**Data Science Capstone Project Report**

Holly Tibble

Notebook link:

<https://github.com/hollytibble/Coursera_Capstone/blob/master/Holly's%20Data%20Science%20Capstone%20Project.ipynb>

**Background**

Over the limited warm summer months in the Scottish city of Edinburgh, it is popular for locals to spend much of their day in the public parks – socialising, drinks and barbequing. Having recently procured a license for a mobile coffee shop, I am trying to identify the best spots in the city as a vendor.

As I will be operating out of a vehicle, I will not be able to provide customer toilets. As such, I would like to find spots with nearby public (or publicly accessible) toilets (maximum range 500 metres). My license allows me to park my cart on the edge of public parks, but I prefer to be nearby to park play areas as I think my target audience is probably parents and families, rather than young adults who are perhaps more likely to be consuming alcohol.

**Data**

The data I will be using to find suggested locations for my coffee cart will be the locations of public play areas, public toilets, and other local shops.

The play areas data was extracted from data.edinburghopendata.info, and was most recently updated in August 2017. The data contains the following variables:

* Site: the name of the play area (string)
* Play facilities: a description of the play area (string)
* Address: the street address of the play area (string)
* Postcode: the postcode of the play area (string)
* Telephone: the telephone number of the local council office responsible for the management of the play area (numeric)
* Location: the coordinates of the play area (numeric tuple)

The public toilets data is also from data.edinburghopendata.info, and was updated in July 2017. The data contains the following variables:

* Toilet: the name of the toilet (string)
* Type of facility: either public toilets, or the type of establishment if publicly accessible, such as ‘Library’ (string)
* Refurbished in 2013: a note of whether or not this toilet (public only) was refurbished in 2013 (string)
* Charge: the cost of using the toilet, where applicable (string)
* Refurbishment status: unknown, missing for all entries
* Facilities: description of the available facilities, such as baby changing and disabled access toilets (string)
* Opening times: the days and times for which the toilet is available for use (string)
* Telephone: the telephone number of the council office or establishment responsible for the upkeep of the facilities (numeric)
* Email: an email address for the council office or establishment responsible for the upkeep of the facilities (string)
* Website: a website for the establishment responsible for the upkeep of the facilities (string)
* Address: the street address of the toilet (string)
* Location: the coordinates of the toilet (numeric tuple)

Finally, foursquare data was used to identify nearby establishments, within a radius of 1km of the search query coordinates. A limit of 50 responses was applied, and the name, coordinates, and category of each venue was extracted from the json file returned by the query.

**Methodology**

First, I conducted a many-to-many merge on the play areas (henceforth ‘sites’) and public toilets, such that all possible pairs were generated. In this merged data I calculated the distance between the site and the toilet, using geodesic distance implemented using the *geopy* package. Geodesic distance, or great-circle distance, is defined as the distance between any two points on a sphere. While the earth is not quite perfectly spherical, geodesics on an ellipsoid are more computationally intensive, and rarely improve accuracy by more than 0.5% [[1]](#footnote-1). On such a small scale as we are using, the difference is negligible. Toilets over 500 metres from the site were excluded, and the unique sites with public toilets within range were identified and retained.

Venues of any type within a 1km range of each retained site were identified from the foursquare database. Coffee shops and cafés were separated, and the number within radius of each site was counted. Next, other venues were assessed by converting the category description into a binary variables, using one-hot encoding. For each site, the mean of each binary category variable was calculated for the neighbouring venues, to estimate the relative frequency of the venue type in the vicinity.

My aim was to identify areas in which a coffee shop would be likely to be successful, but in which there were fewer than expected currently operating establishments. To identify these areas, I created a random forest regression model (from *sklearn*) to estimate the number of coffee shops within range of each site, using their other neighbouring venues as predictor variables. I then ran the data back through the model to create the model’s predictions of the number of coffee shops, and identified areas with fewer coffee shops than expected, defined as having a regression output of at least 0.5 coffee shops higher than the observed value.

I then visualised this data by plotting the candidate sites, their nearby public toilets, and their nearby rival coffee shops on a colour coded map, using the *folium* package.

**Results**

155 public play areas and 98 public or publicly accessible toilets were identified in the Edinburgh open datasets. The coordinates were not available for 1 of the toilets, and 1 of the play areas, resulting in their exclusion. The merged dataset of toilets and sites contained 14,938 candidate pairs on which to measure the distance. 14,777 candidate pairs were excluded as the distance between them exceeded 500 metres, leaving 83 unique sites with a mean distance to the closest public toilet of 290 metres (figure 1).

A picture containing drawing

Description automatically generated

Figure 1: Histogram of the kilometres between the nearest toilet and each site (maximum 0.5km)

Out of 1773 venues local to our 83 sites, 193 were coffee shops (11%). 47 of the sites had a coffee shop within range (57%, figure 2).

A picture containing clock

Description automatically generated

Figure 2: Histogram of the number of coffee shops within 1km of sites which had public toilets within range

Of the 1583 non-coffee venues, the five most common categories (in descending order of frequency) were pubs, parks, grocery stores, supermarkets and bars.

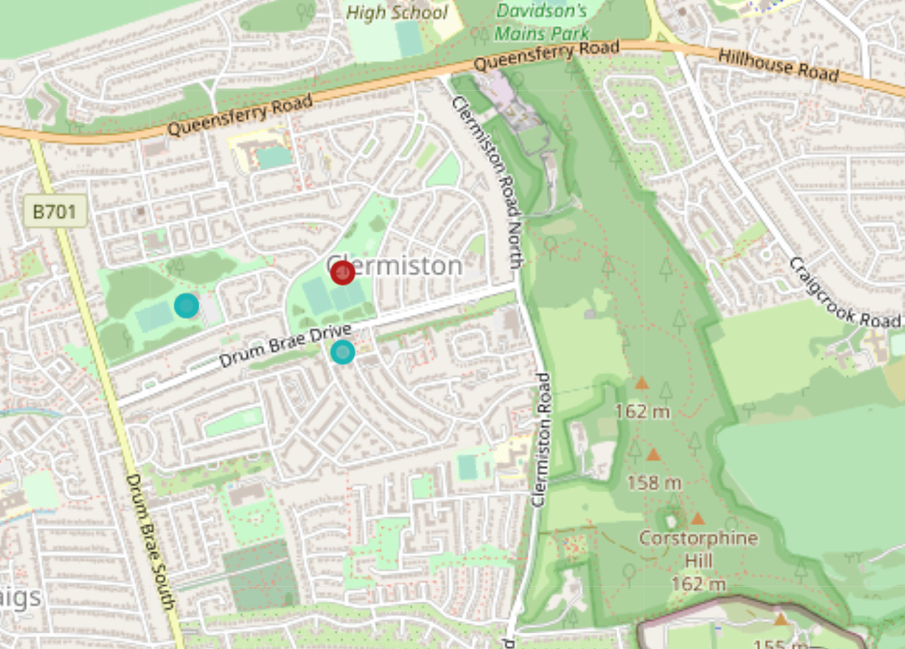
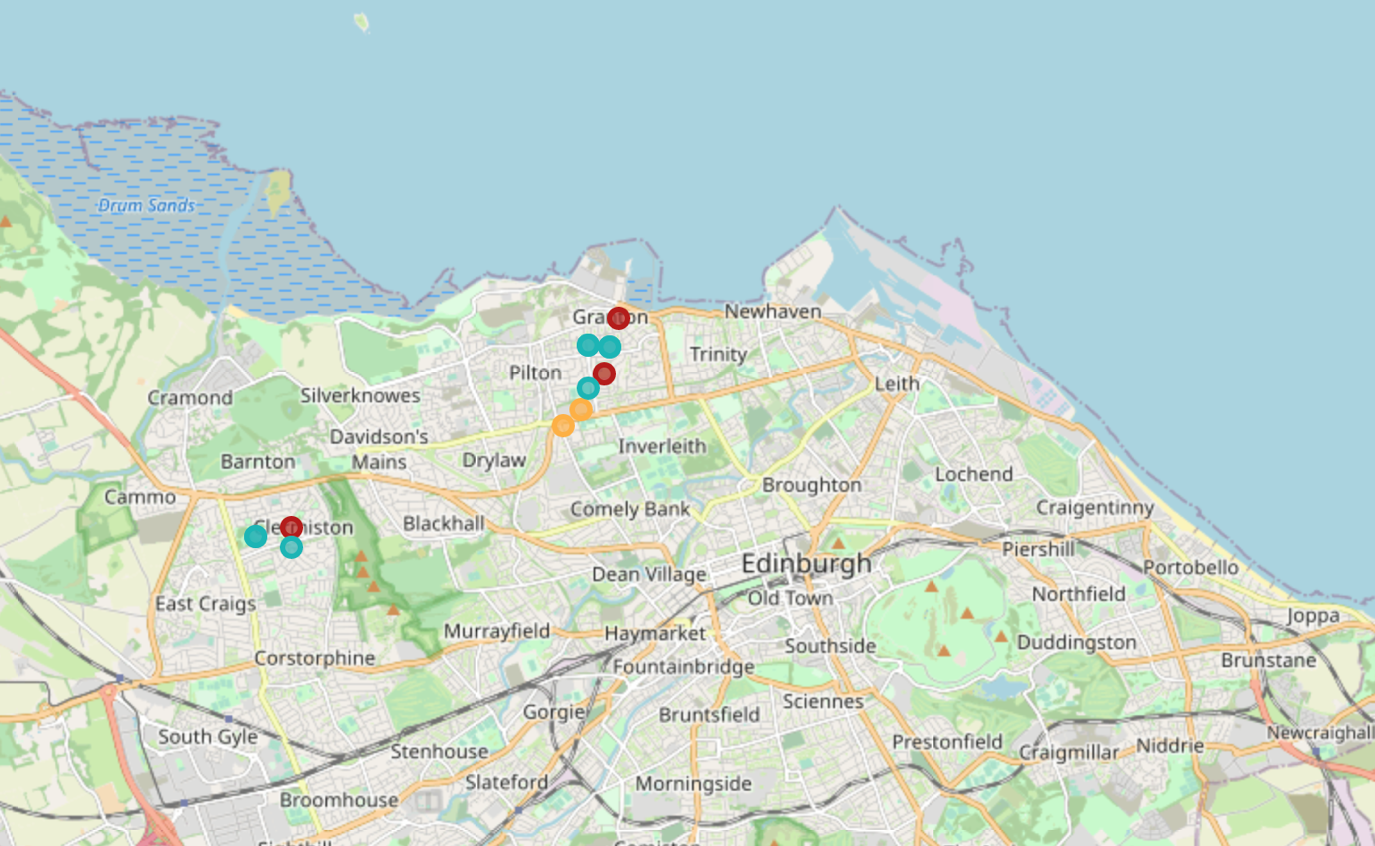
The random forest was run on all of the samples, with no divide for training and testing, as the model’s predictive performance is not the primary aim of the analysis. Regardless, the model had a mean squared error of 0.34 and an R2 of 0.96 in the training data. Figure 3 shows the predicted and observed number of coffee shops in the vicinity of each site. We can see a number of points below the reference line, indicating that they had fewer coffee shops than the model would predict.

A picture containing sitting, table, different, man

Description automatically generated

Figure 3: Scatterplot of predicted and observed coffee shops in the vicinity of each site which had public toilets within range

Three sites were identified which had more than 0.5 fewer than predicted coffee shops in the vicinity; Clermiston Park, East Pilton Park, and Granton Crescent (Figure 4).



Legend

Site

Toilet

Coffee

Figure 4: Locations of identified candidate sites, their nearby public toilets, and potential rival coffee establishments

**Discussion**

In this project, I have used random forests to identify potential spots for a mobile coffee cart, in public park play areas, with nearby public toilets, and with fewer than expected neighbouring coffee shops, as predicted by their other neighbouring venues.

A strength of this study was that it made use of well-maintained datasets, and thus identified all relevant sites in the city matching the criteria. A limitation is that the foot-traffic and reviews of the neighbouring coffee shops were not included, which could have further informed demand in a given area.

**Conclusion**

Three spots were identified as candidate locations for a mobile coffee cart in the city of Edinburgh, Scotland. These sites will be investigated further, and trialled in order to establish their suitability as regular spots.

1. For more information, refer to Admiralty Manual of Navigation, Volume 1, The Stationery Office, 1987, p. 10, The errors introduced by assuming a spherical Earth based on the international nautical mile are not more than 0.5% for latitude, 0.2% for longitude. [↑](#footnote-ref-1)