HDB Resale Price

Analysis on ETL / Data Prep

By Arif

1. There are 5 files data from Jan 1990 to May 2023  
   I combined all of them into 1 file
2. Some common things across those 5 files
   1. Field ‘Month’ has value of ‘YYYY-MM’  
      I transformed it to ‘YYYY-MM-01’ , with hope that it’s easier later in the next process when the field is changed to a date field.
   2. Field ‘lease\_commence\_date’ has value of ‘YYYY’  
      I transformed it to ‘YYYY-01-01’, with the same objective above.
   3. Field ‘flat\_model’ has 2 inconsistent values : ‘MULTI GENERATION’ and ‘MULTI-GENERATION’  
      I changed the values to become ‘MULTI GENERATION’
3. Some common things across the first 3 files, for data 1990-2014
   1. There is no field ‘remaining\_lease’ as we have in other file  
      I calculated the value with formula : lease\_commence\_date - month
4. For the fourth file 2015-2016 data
   1. Field ‘remaining\_lease’ is available with 2 digit format ‘nn’ - in years
5. For the fifth file 2017 onwards
   1. Field ‘remaining\_lease’ has value of ‘YY years MM months’  
      I transformed it to become numeric