Online CDS 101 *Introduction to Computational and Data Sciences*

Data transformation

Other helpful dplyr verbs

Other helpful `dplyr` verbs

- `transmute`: Like `mutate`, except the transformed output is placed in a new data frame
- `pull`: Extract column into the base R vector data type
- `rename`: Convenient way to change the name of a variable (column)
- `distinct`: Determines unique values across one or more columns in the dataset
- `count`: Group by category and count the number of group members

transmute example

Example from mutate lectures for computing term lengths:

```
presidential %>%
  mutate(term_length = interval(start, end) / dyears(1))
```

```
## # A tibble: 11 x 5
##    name       start      end        party      term_length
##    <chr>      <date>     <date>     <chr>            <dbl>
##  1 Eisenhower 1953-01-20 1961-01-20 Republican        8.01
##  2 Kennedy    1961-01-20 1963-11-22 Democratic        2.84
##  3 Johnson    1963-11-22 1969-01-20 Democratic        5.17
##  4 Nixon      1969-01-20 1974-08-09 Republican        5.55
##  5 Ford       1974-08-09 1977-01-20 Republican        2.45
##  6 Carter     1977-01-20 1981-01-20 Democratic        4.00
##  7 Reagan     1981-01-20 1989-01-20 Republican        8.01
##  8 Bush       1989-01-20 1993-01-20 Republican        4.00
##  9 Clinton    1993-01-20 2001-01-20 Democratic        8.01
## 10 Bush       2001-01-20 2009-01-20 Republican        8.01
## 11 Obama      2009-01-20 2017-01-20 Democratic        8.01
```

transmute example

Example from mutate lectures for computing term lengths:

```
presidential %>%
  transmute(term_length = interval(start, end) / dyears(1))
```

```
## # A tibble: 11 x 1
##    term_length
##          <dbl>
##  1        8.01
##  2        2.84
##  3        5.17
##  4        5.55
##  5        2.45
##  6        4.00
##  7        8.01
##  8        4.00
##  9        8.01
## 10        8.01
## 11        8.01
```

`pull` example

`pull` extracts a column and converts it into the base R vector data type:

```
presidential %>%
  pull(name)
```

```
##  [1] "Eisenhower" "Kennedy"    "Johnson"    "Nixon"      "Ford"
##  [6] "Carter"     "Reagan"     "Bush"       "Clinton"    "Bush"
## [11] "Obama"
```

`rename` example

`rename` lets us rename the columns in the dataset:

```
presidential
```

```
## # A tibble: 11 x 4
##    name       start      end        party
##    <chr>      <date>     <date>     <chr>
##  1 Eisenhower 1953-01-20 1961-01-20 Republican
##  2 Kennedy    1961-01-20 1963-11-22 Democratic
##  3 Johnson    1963-11-22 1969-01-20 Democratic
##  4 Nixon      1969-01-20 1974-08-09 Republican
##  5 Ford       1974-08-09 1977-01-20 Republican
##  6 Carter     1977-01-20 1981-01-20 Democratic
##  7 Reagan     1981-01-20 1989-01-20 Republican
##  8 Bush       1989-01-20 1993-01-20 Republican
##  9 Clinton    1993-01-20 2001-01-20 Democratic
## 10 Bush       2001-01-20 2009-01-20 Republican
## 11 Obama      2009-01-20 2017-01-20 Democratic
```

`rename` example

`rename` lets us rename the columns in the dataset:

```
presidential %>%
  rename(term_begin = start, term_end = end)
```

```
## # A tibble: 11 x 4
##    name       term_begin term_end   party
##    <chr>      <date>     <date>     <chr>
##  1 Eisenhower 1953-01-20 1961-01-20 Republican
##  2 Kennedy    1961-01-20 1963-11-22 Democratic
##  3 Johnson    1963-11-22 1969-01-20 Democratic
##  4 Nixon      1969-01-20 1974-08-09 Republican
##  5 Ford       1974-08-09 1977-01-20 Republican
##  6 Carter     1977-01-20 1981-01-20 Democratic
##  7 Reagan     1981-01-20 1989-01-20 Republican
##  8 Bush       1989-01-20 1993-01-20 Republican
##  9 Clinton    1993-01-20 2001-01-20 Democratic
## 10 Bush       2001-01-20 2009-01-20 Republican
## 11 Obama      2009-01-20 2017-01-20 Democratic
```

`distinct` example

`distinct` can find all the unique political parties in the party column:

```
presidential %>%
  distinct(party)
```

```
## # A tibble: 2 x 1
##   party
##   <chr>
## 1 Republican
## 2 Democratic
```

`count` example

`count` finds the number of presidents in the two political parties in this dataset:

```
presidential %>%
  count(party)
```

```
## # A tibble: 2 x 2
##   party          n
##   <chr>      <int>
## 1 Democratic     5
## 2 Republican     6
```

Note that `count` is identical to the following `group_by` and `summarize` command:

```
presidential %>%
  group_by(party) %>%
  summarize(n = n())
```

```
## # A tibble: 2 x 2
##   party          n
##   <chr>      <int>
## 1 Democratic     5
## 2 Republican     6
```

Credits

License

Acknowledgments

Ideas and examples for the dplyr demos adapted from *Modern Data Science with R* by Benjamin Baumer, Daniel Kaplan, and Nicholas Horton, chapter 4.