dplyr

The Five Verbs of dplyr

The dplyr package contains five key data manipulation functions, also called verbs:

- select(), which returns a subset of the columns,
- filter(), that is able to return a subset of the rows,
- arrange(), that reorders the rows according to single or multiple variables,
- mutate(), used to add columns from existing data and
- summarise(), which reduces each group to a single row by calculating aggregate measures.

What order of operations should we use to find the average value of the ArrDelay (arrival delay) variable for all American Airline flights in the hflights tbl?

Questions:

Manipulating Variables (Select and Mutate)

2. Return a copy of hflights that contains the four columns related tod elay (ActualElapsedTime, AirTime, ArrDelay, DepDelay).

3. Return a copy of hflights containing the columns Origin up to Cancelled

4. Find the most concise way to select: columns Year up to and including DayOfWeek, columns ArrDelay up to and including Diverted.

5. dplyr comes with a set of helper functions that can help you select variables. These functions find groups of variables to select, based on their names. dplyr provides 6 helper functions, each of which only works when used inside select().

- starts_with("X"): every name that starts with "X",
- ends_with("X"): every name that ends with "X",
- contains("X"): every name that contains "X",
- matches("X"): every name that matches "X", which can be a regular expression,
- num_range("x", 1:5): the variables named x01, x02, x03, x04 and x05,
- one_of(x): every name that appears in x, which should be a character vector.

Use select and a helper function to return a tbl copy of hflights that contains just ArrDelay and DepDelay.

6. Use a combination of helper functions and variable names to return the UniqueCarrier, FlightNum, TailNum, Cancelled, and CancellationCode columns of hflights.

7. Which variables in hflight do you think count as a plane's "ground time"? Use mutate() to add these variables together and save them as GroundTime. Save your results as g.

Manipulating Observations (Filter and Arrange)

When manipulating observations, we should know the following operations: (Revision)

- x < y, TRUE if x is less than y
- x <= y, TRUE if x is less than or equal to y
- x == y, TRUE if x equals y
- x != y, TRUE if x does not equal y
- x >= y, TRUE if x is greater than or equal to y
- x > y, TRUE if x is greater than y
- x %in% c(a, b, c), TRUE if x is in the vector c(a, b, c)

8. Return a copy of all flights that traveled 3000 miles or more. Save it in f1.

9. Return a copy of all flights flights where taxiing took longer than flying. Save it in f3.

10. Return a copy of all cancelled weekend flights

11. Arrange according to carrier and decreasing departure delays

12. Arrange flights by total delay (normal order).

13. Filter out flights leaving to DFW before 8am and arrange according to decreasing AirTime

Manipulating Groups of Observation (summarize and group_by)

14. Determine the shortest and longest distance flown and save statistics to min_dist and max_dist resp.

15. Determine the longest distance for diverted flights, save statistic to max_div. Use a one-liner!

16. dplyr provides several helpful aggregate functions of its own, in addition to the ones that are already defined in R. These include:

- first(x) - The first element of vector x.
- last(x) - The last element of vector x.
- nth(x, n) - The nth element of vector x.

- n() - The number of rows in the data.frame or group of observations that summarise() describes.
- n_distinct(x) - The number of unique values in vector x.

Create a table with the following variables (and variable names): the total number of observations in hflights (n_obs), the total number of carriers that appear in hflights (n_carrier), the total number of destinations that appear in hflights (n_dest), and the destination of the flight that appears in the 100th row of hflights (dest100).

17. Use Piping:

    (1) Take the hflights data set and then,

    (2) Add a variable named diff that is the result of subtracting TaxiIn from TaxiOut, and then

    (3) pick all the rows whose diff value does not equal NA, and then

    (4) summarise the data set with a value named avg that is the mean diff value.

Store the result in the variable p.

18. Use Piping:

    Define a data set named d that contains just the Dest, UniqueCarrier, Distance, and ActualElapsedTime columns of hflights as well an additional variable: RealTime which is equal the actual elapsed time plus 100 minute.

19. Use Piping to compare the individual carriers.

    For each carrier, count the total number of flights flown by the carrier (n_flights), the total number of cancelled flights (n_canc), and the average arrival delay of the flights whose delay does not equal NA (avg_delay). Once you've calculated these results, arrange() the carriers from low to high by their average arrival delay. Use number of flights cancelled to break any ties. Which airline scores best based on these statistics?