Somerville Life Satisfaction Project-Summary

Citlally Reynoso

11/20/2020

About Somerville

Somerville is a city located in Massachusetts, United States. It is about three miles northwest of the city of Boston and has a population of 81,360 people.

Data

The residents of Somerville responded to a survey that asked about the level of satisfaction they had towards their life in general. Additionally, a series of questions were asked about personal and environmental aspects of the individuals' lives.
Somerville Happiness Survey responses - 2011, 2013, 2015 link: https://catalog.data.gov/dataset/somerville-happiness-survey-responses-2011-2013-2015

Objective

My goal is to investigate the questions asked to the residents of Somerville in order to identify the aspects of life with the strongest relationship to happiness. Happiness is a state which is rather challenging to quantify and attribute, so I do not intend to find the source of happiness. However, I am excited to learn more about what the happy people of Somerville have in common. I am thrilled to see what interesting connections I will discover in this happiness survey.

Data Cleaning

1. There are observations from the years 2011-2015 in this dataset, but this range in years may result in analytical issues.

- An example may be: issues interpreting the effect of income due to inflation
- remove columns with measurements exclusively for years **after** 2011

2. Answers should range from 1-10 or from 1-5.

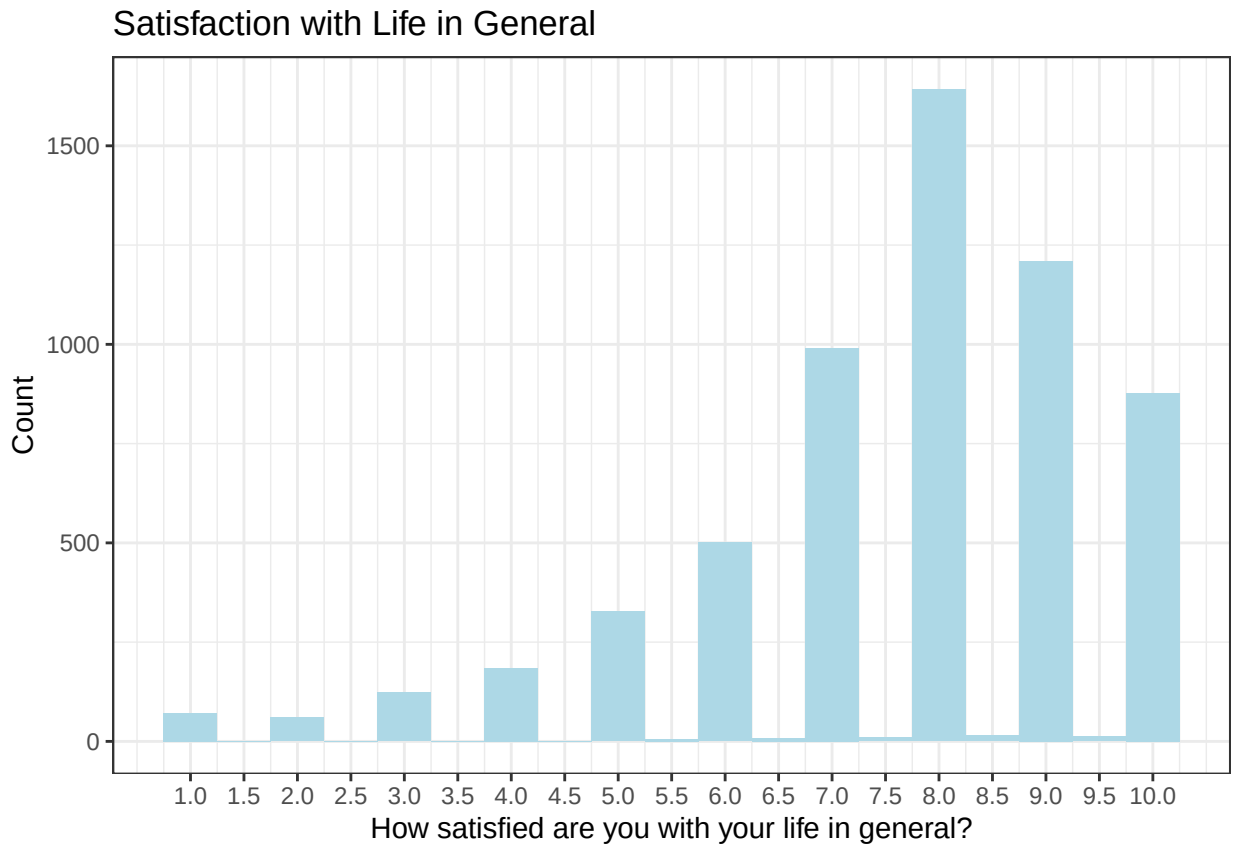
- responses outside of the appropriate range will be identified and replaced with NA.

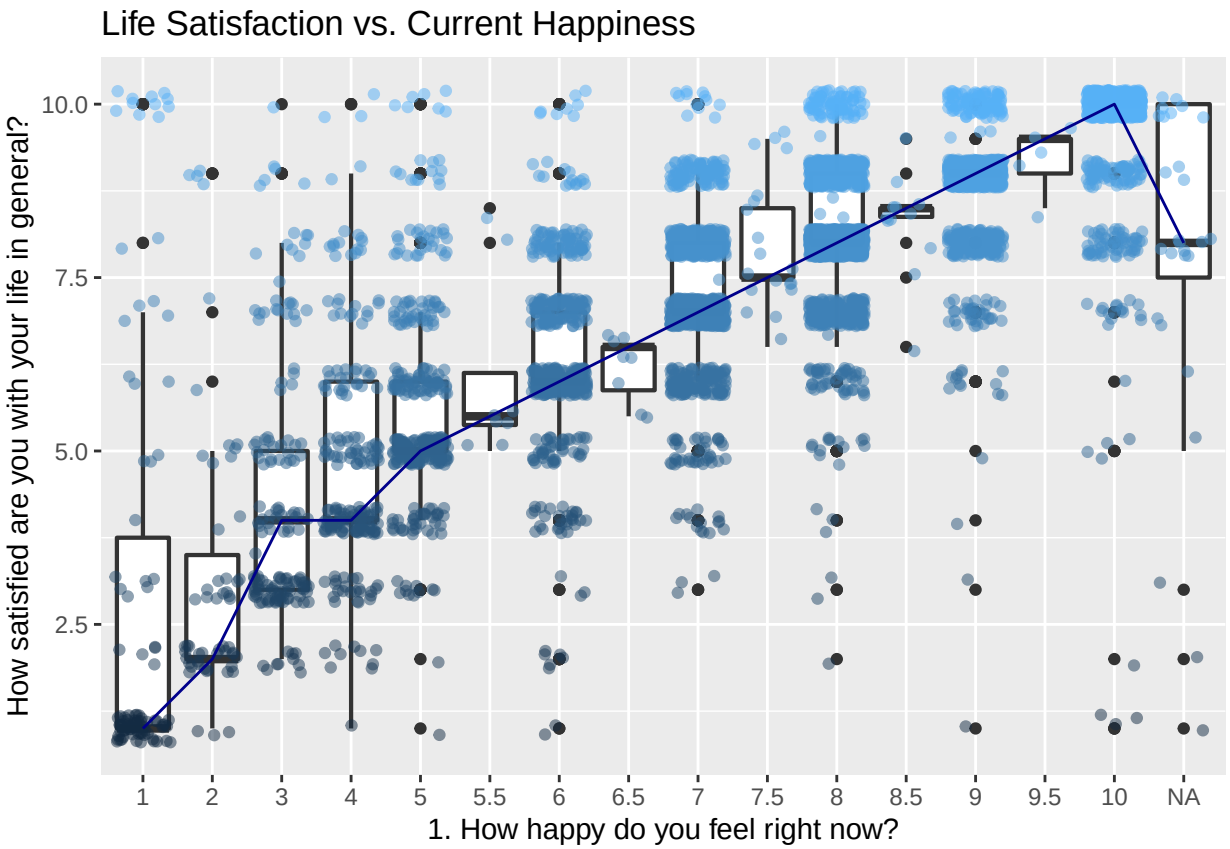
3. Our question will revolve around how happy you are with your life as a whole. Therefore, any rows missing data for the following question will be omitted. : How satisfied are you with your life in general?
4. Change Column names to facilitate the process of subsetting the data set
5. Drop levels not used in factor variables

Table 1: Survey Questions and their New, Corresponding Column Names

Survey_Questions	Column_Name
1. How happy do you feel right now?	happy_now
2. How satisfied are you with your life in general?	satisfied_life
3. How satisfied are you with Somerville as a place to live?	satisfied_somerville
4. In general how similar are you to other people you know?	similar
5. When making decisions are you more likely to seek advice or decide for yourself?	decisions
6. The availability of affordable housing?	affordable_housing
7. How would you rate the following The overall quality of public schools in your community?	public_schools
8. How would you rate the following The beauty or physical setting?	setting_beauty
9. How would you rate the following The effectiveness of the local police?	local_police
10. What is your gender?	gender
11. Age?	age
12. Marital status?	marital_status
13. What is your race?	race
14. How long have you lived here?	years_here
15. What is your annual household income?	annual_income

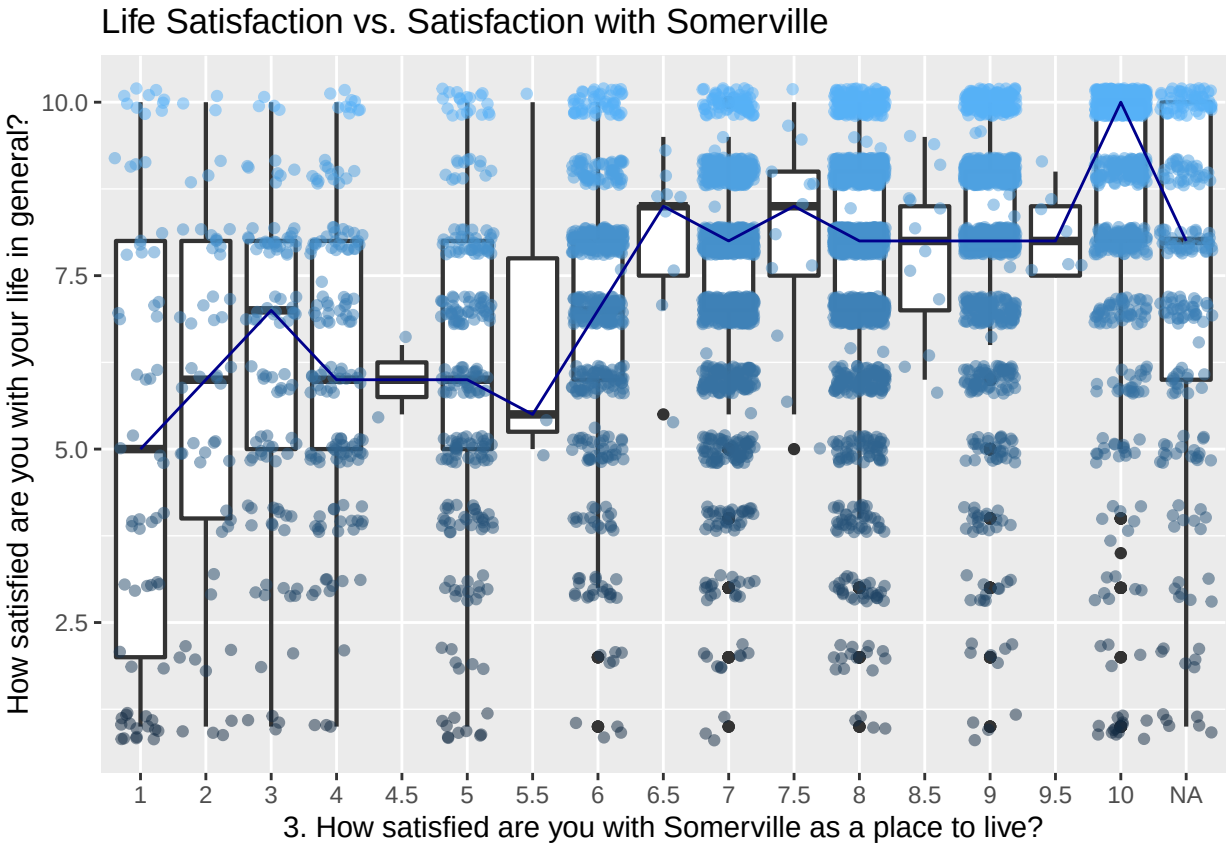
Exploratory Data Analysis & Data Visualization



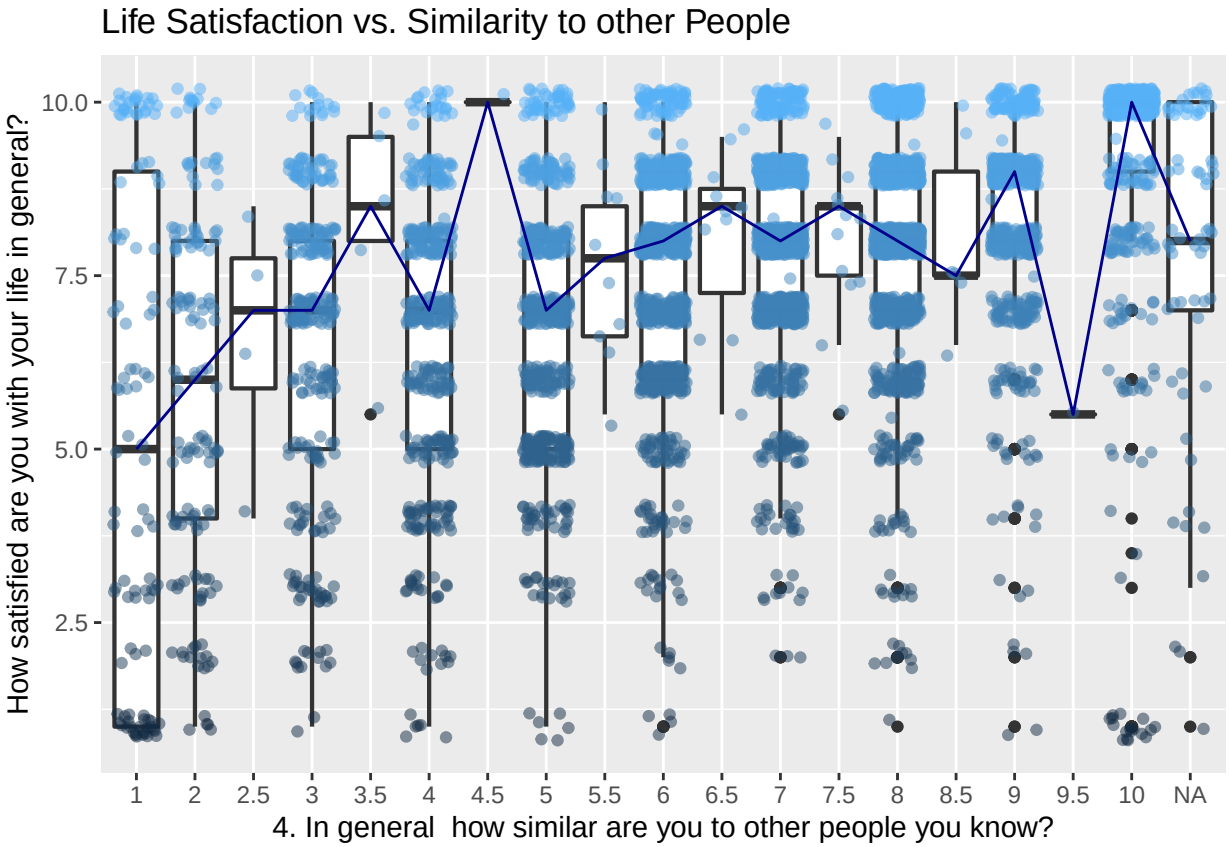


How happy do you feel right now?

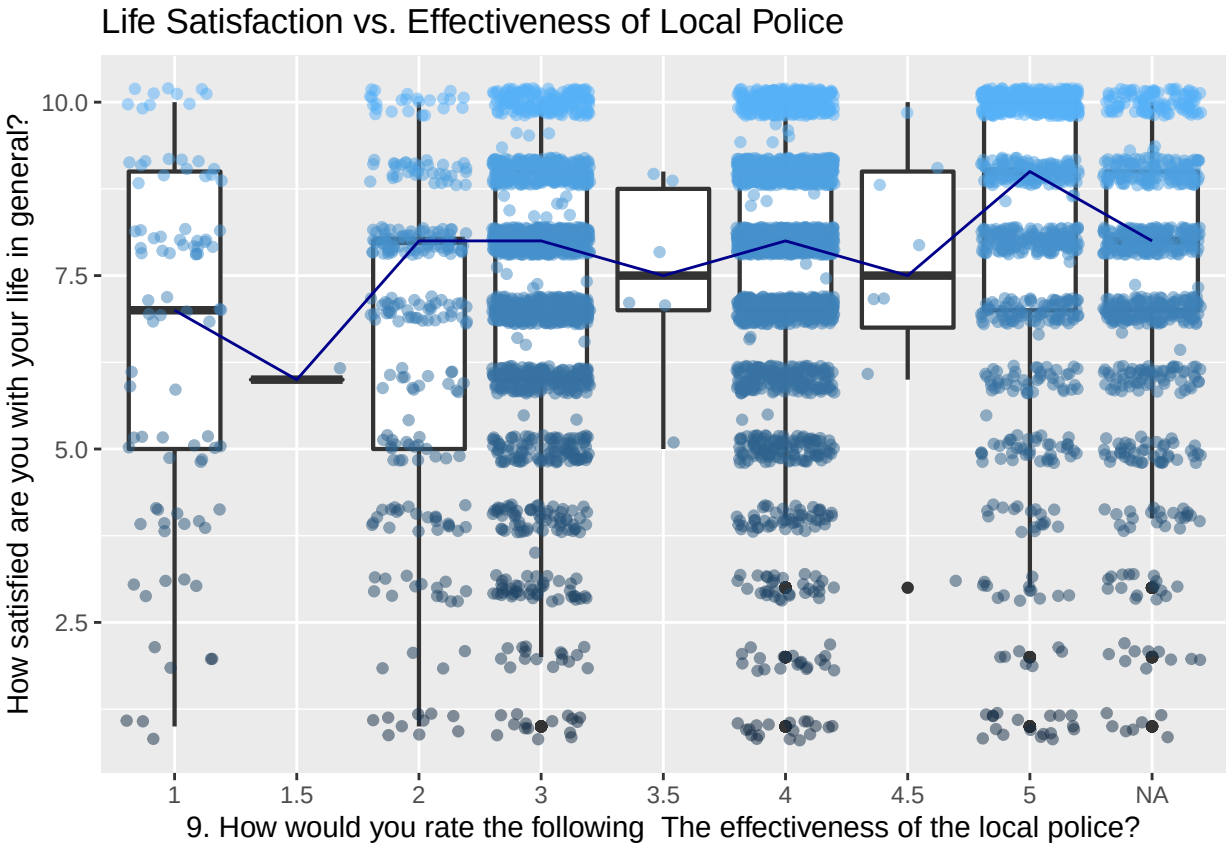
The boxplot graph above clearly showcases a strong relationship between general happiness and current happiness. As the current happiness of the individual rises, their general happiness increases as well. This is the most promising predictor.



How satisfied are you with Somerville as a place to live?
The graph above illustrates a positive relationship between satisfaction with Somerville and life satisfaction. There is evidence to suggest that as your satisfaction with Somerville goes up, your life satisfaction goes up. Even though this graph does not show a perfectly positive relationship, this variable will most likely be a significant predictor.



In general, how similar are you to other people you know?
Showcased above is another positive trend. The more similar you are to the people around you, the happier you tend to be in your overall life. There is a sharp rise at 4.5 that does not seem to follow with the trend. However, this is but a single observation. We see another single observation at 9.5 that really changes the path of the trend, but this is only one observation. When modeling the data, this single point will not hold as much weight as is illustrated above.



How would you rate the following: The effectiveness of the local police?
This variable is interesting because it seems to have a positive relationship with life satisfaction but it is far from strong or clear. It is important to point out that 766 people failed to answer this question, and including this variable in the analysis would cause too many observations to be lost.

- Tip: if you have a variable with a small set of levels and many NAs, analyzing it as a categorical predictor might be a wonderful opportunity. You can use NA as a category and this saves those observations from being lost in the analysis. If the NAs stand out compared to the rest of the boxplots, then looking into a relationship can prove fruitful.
- However, in this case, the NAs do not stand out and seem to spread out through all the ratings of life satisfaction There is no indication that people who fail to answer are more or less happy in life. The categorical transformation of the local_police variable will not help to explain the variation in life satisfaction.

Linear Regression Model

This model is a wonderful place to start because it provides results that are very easy to interpret. The goal of this paper is to explore and find the variables which have the greatest effect on the life satisfaction experienced by the residents of Somerville. Consequently, it is very helpful to create a model with comprehensible results that can appeal to people with various levels of statistical literacy.

Final Linear Regression Model
Output:
Satisfied life: How satisfied you are with your life in general on a scale of 1-10. Where 1 is the least amount of satisfaction and ten is the largest amount of satisfaction.
Predictors:

1. happy now
2. satisfied Somerville
3. similar
4. marital status
5. annual income

Summary of the Model

```
##
## Call:
## lm(formula = satisfied_life ~ happy_now + satisfied_somerville +
##     similar + marital_status + annual_income, data = happy2011)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7160 -0.5003  0.0077  0.5263  7.5249
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        0.953304   0.109250   8.726  < 2e-16 ***
## happy_now                          0.698868   0.008909  78.441  < 2e-16 ***
## satisfied_somerville               0.048572   0.009114   5.330 1.02e-07 ***
## similar                            0.115303   0.007672  15.030  < 2e-16 ***
## marital_statusMarried              0.135285   0.057714   2.344 0.019108 *
## marital_statusR                    0.212946   0.126964   1.677 0.093555 .
## marital_statusSingle, Never Married -0.015588   0.055194  -0.282 0.777627
## marital_statusWidowed              0.221571   0.086398   2.565 0.010356 *
## annual_income100,000 and up        0.372503   0.072901   5.110 3.33e-07 ***
## annual_income20,000 - $29,999      0.176611   0.085256   2.072 0.038353 *
## annual_income30,000 - $39,999      0.101474   0.083563   1.214 0.224668
## annual_income40,000 - $49,999      0.197480   0.082039   2.407 0.016109 *
## annual_income50,000 - $59,999      0.266778   0.081493   3.274 0.001068 **
## annual_income60,000 - $69,999      0.293557   0.083678   3.508 0.000455 ***
## annual_income70,000 - $79,999      0.329463   0.086920   3.790 0.000152 ***
## annual_income80,000 - $89,999      0.385820   0.088985   4.336 1.48e-05 ***
## annual_income90,000 - $99,999      0.227779   0.090988   2.503 0.012328 *
## annual_incomeLess than $10,000     0.164609   0.097924   1.681 0.092819 .
## annual_incomeR                     0.245397   0.092239   2.660 0.007826 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.115 on 5703 degrees of freedom
##   (316 observations deleted due to missingness)
## Multiple R-squared:  0.6446, Adjusted R-squared:  0.6434
## F-statistic: 574.6 on 18 and 5703 DF,  p-value: < 2.2e-16
```

Coefficient and Model Analysis

Using the predictors enumerated below, the final model is able to explain 64.46% of the variation in the life satisfaction variable. Before moving on to cross-validating this model, I will analyze what some of the model coefficients are telling us:

1. **Happy now:** For every one-point increase in the level of current happiness, there is, on average, a 0.698868 increment in general life satsifaction.

- The current level of happiness of the participants is, by far, the strongest predictor of their general life satisfaction

2. **Satisfied Somerville:** A one-point increment in the participants' satisfaction with Somerville resulted in an average increment of 0.048572 in their life satisfaction.

3. **Similar:** For every point increase in similarity, there is a 0.115 increase in general life satisfaction.

- For the residents of Somerville, feeling more similar to the people around correlates to a higher level of life satisfaction.

4. **Marital status:** Below is an ordered list of marital status based on levels of life satisfaction. Single/never married people being the least satisfied and widowed people experiencing the greatest satisfaction:
single/never married < Divorced < Married < Refused Response < Widowed

- This is an interesting categorical variable which allows us to investigate how marital status affects the life satisfaction of the people of Somerville. The least happy people are those that are single and have never been married. These people experience a decrease in life satisfaction of 0.01 point compared to those who are divorced. Next we have people who have been divorced. Those who are married come in at an average increment of 0.13 in life satisfaction compared to the divorced group. Those who are widowed or refused to answer the question are 0.2 points more satisfied with their life than those who have been divorced.

5. *Annual income:* is categorized in 10,000 dollar increments. The lowest category is less than $10,000 and the highest is $100,000 or more.

- Those who earn between $10,000 and $20,000 seem to be the most unhappy members of the Somerville community. They are even more unhappy than the group who earn less than $10,000.
- People who earn between $20,000-$50,000 tend to be 0.1-0.2 points happier than those who earn less that between $10,000 and $20,000 annually.
- People who earn $50,000-$100,000 or more tend to be 0.25-0.4 point happier than those who earn between $10,000 and $20,000 annually

Cross Validation

```
## Linear Regression
##
## 5722 samples
##    5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5150, 5150, 5150, 5149, 5150, 5151, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.117795  0.6425041  0.75637
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

I used 10-Fold resampling cross validation to measure the accuracy with which the final linear model can predict the general happiness of Somerville Residents.

- **The Final Model has a 64% accuracy** in predicting life satisfaction when given: current level of happiness, satisfaction with Somerville, perceived similarity level, marital status, and annual income.

Ordinal Life Satisfaction Variable

I created a new ordinal variable called life_satisfaction that classifies the general life satisfaction of the subjects based on the following scale:

- 1-2.5: extremely unsatisfied
- 3-4.5: unsatisfied
- 5-6.5: somewhat satisfied
- 7-8.5: satisfied
- 9-10: very satisfied

Table of life_satisfaction Category Counts

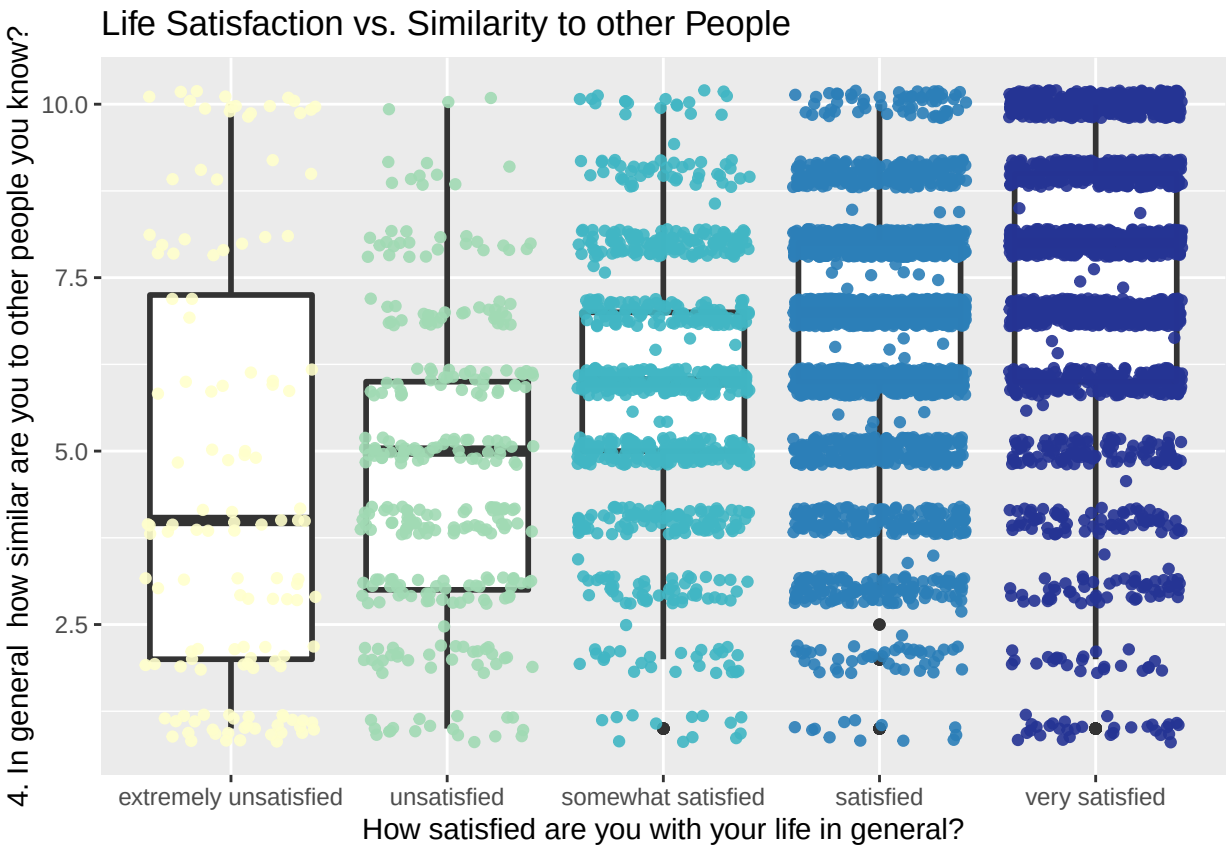
```
##
## extremely unsatisfied           unsatisfied    somewhat satisfied
##                   131                   308                   842
##             satisfied        very satisfied
##                  2657                  2100
```

Exploratory Data Anlysis for Ordered Categorical Variable

For your convenience, only the most important graphs have been included in this pdf

- Current happiness and life satisfaction share a strong positive relationship. Those who are currently happy tend to express higher levels of life satisfaction.

- Somerville residents who are happier with the city seem to experience higher levels of life satisfaction. The variation of how satisfied residents are with Somerville increases as general life satisfaction decreases. On the other hand, residents who are very satisfied with their life seem to generally rate their satisfaction with Somerville highly.



- People who feel that others are similar to them, tend to experience higher levels of life satisfaction. The group of people who are extremely unsatisfied with their life tend to have a large variation in how similar they feel to others, and they center around a level 4.4 out of 10. The higher the life satisfaction, the less variation we see in levels of perceived similarity. Those who are very satisfied with life average at a perceived similarity level of 7.5 out of 10.

Ordinal Logistic Regression Model

Final Linear Regression Model

Output:

Life Satisfaction: How satisfied you are with your life in general: extremely unsatisfied, unsatisfied, somewhat satisfied, satisfied, or very satisfied

Predictors:

1. happy now
2. satisfied Somerville
3. similar

Table of Variable Significance

```
##                                        Value       p value
## happy_now                          1.20186252  0.000000e+00
## satisfied_somerville               0.08590793  7.145497e-07
## similar                            0.17236175  8.240952e-31
## extremely unsatisfied|unsatisfied  3.55070318  4.433030e-85
## unsatisfied|somewhat satisfied     6.04524103 1.703551e-255
## somewhat satisfied|satisfied       8.61011905  0.000000e+00
```

```
## satisfied|very satisfied          12.14821071  0.000000e+00
```

The variables fitted in the final ordinal logistic model are all listed in the table above.

- All the chosen predictors have a p-value smaller than the established 0.05 threshold, hence all the variables are significant predictors that help us explain the variation in the life satisfaction of Somerville residents.
- Additionally, each category of happiness is significantly different from the others. The data shows evidence to suggest that there is a difference in the lives of those who belong to different categories of life satisfaction.
- The log-odds coefficients for the model have also been provided. Because this is a logistic model, an exponential transformation is necessary to interpret the coefficients in the point system they were originally in.

Coefficients and their Confidence Intervals

```
## Waiting for profiling to be done...
```

```
##                                  2.5 %   97.5 %
## happy_now            3.326306 3.171135 3.492112
## satisfied_somerville 1.089706 1.053305 1.127351
## similar              1.188108 1.153866 1.223445
```

1. **Happy Now**: For every one-point increase in the level of current happiness, there is an average increase of 3.326 in general life satisfaction.

- Current happiness has the strongest, positive relationship with life satisfaction. Those who are currently happy highly rate their life satisfaction, and those who are sad tend to poorly rate their life satisfaction.

2. **Satisfied Somerville**: When there is a one-point increase in the level of satisfaction with Somerville, general life satisfaction increases by 1.089 points.

- Individuals who are more satisfied with Somerville tend to be more satisfied with life.

3. **Similar**: People who view those around them as more similar to them, generally live a more satisfied life. For a one-point increase in perceived similarity, there is a 1.188 increase in life satisfaction.

- For the residents of Somerville, feeling more similar to the people around them correlates to experiencing a more satisfying life.

Cross Validation

```
## Ordered Logistic or Probit Regression
##
## 5722 samples
##    3 predictor
##    5 classes: 'extremely unsatisfied', 'unsatisfied', 'somewhat satisfied', 'satisfied', 'very satis
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5150, 5149, 5150, 5151, 5149, 5150, ...
## Resampling results across tuning parameters:
##
##   method    Accuracy   Kappa
##   cauchit   0.7118217  0.5560474
##   cloglog   0.5781187  0.3279161
##   logistic  0.7013365  0.5309041
##   loglog    0.6732048  0.4781372
##   probit    0.6910321  0.5093566
```

```
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was method = cauchit.
```

Using 10-fold resampling cross validation, **the ordinal logistic regression (olr) model has shown to have a 71% accuracy rate.** This model can accurately predict the life satisfaction category a person falls under 71% of the time. It can accomplish this using only three predictors: current level of happiness, satisfaction with Somerville, and similarity to others.
- This model has outperformed the linear model which included two additional predictors: annual_income and marital status

Insights and Conclusions

The conclusions reached in this paper extend only to the residents of Somerville. Any recommendations made are based on reasoning but are not backed by statistical evidence given that the residents of Somerville are not representative of the general population. Please take these recommendations at your own risk and with a grain of salt.

1. Current happiness affects how satisfied Somerville residents are with their life.

- This comes as no surprise, considering that general life satisfaction is but an extension of momentary happiness. However, it is interesting to question to what extent our current state of happiness may affect our perception of life satisfaction.

2. The people of Somerville are more satisfied with their life when they are satisfied with the place where they live.

- Although this conclusion cannot be extrapolated to everyone else in the world because Somerville is not a representative sample, there is something intuitive about the conclusion.
- Next time you are looking to improve your life, maybe you can move to a place which you are highly satisfied with!
- If this is not a feasible option, then you can redecorate your house or room. This satisfaction could permeate to your state of happiness.

3. The people of Somerville are happier when they deem that those around them are similar to them.

- Oddly enough, this question does not specify whether people are similar in terms of ethnicity, socioeconomic background, or general interests. Therefore, I presume that the question is open to interpretation. Responses might simply translate to how much people feel like they fit in with the members of their community.
- Maybe they share your cultural background, interests, or simply your sense of humor. If you are looking to improve your life then finding a group of people that you feel similar to is a good start!

Study Limitations and Further Questions

Limitations:

1. The data had some promising variables, such as effectiveness of local police, that I was not able to use due to missingness. These variables were potentially important and could have told us more about what affects the life satisfaction of the residents of Somerville, but they had to be dropped.
2. The residents of Somerville are not representative of the general population of the United States. Somerville, Massachusetts is no representative in terms of racial composition, socioeconomic background, and many other important characteristics. This limits the implications of our findings to only the people of Somerville.

Further Questions

1. How would a sample representative of the population of the United States respond to the questions asked in this study?
2. Is there a way to garner a higher response rate for the questions that seemed potentially important but that resulted in a lot of NAs? Could we redesign the questions or response type to make the questions more welcoming to a response?