

# Flights Analysis

Monica Buczynski

9/12/2020

Note: The purpose of this document is to showcase a sample of skills that I learned in R for Data Science (chapter: Model Building) by Garrett Golemund and Hadley Wickham. Some scripts were taken from <https://r4ds.had.co.nz/model-building.html>. The code for each exercise was studied carefully for understanding and then was retyped manually into R to maximize the learning experience; however, many of the scripts were altered for further analysis and presentation aesthetics or I added my own code for further analysis.

## Question: What affects the number of daily flights?

Counting the number of flights per day and visualising it with ggplot2:

```
daily <- flights %>%  
  mutate(date = make_date(year, month, day)) %>%  
  group_by(date) %>%  
  summarise(n = n())
```

*# same as ...*

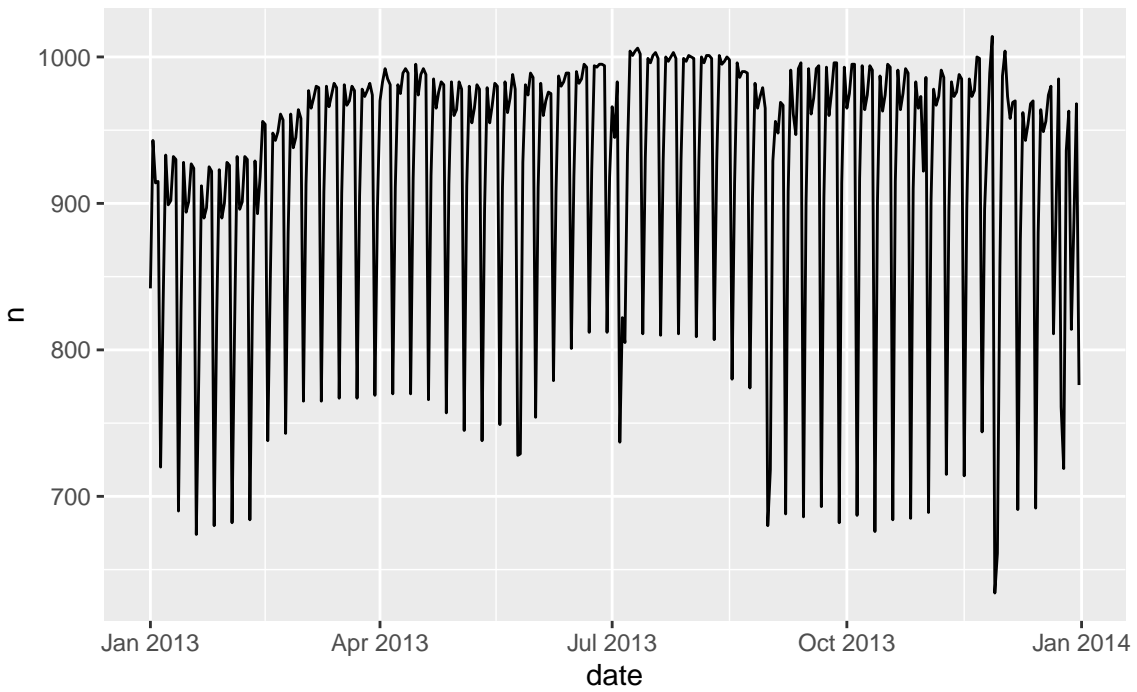
```
daily <- flights %>%  
  mutate(date = make_date(year, month, day)) %>%  
  group_by(date) %>%  
  count()
```

*#View preview of data*

```
head(daily,4)
```

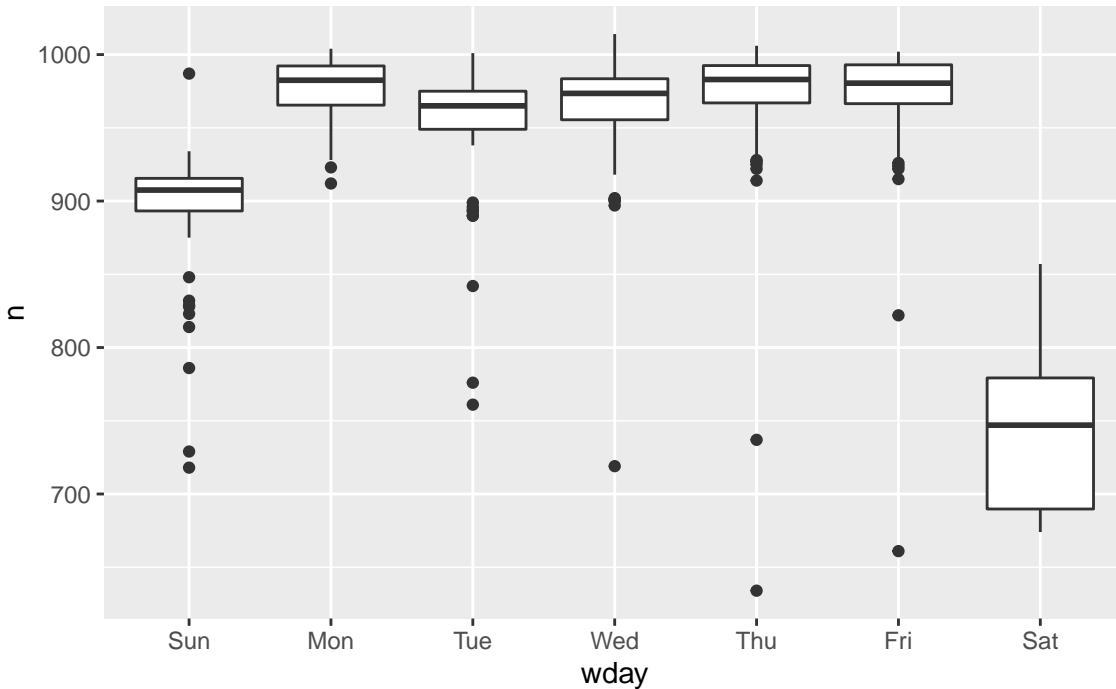
```
## # A tibble: 4 x 2  
## # Groups:   date [4]  
##   date          n  
##   <date>      <int>  
## 1 2013-01-01   842  
## 2 2013-01-02   943  
## 3 2013-01-03   914  
## 4 2013-01-04   915
```

```
# Graphical representation  
ggplot(daily,aes(date,n)) +  
  geom_line()
```

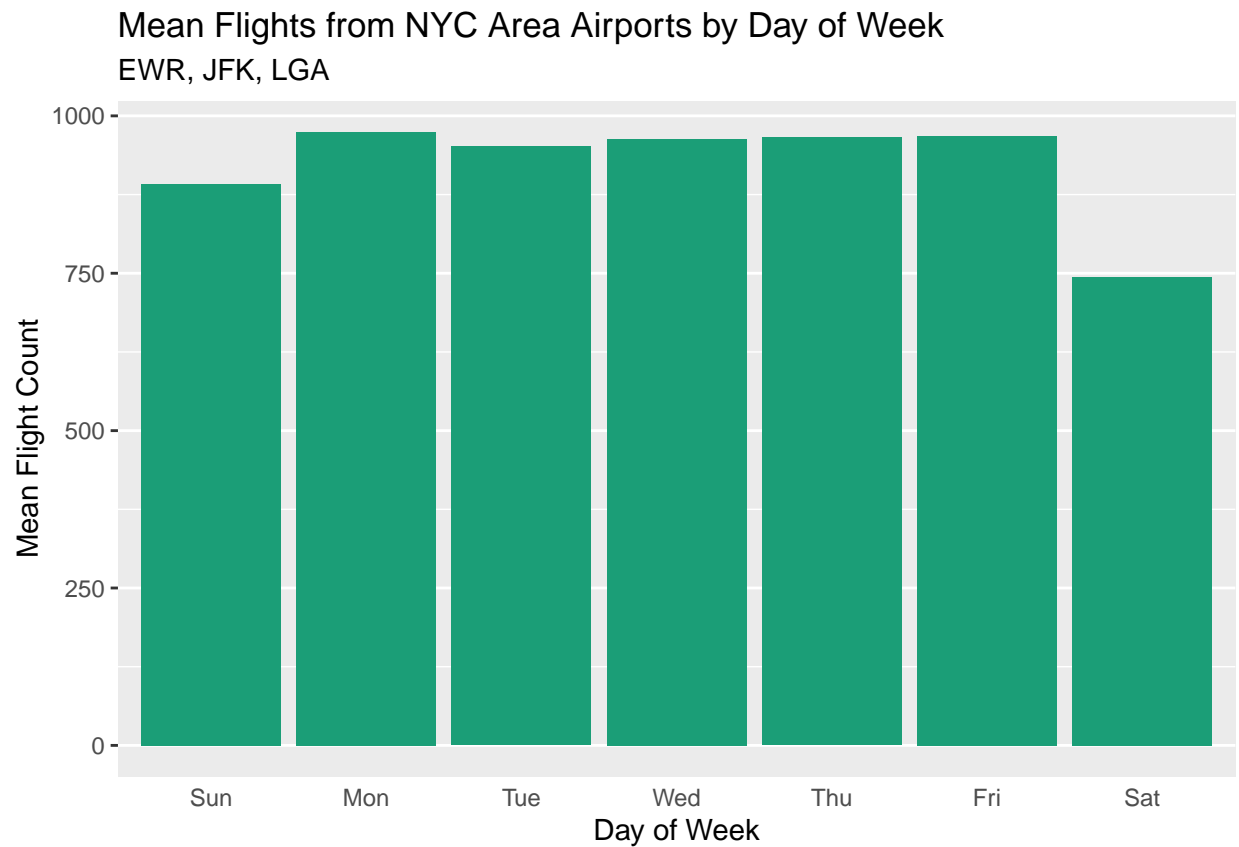


**Day-of-week effect:** There are fewer flights on weekends because most travel is for business. It would be rare for someone to leave on a Saturday for a Monday meeting as they would probably like to spend their weekend focused on leisure activities.

```
daily1 <- daily %>%  
  mutate(wday = wday(date, label = TRUE))  
  
ggplot(daily1, aes(wday, n)) +  
  geom_boxplot()
```



```
daily1 %>%
  group_by(wday) %>%
  summarise(mean_n = mean(n)) %>%
  ggplot(aes(x = wday, y = mean_n)) +
  geom_bar(stat = "identity", fill = "#1B9E77") +
  theme(axis.ticks.x = element_blank(), panel.grid.major.x = element_blank(), panel.grid.minor.x = element_blank())
labs(title = "Mean Flights from NYC Area Airports by Day of Week",
     subtitle = "EWR, JFK, LGA",
     x = "Day of Week",
     y = "Mean Flight Count")
```



Compute and visualise the residuals: We see that we have removed the day-of-week effect, but there still may be other patterns that exist.

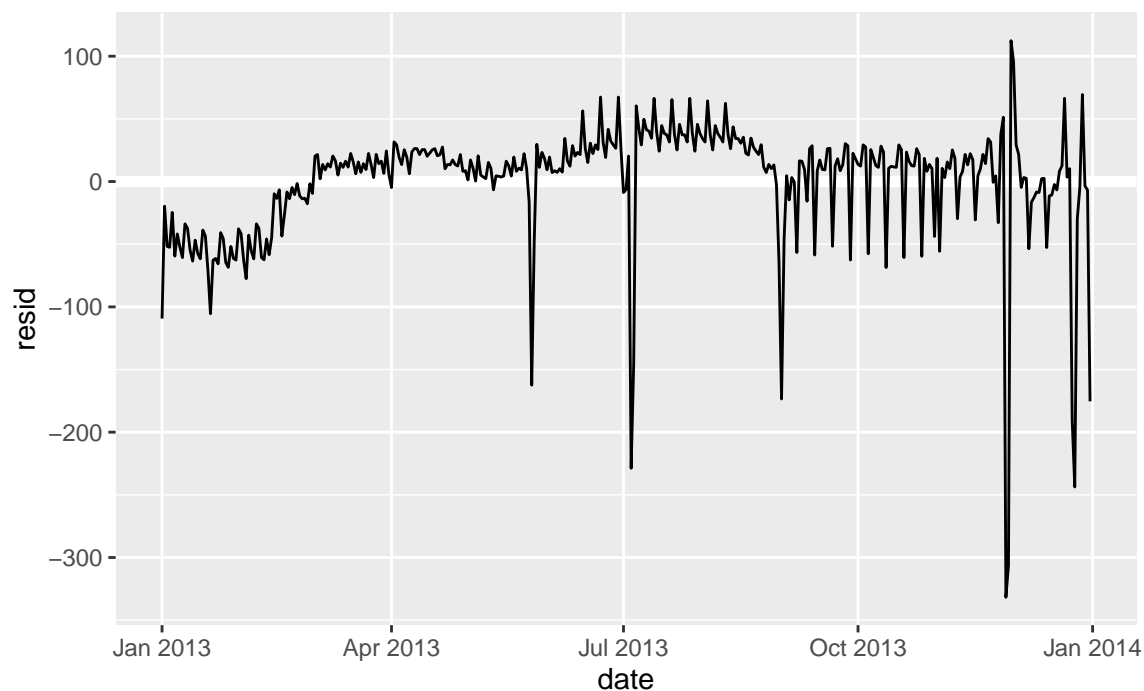
```
# Remove day-of-week effect by fitting a model
```

```
mod <- lm(n ~ wday, data = daily1)
```

```
(head(daily2 <- daily1 %>%  
  add_residuals(mod)))
```

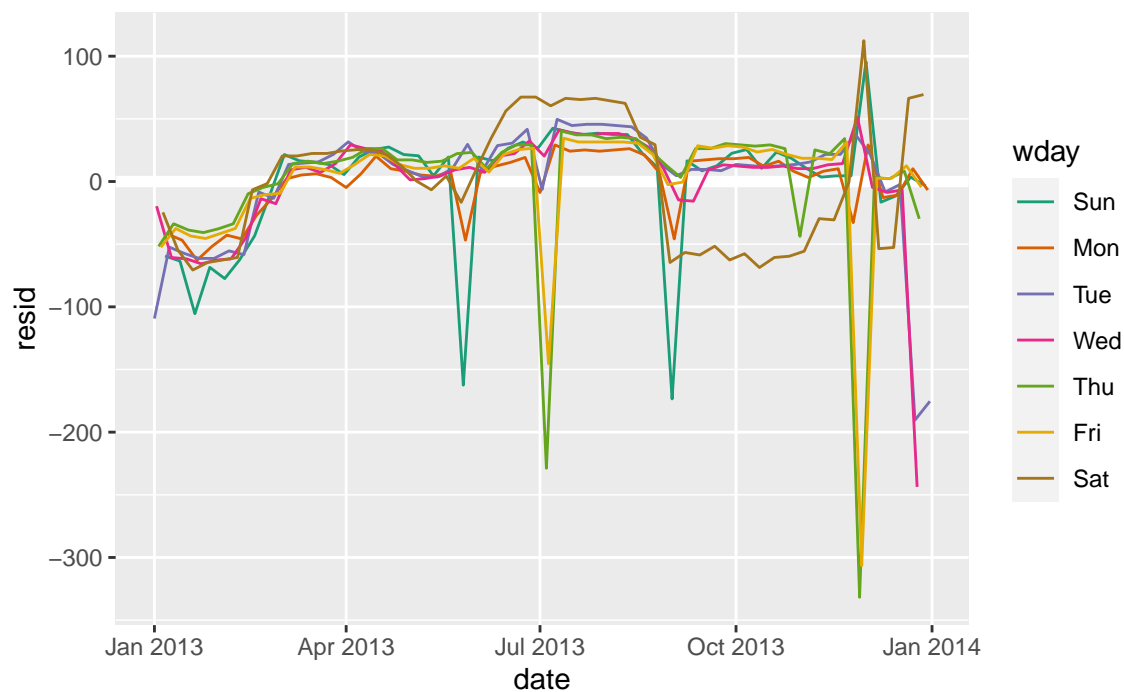
```
## # A tibble: 6 x 4  
## # Groups:   date [6]  
##   date          n wday  resid  
##   <date>      <int> <ord> <dbl>  
## 1 2013-01-01   842 Tue   -109.  
## 2 2013-01-02   943 Wed    -19.7  
## 3 2013-01-03   914 Thu   -51.8  
## 4 2013-01-04   915 Fri   -52.5  
## 5 2013-01-05   720 Sat   -24.6  
## 6 2013-01-06   832 Sun   -59.5
```

```
daily2 %>%  
  ggplot(aes(date, resid)) +  
  geom_ref_line(h = 0) +  
  geom_line() +  
  theme(panel.grid.minor.x = element_blank())
```



Drawing a plot with one line for each day of the week to more clearly see that our model has taken out the day-of-week effect, but that other patterns may still be present. It seems that Saturday flights change with season.

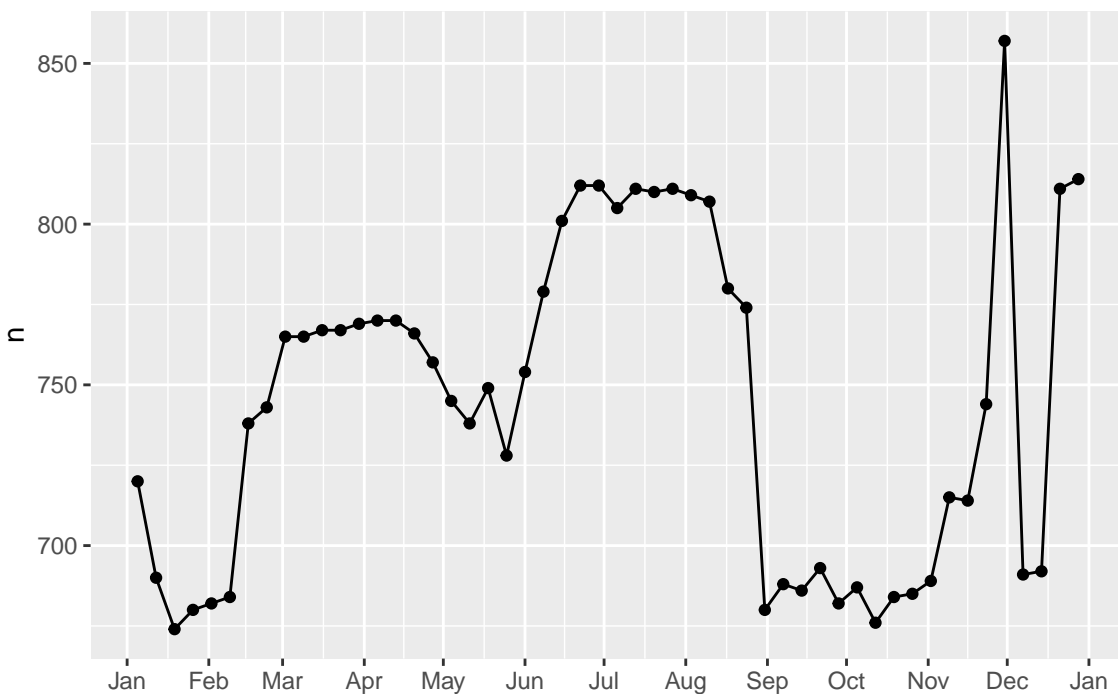
```
ggplot(daily2, aes(x = date, y = resid, colour = wday)) +  
  geom_ref_line(h = 0) +  
  geom_line() +  
  theme(panel.grid.minor.x = element_blank()) +  
  scale_color_brewer(palette = "Dark2")
```



Seasonal Saturday effect: Increased Saturday flights in the summer may be due to the increase of people taking summer vacations. For NY state, summer break in 2013 was Jun 26–Sep 9. A working hypothesis of why summer Saturday flights are more frequent in the summer than in the fall may be due to school being in session as well as the Thanksgiving, Christmas and New Years holidays.

```
# Visualization of Saturdays
```

```
daily1 %>%  
  filter(wday == "Sat") %>%  
  ggplot(aes(date, n)) +  
  geom_point() +  
  geom_line() +  
  scale_x_date(NULL, date_breaks = "1 month", date_labels = "%b")
```



```
# terms variable captures the fall, spring and summer school terms
```

```
term <- function(date) {  
  cut(date,  
    breaks = ymd(20130101, 20130605, 20130825, 20140101),  
    labels = c("spring", "summer", "fall")  
  )  
}
```

```
daily2 <- daily1 %>%  
  mutate(term = term(date))
```

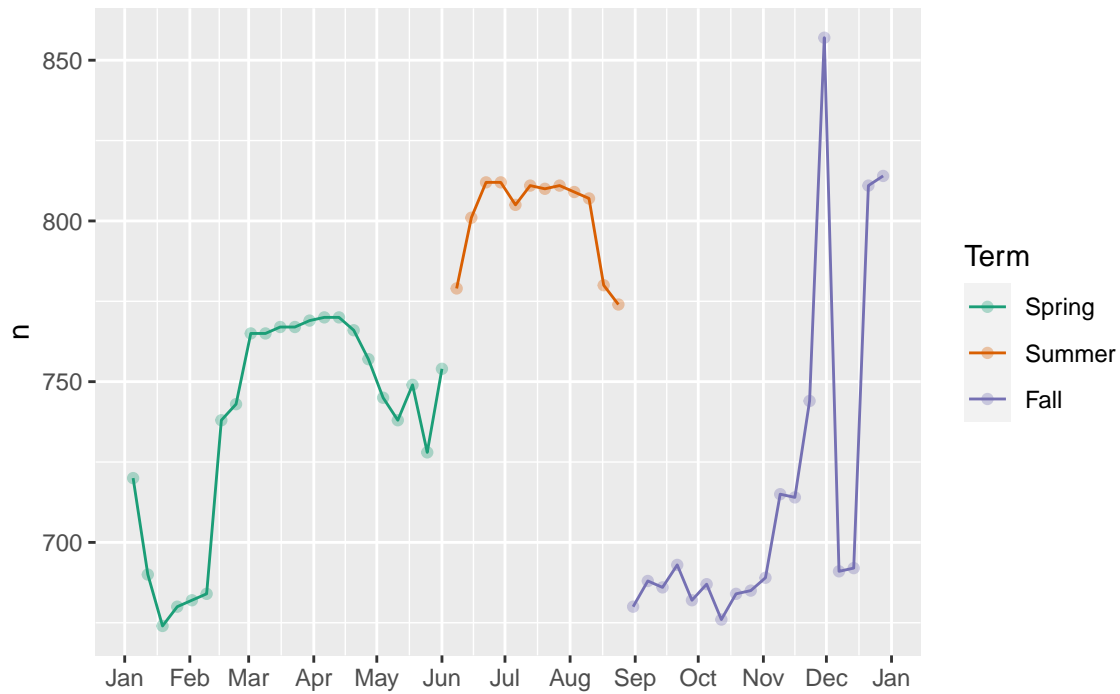
```
# View of first three rows
```

```
head(daily2, 3)
```

```
## # A tibble: 3 x 4  
## # Groups:   date [3]  
##   date           n wday term  
##   <date>         <int> <ord> <fct>  
## 1 2013-01-01     842 Tue  spring  
## 2 2013-01-02     943 Wed  spring  
## 3 2013-01-03     914 Thu  spring
```

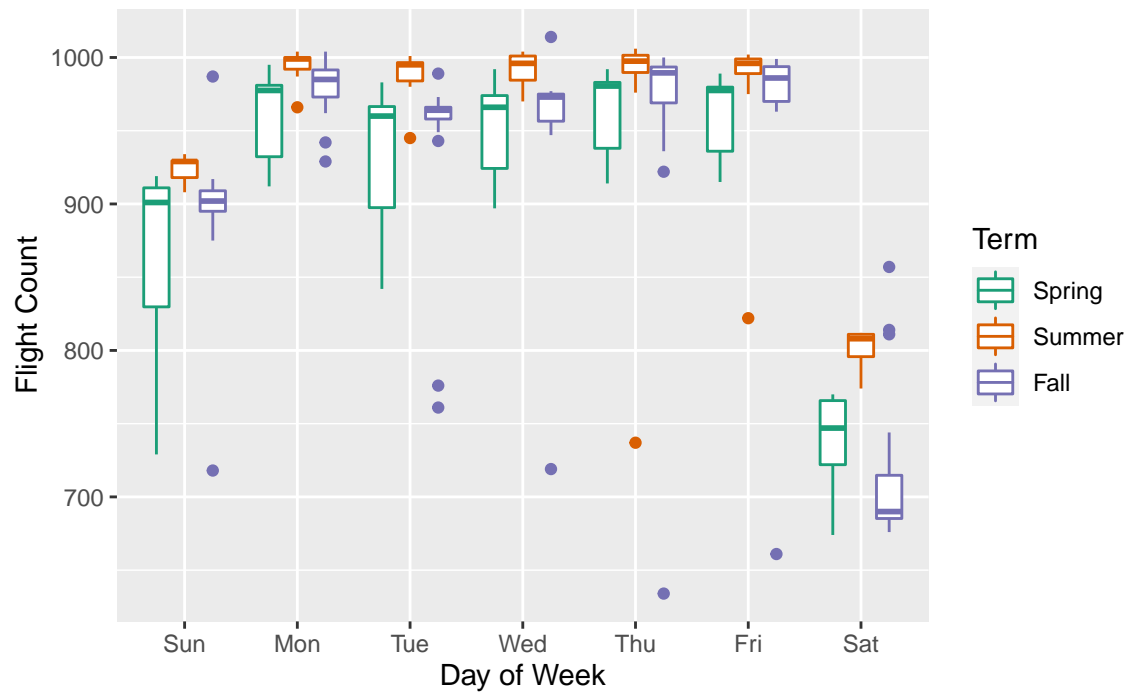


```
daily2 %>%
  filter(wday == "Sat") %>%
  ggplot(aes(x = date, y = n, colour = term)) +
  geom_line() +
  geom_point(alpha = 1/3) +
  scale_x_date(NULL, date_breaks = "1 month", date_labels = "%b") +
  scale_color_brewer(palette = "Dark2",
                    name = "Term",
                    labels = c("Spring", "Summer", "Fall")) # use "%b" for month abbreviations
```



```
# How term variable affects the other days of the week
```

```
daily2 %>%  
  ggplot(aes(x = wday, y = n, colour = term)) +  
  geom_boxplot() +  
  scale_color_brewer(palette = "Dark2", name = "Term", labels = c("Spring", "Summer", "Fall")) +  
  labs(x = "Day of Week",  
       y = "Flight Count")
```



```

# Make Monday first so that the weekend can be side by side

monday_first <- function(x) {
  fct_relevel(x, levels(x)[-1])
}

ggplot(daily2, aes(x = monday_first(wday), y = n, colour = term)) +
  geom_boxplot() +
  labs(x = "Day of Week", y = "Number of Flights") +
  scale_color_brewer(palette = "Dark2", name = "Term", labels = c("Spring", "Summer", "Fall"))

```

