R Coding Practice

Basic

Ralf Becker

June 2020

Aim for today

- Become familiar with some Covid related data
- Upload data from csv
- Undertake some basic data exploration
- Use time formats
- Create time-series graphs

Why, as economists, should we look at Covid-19

- Understanding the current and future needs are important for business and government for planning (toilet paper producers, fresh food importers, pasta retailers, NHS hospitals, etc)
- An event which allows us to reconsider the interplay between markets, government, and civil society

Data used

Today we will use two data sources.

1. Google mobility data, from https://www.google.com/covid19/mobility/.
2. Data from https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker. This dataset combines a range of Covid statistics, like the number of affected and the number of deseased.

Setup

Load the libraries needed.

```
library(forecast)     # used for some data smoothing
library(tidyverse)    # for almost all data handling tasks
library(ggplot2)      # plotting toolbox
library(xts)
```

Setup your working directory.

```
setwd("YOUR_WORKING_DIRECTORY")
```

or by using the menu - Session - Set Working Directory - To Source File Location

Import and examine Google mobility data

Go to https://www.google.com/covid19/mobility/ and download the "Global_Mobility_Report.csv" and save it in your working directory folder. We use the `csv_read` function to load the data andsave them into the `mob_data` dataframe. Also run the `str` function to see what variable are included.

```
# Data from https://www.google.com/covid19/mobility/
mob_data <- read.csv("Global_Mobility_Report.csv")
str(mob_data)
```

```
'data.frame':   477322 obs. of  11 variables:
 $ country_region_code                                : Factor w/ 131 levels "AE","AF","AG",..: 1 1 1 1 1
 $ country_region                                     : Factor w/ 132 levels "Afghanistan",..: 124 124 12
 $ sub_region_1                                       : Factor w/ 1845 levels "","Ã-rebro County",..: 1 1
 $ sub_region_2                                       : Factor w/ 1716 levels "","Abbeville County",..: 1
 $ date                                               : Factor w/ 105 levels "01/03/2020","01/04/2020",.
 $ retail_and_recreation_percent_change_from_baseline: int  0 1 -1 -2 -2 -2 -3 -2 -1 -3 ...
 $ grocery_and_pharmacy_percent_change_from_baseline : int  4 4 1 1 0 1 2 2 3 0 ...
 $ parks_percent_change_from_baseline                : int  5 4 5 5 4 6 6 4 3 5 ...
 $ transit_stations_percent_change_from_baseline     : int  0 1 1 0 -1 1 0 -2 -1 -1 ...
 $ workplaces_percent_change_from_baseline           : int  2 2 2 2 2 1 -1 3 4 3 ...
 $ residential_percent_change_from_baseline          : int  1 1 1 1 1 1 1 1 1 1 ...
```

There are 477322 rows of data and 11 variables. We have geographical information and information on activity indices. More detail on the latter soon.

You can see that there is a `date` variable which contains date information. This is currently formatted as a `fctor` variable, i.e. a categorical variable. Let's use the `as.Date` function to convert this variable into a date format such that R recognises that these are dates.

```
mob_data$date <- as.Date(as.character(mob_data$date),"%d/%m/%Y")
```

Let's look at a small subset of the data, in particular we pick out three comparable city regions. The regional information is saved in `sub_region_1`.

```
mob_data_sel <- mob_data %>% filter(sub_region_1 %in% c("Greater Manchester", "Stockholm County", "Berli
```

Let's look at some summary statistics.

```
summary(mob_data_sel)
```

```
 country_region_code             country_region                sub_region_1
 DE     :105         Germany             :105    Berlin                :105
 GB     :105         Sweden              :105    Greater Manchester    :105
 SE     :105         United Kingdom      :105    Stockholm County      :105
 AE     :  0         Afghanistan         :  0                          :  0
 AF     :  0         Angola              :  0    Ã-rebro County        :  0
 AG     :  0         Antigua and Barbuda:  0    Ã-stergÃ¶tland County:  0
 (Other):  0         (Other)             :  0    (Other)               :  0
           sub_region_2      date
                 :315   Min.   :2020-02-15
 Abbeville County:  0   1st Qu.:2020-03-12
 Acadia Parish   :  0   Median :2020-04-07
 Accomack County :  0   Mean   :2020-04-07
 Ada County      :  0   3rd Qu.:2020-05-03
 Adair County    :  0   Max.   :2020-05-29
 (Other)         :  0
 retail_and_recreation_percent_change_from_baseline
```

```
Min.   :-90.00
1st Qu.:-63.00
Median :-30.00
Mean   :-35.37
3rd Qu.: -8.00
Max.   : 11.00

grocery_and_pharmacy_percent_change_from_baseline
Min.   :-93.000
1st Qu.:-17.000
Median : -6.000
Mean   : -9.381
3rd Qu.:  0.000
Max.   :107.000
NA's   :3
parks_percent_change_from_baseline
Min.   :-63.000
1st Qu.:-13.000
Median :  6.000
Mean   :  8.594
3rd Qu.: 22.500
Max.   :148.000

transit_stations_percent_change_from_baseline
Min.   :-82.00
1st Qu.:-59.00
Median :-43.00
Mean   :-39.18
3rd Qu.:-11.00
Max.   :  6.00

workplaces_percent_change_from_baseline
Min.   :-87.00
1st Qu.:-53.00
Median :-35.00
Mean   :-32.23
3rd Qu.: -6.00
Max.   :  2.00

residential_percent_change_from_baseline
Min.   :-1.00
1st Qu.: 3.00
Median :12.00
Mean   :11.84
3rd Qu.:18.00
Max.   :32.00
```

You can see that there are 6 activity indices. Google extract these from the detailed user data they. For instance the `workplaces_percent_change_from_baseline` provides information on the extend to which they detected activities in workplaces. The numbers are percentage changes relative to a baseline. It is important to understand what the baseline is. This is a general point, you need to understand data definitions. On the https://www.google.com/covid19/mobility/data_documentation.html?hl=en you can find the following:

> Changes for each day are compared to a baseline value for that day of the week: The baseline is

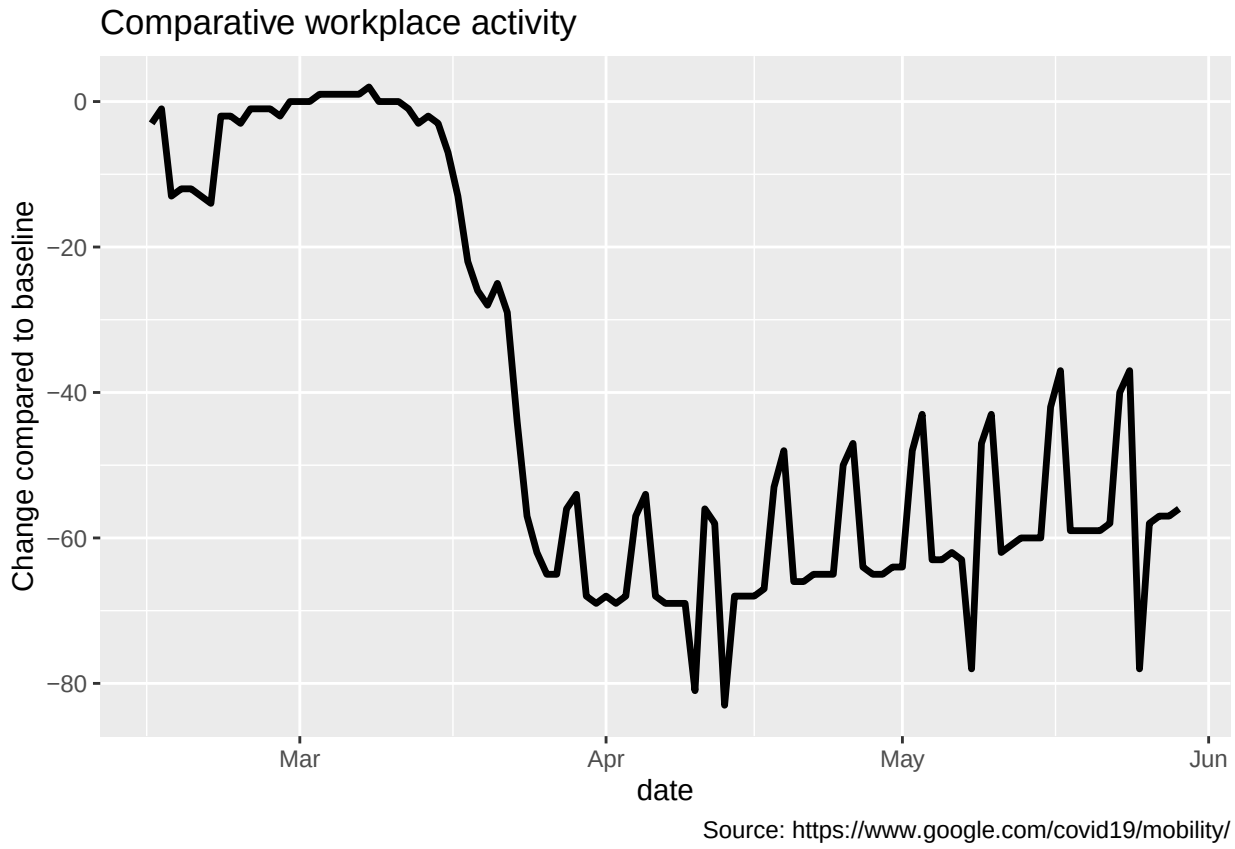
the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020.

Some data plots

Let's plot a few of the activity inices using the `ggplot` function.

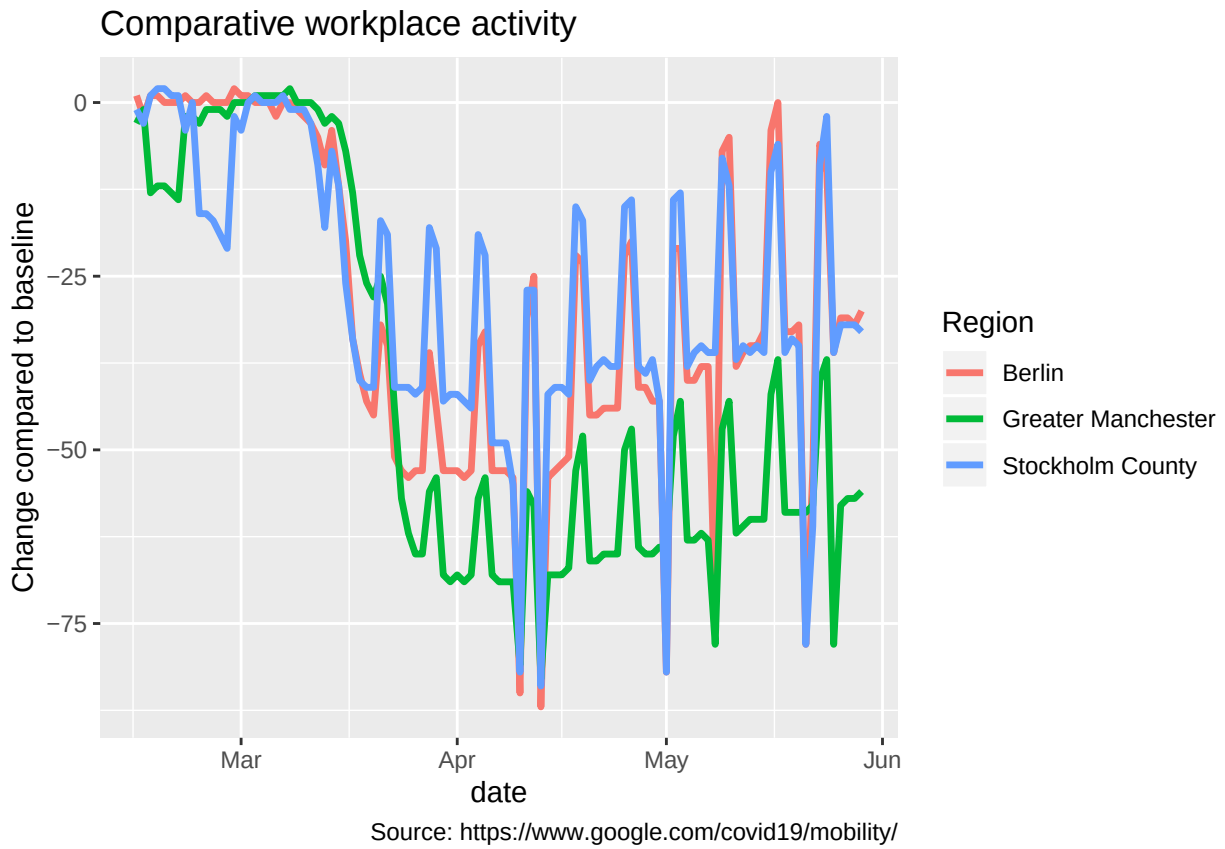
First we wpick out one of the locations, `Greater Manchester` and plot `workplaces_percent_change_from_baseline`. We achieve this by first creating a new (temporary) dataset, `temp` which only contains data from Manchester.

```
temp <- mob_data_sel %>% filter(sub_region_1 == "Greater Manchester")
ggplot(temp,aes(x =date, y=workplaces_percent_change_from_baseline)) +
  geom_line(size = 1.2) +
  labs(title = "Comparative workplace activity",
       caption = "Source: https://www.google.com/covid19/mobility/") +
  ylab("Change compared to baseline")
```



Let's add the same information for the Berlin and Stockholm. Hence we are using `mob_data_sel`.

```
ggplot(mob_data_sel,aes(x =date, y=workplaces_percent_change_from_baseline, color=sub_region_1)) +
  geom_line(size = 1.2) +
  labs(title = "Comparative workplace activity",
       caption = "Source: https://www.google.com/covid19/mobility/") +
  ylab("Change compared to baseline") +
  scale_color_discrete(name="Region")
```



You can clearly see the dip due to lockdowns, the gradual increase of the activity since and the weekly seasonality pattern.

Import policy and outcome data

Let's use another dataset which contains measures of how stringet a countrie's policies were to restrict the spread of the pandemic, but also contains some basic health indicators. Go to the https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker page of Blavatnik Schoool of Government (Uni of Oxford) and download the latest available data into your working directory.

```
# Data from https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
policy_data <- read.csv("OxCGRT_latest.csv")
# policy_data <- read.csv("https://raw.githubusercontent.com/OxCGRT/covid-policy-tracker/master/data/Ox
```

Let's restrict ourselves to the three countries which correspond to the above cities.

```
policy_data_sel <- policy_data %>% filter(CountryName %in% c("United Kingdom", "Sweden", "Germany"))
names(policy_data_sel)
```

```
 [1] "CountryName"
 [2] "CountryCode"
 [3] "Date"
 [4] "C1_School.closing"
 [5] "C1_Flag"
 [6] "C2_Workplace.closing"
 [7] "C2_Flag"
```

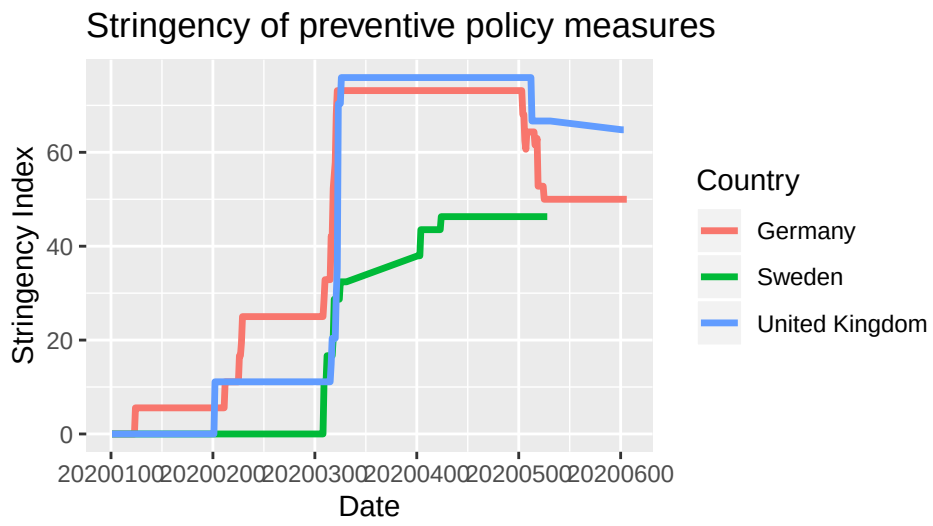
```
 [8] "C3_Cancel.public.events"
 [9] "C3_Flag"
[10] "C4_Restrictions.on.gatherings"
[11] "C4_Flag"
[12] "C5_Close.public.transport"
[13] "C5_Flag"
[14] "C6_Stay.at.home.requirements"
[15] "C6_Flag"
[16] "C7_Restrictions.on.internal.movement"
[17] "C7_Flag"
[18] "C8_International.travel.controls"
[19] "E1_Income.support"
[20] "E1_Flag"
[21] "E2_Debt.contract.relief"
[22] "E3_Fiscal.measures"
[23] "E4_International.support"
[24] "H1_Public.information.campaigns"
[25] "H1_Flag"
[26] "H2_Testing.policy"
[27] "H3_Contact.tracing"
[28] "H4_Emergency.investment.in.healthcare"
[29] "H5_Investment.in.vaccines"
[30] "M1_Wildcard"
[31] "ConfirmedCases"
[32] "ConfirmedDeaths"
[33] "StringencyIndex"
[34] "StringencyIndexForDisplay"
[35] "StringencyLegacyIndex"
[36] "StringencyLegacyIndexForDisplay"
[37] "GovernmentResponseIndex"
[38] "GovernmentResponseIndexForDisplay"
[39] "ContainmentHealthIndex"
[40] "ContainmentHealthIndexForDisplay"
[41] "EconomicSupportIndex"
[42] "EconomicSupportIndexForDisplay"
```

Dates are in the `Date` variable. They are formatted as `20200521` for 21 May 2020. Let's translate these into date format.

```
policy_data$Date <- as.Date(as.character(policy_data$Date),"%Y%m%d")
```

The variable `StringencyIndex` contains an index describing teh severity of the policy measures imposed.

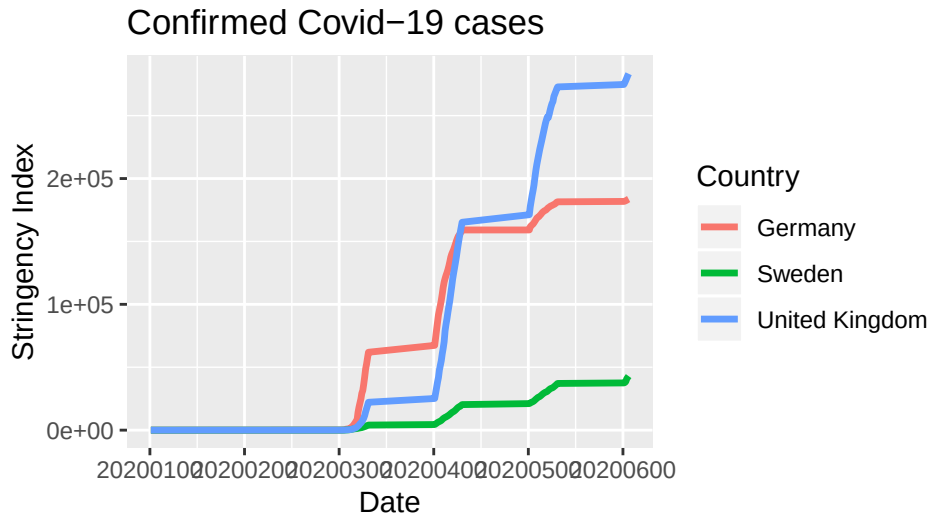
```
ggplot(policy_data_sel,aes(x =Date, y=StringencyIndex, color=CountryName)) +
  geom_line(size = 1.2) +
  labs(title = "Stringency of preventive policy measures", caption = "Source: Univ of Oxford, Blavatnik
  ylab("Stringency Index") +
  scale_color_discrete(name="Country")
```



Source: Univ of Oxford, Blavatnik Schoool of Government

Let's also look at some infection numbers.

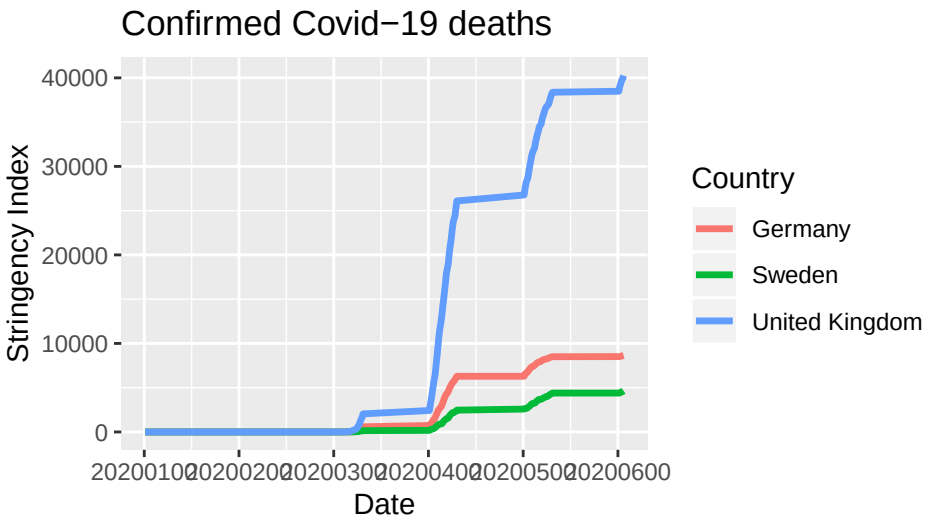
```
ggplot(policy_data_sel,aes(x =Date, y=ConfirmedCases, color=CountryName)) +
  geom_line(size = 1.2) +
  labs(title = "Confirmed Covid-19 cases", caption = "Source: Univ of Oxford, Blavatnik Schoool of Gover
  ylab("Stringency Index") +
  scale_color_discrete(name="Country")
```



Source: Univ of Oxford, Blavatnik Schoool of Government

Or now the number of confirmed deaths.

```
ggplot(policy_data_sel,aes(x =Date, y=ConfirmedDeaths, color=CountryName)) +
  geom_line(size = 1.2) +
  labs(title = "Confirmed Covid-19 deaths", caption = "Source: Univ of Oxford, Blavatnik Schoool of Gov
  ylab("Stringency Index") +
  scale_color_discrete(name="Country")
```



Source: Univ of Oxford, Blavatnik Schoool of Government

When looking at the number of deaths one would have to concluded that the UK has fared worse so far. This, however, does not take the size of the population into account. While the UK population is about 66 million, that of Sweden is about 10 million. If you adjust for this, then, in terms of deaths Sweden and the UK have about similar numbers.