Intro to Graphics with ggplot2

ggplot2 in a nutshell

- Package for statistical graphics
- Developed by Hadley Wickham (An ISU Alumni)
- Designed to adhere to good graphical practices
- Supports a wide variety plot types
- Constructs plots using the concept of layers

- `http://docs.ggplot2.org/current/` for reference material
- Hadley's book *ggplot2: Elegant Graphics for Data Analysis*

`ggplot()`

`ggplot()` function is the starting point for plots using the package

- This is the "blank canvas" function
- Can set default data scales for the plot here
- creates an object that can be saved
- plot layers can be added to modify plot complexity

ggplot() structure

ggplot() function has a basic syntax

```
ggplot(aes(variables=scales), dataset)
```

- The aes(..) statement: defines connection of variables to scales
- variables: and data column we want to plot
- scales: x, y, color, size, shape, groupings, orderings, etc.
- dataset: specified with a data= statement

Adding Layers to ggplot()

Now that aesthetic scales have been defined
we need to add geometric or statistical layers

```
ggplot(aes(variables=scales), dataset) +
     geom_point(aes(...),dataset) +
      stat_smooth(aes(...),dataset)
```

- `aes(..)` : Define in layers if different from default in textttggplot()
- dataset: Define in layers if different from default in textttggplot()
- This allows layers to be built from multiple data sources
- http://docs.ggplot2.org/current/ for reference material

Diamonds Data

We will explore the diamonds data set (preloaded along with ggplot2) using qplot for basic plotting.

The data set was scraped from a diamond exchange company data base by Hadley. It contains the prices and attributes of over 50,000 diamonds

Examining the Diamonds Data

What does the data look like?

Lets look at the top few rows of the diamond data frame to find out!

```
head(diamonds)
```

```
##   carat       cut color clarity depth table price    x    y    z
## 1  0.23     Ideal     E     SI2  61.5    55   326 3.95 3.98 2.43
## 2  0.21   Premium     E     SI1  59.8    61   326 3.89 3.84 2.31
## 3  0.23      Good     E     VS1  56.9    65   327 4.05 4.07 2.31
## 4  0.29   Premium     I     VS2  62.4    58   334 4.20 4.23 2.63
## 5  0.31      Good     J     SI2  63.3    58   335 4.34 4.35 2.75
## 6  0.24 Very Good     J    VVS2  62.8    57   336 3.94 3.96 2.48
```

qplot() demo

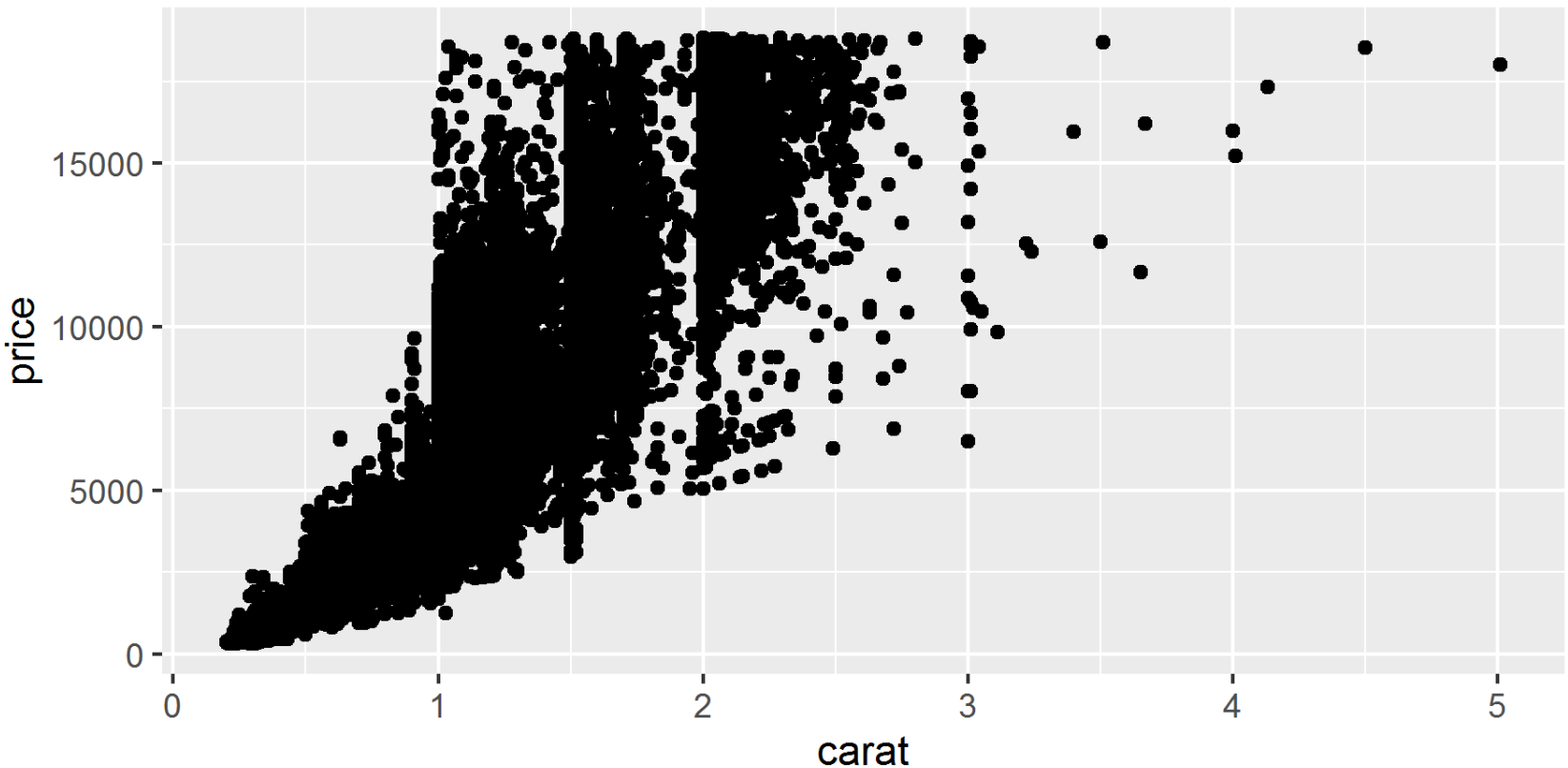
Demo of basic plot types and options using ggplot()!

Follow along with the demo by opening GraphicsIntro.R in your own R environment

Scatterplot

Basic scatter plot of diamond price vs carat weight

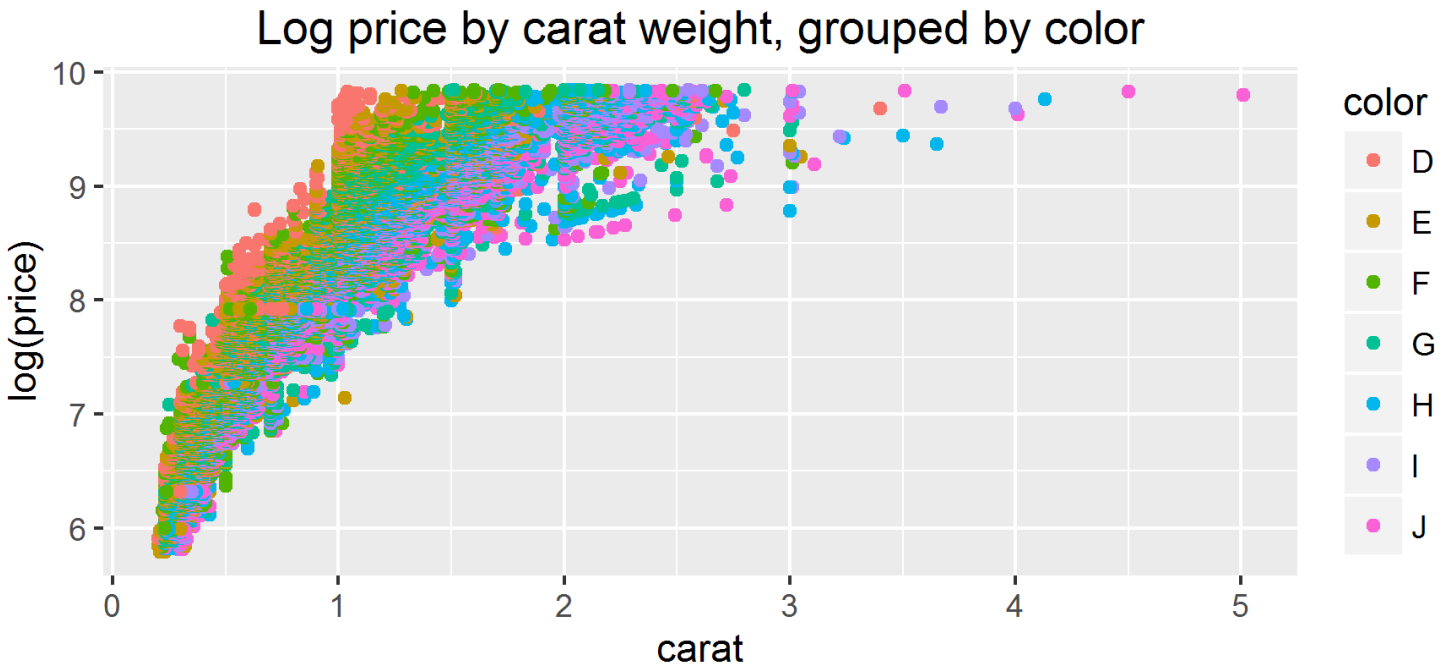
```
ggplot(aes(x=carat, y=price), data=diamonds) +
  geom_point()
```



Scatterplot

Scatter plot of diamond price vs carat weight showing versitility of options in qplot

```
ggplot(aes(x=carat, y=log(price), color=color), data=diamonds, alpha=I(
  geom_point() + ggtitle("Log price by carat weight, grouped by color")
```



Your Turn

All of the your turns for this section will use the tips data set (loaded in with reshape package)

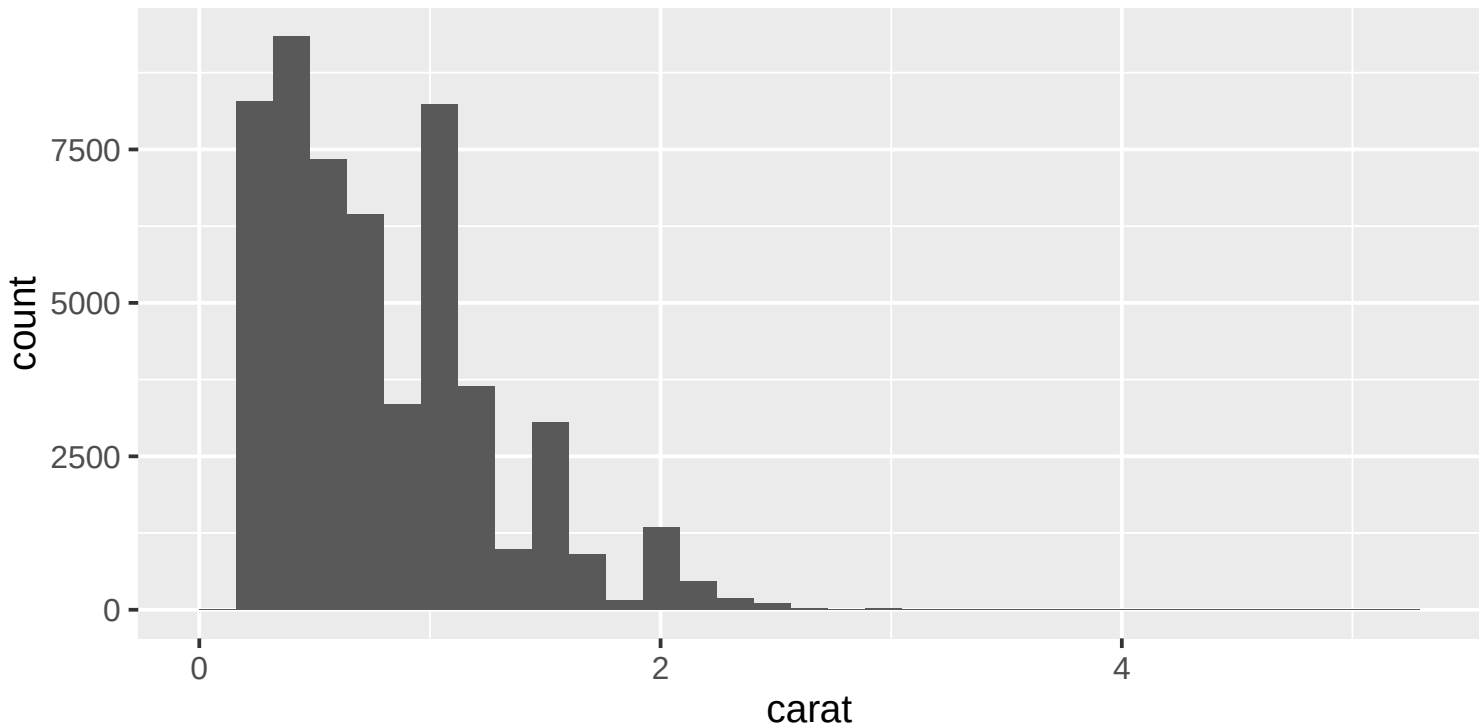
```
data(tips, package="reshape2")
```

- Use qplot to build a scatterplot of variables tips and total bill
- Use options within qplot to color points by smokers
- Clean up axis labels and add main plot title

Histograms

Basic histogram of carat weight

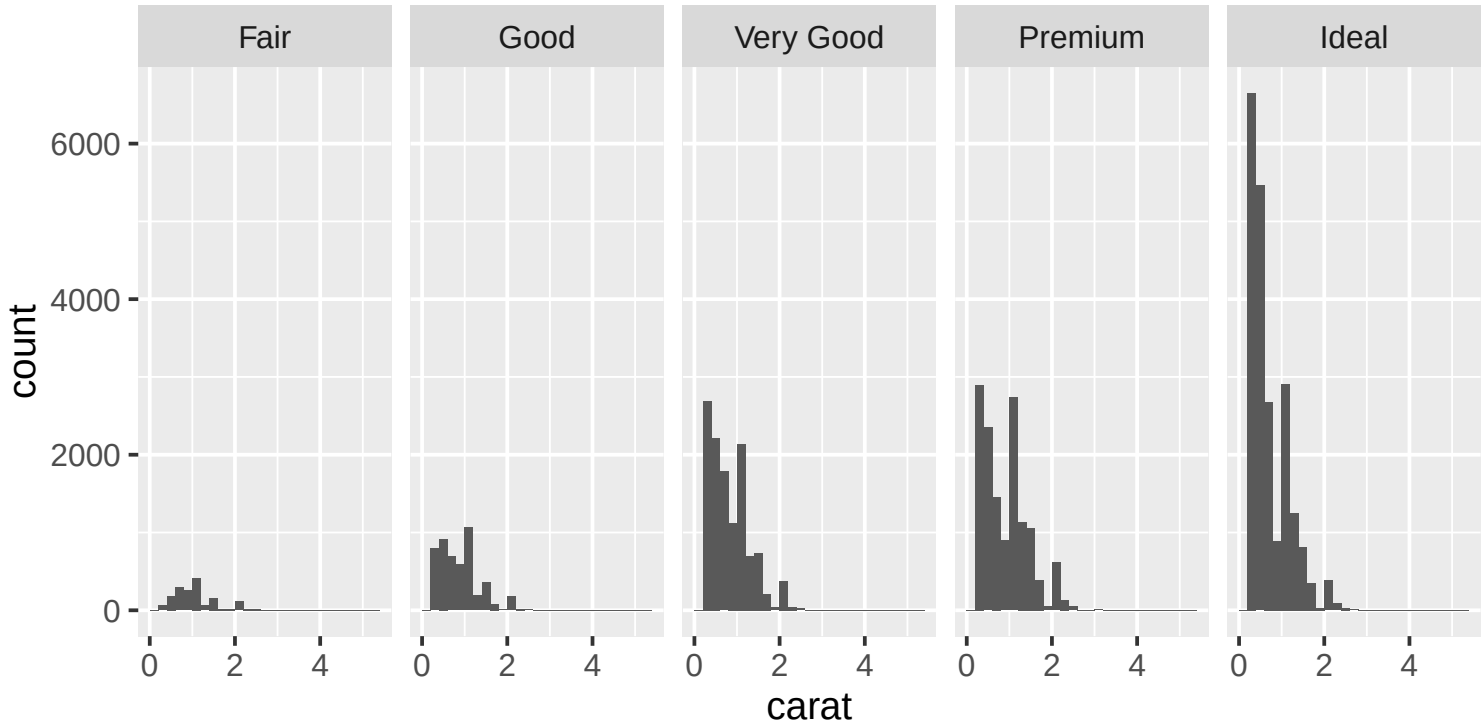
```
ggplot() +
  geom_histogram(aes(x=carat), data=diamonds)

## `stat_bin()` using `bins = 30`.  Pick better value with
`binwidth`.
```



Histograms

Carat weight histograms faceted by cut

```
ggplot(aes(x=carat), data=diamonds) +
  geom_histogram(binwidth=.2) +
  facet_grid(.~cut )
```



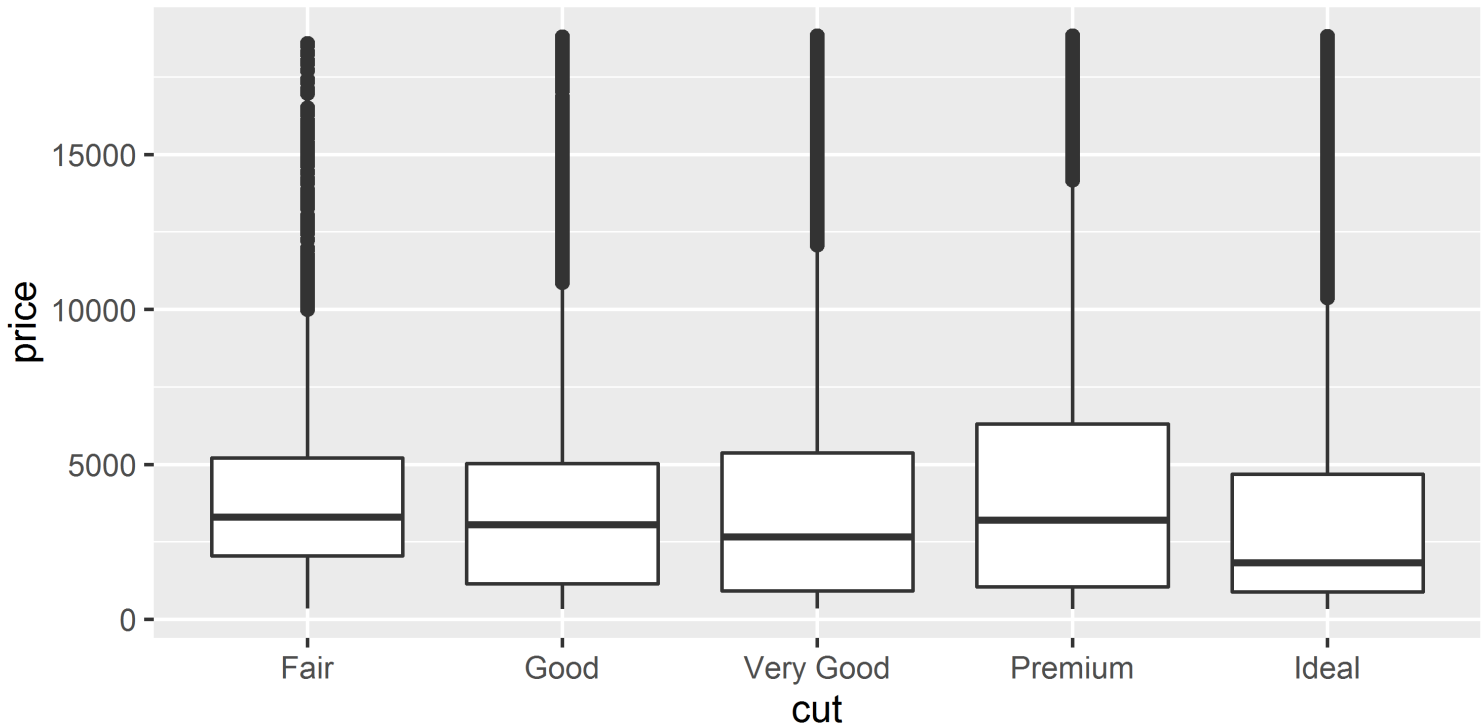
Your Turn

- Create a new variable in tips data frame rate = tip/total bill
- Use qplot to create a histogram of rate
- Change the bin width on that histogram to 0.05
- Facet this histogram by size of the group

Boxplots

Side by side boxplot of diamond prices within cut groupings

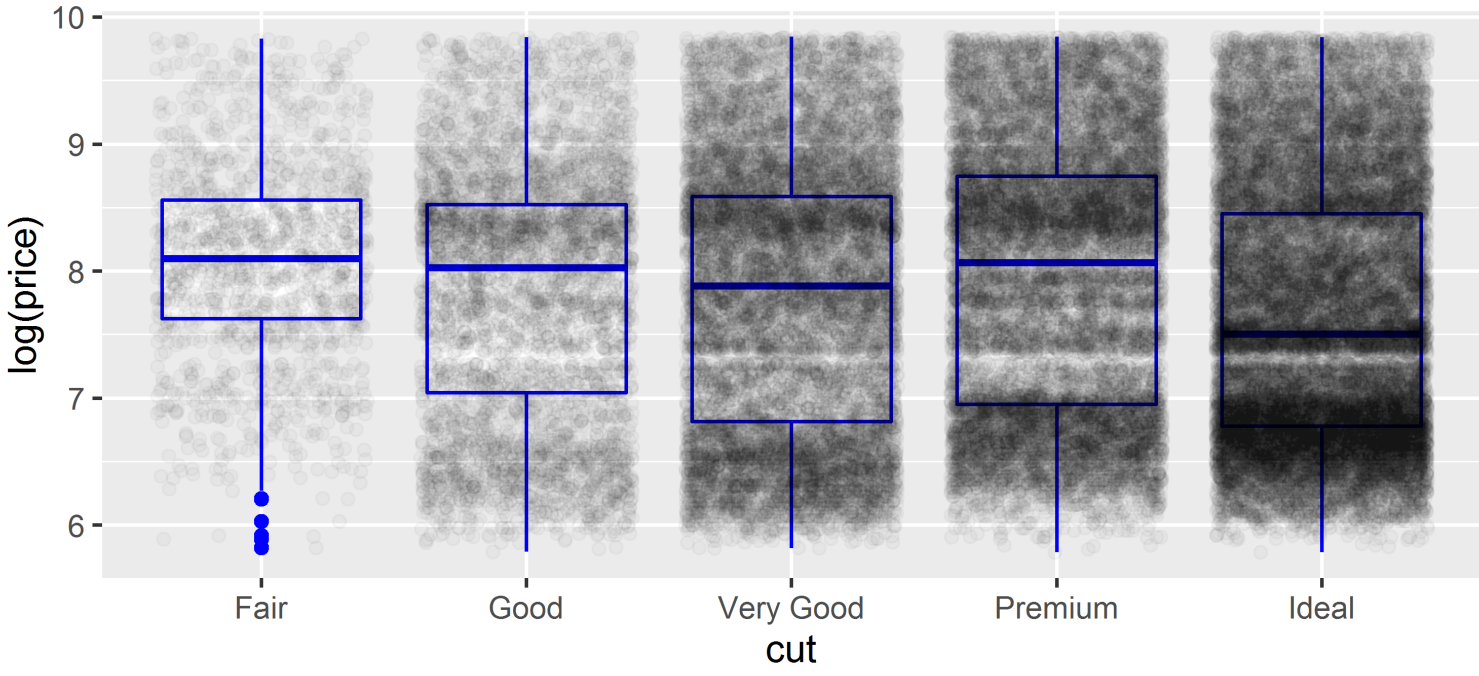
```
ggplot(aes(x=cut, y=price), data=diamonds) +
  geom_boxplot()
```



Boxplots

Side by side boxplot of log prices within cut groupings with jittered values overlay

```
ggplot(aes(x=cut, y=log(price)), data=diamonds,
        main="Boxplots of log Diamond Prices Grouped by Cut Quality") +
  geom_boxplot(color="blue") +
        geom_jitter(alpha=I(.025))
```



Your Turn

- Make side by side boxplots of tipping rate for males and females
- Overlay jittered points for observed values onto this boxplot

Bar plots

To investigate bar plots we will switch over to the Titanic data set

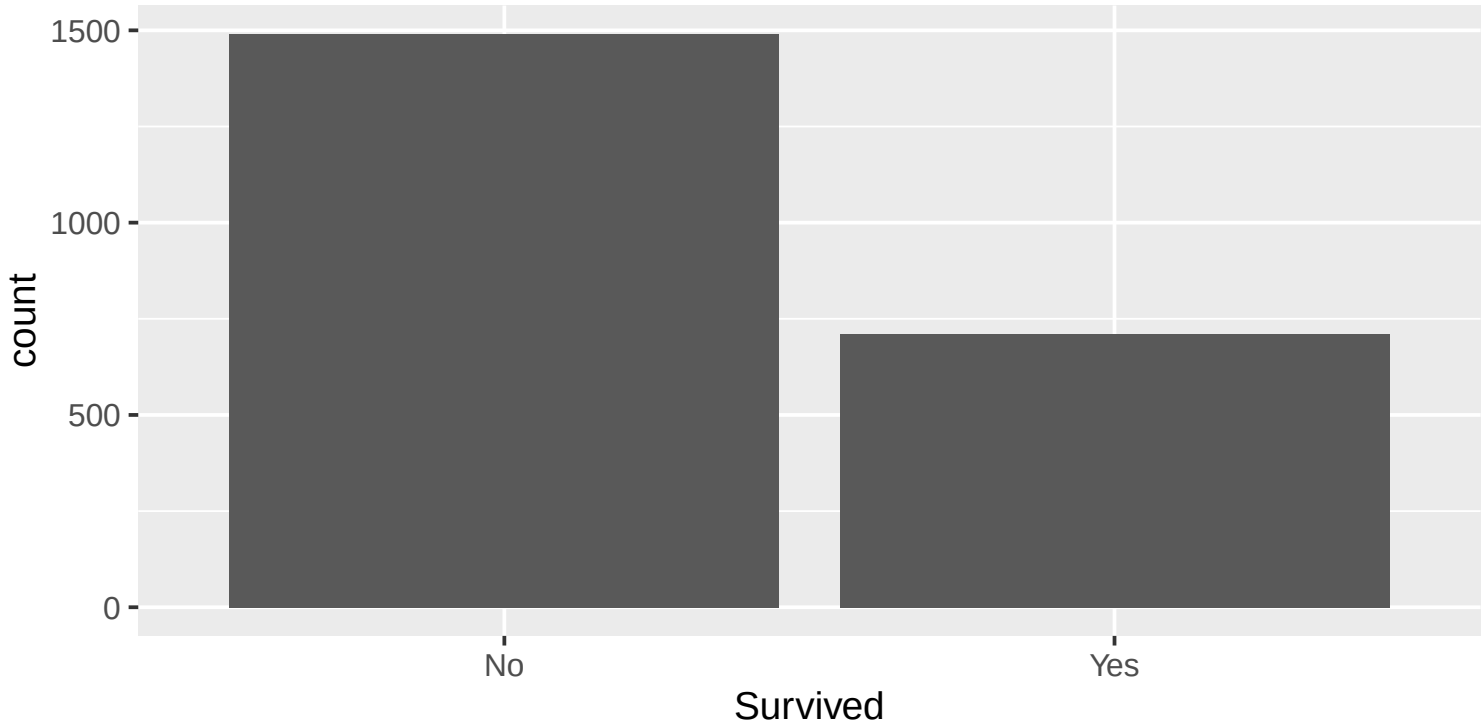
```
titanic <- as.data.frame(Titanic)
```

Data includes passenger characteristics and survival outcomes for those aboard the RMS Titanics ill fated maiden voyage

Bar Plots

Basic bar plot of survival outcomes

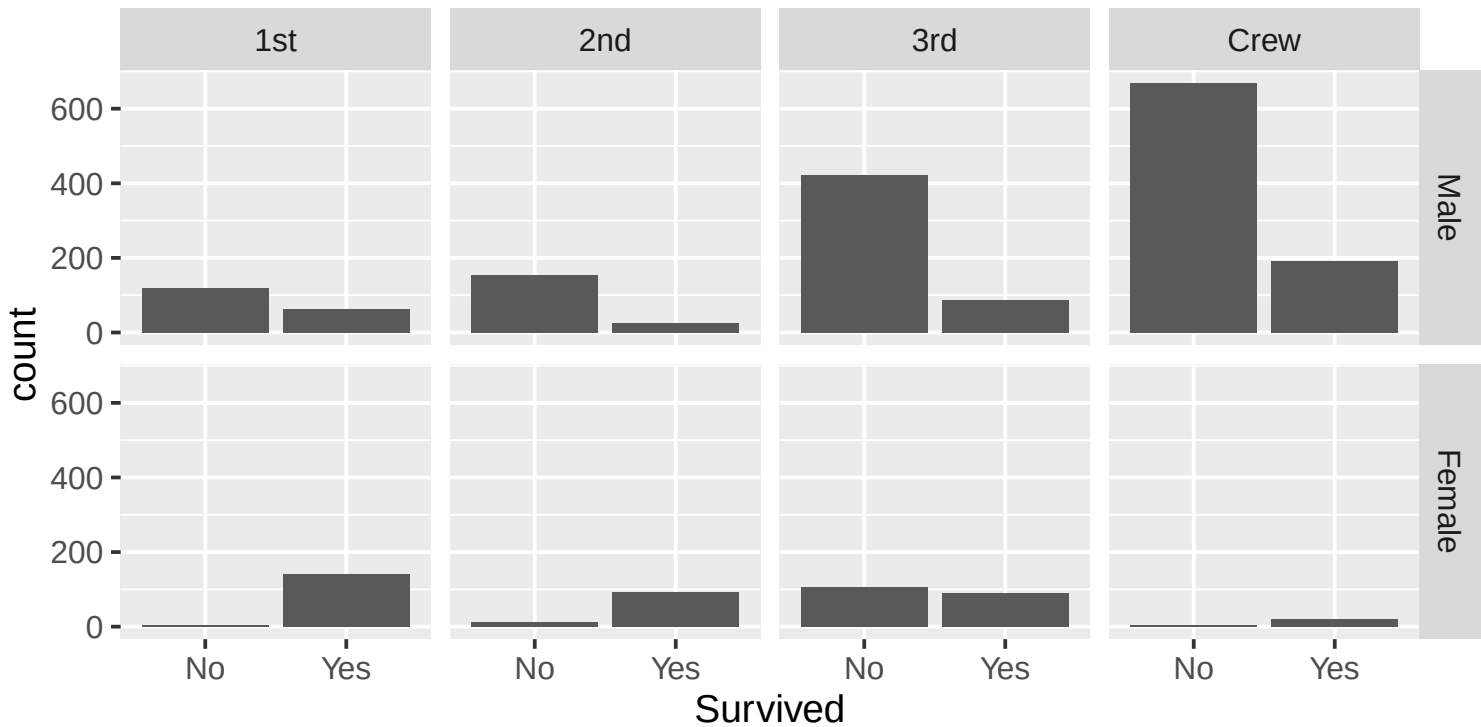
```
ggplot(aes(x=Survived, weight=Freq), data=titanic) +
  geom_bar()
```



Bar Plots

Bar plot faceted by gender and class

```
ggplot(aes(x=Survived, weight=Freq), data=titanic) +
  geom_bar()+
  facet_grid(Sex~Class)
```



Your Turn

- Use the tips data to make a barplot for counts of smoking and non smoking customers
- Facet using day of week and time of day to view how smoking status changes for different meal times