Introduction to dplyr

Paul M. Magwene

What is `dplyr`?

`dplyr` is a package that provides a "grammar for data manipulation"

Key "verbs" in the `dplyr` package:

- `select()`
- `filter()`
- `mutate()`
- `arrange()`
- `summarize()`
- `group_by()`

select() subsets columns

```
names(iris)
```

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length"
[4] "Petal.Width"  "Species"
```

```
# select two columns
select(iris, Sepal.Length, Petal.Length) %>% head(3)
```

```
  Sepal.Length Petal.Length
1          5.1          1.4
2          4.9          1.4
3          4.7          1.3
```

```
# select everything BUT the species column
select(iris, -Species) %>% head(3)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
```

`select()` has some specialized functions for powerful filtering

```
select(iris, starts_with("Petal")) %>% head(3)
```

```
  Petal.Length Petal.Width
1          1.4         0.2
2          1.4         0.2
3          1.3         0.2
```

```
select(iris, ends_with("Length")) %>% head(3)
```

```
  Sepal.Length Petal.Length
1          5.1          1.4
2          4.9          1.4
3          4.7          1.3
```

`filter()` selects rows that match criteria

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa