Data Analysis Framework

Framework: Data Processing

Data Processing is the process of converting data to information on which analysis can be performed. It involves multiple processes.

a) Data Merging

Data might reside in multiple files. To perform analysis, you require those data to be in one place. The Data analyst will perform merging. Files should have at least one common column by which they can be merged.

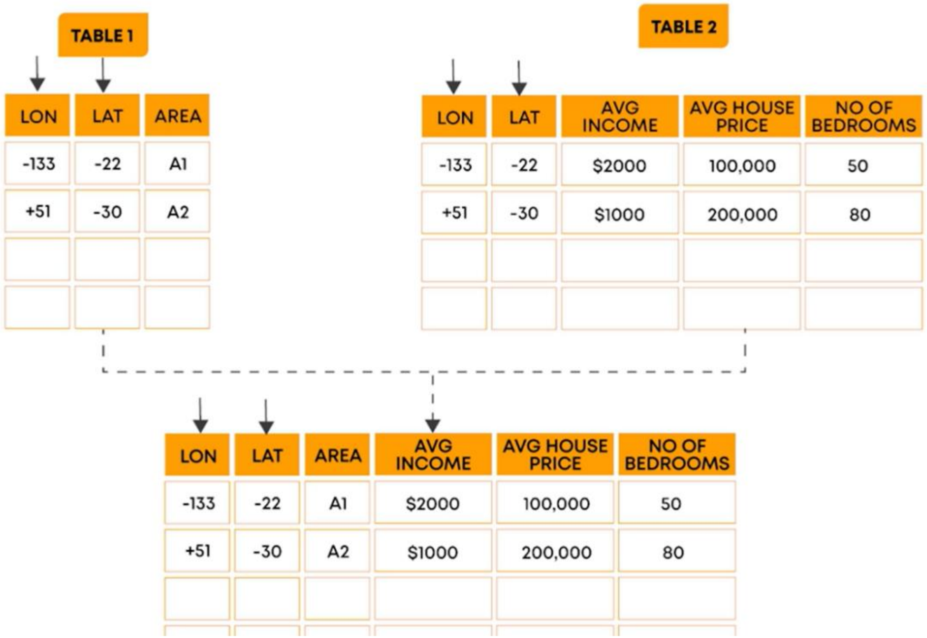


TABLE 1

LON	LAT	AREA
-133	-22	A1
+51	-30	A2

TABLE 2

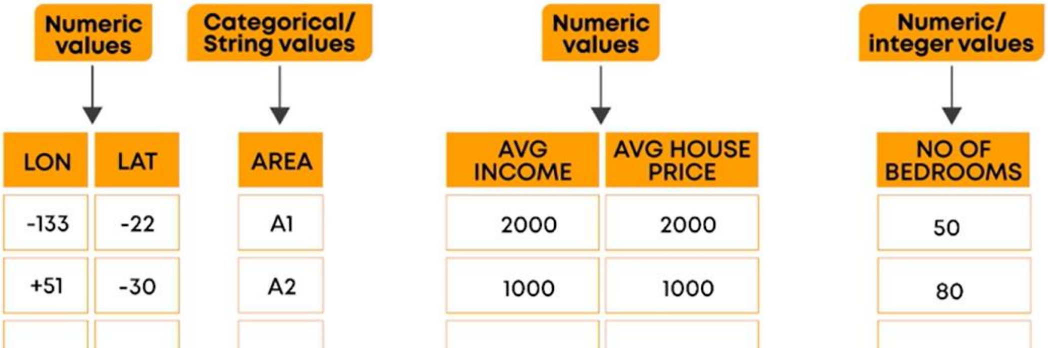
LON	LAT	AVG INCOME	AVG HOUSE PRICE	NO OF BEDROOMS
-133	-22	$2000	100,000	50
+51	-30	$1000	200,000	80

LON	LAT	AREA	AVG INCOME	AVG HOUSE PRICE	NO OF BEDROOMS
-133	-22	A1	$2000	100,000	50
+51	-30	A2	$1000	200,000	80

Figure 1 The content of Table1 is merged with Table2. LON & LAT are the common features.

b) Data Type Validation

All the features should be in an expected format. For example, all the date columns, like birth date, transaction date, etc., need to be in date format.



LON	LAT	AREA	AVG INCOME	AVG HOUSE PRICE	NO OF BEDROOMS
-133	-22	A1	2000	2000	50
+51	-30	A2	1000	1000	80

Figure 2 Different Datatypes (numeric, string) in a Dataset

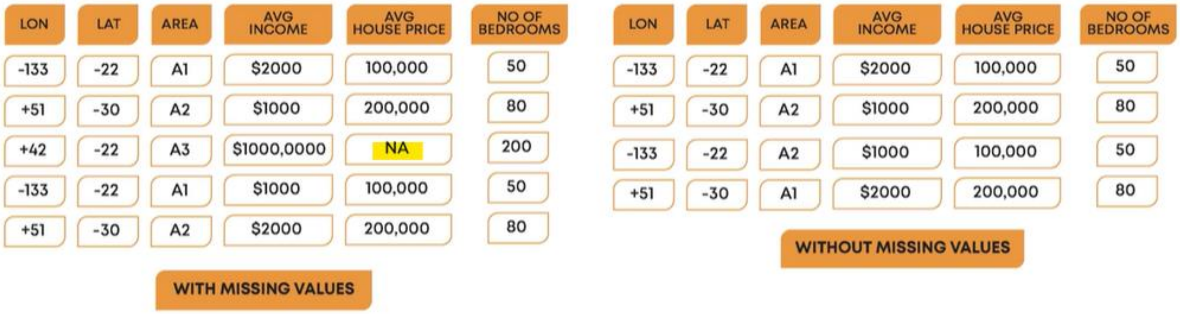
The correct format helps in the analysis and feature engineering methods.

c) Handling the Missing Values

In the dataset there are chances some data is not available for any attribute. That data is known as missing values. You can ignore those values if that particular attribute with missing values is not required for your data analysis. Still, if that specific column is necessary for the analysis, you must handle those missing values.

There are various method lines to handle these missing values.

- Deleting those rows (if missing value rows are too few compared to the whole dataset).
- Putting zero or replacing the value using the statistics method (Mean/Median/Mod etc., all these methods depend on the dataset and the need for the attribute in the analysis).



LON	LAT	AREA	AVG INCOME	AVG HOUSE PRICE	NO OF BEDROOMS
-133	-22	A1	$2000	100,000	50
+51	-30	A2	$1000	200,000	80
+42	-22	A3	$1000,0000	NA	200
-133	-22	A1	$1000	100,000	50
+51	-30	A2	$2000	200,000	80

WITH MISSING VALUES

LON	LAT	AREA	AVG INCOME	AVG HOUSE PRICE	NO OF BEDROOMS
-133	-22	A1	$2000	100,000	50
+51	-30	A2	$1000	200,000	80
-133	-22	A2	$1000	100,000	50
+51	-30	A1	$2000	200,000	80

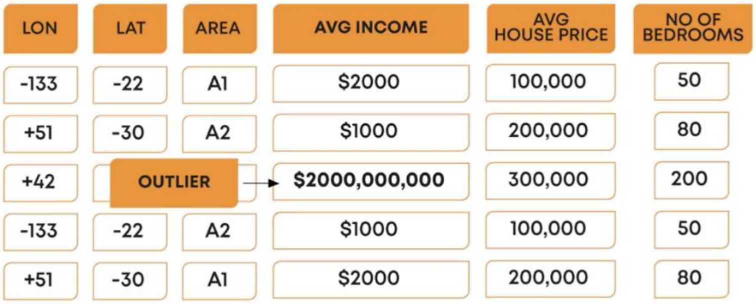
WITHOUT MISSING VALUES

Figure 3 Tables with missing value (represented as NA) and without missing values

Handling this missing value is required as it impacts the final analysis, giving wrong insight if data is not available.

d) Handling Outliers

Outliers are the data points or values which do not fall into all the other values. For example, if all values in a column are between 1 to 100 and there are two values which are 2000 and 3000, then these two will be outliers. Outliers can be an error or exceptional cases in any particular data.



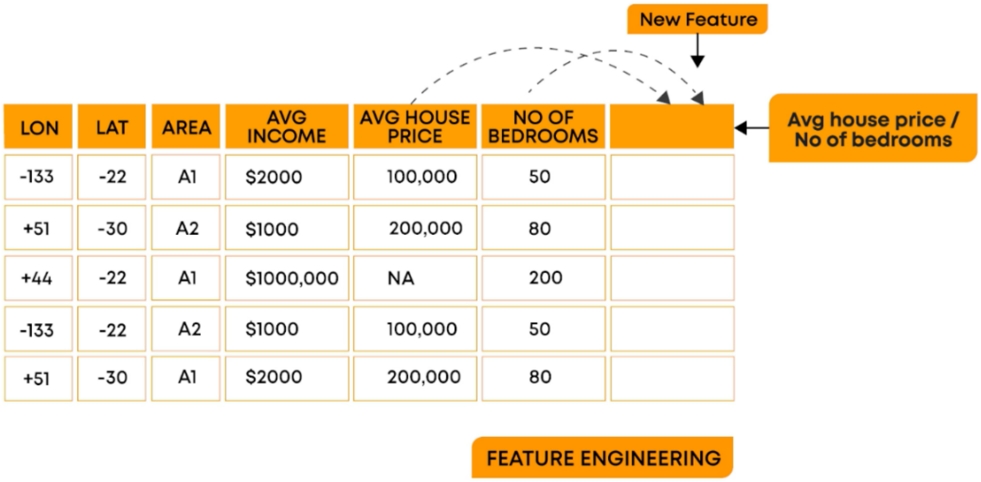
LON	LAT	AREA	AVG INCOME	AVG HOUSE PRICE	NO OF BEDROOMS
-133	-22	A1	$2000	100,000	50
+51	-30	A2	$1000	200,000	80
+42			$2000,000,000	300,000	200
-133	-22	A2	$1000	100,000	50
+51	-30	A1	$2000	200,000	80

Figure 4 In the Average Income column, the value $2000,000,000 is an outlier

You need to handle these types of values as it impacts the result of the analysis. These kinds of outliers can impact mathematical computation. You can handle these by removing those rows or by bucketing methods.

e) Feature Engineering

Feature Engineering is a critical part of the success of data analytics projects. To come up with a good set of features is important. Feature engineering involves the following steps: Feature selection is the process of selecting the most useful features among existing features, and Feature extraction combines existing features to produce a more useful one. These new features can be helpful in analysis.



LON	LAT	AREA	AVG INCOME	AVG HOUSE PRICE	NO OF BEDROOMS	
-133	-22	A1	$2000	100,000	50	
+51	-30	A2	$1000	200,000	80	
+44	-22	A1	$1000,000	NA	200	
-133	-22	A2	$1000	100,000	50	
+51	-30	A1	$2000	200,000	80	

New Feature
Avg house price / No of bedrooms
FEATURE ENGINEERING

Figure 5 New feature (column) is created, using “Average House Price” & “No. of Bedrooms” columns