Finding Good Datasets

Find Data you Can Live With

- It doesn’t have to be perfect

- It’s better to have GOOD, USEFUL data than bad data, even it’s not quite on the right topic

- It’s better to choose data relatively quickly!

Most of You will Use Internet Data

- Kaggle
- Google Dataset Search
- Machine Learning Repo
- Data.gov
- Data World

Data Considerations

Data Wrangling

- Do you like wrangling?
- How much are you willing to do?
- Is data primarily text?
- Does data need recoding?
- Is data “messy?”
- Is there a lot of missing data?
- Do you need to generate new features / columns to make the best use of the data?

Amount of Data

- Too little?
  - Can you meet all sample size assumptions for your chosen analysis?
  - Maximum sample size requirement is 200
  - Do you have enough columns to report something interesting if one analysis has no significant results?

- Too much?
  - Will you bog down your computer using standard programs like R and Python?
  - Would you need to pay for commodity computing?
  - No more than 500,000 rows typically

Stipulations for Use

- Do you need to pay a fee?
- Do you need to report your findings?
- Do you need to anonymize the data?
- Do you need to reference the data source?

Project Uniqueness

- When possible it is better to be unique
- But unique can be difficult!
- Avoid built-in R / Python / SQL datasets
- Most things on Kaggle, etc. should be ok

Do a Kaggle Data Search…