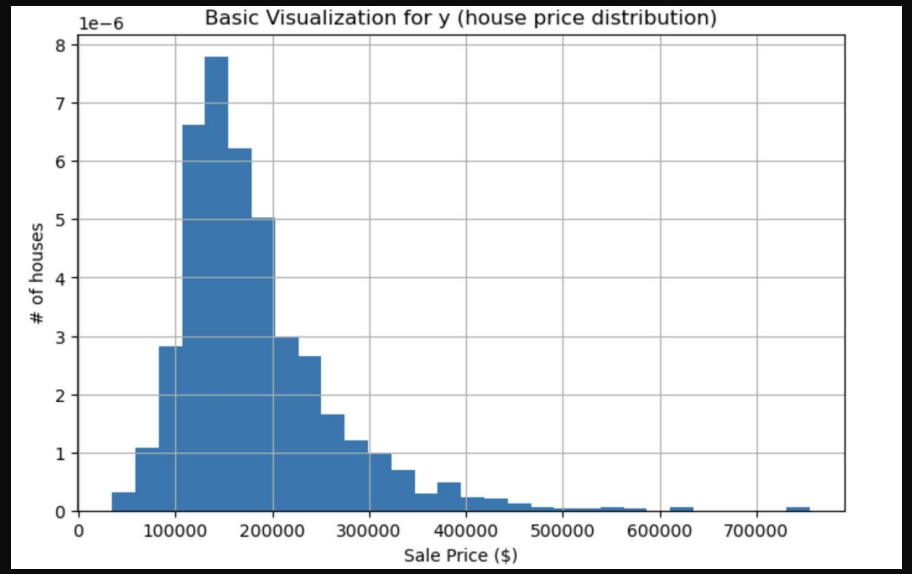
DATA ANALYSIS AND VISUALIZATION

AML Team 15: Sarah Kim, Chengyi Gong, Zhongyuan Ye, Lucy Zhang, Ziyu Liu

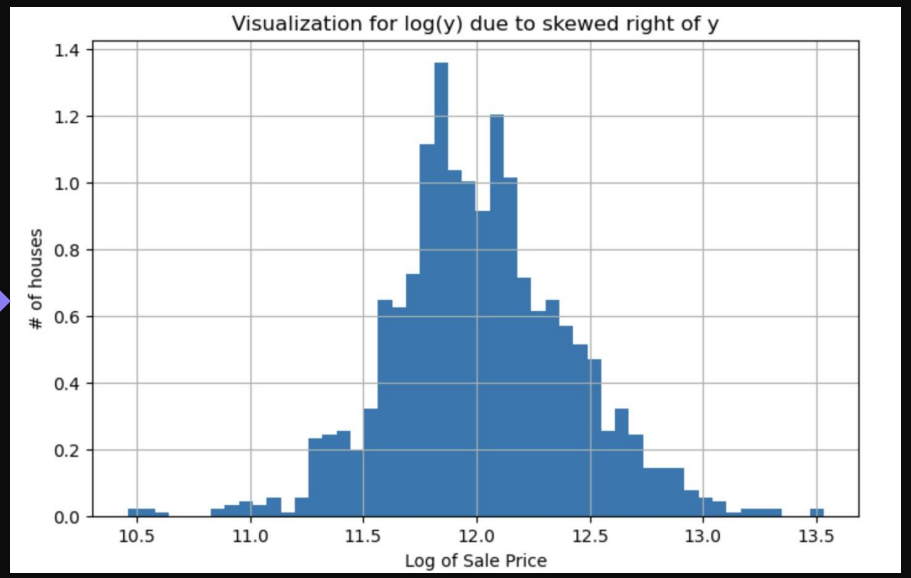
INITIAL DATA EXPLORATION & INSIGHTS

- Source: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview
- 81 columns (38 numerical (including 'SalePrice' and 'id' columns), 43 categorical) and 1460 rows
- Numerical features: 37 numerical columns (including ‘id’ column)
- Categorical features: 43 categorical columns
- Categorical features:
  - 9 of the categorical columns can be mapped to numerical values using ordinal mappings
  - 14 categorical columns had less than 5 unique values, so they can be encoded (as numerical values) by one-hot-encoding
  - Rest columns can be encoded by target encoding

VISUALIZATION 1: TARGET VALUE DISTRIBUTION

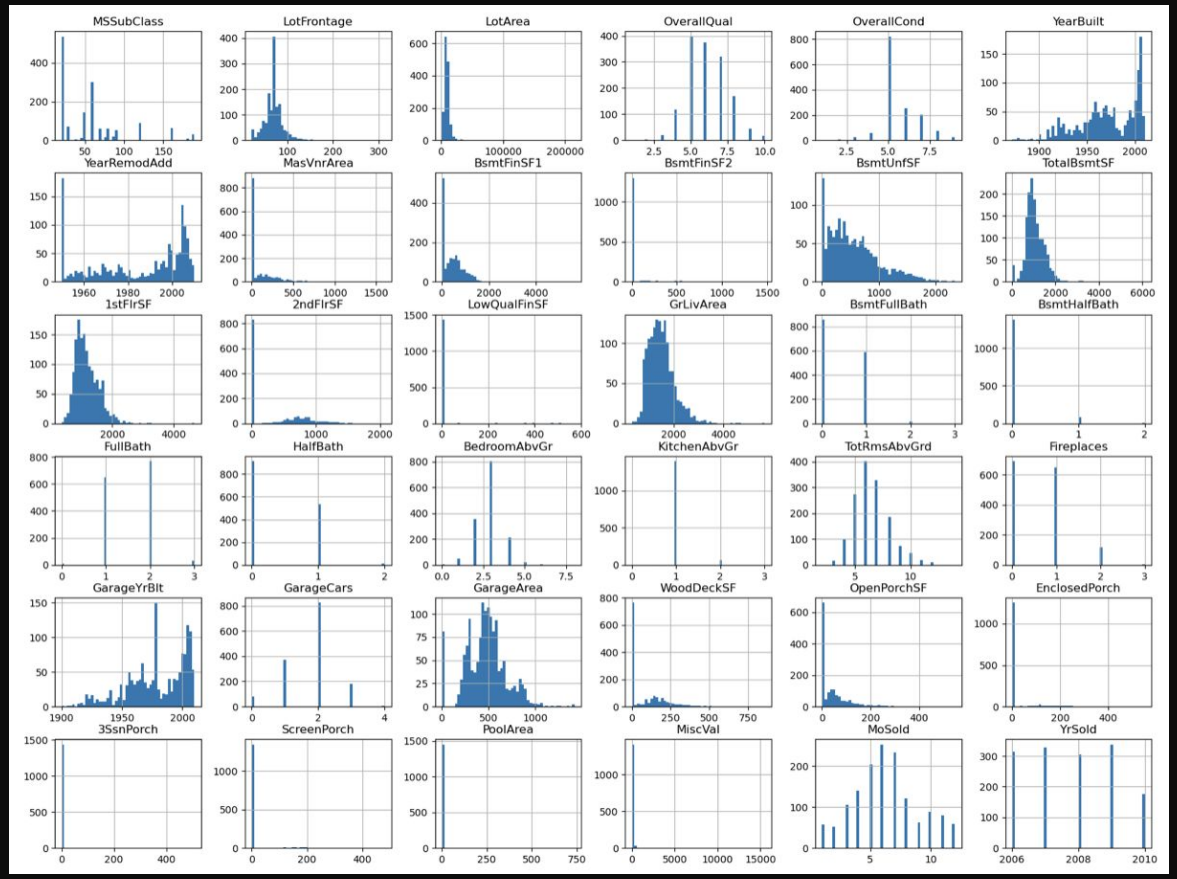


Log Transformation



The original target values (house price) are skewed to the right. However, the logs of target values are normally distributed, which is a good sign and indicates we don't need to sample

VISUALIZATION 2: NUM COL DISTRIBUTIONS



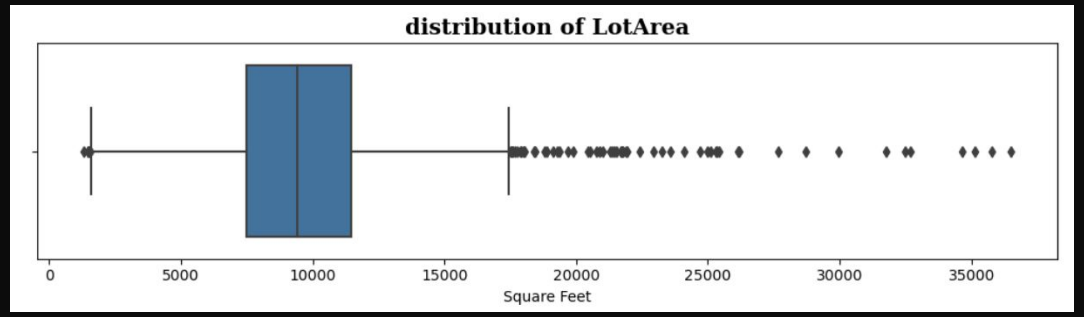
Distribution of all numeric columns

x-axis = value

y-axis = count

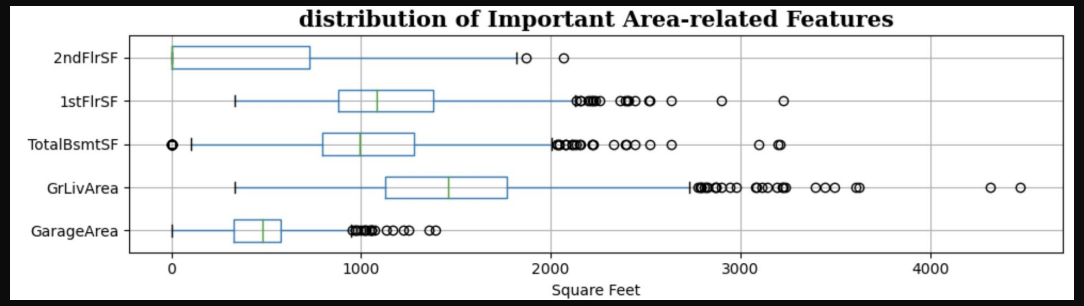
- Columns such as 2ndFlrSF, PoolArea, ScreenPorch, BsmtHalfBath, 3SsnPorch, EnclosedPorch, BsmtFinSF2, OpenPorchSF, MiscVal, WoodDeckSF, and LowQualFinSF are highly skewed to the right
- The numerical columns have **various scales.**
- 'LotArea' is the most skewed column and what we deem to be the most important feature → sorted every record based on the 'LotArea' of each record and kept only the rows within the bottom 99% of the record values, because extreme outlier values tend to lower model performance. We chose not to take out the outliers because then too many values would be taken out.

VISUALIZATION 3: DISTRIBUTIONS BY COLUMN TYPE



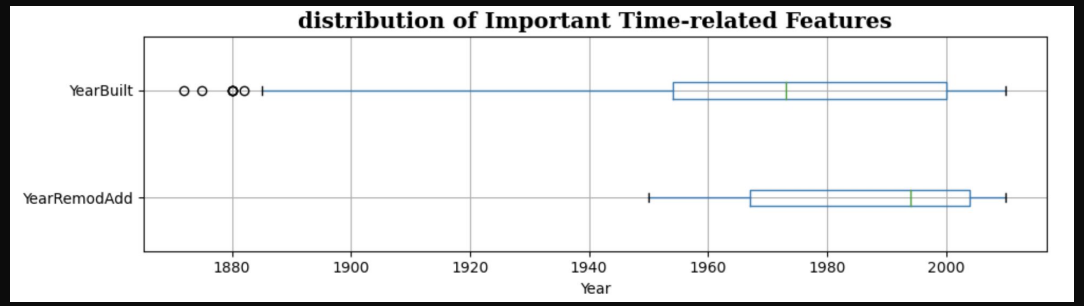
Distribution of lot size in square feet

- Most of the houses have a total area between 7,500 and 11,50 ft^2, with the median around 9,500 ft^2
- Outliers are detected, ranging from 17,500 to over 50,000 ft^2, but they will be left alone because there are too many outliers to take out all of them



Distribution of important area-related features

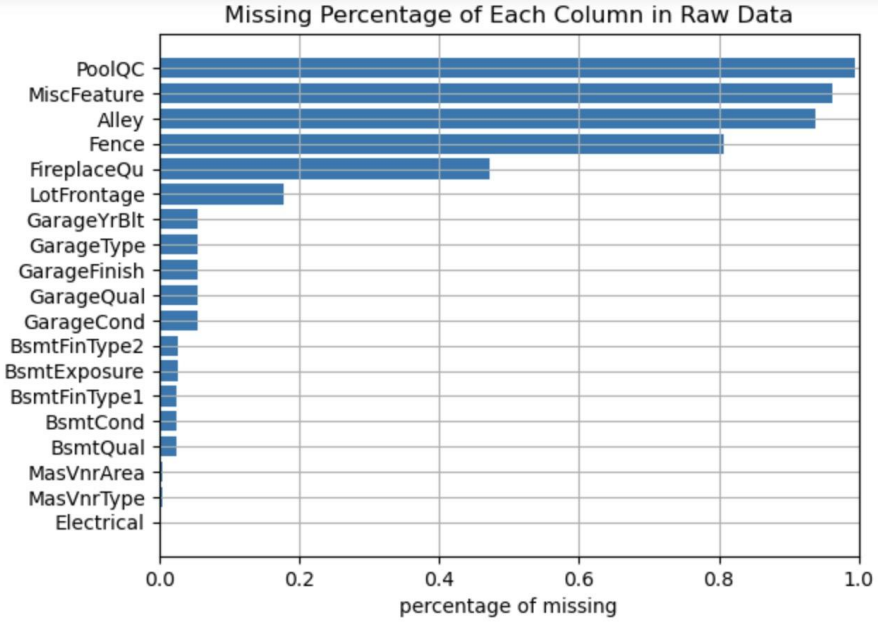
- The 1st floor generally has a much larger area than the 2nd
- The areas of the first floor and basement have a similar range, and might also be highly correlated
- Most of the square footages of second floor are near 0



Distribution of important time-related features

- Most of the original construction dates of houses are ranged from 1950 to 2000, with the median in the 1970s
- Most of the houses were remodeled from the late-1960s to mid-2000s, with the median in the 1990s

VISUALIZATION 4: MISSING VALUES

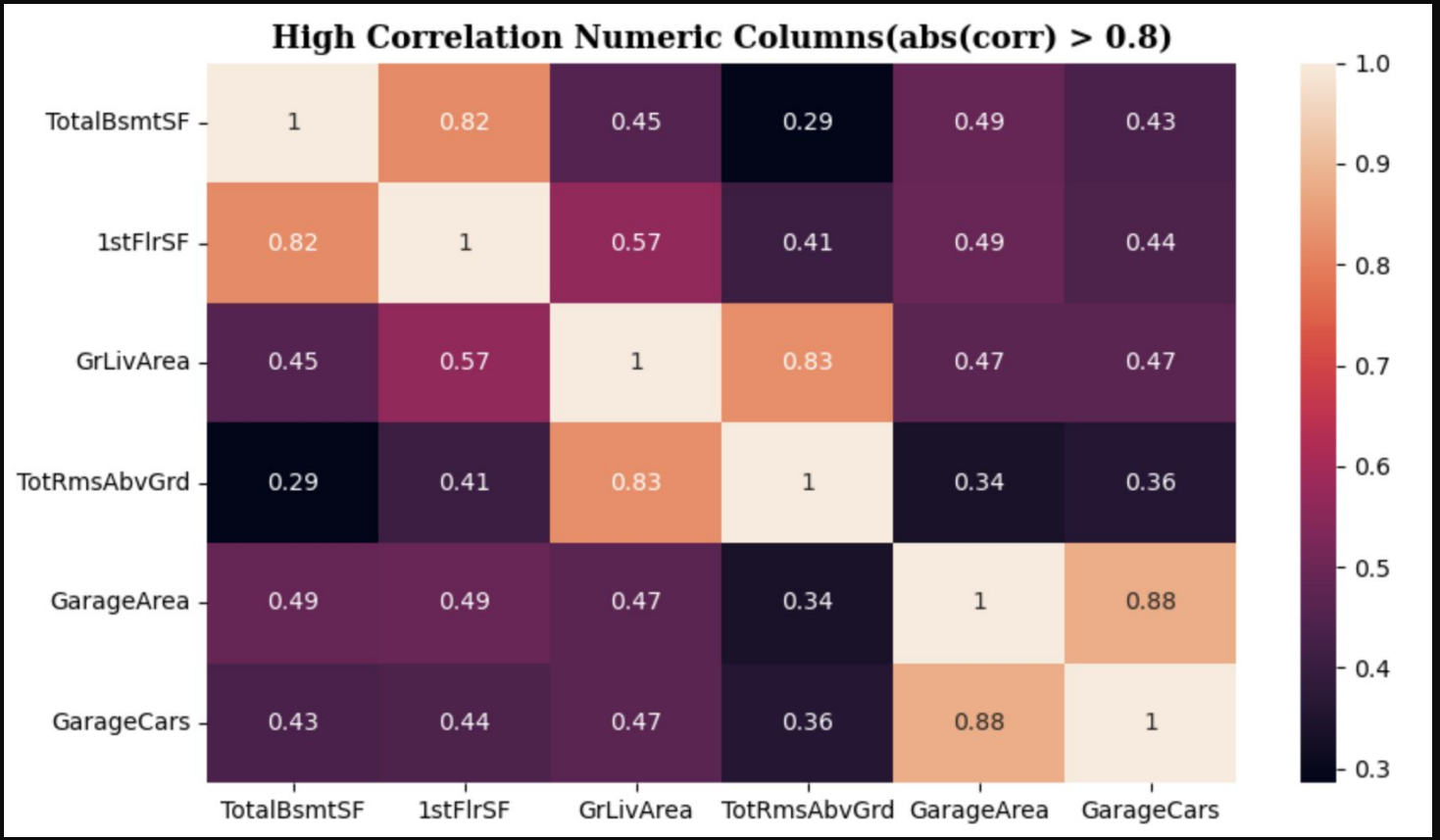


Top 4 features with most missing values:

Feature	Meaning	Proportion of NaN values
PoolQC	Pool quality	99.52%
MiscFeature	Miscellaneous factors that are not covered in other columns	96.30%
Alley	Type of alley access	93.77%
Fence	Fence quality	80.75%

Drop the top 4 columns with the highest number of missing values, since 80.8 to 99.5% of the values of those columns are NaN

VISUALIZATION 5: CORRELATION MATRIX



Drop one feature for each highly correlated pair

CORRELATION HEATMAP ANALYSIS

Highly correlated numerical columns: 3 pairs (Threshold > 0.8)

- **GarageCars vs. GarageArea: 0.88.** As the total area of the garage increases, it is likely that more cars can fit inside → drop 'GarageCars' column (not GarageArea, because the GarageArea is actually part of the house and should be more directly related to its price)
- **TotRmsAbvGrd vs. GrLivArea: 0.83.** As the total number of rooms in a house increases, the total living area is also likely to increase → drop 'GrLivArea' column (not TotRmsAbvGrd, because the room sizes can vary, so just the total number of rooms is less likely to be correlated to the house price than the total above-floor living area)
- **TotalBsmtSF vs. 1stFlrSF: 0.82.** They share a common area of land, where the basement's ceiling tends to be 1st floors floor, so it's likely that they have a similar size → drop 'TotalBsmtSF' column (not 1st floor, because the 1st floor tends to a more significant part of the house than the basement)

CLEANING (some parts already mentioned previously)

- Dropped the ‘Id’ column
- Removed the columns w/ the top 4 highest percentage of missing values (PoolQC, MiscFeature, Alley, and Fence)
- Replaced missing values of remaining numerical columns w/ column’s mean
- Replaced missing values of remaining categorical columns w/ most frequently occurring category
- Sorted every record based on the ‘LotArea’ of each record (because we presumed LotArea (the area of the house) was the most important feature in determining the price) and kept only the rows within the bottom 99% of the record values, since ‘Lot Area’ was highly skewed to the right
- Dropped 23 columns due to high correlation: 3 columns are numeric columns, 20 columns are columns resulting from columns that were from encoding

PROPOSED ML TECHNIQUES

- Goal: predict housing price in Ames, Iowa given the house's various features.
- Techniques:
  - Linear / Ridge regression: examine the relationship between the housing price and the number of bathrooms, number of bedrooms, size of garage, shape of the property, etc.
  - Random forest, XGBoost, Gaussian Process: use ensemble methods to enhance model performance
  - Neural Network