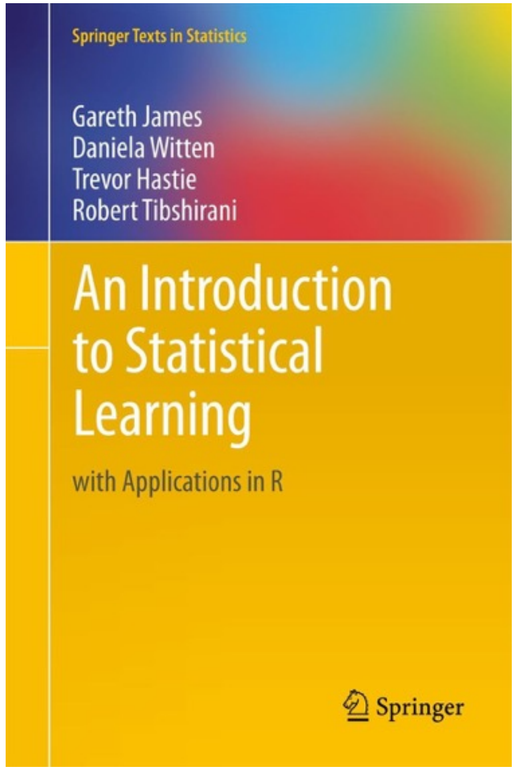
Introduction to Data Science
CS 5963 / Math 3900

Lecture 5: Linear Regression

Recommended Reading: ISLR, Ch.3
Available digitally here: link



THE UNIVERSITY OF UTAH

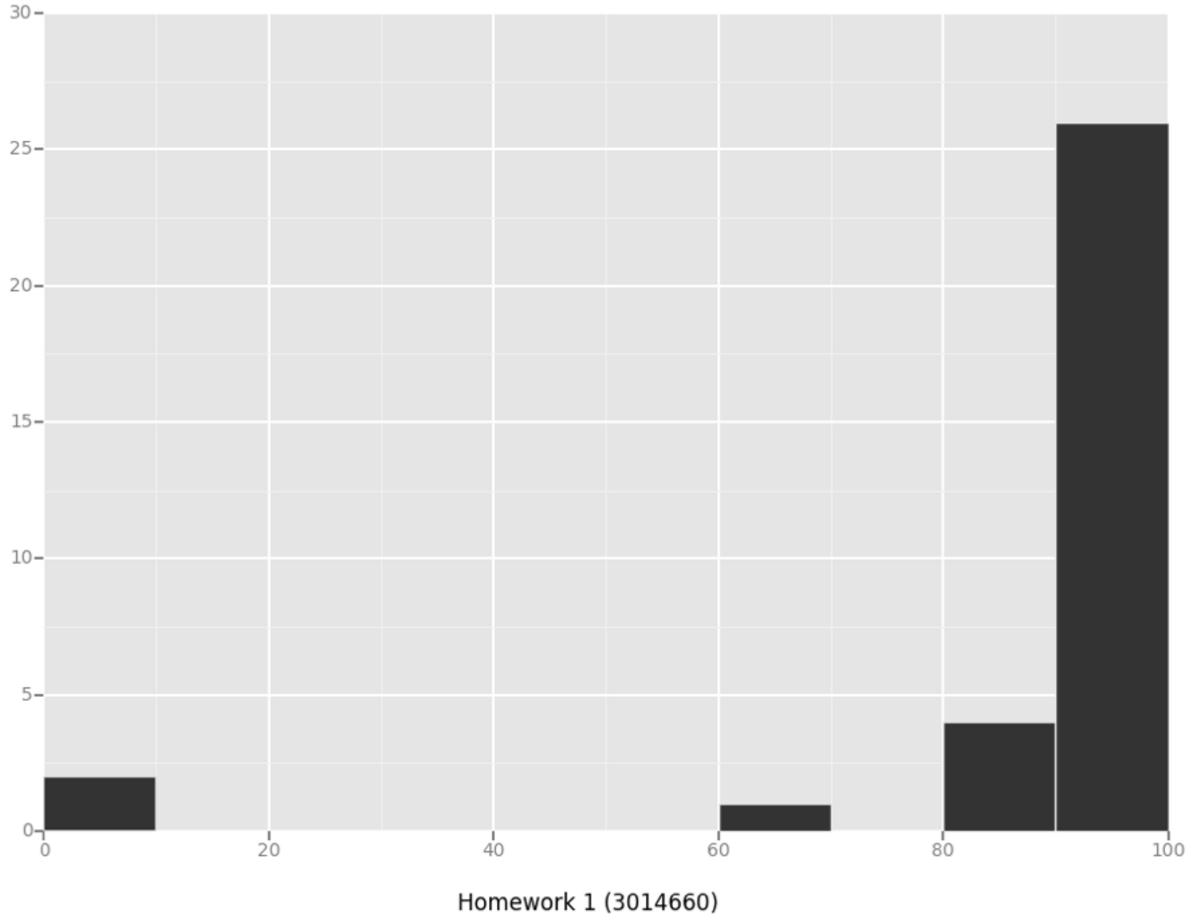
hw1 grades

In [30]:
```
filtered_hw1_grades.describe()
```

Out[30]:

	Homework 1 (3014660)
count	33.000000
mean	88.424242
std	23.971377
min	0.000000
25%	91.000000
50%	95.000000
75%	99.000000
max	100.000000

In [32]:
```
ggplot(filtered_hw1_grades, aes(x=hw_label)) + geom_histogram()
```



Linear Regression

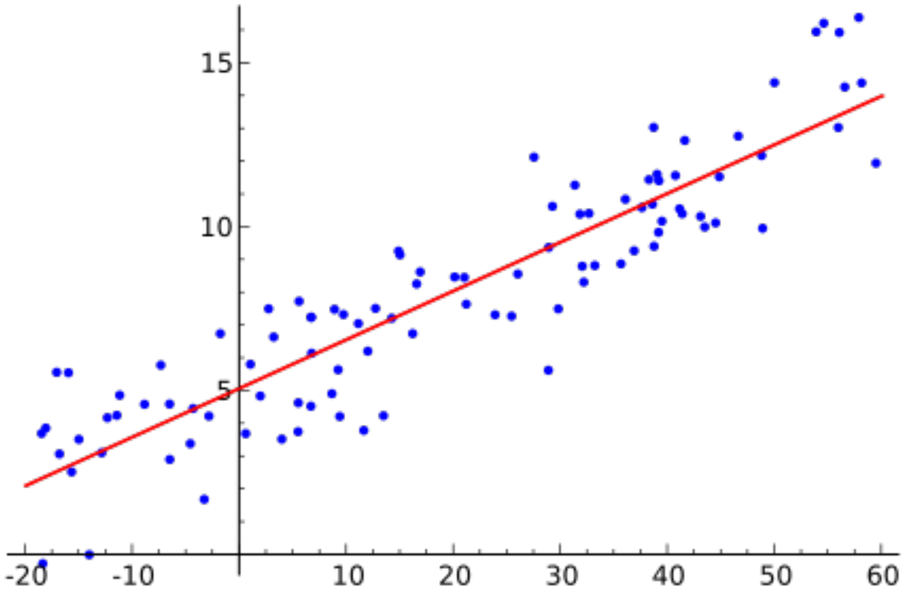
Linear regression models the relationship between a (real-valued) dependent value Y and an independent (explanatory) variable X.

Examples:

explanatory variable	dependent variable
square footage	house price
advertising dollars spent	profit
stress	lifespan
?	?

Simple Linear Regression

$$y \sim \beta_0 + \beta_1 x$$



We have samples (x_1, y_1), (x_2, y_2),. . . ,(x_n, y_n).

Goal: Find the best values of β_0 and β_1, denoted $\hat{\beta}_0$ and $\hat{\beta}_1$, so that the prediction $y = \hat{\beta}_0 + \hat{\beta}_1 x$ "best fits" the data.

Theorem. The best parameters in the least squares sense are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$