A1_31378

January 17, 2022

Name: Sourav Kotkar

Roll No: 31378

0.1 Assignment-1 : Data Wrangling I

Perform the following operations using Python on any open-source dataset (e.g., data.csv)

1. Import all the required Python Libraries.

2. Locate an open-source data from the web (e.g. https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).

3. Load the Dataset into pandas' data frame.

4. Data Preprocessing: check for missing values in the data using pandas insult (), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.

5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.

6. Turn categorical variables into quantitative variables in Python.

In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set.

0.1.1 Importing libraries

[1]:
```
import pandas as pd
```

0.1.2 Loading the dataset

[2]:
```
df = pd.read_csv('melb_data.csv') #Loading the dataset
```

0.1.3 Data Preprocessing

[3]:
```
df.head(4) #returns first n rows
```

[3]:
```
       Suburb           Address  Rooms Type      Price Method SellerG  \
0  Abbotsford      85 Turner St      2    h  1480000.0      S  Biggin
1  Abbotsford   25 Bloomburg St      2    h  1035000.0      S  Biggin
2  Abbotsford      5 Charles St      3    h  1465000.0     SP  Biggin
3  Abbotsford  40 Federation La      3    h   850000.0     PI  Biggin

         Date  Distance  Postcode  …  Bathroom  Car  Landsize  BuildingArea  \
0  3/12/2016       2.5    3067.0  …       1.0  1.0     202.0           NaN
1  4/02/2016       2.5    3067.0  …       1.0  0.0     156.0          79.0
2  4/03/2017       2.5    3067.0  …       2.0  0.0     134.0         150.0
3  4/03/2017       2.5    3067.0  …       2.0  1.0      94.0           NaN

   YearBuilt  CouncilArea Lattitude  Longtitude             Regionname  \
0        NaN        Yarra  -37.7996    144.9984  Northern Metropolitan
1     1900.0        Yarra  -37.8079    144.9934  Northern Metropolitan
2     1900.0        Yarra  -37.8093    144.9944  Northern Metropolitan
3        NaN        Yarra  -37.7969    144.9969  Northern Metropolitan

  Propertycount
0        4019.0
1        4019.0
2        4019.0
3        4019.0

[4 rows x 21 columns]
```

[4]:
```
df.tail(6) #returns last n rows
```

[4]:
```
             Suburb        Address  Rooms Type      Price Method   SellerG  \
13574    Westmeadows     9 Black St      3    h   582000.0      S       Red
13575  Wheelers Hill   12 Strada Cr      4    h  1245000.0      S     Barry
13576   Williamstown  77 Merrett Dr      3    h  1031000.0     SP  Williams
13577   Williamstown    83 Power St      3    h  1170000.0      S     Raine
13578   Williamstown   96 Verdon St      4    h  2500000.0     PI   Sweeney
13579     Yarraville     6 Agnes St      4    h  1285000.0     SP   Village

             Date  Distance  Postcode  …  Bathroom  Car  Landsize  \
13574  26/08/2017      16.5    3049.0  …       2.0  2.0     256.0
13575  26/08/2017      16.7    3150.0  …       2.0  2.0     652.0
13576  26/08/2017       6.8    3016.0  …       2.0  2.0     333.0
```

```
13577  26/08/2017       6.8    3016.0  …       2.0  4.0     436.0
13578  26/08/2017       6.8    3016.0  …       1.0  5.0     866.0
13579  26/08/2017       6.3    3013.0  …       1.0  1.0     362.0

       BuildingArea  YearBuilt  CouncilArea Lattitude  Longtitude  \
13574           NaN        NaN          NaN -37.67917   144.89390
13575           NaN     1981.0          NaN -37.90562   145.16761
13576         133.0     1995.0          NaN -37.85927   144.87904
13577           NaN     1997.0          NaN -37.85274   144.88738
13578         157.0     1920.0          NaN -37.85908   144.89299
13579         112.0     1920.0          NaN -37.81188   144.88449

                     Regionname Propertycount
13574       Northern Metropolitan        2474.0
13575  South-Eastern Metropolitan        7392.0
13576       Western Metropolitan        6380.0
13577       Western Metropolitan        6380.0
13578       Western Metropolitan        6380.0
13579       Western Metropolitan        6543.0

[6 rows x 21 columns]
```

[5]:
```
df.describe() #provides quick overview of the numerical data in a DataFrame
```

[5]:
```
              Rooms         Price      Distance      Postcode      Bedroom2  \
count  13580.000000  1.358000e+04  13580.000000  13580.000000  13580.000000
mean       2.937997  1.075684e+06     10.137776   3105.301915      2.914728
std        0.955748  6.393107e+05      5.868725     90.676964      0.965921
min        1.000000  8.500000e+04      0.000000   3000.000000      0.000000
25%        2.000000  6.500000e+05      6.100000   3044.000000      2.000000
50%        3.000000  9.030000e+05      9.200000   3084.000000      3.000000
75%        3.000000  1.330000e+06     13.000000   3148.000000      3.000000
max       10.000000  9.000000e+06     48.100000   3977.000000     20.000000

           Bathroom           Car       Landsize  BuildingArea    YearBuilt  \
count  13580.000000  13518.000000   13580.000000   7130.000000  8205.000000
mean       1.534242      1.610075     558.416127    151.967650  1964.684217
std        0.691712      0.962634    3990.669241    541.014538    37.273762
min        0.000000      0.000000       0.000000      0.000000  1196.000000
25%        1.000000      1.000000     177.000000     93.000000  1940.000000
50%        1.000000      2.000000     440.000000    126.000000  1970.000000
75%        2.000000      2.000000     651.000000    174.000000  1999.000000
max        8.000000     10.000000  433014.000000  44515.000000  2018.000000

          Lattitude    Longtitude  Propertycount
count  13580.000000  13580.000000   13580.000000
mean     -37.809203    144.995216    7454.417378
```

```
std        0.079260      0.103916    4378.581772
min      -38.182550    144.431810     249.000000
25%      -37.856822    144.929600    4380.000000
50%      -37.802355    145.000100    6555.000000
75%      -37.756400    145.058305   10331.000000
max      -37.408530    145.526350   21650.000000
```

[6]:
```
df.info() #provides technical info about dataframe
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13580 entries, 0 to 13579
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Suburb         13580 non-null  object
 1   Address        13580 non-null  object
 2   Rooms          13580 non-null  int64
 3   Type           13580 non-null  object
 4   Price          13580 non-null  float64
 5   Method         13580 non-null  object
 6   SellerG        13580 non-null  object
 7   Date           13580 non-null  object
 8   Distance       13580 non-null  float64
 9   Postcode       13580 non-null  float64
 10  Bedroom2       13580 non-null  float64
 11  Bathroom       13580 non-null  float64
 12  Car            13518 non-null  float64
 13  Landsize       13580 non-null  float64
 14  BuildingArea   7130 non-null   float64
 15  YearBuilt      8205 non-null   float64
 16  CouncilArea    12211 non-null  object
 17  Lattitude      13580 non-null  float64
 18  Longtitude     13580 non-null  float64
 19  Regionname     13580 non-null  object
 20  Propertycount  13580 non-null  float64
dtypes: float64(12), int64(1), object(8)
memory usage: 2.2+ MB
```

[7]:
```
df.size #returns total number of elements
```

[7]:
```
285180
```

[8]:
```
df.shape #returns a tuple representing the dimensionality of the DataFrame
```

[8]:
```
(13580, 21)
```

[9]:
```
df['Car'].shape
```

[9]: (13580,)

[10]:
```
df['Address'].shape
```

[10]: (13580,)

[11]:
```
df.isnull() #detect missing values for an array-like object
```

[11]:

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	\
0	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	
…	…	…	…	…	…	…	…	…		
13575	False	False	False	False	False	False	False	False	False	
13576	False	False	False	False	False	False	False	False	False	
13577	False	False	False	False	False	False	False	False	False	
13578	False	False	False	False	False	False	False	False	False	
13579	False	False	False	False	False	False	False	False	False	

	Postcode	…	Bathroom	Car	Landsize	BuildingArea	YearBuilt	\
0	False	…	False	False	False	True	True	
1	False	…	False	False	False	False	False	
2	False	…	False	False	False	False	False	
3	False	…	False	False	False	True	True	
4	False	…	False	False	False	False	False	
…	…	…	…	…	…	…	…	
13575	False	…	False	False	False	True	False	
13576	False	…	False	False	False	False	False	
13577	False	…	False	False	False	True	False	
13578	False	…	False	False	False	False	False	
13579	False	…	False	False	False	False	False	

	CouncilArea	Lattitude	Longtitude	Regionname	Propertycount
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
…	…	…	…	…	…
13575	True	False	False	False	False
13576	True	False	False	False	False
13577	True	False	False	False	False
13578	True	False	False	False	False
13579	True	False	False	False	False

```
[13580 rows x 21 columns]
```

[12]:
```
df.isnull().sum()
```

[12]:
```
Suburb             0
Address            0
Rooms              0
Type               0
Price              0
Method             0
SellerG            0
Date               0
Distance           0
Postcode           0
Bedroom2           0
Bathroom           0
Car               62
Landsize           0
BuildingArea    6450
YearBuilt       5375
CouncilArea     1369
Lattitude          0
Longtitude         0
Regionname         0
Propertycount      0
dtype: int64
```

[13]:
```
df['Car'].fillna(value=df['Car'].mean(), inplace=True) #Replace null values by␣
↪mean
df['BuildingArea'].fillna(value=df['BuildingArea'].mean(), inplace=True)
df['YearBuilt'].fillna(value=df['YearBuilt'].mean(), inplace=True)
```

[14]:
```
df.dropna(subset = ["CouncilArea"], inplace=True) #Remove rows with null values
```

[15]:
```
df.isnull().sum()
```

[15]:
```
Suburb          0
Address         0
Rooms           0
Type            0
Price           0
Method          0
SellerG         0
Date            0
Distance        0
Postcode        0
Bedroom2        0
```

```
Bathroom         0
Car              0
Landsize         0
BuildingArea     0
YearBuilt        0
CouncilArea      0
Lattitude        0
Longtitude       0
Regionname       0
Propertycount    0
dtype: int64
```

[16]:
```
df.dtypes #returns datatypes
```

[16]:
```
Suburb            object
Address           object
Rooms              int64
Type              object
Price            float64
Method            object
SellerG           object
Date              object
Distance         float64
Postcode         float64
Bedroom2         float64
Bathroom         float64
Car              float64
Landsize         float64
BuildingArea     float64
YearBuilt        float64
CouncilArea       object
Lattitude        float64
Longtitude       float64
Regionname        object
Propertycount    float64
dtype: object
```

0.1.4 Data Formatting

[17]:
```
#Type conversion
df['Price'] = df['Price'].astype(int)
df['Bathroom'] = df['Bathroom'].astype(int)
df['Bedroom2'] = df['Bedroom2'].astype(int)
df['Propertycount'] = df['Propertycount'].astype(int)
```

[18]:
```
df.dtypes
```

[18]:
```
Suburb            object
Address           object
Rooms              int64
Type              object
Price              int32
Method            object
SellerG           object
Date              object
Distance         float64
Postcode         float64
Bedroom2           int32
Bathroom           int32
Car              float64
Landsize         float64
BuildingArea     float64
YearBuilt        float64
CouncilArea       object
Lattitude        float64
Longtitude       float64
Regionname        object
Propertycount      int32
dtype: object
```

0.1.5 Data Normalization

[19]:
```
normalized_df = df.copy()
normalized_df['Price'] = normalized_df['Price'] / normalized_df['Price'].abs().
 ↪max()
```

[20]:
```
normalized_df
```

[20]:

	Suburb	Address	Rooms	Type	Price	Method	\
0	Abbotsford	85 Turner St	2	h	0.164444	S	
1	Abbotsford	25 Bloomburg St	2	h	0.115000	S	
2	Abbotsford	5 Charles St	3	h	0.162778	SP	
3	Abbotsford	40 Federation La	3	h	0.094444	PI	
4	Abbotsford	55a Park St	4	h	0.177778	VB	
…	…	…	…	…	…	…	
12208	Williamstown	87 Pasco St	3	h	0.142778	S	
12209	Windsor	201/152 Peel St	2	u	0.062222	PI	
12210	Wollert	60 Saltlake Bvd	3	h	0.058367	S	
12211	Yarraville	2 Adeney St	2	h	0.083333	SP	
12212	Yarraville	54 Pentland Pde	6	h	0.272222	VB	

	SellerG	Date	Distance	Postcode	…	Bathroom	Car	\
0	Biggin	3/12/2016	2.5	3067.0	…	1	1.0	
1	Biggin	4/02/2016	2.5	3067.0	…	1	0.0	

```
2             Biggin   4/03/2017       2.5    3067.0  …         2  0.0
3             Biggin   4/03/2017       2.5    3067.0  …         2  1.0
4             Nelson   4/06/2016       2.5    3067.0  …         1  2.0
…               …         …       …       …  …        …  …
12208           Jas  29/07/2017       6.8    3016.0  …         1  0.0
12209  hockingstuart  29/07/2017       4.6    3181.0  …         1  1.0
12210      Stockdale  29/07/2017      25.5    3750.0  …         2  2.0
12211  hockingstuart  29/07/2017       6.3    3013.0  …         1  2.0
12212        Village  29/07/2017       6.3    3013.0  …         3  2.0

       Landsize  BuildingArea    YearBuilt  CouncilArea Lattitude  Longtitude  \
0         202.0     151.96765  1964.684217        Yarra -37.79960   144.99840
1         156.0      79.00000  1900.000000        Yarra -37.80790   144.99340
2         134.0     150.00000  1900.000000        Yarra -37.80930   144.99440
3          94.0     151.96765  1964.684217        Yarra -37.79690   144.99690
4         120.0     142.00000  2014.000000        Yarra -37.80720   144.99410
…           …           …           …       …        …         …
12208     296.0     151.96765  1964.684217  Hobsons Bay -37.86335   144.89487
12209       0.0      61.60000  2012.000000  Stonnington -37.85581   144.99025
12210     400.0     151.96765  1964.684217   Whittlesea -37.61387   145.03850
12211     269.0     151.96765  1964.684217  Maribyrnong -37.81649   144.86731
12212    1087.0     388.50000  1920.000000  Maribyrnong -37.81038   144.89389

                 Regionname Propertycount
0      Northern Metropolitan          4019
1      Northern Metropolitan          4019
2      Northern Metropolitan          4019
3      Northern Metropolitan          4019
4      Northern Metropolitan          4019
…                        …           …
12208   Western Metropolitan          6380
12209  Southern Metropolitan          4380
12210  Northern Metropolitan          2940
12211   Western Metropolitan          6543
12212   Western Metropolitan          6543

[12211 rows x 21 columns]
```

0.1.6 Categorical variables to Quantitative variables

[21]:
```
df.Type.unique() #returns unique values for a column
```

[21]:
```
array(['h', 'u', 't'], dtype=object)
```

[22]:
```
df.Method.unique()
```

[22]:
```
array(['S', 'SP', 'PI', 'VB', 'SA'], dtype=object)
```

[23]:
```
df2 = df.copy() #copy dataframe
```

[24]:
```
df2 = pd.get_dummies(df2, columns=['Type','Method']) #Categorical variables to␣
↪Quantitative variables
```

[25]:
```
df2
```

[25]:

	Suburb	Address	Rooms	Price	SellerG	\
0	Abbotsford	85 Turner St	2	1480000	Biggin	
1	Abbotsford	25 Bloomburg St	2	1035000	Biggin	
2	Abbotsford	5 Charles St	3	1465000	Biggin	
3	Abbotsford	40 Federation La	3	850000	Biggin	
4	Abbotsford	55a Park St	4	1600000	Nelson	
…	…	…	…	…	…	
12208	Williamstown	87 Pasco St	3	1285000	Jas	
12209	Windsor	201/152 Peel St	2	560000	hockingstuart	
12210	Wollert	60 Saltlake Bvd	3	525300	Stockdale	
12211	Yarraville	2 Adeney St	2	750000	hockingstuart	
12212	Yarraville	54 Pentland Pde	6	2450000	Village	

	Date	Distance	Postcode	Bedroom2	Bathroom	…	\
0	3/12/2016	2.5	3067.0	2	1	…	
1	4/02/2016	2.5	3067.0	2	1	…	
2	4/03/2017	2.5	3067.0	3	2	…	
3	4/03/2017	2.5	3067.0	3	2	…	
4	4/06/2016	2.5	3067.0	3	1	…	
…	…	…	…	…	…	…	
12208	29/07/2017	6.8	3016.0	3	1	…	
12209	29/07/2017	4.6	3181.0	2	1	…	
12210	29/07/2017	25.5	3750.0	3	2	…	
12211	29/07/2017	6.3	3013.0	2	1	…	
12212	29/07/2017	6.3	3013.0	6	3	…	

	Regionname	Propertycount	Type_h	Type_t	Type_u	Method_PI	\
0	Northern Metropolitan	4019	1	0	0	0	
1	Northern Metropolitan	4019	1	0	0	0	
2	Northern Metropolitan	4019	1	0	0	0	
3	Northern Metropolitan	4019	1	0	0	1	
4	Northern Metropolitan	4019	1	0	0	0	
…	…	…	…	…	…	…	
12208	Western Metropolitan	6380	1	0	0	0	
12209	Southern Metropolitan	4380	0	0	1	1	
12210	Northern Metropolitan	2940	1	0	0	0	
12211	Western Metropolitan	6543	1	0	0	0	
12212	Western Metropolitan	6543	1	0	0	0	

	Method_S	Method_SA	Method_SP	Method_VB

```
0             1         0          0          0
1             1         0          0          0
2             0         0          1          0
3             0         0          0          0
4             0         0          0          1
…           …         …          …          …
12208         1         0          0          0
12209         0         0          0          0
12210         1         0          0          0
12211         0         0          1          0
12212         0         0          0          1

[12211 rows x 27 columns]
```

[ ]: