Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication

Arjun Srinivasan  Steven M. Drucker  Alex Endert  John Stasko
What are data facts?

How can we integrate data facts into visualization tools?

Why is this integration beneficial?
What are data facts?

How can we integrate data facts into visualization tools?

Why is this integration beneficial?
BRACE YOURSELVES
AUTO-INSIGHT SYSTEMS ARE COMING
Auto-insight Systems

Statistics + Natural Language Generation (NLG)

Sales for US is 3 times Europe
There is a rising trend of SUV’s market share
There is a falling trend of Sedan’s market share
Acceleration and Horsepower exhibit a moderate correlation
Displacement and Horsepower exhibit a strong correlation
Auto-insight Systems > Visualization Systems
For every day, “Dinner” increases by about 2.41.

30 September has the lowest value for “Food total” (10.86) and the second-lowest value for “Dinner” (10.86).

There is a correlation between “Total Downtime Minutes Max” and “Total Defect Qty Max” for 6.
For every day, “Dinner” increases by about 2.41.

VIS ‘17 expenses
The analysis measures Total Revenue and Sum of Labor Costs Variable across 50 cities.

- As Sum of Labor Costs Variable increased, Total Revenue increased based on the data provided. Specifically, when Sum of Labor Costs Variable increased by $1, Total Revenue increased by $4. There may be other factors contributing to Total Revenue, but there is evidence of a very strong relationship.
- When organized into groups of similar Sum of Labor Costs Variable and Total Revenue values, one distinct group stands out. There were 44 cities that had values of Sum of Labor Costs Variable between $0 and $1.5 million and Total Revenue between $0 and $4.8 million.
- The distribution of Sum of Labor Costs Variable ranges from $0 to $2 million. The average Sum of Labor Costs Variable per city is $287,794 and the median is $84,637.
- The minimum value for Total Revenue is $0 and the maximum value is $14.4 million. The average Total Revenue per city is $2.1 million and the median is $687,266.

In July 2014, 54 percent global mobile operating system market share was owned by Android. Goodness gracious! That’s more than half!

Between December 2008 and December 2014, their peak month was November 2014 at 60 percent. The leader in usage at the time was Android.
In July 2014, 54 percent global global mobile operating system market share was owned by Android. Goodness gracious! That's more than half!

Between December 2008 and December 2014, their peak month was November 2014 at 60 percent.

The leader in usage at the time was Android.
• For every day, “Dinner” increases by about 2.41

• There is a correlation between Acceleration and Horsepower

• Between December 2008 and December 2014, Android’s peak month was November 2014 at 60%

• Five cylinder cars have highest average Acceleration

“Insights”?
Defining “Insights”

• Toward measuring visualization insight. North (2006)

• Defining insight for visual analytics. Chang et al. (2009)
• For every day, “Dinner” increases by about 2.41

• There is a correlation between Acceleration and Horsepower

• Between December 2008 and December 2014, Android’s peak month was November 2014 at 60%

• Five cylinder cars have highest average Acceleration
Data Fact:

“a textual description of the result of one or more statistical functions applied to the data used to create a visualization.”
“a textual description of the result of one or more statistical functions applied to the data used to create a visualization.”

Can we do more with data facts?
Can data facts present communication-oriented alternatives?

US manufactures highest number of cars. The highest number of US cars were manufactured in 1973.
Can data facts present communication-oriented alternatives?

US manufactures highest number of cars. The highest number of US cars were manufactured in 1973.
Can data facts aid exploratory data analysis?

Cylinders: 4 has item (peugeot 504) with highest value for Acceleration
Can data facts aid exploratory data analysis?

Cylinders: 4 has item (peugeot 504) with highest value for Acceleration

Cylinders: 5 has highest average Acceleration
Can data facts aid exploratory data analysis?

- Cylinders: 4 has item (peugeot 504) with highest value for Acceleration
- Cylinders: 5 has highest average Acceleration
What are data facts?

How can we integrate data facts into visualization tools?

Why is this integration beneficial?
Interactive Data Facts

Voder

Manual View Specification

Interactive Data Facts

Disc-shaped voice-box translation devices that enabled the Phylosians to converse in audible language with humanoid visitors.
Data Fact Search Panel
Demo
Visualization → Facts \(\times\) Embellishments

Heuristics
<table>
<thead>
<tr>
<th>Attribute Combination</th>
<th>Example Data Fact</th>
<th>Task(s)</th>
<th>Visualization</th>
<th>Embellishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>pontiac grand prix has highest value for Horsepower</td>
<td>Find Extremum</td>
<td>Strip plot</td>
<td>O S IL TH RL CH QL</td>
</tr>
<tr>
<td></td>
<td>Most values for Horsepower are in the range 75 - 125</td>
<td>Characterize Distribution</td>
<td>Box plot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pontiac catalina appears to be an outlier</td>
<td>Find Anomalies</td>
<td>Strip plot</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Europe has the least number of items</td>
<td>Find Extremum</td>
<td>Bar chart</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of items in US is 2.57 times the number of items in Europe</td>
<td>Characterize Distribution</td>
<td>Donut chart</td>
<td></td>
</tr>
<tr>
<td>N x N</td>
<td>Acceleration and Displacement have a strong inverse correlation</td>
<td>Correlation</td>
<td>Scatterplot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Most items in the dataset have high Horsepower and low MPG</td>
<td>Characterize Distribution</td>
<td>Barplot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average Retail Price of SUV is 1.76 times Sedan</td>
<td>Characterize Distribution</td>
<td>Barplot</td>
<td></td>
</tr>
<tr>
<td>C x N</td>
<td>Japan has highest average MPG</td>
<td>Find Extremum</td>
<td>Bar chart</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(+Derived Value)</td>
<td>Donut chart</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Europe has item (flat 128) with lowest value for Displacement</td>
<td>Find Extremum</td>
<td>Strip plot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(+Derived Value)</td>
<td>Stacked bar chart</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C x C</td>
<td>The largest group of items in the dataset have Origin:Europe and Cylinders:5</td>
<td>Find Extremum</td>
<td>Stacked bar chart</td>
<td></td>
</tr>
<tr>
<td></td>
<td>US has most number of items. Most items in US belong to 8 for Cylinders</td>
<td>Characterize Distribution</td>
<td>Scatterplot + Size</td>
<td></td>
</tr>
<tr>
<td>N x N x N</td>
<td>Most items with low MPG and low Weight also have low Horsepower</td>
<td>Characterize Distribution</td>
<td>Scatterplot + Size</td>
<td></td>
</tr>
<tr>
<td>C x N x N</td>
<td>Overall, Displacement and Weight have a strong correlation</td>
<td>Correlation</td>
<td>Scatterplot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Items with Origin:Japan exhibit a strong correlation between Displacement and Weight</td>
<td>Correlation</td>
<td>Scatterplot + Color</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Most items with Origin:Europe have low Displacement and low Weight</td>
<td>Characterize Distribution</td>
<td>Scatterplot + Color</td>
<td></td>
</tr>
<tr>
<td></td>
<td>datsum 1200 with lowest weight for Weight has Cylinders:4 and Origin:Japan</td>
<td>Find Extremum</td>
<td>Strip plot + Color</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(+Derived Value)</td>
<td>Scatterplot + Color</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C x C x N</td>
<td>Items with Origin: Japan and Cylinders: 4 have lowest AVG(Weight)</td>
<td>Find Extremum</td>
<td>Strip plot + Color</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(+Derived Value)</td>
<td>Scatterplot + Color</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q x Q x C
Horsepower x
Weight x
Origin

Visualization

Facts ¬ Embellishments
$Q \times Q \times C$

Horsepower x Weight x Origin

Visualization

Correlation
Pearson's $r < -0.75$ or $> 0.75$

Distribution
(#points in BIN > 75%)
Overall, Horsepower and Weight have a strong correlation.

Items with Origin: Japan exhibit a strong correlation between Horsepower and Weight.

Items with Origin: US exhibit a strong correlation between Horsepower and Weight.

Most items with Origin: Japan have low Horsepower and low Weight.
• Overall, Horsepower and Weight have a strong correlation

• Items with Origin: Japan exhibit a strong correlation between Horsepower and Weight

• Items with Origin: US exhibit a strong correlation between Horsepower and Weight

• Most items with Origin: Japan have low Horsepower and low Weight
• Overall, Horsepower and Weight have a strong correlation

• Items with Origin: Japan exhibit a strong correlation between Horsepower and Weight

• Items with Origin: US exhibit a strong correlation between Horsepower and Weight

• Most items with Origin: Japan have low Horsepower and low Weight
Overall, Horsepower and Weight have a strong correlation. Items with Origin: Japan exhibit a strong correlation between Horsepower and Weight. Items with Origin: US exhibit a strong correlation between Horsepower and Weight. Most items with Origin: Japan have low Horsepower and low Weight.
What are data facts?

How can we integrate data facts into visualization tools?

Why is this integration beneficial?
User Study

Task: Explore the data and present your findings.
Iterating between Visualizations and Facts during Exploration
Start → V → F → Start
Participant Feedback

• Varied preferences for using facts for interpretation.

“It’s almost like this tool is training me by showing facts based on a visualization. Now I can use this the other way around like if I wanted to show a fact, I know which visualization I need to check.”

(P11 - novice)

“The facts shown were useful given that I didn’t know anything about the dataset. But if this was a type of dataset that I use on a regular basis, I’d want the system to tell me facts specific to the domain of the dataset.”

(P4 - expert)
Participant Feedback

• Some experts (2/4) felt suggestion of alternative visualizations was unnecessary.

“*I feel the system should be smart enough to select the best visualization for a statement automatically. In fact, when we’re working on building some sort of report, I tell my team not to worry about the visualization and always add the default one suggested by the system.*”

(P3 - expert)
Participant Feedback

- All participants liked suggestion of alternative embellishments and highlighting via data facts.

“it was nice to have the system consistently show facts in the visualization by fading things out. Since getting to possible styling options was easy, I could simply go in and format a chart further when I wanted to.”

(P6 - intermediate)
Research Opportunity: Integrating Natural Language Understanding (NLU) and Generation (NLG)

Articulate, Sun et al. 2011

DataTone, Gao et al. 2015

Eviza, Setlur et al. 2016

Evizeon, Hoque et al. 2017

Orko, Srinivasan & Stasko 2017
Research Opportunity: Integrating Natural Language Understanding (NLU) and Generation (NLG)

**Articulate**, Sun et al. 2011

**DataTone**, Gao et al. 2015


**Evizeon**, Hoque et al. 2017

**Orko**, Srinivasan & Stasko 2017
Compare weight and acceleration for cars with different cylinder counts.
Here are some commands you might want to try:

- Is there a group of cars that exhibits a strong correlation between acceleration and weight?
- Highlight heavy but fast cars.
Augmenting visualizations with interactive data facts helps:

- Interpret visualizations and explore alternatives for communication
- Facilitate flexible exploration strategies allowing users to switch between top-down and bottom-up exploration
Augmenting visualizations with interactive data facts helps:

- Interpret visualizations and explore alternatives for communication
- Facilitate flexible exploration strategies allowing users to switch between top-down and bottom-up exploration

Arjun Srinivasan (@10_arjun)
Steven M. Drucker
Alex Endert
John Stasko
Backup slides...
Potential Risks

• Participants skipped visualizations when the system did not show facts (Trust)

• Ability to search for facts and select visualizations and embellishments to show those can be misused (Deception)
## Usage Summary

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visualizations created</td>
<td>86 (4-12 per session)</td>
</tr>
<tr>
<td>Corresponding data facts saved</td>
<td>119 (4-17 data facts per session)</td>
</tr>
<tr>
<td>System generated</td>
<td>102</td>
</tr>
<tr>
<td>Manually entered</td>
<td>17</td>
</tr>
<tr>
<td>Search queries executed</td>
<td>31 (9 sessions, 1-9 per session)</td>
</tr>
</tbody>
</table>
Overall, Horsepower and Weight have a strong correlation

Most items with Origin: Japan have low Horsepower and low Weight

**Attributes:** Horsepower, Weight  
**Tasks:** Correlate

**Attributes:** Horsepower, Weight  
**Tasks:** Characterize Distribution, Filter  
**Value:** Japan
Future Work

• Integrating NLU and NLG
• Integration with partial view specification-based tools
• Recommending exploratory facts and visualizations based on user interest
Visual Data Exploration Tools (w/ Interactive Data Facts)

Voyager: Wongsuphasawat et al. (2016)

ChartAccent: Ren et al. (2017)

VisualQ: Kong et al. (2017)
Static Data Facts: **Pros** and **Cons**

+ Help detect salient facts or confirm inferences
+ Provide richer information about user intent
+ Aid communication or sharing of findings

- Difficult to read facts and mentally map to visualization
- Alternative ways to show facts remain unexplored
How it works?

- US has most items
- Number of items in US is 2.57 times Europe
- US has most items
- Number of items in US is 2.57 times Europe

- Displacement and Horsepower have a strong correlation
- Most items in the dataset have low Displacement and Horsepower

- Stroke, Opacity, Hull, Quadrant Lines, Regression Line
- Stroke, Opacity, Text highlight
- Stroke, Opacity, Text highlight
- Stroke, Opacity, Text highlight
Behavior highlights

• Experts:
  • More V->V than non-experts

• Novices:
  • Used search most
  • Almost no V->V

• Search feature:
  • 3 participants started with search (one in each group).
  • 2 used it more after they tried it once.
  • Results were clicked 18/31 times (58%)