

GDA of *zagat* (from *smss*)

Background Zagat ratings of Italian restaurants in Boston, London, and New York.

Aims Are ratings in the three cities different? What patterns of ratings are there? How are ratings on different criteria related?

Source Agresti and Finlay's "Statistical Methods for the Social Sciences" (4th edition)

Structure 193 observations on 6 variables (4 numeric, 1 name, 1 factor)

Zagat ratings use four criteria (Food, Decor, Service, all reported on a scale of 0 to 30, and the Cost of an average meal in dollars). The numbers of restaurants in each city and the Cost distributions are shown below:

```
library(gridExtra)
data(zagat, package="smss")
a1 <- ggplot(zagat, aes(City)) + geom_bar() + ylab("") + xlab("") +
  ggtitle("Numbers of restaurants rated")
a2 <- ggplot(zagat, aes(City, Cost)) + geom_boxplot() + ylab("") + xlab("") +
  ggtitle("Average cost of a meal") + ylim(0, 150)
grid.arrange(a1, a2, nrow=1)
```

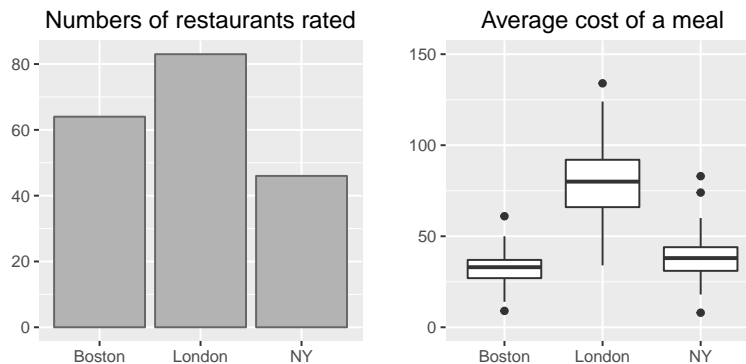


Fig 1: More Italian restaurants were rated in London than in either of New York or Boston. The meals were most expensive in London. (The dates the ratings were made and the exchange rates used are not given.)

The distributions of the three other criteria are as follows:

```
boxplot(zagat[, 3:5], pch=16, horizontal = TRUE, ylim=c(0,30), las=1)
```

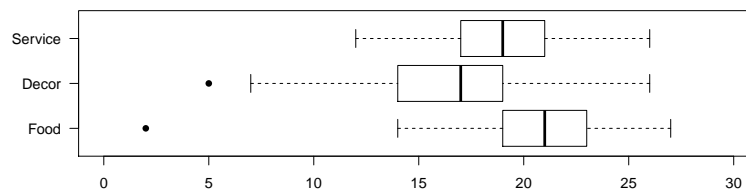


Fig 2: Boxplots of the ratings on the three criteria other than Cost. The low outlier on Food looks suspicious and that restaurant will be left out of further analyses.

```
library(dplyr)
zagatX <- zagat %>% filter(Food > 2)
```

For a small dataset like this a lot can be shown in a scatterplot matrix display.

```
library(GGally)
ggpairs(zagatX, columns=c(1, 3:6), upper = list(continuous = "cor",
  combo = "facetdensity"), lower = list(continuous = "points", combo = "box"),
  diag=list(continuous = "density", discrete="bar"))
```

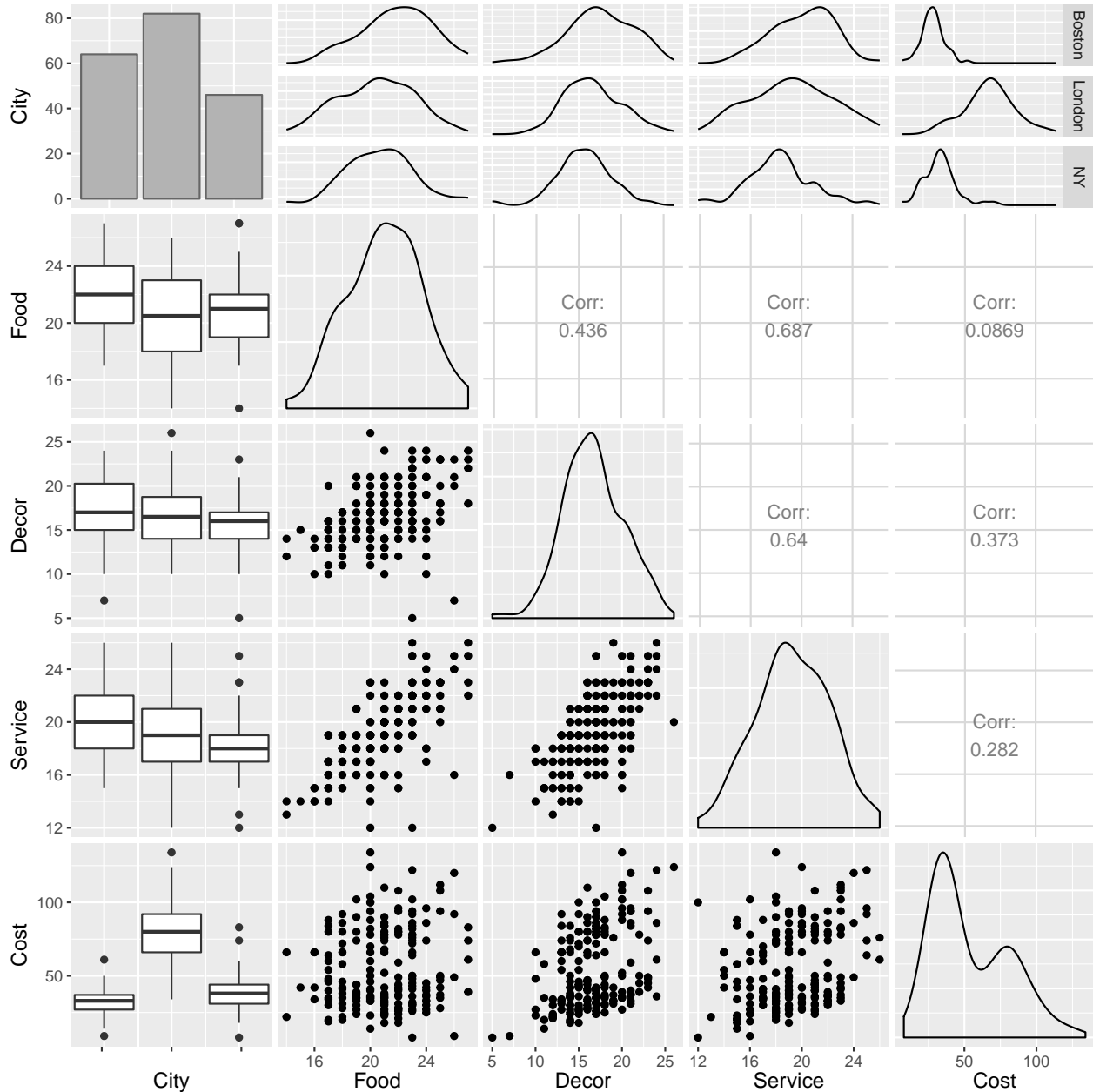


Fig 3: Food, Decor, and Service are fairly positively correlated. The distributions of these three criteria do not appear to vary much by city, but there are significant differences between some of the means as the next figure shows.

One way to get confidence intervals for each of the means is to fit linear models without an intercept and plot the resulting coefficients with intervals depending on their standard errors. This does assume that the variability is the same for each group.

```
library(coefplot)
library(gridExtra)
m1 <- lm(data=zagatX, Food~0 + City)
m2 <- lm(data=zagatX, Decor~0 + City)
m3 <- lm(data=zagatX, Service~0 + City)
g1 <- coefplot(m1, predictors="City", lwdOuter=0.5, title=
  "CIs for average Food scores") + xlab("") + ylab("") + xlim(13,23) + coord_flip()
g2 <- coefplot(m2, predictors="City", lwdOuter=0.5, title=
  "CIs for average Decor scores") + xlab("") + ylab("") + xlim(13,23) + coord_flip()
g3 <- coefplot(m3, predictors="City", lwdOuter=0.5, title=
  "CIs for average Service scores") + xlab("") + ylab("") + xlim(13,23) + coord_flip()

grid.arrange(g1, g2, g3, nrow=1)
```

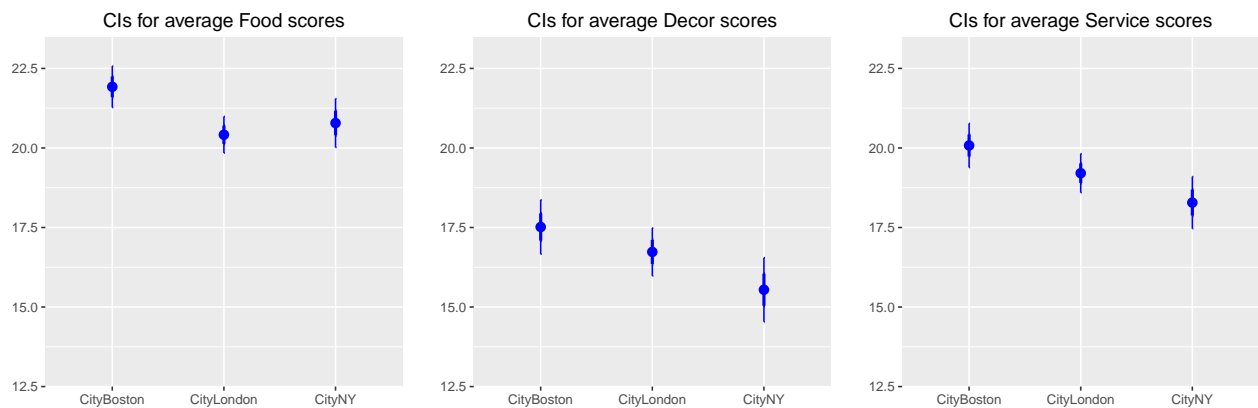


Fig 4: Confidence intervals for the mean ratings with widths equal to two standard errors. The average ratings are highest on Food and lowest on Decor. Boston has the highest average scores on all three criteria and is significantly better than New York on all and significantly better than London on Food.

These statements are apparent from the graphics (especially when you zoom in) and they can be confirmed by looking at summaries of the linear models.

For many people the relationship of Food to Cost is of most interest. Since London is much more expensive than the other two cities, separate highlighted scatterplots have been drawn.

```
library(extracat)
library(scales)
facetshade(data = zagatX, aes(x = Cost, y = Food), f = .~City) +
  geom_point(colour = alpha("black", 0.05)) + geom_point(data = zagatX,
  colour = "red") + facet_wrap(f=~City, nrow=1) + theme(legend.position="none")
```

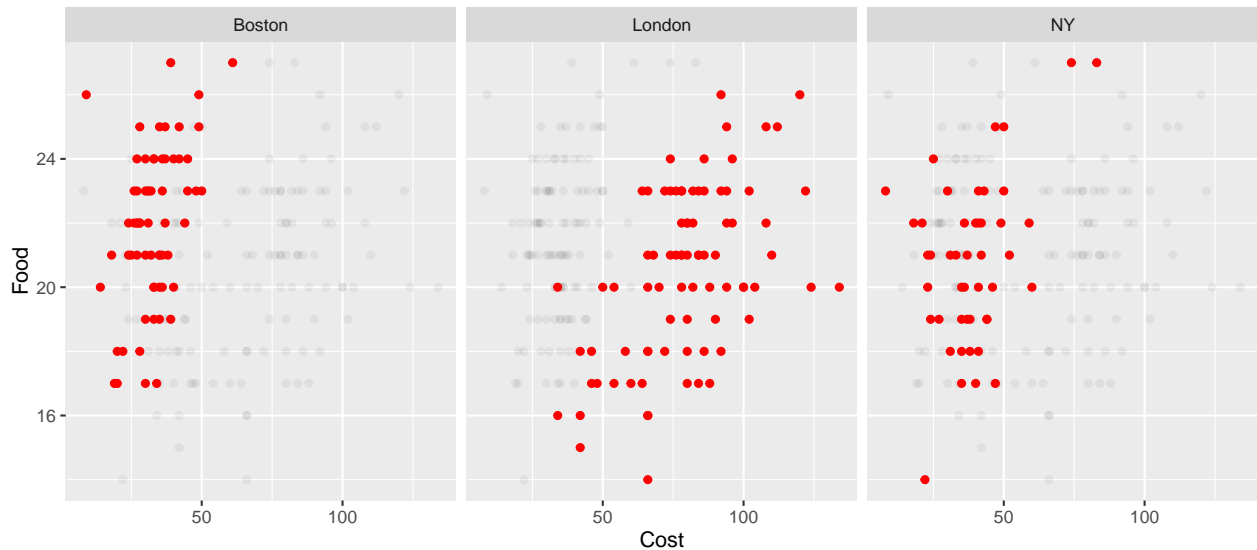


Fig 5: Scatterplots of Food against Cost for the three cities. There is some evidence of increasing food scores with increasing cost, especially for London, but there is considerable variability.

Fitting spline models to each City's data gives the following display:

```
ggplot(zagatX, aes(Cost, Food)) + geom_point() + facet_wrap(~City) + geom_smooth(method="gam")
```

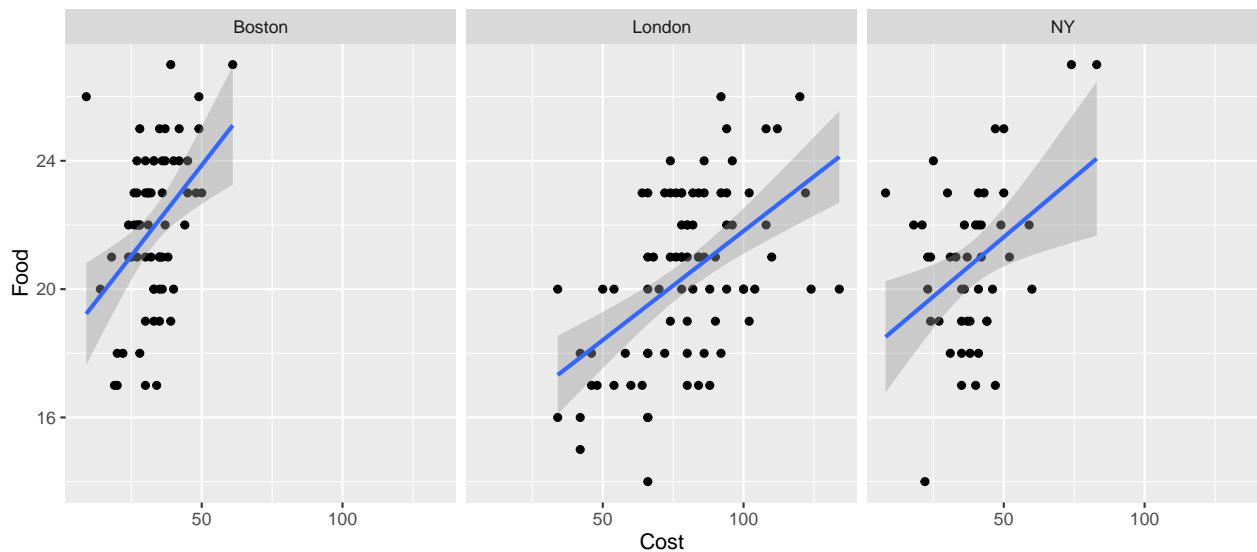


Fig 6: Scatterplots of Food against Cost for the three cities with spline smooths added. Nonlinear models do not seem to be necessary, but none of the three fits is very good.

The dataset is used in three exercises in Agresti and Finlay's book.