# Exploratory_Data_Analysis

Louise Braithwaite

04/09/2020

## Introduction

Exploratory data analysis was conducted to:

- identify data quality issues
- inform the data preparation process
- analyise the data to identify how it can be used in the final dashboard design

The final analysis work was directed by the aim of the project, to tell a story about COVID-19 in the North East of England.

## Data Selection

The final dashboard uses thirteen data frames, which originate from eight original files. Table 1 provides a summary of the data frames.

**Table 1: data frame summary**

| Dashboard section | Figure | Shiny data frame | Source data frame (source) |
|---|---|---|---|
| Cases | Figure 1.1 | nationCases | cases (gov.uk) |
| Cases | Figure 1.2 | nationCases | cases (gov.uk) |
| Cases | Figure 1.3 | regionCases | cases (gov.uk) |
| Cases | Figure 1.4 | regionCases | cases (gov.uk) |
| Cases | Figure 1.5 | regionCases | cases (gov.uk) |
| Cases | Figure 1.6 | NEltlaCases | cases (gov.uk) |
| Cases | Figure 1.7 | NEltlaCases | cases (gov.uk) |
| Cases | Figure 1.8 | NEltlaCases | cases (gov.uk) |
| Deaths | Figure 2.1 | nationdeaths | deaths (gov.uk) |
| Deaths | Figure 2.2 | nationdeaths | deaths (gov.uk) |
| Deaths | Figure 2.3 | regiondeaths | deaths (gov.uk) |
| Deaths | Figure 2.4 | regiondeaths | deaths (gov.uk) |
| Traffic | Figure 3.1 | mediantrafficLONG | mediantraffic (Urban Observatory) |
| Traffic | Map 3.1 | Commutingtrafficmap | anpr.volumes.point.meta (Urban Observatory) |
| Traffic | Figure 3.2 | CommutingtrafficDAY | anpr.volumes.16min (Urban Observatory) |
| Car Parks | Map 4.1 | carpark.meta | carpark.meta (Urban Observatory) |
| Car Parks | Map 4.2 | carpark.meta | carpark.meta (Urban Observatory) |

| Dashboard section | Figure | Shiny data frame | Source data frame (source) |
|---|---|---|---|
| Car Parks | Figure 4.1 | carparkSummary | carpark (Urban Observatory) |
| Pedestrians | Map 5.1 | N/A | N/A |
| Pedestrians | Figure 5.1 | ped.daily.totals | pedestrian.flow (Urban Observatory) |
| Pedestrians | Figure 5.2 | NorthumberlandSt.east | pedestrian.flow (Urban Observatory) |
| Pedestrians | Figure 5.3 | NorthumberlandSt.west | pedestrian.flow (Urban Observatory) |

## Data Selection: Gov.uk

The cases and deaths data was pulled from the government's coronavirus api https://api.coronavirus.data.gov.uk/v1/data The data is prepared for a number of different geographic regions, referred to as areaType. A summary is provided in table 3.

**Table 2: an overview of the valid values for the areaType metric in the gov.uk coronavirus api**

| areaType value | Description |
|---|---|
| overview | Overview data for the United Kingdom |
| nation | Nation data (England, Northern Ireland, Scotland and Wales) |
| region | Region data (east midlands, east of england, london, north east, north west, south east, south west, west midlands, yorkshire and the humber |
| nhsregion | NHS region data (East of England, London, Midlands, North East and Yorkshire, North West, South East, South West) |
| utla | Upper-tier local authority data (there are 149 utla regions) |
| ltla | Lower-tier local authority data (there are 315 ltla regions) |

https://www.gov.uk/understand-how-your-council-works

### Cases

The Government's api provides seven metrics for cases:

- newCasesByPublishDate
- cumCasesByPublishDate
- cumCasesBySpecimenDateRate
- newCasesBySpecimenDate
- cumCasesBySpecimenDate
- maleCases
- femaleCases

The cases

for regions, upper tier local authorities and lower tier local authorities is presented by specimen date (the date when the sample was taken from the person being tested) and the case data for Nations is presented by reporting date (the date the case was first included in the published totals) The cases data can be presented

by specimen date (the date when the sample was taken from the person being tested) or by reporting date (the date the case was first included in the published totals). The availability of each of these time series varies by area. Nation date = “date”, name = “areaName”, code = “areaCode”, cases = list( daily = “newCasesBySpecimenDate”, cumulative = “cumCasesBySpecimenDate”, rate = “cumCasesBySpecimenDateRate”, dailyPublish = “newCasesByPublishDate”, cumulativePublish = “cumCasesByPublishDate” Region, Upper Tier Local Authority, Lower Tier Local Authority date = “date”, name = “areaName”, code = “areaCode”, cases = list( daily = “newCasesBySpecimenDate”, cumulative = “cumCasesBySpecimenDate”, rate = “cumCasesBySpecimenDateRate”,

## Data Quality Issues

## Data Preparation

## Data Analysis for Dashboard Design