



Northeastern University

College of Engineering

IE-7945 Master's Project

PROJECT REPORT

TABLEAU DASHBOARD OF NEIGHBORHOOD SELECTION TOOL

By
Sahar Tariq
12/19/2020

Contents

1	ABSTRACT	3
2	INTRODUCTION.....	3
3	SCOPE	3
4	DATA DESCRIPTION.....	3
5	DATA PREPROCESSING	5
6	TABLEAU FRAMEWORK.....	5
6.1	CALCULATIONS	5
6.2	VISUALIZATION	6
7	FUTURE IMPROVEMENTS.....	7

1 ABSTRACT

This report describes the backend data and the front-end Tableau UI of a personalized neighborhood selection tool based on housing, socioeconomic and lifestyle metrics.

2 INTRODUCTION

Surveys done by the most popular housing websites in the US (such as Zillow, Trulia, Housing Wire, Red Fin) show that millennials move once every two years. Millennials don't live their whole lives in the same neighborhoods they grew up in, focusing rather on curated experiences that perfectly suit their needs.

Although there is ample data available online about most neighborhoods in cities, it is difficult for a non-analyst to know which is the most reputable source for their decision making or where exactly to get this data from and how to interpret it. Also, important variables may be overlooked when brainstorming what variables one should consider. And the most difficult task - there may be many variables to consider when choosing the right neighborhood, but it is difficult to mentally compare all these variables objectively.

The purpose of this tool is to provide a survey to users about their neighborhood preferences with pre-built options. As the user chooses their preferences, dynamic visualizations display calculated metrics and ratings related to which neighborhood is the best for the user based on their preferences. The tool uses parameters, and a multiple linear regression whose coefficients are dependent on the user's inputs.

3 SCOPE

This proof-of-concept dashboard is limited to 5 zipcodes (02138, 02139, 02140, 02141, 02142) in Cambridge, Massachusetts, USA.

Data cleaning is done using R programming software and UI is created in Tableau.

4 DATA DESCRIPTION

The following table lists the data sources and data characteristics of the variables.

The main sources is the City of Cambridge's open data initiative.

Table 1: Data Description

	Variable	Data Description	Data Source
1	Land Mass (sq-mile)	By zipcode.	http://www.city-data.com/zip/02142.html
2	Population Density (people per sq mile)	By zipcode	http://www.city-data.com/zip/02142.html
3	Cleanliness Rating (0-5 scale)	Average number of rodent sightings reported to city 311 helpline per year by zipcode. This value is inversed and normalized to a 0-5 scale with 5 being score for least number of rodents seen and 0 being score for most number of rodents seen	https://data.cambridgema.gov/Public-Works/Commonwealth-Connect-Service-Requests-Rodent-Sight/gzbv-wgij
4	Safety (0-5 scale)	The crime rate per 10,000 people of 3 crime categories is aggregated by reporting zipcode.	https://data.cambridgema.gov/Public-Safety/Crime-Reports/xuad-73uj

	<ul style="list-style-type: none"> Disorderly Behavior and Drugs Rate (per 10,000 People) Theft Rate (per 10,000 People) Street Harassment, Assault, and Robbery Rate (per 10,000 People) 	This value is inversed and normalized to a 0-5 scale with 5 being score for least crime and 0 being score for most crime	
5	Car road safety (0-5 scale) <ul style="list-style-type: none"> Auto Theft Auto Accidents Maximum Traffic Count 	The crime rate per 10,000 people of 2 crime categories per reporting zipcode, and the average number of cars that cross through the zipcode per day is aggregated. This value is inversed and normalized to a 0-5 scale with 5 being score for least crime and traffic, and 0 being score for most crime and traffic	https://data.cambridgema.gov/Public-Safety/Crime-Reports/xuad-73uj https://data.cambridgema.gov/Traffic-Parking-and-Transportation/Average-Daily-Traffic-Counts-1972-to-2017/v43b-kqeq
6	Socioeconomic Rating (0-5 scale) <ul style="list-style-type: none"> Median Household Income (\$) Percent Residents Below Poverty Line (%) Percent Residents Unemployed (%) 	The median household income, and the inverses of the percent residents below poverty and of the percent residents unemployed per zipcode are aggregated. This value is normalized to a 0-5 scale with 5 being score for best socioeconomic standing, and 0 being score for worst socioeconomic standing	http://www.city-data.com/zip/02142.html
7	Higher Education Level (0-5 scale) <ul style="list-style-type: none"> Percent of population with a Bachelors degree (%) Percent of population with a Masters degree (%) 	The two variables per zipcode are aggregated and normalized to a 0-5 scale, with 5 being best	http://www.city-data.com/zip/02142.html
8	Median Age of Residents	By zipcode	http://www.city-data.com/zip/02142.html
9	Number of Universities	By zipcode	https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_metropolitan_Boston
10	Number of PreK-Grade5 Schools	By zipcode	https://data.cambridgema.gov/General-Government/Cambridge-Public-School-Locations/jhbq-dj88
11	Number of Grade6-8 Schools	By zipcode	
12	Number of Grade9-12 Schools	By zipcode	
13	Number of Grocery and Convenience Stores	By zipcode	https://data.cambridgema.gov/Economic-Development/Open-and-Closed-Businesses-During-Covid-19-Pandemi/9q33-qjp4
14	Number of ER/Urgent Care	By zipcode	Google Maps
15	Number of Restaurants	By zipcode	https://data.cambridgema.gov/Licensing/Eating-Establishment-Business-Permits/x6ni-wtbi
16	Number of Parks	By zipcode	https://data.cambridgema.gov/Geographic-Information-GIS-/Master-Addresses-List/vup6-kpww
17	Average Rent of 1-Bedroom House (\$)	By zipcode	http://www.city-data.com/zip/02142.html

18	Median Price of Buying a House (\$)	By zipcode	http://www.city-data.com/zip/02142.html
19	Median Property Tax (\$)	By zipcode	http://www.city-data.com/zip/02142.html

5 DATA PREPROCESSING

Data Preprocessing steps taken:

- All variables are aggregated by zipcode level
- For data obtained by neighborhood level, they are converted to zipcode level by cross referencing
- All variables are normalized to 0-1 scale with the following formula, prior to any calculations to avoid issues with widely variant variable scales

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

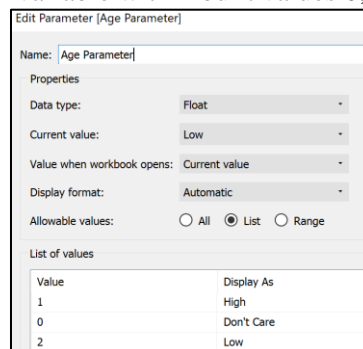
- Data is cleaned of any symbols or discrepancy in formatting

6 TABLEAU FRAMEWORK

6.1 CALCULATIONS

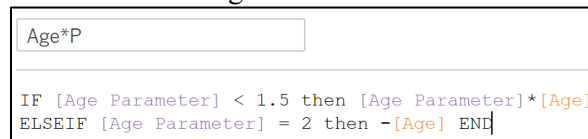
To build the user preferences input framework, the following steps were done:

- A parameter was created for each variable with inbuilt values 0,1,2.



List of values	
Value	Display As
1	High
0	Don't Care
2	Low

- A calculation was created for each variable multiplied by the parameter, with the condition that if the user chooses option “High” then a 1 is assigned to the variable’s coefficient, if user chooses option “Don’t Care” then a 0 is assigned to the variable’s coefficient to remove it, and if user chooses option “Low” then a -1 is assigned to the variable’s coefficient to penalize for it.



```

Age*P

IF [Age Parameter] < 1.5 then [Age Parameter]*[Age]
ELSEIF [Age Parameter] = 2 then -[Age] END

```

- Area Score metric was created with all the variables by using multiple linear regression where $y = \sum \text{parameter}_i * \text{variable}_i$

$$\begin{aligned}
 & [\text{Land} * P] + [\text{PopDensity} * P] + [\text{Cleanliness} * P] + [\text{Safety} * P] + [\text{CarRoad} * P] \\
 & + [\text{Socioeconomic} * P] + [\text{Age} * P] + [\text{HigherEdu} * P] + [\text{PreK-5} * P] + [\text{Grade6-8} * P] \\
 & + [\text{Grade9-12} * P] + [\text{University} * P] + [\text{Grocery} * P] + [\text{ER} * P] + [\text{Restaurant} * P] \\
 & + [\text{Park} * P] + [\text{Rent} * P] + [\text{Buy} * P] + [\text{Tax} * P]
 \end{aligned}$$

- The same method was also used to create Category Scores for each grouping of variables
- Both Area Score metric and also Category Scores were normalized such that every time new values are calculated for the zipcodes based on new user preferences, the values are scaled to a 0-5 range, with 0 being worst performer and 5 being best performer, per that user's needs.

Totals summarize values from Table (across).

$$\frac{((\text{MEDIAN}([\text{xMetric}]) - \text{TOTAL}(\text{MIN}([\text{xMetric}]))) / (\text{TOTAL}(\text{MAX}([\text{xMetric}]) - \text{TOTAL}(\text{MIN}([\text{xMetric}]))) * 5)}{1}$$

- A ranking of the Area Score metric was also created to display 1st, 2nd, 3rd, 4th, and 5th performing zipcode per the user's preferences

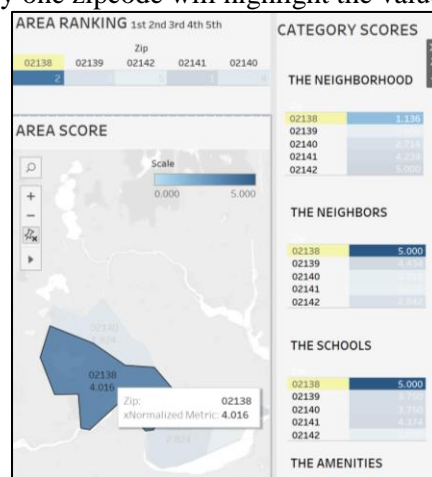
Results are computed along Table (across).

$$\text{RANK_DENSE}([\text{xNormalized Metric}])$$

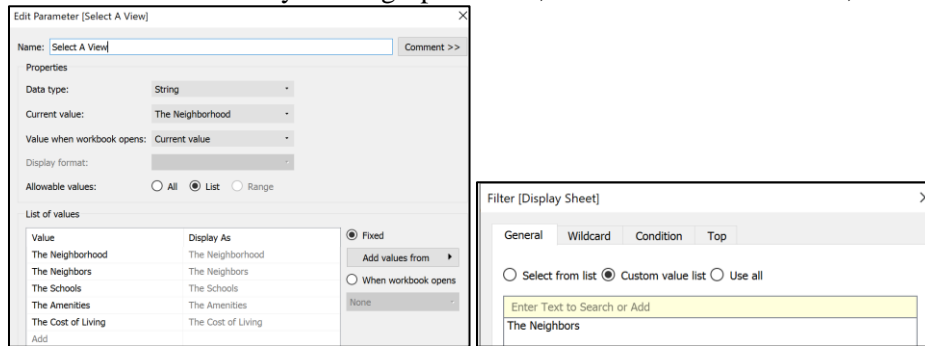
6.2 VISUALIZATION

The calculations described in section 6 were then visualized such that:

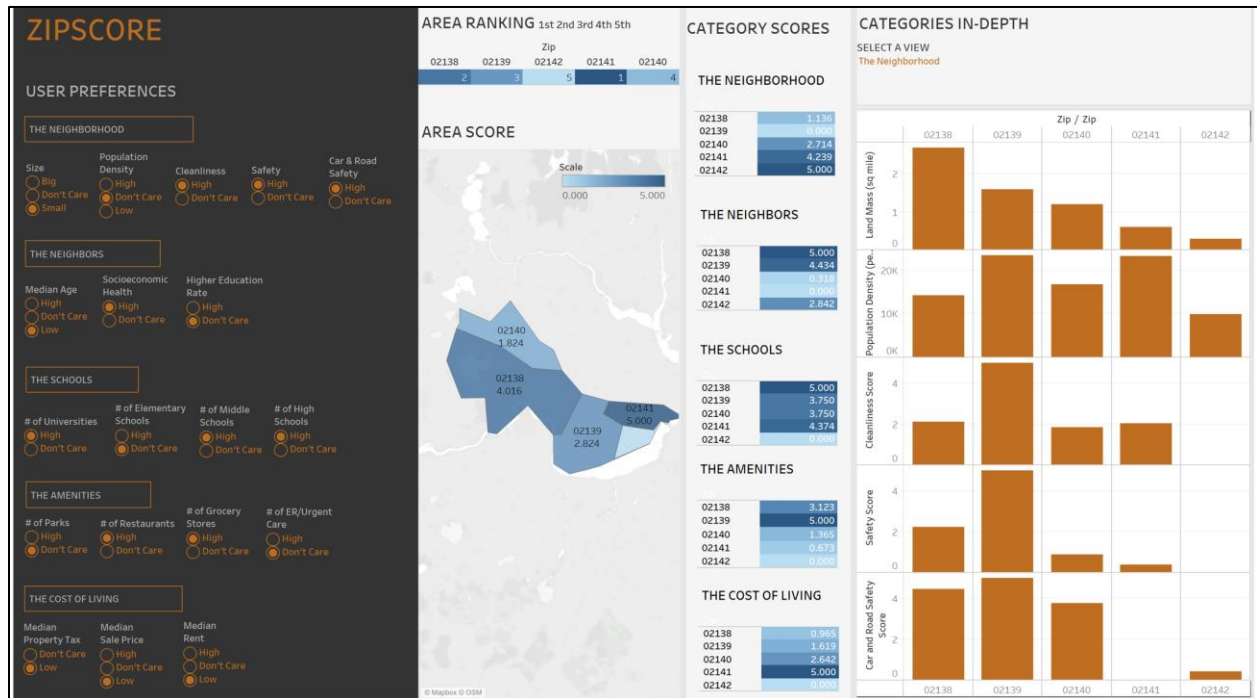
- Black left panel is for user input
 - User can click to choose their preferences
- White mid panel (blue) shows calculated values per those inputs
 - Area Score metric and Category Scores metrics, normalized to 0-5 range are displayed, with 5 (and darkest blue) being best score.
 - Ranking of Area Score metric 1st, 2nd, 3rd, 4th, 5th is also displayed
 - Hovering over any one zipcode will highlight the values for that zipcode in all the plots



- White right panel (orange) shows in depth raw values of each variable
 - Drop down menu allows user to choose which category of variables to view the in depth raw data for. This is done by creating a parameter, worksheets for each view, and filters



The final Dashboard is the following:



7 FUTURE IMPROVEMENTS

- Order Area Ranking ribbon by Rank instead of by Zip. Currently Tableau does not allow for sorting by a Table Calculation
- Edit orientation of Categories in Depth plots to be sideways instead
- Include more zipcodes
- User survey and record how users interact with UI to understand how to further optimize UI