Assignment 2

Anthony Cunningham

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)

# Change working dir in RMarkdown cell
knitr::opts_knit$set(root.dir =
'C:/Users/AC069015/kumc_applied_stats/data_824_data_viz_and_acquisition/2_data_wrangling'
)

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
```

```
library(stringr)
library(tidyr)
library(ggplot2)

df <- hflights
```

Exercise 1

```
total_flights <- nrow(df)

t11 <- df %>%
    mutate(Cancelled = if_else(Cancelled == "1", TRUE, FALSE, missing = FALSE)) %>%
    group_by(Cancelled) %>%
    summarise(
        n_flights = n(),
        perc_flights = round(100 * (n() / total_flights), 2),
    ) %>%
    arrange(desc(Cancelled))

t11
```

```
## # A tibble: 2 × 3
##   Cancelled n_flights perc_flights
##   <lgl>         <int>        <dbl>
## 1 TRUE           2973         1.31
## 2 FALSE        224523        98.7
```

```
carriers <- df %>%
    group_by(UniqueCarrier) %>%
    summarise(n_flights = n()) %>%
    arrange(UniqueCarrier)

t12 <- df %>%
    mutate(Cancelled = as.character(Cancelled)) %>%
    mutate(Cancelled = if_else(Cancelled == "1", "Cancelled", "Not Cancelled", missing = "Not Ca
ncelled")) %>%
    left_join(carriers, by = "UniqueCarrier") %>%
    group_by(UniqueCarrier, Cancelled) %>%
    summarise(perc_flights = round(100 * (n() / n_flights), 2)) %>%
    arrange(UniqueCarrier, Cancelled) %>%
    distinct() %>%
    pivot_wider(names_from = Cancelled, values_from = perc_flights) %>%
    arrange(desc(Cancelled))

t12
```

```
## # A tibble: 15 × 3
## # Groups:   UniqueCarrier, Cancelled [15]
##    UniqueCarrier Cancelled `Not Cancelled`
##    <chr>             <dbl>           <dbl>
##  1 EV                 3.45            96.6
##  2 MQ                 2.9             97.1
##  3 B6                 2.59            97.4
##  4 AA                 1.85            98.2
##  5 UA                 1.64            98.4
##  6 DL                 1.59            98.4
##  7 WN                 1.55            98.4
##  8 XE                 1.55            98.4
##  9 OO                 1.39            98.6
## 10 YV                 1.27            98.7
## 11 US                 1.13            98.9
## 12 FL                 0.98            99.0
## 13 F9                 0.72            99.3
## 14 CO                 0.68            99.3
## 15 AS                NA              100
```

```
t13 <- df %>%
    mutate(Cancelled = as.character(Cancelled)) %>%
    mutate(Cancelled = if_else(Cancelled == "1", "Cancelled", "Not Cancelled", missing = "Not Ca
ncelled")) %>%
    filter(Cancelled == "Cancelled") %>%
    group_by(UniqueCarrier, CancellationCode) %>%
    summarise(n_flights = n()) %>%
    pivot_wider(names_from = CancellationCode, values_from = n_flights) %>%
    arrange(UniqueCarrier)

t13
```

```
## # A tibble: 14 × 5
## # Groups:   UniqueCarrier [14]
##    UniqueCarrier     A     B     C     D
##    <chr>         <int> <int> <int> <int>
##  1 AA               20    29    11    NA
##  2 B6                5    13    NA    NA
##  3 CO               37   436     2    NA
##  4 DL               13    27     2    NA
##  5 EV               60    14     2    NA
##  6 F9                2     4    NA    NA
##  7 FL                8    12     1    NA
##  8 MQ               39    71    25    NA
##  9 OO              121    87    15     1
## 10 UA               21    10     3    NA
## 11 US               27    17     2    NA
## 12 WN              517   181     5    NA
## 13 XE              331   751    50    NA
## 14 YV                1    NA    NA    NA
```

Exercise 2

```
t21 <- df %>%
    mutate(
        Month = str_pad(Month, width = 2, side = "left", pad = "0"),
        DayofMonth = str_pad(DayofMonth, width = 2, side = "left", pad = "0")
        ) %>%
    unite("Date", Year, Month, DayofMonth, sep = "-") %>%
    mutate(Date = as.Date(Date, format = "%Y-%m-%d")) %>%
    group_by(Date, Origin) %>%
    summarise(n_flights = n()) %>%
    arrange(Date, Origin)

print("Only first 20 rows are printed out.")
```

```
## [1] "Only first 20 rows are printed out."
```

```
paste0("Total size of table is: ", nrow(t21), " rows by ", ncol(t21), " columns")
```

```
## [1] "Total size of table is: 730 rows by 3 columns"
```

```
head(t21, 20)
```

```
## # A tibble: 20 × 3
## # Groups:   Date [10]
##    Date       Origin n_flights
##    <date>     <chr>      <int>
##  1 2011-01-01 HOU          100
##  2 2011-01-01 IAH          452
##  3 2011-01-02 HOU          141
##  4 2011-01-02 IAH          537
##  5 2011-01-03 HOU          151
##  6 2011-01-03 IAH          551
##  7 2011-01-04 HOU          151
##  8 2011-01-04 IAH          432
##  9 2011-01-05 HOU          146
## 10 2011-01-05 IAH          444
## 11 2011-01-06 HOU          144
## 12 2011-01-06 IAH          516
## 13 2011-01-07 HOU          146
## 14 2011-01-07 IAH          515
## 15 2011-01-08 HOU          108
## 16 2011-01-08 IAH          392
## 17 2011-01-09 HOU          129
## 18 2011-01-09 IAH          473
## 19 2011-01-10 HOU          148
## 20 2011-01-10 IAH          511
```