Understanding and Communicating Multimodal Transportation Data

First Edition

Development, Deployment, and Assessment of a New Educational Paradigm for Transportation Professionals and University Students (A Collaboration of the Region X Transportation Consortium)

by Chris Monsere

published by
Pacific Crest
Plainfield, IL
Table of Contents

Table of Contents ........................................................................................................... iii
Preface .................................................................................................................................................................................. ix

Chapter 1  Principles of Scientific Graphical Display ................................................................. 1
  Activity #1  Principles of Graphics ......................................................................................... 3
  Activity #2  Overview of the Datasets Available for the Class .............................................. 5
  Activity #3  An Experiment in Graphical Perception ............................................................. 7
  Activity #4  Critiquing a Graphic for Graphicacy ................................................................. 11

Chapter 2  Getting Started with Data ............................................................................................ 13
  Activity #5  Setting up Your Accounts .................................................................................. 15
  Activity #6  Creating a Simple Database – An Excel Strawman ........................................ 17
  Activity #7  Overview of SQL ................................................................................................. 21
  Activity #8  Creating a Simple Database – Now with PostgreSQL ................................. 23
  Activity #9  Simple SQL ........................................................................................................ 29

Chapter 3  Introduction to R ...................................................................................................... 33
  Activity #10  An Introduction to R and R Studio ................................................................. 35
  Activity #11  Setting Up R .................................................................................................... 37
  Activity #12  A Starting Point – Some Simple R ................................................................. 39
  Activity #13  Reading in Data Files ....................................................................................... 45
  Activity #14  R Plots ................................................................................................................ 49
  Activity #15  Learning Some Simple Plotting Features of R ................................................ 51
  Activity #16  Your First Advanced Plot .............................................................................. 67
  Activity #17  Code Sharing .................................................................................................... 69
  Activity #18  Thinking like a Computer – Pseudo-coding and Functions........................ 71
  Activity #19  Write Your Own Function ............................................................................... 79
  Activity #20  Connecting to the Class Database via RODBC and PostgresSQL Drivers .... 81
  Activity #21  Using R with PostgresSQL ............................................................................ 83
  Activity #22  Packages .......................................................................................................... 89
  Activity #23  Working with Time in R ................................................................................. 91

Chapter 4  Using Graphics for Exploratory Data Analysis .......................................................... 97
  Activity #24  Interactive Review of Basic Statistics Using R ............................................... 99
  Activity #25  Basic Charts for Single Discrete Variable ....................................................... 103
  Activity #26  Exploring Single Discrete Variable Plots ......................................................... 105
<table>
<thead>
<tr>
<th>Activity #27</th>
<th>Exploratory and Diagnostic Plots for the Distribution of a Single Continuous Variable</th>
<th>108</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity #28</td>
<td>Probability Distributions</td>
<td>111</td>
</tr>
<tr>
<td>Activity #29</td>
<td>Kernel Density Estimates and Histograms</td>
<td>113</td>
</tr>
<tr>
<td>Activity #30</td>
<td>Diagnosing a Distribution</td>
<td>117</td>
</tr>
<tr>
<td>Activity #31</td>
<td>Depicting the Distribution Involving Discrete Variables</td>
<td>123</td>
</tr>
<tr>
<td>Activity #32</td>
<td>Depicting the Distribution of Two Continuous Variables</td>
<td>125</td>
</tr>
<tr>
<td>Activity #33</td>
<td>Introduction of Final Project Topics</td>
<td>127</td>
</tr>
<tr>
<td>Activity #34</td>
<td>One and Two Sample Tests</td>
<td>129</td>
</tr>
<tr>
<td>Activity #35</td>
<td>Exploring Confidence Intervals and Simple Hypothesis Testing</td>
<td>131</td>
</tr>
<tr>
<td>Activity #36</td>
<td>An Application of Hypothesis Testing</td>
<td>139</td>
</tr>
</tbody>
</table>

**Chapter 5** Data Exploration for Understanding | 143 |
| Activity #37 | Advanced Multivariate Continuous Displays and Diagnostics | 145 |
| Activity #38 | Introduction to Random Sampling | 155 |

**Chapter 6** Putting it All Together: An Independent Structured Analysis | 157 |
| Topic 1 | Assessing The Accessibility Of Trimet Bus Stops | 159 |
| Topic 2 | Assessing Trimet Bus Headway Reliability | 161 |
| Topic 3 | Bicycle Performance | 163 |
| Topic 4 | Freeway Data and Incidents | 165 |
| Topic 5 | Freeway Data and Weather | 167 |
| Topic 6 | WIM Data – Side By Side Loadings | 169 |

**Appendix** Dataset Narratives | A1 |

**List of Figures**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Original Cleveland and McGill experiment stimuli (Cleveland and McGill, 1986)</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Placement of observation values on the Answer Log</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Grading rubric for a peanut butter and jelly sandwich</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Sample Database</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Excel Screen Capture of Stop Table</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Exporting the Excel file to a CSV format</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Screen capture of the phpPgAdmin login screen</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Screen capture of phpPgAdmin database screen</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Screen Capture of the phpPgAdmin add table screen</td>
</tr>
<tr>
<td>Figure 10</td>
<td>phpPgAdmin table browser screen capture</td>
</tr>
</tbody>
</table>
Figure 11  phpPGAdmin SQL window screen capture .......................................................... 30
Figure 12  R Studio Interface ............................................................................................... 37
Figure 13  Screen Capture of R Studio (with R Script File open) ....................................... 40
Figure 14  Screen capture of table view of data set in RStudio ........................................... 46
Figure 15  Plot of trimet$stop_time vs trimet$est_load .................................................... 52
Figure 16  Plot of trimet$est_load ....................................................................................... 52
Figure 17  Plot of trimet$service_day ................................................................................ 53
Figure 18  Load by time of day, plot type="l" ...................................................................... 54
Figure 19  Load by time of day, x and y labels and title added ............................................ 54
Figure 20  Load by time of day ............................................................................................ 56
Figure 21  Plot of trimet$stop_time vs trimet$est_load with modified x-axis (0,25) .............. 56
Figure 22  Plot of trimet$stop_time vs trimet$est_load with modified x-axis (14, 18) and y-axis limits (0,30) ...................................................................................... 56
Figure 23  Line Type and Symbol Type (from Murrell, R Graphics) ..................................... 57
Figure 24  Selection of Color Palettes (from Maindonald & Braun, Data Analysis and Graphics Using R) ........................................................................................................... 58
Figure 25  Plot of trimet$stop_time vs trimet$est_load pch=16 and col="red" with a 10% transparency value………………………………………………………………………………. 58
Figure 26  Default color palettes in R, here shown with n=16 elements ................................. 59
Figure 27  Description of margin layout surrounding the plot region .................................. 60
Figure 28  Plot of trimet$stop_time versus trimet$dwell with 2 row and 2 column layout ............................................................................................................................... 61
Figure 29  Plot of trimet$stop_time versus trimet$dwell with 2 row and 3 column layout ............................................................................................................................... 61
Figure 30  Plot of trimet$stop_time versus trimet$dwell with 2 row and 3 column layout with ordering completed down columns .................................................................................. 62
Figure 31  Plot of trimet$stop_time vs trimet$dwell and trimet$load ................................... 63
Figure 32  Plot of s_plot$stop_time vs. s_plot$dwell with third dimension of sched_status differentiated by color ................................................................. 64
Figure 33  Plot of trimet$stop_time vs trimet$dwell with schedule_status as 3rd dimension ............................................................................................................................... 65
Figure 34  Plot of trimet$stop_time vs trimet$dwell with schedule_status as 3rd dimension ............................................................................................................................... 66
Figure 35  Plot outputs from sample loop code ..................................................................... 72
Figure 36  Headways at Stop ID 2107, March 8, 2007 ............................................................ 79
Figure 37  ODBC Data Source Administrator pop-up window .................................................. 81
Figure 38  PostgreSQL ANSI driver pop-up window ................................................................. 82
Figure 39  Num Recs 19168 Avg GVW 59.9374478326019 ......................................................... 84
Figure 40  Three barplots ........................................................................................................ 85
Figure 41  From SQL, avg() with GROUP BY and From R, tapply .......................................... 86
Figure 42  SQL and R comparison plots .................................................................................. 87
Figure 43  Plots showing R time ............................................................................................. 95
Figure 44  Volume on SB OR-217 at OR-10 2:00PM-6:00PM in five minute intervals for a seven day period ........................................................................... 96
Figure 45  Side by Side box plots of axle space 1(left) and 2(right) ....................................... 101
Figure 46  Box plots of axle space 2 by station ....................................................................... 101
Figure 47  incidentid (first plot frames) .................................................................................. 106
Figure 48  Incident type stripplot ........................................................................................... 107
Figure 49  Sample Plots of incident type ................................................................................. 108
Figure 50  Axle 1, 2 and 3 (Picture: Andrew Nichols) ............................................................. 113
Figure 51  Histograms of Station 8, Axle 1 Weight in Kips, August 2009 ............................... 114
Figure 52  KDE plots of Station 8, Axle 1 Weight in Kips, August 2009 ................................. 115
Figure 53  Normal distributions with varying means and equal standard deviations (left) and varying standard deviations and equal means (right) ........................................ 118
Figure 54  Normal distribution density function plot (left) and cumulative density function plot (right) .................................................................................................................. 119
Figure 55  (top left to right) Scatter plot of random numbers generated for sample n=200, histogram of same random number generated sample with KDE overlaying histogram, boxplot of same data, and empirical cumulative distribution function of the sample. (bottom left to right) Theoretical distribution overlain by KDE, and theoretical cumulative distribution function overlain by empirical distribution function .......................................................................................... 120
Figure 56  Quantile-Quantile plot of the random number generated sample n=200 ............. 120
Figure 57  Summary Plots Comparing the Distributions .......................................................... 132
Figure 58  Q-Q Plots of the Distributions ................................................................................. 132
Figure 59  Plots of PDF and CDF for t-distribution with dof=249 ........................................ 133
Figure 60  Plot of Mean and 95th Percentile Confidence Interval .......................................... 134
Figure 61  Z and t distributions with varying degrees of freedom ......................................... 136
Figure 62  Scatterplots of speed versus volume (left) and speed versus occupancy (right) ...... 145
Figure 63  Scatterplot with third dimension as color .............................................................. 146
Figure 64  Scatterplot with third dimension as size ............................................................... 147
Figure 65  Scatterplot with third dimension as second axis ................................................... 148
Figure 66  Scatterplot - overlay points with transparency .................................................. 149
Figure 67  Scatterplot with sunflower overlay describing multiplicity of data points .......... 151
Figure 68  Bivariate plots of volume and speed from the loop dataset ............................... 151
Figure 69  Graphic of speed and volume data from loop dataset  
          using xyplot() in the lattice package ........................................................................ 152
Figure 70  Filled contour plot of volume versus speed  
          with third dimension of kernel density estimation ...................................................... 153
Figure 71  Random sample graphics of p-values for 100 t-tests ........................................ 155
Figure 72  Bicycle Performance dataset information .......................................................... 163
This course will introduce students to appropriate research methods for using transportation data sets and communicating the results of their work to a broad audience. The course content includes:

(a) selections of the appropriate graphical method (making knowledge-based decisions on selections for best perceptions)
(b) managing, extracting, and filtering large-scale data
(c) understanding types and dimensions of data (time resolution, discrete, continuous, and aggregations)
(d) techniques for visualizing data and exploratory analysis
(e) basic statistical analysis applied to transportation problems (public transportation, traffic, safety, freight, bicycle performance) using open-source script-based statistical tools (R) and databases (PostgresSQL)
(f) selection of appropriate analysis technique
(g) presentation of material in a technical summary

This is a gateway course; the knowledge gained in this course will be applied throughout the remaining graduate curriculum.

Students taking this course will have had an introductory transportation course, an undergraduate course in statistics and probability, and an engineering problem solving course with an exposure to programming logic. Three audiences are envisioned for this course:

(a) Graduate-level civil engineering students with an emphasis in transportation, in their first quarter.
(b) Transportation professionals with a desire to expand their knowledge of data analysis
(c) Advanced senior undergraduate civil engineering students with necessary skills and permission of the instructor.

The course uses the open source language R. Use of the PostgreSQL database will require comfort with various computing platforms (Unix, Windows) including the installation of software, downloading and installing web-based technologies.

The long-term behaviors, roles, and way of being will be supported by this course:

(a) Problem solver
(b) Researcher
(c) Communicator
(d) Collaborator
(e) Open-source software
   i. R
   ii. PostgreSQL

Reference books (required)

i. Keen, Kevin. Graphics for Statistics and Data Analysis with R
iii. Scientific Approaches to Transportation Research Volumes 1 and 2, web book
In this activity textbook, each activity includes an overview, a description of the task, a description of the deliverable, and the assessment method. Activities are also shown as in class or out of class. The following structure will be used to assess activities:

1. **Participation Activities**
   a. These activities require quick assessment and feedback. You will receive credit for completing these activities.

2. **Annotated Code Activities**
   a. In these activities you be asked to only submit a script or code file that contains comments and demonstrates active exploration of the objectives within the activity.

3. **Peer Assessment Activities**
   a. Some activities will require you to assess the work of your fellow students. In these activities, your performance will be based on your work assessed by the instructor, your feedback to peers, and your peers’ assessment of your work.

4. **Short Response Activities**
   a. Many activities are structured such that you respond to a set of questions. You will receive credit both for completing these activities and for the depth and detail of your responses. We will attempt electronic submittal and feedback for these activities.

5. **Discovery Activities**
   a. These activities require you to build on knowledge and skills introduced to you in previous activities. These activities will be open ended and you will receive credit for completing these activities and for the creativity of your exploration.

There will also be a final project which is an independent structured analysis which you will select from a set of open-ended questions devised by the instructor presented in Chapter 6. This project will serve as the final assessment that the student has made progress in developing knowledge and skills in this class. The project is due during the final exam period, where students will make a brief presentation on their results to the class. You are encouraged to make an early selection of the project topic to begin your work in advance.

A number of people have contributed to make this version of the course document. The early version of this course and activities were developed by the primary author, Chris Monsere. Contributions include those from Ashley Haire, Chengxin Dai, and Joel Barnett (who did a lot of work doing the final editing of the course design document. This workbook benefited from the collaboration and input of Michael Kyte and Steve Berylein, University of Idaho, Kelly Pitera, University of Washington, Shane Brown, Washington State University, and Ming Lee, University of Alaska. The work was funded by FHWA TDEDP program. All errors and omissions are the responsibility of the primary author. Robert Bertini initiated Portland State’s collaboration on this project.