Investigating Real World Data with Online Visualization Tools: Building Future Data Scientists

Hollylynne S. Lee, Gemma Mojica, Christina Azmy
NC State University
April 26, 2018
NCTM Annual Conference
Washington, DC

Presentation partially supported by grants from Hewlett Foundation and NSF DUE 1625713.
Opinions and findings are those of the presenter.
Statistics and Data Science Careers

Highly ranked careers in the past 5 years

Growth in demands expected to be over 30% through 2024

http://thisisstatistics.org
What is in our mathematics curriculum to prepare students for these careers?
To engage in data science and statistics, one uses ____________ skills and understandings to create insight from data?
Go to Menti.com Use code 809206

To engage in data science and statistics, one uses _________ skills and understandings to create insight from data?

Enter a word

Enter another word

Enter another word

Submit
What is data science?

Drew Conway, 2013
David Taylor, 2016
Statistical and Data Literacy should include knowing ...

- who collects data about us, why and how they collect it;
- how to analyze and interpret data from random and non-random samples;
- issues of data privacy and ownership;
- how to create representations of data to answer questions about real-life processes;
- the importance of the provenance of data and how it is stored;
- why data must sometimes be altered before analysis; and
- some aspects of predictive modeling.

Gould (2017)
CODAP: Common Online Data Analysis Platform
https://codap.concord.org/
Our Goal: Increase opportunities for learning data science and statistics.....

within existing curricula constraints
Core Design Principles for Data, Tools, and Tasks

• **Data** is real (collected by students or authenticated by teacher), multivariate (categorical & quantitative), “large”, and sometimes messy
• **Data** contexts are engaging to students

• **Tools** facilitate data moves, in tabular and graphical form
• **Tools** support links among representations of data

• **Tasks** have multiple entry points for different levels of sophistication
• **Tasks** provoke curiosity and promote different ways of engaging with data
Setting the context....

Ever worry about the cost of gas to support your commute to work and active lifestyle?

Do you wonder how the automobile industry may have made vehicles more or less fuel efficient?

Ever worry about how the automobile industry impacts our environment?
Investigating Fuel Economy of Vehicles

GETTING TO KNOW THE NEW FUEL ECONOMY AND ENVIRONMENT LABEL

The diagram below shows a sample label for a gasoline vehicle. Slightly different designs are used for flexible-fuel vehicles, electric vehicles, and plug-in hybrids. For more in-depth descriptions of label information for all vehicle types, visit www.fueleconomy.gov.

Fuel economy for city, highway, and combined city/highway driving (55% city and 45% highway) in miles per gallon.

Combined city/highway fuel economy range for vehicles in the same EPA vehicle class and fuel economy for most efficient vehicle overall for model year.

Technology and fuel type, such as gasoline, diesel, flexible fuel (gasoline–ethanol (E85)), plug-in hybrid (electricity–gasoline), and electric vehicle.

Fuel consumption in gallons per 100 miles in combined city/highway driving.

Annual fuel cost based on 15,000 miles per year at $2.45 per gallon for regular unleaded and $2.80 for premium.

Annual mileage, fuel cost, and other assumptions used to calculate estimates on the sticker.

Fuel economy and tailpipe greenhouse gas emissions to those of other vehicles of the same model year on a scale of 1 (worst) to 10 (best).

Upstream emissions estimates are available at fueleconomy.gov.

Compare tailpipe emissions of smog-producing pollutants to those of other vehicles of the same model year on a scale of 1 (worst) to 10 (best).

Fuel economy & greenhouse gas rating. High ratings reduce fuel use and smog-producing pollutants. This vehicle emits 70% fewer grams of pollutants per mile than the average new vehicle. See fueleconomy.gov for more information.

Fuel cost savings or extra expenditure compared to an "average" new vehicle of the same model year (25 mpg for model year 2017).

You save $1,750 in fuel costs over 5 years compared to the average new vehicle.

QR Code: Scan the code with your smartphone to link to official information on fueleconomy.gov (requires scanner app).
Look up a few cases!

www.fueleconomy.gov/feg/findacar.shtml
A Vehicle: 2007 Toyota Sienna

Fuel Economy  | Energy and Environment  | Safety  | Sp
--- | --- | --- | ---

**2007 Toyota Sienna 2WD**

---

**Regular Gasoline**

MPG estimates for 2016 and older vehicles may have been revised. Learn more.

Unofficial MPG Estimates from Vehicle Owners. Learn more about "My MPG" feature.

You save or spend*

Note: The average 2018 vehicle gets 27 MPG.

You spend $3,000 more in fuel costs over 5 years compared to the average new vehicle.

- U.S. barrel
- Imported barrel
1 barrel = 42 gallons

**Energy Impact Score**

Annual Petroleum Consumption

- Regular Gasoline
  - 17.3 barrels

Greenhouse Gas Emissions

- Regular Gasoline
  - 468 grams per mile

Units:

- Grams per mile
- Tailpipe CO2

For model years 2012 and earlier, tailpipe CO2 is estimated using an EPA emissions factor and does not reflect direct test results.

**EPA Smog Rating**

State of purchase:

- Select State

**Vehicle Specification Data**

- EPA Size Class: Minivan - 2WD
- Drive: Front-Wheel Drive
- Start-Stop Technology: Not Available
- Gas Guzzler: No
- Turbocharger: No
- Supercharger: No
- Passenger Volume: Regular
- Luggage Volume: Regular
- Fuel Type: Regular Gasoline
- Engine Descriptor: CLKUP
- Transmission Descriptor: 5sp
Going beyond a few cases....

If we looked at data *for many vehicles*, what could we learn about fuel economy in vehicles?
Let’s explore some data from 2015

What vehicle attributes may be of interest?
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Division</td>
<td>Company name</td>
</tr>
<tr>
<td>Carline</td>
<td>Model name of the vehicle</td>
</tr>
<tr>
<td>Vehicle_Type</td>
<td>Identifies the vehicle as car, SUV, van, minivan, or truck</td>
</tr>
<tr>
<td>Eng_Displacement</td>
<td>A measure of volume (in liters) of the total volume of air in all the cylinders in an engine as air is swept through the chambers in a complete cycle.</td>
</tr>
<tr>
<td>No_Cylinders</td>
<td>Number of cylinders in an engine</td>
</tr>
<tr>
<td>City_MPG</td>
<td>Estimated miles per gallon in city driving</td>
</tr>
<tr>
<td>Hwy_MPG</td>
<td>Estimated miles per gallon in highway driving</td>
</tr>
<tr>
<td>Comb_MPG</td>
<td>Estimated miles per gallon in a combination of city driving (55%) and highway driving (45%)</td>
</tr>
<tr>
<td>Guzzler</td>
<td>Identifies whether the vehicle has exceptionally low fuel economy (Y, N)</td>
</tr>
<tr>
<td>Transmission</td>
<td>Identifies vehicles as manual or automatic transmission</td>
</tr>
<tr>
<td>Detail_Transmission</td>
<td>Detailed description of transmission type</td>
</tr>
<tr>
<td>No_Gears</td>
<td>Number of transmission gears</td>
</tr>
<tr>
<td>Drive_Desc</td>
<td>Drivetrain (2-wheel, 4-wheel, or all-wheel)</td>
</tr>
<tr>
<td>Fuel_Usage</td>
<td>Type of fuel (premium, midgrade, regular, diesel)</td>
</tr>
<tr>
<td>AnnualFuel_Cost</td>
<td>Estimated annual fuel cost assuming 15,000 miles per year (55% city and 45% highway) and average fuel price</td>
</tr>
<tr>
<td>Hybrid</td>
<td>Identifies whether the vehicle has a hybrid engine (Y, N) such that it utilizes more than one form of onboard energy to achieve propulsion. A hybrid will have a traditional engine and fuel tank, as well as one or more electric motors and a battery pack</td>
</tr>
</tbody>
</table>
1226 vehicles from 2015

We could use all data
Or
We could use a random sample of the data
Or
We could use random samples from the data
Or
We could take random samples from the data
Different Data Formats

Sample of 300 vehicles (Option 2)

[link]
tinyurl.com/2015VehiclesSample

Option 3:   Download CSV file and import into CODAP

tinyurl.com/vehiclescsv   codap.concord.org

Option 4:   Data in CODAP with a Sampler designed

tinyurl.com/sampler2015vehicles
Examining Relationships in Samples of Data and Building Inference Ideas
Discussion Questions

What were the benefits of using this large multivariate data set? What about drawbacks?

What data moves did students use to model and reason about the linear relationship?

To what extent does this task promote curiosity?
Discussion of Data Tools and Task

- **Data** is real (collected by students or authenticated by teacher), multivariate (categorical & quantitative), “large”, and sometimes messy
- **Data** contexts are engaging to students

- **Tools** facilitate data moves, in tabular and graphical form
- **Tools** support links among representations of data

- **Tasks** have multiple entry points for different levels of sophistication
- **Tasks** provoke curiosity and promote different ways of engaging with data
Sampling Vehicles and Collecting Slopes

https://tinyurl.com/samplingVehiclesSlope
Ready to Learn More?

Join educators from around the world in FREE online professional development courses!

Teaching Statistics Through Inferential Reasoning--Going on now! Register by May 1 and course ends May 28th

go.ncsu.edu/tsir
Contact Us!

Hollylynne Lee  
Professor of Mathematics & Statistics Education  
Director of HI-RiSE  
hollylynne@ncsu.edu

hirise.fi.ncsu.edu