Happiness

Julian, Anjali, Taneisha

Introduction

Happiness is a vital aspect of life, but have we ever thought about what it is that truly makes us happy?

A study done in Somerville, Massachusetts analyzed various factors that were thought to contribute to people's overall happiness.

6 How would you rate the following?

	VERY BAD				VERY GOOD
The availability of information about city services					
The cost of housing					
The overall quality of public schools					
Your trust in the local police					
The maintenance of streets and sidewalks					
The availability of social community events*					

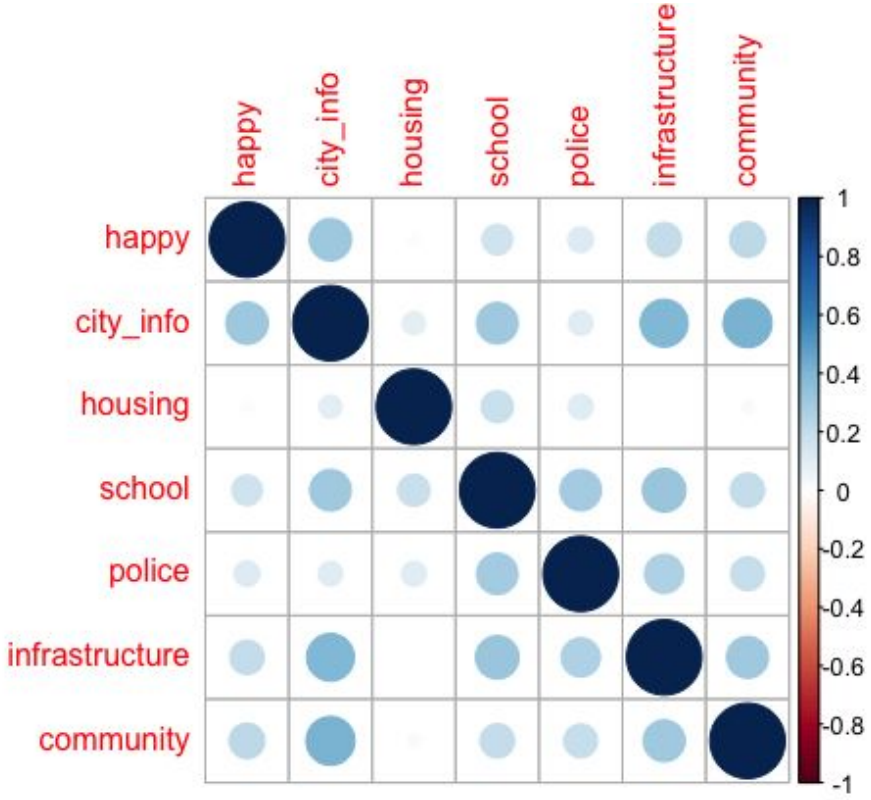
*such as festivals, picnics, parades, and street fairs (e.g., SomerStreets)

Dataset

We obtained our dataset from the UCI Machine Learning Repository. Our dataset contains information on 143 individuals in Somerville and has the following variables:

- Predictors:
  - **city_info**: the availability of information about the city services
  - **housing**: the cost of housing
  - **school**: the overall quality of public schools
  - **police**: trust in the local police
  - **infrastructure**: the maintenance of streets and sidewalks
  - **community**: the availability of social community events
  - All of the predictors have values 1-5
- Response:
  - **happy**: decision attribute with values 0 (unhappy) and 1 (happy)

Correlation Plot

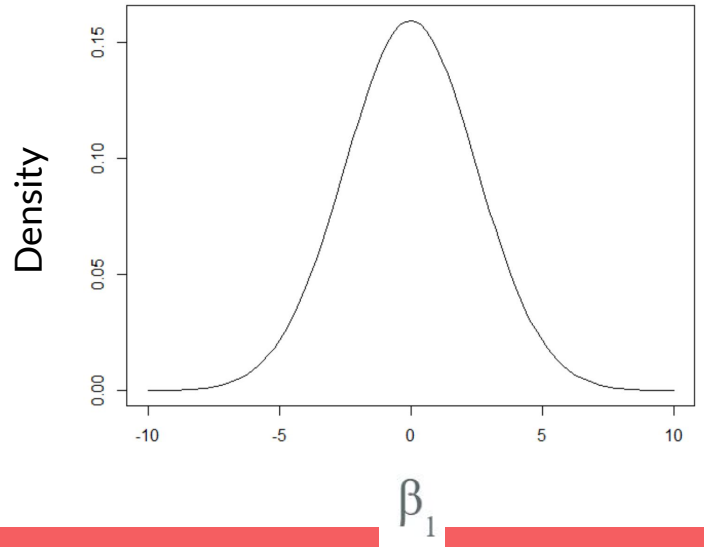


Priors

We decided to stick to default priors for our coefficient parameters. These “defaults” were taken from the rstanarm documentation.

All our beta coefficient parameters were set to have the default normal prior defined below:

$$Normal \sim (0, 2.5)$$



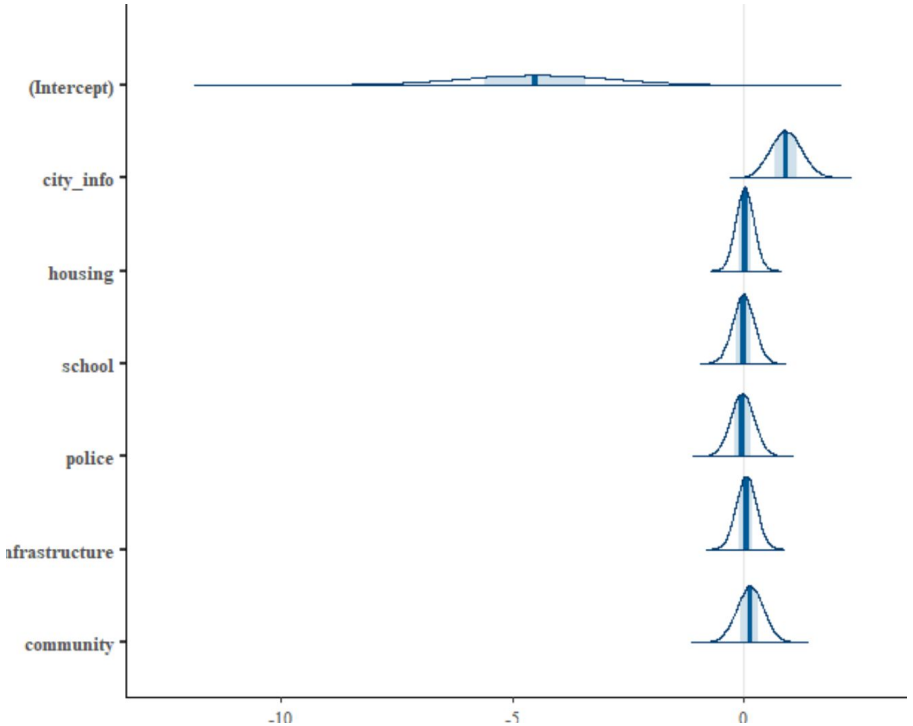
Full Model

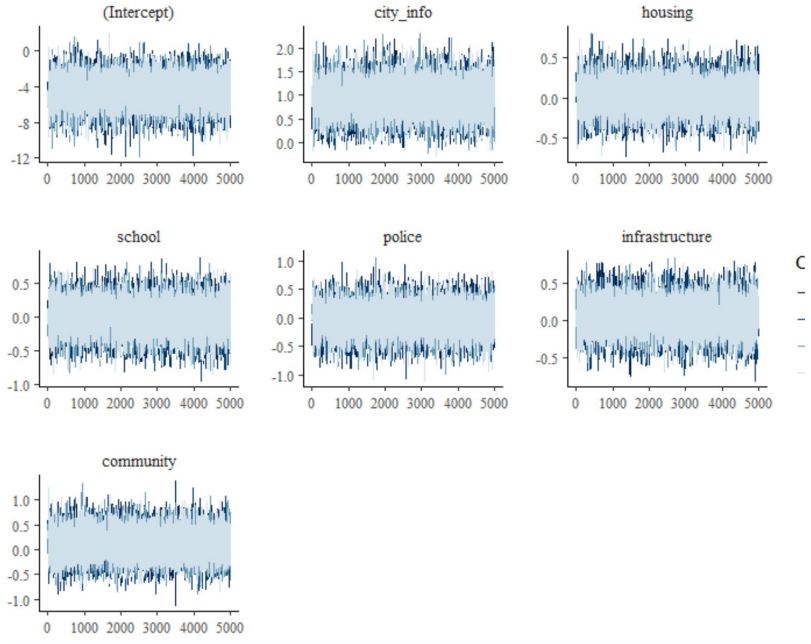
$$Y \sim \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \beta_5 * X_5 + \beta_6 * X_6$$

where, Y is logit(π) and X1, X2, X3, X4, X5, X6 are happiness scores on a scale from 1-5, attributed to: City Info, Housing, School, Police, Infrastructure, and Community, respectively.

The full model contains all the fields in our dataset as features.

Full Model Diagnostics





Reduced Model

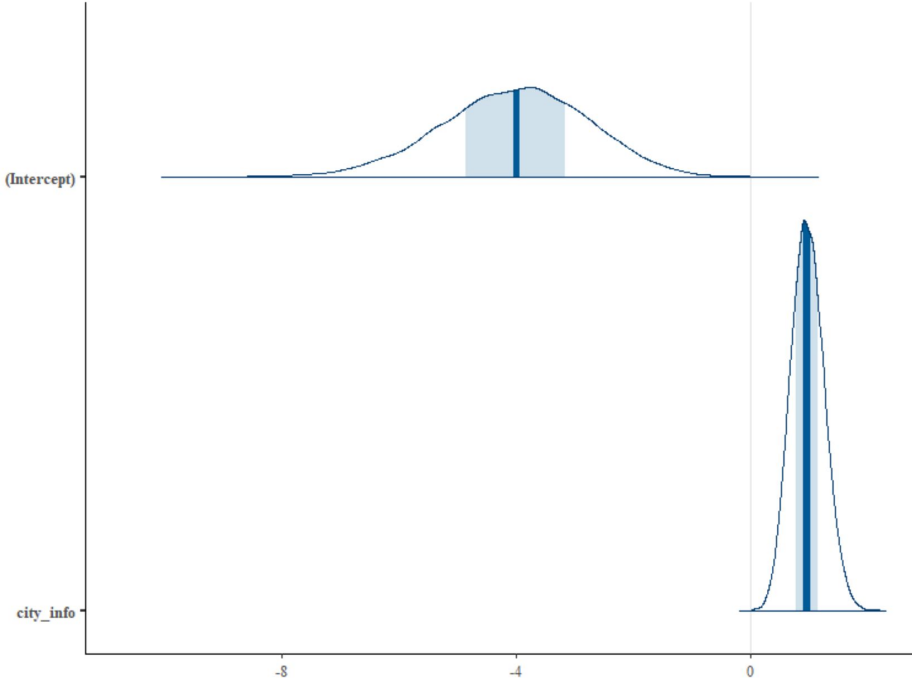
$$Y \sim \beta_0 + \beta_1 * X_1$$

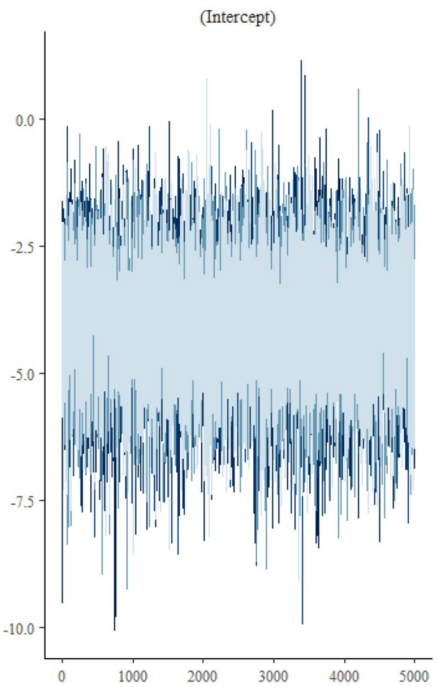
where, Y is logit(π) and X1 is the happiness score on a scale from 1-5, attributed to City Info.

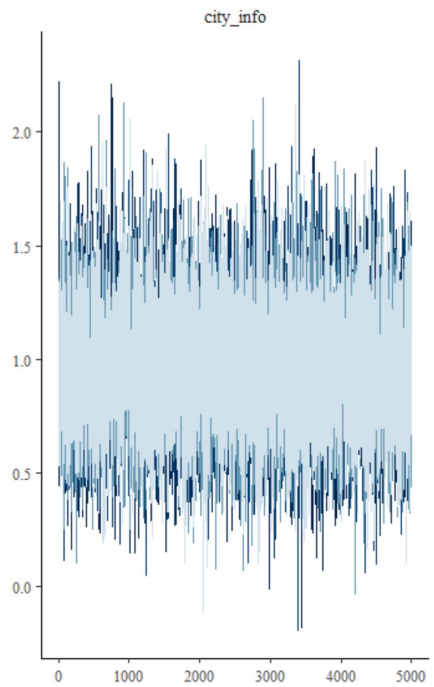
The reduced model contains a subset of all the fields that are a part of the full model.

β_1

Reduced Model Diagnostics







Model Evaluation

Error rate = proportion of misclassified observations

5-fold cross validation

- Train and evaluate the models on five, 80-20 splits
- Compute the error rate for each iteration
- Take the mean of the error rate computed for each of the five train-test splits

Results

Full Model:

5-fold Error rate: 0.420197 (rstanarm)

5-fold Error rate: 0.4285714 (custom function + rstan model)

Reduced Model:

5-fold Error rate: 0.3492611 (rstanarm)

5-fold Error rate: 0.3428571 (custom function + rstan model)

Conclusion

The reduced model had a lower error rate for both rstan and rstanarm. This suggests that the reduced model performs better than the full model.

While more features generally leads to a lower error rate, it could also lead to overfitting. So, to prevent this, we trained and tested our models on separate data using K-Fold cross validation (K = 5).

Using a train-test split to evaluate the two models allowed us to notice that the model including all 6 features could potentially be overfitting.

Finally, we can conclude that the more informed a citizen of Somerville is about their city services, the more likely they are to be happy.

Source Code

GitHub repo with documented source code in Rmd: Source Code