Day 1 - Intro to R

Sneak Peek!

Karsten Maurer, Eric Hare, and Susan VanderPlas

Iowa State University

Motivating Example

- Kick off the workshop by exploring a real data set using R!
- Goal: get the flavor of using R for data management and exploration
- Don't worry about understanding all the coding right away
- We will go back and explain how it all works in detail

Tips Data Set

- Tips data set recorded by a server in a restaurant over a span of about 10 weeks
- Server recorded several variables about groups they served:
  - Amount they were tipped
  - Cost of the total bill
  - Several characteristics about the groups being served
- Primary Question: How do these variable influence the amount being tipped?
- Follow along using RWorkshop1Tips.R

First look at data in R

Lets use R to look at the top few rows of the tips data set

```
# head() will pull the first few rows
head(tips)
```

```
##   total_bill  tip    sex smoker day   time size
## 1      16.99 1.01 Female     No Sun Dinner    2
## 2      10.34 1.66   Male     No Sun Dinner    3
## 3      21.01 3.50   Male     No Sun Dinner    3
## 4      23.68 3.31   Male     No Sun Dinner    2
## 5      24.59 3.61 Female     No Sun Dinner    4
## 6      25.29 4.71   Male     No Sun Dinner    4
```

Tips data attributes

How big is this data set and what types of variables are in each column?

```
  #look at the structure of the tips data set
str(tips)
```

```
## 'data.frame': 244 obs. of  7 variables:
##  $ total_bill: num  17 10.3 21 23.7 24.6 ...
##  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ..
##  $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2
##  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
##  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3
##  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

Tips Variables

Let's get a summary of the values for each variable in tips

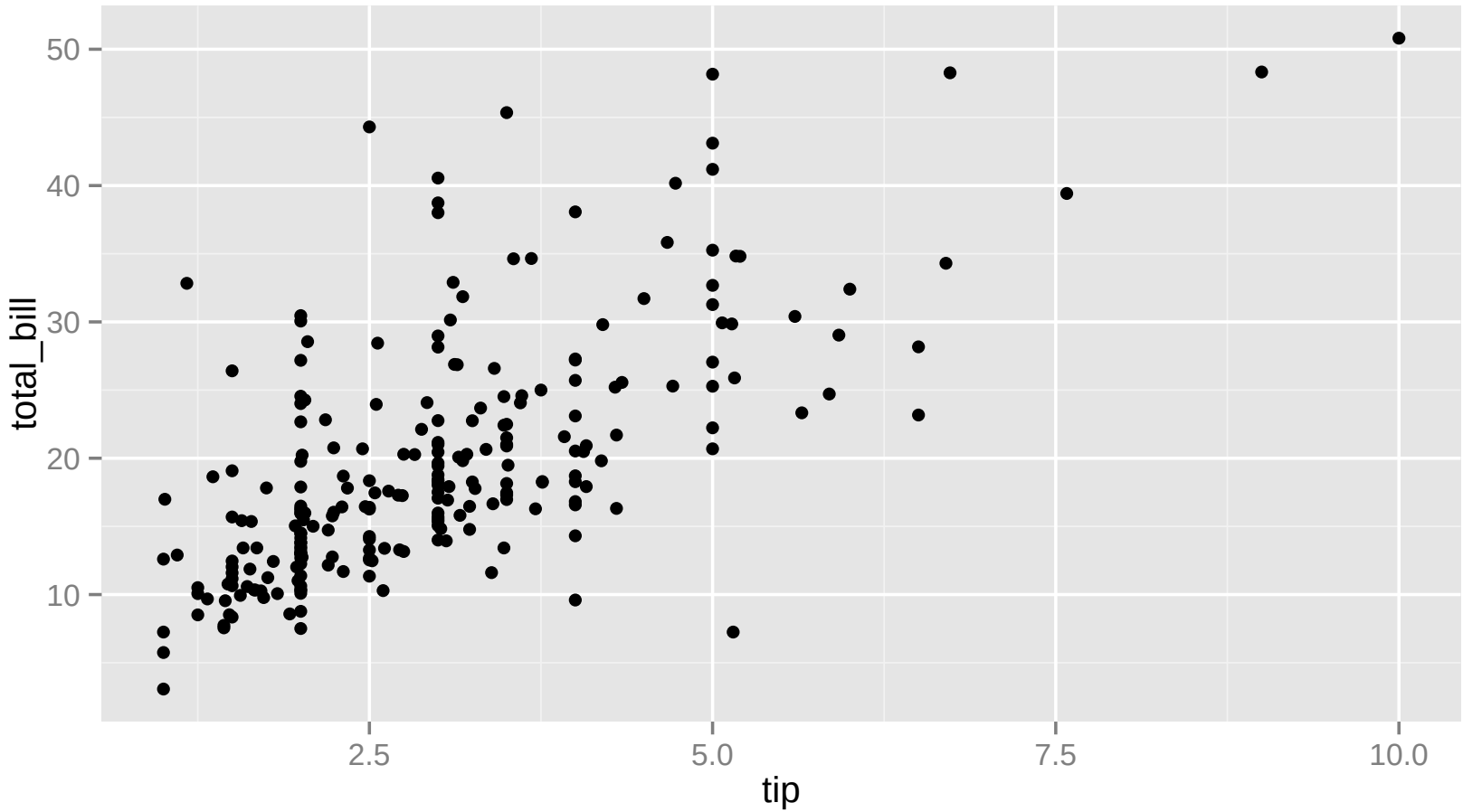
```
summary(tips)
```

```
##    total_bill          tip             sex      smoker      day
##  Min.   : 3.07   Min.   : 1.00   Female: 87   No :151   Fri :19
##  1st Qu.:13.35   1st Qu.: 2.00   Male  :157   Yes: 93   Sat :87
##  Median :17.80   Median : 2.90                           Sun :76
##  Mean   :19.79   Mean   : 3.00                           Thur:62
##  3rd Qu.:24.13   3rd Qu.: 3.56
##  Max.   :50.81   Max.   :10.00
##      time           size
##  Dinner:176   Min.   :1.00
##  Lunch : 68   1st Qu.:2.00
##               Median :2.00
##               Mean   :2.57
##               3rd Qu.:3.00
##               Max.   :6.00
```

Scatterplots

Lets look at the relationship between total bill and tip value

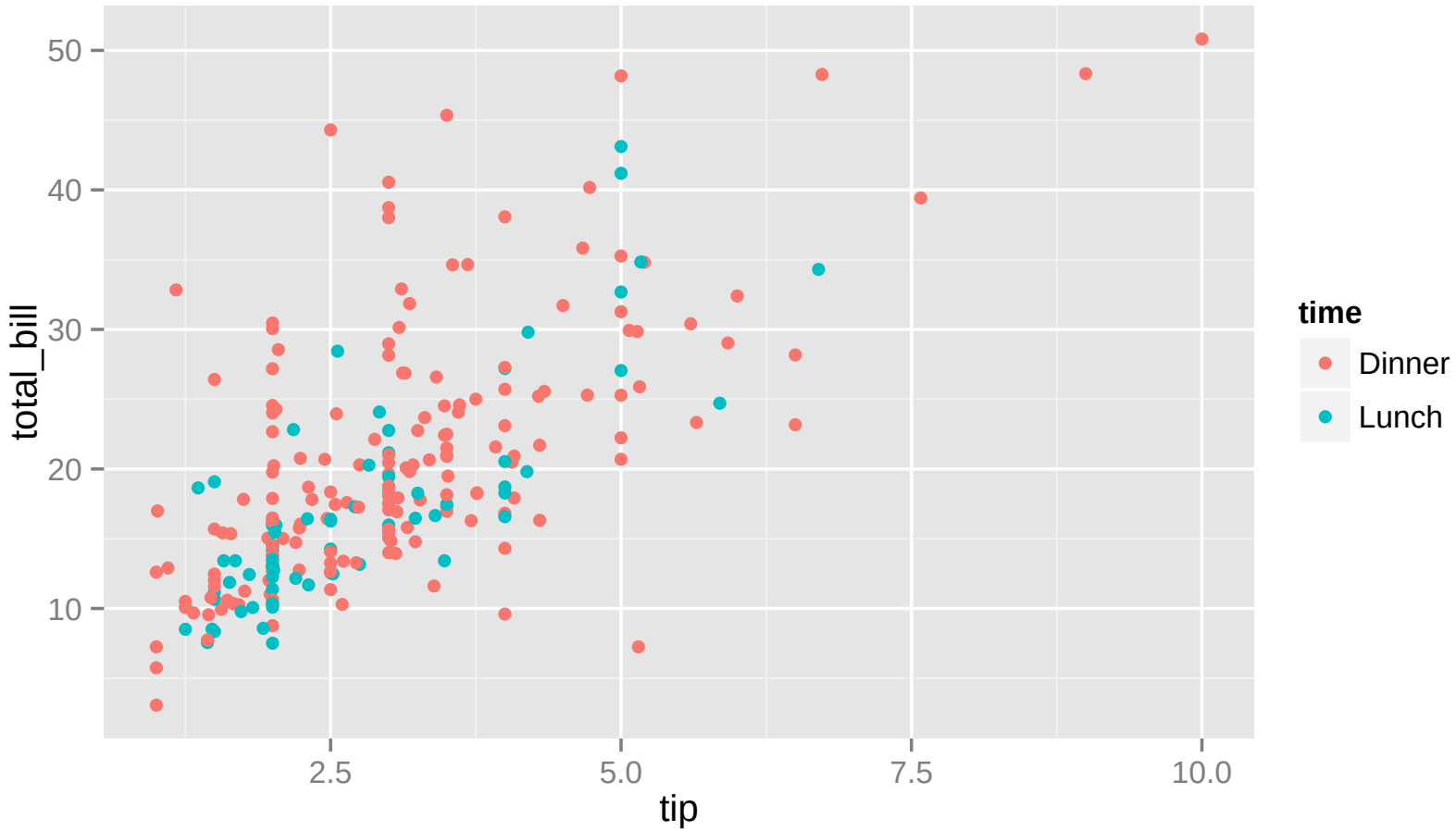
```
qplot(tip, total_bill, geom="point", data=tips)
```



Scatterplots

Color the points by lunch and dinner groups

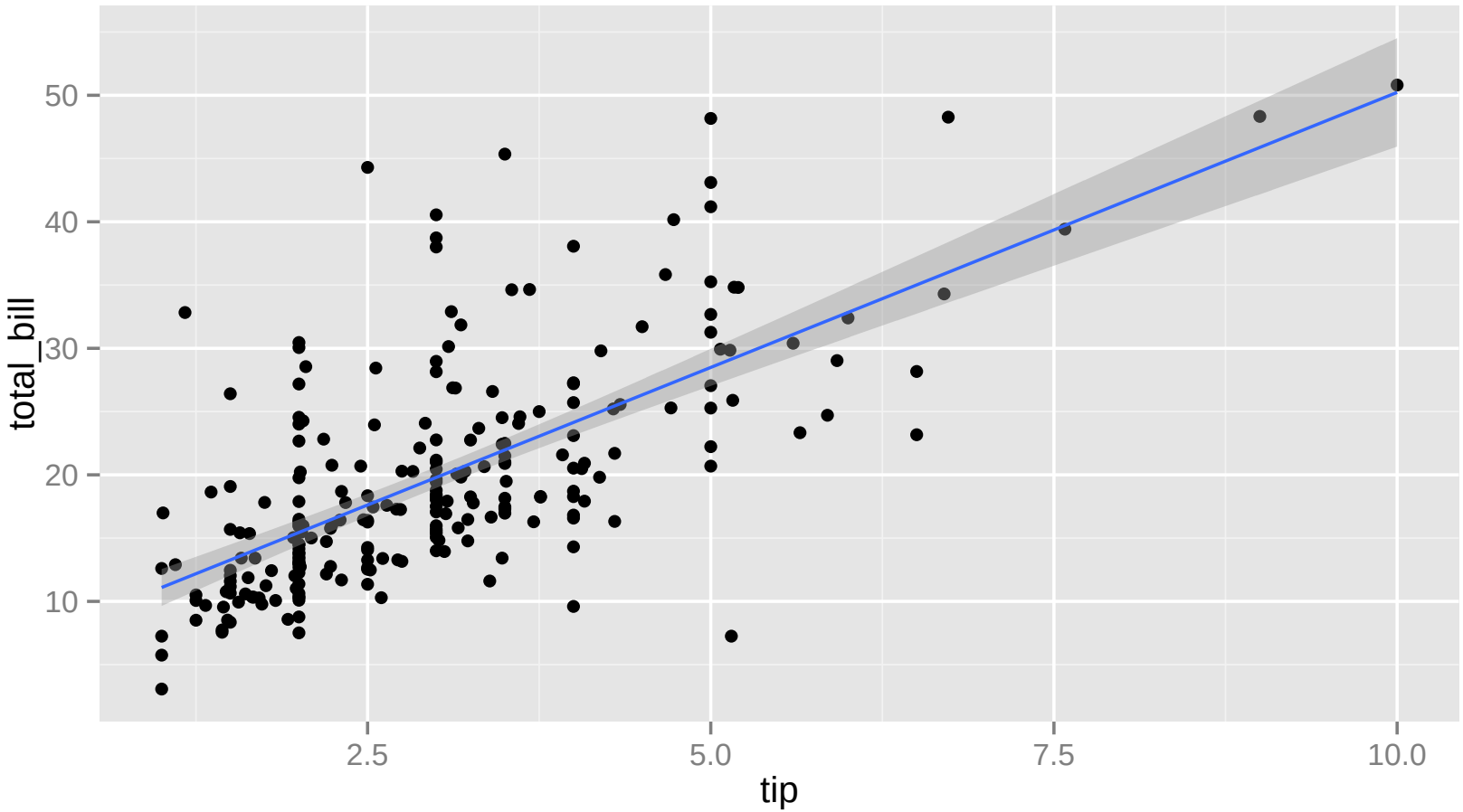
```
qplot(tip, total_bill, geom="point", data=tips, colour=time)
```



Scatterplots

Add linear regression line to the plot

```
qplot(tip, total_bill, geom="point", data=tips) + geom_smooth(method="lm
```



Rate of Tipping

Tipping generally done using a rule of thumb based on a percentage of the total bill. We will make a new variable in the data set for the tipping rate = tip / total bill

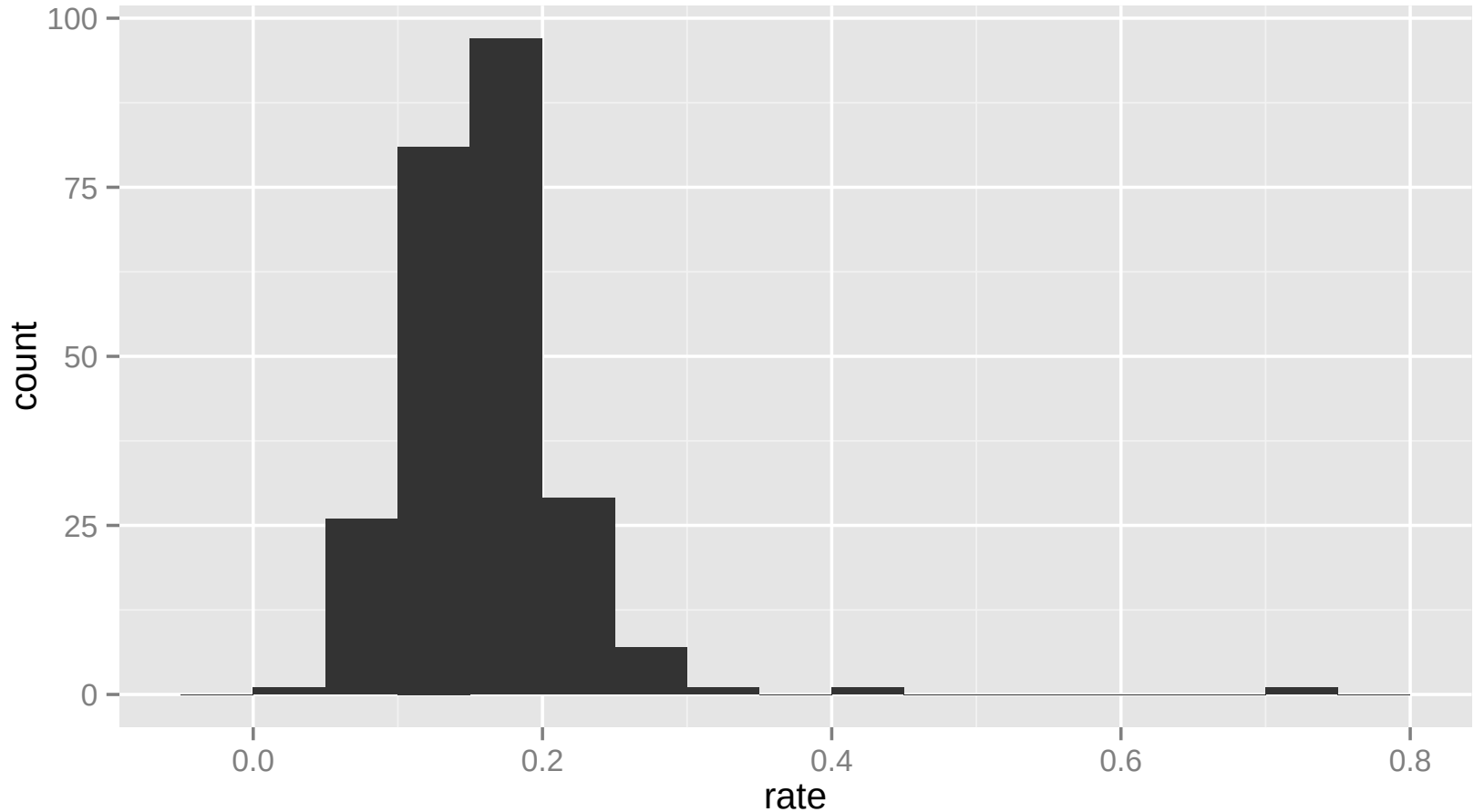
```
tips$rate <- tips$tip / tips$total_bill
# What are the properties of this new variable for tipping rate?
summary(tips$rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0356  0.1290  0.1550  0.1610  0.1910  0.7100
```

Tipping Rate Histogram

Lets look distribution of tipping rate values with a histogram

```
qplot(rate, data=tips, binwidth=.05)
```



Rate of Tipping

One person tipped over 70%, who are they?

```
tips[which.max(tips$rate),]
```

```
##     total_bill  tip  sex smoker day   time size   rate
## 173       7.25 5.15 Male    Yes Sun Dinner    2 0.7103
```

Rates by Gender

Look at the average tipping rate for men and women seperately

```
mean(tips$rate[tips$sex=="Male"])
```

```
## [1] 0.1577
```

```
mean(tips$rate[tips$sex=="Female"])
```

```
## [1] 0.1665
```

t-test

There is a difference but is it statistically significant?

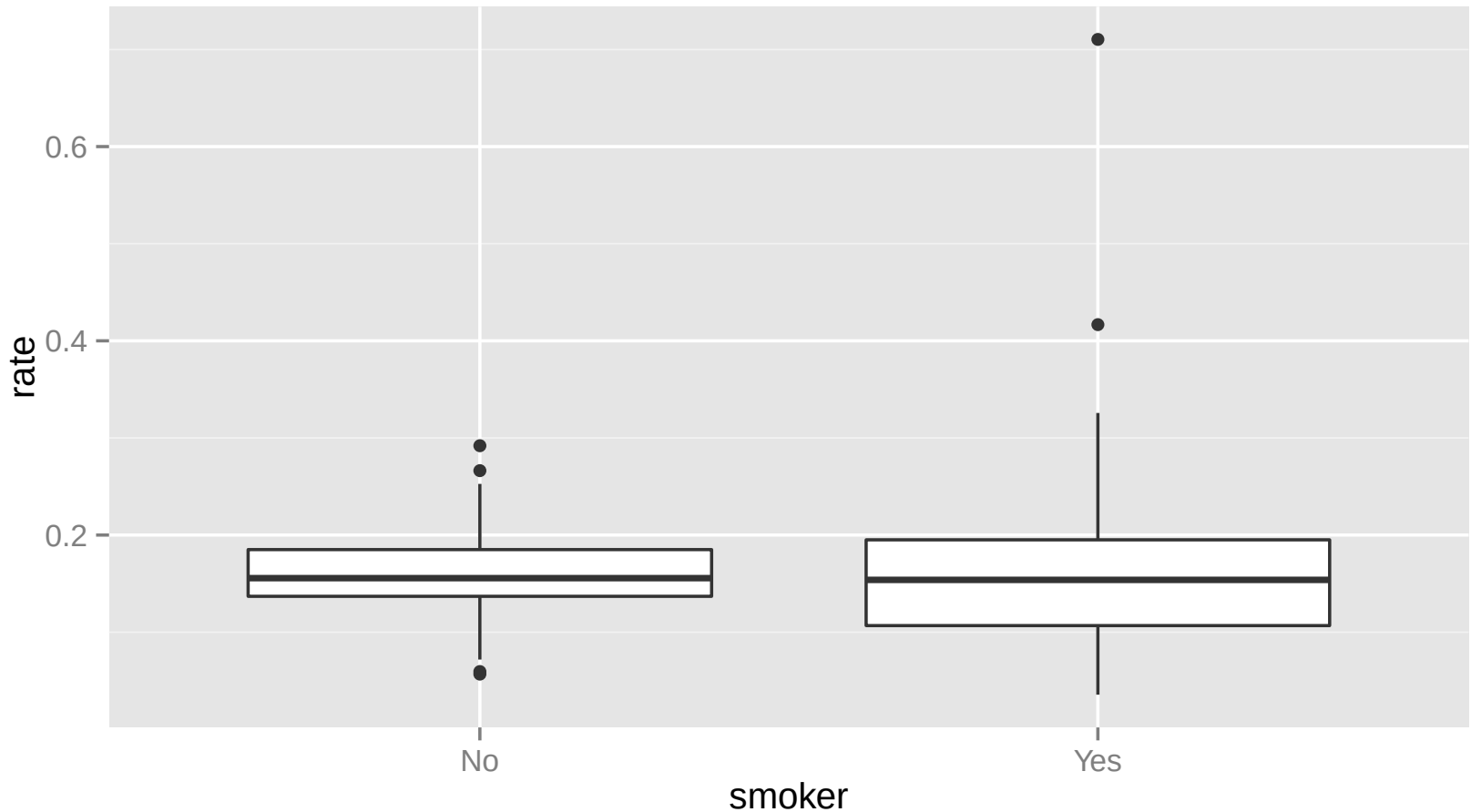
```
t.test(rate ~ sex , data=tips)
```

```
##
##  Welch Two Sample t-test
##
## data:  rate by sex
## t = 1.143, df = 206.8, p-value = 0.2542
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.006404  0.024084
## sample estimates:
## mean in group Female   mean in group Male
##               0.1665                0.1577
```

Boxplots

Perhaps we are interested if smokers tip at a different rate than non-smokers. We could compare the rate values of each group with a side by side boxplot!

```
qplot(smoker, rate, geom="boxplot",data=tips)
```



Your Turn

Try playing with chunks of code from RWorkshop1Tips.R to further investigate the tips data

- Get a summary of the total bill values
- Make side by side boxplots of tip rates for different days of the week
- Find the average tip value for smokers