------------------------------------------------------------------------------

**Subject: 109201-MTH-650-NWMSB Introduction to Data Analytics Graduate Midland Fall 2022-2023**

**Professor: Dr. Itauma Itauma**

Team Members: Sandeep, Satheesh & Nasim

**Data cleaning Process and Automation for the ‘partial fulfillment’ of ‘Data Analytics Project’ showcase and submission**

------------------------------------------------------------------------------

**Data Cleaning Problem Statement:**

Sometimes our datasets have missing values. Machine learning algorithms don’t deal well with missing values.

**Solutions:**

**Solution 1:**

Drop each feature which contains missing values (drop the column)

**Solution 2:**

Drop each entry which contains missing values (drop the row)

**Solution 3:**

Imputation (fill in the missing values)

**Imputation:**

Deal with missing data points by substituting new values.

Common strategy: replace each missing value in a feature with the mean, median, or mode of the feature.

**1. Get the data**

**From kaggle.com:**

<https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot>

**Meta Data**

**Columns details:**

**Rooms:** Number of rooms

**Price:** Price in dollars

**Method:** S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.

**Type:** br - bedroom(s); h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.

**SellerG:** Real Estate Agent

**Date:** Date sold

**Distance:** Distance from CBD In Kilometers

**Regionname:** General Region (West, North West, North, North east …etc)

**Propertycount:** Number of properties that exist in the suburb.

**Bedroom2 :** Scraped # of Bedrooms (from different source)

**Bathroom:** Number of Bathrooms

**Car:** Number of carspots

**Landsize:** Land Size in Metres

**BuildingArea:** Building Size in Metres

**YearBuilt:** Year the house was built

**CouncilArea:** Governing council for the area

**Latitude:** Self explanatory

**Longitude:** Self explanatory

**Import pandas**

****

**Read Data:**

****

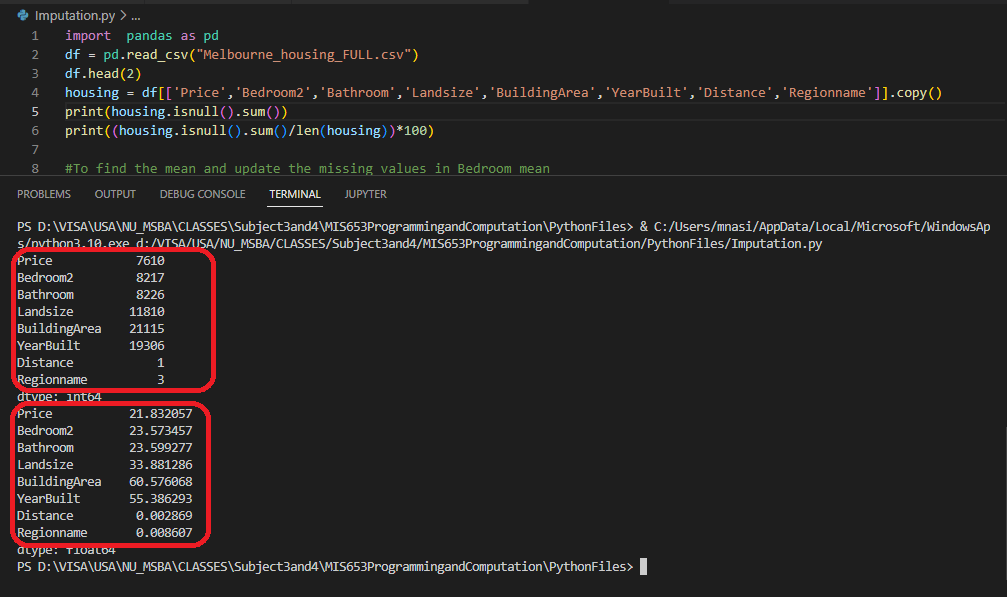
**Create subset of the data to work with**

****

**2. Explore the missing data (Completeness of Data)**

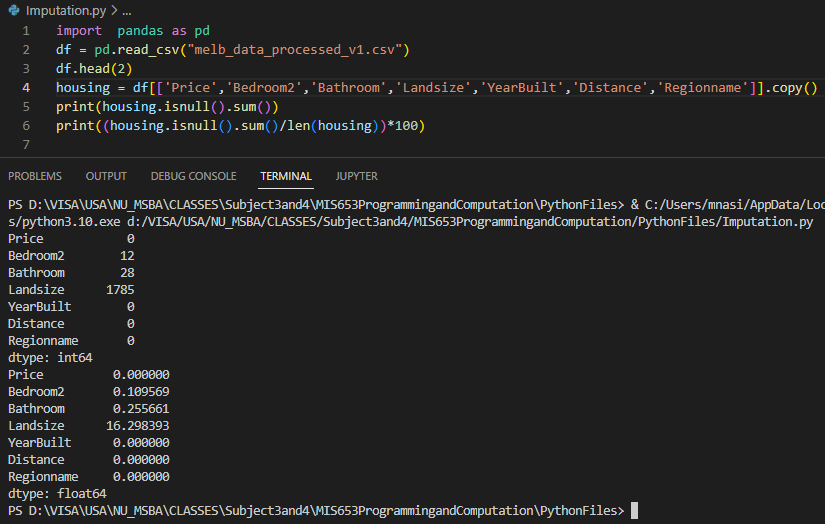
Melbourne\_housing\_FULL.csv contains: 34858 records.

**Examine missing data**

****

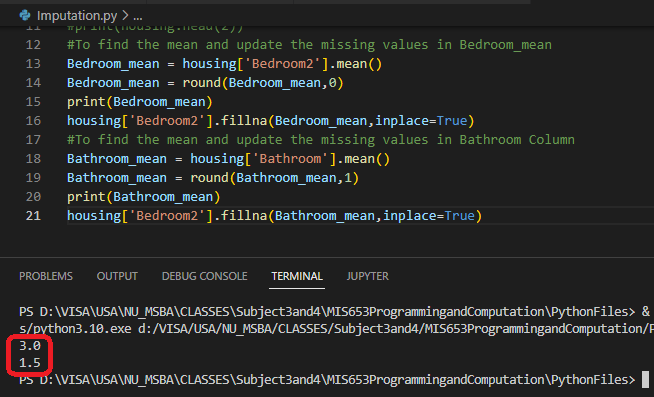
**Drop columns with more than 25% of missing data**

****

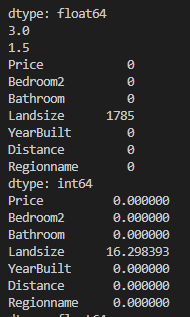
****

**3. Impute Substitute Values**

**Strategy 1: Impute Mean**

****

**Check for missing data**

****

**Future Learning and implementation**

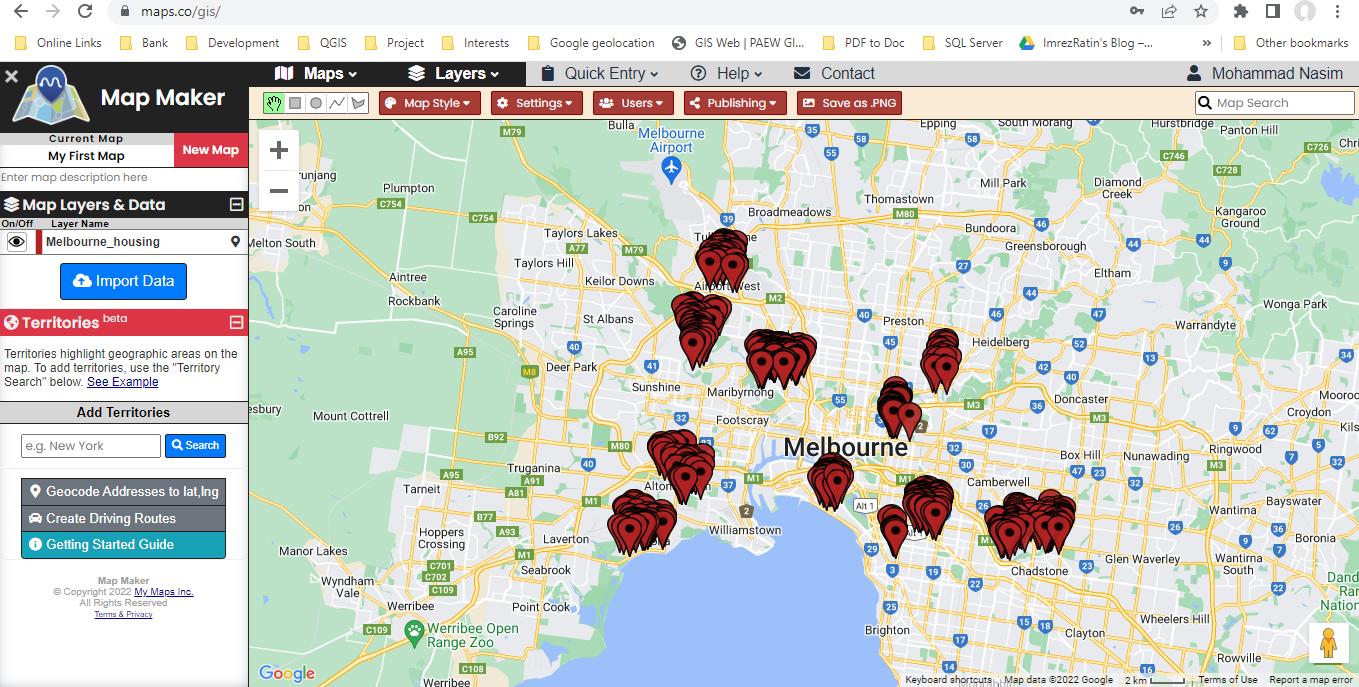
Scikit-Learn Application:

Scikit-learn include imputation functionality in its preprocessing module.

http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Imputer.html

**3. Ensuring Random Normal Distribution of the data**

**Data Visualization in GIS.**

****

A sample Data of 500 records, just to show the skewness of the records.

Published Link : <https://maps.co/map/638ddc953373e450601462wrh44e594>

**4. Conduct your data analysis to clarify your hypotheses**

**5. Test your hypotheses.**

**6. Develop Alternative Solutions and their Probable Consequences (Do a Cost/Benefit Analysis for each)**

**7. Decide/Recommend which Solution to choose.**

**8. Identify measures for Success**

**9. Literature Review**