Model Building Homework

1. (1 pt) What do you think is the cause for the days with the highest and lowest residual values? Look at all days with abs(resid) > 80.

term <- function(date) {

cut(date,

breaks = ymd(20130101, 20130301, 20130605, 20130825, 20140101),

labels = c("winter","spring", "summer", "fall")

)

}

# add day of week and term to count by date

daily <- flights %>%

mutate(date = make\_date(year, month, day)) %>%

count(date) %>%

mutate(wday = wday(date, label = TRUE)) %>%

mutate(term = term(date))

mod1 <- lm(n ~ wday \* term, data = daily)

daily\_res <- daily %>%

add\_residuals(mod1, "resid")

1. (2 pts) Create a new variable that splits the wday variable into terms (seasons), but only for Saturdays, i.e. it should have Thurs, Fri, but Sat-winter, Sat-summer, Sat-spring, Sat-fall. Use cutoff dates of March 1, June 5, Aug 25 to separate into seasons. How does this model compare with the model with every combination of wday and term? Plot both model residuals side by side.
2. (2 pts) Create a new variable that combines the day of week, term (for Saturdays), and public holidays that you identified in number 1. What do the residuals of that model look like? Do this first as one factor called holiday for all the dates with large absolute value residuals. Then do it again with one factor for the high residual dates and another for the low residual dates. Which model works better?
3. (2 pts) Create a variable that contains the day of the week unless it is one of the two types you identified in the second part of number 3. Produce a model that models n based on both that variable and the term as well as their interaction. Plot the residuals.
4. (3 pts) Use what you have learned above to predict the number of flights for 2019 per day. Print a graph that overlays the number of flights in 2013 with your number of predicted flights in 2019. How many flights do you predict for each day June 20 - July 10 of 2019?