

Sydney Liveability Analysis

Bohan Zhang¹ and Yi Ji¹

¹ The University of Sydney, NSW, 2008

This manuscript was compiled on May 23, 2022

This study is aimed to investigate the liveability for each suburb in SA2(Greater Sydney) area through computing liveability score by the Sigmoid function of z-score using multiple datasets such as census-based data and geospatial data. The final liveability score yields by calculating z-score for five dimensions of school catchments, accommodation and food service facilities, retail service facilities, crime frequency and health services. Through this study, the question “which area of Sydney is more suitable for living” is answered by the visualization of GIS-based analysis. Also, the correlation between liveability scores and median income and median rent are explored. Specifically, the results show that the top 5 suburb of high liveability are “sydney - haymarket - the rocks,” “badgreys creek,” “north sydney - lavender bay,” “chullora” and “darlinghurst,” where the liveability is ranked in order of priority. Furthermore, the Pearson correlation coefficient test results suggest that the liveability score is proportionate to the average annual household income and average rent, implying that the liveability score tends to rise in tandem with the growth of income and rent in the region. Finally, the liveability score of each region of the City of Sydney is assessed by incorporating a car park and a traffic dimension. The goal is to assist stakeholders in making better decisions, such as giving a reference for address selection for big families with numerous automobiles and persons with impairments, as well as guidance on the placement of small family stores. Additionally, to propose liveability improvement references for urban planners in diverse areas.

liveability | liveability score | census based data | correlation analysis

Introduction

Background. Liveability is a widely used term for describing the quality of life and communal well-being. While specific definitions differ, healthy communities, environmental sustainability, social capital, and social cohesiveness are widely regarded to be the foundations of liveability. The relative comfort of an area is often judged based on more than 30 qualitative and quantitative characteristics, divided into five main categories: stability, healthcare, culture and environment, education, and infrastructure. The liveability score is intended to assist individuals in quickly and simply assessing the quality of a location. The widely used ranking now consists of seven areas, each of which is assessed independently based on data (amenity, cost of living, crime, employment, housing, schools, and user ratings). Assign a letter grade to each category to make it easier to assess scores and establish the level of comfort in the area.

Dataset Description. This section is aimed to introduce the information about datasets used in the following analysis including the source and content of data, the format of data and the pre-processing operations such as data cleaning.

The *Neighborhoods.csv* data set is sourced from Australian Bureau of Statistics (ABS) which was gathered during the 2016 Australian government census in the Greater Sydney area. It contains 321 entries and 12 attributes, which includes

basic information data about each SA-2 (Statistical Area 2) level area, such as the id and name of area and population number; besides, young people aged 0-19 are divided into intervals every 4 years, and the specific population of each interval is also recorded, such as the number of children aged 0-4 in the area; also contains information about residents' livelihood such as the number of business, median income, rent, and so on. Data cleaning includes: filtering missing data; unifying semantic data representations; and type conversion. Rows with missing values are dropped, and the space is replaced with 'nan,' which stands for 'not a number.' Because it represents the identification of the related region, the 'area id' attribute is transformed from quantitative data type integer to categorical data type object. In addition, for the two numeric attributes *population* and *number_of_dwellings* is unified by removing the comma ',' in the value which is inconsistent with numerical data type, and converting these two variable to floating numbers since they both shows mathematically meaningful statistics representing the number of population and the number of dwellings.

The *BusinessStats.csv* data set is obtained from the 2016 census data for the Greater Sydney area released by the Australian Bureau of Statistics (ABS). The data set contains 2300 records and 9 attributes, three of which are the fundamental information about the suburb, such as *area_id*, *area_name* and number of business, while the other six columns record the number of business in each industrial category in each SA-2 level, Greater Sydney area, such as *accommodataion_and_food_service* or *retail_trade*. To clean the data for the further analysis, cast the data type of *AR-EASQKM16* to float, since it stores the mathematical area size in unit square kilometers of each region in SA2 area.

The *SA2_2016_AUST.shp* shape data set is taken from the Australian Bureau of Statistics (ABS), which contains information about each SA2 area and its associated parent areas. It's a shape file has 2309 items and 13 columns, including the name and matching code of each region, such as SA2, Greater Sydney area, and SA3 area, among others. And one of those attributes, *geometry* holds geographical data relating to a specific position in each region through polygon and multipolygon. The Spatial Reference Identifier (SRID) in this data collection is specified 4326, which refers the WGS84 world geodetic coordinate system is used. Furthermore, the polygons in the geometry column in SA2 data frame is converted by conducting the same WKT conversion. To clean the data, the rows which contains missing value are dropped; SRID is referenced in the geometry column; rename attributes for readability.

The *catchments_future.shp* geospatial data is provided by NSW Government. However, the original data source and release year remain unknown. It contains 43 rows and 18 columns, including the basic information of each area such as

USE_ID, ADD_DATE. Furthermore, the data set includes the shape data for schools from kindergarten to high school (year 12) as well as prospective Government school catchments in that corresponding area. The *geometry* column represents the precise position of the region on the Earth. Data cleaning includes: copy the original data for future use; reference the spatial column *geometry* with SRID4326 refers to WSG84 world geodetic coordinate system.

The *BreakEnterDwelling_JanToDec2021.shp* geospatial data set is obtained from NSW Bureau of Crime Statistics and Research (BOCSAR) which contains the shape data of theft “hotspots” in NSW in 2021. It contains 2593 entries and 7 attributes, including the OBJECTID attribute, which is the object’s identifier. The specific location of the theft is recorded in the *geometry* column, which is represented by polygon and multi-polygon with SRID 4326, indicating that the WGS84 global geodetic coordinate system is used. Data pre-processing includes: reference the geometry column with SRID and rename the attribute for readability.

Extra Datasets Description. The *Mobility_parking.geojson* data set is gathered from the City of Sydney Open Data Hub which is public in August 2019 and update in September 2020. It’s a GeoJSON format data set which encoding a variate of geographic data structures that keeps track of available parking spaces for persons who have a mobility parking scheme permit for a defined period of time. The data set contains 201 rows and 15 columns, including the unique identifier *OBJECTID*, and also comprehensive information about parking locations such as *Address*, *Street*, and *Suburb*. Furthermore, the specific geospatial information of that parking place on Earth is stored in the *geometry* column represented by polygons. Since the data is well-formatted, only rename the attributes to gain more readability.

The *Parking_meters.geojson* data set is sourced from the City of Sydney Open Data Hub which is public in August 2018 and update in September 2021. It’s a GeoJSON type data set that represents geospatial features and their non-spatial attributes. It records the information about the metered parking in the City, time limits and costs. The data set contains 1377 entries and 12 attributes, such as the identify of meter and location, the payment method accepted, and the location in detailed including the name of *Street* and *Suburb*. Besides, the *geometry* attributes represents the specific location of that metered parking place on Earth. To clean the data, the attributes of the data set is renamed for further analysis.

Database Description. The objective of this section is to introduce the database about Sydney liveability, including the schema of database, the relationship between tables and the creation of those indexes.

This dataset is based on real data obtained by the Australian Bureau of Statistics during the 2016 census to study the living environment of residents in Sydney’s SA2 area and estimate liveability scores for each region. according to “Australian Statistical Geography Standard(ASGS)”, Statistical Areas Level 2 (SA2) are medium-sized general purpose areas built up from whole Statistical Areas Level 1. Their purpose is to represent a community that interacts together socially and economically. SA2 is the same size as Suburb and is named after it. Statistics on the living level of each SA2 area in the

Greater Sydney area, such as the number of businesses, rent, etc, might aid in analyzing Sydney’s liveability.

This report is based on the liveability dataset which represents data that has been collected by ABS 2016 census, and used to assess the liveability of each suburb in the following study. The database consists of five tables: *neighbourhoods2*, *businessstats*, *catchments*, *crime* and *sa2*.

These five tables are linked via the following relationships: The SA2 area is divided into many sub-areas according to location, such as Surry Hills and Darlinghurst. Basic information about these sub-areas, such as population in each 4-year-period age group is recorded in the *neighbourhoods2* table. Furthermore, business-related information within each sub-region, such as the number of retail trade business, the number of health care and social assistance institution, is documented in the *BusinessStats* database. Moreover, the *sa2* table contains geographic graphic data for corresponding to each suburb, which indicates the region’s shape and specific location on the Earth. While the *crime* table contains information on the frequency of theft offenses in the area through shape data. The *catchments* table demonstrates the divisions of kindergartens, primary schools, and high schools, as well as future school catchments data. Although the graphical data type for each area may be polygons or multi-polygons, since some shapes may contain multiple exterior rings, all geographic data is entered in the database as a multi-polygon data type.

Fig.1 shows the ERD (Entity Relationship Diagram) of schema, where the PK (primary key) is denoted by bold underline, and the corresponding FK (foreign key), constraints and table links are showed through connecting lines. According to the ERD, the *neighbourhoods2*, *businessStats* and *sa2* tables can be related to each other by their primary key. Note that there is no primary key defined for *catchments* table or *crime* table since the column holding ID data has duplicate values, which does not satisfy the uniqueness requirement of PK. However, the two tables may still be connected with *sa2* table via the *geom* column.

In this case, the index is created on the *geom* column in the three tables which contains geospatial data. The index form of the “Generalized Search Tree” AA spatial database consists of a collection of tuples representing spatial objects. Each tuple has a unique identifier that can be used to retrieve it. Every leaf node contains between m and M index. So with using the tree structure, SQL queries don’t need to search the whole column of multi-polygon instead, Query can find out the contain and overlap relation between the multi-polygons. Therefore, the searching process speeds up after creating indexes on the geometry attributes.

Greater Sydney liveability Analysis

This section presents the method for identifying the liveability of each suburb in Greater Sydney, as well as interpreting the results.

Method.

$$Score = S(z_{school} + z_{accomm} + z_{retail} - z_{crime} + z_{health})$$

The liveability score is calculated using the above method by taking the Sigmoid function of the sum of the z-scores for the five measurement factors. Since the relationship between

the specific value and the mean value of the group can be expressed by computing the z-score. In this formula, the *school* factor represents the number of schools catchment areas per 1000 0-19-year-old young people, which suggests the allocation of school and educational resource; the *accom* factor represents the number of accommodation and food services per 1000 people, which documents the infrastructure configuration and community size; the *retail* factor represents the number of retail services per 1000 people, which shows the scale of business and convenience of living; the *crime* factor represents the sum of hotspot areas divided by total area, which reflects the theft crime frequency and further demonstrates the security and stability; finally, the health factor represents the number of health services per 1000 people, it presents the medical condition per capita. The final score ranges from -1 to 1, with 1 being more livable and -1 being less livable. However, some measurements for some regions are lacking, such as the corresponding retail shop statistics for that suburb. Removing all suburbs with any missing values causes an immediate decrease in the amount of data by a substantial margin, resulting in misleading assessment findings; hence, these true values are padded with 0s in the score computation.

Results and Discussion. Table 1 shows that Greater Sydney's top five liveability suburb are as follows: "sydney-hey-market-the rocks," "badgerys creek," "north sydney - lavender bay," "chullora" and "darlinghurst." The scores of these five areas are all close to one, indicating that they perform well in the five aspects of participating in the liveability evaluation, and that, when compared to other areas, these five areas have a good level of security, adequate educational resources and health facilities, and business services related to people's livelihood are relatively complete. To summarize, the five regions in Table 1 are considered to offer higher living comfort and convenience. Remarkably, the final ranking results overlapped with three regions, including A, B, and C, with the top five rankings in only three dimensions: general health, retail, accommodation and food services. This demonstrates that the benefits of their high scores are found in two aspects: medical conditions and people's livelihood services. This demonstrates that the creation of regional livelihood service facilities has a favourable influence on the area's liveability, which is consistent with expectations. Surprisingly, the suggested liveable places identified by the liveability score did not coincide with the top five crime-free zones. However, the significance of security to liveability cannot be emphasized. Because safety is one of the top concerns for inhabitants, a lack of security can have a detrimental influence on a site's liveability. Table 2 displays the five regions with the lowest liveability scores, which also happen to be the most crime-ridden similar to the results from crime score computation, demonstrating that safety is an essential factor in liveability. This study found extreme values for school district allocation as a measure of educational resources. Schools in the "Badgerys Creek" district scored seventy times higher than the second place, and despite the fact that there is only one school in the district (or a school catchment to which a school is planned to be allocated in the future), the per capita allocation is high because the district has only 13 people of educational age. Overabundance of educational resources results in overrated schools. Despite unexceptional performance in other indicators such as business, health care, and crime, extremely high school ratings are also

a big component in "Badgerys Creek" 's top five liveability rankings.

Correlation Analysis

This section is aimed to investigate the correlation between median rent, median income, and liveability score in the Sydney SA2 area based on the Pearson correlation coefficient test.

Assumptions. The data has to meet the assumptions of the Pearson correlation test in order for the results to be valid. First, the liveability score, median rent, and income all meet the assumption of continuous numerical variables. Besides, the adoption of a median to estimate family income and rent in each location decreases the possibility of outliers. According to the scatter plot shown in Fig. 2, there is an uphill pattern moving from left to right, which indicates there might be a positive relationship between liveability score and income. It also proves that the data satisfies the correlation assumption. Fig. 3 depicts a similar situation, with a link between the median rent and the liveability score. The data point distribution indicates a tendency for rent to increase with the score, suggesting a positive correlation between the two variables. In conclusion, the data generally meet all the assumptions of the Pearson correlation test.

Results and Discussions. The correlation coefficient between median income and liveability score is 0.33, and the p-value is smaller than 4.72×10^{-9} , indicating a statistically significant correlation. As a consequence, the median income and liveability score have a moderately positive linear relationship. That is, both the median rent and liveability score change in the same direction, the improvement of liveability has an impact on a particular level of income growth.

The correlation coefficient of median rent and liveability score is 0.49, And the p-value is 1.57×10^{-19} which is far less than the significant level 0.05. It indicates that the correlation is statistically significant. Therefore, it can be conclude that there is a medium positive correlation between median rent and liveability. In another words, the the median rent increases with liveability scores to a certain extent.

In general, rent is more closely tied to liveability than income since the correlation coefficient is 0.16 higher.

However, while the mean rent and average income are positively correlated with the area's liveability, it is hard to tell which is the dependent variable and which is the independent variable. Because high-income individuals will have more options for where to live, they may prefer to live in regions that are objectively more livable, such as residential districts with higher public protection or more established commerce. As a result, such residential values are likewise greater. Because persons with higher incomes have a wider range of liveability levels, average income may be regarded the dependent variable in this scenario. Yet, determining the relationship between average rent and liveability scores is more ambiguous. Because the liveability of the region in which the property is located is employed as a dependent variable from the standpoint of the homeowner, the better the liveability, the greater the rent renters are prepared to pay. In contrast, the higher the rental cost, the better the location of the rented property, and the more extensive the surrounding living facilities are from the standpoint of renters. The rent appears to be the dependent

variable at this moment. To recapitulate, whether it is the link between the score and the average high income or the correlation with the average rent, which side is impacted and which side affects the other varies in real life depending on different views. But one thing is certain: there is a statistically significant positive correlation between them.

City of Sydney Analysis

This section aims to optimize and briefly explain the liveability scores for each of the City of Sydney areas based on stakeholder demands.

Method.

$$Score = S(z_{school} + z_{acomm} + z_{retail} - z_{crime} + z_{health} + z_{parkingfree} + z_{parkingcharge})$$

Based on the basic liveability score evaluation framework, a new evaluation feature, traffic, is introduced by adding parking space data to generate the livvability in City of Sydney. Since the parking spaces reflect the liveability of public spaces, *parkingfree* denotes the distribution of free parking spaces in the suburb for those with parking permits; *parkingcharge* reflects the distribution of metered parking lots in the area. Together, these two variables represent the distribution of car park resources in the suburb. Similarly, their z-score will be utilized for the Sigmoid function to determine the liveability score at the end. Since the z-score indicates the difference between parking resources in each suburb and the average parking resources in the Sydney city area to establish the area's liveability rating.

Results and Discussion. Table 3 presents the five most livable areas in the City of Sydney are "surry hills," "sydney," "pyrmont," "darlinghurst" and "ultimo." The addition of parking spaces data not only takes into account parking issues that are closely related to stakeholders who use private cars as their primary mode of transportation, but it also increases the proportion of the dimension that measures the region's commercial development in the score calculation formula. Since metered parking lots are frequently densely located in commercially crowded places such as shopping malls or pedestrian streets.

As a consequence, this research will serve as a reference for the three stakeholders of resident, business and government, based on the results of the liveability score analysis of Sydney of City and the results of the liveability analysis above.

First, extended families with multiple cars and persons with impairments are the most direct stakeholders in terms of housing demands. The major reason of street traffic congestion is a lack of parking places, namely on-street parking, which causes traffic congestion and increases traffic difficulties. This will make it difficult for individuals to travel. Besides, apartment parking spots are restricted, and legal temporary parking places are always in short supply. As a result, a parking lot near the family can be a viable solution to the problem of redundant vehicle placement, whether it is the second car in the house or the car driven by relatives and friends when they come to visit. Choosing a home with plenty of parking may considerably lessen the hassle of potential unlawful parking. Furthermore, many people with disabilities rely on private transportation for their transportation, and free parking for

licensed individuals might significantly alleviate the problem of parking for people with disabilities in their daily travel. Choosing to live in an area like this might relieve them of the burden of thinking about transportation choices near their house.

Second, small businesses who cannot afford to run their own parking lots might use the liveability score for site selection following the addition of parking lot data. Because a high score indicates a rich neighborhood with extensive commerce and high pedestrian flow. Not only that, but parking facilities are critical to a small business's success since no consumer likes to waste time seeking a parking place. As a result, in the absence of the capacity to run the company's own parking space, selecting a location with adequate parking resources is a viable option.

Finally, when the ranking results of the City of Sydney are compared to the ranking results of the Greater of Sydney, it is discovered that "sydney" suburb and "darlinghurst" are in the top five of both rankings. This demonstrates that traffic management in these two locations is excellent. Today, one of the most significant variables influencing the liveability of city streets is urban traffic management. As a consequence, the findings may be utilized as a reference for government regional planning organizations to build transportation infrastructure and street grids with multidimensional aspects such as parking space planning to improve the area's liveability.

Limitations

This analysis relies on 2016 census data from several ABS, however the 2019-2021 COVID-19 epidemic will have a substantial impact on the real situation in various areas of Sydney, such as the likely closure of numerous stores. Because of the data latency, the computed liveability score may deviate from reality.

In some suburbs, a measuring variable, such as statistics on health services, may be lacking. Those missing observations are replaced by 0 in the computation to prevent cutting too many areas. This may result in discrepancies between the projected score and real liveability.

Because educational resources in districts will be evaluated based on the number of schools allocated per 1,000 school-age pupils, some districts with relatively small school-age populations may obtain high school scores despite having a few schools.

Fig.4 depicts a heat map based on the liveability score, where 1 denotes good living comfort and -1 indicates that the location is uninhabitable. The heatmap findings are identical to those reported by the SMH (Sydney Morning Herald), with Sydney's eastern suburbs being rated as more liveable, while western Sydney residents rate their suburbs as worse places to live than people in their parts of Sydney.

Conclusion

To guarantee the comprehensiveness of result, further research should try to introduce data from more dimensions, not only the objective physical data used in this study but also the subjective evaluation of households in the region through statistical forms such as questionnaires as a reference, in order to obtain a more appropriate regional liveability score.

Appendix

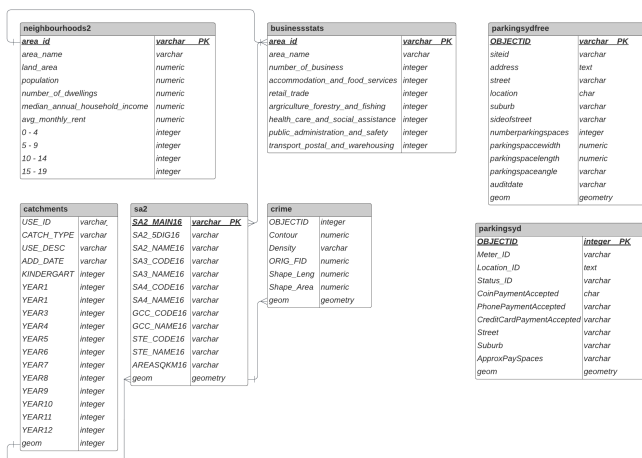


Fig. 1. Entity Relationship Diagram

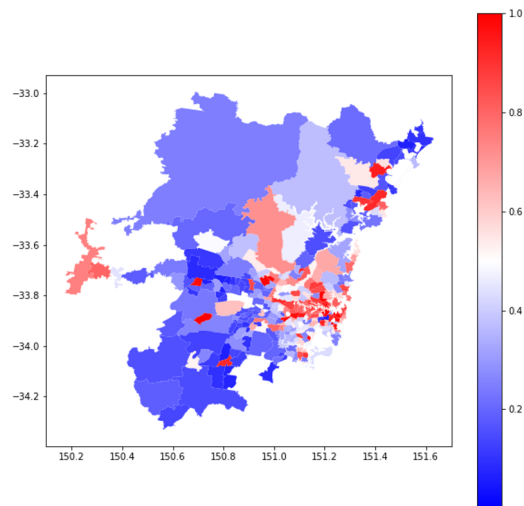


Fig. 4. liveability score heatmap of Sydney

Table 2. liveability score last 5

area_name	score
blue haven - san remo	0.065231
claymore - eagle vale - raby	0.063820
lethbridge park - tregear	0.058963
surry hills	0.027200
redfern - chippendale	0.000110

Table 3. liveability (with parking) top 5

area_name	score
surry hills	0.9951965
sydney	0.9943203
pyrmont	0.9628987
darlinghurst	0.9399703
ultimo	0.7843653

References

Jupyter notebook

- [1] Arefi, M., & Nasser, N. (2021, February 26). Urban Design, safety, Livability, & Accessibility - Urban Design International. Retrieved May 23, 2022, from <https://link.springer.com/article/10.1057/s41289-021-00155-9>
- [2] The importance of parking in planning for Livable Communities - CMAP. (n.d.). Retrieved May 23, 2022, from https://www.cmap.illinois.gov/updates/all/-/asset_publisher/UImfSLnFfMB6/content/the-importance-of-parking-in-planning-for-livable-communities
- [3] Liveability dimensions and attributes: Their relative importance in the ... (n.d.). Retrieved May 22, 2022, from https://www.researchgate.net/publication/46817848_Liveability_dimensions_and_attributes_Their_relative_importance_in_the_eyes_of_neighbourhood_residents
- [4] Measuring the livability of an urban centre: An exploratory study of ... (n.d.). Retrieved May 22, 2022,

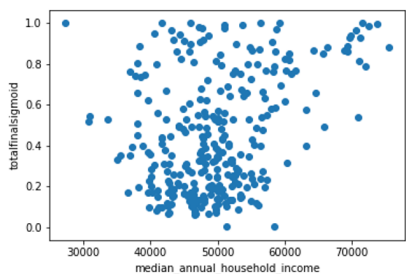


Fig. 2. Scatter plot of median income and liveability score

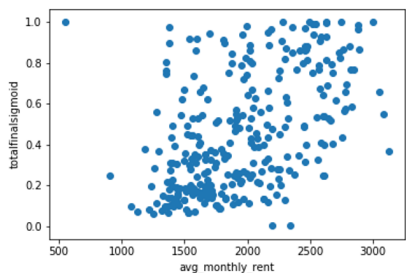


Fig. 3. Scatter plot of monthly rent and score

Table 1. liveability score top 5

area_name	score
sydney - haymarket - the rocks	1.000000
badgerys creek	0.999999
north sydney - lavender bay	0.999998
chullora	0.999985
darlinghurst	0.999866

from https://www.researchgate.net/publication/232970553_Measuring_the_livability_of_an_urban_centre_An_exploratory_study_of_key_performance_indicators

[5] National Academies of Sciences, Engineering, and Medicine. 2002. Community and Quality of Life: Data Needs for Informed Decision Making. Washington, DC: The National Academies Press. <https://doi.org/10.17226/10262>.

[6] Neighborhood Livability in Northwest Portland: A case study of portland ... (n.d.). Retrieved May 22, 2022, from https://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?httpsredir=1&article=1125&context=cus_pubs

[7] Street livability and beautification in an era of disruptive transport ... (n.d.). Retrieved May 22, 2022, from https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID4006498_code2568092.pdf?abstractid=4006498&type=2