Lab 11 - Data, Aesthetics, & Geometries

Your Name Here

November 9, 2017

Complete the following exercises below. Knit together the PDF document and commit both the Lab 11 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Which variables in your dataset are you interested in visualizing? Describe the level of measurement of these variables and what type of geography you think is appropriate to represent these variables. Give your reasoning for choosing the `geom_()` you selected.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: lattice

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'memisc'

## The following objects are masked from 'package:dplyr':
##
##     collect, recode, rename

## The following objects are masked from 'package:stats':
##
##     contr.sum, contr.treatment, contrasts

## The following object is masked from 'package:base':
##
##     as.array
```

I want to look at the overall satisfaction with life data and a number of variables such as death of child, widow sataus or has been subject to physical attack. This would be a combination of both ordinal data and nominal data. Im thinking the best way to represend this data would be to graph the nominal variable on the x axis then make two discrete bar charts for each x variable for each of the ordinal satisfaction levels.

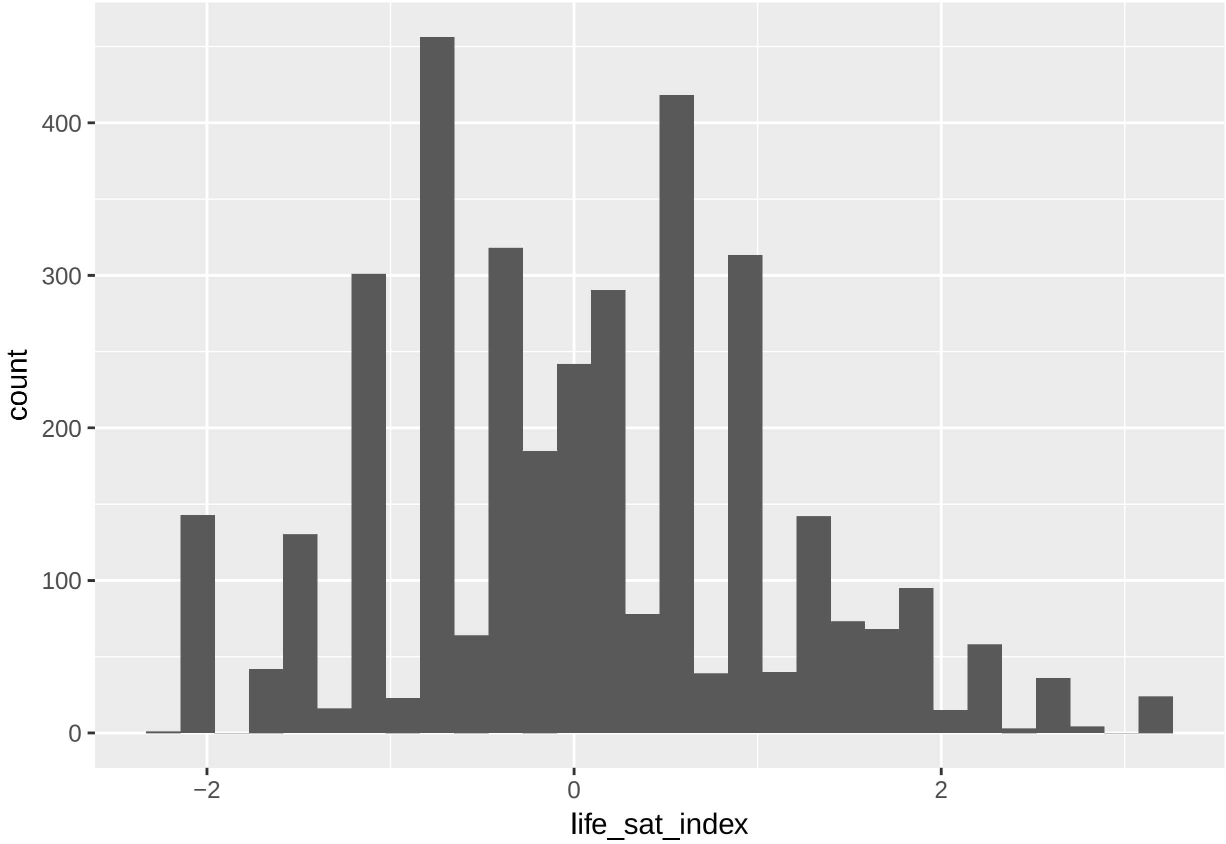
2. Is your data in the proper format to visualize the data in the way you want? Why or why not? *If you need/want to change the structure of your data, do it below.*

I think the data is in the proper format, the observations and varaibles are discrete in their measuremnts.

3. Create at least two different exploratory plots of the variables you chose using the skills we covered in class today. What types of mapping aesthetics did you choose and why? What do these plots tell you about your data?

```r
ggplot(data = changing_lives_subset,
       aes(x = life_sat_index)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

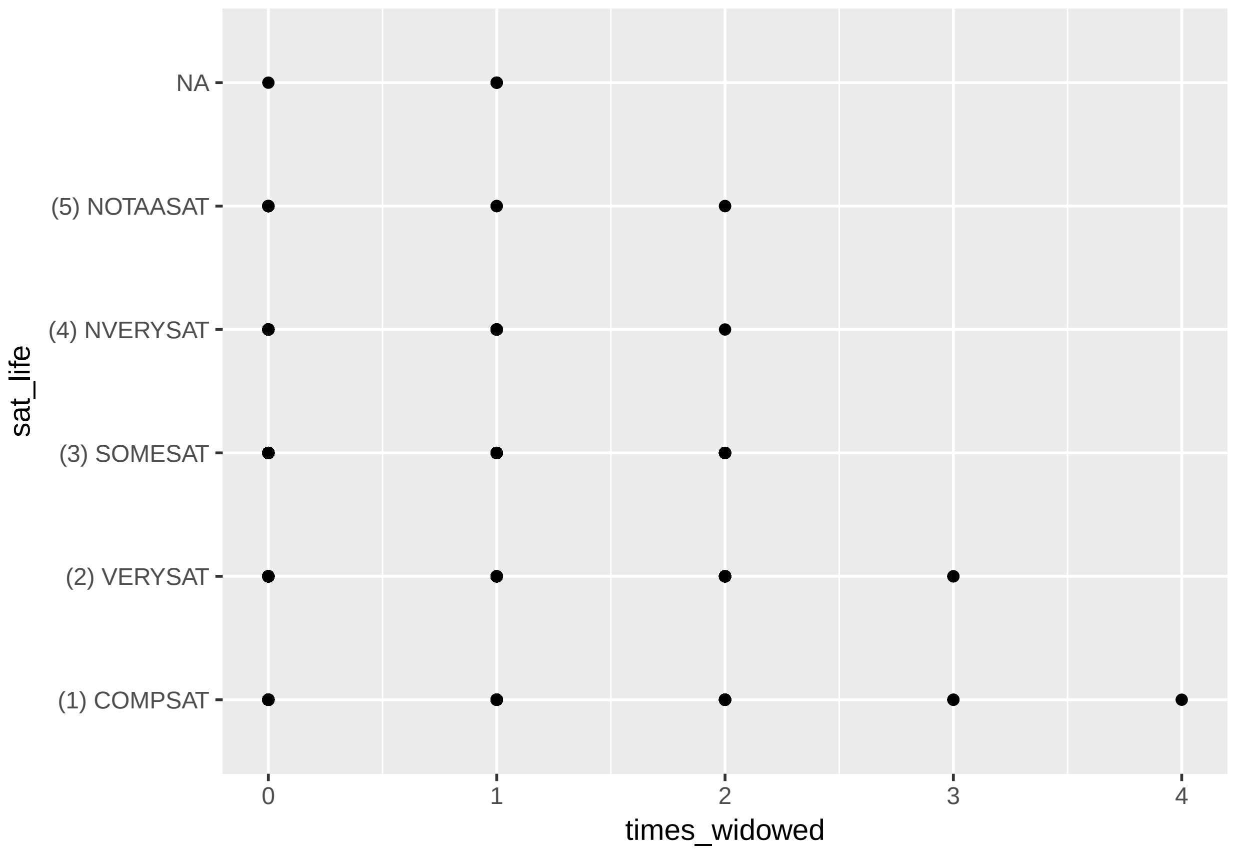


```r
widow_subset <- subset(changing_lives_subset_final, widowed_y_n == "\"YES\"")

#I am trying to subset my data based on a variable, I feel like this should work but its not showing an

ggplot(data = changing_lives_subset,
       aes(x = times_widowed, y = sat_life)) +
         geom_point()
```

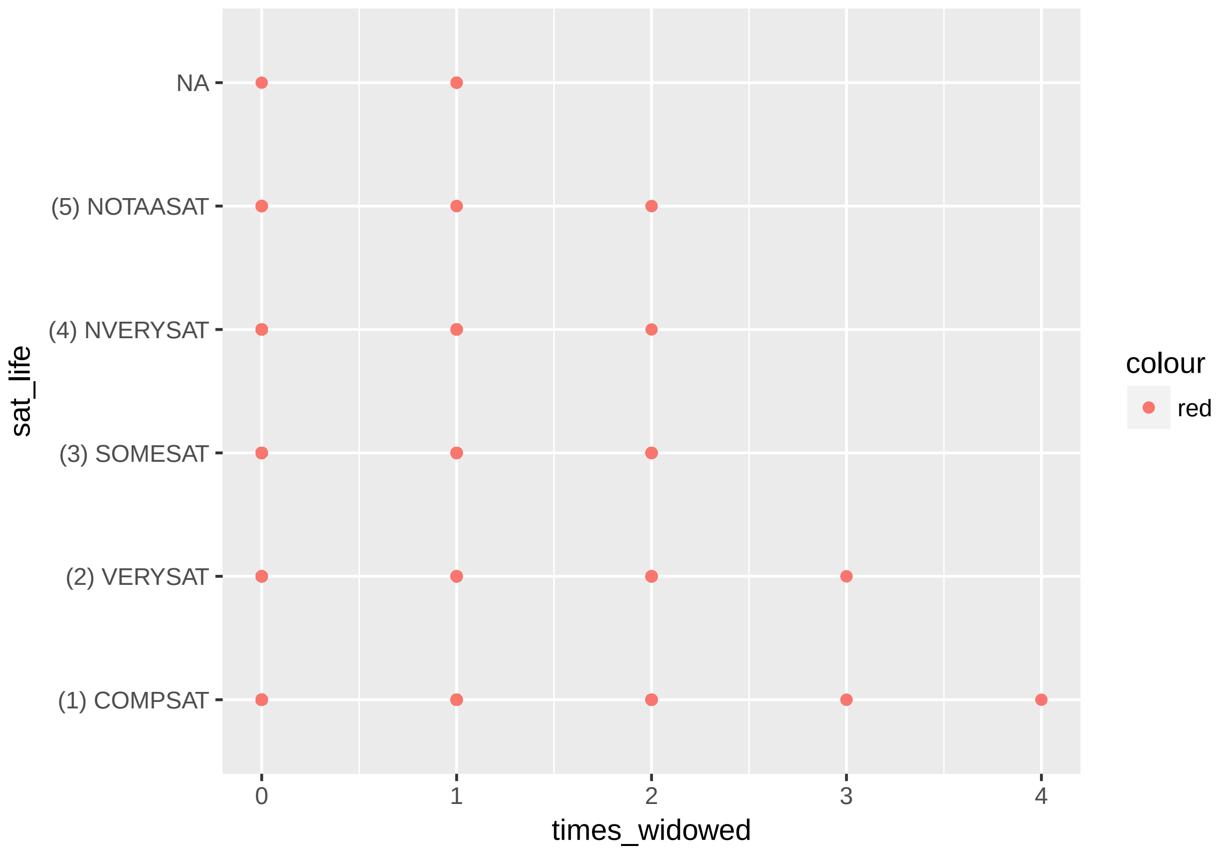
```
## Warning: Removed 2795 rows containing missing values (geom_point).
```



4. Create at least three variations of the plots you've already made by modifying some of the arguments we covered in class (i.e. `position`, `scale`, `size`, `linetype` etc.). Do any of these modifications help you understand your data better? Why or why not? Do any of them create a misleading interpretation of the relationships between your variables? If yes, how so?

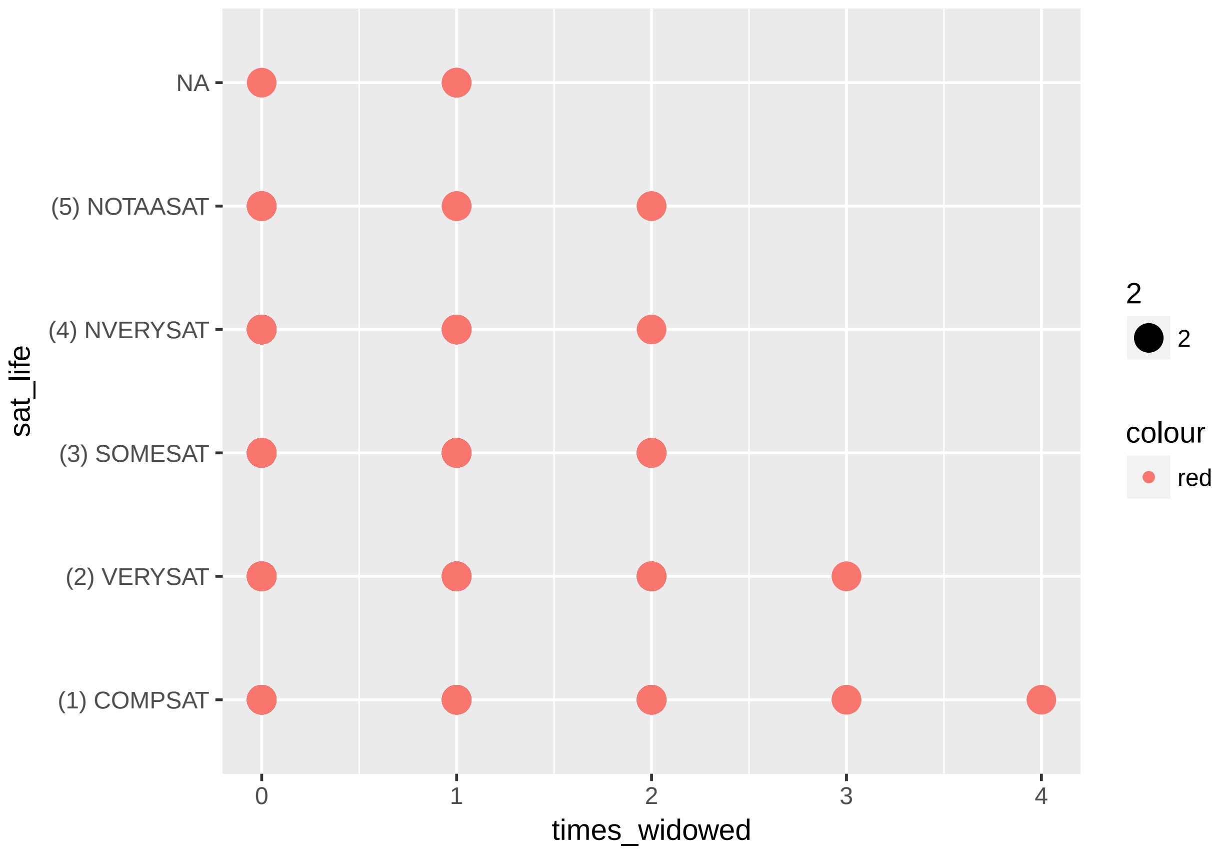
```
ggplot(data = changing_lives_subset,
       aes(x = times_widowed, y = sat_life, col = "red",)) +
         geom_point()
```

```
## Warning: Removed 2795 rows containing missing values (geom_point).
```



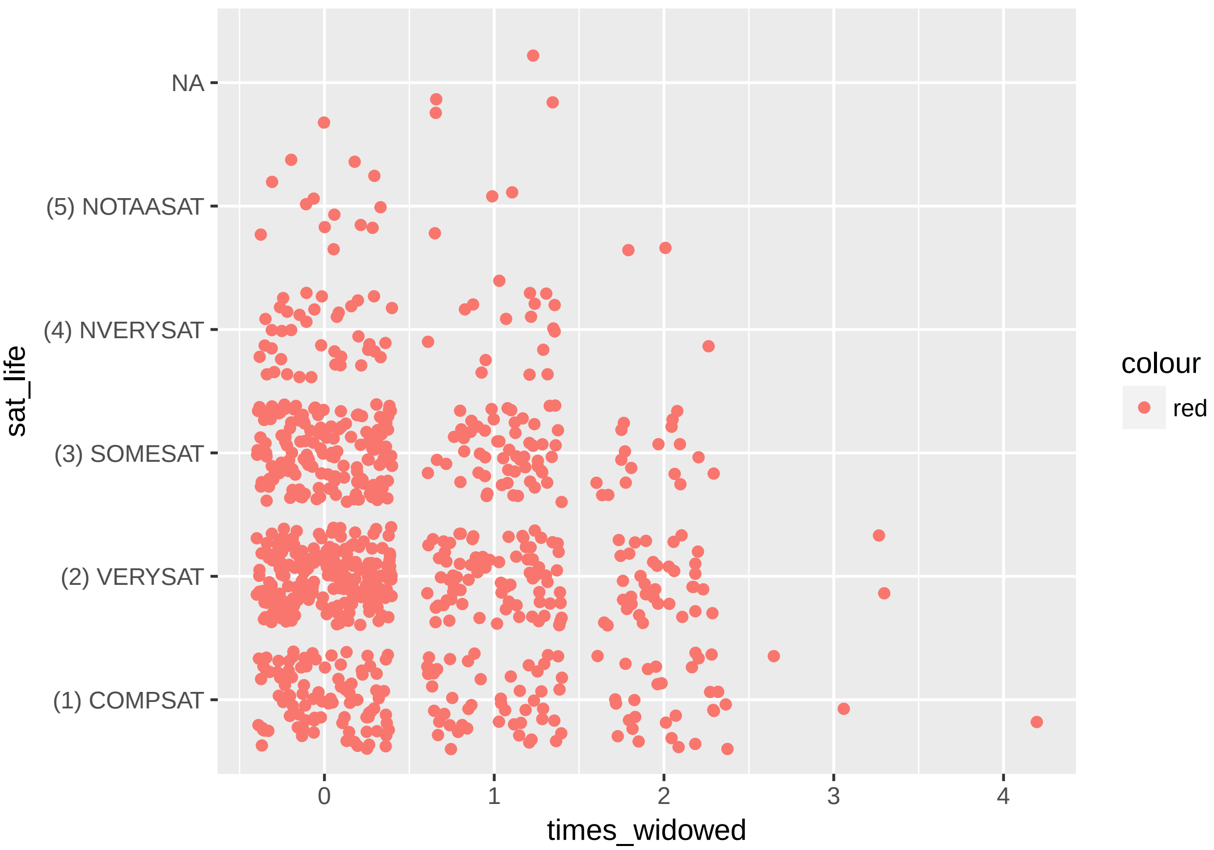
```
ggplot(data = changing_lives_subset,
       aes(x = times_widowed, y = sat_life, col = "red", size = 2)) +
         geom_point()
```

```
## Warning: Removed 2795 rows containing missing values (geom_point).
```

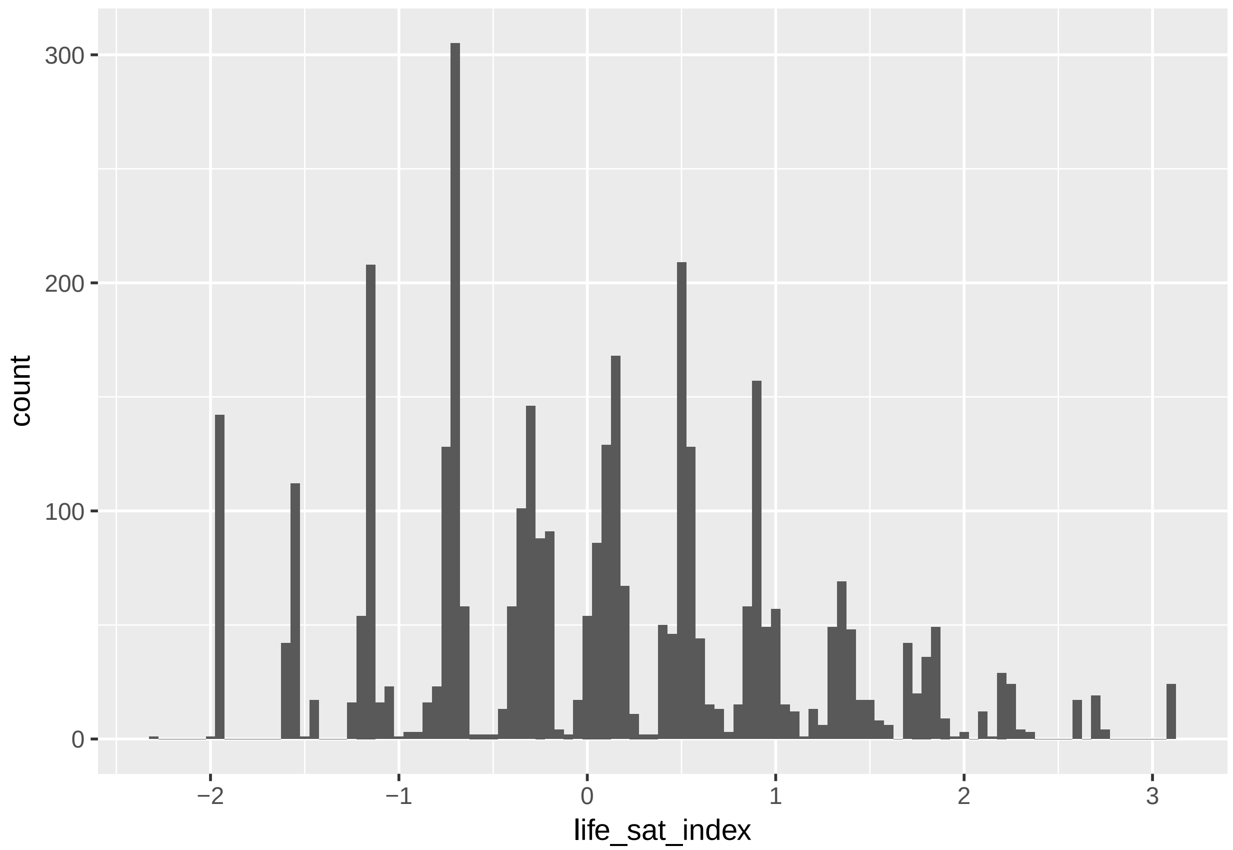


```r
ggplot(data = changing_lives_subset,
       aes(x = times_widowed, y = sat_life, col = "red")) +
         geom_point(position = "jitter")
```

```
## Warning: Removed 2795 rows containing missing values (geom_point).
```

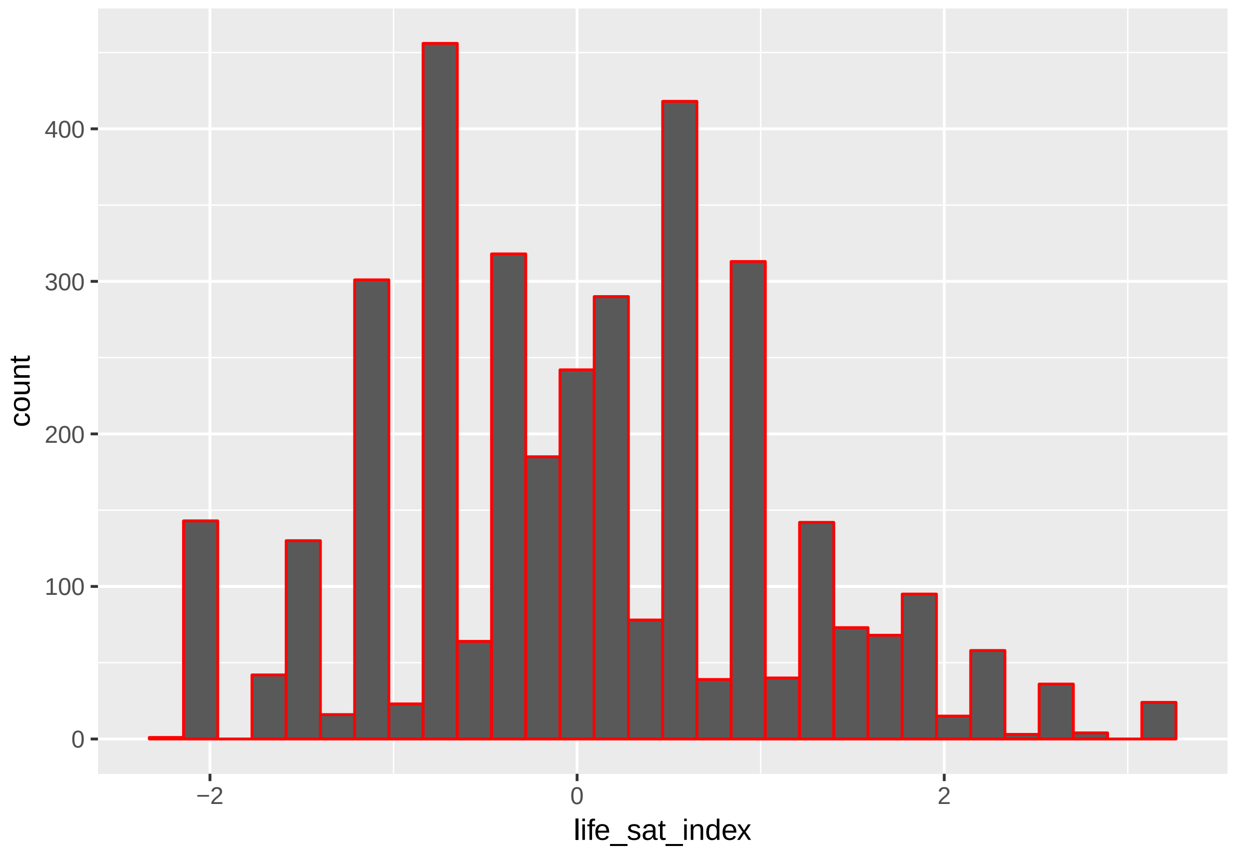


```
ggplot(data = changing_lives_subset,
       aes(x = life_sat_index)) +
  geom_histogram(binwidth = .05)
```

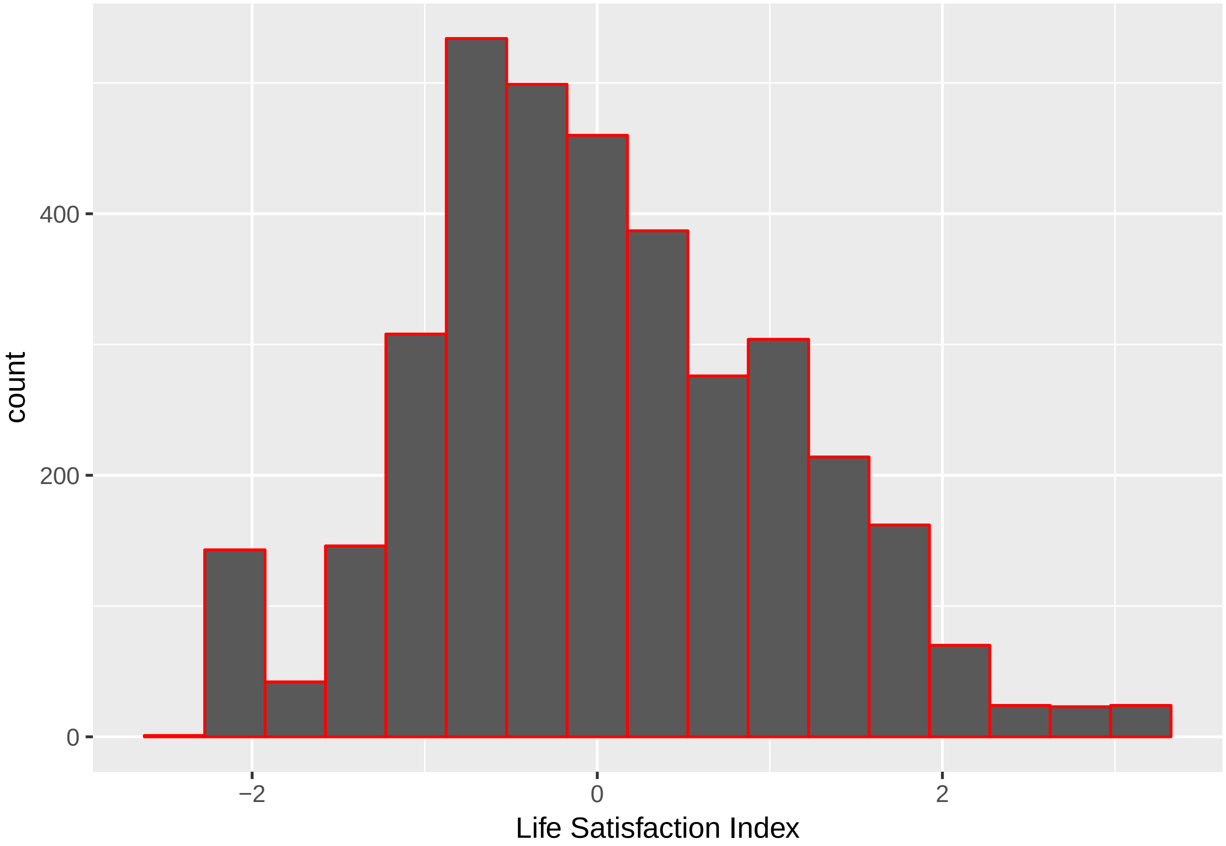


```
ggplot(data = changing_lives_subset,
       aes(x = life_sat_index)) +
  geom_histogram(col = "red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data = changing_lives_subset,
       aes(x = life_sat_index,)) +
  geom_histogram(binwidth = .35, col = "red") +
  labs(x = "Life Satisfaction Index")
```



chaning the bin size of the histrogram is useful to view the distribution (~normal). the jitter in the plot allows the audience to see the number of cases at each level. The jitter is absolutely crutial, considering all the points fall into one of 4 catagories. I do not think it is misleading considering the data wold be undecipherable without it.

5. From the plots you've created thus far, do any of them seem appropriate for a general audience? Why or why not? If so, what do you think you'd still need to do to make them more suitable as explanatory visualizations?

I think I need to do better to communicate what the audience is looking at. This can be accomplished through ascetic modifications and labeling (titles etc). Manipulating the data further to create something that is representative of the data but not manipulative towards the audience is also paramount.