**PDF REPORT DATA2001**

Dataset Description: What are your data sources and how did you obtain and pre-process the data?

Most of the datasets were provided by the University of Sydney. 2 of our datasets came from the City of Sydney Open Data Hub, a public website that has open data that anyone can use.

<https://data.cityofsydney.nsw.gov.au/>

|  |  |
| --- | --- |
| DATASET | Information |
| Neighbourhoods.csv | Census data of NSW neighbourhoods providing demographical information such as population size, land area and number of dwellings in neighbourhood |
| BusinessStats.csv | Looking at the Sydney business aspect, relating are to the number of businesses and what sorts of businesses there are in each district. |
| SA2\_2016\_AUST.zip | Provides geometries to be used for POSTGIS relating to SA2 areas and their boundaries for spatial joins |
| break\_and\_enter.zip | - shape data of theft ’hotspots’ in NSW as determined by BOCSAR. |
| school\_catchments.zip | Contains data about primary, secondary and future Government schools catchments |
| Walking\_count\_sites.geojson | Looking at how pedestrians get around the city with results from twice yearly walking count studies |

Pre-processing

We first connected the PostgreSQL database from Python and then created schemas with appropriate domain types for each table corresponding to the relevant datasets. We also cleaned the datasets to use it.

|  |  |
| --- | --- |
| DATASET | Cleaning Method |
| Neighbourhoods.csv | * Set empty cells to NaN using numpy * Remove regions not an sa2 |
| BusinessStats.csv | * Remove regions not in sa2 * Convert to string |
| SA2\_2016\_AUST.zip | * Make all columns names lowercase * Filter rows that are only part of Greater Sydney * Removing unnecessary columns * Checked for entries with no geometry which are then removed * Convert values from object to more usable forms * Make column names neater * Convert to postGIS |
| break\_and\_enter.zip | * Make all columns lower case * Cast density as string * Change geometry datatypes as with SA2 data * Rename columns to correct formatting |
| school\_catchments.zip | * Combine 3 catchment shapefile into one and remove unnecessary columns * Make all rows and columns lower case * Cast again * Make polygon nice |
| Walking\_count\_sites.geojson | * Making location of sites into a string * Make the columns all lower case * Renaming the columns |

(b) Database Description: Into which database schema did you integrate your data (preferable shown with a

diagram)? Which index(es) did you create, and why?

|  |
| --- |
| neighbourhoods |
| Sa2\_code integer |
| Sa2\_name VARCHAR(100) |
| Land\_area FLOAT |
| Population FLOAT |
| Number\_of\_dwellings FLOAT |
| Median\_annual\_household\_income FLOAT |
| Avg\_monthly\_rent FLOAT |
| total\_young\_people INTEGER |

|  |
| --- |
| Business\_stats |
| Sa2\_code Integer |
| Sa2\_name VARCHAR |
| number\_of\_businesses INTEGER |
| Accommodation\_and\_food\_services INTEGER |
| Health\_care\_and\_social\_assistance INTEGER |

|  |
| --- |
| Break\_and\_enter |
| Object\_id INTEGER |
| Density varchar(50) |
| Shape\_length FLOAT |
| Shape\_area FLOAT |

(c) Greater Sydney Score Analysis: Show which formula you applied to compute the liveability score per neighbourhood, and give an overview of the results through

(d) Correlation Analysis: How well does your score correlate with the median rent and median income in each

neighbourhood?

(e) City of Sydney Analysis: Propose a stakeholder and give a brief introduction. Show how you tailored your

score for their needs. Demonstrate the results on a map.