DATA WRANGLING AND VISUALIZATION IN ACCESS SERVICES

Ellie Kohler, MA, MLIS
Access and Learning Services Librarian
Rockhurst University Library
November 17, 2016

2016 Access Services Conference
Atlanta, Georgia
HTTP://TINY.CC/DATAWRANGLE

Links to some of the tools I’m talking about today.
Your suggestions!
Disclaimer:
This presentation will not make you an expert.
You know data

- DS 63.04 .M38 2015
- The Ottoman Endgame
- About the Ottoman Empire
- It’s yellow
Where do you get your data?

- Google sheets: internal reports
- Web Management Reports
- ILS System: III (Sierra)
- MOBIUS Statistics
- OCLC Web Reports
- Wufoo
- Gimlet
- The list goes on...
All identifiable information is blocked or stripped from these transactions.

Measures of Student Success
- Grades
- Graduation

- Checked out books
- Asked for library help
- Was in a library instruction class
- Used library chat
- Used databases
- Studied in library
- Accessed ebooks
- ILL Books and articles
What is Data Wrangling?

- Obtain
- Understand
- Transform
- Augment
- Shape
- Share
Hell is other people’s data
80 / 20 Rule

- Languages
- Dates
- “Revisions”
- Mistakes - Typos
- Blanks
- Aliasing
- Wrong format
- Too specific
- Too broad
Free Tools That Can Help

- Free
- Download

Refine
- Free version
- Online
- Bigger datasets

TRIFACTA
- Free version
- Download

Talend
- Free version
- Download
- Version 1.3
### Cluster & Edit column "Complete Journal Title"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find

<table>
<thead>
<tr>
<th>Method</th>
<th>key collision</th>
<th>Keying Function</th>
<th>fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>International Review for the Sociology of Sport (1 rows)</td>
<td>International Review for the Sociology of Sport (1 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>International Review for the Sociology of Sport (1 rows)</td>
<td>International Review for the Sociology of Sport (1 rows)</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neuropsychologia (4 rows)</td>
<td>Neuropsychologia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neuropsychologia (2 rows)</td>
<td>Neuropsychologia</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Journal of Research in Childhood Education (1 rows)</td>
<td>Journal of Research in Childhood Education (1 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Journal of Research in Childhood Education (1 rows)</td>
<td>Journal of Research in Childhood Education (1 rows)</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Equity and Excellence in Education (3 rows)</td>
<td>Equity and Excellence in Education (3 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Equity and Excellence in Education (3 rows)</td>
<td>Equity and Excellence in Education (3 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1 rows)</td>
<td>(1 rows)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BioEssays (1 rows)</td>
<td>BioEssays</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BioEssays (1 rows)</td>
<td>BioEssays</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Psychiatry Research (7 rows)</td>
<td>Psychiatry Research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Psychiatry Research (1 rows)</td>
<td>Psychiatry Research</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brain Research Bulletin (2 rows)</td>
<td>Brain Research Bulletin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brain research bulletin (1 rows)</td>
<td>Brain research bulletin</td>
</tr>
</tbody>
</table>
When clustering fails you....
R: Free Tool for Coding

Packages:
- dplyr
- tidyr
- data.table
- psych
- zoo
- janitor
- gmodels

Diamond Pricing

```r
library(ggplot2)
sources('plots/FormatPlot.R')
library(dplyr)
library(tidyr)
library(data.table)
library(psych)
library(zoo)
library(janitor)
library(gmodels)

diamonds <- read.csv('diamonds.csv')
summary(diamonds)
aveCarat <- ave(clarity, diamonds$carat)
avePrice <- ave(price, diamonds$carat)

p <- qplot(carat, price, data=diamonds, color=clarity, main='Diamond Pricing')
format.plot(p, size=24)
```
Python: Free Tool for Coding

Packages:
- pandas
- numpy
- matplotlib
VISUALIZATION
#1 Correlation does not equal causation

**Definitive Proof that Jim Carrey Causes Autism**

- Autism prevalence (cases per 1,000 population)
- Jim Carrey movies (total)

Sources: CDC; IMDb.

$r^2 = 0.8891$
Rule #2: Put Your Pie Charts on a Diet
Rule #3: Don’t Mislead
It’s All About the Data
Make interactive, quality graphs online

- R
- Python (Jupyter)
- D3.js

….or just make graphs
Story That Made Headlines

"Numbers have an important story to tell. They rely on you to give them a voice."

YEAR
Tableau

- Data Visualization standard
- Data sharing
- Some analytics
- Public, Educational, Paid
- Steep Learning Curve
- Learn: Coursera, Tableau website, Tableau blog, Youtube videos, Udemy
- [https://public.tableau.com/profile/ellie.kohler/](https://public.tableau.com/profile/ellie.kohler/)
Overdues

Top Checkout Subjects This Month

- Reserves
- Higher Education
- Religion and Theology - Christianity, the Bible, and...
- Literature - Drama
- Language - Basque (Uralic)

Liaison Ordering

- Department
  - Communications: 0.0%
  - Engineering: 0.0%
  - Exercise Sci.: 19.4%
  - Mathematics: 0.0%
  - Physics: 0.0%

Updated Monthly
Ordering Deadline: Saturday March 4, 2017
Instruction, Reference and ILL Article Requests

Chose a Date Range
12/30/2015 to 6/30/2016

Instruction and Article Requests (Dual Axis)

Instruction and Reference Questions (Dual Axis)
THANK YOU!

QUESTIONS?

Ellie.Kohler@Rockhurst.edu