AN ANALYSIS OF ASIAN AMERICAN QUALITY OF LIFE

DATA SCIENCE MAJOR CAPSTONE, VALERIE TSENG '23

WELLESLEY

MOTIVATION

Asian Americans are the fastest growing racial group in the United States, with a population size nearing 22 million. In order to improve the quality of life for Asian Americans, it is important to understand the group's diversity. Closely analyzing the extent to which variables such as economic status, education levels, and feelings of community affect the quality of life for individuals helps inform decisions to improve conditions for the community as a whole.

DATA SET

The dataset used in this project is titled "Final Report of Asian American Quality of Life" (AAQoL), taken from the City of Austin Open Data Portal. The data itself consists of 2,609 rows and 231 columns. Units answered 231 questions on various topics, from demographic data like household income, to subjective measures like personal satisfaction with healthcare and closeness with family and friends. Furthermore, the survey includes questions about Satisfaction with Life, Quality of Life, Identity with Ethnicity, Familiarity with America, and dozens more.

EXPLORATORY VISUALS

Fig 1: Quality of Life Distribution

Figure 1 shows a relatively normal distribution with a slight left skew. The average is 7.83 and the median is 8.0.

Fig 2: Annual Income Distribution

Figure 2 has a significantly left skewed distribution, with a significant amount of survey takers earning over $70k annually.

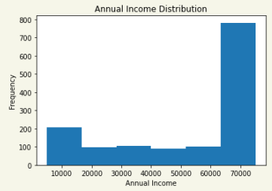
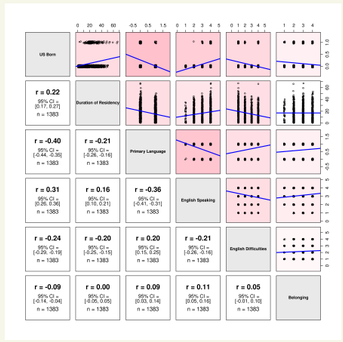


Fig 3: Correlation between Immigration Variables



RESEARCH QUESTIONS

Research Question 1 (RQ1): To what extent can the variables present in this dataset help train a model to accurately predict Quality of Life ratings?

Research Question 2 (RQ2): How closely correlated are variables within each category of variables (demographic, economic, immigration, healthcare, friends and family, and community)??

Research Question 3 (RQ3): Which of the following categories of variables correlate most strongly with quality of life rating among Asian Americans?

LINEAR REGRESSION MODEL

After data cleaning, I trained a linear regression model using the remaining 46 variables to identify which variables contributed most strongly to predicting the Quality of Life indicator, as outlined in **RQ1**. The highest coefficients are shown below. The linear regression model yielded an R-squared of 0.54.

Figure 4: Top 5 Coefficients

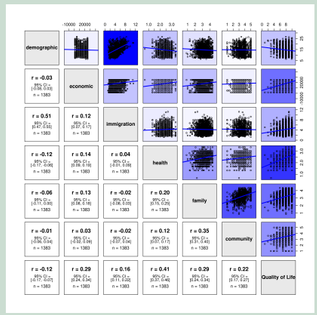
	0	Coefs
0	Satisfaction	0.245174
1	Present Mental Health	0.247584
2	US Born	-0.249792
3	Satisfied With Life 1	0.250847
4	Achieving Ends Meet	-0.318410

ADDRESSING MULTICOLLINEARITY

As shown in Figure 3, there is significant correlation between variables in the same category, prompting **RQ2**. After creating similar multiple correlation plots in all six categories (demographic, economic, immigration, health, family and friends, community), it became clear that each category contained variables that were significantly correlated. To address this multicollinearity, I linearly combined the variables in each category. After combining correlated variables within the same category, a new correlation plot was created, this time between categories.

CATEGORY CORRELATION

Fig 5: Correlation between Categories



MODEL COMPARISON

For the model comparison process in **RQ3**, I tested the accuracy of various models to determine which model would yield most informative results. The models I incorporated are random forest classifier, k-nearest neighbors classifier, and support vector machines.

The random forest classifier model led to the highest accuracy rates overall, so I chose this model to test which specific category on its own will lead to greatest accuracy.

Of the six different categories, the health - related features yielded the highest accuracy of 0.347, and the demographic feature led to the lowest accuracy of 0.23.

Figure 6: Category Models Accuracy

Category	Accuracy
Demographic	0.231214
Economic	0.315029
Immigration	0.268786
Health	0.346821
Family	0.289017
Community	0.274566

DATA ETHICS AND LIMITATIONS

- An important consideration of the data ethics of this project is the manner in which the data was collected. Surveys were volunteer-based, with each respondent receiving $10 for their time. Because many of these respondents are non-native English speakers, there could be considerations about collecting their personal data and whether it was clearly communicated what would be done with collected data.
- Limitations of the survey include: Overrepresentation of units who make over $70k, voluntary response sampling instead of random sampling, and limited geographic representation as the survey was conducted by the city of Austin for Austin residents. Overall, these findings will not be generalizable across all Asian Americans.

CONCLUSION

- Many disparities exist within the Asian American community in Austin, including economic status, with some earning as low as <$10k annual salary and a majority earning $70k+.
- Linear regression model has Achieving Ends Meet as most significant coefficient and an overall R-squared score of 0.54. This suggests substantial predictability of the Quality of Life variable.
- Presence of notable multicollinearity between variables in the same category, especially in immigration category motivated linearly combining variables.
- A random forest classifier model trained on health-related variables leads to the most accurate classification predictors for Quality of Life indicator.
- Overall, this project identified the most important variables and most accurate models for predicting Quality of Life in Asian Americans living in Austin, TX