**UK HOUSING PRICE**

# Introduction

This original dataset provides comprehensive information on property sales in England and Wales in 1995 - 2023, as sourced from the UK government's HM Land Registry. It offers valuable insights into property transactions, including sale prices, locations, and types of properties sold.

Although the original dataset has the period from 1995 to 2023, but in this project, I only focus on the period 2019 – 2023.

# Programming languages

Python and T-SQL

# Download Dataset

Link: UK Housing Prices 1995- 2023 (Github)

Or Link: <https://www.kaggle.com/datasets/lorentzyeung/price-paid-data-202304/data> (Kaggle)

**Address data explanation**

Postcode: The postal code where the property is located.

Street: The street name where the property is located.

Locality: Additional locality information.

Town/City: The town or city where the property is located.

District: The district in which the property resides.

County: The county where the property is located.

Price: The price for which the property was sold.

# Normalizing the dataset

Purposes:

- Reduce the size of dataset from 4GB to under 1GB

- Decrease the period of time from 1995 - 2023 to 2019 - 2023

- Original dataset has 16 columns, drop out 4 columns to become 12 columns

- Reduce one csv 4 GB to 3 csv files as MSSQL not allow to upload file over 300 KB

**Step 1: Open new Notebook from Kaggle**

<https://www.kaggle.com/code>

Note: you can use another environment to normalize the data like Google Colab, etc, but for convenience, Kaggle will allow bigdata file uploads without setup time.

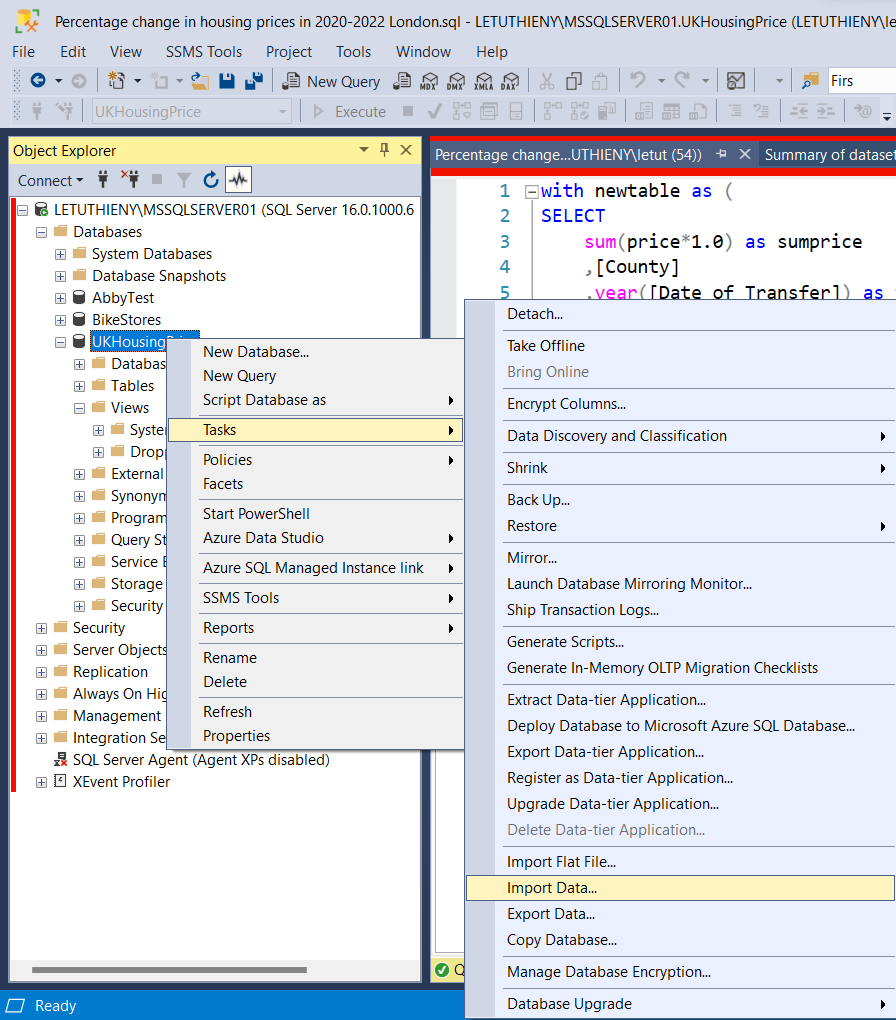
**Step 2: Open file I already adjusted**

Link: greaterlondon-property-price-autoviz-catboost-shap.ipynb

**Step 3: Download the CSV files to your computer**

Link:

# Upload the CSV files to MSSQL



# Clean and prepare data

# Insight Queries

\*\* Files need to download

Main file

https://www.kaggle.com/datasets/lorentzyeung/price-paid-data-202304/data

Additional files:

Index of Place Names in Great Britain (December 2022)

https://geoportal.statistics.gov.uk/datasets/8f8b561f256b40c3a6df71e400bb54f0/about

Summary of Results

The dataset contains records of property sales dating back to January 1995, up to the most recent monthly data. It covers various types of transactions, from residential to commercial properties, providing a holistic view of the real estate market in England and Wales.

In this paper, i will only forcus to the data from 01 Jan 2019 to 2023

MS SQL - Analysing data

- Add ID for tables

ALTER TABLE YourTable

ADD ID INT IDENTITY(1,1);

- Create empty table name UK\_Housing\_Prices with matched columns and data types of existed 3 tables

- Combines 3 tables to 1 table UK\_Housing\_Prices, using Insert and Select and Left Join

Finding insight of data

- What is the top 10 streets have the most expensive property in the UK?

- Increasing percentage of properties since 2019?

# Upload big files to Github

**Step 1: Download Github for desktop**

<https://desktop.github.com/>

**Step 2: Download GIT**

<https://gitforwindows.org/>

**Step 3: Download Git Large File Storage (LFS)**

It replaces large files such as audio samples, videos, datasets, … with text pointers inside Git, while storing the file contents on a remote server like GitHub.com or GitHub Enterprise.

<https://git-lfs.com/>

**Step 4: Navigate the file direction for LFS get the files**

Open command Prompt

🡺 input

cd A:\Abby Portfolio\UK\_Housing\_Price

**Step 5: Initializes a new Git repository**

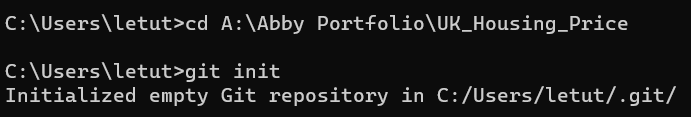
Open command Prompt

🡺 input

git init

🡺 output

Initialized empty Git repository in …



**Step 6: Initialize Git LFS**

Open Git Bash

🡺 input

git lfs install

🡺 output

Updated Git hooks

Git LFS initialized

