Appendix - Data

Data source 1

Background
Data comes from Scottish Government: https://statistics.gov.scot/data/green-or-blue-space-shs

The Scottish Household Survey is an annual survey of over 10,000 households in Scotland, asking questions about people's homes, neighbourhoods and views on local public services. The survey *'provides robust evidence on the composition, characteristics, attitudes and behaviour of private households and individuals as well as evidence on the physical condition of Scotland's homes'*. 1

The Scottish Government, local councils and various charities use the results to improve the lives of people in your area and across Scotland. The survey has been running since 1999 and is independent of all political parties. 2

More about the survey here: http://www.scottishhouseholdsurvey.com/

Info about data
A csv file with ~33k rows and 13 columns. Has columns with Local Authority code, year, characteristics (age, gender, ethnicity, Scottish Index Multiple Deprivation etc.), distance banding, measurement and value.

FeatureCode	DateCode	Measurement	Units	Value	Distance to Nearest Green or Blue Space	Age	Gender	Urban Rural Classification	SIMD quintiles	Type Of Tenure	Household Type	Ethnicity
S12000050	2013	95% Upper Confidence Limit, Percent	Percent Of Adults	2.6	Don't Know	35-64 years	All	All	All	All	All	All
S92000003	2018	95% Lower Confidence Limit, Percent	Percent Of Adults	1.7	Don't Know	All	All	All	All	All	Pensioners	All
S12000013	2017	95% Lower Confidence Limit, Percent	Percent Of Adults	0.0	Don't Know	35-64 years	All	All	All	All	All	All
S12000041	2018	95% Upper Confidence Limit, Percent	Percent Of Adults	3.6	Don't Know	All	Female	All	All	All	All	All
S12000028	2014	Percent	Percent Of Adults	73	A 5 minute walk or less	All	Male	All	All	All	All	All
S12000014	2013	95% Lower Confidence Limit, Percent	Percent Of Adults	0.0	Don't Know	35-64 years	All	All	All	All	All	All
S12000042	2018	Percent	Percent Of Adults	9	Don't Know	All	Female	All	All	All	All	All
S12000028	2014	Percent	Percent Of Adults	68	A 5 minute walk or less	65 years and over	All	All	All	All	All	All
S12000027	2017	Percent	Percent Of Adults	80	A 5 minute walk or less	35-64 years	All	All	All	All	All	All
S12000050	2013	95% Upper Confidence Limit, Percent	Percent Of Adults	79.4	A 5 minute walk or less	All	All	All	All	All	With Children	All
S12000049	2017	95% Lower Confidence Limit, Percent	Percent Of Adults	0.9	Don't Know	All	All	All	All	All	All	All
S12000019	2014	Percent	Percent Of Adults	0	Don't Know	All	All	All	All	All	All	All
S12000049	2014	Percent	Percent Of Adults	2	Don't Know	16-34 years	All	All	All	All	All	All
S92000003	2015	95% Upper Confidence Limit, Percent	Percent Of Adults	2.0	Don't Know	16-34 years	All	All	All	All	All	All
S12000020	2014	95% Upper Confidence Limit, Percent	Percent Of Adults	1.6	Don't Know	All	All	All	All	All	Adults	All

Data source 2

Background
Data comes from NHS: https://www.opendata.nhs.scot/sl/dataset/geography-codes-and-labels/resource/967937c4-8d67-4f39-974f-fd58c4acfda5

Open data NHS dataset containing 32 Council Areas names and codes.

Info about data
A csv file with 32 rows and 13 columns, but only using first 2 (CA and CA name)

ca11_ca19

CA	CAName	CADateEnacted	CADateArchived	HSCP	HSCPName	HSCPDateEnacted	HSCPDateArchived	HB	HBName	HBDateEnacted	HBDateArchived	Country
S12000005	Clackmannanshire	19960401		S37000005	Clackmannanshire and Stirling	20160401		S08000019	NHS Forth Valley	20140401		S92000003
S12000006	Dumfries and Galloway	19960401		S37000006	Dumfries and Galloway	20160401		S08000017	NHS Dumfries and Galloway	20140401		S92000003
S12000008	East Ayrshire	19960401		S37000008	East Ayrshire	20150402		S08000015	NHS Ayrshire and Arran	20140401		S92000003
S12000010	East Lothian	19960401		S37000010	East Lothian	20160401		S08000024	NHS Lothian	20140401		S92000003
S12000011	East Renfrewshire	19960401		S37000011	East Renfrewshire	20151007		S08000021	NHS Greater Glasgow and Clyde	20140401	20190331	S92000003
S12000011	East Renfrewshire	19960401		S37000011	East Renfrewshire	20151007		S08000031	NHS Greater Glasgow and Clyde	20190401		S92000003
S12000013	Na h-Eileanan Siar	19960401		S37000031	Western Isles	20160401		S08000028	NHS Western Isles	20140401		S92000003
S12000014	Falkirk	19960401		S37000013	Falkirk	20160401		S08000019	NHS Forth Valley	20140401		S92000003
S12000015	Fife	19960401	20180201	S37000014	Fife	20160401	20180201	S08000018	NHS Fife	20140401	20180201	S92000003
S12000017	Highland	19960401		S37000016	Highland	20150401		S08000022	NHS Highland	20140401		S92000003

Data pre-processing

Data processing was done in R. I have attached the R script used to pre-process the data. Different pre-processing was done to use the different data visitation tools tested (see later section) but have only kept the processing used for the final chosen tool.

Pre-processing/EDA steps taken

1. Filtered so only had 'totals' rows - as information split by other variables was available (gender, ethnicity, Scottish Index Multiple Deprivation etc.) - and only the short walk distance and for the start and end time periods (2013 and 2018). Joined on data source 2 to get LA names (instead of having to identify by code).

Data cleaning

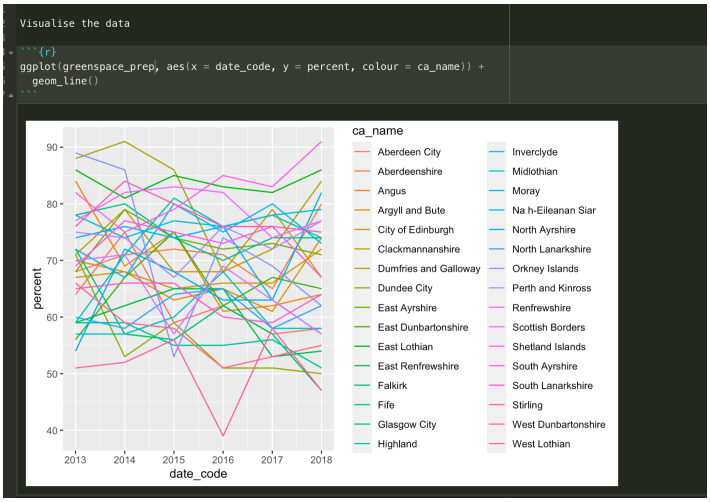
```r
greenspace_prep <- greenspace_all %>%
  rename(distance = distance_to_nearest_green_or_blue_space) %>%
  select(feature_code, date_code, measurement, value, distance) %>%
  left_join(la_names, by = c("feature_code" = "ca")) %>% #join on local authority names
  arrange(feature_code, date_code, distance) %>%
  rename(percent = value) %>%
  filter(feature_code != "S92000003") %>% #remove scotland overall rows
  filter(distance == "A 5 minute walk or less") %>% #only want 5 mins or less
  select(c(date_code, ca_name, percent))

greenspace_prep
```

date_code <dbl>	ca_name <chr>	percent <dbl>
2013	Clackmannanshire	71
2014	Clackmannanshire	79
2015	Clackmannanshire	68
2016	Clackmannanshire	68
2017	Clackmannanshire	79
2018	Clackmannanshire	67
2013	Dumfries and Galloway	88
2014	Dumfries and Galloway	91
2015	Dumfries and Galloway	86
2016	Dumfries and Galloway	68

2. Made a line chart to briefly visualise the data (EDA). From the initial view found would need to rule out a line chart showing all LAs as far too cluttered and hard to gain insights clearly.

Visualise the data

```r
ggplot(greenspace_prep, aes(x = date_code, y = percent, colour = ca_name)) +
  geom_line()
```



3. Pivoted the data so in wide format and calculated the difference between 2013 and 2018 and the sign of difference (positive/negative/zero)

Reshape the data so it's in wide format for Datawrapper

```{r}
two_time_points_wide <- greenspace_prep %>%
  filter(date_code == 2013 | date_code == 2018) %>%
  distinct(date_code, ca_name, percent) %>% #seems to be double rows so group by to remove
  select(ca_name, date_code, percent) %>%
  pivot_wider(names_from = date_code, values_from = percent) %>%
  mutate(change = round(`2018` - `2013`,2), sign = ifelse(change >0, "positive", ifelse(change == 0, "none", "negative")))
#calculate differences and sign of differences

two_time_points_wide
```

ca_name <chr>	2013 <dbl>	2018 <dbl>	change <dbl>	sign <chr>
Clackmannanshire	71	67	-4	negative
Dumfries and Galloway	88	84	-4	negative
East Ayrshire	68	71	3	positive
East Lothian	86	86	0	none
East Renfrewshire	59	54	-5	negative
Na h-Eileanan Siar	74	82	8	positive
Falkirk	72	47	-25	negative
Highland	59	74	15	positive
Inverclyde	57	58	1	positive
Midlothian	78	79	1	positive

4. Saved cleaned dataset to a csv file (to be uploaded into data visualisation tool)

Tools used

1. Pre-processing: R
2. Core visualisation: Datawrapper
3. Customisation of visualisation: Sketch
4. Presentation: R markdown/Shiny/shinyapps.io

Tool matrix attached.