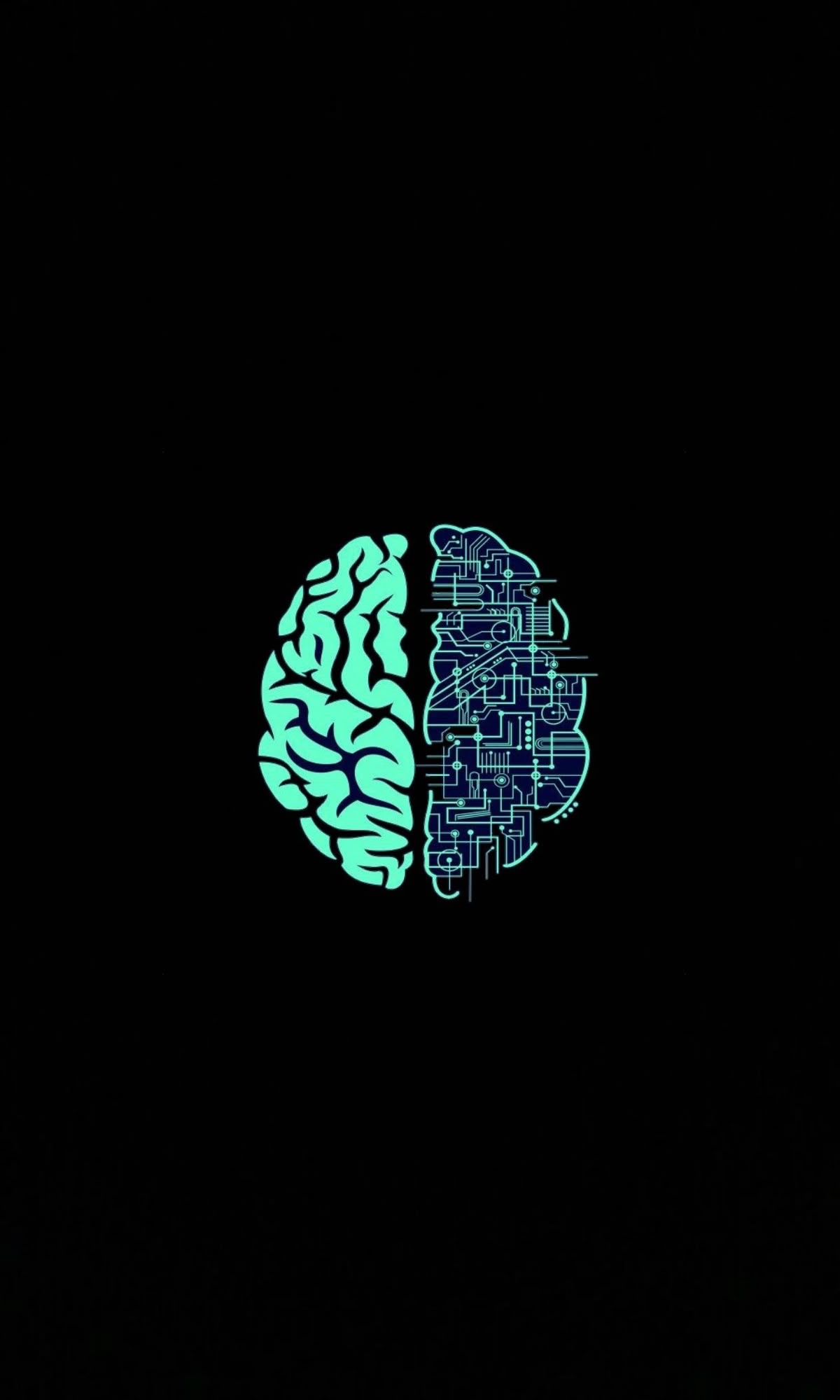
**Artificial Intelligence**

**Machine Learning**

**HW02-Regression**

Name: Masoud Lotfizadeh Sepehri

Student ID: 810603134

Email: [mlotfizadeh@ut.ac.ir](mailto:mlotfizadeh@ut.ac.ir)

Course Instructor**:** Professor Shariatpanahi

**فهرست مطالب**

مقدمه ......................................................................................................................................................................................... 3

بارگذاری و بررسی اولیه داده‌ها (DataFrame و info) ............................................................................................ 4

حذف داده‌های پرت و جایگزینی مقادیر گمشده ............................................................................................................... 5

محاسبه آمار توصیفی (کمینه، بیشینه، انحراف معیار) .................................................................................................... 7

ماتریس همبستگی و شناسایی ویژگی‌های مؤثر بر قیمت خانه ................................................................................. 8

ترسیم Jointplot برای ویژگی‌های با بیشترین همبستگی ...................................................................................... 12

انتخاب ویژگی‌ها با SelectKBest (روش f\_regression) ........................................................................... 17

تقسیم داده‌ها به آموزش و آزمون (train-test split) .......................................................................................... 20

آموزش مدل‌های رگرسیون (خطی، ریج، لاسو، چندجمله‌ای) ..................................................................................... 20

ارزیابی مدل‌ها (محاسبه R² و RMSE) ..................................................................................................................... 21

توضیح Bias–Variance trade-off و مثال عملی با چندجمله‌ای‌ها .............................................................. 23

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.

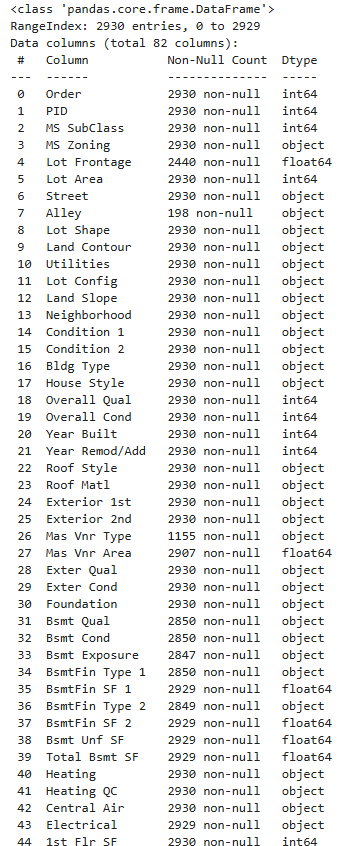
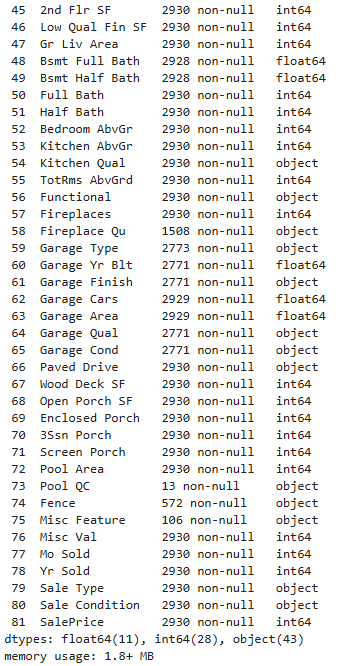
پس از آماده‌سازی داده‌ها، چهار مدل مختلف شامل Linear Regression، Ridge Regression، Lasso Regression و Polynomial Regression آموزش داده شدند و عملکرد آن‌ها با معیارهای R² و RMSE ارزیابی گردید. در نهایت نیز مفهوم Bias-Variance Trade-off بررسی شد تا نشان داده شود که چگونه افزایش پیچیدگی مدل می‌تواند بر بایاس و واریانس تأثیر بگذارد و منجر به بهبود یا افت عملکرد مدل شود.

**بارگذاری و بررسی اولیه داده‌ها (DataFrame و info)**

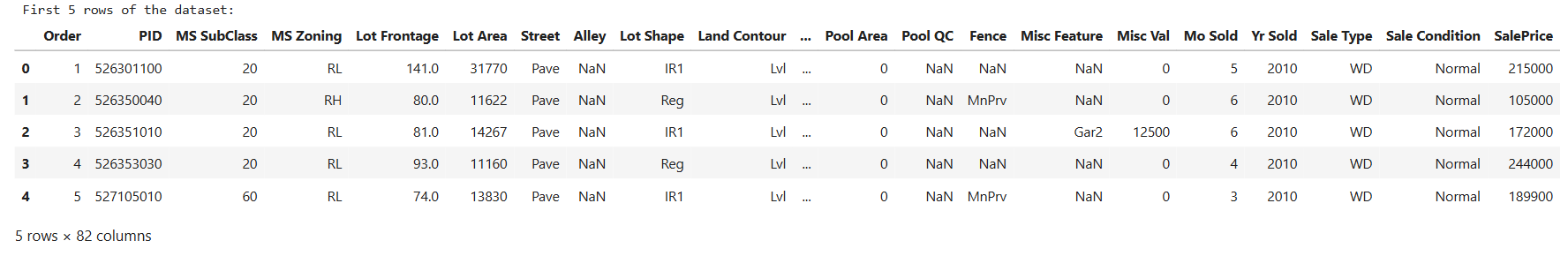
داده مورد استفاده شامل ۲۹۳۰ ردیف و ۸۲ ستون است. حجم داده در حافظه حدود ۱.۸ مگابایت می‌باشد. ستون‌ها از نظر نوع داده به سه گروه تقسیم می‌شوند:

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

ستون SalePrice متغیر 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 Order و PID که نقش توصیفی دارند و برای مدل‌سازی مورد استفاده قرار نمی‌گیرند.



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.

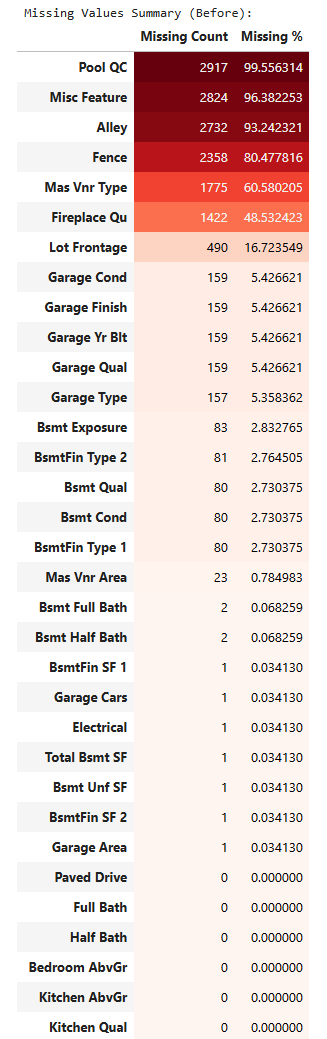
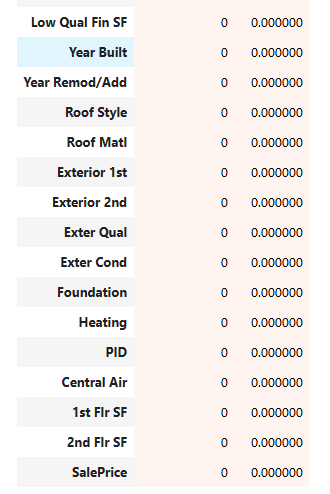
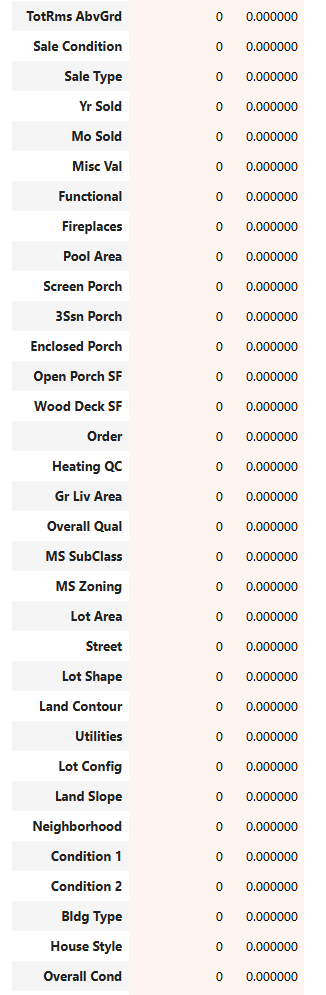


**حذف داده‌های پرت و جایگزینی مقادیر گمشده**

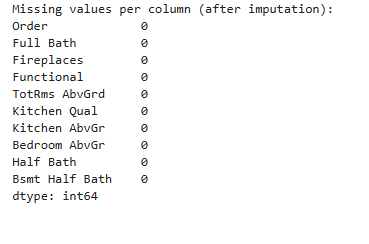
در این مرحله داده‌ها از نظر وجود مقادیر گمشده (Missing Values) بررسی شدند. نتایج اولیه نشان داد که برخی از ستون‌ها دارای درصد بالایی از داده‌های مفقود هستند. به عنوان نمونه:

* 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.



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.



**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.



**ماتریس همبستگی و شناسایی ویژگی‌های مؤثر بر قیمت خانه**

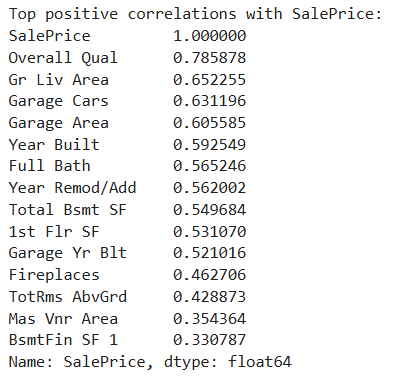
در این مرحله به کمک ماتریس همبستگی، ارتباط بین ویژگی‌های عددی و متغیر هدف یعنی قیمت فروش خانه (SalePrice) بررسی شد. هدف، شناسایی مهم‌ترین عواملی است که بر قیمت خانه اثرگذارند.

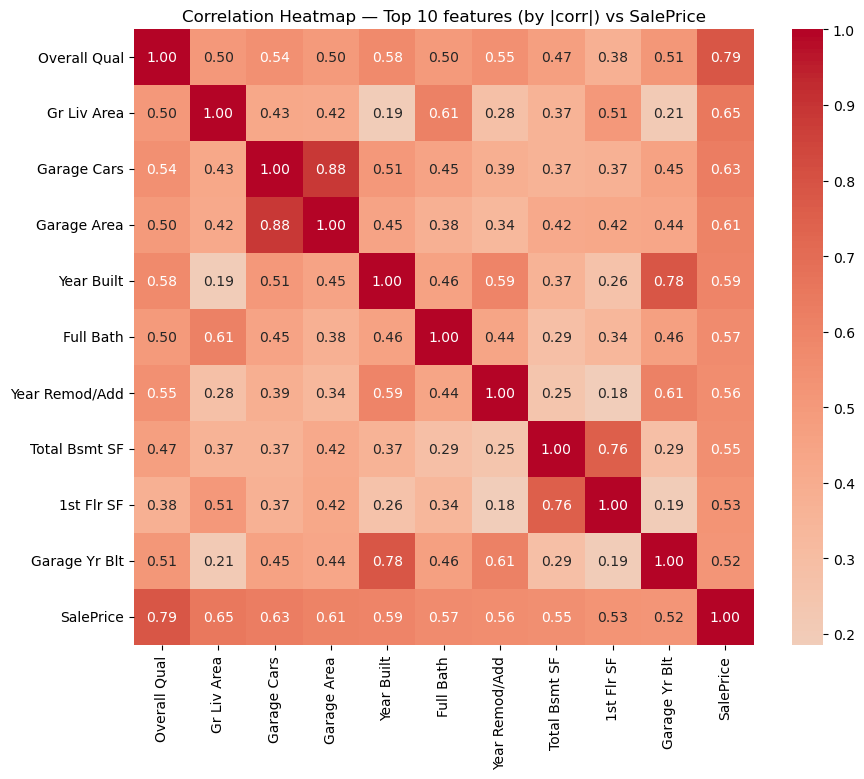
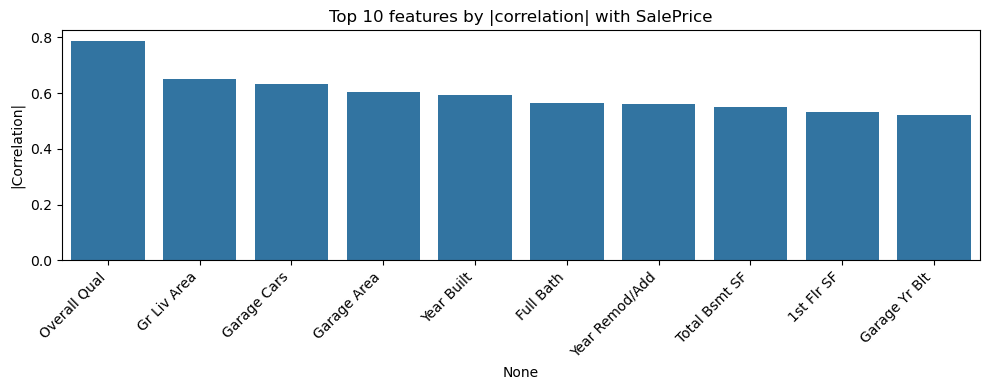
**نتایج همبستگی مثبت**

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 Cars و 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 Built و 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 Bath و Total Bsmt SF نیز 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.



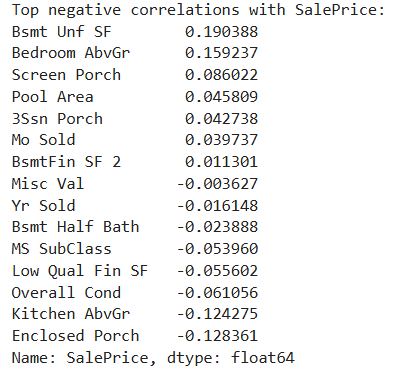


**نتایج همبستگی منفی**

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 SubClass نیز همبستگی منفی ناچیزی دارند.

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



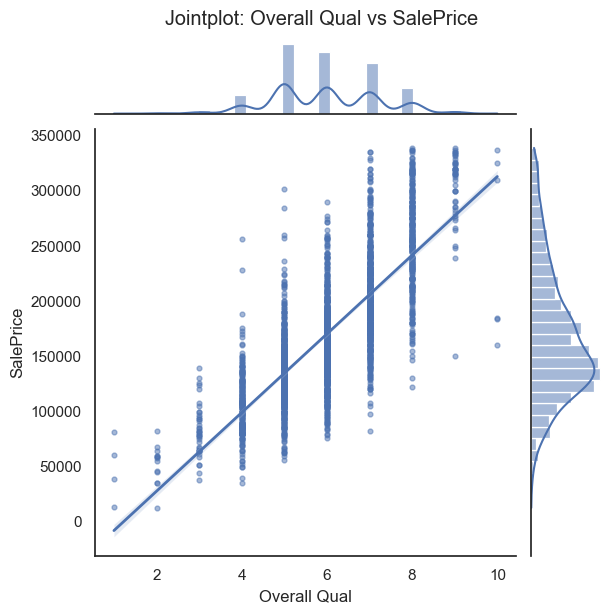
نتایج نشان می‌دهد که ویژگی‌های کیفی و کمی مرتبط با کیفیت ساخت و مساحت زیربنا بیشترین ارتباط را با قیمت فروش دارند. در مقابل، متغیرهای منفی اثر اندکی بر قیمت می‌گذارند و نقش اصلی در مدل‌سازی باید بر متغیرهای مثبت و پرقدرت متمرکز شود.

**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:

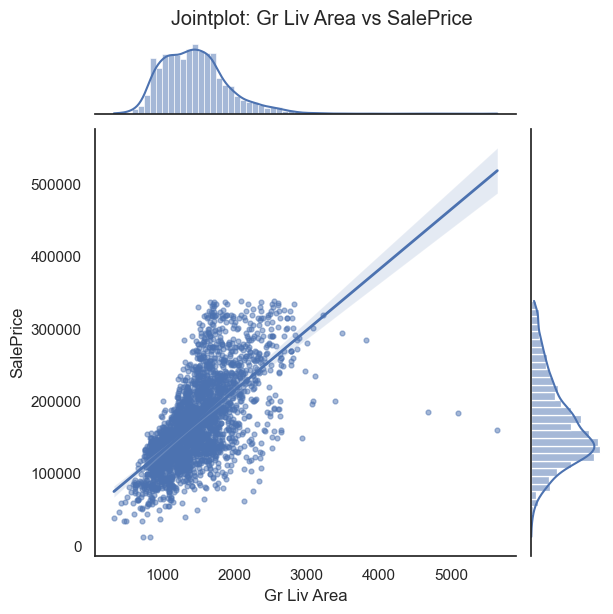
**Overall Qual (کیفیت کلی ساختمان)**

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.



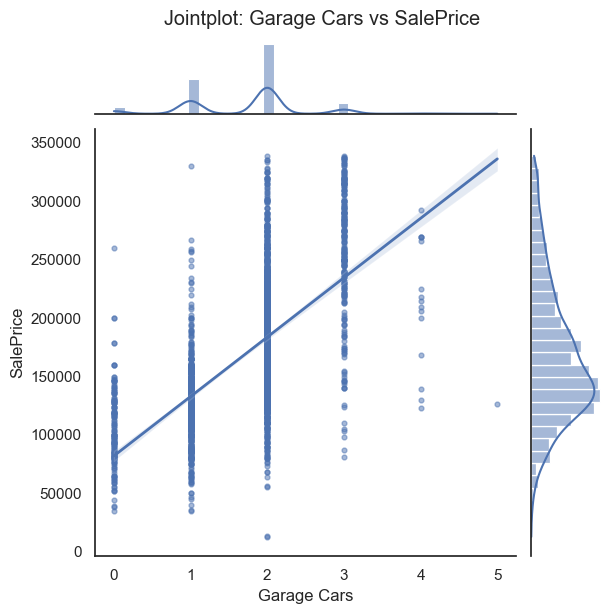
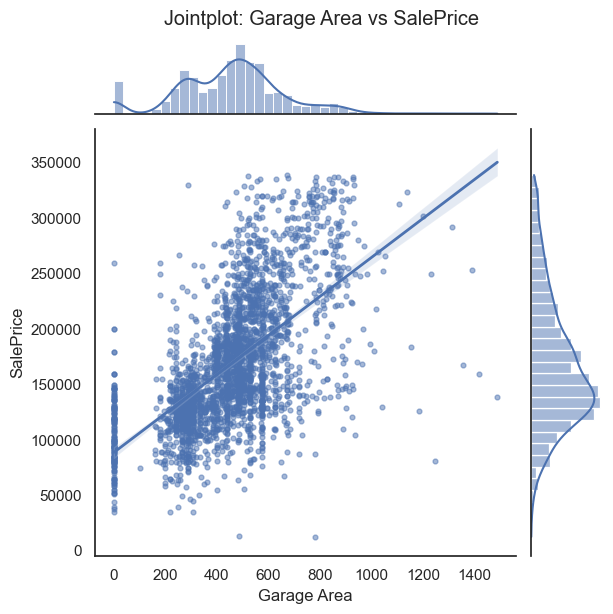
**Gr Liv Area (مساحت زیربنای بالای زمین)**

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

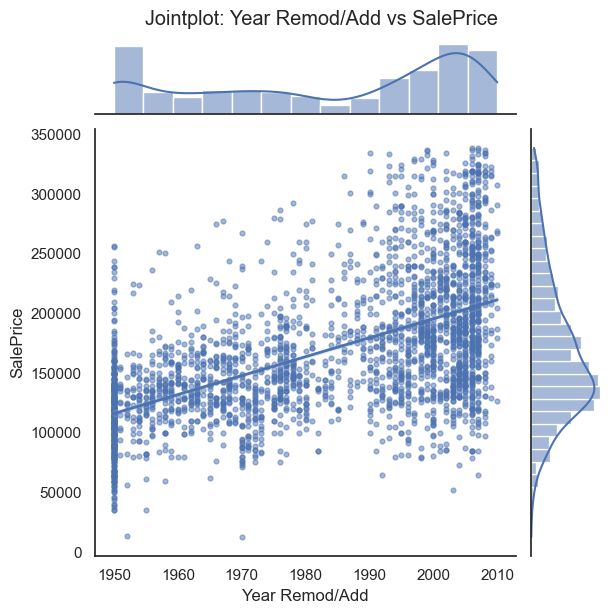
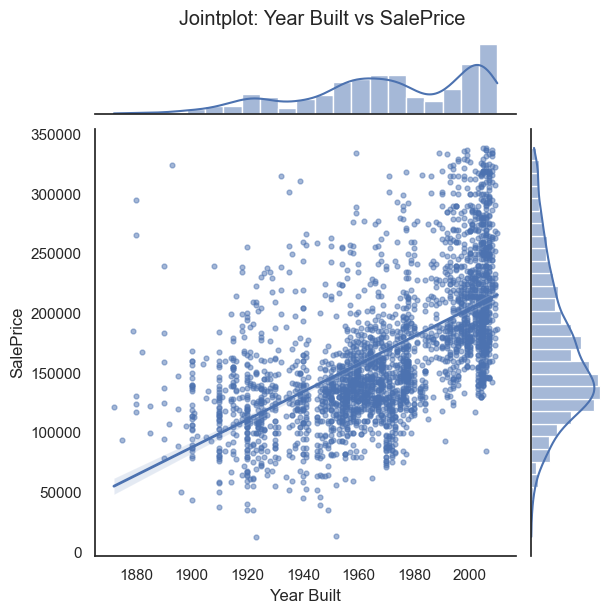


**Garage Cars و Garage Area (تعداد خودروهای قابل پارک و مساحت گاراژ)**

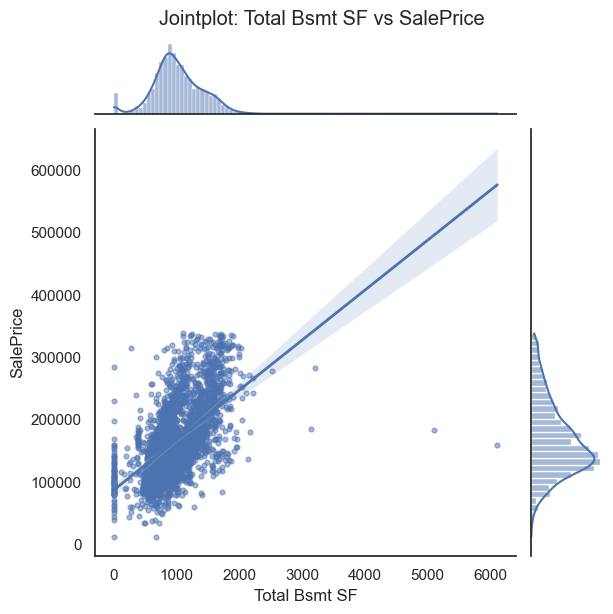
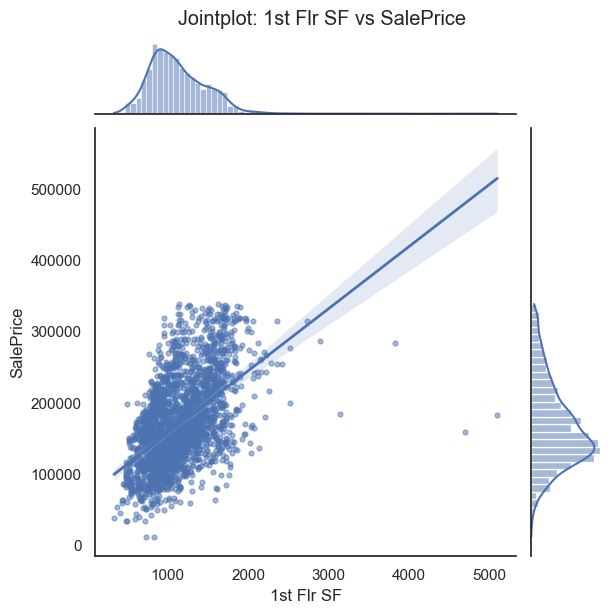
هر دو ویژگی رابطه مثبتی با قیمت دارند. خانه‌هایی با ظرفیت گاراژ بیشتر و گاراژهای بزرگ‌تر، به طور میانگین قیمت بالاتری دارند. این رابطه به‌ویژه در شهرهای کوچک یا محلات مسکونی اهمیت بیشتری پیدا می‌کند.



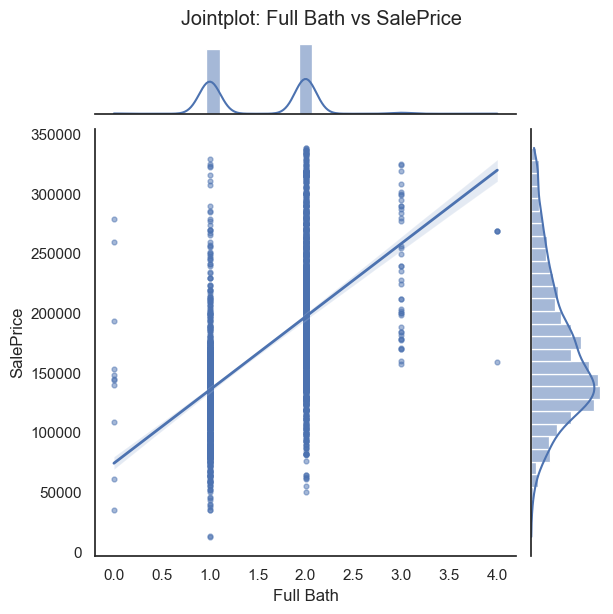
**Year Built و Year Remod/Add (سال ساخت و سال بازسازی)**

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.

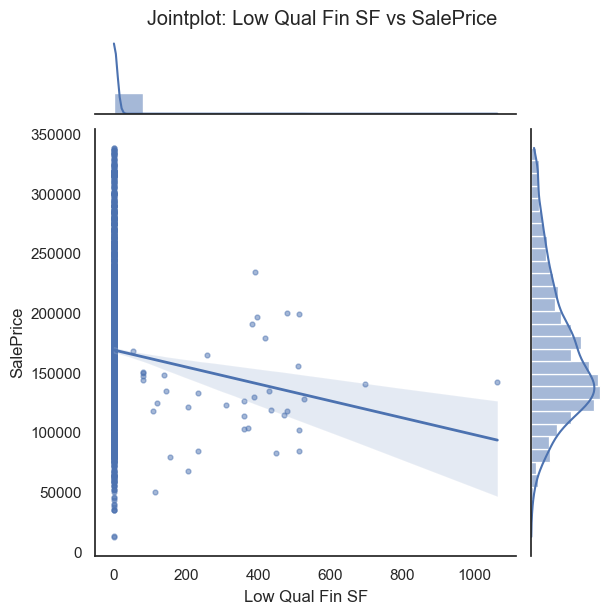


**Full Bath (تعداد حمام کامل)**  
همانطور که انتظار می‌رفت، افزایش تعداد حمام‌ها به‌طور مستقیم باعث افزایش قیمت خانه می‌شود. در این نمودار نیز رابطه خطی نسبتا واضحی دیده می‌شود.

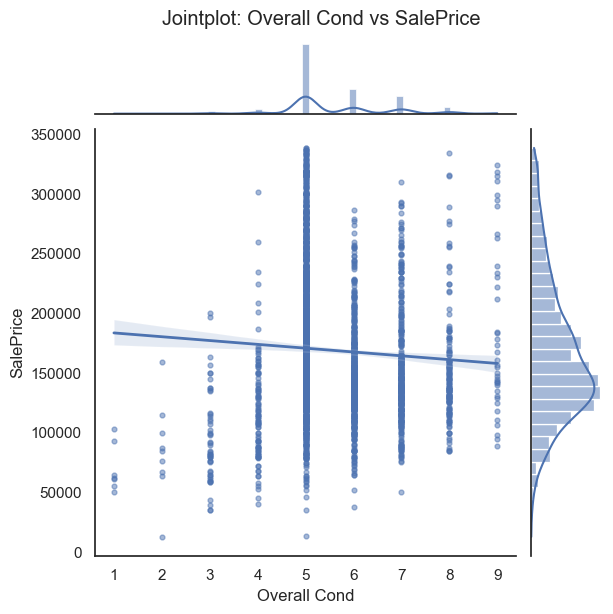
****

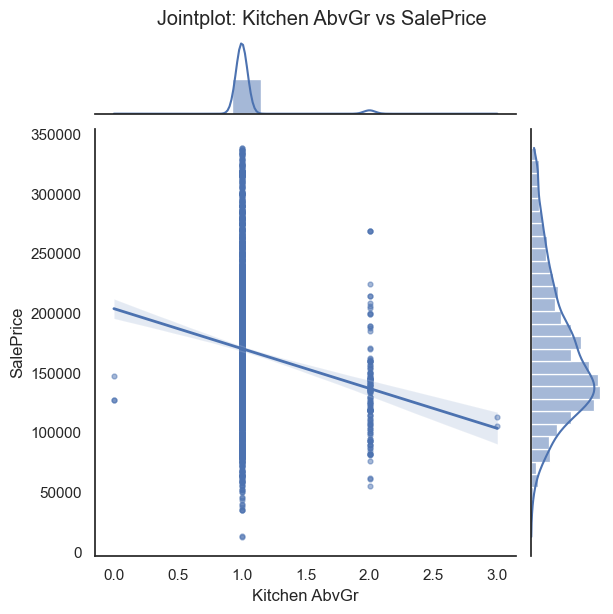
**ویژگی‌های با همبستگی منفی:**

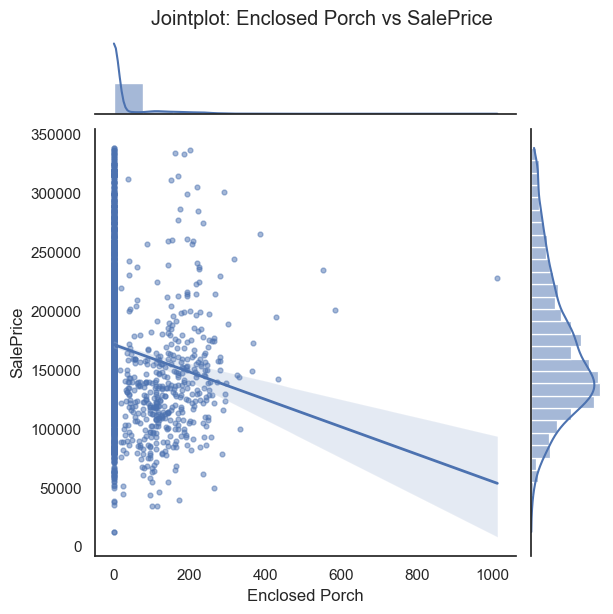
Low Qual Fin SF : هرچه متراژ فضای تکمیل‌شده با کیفیت پایین بیشتر باشد، قیمت کاهش می‌یابد. این فضاها به‌جای ایجاد ارزش، معمولاً علامتی از ساخت ضعیف محسوب می‌شوند.



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.

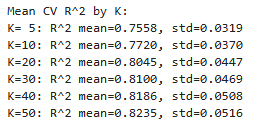
Enclosed Porch **:** افزایش مساحت ایوان بسته با کاهش قیمت همراه است؛ به‌نظر می‌رسد این فضاها به‌جای ارزش‌افزوده، بخشی از فضای خانه را اشغال کرده و کارایی پایین‌تری دارند.

**انتخاب ویژگی‌ها با SelectKBest (روش f\_regression)**

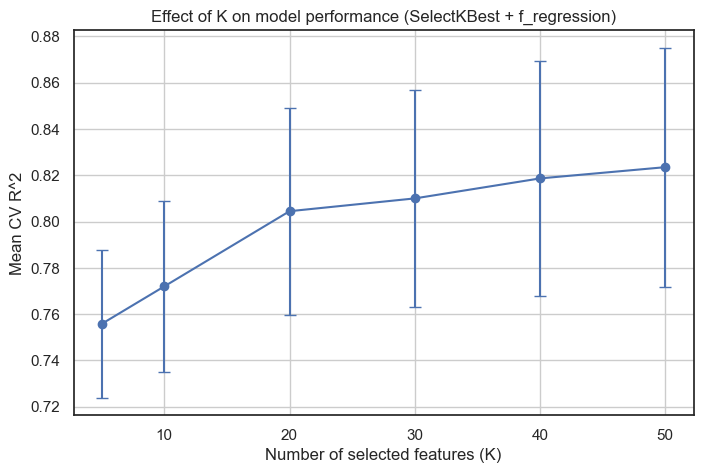
در این بخش هدف انتخاب مؤثرترین ویژگی‌ها از میان متغیرهای موجود است تا عملکرد مدل‌های رگرسیون بهینه شود. برای این کار از روش SelectKBest همراه با آزمون آماری f\_regression استفاده شده است. این روش به هر ویژگی یک امتیاز (F-score) اختصاص می‌دهد که بیانگر قدرت رابطه آن ویژگی با متغیر هدف (قیمت خانه) است.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

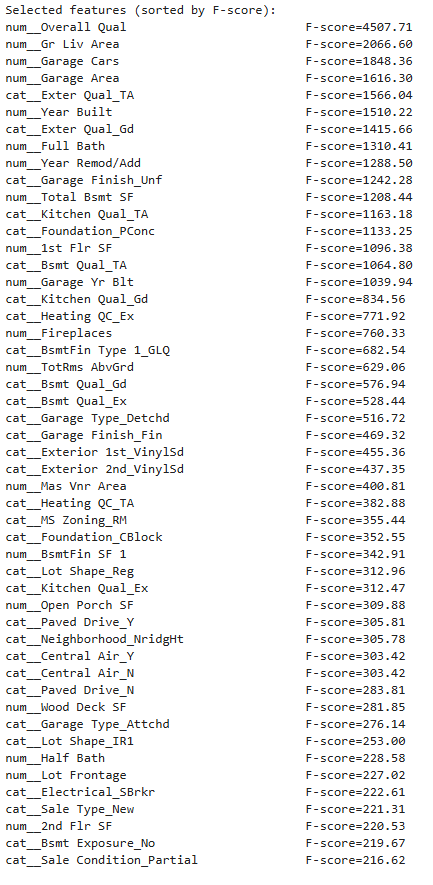
**نتایج انتخاب تعداد ویژگی‌ها (K)**

* مقایسه مقادیر مختلف K نشان داد که با افزایش تعداد ویژگی‌های انتخاب‌شده، دقت مدل )معیار (R2 بهبود می‌یابد.
* بهترین مقدار به‌دست آمده برای K=50 بود که منجر به میانگین R2=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.



نمودار تغییرات میانگین R2 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.



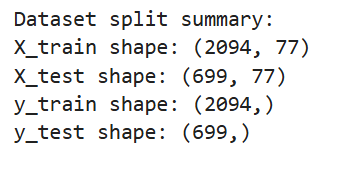
**ویژگی‌های انتخاب شده ( بر اساس F-score) :**

**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.



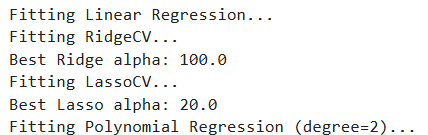
**آموزش مدل‌های رگرسیون (خطی، ریج، لاسو، چندجمله‌ای)**

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 (باhandle\_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 100 جستجو 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 1 تا 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 بررسی bias-variance trade-off، اهمیت آن روشن‌تر خواهد شد.



**ارزیابی مدل‌ها (محاسبه R² و RMSE)**

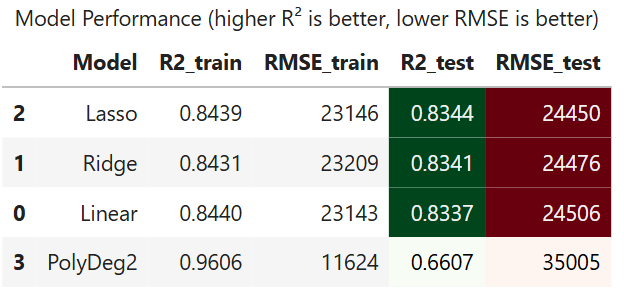
در این بخش، عملکرد مدل‌های آموزش‌دیده (رگرسیون خطی ساده، Ridge، Lasso و Polynomial درجه‌۲) بر روی داده‌های آزمون ارزیابی شد. برای این منظور از دو معیار متداول در مسائل رگرسیون استفاده گردید:

**ریشه میانگین مربعات خطا (RMSE)**

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 (R2)**

این معیار بیان می‌کند که مدل چه نسبتی از تغییرات متغیر وابسته را توضیح می‌دهد. مقدار R2 بین ۰ و ۱ قرار دارد و هرچه به ۱ نزدیک‌تر باشد، مدل عملکرد بهتری دارد.



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، Ridge و 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.

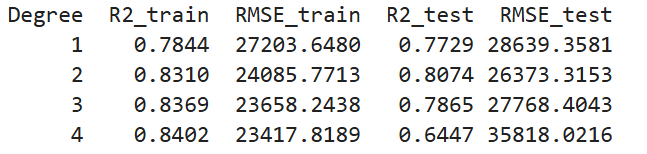
**توضیح Bias–Variance trade-off و مثال عملی با چندجمله‌ای‌ها**

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

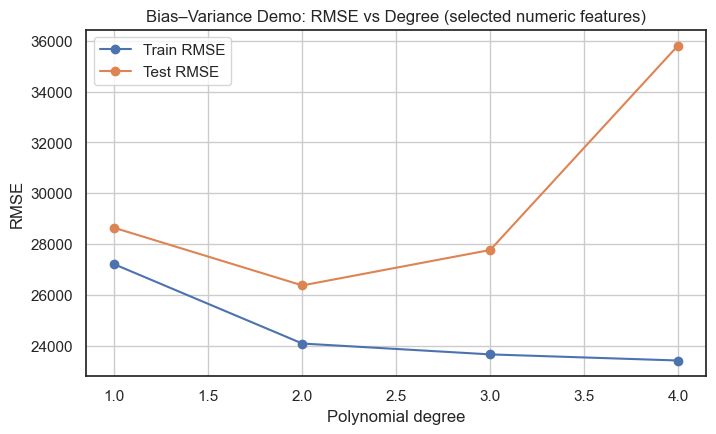
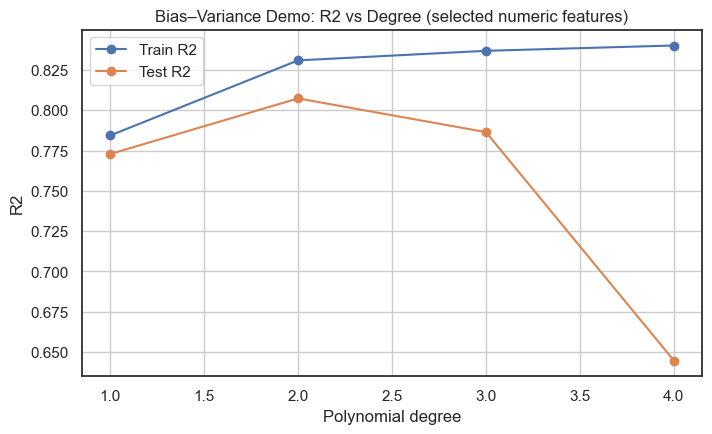
* بایاس (Bias) خطایی است که ناشی از ساده‌سازی بیش از حد مدل است. در این حالت مدل نمی‌تواند الگوهای پیچیده داده‌ها را بیاموزد (Underfitting).
* واریانس (Variance) خطایی است که به دلیل پیچیدگی بیش از حد مدل و وابستگی شدید به داده‌های آموزش رخ می‌دهد. در این وضعیت مدل روی داده‌های آموزش عملکرد بسیار خوبی 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:**



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 R² 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 (R²\_train ≈ 0.84)، ولی روی داده‌های آزمون عملکرد به شدت افت می‌کند .(R²\_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