Bivariate Data
Univariate (one variable) data

- Involves only a single variable
  - So cannot describe associations or relationships
- Descriptive Statistics
  - Central tendencies: mean, median, mode
  - Dispersion: range, max, min, quartiles, variance, standard deviation
- Visualizations
  - Pie charts, Bar charts, Line charts, Histograms, Box plots
Bivariate (two variables) data

- Involves two variables
  - So *can* describe associations or relationships

- Descriptive Statistics
  - Central tendencies: mean, median, mode
  - Dispersion: variance, standard deviation, covariance, correlation

- Visualizations
  - Scatter plots
Bivariate data and scatter plots
Bivariate data and scatter plots

- A scatter plot of bivariate data shows one variable vs. the other on a 2-dimensional graph
- If there is an explanatory variable, it is plotted on the horizontal (x) axis, and the response variable is plotted on the vertical (y) axis
  - If there is no explanatory-response distinction either variable can be plotted on either axis
- \( x \) is also known as the independent variable, and \( y \) the dependent
Covariance & Correlation
Hybrid cars sold in the U.S. from 1997 to 2013

The variables:

- Model of the car
- Year of manufacture
- MSRP (manufacturer's suggested retail price) in 2013 dollars
- Acceleration rate in km per second
- Fuel economy in miles per gallon
- Model’s class

First 15 rows of data

<table>
<thead>
<tr>
<th>vehicle</th>
<th>year</th>
<th>msrp</th>
<th>acceleration</th>
<th>mpg</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prius (1st Gen)</td>
<td>1997</td>
<td>24509.74</td>
<td>7.46</td>
<td>41.26</td>
<td>C</td>
</tr>
<tr>
<td>Tino</td>
<td>2000</td>
<td>35354.97</td>
<td>8.20</td>
<td>54.10</td>
<td>C</td>
</tr>
<tr>
<td>Prius (2nd Gen)</td>
<td>2000</td>
<td>26832.25</td>
<td>7.97</td>
<td>45.23</td>
<td>C</td>
</tr>
<tr>
<td>Insight</td>
<td>2000</td>
<td>18936.41</td>
<td>9.52</td>
<td>53.00</td>
<td>TS</td>
</tr>
<tr>
<td>Civic (1st Gen)</td>
<td>2001</td>
<td>25833.38</td>
<td>7.04</td>
<td>47.04</td>
<td>C</td>
</tr>
<tr>
<td>Insight</td>
<td>2001</td>
<td>19036.71</td>
<td>9.52</td>
<td>53.00</td>
<td>TS</td>
</tr>
<tr>
<td>Insight</td>
<td>2002</td>
<td>19137.01</td>
<td>9.71</td>
<td>53.00</td>
<td>TS</td>
</tr>
<tr>
<td>Alphard</td>
<td>2003</td>
<td>38084.77</td>
<td>8.33</td>
<td>40.46</td>
<td>MV</td>
</tr>
<tr>
<td>Insight</td>
<td>2003</td>
<td>19137.01</td>
<td>9.52</td>
<td>53.00</td>
<td>TS</td>
</tr>
<tr>
<td>Civic</td>
<td>2003</td>
<td>14071.92</td>
<td>8.62</td>
<td>41.00</td>
<td>C</td>
</tr>
<tr>
<td>Escape</td>
<td>2004</td>
<td>36676.10</td>
<td>10.32</td>
<td>31.99</td>
<td>SUV</td>
</tr>
<tr>
<td>Insight</td>
<td>2004</td>
<td>19237.31</td>
<td>9.35</td>
<td>52.00</td>
<td>TS</td>
</tr>
<tr>
<td>Prius</td>
<td>2004</td>
<td>20355.64</td>
<td>9.90</td>
<td>46.00</td>
<td>M</td>
</tr>
<tr>
<td>Silverado 15 2WD</td>
<td>2004</td>
<td>30089.64</td>
<td>9.09</td>
<td>17.00</td>
<td>PT</td>
</tr>
<tr>
<td>Lexus RX400h</td>
<td>2005</td>
<td>58521.14</td>
<td>12.76</td>
<td>28.23</td>
<td>SUV</td>
</tr>
</tbody>
</table>
Positive association

- The points are scattered in an upward direction, indicating that cars with greater acceleration tend to cost more.
- Conversely, cars that cost more tend to have greater acceleration.
- This is an example of positive association: above average values of one variable tend to be associated with above average values of the other.
Negative association

- There is a clear downward trend
- Hybrid cars with higher mpg tend to cost less; conversely, cars that cost more tend to have lower mpg
- This might seem confusing at first until we consider that cars that accelerate faster tend to be less fuel efficient and have lower mpg
Covariance

Measure of how changes in one variable are associated with changes in another: \( \text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - \mu_X \mu_Y \)

- If two variables are independent \((E[XY] = E[X]E[Y])\), then their covariance is 0.
- But if the covariance of two variables is zero, they are not necessarily independent, because covariance captures only linear associations.
  - These data sets exhibit relatively little covariance.
Properties of Correlation

- Symmetric measure: makes no distinction between explanatory and response variables (switching the axes will not affect $r$)
- Measures strength of linear relationships only
- Both variables must be quantitative

**Problem:** cannot compare covariances (i.e., Fahrenheit degrees $x$ dollars vs. Celsius degrees $x$ euros)

**Solution:** correlation = normalized covariance
Correlation

**Normalized** measure of how changes in one variable are associated with changes in another: \( \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \)

- If two variables are independent (\( E[XY] = E[X]E[Y] \)), then their correlation is 0.
- But if the correlation of two variables is zero, they are not necessarily independent, because correlation captures only **linear** associations.
  - These data sets exhibit zero or near-zero correlation.
Properties of Correlation

- Symmetric measure: makes no distinction between explanatory and response variables (switching the axes will not affect $r$)
- Measures strength of linear relationships only
- Both variables must be quantitative
- Is a number between -1 and 1
- Is invariant to change of units (Corr($X$, $Y$) is calculated using standard units)
Sample Covariance

c is the average product, across all observations, of deviations

\[ \gamma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \]

\[ c_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \]
Sample Correlation

is the average product, across all observations, of deviations (measured in standard units)

\[ \rho_{XY} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{x_i - \mu_x}{\sigma_x} \right) \left( \frac{x_i - \mu_y}{\sigma_y} \right) \]

\[ r_{XY} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{x_i - \bar{x}}{s_{xx}} \right) \left( \frac{x_i - \bar{x}}{s_{yy}} \right) \]
A toy example: Ice cream sales

Based on the scatter plot of temperature vs. ice cream sales, we expect $r$ to be positive and close to 1, as there seems to be a strong linear relationship.
Step 1: Convert values to standard units

- Let’s calculate temperature in standard units, starting with 16.4
- mean of temperature: 19.08182
- standard deviation of temperature: 3.755437

Value in standard units = (value - average) / SD

Temperature 16.4 in standard units = (16.4 - 19.08182) / 3.755437 = -0.714116

All the values in the columns of temperatures and sales in standard units were calculated this way.
Step 2: Multiply corresponding pairs of values in standard units
Step 3: $r$ is the average of the products of the values in standard units

- We compute the mean of the products in standard units
- In this case $r$ is 0.9585728
- This confirms our conjecture that $r$ is positive and close to 1
Computing $r$ in R

```r
> hybrid <- read.csv("hybrid.csv")
> cor(hybrid$msrp, hybrid$accelrate)
[1] 0.6955779
```
Computing $r$ in R

```r
> hybrid <- read.csv("hybrid.csv")
> cor(hybrid$msrp, hybrid$mpg)
[1] -0.5318264
```
Summary

Correlation measures the **direction** and **strength** of a **linear** relationship between two quantitative variables.
Correlation is powerful and simple but easy to misinterpret:

- *Correlation does not imply causation!*
  - Correlation only measures association.
- Correlation only measures **linear association**.
- **Outliers** can have a significant effect on correlation.
- Correlation can be misleading when data are **aggregated**.
Correlation does not imply causation

Intuitive Example:
- Imagine a positive correlation between math abilities and kids’ weight.
- Does this imply that students who are better at math gain weight easily?
- Or that gaining weight can improve a student’s math abilities?
- No! Of course not!
- Age is a confounding variable, which explains the correlation.

Older children both weigh more and are better at math than younger children.
Correlation measures linear associations

\[
x <- \text{seq}(-3, 3, \text{by} = 0.5)
\]

\[
y <- x ** 2
\]

\[
cor(x, y)
\]

\[
[1] 0
\]

\[
\text{plot}(x, y, \text{xlab} = "x", \text{ylab} = "y")
\]
Outliers can gravely impact correlation

> cor(line$x, line$y)
[1] 1

> cor(outlier$x, outlier$y)
[1] 0
Correlations across aggregated data

- When the data for each country are collapsed to a single point, the variables seem strongly correlated.
- But individuals vary, so there is really a cloud of points per country.
- Thus, the true correlation is lower than the value calculated using aggregate data.
- Aggregate (and possibly spurious) correlations like these are known as ecological correlations.
Correlations across aggregated data

First 15 rows of average SAT scores by state in 2015

<table>
<thead>
<tr>
<th>State</th>
<th>Critical Reading</th>
<th>Mathematics</th>
<th>Writing</th>
<th>Total SAT score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>545</td>
<td>538</td>
<td>533</td>
<td>1616</td>
</tr>
<tr>
<td>Alaska</td>
<td>509</td>
<td>503</td>
<td>482</td>
<td>1494</td>
</tr>
<tr>
<td>Arizona</td>
<td>523</td>
<td>527</td>
<td>502</td>
<td>1552</td>
</tr>
<tr>
<td>Arkansas</td>
<td>568</td>
<td>569</td>
<td>551</td>
<td>1688</td>
</tr>
<tr>
<td>California</td>
<td>495</td>
<td>506</td>
<td>491</td>
<td>1492</td>
</tr>
<tr>
<td>Colorado</td>
<td>582</td>
<td>587</td>
<td>567</td>
<td>1736</td>
</tr>
<tr>
<td>Connecticut</td>
<td>504</td>
<td>505</td>
<td>504</td>
<td>1514</td>
</tr>
<tr>
<td>Delaware</td>
<td>462</td>
<td>461</td>
<td>445</td>
<td>1366</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>441</td>
<td>440</td>
<td>432</td>
<td>1313</td>
</tr>
<tr>
<td>Florida</td>
<td>485</td>
<td>480</td>
<td>468</td>
<td>1434</td>
</tr>
<tr>
<td>Georgia</td>
<td>490</td>
<td>485</td>
<td>475</td>
<td>1450</td>
</tr>
<tr>
<td>Hawaii</td>
<td>487</td>
<td>508</td>
<td>477</td>
<td>1472</td>
</tr>
<tr>
<td>Idaho</td>
<td>467</td>
<td>463</td>
<td>442</td>
<td>1372</td>
</tr>
<tr>
<td>Illinois</td>
<td>599</td>
<td>616</td>
<td>587</td>
<td>1802</td>
</tr>
<tr>
<td>Indiana</td>
<td>496</td>
<td>499</td>
<td>478</td>
<td>1473</td>
</tr>
</tbody>
</table>
Correlations across aggregated data

```
> cor(sat$Critical.Reading, sat$Mathematics)
[1]  0.9843772
```

- This correlation does not reflect the true relationship between students’ Math and Critical Reading scores
- States do not take tests; students do
A 2012 paper in the New England Journal of Medicine

Some responded harshly:
http://blogs.scientificamerican.com/the-curious-wavelength/chocolate-consumption-and-nobel-prizes-a-bizarre-juxtaposition-if-there-ever-was-one/

Other responses were more relaxed: