Context

Business intelligence and data-driven decision making

This project provides insights on how different groups of people have accessing to green space and how the government can help increase the neighbourhood ratings

Domain knowledge and the business context

Scottish Household servers is run by Scottish government. It is designed to help the Scottish Government and other bodies to plan services and policies for Scotland. The project helps the government to understand the differences in how people accessing green space in order to make better policies.

Questions to answer:

- Are there certain groups that have local access to green space?
- Are there groups that are lacking access?
- What there big differences in how far people have to walk to access their green space?
- Are there any differences between rural and urban areas?
- How do people in neighbourhoods with good access to green space differ from those who have no good access? Are there differences in how they rate their neighbourhoods? Are there differences in how they rate their communities?
- Is there any way to predict which households would have higher ratings?

Data

Internal and external data sources

All data used were from the organisation and are open.

I used ‘neighboourhood_rate.csv’, (38055 rows x 15 columns)

‘dis_green_blue.csv’, (38451 rows × 15 columns)

‘community_belonging.csv’, (43611rows x 15 columns)

‘shs_aggregate_rasponses.csv’ (58530 rows x 13 columns)

dis_green_blue, community belong and neighbourhood rate are very similar, they all includes:

featurecode', 'featurename', 'featuretype', 'datecode', 'measurement', 'units','value','distance_to_nearest_green_or_blue_space' 'gender', 'urban_rural_classification', 'simd_quintiles', 'type_of_tenure', 'household_type', ‘ethnicity’; Community belong and neighbourhood rate, don’t have age columns but they have community belong and neighbourhood rate instead.

shs_aggregate_responses has:

'year', 'community_belonging', 'neighbourhood_rating', 'distance_to_nearest_green_space', 'satisfaction_with_nearest_green_space', 'age', 'gender', 'economic_status', 'household_size', 'highest_education_level','nearest_green_space_use', ‘volunteering_last_twelve_months','n_persons'.

Types of data

Mostly are categorical strings. Also has numerical types.

featurecode,featurename,featuretype,measurement,distance_to_nearest_green_or_blue_space,age,gender,urban_rural_classification,simd_quintiles,type_of_tenure,household_type,ethnicity are Categorical. Each of they has a number of categories and each recode has one of those categories.

datecode, value are Numerical. They are numbers.

Data formats

CSV formats

Data quality and bias

For Ethnicity groups, all the 'Other' groups are from Edinburgh and Glasgow, The data has a lot of missing values. Data was pre-grouped and can not compare between columns.

Ethics

Ethical issues in data sourcing and extraction

No, The Scottish Household Survey (SHS) is a continuous survey based on a random sample of the general population in private residences in

Scotland.Data used for this project is open to public. There is no identifiable information.

Ethical implications of business requirements

No, the business requirement were to gain insights into how people feel about their local communities in Scotland. In particular, we are interested in the relationship between distance to outdoor space, and neighbourhood ratings. The target and aim are to use the insights to make better policy that improves community belonging and neighbourhood rating. The requirements and aim are using data to contribute to society so I don't think there are any ethical issue.

Analysis

Stages in the data analysis process

1.Data exploring and cleaning 2.Feature engineering 3.Converting 'values' to proportions 4.Creating and interpreting plots

5. Hypothesis test 6. Building models

Tools for data analysis

I used Python for my analysis, within it, I used Pandas, Numpy, Seaborn, Matplotlib, scitick-learn

Descriptive, diagnostic, predictive and prescriptive analysis

My project fall under all the following categories.

Descriptive Analytics tells you what happened in the past. In my project, I created many charts and visualisation that describes what happened during 2013 and 2019 in terms of the rating. It shows how different groups rated their neighbourhood.

Diagnostic Analytics helps you understand why something happened in the past. Combined with my visualisations and decision tree model, my project shows how important each feature are and at which level they start to give different ratings of neighbour. This helps us to understand why we had the rating results from 2013 and 2019.

Predictive Analytics predicts what is most likely to happen in the future. Based on the 2013 to 2019 data, in my project, my decision tree model can be used to predict what may happen in future based on similar scenarios in the past.

Prescriptive Analytics recommends actions you can take to affect those outcomes. Combined with my analysis and interpretation of the decision tree, it suggested who are likely to give higher rate, which can help government to make reactive policies based on it. For example, my project suggests people who has good access to green space are more likely to rate their neighbourhood highly and we spotted people over 65 lack green space accessing. This suggest that government should provide more green space to people over 65 to improve their neighbourhood ratings.