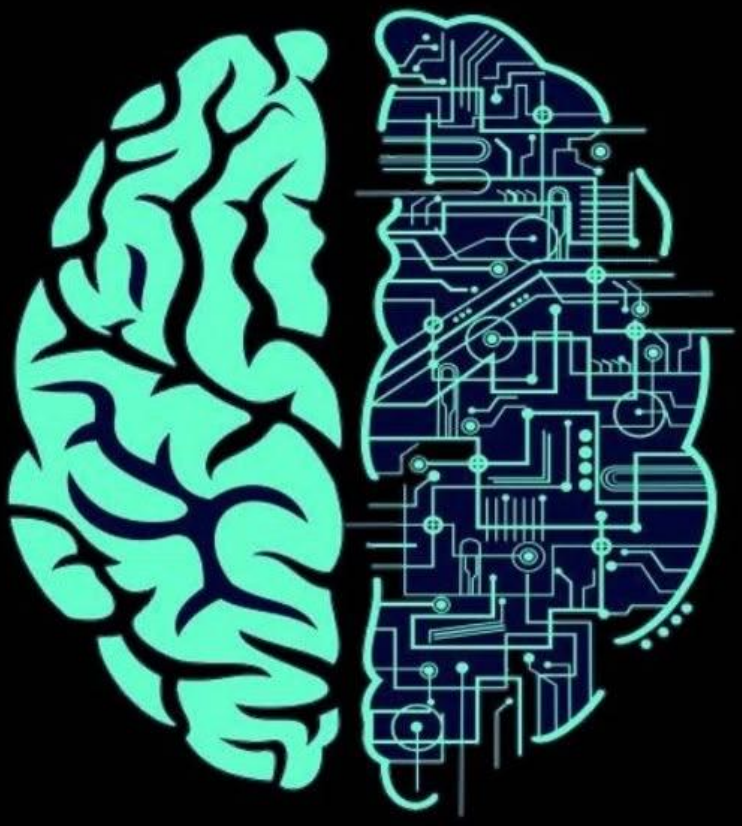
Artificial Intelligence

Machine Learning

HW02-Regression



Name: Masoud Lotfizadeh Sepehri

Student ID: 810603134

Email: mlotfizadeh@ut.ac.ir

Course Instructor: Professor Shariatpanahi

فهرست مطالب

در این تمرین هدف اصلی آشنایی با فرایند تحلیل داده‌ها و ساخت مدل‌های رگرسیونی برای پیش‌بینی قیمت فروش خانه‌ها است. بدین منظور، ابتدا داده‌ها بررسی و آماده‌سازی شدند؛ شامل شناسایی ویژگی‌ها، مدیریت مقادیر گمشده و محاسبه آمار توصیفی. سپس با تحلیل همبستگی و ترسیم نمودارهای مختلف، مهم‌ترین ویژگی‌های مؤثر بر قیمت شناسایی گردید. در ادامه، با استفاده از روش انتخاب ویژگی (SelectKBest) و معیار آماری f_regression ، مجموعه‌ای از ویژگی‌های کلیدی انتخاب شد تا از بیش‌برازش و پیچیدگی غیرضروری جلوگیری شود.

پس از آماده‌سازی داده‌ها، چهار مدل مختلف شامل Linear Regression، Ridge Regression، Lasso Regression و Polynomial Regression آموزش داده شدند و عملکرد آن‌ها با معیارهای R^2 و RMSE ارزیابی گردید. در نهایت نیز مفهوم Bias-Variance Trade-off بررسی شد تا نشان داده شود که چگونه افزایش پیچیدگی مدل می‌تواند بر بایاس و واریانس تأثیر بگذارد و منجر به بهبود یا افت عملکرد مدل شود.

بارگذاری و بررسی اولیه داده‌ها (DataFrame و info)

داده مورد استفاده شامل ۲۹۳۰ ردیف و ۸۲ ستون است. حجم داده در حافظه حدود ۱.۸ مگابایت می‌باشد. ستون‌ها از نظر نوع داده به سه گروه تقسیم می‌شوند:

- ۲۸ ستون عدد صحیح(int64) ،
- ۱۱ ستون عدد اعشاری(float64) ،
- ۴۳ ستون متنی یا دسته‌ای.(object)

ستون SalePriceمتغیر هدف است که قیمت فروش خانه‌ها را نشان می‌دهد. این ستون کامل بوده و ۲۹۳۰ مقدار بدون داده گمشده دارد. در میان ستون‌ها، تعدادی متغیر صرفاً شناسه‌ای هستند، مانند Orderو PIDکه نقش توصیفی دارند و برای مدل‌سازی مورد استفاده قرار نمی‌گیرند.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2930 entries, 0 to 2929
Data columns (total 82 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Order            2930 non-null   int64
 1   PID              2930 non-null   int64
 2   MS SubClass      2930 non-null   int64
 3   MS Zoning        2930 non-null   object
 4   Lot Frontage     2440 non-null   float64
 5   Lot Area         2930 non-null   int64
 6   Street           2930 non-null   object
 7   Alley            198 non-null    object
 8   Lot Shape        2930 non-null   object
 9   Land Contour     2930 non-null   object
 10  Utilities        2930 non-null   object
 11  Lot Config       2930 non-null   object
 12  Land Slope       2930 non-null   object
 13  Neighborhood     2930 non-null   object
 14  Condition 1      2930 non-null   object
 15  Condition 2      2930 non-null   object
 16  Bldg Type        2930 non-null   object
 17  House Style      2930 non-null   object
 18  Overall Qual     2930 non-null   int64
 19  Overall Cond     2930 non-null   int64
 20  Year Built       2930 non-null   int64
 21  Year Remod/Add   2930 non-null   int64
 22  Roof Style       2930 non-null   object
 23  Roof Matl        2930 non-null   object
 24  Exterior 1st     2930 non-null   object
 25  Exterior 2nd     2930 non-null   object
 26  Mas Vnr Type     1155 non-null   object
 27  Mas Vnr Area     2907 non-null   float64
 28  Exter Qual       2930 non-null   object
 29  Exter Cond       2930 non-null   object
 30  Foundation       2930 non-null   object
 31  Bsmt Qual        2850 non-null   object
 32  Bsmt Cond        2850 non-null   object
 33  Bsmt Exposure    2847 non-null   object
 34  BsmtFin Type 1   2850 non-null   object
 35  BsmtFin SF 1     2929 non-null   float64
 36  BsmtFin Type 2   2849 non-null   object
 37  BsmtFin SF 2     2929 non-null   float64
 38  Bsmt Unf SF      2929 non-null   float64
 39  Total Bsmt SF    2929 non-null   float64
 40  Heating          2930 non-null   object
 41  Heating QC       2930 non-null   object
 42  Central Air      2930 non-null   object
 43  Electrical       2929 non-null   object
 44  1st Flr SF       2930 non-null   int64
 45  2nd Flr SF       2930 non-null   int64
 46  Low Qual Fin SF  2930 non-null   int64
 47  Gr Liv Area      2930 non-null   int64
 48  Bsmt Full Bath   2928 non-null   float64
 49  Bsmt Half Bath   2928 non-null   float64
 50  Full Bath        2930 non-null   int64
 51  Half Bath        2930 non-null   int64
 52  Bedroom AbvGr    2930 non-null   int64
 53  Kitchen AbvGr    2930 non-null   int64
 54  Kitchen Qual     2930 non-null   object
 55  TotRms AbvGrd    2930 non-null   int64
 56  Functional       2930 non-null   object
 57  Fireplaces       2930 non-null   int64
 58  Fireplace Qu     1508 non-null   object
 59  Garage Type      2773 non-null   object
 60  Garage Yr Blt    2771 non-null   float64
 61  Garage Finish    2771 non-null   object
 62  Garage Cars      2929 non-null   float64
 63  Garage Area      2929 non-null   float64
 64  Garage Qual      2771 non-null   object
 65  Garage Cond      2771 non-null   object
 66  Paved Drive      2930 non-null   object
 67  Wood Deck SF     2930 non-null   int64
 68  Open Porch SF    2930 non-null   int64
 69  Enclosed Porch   2930 non-null   int64
 70  3Ssn Porch       2930 non-null   int64
 71  Screen Porch     2930 non-null   int64
 72  Pool Area        2930 non-null   int64
 73  Pool QC          13 non-null     object
 74  Fence            572 non-null    object
 75  Misc Feature     106 non-null    object
 76  Misc Val         2930 non-null   int64
 77  Mo Sold          2930 non-null   int64
 78  Yr Sold          2930 non-null   int64
 79  Sale Type        2930 non-null   object
 80  Sale Condition   2930 non-null   object
 81  SalePrice        2930 non-null   int64
dtypes: float64(11), int64(28), object(43)
memory usage: 1.8+ MB
```

مرور پنج ردیف اول داده نشان می‌دهد که اطلاعات ترکیبی از ویژگی‌های عددی (مانند Lot Area, Gr Liv Area) و ویژگی‌های کیفی (مانند MS Zoning, Sale Condition) هستند. همچنین در همان ابتدا می‌توان گستره تقریبی قیمت‌ها را مشاهده کرد که از حدود 105000 تا 244000 دلار در داده‌های نمونه متغیر است.

First 5 rows of the dataset:

	Order	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	...	Pool Area	Pool QC	Fence	Misc Feature	Misc Val	Mo Sold	Yr Sold	Sale Type	Sale Condition	SalePrice
0	1	526301100	20	RL	141.0	31770	Pave	NaN	IR1	Lvl	...	0	NaN	NaN	NaN	0	5	2010	WD	Normal	215000
1	2	526350040	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	...	0	NaN	MnPrv	NaN	0	6	2010	WD	Normal	105000
2	3	526351010	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	...	0	NaN	NaN	Gar2	12500	6	2010	WD	Normal	172000
3	4	526353030	20	RL	93.0	11160	Pave	NaN	Reg	Lvl	...	0	NaN	NaN	NaN	0	4	2010	WD	Normal	244000
4	5	527105010	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	...	0	NaN	MnPrv	NaN	0	3	2010	WD	Normal	189900

5 rows × 82 columns

حذف داده‌های پرت و جایگزینی مقادیر گمشده

در این مرحله داده‌ها از نظر وجود مقادیر گمشده (Missing Values) بررسی شدند. نتایج اولیه نشان داد که برخی از ستون‌ها دارای درصد بالایی از داده‌های مفقود هستند. به عنوان نمونه:

- ستون Pool QC با حدود ۹۹.۶٪ داده گمشده،
- ستون Misc Feature با بیش از ۹۶٪ داده گمشده،
- ستون Alley با بیش از ۹۳٪ داده گمشده،
- ستون Fence با حدود ۸۰٪ داده گمشده،
- ستون Mas Vnr Type با حدود ۶۰٪ داده گمشده.

وجود چنین مقادیر گمشده گسترده‌ای نشان می‌دهد که برخی ویژگی‌ها به‌طور کلی در مجموعه داده کم‌اهمیت یا کم‌استفاده هستند و می‌توانند در مراحل بعدی حذف شوند. علاوه بر این، تعدادی از ستون‌ها مانند Fireplace Qu(۴۸٪ داده گمشده) یا Lot Frontage(۱۶٪ داده گمشده) نیز نیاز به جایگزینی یا تخمین دارند.

Missing Values Summary (Before):

	Missing Count	Missing %
Pool QC	2917	99.556314
Misc Feature	2824	96.382253
Alley	2732	93.242321
Fence	2358	80.477816
Mas Vnr Type	1775	60.580205
Fireplace Qu	1422	48.532423
Lot Frontage	490	16.723549
Garage Cond	159	5.426621
Garage Finish	159	5.426621
Garage Yr Blt	159	5.426621
Garage Qual	159	5.426621
Garage Type	157	5.358362
Bsmt Exposure	83	2.832765
BsmtFin Type 2	81	2.764505
Bsmt Qual	80	2.730375
Bsmt Cond	80	2.730375
BsmtFin Type 1	80	2.730375
Mas Vnr Area	23	0.784983
Bsmt Full Bath	2	0.068259
Bsmt Half Bath	2	0.068259
BsmtFin SF 1	1	0.034130
Garage Cars	1	0.034130
Electrical	1	0.034130
Total Bsmt SF	1	0.034130
Bsmt Unf SF	1	0.034130
BsmtFin SF 2	1	0.034130
Garage Area	1	0.034130
Paved Drive	0	0.000000
Full Bath	0	0.000000
Half Bath	0	0.000000
Bedroom AbvGr	0	0.000000
Kitchen AbvGr	0	0.000000
Kitchen Qual	0	0.000000
TotRms AbvGrd	0	0.000000
Sale Condition	0	0.000000
Sale Type	0	0.000000
Yr Sold	0	0.000000
Mo Sold	0	0.000000
Misc Val	0	0.000000
Functional	0	0.000000
Fireplaces	0	0.000000
Pool Area	0	0.000000
Screen Porch	0	0.000000
3Ssn Porch	0	0.000000
Enclosed Porch	0	0.000000
Open Porch SF	0	0.000000
Wood Deck SF	0	0.000000
Order	0	0.000000
Heating QC	0	0.000000
Gr Liv Area	0	0.000000
Overall Qual	0	0.000000
MS SubClass	0	0.000000
MS Zoning	0	0.000000
Lot Area	0	0.000000
Street	0	0.000000
Lot Shape	0	0.000000
Land Contour	0	0.000000
Utilities	0	0.000000
Lot Config	0	0.000000
Land Slope	0	0.000000
Neighborhood	0	0.000000
Condition 1	0	0.000000
Condition 2	0	0.000000
Bldg Type	0	0.000000
House Style	0	0.000000
Overall Cond	0	0.000000
Low Qual Fin SF	0	0.000000
Year Built	0	0.000000
Year Remod/Add	0	0.000000
Roof Style	0	0.000000
Roof Matl	0	0.000000
Exterior 1st	0	0.000000
Exterior 2nd	0	0.000000
Exter Qual	0	0.000000
Exter Cond	0	0.000000
Foundation	0	0.000000
Heating	0	0.000000
PID	0	0.000000
Central Air	0	0.000000
1st Flr SF	0	0.000000
2nd Flr SF	0	0.000000
SalePrice	0	0.000000

پس از اعمال روش‌های جایگزینی (Imputation) برای داده‌های عددی و دسته‌ای (مثلاً میانگین برای ویژگی‌های عددی و مد برای ویژگی‌های متنی) و در برخی موارد حذف ستون‌های با داده بسیار ناقص، بررسی مجدد نشان داد که تعداد مقادیر گمشده در تمامی ستون‌های باقی‌مانده به صفر رسید. این امر به این معناست که مجموعه‌داده نهایی کاملاً پاک‌سازی شده و هیچ مقدار گمشده‌ای باقی نمانده است و در ادامه می‌تواند بدون مشکل برای مراحل بعدی مدل‌سازی مورد استفاده قرار گیرد.

```
Missing values per column (after imputation):
Order             0
Full Bath         0
Fireplaces        0
Functional        0
TotRms AbvGrd     0
Kitchen Qual      0
Kitchen AbvGr     0
Bedroom AbvGr     0
Half Bath         0
Bsmt Half Bath    0
dtype: int64
```

محاسبه آمار توصیفی (کمینه، بیشینه، انحراف معیار)

در این بخش آماره‌های توصیفی متغیرهای عددی محاسبه شد تا درک اولیه‌ای از گستره، پراکندگی و ویژگی‌های داده‌ها حاصل شود. نتایج نشان داد:

- ویژگی‌های زمینی و زیربنایی: متغیر *Lot Area*(مساحت زمین) بین ۱۳۰۰ تا ۱۶۴,۶۶۰ فوت مربع تغییر می‌کند و نشان‌دهنده تنوع بالای اندازه‌ی قطعات است. متغیر *Gr Liv Area*(زیربنای قابل سکونت) حداکثر تا ۵۶۴۲ فوت مربع می‌رسد که بیانگر وجود خانه‌های بسیار بزرگ در کنار خانه‌های کوچک‌تر است.

- کیفیت و وضعیت کلی ساختمان: متغیر *Overall Qual*در بازه ۱ تا ۱۰ تعریف شده و با انحراف معیار ۱.۳ توزیع نسبتاً متمرکزی دارد؛ اکثر خانه‌ها کیفیتی در حد متوسط به بالا دارند. متغیر *Overall Cond*نیز از ۱ تا ۹ متغیر است اما پراکندگی کمتری دارد.

- سال ساخت و بازسازی: خانه‌ها از سال ۱۸۷۲ تا ۲۰۱۰ ساخته شده‌اند و متغیر *Year Remod/Add*نشان می‌دهد بسیاری از خانه‌ها بین سال‌های ۱۹۵۰ تا ۲۰۱۰ بازسازی یا تغییر یافته‌اند. این گستره‌ی زمانی باعث ایجاد ناهمگنی زیادی در داده‌ها می‌شود.

- امکانات داخلی و رفاهی: متغیرهایی نظیر تعداد حمام کامل(*Full Bath*) ، نیم‌حمام (*Half Bath*) و شومینه (*Fireplaces*) توزیع متنوعی دارند (به‌طور مثال تعداد حمام کامل بین ۰ تا ۴ متغیر است). ظرفیت گاراژ (*Garage Cars*) نیز تا حداکثر ۵ خودرو ثبت شده است.

- قیمت فروش(SalePrice): قیمت خانه‌ها در بازه ۱۲,۷۸۹ تا ۳۳۸,۹۱۰ دلار قرار دارد. انحراف معیار بالا (≈۵۹,۰۰۰) بیانگر پراکندگی شدید قیمت‌ها و وجود بازار ناهمگون در داده‌هاست.

به طور کلی، تحلیل آماری توصیفی نشان داد داده‌ها از نظر ابعاد فیزیکی، کیفیت، امکانات و قیمت بسیار متنوع هستند. این تنوع در مراحل بعدی مدل‌سازی اهمیت بالایی دارد زیرا می‌تواند هم موجب افزایش توان پیش‌بینی شود و هم چالش‌هایی مانند وجود داده‌های پرت یا نیاز به نرمال‌سازی ایجاد کند.

Descriptive Statistics (Numeric Columns):

	min	max	std
Order	1.0	2.930000e+03	8.476762e+02
PID	526301100.0	1.007100e+09	1.886559e+08
MS SubClass	20.0	1.900000e+02	4.314435e+01
Lot Frontage	21.0	3.130000e+02	2.067421e+01
Lot Area	1300.0	1.646600e+05	6.816410e+03
Overall Qual	1.0	1.000000e+01	1.300120e+00
Overall Cond	1.0	9.000000e+00	1.124431e+00
Year Built	1872.0	2.010000e+03	3.004981e+01
Year Remod/Add	1950.0	2.010000e+03	2.085629e+01
Mas Vnr Area	0.0	1.600000e+03	1.551841e+02
BsmtFin SF 1	0.0	5.644000e+03	4.207758e+02
BsmtFin SF 2	0.0	1.526000e+03	1.677021e+02
Bsmt Unf SF	0.0	2.062000e+03	4.291713e+02
Total Bsmt SF	0.0	6.110000e+03	4.051508e+02
1st Flr SF	334.0	5.095000e+03	3.594654e+02
2nd Flr SF	0.0	1.818000e+03	4.135263e+02
Low Qual Fin SF	0.0	1.064000e+03	4.619043e+01
Gr Liv Area	334.0	5.642000e+03	4.603045e+02
Bsmt Full Bath	0.0	3.000000e+00	5.185825e-01
Bsmt Half Bath	0.0	2.000000e+00	2.488580e-01
Full Bath	0.0	4.000000e+00	5.425578e-01
Half Bath	0.0	2.000000e+00	4.998199e-01
Bedroom AbvGr	0.0	8.000000e+00	8.171888e-01
Kitchen AbvGr	0.0	3.000000e+00	2.190344e-01
TotRms AbvGrd	2.0	1.500000e+01	1.498345e+00
Fireplaces	0.0	4.000000e+00	6.382217e-01
Garage Yr Blt	1895.0	2.207000e+03	2.473035e+01
Garage Cars	0.0	5.000000e+00	7.322353e-01
Garage Area	0.0	1.488000e+03	2.031869e+02
Wood Deck SF	0.0	1.424000e+03	1.232297e+02
Open Porch SF	0.0	7.420000e+02	6.671901e+01
Enclosed Porch	0.0	1.012000e+03	6.504984e+01
3Ssn Porch	0.0	5.080000e+02	2.494264e+01
Screen Porch	0.0	5.760000e+02	5.486006e+01
Pool Area	0.0	8.000000e+02	3.482015e+01
Misc Val	0.0	1.700000e+04	5.799590e+02
Mo Sold	1.0	1.200000e+01	2.707765e+00
Yr Sold	2006.0	2.010000e+03	1.318851e+00
SalePrice	12789.0	3.389310e+05	5.898905e+04

ماتریس همبستگی و شناسایی ویژگی‌های مؤثر بر قیمت خانه

در این مرحله به کمک ماتریس همبستگی، ارتباط بین ویژگی‌های عددی و متغیر هدف یعنی قیمت فروش خانه (SalePrice) بررسی شد. هدف، شناسایی مهم‌ترین عواملی است که بر قیمت خانه اثرگذارند.

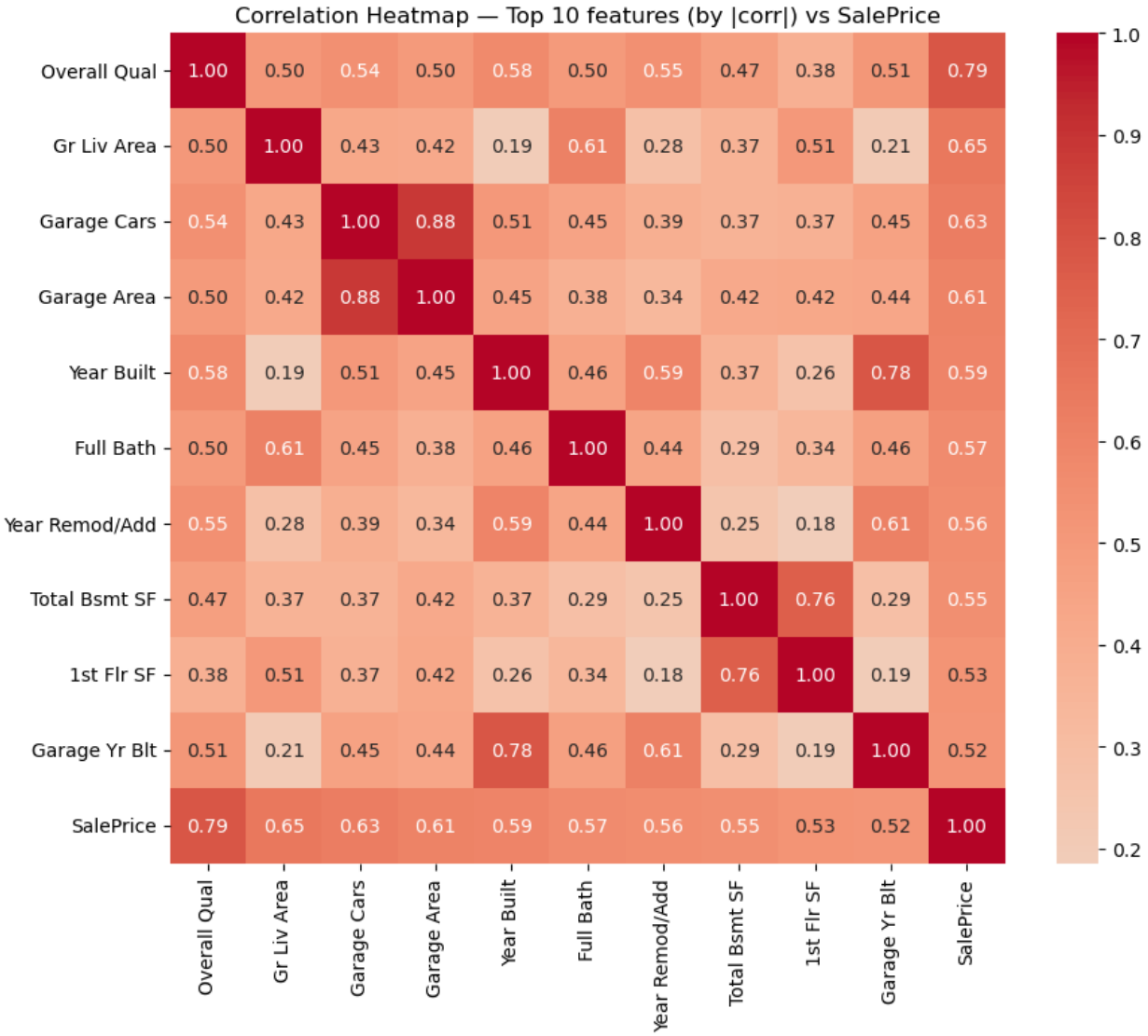
نتایج همبستگی مثبت

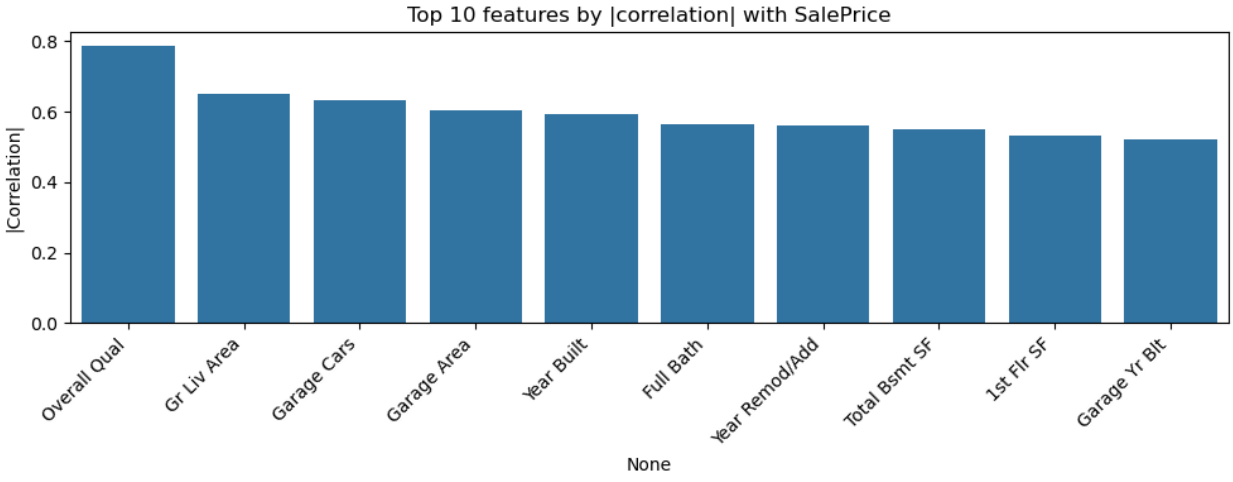
بر اساس نتایج، چند ویژگی دارای بالاترین همبستگی مثبت با قیمت فروش هستند:

- Overall Qual (کیفیت کلی ساختمان) با ضریب همبستگی حدود 0.79، مهم‌ترین عامل در تعیین قیمت است.
- Gr Liv Area (مساحت فضای زندگی بالای زمین) با مقدار 0.65، نشان‌دهنده تأثیر مستقیم افزایش مساحت بر قیمت است.
- Garage Carsو Garage Area (تعداد و مساحت پارکینگ) به ترتیب با ضرایب 0.63 و 0.61، بیانگر اهمیت ظرفیت و اندازه پارکینگ در ارزش خانه هستند.
- Year Builtو Year Remod/Add (سال ساخت و سال بازسازی) به ترتیب با ضرایب 0.59 و 0.56، بیان می‌کنند که خانه‌های جدیدتر یا بازسازی‌شده معمولاً ارزش بیشتری دارند.
- سایر متغیرها مانند Full Bathو Total Bsmt SFنیز با ضرایب بالای 0.55، نقش مهمی در افزایش قیمت ایفا می‌کنند.

به‌طور کلی، کیفیت کلی، مساحت زیربنا و امکانات اصلی بیشترین ارتباط مثبت با قیمت دارند.

```
Top positive correlations with SalePrice:
SalePrice         1.000000
Overall Qual      0.785878
Gr Liv Area       0.652255
Garage Cars       0.631196
Garage Area       0.605585
Year Built        0.592549
Full Bath         0.565246
Year Remod/Add    0.562002
Total Bsmt SF     0.549684
1st Flr SF        0.531070
Garage Yr Blt     0.521016
Fireplaces        0.462706
TotRms AbvGrd     0.428873
Mas Vnr Area      0.354364
BsmtFin SF 1      0.330787
Name: SalePrice, dtype: float64
```


Correlation Heatmap — Top 10 features (by |corr|) vs SalePrice
Overall Qual 1.00 0.50 0.54 0.50 0.58 0.50 0.55 0.47 0.38 0.51 0.79
Gr Liv Area 0.50 1.00 0.43 0.42 0.19 0.61 0.28 0.37 0.51 0.21 0.65
Garage Cars 0.54 0.43 1.00 0.88 0.51 0.45 0.39 0.37 0.37 0.45 0.63
Garage Area 0.50 0.42 0.88 1.00 0.45 0.38 0.34 0.42 0.42 0.44 0.61
Year Built 0.58 0.19 0.51 0.45 1.00 0.46 0.59 0.37 0.26 0.78 0.59
Full Bath 0.50 0.61 0.45 0.38 0.46 1.00 0.44 0.29 0.34 0.46 0.57
Year Remod/Add 0.55 0.28 0.39 0.34 0.59 0.44 1.00 0.25 0.18 0.61 0.56
Total Bsmt SF 0.47 0.37 0.37 0.42 0.37 0.29 0.25 1.00 0.76 0.29 0.55
1st Flr SF 0.38 0.51 0.37 0.42 0.26 0.34 0.18 0.76 1.00 0.19 0.53
Garage Yr Blt 0.51 0.21 0.45 0.44 0.78 0.46 0.61 0.29 0.19 1.00 0.52
SalePrice 0.79 0.65 0.63 0.61 0.59 0.57 0.56 0.55 0.53 0.52 1.00
Overall Qual
Gr Liv Area
Garage Cars
Garage Area
Year Built
Full Bath
Year Remod/Add
Total Bsmt SF
1st Flr SF
Garage Yr Blt
SalePrice
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2


Top 10 features by |correlation| with SalePrice
|Correlation|
0.0
0.2
0.4
0.6
0.8
Overall Qual
Gr Liv Area
Garage Cars
Garage Area
Year Built
Full Bath
Year Remod/Add
Total Bsmt SF
1st Flr SF
Garage Yr Blt
None

نتایج همبستگی منفی

برخی ویژگی‌ها نیز دارای همبستگی منفی یا بسیار ضعیف با قیمت فروش هستند:

- Enclosed Porch(ایوان بسته) با ضریب 0.13- ، اثر کاهشی اندکی دارد.
- Kitchen AbvGr (تعداد آشپزخانه بالای زمین) با ضریب 0.12- ، افزایش غیرمعمول تعداد آشپزخانه را به‌عنوان عاملی منفی نشان می‌دهد.
- Overall Cond (وضعیت کلی ساختمان) با ضریب 0.06- ، تأثیر منفی بسیار کم دارد.
- سایر ویژگی‌ها مانند Low Qual Fin SF و MS SubClassنیز همبستگی منفی ناچیزی دارند.

این متغیرها در مقایسه با عوامل مثبت اثر کمتری بر قیمت نهایی دارند.

```
Top negative correlations with SalePrice:
Bsmt Unf SF        0.190388
Bedroom AbvGr      0.159237
Screen Porch       0.086022
Pool Area          0.045809
3Ssn Porch         0.042738
Mo Sold            0.039737
BsmtFin SF 2       0.011301
Misc Val          -0.003627
Yr Sold           -0.016148
Bsmt Half Bath    -0.023888
MS SubClass       -0.053960
Low Qual Fin SF   -0.055602
Overall Cond      -0.061056
Kitchen AbvGr     -0.124275
Enclosed Porch    -0.128361
Name: SalePrice, dtype: float64
```

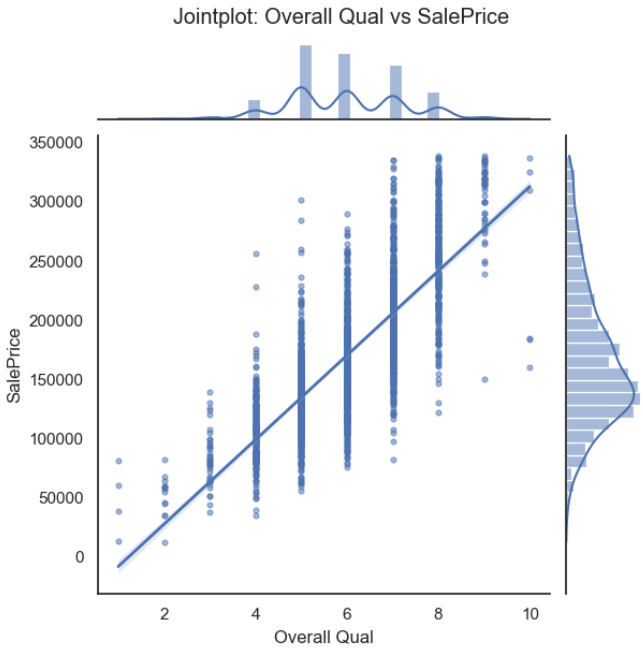
نتایج نشان می‌دهد که ویژگی‌های کیفی و کمی مرتبط با کیفیت ساخت و مساحت زیربنا بیشترین ارتباط را با قیمت فروش دارند. در مقابل، متغیرهای منفی اثر اندکی بر قیمت می‌گذارند و نقش اصلی در مدل‌سازی باید بر متغیرهای مثبت و پرقدرت متمرکز شود.

ترسیم Jointplot برای ویژگی‌های با بیشترین همبستگی

برای بررسی دقیق‌تر تأثیر هر ویژگی بر قیمت فروش، از نمودارهای Jointplot استفاده شد. این نمودارها همزمان پراکندگی نقاط و روند رابطه خطی میان متغیرها با متغیر وابسته (SalePrice) را نمایش می‌دهند. نتایج حاصل نشان می‌دهد که ویژگی‌های انتخاب‌شده در بخش قبل (با بالاترین همبستگی مثبت) الگوهای مشخص و معناداری با قیمت خانه دارند:

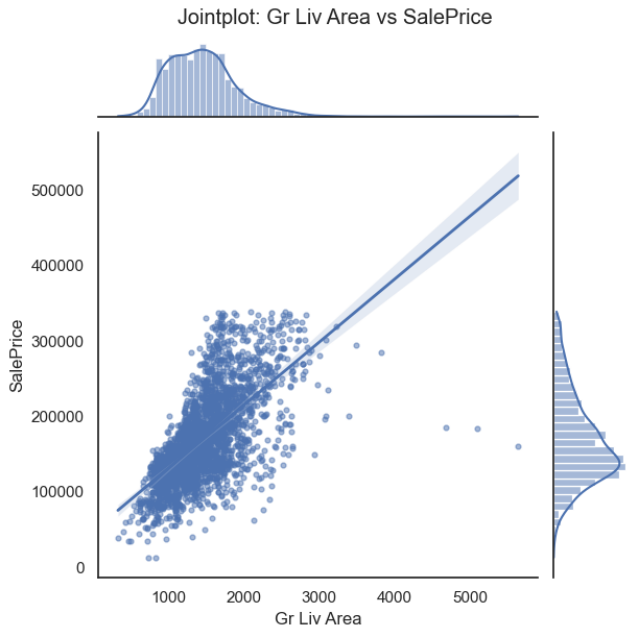
Overall Qual (کیفیت کلی ساختمان)

این ویژگی قوی‌ترین رابطه را با قیمت خانه نشان می‌دهد. همانطور که در نمودار مشخص است، با افزایش کیفیت کلی، قیمت به‌صورت خطی و یکنواخت افزایش می‌یابد. این موضوع تأیید می‌کند که کیفیت ساخت یکی از مهم‌ترین عوامل تعیین‌کننده ارزش ملک است.



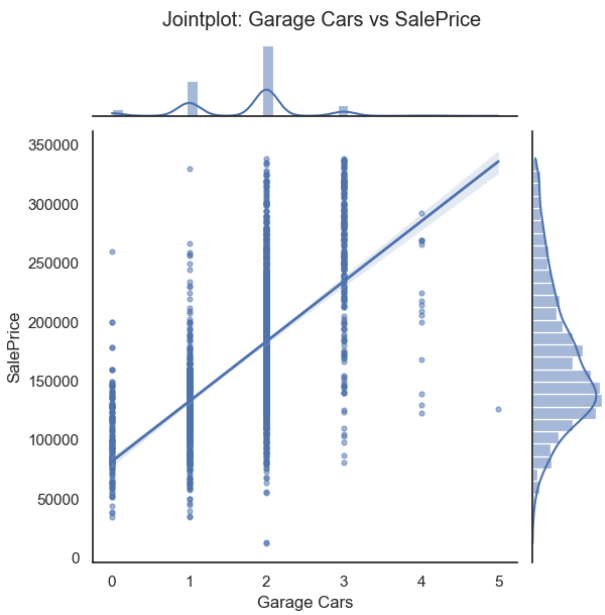
Gr Liv Area (مساحت زیربنای بالای زمین)

پراکندگی نقاط رابطه‌ای نزدیک به خط رگرسیون را نشان می‌دهد. هرچه مساحت زیربنای قابل‌استفاده بیشتر باشد، قیمت نیز بالاتر است. در مقادیر بالا، چند نمونه پرت مشاهده می‌شود که بیانگر خانه‌های بسیار بزرگ با قیمت‌های غیرمتناسب است



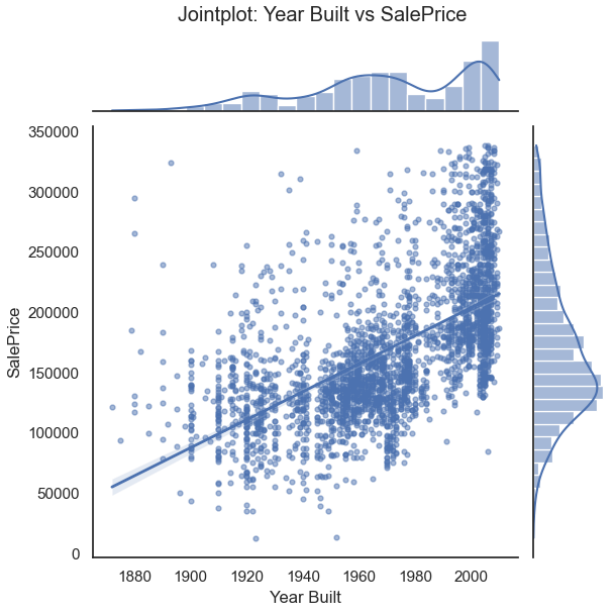
Garage Cars و Garage Area (تعداد خودروهای قابل پارک و مساحت گاراژ)

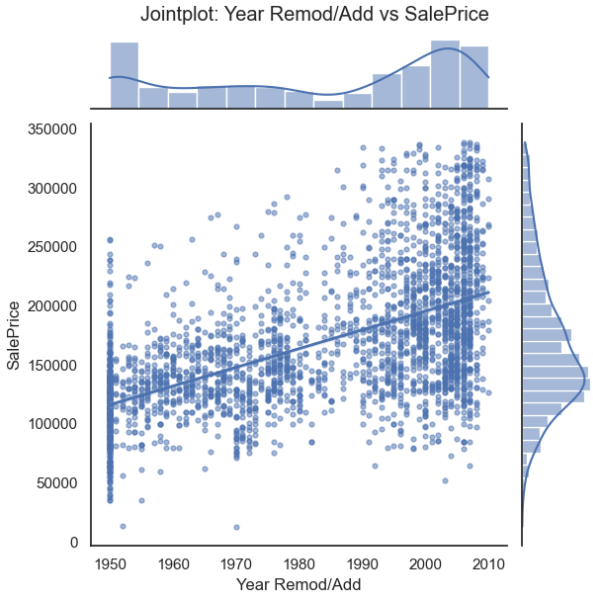
هر دو ویژگی رابطه مثبتی با قیمت دارند. خانه‌هایی با ظرفیت گاراژ بیشتر و گاراژهای بزرگ‌تر، به طور میانگین قیمت بالاتری دارند. این رابطه به‌ویژه در شهرهای کوچک یا محلات مسکونی اهمیت بیشتری پیدا می‌کند.



Year Built و Year Remod/Add (سال ساخت و سال بازسازی)

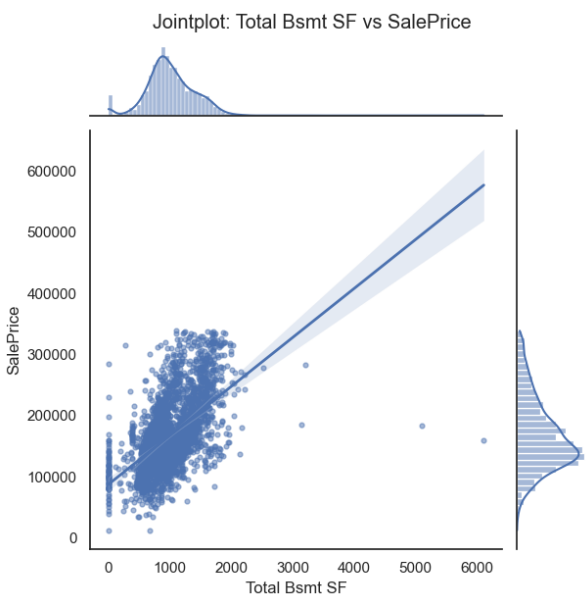
خانه‌های جدیدتر یا بازسازی‌شده قیمت بیشتری دارند. نمودارها نشان می‌دهند که پس از دهه ۱۹۸۰ روند صعودی قیمت‌ها با شیب بیشتری همراه است.

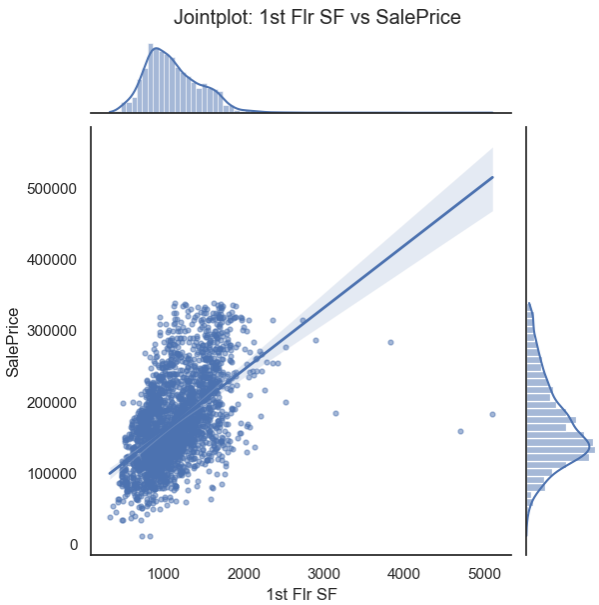




Total Bsmt SF و 1st Flr SF (مساحت زیرزمین و مساحت طبقه اول)

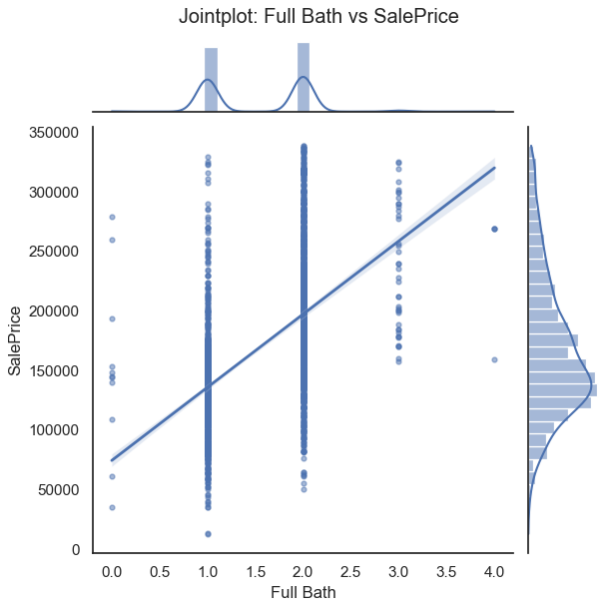
این متغیرها نیز ارتباط مثبتی با قیمت دارند؛ هر چه فضای بیشتری در اختیار خریدار قرار گیرد، ارزش ملک افزایش می‌یابد. در هر دو نمودار، مقادیر پرت مربوط به خانه‌های بسیار بزرگ قابل مشاهده است.





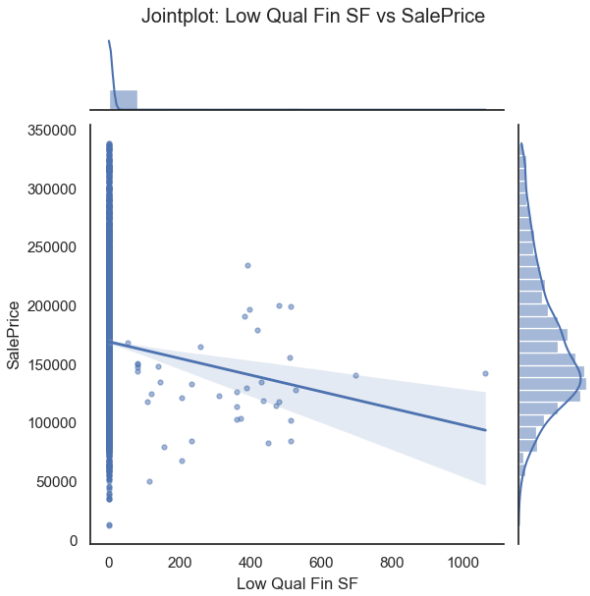
Full Bath (تعداد حمام کامل)

همانطور که انتظار می‌رفت، افزایش تعداد حمام‌ها به‌طور مستقیم باعث افزایش قیمت خانه می‌شود. در این نمودار نیز رابطه خطی نسبتا واضحی دیده می‌شود.

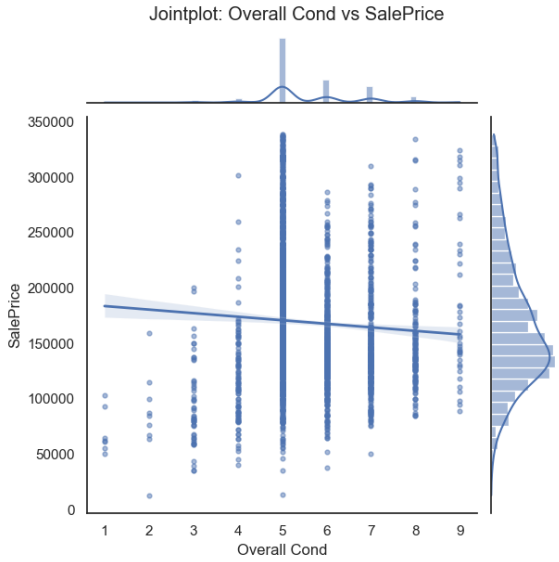


ویژگی‌های با همبستگی منفی:

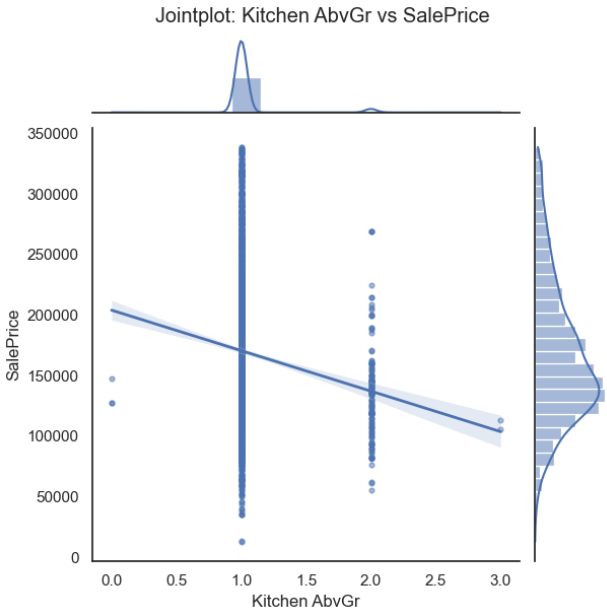
Low Qual Fin SF : هرچه متراژ فضای تکمیل‌شده با کیفیت پایین بیشتر باشد، قیمت کاهش می‌یابد. این فضاها به‌جای ایجاد ارزش، معمولاً علامتی از ساخت ضعیف محسوب می‌شوند.



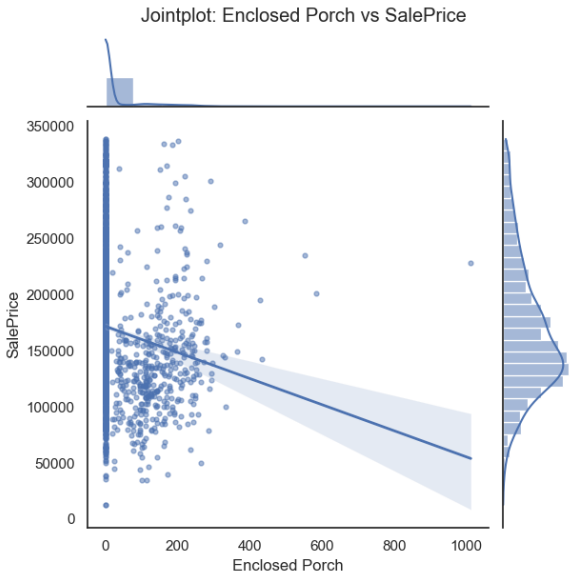
Overall Cond : برخلاف کیفیت کلی، وضعیت کلی خانه اثر منفی جزئی بر قیمت دارد. این موضوع می‌تواند ناشی از این باشد که ارزیابی کیفی (Overall Qual) اهمیت بیشتری دارد و اثر وضعیت (Overall Cond) در حضور آن کمرنگ یا معکوس می‌شود.



Kitchen AbvGr : افزایش تعداد آشپزخانه‌ها (بیش از یک) با کاهش قیمت همراه بوده است. احتمالاً به دلیل این است که وجود چند آشپزخانه بیشتر در خانه‌های کوچک یا غیرمعمول مشاهده می‌شود که بازارپسندی کمتری دارند.



Enclosed Porch : افزایش مساحت ایوان بسته با کاهش قیمت همراه است؛ به‌نظر می‌رسد این فضاها به‌جای ارزش‌افزوده، بخشی از فضای خانه را اشغال کرده و کارایی پایین‌تری دارند.



انتخاب ویژگی‌ها با SelectKBest (روش f_regression)

در این بخش هدف انتخاب مؤثرترین ویژگی‌ها از میان متغیرهای موجود است تا عملکرد مدل‌های رگرسیون بهینه شود. برای این کار از روش SelectKBest همراه با آزمون آماری f_regression استفاده شده است. این روش به هر ویژگی یک امتیاز (F-score) اختصاص می‌دهد که بیانگر قدرت رابطه آن ویژگی با متغیر هدف (قیمت خانه) است. سپس بهترین تعداد ویژگی‌ها (K) با استفاده از اعتبارسنجی متقاطع (Cross-Validation) انتخاب شده است.

نتایج انتخاب تعداد ویژگی‌ها(K)

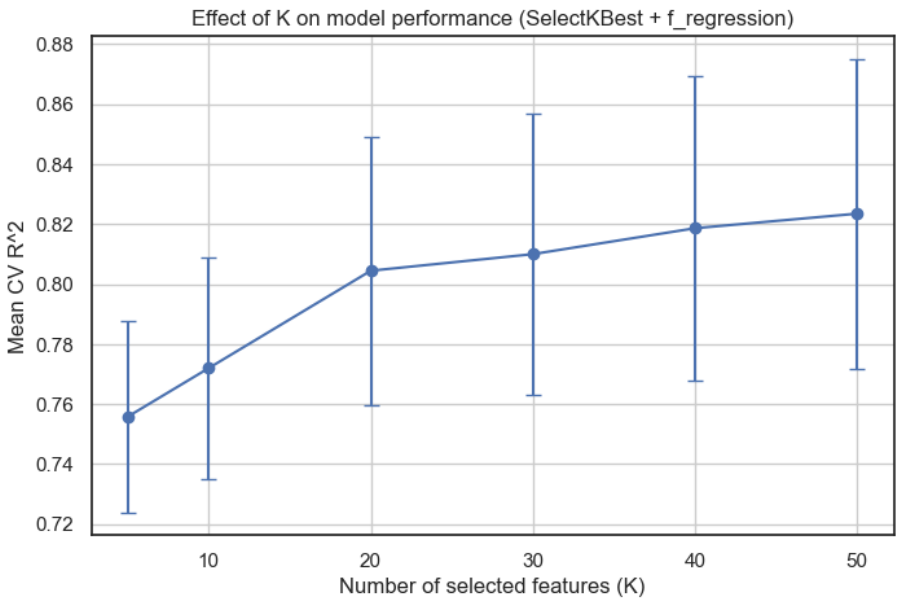
- مقایسه مقادیر مختلف K نشان داد که با افزایش تعداد ویژگی‌های انتخاب‌شده، دقت مدل (معیار R^2) بهبود می‌یابد.
- بهترین مقدار به‌دست آمده برای K=50 بود که منجر به میانگین R^2=0.8235 در داده‌های اعتبارسنجی شد.

- روند تغییرات نشان می‌دهد که اضافه کردن ویژگی‌ها باعث بهبود عملکرد مدل می‌شود، اما پس از حدود 30 ویژگی، شیب بهبود کندتر شده و در نهایت در 50 ویژگی به بیشترین مقدار می‌رسد.

```
Mean CV R^2 by K:
K= 5: R^2 mean=0.7558, std=0.0319
K=10: R^2 mean=0.7720, std=0.0370
K=20: R^2 mean=0.8045, std=0.0447
K=30: R^2 mean=0.8100, std=0.0469
K=40: R^2 mean=0.8186, std=0.0508
K=50: R^2 mean=0.8235, std=0.0516
```

```
Best K via GridSearchCV:
Best K: 50
Best CV R^2: 0.8235
```

نمودار تغییرات میانگین R^2 بر اساس تعداد ویژگی‌ها نشان می‌دهد انتخاب تعداد کم ویژگی (مثلاً 5 یا 10) منجر به مدل ضعیف‌تر و بایاس بالا می‌شود. افزایش تدریجی تعداد ویژگی‌ها دقت مدل را افزایش می‌دهد و خطای سیستماتیک کاهش می‌یابد. در نهایت، انتخاب بیش از 50 ویژگی تفاوت چندانی ایجاد نمی‌کند و ممکن است منجر به افزایش واریانس شود.



ویژگی‌های انتخاب شده (بر اساس F-score) :

```
Selected features (sorted by F-score):
num__Overall Qual                          F-score=4507.71
num__Gr Liv Area                           F-score=2066.60
num__Garage Cars                           F-score=1848.36
num__Garage Area                           F-score=1616.30
cat__Exter Qual_TA                         F-score=1566.04
num__Year Built                            F-score=1510.22
cat__Exter Qual_Gd                         F-score=1415.66
num__Full Bath                             F-score=1310.41
num__Year Remod/Add                        F-score=1288.50
cat__Garage Finish_Unf                     F-score=1242.28
num__Total Bsmt SF                         F-score=1208.44
cat__Kitchen Qual_TA                       F-score=1163.18
cat__Foundation_PConc                      F-score=1133.25
num__1st Flr SF                            F-score=1096.38
cat__Bsmt Qual_TA                          F-score=1064.80
num__Garage Yr Blt                         F-score=1039.94
cat__Kitchen Qual_Gd                       F-score=834.56
cat__Heating QC_Ex                         F-score=771.92
num__Fireplaces                            F-score=760.33
cat__BsmtFin Type 1_GLQ                    F-score=682.54
num__TotRms AbvGrd                         F-score=629.06
cat__Bsmt Qual_Gd                          F-score=576.94
cat__Bsmt Qual_Ex                          F-score=528.44
cat__Garage Type_Detchd                    F-score=516.72
cat__Garage Finish_Fin                     F-score=469.32
cat__Exterior 1st_VinylSd                  F-score=455.36
cat__Exterior 2nd_VinylSd                  F-score=437.35
num__Mas Vnr Area                          F-score=400.81
cat__Heating QC_TA                         F-score=382.88
cat__MS Zoning_RM                          F-score=355.44
cat__Foundation_CBlock                     F-score=352.55
num__BsmtFin SF 1                          F-score=342.91
cat__Lot Shape_Reg                         F-score=312.96
cat__Kitchen Qual_Ex                       F-score=312.47
num__Open Porch SF                         F-score=309.88
cat__Paved Drive_Y                         F-score=305.81
cat__Neighborhood_NridgHt                  F-score=305.78
cat__Central Air_Y                         F-score=303.42
cat__Central Air_N                         F-score=303.42
cat__Paved Drive_N                         F-score=283.81
num__Wood Deck SF                          F-score=281.85
cat__Garage Type_Attchd                    F-score=276.14
cat__Lot Shape_IR1                         F-score=253.00
num__Half Bath                             F-score=228.58
num__Lot Frontage                          F-score=227.02
cat__Electrical_SBrkr                      F-score=222.61
cat__Sale Type_New                         F-score=221.31
num__2nd Flr SF                            F-score=220.53
cat__Bsmt Exposure_No                      F-score=219.67
cat__Sale Condition_Partial                F-score=216.62
```

تقسیم داده‌ها به آموزش و آزمون (train-test split)

در این مرحله داده‌ها برای ارزیابی مدل‌ها به دو بخش مجزا تقسیم شدند. این کار به‌منظور سنجش عملکرد واقعی مدل و جلوگیری از یادگیری صرفاً بر روی داده‌های آموزشی انجام می‌شود.

- نسبت تقسیم‌بندی : ٪۷۵ داده‌ها برای آموزش و ٪۲۵ برای آزمون در نظر گرفته شد.
- تعداد نمونه‌ها : از مجموع داده‌های پردازش‌شده، تعداد ۲۰۹۴ نمونه در مجموعه آموزش و تعداد ۶۹۹ نمونه در مجموعه آزمون قرار گرفتند.
- تعداد ویژگی‌ها : هر دو مجموعه دارای ۷۸ ویژگی هستند که نشان‌دهندهٔ ثبات ساختار داده پس از تقسیم است.
- برنامه‌ریزی برای ارزیابی : داده‌های آموزش برای یادگیری مدل و داده‌های آزمون برای سنجش دقت پیش‌بینی به کار می‌رود. به این ترتیب، ارزیابی نهایی مدل روی داده‌هایی انجام می‌شود که در فرایند آموزش دیده نشده‌اند.

این تقسیم‌بندی نشان می‌دهد که پایگاه داده به‌صورت متوازن و مناسب آماده آموزش مدل‌های رگرسیون شده است. وجود تعداد بالای نمونه در بخش آموزش باعث می‌شود مدل بتواند روابط بین متغیرها و متغیر هدف (قیمت فروش) را به‌خوبی فراگیرد، در حالی که داده‌های آزمون معیار معتبری برای مقایسه مدل‌ها و انتخاب بهترین روش فراهم می‌آورند. استفاده از بذر تصادفی ثابت نیز موجب بازتولیدپذیری نتایج در اجرای‌های بعدی خواهد شد.

```
Dataset split summary:
X_train shape: (2094, 77)
X_test shape: (699, 77)
y_train shape: (2094,)
y_test shape: (699,)
```

آموزش مدل‌های رگرسیون (خطی، ریج، لاسو، چندجمله‌ای)

در این مرحله چهار مدل رگرسیونی متفاوت با استفاده از کتابخانه scikit-learn روی داده‌های آموزش به کار گرفته شدند. پیش از ورود به جزئیات مدل‌ها، لازم است اشاره شود که برای کنترل ابعاد داده و تمرکز روی ویژگی‌های مؤثر، از روش SelectKBest با معیار f_regression و مقدار بهینه K=50 (یافته‌شده در مرحله ششم) استفاده شد. این کار سبب شد تنها ۵۰ ویژگی مهم وارد فرآیند مدل‌سازی شوند.

برای پیش‌پردازش داده‌ها، ستون‌های عددی به همان صورت اصلی و ستون‌های دسته‌ای با OneHotEncoder (با handle_unknown="ignore") کدگذاری شدند. سپس داده‌ها با StandardScaler نرمال‌سازی شدند تا مقیاس ویژگی‌ها همگن گردد. این روند در همه مدل‌ها ثابت بود تا مقایسه عادلانه صورت گیرد.

مدل اول Linear Regression بود که به عنوان خط پایه (baseline) بدون هیچ منظم‌سازی اجرا شد. مدل دوم Ridge Regression با روش RidgeCV آموزش داده شد؛ در این مدل مقدار α در بازه لگاریتمی 0.001 تا 100جستجو شد و در نهایت مقدار بهینه α=100.0 انتخاب گردید. این نتیجه نشان می‌دهد که داده‌ها برای جلوگیری از بیش‌برازش نیاز به منظم‌سازی قوی داشتند. مدل سوم Lasso Regression با استفاده از LassoCV و جستجو در بازه خطی α بین 1تا 20 آموزش داده شد و مقدار بهینه α=20.0 به دست آمد. این انتخاب نشان‌دهنده تمایل مدل به کاهش ضرایب و انتخاب ویژگی‌های کمتر برای مقابله با پیچیدگی داده است. مدل چهارم Polynomial Regression با درجه ۲ پیاده‌سازی شد که در آن تعاملات و توان دوم ویژگی‌ها تولید و سپس با رگرسیون خطی برازش داده شدند. هدف از این مدل بررسی تأثیر افزایش پیچیدگی بر عملکرد بود.

در مجموع می‌توان گفت نتایج به دست آمده حاکی از آن است که مدل‌های Ridge و Lasso با منظم‌سازی قوی توانستند داده‌ها را بهتر کنترل کنند و پارامترهای بزرگ انتخاب‌شده برای α نشانه وجود نویز و همبستگی بالا میان ویژگی‌ها است. مدل خطی پایه ساده‌ترین حالت را ارائه داد و مدل چندجمله‌ای با درجه ۲ پیچیدگی بیشتری به مدل افزود که در مراحل بعدی، هنگام بررسیbias-variance trade-off ، اهمیت آن روشن‌تر خواهد شد.

```
Fitting Linear Regression...
Fitting RidgeCV...
Best Ridge alpha: 100.0
Fitting LassoCV...
Best Lasso alpha: 20.0
Fitting Polynomial Regression (degree=2)...
```

ارزیابی مدل‌ها (محاسبه R^2 و RMSE)

در این بخش، عملکرد مدل‌های آموزش‌دیده (رگرسیون خطی ساده، Ridge، Lasso و Polynomial درجه۲) بر روی داده‌های آزمون ارزیابی شد. برای این منظور از دو معیار متداول در مسائل رگرسیون استفاده گردید:

ریشه میانگین مربعات خطا(RMSE)

$$\text{RSME} = \sqrt{(\hat{y}_i - y_i)^2 \sum_{i=1}^{n} \frac{1}{n}}$$

این معیار میانگین خطای پیش‌بینی‌ها را در واحد اصلی متغیر هدف (اینجا: دلار) نشان می‌دهد. هرچه مقدار آن کمتر باشد، پیش‌بینی‌های مدل به مقادیر واقعی نزدیک‌تر هستند.

ضریب تعیین(R^2)

$$R^2 = \frac{(\hat{y}_i - y_i)^2 \sum_{i=1} n}{(\bar{y} - y_i)^2 \sum_{i=1} n} - 1$$

این معیار بیان می‌کند که مدل چه نسبتی از تغییرات متغیر وابسته را توضیح می‌دهد. مقدار R^2 بین ۰ و ۱ قرار دارد و هرچه به ۱ نزدیک‌تر باشد، مدل عملکرد بهتری دارد.

Model Performance (higher R^2 is better, lower RMSE is better)

	Model	R2_train	RMSE_train	R2_test	RMSE_test
2	Lasso	0.8439	23146	0.8344	24450
1	Ridge	0.8431	23209	0.8341	24476
0	Linear	0.8440	23143	0.8337	24506
3	PolyDeg2	0.9606	11624	0.6607	35005

نتایج نشان داد که سه مدل خطی عملکرد تقریباً مشابهی دارند. هر سه در داده‌های آزمون توانستند حدود ۸۳٪ از تغییرات قیمت فروش را توضیح دهند و خطای آن‌ها نزدیک به ۲۴ هزار دلار باقی ماند. این شباهت نشان می‌دهد که اضافه‌کردن منظم‌سازی) در Ridge و (Lasso در این مجموعه داده نسبت به رگرسیون خطی ساده برتری محسوسی ایجاد نکرده است، اما به کنترل ضرایب و پایداری کمک کرده است.

در مقابل، مدل Polynomial درجه ۲ اگرچه بر داده‌های آموزش عملکرد عالی داشت، اما روی داده‌های آزمون دچار افت شدید شد. این اختلاف بزرگ نشانه‌ی بیش‌برازش (Overfitting) است؛ یعنی مدل جزئیات داده‌های آموزش را بیش از حد یاد گرفته و توان تعمیم‌دهی آن به داده‌های جدید پایین آمده است.

در جمع‌بندی، می‌توان گفت که مدل‌های خطی ساده، Ridgeو Lasso بهترین تعادل میان دقت و تعمیم‌پذیری را نشان دادند، در حالی که Polynomial Regression به دلیل پیچیدگی بیش از حد گزینه‌ی مناسبی برای این مسئله نبود.

توضیح Bias–Variance trade-off و مثال عملی با چندجمله‌ای‌ها

در این بخش تمرین به مفهوم مهم Bias-Variance Trade-offپرداخته می‌شود که از مبانی اصلی یادگیری ماشین است. بصورت کلی:

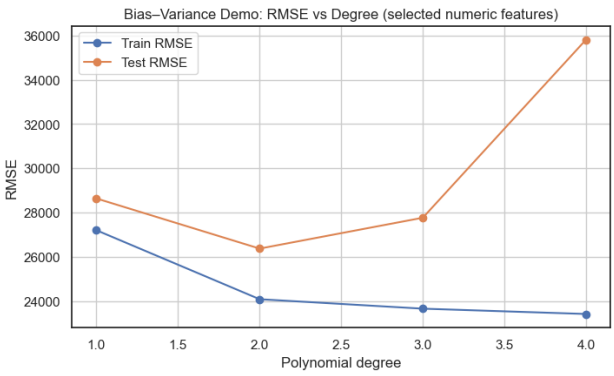
- بایاس (Bias) خطایی است که ناشی از ساده‌سازی بیش از حد مدل است. در این حالت مدل نمی‌تواند الگوهای پیچیده داده‌ها را بیاموزد(Underfitting) .
- واریانس (Variance) خطایی است که به دلیل پیچیدگی بیش از حد مدل و وابستگی شدید به داده‌های آموزش رخ می‌دهد. در این وضعیت مدل روی داده‌های آموزش عملکرد بسیار خوبی دارد اما روی داده‌های آزمون ضعیف عمل می‌کند(Overfitting) .
- هدف یادگیری ماشین یافتن تعادل بهینه بین بایاس و واریانس است تا مدل نه بیش از حد ساده باشد و نه بیش از حد پیچیده.

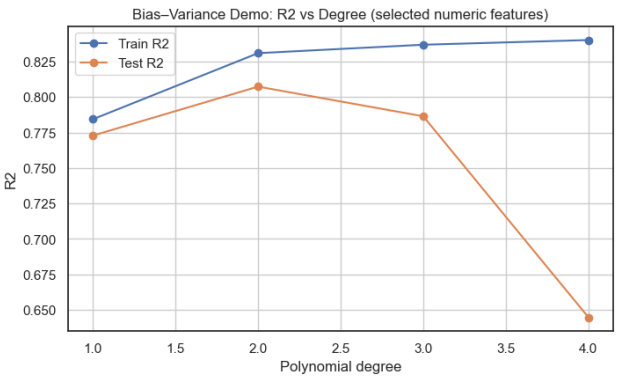
در این تمرین، برای نشان دادن این پدیده از رگرسیون چندجمله‌ای (Polynomial Regression) با درجات مختلف (از ۱ تا ۴) استفاده شد.

جدول عملکرد مدل‌ها:

Degree	R2_train	RMSE_train	R2_test	RMSE_test
1	0.7844	27203.6480	0.7729	28639.3581
2	0.8310	24085.7713	0.8074	26373.3153
3	0.8369	23658.2438	0.7865	27768.4043
4	0.8402	23417.8189	0.6447	35818.0216

در درجه ۱، مدل خطی است. هم خطای آموزش و هم خطای آزمون نسبتاً زیاد است که نشان‌دهنده بایاس بالا (Underfitting) است. در درجه ۲، هر دو معیار R^2 و RMSE بهبود می‌یابند. این نشان می‌دهد که مدل توانسته الگوهای داده را بهتر یاد بگیرد و به یک نقطه تعادل نزدیک شود. در درجه ۳، اگرچه عملکرد آموزش کمی بهتر می‌شود، ولی روی داده‌های آزمون اندکی افت مشاهده می‌شود. این نشانه شروع افزایش واریانس است و در درجه ۴، مدل بیش از حد پیچیده شده و اگرچه روی داده‌های آموزش عملکرد خوبی دارد (R^2_train ≈ 0.84)، ولی روی داده‌های آزمون عملکرد به شدت افت می‌کند (R^2_test ≈ 0.64). این وضعیت واضحاً Overfitting و واریانس بالا را نشان می‌دهد. این آزمایش نشان داد که با افزایش پیچیدگی مدل خطای بایاس کاهش می‌یابد (مدل بهتر الگوهای آموزش را یاد می‌گیرد) اما پس از یک نقطه، خطای واریانس افزایش می‌یابد و دقت روی داده‌های آزمون افت می‌کند. درنتیجه انتخاب مناسب درجه پیچیدگی مدل (اینجا درجه ۲ بهترین حالت بود) برای دستیابی به تعادل بین بایاس و واریانس اهمیت حیاتی دارد.





The End