BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

✡ ✡

HCMUTE

BÁO CÁO ĐỒ ÁN CUỐI KỲ

ĐỀ TÀI: HOUSE RENT PRICING

Môn học: KHAI PHÁ DỮ LIỆU
Mã lớp học phần: 222DAMI330484
GVHD: ThS. Nguyễn Văn Thành

Nhóm sinh viên thực hiện: Nhóm 11

Trần Lê Ngọc Gia Hân	20133038
Nguyễn Thị Hoàng Trang	20133099
Nguyễn Thị Thanh Ngân	20133018

TP. Hồ Chí Minh, 24 tháng 05 năm 2023

DANH SÁCH THÀNH VIÊN THAM GIA

THỰC HIỆN ĐỀ TÀI VÀ VIẾT BÁO CÁO

Môn: Khai Phá Dữ Liệu - *HỌC KÌ II – NĂM HỌC 2022 – 2023*

STT	HỌ VÀ TÊN	MSSV	TỶ LỆ ĐÓNG GÓP
1	Trần Lê Ngọc Gia Hân	20133038	100%
2	Nguyễn Thị Hoàng Trang	20133099	100%
3	Nguyễn Thị Thanh Ngân	20133018	100%

Nhận xét của giảng viên:

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Ngày … tháng 05 năm 2023

Giảng viên chấm điểm

Ths. Nguyễn Văn Thành

LỜI CẢM ƠN

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến **Trường Đại học Sư phạm Kỹ thuật TPHCM** đã đưa môn học **Khai Phá Dữ Liệu** vào chương trình giảng dạy. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – **Th.S Thầy Nguyễn Văn Thành** đã dạy dỗ, truyền đạt những kiến thức quý báu cho chúng em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học Kho Dữ Liệu của Thầy, nhóm em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để chúng em có thể vững bước sau này. Bộ môn Khai Phá Dữ Liệu là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài đồ án này khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong Thầy xem xét và góp ý để bài đồ án của chúng em được hoàn thiện hơn.

Nhóm chúng em xin chân thành cảm ơn!

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI 11
1.1. Lý do chọn đề tài 11
1.2. Tổng quan về tập dữ liệu 11
1.2.1. Nguồn dữ liệu 11
1.2.2. Mô tả chi tiết tập dữ liệu 11
1.3. Giới thiệu về các công cụ được sử dụng 12
CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU 13
2.1. Tải dữ liệu 13
2.2. Tải dữ liệu vào SQL Server 2019 15
CHƯƠNG 3: TRỰC QUAN HÓA DỮ LIỆU 18
3.1. Trực quan hóa dữ liệu trên công cụ Python 18
3.1.1. Quy mô của các Houses/Apartments/Flats cho thuê có ảnh hưởng đến giá thuê không? 18
3.1.2. Loại khu vực của các Houses/Apartments/Flats cho thuê có liên quan đến giá cho thuê hay không? 22
3.1.3. Những loại người thuê nhà được ưa thích? 23
3.2. Trực quan hóa dữ liệu trên công cụ Power BI 25
3.2.1. Dashboard 25
3.2.2. Đánh giá giá thuê nhà giữa các thành phố 26
3.2.3. Tương quan giữa các bên cho thuê và giá thuê nhà 26
3.2.4. Tương quan giữa diện tích và giá thuê nhà 29
3.2.5. Tương quan giữa tình trạng nội thất và giá thuê nhà 30
3.2.6. Tương quan giữa số phòng của căn nhà và giá thuê nhà 32
CHƯƠNG 4: KHAI PHÁ DỮ LIỆU BẰNG SSAS 33
4.1. Quá trình tạo mô hình 33
4.2. Association Rule Mining 35
4.2.1. Giới Thiệu 35
4.2.2. Tiến Hành Khai Phá Luật Kết Hợp 35
4.3. Clustering 46
4.3.1. Giới Thiệu 46
4.3.2. Tiến Hành Khai Phá Phân Cụm 46
4.4. Classification 55
4.4.1. Giới Thiệu 55
4.4.2. Decision Tree 55
4.4.2.1. Khai phá dữ liệu **Error! Bookmark not defined.**
4.4.2.2. Đánh giá mô hình Decision tree 61
CHƯƠNG 5: KẾT LUẬN 76
5.1. Kết quả đạt được 76
5.2. Những hạn chế 76
5.3. Tài liệu tham khảo 76

PHÂN CÔNG NHIỆM VỤ

Nhiệm vụ	**Trần Lê Ngọc Gia Hân**	**Nguyễn Thị Thanh Ngân**	**Nguyễn Thị Hoàng Trang**
Tìm kiếm tập dữ liệu	x		
Tiền xử lý dữ liệu	x	x	x
Trực quan hóa dữ liệu			
+ Dùng ngôn ngữ python		x	
+ Dùng công cụ Power BI			x
Tìm hiểu và khai phá dữ liệu			
+ Sử dụng giải thuật Microsoft Association Rules		x	
+ Sử dụng giải thuật Microsoft Clustering	x		
+ Sử dụng giải thuật Microsoft Decision Trees			x

CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

1.1. Lý do chọn đề tài

- Nhà ở tại Ấn Độ phát triển lâu đời từ những lâu đài cổ xưa cho đến những tòa nhà cao tầng hiện đại trong các thành phố lớn đến những túp lều nhỏ ở các thị trấn xa xôi. Với mức thu nhập tăng lên theo từng giai đoạn, lĩnh vực nhà ở ở Ấn Độ cũng có sự tăng trưởng vượt bậc. Thuê nhà, hay mướn nhà là hình thức để sử dụng nhà ở một các tạm thời mà ngôi nhà hay căn hộ đó được sở hữu bởi một người chủ khác. Thuê nhà là một cách để chia sẻ về mặt kinh tế khi người thuê thì trả tiền thuê để ở và người chủ thì chi trả cho phí tài sản phát sinh bởi quyền sở hữu. Chi phí phải trả để thuê một căn nhà có thể được định giá dựa trên nhiều yếu tố, từ vị trí thành phố nơi nhà hay căn hộ đó tọa lạc, kích thước nhà hay căn hộ, đến số lượng các phòng trong căn nhà như số phòng tắm, phòng ngủ,... mà chi phí đó có thể biến thiên từ rất thấp đến rất cao. Với mục tiêu tìm hiểu và dự đoán xem giá thuê của một căn nhà có phải là phù hợp hay không, hay giá thuê đó đã bị phóng đại thành một con số khổng lồ, nhóm chúng em đã chọn tập dữ liệu 'House Rent Dataset' để thực hiện trực quan và dự đoán giá thuê này của hơn 4700 nhà cho thuê tại Ấn Độ được đăng năm 2022.

1.2. Tổng quan về tập dữ liệu

1.2.1. Nguồn dữ liệu

- Nhóm sử dụng tập dữ liệu **House Rent Dataset** trên Kaggle. Dữ liệu này cung cấp thông tin về các căn hộ/phòng cho thuê ở Ấn Độ được mô tả dựa trên nhiều đặc điểm khác nhau như kích thước phòng, vị trí căn hộ, số lượng phòng cho thuê, đối tượng ưu tiên,.. giá cho thuê.
- Nguồn dữ liệu: House Rent Dataset Kaggle

1.2.2. Mô tả chi tiết tập dữ liệu

- Tập dữ liệu gồm có hơn 4.700 dòng (rows) là các nhà cho thuê và 12 cột (columns) là thông tin về căn hộ cho thuê đó. Cụ thể là

- BHK: Số lượng phòng ngủ(Bedrooms), sảnh(Hall), phòng bếp (Kitchen).
- Rent: Giá cho thuê của nhà/ căn hộ.
- Size: Diện tích nhà/ căn hộ (feet vuông).
- Floor: Tầng đang cho thuê trong tổng số các tầng (Example: Ground out of 2, 3 out of 5, etc.)
- Area Type: Loại khu vực.
- Area Locality: Vị trí của nhà/ căn hộ.
- City: Thành phố nơi nhà/ căn hộ tọa lạc.
- Furnishing Status: Tình trạng nội thất trong nhà/ căn hộ (Nội thất đầy đủ > Nội thất chưa đầy đủ > Không có nội thất).

- Tenant Preferred: Đối tượng ưu tiên cho thuê.
- Bathroom: Số lượng phòng tắm.
- Point of Contact: Người cần liên lạc nếu cần biết thêm thông tin cho thuê nhà/ căn hộ.

1.3. Giới thiệu về các công cụ được sử dụng

- Jupyter Notebook: Sử dụng python để tiền xử lý dữ liệu
- SQL Server 2019: Tải dữ liệu từ file .csv vào database
- Visual Studio 2019 + Analysis Services Multidimensional and Data Mining Project (SSAS): Thực hiện khai phá dữ liệu
- Power BI: Trực quan hóa dữ liệu.

CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU

2.1. Tải dữ liệu

- Thư viện chính được sử dụng để tiền xử lý dữ liệu là thư viện pandas
- Tải dữ liệu

```
In [13]: house_rent_dataset = pd.read_csv('House_Rent_Dataset.csv')

         house_rent_dataset.head()
```

Out[13]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner

- Xem mô tả về tập dữ liệu

```
house_rent_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4746 entries, 0 to 4745
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Posted On         4746 non-null   object
 1   BHK               4746 non-null   int64
 2   Rent              4746 non-null   int64
 3   Size              4746 non-null   int64
 4   Floor             4746 non-null   object
 5   Area Type         4746 non-null   object
 6   Area Locality     4746 non-null   object
 7   City              4746 non-null   object
 8   Furnishing Status 4746 non-null   object
 9   Tenant Preferred  4746 non-null   object
 10  Bathroom          4746 non-null   int64
 11  Point of Contact  4746 non-null   object
dtypes: int64(4), object(8)
memory usage: 445.1+ KB
```

-> Ta thấy rằng tập dữ liệu không có missing value

- Đổi tên cột ‘BHK’ thành ‘Number of Bedrooms-Halls-Kitchens’

```python
house_rent_dataset = house_rent_dataset.rename(columns={'BHK': 'Number of Bedrooms-Halls-Kitchens'})
house_rent_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4746 entries, 0 to 4745
Data columns (total 12 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Posted On                       4746 non-null   object
 1   Number of Bedrooms-Halls-Kitchens 4746 non-null   int64
 2   Rent                            4746 non-null   int64
 3   Size                            4746 non-null   int64
 4   Floor                           4746 non-null   object
 5   Area Type                       4746 non-null   object
 6   Area Locality                   4746 non-null   object
 7   City                            4746 non-null   object
 8   Furnishing Status               4746 non-null   object
 9   Tenant Preferred                4746 non-null   object
 10  Bathroom                        4746 non-null   int64
 11  Point of Contact                4746 non-null   object
dtypes: int64(4), object(8)
memory usage: 445.1+ KB
```

- Thêm hai cột ‘Recent Floor’ và ‘Total Floors’ được tách từ cột ‘Floor’

```python
house_rent_dataset[['Recent Floor', 'Total Floors']] = house_rent_dataset['Floor'].str.split(' out of ', expand=True)
house_rent_dataset.head()
```

	Posted On	Number of Bedrooms-Halls-Kitchens	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact	Recent Floor	Total Floors
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner	Ground	2
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner	1	3
2	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner	1	3
3	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner	1	2
4	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner	1	2

- Sau đó loại bỏ cột ‘Floor’ khỏi tập dữ liệu bằng lệnh drop()

```python
house_rent_dataset.drop('Floor', axis=1, inplace=True)
house_rent_dataset.head()
```

Python

	Posted On	Number of Bedrooms-Halls-Kitchens	Rent	Size	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact	Recent Floor	Total Floors
0	2022-05-18	2	10000	1100	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner	Ground	2
1	2022-05-13	2	20000	800	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner	1	3
2	2022-05-16	2	17000	1000	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner	1	3
3	2022-07-04	2	10000	800	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner	1	2
4	2022-05-09	2	7500	850	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner	1	2

- Sắp xếp lại thứ tự các thuộc tính

```python
house_rent_dataset = house_rent_dataset.reindex(columns=['Posted On', 'Number of Bedrooms-Halls-Kitchens', 'Recent Floor', 'Total Floors', 'Rent',
'Size', 'Area Type', 'Area Locality', 'City', 'Furnishing Status', 'Tenant Preferred', 'Bathroom', 'Point of Contact',])
house_rent_dataset.head()
```

Python

	Posted On	Number of Bedrooms-Halls-Kitchens	Recent Floor	Total Floors	Rent	Size	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2022-05-18	2	Ground	2	10000	1100	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2022-05-13	2	1	3	20000	800	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2022-05-16	2	1	3	17000	1000	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2022-07-04	2	1	2	10000	800	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2022-05-09	2	1	2	7500	850	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner

- Tiếp theo ta tiến hành đổi dữ liệu cho cột ‘Recent Floor’.
  - Nếu Recent Floor = ‘Ground’ thì đổi thành 0
  - Nếu Recent Floor = ‘Upper Basement’ thì đổi thành -1
  - Nếu Recent Floor = ‘Lower Basement’ thì đổi thành -2

```python
house_rent_dataset.loc[house_rent_dataset['Recent Floor'] == 'Ground', 'Recent Floor'] = 0
house_rent_dataset.loc[house_rent_dataset['Recent Floor'] == 'Upper Basement', 'Recent Floor'] = -1
house_rent_dataset.loc[house_rent_dataset['Recent Floor'] == 'Lower Basement', 'Recent Floor'] = -2
house_rent_dataset.head()
```

Python

	Posted On	Number of Bedrooms-Halls-Kitchens	Recent Floor	Total Floors	Rent	Size	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2022-05-18	2	0	2	10000	1100	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2022-05-13	2	1	3	20000	800	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2022-05-16	2	1	3	17000	1000	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2022-07-04	2	1	2	10000	800	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2022-05-09	2	1	2	7500	850	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner

- Kiểm tra và xóa các giá trị na trong cột 'Total Floors' đã xử lý ở trên
- Tạo index cho các dòng, đánh số từ 1 đến hết.

```python
house_rent_dataset[house_rent_dataset['Total Floors'].isna()]
house_rent_dataset = house_rent_dataset.dropna()
```

Python

```python
house_rent_dataset = house_rent_dataset.reset_index(names='RentID')
house_rent_dataset['RentID'] = house_rent_dataset['RentID'] + 1
house_rent_dataset.head()
```

Python

	RentID	Posted On	Number of Bedrooms-Halls-Kitchens	Recent Floor	Total Floors	Rent	Size	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	1	2022-05-18	2	0	2	10000	1100	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2	2022-05-13	2	1	3	20000	800	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	3	2022-05-16	2	1	3	17000	1000	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	4	2022-07-04	2	1	2	10000	800	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	5	2022-05-09	2	1	2	7500	850	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner

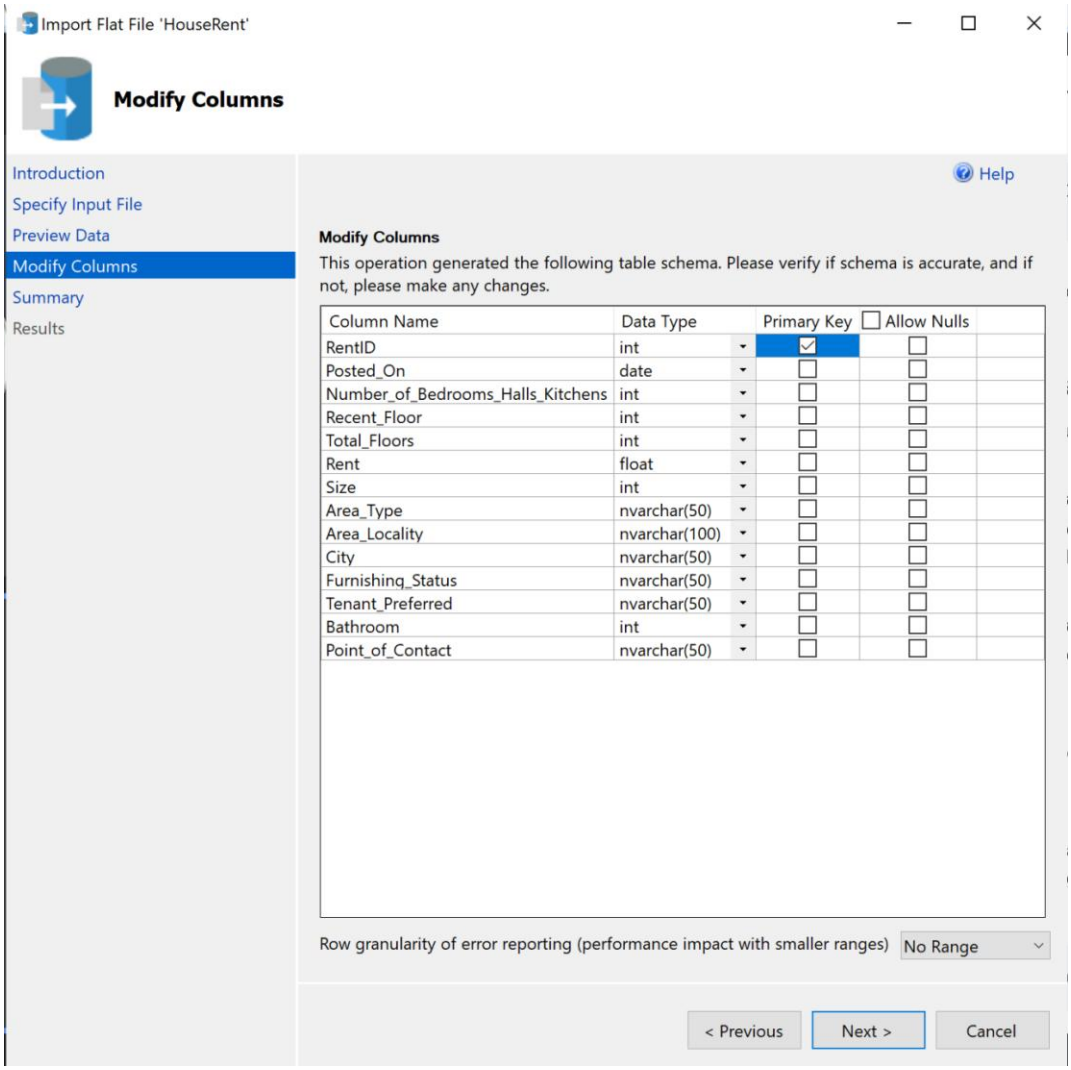
- Xuất DataFrame sau khi đã xử lý xong ra file csv

```python
house_rent_dataset.to_csv('Preprocessed_House_Rent.csv', index=False)
```

DataPreProcess_House_Rent.ipynb	5/17/2023 2:56 AM	Jupyter Source File	42 KB
House_Rent_Dataset.csv	5/17/2023 1:19 AM	XLS Worksheet	554 KB
Preprocessed_House_Rent.csv	5/17/2023 2:56 AM	XLS Worksheet	539 KB

2.2. Tải dữ liệu vào SQL Server 2019

- Import Flat File vào Database HouseRent
- Thiết lập lại kiểu dữ liệu cho các cột và set cột **RentID** là Primary Key


Import Flat File 'HouseRent'
Modify Columns
Introduction
Specify Input File
Preview Data
Modify Columns
Summary
Results
Help
Modify Columns
This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.
Column Name
Data Type
Primary Key
Allow Nulls
RentID int
Posted_On date
Number_of_Bedrooms_Halls_Kitchens int
Recent_Floor int
Total_Floors int
Rent float
Size int
Area_Type nvarchar(50)
Area_Locality nvarchar(100)
City nvarchar(50)
Furnishing_Status nvarchar(50)
Tenant_Preferred nvarchar(50)
Bathroom int
Point_of_Contact nvarchar(50)
Row granularity of error reporting (performance impact with smaller ranges)
No Range
< Previous
Next >
Cancel

- Xem dữ liệu

SQLQuery3.sql - DE...IB57SI\hoang (63)) | SQLQuery2.sql - DE...IB57SI\hoang (54))*

```sql
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [RentID]
      ,[Posted_On]
      ,[Number_of_Bedrooms_Halls_Kitchens]
      ,[Recent_Floor]
      ,[Rent]
      ,[Size]
      ,[Area_Type]
      ,[Area_Locality]
      ,[City]
      ,[Furnishing_Status]
      ,[Tenant_Preferred]
      ,[Bathroom]
      ,[Point_of_Contact]
  FROM [HouseRent].[dbo].[Preprocessed_House_Rent]
```

127 %

Results | Messages

	RentID	Posted_On	Number_of_Bedrooms_Halls_Kitchens	Recent_Floor	Rent	Size	Area_Type	Area_Locality	City	Furnishing_Status	Tenant_Preferred	Bathroom	Point_of_Contact
1	1	2022-05-18	2	0	10000	1100	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
2	2	2022-05-13	2	1	20000	800	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	3	2022-05-16	2	1	17000	1000	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
4	4	2022-07-04	2	1	10000	800	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
5	5	2022-05-09	2	1	7500	850	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner
6	6	2022-04-29	2	0	7000	600	Super Area	Thakurpukur	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
7	7	2022-06-21	2	0	10000	700	Super Area	Malancha	Kolkata	Unfurnished	Bachelors	2	Contact Agent
8	8	2022-06-21	1	1	5000	250	Super Area	Malancha	Kolkata	Unfurnished	Bachelors	1	Contact Agent
9	9	2022-06-07	2	1	26000	800	Carpet Area	Palm Avenue Kolkata, Ballygunge	Kolkata	Unfurnished	Bachelors	2	Contact Agent
10	10	2022-06-20	2	1	10000	1000	Carpet Area	Natunhat	Kolkata	Semi-Furnished	Bachelors/Family	2	Contact Owner
11	11	2022-05-23	3	1	25000	1200	Carpet Area	Action Area 1, Rajarhat Newtown	Kolkata	Semi-Furnished	Bachelors/Family	2	Contact Agent
12	12	2022-06-07	1	1	5000	400	Carpet Area	Keshtopur	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
13	13	2022-05-14	1	1	6500	250	Carpet Area	Tarulia, Keshtopur	Kolkata	Furnished	Bachelors	1	Contact Owner
14	14	2022-05-09	1	1	5500	375	Carpet Area	Dum Dum Metro	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
15	15	2022-05-05	3	0	8500	900	Carpet Area	Paschim Barisha	Kolkata	Unfurnished	Bachelors	2	Contact Owner
16	16	2022-06-01	3	1	40000	1286	Carpet Area	New Town Action Area 1	Kolkata	Furnished	Bachelors/Family	2	Contact Owner
17	17	2022-05-17	2	1	6000	600	Super Area	Barasat	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
18	18	2022-06-20	2	0	10000	800	Super Area	Behala	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
19	19	2022-06-09	2	0	11000	2000	Carpet Area	Behala Chowrasta	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
20	20	2022-06-09	2	1	6000	660	Super Area	Behala	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
21	21	2022-07-02	2	1	7900	650	Carpet Area	Santoshpur	Kolkata	Unfurnished	Family	1	Contact Owner

CHƯƠNG 3: TRỰC QUAN HÓA DỮ LIỆU

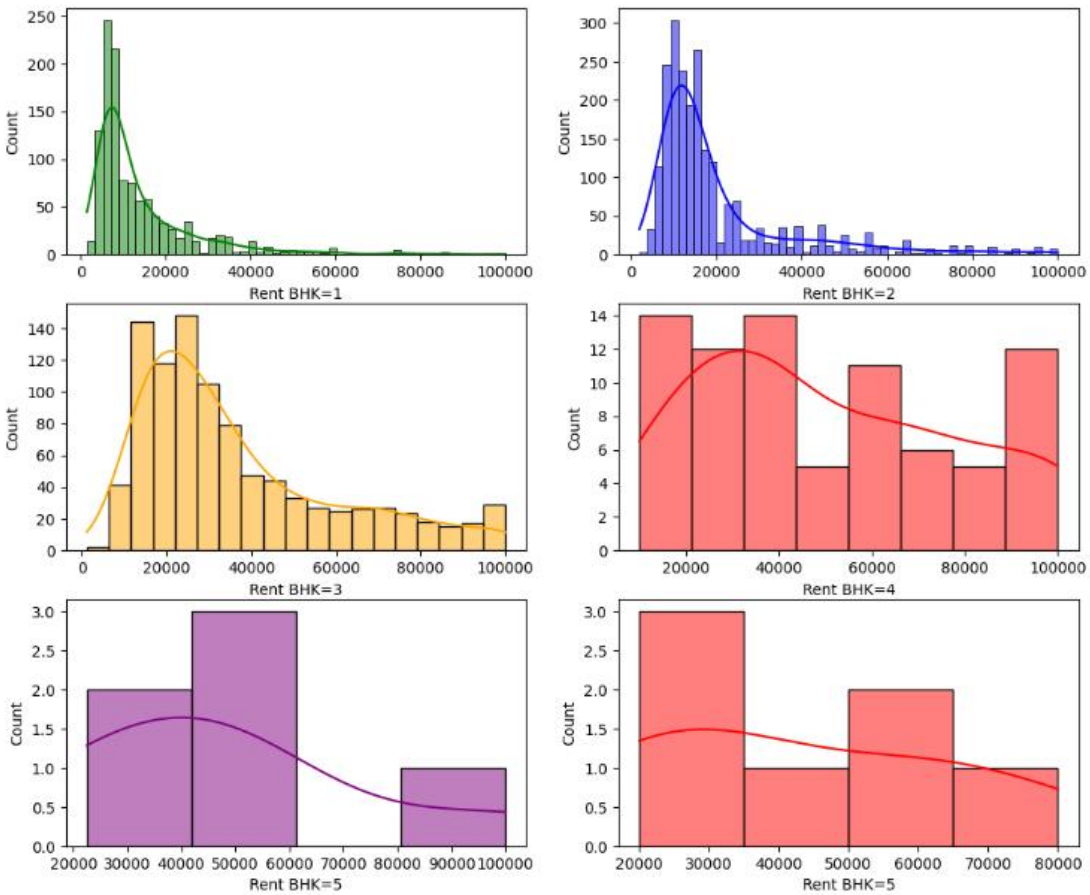
3.1. Trực quan hóa dữ liệu trên công cụ Python

3.1.1. Quy mô của các Houses/Apartments/Flats cho thuê có ảnh hưởng đến giá thuê không?

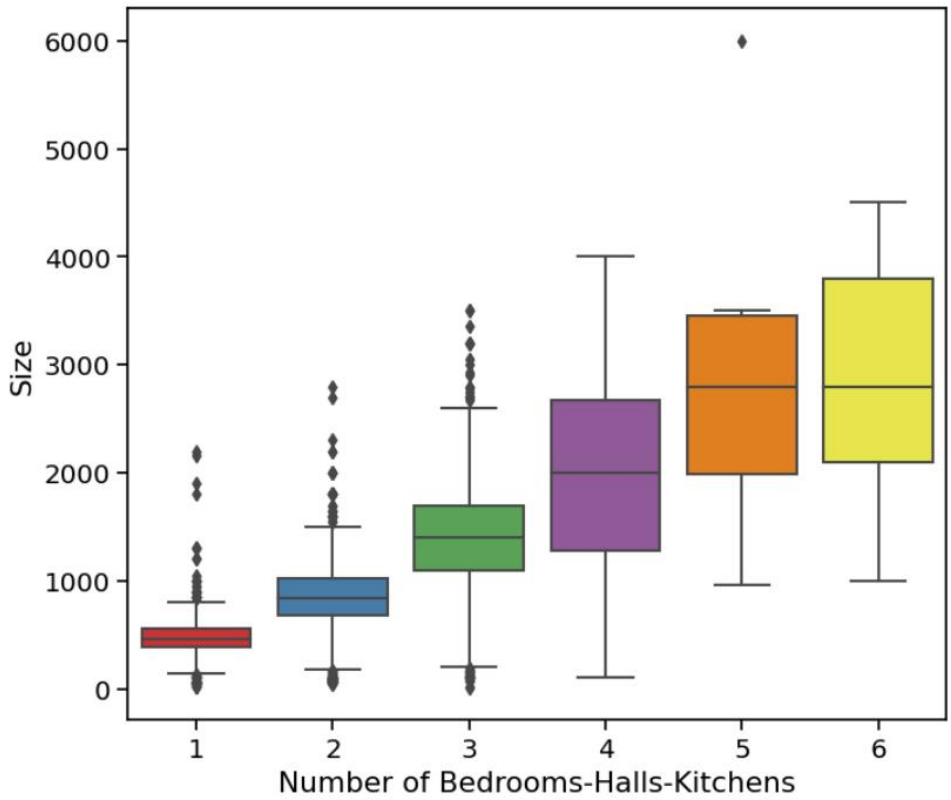
- Số lượng phòng cho thuê trong các Houses/Apartments/Flats

```
# Số lượng 'Number of Bedrooms-Halls-Kitchens'
fig, ax = plt.subplots()
sns.countplot(data=df, x='Number of Bedrooms-Halls-Kitchens', ax=ax)
for p in ax.patches:
    ax.annotate(int(p.get_height()), (p.get_x()+0.2, p.get_height()+5))
plt.show()
```

- Hầu hết các Houses/Apartments/Flats cho thuê có số lượng phòng tắm - sảnh - phòng ăn = 2 là phổ biến nhất, khá ít nơi cho thuê có quy mô lớn hơn 5 phòng trở lên.
- Mối liên quan giữa Number of Bedroom-Halls-Kitchens với giá cho thuê Rent

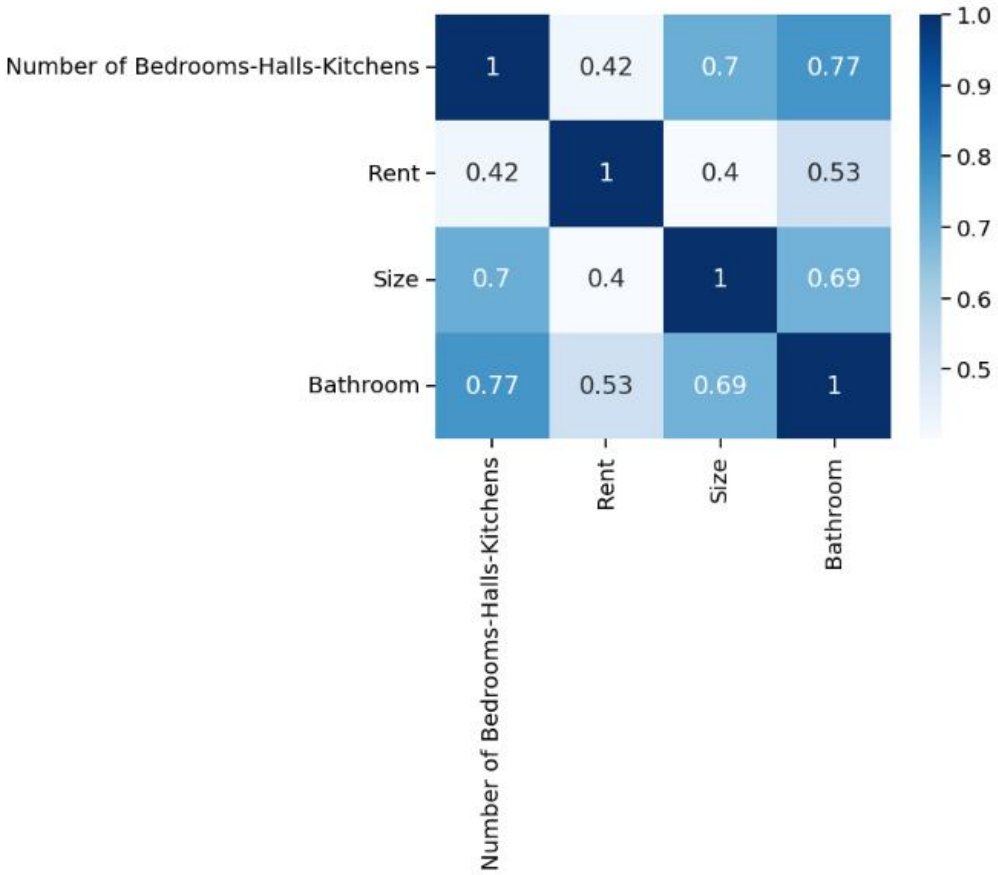


- Nhìn vào hình bên trên ta thấy với Số lượng Bedroom-Halls-Kitchens cho thuê bằng 1 và 2 thì mức giá cho thuê sẽ dao động nhiều hơn ở những khoảng thấp hơn 40000. Với số lượng phòng Bedroom-Halls-Kitchens cho thuê từ 3 đến 4 phòng thì sẽ có mức giá dao động ổn định hơn và những nơi cho thuê có quy mô từ 5 phòng thì chắc chắn giá thuê sẽ từ 20000 trở lên.



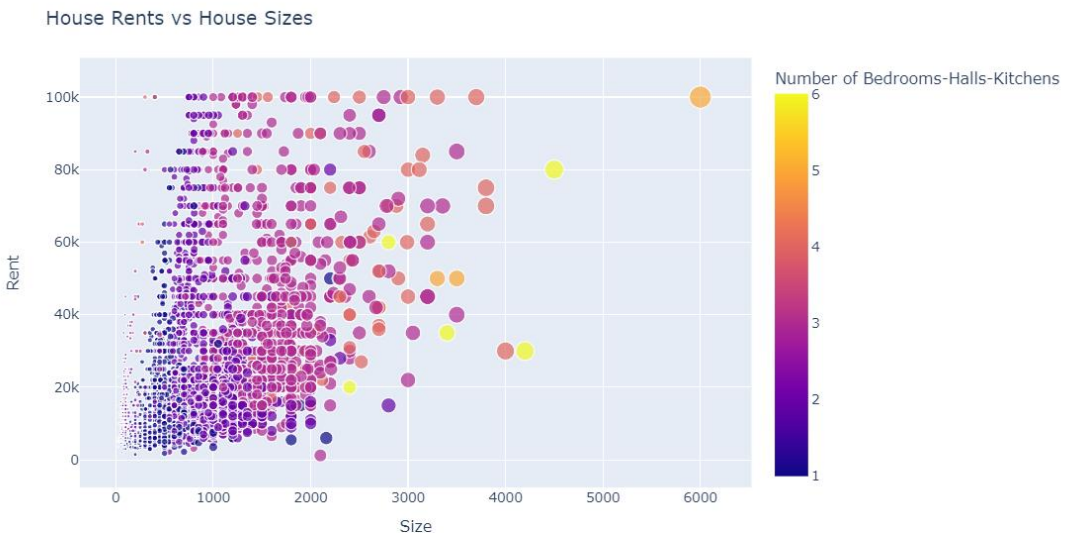
- Nhìn chung thì các Houses/Apartments/Flats cho thuê có số lượng Bedrooms - Halls - Kitchen nhiều thì diện tích cũng sẽ lớn theo.
- Giá thuê trung bình theo số lượng phòng tắm của các nơi cho thuê

- Những nơi cho thuê có 4 đến 6 phòng tắm sẽ có giá thuê trung bình cao nhất hơn 50000.



- Biểu đồ nhiệt thể hiện sự tương quan giữa các biến Number of Bedrooms - Halls - Kitchens, Size, Bathroom và Rent. Ta có thể thấy có sự tương quan mạnh giữa ba biến Number of Bedrooms - Halls - Kitchens, Bathroom và Size với nhau. Cho thấy chúng có tỉ lệ thuận đồng nghĩa là quy mô Houses/Apartments/Flats lớn thì số lượng phòng Bedrooms - Halls - Kitchens nói chung hay số lượng phòng Bathroom và diện tích chung cũng sẽ lớn theo.
- Biểu đồ tương quan giữa Diện tích, Số lượng Bedrooms - Halls - Kitchens cho thuê với giá thuê.

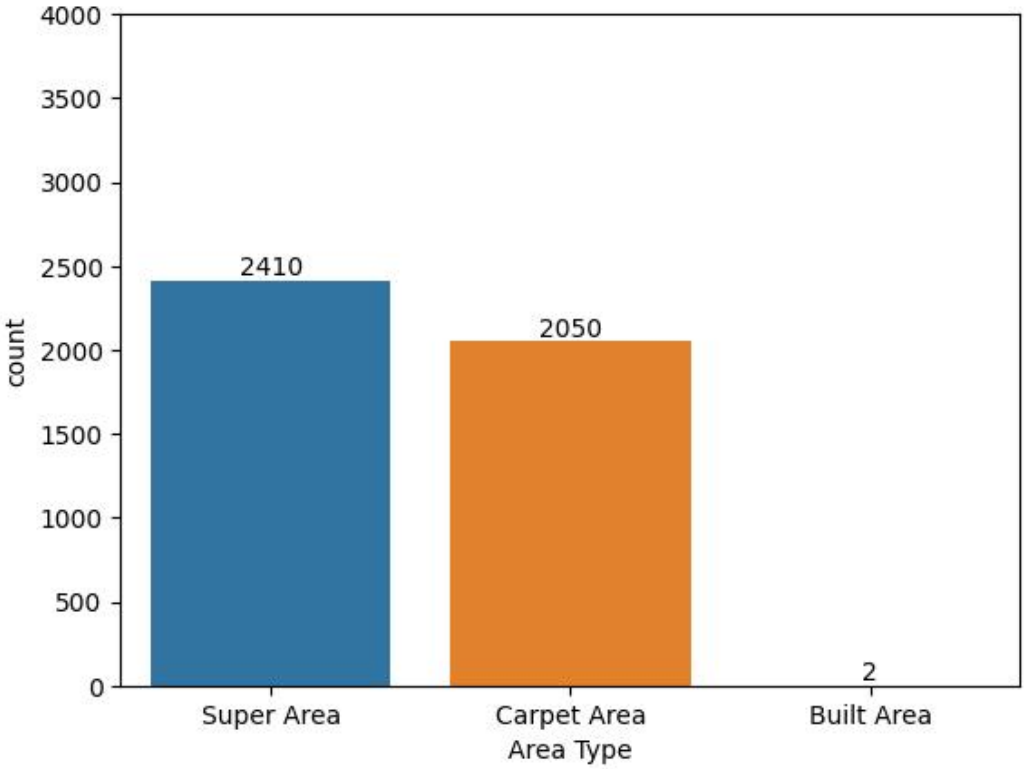
```
fig = px.scatter(df, x='Size', y='Rent', color='Number of Bedrooms-Halls-Kitchens', size='Size', hover_data=['Rent'])
fig.update_layout(title='House Rents vs House Sizes',
                  yaxis_zeroline=False, xaxis_zeroline=False)
fig.show()
```



- Khi kết hợp các yếu tố lại với nhau ta thấy nếu diện tích cho thuê lớn với nhiều phòng cho thuê thì chắc chắn giá cho thuê sẽ mắc. Tuy nhiên với những nơi cho thuê có diện tích nhỏ ít số phòng thì không thể khẳng định chúng có giá cho thuê rẻ.

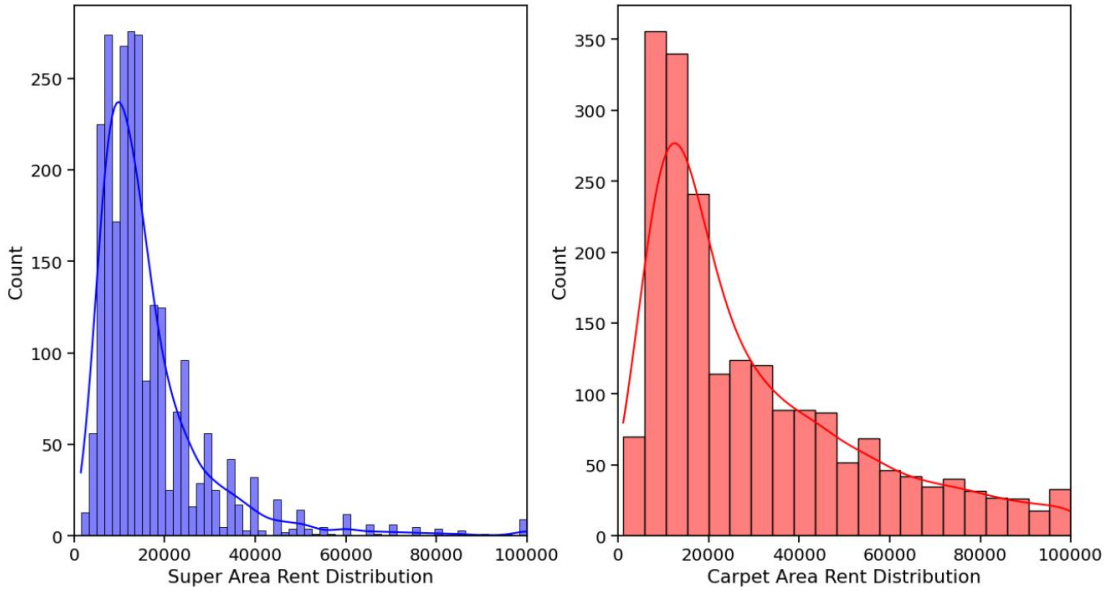
3.1.2. Loại khu vực của các Houses/Apartments/Flats cho thuê có liên quan đến giá cho thuê hay không?

- Phân loại các loại khu vực tại các Houses/Apartments/Flats cho thuê



- Ta thấy các nơi cho thuê được chia ra 2 khu vực chính là Super Area và Carpet Area.
- Trung bình mức giá thuê theo khu vực

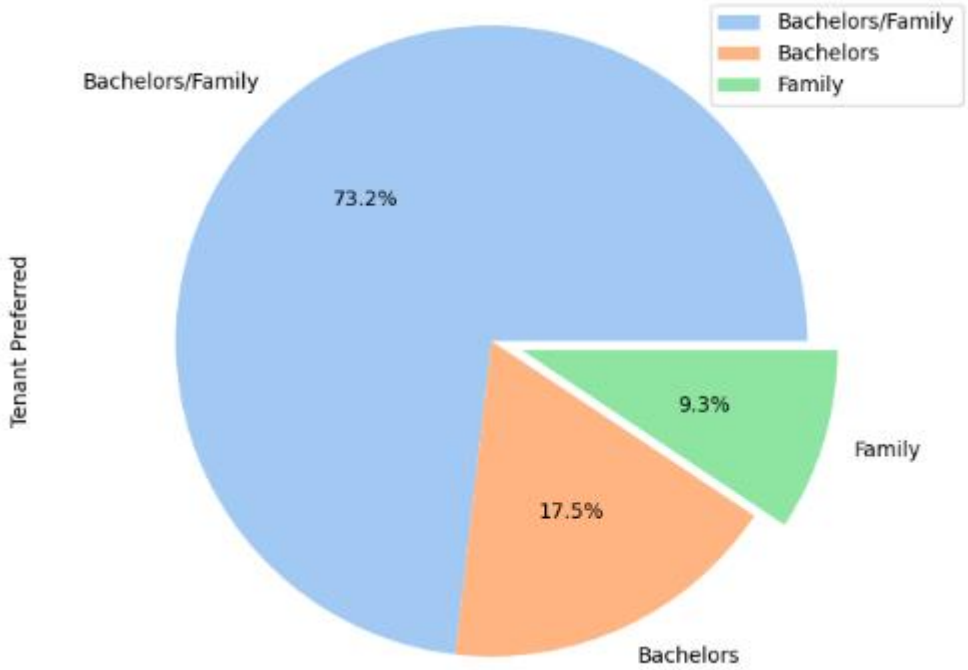
- Có vẻ như Area Type là Carpet thì mức giá cho thuê sẽ mắc hơn.
- Sự phân bố giá thuê giữa khu Super và Carpet



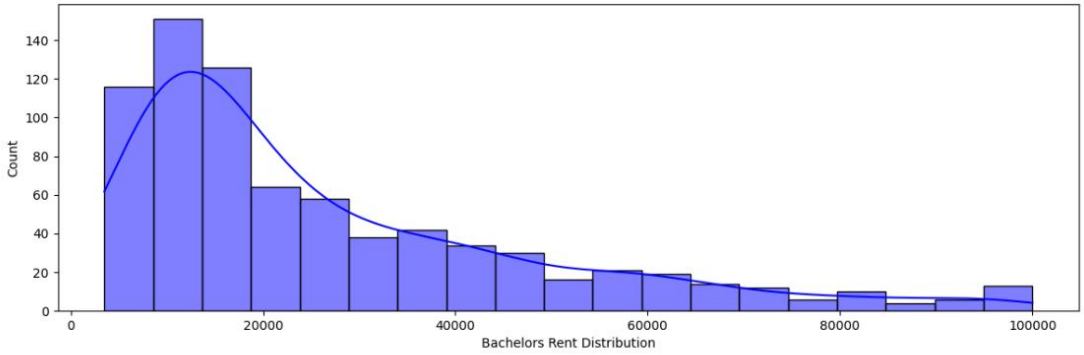
- Hầu hết các hầu các chỗ thuê thuộc khu vực Super có mức giá từ trung bình đến thấp. Còn khu vực **Carpet** sở hữu nhiều nơi cho thuê có giá **khá cao** so với khu vực **Super. Carpet** sở hữu nhiều những căn hộ/ phòng/ nhà cho thuê khá đắt đỏ.

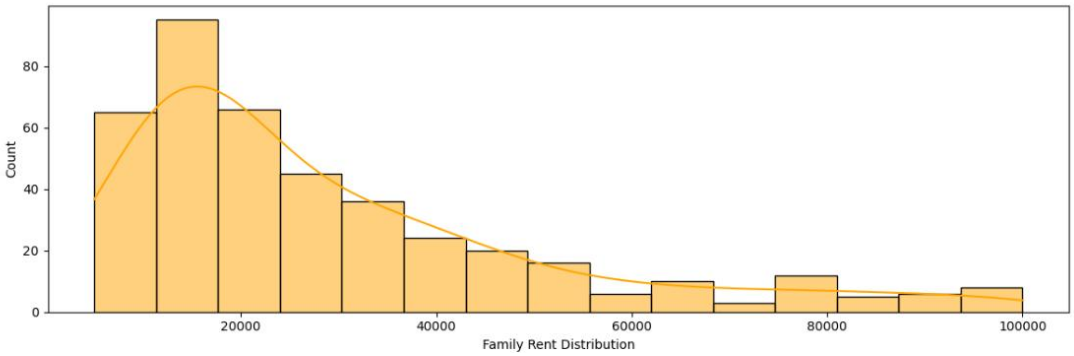
3.1.3. Những loại người thuê nhà được ưa thích?

- Các loại đối tượng ưu tiên cho thuê nhà



- **72,6%** chủ sở hữu nhà sẵn sàng cho cả **gia đình** và **người độc thân** thuê nhà của họ.
- Phân bố giá thuê nhà giữa đối tượng ‘Family’ và ‘Bachelors’

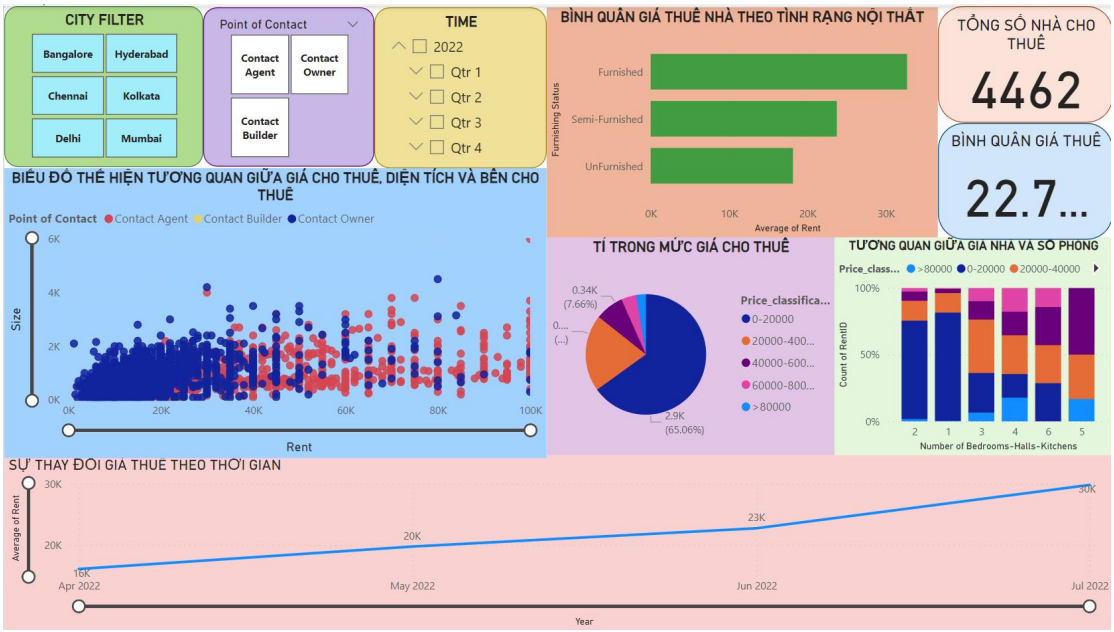




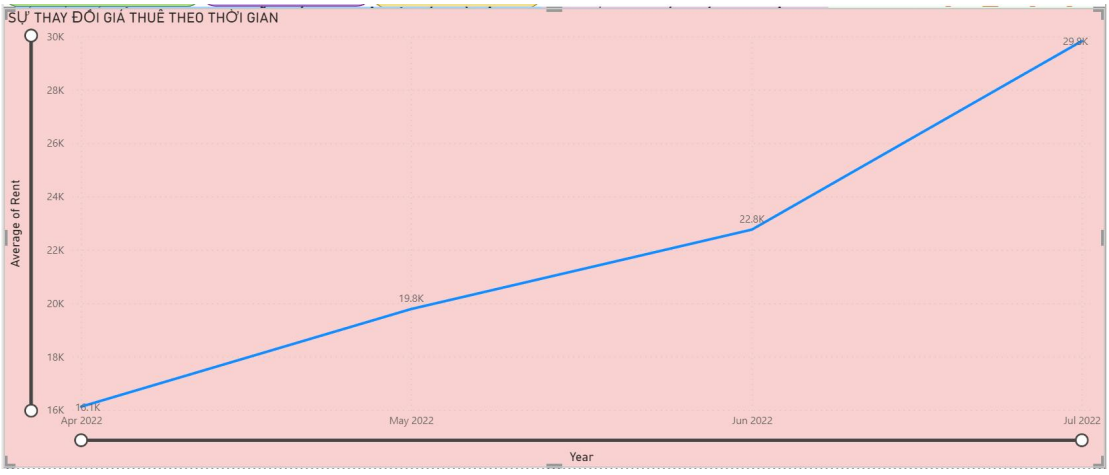
- Ta thấy không có nhiều sự khác biệt về giá thuê giữa hai đối tượng ‘Bachelors’ và ‘Family’.

3.2. Trực quan hóa dữ liệu trên công cụ Power BI

3.2.1. Dashboard

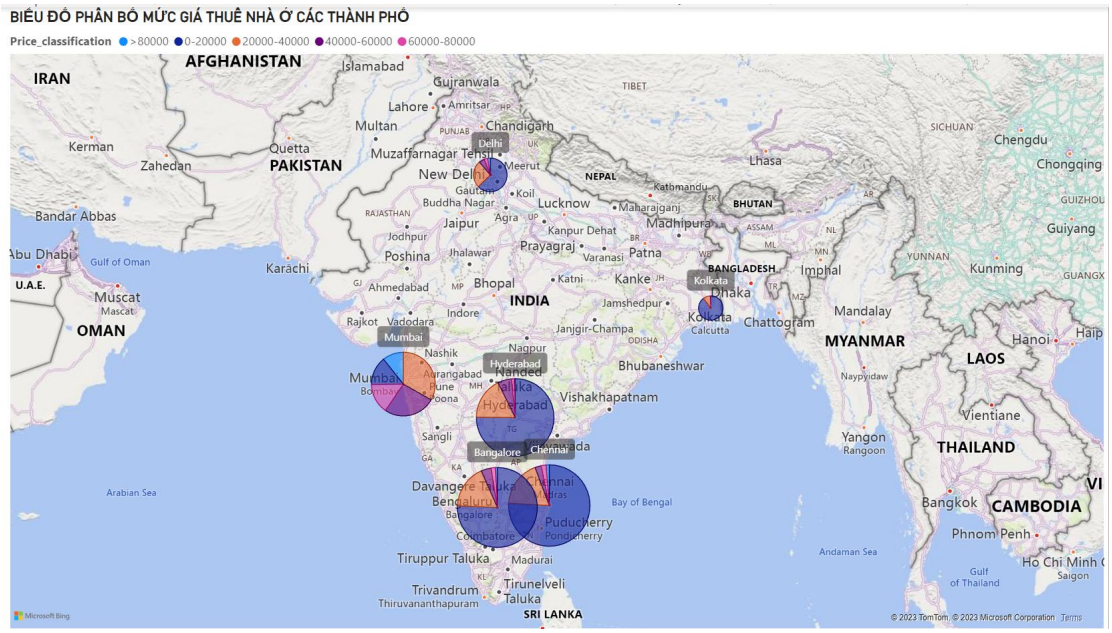


- Bình quân giá thuê nhà luôn tăng dần từ tháng 4 cho đến tháng 7.



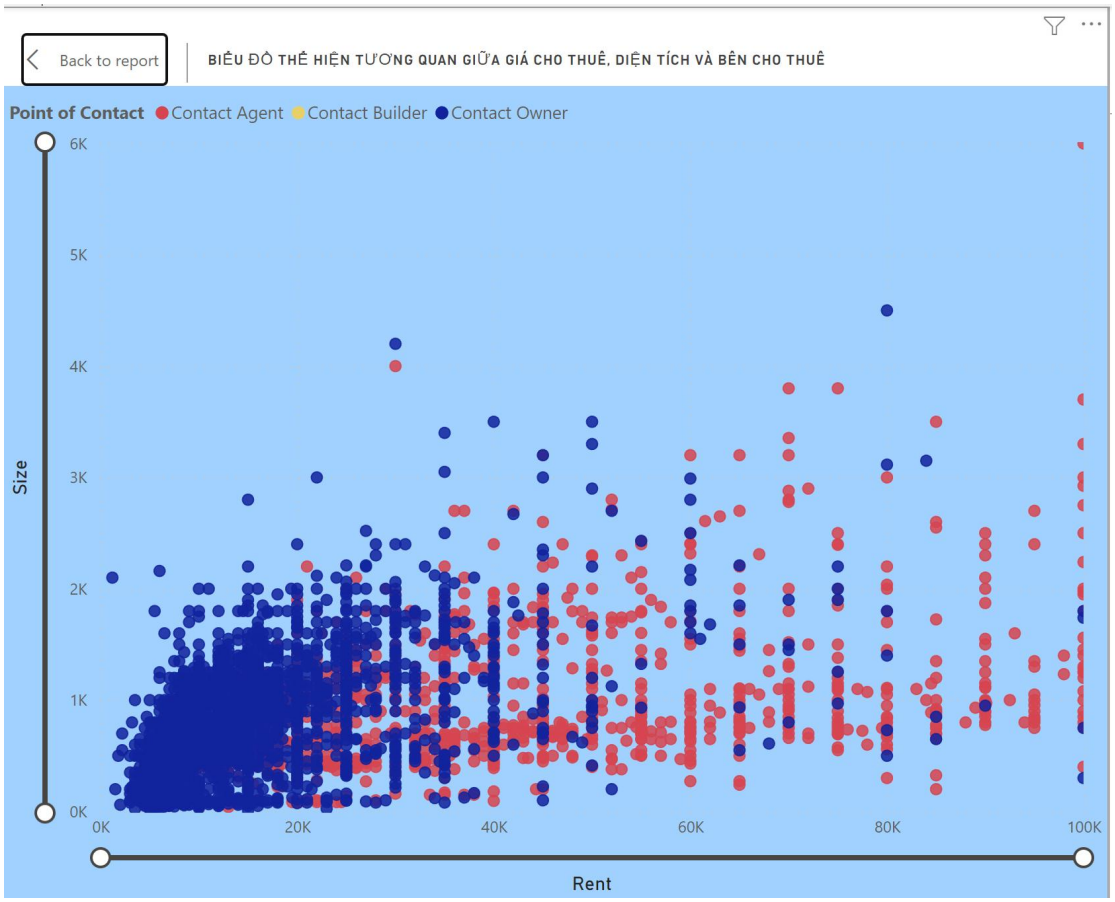
3.2.2. Đánh giá giá thuê nhà giữa các thành phố

- Số liệu tổng hợp giá nhà cho thuê ở 6 thành phố. Đa số các ngôi nhà cho thuê đều trong tầm giá dưới 20000.
- Giá nhà ở thành phố Mumbai tương đối cao so với các thành phố khác, tỉ trong các căn nhà có mức giá cho thuê từ 60000-80000 và >80000 ở thành phố này cao hơn ở các thành phố khác.

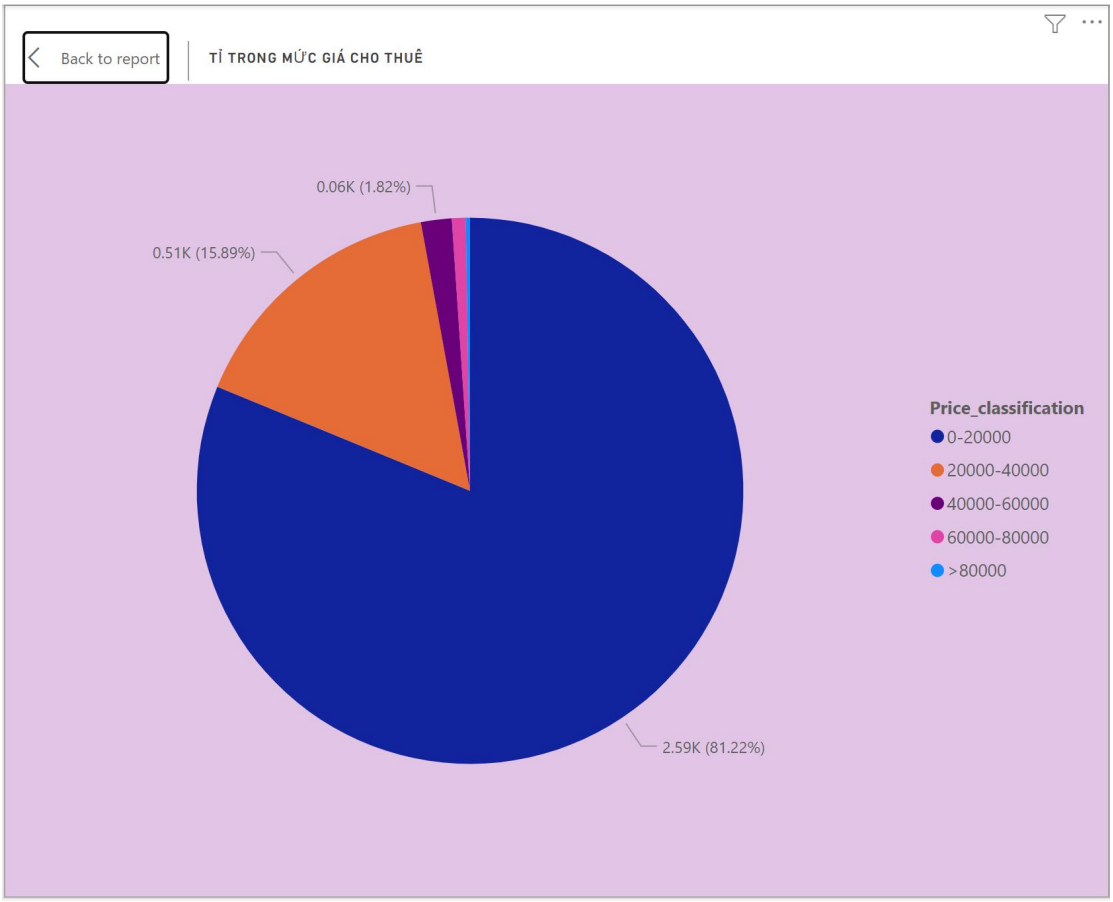


3.2.3. Tương quan giữa các bên cho thuê và giá thuê nhà

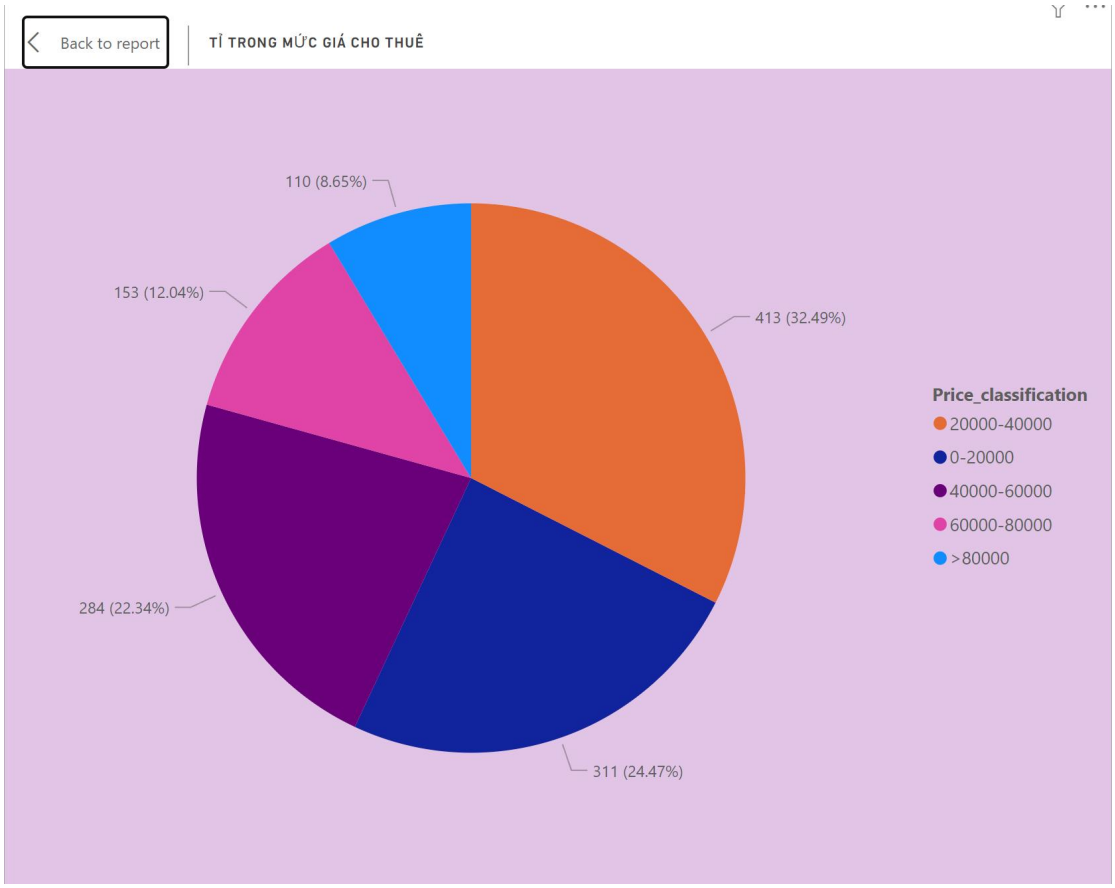
- Với cùng một diện tích, những căn nhà được thuê trực tiếp từ chủ nhà sẽ rẻ hơn so với nhà được cho thuê bởi công ty cho thuê.



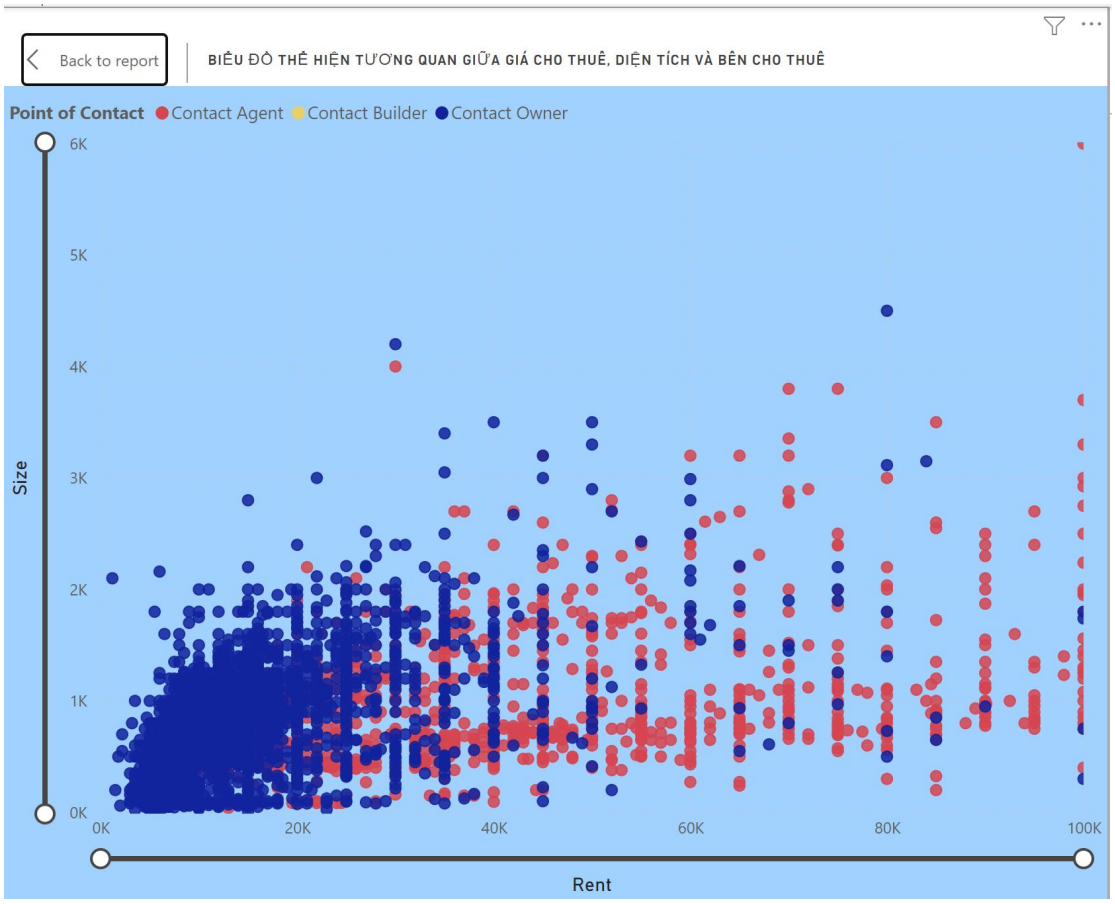
+ Giá nhà được cho thuê trực tiếp từ chủ nhà:



+ Giá nhà được cho thuê trừ công ty cho thuê:

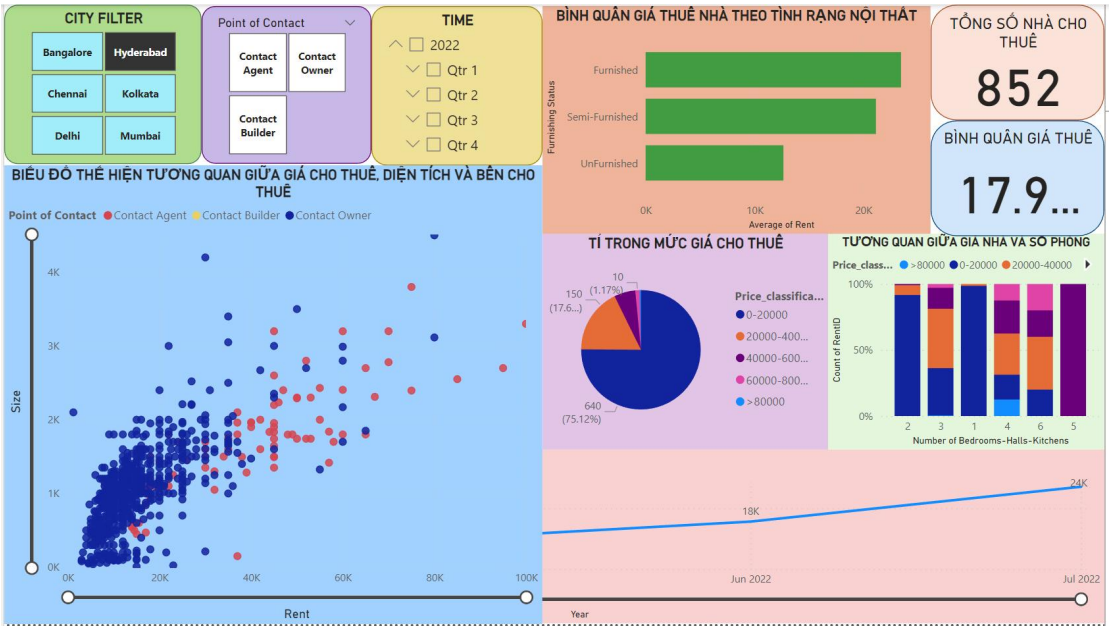


3.2.4. Tương quan giữa diện tích và giá thuê nhà

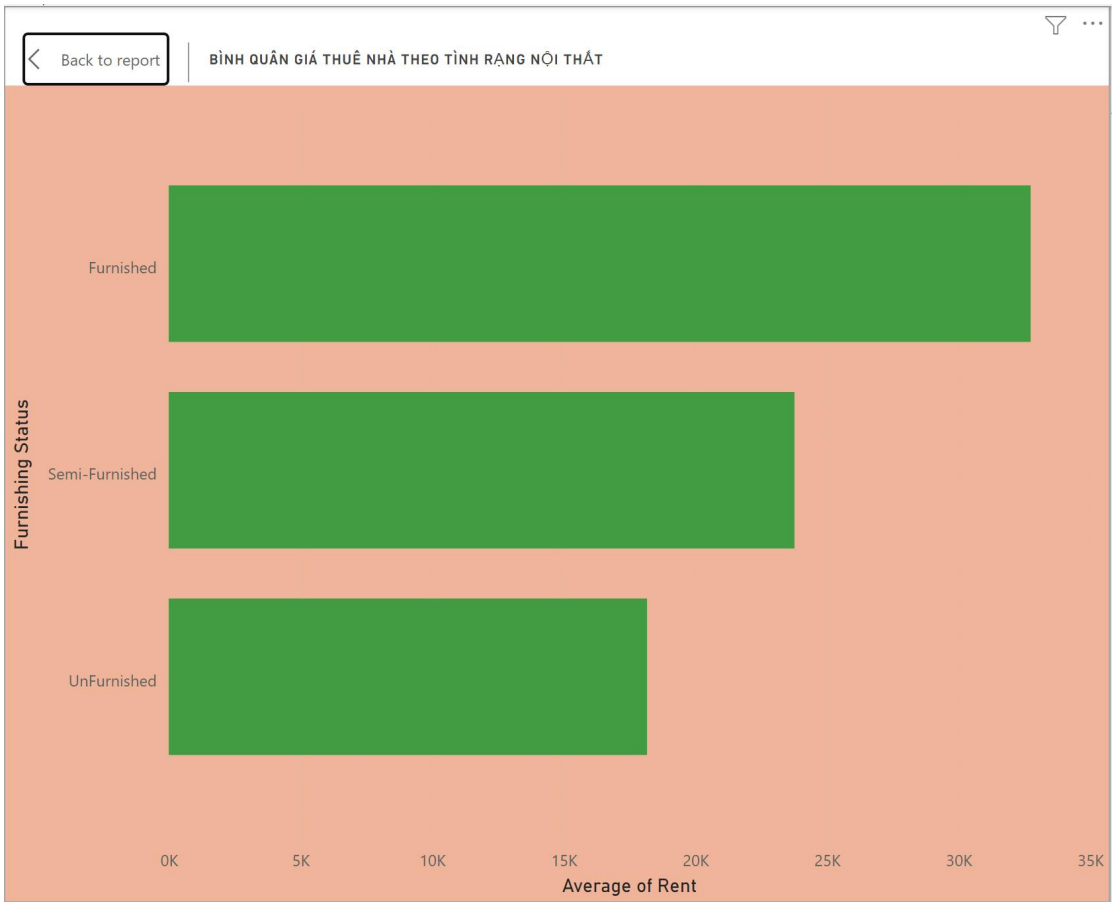


- Từ biểu đồ phân tán phía trên, ta thấy diện tích nhà không ảnh hưởng đến giá cho thuê một các đơn lẻ, nó cần phải kết hợp với những yếu tố khác như thành phố, số phòng,...

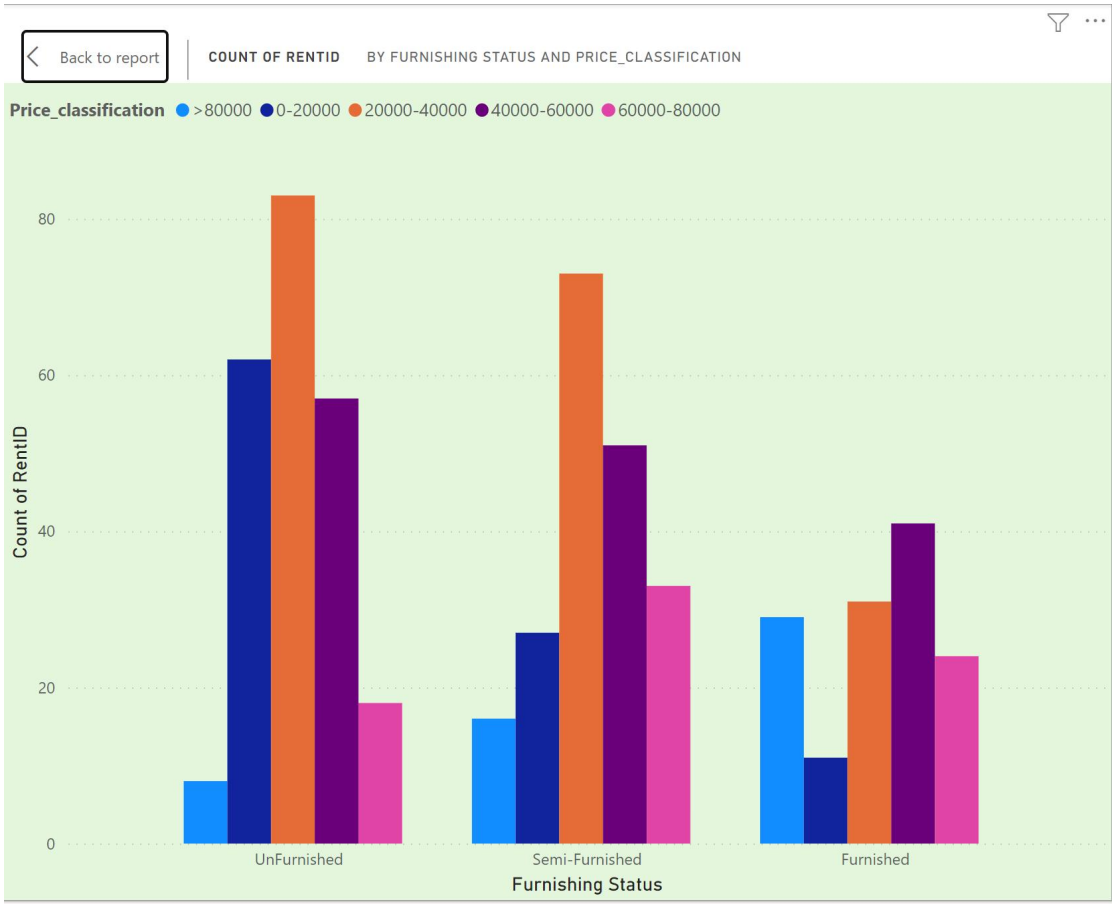
- Nhìn vào biểu đồ phân tán giá nhà theo size và bên cho thuê ở thành phố Hyderabad, ta thấy nhà có diện tích càng lớn thì giá thuê càng cao.



3.2.5. Tương quan giữa tình trạng nội thất và giá thuê nhà

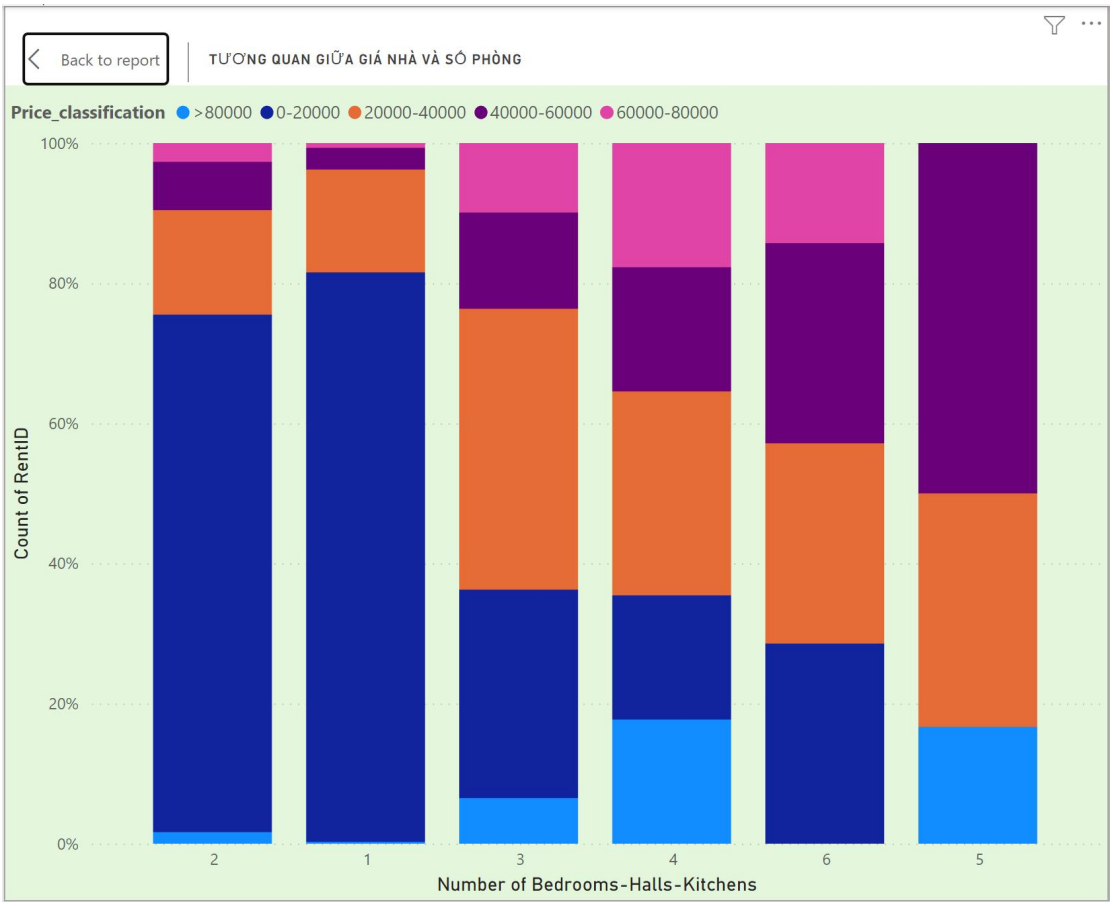


- Những căn hộ có trang bị nội thất đầy đủ sẽ có bình quân giá thuê cao nhất và thấp nhất là những căn hộ không trang bị nội thất.



- Ở Mumbai, nhà đã được trang bị nội thất đầy đủ ít khi có mức giá dưới 20000. Nhà được trang bị nội thất nhưng chưa đầy đủ chủ yếu có mức giá từ 20000-40000 và nhà chưa được trang bị nội thất phần lớn có mức giá từ 0-40000.
- Nhìn chung giá nhà được trang bị nội thất đầy đủ sẽ cao hơn hai loại còn lại.

3.2.6. Tương quan giữa số phòng của căn nhà và giá thuê nhà

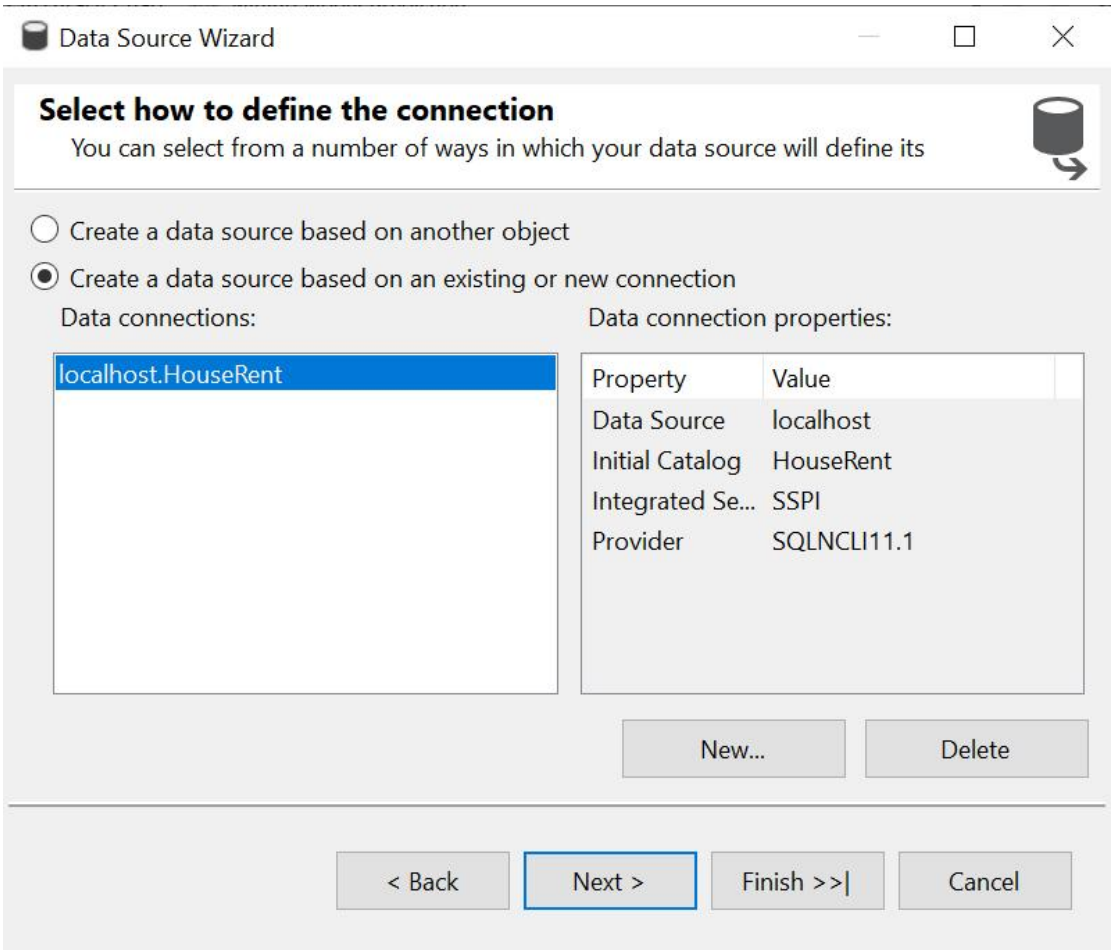


- Các căn hộ có 1-2 phòng thường có giá trong khoảng dưới 40000.
- Những căn hộ có giá thuê trong khoảng từ 40000 trở lên thường là những căn hộ có 3 phòng trở lên.
- Nhìn chung, số phòng càng nhiều thì mức giá cho thuê càng cao.

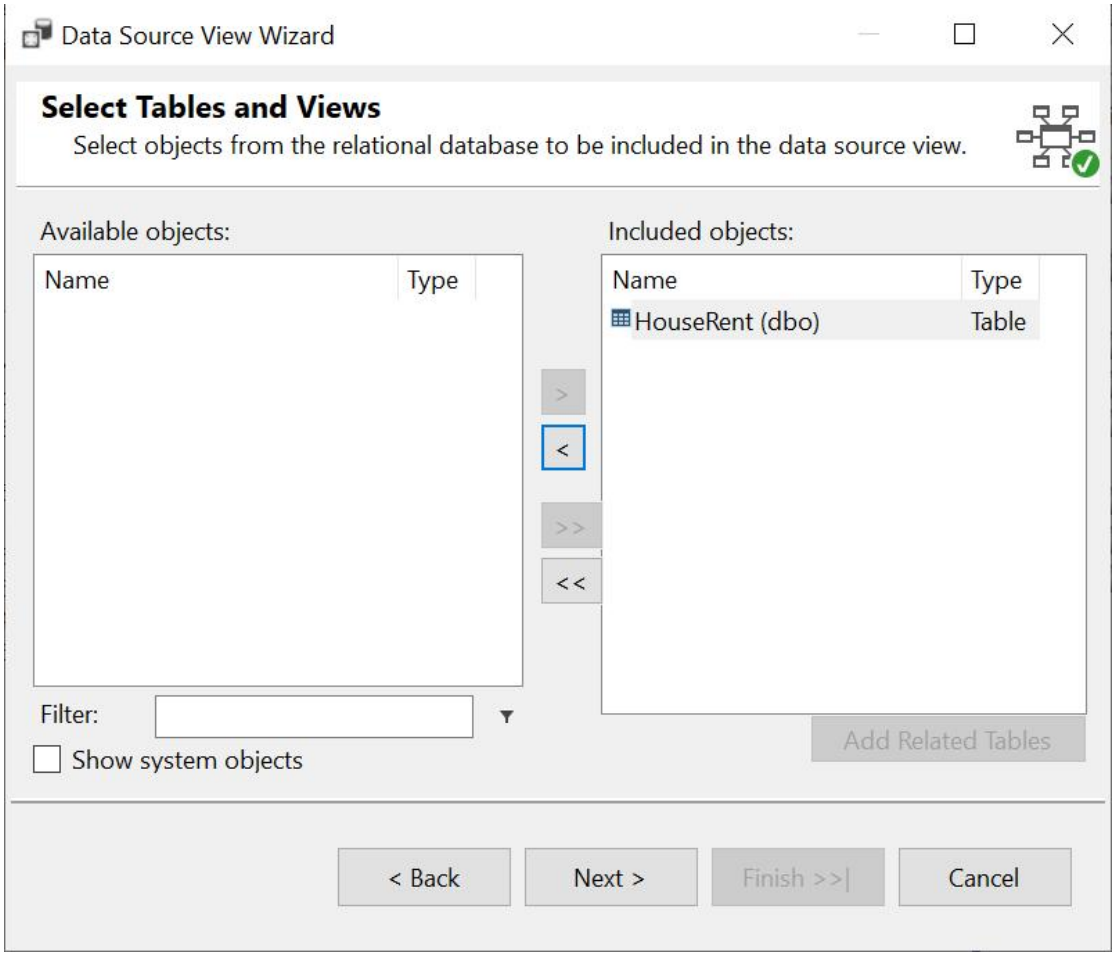
CHƯƠNG 4: KHAI PHÁ DỮ LIỆU BẰNG SSAS

4.1. Quá trình tạo mô hình

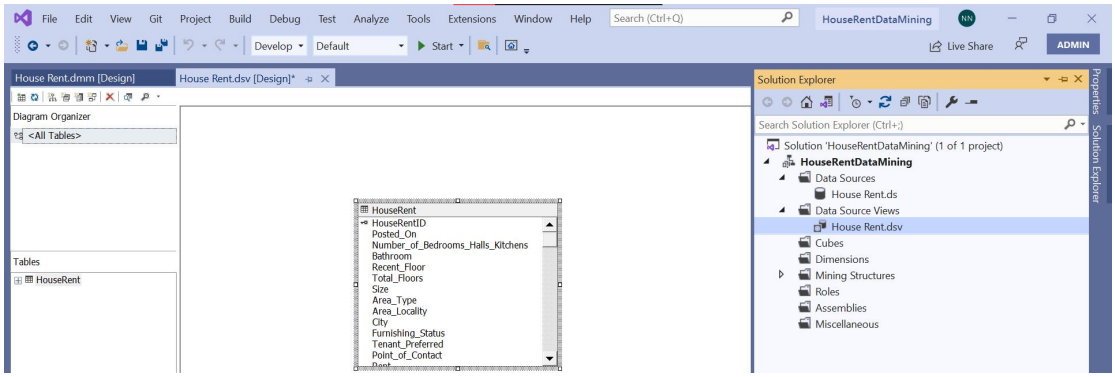
- Mở Visual Studio 2019 để khởi tạo project **Analysis Services Multidimensional and Data Mining Project**
- Khởi tạo DataSource từ database **HouseRent** và đặt tên cho Data Source là **HouseRent**



- Khởi tạo Data Source View và load table **HouseRent** vào Data Source View



- Xuất hiện table House Rent trong Data Source View



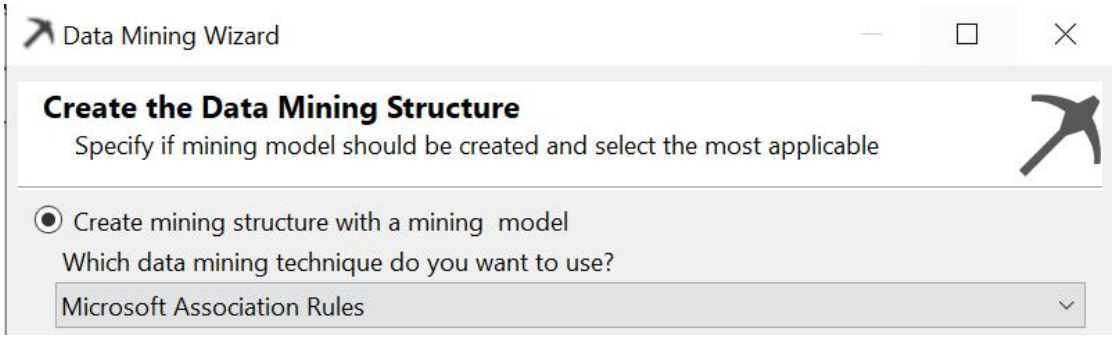
4.2. Association Rule Mining

4.2.1. Giới Thiệu

- Association Rule: là quá trình tìm ra các mối quan hệ (tương quan) giữa các đối tượng trong tập dữ liệu. Có dạng X => Y với:
  - X: antecedent (tiền đề)
  - Y: consequent (hệ quả)
- Hai tham số quan trọng được dùng để đo lường luật kết hợp là:
  - Support (độ hỗ trợ): Xác suất mà X và Y đồng thời cùng xuất hiện.
  - Confidence (độ tin cậy): Xác suất xảy ra Y khi có X
- Ngoài hai tham số chính là Minimum probability và Minimum support thì thuật toán Microsoft Association Rule còn cung cấp thêm một tham số là **Importance** (độ quan trọng) - xác định sự tương quan giữa sự kiện X và sự kiện Y. Nếu
  - Importance > 0 (dương) cho thấy khi X xảy ra thì khả năng xuất hiện Y cũng sẽ **tăng** lên. Importance có giá trị dương lớn cho thấy X có ảnh hưởng **tích cực** đến sự xuất hiện của Y.
  - Importance < 0 (âm) cho thấy khi X xảy ra thì khả năng xuất Y sẽ **giảm** đi. Importance có giá trị âm lớn cho thấy X có ảnh hưởng **tiêu cực** đến sự xảy ra của Y.
  - Importance = 0 thì điều này cho thấy **không có mối liên hệ giữa X và Y**.
- Đối với tập dữ liệu House Rent - Giá thuê nhà ở Ấn Độ, thì sử dụng khai phá luật Kết hợp để tìm ra các mối quan hệ giữa các thuộc tính, các quy luật kết hợp ảnh hưởng đến giá cho thuê (Rent), ví dụ như nếu nơi cho thuê thuộc thành phố lớn và diện tích rộng thì giá thuê sẽ mắc.
- Khi tìm ra các quy luật kết hợp, ta có thể sử dụng chúng để phân tích và đưa ra dự đoán giá thuê nhà dựa trên các mối quan hệ phụ thuộc đó.

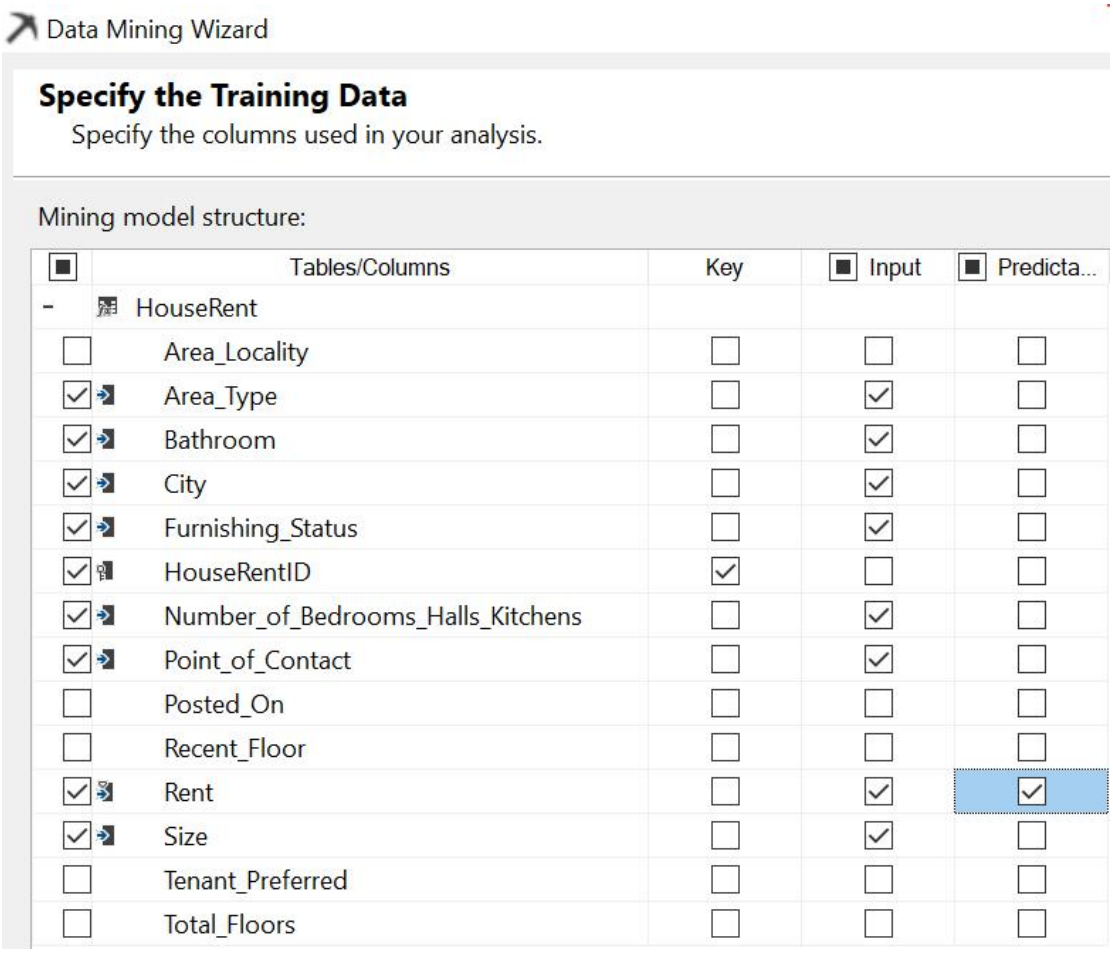
4.2.2. Tiến Hành Khai Phá Luật Kết Hợp

- Sau khi đã khởi tạo Data Source và Data Source View ta tiến hành khởi tạo **Mining Structure** và chọn thuật toán **Microsoft Association Rules** để khai phá luật kết hợp.

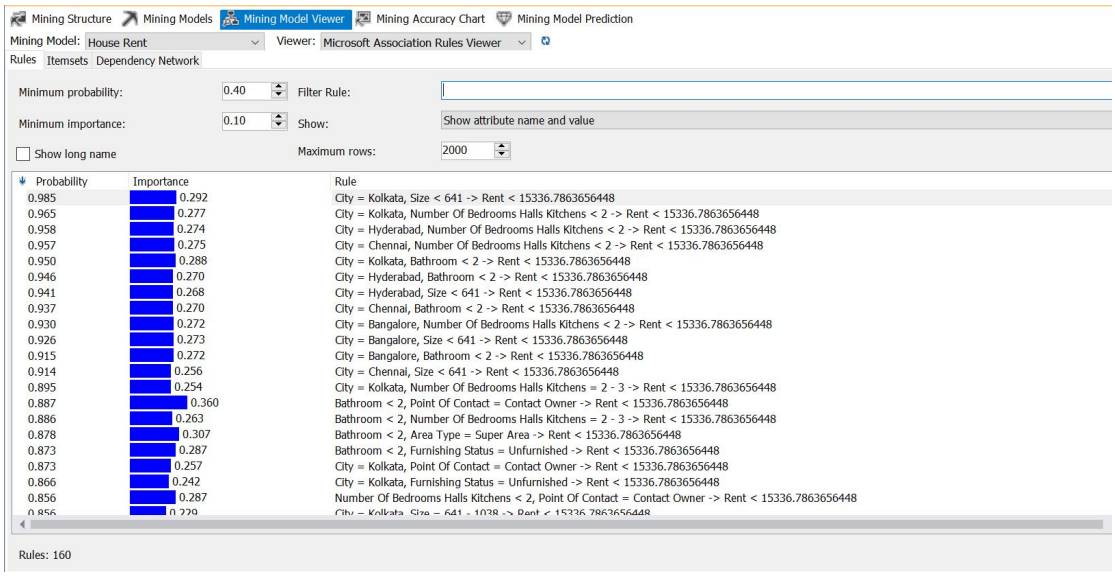


- Sau khi trực quan hóa dữ liệu thể hiện tương quan giữa các biến thì ta sẽ chọn ra các thuộc tính có ảnh hưởng đến giá thuê nhà là: *Area_Type, Bathroom, City, Furnishing_Status, Number of Bedrooms_Halls_Kitchens, Point _of_Contact, Size,*

và Rent. Đồng thời chọn **HouseRentID** làm khóa chính (Key), và chọn thuộc tính **Rent** làm biến dự đoán (Predictable). Mô tả như hình dưới đây.

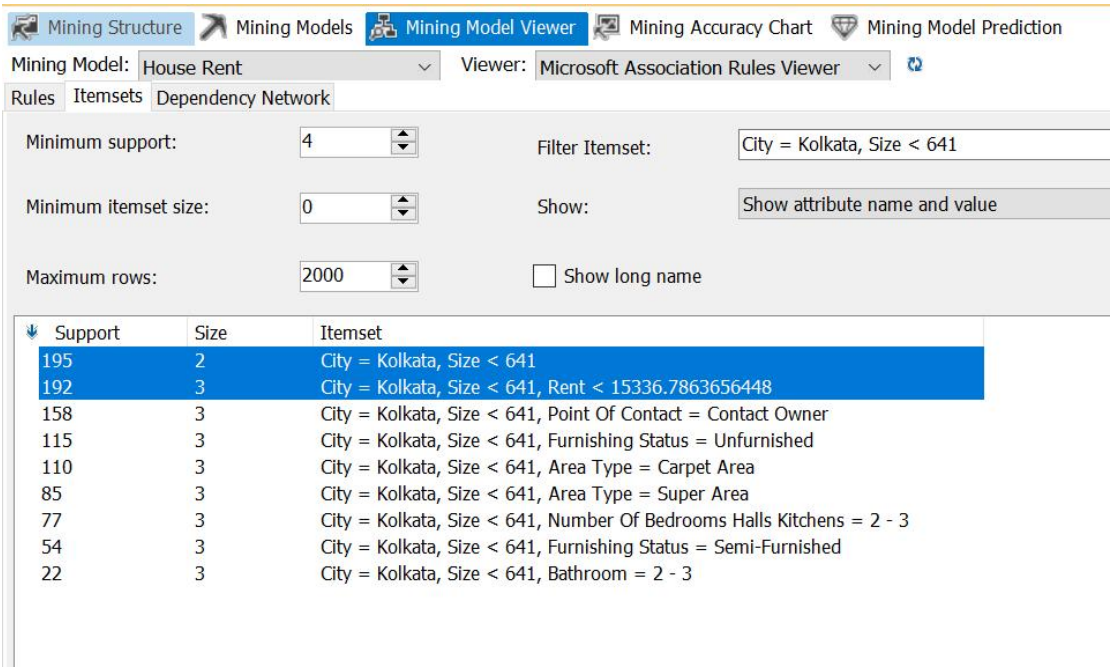


- Tiến hành deploy project và process để xem kết quả.
- Chọn tab **Mining Model Viewer -> Rules** để xem kết quả các luật kết hợp.

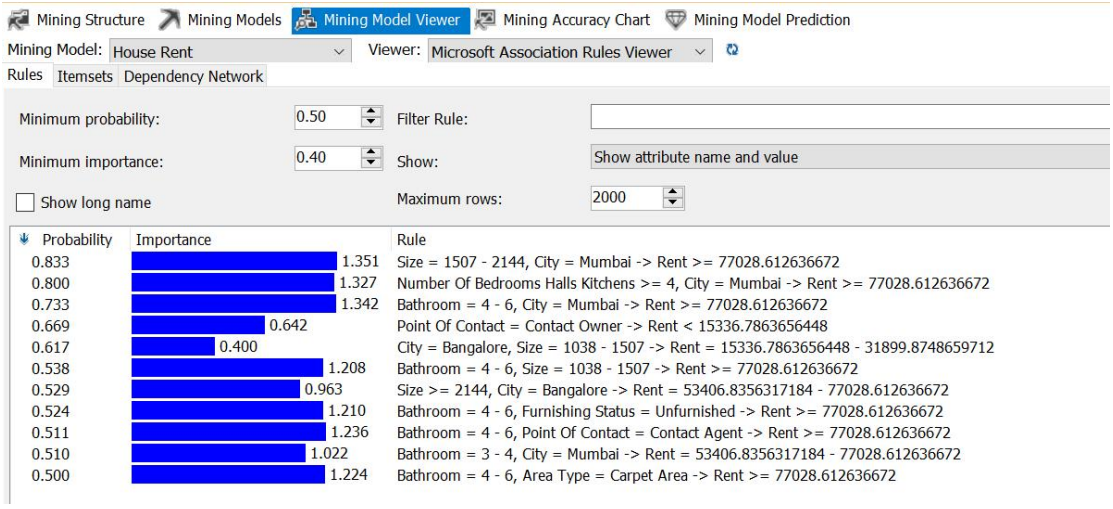


- Ta chọn Min probability = 0.4 - một giá trị trung bình để có cái nhìn toàn diện hơn về mối quan hệ giữa các yếu tố. Đặt Min importance > 0 cụ thể là 0.1 để xem xét các mối quan hệ có mức độ tương quan thuận giữa các biến.

- Ta thấy thuật toán sau khi thực thi tìm kiếm ra được 160 các luật kết hợp khác nhau với Min probability là 0.4 và Min importance là 0.1.
- Trong đó luật đầu tiên được tìm ra có xác suất là cao nhất có ý nghĩa với các căn hộ/ nhà/ phòng cho thuê ở trong thành phố Kolkata và có diện tích nhỏ hơn 641 thì giá thuê sẽ thấp hơn khoảng 15336. Và luật này có xác suất đúng đến 98.5%.
- Ở tab **Itemsets** ta lọc theo điều kiện **'City = Kolkata, Size < 641'**



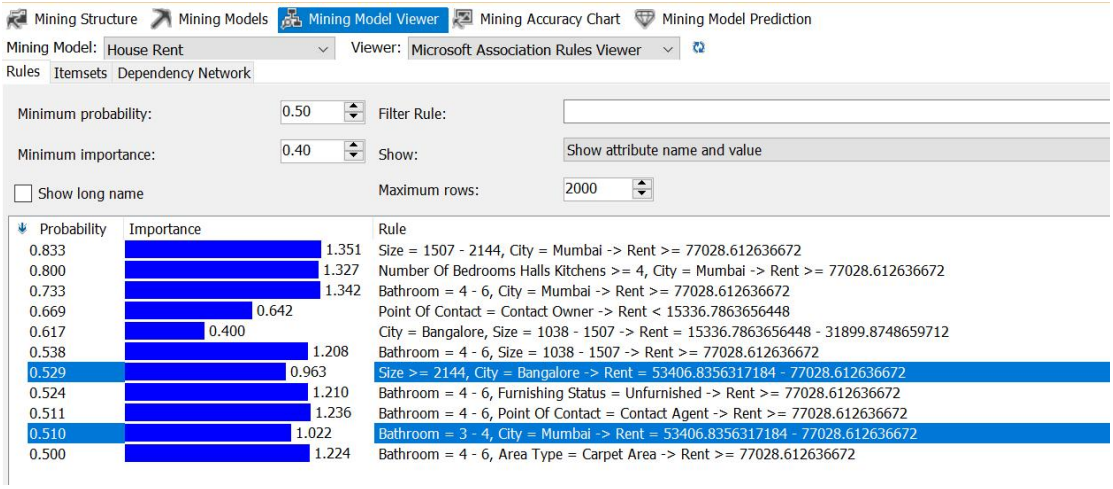
- Nhận thấy có 195 căn hộ/ nhà/ phòng cho thuê thuộc thành phố Kolkata có diện tích nhỏ hơn 641 thì có tới 192 trong tổng số đó có mức giá thuê thấp hơn khoảng 15336. Đó là lí do tại sao luật kết hợp **'City = Kolkata, Size < 641 -> Rent < 15336.7863656448'** lại chính xác cao đến vậy.
- Điều chỉnh tăng một số thông số để tìm ra các luật có ý nghĩa hơn
  - Minimum probability = 0.5 với ý nghĩa loại bỏ những yếu tố có xác suất xảy ra thấp hơn 0.5 và tập trung vào những biến có xác suất xảy ra cao để phân tích vào những yếu tố quan trọng ảnh hưởng đến giá thuê nhà.
  - Minimum importance = 0.4 với ý nghĩa muốn tập trung vào những mối quan hệ có mức độ tương quan đáng kể và bỏ qua những mối quan hệ yếu. Từ đó tìm ra những những yếu tố có ảnh hưởng quan trọng đến giá cho thuê nhà.



- Với lựa chọn Min probability = 0.5 và Min importance = 0.4 ta tìm được 11 luật kết hợp như hình bên trên. Các luật đưa ra sẽ được sắp xếp theo Probability (xác suất) giảm dần. Với luật đầu tiên được tìm ra ta thấy nếu Houses/ Apartment/ Flats có Size = 1507 - 2144 và có City thuộc 'Mumbai' thì Giá cho thuê Rent >= 77028 với xác suất probability = 0.833 và mức độ quan trọng importance = 1.351. Luật này đưa ra khá đúng với những gì trực quan hóa cho rằng những nơi cho thuê thuộc thành phố Mumbai đều có giá cho thuê khá cao. Hai luật kế tiếp được tìm ra cũng cho thấy các khu vực cho thuê thuộc 'Mumbai' có giá cho thuê hơn 70000.

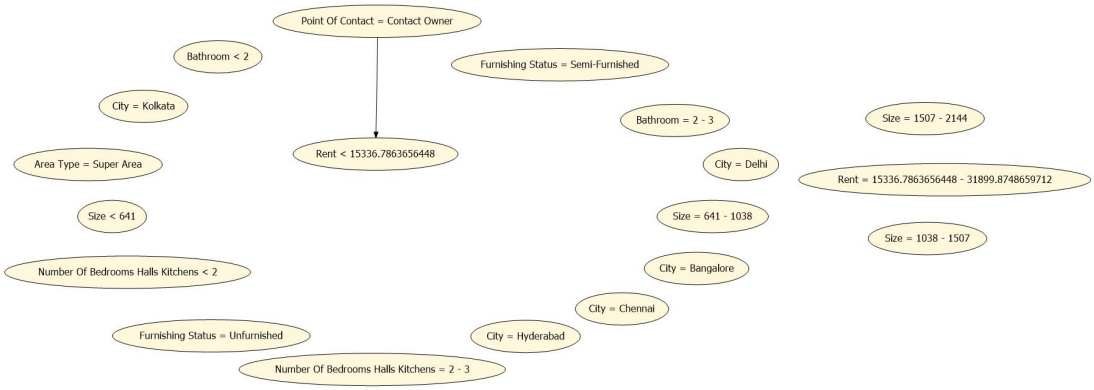
Nhận Xét: Khi ta lựa chọn các tham số có giá trị cao thì thuật toán tìm ra cho ta những luật kết hợp có ý nghĩa lớn trong việc phân tích những yếu tố ảnh hưởng đến giá thuê nhà

- Với giá thuê **Rent < 15336**: thì sẽ có sự ảnh hưởng của yếu tố **Point of Contact** cụ thể là Contact Owner (chủ nhà cho thuê) với mức độ tương quan là 0.642 và xác suất có thể xảy ra là 0.617.
- Với giá thuê **Rent = 15336 - 31899**: thì yếu tố ảnh hưởng đến sẽ là yếu tố **City** và **Size**. Hay nói cách khác có thể nói những nơi cho thuê thuộc thành phố Bangalore và có diện tích từ 1038 - 1507 thì sẽ có 61.7% khả năng giá cho thuê sẽ dao động từ 15336 - 31899.
- Với giá thuê **Rent = 53406 - 77028**: thì sẽ có yếu tố của **City, Size, Bathroom.** Cho thấy với các địa điểm nằm trong thành phố Mumbai, hoặc Bangalore đồng thời có số lượng phòng tắm cho thuê từ 3 - 4 phòng với diện tích hơn 2144 thì giá thuê sẽ dao động từ trung bình đến cao.

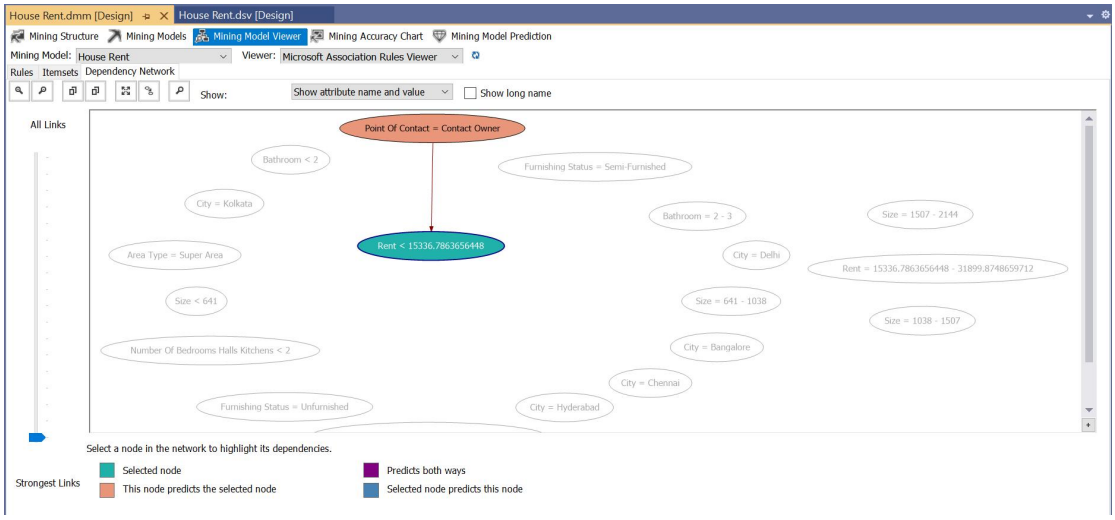


- Với giá thuê **Rent >= 77028:** phụ thuộc vào nhiều yếu tố như Size, City, Number of Bedrooms Halls Kitchens, Bathroom. Cho thấy những nơi cho thuê vừa có quy mô lớn với số lượng phòng cho thuê nhiều từ 4 - 6 với diện tích lớn hơn 1000 mà nằm trong thành phố đắt đỏ là Mumbai thì khả năng giá thuê sẽ rất cao.

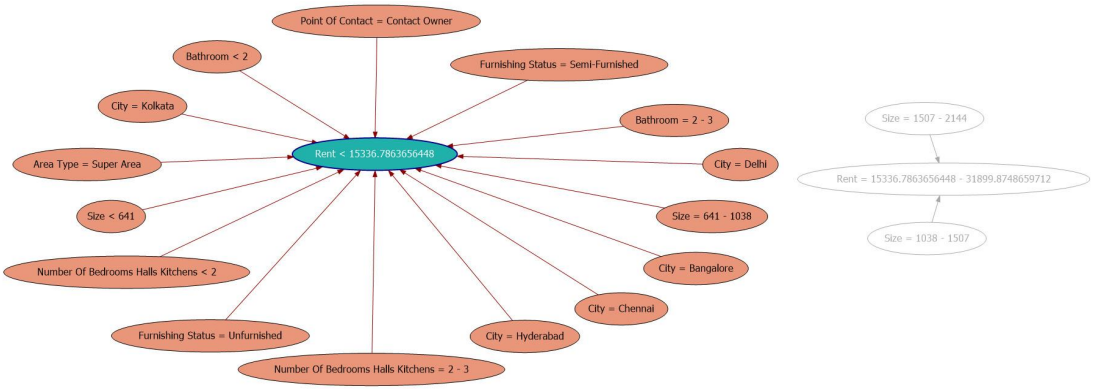
- Để dễ dàng xem các mức độ phụ thuộc giữa các itemset với nhau thì ta có thể xem ở tab **Dependency Network.**



- Ở Dependency Network, ta có thể xem được các mô tả mức độ phụ thuộc giữa các hạng mục trong luật kết hợp và mô tả theo độ mạnh (Strongest Links) đến giảm dần (kéo gần đến All Links).

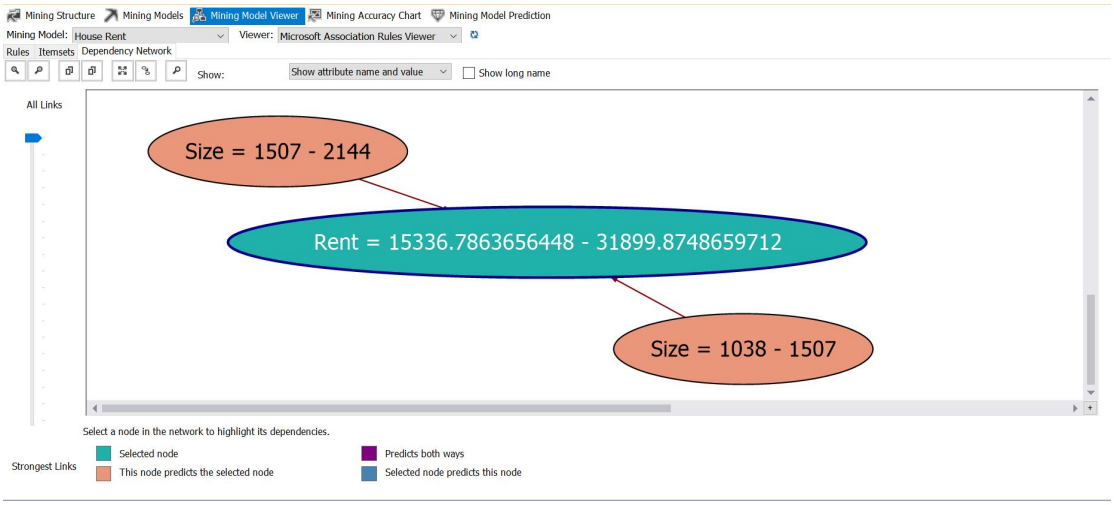


- Với liên kết mạnh mẽ nhất (Strongest Link) thì ta thấy luật Người đứng ra cho thuê nhà là chính chủ (Point of Contact = Contact Owner) thì mức giá thuê dưới 15336 (Rent < 15336.7863656448).



Nhận xét: Dựa vào tất cả những liên kết mà nó phụ thuộc đến giá thuê nhà thấp hơn 15336, cho thấy rằng những nơi cho thuê có **quy mô nhỏ** với diện tích dưới 1000 có từ 2 3 phòng cho thuê với tình trạng nội thất là không trang bị hay chỉ trang bị một ít, bên cạnh đó những nơi này thuộc khu vực 'Super' trong các thành phố như Kolkata, Hyderabad, Chennai, hay Bangalore, thậm chí trong thành phố Delhi mà do chính chủ đứng ra cho thuê thì khả năng giá thuê sẽ khá thấp.

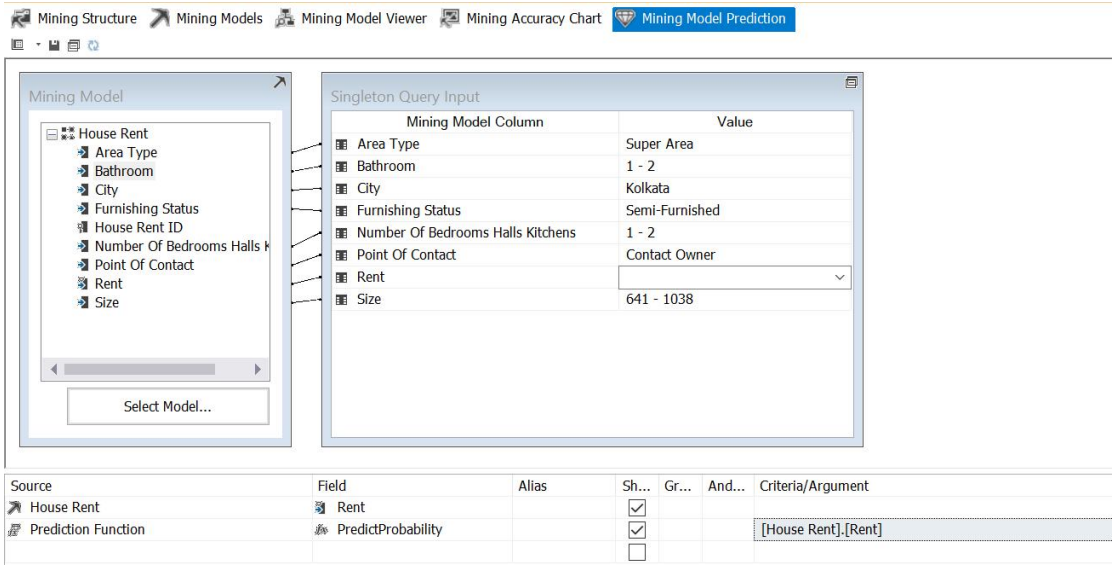
- Mô hình biểu diễn các yếu tố phụ thuộc ảnh hưởng đến mức giá cho thuê **từ 15336 - 31899**.



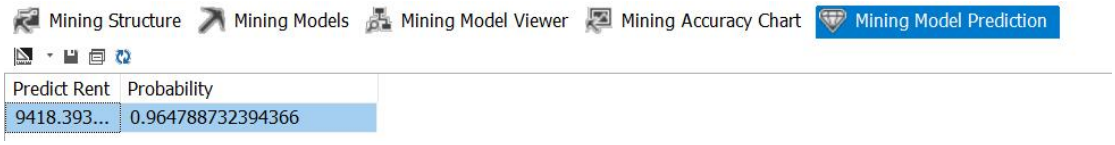
Nhận xét: Ta thấy mô hình không thể hiện rõ những yếu tố ảnh hưởng đến giá thuê từ 15336 - 31899. Duy chỉ có những thông tin liên quan đến Diện tích (từ 1038 - 2144). Tuy nhiên dựa vào bấy nhiêu thông tin thì không thể suy ra được rằng những nơi cho thuê có diện tích như thế thì giá thuê sẽ rơi vào 15336 - 31899.

4.2.3. Áp dụng Association Rule vào mô hình dự đoán giá thuê

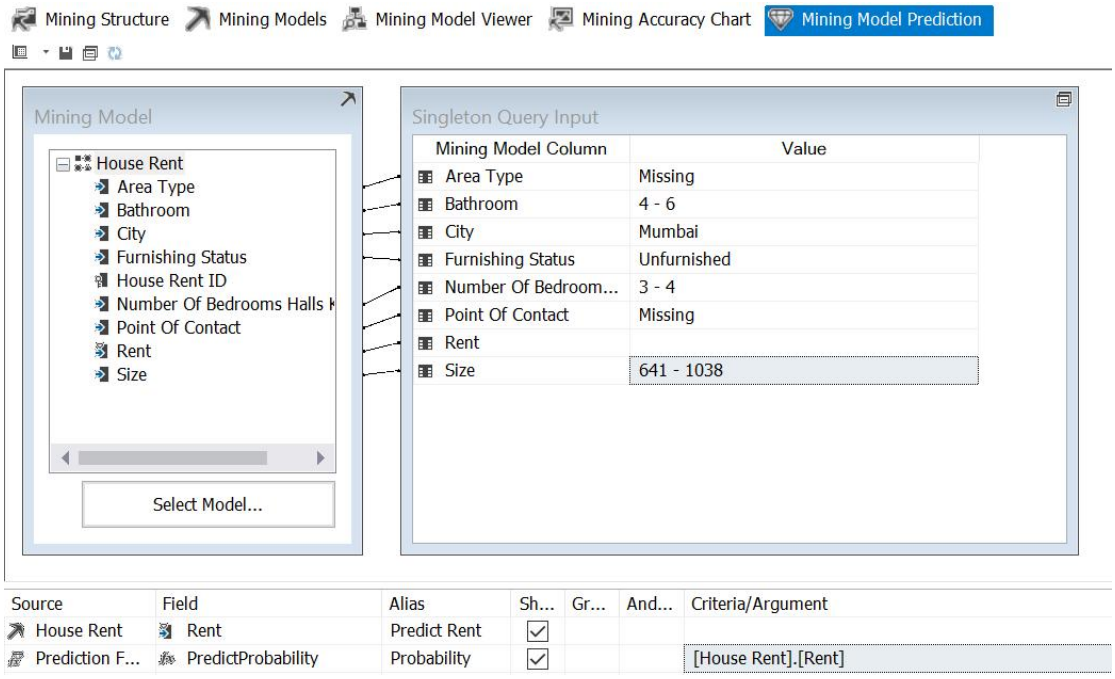
- Chọn tab **Mining Model Prediction** để tiến hành dự đoán dựa trên model của luật kết hợp.
- Điền các giá trị cho các thuộc tính trừ cột Rent (vì cột này dùng để dự đoán giá cho thuê). Các giá trị input được mô tả ở hình bên dưới.



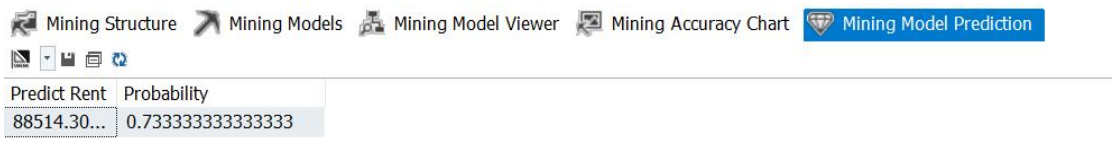
- Chọn **Result** để xem kết quả dự đoán Rent



- Với những giá trị đã chọn thì kết quả dự đoán giá thuê là khoảng **9418** với xác suất là 0.964. Ta thấy mô hình dự đoán đúng so với kết quả đã phân tích trước đó.
- Thử một số giá trị khác với thành phố Mumbai và số lượng cho thuê phòng lớn.



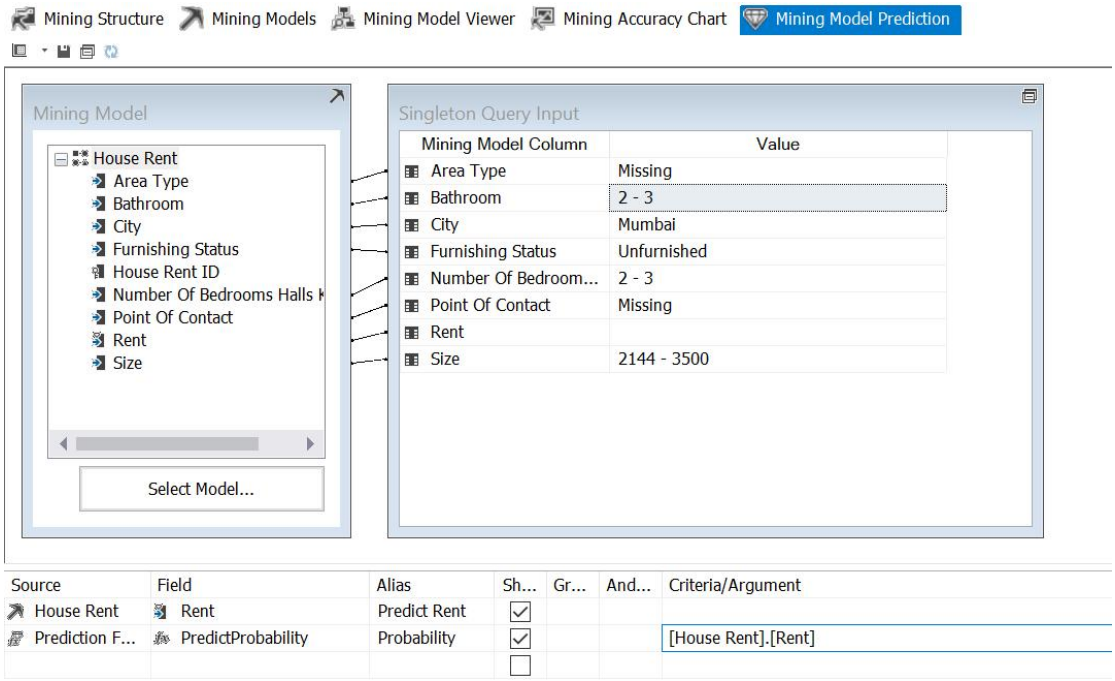
- Kết quả thu được



- Ta thấy nơi được chọn thuộc thành phố Mumbai và với số lượng phòng cho thuê nhiều và diện tích khá bé thì giá thuê dự đoán được rơi vào 88514 với khả năng là

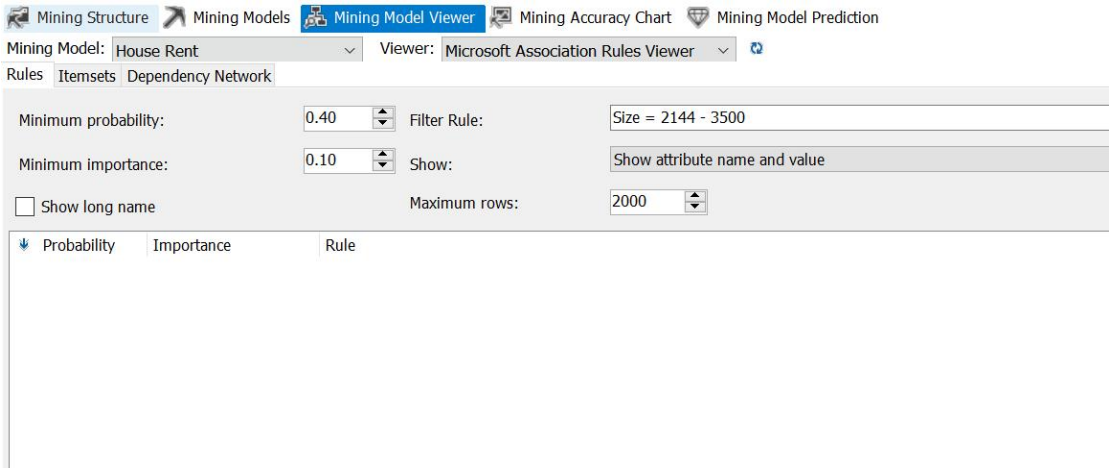
73.33%. Dự đoán đúng như kết quả phân tích được với những nơi cho thuê hơn 77 nghìn.

- Giữ nguyên thành phố Mumbai, thử thay đổi một số giá trị khác như Bathroom, Number of Bedrooms, Halls, Kitchens, và Size.

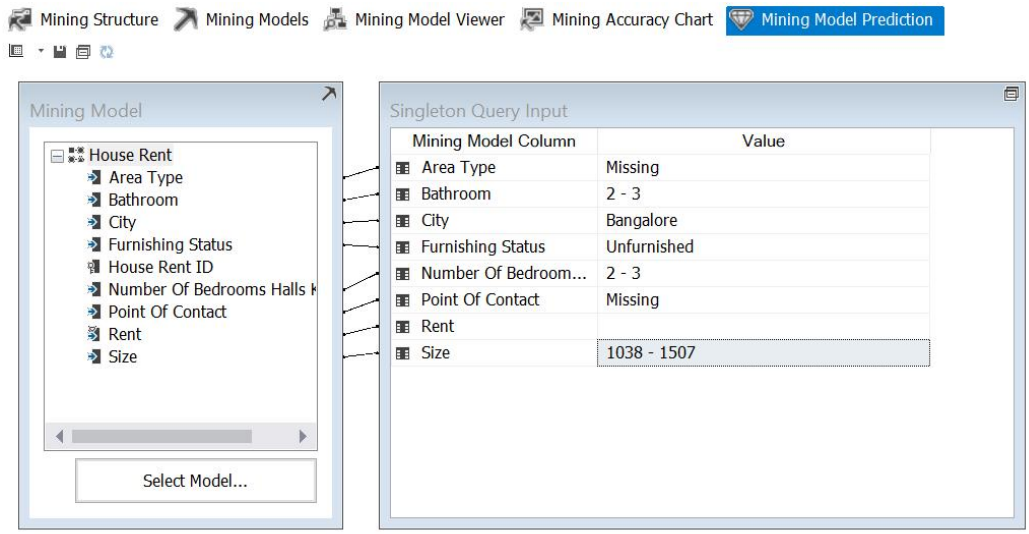


- Kết quả như sau

- Ta thấy kết quả đưa ra khác với những gì đã phân tích mặc dù quy mô cho thuê lớn với diện tích hơn 2000 feet vuông, tọa lạc ở thành phố Mumbai mà giá thuê dự đoán lại thấp hơn 15000. Lý do là vì những luật kết hợp mà thuật toán đưa ra thì không có luật nào thể hiện tương quan giữa Size = 2144 - 3500 với giá thuê Rent. Nên khi đưa ra kết quả dự đoán sẽ sai so với kết quả đã phân tích được.

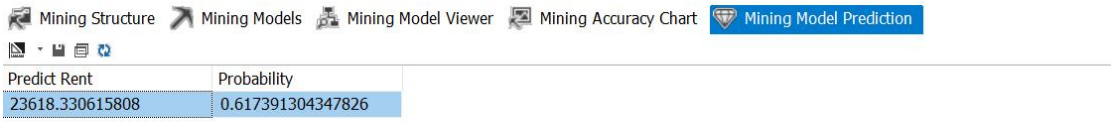


- Thử dự đoán với một thành phố khác ngoài Mumbai



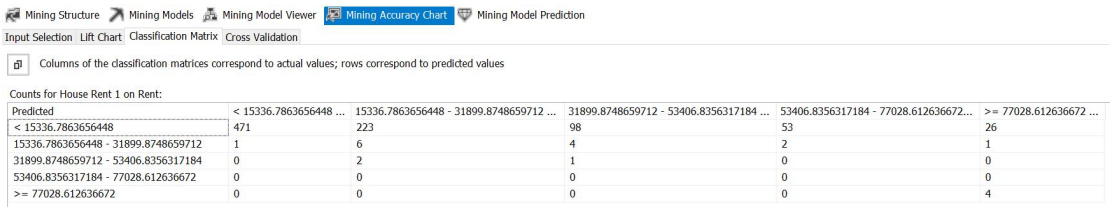
Source	Field	Alias	Sh...	Gr...	And...	Criteria/Argument
House Rent	Rent	Predict Rent	☑			
Prediction F...	PredictProbability	Probability	☑			[House Rent].[Rent]
			☐			

- Kết quả dự đoán



Predict Rent	Probability
23618.330615808	0.617391304347826

- Ta thấy với địa điểm tọa lạc thuộc thành phố Bangalore với quy mô và diện tích vừa thì giá cho thuê dự đoán là khoảng 23618 với xác suất là 61.73%.
- Đánh giá về khả năng dự đoán trên tập test (20% của tập dữ liệu)



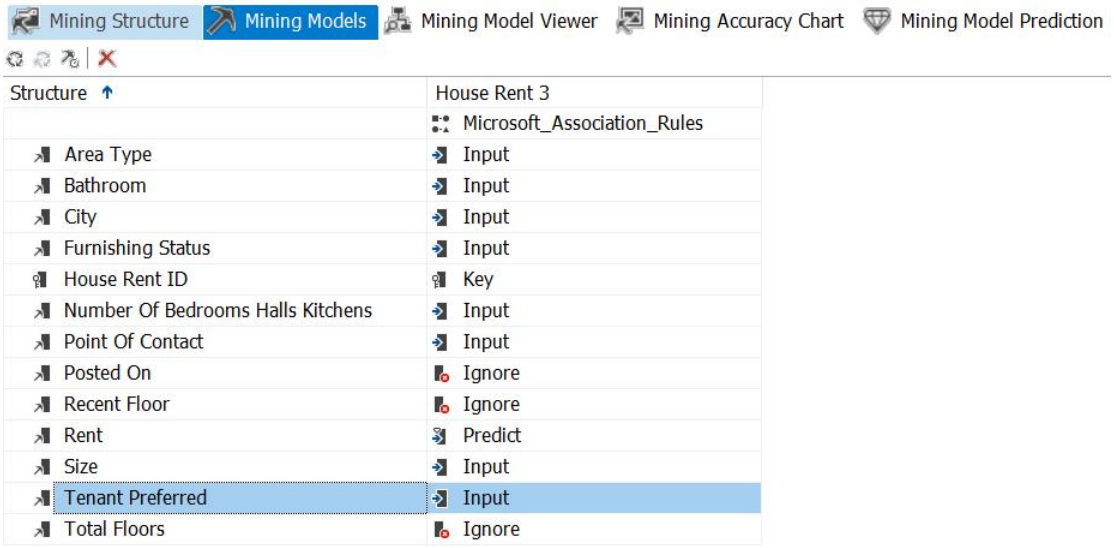
Predicted	< 15336.7863656448 ...	15336.7863656448 - 31899.8748659712 ...	31899.8748659712 - 53406.8356317184 ...	53406.8356317184 - 77028.612636672...	>= 77028.612636672 ...
< 15336.7863656448	471	223	98	53	26
15336.7863656448 - 31899.8748659712	1	6	4	2	1
31899.8748659712 - 53406.8356317184	0	2	1	0	0
53406.8356317184 - 77028.612636672	0	0	0	0	0
>= 77028.612636672	0	0	0	0	4

  + Tổng số trường hợp mà mô hình dự đoán đúng là 482 trên tổng số 892 case
  + Accuracy: 0.5403 ~ 54%

Nhận xét: Như vậy thuật toán chỉ dự đoán đúng được những trường hợp xuất hiện trong các luật kết hợp. Còn với những trường hợp ngoài thì sẽ cho ra kết quả dự đoán không chính xác.

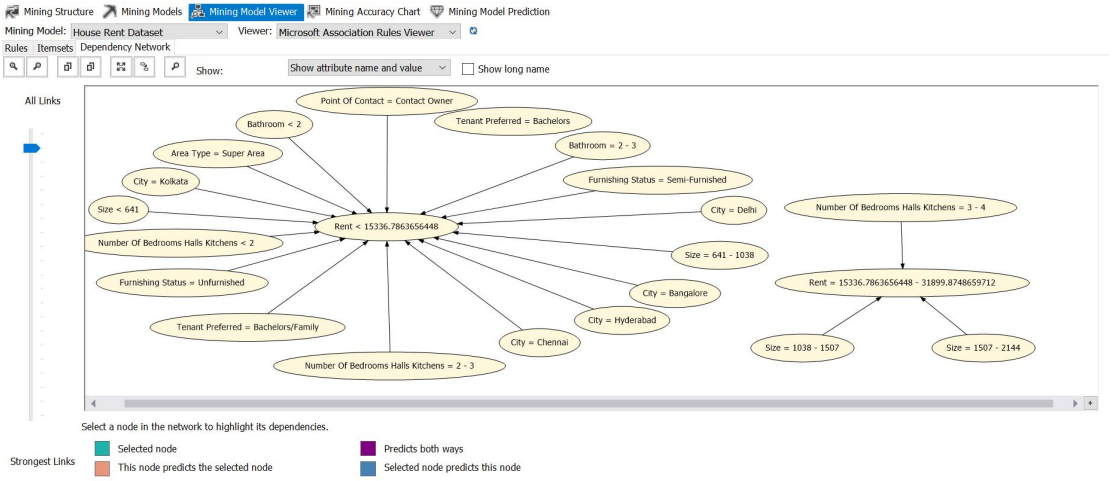
4.2.4. Khai phá dữ liệu với sự thay đổi các biến đầu vào

- Lần này nhóm sẽ tiến hành thử nghiệm với dữ liệu đầu vào có nhiều thuộc tính hơn ban đầu. Thêm thuộc tính Tenant Preferred vào mô hình.

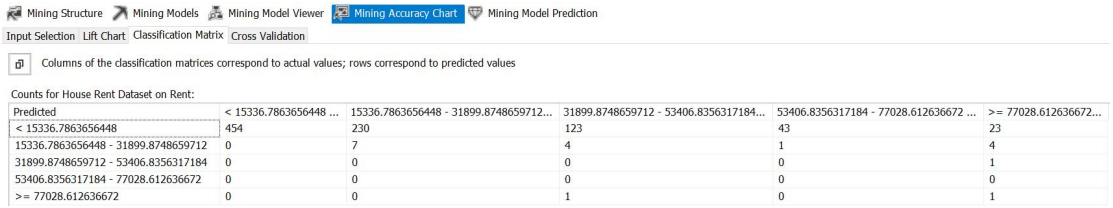


Structure	House Rent 3
	Microsoft_Association_Rules
Area Type	Input
Bathroom	Input
City	Input
Furnishing Status	Input
House Rent ID	Key
Number Of Bedrooms Halls Kitchens	Input
Point Of Contact	Input
Posted On	Ignore
Recent Floor	Ignore
Rent	Predict
Size	Input
Tenant Preferred	Input
Total Floors	Ignore

- Kết quả thu được



- Ta thấy không có sự khác biệt đáng kể giữa các đối tượng ưu tiên cho thuê với giá thuê nhà.
- Đánh giá khả năng dự đoán trên mô hình (tập test)



Predicted	< 15336.7863656448 ...	15336.7863656448 - 31899.8748659712...	31899.8748659712 - 53406.8356317184...	53406.8356317184 - 77028.612636672 ...	>= 77028.612636672...
< 15336.7863656448	454	230	123	43	23
15336.7863656448 - 31899.8748659712	0	7	4	1	4
31899.8748659712 - 53406.8356317184	0	0	0	0	1
53406.8356317184 - 77028.612636672	0	0	0	0	0
>= 77028.612636672	0	0	1	0	1

  + Tổng số trường hợp mà mô hình dự đoán đúng là 462 trên tổng số 892 case
  + Accuracy: 0.5179 ~ 52%
- Nhận thấy khả năng dự đoán của mô hình này so với mô hình trước là thấp hơn.

4.2.5. Đánh giá mô hình Association Rule

4.2.5.1. Nhận xét chung

- Những luật kết hợp mà thuật toán đem lại đã đưa ra những kết quả khá có ý nghĩa khi tìm ra được những yếu tố ảnh hưởng đến giá thuê nhà cụ thể là mức dưới 15336 và mức trên 77028. Điều này sẽ giúp cho ta có cái nhìn rõ hơn về những yếu tố ảnh hưởng mạnh đến hai mức giá thuê trên. Và thấy được những yếu tố nào gây ra sự khác biệt rõ rệt giữa hai mức giá thuê này.

- Đánh giá về khả năng dự đoán trên tập test (20% của tập dữ liệu)
  - Mô hình thứ nhất

Mining Structure | Mining Models | Mining Model Viewer | Mining Accuracy Chart | Mining Model Prediction

Input Selection | Lift Chart | Classification Matrix | Cross Validation

Columns of the classification matrices correspond to actual values; rows correspond to predicted values

Counts for House Rent 1 on Rent:

Predicted	< 15336.7863656448 ...	15336.7863656448 - 31899.8748659712 ...	31899.8748659712 - 53406.8356317184 ...	53406.8356317184 - 77028.612636672...	>= 77028.612636672 ...
< 15336.7863656448	471	223	98	53	26
15336.7863656448 - 31899.8748659712	1	6	4	2	1
31899.8748659712 - 53406.8356317184	0	2	1	0	0
53406.8356317184 - 77028.612636672	0	0	0	0	0
>= 77028.612636672	0	0	0	0	4

    + Tổng số trường hợp mà mô hình dự đoán đúng là 482 trên tổng số 892 case
    + Accuracy: 0.5403 ~ 54%
  - Mô hình thứ hai

Mining Structure | Mining Models | Mining Model Viewer | Mining Accuracy Chart | Mining Model Prediction

Input Selection | Lift Chart | Classification Matrix | Cross Validation

Columns of the classification matrices correspond to actual values; rows correspond to predicted values

Counts for House Rent Dataset on Rent:

Predicted	< 15336.7863656448 ...	15336.7863656448 - 31899.8748659712...	31899.8748659712 - 53406.8356317184...	53406.8356317184 - 77028.612636672 ...	>= 77028.612636672...
< 15336.7863656448	454	230	123	43	23
15336.7863656448 - 31899.8748659712	0	7	4	1	4
31899.8748659712 - 53406.8356317184	0	0	0	0	1
53406.8356317184 - 77028.612636672	0	0	0	0	0
>= 77028.612636672	0	0	1	0	1

    + Tổng số trường hợp mà mô hình dự đoán đúng là 462 trên tổng số 892 case
    + Accuracy: 0.5179 ~ 52%
- Như vậy có thể thấy thuật toán Luật kết hợp không thích hợp để dự đoán kết quả vì nó chỉ áp dụng với những trường hợp đúng với các luật kết hợp mà nó sinh ra còn những trường không tồn tại thì nó không đem lại kết quả đúng.

4.2.5.2. Ứng dụng

- Nếu đóng vai trò là một người đang tìm thuê nhà đang muốn tìm những căn nhà phù hợp với kinh phí của họ thì cần cân nhắc các yếu tố ảnh hưởng đến giá cho thuê như đã phân tích là thành phố mà nơi cho thuê tọa lạc, hay qui mô căn hộ.

+ Ví dụ họ muốn tìm thuê nhà có giá tiền dưới 15336 thì nên lựa chọn các nơi có đặc điểm như quy mô căn hộ nhỏ, diện tích bé, số lượng phòng cho thuê không nhiều từ 2 - 3 phòng, không trang bị nội thất hoặc chỉ trang bị những thứ cơ bản và tọa lạc ở những thành phố nhỏ như Kolkata, Hyderbad,.. và nên thuê từ chủ nhà để không phải chịu những chi phí môi giới,...

- Nếu đóng vai tròn là chủ đầu tư muốn tăng lợi nhuận thì dựa vào những luật kết hợp từ đó tìm hiểu được những điều kiện giúp tăng giá cho thuê cụ thể là nên đầu tư những căn hộ phòng cho thuê ở thành phố Mumbai hoặc Bangalore, có qui mô vừa hoặc lớn với số phòng cho thuê nhiều,.. thì chắc chắn sẽ đem lại lợi nhuận cao.

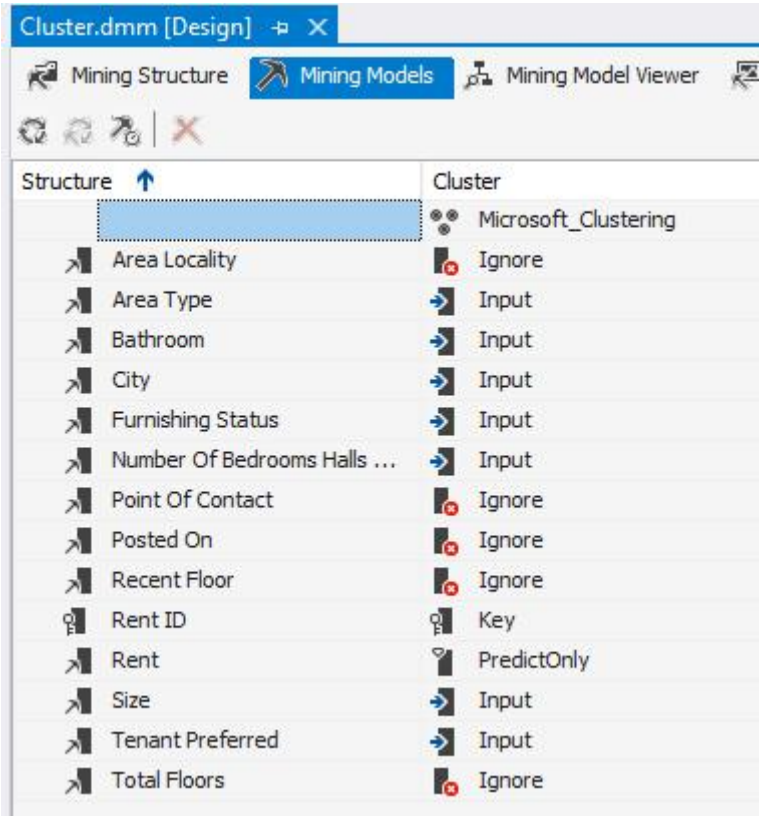
4.3. Clustering

4.3.1. Giới Thiệu

- Microsoft Clustering là một kỹ thuật phân cụm học không giám sát. Áp dụng thuật toán phân cụm để phân nhóm các điểm dữ liệu dựa trên một số đặc trưng. Các cụm được tìm ra chứa các điểm dữ liệu có tính chất tương tự nhau.
- Với tập dữ liệu House Rent của nhóm, việc áp dụng Clustering giúp cho việc phân nhóm các căn hộ/ nhà/ phòng cho thuê có tính chất tương đồng nhau dựa trên các đặc trưng như diện tích (size), số phòng ngủ (bathroom), khu vực (area type), thành phố (city),... Điều này giúp chúng ta có cái nhìn tổng thể về thị trường nhà cho thuê ở Ấn Độ, đồng thời giúp các chủ đầu tư, chủ hộ đưa ra các quyết định kinh doanh phù hợp.
- Sau khi đã phân cụm các căn hộ/ nhà/ phòng cho thuê dựa trên các đặc trưng cụ thể thì có thể áp dụng dự đoán thông tin của căn hộ mới thông qua cụm mà nó được gán nhãn.

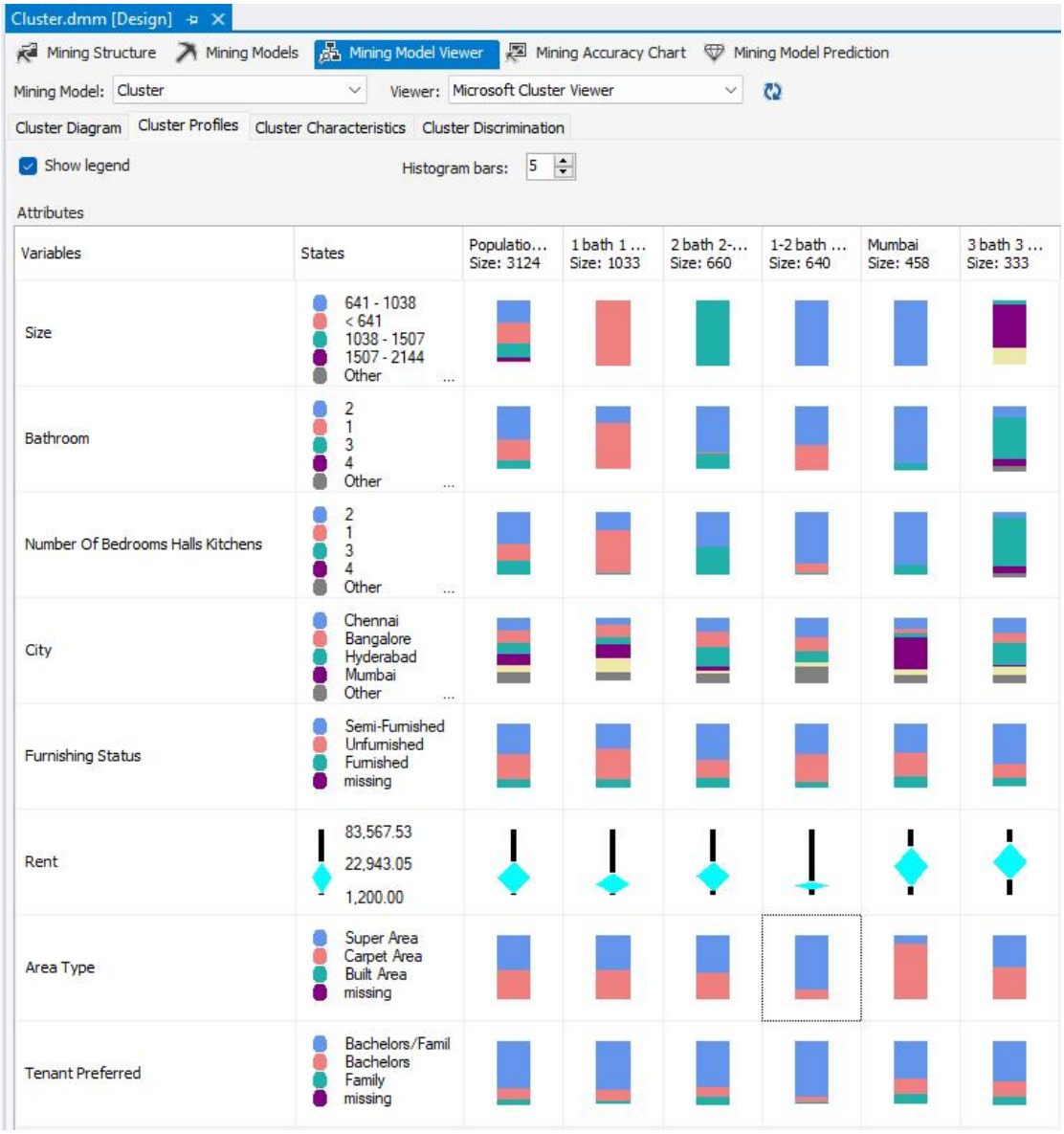
4.3.2. Tiến Hành Khai Phá Phân Cụm

- Khởi tạo Mining Structure chọn thuật toán Microsoft Clustering để khai phá luật kết hợp.
- Thực hiện lựa chọn các biến dự đoán cho mô hình là: Area Type, Bathroom, City, Furnishing Status, Number of Bedrooms-Halls-Kitchens, Size, Tenant Preferred.
- Thực hiện lựa chọn biến kết quả cho mô hình là: Rent.



- Process mô hình sau đó vào tab Cluster Profiles, thực hiện đặt lại tên cho các cụm theo các đặc tính thể hiện cho cụm ấy. Thứ tự từ cụm 1 đến 5 được đặt tên như sau:

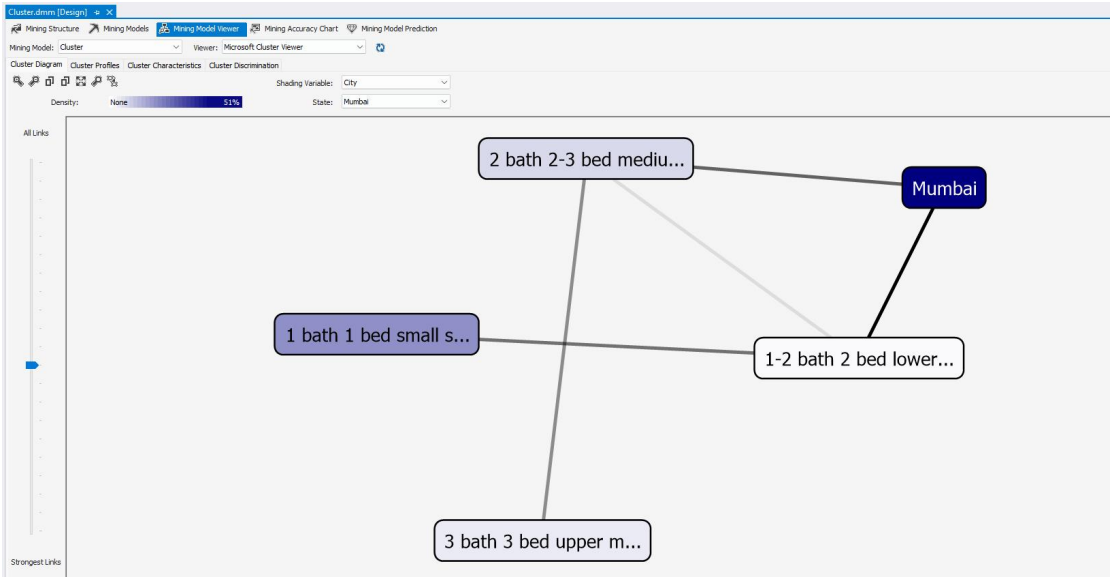
+ 1 bath 1 bed small size;
+ 2 bath 2-3 bed medium size;
+ 1-2 bath 2 bed lower medium size;
+ Mumbai 2 bath 2 bed;
+ 3 bath 3 bed upper medium size.



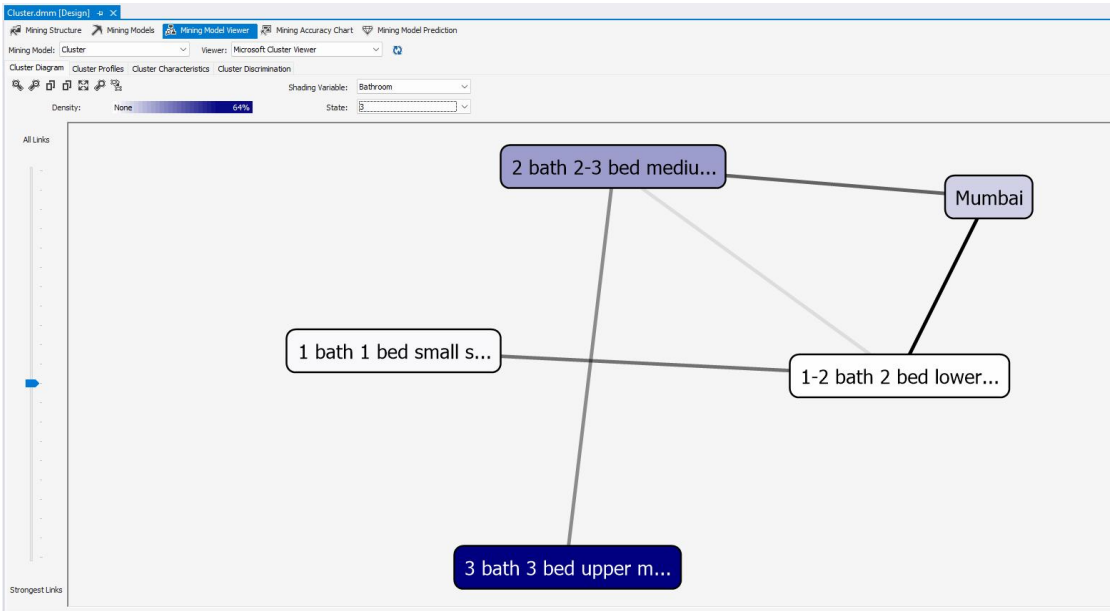
- Vào tab Cluster Diagram để xem sự tương quan của các cụm theo các biến dự đoán và giá trị khác nhau.

+ Khi chọn Shading Variable là ‘City’ và State là ‘Mumbai’, ta thấy được cụm Mumbai có độ đậm cao nhất, cho thấy các dòng có biến dự đoán ‘City’ với giá

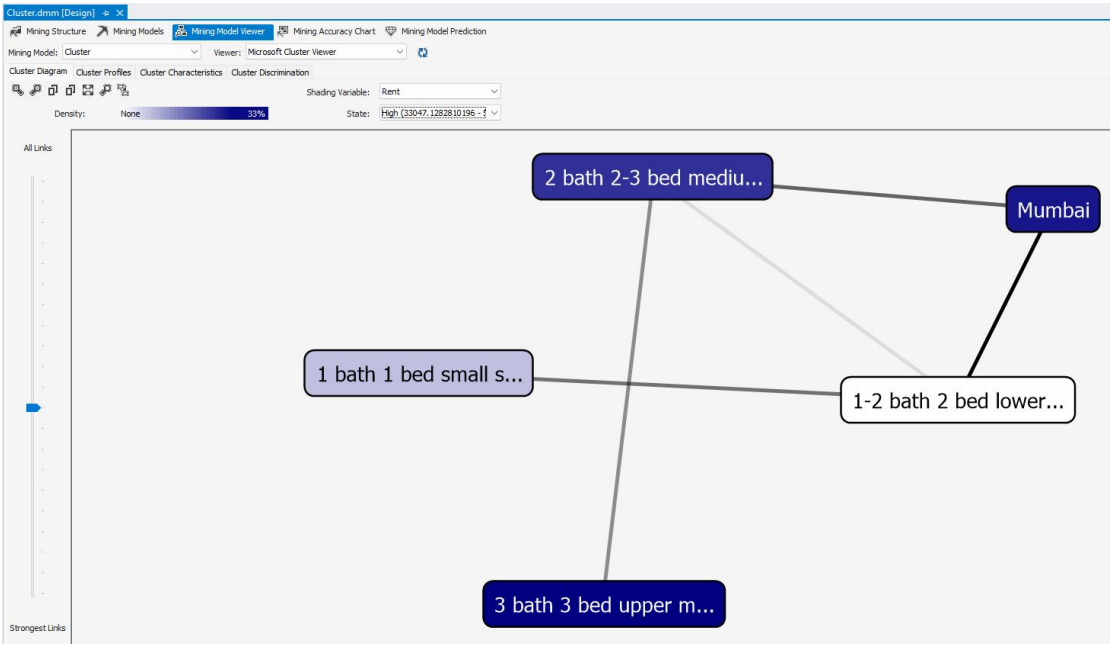
trị là ‘Mumbai’ thuộc nhiều vào cụm Mumbai 2 bath 2 bed.



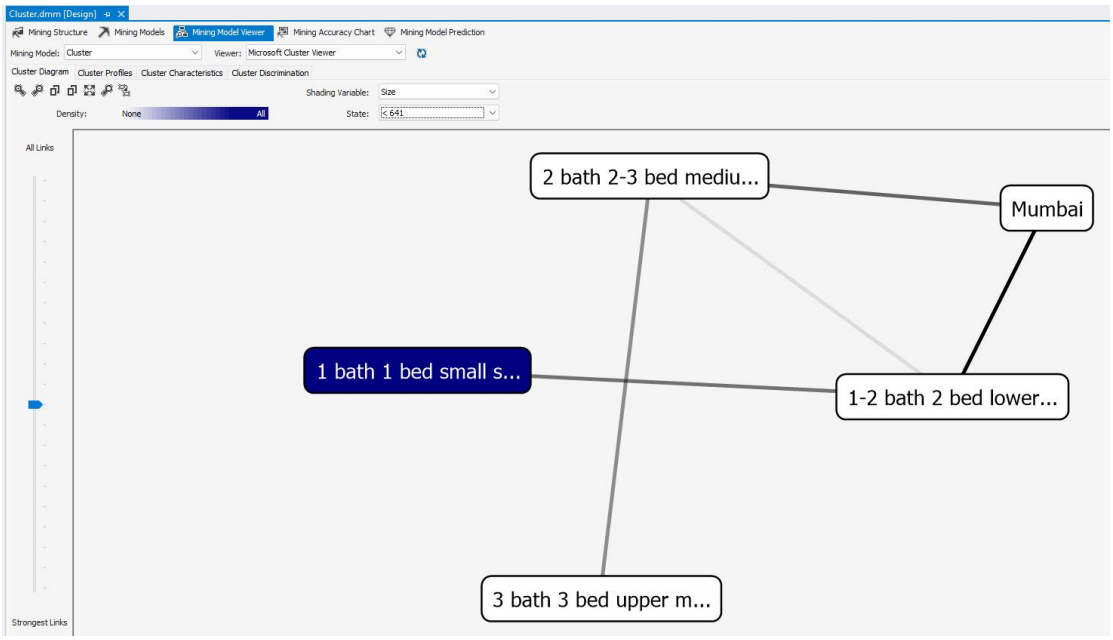
+ Khi chọn Shading Variable là ‘Bathroom’ và State là ‘3’, ta thấy được cụm 3 bath 3 bed upper medium size có độ đậm cao nhất, cho thấy các dòng có biến dự đoán ‘Bathroom’ với giá trị là ‘3’ thuộc nhiều vào cụm 3 bath 3 bed upper medium size.



+ Khi chọn Shading Variable là ‘Rent’ và State là ‘High’, ta thấy được 3 cụm Mumbai, 3 bath 3 bed upper medium size, 2 bath 2-3 bed medium size có độ đậm cao hơn 2 cụm còn lại, cho thấy các dòng thuộc tính Rent từ 33000-53000 thuộc nhiều vào 3 cụm này.



+ Khi chọn Shading Variable là ‘Size’ và State là ‘<641’, ta thấy được cụm 1 bath 1 bed small size có độ đậm cao nhất, cho thấy các dòng có biến dự đoán ‘Size’ với giá trị <641 thuộc nhiều vào cụm 1 bath 1 bed small size.



- Vào tab Mining Model Prediction, điền vào Singleton Query Input các giá trị cho các biến dự đoán và mong đợi rằng tập giá trị này sẽ rơi vào cụm này. Các giá trị theo các biến như sau:

+ Area type = ‘Carpet Area’

+ Bathroom = 2

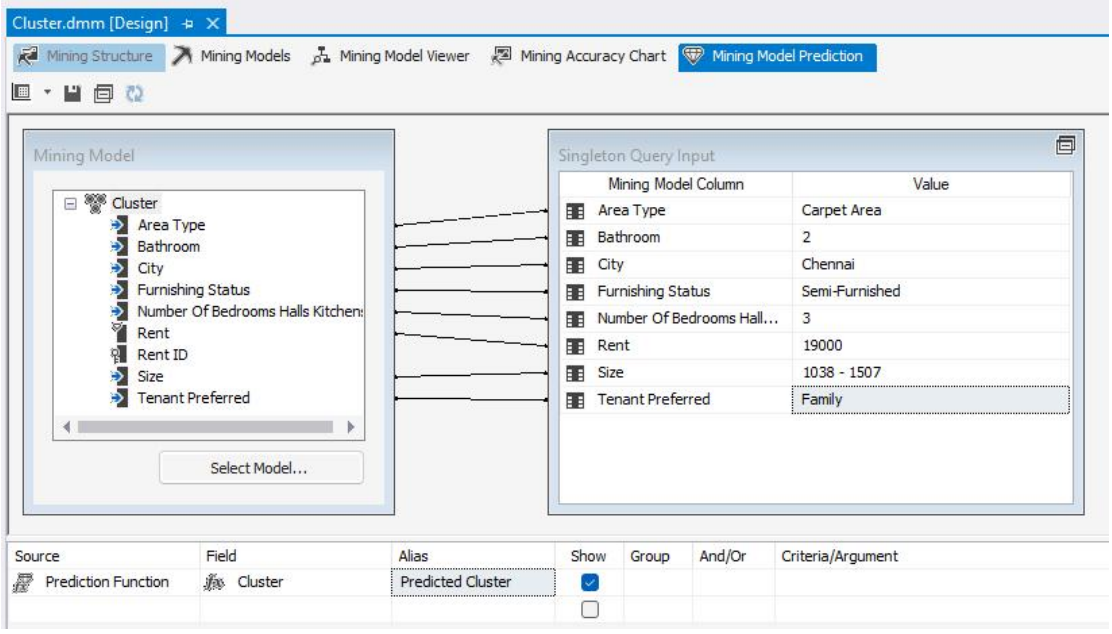
+ City = ‘Chennai’

+ Furnishing Status = ‘Semi - Furnished’

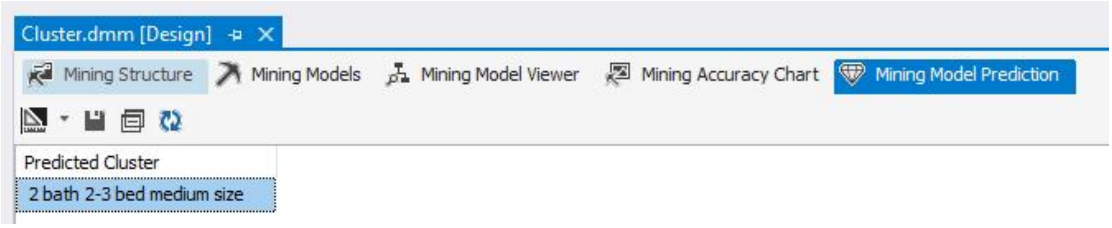
+ Number of Bedrooms-Halls-Kitchens = 3

+ Rent = 19000

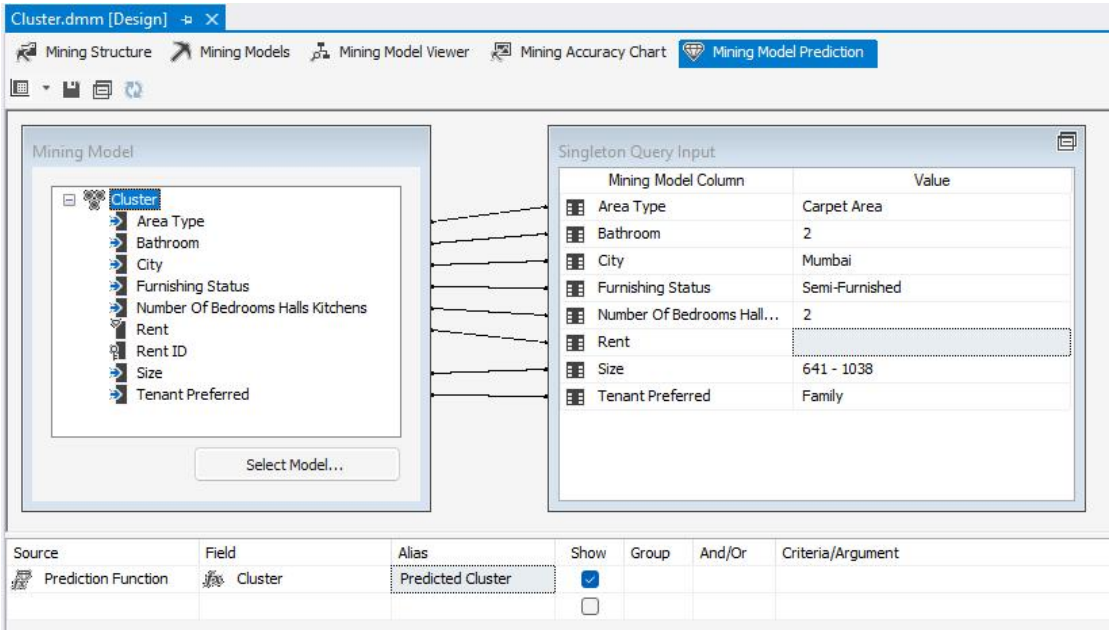
+ Size = ‘1038-1507’
+ Tenant Preferred = ‘Family’
- Ta mong đợi rằng tập giá trị này sẽ rơi vào cụm 2 bath 2-3 bed medium size vì đã cố tình đặt giá trị cho Bathroom là 2 và Number of Bedrooms-Halls-Kitchens là 3.



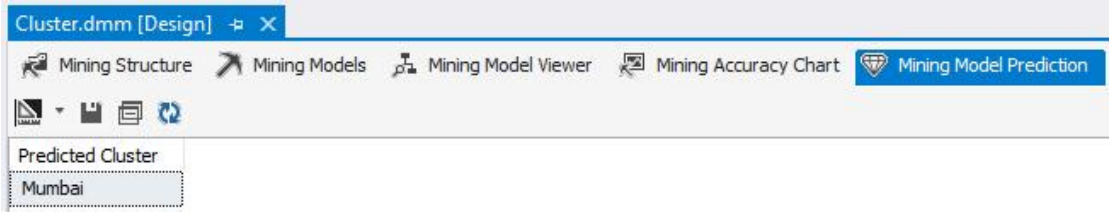
- Kết quả mô hình dự đoán được là cụm 2 bath 2-3 bed medium size, đúng với mong đợi ban đầu.



- Tiếp tục thử một bộ giá trị khác trong Mining Model Prediction.
Các giá trị theo các biến như sau:
+ Area type = ‘Carpet Area’
+ Bathroom = 2
+ City = ‘Mumbai’
+ Furnishing Status = ‘Semi - Furnished’
+ Number of Bedrooms-Halls-Kitchens = 2
+ Rent = ?
+ Size = ‘647-1038’
+ Tenant Preferred = ‘Family’
- Ta mong đợi rằng tập giá trị này sẽ rơi vào cụm Mumbai 2 bath 2 bed.

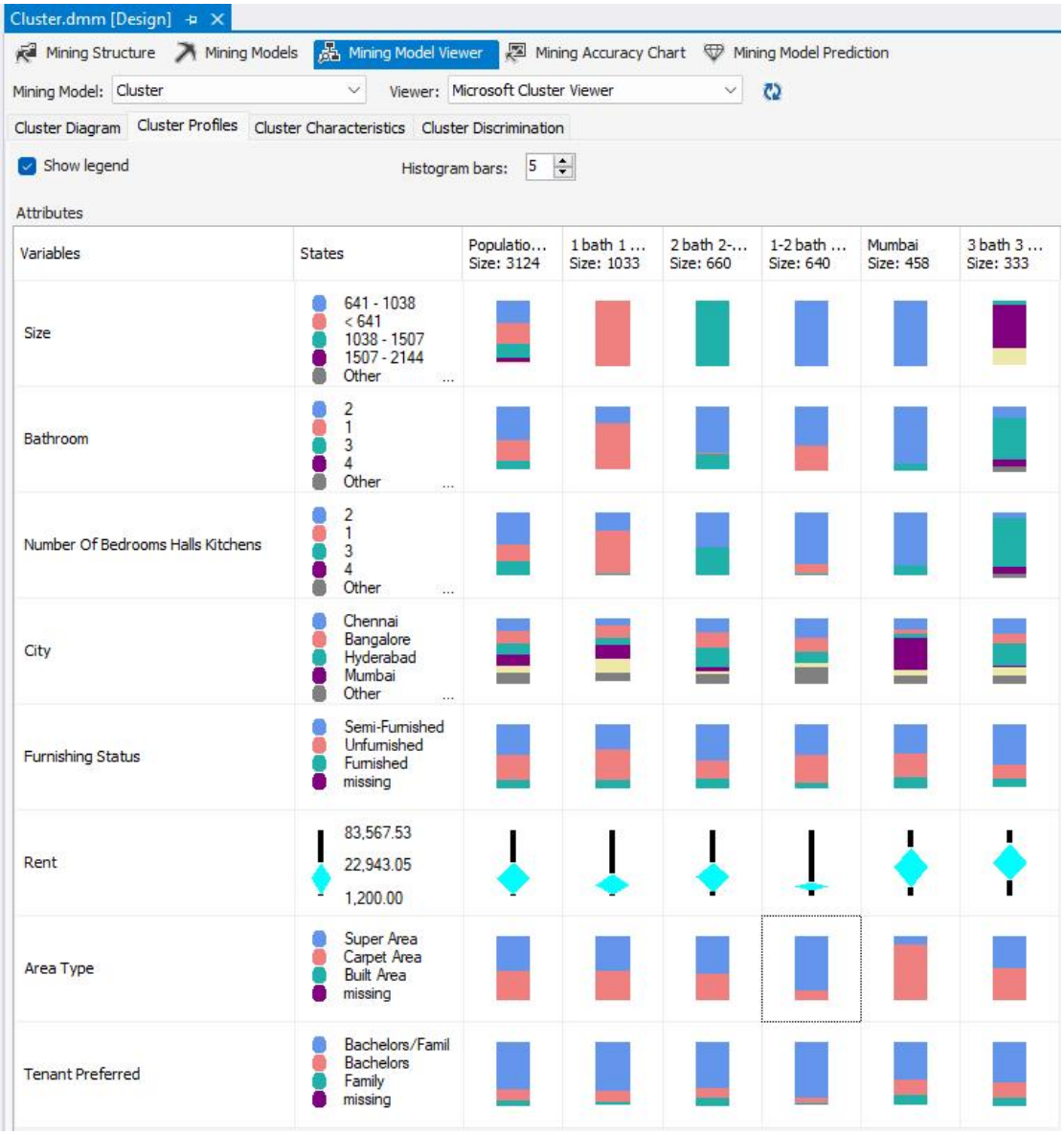


- Kết quả mô hình dự đoán được là cụm Mumbai 2 bath 2 bed, đúng với mong đợi ban đầu.



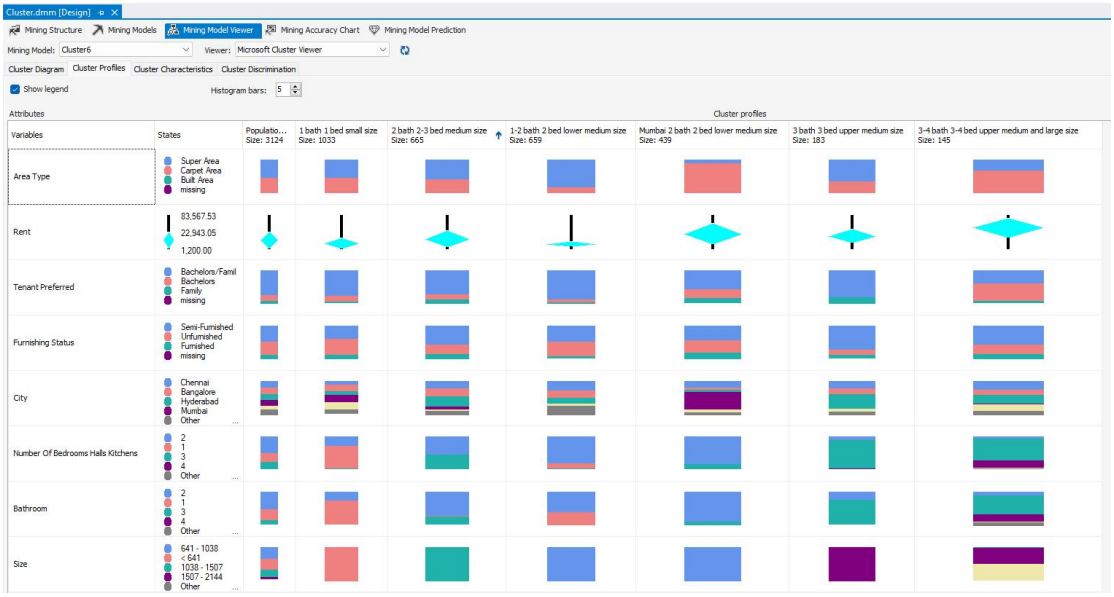
4.3.3. Khai phá với số cụm khác nhau

- Ban đầu nhóm thực hiện phân cụm bằng thuật toán Microsoft Clustering với số lượng cụm tối ưu tiêu chuẩn là 5 cụm. Các cụm được phân tương đối khác nhau và có những đặc tính riêng tượng trưng cho mỗi cụm. Tuy vậy, các đặc tính ấy ở một số cụm vẫn có sự trùng lắp vào nhau.



(Cụm 1-2 bath 2 bed lower medium size và cụm Mumbai 2 bath 2 bed có đặc tính size trùng lắp nhau nhiều)

- Nhóm phát hiện ra rằng khi giảm số lượng cụm và process lại mô hình thì sự trùng lắp các đặc tính này xảy ra nhiều hơn. Và ngược lại khi tăng số cụm lên, sự trùng lắp giảm đi và các cụm nổi bật lên bởi những giá trị khác nhau giữa các đặc tính. Nhóm đã tìm ra số cụm phù hợp nhất với tập dữ liệu là 6 cụm.



- Thực hiện đặt lại tên 6 cụm để hiểu rõ về đặc điểm của cụm như sau:
    + 1 bath 1 bed small size;
    + 2 bath 2-3 bed medium size;
    + 1-2 bath 2 bed lower medium size;
    + Mumbai 2 bath bed lower medium size;
    + 3 bath 3 bed upper medium size;
    + 3-4 bath 3-4 bed upper medium and large size.

4.3.4. Ứng dụng mô hình

- Với mô hình phân cụm dùng Microsoft Clustering, tạo ra các cụm với những đặc trưng khác nhau, ở vai trò người dùng, nếu họ chỉ có một vài thông tin không đầy đủ, việc phân cụm sẽ giúp họ có được những dự đoán về các thông tin còn lại dựa vào cụm được phân vào với những thông tin đã có. Chẳng hạn với những thông tin như phòng đang cho thuê có 4 phòng ngủ, 4 phòng tắm, người dùng có thể mong đợi căn phòng ấy sẽ có kích thước upper medium hoặc large và có giá phòng trung bình cao. Mô hình này cũng có thể được dùng theo hướng ngược lại, thay vì dự đoán giá thuê, người dùng có thể xem xét giá thuê mong muốn của họ thuộc vào cụm vào và họ sẽ biết được với giá thuê đó, phòng thuê sẽ có những đặc trưng gì về số phòng, thành phố, kích thước,...

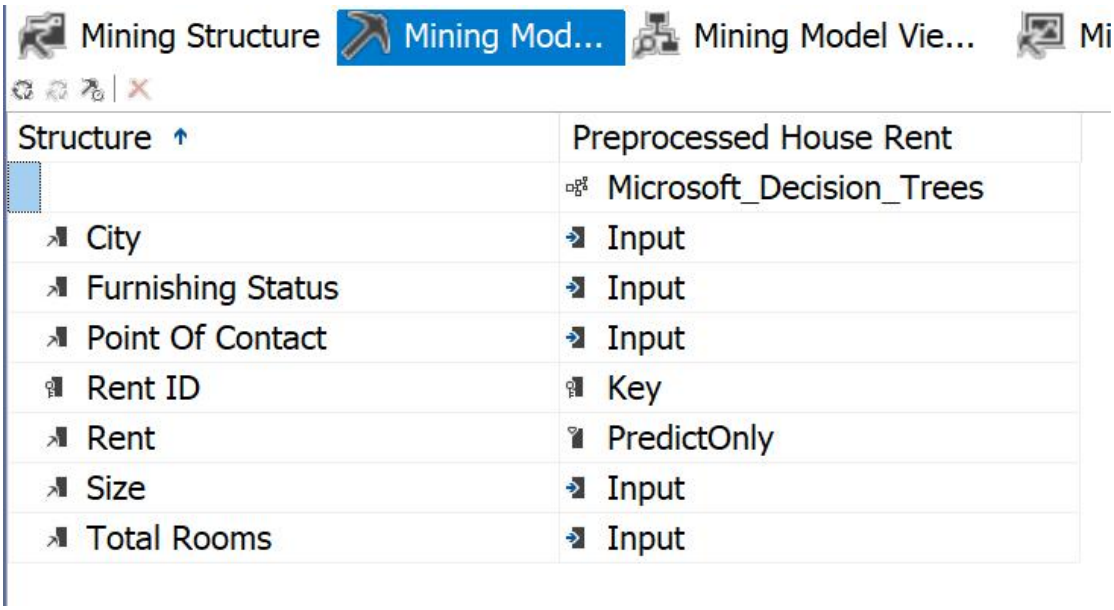
4.4. Classification

4.4.1. Giới Thiệu

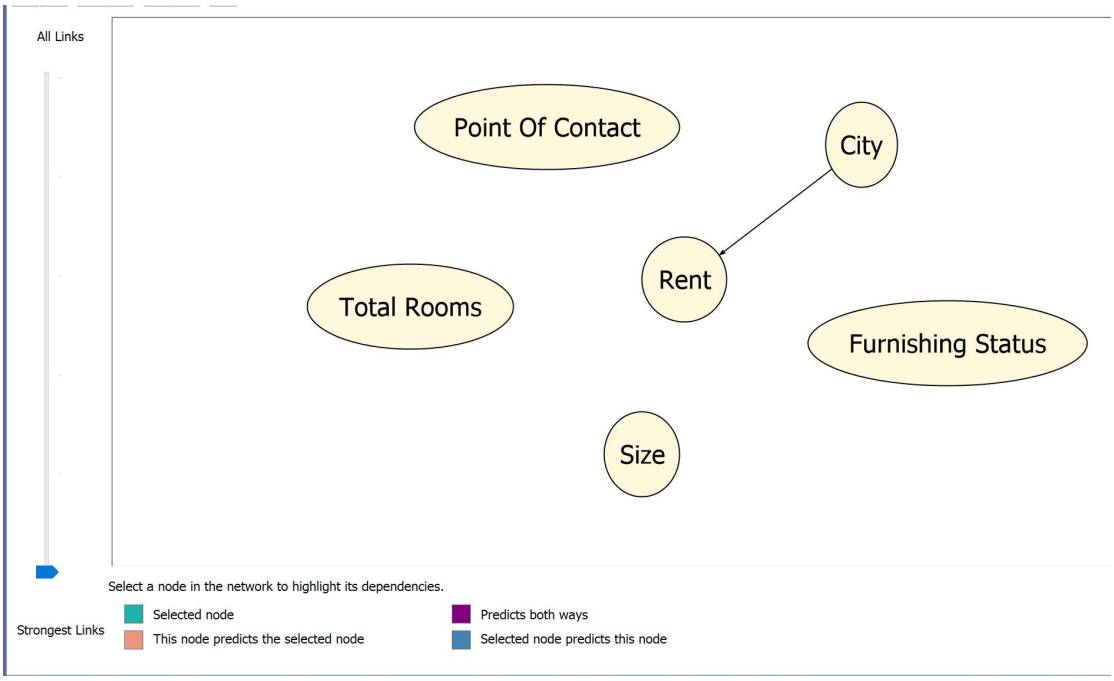
- Thuật toán phân loại (Classification) là một kỹ thuật học máy được sử dụng để dự đoán dựa trên các đặc trưng của điểm dữ liệu và các nhãn hoặc lớp đã biết trước đó.
- Với tập dữ liệu về giá thuê nhà ở Ấn Độ, thì ta có thể phân loại các căn hộ/ nhà/ phòng cho thuê theo các mức giá khác nhau. Từ đó, khi có một căn hộ/ nhà/ phòng cần cho thuê, việc xác định loại giá thuê của căn hộ này sẽ giúp cho chủ hộ hay chủ đầu tư có thể điều chỉnh mức giá thuê thích hợp hoặc người thuê nhà có thể lựa chọn những căn nhà có đặc điểm phù hợp trong mức giá phù hợp với mình.
- Thuật toán nhóm sử dụng chính là:
  + Decision Tree: Xây dựng cây quyết định dựa trên các quy tắc phân loại để dự đoán loại giá thuê của một căn hộ mới. Các quy tắc này được học từ tập dữ liệu bao gồm các căn hộ đã biết giá thuê của chúng.

4.4.2. Khai phá dữ liệu

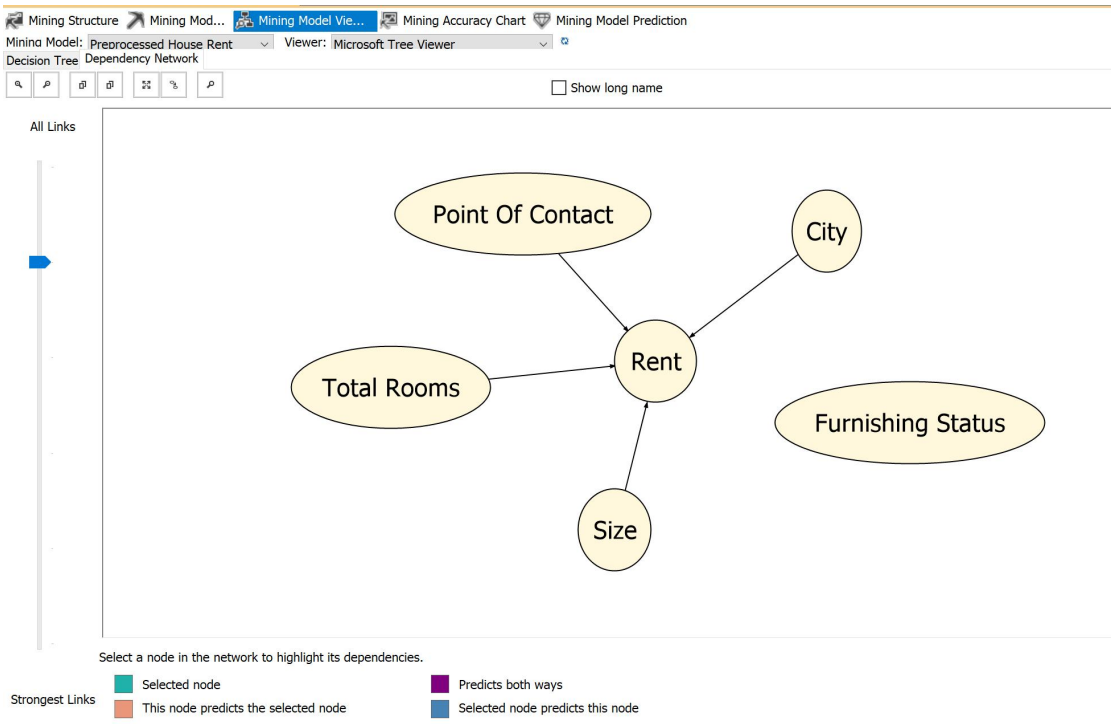
- Khởi tạo Mining Structure chọn thuật toán Microsoft Decision Tree để thực hiện phân loại.
- Qua quá trình trực quan hóa dữ liệu, chọn ra 5 biến đầu vào cho mô hình là: City, Furnishing Status, Size, Point Of Contact, Total Rooms (Number of Bedrooms-Halls-Kitchens + Bathroom).
- Thực hiện lựa chọn biến kết quả cho mô hình là: Rent.



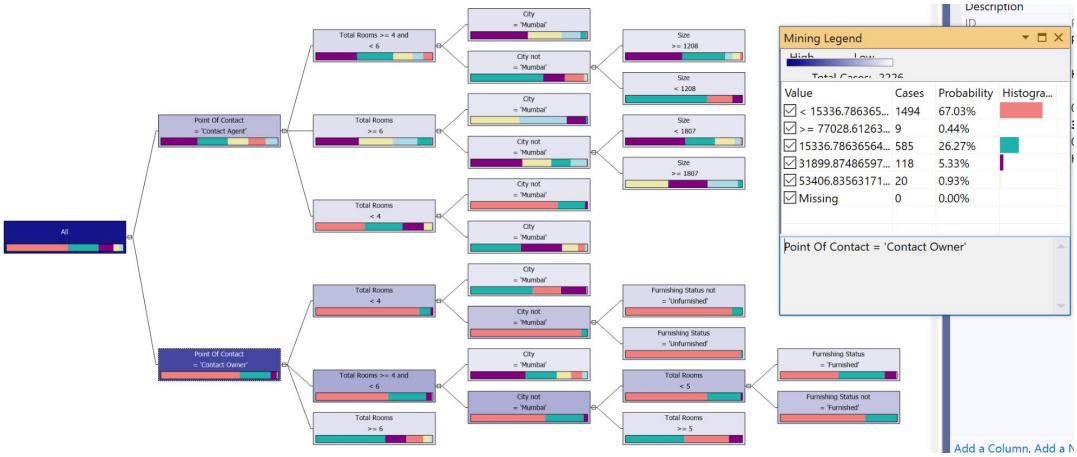
- Bật tab Mining Model Viewer → Dependency Network.



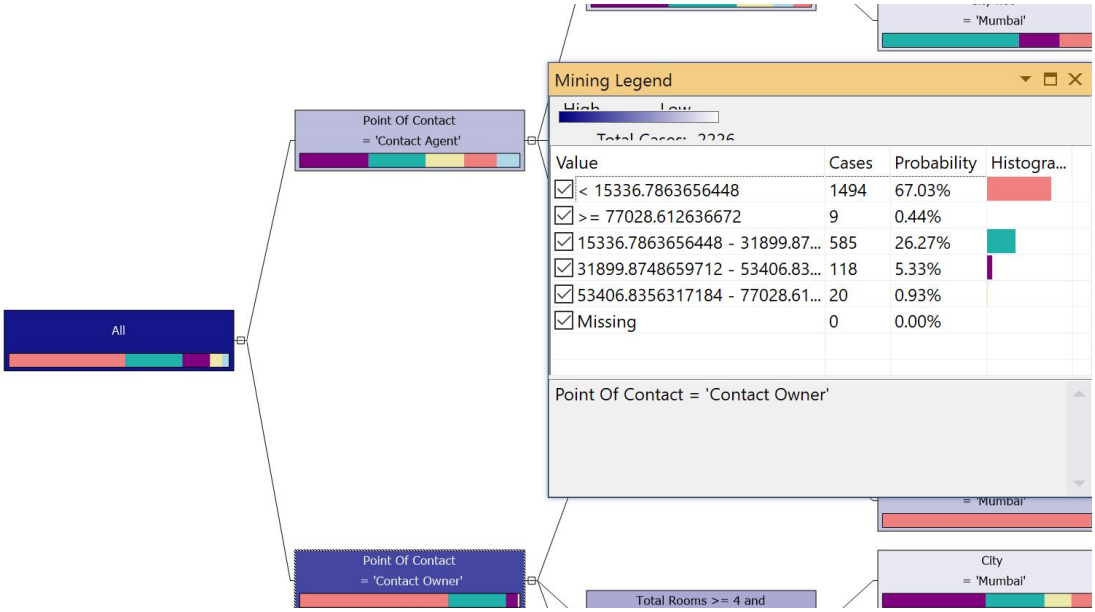
- Ta thấy biến đầu vào có liên kết mạnh nhất với biến dự đoán là 'City'. Kéo lên trên để xem mức độ liên kết của từng biến đầu vào với biến dự đoán.



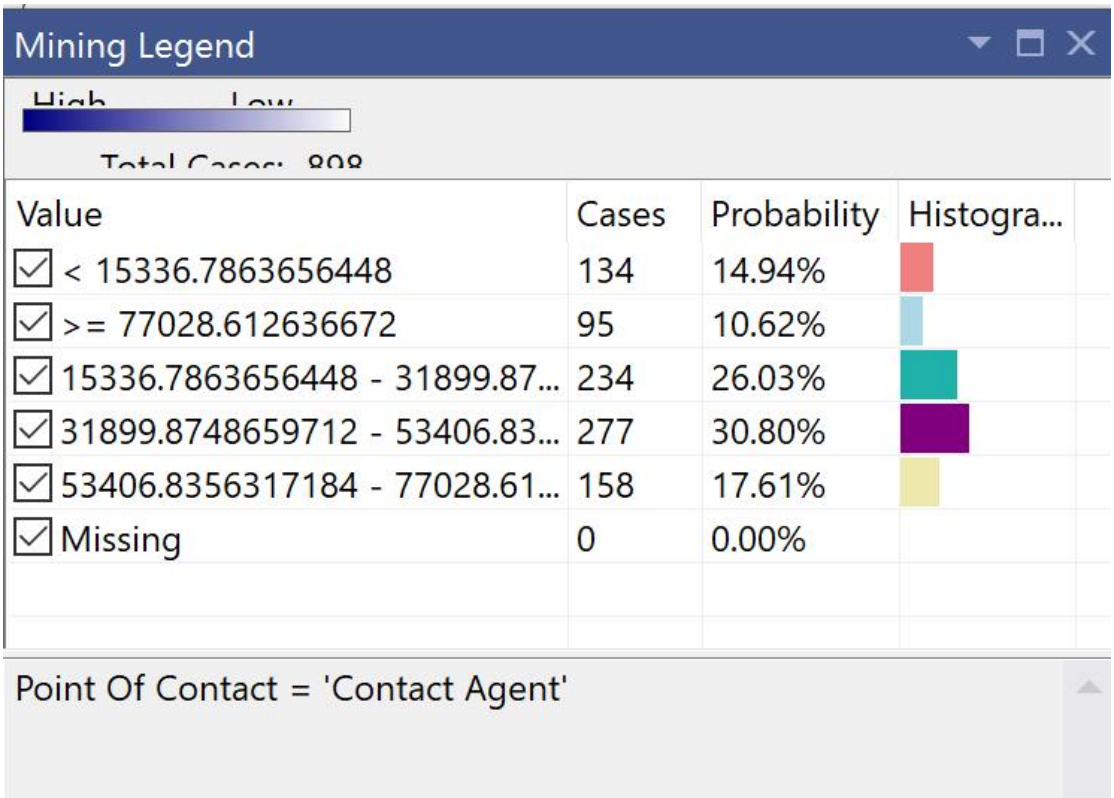
- Bật tab Mining Model Viewer → Decision Tree để xem cây phân loại.
- Decision tree gồm có 6 level.



- Level 1: Point Of Contact:



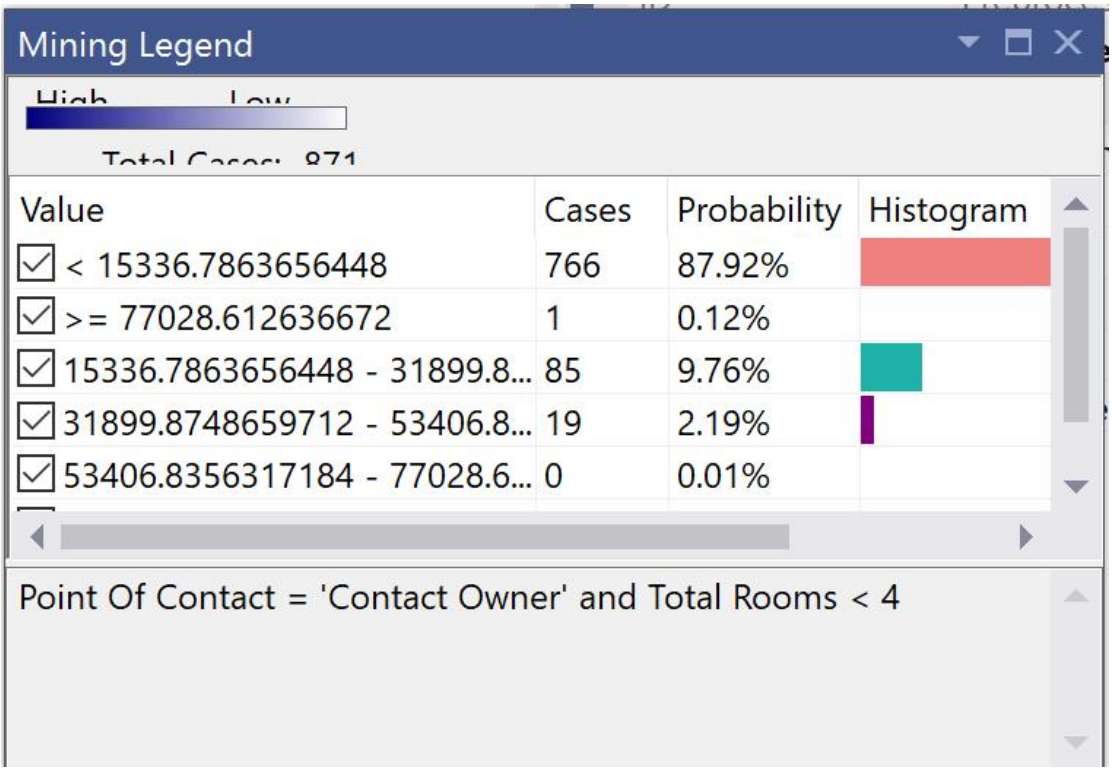
+ Nhà được cho thuê trực tiếp bởi chủ nhà sẽ có giá thuê rẻ hơn so với nhà được cho thuê bởi công ty cho thuê. Hơn 67% căn nhà được cho thuê bởi chủ nhà có giá dưới 15.400, phần còn lại có giá từ 15.400-32.000 là chủ yếu.



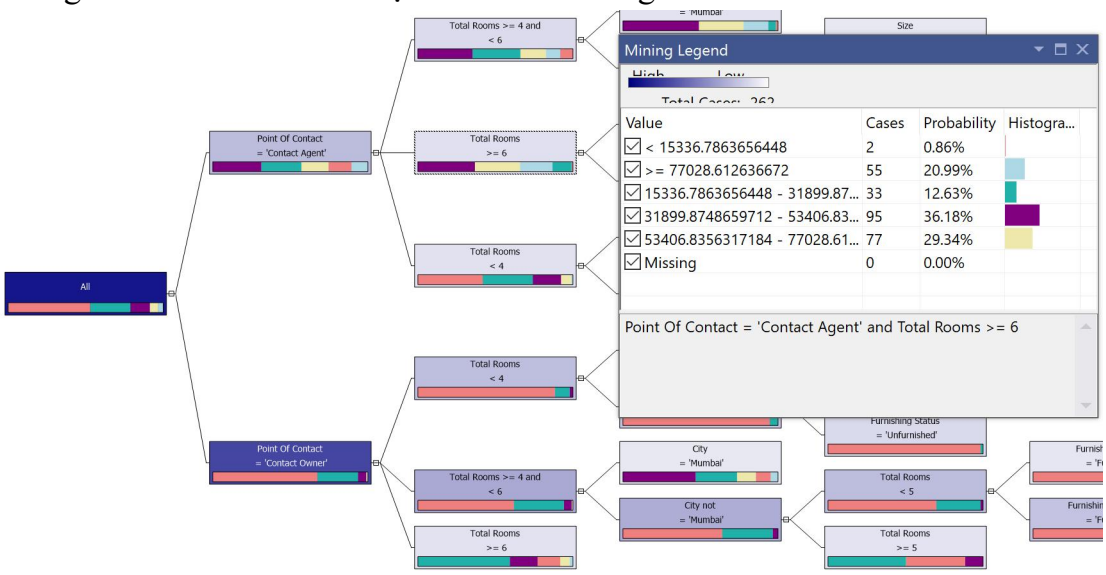
+ Ngược lại, nhà được cho thuê bởi công ty cho thuê sẽ có giá đắt hơn, đa phần sẽ có giá trên 15.300 và có hơn 48% số nhà do công ty cho thuê có giá thuê lớn hơn 31900.

  Kết luận: Nhà được cho thuê bởi chính chủ nhà sẽ có giá thuê rẻ hơn so với nhà được cho thuê bởi công ty cho thuê.

- Level 2: Total Rooms:



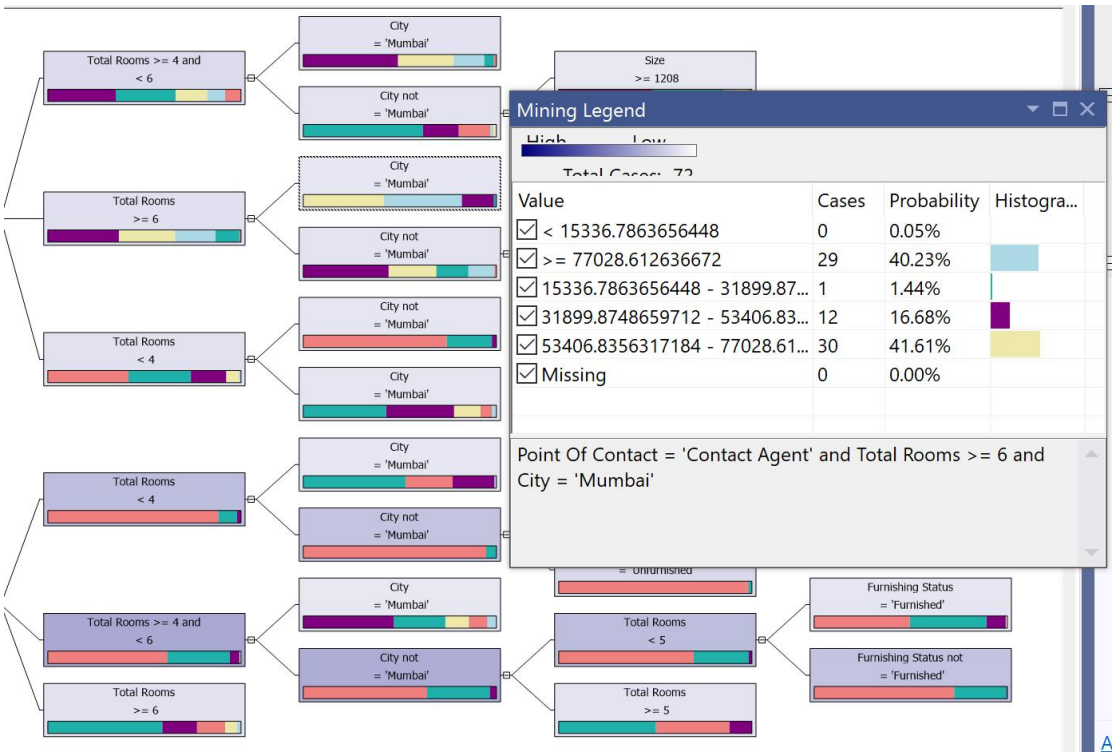
+ Những ngôi nhà có số phòng ít (dưới 4 phòng) sẽ có giá thuê thấp hơn những ngôi nhà có nhiều phòng. Có đến 87.92% căn nhà do chủ nhà cho thuê và có ít phòng đều có giá cho thuê dưới 15.400 và hầu như không có căn nhà nào có 2 đặc tính kể trên có giá thuê trên 54.000.



+ Ngược lại, những ngôi nhà được cho thuê bởi công ty cho thuê và có nhiều phòng (>=6) là những ngôi nhà có giá thuê đắt nhất, hơn 80% có giá thuê trên 32.000, 50% có giá trên 53.000, trong đó có đến 20.99% có giá thuê trên 77000.
+ Cùng được cho thuê bởi công ty thuê nhà, xác suất nhà có giá cho thuê cao ở nhóm nhà có trên 6 phòng (màu xanh dương và màu vàng) lớn hơn hẳn so với nhà có từ 4 đến 6 phòng và có có dưới 4 phòng.

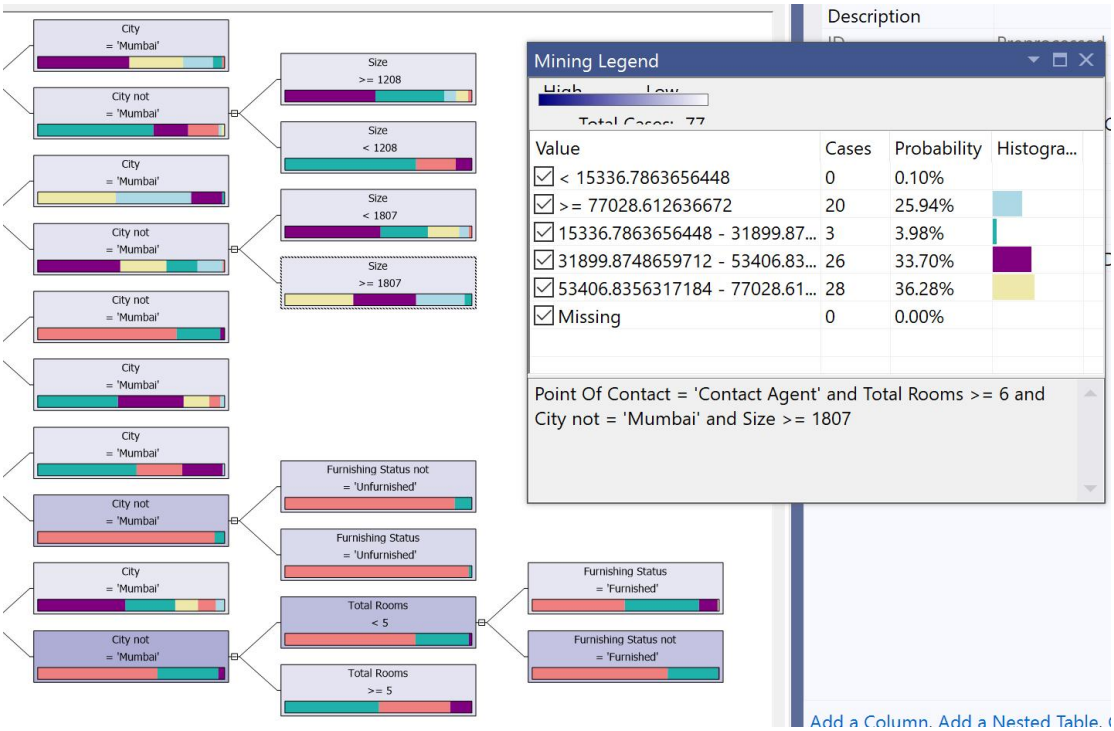
Kết luận: Khi các căn nhà được cho thuê bởi cùng 1 loại bên cho thuê (hoặc là cùng cho thuê bởi chính chủ nhà, hoặc là cùng được cho thuê bởi công ty cho thuê nhà) thì căn nhà nào có số lượng phòng càng nhiều thì giá thuê càng đắt.

- Level 3: City



+ Ở tất cả các nhánh đều thể hiện giá thuê nhà ở thành phố ‘Mumbai ’ đắt hơn so với những thành phố còn lại. Xác suất giá thuê nhà trên 53400 ở nhánh những căn nhà do công ty cho thuê, có số phòng trên 6 và nằm ở thành phố Mumbai là rất lớn (82%).
  Kết luận: Thành phố Mumbai là thành phố có giá thuê nhà đắt đỏ nhất trong các thành phố.

- Level 4: Size



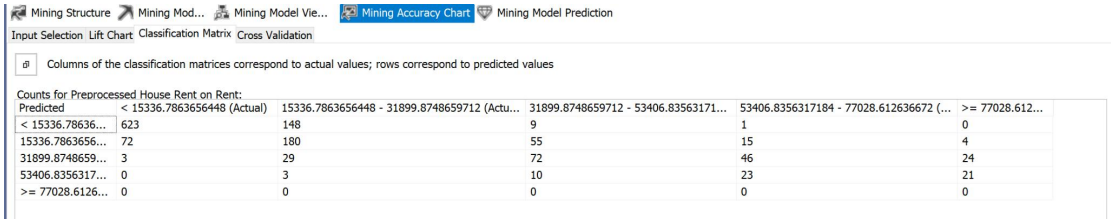
+ Những căn nhà có diện tích lớn hơn 1807 feet, được cho thuê bởi công ty môi giới và có trên 6 phòng không có căn nào có giá thuê dưới 15400. Đa số chúng đều trên 32000.
+ Trong khi đó, dù đều không nằm trong thành phố Mumbai, những căn nhà có diện tích 1208-1807 feet sẽ có giá cho thuê từ 15000 đến 53000, còn những căn có diện tích dưới 1208 feet sẽ có giá thuê rẻ hơn nhiều, chỉ dưới 31900.

  Kết luận: Những căn nhà có diện tích càng lớn thì giá thuê càng cao.

- Level 5 và level 6:
+ Những căn nhà không trang bị nội thất thì sẽ có giá thuê rẻ hơn những ngôi nhà có trang bị nội thất.

4.4.3. Đánh giá mô hình Decision Tree

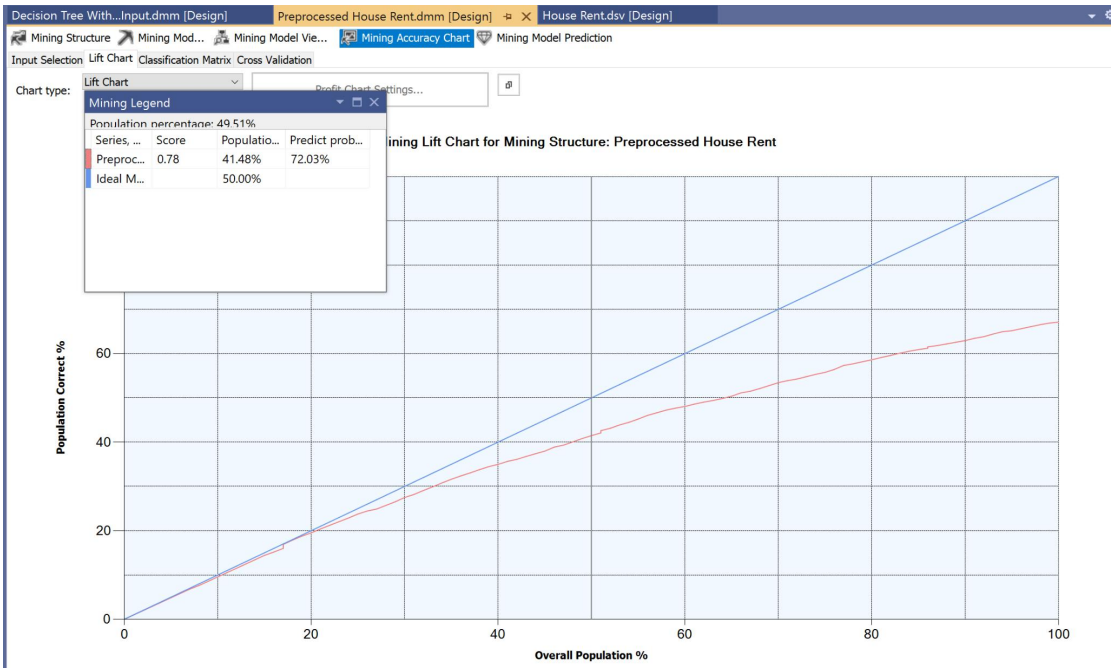
- Chọn tab Mining Accuracy Chart, quan sát Classification Matrix



Predicted	< 15336.7863656448 (Actual)	15336.7863656448 - 31899.8748659712 (Actu...	31899.8748659712 - 53406.83563171...	53406.8356317184 - 77028.612636672 (...	>= 77028.612...
< 15336.78636...	623	148	9	1	0
15336.7863656...	72	180	55	15	4
31899.8748659...	3	29	72	46	24
53406.8356317...	0	3	10	23	21
>= 77028.6126...	0	0	0	0	0

+ Tổng số case mô hình đoán đúng là 898 trên tổng số 1338 case
+ Accuracy: 68%

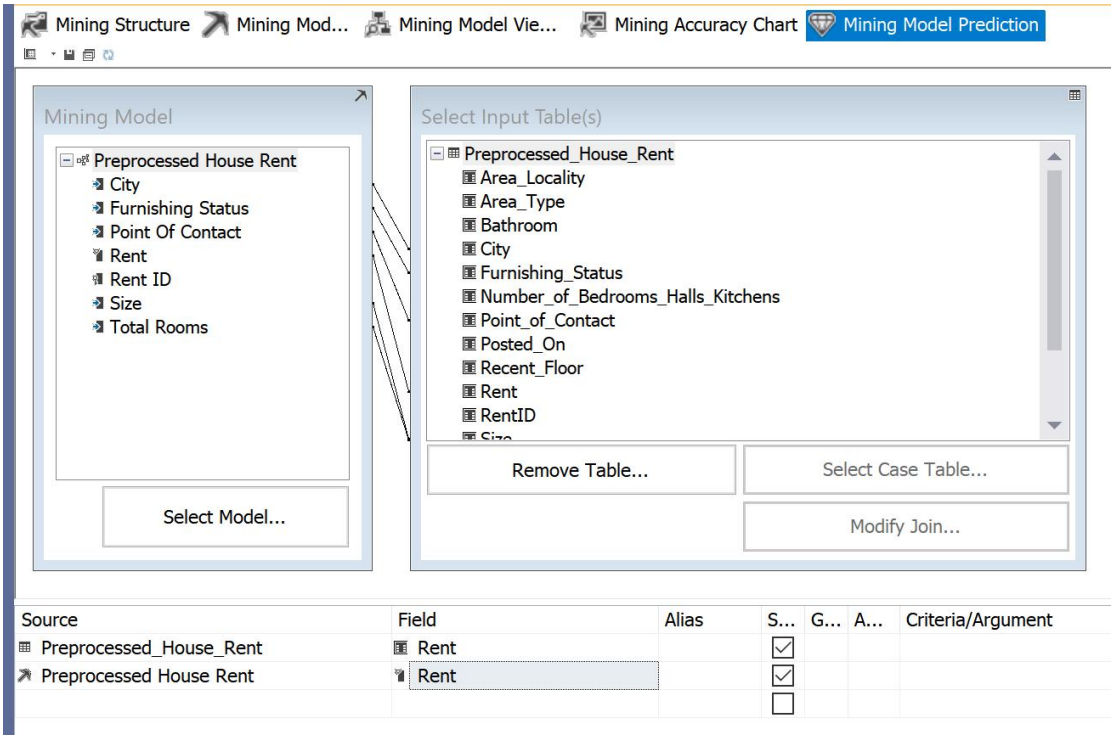
- Score của mô hình dự đoán khoảng giá thuê nhà Decision Tree là 0.78



- Biểu đồ thể hiện tính chính xác của model trong việc dự đoán chính xác giá nhà.

- Mô hình có score là 2.28. Những chấm tròn nằm xung quanh đường thẳng, có thể ứng dụng mô hình trong việc dự đoán khoảng giá thuê nhà dựa vào các đặc tính của ngôi nhà đó.

- Chọn tab Mining Model Prediction để xem giá trị dự đoán của model Decision Tree so với giá thực tế trong tập dữ liệu.



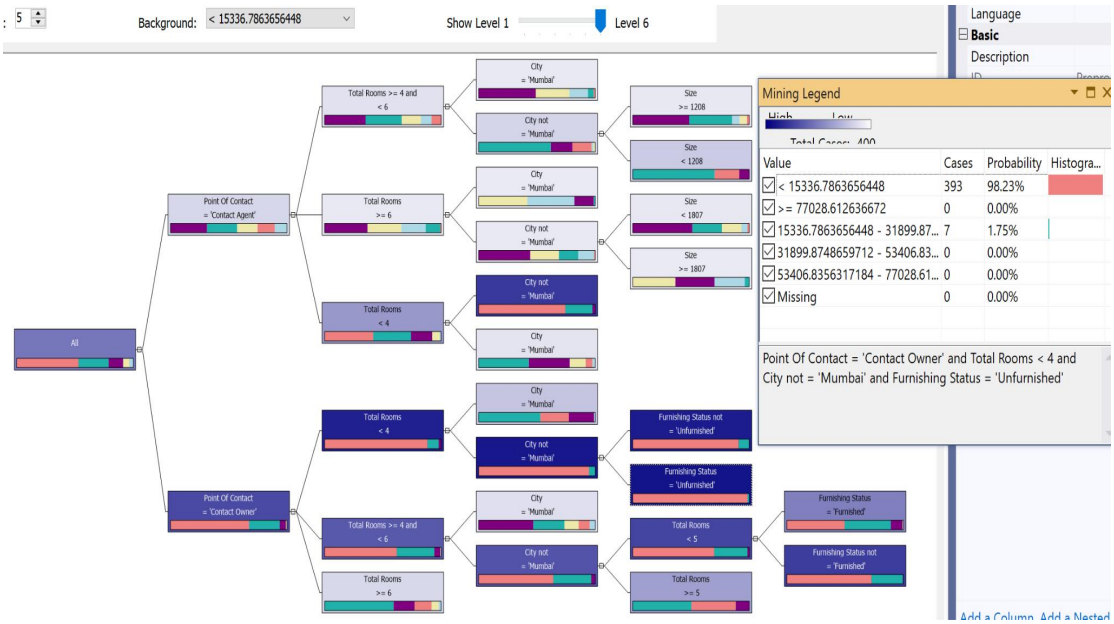
+ Chọn result để xem kết quả dự đoán:

10000	7952.78730232927
7500	8047.81439616426
7000	9449.72986197477
10000	13272.4284769447
5000	5153.40584185182
26000	14476.2572657782
10000	13273.4239311891
25000	24250.372486516
5000	6977.57640828321
6500	5743.35911417979
5500	6673.54798054464
8500	13752.6950567209
40000	26688.9588200023
6000	8773.08668678146
10000	7952.78730232927
11000	10233.437554369
6000	6556.7149769463
7900	6223.30526053453
9000	4760.54149931705
4000	4878.37566825914
6500	18950.5147499708
8000	6250.29895955707
7000	8332.89567766922
5300	6004.70057806372
6000	11309.1215736313
8500	7952.78730232927
12500	11329.940864209
7500	5174.316185853
15000	12021.4338771046
6000	8456.08232592524
5000	7105.13568064591

4.4.4. Ứng dụng mô hình

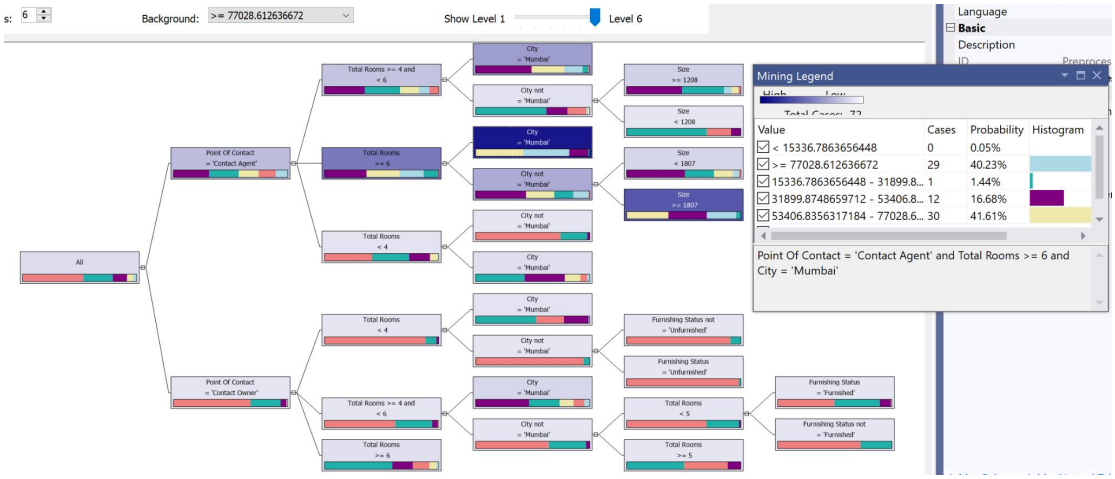
- Nếu đóng vai trò là một người đang tìm thuê nhà đang muốn tìm những căn nhà phù hợp với kinh phí của họ, chọn background là nhóm giá thuê tương ứng với khoảng giá thuê mà họ có thể chi trả mỗi tháng. Ví dụ họ muốn tìm thuê nhà có giá tiền dưới 15400, như vậy cần chọn background dưới

15400.



=> Nếu muốn chọn những căn nhà có giá thuê rẻ (<15400) thì nên lựa chọn những căn nhà dược cho thuê bởi chính chủ nhà, có tổng số phòng <4, không nên lựa chọn thuê nhà ở thành phố Mumbai và nên chọn những căn nhà không bao gồm nội thất.

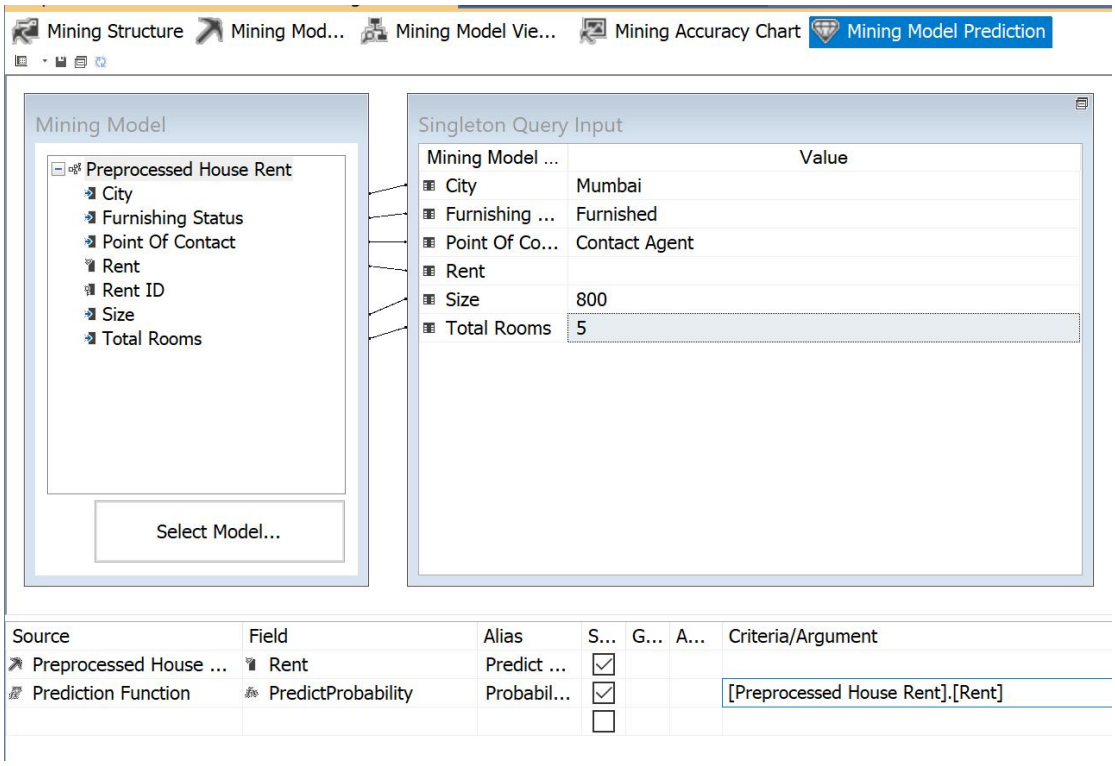
- Nếu đóng vai tròn là chủ đầu tư muốn chọn một nơi để xây nhà cho thuê, họ sẽ cần lựa chọn những yếu tố giúp tăng giá thuê nhà. Ví dụ mức giá họ mong muốn là trên 77000, chọn background là nhóm >=77000, ta thấy:



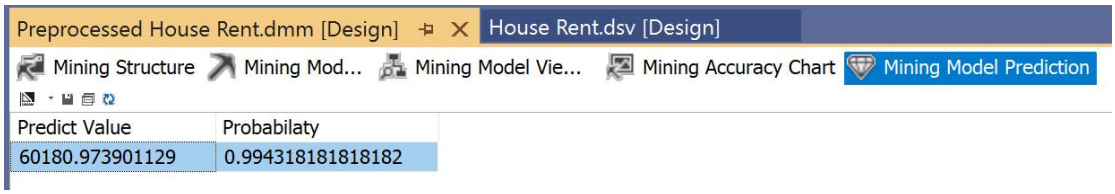
=> Nếu muốn có mức giá cho thuê đó thì chủ đầu tư nên chọn thành phố Mumbai, diện tích căn nhà trên 1807 feet, tuy nhiên cần cân nhắc đến việc giá đất ở Mumbai cũng sẽ khá đắt thì chủ đầu tư nên xây nhiều phòng trong căn nhà cho thuê và nếu có thể thì trực tiếp cho khách hàng thuê chứ không qua công ty môi giới để giảm thiểu chi phí.

- Cũng có thể sử dụng mô hình để dự đoán khoảng giá thuê nhà theo những yêu cầu về các đặc tính căn nhà theo nhu cầu của người đang tìm thuê:

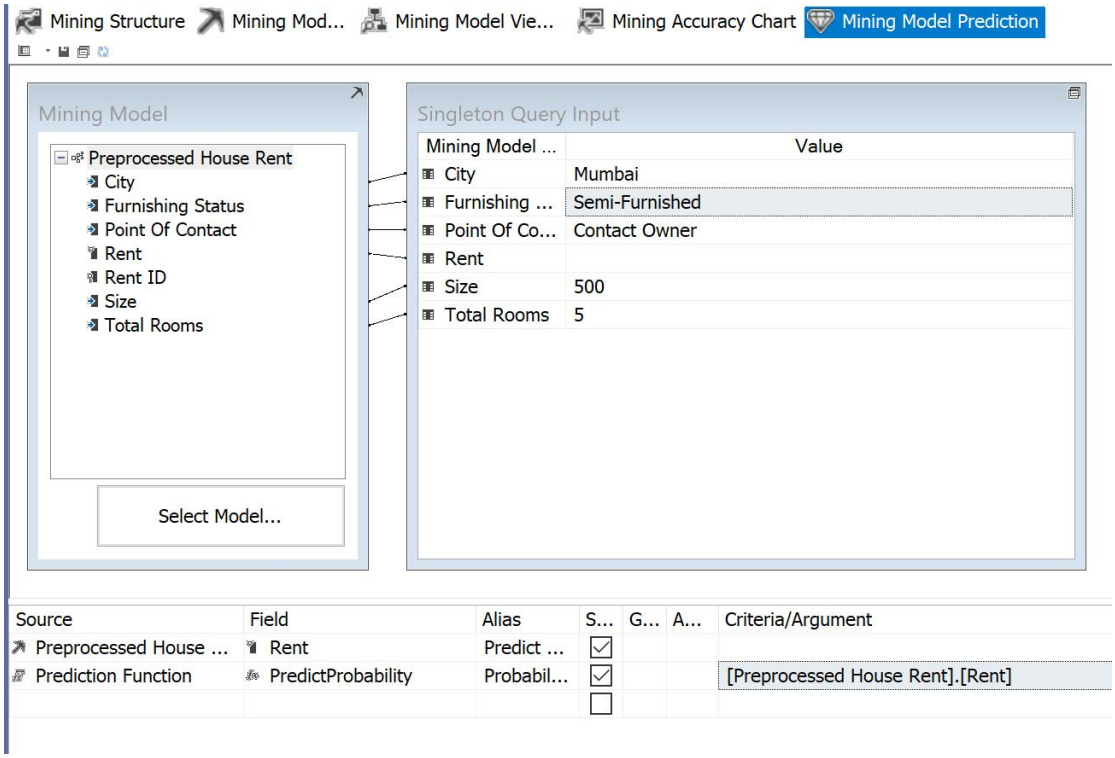
+ Ví dụ: Một người cần thuê 1 căn hộ ở Mumbai. có tình trạng nội thất đầy đủ, thuê trực tiếp từ chủ nhà, có diện tích là 800 feet và có 5 phòng thì người đó nên chuẩn bị tài chính khoảng bao nhiêu.



+ Chọn Result để xem kết quả dự đoán. Với những yêu cầu về căn nhà mà người thuê đang tìm kiếm thì họ sẽ phải chuẩn bị mức giá thuê khoảng 60000, xác suất cao đến 99%.



+ Nếu giá này còn khá cao đối với khả năng của họ thì người thuê có thể hạ tiêu chuẩn thuê xuống một chút, ví dụ họ nên chọn thuê trực tiếp từ chủ nhà thay vì qua công ty môi giới để giảm giá thuê, chọn thuê căn nhà nhỏ hơn và được trang bị một phần nội thất cần thiết thay vì đầy đủ nội thất như ban đầu.



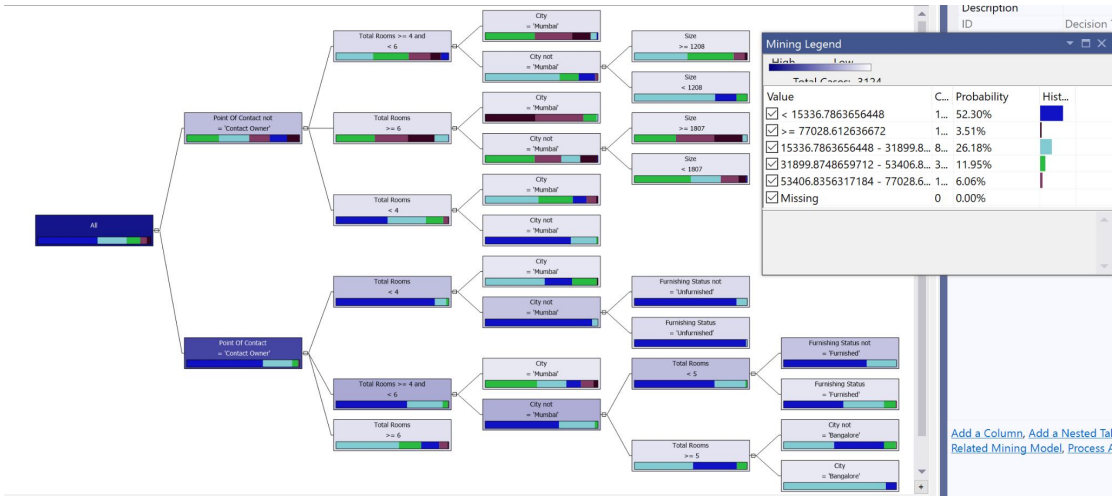
+ Bằng cách này họ đã giảm được chi phí thuê nhà đáng kể, từ 60000 xuống còn 37300. Xác suất căn nhà có các đặc tính mà họ yêu cầu được định giá 37300 là 97%.

4.4.5. Khai phá dữ liệu với tập biến đầu vào thay đổi

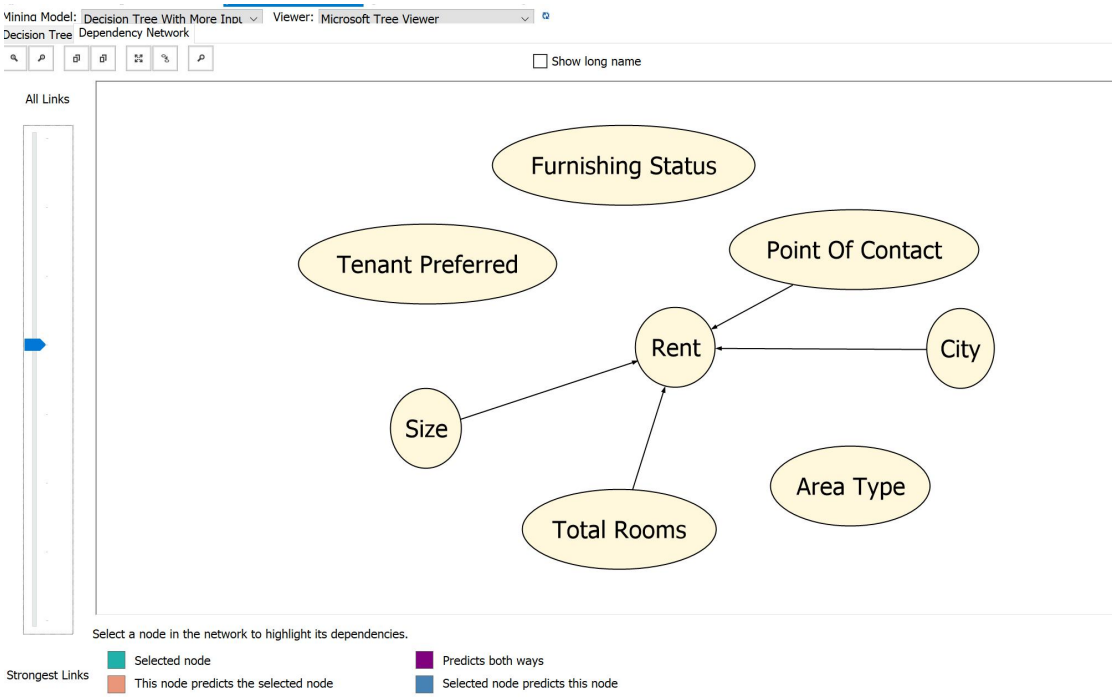
4.4.5.1. Khi thêm thuộc tính đầu vào

- Cũng với các bước khai phá và đánh giá như trên nhưng thêm một biến đầu vào ‘Area Type’ và ‘Tenant Preferred’ vào mô hình.

Structure	Decision Tree With More Input
	Microsoft_Decision_Trees
Area Type	Input
City	Input
Furnishing Status	Input
Point Of Contact	Input
Recent Floor	Ignore
Rent ID	Key
Rent	PredictOnly
Size	Input
Tenant Preferred	Input
Total Floors	Ignore
Total Rooms	Input

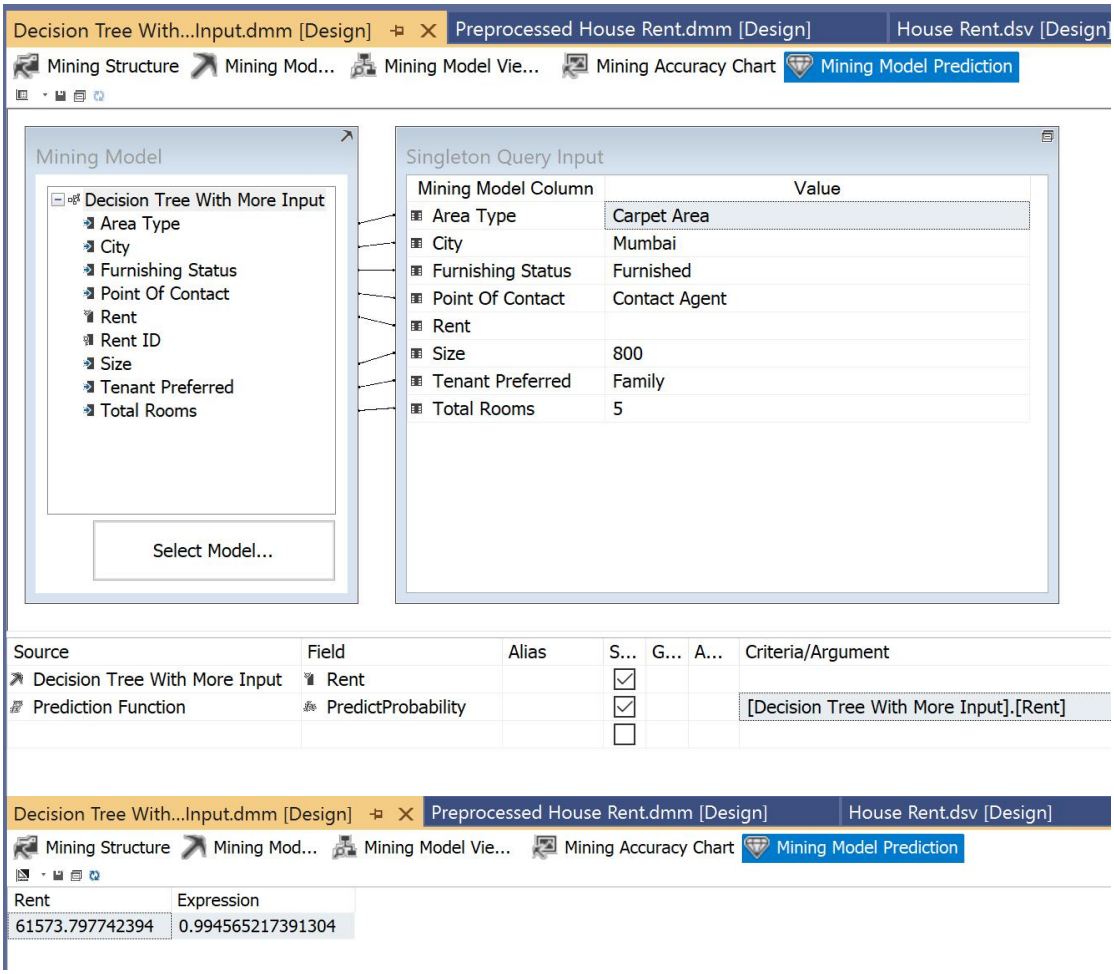


- Ta thấy cây quyết định cũng không có quá nhiều thay đổi, các quy luật rút ra được so với tập 5 biến đầu vào ban đầu cũng không thay đổi, chỉ có xác suất của các khoảng giá thuê trong từng nhánh là có thay đổi nhưng không đáng kể.

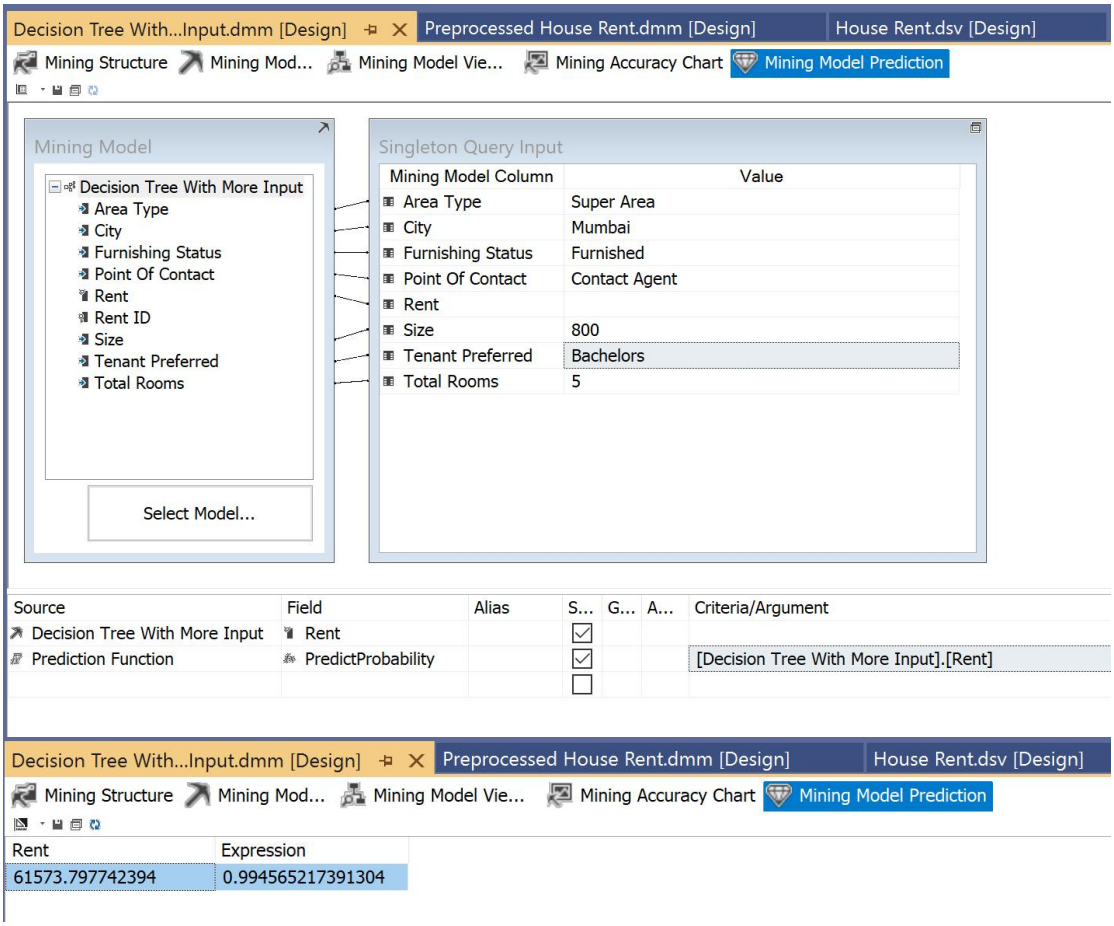


- Trong Dependency Network, mức độ liên kết của hai biến ‘Tenant Preferred’ và ‘Area Type’ cũng không quá mạnh.

- Khi thực hiện dự đoán giá nhà: Giữ nguyên các giá trị của 5 Input ban đầu, chỉ thay đổi giá trị của biến ‘ Area Type’ và ‘Tenant Preferred’, ta thấy giá thuê nhà thay đổi không đáng kể


Decision Tree With...Input.dmm [Design]
Preprocessed House Rent.dmm [Design]
House Rent.dsv [Design]
Mining Structure
Mining Mod...
Mining Model Vie...
Mining Accuracy Chart
Mining Model Prediction
Mining Model
Decision Tree With More Input
Area Type
City
Furnishing Status
Point Of Contact
Rent
Rent ID
Size
Tenant Preferred
Total Rooms
Select Model...
Singleton Query Input
Mining Model Column
Value
Area Type
Carpet Area
City
Mumbai
Furnishing Status
Furnished
Point Of Contact
Contact Agent
Rent
Size
800
Tenant Preferred
Family
Total Rooms
5
Source
Field
Alias
S...
G...
A...
Criteria/Argument
Decision Tree With More Input
Rent
Prediction Function
PredictProbability
[Decision Tree With More Input].[Rent]

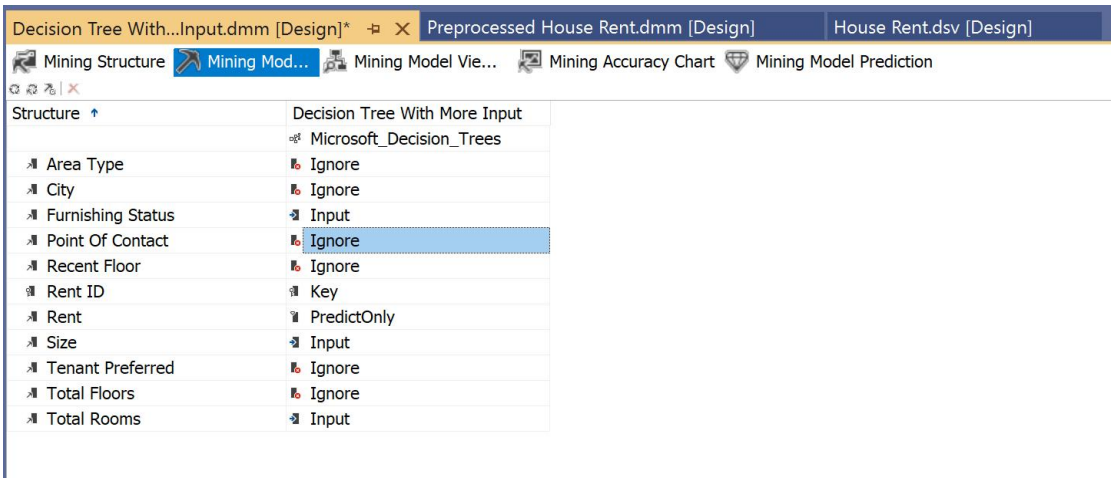
Decision Tree With...Input.dmm [Design]
Preprocessed House Rent.dmm [Design]
House Rent.dsv [Design]
Mining Structure
Mining Mod...
Mining Model Vie...
Mining Accuracy Chart
Mining Model Prediction
Rent
Expression
61573.797742394
0.994565217391304



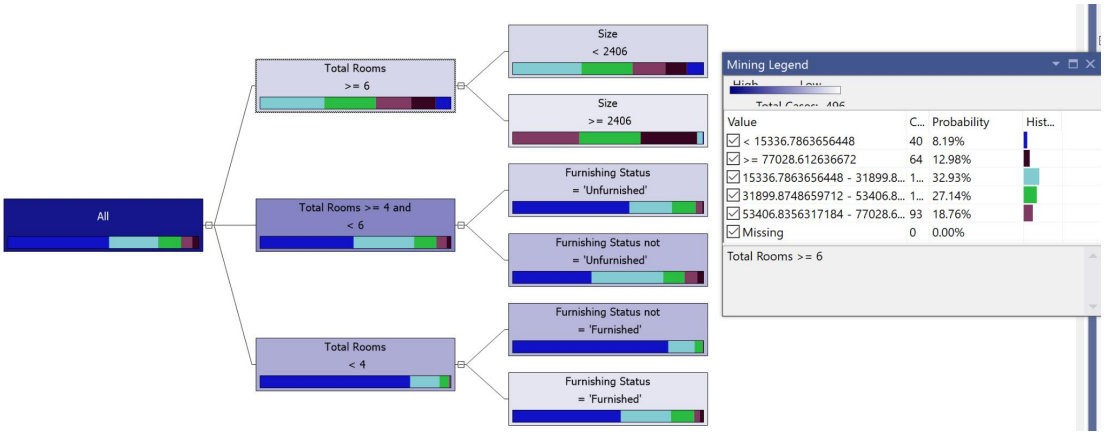
=> Hai biến đầu vào vừa thêm vào không ảnh hưởng quá nhiều đến mô hình dự đoán.

4.4.5.2. Khi bớt thuộc tính đầu vào

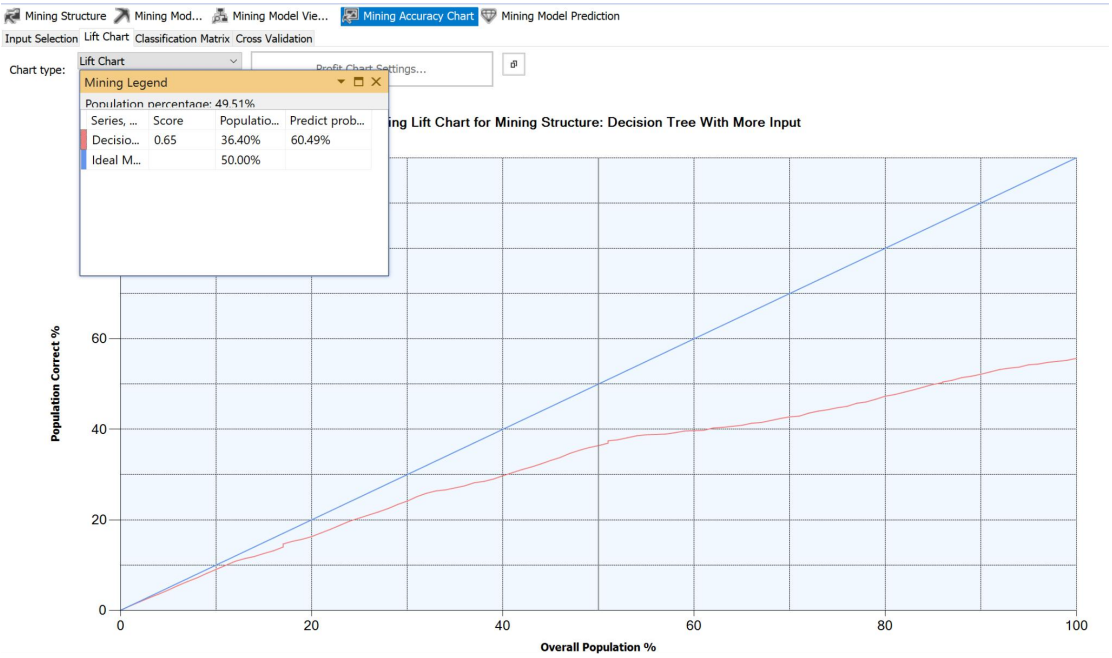
- Với cùng phương pháp khai phá và đánh giá như trên nhưng lần này chỉ sử dụng 3 biến đầu vào là ‘Furnishing Status’, ‘Size’, ‘Total Rooms’



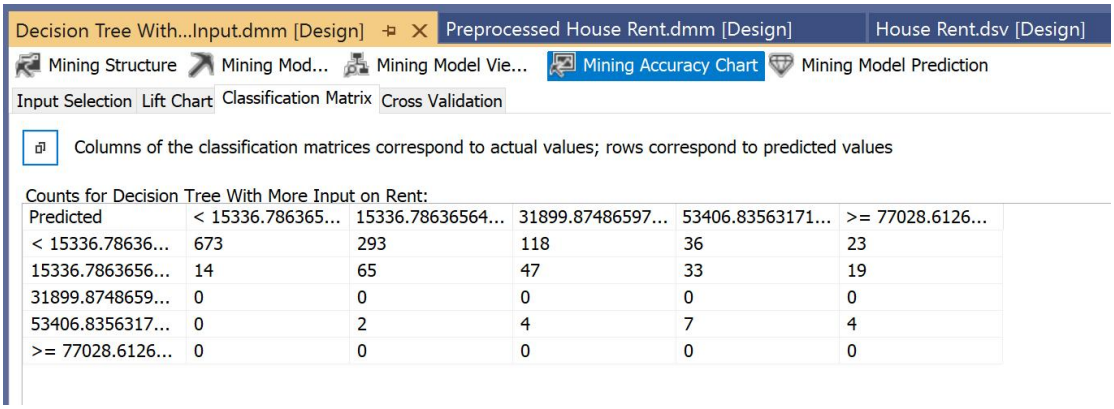
- Cây quyết định chỉ còn lại 3 level



- Vẫn với các quy luật cũ: Số phòng càng nhiều thì giá thuê càng cao, diện tích càng lớn thì giá thuê cao, những căn nhà được trang bị đầy đủ nội thất thì giá thuê cao hơn những căn nhà chưa có nội thất.
- Tuy nhiên độ chính xác của mô hình giảm đi đáng kể, score của mô hình chỉ còn 0.65 thay vì 0.78 như cách chọn 5 biến đầu vào như ban đầu.



- Các trường hợp mô hình dự đoán đúng mức giá cũng không cao, chỉ đúng 745/1338 trường hợp, accuracy khoảng 55%, thấp hơn so với cách chọn 5 biến đầu vào (68%). Mô hình cho quá nhiều kết quả nhà có giá thuê dưới 15300.



Predicted	< 15336.786365...	15336.78636564...	31899.87486597...	53406.83563171...	>= 77028.6126...
< 15336.78636...	673	293	118	36	23
15336.7863656...	14	65	47	33	19
31899.8748659...	0	0	0	0	0
53406.8356317...	0	2	4	7	4
>= 77028.6126...	0	0	0	0	0

=> Nếu bỏ bớt đi các biến đầu vào có ảnh hưởng mạnh đến biến dự đoán thì kết quả dự đoán sẽ bị giảm đi độ chính xác.

CHƯƠNG 5: ĐÁNH GIÁ VÀ SO SÁNH BA THUẬT TOÁN

5.1. Classification

- Là lựa chọn tối ưu nhất trong 3 thuật toán nếu người dùng muốn phân loại căn nhà với những đặc điểm mà họ mong muốn sẽ thuộc vào khoảng giá thuê từ bao nhiêu đến bao nhiêu, từ đó chuẩn bị ngân sách cho việc thuê nhà hoặc các phương án đầu tư phù hợp.
- Người dùng cũng có thể lựa chọn mức giá thuê nhà mà mình mong muốn và chiếu theo các đặc điểm mà mô hình đưa ra để chọn căn nhà phù hợp với kinh phí của mình trong thực tế.
- Đối với tập dữ liệu 'Giá nhà cho thuê ở Ấn Độ' và thuật toán được chọn để dự đoán là Decision Tree thì tập các biến đầu vào nên lựa chọn là 'Point To Contact'- Bên cho thuê nhà, 'Total Rooms'- Tổng số phòng, 'City'- Thành phố, 'Size'- Diện tích của ngôi nhà, 'Furnishing Status'- Tình trạng nội thất.

5.2. Clustering

- Phù hợp với người dùng muốn có cái nhìn tổng quát cả về giá thuê lẫn các đặc trưng tương ứng với từng mức giá thuê khác nhau. Từ đó sẽ giúp ích cho việc quản lý và phân tích thị trường cho thuê nhà và giúp đưa các quyết định kinh doanh như giá cả, chiến lược quảng cáo,... với từng nhóm nhà. Ví dụ như đối với những nhóm nhà thuê có diện tích nhỏ đến vừa, giá thuê rẻ thì ta có thể tập trung quảng cáo cho đối tượng là sinh viên. Còn với những nhóm căn hộ cao cấp hơn với diện tích lớn thì có thể xây dựng chiến lược quảng bá đến đối tượng là những người có thu nhập cao.
- Khi dự đoán được căn nhà của người dùng thuộc cụm nào thì có thể suy ra các đặc tính còn lại không được đề cập của nó dựa vào đặc tính chung của cụm.

5.3. Association Rule

- Việc áp dụng khai phá luật kết hợp vào bài toán 'House Rent' ở Ấn Độ đã tìm ra được các mối quan hệ và kết nối giữa các tiêu chí như diện tích, thành phố, số phòng ngủ, số phòng tắm,... với giá cả. Từ các quan hệ giữa các tiêu chí đó thì có thể đưa ra những lời khuyên cho khách hàng

về những lựa chọn phù hợp khi tìm kiếm nhà cho thuê dựa trên các kết quả phân tích. Tuy nhiên vì tập dữ liệu còn giới hạn nên có thể các luật sinh ra chưa phản ánh được đầy đủ cũng như sự khác biệt giữa các mối liên hệ với các mức giá thuê khác nhau. Đồng thời qua quá trình đánh giá và phân tích nhóm nhận thấy rằng thuật toán này không thích hợp để dự đoán giá thuê vì nó chỉ dự đoán đúng khi nó xuất hiện trong các luật kết hợp còn với những trường hợp khác với tập dữ liệu thì nó sẽ không đem lại kết quả chính xác. Vì bản chất thuật toán này thích hợp với những bài toán cho 'phân tích giỏ hàng' nên nó sẽ không đem lại kết quả thật sự tối ưu.

CHƯƠNG 6: KẾT LUẬN

6.1. Kết quả đạt được

- Thực hiện khai phá dữ liệu thông qua các giải thuật được cung cấp bởi Microsoft Analysis Services Multidimensional and Data Mining Project như giải thuật Microsoft Association Rules, Microsoft Clustering và Microsoft Decision Trees để khai phá *'Dữ liệu và dự đoán giá thuê nhà ở Ấn Độ'* giúp chúng em hiểu sâu hơn về kiến thức môn học Khai phá dữ liệu và thành thạo hơn trong việc sử dụng các công cụ hỗ trợ quá trình tiền xử lý dữ liệu như ngôn ngữ python và các thư viện hỗ trợ như matplotlib, seaborns, pandas, os,..., công cụ để trực quan hóa dữ liệu như Power BI Desktop, matplotlib.pyplot. Nâng cao kỹ năng nhìn nhận và tiếp thu, xử lý thông tin đầu vào, đặt những câu hỏi tiền đề và thảo luận hướng đi để giải quyết và trả lời câu hỏi.

6.2. Những hạn chế

- Trong quá trình thực hiện chọn giải thuật và khai phá dữ liệu, nhóm chúng em không tránh khỏi gặp một số những khó khăn vì còn hạn chế về mặt kiến thức và tính phù hợp của tập dữ liệu. Tập dữ liệu còn nhiều dữ kiện chưa được khai thác hết. Những thực hiện của chúng em chưa đạt đến mức nâng cao trong tổng thể quá trình khai phá dữ liệu một cách hoàn chỉnh.
- Mô hình dự đoán trong SSAS sẽ không thể dự đoán những trường hợp không xuất hiện trong tập dữ liệu training.

6.3. Tầm nhìn và hướng cải tiến mô hình

- Để tăng độ chính xác và tính ứng dụng của các mô hình thì cần phải tăng tính đa dạng của tập dữ liệu. Tăng lên về số dòng và đa dạng về số trường hợp có thể xảy ra, bổ sung thêm dữ liệu của nhiều thành phố trên Ấn Độ và nhiều thuộc tính khác ảnh hưởng đến giá thuê nhà vào tập dữ liệu như địa điểm có gần trường học, bệnh viện, trung tâm mua sắm hay không, thuộc khu vực ngoại thành hay nội thành,..
- Tùy theo từng nhu cầu và từng vai trò mà lựa chọn thuật toán phù hợp.
- Đối với tập dữ liệu 'Giá thuê nhà ở Ấn Độ' này thì nên chọn những thuộc tính có mức độ ảnh hưởng lớn tới giá thuê nhà như 'Point of contact', 'City', 'Total Rooms', 'Size', 'Furnishing Status'.

6.4. Tài liệu tham khảo

[1] Dinesh Asanka (2020), *Association Rule Mining in SQL Server.* [online] sqlshack.com. Available at: https://www.sqlshack.com/the-association-rule-mining-in-sql-server/. [Accessed 17 May 2023].

[2] Nguyễn Văn Chúc (no date). *Building Classification model with Decision Tree Using SSAS*. [online] youtube.com. Available at: https://www.youtube.com/watch?v=o8tEEBy5zjQ&ab_channel=ChucNguyenVan. [Accessed 21 May 2023].

[3] Dinesh Asanka (2019), *Microsoft Decision Trees in SQL Server.* [online] sqlshack.com. Available at: https://www.sqlshack.com/microsoft-decision-trees-in-sql-server/. [Accessed 18 May 2023].

[4] Nguyễn Văn Chúc (no date). *Building Clustering model with SSAS*. [online] youtube.com. Available at: https://www.youtube.com/watch?v=gBxCexfKDFo&t=434s&ab_channel=ChucNguyenVan. [Accessed 21 May 2023].

[5] Dinesh Asanka (2020), *Microsoft Clustering in SQL Server.* [online] sqlshack.com. Available at: https://www.sqlshack.com/microsoft-clustering-in-sql-server/. [Accessed 18 May 2023].