Notes on Ch4 - The Ames Housing Data

N_Lim

2025-06-25

Loading the data:

```
library(modeldata)
data(ames)
```

checking the data

```
dim(ames)
```

```
## [1] 2930   74
```

basic EDA on ames data

```
# Load packages
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.3.0 --
```

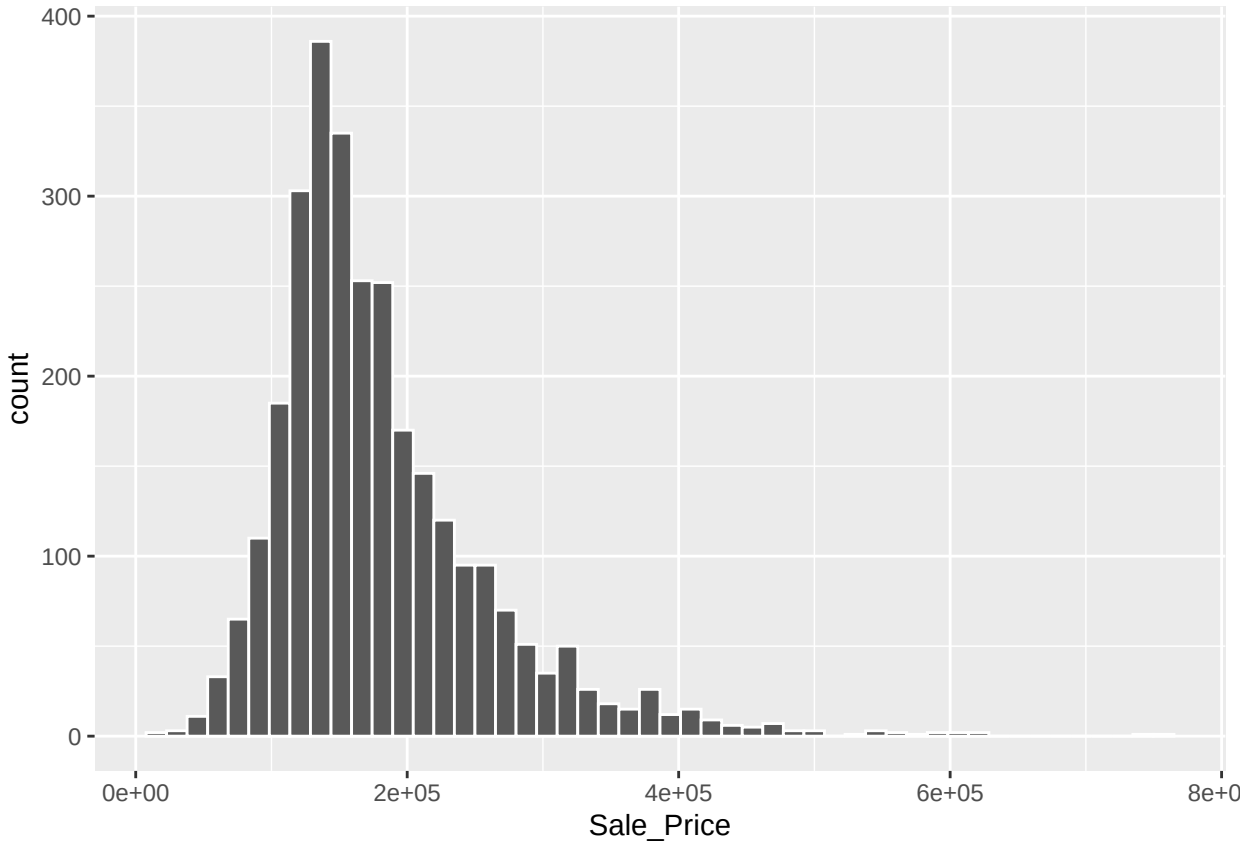
```
## v broom        1.0.8     v rsample      1.2.1
## v dials        1.4.0     v tibble       3.2.1
## v dplyr        1.1.4     v tidyr        1.3.1
## v ggplot2      3.5.2     v tune         1.3.0
## v infer        1.0.8     v workflows    1.2.0
## v parsnip      1.3.2     v workflowsets 1.1.1
## v purrr        1.0.4     v yardstick    1.3.2
## v recipes      1.3.1
```

```
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
```

```
tidymodels_prefer()
```

Actual start of EDA

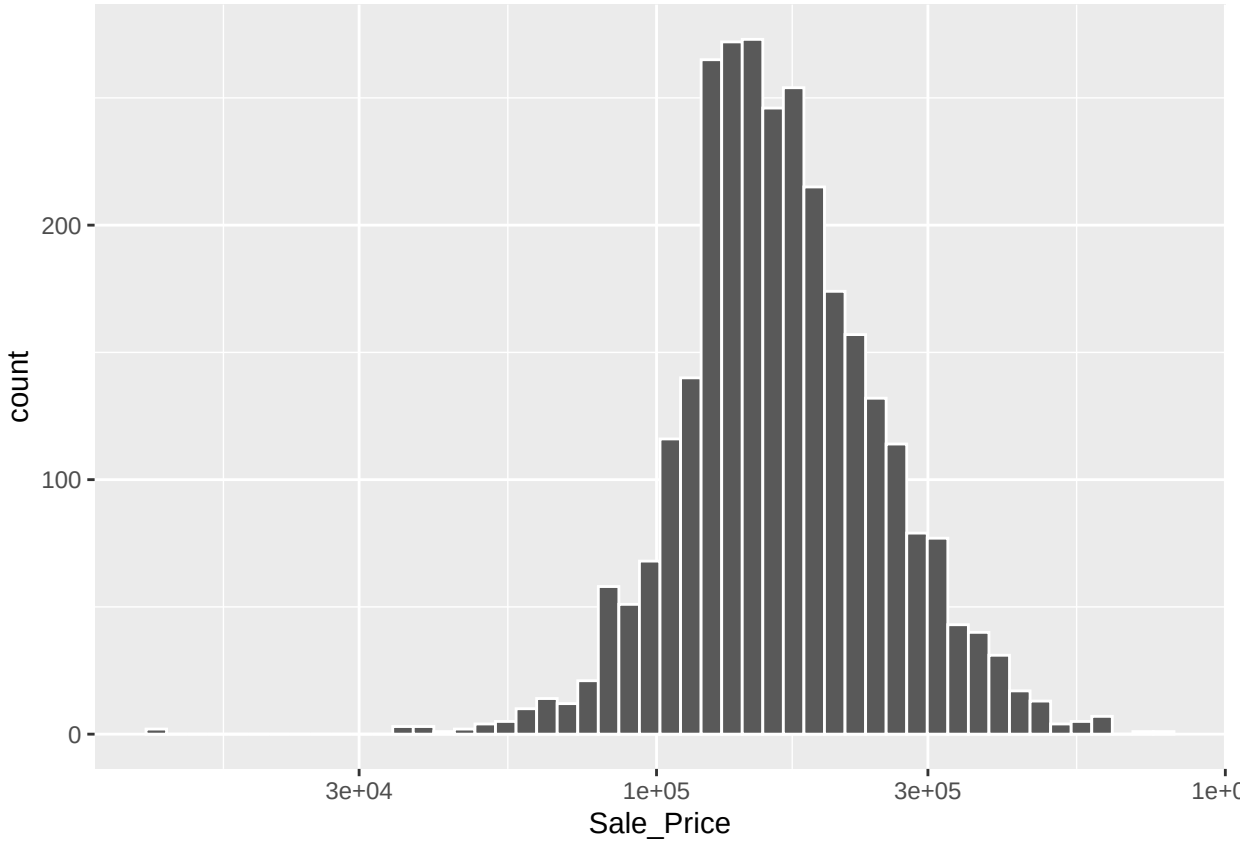
```
ggplot(ames, aes(x = Sale_Price)) +
  geom_histogram(bins = 50, col = "white")
```



Observations on the plot: - data are right-skewed - there are more expensive houses in the dataset than inexpensive ones - log-transforming the data first before modeling is a good decision - log-transformation will also stabilize the variance in the data, making our inference more 'legit'

Plot of log-transformed data:

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



While not perfect, the data looks closer to bell curve now...

Proceeding to log-transform the data:

```
ames <- ames |>
  mutate(Sale_Price = log10(Sale_Price))
```