

# Daegu Apartments

Data Exploration and Visualization

Presented by:  
Kelas : KASDD D

1. Caryn Hanuga - 2006596541
2. Graciella Regina Indria Suwono - 2006463
3. Helga Syahda Elmira - 2006463686
4. Metta Permatasari - 2006463761





HELLO PANDA TEAM

---

# Team Members:



Caryn Hanuga

**2006596541**



Graciella Regina

**20064633282**



Helga Syahda Elmira

**2006463686**



Metta Permatasari

**2006463761**



# Table of Contents

---

- 01 Data Descriptions
  - 02 Eksplorasi dan Analisis (EDA)
  - 03 Eksplorasi 1
  - 04 Eksplorasi 2
  - 05 Pre-Processing
  - 06 Analisis Model
-



---

# Business Understanding

- Mengetahui fasilitas yang dapat mempengaruhi peningkatan harga penjualan apartemen sehingga penjual unit apartemen bisa lebih memahami harga jual dari unit apartemen
- Dengan mengetahui tren waktu dengan penjualan unit apartemen tertinggi orang yang ingin menjual apartemen bisa menjual atau mempromosikan penjualan diwaktu waktunya tersebut
- Dengan mengetahui signifikansi perbedaan harga apartemen yang dibangun sebelum dan setelah tahun 2000, maka kedepannya dapat membantu pengembang untuk memprediksi peningkatan harga yang terjadi di lain waktu
- Dengan mengetahui keterhubungan antara tinggi posisi lantai dengan harga jual, maka dapat membantu pengembang untuk memberikan harga yang sesuai dengan lantai yang dipilih
- Dengan mengetahui masa penjualan terbaik untuk menjual suatu unit (durasi dari dimulainya pembangunan). Penjual bisa mengetahui kira-kira kapan apartemennya akan terjual
- Dengan membuat model klasifikasi, maka pengembang dapat mengetahui apakah suatu apartemen family friendly atau tidak. Sedangkan, model regresi dibuat untuk membantu pengembang memprediksi harga jual apartemen



HELLO PANDA TEAM



# Data Description

---



# 5891 ENTRIES

31 Columns

01 SalePrice	07 HallwayType	07 N_APT	07 HallwayType
02 YearBuilt	08 HeatingType	08 N_manager	08 HeatingType
03 YrSold	09 AptManageType	09 N_elevators	09 AptManageType
04 MonthSold	10 N_Parkinglot(Ground)	10 SubwayStation	10 N_Parkinglot(Ground)
05 Size(sqf)	11 N_Parkinglot(Basement)	11 N_Parkinglot(Basement)	11 N_Parkinglot(Basement)
06 Floor	12 TimeToBusStop	12 TimeToBusStop	12 TimeToBusStop



# Eksplorasi dan Analisis (EDA)

---

Dilakukan untuk lebih memahami dataset. Pemahaman terhadap data akan membantu dalam menentukan teknik pra-proses dan analisis data



# EDA yang dilakukan:

- Fasilitas apa saja yang paling berpengaruh terhadap harga apartemen?
- Apakah terdapat perbedaan harga yang signifikan antara apartemen yang dibangun sebelum dan setelah tahun 2000?
- Adakah waktu tertentu di mana penjualan apartemen lebih tinggi daripada biasanya?
- Jelaskan tren perubahan harga penjualan apartemen dari tahun ke tahun!
- Pengaruh tinggi lantai dengan harga jual (Eksplorasi 1)
- Durasi Penjualan Unit Apartemen di Daegu (Pembangunan s/d Penjualan) (Eksplorasi 2)



1a

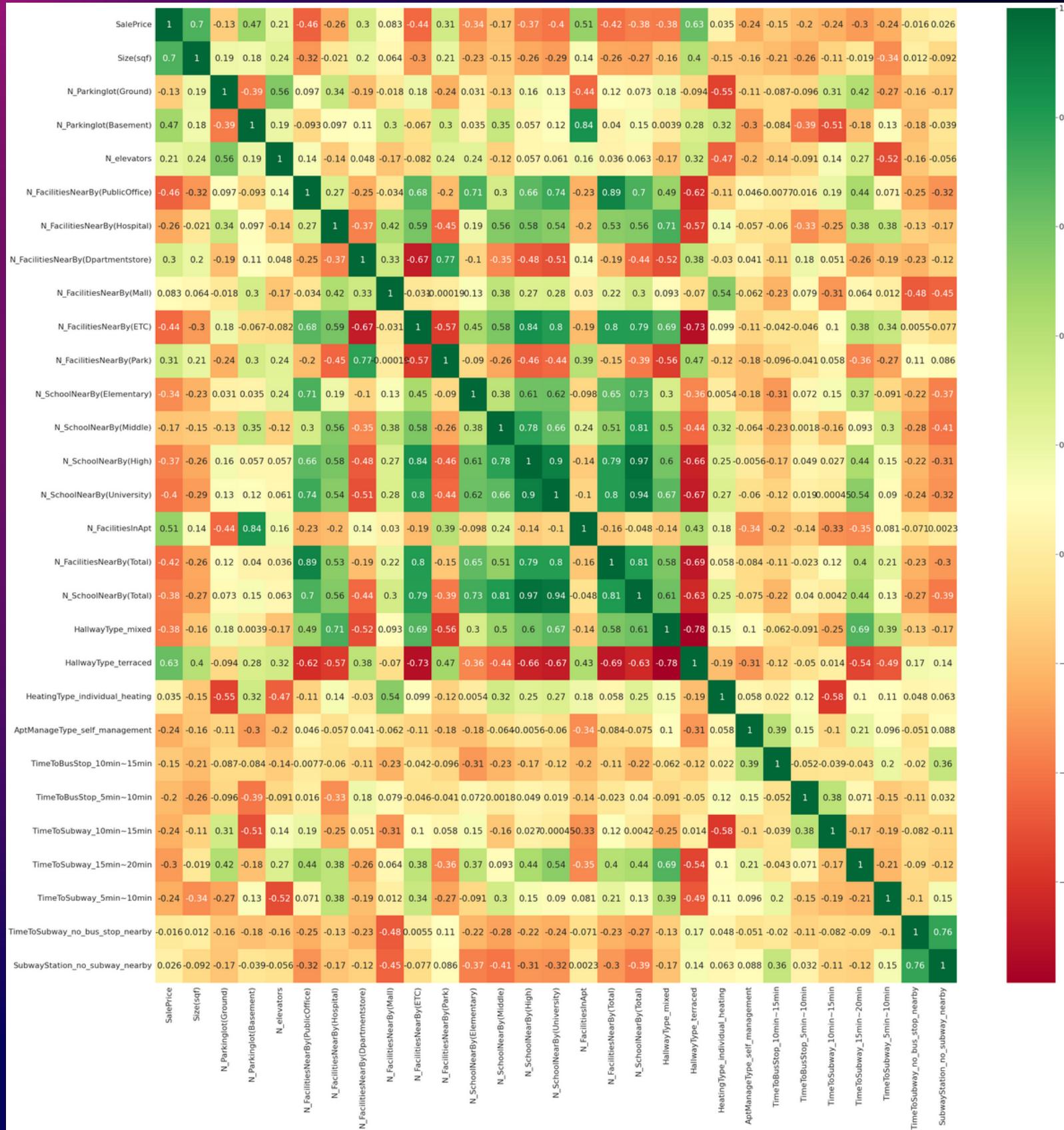
# Fasilitas apa saja yang paling berpengaruh terhadap harga apartemen?

Disini kami melakukan **uji korelasi dengan heatmap**. Sebelumnya dilakukan one-hot encoding terlebih dahulu untuk mendapatkan atribut fasilitas dalam bentuk numerik sehingga dapat dieksplor. Setelah itu dilakukan *drop* data-data yang **bukan fasilitas** seperti,

- YearBuilt,
- YrSold,
- MonthSold,
- N\_APT,
- N\_manager,
- FamilyFriendly,
- Floor,
- SubwayStation\_Banwoldang,
- SubwayStation\_Chil-sung-market,
- SubwayStation\_Daegu,
- SubwayStation\_Kyungbuk\_uni\_hospital,
- SubwayStation\_Myung-duk,
- SubwayStation\_Sin-nam



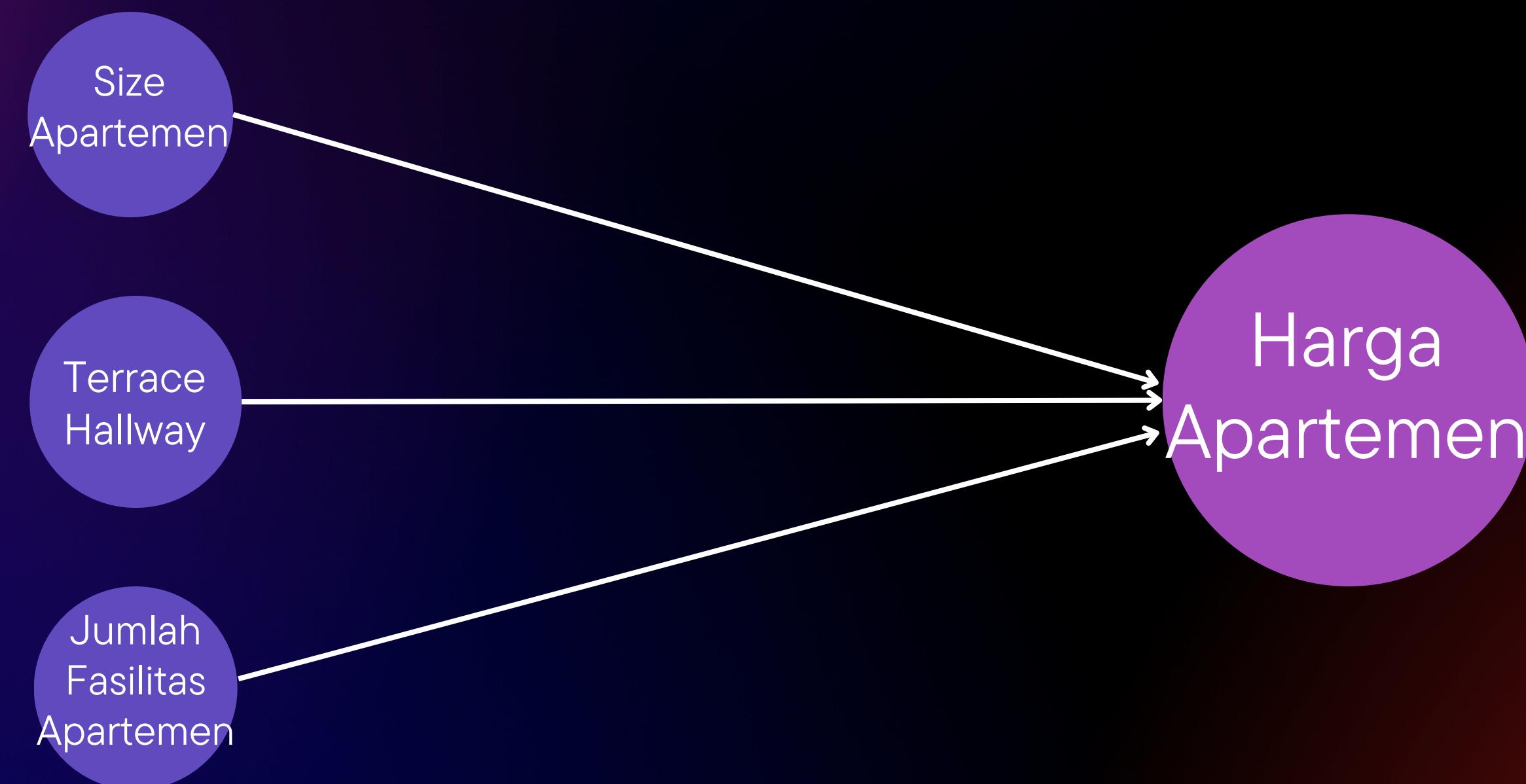
1a





1a

Fasilitas apa saja yang paling berpengaruh terhadap harga apartemen?





1b

# Apakah terdapat perbedaan harga yang signifikan antara apartemen yang dibangun sebelum dan setelah tahun 2000?

Signifikansi perbedaan harga apartemen yang dibangun sebelum dan setelah tahun 2000 akan divisualisasikan melalui plot **line chart**.

Secara umum step yang dilakukan adalah sebagai berikut:

- Memisahkan data apartemen yang dibangun pada tahun sebelum dan setelah tahun 2000
  - > jumlah data sebelum tahun 2000 : **1761** unit apartemen
  - > jumlah data setelah tahun 2000 : **4130** apartemen
- Mencari rata-rata dari harga apartemen yang dibangun sebelum dan setelah tahun 2000
  - > rata-rata harga apartemen sebelum tahun 2000 : **141197.86598523566**
  - > rata-rata harga apartemen setelah tahun 2000 : **255338.12566585955**



1b

# Apakah terdapat perbedaan harga yang signifikan antara apartemen yang dibangun sebelum dan setelah tahun 2000?

- Visualisasi plot dan melihat perbedaan harga rata-rata nya



Setelah melakukan perhitungan mengenai perbedaan rata-rata harga apartemen, kami mendapatkan bahwa perbedaan harga yang terlihat cukup jauh sebanyak **80.83709968574097%**

Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa **terdapat perbedaan harga yang signifikan** antara apartemen yang dibangun sebelum dan setelah tahun 2000.

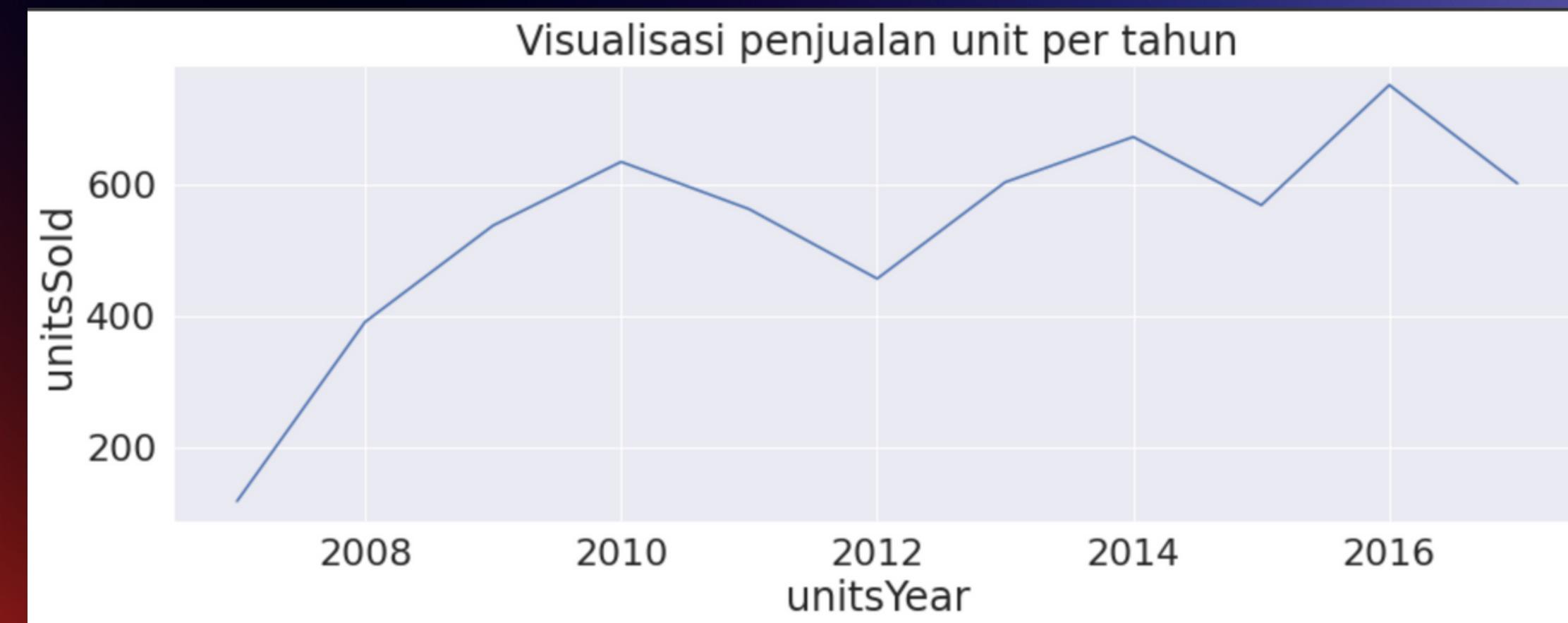


1c

# Adakah waktu tertentu di mana penjualan apartemen lebih tinggi daripada biasanya?

Disini kami mengasumsikan penjualan apartemen yang dimaksud adalah penjualan **unit apartemen**

## 1) Tahun dengan penjualan unit apartemen tertinggi



unit apartemen terjual paling banyak di tahun 2016 dengan jumlah 751

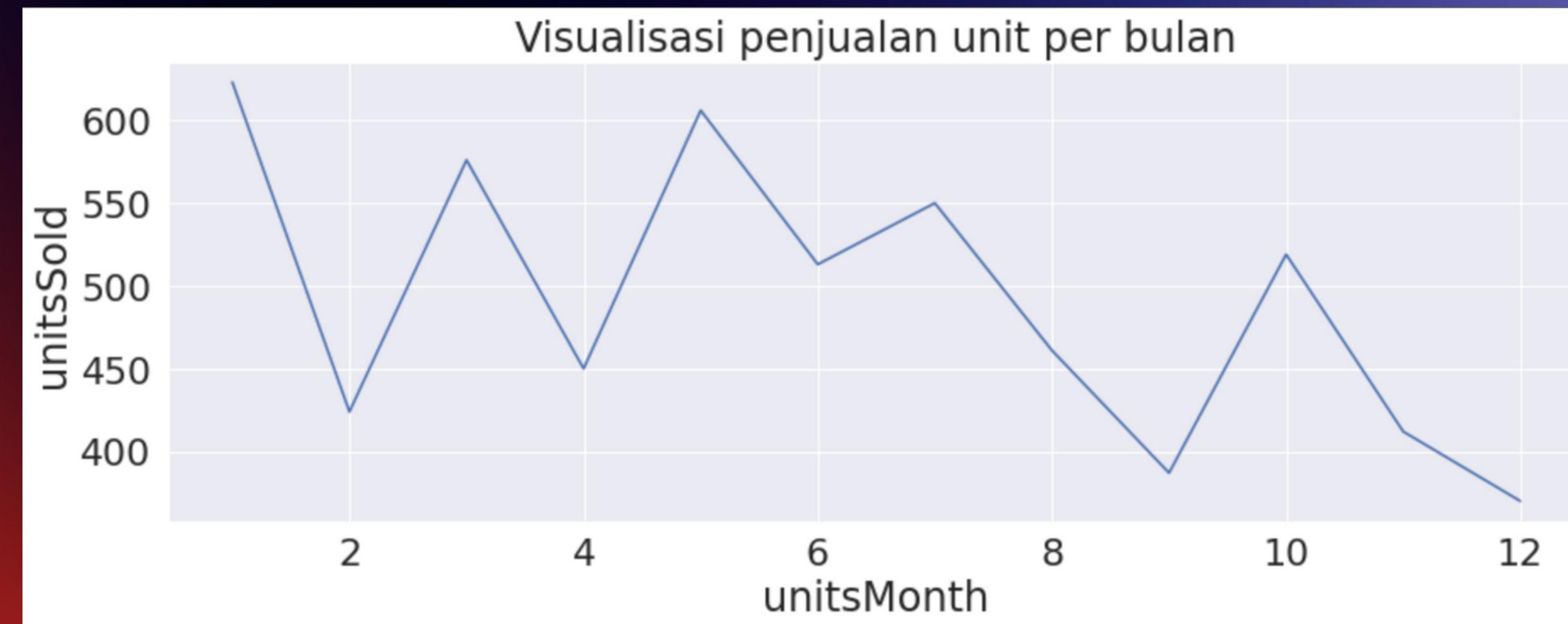
Interpretasi: Tahun 2016 adalah tahun dengan penjualan unit tertinggi dengan jumlah 751 unit



1c

Adakah waktu tertentu di mana penjualan apartemen lebih tinggi daripada biasanya?

2) Bulan (akumulasi dari 10 tahun) dengan penjualan unit apartemen tertinggi



unit apartemen terjual paling banyak di bulan 1 dengan jumlah 623

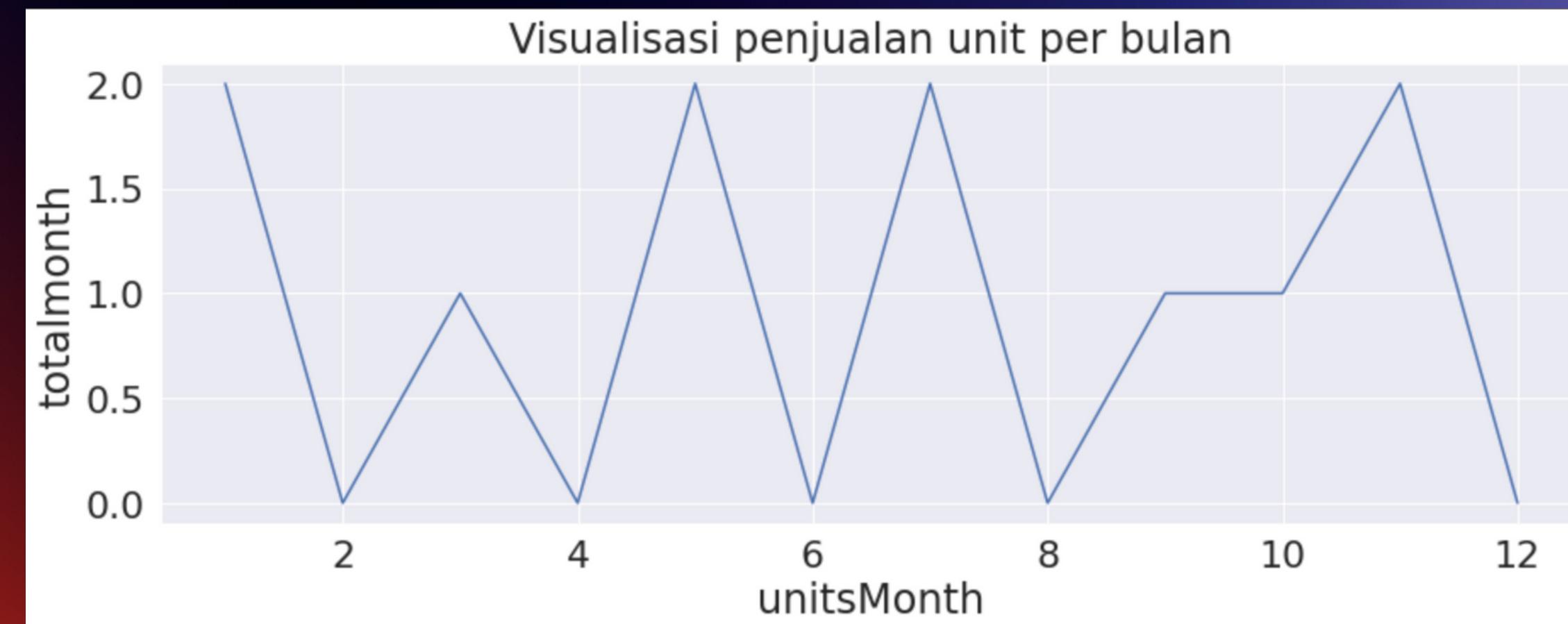
Interpretasi: Bulan Januari adalah bulan dengan penjualan unit tertinggi dengan jumlah 623 unit



1c

Adakah waktu tertentu di mana penjualan apartemen lebih tinggi daripada biasanya?

3) Bulan (akumulasi dari 10 tahun) dengan penjualan unit apartemen tertinggi dalam setahun



Interpretasi: Disini dapat dilihat bahwa bulan Januari, Mei, Juli, dan November pernah menjadi bulan dengan unit penjualan apartemen tertinggi dalam jangka 1 tahun selama 2x yang mana merupakan angka tertinggi suatu bulan bisa mencapai bulan dengan penjualan tertinggi selama setahun.



1c

Adakah waktu tertentu di mana penjualan apartemen lebih tinggi daripada biasanya?

Jadi dapat dilihat bahwa tren penjualan **meningkat pada bulan Januari, Mei, Juli, dan November.**

Kalau melihat dari keseluruhan tahun maka **tahun dengan penjualan tertinggi ada di tahun 2016** sebanyak 751 unit yang terjual. Jika melihat dari keseluruhan tahun dan bulan, maka **bulan Januari memiliki tingkat penjualan tertinggi selama 10 tahun** dengan jumlah 623 unit yang terjual.

Maka dari itu, jika melihat secara garis besar maka waktu dimana penjualan unit apartemen tertinggi untuk tahun jatuh pada tahun 2016, untuk bulan jatuh di bulan Januari. Namun, secara keseluruhan bulan-bulan yang memiliki penjualan unit tinggi ada pada bulan Januari, Mei, Juli, dan November.



1d

# Jelaskan tren perubahan harga penjualan apartemen dari tahun ke tahun!

Tren perubahan harga penjualan dari tahun ke tahun akan divisualisasikan melalui plot **line chart**.

Secara umum step yang dilakukan adalah sebagai berikut:

- Memisahkan terlebih dahulu **per tahunnya**
- Membuat array kosong, menyimpan **mean harga penjualan tiap tahunnya**
- **Visualisasi plot** dan melihat tren perubahan harganya

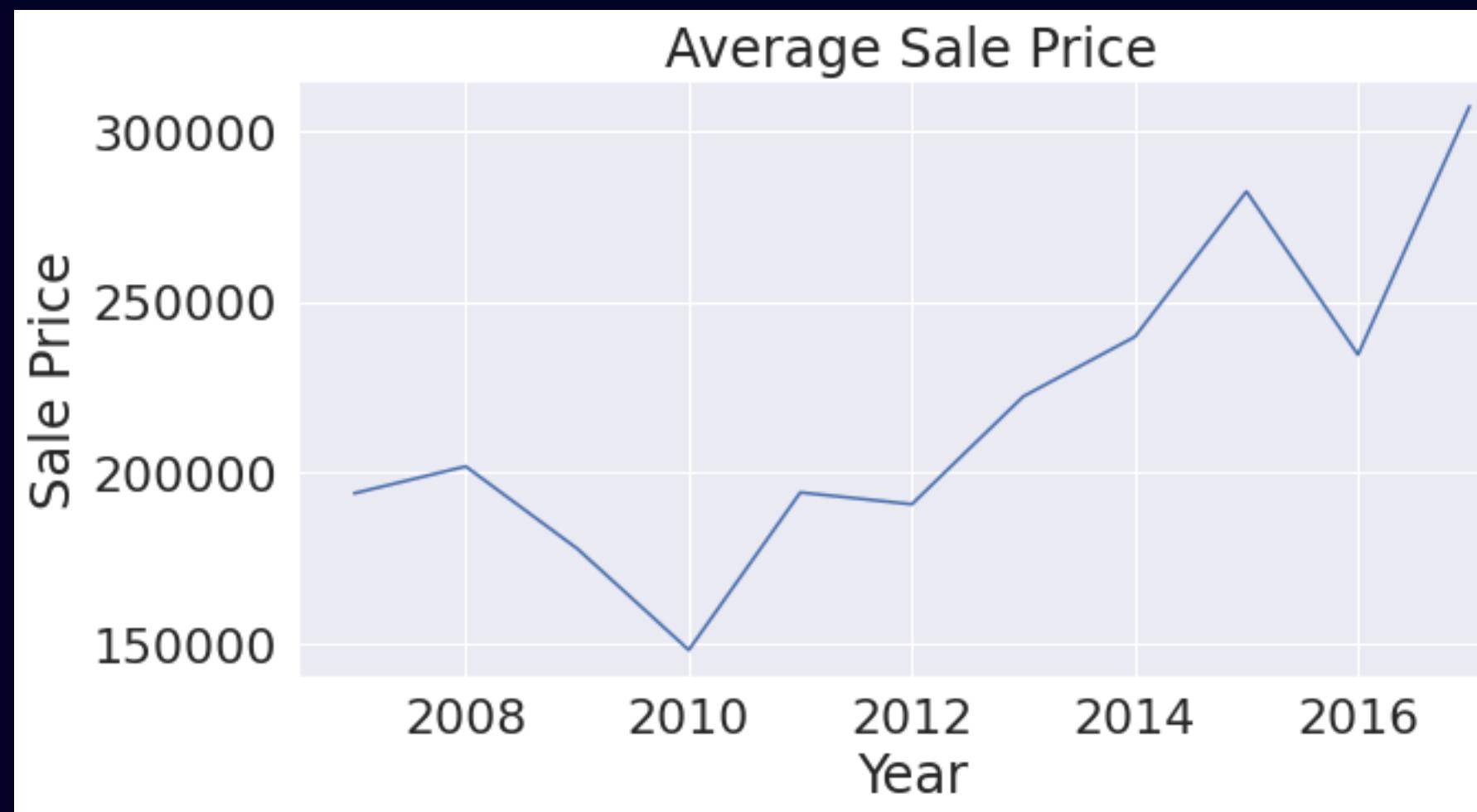
Didapat:

- Tahun terkecil yaitu = **2007**
- Tahun terbesar yaitu = **2017**



1d

Tren perubahan harga penjualan dari tahun ke tahun melalui **Line Chart**



Means harga penjualan per tahun

- **2007:** 193989.0341880342
- **2008:** 201853.05897435898
- **2009:** 177825.6573556797
- **2010:** 148217.15141955836
- **2011:** 194249.20996441282
- **2012:** 190801.32894736843
- **2013:** 222325.98341625207
- **2014:** 239752.70386904763
- **2015:** 282221.4419014084
- **2016:** 234509.40213049267
- **2017:** 307065.3111480865

Maka, jika dilihat melalui visualiasi tren perubahan harga penjualan dari tahun ke tahun, dapat disimpulkan bahwa harga penjualan mengalami **penaikan seiring berjalananya tahun**. Dan dengan mengalami **penurunan** yang sangat signifikan mulai di tahun **2009-2010** (dengan titik terendah yaitu Sale Price 2010 sebesar **148.217**)



# Eksplorasi 1

## Pengaruh tinggi lantai dengan harga jual

Pengaruh tinggi lantai dengan harga jual apartemen akan divisualisasikan melalui plot **line chart**.

Secara umum step yang dilakukan adalah sebagai berikut:

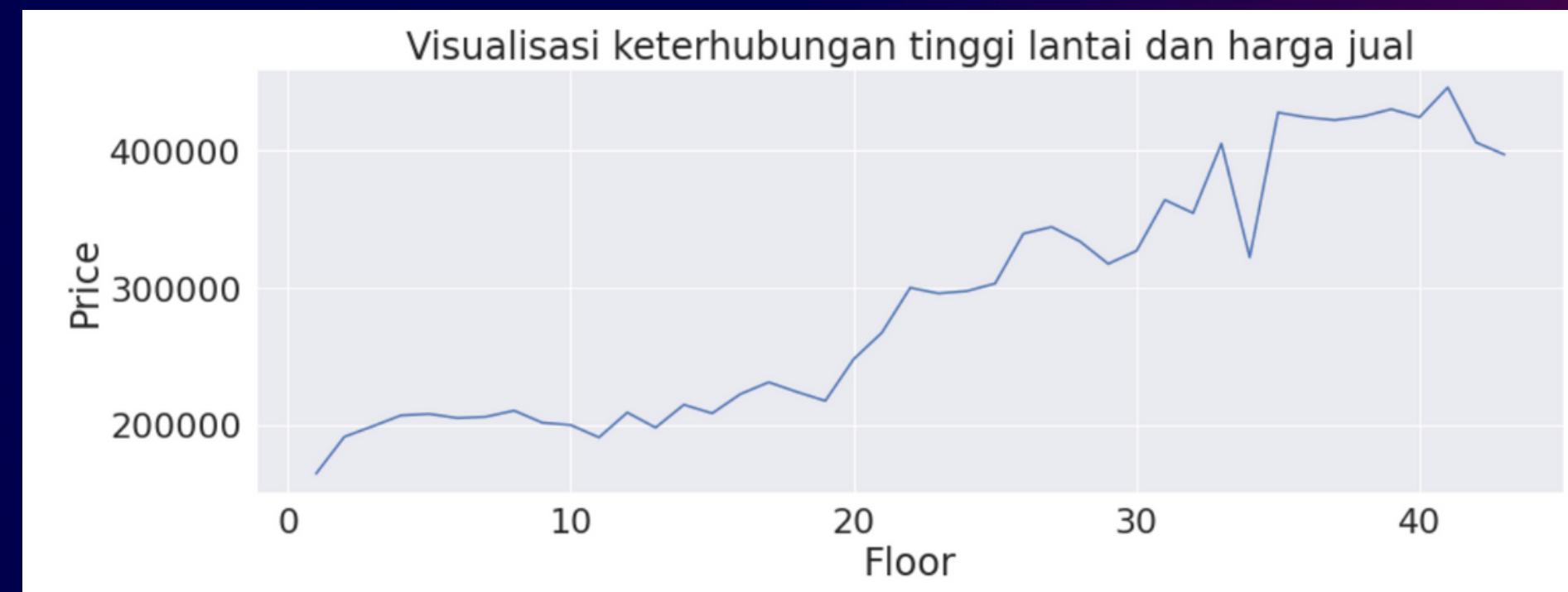
- Membuat list berisikan tinggi lantai sebuah apartemen (list berisikan *unique value*)
- Mengelompokkan apartemen berdasarkan tinggi lantai apartemen
- Menghitung rata-rata harga di tiap kelompok (berdasarkan tinggi apartemen)
- Memasukkan rata-rata tinggi apartemen ke dalam sebuah list kosong



# Eksplorasi 1

## Pengaruh tinggi lantai dengan harga jual

Memvisualisasikan keterhubungan tinggi lantai dengan harga jual pada sebuah line chart.



Berdasarkan line chart yang terlihat pada gambar diatas, dapat dijelaskan bahwa secara garis besar, chart bergerak semakin ke atas seiring dengan semakin tingginya lantai.

Oleh karena itu, dapat disimpulkan bahwa **semakin tinggi lantai, maka harga jual akan semakin tinggi**.



## Eksplorasi 2

### Durasi Penjualan Unit Apartemen di Daegu (Pembangunan s/d Penjualan)

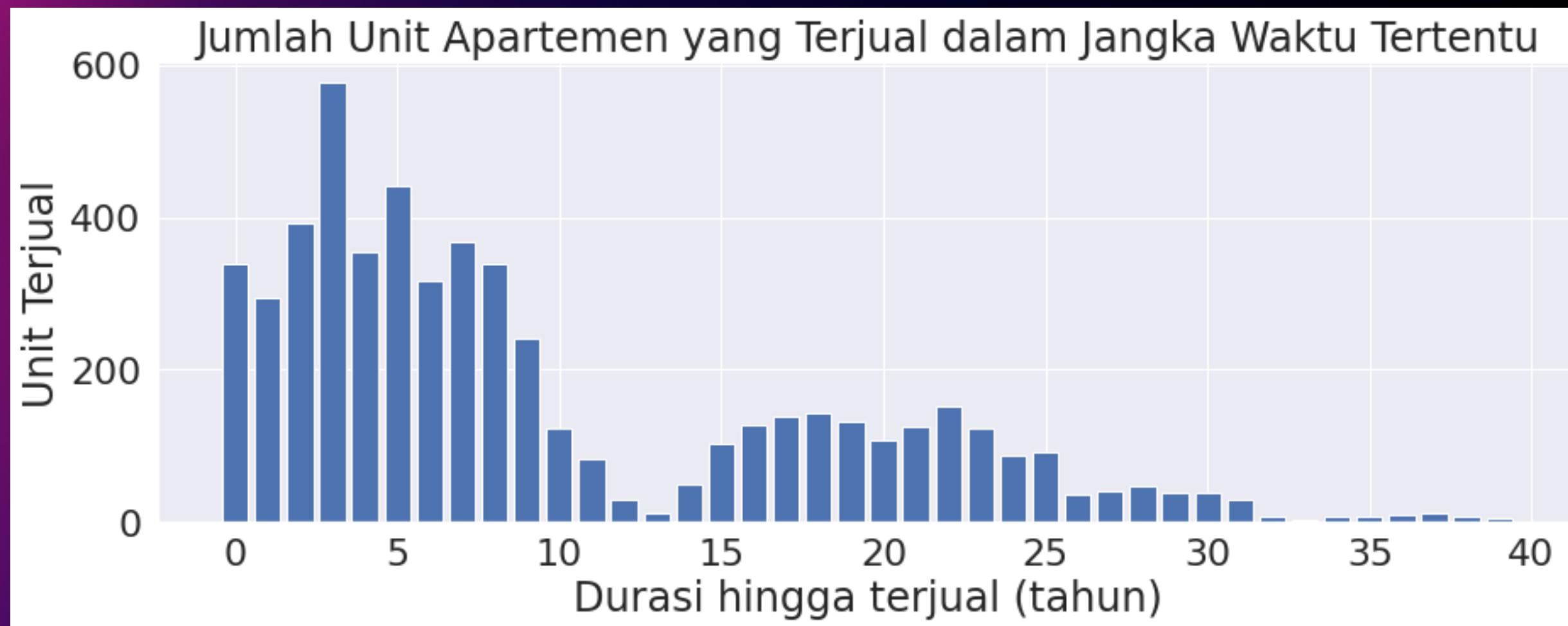
Durasi terjualnya unit apartemen di Daegu diasumsikan dimulai dari **proses pembangunan sampai dengan terjual**. Melalui atribut **YearBuilt** untuk asumsi tahun dimulainya pembangunan, dan atribut **YrSold** yang menandakan tahun terjualnya suatu unit.

Didapat:

- Apartemen terbanyak yang terjual dalam durasi pembangunan - penjualan selama **3 tahun**, yaitu sebanyak **576** unit
- Rata-rata bangunan yang ada terjual dalam durasi selama **9 tahun (9.690582959641256)**
- Terdapat **40** range durasi unik (durasi dari pembangunan - penjualan)
- Dengan list unit terjual dalam 40 durasi tersebut: [340, 294, 392, 576, 355, 442, 316, 368, 338, 240, 124, 83, 29, 11, 49, 103, 128, 139, 144, 132, 107, 125, 152, 122, 88, 91, 36, 41, 48, 38, 38, 29, 7, 2, 7, 8, 9, 12, 7, 5]



Durasi Penjualan Unit Apartemen di Daegu (Pembangunan s/d Penjualan) menggunakan visualisasi **Bar Chart**



Dengan gambar diatas dapat dilihat bahwa unit apartemen paling banyak terjual dalam jangka waktu **3 tahun** setelah dibangun. Dapat dilihat juga bahwa lebih banyak apartemen yang terjual dalam jangka waktu **dibawah 10 tahun** setelah pembangunan. Dan menunjukkan **kecenderungan lebih sedikit** unit apartemen yang terjual apabila telah **melewati durasi 10 tahun pertama** setelah dibangun



# Pre- Processing

---

Dilakukan pembersihan pada  
dataset DaeguApartments



# One Hot Encoding

One Hot Encoding dilakukan pada bagian EDA namun, ini sebenarnya masuk dalam *pre-processing*. Hal ini dikarenakan pada EDA 1 dibutuhkan tipe data yang sama untuk setiap atribut fasilitas.

One Hot Encoding dilakukan untuk mengubah semua tipe data kategorikal menjadi numerik sehingga data lebih mudah diolah. Khususnya untuk pengolahan yang butuh mencari korelasi atribut.

11	N_FacilitiesNearBy(PublicOffice)	5575	non-null	int64
12	N_FacilitiesNearBy(Hospital)	5575	non-null	int64
13	N_FacilitiesNearBy(Dpartmentstore)	5575	non-null	int64
14	N_FacilitiesNearBy(Mall)	5575	non-null	int64
15	N_FacilitiesNearBy(ETC)	5575	non-null	int64
16	N_FacilitiesNearBy(Park)	5575	non-null	int64

25	HallwayType_mixed	5575	non-null	uint8
26	HallwayType_terraced	5575	non-null	uint8

Contoh hasil One Hot Encoding pada atribut N\_FacilitiesNearBy dan HallwayType



# Check Missing Values

Setelah dilakukan pengecekan *missing values* ditemukan terdapat 152 *missing values* pada atribut FamilyFriendly.

- Cara Penanganan: **Mengisi semua row yang memiliki missing values dengan nilai modus atribut FamilyFriendly.** Hal ini dilakukan karena FamilyFriendly memiliki tipe nilai biner 0-1 yang artinya nilainya hanya iya atau tidak sehingga paling tepat menggunakan modus untuk mengisi missing values.

cek_missing_values(daegu_df)		
	Total	Percent
FamilyFriendly	152	0.025802

Gambar sebelum penanganan  
*missing values*

cek_missing_values(daegu_df)		
	Total	Percent

Gambar setelah penanganan  
*missing values*



# Check Duplicate

Setelah dilakukan pengecekanan duplikasi ditemukan terdapat 486 duplikasi data.

- Cara Penanganan: **Drop semua row yang merupakan duplikasi data.** Hal ini dilakukan untuk menghilangkan redundansi data dan bias pada model yang dihasilkan.





# Check Outliers

Setelah dilakukan pengecekan *outlier* ditemukan terdapat beberapa atribut yang memiliki *outlier*. Namun, atribut yang memiliki *outlier* adalah atribut yang memiliki tipe data biner (Yes/No) sehingga *outlier* pada jenis atribut ini **diabaikan**.

*Outliers* lain ditemukan di atribut **Size(sqf), Floor, dan SalePrice**. Namun, kami mengasumsikan bahwa nilai yang ada memang nilai *real* dimana memang di daerah Daegu terdapat beberapa unit apartemen yang memiliki ukuran unit, tinggi lantai apartemen, serta harga jual yang jauh berbeda dengan unit-unit lain. Maka itu, *outlier* pada atribut-atribut ini juga **diabaikan**.

- Cara Penanganan: **Diabaikan**

# Analisis Model

- Klasifikasi Apartemen Family Friendly
- Prediksi Harga Square Per Ft
- Clustering Apartemen





# Klasifikasi Apartemen Family Friendly

Klasifikasi ini digunakan untuk menentukan apakah suatu apartemen family friendly atau tidak, berdasarkan fitur-fitur yang ada

## Algoritma yang Digunakan      Performance Metrics

- 
- KFold Cross Validation untuk membagi training dan testing data. Didapatkan  $n\_split = 16$
  - Random Forest, menggunakan GridSearchCV untuk menentukan best parameter
  - Accuracy, Precision, Recall
  - F1-Score
  - MAE,MSE
  - Confusion Matrix



# Evaluasi Hasil Klasifikasi

**Akurasi:** 0.9942528735632183

**F1-score:** 0.9910585817060638

**Precision:** 0.9859154929577465

**MAE:** 0.005747126436781609

**Recall:** 0.9964157706093191

**MSE:** 0.005747126436781609

## Confusion Matrix

<b>prediction</b>	0.0	1.0
<b>actual</b>		
0.0	277	2
1.0	0	69

### Penjelasan:

Nilai akurasi, precision, recall, dan F1-score yang didapatkan sudah sangat tinggi. Nilai mean absolute error dan mean squared error yang didapatkan juga rendah. Jika dilihat dari confusion matrix, hanya ada 2 data yang diklasifikasikan di kelas yang salah. Sehingga, model klasifikasi ini sudah cukup akurat.



# Prediksi Harga per Square ft

## Algoritma yang Digunakan

---

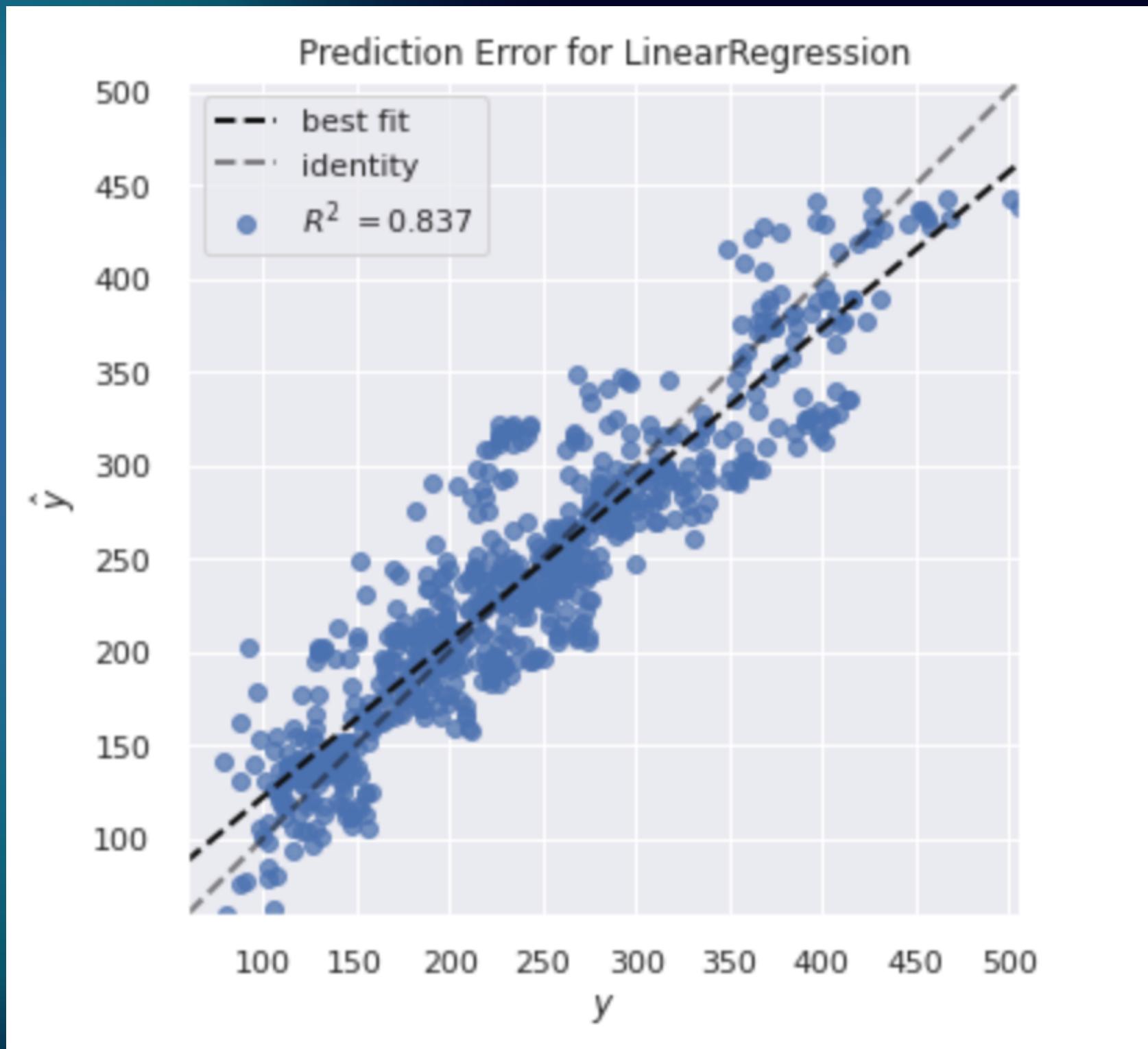
- Feature Selection menggunakan **SFS** ( Sequential Feature Selector)
- **KFold Cross Validation** untuk membagi training dan testing data
- **Linear Regression** untuk membuat model

## Performance Metrics

---

- R-Squared
- Mean Absolute Error
- Mean Squared Error

# Regression Model



**R-squared:** 0.8373076700254203  
**MAE:** 26.51511966221499  
**MSE:** 1218.146888726243

# Prediction vs Actual Data Comparison

	<b>Prediction</b>	<b>Actual</b>
14	78.194861	102.595604
15	115.910275	144.285714
20	80.265770	107.191549
31	59.710589	79.901193
42	112.171752	147.169584
...	...	...
5863	309.305422	369.762238
5865	225.007648	199.425352
5871	233.566302	262.644518
5874	442.928019	467.211180
5882	344.726156	296.056017

696 rows × 2 columns

## Penjelasan:

Jika dilihat dari perbandingan pada tabel ini, bisa dilihat bahwa hasil prediksi dan data asli mempunyai perbedaan yang cukup signifikan. Nilai MAE yang didapatkan juga masih cenderung tinggi sehingga model linear regression ini masih mempunyai error untuk memprediksi SalePrice per sqf.



# Apartment Clustering

---

Untuk clustering ini, akan ada 3 fitur yang digunakan yaitu **SalePrice**, **Size(sqf)**, dan **Floor**.

Pemilihan ketiga fitur tersebut berdasarkan hasil percobaan dari beberapa kombinasi fitur lain, dan didapatkan kombinasi ketiga fitur tersebut menghasilkan cluster yang paling baik.

## Langkah-langkah

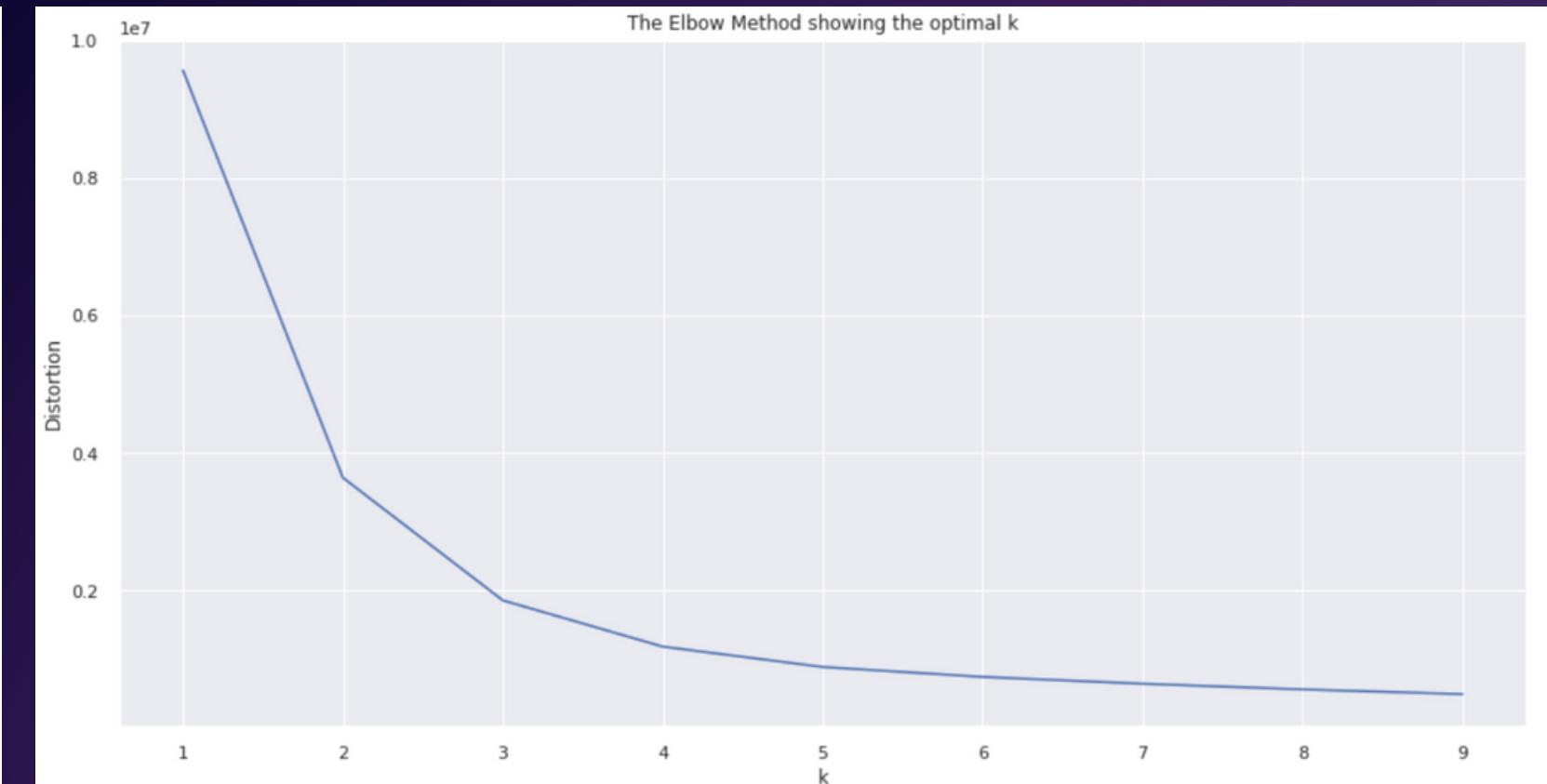
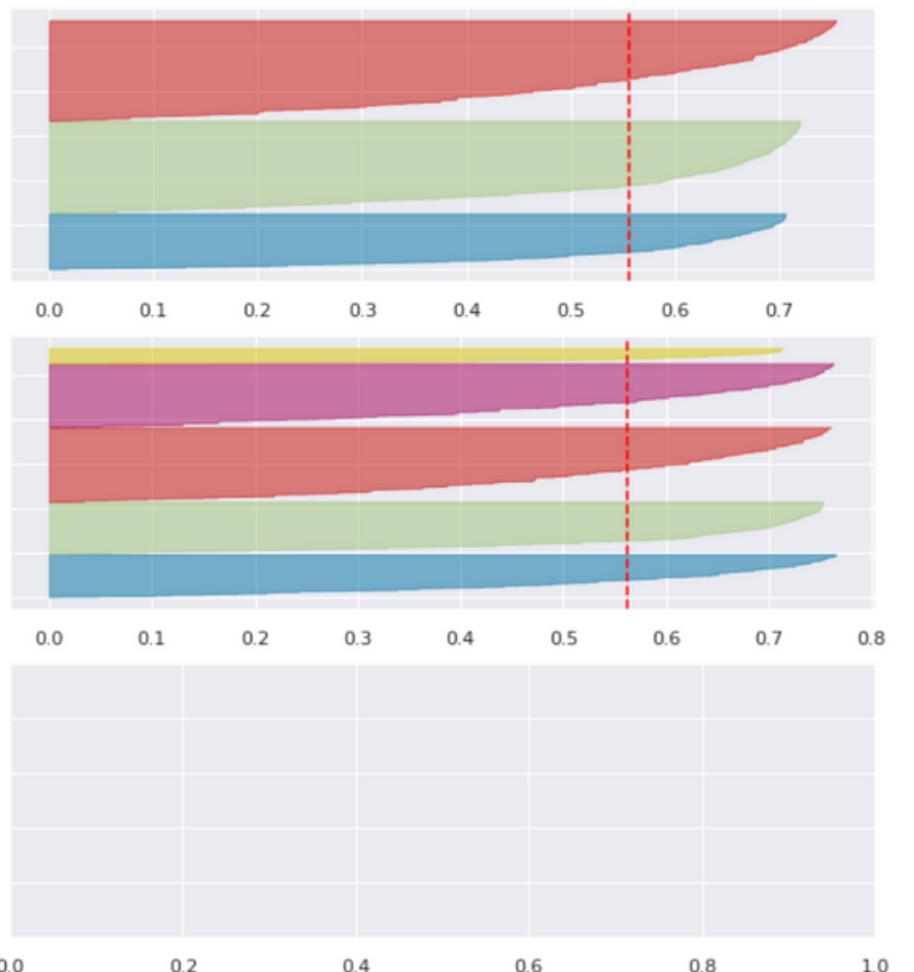
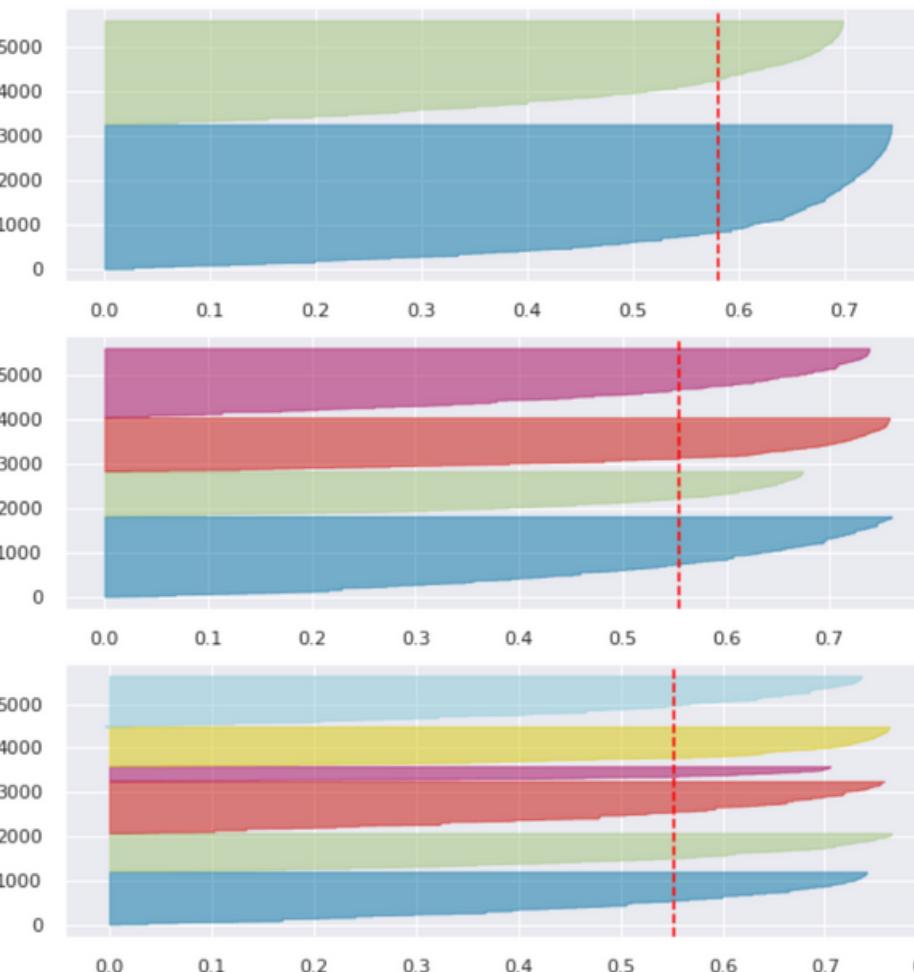
---

- Menentukan **jumlah cluster** dengan memeriksa **silhouette score** dan menggunakan **Elbow Method**. Didapatkan **jumlah cluster yang paling baik = 2**
- Menggunakan **K-Means Clustering** untuk membuat cluster



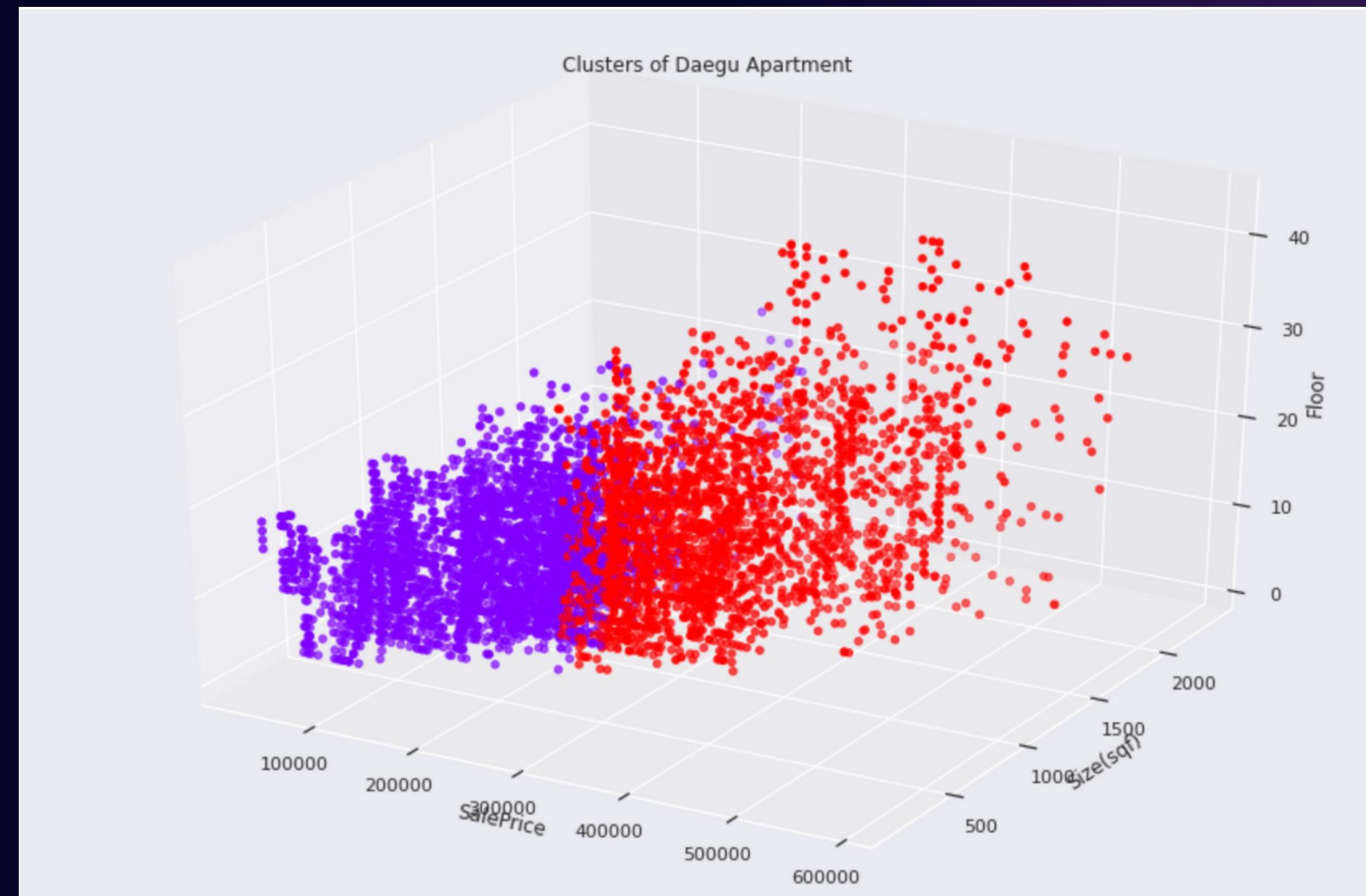
# Menentukan jumlah cluster

```
For n_clusters = 2 The average silhouette_coefficient is : 0.5798154137838014
For n_clusters = 3 The average silhouette_coefficient is : 0.5567326077850978
For n_clusters = 4 The average silhouette_coefficient is : 0.5539545780483202
For n_clusters = 5 The average silhouette_coefficient is : 0.5614843886282063
For n_clusters = 6 The average silhouette_coefficient is : 0.5517454887294481
```





# Visualisasi Cluster





# Visualisasi Cluster





# Visualisasi Cluster



Batas antara cluster 0 dan cluster 1 bisa terlihat jelas pada suatu nilai pada SalePrice yaitu di antara 20000 - 300000

Bisa dilihat juga pada ketinggian floor tertentu hanya dimiliki oleh data pada cluster 1

# Karakteristik Cluster 0

	SalePrice	Size(sqf)	Floor	Clusters
count	3245.000000	3245.000000	3245.000000	3245.0
mean	151491.786749	808.711864	10.279199	0.0
std	54838.562371	307.874810	6.072239	0.0
min	32743.000000	135.000000	1.000000	0.0
25%	106194.000000	644.000000	5.000000	0.0
50%	158584.000000	829.000000	10.000000	0.0
75%	197601.000000	914.000000	15.000000	0.0
max	239823.000000	2337.000000	30.000000	0.0

## SalePrice

Average : 151491.786749  
Max : 239823  
Min : 32743

## Size(sqf)

Average : 808.711864  
Max : 2337  
Min : 135

## Floor

Mode : 11  
Average : 10

# Karakteristik Cluster 1

	SalePrice	Size(sqf)	Floor	Clusters
<b>count</b>	2330.000000	2330.000000	2330.000000	2330.0
<b>mean</b>	328840.231330	1194.042489	14.724893	1.0
<b>std</b>	67778.570181	366.269933	8.736559	0.0
<b>min</b>	240265.000000	636.000000	1.000000	1.0
<b>25%</b>	269911.000000	910.000000	8.000000	1.0
<b>50%</b>	317699.000000	1103.000000	14.000000	1.0
<b>75%</b>	371681.000000	1448.000000	21.000000	1.0
<b>max</b>	585840.000000	2337.000000	43.000000	1.0

## SalePrice

Average : 328840.231330  
Max : 58584  
Min : 240265

## Size(sqf)

Average : 1194.042489  
Max : 2337  
Min : 636

## Floor

Mode : 7  
Average : 14

# Kesimpulan Cluster

Apartemen pada cluster 0 **cenderung memiliki SalePrice dan Size(sqf) lebih rendah** dibanding dengan cluster 1. Untuk lantai, **tidak ada perbedaan yang signifikan** antara cluster 0 dan cluster 1, keduanya tersebar dari lantai bawah sampai lantai atas. Namun, apartemen pada cluster 0 paling tinggi ada di lantai 30, sedangkan apartemen pada cluster 1 ada yang berada di lantai 43.

Thank  
you!