**CUSTOMER SEGMENTATION USING DATA SCIENCE**

**Phase 2:** Innovation

To turn our concepts into innovations, we first perform data wrangling before deploying the supervised learning technique - Regression.

**Data Wrangling:**

The first step is importing the dataset into our primary data analysis tool, sypder, a popular integrated programming environment (IDE) for Python that makes the most of Python modules such as Pandas, NumPy, and Matplotlib.

For preliminary data exploration, we employ Pandas and Matplotlib. We use Pandas methodologies like dropna(), fillna(), and so on to handle missing values. The drop\_duplicates() method is utilized for eliminating duplicates. The dataset is then aggregated, which results in data grouping, calculating the number of occurrences inside a category, minimum and maximum values within a category, identifying the frequency or occurrence of each unique value within a category, and so on.

**Regression:**

We commence by loading the cleaned dataset that was generated throughout the data-wrangling procedure. The predictor (independent) variable(s) and target (dependent) variable(s) are then determined for regression analysis.

Independent variables: Age, Annual Income

Dependent Variable: Spending score

Because of the identifiers or categorical variables, CustomerID and Genre are typically not considered independent or dependent variables in the context of regression analysis.

The principal objective of regression analysis is to forecast the "Spending Score" based on "Age" and "Annual Income".

**Outcome of this phase:**

Progressively it turns out, quality statistics and consistency, as accurate and reliable models will be achieved, dimensionality reduction, improved insights, and beyond. The interpretation of the regression results and their significance becomes more reliable with clean data.