According to a survey 3 of 5 data scientists spend most of the time of their day working on Data pre processing i.e Data Preparation .

This Phase has following Output directly apart from Outputs derived from tasks.

1.Data set which is prepared for Modelling or the major analysis work of the project. Data set description also to be included along with Data set which will be used for modelling .

**Task1: Select Data:**

In this task We select Data (Feature selection and Sampling from Data)

Output: rationale for inclusion or exclusion

We have checked correlation between explanatory varaibles and correlation is too low that We did not exclude any explanatory variables from our analysis

Task2: Clean Data:

We once again check for the Data Quality of the selected data and if found any errors , We need to reverify the Data Quality report from Data Understanding phase.

Output: Data Cleaning Report.

Checked for errors and did not find any.

Task 3: Construct Data:

This task includes creation of new attributes from existing attributes for the sake of Modelling.

Ex: We define area when given length and breadth for a product.

Output: Derived attributes and generated records.

Task4:Integrate Data:

These are methods whereby information is combined from multiple tables or records to create new records or values.

Output:

Merged data

Task 5: Format Data:

Formatting transformations refer to primarily syntactic modification made to the data that do not change its meaning.

Output:

Reformatted Data.