Data science- The first report

Sadaf Fatollahy

۶ آبان ۱۴۰۲

نام استاد: دکتر خردپیشه
نام درس : مبانی علوم داده

فهرست مطالب

۱.۰ مقدمه

در این تمرین ما قصد داریم ازمون های اماری را بر روی دو دیتاست بررسی کنیم دیتاست اول مربوط به پیش بینی قیمت خانه است.در این دیتاست ابتدا مقداری فرایند های EDA و Preprocessing را انجام میدهیم ، بعد از ان به سوالات داخل تمرین پاسخ داده وتعدادی سوال را خودمان مطرح میکنیم.

۲.۰ درباره دیتاست

هر فیچر نشان دهنده چیست؟

- SalePrice :قیمت فروش ملک به دلار این متغیر هدفی است که می‌خواهید پیش‌بینی کنید.
- MSSubClass :کلاس ساختمان
- MSzoning: طبقه بندی کلی منطقه بندی
- LotFrontage : اندازه خطی خیابان متصل به ملک
- LotArea : اندازه قطعه در فوت مربع
- Street : نوع دسترسی جاده
- Alley :نوع دسترسی به کوچه
- LotShape: شکل کلی ملک
- LandContour : مسطح بودن ملک
- Utilities : نوع تاسیسات موجود
- LotConfig : پیکربندی قطعه
- LandSlope :شیب ملک
- Neighborhood : مکان های فیزیکی در محدوده شهر ایمز
- Condition1 : نزدیکی به جاده اصلی یا راه آهن
- Condition2 : نزدیکی به جاده اصلی یا راه آهن (در صورت وجود ثانیه)
- BldgType : نوع مسکن
- HouseStyle : سبک سکونت
- OverallQual : مواد کلی و کیفیت پایان
- OverallCond : رتبه بندی وضعیت کلی
- YearBuilt : تاریخ ساخت اصلی
- YearRemodAdd : تاریخ بازسازی

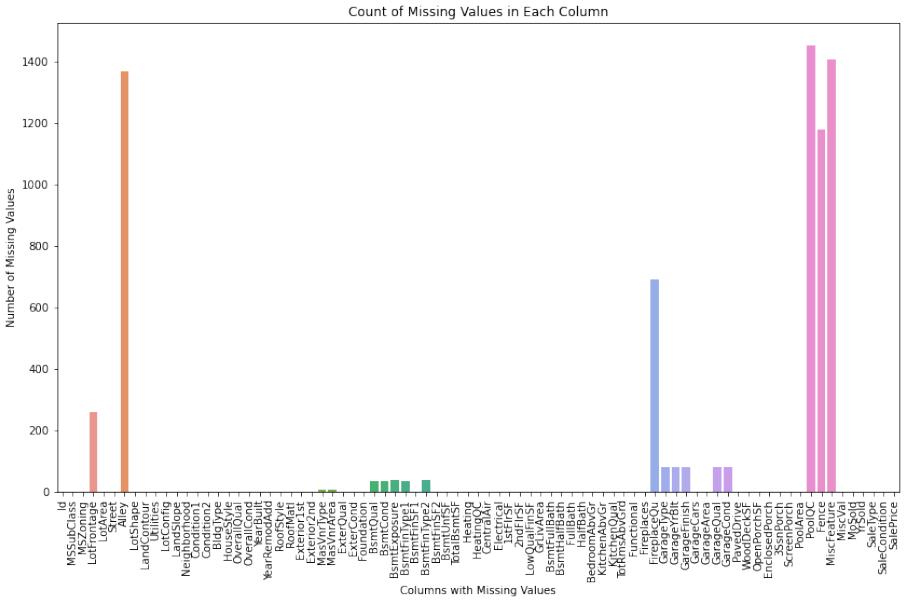
- RoofStyle: نوع سقف
- RoofMatl : مواد سقف
- Exterior1st : پوشش بیرونی خانه
- Exterior2nd : پوشش بیرونی خانه (اگر بیش از یک ماده باشد)
- MasVnrType : نوع روکش بنایی
- MasVnrArea : سطح روکش بنایی در فوت مربع
- ExterQual : کیفیت مواد بیرونی
- ExterCond :وضعیت فعلی مواد در نمای بیرونی
- Foundation : نوع فونداسیون
- BsmtQual : ارتفاع زیرزمین
- BsmtCond : وضعیت عمومی زیرزمین
- BsmtExposure : دیوارهای زیرزمین در سطح باغ یا پیاده روی
- BsmtFinType1 : کیفیت زیرزمین مساحت تمام شده
- BsmtFinSF1 : فوت مربع تمام شده نوع ۱
- BsmtFinType2 : کیفیت منطقه تکمیل شده دوم (در صورت وجود)
- BsmtFinSF2 : فوت مربع تمام شده نوع ۲
- BsmtUnfSF : متر مربع ناتمام مساحت زیرزمین
- TotalBsmtSF: کل متر مربع مساحت زیرزمین
- Heating: نوع گرمایش
- HeatingQC : کیفیت و وضعیت گرمایش
- CentralAir: تهویه مطبوع مرکزی
- Electrical : سیستم الکتریکی
- 1stFlrSF: طبقه اول فوت مربع
- 2ndFlrSF : طبقه دوم فوت مربع
- LowQualFinSF : فوت مربع تمام شده با کیفیت پایین (همه طبقات)
- GrLivArea : بالاتر از درجه (زمین) منطقه نشیمن فوت مربع
- BsmtFullBath: حمام های کامل زیرزمین

- BsmtHalfBath : نیم حمام زیرزمین
- FullBath :حمام کامل بالاتر از درجه
- HalfBath: نیم حمام بالاتر از درجه
- Bedroom : تعداد اتاق خواب بالاتر از سطح زیرزمین
- Kitchen :تعداد آشپزخانه
- KitchenQual : کیفیت آشپزخانه
- TotRmsAbvGrd : مجموع اتاق های بالاتر از درجه (شامل حمام نمی شود)
- Functional : رتبه بندی عملکرد خانه
- Fireplaces : تعداد شومینه
- FireplaceQu : کیفیت شومینه
- GarageType : محل گاراژ
- GarageYrBlt : سال گاراژ ساخته شد
- GarageFinish : نمای داخلی گاراژ
- GarageCars : اندازه گاراژ در ظرفیت ماشین
- GarageArea : اندازه گاراژ در فوت مربع
- GarageQual : کیفیت گاراژ
- GarageCond : وضعیت گاراژ
- PavedDrive : مسیر آسفالت شده
- WoodDeckSF : مساحت عرشه چوبی در فوت مربع
- OpenPorchSF: مساحت ایوان باز در فوت مربع
- EnclosedPorch :محوطه ایوان محصور در فوت مربع
- 3SsnPorch: مساحت ایوان سه فصل به متر مربع
- ScreenPorch :مساحت ایوان پرده در فوت مربع
- PoolArea : مساحت استخر در فوت مربع
- PoolQC : کیفیت استخر
- Fence : کیفیت حصار
- MiscFeature : ویژگی متفرقه که در دسته های دیگر پوشش داده نمی شود

- Value of miscellaneous feature:MiscVal
- MoSold : ماه فروش
- YrSold : سال فروش
- SaleType : نوع فروش
- SaleCondition : شرایط فروش

۳.۰ EDA and pre proccesing

این دیتاست شامل ۱۴۶۰ سطر و ۸۱ ستون فیچر است. در ابتدا مقدار missing value را در این دیتاست بررسی کردیم.



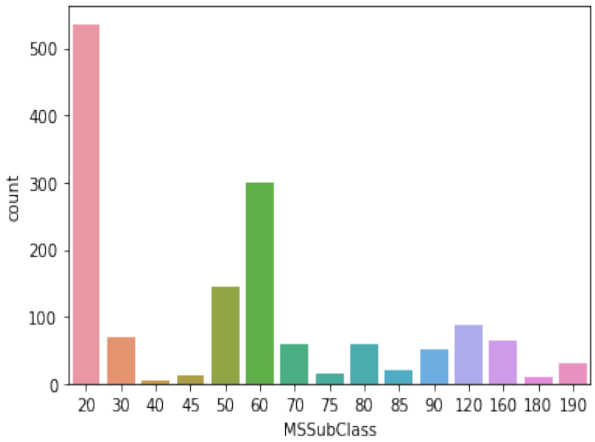
شکل ۱: فیجر های دارای مقدار missing

با توجه به شکل بالا میتوانیم مقدار مقادیر missing را در هر فیچر متوجه شویم. حال باید فکری به حال این مقادیر بکنیم.فیچر های ['MiscFeature','PoolQC','Fence','Alley'] تقریبا دارای ۹۰ درصد مقدار null هستند پس ما این فیچر ها را حذف میکنیم. اما دیگر فیچر ها را بسته به اینکه از نوع عددی هستند یا غیر عددی به ترتیب با مقدار میانگین و مقدار مد جایگزاری کردیم. به این ترتیب دیگر missing value در دیتاست موجود نیست.

۴.۰ سوال

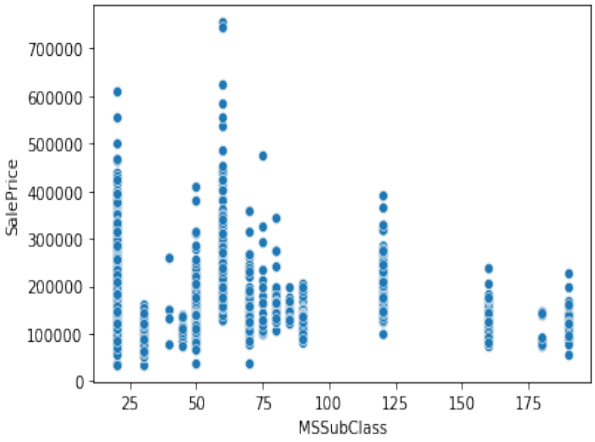
:

۱. فیچر MSSubClass چیست و چرا مهم است؟ در ابتدا نگاهی به این فیچر می اندازیم



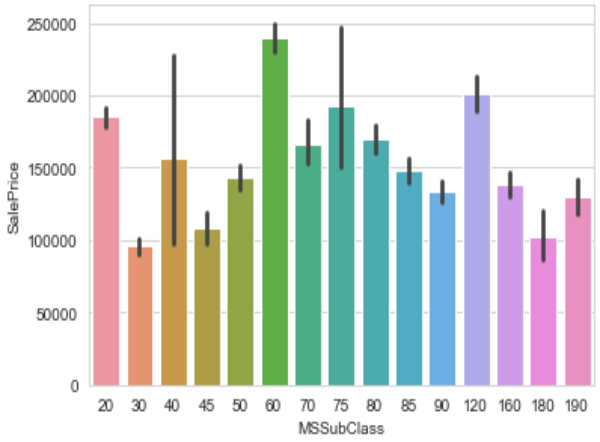
شکل ۲: تعداد کلاس های MSSubClass

این فیچر از نوع عددی و گسسته است که با توجه به شکل کلاس ۲۰ و ۶۰ بیشترین مقدار را دارند.
حال ارتباط ان با target یعنی house price را بررسی میکنیم.



شکل ۳: relation between house price and MSSubclass

با توجه به شکل ارتباط خطی بین این دو فیچر مشاهده نمیشود فقط میتوان نتیجه گرفت که تنوع قیمت خانه در کلاس ۲۰ و ۶۰ بیشتر است.



شکل ۴: The mean sale price for each class in MSSubclass.

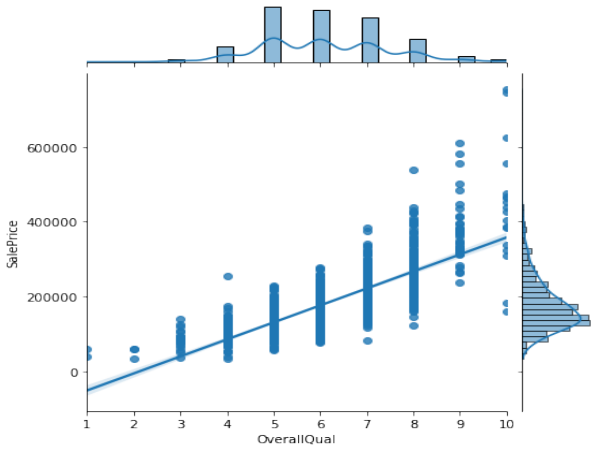
در این شکل نیز میتوان میانگین قیمت فروش خانه را در هر کلاس از این فیچر مشاهده کرد که کلاس ۶۰ بیشترین میانگین فروش را داشته است.

در نهایت ما همبستگی این فیچر را با saleprice بررسی کردیم که مقدار ان 0.08- شد. که این به این معنی است که هنگامی که یک متغیر افزایش می‌یابد، متغیر دیگر کاهش می‌یابد. اما این همبستگی ضعیف است و نمی‌توان از آن به طور قطعی نتیجه گرفت که یک متغیر تغییر می‌کند، دیگری نیز تغییر می‌کند.

۲. کیفیت کلی OverallQual یک خانه چه ارتباطی با قیمت فروش آن دارد؟

برای این منظور ابتدا پلات های زیر را رسم کردیم:

در این پلات توزیع saleprice و overallqual مشخص شده است همانطوز که معلوم است . overallqual توزیع نرمال ندارد اما به نظر میرسد saleprice دارای توزیع نرمال به همراه چولگی وجود داشته باشد .اما به همین شکل اکتفا نمیکنیم و با استفاده از ازمون های اماری فرضیه نرمال بودن این فیچر ها را بررسی میکنیم. نکته بعدی وجود ارتباط خطی بین این دو فیچر در شکل است که باز هم با استفاده از ازمون اماری ان را بررسی کردیم.



شکل ۵: relation between house price and OverallQuall

بررسی نرمال بودن دوتا فیچر ذکر شده با استفاده از ازمون shapiro:

- ازمون shapiro

آزمایش می کند که آیا یک نمونه داده توزیع گاوسی دارد یا خیر.

مفروضات:

مشاهدات در هر نمونه مستقل و به طور یکسان توزیع شده است (iid).

تفسیر:

H_0: نمونه دارای توزیع گاوسی است.

H_1: نمونه توزیع گاوسی ندارد.

نتیجه این ازمون به این صورت بود که هیچ کدام داری توزیع نرمال نیستند.

اما ایا رابطه معناداری بین انها از طریق ازمون فرض اثبات میشود؟

به این منظور از ازمون Spearman استفاده میکنیم. این ازمون زمانی استفاده میشود که دو نتغیر مورد بررسی عددی باشند و توزیع نرمال نداشته باشند.

- ازمون Spearman

آزمایش می کند که آیا یک نمونه داده به صورت یکپارچه قابل تفکیک است یا خیر.

مفروضات:

الف) مشاهدات در هر نمونه مستقل هستند و به طور یکسان توزیع می شوند.

ب) مشاهدات در هر نمونه رتبه بندی می شوند.

تفسیر:

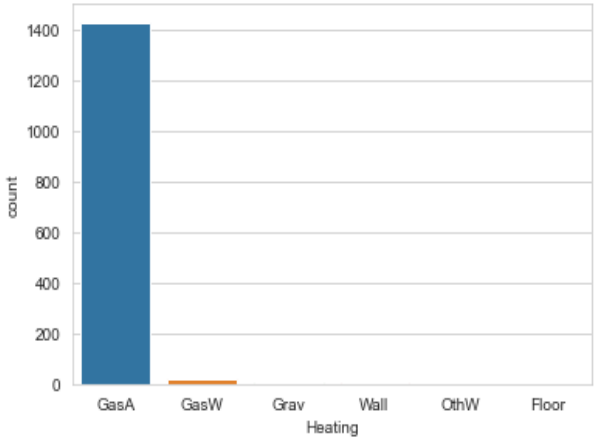
H_0:: نمونه ها همبستگی دارند.

H_1: نمونه هیچ همبستگی ندارد.

این ازمون مشخص کرد که ارتباط بین این دو فیچر وجود دارد.

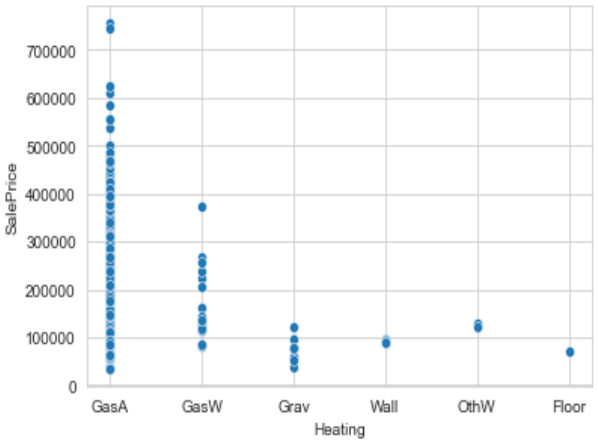
۳. انواع مختلف گرمایش چه تاثیری بر قیمت فروش دارد؟

برای این منظور ابتدا پلات زیر را رسم کردیم:



شکل ۶: The number of heating feature classes

در این پلات ابتدا تعداد انواع سوخت گرمایشی را برای فیچر Heating بررسی کردیم که گرمایشی GasA تقریبا در همه خانه ها مورد استفاده قرار گرفته است. سپس ارتباط ان با sale price را بررسی کردیم.



شکل ۷: relation between house price and Heating

انطور که مشخص است خانه هایی با گرمایشی از نوع GasA تنوع قیمت بیشتری دارند و همچنین گرانترین خانه مربوط به این نوع است از طرفی خانه هایی که گرمایشی انها از نوع Floor بوده است

جز ارزان ترین خانه ها بوده اند. حال این مورد را با تست های اماری بررسی میکنیم. با توجه به نوع فیچر ها از ازمون kruskal استفاده میکنیم.

- ازمونkruskal

این ازمون ارزیابی می‌کند که آیا تفاوت‌های آماری معنی‌داری در قیمت‌های فروش در انواع مختلف گرمایش وجود دارد، بدون اینکه توزیع خاصی را فرض کنیم.

مفروضات:

الف)مشاهدات هر نمونه داده مستقل و توزیع می شود.

ب)مشاهدات را می توان رتبه بندی کرد.

تفسیر:

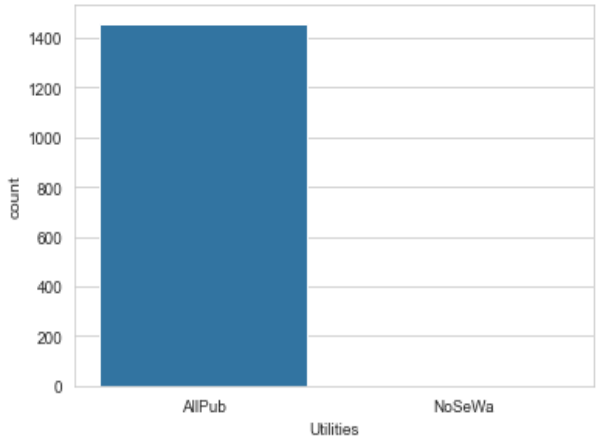
H_0: تفاوت قابل توجهی در قیمت فروش در انواع مختلف گرمایش وجود دارد.

H_1: ممکن است تفاوت قابل توجهی در قیمت فروش در انواع مختلف گرمایش وجود نداشته باشد.

این تست اثبات کرد که تفاوت قابل توجهی در قیمت فروش در انواع مختلف گرمایش وجود دارد.

۴. چگونه انواع مختلف خدمات شهری موجود در یک ملک با قیمت های فروش مرتبط است؟

در ابتدا تعداد هر کلاس از فیچر utilities را بررسی کردیم.

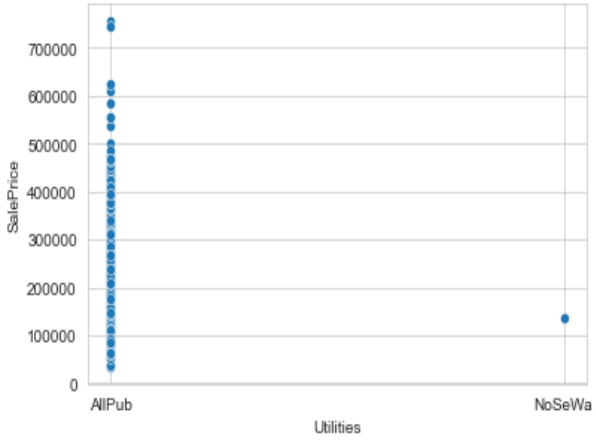


شکل ۸: count of each category in utilities

مشخص شد که فقط یک مورد از نوع NoSeWa وجود دارد و بقیه از نوع AllPub است. حال ارتباط ان با saleprice را بررسی میکنیم.

مشاهده میشود همانطور که توقع میرفت بیشترین فروش از نوع AllPub است. اما ما از نوع NoSeWa نمونه بیشتری نداریم که به طور قطعی نظر بدهیم.حال با تست اماری kruskal ان را

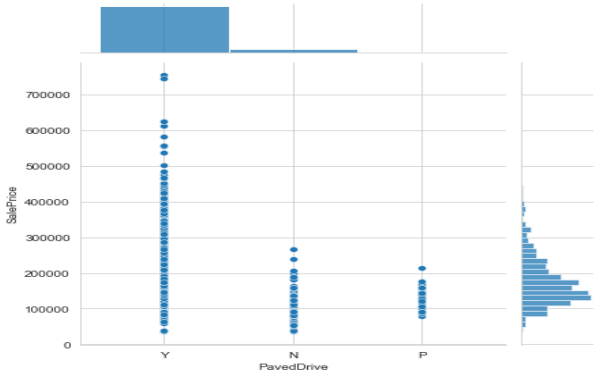
بررسی میکنیم. این تست بیان کرد که ممکن است تفاوت قابل توجهی در قیمت فروش در انواع مختلف خدمات شهری وجود نداشته باشد.



شکل ۹: relation between house price and utilities

۵. ایا مسیر اسفالت شده بر روی قیمت خانه تاثیر به سزایی دارد؟

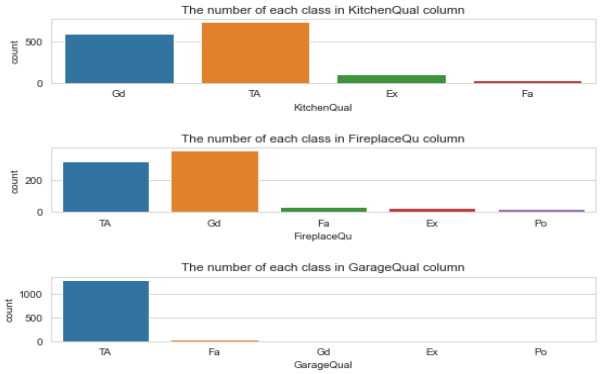
برای پاسخ به این سوال پلات زیر را رسم کردیم که علاوه بر تعداد انواع اسفالت ها را در ان فیچر نشان میدهد ، نشان میدهد که قیمت خانه هایی که اسفالت خیابان انها از نوع p است از همه کمتر است و نوع y تنوع قیمت بیشتری دارد و گران ترین خانه ها نیز دارای این نوع اسفالت هستند .حال با تست اماری kruskal ان را بررسی میکنیم. این تست بیان کرد که تفاوت قابل توجهی در قیمت خانه ها بسته به نوع اسفالت وجود دارد.



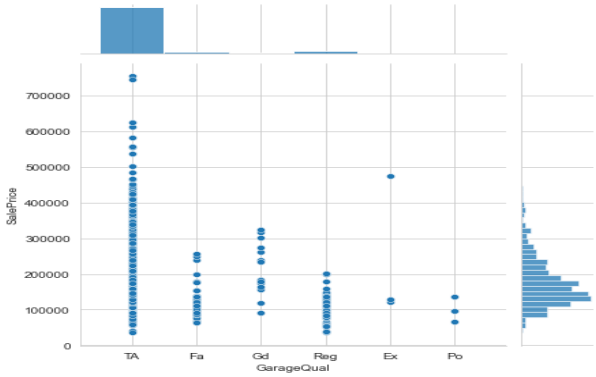
شکل ۱۰: relation between house price and paveddrive

۶. ایا تفاوت قابل توجهی در قیمت خانه ها بر مبنای کیفیت اشپزخانه ، کیفیت گاراژ و کیفیت شومینه ها وجود دارد؟

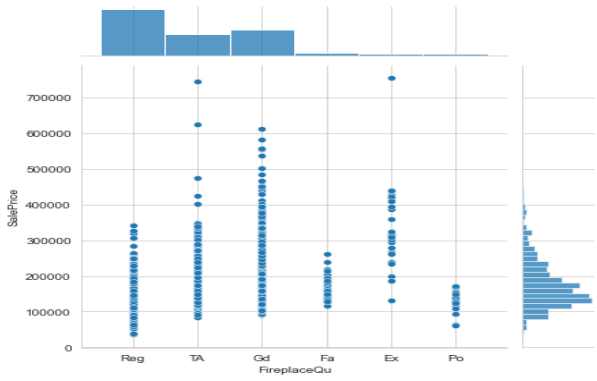
در نگاه اول ممکن است تصور شود که بالا بودن کیفیت هر یک از انها به تنهایی سبب بالارفتن قیمت خانه میشود ، اما با ازمون زیر به این نتیجه میرسیم که هر سه انها باهم سبب افزایش قیمت خانه میشود.
در ابتدا پلات زیر را رسم کردیم که بسته به نوع کیفیت هر کدام از این فیچر تعداد انها را نشان میدهد.



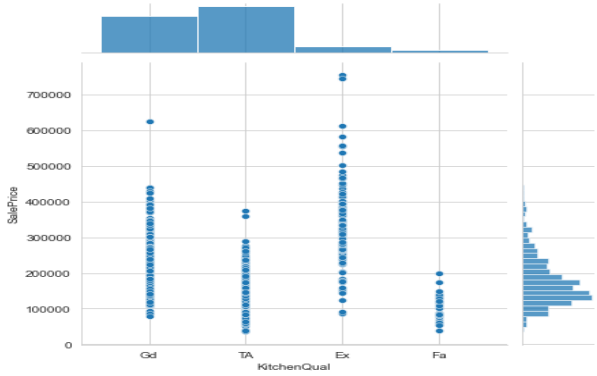
در پلات های زیر قیمت خانه را بر مبنای هر کدام از این سه فیچر بررسی کردیم.



شکل ۱۱: relation between house price and garagequal



شکل ۱۲: relation between house price and fireplacequal

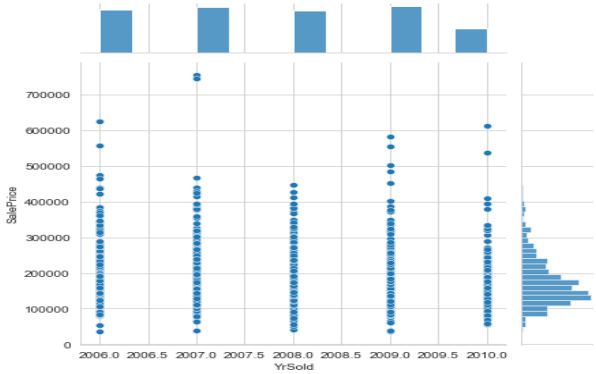


شکل ۱۳: relation between house price and kitchenqual

شاید بتوان دو به دو به این پلات ها ارتباط پیدا کرد اما پیدا کردن ارتباط بین هر سه کمی مشکل است. پس از انکه ازمون kruskal را انجام دادیم فرض ما تایید شد. و هر سه این فیچر ها با هم سبب افزایش قیمت خانه میشود..

۷. ایا تفاوت قابل توجهی در قیمت خانه ها بر مبنای سال فروش خانه وجود دارد؟

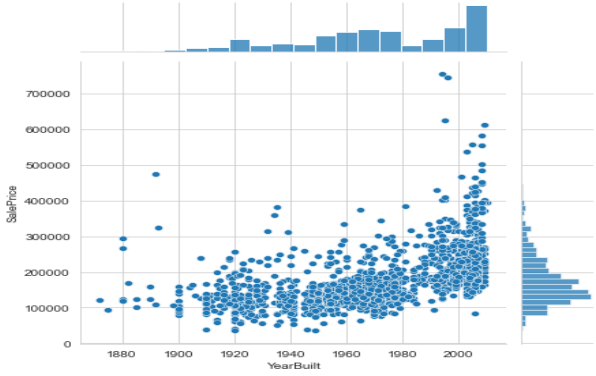
پلات زیر اثبات میکند که تفاوتی وجود ندارد و این به این معنی است در این کشور تورمی وجود ندارد. ازمون Spearman نیز این موضوع را تایید کرد.



شکل ۱۴: relation between house price and yearsold

۸. ایا تفاوت قابل توجهی در قیمت خانه ها بر مبنای سال ساخت خانه وجود دارد؟

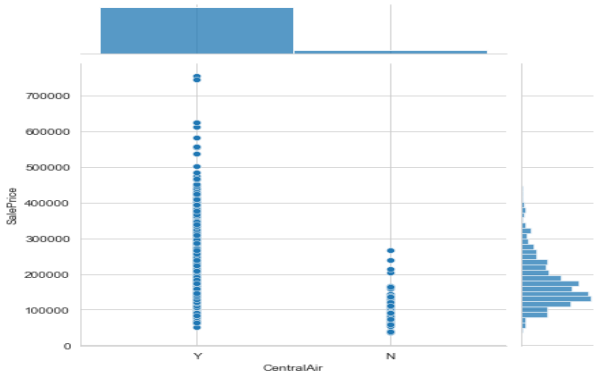
پلات زیر اثبات میکند که تفاوتی اندکی وجود دارد و این به دلیل نوساز بودن خانه هاست. ازمون Spearman نیز این موضوع را تایید کرد. و همبستگی این دو مورد با هم ۵۰٪ بود.



شکل ۱۵: relation between house price and year built

۹. ایا تفاوت قابل توجهی در قیمت خانه ها بر مبنای تهویه مطبوع مرکزی وجود دارد؟

طبیعی است که باید وجود داشته باشد پلات زیر و ازمون kruskal نیز این موضوع را تایید کرد. قیمت خانه هایی که تهویه انها از نوع N است بسیار کمتر است.



شکل ۱۶: relation between house price and central air