IS 607 Project 2

Derek G. Nokes

Saturday, March 14, 2015

Contents

```
# load the libraries
library('ggplot2')
library('ggthemes')
library('RColorBrewer')
library('Hmisc')
```

Load Data

We begin by loading the data.

```
# load the data
inputFile<-"C:/Users/dgn2/Documents/R/IS607/Project_2/project2_data.csv"
# load data
data <- read.csv(inputFile, header=TRUE)
```

The data is displayed in the following table:

I X	I Y	II X	II Y	III X	III Y	IV X	IV Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Table 1: X and Y by Quarter

Exploratory Data Analysis

Data Description

Once the data is loaded, we take a quick look at the attributes of the data:

```
str(data)
```

```
## 'data.frame':    44 obs. of  3 variables:
##  $ quarter: Factor w/ 4 levels "I","II","III",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ x      : int  10 8 13 9 11 14 6 4 12 7 ...
##  $ y      : num  8.04 6.95 7.58 8.81 8.33 ...
```

We can see that there are 44 rows and 3 columns in our data set (i.e., *quarter*, x, and y).

variable	description
quarter	is a factor with 4 levels (I, II, III, & IV)
x	is an int
y	is a decimal number

Table 2: Data Description

A quick summary of the data by the *quarter* factor does not provide much insight.

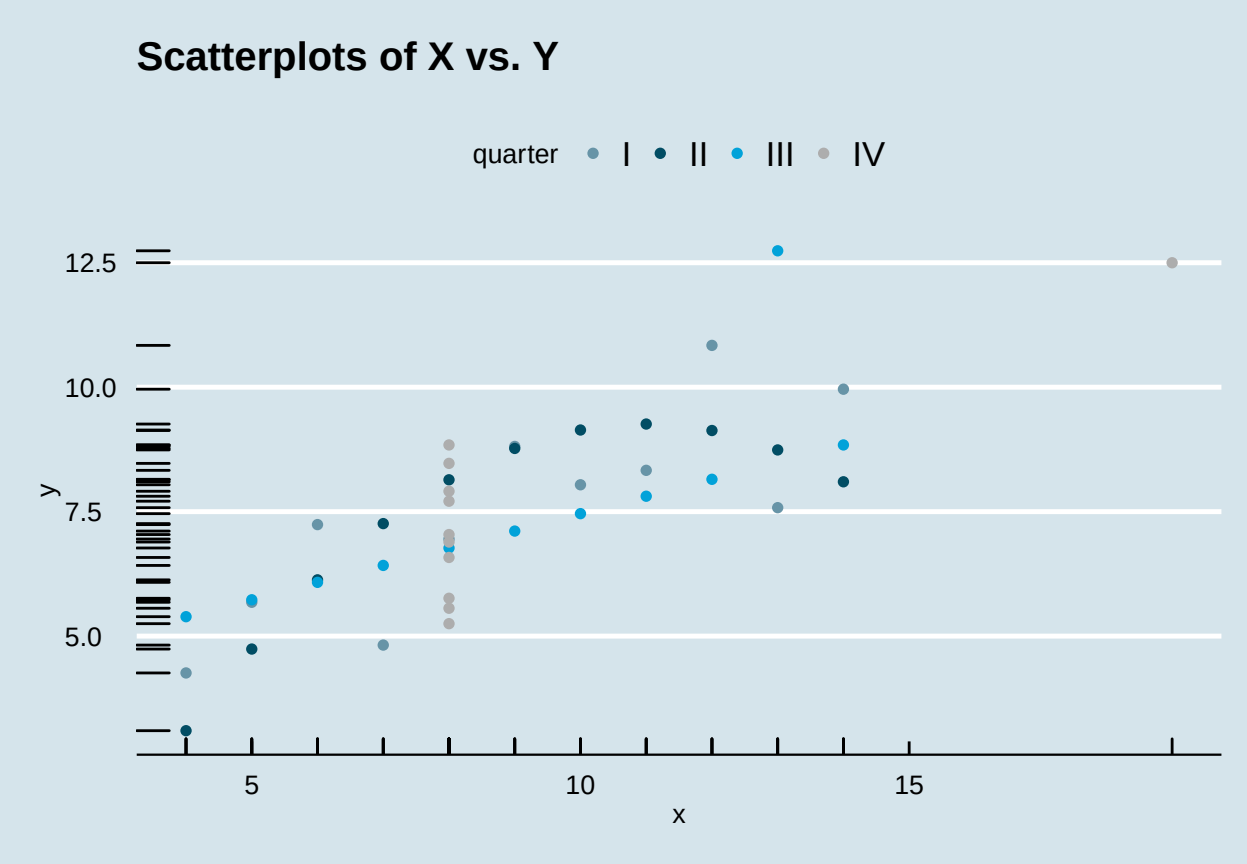
```
# create the data summary
summaries <- aggregate(data$y, by=list(data$quarter),FUN=summary)
# relabel the data summary
colnames(summaries)<-c('Group','')
summaries
```

```
##   Group  .Min. .1st Qu. .Median  .Mean .3rd Qu.  .Max.
## 1     I  4.260    6.315   7.580  7.501    8.570 10.840
## 2    II  3.100    6.695   8.140  7.501    8.950  9.260
## 3   III  5.390    6.250   7.110  7.500    7.980 12.740
## 4    IV  5.250    6.170   7.040  7.501    8.190 12.500
```

Graphical Exploration

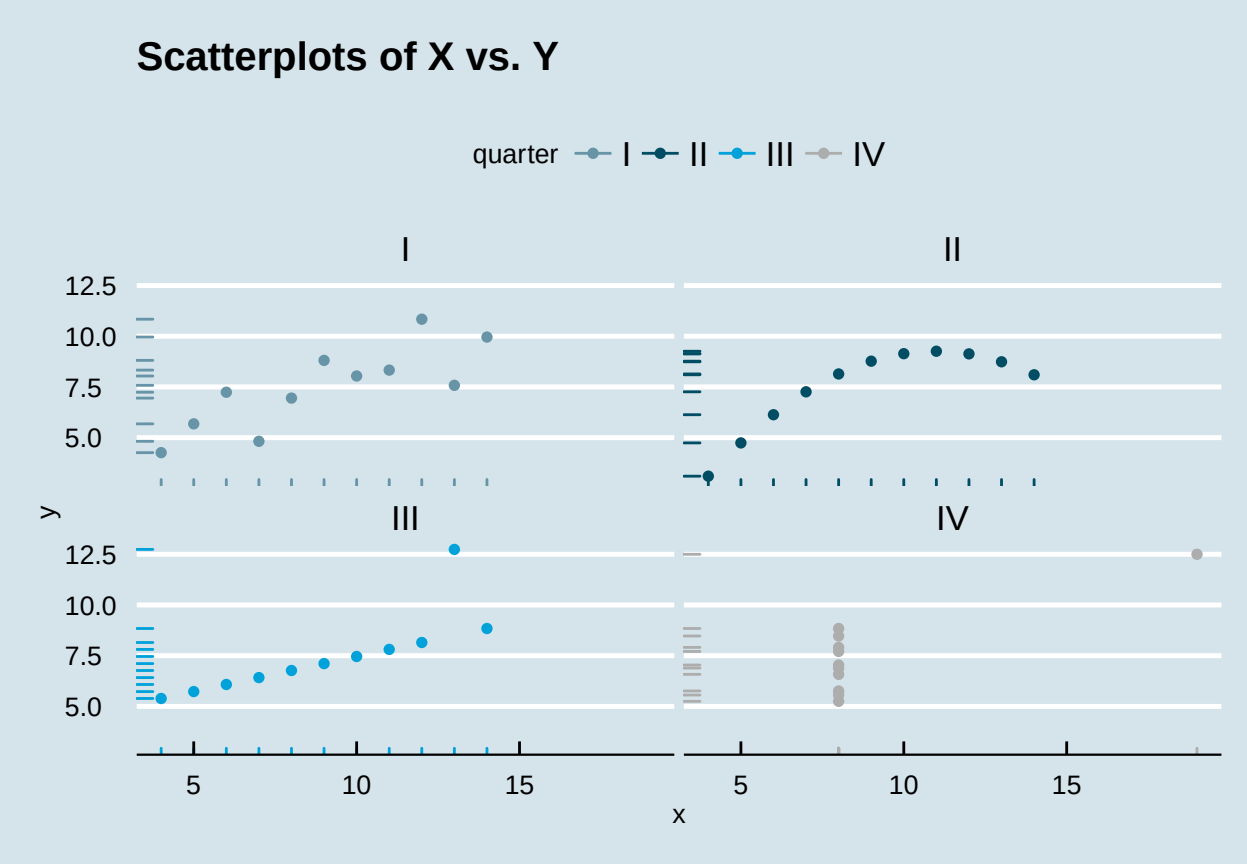
First, we create a single scatter plot varying the color by quarter.

```
# graph the data
p <- ggplot(data, aes(x, y))+theme_economist() +
  scale_colour_economist() + ggtitle("Scatterplots of X vs. Y") +
  geom_rug()
p + geom_point(aes(colour = quarter))
```



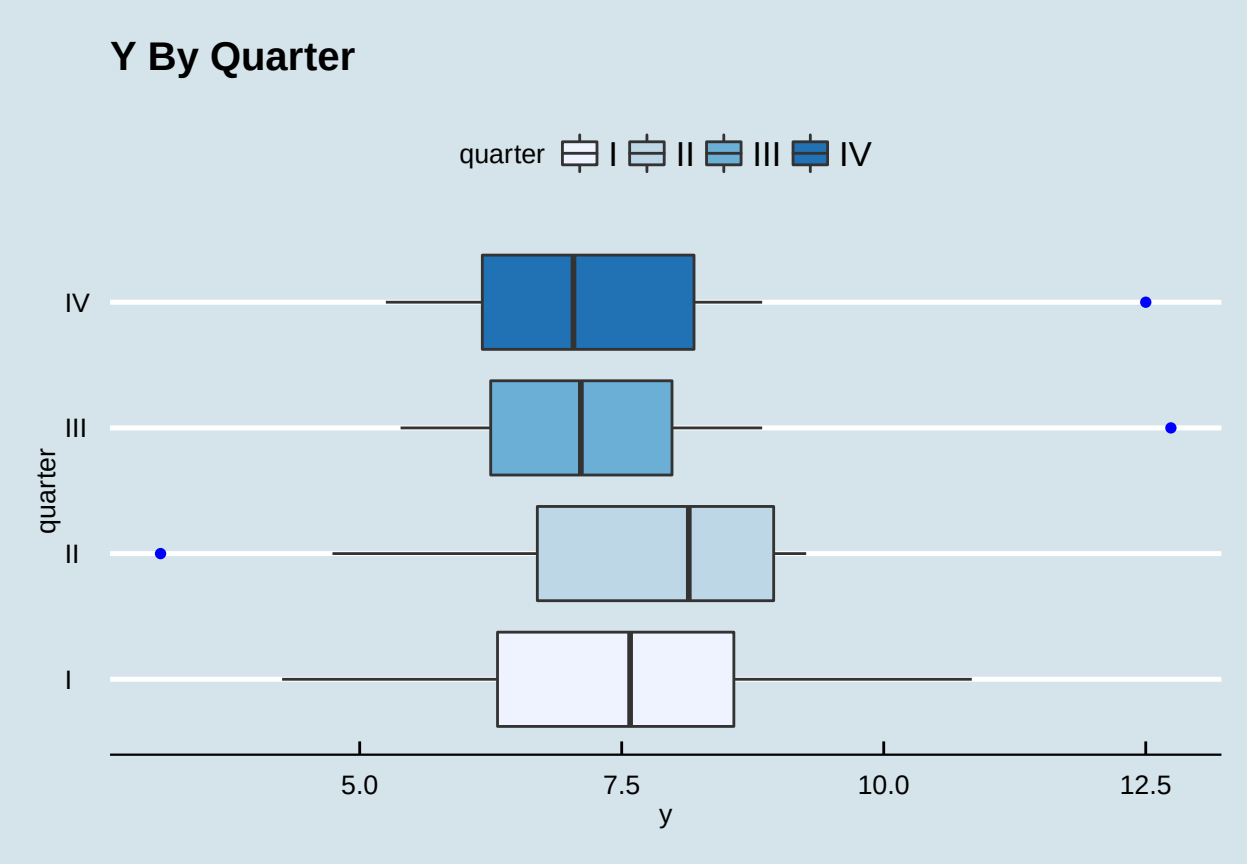
The patterns associated with each quarter are difficult to see on a single scatter plot, so we split out the quarters into separate scatter plots.

```
qplot(x,y, data=data,color=quarter,facets=~quarter,
      xlab="x", ylab="y",main="Scatterplots of X vs. Y") +
  geom_rug() + scale_fill_brewer() +
  theme_economist() + scale_colour_economist()
```



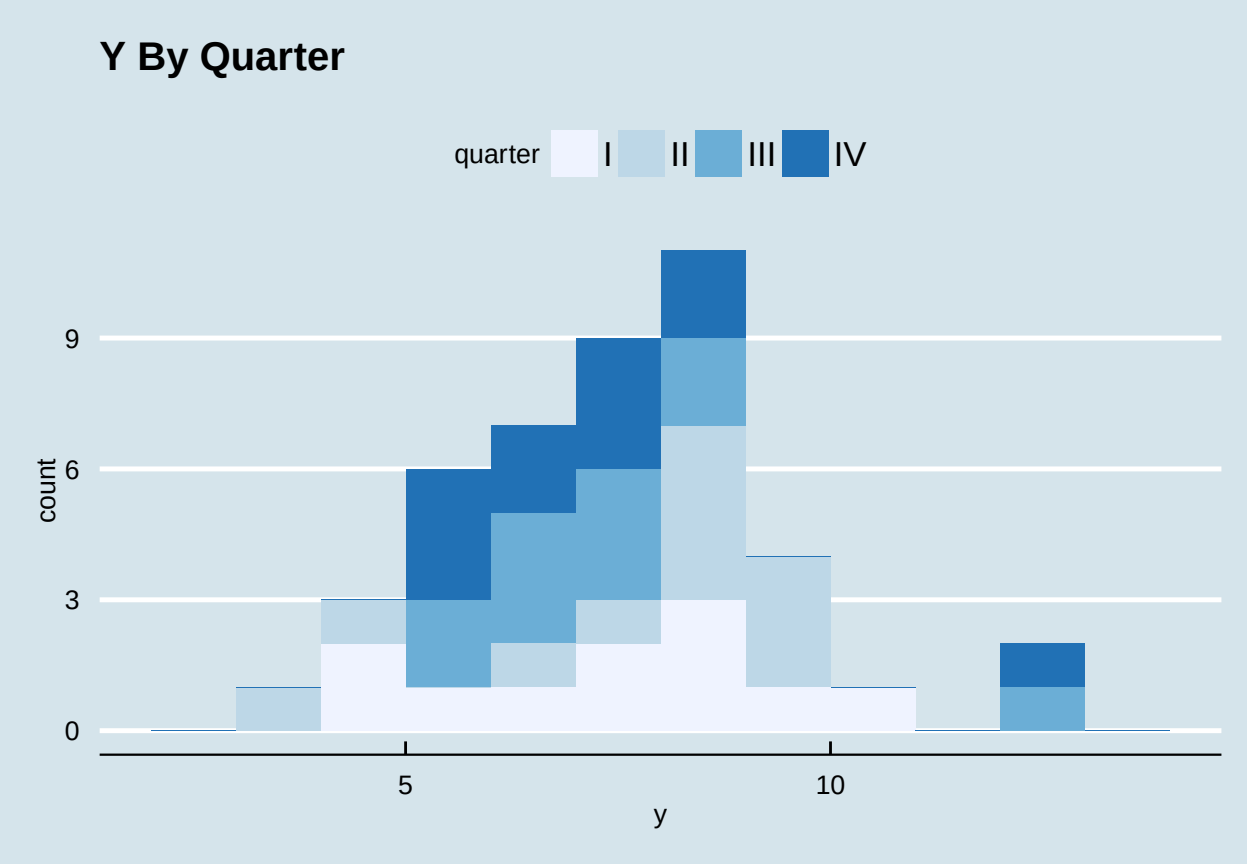
The patterns in each data set are clearer. Both III and IV appear to have outliers.

```
p <- ggplot(data, aes(quarter,y))
p + geom_boxplot(outlier.colour = "blue",aes(fill = quarter)) +
  coord_flip() + scale_fill_brewer() + ggtitle('Y By Quarter')+
  theme_economist() + scale_colour_economist()
```



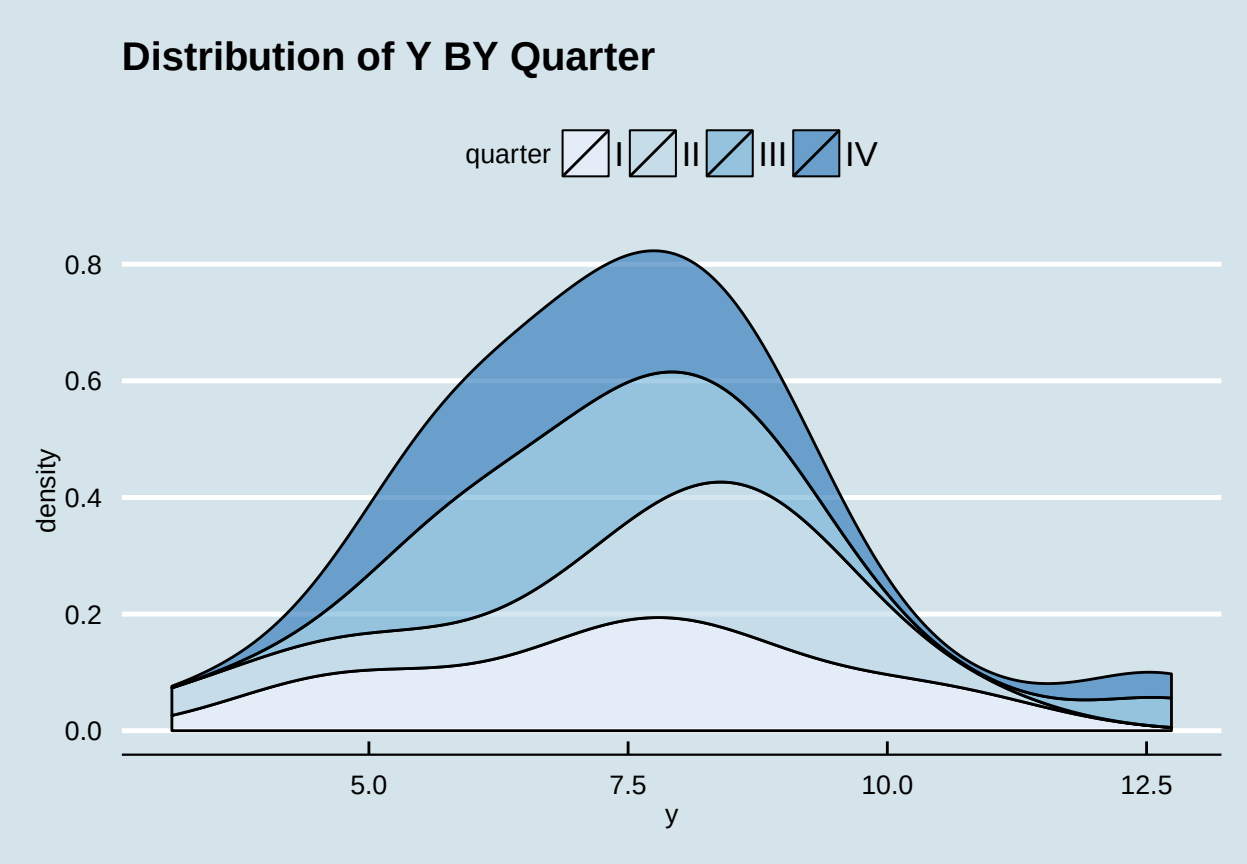
We now examine the distribution of Y, first by creating a histogram where color denotes the contribution by each quarter

```
# create a histogram where color denotes contribution by
# each quarter
dplot <- ggplot(data, aes(y, fill = quarter))
dplot + geom_bar(position = "stack",binwidth=1) +
  theme_economist() + scale_colour_economist() +
  scale_fill_brewer() + ggtitle('Y By Quarter')
```



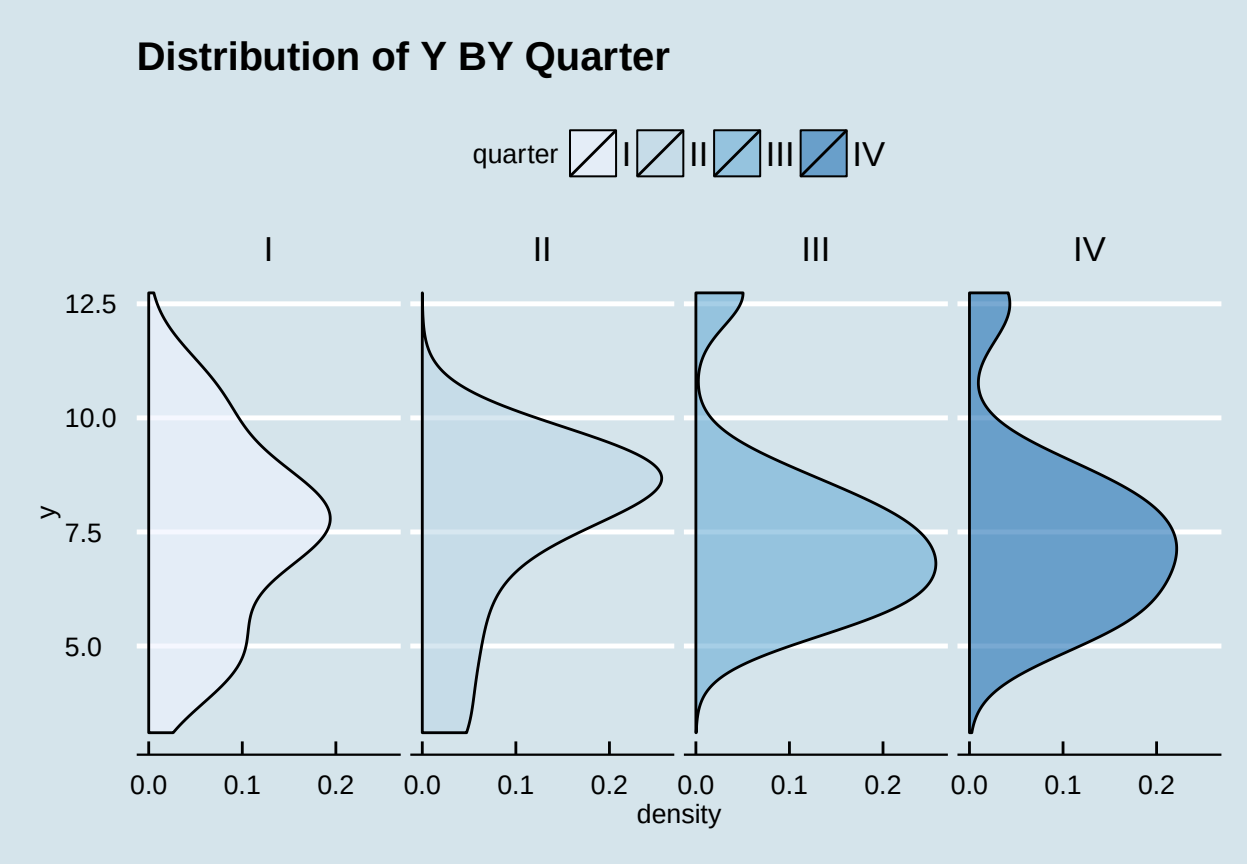
We continue to examine the distribution of Y, focusing next on kernel smoothed densities for each respective quarter.

```
# plot the kernal smoothed density by quarter
qplot(y, data=data, geom="density",position="stack", fill=quarter,
      alpha=I(.6), main="Distribution of Y BY Quarter",
      xlab="y", ylab="density") + scale_fill_brewer() +
  theme_economist() + scale_colour_economist()
```



The shape of the distribution of Y for each quarter is more clear once we look at the distribution for each quarter separately.

```
# plot the kernal smoothed density by quarter
qplot(y, data=data, geom="density",position="stack", fill=quarter,
      alpha=I(.6), main="Distribution of Y BY Quarter",
      xlab="y", ylab="density") + scale_fill_brewer() +
  theme_economist() + scale_colour_economist() +
  facet_grid(. ~ quarter) +  coord_flip()
```



The density differs significantly by quarter. The mode of the distribution shifts.