# Data

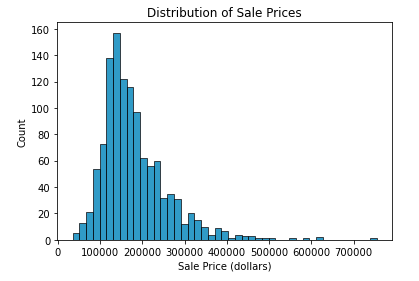
Our primary data source is the [Ames, Iowa housing sales dataset](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data) from Kaggle. The dataset contains data for 1,460 homes, spans 25 neighborhoods in Ames, and covers the years 2006 through 2010. The attributes include sale price, square footage, number of bedrooms, and 75 other features for each home.

We merged our primary dataset with neighborhood-level data on school quality, crime, walkability, median income, and unemployment rate. We sourced the school quality data from [GreatSchools.org](https://www.greatschools.org/), which assigns a numerical rating of 1 to 10 for each school district. We sourced the crime data from [www.Realtor.com](http://www.Realtor.com), which rates neighborhood crime levels on a scale of 1 to 5. Next, we sourced the walkability data from [WalkScore.com](https://www.walkscore.com/). For each neighborhood, we pulled both the walk score and the bike score, which respectively measure a neighborhood’s walkability and bike-ability on a numerical scale from 0 to 100. Finally, we sourced median income and unemployment rate data from the 2019 5-year American Community Survey.

Combining the neighborhood-level data with our primary dataset yielded a total of 83 features, which we will use to predict the sale price of homes.

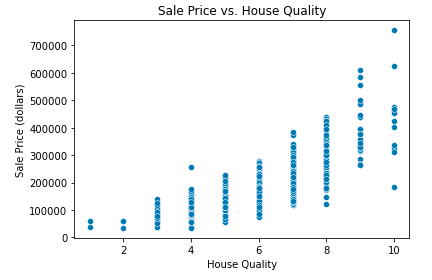
# Data Exploration: Distribution of Sale Price

The predicted variable (sale price) ranges from $34,900 to $755,000 in our training dataset. The distribution of sale price is plotted below.

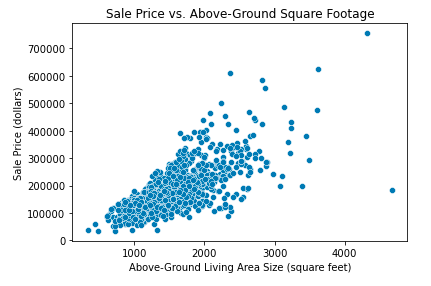


# Data Exploration: Home Features and Sale Price

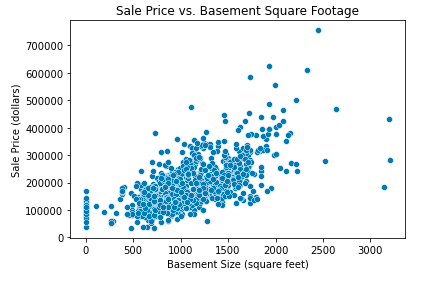
A number of numerical housing features in our dataset appear to have a strong correlation with sale price. One of these features is house quality, an ordinal variable measured on a scale of 1 to 10. Below, we present a plot that depicts a positive correlation between sale price and house quality.



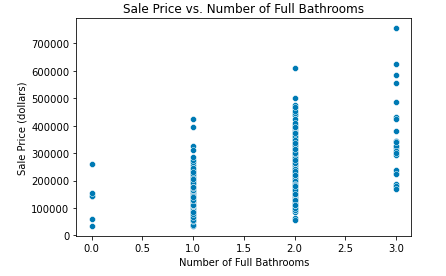
Another feature that exhibits a strong positive correlation with sale price is square footage. This is true of both above- and below-ground square footage. Below, we present a plot that depicts the relationship between sale price and above-ground square footage.



Next, we present a plot that displays the relationship between sale price and below-ground (basement) square footage.



Another feature that appears to be strongly correlated with sale price is the number of full bathrooms (i.e. the number of bathrooms containing a shower/tub). The positive correlation between sale price and number of full bathrooms is plotted below.



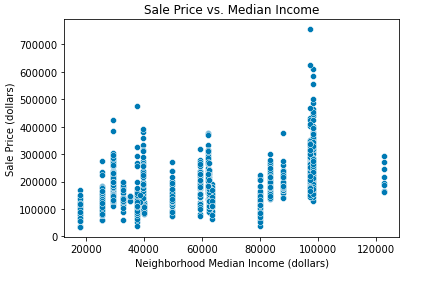
Interestingly, the relationship between sale price and number of bedrooms does not follow the same monotonically increasing pattern. As the number of bedrooms increases from 0 to 4, the sale price increases, but the sales price steeply drops off after the number of bedrooms increases to 5 and above. The relationship between sale price and number of bedrooms is plotted below.



# Data Exploration: Neighborhood-Level Features and Sale Price

In this section, we consider the relationship between sale price and our neighborhood-level features (i.e. school quality, crime, walk score, bike score, median income, unemployment rate).

First, let’s consider median income at the neighborhood level. As we can see form the plot below, sale price appears to have a weakly positive relationship with median income, with a few neighborhoods that have relatively high median income but relatively low home prices.



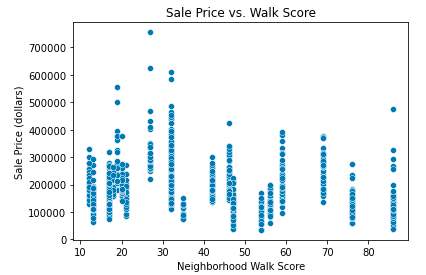
Next, let’s look at the relationship between sale price and the neighborhood-level unemployment rate. From the plot below, we can see that sale price exhibits a weak negative correlation with the unemployment rate. This makes sense as one would expect neighborhoods with more unemployed individuals to also have lower home prices. A few neighborhoods constitute notable exceptions, exhibiting a low unemployment rate and low housing prices.



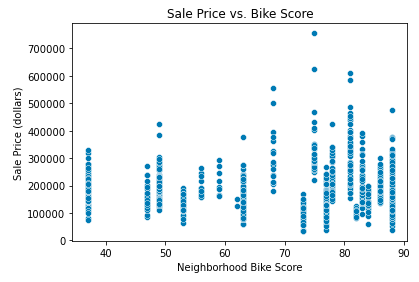
The crime rating is measured on a scale of 1 to 5. We can see from the plot below that sale price appears to be weakly decreasing in crime level.



Next, let’s look at the neighborhood-level walk score, which is measured on a scale of 0 to 100. Based on the plot below, the relationship between walk score and sale price is a bit ambiguous. At lower walk scores below 20, sale price remains low. As the walk scores increase from 20 to 30, sale price also increases substantially, but at higher walk scores above 30, there is not noticeable pattern in the relationship between sale price and walk score.



Unlike the walk score, the neighborhood-level bike score appears to exhibit a stronger correlation with sale price. Neighborhoods with higher bike scores (above 65) tend to also have higher home sale prices.



Finally, let’s consider school quality, which is measured on a scale of 1 to 10 in our dataset. There are only five school districts in Ames, so there is limited variation in school quality across neighborhoods. Nevertheless, we have plotted the relationship between school quality and sale price, and we observe that the highest-rated school district contains the most expensive homes, but in the other school districts, there is no observable relationship between sale price and school quality.

