greatlearning
Learning for Life

Exploratory Data Analysis

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Topics of the week

- Introduction to EDA
- Stages of EDA and EDA in data science life cycle
- Distribution of Data
- Types of variables and missing values
- Descriptive data exploration – five point summary
- Data Distribution – pairplot and heatmap
- EDA Summary
- Case Study

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Introduction - Exploratory Data Ana

- Exploratory data analysis is one of the crucial steps in the data analysis process. It's the first thing that you'd do before you start with modelling your data. It provides all the necessary context to develop a deeper understanding of the data you're dealing with and help create an appropriate model and interpret the results correctly.

- To give an idea of the importance EDA holds in a project, note that In a typical project life-cycle, upwards of 50% of the time is spent on procuring, cleaning, and exploring the data.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Introduction - Exploratory Data Ana

EDA Comprises of:

- Analysing dataset.
- Summarizing the main characteristics
- Detecting the anomalies and outliers
- Visualise patterns / trends

Objectives of EDA:

- To understand the spread of variables in the dataset
- To obtain cues on relationship between variables in a dataset
- To detect any outliers in the dataset
- Spot missing values in the dataset

Please note:

- EDA is the initial examination of the dataset and hence, the observations from EDA need not necessarily be statistically significant.
- Further statistical models need to be applied to confirm statistical significance.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

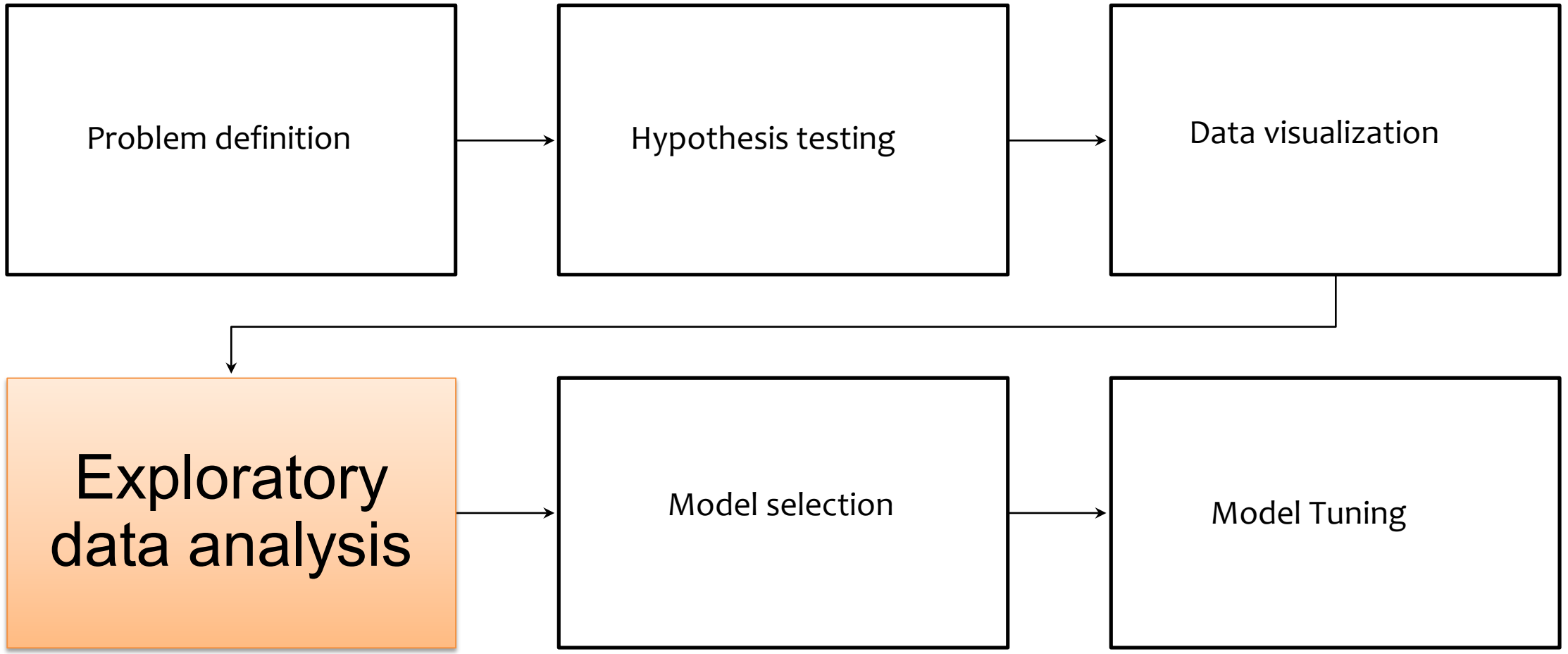
Introduction - Exploratory Data Ana

Stages of EDA

- Univariate analysis — provides summary statistics for each field in the raw data set
- Bivariate analysis — to find the relationship between each variable in the dataset and the target variable of interest
- Multivariate analysis — is performed to understand interactions between different variables in the dataset

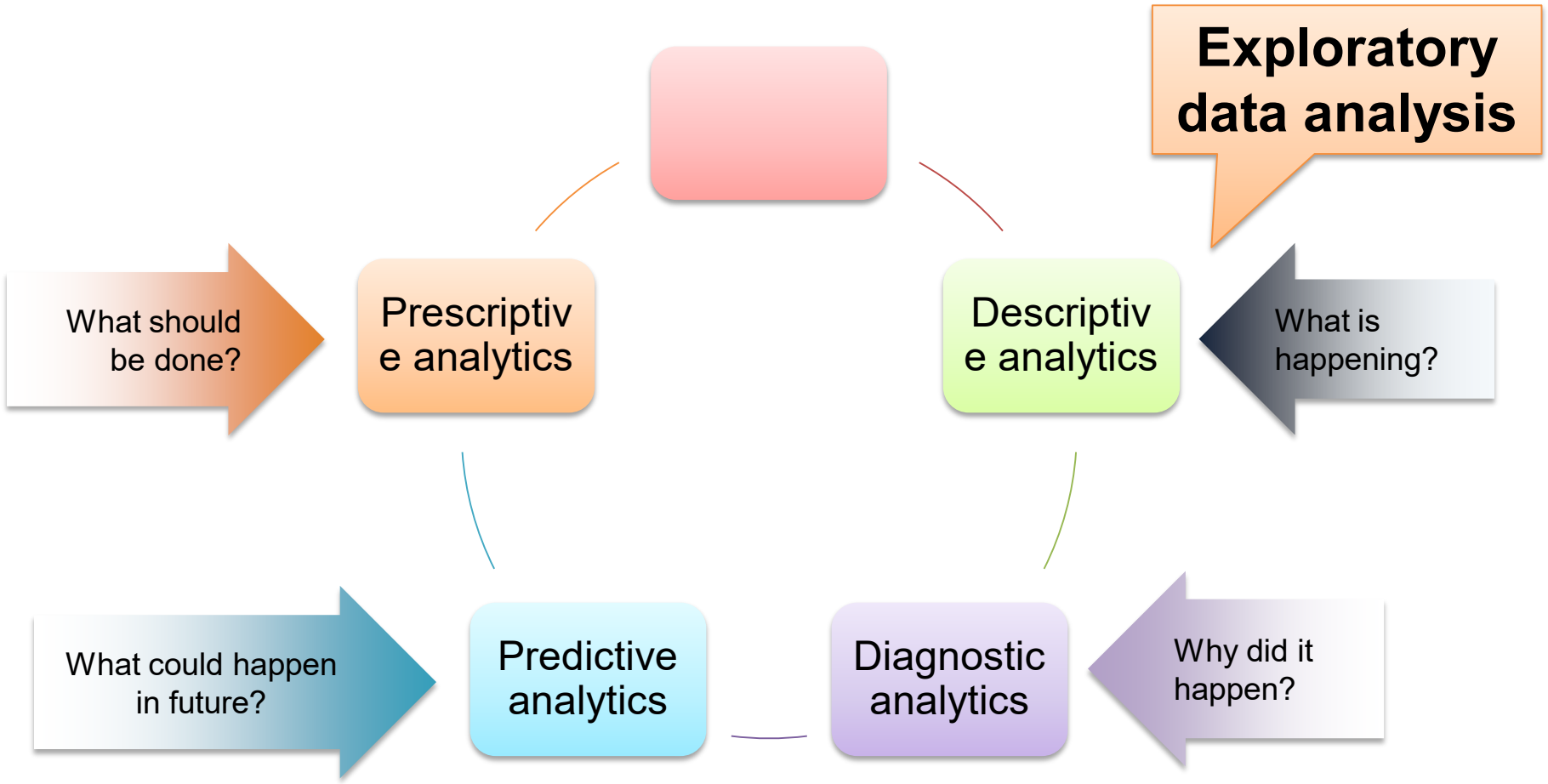
Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Data Science life Cycle and EDA



Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Data Science life Cycle and EDA



Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Introduction to Exploratory Data Ar

Now let us discuss a few important aspects of EDA.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Types of variables

- Types of Variables - Categorical and Continuous
  - Categorical Data
    - Data described in categories, example - Gender, Nationality
  - Continuous Data
    - Continuous distribution of the data points, example - salary

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Missing values

- Dealing with missing values is an important aspect in the exploratory data analysis.
- Missing data should be treated accordingly before being used to create a model.
- There are various methods to deal with the missing data such as imputation, multiple imputation, filling with the central tendencies etc.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Descriptive Data Exploration

greatlearning
Learning for Life

Descriptive Data Measures

- Measures of central tendency - Mean, median, mode
- Measures of dispersion – Quartiles, percentiles, Standard deviation, variance, coefficient of variation
- The above measures give a fair idea of how the variable is distributed.
- The describe method of pandas helps in this analysis(discussed in the next slide)

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Descriptive Data Exploration - Five

- It is a common EDA practice to perform five point summary on the dataset at hand.
- The Five point summary gives us valuable insights regarding the distribution of the variables
- Using pandas, this can be performed using df.describe() function. Below is a snippet of the describe() function applied on a dataframe.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Distribution of Data

- Shape of Data
  - Skewness - Data might not always be normally distributed. There might be some degree of skewness involved.
  - Df.skew() function gives the skewness of all the features at hand.
  - Example Snippet ->

```
pima_df.skew()
```

```
Pregnancies                 0.901674
Glucose                     0.173754
BloodPressure              -1.843608
SkinThickness               0.109372
Insulin                     2.272251
BMI                        -0.428982
DiabetesPedigreeFunction    1.919911
Age                         1.129597
Outcome                     0.635017
```

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

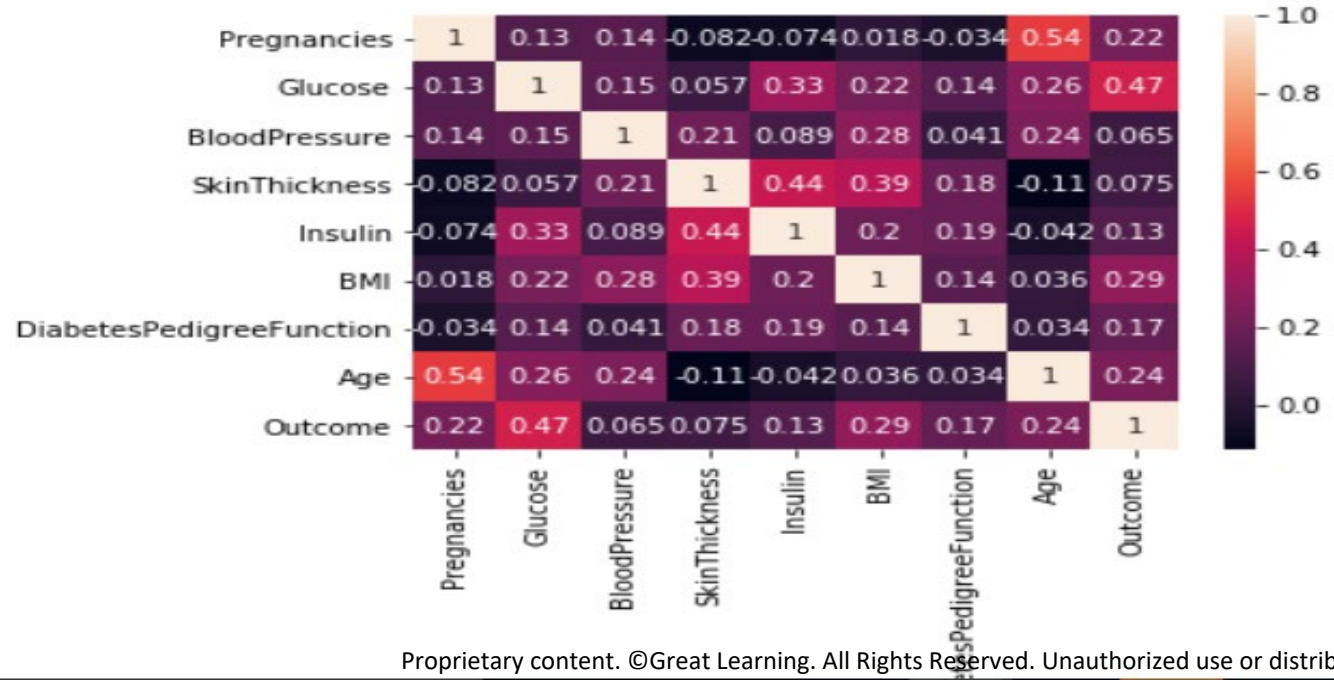
Distribution Of Data - Pairplot

- Pairplot is a powerful visualization for the bivariate analysis. It gives an understanding of the variation of a variable with another.
- The same can be used for visual analysis of outliers, another important concept w.r.t. to exploratory data analysis.
- Given below is a snippet for a seaborn pairplot.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Correlation

- A common practice in EDA is to identify the correlation between all the variables that are there in the dataset.
- Correlation provides a measure of association between two variables.
- Correlation, however, does not imply causation and should not be confused with the same.
- Correlation can be plotted best by a correlation heatmap. Below is a snapshot.



Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Summarizing EDA

- Exploratory data analysis is a very important and powerful analysis that aids in understanding the data at hand for modelling.
- EDA steps vary with the data at hand. A clean and structured data need to go through lesser steps while an unstructured one needs a lot of time on EDA.
- Visualization is an important step in Exploatory data Analysis. A visual analysis of a complex data might make extracting insights easy.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Case Study

Pima Indian Diabetes Analysis

Pima Indians diabetes dataset is a popular real life datasets used to study diabetes. The Pima Indians show a high tendency of having diabetes and In particular, all patients here are females at least 21 years old of Pima Indian heritage. Let us perform EDA to analyze the attributes and their effects on the outcome variable.

Attribute Information:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function 8. Age (years) 9. Class variable (0 or 1)

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

greatlearning
Learning for Life

HappyLearning

Questions?

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited