

A Case Study in Reproducibility: Detecting Data Analysis Patterns in Text and Graphs to Characterize Student Workflows

Ron Yurko Rebecca Nugent Philipp Burckhardt

Department of Statistics & Data Science
Carnegie Mellon University

Classification Society 2018

The Science of Data Science

Growing interest in **reproducible** research

Need for understanding the process - why someone takes certain steps and reaches different conclusions

*Many analysts, one dataset*¹- 29 teams of analysts reaching vastly different conclusions with same dataset

Goal: Understand the reproducibility of data analysis workflows

¹[Silberzahn et al., 2017]

Case study: revamped intro stats & data science course

Required course in Dietrich College of Humanities and Social Sciences general education curriculum for students in several programs:

- Economics, English, History, Information Systems, International Relations, Modern Languages, Philosophy, Psychology, Social & Decision Sciences, and Statistics & Data Science
- Also taken by majors across campus

New emphasis on student inquiry and writing about data analysis with non-traditional data types and interdisciplinary case studies

Students **interact with the ISLE e-learning framework**
(browser-based Interactive Statistics Learning Environment)
lead developer is Philipp Burckhardt

Students engage in the data analysis workflow with an interactive explorer that records answers and actions

Questions

« < 1 2 3 4 > »

For this last scenario, you'll work with a partner to choose and calculate summary measures, design and share a graph, and write up a description including a conclusion.

Scenario #4: It is thought that there is a relationship between the age of the student and the level of weekday alcohol use. Specifically, the older a student, the higher the level of weekday alcohol consumption.

Your Description

Your answer:

Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.

Toolbox

Data Statistics Tables **Plots**

Models Distributions

Scatterplot

Variable on x-axis:
Age

Variable on y-axis:
WkdyAlc

Color: Type: Size:

Select... Select... Select...

☐ Show Regression Model

Split By: Method:

Select... linear

Generate

Output

Age	WkdyAlc
16	1
16	2
16	3
16	4
16	5
17	1
17	2
17	3
17	4
17	5
18	1
18	2
18	3
18	4
18	5
19	1
19	2
19	3
19	4
19	5
20	1
20	2
20	3
20	4
20	5
21	1
21	2
21	3
21	4
21	5
22	1
22	2
22	3
22	4
22	5

Clear All

Students engage in the data analysis workflow with an interactive explorer that records answers and actions

Time: 11:30:22 PM | **User:** ryurko@andrew.cmu.edu

ID: description_scenario4 | **Type:** FREE_TEXT_QUESTION_SUBMIT_ANSWER

Value: Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.

Time: 11:24:33 PM | **User:** ryurko@andrew.cmu.edu

ID: schoolabsence | **Type:** DATA_EXPLORER:SCATTERPLOT

Value: {

```
"xval": "Age",  
"yval": "WkdyAlc",  
"color": null,  
"type": null,  
"regressionLine": false,  
"regressionMethod": "linear",  
"lineBy": null  
}
```

Students engage in the data analysis workflow with an interactive explorer that records answers and actions

Time: 11:30:22 PM | User: ryurko@andrew.cmu.edu

ID: description_scenario4 | Type: FREE_TEXT_QUESTION_SUBMIT_ANSWER

Value: Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.

Time: 11:24:33 PM | User: ryurko@andrew.cmu.edu

ID: schoolabsence | Type: DATA_EXPLORER:SCATTERPLOT

Value: {

```
"xval": "Age",  
"yval": "WkdyAlc",  
"color": null,  
"type": null,  
"regressionLine": false,  
"regressionMethod": "linear",  
"lineBy": null
```

}

Given these action logs, how can we characterize how students approach data analysis? The way they write about data?

Text analysis of students' answers

Starting with simple **bag-of-words** techniques - ignorant of order

Represent student text answers in a matrix where rows are individual student answers and columns are unique words

Values in matrix could be:

- does the student write the word: Yes / No
- number of times the word appears in each answer

²[Salton and McGill, 1986]

Text analysis of students' answers

Starting with simple **bag-of-words** techniques - ignorant of order

Represent student text answers in a matrix where rows are individual student answers and columns are unique words

Values in matrix could be:

- does the student write the word: Yes / No
- number of times the word appears in each answer
- penalized frequency by how many answers the word appears in

Term Frequency - Inverse Document Frequency (TF-IDF)²:

$$\text{TF-IDF} = \# \text{ of times word is in answer} \cdot \log\left(\frac{\# \text{ of answers}}{\# \text{ of answers with word}}\right)$$

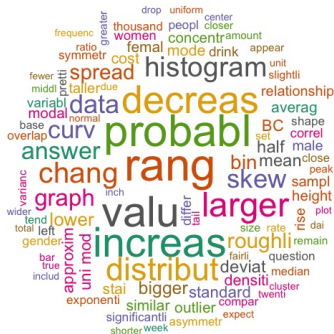
e.g. "the" would have a very low TF-IDF value

²[Salton and McGill, 1986]

Word cloud comparison for graphs

Created word clouds where size reflects the sum of TF-IDF values from answers where students made histograms compared to boxplots

Histograms



Boxplots



Overview of example lab session

Lab session in week five of class uses a single dataset about school absences in Portugal but consists of four question scenarios:

- **Scenario 1:** Number of absences by location, urban or rural?
- **Scenario 2:** Older students more likely to miss school?
- **Scenario 3:** Academic performance by number of classes failed, differences between males and females?
- **Scenario 4:** Relationship between age and alcohol use?

Scenarios 1-3: critique and write description with **explicit instructions** on what stats and graphs to edit/create

Scenario 4: only write description with **no guidance**

Refer to as: S1 Critique, S1 Description,..., S4 Description

Spherical K-means for clustering text

Spherical K-Means³: K-Means but minimizing cosine dissimilarity

$$d(x_1, x_2) = 1 - \cos(x_1, x_2) = 1 - \frac{\langle x_1, x_2 \rangle}{\|x_1\| \|x_2\|}$$

Spherical K-Means with $K = 4$ for each semester separately after pre-processing:

- Removed common stop words (e.g. a, an, but, the, ...)
- Removed numbers
- Spell-checked
- Stemming (e.g. cats \rightarrow cat, dependent \rightarrow depend)
- Used TF-IDF values (with respect to entire lab)

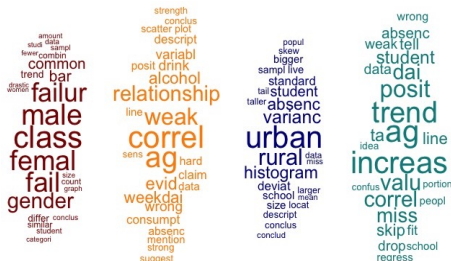
³[Dhillon and Modha, 2001]

Similar word clouds for clusters in each semester

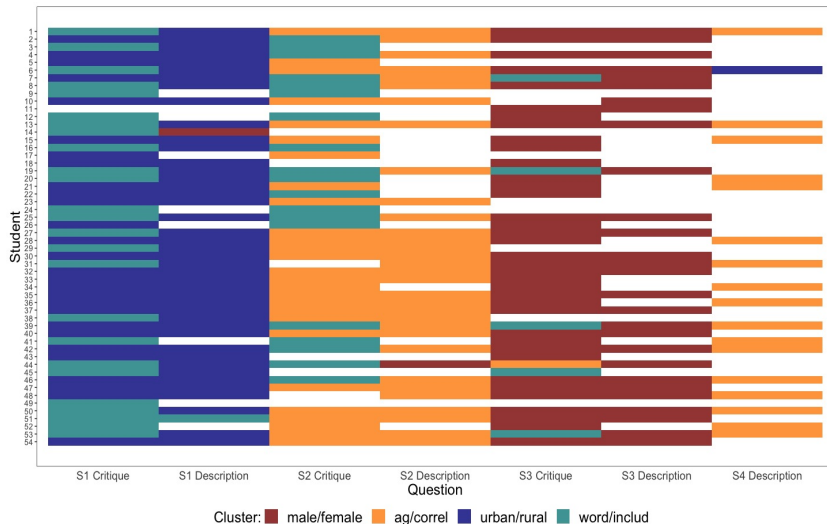
Fall 2017 - 266 answers by 54 students, 529 unique words



Spring 2018 - 543 answers by 101 students, 743 unique words



Fall semester clustering results



Example of difference in answers for S1 Description

Student 29's answer:

"The mean absences for urban students is greater than the mean absences for rural students, but the variance is also significantly higher for urban students than for rural students. Therefore, our results are inconclusive based solely on mean and variance."

Example of difference in answers for S1 Description

Student 29's answer:

"The mean absences for urban students is greater than the mean absences for rural students, but the variance is also significantly higher for urban students than for rural students. Therefore, our results are inconclusive based solely on mean and variance."

Meanwhile, Student 14's?

Example of difference in answers for S1 Description

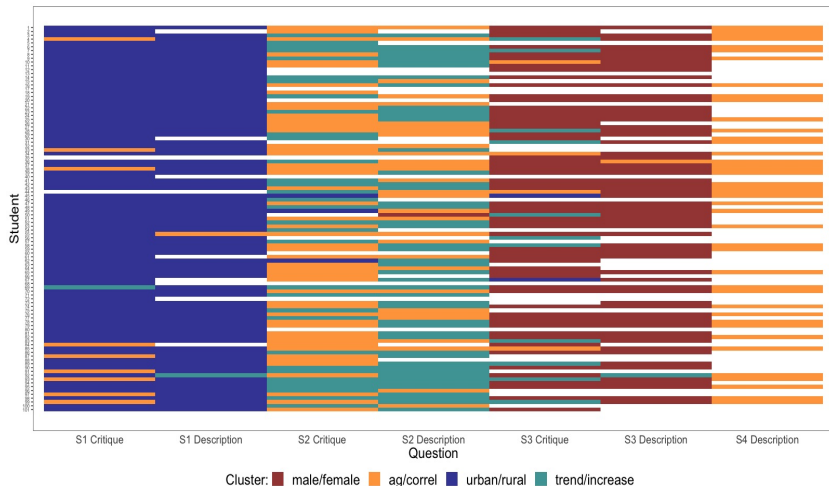
Student 29's answer:

"The mean absences for urban students is greater than the mean absences for rural students, but the variance is also significantly higher for urban students than for rural students. Therefore, our results are inconclusive based solely on mean and variance."

Meanwhile, Student 14's?

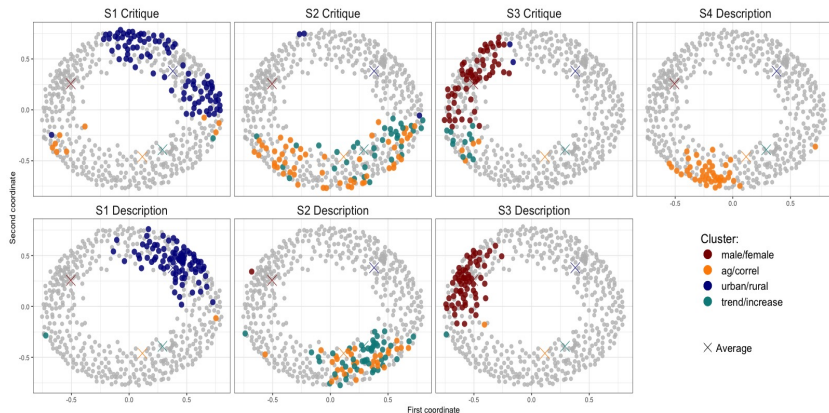
"Skipped for time."

Spring semester clustering results



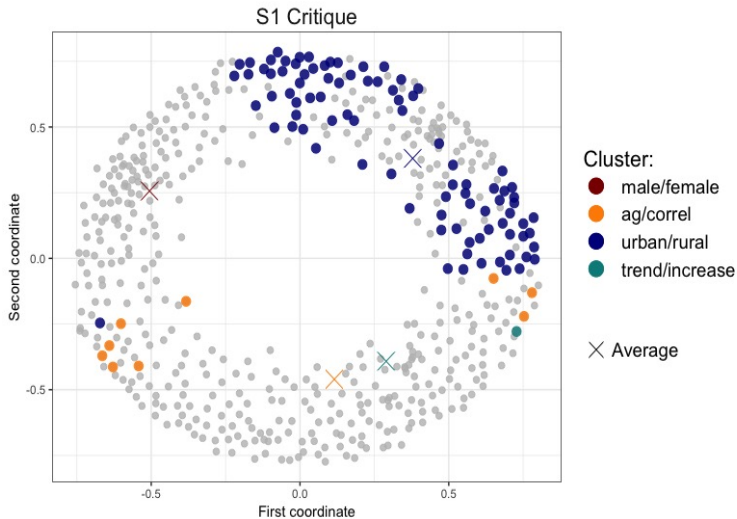
Structure of spring semester answers

MDS spherical projection⁴ using cosine dissimilarity matrix, including the averages from spherical K -means faceted by question:



⁴[de Leeuw and Mair, 2009]

Differences in student answers to S1 Critique



Differences in student answers to S1 Critique

Incorrect example description provided to students to critique:

"We have 3.885 and 24.049 for urban and 3.142 and 15.480 for rural. Because the mean and variance for urban are bigger than the mean and variance for rural, we definitely have more absences for the students in urban areas. When looking at the histograms, they both have right tails. However, because the urban histogram is taller than the rural histogram, we know there are more absences from the urban areas. So we conclude that living in a urban area causes students to miss more school."

Differences in student answers to S1 Critique

Word clouds for each cluster of answers:

explicit
causat
signific
signifi
mention
sentenc
hard unit
follow claim
bit draw
implic
specif told
variabl
conclus
suggestrepres
causalneutral
confoundreferenc
explan
exploratori

necessarili
standard
popul bigger
peopl conclus
miss student
school rural
sampl varianc
urban
descript causat
histogram
unit live deviat
definit taller
absenc
size
account
describ

fit
valu
chang
actual
data

Gold clusters answers focused on the manner of the description's conclusions, while the blue cluster focused on differences between urban and rural locations

Differences in student answers to S1 Critique

Example of answer from gold cluster:

"They didn't include the units or meaning of any of the numbers. Also, they drew conclusions without convincing enough evidence that the conclusions were true. They also didn't take into account the sample size for both categories. Correlation does not imply causation!"

Differences in student answers to S1 Critique

Example of answer from gold cluster:

"They didn't include the units or meaning of any of the numbers. Also, they drew conclusions without convincing enough evidence that the conclusions were true. They also didn't take into account the sample size for both categories. Correlation does not imply causation!"

Example of answer from blue cluster:

"One issue with this conclusion is that the number of students in each category is unequal. It makes sense that urban schools would see more or less absences compared to rural schools because they have so many more students. Looking at the graphs overlapping, the urban schools have a large majority of students that miss less than 5 days. The height of the histogram just reflects that number of students in that bin and the urban category has a lot more students."

Differences in cluster answers for S1 Critique

Answer from cyan cluster?

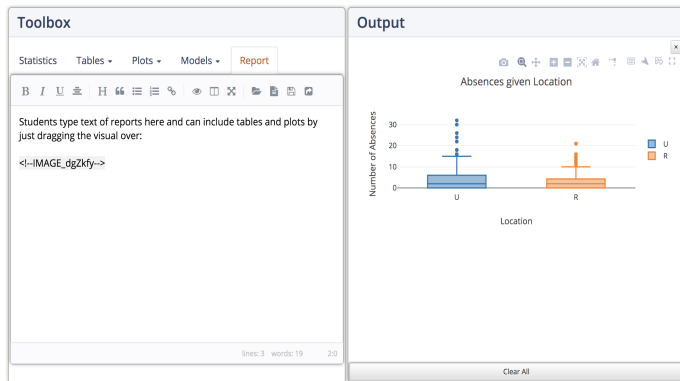
“Changed the values to fit the actual values from the data.”

Differences in cluster answers for S1 Critique

Answer from cyan cluster?

“Changed the values to fit the actual values from the data.”

In the spring semester, this lab session was completed in an **interactive markdown editor** - so students could change prompts



Clustering individual questions in lab five scenarios

Computed cosine dissimilarity matrix using TF-IDF values with respect to the question between all of the student answers, **incorrect example**, and **instructor solution**

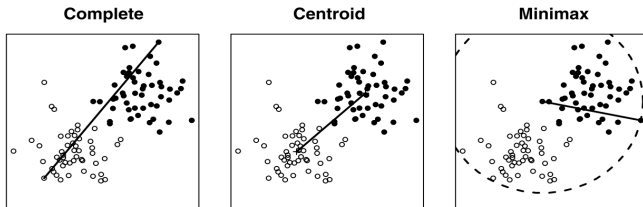
⁵[Bien and Tibshirani, 2011]

Clustering individual questions in lab five scenarios

Computed cosine dissimilarity matrix using TF-IDF values with respect to the question between all of the student answers, **incorrect example**, and **instructor solution**

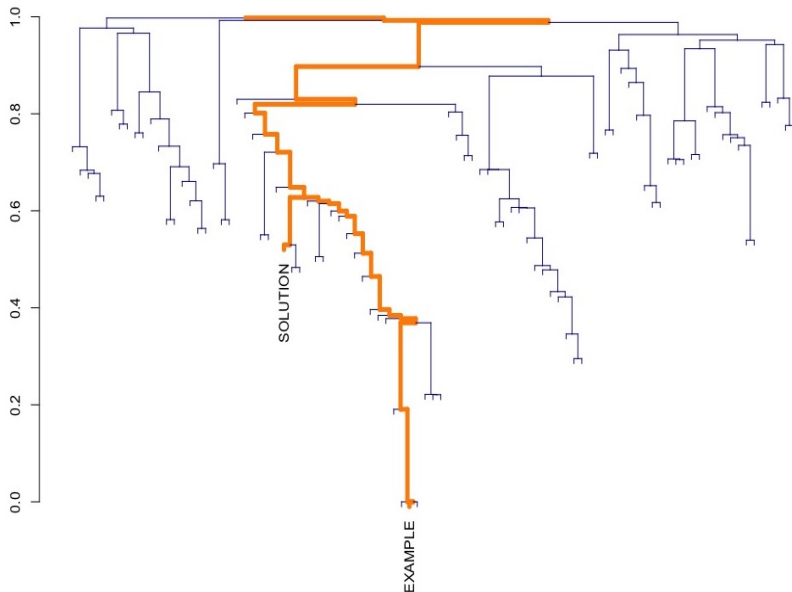
Applied **minimax linkage**⁵, each cluster has **prototype** answer

$$d(G, H) = \min_{x \in GUH} \left[\max_{x' \in GUH} d(x, x') \right]$$



⁵[Bien and Tibshirani, 2011]

Dendrogram for S1 Description

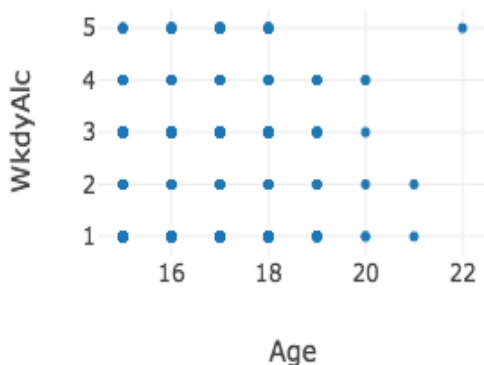


Open-ended question in fall semester - S4 Description

Question for students: Relationship between age and alcohol use?

Open-ended question in fall semester - S4 Description

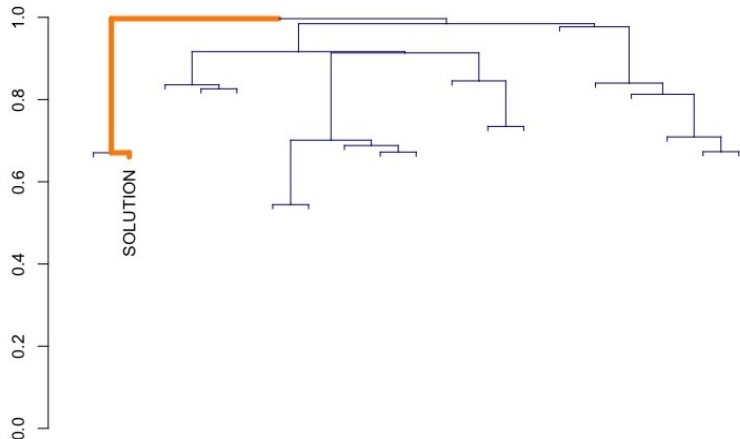
Question for students: Relationship between age and alcohol use?



Instructor solution described boxplots!

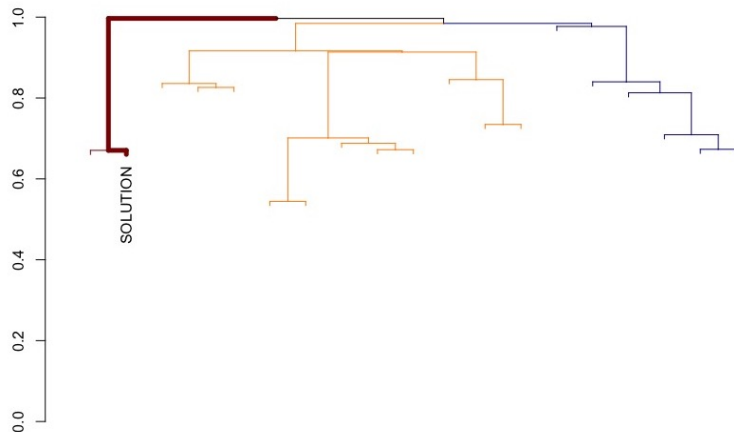
Dendrogram for S4 Description

19 answers, 93 unique words (after pre-processing)

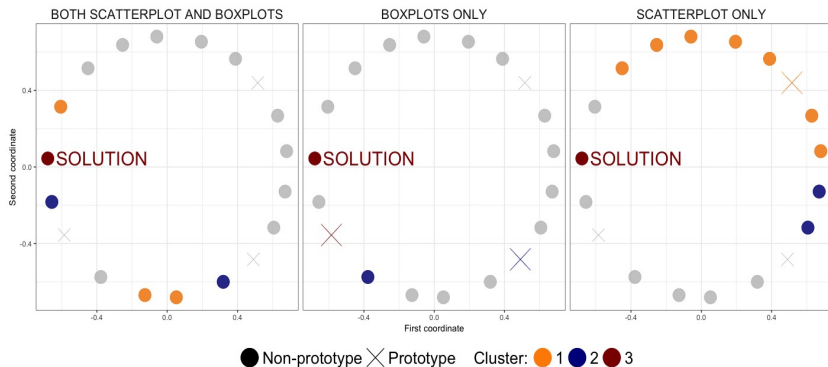


Dendrogram for S4 Description, $K = 3$

19 answers, 93 unique words (after pre-processing)



Students that made scatterplots described the relationship differently than those that only made boxplots



Comparison of prototype answers:

Prototype answer from gold cluster:

“The correlation value for age and weekday alcohol use is 0.086 which indicates a very weak positive linear relationship. The scatterplot shows a pretty even distribution of values, therefore I would not conclude that the older a student, the higher the level of weekday alcohol consumption.”

Comparison of prototype answers:

Prototype answer from gold cluster:

“The correlation value for age and weekday alcohol use is 0.086 which indicates a very weak positive linear relationship. The scatterplot shows a pretty even distribution of values, therefore I would not conclude that the older a student, the higher the level of weekday alcohol consumption.”

Comparison of prototype answers:

Prototype answer from gold cluster:

"The correlation value for age and weekday alcohol use is 0.086 which indicates a very weak positive linear relationship. The scatterplot shows a pretty even distribution of values, therefore I would not conclude that the older a student, the higher the level of weekday alcohol consumption."

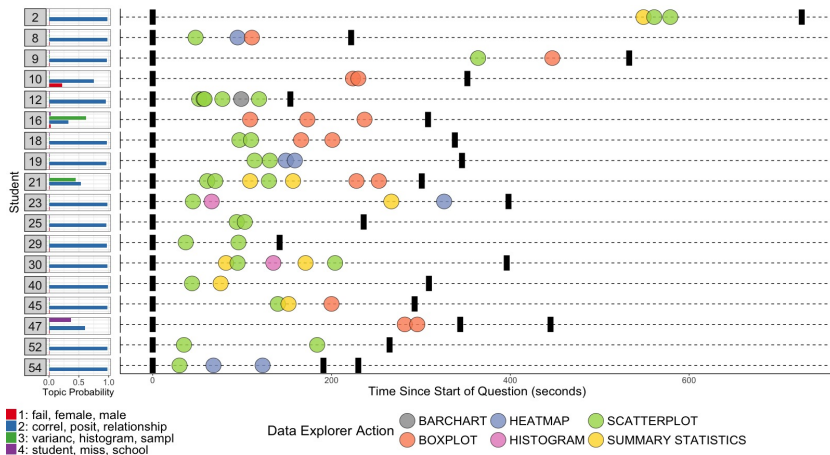
Prototype answer from blue cluster:

"The boxplot shows that the median of are similar in each alcohol use."

Prototype answer from red cluster:

"I believe that the outliers can be misleading and that the data suggest there is not a strong rel"

Link topic modeling of student answers to the timeline of their actions to understand why they answered differently



Discussion and next steps

Can cluster text of answers by students to reveal differences

Not just right or wrong, but full text of answers -
gaining insight into the process of student thinking

Discussion and next steps

Can cluster text of answers by students to reveal differences

Not just right or wrong, but full text of answers -
gaining insight into the process of student thinking

Continue to explore use of topic modeling procedures

Test sensitivity of vocabulary selection

Analysis with the ISLE platform **can be done in real-time**

Discussion and next steps

Can cluster text of answers by students to reveal differences

Not just right or wrong, but full text of answers -
gaining insight into the process of student thinking

Continue to explore use of topic modeling procedures

Test sensitivity of vocabulary selection

Analysis with the ISLE platform **can be done in real-time**

Move beyond bag-of-words and detect structure in their argument
with NLP techniques - **we have access to full data analysis reports**

Can lead to greater understanding of reproducible research

Acknowledgements

Advisor Rebecca Nugent

ISLE development team: Philipp Burckhardt and Frank Kovacs

Teaching Statistics Group

Science of Data Science / Clustering, Classification, and
Record Linkage Research Group

Contact info and thanks!

Email: ryurko@stat.cmu.edu

Website: <http://www.stat.cmu.edu/~ryurko/>




GitHub: [ryurko](#)

Twitter: [@Stat_Ron](#) [#CS2018](#)

References I

-  Aggarwal, C. C. (2018).
Machine Learning for Text.
Springer International Publishing, 1 edition.
-  Bien, J. and Tibshirani, R. (2011).
Hierarchical clustering with prototypes via minimax linkage.
Journal of the American Statistical Association, 106:1075–1084.
-  Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
Journal of Machine Learning Research, 3:993–1022.
-  de Leeuw, J. and Mair, P. (2009).
Multidimensional scaling using majorization: SMACOF in R.
Journal of Statistical Software, 31(3):1–30.

References II

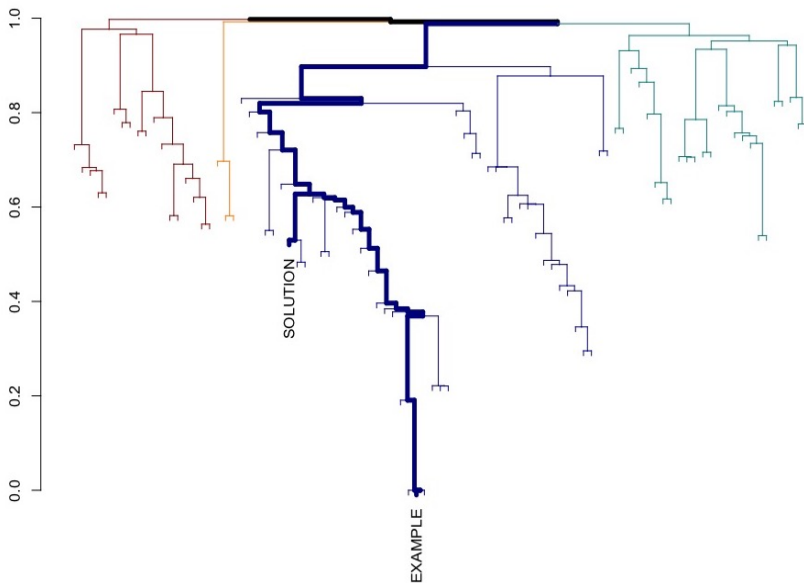
-  Dhillon, I. S. and Modha, D. S. (2001).
Concept decompositions for large sparse text data using clustering.
Machine Learning, 42(1):143–175.
-  Salton, G. and McGill, M. J. (1986).
Introduction to Modern Information Retrieval.
McGraw-Hill, Inc., New York, NY, USA.
-  Silberzahn, R., Uhlmann, E., Daniel, Pasquale, M., Frederik, A., Awtrey, A. E., Štěpán Bahník, Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Christensen, F. C. G., Clay, R., Craig, M., Dalla, A., Lammertjan, R., Mathew, D., Ismael, E., Cervantes, F., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T., Eriksson, K. H., Heene, M., Mohr, A. H., Hogden, F., Huiand, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy,

References III

D., Lei, R., Lindsay, T., Liverani, S., Madan, C., Molden, D., Molleman, E., Morey, R., Mulder, L., Nijstad, B., Pope, B., Pope, N., Prenoveau, J., Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlueter, E., S, F., Sherman, M., Sommer, S. A., Sotak, K., Spain, S., Sporlein, C., Stafford, T., Stefanutti, L., Täuber, S., Ullrich, J., Vianello, M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S., and Nosek, B. (2017).

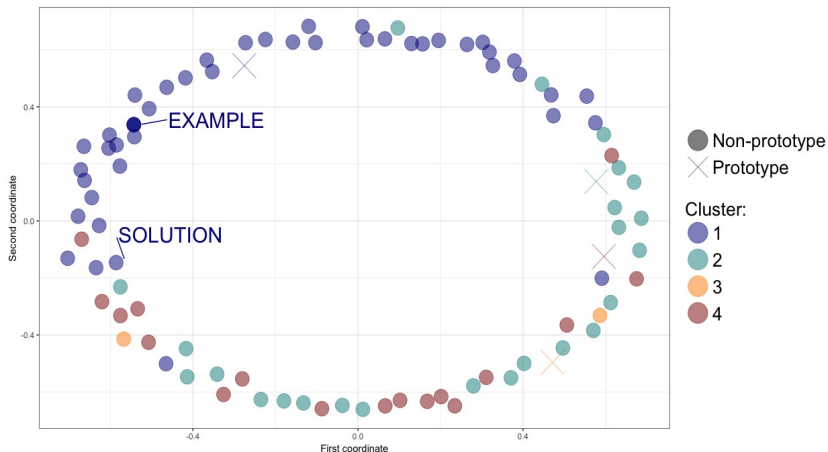
Many analysts, one dataset: Making transparent how variations in analytical choices affect results.

Dendrogram for S1 Description with $K = 4$



MDS projection of S1 Description answers

Answers in clusters one and two focus on similar points regarding differences in means and variances (limited by bag-of-words)



Prototype answers compared to instructor solution

Prototype answer for cluster three:

"Define the numbers used in the first sentence and what they imply. A greater variance also does not imply more absences, just a greater spread. Remove the causal implication in the last line and suggest a relationship of another kind."

Prototype answers compared to instructor solution

Prototype answer for cluster three:

"Define the numbers used in the first sentence and what they imply. A greater variance also does not imply more absences, just a greater spread. Remove the causal implication in the last line and suggest a relationship of another kind."

From prototype answer for cluster four:

"We have 3.885 and 4.904 mean and standard deviation for urban and 3.142 and 3.934 mean and standard deviation for rural... they both have right tails, and so they are right-skewed, with the mean larger than the median..."

Cluster four answers don't change the meaning of the example, just add/change certain phrases