Exploring the Effect of Holidays on Airbnb Reviews: A Data Analysis Project

Project 2 - MAT214

Joel Joseph Jaison

2023-05-04

Introduction And Description:

The Airbnb dataset is a comprehensive collection of information on Airbnb listings from various cities in New York, including Brooklyn, Manhattan, Queens, Staten Island, and the Bronx. This dataset is widely used for data analysis and visualization to gain insights into the pricing, popularity, and performance of Airbnb listings in different areas. It contains a wealth of information on over 48,000 listings, including the property type, room type, location, price, availability, minimum nights, number of reviews and review scores. Moreover, the dataset also includes geographic data, such as longitude and latitude coordinates, which can be used to visualize the listings on a map. With data covering a period of eight years from 2011 to 2019, this dataset provides a vast time frame to analyze the changes in the Airbnb market in New York City.I chose the Airbnb dataset because it provides a rich and diverse set of variables that can be analyzed to uncover trends and patterns in the short-term rental market in New York City. As a data analyst, I am interested in understanding how various factors, such as location, amenities, and host characteristics, affect the pricing and popularity of Airbnb listings. The dataset's geographic data also enables me to map out the listings and analyze spatial patterns in the data. Moreover, the size and scope of the dataset make it an ideal resource for exploratory data analysis and statistical modeling. Overall, the Airbnb dataset offers a unique opportunity to gain insights into the dynamic and rapidly evolving short-term rental market in New York City.

Format:

Variables	Description
id	ID number of the listing
name	Name of the listing
host_id	ID number of the host
host_name	Name of the host
neighbourhood_group	Boroughs in which the listing is located
neighbourhood	Neighborhood in which the listing is located

Variables	Description
latitude	Latitude of the listing
longitude	Longitude of the listing
room_type	Type of room (e.g. Entire home/apt, Private room)
price	Price per night
minimum_nights	Minimum number of nights for a stay
number_of_reviews	Number of reviews for the listing
last_review	Date of the latest review
reviews_per_month	Number of reviews per month
calculated_host_listings_count	Number of listings by the same host
availability_365	Number of days in the year that the listing is available

Source:

https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data (https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data)

Task 1:

1.Parsing Data

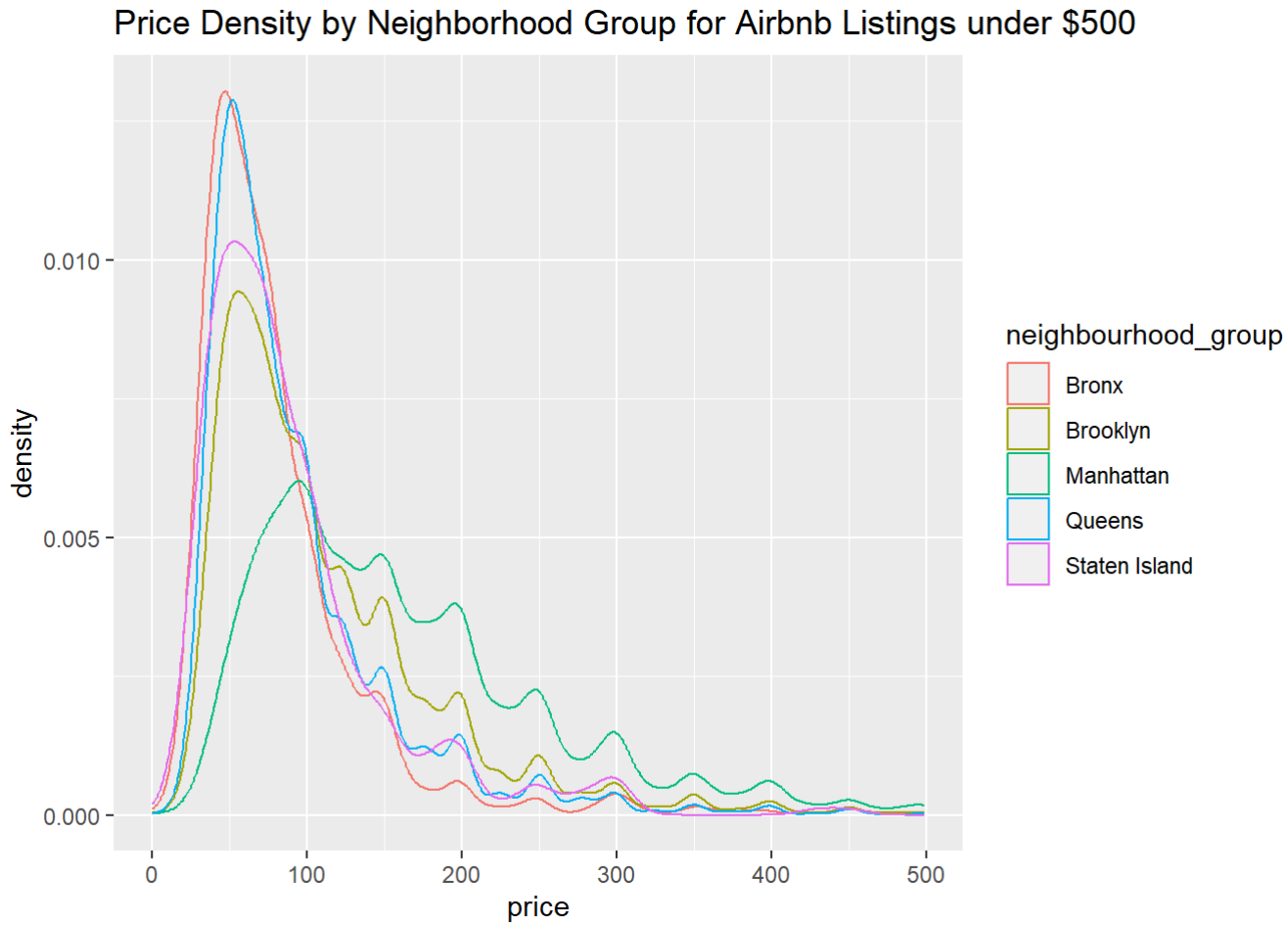
```
airbnb <- read_csv("C:/Users/joelj/OneDrive/Desktop/AB_NYC_2019.csv")
```

This line of code loads the data stored in the file “AB_NYC_2019.csv” into R as a dataframe and assigns it to the variable name “airbnb”.This line of code assumes that our data set is in .csv format. The filepath of the file is specified as “C:/Users/joelj/OneDrive/Desktop/” and may need to be adjusted depending on where the file is located on the user’s computer. Once loaded, the data can be manipulated and analyzed using various R functions and packages.

Data Wrangling:

```
airbnb$neighbourhood_group <- factor(airbnb$neighbourhood_group)
airbnb$room_type <- factor(airbnb$room_type)
airbnb$price <- as.numeric(airbnb$price)
airbnb$last_review <- as.Date(airbnb$last_review)
airbnb <- airbnb
```

Exploratory Data Analysis



This plot shows the density distribution of Airbnb listing prices for neighborhoods grouped by their neighborhood group, but only for listings with a price less than $500. The x-axis represents the price range, while the y-axis shows the density of listings falling within each price range. The different colors of the density curves represent the different neighborhood groups. This plot allows us to visually compare the price distribution of listings across different neighborhood groups for lower-priced listings.

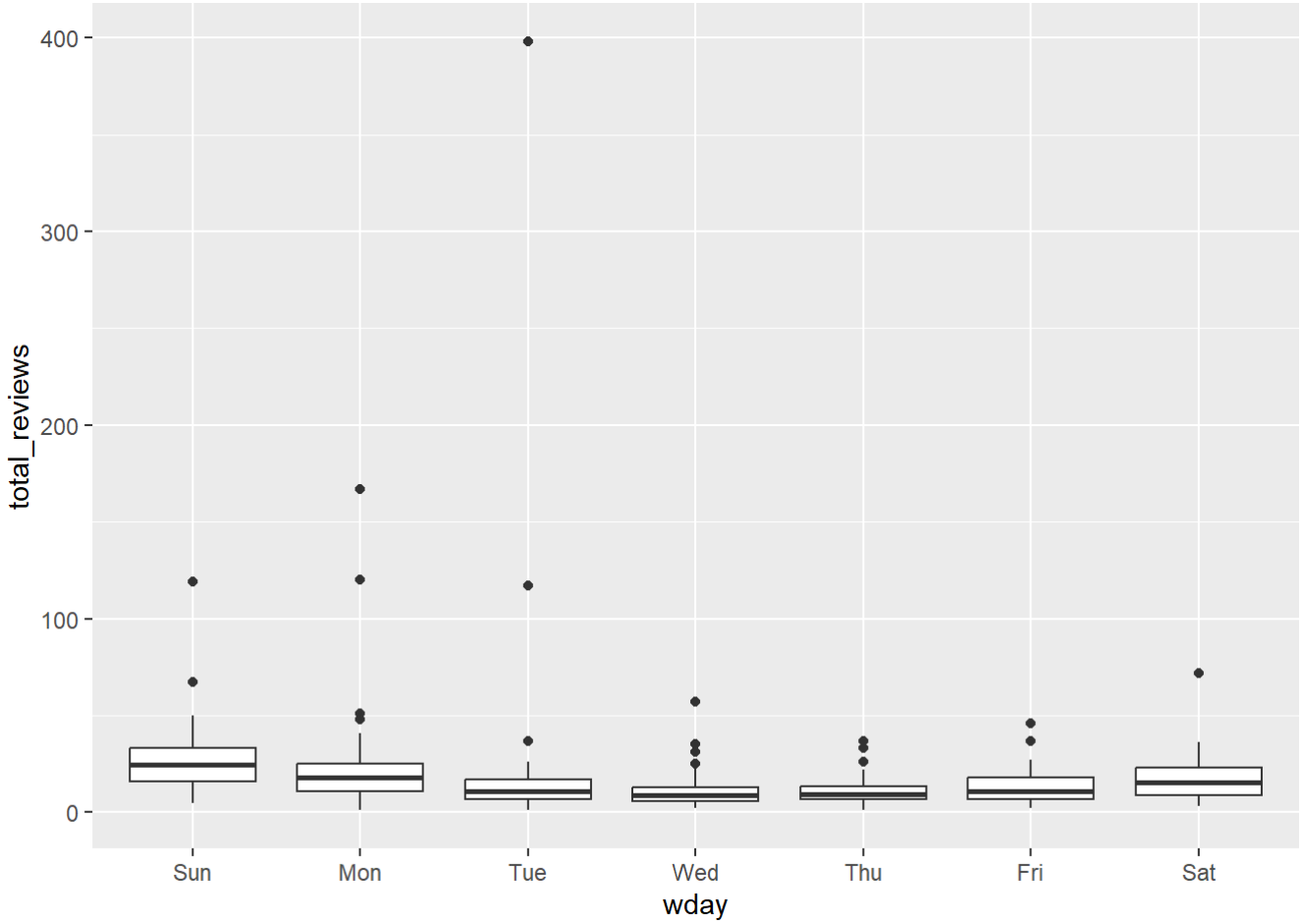
Question: What affects the total number of last reviews?

To answer this question,I had created a new data frame using my data set called airbnb_reviews. Airbnb_reviews contains the date,week day and year of the last_review variable. To make my analysis, more easier to interpret, I filtered the data frame to year 2018 only.

```
airbnb_reviews <- airbnb%>%
  filter(!is.na(last_review)) %>%
  mutate(date = ymd(last_review),
         wday = wday(last_review),
         year = year(last_review))%>%
  group_by(date,wday,year) %>%
  summarize(total_reviews = n()) %>%
  filter(date >= ymd("2018-1-1") & date <= ymd("2019-1-1"))
```

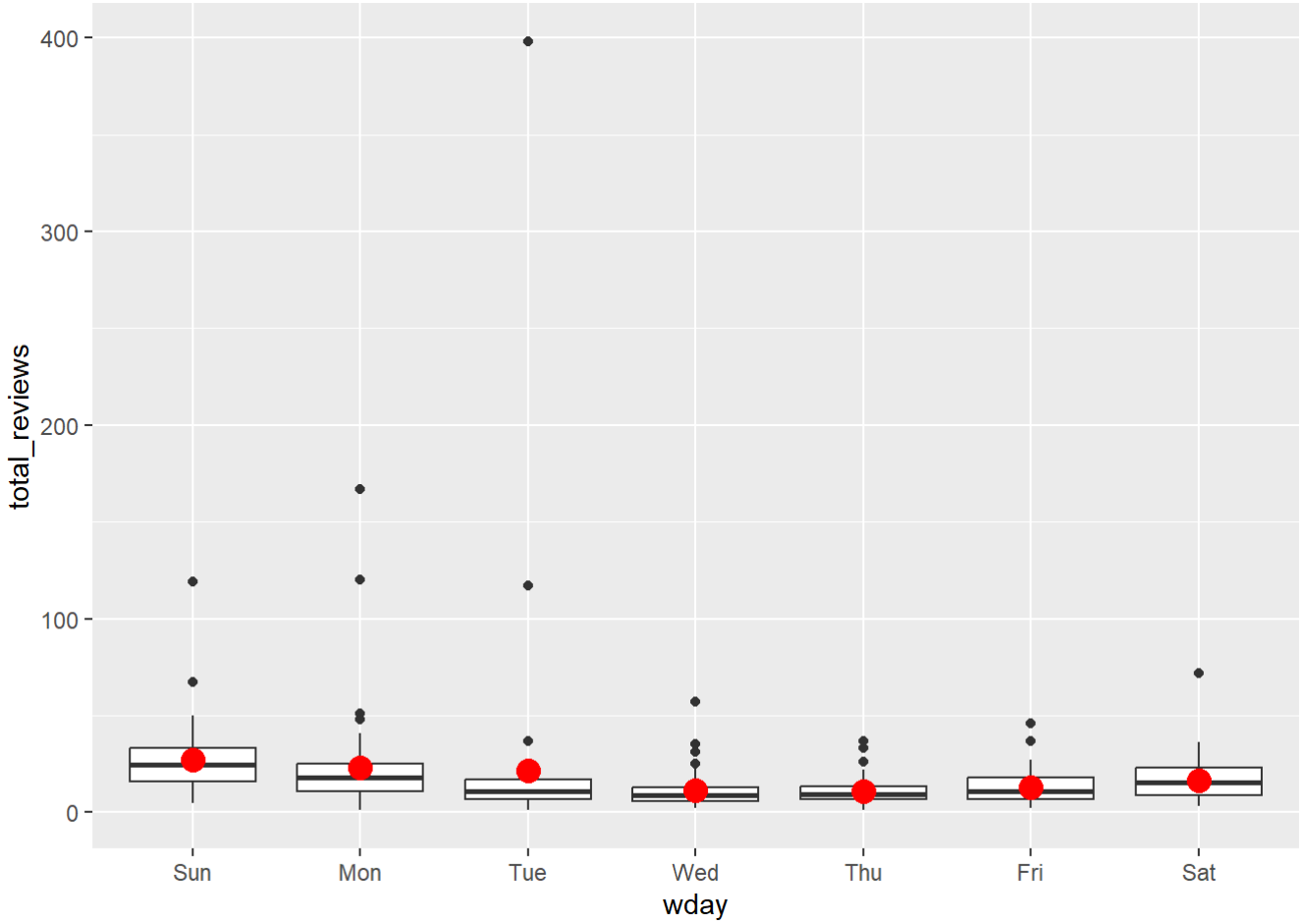
To understand our question we need to visualize the the variables we are going to use to get a better understanding.

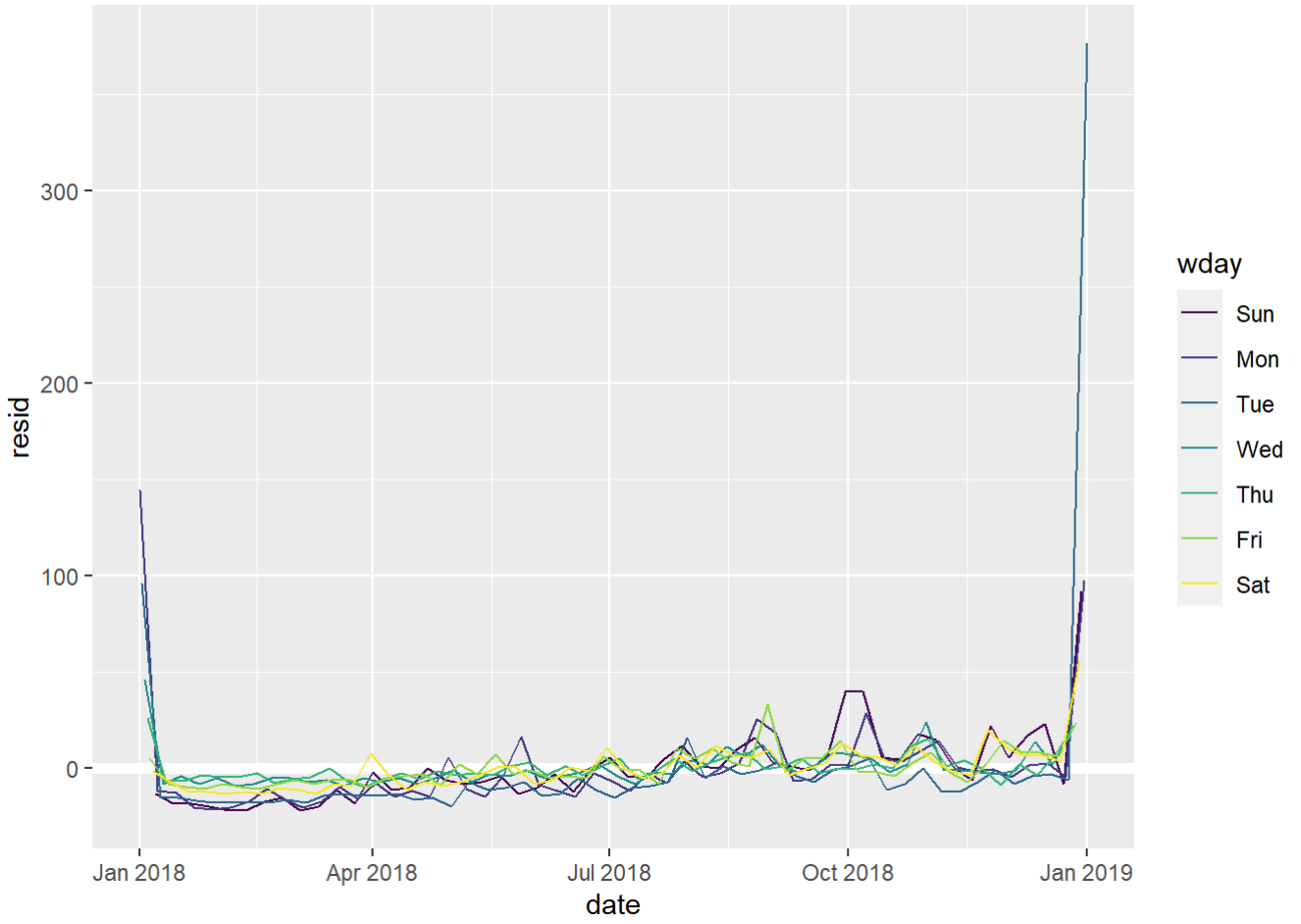
To do this we create a box plot with the total number of reviews against each day of the week. I had assumed that there would surely be a higher number of reviews in days such as Sunday and Saturday as these are holidays.



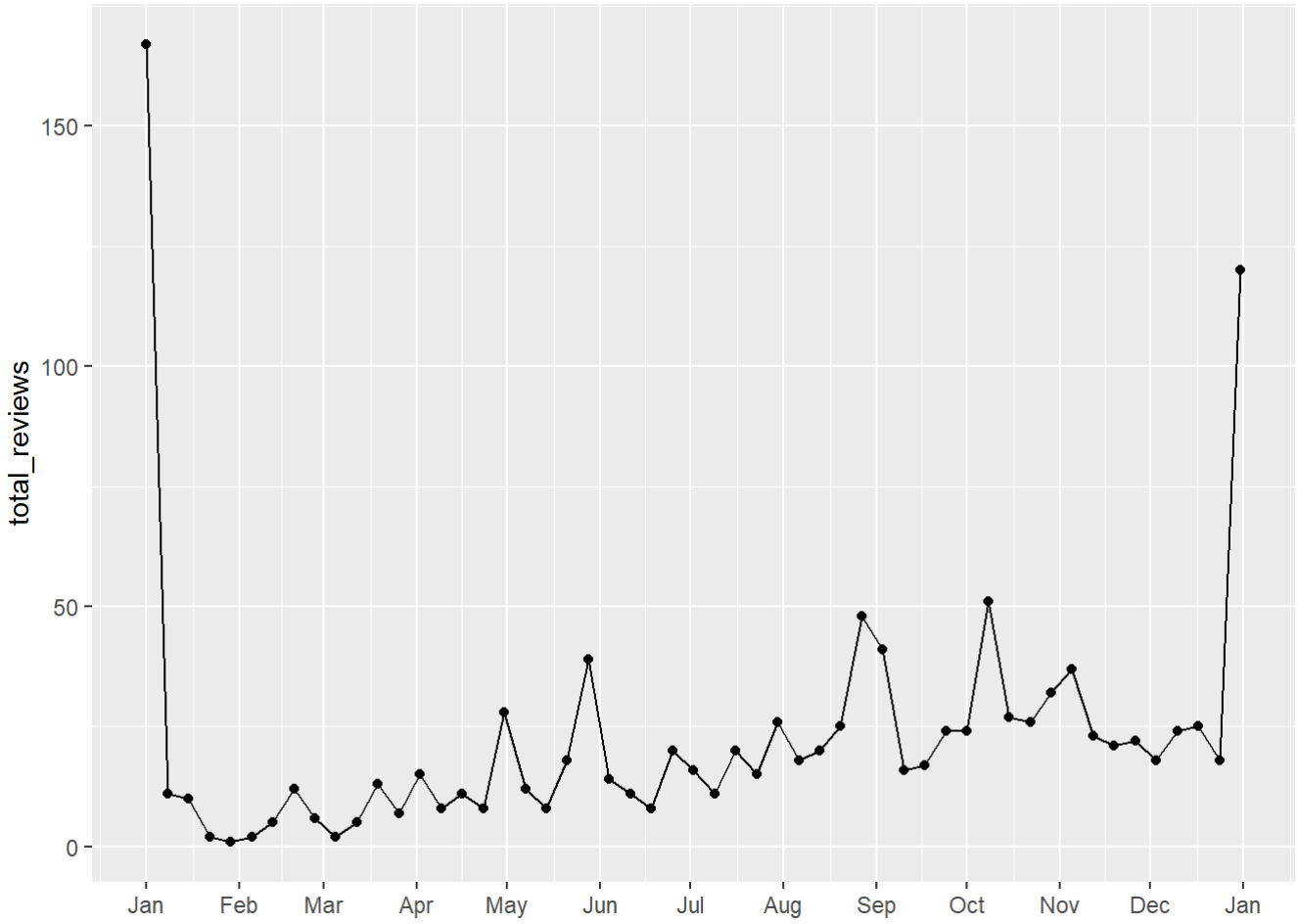
To my surprise, my assumption was inaccurate. It was found that Sunday and Monday have the highest number of reviews.Therefore it was evident that I should try to find the reasons why Monday is so high even though it is not a public holiday.To understand this, I started by using a linear regression model to further evaluate my course of action. And then plot the predictions on a boxplot and visualize the residuals against week days.

```
mod <- lm(total_reviews ~ wday, data = airbnb_reviews)
```


400
300
200
100
0
total_reviews
Sun
Mon
Tue
Wed
Thu
Fri
Sat
wday


300
200
100
0
resid
Jan 2018
Apr 2018
Jul 2018
Oct 2018
Jan 2019
date
wday
Sun
Mon
Tue
Wed
Thu
Fri
Sat

Even in the residuals plot we still do not see a great variation across any of the days but it still does not answer our question regarding Monday. To further understand this effect,I plotted the the linear model only for Monday.

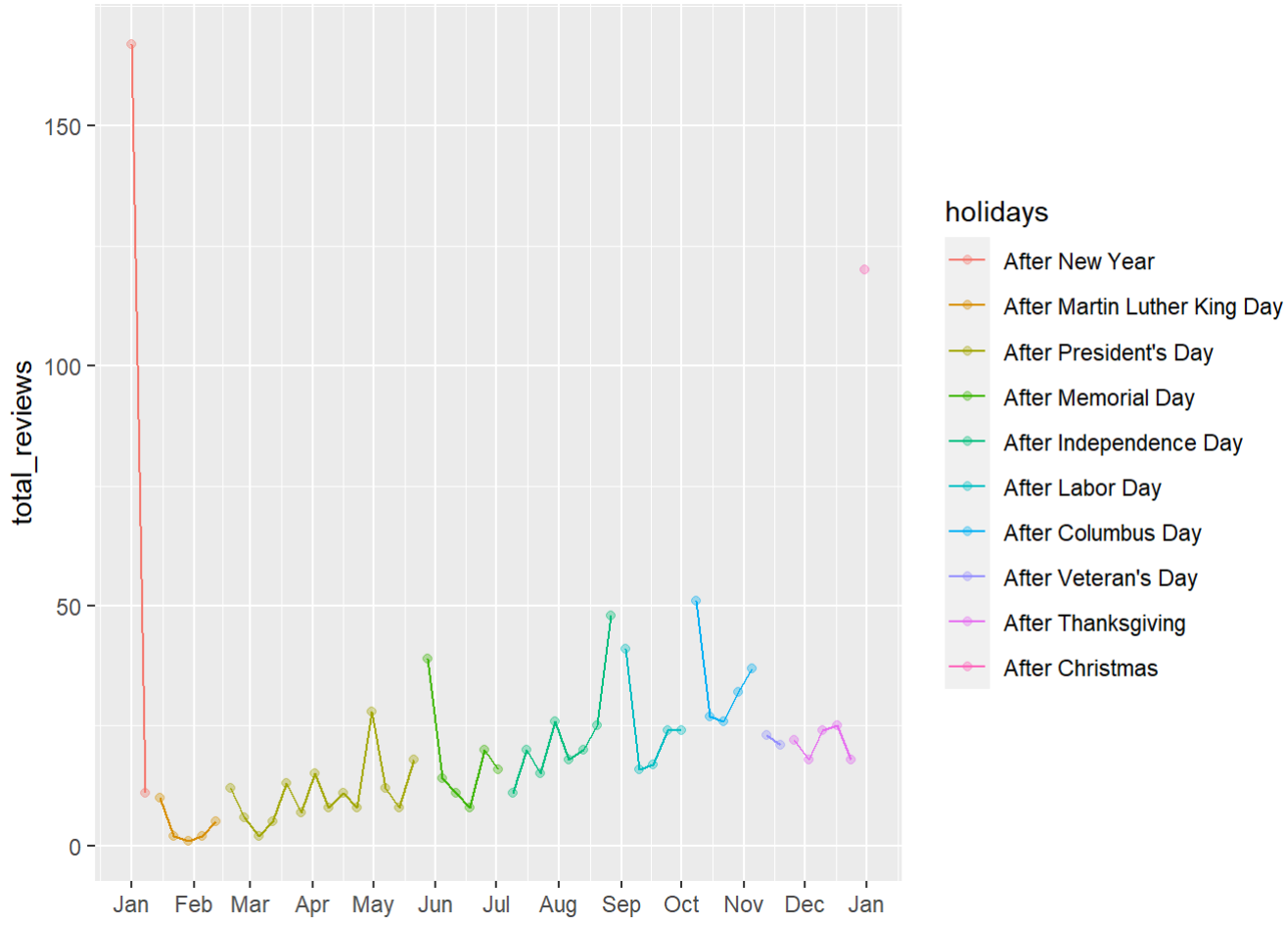


From the plot we were able to understand that there is a high level of reviews in January and towards December. Why would it be so ?. With regard to January and December months, two main holidays are observed in US which are New Year and Christmas respectively. Therefore I was able to come to an assumption that the reason why Monday would have a high number of reviews could be because of the effect of Holidays in the United States. To further understand this effect I researched about the holidays in United States.And using these holidays I set up breaks in the model.

Federal Holidays 2018

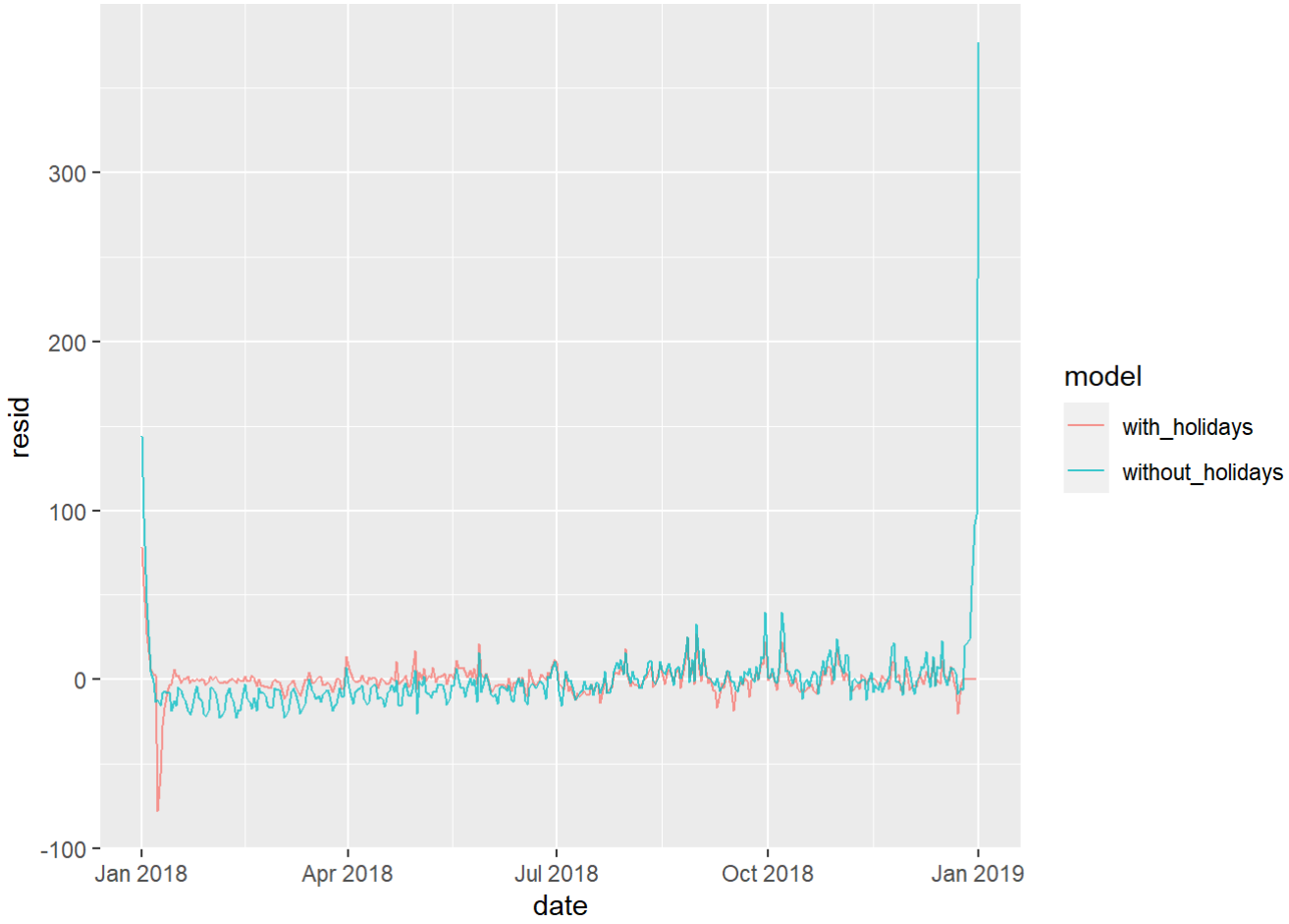
Date	Federal holiday	Day of the week
January 1, 2018	New Year's Day	Monday
January 15, 2018	Martin Luther King Day	Monday
February 19, 2018	Presidents' Day	Monday
May 28, 2018	Memorial Day	Monday
July 4, 2018	Independence Day	Wednesday
September 3, 2018	Labor Day	Monday
October 8, 2018	Columbus Day	Monday
November 11, 2018	Veterans Day	Sunday
November 12, 2018	Veterans Day (observed)	Monday
November 22, 2018	Thanksgiving Day	Thursday
December 25, 2018	Christmas Day	Tuesday

Figure 1: Holidays in 2018 (1)



We see from the line graph that certain of the breaks are a good effect on the plot but not all of the breaks are a good decision maker. To understand what would happen if holiday effect is considered , I made two linear regression models, one for the data with holidays and another without it.

```
mod3 <- lm(total_reviews ~ wday, data = airbnb_reviews)
mod4 <- lm(total_reviews ~ wday * holidays, data = airbnb_reviews)
```



From the plot with the residuals it is evident that there certainly as been improvements in the model with holidays but in the case of January and December there still has not been any proper changes in the residuals. So definitely we are not able to say that holidays have a strong effect. Therefore,we surely need a different approach to this.

The following federal holidays are observed by the majority of private businesses with paid time off:

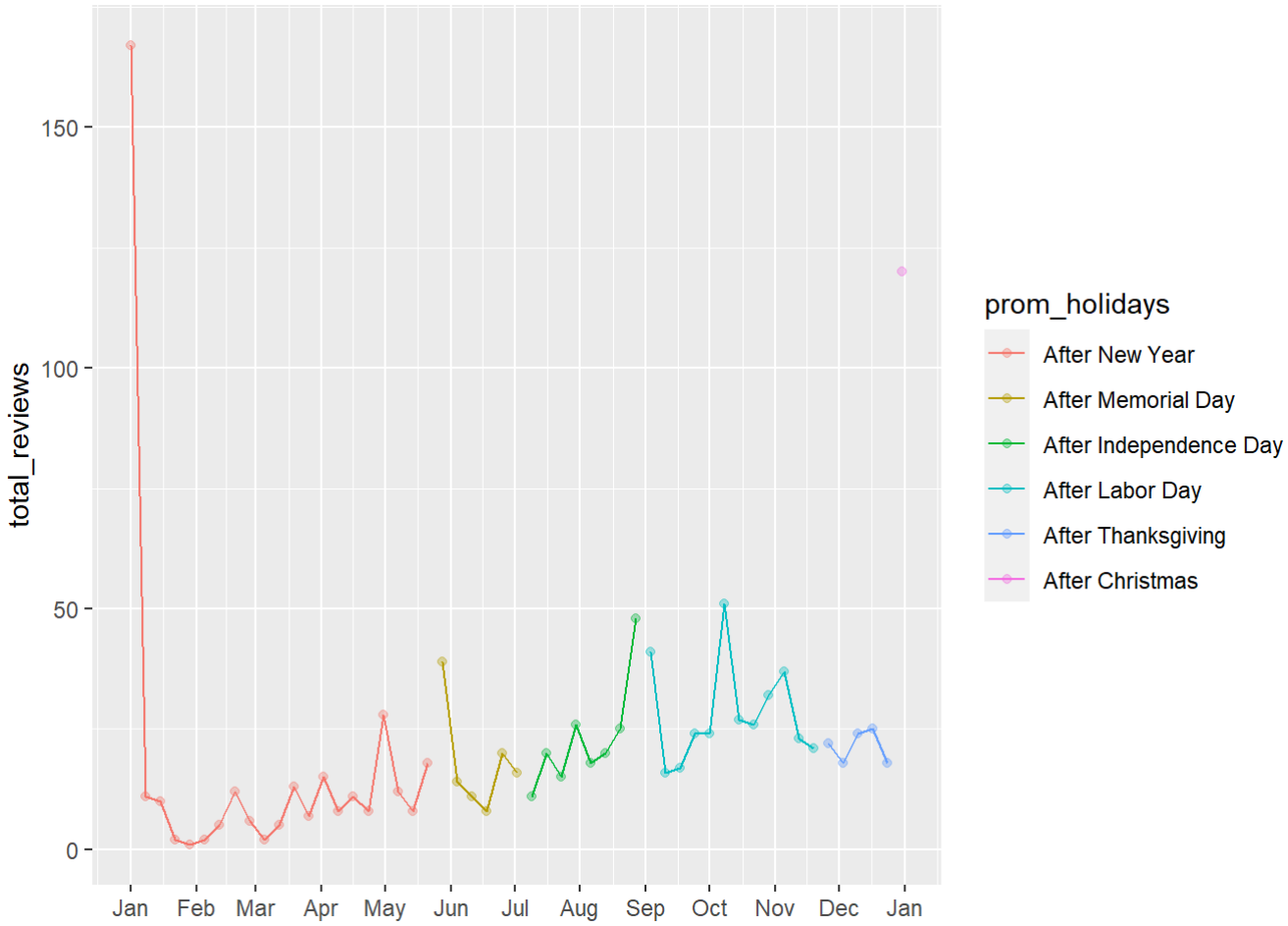
- New Year's Day (January 1)[10]
- Memorial Day (May 25–31, floating Monday)
- Independence Day (July 4)
- Labor Day (September 1–7, floating Monday)
- Thanksgiving (November 22–28, floating Thursday)
- Christmas (December 25)

Other federal holidays are less widely observed by businesses. These include:

- Martin Luther King Jr. Day (January 15–21, floating Monday)
- Washington's Birthday (February 15–21, floating Monday)
- Juneteenth National Independence Day (June 19)
- Columbus Day (October 8–14, floating Monday)
- Veterans Day (November 11)

Figure 2: Holidays in 2018 (2)

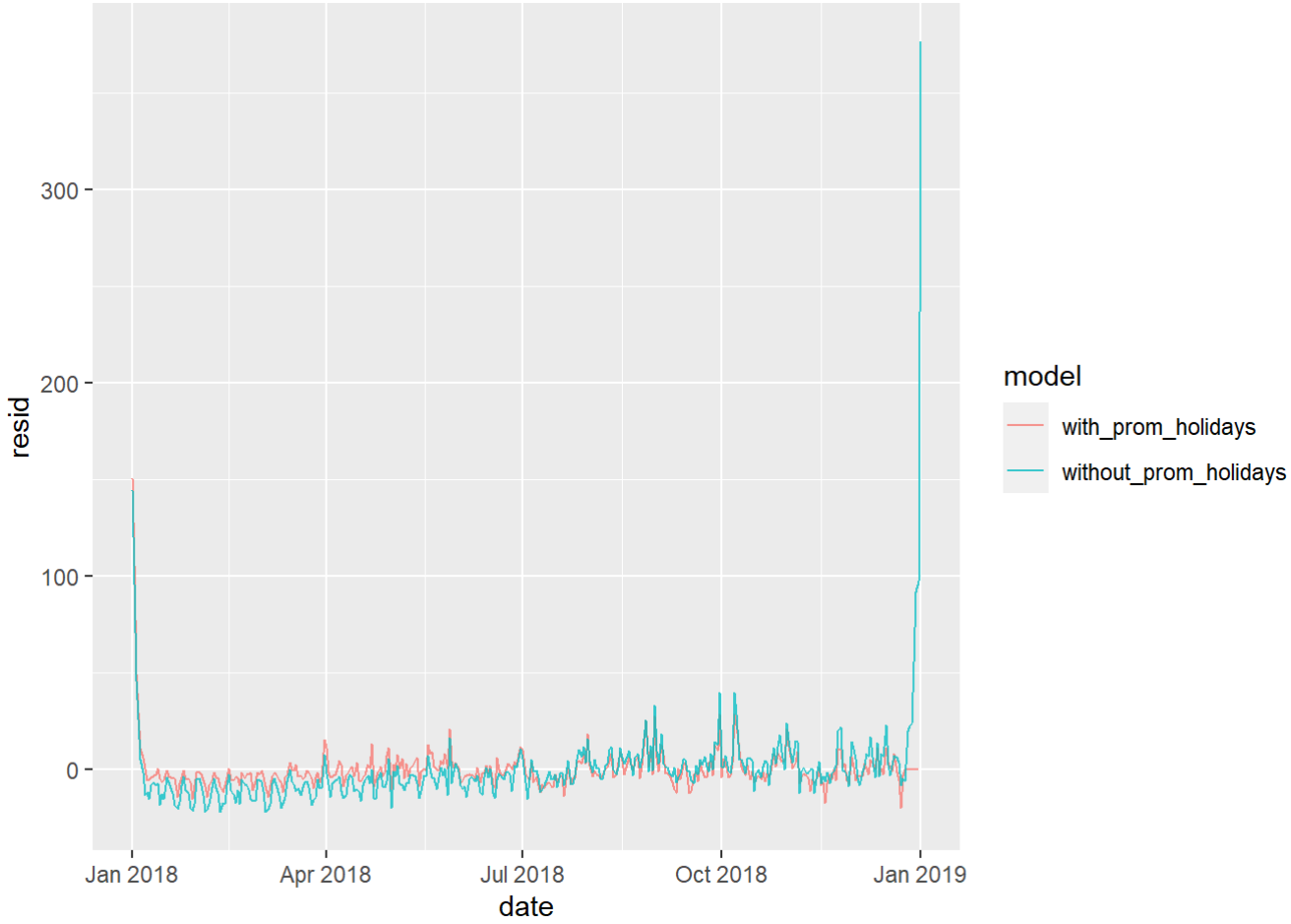
To further my analysis, I tried looking more about the holidays in United States. I was able to find that there are certain holidays which are not considered by all businesses. Therefore, I should only consider New Year, Memorial Day,Independence Day,Labor Day,Thanksgiving and Christmas for the breaks.



Here again we can see that these holiday breaks are essential and possibly give us a good idea about the effect of holidays. Also because the number of breaks are lesser in amount compared to the last, we could say that it would create a better model compared to the previous one.

Now we start with the modeling part for these prominent holidays. We create two models one for the original week day parameter and another which includes the prominent holiday parameter. And with these models we create a residual plot comparing the effect of both models.

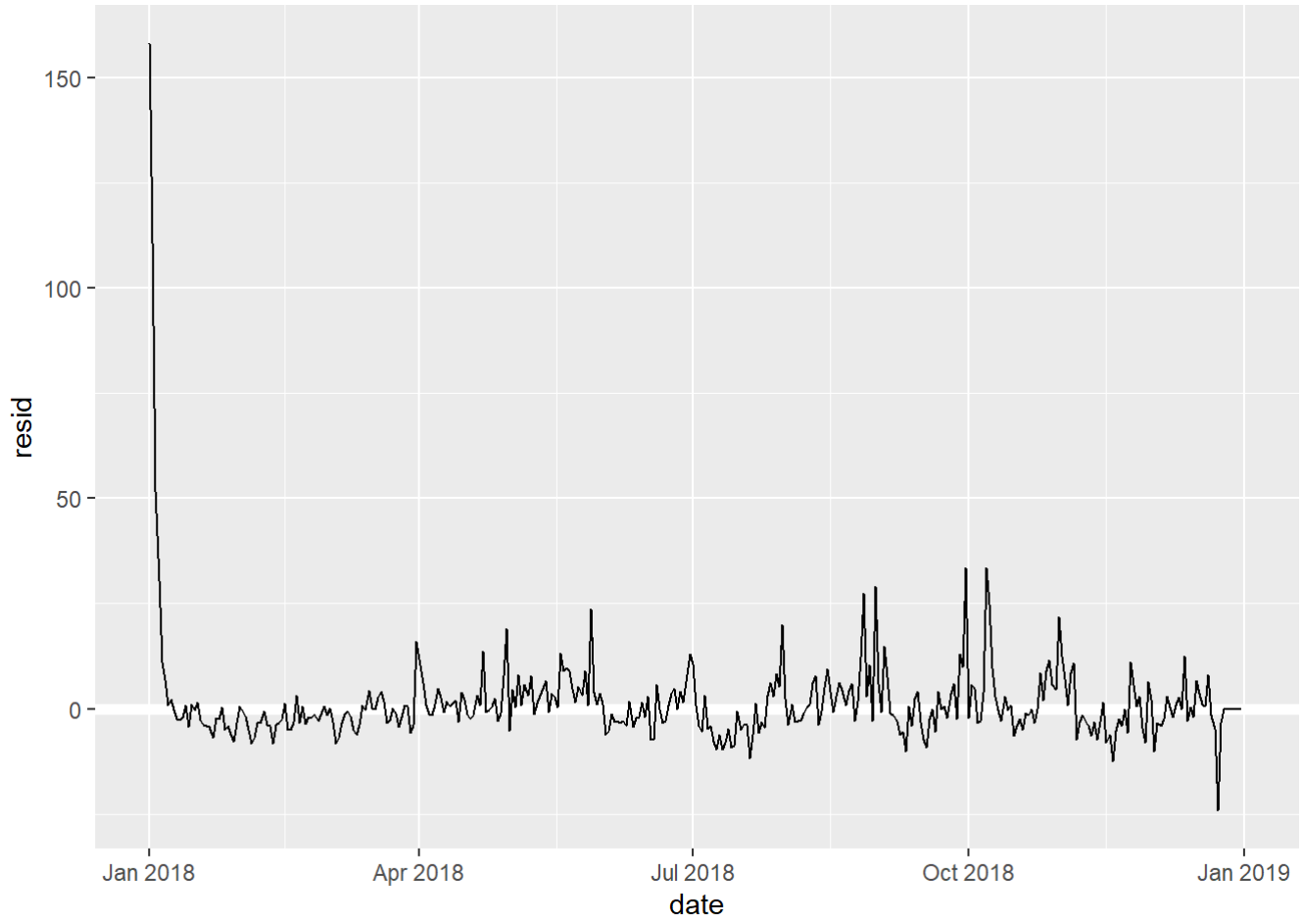
```
model1 <- lm(total_reviews ~ wday, data = airbnb_reviews)
model2 <- lm(total_reviews ~ wday * prom_holidays, data = airbnb_reviews)
```



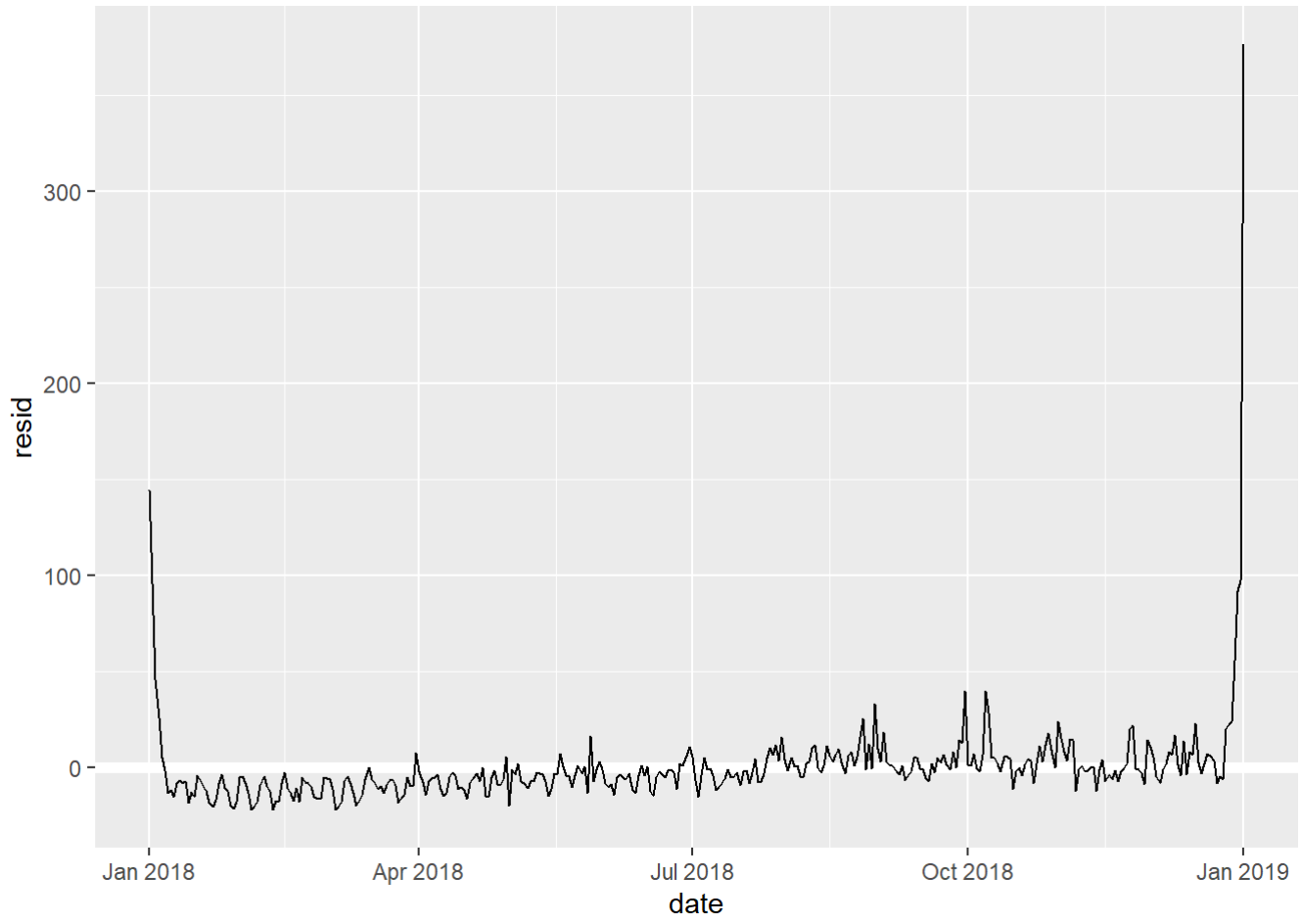
Here we are able to understand that the effect of the residuals are better compared to the model with all holidays. But yet again we see that January is not skewed towards the zero base line and we can not still say that it is the best fit.

I have done two linear models to understand this effect, but both the models are not the best method to go through for future analysis.Therefore I decided to try one more model with the help of robust linear modelling and compared it with the original linear model.

```
mod_2 <- rlm(total_reviews ~ wday * prom_holidays, data = airbnb_reviews)
```




```
modu2 <- lm(total_reviews ~ wday, data = airbnb_reviews)
```



Here we see that the residuals in the robust linear modelling is far better that linear model one and this approach of taking the prominent holidays surely has helped to remove the effect in the month of December but still the effect of January is still very noticeable. Therefore we can conclude that our model has not been able to identify the effect of holidays.

Discussion

In this analysis, I tried to take the effect of the days of the week with the holidays and this has not been ideal. I had tried two different types of model and also tried different variations with the holidays,but the effect did not change a lot.Therefore a different approach either by finding the effect of a different factor or taking a different scale other than the days of the week or maybe even a different type of model is required.

One limitation of my study is that I only looked at data from the year 2018. It is possible that the patterns I observed may differ in other years or in different locations. Future studies could explore this further by analyzing data from different years or regions. Additionally, it would be interesting to explore the impact of other factors, such as month of the year effect, on the total number of last reviews which are in the Airbnb dataset. Even taking into consideration of other factors which are outside of this dataset can be ideal.

Overall, our study was not able provide valuable insights to answer the question. But surely it did set the tone for future data analysts to do further research on it.

Refereneces:

1. https://www.calendarpedia.com/holidays/federal-holidays-2018.html (https://www.calendarpedia.com/holidays/federal-holidays-2018.html)
2. https://en.wikipedia.org/wiki/Public_holidays_in_the_United_States (https://en.wikipedia.org/wiki/Public_holidays_in_the_United_States)

`