The R Language and Environment

What is R? (www.r-project.org)

> *R is a language and environment for statistical computing and graphics*

– `http://www.r-project.org/about.html`

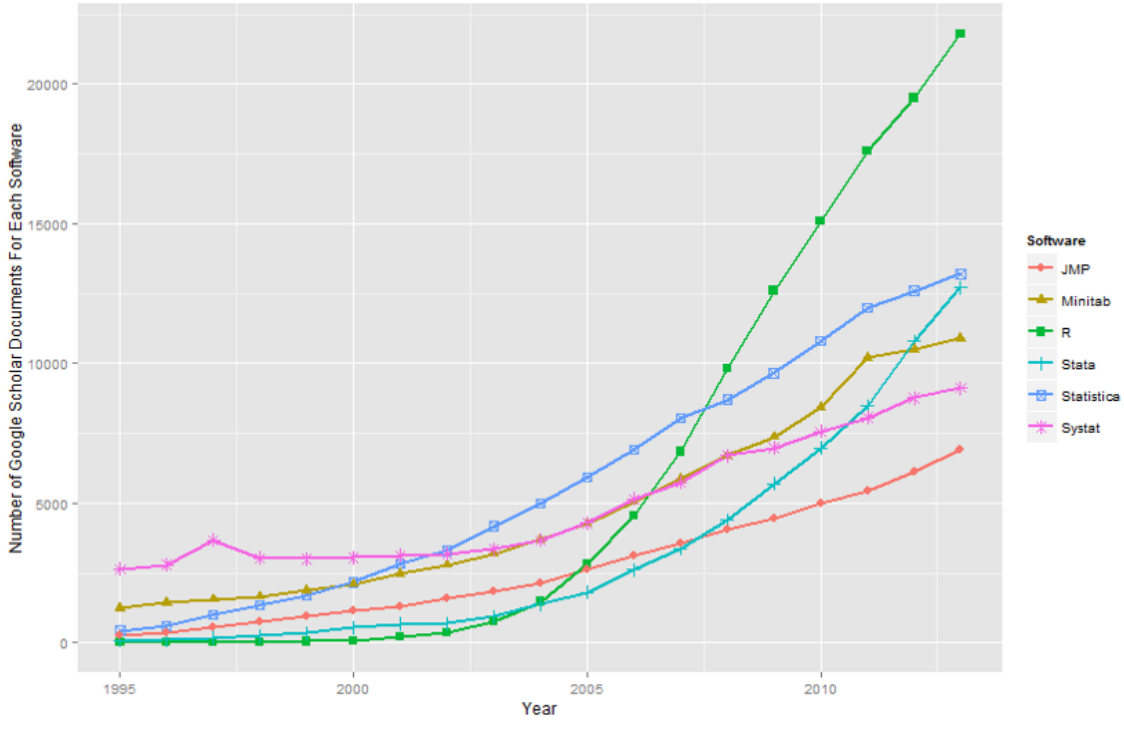
Why use R?

R offers many advantages:

- **Free**: As an open-source project, you can use R free of charge
- **A language**: In R, you do data analysis by writing functions and scripts, not by pointing and clicking.
- **Flexible statistical analysis toolkit**: All of the data analysis tools that you can think of are built into the R language
- **Powerful, cutting-edge analytics**: Leading academics and researches use R to develop the latest methods in statistics, machine learning, and predictive modeling
- **A robust, growing community**: thousands of contributors and over two millions users. If you have a question about R, someone's answered it (or they can and will)

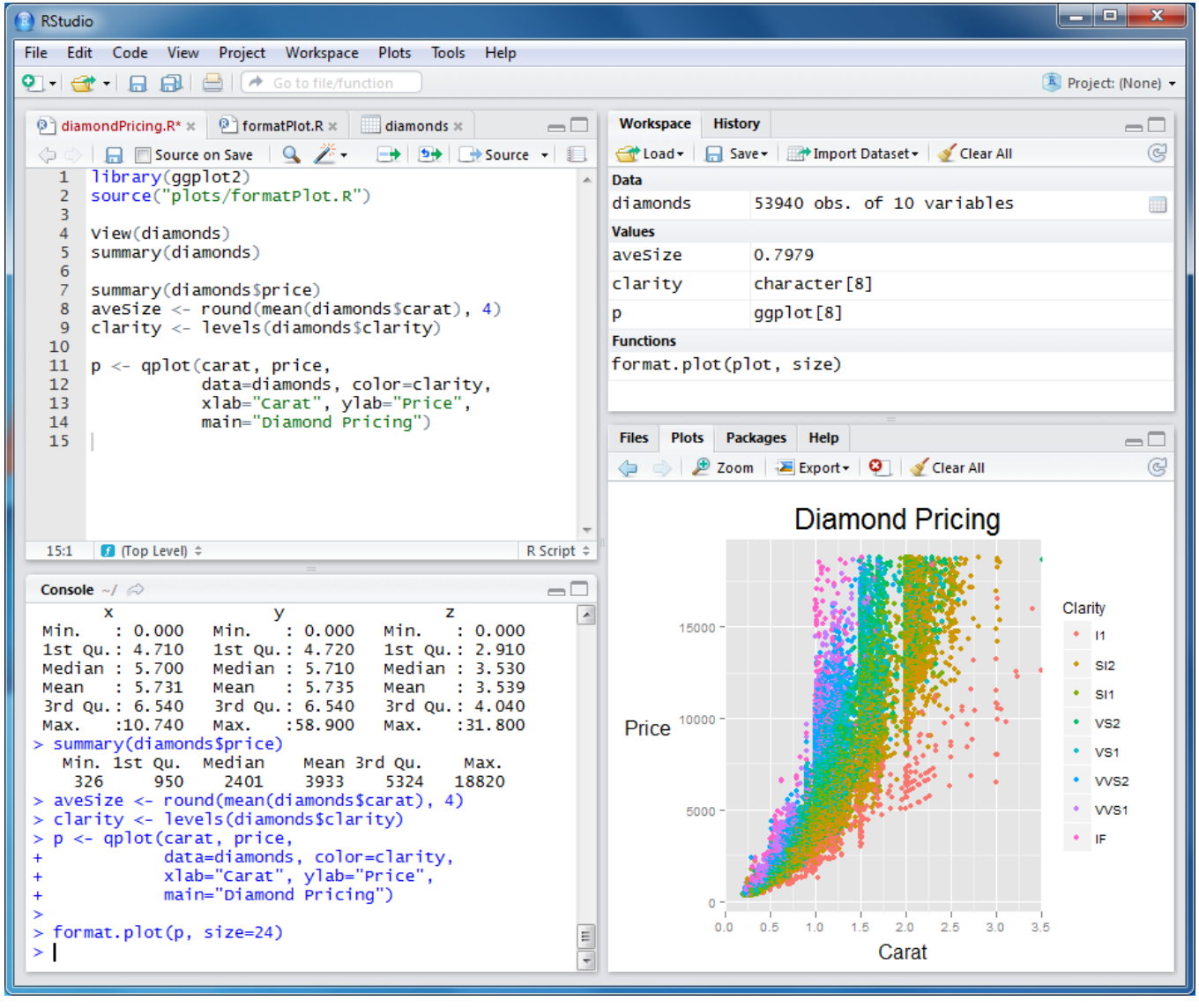
Why use R?

R is growing in the research and business community



Note: Excludes the incumbents SAS and SPSS, which the most used but in strong decline

RStudio (www.rstudio.com)



RStudio Hands On

Generating Random Variables in R

Random Variable: Number of heads appearing in 10 flips of a coin

- simulate 10 coin flips: coin.flip

```
s <- rbinom(n=10, size=1, prob=0.5) coin.flip s
```

```
[1] 1 0 1 1 0 0 0 1 0 1
```

- count the number of heads (number of "1"s) x <-

```
sum( coin.flip s )
```

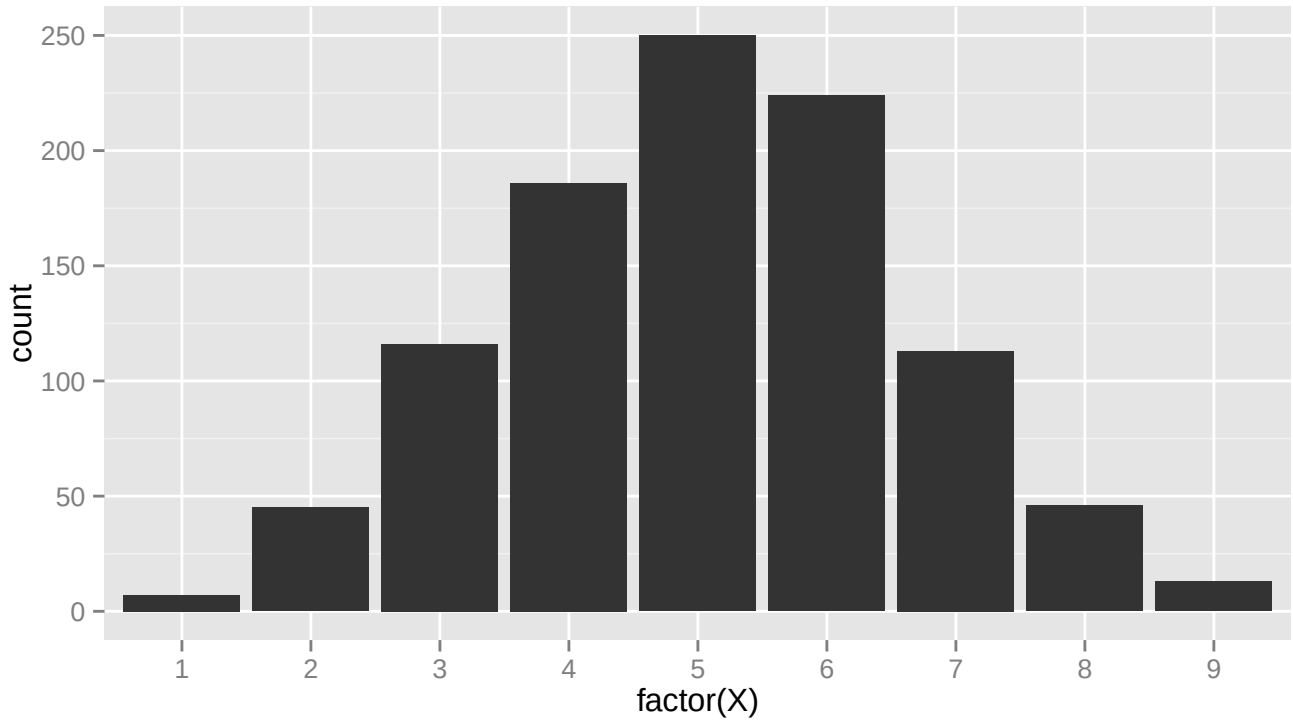
```
X
```

```
[1] 5
```

Generating Random Variables in R

Simulate and plot 1000 coin flips

```
library(ggplot2)
X <- rbinom(n=1000, size=10, prob=0.5)
qplot(factor(X), geom="histogram")
```



Example: Course Grades

Example: Course Grades

Display first rows

```
dt.grades[1:10]
```

	student_id	hw1	hw2	hw3	hw4	hw5	hw6	hw7	hw8	hw_avg	particip	midterm	final_exam	final_grade	pass
1:	1	20	16.3	17.5	19.5	16.8	16.0	14.7	15.3	17.0	1	17.0	18.6	18	TRUE
2:	2	20	20.0	19.3	19.0	14.0	15.6	17.2	16.9	17.8	NA	15.1	14.3	16	TRUE
3:	3	20	15.0	16.7	19.0	15.5	15.8	17.3	17.2	17.1	1	7.7	15.6	15	TRUE
4:	4	20	18.8	18.0	20.0	14.3	16.0	16.5	18.9	17.8	1	14.9	17.8	17	TRUE
5:	5	20	16.3	17.1	16.5	16.8	16.0	14.7	13.6	16.4	1	7.5	13.0	14	TRUE
6:	6	20	17.5	17.0	14.9	15.3	16.2	16.1	16.4	16.7	NA	12.5	8.1	14	TRUE
7:	7	20	20.0	19.3	19.0	14.0	15.6	17.2	16.4	17.7	NA	17.2	18.3	18	TRUE
8:	8	20	16.3	20.0	19.5	13.3	15.8	18.9	19.1	17.8	1	13.9	19.7	18	TRUE
9:	9	20	18.8	16.5	20.0	16.5	15.3	16.7	15.3	17.4	1	14.7	18.6	17	TRUE
10:	10	20	18.8	16.3	17.0	15.5	15.8	17.3	18.1	17.3	1	11.0	16.1	16	TRUE

Example: Course Grades

Summarize the data

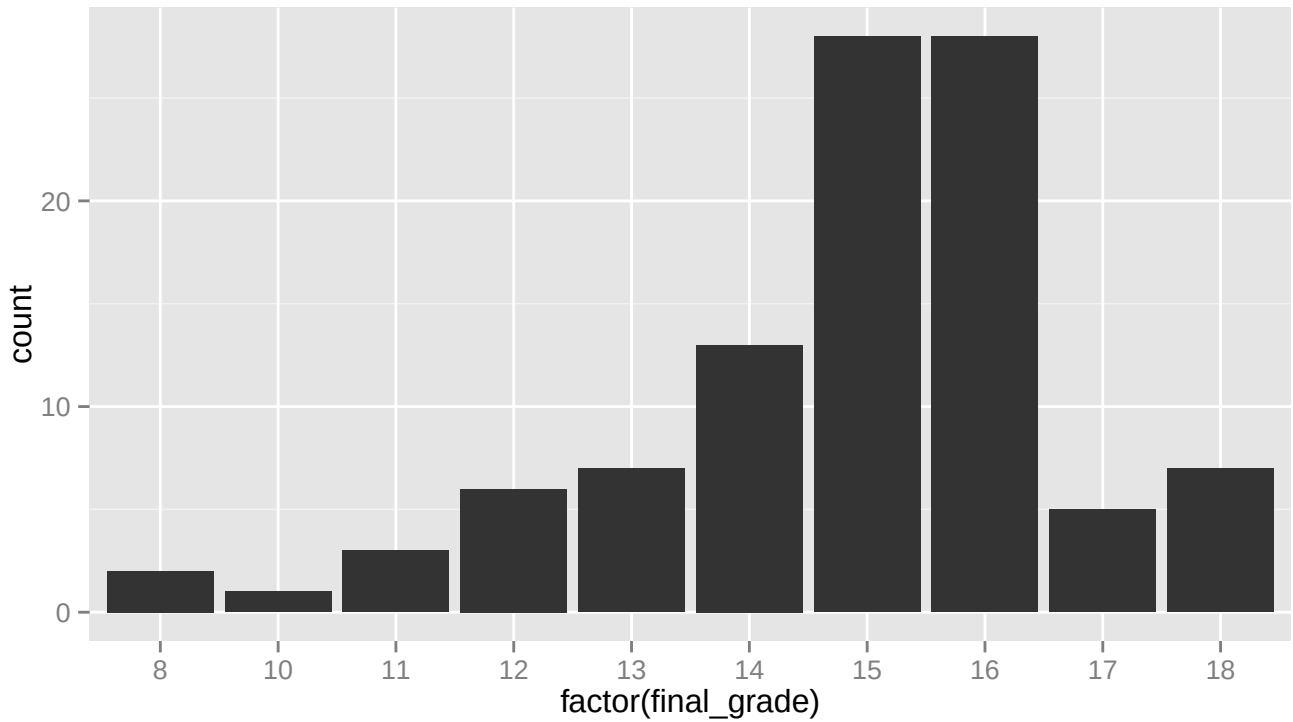
```
summary(dt.grades)
```

```
  student_id          hw1          hw2             hw3             hw4             hw5             hw6       
Min.   :  1.00   Min.   :20   Min.   :12.50   Min.   : 0.00   Min.   :10.90   Min.   :10.3   Min.   :12.20  
1st Qu.: 25.75   1st Qu.:20   1st Qu.:15.00   1st Qu.:14.30   1st Qu.:16.50   1st Qu.:13.3   1st Qu.:14.40  
Median : 50.50   Median :20   Median :17.50   Median :16.30   Median :18.00   Median :14.3   Median :15.80  
Mean   : 50.50   Mean   :20   Mean   :17.17   Mean   :15.69   Mean   :17.42   Mean   :14.5   Mean   :15.12  
3rd Qu.: 75.25   3rd Qu.:20   3rd Qu.:18.80   3rd Qu.:17.10   3rd Qu.:19.00   3rd Qu.:16.0   3rd Qu.:16.00  
Max.   :100.00   Max.   :20   Max.   :20.00   Max.   :20.00   Max.   :20.00   Max.   :20.0   Max.   :16.60  

      hw7             hw8             hw_avg          particip     midterm         final_exam      final_grade   
Min.   :10.20   Min.   : 0.00   Min.   :13.20   Min.   :1    Min.   : 2.7   Min.   : 2.70   Min.   : 8.00  
1st Qu.:14.70   1st Qu.:12.00   1st Qu.:15.40   1st Qu.:1    1st Qu.: 8.3   1st Qu.:11.00   1st Qu.:13.00  
Median :16.50   Median :13.90   Median :16.40   Median :1    Median :12.2   Median :13.65   Median :14.00  
Mean   :15.92   Mean   :13.25   Mean   :16.12   Mean   :1    Mean   :11.2   Mean   :13.09   Mean   :14.44  
3rd Qu.:17.20   3rd Qu.:16.30   3rd Qu.:16.70   3rd Qu.:1    3rd Qu.:13.8   3rd Qu.:15.62   3rd Qu.:16.00  
Max.   :18.90   Max.   :19.10   Max.   :17.80   Max.   :1    Max.   :17.2   Max.   :19.70   Max.   :18.00  
                                                NA's   :70                                                 

   pass        
Mode :logical  
FALSE:1        
TRUE :99       
NA's :0        
```

Example: Course Grades

Plot Final Grades

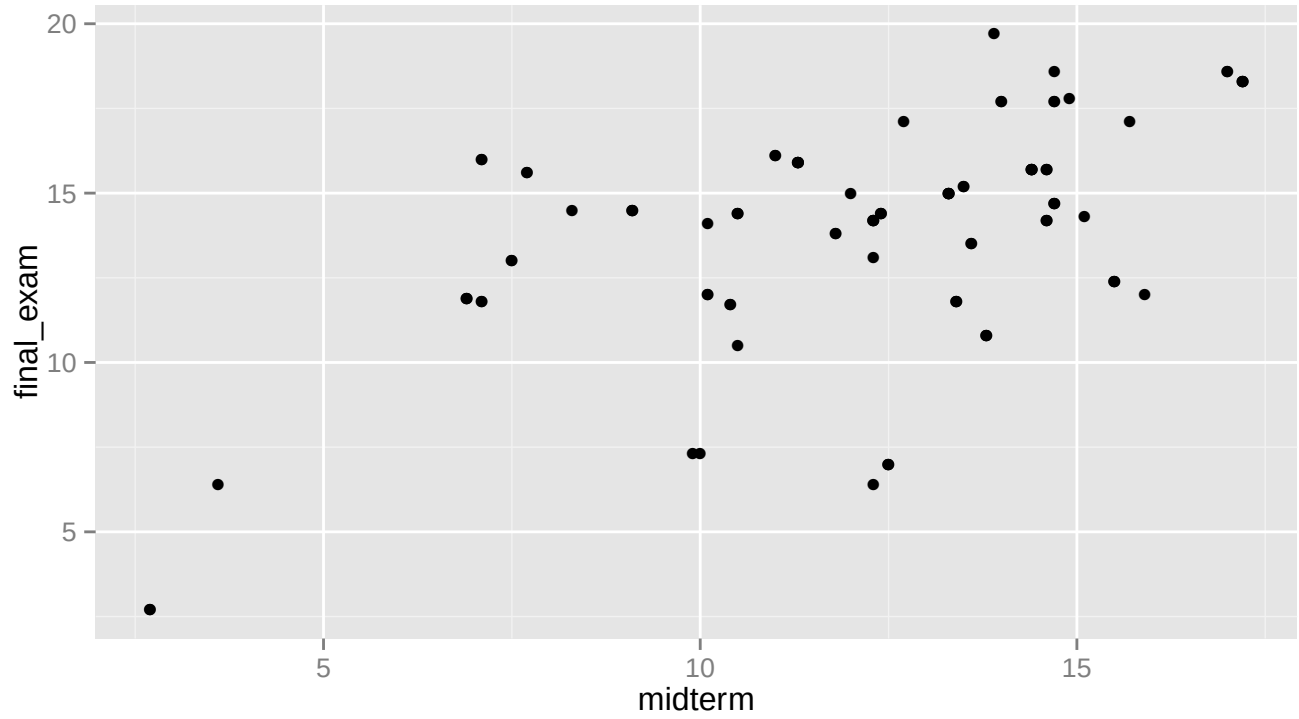
```
qplot(factor(final_grade), geom="histogram", data=dt.grades)
```



Example: Course Grades

Plot Midterm Grades against Final Exam Grades

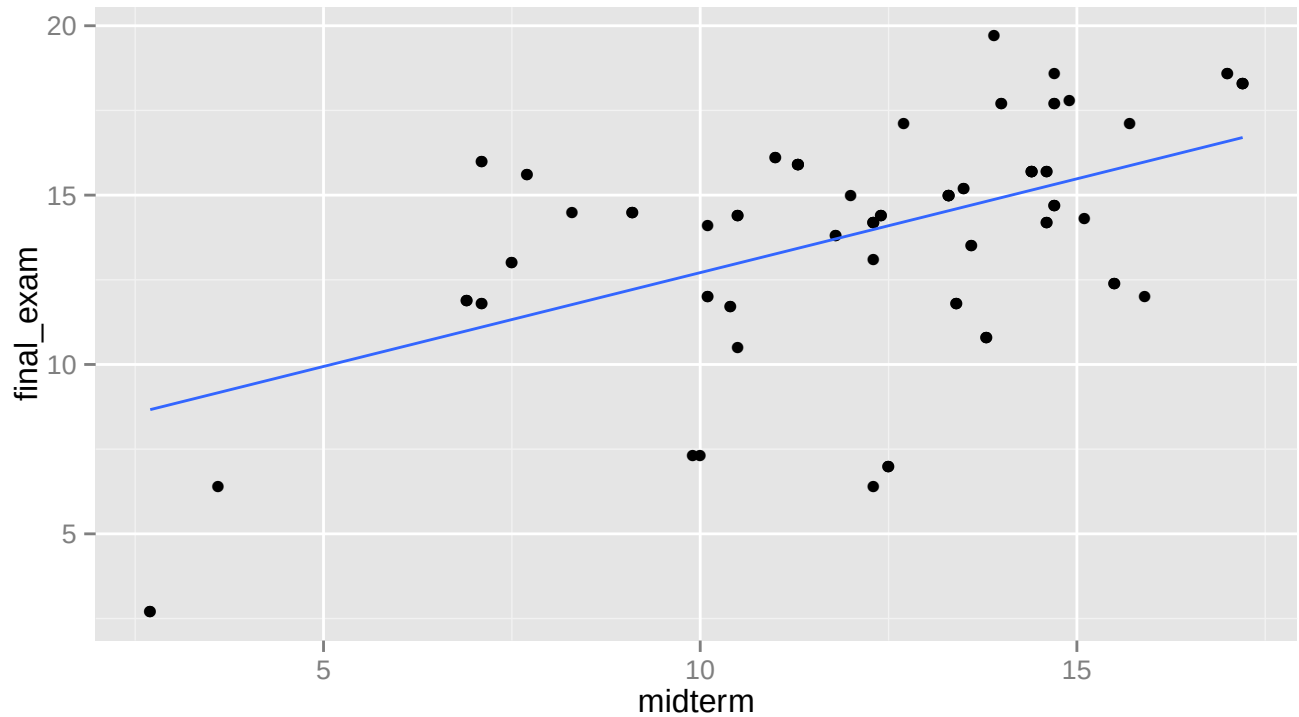
```
qplot(midterm, final_exam, geom="point", data=dt.grades)
```



Example: Course Grades

Plot Midterm Grades against Final Exam Grades

```
qplot(midterm, final_exam, geom="point", data=dt.grades) + geom_smooth(method=lm, se=FALSE)
```



Example: Course Grades

Sample Covariance between midterm and final exam grades

```
dt.grades[, cov(final_exam, midterm)]
```

```
[1] 5.284127
```

Sample Conditional Mean

```
dt.grades[midterm < 10, mean(final_exam)]
dt.grades[midterm >= 10, mean(final_exam)]
```

```
[1] 11.87368
[1] 14.38025
```

Example: Course Grades

Sample Conditional Mean

```
ggplot(dt.grades[, list(avg_final_exam = mean(final_exam)), by=list(midterm=round(midterm))]) +
geom_point(aes(midterm, avg_final_exam))
```