This test tells whether mean of two groups are statistically different from each other or not. Here we take a qualitative (categorical) and quantitative (continuous) features. It is found by:**



**Where**

**equation and equation are the means of the two samples**

**σ 1 and σ 2 are the standard deviations of the two samples**

**n1 and n2 are the sizes of the two samples**

**Multivariate graphical**

**Categorical and Continuous data**:

When we have one categorical (usually explanatory) and one continuous (usually outcome) feature, graphical EDA usually takes the form of conditioning on the categorical random feature. This simply indicates that we focus on all of the subjects with a particular level of the categorical random feature, then make plots of the continuous feature for those subjects. We repeat this for each level of the categorical feature, then compare the plots. The most commonly used of these are side-by-side boxplots

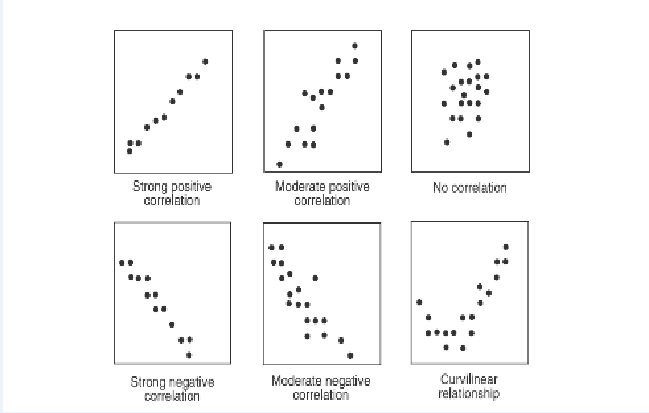
**Continuous and Continuous data**:

Scatter plot is a way to find out the relationship between two continuous features.

The relationship can be linear or non-linear.

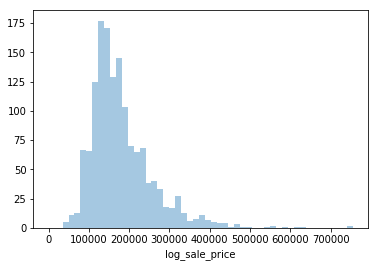
This is good plot for data visualization as we can see data points as dots.

Scatter plot:



Code file reference: scattercorr.py, Scatterplot.py

Dist plot: Visualization of continuous columns with log values



Code file reference: Joint&DistPlots.py

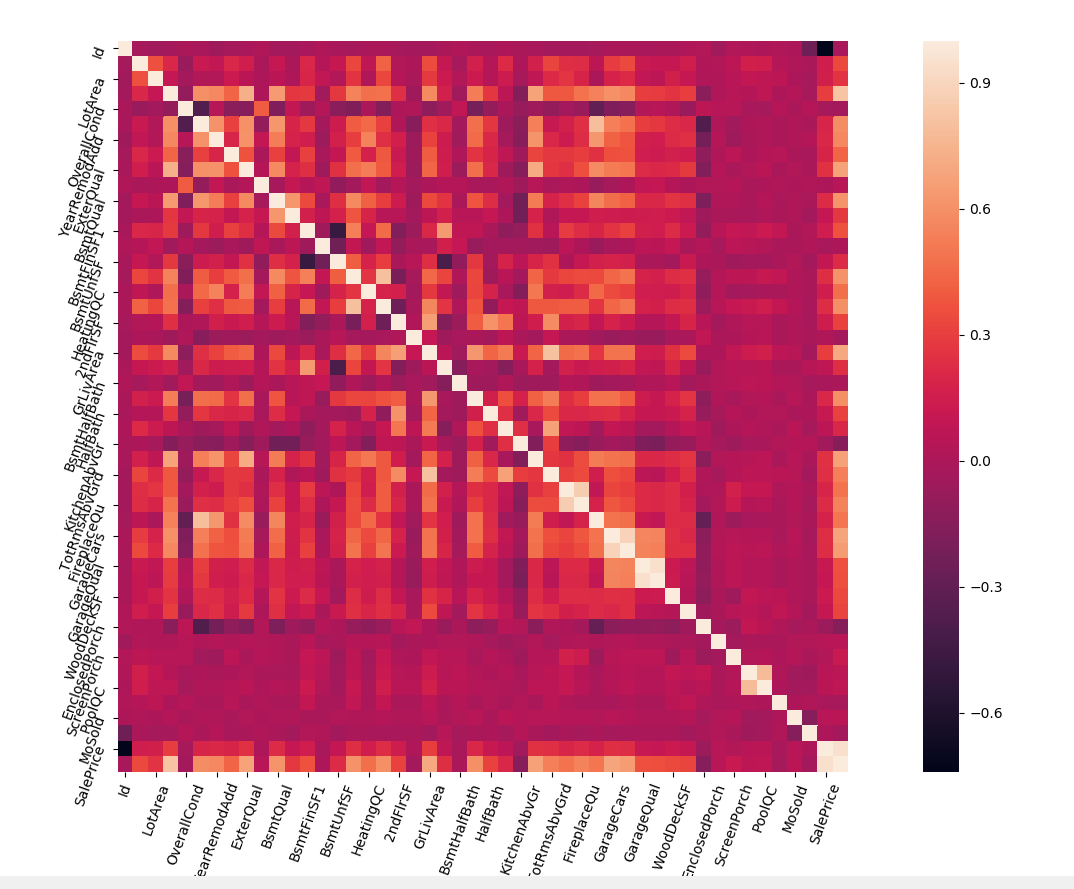
**Heat Map**

1. Heat map is used to understand the correlation between features
2. Heat map can be applied only on number column
3. Correlation boundary is from -1 to +1
4. -0.5 to +0.5 correlation columns are considered

We can use heat map, LASSO, Ridge to find the highly correlated features. We need to use Zscore to convert the values if not present in same units.

**Boundary: -1 to 0 to +1**

We consider features with the range from **– 0.5 to + 0.5**



**Heat Map**

Code file reference: HeatMap.py

Suppose there are five columns/features, we need to find each column relation to rest of columns, So number of relations are 5\*5 = 25

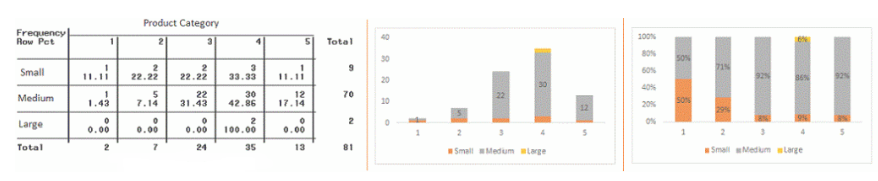
Ex: In house prediction lot area, lot size, sale price as lot area increases then lot size increases , sale price increases.

So, inorder to consider both lot area and lot size, we can have one column out of them to predict sale price, rather than considering columns driving towards same, consider unique columns that helps us predict ouput.(consider less correlated columns)...

in heat map diagram dark color is less correlated. consider less correlated columns than more correlated columns for prediction

**Categorical and Categorical data**:

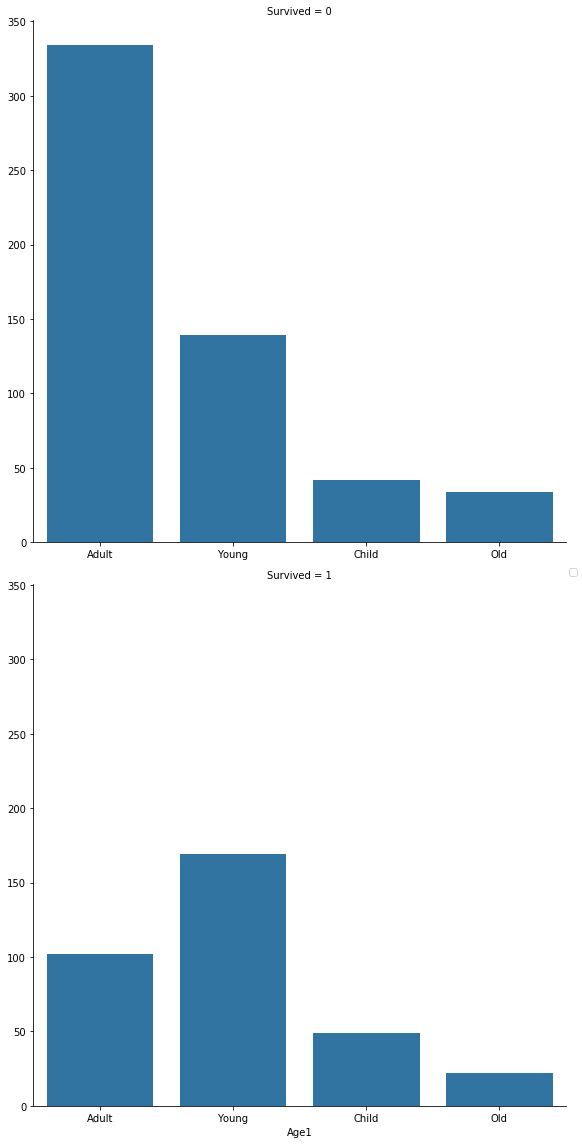
Stacked Column Chart is used to visualise graphically.



Code file reference: Stacked.py

**Continuous & Categorical:**

**Facet Grid:** It is used to draw plots with multiple Axes where each Axes shows the same relationship conditioned on different levels of some variable. It’s possible to condition on up to three variables by assigning variables to the rows and columns of the grid and using different colors for the plot elements. Facetgrid we use for number/numeric/continuous data. Hue property is used on Y axis estimator



**Facet Grid**

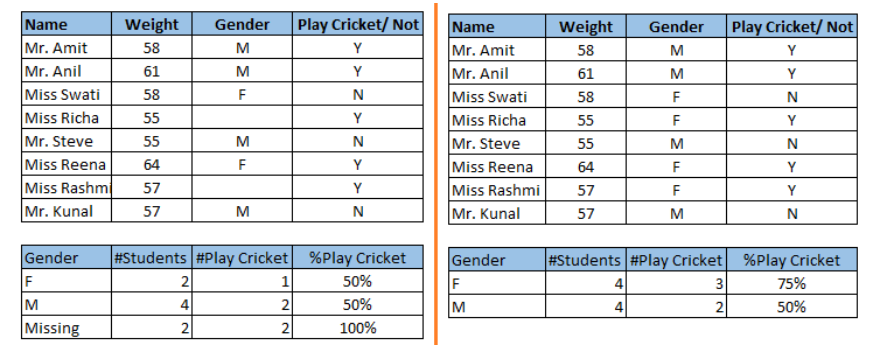
Factorplot we use for categorical data (extracted and converted column)

Extracted means getting first name, last name from title in titanic data

Converted means age split into groups in titanic data.

Code file reference: Facet&FactorPlots.py, factor.py

**Missing value treatment**

****Missing data can reduce the accuracy of model and can also lead to a biased model.

From the above dataset:

On the left side, we have not treated missing values and as a result playing cricket by males is higher than females.

But after treatment of missing values (based on gender) on the right side, we can see that females have higher chances of playing cricket compared to males.

Ways to treat missing values

**Deletion:** There are two types

* **List Wise Deletion**
* **Pair Wise Deletion**

**List wise Deletion:** In list wise deletion, we delete rows where any of the variable is missing.

**Pair Wise Deletion:** In Pair wise deletion, we ignore missing values.

Code file reference: listwise.py

**Mean/Mode/Median Imputation**

We can also use Mean/Median/Mode to fill the missing data.