<u>**Project Pruposal - Type 3**</u>

Matan Solomon 209339894,   Lotan Amit 208800342

We chose to investigate and explore the problem of missing and unreasonable values in the dataset and its effect on training prediction models.

Problem Explained:

In each raw dataset there are some missing values, some of them caused by mistakes and some caused by the properties of the features.
The AI models can not ignore a missing value so before training the model on the data we need a full dataset. The default solutions are removing all items with missing features or giving a default value for each feature, 0 for example. This approach may lead to a shift in the training dataset which could harm the model accuracy.

Our Approach:

After finding a dataset with some missing values we will train a simple
First, we will classify the type of missing values into one of the following:
1) Missing Completely at Random (MCAR).
2) Missing at Random (MAR).
 3) Missing Not at Random (MNAR).
After that, we will check different ways to fill in the missing values and how each way affects the model training.

The dataset we will use to explore the problem is ‘*Housing Prices Competition for Kaggle Learn Users*‘ (https://www.kaggle.com/competitions/home-data-for-ml-course), this is a big dataset with a lot of features and some of them have a large percentage of NULL values.