It’s mid-September and you’re wandering around your preferred supermarket when you stumble across the Christmas section. “Already”, you think. “It wasn’t like this back when I was a kid”. Well, with the power of data science maybe, just maybe, we can definitively say whether Christmas does indeed start earlier every year.

Anyone involved in marketing (or who wants to know how much their name gets googled) has probably checked out Google Trends. The main feature on offer is the ability to see the ‘Interest over time’ of a search term, relative to its peak popularity. So at each period, there is a “hits” value between 1 and 100, where 100 is the most popular that search term was in the time frame. Here’s an example of the search term “Trump”:

Unsurprisingly we see that the search term ‘Trump’ was most popular around November 2016, the same time as the US Presidential Election of that year where someone with the last name ‘Trump’ performed rather well. Maybe more surprising is how the searches for ‘Trump’ have been relatively low since the election, despite it feeling as if Donald has been a constant fixture in the news.

It is possible to download this information as a CSV file, which could then be read into R and we could have some fun with it. Luckily for us, there is an R package which makes getting this information into R even easier! **gtrendsR** provides an interface for retrieving this information provided by Google Trends. All the flexibility offered by the website is available via the package: we can look at trends for different areas, change the time range we see data for, even compare up to 5 search terms at once.

Sadly, as you can probably imagine Google Trends can only give us data going back so far, in this case being 2004. I’ve decided that we’ll focus on the last 10 years of data we can get from Google Trends; hopefully any patterns will present themselves over this period.

**Some Trendy Data Science**

Let’s start by loading in the **gtrendsR** package and using its main function, gtrends, to pull the search pattern for ‘Christmas’ over the last 10 years. We pass the search term through as a string, to the argument keyword, while we define the time range we want to gather data for to the argument time. There are lots of predefined strings you can pass to time, whether you want the last hour, last day or previous week. If none of these capture what you’re after, you can define your own time range or set time = "all" to get all available data.

library(gtrendsR)

xmas <- gtrends(keyword = "Christmas",

time = "2009-01-01 2018-12-01")

class(xmas)

## [1] "gtrends" "list"

We see that xmas is an object of type "gtrend" "list". It’s actually a list made up of seven data frames.

names(xmas)

## [1] "interest\_over\_time" "interest\_by\_country" "interest\_by\_region"

## [4] "interest\_by\_dma" "interest\_by\_city" "related\_topics"

## [7] "related\_queries"

If we look at the related\_topics data frame we see that these results are not just for explicitly searching “Christmas”. It also takes into account related topics, such as ‘Christmas Day’, ‘Gift’ and ‘Tree’.

library(dplyr)

xmas$related\_topics %>% select(-category) %>% head()

## subject related\_topics value keyword

## 1 100 top Christmas Day Christmas

## 2 8 top Christmas tree Christmas

## 3 8 top Tree Christmas

## 4 5 top Gift Christmas

## 5 5 top Christmas music Christmas

## 6 3 top Christmas lights Christmas

At the moment we’re interested in the interest\_over\_time data frame. interest\_over\_time provides us with the *hits* over time for the keyword provided. Remember, this is all relative to the peak popularity over the ten years we’ve pulled data for. The information from this data frame is what we see when we pass the gtrends object through the plot function.

plot(xmas)

Note that the intervals between the observations depend on the size of the time range you provide. Here we’ve provided a large time range of 10 years, which gives us a *hits* value for every month. This is okay when we’re looking at the data like this, but we want to compare the trend in more detail for each year, to see if it’s changed over the years. Hence we will pull the data for each year separately so that we get a *hits* value for each week. The downside is that now instead of being relative to the most popular point over the 10 years, the value of *hits* will only be relative to the most popular point in each year, which is always the observation closest to Christmas day. We can still use this to get a good idea of the trend over the year however, in particular the run-up to Christmas.

**So does Christmas start earlier every year?**

First let’s pull the data for each full year individually, going back to 2009. We’ll do this using a nice for loop that calls gtrends for each time frame then rbind them all together.

dates <- c("2017-01-01 2017-12-31", "2016-01-01 2016-12-31", "2015-01-01 2015-12-31",

"2013-01-01 2013-12-31", "2012-01-01 2012-12-31", "2011-01-01 2011-12-31",

"2010-01-01 2010-12-31", "2009-01-01 2009-12-31")

allXmas <- data.frame(date = character(0),

hits = numeric(0),

keyword = character(0),

geo = character(0),

gprop = character(0),

category = character(0))

for(i in dates) {

trendData <- gtrends(keyword = "Christmas",

time = i)

allXmas <- rbind(allXmas,

trendData$interest\_over\_time)

}

We need to create a few columns now, one which simply defines the year of that observation and another which details how far into the year we are, we’ll do this by obtaining the day of the year. The year function from **lubridate** makes extracting the year easy enough, while after a bit of Googling of my own I discover the function strftime which does the job for the day of the year.

library(lubridate)

allXmas <- mutate(allXmas, year = year(date),

doy = as.numeric(strftime(date, format = "%j")))

head(allXmas)

## date hits keyword geo gprop category year doy

## 1 2017-01-01 01:00:00 12 Christmas world web 0 2017 1

## 2 2017-01-08 01:00:00 5 Christmas world web 0 2017 8

## 3 2017-01-15 01:00:00 3 Christmas world web 0 2017 15

## 4 2017-01-22 01:00:00 3 Christmas world web 0 2017 22

## 5 2017-01-29 01:00:00 2 Christmas world web 0 2017 29

## 6 2017-02-05 01:00:00 2 Christmas world web 0 2017 36

Just what we were after.

Now we get to the visualisation, we’ll use ggplot to offer a bit more flexibility.

library(ggplot2)

ggplot(data = allXmas,

mapping = aes(x = doy, y = hits, colour = factor(year))) +

geom\_line(size = 0.5)

Ah…. There doesn’t seem to be much difference there, let’s focus on the latter part of the year.

ggplot(data = allXmas,

mapping = aes(x = doy, y = hits, colour = factor(year))) +

geom\_line(size = 0.5) +

xlim(200, 365)

We can see slightly more in this plot, still it’s hard to discern any differences between the escalation of Christmas hype from 2009 up to 2016. The only clear difference seems to be between 2017 and the other years, where the search trend for Christmas clearly increased sooner than for any other year. The fact that this is the most recent year may just be coincidence, we don’t see a clear scale where year upon year the searches for Christmas have increased earlier in the year.

With 2018 still being in the early stages of the festivities, it doesn’t make sense to pull its Christmas trend data through on its own: the hits value of 100 would be the most recent observation. Instead, we can pull it through alongside the 2017 trend, so we’ll get information relative to the volume of Christmas related searches in 2017.

twenty1718 <- gtrends(keyword = "Christmas",

time = "2017-01-01 2018-12-01")

overTime <- mutate(twenty1718$interest\_over\_time,

year = year(date),

doy = as.numeric(strftime(date, format = "%j")))

ggplot(data = overTime,

mapping = aes(x = doy, y = hits, colour = factor(year))) +

geom\_line(size = 1)

The popularity of Christmas this year appears to be increasing at an almost identical rate to last year up to the start of December.

So unfortunately just looking at the visualisations of the trends we can’t deduce much about whether Christmas does come earlier every year. When we looked at the trends from 2009 up to 2017 it did appear as if 2017 had an earlier build up to Christmas and 2018 so far is following the same pattern. If you felt like Christmas started earlier this year and last then you were right! Maybe 2017 was the tipping point and every year from now will either follow the same pattern or the trend will start even earlier? We’ll have to check back in a few years to see how future Christmas’ panned out.

Luckily we don’t have to come away from this analysis empty-handed, we can have a mess around with a few other features **gtrendsR** has to offer!

**Location, Location, Location**

Another great feature of Google Trends is the ability to see how the popularity of a search term varied for a particular location, as well as the ability to compare across multiple locations. For example, let’s compare the relative popularity of Christmas in the UK to the US. To select a particular location we will need its,country\_code often a shortening of its name, which we then pass through the gtrends function to the argument geo. To obtain this code we can look it up in the countries dataset that comes with **gtrendsR**.

data("countries")

filter(countries, sub\_code == "") %>%

head()

## country\_code sub\_code name

## 1 AF AFGHANISTAN

## 2 AL ALBANIA

## 3 DZ ALGERIA

## 4 AS AMERICAN SAMOA

## 5 AD ANDORRA

## 6 AO ANGOLA

filter(countries, name %in% c("UNITED KINGDOM", "UNITED STATES"))

## country\_code sub\_code name

## 1 GB UNITED KINGDOM

## 2 US UNITED STATES

The code for the United States is “US” as expected, while the code for the UK is actually “GB”. Next job is to pull the data for each country at the same time so we have a direct comparison. We’re going to look at the trends over 2017, as we’re only looking at one year now we don’t need to create a variable for ‘day of the year’ this time.

byCountry <- gtrends(keyword = "Christmas",

time = "2017-01-01 2017-12-31",

geo = c("US", "GB"))

ggplot(data = byCountry$interest\_over\_time,

mapping = aes(x = date, y = hits, colour = factor(geo))) +

geom\_line(size = 1)

Looks like Christmas is a bigger thing here in the UK than over in America. Weirdly when the search frequency for ‘Christmas’ peaks in the US, its popularity in the UK has already started to decrease. Not only is there a higher peak in the use of ‘Christmas’ as a search term but the increase in popularity also begins sooner in the year.

It’s easy to see how knowing this could make a marketing campaign for Christmas much more efficient. For example, if an ad campaign based around Christmas was released at the start of October, then in theory, this should receive a lot more interest in the UK than in the US. A staggered release could mean that the campaign is most prominent when the sharpest increase in interest in Christmas is happening in both the UK and America.

One possible explanation of the difference in Christmas search trends on either side of the Atlantic could be other holidays, Halloween and Thanksgiving are more popular in America as we see here.

otherHols <- gtrends(keyword = c("Halloween", "Halloween", "Thanksgiving", "Thanksgiving"),

time = "2017-01-01 2017-12-31",

geo = c("US", "GB", "US", "GB"))

otherHols <- otherHols$interest\_over\_time %>%

mutate(hits = as.numeric(ifelse(hits == "<1", 0, hits)))

ggplot(data = otherHols,

mapping = aes(x = date, y = hits, colour = geo,

linetype = keyword)) +

geom\_line(size = 1)

Note here that we have to do a bit of juggling; anything with a hits value of less than 1 is reported as “<1”, but as this is a string R then takes hitsto be a categorical variable rather than numerical variable. So we convert all instances of “<1” to 0 and explicitly tell R that hits is numeric.

Halloween is only marginally more searched in the US than the UK, Thanksgiving is where the main difference can be found. Happening on the fourth Thursday of November every year Thanksgiving could be a large part of the reason why there is a delay in the interest in Christmas in the States. Another event which isn’t exactly a holiday but goes hand in hand with Thanksgiving is ‘Black Friday’. Taking place the Friday after Thanksgiving this shopping spectacular is growing in popularity (you only have to look at its Google Trend to see that).

bf <- gtrends(keyword = "Black Friday")

bf <- bf$interest\_over\_time %>%

mutate(hits = as.numeric(ifelse(hits == "<1", 0, hits)))

ggplot(data = bf,

mapping = aes(x = date, y = hits)) +

geom\_line(size = 1, colour = "red")

Searches for ‘Black Friday’ doubled from 2014 to 2017, but is this a distraction from Christmas or an event which makes us think about our Christmas shopping earlier? Let’s compare the search trends for Christmas, Thanksgiving, Halloween and Black Friday, splitting by location.

#First pull the data for the UK

allEventsUK <- gtrends(keyword = c("Christmas", "Thanksgiving",

"Halloween", "Black Friday"),

time = "2017-01-01 2017-12-31",

geo = "GB")

allEventsUK <- allEventsUK$interest\_over\_time %>%

mutate(hits = as.numeric(ifelse(hits == "<1", 0, hits)))

# Then for the US

allEventsUS <- gtrends(keyword = c("Christmas", "Thanksgiving",

"Halloween", "Black Friday"),

time = "2017-01-01 2017-12-31",

geo = "US")

allEventsUS <- allEventsUS$interest\_over\_time %>%

mutate(hits = as.numeric(ifelse(hits == "<1", 0, hits)))

# Combine UK and US data

allEvents <- rbind(allEventsUK,

allEventsUS)

# Plot

ggplot(data = allEvents,

mapping = aes(x = date, y = hits, colour = keyword)) +

geom\_line(size = 1) +

facet\_wrap(~ geo) +

scale\_x\_datetime(limits = c(as.POSIXct("2017-07-01"), NA))

There are lots of differences between search patterns in the UK and US. The main difference in a single holiday being with Thanksgiving, more searched than Halloween in America but barely featured on the UK plot. Surprisingly Black Friday was searched more at its peak than Christmas in the US, although it is a sharper increase and decrease than the gradual rise of Christmas. In both plots, we find an increase in searches relating to Christmas as the popularity of Halloween decreases. This is one of the two sharpest increases in Christmas searches in the UK, with the other occurring as searches for Black Friday decreased.

The UK seems to show that we don’t focus on more than one holiday at once, with Christmas searches plateauing during peak periods for Halloween and Black Friday. Only to rise again once these holidays were over. If the popularity of Black Friday continues to increase at its current pace it will be interesting to see the effect this has on Christmas. From an earlier plot we know that Christmas is less popular at its peak in America than the UK, maybe this is related to the difference in Black Friday popularity in each country. Will Black Friday continue its rapid rise in the UK and result in the overall popularity of Christmas decreasing?

**All I want for Christmas is… some more data!**

Currently, it’s hard to determine whether any of these possible patterns will continue, at the moment 2017 appears to be the anomaly. Confirming any of our hunches will take a few more years of closely monitoring the Google Trends.

I would definitely recommend checking out Google Trends and **gtrendsR**, it’s not just for Christmas (data). Maybe you’ll be able to make some more solid conclusions!