Gerry William Nanlohy / 2440018854 / LB01 / Assignment 02

Importing Data from Various Sources

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

In [2]:
```
import numpy as np
```

In [3]:
```
import matplotlib.pyplot as plt
```

Reading Excel Files

In [4]:
```
dataset = pd.read_csv('listings.csv')
```

In [5]:
```
dataset.head()
```

Out[5]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	a
0	49091	COZICOMFORT LONG TERM STAY ROOM 2	266763	Francesca	North Region	Woodlands	1.44255	103.79580	Private room	83	180	1	2013-10-21	0.01	2	
1	50646	Pleasant Room along Bukit Timah	227796	Sujatha	Central Region	Bukit Timah	1.33235	103.78521	Private room	81	90	18	2014-12-26	0.28	1	
2	56334	COZICOMFORT	266763	Francesca	North Region	Woodlands	1.44246	103.79667	Private room	69	6	20	2015-10-01	0.20	2	
3	71609	Ensuite Room (Room 1 & 2) near EXPO	367042	Belinda	East Region	Tampines	1.34541	103.95712	Private room	206	1	14	2019-08-11	0.15	9	
4	71896	B&B Room 1 near Airport & EXPO	367042	Belinda	East Region	Tampines	1.34567	103.95963	Private room	94	1	22	2019-07-28	0.22	9	

In [6]:
```
dataset.shape
```

Out[6]:
```
(7907, 16)
```

1. Extracting Independent Variables & Dependent Variables

Independent Variables

In [7]:
```
x = dataset.iloc[:, :-1].values
```

In [8]:
```
x
```

Out[8]:
```
array([[49091, 'COZICOMFORT LONG TERM STAY ROOM 2', 266763, ...,
        '2013-10-21', 0.01, 2],
       [50646, 'Pleasant Room along Bukit Timah', 227796, ...,
        '2014-12-26', 0.28, 1],
       [56334, 'COZICOMFORT', 266763, ..., '2015-10-01', 0.2, 2],
       ...,
       [39109336, '[ Farrer Park ] New City Fringe CBD Mins to MRT',
        2814488565, ..., nan, nan, 3],
       [38110493, 'Cheap Master Room in Central of Singapore', 243835202,
        ..., nan, nan, 2],
       [38112762, 'Amazing room with private bathroom walk to Orchard',
        28788520, ..., nan, nan, 7]], dtype=object)
```

Dependent Variables

In [9]:
```
y = dataset.iloc[:, 10].values
```

In [10]:
```
y
```

Out[10]:
```
array([180,  90,   6, ...,  30,  14,  90], dtype=int64)
```

2. Handling missing data

(replacing missing data with the mean value)

In [11]:
```
from sklearn.impute import SimpleImputer
```

In [12]:
```
imputer = SimpleImputer(missing_values = np.nan,strategy = 'mean')
```

In [13]:
```
selection_convert = dataset.select_dtypes(np.number)
```

In [14]:
```
imputer = imputer.fit(selection_convert)
```

In [15]:
```
selection_convert = imputer.transform(selection_convert)
```

In [16]:
```
selection_convert
```

Out[16]:
```
array([[4.90910000e+04, 2.66763000e+05, 1.44255000e+00, ...,
        1.00000000e-02, 2.00000000e+00, 3.65000000e+02],
       [5.06460000e+04, 2.27796000e+05, 1.33235000e+00, ...,
        2.80000000e-01, 1.00000000e+00, 3.65000000e+02],
       [5.63340000e+04, 2.66763000e+05, 1.44246000e+00, ...,
        2.00000000e-01, 2.00000000e+00, 3.65000000e+02],
       ...,
       [3.91093360e+07, 2.81448565e+08, 1.31286000e+00, ...,
        1.04366867e+00, 3.00000000e+00, 1.73000000e+02],
       [3.81104930e+07, 2.43835202e+08, 1.29543000e+00, ...,
        1.04366867e+00, 2.00000000e+00, 3.60000000e+01],
       [3.81127620e+07, 2.87885200e+07, 1.29672000e+00, ...,
        1.04366867e+00, 7.00000000e+00, 3.65000000e+02]])
```

In [31]:
```
dataset.shape
```

Out[31]:
```
(7907, 16)
```

3. Encoding Categorical data for neighbourhood_group variable & room_type variable

In [17]:
```
from sklearn.preprocessing import LabelEncoder
```

In [18]:
```
labelencoder = LabelEncoder()
```

neighbourhood_group variable

In [19]:
```
dataset['neighbourhood_group'] = labelencoder.fit_transform(dataset['neighbourhood_group'
```

```
  File "C:\Users\tekni\AppData\Local\Temp/ipykernel_6740/3358154829.py", line 1
    dataset['neighbourhood_group'] = labelencoder.fit_transform(dataset['neighbourhood_group'
                                                                                            ^
SyntaxError: unexpected EOF while parsing
```

In [20]:
```
dataset
```

Out[20]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings
0	49091	COZICOMFORT LONG TERM STAY ROOM 2	266763	Francesca	North Region	Woodlands	1.44255	103.79580	Private room	83	180	1	2013-10-21	0.01	
1	50646	Pleasant Room along Bukit Timah	227796	Sujatha	Central Region	Bukit Timah	1.33235	103.78521	Private room	81	90	18	2014-12-26	0.28	
2	56334	COZICOMFORT	266763	Francesca	North Region	Woodlands	1.44246	103.79667	Private room	69	6	20	2015-10-01	0.20	
3	71609	Ensuite Room (Room 1 & 2) near EXPO	367042	Belinda	East Region	Tampines	1.34541	103.95712	Private room	206	1	14	2019-08-11	0.15	
4	71896	B&B Room 1 near Airport & EXPO	367042	Belinda	East Region	Tampines	1.34567	103.95963	Private room	94	1	22	2019-07-28	0.22	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7902	38105126	Loft 2 pax near Haw Par / Pasir Panjang. Free ...	278109833	Belle	Central Region	Queenstown	1.27973	103.78751	Entire home/apt	100	3	0	NaN	NaN	
7903	38108273	3bedroom luxury at Orchard	238891646	Neha	Central Region	Tanglin	1.29269	103.82623	Entire home/apt	550	6	0	NaN	NaN	
7904	38109336	[Farrer Park] New City Fringe CBD Mins to MRT	2814488565	Mindy	Central Region	Kallang	1.31286	103.85996	Private room	58	30	0	NaN	NaN	
7905	38110493	Cheap Master Room in Central of Singapore	243835202	Huang	Central Region	River Valley	1.29543	103.83801	Private room	56	14	0	NaN	NaN	
7906	38112762	Amazing room with private bathroom walk to Orc...	28788520	Terence	Central Region	River Valley	1.29672	103.83325	Private room	65	90	0	NaN	NaN	

7907 rows × 16 columns

In [32]:
```
dataset.shape
```

Out[32]:
```
(7907, 16)
```

room_type variable

In [21]:
```
dataset['room_type'] = labelencoder.fit_transform(dataset['room_type'])
```

In [22]:
```
dataset
```

Out[22]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings
0	49091	COZICOMFORT LONG TERM STAY ROOM 2	266763	Francesca	North Region	Woodlands	1.44255	103.79580	1	83	180	1	2013-10-21	0.01	
1	50646	Pleasant Room along Bukit Timah	227796	Sujatha	Central Region	Bukit Timah	1.33235	103.78521	1	81	90	18	2014-12-26	0.28	
2	56334	COZICOMFORT	266763	Francesca	North Region	Woodlands	1.44246	103.79667	1	69	6	20	2015-10-01	0.20	
3	71609	Ensuite Room (Room 1 & 2) near EXPO	367042	Belinda	East Region	Tampines	1.34541	103.95712	1	206	1	14	2019-08-11	0.15	
4	71896	B&B Room 1 near Airport & EXPO	367042	Belinda	East Region	Tampines	1.34567	103.95963	1	94	1	22	2019-07-28	0.22	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7902	38105126	Loft 2 pax near Haw Par / Pasir Panjang. Free ...	278109833	Belle	Central Region	Queenstown	1.27973	103.78751	0	100	3	0	NaN	NaN	
7903	38108273	3bedroom luxury at Orchard	238891646	Neha	Central Region	Tanglin	1.29269	103.82623	0	550	6	0	NaN	NaN	
7904	38109336	[Farrer Park] New City Fringe CBD Mins to MRT	2814488565	Mindy	Central Region	Kallang	1.31286	103.85996	1	58	30	0	NaN	NaN	
7905	38110493	Cheap Master Room in Central of Singapore	243835202	Huang	Central Region	River Valley	1.29543	103.83801	1	56	14	0	NaN	NaN	
7906	38112762	Amazing room with private bathroom walk to Orc...	28788520	Terence	Central Region	River Valley	1.29672	103.83325	1	65	90	0	NaN	NaN	

7907 rows × 16 columns

In [30]:
```
dataset.shape
```

Out[30]:
```
(7907, 16)
```

4. Splitting the Dataset into the Training set & Test set

In [23]:
```
from sklearn.model_selection import train_test_split
```

In [25]:
```
x_train, x_test, y_train, y_test = train_test_split(x, y,test_size=1/3, random_state=0)
```

5. Print x_train, x_test, y_train & y_test

In [26]:
```
print(x_train)
```

```
[[16345097 '2 Bedrooms-Buses to Orchard, Sentosa, Marina Bay' 27869744
  ... '2018-05-02' 0.21 1]
 [19039775 'NEW Cozy Bedroom Suite/ WIFI@Orchard/Somerset Area' 46685310
  ... '2019-07-19' 2.21 28]
 [18677985 "Business Traveler's Haven near Orchard Road" 129831064 ...
  nan nan 1]
 ...
 [13735926 'Large bedroom w/ use of whole 1st floor of duplex' 80393016
  ... '2016-07-11' 0.03 1]
 [18954823 'NEW LIST: Nice & big 1 Bedrm apartment near to MRT' 132164074
  ... nan nan 1]
 [19615585 'City Fringe Apartment for Rent' 21038840 ... nan nan 1]]
```

In [27]:
```
print(x_test)
```

```
[[24609199 'Cozy Deluxe Queen Private Room' 164189015 ... '2019-04-16'
  0.23 1]
 [28422606 'Fully furnish,clean&bright private room,near NTU.' 60077959
  ... '2019-08-22' 0.81 4]
 [29252819 'Cosy Private Room Suite @ Orchard/ Central Area.' 61619807
  ... '2019-06-23' 1.3 27]
 ...
 [31132529 'For 14 paxes, Entire 4 bedroom Apartment near MRT!' 211434562
  ... '2019-07-09' 0.59 64]
 [30053313 'A - Private, Cozy Apt, 3 mins to Orchard Road' 225311799 ...
  '2019-06-12' 1.16 30]
 [10626378 'Master room next to bencoolen station' 11390076 ...
  '2019-05-12' 0.34 18]]
```

In [28]:
```
print(x_test)
```

```
[[24609199 'Cozy Deluxe Queen Private Room' 164189015 ... '2019-04-16'
  0.23 1]
 [28422606 'Fully furnish,clean&bright private room,near NTU.' 60077959
  ... '2019-08-22' 0.81 4]
 [29252819 'Cosy Private Room Suite @ Orchard/ Central Area.' 61619807
  ... '2019-06-23' 1.3 27]
 ...
 [31132529 'For 14 paxes, Entire 4 bedroom Apartment near MRT!' 211434562
  ... '2019-07-09' 0.59 64]
 [30053313 'A - Private, Cozy Apt, 3 mins to Orchard Road' 225311799 ...
  '2019-06-12' 1.16 30]
 [10626378 'Master room next to bencoolen station' 11390076 ...
  '2019-05-12' 0.34 18]]
```

In [29]:
```
print(x_test)
```

```
[[24609199 'Cozy Deluxe Queen Private Room' 164189015 ... '2019-04-16'
  0.23 1]
 [28422606 'Fully furnish,clean&bright private room,near NTU.' 60077959
  ... '2019-08-22' 0.81 4]
 [29252819 'Cosy Private Room Suite @ Orchard/ Central Area.' 61619807
  ... '2019-06-23' 1.3 27]
 ...
 [31132529 'For 14 paxes, Entire 4 bedroom Apartment near MRT!' 211434562
  ... '2019-07-09' 0.59 64]
 [30053313 'A - Private, Cozy Apt, 3 mins to Orchard Road' 225311799 ...
  '2019-06-12' 1.16 30]
 [10626378 'Master room next to bencoolen station' 11390076 ...
  '2019-05-12' 0.34 18]]
```

End of Code