data cleaning principles

Karl Broman
Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kbroman@fosstodon.org
bit.ly/datacleaning2023
Tidy data are all alike, but every messy dataset is messy in its own way.

– Hadley Wickham

r4ds.had.co.nz/tidy-data.html
If I clean up [Medicare] data ... does any of the knowledge I gain ... apply to the processing of RNA-seq data?

– Roger Peng
Join us for the first inaugural Data Mishaps Night! We will feature a lineup of data mistake stories with a focus on the human aspect of data work and lessons learned the hard way.
Data cleaning

► tedious
► embarrassing
► needs context
► doesn’t feel like progress
Data cleaning

- tedious
- embarrassing
- needs context
- doesn’t feel like progress

- requires creativity
- requires coding prowess
- source of many problems
Data cleaning principles

fundamentals

verify
explore

ask
document
1. Don’t clean data when you’re tired or hungry.

(paraphrasing Ghazal Gulati)
2. Don’t trust anyone (even yourself)
2. Don’t trust anyone (even yourself)

“my motto is ‘trust no one’
...except maybe @kwbronian?”

– Jenny Bryan
3. Think about what might have gone wrong and how it might be revealed

doi:10/gpfzs8
4. Use care in merging
fundamentals

5. Dates & categories suck
Principle:

a fundamental truth that guides our thinking
5. Dates & categories suck
6. Check that distinct things are distinct

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WisclID</td>
<td>ID</td>
<td>NEOID</td>
<td>Fem_CA</td>
<td>Fem_Imax</td>
<td>Fem_Lmin</td>
<td>Fem_J</td>
</tr>
<tr>
<td>2</td>
<td>F2.C1W.F.1248</td>
<td>1248</td>
<td>NEO183</td>
<td>0.7524</td>
<td>0.1427</td>
<td>0.1006</td>
<td>0.2433</td>
</tr>
<tr>
<td>3</td>
<td>F2.C1W.M.1250</td>
<td>1250</td>
<td>NEO184</td>
<td>0.7669</td>
<td>0.1556</td>
<td>0.09652</td>
<td>0.2521</td>
</tr>
<tr>
<td>4</td>
<td>F2.C1W.F.1251</td>
<td>1251</td>
<td>NEO185</td>
<td>0.7613</td>
<td>0.1549</td>
<td>0.09659</td>
<td>0.2515</td>
</tr>
<tr>
<td>5</td>
<td>F2.C1W.F.1254</td>
<td>1254</td>
<td>NEO186</td>
<td>0.7475</td>
<td>0.1503</td>
<td>0.08603</td>
<td>0.2363</td>
</tr>
<tr>
<td>6</td>
<td>F2.C1W.M.1257</td>
<td>1257</td>
<td>NEO187</td>
<td>0.8197</td>
<td>0.1849</td>
<td>0.1056</td>
<td>0.2905</td>
</tr>
<tr>
<td>7</td>
<td>F2.___.F.715</td>
<td>715</td>
<td>NEO764</td>
<td>0.6017</td>
<td>0.09662</td>
<td>0.05969</td>
<td>0.1563</td>
</tr>
<tr>
<td>8</td>
<td>F2.___.F.751</td>
<td>751</td>
<td>NEO765</td>
<td>0.7273</td>
<td>0.1304</td>
<td>0.08735</td>
<td>0.2178</td>
</tr>
<tr>
<td>9</td>
<td>F2.___.F.1251</td>
<td>1251</td>
<td>NEO766</td>
<td>0.6675</td>
<td>0.1157</td>
<td>0.07814</td>
<td>0.1938</td>
</tr>
<tr>
<td>10</td>
<td>F2.___.M.1340</td>
<td>1340</td>
<td>NEO768</td>
<td>0.6656</td>
<td>0.1387</td>
<td>0.08122</td>
<td>0.2199</td>
</tr>
<tr>
<td>11</td>
<td>F2.C1W.M.739</td>
<td>739</td>
<td>NEO779</td>
<td>0.9336</td>
<td>0.2828</td>
<td>0.1628</td>
<td>0.4456</td>
</tr>
</tbody>
</table>
7. Check that matching things match

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>id</td>
<td>sex</td>
<td>n_gen</td>
<td>age_days</td>
</tr>
<tr>
<td>2</td>
<td>F20.25</td>
<td>M</td>
<td>20</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td>F21.30</td>
<td>M</td>
<td>21</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>F21.68</td>
<td>M</td>
<td>21</td>
<td>71</td>
</tr>
<tr>
<td>5</td>
<td>F22.52</td>
<td>M</td>
<td>22</td>
<td>73</td>
</tr>
<tr>
<td>6</td>
<td>F21.71</td>
<td>F</td>
<td>22</td>
<td>63</td>
</tr>
<tr>
<td>7</td>
<td>F22.116</td>
<td>F</td>
<td>22</td>
<td>57</td>
</tr>
<tr>
<td>8</td>
<td>F21.F20.9.M5</td>
<td>M</td>
<td>20</td>
<td>82</td>
</tr>
<tr>
<td>10</td>
<td>F20.26</td>
<td>M</td>
<td>20</td>
<td>75</td>
</tr>
<tr>
<td>11</td>
<td>F21.62</td>
<td>M</td>
<td>21</td>
<td>72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>id</td>
<td>sex</td>
<td>age_at_dosing</td>
<td>n_gen</td>
</tr>
<tr>
<td>2</td>
<td>F22.69</td>
<td>F</td>
<td>67</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>F22.106</td>
<td>F</td>
<td>69</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>F22.70</td>
<td>F</td>
<td>67</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>F22.107</td>
<td>F</td>
<td>69</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>F21.71</td>
<td>F</td>
<td>65</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>F22.116</td>
<td>F</td>
<td>62</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>F22.73</td>
<td>F</td>
<td>65</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>F22.117</td>
<td>F</td>
<td>62</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>F21.108</td>
<td>F</td>
<td>62</td>
<td>21</td>
</tr>
<tr>
<td>11</td>
<td>F22.118</td>
<td>F</td>
<td>59</td>
<td>22</td>
</tr>
</tbody>
</table>
8. Check calculations
9. Look for other instances of a problem
explore

10. Make lots of plots
10. Make lots of plots

![Graph showing body weight comparison between 6 wk and 10 wk]
10. Make lots of plots
11. Look at missing value patterns
12. With massive data, make more plots not fewer.
12. With massive data, make more plots not fewer
explore

12. With massive data, make more plots not fewer

[Graph showing a scatter plot with median and interquartile range values]
12. With massive data, make more plots not fewer
explore

13. Follow up all artifacts
14. Ask questions

When were the data gathered?
How, and by whom?
Was the data gathered in batches?
How were the data files created?
Was any calibration or normalization done?
15. Ask for the primary data

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MouseNum</td>
<td>Wean Date</td>
<td>Assay Date</td>
<td>Weight</td>
<td>Glucose</td>
<td>Insulin</td>
<td>Trigly</td>
<td>HOMA</td>
</tr>
<tr>
<td>2</td>
<td>Mouse3001</td>
<td>6/22/2005</td>
<td>8/18/2005</td>
<td>47.3</td>
<td>617</td>
<td>11.7</td>
<td>175.1</td>
<td>321.9</td>
</tr>
<tr>
<td>3</td>
<td>Mouse3002</td>
<td>6/22/2005</td>
<td>8/18/2005</td>
<td>51</td>
<td>256.5</td>
<td>50.6</td>
<td>97.6</td>
<td>576.6</td>
</tr>
<tr>
<td>4</td>
<td>Mouse3003</td>
<td>6/22/2005</td>
<td>8/18/2005</td>
<td>50.6</td>
<td>274.9</td>
<td>52.5</td>
<td>160.5</td>
<td>641.3</td>
</tr>
<tr>
<td>5</td>
<td>Mouse3004</td>
<td>6/22/2005</td>
<td>8/18/2005</td>
<td>46</td>
<td>615.1</td>
<td>9</td>
<td>238.7</td>
<td>246</td>
</tr>
<tr>
<td>6</td>
<td>Mouse3005</td>
<td>6/30/2005</td>
<td>b</td>
<td>42.4</td>
<td>NA</td>
<td>NA</td>
<td>102.3</td>
<td>587.1</td>
</tr>
<tr>
<td>7</td>
<td>Mouse3006</td>
<td>6/30/2005</td>
<td>b</td>
<td>39.7</td>
<td>NA</td>
<td>NA</td>
<td>209.4</td>
<td>338.7</td>
</tr>
<tr>
<td>8</td>
<td>Mouse3007</td>
<td>6/30/2005</td>
<td>b</td>
<td>36.9</td>
<td>NA</td>
<td>NA</td>
<td>69.8</td>
<td>140.6</td>
</tr>
<tr>
<td>9</td>
<td>Mouse3008</td>
<td>6/30/2005</td>
<td>b</td>
<td>50</td>
<td>195.5</td>
<td>45.3</td>
<td>142.1</td>
<td>393.4</td>
</tr>
<tr>
<td>10</td>
<td>Mouse3009</td>
<td>6/30/2005</td>
<td>b</td>
<td>40.1</td>
<td>569.4</td>
<td>12.4</td>
<td>411</td>
<td>312.9</td>
</tr>
<tr>
<td>11</td>
<td>Mouse3010</td>
<td>6/30/2005</td>
<td>b</td>
<td>40.7</td>
<td>593.8</td>
<td>15.6</td>
<td>333.6</td>
<td>411.8</td>
</tr>
</tbody>
</table>
“What the heck is ‘FAD_NAD SI 8.3_3.3G’?”
16. Ask for metadata

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>name</td>
<td>plot_name</td>
<td>group</td>
<td>description</td>
</tr>
<tr>
<td>2</td>
<td>mouse</td>
<td>Mouse</td>
<td>demographic</td>
<td>Animal identifier</td>
</tr>
<tr>
<td>3</td>
<td>sex</td>
<td>Sex</td>
<td>demographic</td>
<td>Male (M) or Female (F)</td>
</tr>
<tr>
<td>4</td>
<td>sac_date</td>
<td>Date of sac</td>
<td>demographic</td>
<td>Date mouse was sacrificed</td>
</tr>
<tr>
<td>5</td>
<td>partial_inflation</td>
<td>Partial inflation</td>
<td>clinical</td>
<td>Indicates if mouse showed partial pancreatic inflation</td>
</tr>
<tr>
<td>6</td>
<td>coat_color</td>
<td>Coat color</td>
<td>demographic</td>
<td>Coat color, by visual inspection</td>
</tr>
<tr>
<td>7</td>
<td>crumblers</td>
<td>Crumblers</td>
<td>clinical</td>
<td>Indicates if mouse stored food in their bedding</td>
</tr>
<tr>
<td>8</td>
<td>diet_days</td>
<td>Days on diet</td>
<td>clinical</td>
<td>Number of days on high−fat diet</td>
</tr>
</tbody>
</table>
17. Ask why data are missing

Assay failed?
Below detection limit?
Viewed as outliers?
Subjects dropped out?
18. Create checklists & pipelines

- Percent missing genotypes
- Sample duplicates
- Sex and X/Y genotypes
- Heterozygosity
- Genotype frequencies
- Crossover counts
- Genotyping error rates
19. Document not just what but why

Gough project diagnostics

Karl Broman, 3 March 2014

Combine genotypes and phenotypes

I've combined the initial genotypes (using the re-clustered genotypes for plates 14-16) with the well-behaved portion of the re-run genotypes. I'm focusing on 36813 markers that are informative (though, as we'll see, there are still a lot of badly behaved and basically non-informative markers that need to be removed). I've combined data on replicate samples, to give one set of genotype calls for each sample.

There are 1497 genotyped mice and 1464 phenotyped mice. All of the mice in the phenotype data have genotypes, but there are 33 genotyped mice with no phenotypes, including 3 Gough mice and 30 F2 progeny.
20. Expect to recheck
Data cleaning principles

**fundamentals**
1. Don't clean data when tired or hungry
2. Don't trust anyone (even yourself)
3. Think about what might have gone wrong
4. Use care in merging
5. Dates & categories suck

**verify**
6. Verify that distinct things are distinct
7. Verify that matching things match
8. Check calculations
9. Look for other instances of problems

**explore**
10. Make lots of plots
11. Look at missing value patterns
12. With big data make more plots
13. Follow up all artifacts

**ask**
14. Ask questions
15. Ask for the primary data
16. Ask for metadata
17. Ask why data are missing

**document**
18. Create checklists & pipelines
19. Document not just what but why
20. Expect to recheck
I will let the data speak for itself when it cleans itself.

– Allison Reichel
**fundamentals**
1. Don't clean data when tired or hungry
2. Don't trust anyone (even yourself)
3. Think about what might have gone wrong
4. Use care in merging
5. Dates & categories suck

**verify**
6. Verify that distinct things are distinct
7. Verify that matching things match
8. Check calculations
9. Look for other instances of problems

**explore**
10. Make lots of plots
11. Look at missing value patterns
12. With big data make more plots
13. Follow up all artifacts

**ask**
14. Ask questions
15. Ask for the primary data
16. Ask for metadata
17. Ask why data are missing

**document**
18. Create checklists & pipelines
19. Document not just what but why
20. Expect to recheck