利用MAOVA檢測
印度租屋價格及房間坪數

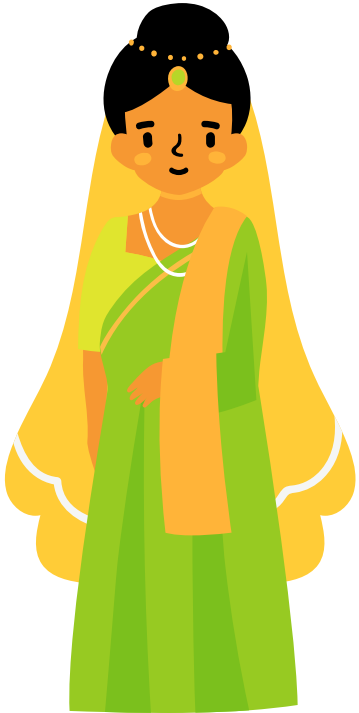
111426025 盧盈穎

2022/10/11

目錄

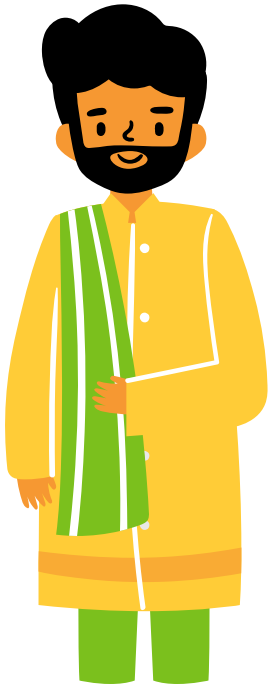
01

研究動機



01 研究動機
Research motivation

現今越來越多人選擇租屋，但租屋的價格、房屋坪數可能是每個人關注的點，那影響價格和坪數，可能會是房屋風水、房間朝向、同住租客類型等因素，因此，本研究從kaggle上收集到印度租屋價格、坪數是否會受房屋朝向和同住租客類型的影響。



02

資料集介紹

02. 資料集介紹 / 資料前處理

Intro. to Datasets

使用kaggle所提供Rental price of India's IT Capital - Pune, MH, IND資料集

此資料集是印度IT城市-浦納的租屋資料。

https://www.kaggle.com/datasets/anantsakhare/rental-price-of-indias-it-capital-pune-mh-ind

02. 資料集介紹 / 資料前處理 / 資料觀察

Intro. to Datasets

資料欄位:

```
df.columns
```

```
Index(['bedroom', 'bathrooms', 'area', 'furnishing', 'avalable_for', 'address',
       'floor_number', 'facing', 'floor_type', 'gate_community', 'corner_pro',
       'parking', 'wheelchairadption', 'petfacility', 'aggDur', 'noticeDur',
       'lightbill', 'powerbackup', 'propertyage', 'no_room', 'pooja_room',
       'study_room', 'others', 'servant_room', 'store_room', 'maintenance_amt',
       'brok_amt', 'deposit_amt', 'mnt_amt', 'rent'],
      dtype='object')
```

最後使用欄位 : area、rent、facing、available_for

報導資料:

https://kknews.cc/zh-tw/world/ql5bqrr.html

https://dq.yam.com/post/10362

選用這兩個因素是因為在印度，他們的信仰對於房子的朝向非常重視，而在租屋方面，對於租客類型也極其重視。

主欄位	類別	個數
facing	East	6
	West	
	South	
	North	
	North-East	
	North-West	
available_for	All	2
	Family Only	

1、檢查資料是否有空值

```
df.isnull().sum()
```

```
bedroom              0
bathrooms            0
area                 0
furnishing           0
avalable_for         0
address              0
floor_number         0
facing               0
floor_type           0
gate_community       0
corner_pro           0
parking              0
wheelchairadption    0
petfacility          0
aggDur               0
noticeDur            0
lightbill            0
powerbackup          0
propertyage          0
no_room              0
pooja_room           0
study_room           0
others               0
servant_room         0
store_room           0
maintenance_amt      0
brok_amt             0
deposit_amt          0
mnt_amt              0
rent                 0
dtype: int64
```

2、清除facing、available_for 、area中的 No Direction、None、0

```
for i in range(len(df)):
    if df['facing'][i]=='No Direction':
        df=df.drop(index=[i])
    elif df['avalable_for'][i]== 'None':
        df=df.drop(index=[i])
    elif df['area'][i]== 0:
        df=df.drop(index=[i])
```

3、清除rent中的極端值

```
import numpy as np
print ("Shape Of The Before Ouliers: ",df.shape)
n=1.2
#IQR = Q3-Q1
IQR = np.percentile(df['rent'],75) - np.percentile(df['rent'],25)
#outlier = Q3 + n*IQR
df=df[df['rent'] < np.percentile(df['rent'],75)+n*IQR]
#outlier = Q1 - n*IQR
df=df[df['rent'] > np.percentile(df['rent'],25)-n*IQR]
print ("Shape Of The After Ouliers: ",df.shape)
```

```
Shape Of The Before Ouliers:  (4583, 30)
Shape Of The After Ouliers:  (4280, 30)
```

4、清除area中的極端值

```
import numpy as np
print ("Shape Of The Before Ouliers: ",df.shape)
n=1.8
#IQR = Q3-Q1
IQR = np.percentile(df['area'],75) - np.percentile(df['area'],25)
#outlier = Q3 + n*IQR
df=df[df['area'] < np.percentile(df['area'],75)+n*IQR]
#outlier = Q1 - n*IQR
df=df[df['area'] > np.percentile(df['area'],25)-n*IQR]
print ("Shape Of The After Ouliers: ",df.shape)
```

```
Shape Of The Before Ouliers:  (4280, 30)
Shape Of The After Ouliers:  (4210, 30)
```

5、取出目標欄位

```
df_F = df['facing'].to_numpy()
df_Y = df['avalable_for'].to_numpy()
df_R = df['rent'].to_numpy()
df_B = df['area'].to_numpy()
```

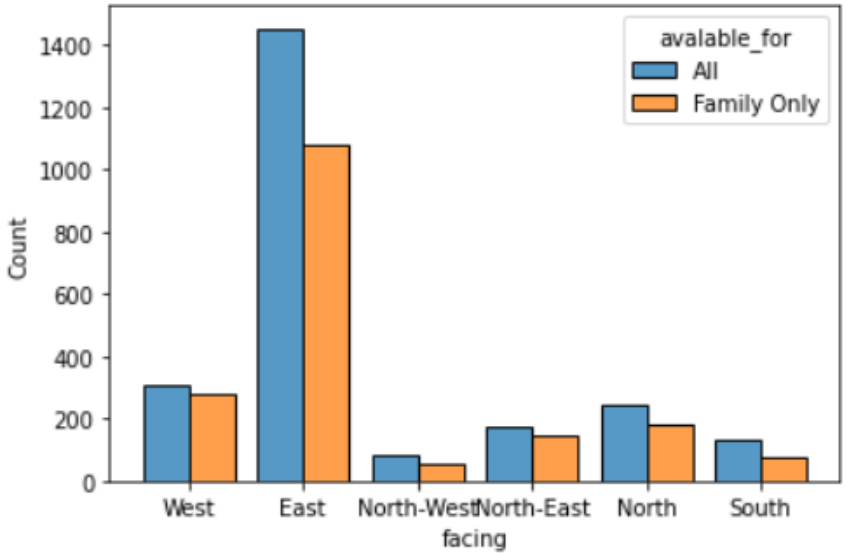
```
df_R4 = pd.DataFrame({'facing':df_F,'avalable_for':df_Y,'rent':df_R,'area':df_B})
df_R4
```

6、最終資料

	facing	avalable_for	rent	area
0	West	All	20000.0	1050.0
1	East	All	14000.0	760.0
2	East	Family Only	13000.0	628.0
3	East	Family Only	28000.0	1530.0
4	North-West	All	25999.0	1400.0
...	...	...	...	...
4205	East	All	11000.0	800.0
4206	North	All	20000.0	805.0
4207	North	Family Only	15000.0	900.0
4208	South	Family Only	15000.0	750.0
4209	East	All	23000.0	500.0

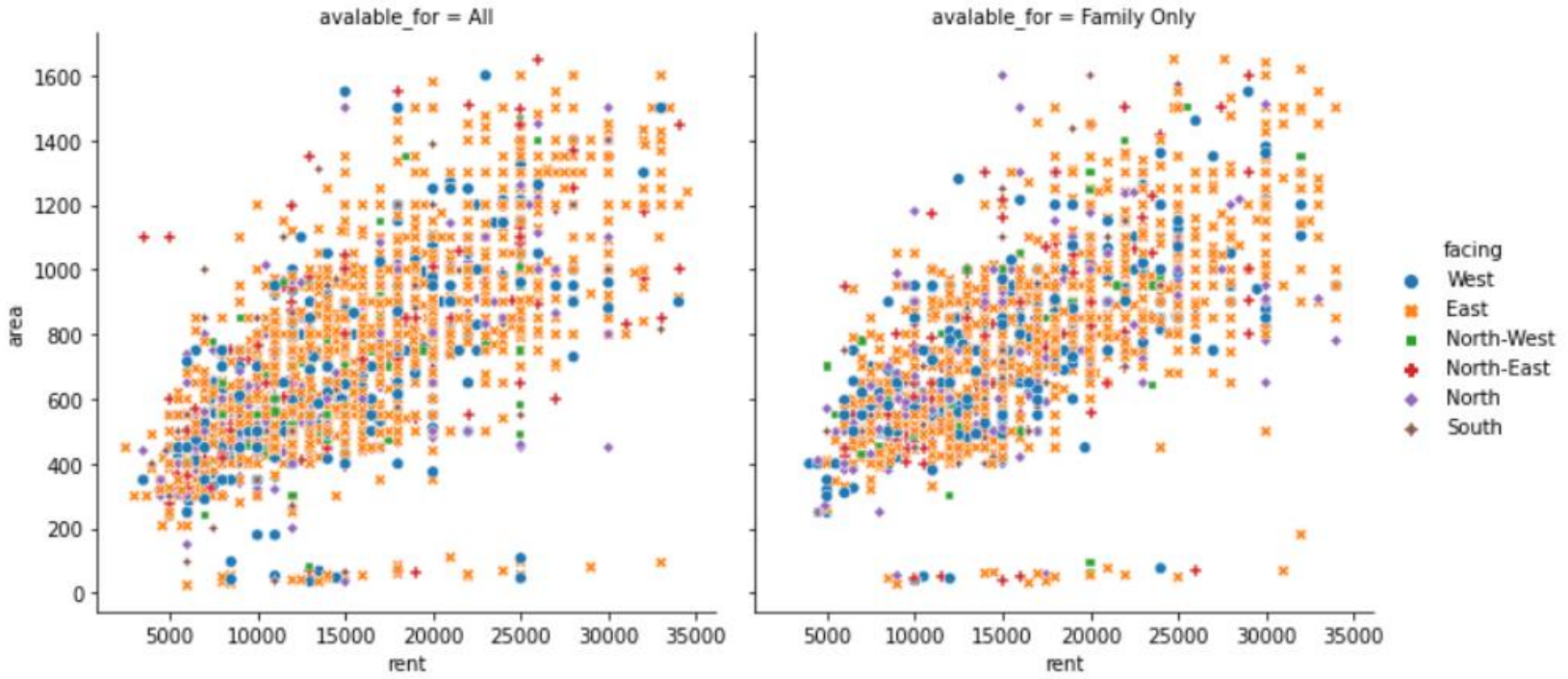
4210 rows x 4 columns

各種類別直方圖比較



可以看出東方佔這份資料的大多數，如同報導所述，印度人在買房、租房所在意之方位。

各種類別散佈圖比較


avalable_for = All
avalable_for = Family Only
area
rent
facing
West
East
North-West
North-East
North
South

03

研究方法

03. 研究方法
Methods

目的：

探討印度租屋的房屋朝向和租客類型是否會影響租屋價格和租屋坪數?

統計方法：

Two-way MANOVA檢測房屋朝向和租客類型是否會影響租屋價格和租屋坪數。

03. 研究方法 two-way MANOVA

Methods

1. 檢查資料集是否符合常態分配 – facing因子

根據中央極限定理，樣本數夠大，樣本和減去平均數再除以標準差，將會趨近平均數為0。

```
fa_E  =  df_R4[df_R4['facing']=='East']
fa_W  =  df_R4[df_R4['facing']=='West']
fa_S  =  df_R4[df_R4['facing']=='South']
fa_N  =  df_R4[df_R4['facing']=='North']
fa_NE  =  df_R4[df_R4['facing']=='North-East']
fa_NW  =  df_R4[df_R4['facing']=='North-West']
```

Facing 因子各類別的數量，皆大於30，根據中央極限定理，故符合常態分配。

```
print('East:',len(fa_E),'\n','West:',len(fa_W),'\n','South:',len(fa_S),
            '\n','North:',len(fa_N),'\n','North-East:',len(fa_NE),'\n','North-West:',len(fa_NW))
```

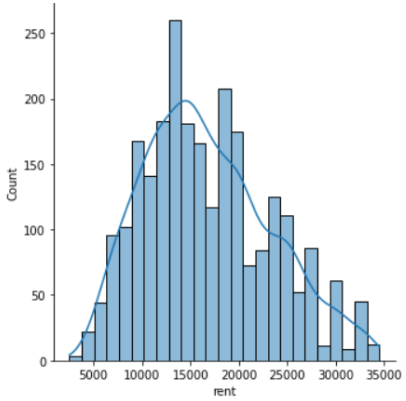
```
East: 2535
 West: 583
 South: 207
 North: 422
 North-East: 320
 North-West: 143
```

03. 研究方法 two-way MANOVA

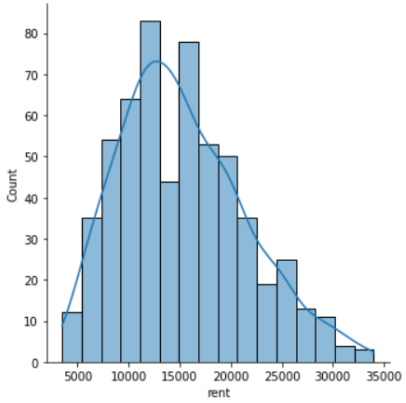
Methods

facing因子 – 分布圖 - rent

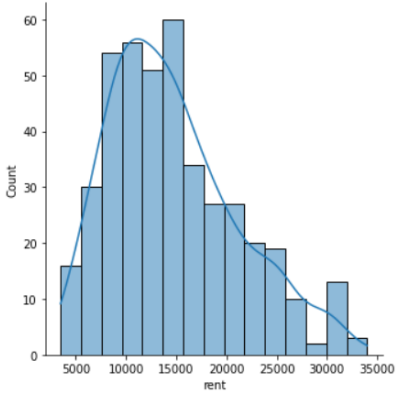
East



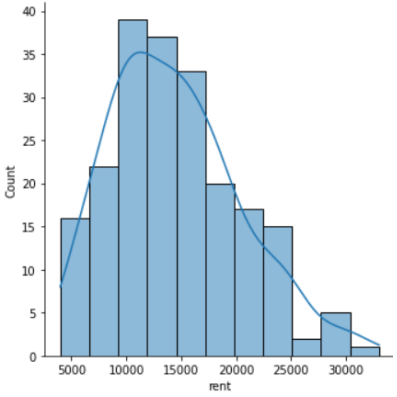
West



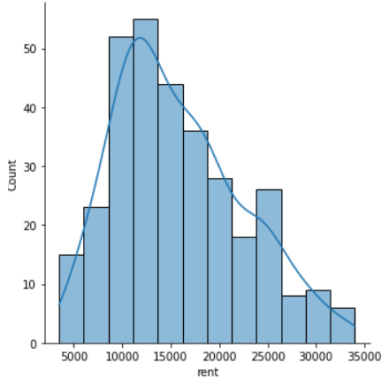
North



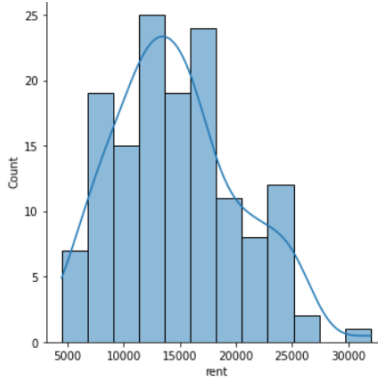
South



North-East



North-West

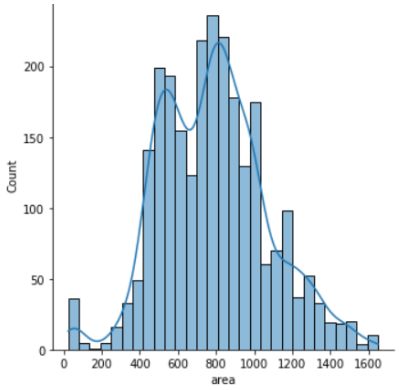


03. 研究方法 two-way MANOVA

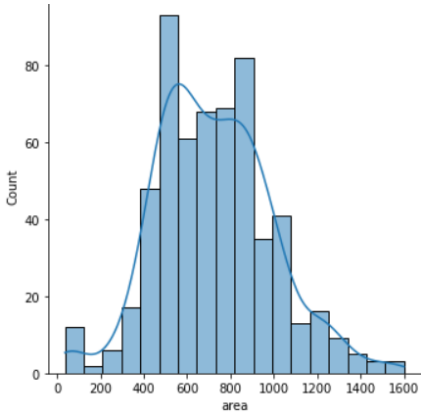
Methods

facing因子 – 分布圖 - area

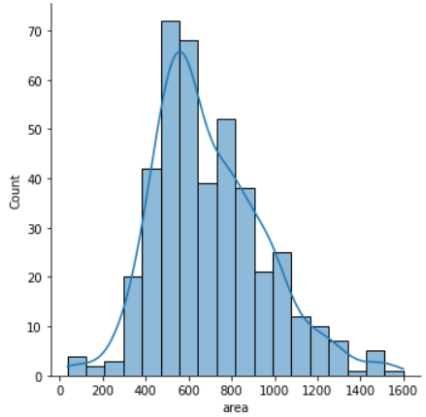
East



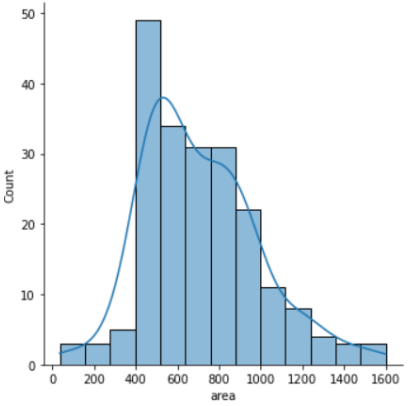
West



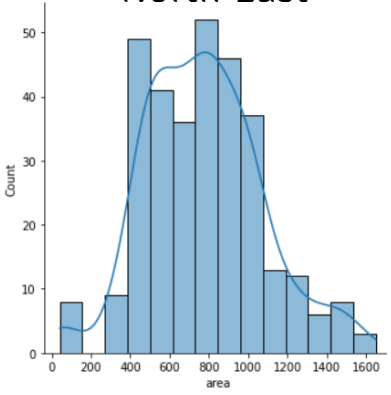
North



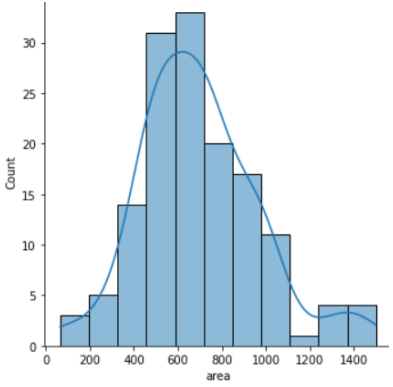
South



North-East



North-West



03. 研究方法 two-way MANOVA
Methods

2. 檢查資料集是否符合常態分配 – avalable_for因子

根據中央極限定理，樣本數夠大，樣本和減去平均數再除以標準差，將會趨近平均數為0。

```
ye_1  =  df_R4[df_R4['avalable_for']=='All']
ye_2  =  df_R4[df_R4['avalable_for']=='Family  Only']
```

avalable_for 因子各類別的數量，皆大於30，根據中央極限定理，故符合常態分配。

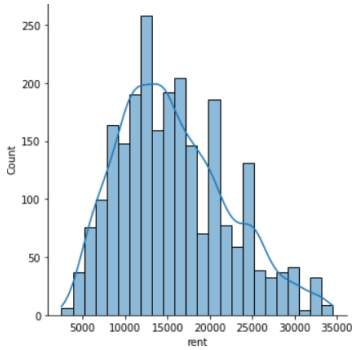
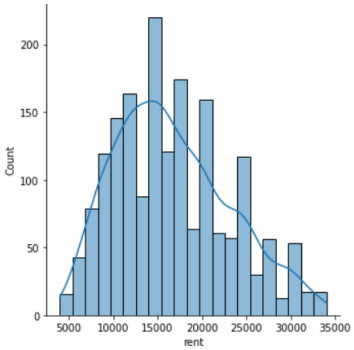
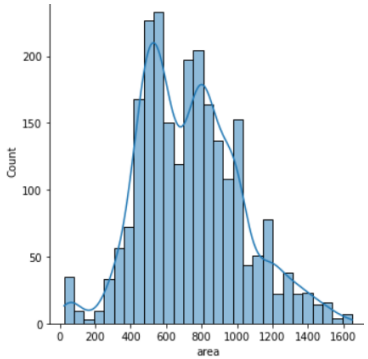
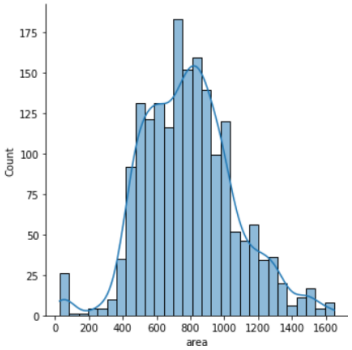
```
print('All:',len(ye_1),'\n','Family  Only:',len(ye_2))
```

```
All: 2396
 Family Only: 1814
```

03. 研究方法 two-way MANOVA

Methods

avalable_for因子 – 分布圖

	All	Family Only
rent		
area		

03. 研究方法 two-way MANOVA

Methods

3.虛無假設

$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$

$i =$ East、West、North、South、North-East、North-West

$j =$ All、Family Only

$\tau_i = \tau_1$、τ_2、τ_3、τ_4、τ_5、τ_6

$\beta_j = \beta_1$、β_2

$\tau\beta_{ij} = \tau\beta_{11}$、$\tau\beta_{12}$、$\tau\beta_{21}$、$\tau\beta_{22}$、$\tau\beta_{31}$、$\tau\beta_{32}$、$\tau\beta_{41}$、$\tau\beta_{42}$、$\tau\beta_{51}$、$\tau\beta_{52}$、$\tau\beta_{61}$、$\tau\beta_{62}$

$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = \tau_6$

H_A：H_0 is not true

$H_0 : \beta_1 = \beta_2$

H_A：H_0 is not true

$H_0 : \tau\beta_{11} = \tau\beta_{12} = \tau\beta_{21} = \tau\beta_{22} = \tau\beta_{31} = \tau\beta_{32} = \tau\beta_{41} = \tau\beta_{42} = \tau\beta_{51} = \tau\beta_{52} = \tau\beta_{61} = \tau\beta_{62}$

H_A：H_0 is not true

03. 研究方法 two-way MANOVA

Methods

結論：

```
from statsmodels.multivariate.manova import MANOVA

maov = MANOVA.from_formula('rent + area ~ facing + avalable_for + facing*avalable_for'\
,data = df_R4)
```

- 房屋朝向， p<0.05，拒絕H_0 ，因此房屋朝向和租金、房屋坪數存在差異。

  因此，透過one way ANOVA檢測存在差異性和 Tukey檢查類別間的交互關係。
- 租客類型，p<0.05，拒絕H_0 ，因此租客類型和租金、房屋坪數存在差異。

  因此，透過Two sample T-test檢查類別間的差異。
- 房屋朝向、租客類型之間交互關係， p>0.05，不拒絕H_0 ，因此兩者之間並沒有互相影響。

Multivariate linear model

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.2536	2.0000	4197.0000	6176.7522	0.0000
Pillai's trace	0.7464	2.0000	4197.0000	6176.7522	0.0000
Hotelling-Lawley trace	2.9434	2.0000	4197.0000	6176.7522	0.0000
Roy's greatest root	2.9434	2.0000	4197.0000	6176.7522	0.0000

facing	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9886	10.0000	8394.0000	4.8272	0.0000
Pillai's trace	0.0114	10.0000	8396.0000	4.8180	0.0000
Hotelling-Lawley trace	0.0115	10.0000	6292.7506	4.8369	0.0000
Roy's greatest root	0.0108	5.0000	4198.0000	9.0274	0.0000

avalable_for	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9931	2.0000	4197.0000	14.5301	0.0000
Pillai's trace	0.0069	2.0000	4197.0000	14.5301	0.0000
Hotelling-Lawley trace	0.0069	2.0000	4197.0000	14.5301	0.0000
Roy's greatest root	0.0069	2.0000	4197.0000	14.5301	0.0000

facing:avalable_for	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9982	10.0000	8394.0000	0.7383	0.6888
Pillai's trace	0.0018	10.0000	8396.0000	0.7384	0.6888
Hotelling-Lawley trace	0.0018	10.0000	6292.7506	0.7383	0.6888
Roy's greatest root	0.0014	5.0000	4198.0000	1.1445	0.3344

03. 研究方法 One-way ANOVA
Methods

facing 因子 - 針對rent

虛無假設

μ_{R1} : East　　μ_{R4} : North

μ_{R2} : West　　μ_{R5} : North−East

μ_{R3} : South　　μ_{R6} : North−West

H_0: $\mu_{R1} = \mu_{R2} = \mu_{R3} = \mu_{R4} = \mu_{R5} = \mu_{R6}$

H_a: 至少有一組的租金平均值不完全相同

α = 0.03

結論：

```
mod = ols('rent~facing ', data = df_R6).fit()
sm.stats.anova_lm(mod, typ = 2,alpha=0.03)
```

	sum_sq	df	F	PR(>F)
facing	3.231291e+09	5.0	15.684157	2.508030e-15
Residual	1.732238e+11	4204.0	NaN	NaN

- p<0.03，拒絕H_0 ，因此不同的房屋朝向對於租金有存在顯著差異。

03. 研究方法 Tukey-HSD

Methods

facing 因子 - 針對rent

Multiple Comparison of Means - Tukey HSD, FWER=0.03

group1	group2	meandiff	p-adj	lower	upper	reject
East	North	-2096.8196	0.001	-3118.0287	-1075.6105	True
East	North-East	-931.2886	0.1408	-2083.6091	221.032	False
East	North-West	-2063.2453	0.0026	-3732.7303	-393.7604	True
East	South	-2306.4201	0.001	-3710.5063	-902.3338	True
East	West	-1477.8425	0.001	-2370.0154	-585.6696	True
North	North-East	1165.5311	0.1395	-274.2782	2605.3403	False
North	North-West	33.5743	0.9	-1845.8905	1913.0391	False
North	South	-209.6004	0.9	-1857.8317	1438.6308	False
North	West	618.9771	0.6376	-622.4658	1860.42	False
North-East	North-West	-1131.9568	0.4963	-3085.764	821.8505	False
North-East	South	-1375.1315	0.1557	-3107.6559	357.393	False
North-East	West	-546.5539	0.8021	-1897.907	804.7991	False
North-West	South	-243.1747	0.9	-2355.2775	1868.9281	False
North-West	West	585.4028	0.9	-1227.1899	2397.9956	False
South	West	828.5775	0.5876	-742.9729	2400.128	False

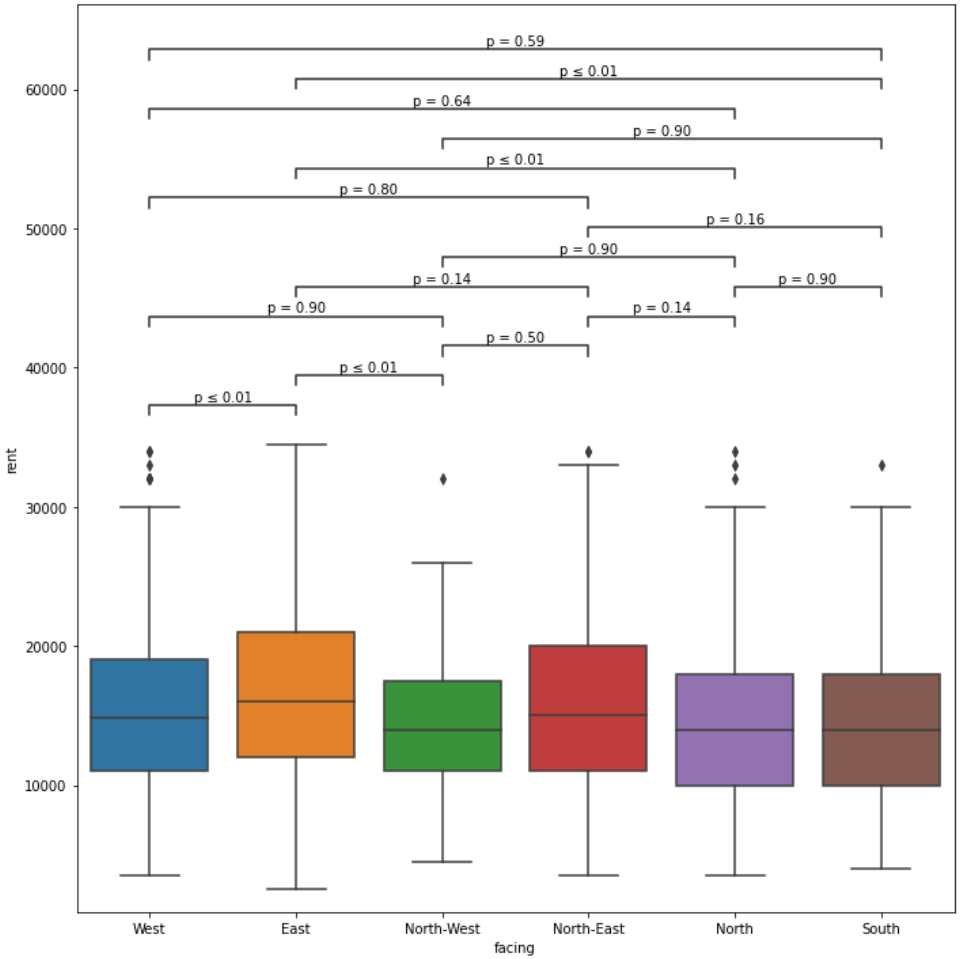
- 黃色底線的類別，由於他們的p值小於0.03，他們之間存在差異。

以下會以箱形圖顯示他們之間的交互關係是否存在差異

03. 研究方法 Tukey-HSD
Methods

facing 因子 - 針對rent

```
tukey_df = posthoc_tukey(df_R6,val_col="rent", group_col="facing")
tukey_df

remove = np.tril(np.ones(tukey_df.shape), k=0).astype("bool")
tukey_df[remove] = np.nan

molten_df = tukey_df.melt(ignore_index=False).reset_index().dropna()
molten_df

plt.figure(figsize=(10,10))
ax = sns.boxplot(data=df_R6, x="facing", y="rent")

pairs = [(i[1]["index"], i[1]["variable"]) for i in molten_df.iterrows()]
p_values = [i[1]["value"] for i in molten_df.iterrows()]

annotator = Annotator(
        ax, pairs, data=df_R6, x="facing", y="rent")
annotator.configure(text_format="simple", loc="inside")
annotator.set_pvalues_and_annotate(p_values)
plt.tight_layout()
```



03. 研究方法 One-way ANOVA
Methods

facing 因子 - 針對area

虛無假設

μ_{D1} : East　　μ_{D4} : North

μ_{D2} : West　　μ_{D5} : North−East

μ_{D3} : South　　μ_{D6} : North−West

H_0: $\mu_{D1} = \mu_{D2} = \mu_{D3} = \mu_{D4} = \mu_{D5} = \mu_{D6}$

H_a: 至少有一組的房屋坪數平均值不完全相同

α = 0.03

結論：

```
mod = ols('area~facing', data = df_R6).fit()
sm.stats.anova_lm(mod, typ = 2,alpha=0.03)
```

	sum_sq	df	F	PR(>F)
facing	5.365940e+06	5.0	13.870699	1.810255e-13
Residual	3.252671e+08	4204.0	NaN	NaN

- p<0.03，拒絕H_0 ，因此不同的房屋朝向對於房屋坪數存在差異。

03. 研究方法 Tukey-HSD

Methods

facing 因子 - 針對area

Multiple Comparison of Means - Tukey HSD, FWER=0.03

group1	group2	meandiff	p-adj	lower	upper	reject
East	North	-89.7513	0.001	-134.0031	-45.4995	True
East	North-East	-13.4883	0.9	-63.4215	36.445	False
East	North-West	-86.7801	0.0039	-159.1235	-14.4368	True
East	South	-76.3904	0.002	-137.2333	-15.5475	True
East	West	-63.8223	0.001	-102.4826	-25.162	True
North	North-East	76.263	0.003	13.8721	138.6539	True
North	North-West	2.9712	0.9	-78.4712	84.4136	False
North	South	13.3609	0.9	-58.0615	84.7833	False
North	West	25.929	0.6666	-27.8661	79.7241	False
North-East	North-West	-73.2919	0.0926	-157.9557	11.372	False
North-East	South	-62.9021	0.1144	-137.9772	12.1729	False
North-East	West	-50.334	0.0971	-108.8919	8.2238	False
North-West	South	10.3897	0.9	-81.1335	101.9129	False
North-West	West	22.9578	0.9	-55.5868	101.5025	False
South	West	12.5681	0.9	-55.5315	80.6677	False

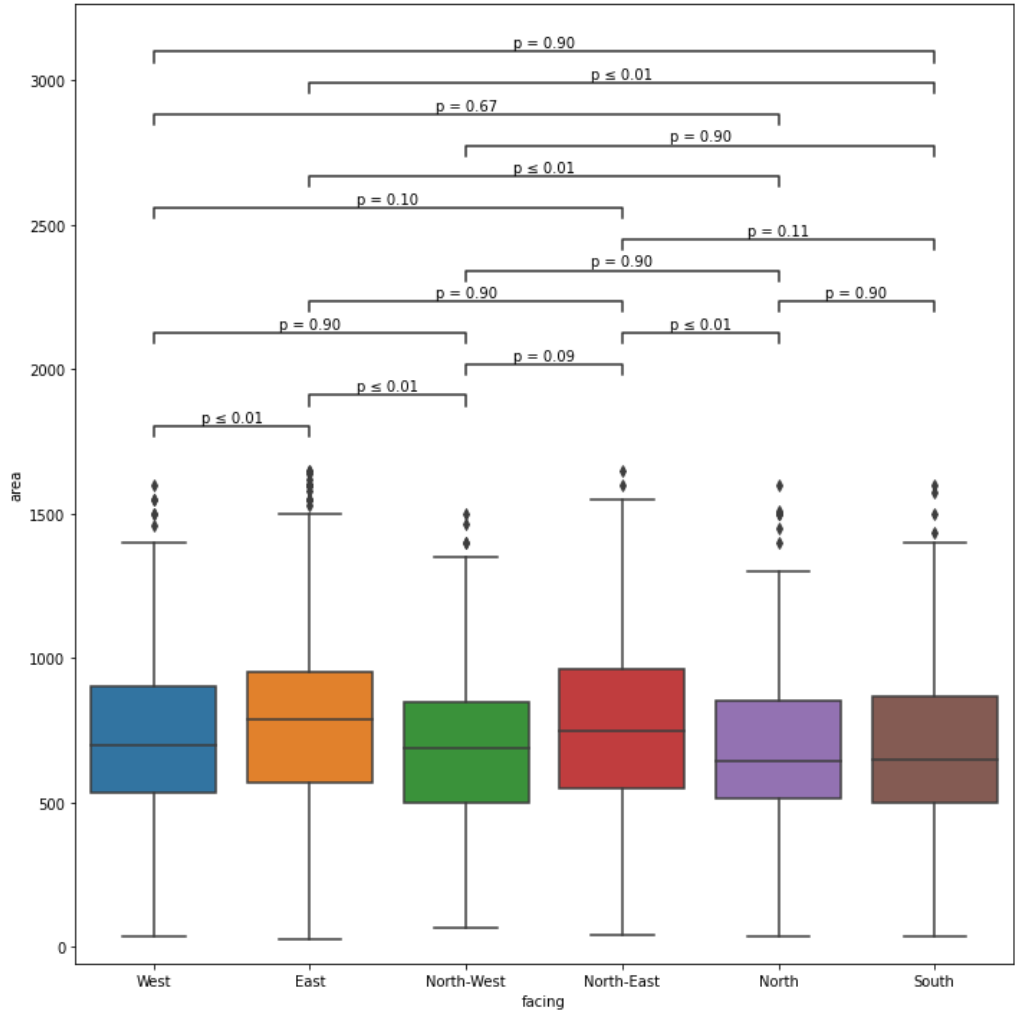
- 黃色底線的類別，由於他們的p值小於0.03，他們之間存在差異。

以下會以箱形圖顯示他們之間的交互關係是否存在差異

03. 研究方法 Tukey-HSD
Methods

facing 因子 - 針對area

```
tukey_df = posthoc_tukey(df_R6,val_col="area", group_col="facing")
tukey_df

remove = np.tril(np.ones(tukey_df.shape), k=0).astype("bool")
tukey_df[remove] = np.nan

molten_df = tukey_df.melt(ignore_index=False).reset_index().dropna()
molten_df

plt.figure(figsize=(10,10))
ax = sns.boxplot(data=df_R6, x="facing", y="area")

pairs = [(i[1]["index"], i[1]["variable"]) for i in molten_df.iterrows()]
p_values = [i[1]["value"] for i in molten_df.iterrows()]

annotator = Annotator(
        ax, pairs, data=df_R6, x="facing", y="area")
annotator.configure(text_format="simple", loc="inside")
annotator.set_pvalues_and_annotate(p_values)
plt.tight_layout()
```



03. 研究方法 Two-sample T-test

Methods

avalable_for 因子 - 針對rent

虛無假設

μ_{C1} : All

μ_{C2} : Family Only

H_0: $\mu_{C1} = \mu_{C2}$

H_a: 租客類型不同租金平均值不完全相同

α = 0.03

結論：

```
df_ALL  =  df_R5[df_R5['avalable_for']=='All']
df_B  =  df_R5[df_R5['avalable_for']=='Family  Only']
```

```
df_ALL_R  =  df_ALL.iloc[:,1:2]
df_B_R  =  df_B.iloc[:,1:2]
df_ALL_R  =  df_ALL_R.to_numpy()
df_B_R  =  df_B_R.to_numpy()
```

```
mean1  =  np.mean(df_ALL_R)
mean2  =  np.mean(df_B_R)
std1  =  np.std(df_ALL_R)
std2  =  np.std(df_B_R)
nobs1  =  len(df_ALL_R)
nobs2  =  len(df_B_R)
modified_std1  =  np.sqrt(np.float32(nobs1)/np.float32(nobs1-1))  *  std1
modified_std2  =  np.sqrt(np.float32(nobs2)/np.float32(nobs2-1))  *  std2
statistic,  pvalue  =  stats.ttest_ind_from_stats(  mean1=mean1,  std1=modified_std1,  nobs1=nobs1,
                          mean2=mean2,  std2=modified_std2,  nobs2=nobs2  )
print('statistic:',statistic,'\n','pvalue:',pvalue)
```

```
statistic: -4.98364682582449
 pvalue: 6.490459281723903e-07
```

- p<0.03，拒絕H_0 ，因此租客類型是全部和只能租給家庭之間的租金存在差異。

03. 研究方法 Two-sample T-test

Methods

avalable_for 因子 - 針對area

虛無假設

μ_{E1} : All

μ_{E2} : Bachelors

H_0: $\mu_{E1} = \mu_{E2}$

H_a:租客類型不同房屋坪數平均值不完全相同

α = 0.03

結論：

```
mean1 = np.mean(df_ALL_A)
mean2 = np.mean(df_B_A)
std1 = np.std(df_ALL_A)
std2 = np.std(df_B_A)
nobs1 = len(df_ALL_A)
nobs2 = len(df_B_A)
modified_std1 = np.sqrt(np.float32(nobs1)/np.float32(nobs1-1)) * std1
modified_std2 = np.sqrt(np.float32(nobs2)/np.float32(nobs2-1)) * std2
statistic, pvalue = stats.ttest_ind_from_stats( mean1=mean1, std1=modified_std1, nobs1=nobs1,
                                                mean2=mean2, std2=modified_std2, nobs2=nobs2 )
print('statistic:',statistic,'\n','pvalue:',pvalue)
```

```
statistic: -6.184257285767972
 pvalue: 6.832969045034287e-10
```

- p<0.03，拒絕H_0 ，因此租客類型是全部和只能租給家庭之間的房屋坪數存在差異。

03. 研究方法
Methods

總結論:

- 房屋朝向不同，對於租金來說，存在顯著差異；對於房屋坪數來說，存在顯著差異。

- 租客類型不同，對於租金來說，存在顯著差異；對於房屋坪數來說，存在顯著差異。

謝謝大家

111426025 盧盈穎