Exploration

Austin Palmer | ajp4344

2023-10-20

```
# Read in Dataset
data <- read.csv("~/Classes/sds322E/Project/Final_Report_of_the_Asian_American_Quality_of_Life__AAQoL_.

# Select Columns of Interest
 col_filtered <- data|>
  select(Quality.of.Life, Belonging, English.Speaking, Income, Age, Household.Size, Familiarity.with.Am

# Create numeric variable representing a 4 option categorical variable. Easier to work compared to stri
col_filtered |>
  mutate(familiarity_num = case_when(
    Familiarity.with.America == "Very low" ~ 0,
    Familiarity.with.America == "Low" ~ 1,
    Familiarity.with.America == "High" ~ 2,
    Familiarity.with.America == "Very high" ~ 3,
    TRUE ~ NA_integer_)) -> col_filtered

# Filter out rows with N/A values in columns of interest
working_dataset <- col_filtered |>
  filter(!is.na(Quality.of.Life) & !is.na(Belonging) & !is.na(English.Speaking) & !is.na(Income)   &!is

# Get init/post-filter dimensions
dim(data)
```

```
## [1] 2609  231
```

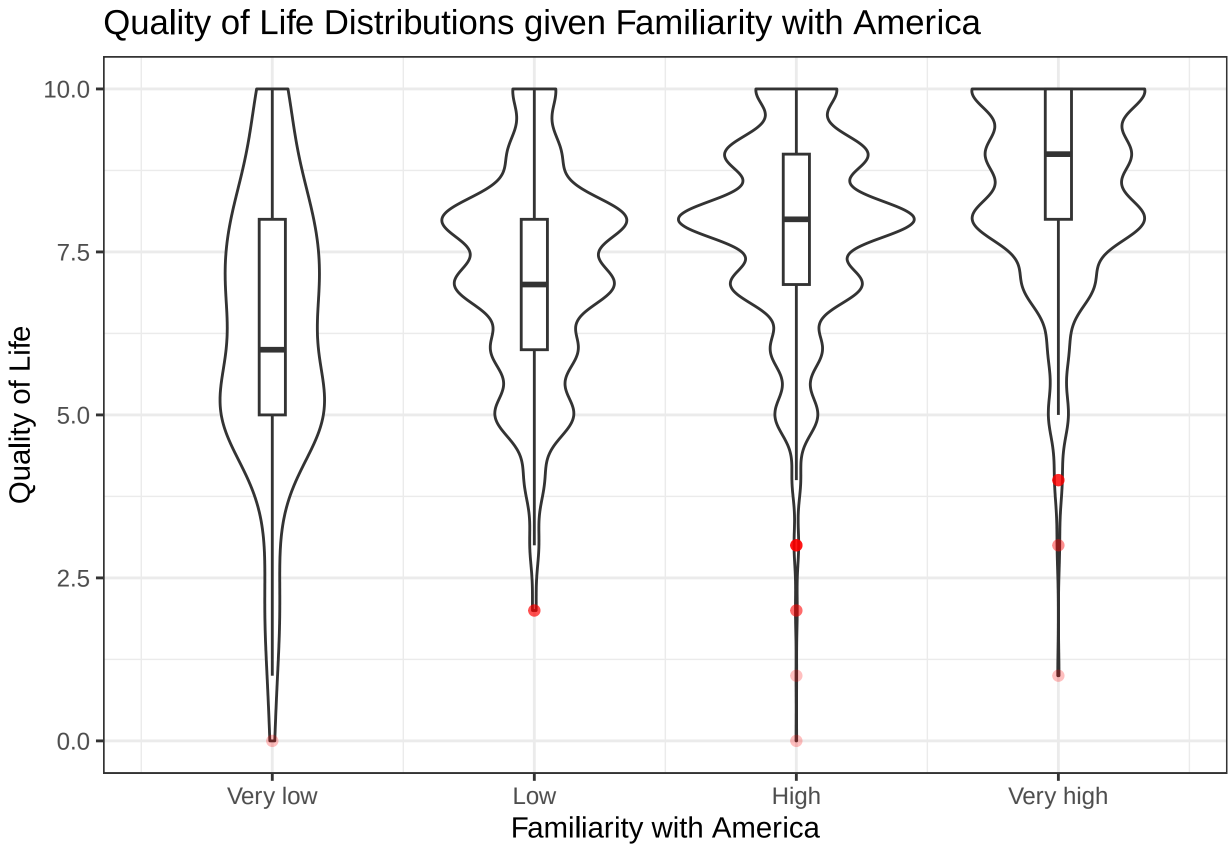
```
dim(working_dataset)
```

```
## [1] 2539    8
```

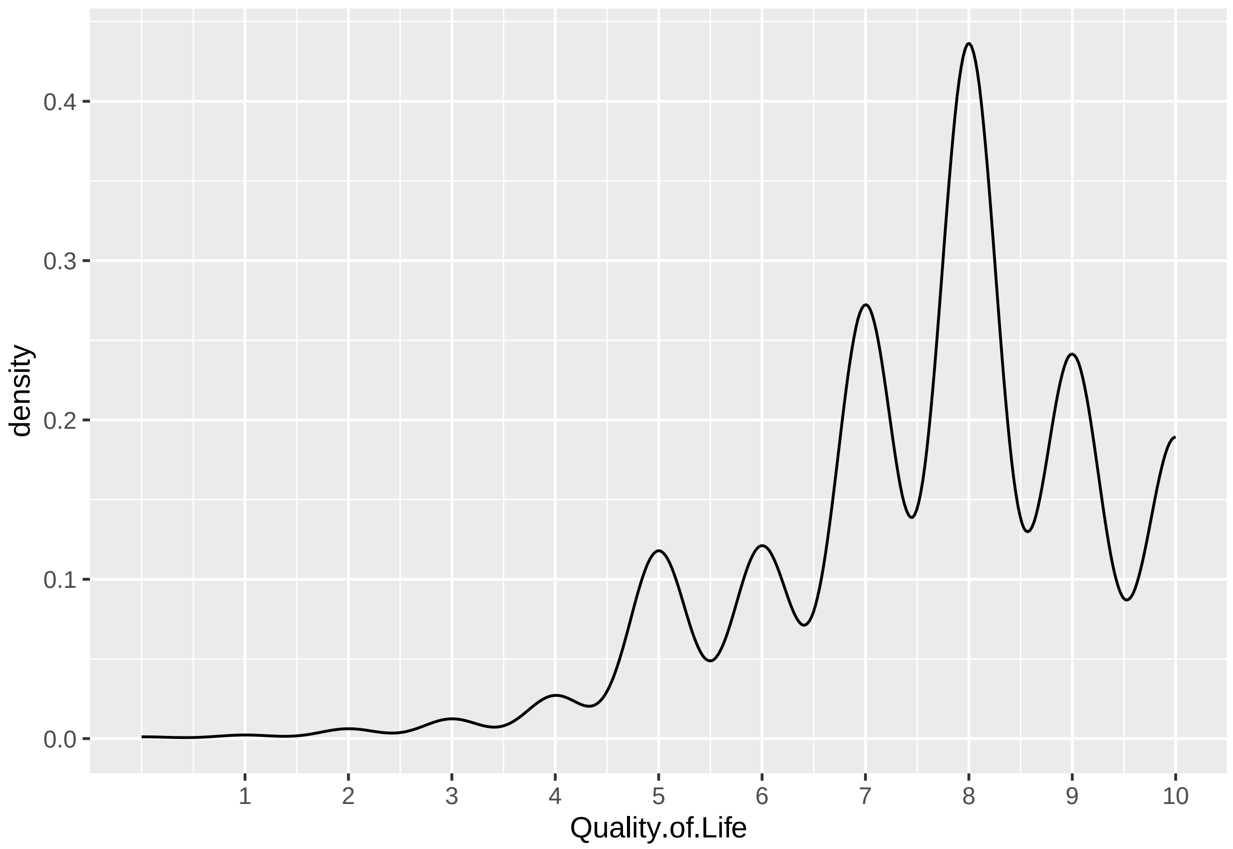
Save familiarity null value rows to predict later

```
familiarity_NA_rows <- col_filtered |>
  filter(is.na(familiarity_num))
```

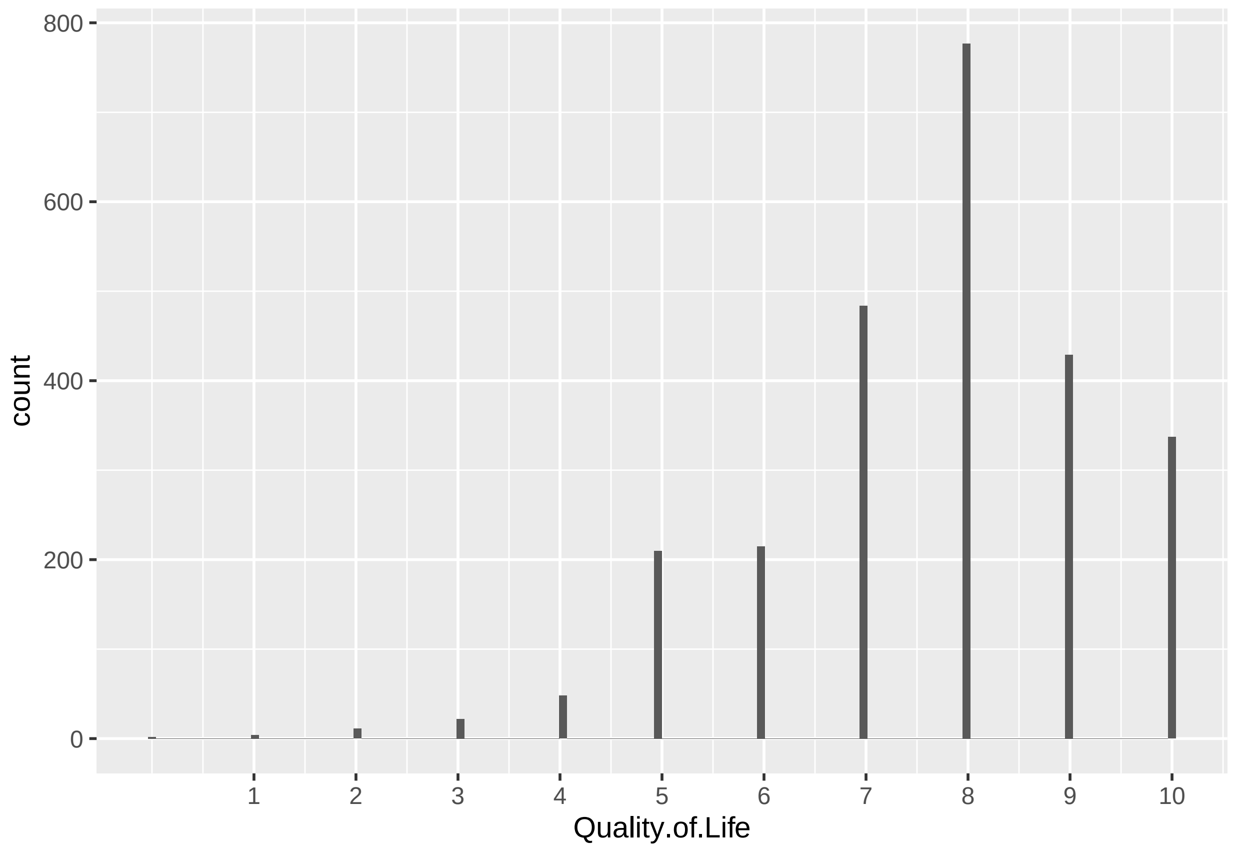
```
desired_order <- c("Very low", "Low", "High", "Very high")
desired_order_num <- c(0, 1, 2, 3)
working_dataset |>
  ggplot(aes(x = familiarity_num, y=Quality.of.Life, group=Familiarity.with.America)) +
    geom_violin() +
    geom_boxplot(width = 0.1, outlier.alpha = .25, outlier.color = "red") +
    scale_y_continuous(labels = scales::comma) +
    labs(
      title = "Quality of Life Distributions given Familiarity with America",
      x = "Familiarity with America",
      y = "Quality of Life"
    ) +
  scale_x_continuous(breaks = desired_order_num, labels = desired_order) +
  theme_bw()
```



```
working_dataset |>
  ggplot() +
  geom_density(aes(x=Quality.of.Life)) +
  scale_x_continuous(breaks=seq(1,10))
```



```
working_dataset |>
  ggplot() +
  geom_histogram(aes(x=Quality.of.Life), bins = 130) +
  scale_x_continuous(breaks=seq(1,10))
```



Linear Model

```
model <- lm(formula=Quality.of.Life~familiarity_num, data=working_dataset)
summary(model)
```

```
##
## Call:
## lm(formula = Quality.of.Life ~ familiarity_num, data = working_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7874 -0.7874  0.2126  1.2126  3.5895
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.41051    0.07678    83.5   <2e-16 ***
## familiarity_num  0.68845    0.03868    17.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.539 on 2537 degrees of freedom
## Multiple R-squared:  0.111,  Adjusted R-squared:  0.1106
## F-statistic: 316.7 on 1 and 2537 DF,  p-value: < 2.2e-16
```

The dataset is tidy because each row is an individual observation (Asian American) and each variable (explanatory and outcome) has its own columns.There is also no rows with any NA values in any of the variables and there is only the six variable of interest present in the new

dataset.

---

3. Results

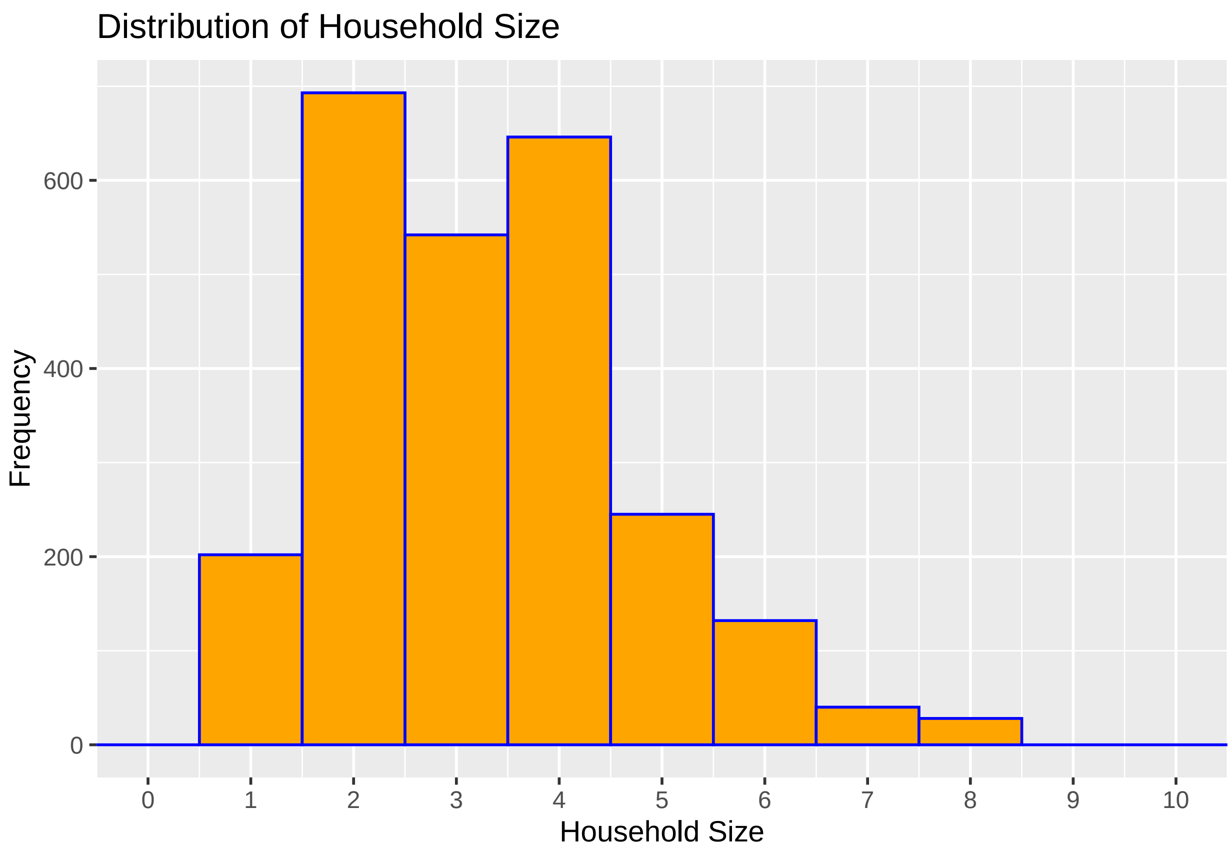
Question 1: How does household size affect the quality of life for Asian Americans?

```r
# Numeric univariate visualization
head(working_dataset)
```

```
##   Quality.of.Life     Belonging English.Speaking            Income Age
## 1               8 Not very much             Well       $0 - $9,999  23
## 2               5     Very much        Very well                    34
## 3               8 Not very much             Well  $70,000 and over  28
## 4              10 Not very much        Very well $50,000 - $59,999  25
## 5               9 Not very much        Very well  $70,000 and over  60
## 6               6 Not very much        Very well  $70,000 and over  43
##   Household.Size Familiarity.with.America familiarity_num
## 1              3                      Low               1
## 2              3                Very high               3
## 3              2                     High               2
## 4              1                Very high               3
## 5              3                      Low               1
## 6              3                     High               2
```

```r
ggplot(working_dataset) +
  geom_histogram(aes(x = Household.Size), color = "blue", fill = "orange", # add color and fill
                 bins= 10, binwidth = 1, center = 1) + # adjust the binwidth + center of the first bin
  scale_x_continuous(oob = scales::oob_keep, limits = c(0,10), breaks = seq(0,10,1)) + # adjust the tic
  labs(title = "Distribution of Household Size",
       x = "Household Size",
       y = "Frequency")
```

```
## Warning: Removed 11 rows containing non-finite values (`stat_bin()`).
```



```
#Skewed to the right, thus median and IQR
median(working_dataset$Age)
```

```
## [1] 40
```

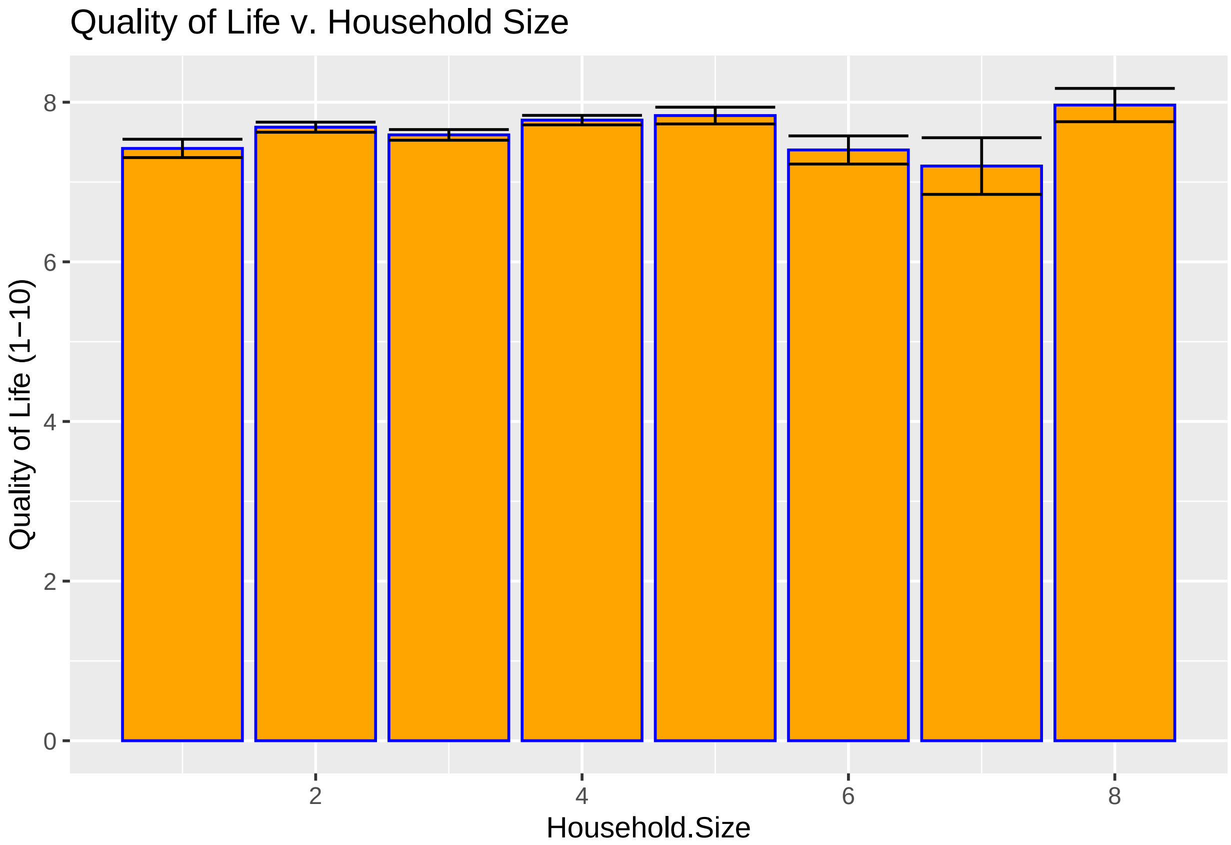
```
IQR(working_dataset$Age)
```

```
## [1] 27
```

Because the distribution of household size is skewed to the right, the median and IQR will be reported. The median and IQR are 40 +/- 25 years old.

```
#Bivariate visualization of age and quality of life
ggplot(data = working_dataset, aes(x = Household.Size, y = Quality.of.Life)) +
  # By default a bar represents a count but we can change what the height of a bar represents
  # Represent a summary stat using the mean function
  geom_bar(stat = "summary", fun = "mean",  color = "blue", fill = "orange") +
  # Adjust the label of the y-axis accordingly
  labs(y = "Quality of Life (1-10)",
       y = "Household Size",
       title = "Quality of Life v. Household Size") +
  # Add error bars
  geom_errorbar(stat = "summary", fun.data = "mean_se")
```

```
## Warning: Removed 11 rows containing non-finite values (`stat_summary()`).
## Removed 11 rows containing non-finite values (`stat_summary()`).
```



Using the difference between the error bars for each household size as comparison, it seems that there isn't any correlation between household size and quality of life. The only significant difference can be seen with household size of 7 vs. 2, 4, 5, and 8 which may just be by random chance.

Model?

```
model <- lm(formula=Quality.of.Life~familiarity_num, data=working_dataset)
summary(model)
```

```
##
## Call:
## lm(formula = Quality.of.Life ~ familiarity_num, data = working_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7874 -0.7874  0.2126  1.2126  3.5895
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.41051    0.07678    83.5   <2e-16 ***
## familiarity_num  0.68845    0.03868    17.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.539 on 2537 degrees of freedom
## Multiple R-squared:  0.111,  Adjusted R-squared:  0.1106
## F-statistic: 316.7 on 1 and 2537 DF,  p-value: < 2.2e-16
```

---

4. Discussion

Question 1: How does household size affect the quality of life for Asian Americans?

The visualization showed that there isn't any correlation between household size and quality of life as seen in visual (#). As seen in the bar graph, there isn't a significant difference in the mean quality of life (1-10) between the different household sizes. The only exception is household size of 7 vs. 2, 4, 5, and 8 which may just be by random chance. The data didn't match my expectations because I believe that lower household size would have higher quality of life. A possible reason for why it fails to meet my prediction is that different household sizes each have their own problems and enjoyment. The only result I'm curious about is why the quality of life for household size of 7 dipped compared to the other household size. The implications of my study is that Asian Americans wouldn't be too worried about having too small or too large of a household size because according to this data, household size didn't have a noticeable effect on their quality of life. Therefore, Asian Americans can have any household size they want as it wouldn't affect their happiness. The main takeaway for these findings is that household size doesn't affect quality of life for Asian Americans living in Austin.

Question 3: What effect does familiarity with America have on Asian American quality of life?

---

5. Reflection, Acknowledgements, and References

The challenging part was cleaning up the data to ensure that it can be analyzed with the visualization to answer our research questions. Thankfully, the data collected by UT Austin AAQoL Research Team was already tidy to begin with which made the process so much easier. To further clean up the data by removing NA value and only selecting the variables of interest, it was due to Dr. Layot's lectures on data wrangling that this step went a lot smoother. Additionally, the making of the ggplots for visualizations was also made easier to the lecture material provided by Dr. Layot.

Contribution: Wendi (Introduction, everything for question #1, reflection), Kevin (Introduction, result), Austin ()

Reference: Link for dataset: https://data.austintexas.gov/dataset/Final-Report-of-the-Asian-American-Quality-of-Life/hc5t-p62z

Link for original study for context: https://www.austintexas.gov/sites/default/files/files/Boards_and_Commissions/Asian%20American%20Quality%20of%20Life%20Study%20(2016).pdf