Descriptive Statistics

POLS 602

Dr. Mike Burnham
Texas A&M Political Science

Announcements

Data Exploration

STAR

- Student Teacher Achievement Ratio
- class size -> student achievement

```{r}
star <- read.csv('https://raw.githubusercontent.com/MLBurnham/pols_602/refs/heads/main/data/STAR.csv')
head(star)
```

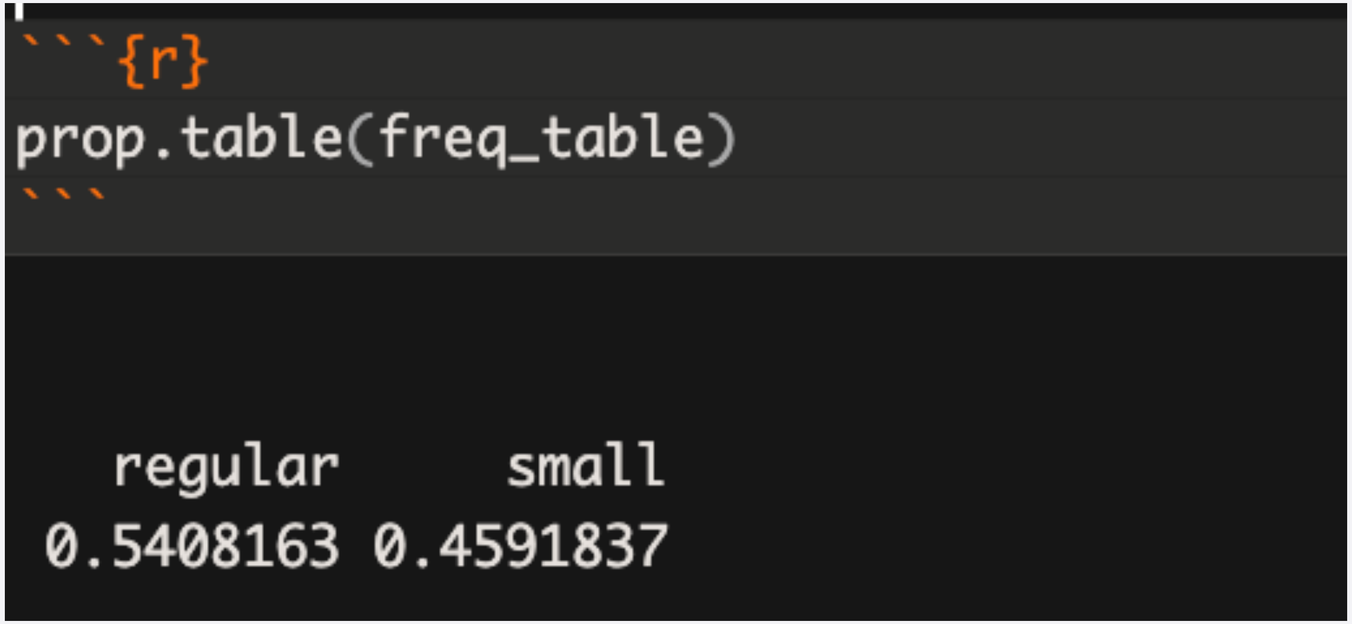
	classtype <chr>	reading <int>	math <int>	graduated <int>
1	small	578	610	1
2	regular	612	612	1
3	regular	583	606	1
4	small	661	648	1
5	small	614	636	1
6	regular	610	603	0

Frequency Table

```r
freq_table <- table(star$classtype)
freq_table
```

```
 regular    small
     689      585
```

Proportion Table



Central Tendency

Arithmetic Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

\bar{x} = \frac{\sum_{i=1}^{n} x_{i}}{n}

- What people are generally referring to when they say “average”
- Includes every data point in its calculation
- Not a robust statistic
- Influenced by outliers
- A bar over a variable indicates the arithmetic mean of that variable: $\bar{x}$ (x bar)

Arithmetic Mean

```r
mean(star$reading)
# If there is missing data in your vector:
mean(star$reading, na.rm = TRUE)
```

```
[1] 628.803
[1] 628.803
```

Median

- The middle number in a sorted list of all the numbers
- Robust statistic

Median

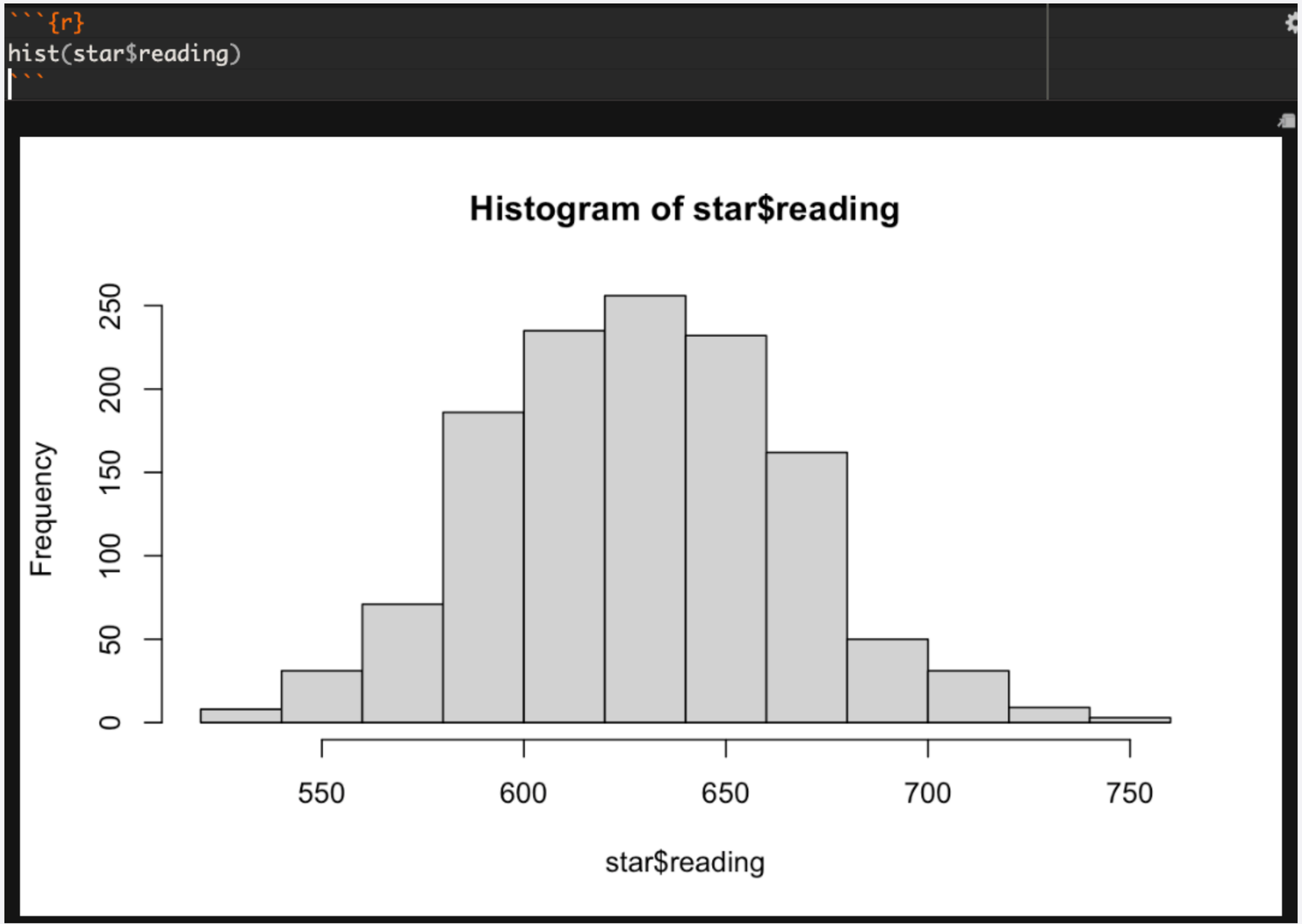
```r
median(star$reading)
# If there is missing data in your vector:
median(star$reading, na.rm = TRUE)
```

```
[1] 629
[1] 629
```

Spread

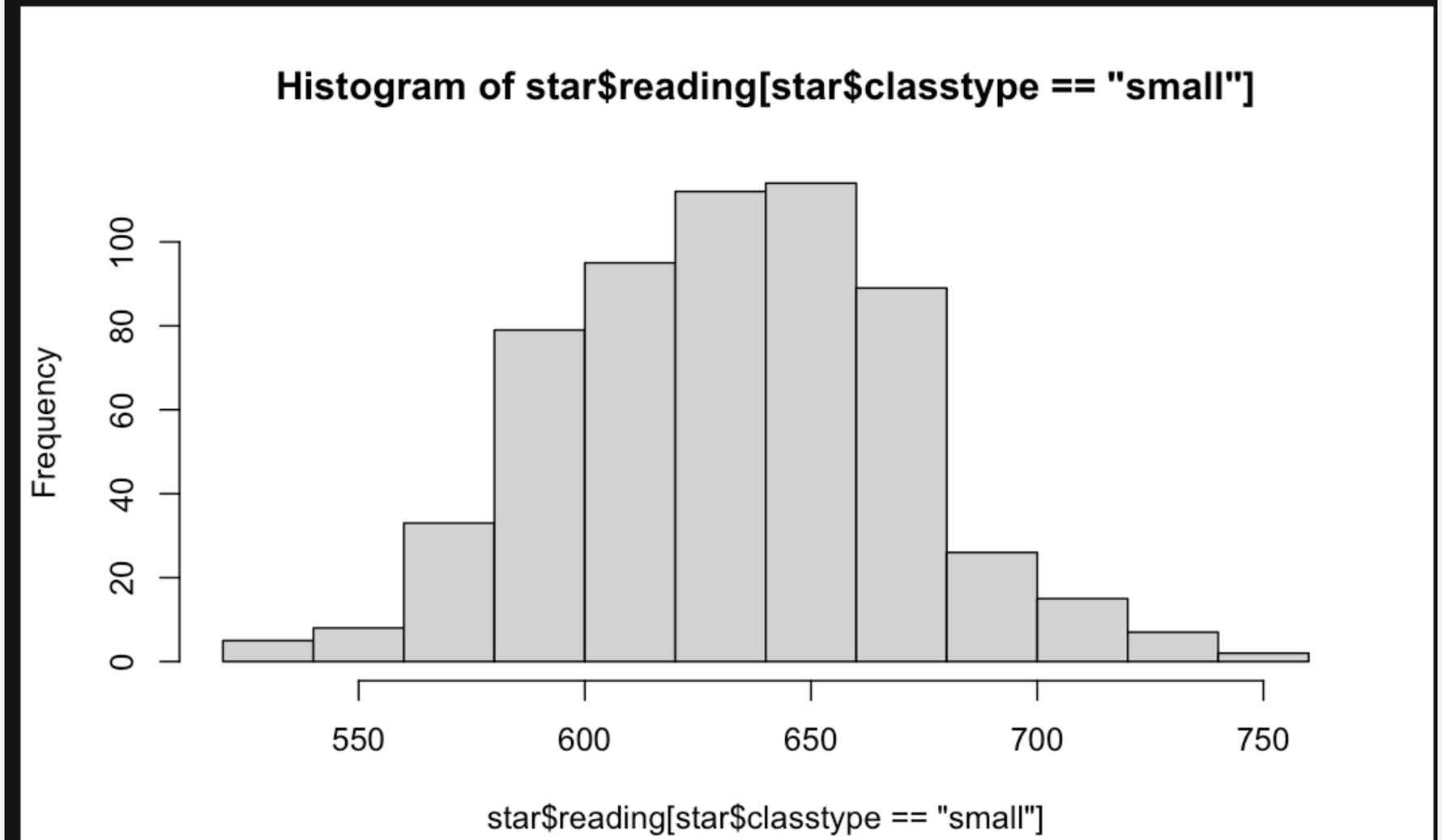
Visualizing Spread

```r
hist(star$reading)
```



Visualizing Spread

```r
hist(star$reading[star$classtype == 'small'])
```



Standard Deviation

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}

$$\sigma = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)$$

\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}

Standard Deviation

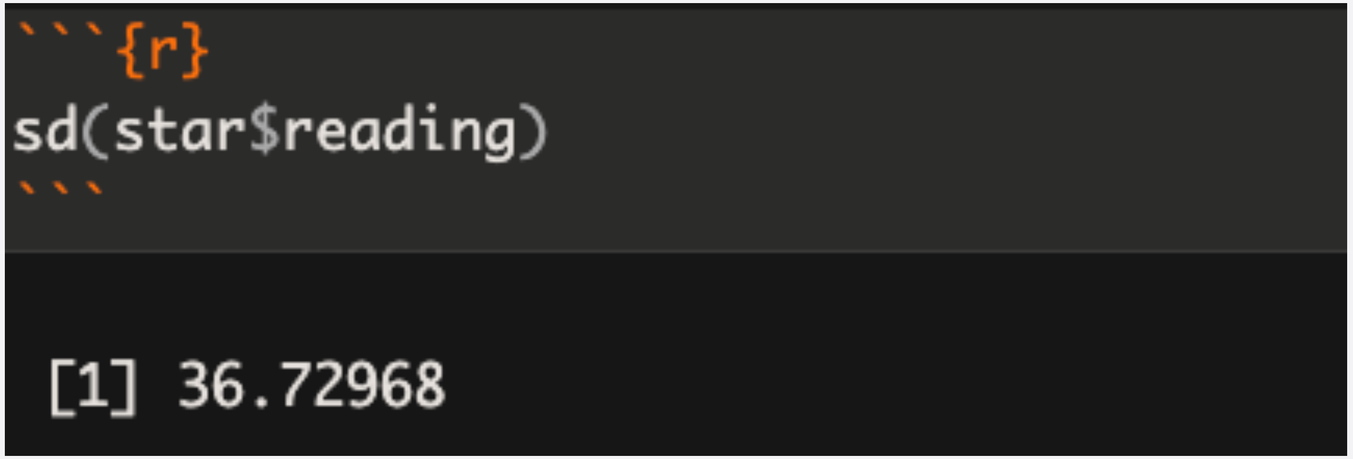
$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

```
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}
```

$$s = \sqrt{\frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

```
s = \sqrt{\frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2}
```

Standard Deviation



Variance

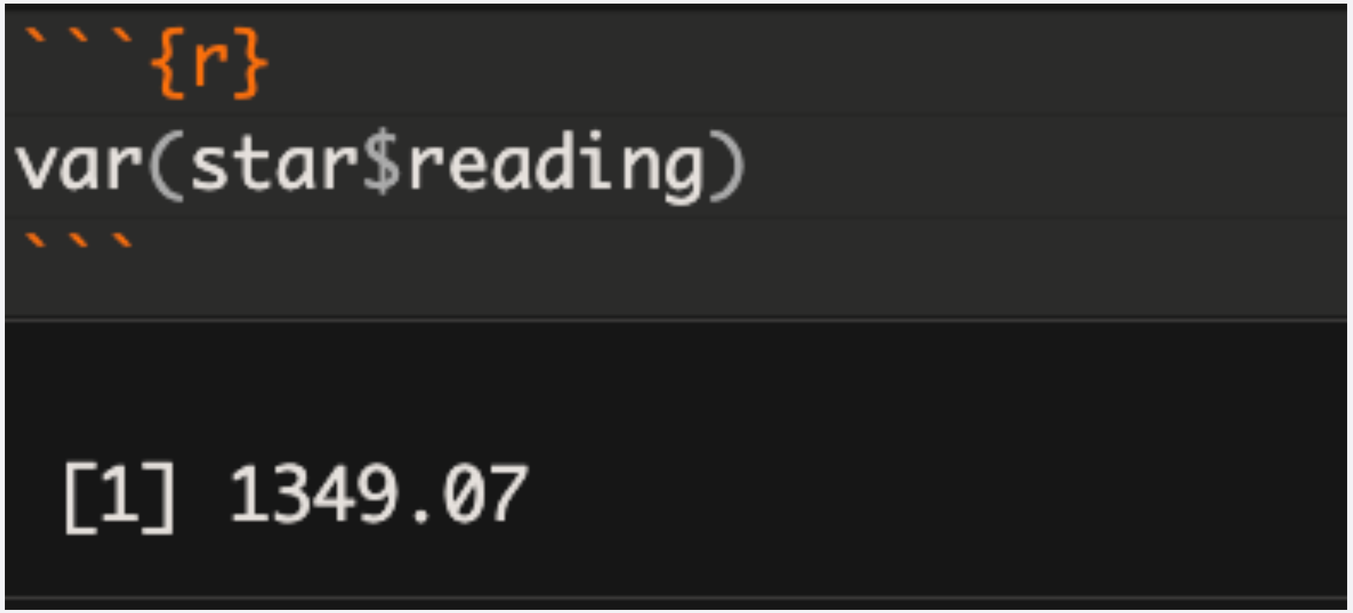
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

s^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2

Variance



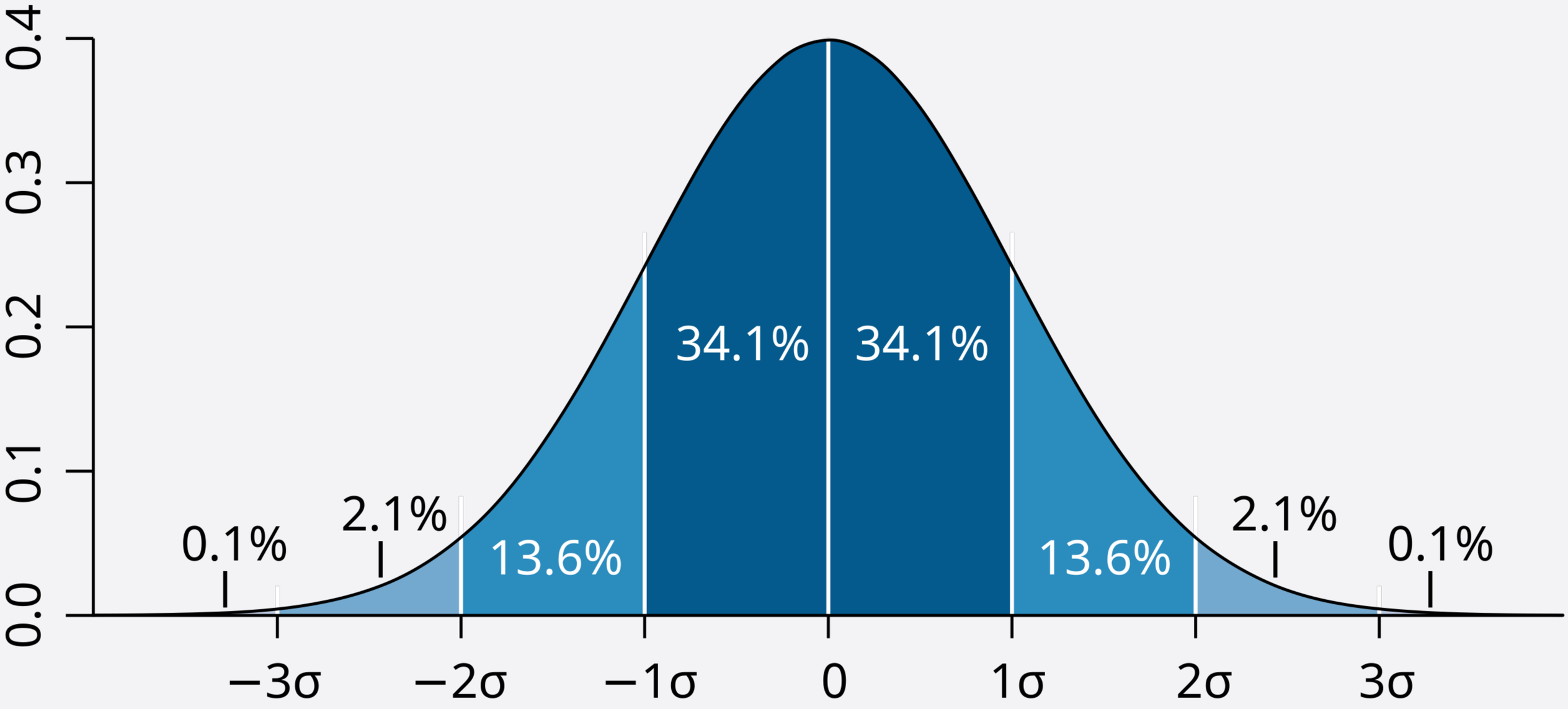
Standard Deviation

```r
sd(star$reading)
```

```
[1] 36.72968
```

```r
sqrt(var(star$reading))
```

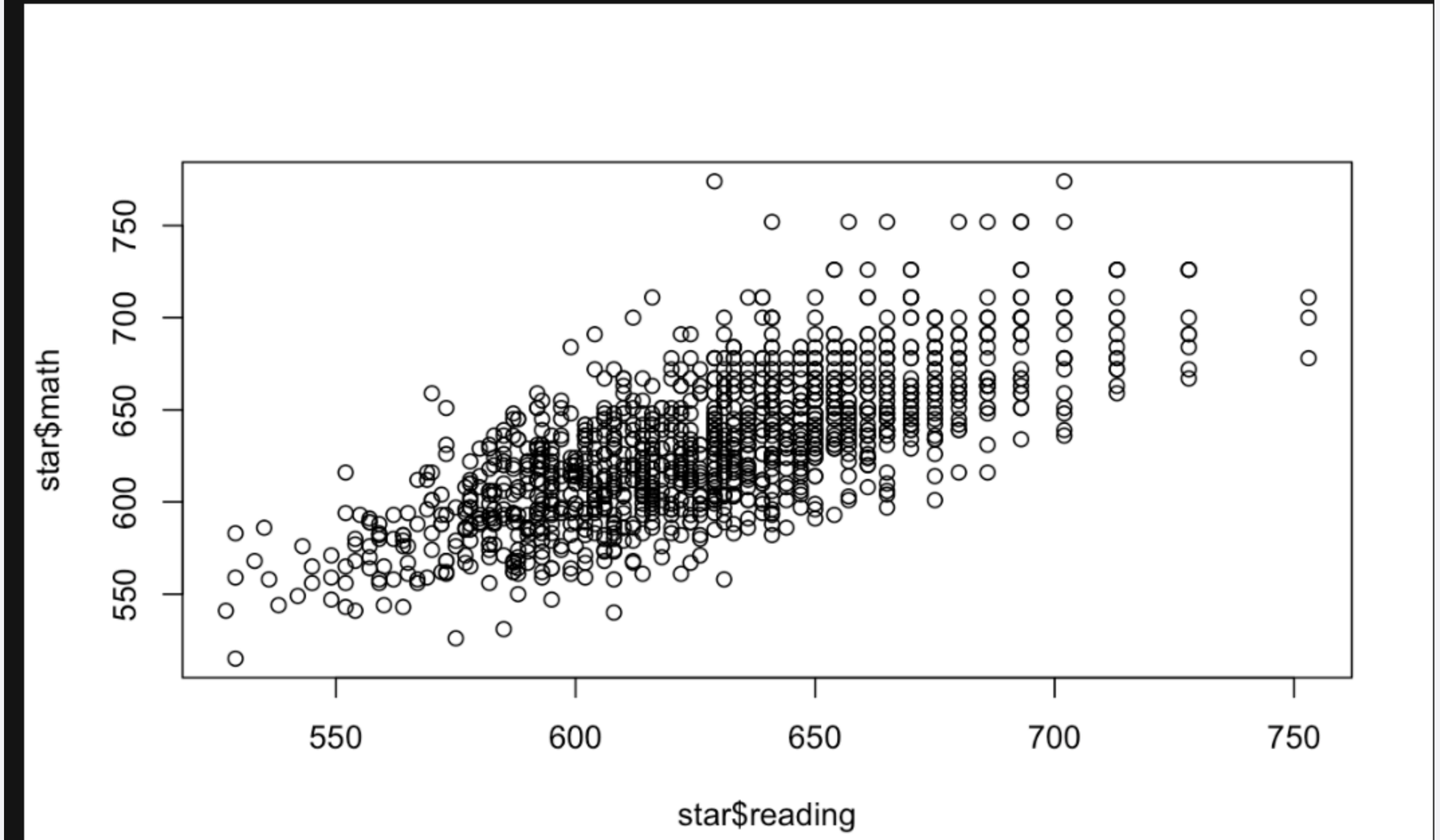
```
[1] 36.72968
```


0.0
0.1
0.2
0.3
0.4
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

Correlation

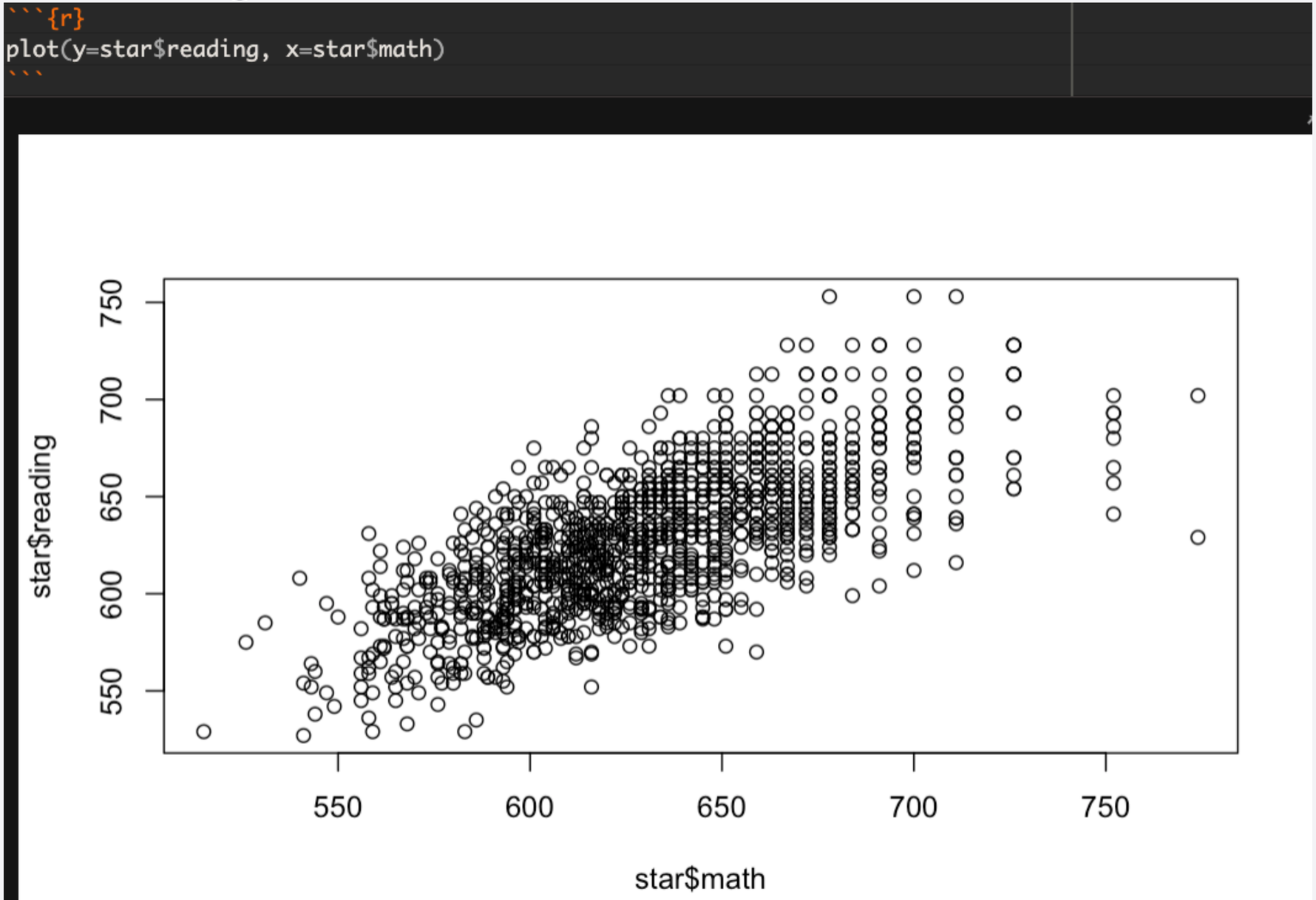
Visualizing Correlation

```r
plot(star$reading, star$math)
```



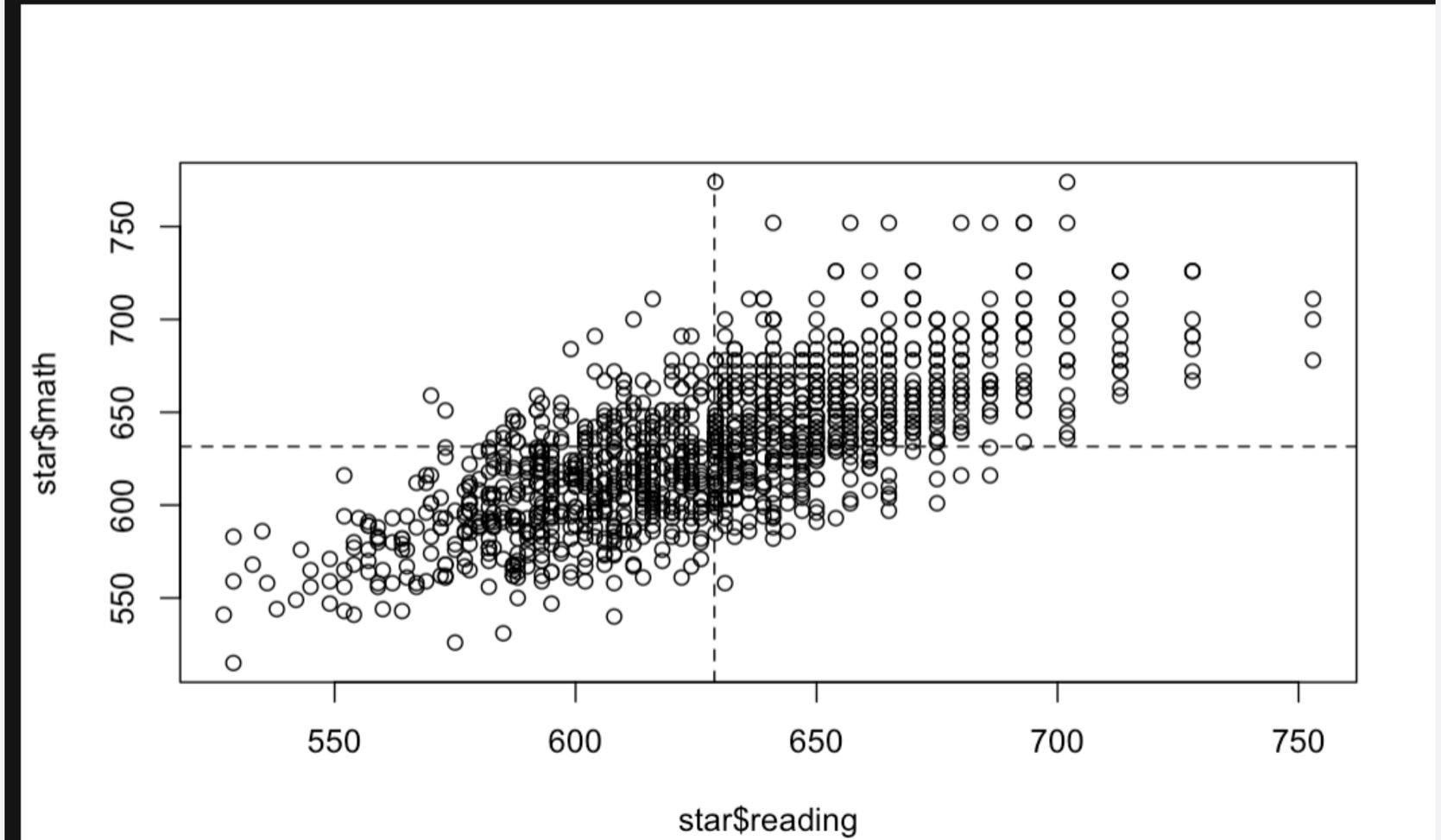
Visualizing Correlation

```r
plot(y=star$reading, x=star$math)
```



Visualizing Correlation

```r
plot(star$reading, star$math)
abline(v=mean(star$reading), lty='dashed')
abline(h=mean(star$math), lty='dashed')
```



Z-Scores

$$Z_i^X = \frac{(X_i - \bar{X})}{sd(X)}$$

Z_{i}^{X} = \frac{(X_i - \bar{X})}{sd(X)}

Correlation Coefficient

$$cor(X, Y) = \frac{\sum_{i=1}^{n} Z_i^X \times Z_i^Y}{n}$$

cor(X,Y) = \frac{\sum_{i=1}^n Z_{i}^X \times Z_{i}^Y}{n}

Correlation Coefficient