Extra Review

UNIVERSITY OF
SAN FRANCISCO

Abbie M Popa

BSDS 100 - Intro to Data Science with R

Outline

- Something Old
  - Tidy Data
  - Suggestions for extra practice (functions, plotting, string analysis)
- Something New
  - Web-scraping
  - Machine Learning
- A wrap-up activity
- Course review

Request for Extra Review

Tidy Data

- Two Approaches:
  1. Formal Definition
  2. Practical Definition

Tidy Data - The Shelter Animal Example

species	wgt_admitted	wgt_adopted	supp
"cat"	8	15	"turkey"
"cat"	9	11	"kibbles"
"dog"	18	27	"turkey"

These data are not tidy, let's find out why!

Tidy Data - Formal Definition Approach

Jeff Leek, *The Elements of Data Analytic Style*

1. Each variable should be one column.
2. Each different observation should be a different row.
3. There should be one table for each "kind" of variable.
4. If you have multiple tables, they should include a column that allows them to be linked

Tidy Animals - Formal Definition Approach

species	wgt_admitted	wgt_adopted	supp
"cat"	8	15	"turkey"
"cat"	9	11	"kibbles"
"dog"	18	27	"turkey"

1. Is each variable one column?

Tidy Animals - Formal Definition Approach

species	wgt_admitted	wgt_adopted	supp
"cat"	8	15	"turkey"
"cat"	9	11	"kibbles"
"dog"	18	27	"turkey"

1. Is each variable one column? No! Weight is split across two columns, so is time of measurement
2. Is each observation one row?

Tidy Animals - Formal Definition Approach

species	wgt_admitted	wgt_adopted	supp
"cat"	8	15	"turkey"
"cat"	9	11	"kibbles"
"dog"	18	27	"turkey"

1. Is each variable one column? No! Weight is split across two columns, so is time of measurement
2. Is each observation one row? No! Each row contains two observations of the animals weight
3. Is there one table for each type of variable?

Tidy Animals - Formal Definition Approach

species	wgt_admitted	wgt_adopted	supp
"cat"	8	15	"turkey"
"cat"	9	11	"kibbles"
"dog"	18	27	"turkey"

1. Is each variable one column? No! Weight is split across two columns, so is time of measurement
2. Is each observation one row? No! Each row contains two observations of the animals weight
3. Is there one table for each type of variable? n/a, only one table
4. If there are multiple tables, is their a linking column?

Tidy Animals - Formal Definition Approach

species	wgt_admitted	wgt_adopted	supp
"cat"	8	15	"turkey"
"cat"	9	11	"kibbles"
"dog"	18	27	"turkey"

1. Is each variable one column? No! Weight is split across two columns, so is time of measurement
2. Is each observation one row? No! Each row contains two observations of the animals weight
3. Is there one table for each type of variable? n/a, only one table
4. If there are multiple tables, is their a linking column? n/a, only one table

Note on multiple tables...

- There are at least three approaches to how to decide how many tables to split data across
- This will be covered in more detail in the databases class, CS 333

Tidy Data - Practical Approach

Format the data so that you can complete the goal you have in mind

- I want to make a plot showing what influences change in weight over time
- I want to formally test what affects weight using a linear model

Extra Practice

- Functions:
  - https://r4ds.had.co.nz/index.html chapter 19
  - https://exercism.io/tracks/r - requires login, but free
  - https://www.hackerrank.com - requires login, but free, select `R` from dropdown
- Plotting:
  - https://r4ds.had.co.nz/index.html chapter 3
  - http://www.cookbook-r.com/Graphs/ type their code, try on your own data
- String Analysis:
  - https://r4ds.had.co.nz/index.html chapter 14
  - https://regexr.com/, just remember, `R` requires the double backslash where most languages use the single backslash!

Extra Practice

But remember, the best way to practice R (or any programming language) is to complete projects!

- Enter a kaggle competition https://www.kaggle.com/
- Enter a driven data competition https://www.drivendata.org/
- Solve your own problem with data! Then share it...

  https://github.com/

  https://medium.com/

  Your personal website!
- Examples:
  - https://pudding.cool/
  - https://flowingdata.com/
  - https://towardsdatascience.com/

New Topics (Very Brief)

Web Scraping

- Types:
  - Some websites have an API, where you can directly request data be served to you in a convenient format (e.g., CSV)
  - Most websites are written in `html`, a structured way to format and display text
  - Some website rely heavily on `javascript`, which is code that executes in your browser
- We will focus on the second, which is common and doable

Web Scraping

- Examine the web page using the browser's developer tools
- Import and parse the webpage using the `rvest` package
- Example, build a corpus of legal terms to see how many lawyers are using our web app

Web Scraping

- Tutorial

  https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/

- Major sticking points:
  - Javascript (requires Selenium)
  - If they change their webpage, you'll need to update!
- Data Mining Class: CS 451

Machine Learning

- Goal: Use a collected data set, to build a model, that can be used to make predictions about new data
- Many approaches all with different pros and cons (no free lunch!)
- A linear regression example

Machine Learning

- Machine Learning is a huge topic!
- Several Types
  - Supervised Learning - data has labels
  - Unsupervised Learning - data does not have labels
  - Regression - Predicted Value is continuous
  - Classification - Predicted Value is categorical
- Different levels of flexibility
  - Linear regression is not very flexible, the output model is always a straight line
  - Other options, such as decision trees, are more flexible

Machine Learning

- Workflow
  1. Split data into train and test sets
  2. Train data on the training set
  3. Check accuracy on test set
  4. Try with different algorithms or hyperparameters
- Considerations
  - More flexible models have a higher risk of "overfitting", that is, they will work well on the training data but not new data
  - This is why with more flexible models it is important to have a larger training dataset

Machine Learning

- Popular Algorithms
  - Linear Regression (seen earlier), not very flexible, but works well with small, normal, data
  - Random Forests, get a lot of flexibility but use some tricks to avoid overfitting
  - Artificial Neural Networks, self-engineer features, often produce state of the art results, but you need more data

Machine Learning

- Free online: https://www-bcf.usc.edu/ gareth/ISL/
- Not in R, but beloved
  https://www.coursera.org/learn/machine-learning
- Class at USF: MATH 373

Wrap-Up

Wrap-Up Activity

- In groups, complete the DS in the Wild Song Lyrics activity from github
- We will vote on the quality of their report together

Course Evals

You rate your Uber driver, why not rate your classes?

Thank you!