DATA SCIENTIST RELOCATION IDENTIFIER PROGRAM

A PILOT BY HYUN GU KANG, PHD

DECEMBER 4, 2019

WHERE TO GO?

Newly minted data scientists likely need to or have the ability to relocate to new places that meet their professional and personal interests.

Where are the likely places to which they would relocate?

This question is important for both data scientists and employers alike.

STUDY DESIGN AND GOALS

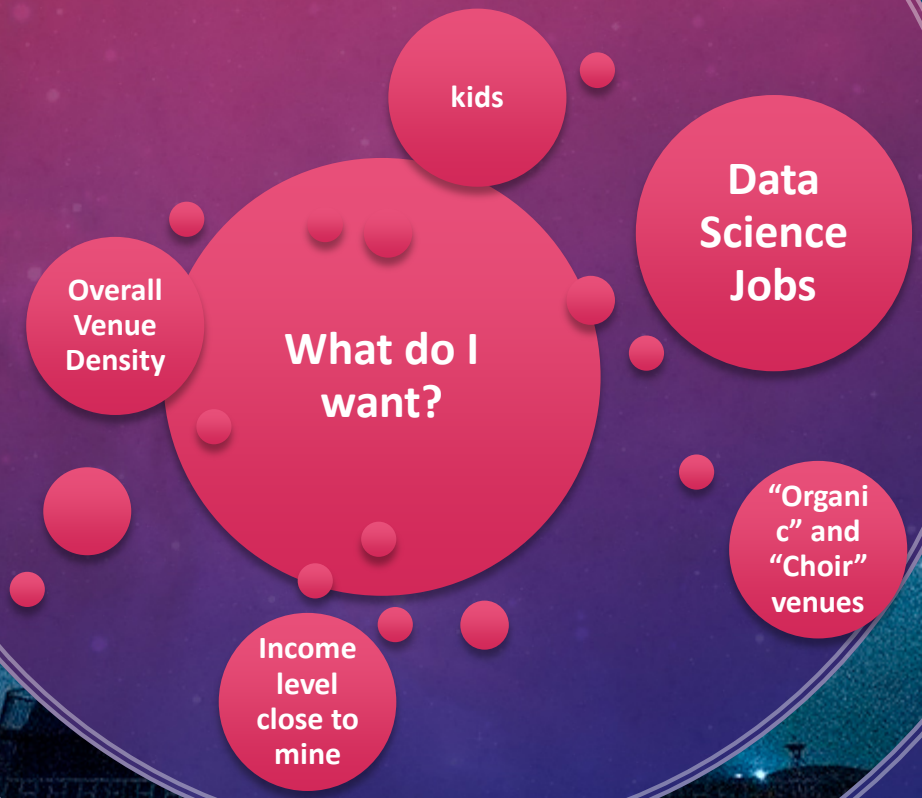
- identify neighborhoods in Texas – as defined zip codes – that offer different amenities or cater to different demographics which will help decision making for the individual.

- Why Texas?
  - Diverse population and urbanization
  - Large land area
  - Potentially business-friendly

DATA SOURCES

- Tax Returns (US IRS)
- FourSquare
- Map Shapes (US Census)
- Zipcodes library

SPECIFY DESIRABLE FEATURES



IDENTIFY KINDS OF NEIGHBORHOODS

Clustering Results

- Cluster 0: lower income, lower home in the household, high density, high venue density, including data, organic, and choir: up and coming urban
- Cluster 1: lower income, low density. - I call this suburban with kids
- Cluster 2: higher income, lower kids, or home with high density - Yuppie neighborhood
- Cluster 3: high income, lower venues density - The estates
- Cluster 4: lower income, lower home, lower density or venues, and low data science - rural

Give these a desirability score by manually rank ordering them according to my preferences:

- Cluster 0 [up and coming) - score = 4 (affordable, has desirable venues)
- Cluster 1 (suburban) - score = 2 (lacks desirable venues)
- Cluster 2 (yuppie) - score = 3 (has desirable venues)
- Cluster 3 (estates) = score = 1 (too expensive, even with a data scientist income)
- Cluster 4 (rural) - score = 0 (lacks jobs)

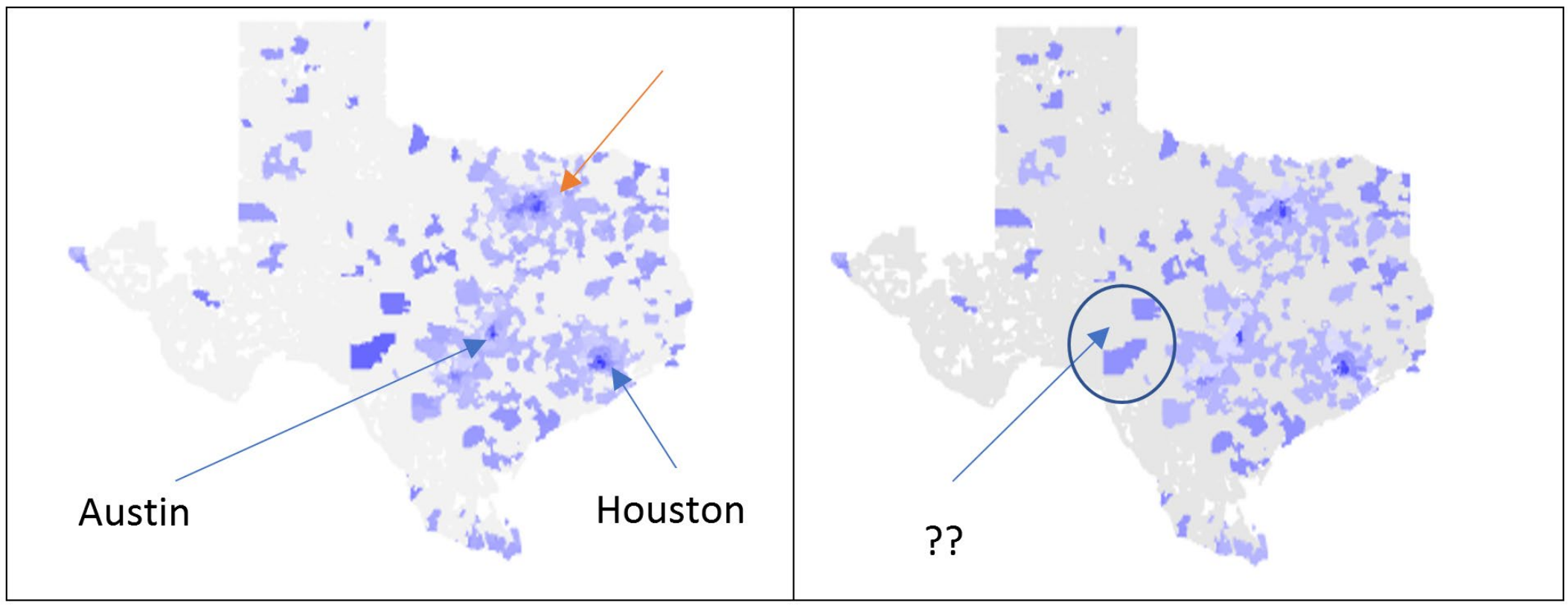
REGRESSION TO FIND HIGHLY DESIRABLE AREAS

city	lr prediction
Austin	5.461262
Houston	4.970501
Austin	4.745056
Houston	4.709116
Houston	4.506680
Houston	4.506342
Houston	4.455803
Austin	4.397563
Houston	4.368335
Austin	4.327493
Austin	4.233221
Houston	4.178711
Houston	4.133718
Houston	4.129572
Austin	4.119329

Linear Regression Predictions	**Random Forest Regression Predictions**
Fit = 75% accurate Identifes Austin and Houston as having the top areas for me	Fit = 90% accurate Identifies Dallas, Houston and Austin as top areas for me

	Zip	city	rf prediction
108	75219	Dallas	4.0
99	75209	Dallas	4.0
441	77057	Houston	4.0
440	77056	Houston	4.0
431	77046	Houston	4.0
427	77042	Houston	4.0
807	78702	Austin	4.0
415	77030	Houston	4.0
412	77027	Houston	4.0
841	78751	Austin	4.0
404	77019	Houston	4.0
481	77098	Houston	4.0
806	78701	Austin	4.0
395	77010	Houston	4.0
96	75206	Dallas	4.0

URBAN AREAS ARE BEST FOR ME… AND WHAT'S THIS?



Darker blue = more desirable

San Angelo and Rock Springs – unclear why they are showing up?

FUTURE WORK AND CONCLUSION

- Based on my rankings, the major urban areas were selected as the most desirable neighborhoods for my tastes and preferences. This is not unexpected, as major urban areas tend to offer both venues as well as jobs.
- West Texas spots could be undiscovered goldmines, or spurious findings
- Interactive interface would be useful to other data scientists
- More precise job, demographic, real estate information could be incorporated
- Enough data of this type from data scientists could inform employers for opening new offices