INTRODUCTION

This report presents an analysis of three SA2 regions within New South Wales (NSW): Sutherland, Ryde, and Central Coast. Using z-scores for four key indications: schools, transportation stops, businesses, and points of interest (POI), a logistic score was calculated to assess the overall accessibility and desirability of each SA2 region. Spatial maps were generated to visualise the logistic score geographically. The purpose is to evaluate and compare each region's relative accessibility, service availability, and business and infrastructure density through a composite score derived from multiple data features.

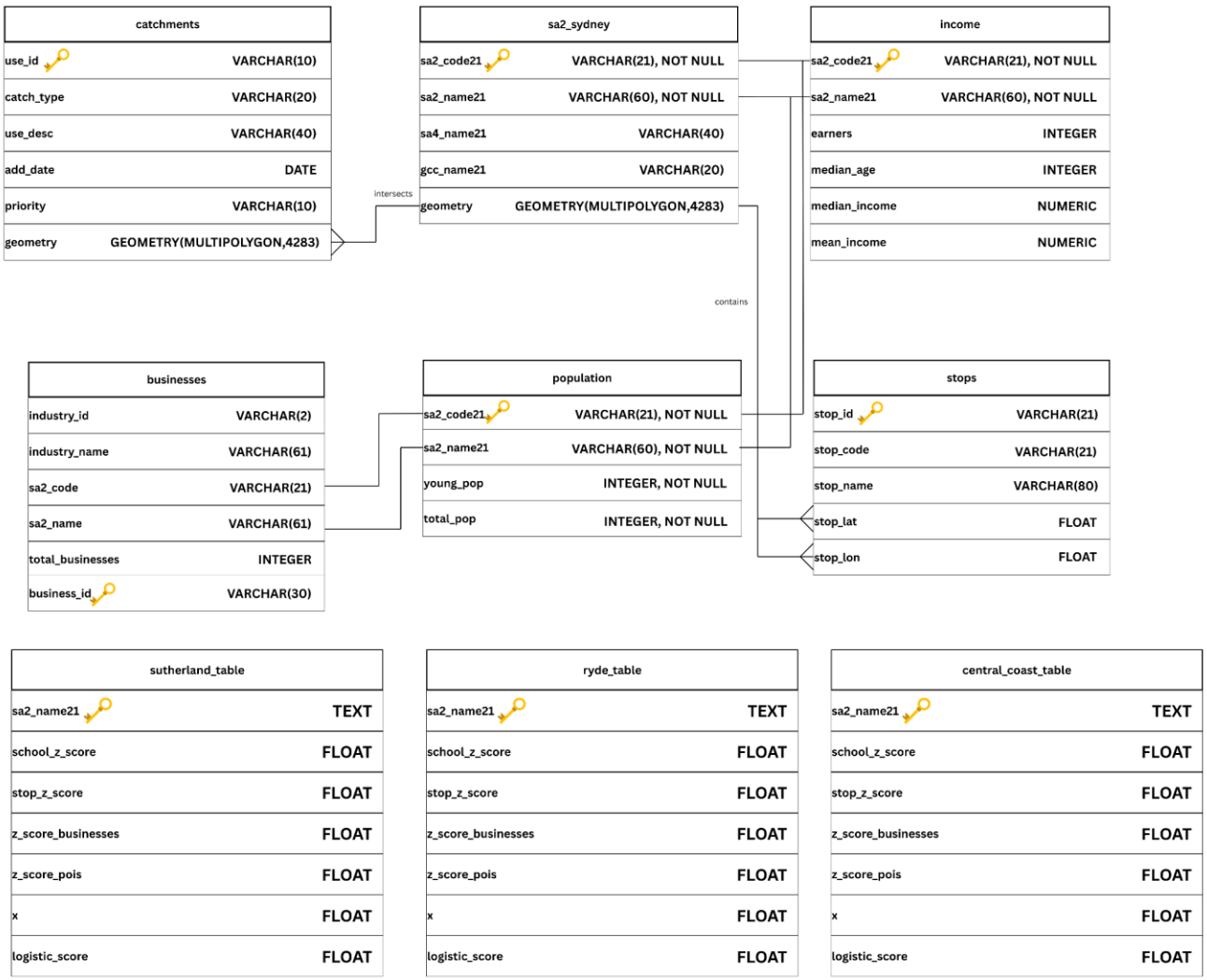
DATASET DESCRIPTION

This analysis was conducted using data from multiple sources, each of which have been processed and normalised.

Data sources	Description
School Locations	Point data indicating all primary and secondary schools within NSW. These were spatially joined to SA2 regions to generate school counts and z-scores. Sourced from NSW Department of Education, the data was collected from the government school enrolments census and the department's administrative records system.
Public Transport Stops	Stop locations from general transit feeds (GTFS) and filtered for stop types within the SA2 boundary, with z-scores calculated for their density. The data is sourced from Transport for NSW where the data was collected automatically by the TfNSW pipeline
Businesses	Business location data aggregated by the respective SA2 which provides a measure of commercial activity per density. This dataset is provided by the Australian Bureau of Statistics and combines annual administrative records from the Australian Business Register with business data from the Australian Taxation Office.
Points of Interest (POIs)	Locations such as parks, community centres, and libraries were all sourced and aggregated by SA2, informing social and cultural service ability. The data is generated from NSW Points of Interest API, which provides geolocated features for given POIs.
SA2 Boundaries	Spatial shapefiles for all SA2 regions in NSW from ABS, used for mapping and spatial joins.
Income	Overall data for each sa2 region. The original source is not provided.

For preprocessing, we replaced the missing value in the income table with the column mean so that the data is valid while preserving the overall variance. Next, for the business and population table, we filtered out rows having non-positive total business/population to ensure the data validity. We then added an extra column in the population table to store the young population. Furthermore, we drop entries with invalid geometries in the school and SA2 boundaries dataset, ensuring that subsequent spatial joins and calculations remain accurate. Moreso, when loading the school data, we merged the primary, secondary, and future school datasets into a single table and removed duplicate records based on use_id. This is because institutions serving both primary and secondary levels likely share the same catchment area and should be counted only once.

Database Description



As provided by the above database schema diagram, each table has a primary key to uniquely identify the entries. In particular, the **catchments** table uses ***use_id*** as its primary key, while the **sa2_sydney**, **income**, **population**, and the **regional summary tables** all use ***sa2_code21*** as their primary key. Also, the **stops** table uses ***stop_id***. Furthermore, the **businesses** table utilises ***business_id*** which is a composite key of sa2_code and industry_code to uniquely identify each entry. Moreso, as indicated by the connecting line in the diagram, foreign keys were used to join different tables to produce an integrated summary.

Additionally, we implemented spatial indexes in the column geometry in both catchments and sa2_sydney tables. With these indexes in place, the following spatial queries such as ST_Contains and ST_Intersect can be executed efficiently. We joined sa2_sydney with stops using ST_Contains because the stop's position is a point and we count the number of stops inside a given SA2 region. We used ST_Intersects for joining catchments and sa2_sydney because each school can recruit students across multiple SA2 regions and the catchment area should only include the intersected region within the SA2 boundaries.

Score Analysis

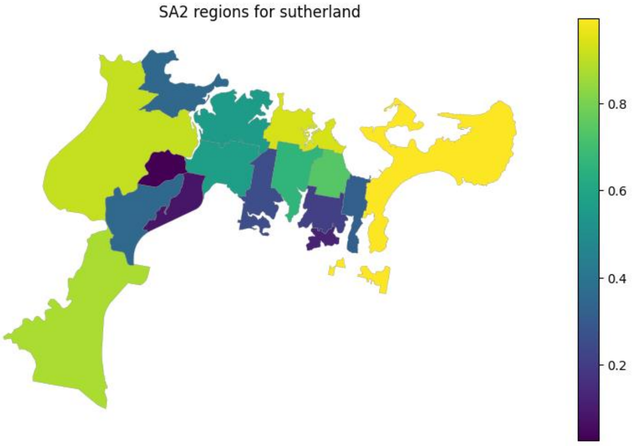
Selection of industries and POI group

In this report, we included all industries and POI within each SA2 region because each region may emphasise different industries and POIs. By having the full dataset, it provides an comprehensive and unbiased view of each SA2 area, allowing consistent and meaningful comparisons across regions.

For each selected SA4 region, we sum the z scores for businesses, stops, schools, and POIs and apply the sigmoid function to the sum to compute each SA2 region's score, as indicated in the logistic_score column of the following tables. The sigmoid function is $e^x/(1 + e^x)$ which provides diminishing returns bounded by 0 - 1, reducing the impact of extreme values and allowing insightful comparison amongst regions.

SUTHERLAND

The distribution within Sutherland shows a mixed spread of scores across its SA2 regions with high performing zones such as Cronulla, Kurnell, and Bundeena exhibiting consistently positive z-scores across all categories. This is exemplified by points of interest and transportation stops with its range being the highest within Sutherland, contributing to a logistic score close to 0.99. However, low scoring regions like Woronora Heights held strongly negative z-scores across all data sources, despite having moderate school access, resulting in a logistic score as low as 0.02.



The trends and insights found within the data showcase that regions with balanced scores across multiple metrics, such as Cronulla and Sylvania, performed consistently well, which could indicate the high level of government funding or higher tax residential payment area. Meanwhile, Heathcote and Waterfall exhibited the highest individual z-score for schools, highlighting exceptional educational infrastructure, but had low scores in other categories like businesses and POIs, ultimately reducing its logistic ranking. Furthermore, Loftus, Yarrawarrah, Gymea and Grays Point displayed consistently negative or low z-scores across all variables, indicating potential underinvestment within the areas in terms of infrastructure and businesses.

Impact of Components:

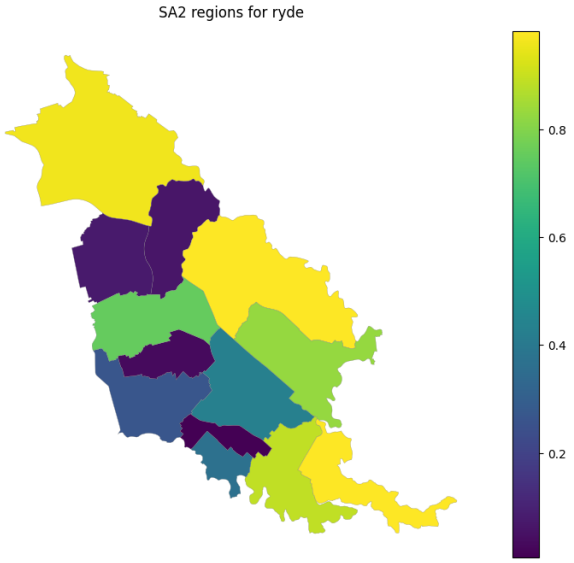
- The z-score for schools played a significant role in separating higher-tier suburbs like Heathcote and Menai from others.
- Transport stop density was a differentiating factor for suburbs nearer to rail lines or bus hubs.
- POIs such as parks and libraries made notable contributions in highly walkable areas like Cronulla.

sa2_name21	school_z_score	stop_z_score	z_score_businesses	z_score_pois	x	logistic_score
Cronulla - Kurnell - Bundeena	0.616472	1.323554	0.542404	2.580589	5.063018	0.993713
Sylvania - Taren Point	-0.377079	0.161964	1.20838	1.667695	2.66096	0.934683
Menai - Lucas Heights - Woronora	0.378459	1.509981	-0.72543	1.135174	2.298185	0.908727
Heathcote - Waterfall	3.666378	-0.669791	-1.060143	-0.043981	1.892464	0.869036
Caringbah	-0.453604	-0.024463	2.156969	-0.633558	1.045344	0.73988
Miranda - Yowie Bay	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.660194
Sutherland - Kirrawee	-0.40342	0.635204	-0.113993	0.155715	0.273506	0.567953
Oyster Bay - Como - Jannali	-0.429131	1.4813	-0.858739	0.041603	0.235034	0.558489
Engadine	-0.342719	1.00806	-1.069786	-0.215148	-0.619592	0.349874
Illawong - Alfords Point	-0.052478	-0.469022	0.075691	-0.177111	-0.62292	0.349118
Woolooware - Burraneer	-0.408504	-0.296935	0.568738	-0.614539	-0.751239	0.320551
Gymea - Grays Point	-0.494418	0.377074	-0.535993	-0.414844	-1.068181	0.255749
Caringbah South	-0.506676	-0.339957	0.053881	-0.519446	-1.312197	0.212119
Lilli Pilli - Port Hacking - Dolans Bay	-0.467251	-1.644952	1.160621	-1.004421	-1.956003	0.1239
Loftus - Yarrawarrah	-0.314107	-0.096166	-1.179047	-0.852272	-2.441593	0.080056
Woronora Heights	0.067421	-1.558908	-1.024607	-1.099514	-3.615608	0.026196

Sutherland's scores demonstrate how specific infrastructure can elevate a suburb, but balanced accessibility across all features is critical for higher overall logistic performance.

RYDE

Ryde is the most consistently highest scoring SA2 region in comparison to Sutherland and Central Coast. Its suburbs, Hunters Hill - Woolwich and Macquarie Park - Marsfield, are within the top end of the logistic scale with scores above 0.98, especially concerning business and POI density.



The trends and insights found that areas with high business and transport z-scores like Macquarie Park are highly favoured in the logistic transformation, whilst Epping has a positive school z-score and business access, but is impacted by negative POI scores. Ryde - South and Denistone are among the lowest ranked SA2s in the region with negative z-scores in almost every category, particularly in school and transportation stop access.

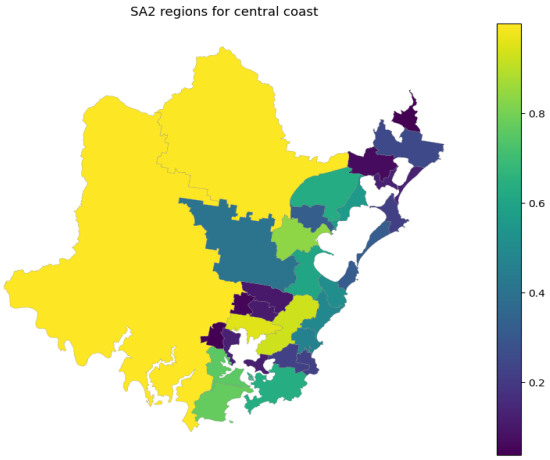
Impact of Components:

- Business density is a clear differentiator here; Ryde's commercial zones significantly outperformed others.
- Transport z-scores contributed heavily to Eastwood and Pennant Hills' rise in the ranks.
- Schools and POIs provided a smaller but still meaningful influence, especially in suburban areas like Putney.

In summary, with Ryde having the highest positive z-scores due to its high urban density and consistent access to infrastructure in between Sutherland and Central Coast. However, a few outlier regions reveal areas where transportation infrastructure would need improving.

sa2_name21	school_z_score	stop_z_score	z_score_businesses	z_score_pois	x	logistic_score
Hunters Hill - Woolwich	0.31407	-0.023218	1.506601	2.162914	3.960367	0.9813
Macquarie Park - Marsfield	2.032475	1.471593	-0.216723	0.588747	3.876091	0.979689
Pennant Hills - Cheltenham	1.842532	0.939785	-0.470851	0.869202	3.180668	0.9601
Gladesville - Huntleys Point	-0.229394	-0.253189	1.199444	1.357736	2.074597	0.88841
North Ryde - East Ryde	0.587201	0.73856	0.058229	0.172588	1.556577	0.825862
Eastwood	-0.23594	1.05477	0.680661	-0.406416	1.093074	0.74896
Ryde - North	-0.457931	1.069144	-0.839334	-0.07168	-0.299801	0.425606
Putney	0.078421	-1.460536	1.864044	-1.030655	-0.548726	0.36616
West Ryde - Meadowbank	-0.573603	0.120514	-0.412217	-0.144055	-1.009361	0.267105
Epping (NSW) - West	-1.064626	-0.16695	-0.330107	-0.967326	-2.52901	0.073849
Epping (East) - North Epping	-0.484572	-0.411294	-1.122533	-0.732106	-2.750504	0.060058
Denistone	-0.075552	-1.546775	-0.949594	-0.940185	-3.512106	0.02897
Ryde - South	-1.733081	-1.532402	-0.967618	-0.858763	-5.091865	0.006109

CENTRAL COAST



Central Coast exhibited the broadest variation in logistic scores with Calga and Kulnura emerging as the top two performing SA2 regions across all three LGAs with a logistic score of ~0.999. This is due to their extremely high business and POI z-scores, indicating high funding and infrastructural support.

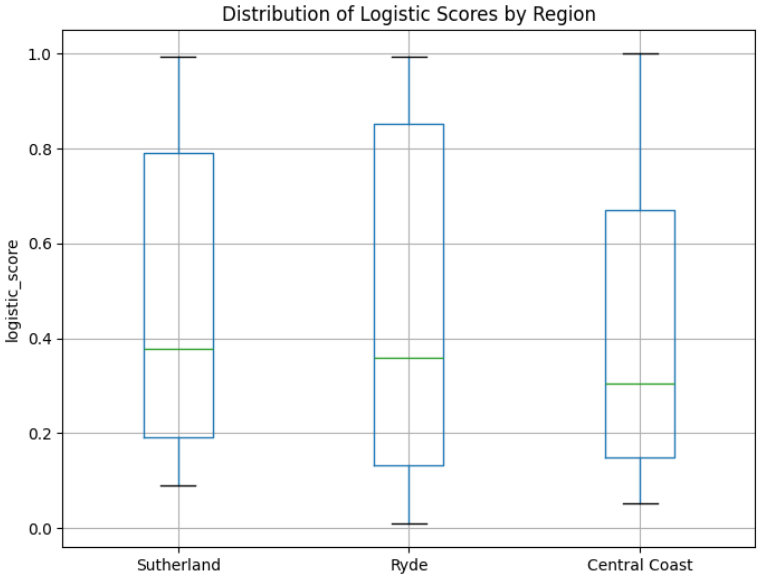
The trends and insights within the Central Coast can be observed within the northern and western areas such as Gosford, Springfield, Jilliby and Yarramalong with the suburbs scoring well due to strong transport and POI accessibility. Likewise, southern inland regions like Kariong, Narara, Summerland Point and Gwandalan consistently scored the lowest (<0.06) which could be attributed to the poor access to schools and fewer POIs and transportation stops. Terrigal and Erina had moderate high schools and stop z-scores but lacked in business and POI features, leading to its average score.

Impact of Components:

- POIs had a significant impact due to their wide variation across the Central Coast. Suburbs with active centres or tourism-oriented infrastructure outperformed residential-only zones.
- Business and stop z-scores were often strong predictors of whether an area broke into the top tier.
- Some areas had school z-scores below -0.2, reflecting fewer education institutions within regional inland zones.

In summary, the central coast is an example of a volatile region with it holding the highest scoring SA2 across all LGAS, but also having the most SA2s with scores under 0.2. This could be attributed to infrastructure concentration in certain coastal pockets skewing the distribution heavily.

sa2_name21	school_z_score	stop_z_score	z_score_businesses	z_score_pois	x	logistic_score
Calga - Kulnura	4.8198	-0.308218	3.70423	4.204269	12.42008	0.999996
Jilliby - Yarramalong	2.172393	0.385273	0.796835	2.097493	5.451994	0.995731
Gosford - Springfield	-0.283631	1.806931	0.685837	0.773535	2.982671	0.951785
Erina - Green Point	-0.240832	0.997857	0.908238	0.852342	2.517605	0.925367
Tuggerah - Kangy Angy	-0.183968	-0.319777	2.381606	-0.250957	1.626904	0.835745
Umina - Booker Bay - Patonga	-0.278016	2.165235	-0.512883	-0.140627	1.233709	0.774467
Woy Woy - Blackwall	-0.270119	1.229021	-0.59687	0.773535	1.135567	0.756865
Box Head - MacMasters Beach	-0.22524	0.177226	0.172554	0.416276	0.540816	0.632002
Warnervale - Wadalba	-0.265605	1.194347	-0.670181	0.26917	0.527729	0.628953
Chittaway Bay - Tumbi Umbi	-0.270338	1.286812	-0.147849	-0.445347	0.423278	0.604267
Gorokan - Kanwal - Charmhaven	-0.294816	1.541092	-0.894331	-0.114358	0.237587	0.559119
Bateau Bay - Killarney Vale	-0.296614	1.217463	-0.582617	-0.319256	0.018975	0.504744
Wamberal - Forresters Beach	-0.278397	0.073202	0.33593	-0.182657	-0.051922	0.487022
Terrigal - North Avoca	-0.290334	0.223458	0.407579	-0.508393	-0.16769	0.458175
Ourimbah - Fountaindale	0.275042	-0.828337	0.344346	-0.177403	-0.386353	0.404596
Wyong	-0.26657	-0.088613	0.043547	-0.429586	-0.741221	0.322737
The Entrance	-0.279219	-0.134846	-0.176688	-0.17215	-0.762902	0.318016
Lake Munmorah - Mannering Park	-0.204425	0.038527	-0.824501	-0.109104	-1.099503	0.249833
Avoca Beach - Copacabana	-0.289031	-0.828337	0.51428	-0.629231	-1.232319	0.225776
Kincumber - Picketts Valley	-0.266207	-0.678081	-0.036774	-0.292987	-1.274049	0.218565
Toukley - Norah Head	-0.276629	-0.285102	-0.377143	-0.340271	-1.279145	0.217696
Point Clare - Koolewong	-0.279704	-0.654964	-0.455038	-0.534662	-1.924368	0.127375
Budgewoi - Buff Point - Halekulani	-0.281395	-0.412242	-0.759209	-0.47687	-1.929716	0.126782
Saratoga - Davistown	-0.291042	-1.140408	-0.07542	-0.508393	-2.015263	0.11761
Wyoming	-0.289565	-0.458475	-0.715862	-0.666007	-2.129909	0.106224
Niagara Park - Lisarow	-0.261775	-1.105734	-0.434715	-0.392809	-2.195032	0.100197
Blue Haven - San Remo	-0.268558	-0.839895	-1.001884	-0.54517	-2.655507	0.06565
Narara	-0.283356	-1.094176	-0.794181	-0.713291	-2.885004	0.0529
Kariong	-0.281576	-1.544945	-0.564155	-0.765829	-3.156505	0.040836
Summerland Point - Gwandalan	-0.270271	-1.614294	-0.67468	-0.671261	-3.230506	0.038034



The boxplot showcases the spread and central tendency of logistic scores for each SA2 region with the three LGAs: Sutherland, Ryde, and Central Coast.

Sutherland can be observed to have a median score of approximately 0.38, indicating a slightly above average accessibility overall whilst its spread consists of a fairly wide distribution, ranging from approximately 0.08 to 1. Sutherland's boxplot highlights a balanced spread with a long upper tail which suggests that the LGA contains a few high performing areas, but most regions cluster around the middle. It also has moderate variation with it not being the most stable or the most extreme.

Similarly to Sutherland, Ryde's median score is approximately 0.36 with it having the widest spread amongst Sutherland and the Central Coast with scores ranging from approximately 0.01 to 1. This reveals that although some regions in Ryde perform extremely well, others perform quite poorly, indicating a large disparity within the LGA. It's also the least consistent LGA with large gaps between the SA2s.

However, Central Coast has the lowest median score out of the LGAs with it being approximately 0.30, showcasing that most areas around the region rank lower in terms of combined infrastructure and amenity accessibility. Central Coast's spread is fairly wide with the range being approximately 0.05 to 1.0 highlight that while some of the highest and lowest scoring SA2s, many areas are towards the lower end which pulls down the median. Its consistency is moderate with a clear lean towards underperforming zones.

CORRELATION ANALYSIS

Region	Correlation Coefficient	Interpretation
Sutherland	–0.470	Moderate negative correlation
Ryde	0.331	Moderate positive correlation
Central Coast	–0.130	Weak negative correlation

The statistical test for correlation between computed scores and the area's median outcome revealed Sutherland's negative correlation which suggests that higher logistic scores may be found in lower income areas. A possible reason for this would be the accessibility of good public services and POIs in density heavy suburbs such as Cronulla but these areas have relatively lower median income compared to Sydney averages; suggesting that infrastructure is more equitably distributed.

Furthermore, Ryde's positive correlation indicates that the wealthier areas tend to have better infrastructure and amenity access, suggesting that wealth is the main motivator to reinforce accessibility advantages within the region. This makes sense as within urban planning, higher income suburbs often attract more development as seen in suburbs: Macquarie Park and Hunters Hill.

While the negative correlation for the Central Coast is lower than Sutherland, it still suggests that higher income areas tend to have better infrastructure. However, the regions show more variability in terms of the relationship between income and accessibility with some high scoring SA2s not necessarily being wealthier. Hence, while income is a factor reflected in accessibility, regional planning and the clustering of services also influence scores.

In summary, the magnitudes of all 3 correlation coefficients is greater than $\alpha(0.05)$, which fails to reject the null hypothesis of no linear relationship between median income and score. However, the fact that 2 regions show negative trends while one shows positive suggests regional heterogeneity in how income relates to the sigmoid score, and it requires further evidence to discover how they are correlated in general.

Overall, the logistic score provides an insightful and intuitive ranking of SA2 regions and simplifies comparisons. However, it equally weighs all four z-scores and among them, schools and business are density measures that are prone to distortion in regions with very low populations or large geographic areas. For example, the sum of the z-scores for Calga is 12.4, which is significantly higher than others. Therefore, for further study, incorporating more z-scores from additional dimensions and applying differential weighting can produce a more balanced and objective summary of the region.