THE UNIVERSITY OF SYDNEY

DATA2901: Data Science, Big Data and Data Variety (Adv)

Practical Assignment Report:

PARRAMATTA & EASTERN SUBURB ANALYSIS

ADV 01 - Group 2

Bao Ngo - 540916379

Cong Thanh Vu - 530784058

1. DATASET DESCRIPTION

1.1. Data Sources

Table	Source	Description
sa2_boundaries	ABS SA2 shapefile via ABS website	The geographic boundaries of each SA2 region in Greater Sydney.
businesses	ABS Business Register	Number of businesses by SA2 and industry code.
stops	Transport for NSW Open Data	Public transport stops (buses/ trains) in GTFS format (TXT file).
catchments (school)	NSW Department of Education	Polygons defining catchment areas for public schools in NSW.
population	ABS Census 2021	SA2-level demographic data including total population and age-grouped population.
income	ABS Census 2021	Median income for each SA2_region.
sa2_pois	NSW POI API	POIs per SA2s (e.g. supermarkets, libraries, etc.)

1.2. Key Fields

Key fields retained across datasets include `sa2_code` identifier, `geom` for spatial joins, and numeric fields like `total_population`, `young_population`, `total_businesses`, and `median_income` for score calculations.

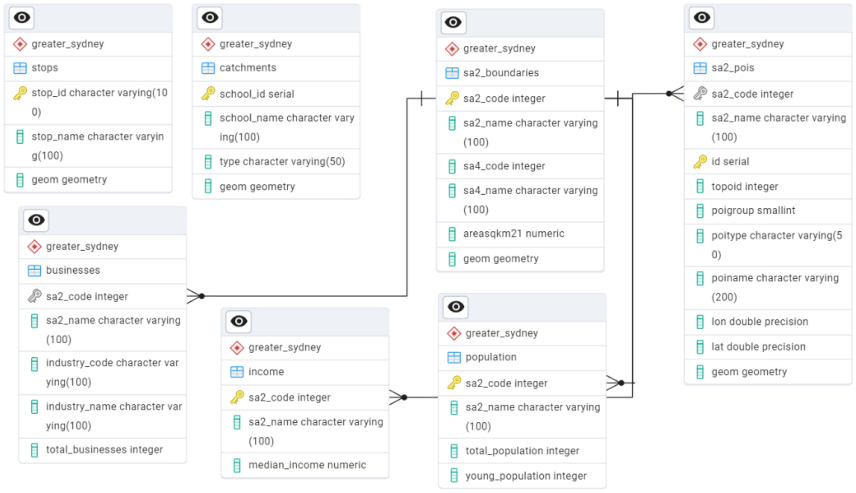
1.3. Data Preprocessing

SRID: all geom columns in the datasets were transformed to EPSG:4326 for compatibility with PostGIS functions (e.g. ST_Intersect(), ST_Contains, etc.)
Geographic Filter: restricted SA2 regions to Greater Sydney.
Type & Schema Alignment: pruned columns that are not needed for spatial joins and scoring to reduce table bloating, checked for ID uniqueness (duplicates), casted all ID to INT and renamed columns to match with database schema (e.g. “SA2_CODE21” to “sa2_code”).
Missing & Unassigned Records: dropped rows with non-numeric keys (e.g. “np”), and filtered child tables to only sa2_code(s) that are in the master sa2_boundaries.

2. DATABASE DESCRIPTION

2.1. Schema Overview

We designed a PostgreSQL database with PostGIS to support spatial integration and scoring. The schema includes the following key tables. All spatial tables use EPSG:4326, and `sa2_code` serves as the primary key for linking most tables. This structure enabled efficient spatial joins and metric computation for scoring and analysis.



Database Schema Diagram

2.2. Indexing Strategy

Spatial: to enable efficient spatial joins, GiST indexes were created for all geometry columns.
B-Tree: because PostgreSQL only automatically generates indexes on unique constraints or primary keys. B-Tree indexes were created on `sa2_code` of foreignly linked child tables to improve join & filtering operations.

3. RESULTS & CORRELATION ANALYSIS

3.1. Scoring Methodology

Given the **z-score** formula:

$$z = \frac{x - \mu}{\sigma}$$

x : the raw value for the SA2 (e.g., number of schools per 1000 youth)

μ : the mean value of that component across all SA2s within the same SA4,

σ : the standard deviation for the component within the SA4

This **z-score** formula standardized each component by converting them into a common scale (mean = 0, sd = 1). This ensures that each component contributes equally to the aggregated sigmoid score, preventing any metric from dominating due to its numerical magnitude. A positive z-score indicates above-average performance, while a negative score indicates below-average.

We use the sigmoid function to extend the z-score approach to produce a meaningful indicator for the well-resourcefulness of each SA2 region:

$$Score_{SA2} = S(z_{business} + z_{stops} + z_{schools} + z_{POIs})$$

- $S(x) = \frac{1}{1 + e^{-x}}$: the **sigmoid function** transforms the aggregated sum of z-scores into a value between 0 and 1.
- $z_{business}$: z-score of businesses per 1000 residents, focusing on the Transport, Postal and Warehousing industry.
- z_{stops} : z-score of the number of public transport stops within the SA2.
- $z_{schools}$: z-score of school catchments per 1000 'young population' (ages 0-19).
- z_{POIs} : z-score of number of points of interest within the SA2.

Because z-score standardization ensures that all components contribute equally, and one standard deviation of any component makes the same impact on the aggregated score. All components have proportional contribution and components with higher z-scores have greater influence on the aggregated sigmoid score.

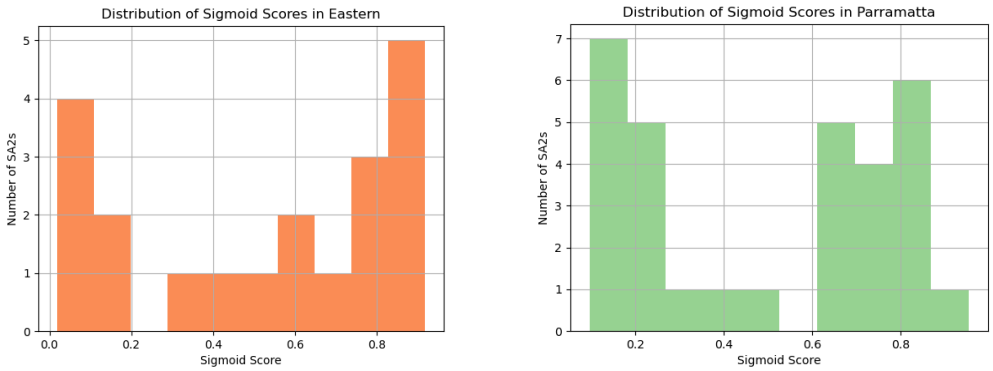
Moreover, the shape of the sigmoid function makes the overall score very sensitive around midpoint, allowing for clearer differentiation between moderately well-resourced SA2 regions. While at the extremes (0 and 1), the curve flattens, preventing very high or very low inputs to disproportionately influence the overall score. This provides stability and reduces the impact of outliers.

To provide meaningful per-capita calculations and fair comparisons, SA2 regions with a total population under 100 are excluded from the scoring process. Such areas often serve non-residential or special-purpose functions and do not reflect typical population-based dynamics central to our analysis. The excluded SA2s were: *Centennial Park* (Eastern Suburbs), *Rookwood Cemetery*, *Smithfield Industrial*, and *Yennora Industrial* (Parramatta).

It is also important to note that, since scores are calculated relative to other SA2s within a specific SA4 region, comparisons of SA2 scores are valid only within the SA4 and should not be used for cross-region evaluations.

3.2. Result Analysis

3.2.1. Distribution



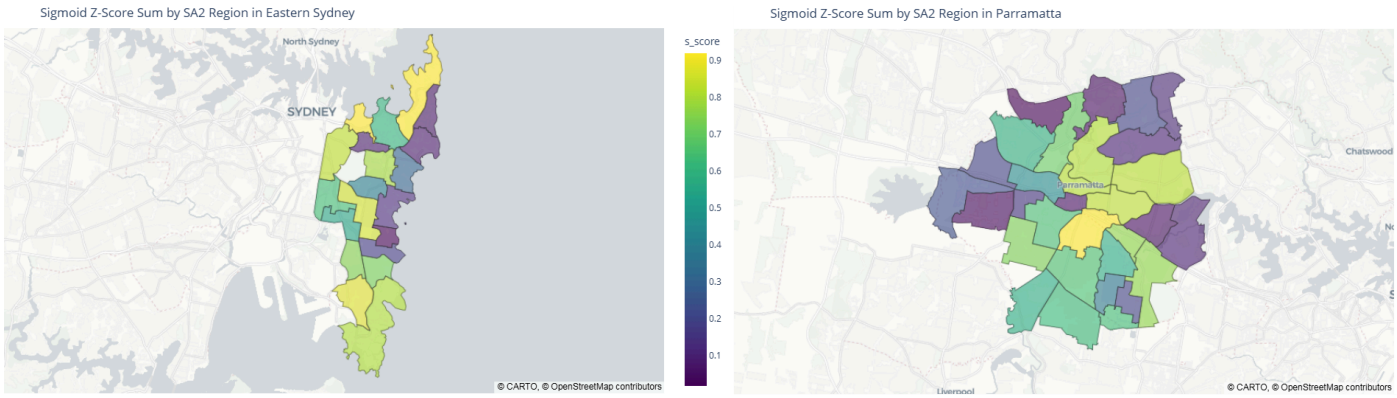
Sigmoid Score Distribution in Eastern Suburbs & Parramatta

The histograms above reveal a bimodal distribution of sigmoid scores for both **Eastern Suburbs** and **Parramatta**. This pattern suggests a polarisation in well-resourcefulness between SA2 regions within each SA4. Many SA2s scored either quite high or very low, with relatively few falling in the middle range. In the **Eastern Suburbs**, for example, SA2 regions such as Rose Bay - Vaucluse – Watsons Bay or Double Bay – Darling Point scored well above 0.8, while several other regions scored below 0.2. (see Appendix A) Parramatta displays a similar polarity, but with a slightly greater number of middle-ranked SA2s. The bimodality of distribution indicates that access to infrastructure and services is uneven within these regions - some SA2s are highly-resourced, while others significantly lag behind.

3.2.2. Spatial Patterns

In the **Eastern Suburbs**, many of the low-scoring SA2s are concentrated along the coastline. As shown in the map above, this includes areas such as Bondi Beach, South Coogee, and Dover Heights, which tend to

have fewer public transport stops, limited school catchments, and lower density of POIs - all key factors contributing to their lower scores.



Choropleth Map: Sigmoid Score by SA2 Region in Eastern Suburbs & Parramatta

However, this coastal pattern is not uniform. Coastal SA2s like Rose Bay – Vaucluse – Watsons Bay, Maroubra – South, and Malabar scored relatively high, driven by better infrastructure coverage and particularly high POI counts. This suggests that POI density is the dominant driver of score variation in Eastern Suburbs, with well-connected SA2s showing consistently higher performance.

In **Parramatta**, lower-scoring SA2s like Winston Hills, North Rocks, and Carlingford are mostly located on the outer edges of the SA4, where public transport, POIs, and school access are limited. In contrast, higher scores cluster near the centre, in areas like Granville – Clyde, Rosehill – Harris Park, and Parramatta – North, which are better connected and commercially active.

Notably, top SA2s don't require all components to be above average. For example, the region with the highest score, Granville - Clyde had a slightly negative schools' score (-0.28), while Ermington – Rydalmere had a negative score in business (-0.88), yet both scored highly due to strengths in other areas. Conversely, low-scoring regions, such as Greystanes - South, can still have isolated strengths (z-stops = 1.36), but these aren't enough to offset major weaknesses. (see Appendix A)

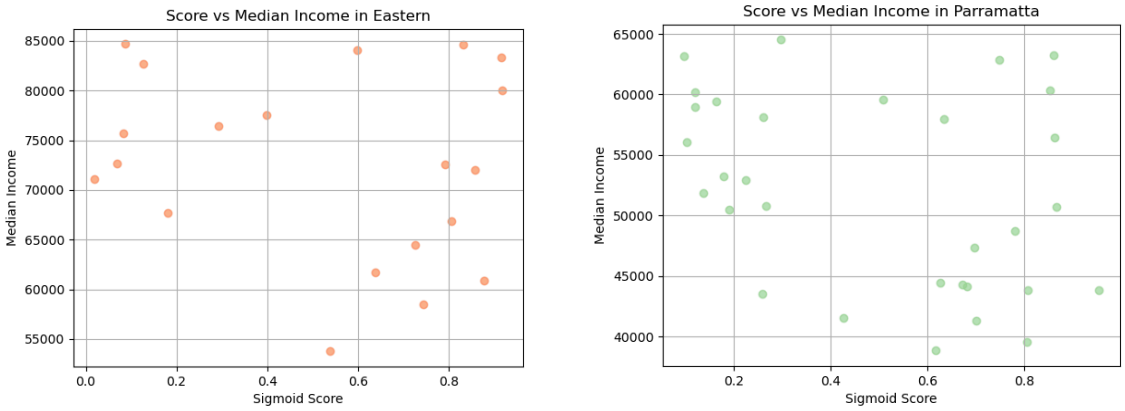
Overall, the results highlight that balanced performance across components is more valuable than excelling in just one. The sigmoid function reinforces this by compressing the influence at the extremes.

3.3 Correlation Analysis

To evaluate the relationship between well-resourcefulness and income in a SA2 region , we conducted a correlation analysis by comparing each SA2's sigmoid score with its median income.

In both **Parramatta** and **Eastern Suburbs**, we observe a consistent pattern: lower sigmoid scores often occur in high-income suburbs, while higher scores are more common in lower-income areas. In Parramatta, the correlation between score and income is –0.33, while the correlation in Eastern Suburbs is slightly weaker, –0.17 (see Appendix E). Wealthier areas such as Woollahra and South Coogee tend to have limited public infrastructure, while regions like Granville - Clyde or Matraville - Chifley score highly due to better access to transport, schools, and POIs.

These findings demonstrate a key insight: a region's economic wealth does not necessarily correspond to its public service access. The scoring method provides a complementary lens to traditional income-based metrics by revealing under-serviced areas that may be overlooked when income is the indicator. Thus, this method presents as a valuable tool for equity-focused urban planning and future infrastructure investment.



Scatter Plot between Sigmoid Score & Median Income in Eastern Suburbs & Parramatta

3.4 Rank Based Scoring Method

To complement the original scoring system, we designed a rank-based alternative that evaluates SA2 regions based on their relative standing across key components:

$$avg\_rank_{SA_2} = \frac{r_{business} + r_{stops} + r_{POIs} + r_{schools}}{4}$$

- $r_{business}$: rank of businesses per 1,000 residents, focusing on the Transport, Postal and Warehousing industry.
- r_{stops} : rank of public transport stops (train and bus).
- r_{POIs} : rank of number of Points of Interest (from the NSW POI API).
- $r_{schools}$: rank of school catchments per 1,000 youth (ages 0–19).

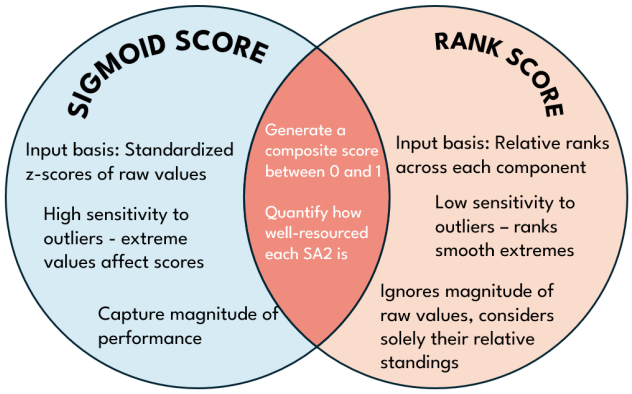
Next, to standardise the result on a 0-1 scale, we applied min-max normalization, producing the final rank-based score:

$$r\_score_{SA_2} = 1 - \frac{avg\_rank_{SA_2} - min(avg\_rank)}{max(avg\_rank) - min(avg\_rank)}$$

This scoring method ensures that SA2s with better (lower) average ranks receive higher scores, and the final score reflects balanced, consistent performance across all components. (see Appendix C)

Comparison to original results: The rank-based and sigmoid scoring methods generally agree on the top and bottom SA2s. Regions like Granville – Clyde and Double Bay – Darling Point consistently score highly in both methods, while Winston Hills and South Coogee rank low in both.

In general, a high rank score tends to align with a high sigmoid score, but not necessarily the other way around. For example, Rose Bay – Vaucluse – Watsons Bay ranks 1st by sigmoid score due to extremely high POIs and transport stops, yet only 10th by rank score because of weaker school and business presence. Similarly, Auburn North ranks 2nd by sigmoid (driven by high business density), but drops to 13th in rank score due to imbalance



Sigmoid Score & Rank Score method comparision

across metrics. (See Appendix B & Appendix D). This suggests that rank-based scores reward consistency, while sigmoid scores emphasize individual strengths, making the rank-based method more reliable for identifying well-rounded, well-serviced regions.

3.5 Building Regression Model

To explore how public infrastructure and geography relate to median income across regions, we developed a multiple linear regression model. We selected the following features based on their relevance to urban accessibility and regional development:

- `businesses_per_1000`: number of Transport, Postal and Warehousing businesses per 1,000 residents in each SA2.
- `num_stops`: number of public transport stops within each SA2.
- `num_pois`: number of Points of Interest in each SA2.
- `areasqkm21`: area of each SA2 region in square kilometers.

An 80/20 train-test split was applied, and we used backward stepwise selection on the training set to retain only statistically significant predictors. The final model retained all four predictors (see Appendix F), the estimated regression equation is:

$$\widehat{median\_income} = 78400 - 2154 \times areasqkm21 - 44.3 \times num\_stops - 1096.7 \times businesses\_per\_1000 + 94.2 \times num\_pois$$

The model performed well on the training set, with an adjusted R-squared of 0.776, indicating that approximately 78% of the variability in median income can be explained by the model. On the test set, the model achieved a Test R^2 of 0.35 and a Root Mean Squared Error (RMSE) of 7164, indicating moderate predictive performance.

Assumptions Checking: (see Appendix G)
The residuals vs fitted values plot shows no major curvature or funneling, suggesting that the relationship between predictors and income is approximately linear, and the residuals display homoscedasticity.
The Q-Q plot indicates that residuals are approximately normally distributed, with only minor deviations. This supports the assumption of normal error terms.
We also calculated the Variance Inflation Factor (VIF) for each predictor. All VIF values were well below 5, indicating no significant multicollinearity.

Limitations: A key limitation of this analysis is the small dataset size. We only included SA2s from two SA4 regions (Eastern Suburbs and Parramatta), resulting in just around 50 observations. This restricts the model's generalisability and statistical power, particularly on the test set.

Conclusion: Interestingly, three of the four predictors - `areasqkm21, businesses_per_1000`, and `num_stops` - have negative coefficients, indicating that larger, more industrial, or more connected areas tend to be associated with lower median incomes. In contrast, `num_pois` is positively associated with income, suggesting that access to public amenities such as parks and community facilities is linked to higher-income areas. These findings suggest that not all infrastructure contributes equally to socioeconomic outcomes, and in some cases, high density or industrial development may coincide with lower-income populations.

4. CONCLUSION

Overall, this project demonstrates how spatial data and public infrastructure metrics can be combined to evaluate regional equity and identify underserved areas. While both scoring methods identified well-resourced SA2s, the rank-based approach more reliably captured balanced infrastructure access. We found that high service access did not always align with income, and our regression model showed that POI access significantly influenced socioeconomic outcomes. These findings underscore the value of data-driven approaches in supporting equity-focused urban planning.

THE UNIVERSITY OF SYDNEY

APPENDIX A: Sigmoid Score's Results

	sa2_code	sa2_name	businesses_per_1000	num_pois	num_stops	schools_per_1000	z_business	z_stops	z_pois	z_schools	s_score
0	118011346	Rose Bay - Vaucluse - Watsons Bay	2.660680	233	191	1.664447	-1.027254	2.440997	2.398938	-1.386235	0.918822
1	118011650	Double Bay - Darling Point	4.678479	174	74	6.784261	0.047784	-0.849287	1.328308	1.868944	0.916503
2	118021653	Matraville - Chifley	8.686614	94	129	2.617040	2.183228	0.697428	-0.123395	-0.780576	0.878327
3	118021570	Randwick - South	3.505492	139	129	5.383023	-0.577157	0.697428	0.693188	0.978037	0.857110
4	118011345	Paddington - Moore Park	2.545723	177	101	6.038647	-1.088501	-0.089991	1.382747	1.394883	0.831898
5	118021652	Malabar - La Perouse	3.088442	210	127	3.222836	-0.799352	0.641183	1.981574	-0.395411	0.806589
6	118011341	Bondi Junction - Waverly	3.806112	106	144	4.689332	-0.416993	1.119259	0.094361	0.536988	0.791438
7	118021568	Maroubra - West	8.273635	56	64	5.500000	1.963203	-1.130508	-0.812953	1.052411	0.745006
8	118021567	Maroubra - South	6.023013	57	122	4.639393	0.764122	0.500573	-0.794807	0.505236	0.726140
9	118021564	Kensington (NSW)	6.012526	70	73	5.798394	0.758534	-0.877409	-0.558905	1.242130	0.637458
10	118011649	Bellevue Hill	3.857812	117	159	2.195734	-0.389449	1.541090	0.293970	-1.048443	0.598007
11	118021565	Kingsford	7.985481	61	89	3.052270	1.809680	-0.427456	-0.722222	-0.503857	0.538957
12	118021569	Randwick - North	4.848708	53	110	4.093199	0.138478	0.163108	-0.867392	0.157966	0.399430
13	118011339	Bondi - Tamarama - Bronte	4.456651	80	102	3.255562	-0.070401	-0.061869	-0.377443	-0.374604	0.292284
14	118021566	Maroubra - North	4.574565	70	72	3.773585	-0.007579	-0.905531	-0.558905	-0.045244	0.179865
15	118021651	Coogee - Clovelly	2.839757	87	116	2.147651	-0.931846	0.331841	-0.250419	-1.079014	0.126813
16	118011347	Woollahra	2.404488	59	53	5.416385	-1.163747	-1.439851	-0.758515	0.999248	0.086049
17	118011340	Bondi Beach - North Bondi	4.039732	78	96	1.530222	-0.292526	-0.230601	-0.413735	-1.471575	0.082532
18	118011344	Dover Heights	4.828326	55	84	1.750700	0.127619	-0.568066	-0.831100	-1.331395	0.068949
19	118021654	South Coogee	2.659574	40	49	3.342246	-1.027843	-1.552339	-1.103294	-0.319490	0.017934

Sigmoid Score in Eastern Suburbs, ordered by s-score

	sa2_code	sa2_name	businesses_per_1000	num_pois	num_stops	schools_per_1000	z_business	z_stops	z_pois	z_schools	s_score
0	125031481	Granville - Clyde	25.476892	185	239	3.299120	1.007516	1.234396	1.081511	-0.289826	0.954069
1	125041719	Rosehill - Harris Park	33.004386	134	59	5.208333	1.938641	-1.339569	0.144083	1.116391	0.865244
2	125021477	Ermington - Rydalmere	10.197912	168	294	3.598299	-0.882441	2.020885	0.769035	-0.069468	0.862713
3	125041717	Parramatta - North	16.389412	167	60	7.102273	-0.116574	-1.325269	0.750654	2.511359	0.860587
4	125041489	North Parramatta	13.307326	213	198	3.702942	-0.497817	0.648104	1.596178	0.007606	0.852465
5	125011586	Lidcombe	17.155714	173	181	3.938041	-0.021785	0.405007	0.860940	0.180767	0.806110
6	125011583	Auburn - North	41.370938	59	80	4.672897	2.973554	-1.039273	-1.234488	0.722020	0.805622
7	125031484	Guildford West - Merrylands West	16.433143	180	262	2.098951	-0.111165	1.563292	0.989606	-1.173803	0.780388
8	125041491	Northmead	14.738285	247	103	3.550543	-0.320813	-0.710377	2.221130	-0.104642	0.747495
9	125031483	Guildford - South Granville	19.817344	144	242	2.254156	0.307449	1.277296	0.327892	-1.059487	0.701228
10	125031714	Merrylands - Holroyd	20.369059	179	199	2.097789	0.375695	0.662404	0.971226	-1.174658	0.697341
11	125011587	Regents Park	24.409144	44	66	7.269790	0.875439	-1.239470	-1.510202	2.634743	0.681464
12	125031479	Chester Hill - Sefton	15.913978	151	256	2.283850	-0.175383	1.477493	0.456559	-1.037616	0.672839
13	125041493	Toongabbie - Constitution Hill	12.867327	180	194	3.041145	-0.552244	0.590905	0.989606	-0.479836	0.633772
14	125031480	Fairfield - East	15.132329	132	230	3.115265	-0.272071	1.105698	0.107321	-0.425243	0.626143
15	125011582	Auburn - Central	31.887301	100	141	2.775850	1.800461	-0.166985	-0.480869	-0.675237	0.617127
16	125041589	Wentworthville - Westmead	20.639882	127	174	2.744237	0.409195	0.304909	0.015416	-0.698521	0.507749
17	125011584	Auburn - South	22.801303	52	83	5.577689	0.676555	-0.996374	-1.363155	1.388438	0.426894
18	125031715	Pemulwuy - Greystanes (North)	10.176951	168	93	3.840000	-0.885034	-0.853376	0.769035	0.108555	0.297168
19	125021712	Carlingford - West	10.416127	118	211	2.561072	-0.855449	0.834002	-0.150012	-0.833430	0.267981
20	125041588	Pendle Hill - Girraween	18.729001	100	129	3.157064	0.172825	-0.338582	-0.480869	-0.394457	0.260941
21	125011585	Berala	19.081762	50	96	4.975124	0.216461	-0.810476	-1.399917	0.944623	0.259358
22	125031716	South Wentworthville	15.680393	45	120	4.955401	-0.204277	-0.467281	-1.491821	0.930096	0.225607
23	125021711	Carlingford - East	7.892931	99	121	4.608295	-1.167559	-0.452981	-0.499250	0.674437	0.190718
24	125021478	Oatlands - Dundas Valley	9.538180	141	152	2.587880	-0.964048	-0.009687	0.272750	-0.813685	0.180248
25	125011710	Wentworth Point - Sydney Olympic Park	11.184816	165	71	3.141690	-0.760364	-1.167971	0.713892	-0.405780	0.165174
26	125041718	Parramatta - South	27.108434	34	59	3.680336	1.209332	-1.339569	-1.694012	-0.009044	0.137846
27	125011709	Silverwater - Newington	14.138817	64	80	4.506534	-0.394965	-1.039273	-1.142583	0.599486	0.121603
28	125041490	North Rocks	6.496575	87	139	4.065041	-1.340284	-0.195584	-0.719822	0.274307	0.121172
29	125031713	Greystanes - South	8.769822	79	248	1.523395	-1.059091	1.363094	-0.866869	-1.597724	0.103346
30	125041494	Winston Hills	6.161264	126	153	2.538071	-1.381760	0.004613	-0.002965	-0.850371	0.097046

Sigmoid Score in Parramatta, ordered by s-score

APPENDIX B: Rank Score's Results

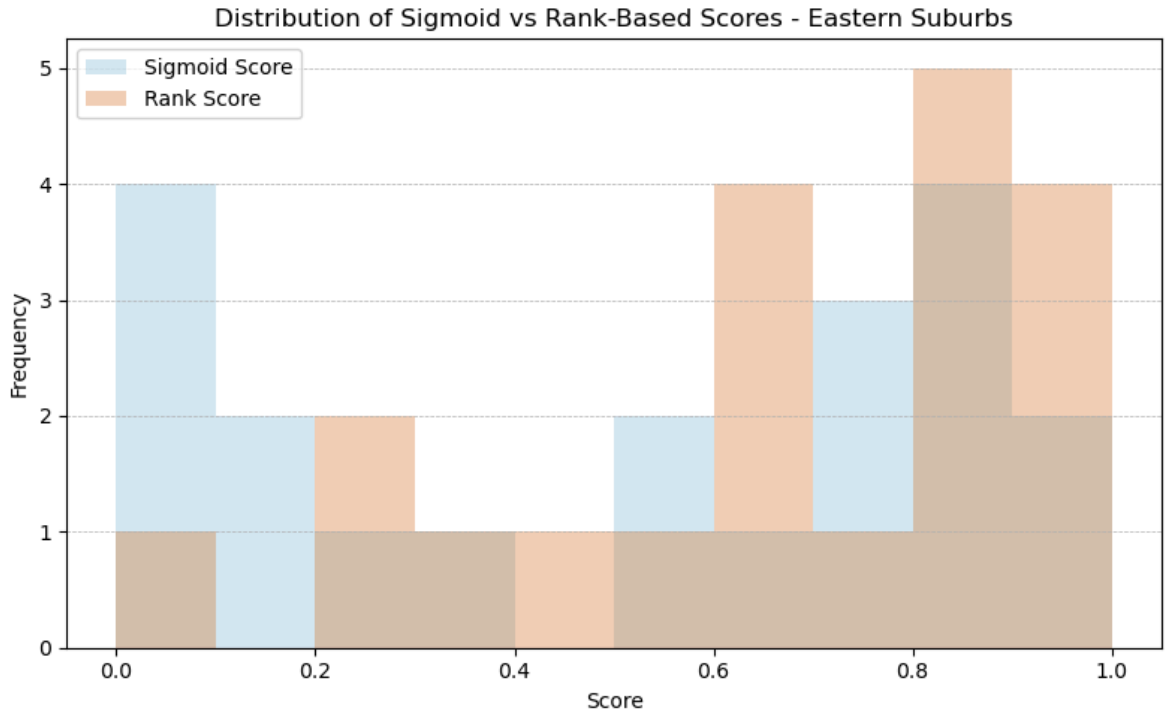
	sa2_code	sa2_name	r_business	r_poi	r_stops	r_schools	avg_rank	r_score
0	118011650	Double Bay - Darling Point	8	4	15	1	7.00	1.000000
1	118021653	Matraville - Chifley	1	8	4	15	7.00	1.000000
2	118021570	Randwick - South	14	5	4	6	7.25	0.975610
3	118011341	Bondi Junction - Waverly	13	7	3	7	7.50	0.951220
4	118011345	Paddington - Moore Park	19	3	11	2	8.75	0.829268
5	118021567	Maroubra - South	4	16	7	8	8.75	0.829268
6	118021652	Malabar - La Perouse	15	2	6	13	9.00	0.804878
7	118021564	Kensington (NSW)	5	12	16	3	9.00	0.804878
8	118011649	Bellevue Hill	12	6	2	16	9.00	0.804878
9	118011346	Rose Bay - Vaucluse - Watsons Bay	17	1	1	19	9.50	0.756098
10	118021568	Maroubra - West	2	17	18	4	10.25	0.682927
11	118011339	Bondi - Tamarama - Bronte	10	10	10	12	10.50	0.658537
12	118021569	Randwick - North	6	19	9	9	10.75	0.634146
13	118021565	Kingsford	3	14	13	14	11.00	0.609756
14	118021566	Maroubra - North	9	12	17	10	12.00	0.512195
15	118021651	Coogee - Clovelly	16	9	8	17	12.50	0.463415
16	118011340	Bondi Beach - North Bondi	11	11	12	20	13.50	0.365854
17	118011344	Dover Heights	7	18	14	18	14.25	0.292683
18	118011347	Woollahra	20	15	19	5	14.75	0.243902
19	118021654	South Coogee	18	20	20	11	17.25	0.000000

Rank Score in Eastern Suburbs, ordered by r-score

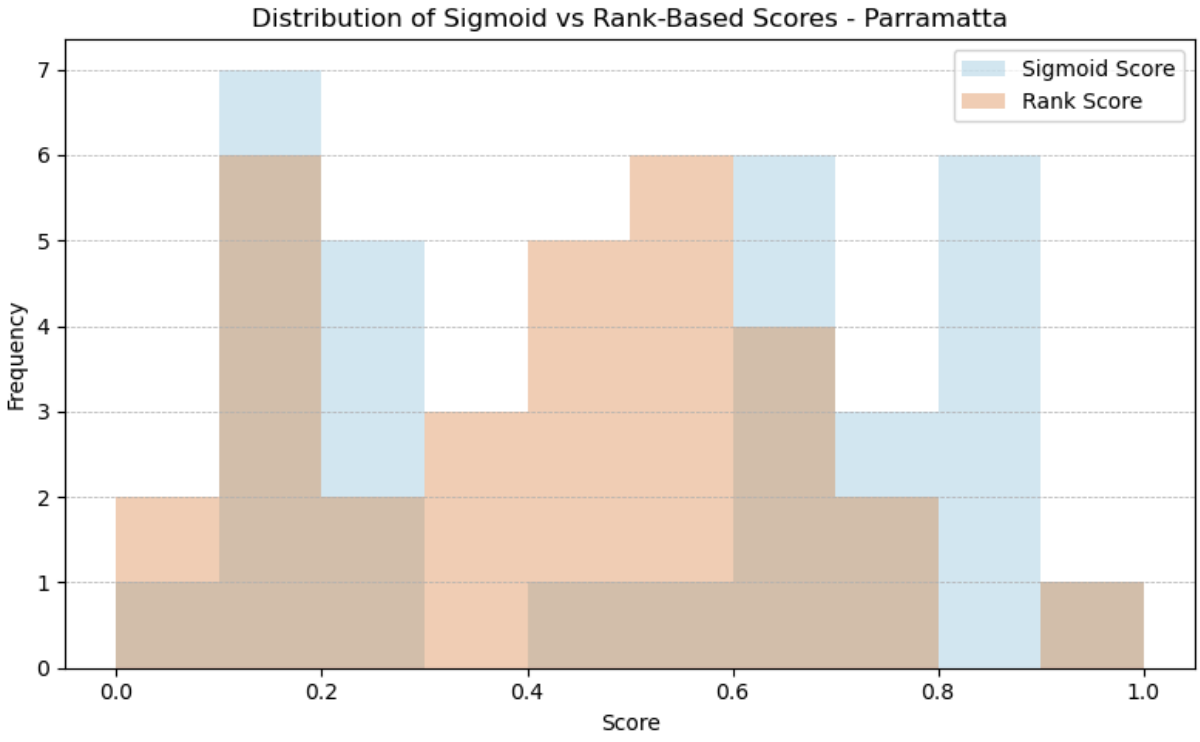
	sa2_code	sa2_name	r_business	r_poi	r_stops	r_schools	avg_rank	r_score
0	125031481	Granville - Clyde	5	3	6	17	7.75	1.000000
1	125011586	Lidcombe	13	7	12	11	10.75	0.793103
2	125041489	North Parramatta	21	2	10	13	11.50	0.741379
3	125031484	Guildford West - Merrylands West	14	4	2	29	12.25	0.689655
4	125021477	Ermington - Rydalmere	25	8	1	15	12.25	0.689655
5	125041719	Rosehill - Harris Park	2	15	30	4	12.75	0.655172
6	125031714	Merrylands - Holroyd	9	6	9	30	13.50	0.603448
7	125031483	Guildford - South Granville	10	13	5	28	14.00	0.568966
8	125041717	Parramatta - North	15	10	29	2	14.00	0.568966
9	125041491	Northmead	19	1	21	16	14.25	0.551724
10	125031479	Chester Hill - Sefton	16	12	3	27	14.50	0.534483
11	125041493	Toongabbie - Constitution Hill	22	4	11	21	14.50	0.534483
12	125011583	Auburn - North	1	26	25	7	14.75	0.517241
13	125011582	Auburn - Central	3	20	16	22	15.25	0.482759
14	125041589	Wentworthville - Westmead	8	17	13	23	15.25	0.482759
15	125031480	Fairfield - East	18	16	7	20	15.25	0.482759
16	125011584	Auburn - South	7	27	24	3	15.25	0.482759
17	125011587	Regents Park	6	30	28	1	16.25	0.413793
18	125011585	Berala	11	28	22	5	16.50	0.396552
19	125041588	Pendle Hill - Girraween	12	20	18	18	17.00	0.362069
20	125031715	Pemulwuy - Greystanes (North)	26	8	23	12	17.25	0.344828
21	125031716	South Wentworthville	17	29	20	6	18.00	0.293103
22	125021712	Carlingford - West	24	19	8	25	19.00	0.224138
23	125021711	Carlingford - East	29	22	19	8	19.50	0.189655
24	125041718	Parramatta - South	4	31	30	14	19.75	0.172414
25	125011709	Silverwater - Newington	20	25	25	9	19.75	0.172414
26	125011710	Wentworth Point - Sydney Olympic Park	23	11	27	19	20.00	0.155172
27	125021478	Oatlands - Dundas Valley	27	14	15	24	20.00	0.155172
28	125041490	North Rocks	30	23	17	10	20.00	0.155172
29	125031713	Greystanes - South	28	24	4	31	21.75	0.034483
30	125041494	Winston Hills	31	18	14	26	22.25	0.000000

Rank Score in Parramatta, ordered by r-score

APPENDIX C: Distribution of Rank Score and Sigmoid Score

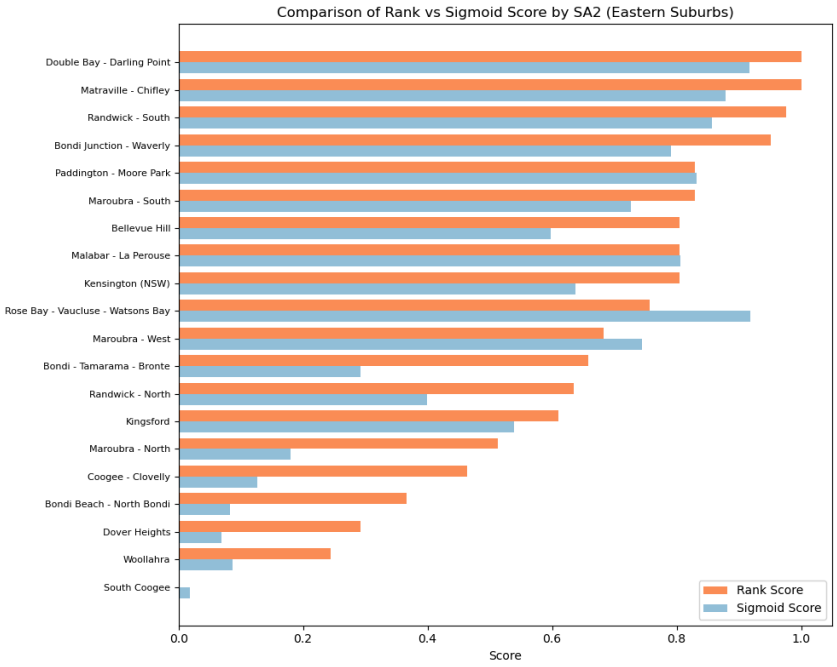


Rank Score & Sigmoid Score Distribution in Eastern Suburbs

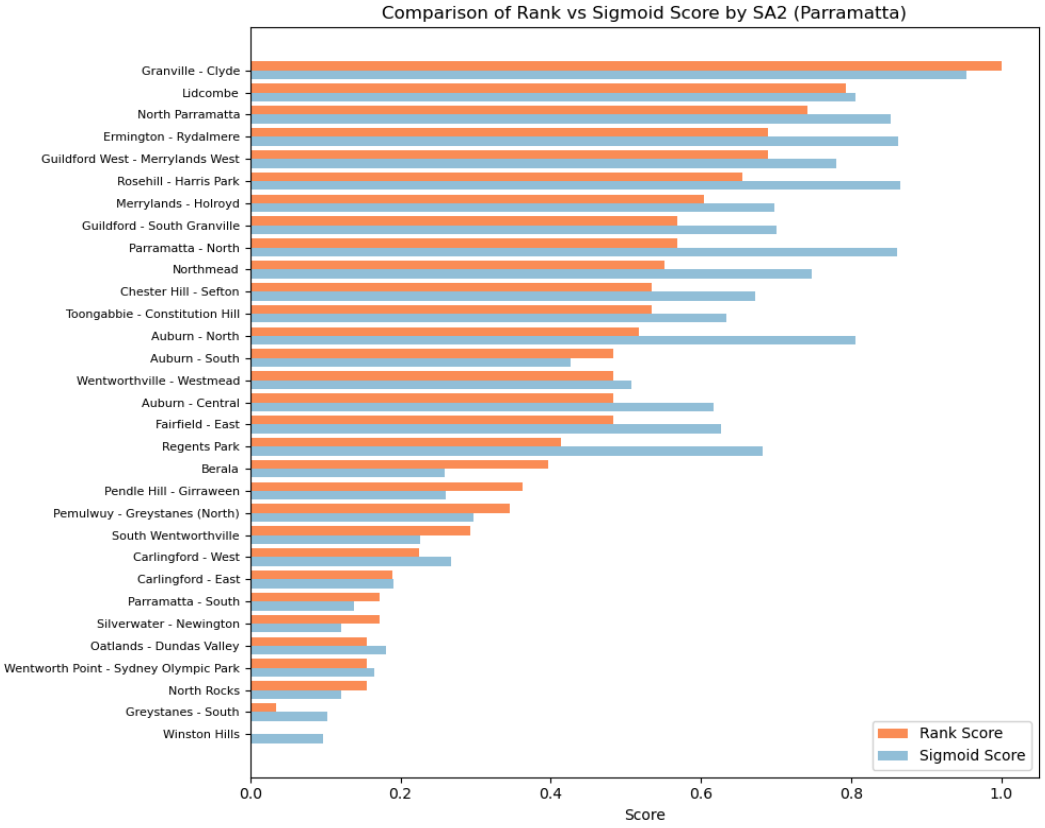


Rank Score & Sigmoid Score Distribution in Parramatta

APPENDIX D: Rank Scores and Sigmoid Score by SA2

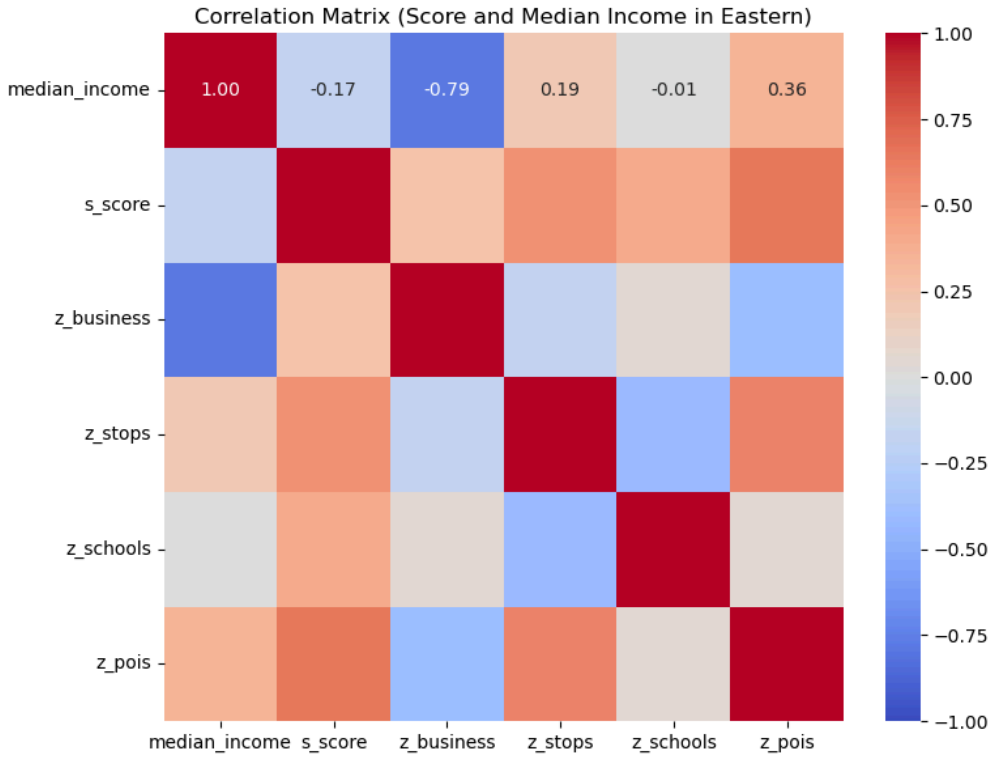


Bar Graph of Rank Scores vs Sigmoid Score by SA2 in the Eastern Suburbs

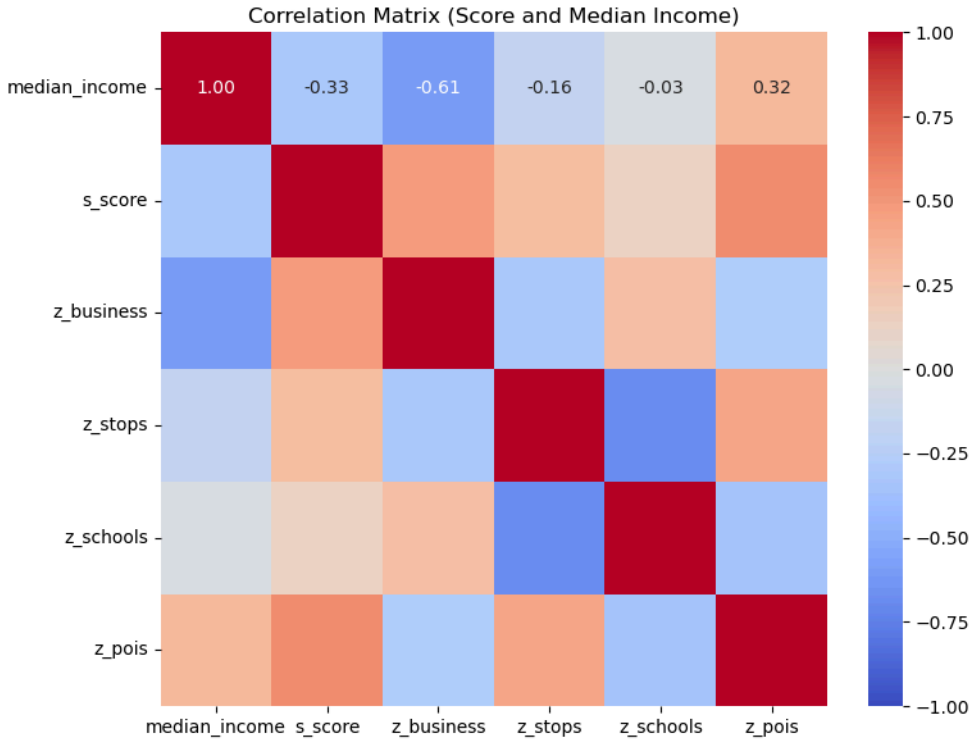


Bar Graph of Rank Score vs Sigmoid Scores by SA2 in Parramatta

APPENDIX E: Correlation Matrix



Correlation Heat Map: Sigmoid Score vs Median Income - Eastern Suburbs



Correlation Heat Map: Sigmoid Score vs Median Income (Parramatta)

APPENDIX F: Model Summary

OLS Regression Results

Dep. Variable:	median_income	R-squared:	0.799
Model:	OLS	Adj. R-squared:	0.776
Method:	Least Squares	F-statistic:	34.78
Date:	Sun, 18 May 2025	Prob (F-statistic):	9.60e-12
Time:	14:12:21	Log-Likelihood:	-406.22
No. Observations:	40	AIC:	822.4
Df Residuals:	35	BIC:	830.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.84e+04	3087.241	25.395	0.000	7.21e+04	8.47e+04
areasqkm21	-2153.9613	823.127	-2.617	0.013	-3824.997	-482.925
num_stops	-44.2550	22.794	-1.942	0.060	-90.529	2.019
businesses_per_1000	-1096.7140	110.241	-9.948	0.000	-1320.515	-872.913
num_pois	94.1517	27.370	3.440	0.002	38.588	149.715

Omnibus:	1.733	Durbin-Watson:	1.595
Prob(Omnibus):	0.421	Jarque-Bera (JB):	1.168
Skew:	0.109	Prob(JB):	0.558
Kurtosis:	2.192	Cond. No.	561.

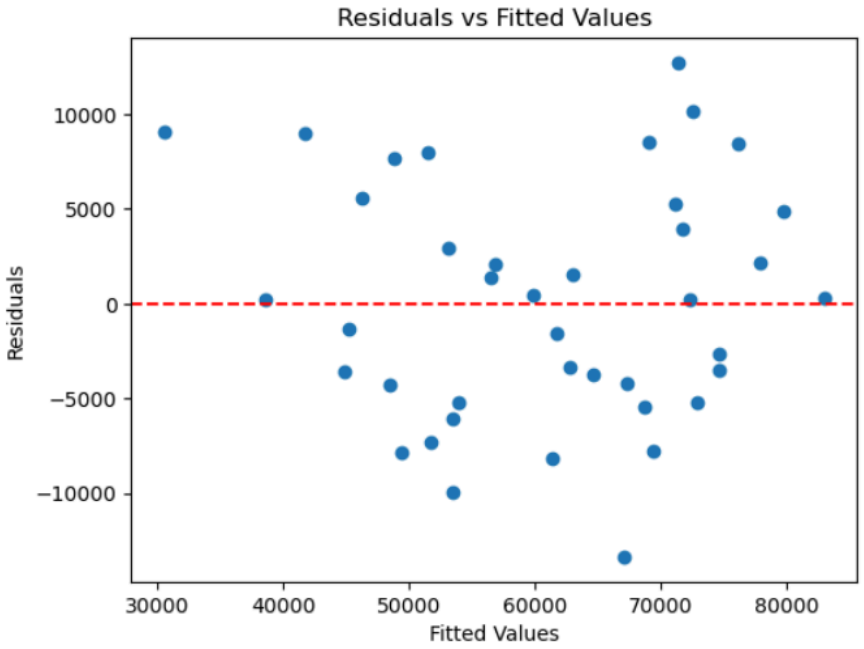
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Summary of Final Regression Model Predicting Median Income

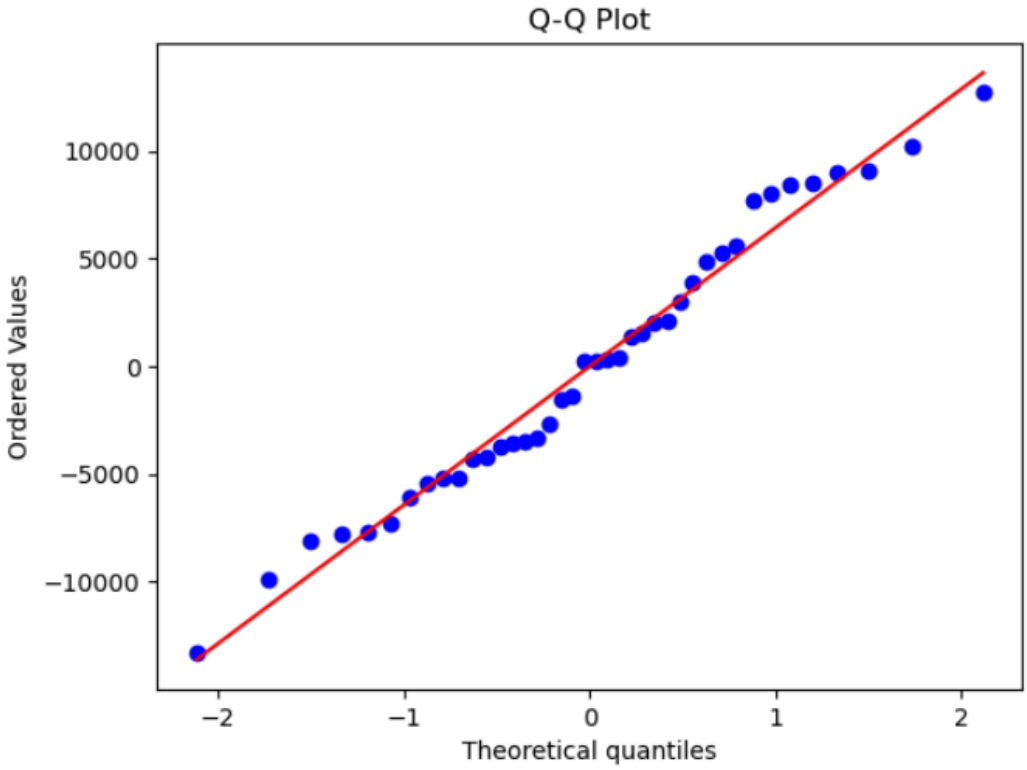
Test R^2: 0.3500337315030242
Test RMSE: 7163.80135118231

Statistics on testing data

APPENDIX G: Assumptions Checking



Residuals vs Fitted Values Plot for Income Regression Model



Q-Q Plot of Residuals for Income Regression Model

	Feature	VIF
0	const	8.727391
1	businesses_per_1000	1.019778
2	num_pois	2.137954
3	num_stops	1.892515
4	areasqkm21	3.083768

Variance Inflation Factor (VIF) for Regression Predictors