Rental Prices- Statistical Approach

In [3]:

```python
# Importing packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# Loading data and printing out a few lines.
df = pd.read_csv('rentals_cleaned_csv')
df.head()
```

Out[3]:

	Unnamed: 0	id	latitude	longitude	property_type	room_type	bathrooms	bedrooms	minimum_nights	price
0	0	958	37.76931	-122.43386	Apartment	Entire home/apt	1	1.0	1	17000.0
1	1	3850	37.75402	-122.45805	House	Private room	1	1.0	1	9900.0
2	2	5858	37.74511	-122.42102	Apartment	Entire home/apt	1	2.0	30	23500.0
3	3	7918	37.76669	-122.45250	Apartment	Private room	4	1.0	32	6500.0
4	4	8142	37.76487	-122.45183	Apartment	Private room	4	1.0	32	6500.0

In [4]:

```python
# Linear simple regression between bedrooms and price
df['intercept'] = 1
from statsmodels.api import OLS
lm = OLS(df['price'], df[['bedrooms','intercept']])
lm.fit().summary()
```

Out[4]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.142
Model:	OLS	Adj. R-squared:	0.141
Method:	Least Squares	F-statistic:	1334.
Date:	Fri, 08 Oct 2021	Prob (F-statistic):	1.68e-270
Time:	16:11:19	Log-Likelihood:	-94209.
No. Observations:	8095	AIC:	1.884e+05
Df Residuals:	8093	BIC:	1.884e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
bedrooms	1.202e+04	329.185	36.525	0.000	1.14e+04	1.27e+04
intercept	5551.6158	537.885	10.321	0.000	4497.224	6606.008

Omnibus:	15552.285	Durbin-Watson:	1.898
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47600476.124
Skew:	14.785	Prob(JB):	0.00
Kurtosis:	377.501	Cond. No.	3.69

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [5]:

```python
# Linear simple regression between bathrooms and price
lm = OLS(df['price'], df[['bathrooms','intercept']])
lm.fit().summary()
```

Out[5]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.015
Model:	OLS	Adj. R-squared:	0.015
Method:	Least Squares	F-statistic:	122.3
Date:	Fri, 08 Oct 2021	Prob (F-statistic):	3.08e-28
Time:	16:11:19	Log-Likelihood:	-94765.
No. Observations:	8095	AIC:	1.895e+05
Df Residuals:	8093	BIC:	1.895e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
bathrooms	3878.1668	350.613	11.061	0.000	3190.876	4565.458
intercept	1.624e+04	595.078	27.285	0.000	1.51e+04	1.74e+04

Omnibus:	14538.840	Durbin-Watson:	1.890
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30433162.947
Skew:	12.820	Prob(JB):	0.00
Kurtosis:	302.284	Cond. No.	3.91

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [6]:

```python
# Linear simple regression between minimum_nights and price
lm = OLS(df['price'], df[['minimum_nights','intercept']])
lm.fit().summary()
```

Out[6]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	9.194
Date:	Fri, 08 Oct 2021	Prob (F-statistic):	0.00244
Time:	16:11:20	Log-Likelihood:	-94822.
No. Observations:	8095	AIC:	1.896e+05
Df Residuals:	8093	BIC:	1.897e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
minimum_nights	42.9292	14.158	3.032	0.002	15.175	70.683
intercept	2.104e+04	400.824	52.503	0.000	2.03e+04	2.18e+04

Omnibus:	14333.857	Durbin-Watson:	1.901
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27035148.737
Skew:	12.462	Prob(JB):	0.00
Kurtosis:	285.015	Cond. No.	34.5

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

From this simple linear regression model we see that bedrooms, bathrooms and minimum nights are statsticaly significant, why not cheking multiple regression

In [7]:

```python
# Multiple regression
lm = OLS(df['price'], df[['bathrooms', 'bedrooms', 'minimum_nights', 'intercept']])
lm.fit().summary()
```

Out[7]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.143
Model:	OLS	Adj. R-squared:	0.143
Method:	Least Squares	F-statistic:	450.5
Date:	Fri, 08 Oct 2021	Prob (F-statistic):	1.17e-270
Time:	16:11:20	Log-Likelihood:	-94201.
No. Observations:	8095	AIC:	1.884e+05
Df Residuals:	8091	BIC:	1.884e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
bathrooms	-49.1803	345.982	-0.142	0.887	-727.394	629.034
bedrooms	1.206e+04	347.943	34.664	0.000	1.14e+04	1.27e+04
minimum_nights	51.0530	13.122	3.891	0.000	25.331	76.776
intercept	4743.8719	669.126	7.090	0.000	3432.212	6055.532

Omnibus:	15392.362	Durbin-Watson:	1.894
Prob(Omnibus):	0.000	Jarque-Bera (JB):	44197259.302
Skew:	14.463	Prob(JB):	0.00
Kurtosis:	363.831	Cond. No.	66.3

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Badrooms have the higest effect on rental peice, followed by the minimum nights. For the bathroom its p value suggests it has no effect on rental price, lets check for multicollinearity.

In [8]:

```python
# Cehck for multicollinearity
df[['price','bathrooms','bedrooms','minimum_nights']].corr()
```

Out[8]:

	price	bathrooms	bedrooms	minimum_nights
price	1.000000	0.122035	0.376182	0.033685
bathrooms	0.122035	1.000000	0.325289	0.020778
bedrooms	0.376182	0.325289	1.000000	-0.016807
minimum_nights	0.033685	0.020778	-0.016807	1.000000

It seems there is a moderat linear relation betweem bathrooms and bed rooms, lets check the variance infation factors.

In [9]:

```python
# Cehck for multicollinearity #VIFS
from patsy import dmatrices
from statsmodels.stats.outliers_influence import variance_inflation_factor
y, x = dmatrices('price ~ bathrooms + bedrooms + minimum_nights', df, return_type = 'dataframe')
vif = pd.DataFrame()
vif['VIF'] = [variance_inflation_factor(x.values,i) for i in range(x.shape[1])]
vif['features'] = x.columns
vif
```

Out[9]:

	VIF	features
0	4.828112	Intercept
1	1.119196	bathrooms
2	1.119029	bedrooms
3	1.001054	minimum_nights

VIFS are lower than ten, There is no multicollinearity, however if we add an interaction what will happen?

In [10]:

```python
# Add an interaction term
df['bath_bed_rooms'] = df['bathrooms'] * df['bedrooms']
lm = OLS(df['price'], df[['bathrooms', 'bedrooms','bath_bed_rooms','minimum_nights', 'intercept']])
lm.fit().summary()
```

Out[10]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.144
Model:	OLS	Adj. R-squared:	0.143
Method:	Least Squares	F-statistic:	339.4
Date:	Fri, 08 Oct 2021	Prob (F-statistic):	1.61e-270
Time:	16:11:20	Log-Likelihood:	-94198.
No. Observations:	8095	AIC:	1.884e+05
Df Residuals:	8090	BIC:	1.884e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
bathrooms	-614.3271	419.728	-1.464	0.143	-1437.102	208.448
bedrooms	1.125e+04	488.173	23.039	0.000	1.03e+04	1.22e+04
bath_bed_rooms	380.7712	160.202	2.377	0.017	66.734	694.808
minimum_nights	51.3136	13.119	3.911	0.000	25.598	77.030
intercept	5803.4406	803.868	7.219	0.000	4227.652	7379.229

Omnibus:	15377.895	Durbin-Watson:	1.895
Prob(Omnibus):	0.000	Jarque-Bera (JB):	44374515.627
Skew:	14.428	Prob(JB):	0.00
Kurtosis:	364.564	Cond. No.	86.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

From the above statstical approach we can say that there is a linear relation between price, and numerical features except for the bathrooms that is replaced by the interaction term of bathrooms multiplied by bedrooms, actualy we don't need bathrooms in our model, also there is no multimulticollinearity.