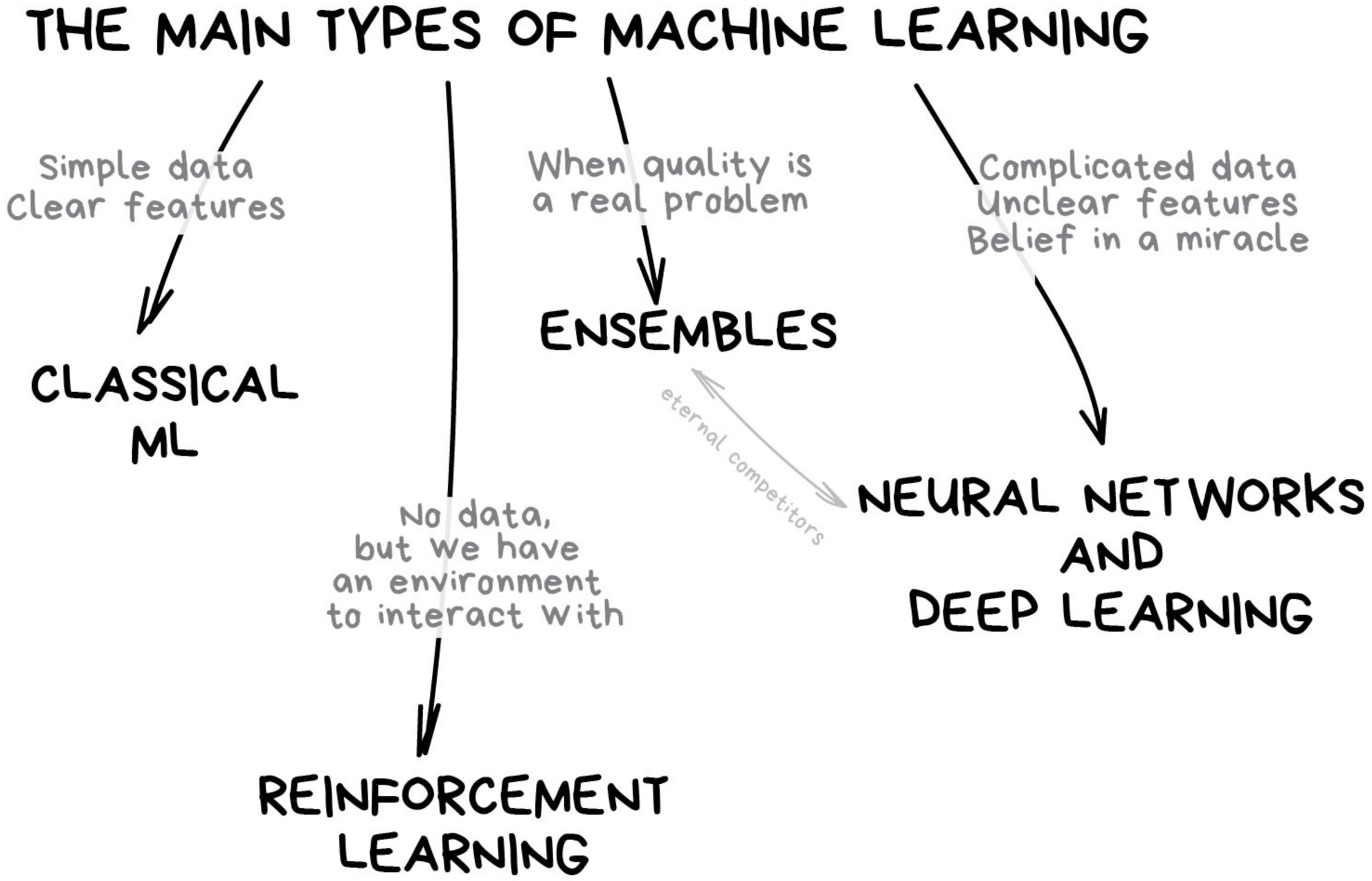
# Linear Regression

## And some ML


THE MAIN TYPES OF MACHINE LEARNING
Simple data
Clear features
CLASSICAL ML
No data,
but we have
an environment
to interact with
REINFORCEMENT LEARNING
When quality is
a real problem
ENSEMBLES
eternal competitors
Complicated data
Unclear features
Belief in a miracle
NEURAL NETWORKS AND DEEP LEARNING

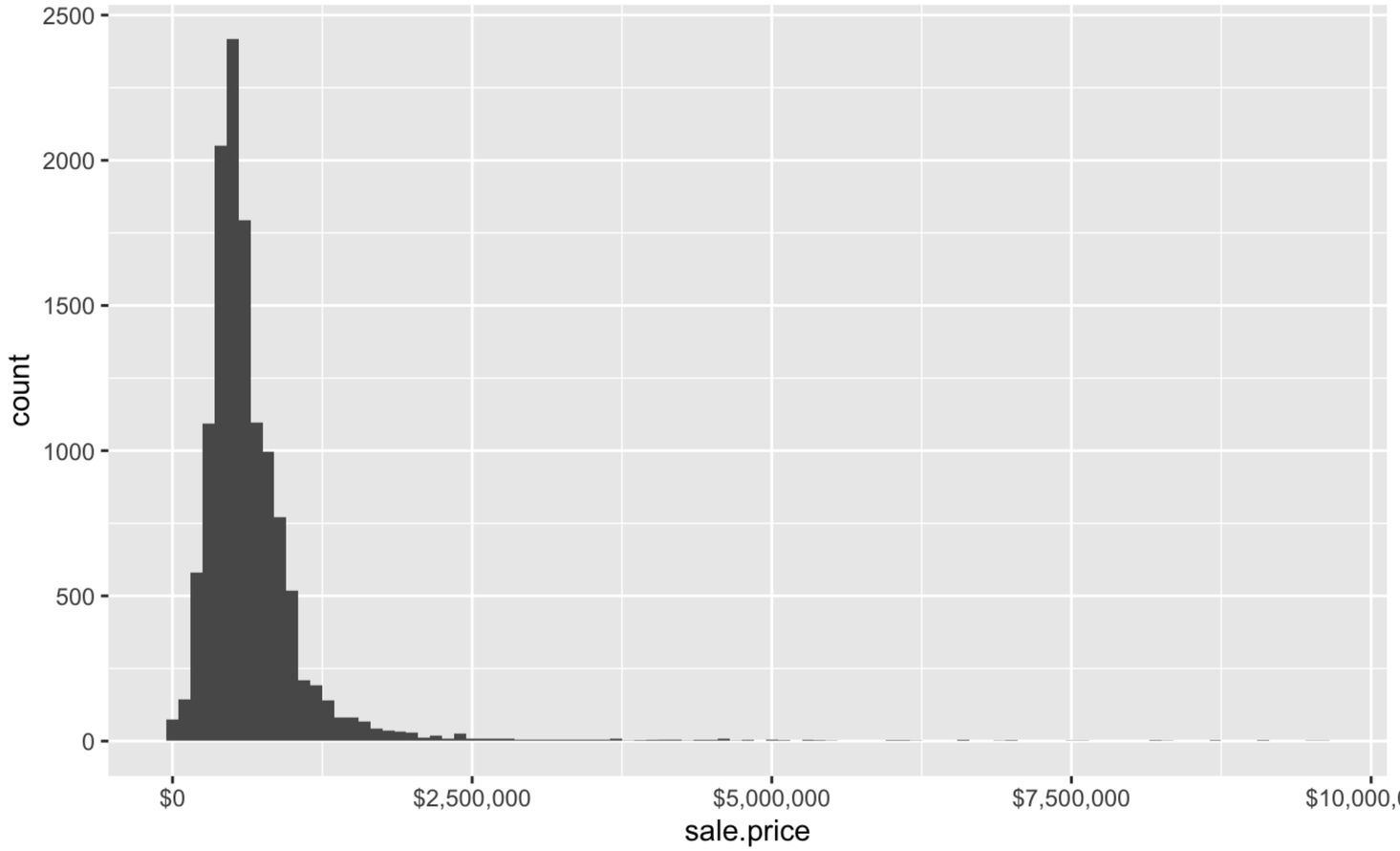
# Classical ML

- Can be broken down into two main categories
  - Supervised ML
    - Regression - Predict **value** based on past
    - Classification - Predict **group** based on past
  - Unsupervised ML
    - Clustering - Break into **groups**
    - Dimensionality Reduction - “Important” components
    - Anomaly Detection - “Weirdness”

# Linear Regression

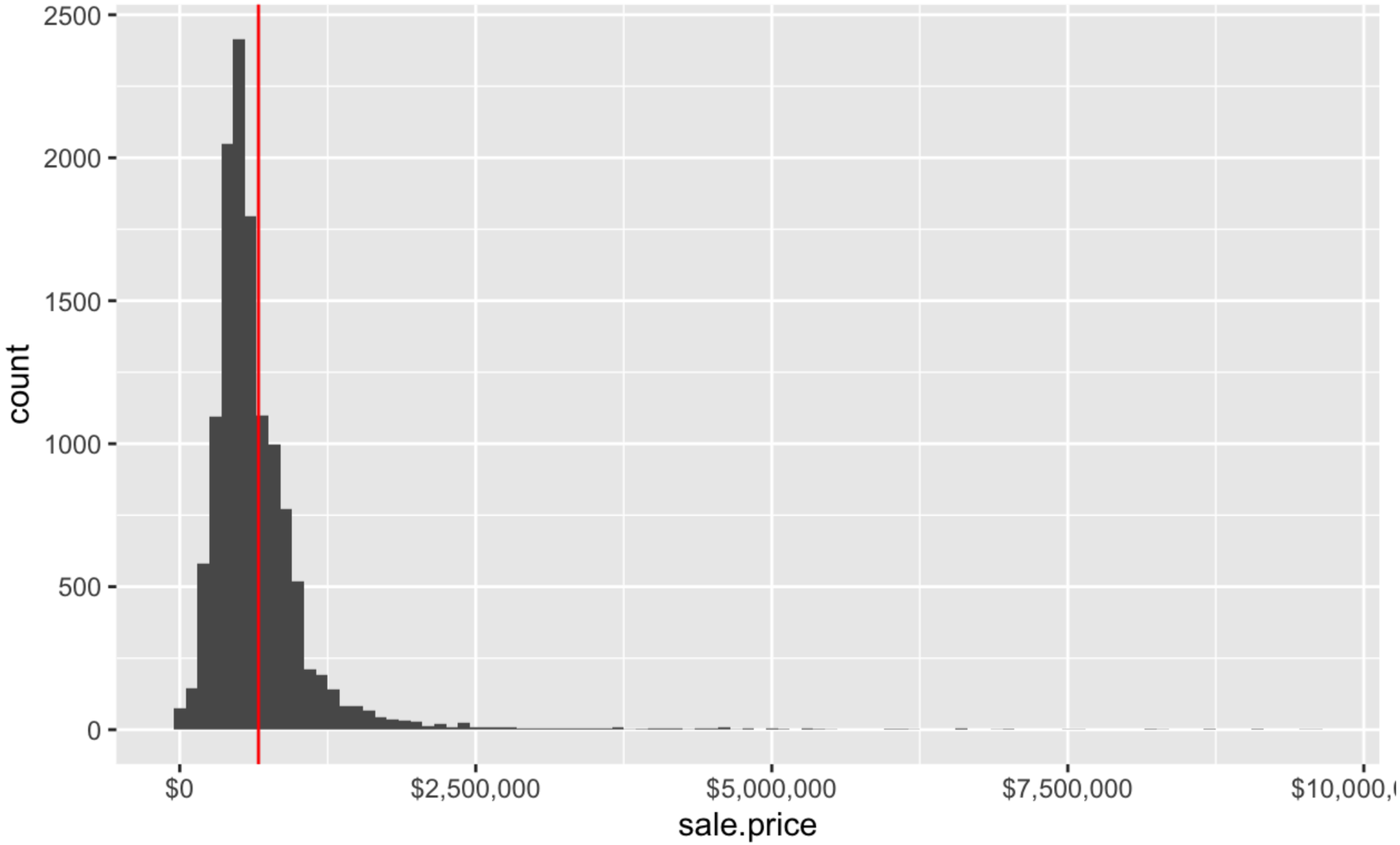
## Intuition

```
[1]  241500  180000  246000  255000  380250  325000  340000  210000  305000  365000  365000  375000  420000
```



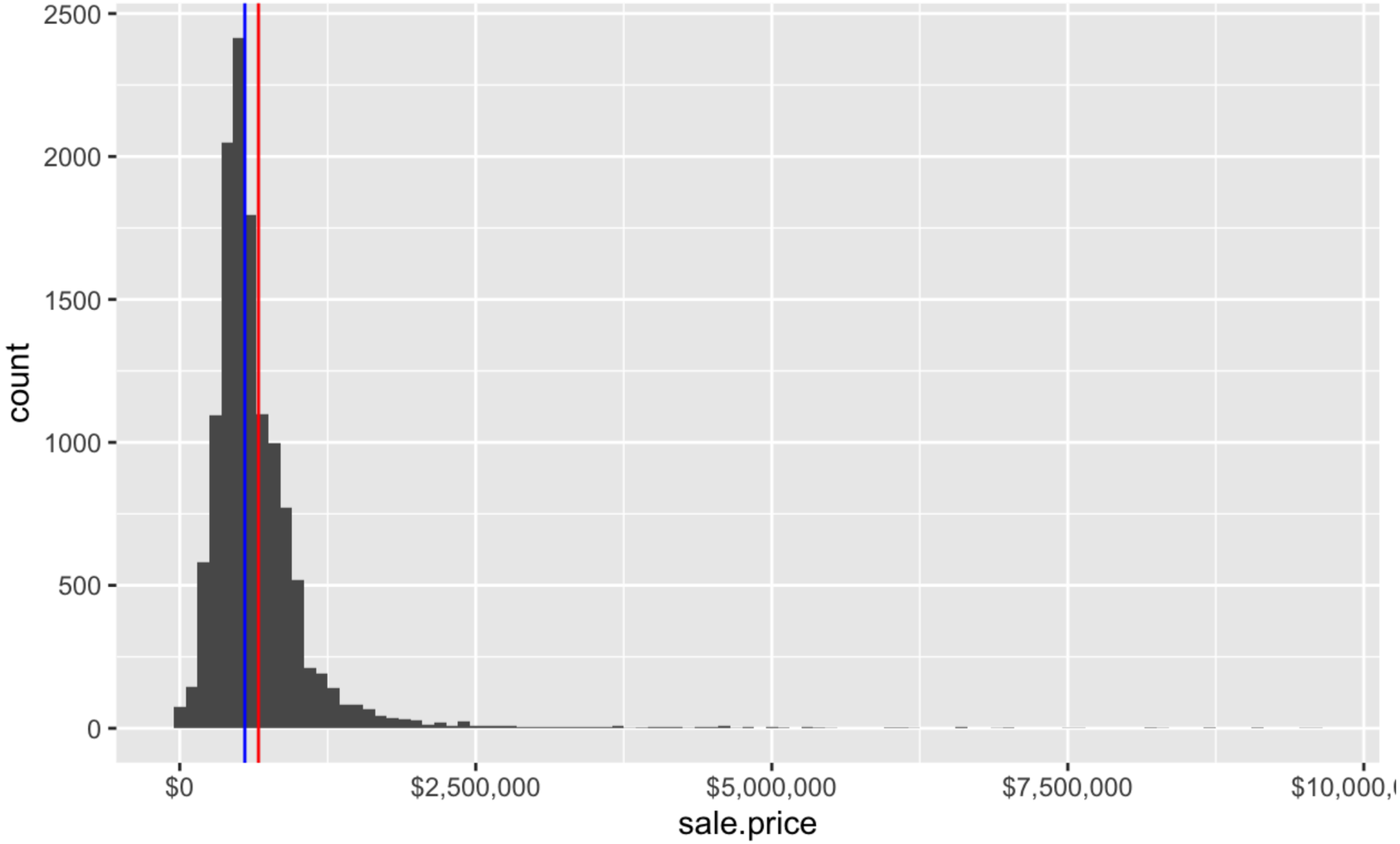
# Linear Regression

## Intuition



# Linear Regression

## Intuition



# What is the error?

```
home_sales_nyc %>%
  mutate(
    error = (sale.price - mean(sale.price)),
   ) %>%
  summarise(mean_error = mean(error))
```

A tibble: 1 × 3

| mean_error <dbl> |
|---|
| 0 |

1 row

# What is the error?

```
home_sales_nyc %>%
  mutate(
    sq_error = (sale.price - mean(sale.price))^2,
   ) %>%
  summarise(mean_sq_error = mean(sq_error))
```

A tibble: 1 × 3

| mean_error <dbl> | mean_sq_error <dbl> |
|---|---|
| 0 | 288431380698 |

1 row

# What is the error?

```
home_sales_nyc %>%
  mutate(
    sq_error = (sale.price - mean(sale.price))^2,
   ) %>%
  summarise(root_mean_sq_error = sqrt(mean(sq_error)))
```

A tibble: 1 × 3

| mean_error <dbl> | mean_sq_error <dbl> | root_mean_sq_error <dbl> |
|---|---|---|
| 0 | 288431380698 | 537058 |

1 row

# Linear Regression

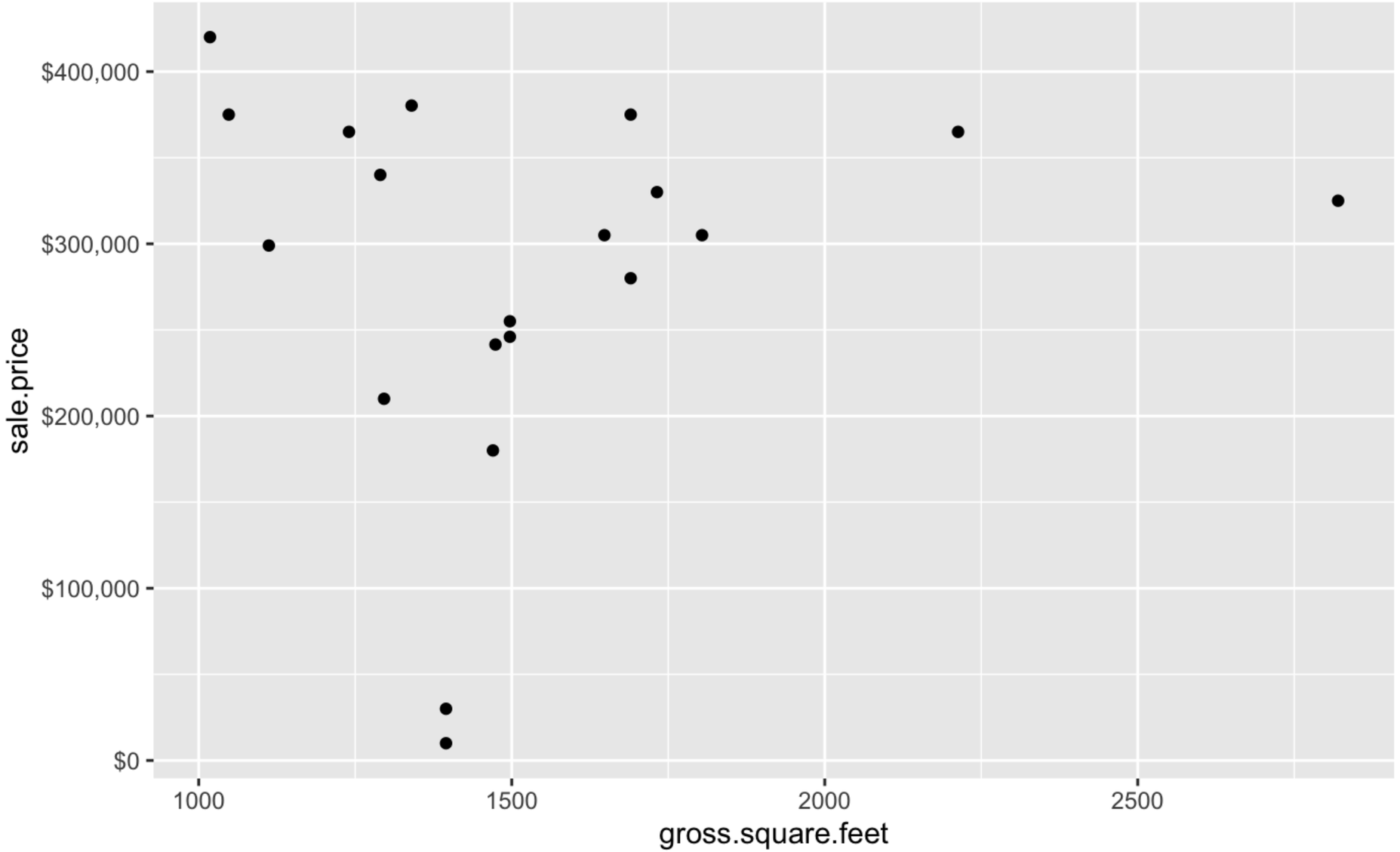
## Intuition

A tibble: 12,668 × 2

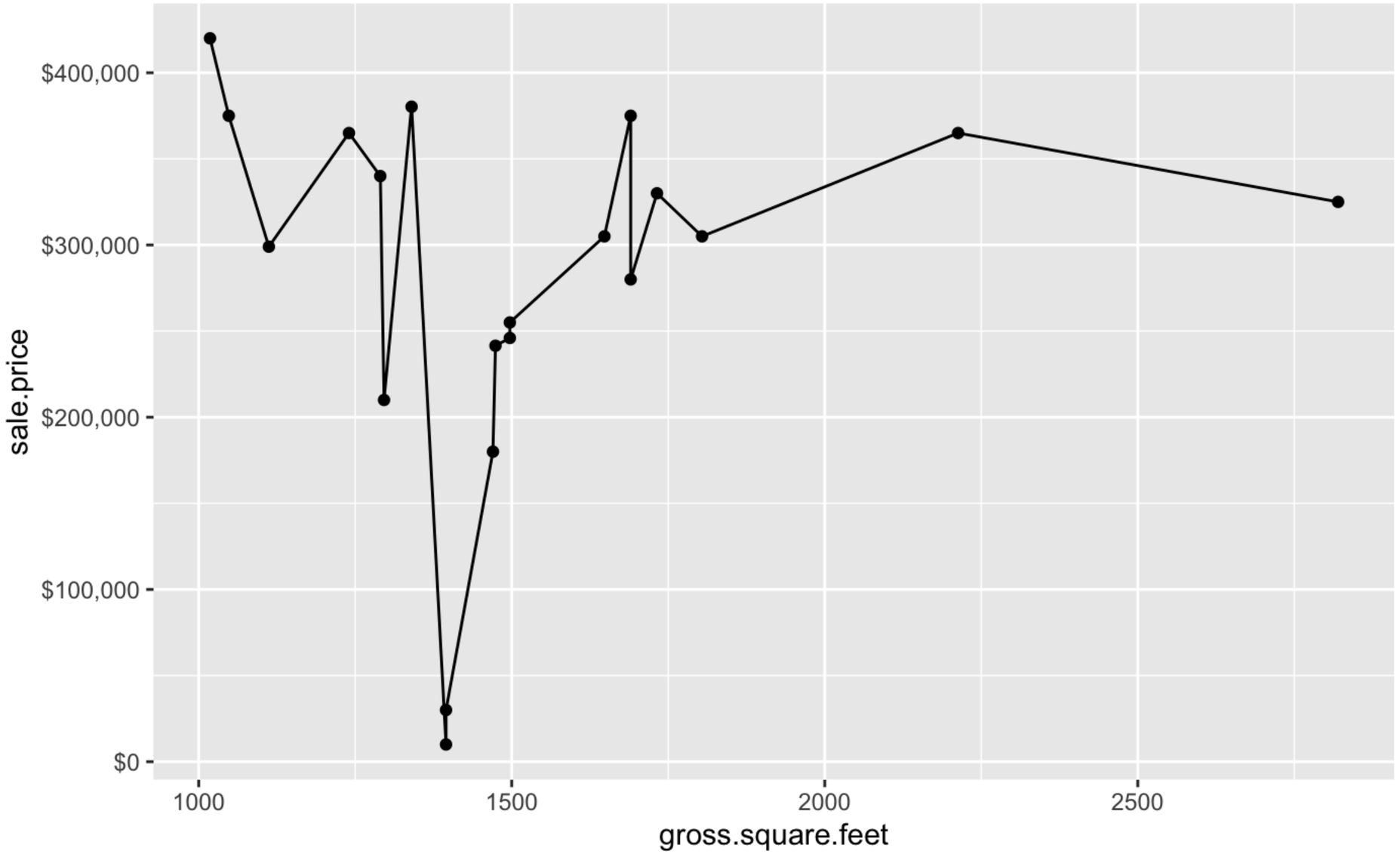
| land.square.feet <dbl> | sale.price <dbl> |
|---|---|
| 2500 | 241500 |
| 1578 | 180000 |
| 1694 | 246000 |
| 1694 | 255000 |
| 3525 | 380250 |
| 1293 | 325000 |
| 1103 | 340000 |
| 1650 | 210000 |
| 1791 | 305000 |
| 2356 | 365000 |

1–10 of 12,668 rows

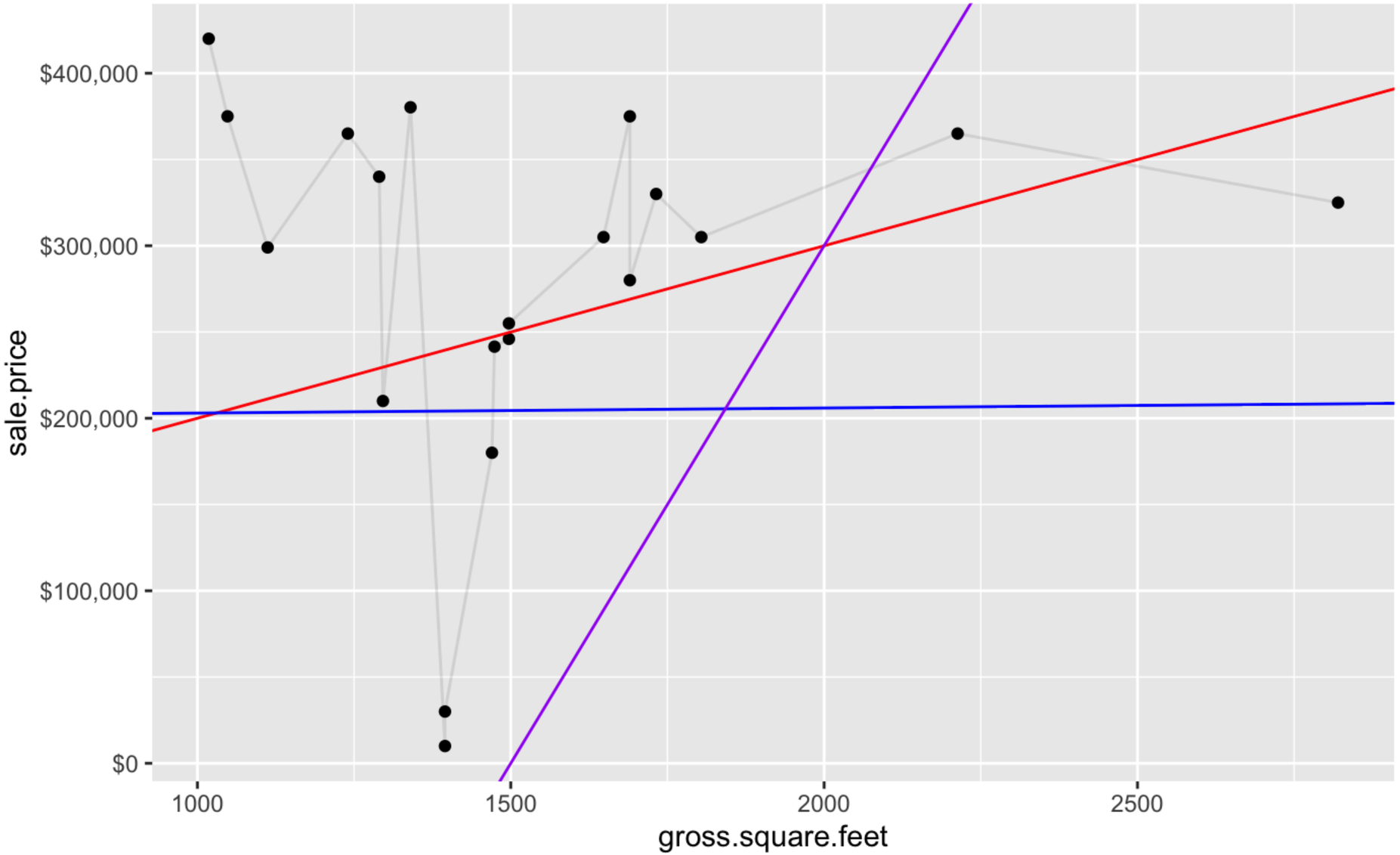
# Relationship?



# Relationship?



# Relationship?



# Let’s define this process

- Let there be some ***feature(s)*** $x$
- Let there be some ***outcome*** $y$
- Then, there exist some ***hypothesis*** $h$ of the relationship between $x$ and $y$
- This hypothesis relies on ***parameters*** $K$
- Using some ***loss*** $L$ that we try to minimize, we estimate the best ***parameters*** for the given ***hypothesis***

# Bias Variance tradeoff

## overfitting and underfitting