data cleaning principles

Karl Broman
Biostatistics & Medical Informatics, UW–Madison

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

– Hadley Wickham
If I clean up [Medicare] data ... does any of the knowledge I gain ... apply to the processing of RNA-seq data?

– Roger Peng

doi:10/jz69
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 @kw broman?”

– Jenny Bryan
3. Think about what might have gone wrong and how it might be revealed.
4. Use care in merging

<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>id</td>
<td>glucose.0</td>
<td>glucose.5</td>
<td>glucose.15</td>
<td>glucose.30</td>
<td>insulin.0</td>
<td>insulin.5</td>
</tr>
<tr>
<td>2</td>
<td>DO−221</td>
<td>145.742786</td>
<td>206.452638</td>
<td>216.640608</td>
<td>299.55501</td>
<td>0.74455</td>
<td>2.0264</td>
</tr>
<tr>
<td>3</td>
<td>DO−222</td>
<td>226.92224</td>
<td>303.803088</td>
<td>299.55501</td>
<td>0.74455</td>
<td>2.0264</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>DO−223</td>
<td>121.051535</td>
<td>0.0882</td>
<td>407.355505</td>
<td>0.63475</td>
<td>470.541525</td>
<td>0.8195</td>
</tr>
<tr>
<td>5</td>
<td>DO−224</td>
<td>122.95695</td>
<td>0.19155</td>
<td>298.193665</td>
<td>0.6467</td>
<td>323.148455</td>
<td>0.40515</td>
</tr>
<tr>
<td>6</td>
<td>DO−225</td>
<td>143.60919</td>
<td>0.23435</td>
<td>438.88705</td>
<td>0.70505</td>
<td>406.249135</td>
<td>0.2498</td>
</tr>
<tr>
<td>7</td>
<td>DO−226</td>
<td>125.29262</td>
<td>0.04</td>
<td>543.74634</td>
<td>1.7366</td>
<td>520.205245</td>
<td>0.8498</td>
</tr>
<tr>
<td>8</td>
<td>DO−227</td>
<td>135.61874</td>
<td>0.91275</td>
<td>393.03416</td>
<td>3.73095</td>
<td>454.62209</td>
<td>1.7325</td>
</tr>
</tbody>
</table>
5. Dates & categories suck
Principle:

a fundamental truth that guides our thinking
5. Dates & categories suck
verify

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

![Graph showing body weight data for 6 and 10 weeks.]
explore

10. Make lots of plots

Mouse index
Adipose weight (mg)

0 100 200 300 400 500
0 200 400 600 800 1000 1200 1400
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
12. With massive data, make more plots not fewer
explore

13. Follow up all artifacts

kbroman.org/blog/2012/04/25/microarrays-suck
14. Ask questions
15. Ask for the primary data
16. Ask for metadata
17. Ask why data are missing
18. Create checklists & pipelines
19. Document not just what but why
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
6. Verify that distinct things are distinct
7. Verify that matching things match
8. Check calculations
9. Look for other instances of problems

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

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

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

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

explore

ask

document