2 Data Wrangling with R

Exercise 1

Read in the "Child_Variants.csv" file into a variable called "child".

- Create a new variable called "Type" containing the value "SNP" if both "REF" and "ALT" contain only one letter (you can use the `nchar()` function for this), and "INDEL", otherwise. Overwrite the original variable.
- Find genes which have at least three novel SNPs in them and calculate their average "COVERAGE". This means that
  - The "Type" of the observation must be "SNP".
  - The observation should not have a dbSNP identifier (it should be a dot).

Exercise 2

Load the contents of "gadata_example_2.csv" into a tidy data structure.

- Calculate the average pageviews per day by channel.
- Add two new variables: page views per session ("pvs_per_session") and the bounce rate ("bounce_rate"), the number of bounces per entrance.
- Calculate the average bounce rate by device category ("deviceCategory").

3 Data Visualization with R and `ggplot2`

- Unless explicitly stated, please use the `ggplot` function in the `ggplot2` package.
- Depending on the exercise, you may need to reformat the data before plotting.

Exercise 1

1. Read the contents of `brain_bodyweight.txt` into a data frame.
2. Create a scatter plot with the body weight on the x-axis and the brain weight on the y-axis.
3. Use the `geom_text_repel()` function in the `ggrepel` package to add non-overlapping labels indicating the species.

Exercise 2

1. Read the contents of `chr_data.txt` into a data frame.
2. Create a scatter plot with `GM06990_ABL1` on the x-axis and `GM06990_MLLT3` on the y-axis.
3. Create several plots:
   (a) Color the points blue.
   (b) Color the points according to the genomic location.
4. Create boxplots summarizing the distribution of `GM06990_ABL1` by genomic location.
5. Overlay the data points on top of the boxplots; jitter the points to avoid overlaps.
6. Create a violin plot combined with a boxplot, similar to Figure 1 (note that this is a different dataset!).

Exercise 3

Figure 1: Source: `http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization`

1. Read the contents of "gene_expression.txt" into a data frame. The table represents the expression values measured under three conditions ("control", "cond1", and "cond2") for one particular gene in cell lines derived from four different patients.
2. Recreate the graph in Figure 2.

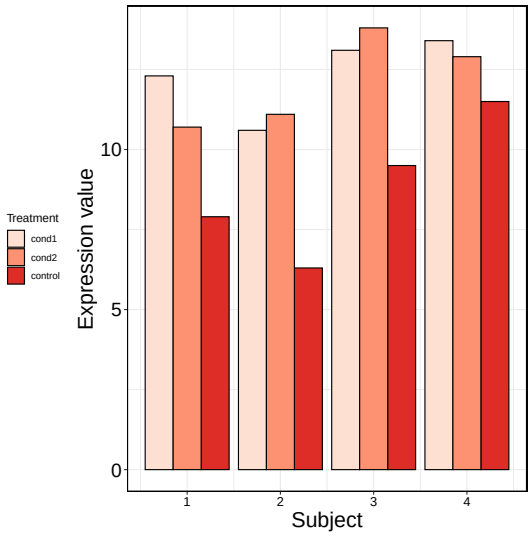

Figure 2: Gene expression values for four patients.

Exercise 4

Using the "iris" dataset (`https://en.wikipedia.org/wiki/Iris_flower_data_set`):

1. Create a scatter plot of sepal length vs. sepal width; color the points according to the species.
2. Create a data frame containing the mean and the standard error of the mean (SEM) for each species and measurement. Create a clustered bar chart visualizing each measurement for each of the four species. Add an error bar to the chart showin the SEM.
3. Instead of computing the summary statistics by yourself, let `ggplot` do it. Use the `stat_summary()` function for this purpose.

Exercise 5

1. Load the data in `mammals.txt`. This dataset is described here: `http://search.r-project.org/library/ggplot2/html/msleep.html`.
2. Produce a figure similar to Figure 3.

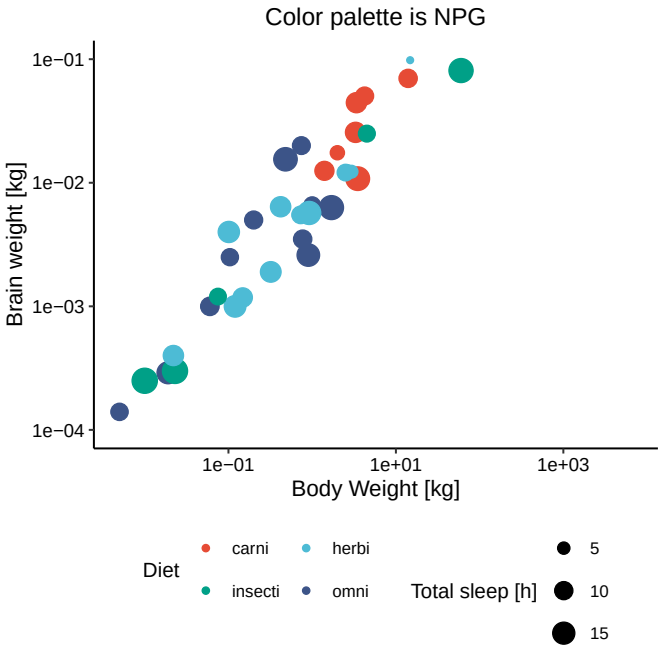

Figure 3: Sleep mammals dataset

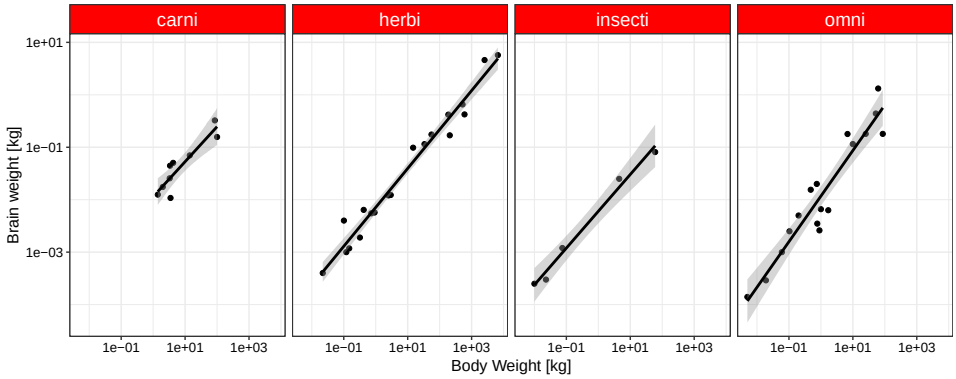

Figure 4: Sleep mammals dataset (per group)

3. Produce a figure similar to Figure 4.

4. What relationship do these data suggest for brain and body weight (`brainwt` and `bodywt`)? What does it indicate about humans? How about water opossums?