Life expectancy analysis

Cillian, James, Hope and Peter

Setting Up

Principles:

- Remote Collaboration
- Reproducible Research

Tools:

- Python (can open SAS files and SPSS files!)
- Jupyter Notebook
- Docker Container: https://hub.docker.com/r/darribas/gds_py
- Git Repository: https://github.com/peterprescott/ward-lifetimes

ward_lifetimes.ipynb

Code

Python 3

Ward Life Expectancies

Group project done by Cillian Berragan, James Murphy, Hope Bleasdale, and Peter Prescott, for the Manchester module of the Data Analytics MSc/PhD.

Task

Setting Up Computing Environment

For the sake of simple collaboration and reproducible research [@CBrunsdonASingleton2015], this task is presented here as a Jupyter notebook [@FPerezBGranger2015] that can be run in a Docker container [@CBoettiger2015]; specifically, version 4.0 of the GDS Python stack maintained by @Arribas-Bel2020, and based on the standard Jupyter container.

We then need use conda to install a few extra packages that aren't included in the gds_py:4.0 container (TODO: write Dockerfile to create a Docker image including these), and clone the git repository [@KRam2013] including this notebook.

Once you have installed Docker (if you are on Windows, you may need to upgrade to Windows 10 Pro), go to your Terminal (Powershell for Windows), and run these commands:

```
docker container run -it -p 8888:8888 darribas/gds_py:4.0 bash
conda install pyreadstat -y
conda install geoplot -y
conda install mlxtend --channel conda-forge -y
git clone https://github.com/peterprescott/ward-lifetimes
cd ward-lifetimes
jupyter lab ward_lifetimes.ipynb
```

This will take a little while to download. When it is ready, you will then be instructed to copy into your browser a URL that looks like this: http://127.0.0.1:8888/?token=39dd92f7720d42d5f9abab59485ca208a4dafb877852f1be (though your security token at the end will be different). Do that, click `ward_lifetimes.ipynb` on the left sidebar, and you should find yourself looking at a live version of this notebook.

We then begin by importing the packages that we are going to use.

[1]:
```
import pandas as pd # for data manipulation
import matplotlib.pyplot as plt # for visualization
import numpy as np
import os.path
from IPython.display import Image, display, HTML
```

Review the Data

Combining the Datasets

Visualize

Use Linear Regression to Predict Life-Expectancy

Data Preprocessing

Initially the data was separated into 5 varying file types containing information on London Boroughs. In order for the data to be joined together it had to first be cleansed by fixing all discrepancies between data sets. This process is available to view at https://github.com/peterprescott/ward-lifetimes/blob/master/ward_lifetimes.ipynb

Quick-look visualisation methods were used within each data set to observe any abnormalities within the data.

In this case KDE plots were used to measure the spread of the data.

In one case the raw data identified a life expectancy of ‘178’ which, unless the Sutton ward had undergone miraculous biological experiments, was not possible.

Exploring Datasets

Five files:

- London_District_codes.csv
- London_ward_data_socioeconomic.sav (SPSS file)
- London_ward_data_environment.csv (6char wardcode ID)
- London_ward_data_health.sas7bdat (SAS file)
- London_ward_data_demographics.dat (text file)

Merging Datasets

- Merge socio and env datasets on shared Wardcode

- Merge health and demo datasets on shared Wardname

- Separate Wardname into distinct names of District and Ward

- Create Districtcode from Wardcode and create District-Level Dataset

- Combine on Unique Population within District

Cleaning

- Remove whitespace from codes
- Extra characters at end of wardcode
- 178 year life-expectancy?
- & for ‘and’
- ‘Street’ and ‘St.’

Outcome Variables

Created new `Lifeexpectancy` variable:

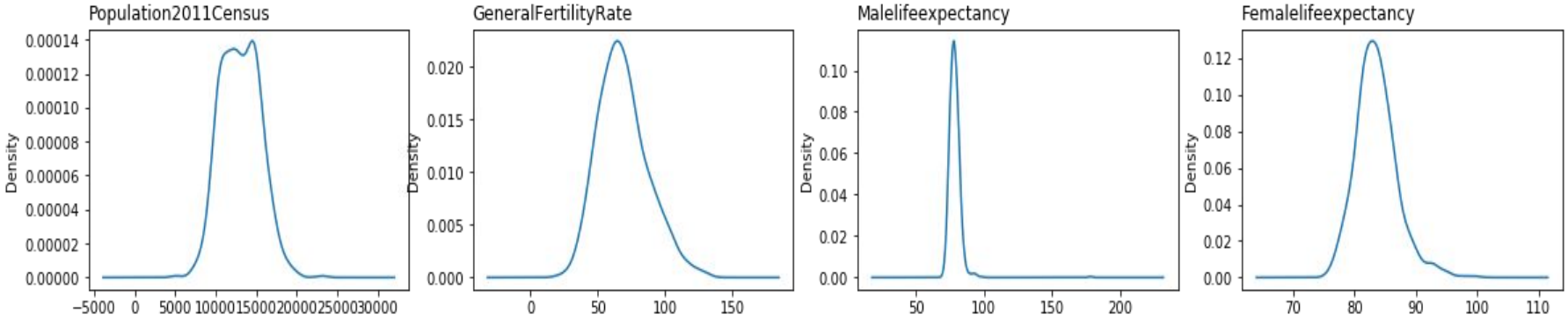
`Lifeexpectancy` = (`Malelifeexpectancy` + `Femalelifeexpectancy`)/2

Assumption: Men and women are equally distributed in each ward.

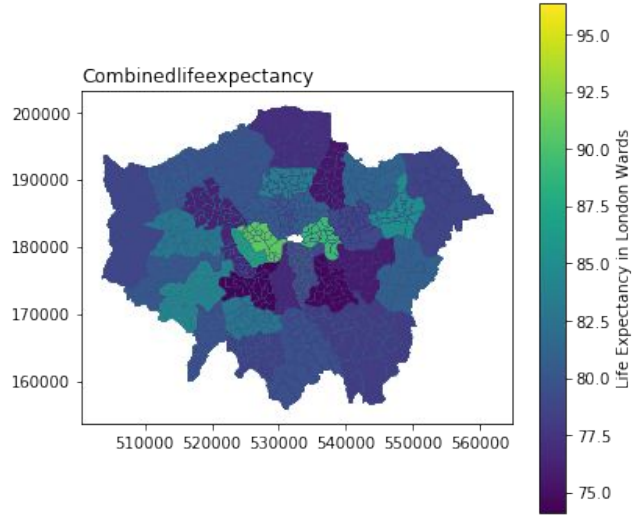
May not be true, so ran analysis for males and females separately as well as combined. But here we will assume our assumption is true and focus on the joint case.

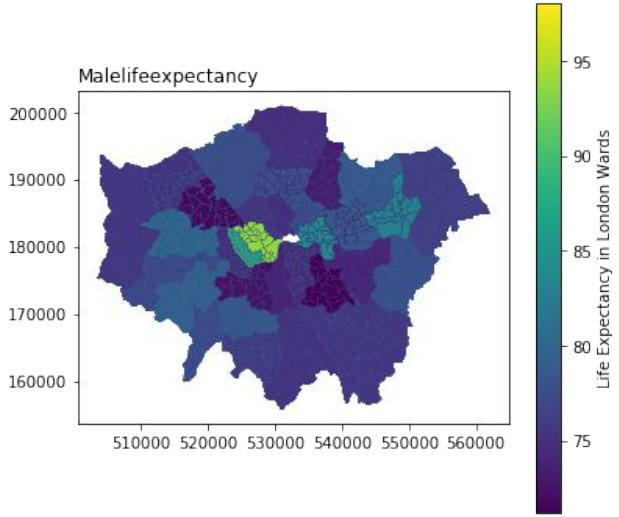
Variable Distribution

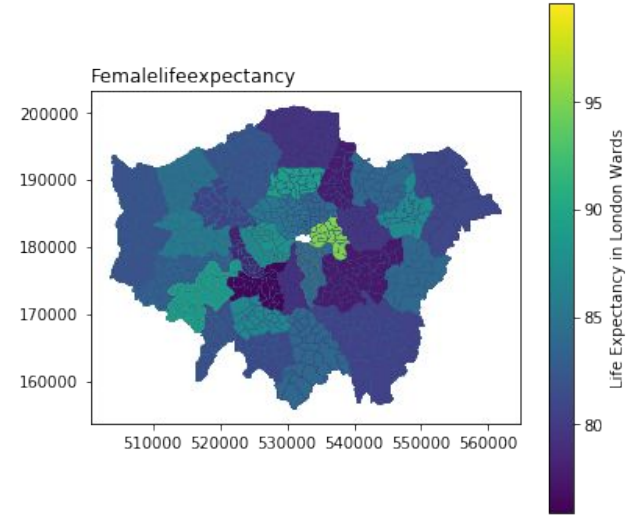
Visualization showed that there was an impossible outlier in the `Malelifeexpectancy` variable.



Mapping Life Expectancy Across Wards

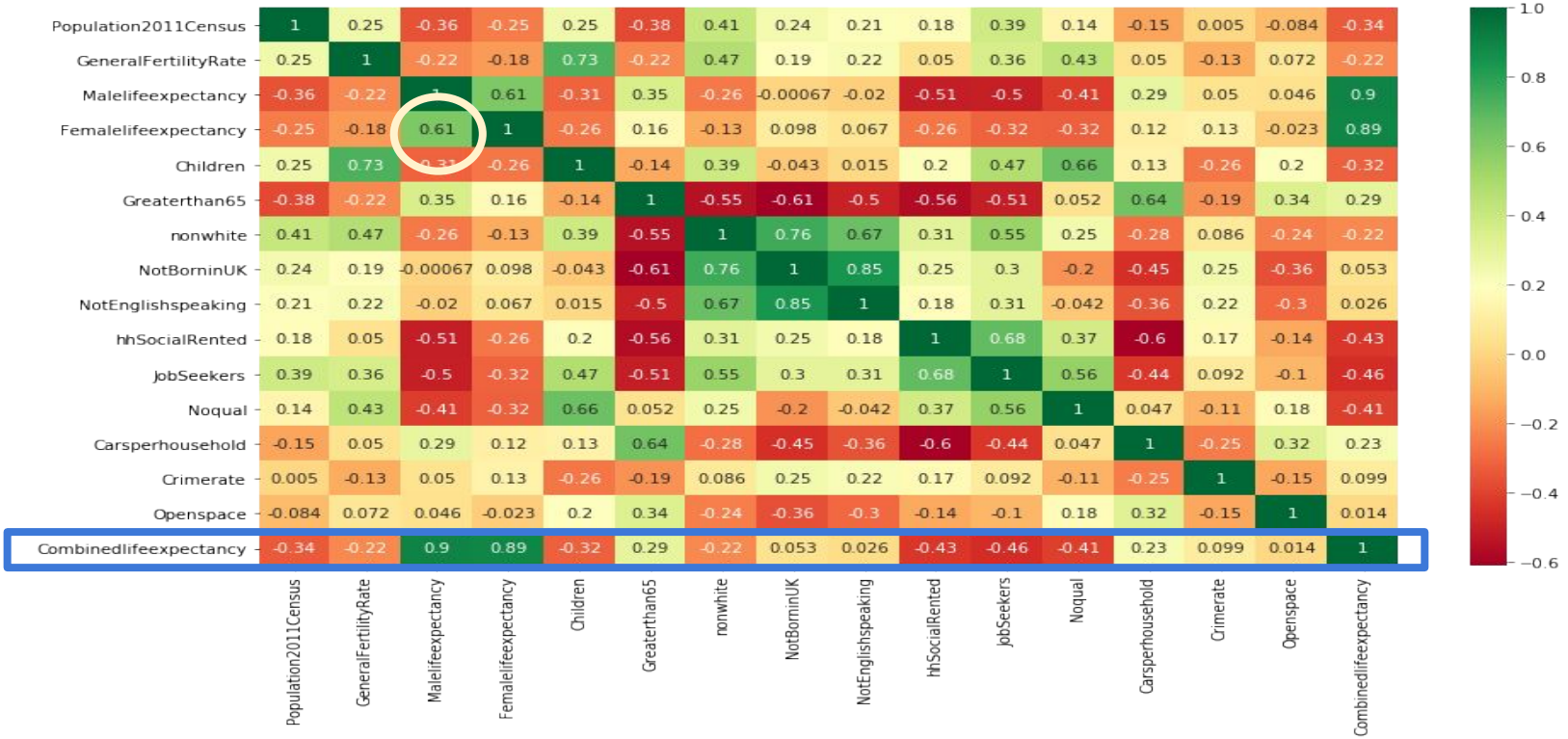




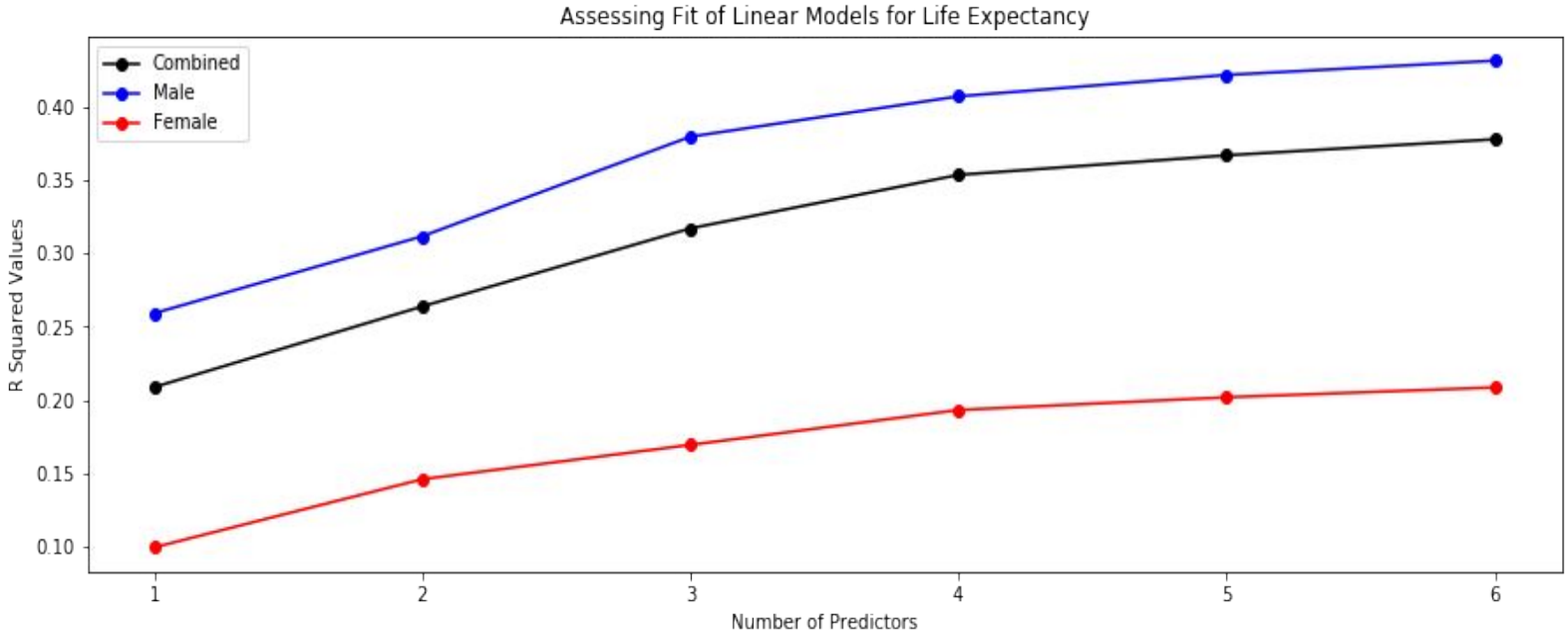


- Male and Female life expectancies are very different geographically
- Values are clustered geographically without having been binned into quantiles, which suggests that although the data has been given for each ward, it has actually been collected at a more aggregated level.

Correlation Matrix



How many dimensions are worth considering?



Finding a Best Predictive Linear Model

Table 1: Linear Models of Increasing Complexity, predicting Life Expectancy in London Wards

No. of Var.s	Intercept	Q("Greaterthan65")	Q("JobSeekers")	Q("Noqual")	Q("NotBorninUK")	Q("Population2011Census")	Q("hhSocialRented")	Q("nonwhite")	rSquared
1	-6.160003e-15	NaN	-0.457224	NaN	NaN	NaN	NaN	NaN	0.208718
2	-6.160003e-15	0.311499	NaN	-0.425129	NaN	NaN	NaN	NaN	0.263851
3	-6.160003e-15	NaN	NaN	-0.268190	NaN	-0.249514	-0.285302	NaN	0.317149
4	-6.160003e-15	NaN	NaN	NaN	0.481906	-0.233923	-0.392284	-0.365448	0.353673
5	-6.160003e-15	0.180240	NaN	NaN	0.567652	-0.198683	-0.317903	-0.367945	0.367042
6	-6.160003e-15	0.231887	NaN	-0.156925	0.449098	-0.196762	-0.243353	-0.234762	0.378022

```
lm_df['Male']
```

Table 2: Linear Models of Increasing Complexity, predicting Male Life Expectancy in London Wards

No. of Var.s	Intercept	Q("Greaterthan65")	Q("Noqual")	Q("NotBorninUK")	Q("Population2011Census")	Q("hhSocialRented")	Q("nonwhite")	rSquared
1	-9.384854e-16	NaN	NaN	NaN	NaN	-0.509492	NaN	0.259165
2	-9.384854e-16	0.376139	-0.432785	NaN	NaN	NaN	NaN	0.311769
3	-9.384854e-16	NaN	-0.238929	NaN	-0.254664	-0.373766	NaN	0.379788
4	-9.384854e-16	NaN	NaN	0.425544	-0.235421	-0.465212	-0.338226	0.407315
5	-9.384854e-16	0.188556	NaN	0.515247	-0.198555	-0.387399	-0.340838	0.421946
6	-9.384854e-16	0.237215	-0.147846	0.403551	-0.196746	-0.317163	-0.215361	0.431692

```
lm_df['Female']
```

Table 3: Linear Models of Increasing Complexity, predicting Female Life Expectancy in London Wards

No. of Var.s	Intercept	Q("Greaterthan65")	Q("JobSeekers")	Q("Noqual")	Q("NotBorninUK")	Q("Population2011Census")	Q("hhSocialRented")	Q("nonwhite")	rSquared
1	-1.845746e-15	NaN	-0.315381	NaN	NaN	NaN	NaN	NaN	0.099305
2	-1.845746e-15	NaN	NaN	-0.290420	NaN	-0.210802	NaN	NaN	0.145850
3	-1.845746e-15	NaN	-0.314430	NaN	0.237121	-0.186854	NaN	NaN	0.169297
4	-1.845746e-15	NaN	NaN	NaN	0.440199	-0.183960	-0.235979	-0.317838	0.193127
5	-1.845746e-15	0.256251	NaN	-0.195080	0.381962	-0.148651	NaN	-0.171176	0.201810
6	-1.845746e-15	0.178401	NaN	-0.133792	0.403044	-0.156052	-0.116968	-0.206149	0.208537

[100]:

OLS Regression Results

Dep. Variable:	Q("Combinedlifeexpectancy")	**R-squared:**	0.354
Model:	OLS	**Adj. R-squared:**	0.350
Method:	Least Squares	**F-statistic:**	84.95
Date:	Fri, 03 Apr 2020	**Prob (F-statistic):**	1.55e-57
Time:	15:28:55	**Log-Likelihood:**	-749.64
No. Observations:	626	**AIC:**	1509.
Df Residuals:	621	**BIC:**	1531.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.16e-15	0.032	-1.92e-13	1.000	-0.063	0.063
Q("Population2011Census")	-0.2339	0.036	-6.572	0.000	-0.304	-0.164
Q("nonwhite")	-0.3654	0.054	-6.831	0.000	-0.471	-0.260
Q("NotBorninUK")	0.4819	0.050	9.700	0.000	0.384	0.579
Q("hhSocialRented")	-0.3923	0.034	-11.523	0.000	-0.459	-0.325

Omnibus:	130.358	**Durbin-Watson:**	1.859
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	437.000
Skew:	0.960	**Prob(JB):**	1.28e-95
Kurtosis:	6.614	**Cond. No.**	3.12

Configuring the model

$$LifeExpectancy = 0.48(notborninuk) - 0.20(populationcensus) - 0.32(socialrented) - 0.37(nonwhite) + \epsilon$$

After running a number of exhaustive regression models using the ***mlxtend*** and ***statsmodels*** packages, an optimum model was chosen based on the R^2 score.

6 models were created for each variable: female life expectancy, male life expectancy and a combined (mean) life expectancy for both.

The 4 variable model was chosen as there was little statistical benefit in adding complexity beyond this.

To note: Female life expectancy was much more difficult to model - lower R^2. Conversely male life exp rather simpler to model.

Configuring the model - relevant literature

Deprivation

- Woods et al. (2004) showed that geographical variation in life expectancy is largely explained by deprivation (based on IMD score). This supports our findings as the variables **job_seekers** and **social_rented** were found to be important predictors of life expectancy in our model.

Socio-economic status

- Love-Koh et al. (2015) found that socio-economic status was an important attribute in Quality Adjusted Life Expectancy (QALE), along with sex and age. (Again related to variables **job_seekers** and **social_rented**).

Ethnicity

- Wohland et al. (2014) found the majority of ethnic groups in England and Wales have significantly lower disability-free life expectancy than white-British. This was also picked up in our model with the variable **Non-white**.

Education

- Meara et al. (2008) found higher education was associated with higher life expectancy - this notion was picked up by our model in the variable **no_quals**.

Open space

- A wealth of research demonstrates the link between health and proximity to open space - for mental health and physical health (Groeonwegan et al., 2006, Villanueva et al., 2015). Our model **did not** pick this up - the variable openspace was not recognised as a key predictor variable.

References

Groenewegen, P.P., Van den Berg, A.E., De Vries, S. and Verheij, R.A., 2006. Vitamin G: effects of green space on health, well-being, and social safety. BMC public health, 6(1), p.149.

Love-Koh, J., Asaria, M., Cookson, R. and Griffin, S., 2015. The Social Distribution of Health: Estimating Quality-Adjusted Life Expectancy in England. *Value in Health*, 18(5), pp.655-662.

Meara, E., Richards, S. and Cutler, D., 2008. The Gap Gets Bigger: Changes In Mortality And Life Expectancy, By Education, 1981–2000. Health Affairs, 27(2), pp.350-360.

Villanueva, K., Badland, H., Hooper, P., Koohsari, M.J., Mavoa, S., Davern, M., Roberts, R., Goldfeld, S. and Giles-Corti, B., 2015. Developing indicators of public open space to promote health and wellbeing in communities. Applied geography, 57, pp.112-119.

Wohland, P., Rees, P., Nazroo, J. and Jagger, C., 2014. Inequalities in healthy life expectancy between ethnic groups in England and Wales in 2001. *Ethnicity & Health*, 20(4), pp.341-353.

Woods, L. M., Rachet, B., Riga, M. Stone, N. Shah, A. and Coleman, M. P.. "Geographical variation in life expectancy at birth in England and Wales is largely explained by deprivation." *Journal of Epidemiology & Community Health* 59, no. 2, pp. 115-120.