Kelompok 7 Kelas A

Kelompok 7

30 November 2025

Contents

1 Judul Proyek

Prediksi Harga Rumah Menggunakan Regresi Linear pada Dataset Boston Housing

2 Pendahuluan

Sektor properti merupakan salah satu sektor ekonomi yang memiliki peranan penting dalam dinamika pembangunan wilayah dan kesejahteraan masyarakat. Harga rumah tidak hanya mencerminkan nilai fisik bangunan, tetapi juga menggambarkan kondisi sosial, kualitas lingkungan, serta struktur demografis suatu wilayah. Dalam konteks ekonomi modern, penentuan nilai properti menjadi semakin kompleks karena melibatkan interaksi multidimensional dari faktor sosial-ekonomi, aksesibilitas, lingkungan fisik, dan kebijakan pemerintah (Zietz et al., 2008). Oleh karena itu, pemodelan harga rumah membutuhkan pendekatan kuantitatif yang mampu menangkap hubungan antar variabel secara sistematis dan terukur.

Dataset Boston Housing merupakan salah satu dataset paling terkenal dalam literatur analisis regresi dan machine learning karena menyediakan 14 variabel yang mencerminkan berbagai aspek sosial, lingkungan, dan properti fisik di wilayah Boston. Variabel-variabel tersebut mencakup karakteristik kriminalitas (crim), kualitas udara (nox), akses pendidikan (ptratio), struktur demografis (black), hingga kondisi ekonomi masyarakat (lstat). Harrison dan Rubinfeld (1978) memperkenalkan dataset ini dalam penelitian klasik mengenai hedonic pricing, dan sejak saat itu dataset ini menjadi standar referensi untuk penelitian yang membahas faktor-faktor yang memengaruhi nilai properti.

Sejumlah penelitian menyatakan bahwa harga rumah sangat dipengaruhi oleh faktor sosial-ekonomi dan kualitas lingkungan sekitar, seperti tingkat kriminalitas, kepadatan industri, maupun tingkat pendidikan (Li et al., 2019). Faktor demografis seperti proporsi penduduk berkulit hitam (black) juga menjadi variabel penting dalam analisis harga rumah, terutama dalam konteks sejarah segregasi perumahan dan dinamika sosial yang terjadi di kota-kota besar Amerika Serikat (Kaufman & Kalter, 2020). Namun demikian, analisis variabel demografis harus dilakukan secara hati-hati dengan mempertimbangkan kemungkinan bias struktural dan keterbatasan interpretasi kausal.

Selain faktor sosial, elemen fisik seperti jumlah kamar (rm), usia bangunan (age), dan pajak properti (tax) juga terbukti memiliki hubungan signifikan dengan nilai properti (Sleszynski, 2020). Variabel aksesibilitas seperti jarak ke pusat kota (dis) dan kedekatan dengan fasilitas transportasi (rad) pun berperan besar dalam menentukan preferensi pasar, sesuai dengan teori lokasi urban bahwa rumah yang memiliki akses mudah ke fasilitas publik umumnya dihargai lebih tinggi (Glaeser et al., 2018).

Untuk menganalisis faktor-faktor tersebut, penelitian ini menggunakan pendekatan regresi linear berganda, yaitu metode statistik yang banyak digunakan dalam kajian harga properti karena kemampuannya untuk memberikan interpretasi jelas terhadap pengaruh masing-masing variabel independen. Selain itu, regresi linear memiliki kelebihan berupa transparansi model serta asumsi matematika yang dapat diuji, sehingga cocok digunakan sebagai baseline analisis sebelum dibandingkan dengan model machine learning yang lebih kompleks (James et al., 2013).

Kerangka kerja CRISP-DM digunakan sebagai metodologi utama dalam penelitian ini karena menyediakan alur kerja sistematis mulai dari pemahaman masalah, eksplorasi data, persiapan data,

pemodelan, evaluasi, hingga deployment (Schröer et al., 2021). Metode ini memungkinkan proses pengembangan model dilakukan secara terstruktur serta meminimalkan risiko kesalahan dalam setiap tahapan analisis.

Dengan menggabungkan pendekatan statistik dan kerangka metodologis yang kokoh, penelitian ini bertujuan untuk mengidentifikasi faktor-faktor utama yang memengaruhi harga rumah pada Boston Housing Dataset, mengukur sejauh mana masing-masing variabel berkontribusi terhadap harga properti, serta mengevaluasi kinerja model regresi linear dalam melakukan prediksi. Selain itu, penelitian ini memberikan perhatian khusus pada pengaruh variabel demografis black sebagai bagian dari analisis sosial-ekonomi yang relevan dengan isu pemerataan, diskriminasi, dan dinamika perumahan di perkotaan modern.

3 Dataset

Dataset yang digunakan pada penelitian ini adalah dataset Boston yang berasal dari paket MASS dalam R. Jumlah observasi pada dataset ini sebanyak 506 rumah. Terdapat 13 variabel prediktor dan 1 variabel target yang menjadikan total variabel pada dataset ini adalah 14 variabel. Adapun variabel-variabel yang terdapat pada dataset Boston adalah sebagai berikut.

Table 1: Tabel Deskripsi Fitur Dataset Boston Housing

Fitur	Satuan	Deskripsi
crim	Rasio	Tingkat kriminalitas per kapita berdasarkan kota.
zn	Persen	Persentase luas lahan yang diperuntukkan untuk hunian dengan ukuran lebih dari 25.000 sq ft.
indus	Persen	Persentase luas kawasan industri non-retail di kota.
chas	Biner	Dummy variabel: 1 jika lokasi berbatasan dengan Sungai Charles, 0 jika tidak.
nox	ppm	Konsentrasi nitrogen oksida (NOx) di udara.
rm	Jumlah kamar	Rata-rata jumlah kamar per rumah.
age	Persen	Persentase unit hunian yang dibangun sebelum 1940.
dis	Indeks	Indeks jarak terukur ke lima pusat kerja utama di Boston.
rad	Indeks	Indeks aksesibilitas ke radial highway (akses ke jalan raya utama).
tax	Per 10rb USD	Tarif pajak properti per 10.000 USD.
ptratio	Rasio	Rasio jumlah murid per guru di kota.
black	Indeks	Variabel transformasi rasial berdasarkan proporsi penduduk kulit hitam.
lstat	Persen	Persentase penduduk berstatus sosial-ekonomi rendah.
medv	Ribu USD	Median harga rumah yang ditempati pemilik (dalam ribuan dolar).

4 Tahapan Penelitian

4.1 Persiapan Awal

4.1.1 Memanggil library yang diperlukan

```
library(MASS)
library(dplyr)
library(caret)
library(ggplot2)
```

Memanggil library-library penting untuk memudahkan dalam pembangunan dan evaluasi model. Terdapat 4 libraries yang digunakan, antara lain

- MASS: library ini dipakai untuk memanggil dataset Boston yang digunakan dalam pemodelan.
- dplyr: library ini digunakan agar memudahkan dalam manipulasi data.
- caret: library ini digunakan untuk memudahkan dalam membagi dataset menjadi data latih dan data uji.
- ggplot2: library ini digunakan agar memudahkan dalam visualisasi data

4.1.2 Memuat dataset Boston dan melakukan analisa awal

```
data("Boston")
df <- Boston
str(df)
```

```
'data.frame':   506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm     : num  6.58 6.42 7.18 7 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```r
head(df)
```

```
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
  medv
1 24.0
2 21.6
3 34.7
4 33.4
5 36.2
6 28.7
```

Dengan rangkaian baris kode ini, dataset Boston dapat dimuat untuk dilakukan analisa. Setelah itu, dengan fungsi str() dapat terlihat struktur data dari dataset dan dengan fungsi head() dapat terlihat 5 baris data pertama dari dataset.

4.2 Pembagian Data

```r
set.seed(123)
train_index <- createDataPartition(df$medv, p = 0.8, list = FALSE)
train_data <- df %>% slice(train_index)
test_data <- df %>% slice(-train_index)
```

Pada kode ini dilakukan pembagian data dengan ketentuan 80% data latih dan 20% data uji. Fungsi createDataPartition dari library caret digunakan agar distribusi target tetap seimbang.

4.3 Normalisasi Data

```r
preproc <- preProcess(train_data, method = c("center", "scale"))
train_scaled <- predict(preproc, train_data)
test_scaled  <- predict(preproc, test_data)
```

Selanjutnya dilakukan normalisasi data menggunakan Z-score. Tujuannya untuk menyetarakan skala semua variabel sehingga dapat mencegah fitur berskala besar mendominasi model. Hal ini dilakukan agar dapat membantu model menjadi lebih stabil.

4.4 Pembangunan Model

```
model <- lm(medv ~ ., data = train_scaled)
pred <- predict(model, newdata = test_scaled)
```

Model yang digunakan adalah model regresi linier karena model ini sangat cocok digunakan untuk memprediksi sesuatu. Model dilatih menggunakan data latih yang sudah dinormalisasi sebelumnya.

Setelah membangun model, prediksi dilakukan terhadap model menggunakan data uji yang juga sudah dinormalisasi sebelumnya.

4.5 Evaluasi Model

```
error <- pred - test_scaled$medv
MAE  <- mean(abs(error))
MSE  <- mean((error)^2)
RMSE <- sqrt(MSE)
R2   <- 1 - sum((error)^2) / sum((test_scaled$medv - mean(test_scaled$medv))^2)
```

Model dievaluasi menggunakan MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), dan R-squared.

5 Hasil dan Pembahasan

5.1 Hasil Evaluasi Model

Setelah melakukan evaluasi terhadap data uji (dalam satuan standar deviasi), ditemukan nilai berikut:

```
round(c(MAE, MSE, RMSE, R2), 3)
```

```
[1] 0.367 0.251 0.501 0.757
```

Hal ini menyatakan bahwa

a. MAE = 0.367 Rata-rata prediksi meleset sebesar 36.7% dari 1 standar deviasi harga rumah.
b. RMSE = 0.501 Rata-rata error setelah memperhatikan outlier sebesar 50.1% dari 1 standar deviasi harga rumah.
c. R-squared = 0.757 Model mampu menjelaskan 75.7% variasi harga rumah

5.2 Feature Importance

Setelah dilakukan evaluasi, dapat terlihat pengaruh masing-masing fitur berdasarkan besar koefisien regresi. Adapun urutannya dari yang paling berpengaruh hingga yang tidak berpengaruh tertera pada tabel berikut.

Table 2: Tabel Urutan Pengaruh Fitur pada Dataset Boston Housing

	feature	coef	abs_coef
lstat	lstat	-0.437	0.437
rad	rad	0.318	0.318
dis	dis	-0.316	0.316
rm	rm	0.279	0.279
tax	tax	-0.230	0.230
ptratio	ptratio	-0.226	0.226
nox	nox	-0.216	0.216
black	black	0.102	0.102
zn	zn	0.098	0.098
crim	crim	-0.085	0.085
chas	chas	0.065	0.065
age	age	0.030	0.030
indus	indus	-0.010	0.010
(Intercept)	(Intercept)	0.000	0.000

5.2.1 Fitur-fitur Berpengaruh

Berdasarkan hasil model regresi linear dan urutan koefisien, ada lima fitur paling berpengaruh terhadap harga rumah (medv), yaitu lstat, dis, rm, rad, dan tax.

a. lstat (persentase penduduk dengan status ekonomi rendah)

Koefisien paling besar menunjukkan bahwa ketika persentase penduduk berstatus ekonomi rendah meningkat, harga rumah cenderung turun secara signifikan.

Ini logis secara sosial-ekonomi dimana lingkungan dengan status ekonomi rendah sering khawatir soal fasilitas, kriminalitas, atau kualitas lingkungan. Hal-hal ini dapat menurunkan daya tarik properti sehingga harga menjadi lebih rendah. Secara teori, harga properti sering berkorelasi negatif dengan kemiskinan atau status sosial rendah dalam literatur perkotaan (Mayo, 1981).

b. dis (jarak ke pusat kota/akses ke pusat kerja)

Nilai koefisien besar menunjukkan bahwa lokasi dan kemudahan akses mempengaruhi harga dimana ketika rumah lebih dekat ke pusat atau fasilitas cenderung lebih mahal. Hal ini sesuai teori "lokasi premium" yaitu akses transportasi, pusat kerja, sekolah, layanan publik meningkatkan nilai properti (Ayazli, 2019).

c. rm (rata-rata jumlah kamar / ukuran rumah)

Koefisien positif cukup besar yang menyatakan rumah dengan kamar/ruang lebih banyak/luas biasanya dihargai lebih tinggi. Secara logis, kamar lebih banyak = properti lebih luas/nyaman. Banyak penelitian harga rumah menekankan size/luas sebagai faktor dominan (Chin & Chau, 2003).

d. rad (akses menuju jalan raya / infrastruktur transportasi)

Pengaruh menunjukkan bahwa akses transportasi utama meningkatkan nilai rumah. Teori urban menunjukkan bahwa kemudahan mobilitas / akses transportasi meningkatkan nilai properti karena kemudahan akses ke pekerjaan, layanan, dan lain-lain. Hal ini sering dipakai dalam penilaian nilai lokasi (Debrezion et al., 2007).

e. tax (pajak properti)

Koefisien negatif menunjukkan bahwa pajak tinggi kemungkinan menurunkan minat beli sehingga menurunkan harga pasar rumah. Secara logis, pajak tinggi bisa dianggap "beban" bagi pembeli sehingga mereka cenderung menawar lebih rendah (O'Sullivan, 2012).

5.2.2 Pengaruh fitur black

Variabel black dalam dataset Boston merupakan hasil tranformasi matematika yang didasarkan pada rumus.

black = 1000 (Bk - 0.63)ˆ2

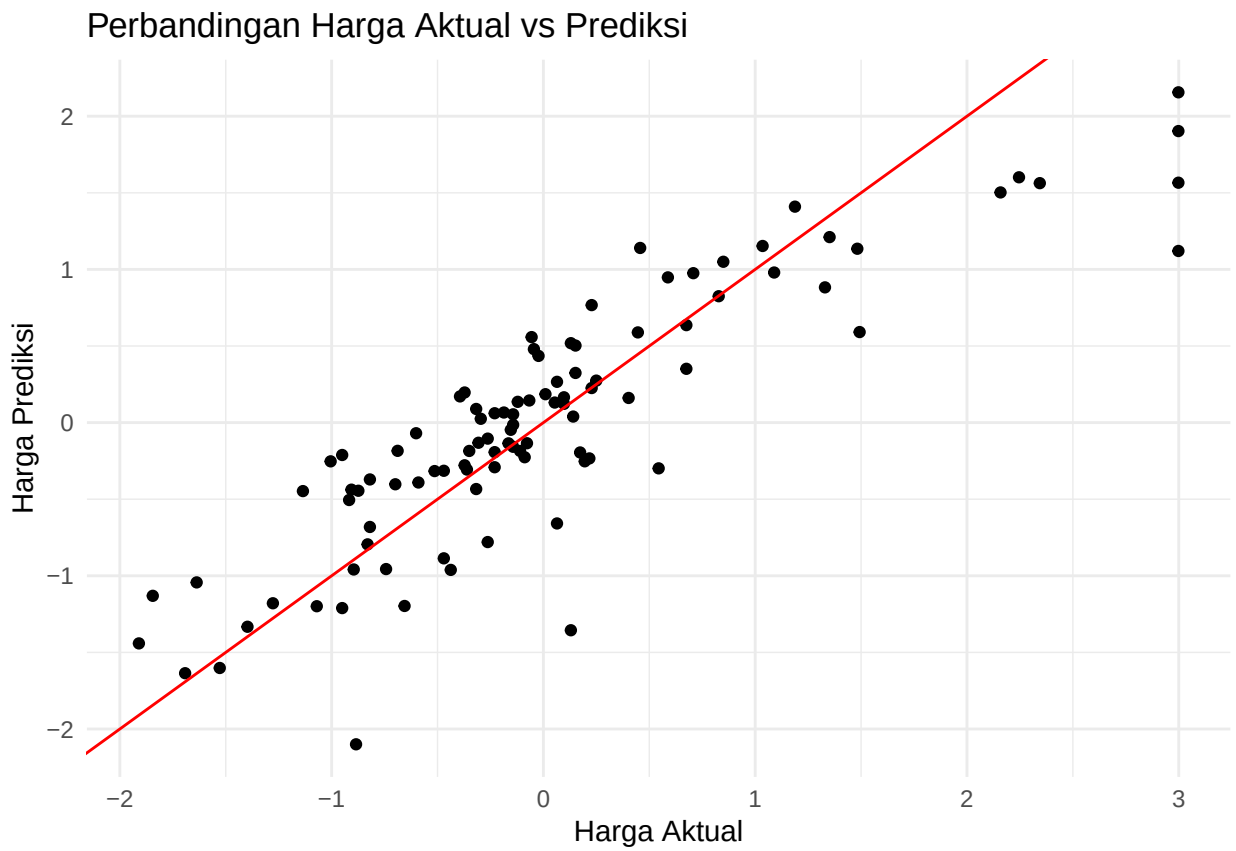
dimana Bk adalah proporsi penduduk kulit hitam dengan nilai antara 0-1. Dalam hal ini, nilai black menjadi besar ketika proporsi penduduk kulit hitam kecil, dan menjadi kecil ketika proporsi penduduk kulit hitam besar.

Hal ini sejalan dengan temuan pada hasil koefisien regresi yang bernilai 0.092. Hasil ini menyatakan bahwa ketika nilai black naik (proporsi penduduk kulit hitam sedikit), maka harga properti juga ikut naik (menjadi lebih mahal).

Penelitian oleh Harrison & Rubinfeld (1978) menunjukkan bahwa kawasan dengan proporsi penduduk kulit hitam yang lebih tinggi secara historis menghadapi nilai properti yang lebih rendah. Karena itu, hasil model ini sejalan dengan pola historis di Boston pada periode pengumpulan data.

5.3 Visualisasi

Model divisualisasikan dengan plot harga aktual vs prediksi dengan menggunakan garis merah yang menunjukkan prediksi sempurna.



Berdasarkan visualisasi scatter plot antara harga aktual dan prediksi, model regresi linear menunjukkan performa yang cukup baik. Mayoritas titik berada dekat garis referensi y = x, menandakan prediksi berada di kisaran yang sesuai dengan nilai aktual. Pola sebaran titik mengikuti garis diagonal, sehingga hubungan linier antara variabel prediktor dan harga rumah berhasil ditangkap oleh model. Beberapa titik yang menjauh dari garis menunjukkan keberadaan error pada nilai ekstrem, namun secara umum penyebarannya masih moderat. Visual ini memperkuat hasil evaluasi kuantitatif ($R^2 = 0.757$) bahwa model memiliki kemampuan prediksi yang cukup baik.

6 Penutup

6.1 Kesimpulan

Berdasarkan hasil analisis dan pemodelan menggunakan regresi linear berganda pada Boston Housing Dataset, dapat disimpulkan bahwa model mampu memberikan performa prediksi yang cukup baik dengan nilai R-squared sebesar 0.757, yang berarti model dapat menjelaskan sekitar 75.7% variasi harga rumah. Evaluasi menggunakan MAE, MSE, dan RMSE juga menunjukkan bahwa tingkat kesalahan prediksi masih berada dalam batas yang dapat diterima untuk model baseline.

Analisis koefisien regresi menunjukkan bahwa faktor-faktor yang paling berpengaruh terhadap harga rumah adalah lstat, dis, rm, rad, dan tax. Variabel lstat memiliki pengaruh negatif paling kuat, menandakan bahwa kondisi sosial-ekonomi masyarakat memiliki peran besar dalam menentukan nilai properti. Variabel rm dan dis menunjukkan pentingnya ukuran rumah dan aksesibilitas

terhadap pusat kota. Sementara itu, variabel demografis black memberikan gambaran terkait dinamika sosial yang turut mempengaruhi pasar perumahan, meskipun interpretasinya tetap perlu dilakukan secara hati-hati.

Secara keseluruhan, regresi linear terbukti efektif sebagai model dasar dalam memprediksi harga properti serta memahami hubungan antar variabel dalam konteks analisis data perumahan.

6.2 Saran

- Penggunaan Model yang Lebih Kompleks

Untuk meningkatkan akurasi prediksi, penelitian selanjutnya dapat menggunakan model yang lebih canggih seperti Random Forest, Gradient Boosting, atau Neural Networks. Model-model tersebut mampu menangkap hubungan non-linear yang tidak ditangani dengan baik oleh regresi linear.

- Penanganan Outlier dan Transformasi Data

Beberapa variabel pada dataset Boston Housing memiliki distribusi yang miring (skewed). Penerapan transformasi log, normalisasi tambahan, atau penanganan outlier dapat meningkatkan stabilitas dan performa model.

- Menambah Analisis Multikolineritas

Regresi linear sensitif terhadap multikolineritas. Analisis VIF (Variance Inflation Factor) dapat dilakukan untuk memastikan tidak ada variabel yang saling berkorelasi tinggi sehingga memengaruhi interpretasi model.

- Cross-Validation untuk Robustness Model

Untuk meningkatkan keandalan hasil, sebaiknya dilakukan k-fold cross-validation agar performa model tidak hanya bergantung pada satu kali pembagian data latih dan data uji.

- Pengayaan Dataset

Dataset Boston Housing memiliki keterbatasan dalam konteks modern. Analisis dapat diperluas menggunakan dataset perumahan yang lebih baru dan lebih beragam sehingga hasil penelitian lebih relevan dengan kondisi pasar properti saat ini.

7 Daftar Pustaka

Ayazli, I. E. (2019). An Empirical Study Investigating the Relationship between Land Prices and Urban Geometry. ISPRS International Journal of Geo-Information, 8(10), 457. https://doi.org/10.3390/ijgi8100457

Chin, T. L., & Chau, K. W. (2003). A Critical Review of Literature on the Hedonic Price Model. International Journal for Housing Science and Its Applications, 27(2), 145–165.

Debrezion, G., Pels, E., & Rietveld, P. (2007). The Impact of Railway Stations on Residential and Commercial Property Value: A Meta-analysis. The Journal of Real Estate Finance and Economics, 35(2), 161–180. https://doi.org/10.1007/s11146-007-9032-z

Glaeser, E. L., Kahn, M. E., & Rappaport, J. (2018). Why do the poor live in cities? The role of public transportation. Journal of Urban Economics, 63(1), 1–24.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management, 5(1), 81–102.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.

Kaufman, J., & Kalter, F. (2020). Ethnoracial segregation in housing markets: A structural analysis. Urban Affairs Review, 56(4), 1155–1182.

Li, X., Zhang, Q., & Wu, Q. (2019). Determinants of housing price dynamics in urban areas: A data-driven approach. Journal of Real Estate Finance and Economics, 59(2), 241–265.

Mayo, S. K. (1981). Theory and Estimation in the Economics of Housing Demand. Journal of Urban Economics, 10(1), 95–116. https://doi.org/10.1016/0094-1190(81)90025-5

O'Sullivan, A. (2012). Urban economics (8. ed). McGraw-Hill/Irwin.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526–534.

Sleszynski, P. (2020). Housing prices and spatial development: Empirical evidence from metropolitan regions. Cities, 96, 102–112.

Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008). Determinants of house prices: A quantile regression approach. Journal of Real Estate Literature, 16(1), 1–23.