**Title**

Building a database by extracting data from PDF files using Python and Excel.

1. **Introduction**

In this project, we created a database for the City of Toronto from data extracted from the Toronto Region Real Estate Board (TRREB) website. We performed various data preprocessing techniques to clean, format and analyze the data. We deal with missing values and improper formats using the appropriate techniques. All of this was done using python and excel. The main goal was to create an end-to-end ETL pipeline from the data source to the database.

The challenges faced and the solutions utilized are also explained. For this project we are not necessarily interested in the nature of the data but more so creating a dataset. In the next project we deep dive into understanding the data that we have extracted.

1. **Data Collection**

We are creating our own data set from TRREB website.

We will be downloading the PDFs from this link [TRREB](https://www.trreb.ca/index.php/market-news/market-watch/market-watch-archive) and then using excel and python to transform the data in the correct format. The PDFs files that we use are under the Market Watch Archive.

* 1. **Web scraping the PDF files.**

To automate this process, we have written a python code where you can enter the years and months for PDF files you want to extract from TRREB website under the Market Watch Archive. We have named the file as **Script0-file\_extraction.py**. The output of the code that is the PDF files will be stored in the raw folder in the path ‘**City of Toronto\src\data\housing\_market\_Q1\raw’**.

* 1. **Modifying the PDF file**

Upon examining the file, you will see that there are 27 pages but not all of them are required for this project. So, we will have to extract only the pages that we require. The data we require is for the City of Toronto, so we extract specific pages through a python script we have written called **Script1-extract\_pdf\_pages.py.** In this code, we choose which pages we require and store the new PDF in the preprocessed folder in the path ‘**City of Toronto\src\data\housing\_market\_Q1\preprocessed’**.

* 1. **Extracting tables from the PDF**

1. **Data Exploration and Preprocessing:**