exam1

March 11, 2024

[ ]:
```
import pandas as pd
df = pd.read_csv('melb_data.csv')
```

```
C:\Users\woosh\AppData\Local\Temp\ipykernel_10176\267251470.py:1:
DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of
pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better
interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

  import pandas as pd
```

[ ]:
```
df.head()
```

[ ]:

	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	
4	Abbotsford	55a Park St	4	h	1600000.0	VB	Nelson	

	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	
4	4/06/2016	2.5	3067.0	…	1.0	2.0	120.0	142.0	

	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	
4	2014.0	Yarra	-37.8072	144.9941	Northern Metropolitan	

```
  Propertycount
0        4019.0
1        4019.0
2        4019.0
3        4019.0
4        4019.0

[5 rows x 21 columns]
```

[ ]:
```
print(df.size)
print(df.shape)
print(df.ndim)
print(df.info())
```

```
285180
(13580, 21)
2
<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
None
```

[ ]:
```
print(df.describe())
```

```
              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
```

[ ]:
```
df.loc[:,['Price','Landsize','Propertycount']].describe()
```

[ ]:
```
              Price       Landsize  Propertycount
count  1.358000e+04   13580.000000   13580.000000
mean   1.075684e+06     558.416127    7454.417378
std    6.393107e+05    3990.669241    4378.581772
min    8.500000e+04       0.000000     249.000000
25%    6.500000e+05     177.000000    4380.000000
50%    9.030000e+05     440.000000    6555.000000
75%    1.330000e+06     651.000000   10331.000000
max    9.000000e+06  433014.000000   21650.000000
```

[ ]:
```
df.loc[(df['Landsize']<500)]
```

[ ]:
```
             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
4        Abbotsford       55a Park St      4    h  1600000.0     VB    Nelson
…               …                …    …  …       …    …       …
13572      Watsonia     76 Kenmare St      2    h   650000.0     PI  Morrison
13574   Westmeadows        9 Black St      3    h   582000.0      S       Red
13576  Williamstown     77 Merrett Dr      3    h  1031000.0     SP  Williams
13577  Williamstown       83 Power St      3    h  1170000.0      S     Raine
13579    Yarraville        6 Agnes St      4    h  1285000.0     SP   Village

              Date  Distance  Postcode  …  Bathroom  Car  Landsize  \
0        3/12/2016       2.5    3067.0  …       1.0  1.0     202.0
1        4/02/2016       2.5    3067.0  …       1.0  0.0     156.0
2        4/03/2017       2.5    3067.0  …       2.0  0.0     134.0
3        4/03/2017       2.5    3067.0  …       2.0  1.0      94.0
4        4/06/2016       2.5    3067.0  …       1.0  2.0     120.0
…              …       …       …  …       …  …       …
13572  26/08/2017      14.5    3087.0  …       1.0  1.0     210.0
13574  26/08/2017      16.5    3049.0  …       2.0  2.0     256.0
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
13579  26/08/2017       6.3    3013.0  …       1.0  1.0     362.0

       BuildingArea  YearBuilt  CouncilArea Lattitude  Longtitude  \
0               NaN        NaN        Yarra -37.79960   144.99840
1              79.0     1900.0        Yarra -37.80790   144.99340
2             150.0     1900.0        Yarra -37.80930   144.99440
3               NaN        NaN        Yarra -37.79690   144.99690
4             142.0     2014.0        Yarra -37.80720   144.99410
…               …        …          …       …          …
13572          79.0     2006.0          NaN -37.70657   145.07878
13574           NaN        NaN          NaN -37.67917   144.89390
13576         133.0     1995.0          NaN -37.85927   144.87904
13577           NaN     1997.0          NaN -37.85274   144.88738
13579         112.0     1920.0          NaN -37.81188   144.88449

                  Regionname Propertycount
0      Northern Metropolitan        4019.0
1      Northern Metropolitan        4019.0
2      Northern Metropolitan        4019.0
3      Northern Metropolitan        4019.0
4      Northern Metropolitan        4019.0
…                        …           …
13572  Northern Metropolitan        2329.0
13574  Northern Metropolitan        2474.0
13576   Western Metropolitan        6380.0
13577   Western Metropolitan        6380.0
13579   Western Metropolitan        6543.0
```

```
[7392 rows x 21 columns]
```

[ ]:
```
df.loc[(df['Bedroom2'] == 2) & (df['Bathroom'] == 1) & (df['Car'] == 1)]
```

[ ]:
```
             Suburb             Address  Rooms Type      Price Method  \
0         Abbotsford        85 Turner St      2    h  1480000.0      S
13        Abbotsford       45 William St      2    h  1172500.0      S
17        Abbotsford         78 Yarra St      3    h  1176500.0      S
19        Abbotsford       42 Valiant St      2    h   890000.0      S
23        Abbotsford  6/219 Nicholson St      2    u   500000.0      S
…                …                  …    …  …         …    …
13482  Malvern East     2002 Malvern Rd      2    u   651000.0     SP
13495  Moonee Ponds     1/53 Buckley St      2    u   435000.0      S
13510    Nunawading       3/39 Lemon Gr      2    u   710000.0      S
13511      Oak Park        18 Jessie St      2    h  1006000.0      S
13572      Watsonia       76 Kenmare St      2    h   650000.0     PI

          SellerG        Date  Distance  Postcode  …  Bathroom  Car  \
0          Biggin   3/12/2016       2.5    3067.0  …       1.0  1.0
13         Biggin  13/08/2016       2.5    3067.0  …       1.0  1.0
17         LITTLE  16/07/2016       2.5    3067.0  …       1.0  1.0
19         Biggin  17/09/2016       2.5    3067.0  …       1.0  1.0
23        Collins  18/06/2016       2.5    3067.0  …       1.0  1.0
…               …           …       …         …  …         …  …
13482      Jellis  26/08/2017       8.4    3145.0  …       1.0  1.0
13495      Nelson  26/08/2017       6.2    3039.0  …       1.0  1.0
13510      Jellis  26/08/2017      15.4    3131.0  …       1.0  1.0
13511   Stockdale  26/08/2017      11.2    3046.0  …       1.0  1.0
13572    Morrison  26/08/2017      14.5    3087.0  …       1.0  1.0

       Landsize  BuildingArea  YearBuilt  CouncilArea Lattitude  Longtitude  \
0         202.0           NaN        NaN        Yarra -37.79960   144.99840
13        195.0           NaN        NaN        Yarra -37.80840   144.99730
17        138.0         105.0     1890.0        Yarra -37.80210   144.99650
19        150.0          73.0     1985.0        Yarra -37.80110   145.00040
23          0.0          60.0     1970.0        Yarra -37.80150   144.99720
…           …             …          …            …         …           …
13482     129.0          97.0     1940.0          NaN -37.87798   145.06731
13495    1475.0          66.0     1970.0          NaN -37.75799   144.92354
13510     903.0           NaN     1985.0          NaN -37.80640   145.18452
13511     716.0           NaN        NaN          NaN -37.71589   144.92176
13572     210.0          79.0     2006.0          NaN -37.70657   145.07878

                Regionname Propertycount
0      Northern Metropolitan        4019.0
13     Northern Metropolitan        4019.0
```

```
17     Northern Metropolitan        4019.0
19     Northern Metropolitan        4019.0
23     Northern Metropolitan        4019.0
…                        …           …
13482  Southern Metropolitan        8801.0
13495   Western Metropolitan        6232.0
13510   Eastern Metropolitan        4973.0
13511  Northern Metropolitan        2651.0
13572  Northern Metropolitan        2329.0

[2137 rows x 21 columns]
```

[ ]:
```
drop_row = df.dropna()
print(drop_row.shape)
print(drop_row.info())
```

```
(6196, 21)
<class 'pandas.core.frame.DataFrame'>
Index: 6196 entries, 1 to 12212
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Suburb         6196 non-null   object
 1   Address        6196 non-null   object
 2   Rooms          6196 non-null   int64
 3   Type           6196 non-null   object
 4   Price          6196 non-null   float64
 5   Method         6196 non-null   object
 6   SellerG        6196 non-null   object
 7   Date           6196 non-null   object
 8   Distance       6196 non-null   float64
 9   Postcode       6196 non-null   float64
 10  Bedroom2       6196 non-null   float64
 11  Bathroom       6196 non-null   float64
 12  Car            6196 non-null   float64
 13  Landsize       6196 non-null   float64
 14  BuildingArea   6196 non-null   float64
 15  YearBuilt      6196 non-null   float64
 16  CouncilArea    6196 non-null   object
 17  Lattitude      6196 non-null   float64
 18  Longtitude     6196 non-null   float64
 19  Regionname     6196 non-null   object
 20  Propertycount  6196 non-null   float64
dtypes: float64(12), int64(1), object(8)
memory usage: 1.0+ MB
None
```

[ ]:
```
df.isnull().any()
```

[ ]:
```
Suburb           False
Address          False
Rooms            False
Type             False
Price            False
Method           False
SellerG          False
Date             False
Distance         False
Postcode         False
Bedroom2         False
Bathroom         False
Car               True
Landsize         False
BuildingArea      True
YearBuilt         True
CouncilArea       True
Lattitude        False
Longtitude       False
Regionname       False
Propertycount    False
dtype: bool
```

[ ]:
```
fill_zero = df.fillna(0)
df.isnull().any()
print(df.loc[:,['Car','BuildingArea','YearBuilt']].mean())
print(fill_zero.loc[:,['Car','BuildingArea','YearBuilt']].mean())
# I think it is not reasonable because year that was calculated was inaccurate.
```

```
Car               1.610075
BuildingArea    151.967650
YearBuilt      1964.684217
dtype: float64
Car               1.602725
BuildingArea     79.788611
YearBuilt      1187.056996
dtype: float64
```

[ ]:
```
print(df.select_dtypes(include='number').mean())
# Impute missing values for numeric columns using the average
numeric_columns = df.select_dtypes(include='number').columns
df[numeric_columns] = df[numeric_columns].fillna(df[numeric_columns].mean())
print(df.select_dtypes(include='number').mean())
# Remove rows with missing values for non-numeric columns
df = df.dropna(subset=df.select_dtypes(exclude='number').columns)
# I think it is quite impressive by seeing that mean are all the same.. Should␣
 ↪be used because data is stll accurate
```

```
Rooms             2.937997e+00
Price             1.075684e+06
Distance          1.013778e+01
Postcode          3.105302e+03
Bedroom2          2.914728e+00
Bathroom          1.534242e+00
Car               1.610075e+00
Landsize          5.584161e+02
BuildingArea      1.519676e+02
YearBuilt         1.964684e+03
Lattitude        -3.780920e+01
Longtitude        1.449952e+02
Propertycount     7.454417e+03
dtype: float64
Rooms             2.937997e+00
Price             1.075684e+06
Distance          1.013778e+01
Postcode          3.105302e+03
Bedroom2          2.914728e+00
Bathroom          1.534242e+00
Car               1.610075e+00
Landsize          5.584161e+02
BuildingArea      1.519676e+02
YearBuilt         1.964684e+03
Lattitude        -3.780920e+01
Longtitude        1.449952e+02
Propertycount     7.454417e+03
dtype: float64
```

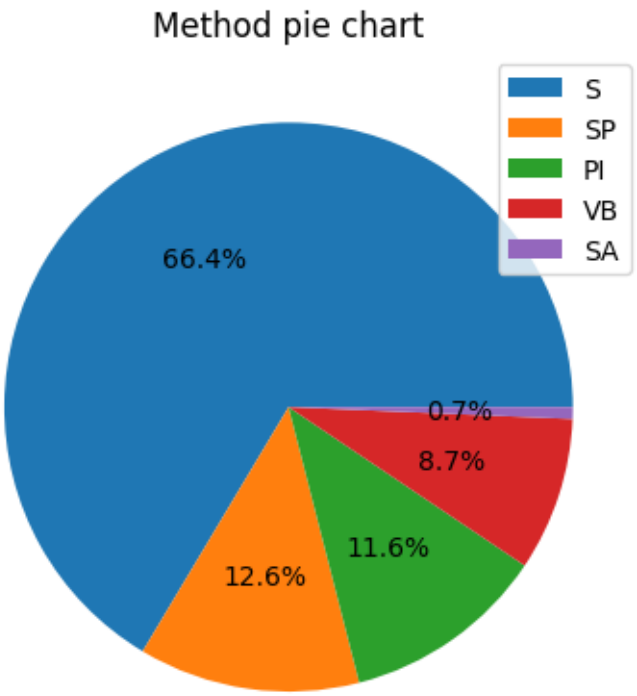
[ ]:
```
df['Date'] = pd.to_datetime(df['Date'],format = 'mixed')
df['Date'] = df['Date'].dt.strftime('%Y/%m/%d')
```

[ ]:
```
df['Date'].head()
```

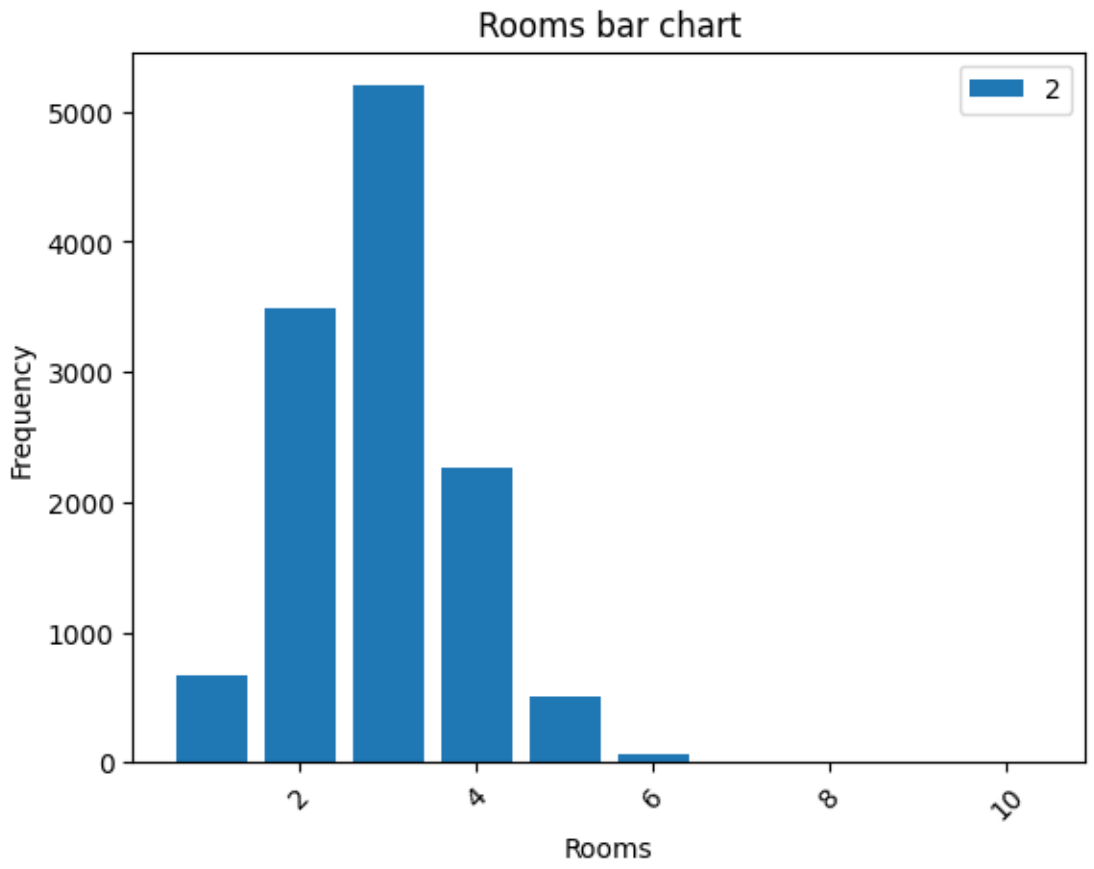
[ ]:
```
0    2016/03/12
1    2016/04/02
2    2017/04/03
3    2017/04/03
4    2016/04/06
Name: Date, dtype: object
```

[ ]:
```
import matplotlib.pyplot as plt

plt.pie(df['Method'].value_counts(),autopct='%1.1f%%')
plt.title('Method pie chart')
plt.legend(df['Method'].value_counts().index)
plt.show()
```



[ ]:
```
plt.bar(df['Rooms'].unique(),df['Rooms'].value_counts(sort=False))
plt.title('Rooms bar chart')
plt.xlabel('Rooms')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.legend(df['Rooms'].value_counts(sort=False).index)
plt.show()
```



[ ]:
```
df.
  ↪groupby(['Regionname','Type'])[['Price','Bedroom2','Bathroom','Car','Landsize']].
  ↪sum()
```

[ ]:

Regionname	Type	Price	Bedroom2	Bathroom	Car	\
Eastern Metropolitan	h	1.133086e+09	3261.0	1636.0	1763.0	
	t	9.422715e+07	328.0	198.0	180.0	
	u	1.095653e+08	399.0	211.0	217.0	
Eastern Victoria	h	2.889698e+07	142.0	78.0	88.0	
	u	1.384000e+06	8.0	3.0	4.0	
Northern Metropolitan	h	2.528641e+09	7352.0	3439.0	3789.0	
	t	2.139878e+08	720.0	461.0	375.0	
	u	4.441437e+08	1529.0	946.0	878.0	
Northern Victoria	h	1.455350e+07	91.0	46.0	47.0	
South-Eastern Metropolitan	h	2.575009e+08	963.0	467.0	573.0	
	t	1.596875e+07	49.0	30.0	30.0	
	u	1.952750e+07	79.0	43.0	47.0	
Southern Metropolitan	h	4.317675e+09	7990.0	4366.0	4307.0	

```
                           t     4.768569e+08    1175.0     774.0   693.0
                           u     1.015503e+09    2955.0    1818.0  1724.0
Western Metropolitan       h     1.943246e+09    6391.0    3072.0  3634.0
                           t     1.657553e+08     661.0     421.0   350.0
                           u     1.986504e+08     850.0     482.0   473.0
Western Victoria           h     9.572750e+06      83.0      38.0    49.0
```

```
                                  Landsize
Regionname                 Type
Eastern Metropolitan       h      675951.0
                           t       28051.0
                           u       53761.0
Eastern Victoria           h      147098.0
                           u         886.0
Northern Metropolitan      h     1537299.0
                           t       89956.0
                           u      385133.0
Northern Victoria          h       62746.0
South-Eastern Metropolitan h      177101.0
                           t        3742.0
                           u       11829.0
Southern Metropolitan      h     1360338.0
                           t      104129.0
                           u      706538.0
Western Metropolitan       h     1011765.0
                           t       56754.0
                           u      226520.0
Western Victoria           h       15942.0
```

[ ]: