Gerry William Nanhoy / 2440018854 / L801 / Assignment 06

Importing Data from Various Sources

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Reading Excel Files

```
dataset = pd.read_csv('listings.csv')
dataset.head()
```

	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
0	49091	COZY COMFORTABLE LONG TERM STAY	266761	Francesca	North Region	Woodlands	1.44255	103.79980	Private room	83	180	1	2013-10-21	0.01	2
1	50646	Pleasant Room along Bedok	227796	Sophia	Central Region	Bukit Timah	1.33235	103.78521	Private room	81	90	18	2014-12-26	0.26	1
2	56334	COZY COMFORTABLE	266761	Francesca	North Region	Woodlands	1.44246	103.79967	Private room	69	6	20	2015-10-01	0.20	2
3	71609	Ensuite Room (Room 1 on 2 floor)	367042	Belinda	East Region	Tampines	1.34541	103.95812	Private room	206	1	14	2019-06-11	0.15	9
4	75235	B&B Room 1 near Nanyang & EXPO	367042	Belinda	East Region	Tampines	1.34567	103.95963	Private room	94	1	22	2019-07-28	0.22	9

```
dataset.shape
```

```
(7907, 16)
```

Data Reading

```
dataset.info()
```

```
<class'pandas.core.frame.DataFrame'>
RangeIndex: 7907 entries, 0 to 7906
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              7907 non-null   int64
 1   name                            7905 non-null   object
 2   host_id                         7907 non-null   int64
 3   host_name                       7907 non-null   object
 4   neighbourhood_group             7907 non-null   object
 5   neighbourhood                   7907 non-null   object
 6   latitude                        7907 non-null   float64
 7   longitude                       7907 non-null   float64
 8   room_type                       7907 non-null   object
 9   price                           7907 non-null   int64
 10  minimum_nights                  7907 non-null   int64
 11  number_of_reviews               7907 non-null   int64
 12  last_review                     5149 non-null   object
 13  reviews_per_month               5149 non-null   float64
 14  calculated_host_listings_count  7907 non-null   int64
 15  availability_365                7907 non-null   int64
dtypes: float64(3), int64(7), object(6)
memory usage: 988.5+ KB
```

```
dtypes.dtypes
```

```
id                                  int64
name                                object
host_id                             int64
host_name                           object
neighbourhood_group                 object
neighbourhood                       object
latitude                            float64
longitude                           float64
room_type                           object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                         object
reviews_per_month                   float64
calculated_host_listings_count      int64
availability_365                    int64
dtype: object
```

```
dataset.isnull().sum()
```

```
id                                  0
name                                2
host_id                             0
host_name                           0
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                      2758
reviews_per_month                2758
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

```
dataset['neighbourhood_group'].unique()
```

```
array([' North Region', 'Central Region', 'East Region', 'West Region',
       'North East Region'], dtype=object)
```

Discuss to find the correlation and covariance of the data (the variables used can be categorical vs categorical, categorical vs numeric, or numeric vs numeric variables).

Correlation

```
from sklearn.preprocessing import LabelEncoder
```

```
temp_dataset = dataset.copy()
encoder = LabelEncoder()
temp_dataset['name'] = encoder.fit_transform(temp_dataset['name'])
temp_dataset['host_name'] = encoder.fit_transform(temp_dataset['host_name'])
temp_dataset['neighbourhood_group'] = encoder.fit_transform(temp_dataset['neighbourhood_group'])
temp_dataset['neighbourhood'] = encoder.fit_transform(temp_dataset['neighbourhood'])
temp_dataset['room_type'] = encoder.fit_transform(temp_dataset['room_type'])
temp_dataset['last_review'] = encoder.fit_transform(temp_dataset['last_review'])
```

```
temp_dataset.head()
```

	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	availability_365
0	49091	1955	266762	499	2	41	1.44255	103.79980	1	83	180	1	0	0.01	2	2
1	50646	4951	227796	1446	0	9	1.33235	103.78521	1	81	90	18	9	0.26	1	0
2	56334	1956	266762	499	2	41	1.44246	103.79967	1	69	6	20	47	0.20	2	2
3	71609	3126	367042	1294	1	26	1.34541	103.95812	1	206	1	14	982	0.15	9	1
4	75235	1457	367042	1294	1	26	1.34567	103.95963	1	94	1	22	970	0.22	9	2

```
plt.figure(figsize=(14, 14))
sns.heatmap(temp_dataset.corr(), annot=True)
```

```
plt.tight_layout()
plt.show()
```

Covariance

```
plt.figure(figsize=(14, 14))
sns.heatmap(temp_dataset.cov(), annot=True)
```

```
plt.tight_layout()
plt.show()
```


Normalized the data using Standardization or Range Scaling method.

```
from python_nominal import identify_nominal_columns
nominal_features = identify_nominal_columns(dataset)
categorical_features
```

```
['name',
 'host_name',
 'neighbourhood_group',
 'neighbourhood',
 'room_type',
 'last_review']
```

```
dataset.head()
```

	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
0	49091	COZY COMFORTABLE LONG TERM STAY	266761	Francesca	North Region	Woodlands	1.44255	103.79980	Private room	83	180	1	2013-10-21	0.01	2
1	50646	Pleasant Room along Bedok	227796	Sophia	Central Region	Bukit Timah	1.33235	103.78521	Private room	81	90	18	2014-12-26	0.26	1
2	56334	COZY COMFORTABLE	266761	Francesca	North Region	Woodlands	1.44246	103.79967	Private room	69	6	20	2015-10-01	0.20	2
3	71609	Ensuite Room (Room 1 on 2 floor)	367042	Belinda	East Region	Tampines	1.34541	103.95812	Private room	206	1	14	2019-06-11	0.15	9
4	75235	B&B Room 1 near Nanyang & EXPO	367042	Belinda	East Region	Tampines	1.34567	103.95963	Private room	94	1	22	2019-07-28	0.22	9

```
dataset.drop(['id', 'name', 'host_name', 'last_review'], axis=1, inplace=True)
```

```
help_dataset = pd.get_dummies(dataset)
help_dataset.head()
```

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	neighbourhood_group_Central Region	neighbourhood_group_East Region	...	neighbourhood_Tampines	neig
0	1.44255	103.79980	83	180	1	0.01	2	2	0	0	...	0	0
1	1.33235	103.78521	81	90	18	0.26	1	0	1	0	...	0	0
2	1.44246	103.79967	69	6	20	0.20	2	2	0	0	...	0	0
3	1.34541	103.95812	206	1	14	0.15	9	1	0	0	...	1	0
4	1.34567	103.95963	94	1	22	0.22	9	2	0	0	...	1	0

5 rows × 18 columns

```
help_dataset.columns
```

```
Index(['latitude', 'longitude', 'price', 'minimum_nights', 'number_of_reviews',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'neighbourhood_group_Central Region',
       'neighbourhood_group_East Region', 'neighbourhood_group_North Region',
       'neighbourhood_group_North East Region',
       'neighbourhood_group_West Region', 'neighbourhood_Ang Mo Kio',
       'neighbourhood_Bedok', 'neighbourhood_Bishan',
       'neighbourhood_Bukit Batok', 'neighbourhood_Bukit Merah',
       'neighbourhood_Bukit Panjang', 'neighbourhood_Bukit Timah',
       'neighbourhood_Central Area', 'neighbourhood_Chua Chu Kang',
       'neighbourhood_Clementi', 'neighbourhood_Downtown Core',
       'neighbourhood_Geylang', 'neighbourhood_Hougang',
       'neighbourhood_Jurong East', 'neighbourhood_Jurong West',
       'neighbourhood_Kallang', 'neighbourhood_Lim Chu Kang',
       'neighbourhood_Mandai', 'neighbourhood_Marine Parade',
       'neighbourhood_Marine Parade', 'neighbourhood_Novena',
       'neighbourhood_Orchard', 'neighbourhood_Outram', 'neighbourhood_Orchard',
       'neighbourhood_Punggol', 'neighbourhood_Queenstown',
       'neighbourhood_Sembawang', 'neighbourhood_Sembawang',
       'neighbourhood_Serangoon', 'neighbourhood_Singapore River',
       'neighbourhood_Southern Islands', 'neighbourhood_Tanglin Rubbr',
       'neighbourhood_Tampines', 'neighbourhood_Tampines',
       'neighbourhood_Tua Pasoh', 'neighbourhood_Tuas',
       'neighbourhood_Tuas', 'neighbourhood_Tuas', 'neighbourhood_woodlands',
       'neighbourhood_Yishun', 'room_type_Entire home/apt',
       'room_type_Private room', 'room_type_Shared room'],
      dtype='object')
```

Do feature selection and extraction.

```
x = dataset.iloc[:, :1]
y = dataset.iloc[:, -1]
```

```
print(x)
```

```
       neighbourhood_group neighbourhood  latitude  longitude        room_type \
0                North Region    Woodlands  1.44255  103.79980     Private room
1               Central Region   Bukit Timah  1.33235  103.78521     Private room
2                North Region    Woodlands  1.44246  103.79967     Private room
3                East Region     Tampines  1.34541  103.95812     Private room
4                East Region     Tampines  1.34567  103.95963     Private room
...                      ...          ...       ...       ...              ...
7902             Central Region    Queenstown  1.27973  103.7851  Entire home/apt
7903             Central Region       Bedok  1.32404  103.93009     Private room
7904             Central Region      Kallang  1.31296  103.85996     Private room
7905             Central Region  River Valley  1.28543  103.83981     Private room
7906             Central Region  River Valley  1.28972  103.83376     Private room

      price  minimum_nights  number_of_reviews  reviews_per_month  \
0        83             180                  1               0.01
1        81              90                 18               0.26
2        69               6                 20               0.20
3       206               1                 14               0.15
4        94               1                 22               0.22
...     ...             ...                ...                ...
7902     59               1                  0                NaN
7903     60               2                  0                NaN
7904     50               1                  0                NaN
7905     45              30                  0                NaN
7906     55              90                  0                NaN

      calculated_host_listings_count  availability_365
0                                  2                2
1                                  1                0
2                                  2                2
3                                  9                1
4                                  9                2
...                              ...              ...
7902                               1              365
7903                               1              365
7904                               1              365
7905                               1              365
7906                               1              365

[7907 rows x 10 columns]
```

```
dataset.info()
```

```
<class'pandas.core.frame.DataFrame'>
RangeIndex: 7907 entries, 0 to 7906
Data columns (total 10 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              7907 non-null   int64
 1   name                            7905 non-null   object
 2   host_id                         7907 non-null   int64
 3   host_name                       7907 non-null   object
 4   neighbourhood_group             7907 non-null   object
 5   neighbourhood                   7907 non-null   object
 6   latitude                        7907 non-null   float64
 7   longitude                       7907 non-null   float64
 8   room_type                       7907 non-null   object
 9   price                           7907 non-null   int64
 10  minimum_nights                  7907 non-null   int64
 11  number_of_reviews               7907 non-null   int64
 12  last_review                     5149 non-null   object
 13  reviews_per_month               5149 non-null   float64
 14  calculated_host_listings_count  7907 non-null   int64
 15  availability_365                7907 non-null   int64
dtypes: float64(3), int64(7), object(6)
memory usage: 988.5+ KB
```

```
dataset.dtypes
```

```
id                                  int64
name                                object
host_id                             int64
host_name                           object
neighbourhood_group                 object
neighbourhood                       object
latitude                            float64
longitude                           float64
room_type                           object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                         object
reviews_per_month                   float64
calculated_host_listings_count      int64
availability_365                    int64
dtype: object
```

```
dataset.isnull().sum()
```

```
id                                  0
name                                2
host_id                             0
host_name                           0
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                      2758
reviews_per_month                2758
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

```
dataset['neighbourhood_group'].unique()
```

```
array([' North Region', 'Central Region', 'East Region', 'West Region',
       'North East Region'], dtype=object)
```

```
print(y)
```

```
0    365
1      0
2    365
3      1
4      2
...  ...
7902   1
7903   1
7904   1
7905   1
7906   1
Name: availability_365, Length: 7907, dtype: int64
```

End of Code