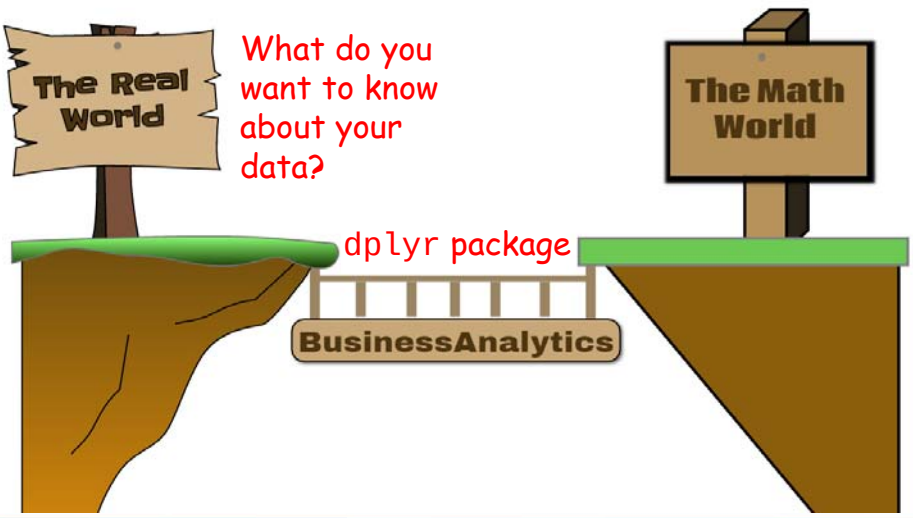
Data Manipulation Part 1

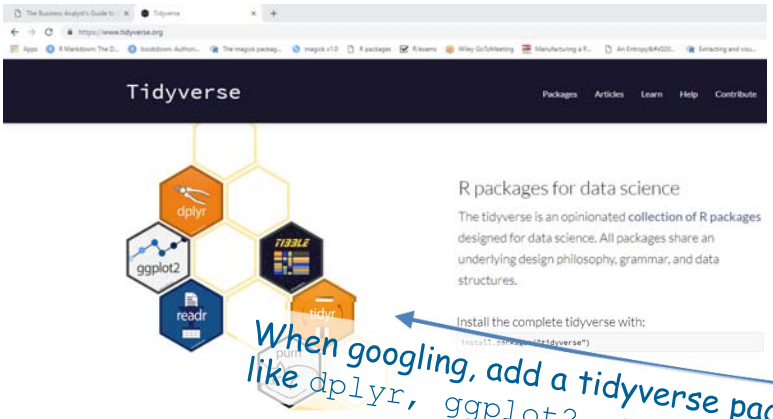
Six Magic Words + Indicator Functions

UNIVERSITY OF DELAWARE

dplyr: The Shortest Data Manipulation Bridge



As a rule of thumb, tidyverse packages lead to the shortest bridge crossings



When googling, add a tidyverse package name like `dplyr`, `ggplot2`, `tidyr`, `stringr`, `readr`, etc…

`dplyr` for data manipulation

What do you want to know about your data?

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

dp

Materials Adapted From: Hadley Wickham, "dplyr tutorial", June 2014

dplyr::filter()

Show me the Droids I am looking for.

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

```
filter(df, species == "Droid")
```

name	height	species
C-3PO	167	Droid
R2-D2	96	Droid

dplyr::filter()

Show me the non-humans.

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

```
filter(df, species != "Human")
```

or equivalently

```
filter(df, species %in% c("Droid","Wookiee", "Yoda's species"))
```

name	height	species
C-3PO	167	Droid
R2-D2	96	Droid
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

dplyr::select()

Let's anonymize the data and just look at height and species.

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
		a's species

```
select(df, height, species)
    or
select(df, -name)
```

height	species
172	Human
167	Droid
96	Droid
202	Human
182	Human
228	Wookiee
66	Yoda's species

Materials Adapted From: Hadley Wickham, "dplyr tutorial", June 2014

dplyr::arrange()

Let's arrange everyone by height.

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee

```
arrange(df, height)
```

name	height	species
Yoda	66	Yoda's species
R2-D2	96	Droid
C-3PO	167	Droid
Luke Skywalker	172	Human
Obi-Wan Kenobi	182	Human
Darth Vader	202	Human
Chewbacca	228	Wookiee

Materials Adapted From: Hadley Wickham, "dplyr tutorial", June 2014

dplyr::arrange()

Let's arrange everyone by descending height.

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee

```
arrange(df, desc(height))
```

name	height	species
Chewbacca	228	Wookiee
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Yoda	66	Yoda's species

dplyr::mutate()

Let's get height in inches.

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid

name	height	species	inches
Luke Skywalker	172	Human	67.72
C-3PO	167	Droid	65.75
R2-D2	96	Droid	37.80
Darth Vader	202	Human	79.53
Obi-Wan Kenobi	182	Human	71.65
Chewbacca	228	Wookiee	89.76
Yoda	66	Yoda's species	25.98

```
mutate(df, inches = height / 2.54)
```

dplyr::summarize()

What is the **average** height in the dataset?

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

avgHeight
159

```
summarize(df, avgHeight = mean(height))
```


dplyr::group_by()

What is the **average** height of each species in the dataset?

df =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

newDF =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

Step1: Create newDF which tells the dataframe to create groups

```
newDF = group_by(df, species)
```

The power of `summarize()` combined with: `group_by()`

newDF =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

Step1: Create newDF which tells the dataframe to create groups

```
newDF = group_by(df, species)
```


The power of `summarize()` combined with: `group_by()`

Step1: Create newDF which tells the dataframe to create groups

newDF =

name	height	species
Luke Skywalker	172	Human
C-3PO	167	Droid
R2-D2	96	Droid
Darth Vader	202	Human
Obi-Wan Kenobi	182	Human
Chewbacca	228	Wookiee
Yoda	66	Yoda's species

```
newDF = group_by(df, species)
```

Step2: Now aggregate rows by finding the mean height.

species	height
Droid	132
Human	185
Wookiee	228
Yoda's species	66

```
summarize(newDF, avgHeight = mean(height))
```

Each function has two properties

```
 filter(df, species == "Droid")
 filter(df, species != "Human")
 select(df, height, species)
 select(df, -name)
arrange(df, height)
arrange(df, desc(height))
 mutate(df, inches = height / 2.54)
summarize(df, avgHeight = mean(height))
group_by(df, species)
```


Each function has two properties

```
 filter(df, species == "Droid")
 filter(df, species != "Human")
 select(df, height, species)
 select(df, -name)
arrange(df, height)
arrange(df, desc(height))
 mutate(df, inches = height / 2.54)
summarize(df, avgHeight = mean(height))
group_by(df, species)
```

PROPERTY #1: THE FIRST ARGUMENT IS ALWAYS A DATA FRAME

Each function has two properties

```
filter(df, species == "Droid")
filter(df, species != "Human")
select(df, height, species)
select(df, -name)
arrange(df, height)
arrange(df, desc(height))
mutate(df, inches = height / 2.54)
summarize(df, avgHeight = mean(height))
group_by(df, species)
```

If input and outputs are data frames, then…

code that uses these functions might look like this…

```
library(dplyr)  ### use dplyr functions in this R session
df  = starwars  ### starwars is a built-in data frame

df1 =    select( df, name, height, species)
df2 =  group_by(df1,  species)
df3 = summarize(df2, avgHeight = mean(height))
df3
```

What if I don't need all those intermediate data frames (e.g. `df1`, `df2`, etc.) to be clogging up my environment?

If input and outputs are data frames, then…

code that uses these functions might look like this…

```
summarize(group_by(select(starwars,
                          name, height, species),
                   species),
          avgHeight = mean(height))
```

read this inside-out

If input and outputs are data frames, then…

code that uses these functions might look like this…

```
summarize(group_by(select(starwars,
                          name, height, species),
                   species),
          avgHeight = mean(height))
```

read this inside-out

Rescued by The Chaining Operator %>%

%>% takes what is to its left and makes it the first argument of the function on its right

Assume: $f(x, y) = x + y^2$

What is $f(3,2)$?

What is 6 %>% $f(2)$?

Rescued by The Chaining Operator %>%

%>% takes what is to its left and makes it the first argument of the function on its right

```
 df = starwars
df1 =    select( df, name, height, species)
df2 =  group_by(df1, species)
df3 = summarize(df2, avgHeight = mean(height))
df3
```

```
starwars %>%
  select(name, height, species) %>%
  group_by(species) %>%
  summarize(avgHeight = mean(height))
```

Indicator Functions in dplyr


```
## what percentage of characters are droids

starwars[1:10,] %>%
   select(name, species)
```

```
➢ starwars[1:10,] %>%
    select(name, species)
 # A tibble: 10 x 2
 name                 species
 <chr>                <chr>
 1 Luke Skywalker     Human
 2 C-3PO              Droid
 3 R2-D2              Droid
 4 Darth Vader        Human
 5 Leia Organa        Human
 6 Owen Lars          Human
 7 Beru Whitesun lars Human
 8 R5-D4              Droid
 9 Biggs Darklighter  Human
10 Obi-Wan Kenobi     Human
```

```
starwars[1:10,] %>%
  select(name, species) %>%
  mutate(isDroidFlag =
           ifelse(species == “Droid”, TRUE,
FALSE))
```

```
> starwars[1:10,] %>%
    select(name, species)
# A tibble: 10 x 2
   name                species
   <chr>               <chr>
 1 Luke Skywalker      Human
 2 C-3PO               Droid
 3 R2-D2               Droid
 4 Darth Vader         Human
 5 Leia Organa         Human
 6 Owen Lars           Human
 7 Beru Whitesun lars  Human
 8 R5-D4               Droid
 9 Biggs Darklighter   Human
10 Obi-Wan Kenobi      Human
```

```
> starwars[1:10,] %>%
    select(name, species) %>%
    mutate(isDroidFlag =
             ifelse(species == "Droid", TRUE, FALSE))
# A tibble: 10 x 3
   name                species    isDroidFlag
   <chr>               <chr>      <lgl>
 1 Luke Skywalker      Human      FALSE
 2 C-3PO               Droid      TRUE
 3 R2-D2               Droid      TRUE
 4 Darth Vader         Human      FALSE
 5 Leia Organa         Human      FALSE
 6 Owen Lars           Human      FALSE
 7 Beru Whitesun lars  Human      FALSE
 8 R5-D4               Droid      TRUE
 9 Biggs Darklighter   Human      FALSE
10 Obi-Wan Kenobi      Human      FALSE
```



```
starwars[1:10,] %>%
  select(name, species) %>%
  mutate(isDroidFlag =
           ifelse(species == "Droid", TRUE, FALSE)) %>%
  summarize(countDroids = sum(isDroidFlag, na.rm = TRUE))
```

```
> starwars[1:10,] %>%
    select(name, species) %>%
    mutate(isDroidFlag =
             ifelse(species == "Droid", TRUE,
FALSE))
# A tibble: 10 x 3
   name                species    isDroidFlag
   <chr>               <chr>      <lgl>
 1 Luke Skywalker      Human      FALSE
 2 C-3PO               Droid      TRUE
 3 R2-D2               Droid      TRUE
 4 Darth Vader         Human      FALSE
 5 Leia Organa         Human      FALSE
 6 Owen Lars           Human      FALSE
 7 Beru Whitesun lars  Human      FALSE
 8 R5-D4               Droid      TRUE
 9 Biggs Darklighter   Human      FALSE
10 Obi-Wan Kenobi      Human      FALSE
```

```
> starwars[1:10,] %>%
    select(name, species) %>%
    mutate(isDroidFlag =
             ifelse(species == "Droid", TRUE, FALSE)) %>%
    summarize(countDroids =
                sum(isDroidFlag, na.rm = TRUE))
# A tibble: 1 x 1
  countDroids
        <int>
1           3
```

```
starwars[1:10,] %>%
  select(name, species) %>%
  mutate(isDroidFlag =
           ifelse(species == "Droid", TRUE, FALSE)) %>%
  summarize(countDroids = sum(isDroidFlag, na.rm = TRUE),
            pctDroids = mean(isDroidFlag, na.rm = TRUE)
```

```
> starwars[1:10,] %>%
    select(name, species) %>%
    mutate(isDroidFlag =
             ifelse(species == "Droid", TRUE,
FALSE))
# A tibble: 10 x 3
   name               species    isDroidFlag
   <chr>              <chr>      <lgl>
 1 Luke Skywalker     Human      FALSE
 2 C-3PO              Droid      TRUE
 3 R2-D2              Droid      TRUE
 4 Darth Vader        Human      FALSE
 5 Leia Organa        Human      FALSE
 6 Owen Lars          Human      FALSE
 7 Beru Whitesun lars Human      FALSE
 8 R5-D4              Droid      TRUE
 9 Biggs Darklighter  Human      FALSE
10 Obi-Wan Kenobi     Human      FALSE
```

```
> starwars[1:10,] %>%
    select(name, species) %>%
    mutate(isDroidFlag =
             ifelse(species == "Droid", TRUE, FALSE)) %>%
    summarize(countDroids = sum(isDroidFlag, na.rm = TRUE),
              pctDroids = mean(isDroidFlag, na.rm = TRUE))
# A tibble: 1 x 2
  countDroids pctDroids
        <int>     <dbl>
1           5       0.3
```


Switch to R-Script