4 The Ames housing data

The Ames housing data set (De Cock 2011) is an excellent resource for learning about models that we will use throughout this book. It contains data on 2,930 properties in Ames, Iowa, including columns related to

- house characteristics (bedrooms, garage, fireplace, pool, porch, etc.),
- location (neighborhood),
- lot information (zoning, shape, size, etc.),
- ratings of condition and quality, and
- sale price.

> Our goal for these data is to predict the sale price of a house based on its other characteristics.

The raw data are provided by the authors, but in our analyses in this book, we use a transformed version available in the **modeldata** package. This version has several changes and improvements to the data[8]. For example, the longitude and latitude values have been determined for each property. Also, some columns were modified to be more analysis ready. For example:

- In the raw data, if a house did not have a particular feature, it was implicitly encoded as missing. For example, there were 2,732 properties that did not have an alleyway. Instead of leaving these as missing, they were relabeled in the transformed version to indicate that no alley was available.
- The categorical predictors were converted to R's factor data type. While both the tidyverse and base R have moved away from importing data as factors by default, this data type is a better approach for storing qualitative data for *modeling* than simple strings.

- We removed a set of quality descriptors for each house since they are more like outcomes than predictors.

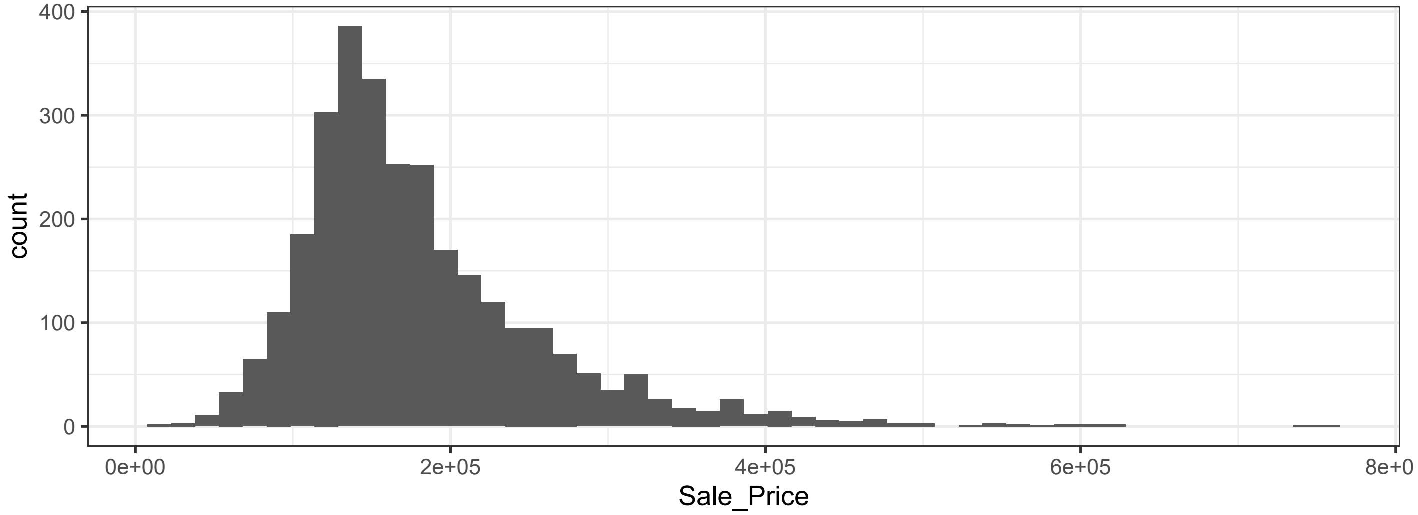
To load the data:

```r
library(modeldata) # This is also loaded by the tidymodels package
data(ames)

# or, in one line:
data(ames, package = "modeldata")

dim(ames)
#> [1] 2930   74
```

4.1 EXPLORING IMPORTANT FEATURES

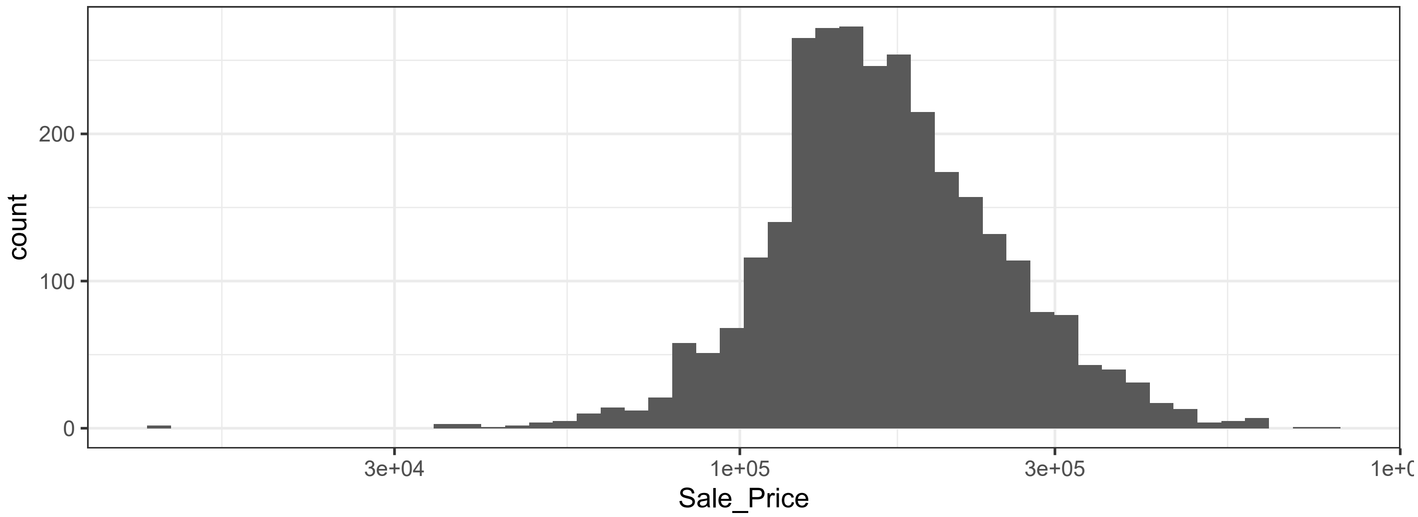
It makes sense to start with the outcome we want to predict: the last sale price of the house (in USD):

```r
library(tidymodels)
tidymodels_prefer()

ggplot(ames, aes(x = Sale_Price)) +
  geom_histogram(bins = 50)
```



The data are right-skewed; there are more inexpensive houses than expensive ones. The median sale price was $160,000 and the most expensive house was $755,000. When modeling this outcome, a strong argument can be made that the price should be log-transformed. The advantages of doing this are that no houses would be predicted with negative sale prices and that errors in predicting expensive houses will not have an undue influence on the model. Also, from a statistical perspective, a logarithmic transform may also *stabilize the variance* in a way that makes inference more legitimate. Let's visualize the transformed data:

```
ggplot(ames, aes(x = Sale_Price)) +
  geom_histogram(bins = 50) +
  scale_x_log10()
```



While not perfect, this will probably result in better models than using the untransformed data.

> The downside to transforming the outcome is mostly related to interpretation.

The units of the model coefficients might be more difficult to interpret, as will measures of performance. For example, the root mean squared error (RMSE) is a common performance metric that is used in regression models. It uses the difference between the observed and predicted values in its calculations. If the sale price is on the log scale, these differences (i.e. the residuals) are also in log units. For this reason, it can be difficult to understand the quality of a model whose RMSE is 0.15 log units.

Despite these drawbacks, the models used in this book utilize the log transformation for this outcome. *From this point on*, the outcome column is pre-logged in the `ames` data frame:

```r
ames <- ames %>% mutate(Sale_Price = log10(Sale_Price))
```

Another important aspect of these data for our modeling are their geographic locations. This spatial information is contained in the data in two ways: a qualitative `Neighborhood` label as well as quantitative longitude and latitude data. To visualize the spatial information, let’s use both together to plot the data on a map and color by neighborhood:

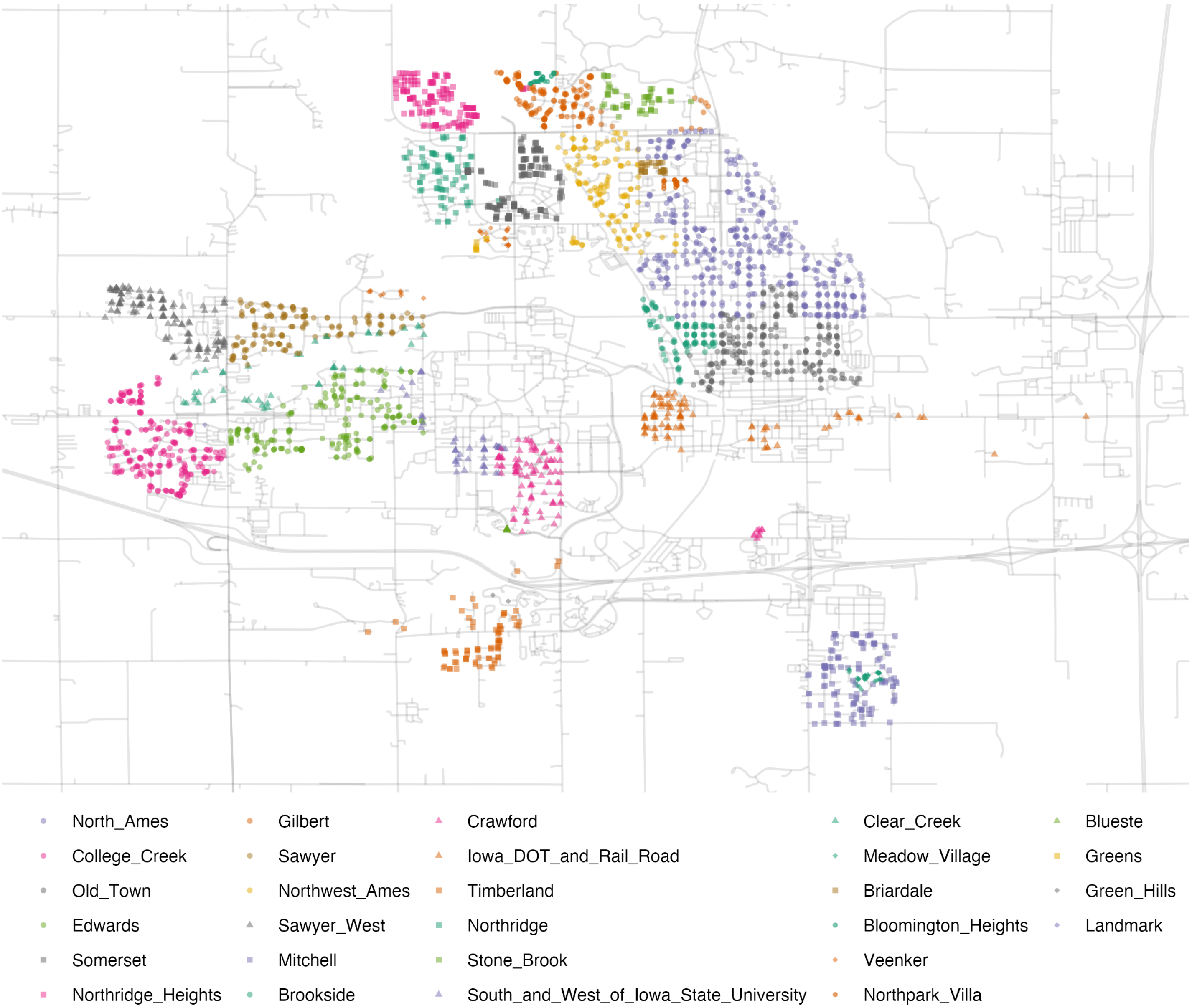
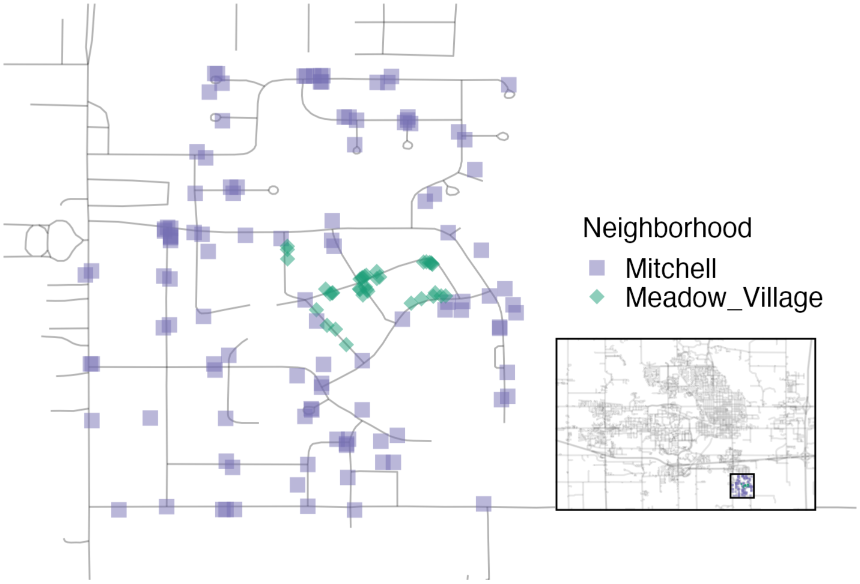


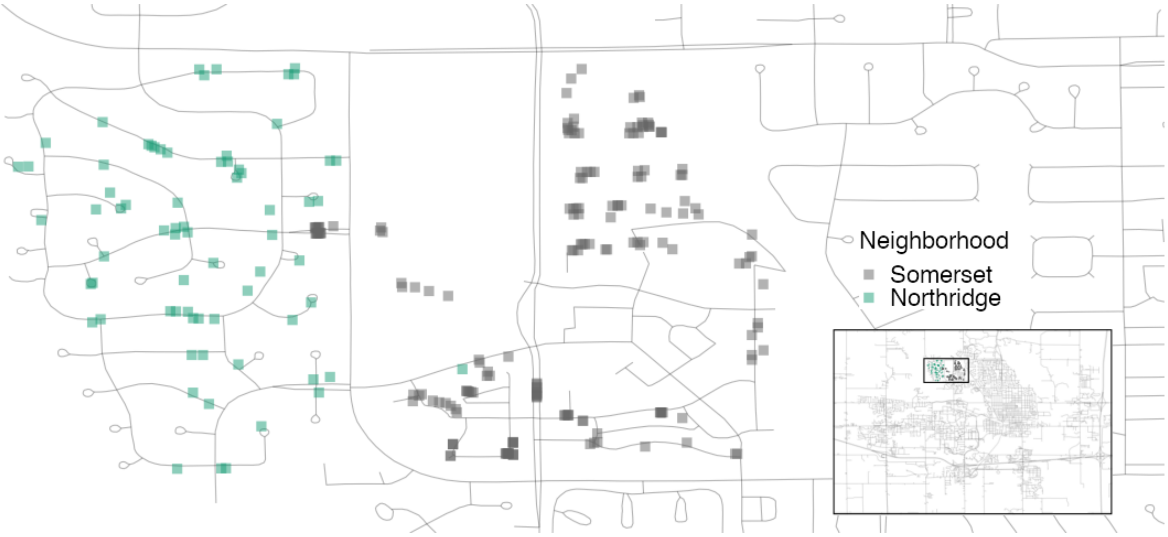
Figure 4.1: Neighborhoods in Ames IA

We can see a few noticeable patterns. First, there is a void of data points in the center of Ames. This corresponds to Iowa State University. Second, while there are a number of neighborhoods that are geographically isolated, there are others that are adjacent to each other. For example, Timberland is located apart from almost all other neighborhoods:

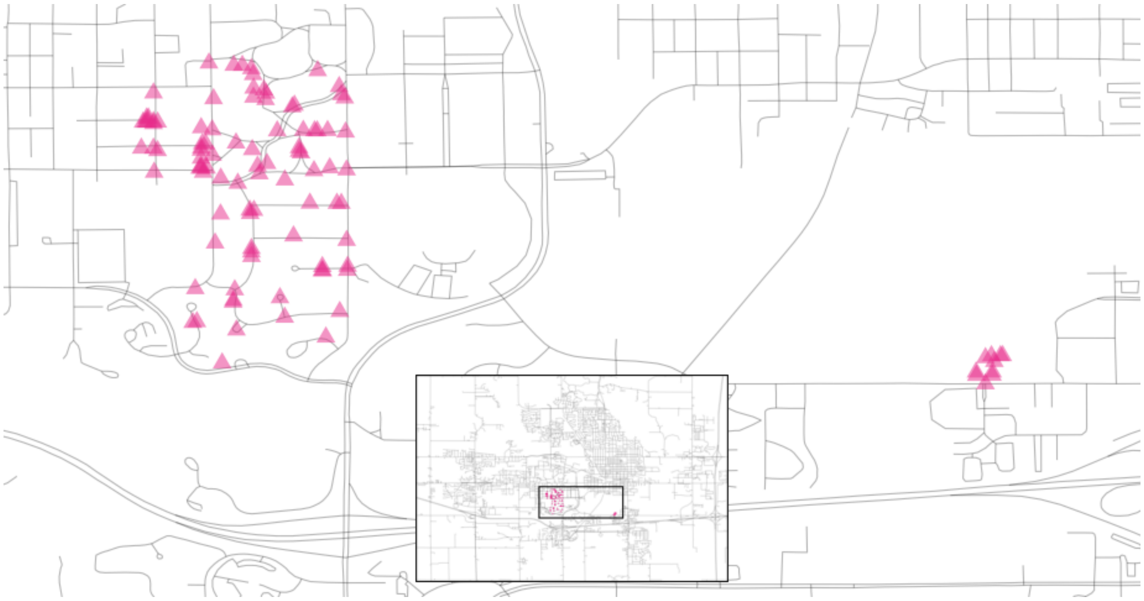
The Meadow Village neighborhood in Southwest Ames is like an island of properties ensconced inside the sea of properties that make up the Mitchell neighborhood:



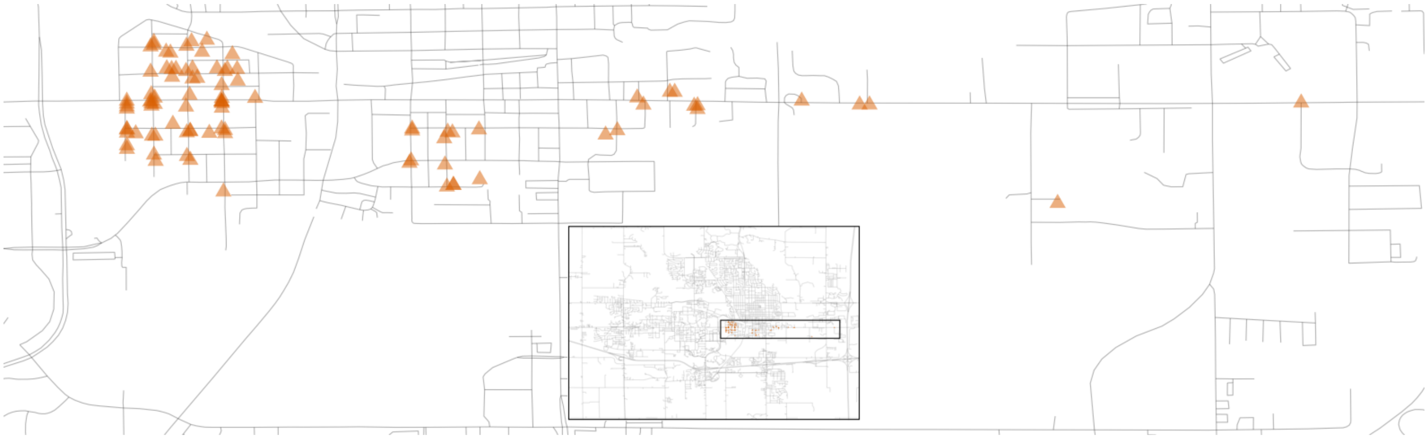
A detailed inspection of the map also shows that the neighborhood labels are not completely reliable. For example, there are some properties labeled as being in Northridge that are surrounded by houses in the adjacent Somerset neighborhood:



Also, there are ten isolated houses labeled as being in Crawford but are not close to the majority of the other houses in that neighborhood:



Also notable is the “Iowa Department of Transportation (DOT) and Rail Road” neighborhood adjacent to the main road on the east side of Ames. There are several clusters of houses within this neighborhood as well as some longitudinal outliers; the two houses furthest east are isolated from the other locations.



As previously described in Chapter 1, it is critical to conduct *exploratory data analysis* prior to beginning any modeling. These housing data have characteristics that present interesting challenges about how the data should be processed and modeled. We describe many of these in later chapters. Some basic questions that could be examined include:

- Are there any odd or noticeable things about the distributions of the individual predictors? Is there much skewness or any pathological distributions?
- Are there high correlations between predictors? For example, there are multiple predictors related to the size of the house. Are some redundant?
- Are there associations between predictors and the outcomes?

Many of these questions will be revisited as these data are used in upcoming examples.

4.2 CHAPTER SUMMARY

This chapter introduced a data set used in later chapters to demonstrate tidymodels syntax and investigated some of its characteristics. Exploratory data analysis like this is an essential component of any modeling project; EDA uncovers information that contributes to better modeling practice.

The important code that we will carry forward into subsequent chapters is:

```
library(tidymodels)
data(ames)
ames <- ames %>% mutate(Sale_Price = log10(Sale_Price))
```

REFERENCES

De Cock, D. 2011. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” *Journal of Statistics Education* 19 (3).

[8] For a complete account of the differences, see this script.↩