Module 3

Data Analytics with Python - Applied analytics

Section: Exploratory Data Analytics

Exploratory Data Analytics

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights as possible from it. EDA is all about making sense of data in hand, before getting them dirty with it.

Derived Variables

- In the context of statistical analysis, derived variable is one that is derived from two or more primary variable.
- Derived variables are created by calculating or categorizing the existing variables in the dataset.
- Some of the examples for derived variables are ratios, percentages, indices and rates.
- Derived variables are based on primary variables, any change in the primary variables will have an impact on the derived variables.
- Derived variables can have some unexpected properties and may require different handling methods.

Analysis of Derived Variables

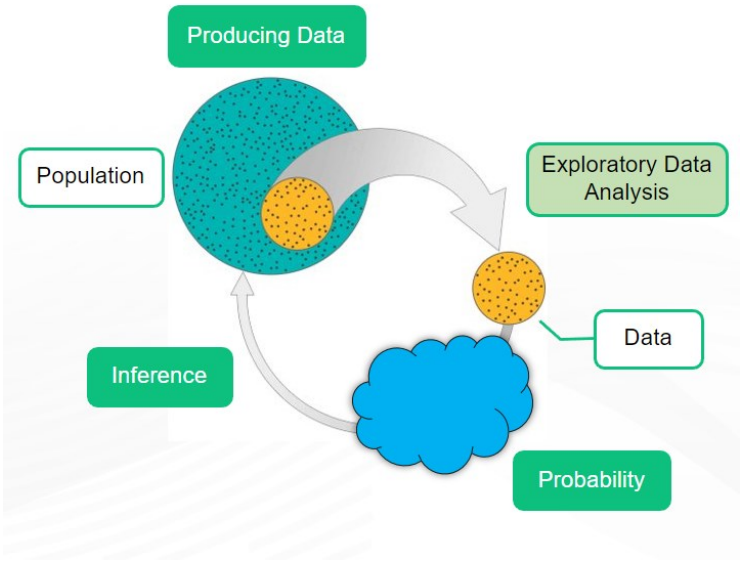
- Analysis of derived variables includes methods that take a collection of measurements and combines them into a single meaningful summary feature.
- Statistical methods used to form a model or prediction, produce both predicted values and residuals.
- The predicted values are the ones predicted by the model, which in most cases is a single variable. We can analyse the variable in relation to other variables in the model or outside the model.
- Residuals are the ones that the model couldn't predict. Residuals are used to diagnose the model and identify the problems associated with it.

- In practice, derived variables are produced during statistical analysis, with many variations, which depend on the analysis technique employed.
- There are many tools available to analyse the derived variables.

Exploratory data analysis (EDA) is done in data as a set of initial investigations, for discovering patterns, detecting anomalies, hypothesis testing and to validate the assumptions using summary statistics and other visual representations.

EDA is done in order to make sure that the data in hand is ready to be used by machine learning algorithms, and to determine the most suitable algorithms for the dataset.

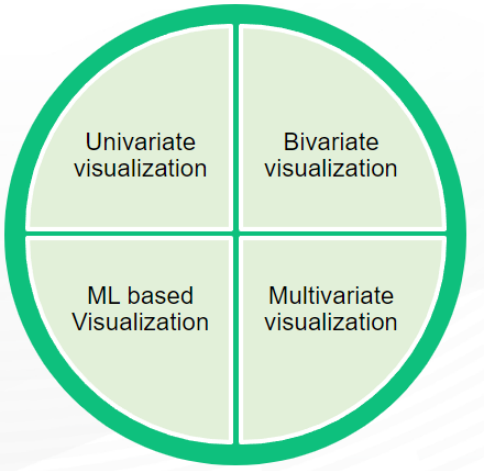
EDA is also used to determine the feature variables that can be used for machine learning.



Methods of Exploratory Data Analysis

There are different exploratory techniques available for doing EDA and the results can be compared. After doing data analysis and understanding it, we can do data collection and cleaning steps for transforming data according to the business requirements.

Some of the methods available for EDA are as follows



EDA with Data Visualization

- Univariate visualization of and summary statistics for each field in the raw dataset
- Bivariate visualization and summary statistics for assessing the relationship between each variable in the dataset and the target variable of interest (e.g. time until churn, spend)
- Multivariate visualizations to understand interactions between different fields in the data
- Dimensionality reduction to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data
- Clustering of similar observations in the dataset into differentiated groupings, which by collapsing the data into a few small data points, patterns of behavior can be more easily identified.

Approaches and Techniques for EDA

Categorical V/s Continuous	**1. Probability Distribution analysis - using distplot** **2. Swarm plot and boxplot** **3. ANOVA - f_classif**
Continuous v/s continuous	**1. Scatterplot** **2. Correlation Analysis** **3. ANOVA - f_regress**
Categorical v/s Categorical	**1. Barplot - of ratio of frequency/count in multiple groups** **2. Chi Square Test** **3. Histogram/countplot**

Summary

- It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.
- Derived variables are created by calculating or categorizing the existing variables in the dataset
- Analysis of derived variables includes methods that take a collection of measurements and combines them into a single meaningful summary feature.
- Visualizing data helps to identify pattern and understand data effectively.