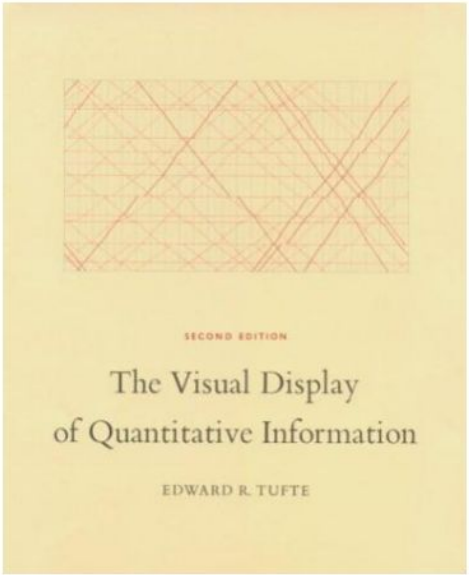
Data Visualization

Lei Xiao, PhD, PENG
Data Science & Analytics Engineer @Husky Energy

Must Read on Visualization

Tufte 1: *The Visual Display of Quantitative Information, 2e*



Tufte, E. R. (2001). *The Visual Display of Quantitative Information.* Cheshire, CT, USA: Graphics Press.
http://www.edwardtufte.com/tufte/books_vdqi

The ability to take data–to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it–that's going to be a hugely important skill in the next decades, … because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

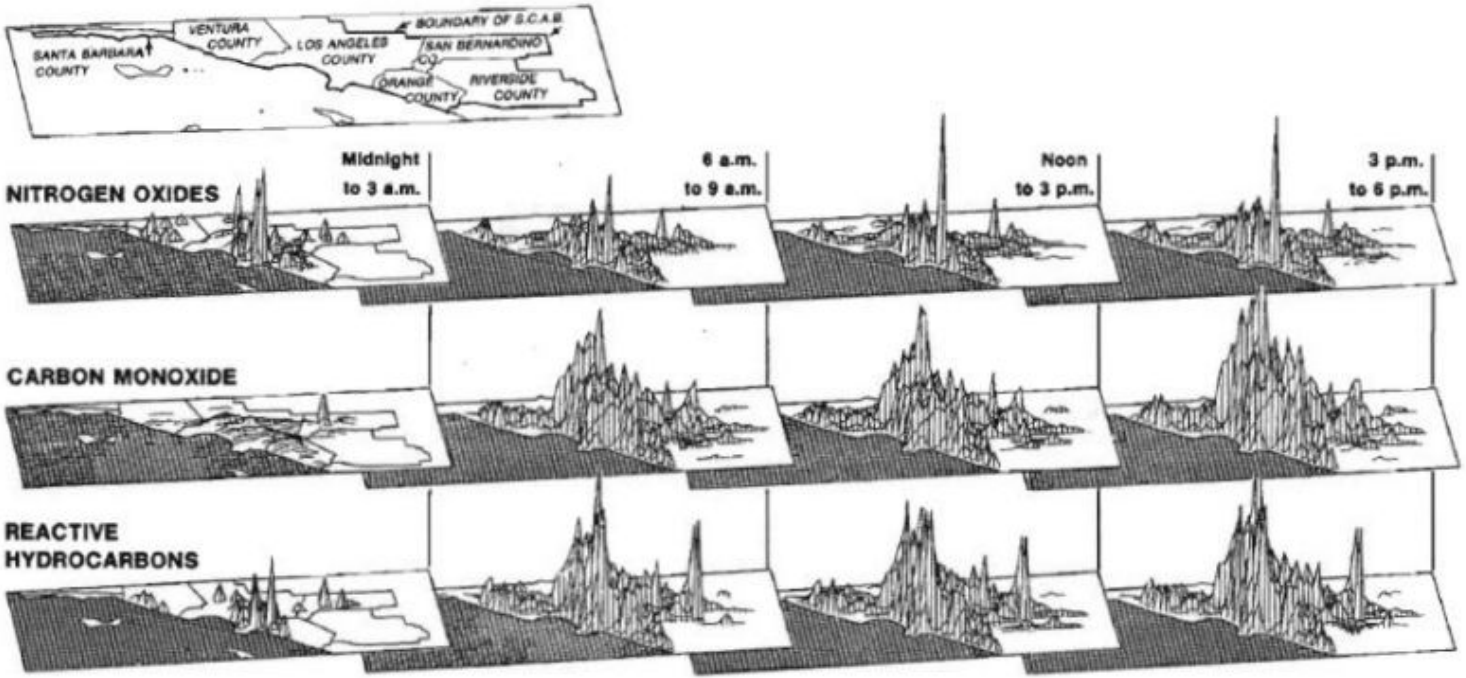
Good Visualization Should Do's

1. Show the data
2. Induce the reader to think about the *substance* rather than about
   - Methodology
   - Graphic design
   - Technology of graphic production
   - Something else
3. Avoid distorting what the data have to say
4. Present many numbers in a small space
5. Make large data sets coherent
6. Encourage the eye to compare different pieces of data
7. Reveal the data at different levels of detail, broad to fine
8. Serve a clear purpose: description, evaluation, tabulation, decoration
9. Be closely integrated with statistical and verbal descriptions of data

Data in Context: Cholera Outbreak

Used map to hypothesize that pump on Broad St. was the cause. [from Tufte 83]

Preview:
Small Multiple – Air Pollution Map

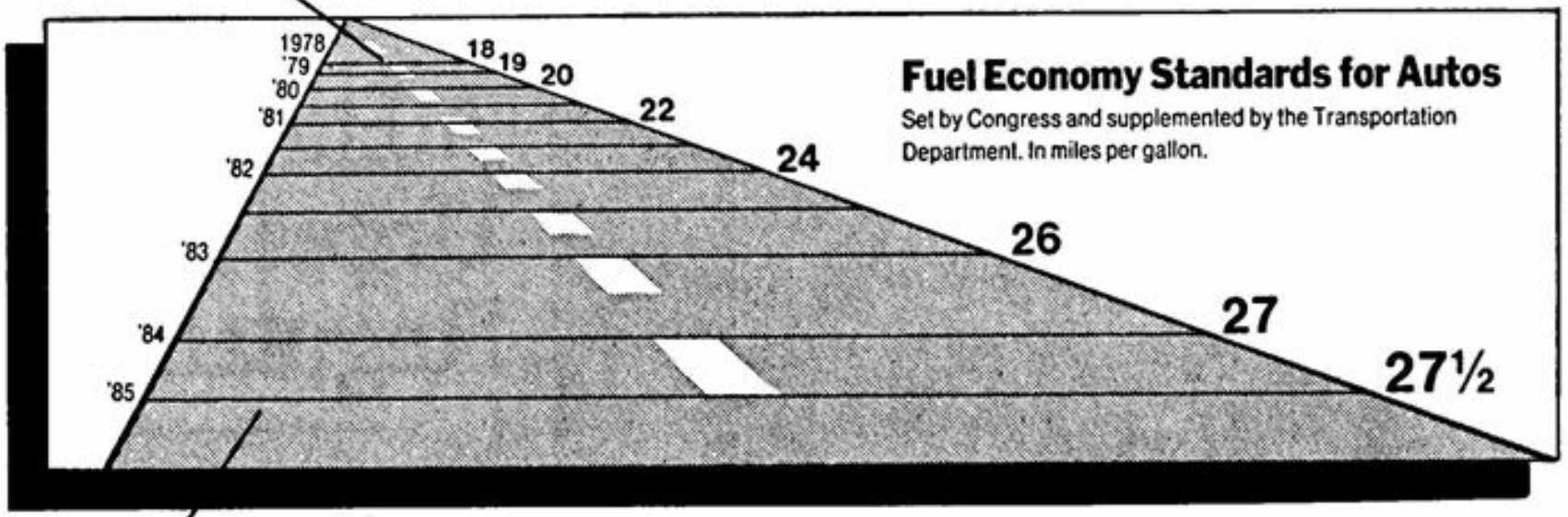


p. 42 (Tufte 1 *aka* VDQI, 1e & 2e)

Graphic Integrity Principles

- See also *How to Lie With Statistics* (Huff, 1984): http://bit.ly/3wAgS0
- 1. Make Representation of Numbers Proportional to Quantities
  - Ratio of size to numerical value should be close to 1
  - As physically measured on surface of graphic
- 2. Use Clear, Detailed, Thorough Labeling
  - Don't introduce or propagate graphical distortion, ambiguity
  - Write out explanations of the data *on the graphic itself*
  - Label important events in the data
- 3. Show Data Variation, Not Design Variation
- 4. Use Standardized (*e.g.*, Inflation-Adjusted) Units, Not Nominal
- 5. Depict *N* Data Dimensions with ≤ *N* Variable Dimensions
  - Don't use more than *N* information-carrying dimensions for *N*-D data
  - When graphing data in *N*-D, use *N*-D ratio (see #1 above)
- 6. Quote Data in Full Context (Don't Quote Out of Context)

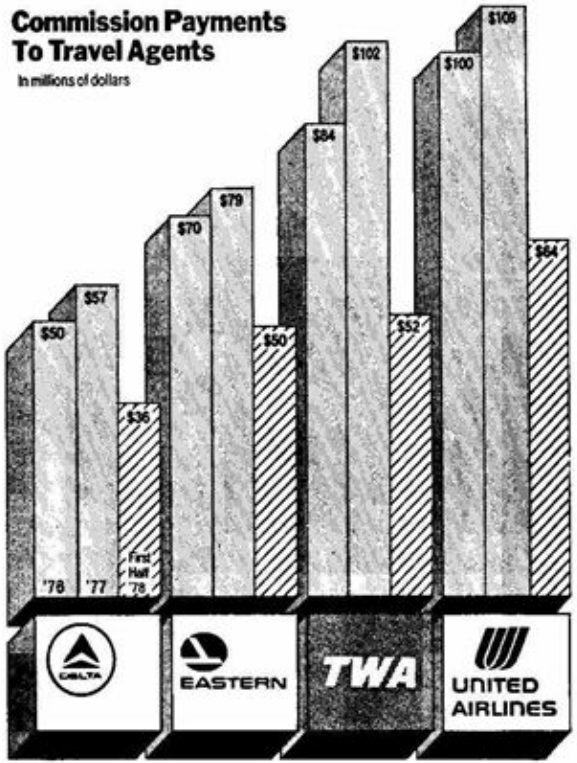
This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

New York Times, August 9, 1978, p. D-2.

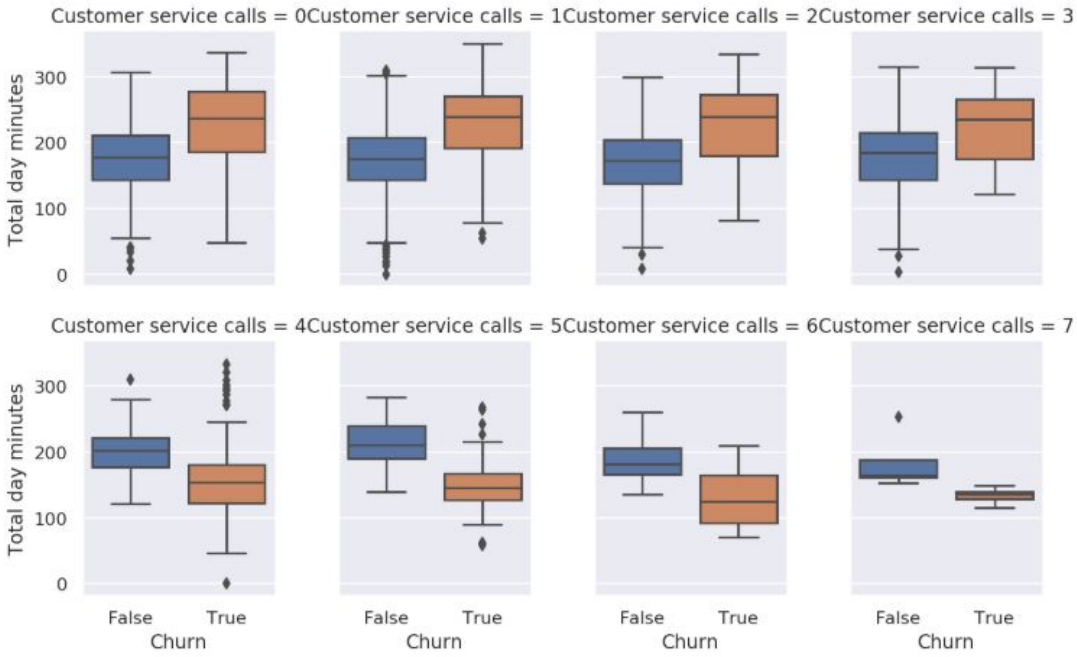
This pseudo-decline was created by comparing six months' worth of payments in 1978 to a full year's worth in 1976 and 1977, with the lie repeated four times over.

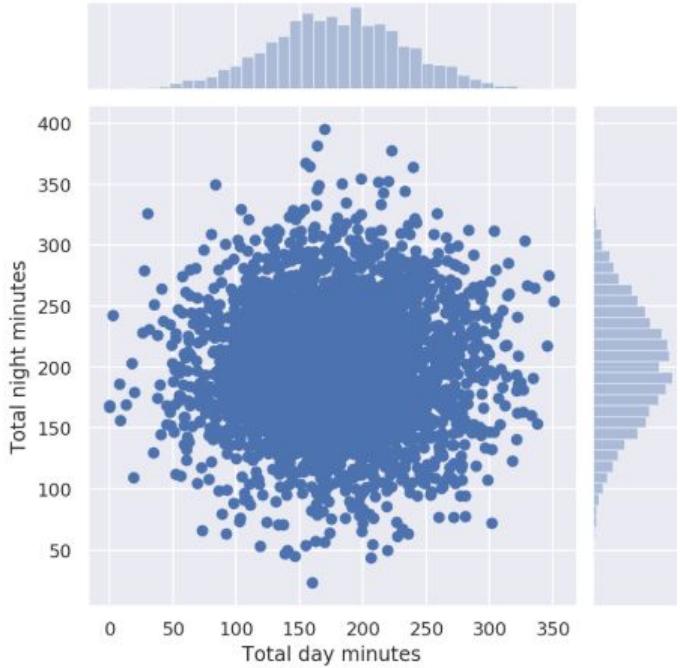


New York Times, August 8, 1978, p. D-1.

Multivariate Visualization

Multivariate plots allow us to see relationships between two and more different variables, all in one figure. Just as in the case of univariate plots, the specific type of visualization will depend on the types of the variables being analyzed.

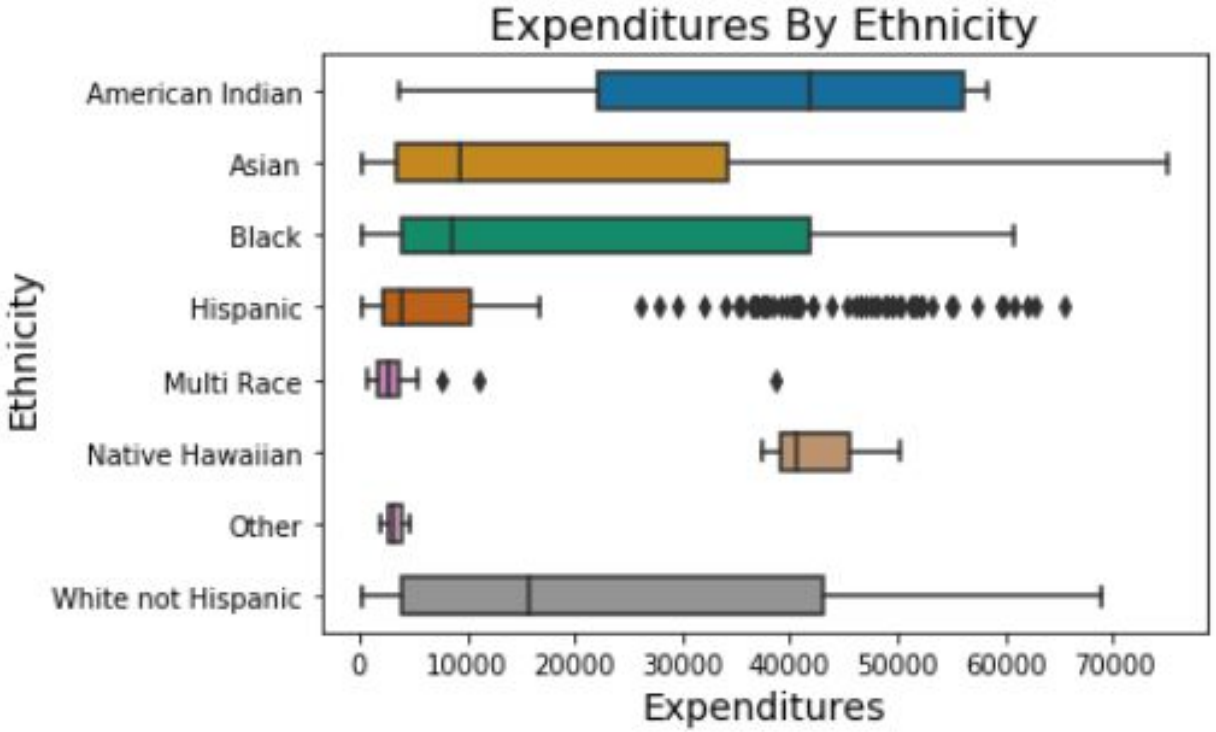

Customer service calls = 0
Customer service calls = 1
Customer service calls = 2
Customer service calls = 3
Customer service calls = 4
Customer service calls = 5
Customer service calls = 6
Customer service calls = 7
Total day minutes
0
100
200
300
False
True
Churn


Total night minutes
Total day minutes
0
50
100
150
200
250
300
350
400

Simpson's Paradox

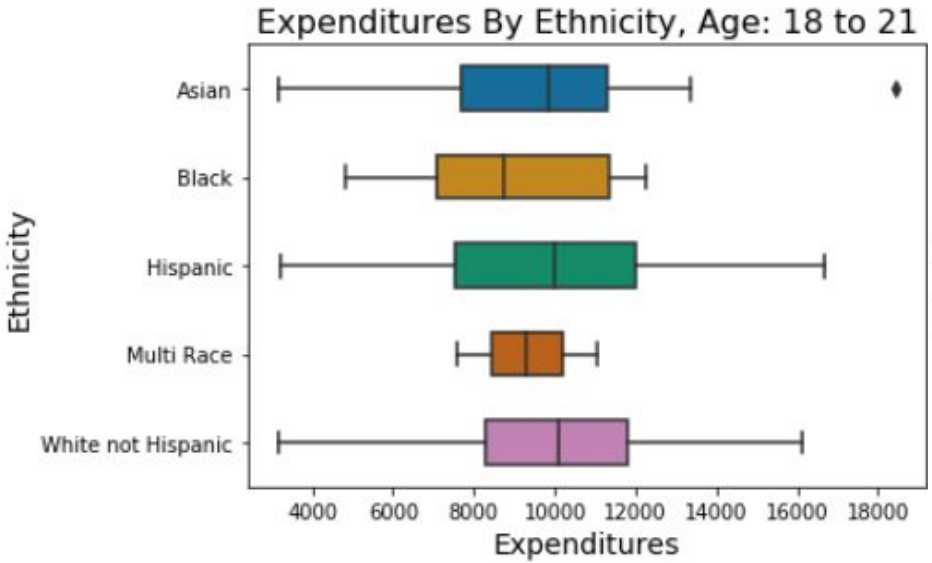
It is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined. In other words, Simpson's Paradox occurs when groups of data show on a particular trend; however this trend is reversed when the groups are combined together.

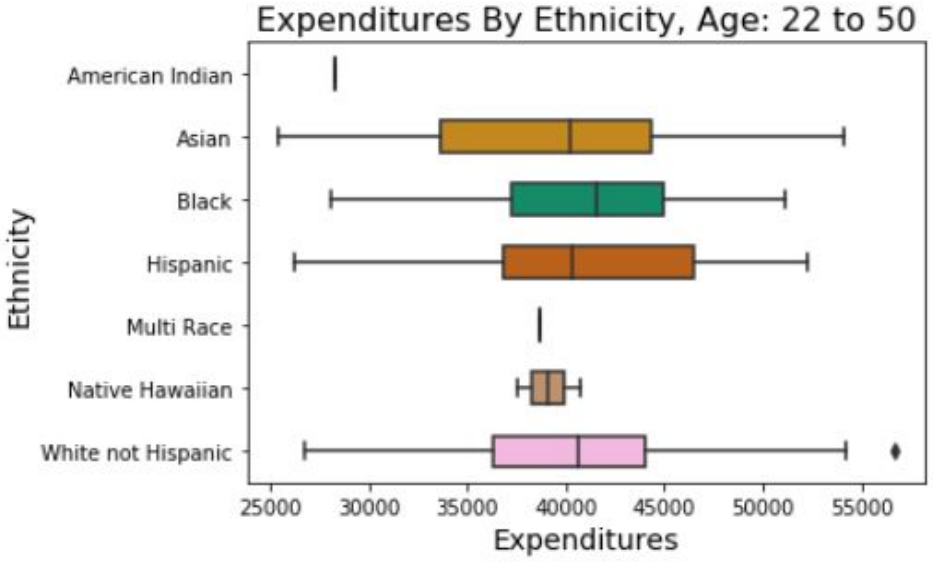
Stats Considering All Age Group



Hispanic spends much less than White not Hispanic

Stats in Various Age Groups



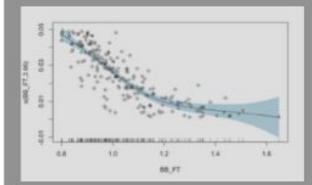


Hispanic spends as much as White not Hispanic

Model Interpretability

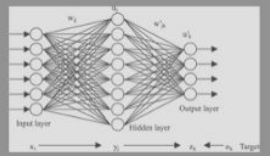
Interpretable Model

- Those models which are inherently explainable
- Typically have lower accuracy
- Simple in computation, one might even argue that they are elegant.
- For eg Linear Regression, Logistic Regression, GLM, GAM, Decision Tree etc



Black Box Model

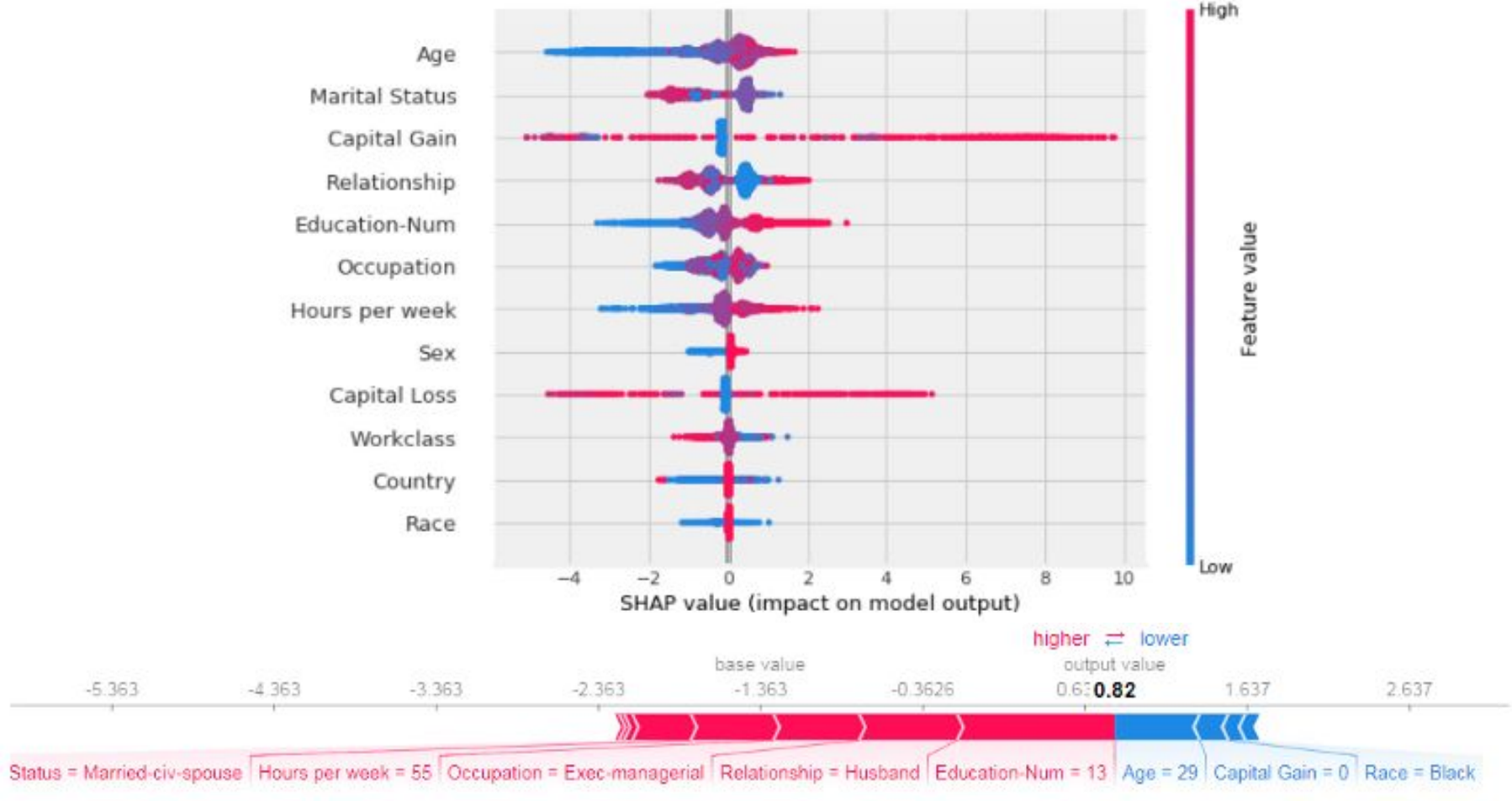
- Models whose internal complexities render their internal interactions unexplainable
- Typically have higher accuracy
- Complex computation.
- For eg Neural Networks, Gradient Boosting Machines etc



Global: Explaining the entire model as a whole;
Local: Explaining individual predictions
Per Feature: How does a feature behave w.r.t the model predictions or the actual target variable

SHAP @ Global & Local

SHapely **A**dditive ex**P**lanations Values, motivated by the work done by Lloyd Shapley(1953) in Game Theory. The Shapley value is the average marginal contribution of a feature value across all possible coalitions.


Age
Marital Status
Capital Gain
Relationship
Education-Num
Occupation
Hours per week
Sex
Capital Loss
Workclass
Country
Race
−4
−2
0
2
4
6
8
10
SHAP value (impact on model output)
High
Feature value
Low
higher ⇄ lower
base value
output value
-5.363
-4.363
-3.363
-2.363
-1.363
-0.3626
0.82
1.637
2.637
Status = Married-civ-spouse
Hours per week = 55
Occupation = Exec-managerial
Relationship = Husband
Education-Num = 13
Age = 29
Capital Gain = 0
Race = Black

“ Talk is cheap. Show me the code. ”

Linus Torvalds

References

http://www.kdd.cs.ksu.edu/Courses/CIS736/Lectures/Slides/Lecture-34-Main_6up.pdf

https://courses.cs.washington.edu/courses/cse512/18sp/

https://www.kaggle.com/saicataram/simpson-s-paradox-in-python

https://github.com/Yorko/mlcourse.ai

https://medium.com/@ag.ds.bubble/model-interpretability-a4244d82ffb2

https://towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608