*Data science*

What is DATA SCIENCE ?

Data Science Is An Interdisciplinary Field That Uses Scientific Methods , Processes ,Algorithms And System To Extract Knowledge And Insights From Structured And Unstructured Data , And Apply Knowledge And Actionable Insights From Data Across A Broad Range Of Application Domains. Data Science Is Related To Data Mining , Machine Learning And Big Data.

**Data science is a “concept to unify** statistics ,data analysis, informatics, and **their related methods” in** **order to “understand and analyze actual phenomena” with data. It uses techniques and theories drawn form many fields within the context of** mathematics ,statistics, computer science ,information science, and domain knowledge.

**What is exploratory data analysis?**

***Exploratory data analysis(EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best manipulate data sources to get the answers you need ,making it easier for data scientists to discover patterns , spot anomalies, test a hypothesis, or check assumptions.***

***EDA*** *Is Primarily Used To See What Data Can Beyond The Formal Modeling Or Hypothesis Testing Task And Provides A Better Understanding Of Data Set Variables And The Relationships Between Them. It Can Also Help Determine If The Statistical Techniques You Are Considering For Data Analysis Are Appropriate.Originally Developed By American Mathematician John Tukey In The 1970s. EDA Techniques Continue To Be A Widely Used Method In The Data Discovery Process Today.*

**Why** *Is Exploratory Data Analysis Important In Data Science? The Main Purpose Of EDA Is To Help Look At Data Before Making Any Assumptions.It Can Help Identify Obvious Errors , As Well As Better Understand Patterns Within The Data, Detect Outliers Or Anomalous Events, Find Interesting Relations Among The Variables.*

**D***ATA Scientists Can Use Exploratory Analysis To Ensure The Results They Produce Are Valid And Applicable To Any Desired Business Outcomes And Goals .EDA Also Helps Stakeholders By Confirming They Are Asking The Right Questions .EDA Cam Help Answer Questions About Standard Deviations, Categorical Variables, And Confidence Intervals. Once EDA Is Complete And Insights Are Drawn, Its Features Can Then Be Used For More Sophisticated Data Analysis Or Modeling, Including Machine Learning.*

***Exploratory data analysis tools specific statistical functions and techniques you can perform with EDA tools include:***

*Clustering And Dimension Reduction Techniques, Which Helps Create Graphical Displays Of High-Dimensional Data Containing Many Variables. Univariate Visualization Of Each Field In The Raw Datasets, With Summary Statistics. Bivariate Visualization And Summary Statistics That Allow You To Assess The Relationship Between Each Variable In The Datasets And The Targets Variables You’re Looking At .Multivariate Visvualizations , For Mapping And Understanding Interactions Between Different Fields In The Data.*

* *K-Means Clustering Is A Clustering Method In Unsupervised Learning Where Data Points Are Assigned Into K Groups, i.e. The Number Of Clusters,Based On The Distance From Each Group’s Centroid. The Data Points Closest To A Particular Centroid Will Be Clustered Under The Same Category.*

*Predictive Models ,Such As Linear Regression, Use Statistics And Data To Predict Outcomes.*

**Types of exploratory data analysis**

***There are Four primary types of EDA:***

1. ***Univariate non-graphical-*** *This is simplest form of data analysis, where the data being analyzed consists of just one variable.Since it’s a single variable, it doesn’t deal with causes or relationships.The main purpose of univariate analysis is to describe the data and find patterns that exists within it.*
2. ***Univariate Graphical-****Non Graphical Methods Don’t Provide A Full Picture Of The Data. Graphical Methods Are Therefore Required. Common Types Of Univariate Graphics Include:*

*Stem- And Leaf Plots, Which Show All Data Values And The Shape Of The Distribution.*

*Histograms, A Bar Plot In Which Each Represents The Frequency (Count)Or Proportion (Count/Tota Count) Of Cases For A Range Of Values.*

*Box Plots ,Which Graphically Depicts The Five-Number Summary Of Minimum, First Quartile, Median, Third Quartile, And Maximum*

***3.Multivariate nongraphical :*** *Multivariate Data Arises From More Than One Variable . Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.*

*Multivariate graphical:*

*Multivariate Data Uses Graphics To Display Relationships Between Two Or More Sets Of Data.The Most Used Graphics Is A Grouped Bar Plot Or Bar Chart With Each Group Representing One Level Of One Of The Variable And Each Bar Within A Group Representing The Levels Of The Other Variable.*

* *Other types of multiple graphics include:*
* *Scatter plot,which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.*
* *Multivariate chart, which is a line graph of data plotted over time.*
* *Bubble chart ,which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.*
* *Heatmap, which is a graphical represtation of data where values are depicted by color.*

*Exploratory data analysis tools some of the most common data science tools used to create an EDA include:*

***Python:*** *An Interpreted , Object Oriented Programming Language With Dynamics Sementics. Its High-Level ,Built- In Data Structures, Combined With Dynamic Typing And Dynamic Binding, Make It Very Attractive for Rapid Application Development , As Well As For Use As A Scripting Or Glue Language To Connect Existing Components Together. Python And EDA Can Be Used Together. Python And EDA Can Be Used Together To Identify Missing Values In Datasets ,Which Is Important So You Can Decide How To Nhandle Missing Values For Machine Learning .*

***What is the EDA process in Data science?***

There Are Several Models That Data Can Fit Into A Through Analysis. But Before You Do So,You Have To Determine Which Model Is An Ideal Fit For The Data At Hand.For This Reason, You End Up Exploring The Data , Its Shape, Characterstics And Come Up With A Summary That Describe The Current State Of Data At Hand And Whether The Data Needs Further Processing Before It Can Be Modeled With Statistical And Scientific Techniques. This Exploration Of Data , Usually With The Help Of Descriptive Statistics ,Visualization Tools , And Presentation Techniques, Make Up For What We Call EDA Analysis In Data Science.

Eda In Data Science Is Quite Like The Service Advisor Doing A Rough Inspection Of Your Car, Asking The Few Premilinary Questions, Setting Expections And Then Taking The Car In For Service It Is One Of The First Things Done With The Data , So It Is Critical Phase ,As Many Inferences And Consequent Actions Depends On This Explorations.

**What is EDA?**

**EDA** *Can Be Quite Extensive And Time\_Consuming Depending On What And How Much Data You Have .Unfortunately , There Is No Structured Way To Perform EDA, Although There Are A Few Techinques That Will Give You The Best Results Out Of EDA .The Important Ones That One Should Try Get From The Data Are,*

* ***Detect outliers and anomalies***
* ***Dtermine the quality of data***
* ***Determine what statistical models can fit the data***
* ***Find out if the assumptionsabout the data,that you or your team started out with is correct or way off.***
* ***Extracted variables or dimensions on which the data can be pivoted.***
* ***Determine whether to apply univariate or multivariate analytical techniques.***

*That’s why EDA is the most important port of data science*

Conclusion

EDA is essential in the analysis of massive data sets ,to be able to ensure that you have the right data for the chosen statistical model. You certainly would not want to figure out at a later stage that the data is not a good fit for the statistical model you are trying to build. A sound EDA must be performed before any data mining, data analysis, or dat modeling occurs.

*Honey Chaudhary*