Onboard and Wayside Performance Detectors

EDA

1st Step of Data Analysis

or

RB’s Data Analysis Rule #1

RB Wiley
Principal Investigator

Tuesday October 1, 2013
Exploratory Data Analysis (EDA)

♦ What is it?

♦ What does it have to do with rail vehicle performance data?

♦ Why is it so important

♦ Who cares?

♦ YOU DO!
EDA: What is it?

♦ ‘Exploratory Data Analysis (EDA) is an **approach**/philosophy for data analysis that employs a **variety of techniques** (mostly **graphical**).’ NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 2013.

- ‘EDA … **postpones the usual assumptions** … allowing the **data itself to reveal** its underlying structure and model’

- ‘…main role of EDA is to **open-mindedly explore**, … enticing the data to reveal its structural secrets, … **gain some new, often unsuspected, insight into the data**…’

- ‘EDA makes heavy use of the human’s **natural pattern-recognition capabilities**…’
EDA: What is it? (cont)

♦ EDA sometimes described as an alternative to:
  ● Classical data analysis
  ● Bayesian data analysis

♦ All three approaches start with a problem; have data analysis; & end with a conclusion

♦ RB’s Rules of Data Analysis
  ● 1st: EDA (the *objective* data analysis phase)
  ● 2nd: following a thorough EDA, recheck the EDA
  ● 3rd: after EDA (the *subjective* data analysis phase)

Wikipedia: Exploratory Data Analysis
EDA: What is it? (cont)

♦ Maximize insight into the data...
  ● Often, 80% of the questions of interest are answered in EDA

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Cumulative Count</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loaded</td>
<td>4075</td>
<td>4075</td>
<td>49.7679</td>
<td>49.7680</td>
</tr>
<tr>
<td>Empty</td>
<td>4113</td>
<td>8188</td>
<td>50.2321</td>
<td>100.0000</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>8188</td>
<td>0.0000</td>
<td>100.0000</td>
</tr>
</tbody>
</table>

Distribution charts include:
- Frequency tables
- Descriptive statistics
- Distribution characteristics
  - Parameters (median, mean, SD, quartiles)
  - Histograms
  - Box & Whisker Plots
  - Runs plots (scatterplots)

John Tukey: Father of EDA (& boxplots): 1977
EDA: What is it? (cont)

- Maximize insight into the data...
- Detect outliers, trends

- Distribution characteristics
  - Histograms
  - Box & Whisker Plots
  - Runs plots (scatterplots)
  - Probability plots (P-P plots)
  - Quantile plots (Q-Q plots)

Statsoft: Online Statistical Textbook: Box & Whisker Plots
EDA: What is it? (cont)

- Maximize insight into the data...
- Detect outliers and anomalies
- Identify Structure (subgroups)
  - Distribution characteristics
  - Categorized Histograms
  - Compound Box & Whisker Plots
  - Categorized Scatterplots
  - Categorized P plots
  - Categorized Q plots

www.isixsigma.com : Tools - Graphical Analysis
EDA: The Most Important Data Analysis

Histogram of R-L Imbalance
Unstacked.sta 36v*8188c

www.isixsigma.com : Tools - Histograms
Box Plot of R-L Imbalance grouped by axle_no; categorized by lead_end and Empty/Loaded

Unstacked.sta 36v*8188c
EDA: Why should I care?

♦ Nobody wants to know just half of the truth

♦ No surprises!

♦ No Oops!

♦ No Gotchas!

EDA: Carnegie-Mellon University - textbook
EDA: Cal-Berkeley - © 2011 Sage Publications
EDA, A Primer... Purdue University
QUESTIONS?

Roy Brooks (RB) Wiley, Principal Investigator, Engineering Services

Transportation Technology Center, Inc. Association of American Railroads

55500 D.O.T. Rd
Pueblo, CO 81001

rb_wiley@aar.com

(719) 584-0590 phone/fax
(719) 250-5299 wireless