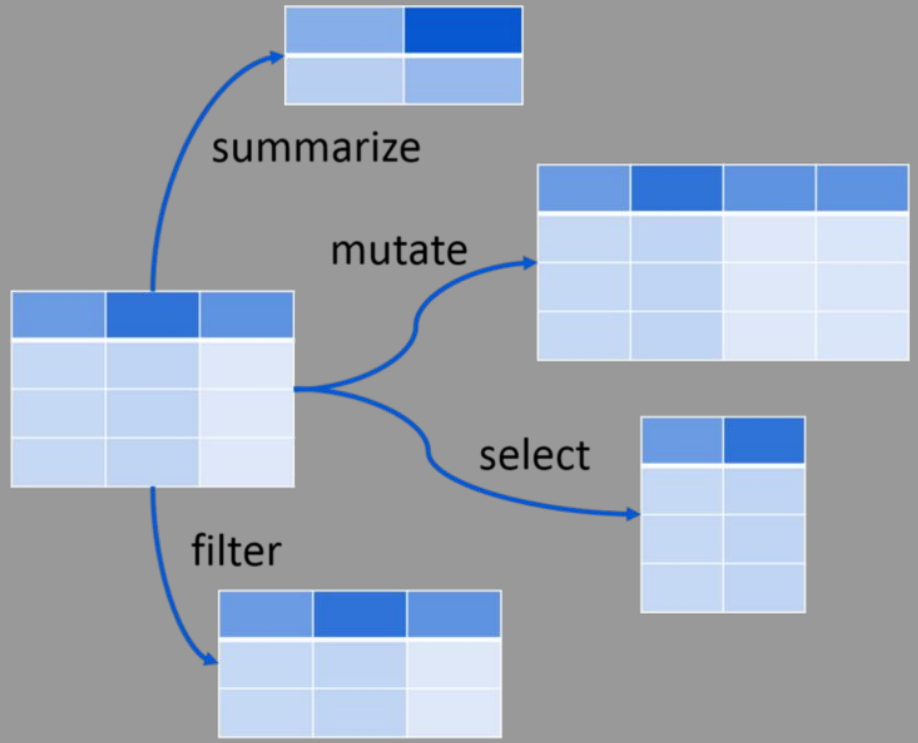
Data Transformation Using dplyr in R

Vaishnavi Kashyap, MSDS 610

Clean Data Is Rare!!!

“A grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges”



https://dplyr.tidyverse.org/

dplyr Basics

All 5 verbs work in a similar manner:

1. The first argument is a data frame.
2. The subsequent arguments describe what to do with the data frame, using the variable names (without quotes).
3. The output is a new dataframe, printed on its own or assigned to a variable

They can be used with the group_by() function as well in order to further narrow down the exact data you want.

filter() - Pick Observations by Their Values

```
> filter(flights, month==8, day==30)
# A tibble: 965 x 19
```

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier
	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<int>	<dbl>	<chr>
1	2013	8	30	450	500	-10	621	642	-21	US
2	2013	8	30	521	529	-8	729	809	-40	UA
3	2013	8	30	539	545	-6	932	921	11	B6
4	2013	8	30	539	545	-6	801	830	-29	UA
5	2013	8	30	541	545	-4	827	855	-28	AA
6	2013	8	30	550	600	-10	815	901	-46	UA
7	2013	8	30	551	608	-17	656	719	-23	B6
8	2013	8	30	551	600	-9	650	716	-26	EV
9	2013	8	30	553	600	-7	808	826	-18	DL
10	2013	8	30	553	600	-7	706	722	-16	UA

```
# … with 955 more rows, and 9 more variables: flight <int>, tailnum <chr>, origin <chr>,
#   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

arrange() - Reorder Rows

```
> arrange(aug30, desc(dep_time))
# A tibble: 965 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>
 1  2013     8    30     2359           2359         0      340            340         0 B6
 2  2013     8    30     2358           2359        -1      334            344       -10 B6
 3  2013     8    30     2355           2359        -4      339            350       -11 B6
 4  2013     8    30     2308           2155        73      157             43        74 B6
 5  2013     8    30     2303           2305        -2        7             13        -6 B6
 6  2013     8    30     2253           2255        -2       20             19         1 B6
 7  2013     8    30     2251           2059       112     2348           2211        97 UA
 8  2013     8    30     2249           2255        -6        7             14        -7 B6
 9  2013     8    30     2241           2245        -4     2357              1        -4 B6
10  2013     8    30     2240           2245        -5     2356           2359        -3 B6
# … with 955 more rows, and 9 more variables: flight <int>, tailnum <chr>, origin <chr>,
#   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

select() - Narrow Down the Dataset

```
> arrange(select(aug30, dep_time, sched_dep_time), desc(dep_time))
# A tibble: 965 x 2
   dep_time sched_dep_time
      <int>          <int>
 1     2359           2359
 2     2358           2359
 3     2355           2359
 4     2308           2155
 5     2303           2305
 6     2253           2255
 7     2251           2059
 8     2249           2255
 9     2241           2245
10     2240           2245
# … with 955 more rows
```

mutate() - Add New Variables

```
> arrange(mutate(diff, time_diff=dep_time-sched_dep_time), time_diff)
# A tibble: 965 x 3
   dep_time sched_dep_time time_diff
      <int>          <int>     <int>
 1      551            608       -57
 2     1656           1709       -53
 3     1847           1900       -53
 4     2047           2100       -53
 5     1553           1605       -52
 6     1859           1910       -51
 7     1949           2000       -51
 8     1959           2010       -51
 9      450            500       -50
10      550            600       -50
```

summarize() - Grouped Summaries

```
> summarize(daily, delay = mean(arr_delay, na.rm = TRUE))
`summarise()` regrouping output by 'year', 'month' (override with `.groups` argument)
# A tibble: 365 x 4
# Groups:   year, month [12]
    year month   day  delay
   <int> <int> <int>  <dbl>
 1  2013     1     1 12.7
 2  2013     1     2 12.7
 3  2013     1     3  5.73
 4  2013     1     4 -1.93
 5  2013     1     5 -1.53
 6  2013     1     6  4.24
 7  2013     1     7 -4.95
 8  2013     1     8 -3.23
 9  2013     1     9 -0.264
10  2013     1    10 -5.90
# … with 355 more rows
```


NOT COOL BRO
CLEAN YOUR DATA

References

https://r4ds.had.co.nz/transform.html

https://rdrr.io/cran/nycflights13/man/flights.html

https://dplyr.tidyverse.org/

http://statseducation.com/Introduction-to-R/modules/getting%20data/tibbles/