CSE 6242 – Where would you LIVE next?

Team Number: **60** Team Name: **GA-30223**
Sanjay Mathan, Sreejisha Purushotham, Rukumani Rimal, Kha D Tran

Final Report

Introduction and Problem Statement

As our urban environments expand and adapt, the notion of a "livable city" has become paramount (Camanho et. al, 2015; Mittal et. al, 2020; Kovacs-Györi et.al, 2020). A city's livability is not just determined by tangible metrics; it is also shaped by the feelings and preferences of its inhabitants, considering what they deem as the ideal city for their distinct lifestyles, preferences, and needs. Traditionally, the livability of cities has been evaluated based on sociodemographic data only, focusing on elements like healthcare, infrastructure, education, and safety (Camanho et. al, 2015; Mittal et. al, 2020; Kovacs-Györi et.al, 2020; C. J. L, Balsas, 2004; Greenberg et. Al, 2001, Greenberg, 2002, Kutty et. Al, 2022). While these metrics are vital, they may not always encapsulate the genuine feelings and experiences of the city's residents.

Our team will explore, research, and create a visualization to find the best US city to live in based on a variety of livable criteria ranging from weather, crime rate, education, politics, etc.

Literature Survey

Li argues that non-native residents tend to have higher satisfaction and livability scores than native residents. Foreign-born residents tend to live closer to their workplaces, shopping, transportation, and educational institutions. Cultural and language barrier often motivates immigrants to live closer to more concentered areas **(Li, 2012)**.

Earlier studies also argue metrics of livable places using the urbanization design such as imageability, legibility, and linkage. **(Clemente et.al, 2013)** While these have their own merits, we have chosen not to incorporate them into our project, as we are inclined to use other metrics that are data-dependent for a more subjective measurement.

Many researchers have explored the city's livability using socio-economic data **(Bertrand et.al, 2013; Balsas, 2004)**. Balsas analyzed a set of KPI to measure the city-center livability. He discussed how urban livability has transformed in the Western world over decades **(Balsas et. al, 2004)**. These KPIs can be complex, yet vital to gain insights into the health and well-being of a society. Urban livability indicators must consider qualitative and quantitative aspects for com- prehensive coverage. **Okulicz-Kozaryn et. al.** compared urban livability versus satisfaction to measure the quality of life. The city rankings often focus on objective factors, such as education, housing, transportation, and health care. However, what matters as well are the subjective factors such as openness, trust, creativity, and innovation. Livability rankings would be different for each person based on a combination of both objective and subjective factors **(Okulicz-Kozaryn et. al, 2013).** Zanella proposes a conceptual model based on broad social and economic factors along with the principles of environmental sustainability. This model factors in

unconventional attributes such as solid waste and air pollution to represent environmental impact **(Zanella, 2015).**

However, there is only a handful of research done using the big data approaches. **Kovacs-Györi et.al.** tried to assess urban livability analysis in the age of big data from disparate platforms like social media, IoT, etc. In his research, he analyzed geospatial data to understand human behavior and improve urban living. Mittal derived 8 fundamental criteria for addressing the preferred attributes for an ideal model that could typically be characterized as a tool. Mittal conducted extensive review of 26 models related to urban quality of life. **(Mittal, 2020)**

Another key aspect of city livability is the crime rate, as explored by Chang and Kim's research. Their research reveals that violent crimes tend to be higher than property crimes overall. The research found crime rates tend to increase at a slower pace relative to population. We plan to utilize that research in our project, separate out the property and violent crime, as well as focus on the top 50 cities by population for our filter. **(Chang and Kim, 2019)**

Finally, we observed most of the projects aimed to rank cities using similar criteria, each with its own ranking and weighted calculation, and the results are static. As noted by Garret, 'user experience is vital to all kinds of products and services.' While Garret's book doesn't provide specific details on how to implement a good UI design for our project, it offers valuable concepts that we can leverage to achieve our goals **(Garret, 2010)**.

Proposed Method

List of Innovation

1. To rate any given geographical area/county/city on the quality of the educational system, we pulled datasets from the US Department of Education that provided assessment proficiency in Mathematics and Reading/Language Arts for every single elementary, middle, and high school at every grade. We distilled the necessary percentage of students who attained proficiency and aggregated it at the county level. Further, MSA to county mapping from the US Department of Commerce – BEA helped in rolling up the educational system rating for any given MSA.
2. Additionally, we pulled data from the US Department of Housing & Urban Development (HUD) for the urban living model, the National Weather Service (NWS) for the weather model, and the US Environment Protection Agency (EPA) for the commute scoring model, as well as the US Department of Justice for the crime scoring model.
3. Our goal was to offer a simple interactive web interface with minimal control variables using sliders over the US Map on key livability factors such as Weather, Crime, Entertainment, and Education. The control variables connected to a condensed score per variable extracted from densely populated publicly available datasets. This framework could be extended to add more control variables to offer users the flexibility to customize subjective individual preferences on the ideal city to live in.

Data Preparation, Analysis and Curation

For this project we pulled large datasets from various US Government sources such as US Department of Education (ED), US Census Bureau, Bureau of Labor Statistics (BLS), Environmental Protection Agency (EPA) and Nation Weather Services. We then cleansed and curated the datasets to geographical area/county/city information. We included larger metropolitan statistical area (MSA) for the final visualization and excluded smaller cities by population as they do not have all the data that are publicly available.

During our analysis, we ran into a variety of issues dealing with publicly available data. For instance, crime data is not completed, with some jurisdictions failing to report the data to a centralized source (FBI). (source here). As such, we scrap the data from independent crime data sources such as neighborhood scout, as they gather these data by themselves from 18,000+ local law enforcement agencies. In practice, any projects should have funding allocated to purchase these data directly from their API.

We gathered data from the Department of Education (ED) on math and reading proficiency percentage for every single school in the US. Over 7.6 million proficiency scores were evaluated at each grade and school levels. This dataset does not contain MSA related information however, it contains the school district ID, city, state & zip code information. To get the MSA information, dataset from US Census Bureau was combined. We experimented enriching our educational dataset with additional data from National Center for Educational Statistics (NCES) on the number of Student and Teachers per school. Upon evaluating this merged dataset, we found there were many inconsistencies at grade and school levels with missing and seemingly incorrect number of student and teacher ratio that we decided not to use the merged dataset.

Various data sources we use such as “Data.gov” and “Census.gov” are informative to reflect social and economic attributes in our datasets. However, they can be time-consuming with diminished return as some of the data points are not relevant for this project. As a prime example of this, one of the datasets with 8 million records of weather events from 2016-2022, did not make it to the curated version due to much missing data on both the city and MSA level. As a result, there are a few datasets we must gather by another methodology such as web-scraping.

Then we assign a ranking to each category and calculate the final score, or the livability index. The ranking helps normalize various metrics, for instance, the shorter commute time is better, shorter crime rate is better, but more college degree rate or more population above poverty level rate is better. Some categorical data is for information purposes only, to help align with the user’s unique interest, such as weather, population, etc.

One of the main goals for our project is to emphasize the user experience of the dataset and visualization. We transformed all the datasets to a condensed scoring model for each control variable identified. These control variables are presented as sliders in the Tableau User Interface.

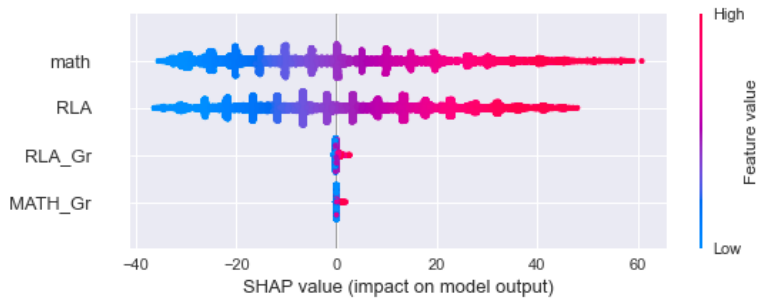
Experiment and Evaluation

In the curated version of the data for visualization, we experimented with different methodologies of looking at the data, and we opt to use different approaches depending on the availability of the datasets, and what the datasets are about.

We experimented with different approaches on aggregating school level proficiency rate at the MSA level. Simply taking the average rating at the MSA level for all schools would bring down the score when there were a few high proficiency schools mixed with many low proficiency ones. However, using the weighted average, we could apply higher weightage to the high proficiency school and lower weightage to the lesser proficiency schools. As part of feature engineering, we added the grade level based upon proficiency percentage and we tested various options of using either using raw percentages or graded score or a combination of both. The combination dataset performed better on the resulting model scoring, which was finally selected to plot the visualization on Tableau.

To evaluate the boosting algorithms using XGBoost, the educational dataset was split into training and testing subsets. We experimented with different combinations of 60/40, 70/30, 80/20 of training and testing the dataset. The Final set of 20% containing 9,839 records used for testing and the 80% with 39,353 records used for training. To minimize Bais vs Variance tradeoff, we decided to keep simple features in the model for better ability to fit the training data and resulting test score.

We used SHAP method (SHapley Additive exPlanations) which is based on cooperative game theory and often used to increase transparency and interpret significant feature contributing to the machine learning models.



As hyper parameter tuning effort we evaluated different objective functions such as:

XGBoost Model - Learning Task Parameters Tuning	Results
objective='rank:pairwise	0.97366203
objective='rank:ndcg'	0.97033936
objective='rank:map'	0.95123687

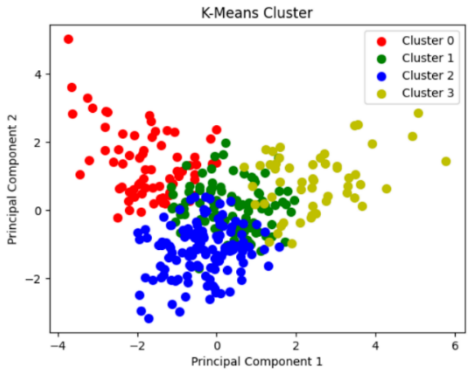
– rank:pairwise: where the pairwise loss is minimized

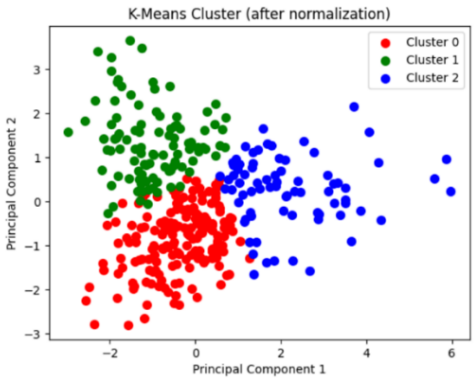
– rank:ndcg: where Normalized Discounted Cumulative Gain (NDCG) is maximized

– rank:map: where Mean Average Precision (MAP) is maximized

Objective function “rank:pairwise” was finally selected for better accuracy.

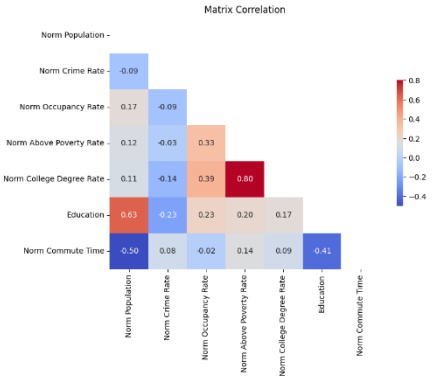
For the final curation of the data, prior to implementing it into our dashboard, we experimented with various techniques when exploring the data. We decided to go with Min-Max Normalization, as it fits better with our goal to have a simple to use visualization. As shown in figure below, when trying to experiment with different methodology for user to pick out the best cities to live in, K-Mean Clusters does an average job of associating some of these attributes together. However, there are way too many outliers for it to become a good method to use. The normalization helps reduce the outliers.





Post normalization also helps make it clearer for the elbow method to use 3 clusters instead of 4 (Appendix 5&6).

Additionally, normalization also helps us see the correlation between each value better. For instance, the correlation between normalized population and normalized crime rate is only -0.09, revalidating the result found in research by Chang and Kim as very large cities may be experiencing lower crime rates compared to middle size city. **(Chang and Kim, 2019).** Further analysis between New York City and Atlanta show that larger cities have lower crime rate compared to middle-sized cities.



Atlanta Annual Crimes			
	Violent	Property	Total
Number of Crimes	4,720	19,914	24,634
Crime Rate (per 1,000 residents)	9.51	40.11	49.62

New York Annual Crimes			
	Violent	Property	Total
Number of Crimes	8,827	33,234	42,061
Crime Rate (per 1,000 residents)	5.21	19.62	24.83

Visualization

The outcome of this project is the UI built on Tableau as an interactive mapping interface that allows users to customize the criteria based on their personal preferences. It also includes a composite livability index that allows users to filter for different ranges of the livability index.

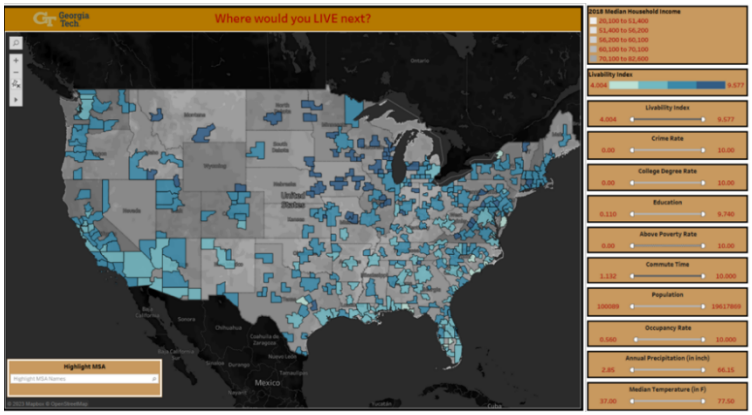


Exhibit 1: View of MSAs with no filtering criteria on where would you LIVE next?

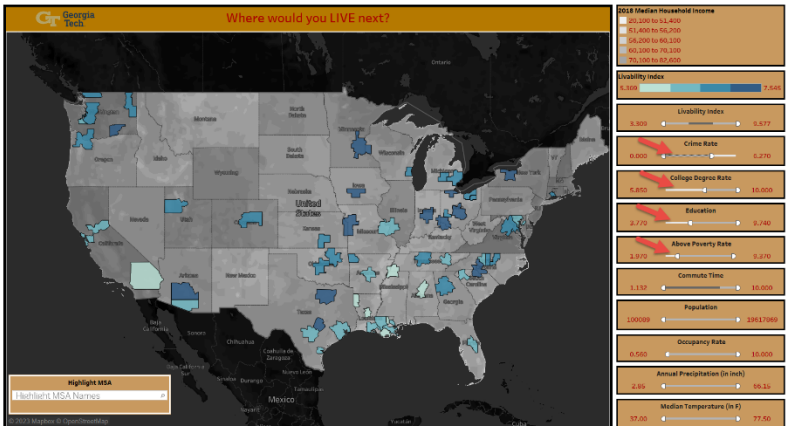


Exhibit 2: View of MSAs matching the desired filtering criteria on where would you LIVE next?

Exhibit 1 & 2, are the Interactive digital map showcasing multiple Metropolitan Statistical Areas (MSAs) matching the user filtering criteria on the custom choices on "Where would you LIVE next?"

On the right side of the map is a filter tool equipped with adjustable slides. When these sliders are set to specific criteria, only certain cities that match these criteria are shown on the map. The map itself is detailed, displaying geographical features, and the filter tool appears modern and user-friendly, emphasizing the map's digital interactivity and functionality.

The image features an interactive component where tooltips appear over each Metropolitan Statistical Area (MSA) on the map. As users navigate the map and hover their cursor over different MSAs, detailed information about each specific area is displayed. This pop-up information likely includes data relevant to the criteria set in the filter tool, such as population statistics, economic indicators, geographic details, or other pertinent attributes that define each MSA. This feature adds an educational and informative layer to the map, allowing users to gain a comprehensive understanding of each area based on the predefined criteria. The design suggests a user-friendly and informative tool, ideal for analysis or educational purposes.

MSA Code: 28140
MSA Names: KANSAS CITY, MO-KS
Principal Cities: Kansas City, Lenexa, Overland Park, Kansas City
Livability Index: 7.4888
Crime Rate: 4.083
College Degree Rate: 8.4472
Commute Time:6.166
Education:7.392
Population: 2,209,494
Occupancy Rate: 8.8258
Above Povery Rate: 7.8105
Average median Temperatire: 56.50

Effort Statement

All team members contributed a similar amount of effort for the project.

Conclusion and Further Research

We developed an initial prototype for the identified control variables and visualized the data. Each of the attributes that we currently use in our visualization can be expanded for detailed analysis and better data from a paid source. While it is a proof of concept, these data points have the potential to expand much further. For instance, just weather alone, there are a variety of different data such as severe weather events, precipitation, snowfall, sunlight, etc. for a user who is interested in a particular type of weather.

In addition, this model can be scaled for other control variables such as utilities, transportation, parks, and recreation etc, accommodation, political party, etc. making it more robust and scalable for further research.

References

Balsas, C. J. L. (2004). Measuring the livability of an urban centre: An exploratory study of key performance indicators. planning, practice research. 19(1). https://doi.org/10.1080/0269745042000246603

Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in new york city: A high resolution spatial and temporal view. https://doi.org/10.48550/arXiv.1308.5010

Chang, S., Y., & Kim, E., H. (2019). Do larger cities experience lower crime rates? a scaling analysis of 758 cities in the us. 11(11), 3111. https://www.proquest.com/docview/2322181841?pq-origsite=primo

Ewing, R., Clemente, O., Neckerman, K., Purciel-Hill, M., Quinn, J., & Rundle, A. (2013). Measuring urban design metrics for livable places, 192. https://link.springer.com/book/10.5822/978-1-61091-209-9

Garrett, J. J. (2010). The elements of user experience, second edition: User-centered design for the web and beyond. O'Reilly. https://learning.oreilly.com/library/view/theelements-of/9780321688651/

Jianxiao, L., Han, B. I., & Wang, M. (2020). Using multi-source data to assess livability in hong kong at the community-based level: A combined subjective-objective approach, 284–294.

Kovacs-Györi, A., Ristea, A., Havas, C., Mehaffy, M., Hochmair, H. H., Resch, B., Juhasz, L., Lehner, A., Ramasubramanian, L., & Blaschke, T. (2022). Opportunities and challenges of geospatial analysis for promoting urban livability in the era of big data and machine (Vol. 9(12)). https://doi.org/10.3390/ijgi9120752

Kutty, A. A., Wakjira, T. G., Kucukvar, M., Abdella, G. M., & Onat, N. C. (2022). Urban resilience and livability performance of european smart cities: A novel machine learning approach. 378. https://doi.org/10.1016/j.jclepro.2022.134203

Li, Y. (2012). Neighborhood amenities, satisfaction, and perceived livability of foreign born and native-born u.s. residents. Journal of identity and migration studies. 6(1), 115–137.

Mittal, S., Chadchan, J., & Mishra, S. K. (2020). Review of concepts, tools, and indices for the assessment of urban quality of life. Social indicators research. 226(1), 187–214. https://doi.org/10. 1007/s11205-019-02232-7

Okulicz-Kozaryn, A. (2020). City life: Rankings (livability) versus perceptions (satisfaction). social indicators research. 110(2), 433–3451. https://doi.org/10.1007/s11205-019-02232-7

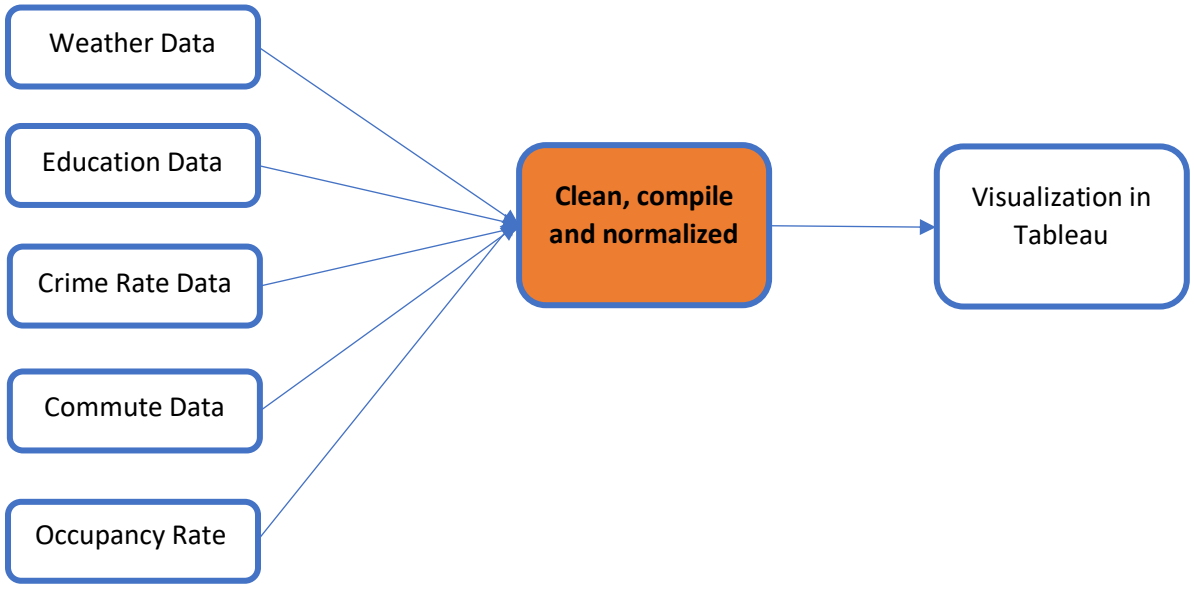
Reid Ewing ewing@arch.utah.edu & Robert Cervero robertc@berkeley.edu (2010) Travel and the Built Environment, Journal of the American Planning Association, 76:3, 265-294, https://doi.org/10.1080/01944361003766766

Sun, Y., & Du, Y. (2017). Big data and sustainable cities: Applications of new and emerging forms of geospatial data in urban studies. O'Reilly. https://opengeospatialdata.springeropen.com/articles/10.1186/s40965-017-0037-0

Zanella, A., Camanho, A. S., & Dias, T. G. (2015). The assessment of cities livability integrating human wellbeing and environmental impact. annals of operations research. 149(1), 695–726. https://doi.org/10.1007/s10479-014-1666-7

Appendix

1. Data Extraction and Ingestion strategy



2. Project Plan

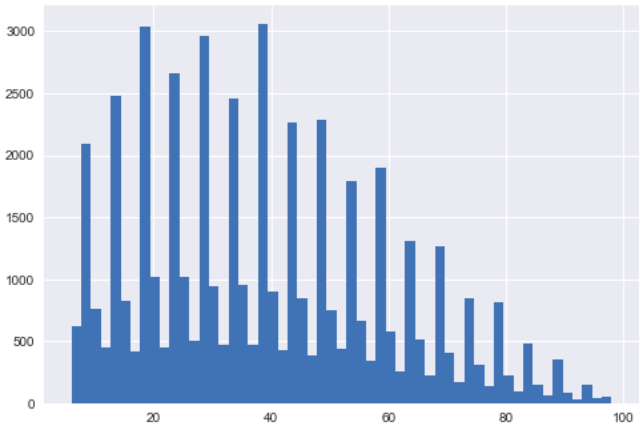
Project Plan

Key Deliverable

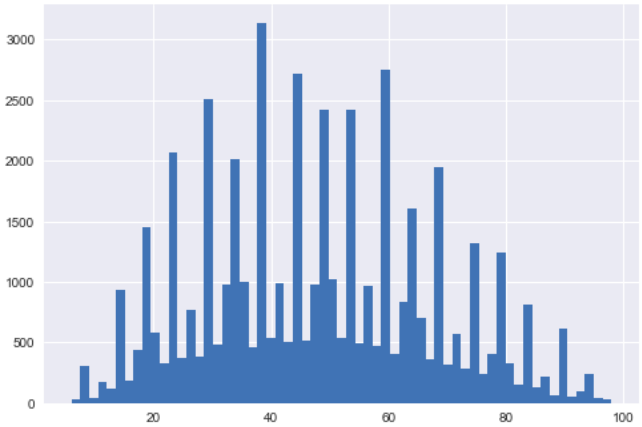
Title	Best City to Live in USA
Team Name	Team #60: GA-30332
DATE	9/20/23

WBS NUMBER	TASK TITLE	TASK Lead	START DATE	DUE DATE	DURATION	PCT OF TASK COMPLETE
1	**Project Conception and Initiation**					
1.1	Team Formation	Kha T	9/20/23	9/22/23	2	100%
1.1.1	Assign Team Name	Sanjay M	9/21/23	9/22/23	1	100%
1.2	Meet & Greet	Team	10/5/23	10/5/23	0	100%
1.3	Choose a Topic	Team	10/9/23	10/9/23	0	100%
1.6	Project Initiation	Team	10/9/23	10/10/23	1	100%
2	**Project Proposal**					
2.1	Scope and Goal Setting	Team	10/9/23	10/11/23	2	85%
2.2	Documentation	Ruku R	10/11/23	10/13/23	2	80%
2.4	Proposal submit	Sree P	10/12/23	10/13/23	1	5%
3	**Project Monitoring-Midterm**					
3.1	Data Cleanse and Extraction	Kha T	10/13/23	10/20/23	7	0%
3.2	Data Analysis	Sanjay M	10/13/23	10/20/23	7	0%
3.2.1	Data Model	Kha T	10/20/23	10/27/23	7	0%
3.2.2	Build visualization	Ruku R	10/20/23	10/27/23	7	0%
3.3	Project Updates	Sree P	10/27/23	11/3/23	6	0%
3.3.1	Updates and documentation	Team	10/27/23	11/3/23	6	0%
4	**Project Performance -Final**					
4.1	Visual Analytics	Ruku R	11/3/23	11/10/23	7	0%
4.2	Test Data	Sree P	11/11/23	11/15/23	4	0%
4.3	Make Poster	Team	11/15/23	11/30/23	15	0%
4.4	Final Poster Presentation	Team	11/20/23	12/1/23	11	0%

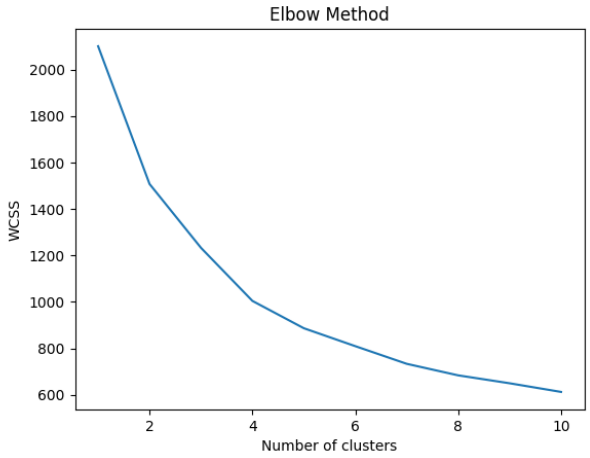
3. Histogram of Educational: Math proficiency scores vs number of observations.

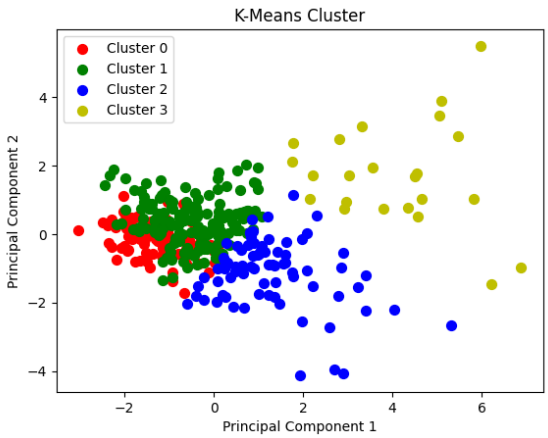


4. Histogram of Educational: Reading proficiency scores vs number of observations.



5. Before normalization to pick the best cities to live:





6. After Normalization to pick the best cities to live: