Analysis of Flights

I will be analyzing the flights data set and answering various prompts

Libraries

```
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
```

```
library(readr)
library(memoise)
```

```
## Warning: package 'memoise' was built under R version 4.4.3
```

Data set

```
flights = as_tibble(data.table::fread("https://github.com/Rdatatable/data.table/blob/mas
flights
```

```
## # A tibble: 253,316 x 11
##     year month   day dep_delay arr_delay carrier origin dest  air_time distance
##    <int> <int> <int>     <int>     <int> <chr>   <chr>  <chr>    <int>    <int>
##  1  2014     1     1        14        13 AA      JFK    LAX        359     2475
```

```
##  2  2014     1     1        -3         13 AA      JFK    LAX         363     2475
##  3  2014     1     1         2          9 AA      JFK    LAX         351     2475
##  4  2014     1     1        -8        -26 AA      LGA    PBI         157     1035
##  5  2014     1     1         2          1 AA      JFK    LAX         350     2475
##  6  2014     1     1         4          0 AA      EWR    LAX         339     2454
##  7  2014     1     1        -2        -18 AA      JFK    LAX         338     2475
##  8  2014     1     1        -3        -14 AA      JFK    LAX         356     2475
##  9  2014     1     1        -1        -17 AA      JFK    MIA         161     1089
## 10  2014     1     1        -2        -14 AA      JFK    SEA         349     2422
## # i 253,306 more rows
## # i 1 more variable: hour <int>
```

Analysis

This section will mainly involve demonstrating an understanding of different aspects of dplyr rather than focusing on a prompt.

Finding the average dep_delay by carrier

```
flights %>%
  group_by(carrier) %>%
  summarise(avg_dep_delay = mean(dep_delay))
```

```
## # A tibble: 14 x 2
##    carrier avg_dep_delay
##    <chr>           <dbl>
##  1 AA               8.51
##  2 AS               8.83
##  3 B6              12.0
##  4 DL              12.2
##  5 EV              17.6
##  6 F9              24.7
##  7 FL              20.6
##  8 HA               8.49
##  9 MQ               8.06
## 10 OO              12.6
## 11 UA              14.3
## 12 US               3.52
## 13 VX              10.4
## 14 WN              18.9
```

Finding total delay of each flight

```r
flights %>%
  mutate(tot_delay = arr_delay + dep_delay) %>%
  select(carrier, origin, tot_delay)
```

```
## # A tibble: 253,316 x 3
##    carrier origin tot_delay
##    <chr>   <chr>      <int>
##  1 AA      JFK           27
##  2 AA      JFK           10
##  3 AA      JFK           11
##  4 AA      LGA          -34
##  5 AA      JFK            3
##  6 AA      EWR            4
##  7 AA      JFK          -20
##  8 AA      JFK          -17
##  9 AA      JFK          -18
## 10 AA      JFK          -16
## # i 253,306 more rows
```

Top 5 origin-dest pairs with most flights?

```r
flights %>%
  group_by(origin, dest) %>%
  summarise(N = n()) %>%
  arrange(desc(N)) %>%
  head(5)
```

```
## `summarise()` has grouped output by 'origin'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 5 x 3
## # Groups:   origin [2]
##   origin dest      N
##   <chr>  <chr> <int>
## 1 JFK    LAX   10208
## 2 JFK    SFO    7368
## 3 LGA    ORD    7052
## 4 LGA    ATL    6925
## 5 LGA    MIA    5084
```

Average departure and arrival delay by month

```r
flights %>%
  group_by(month) %>%
  summarise(avg_dep_delay = mean(dep_delay),
            avg_arr_delay = mean(arr_delay))
```

```
## # A tibble: 10 x 3
##    month avg_dep_delay avg_arr_delay
##    <int>         <dbl>         <dbl>
##  1     1         23.0         20.8
##  2     2         17.8         17.4
##  3     3          8.93         4.58
##  4     4         10.2          7.26
##  5     5         13.7          7.78
##  6     6         14.1          8.82
##  7     7         16.5         12.3
##  8     8         10.0          3.59
##  9     9          4.74         0.478
## 10    10          7.85         1.76
```

Which days of the month have the most on-time arrivals?

```r
flights %>%
  group_by(day) %>%
  filter(arr_delay <= 0) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 31 x 2
##      day count
##    <int> <int>
##  1    20  5820
##  2    19  5571
##  3    25  5439
##  4    28  5396
##  5    17  5378
##  6    24  5375
##  7    18  5191
##  8    27  5136
##  9     7  5007
## 10    26  4992
## # i 21 more rows
```

Delay percentiles by carrier

```r
flights %>%
  group_by(carrier) %>%
  summarise(arr_delay_quan = quantile(arr_delay)) # fix later
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated i
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `summarise()` has grouped output by 'carrier'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 70 x 2
## # Groups:   carrier [14]
##    carrier arr_delay_quan
##    <chr>            <dbl>
##  1 AA               -72
##  2 AA               -17
##  3 AA                -5
##  4 AA                12
##  5 AA              1494
##  6 AS               -70
##  7 AS               -28.8
##  8 AS               -14.5
##  9 AS                 3
## 10 AS               290
## # i 60 more rows
```