DATA2001 Assignment - Report

Lab-30 Group 5
Name: Syed Ahmad Sabaat SID: 510415790
Name: Frederic Max Serisier SID: 530490355
Name: Duc Viet Hoang Vu SID: 530016546

Data Description:

Business.csv
This dataset was given to us as part of the assignment and so was obtained through canvas. This dataset did not need much cleaning, with the only change being the renaming of the 'sa2_code' column to 'SA2_CODE21' to remain constant with our other dataframes. It contains data on different types of businesses in each SA2 region in NSW.

Income.csv
This dataset was given as part of the assignment. The pre-processing involved renaming the SA2 code column to 'SA2_CODE21' and also replacing any cells with value 'np' with NULL values, since 'np' means 'not provided'. This dataset contains data on income per SA2 region in NSW.

Population.csv
This dataset was given as part of the assignment. The pre-processing involved renaming the 'sa2_code' column to 'SA2_CODE21'. It contains data on population number for each SA2 region in NSW.

Stops.txt
This dataset was also obtained through canvas as part of the assignment. The pre-processing of this dataset came in the form of some column renaming for our own ease and also converting the latitude and longitude columns into POINT geometries with its own column 'geom'. Further, the SRID of the dataset was made to match the SRID that we would be using across our database (4326). It contains data of all public transport stops and their locations in the Sydney area.

PollingPlace2019.csv
This dataset was provided as a part of the assignment via canvas. This dataset required some cleaning such as removing the unwanted columns, which were the "the_geom" and the "premises_state_abbrevation". In addition, we also convert the "latitude" and longitude" into POINT geometries in the "geom" column. And we ensure that the SRID of the "geom" column matches the SRID that we use across the database (4326). It contains data on polling places and their locations in the Sydney area.

Schools.shp
This dataset was an aggregation of 3 dataframes provided as part of the assignment. It contains a

shapefile of all primary, secondary and future catchment areas for schools in NSW. The datasets were combined together into one school dataset on the condition that future catchments were prioritised.

SA2 digital boundaries

This dataset was provided as a part of the assignment via canvas. This dataset doesn't require much cleaning, with the only change being that it's filtered to include only Greater Sydney. It contains data on the geometries of each SA2 region in NSW and also certain attributes such as the area of the region in square kilometres.

Traffic-lights-location-data-may-2021.csv

This dataset was sourced from data.nsw.gov.au. The pre-processing of this dataset involved small tweaks to the naming of some columns and also the dropping of columns that would not be needed later on. Like many of the other datasets, we convert the latitude and longitude columns into POINT geometries in a new 'geom' column and ensure that the SRID of this column matches the SRID that we use across the database. This dataset contains data for locations of all traffic lights in NSW.

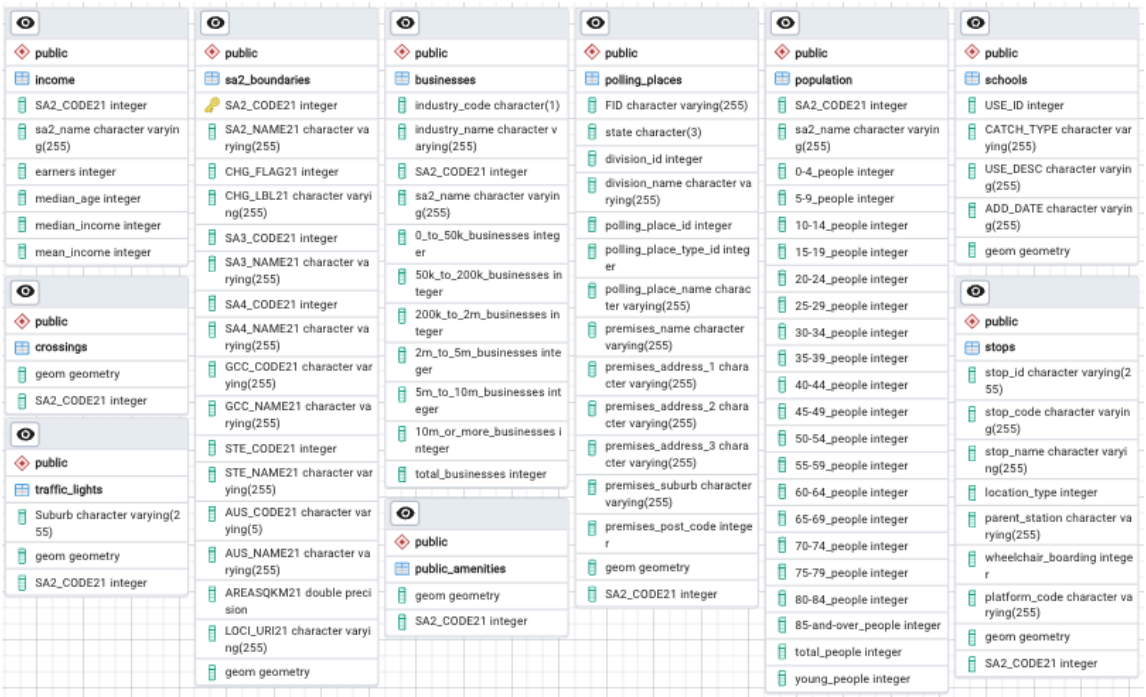
Publicamenities.geojson

This dataset was exported from overpass-turbo after using a code snippet that finds the locations of all water fountains and public toilets in NSW. Pre-processing involved dropping all redundant columns and converting latitude and longitude columns into POINT geometries.

Crossings.geojson

This dataset was exported from overpass-turbo after using a code snippet that finds the locations of all pedestrian crossings in NSW. Pre-processing involved dropping all redundant columns and converting latitude and longitude columns into POINT geometries.

Database Description



Score Analysis

The ‘bustling’ score for each SA2 area was calculated according to the following formula:

$$\text{Score} = S\left(\frac{z_{\text{business}} + z_{\text{stops}} + z_{\text{polls}} + z_{\text{schools}} + z_{\text{traffic lights}} + z_{\text{public amenities}} + z_{\text{crossings}}}{7}\right)$$

We add up each of the z-scores and then take their average. We then finally put this average through the sigmoid function, which maps any input to a number between 0 and 1, which will be our score for how ‘bustling’ the area is.

This final score also ignores any SA2 region with a ‘young people’ population of less than 100 - This decision was made so that it ignores regions with populations less than 100 but also ignores regions with a very small number of kids so that the z-schools score doesn’t have too many outliers.

Below are each of the calculations for coming up with each individual z-score:

Z-business

For the businesses z-score, we needed to select certain industries that we thought would most aptly represent a ‘bustling’ area. Because of this, we decided to choose 3 industries based on this logic and find the total number of these industries in each sa2 region per 1000 people. The business types were also weighted according to how important we thought they were to our ‘bustling’ metric.
The industries we chose were:

- Retail trade

Lots of Retail trade businesses in one region is a hallmark of a ‘bustling’ area. It is one of the largest in-person customer service industries and so of course should be selected. It was weighted at 30% of total businesses

- Arts and Recreation Services

This industry type includes concerts, events etc that would generate a lot of foot traffic and thus contribute to a more bustling area. This was weighted at 20% of total businesses

- Accommodation and Food Services

This includes hotels, bars, restaurants etc which would be the greatest contributor to our bustling metric and so was weighted at 50% of total businesses.

Z-stops

The number of public transport stops in a region is possibly one of the greatest metrics for finding out how bustling an area is, since more movement = more bustle. So the number of stops for each SA2 region was found and a z-score was assigned appropriately, however we ran into a small problem. Because SA2 regions are not all the same size, the larger regions were spitting out very large numbers simply because more area = more area for stops. This caused the smaller areas (that would normally be considered quite bustling) to have much smaller values and thus smaller bustling scores. This problem was eradicated by calculating this metric per square kilometre. This fixed the problem through

normalisation and allowed us to see density of stops rather than just number of stops, which is a much more valuable number to us for this scenario.

Z-polls

The number of polling places in a region is another great metric for figuring out a bustling score since polling places are chosen based on where people are travelling around. Similar to stops, the number of polls for each SA2 was found along with its z-score but the same problem as before arose regarding larger areas having disproportionately large bustling scores. And so this metric too was calculated per square kilometre to allow for more accurate bustling scores and normalisation.

Z-schools

Schools are areas in which the word 'bustling' is a perfect description for how they are, and so any data regarding them will be useful in our bustling metric. The data in particular were their catchment areas, and even more specifically, how many catchment areas intersect with each of the SA2 regions.
The more intersecting catchment areas, the more schools and therefore the more 'bustle' in the region. However, we needed to normalise the data due to the fact that different SA2 areas will have varying student populations. In light of this, we calculated this metric per 1000 young people (where young people are people aged 0-19 eg. our students).

Additional datasets:

We can extend our formula by adding in 3 more z-scores for: traffic lights, public amenities and pedestrian crossings in each SA2 region.
Extending our score with additional datasets should (in theory) increase the accuracy of our 'bustling' metric. This is because adding more z-scores inside our sigmoid function will reduce the impact of each individual z-score on the total bustling score.
However, adding more z-scores inside our sigmoid function also means that there will be an increase in bustling scores at the extreme values of 0 or 1 (since bustling areas will have all high z-scores and therefore higher value inside sigmoid and vice versa). To mitigate this, we take the average of the z-scores before applying the sigmoid function. This allows regions with all-round high z-scores to have a higher bustling score, which is exactly what we want since for a place to be 'bustling' it has to have a combination of these individual bustling metrics.

Z-Traffic lights

Traffic lights themselves don't necessarily mean an area is bustling, however, lots of traffic lights in one area generally mean there is more traffic movement and therefore more 'bustle'. This is the reason we chose the number of traffic lights per square kilometre per SA2 region (again, we use per square kilometre to normalise the data as mentioned previously) as another addition to our bustling formula.

Z-Public amenities

Public water fountains and bathrooms are generally introduced into an area when the government notices that there is a lot of general public walking around the specific area. Therefore, wherever there are lots of these, there will be lots of people - And lots of people means a more 'bustling' area. Again, we

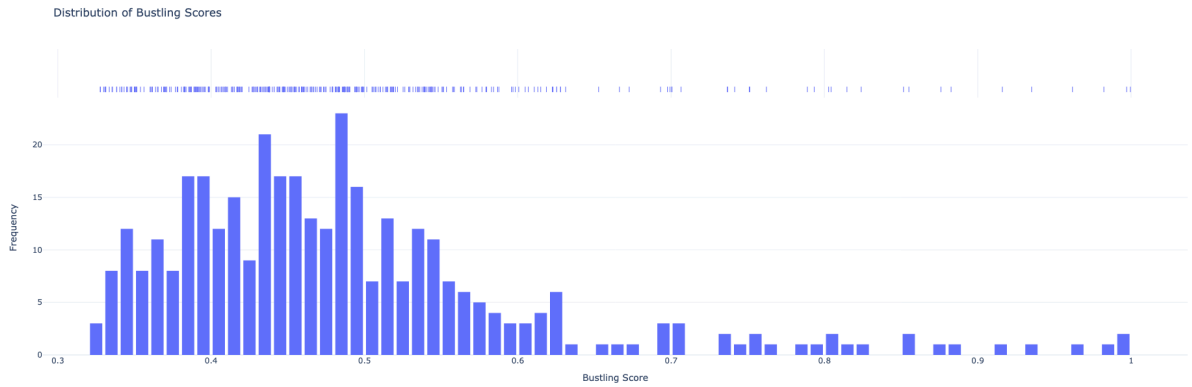
found the number of water fountains and bathrooms per SA2 and normalised by calculating the metric per square kilometre.

Z-Crossings

Pedestrian crossings are more likely to be in places where there is lots of pedestrian movement, and as we know from before: more movement = more 'bustle'. This is why it is another great metric for working out how 'bustling' an area is. Again we normalise this by calculating the metric per square kilometre and then find its corresponding z-score for each SA2.

Analysis of our final 'bustling' metric

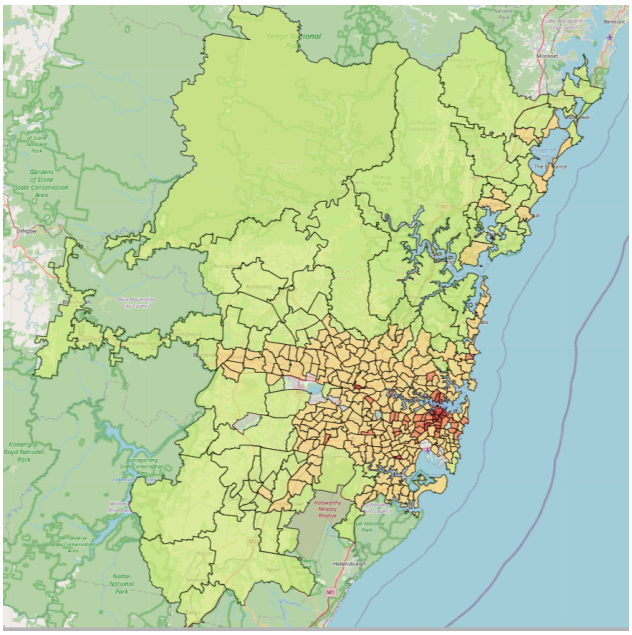
After calculating the bustling score of each SA2 region with our formula, we got the following distribution:



In Sydney, the bustling scores for different regions mostly fall between 0.3 and 1. Most of the scores are clustered between 0.4 and 0.6 and peak at around 0.5. This means most areas have a moderate level of activity.

The distribution somewhat resembles a normal curve but is a bit skewed to the left, with more scores in the 0.4 to 0.5 range. There are a few outliers, with some regions scoring really high (close to 1) but not many regions scoring very low, showing there are some very high bustling regions and not many regions of very low bustling (which is what should be expected since Sydney is the most active city in Australia). Beyond the central cluster, the frequency of scores drops significantly, and scores above 0.7 are less common. Overall, most regions have moderate bustling scores between 0.4 and 0.6, indicating moderate activity levels across Sydney, with a slight lean towards lower scores.

In addition, a graph was also made that colours each SA2 region according to its bustling score.



The redder a region is, the higher the score. What we can see is expected when mapping 'bustling' regions. The regions that are closer to the Sydney CBD are far more bustling whereas the regions on the outskirts are far less bustling. Areas closer to the CBD are more densely populated and therefore have more stops, polls, businesses, etc. per square kilometre leading to our high bustling score.

Correlation Analysis

	correlation_coefficient
0	0.158244

When comparing how bustling a region is with its median income, we found a correlation coefficient of around 0.158. This means there's a fairly weak positive link between these two factors. So, while there is some relationship between how lively a place is and the median income there, it's not a strong predictor.

	SA2_CODE21	SA2_NAME21	median_income	bustling_score
0	120021389	Lilyfield - Rozelle	88220	0.568858
1	120021387	Balmain	87932	0.529469
2	117031330	Erskineville - Alexandria	87640	0.595824
3	121041417	North Sydney - Lavender Bay	85147	0.855091
4	118011347	Woollahra	84677	0.697588
...	...	...	...	...
354	125011583	Auburn - North	39571	0.614967
355	119021574	Wiley Park	39550	0.519122
356	119021573	Lakemba	39413	0.564759
357	125011582	Auburn - Central	38824	0.513616
358	117031645	Sydney (South) - Haymarket	35875	0.997146

The median income of an area might not always match up with how bustling it is for several reasons (some regions with a high median income have lower bustling scores compared to regions in CBD, which was illustrated by the table). Tourist attractions, busy commercial districts, big public transportation hubs, universities, and frequent cultural events can make an area lively even if people living there don't earn a lot. Moreover, places known for shopping, dining, and nightlife, high population density, and investments in infrastructure and public spaces by the government can all contribute to an area's hustle and bustle, regardless of how much the locals make. So, many different factors can make a place busy and lively beyond just the income levels of its residents.

Moreover, the bustling scores give us a good idea about the activity levels in different regions, but they're not great at predicting economic factors like median income. This weak correlation shows that just looking at how bustling an area is doesn't fully explain its economic situation.