Advanced Data Journalism: Doing More with R

Class 1: Exploring data

Andrew Ba Tran

dplyr verbs/functions for wrangling data:

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

Importing data

```r
df <- read_csv("https://www.fema.gov/api/open/v2/DisasterDeclarationsSummaries.csv")

df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
```

```r
glimpse(df)
```

```
Rows: 63,167
Columns: 24
$ femaDeclarationString    <chr> "FM-5444-TX", "FM-5436-NE", "FM-5444-TX", "FM…
$ disasterNumber           <dbl> 5444, 5436, 5444, 5436, 5436, 5435, 5434, 543…
$ state                    <chr> "TX", "NE", "TX", "NE", "NE", "AZ", "AZ", "NM…
$ declarationType          <chr> "FM", "FM", "FM", "FM", "FM", "FM", "FM", "FM…
$ declarationDate          <dttm> 2022-07-19, 2022-04-23, 2022-07-19, 2022-04-…
$ fyDeclared               <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 202…
$ incidentType             <chr> "Fire", "Fire", "Fire", "Fire", "Fire", "Fire…
$ declarationTitle         <chr> "CHALK MOUNTAIN FIRE", "ROAD 702 FIRE", "CHAL…
$ ihProgramDeclared        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ iaProgramDeclared        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ paProgramDeclared        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ hmProgramDeclared        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ incidentBeginDate        <dttm> 2022-07-18, 2022-04-22, 2022-07-18, 2022-04-…
$ incidentEndDate          <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ disasterCloseoutDate     <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ fipsStateCode            <chr> "48", "31", "48", "31", "31", "04", "04", "35…
$ fipsCountyCode           <chr> "221", "063", "425", "065", "145", "025", "00…
$ placeCode                <dbl> 99221, 99063, 99425, 99065, 99145, 99025, 990…
$ designatedArea           <chr> "Hood (County)", "Frontier (County)", "Somerv…
$ declarationRequestNumber <dbl> 22060, 22034, 22060, 22034, 22034, 22032, 220…
$ lastIAFilingDate         <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ hash                     <chr> "373c5ec27998afc08a53302dae796f476b1a6546", "…
$ id                       <chr> "867be42a-71d5-4f13-aa21-d91e0a6fd577", "e671…
$ lastRefresh              <dttm> 2022-07-20 21:21:23, 2022-07-20 21:21:23, 20…
```

range()

```r
range(1, 4, 6, 22, 2002)
```

```
[1]    1 2002
```

```
range(df$incidentBeginDate)
```

```
[1] "1953-05-02 UTC" "2022-07-26 UTC"
```

table()

```r
table(df$state)
```

```

  AK   AL   AR   AS   AZ   CA   CO   CT   DC   DE   FL   FM   GA   GU   HI   IA
 310 1652 1593   75  333 1466  646  255   23   53 2091   31 2269   19  100 1848
  ID   IL   IN   KS   KY   LA   MA   MD   ME   MH   MI   MN   MO   MP   MS   MT
 357 1282 1451 1759 2576 2493  398  448 1013   53  796 1540 2700   63 1901  605
  NC   ND   NE   NH   NJ   NM   NV   NY   OH   OK   OR   PA   PR   PW   RI   SC
1995 1352 1485  297  625  512  273 1485 1281 2472  583 1239 1831    1  114  855
  SD   TN   TX   UT   VA   VI   VT   WA   WI   WV   WY
1405 1594 5173  249 2522   80  330  965  892 1230  128
```

count()

```
counted <- count(df, state)

counted
```

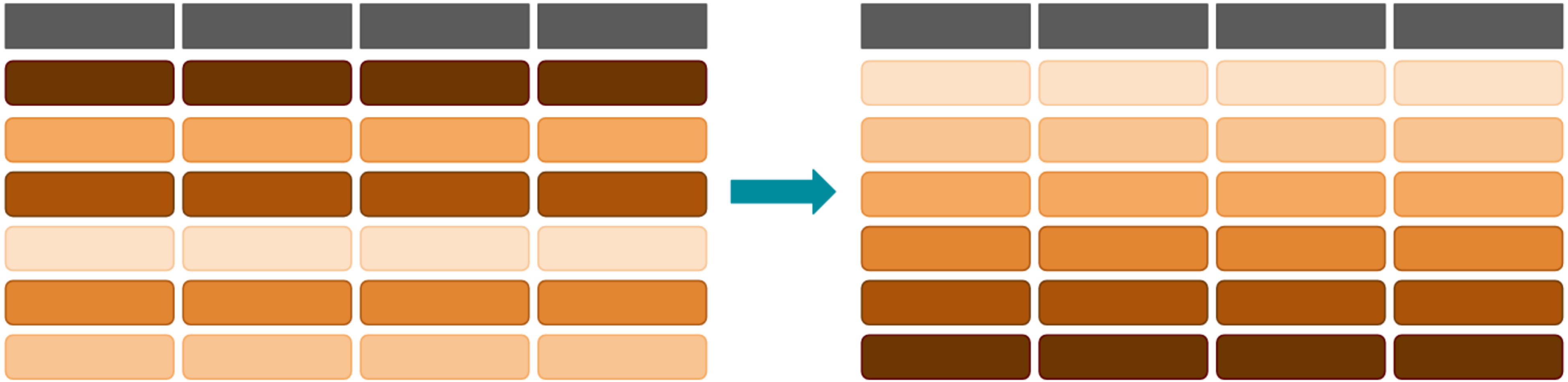
```
# A tibble: 59 × 2
   state     n
   <chr> <int>
 1 AK      310
 2 AL     1652
 3 AR     1593
 4 AS       75
 5 AZ      333
 6 CA     1466
 7 CO      646
 8 CT      255
 9 DC       23
10 DE       53
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

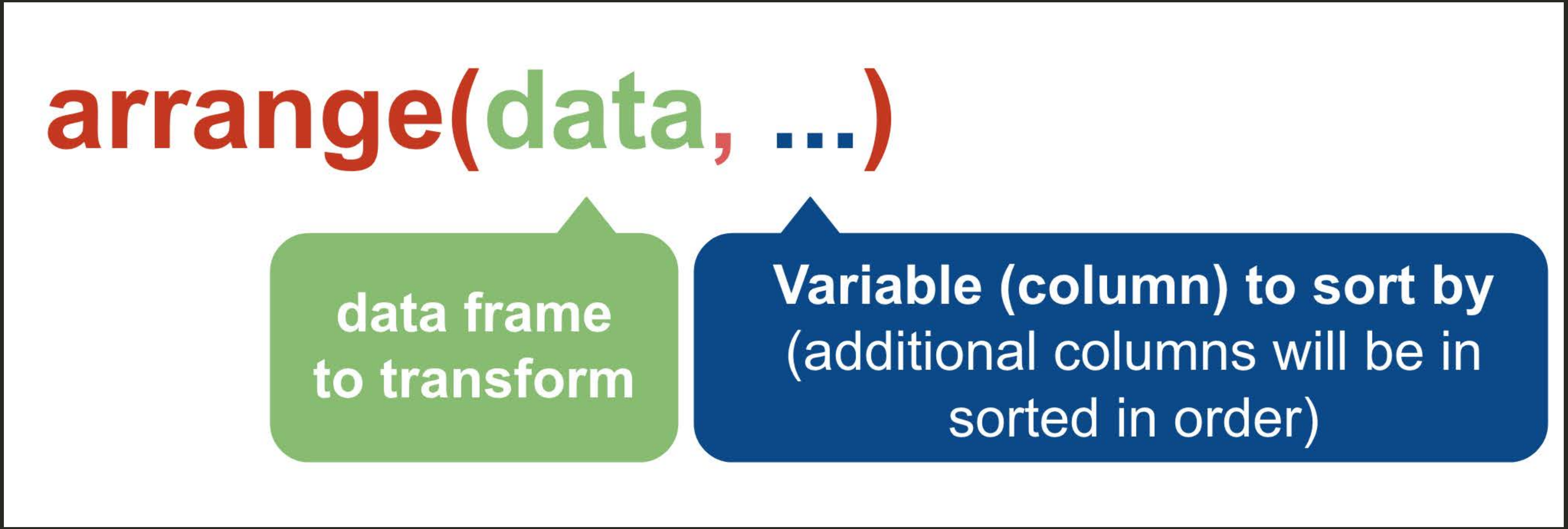
```r
counted <- count(df, state, name="disasters")

counted
```

```
# A tibble: 59 × 2
   state disasters
   <chr>     <int>
 1 AK          310
 2 AL         1652
 3 AR         1593
 4 AS           75
 5 AZ          333
 6 CA         1466
 7 CO          646
 8 CT          255
 9 DC           23
10 DE           53
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

arrange()

Reorder rows with **arrange()**





arrange(data, ...)

data frame to transform

Variable (column) to sort by
(additional columns will be in sorted in order)

```
arrange(counted, disasters)
```

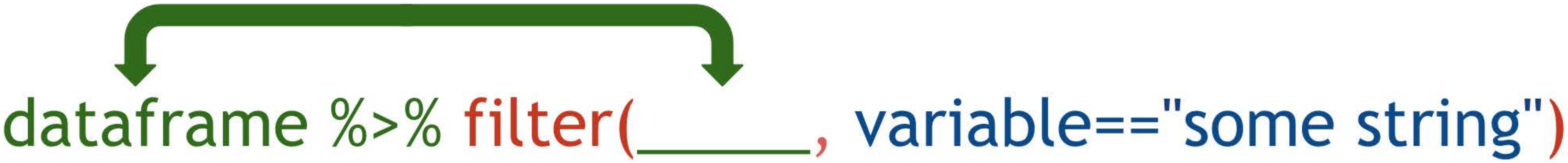
```
# A tibble: 59 × 2
   state disasters
   <chr>     <int>
 1 PW            1
 2 GU           19
 3 DC           23
 4 FM           31
 5 DE           53
 6 MH           53
 7 MP           63
 8 AS           75
 9 VI           80
10 HI          100
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
arrange(counted, desc(disasters))
```

```
# A tibble: 59 × 2
   state disasters
   <chr>     <int>
 1 TX         5173
 2 MO         2700
 3 KY         2576
 4 VA         2522
 5 LA         2493
 6 OK         2472
 7 GA         2269
 8 FL         2091
 9 NC         1995
10 MS         1901
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Pipes

%>%



```
dataframe %>% filter(______, variable=="some string")
```

```r
counted <- count(df, state, name="disasters")
sorted_count <- arrange(counted, desc(disasters))

sorted_count
```

```
# A tibble: 59 × 2
   state disasters
   <chr>     <int>
 1 TX         5173
 2 MO         2700
 3 KY         2576
 4 VA         2522
 5 LA         2493
 6 OK         2472
 7 GA         2269
 8 FL         2091
 9 NC         1995
10 MS         1901
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable na
```

```r
df %>%
  count(state, name="disasters")
```

```
# A tibble: 59 × 2
   state disasters
   <chr>     <int>
 1 AK          310
 2 AL         1652
 3 AR         1593
 4 AS           75
 5 AZ          333
 6 CA         1466
 7 CO          646
 8 CT          255
 9 DC           23
10 DE           53
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  count(state, name="disasters") %>%
    arrange(desc(disasters))
```

```
# A tibble: 59 × 2
   state disasters
   <chr>     <int>
 1 TX         5173
 2 MO         2700
 3 KY         2576
 4 VA         2522
 5 LA         2493
 6 OK         2472
 7 GA         2269
 8 FL         2091
 9 NC         1995
10 MS         1901
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

dplyr verbs/functions for wrangling data:

- **arrange()**
- filter()
- select()
- mutate()
- **summarize() (pretty much count())**
- group_by()

Advanced Data Journalism: Doing More with R

Class 1: Filtering and selecting

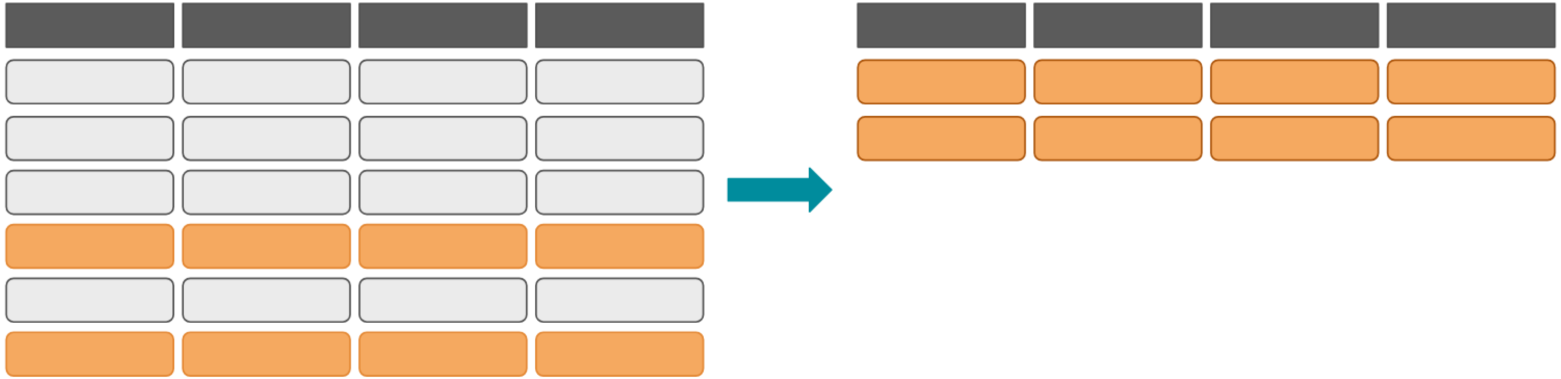
Andrew Ba Tran

dplyr verbs/functions for wrangling data:

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

filter()

Extract cases with **filter()**



You can filter based on values in a column/vector with these operators:

- `>` `<` greater than, less than
- `>=` `<=` greater than or equal to, less than or equal to
- `==` tests whether the objects on either end are equal
- `!=` not equal to
- `%in%` equals (one value match out of multiple options)

```
df <- read_csv("https://www.fema.gov/api/open/v2/DisasterDeclarationsSummaries.csv")

glimpse(df)
```

```
Rows: 63,167
Columns: 24
$ femaDeclarationString    <chr> "FM-5444-TX", "FM-5436-NE", "FM-5444-TX", "FM…
$ disasterNumber           <dbl> 5444, 5436, 5444, 5436, 5436, 5435, 5434, 543…
$ state                    <chr> "TX", "NE", "TX", "NE", "NE", "AZ", "AZ", "NM…
$ declarationType          <chr> "FM", "FM", "FM", "FM", "FM", "FM", "FM", "FM…
$ declarationDate          <dttm> 2022-07-19, 2022-04-23, 2022-07-19, 2022-04-…
$ fyDeclared               <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 202…
$ incidentType             <chr> "Fire", "Fire", "Fire", "Fire", "Fire", "Fire…
$ declarationTitle         <chr> "CHALK MOUNTAIN FIRE", "ROAD 702 FIRE", "CHAL…
$ ihProgramDeclared        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ iaProgramDeclared        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ paProgramDeclared        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ hmProgramDeclared        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ incidentBeginDate        <dttm> 2022-07-18, 2022-04-22, 2022-07-18, 2022-04-…
$ incidentEndDate          <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ disasterCloseoutDate     <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ fipsStateCode            <chr> "48", "31", "48", "31", "31", "04", "04", "35…
$ fipsCountyCode           <chr> "221", "063", "425", "065", "145", "025", "00…
$ placeCode                <dbl> 99221, 99063, 99425, 99065, 99145, 99025, 990…
$ designatedArea           <chr> "Hood (County)", "Frontier (County)", "Somerv…
$ declarationRequestNumber <dbl> 22060, 22034, 22060, 22034, 22034, 22032, 220…
```

```r
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>  
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
df %>%
  filter(incidentType=="Hurricane")
```

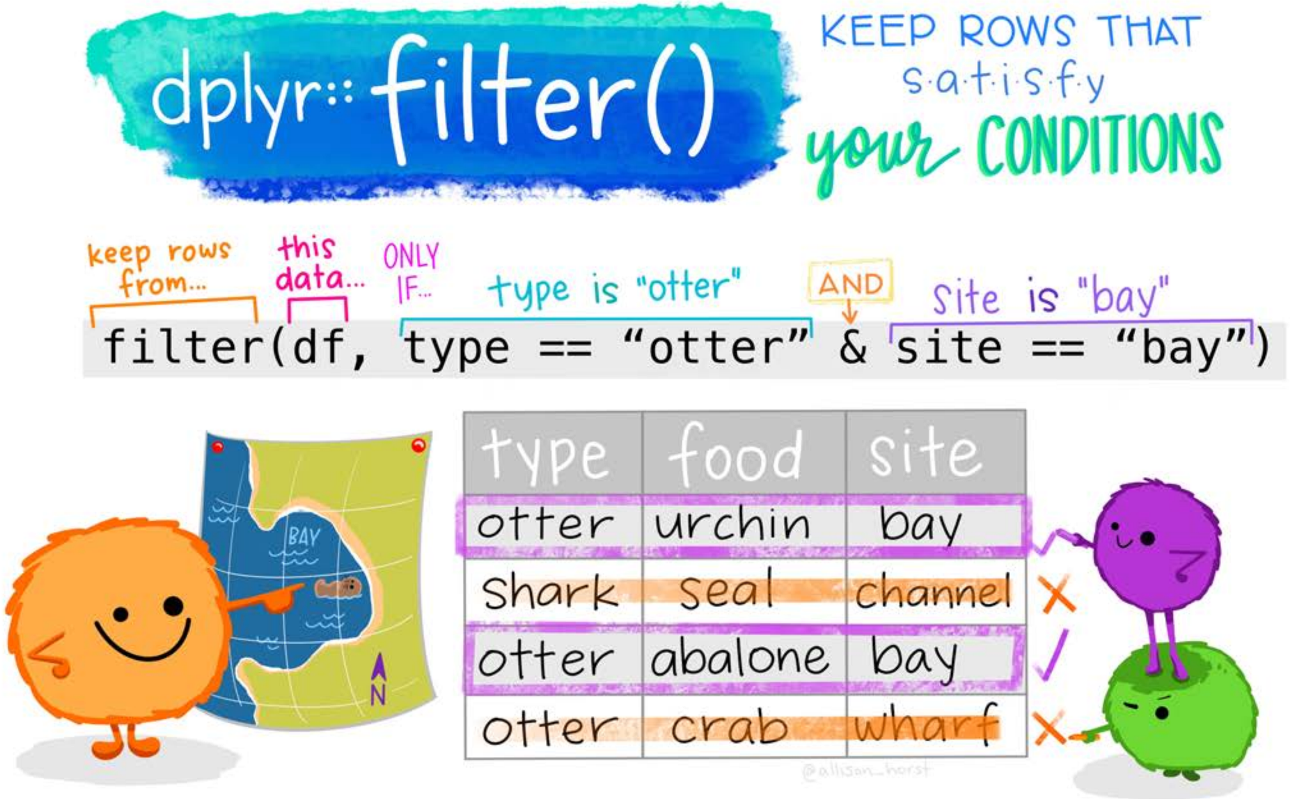
```
# A tibble: 12,489 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>  
 1 DR-4627-DE     4627 DE    DR      2021-10-24 00:00:00    2022 Hurric… REMNAN…
 2 DR-4626-MS     4626 MS    DR      2021-10-22 00:00:00    2022 Hurric… HURRIC…
 3 DR-4626-MS     4626 MS    DR      2021-10-22 00:00:00    2022 Hurric… HURRIC…
 4 DR-4626-MS     4626 MS    DR      2021-10-22 00:00:00    2022 Hurric… HURRIC…
 5 DR-4626-MS     4626 MS    DR      2021-10-22 00:00:00    2022 Hurric… HURRIC…
 6 DR-4626-MS     4626 MS    DR      2021-10-22 00:00:00    2022 Hurric… HURRIC…
 7 DR-4629-CT     4629 CT    DR      2021-10-30 00:00:00    2022 Hurric… REMNAN…
 8 DR-4629-CT     4629 CT    DR      2021-10-30 00:00:00    2022 Hurric… REMNAN…
 9 DR-4626-MS     4626 MS    DR      2021-10-22 00:00:00    2022 Hurric… HURRIC…
10 DR-4626-MS     4626 MS    DR      2021-10-22 00:00:00    2022 Hurric… HURRIC…
# … with 12,479 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Extra operators

Filter multiple values

What if you want to filter multiple items? Well, you'd have to use Boolean logic operators such as:

- `&` means AND, in Boolean logic
- `|` means OR, in Boolean logic
- `!` means NOT, in Boolean logic



```
filter(df, type == “otter” & site == “bay”)
```

type	food	site
otter	urchin	bay
shark	seal	channel
otter	abalone	bay
otter	crab	wharf

```r
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
df %>%
  filter(incidentType=="Hurricane" |
           incidentType == "Fire")
```

```
# A tibble: 16,085 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 16,075 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

%in%

```r
disaster_list <- c("Flood", "Hail", "Typhoon")

df %>%
  filter(incidentType %in% disaster_list)
```

```
# A tibble: 10,678 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 DR-4659-MN     4659 MN    DR      2022-07-13 00:00:00    2022 Flood   SEVERE…
 2 DR-4659-MN     4659 MN    DR      2022-07-13 00:00:00    2022 Flood   SEVERE…
 3 DR-4659-MN     4659 MN    DR      2022-07-13 00:00:00    2022 Flood   SEVERE…
 4 DR-4659-MN     4659 MN    DR      2022-07-13 00:00:00    2022 Flood   SEVERE…
 5 DR-4655-MT     4655 MT    DR      2022-06-16 00:00:00    2022 Flood   SEVERE…
 6 DR-4659-MN     4659 MN    DR      2022-07-13 00:00:00    2022 Flood   SEVERE…
 7 DR-4650-WA     4650 WA    DR      2022-03-29 00:00:00    2022 Flood   SEVERE…
 8 DR-4650-WA     4650 WA    DR      2022-03-29 00:00:00    2022 Flood   SEVERE…
 9 DR-4650-WA     4650 WA    DR      2022-03-29 00:00:00    2022 Flood   SEVERE…
10 DR-4659-MN     4659 MN    DR      2022-07-13 00:00:00    2022 Flood   SEVERE…
# … with 10,668 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
```

select()

```r
colnames(df)
```

```
 [1] "femaDeclarationString"    "disasterNumber"          
 [3] "state"                    "declarationType"         
 [5] "declarationDate"          "fyDeclared"              
 [7] "incidentType"             "declarationTitle"        
 [9] "ihProgramDeclared"        "iaProgramDeclared"       
[11] "paProgramDeclared"        "hmProgramDeclared"       
[13] "incidentBeginDate"        "incidentEndDate"         
[15] "disasterCloseoutDate"     "fipsStateCode"           
[17] "fipsCountyCode"           "placeCode"               
[19] "designatedArea"           "declarationRequestNumber"
[21] "lastIAFilingDate"         "hash"                    
[23] "id"                       "lastRefresh"             
```

```
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType)
```

```
# A tibble: 63,167 × 4
   femaDeclarationString state declarationDate     incidentType
   <chr>                 <chr> <dttm>              <chr>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

slice()

```r
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla
   <chr>          <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX      5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALI
 2 FM-5436-NE      5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 
 3 FM-5444-TX      5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALI
 4 FM-5436-NE      5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 
 5 FM-5436-NE      5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 
 6 FM-5435-AZ      5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOI
 7 FM-5434-AZ      5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNI
 8 FM-5433-NM      5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAI
 9 FM-5432-NM      5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRI
10 FM-5431-NM      5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMI
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variabl
```

```r
df %>%
  arrange(desc(declarationDate))
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 2 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 3 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 4 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 5 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 6 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 7 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 8 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 9 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
10 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variabl
```

```r
df %>%
  arrange(desc(declarationDate)) %>%
  filter(incidentType=="Flood")
```

```
# A tibble: 10,548 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 2 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 3 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 4 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 5 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 6 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 7 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 8 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
 9 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
10 DR-4663-KY     4663 KY    DR      2022-07-29 00:00:00    2022 Flood   SEVER
# … with 10,538 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variabl
```

```r
df %>%
  arrange(desc(declarationDate)) %>%
  filter(incidentType=="Flood") %>%
  select(state, declarationDate, designatedArea)
```

```
# A tibble: 10,548 × 3
   state declarationDate     designatedArea
   <chr> <dttm>              <chr>
 1 KY    2022-07-29 00:00:00 Breathitt (County)
 2 KY    2022-07-29 00:00:00 Clay (County)
 3 KY    2022-07-29 00:00:00 Floyd (County)
 4 KY    2022-07-29 00:00:00 Johnson (County)
 5 KY    2022-07-29 00:00:00 Knott (County)
 6 KY    2022-07-29 00:00:00 Leslie (County)
 7 KY    2022-07-29 00:00:00 Letcher (County)
 8 KY    2022-07-29 00:00:00 Magoffin (County)
 9 KY    2022-07-29 00:00:00 Martin (County)
10 KY    2022-07-29 00:00:00 Owsley (County)
# … with 10,538 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  arrange(desc(declarationDate)) %>%
  filter(incidentType=="Flood") %>%
  select(state, declarationDate, designatedArea) %>%
  slice(1)
```

```
# A tibble: 1 × 3
  state declarationDate     designatedArea
  <chr> <dttm>              <chr>
1 KY    2022-07-29 00:00:00 Breathitt (County)
```

```
df %>%
  arrange(desc(declarationDate)) %>%
  filter(incidentType=="Flood") %>%
  select(state, declarationDate, designatedArea) %>%
  slice(2)
```

```
# A tibble: 1 × 3
  state declarationDate     designatedArea
  <chr> <dttm>              <chr>
1 KY    2022-07-29 00:00:00 Clay (County)
```

```
df %>%
  arrange(desc(declarationDate)) %>%
  filter(incidentType=="Flood") %>%
  select(state, declarationDate, designatedArea) %>%
  slice(1:5)
```

```
# A tibble: 5 × 3
  state declarationDate     designatedArea
  <chr> <dttm>              <chr>
1 KY    2022-07-29 00:00:00 Breathitt (County)
2 KY    2022-07-29 00:00:00 Clay (County)
3 KY    2022-07-29 00:00:00 Floyd (County)
4 KY    2022-07-29 00:00:00 Johnson (County)
5 KY    2022-07-29 00:00:00 Knott (County)
```

dplyr verbs/functions for wrangling data:

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

Advanced Data Journalism: Doing More with R

Class 1: Mutate and Summarize

Andrew Ba Tran

dplyr verbs/functions for wrangling data:

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

Importing data

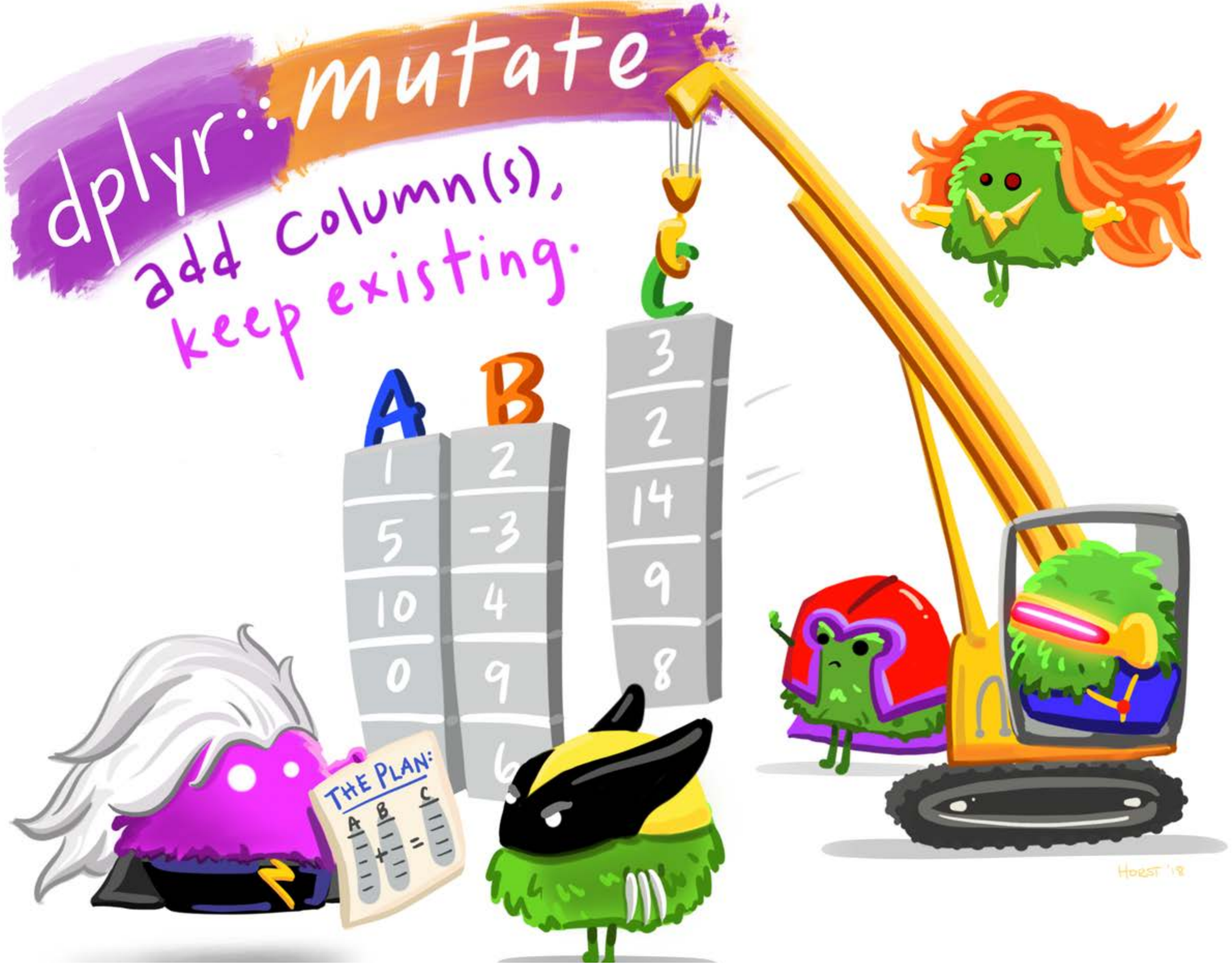
```r
df <- read_csv("https://www.fema.gov/api/open/v2/DisasterDeclarationsSummaries.csv")

df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
glimpse(df)
```

```
Rows: 63,167
Columns: 24
$ femaDeclarationString    <chr> "FM-5444-TX", "FM-5436-NE", "FM-5444-TX", "FM…
$ disasterNumber           <dbl> 5444, 5436, 5444, 5436, 5436, 5435, 5434, 543…
$ state                    <chr> "TX", "NE", "TX", "NE", "NE", "AZ", "AZ", "NM…
$ declarationType          <chr> "FM", "FM", "FM", "FM", "FM", "FM", "FM", "FM…
$ declarationDate          <dttm> 2022-07-19, 2022-04-23, 2022-07-19, 2022-04-…
$ fyDeclared               <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 202…
$ incidentType             <chr> "Fire", "Fire", "Fire", "Fire", "Fire", "Fire…
$ declarationTitle         <chr> "CHALK MOUNTAIN FIRE", "ROAD 702 FIRE", "CHAL…
$ ihProgramDeclared        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ iaProgramDeclared        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ paProgramDeclared        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ hmProgramDeclared        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ incidentBeginDate        <dttm> 2022-07-18, 2022-04-22, 2022-07-18, 2022-04-…
$ incidentEndDate          <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ disasterCloseoutDate     <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ fipsStateCode            <chr> "48", "31", "48", "31", "31", "04", "04", "35…
$ fipsCountyCode           <chr> "221", "063", "425", "065", "145", "025", "00…
$ placeCode                <dbl> 99221, 99063, 99425, 99065, 99145, 99025, 990…
$ designatedArea           <chr> "Hood (County)", "Frontier (County)", "Somerv…
$ declarationRequestNumber <dbl> 22060, 22034, 22060, 22034, 22034, 22032, 220…
$ lastIAFilingDate         <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ hash                     <chr> "373c5ec27998afc08a53302dae796f476b1a6546", "…
$ id                       <chr> "867be42a-71d5-4f13-aa21-d91e0a6fd577", "e671…
$ lastRefresh              <dttm> 2022-07-20 21:21:23, 2022-07-20 21:21:23, 20…
```

mutate()



dplyr::mutate
add column(s),
keep existing.
A
1
5
10
0
B
2
-3
4
9
6
C
3
2
14
9
8
THE PLAN:
A
B
+
=
C
Horst '18

```r
library(lubridate)
```

- extract year! month! day!
- also convert strings into date format recognized by R

```r
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>  
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType)
```

```
# A tibble: 63,167 × 4
   femaDeclarationString state declarationDate     incidentType
   <chr>                 <chr> <dttm>              <chr>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate))
```

```
# A tibble: 63,167 × 5
   femaDeclarationString state declarationDate     incidentType  year
   <chr>                 <chr> <dttm>              <chr>        <dbl>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire          2022
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire          2022
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire          2022
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire          2022
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire          2022
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

summarize()

```r
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
df %>%
  summarize(disasters=n())
```

```
# A tibble: 1 × 1
  disasters
      <int>
1     63167
```

group_by()

```r
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType)
```

```
# A tibble: 63,167 × 4
   femaDeclarationString state declarationDate     incidentType
   <chr>                 <chr> <dttm>              <chr>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate))
```

```
# A tibble: 63,167 × 5
   femaDeclarationString state declarationDate     incidentType  year
   <chr>                 <chr> <dttm>              <chr>        <dbl>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire          2022
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire          2022
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire          2022
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire          2022
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire          2022
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  group_by(state)
```

```
# A tibble: 63,167 × 5
# Groups:   state [59]
   femaDeclarationString state declarationDate     incidentType  year
   <chr>                 <chr> <dttm>              <chr>        <dbl>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire          2022
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire          2022
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire          2022
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire          2022
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire          2022
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  group_by(state) %>%
  summarize(disasters=n())
```

```
# A tibble: 59 × 2
   state disasters
   <chr>     <int>
 1 AK          310
 2 AL         1652
 3 AR         1593
 4 AS           75
 5 AZ          333
 6 CA         1466
 7 CO          646
 8 CT          255
 9 DC           23
10 DE           53
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  group_by(state) %>%
  summarize(disasters=n())
```

```
# A tibble: 59 × 2
   state disasters
   <chr>     <int>
 1 AK          310
 2 AL         1652
 3 AR         1593
 4 AS           75
 5 AZ          333
 6 CA         1466
 7 CO          646
 8 CT          255
 9 DC           23
10 DE           53
# … with 49 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  group_by(incidentType) %>%
  summarize(disasters=n())
```

```
# A tibble: 23 × 2
   incidentType    disasters
   <chr>               <int>
 1 Biological           7857
 2 Chemical                9
 3 Coastal Storm         637
 4 Dam/Levee Break        13
 5 Drought              1292
 6 Earthquake            227
 7 Fire                 3596
 8 Fishing Losses         42
 9 Flood               10548
10 Freezing              301
# … with 13 more rows
# ℹ Use `print(n = ...)` to see more rows
```

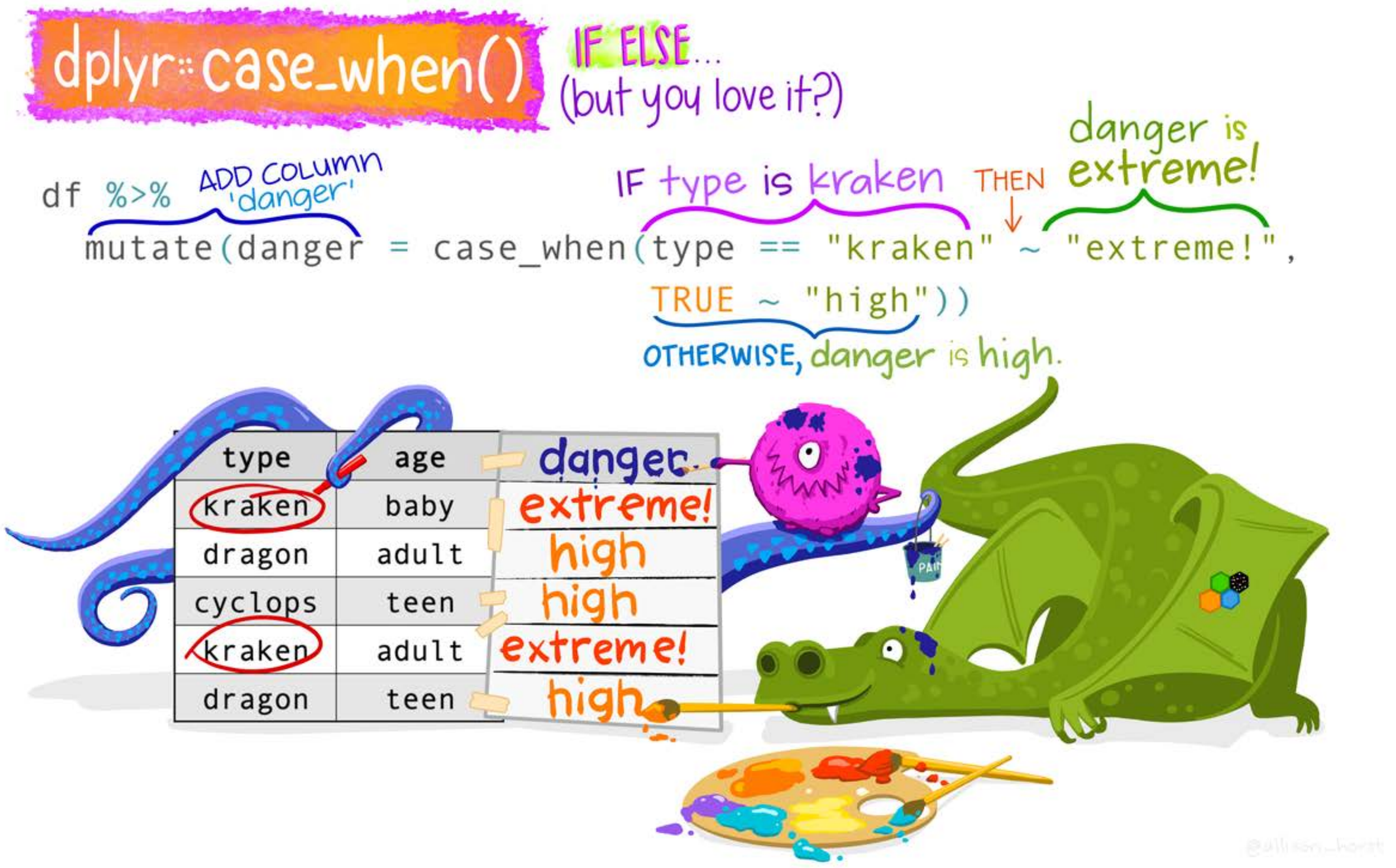
```
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  group_by(state, incidentType) %>%
  summarize(disasters=n())
```

```
# A tibble: 490 × 3
# Groups:   state [59]
   state incidentType    disasters
   <chr> <chr>               <int>
 1 AK    Biological            121
 2 AK    Coastal Storm           2
 3 AK    Earthquake             13
 4 AK    Fire                   30
 5 AK    Flood                  47
 6 AK    Freezing               14
 7 AK    Mud/Landslide           6
 8 AK    Other                   4
 9 AK    Severe Storm(s)        69
10 AK    Snow                    4
# … with 480 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  group_by(state, incidentType, year) %>%
  summarize(disasters=n())
```

```
# A tibble: 2,785 × 4
# Groups:   state, incidentType [490]
   state incidentType   year disasters
   <chr> <chr>         <dbl>     <int>
 1 AK    Biological     2020       121
 2 AK    Coastal Storm  2015         1
 3 AK    Coastal Storm  2018         1
 4 AK    Earthquake     1964         1
 5 AK    Earthquake     2002         6
 6 AK    Earthquake     2018         3
 7 AK    Earthquake     2019         3
 8 AK    Fire           1970         1
 9 AK    Fire           1971         2
10 AK    Fire           1973         1
# … with 2,775 more rows
# ℹ Use `print(n = ...)` to see more rows
```

case_when()



```
df %>%
  mutate(danger = case_when(type == "kraken" ~ "extreme!",
                            TRUE ~ "high"))
```

type	age	danger
kraken	baby	extreme!
dragon	adult	high
cyclops	teen	high
kraken	adult	extreme!
dragon	teen	high

```r
df
```

```
# A tibble: 63,167 × 24
   femaDecla…¹ disas…² state decla…³ declarationDate     fyDec…⁴ incid…⁵ decla…⁶
   <chr>         <dbl> <chr> <chr>   <dttm>                <dbl> <chr>   <chr>
 1 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 2 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 3 FM-5444-TX     5444 TX    FM      2022-07-19 00:00:00    2022 Fire    CHALK …
 4 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 5 FM-5436-NE     5436 NE    FM      2022-04-23 00:00:00    2022 Fire    ROAD 7…
 6 FM-5435-AZ     5435 AZ    FM      2022-04-19 00:00:00    2022 Fire    CROOKS…
 7 FM-5434-AZ     5434 AZ    FM      2022-04-19 00:00:00    2022 Fire    TUNNEL…
 8 FM-5433-NM     5433 NM    FM      2022-04-12 00:00:00    2022 Fire    NOGAL …
 9 FM-5432-NM     5432 NM    FM      2022-04-12 00:00:00    2022 Fire    MCBRID…
10 FM-5431-NM     5431 NM    FM      2022-04-12 00:00:00    2022 Fire    HERMIT…
# … with 63,157 more rows, 16 more variables: ihProgramDeclared <dbl>,
#   iaProgramDeclared <dbl>, paProgramDeclared <dbl>, hmProgramDeclared <dbl>,
#   incidentBeginDate <dttm>, incidentEndDate <dttm>,
#   disasterCloseoutDate <dttm>, fipsStateCode <chr>, fipsCountyCode <chr>,
#   placeCode <dbl>, designatedArea <chr>, declarationRequestNumber <dbl>,
#   lastIAFilingDate <dttm>, hash <chr>, id <chr>, lastRefresh <dttm>, and
#   abbreviated variable names ¹femaDeclarationString, ²disasterNumber, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType)
```

```
# A tibble: 63,167 × 4
   femaDeclarationString state declarationDate     incidentType
   <chr>                 <chr> <dttm>              <chr>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate))
```

```
# A tibble: 63,167 × 5
   femaDeclarationString state declarationDate     incidentType  year
   <chr>                 <chr> <dttm>              <chr>        <dbl>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire          2022
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire          2022
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire          2022
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire          2022
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire          2022
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  mutate(year_group=case_when(
    year < 1960 ~ "1950 - 1959",
    year >=1960 & year < 1969 ~ "1960-1969",
    year >=1970 & year < 1979 ~ "1970-1979",
    year >=1980 & year < 1989 ~ "1980-1989",
    year >=1990 & year < 1999 ~ "1990-1999",
    year >=2000 & year < 2009 ~ "2000-2009",
    year >=2010 & year < 2019 ~ "2010-2019",
    TRUE ~ "2020+"
  ))
```

```
# A tibble: 63,167 × 6
   femaDeclarationString state declarationDate     incidentType  year year_group
   <chr>                 <chr> <dttm>              <chr>        <dbl> <chr>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022 2020+
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022 2020+
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022 2020+
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022 2020+
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022 2020+
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire          2022 2020+
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire          2022 2020+
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire          2022 2020+
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire          2022 2020+
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire          2022 2020+
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  mutate(year_group=case_when(
    year < 1960 ~ "1950 - 1959",
    year >=1960 & year < 1969 ~ "1960-1969",
    year >=1970 & year < 1979 ~ "1970-1979",
    year >=1980 & year < 1989 ~ "1980-1989",
    year >=1990 & year < 1999 ~ "1990-1999",
    year >=2000 & year < 2009 ~ "2000-2009",
    year >=2010 & year < 2019 ~ "2010-2019",
    TRUE ~ "2020+"
  )) %>%
  group_by(year_group)
```

```
# A tibble: 63,167 × 6
# Groups:   year_group [8]
   femaDeclarationString state declarationDate     incidentType  year year_group
   <chr>                 <chr> <dttm>              <chr>        <dbl> <chr>
 1 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022 2020+
 2 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022 2020+
 3 FM-5444-TX            TX    2022-07-19 00:00:00 Fire          2022 2020+
 4 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022 2020+
 5 FM-5436-NE            NE    2022-04-23 00:00:00 Fire          2022 2020+
 6 FM-5435-AZ            AZ    2022-04-19 00:00:00 Fire          2022 2020+
 7 FM-5434-AZ            AZ    2022-04-19 00:00:00 Fire          2022 2020+
 8 FM-5433-NM            NM    2022-04-12 00:00:00 Fire          2022 2020+
 9 FM-5432-NM            NM    2022-04-12 00:00:00 Fire          2022 2020+
10 FM-5431-NM            NM    2022-04-12 00:00:00 Fire          2022 2020+
# … with 63,157 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
df %>%
  select(femaDeclarationString, state,
         declarationDate, incidentType) %>%
  mutate(year=year(declarationDate)) %>%
  mutate(year_group=case_when(
    year < 1960 ~ "1950 - 1959",
    year >=1960 & year < 1969 ~ "1960-1969",
    year >=1970 & year < 1979 ~ "1970-1979",
    year >=1980 & year < 1989 ~ "1980-1989",
    year >=1990 & year < 1999 ~ "1990-1999",
    year >=2000 & year < 2009 ~ "2000-2009",
    year >=2010 & year < 2019 ~ "2010-2019",
    TRUE ~ "2020+"
  )) %>%
  group_by(year_group) %>%
  summarize(disasters=n())
```

```
# A tibble: 8 × 2
  year_group  disasters
  <chr>           <int>
1 1950 - 1959        94
2 1960-1969        1108
3 1970-1979        5075
4 1980-1989        1735
5 1990-1999        8806
6 2000-2009       16348
7 2010-2019       12087
8 2020+           17914
```

dplyr verbs/functions for wrangling data:

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