Types of data science questions

The Data Scientist’s Toolbox

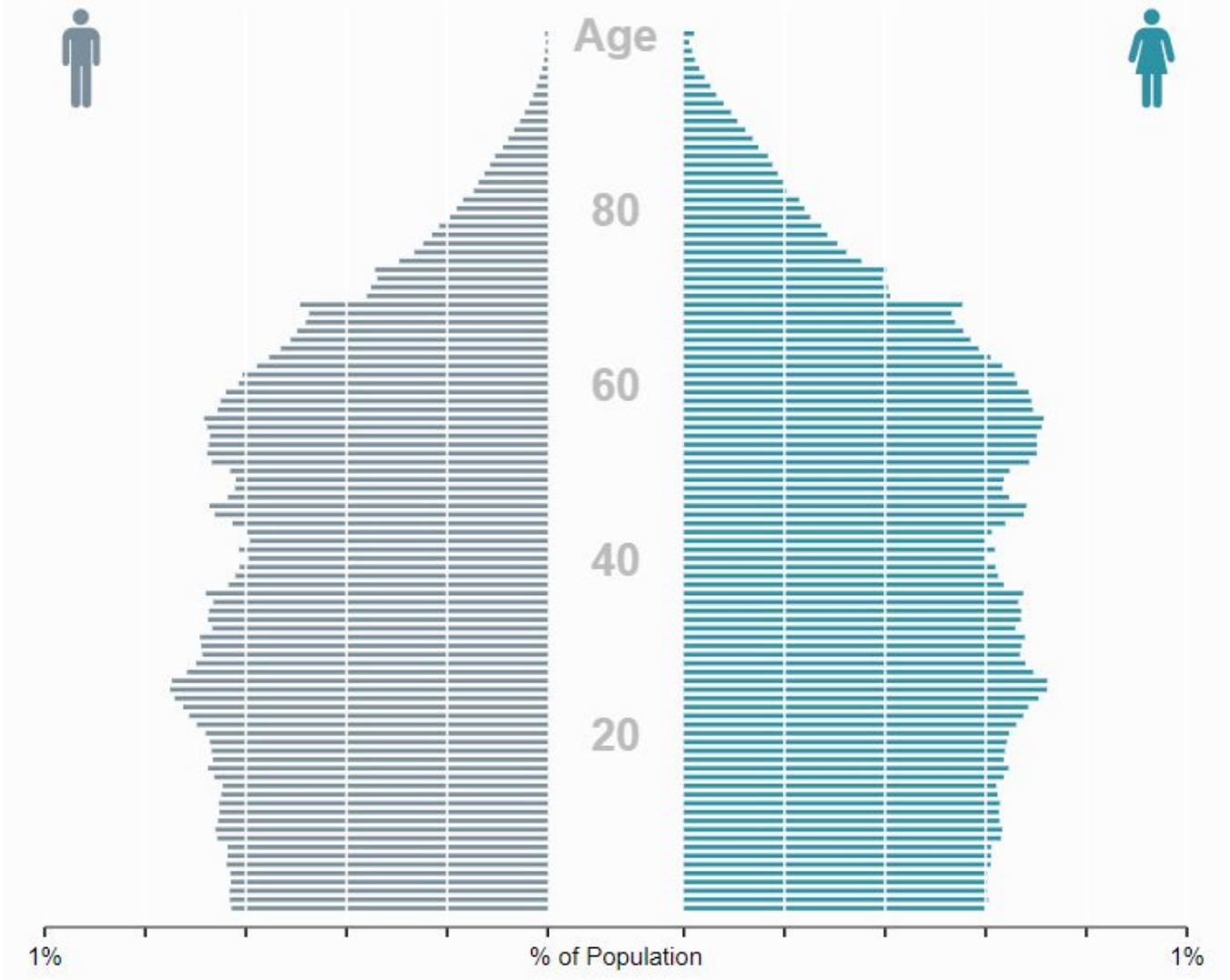
Types of data analysis

1) Descriptive
2) Exploratory
3) Inferential
4) Predictive
5) Causal
6) Mechanistic

Descriptive analysis

Goal: Describe or summarize a set of data

- Early analysis when receive new data
- Generate simple summaries about the samples and their measurements
  - Eg: measures of central tendency or measures of variability
- NOT for generalizing the results of the analysis to a larger population or trying to make conclusions



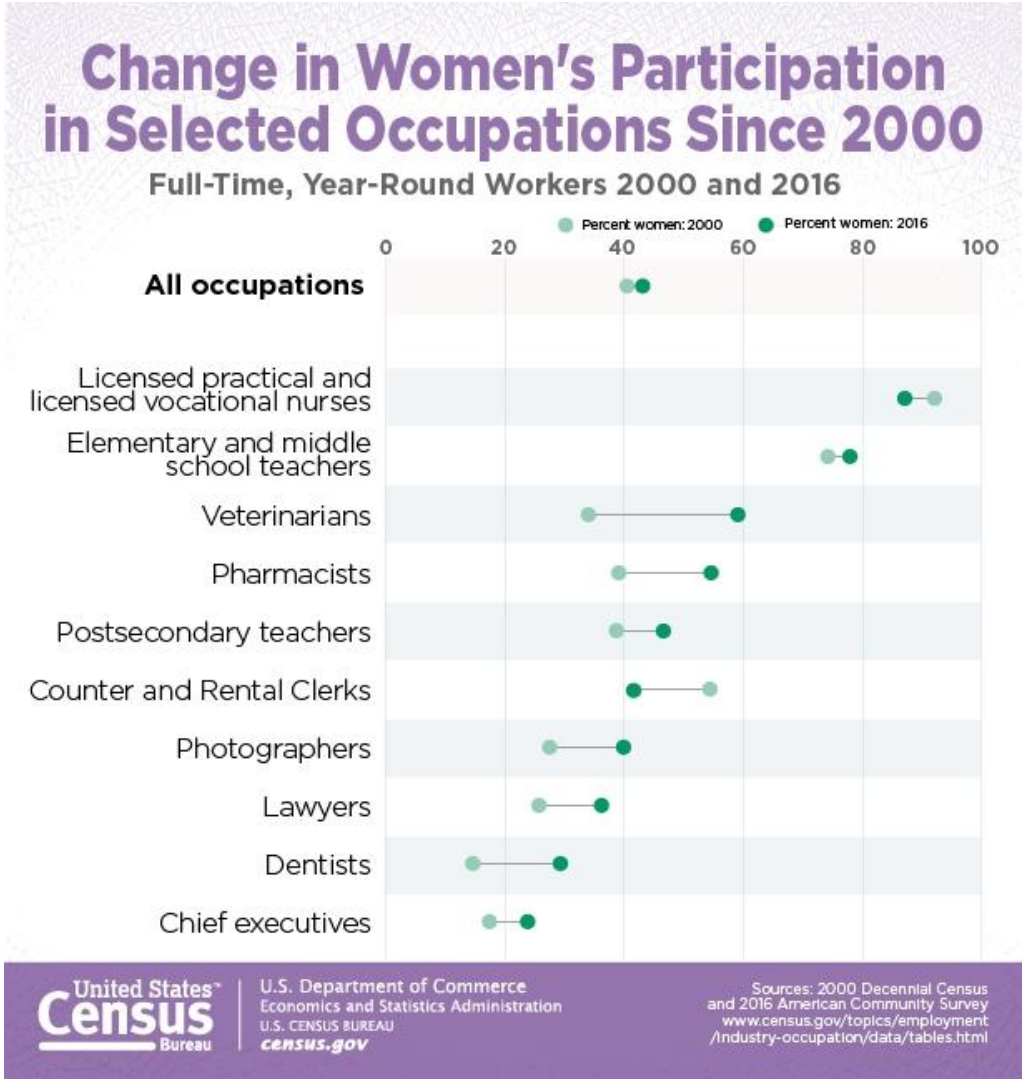
https://www.census.gov/popclock/

Exploratory analysis

Goal: Examine the data and find relationships that weren't previously known

- Explore how different variables might be related
- Useful for discovering new connections
- Help to formulate hypotheses and drive the design of future studies and data collection

Correlation does not imply causation!



https://www.census.gov/library/visualizations/2018/comm/women-occupations.html

Inferential analysis

Goal: Use a relatively small sample of data to say something about the population at large

- Provide your estimate of the variable for the population and provide your uncertainty about your estimate
- Ability to accurately infer information about the larger population depends heavily on sampling scheme

Epidemiology. 24(1):23–31, JAN 2013
DOI: 10.1097/EDE.0b013e3182770237,
, PMID: 23211349
Issn Print: 1044-3983
Publication Date: 2013/01/01

Share Print

Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 U.S. Counties for the Period from 2000 to 2007

Andrew W. Correia; C. Arden Pope; Douglas W. Dockery; Yun Wang; Majid Ezzati; Francesca Dominici

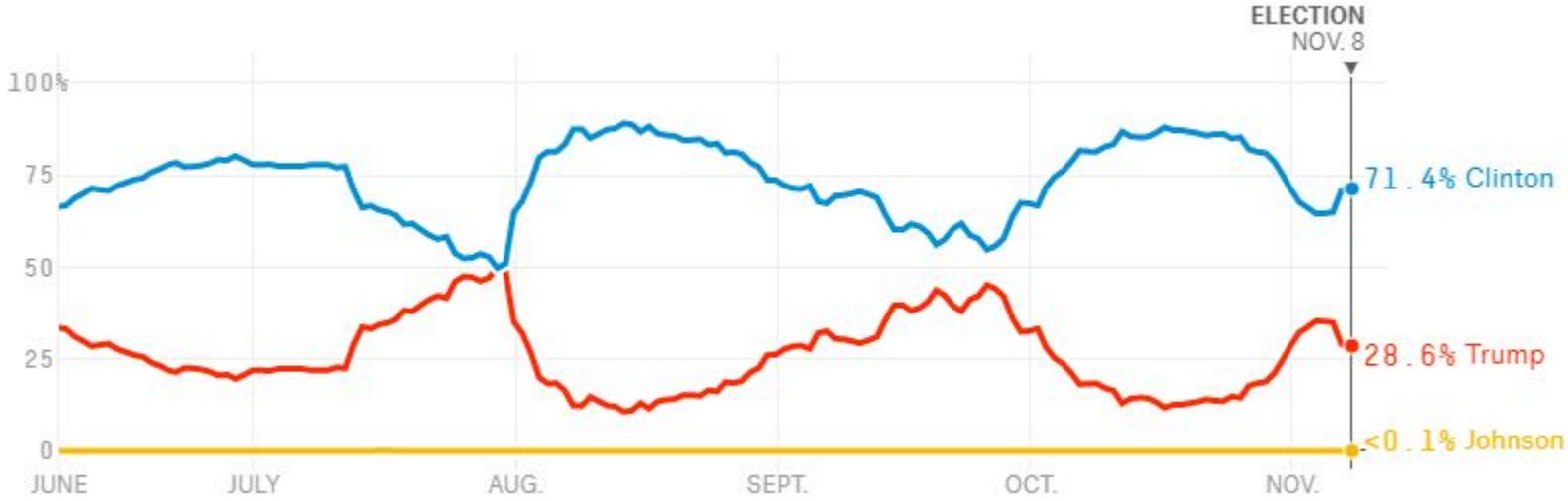
Predictive analysis

Goal: Use current and historical data to make predictions about future data

- Accuracy in predictions is dependent on measuring the right variables
- Many ways to build up prediction models with some being better or worse for specific cases,
  - More data and a simple model generally performs well at predicting future outcomes

Just because one variable may predict another, it does not mean that one causes the other

FiveThirtyEight's predictions of the 2016 US election



https://projects.fivethirtyeight.com/2016-election-forecast/

Causal analysis

Goal: See what happens to one variable when we manipulate another variable

- Gold standard in data analysis
- Often applied to the results of randomized studies that were designed to identify causation
- Usually analysed in aggregate and observed relationships are usually average effects

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy

R.S. Finkel, E. Mercuri, B.T. Darras, A.M. Connolly, N.L. Kuntz, J. Kirschner, C.A. Chiriboga, K. Saito, L. Servais, E. Tizzano, H. Topaloglu, M. Tulinius, J. Montes, A.M. Glanzman, K. Bishop, Z.J. Zhong, S. Gheuens, C.F. Bennett, E. Schneider, W. Farwell, and D.C. De Vivo, for the ENDEAR Study Group*

http://www.nejm.org/doi/full/10.1056/NEJMoa1702752

Mechanistic analysis

Goal: Understand the exact changes in variables that lead to exact changes in other variables

- Applied to simple situations or those that are nicely modeled by deterministic equations
- Commonly applied to physical or engineering sciences
  - Eg: Biological sciences, are far too noisy to use mechanistic analysis
- Often, the only noise in the data is measurement error

Polymer Testing

Volume 61, August 2017, Pages 364-372

Compatibilization of toughened polypropylene/biocarbon biocomposites: A full factorial design optimization of mechanical properties

Ehsan Behazin [a, b], Manjusri Misra [a, b], Amar K. Mohanty [a, b]

https://www.sciencedirect.com/science/article/pii/S0142941817303422

Summarizing: Types of data science questions

The Data Scientist’s Toolbox