## به نام خدا

مرضیه علیدادی – ۸۱۰۱۰۱۲۳۶

تکلیف کامپیوتری ۱ داده کاوی

# سوال ۱:

## ❖ پیش پردازش:

- **بخش ۱:** ۵ سطر ابتدایی دیتاست: (تعدادی از ستون‌ها نمایش داده نشده‌اند.)

| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley | LotShape | LandContour | Utilities | ... | PoolArea | PoolQC | Fence | MiscFeature | MiscVal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 60 | RL | 65.0 | 8450 | Pave | NaN | Reg | Lvl | AllPub | ... | 0 | NaN | NaN | NaN | 0 |
| **1** | 2 | 20 | RL | 80.0 | 9600 | Pave | NaN | Reg | Lvl | AllPub | ... | 0 | NaN | NaN | NaN | 0 |
| **2** | 3 | 60 | RL | 68.0 | 11250 | Pave | NaN | IR1 | Lvl | AllPub | ... | 0 | NaN | NaN | NaN | 0 |
| **3** | 4 | 70 | RL | 60.0 | 9550 | Pave | NaN | IR1 | Lvl | AllPub | ... | 0 | NaN | NaN | NaN | 0 |
| **4** | 5 | 60 | RL | 84.0 | 14260 | Pave | NaN | IR1 | Lvl | AllPub | ... | 0 | NaN | NaN | NaN | 0 |

5 rows × 81 columns

- **بخش ۲**: تعداد مقادیر گمشده در هر ستون بدین صورت است:

| | Total | Percent |
|---|---|---|
| PoolQC | 2909 | 0.996574 |
| MiscFeature | 2814 | 0.964029 |
| Alley | 2721 | 0.932169 |
| Fence | 2348 | 0.804385 |
| FireplaceQu | 1420 | 0.486468 |
| LotFrontage | 486 | 0.166495 |
| GarageFinish | 159 | 0.054471 |
| GarageQual | 159 | 0.054471 |
| GarageCond | 159 | 0.054471 |
| GarageYrBlt | 159 | 0.054471 |
| GarageType | 157 | 0.053786 |
| BsmtExposure | 82 | 0.028092 |
| BsmtCond | 82 | 0.028092 |
| BsmtQual | 81 | 0.027749 |
| BsmtFinType2 | 80 | 0.027407 |
| BsmtFinType1 | 79 | 0.027064 |
| MasVnrType | 24 | 0.008222 |
| MasVnrArea | 23 | 0.007879 |
| MSZoning | 4 | 0.001370 |
| Functional | 2 | 0.000685 |
| Utilities | 2 | 0.000685 |
| BsmtHalfBath | 2 | 0.000685 |
| BsmtFullBath | 2 | 0.000685 |
| GarageArea | 1 | 0.000343 |
| BsmtFinSF1 | 1 | 0.000343 |
| SaleType | 1 | 0.000343 |
| GarageCars | 1 | 0.000343 |
| BsmtUnfSF | 1 | 0.000343 |
| Electrical | 1 | 0.000343 |
| Exterior2nd | 1 | 0.000343 |
| Exterior1st | 1 | 0.000343 |
| KitchenQual | 1 | 0.000343 |
| TotalBsmtSF | 1 | 0.000343 |
| BsmtFinSF2 | 1 | 0.000343 |
| TotRmsAbvGrd | 0 | 0.000000 |
| Fireplaces | 0 | 0.000000 |
| Id | 0 | 0.000000 |
| BedroomAbvGr | 0 | 0.000000 |
| PavedDrive | 0 | 0.000000 |
| WoodDeckSF | 0 | 0.000000 |
| OpenPorchSF | 0 | 0.000000 |
| EnclosedPorch | 0 | 0.000000 |
| 3SsnPorch | 0 | 0.000000 |
| ScreenPorch | 0 | 0.000000 |
| PoolArea | 0 | 0.000000 |
| MiscVal | 0 | 0.000000 |
| MoSold | 0 | 0.000000 |
| YrSold | 0 | 0.000000 |
| SaleCondition | 0 | 0.000000 |
| KitchenAbvGr | 0 | 0.000000 |
| HeatingQC | 0 | 0.000000 |
| HalfBath | 0 | 0.000000 |
| FullBath | 0 | 0.000000 |
| LotArea | 0 | 0.000000 |
| Street | 0 | 0.000000 |
| LotShape | 0 | 0.000000 |
| LandContour | 0 | 0.000000 |
| LotConfig | 0 | 0.000000 |
| LandSlope | 0 | 0.000000 |
| Neighborhood | 0 | 0.000000 |
| Condition1 | 0 | 0.000000 |
| Condition2 | 0 | 0.000000 |
| BldgType | 0 | 0.000000 |
| HouseStyle | 0 | 0.000000 |
| OverallQual | 0 | 0.000000 |
| OverallCond | 0 | 0.000000 |
| YearBuilt | 0 | 0.000000 |
| YearRemodAdd | 0 | 0.000000 |
| RoofStyle | 0 | 0.000000 |
| RoofMatl | 0 | 0.000000 |
| ExterQual | 0 | 0.000000 |
| ExterCond | 0 | 0.000000 |
| Foundation | 0 | 0.000000 |
| Heating | 0 | 0.000000 |
| MSSubClass | 0 | 0.000000 |
| CentralAir | 0 | 0.000000 |
| 1stFlrSF | 0 | 0.000000 |
| 2ndFlrSF | 0 | 0.000000 |
| LowQualFinSF | 0 | 0.000000 |
| GrLivArea | 0 | 0.000000 |
| SalePrice | 0 | 0.000000 |

- **بخش ۳:** در مبحث جایگزینی داده‌های مفقود شده، یکی از روش های پرکاربرد رسیدگی به این مقادیر، در نظر نگرفتن آن هاست. ولی این روش در مواردی، ممکن است خیلی مناسب نباشد.
روش دیگری که پر کاربرد است، استفاده از imputation است. در این روش، داده های مفقود را با یک سری تخمین، با داده های دیگری جایگزین می کنیم. که در این صورت، کل داده ها را برای تحلیل در اختیار داریم و از کل آن ها استفاده می کنیم؛ گویی این داده های تخمین زده شده، واقعا همان داده‌های مشاهده شده هستند.
یک سری روش های معمول، برای این تخمین ها وجود دارد:
**۱.** Mean imputation: میانگین ستون را با استفاده از داده های مشاهده شده ی موجود، محاسبه می کنیم. و این مقدار را جایگزین فیلد های مفقود در آن ستون می کنیم. مزیت این روش، این است که، میانگین داده های آن ستون ثابت باقی می ماند. اما این روش مضرات خیلی خیلی زیادی دارد و نسبت به بقیه‌ی روش ها که در ادامه معرفی می‌شود، بدترین است.
**۲.** Substitution: یک نمونه‌ی جدید از جنس داده‌هایی که داریم، بررسی می‌کنیم. و نتیجه‌ی مشاهدات را جایگزین مقادیر مفقود شده می‌کنیم.
**۳.** Hot deck imputation: بقیه ی سمپل های موجود در دیتاست را در نظر می گیریم. از بین آن ها، آن سمپل هایی که از نظر بقیه‌ی متغیرها، با سمپل مورد نظر ما که دارای داده‌ی مفقود است، مشابه است را، مد نظر قرار می دهیم. یکی از بین آن‌ها به صورت تصادفی انتخاب می کنیم و مقدار موجود در آن سمپل که نظیر داده‌ی مفقود است را انتخاب می کنیم و جایگزین داده‌ی مفقودِ موردنظر می کنیم. یک مزیت این روش این است که همواره از داده های معتبر استفاده خواهیم کرد؛ مثلا اگر یک بازه‌ی مجاز برای متغیری داریم، در این روش، همواره این شرطِ قرار گیری در این بازه، رعایت خواهد شد. مزیت دیگر این است که، استفاده از مولفه‌ی تصادفی بودن، باعث ایجاد تنوع در داده ها می شود.
**۴.** Cold deck imputation: بقیه ی سمپل های موجود در دیتاست را در نظر می گیریم. از بین آن‌ها، آن سمپل‌هایی که از نظر بقیه ی متغیرها، با سمپل مورد نظر ما که دارای داده‌ی مفقود است، مشابه است را، مد نظر قرار می دهیم. یکی از بین آن ها با استفاده از روشی سیستماتیک، انتخاب می کنیم و مقدار موجود در آن سمپل که نظیر داده‌ی مفقود است را انتخاب می کنیم و جایگزین داده‌ی مفقودِ موردنظر می‌کنیم. این روش، مشابه روش قبلی است؛ با این تفاوت که به جای انتخاب تصادفی از بین سمپل ها، انتخابی سیستماتیک خواهیم داشت.
**۵.** Regression imputation: مقدار مفقود شده را برحسب بقیه‌ی مقادیر حدس می زنیم. در این صورت، میانگین ثابت نمی ماند؛ ولی روابط بین متغیرها حفظ می شود.

۶. Stochastic regression imputation: مثل روش قبلی، مقدار مفقود شده را برحسب بقیه‌ی مقادیر حدس می زنیم؛ اما اینجا مولفه‌ی تصادفی بودن را نیز دخیل می کنیم. این روش، مزایای روش قبلی به علاوه‌ی مزایای تصادفی بودن را همزمان دارد.
۷. Interpolation and extrapolation: یک مقدار برای متغیر مفقود شده، با استفاده از مقادیر بقیه‌ی متغیرهای همان سمپل، تخمین می‌زنیم. این دو روش، معمولاً فقط در داده‌های طولی کاربرد دارد. برای استفاده از این روش ها، باید احتیاط کرد. و گاهاً قبل از استفاده از آن ها، نیاز است تا یک سری پیش فرض های اولیه در نظر گرفته شود.

ما به طور کلی دو رویکرد برای imputation داریم. یا می توانیم از یکی از این ۷ روش به تنهایی استفاده کنیم؛ و یا اینکه از ترکیبی از آن ها در کنار هم استفاده کنیم. مزیت روش اول این است که، مفهوم ساده تری دارد و معمولاً داده‌ی تخمین زده شده، در بازه‌ی موردنیاز برای متغیر قرار دارد. و مزیت روش دوم این است که، تخمین دقیق تر و بهتری را نتیجه خواهد داد.

درباره‌ی ستون‌های دارای مقادیر گمشده در دیتاست موردنظر:

- اگر ستون‌ها را به صورت مرتب شده‌ی نزولی براساس تعداد مقادیر گمشده در نظر بگیریم (به همان ترتیبی که در بخش قبل نمایش داده شده‌اند)، ۴ ستون با بیشترین تعداد مقادیر گمشده (یعنی PoolQC، MiscFeature، Alley و Fence)، بیش از ۸۰ درصد مقادیرشان گمشده هستند. ممکن است به نظر برسد که این ستون‌ها عملاً کمکی به تحلیل‌هایمان بر این دیتاست نخواهندکرد و به همین دلیل حذف شوند. ولی با بررسی علت عدم وجود مقادیر گمشده، می‌توان آن‌ها را با مقادیر مناسبی جایگزین کرد و این ستون‌ها را همچنان برای تحلیل دیتاست نگه داشت.
- ستون اول، یعنی PoolQC، ۹۹,۶ درصد مقادیرش گمشده هستند. این ستون، کیفیت استخر خانه‌ی نظیر هر رکورد را نشان می‌دهد. و این مقادیر گمشده، مربوط به خانه‌هایی‌ست که استخر ندارند. داده‌های این ستون از نوع ordinal با ۴ مقدار مختلف هستند. همه‌ی مقادیر گمشده‌ی این ستون را با یک مقدار ثابت جدید، به عنوان دسته‌ی پنجم و در پایین‌ترین سطح درنظر می‌گیریم و مفهوم آن، نداشتن استخر است. از آن به بعد، داده‌های این ستون از نوع ordinal با ۵ مقدار مختلف خواهندبود.
- ستون دوم، یعنی MiscFeature، ۹۶,۴ درصد مقادیرش گمشده هستند. این ستون، امکانات متفرقه‌ای را نشان می‌دهد، که در بقیه‌ی فیلدها، صحبتی از آن‌ها نشده‌است. و این مقادیر گمشده، مربوط به خانه‌هایی‌ست که امکانات اضافه‌ای ندارند. داده‌های این ستون از نوع ordinal با ۵ مقدار مختلف هستند. همه‌ی مقادیر گمشده‌ی این ستون را با یک مقدار ثابت جدید، به عنوان

دسته‌ی ششم و در پایین‌ترین سطح درنظر می‌گیریم و مفهوم آن، نداشتن امکانات اضافه است. از آن به بعد، داده‌های این ستون از نوع ordinal با ۶ مقدار مختلف خواهندبود.

- ستون سوم، یعنی Alley، ۹۳ درصد مقادیرش گمشده هستند. این ستون، نحوه‌ی دسترسی خانه‌ی نظیر آن رکورد به کوچه را نشان می‌دهد. و این مقادیر گمشده، مربوط به خانه‌هایی‌ست که به کوچه دسترسی ندارند. داده‌های این ستون از نوع ordinal با ۲مقدار مختلف هستند. همه‌ی مقادیر گمشده‌ی این ستون را با یک مقدار ثابت جدید، به عنوان دسته‌ی سوم و در پایین‌ترین سطح درنظر می‌گیریم و مفهوم آن، نداشتن دسترسی به کوچه است. از آن به بعد، داده‌های این ستون از نوع ordinal با ۳ مقدار مختلف خواهندبود.
- ستون چهارم، یعنی Fence، ۸۰ درصد مقادیرش گمشده هستند. این ستون، کیفیت حصار خانه‌ی نظیر هر رکورد را نشان می‌دهد. و این مقادیر گمشده، مربوط به خانه‌هایی‌ست که حصار ندارند. داده‌های این ستون از نوع ordinal با ۴ مقدار مختلف هستند. همه‌ی مقادیر گمشده‌ی این ستون را با یک مقدار ثابت جدید، به عنوان دسته‌ی پنجم و در پایین‌ترین سطح درنظر می‌گیریم و مفهوم آن، نداشتن حصار است. از آن به بعد، داده‌های این ستون از نوع ordinal با ۵ مقدار مختلف خواهندبود.
- ستون پنجم، یعنی FireplaceQu، ۴۰ درصد مقادیرش گمشده هستند. این ستون، کیفیت شومینه‌ی خانه‌ی نظیر هر رکورد را نشان می‌دهد. و این مقادیر گمشده، مربوط به خانه‌هایی‌ست که شومینه ندارند و فیلد مربوط به Fireplace شان برابر ۰ است. داده‌های این ستون از نوع ordinal با ۵ مقدار مختلف هستند. همه‌ی مقادیر گمشده‌ی این ستون را با یک مقدار ثابت جدید، به عنوان دسته‌ی ششم و در پایین‌ترین سطح درنظر می‌گیریم و مفهوم آن، نداشتن شومینه است. از آن به بعد، داده‌های این ستون از نوع ordinal با ۶ مقدار مختلف خواهندبود.
- ستون ششم، یعنی LotFrontage، ۱۶,۶ درصد مقادیرش گمشده هستند. این ستون، اندازه‌ی خیابان متصل به خانه‌ی نظیر هر رکورد را نشان می‌دهد. داده‌های این ستون از نوع numeric هستند. باید داده‌های گمشده را به طریق مناسبی تخمین زد. با توجه به این‌که این فیلد، به بقیه‌ی ویژگی‌های خانه، ارتباطی ندارد، لازم نیست از روی رکوردهای مشابه، آن را تخمین زد. می‌توان میانگین مقادیر موجود در این ستون را در این جایگاه‌ها قرار داد، تا میانگین داده‌های این ستون نیز تغییری نکند.
- ستون هفتم، هشتم، نهم و یازدهم، یعنی GarageFinish، GarageQual، GarageCond و GarageType، حدوداً ۵,۴ درصد مقادیرشان گمشده هستند. این ستون‌ها، به ترتیب، وضعیت تکمیل نمای داخلی گاراژ، کیفیت گاراژ، وضعیت گاراژ و مکان گاراژ خانه‌ی نظیر هر رکورد را نشان می‌دهند. این مقادیر گمشده، مربوط به خانه‌هایی‌ست که گاراژ ندارند. داده‌های این ستون‌ها،

از نوع ordinal و به ترتیب، با ۳، ۵، ۵ و ۶ مقدار مختلف هستند. همه‌ی مقادیر گمشده‌ی این ستون‌ها را با یک مقدار ثابت جدید، به عنوان دسته‌ی چهارم، ششم، ششم و هفتم، و در پایین‌ترین سطح درنظر می‌گیریم و مفهوم آن‌ها، نداشتن گاراژ است. از آن به بعد، داده‌های این ستون‌ها از نوع ordinal و به ترتیب، با ۴، ۶، ۶ و ۷ مقدار مختلف خواهندبود.

- ستون دهم، یعنی GarageYrBlt، ۵,۴ درصد مقادیرش گمشده هستند. این ستون، سال ساخت گاراژ خانه‌ی نظیر هر رکورد را نشان می‌دهد. با توجه به اینکه تعداد مقادیر گمشده در این ستون، دقیقاً برابر با تعداد مقادیر گمشده در ۳ ستون قبلی است، که مقادیر گمشده‌ی آن‌ها نشان‌دهنده‌ی عدم وجود گاراژ بود، مقادیر گمشده‌ی این ستون نیز، مربوط به خانه‌هایی‌ست که گاراژ ندارند. داده‌های این ستون از نوع numeric هستند. این مقادیر گمشده با یک مقدار جدید، برای مثال صفر، جایگزین می‌شوند.
- ستون‌های دوازدهم تا شانزدهم، یعنی BsmtExposure، BsmtCond، BsmtQual، BsmtFinType2 و BsmtFinType1، حدوداً ۲,۸ درصد مقادیرشان گمشده هستند. این ستون‌ها، به‌ترتیب، میزان نورگیر بودن زیرزمین، وضعیت کلی زیرزمین، ارتفاع زیرزمین، رتبه‌بندی سطح دوم تکمیل‌شده‌ی زیرزمین(در صورت وجود بیش از یک نوع) و رتبه‌بندی سطح تکمیل‌شده‌ی زیرزمین خانه‌ی نظیر هر رکورد را نشان می‌دهند. این مقادیر گمشده، مربوط به خانه‌هایی‌ست که زیرزمین ندارند. داده‌های این ستون‌ها، از نوع ordinal و به ترتیب، با ۴، ۵، ۵، ۶ و ۶ مقدار مختلف هستند. همه‌ی مقادیر گمشده‌ی این ستون‌ها را با یک مقدار ثابت جدید، به عنوان دسته‌ی پنجم، ششم، ششم، هفتم و هفتم، و در پایین‌ترین سطح درنظر می‌گیریم و مفهوم آن‌ها، نداشتن زیرزمین است. از آن به بعد، داده‌های این ستون‌ها از نوع ordinal و به ترتیب، با ۵، ۶، ۶، ۷ و ۷ مقدار مختلف خواهندبود.
- برای بقیه‌ی ستون‌های دارای مقادیر گمشده، با توجه به این‌که مجموع سطرهایی که این ستون‌ها در آن‌ها دارای مقادیر گمشده هستند، مجموعاً ۳۷ سطر، یعنی حدود ۰,۰۱ درصد کل سطرهاست، این سطرهای دارای مقادیر گمشده حذف می‌شوند. و با توجه به کم بودن تعداد این سطرهای حذف شده، مشکلی برای تحلیل روی دیتاست موردنظر، ایجاد نخواهدشد.

- **بخش ۴**: مقادیر گمشده در این دیتاست، به روشی که در بخش قبل توضیح داده‌شد، برطرف شدند. در نهایت دیتاست مورد به نظر به شکل زیر با تعداد سطر و ستون‌های مشخص شده، برای تحلیل‌هایمان باقی ماند:

| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley | LotShape | LandContour | Utilities | ... | PoolArea | PoolQC | Fence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 60 | RL | 65.0 | 8450 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **1** | 2 | 20 | RL | 80.0 | 9600 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2** | 3 | 60 | RL | 68.0 | 11250 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **3** | 4 | 70 | RL | 60.0 | 9550 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **4** | 5 | 60 | RL | 84.0 | 14260 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2914** | 2915 | 160 | RM | 21.0 | 1936 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2915** | 2916 | 160 | RM | 21.0 | 1894 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2916** | 2917 | 20 | RL | 160.0 | 20000 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2917** | 2918 | 85 | RL | 62.0 | 10441 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | MnPrv |
| **2918** | 2919 | 60 | RL | 74.0 | 9627 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |

2882 rows × 81 columns

- **بخش ۵**: از دو روش برای شناسایی و حذف داده‌های پرت استفاده شد:

**۱.** روش اول: محا سبه‌ی چارک اول و سوم و مقدار inter quartile range و ا ستفاده از منطق موجود در نمودار جعبه ای، برای تشـخیص مقادیر با فاصـله‌ی بیش از ۱,۵ برابر inter quartile range با چارک اول و چارک سوم، به عنوان مقادیر پرت:
در این روش، ابتدا برای هر یک از سـتون‌های با مقادیر عددی، مقادیر پرت، به شـرح گفته شـده تشـخیص داده شـد؛ سـپس تمامی رکوردهایی که حداقل، مقدار یکی از فیلدهای موجود در آن‌ها، مقدار پرت بود، حذف شدند.

**۲.** روش دوم: مقادیر نرمال شده با ا ستفاده از روش z-score، برای هر یک از ستون‌های با مقادیر عددی، در نظر گرفته شد؛ و سپس تمامی رکوردهایی که حداقل، مقدار یکی از فیلدهای موجود در آن‌ها، در حالت نرمال شده، بیشتر از ۳ یا کمتر از ۳- بود، حذف شدند.

در روش اول، ن سبت به روش دوم، تعداد داده‌های بی شتری به عنوان داده‌های پرت ت شخیص داده شدند. برای مثال، تعداد داده‌های پرت شناخته شده برای ستون MSSubClass، در روش اول، برابر ۲۰۷ مقدار ا ست. و در روش دوم، برابر ۱۱ مقدار ا ست. دو شکل زیر، به ترتیب، داده‌های پرت ت شخیص داده شده تو سط روش اول برای این ستون، و داده‌های پرت ت شخیص داده شده تو سط روش دوم برای این ستون را نمایش می‌دهند:

| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley | LotShape | LandContour | Utilities | ... | PoolArea | PoolQC | Fence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **9** | 10 | 190 | RL | 50.0 | 7420 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **48** | 49 | 190 | RM | 33.0 | 4456 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **56** | 57 | 160 | FV | 24.0 | 2645 | Pave | Pave | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **75** | 76 | 180 | RM | 21.0 | 1596 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | GdWo |
| **87** | 88 | 160 | FV | 40.0 | 3951 | Pave | Pave | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2910** | 2911 | 160 | RM | 21.0 | 1484 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2912** | 2913 | 160 | RM | 21.0 | 1533 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2913** | 2914 | 160 | RM | 21.0 | 1526 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | GdPrv |
| **2914** | 2915 | 160 | RM | 21.0 | 1936 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2915** | 2916 | 160 | RM | 21.0 | 1894 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |

207 rows × 81 columns

| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley | LotShape | LandContour | Utilities | ... | PoolArea | PoolQC | Fence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **75** | 76 | 180 | RM | 21.000000 | 1596 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | GdWo |
| **472** | 473 | 180 | RM | 35.000000 | 3675 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **489** | 490 | 180 | RM | 21.000000 | 1526 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **1039** | 1040 | 180 | RM | 21.000000 | 1477 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **1297** | 1298 | 180 | RM | 35.000000 | 3675 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **1452** | 1453 | 180 | RM | 35.000000 | 3675 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **1890** | 1891 | 180 | RM | 35.000000 | 3675 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2243** | 2244 | 180 | RM | 21.000000 | 1974 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2550** | 2551 | 180 | RM | 35.000000 | 3675 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2602** | 2603 | 180 | RM | 69.305795 | 1533 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2864** | 2865 | 180 | RM | 35.000000 | 3675 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |

11 rows × 81 columns

در نهایت، تعداد ســـطرهای باقی مانده در روش اول، برابر با ۱۰۲۹ ســـطر، و در روش دوم، برابر با ۱۸۵۵ سطر است. دو شکل زیر، به ترتیب، سطرهای باقی مانده پس از اعمال روش اول، و سطرهای باقی مانده پس از اعمال روش دوم را نمایش می‌دهند:

| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley | LotShape | LandContour | Utilities | ... | PoolArea | PoolQC | Fence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 60 | RL | 65.0 | 8450 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2** | 3 | 60 | RL | 68.0 | 11250 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **4** | 5 | 60 | RL | 84.0 | 14260 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **6** | 7 | 20 | RL | 75.0 | 10084 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **10** | 11 | 20 | RL | 70.0 | 11200 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2890** | 2891 | 50 | RM | 75.0 | 9060 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2898** | 2899 | 20 | RL | 70.0 | 9116 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2902** | 2903 | 20 | RL | 95.0 | 13618 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2907** | 2908 | 20 | RL | 58.0 | 10172 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2918** | 2919 | 60 | RL | 74.0 | 9627 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |

1029 rows × 81 columns

| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley | LotShape | LandContour | Utilities | ... | PoolArea | PoolQC | Fence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 60 | RL | 65.0 | 8450 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2** | 3 | 60 | RL | 68.0 | 11250 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **4** | 5 | 60 | RL | 84.0 | 14260 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **6** | 7 | 20 | RL | 75.0 | 10084 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **10** | 11 | 20 | RL | 70.0 | 11200 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2906** | 2907 | 160 | RM | 41.0 | 2665 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2907** | 2908 | 20 | RL | 58.0 | 10172 | Pave | NoAlleyAccess | IR1 | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2912** | 2913 | 160 | RM | 21.0 | 1533 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2915** | 2916 | 160 | RM | 21.0 | 1894 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |
| **2918** | 2919 | 60 | RL | 74.0 | 9627 | Pave | NoAlleyAccess | Reg | Lvl | AllPub | ... | 0 | NoPool | NoFence |

1855 rows × 81 columns

✓ در نتیجه، برای این‌که اطلاعات دیتا ست موردنظر، آ سیب نبیند؛ دیتا ست حا صل از روش دوم، به عنوان دیتاست خواسته شده‌ی بدون داده‌های پرت، در نظر گرفته خواهد شد.

- **بخش ۶:**

بخشی از لیست مقایسه‌ای جفت خانه‌های با این ویژگی: (لیست کامل، در فایل کد وجود دارد.)

```
Id1 SP1  LA1  Id2 SP2  LA2
1 208500 8450 11 129500 11200
1 208500 8450 13 144000 12968
1 208500 8450 15 157000 10920
1 208500 8450 17 149000 11241
1 208500 8450 19 159000 13695
1 208500 8450 31 40000 8500
1 208500 8450 32 149350 8544
1 208500 8450 33 179900 11049
1 208500 8450 37 145000 10859
1 208500 8450 41 160000 8658
1 208500 8450 43 144000 9180
1 208500 8450 44 130250 9200
1 208500 8450 58 196500 11645
1 208500 8450 61 158000 13072
1 208500 8450 64 140000 10300
1 208500 8450 74 144900 10200
1 208500 8450 77 135750 8475
1 208500 8450 78 127000 8635
```

**۱.** مقایسه‌ی بین خانه با id مساوی ۱ و خانه با id مساوی ۱۱:

خانه شماره ۱ نسبت به خانه شماره ۱۱ گران‌تر است، با این‌که متراژ کمتری دارد. با دقت در جزئیات هر یک، می‌توان به دلیل آن پی برد:

- کیفیت کلی خانه شماره ۱، "خوب" براورد شده؛ در حالی که کیفیت کلی خانه شماره ۱۱، "زیر متوسط" براورد شده است.
- سال ساخت خانه شماره ۱ برابر ۲۰۰۳ است؛ در حالی که ساخت خانه شماره ۱۱ برابر سال ۱۹۶۵ است. یعنی خانه‌ی شماره ۱، ۳۸ سال جدیدتر است.
- نوع روکش خانه شماره ۱ از نوع آجر است؛ در حالی که خانه شماره ۱۱، روکش ندارد.
- کیفیت مواد استفاده شده در نمای بیرونی خانه شماره ۱، "خوب" براورد شده؛ در حالی که کیفیت مواد استفاده شده در نمای بیرونی خانه شماره ۱۱، "متوسط" براورد شده است.
- ارتفاع زیرزمین خانه شماره ۱، بین ۹۰ تا ۹۹ اینچ است و "خوب" براورد شده؛ در حالی که ارتفاع زیرزمین خانه شماره ۱۱، بین ۸۰ تا ۸۹ اینچ است و "متوسط" براورد شده است.
- متراژ طبقات اول و دوم خانه شماره ۱، به ترتیب ۸۵۶ و ۸۵۴ فوت است؛ در حالی که خانه شماره ۱۱ فقط یک طبقه دارد و ۱۰۴۰ فوت است.
- کیفیت آ شپزخانه‌ی خانه شماره ۱، "خوب" براورد شده؛ در حالی که کیفیت آ شپزخانه‌ی خانه شماره ۱۱، "متوسط" براورد شده است.
- مساحت گاراژ خانه شماره ۱، برابر ۵۴۸ فوت است؛ در حالی که مساحت گاراژ خانه شماره ۱۱ برابر ۳۸۴ فوت است. گاراژ خانه ۱ بزرگ‌تر است.

| Id | OverallQual | YearBuilt | MasVnrType | ExterQual | BsmtQual | 1stFlrSF | 2ndFlrSF | KitchenQual | GarageArea | LotArea | SalePrice |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7 | 2003 | BrkFace | Gd | Gd | 856 | 854 | Gd | 548.0 | 8450 | 208500 |
| 11 | 5 | 1965 | None | TA | TA | 1040 | 0 | TA | 384.0 | 11200 | 129500 |

**۲.** مقایسه‌ی بین خانه با id مساوی ۱۴۰۳ و خانه با id مساوی ۱۴۲۵:

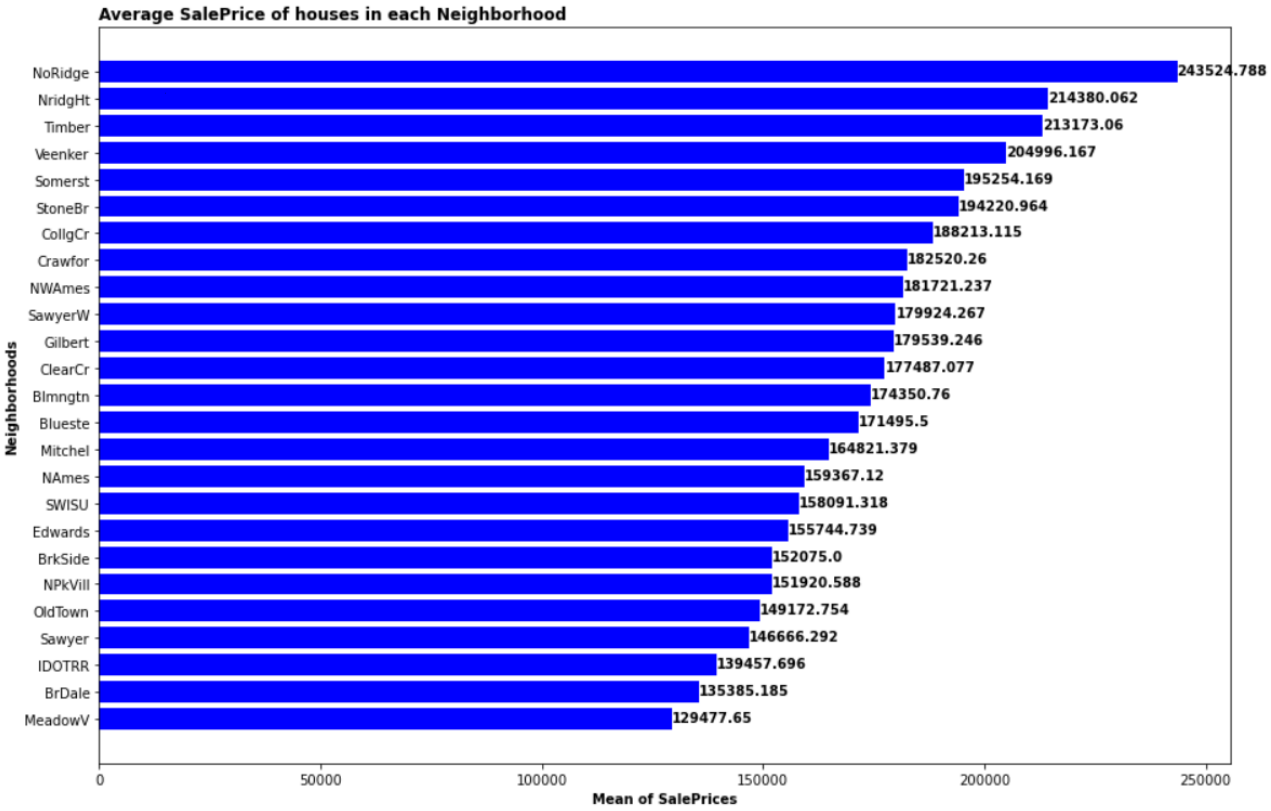
- کیفیت کلی خانه شماره ۱۴۰۳، "خوب" براورد شده؛ در حالی که کیفیت کلی خانه شماره ۱۴۲۵، "زیر متوسط" براورد شده است.
- سال ساخت خانه شماره ۱۴۰۳ برابر ۲۰۰۶ است؛ در حالی که ساخت خانه شماره ۱۴۲۵ برابر سال ۱۹۵۸ است. یعنی خانه‌ی شماره ۱۴۰۳، ۴۸ سال جدیدتر است.
- کیفیت مواد استفاده شده در نمای بیرونی خانه شماره ۱۴۰۳، "خوب" براورد شده؛ در حالی که کیفیت مواد استفاده شده در نمای بیرونی خانه شماره ۱۴۲۵، "متوسط" براورد شده است.
- ارتفاع زیرزمین خانه شماره ۱۴۰۳، بین ۹۰ تا ۹۹ اینچ است و "خوب" براورد شده؛ در حالی که ارتفاع زیرزمین خانه شماره ۱۴۲۵، بین ۸۰ تا ۸۹ اینچ است و "متوسط" براورد شده است.
- وضعیت عمومی زیرزمین خانه شماره ۱۴۰۳، "خوب" براورد شده؛ در حالی که وضعیت عمومی زیرزمین خانه شماره ۱۴۲۵، "متوسط" براورد شده است.
- متراژ زیرزمین خانه شماره ۱۴۰۳، ۱۲۸۶ فوت مربع است؛ در حالی که متراژ زیرزمین خانه شماره ۱۴۲۵، ۱۰۲۴ فوت مربع است. زیرزمین خانه ۱۴۰۳ بزرگ‌تر است.
- وضعیت و کیفیت گرمایش خانه شماره ۱۴۰۳، "عالی" براورد شده؛ در حالی که وضعیت و کیفیت گرمایش خانه شماره ۱۴۲۵، "متوسط" براورد شده است.
- کیفیت آشپزخانه خانه شماره ۱۴۰۳، "خوب" براورد شده؛ در حالی که کیفیت آشپزخانه خانه شماره ۱۴۲۵، "متوسط" براورد شده است.
- کیفیت شومینه خانه شماره ۱۴۰۳، "خوب" براورد شده؛ در حالی که کیفیت شومینه خانه شماره ۱۴۲۵، "متوسط" براورد شده است.
- گاراژ مربوط به خانه شماره ۱۴۰۳، ظرفیت ۲ خودرو را دارد؛ در حالی که گاراژ خانه شماره ۱۴۲۵، ظرفیت ۱ خودرو را دارد.
- مساحت گاراژ خانه شماره ۱۴۰۳، برابر ۶۶۲ فوت است؛ در حالی که مساحت گاراژ خانه شماره ۱۴۲۵ برابر ۴۸۴ فوت است. گاراژ خانه ۱۴۰۳ بزرگ‌تر است.

| Id | OverallQual | YearBuilt | ExterQual | BsmtQual | BsmtCond | TotalBsmtSF | HeatingQC | KitchenQual | FireplaceQu | GarageCars | GarageArea | LotArea | SalePrice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1403 | 7 | 2006 | Gd | Gd | Gd | 1286.0 | Ex | Gd | Gd | 2.0 | 662.0 | 6762 | 193879 |
| 1425 | 5 | 1958 | TA | TA | TA | 1024.0 | TA | TA | TA | 1.0 | 484.0 | 9503 | 144000 |

- ✓ به طور کلی، متراژ خانه، تنها عامل موثر بر قیمت آن خانه نیست. بلکه پارامترها و امکانات دیگر، نظیر مواردی که برای دو مثال بالا ذکر شد، نیز بر قیمت خانه تاثیرگذار است.

## ❖ نمایش دادگان:

- **بخش ۱:** نمودار میانگین قیمت خانه‌ها در محله‌ها به صورت نزولی:




- **بخش ۲:**

✓ اگر منظور از ۵ محله با گران‌ترین خانه‌ها، محله‌هایی باشد که بیشترین میانگین قیمت خانه‌ها را داشته باشند، ۵ محله‌ی اول در نمودار بخش قبل، محله‌های مورد نظر هستند. یعنی به ترتیب:

1. NoRidge → Northridge
2. NridgHt → Northridge Heights
3. Timber → Timberland
4. Veenker → Veenker
5. Somerst → Somerset

اگر منظور از ۵ محله با گران‌ترین خانه‌ها، محله‌هایی باشد که گران‌ترین خانه‌ها در آن‌ها قرار دارند، این ۵ محله‌ی مورد نظر به ترتیب عبارتند از:

| | Id | SalePrice | Neighborhood |
|---|---|---|---|
| **1924** | 1925 | 381000 | Timber |
| **2545** | 2546 | 378500 | Edwards |
| **2430** | 2431 | 378500 | NAmes |
| **1388** | 1389 | 377500 | Gilbert |
| **1482** | 1483 | 377426 | Gilbert |
| **2883** | 2884 | 377426 | Crawfor |
| **336** | 337 | 377426 | StoneBr |
| **2338** | 2339 | 375000 | Somerst |
| **481** | 482 | 374000 | NridgHt |
| **2652** | 2653 | 372402 | NridgHt |

1. Timber → Timberland
2. Edwards → Edwards
3. NAmes → North Ames
4. Gilbert → Gilbert
5. Crawfor → Crawford

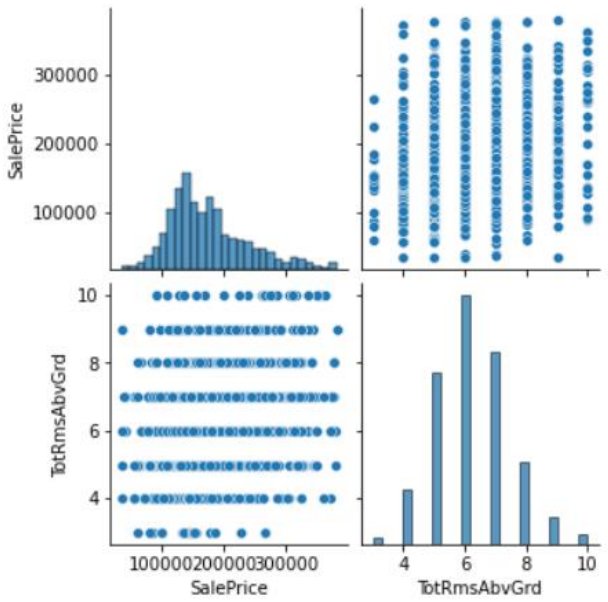
✓ ۵ محله با بیشترین تعداد خانه‌ها عبارتند از:

```
NAmes      267
CollgCr    235
Somerst    142
Gilbert    138
OldTown    126
Edwards    119
NridgHt    112
Sawyer      89
NWAmes      76
SawyerW     75
BrkSide     73
Mitchel     58
Crawfor     50
Timber      50
IDOTRR      46
NoRidge     33
StoneBr     28
BrDale      27
Blmngtn     25
SWISU       22
MeadowV     20
NPkVill     17
ClearCr     13
Blueste      8
Veenker      6
```

1. NAmes → North Ames
2. CollgCr → College Creek
3. Somerst → Somerset
4. Gilbert → Gilbert
5. OldTown → Old Town

- **بخش ۳:**

✓ نمودار pairPlot برای این دو ستون بدین شکل است:



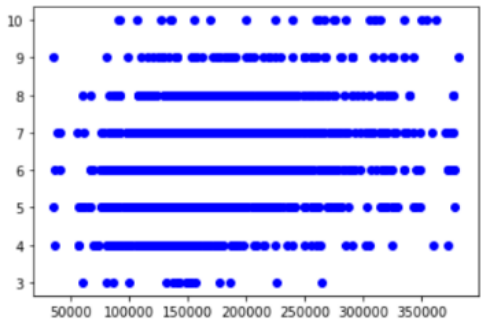

همانطور که مشخص است، نمودار pairPlot دو نوع نمودار برای ما نمایش می دهد: histogram و scatter plot.

نمودار histogram در یک نمودار دوبعدی، طریقه‌ی توزیع یک متغیر واحد را نشان می دهد. و scatter plot رابطه ی بین دو متغیر را نشان می‌دهد.

نمودار بالا سمت راست و نمودار پایین سمت چپ، نمودار scatter plot بین دو متغیر مورد نظر را نشان می‌دهند. می‌توان مشاهده کرد که این دو متغیر تقریبا از هم مستقل هستند. و تعداد اتاق‌های یک خانه، تاثیری بر قیمت آن خانه ندارد.

همچنین، از روی دو نمودار دیگر که histogram هستند، می‌توان دریافت که توزیع متغیر مربوط به تعداد اتاق خواب‌ها از نوع نرمال است، و توزیع قیمت خانه‌ها، دارای کجی راست است.

- ✓ نمودار scatterplot بهترین نمودار برای نمایش تغییرات دو متغیر نسبت به هم است:



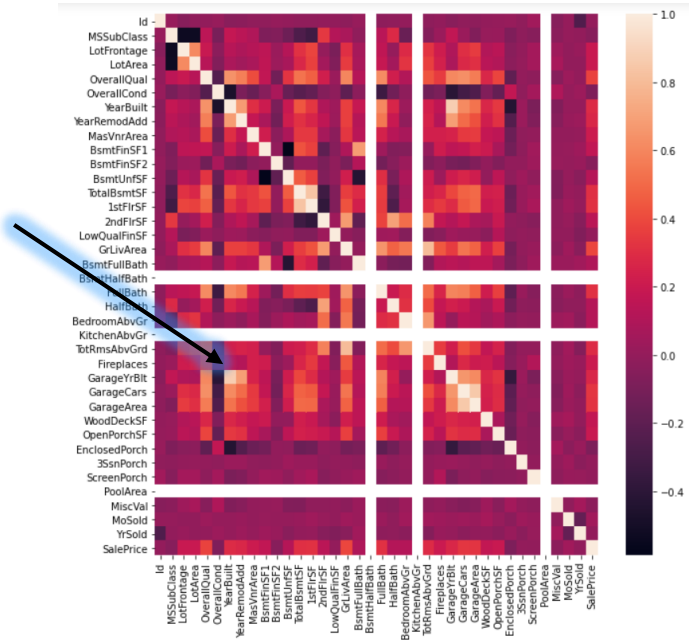

- ✓ همچنین، correlation آن‌ها با استفاده از روش pearson، برابر ۰,۲۹ بدست آمد:

| | SalePrice | TotRmsAbvGrd |
|---|---|---|
| **SalePrice** | 1.000000 | 0.291969 |
| **TotRmsAbvGrd** | 0.291969 | 1.000000 |

این عدد نشان دهنده‌ی این است که این دو متغیر ارتباطی با هم ندارند.

- ✓ به علاوه، correlation آن‌ها با استفاده از روش spearman هم محاسبه شد و برابر ۰,۳ بدست آمد. این روش هم نشان داد که این دو متغیر ارتباطی با هم ندارند.

- **بخش ۴:**




- ✓ نمودار heatmap:

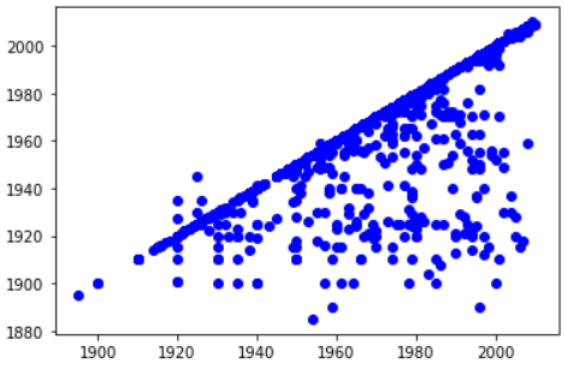
در این نمودار، هرچه مربعِ رنگیِ نشان دهنده‌ی رابطه‌ی دو متغیر، کم‌رنگ‌تر باشد، این دو با هم correlation بیشتری دارند.

در شکل مقابل، کم‌رنگ‌ترین مربع، که با فلش به آن اشاره شده است، نشان‌دهنده‌ی رابطه‌ی بین GarageYrBlt و YearBuilt است. این دو متغیر، بیشترین ارتباط را با هم دارند.

✓ برای هر زوج متغیر، مقدار pearson correlation محا سبه شد. بی شترین عدد مربوط به رابطه‌ی دو متغیر GarageYrBlt و YearBuilt ا ست. این دو متغیر، بی شترین ارتباط را با هم دارند:

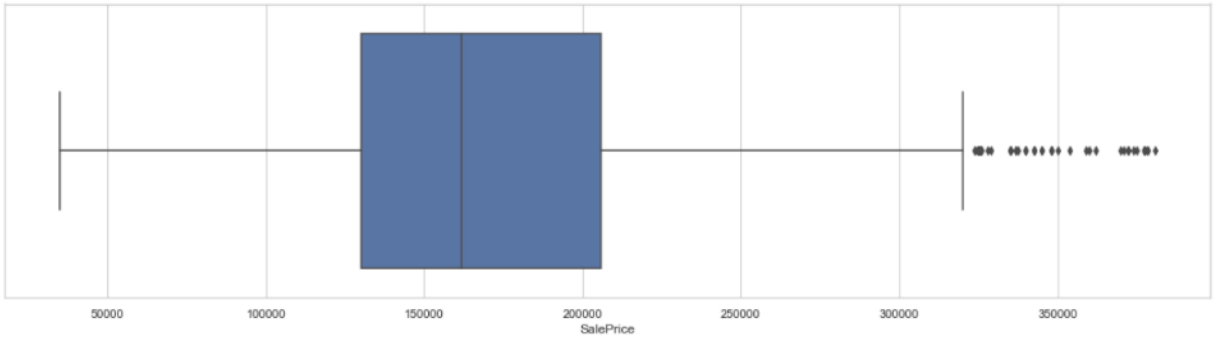
| | | |
|---|---|---|
| MoSold | MoSold | 1.000000 |
| YrSold | YrSold | 1.000000 |
| LotFrontage | LotFrontage | 1.000000 |
| BsmtFullBath | BsmtFullBath | 1.000000 |
| SalePrice | SalePrice | 1.000000 |
| GarageYrBlt | YearBuilt | 0.872805 |
| YearBuilt | GarageYrBlt | 0.872805 |
| GarageArea | GarageCars | 0.852404 |
| GarageCars | GarageArea | 0.852404 |
| TotalBsmtSF | 1stFlrSF | 0.835231 |
| 1stFlrSF | TotalBsmtSF | 0.835231 |
| GrLivArea | TotRmsAbvGrd | 0.806731 |
| TotRmsAbvGrd | GrLivArea | 0.806731 |
| HalfBath | 2ndFlrSF | 0.701133 |
| 2ndFlrSF | HalfBath | 0.701133 |
| YearRemodAdd | GarageYrBlt | 0.686989 |
| GarageYrBlt | YearRemodAdd | 0.686989 |
| BsmtFinSF1 | BsmtFullBath | 0.681065 |
| BsmtFullBath | BsmtFinSF1 | 0.681065 |
| YearBuilt | YearRemodAdd | 0.678332 |

✓ نمودار پراکندگی این دو متغیر نسبت به هم رسم شد، که نشان‌دهنده‌ی correlation مثبت این دو است: (این دو متغیر، به ترتیب، نشان دهنده‌ی سال ساخت گاراژ و سال ساخت خانه هستند.)

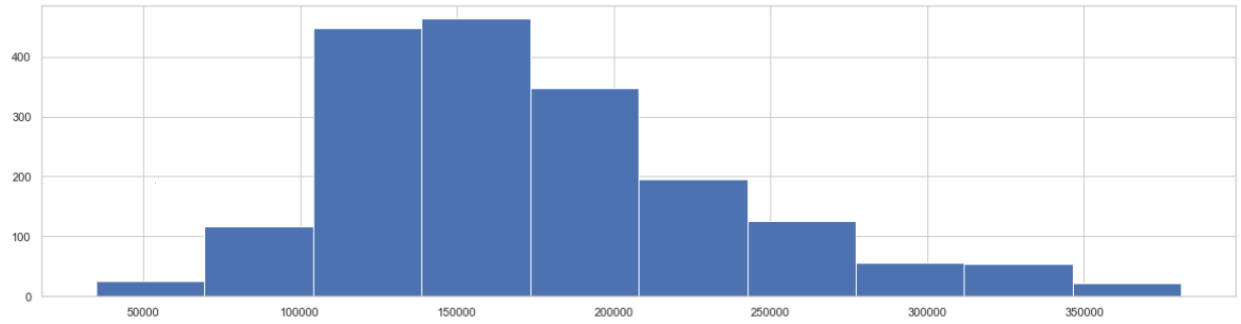



- **بخش ۵:**

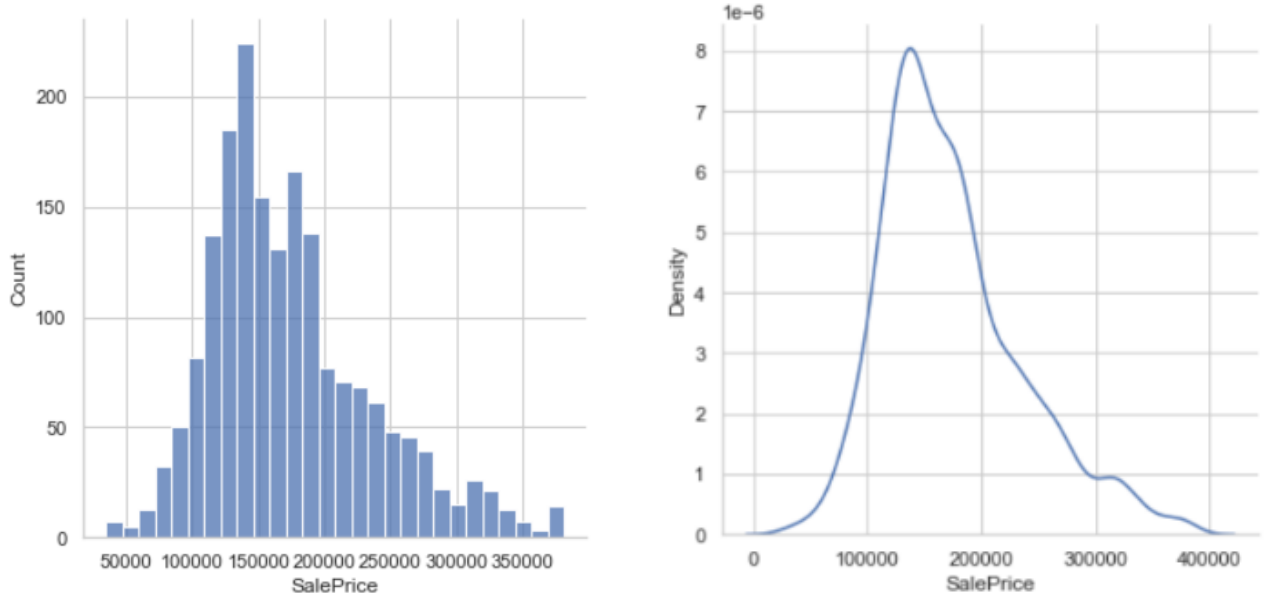
۱. نمودار جعبه‌ای:



۲. نمودار فراوانی:




۳. نمودار توزیعی:




- ✓ هر سه نمودار، چولگی راست توزیع متغیر را نشان می‌دهند. دو نمودار فراوانی و توزیعی، نسبت به نمودار جعبه ای اطلاعات بیشتری درباره‌ی نحوه‌ی توزیع این متغیر به ما می‌دهند. نمودار فراوانی، درواقع مثل نمودار توزیعی، این را نشان می‌دهد که مقادیر با چه فراوانی‌ای وجود دارند. تفاوت این دو نمودار در این است، که نمودار فراوانی، مقادیر را در دسته‌هایی قرار می‌دهد، و سپس توزیع آن دسته‌ها را نشان می‌دهد. این دسته‌ها می‌توانند دارای عرض دلخواه ما باشند. اگر مقادیر را دسته‌بندی نکنیم، نمودار فراوانی، همانند نمودار توزیعی می‌شود.
- ✓ به طور کلی، نمودار توزیعی اطلاعات بیشتری درباره‌ی نحوه‌ی توزیع و پراکندگی مقادیر متغیرها در اختیار ما قرار می‌دهد.

- **بخش ۶:**

می‌توان نمودار پراکندگی را برای ۳ متغیر، درصورتی‌که دو تا از متغیرها به صورت عددی و یکی آن‌ها categorical باشد، رسم کرد. به این شکل که، طول و عرض نمودار، به هر یک از دو متغیر عددی نسبت داده می‌شود و رنگ هر یک از داده‌هایی که در صفحه‌ی نمودار نمایش داده می‌شوند، بسته به اینکه از نظر متغیر سوم در کدام category قرار می‌گیرد، تعیین می‌شود.
برای مثال نمودار زیر برای سه متغیر SalePrice، YearBuilt و OverallQual رسم شده‌است.
دو متغیر عددی SalePrice و YearBuilt به ترتیب در طول و عرض نمودار قرار گرفته‌اند و متغیر دسته‌ای OverallQual، در رنگ داده‌های روی نمودار، اثرگذار بوده است:


نمودار رسم شده، نشان‌گر این است که هر چه خانه جدیدتر ساخته شده باشد، احتمالاً قیمت بیشتری خواهد داشت، و همچنین، کیفیت کلی بیشتری هم خواهدداشت.
در کل، این برداشت می‌شود که این سه متغیر (سال ساخت خانه، قیمت خانه و کیفیت کلی خانه) با یکدیگر تقریباً ارتباط مثبتی دارند.

---

## سوال ۲:

### ❖ پیش پردازش:

- **بخش ۱:** به کمک API موردنظر، exchange Rate مربوط به ارزها، نسبت به دلار به دست آمد و یک ستون جدید، نشان دهنده‌ی تبدیل شده‌ی مقادیر به دلار، به دیتاست اضافه شد. دیتاست تغییریافته، به شکل مقابل است:

| | user_id | SKU | AddedTime | Price | CurrencyISO | USD_Price |
|---|---|---|---|---|---|---|
| 0 | 6192636 | personal_offer_starter_pack | 2023-01-31 13:16:55.991756+00:00 | 2.29 | GBP | 2.760699 |
| 1 | 5954105 | bundle_pack_1 | 2023-01-06 19:20:33.631714+00:00 | 1.79 | EUR | 1.891978 |
| 2 | 5954105 | coin_pack_1 | 2023-01-07 15:56:47.792655+00:00 | 0.79 | EUR | 0.835007 |
| 3 | 5903715 | bundle_pack_1 | 2023-01-01 18:48:38.391356+00:00 | 1.99 | USD | 1.990000 |
| 4 | 5984323 | golden_ticket_season_pass | 2023-01-11 21:13:47.161073+00:00 | 9.99 | USD | 9.990000 |
| ... | ... | ... | ... | ... | ... | ... |
| 1408 | 5964679 | bundle_pack_1 | 2023-02-01 13:13:43.680081+00:00 | 1.49 | USD | 1.490000 |
| 1409 | 5964679 | coin_pack_1 | 2023-01-19 13:51:50.503887+00:00 | 0.99 | USD | 0.990000 |
| 1410 | 5964679 | bundle_pack_1 | 2023-02-04 23:15:25.865344+00:00 | 1.49 | USD | 1.490000 |
| 1411 | 5964679 | coin_pack_1 | 2023-01-20 15:43:32.899227+00:00 | 0.99 | USD | 0.990000 |
| 1412 | 4756103 | golden_ticket_season_pass | 2023-02-03 14:45:07.543052+00:00 | 9.99 | USD | 9.990000 |

1413 rows × 6 columns

- **بخش ۲:** این دو دیتاست، دارای ۳۰۵ مقدار مشترک در شناسه‌ی کاربر (user_id) بودند. با ترکیب آن‌ها با استفاده از این متغیر، سطرهایی که شناسه کاربر مشترک در دو دیتاست داشتند، در کنار هم قرار گرفتند. و سطرهایی که در یکی از دیتاست‌ها، شناسه‌ی کاربر نظیر نداشتند، با مقدار null برای اطلاعات دیتاست دیگر، در دیتاستِ حاصل، ظاهر شدند.

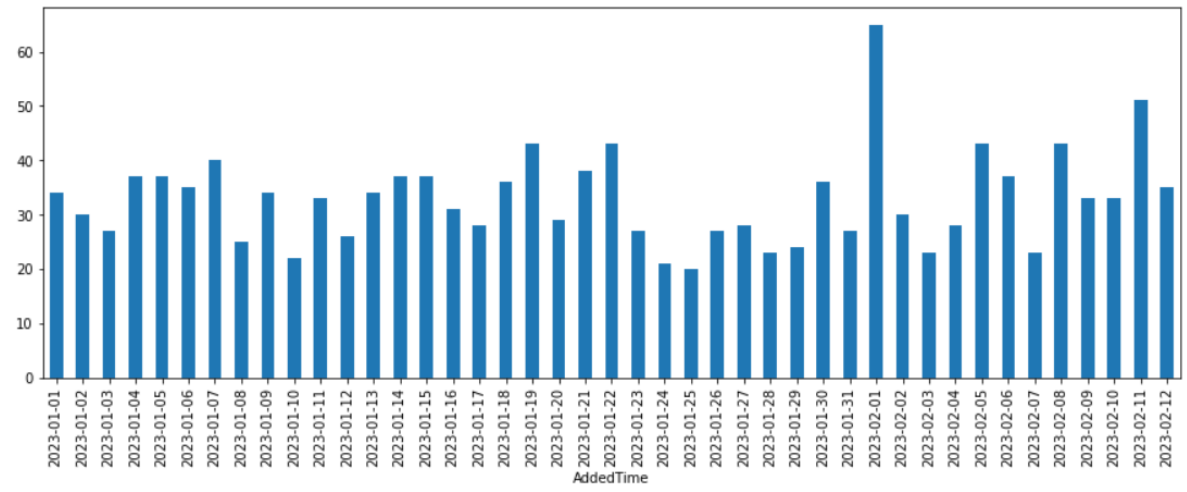
دیتاست حاصل:

| | user_id | SKU | AddedTime | Price | CurrencyISO | USD_Price | registered_time | country_code |
|---|---|---|---|---|---|---|---|---|
| **0** | 6192636 | personal_offer_starter_pack | 2023-01-31 13:16:55.991756+00:00 | 2.29 | GBP | 2.760699 | 2023-01-30 16:12:02.731706+01:00 | GB |
| **1** | 5954105 | bundle_pack_1 | 2023-01-06 19:20:33.631714+00:00 | 1.79 | EUR | 1.891978 | 2023-01-05 20:33:54.584158+01:00 | BE |
| **2** | 5954105 | coin_pack_1 | 2023-01-07 15:56:47.792655+00:00 | 0.79 | EUR | 0.835007 | 2023-01-05 20:33:54.584158+01:00 | BE |
| **3** | 5903715 | bundle_pack_1 | 2023-01-01 18:48:38.391356+00:00 | 1.99 | USD | 1.990000 | 2022-12-30 18:47:41.108888+01:00 | US |
| **4** | 5984323 | golden_ticket_season_pass | 2023-01-11 21:13:47.161073+00:00 | 9.99 | USD | 9.990000 | 2023-01-09 18:57:06.353154+01:00 | US |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **57617** | 5983087 | NaN | NaT | NaN | NaN | NaN | 2023-01-09 16:36:16.596704+01:00 | CN |
| **57618** | 6132178 | NaN | NaT | NaN | NaN | NaN | 2023-01-24 15:28:11.102926+01:00 | GB |
| **57619** | 5968220 | NaN | NaT | NaN | NaN | NaN | 2023-01-07 15:21:56.387803+01:00 | AF |
| **57620** | 6221518 | NaN | NaT | NaN | NaN | NaN | 2023-02-02 22:07:30.089388+01:00 | IQ |
| **57621** | 6043727 | NaN | NaT | NaN | NaN | NaN | 2023-01-15 14:13:42.715729+01:00 | TR |

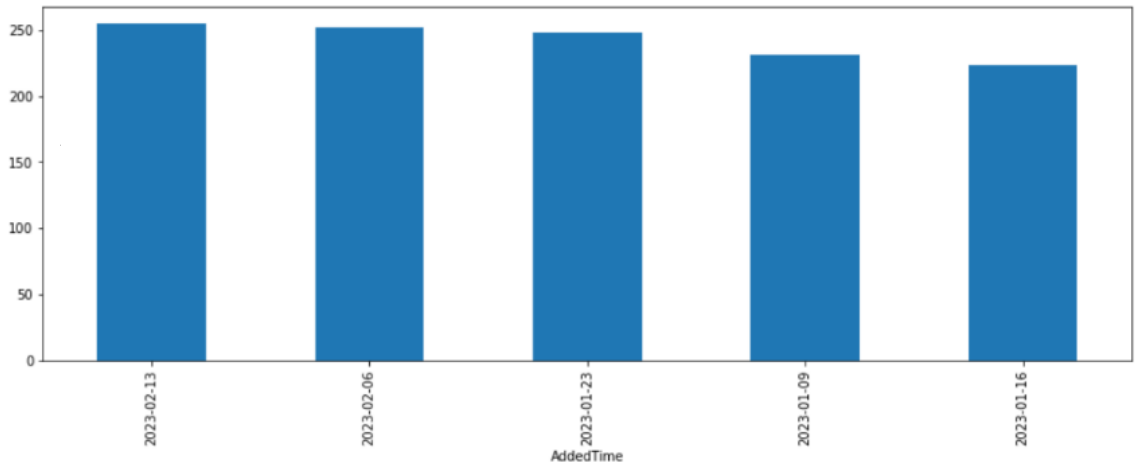
57622 rows × 8 columns

## ❖ نمایش دادگان:

- **بخش ۱:** نمودار میله‌ای نشان‌گر مجموع خریدها در هر روز، مطابق شکل زیر رسم شد:



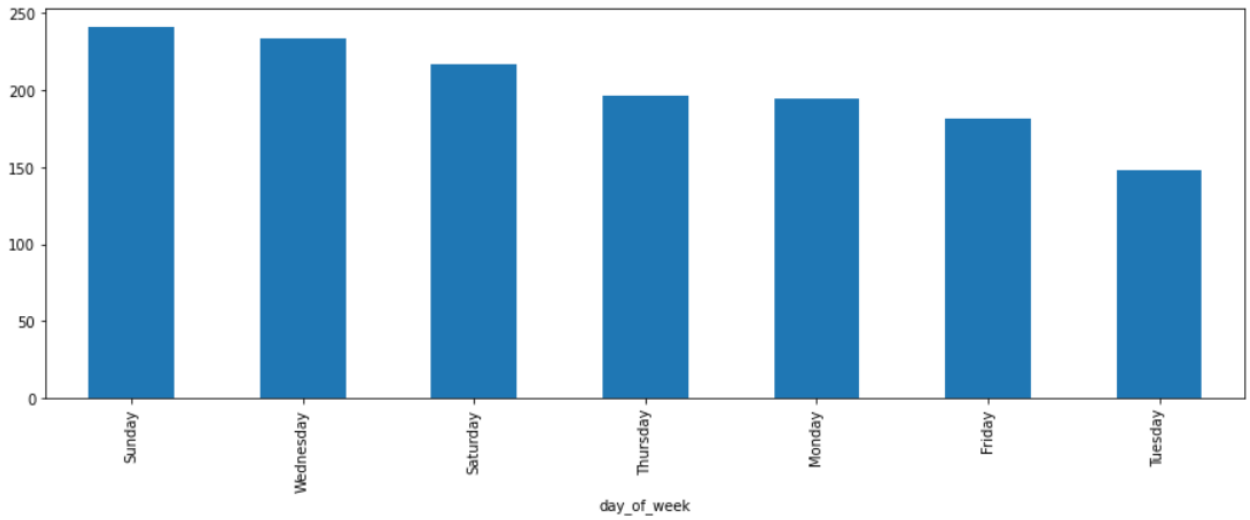
- **بخش ۲:** در هفته‌ی شروع شونده از ۲۰۲۳/۰۲/۱۳ بیشترین خرید (۲۵۵ مورد) ثبت شده‌است.
۵ هفته با بیشترین خرید، در نمودار زیر نمایش داده شده‌اند:




- **بخش ۳:** یک ستون جدید، نشان‌دهنده‌ی روزهای هفته، به دیتاست اضافه شد:

| | user_id | SKU | AddedTime | Price | CurrencyISO | USD_Price | date | registered_time | country_code | day_of_week |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 6192636 | personal_offer_starter_pack | 2023-01-31 13:16:55.991756+00:00 | 2.29 | GBP | 2.760699 | 2023-01-24 13:16:55.991756+00:00 | 2023-01-30 16:12:02.731706+01:00 | GB | Tuesday |
| **1** | 5954105 | bundle_pack_1 | 2023-01-06 19:20:33.631714+00:00 | 1.79 | EUR | 1.891978 | 2022-12-30 19:20:33.631714+00:00 | 2023-01-05 20:33:54.584158+01:00 | BE | Friday |
| **2** | 5954105 | coin_pack_1 | 2023-01-07 15:56:47.792655+00:00 | 0.79 | EUR | 0.835007 | 2022-12-31 15:56:47.792655+00:00 | 2023-01-05 20:33:54.584158+01:00 | BE | Saturday |
| **3** | 5903715 | bundle_pack_1 | 2023-01-01 18:48:38.391356+00:00 | 1.99 | USD | 1.990000 | 2022-12-25 18:48:38.391356+00:00 | 2022-12-30 18:47:41.108888+01:00 | US | Sunday |
| **4** | 5984323 | golden_ticket_season_pass | 2023-01-11 21:13:47.161073+00:00 | 9.99 | USD | 9.990000 | 2023-01-04 21:13:47.161073+00:00 | 2023-01-09 18:57:06.353154+01:00 | US | Wednesday |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **57617** | 5983087 | NaN | NaT | NaN | NaN | NaN | NaT | 2023-01-09 16:36:16.596704+01:00 | CN | NaN |

سپس خریدها بر اساس روزهای هفته دسته بندی شدند و نمودار متناظر به شکل زیر رسم شد:




بیشترین خرید مربوط به یکشنبه‌هاست. ۲۴۱ مورد از خریدها در این روز انجام شده‌اند.

- **بخش ۴:** ۱۰ کاربر با بیشترین خرید:

```
user_id
2474953    50
6039515    36
5394301    29
4913028    28
6294656    27
1507884    25
6029577    25
4627239    24
3265117    23
5991949    22
```

کاربر با بیشترین خرید، مربوط به کشور آمریکا است. اطلاعات این کاربر:

| | user_id | registered_time | country_code |
|---|---|---|---|
| **49302** | 2474953 | 2021-10-10 01:51:08.762136+02:00 | US |

- **بخش ۵:** ۱۵ کاربر مربوط به ۱۵ کشور، خریداران یکتای آن کشورها هستند:

```
country_code
AD    1
LI    1
KI    1
FM    1
BB    1
TK    1
FO    1
TO    1
GG    1
VI    1
GU    1
AG    1
YT    1
GW    1
SB    1
```

- **بخش ۶:** ۵ کشور با بیشترین مبلغ خرید عبارتند از:

```
US    4226.720000
GB     378.722122
AU     248.127279
DE     207.183281
CA     128.843221
```

۱. آمریکا
۲. انگلیس
۳. استرالیا
۴. آلمان
۵. کانادا

- **بخش ۷:** بهترین خریداران مربوط به کشور آمریکا هستند. از بین ۱۵ خریدار که بیشترین مجموع مبلغ خریده شده را داشتند، ۱۴ خریدار مربوط به آمریکا هستند. اطلاعات ۵ خریدار برتر از این منظر:

| | user_id | registered_time | country_code |
|---|---|---|---|
| **49302** | 2474953 | 2021-10-10 01:51:08.762136+02:00 | US |

| | user_id | registered_time | country_code |
|---|---|---|---|
| **51139** | 3265117 | 2022-01-20 19:26:04.577517+01:00 | US |

| | user_id | registered_time | country_code |
|---|---|---|---|
| **34217** | 6029577 | 2023-01-14 03:31:57.876848+01:00 | US |

| | user_id | registered_time | country_code |
|---|---|---|---|
| **37540** | 5942278 | 2023-01-04 15:57:12.140747+01:00 | US |

| | user_id | registered_time | country_code |
|---|---|---|---|
| **34440** | 1507884 | 2021-04-13 01:05:27.393719+02:00 | US |

- **بخش ۸:** اگر بخواهیم با استفاده از تبلیغات، تعدادی کاربر جدید از یکی از کشورهای این دیتاست جذب کنیم، بهتر است از کشور چین انتخاب شوند. تعداد ۱۹۴ کاربر در این دیتاست، مربوط به کشور چین هستند؛ که با توجه به جمعیت زیاد این کشور، همچنان پتانسیل افزایش کاربر در این کشور وجود دارد. مثلا دو کشور هند و آمریکا که به ترتیب، دومین و سومین کشورهای پرجمعیت دنیا هستند، در این دیتاست از نظر تعداد کاربران، با بیش از ۵۵۰۰ کاربر، در رده‌های اول و دوم قرار گرفته‌اند. ولی کشور چین، تنها ۱۹۴ کاربر در این دیتاست دارد.

```
country_code
IN     6960
US     5585
CN      194
```

- **بخش ۹:** تعداد کاربران ثبت شده‌ی هر کشور بر روی نقشه:

نمایش بزرگتر اعداد: