DSO 545: Statistical Computing and Data Visualization

Data Manipulation with dplyr : Analyzing Flights Data (hflights)

Fall 2018

LAB 03

The Five Verbs of `dplyr`

1. 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,
- `summarise()`, which reduces each group to a single row by calculating aggregate measures.

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

```
# The order is: filter(), then summarise()
```

Manipulating Variables (Select and Mutate)

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

```
library(hflights)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df2 = select(hflights, ActualElapsedTime, AirTime, ArrDelay, DepDelay)
head(df2)
```

```
##      ActualElapsedTime AirTime ArrDelay DepDelay
## 5424                60      40      -10        0
## 5425                60      45       -9        1
## 5426                70      48       -8       -8
## 5427                70      39        3        3
## 5428                62      44       -3        5
## 5429                64      45       -7       -1
```

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

```
df3= select(hflights, Origin:Cancelled)
head(df3)
```

```
##      Origin Dest Distance TaxiIn TaxiOut Cancelled
## 5424    IAH  DFW      224      7      13         0
## 5425    IAH  DFW      224      6       9         0
## 5426    IAH  DFW      224      5      17         0
## 5427    IAH  DFW      224      9      22         0
## 5428    IAH  DFW      224      9       9         0
## 5429    IAH  DFW      224      6      13         0
```

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

```
df4 = select(hflights, -(DepTime:AirTime))
head(df4)
```

```
##      Year Month DayofMonth DayOfWeek ArrDelay DepDelay Origin Dest
## 5424 2011     1          1         6      -10        0    IAH  DFW
## 5425 2011     1          2         7       -9        1    IAH  DFW
## 5426 2011     1          3         1       -8       -8    IAH  DFW
## 5427 2011     1          4         2        3        3    IAH  DFW
## 5428 2011     1          5         3       -3        5    IAH  DFW
## 5429 2011     1          6         4       -7       -1    IAH  DFW
##      Distance TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 5424      224      7      13         0                          0
## 5425      224      6       9         0                          0
## 5426      224      5      17         0                          0
## 5427      224      9      22         0                          0
## 5428      224      9       9         0                          0
## 5429      224      6      13         0                          0
```

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`.

```
df5 = select(hflights, ends_with("Delay"))
head(df5)
```

```
##      ArrDelay DepDelay
## 5424      -10        0
## 5425       -9        1
## 5426       -8       -8
## 5427        3        3
## 5428       -3        5
## 5429       -7       -1
```

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

```
df6 = select(hflights, UniqueCarrier, ends_with("Num") , starts_with("Cancel"))
head(df6)
```

```
##      UniqueCarrier FlightNum TailNum Cancelled CancellationCode
## 5424            AA       428  N576AA         0
## 5425            AA       428  N557AA         0
## 5426            AA       428  N541AA         0
## 5427            AA       428  N403AA         0
## 5428            AA       428  N492AA         0
## 5429            AA       428  N262AA         0
```

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`.

```
df7 = mutate(hflights, GroundTime = TaxiIn + TaxiOut)
head(df7)
```

```
##   Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1 2011     1          1         6    1400    1500            AA       428
## 2 2011     1          2         7    1401    1501            AA       428
## 3 2011     1          3         1    1352    1502            AA       428
## 4 2011     1          4         2    1403    1513            AA       428
## 5 2011     1          5         3    1405    1507            AA       428
## 6 2011     1          6         4    1359    1503            AA       428
##   TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 1  N576AA                60      40      -10        0    IAH  DFW      224
## 2  N557AA                60      45       -9        1    IAH  DFW      224
## 3  N541AA                70      48       -8       -8    IAH  DFW      224
## 4  N403AA                70      39        3        3    IAH  DFW      224
## 5  N492AA                62      44       -3        5    IAH  DFW      224
## 6  N262AA                64      45       -7       -1    IAH  DFW      224
##   TaxiIn TaxiOut Cancelled CancellationCode Diverted GroundTime
## 1      7      13         0                         0         20
## 2      6       9         0                         0         15
## 3      5      17         0                         0         22
## 4      9      22         0                         0         31
## 5      9       9         0                         0         18
## 6      6      13         0                         0         19
```

Manipulating Observations (Filter and Arrange)

When manipulating observations, we should know the following operations:

- `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.

```
df8 = filter(hflights, Distance > 3000)
head(df8)
```

```
##   Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1 2011     1         31         1     924    1413            CO         1
## 2 2011     1         30         7     925    1410            CO         1
## 3 2011     1         29         6    1045    1445            CO         1
## 4 2011     1         28         5    1516    1916            CO         1
## 5 2011     1         27         4     950    1344            CO         1
## 6 2011     1         26         3     944    1350            CO         1
##   TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 1  N69063               529     492       23       -1    IAH  HNL     3904
## 2  N76064               525     493       20        0    IAH  HNL     3904
## 3  N69063               480     459       55       80    IAH  HNL     3904
## 4  N77066               480     463      326      351    IAH  HNL     3904
## 5  N76055               474     455       -6       25    IAH  HNL     3904
## 6  N76065               486     471        0       19    IAH  HNL     3904
##   TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 1      6      31         0                         0
## 2     13      19         0                         0
## 3      4      17         0                         0
## 4      7      10         0                         0
## 5      4      15         0                         0
## 6      5      10         0                         0
```

9. Return a copy of all flights flown by one of American(AA), Alaska (AS), or JetBlue (B6) airlines.

```
df9 = filter(hflights, UniqueCarrier %in% c("AA","AS", "B6"))
## OR
df9 = filter(hflights, UniqueCarrier =="AA" | UniqueCarrier =="AS" | UniqueCarrier =="B6" )

head(df9)
```

```
##   Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1 2011     1          1         6    1400    1500            AA       428
## 2 2011     1          2         7    1401    1501            AA       428
## 3 2011     1          3         1    1352    1502            AA       428
## 4 2011     1          4         2    1403    1513            AA       428
## 5 2011     1          5         3    1405    1507            AA       428
## 6 2011     1          6         4    1359    1503            AA       428
##   TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 1  N576AA               60      40      -10        0    IAH  DFW      224
## 2  N557AA               60      45       -9        1    IAH  DFW      224
## 3  N541AA               70      48       -8       -8    IAH  DFW      224
## 4  N403AA               70      39        3        3    IAH  DFW      224
## 5  N492AA               62      44       -3        5    IAH  DFW      224
## 6  N262AA               64      45       -7       -1    IAH  DFW      224
##   TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 1      7      13         0                         0
## 2      6       9         0                         0
## 3      5      17         0                         0
## 4      9      22         0                         0
## 5      9       9         0                         0
## 6      6      13         0                         0
```

10. Return a copy of all flights where taxi-ing took longer than flying.

```
df10 = filter(hflights, TaxiIn + TaxiOut > AirTime)
head(df10)
```

```
##   Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1 2011     1         24         1     731     904            AA       460
## 2 2011     1         30         7    1959    2132            AA       533
## 3 2011     1         24         1    1621    1749            AA      1121
## 4 2011     1         10         1     941    1113            AA      1436
## 5 2011     1         31         1    1301    1356            CO       241
## 6 2011     1         31         1    2113    2215            CO      1533
##   TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 1  N545AA                93      42       29       11    IAH  DFW      224
## 2  N455AA                93      43       12       -6    IAH  DFW      224
## 3  N484AA                88      43        4       -9    IAH  DFW      224
## 4  N591AA                92      45       48       31    IAH  DFW      224
## 5  N14629                55      27       -2       -4    IAH  AUS      140
## 6  N72405                62      30       20       13    IAH  AUS      140
##   TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 1     14      37         0                         0
## 2     10      40         0                         0
## 3     10      35         0                         0
## 4     27      20         0                         0
## 5      5      23         0                         0
## 6      7      25         0                         0
```

11. Return a copy of all cancelled weekend flights

```
df11= filter(hflights, DayOfWeek %in% c(6,7) & Cancelled == 1)
#OR
df11= filter(hflights, (DayOfWeek == 6 | DayOfWeek == 7) & Cancelled == 1)

head(df11)
```

```
##   Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1 2011     1          9         7      NA      NA            AA      1820
## 2 2011     1         29         6      NA      NA            CO       408
## 3 2011     1          9         7      NA      NA            CO       755
## 4 2011     1          9         7      NA      NA            DL         8
## 5 2011     1          9         7      NA      NA            OO      6726
## 6 2011     1          2         7      NA      NA            WN      1629
##   TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 1  N4XCAA                NA      NA       NA       NA    IAH  DFW      224
## 2                        NA      NA       NA       NA    IAH  EWR     1400
## 3                        NA      NA       NA       NA    IAH  ATL      689
## 4  N933DL                NA      NA       NA       NA    IAH  ATL      689
## 5  N779SK                NA      NA       NA       NA    IAH  ASE      914
## 6  N749SW                NA      NA       NA       NA    HOU  DAL      239
##   TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 1     NA      NA         1                B        0
## 2     NA      NA         1                A        0
## 3     NA      NA         1                B        0
## 4     NA      NA         1                B        0
## 5     NA      NA         1                B        0
## 6     NA      NA         1                A        0
```

12. Arrange according to carrier and decreasing departure delays.

```
df12 = arrange(hflights, UniqueCarrier, desc(DepDelay))
```

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

```
df13 = arrange(hflights, DepDelay + ArrDelay)
head(df13)
```

```
##   Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1 2011     7          3         7    1914    2039            XE      2804
## 2 2011     8         31         3     934    1039            OO      2040
## 3 2011     8         21         7     935    1039            OO      2001
## 4 2011     8         28         7    2059    2206            OO      2003
## 5 2011     8         29         1     935    1041            OO      2040
## 6 2011    12         25         7     741     926            OO      4591
##   TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 1  N12157                85      66      -70       -1    IAH  MEM      468
## 2  N783SK               185     172      -56      -11    IAH  BFL     1428
## 3  N767SK               184     171      -56      -10    IAH  BFL     1428
## 4  N783SK               187     171      -54      -11    IAH  BFL     1428
## 5  N767SK               186     169      -54      -10    IAH  BFL     1428
## 6  N814SK               165     147      -57       -4    IAH  SLC     1195
##   TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 1      4      15         0                         0
## 2      3      10         0                         0
## 3      3      10         0                         0
## 4      5      11         0                         0
## 5      4      13         0                         0
## 6      4      14         0                         0
```

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

```
df14 = arrange(filter(hflights, Dest == "DFW" & DepTime < 800), desc(AirTime))

## OR

df14 = hflights %>%
  filter(Dest == "DFW" & DepTime < 800) %>%
  arrange(desc(AirTime))

head(df14)
```

```
##   Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 1 2011    11         22         2     635     825            AA      1903
## 2 2011     8         25         4     602     758            MQ      3265
## 3 2011    10         12         3     559     738            MQ      3265
## 4 2011     5          2         1     716     854            AA      2237
## 5 2011     4          4         1     741     949            AA      1225
## 6 2011     4          4         1     627     742            MQ      3265
##   TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 1  N477AA               110      81       40        0    IAH  DFW      224
## 2  N633MQ               116      74       53        2    HOU  DFW      247
## 3  N632MQ                99      71       33       -1    HOU  DFW      247
## 4  N552AA                98      70       29        1    IAH  DFW      224
## 5  N4XVAA               128      63       89       31    IAH  DFW      224
## 6  N939MQ                75      62       37       27    HOU  DFW      247
##   TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 1     11      18         0                         0
## 2     21      21         0                         0
```

```
## 3      8      20         0                        0
## 4     11      17         0                        0
## 5      6      59         0                        0
## 6      3      10         0                        0
```

Manipulating Groups of Observation (summarize and group_by)

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

```
df15 = hflights %>%
  summarise(min_dist = min(Distance), max_dist = max(Distance))
head(df15)
```

```
##   min_dist max_dist
## 1       79     3904
```

16. Determine the longest distance for diverted flights, save statistic to max_div.

```
df16 = hflights %>%
  filter(Diverted == 1) %>%
  summarise(max_div = max(Distance))
head(df16)
```

```
##   max_div
## 1    3904
```

17. `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`).

```
df17 = hflights %>%
  summarise(n_obs = n(),
            n_carrier = n_distinct(UniqueCarrier),
            n_dest = n_distinct(Dest),
            dest100 = nth(Dest, 100))

head(df17)
```

```
##    n_obs n_carrier n_dest dest100
## 1 227496        15    116     DFW
```

18. 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 of 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.

```
df18 = hflights %>%
  mutate(diff = TaxiIn - TaxiOut) %>%
  filter(!is.na(diff)) %>%
```

```
  summarise(avg = mean(diff))
head(df18)
```

```
##         avg
## 1 -8.992064
```

19. 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.

```
df19 = hflights %>%
  select(Dest, UniqueCarrier, Distance, ActualElapsedTime) %>%
  mutate(RealTime = ActualElapsedTime + 100)

head(df19)
```

```
##   Dest UniqueCarrier Distance ActualElapsedTime RealTime
## 1  DFW            AA      224                60      160
## 2  DFW            AA      224                60      160
## 3  DFW            AA      224                70      170
## 4  DFW            AA      224                70      170
## 5  DFW            AA      224                62      162
## 6  DFW            AA      224                64      164
```

Grouped Summaries using 'group_by()' and 'summarize()'

20. For each destination, find the number of flights, the mean distance travelled, and the mean arrival delay.

```
library(ggplot2)
df20 = hflights %>%
  group_by(Dest) %>%
  summarize(count = n(),
            dist = mean(Distance, na.rm = T),
            delay = mean(ArrDelay, na.rm = T))

head(df20)
```

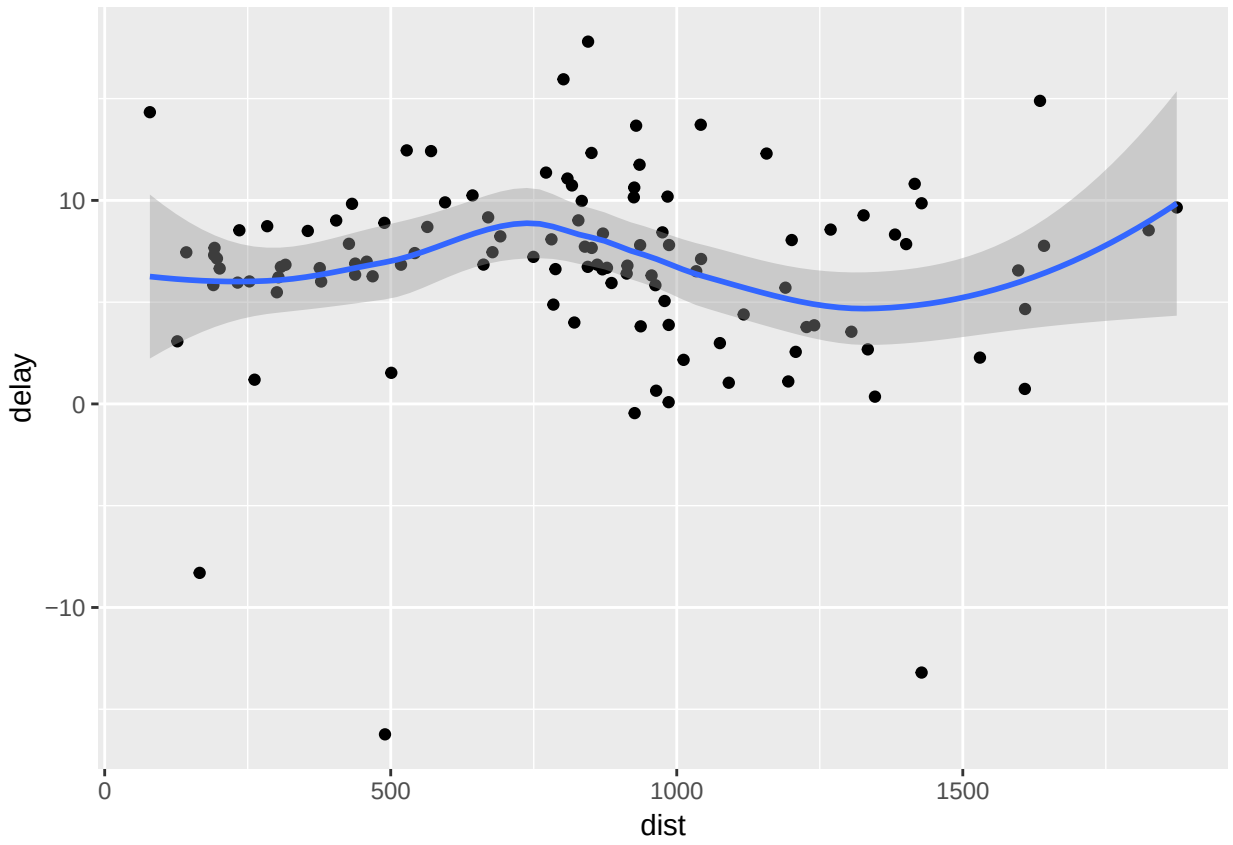
```
## # A tibble: 6 x 4
##   Dest  count  dist delay
##   <chr> <int> <dbl> <dbl>
## 1 ABQ    2812  749.  7.23
## 2 AEX     724  190   5.84
## 3 AGS       1  821   4
## 4 AMA    1297  518.  6.84
## 5 ANC     125 3266  26.1
## 6 ASE     125  914.  6.79
```

21. Investigate visually if there is any relationship between distance travelled and arrival delays. You can get rid of the outliers if any.

```
library(ggplot2)
hflights %>%
  group_by(Dest) %>%
  summarize(count = n(),
            dist = mean(Distance, na.rm = T),
            delay = mean(ArrDelay, na.rm = T)) %>%
```

```
  filter(dist < 2000) %>%
  ggplot(aes(dist, delay)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



22. Use `group_by()` and `summarise()` 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?

```
df21 = hflights %>%
   group_by(UniqueCarrier) %>%
   summarise(n_flights = n(),
             n_canc = sum(Cancelled == 1),
             avg_delay = mean(ArrDelay, na.rm = TRUE)) %>%
   arrange(avg_delay, n_canc)

head(df21)
```

```
## # A tibble: 6 x 4
##   UniqueCarrier n_flights n_canc avg_delay
##   <chr>             <int>  <int>     <dbl>
## 1 US                 4082     46    -0.631
## 2 AA                 3244     60     0.892
## 3 FL                 2139     21     1.85
```

```
## 4 AS                 365      0     3.19
## 5 YV                  79      1     4.01
## 6 DL                2641     42     6.08
```