Cleaning up Data Visually using R

(Sanity Check before running statistical analysis.)

Import data into R

\[
> \text{file.name} \leftarrow \text{file.choose()}
\]

\[
> \text{my.data} \leftarrow \text{read.csv(file.name, row.names=NULL)}
\]

My.data
This is the check basic things about the data

Class Example: Case Study: 8 undergrads who used the mouse and leap motion control to create a circle; Between subject Design.

Sample Data

<table>
<thead>
<tr>
<th>Participant</th>
<th>Interface</th>
<th>Time Taken</th>
<th>Contact Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mouse</td>
<td>68</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Mouse</td>
<td>71</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Mouse</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Mouse</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Leap</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>Leap</td>
<td>47</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Leap</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>Leap</td>
<td>56</td>
<td>0</td>
</tr>
</tbody>
</table>

Independent Variable – 2 levels of interface: Mouse, Leap

Dependent Variable – Time Taken

Explore by looking at the names of the fields and the dimension

- For a between subjects make sure Numbers of rows = number participants
- For Within subjects make sure number of rows-number participants or number of participants times number of conditions

Sanity check to make sure all data correct

1. Check Overall mean and standard deviation

- Mean of the time taken- mean(my.data [,3]) 57.626
- Std Deviation- sd (my.data [,3]) 16.03512
- Min – min(my.data [,3] 31
- Max – max(my.data [,3] 80

Distribution of the data
Visually check to see if it is normally distributed using a histogram

**Histogram** - a graphical representation of the distribution of data. It is an estimate of the probability distribution of a continuous variable

Look at the histogram –
- frequency of a particular value
- is data approximately normally distributed
- range is well covered

If data is not normal then something is quirky about the data. Must look at the data once more.

*Question from student:* Binary response of data – 1,0 (yes, no) data will not have normal distribution because it is not continual data; Look at if there is roughly an equal number of yes/no, if there are too many yes’s- it may be too simple or too hard.

**II. Dependent Variable:** Time taken for each condition

```
hist(my.data[1:4, 3]
```
this is histogram of rows 1-4, column 3 (time taken)

![Histogram of my.data[4]](attachment:image.png)

Repeat this analysis for every condition
The table will visually show if there are differences in the test conditions.

**Boxplot**

Another visual
The box plot is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum

![Boxplot Example]

Prepare a box plot for each condition to compare significant results

**Outliers**
- Outliers mean data points is not from same data points
- More than 1.5 times away are considered outliers
- Set the definition of parameters to show outliers
- Should not have too many outliers
- Need a way to handle outliers; done by case by case basis
  - Class Example: User took longer than other and should not be considered
  - Another Example: User did not understand the instructions and did opposite of what was expected
Another Example: Case Study: 8 undergrads who used the arrow and dart to hit a target Between subject Design.

Sample Data

<table>
<thead>
<tr>
<th>Participant</th>
<th>Interface</th>
<th>Attempts</th>
<th>Contact Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>arrow</td>
<td>136</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>arrow</td>
<td>142</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>arrow</td>
<td>126</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>arrow</td>
<td>160</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>dart</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>dart</td>
<td>94</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>dart</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>dart</td>
<td>112</td>
<td>0</td>
</tr>
</tbody>
</table>

Independent Variable – 2 levels of interface: Arrow, Dart
**Dependent Variable** – Attempts

- Mean of the time taken - mean(my.data[,3]) 115.252.626
- Standard Deviation - sd (my.data[,3]) 16.03512
- Min – min(my.data[,3] 62
- Max – max(my.data [,3] 160

**Figure 3:** histogram of second example

**Figure 4:** Boxplot of second example

**Table of second example:**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean</th>
<th>SD</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: Arrow</td>
<td>141</td>
<td>14.28286</td>
<td>160</td>
<td>126</td>
</tr>
<tr>
<td>II: Dart</td>
<td>89.5</td>
<td>20.6801</td>
<td>112</td>
<td>62</td>
</tr>
</tbody>
</table>
Figure 5: Boxplot of Arrow data

Figure 6: Boxplot of Dart data