Taming Wild Data

Exercises

Data Services @ HSL

EXERCISE 1

1. Use filter to find how many plants of the versicolor species were observed with Sepal.Length >= 6.5. Answer should be 9.
2. Use filter to find the observations with Petal.Length below 1.5 followed by select to show the Petal.Width and Species for these observations.

EXERCISE 2

Use `separate` to create two new columns called 'nutrient' and 'rate' from the old 'nutrientrate' column. Think about how to separate the nutrient from the rate. Look at the help menu for `separate` by calling `?separate`

EXERCISE 3

1. Display the data where the biological process the gene plays a role in (the `bp` variable) is "leucine biosynthesis" (be careful with spelling) *and* the limiting nutrient is Leucine. (Answer should return a 24-by-7 data frame – 4 genes × 6 growth rates).
2. Display the data where the observation had high expression (in the top 1% of expressed genes). *Hint:* see `?quantile` and try `quantile(clean$expression, probs=.99)` to see the expression value which is higher than 99% of all the data, then `filter()` based on that. Try piping your answer into the `View()` function so you can see the whole thing. What does it look like those genes are doing? Answer should return a 1971-by-7 data frame.

EXERCISE 4

1. First, filter the dataframe for genes involved in the "leucine biosynthesis" biological process *and* where the limiting nutrient is Leucine.
2. Pipe the filtered result to `arrange()` where you'll arrange the result of #1 by the gene symbol.
3. Pipe this result in a `View()` statement so you can see the entire result.

EXERCISE 5

Putting it all together

1. Show the limiting nutrient and expression values for the gene ADH2 when the growth rate is restricted to 0.05. *Hint:* 2 pipes: `filter` and `select`.
2. What are the four most highly expressed genes when the growth rate is restricted to 0.05 by restricting glucose? Show only the symbol, expression value, and GO terms (bp and mf). *Hint:* 4 pipes: `filter`, `arrange`, `head`, and `select`.

3. When the growth rate is restricted to 0.05, what is the average expression level across all genes involved in the biological process == "response to stress", separately for each limiting nutrient? What about genes in the "protein biosynthesis" biological process? *Hint:* 3 pipes: `filter`, `group_by`, `summarize`.

EXERCISE 6

** If there is time **

Those were easy, right? How about some tougher ones.

1. Use `n_distinct()` within a `summarize()` call to count the number of different biological processes in the dataset
2. Which 10 biological process annotations have the most genes associated with them? What about molecular functions? *Hint:* 4 pipes: `group_by`, `summarize` with `n_distinct`, `arrange`, `head`.
3. How many distinct genes are there where we know what process the gene is involved in but we don't know what it does? *Hint:* 3 pipes; `filter` where `bp!="biological process unknown" & mf=="molecular function unknown"`, and after `select`ing columns of interest (symbol, bp, mf), pipe the output to `distinct()`. The answer should be **737**
4. When the growth rate is restricted to 0.05 by limiting Glucose, which biological processes are the most upregulated? Show a sorted list with the most upregulated BPs on top, displaying the biological process and the average expression of all genes in that process rounded to two digits. *Hint:* 5 pipes: `filter`, `group_by`, `summarize`, `mutate`, `arrange`.