BIRKBECK, UNIVERSITY OF LONDON

MASTER PROJECT REPORT

Mapping London House Prices through Well-being

Author:
Ben CANDY

Supervisor:
Dr. Alessandro PROVETTI

A project report submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Science and Information Systems

September 12, 2018

Declaration of Authorship

I, Ben CANDY, confirm that:

- 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.

Signed:

Date:

“”

BIRKBECK, UNIVERSITY OF LONDON

Abstract

School of Business, Economics and Informatics
Department of Computer Science and Information Systems

Master of Science

Mapping London House Prices through Well-being

by Ben CANDY

I have always been interested in the spatial analysis of human behaviour. For this project I am going to put together a software project which will allow me to visualise this in a data-driven way. In my literature review I discovered that there are hardly any quantifiable metrics for human happiness in a geographical context. Therefore, I look forward to experimenting with new measures, trying to leverage the wealth of data we now have on "London Living" and Data Science methods such as Principal component analysis and dimensionality reduction.

Acknowledgements

I would like to thank my project supervisor Professor Alessandro Provetti for his support and guidance throughout this project...

Contents

List of Figures

List of Tables

List of Abbreviations

LAH	**L**ist **A**bbreviations **H**ere
WSF	**W**hat (it) **S**tands **F**or

Physical Constants

Speed of Light $c_0 = 2.997\,924\,58 \times 10^8\,\mathrm{m\,s^{-1}}$ (exact)

List of Symbols

a	distance	m
P	power	W ($\mathrm{J\,s^{-1}}$)
ω	angular frequency	rad

Chapter 1

Introduction

1.1 Motivations and problem statement

In recent years, the concept of well-being has been the subject of increasing interests for governments, councils, policy makers and researchers, though research on the subject has been happening since the 1970s. An interesting recent development is a report into well-being in Danish cities by (OECD, 2016) which highlights how well-being tools have become an important tool to identify the needs of citizens and the domains where demand for progress is greatest.

This idea of well-being is often defined using a set of indicators relating to a particular place. There is no universally agreed definition of exactly which factors are the best measures of well-being, however, various indexes and studies show a great deal of commonality on the types of measures that are important to the well-being of citizens.

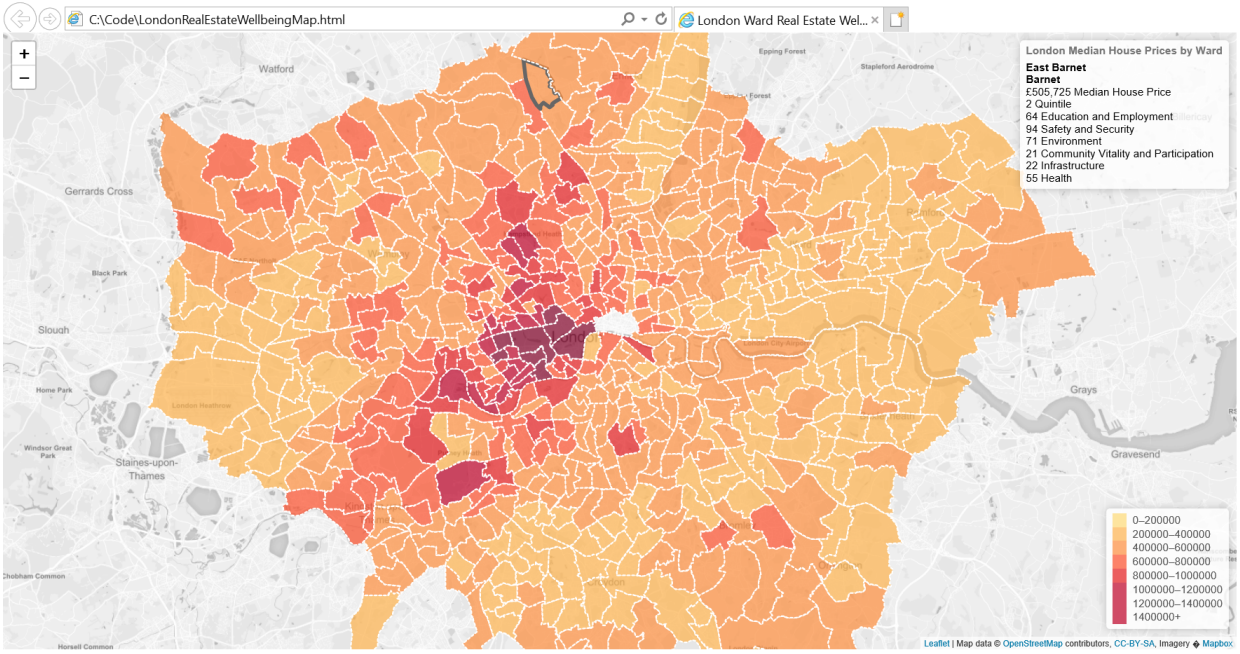
What is less clear is whether, and how, these indicators, and the changes in them over time, affect the real estate values in a specific city. For example, The Greater London Authority created a set of well-being indicators and used them to create a map of well-being by London Ward in 2014 (Greater London Authority, 2014), however, this was not linked back to real estate values. In one study looking at linkages between house prices and mental wellbeing, (Ratcliffe, 2015), points out that when people can move freely, each person will maximize wellbeing by moving to areas that best satisfy their preferences; this results in zero correlation between area characteristics (or house prices) and wellbeing. But if people think they are paying too much for poor quality or too little for high quality, we would observe a positive relationship between area characteristics and wellbeing. The study found a positive correlation between house prices and mental wellbeing.

Another hypothesis found in the Economics literature is that real estate values are the drivers of the well-being indicators, happier people are better at generating wealth (Lyubomirsky, King, and Diener, 2005), which leaves open the possibility that areas with high well-being provide a flow of people migrating to places with higher real estate values. From the point of view of Data Science it would be interesting to consider the real estate values in bandings such as quintiles as well-being indicators may have a very different relationship with the top 20 percent of areas by real estate value than with areas in the lower four quintiles.

Work on urban scaling (Bettencourt, Lobo, Strumsky, and West, 2010) suggests that different types of well-being indicators may have different relationships with real estate values due to the non-linear nature of agglomeration. Bettencourt et al. argues that city indicators are governed by power laws rather than linear per-capita indicators. Social factors tend to be superlinear whereas material infrastructure tends to be sublinear at best. There may therefore be some level of divergence between different well-being factors. This project focuses on the city of London and

tries to determine which well-being indicators show a relationship with real-estate values. The findings will be displayed in the form of an interactive map.

1.2 Project trailer



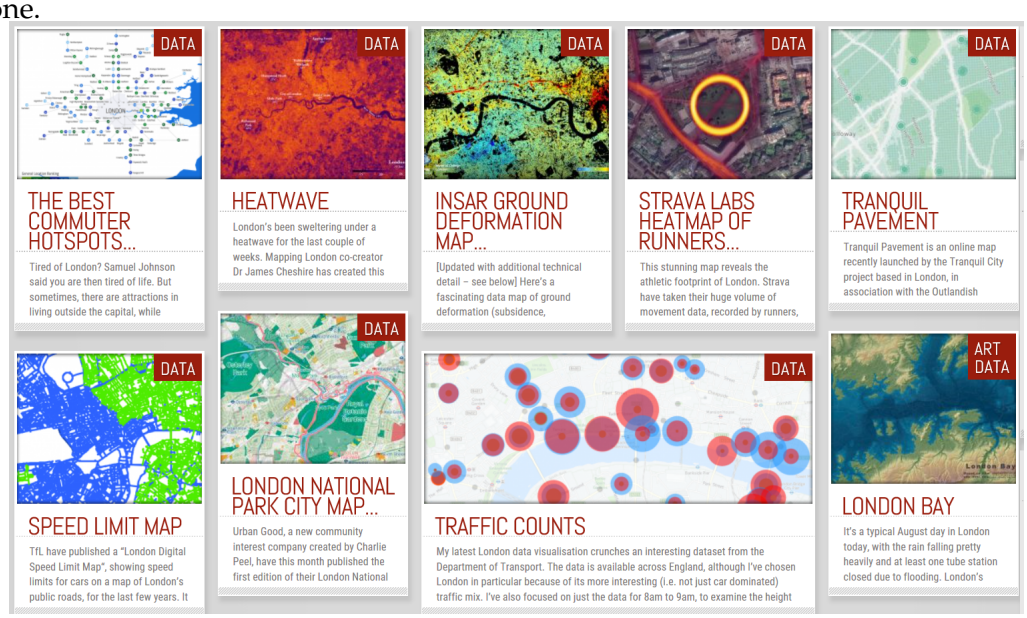
Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor

Chapter 2

State of the art

2.1 Geospatial analysis in Data Science

Maps are an increasingly popular way to visualise data where some form of geographic aspect is an important element of the analysis. The website London Mapping () is a good example of the large amount of data analysis and visualisation taking place through the creation of various types of map for the city of London alone.



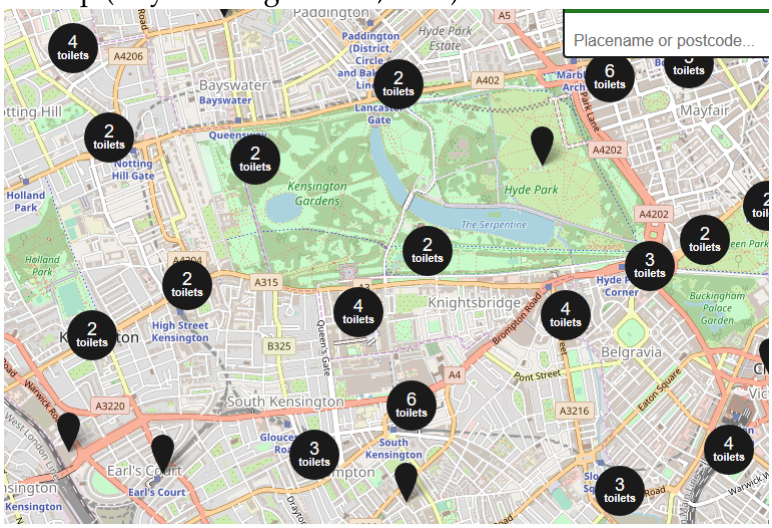
There are several reasons why maps a a good choice for data visualisation which include: They provide a real-world context for the data, helping the audience to understand the analysis They allow users to compare data over different areas or regions at a glance

Data Science techniques play an important role in the creation of this kind of visualisation. Most map visualisations represent locations through coordinate reference systems which represent locations. These locations can be specific locations represented by pairs of coordinates, most often (latitude, longitude), or areas which would be represented by polygons consisting of multiple coordinate pairs. Data Science techniques are required to design functions and algorithms to map, aggregate or disaggregate data between different types of geometry and different coordinate reference systems.

2.1.1 Point Maps

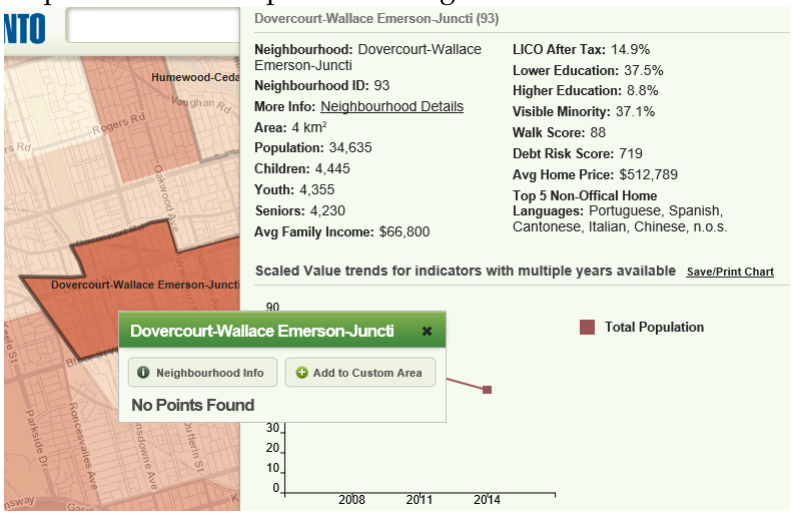
One mapping technique often used in interactive visualisations would be a point map where specific instances of something are plotted as points in the geo-location

in which they occur. This form of mapping would use a coordinate pair, often latitude and longditude but potentially an Ordnance Survey grid reference or other coordinate reference system. An example of this would be the Great British Public Toilet Map (Royal College of Art, 2014)



2.1.2 Choropleth Maps

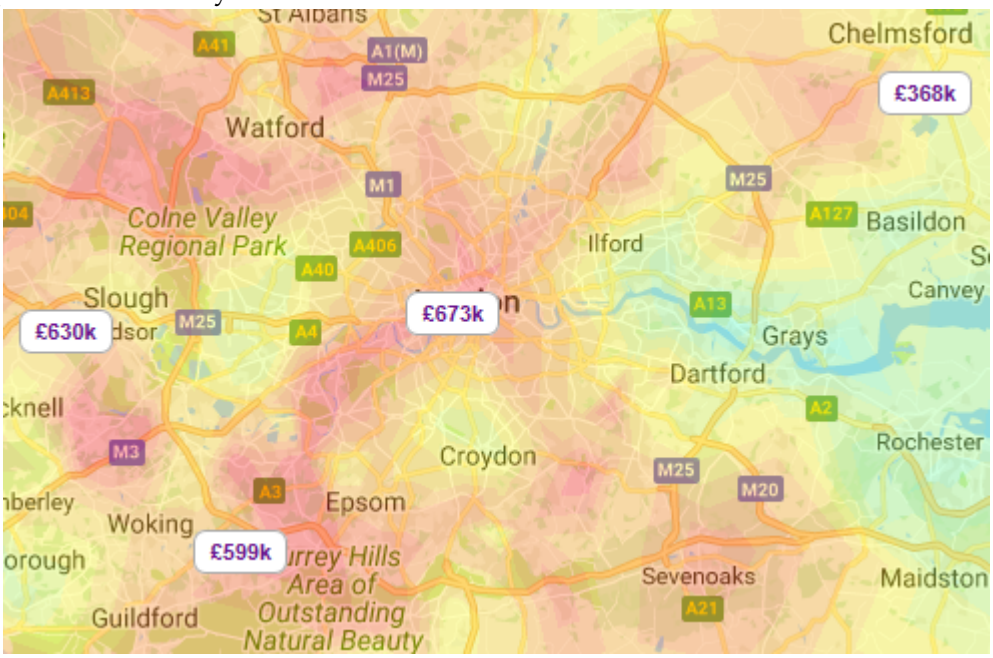
The city of Toronto is a good example of where an informative interactive well-being map has been created (City of Toronto, 2014). On this map a user can click on an area and is presented with a panel showing the values for a number of measures.



The Toronto map uses a thematic mapping technique known as a Choropleth map, using shape files where polygons represent an area of space, in this instance neighbourhoods, with the colour of the polygon related to the value of the area in question. This map is an excellent design example for this project as the geometries used, in this case neighbourhoods of Toronto, are of a similar geospatial representation to London Wards. The project map will be based on median house prices and hovering over an area will display the values of the well-being indicators and the projected values of the model.

2.1.3 Heat Maps

Real estate values have also been an area of interest for mapping. In recent times these visual displays of real estate value have often been created by the commercial sector, for example, UK property price maps created by the estate agent Zoopla (Zoopla, 2011). This is an example of a heat map. Heat maps share similarities with Choropleth maps in that they assign a colour to an area based on the magnitude of a particular value. Where these two types of map differ is that heat maps assign colour to a cluster of connected points whereas for Choropleth maps, the colouration is based on a value for a polygon representing a given area, usually a administrative or political boundary.



2.1.4 Network Maps

An interesting alternative approach has been taken by (Quercia, Aiello, and Schifanello, Smelly Maps, 2016) who have created a series of interactive maps using crowdsourced data and image processing techniques. The locations used in the maps created by Quercia et al. use specific locations represented by images and sound recordings rather than areas of space. Here, rather than using traditional spatial analysis techniques, the maps have been created using graphs which represent London as a network. The location of the images/sounds are the nodes and the routes between them form the vertices.


TRANSPORT
NATURE
HUMAN
MUSIC
BUILDING

Chapter 3

Data Sources

3.1 Flat files

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

3.1.1 London datastore

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

3.1.2 Transport for London

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque. Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elit posuere quam, quis adipiscing urna leo nec orci. Sed nec nulla auctor odio aliquet consequat. Ut nec nulla in ante ullamcorper aliquam at sed dolor. Phasellus fermentum magna in augue gravida cursus. Cras sed pretium lorem. Pellentesque eget ornare odio. Proin accumsan, massa viverra cursus pharetra, ipsum nisi lobortis velit, a malesuada dolor lorem eu neque.

3.1.3 gov.uk

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque. Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elit posuere quam, quis adipiscing urna leo nec orci. Sed nec nulla auctor odio aliquet consequat. Ut nec nulla in ante ullamcorper aliquam at sed dolor. Phasellus fermentum magna in augue gravida cursus. Cras sed pretium lorem. Pellentesque eget ornare odio. Proin accumsan, massa viverra cursus pharetra, ipsum nisi lobortis velit, a malesuada dolor lorem eu neque.

3.2 APIs

A subset of London data is also available via various APIs. For transport information Transport for London have extensive APIs available, unfortunately none of these were able to provide any of the datasets required for the project. The use of APIs became focussed on collecting information on venues to obtain the data for the 'community vitality' and 'access to cultural spaces' datasets. For this data Foursquare was the primary candidate with venue information available with no cost implications.

3.2.1 Foursquare

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

3.2.2 Google Maps

The places API from Google Maps was queried and code written to extract venue data. However, after obtaining some basic search results it was decided that Google's pricing policy and rate limits made this data source unfeasible for the scope of the project.

3.3 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.

3.3.1 Timeliness

Some of the 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. For example, 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.

3.3.2 Boundary changes

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.

3.3.3 API rate limits

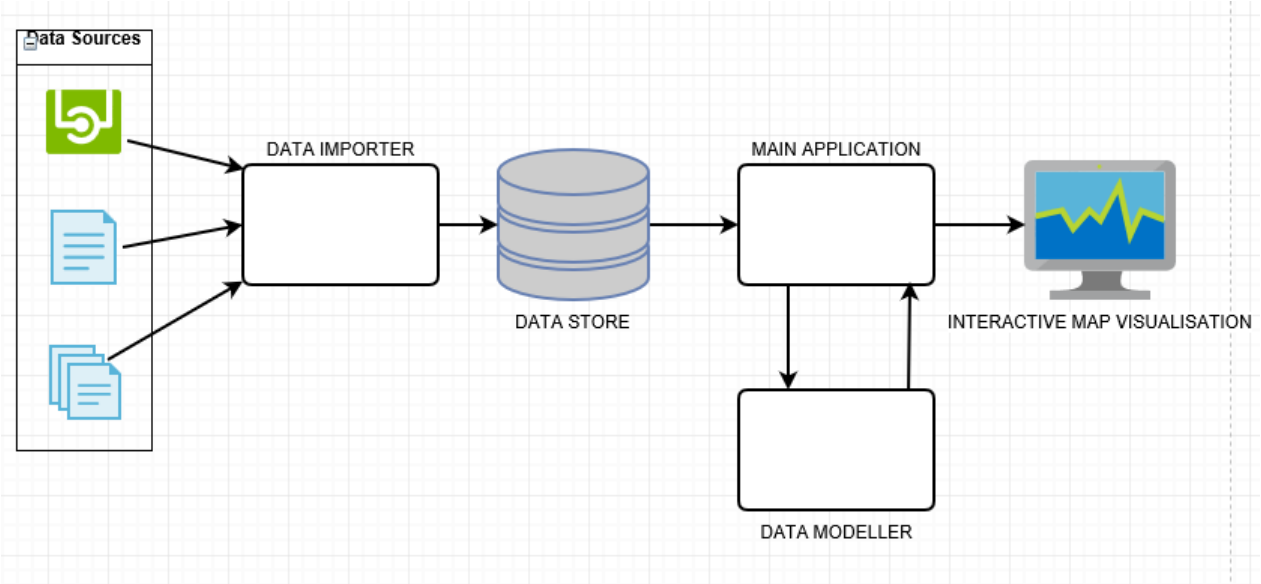
API limits and costs imposed limits on the depth of information that could be obtained from these sources. Whilst I was able to obtain the necessary Foursquare venue and venue location information, venue ratings and check-ins were subject to stringent daily limits which meant that obtaining this information 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. In a commercial setting it may be possible to increase the amount of data that can be obtained from some of the APIs.

Chapter 4

Design

4.1 Architecture of the specification

4.1.1 Archictecture diagram



4.1.2 Architecture choices

For this project two main technologies are being used, Python and Javascript. The majority of the software code has been written in Python, specifically for the data import processes, for the main application which runs the analysis and geocoding and the data modelling component. Python was chosen for these components for a number of factors. Firstly for the data importer, Python has excellent capabilities for data processing allowing the data to be collected and cleaned leveraging the Pandas module. For the main application, Python's geocoding and spatial capabilites were important factors in the choice. The ability to import shape files and reproject coordinate systems leveraging Geopandas and Shapely allowed more flexibility in which datasets were suitable for the project and the option to export to GeoJSON format facilitated the data visualisation. For the visualisation stages Javascript embedded in a html file was chosen. For this final component of the project Javascript was preferred to Python due to the interactive element required on the map. With the aid of the Leaflet.js package in Javascript I was able to make use of more map features to present any end users with a more complete user experience.

4.2 Components

The software architecture design for the project has been created with the aim of being able to isolate individual elements in the interest of performance. The system is not dependent on processing a live data feed so it is important that the component that imports the data does not have to run everytime a user would wish to access the interactive map visualisation. The design means that each component can be modified and without affecting the other parts.

4.2.1 Data inputs

The data input layer is the base layer of the system. This consists of several different types of input including API feeds from the Foursquare platform, csv files collected directly from urls and other data files which have been pre-collected and loaded into the data store. This layer is the most likely to change moving forward as new and updated data sources are published or become available in alternative formats.

4.2.2 Data importer

The data importer software is written in Python and is designed as a series of functions. Each function imports one of the required datasets. Setting these up as seperate functions gives the option to run imports individually. This is an important requirement of the system as certain imports take a considerable amount of time. The iterative nature of the API imports combined with daily rate limits mean that it is only feasible to run the API imports from Foursquare on at most a daily basis. As the system is not reliant on a live feed, the data import software could be run as an overnight batch process.

4.2.3 Data store

The data store is central to the architecture to safeguard performance levels. As the data imports are best suited to a batch process, the cleaned data files should be stored in the data store for the main application software to run efficiently. The data store is also used for any pre-collected data which cannot be collected directly from a url or API. This imports the data for the 18 base indicators which create the 6 well-being domains.

4.2.4 Main application

The main application software performs a variety of tasks and is written in Python. The first task that the software performs is to take the data that has been imported for the 18 base indicators from the data store. The main application then performs any data manipulation and applies a set of geocoding and spatial functions to standardise the data into a format where everything is aggregated for London Ward geometry. Having standardised, the software then creates the 6 well-being domains using functions drawing on Principal Component Analysis and mean calculation functions which also normalise the data to ensure that each score is between 0 and 100. These 6 domains are then combined with data showing median house prices and the related quintiles and passed through the regression and classification models chosen through the data modeller component. The final task performed by this

component is to create a geoJSON file including the well-being domain scores, median house price information, spatial data and predicted regression and classification values. This file contains the data used in the visualisation.

4.2.5 Data modeller

4.2.6 Interactive map

The final component of the software architecture is the interactive map which is built in javascript and html. This component builds a framework for the map using tiles from the MapBox service. A javascript function then takes the geoJSON file and add this as a layer to the map. This part of the architecture can be seen as the user front end as the visualisation allows users to hover over each ward to see the scores, actuals and predictions along with features such as a zoom facility.

Chapter 5

Implementation

5.1 Data collection and cleaning

The key foundation of the project was being able to collect the data that would allow the creation of the well-being domain scores. Collecting the required data and shaping it into the formats that were needed was, in places, one of the most challenging aspects of the project and the one that required the largest proportion of time. This part of the project presented several challenges that needed to be overcome.

5.1.1 Collection of data files

For ward level data, the Greater London Authority's datastore is a primary data source. A significant proportion of the data for the creating of the well-being indicators could be accessed via this source.

5.1.2 Collection from APIs

The API that was key to the success of the project was that of social media platform Foursquare. To represent community vitality, the aim was to create a measure for the restaurants and bars in each ward. Foursquare is a platform based around venue information so holds the information required for this measure. Foursquare also holds information on cultural venues. Therefore the measure of 'Access to Cultural Space' could also be obtained from this platform. To use the API to obtain venue information you are required to search for venues within a specified radius of a certain point, given in latitude and longditude. The challenge with Foursquare was the the API rate limits which limits each search to a maximum of 49 searches. There are also daily and monthly rate limits to stay within. To obtain results of all bars, restaurants and cultural venues accross Greater London with these limits an iterative algorithm had to be created which could travel accross London by latitude and longditude, taking in small enough areas so as to not exceed 49 results a time. To achieve this, I created a latitude and longditude bounding box around London and created a nested loop iteration of 50 points of latitude and 50 points of longditude. The areas created by this had some overlap so as not to have missing areas of the capital. With a counter to check if the search limit of 49 venues was exceeded, this search algorithm put the list of venues into a pandas dataframe and de-duplicated for where areas had ovelapped.

```
def importVenue(category, filename):
    df_Foursquare = pd.DataFrame()
    FrameList = []
    limit_reached = 0
    for i in range(1,50):
        for j in range(15,750,15):
            client_id = "XO1QQIQ02JE0EKQXZ4HBOYPHW2DSWJH5EIBN2AOZG2NRZVPK"
            client_secret = "GJGSIAZR4WXLYOHW4VP1JHTAMLH23EZXCWZDVPCZTYY2RQ4V"
            lat = (51.2 + (i/100.0))
            long = (-0.25 + (j/1000.0))
            if category == 'cultural':
                category_id = '4d4b7104d754a06370d81259' # cultural space
            elif category == 'restaurant':
                category_id = '4d4b7105d754a06374d81259' # food
            elif category == 'bar':
                category_id = '4d4b7105d754a06376d81259' # drink
            distance = 450
            requested_keys = ["categories","id","location","name"]
            url = "https://api.foursquare.com/v2/venues/search?ll=%s,%s&intent=browse&radius=%s&categoryId=%s&limit=49&client_id=%s&client_
            % (lat, long, distance, category_id, client_id, client_secret, time.strftime("%Y%m%d"))
            resp = requests.get(url)
            dataResp = resp.json()
            if dataResp["response"]['venues'] != []:
                data = pd.DataFrame(dataResp["response"]['venues'])[requested_keys]
                df_FoursquareIteration = pd.DataFrame(data)
                if len(df_FoursquareIteration) == 49:
                    limit_reached = limit_reached + 1
                df_FoursquareIteration["categories"] = df_FoursquareIteration["categories"].apply(lambda x: dict(x[0])['name'])
                df_FoursquareIteration["lat"] = df_FoursquareIteration["location"].apply(lambda x: dict(x)["lat"])
                df_FoursquareIteration["long"] = df_FoursquareIteration["location"].apply(lambda x: dict(x)["lng"])
                FrameList.append(df_FoursquareIteration)
    df_Foursquare = pd.concat(FrameList)
    df_Foursquare.drop_duplicates(subset=['id'],keep=False)
    if limit_reached > 0:
        print('Limit Reached')
        print(limit_reached)
    columns = ['name','categories','lat','long']
    df_culture = pd.DataFrame(df_Foursquare, columns = columns)
    df_culture.to_csv(filename, index=False, encoding='utf-8')
```

Because of the rate limits, running the algorithm for such a large area could only be done on a daily basis.

5.2 Well-being domain scores

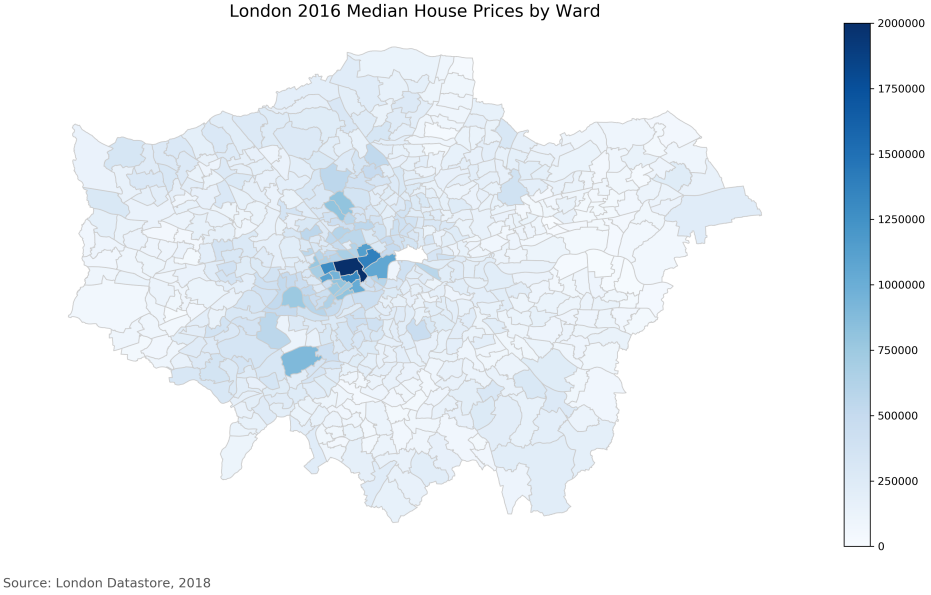
The motivation for creating composite domain scores 6 different measures of well-being was two-fold. Firstly creating a score for each domain over a pre-defined scale would allow users of the front-end interactive map compare scores accross wards and give those scores some contextual meaning and significance. With 6 domains rather than the original 18 datasets, the amount of information presented to the user would not be too large to prevent a non-technical audience from interpreting the information on the map. The second reason for creating the domain indicators was to help to try and create an interpretable model rather than one requiring significant dimension reduction. Whilst some less interpretable models will be tested to validate the quality of the quintile classification model, ideally the project and the user front end should allow us to draw some conclusions about any link between well-being and median house prices in London.

For each domain three appropriate datasets were chosen and then reduced to give a single score for each ward. The 3 datasets for each domain were standardised using a function in Python and those which related to a negative indicator such as emissions or crime were mutliplied by a value of -1. Using this technique of standardisation gave equal weighting to each of the three indicators. Because of the equal weighting given to each indicator, these can be seen as substitutable within the context of the domain which is a important factor in the choice of technique used to reduce the indicators to a score. With indicators that can be assumed to be substitutable suitable techniques would include Principal Component Analysis (PCA) and an arithmetic mean ?INSERT REFERENCE TO PAPER BY ITALIAN ACADEMICS? For the 6 domains, a mixture of the two techniques have been used. For those domains where the 3 indicators have a relatively high correlation ?ABOVE WHAT?

the first Principal Component has been used to try and extract the largest amount of variance within the three indicators. Where the indicators for a domain have low levels of correlation for at least one of the three datasets, the arithmetic mean has been used. This approach gives us three domains which use PCA and three which use the mean. ?INSERT COV MATRICES AND TECHNIQUES FOR EACH DOMAIN? This approach tries to maximise the information contained in all of the datasets. To create an interpretable and comparable score for each domain the resulting single scores have been normalised and then rescaled to be between 0 and 100, this is the score that is presented in the user front end.

5.3 London house prices as a regression problem

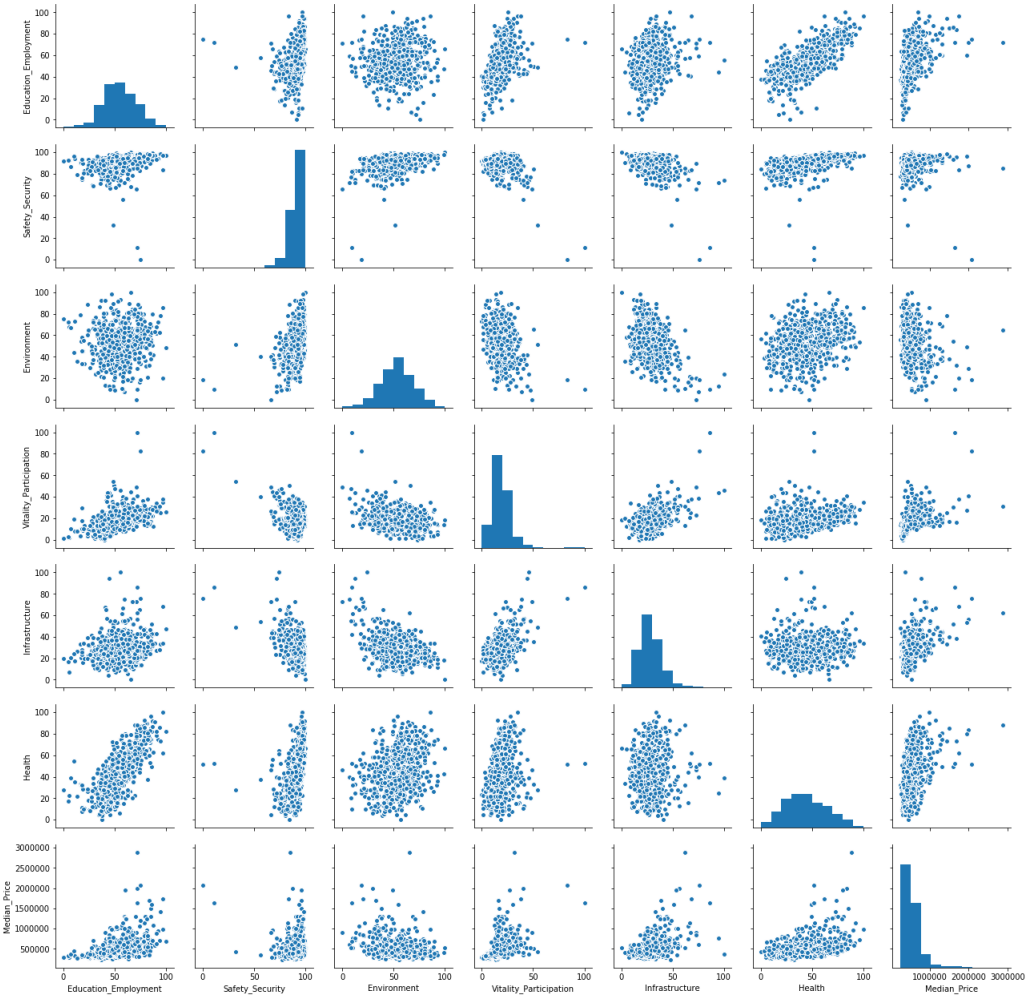
London is known for having very high real estate values but we can see from the map below that there are specific wards, particularly in the western central area, where median house prices are significantly above the Greater London area as a whole.



The choropleth map would suggest a high number of wards fall at the lower end of the range of median house prices with a small number with significantly high prices. From the pairwise plot below, the shape of the histogram of median house prices by ward would suggest that a power law may best describe the distribution.

? FIT POWERLAW TO MEDIAN HOUSE PRICES ?

For the domain scores, the histograms show most of the six as being at least close to a normal distribution. This is less clear in the case of the Safety and Security and Community Vitality and Participation domains with Safety and Security showing a large number of high values and Community Vitality and Participation showing many low scores. This can be attributed to a small number of wards where crime is high and that wards in central London have much greater access to both cultural and social spaces as indicated on the maps below. In both cases, a power law may best describe these distributions.



This is less clear in the case of the Safety and Security and Community Vitality and Participation domains with Safety and Security showing a large number of high values and Community Vitality and Participation showing many low scores. This can be attributed to a small number of wards where crime is high and that wards in central London have much greater access to both cultural and social spaces as indicated on the maps below. In both cases, a power law may best describe these distributions.

Correlations

?INSERT MAPS OF EACH DOMAIN?

To begin the task of trying to model the median house prices from the well-being indicators, a simple multiple linear regression model was used with all six domains as predictors. The results show that this model explains less than half the variance with an R sq score of 0.39. There is enough to suggest that some form of linear model may provide a reasonable model of this problem.

? STATSMODEL ?

Next a polynomial model was tried with degree two. This model explained almost none of the variance with an R sq value of very close to 0. The mean squared error also increased dramatically. The polynomial model does not seem to be a strong candidate for solving the problem.

The next models involved some feature reduction based on the p values of the domains in the original linear model. This meant that first Education and Employment were removed and then Community Vitality and Participation. This resulted in a small improvement in the model fit but little improvement in the MSE.

With the median house price distribution suggesting a power law, a log transformation was performed on the data to try and linearise it with the linear regression model then re-run. This transformation dramatically increased the R sq value of the model, to 0.64, an increase of 0.25 from the equivalent model without the transformation. The model run on the log transformed data explains an additional 25 percent of the variance within the model. The exponent of the MSE is also a substantial reduction. This returned the following model:

?[insert model equation]?

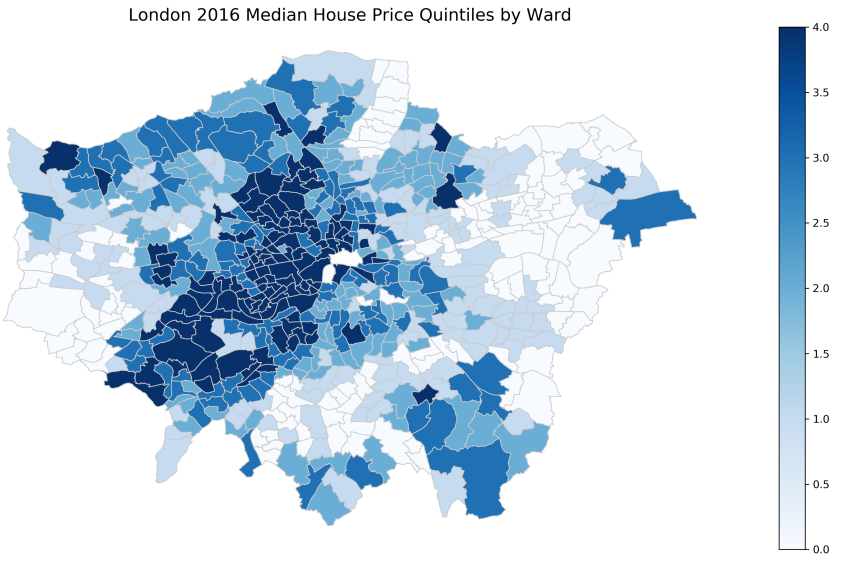
As the best performing regression model this will be used for prediction in the final visulisation.

When the predictions were given the inverse transformation and validated against the actual values, the majority of wards showed that this model had good predictive capability although there were a handful of wards where the prediction had a significant error margin. The wards with significant errors tended to be those which had a much higher or lower score in a particular domain as compared to the neighbouring areas.

Model Type	Av. R sq	Av. MSE
Linear Regression	0.3926146	35326.75098
Linear Regression (poly 2)	0.0067072	59013.81342
Linear Regression (reduced to 5 dim)	0.472768	33907.16412
Linear Regression (reduced to 4 dim)	0.4429811	34608.08413
Linear Regression (Log Transform)	0.6408463	4944.066064

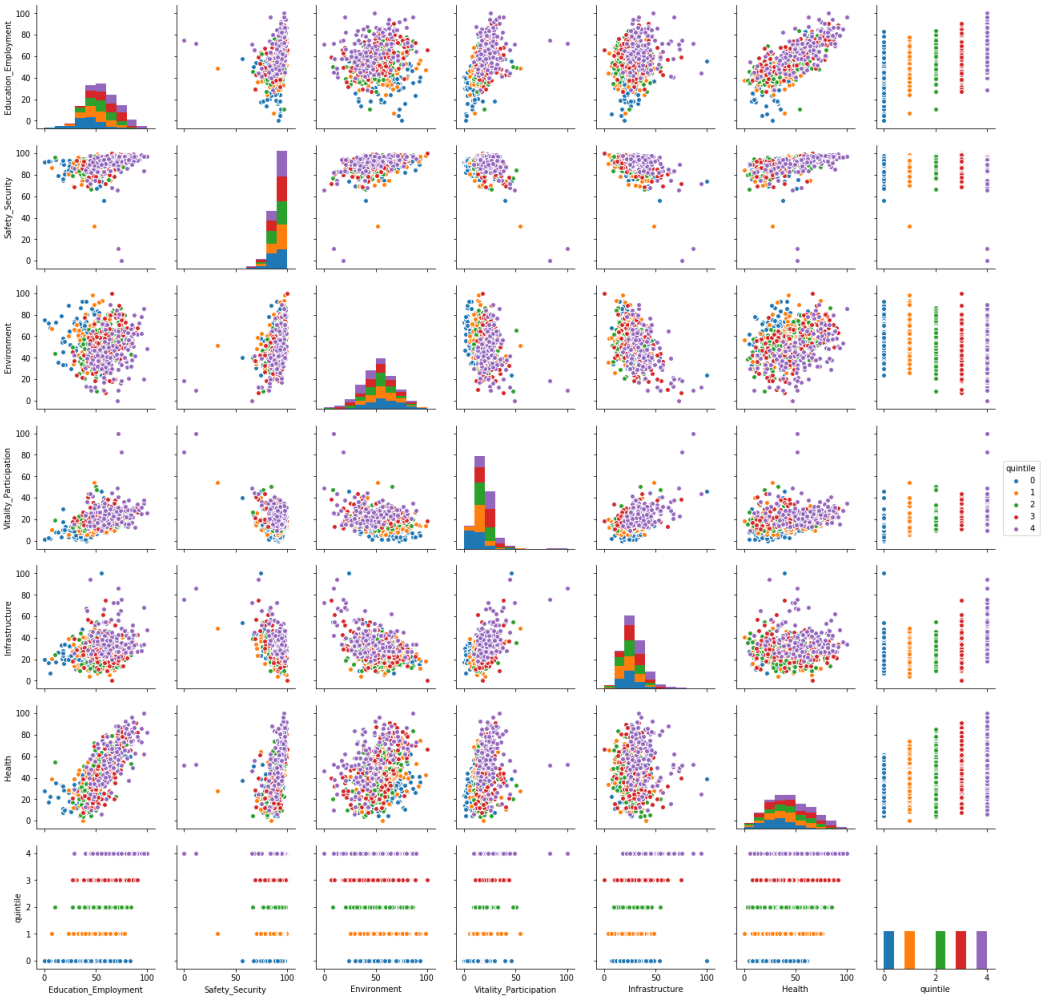
5.4 London house prices as a classification problem

The choropleth map below shows a different picture of London house prices being based on quintiles rather than monetary values, here the fifth quintile covers a larger range of values than the other quintiles due to the extremely expensive areas such as Knightsbridge and Belgravia where the median house price reaches over £2.8 million, 10 times as much as North End ward in Bexley.



? BOXPLOT WITH QUINTILES ?

For the classification problem the aim was to try to find a model that has both a good fit and high interpretability. However, models that often have high accuracy but low interpretability such as support vector machines and neural networks were also used to both discover and validate the best classification model for the problem.



By using the pair-wise plot and colouring the points by quintile, it is clear that there is no clear separation of quintile for any of the indicators with a more gradual movement from first to fifth quintile for most domains. Environment and Saftey and Security have an almost inverse relationship with the quintiles. Because the majority of the high-priced wards are centrally located, this may relate to a lack of greenspace and higher polution centrally as well as providing high-value targets for crime.

? FINISH WRITING UP CLASSIFICATION MODELLING ?

5.5 Encoding as geoJSON

To use the information from the creation of the well-being domains and the predictions obtained from the models in a javascript based visualisation, the information needed to be converted from a pandas dataframe in Python to a geoJSON file in javascript. A geoJSON file is a variation on the JSON format as defined below:

? GEOJSON DEFINITION ?

Whilst a relatively straightforward task within the geopandas module in Python, geoJSON accepts coordinates in latitude and longditude and the shapefile being

used has the coordinates defining the ward polygons in Ordnance Survey grid references. To change this to latitude and longditude required the creation of a function that would take each polygon in the ward geometry in the geo data frame and break this down into the individual points which define the polygon. Once separated into coordinate pairs, these pairs could be reprojected to be latitude and longditude. The function was then required to recombine all of the coordinate pairs into the correct polygon geometries.

? CODE SNIPPET?

The information was then in the required format to be exported to a geoJSON file.

5.6 Interactive maps

For the production of the interactive map which serves as the user front end for this project, the information was moved from the Python environment via geoJSON and into javascript functions embedded into html. This decision was made on the basis of the additional functionality and interactivity that could be smoothly added to the map in javascript. With less experience of javascript as a programming language, the leaflet.js tutorial ?INSERT REFERENCE? was an excellent starting point.

? SCREENSHOT OF MAP (not used in trailer) - MAYBE ZOOMED IN?

The map has key functionality added. This includes a zoom facility and hovering over any of the 625 wards will display the median house price value and quintile, the scores for each of the six well-being domains and the predictions from both the regression and classification models.

Chapter 6

Validation and testing

6.1 Validating the data inputs

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

6.1.1 Subsection 1

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

6.1.2 Subsection 2

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque. Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elit posuere quam, quis adipiscing urna leo nec orci. Sed nec nulla auctor odio aliquet consequat. Ut nec nulla in ante ullamcorper aliquam at sed dolor. Phasellus fermentum magna in augue gravida cursus. Cras sed pretium lorem. Pellentesque eget ornare odio. Proin accumsan, massa viverra cursus pharetra, ipsum nisi lobortis velit, a malesuada dolor lorem eu neque.

6.2 Testing the model

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor

6.3 Testing the outputs

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor

Chapter 7

Conclusions and evaluation

7.1 Lessons learnt

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

7.1.1 Subsection 1

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

7.1.2 Subsection 2

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque. Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elit posuere quam, quis adipiscing urna leo nec orci. Sed nec nulla auctor odio aliquet consequat. Ut nec nulla in ante ullamcorper aliquam at sed dolor. Phasellus fermentum magna in augue gravida cursus. Cras sed pretium lorem. Pellentesque eget ornare odio. Proin accumsan, massa viverra cursus pharetra, ipsum nisi lobortis velit, a malesuada dolor lorem eu neque.

7.2 Possible developments

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.