COMP 333 Data Analytics

Example Tipping in Restaurants

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Data Analytics — Example for Restaurant Tipping

Find the variables which *best predict* the **tip** given to the waiter.

The *variables* (also called *features*, *attributes*) in the data collected:

- the tip amount,
- total bill,
- payer gender,
- smoking/non-smoking section,
- time of day,
- day of the week, and
- size of the party

The approach is to fit a regression model to predict the tip rate.

The fitted **model** is

- $tip\_rate = 0.18 - 0.01 \times party\_size$

if size of the dining party increases by one (leading to a higher bill), the tip rate will decrease by 1%.

`https://en.wikipedia.org/wiki/Exploratory_data_analysis`
D. Cook and D.F. Swayne (with A. Buja, D. Temple Lang, H. Hofmann, H. Wickham, M. Lawrence) (2007), *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*, Springer, 978-0387717616, page 4 et seq.

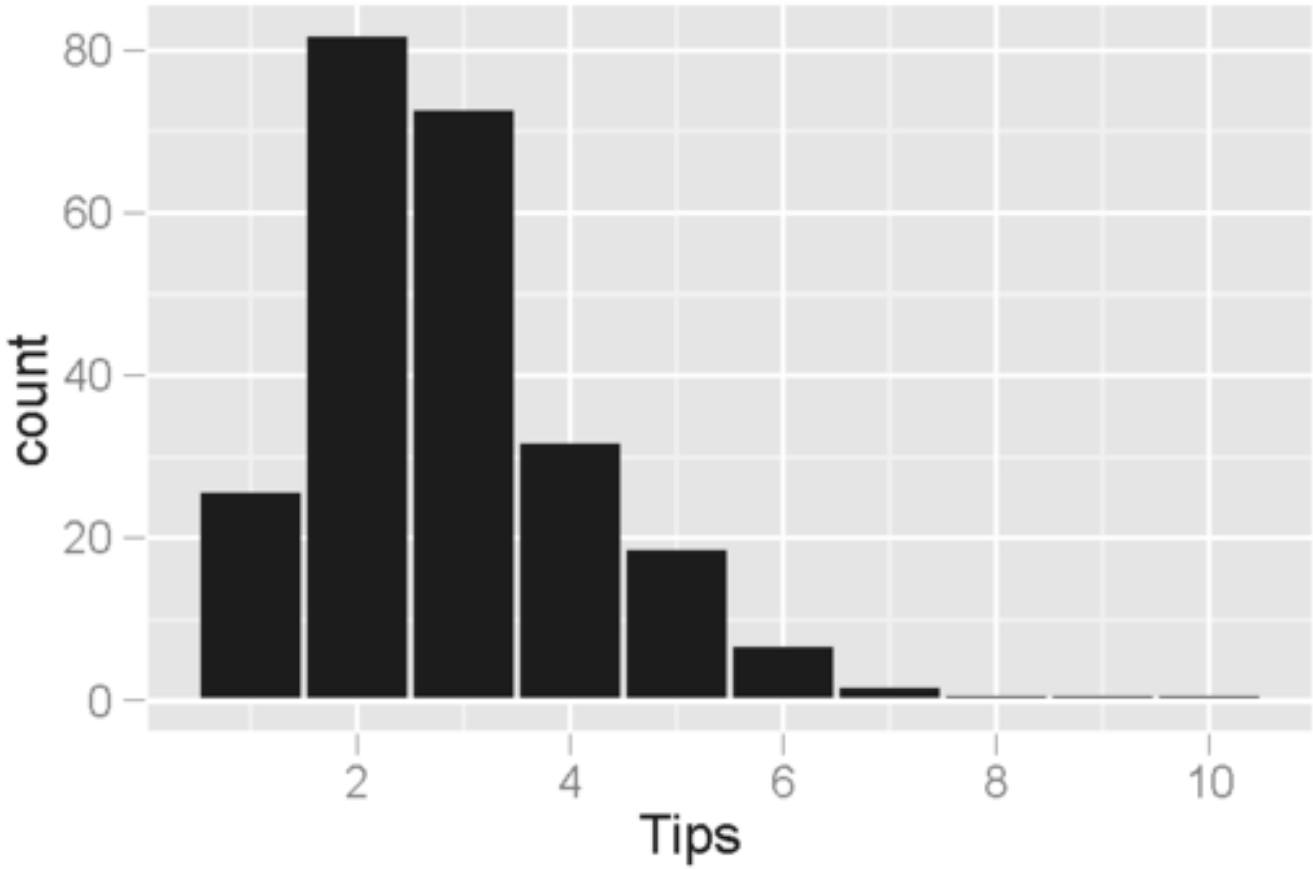
Dataset `http://vincentarelbundock.github.io/Rdatasets/csv/reshape2/tips.csv`

Exploratory Data Analysis

Regression Model does not tell the whole story

This is where Exploratory Data Analysis comes in!

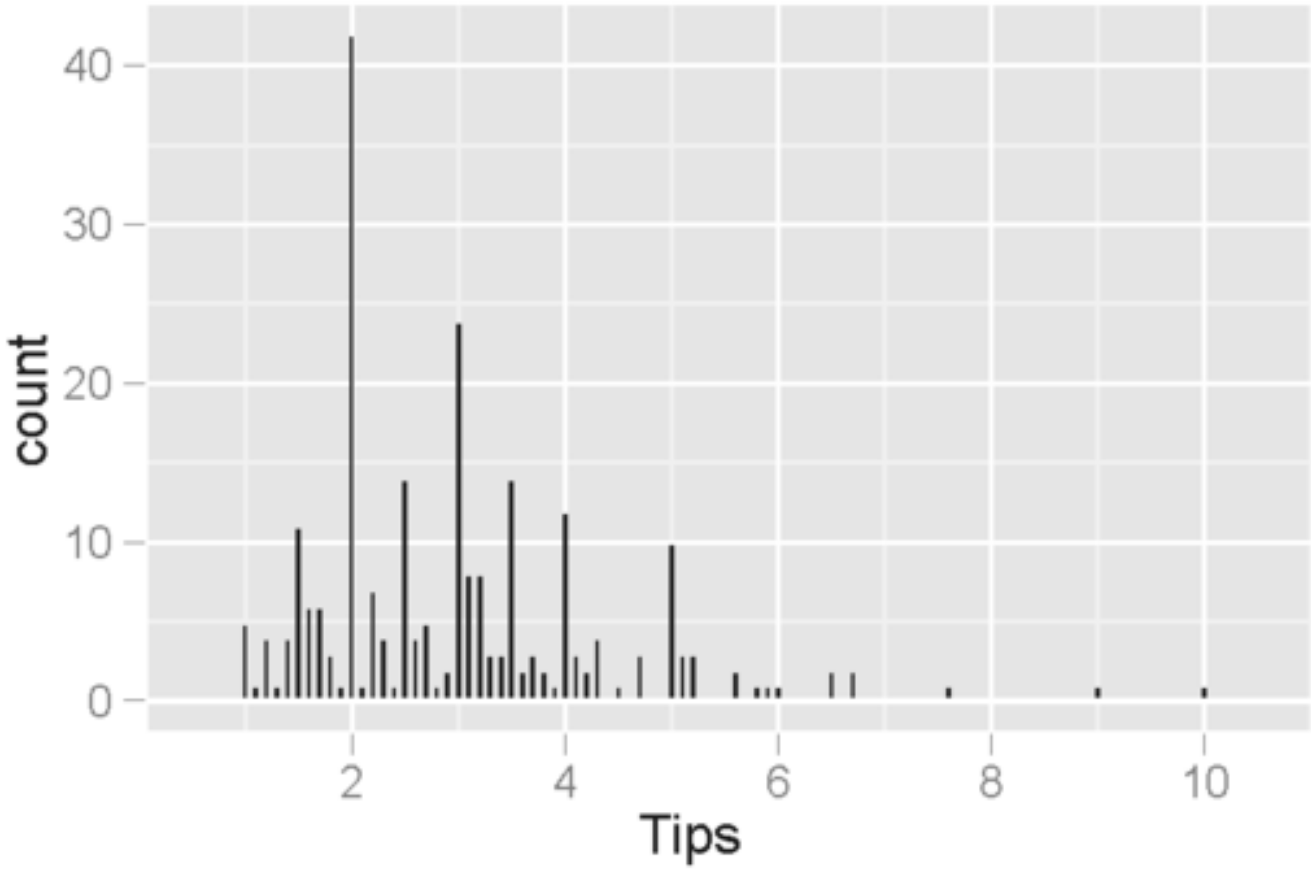
Histogram



Histogram of tip amounts where the bins cover $1 increments.
The distribution of values is skewed right and unimodal,
as is common in distributions of small, non-negative quantities.

https://en.wikipedia.org/wiki/Exploratory_data_analysis

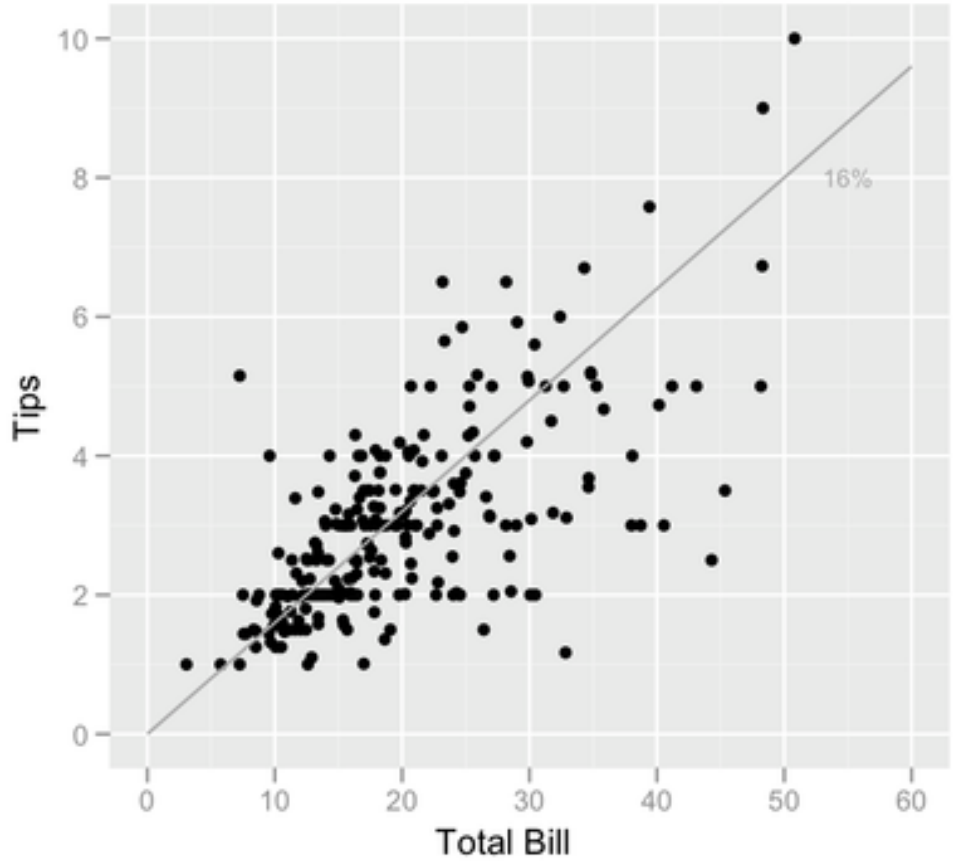
Histogram



Histogram of tip amounts where the bins cover \$0.10 increments.
An interesting phenomenon is visible:
peaks occur at the whole-dollar and half-dollar amounts,
which is caused by customers picking round numbers as tips.

`https://en.wikipedia.org/wiki/Exploratory_data_analysis`

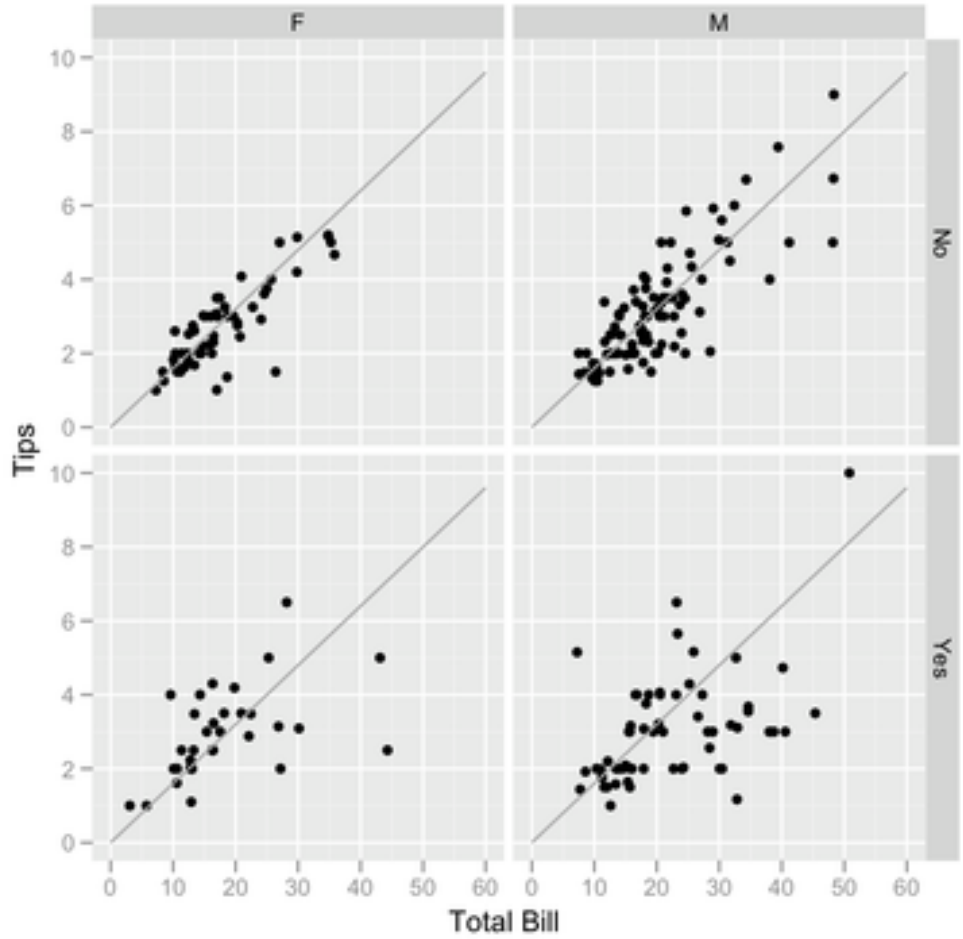
Scatter Plot



Points below the line correspond to tips that are lower than expected (for that bill amount), and points above the line are higher than expected.
We might expect to see a tight, positive linear association,
but instead see variation that increases with tip amount.
More customers are very cheap than very generous.

`https://en.wikipedia.org/wiki/Exploratory_data_analysis`

Scatter Plot



Scatterplot of tips vs. bill separated by payer gender & smoking section status.
Smoking parties have a lot more variability in the tips that they give.

`https://en.wikipedia.org/wiki/Exploratory_data_analysis`