Scottish Household Survey

The characteristics of private households

Ricardo Pulido

Overview

Worked with data 2013-2019

- Reports produced in a week.
- Survey based on a random sample of the general population in private residences in Scotland.
- Informs the Fairer Scotland agenda, National Performance Framework and feeds in to the Scottish Surveys Core Questions (SSCQ) sample.

Key questions

Green Spaces:

- Are there certain groups with local access to green spaces?
- Groups lacking access?
- Rural vs. urban?
- Demographics of people with good access vs. people without good access.

Ratings:

- Predicting neighbourhood ratings
- Predicting community belonging ratings

The Data

Three datasets

- Distance to Green or Blue Space
- Neighbourhood Rating
- Community Belonging

Difficulties

- Unusual format
- Percentage of adults for each variable pair
- Only one numerical variable
- Same variables binned differently in different datasets

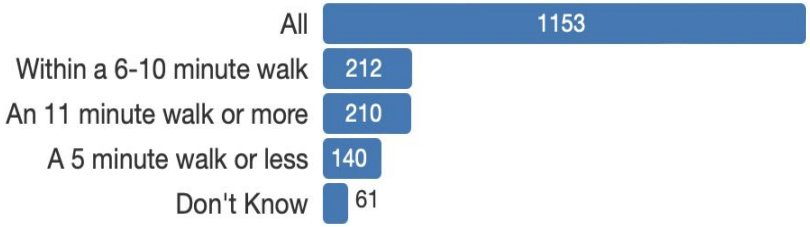
Distance to green or blue space

	featurecode	datecode	measurement	units	value	distance_to_nearest_green_or_blue_space	age	gender	urban_rural_classification	simd_quintiles	type_of_tenure	household_type	ethnicity
0	S12000026	2013	95% Lower Confidence Limit, Percent	Percent Of Adults	71.0	A 5 minute walk or less	All	All	All	All	All	All	All
1	S12000045	2017	Percent	Percent Of Adults	59.0	A 5 minute walk or less	All	All	All	All	All	Pensioners	All
2	S12000026	2014	95% Upper Confidence Limit, Percent	Percent Of Adults	86.9	A 5 minute walk or less	All	All	All	All	All	All	All
3	S12000026	2017	95% Upper Confidence Limit, Percent	Percent Of Adults	80.9	A 5 minute walk or less	All	All	All	All	All	All	All
4	S12000026	2017	95% Upper Confidence Limit, Percent	Percent Of Adults	79.6	A 5 minute walk or less	All	All	All	All	All	Pensioners	All
...	...	...	...	...	...	...	...	...	...	...	...	...	...
38446	S92000003	2018	Percent	Percent Of Adults	26.0	Within a 6-10 minute walk	All	All	All	All	All	All	Other
38447	S92000003	2018	95% Lower Confidence Limit, Percent	Percent Of Adults	20.4	Within a 6-10 minute walk	All	All	All	All	All	All	Other
38448	S92000003	2019	95% Upper Confidence Limit, Percent	Percent Of Adults	7.8	Don't Know	All	All	All	All	All	All	Other
38449	S92000003	2014	95% Lower Confidence Limit, Percent	Percent Of Adults	15.8	Within a 6-10 minute walk	All	All	All	All	All	All	Other
38450	S12000036	2018	95% Upper Confidence Limit, Percent	Percent Of Adults	36.1	Within a 6-10 minute walk	All	All	All	All	All	All	Other

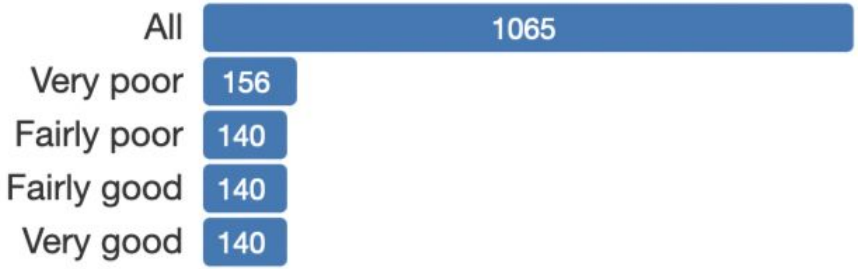
Prepared data

	year	percent_adults	nearest_green_space	age	gender	urban_rural_classification	simd_quintiles	type_of_tenure	household_type	ethnicity	community_belonging	neighbourhood_rating
50001	2017	38.0	All	All	All	Rural	All	All	All	All	Fairly strongly	All
9299	2017	64.0	A 5 minute walk or less	All	All	All	All	All	Adults	All	All	All
213	2018	12.0	An 11 minute walk or more	All	All	Urban	All	All	All	All	All	All
33632	2014	18.0	Within a 6-10 minute walk	All	All	All	All	All	Adults	All	All	All
107271	2018	55.0	All	All	All	All	All	Other	All	All	All	Very good
15027	2018	3.0	All	All	All	All	All	Social Rented	All	All	All	All
113325	2013	9.0	All	All	All	All	All	Social Rented	All	All	All	Fairly poor
13873	2017	15.0	An 11 minute walk or more	All	All	All	All	Other	All	All	All	All
38852	2014	69.0	A 5 minute walk or less	All	All	All	All	All	All	White	All	All
32962	2016	21.0	Within a 6-10 minute walk	16-34 years	All	All	All	All	All	All	All	All

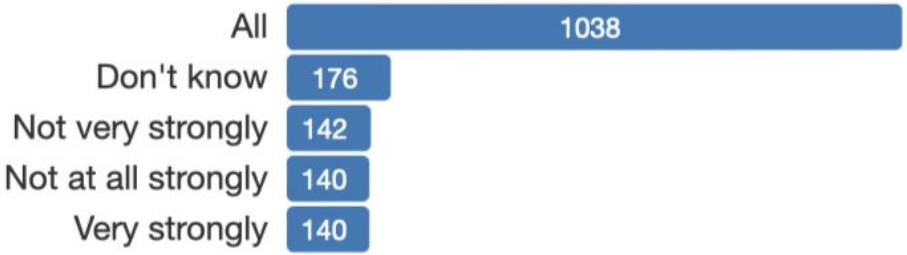
nearest_green_space



neighbourhood_rating

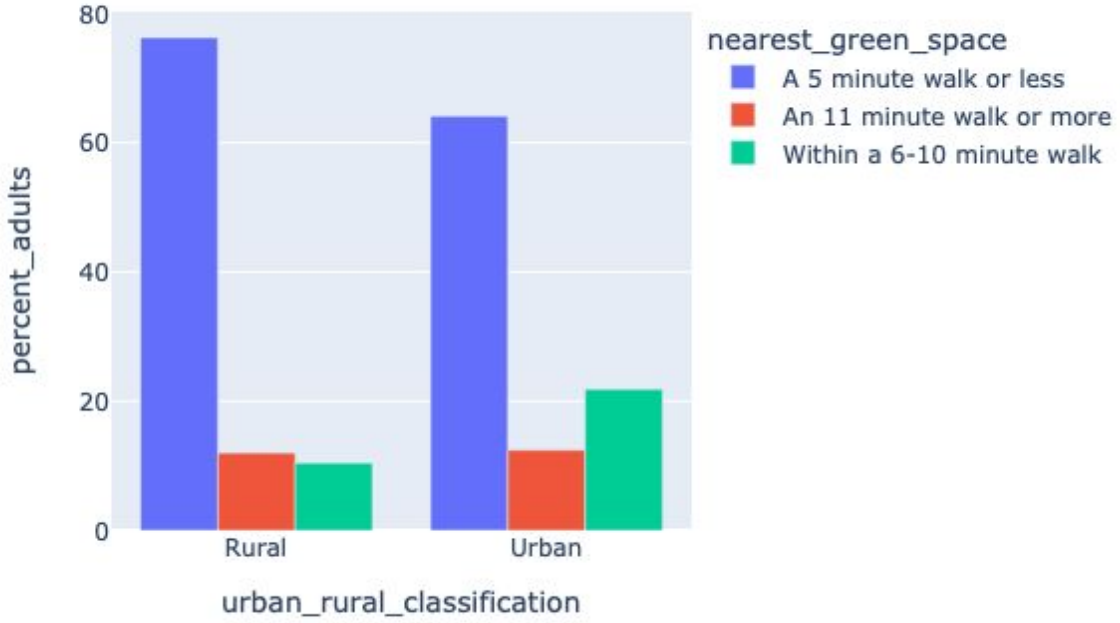


community_belonging



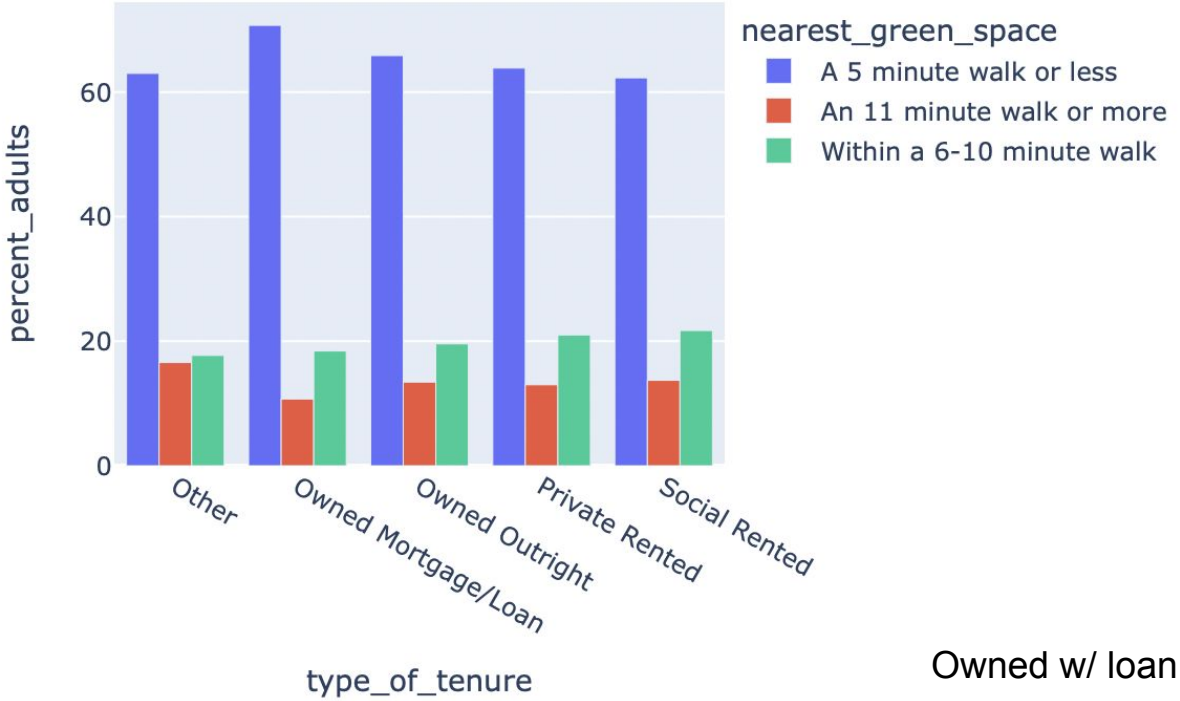
1. Good access to green spaces?

1. Good access to green spaces?



12% difference between rural and urban (5 min or less)

1. Good access to green spaces?

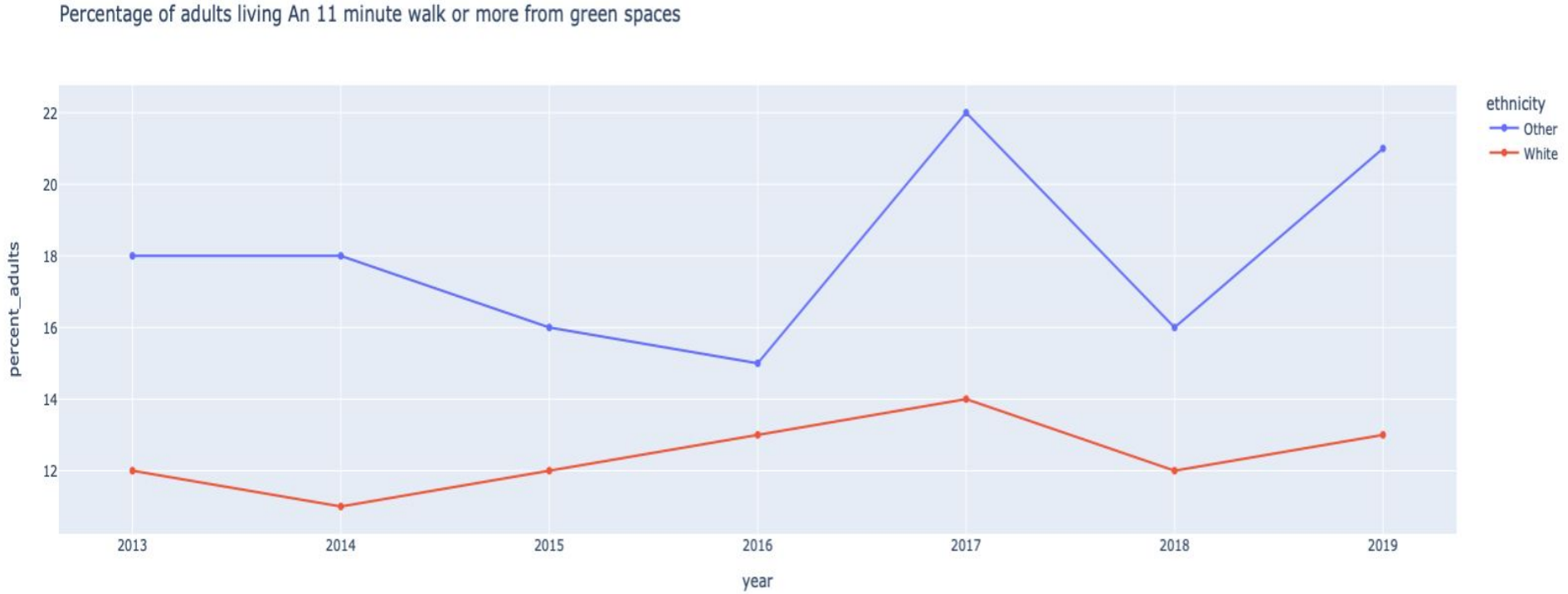


Owned w/ loan just under 70%

2. Lack of access to green spaces?

Scottish index of multiple deprivation

2. Lack of access to green spaces?



2. Lack of access to green spaces?

3. Demographics, good vs. bad access to green spaces

Standard deviation as a way to measure which categories make up for the differences

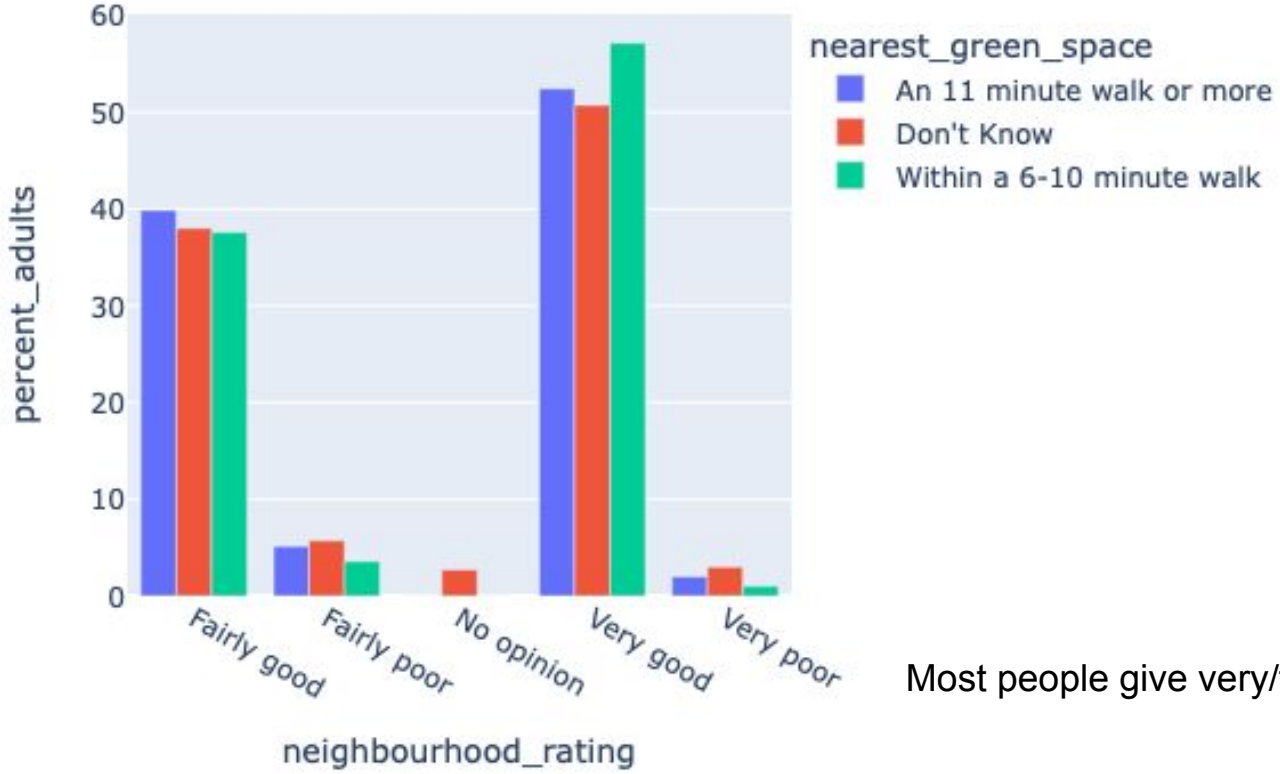
```
df_std(survey, 'nearest_green_space', 'An 11 minute walk or more')
```

	year	percent_adults	nearest_green_space	age	gender	urban_rural_classification	simd_quintiles	type_of_tenure	household_type	ethnicity	community_belonging	neighbourhood_rating
percent_adults	0.843297	NaN	NaN	3.94252	1.313198	0.303046	2.424366	2.091284	2.755329	3.939595	16.640635	24.446819

Out of all the demographic categories, age, ethnicity, household type, SIMD quintiles and type of tenure seem to explain most of the variation.

What is the high SD for neighbourhood rating and community belonging?

3. Demographics, good vs. bad access to green spaces



Most people give very/fairly good ratings

3. Demographics, good vs. bad access to green spaces

Good access

- Rural households
- Owned with mortgage/loan
- Households with children

Lack of access

- 65+ / retired people
- Non-white ethnicity
- The 20% most deprived

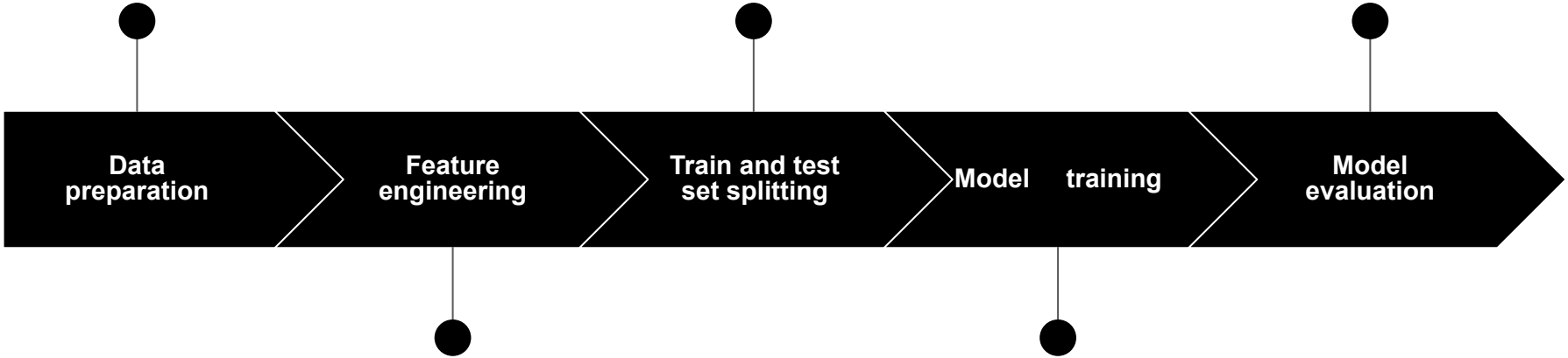
Predictive modelling

Household rating and community belonging

Reading in data, filter out CIs and aggregated values

2013-2018 to train, then predict 2019 to test

Mean absolute error, model accuracy, ROC curve (T/F positive ratio)



Encode variables into binary values

Logistic regression and random forest, via pipelines

4. Modelling. Data preparation

	datecode	value	neighbourhood_rating	gender	urban_rural_classification	simd_quintiles	type_of_tenure	household_type	ethnicity	walking_distance_to_nearest_greenspace	community_belonging
28306	2015	40.0	Fairly good	All	All	All	All	Adults	All	All	All
16190	2018	35.0	Fairly good	Male	All	All	All	All	All	All	All
8521	2015	33.0	Fairly good	All	All	All	Owned Mortgage/Loan	All	All	All	All
55709	2016	43.0	All	All	All	All	All	All	All	All	Fairly strongly
45679	2019	10.0	All	All	Urban	All	All	All	All	All	Not very strongly
73762	2015	22.0	All	All	All	All	Owned Outright	All	All	All	Fairly strongly
50566	2014	38.0	All	All	All	20% most deprived	All	All	All	All	Fairly strongly
32410	2019	28.0	Fairly good	All	All	All	Owned Mortgage/Loan	All	All	All	All
64080	2018	31.0	All	All	All	All	All	All	White	All	Very strongly
57178	2014	1.0	All	All	All	All	All	All	White	All	Don't know

4. Modelling. Feature engineering

	datecode	value	good_neighbourhood	good_community	green_access	owners	tenants	pensioners	children	urban	female	lowest_quintile	white_ethnicity
77832	2018	0.0	0	0	0	0	1	0	0	0	0	0	0
23895	2015	2.0	0	0	0	0	0	0	0	0	0	0	0
57867	2017	18.0	0	0	0	0	0	0	0	0	0	0	1
57124	2014	19.0	0	0	0	0	0	0	1	0	0	0	0
77005	2016	34.0	0	1	0	0	1	0	0	0	0	0	0
77056	2019	48.0	0	1	0	0	1	0	0	0	0	0	0
10556	2018	4.0	0	0	0	0	0	0	0	0	0	1	0
73872	2019	2.0	0	0	0	1	0	0	0	0	0	0	0
14763	2014	40.0	1	0	0	0	0	0	0	0	0	0	0
51121	2016	38.0	0	1	0	0	0	0	1	0	0	0	0
26894	2017	32.0	1	0	0	0	0	0	1	0	0	0	0
15695	2015	26.0	1	0	0	0	0	0	0	0	0	0	0
55677	2016	44.0	0	1	0	0	0	0	0	0	0	0	0
2427	2013	2.0	0	0	1	0	0	0	0	0	0	0	0
60813	2014	6.0	0	0	0	0	0	0	0	0	0	0	0
72522	2015	9.0	0	0	0	1	0	0	0	0	0	0	0
73281	2018	34.0	0	1	0	1	0	0	0	0	0	0	0
27572	2013	93.0	1	0	0	0	0	1	0	0	0	0	0
45460	2013	0.0	0	0	0	0	0	0	0	0	0	0	0
59379	2014	40.0	0	1	0	0	0	0	0	0	1	0	0

4. Modelling. Train and test sets

```
# Train and test split
train = survey.query("datecode != 2019").drop('datecode', axis=1)

test = survey.query("datecode == 2019").drop('datecode', axis=1)

# Now drop both targets from the predictor sets
X_test = test.drop(columns=['good_neighbourhood', 'good_community'])
X_train = train.drop(columns=['good_neighbourhood', 'good_community'])

# Create both targets
community_test = test.good_community
community_train = train.good_community

neighbourhood_test = test.good_neighbourhood
neighbourhood_train = train.good_neighbourhood
```

- Test with 2019, train with the rest
- Drop both response targets in case they’re correlated

4. Modelling. Building and fitting models

1: neighbourhood, LR

- Strong feature engineering
- Pipeline 1 separating categorical and numerical data

2: community, LR

- Strong feature engineering
- Pipeline 1 separating categorical and numerical data

3: neighbourhood, RF

- No feature engineering
- Pipeline 2 separating categorical and numerical data

4: community, RF

- No feature engineering
- Pipeline 2 separating categorical and numerical data

4. Modelling. Scores

1: neighbourhood, LR

- Mean absolute error ~ 0.185
- Cross-validation accuracy ~ 0.66
- AUC ~ 0.70

1: neighbourhood, LR

- Mean absolute error ~ 0.32
- Cross-validation accuracy ~ 0.66
- AUC ~ 0.47

3: neighbourhood, RF

- Cross-validation accuracy ~ 0.43

4: community, RF

- Cross-validation accuracy ~ 0.35

Next steps

Better data preparation / feature engineering
Current format (data cube) too abstract / not amenable to direct analysis and interpretation

Better representation for neighbourhood and community ratings
No clear relationship between good ratings and other household characteristics

Trying and evaluating other models
Possibly nearest neighbours/K-means clustering

Data source:
Community Wellbeing and Social Environment