**Unit-1 Chapter-1 Lecture-1.3**

**Exploratory Data Analysis (EDA)**

**Exploratory data analysis, feature selection for better ML models**

When you’re getting started with a machine learning (ML) project, one critical principle to keep in mind is that data is everything. It is often said that if ML is the rocket engine, then the fuel is the (high-quality) data fed to ML algorithms. However, deriving truth and insight from a pile of data can be a complicated and error-prone job. To have a solid start for your ML project, it always helps to analyze the data upfront, a practice that describes the data by means of statistical and visualization techniques to bring important aspects of that data into focus for further analysis. During that process, it’s important that you get a deep understanding of:

**The properties of the data**, such as schema and statistical properties;

**The quality of the data**, like missing values and inconsistent data types;

**The predictive power of the data**, such as the correlation of features against the target.

This process lays the groundwork for the subsequent feature selection and engineering steps, and it provides a solid foundation for building good ML models.

There are many different approaches to conducting exploratory data analysis (EDA) out there, so it can be hard to know what analysis to perform and how to do it properly. Based on the results of the analysis, you can then determine corresponding feature selection and engineering recommendations.

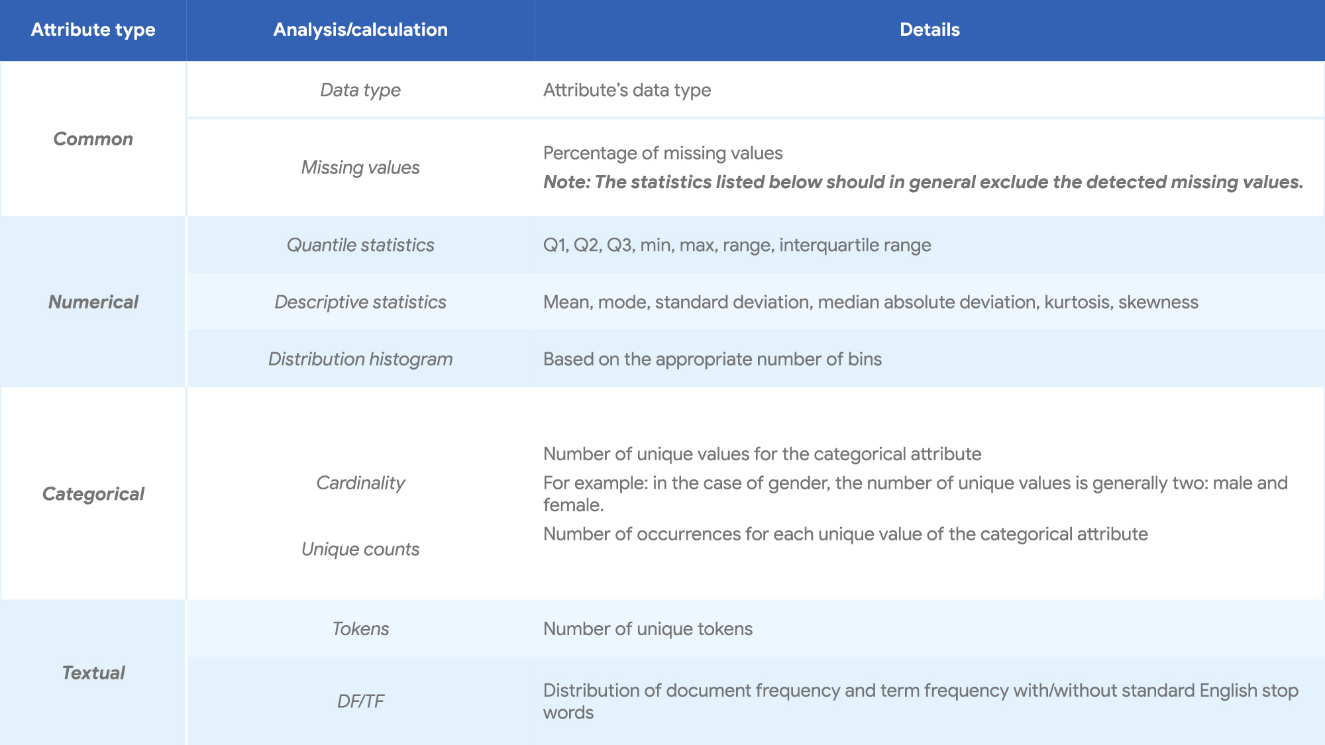
 With the complexity of data and business problems that exist today (such as credit scoring in finance and demand forecasting in retail), how the results of proper EDA can influence your subsequent decisions is a big question.

**Statistical data analysis**

With this type of analysis, data exploration can be conducted from three different angles: descriptive, correlation, and contextual. Each type introduces complementary information on the properties and predictive power of the data, helping you make an informed decision based on the outcome of the analysis.

**1. Descriptive analysis (univariate analysis)**

Descriptive analysis, or univariate analysis, provides an understanding of the characteristics of each attribute of the dataset. It also offers important evidence for feature preprocessing and selection in a later stage. The following table lists the suggested analysis for attributes that are common, numerical, categorical, and textual.



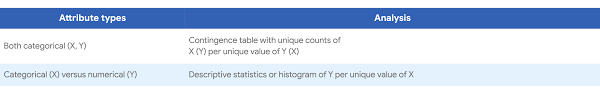
**Fig 3.1- Descriptive analysis**

**2. Correlation analysis (bivariate analysis)**

Correlation analysis (or bivariate analysis) examines the relationship between two attributes, say X and Y, and examines whether X and Y are correlated. This analysis can be done from two perspectives to get various possible combinations:

**Qualitative analysis.**

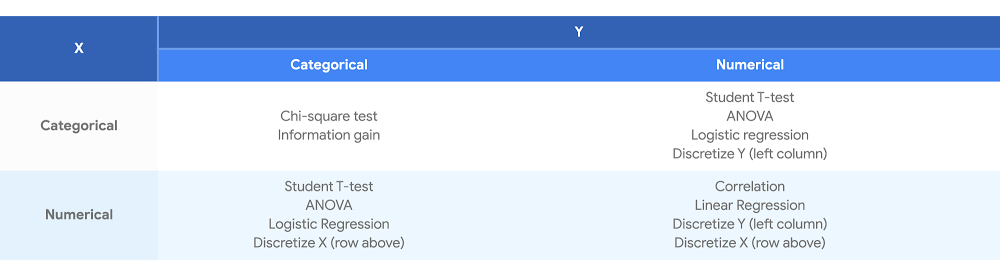
This performs the computation of the descriptive statistics of dependent numerical/categorical attributes against each unique value of the independent categorical attribute. This perspective helps intuitively understand the relationship between X and Y. Visualizations are often used together with qualitative analysis as a more intuitive way of presenting the result.



**Fig 3.2- Qualitative analysis**

**Quantitative analysis.**

This is a quantitative test of the relationship between X and Y, based on the hypothesis testing framework. This perspective provides a formal and mathematical methodology to quantitatively determine the existence and/or strength of the relationship.



**Fig 3.3- Quantitative analysis**

**3. Contextual analysis**

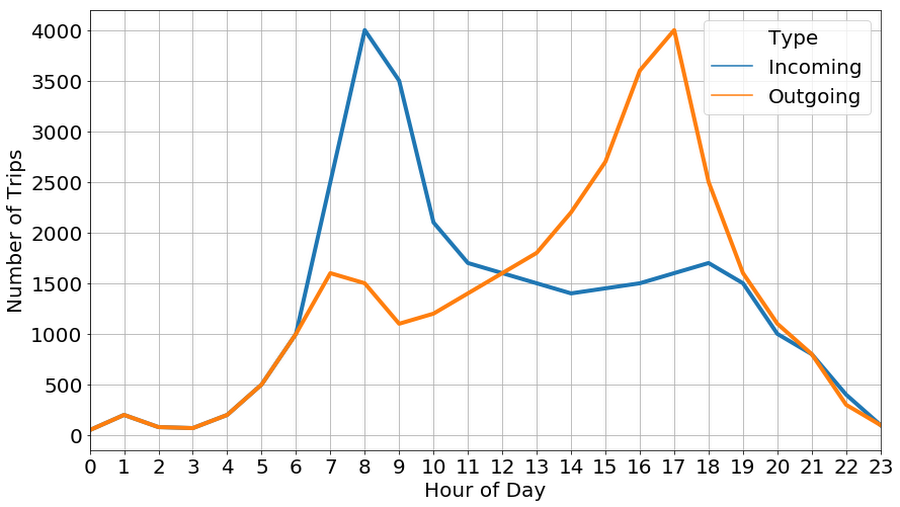
Descriptive analysis and correlation analysis are both generic enough to be performed on any structured dataset, neither of which requires context information. To further understand or profile the given dataset and to gain more domain-specific insights, you can use one of two common contextual information-based analyses:

**Time-based analysis**: In many real-world datasets, the timestamp (or a similar time-related attribute) is one of the key pieces of contextual information. Observing and/or understanding the characteristics of the data along the time dimension, with various granularities, is essential to understanding the data generation process and ensuring data quality

**Agent-based analysis**: As an alternative to the time, the other common attribute is the unique identification (ID, such as user ID) of each record. Analyzing the dataset by aggregating along the agent dimension, i.e., a histogram of the number of records per agent, can further help improve your understanding of the dataset.

**Example of time-based analysis**:

The following figure displays the average number of train trips per hour originating from and ending at one particular location based on a simulated dataset.



**Fig 3.4- Time-based analysis**

From this, we can conclude that peak times are around 8:30 am and 5:30 pm, which is consistent with the intuition that these are the times when people would typically leave home in the morning and return after a day of work.

**Feature selection and engineering**

The ultimate goal of EDA (whether rigorous or through visualization) is to provide insights on the dataset you’re studying. This can inspire your subsequent feature selection, engineering, and model-building process.

The descriptive analysis provides the basic statistics of each attribute of the dataset. Those statistics can help you identify the following issues:

* The high percentage of missing values
* Low variance of numeric attributes
* Low entropy of categorical attributes
* Imbalance of categorical target (class imbalance)
* Skewdistribution of numeric attributes
* High cardinality of categorical attributes

The correlation analysis examines the relationship between two attributes. There are two typical action points triggered by the correlation analysis in the context of feature selection or feature engineering:

* Low correlation between feature and target
* The high correlation between features

Once you’ve identified issues, the next task is to make a sound decision on how to properly mitigate these issues. One such example is for “High percentage of missing values.” The identified problem is that the attribute is missing in a significant proportion of the data points. The threshold or definition of “significant” can be set based on domain knowledge. There are two options to handle this, depending on the business scenario:

* Assign a unique value to the missing value records, if the missing value, in certain contexts, is actually meaningful. For example, a missing value could indicate that a monitored, underlying process was not functioning properly.
* Discard the feature if the values are missing due to misconfiguration, issues with data collection, or untraceable random reasons, and the historic data can’t be reconstituted.

**Book Reading and Video Material**

* Understanding Machine Learning: From Theory to Algorithms by Shai Shalev-Shwartz and Shai Ben-David-Cambridge University Press 2014 [Download](https://www.cse.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf) Buy at Amazon
* Introduction to Machine Learning – the Wikipedia guide [Download](http://datascienceassn.org/sites/default/files/Introduction%20to%20Machine%20Learning.pdf)
* [Weblink (Analytics Vidya)](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/)
* [Online Reading Material(NIST Govt.)](https://www.itl.nist.gov/div898/handbook/toolaids/pff/eda.pdf)
* [NPTEL Video Part-1](https://www.youtube.com/watch?v=-0Au19Kercg)
* [NPTEL Video Part-2](https://www.youtube.com/watch?v=vIwtyJmmxiE)