Assignment 3: Flights of New York

FirstName LastName

2022-06-30

Exercise 1

- Question1 336776 rows and 19columns
- Question 2 Each observation contains year, month, day, dep_time, sched_dep_time, dep_delay, arrive time, sched_arr_time, arr_delay, carrier, flight, tailnum, origin, dest, air_time, distance hour and minute of the plane.
- Question 3 sced_dep_time is Scheduled arrive time. arr_time is the time planes really arrived.

Exercise 2

```
flights %>%
  select(year, month)
```

```
## # A tibble: 336,776 x 2
##     year month
##    <int> <int>
##  1  2013     1
##  2  2013     1
##  3  2013     1
##  4  2013     1
##  5  2013     1
##  6  2013     1
##  7  2013     1
##  8  2013     1
##  9  2013     1
## 10  2013     1
## # ... with 336,766 more rows
```

```
It extract the year and month variable(columns).
```

Exercise 3

```
flights %>%
  select(year:month)
```

```
## # A tibble: 336,776 x 2
##     year month
##    <int> <int>
##  1  2013     1
##  2  2013     1
##  3  2013     1
##  4  2013     1
##  5  2013     1
##  6  2013     1
##  7  2013     1
##  8  2013     1
##  9  2013     1
## 10  2013     1
## # ... with 336,766 more rows
```

colon(n:m) means 'contains columns from n column to m column', or designate the several coloumn

Exercise 4

```
flights %>%
  arrange(air_time, distance)
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1    16     1355           1315        40     1442           1411
##  2  2013     4    13      537            527        10      622            628
##  3  2013     2     3     2153           2129        24     2247           2224
##  4  2013     2    12     2123           2130        -7     2211           2225
##  5  2013     3     8     2026           1935        51     2131           2056
##  6  2013    12     6      922            851        31     1021            954
##  7  2013     2     5     1303           1315       -12     1342           1411
##  8  2013     3    18     1456           1329        87     1533           1426
##  9  2013     3    19     2226           2145        41     2305           2246
## 10  2013     5     8       16           2159       137       53           2304
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

air_time column sorted first.

Exercise 5

```
flights %>%
  arrange(desc(month))
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013    12     1       13           2359        14      446            445
##  2  2013    12     1       17           2359        18      443            437
##  3  2013    12     1      453            500        -7      636            651
##  4  2013    12     1      520            515         5      749            808
##  5  2013    12     1      536            540        -4      845            850
##  6  2013    12     1      540            550       -10     1005           1027
##  7  2013    12     1      541            545        -4      734            755
##  8  2013    12     1      546            545         1      826            835
##  9  2013    12     1      549            600       -11      648            659
## 10  2013    12     1      550            600       -10      825            854
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
flights %>%
  arrange(desc(dep_delay))
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     9      641            900      1301     1242           1530
##  2  2013     6    15     1432           1935      1137     1607           2120
##  3  2013     1    10     1121           1635      1126     1239           1810
##  4  2013     9    20     1139           1845      1014     1457           2210
##  5  2013     7    22      845           1600      1005     1044           1815
##  6  2013     4    10     1100           1900       960     1342           2211
##  7  2013     3    17     2321            810       911      135           1020
##  8  2013     6    27      959           1900       899     1236           2226
##  9  2013     7    22     2257            759       898      121           1026
## 10  2013    12     5      756           1700       896     1058           2020
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

The 51 flight experienced the longest dep_delay

Exercise 6

```
flights %>%
  mutate(
    average_speed = distance / (air_time / 60)
  )
```

```
## # A tibble: 336,776 x 20
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 12 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>,
## #   average_speed <dbl>
```

- Question1 The new column added as a last column.
- Question2 The average_speed is the name of the column. The name of the column is right before the '=' equal sign.

Exercise 7

```
flights %>%
  mutate(
    dep_time_hour = dep_time %/% 100,
    dep_time_minute = dep_time %% 100,
    dep_time_minutes_midnight = dep_time %%1200
  )
```

```
## # A tibble: 336,776 x 22
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 14 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>,
## #   dep_time_hour <dbl>, dep_time_minute <dbl>, dep_time_minutes_midnight <dbl>
```

Exercise 8

```
flights %>%
  filter(arr_delay < 0 & carrier =="AA"
  )
```

```
## # A tibble: 20,769 x 19
##      year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##     <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      606            610        -4      858            910
##  2  2013     1     1      628            630        -2     1137           1140
##  3  2013     1     1      656            659        -3      949            959
##  4  2013     1     1      659            700        -1     1008           1015
##  5  2013     1     1      712            715        -3     1023           1035
##  6  2013     1     1      739            745        -6      918            930
##  7  2013     1     1      753            755        -2     1056           1110
##  8  2013     1     1      803            810        -7      903            925
##  9  2013     1     1      840            845        -5     1311           1350
## 10  2013     1     1      940            945        -5     1119           1130
## # ... with 20,759 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Exercise 9

```
flights %>%
  group_by(carrier) %>%
  mutate(
    average_arr_delay = mean(arr_delay, na.rm = TRUE)
  ) %>% arrange(desc(average_arr_delay))
```

```
## # A tibble: 336,776 x 20
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      833            835        -2     1134           1102
##  2  2013     1     1     1716           1730       -14     1947           1953
##  3  2013     1     2      827            835        -8     1120           1102
##  4  2013     1     2     1728           1730        -2     1952           1953
##  5  2013     1     3      835            835         0     1102           1102
##  6  2013     1     3     1933           1730       123     2131           1953
##  7  2013     1     4      834            835        -1     1059           1102
##  8  2013     1     4     1831           1730        61     2029           1953
##  9  2013     1     5      835            835         0     1057           1102
## 10  2013     1     5     1726           1730        -4     1948           1953
## # ... with 336,766 more rows, and 12 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>,
## #   average_arr_delay <dbl>
```

```
flights %>%
  group_by(carrier) %>%
  mutate(
    average_arr_delay = mean(arr_delay, na.rm = TRUE)
  ) %>% arrange((average_arr_delay))
```

```
## # A tibble: 336,776 x 20
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      724            725        -1     1020           1030
##  2  2013     1     1     1808           1815        -7     2111           2130
##  3  2013     1     2      722            725        -3      949           1030
##  4  2013     1     2     1818           1815         3     2131           2130
##  5  2013     1     3      724            725        -1     1012           1030
##  6  2013     1     3     1817           1815         2     2121           2130
##  7  2013     1     4      725            725         0     1031           1030
##  8  2013     1     4     1808           1815        -7     2101           2130
##  9  2013     1     5      725            725         0     1011           1030
## 10  2013     1     5     1803           1815       -12     2118           2130
## # ... with 336,766 more rows, and 12 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>,
## #   average_arr_delay <dbl>
```

- Question1 "F9" has the longest arrival delays on average.
- Question2 "AS" has the shortest arrival delays on average.

Exercise 10

```
flights_to_miami <- flights %>%
  filter(dest == "MIA" & arr_delay <0) %>%
  select(arr_delay, carrier)
```

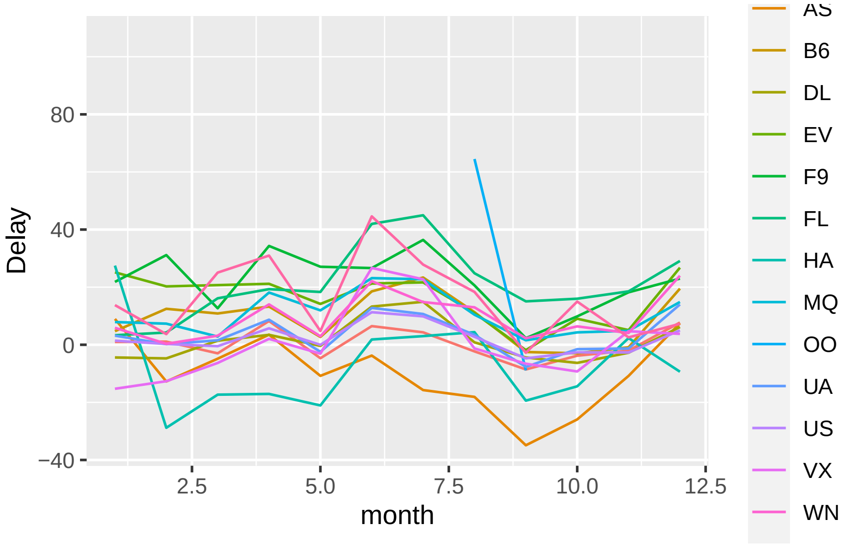
Exercise 11

```
monthly_delays <- flights %>%
  group_by(month, carrier) %>%
  summarize(
    arrival_delay = mean(arr_delay, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  spread(carrier, arrival_delay) %>%
  select(-`9E`)
```

```
pivoted_monthly_delays <- monthly_delays %>%
  pivot_longer(cols = AA:YV, names_to = 'Airline', values_to = "Delay")
```

```
qplot(x = month,
  y = Delay,
  color = Airline,
  geom ="line",
  data = pivoted_monthly_delays
  )
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```


Delay
month
AS
B6
DL
EV
F9
FL
HA
MQ
OO
UA
US
VX
WN