## Methods

### Documents

Two CID reference documents, the CID Asset Information Guide (TFL, n.d. a) and the database schema (TFL, n.d. b) were obtained and assimilated to develop a description of the CID.

### Datasets

R version 4.0.3 was used for data importation and analysis (R Core Team, 2020). The CID data was accessed via the TFL cycling open data portal (TFL, 2017). As the data was available in a JSON file, the function read\_sf from the package sf was used to access and download the data into R as a simple features (sf) tibble (Pebesma, 2018). This function converts longitude and latitude into geometry. To simplify accessing and downloading the CID, an R package was created named CycleInfLnd (Tait and Lovelace, 2019).

Geographical boundaries (full extent of the realm), areas (ONS, 2020a) and estimated population (ONS, 2020b) for London Boroughs were obtained from the Office of National Statistics. Boundary data was imported as an sf data frame using st\_read (package sf).

Road lengths

Number of crossings

### Data manipulation

All spatial data was transformed into the British National Grid (ONS, 2020c) The tidyverse package was used for data cleaning, transformation and manipulation (Wickham et al., 2019). Each data item was examined to see if it was in the correct data format.

The CID variables were examined for completeness.

To ensure that all assets were correctly assigned to a London Borough, all assets were spatially cross-referenced with Borough boundaries. Where there was a mismatch or the CID Borough was NA, these were corrected as follows. Two ASL were assigned to the Borough location where the relevant ASL Traffic signal was located (1 NA, 1 reassigned). 29 Crossings were corrected (28 were NA, 1 reassigned) and a further one Crossing converted into two observations as the Borough Boundary intersected the Road and the Crossing equally. As Cycle lanes and Tracks can cross multiple Borough boundaries as well as cross into Local Authorities outside London, each Cycle Lane and Track was segmented by Borough boundaries and new assets created for each segment. For Cycle Lanes and Tracks, 72 observations were segmented and reassigned the correct Boroughs whilst a further 354 assets with no Borough assigned were segmented and then created into 621 assets. For Signals, there were no Nas, 2 observations were identified as being incorrectly labelled so were relabelled. Traffic calmning – no NAS for Borough, 138 reassigned.

Missing values were reviewed and managed appropriately. Where different datasets were to be joined, relabelling of the variables in common and their values was performed to enable successful joins. For example, consistency in London Borough names as either ‘Kensington and Chelsea’ or ‘Kensington & Chelsea’.

Spatial joins

### Data analysis

The number of infrastructure assets were calculated including by Borough. Borough was chosen as this is the local administrative unit and because whilst TfL builds a large amount of infrastructure, Borough’s also have their own Cycling Plans where they can build cycling infrastructure. For cycle parking, the number of cycle spaces as well as the number of cycle parking sites were calculated. Spatial lengths for cycle lanes/tracks and restricted routes were calculated using st\_length.

Mapview was used to display CID and ONS boundary data on background maps (e.g. street maps) to facilitate data visualisation (e.g. where the assets were in relation to Borough boundaries) and spatial analysis (Appelhans et al., 2019). Parallel coordinate plots were used to visualise multiple dimensions of the CID data (Inselberg, 1985; Wegman, 1990) using the Ggally extension to ggplot2 (Schloerke et al., 2021). To minimise the impact of the differences in scale in number of assets by Borough, the Borough were ranked.

### **Data quality and managing missing data**

Only one observation out of 233596 had an invalid survey date and only one duplicate observation was identified. The commonest missing data was photograph urls ranging from to 1.1% (Signage) to 6.7% (Restricted points). Five datasets had observations that did not have a Borough listed (see Table 3). Whilst the proportion of observations with missing Boroughs was very small, when the length of the observations from the Cycle lanes and tracks and Restricted routes datasets was calculated, it was seen that the missing proportion by length was much higher and thus needed addressing to enable fair comparison between amount of infrastructure by Borough.

**Table 3: Observations missing Borough data by asset type**

|  |  |  |  |
| --- | --- | --- | --- |
| **Asset** | **Number missing** | **% of total** | **% of length of assets** |
| Advanced stop lines | 1 | 0.03% | - |
| Crossings | 28 | 1.6% | - |
| Cycle lanes and tracks | 354 | 1.4% | 6.8% |
| Restricted routes | 18 | 1.3% | 8.2% |
| Signage | 2 | 0% | - |

### **Asset analysis by Borough**

The Boroughs were ranked according the total count or length of assets with one representing the Borough with the most and 34 representing the least. Although there are only 33 Boroughs in London, there was an additional ‘No Borough stated’ category as these charts were constructed before the approach to managing the missing Borough data was developed. As Boroughs vary by size and population, count and length were standardised to count/length (number or kilometre) per area (square kilometre) and per estimated head of population. Again Boroughs were ranked with 1 representing the highest density of infrastructure count/length by area or population.