The Importance of Data Management and Data Quality
About the Presenters

• **Radost Wenman**, FCAS, MAAA, CSPA
  • Consulting Actuary
  • San Francisco, California

• **Greg Frankowiak**, FCAS, MAAA, CSPA, CPCU, MSM
  • Senior Consulting Actuary
  • Bloomington, Illinois
Agenda

- Financial cost of poor data quality
- ASOPs and NAIC guidance
- Data quality best practices
- Review of missing data issues and methods
How much do you believe the total annual financial cost of poor quality data is for organizations in the US?

A. Less than $1 billion
B. $1 billion – $10 billion
C. $10 billion – $100 billion
D. $100 billion – $1 trillion
E. More than $1 trillion

Polling Question
Financial Cost of Poor Data Quality
The Cost of Poor Data Quality

IBM estimates for the entire US economy:

$3.1 \text{ Trillion}

facebook + Apple + Amazon

Combined Market Cap $3.1 \text{ Trillion}
The Cost of Poor Data Quality

Gartner estimates average financial impact per organization:

$9.7 - $14.2 Million

~30% of Revenue
The Cost of Poor Data Quality

- Data Quality Concerns: 84%
- Lost Revenue: 77%
- Lower Productivity: 50%
- Maintenance Costs: 30%
- Inaccurate Data: 25%
- Negative Reputation: 21%
The Cost of Poor Data Quality

Globally, **82%** of organizations are operating without an optimized strategy for data management.
Insurance Survey: % of Time Spent on Data Quality Issues?

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage Responding (%)</th>
<th>Mean Response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Re)Insurer</td>
<td>54</td>
<td>25</td>
</tr>
<tr>
<td>Consulting</td>
<td>23</td>
<td>27</td>
</tr>
<tr>
<td>Other</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>27</td>
</tr>
</tbody>
</table>

Source: GIRO Data Quality Working Party
## Insurance Survey - % of Projects Adversely Affected?

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage Responding (%)</th>
<th>Mean Response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Re)Insurer</td>
<td>54</td>
<td>33</td>
</tr>
<tr>
<td>Consulting</td>
<td>23</td>
<td>38</td>
</tr>
<tr>
<td>Other</td>
<td>23</td>
<td>35</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
<td><strong>34</strong></td>
</tr>
</tbody>
</table>

Source: GIRO Data Quality Working Party
Professional and Regulatory Guidance
ASOP 23 – Data Quality

- Great guidance even if not required
- Section 3.1—Business context
- Section 3.2—Overall data review considerations
- Section 3.3—Specific data review for project
- Section 3.8—Steps *not* required to take
NAIC’s CASTF White Paper

- Casualty Actuarial and Statistical Task Force (CASTF)
- “Regulatory Review of Predictive Models”
- Identifying best practices to guide state insurance departments in their review of predictive models for underlying rating plans

Source: naic.org
NAIC’s CASTF White Paper

• Available Data Sources

• Adjustments to Data

• Data Organization
Modeling ASOP (#56)

- Effective October 2020
- Data section refers to ASOP 23 and ASOP 41
Data Quality Best Practices

Establishing a Data Management Program
What level of focus does your organization have on data management activities?

A. Dedicated cross-department teams and resources
B. Some dedicated resources in some departments
C. People work on it as they have time
D. Considering a data management plan
E. No plans
Establishing a Data Management Program

Support

Metadata

Share

Current State

Governance
Data Quality Best Practices

Continued Quality of Data Stores

Commitment Beyond Numbers
Continued Quality of Data Stores

- Allowed Values/Audits
- Privacy
- Automate/Balance
- Controls
- Refresh
Data Quality Best Practices

Data Quality for Datasets

Commitment Beyond Numbers
Data Quality for Datasets

Assess/Compare

- Document
- Outliers
- Visualize
- Subject Matter Expertise
Review of Missing Data Issues and Methods
Data Profiling – Missing Data

<table>
<thead>
<tr>
<th>Var1</th>
<th>Var2</th>
<th>Var3</th>
<th>Var4</th>
<th>Var5</th>
</tr>
</thead>
<tbody>
<tr>
<td>93</td>
<td>23</td>
<td>M</td>
<td>S</td>
<td>1</td>
</tr>
<tr>
<td>NA</td>
<td>3</td>
<td>NA</td>
<td>S</td>
<td>2</td>
</tr>
<tr>
<td>62</td>
<td>NA</td>
<td>F</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>79</td>
<td>NA</td>
<td>M</td>
<td>M</td>
<td>3</td>
</tr>
<tr>
<td>66</td>
<td>39</td>
<td>M</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Most common issue with “dirty” data
Different mechanisms of “missing-ness”
Affects the accuracy of most ML algorithms
Creates bias and reduces power

Ignoring It Does Not Fix the Issues!
## Data Profiling – Missing Data

The table below shows the proportion of data effectively available for a given number of variables, expressed as a percentage of missing data. Each row represents the proportion of data missing for a specific number of variables.

<table>
<thead>
<tr>
<th>Proportion Missing (%)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99</td>
<td>98</td>
<td>97</td>
<td>96</td>
<td>95</td>
<td>90</td>
<td>86</td>
<td>82</td>
<td>78</td>
<td>74</td>
<td>67</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>98</td>
<td>96</td>
<td>94</td>
<td>92</td>
<td>90</td>
<td>82</td>
<td>74</td>
<td>67</td>
<td>60</td>
<td>55</td>
<td>45</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>97</td>
<td>94</td>
<td>91</td>
<td>89</td>
<td>86</td>
<td>74</td>
<td>63</td>
<td>54</td>
<td>47</td>
<td>40</td>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>96</td>
<td>92</td>
<td>88</td>
<td>85</td>
<td>82</td>
<td>66</td>
<td>54</td>
<td>44</td>
<td>36</td>
<td>29</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>95</td>
<td>90</td>
<td>86</td>
<td>81</td>
<td>77</td>
<td>60</td>
<td>46</td>
<td>36</td>
<td>28</td>
<td>21</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>94</td>
<td>88</td>
<td>83</td>
<td>78</td>
<td>73</td>
<td>54</td>
<td>40</td>
<td>29</td>
<td>21</td>
<td>16</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>93</td>
<td>86</td>
<td>80</td>
<td>75</td>
<td>70</td>
<td>48</td>
<td>34</td>
<td>23</td>
<td>16</td>
<td>11</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>92</td>
<td>85</td>
<td>78</td>
<td>72</td>
<td>66</td>
<td>43</td>
<td>29</td>
<td>19</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>91</td>
<td>83</td>
<td>75</td>
<td>69</td>
<td>62</td>
<td>39</td>
<td>24</td>
<td>15</td>
<td>9</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>90</td>
<td>81</td>
<td>73</td>
<td>66</td>
<td>59</td>
<td>35</td>
<td>21</td>
<td>12</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>85</td>
<td>72</td>
<td>61</td>
<td>52</td>
<td>44</td>
<td>20</td>
<td>9</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>80</td>
<td>64</td>
<td>51</td>
<td>41</td>
<td>33</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>75</td>
<td>56</td>
<td>42</td>
<td>32</td>
<td>24</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>70</td>
<td>49</td>
<td>34</td>
<td>24</td>
<td>17</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Data Profiling – Missing Data

- Remove Missing Data
- Algorithms Adapted for Missing Data
- Impute Missing Values
Data Profiling – Missing Data

Algorithms Adapted for Missing Data

Tree-Based

MARS

k-NN/Kernel
Data Profiling – Missing Data

*ad hoc*
Methods for Missing Data
Imputation

- Dropping Variables
- Listwise/Pairwise Deletion
- Mean / Median / Mode Substitution
- Non-Stochastic Regression
- Last/Baseline Observation Carried Forward
- Indicator Variable Method for Missing Category

**NOT RECOMMENDED**
Data Profiling – Missing Data with Indicator Variable

\[
X' = \begin{cases} 
X & \text{when data are not missing} \\
\text{const} & \text{when data are missing}
\end{cases}
\]

\[
I = \begin{cases} 
0 & \text{when data are not missing} \\
1 & \text{when data are missing}
\end{cases}
\]

\[
Z = \{ \text{set of additional predictors} \}
\]

**True Model**

\[
Y = 1.0 \times X + 1.0 \times Z + \text{error}
\]

**Estimated Model**

\[
Y = a_x \times X' + a_i \times I + a_z \times Z + \text{error}
\]
### Data Profiling – Missing Data with Indicator Variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Full Data Method</th>
<th>Listwise Method</th>
<th>Indicator Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>0.97</td>
<td>0.94</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>1.02</td>
<td>1.05</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>—</td>
<td>—</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

**Z parameter is overestimated**  
**X parameter is underestimated**
Data Profiling – Missing Data

**Advanced Methods for Missing Data Imputation**

- Multiple Imputation
- Maximum Likelihood
- Expectation-Maximization
- Bayesian
- Generative Adversarial Imputation Networks (GAINs)

RECOMMENDED
Data Profiling – Mechanism of GANs

Real Images

Noise

Discriminator

Generator

real

fake
Data Profiling – Missing Data Imputation with GAINs
## Data Profiling – When Not To Impute Missing Data

<table>
<thead>
<tr>
<th>Patient_ID</th>
<th>Sick</th>
<th>Symptom A</th>
<th>Symptom A NA</th>
<th>Symptom B</th>
<th>Symptom B NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>No</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient_ID</th>
<th>Sick</th>
<th>Symptom A Yes</th>
<th>Symptom A NA</th>
<th>Symptom B Yes</th>
<th>Symptom B NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Conclusions

• The cost of poor data quality is high
• Professionalism considerations
• Data quality is not a spectator sport
• Multiple considerations for handling missing data
• **Quality** Data + Analytics leads to success!
Questions
Join Us for the Next APEX Webinar

Thursday, June 11
2:00 p.m. ET
Registration is Open

Medical Professional Liability:
State of the Market in 2020

Tim Mosler
Nick Alicea
Final Notes

• We’d like your feedback and suggestions
  • Please complete our survey

• For copies of this APEX presentation
  • Visit the Resource Knowledge Center at Pinnacleactuaries.com
Thank You for Your Time and Attention

Greg Frankowiak
gfrankowiak@pinnacleactuaries.com

Radost Wenman
rwenman@pinnacleactuaries.com