The impact of Covid19 on the hospitality industry

The Covid19 pandemic has wreaked havoc right across the global economy, with international lockdowns and economic downturns everywhere, one of the most affected industries during this time was hospitality. The hospitality industry as a whole present an effective indicator of the global economy, in a good economic situation greater numbers of people will chose to vacation abroad, consequently boosting the hospitality industry of major tourist hubs around the world. Conversely, during an economic downturn, uncertainty often leads to increased frugality resulting in a reduction in tourism which affects the hospitality industry as a whole.

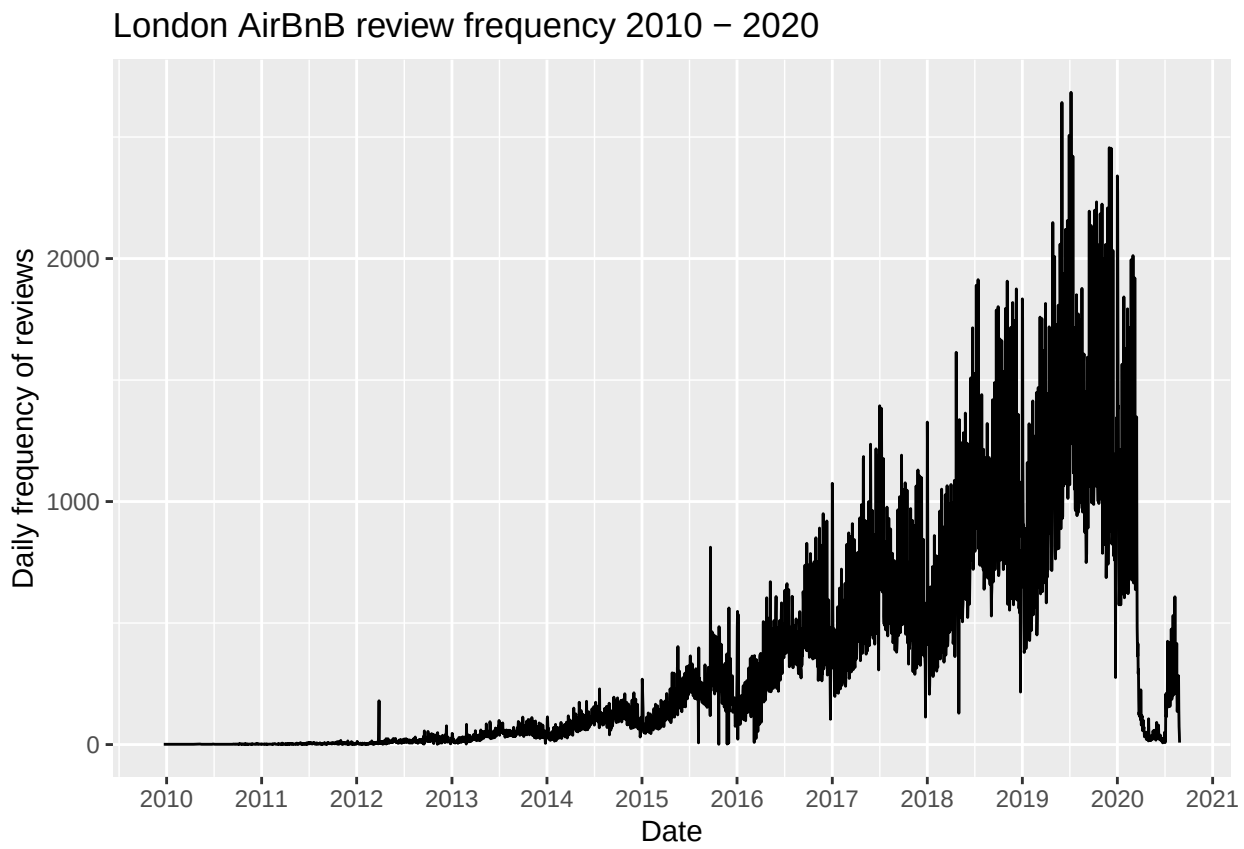
In analysing the global hospitality industry, this project looks at data sourced from AirBnB, a prominent booking application used for the acquisition of lodgings for tourists world wide. Furthermore, the application focuses on smaller businesses renting apartments and rooms as opposed to major hotels, this is beneficial as it provides a closer look at local economies around the world. The data consists of reviews left by customers corresponding to a certain listing within various cities. These can be used as an indicator of the app's usage, however it should be noted that the frequency numbers presented below do not indicate the actual number of booking made but rather only the number of reviews left by the customers.

First, we must import the corresponding libraries used throughout this project. We mainly need libraries with data handling, visualization and model forecasting abilities, these are shown below. Furthermore, to facilitate the effective pipelining of data processing and visualization tasks, several function have been written in a separate file names functions.R, these will be used throughout this project.

To begin, we first load the review data for London.

If we ignore 2020, the data displays numerous seasonal trends, with a general trend which sees a steady increase from around 2013 onwards, indicating the rapid rise in popularity of the application, consequently leading to more users.

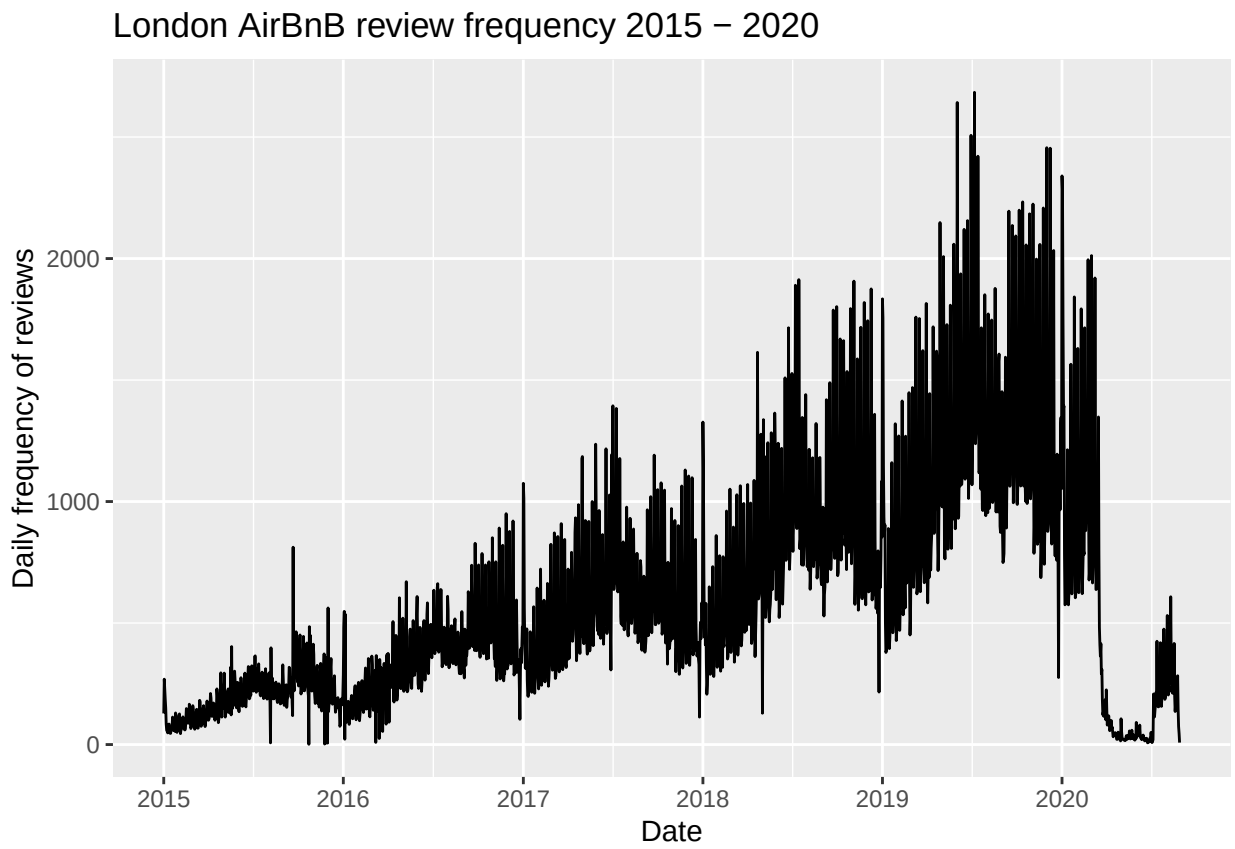
```
# Lets take a closer look by producing a line plot of the data
plot <- ggplot(data_london, aes(x=date, y=Freq)) +
        geom_line()

labs = c("2010", "2011", "2012", "2013", "2014", "2015",
         "2016", "2017", "2018", "2019", "2020", "2021")

breaks = as.Date("2010-01-01") +
         cumsum(c(0, 365, 365, 366, 365, 365,
                  365, 366, 365, 365, 365, 366))

plot +
  ggtitle("London AirBnB review frequency 2010 - 2020") +
  xlab("Date") +
  ylab("Daily frequency of reviews") +
  scale_x_continuous(breaks=breaks,
                     labels=labs)
```



The plot above appears to posses several interesting traits which will now be investigate more closely.

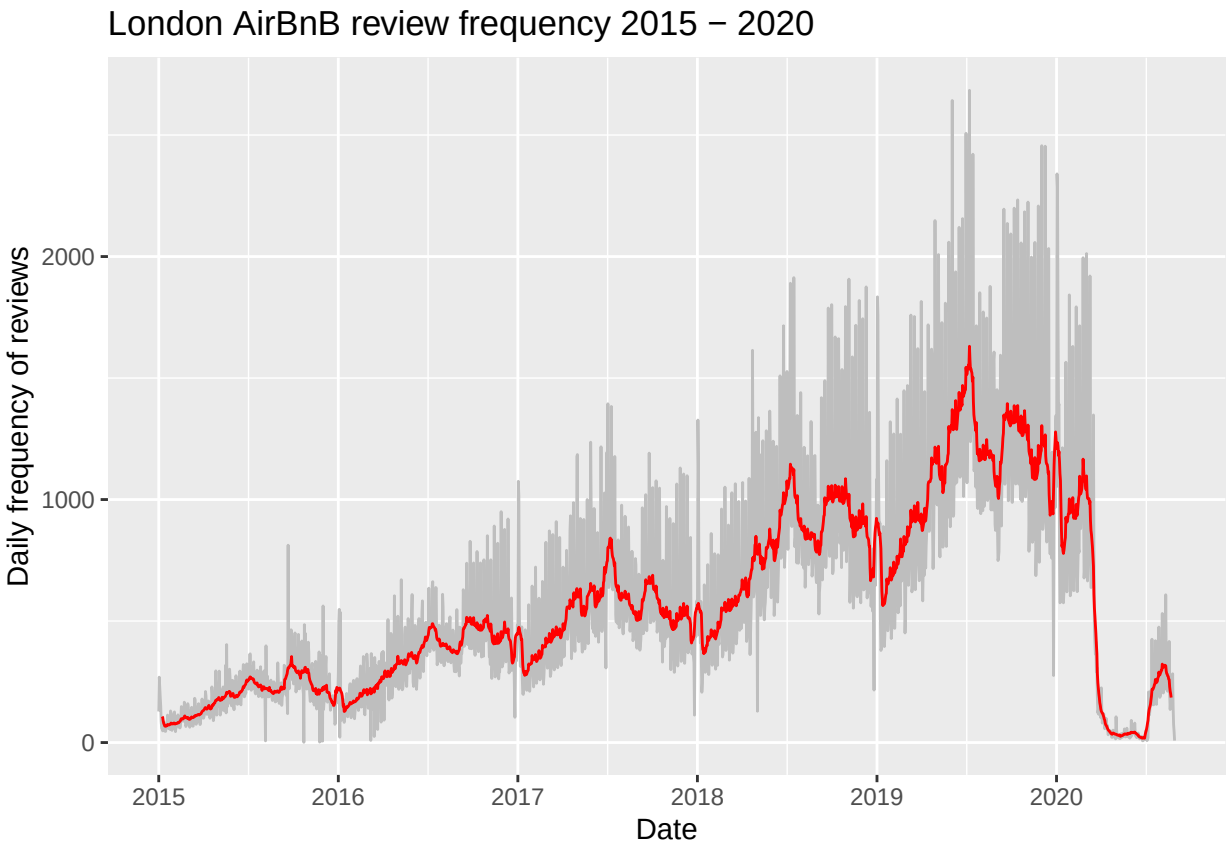
Firstly, we see that years prior to 2015 saw little change, with years after 2015 indicating a rapid rise in the app's popularity. Let us take a closer look at the data post-2015.

```
data_london_post2015 <- date_select("2015-01-01",
                                    "2020-10-01",
                                    data_london)
breaks = as.Date("2015-01-01") + cumsum(c(0, 365, 366, 365, 365, 365))
labs = c("2015", "2016", "2017", "2018", "2019", "2020")

plot <- ggplot(data_london_post2015, aes(x=date, y=Freq)) +
        geom_line() +
        scale_x_continuous(breaks=breaks,
                           labels=labs)

plot +
  ggtitle("London AirBnB review frequency 2015 - 2020") +
  xlab("Date") +
  ylab("Daily frequency of reviews")
```



Looking more closely at the years 2015-2020, we can see some finer seasonal trends. However, these prove quite difficult to identify due to the increase in variance at later years. To visualize these better, let us take moving average of the review frequency in an attempt to extract the general trend of the above-seen data.

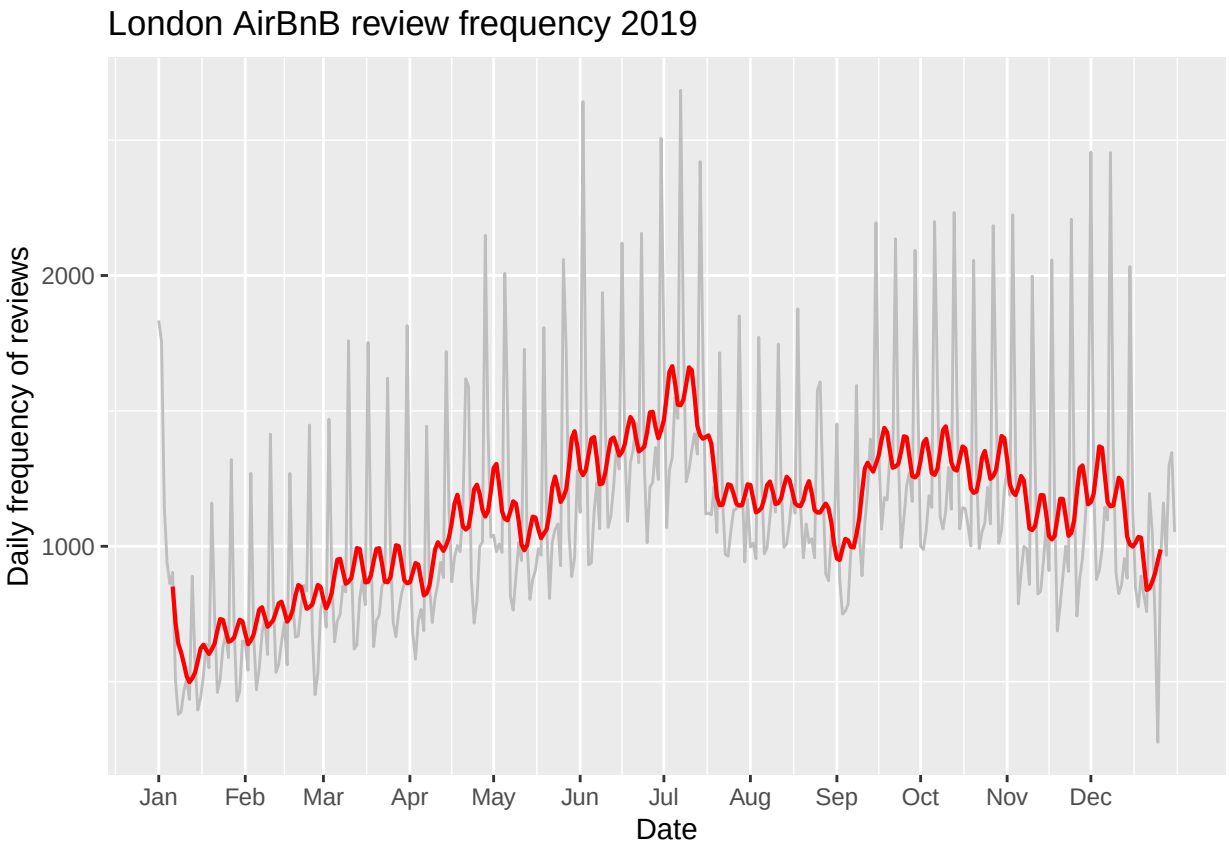
```
plot <- ggplot(data_london_post2015) +
  geom_line(aes(x=date, y=Freq), col="gray") +
  geom_line(aes(x=date, y=ma(Freq, 15)), col="red", size=0.5) +
  scale_x_continuous(breaks=as.Date("2015-01-01") + cumsum(c(0,365, 365, 365, 365, 365)),
                     labels=c("2015", "2016", "2017", "2018", "2019", "2020"))


plot + ggtitle("London AirBnB review frequency 2015 - 2020") +
  xlab("Date") +
  ylab("Daily frequency of reviews")
```



The moving average allows us to more clearly see the general trend within the data, removing much of its variance. There appears to be an annual trend which sees a peak around the middle of the year corresponding to summer period. This is easily explained as the summer vacation period which sees a large number of events and festivals hosted in London, naturally attracting many visitors.

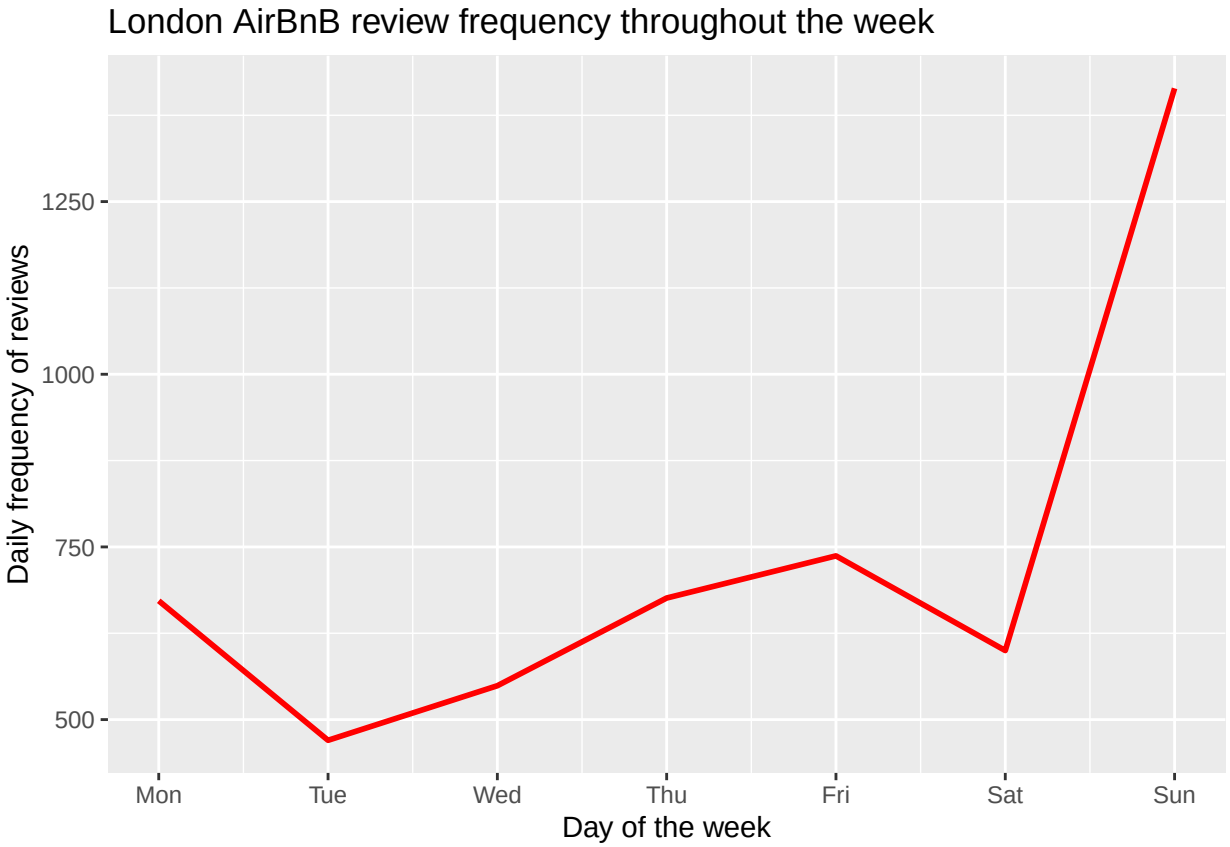
The moving average also indicates a shorter seasonal trend throughout each year, this would be best investigated by looking more closely at the plot above, hence, we now look at the data throught the year 2019.

```
london_20192020 <- date_select("2019-01-01", "2019-12-31", data_london_post2015)

breaks = as.Date("2019-01-01") + cumsum(c(0,31,28,31,30,31,
                                          30,31,31,30,31,30))

plot <- ggplot(london_20192020) +
  geom_line(aes(x=date, y=Freq), col="gray") +
  geom_line(aes(x=date, y=ma(Freq, 10)), col="red", size=0.75) +
  scale_x_continuous(breaks=breaks, labels=month.abb)

plot + ggtitle("London AirBnB review frequency 2019") +
  xlab("Date") +
  ylab("Daily frequency of reviews")
```



Looking at the plot above, we can more clearly see this seasonal trend. There appears to be around 50 of such smaller minima throughout the year, suggesting that this could be a weekly trend. To investigate this further, let us look at the frequency plot over a span of 7 day, corresponding to one week.

```
london_weekly <- date_select("2019-02-04", "2019-02-10", data_london_post2015)
london_weekly$day <- weekdays(london_weekly$date)

days = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")

plot <- ggplot(london_weekly) +
  geom_line(aes(date, Freq), col="red", size=1) +
  scale_x_continuous(breaks=as.Date("2019-02-04") + cumsum(c(0,1,1,1,1,1,1)), labels=days)

plot + ggtitle("London AirBnB review frequency throughout the week") +
  xlab("Day of the week") +
  ylab("Daily frequency of reviews")
```



We can now clearly see the origin of this weekly trend. As expected, the number of reviews peaks during the weekend period with the peak on sunday. This can simply be explained as many booking will involve a weekend-long stay, concluding on the sunday, hence, the most reviews occur on this day.

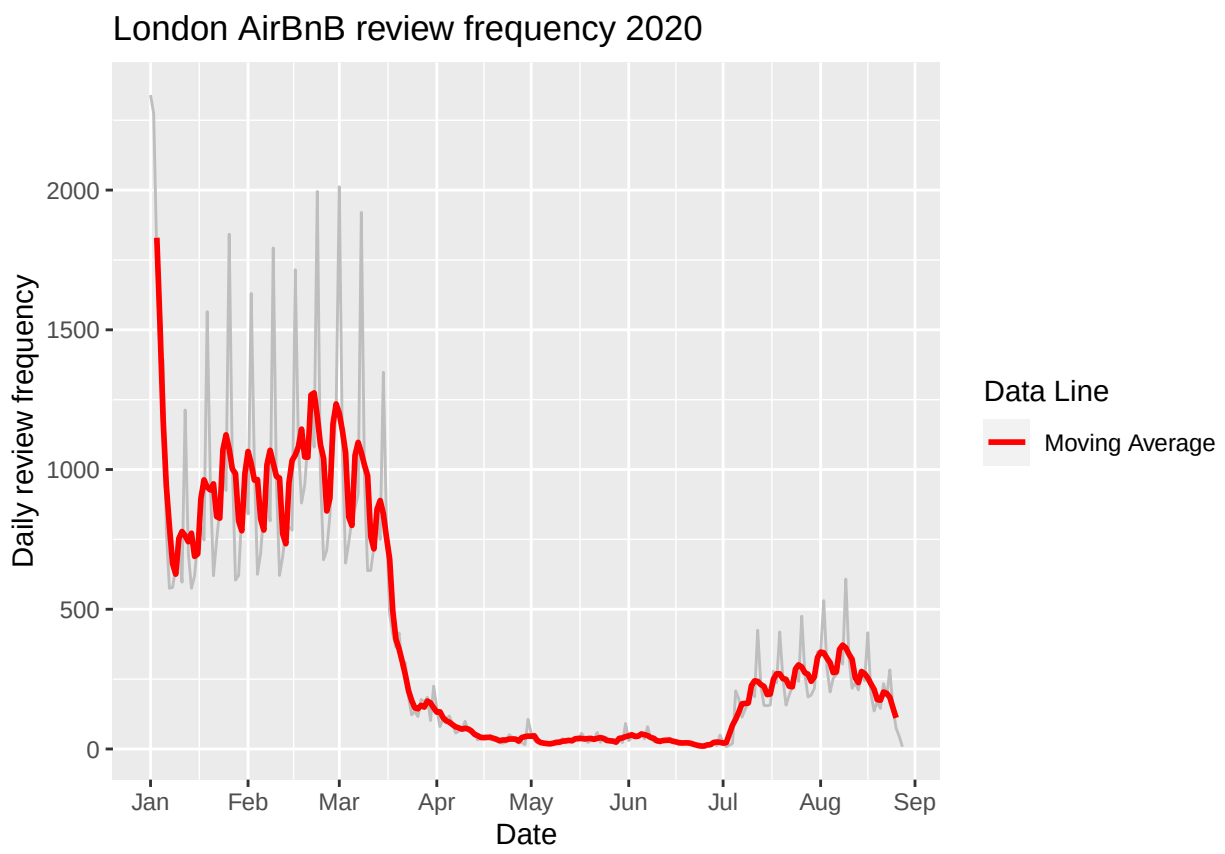
Having looked more closely at the available data we now better understand the seasonal trend found within the data. These are as follows:

- A general trend of increase in the number of reviews throughout the years, indicative of the rising popularity of the application.
- An annual trend with a peak around the middle of the summer period (June - July), caused by various events and festivals coinciding during the summer period.
- Finaly, a weekly trend which sees the most reviews posted on the sunday, likely caused by weekend-long booking which conclude on that day.

With these general and seasonal trends understood, we can now better understand the impact of the pandemic. Looking more closely at data from 2020 onwards, we can see this impact quite clearly.

```
data_london_2020 <- date_select("2020-01-01", "2021-01-01", data_london)

breaks = as.Date("2020-01-01") + cumsum(c(0,31,29,31,30,31,30,31, 30))

ggplot(data_london_2020) +
  geom_line(aes(x=date, y=Freq), col="gray") +
  geom_line(aes(x=date, y=ma(Freq, 5), col="red"), size=1) +
  scale_x_continuous(breaks=breaks,
                     labels=month.abb[1:9]) +
  scale_color_manual(labels = c("Moving Average"),
```

```
                    values = c("red")) +
  labs(title="London AirBnB review frequency 2020",
       x="Date",
       y="Daily review frequency",
       color="Data Line")
```



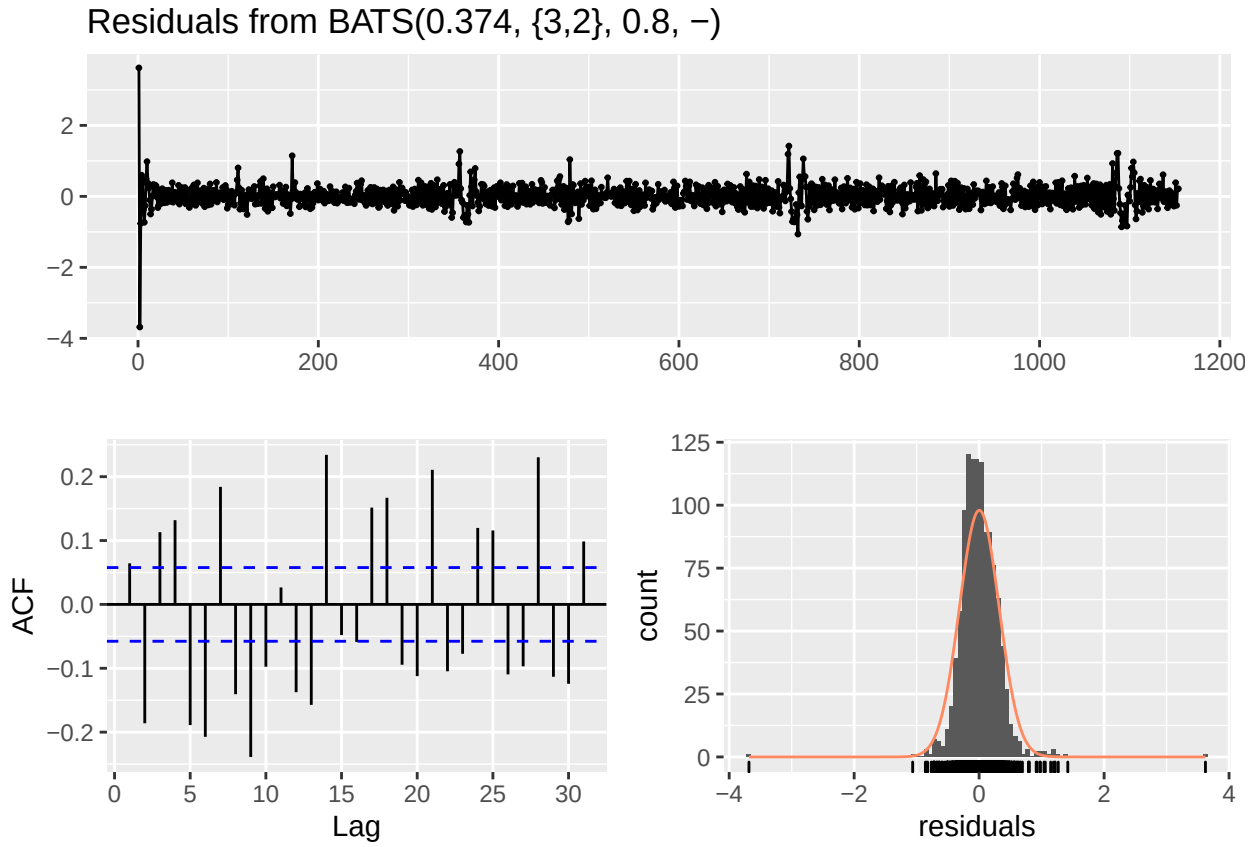
The drop in review numbers is quite stark and dramatic, the initial fall-off point appears to coincide with the first wave of lockdown measures introduce within the UK, with most international and local measures issued around mid-march. During this time, workers across the country, and much of the world, were encouraged to work from home where possible, with additional travel restrictions resulting in record-low numbers of tourists.

This decrease appears to continue throughout much of the lockdown period, with restrictions beginning to be lifted in late june there appears to have been a small bounce-back leading into august however this also appears to fall off. The reason for this second decrease could be the arrival of the second wave within the UK, however this is unlikely as this did not manifest until early september, more likely this is indicative of an incomplete dataset. The dataset is updated every few months by scraping the AirBnB database, this shortfall could be simply caused by some of the more recent reviews not being processed into the database.

Therefore, the dataset is considered reliable up until august, with all subsequent operations considerate of this. To visualise the magnitude of the drop following the pandemic lockdown, we can fit a TBATS model to the data leading up to the lockdown around mid-march. TBATS is a forecasting methods used in the modelling of time-series data, it proves particularily adept at modelling complex seasonal patterns such as those mentioned above.

```
covid_data <- ma(date_select("2017-01-01", "2020-05-10", data_london)$Freq, 10)
train_data <- ma(date_select("2017-01-01", "2020-03-10", data_london)$Freq, 10)
```

```
train_data <- train_data[6:(length(train_data)-6)]

fit.tbats <- tbats(ts(train_data))
f.tbats <- forecast(fit.tbats, h=30)

checkresiduals(f.tbats)
```

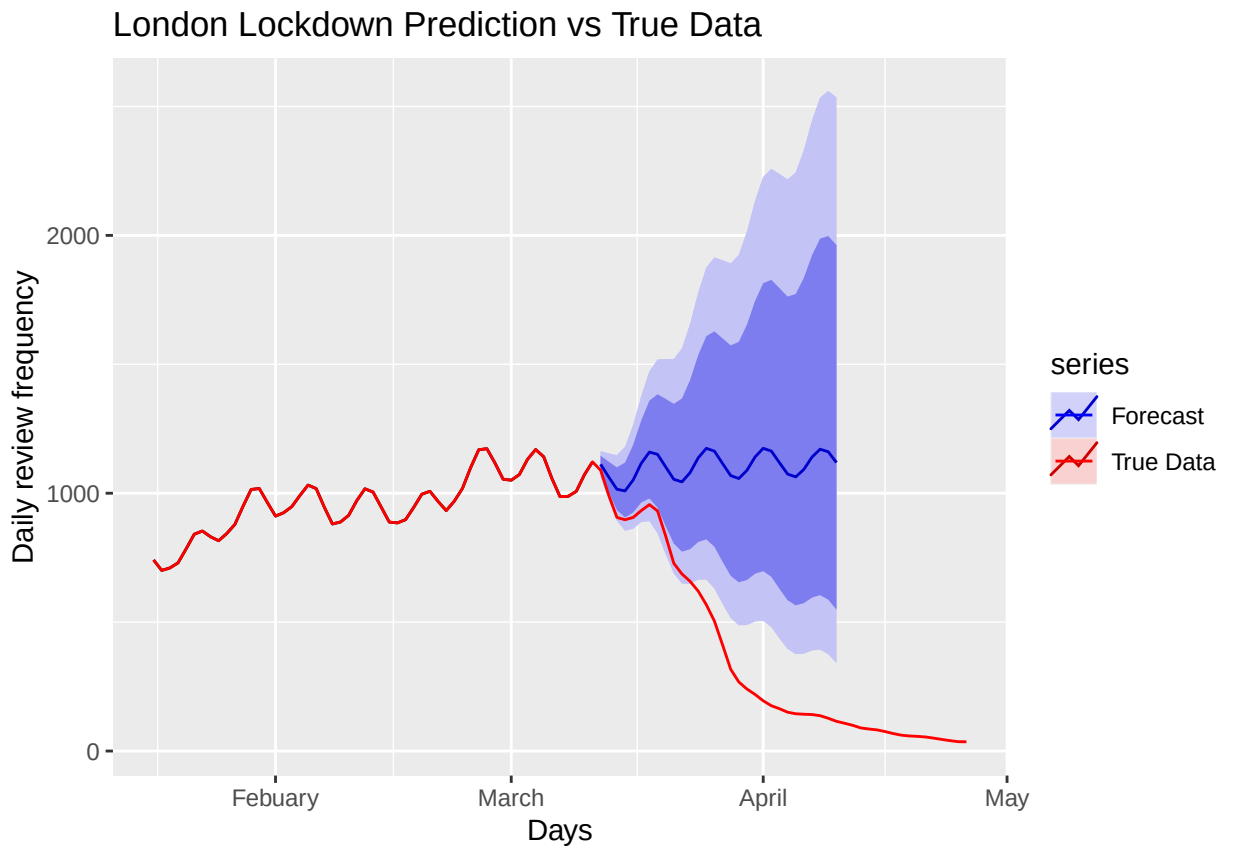



```
##
##  Ljung-Box test
##
## data:  Residuals from BATS(0.374, {3,2}, 0.8, -)
## Q* = 504.13, df = 3, p-value < 2.2e-16
##
## Model df: 16.   Total lags used: 19
```

Checking the residuals, the model appears to be a good fit, with the residuals looking like noise and normally distributed around 0.

```
covid_data <- ma(date_select("2017-01-01", "2020-05-10", data_london)$Freq, 10)
train_data <- ma(date_select("2017-01-01", "2020-03-10", data_london)$Freq, 10)

train_data <- train_data[6:(length(train_data)-6)]

fit.tbats <- tbats(ts(train_data))
```

```
f.tbats <- forecast(fit.tbats, h=30)

plot <- autoplot(f.tbats) + autolayer(f.tbats, series="Forecast") +
  autolayer(ts(covid_data[6:(length(covid_data)-6)]), series="True") +
  scale_x_continuous(breaks = cumsum(c(1115, 29, 31, 30)),
                     labels=c("Febuary", "March", "April", "May"), lim = c(1100, 1200)) +
  scale_color_manual(labels = c("Forecast", "True Data"), values = c("blue", "red"))

plot + labs(x="Days",
            y="Daily review frequency",
            title="London Lockdown Prediction vs True Data")
```



From the plot above, we can see just how drastic this decrease is. The model predicts a slight increase throughout march, much like the previous years, it also does well to model the weekly seasonal trend seen in the data. However, even at the edge of the 95% confidence interval, the model proves unable to even consider such a drop as was caused by the pandemic, indicating just how significant this was.

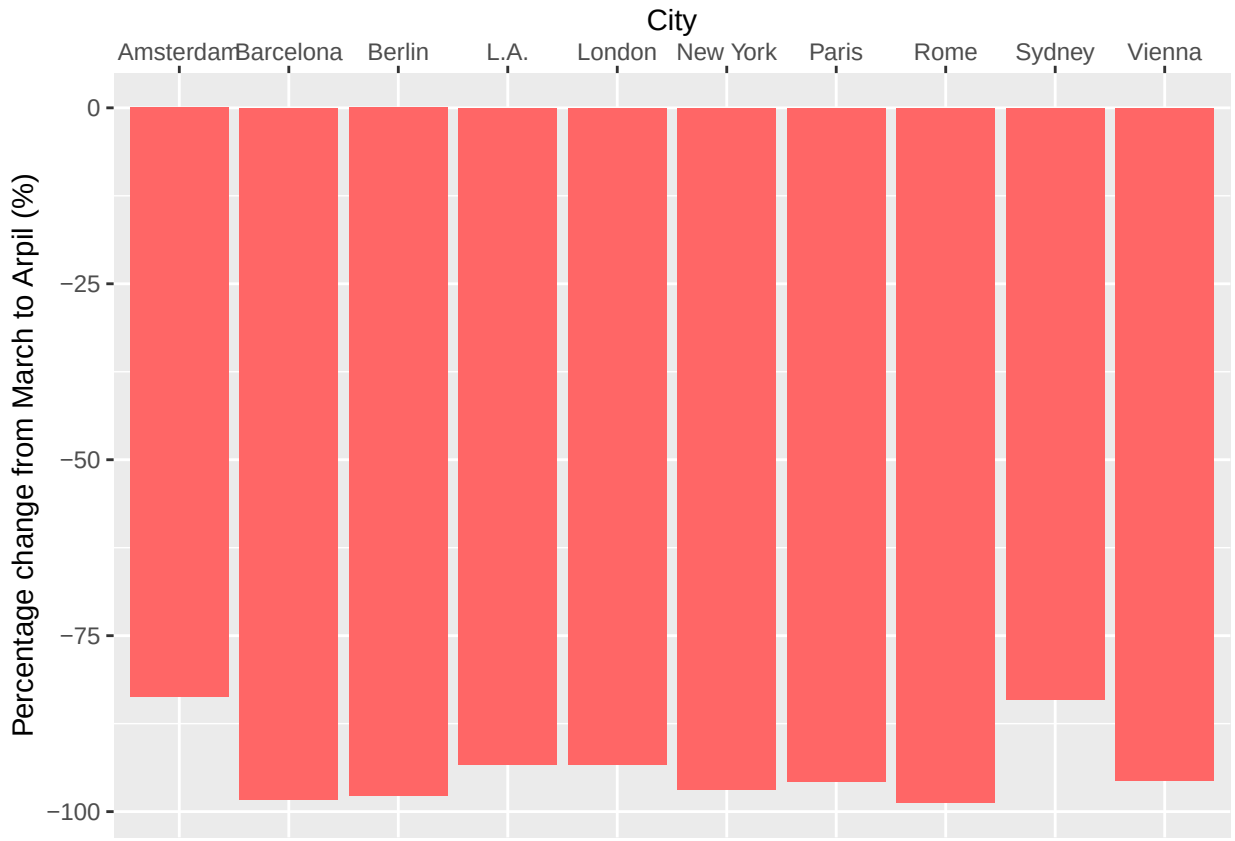
To obtain a better indication of the global impact behind the pandemic we must conisder data from cities around the world. Displaying such plots for each would soon prove tedious for the reader, hence, instead we must find an effective way of indicating the change in frequency between the predicted values and the observed values.

```
paths = c("../../reviews_london.csv",
          "../../reviews_rome.csv",
          "../../reviews_barcelona.csv",
```

```
          "../../reviews_paris.csv",
          "../../reviews_vienna.csv",
          "../../reviews_berlin.csv",
          "../../reviews_amsterdam.csv",
          "../../reviews_sydney.csv",
          "../../reviews_losangeles.csv",
          "../../reviews_newyork.csv")
names = c("London",
          "Rome",
          "Barcelona",
          "New York",
          "Paris",
          "Vienna",
          "Berlin",
          "Amsterdam",
          "Sydney",
          "L.A.",
          "New York"
          )


m <- compare_data(paths, names)
```

```
## [1] 1119.019
## [1] 767.5569
## [1] 567.835
## [1] 1012.195
## [1] 290.7308
## [1] 561.7326
## [1] 285.3766
## [1] 559.4435
## [1] 1177.74
## [1] 1119.019
```

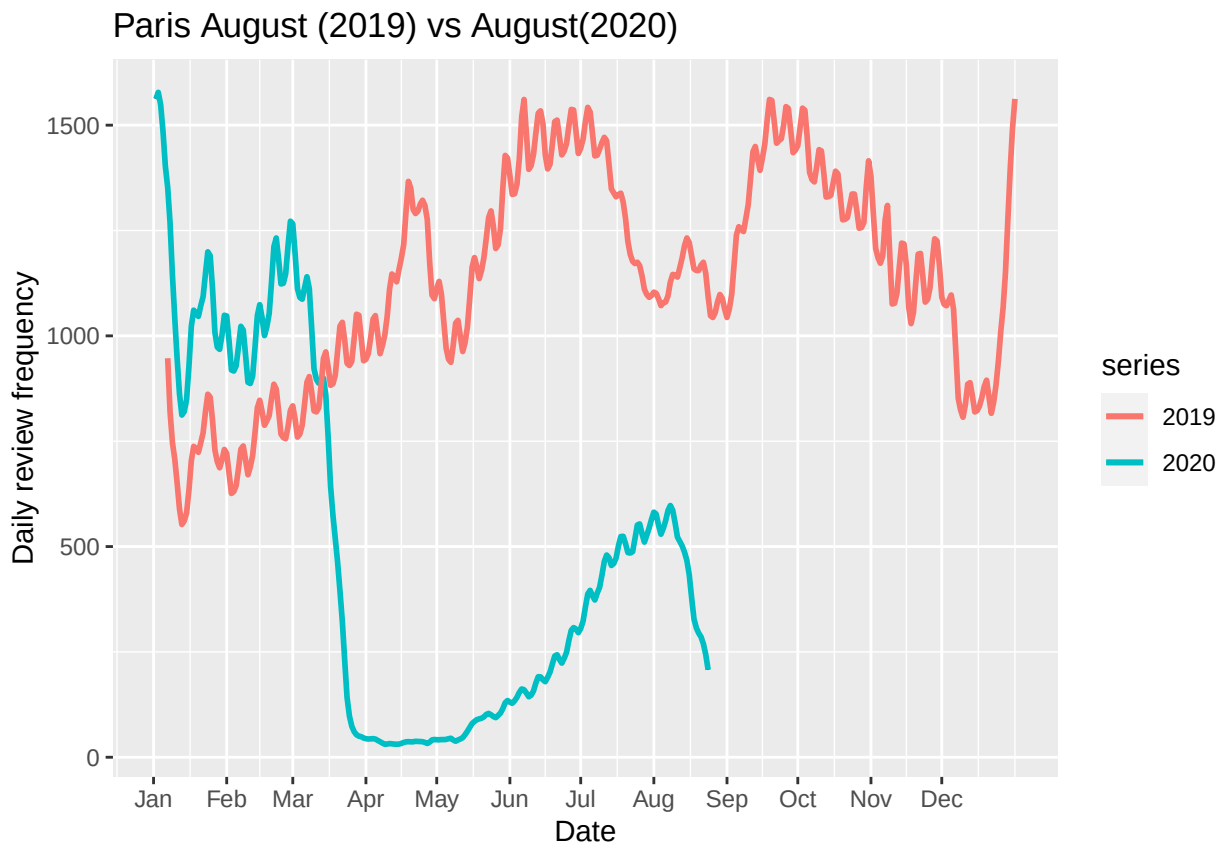
```
percentage_plot(m)
```



From the plot above, we can see that most major tourist hubs around the world saw a significant decline between mid-march to mid-april, as this was the period which saw the initial waves of cases globally. The greatest decline can be observed in regions which were most affected at the time, such as Italy and Spain, with Rome and Barcelona seeing a near 100% decline from the numbers forecasted for that month.

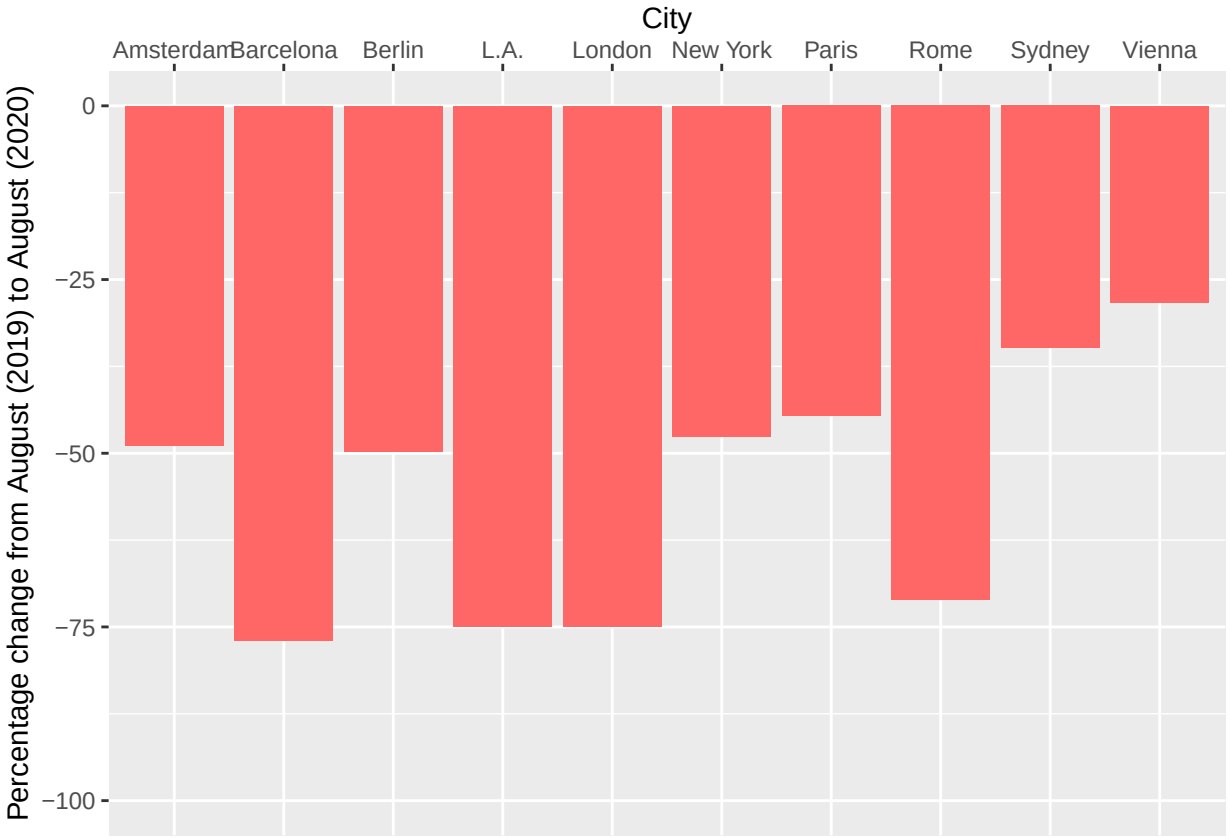
While significant, this was a somewhat expected result as global travel restrictions were certain to cause significant disruptions to the hospitality industry. We should now look to see how the data compares around august, a period where a reduction in cases saw many travel restriction lifted, with many governments attempting to restart their economies. Attempting to extract a model forecast that far into the future could prove uncertain as the confidence bands would grow in size, hence, instead let us take the observed data from the previous year and compare it against that observed during the august of 2020.

```
data_paris <- data_prep("../../reviews_paris.csv")
data <- ma(date_select("2019-01-01", "2021-01-01", data_paris)$Freq, 10)

data_2020 = ts(data[365:length(data)])
data_2019 = ts(data[0:365])

autoplot(data_2020) +
  autolayer(data_2020, series="2020", size=1) +
  autolayer(data_2019, series="2019", size=1) +
  scale_x_continuous(breaks= cumsum(c(0,31,28,31,30,31,30,31,31,30,31,30)),
                     labels=month.abb) +
  labs(title="Paris August (2019) vs August(2020)", x="Date", y="Daily review frequency")
```



We can see that while the numbers begin to rebound, they still fall short of the previous year. To allow for rapid comparison across different countries, we can produce a similar plot of the percentage-change as was seen previously.

```
multi_year <- multi_prev_year(paths, names)

breaks = c(0, -25, -50, -75, -100)
labs = c(0, -25, -50, -75, -100)

ggplot(multi_year, aes(x=city, y=perc)) +
    geom_bar(stat="identity", fill="#FF6666")  +
    scale_x_discrete(position = "top") +
    scale_y_continuous(breaks=breaks,
                       labels=labs,
                       lim=c(-100,0)) +
    xlab("City") +
    ylab("Percentage change from August (2019) to August (2020)")
```



This result proves hopeful, we can see that significant recovery is possible, with cities such as Paris recovering over 50% of the traffic, while most recover at least 25%, all this in only a few months since the lifting of the lockdown measures. The individual variations across cities stem largely from disparate local precautions advised by the government, as well as the domestic economic outlook.

It remains to be seen how the hospitality industry performs in the long run. However, the a data seen in this report indicates that the downturn caused by lockdown measures and travel restrictions can be quickly recovered following the lifting of such restrictions. The main uncertainty which remains is the pandemic itself. As of september there remain fears of a second wave which could see international travel restrictions reinstates in which case we would most likely see the same effect as was witnessed in early april, with the industry coming to a stand still.