Laboratory 1

Daniel Alonso

February 7, 2021

```
library(dplyr)
```

1. Read data

We first pick a time series from the *series.xls* file attached to the project and we create a *.csv* file with the selected time series.

We then select the *bits* column, excluding the time periods.

```
internet_hits<-read.csv(file="./internet_hits.csv", header=TRUE)
```

Looking at our data we notice that June 7th is incomplete, as it starts at 7:00, while the rest of the days start at 00:00, therefore we drop this day for consistency.

```
time <- internet_hits$time[18:length(internet_hits$time)]
internet_hits <- internet_hits$bits[18:length(internet_hits$bits)]
```

2. Define the time series object

We use the *ts* function to define the time series object. We start June 8th, 2005 and our frequency is hourly.

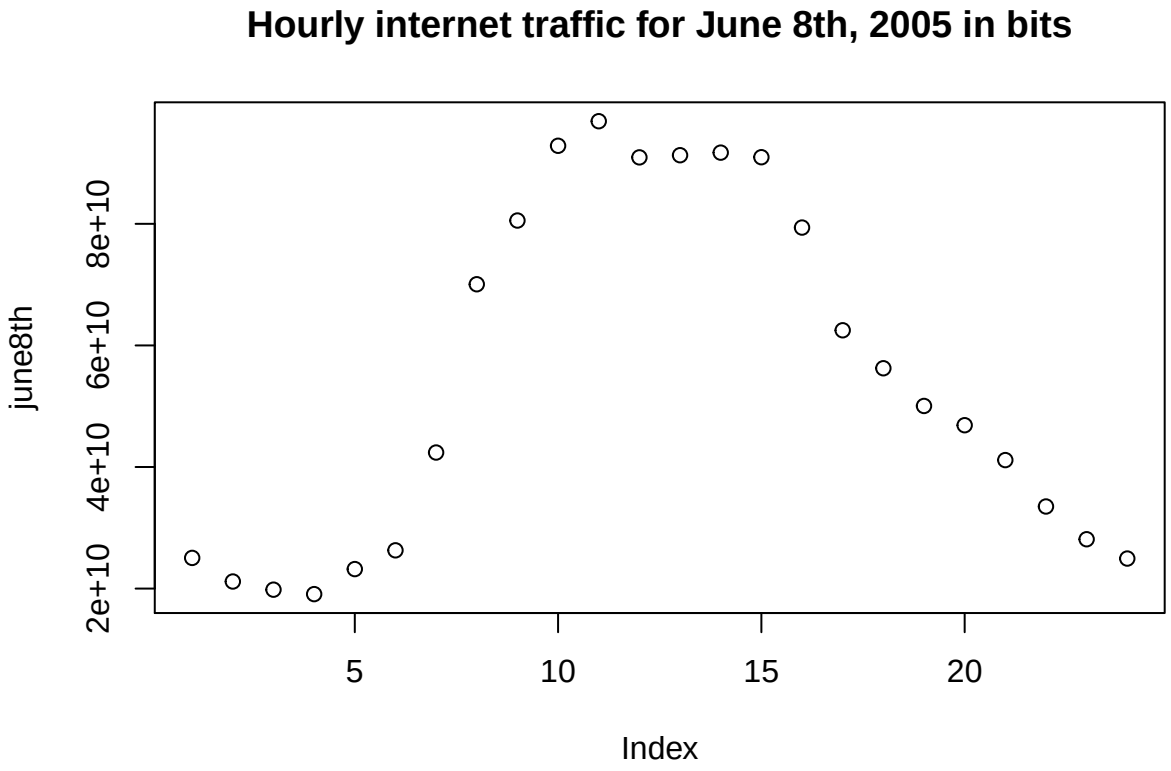
```
hits<-ts(internet_hits, freq=24)
```

3. Plot the time series

We proceed to plot the hourly internet traffic between June 7th and July 28th, 2005.

First let's plot traffic for Monday June 8th, 2005:

```
par(mfrow=c(1,1))
june8th <- hits[1:24]
plot(june8th,main="Hourly internet traffic for June 8th, 2005 in bits")
```



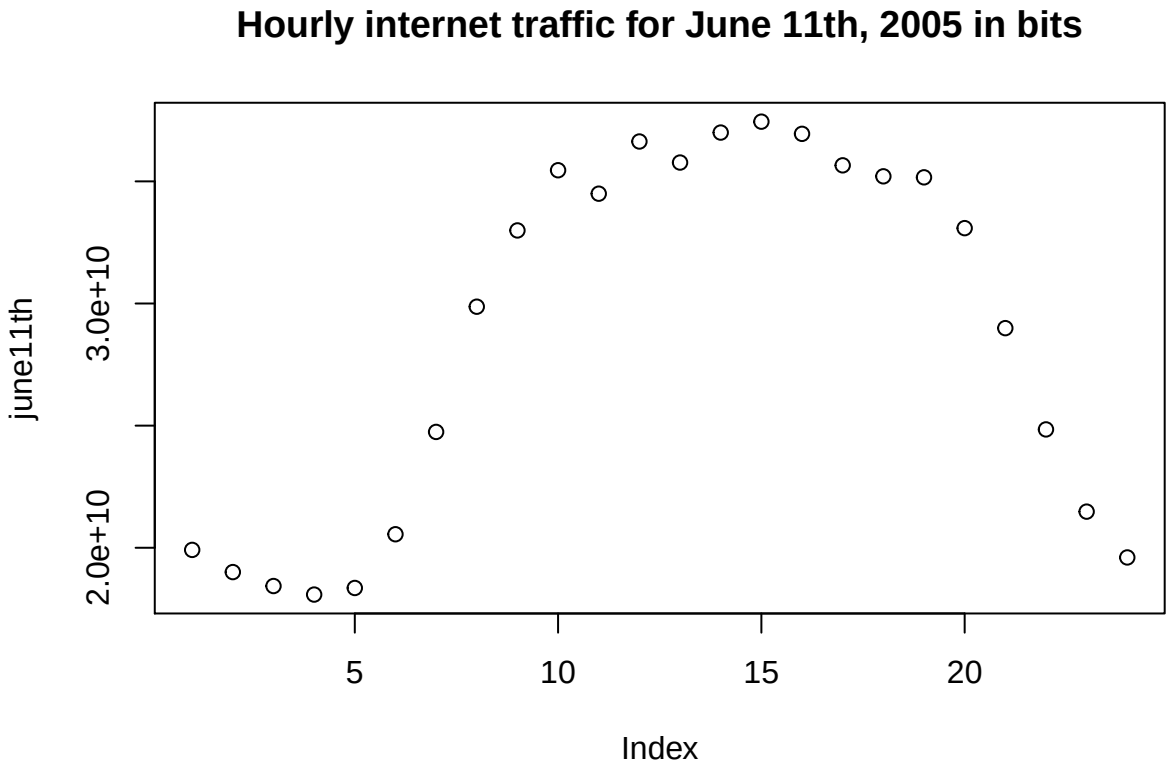
We can see that daily traffic peak hours is around lunchtime, between 11:00 and 14:00, showing significantly more traffic than any other time during the day.

We see that mornings are more active than late night and traffic remains somewhat consistent between 8:00 and 16:00 with a sharp drop around 17:00 and onwards.

The traffic for this day bottoms at around 4:00, as most people are probably sleeping.

Second: let's plot traffic for Saturday June 11th, 2005:

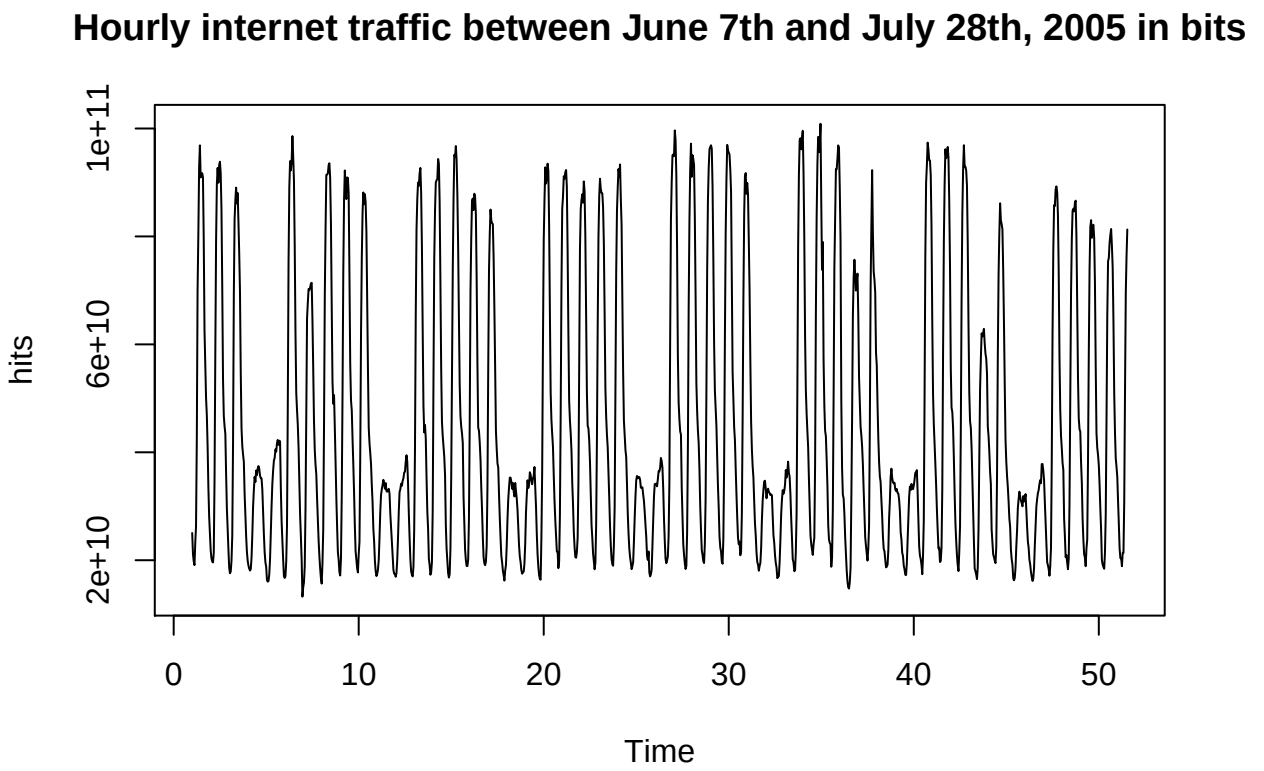
```
par(mfrow=c(1,1))
june11th <- hits[73:96]
plot(june11th,main="Hourly internet traffic for June 11th, 2005 in bits")
```



During weekends activity is reduced significantly. We see that the traffic peak hours now extend up to 19:00, so around 2 hours longer than weekdays, however, overall, total hourly traffic is around half that of weekdays with the exception of nights, where hourly traffic is more similar to that of weekdays.

Overall traffic

```
par(mfrow=c(1,1))
plot(hits,main="Hourly internet traffic between June 7th and July 28th, 2005 in bits")
```



We can clearly see that the seasonal pattern is repeated weekly, where weekdays are high traffic and weekends are low traffic days. As soon as the weekend is over.

4. logarithmic transformation

From the plots we can tell the data is homocedastic, therefore we do not need to perform a log transformation. There's clearly a strong seasonality aspect to the data which might create a perception of heterocedasticity in daily timeframes, but not in hourly timeframes.

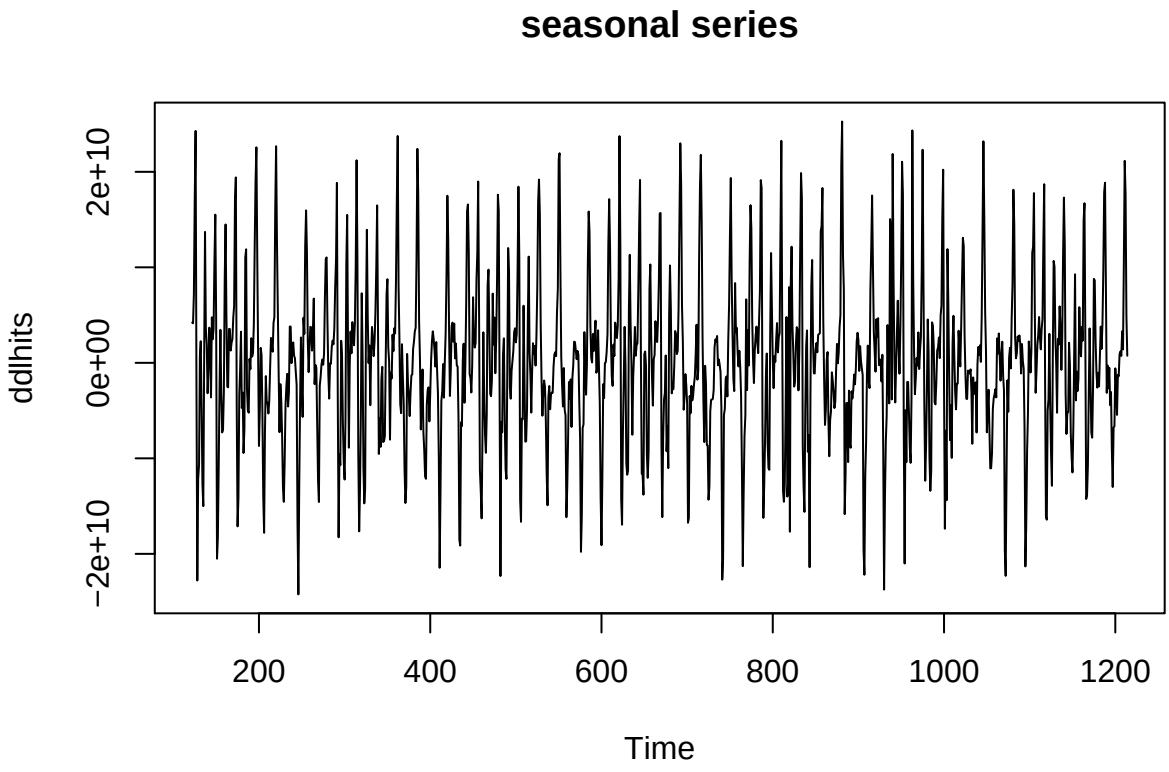
5. First difference

The time series does not have a trend, therefore we do not have to take the first difference.

6. Seasonal difference

As our data is hourly, we take the 120th-difference (5 days) to attempt to eliminate seasonality.

```
ddlhits<-diff(dlhits, lag=120)
plot(ddlhits, main="seasonal series")
```

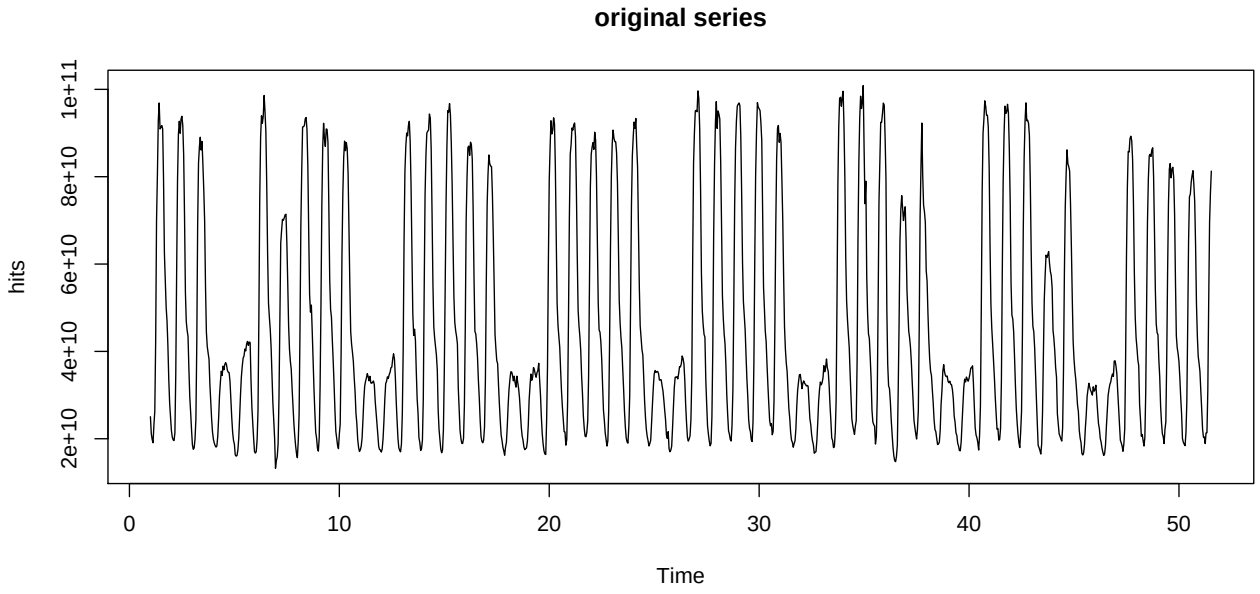


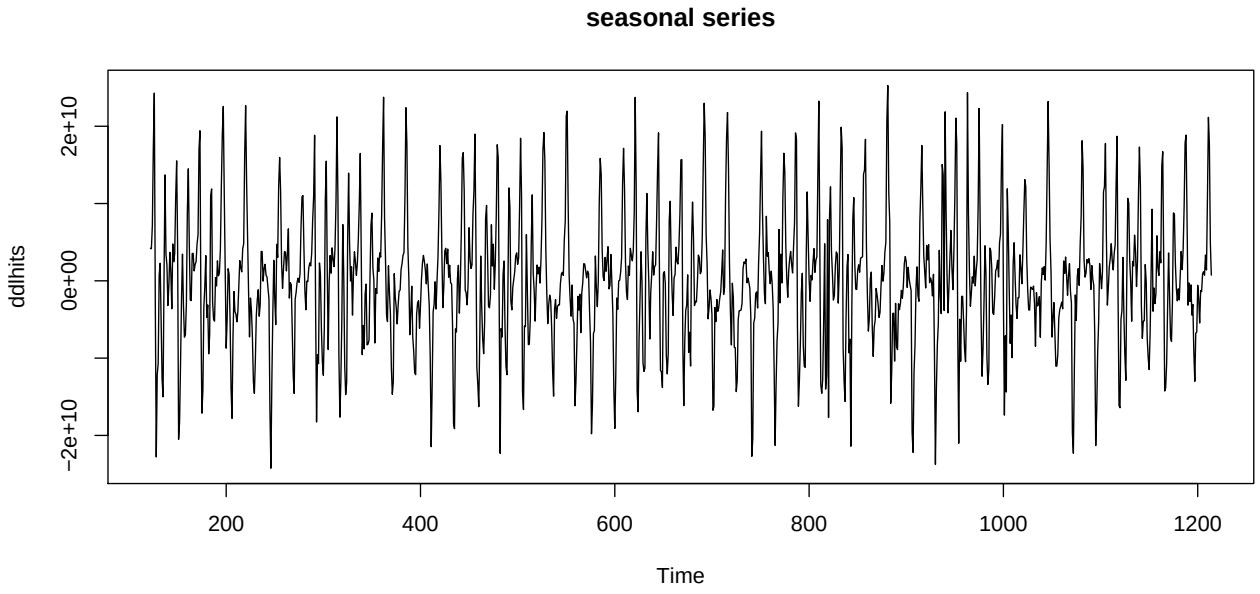
7. plots

We plot the seasonal and original series and we obtain the following:

- We can see the data is homocedastic and has no increasing/decreasing trend, therefore it is not stationary.
- We also notice that the seasonal series plot does not show any pattern, neither trend or seasonality, which we can pretty much deem white noise. We can see this time series is stationary.

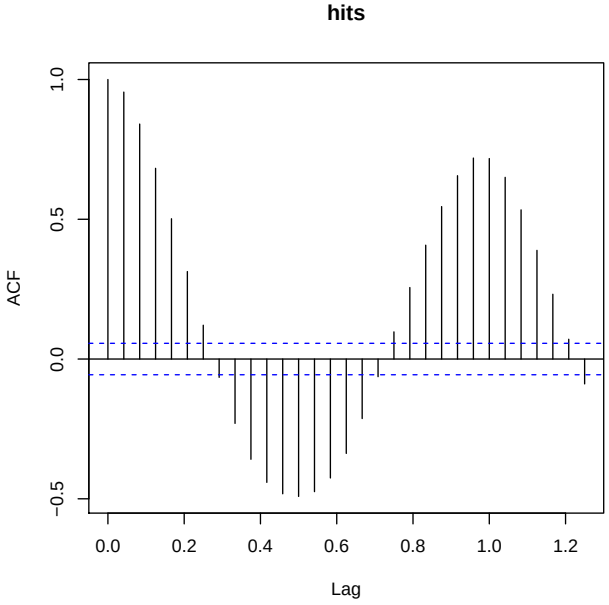
```
par(mfrow=c(2,1))
plot(hits, main="original series")
plot(ddlhits, main="seasonal series")
```

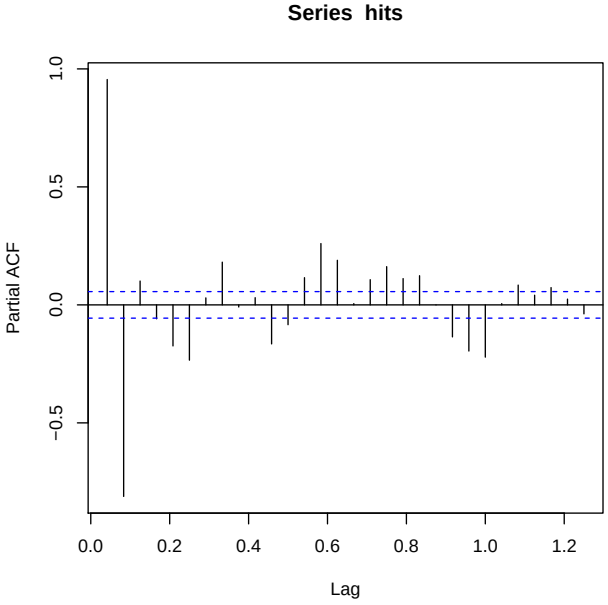


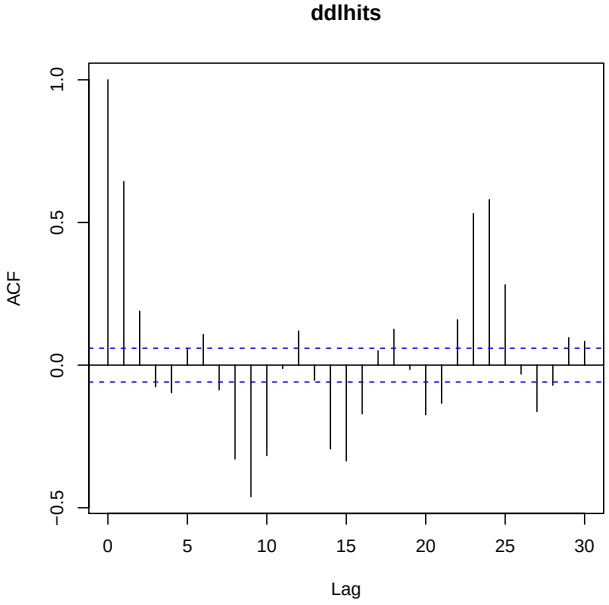


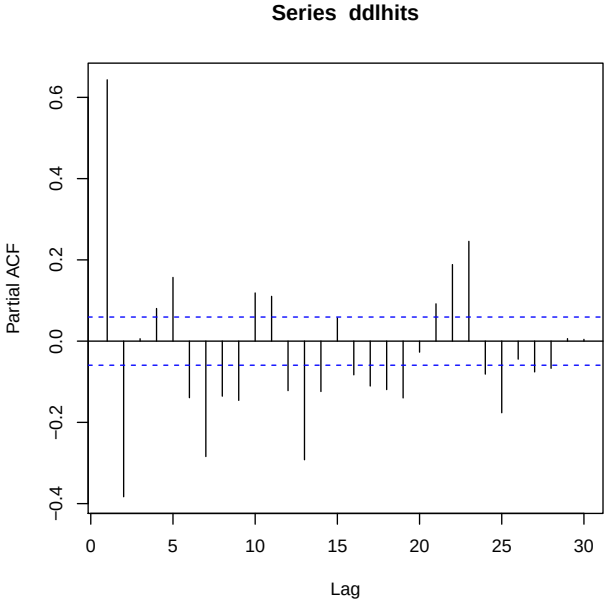
8. acf and pacf

Here we plot autocorrelations and partial autocorrelations. This way we can confirm the accuracy of the patterns we observe in previous steps, but might be hidden from us.









For the acf plot for *hits* we notice that there's a clear pattern, even clearer than before, we confirm that there is certainly a strong seasonality to the data. The event repeats itself consistently. The event has long run memory (in its own timeframe). And we clearly see that essentially all displayed coefficients are significant, as they all cross our significance boundaries.

For the pacf plot for *hits* we see that there's only significant wicks along seasonal limits along with the 2 first wicks.

For the acf and pacf plots for *ddlhits* where we attempt to remove seasonality we can see significant wicks, but this pattern is much less clear than in previous plots. Therefore, it's stationary.

9. Ljung-Box test

For the Ljung-Box test we test the following:

H_0: the model has no dependencies

H_1: the model has dependencies

```
Box.test(hits,lag=24)
#>
#>  Box-Pierce test
#>
#> data:  hits
#> X-squared = 7129.5, df = 24, p-value < 2.2e-16
```

As to the results looking at our P-vals, for the first test we already have observed and pointed out that the seasonality is present quite strongly in the time series. The test has an extremely large test statistic and a p-value of essentially zero. This, once again, confirms seasonality.

```
Box.test(ddlhits,lag=24)
#>
#>  Box-Pierce test
#>
#> data:  ddlhits
#> X-squared = 2037.2, df = 24, p-value < 2.2e-16
```

Our last test shows a significantly decreased test statistic versus the test conducted previously, however, we still reject the null hypothesis.