Where would you LIVE next?

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

INTRODUCTION

A city's livability is determined by tangible metrics such as amenities, commute, cost of living, crime rates, employment, schools, housing and user ratings etc. A user can choose the best place to live based on their distinct lifestyles, preferences, and needs.

GOALS

- Analyze publicly available different datasets and compute using a weighted score to identify ideal city to LIVE in the US, while allowing the user to define the indicators and preferences based on their individual needs.
- Create an easy-to-use visualization to spot the city based on an individual's preferences.

DATASETS

- US Department of Education (ED) (1GB , 7.6M records)
- National Center for Education Statistics (NCES) (20MB)
- US Environment Protection Agency (EPA) (200MB)
- US Department of Commerce – BEA
- Crime data - Web scrapping (70MB)
- Weather Events - (1GB, 8M records)

APPROACH

Curation: We combine all the raw data into one clean dataset, feeding it in Tableau for our visualization.

Normalized: Since the attributes are not in the same scale, we use Min-Max Normalization to normalize the data

Correlation and Clustering: Additional analysis such as correlation and clustering are performed on both the raw and normalized data to inspect the data points.

Visualization: The normalized dataset are feed into tableau dashboard to create an interactive map, allowing users to filter based on their preferences, and showcase the top cities choices.

EXPERIMENT

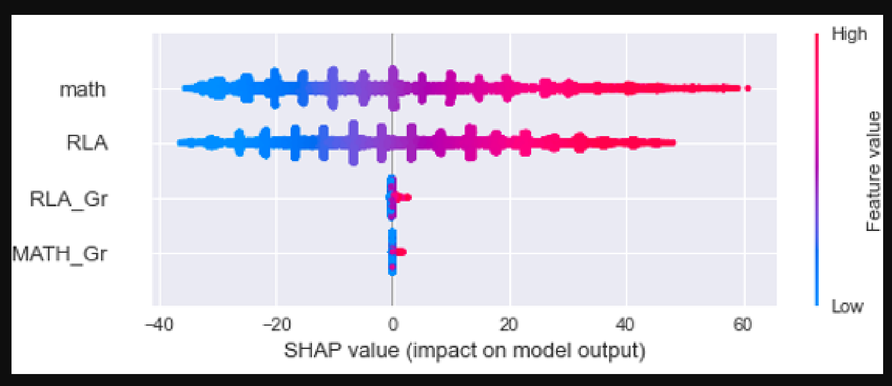
- K-Means clustering are performed to determine number of clusters of MSA. Post normalization shows a reduction in outliers for the clustering.
- Used SHAP method (SHapley Additive exPlanations) to interpret feature contributions to the model.
- Apply Weighted average method keep prominence of higher proficiency schools.
- Used XGBoost to rank the MSAs on educational proficiency.
- Tested with 60/40, 70/30, 80/20 of training and testing the dataset.

EVALUATION

- To minimize Bias vs Variance trade off, we kept simple features in the model for better fit.
- Evaluated hyper parameters to tune different objective functions with “rank” on XGboost.
- We decided to go with Min-Max Normalization, as it fits better with our goal to have a simple to use visualization.
- The normalization helps reduce the outliers, and helps us see the correlation between each variables better.

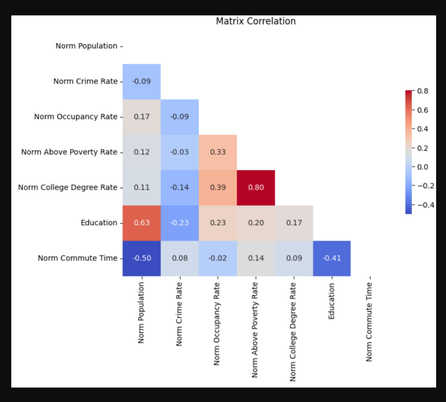
DATA ANALYSIS

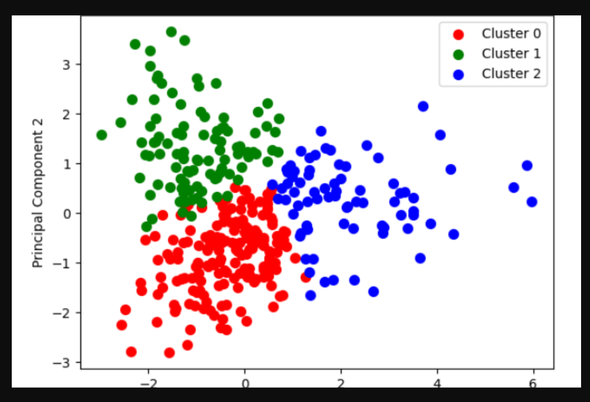
XGBoost Model for Education:



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

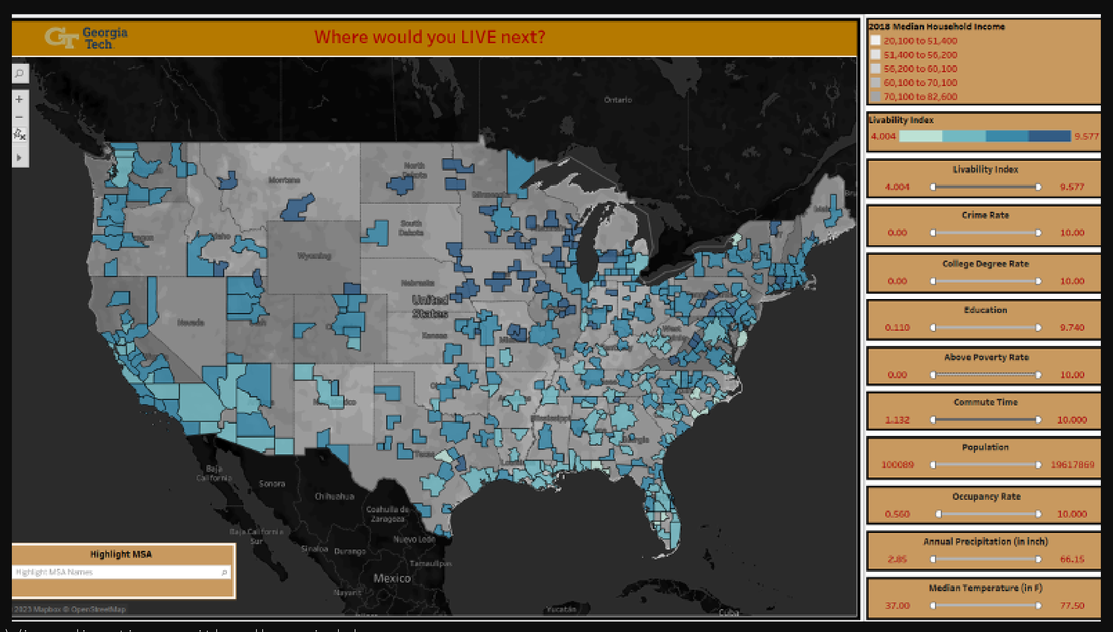
Normalized data correlation



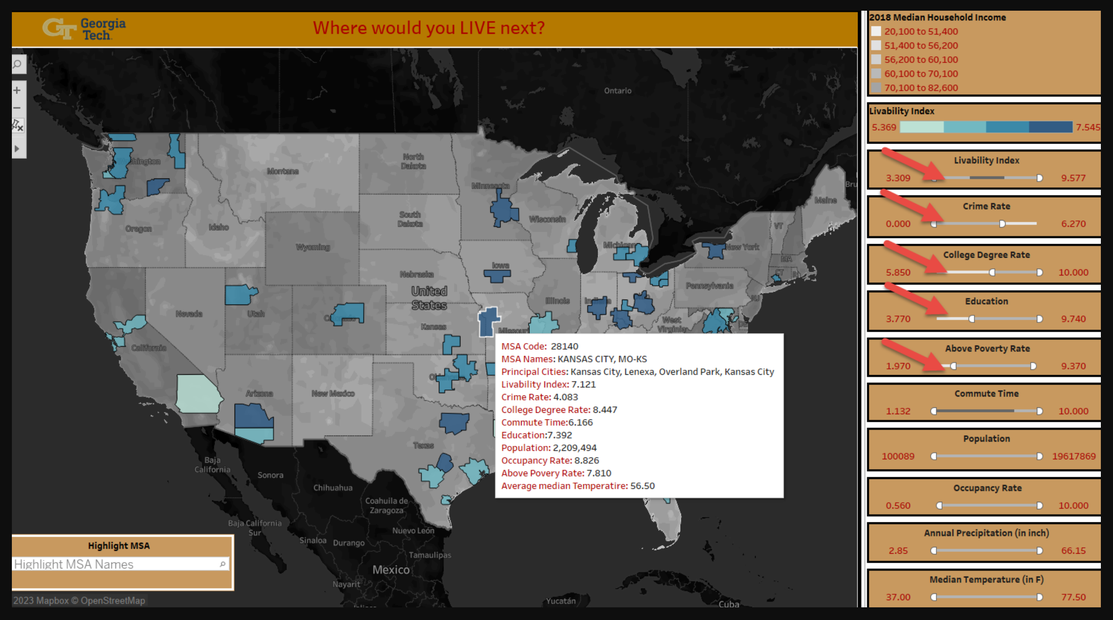


VISUALIZATION

Tableau Map showing all the MSA-area and the curated attributes.



Visualization with all variables



Visualization with variables filter on

SUMMARY

We were able to build a high quality, easy to understand and simple to use controls on determining where would you LIVE next?
This application is capable of scaling with additional data sources and more sophisticated controls on intuitively interpreting the filtered list of MSAs.
Further, this application could be seamlessly integrated with external sites using APIs to consume the curated dataset.