ĐỒ ÁN CUỐI KỲ

NHÓM 6

LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

Chủ đề: Giá thuê phòng trọ ở khu vực thành phố Hồ Chí Minh

GVHD: Thầy Bùi Tiến Lên

DANH SÁCH THÀNH VIÊN

Nhóm 6

HỌ VÀ TÊN	MSSV
Bùi Thanh Tùng	20120398
Đỗ Tấn Tài	20120408
Trần Khắc Bình	20120437
Lê Văn Hùng	20120485

Nội dung chính

1. Thu thập dữ liệu
2. Khám phá dữ liệu
3. Đưa ra các câu hỏi có ý nghĩa cần trả lời
4. Tiền xử lý và phân tích dữ liệu để trả lời cho từng câu hỏi

1 Thu thập dữ liệu

Thu thập dữ liệu

- Đây là bộ dữ liệu về giá thuê phòng trọ từ ngày 04/03/2015 đến ngày 07/01/2023 ở thành phố Hồ Chí Minh từ trang web http://phongtro123.com/
- Thư viện sử dụng: **Scrapy**
- Cách thu thập:

1. Thu thập các link bài post cho thuê trọ ở khu vực thành phố Hồ Chí Minh, lưu vào file post_url.json.

2. Thu thập dữ liệu của từng bài post từ các link đã thu thập, lưu vào file post_info.json.

```
df = pd.read_json('dataset/post_url.json')
print("Số lượng link đã thu thập: ",df.shape[0])
```

```
Số lượng link đã thu thập:  49121
```

```
df = pd.read_json('dataset/post_info.json')
print("Số lượng thông tin giá thuê đã thu thập: ",df.shape[0])
```

```
Số lượng thông tin giá thuê đã thu thập:  49119
```

2 Khám phá dữ liệu

a. Thông tin dữ liệu

- File `post_info.json` cho biết những thông tin về việc thuê phòng trọ khu vực Tp.HCM.
- Dataframe ban đầu:

	Title	Address	Price	Acreage	Content	Id	Type_post	Tenant	Posting_time	End_time	Contact	Phone_number	Zalo	Link
0	4,8tr PHONG FULL NOI THAT DEP GIA RE TAN BINH	Dia chi: 800 Duong Lac Long Quan, Phuong 9, Qu...	4.8 trieu/thang	30m	Phong rieng tu trong toa nha 2 mat tien thoang...	614102.0	Phong tro, nha tro	Tat ca	Thu 5, 11:54 05/01/2023	Thu 3, 11:54 10/01/2023	Huynh Dang Quynh Trang	0938441538	0938441538	https://phongtro123.com/4-8tr-phong-full-noi-t...
1	Cho thue phong tro moi xay, co may lanh, may g...	Dia chi: Duong Thong Nhat, Phuong 11, Quan Go ...	2.8 trieu/thang	20m	PHONG MOI XAY, THOANG MAT, SACH SE || Dia chi:...	612543.0	Phong tro, nha tro	Tat ca	Thu 5, 10:00 05/01/2023	Thu 5, 10:00 12/01/2023	Co Hanh	0919170897	0919170897	https://phongtro123.com/cho-thue-phong-tro-moi...
2	PHONG TRO MOI XAY RAT DEP SO 373/1/2A DUONG LY...	Dia chi: 373/1/2A Pho Ly Thuong Kiet, Phuong 9...	4 trieu/thang	20m	PHONG TRO MOI, DEP SO 373/1/2a LY THUONG KIET,...	212446.0	Phong tro, nha tro	Tat ca	Thu 4, 20:49 04/01/2023	Thu 5, 20:49 12/01/2023	hieuthanh2006 (*)	0918180057	0918180057	https://phongtro123.com/tinh-thanh/ho-chi-minh...

b. Đặc điểm dữ liệu

Số dòng và số cột và lần lượt lưu vào 2 biến là num_rows và num_cols

```
Num rows:  49119
Num cols:  14
```

- Ý nghĩa mỗi dòng: Theo như quan sát sơ bộ về dữ liệu thì một dòng cho biết thông tin về quảng cáo cho thuê phòng trọ của các nơi trên địa bàn Tp.HCM.

b. Đặc điểm dữ liệu

Khi kiểm tra dữ liệu, ta nhận thấy không có dòng bị trùng lặp, do đó ta không cần xử lý vấn đề này

```
have_duplicated_rows = room_df.duplicated().sum() > 0
have_duplicated_rows
```

```
False
```

b. Đặc điểm dữ liệu

Khi kiểm tra dữ liệu, ta nhận thấy có những dòng bị lỗi khi thu thập. Đó là những dòng có thuộc tính nhiều thuộc tính lỗi. Ta cũng sẽ loại bỏ những dòng này.

	Title	Address	Price	Acreage	Content	Id	Type_post	Tenant	Posting_time	End_time	Contact	Phone_number	Zalo	Link
2517						NaN	None	None	None	None	None	None	None	https://phongtro123.com/1-phong-tro-binh-dan-g...
3681	Pass phong o khu Khang Dien Mega Village gia 3...	Dia chi: Khang Dien Mega Village Duong Vo Chi ...	3.3 trieu/thang	\n 25m	Minh can pass 1 phong cho nu phong sang xin mi...	NaN	None	None	None	None	None	None	None	https://phongtro123.com/pass-phong-o-khu-khang...
4925	CHo thue CHDV 234 Quoc lo 13 - 25m2 - Duplex -...	Dia chi: 234 Quoc Lo 13, Phuong 26, Quan Binh ...	4.2 trieu/thang	\n 25m	Dia chi : 234 Quoc lo 13, phuong 26, quan Binh...	NaN	None	None	None	None	None	None	None	https://phongtro123.com/cho-thue-chdv-234-quoc...

b. Đặc điểm dữ liệu

Ta sẽ tiến hành loại bỏ những dòng đó đi.
Kích thước của dữ liệu sau khi xử lý vấn đề trùng lặp và các dòng lỗi:

```
num_rows, num_cols = room_df.shape
print("Num rows: ", num_rows)
print("Num cols: ", num_cols)
```

```
Num rows:  49094
Num cols:  14
```

b. Đặc điểm dữ liệu

Dưới đây là phần mô tả về thông tin các cột trong file post_info.json:

Id:	Mã tin quảng cáo cho thuê trọ.
Title:	Tiêu đề tin quảng cáo cho thuê trọ .
Address:	Địa chỉ cho thuê trọ.
Price:	Mức giá cho thuê trọ.
Acreage:	Diện tích phòng trọ.
Content:	Thông tin mô tả phòng trọ.
Type_post:	Loại tin rao quảng cáo.
Tenant:	Đối tượng có thể thuê trọ.
Posting_time:	Thời gian đăng thông báo cho thuê trọ.
End_time:	Thời gian kết thúc cho thuê trọ.
Contact:	Tên người có thể liên hệ nếu khách muốn thuê trọ.
Phone_number:	Số điện thoại người liên hệ.
Zalo:	Số Zalo người liên hệ.
Link:	Link của bài post

c. Xử lý dữ liệu

Ta tính tỉ lệ phần trăm các giá trị bị thiếu trong các cột.

```
{'Title': 0.0,
 'Address': 0.0,
 'Price': 0.0,
 'Acreage': 0.0,
 'Content': 0.0,
 'Id': 0.0,
 'Type_post': 0.0,
 'Tenant': 0.0,
 'Posting_time': 0.0,
 'End_time': 0.0,
 'Contact': 0.0,
 'Phone_number': 0.00203690878722450 83,
 'Zalo': 0.0020369087872245083,
 'Link': 0.0}
```

c. Xử lý dữ liệu

Có vẻ như sau khi xử lý vấn đề lỗi dòng ở trên thì các cột khá là đầy đủ dữ liệu. Còn 2 cột Phone_number và Zalo có giá trị bị thiếu, ta sẽ thử in nó ra để kiểm tra

	Title	Address	Price	Acreage	Content	Id	Type_post	Tenant	Posting_time	End_time	Contact	Phone_number	Zalo	Link
15236	PHONG CHO THUE FULL NOI THAT-CAO LO-QUAN 8	Dia chi: 190 Duong Cao Lo, Phuong 4, Quan 8, H...	4.3 trieu/thang	35m	* Chi con 1 Phong duy nhat, rat rong rai, co b...	284885.0	Phong tro, nha tro	Tat ca	Thu 7, 15:39 17/08/2019	Thu 6, 12:50 23/08/2019	achin151819@gmail.com	None	None	https://phongtro123.com/tinh-thanh/ho-chi-minh...

➔ Người đăng bài đã dùng email thay vì số điện thoại nên dữ liệu sđt liên hệ bị thiếu.

c. Xử lý dữ liệu

Dưới đây là kiểu dữ liệu của mỗi cột trong dataframe:

```
Title            object
Address          object
Price            object
Acreage          object
Content          object
Id              float64
Type_post        object
Tenant           object
Posting_time     object
End_time         object
Contact          object
Phone_number     object
Zalo             object
Link             object
dtype: object
```

Ta sẽ cần phải thực hiện những việc sau:

- Xóa các đơn vị không cần thiết ở các cột (VD: dong/thang, trieu/thang ở cột Price, thứ ở 2 cột Posting_time và End_time , m và m^2 ở cột Acreage, …)
- Chuyển dtype của cột Id từ float sang dạng chuỗi
- Chuyển dtype của cột Price và Acreage từ object sang float
- Chuyển dtype của cột Posting_time và End_time từ object sang datetime

Sau khi đã xử lý xong, ta sẽ kiểm tra lại kiểu dữ liệu thực sự của các cột:

```
def print_real_dtype(col):
    types = room_df[col].apply(type)
    print(types.unique())
print_real_dtype('Phone_number')
print_real_dtype('Zalo')
print_real_dtype('Price')
print_real_dtype('Acreage')
print_real_dtype('Posting_time')
print_real_dtype('End_time')
```

```
[<class 'str'> <class 'NoneType'>]
[<class 'str'> <class 'NoneType'>]
[<class 'float'>]
[<class 'float'>]
[<class 'pandas._libs.tslibs.timestamps.Timestamp'>]
[<class 'pandas._libs.tslibs.timestamps.Timestamp'>]
```

c. Xử lý dữ liệu

Ta thấy ở 2 cột Phone_number và Zalo có kiểu NoneType (không có giá trị), điều này khớp với dòng có giá trị bị thiếu đã được đề cập ở trên.

Xem lại 1 vài dòng đầu của dữ liệu để kiểm tra kết quả:

Id	Title	Address	Price	Acreage	Content	Type_post	Tenant	Posting_time	End_time	Contact	Phone_number	Zalo	Link
614102	4,8tr PHONG FULL NOI THAT DEP GIA RE TAN BINH	800 Duong Lac Long Quan, Phuong 9, Quan Tan Bi...	4.8	30.0	Phong rieng tu trong toa nha 2 mat tien thoang...	Phong tro, nha tro	Tat ca	2023-01-05 11:54:00	2023-01-10 11:54:00	Huynh Dang Quynh Trang	0938441538	0938441538	https://phongtro123.com/4-8tr-phong-full-noi-t...
612543	Cho thue phong tro moi xay, co may lanh, may g...	Duong Thong Nhat, Phuong 11, Quan Go Vap, Ho C...	2.8	20.0	PHONG MOI XAY, THOANG MAT, SACH SE || Dia chi:...	Phong tro, nha tro	Tat ca	2023-01-05 10:00:00	2023-01-12 10:00:00	Co Hanh	0919170897	0919170897	https://phongtro123.com/cho-thue-phong-tro-moi...
212446	PHONG TRO MOI XAY RAT DEP SO 373/1/2A DUONG LY...	373/1/2A Pho Ly Thuong Kiet, Phuong 9, Quan Ta...	4.0	20.0	PHONG TRO MOI, DEP SO 373/1/2a LY THUONG KIET,...	Phong tro, nha tro	Tat ca	2023-01-04 20:49:00	2023-01-12 20:49:00	hieuthanh2006 (*)	0918180057	0918180057	https://phongtro123.com/tinh-thanh/ho-chi-minh...

Sự phân bố giá trị các cột **numeric**:

- Hiện tại, ta đang có 4 cột thuộc nhóm **numeric** là: Price, Acreage, Posting_time và End_time.
- Với mỗi cột **numeric** ta sẽ tính tỉ lệ % giá trị thiếu (từ 0 đến 100), min, max.

	Price	Acreage	Posting_time	End_time
missing_ratio	0.0	0.0	0.0	0.0
min	0.16	0.0	2015-03-04 15:05:00	2015-12-11 21:49:00
max	800.0	1300.0	2023-01-07 23:15:00	2023-02-14 08:09:00

- Tuy tỉ lệ giá trị thiếu là 0 nhưng lại có những bài đăng để thông tin Acreage là 0 ! (làm cho người thấy bài đăng mơ hồ, không biết rõ diện tích phòng trọ là bao nhiêu)
- Trang web bắt đầu đăng bài từ năm 2015 đến nay.
- Có những phòng trọ rất rẻ nhưng cũng có những phòng cực kì đắt đỏ với diện tích rất lớn (giống cho thuê chung cư hơn là phòng trọ).

Sự phân bố giá trị các cột **categorical**:

- Có 9 cột **categorical** là: "Title", "Address", "Content", "Type_post", "Tenant", "Contact", "Phone_number", "Zalo" và "Link".
- Với mỗi cột **categorical**, ta tính tỉ lệ % giá trị thiếu (từ 0 đến 100), số lượng giá trị khác nhau (không xét giá trị thiếu), list các giá trị khác nhau (không xét giá trị thiếu).

	Title	Address	Content	Type_post	Tenant	Contact	Phone_number	Zalo	Link
missing_ratio	0.0	0.0	0.0	0.0	0.0	0.0	0.002073	0.002073	0.0
num_diff_vals	47188	27711	47129	4	3	20396	23769	23546	48243
diff_vals	[Ki tuc xa 54/20 Bach Dang gan san bay, cong v...	[Duong Nguyen Thi Thap, Phuong Binh Thuan, Qua...	[, , **Tien ich can phong: || - May lanh. || ...	[Phong tro, nha tro, Cho thue can ho, Nha thue...	[Tat ca, Nu, Nam]	[chu nha, Tran Van Tinh, quantri, He Thong 120...	[0796941691, 0917460309, 0974929266, 090271469...	[0917686101, 0796941691, 0917460309, 097492926...	[https://phongtro123.com/4-8tr-phong-full-noi-...

➢ Có vẻ như không có gì bất thường, đúng với những mô tả ở trên.

3

Đưa ra các câu hỏi có ý nghĩa cần trả lời

a. Các câu hỏi có ý nghĩa

Sau khi đã khám phá dữ liệu và hiểu hơn về dữ liệu, ta thấy có một số câu hỏi có thể được trả lời bằng dữ liệu:

- Top 3 quận có giá tiền thuê trọ trung bình cao nhất và thấp nhất?
- Sự biến thiên về số lượng giữa các nhà trọ mà nam giới và nữ giới có thể thuê được theo từng tiêu chí khác nhau ?
- Số lượng nhà trọ cho thuê và giá trung bình 1 m^2 cho thuê qua từng năm ?
- Diễn biến giá trọ của các quận ở mỗi tháng trong năm ?

b. Ý nghĩa các câu hỏi

Các câu hỏi được đặt ra giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 1. Top 3 quận có giá tiền thuê trọ trung bình cao nhất và thấp nhất?

- **Ý nghĩa**: với câu hỏi trên, ta biết được những khu vực tập trung nhiều phòng trọ giá rẻ nhất và mắc nhất để người thuê có thể cân nhắc chọn địa điểm sinh sống phù hợp với tài chính của mình.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 2. Sự biến thiên về số lượng giữa các nhà trọ mà nam giới và nữ giới có thể thuê được theo từng tiêu chí khác nhau ?

- **Ý nghĩa**: Với câu hỏi trên, ta biết được nam và nữ có thể thuê được bao nhiêu nhà trọ dựa trên từng tiêu chí, từ đó so sánh để xem bên nào sẽ có nhiều lợi thế hơn trong việc chọn nhà trọ thích hợp cho mình.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 3. Số lượng nhà trọ cho thuê và giá thuê trung bình cho 1 m^2 cho thuê qua từng năm?

- **Ý nghĩa**: với câu hỏi trên, ta biết được nhu cầu cho thuê phòng trọ và giá thuê ($1m^2$) qua từng năm, từ đó đưa ra quyết định về kinh doanh phòng trọ.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 4. Diễn biến giá trọ của các quận ở mỗi tháng trong năm?

- **Ý nghĩa**: với câu hỏi trên, ta biết được giá thuê ($1m^2$) phòng trọ ở các quận vào các tháng trong năm, từ đó đưa ra quyết định nên thuê phòng trọ vào tháng nào để được giá tốt nhất.

4 Tiền xử lý và phân tích dữ liệu để trả lời cho từng câu hỏi

Tiền xử lý

Để dễ dàng cho việc trả lời các câu hỏi, ta sẽ tạo ra cột District từ cột Address:

Id	Title	District	Address
614102	4,8tr PHONG FULL NOI THAT DEP GIA RE TAN BINH	Tan Binh	800 Duong Lac Long Quan, Phuong 9, Quan Tan Bi...
612543	Cho thue phong tro moi xay, co may lanh, may g...	Go Vap	Duong Thong Nhat, Phuong 11, Quan Go Vap, Ho C...
212446	PHONG TRO MOI XAY RAT DEP SO 373/1/2A DUONG LY...	Tan Binh	373/1/2A Pho Ly Thuong Kiet, Phuong 9, Quan Ta...
603145	Cho thue phong tro moi Duong Luong	Tan Phu	48/13 Duong Luong The Vinh, Phuong

1. Top 3 quận có giá tiền thuê trọ trung bình cao nhất và thấp nhất?

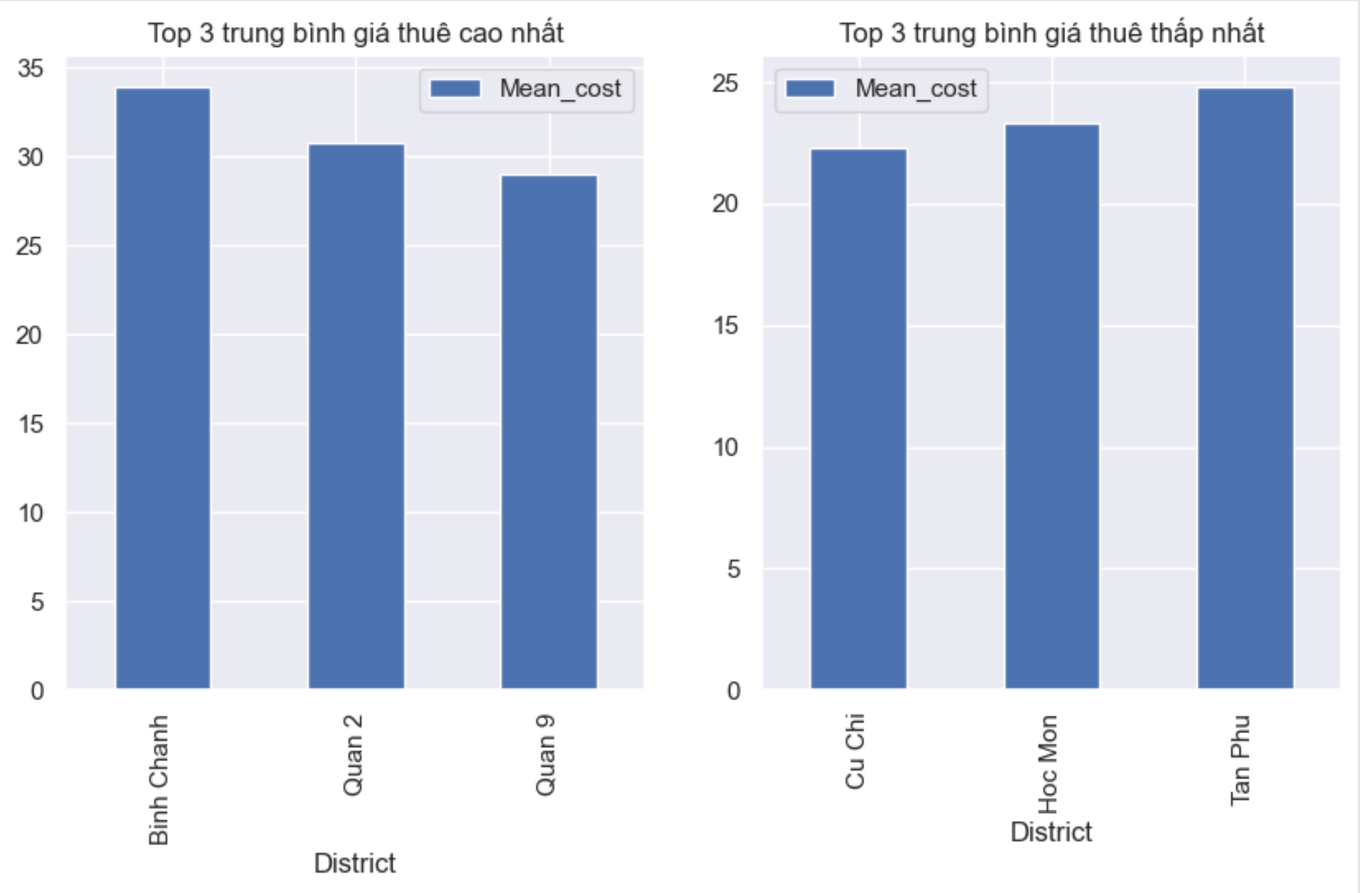
Để trả lời cho câu hỏi này, ta sẽ làm như sau:

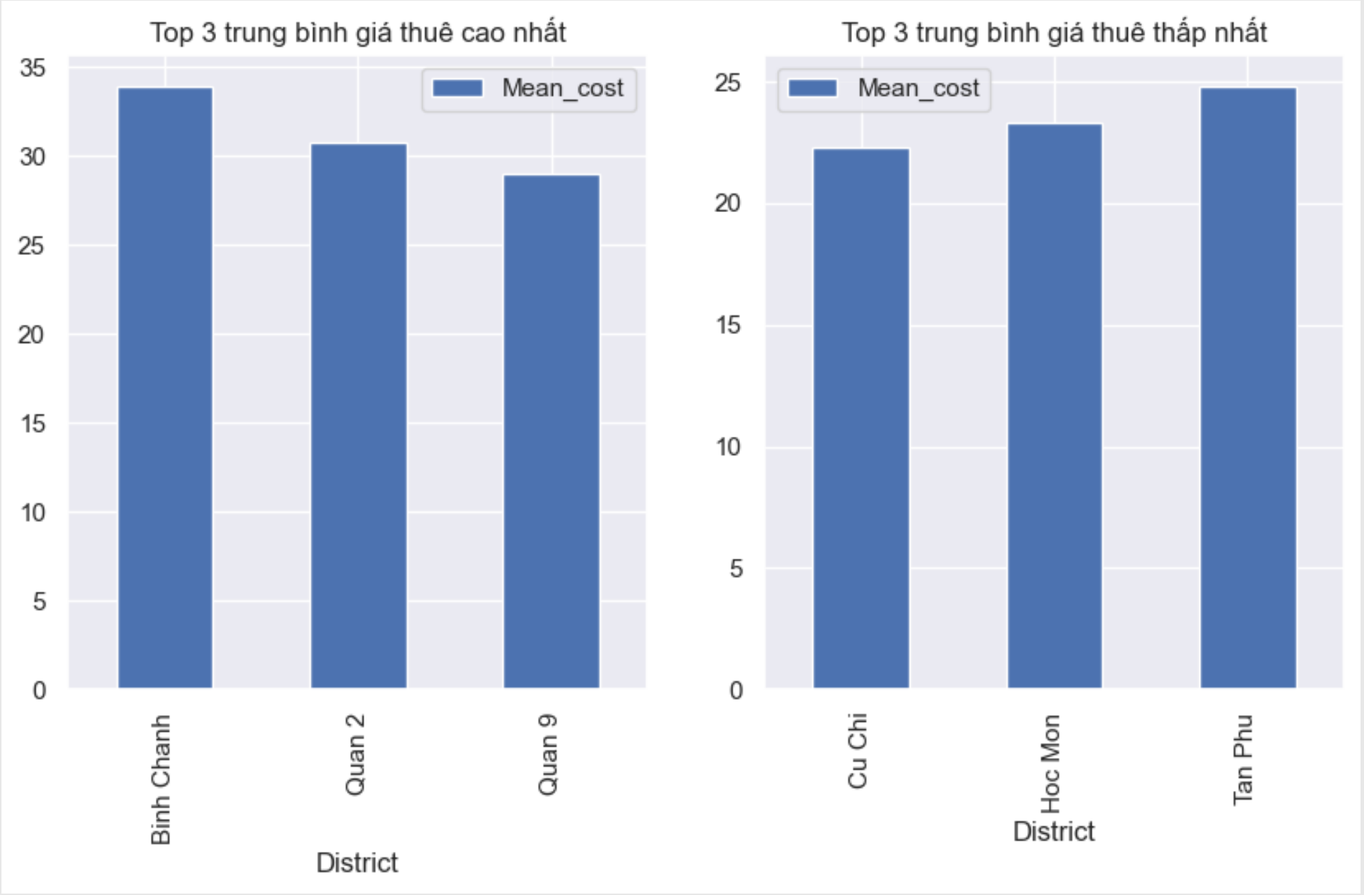
Phân loại tìm tính toán giá tiền thuê trung bình của các quận, chọn ra Top 3 quận có giá tiền thuê trọ trung bình cao nhất và thấp nhất. Ta lưu kết quả vào 2 series df_top3_high và df_top3_low , trong đó index là tên các quận.

	District	Mean_cost
0	Binh Chanh	33.921429
1	Quan 2	30.757220
2	Quan 9	28.958227
3	Nha Be	27.350694
4	Quan 8	27.208967
5	Thu Duc	27.127652
6	Quan 7	27.118022
7	Phu Nhuan	27.116168
8	Binh Thanh	26.939930
9	Go Vap	26.265371
10	Quan 1	26.047996
11	Tan Binh	25.949238
12	Quan 5	25.529217
13	Quan 4	25.275839
14	Quan 3	25.140064
15	Binh Tan	25.020275
16	Quan 6	24.820873
17	Tan Phu	24.815342
18	Hoc Mon	23.336538
19	Cu Chi	22.321429

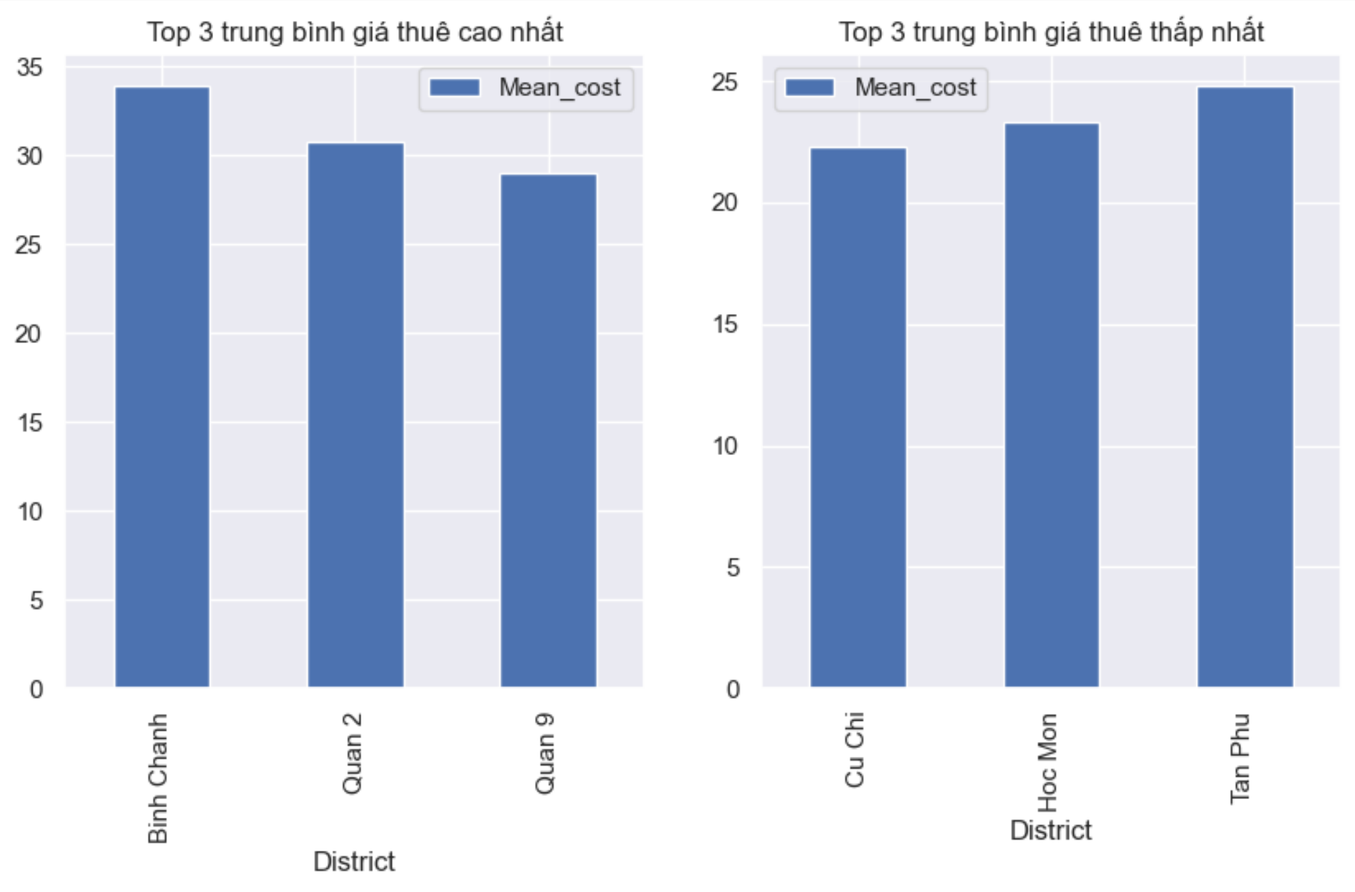
1. Top 5 vận động viên tham gia nhiều năm Thế vận hội nhất?

Từ kết quả ở trên, ta vẽ 2 đồ thị dạng cột, trong đó trục hoành là tên quận và trục tung là giá tiền thuê trung bình. Ta đặt tên trục hoành là "District" và tên trục tung là "Rent_cost".



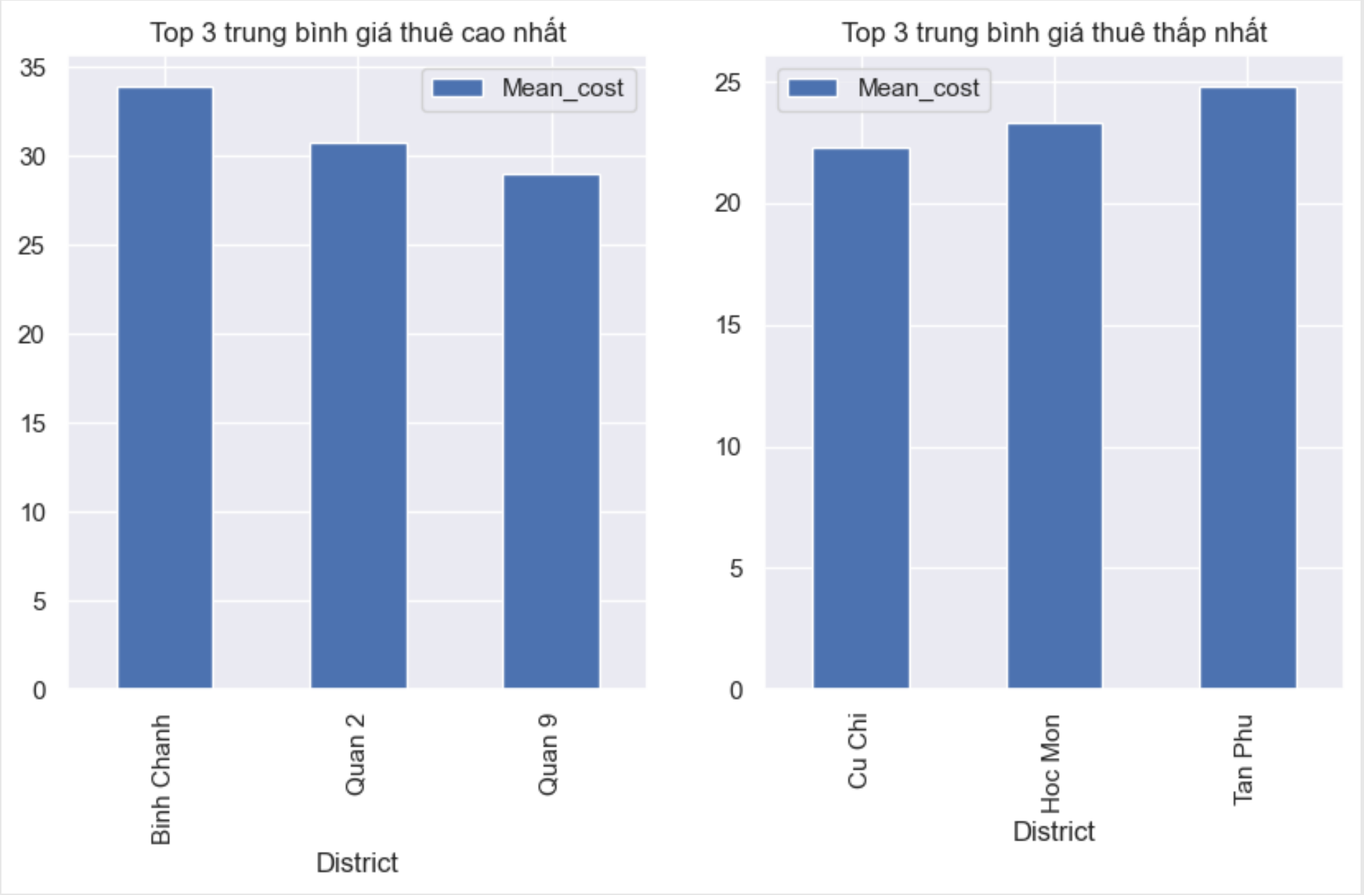


Nhìn chung, để đánh giá được mức giá thuê trọ của một khu vực là cao hay thấp, ta cần căn cứ vào nhiều yếu tố chứ không chỉ là giá tiền thuê. Tuy nhiên kết quả trên cũng phản ánh một phần độ mức chi phí thuê trọ ở các khu vực trên địa bàn Tp.HCM.

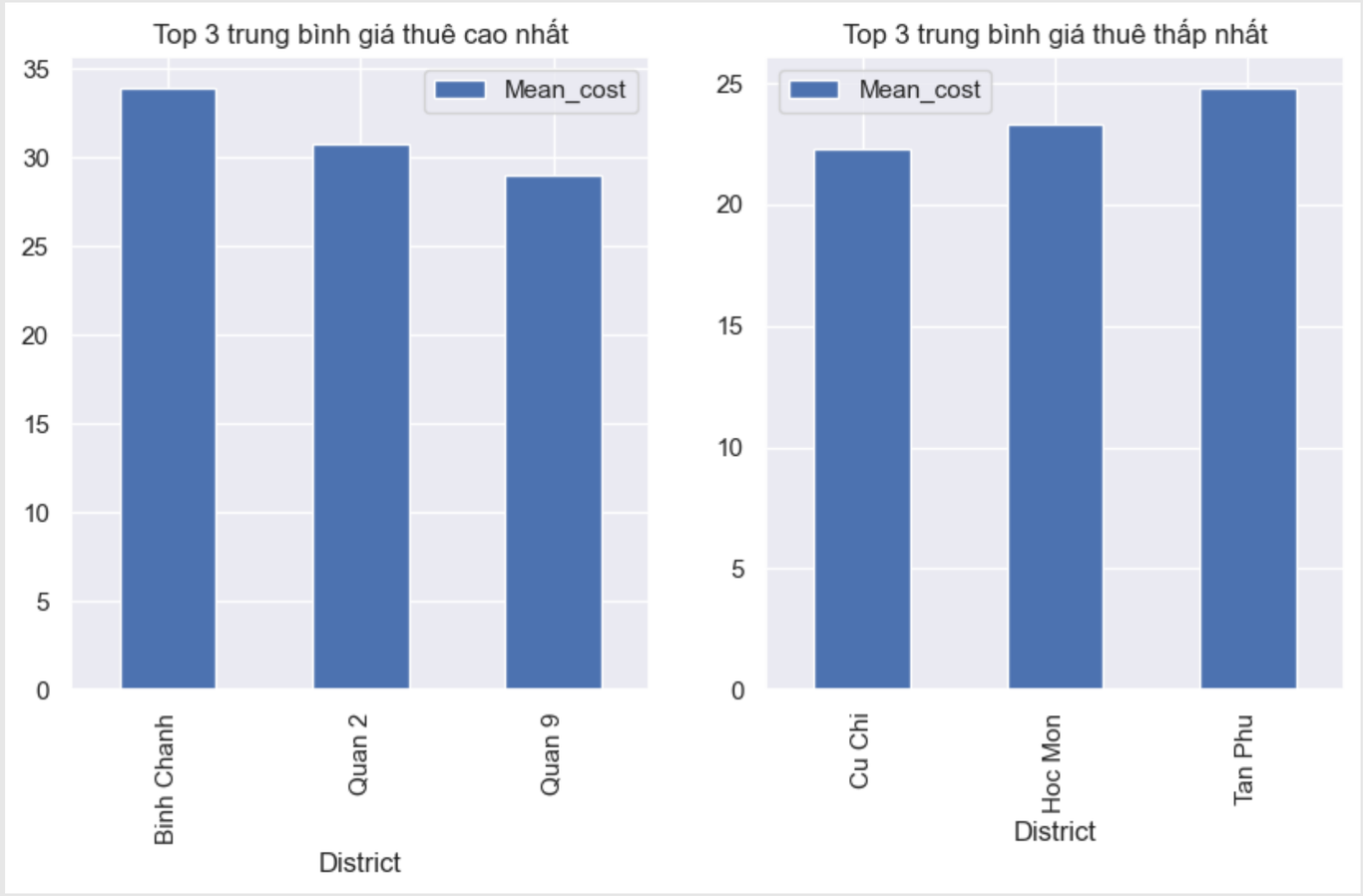


Với top 3 khu vực có mức phí thuê trọ cao nhất, các quận/huyện như Bình Chánh, quận 2, quận 9, ta có một số giải thích cho kết quả này như sau:

- Huyện Bình Chánh là một huyện có diện tích lớn nhất tại TP. Hồ Chí Minh, nó nằm trong khu vực ngoại ô thành phố và có rất nhiều khu dân cư và khu công nghiệp. Huyện Bình Chánh đang phát triển mạnh mẽ và trở thành một trong những huyện có tốc độ phát triển cao nhất tại TP. Hồ Chí Minh và có nhiều dự án đầu tư và xây dựng mới, do đó giá thuê trọ tại đây có thể khá cao so với một số huyện khác tại TP. Hồ Chí Minh.

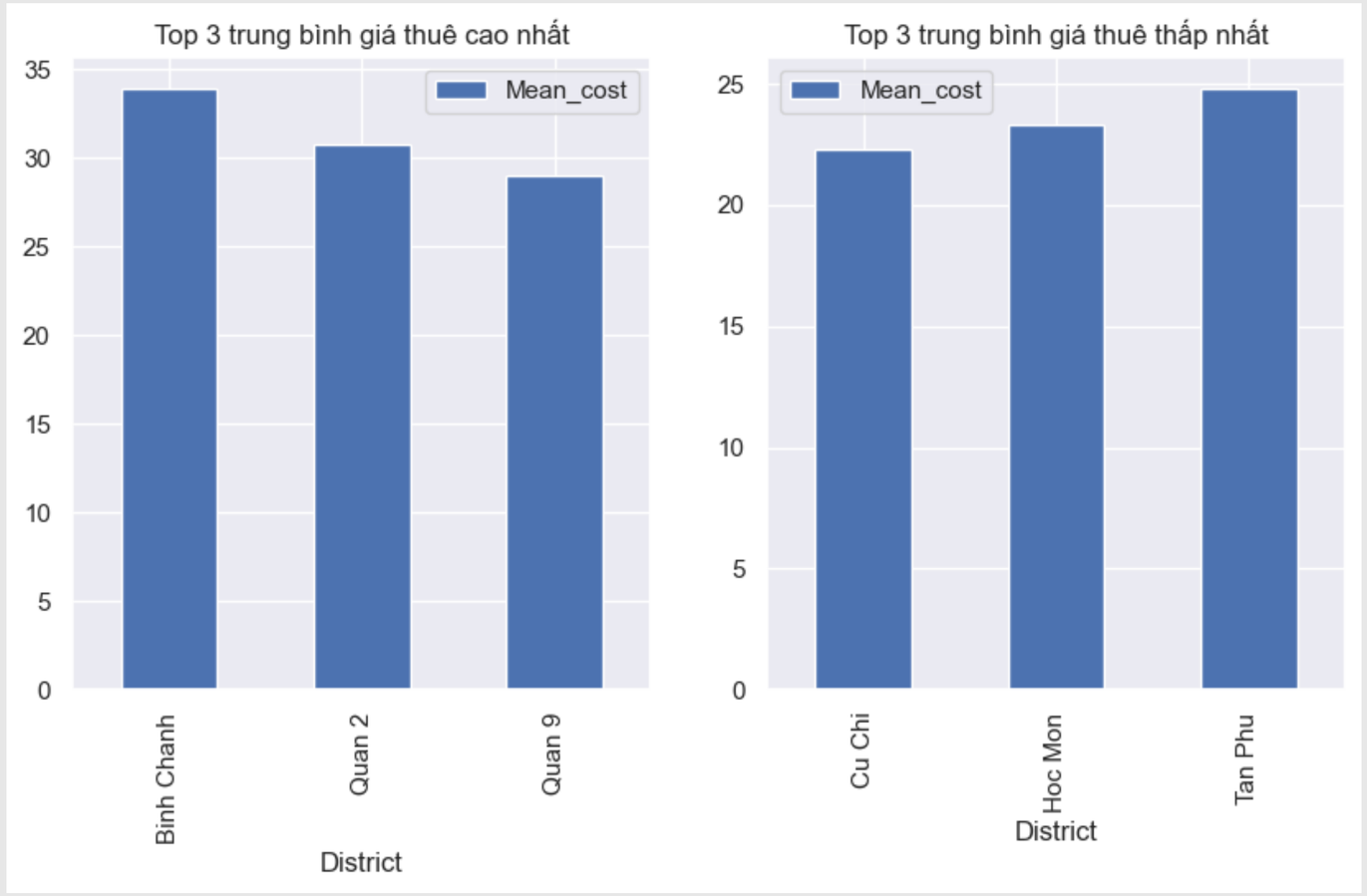


- Quận 9 là một trong những quận có giá thuê trọ cao tại TP. Hồ Chí Minh, do nó nằm trong khu vực ngoại ô thành phố và có rất nhiều tiện ích. Quận 9 có nhiều khu dân cư cao cấp và có một số trường học quốc tế hàng đầu, nên giá thuê trọ tại đây cũng khá cao.



Với top 3 khu vực có mức phí thuê trọ thấp nhất, các quận/huyện như Hóc Môn, quận Tân Phú, quận Củ Chi, ta có một số giải thích cho kết quả này như sau:

- Hóc Môn, Củ Chi là một khu vực ngoại ô của thành phố Hồ Chí Minh.
- Mức độ phát triển công nghiệp và kinh tế thấp hơn so với những khu vực khác trong thành phố. Điều này có thể dẫn đến một khối lượng người thuê trọ thấp hơn, và từ đó dẫn đến mức giá thuê trọ thấp hơn.



Với top 3 khu vực có mức phí thuê trọ thấp nhất, các quận/huyện như Hóc Môn, quận Tân Phú, quận Củ Chi, ta có một số giải thích cho kết quả này như sau:

- Khu vực có tỷ lệ người dân nghèo cao hơn so với một số khu vực khác trong thành phố. Điều này có thể dẫn đến mức giá thuê trọ thấp hơn, vì người cho thuê trọ có thể muốn hạn chế chi phí cho việc thuê trọ để giảm thiểu sự rủi ro cho họ.
- Số lượng trọ cơ sở và căn hộ cho thuê lớn hơn so với một số khu vực khác trong thành phố.

2. Sự biến thiên về số lượng giữa các nhà trọ mà nam giới và nữ giới có thể thuê được theo từng tiêu chí khác nhau?

Để trả lời cho câu hỏi này, ta sẽ làm như sau:

Tạo 2 dataframe dành cho những phòng trọ mà nam giới và nữ giới có thể thuê được:

	Price	Acreage	Tenant
43060	0.16	18.0	Nam
41408	0.18	25.0	Nam
42930	0.20	16.0	Nam
41588	0.25	69.0	Nam
43016	0.25	25.0	Nam
...	...	...	...
23858	430.00	25.0	Nam
18490	450.00	40.0	Nam
18260	450.00	40.0	Nam
18702	500.00	25.0	Nam
7990	550.00	200.0	Nam

43184 rows × 3 columns

Nam giới

	Price	Acreage	Tenant
47184	0.16	18.0	Nu
45348	0.18	25.0	Nu
47031	0.20	16.0	Nu
47128	0.25	25.0	Nu
45612	0.25	12.0	Nu
...	...	...	...
19545	450.00	40.0	Nu
19793	450.00	40.0	Nu
20038	500.00	25.0	Nu
8329	550.00	200.0	Nu
19095	800.00	20.0	Nu

47343 rows × 3 columns

Nữ giới

2. Sự biến thiên về số lượng giữa các nhà trọ mà nam giới và nữ giới có thể thuê được theo từng tiêu chí khác nhau?

Để trả lời cho câu hỏi này, ta sẽ làm như sau:

- Ta nhận thấy với cả 2 dataframe trên đều có những phòng trọ có giá thuê rất lớn. Thường giá thuê trọ với mốc 10 triệu đồng sẽ là mức giá khá đắn đo đối với rất nhiều người.
- Do đó, để tiện cho việc thống kê, ta sẽ chia giá trị cột 'Price' thành từng khoảng giá khác nhau, mỗi khoảng có độ chênh lệch 1 triệu và khoảng lớn nhất sẽ là những phòng trọ có giá cho thuê >= 10 triệu đồng

2. Sự biến thiên về số lượng giữa các nhà trọ mà nam giới và nữ giới có thể thuê được theo từng tiêu chí khác nhau?

2 dataframe lúc này:

	Price	Acreage	Tenant	Price range
43060	0.16	18.0	Nam	0-1
41408	0.18	25.0	Nam	0-1
42930	0.20	16.0	Nam	0-1
41588	0.25	69.0	Nam	0-1
43016	0.25	25.0	Nam	0-1
...	...	...	...	...
23858	430.00	25.0	Nam	>= 10
18490	450.00	40.0	Nam	>= 10
18260	450.00	40.0	Nam	>= 10
18702	500.00	25.0	Nam	>= 10
7990	550.00	200.0	Nam	>= 10

43184 rows × 4 columns

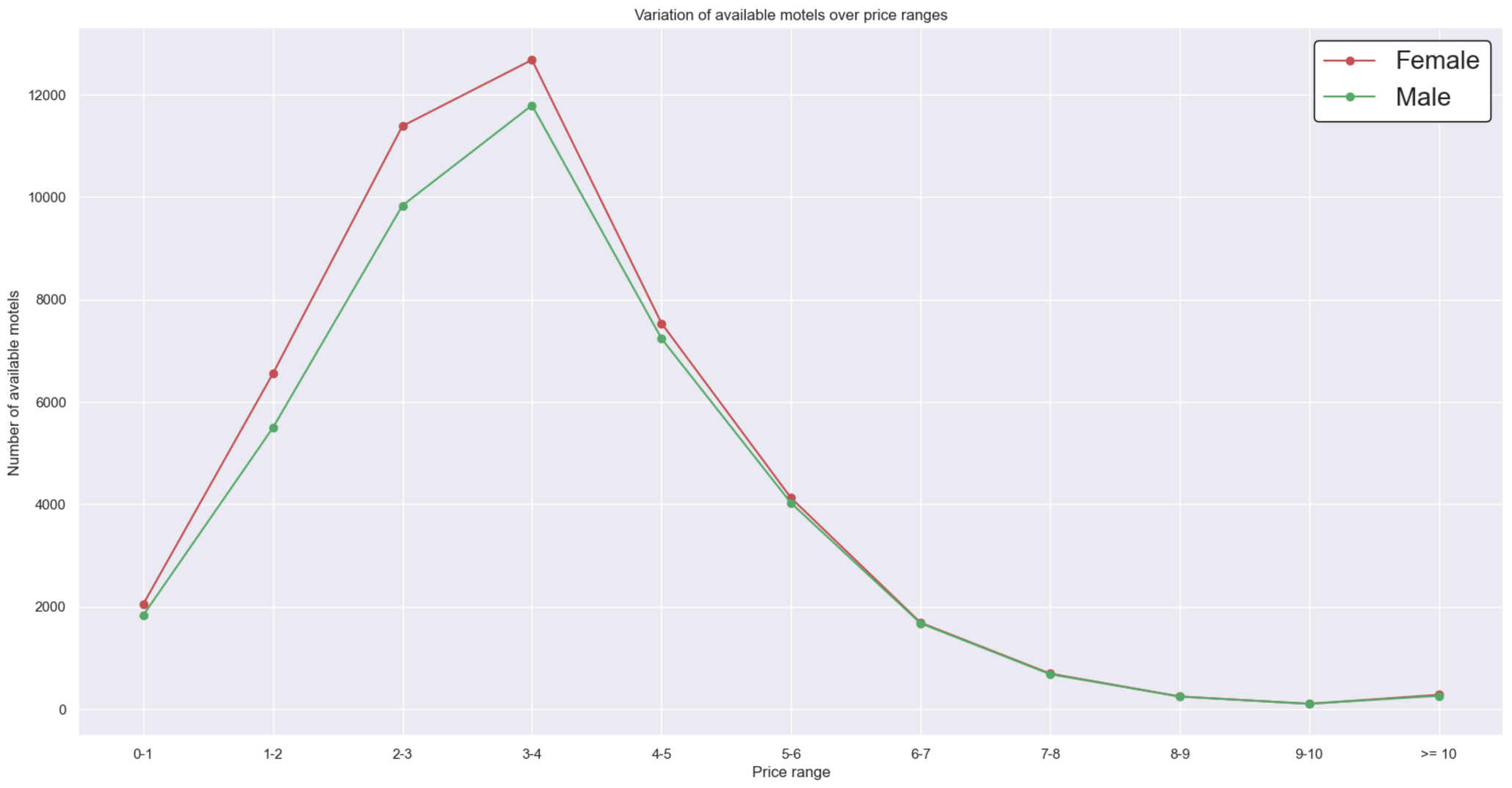
Nam giới

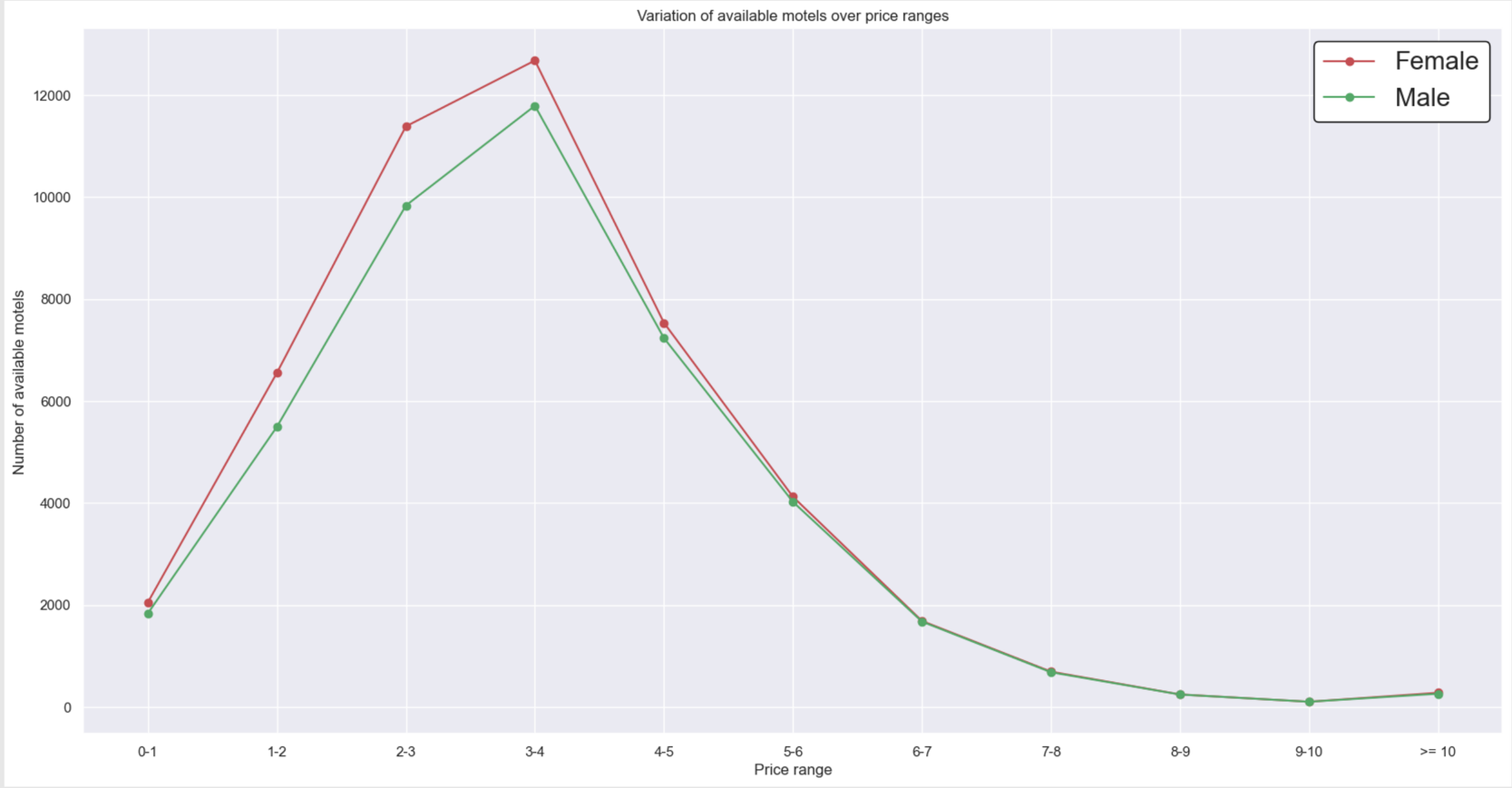
	Price	Acreage	Tenant	Price range
47184	0.16	18.0	Nu	0-1
45348	0.18	25.0	Nu	0-1
47031	0.20	16.0	Nu	0-1
47128	0.25	25.0	Nu	0-1
45612	0.25	12.0	Nu	0-1
...	...	...	...	...
19545	450.00	40.0	Nu	>= 10
19793	450.00	40.0	Nu	>= 10
20038	500.00	25.0	Nu	>= 10
8329	550.00	200.0	Nu	>= 10
19095	800.00	20.0	Nu	>= 10

47343 rows × 4 columns

Nữ giới

Vẽ biểu đồ đường thể hiện sự biến thiên số lượng của phòng trọ mà nam giới và nữ giới có thể thuê theo từng mức giá:

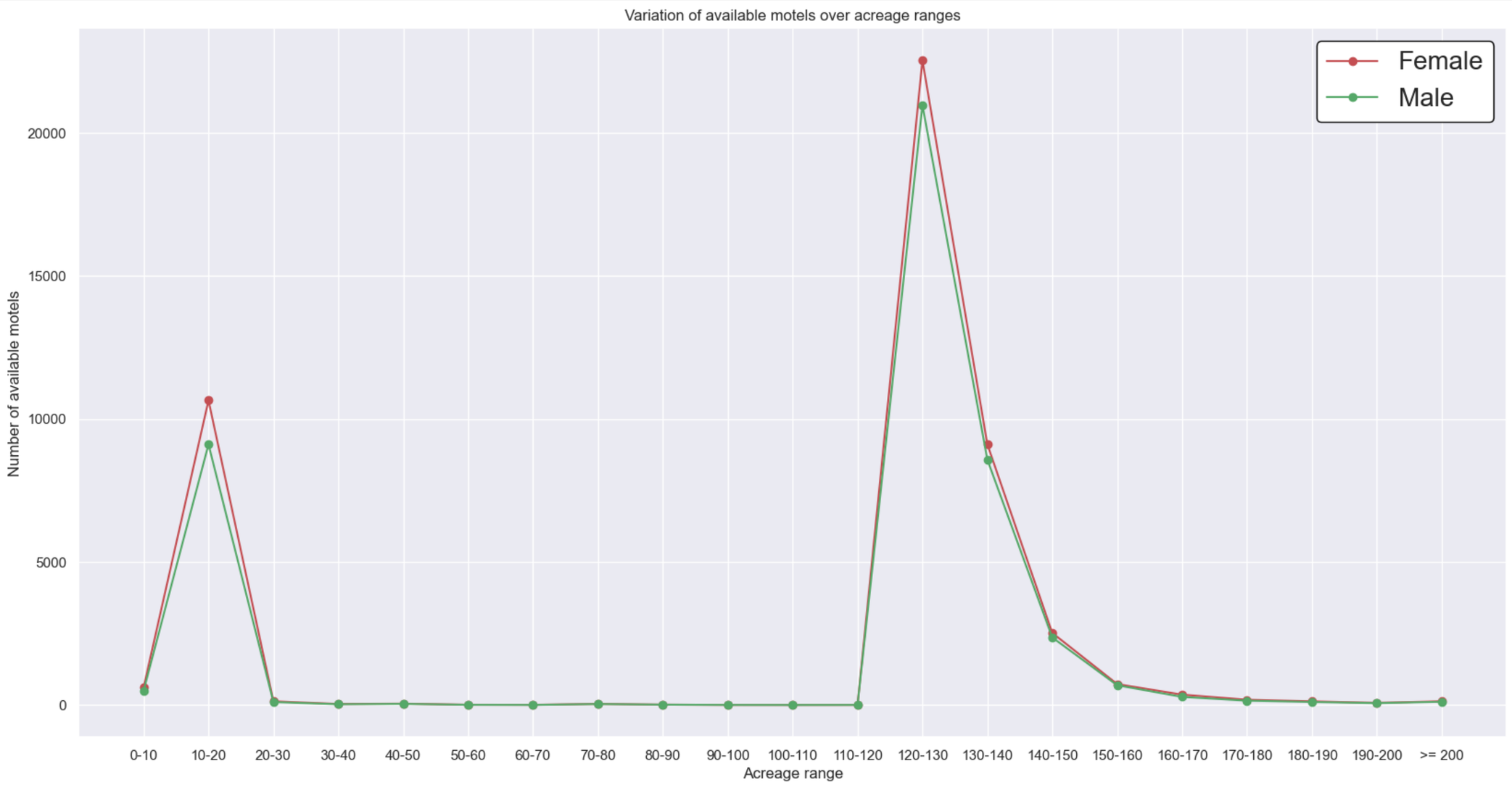


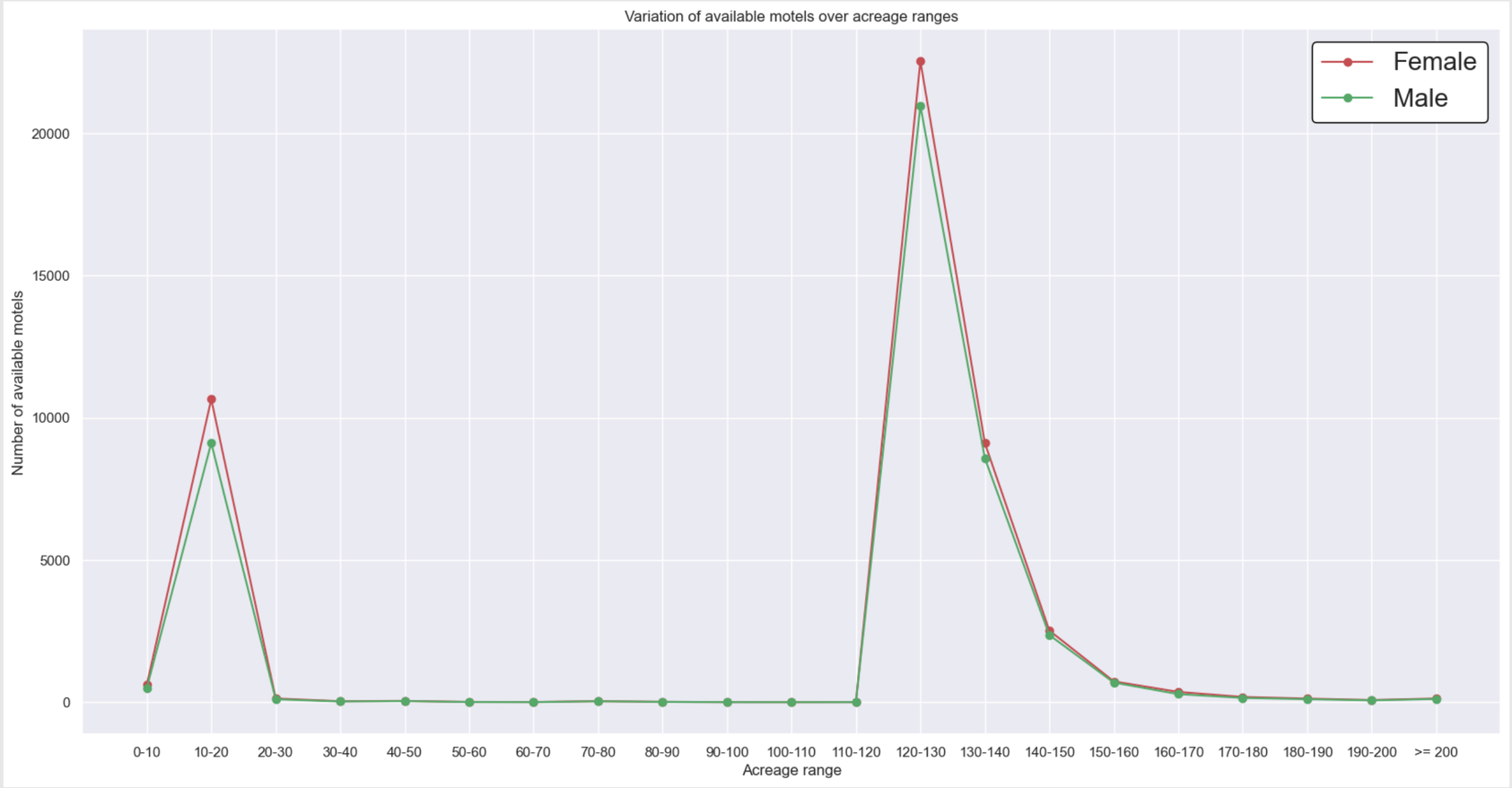


Nhận xét:

- Với những trọ mà nam và nữ có thể ở, số lượng trọ tăng đột biến (xấp xỉ 6 lần từ 2000 lên 12000) qua từng khoảng giá và đỉnh điểm là từ 3-4 triệu đồng, sau đó khi giá càng tăng lên thì số lượng trọ càng giảm mạnh. Có lẽ đây là mức giá phổ biến cho tất cả các trọ trên địa bàn thành phố Hồ Chí Minh.
- Số lượng trọ cho nữ luôn luôn nhiều hơn số trọ mà nam có thể thuê, với đỉnh điểm là chênh lệch gần 1500 trọ ở khoảng giá 2-3 triệu.
- Từ mức giá 4-5 triệu trở đi, giá càng lớn thì số trọ chênh lệch giữa 2 giới là không đáng kể, có lẽ với mức giá này thì trọ sẽ dành cho cả nam và nữ đều có thể đăng ký được.

Cũng làm tương tự, ta sẽ vẽ biểu đồ đường thể hiện sự biến thiên số lượng của phòng trọ mà nam giới và nữ giới có thể thuê theo từng diện tích:

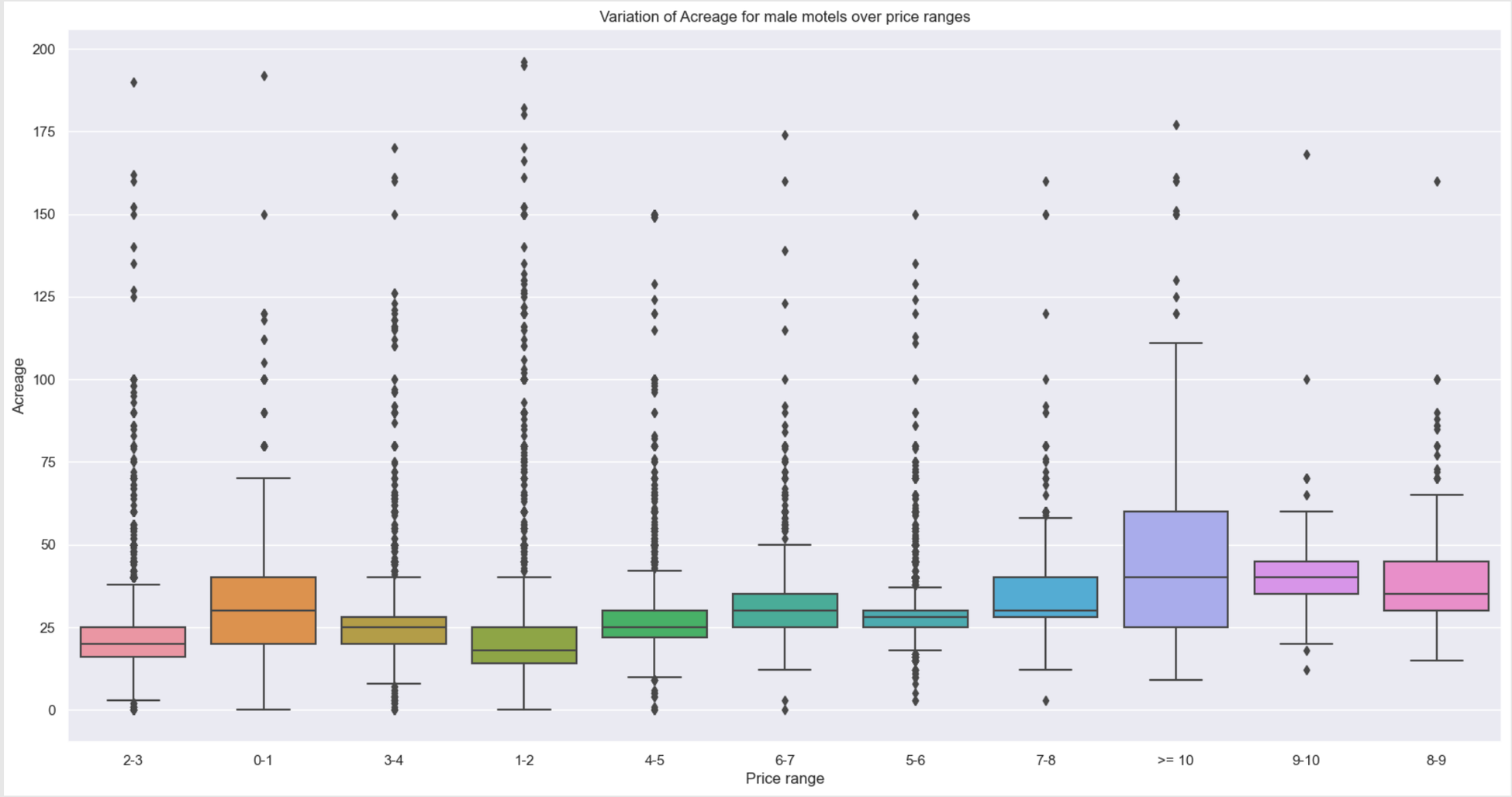




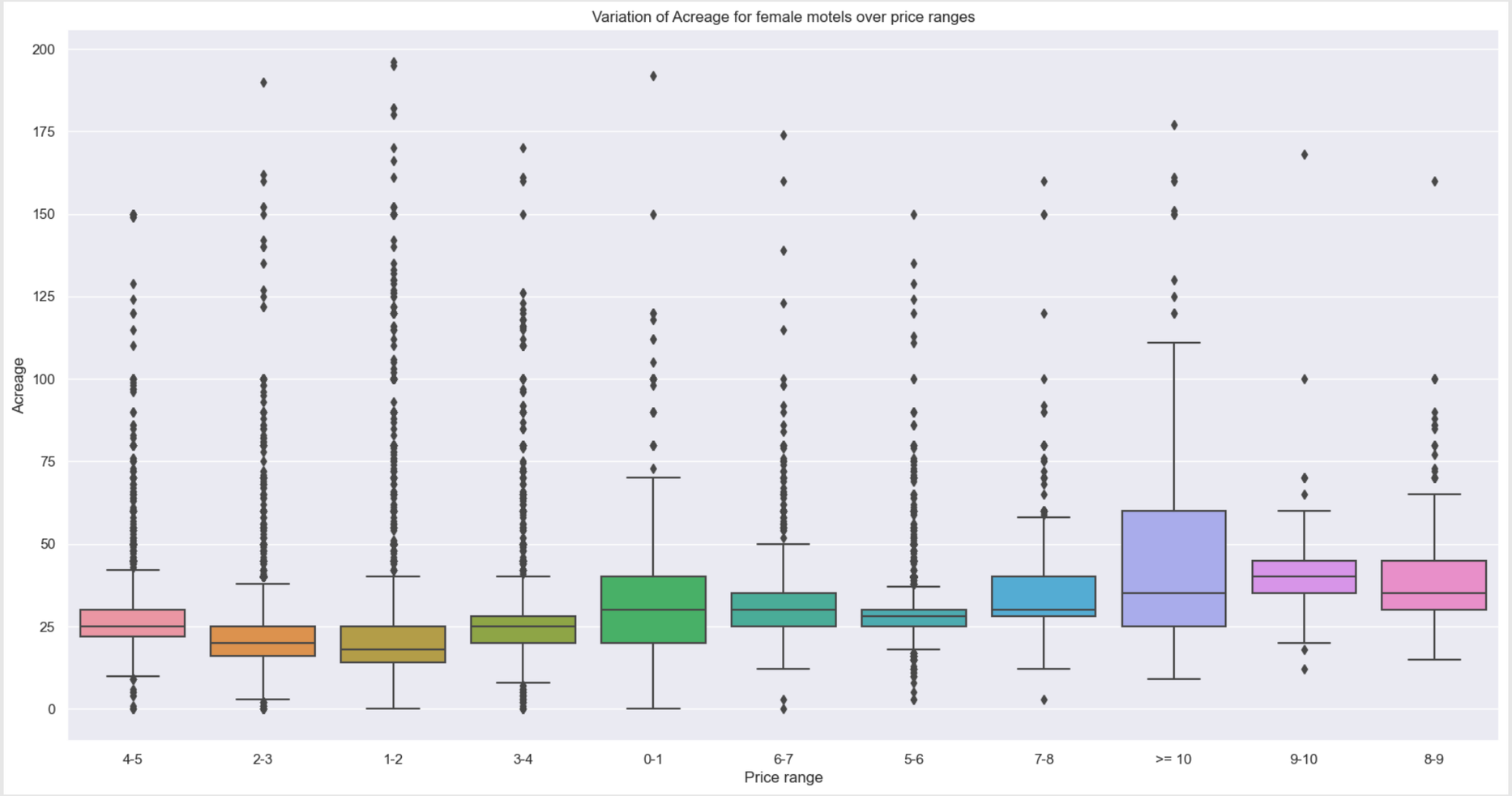
Nhận xét:

- Số lượng trọ tăng mạnh rồi giảm đột ngột ở 2 đoạn : từ diện tích **0-10** m^2 đến **20-30** m^2 (với đỉnh điểm là **10-20** m^2) và từ diện tích **110-120** m^2 đến **160-170** m^2 (đỉnh điểm là **120-130** m^2). Ở các diện tích còn lại thì số lượng trọ rất ít.
- Số lượng trọ cho nữ luôn luôn nhiều hơn số trọ mà nam có thể thuê, với đỉnh điểm là 2 khoảng diện tích đã nêu trên. Tuy nhiên lượng chênh lệch là không quá vượt trội. Điều này thể hiện rằng diện tích trọ sẵn có đối với nam và nữ là tương đương nhau.

Ta có thể quan sát thêm về mối quan hệ giữa khoảng giá và diện tích trọ đối với từng giới tính bằng biểu đồ hộp. Vì theo biểu đồ trên, số lượng trọ diện tích trên 200 m^2 là rất ít nên ta sẽ quan sát với những trọ bé hơn diện tích đấy:



Sự thay đổi diện tích phòng trọ mà nam giới có thể ở theo mức giá



Sự thay đổi diện tích phòng trọ mà nữ giới có thể ở theo mức giá

Nhận xét: Biểu đồ hộp ở cả 2 giới tính là khá tương tự nhau ở cùng 1 mức giá và đa số các khoảng giá từ 0 - 9 triệu luôn có mật độ diện tích dày đặc. Điều này càng thể hiện đây là mức giá hợp lý với đủ phong phú các loại trọ diện tích khác nhau để cả nam và nữ có thể cân nhắc thuê.

3. Số lượng nhà trọ cho thuê và giá thuê trung bình cho 1 m^2 qua từng năm ?

Để trả lời cho câu hỏi này, ta sẽ làm như sau:

- Bước 1: Tính giá trung bình 1 m^2 của mỗi phòng trọ.
- Bước 2: Gom nhóm phòng trọ theo năm.
- Bước 3: Trực quan hóa.

3. Số lượng nhà trọ cho thuê và giá thuê trung bình cho 1 m^2 qua từng năm ?

- Đầu tiên ta tính giá trung bình 1 m^2 của mỗi phòng trọ.

Id	1m^2	Year
614102	0.16000	2023
612543	0.14000	2023
212446	0.20000	2023
603145	0.09000	2023
315940	0.15625	2022

3. Số lượng nhà trọ cho thuê và giá thuê trung bình cho 1 m^2 qua từng năm ?

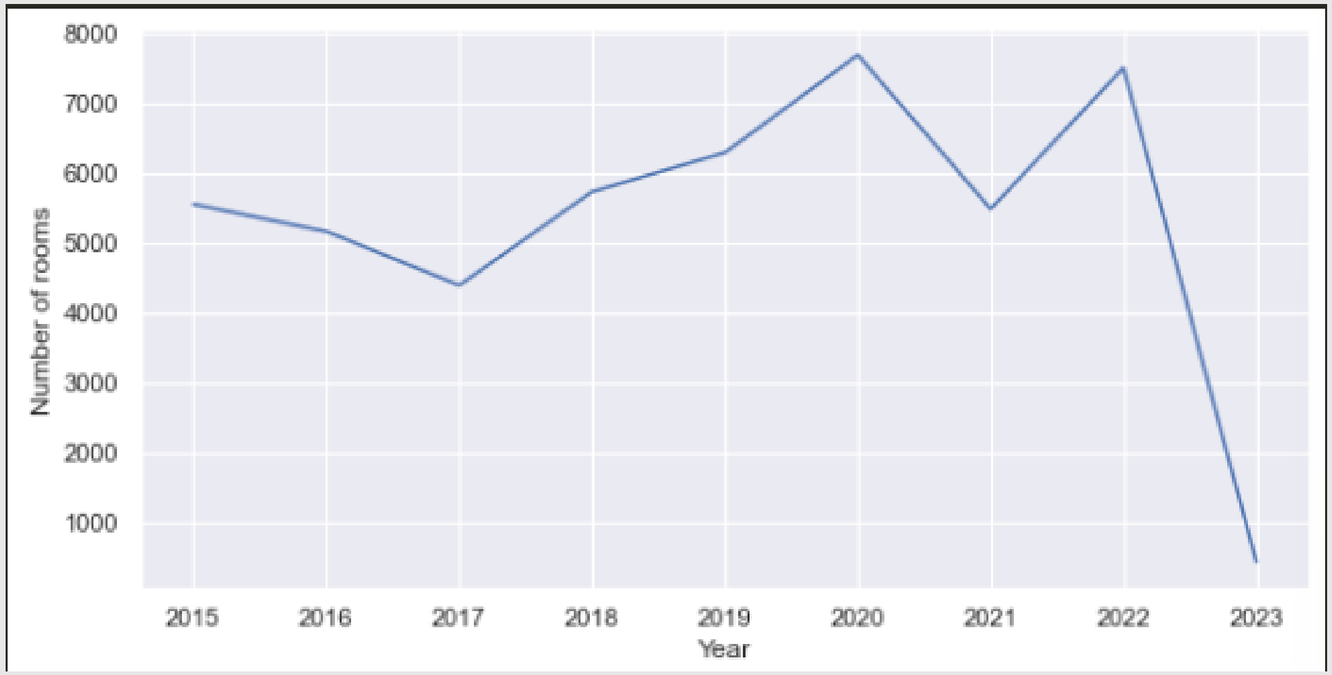
- Gom nhóm phòng trọ theo năm.

	Year	Number of rooms	Average price 1m^2
0	2015	5547	0.142770
1	2016	5158	0.153154
2	2017	4387	0.141204
3	2018	5728	0.162845
4	2019	6287	0.152702
5	2020	7684	0.150147
6	2021	5481	0.147781
7	2022	7501	0.148984
8	2023	430	0.150894

3. Số lượng nhà trọ cho thuê và giá thuê trung bình cho 1 m^2 qua từng năm ?

- Trực quan hóa dữ liệu vừa tính được.

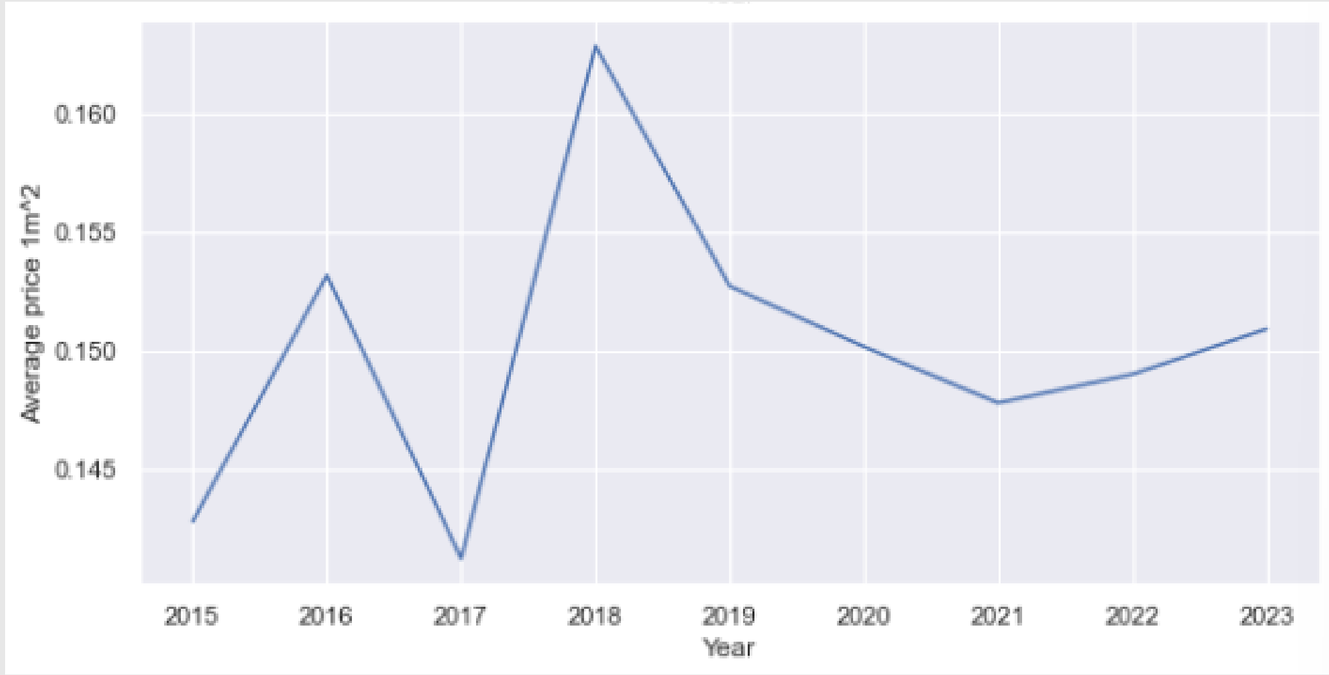
➢ **Nhận xét**: Giá 1m^2 phòng trọ cho thuê tăng giảm liên tục qua từng năm. Năm 2017 có giá 1m^2 cho thuê thấp nhất nhưng qua năm 2018 lại có giá cao nhất. Từ năm 2019 đến nay giá 1m^2 phòng trọ cho thuê đã được ổn định hơn.



3. Số lượng nhà trọ cho thuê và giá thuê trung bình cho 1 m^2 qua từng năm ?

- Trực quan hóa dữ liệu vừa tính được.

➢ **Nhận xét**: Số lượng phòng trọ thay đổi không ổn định nhưng tổng thể là vẫn tăng, năm 2023 có số lượng phòng thấp như vậy là vì đó chỉ là số liệu của tháng 1 năm 2023.



4. Diễn biến giá trọ của các quận ở mỗi tháng trong năm ?

Để trả lời cho câu hỏi này, ta sẽ làm như sau:

- Bước 1: Tính giá trung bình $1m^2$ của mỗi phòng trọ theo quận và tháng.
- Bước 2: Tạo bảng giá trung bình $1m^2$, mỗi dòng ứng với 1 quận, mỗi cột ứng với 1 tháng trong năm.
- Bước 3: Trực quan hóa bằng heatmap.

4. Diễn biến giá trọ của các quận ở mỗi tháng trong năm ?

- Tính giá trung bình 1m^2 của mỗi phòng trọ theo quận và tháng.

Id	District	Month	1m^2
614102	Tan Binh	1	0.16000
612543	Go Vap	1	0.14000
212446	Tan Binh	1	0.20000
603145	Tan Phu	1	0.09000
315940	Quan 2	12	0.15625

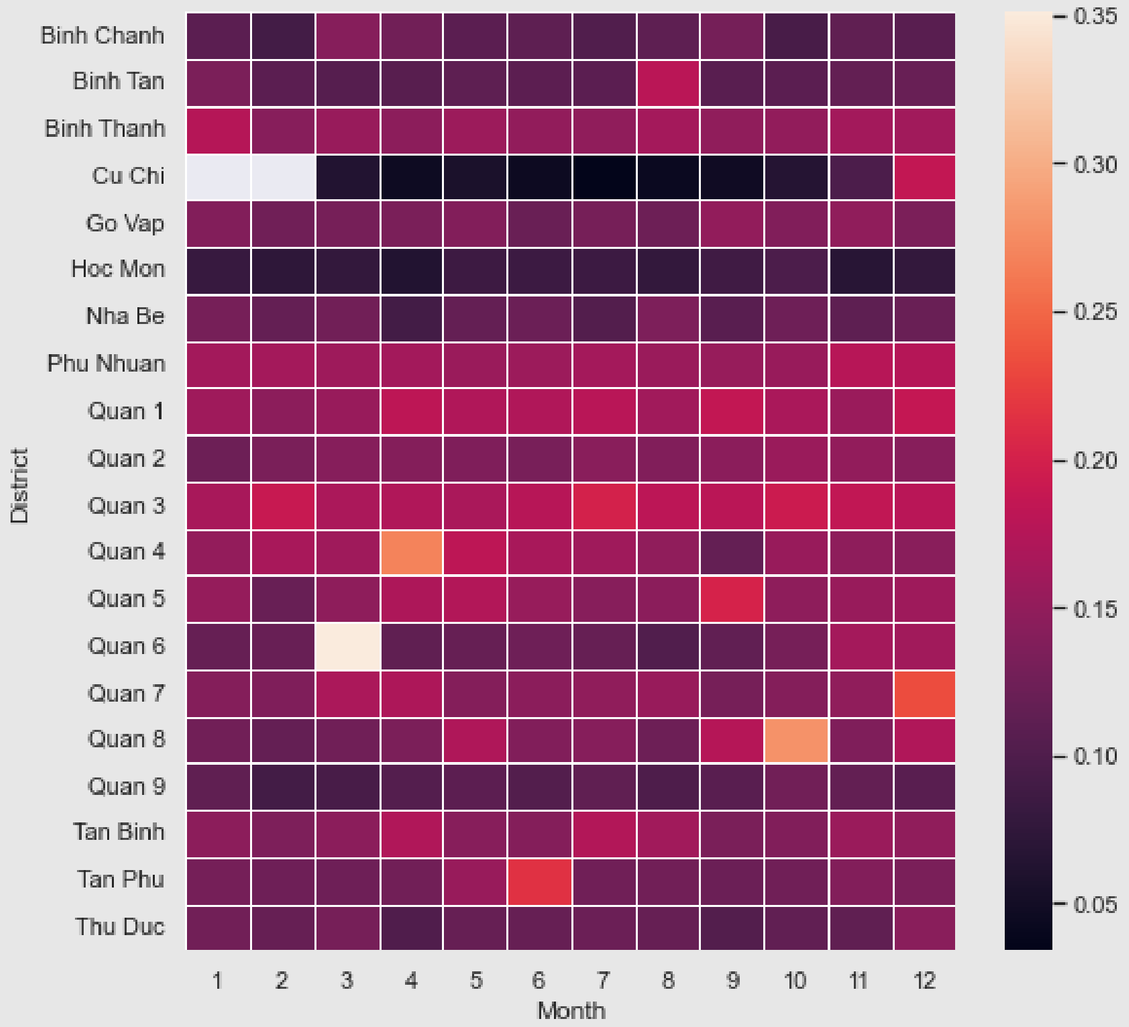
4. Diễn biến giá trọ của các quận ở mỗi tháng trong năm ?

- Tạo bảng giá trung bình $1m^2$, mỗi dòng ứng với 1 quận, mỗi cột ứng với 1 tháng trong năm.

Month	1	2	3	4	5	6	7	8	9	10	11	12
District												
Binh Chanh	0.109284	0.090430	0.140823	0.126296	0.109597	0.111066	0.101945	0.112027	0.128986	0.093798	0.112428	0.107807
Binh Tan	0.133765	0.109291	0.105318	0.106936	0.111331	0.110294	0.108706	0.179080	0.108114	0.109394	0.115759	0.118378
Binh Thanh	0.175604	0.142474	0.155318	0.146277	0.157608	0.149584	0.148471	0.163215	0.148069	0.149416	0.162604	0.160636
Cu Chi	NaN	NaN	0.062917	0.045455	0.056875	0.044701	0.034091	0.042917	0.047667	0.065185	0.097857	0.185417
Go Vap	0.138464	0.124880	0.129697	0.132937	0.139186	0.119735	0.129914	0.122279	0.149270	0.137798	0.148213	0.133508
Hoc Mon	0.080654	0.072937	0.076464	0.062873	0.085305	0.083911	0.084113	0.077350	0.088317	0.097228	0.068316	0.076903
Nha Be	0.130420	0.116846	0.126717	0.091031	0.117020	0.121001	0.103277	0.134659	0.107519	0.123449	0.111947	0.120556
Phu Nhuan	0.162145	0.163113	0.159032	0.162394	0.155651	0.157710	0.163994	0.156454	0.153858	0.155126	0.177168	0.175952
Quan 1	0.160135	0.145669	0.154212	0.182345	0.172084	0.172232	0.178272	0.161570	0.186013	0.167598	0.155630	0.186929
Quan 2	0.122062	0.132195	0.141060	0.140353	0.135682	0.130941	0.143296	0.137047	0.144494	0.156355	0.149713	0.142549
Quan 3	0.166267	0.189726	0.169007	0.171629	0.167645	0.177409	0.200039	0.181050	0.179969	0.193352	0.184710	0.178561
Quan 4	0.151138	0.165480	0.159492	0.269143	0.181420	0.166483	0.160340	0.147977	0.117022	0.154480	0.147812	0.143616
Quan 5	0.151757	0.118808	0.147050	0.169403	0.173675	0.153793	0.142638	0.145213	0.201913	0.147038	0.154647	0.158582
Quan 6	0.117336	0.118965	0.350989	0.112394	0.117657	0.123809	0.117514	0.101656	0.113635	0.129418	0.163368	0.161041
Quan 7	0.140100	0.136415	0.168052	0.169746	0.140292	0.144664	0.149031	0.154931	0.129709	0.139923	0.149054	0.233210
Quan 8	0.126593	0.116592	0.125530	0.134047	0.171326	0.137476	0.140576	0.122386	0.176446	0.280095	0.135686	0.171579
Quan 9	0.113165	0.090016	0.093979	0.104171	0.109635	0.102876	0.112372	0.098476	0.107686	0.126513	0.114778	0.107521
Tan Binh	0.145709	0.134410	0.144946	0.171603	0.142847	0.139833	0.173639	0.161084	0.132570	0.137941	0.156474	0.148338
Tan Phu	0.128251	0.123994	0.123352	0.126227	0.154240	0.214161	0.124540	0.126063	0.121058	0.124642	0.138641	0.132611
Thu Duc	0.126380	0.117371	0.129612	0.100512	0.118080	0.117874	0.121267	0.118195	0.103179	0.113544	0.113019	0.143910

4. Diễn biến giá trọ của các quận ở mỗi tháng trong năm ?

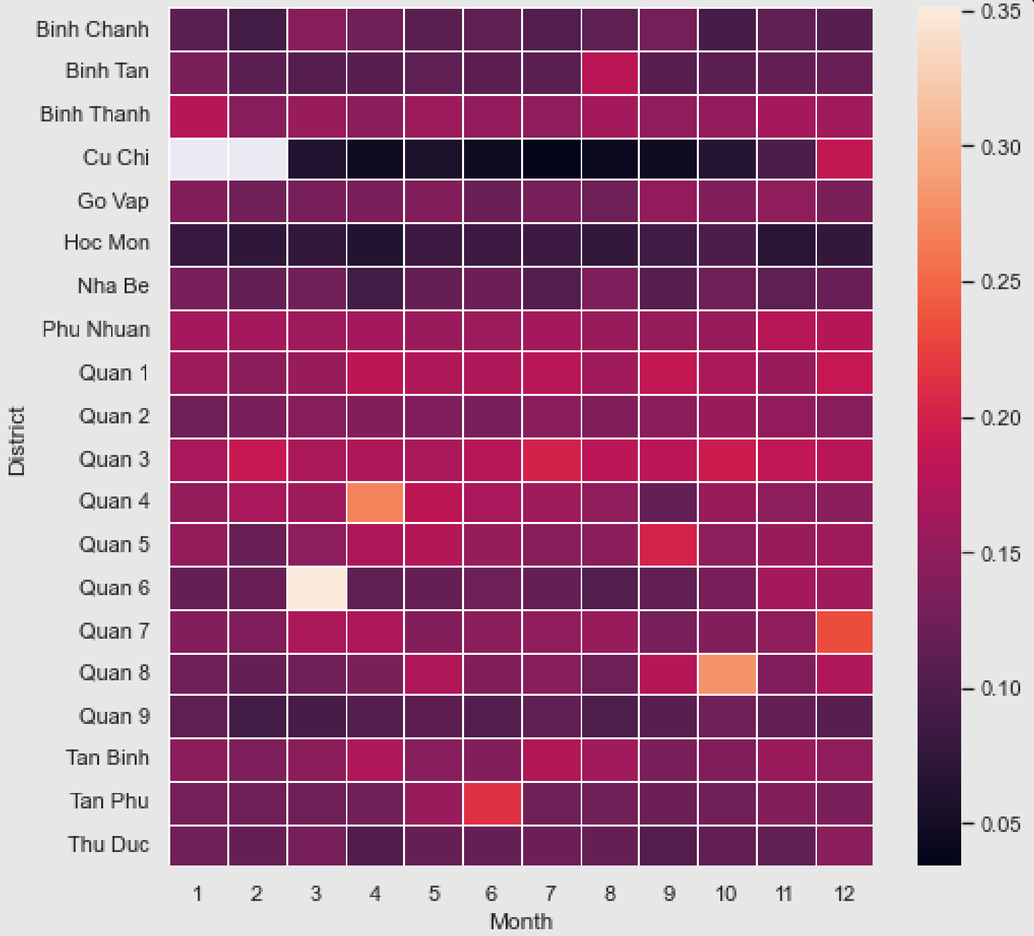
- Trực quan hóa dữ liệu bằng heatmap.



- **Nhận xét:**

- Nhìn vào biểu đồ, ta thấy vào trong khoảng thời gian từ tháng 12 đến tháng 1 (cuối năm - tết nguyên đán) và từ tháng 9 đến tháng 10 (dịp sinh viên nhập học), giá thuê phòng trọ trên mỗi mét vuông là lớn hơn so với các tháng khác trong năm.

- Giá thuê phòng trọ trên mỗi mét vuông qua các tháng ở Củ Chi thấp hơn so với các quận huyện khác. Thậm chí tháng 1,2 không có bài post nào được đăng.



NHÓM 6

THANK YOU
FOR WATCHING