AIAP Question 1

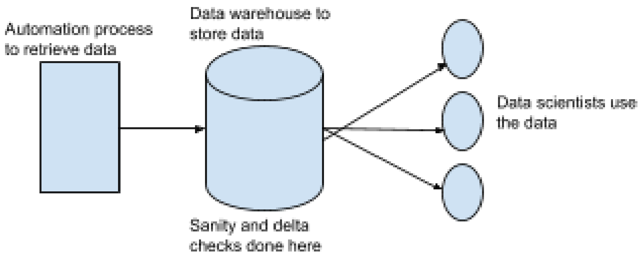
The main task is to deliver good Singapore housing data that is consistently up to date. This can be implemented through an automation process that retrieves the data from the relevant sources and ensures that data is available for the data scientists to use. The key pointers to take note of would be the frequency of data retrieval, any deltas in the data, how to implement the deltas and delivery of data for usage.

My main idea to implement this came as a combination to use TagUI for the automation process to retrieve the data and some simple Python code to verify the quality and structure of the data. The main sources of data I used were from URA for private residential data and data.gov with HDB data for public housing.

The TagUI code simply enters the websites to retrieve the data. Following that the python code provides a basic summary of the data and checks if the data has any missing values and will inform on which columns to take note of for the data scientists. This is to aid the data scientists before their exploratory analysis as they would be required to clean the data and this information will be vital.

The above is mainly a simple way to implement the process in a local machine. However, when we scale the process there may be more processes involved. Beginning with the automation process, we need to set a frequency. If the data retrieval is to be daily, we will need to retrieve the data and compare it with previous data for any deltas. This can be implemented through code. Once the deltas are noted, we will have to update our master copy. This will ensure that the data is consistently up to date. Thus, it is important to note how frequent the data is being retrieved as higher frequencies will lead to higher deltas and equally higher computation.

The sanity check of the data at this stage is fairly sufficient as the main information on missing values and basic data structure is known to the data scientists. It will then be in their realm to clean and use the data as per their requirement. As mentioned, with higher frequencies of data retrieval there might be the need to store this data in a data warehouse for the retrieval by the scientists. This can be managed as a central repository where data is dumped daily and deltas are checked and updated accordingly for use by the data scientists. Below shows a simple illustration of the entire process.


Automation process to retrieve data
Data warehouse to store data
Data scientists use the data
Sanity and delta checks done here