The Power of ‘R’

Building Visualisations of Data with ‘R’
Agenda

• Introduction
  • Bio
  • What is ‘R’?
• Why ‘R’?
• Data Handling in ‘R’
• Data Visualisation
• Case Studies
• Latex (RSweave)
What is ‘R’?

- Open source Software language (1997)
- Aids Developing Statistical software
- Tool for advanced Analytics and Visualisation
- Objected oriented programming
- S - language
Boom in Data Analytics

- Internet boom
- Electronic capturing of data
- Open source/Sharing/Development
- Scale - Economical
- Cloud computing
Why ‘R’?

- Extremely easy to use - Automatically detects type of data
- Open source
- Handles complex tasks
- Parallel Processing
- Supported by cloud computing tools [Spark(R-Spark) etc]
Installing ‘R’

- Download and Install
  - http://cran.us.r-project.org/
- Load basic packages
- You’re Ready to Go!
- Install packages as and when required
Data Handling in ‘R’

• Set working directory
• Read files (into ‘R’)
• Use the functions (or write them)
• Environment variables, etc
• If you need help, type ?/??
  • Demonstration
Why is Visualising important?

> datasets::anscombe

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>y1</th>
<th>y2</th>
<th>y3</th>
<th>y4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>8.04</td>
<td>9.14</td>
<td>7.46</td>
<td>6.58</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>6.95</td>
<td>8.14</td>
<td>6.77</td>
<td>5.76</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>7.58</td>
<td>8.74</td>
<td>12.74</td>
<td>7.71</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8.81</td>
<td>8.77</td>
<td>7.11</td>
<td>8.84</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>8</td>
<td>8.33</td>
<td>9.26</td>
<td>7.81</td>
<td>8.47</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>8</td>
<td>9.96</td>
<td>8.10</td>
<td>8.84</td>
<td>7.04</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>7.24</td>
<td>6.13</td>
<td>6.08</td>
<td>5.25</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>19</td>
<td>4.26</td>
<td>3.10</td>
<td>5.39</td>
<td>12.50</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>8</td>
<td>10.84</td>
<td>9.13</td>
<td>8.15</td>
<td>5.56</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>4.82</td>
<td>7.26</td>
<td>6.42</td>
<td>7.91</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>5.68</td>
<td>4.74</td>
<td>5.73</td>
<td>6.89</td>
</tr>
</tbody>
</table>
Why is Visualising important?

Are the 4 datasets the same?

<table>
<thead>
<tr>
<th>Anscombe's quartet</th>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean of x in each case</td>
<td>9 (exact)</td>
</tr>
<tr>
<td></td>
<td>Sample variance of x in each case</td>
<td>11 (exact)</td>
</tr>
<tr>
<td></td>
<td>Mean of y in each case</td>
<td>7.50 (to 2 decimal places)</td>
</tr>
<tr>
<td></td>
<td>Sample variance of y in each case</td>
<td>4.122 or 4.127 (to 3 decimal places)</td>
</tr>
<tr>
<td></td>
<td>Correlation between x and y in each case</td>
<td>0.816 (to 3 decimal places)</td>
</tr>
<tr>
<td></td>
<td>Linear regression line in each case</td>
<td>$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.1</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
<td>6.6</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.1</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
<td>5.8</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.7</td>
<td>13</td>
<td>12.7</td>
<td>8</td>
<td>7.7</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.8</td>
<td>9</td>
<td>7.11</td>
<td>8</td>
<td>8.8</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.3</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
<td>8.5</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
<td>14</td>
<td>8.84</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.1</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
<td>5.3</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.1</td>
<td>4</td>
<td>5.39</td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>10.8</td>
<td>12</td>
<td>9.1</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
<td>5.6</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.3</td>
<td>7</td>
<td>6.42</td>
<td>8</td>
<td>7.9</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.7</td>
<td>5</td>
<td>5.73</td>
<td>8</td>
<td>6.9</td>
</tr>
</tbody>
</table>
Types of Graphs - Histogram
Types of Graphs - Bar Graphs

- Bar Graphs

1. Factor (cyl): 8, 6, 4
   - Count: 0, 5, 10

2. Factor (cyl): 4, 6, 8
   - Count: 0, 5, 10

3. Factor (gear): 3, 4, 5
   - Count: 0, 5, 10
Types of Graphs - Line Graphs
Types of Graphs - Line Graphs

![Line Graph 1](#)

![Line Graph 2](#)

![Line Graph 3](#)

![Line Graph 4](#)
Types of Graphs - Box Plots
Types of Graphs - Scatter Plots

- Scatter Plot
- 3D Scatterplot
Types of Graphs - Scatter Plots

Three Cylinder Options

- mpg
- disp
- drat
- wt
Resources - ‘R’

- http://stats.stackexchange.com/questions/138/free-resources-for-learning-
- https://cran.r-project.org/doc/manuals/R-intro.pdf
- http://tryr.codeschool.com
- https://www.tutorialspoint.com/r/
- Practice datasets - https://data.gov.in
Resources - Visualisation


