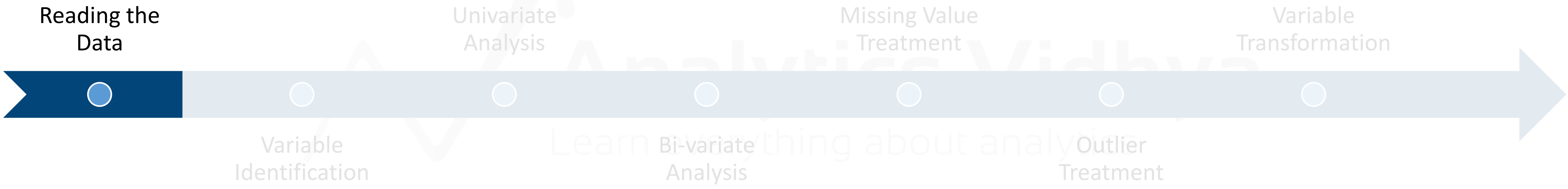
Data Exploration

Stages of Data Exploration



Reading the Data

Variable Identification

Univariate Analysis

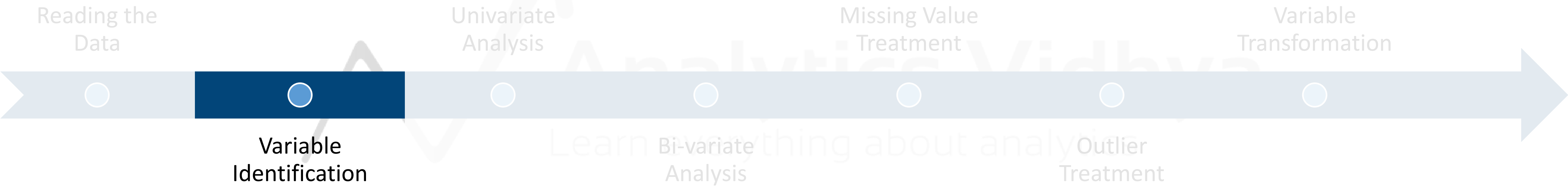
Bi-variate Analysis

Missing Value Treatment

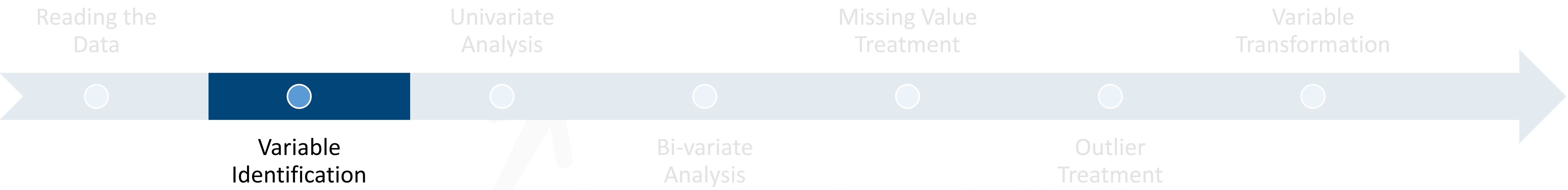
Outlier Treatment

Variable Transformation

Stages of Data Exploration



Stages of Data Exploration



Reading the Data

Variable Identification

Univariate Analysis

Bi-variate Analysis

Missing Value Treatment

Outlier Treatment

Variable Transformation

Topics to be covered:

1. What is Variable Identification?
2. Why do we need to perform Variable Identification?
3. How to perform Variable Identification?

Variable Identification

What is Variable Identification?

1. Independent and Dependent Variables.

2. Continuous and Categorical Variables

Variable Identification

Why do we perform Variable Identification?

1. Techniques like Supervised Learning require identification of Dependent Variable.

2. Different data processing techniques for Categorical and Continuous data.

Variable Identification

Difference between Dependent and Independent Variables

1. Dependent Variable – The variable we are trying to predict.
   Example – Survived variable in titanic problem

2. Independent Variable – The Variables which help in predicting the Dependent Variable.
   Example – Sex, Fare etc. in titanic problem

Variable Identification

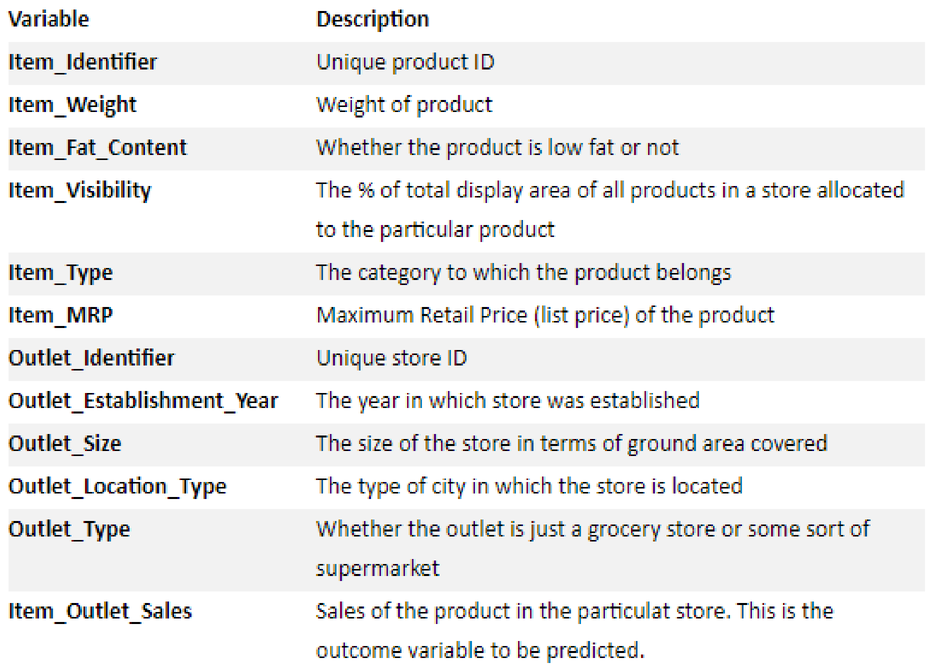
How to identify Dependent and Independent Variables

Can only be identified from the Problem Statement.

Variable Identification

How to identify Dependent and Independent Variables

Problem : The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.



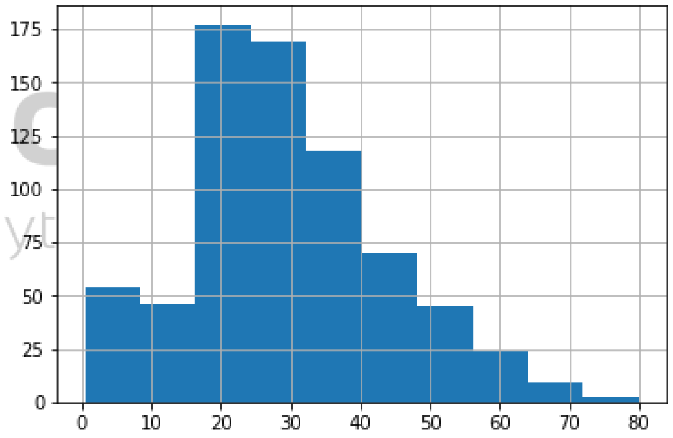
Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particulat store. This is the outcome variable to be predicted.

Variable Identification

Difference between Categorical and Continuous Variables

1. Categorical Variable – Discrete in nature
   Example – Survived, Sex

2. Continuous Variable – Can have infinite number of possible Values
   Example – Fare, Age



Variable Identification

How to identify Categorical and Continuous Variables

Categorical Variables – Stored as **object**

Continuous Variables – Stored as **int** or **float**

dtypes

Thank You