WHERE IS MOST LIVABLE PART IN MELBOURNE

IBM COURSERA CAPSTONE PROJECT

OCT 2018

BY RUIPING DU

CONTENTS

1. Introduction
2. Description of Data
3. Methodology
4. Results
5. Conclusion
6. Discussion

Introduction

The Challenge:
The very first challenge that new immigrants face when they first choose to move to Australia is which suburb that they should choose to settle down. Many factors should be incorporated to make a well-informed decision. They factors could includes: Housing Affordability: How much is the average housing pricing in the suburb? School Accessibility: Are there any good public schools in the suburb? Food Accessibility: Are there any good restaurant with in the suburb? Coffee Accessibility: All Melbournian love coffee. This seems a must-have. Other Facilities: such as parks, movie theatres, etc. In this project, we're aiming to identify the most liveable community in Melbourne filtering by the criteria that we have just listed above. Australia is a big country. To perfectly address the above problem, we need quite a big integrated dataset sourcing from multiple channels. Our project is only aiming to prove the feasibility of such a task. Hence, our focus will be on the metropolitan area of Melbourne (Victoria state) ONLY.

The Target Audience:
Our analysis is aiming to provide informed recommendations to someone who have been new to Melbourne and what to choose where they want to settle down. Of course, people can eaisly find answers online nowadays. However, most of the answers are based on personal experience or preferences. We barely see much data-driven backed analysis and recommendation.
Ideally, someone who has been living in Melbourne for a long while might still find our analysis interesting. As it might provide some extra insights to them or our insights might not reasonate with their gut-feel. This might triger his / her thinking. We'd consider this is a value-add of our work.
Moreover, due to the time constrains and painful process of collecting the data, we'd encourage whoever think that our analysis is interesting and wants to explore more in other regions or other methodology to try to tackle the same challenge. In this case, we hope that our work can my a bit of contribution to the whole data science commjunity

Description of Data

FourSquare.com

- The data help us to explore coffee, food, and many other facilities and factors that can determine whether a community / suburb is liveable or not.
- Source of data : The data were retrieved through Foursqaure.com API. For more details on Foursquare API usage, please visit the documentation for developers pages

Australian Bureau of Statistics

- The data provide Australian location data in the format of geometry (points or multi-polygon) which define the boundary of a given suburb. The below image will show how ABS defined the boundary.
- Source of data: For information of the data and format, please visit ABS website at (http://www.abs.gov.au)

Victorian School Ranking

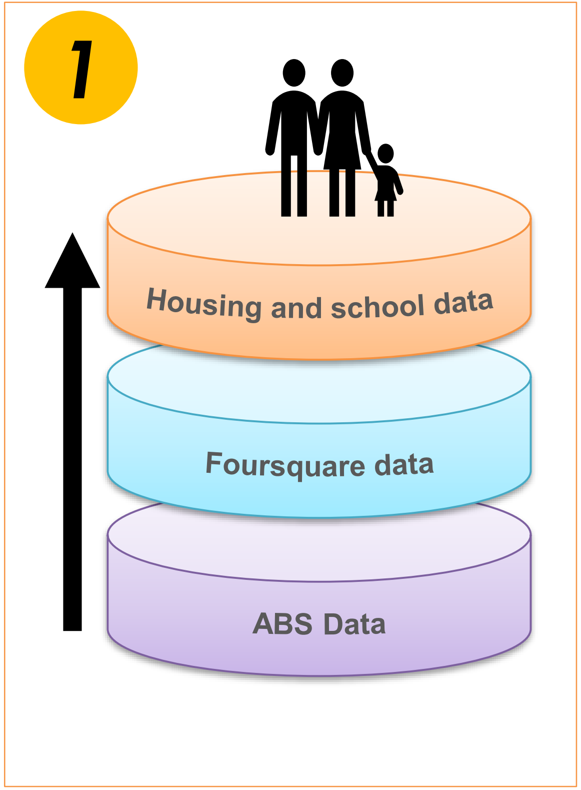
- The data contains the information of all public shcools with geocoded address.
- Source of data : The information were gained from Better Education

Melbourne Residential Property Transaction

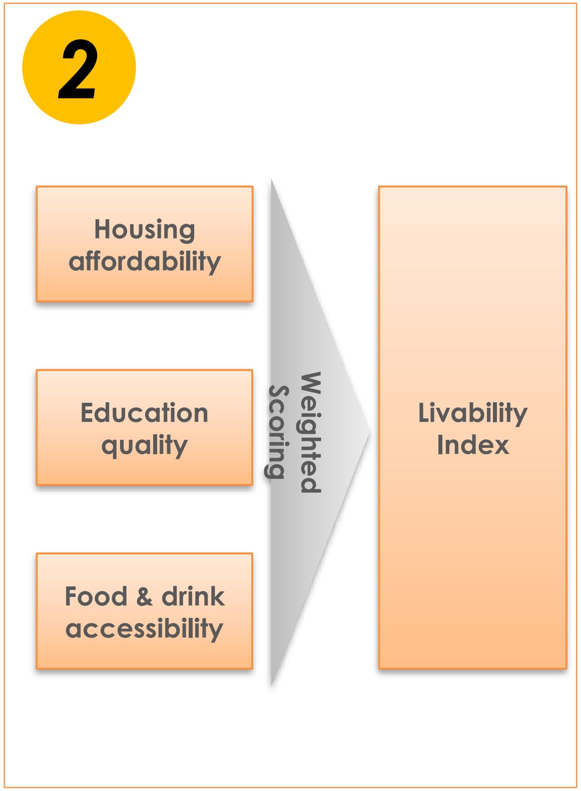
- The data contains all the residential property transaction data in year 2017. The data were retrieved through web scraping. The data contains basic information such as the locations of property, types of property (e.g. house, unit, flat, apartment etc), attributes of property (e.g. number of bedrooms, parking space etc), date sold (timestamp), price sold, agency of the transaction, etc.
- Source of data : Web scraping...

Methodology

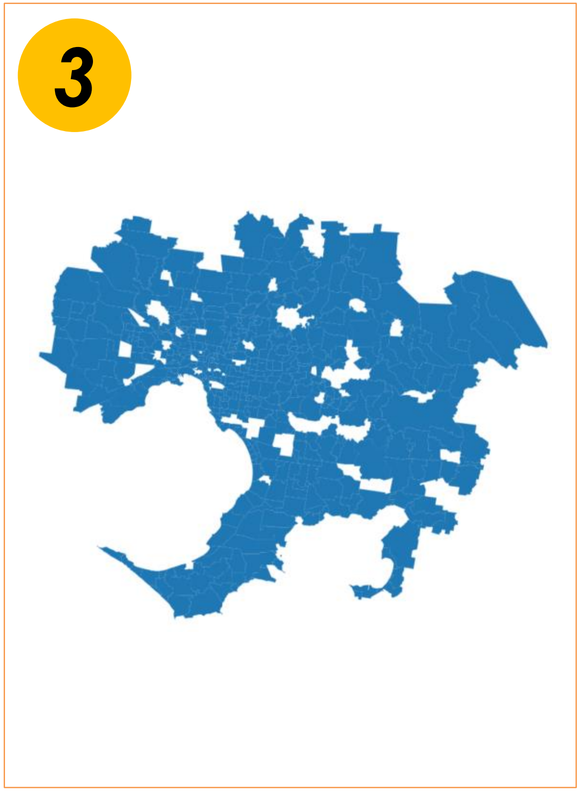
Integrating data from multiple sources



Defining measuring metrics



Visualizing findings



BACKUP

Description of Methodology

Step One: Integrating Data From Multiple Sources

This step is also considered the data processing stage, which is usually the most time consuming stage besides data retrieving. As mentioned above in the data description section. We have data from multiple sources i.e. the property transaction data which will be used to calculate the housing affordability; the school data which will be used to calculate the education quality; the Foursquare data which will be used to calculate the food & drink accessibility. Because the data contains lots of trash data, namely the data that is not considered usuful for our analysis such as transaction agencies, and the data is not in the format we expected. So, we need to spend time cleansing the data, then integrate them into the format that we really want to use. This step contains the sub-steps:

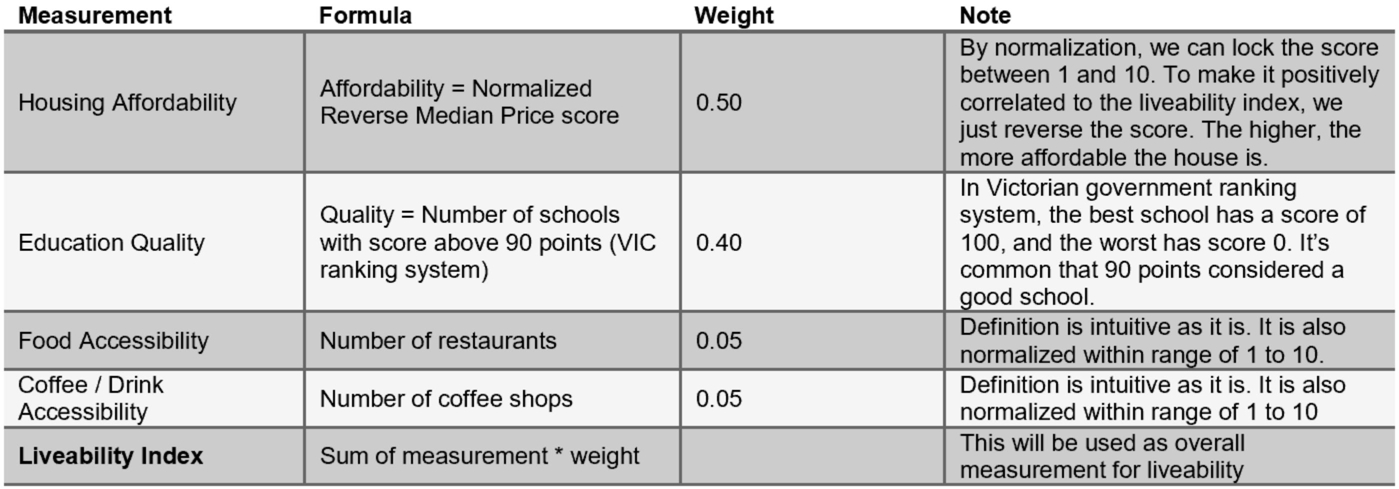
•Data cleansing

•Data integration

Step Two: Defining Measuring Metrics and Calculating The Liveability Index.

This step is the main part of our analysis. Based on the data we retrieved, we will calculate a newly integrated measurement that we call Liveability Index. Intuitively, the high the index, the more liveable the community is.

Here's the approach that demonstrates how we calculated the liveability index on a suburb / community level.



Measurement	Formula	Weight	Note
Housing Affordability	Affordability = Normalized Reverse Median Price score	0.50	By normalization, we can lock the score between 1 and 10. To make it positively correlated to the liveability index, we just reverse the score. The higher, the more affordable the house is.
Education Quality	Quality = Number of schools with score above 90 points (VIC ranking system)	0.40	In Victorian government ranking system, the best school has a score of 100, and the worst has score 0. It's common that 90 points considered a good school.
Food Accessibility	Number of restaurants	0.05	Definition is intuitive as it is. It is also normalized within range of 1 to 10.
Coffee / Drink Accessibility	Number of coffee shops	0.05	Definition is intuitive as it is. It is also normalized within range of 1 to 10
Liveability Index	Sum of measurement * weight		This will be used as overall measurement for liveability

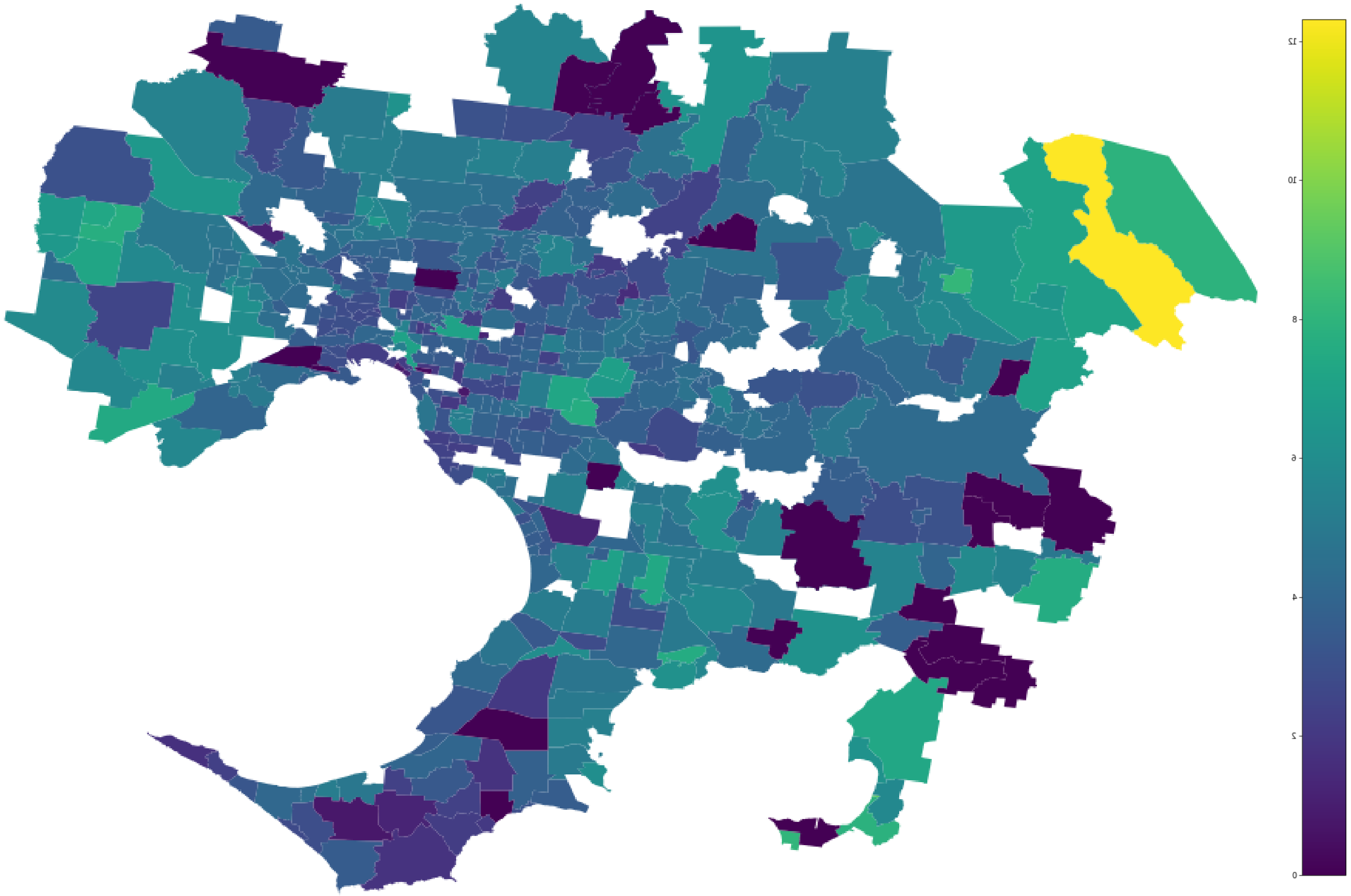
Step Three: Visualizing The Findings

In this step we will use Folium to develop a choropleth map plot where the higher the suburb is liveable. The darker the toner of colour will be.

Results

Top 10 Most Livable Suburb in Melbourne
McMahons Creek
Pioneer Bay
Millgrove
Coronet Bay
Reefton
Grantville
Melton
Wheelers Hill
Longwarry
Cannons Creek

The Liveability of Melbourne



Conclusion and Discussion

Conclusion

As shown in previous slide, we've listed the top 10 most liveable community in Melbourne based on the "Liveability Index" we set and data mining. More details can be viewed via the choropleth plot and the Jupyter notebook that we put online.

Some might found part of the findings contradict their experience. These could mainly due to:

- We used a high weight (60%) on housing affordability as our target audience is for new immigrates
- The limit of data availability. This is mainly due to the restriction of personal Foursqaure subscription. As personal account, we can only retrieve limit amount of data. Hence, in some community, the food and coffee accessibility might be undermined.

In general, we think we've done a good job to provide a data-driven analysis to provide guidance to our target audience.

Discussion

Please consider our analysis a quick-and-dirty approach, which is mainly constrained due to time limitation. For anyone who's reading this report and is interested in performing a similar analysis. The following comments might be worth noting.

Limitations:

•Our analysis didn't consider other factors. As we know from life experience, to define whether a place liveable. It can rely on a wide range of factors such as safety (e.g. low crime rate), culture, other facilities like museums, movie theatres etc. All these factors have not been incorporated in our analysis due to the unavailability of data and time contains. Besides, we're only using a personal account to access data from Foursquare which has its own restrictions

•Calculation of index can be improved. We used a simplified calculation to make the analysis task easier. We'd recommend an refined definition and calculation for anyone who's like to go a extra mile.

Appendix: Resources

Here are some resources that have been used for this task:

Geopandas: Geopandas is a pandas like repository that makes processing and analyzing geospatial task easy. We mainly used its spatial join function to analyze the spatial relationships of the four datasets that we used. It has embed matplotlib as its bundle visualization tool. Most of the plots in our notebook are made by geopandas.
For more information, please visit its official documentation website at http://http://geopandas.org/

Pandas: Needless to say. This is a must-have for python data scientist.

PostgreSQL / PostGIS: The Australian Bureau of Statistics data were provided in Postgis format. We have restored it into our local PostGIS database then convert the data shapely format in our analysis
For more information, please visit its official website at https://postgis.net/

Json: To load and write to Json format file