**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is an approach of analysing or investigation of the dataset using statistical graphics and other data visualization methods. Analysis may include:

* handle missing values,
* maximize insight into a data set and discover patterns,
* extract important variables,
* detect outliers and anomalies,
* find interesting relations among the variables,
* test hypothesis,
* check assumptions,

Drawing reliable conclusions from a massive quantity of data by just gleaning over it is very difficult or almost impossible—instead, you have to look at it carefully through an analytical lens.

**Steps to do EDA**

1. Handle Missing values
   1. Check for Missing values
   2. Replace it with mean, media or mode method whatever is relevant to the problem
      1. I used MODE to fill longitude and latitude.
      2. Our dataset belongs to weather. Assuming, not much variation in the weather in small interval of time, we are using previous values to fill the NULL values
2. Handle outliers
   1. Check the distribution of data.
      1. Use histograms to see the distribution
   2. Delete the values(outliers) which does not lie with in distribution.
      1. If you see any outliers, you can delete those if dropping does not impact much on the total count of instances.
   3. Use box plot to check outliers.
   4. It is up to you whether you want to keep or remove outliers. If you have any domain expert, you can talk to them before taking any decision.
      1. If, for example, 30% of your data is outliers, then it means that there’s something interesting going on with your data that you need to look further into. Mine is also more, so I did not remove any outliers.
3. Feature extraction
   1. You can extract features from the existing data to get some meaningful information.
      1. E.g., I have extracted DATE from DATE & TIME column.
4. Feature selection
   1. Check the co-relation between variables
      1. I tried Heatmap. You can explore more ways to do the same.
      2. Select the features that have high correlation with the target variable. E.g. wind\_spd\_avg , wind\_spd\_max, wave\_ht\_sig are strongly co-related with wave\_ht\_max.

**Things To Remember**

* **There are different methods of handling the missing values: mean, median, mode, remove and many more. See whatever is relevant to your problem. Think logically, what should be the best value for imputing missing values. E.g., for weather attributes, I replace it with previous value assuming not much variation in weather in small interval of time.**
* **If outlier is due to incorrectly entered or measured data, you can drop the outlier.**
* **If the outlier does not change the results but does affect assumptions, you may drop the outlier.**
* **Use Common sense. For example, if you have a really know what range the data should fall in, like people’s ages, you can safely drop values that are outside of that range.**
* **If you have a lot of data, your sample won’t be hurt by dropping a questionable outlier.**
* **If there are a lot of outliers, it is not advisable to drop them, for example, if 30% of your data is outliers, then it actually means that there’s something interesting going on with your data that you need to look further into.**
* **Run your analysis both with and without an outlier — if there’s a substantial change, you should be careful to examine what’s going on before you delete the outlier.**
* **If the response variable follows exponential distribution or is right-skewed, try a transformation, square root and log transformations.  This can make assumptions work better if the outlier is a dependent variable and can reduce the impact of a single point if the outlier is an independent variable.**
* Feature Extraction might become important step in some cases. It aims to reduce the number of features in a dataset by creating new features from the existing ones. These new reduced set of features should then be able to summarize most of the information contained in the original set of features. It also helps to reduce dimensions, prevent overfitting and improve accuracy. E.g., I have extracted DATE from DATE & TIME column.
* While doing the co-relation analysis, consider both highly positive as well as negative co-related features. Two type of co-relations you can observe:
  1. co-relation between independent and dependent,
     1. keep only those variables that are strongly co-related with target variables.
  2. co-relation between independent variables.
     1. if two independent variables are strongly co-related, you can drop one which has less value for co-relation.