Dataset Description

***SA2 Regions:***

*Description:* The geospatial dataset holds information regarding the Australian statistical areas. Primarily the columns of ‘sa2\_code21’, ‘sa2\_name21’ and ‘geom’ identify the unique code, name and enclosed coordinate multi polynomial for each level 2 statistical area in australia.  *Transformations:* The SA2 regions dataset underwent three transformations. First the column names were conformed to lowercase. Second, rows with NA values within the ‘geom’ column were dropped. Third, filtering was performed on ‘ste\_name21' to equal 'New South Wales' and 'gcc\_name21' to equal 'Greater Sydney'. This limits regions to the greater Sydney area. Finally the ‘create\_wkt\_element’ function from the tutorials was used on the ‘geom’ column.

***Business:***

*Description:* The dataset holds information regarding the number of businesses of varying size across SA2 regions in Australia.                                                                                    *Transformations:* The Business dataset was processed by matching the sa2\_name column with the sa2\_name column from the cleaned Greater Sydney and deleting all other rows so only data regarding Greater Sydney was left over.

***Public Transport Stops***:

Description: This dataset contains the locations of all public transport stops (train and bus) in General Transit Feed Specification (GTFS) format

Transformations: The ‘longitude’ and ‘latitude’ columns were used to construct a new column ‘geom’ of POINT objects (using points\_from\_xy()) for each public transport stop. Then the WKTElement() function was applied to the geom point column. This column was then used to merge with the SA2 areas dataset within the SQL query.

***Polling***:

Description: This dataset contains the locations (and other premises details) of polling places for the 2019 Federal election

Transformations: The ‘longitude’ and ‘latitude’ columns were used to construct a new column ‘geom’ of POINT objects (using points\_from\_xy()) for each polling location. Then the WKTElement() function was applied to the geom point column. This column was then used to merge with the SA2 areas dataset within the SQL query.

***School***:

Description: the dataset holds information regarding the geographical regions in which students must live to attend primary, secondary and future Government schools.

Transformations: A ‘school’ dataset was created through the concatenation of the ‘catch\_future’, ‘catch\_primary’ and ‘catch\_secondary’ datasets from the school intake zones dataset webpage. An additional check on the ‘geom’ column was performed for NA values before then being transformed by the ‘create\_wkt\_element’ method from the tutorials.

***Population***:

Description: this dataset contains the estimates of the number of people living in each SA2 by age range.

Transformations: The ‘population’ dataset was initially transformed by renaming the column names to fit postgres column naming conventions. Then a ‘total\_children’ column was created by adding columns 'age\_0\_4\_people', 'age\_5\_9\_people', 'age\_10\_14\_people', and 'age\_15\_19\_people' (this includes all necessary information for the per 1000 young people statistic).

***Income***:

Description: This dataset contains the total earnings statistics by SA2 area

Transformations: The income dataset has columns ‘mean\_income’ and ‘median\_income’ that were coerced into numeric types. Each of those columns were checked for NAs. Finally the dataset was merged with the resulting data frames that held the SA2 area scores.

Additional Datasets:

***House Prices:***

Description:The data set looks at all the suburbs in Sydney and their median house price.The data source for the dataset regarding Sydney house prices is from Sydney Suburb Review and the dataset was obtained from Kaggle.The dataset was cleaned by dropping unwanted columns and then the data was matched to the greater Sydney region dataset by matching the column names suburb names to sa2\_name and thensa2\_name and sa2\_code were  joined with gdf dataset

***Traffic Incidents***

Description

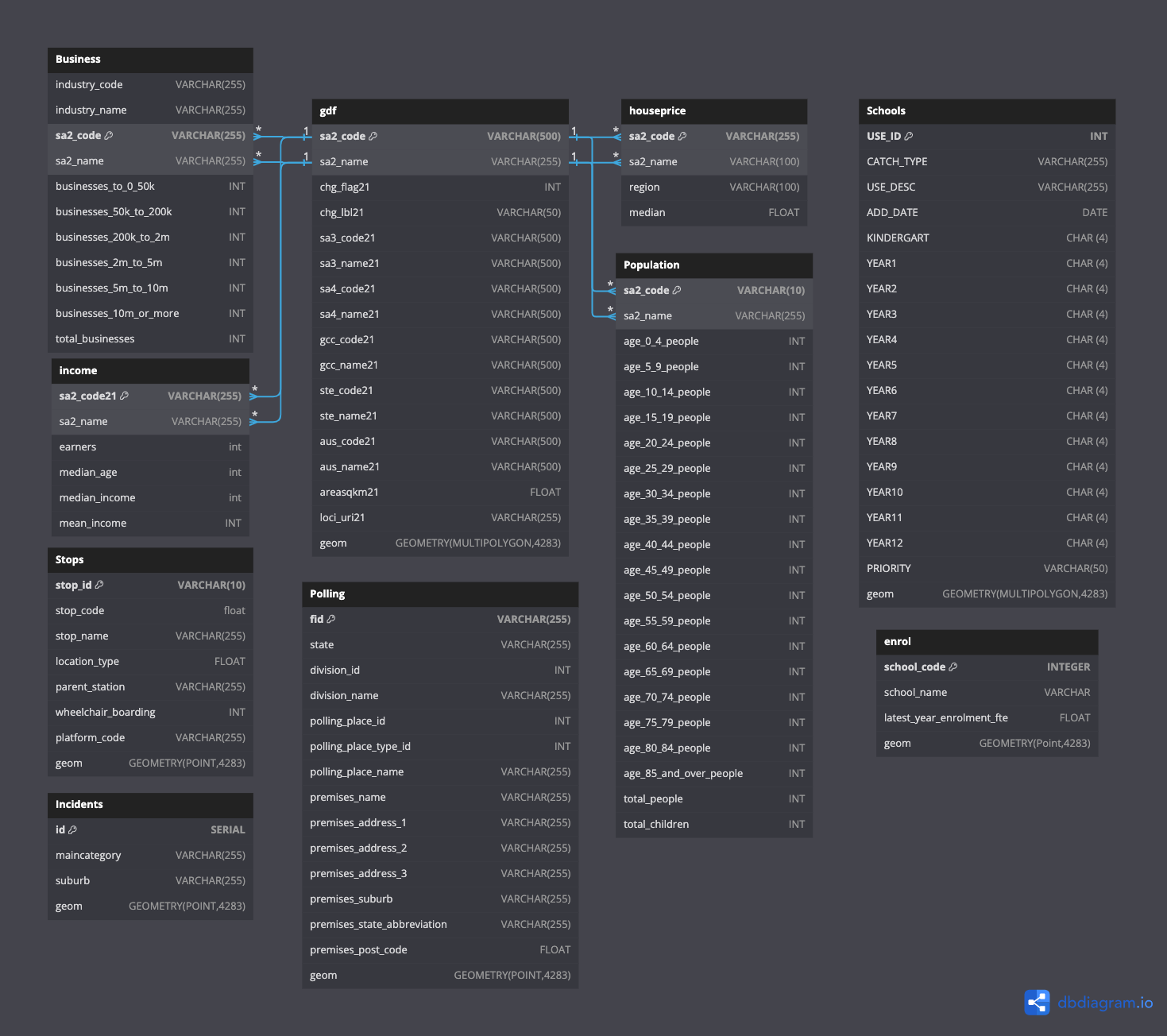
Transformations

***School Enrollments***

Description

Transformations

Database Description



Integration

The primary integration between the datasets in the database occurred during the querying process. Of the datasets which did not have an explicit sa2\_code or sa2\_name attribute, a PostGIS SQL Join was used (ST\_Contains(), ST\_Intersects or ST\_Within()). This would allow for the geom attributes within each dataset to then be used as a joining condition with the geom attribute in the gdf dataset containing the sa2 areas.

Indexes

Two indexes were created in an attempt to speed up the execution time of the queries. First I selected to create an index on the gdf dataset specifically on the geom column. This was because almost every table would be required to join the gdf table using the geom column. The second index I created was on the stops table also using the geom table. This was selected as the stops table had over 100000 tuples which slowed down the query time. The two index creation SQL codes are shown below.

1. “CREATE INDEX gdf\_geom\_index ON gdf USING GIST (geom);”
2. “CREATE INDEX stops\_geom\_index ON Stops USING GIST (geom);”

Score Analysis

A total of four scores were computed based on two scoring functions and applied first on the original task 2 datasets and then on task 3 datasets. First the sigmoid function was used to compute scores. This involved finding the sum of the z scores from each of the datasets before then computing the sigmoid value of the sum for each SA2 area. This was performed both for the original 4 datasets (school, polls, businesses and stops) and then on the additional three (school, polls, businesses, stops, school enrollments, house prices and traffic accidents). An alternative scoring function was used as the sigmoid function produced a large quantity of scores around 0.5 making the map interpretation difficult. So, the RELU function was used instead (MAX(0,x)) again on both the original data sets and additional. The SQL equations for each of these computations are shown below.

**Original Sigmoid**

1/(1+EXP(-(sc.stop\_count\_zscore + sa.school\_area\_1000\_zscore +pc.polling\_count\_zscore + tb.total\_businesses\_zscore)))

**Original RELU**

relu((sc.stop\_count\_zscore + sa.school\_area\_1000\_zscore + pc.polling\_count\_zscore + tb.total\_businesses\_zscore + COALESCE(hp.total\_median\_zscore, 0) + se.school\_enrollment\_count\_zscore + ti.incidents\_count\_zscore))

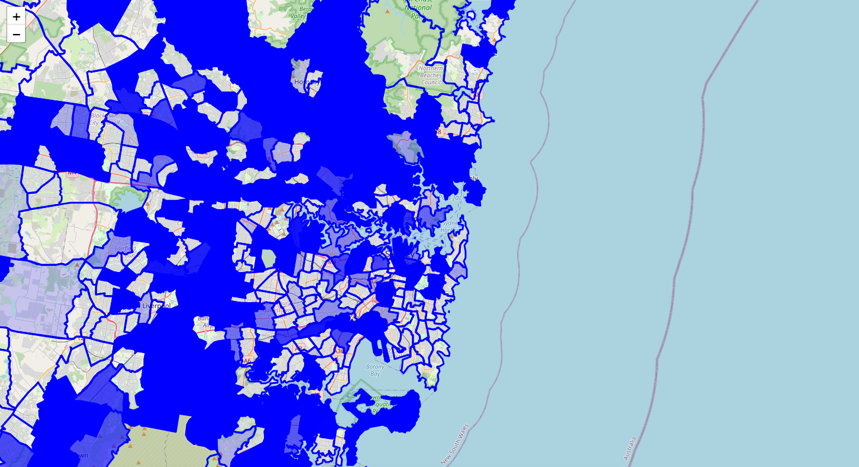
**Additional Sigmoid**

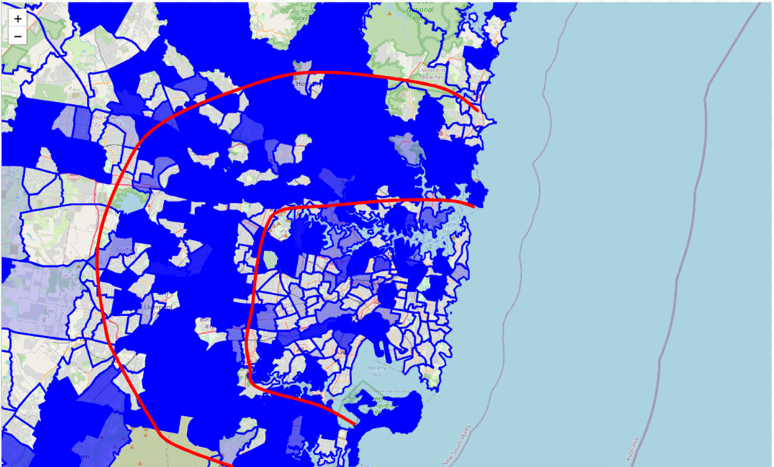
1/(1+EXP(-(sc.stop\_count\_zscore + sa.school\_area\_1000\_zscore + pc.polling\_count\_zscore + tb.total\_businesses\_zscore + COALESCE(hp.total\_median\_zscore, 0) + se.school\_enrollment\_count\_zscore + ti.incidents\_count\_zscore)))

**Additional RELU**

relu((sc.stop\_count\_zscore + sa.school\_area\_1000\_zscore + pc.polling\_count\_zscore + tb.total\_businesses\_zscore + COALESCE(hp.total\_median\_zscore, 0) + se.school\_enrollment\_count\_zscore + ti.incidents\_count\_zscore))

Score Results

Below is an interactive map visualization of greater Sydney and the ‘bustleness’ metric overlaid. The more intense the blue signifies the more busy an SA2 region. On initial glance there seems to be a non-concentrated distribution of busyness throughout greater sydney. With intense blue SA2 regions appearing consistently regardless of North, South, East or West orientations and centrality. With that being said there is a somewhat distinguishable trend that shows an outer ring of Sydney which contains a large proportion of highly busy areas (contained within the redlines in the second image).



This trend has likely emerged for a number of reasons. First, the scores were based, in part, on two datasets regarding school enrollment and school area per 1000 students. This would likely bias the score towards SA2 areas with larger schools which are more likely to be outside the city center. Furthermore, it may also be the case that this outer ring would likely hold a larger total number of traffic accidents due to the higher driving speeds and less congestion. Finally, total businesses is also a large factor, it may be that a larger total number of businesses exist in this outer ring whereas the city may hold a larger amount of

more valuable businesses but fewer overall. All of these factors likely have led to the unintuitive result. This rationale is also validated by the top and bottom 10 SA2 areas as shown in the following tables.

Top 10 Most Busy SA2 Areas (Sigmoid)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sa2 code21** | **Sa2 name21** | **Stop count zscore** | **School area 1000 zscore** | **Polling count zscore** | **Total businesses zscore** | **Total median income zscore** | **School enrollment count zscore** | **Incidents count zscore** | **Final score Sigmoid** |
| 117031644 | Millers Point | 0.34 | -0.11 | 14.54 | 21.54 | 0.00 | -1.13 | 6.91 | 1 |
| 127011592 | Badgerys Creek | -1.61 | 16.56 | -0.99 | -0.79 | 0.00 | 0.00 | -0.70 | 0.999 |
| 117031645 | Haymarket | -0.84 | -0.12 | 5.56 | 4.16 | 0.00 | 0.00 | 0.90 | 0.999 |
| 126011496 | Cheltenham | 0.77 | -0.12 | 0.47 | 0.48 | 0.00 | 3.16 | 3.95 | 0.999 |
| 121041417 | Lavender Bay | -0.33 | -0.12 | -0.02 | 3.16 | 0.00 | 0.18 | 5.73 | 0.999 |
| 115021297 | Wisemans Ferry | 6.33 | -0.06 | 1.20 | 1.46 | 0.00 | -0.41 | 0.02 | 0.999 |
| 115011291 | Bella Vista | 1.95 | -0.12 | 0.23 | 2.65 | 0.54 | 1.60 | -0.22 | 0.998 |
| 102011030 | Kulnura | 0.26 | 0.64 | 0.71 | -0.07 | 0.00 | -1.46 | 6.53 | 0.998 |
| 115011290 | Baulkham Hills | 1.53 | -0.12 | 1.68 | 0.46 | -0.17 | 2.67 | 0.18 | 0.998 |

Top 10 Least Busy SA2 Areas (Sigmoid)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sa2 code21** | **Sa2 name21** | **Stop count zscore** | **School area 1000 zscore** | **Polling count zscore** | **Total businesses zscore** | **Total median income zscore** | **School enrollment count zscore** | **Incidents count zscore** | **Final score Sigmoid** |
| 118021654 | South Coogee | -1.16 | -0.12 | -0.50 | -0.46 | 0.00 | -1.11 | -0.74 | 0.016 |
| 102021052 | Gwandalan | -1.04 | -0.11 | -0.50 | -0.60 | 0.00 | -1.19 | -0.74 | 0.014 |
| 116021562 | Acacia Gardens | -1.28 | -0.12 | -0.74 | -0.64 | 0.00 | -0.82 | -0.61 | 0.014 |
| 127031601 | Warwick Farm | -1.10 | -0.12 | -0.50 | -0.48 | -0.87 | -1.31 | 0.03 | 0.012 |
| 127031730 | Edmondson Park | -1.18 | -0.12 | -0.74 | -0.29 | -0.76 | -1.01 | -0.41 | 0.010 |
| 128011605 | Dolans Bay | -1.31 | -0.12 | -0.74 | -0.52 | 0.00 | -1.19 | -0.74 | 0.009 |
| 125011587 | Regents Park | -0.96 | -0.12 | -0.50 | -0.37 | -0.70 | -1.27 | -0.73 | 0.009 |
| 119041671 | Wolli Creek | -1.59 | -0.12 | -0.99 | -0.28 | -1.82 | 0.00 | 0.12 | 0.009 |
| 123011702 | Spring Farm | -1.18 | -0.12 | -0.74 | -0.54 | -0.90 | -0.73 | -0.74 | 0.007 |

For additional analysis two investigations were carried out, first was score to proximity to the CBD correlation and East vs West Sydney score t-test. Below is a plot of an area’s sigmoid function score and its distance to the CBD. The plot reveals a slight negative correlation between distance and score (-0.0016). Which aligned with the previously established outer ring theory.

A graph with blue dots

Description automatically generated

Furthermore, the second analysis which performed a t-test on the scores of areas within east and west sydney revealed no discernable difference between the regions scores. This likely also serves as evidence for the outer ring as the ring would include areas in both the east and west proportionally.

A screenshot of a graph

Description automatically generated

**Analysis Limitations**

The unintuitive results that gave rise to this outer ring outcome is likely the result of suboptimal dataset selection and scoring function selection. Although a true measure of busyness may be subjective, targeting more practical data such as foot traffic, vehicle traffic, tourism or other datasets may provide a more intuitive result. Furthermore, both the RELU and Sigmoid function implementation placed equal weighting to all of the datasets. Perhaps an adjusted weighting could be implemented on a dataset-by-dataset basis to prioritize which score is more impactful to ‘true’ busyness. Furthermore, the additional datasets did have data quality problems with the house price dataset only holding information for 134 SA2 areas with school enrolment having similar numbers. This likely biased the final results for excluded regions.

**Correlation Analysis**

The correlation coefficient  was utilized to observe how our score correlated with the median income of each SA2 region. The coefficient we got was 0.050244828347737254. This suggests that the relationship between our score and the median income is a very weak positive linear relationship meaning that a higher sigmoid score didn’t necessarily translate to a higher median income, nor can the sigmoid scores be used to predict or provide accurate information about the income of the sa2 regions. This may be due to the nature of our sigmoid score and the metrics of bustling City which were looked at. These results were surprising as we assumed the higher house prices and number of businesses which had higher revenue would have been associated with regions with higher median income, however it seems these factors were undermined by the other metrics of a bustling city we included. When certain metric’s by themselves are compared to the median income  the following  correlation coefficients are observed: -0.01924099627 for school enrolment, 0.0278698124 for traffic incidents, 0.3506018190 for House Prices and  -0.144061333 for public transport stops. This shows that while there was a moderate linear relationship between house prices and median income, all the other metrics had weak associations with median income. Therefore, if a metric did have a strong to moderate relationship with median income it was overshadowed by the majority weak associations all the other metrics had with median income.

Limitations within our dataset may also have contributed to the weak correlation coefficient by influencing the z-scores obtained in the first part of task 3 thus indirectly affecting sigmoid function scores and correlation to median income. The missing rows from house prices data likely skewed the house prices z-scores which would underestimate their impact on the sigmoid score and data quality issues from the school enrolment data may have potentially led to inaccurate z-scores and sigmoid scores therefore the correlation coefficient obtained may not potentially reflect the true relationship between the median income and the metrics selected.

**Appendix**

RELU Head (most busy areas)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sa2 code21** | **Sa2 name21** | **Stop count zscore** | **School area 1000 zscore** | **Polling count zscore** | **Total businesses zscore** | **Total median income zscore** | **School enrollment count zscore** | **Incidents count zscore** | **Final score RELU** |
| 117031644 | Sydney (North) - Millers Point | 0.34 | -0.11 | 14.54 | 21.54 | 0.00 | -1.13 | 6.91 | 42.09 |
| 127011592 | Badgerys Creek | -1.61 | 16.56 | -0.99 | -0.79 | 0.00 | 0.00 | -0.70 | 12.48 |
| 117031645 | Sydney (South) - Haymarket | -0.84 | -0.12 | 5.56 | 4.16 | 0.00 | 0.00 | 0.90 | 9.67 |
| 126011496 | Pennant Hills - Cheltenham | 0.77 | -0.12 | 0.47 | 0.48 | 0.00 | 3.16 | 3.95 | 8.72 |
| 121041417 | North Sydney - Lavender Bay | -0.33 | -0.12 | -0.02 | 3.16 | 0.00 | 0.18 | 5.73 | 8.61 |
| 115021297 | Dural - Kenthurst - Wisemans Ferry | 6.33 | -0.06 | 1.20 | 1.46 | 0.00 | -0.41 | 0.02 | 8.55 |
| 115011291 | Baulkham Hills (West) - Bella Vista | 1.95 | -0.12 | 0.23 | 2.65 | 0.54 | 1.60 | -0.22 | 6.63 |
| 102011030 | Calga - Kulnura | 0.26 | 0.64 | 0.71 | -0.07 | 0.00 | -1.46 | 6.53 | 6.61 |
| 115011290 | Baulkham Hills - East | 1.53 | -0.12 | 1.68 | 0.46 | -0.17 | 2.67 | 0.18 | 6.24 |

RELU Tail (least busy areas)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sa2 code21** | **Sa2 name21** | **Stop count zscore** | **School area 1000 zscore** | **Polling count zscore** | **Total businesses zscore** | **Total median income zscore** | **School enrollment count zscore** | **Incidents count zscore** | **Final score RELU** |
| 127011726 | Cobbitty - Bringelly | -1.07 | -0.07 | -0.74 | -0.49 | -0.85 | -1.23 | -0.66 | 0.00 |
| 127011727 | Gledswood Hills - Gregory Hills | -0.37 | -0.12 | -0.74 | -0.20 | 0.00 | -0.27 | -0.64 | 0.00 |
| 127011728 | Leppington - Catherine Field | 0.45 | -0.11 | -0.02 | -0.12 | -0.82 | -1.16 | -0.24 | 0.00 |
| 127011729 | Oran Park | -0.80 | -0.12 | -0.26 | -0.11 | -0.85 | 2.06 | -0.68 | 0.00 |
| 127021512 | Cabramatta West - Mount Pritchard | 0.85 | -0.12 | -0.02 | -0.13 | -0.64 | -0.02 | -0.11 | 0.00 |
| 127021514 | Edensor Park | -0.84 | -0.12 | -0.50 | -0.31 | 0.00 | -0.70 | -0.60 | 0.00 |
| 127021515 | Fairfield | -0.37 | -0.12 | 0.23 | 0.16 | -0.79 | 0.52 | -0.45 | 0.00 |
| 127021517 | Greenfield Park - Prairiewood | -0.55 | -0.12 | -0.74 | -0.44 | 0.00 | 0.72 | -0.73 | 0.00 |
| 127021518 | Horsley Park - Kemps Creek | 0.66 | -0.07 | -0.74 | -0.17 | 0.00 | -1.50 | 0.87 | 0.00 |

Correlation Coefficient

