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 Grolemund 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.

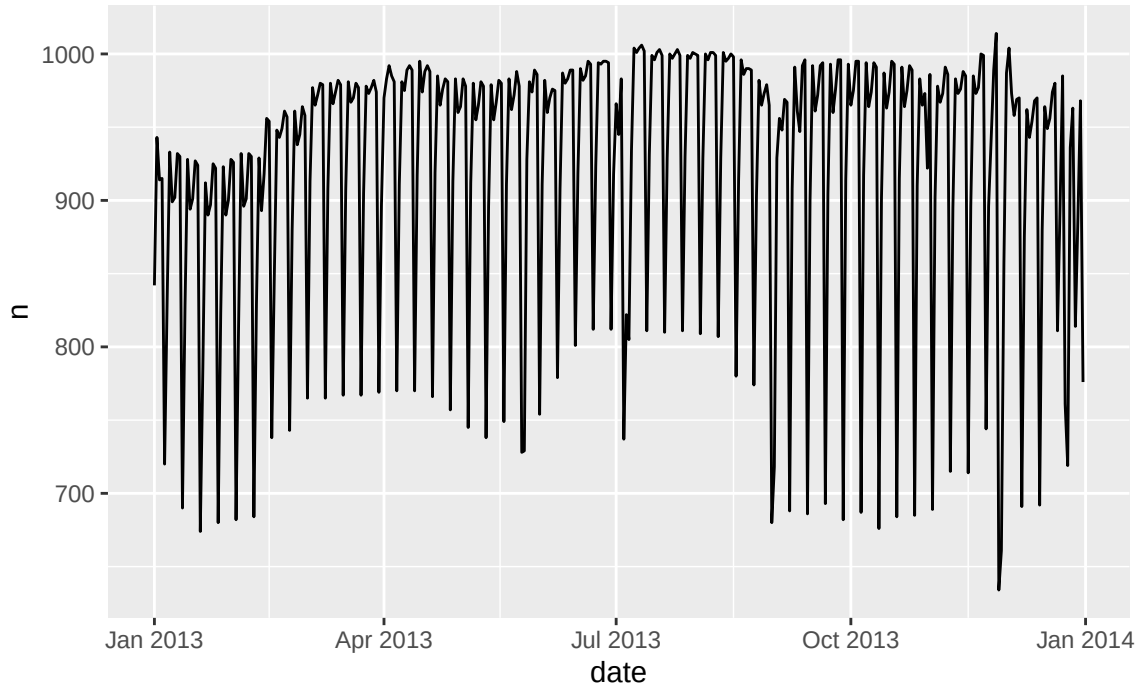
Question: What affects the number of daily flights?

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())

#View preview of data

daily
```

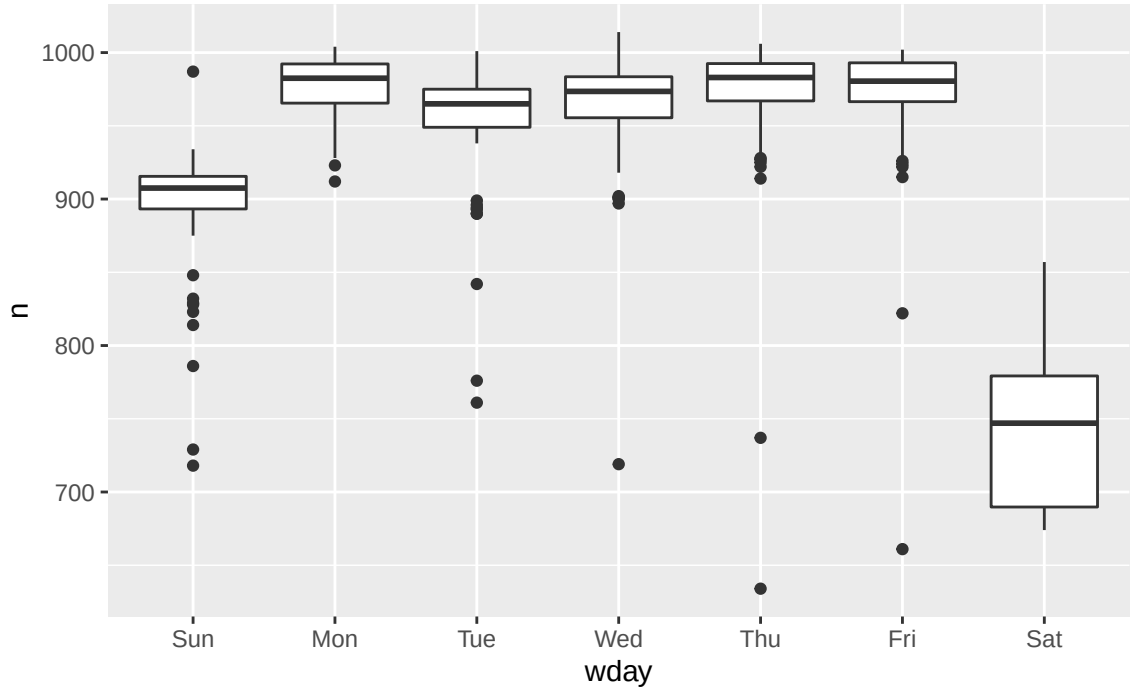
```
## # A tibble: 365 x 2
##    date           n
##    <date>     <int>
##  1 2013-01-01   842
##  2 2013-01-02   943
##  3 2013-01-03   914
##  4 2013-01-04   915
##  5 2013-01-05   720
##  6 2013-01-06   832
##  7 2013-01-07   933
##  8 2013-01-08   899
##  9 2013-01-09   902
## 10 2013-01-10   932
## # ... with 355 more rows
```

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

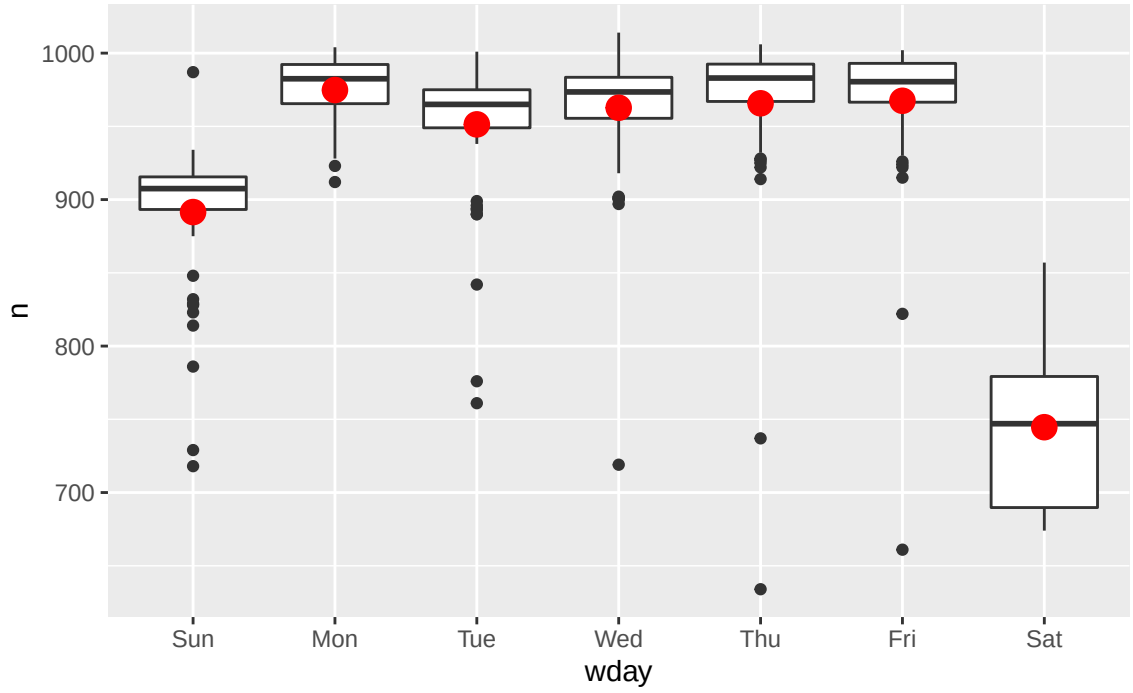


Day-of-week effect: There are fewer flights on weekends because most travel is for business. It would be rare for someone to leave on a Saturday for a Monday meeting as they would probably like to spend their weekend focused on leisure activities.

```
daily <- daily %>%
  mutate(wday = wday(date, label = TRUE))

ggplot(daily, aes(wday, n)) +
  geom_boxplot()
```

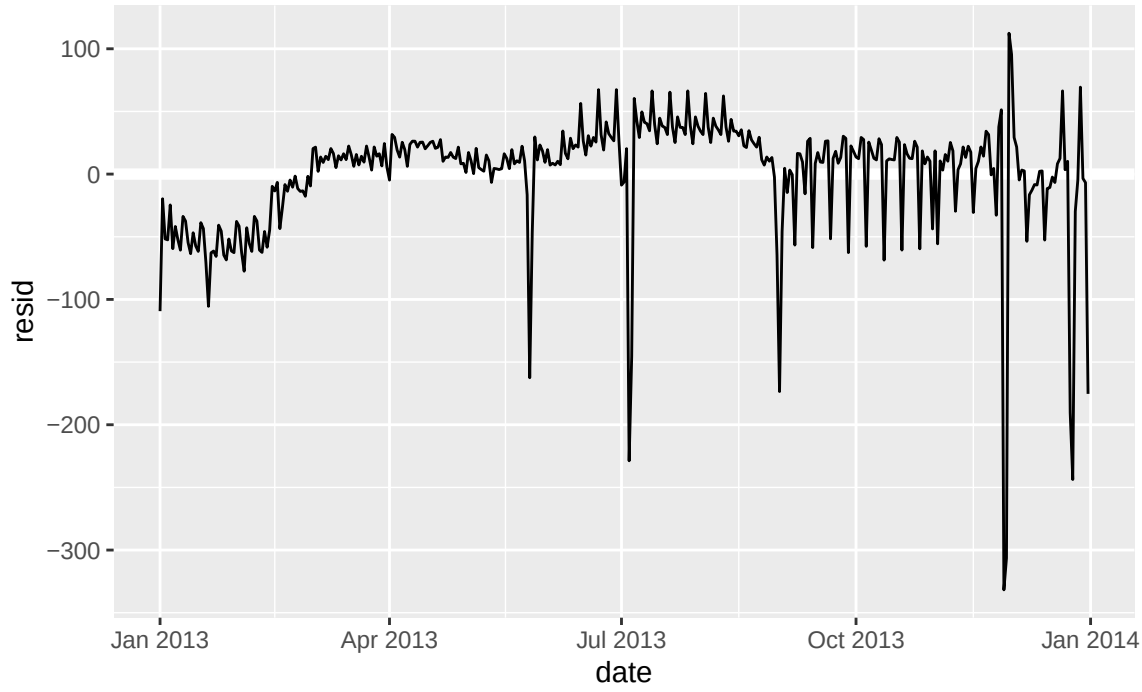


```r
# Remove day-of-week effect by fitting a model
mod <- lm(n ~ wday, data = daily)

grid <- daily %>%
  data_grid(wday) %>%
  add_predictions(mod, "n")


ggplot(daily, aes(wday, n)) +
  geom_boxplot() +
  geom_point(data = grid, colour = "red", size = 4)
```



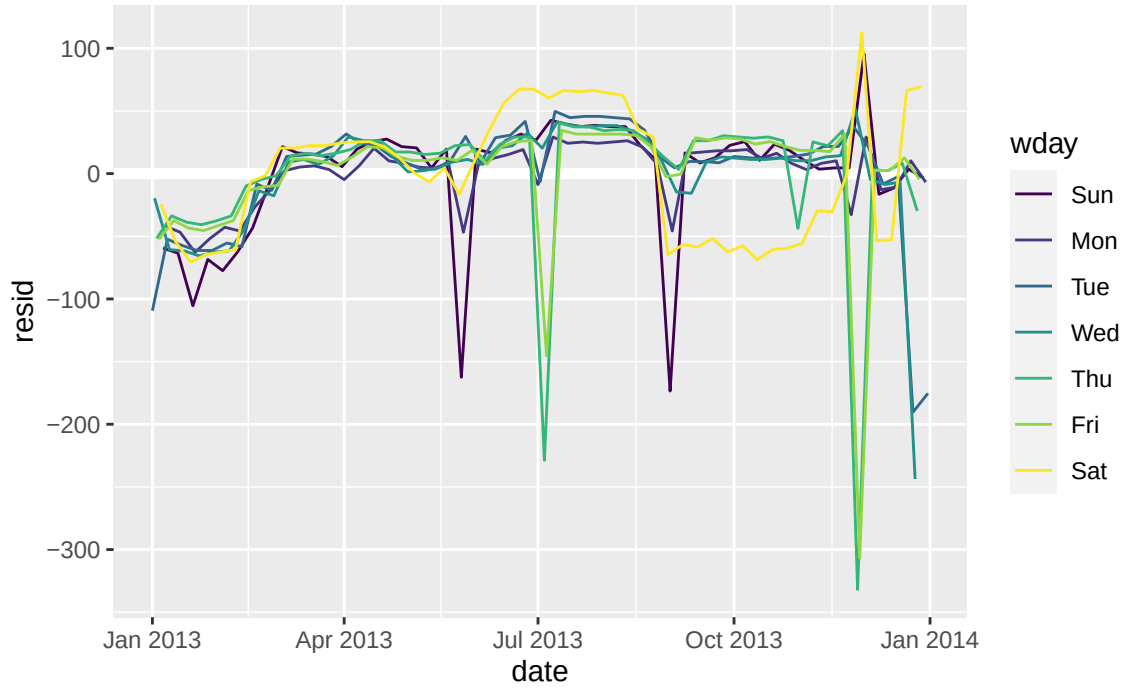
Compute and visualise the residuals: We see that we have removed the day-of-week effect, but there still may be other patterns that exist.

```
daily <- daily %>%
  add_residuals(mod)
daily %>%
  ggplot(aes(date, resid)) +
  geom_ref_line(h = 0) +
  geom_line()
```



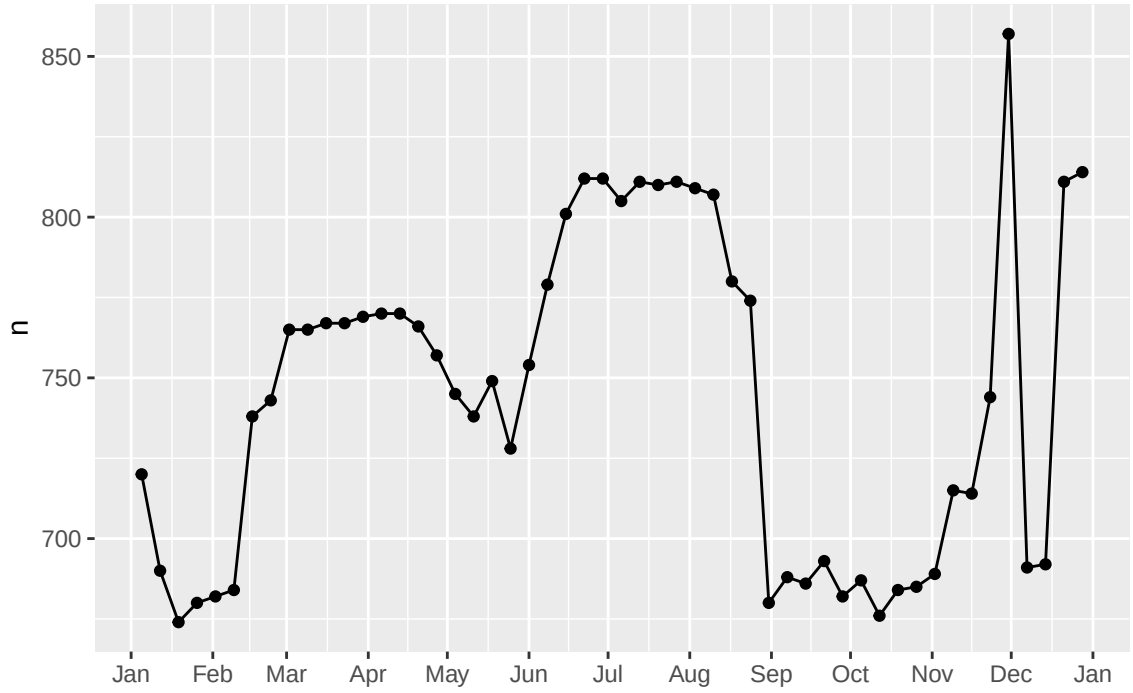
Drawing a plot with one line for each day of the week to more clearly see that our model has taken out the day-of-week effect, but that other patterns may still be present. It seems that Saturday flights change with season.

```
ggplot(daily, aes(date, resid, colour = wday)) +
  geom_ref_line(h = 0) +
  geom_line()
```



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.

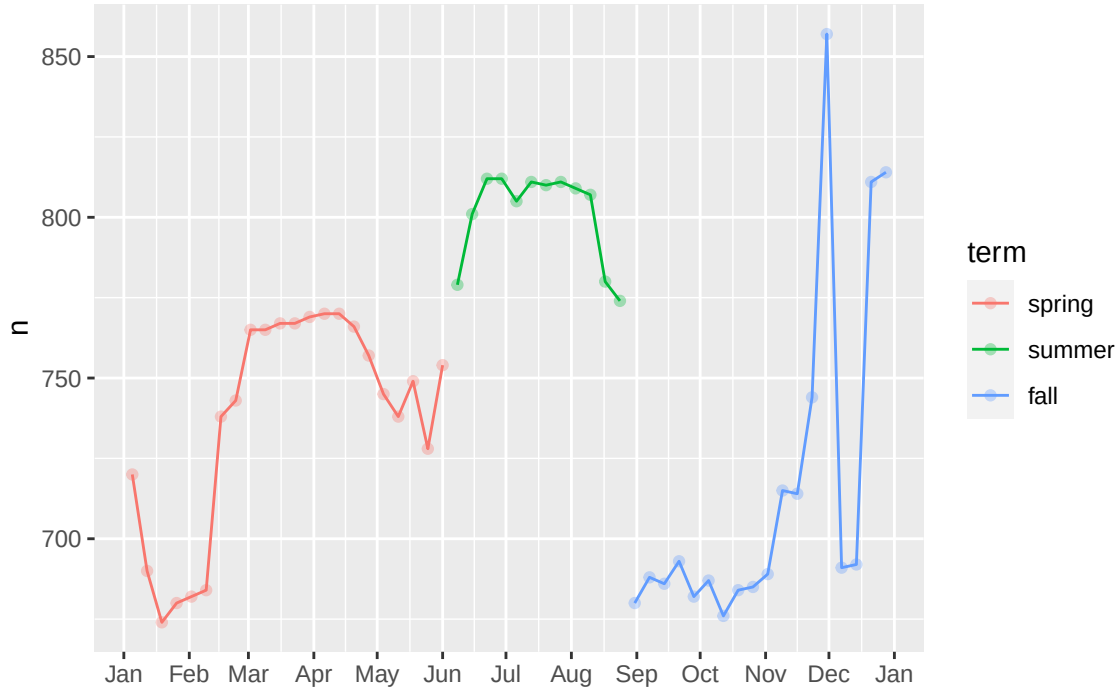
```
# Visualization of Saturdays

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



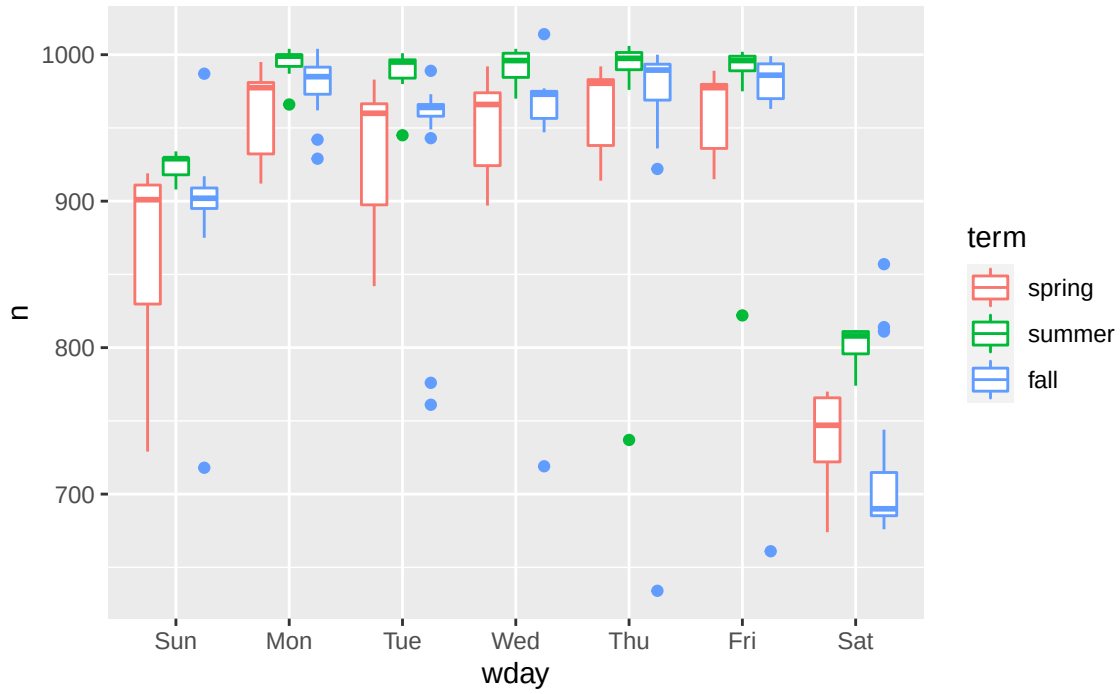
```r
# terms variable captures the fall, spring and summer school terms

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

daily <- daily %>%
  mutate(term = term(date))

# View of first three rows

head(daily, 3)
```

```
## # A tibble: 3 x 5
##   date           n wday   resid term
##   <date>     <int> <ord>  <dbl> <fct>
## 1 2013-01-01   842 Tue   -109.  spring
## 2 2013-01-02   943 Wed    -19.7 spring
## 3 2013-01-03   914 Thu    -51.8 spring
```

```r
daily %>%
  filter(wday == "Sat") %>%
  ggplot(aes(date, n, colour = term)) +
  geom_point(alpha = 1/3) +
  geom_line() +
  scale_x_date(NULL, date_breaks = "1 month", date_labels = "%b") # use "%b" for month abbreviations
```



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

daily %>%
  ggplot(aes(wday, n, colour = term)) +
  geom_boxplot()
```



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

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

daily <- daily %>%
  mutate(wday = wday(date, label = TRUE))
 ggplot(daily, aes(monday_first(wday),n, colour = term)) +
  geom_boxplot() +
   labs(x = "Day of Week", y = "Number of flights")
```