DSCI354-451 Practicum: Question, Tidy, Check, Explore

Roger H. French, JiQi Liu

07 October, 2018

Contents

8.1.2.1 Reading, Homeworks, Projects, SemProjects . . . 1
8.1.2.2 Textbooks . . . 2
8.1.2.3 Syllabus . . . 2
8.1.2.4 Some Tidy Data Analysis Resources . . . 2
8.1.2.4.1 Elements of Data Analytic Style; Jeff Leak . . . 2
8.1.2.4.2 R for Data Science . . . 2
8.1.2.4.3 What is Tidy Data . . . 2
8.1.2.5 Data Analysis Question, Tidying, Checking, Exploratory Data Analysis . . . 2
8.1.2.5.1 Answering the question . . . 2
8.1.2.5.2 The Data Analysis Flow Chart . . . 4
8.1.2.5.3 Common Mistakes . . . 4
8.1.2.6 Tidying the data . . . 7
8.1.2.6.1 These are the components of a processed data set: . . . 7
8.1.2.6.2 Raw data: It is critical that you include the rawest form of the data that you have access to. . . . 7
8.1.2.6.3 Tidy data: . . . 7
8.1.2.6.4 The code (or data) book: . . . 8
8.1.2.6.5 The instruction list or script must be explicit . . . 8
8.1.2.6.6 The ideal instruction list is a script . . . 8
8.1.2.6.7 If there is no script, . . . 8
8.1.2.6.8 Common Mistakes . . . 9
8.1.2.7 Checking the data . . . 9
8.1.2.7.1 How to code variables . . . 9
8.1.2.7.2 In the code book you should explain why censored values are missing. 9
8.1.2.7.3 Avoid coding categorical or ordinal variables as numbers. . . . 9
8.1.2.7.4 Always encode every piece of information about your observations using text. . . . 9
8.1.2.7.5 Identify the missing value indicator . . . 10
8.1.2.7.6 Check for clear coding errors . . . 10
8.1.2.7.7 Check for label switching . . . 10
8.1.2.7.8 If you have data in multiple files, . . . 10
8.1.2.7.9 Check the units (or lack of units) . . . 10
8.1.2.7.10 Common Mistakes . . . 10
8.1.2.8 Cee-lo a good, no house advantage game . . . 10
8.1.2.9 Cee-lo dice game . . . 10
8.1.2.9.1 Cee-lo without a bank (winner take all) . . . 10
8.1.2.9.2 The combinations in Cee-lo . . . 11
8.1.2.9.3 Probabilities[edit] . . . 11
8.1.2.9.4 dice, an R package to calculate dice games . . . 11
8.1.2.9.5 Rolling the Dice on a Warm Night . . . 11

8.1.2.1 Reading, Homeworks, Projects, SemProjects

- Readings:
  - OIS4 for today, Foundations of Inference
  - OIS5 1-4 Numerical Inference for Thursday
- Homeworks
  - HW4 due today
  - HW5 in our repo Thursday
- Data Science Projects:
  -
- 451 SemProjects:
  -
- Friday Comm. Hour
  -

8.1.2.2 Textbooks

- Peng: R Programming for Data Science
- Peng: Exploratory Data Analysis with R
- Open Intro Stats, v3
- Wickham: R for Data Science
- Hastie: Intro to Statistical Learning with R

8.1.2.3 Syllabus

8.1.2.4 Some Tidy Data Analysis Resources

8.1.2.4.1 Elements of Data Analytic Style; Jeff Leak

- https://leanpub.com/datastyle

8.1.2.4.2 R for Data Science

- By Garrett Grolemund, Hadley Wickham
- http://r4ds.had.co.nz/

8.1.2.4.3 What is Tidy Data

- A Wickham paper in your readings subdirectory
- http://vita.had.co.nz/papers/tidy-data.pdf

8.1.2.5 Data Analysis Question, Tidying, Checking, Exploratory Data Analysis

8.1.2.5.1 Answering the question

1. Did you specify the type of data analytic question

```
- (e.g. exploration, assocation causality)
- before touching the data?
```

So here is another Tukey quote.

The data may not contain the answer.

The combination of some data and an aching desire for an answer

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1,**
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

Figure 1: DSCI351-451 Syllabus

```
- does not ensure that a reasonable answer can be extracted
- from a given body of data.
```

John Tukey

8.1.2.5.2 The Data Analysis Flow Chart

1. Types of Data Analyses

- Descriptive:
  - A descriptive data analysis seeks to summarize the measurements
  - in a single data set without further interpretation.
- Exploratory:
  - An exploratory data analysis builds on a descriptive analysis
  - by searching for discoveries, trends, correlations,
  - or relationships between the measurements of multiple variables
    - to generate ideas or hypotheses.
- Inferential:
  - An inferential data analysis goes beyond an exploratory analysis
  - by quantifying whether an observed pattern will likely hold
    - beyond the data set in hand.
  - Inferential data analyses are the most common statistical analysis
    - in the formal scientific literature.
- Predictive:
  - While an inferential data analysis quantifies
  - The relationships among measurements at population-scale,
  - a predictive data analysis uses a subset of measurements (the features)
  - to predict another measurement (the outcome) on a single person or unit.
- Causal:
  - A causal data analysis seeks to find out what happens to one measurement
  - if you make another measurement change.
- Mechanistic:
  - Causal data analyses seek to identify average effects
    - between often noisy variables.
  - For example, decades of data
    - show a clear causal relationship between smoking and cancer.

2. Did you define the metric for success before beginning?
3. Did you understand the context for the question and the scientific or business application?
4. Did you record the experimental design?
5. Did you consider whether the question could be answered with the available data?

8.1.2.5.3 Common Mistakes

- Correlation does not imply causation:
  - Interpreting an inferential analysis as causal.

Most data analyses involve inference or prediction.

- Unless a randomized study is performed,
- it is difficult to infer from The data analytic question
- if there is a relationship between two variables.

A great website to hunt for spurious correlations is

- http://tylervigen.com

Particular caution should be used when applying words

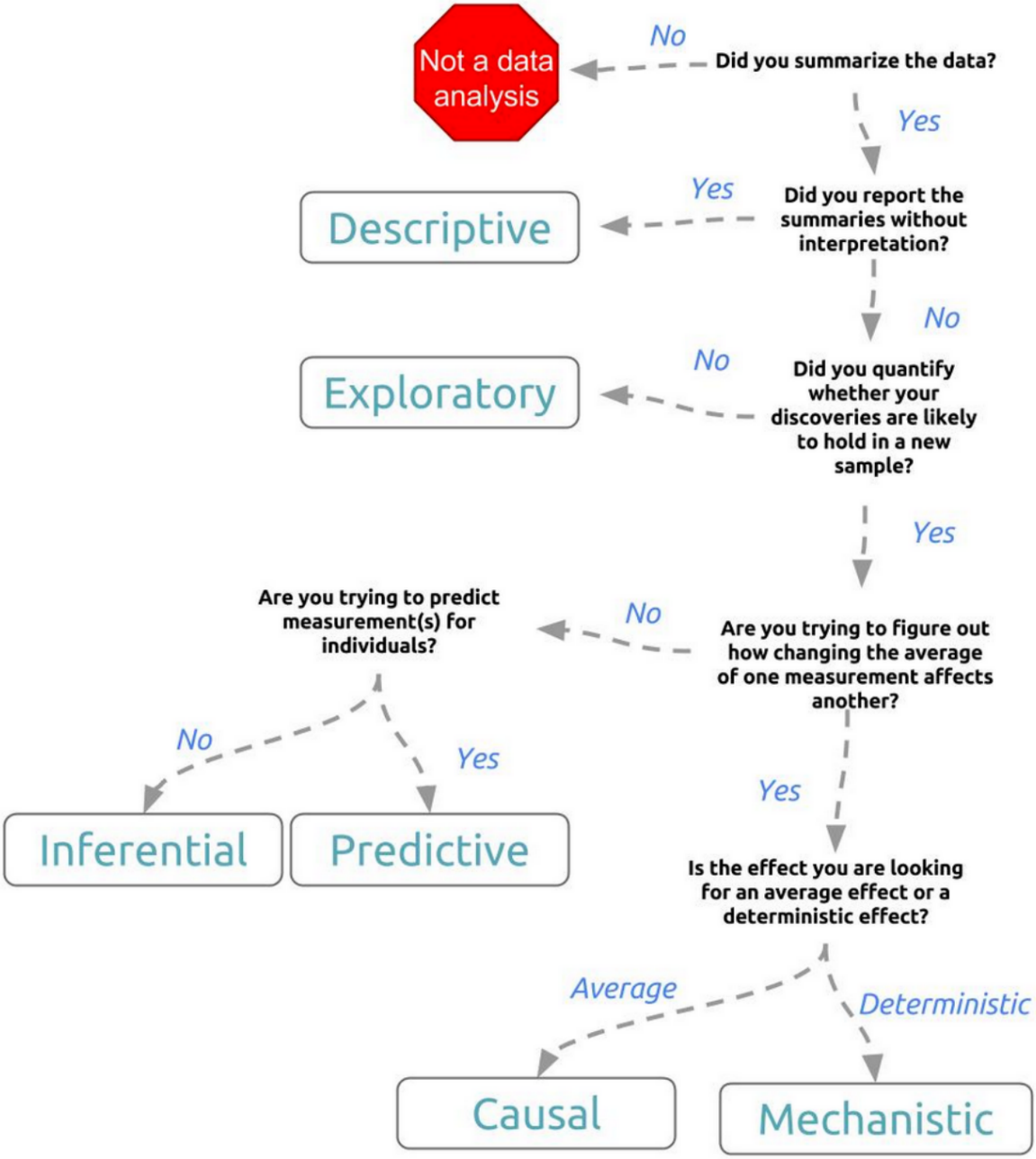

Figure 2.1 The data analysis question type flow chart

Figure 2: the da question flow chart

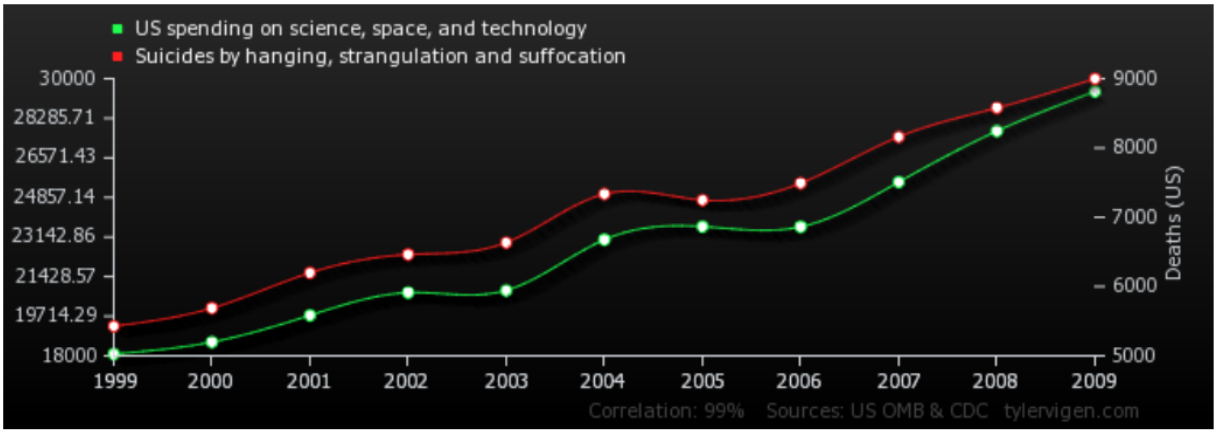

Figure 2.2 A spurious correlation

Figure 3: spurious correlations

- such as “cause” and “effect” when performing inferential analysis.

Inference is not about causation; its infering relationships between variables.

- Overfitting: Interpreting an exploratory analysis as predictive

A common mistake is to use a single, unsplit data set

- for both model building and testing.
- If you apply a predictive model
  - to the same data set used to build the model
  - you can only estimate “resubstitution error” or “training set error”.
- These estimates are very optimistic (Not Good) estimates of the error - you would get if using the model in practice.

If you try enough models on the same set of data,

- you eventually can predict perfectly.
- but this is useless

For a predictive model;

- you need to split your data into training and test datasets,
- and evaluate how well it predicts the test dataset.
- n of 1 analysis: Descriptive versus inferential analysis.

When you have a very small sample size,

- it is often impossible to explore the data,
  - let alone make inference to a larger population.
- Data dredging: Interpreting an exploratory analysis as inferential

Similar to the idea of overfitting,

- if you fit a large number of models to a data set,
  - it is generally possible to identify at least one model
  - that will fit the observed data very well.

As Ronald Coase said:

- "If you torture the data enough, nature will always confess."

8.1.2.6 Tidying the data

The point of creating a tidy data set

- is to get the data into a format
  - that can be easily shared, computed on, and analyzed.
- The components of a data set:

The work of converting the data from raw form

- to directly analyzable form
  - is the first step of any data analysis.
- It is important to see the raw data,
  - understand the steps in the data processing pipeline,
  - and be able to incorporate hidden sources of variability
  - in one's data analysis.

On the other hand, for many data types,

0 the processing steps are well documented and standardized.

8.1.2.6.1 These are the components of a processed data set:

- The raw data.
- A tidy data set.
- A code(data) book describing each variable
  - and its values in the tidy data set.
- An explicit and exact recipe you used
  - to go from raw to tidy and a databook.

1. Is each variable one column?
2. Is each observation one row?
3. Do different data types appear in each table?
4. Did you record the recipe for moving from raw to tidy data?
5. Did you create a code(data) book?
6. Did you record all parameters, units, and functions applied to the data?

8.1.2.6.2 Raw data: It is critical that you include the rawest form of the data that you have access to.

Raw data is relative:

The raw data will be different to each person that handles the data.

One person's raw data, may be some previous persons tidy data!

8.1.2.6.3 Tidy data:

The general principles of tidy data are laid out by Hadley Wickham in

- this paper http://vita.had.co.nz/papers/tidy-data.pdf
- and this video https://vimeo.com/33727555.

The paper and the video are both focused on the tidyverse R packages.

Regardless the four general principles you should pay attention to are:

- Each variable you measure should be in one column
- Each different observation of that variable should be in different row
- There should be one table for each "kind" of variable
- If you have multiple tables, they should include a column in the table
  - that allows them to be linked

Include a row at the top of each data table/spreadsheet that contains full row names.

- If you are sharing your data with the collaborator

They should be shared as csv files, not in Excel

Since Excel files can have buried macros,

- you lost control of the data analysis process.

Also one csv table per file, no workbooks

- No highligting cells.
- csv files are for data only; no code.

- or one Excel file per table.
  - but xls or xlsx files are binary and fragile
  - ascii csv files are more robust

8.1.2.6.4 The code (or data) book:

The measurements you calculate

- will need to be described in more detail
- than you will sneak into the spreadsheet.

The code book contains this information.

At minimum it should contain:

- Information about the variables (including units!)
  - in the data set not contained in the tidy data
  - Information about the summary choices you made
  - Information about the experimental study design you used

8.1.2.6.5 The instruction list or script must be explicit

You may have heard this before,

- but reproducibility is kind of a big deal in computational science.

8.1.2.6.6 The ideal instruction list is a script

The ideal thing for you to do when performing summarization

- is to create a computer script (in R, Python, or something else)
- that takes the raw data as input
  - and produces the tidy data you are sharing as output.

8.1.2.6.7 If there is no script,

be very detailed about parameters, versions, and order of software

8.1.2.6.8 Common Mistakes

- Combining multiple variables into a single column
- Merging unrelated data into a single file
- An instruction list that isn't explicit

8.1.2.7 Checking the data

1. Did you plot univariate and multivariate summaries of the data?
2. Did you check for outliers?
3. Did you identify the missing data code?

Data munging or processing is required

- for basically every data set that you will have access to.

Even when the data are neatly formatted

- like you get from open data sources like http://Data.gov
- you'll frequently need to do things that make it
  - slightly easier to analyze or use the data for modeling.

The first thing to do with any new data set is

- to understand the quirks of the data set and potential errors.

This is usually done with a set of standard summary measures.

The checks should be performed on

- the rawest version of the data set you ave available.

A useful approach is to think of every possible thing that could go wrong

- and make a plot of the data to check if it did.

8.1.2.7.1 How to code variables

When you put variables into a spreadsheet

- there are several main categories you will run into
  - depending on their data type:
- Continuous
- Ordinal
- Categorical
- Missing
- Censored

8.1.2.7.2 In the code book you should explain why censored values are missing.

8.1.2.7.3 Avoid coding categorical or ordinal variables as numbers.

8.1.2.7.4 Always encode every piece of information about your observations using text.

8.1.2.7.5 Identify the missing value indicator

There are a number of different ways

- that missing values can be encoded in data sets.
- The common choices are "NA". Don't use numbers.

8.1.2.7.6 Check for clear coding errors

8.1.2.7.7 Check for label switching

8.1.2.7.8 If you have data in multiple files,

Ensure that data that should be identical across iles is identical

In some cases you will have the same measurements

- recorded twice.

You should check that for each patient

- in the two files the sex is recorded the same.
- This is part of data validation.

8.1.2.7.9 Check the units (or lack of units)

Define the units.

8.1.2.7.10 Common Mistakes

- Failing to check the data at all
- Encoding factors as quantitative numbers
- Not making sufficient plots
- Failing to look for outliers or missing values

8.1.2.8 Cee-lo a good, no house advantage game

- Cee-lo Dice Game
- Cee-lo Probabilities
- Rules and probabilities in readings cee-lo.txt
- Inference (Predicting the Future)

8.1.2.9 Cee-lo dice game

8.1.2.9.1 Cee-lo without a bank (winner take all)

In this version of the game,

- each round involves two or more players of equal status.

A bet amount is agreed upon and

- each player puts that amount in the pile or pot.

Each player then has to roll all three dice at once and

- must continue until a recognized combination is rolled.

Whichever player rolls the best combination

- wins the entire pot, and a new round begins.

In cases where two or more players tie for the best combination,

- they must have a shoot out to determine a single winner.

8.1.2.9.2 The combinations in Cee-lo

The combinations are similar to those described above, and can be ranked from best to worst as:

- 4-5-6
  - The highest possible roll. If you roll 4-5-6, you automatically win.
- Trips
  - Rolling three of the same number is known as rolling "trips".
  - Higher trips beat lower trips,
    - so 4-4-4 is better than * 3-3-3.
    - Any trips beats any established point.
- Point
  - Rolling a pair, and another number,
    - establishes the singleton as a "point".
  - A higher point beats a lower point,
    - so 2-2-6 is better than 5-5-2.
- 1-2-3
  - The lowest possible roll.
  - If you roll 1-2-3, you automatically lose.
- Any other roll is a meaningless combination and
  - must be rerolled until one of the above combinations occurs.

8.1.2.9.3 Probabilities[edit]

- With three six-sided dice there are $6 \times 6 \times 6$ or 216 possible permutations.
  - 4-5-6: 6/216 = 2.777777778% (Automatic Win)
  - Trips: 6/216 = 2.777777778%
  - Point: 90/216 = 41.66666667%
  - 1-2-3: 6/216 = 2.777777778% (Automatic Loss)
  - Meaningless permutations: 108/216 = 50%

8.1.2.9.4 dice, an R package to calculate dice games

dice

8.1.2.9.5 Rolling the Dice on a Warm Night

- Human mystical thinking
- And beware the bank