

Báo Cáo Trực Quan Hoá Dữ Liệu

Team 04

Giảng viên phụ trách: Nguyễn Thanh Bình
Sinh viên thực hiện: Đinh Anh Huy - 18110103
Nguyễn Đức Vũ Duy - 18110004

Mô tả dữ liệu

File dữ liệu đầu vào: *7_Daegu_Real_Estate_data.csv*

Nội dung

Bộ dữ liệu nói về các giao dịch của những căn hộ được tạo từ tháng 8/2007 đến tháng 8/2017 tại phường Daebong, thành phố Daegu, Hàn Quốc.

| | SalePrice | YearBuilt | YrSold | MonthSold | Size(sqf) | Floor | HallwayType | HeatingType | AptManageType | N_Parkinglot(Ground) | ... |
|---|-----------|-----------|--------|-----------|-----------|-------|-------------|--------------------|---------------------|----------------------|-----|
| 0 | 141592 | 2006 | 2007 | 8 | 814 | 3 | terraced | individual_heating | management_in_trust | 111.0 | ... |
| 1 | 51327 | 1985 | 2007 | 8 | 587 | 8 | corridor | individual_heating | self_management | 80.0 | ... |
| 2 | 48672 | 1985 | 2007 | 8 | 587 | 6 | corridor | individual_heating | self_management | 80.0 | ... |
| 3 | 380530 | 2006 | 2007 | 8 | 2056 | 8 | terraced | individual_heating | management_in_trust | 249.0 | ... |
| 4 | 221238 | 1993 | 2007 | 8 | 1761 | 3 | mixed | individual_heating | management_in_trust | 523.0 | ... |

5 rows × 30 columns

Thông tin các cột của dữ liệu

- SalePrice: Giá bán tính bằng đơn vị US dollar (target feature).
- YearBuilt: Năm xây dựng.
- YrSold: Năm giao dịch.
- Size(sqf): Kích thước căn hộ tính bằng đơn vị square feet.
- N_Parkinglot(Ground): Số chỗ đỗ xe ở tầng trệt.
- N_Parkinglot(Basement): Số chỗ đỗ xe ở tầng hầm.
- TimeToBusStop: Thời gian di chuyển từ căn hộ đến trạm xe buýt..
- TimeToSubway: Thời gian di chuyển từ căn hộ đến ga tàu điện ngầm.
- N_APT: Số căn hộ trong khu chung cư.
- N_FacilitiesInApt: Số lượng tiện ích cho người ở chung cư.
- N_manager: Số người quản lý các tiện ích tại căn hộ.

Một số thông tin cơ bản của dữ liệu

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5891 entries, 0 to 5890
Data columns (total 30 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   SalePrice                                5891 non-null   int64
1   YearBuilt                                5891 non-null   int64
2   YrSold                                   5891 non-null   int64
3   MonthSold                               5891 non-null   int64
4   Size(sqft)                              5891 non-null   int64
5   Floor                                    5891 non-null   int64
6   HallwayType                             5891 non-null   object
7   HeatingType                             5891 non-null   object
8   AptManagementType                       5891 non-null   object
9   N_Parkinglot(Ground)                    5891 non-null   float64
10  N_Parkinglot(Basement)                   5891 non-null   float64
11  TimeToBusStop                             5891 non-null   object
12  TimeToSubway                             5891 non-null   object
13  N_APT                                     5891 non-null   float64
14  N_manager                               5891 non-null   float64
15  N_elevators                             5891 non-null   float64
16  SubwayStation                           5891 non-null   object
17  N_FacilitiesNearBy(PublicOffice)         5891 non-null   float64
18  N_FacilitiesNearBy(Hospital)             5891 non-null   int64
19  N_FacilitiesNearBy(Dpartmentstore)       5891 non-null   float64
20  N_FacilitiesNearBy(Mall)                 5891 non-null   float64
21  N_FacilitiesNearBy(ETC)                  5891 non-null   float64
22  N_FacilitiesNearBy(Park)                 5891 non-null   float64
23  N_SchoolNearBy(Elementary)               5891 non-null   float64
24  N_SchoolNearBy(Middle)                   5891 non-null   float64
25  N_SchoolNearBy(High)                     5891 non-null   float64
26  N_SchoolNearBy(University)               5891 non-null   float64
27  N_FacilitiesInApt                        5891 non-null   int64
28  N_FacilitiesNearBy(Total)                5891 non-null   float64
29  N_SchoolNearBy(Total)                    5891 non-null   float64
dtypes: float64(16), int64(8), object(6)
memory usage: 1.3+ MB
```

Một số thông tin cơ bản của dữ liệu

Mô tả dữ liệu dạng số:

| | count | mean | std | min | 25% | 50% | 75% | max |
|------------------------------------|--------|---------------|---------------|---------|----------|----------|----------|----------|
| SalePrice | 5891.0 | 221218.112545 | 106384.186446 | 32743.0 | 144247.0 | 207964.0 | 291150.0 | 585840.0 |
| YearBuilt | 5891.0 | 2002.967408 | 8.811782 | 1978.0 | 1993.0 | 2006.0 | 2008.0 | 2015.0 |
| YrSold | 5891.0 | 2012.691563 | 2.905488 | 2007.0 | 2010.0 | 2013.0 | 2015.0 | 2017.0 |
| MonthSold | 5891.0 | 6.160244 | 3.387752 | 1.0 | 3.0 | 6.0 | 9.0 | 12.0 |
| Size(sqf) | 5891.0 | 955.569173 | 382.464050 | 135.0 | 644.0 | 910.0 | 1149.0 | 2337.0 |
| Floor | 5891.0 | 12.026311 | 7.548743 | 1.0 | 6.0 | 11.0 | 17.0 | 43.0 |
| N_Parkinglot(Ground) | 5891.0 | 195.883551 | 218.597210 | 0.0 | 11.0 | 100.0 | 249.0 | 713.0 |
| N_Parkinglot(Basement) | 5891.0 | 570.761670 | 408.621075 | 0.0 | 184.0 | 536.0 | 798.0 | 1321.0 |
| N_APT | 5891.0 | 5.613648 | 2.811831 | 1.0 | 3.0 | 7.0 | 8.0 | 13.0 |
| N_manager | 5891.0 | 6.310304 | 3.174088 | 1.0 | 5.0 | 6.0 | 8.0 | 14.0 |
| N_elevators | 5891.0 | 11.145985 | 7.806232 | 0.0 | 5.0 | 11.0 | 16.0 | 27.0 |
| N_FacilitiesNearBy(PublicOffice) | 5891.0 | 4.141572 | 1.793496 | 0.0 | 3.0 | 5.0 | 5.0 | 7.0 |
| N_FacilitiesNearBy(Hospital) | 5891.0 | 1.296045 | 0.479760 | 0.0 | 1.0 | 1.0 | 2.0 | 2.0 |
| N_FacilitiesNearBy(Dpartmentstore) | 5891.0 | 0.896282 | 0.810964 | 0.0 | 0.0 | 1.0 | 2.0 | 2.0 |
| N_FacilitiesNearBy(Mall) | 5891.0 | 0.941436 | 0.401355 | 0.0 | 1.0 | 1.0 | 1.0 | 2.0 |
| N_FacilitiesNearBy(ETC) | 5891.0 | 1.941266 | 2.201392 | 0.0 | 0.0 | 1.0 | 5.0 | 5.0 |
| N_FacilitiesNearBy(Park) | 5891.0 | 0.654218 | 0.658320 | 0.0 | 0.0 | 1.0 | 1.0 | 2.0 |
| N_SchoolNearBy(Elementary) | 5891.0 | 3.022407 | 0.954575 | 0.0 | 2.0 | 3.0 | 4.0 | 6.0 |
| N_SchoolNearBy(Middle) | 5891.0 | 2.417756 | 1.037898 | 0.0 | 2.0 | 3.0 | 3.0 | 4.0 |
| N_SchoolNearBy(High) | 5891.0 | 2.659311 | 1.556041 | 0.0 | 1.0 | 2.0 | 4.0 | 5.0 |
| N_SchoolNearBy(University) | 5891.0 | 2.764726 | 1.489289 | 0.0 | 2.0 | 2.0 | 4.0 | 5.0 |
| N_FacilitiesInApt | 5891.0 | 5.809540 | 2.330804 | 1.0 | 4.0 | 5.0 | 7.0 | 10.0 |
| N_FacilitiesNearBy(Total) | 5891.0 | 9.870820 | 3.450319 | 0.0 | 8.0 | 9.0 | 13.0 | 16.0 |
| N_SchoolNearBy(Total) | 5891.0 | 10.864200 | 4.438513 | 0.0 | 7.0 | 10.0 | 15.0 | 17.0 |

Một số thông tin cơ bản của dữ liệu

Mô tả dữ liệu dạng object:

| | count | unique | top | freq |
|----------------------|-------|--------|-----------------------|------|
| HallwayType | 5891 | 3 | terraced | 3564 |
| HeatingType | 5891 | 2 | individual_heating | 5591 |
| AptManageType | 5891 | 2 | management_in_trust | 5542 |
| TimeToBusStop | 5891 | 3 | 0~5min | 4509 |
| TimeToSubway | 5891 | 5 | 0-5min | 2759 |
| SubwayStation | 5891 | 8 | Kyungbuk_uni_hospital | 1644 |

Một số thông tin cơ bản của dữ liệu

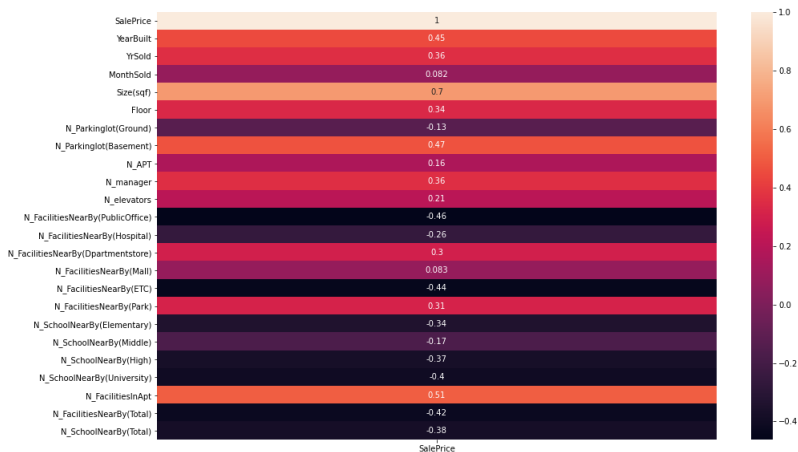
Kiểm tra missing values:

| | Missing values | % Missing |
|------------------------------------|----------------|-----------|
| SalePrice | 0 | 0.0 |
| YearBuilt | 0 | 0.0 |
| YrSold | 0 | 0.0 |
| MonthSold | 0 | 0.0 |
| Size(sqf) | 0 | 0.0 |
| Floor | 0 | 0.0 |
| HallwayType | 0 | 0.0 |
| HeatingType | 0 | 0.0 |
| AptManageType | 0 | 0.0 |
| N_Parkinglot(Ground) | 0 | 0.0 |
| N_Parkinglot(Basement) | 0 | 0.0 |
| TimeToBusStop | 0 | 0.0 |
| TimeToSubway | 0 | 0.0 |
| N_APT | 0 | 0.0 |
| N_manager | 0 | 0.0 |
| N_elevators | 0 | 0.0 |
| SubwayStation | 0 | 0.0 |
| N_FacilitiesNearBy(PublicOffice) | 0 | 0.0 |
| N_FacilitiesNearBy(Hospital) | 0 | 0.0 |
| N_FacilitiesNearBy(Dpartmentstore) | 0 | 0.0 |
| N_FacilitiesNearBy(Mall) | 0 | 0.0 |
| N_FacilitiesNearBy(ETC) | 0 | 0.0 |
| N_FacilitiesNearBy(Park) | 0 | 0.0 |
| N_SchoolNearBy(Elementary) | 0 | 0.0 |
| N_SchoolNearBy(Middle) | 0 | 0.0 |
| N_SchoolNearBy(High) | 0 | 0.0 |
| N_SchoolNearBy(University) | 0 | 0.0 |
| N_FacilitiesInApt | 0 | 0.0 |
| N_FacilitiesNearBy(Total) | 0 | 0.0 |
| N_SchoolNearBy(Total) | 0 | 0.0 |

Ta chọn ra những features quan trọng để trực quan: *SalePrice*, *YearBuilt*, *YrSold*, *TimeToBusStop*, *TimeToSubway*, *N_FacilitiesInApt*.

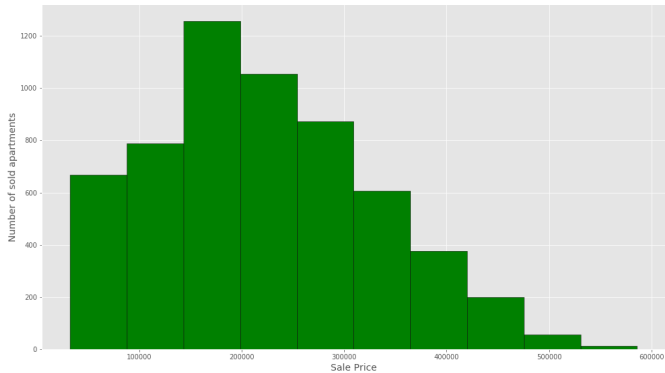
Phân tích dữ liệu

Heatmap mô tả sự tương quan giữa các numerical features với target feature (SalePrice).



Thống kê và vẽ các biểu đồ biểu diễn các features quan trọng

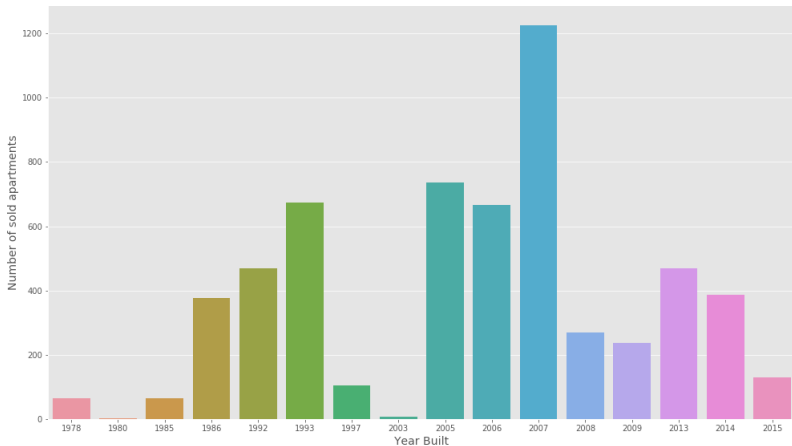
SalePrice



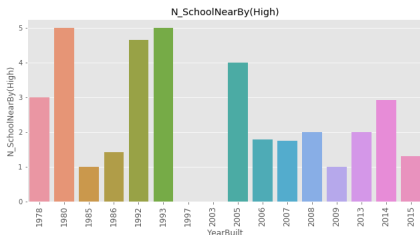
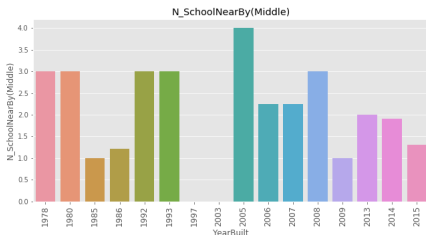
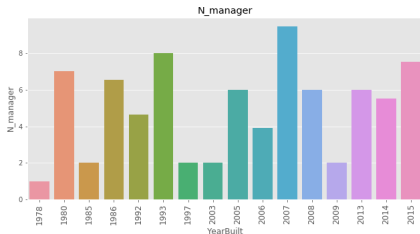
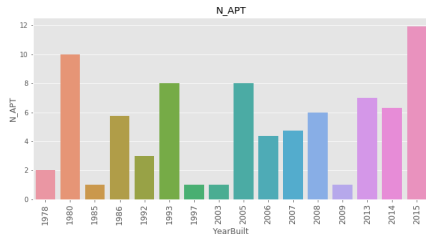
Thống kê số lượng căn hộ giao dịch theo từng năm xây dựng.

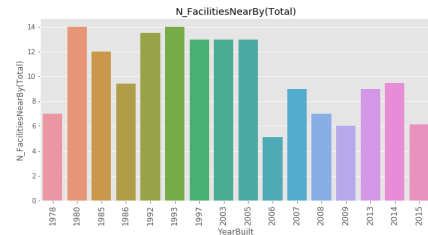
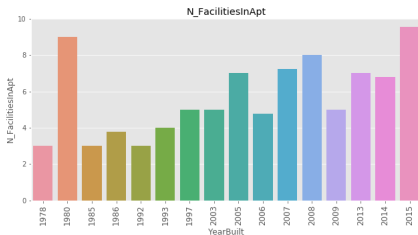
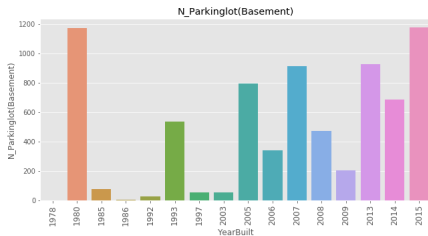
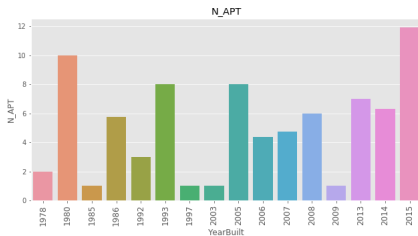
| | YearBuilt | Number of Apt |
|----|-----------|---------------|
| 0 | 1978 | 65 |
| 1 | 1980 | 4 |
| 2 | 1985 | 65 |
| 3 | 1986 | 378 |
| 4 | 1992 | 469 |
| 5 | 1993 | 674 |
| 6 | 1997 | 106 |
| 7 | 2003 | 9 |
| 8 | 2005 | 737 |
| 9 | 2006 | 665 |
| 10 | 2007 | 1225 |
| 11 | 2008 | 270 |
| 12 | 2009 | 238 |
| 13 | 2013 | 468 |
| 14 | 2014 | 387 |

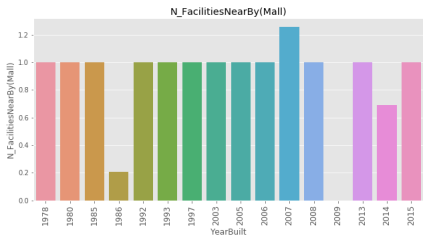
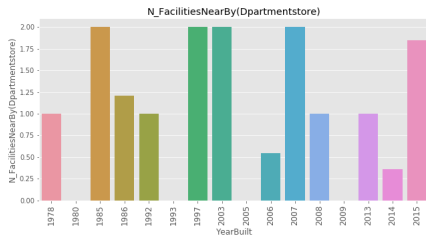
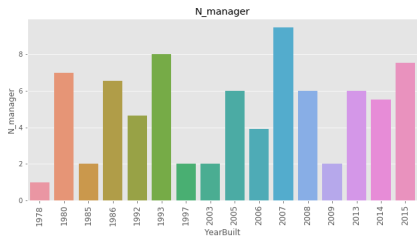
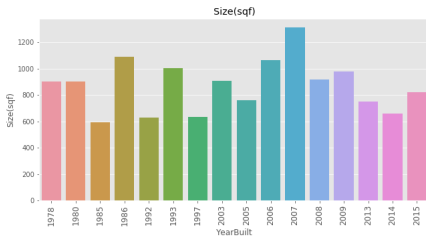
YearBuilt



Để giải thích cho những vấn đề trên, ta sẽ vẽ thêm một số biểu đồ để có cái nhìn bao quát hơn.

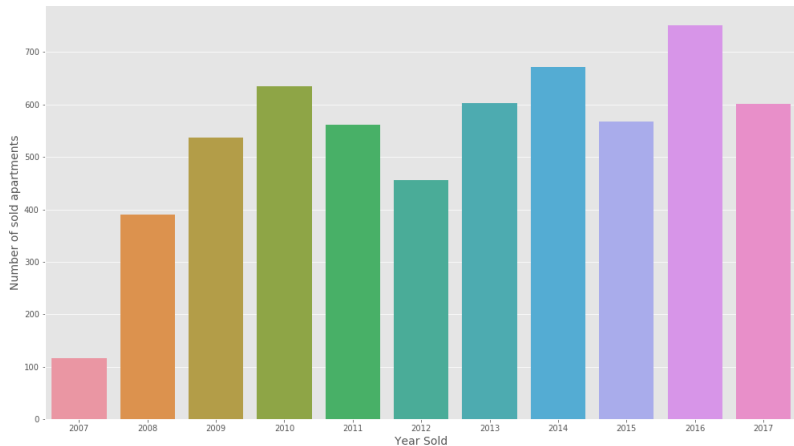






Thống kê số lượng căn hộ giao dịch tính theo năm giao dịch.

| | YrSold | Number of Apt |
|----|--------|---------------|
| 0 | 2007 | 117 |
| 1 | 2008 | 390 |
| 2 | 2009 | 537 |
| 3 | 2010 | 634 |
| 4 | 2011 | 562 |
| 5 | 2012 | 456 |
| 6 | 2013 | 603 |
| 7 | 2014 | 672 |
| 8 | 2015 | 568 |
| 9 | 2016 | 751 |
| 10 | 2017 | 601 |

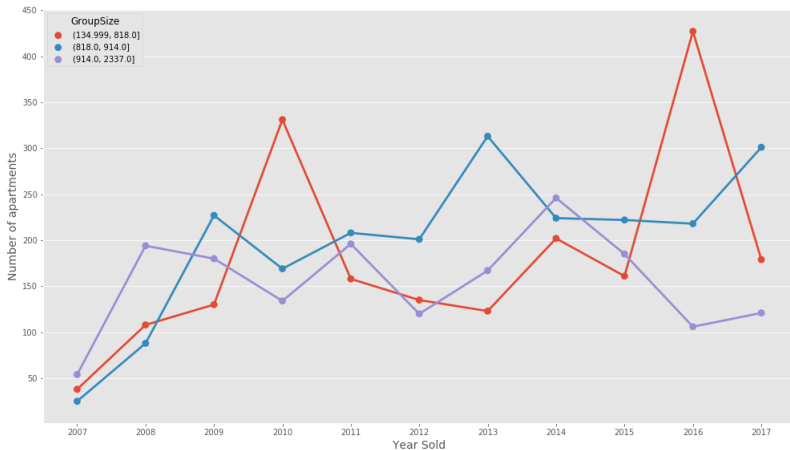


Size(sqf)

Thống kê số lượng căn hộ giao dịch theo kích thước và năm giao dịch.

| | YrSold | GroupSize | Number of Apt |
|---|--------|------------------|---------------|
| 0 | 2007 | (134.999, 818.0] | 38 |
| 1 | 2007 | (818.0, 914.0] | 25 |
| 2 | 2007 | (914.0, 2337.0] | 54 |
| 3 | 2008 | (134.999, 818.0] | 108 |
| 4 | 2008 | (818.0, 914.0] | 88 |

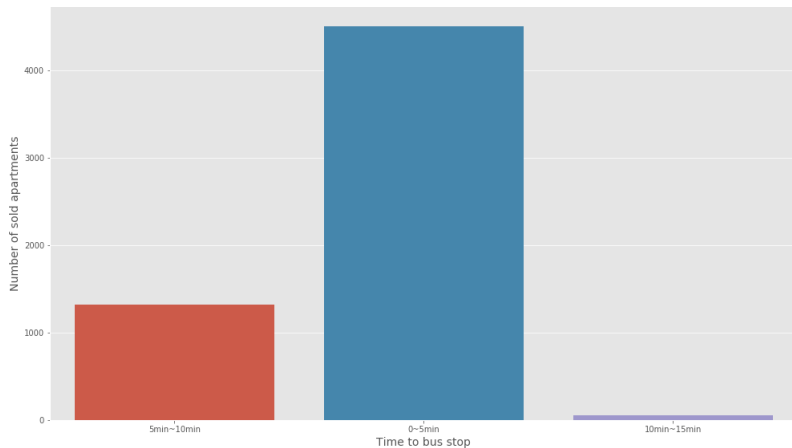
Size(sqf)



Thống kê số lượng căn hộ giao dịch theo thời gian đến các trạm xe buýt.

| | TimeToBusStop | Number of Apt |
|---|---------------|---------------|
| 0 | 0~5min | 4509 |
| 1 | 10min~15min | 55 |
| 2 | 5min~10min | 1327 |

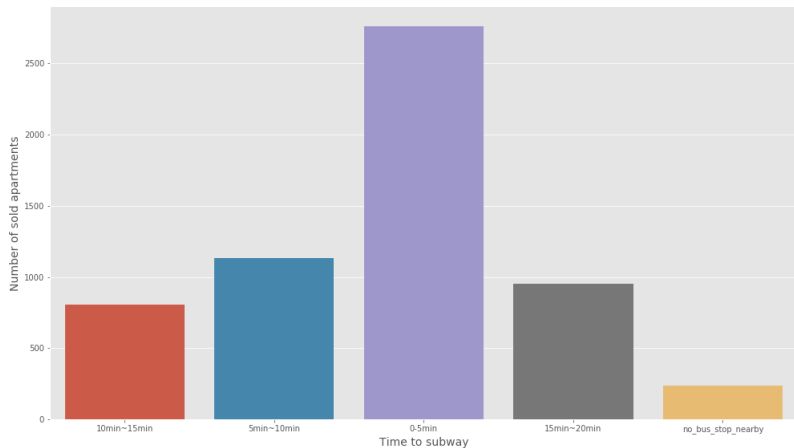
TimeToBusStop



Thống kê số lượng căn hộ giao dịch theo thời gian đến các trạm tàu điện ngầm.

| | TimeToSubway | Number of Apt |
|---|--------------------|---------------|
| 0 | 0-5min | 2759 |
| 1 | 10min~15min | 806 |
| 2 | 15min~20min | 953 |
| 3 | 5min~10min | 1135 |
| 4 | no_bus_stop_nearby | 238 |

TimeToSubway



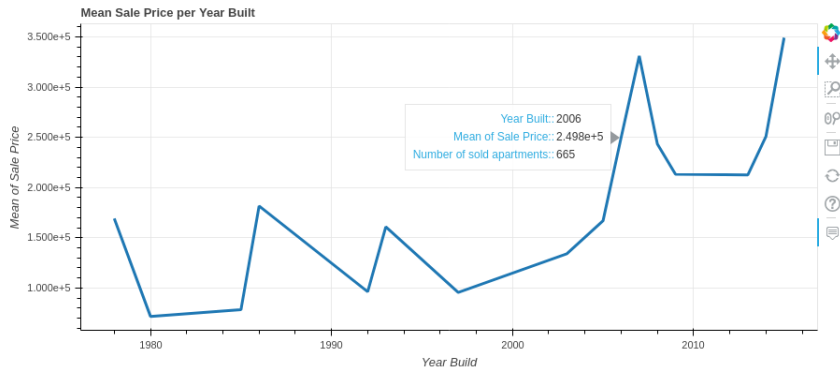
Thống kê và vẽ các biểu đồ biểu diễn mối quan hệ giữa các features

YearuBuilt và Mean of SalePrice

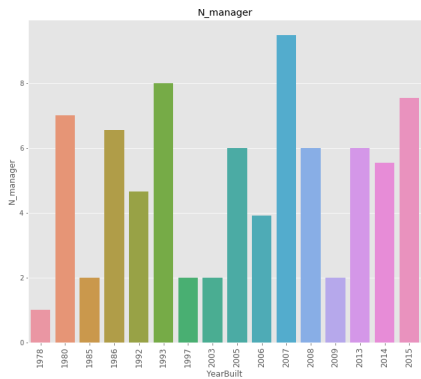
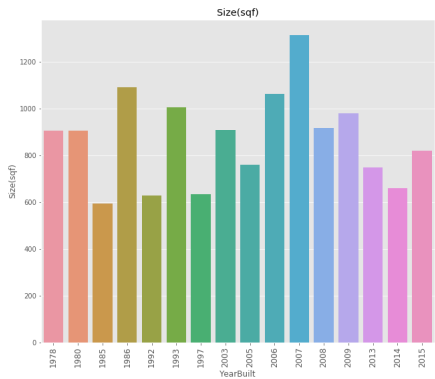
Thống kê số lượng và giá bán trung bình của các căn hộ theo từng năm xây dựng.

| | YearBuilt | MeanSalePrice | NoSoldApt |
|----|-----------|---------------|-----------|
| 0 | 1978 | 169039.615385 | 65 |
| 1 | 1980 | 71681.000000 | 4 |
| 2 | 1985 | 78401.092308 | 65 |
| 3 | 1986 | 181543.283069 | 378 |
| 4 | 1992 | 96215.933902 | 469 |
| 5 | 1993 | 160855.087537 | 674 |
| 6 | 1997 | 95416.122642 | 106 |
| 7 | 2003 | 134021.111111 | 9 |
| 8 | 2005 | 166852.907734 | 737 |
| 9 | 2006 | 249809.592481 | 665 |
| 10 | 2007 | 330971.122449 | 1225 |
| 11 | 2008 | 243339.081481 | 270 |
| 12 | 2009 | 213169.630252 | 238 |
| 13 | 2013 | 212606.918803 | 468 |
| 14 | 2014 | 250990.095607 | 387 |
| 15 | 2015 | 349141.549618 | 131 |

YearuBuilt và Mean of SalePrice



YearuBuilt và Mean of SalePrice

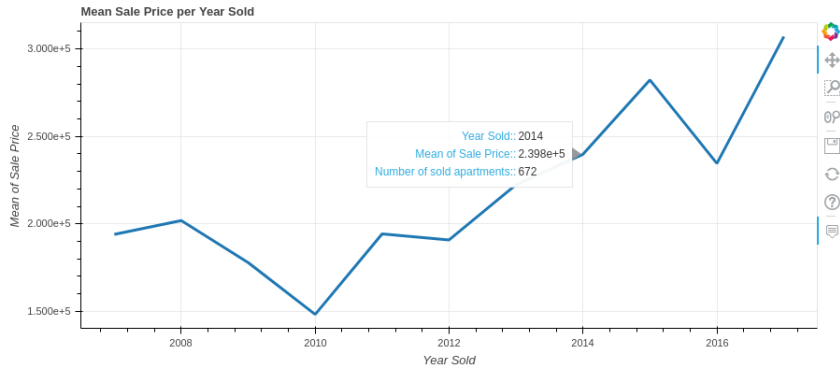


YrSold và Mean of SalePrice

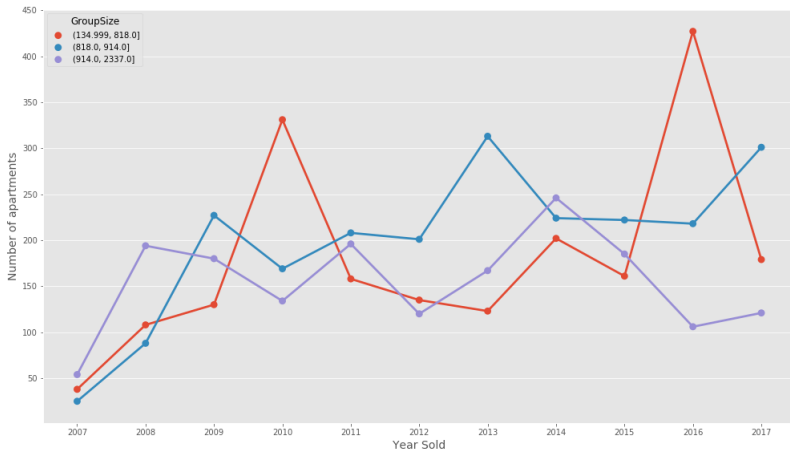
Thống kê số lượng và giá bán trung bình của các căn hộ theo từng năm bán ra.

| | YrSold | MeanSalePrice | NoSoldApt |
|----|--------|---------------|-----------|
| 0 | 2007 | 193989.034188 | 117 |
| 1 | 2008 | 201853.058974 | 390 |
| 2 | 2009 | 177825.657356 | 537 |
| 3 | 2010 | 148217.151420 | 634 |
| 4 | 2011 | 194249.209964 | 562 |
| 5 | 2012 | 190801.328947 | 456 |
| 6 | 2013 | 222325.983416 | 603 |
| 7 | 2014 | 239752.703869 | 672 |
| 8 | 2015 | 282221.441901 | 568 |
| 9 | 2016 | 234509.402130 | 751 |
| 10 | 2017 | 307065.311148 | 601 |

YrSold và Mean of SalePrice



YrSold và Mean of SalePrice

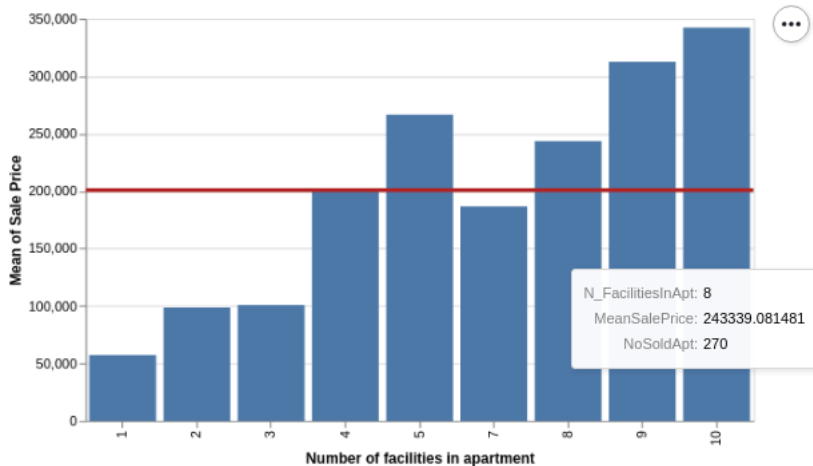


N_Facilities và Mean of SalePrice

Thống kê giá giao dịch trung bình theo số lượng tiện ích có trong căn hộ.

| | N_FacilitiesInApt | MeanSalePrice | NoSoldApt |
|---|-------------------|---------------|-----------|
| 0 | 1 | 57294.836364 | 55 |
| 1 | 2 | 98527.188406 | 69 |
| 2 | 3 | 100854.293944 | 677 |
| 3 | 4 | 200061.286109 | 1447 |
| 4 | 5 | 266346.104310 | 1160 |
| 5 | 7 | 186586.816327 | 1225 |
| 6 | 8 | 243339.081481 | 270 |
| 7 | 9 | 312737.177340 | 203 |
| 8 | 10 | 342371.644586 | 785 |

N_Facilities và Mean of SalePrice

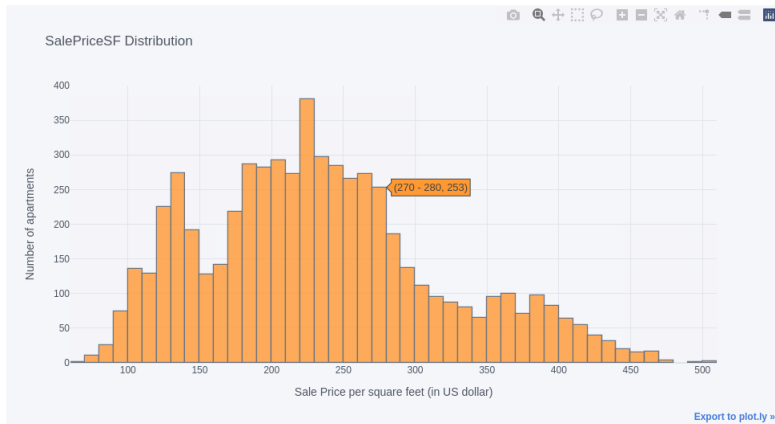


SalePrice và Size(sqf)

Thống kê giá bán trên một đơn vị diện tích (square feet).

| | SalePrice | Size(sqf) | SalePriceSF |
|---|-----------|-----------|-------------|
| 0 | 141592 | 814 | 173.945946 |
| 1 | 51327 | 587 | 87.439523 |
| 2 | 48672 | 587 | 82.916525 |
| 3 | 380530 | 2056 | 185.082685 |
| 4 | 221238 | 1761 | 125.632027 |

SalePrice và Size(sqf)

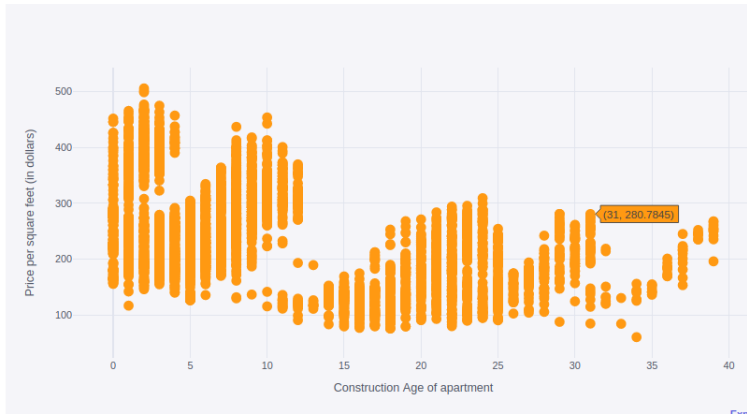


SalePriceSF và ConstructionAge

Thống kê giá giao dịch trên từng square feet theo tuổi thọ căn hộ.

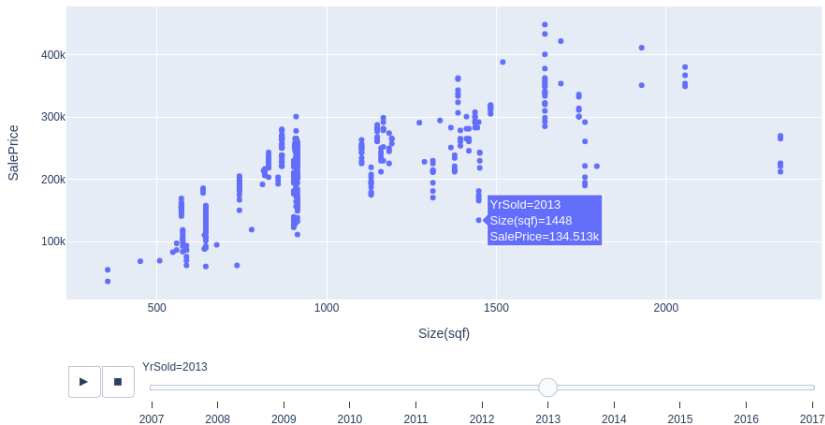
| | SalePriceSF | ConstructionAge |
|---|-------------|-----------------|
| 0 | 173.945946 | 1 |
| 1 | 87.439523 | 22 |
| 2 | 82.916525 | 22 |
| 3 | 185.082685 | 1 |
| 4 | 125.632027 | 14 |

SalePriceSF và ConstructionAge



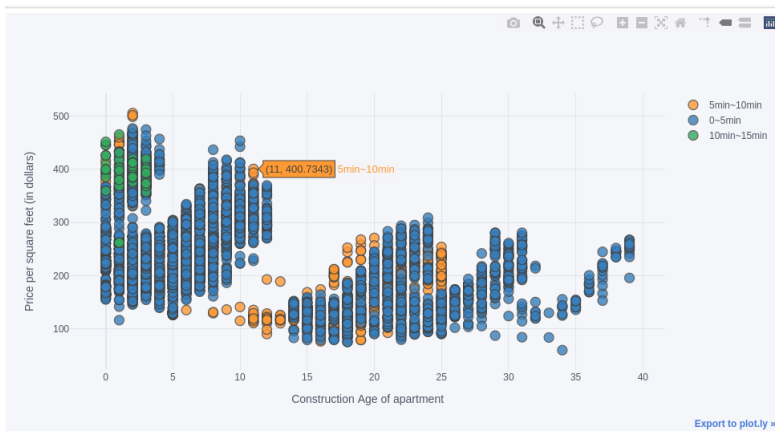
SalePrice, Size(sqf) và YrSold

Biểu đồ thể hiện sự thay đổi giá giao dịch theo kích thước và năm giao dịch.



SalePriceSF, ConstructionAge và TimeToBusStop

Biểu đồ thể hiện mối quan hệ giữa tuổi thọ căn hộ và giá giao dịch trên mỗi square foot theo thời gian đến trạm xe buýt gần nhất.



SalePriceSF, ConstructionAge và TimeToSubway

Biểu đồ thể hiện mối quan hệ giữa tuổi thọ căn hộ và giá giao dịch trên mỗi square feet theo thời gian đến trạm tàu điện ngầm gần nhất.



**Cảm ơn thầy và các bạn đã lắng
nghe!**