Exploratory Data Analysis of Ames Housing Prices

For this assignment we chose to examine the Housing Prices Dataset for Ames Iowa. Here's what we found:

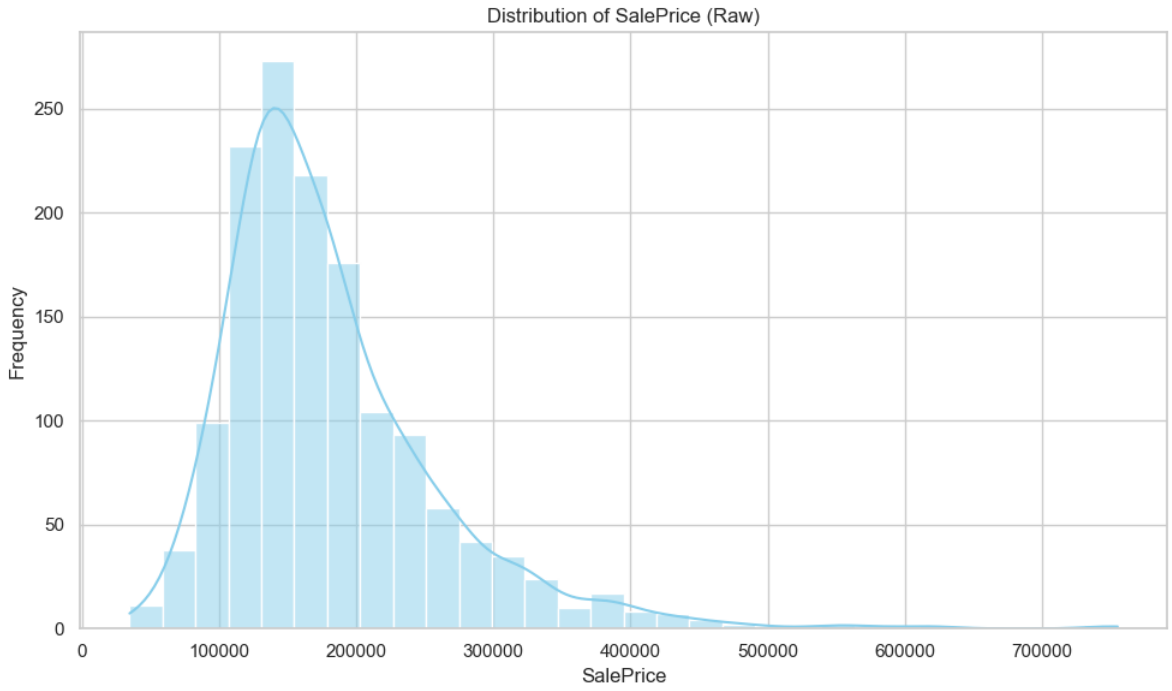
Management/Research Question: "How do housing characteristics such as quality, size, location, and age relate to their sale price?"

This impacts various groups such as home buyers, home sellers, real estate professionals, and policymakers because it identifies which property features most influence value and thus can guide investment, pricing, and development decisions.

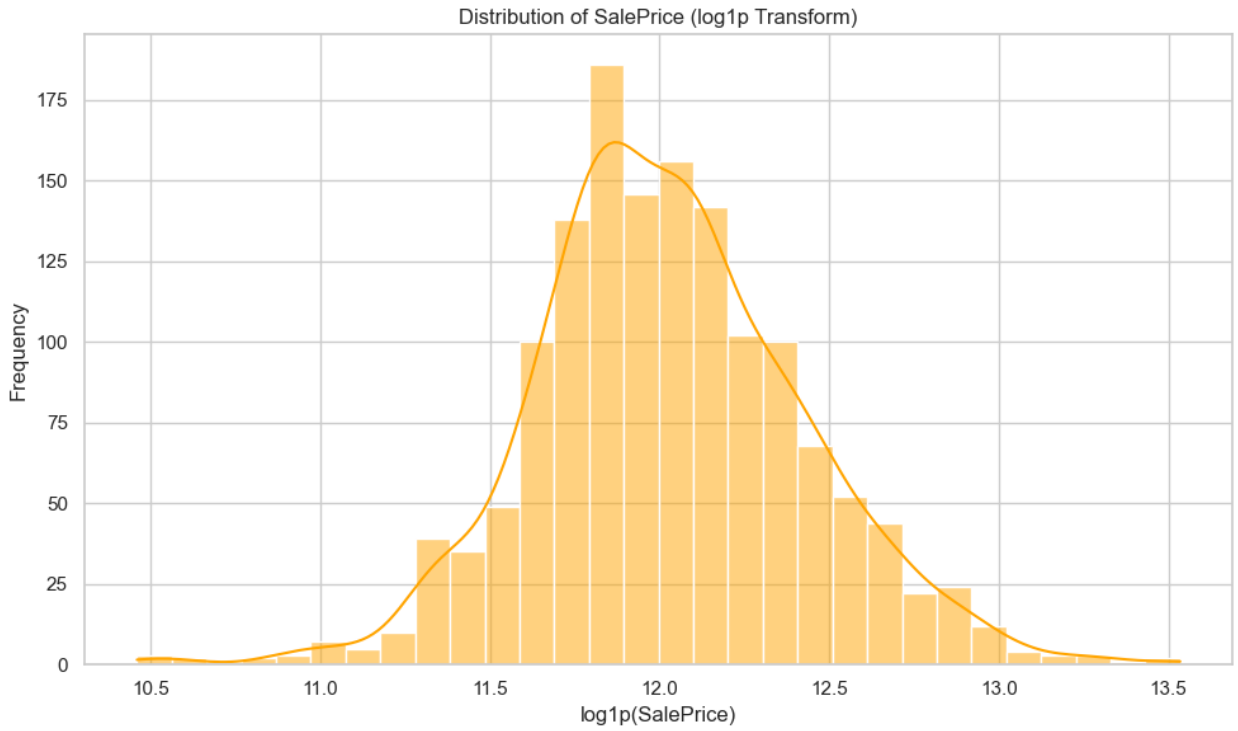
Descriptive Statistics and Target Distribution:

The assignment specifies that we use SalePrice as the dependent variable. When we looked into the data set we saw it contains 1,460 observations with a mean of approximately $180,921 and a median of $163,000.

Prices range from $34,900 to $755,000, with the middle 50% of homes between $129,975 and $214,000 (please see the Data Section Below). The histogram of SalePrice reveals a strong right skew (skewness 1.88, kurtosis 6.51), and the box-plot confirms numerous high-value outliers above $400,000.



We applied a log1p transform to SalePrice to reduce its strong right-skew and compress extreme values. This made the distribution much more symmetric and closer to a normal (bell-shaped) curve. Doing this helps regression models work better because effectively it reduces the influence of very high-priced outliers on the model, stabilizing estimates and making coefficients more interpretable.



The Data:

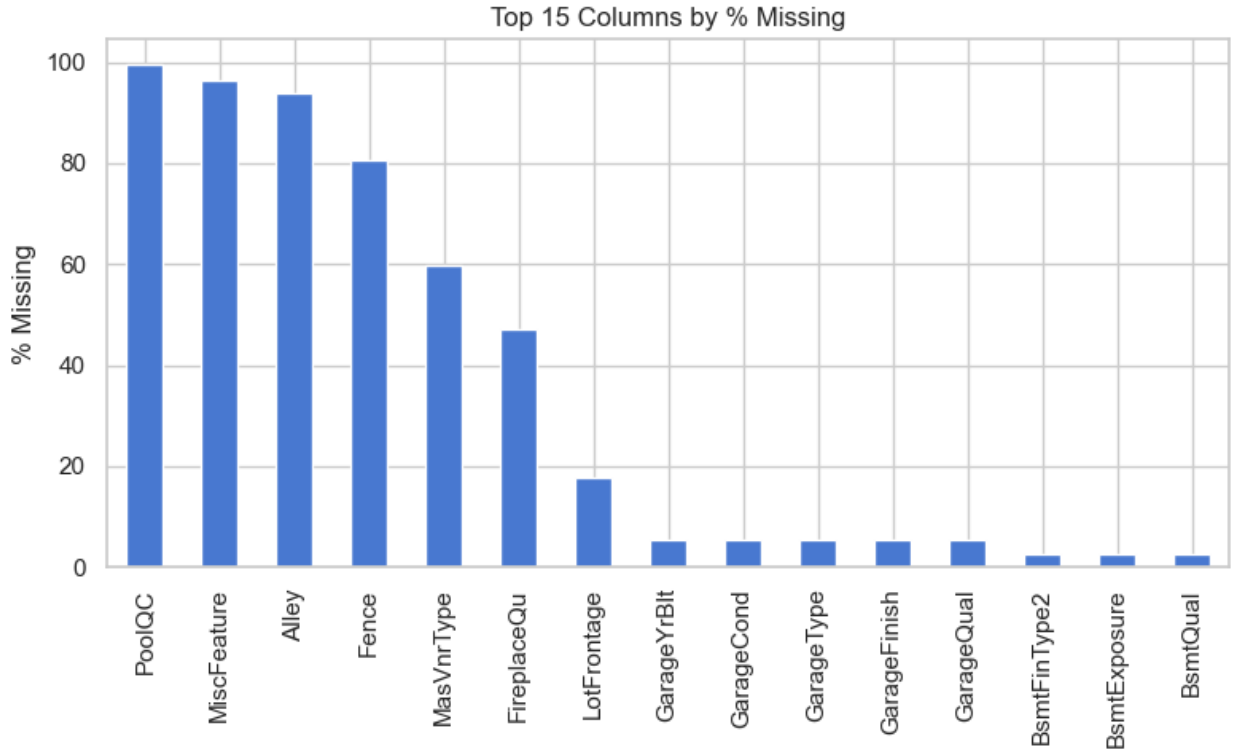
Descriptive statistics for SalePrice:

```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64

Skewness of SalePrice: 1.88
Kurtosis of SalePrice: 6.51
```

Missing Data: Several variables are missing data

PoolQC (99.5%), MiscFeature (96.3%), Alley (93.8%), and Fence (80.8%) are missing for the majority of records, which reflects that the majority of homes do not have a pool, alley or fence. Other variables also had missing value as well; MasVnrType (59.7%), FireplaceQu (47.3%), and LotFrontage (17.7%). 5.5% of homes within the dataset did not have garages. It's important to not that we're not sure if the data is actually missing or if the homes surveyed simply did not have the these items.



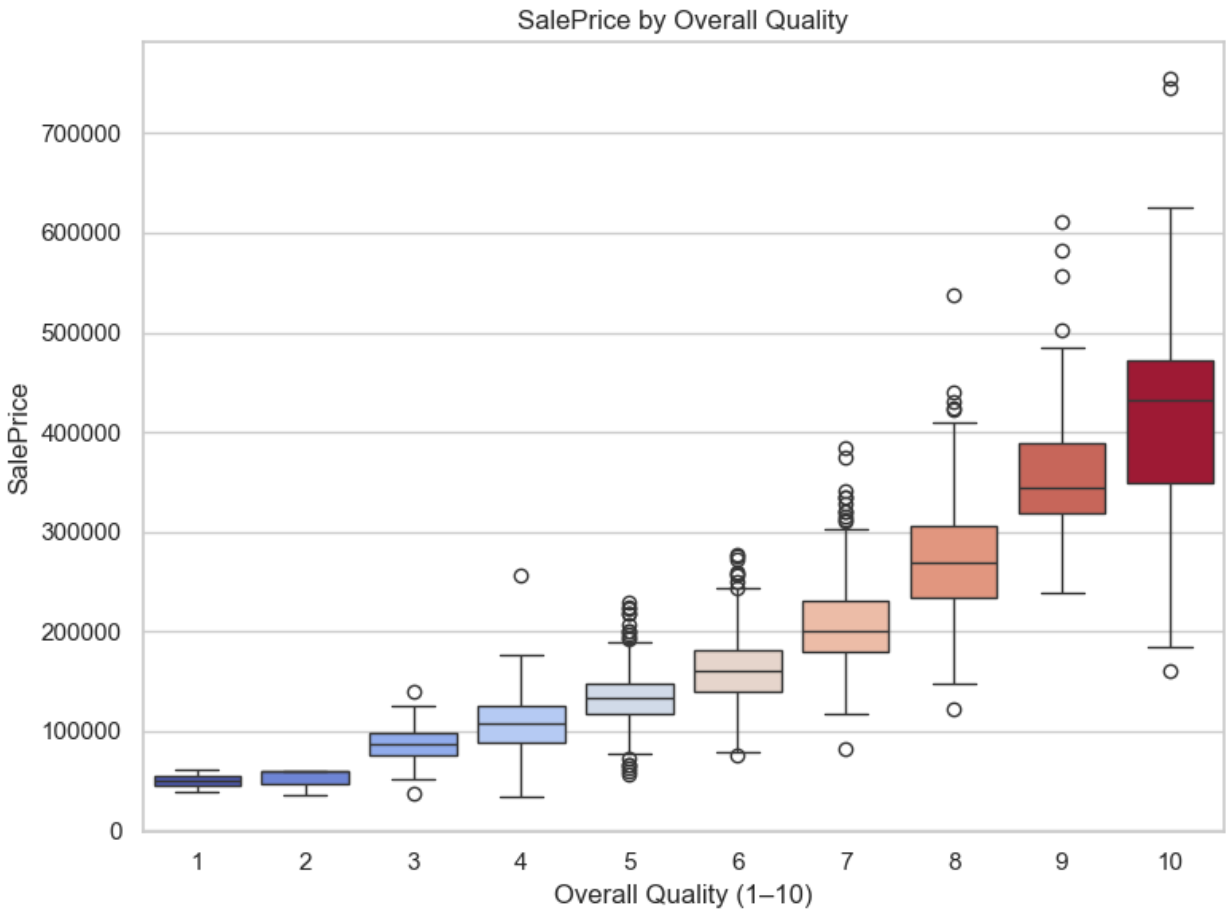
The Data:

```
Columns with missing values (percent missing):
PoolQC          99.520548
MiscFeature     96.301370
Alley           93.767123
Fence           80.753425
MasVnrType      59.726027
FireplaceQu     47.260274
LotFrontage     17.739726
GarageYrBlt      5.547945
GarageCond       5.547945
GarageType       5.547945
GarageFinish     5.547945
GarageQual       5.547945
BsmtFinType2     2.602740
BsmtExposure     2.602740
BsmtQual         2.534247
dtype: float64
```

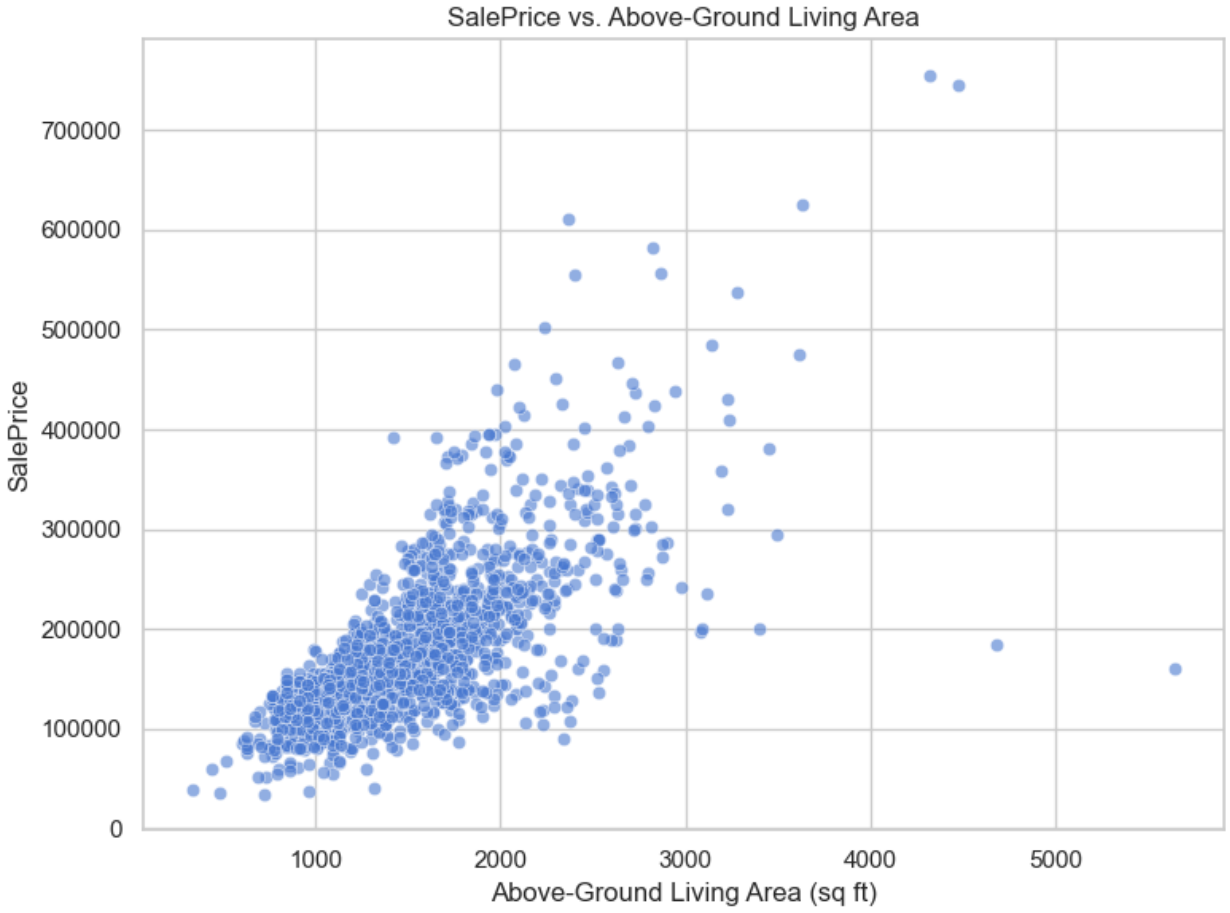
Key Predictors vs. SalePrice:

Three major predictors show strong relationships with SalePrice:

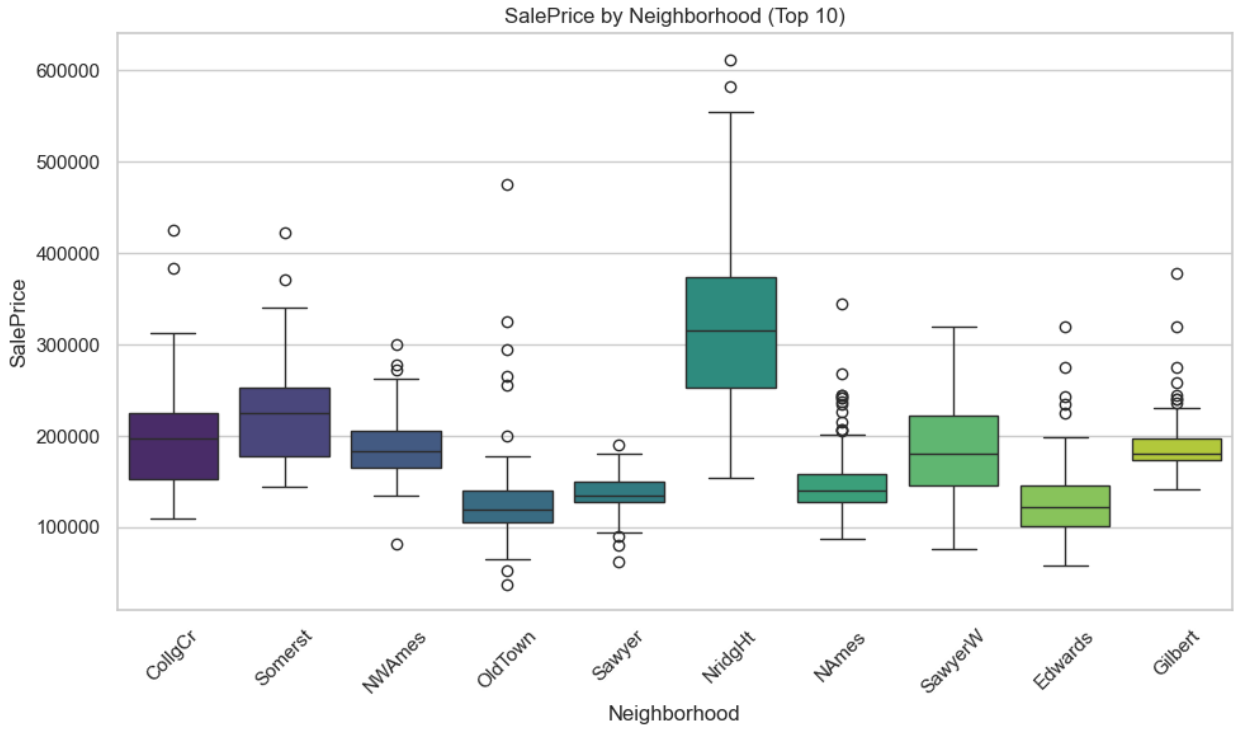
- **OverallQual:** Median prices roughly double between mid-tier homes (ratings 5–6) and top-tier homes (ratings 9–10). High-quality homes frequently sell above $400,000, while low-quality homes cluster below $150,000, confirming quality as a leading driver of price.



- **GrLivArea:** Above-ground living area shows a clear positive relationship with SalePrice. Most homes between 1,000–2,500 square feet sell for $100,000–$300,000, while larger homes up to 5,000 square feet often exceed $400,000, with a few unusual outliers.



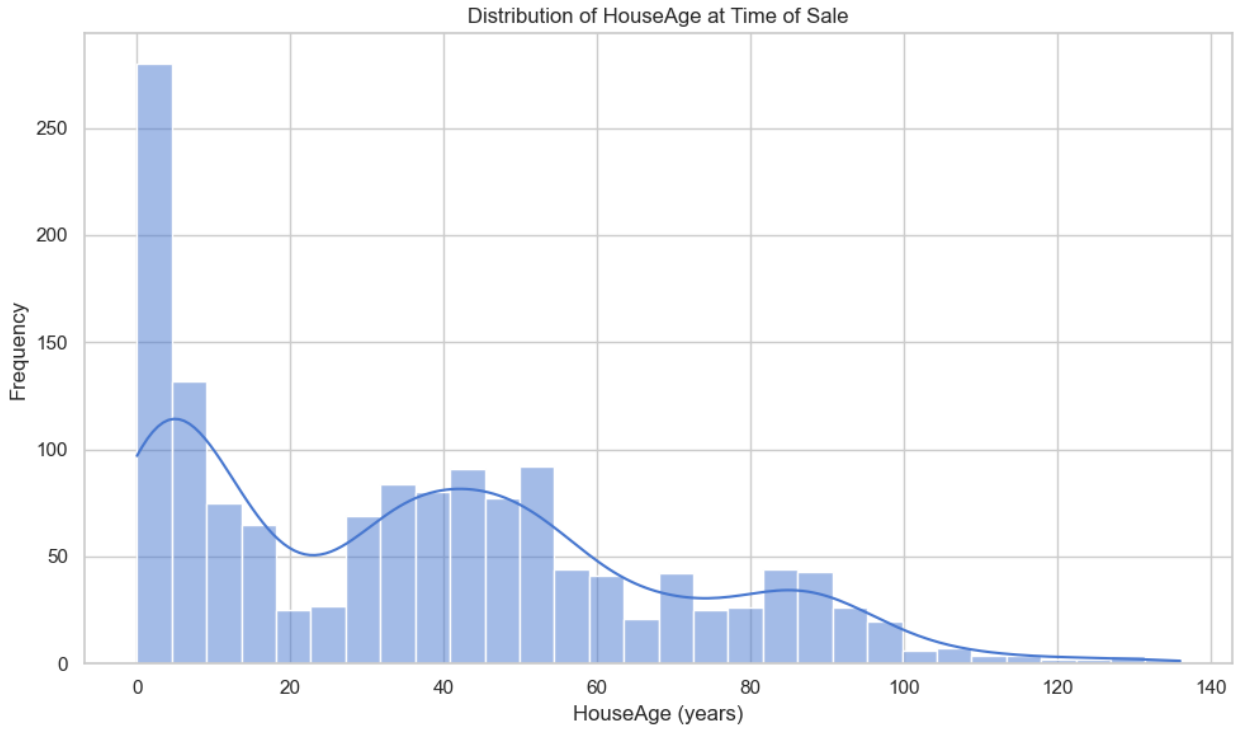
- **Neighborhood:** Location matters significantly. NridgHt and Somerst have median prices above $250,000 (some reaching $500,000+), while OldTown and Sawyer average closer to $150,000.

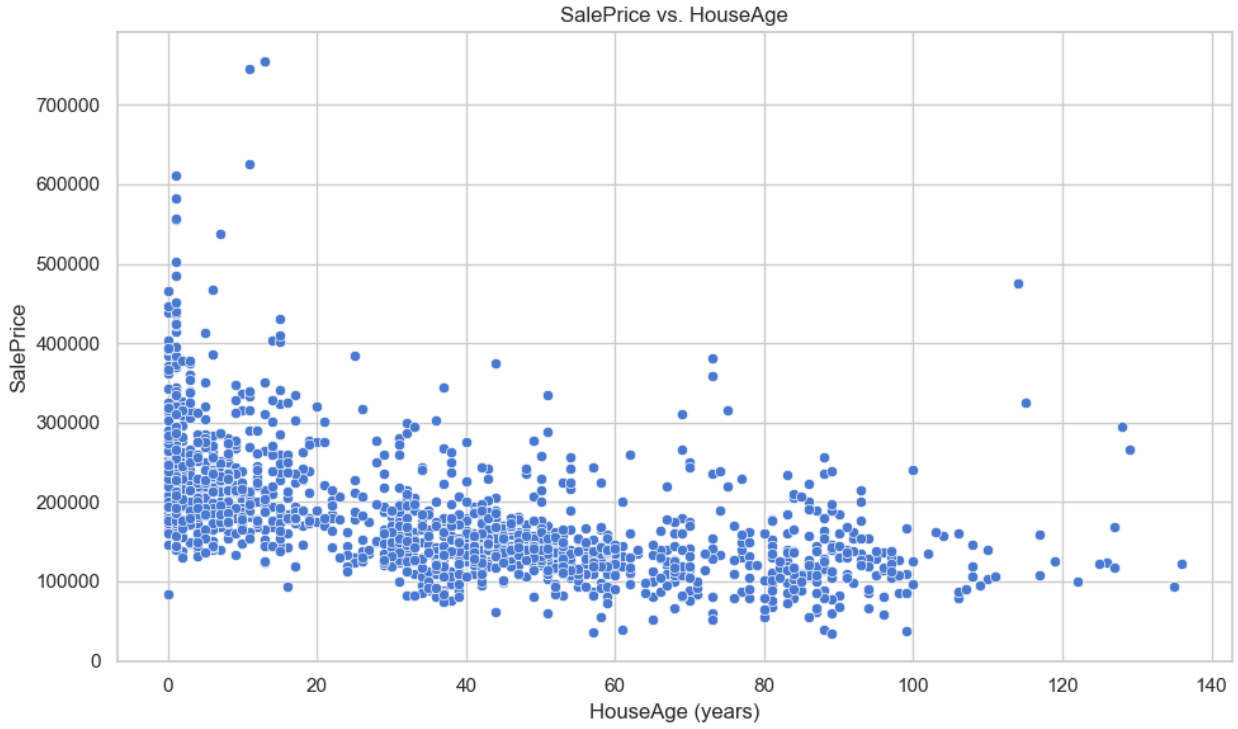


Feature Engineering:

- HouseAge (at sale time) ranges from 0 years for new homes to 136 years for the oldest with a mean of 36.5 years. Its correlation with SalePrice is –0.52, indicating newer homes generally command higher prices.

```
HouseAge feature created. Descriptive stats:
count    1460.000000
mean       36.547945
std        30.250152
min         0.000000
25%         8.000000
50%        35.000000
75%        54.000000
max       136.000000
Name: HouseAge, dtype: float64
```


Distribution of HouseAge at Time of Sale
Frequency
HouseAge (years)
0
50
100
150
200
250
0
20
40
60
80
100
120
140


SalePrice vs. HouseAge
SalePrice
HouseAge (years)
0
100000
200000
300000
400000
500000
600000
700000
0
20
40
60
80
100
120
140

- Remodeled flags whether a home was remodeled after construction.
- TotalBaths aggregates full and half bathrooms above and below grade, creating a single measure of bathroom capacity (sample homes range from 2.0 to 3.5 bathrooms).

Scaling and Comparisons:

Both Min–Max Scaling and Standard Scaling were applied to SalePrice, HouseAge, and TotalBaths. For instance, SalePrice values from $140,000–$250,000 scale to 0.1459–0.2987 (Min–Max) and –0.5153 to +0.8698 (Standard). HouseAge values from 7–95 years scale to 0.05–0.688 (Min–Max) and –1.05 to +1.86 (Standard). TotalBaths values from 2.0–3.5 scale to 0.2–0.5 (Min–Max) and –0.27 to +1.64 (Standard).

```
New feature columns added: HouseAge, Remodeled, TotalBaths
   HouseAge  Remodeled  TotalBaths
0         7          0         3.5
1        34          0         2.5
2         9          1         3.5
3        95          1         2.0
4        10          0         3.5

Sample of raw vs. scaled features:
   SalePrice  HouseAge  TotalBaths  SalePrice_minmax  SalePrice_standard  \
0     208500         7         3.5          0.241078            0.347273
1     181500        34         2.5          0.203583            0.007288
2     223500         9         3.5          0.261908            0.536154
3     140000        95         2.0          0.145952           -0.515281
4     250000        10         3.5          0.298709            0.869843

   HouseAge_minmax  HouseAge_standard  TotalBaths_minmax  TotalBaths_standard
0         0.050725          -1.050994                0.5             1.642256
1         0.246377          -0.156734                0.3             0.368581
2         0.065217          -0.984752                0.5             1.642256
3         0.688406           1.863632                0.2            -0.268257
4         0.072464          -0.951632                0.5             1.642256
```

Insights and Conclusion:

This exploratory analysis highlighted several things:

- SalePrice is highly skewed but can be normalized with a log transform.
- Missing data largely represents absent features, guiding appropriate imputation strategies.
- Quality, size, and location are the strongest predictors of SalePrice, while HouseAge shows a moderate negative correlation (–0.52) with price.
- Feature engineering and scaling add meaningful dimensions and ensure comparability across variables with different units.