**Mapping how changes in well-being indicators have been reflected in real estate values in London**

MSc Project Report

Ben Candy, MSc Data Science

Supervisor: Alessandro Provetti

Department of Computer Science and Information Systems, Birkbeck College, University of London 2018

This report is substantially the result of my own work, expressed in my own words, except where explicitly indicated in the text. I give my permission for it to be submitted to the JISC Plagiarism Detection Service.

The report may be freely copied and distributed provided the source is explicitly acknowledged

Table of Contents

[1. Abstract 3](#_Toc523860857)

[2. Introduction 4](#_Toc523860858)

[3. Project Trailer 5](#_Toc523860859)

[4. Design 6](#_Toc523860860)

[5. Implementation 7](#_Toc523860861)

[5.1. Collection and cleaning of data 7](#_Toc523860862)

[5.2. Creating the well-being domain scores 7](#_Toc523860863)

[5.3. Modelling the data as a regression problem 7](#_Toc523860864)

[5.4. Modelling the data as a classification problem 7](#_Toc523860865)

[5.5. Creating a GeoJSON file for the best model 7](#_Toc523860866)

[5.6. Displaying the project results via a Javascript interactive map 7](#_Toc523860867)

[6. Testing 8](#_Toc523860868)

[7. Evaluation 9](#_Toc523860869)

[7.1. Overall success of the project 9](#_Toc523860870)

[7.2. Data availability 9](#_Toc523860871)

[7.3. Potential extensions of the project 9](#_Toc523860872)

[References 10](#_Toc523860873)

[Appendix A: User Guide 11](#_Toc523860874)

[Appendix B: Additional Tables 12](#_Toc523860875)

[Appendix C: Project Code 12](#_Toc523860876)

# Abstract

# 2. Introduction

# 3. Project Trailer

# 4. Design

# 5. Implementation

## 5.1. Collection and cleaning of data

## 5.2. Creating the well-being domain scores

## 5.3. Modelling the data as a regression problem

## 5.4. Modelling the data as a classification problem

## 5.5. Creating a GeoJSON file for the best model

## 5.6. Displaying the project results via a Javascript interactive map

# 6. Testing

# 7. Evaluation

## 7.1. Overall success of the project

## 7.2. Data availability

In part thanks to the Greater London Authority’s Datastore, a wealth of information is available at London Ward level. Combining this with information from the Foursquare API and Department for Education schools data provided a significant body of information in which to model London as a multiple dataset.

Data availability was such that datasets for each of the indicators feeding into the six domains were available and it was not necessary to use substitute indicators that differed significantly from those originally planned.

One data availability issue encountered was the fact that in many cases the various datasets related to different years, often linked to census years and administrative or electoral changes.

Wherever possible the most recent data has been used to synchronise as closely as possible with the 2017 data used for the median house price data. Emission data is from 2011 as this is the most recently published at ward level, the London Air Quality daily feed run by King’s College does not have enough coverage to sufficiently differentiate over 600 wards. Ideally this data would have been from the same year as the house price information.

Ward boundaries were re-drawn in three London Boroughs in 2014 meaning that a function had to be written that, for data pre-dating 2014, would map old ward codes to the new ward that contained the largest section of the newly defined area. Newer data would also have been helpful here but the commonality of area between the old and new codes in the mapping should ensure that the data is representative of the new area.

The other limiting factors in data availability were the API limits and costs, whilst I was able to obtain the necessary Foursquare venue and venue location information, venue ratings and check-ins were subject to stringent daily limits which meant that obtaining this information would have taken weeks of daily iterations or significant costs neither of which were feasible for this project. I investigated the possibility of using Google Places API but the new pricing model introduced this year also made this unfeasible.

## 7.3. Potential extensions of the project

An extension of this project that would be an interesting addition to the results already displayed would be to look at whether changes in the well-being domains were related to changes in house prices. This could potentially provide insight on whether increasing house prices fuel rises in well-being indicators or whether improvements in local well-being increase demand for houses in a particular area.

To achieve this, more data would need to be found that could represent all of the well-being indicators for regular intervals over a significant period of time. Given that some of the datasets are only produced every few years, this may require looking at some innovative solutions of data collection with crowdsourcing and image analysis possible avenues of investigation.

If being run as a commercial application, a paying subscription to the Foursquare and/or Google API would allow further analysis around assessing venue popularity to feed into the Community Vitality and Participation domain.

# References

# Appendix A: User Guide

# Appendix B: Additional Tables

# Appendix C: Project Code