Problem Statement

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month.  
The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month.

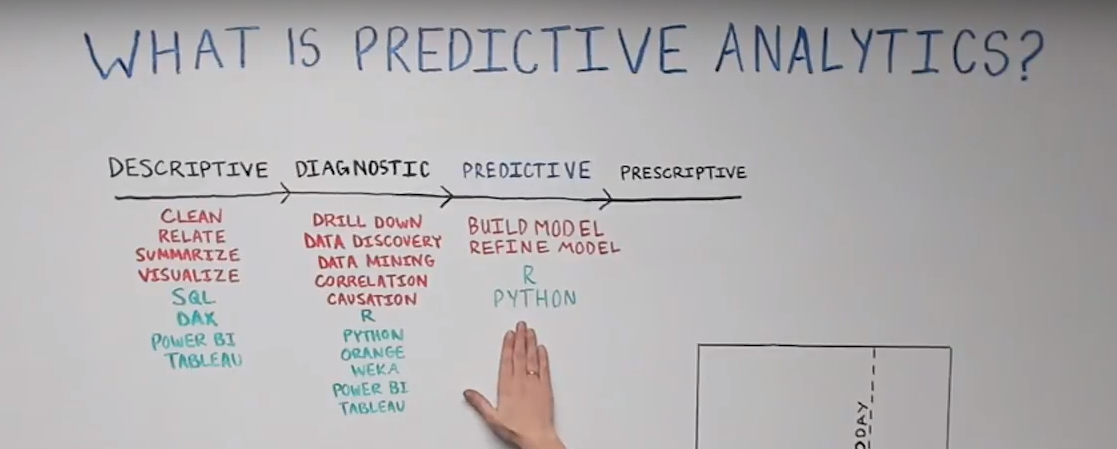
Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

**METRICS**

Purchase amount of customer = Purchase amount

Various products = 3491 product IDs

Age, gender, marital status, city\_type, stay\_in\_current\_city



[2 variable Correlations](https://www.youtube.com/watch?v=Ypgo4qUBt5o)

[Predictive Analytics Intro](https://www.youtube.com/watch?v=4y6fUC56KPw)

1. Understand the problem or ask interesting questions
   1. Understanding problem
      1. Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.
   2. Questions
      1. Who is the audience that will use the results from the analysis? sales people
      2. How will the results be used? They want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.
      3. What questions will the audience have about our analysis?
      4. How should the questions be prioritized to derive the most value?
      5. Which data sources are available to work with? Model training data and model testing data
      6. What is size of each data set and how much data will I need to get from each one? Training=500k rows, Test= 233k rows
2. Data preparation
   1. Understanding Data
      1. Infer that 0 is not married because all users below 17 had 0
      2. Is there a need to fix any value in the database? NO
      3. How is each column’s Data presented (i.e categories represented as numbers)
         1. How many unique categories do we have? 20 unique categories (1 to 20)
   2. Extreme values
      1. Data types and representations?
         1. Dtype code
      2. Is the data Gaussian-like ?normality Test using Q-Q plot [LINK](https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/) Data isn’t Gauusian
      3. Found outliers? No Outliers
   3. Cleaning Data
      1. Finding missing values
         1. How many values are missing for each column “Code” find the blank cell
            1. How much of total % is missing?Only category 2 and 3 have missing values
   4. Descriptive Analysis
      1. Comparing column values based on purchase amount
3. Data Analysis
   1. Predictive Analysis
      1. What kind of model?
      2. Try to get an equation like so Purchase = Beta + x1Age + x2Gender +…..
4. Visualize
   1. Sales by Gender
5. Answer questions with Visual presentation