

Initial Data Collection Report

Louise Braithwaite

Background

The initial data for this project was guided by two existing COVID-19 dashboards. The first was a PowerBI dashboard called ‘COVID-19 in Northern England’ created by Professor Nick Holliman. This dashboard reports on COVID-19 cases and deaths in Northern England and predominately uses data from UK government agencies such as, Public Health England and the Office of National Statistics. The second dashboard was created by the Urban Observatory, ‘Effects of Virus Measures’. This dashboard presents real-time data, which quantifies the impact of virus measures on people’s daily routines, such as pedestrian footfall, vehicle movements and car park data.

In addition to identifying two clear areas to start the data collection effort a third area of interest was also identified. As this project is about telling the story of COVID-19 in the North East the policy response and media coverage is also of importance. A timeline of national policy and health system responses to COVID-19 in England was sourced from The Health Foundation.

- Professor Nick Holliman dashboard
- Urban Observatory dashboard
- COVID-19 policy tracker

The initial data collected for the project is summarised in table 1.

Table 1: Summary of the initial data files

Data file	Data Source	Format	Size
Latest Lab-confirmed Covid-19 cases	Office of National Statistics	csv	3MB
Deaths registered weekly in England and Wales, provisional: 2020	Office of National Statistics	xlsx	190.3KB
Deaths registered weekly in England and Wales, provisional: 2019	Office of National Statistics	xls	159 KB
Deaths registered weekly in England and Wales, provisional: 2018	Office of National Statistics	xls	119 KB
Deaths registered weekly in England and Wales, provisional: 2017	Office of National Statistics	xls	92 KB
Deaths registered weekly in England and Wales, provisional: 2016	Office of National Statistics	xls	93 KB
Deaths registered weekly in England and Wales, provisional: 2015	Office of National Statistics	xls	91 KB

Data file	Data Source	Format	Size
Deaths registered weekly in England and Wales, provisional: 2014	Office of National Statistics	xls	91 KB
Deaths registered weekly in England and Wales, provisional: 2013	Office of National Statistics	xls	90 KB
Deaths registered weekly in England and Wales, provisional: 2012	Office of National Statistics	xls	90 KB
Deaths registered weekly in England and Wales, provisional: 2011	Office of National Statistics	xls	92 KB
Deaths registered weekly in England and Wales, provisional: 2010	Office of National Statistics	xls	88 KB
COVID-19 policy tracker	The Health Foundation	xlsx	754 KB

Daily Lab Confirmed Cases data

Data available for period 31 December 2019 - 22 June 2020. Table x provides an overview of the data when first loaded into R.

Data File	Column Header	Class	Example	Number of NAs
policy.tracker	"Date"	character	12.01.2020	0
policy.tracker	"ID"	double	2	0
policy.tracker	"Milestone"	character	"China shares genetic sequence. . ."	0
policy.tracker	"Detail"	character	"China shares genetic sequence. . ."	30
policy.tracker	"URL"	character	"WHO"	8
policy.tracker	"Category"	character	"Narrative"	8
policy.tracker	"Sub.category"	character	"Global spread and response"	9
policy.tracker	"Notes"	character	NA	182

Daily lab confirmed cases data

Data available for period 31 December 2019 - 22 June 2020. Table x provides an overview of the data when first loaded into R.

Data File	Column Header	Class	Example	Number of NAs
daily.lab.confirmed.cases	"Area.name"	character	"South West"	0
daily.lab.confirmed.cases	"Area.code"	character	"E12000009"	0
daily.lab.confirmed.cases	"Area.type"	character	"Region"	0
daily.lab.confirmed.cases	"Specimen.date"	date	2020-07-01	0
daily.lab.confirmed.cases	"Daily.lab.confirmed.cases"	double	0	4
daily.lab.confirmed.cases	"Previously.reported.daily.cases"	double	NA	40188
daily.lab.confirmed.cases	"Change.in.daily.cases"	double	NA	40188
daily.lab.confirmed.cases	"Cumulative.lab.confirmed.cases"	double	7973	0
daily.lab.confirmed.cases	"Previously.reported.cumulative.cases"	double	NA	40188
daily.lab.confirmed.cases	"Change.in.cumulative.cases"	double	NA	40188
daily.lab.confirmed.cases	"Cumulative.lab.confirmed.cases.rate"	double	142.4	0

The 4 NAs logged for the `Daily.lab.confirmed.cases` variable are missing values for observations on 20 March 2020 in Hackney and the City of London. These values will need to be considered when plotting graphs or if they are included when calculating averages. *Action* note the NAs in `Daily.lab.confirmed.cases` for further analysis. The large number of NAs for variables, `Previously.reported.daily.cases`, `Change.in.daily.cases`, `Previously.reported.cumulative.cases` and `Change.in.cumulative.cases` exist because there are only values for observations where `Area.name` is “England”. *Action* remove variables `Previously.reported.daily.cases`, `Change.in.daily.cases`, `Previously.reported.cumulative.cases` and `Change.in.cumulative.cases`

```
daily.lab.confirmed.cases %>% group_by(Area.type) %>% summarise(unique.values = length(unique(Area.name)))
```

#1 Lower tier local authority 316 #2 Nation 1 #3 Region 9 #4 Upper tier local authority 150 It is worth noting that some (not all) of the ‘upper tier local authority’ and ‘lower tier local authority’ variables have the same `Area.name` so it would not be helpful to filter by `Area.name`, as it often returns duplicate values.

Weekly deaths 2010-2020

These files are provided in xlsx excel format. As they are only going to be processed once, it would be easier to edit the xlsx files rather than automate the data loading process in R. Below are the steps taken to update the excel files.

Steps used when manually updating the Weekly COVID-19 Deaths xls files for 2010-2019

- Step 1: create a copy of the “Weekly Figures [INSERT YEAR]” tab. Add your initials to the new tab.
- Step 2: create a new tab and name it "TRANSPOSED[INSERT INITIALS]"
- Step 3: delete all rows apart from:
 - Week ended
 - Total deaths, all ages
 - North East
 - North West
 - Yorkshire and The Humber
 - East Midlands
 - West Midlands
 - East
 - London
 - South East
 - South West
 - Wales
- Step 4: select the remaining cells and copy
- Step 5: navigate into the “TRANSPOSED” tab and paste transpose (this will make the table wide)
- Step 6: insert new A column and name it “WeekNumber” insert the values 1:52 (note: in 2015 53 weeks are recorded)
- Step 7: there are a couple inconsistencies with the date formats, before saving check they are in the same format (DD/MM/YYYY)
- Step 8: save as xlsx file

After loading the weekly death data the rows are combined to create a single dataframe called `weekly-deaths2010.2020`. A summary of the data is provided below.

Data Frame	Column Header	Class	Example	Number of NAs
<code>weeklydeaths2010.2020</code>	“WeekNumber”	double	1	0
<code>weeklydeaths2010.2020</code>	“Week.ended”	date	“2010-01-08 00:00:00”	0
<code>weeklydeaths2010.2020</code>	“Total.deaths.all.ages”	double	12968	0
<code>weeklydeaths2010.2020</code>	“North.East”	double	648	0
<code>weeklydeaths2010.2020</code>	“North.West”	double	1686	0
<code>weeklydeaths2010.2020</code>	“Yorkshire.and.The.Humber”	double	1317	0
<code>weeklydeaths2010.2020</code>	“East.Midlands”	double	1091	0

Data Frame	Column Header	Class	Example	Number of NAs
weeklydeaths2010.2020	“West.Midlands”	double	1370	0
weeklydeaths2010.2020	“East”	double	1412	0
weeklydeaths2010.2020	“London”	double	1226	0
weeklydeaths2010.2020	“South.East”	double	1904	0
weeklydeaths2010.2020	“South.West”	double	1395	0
weeklydeaths2010.2020	“Wales”	double	879	0

Steps used when manually updating the Weekly COVID-19 Deaths xls files for 2020

- Step 1: create a copy of the “Weekly Figures [INSERT YEAR]” tab. Add your initials to the new tab, eg “Weekly Figure 2010_LB”
- Step 2: create a new tab and name it "TRANSPOSED[INSERT INITIALS]"
- Step 3: delete all rows apart from:
 - Week ended
 - Total deaths, all ages
 - North East
 - North West
 - Yorkshire and The Humber
 - East Midlands
 - West Midlands
 - East
 - London
 - South East
 - South West
 - Wales
- Step 4: copy and paste ‘Week ended’ and ‘Total deaths, all ages’ into column B then delete column A
- Step 5: select the remaining cells and copy
- Step 6: navigate into the “TRANSPOSED” tab and paste transpose (this will make the table wide)
- Step 7: insert new A column and name it “WeekNumber” insert the values 1:52 (note: in 2015 53 weeks are recorded)
- Step 8: there are a couple inconsistencies with the date formats, before saving check they are in the same format (DD/MM/YYYY)
- Step 8: create a copy of the “Covid-19 - Weekly occurrences” tab. Add your initials to the copied tab, eg. “Covid-19 - Weekly occurrencesLB”
- Step 9: create a new tab and name it "TRANSPOSEDcovid[INSERT INITIALS]"
- Step 10: delete all rows apart from:
 - Week ended
 - Deaths involving COVID-19, all ages1
 - North East
 - North West
 - Yorkshire and The Humber
 - East Midlands
 - West Midlands
 - East
 - London
 - South East
 - South West
 - Wales
- Step 11: copy and paste ‘Week ended’ and ‘Deaths involving COVID-19, all ages1’ into column B then delete column A
- Step 12: select the remaining cells and copy
- Step 13: navigate into the “TRANSPOSEDcovid” tab and paste transpose (this will make the table

wide)

- Step 14: * Step 14: insert new A column and name it “WeekNumber” insert the values 1:52 (note: in 2015 53 weeks are recorded)
- Step 15: delete the row of total numbers Week ended = “1 to 25”
- Step 16: there are a couple inconsistencies with the date formats, before saving check they are in the same format (DD/MM/YYYY)

The data collected can be is divided into three sections:

1. Government agency data. This includes data provided by Public Health England, the Office of National Statistics and the NHS. Data sets include:
 - DailyLabConfirmedCases
 - weeklyCovid19deaths
 - AllWeeklyDeaths
2. Urban Observatory data. Data sets include:
 - traffic
 - carpark
 - carpark.meta
 - pedestrian.flow
3. Policy response and media coverage
 - policy.data