Exploration

Wendi Deng (wd4737), Kevin Tran, Austin Palmer (ajp4344)

2023-10-20

1. Title and Introduction

For this project, the dataset we will use is a survey regardings the quality of life of Asian Americans in Austin (https://www.austintexas.gov/sites/default/files/files/Boards_and_Commissions/Asian%20American%20Quality%20of%20Life%20Study%20(2016).pdf), Texas. We are interested in this dataset because we want to know what aspects would either increase or decrease the quality of life for Asian Americans living in Austin. This dataset was created by the UT Austin AAQoL Research Team because the researchers wanted to understand the social and health needs of the Asian American community. They mentioned how Asians have overtaken Hispanic as the largest group of immigrants to the US and how by 2050, 10% of the U.S. population will consist of Asians. As a result of this growing minority group, they argued it is important to understand what their needs are to make their life in America better. For our project, we wanted to know what factors may affect Asian Americans' quality of life. Explanatory variables include household size, whether they speak English well, sense of belonging, income, and age. Each row in the dataset represents the participants in the survey, specifically Asian Americans. We would expect decreased household size, speaking English perfectly, feeling a sense of belonging, high income, and low age would reflect a higher quality of life response in the survey. The following are questions we will answer in this project:

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

(Kevin) Question 2: How does income affect the quality of life for Asian Americans?

(Austin) Question 3: How do cultural identity and social integration dynamics affect the quality of life for Asian Americans?

2. Method

The dataset obtained from UT Austin AAQoL Research Team was already tidy because each row is an individual observation (Asian American) and each variable (explanatory and outcome) has its own columns. However, there still some cleanup left to do such as selecting only the six variable we are interested in and removing rows with NA values in any of these variables.

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

# Select Columns of Interest
 col_filtered <- data|>
  select(Quality.of.Life, Income, Age, Household.Size)


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

# Get init/post-filter dimensions
cat("---", "\nFull Dataset:", "\n  Rows:", nrow(data), "\n  Columns:", ncol(data), "\n---",
    "\nWorking Dataset:", "\n  Rows:", nrow(working_dataset), "\n  Columns:", ncol(working_dataset), "\n---")
```

```
## ---
## Full Dataset:
##   Rows: 2609
##   Columns: 231
## ---
## Working Dataset:
##   Rows: 2561
##   Columns: 4
## ---
```

Our dataset is nearly tidy because each row is an individual observation (Asian American) and each variable (explanatory and outcome) has its own columns. We complete tidying it by filtering columns down to our explanatory and 3 factors relevant for the first and second question, then strip out any rows containing NA values.

To explore the impact of social and cultural dynamics on quality of life we filter columns for variables we expect to possibly be indicative of how individuals are fitting in and interacting with society around them. We are liberal with our initial selection of factors as we will trim the set for relevant factors later in this section. In an effort to get a more accurate estimate of β we include a set of control variables unrelated to our question, intended to capture correlated differences in traits between subgroups of our population: `[Age, Income, Gender, Student, Retired, Regular.Exercise]` . To make modeling inputs and results more intuitive, we transform categorical variables in factors and controls into numeric variables where the low/high value represents the low/high end of the variable.

```
# Select variables for analysis of social/cultural integration variable effects on QoL
# We also include control variables to further isolate the true effect
# Control Variables: Income
social_columns <- data |>
  select(Quality.of.Life, English.Speaking, English.Difficulties, Identify.Ethnically, Familiarity.with.America,
Belonging, Discrimination, Duration.of.Residency, US.Born, Close.Friends,
         # Controls
         Age, Income, Gender, Student, Retired, Regular.Exercise)

# **Create working dataset**
#
# Map Categorical variables to numerical
# Max of 4 values per categorical
# Use numerical mappings for simplified regression estimation and plot ordering
#
# Factor Remapping: {
#    English.Speak == speaking_num ~ [0:3]
#    English.Difficulties == difficulties_num ~ [0:3]
#    Identify.Ethnically == eth_identity_num ~ [0:3]
#    Familiarity.with.America == us_familiarity_num ~ [0:3]
#    Belonging == belonging_num ~ [0:3]
#    US.Born == us_born_num ~ [0:1]

#  }
#
# Control Remapping: {
#    Income == income_num ~ [0:7]
#    Gender == male_num ~ [0:1]
#    Retired == retired_num ~ [0:1]
#    Student == student_num ~ [0:1]
#    Age == age_bin ~ [0:80; by=10]
#
# }
#
```

```
##
## Total Rows:  2609
## Rows w/ NA's: 405
## No-NA Rows:  2204
## Data Loss(%):  15.523
```

We end up losing 405 rows, about 15.5% of our dataset. Excluding controls the loss only falls to $\approx 14\%$. We consider this trade-off to be worthwhile and keep the controls. The rows lost per factor is distributed very tightly around ≈ 38 removed rows for N/A values per explanatory. We save the rows with N/A values for the possibility of predicting the null values and reducing the data loss on filtration.

Additionally, we include a list of helper functions we used to handle common tasks:

```
# Builds string output for vector-like obj
# Returns: "[a, b, c]"
format_vec <- function(vec) {
  vec_str <- paste(as.vector(vec), collapse = ", ")
  final_str <- paste0("[", vec_str, "]")
  return(final_str)
}
```

We use a linear regression as an initial filter for inconsequential factors/controls, and a signal those that deserve further exploration.

```
social_factors <- social_wd |>
  select(speaking_num, difficulties_num, eth_identity_num, us_familiarity_num, belonging_num, us_born_num, Close.
Friends, Duration.of.Residency, Quality.of.Life, income_num, male_num, income_num, retired_num, student_num, age_
bin)

social_model <- lm(formula = Quality.of.Life ~ ., data=social_factors)
summary(social_model)
```

```
##
## Call:
## lm(formula = Quality.of.Life ~ ., data = social_factors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0044 -0.8100  0.0927  0.9227  3.6020
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.594778   0.176759  31.652  < 2e-16 ***
## speaking_num           0.367521   0.045731   8.037 1.51e-15 ***
## difficulties_num      -0.129137   0.029029  -4.449 9.09e-06 ***
## eth_identity_num      -0.014876   0.053385  -0.279 0.780536
## us_familiarity_num     0.295171   0.049498   5.963 2.89e-09 ***
## belonging_num          0.159897   0.045958   3.479 0.000513 ***
## us_born_num           -0.435580   0.127557  -3.415 0.000650 ***
## Close.Friends          0.131369   0.024105   5.450 5.62e-08 ***
## Duration.of.Residency  0.006776   0.003617   1.873 0.061137 .
## income_num             0.093690   0.013977   6.703 2.60e-11 ***
## male_num              -0.119231   0.061686  -1.933 0.053383 .
## retired_num            0.012253   0.140914   0.087 0.930718
## student_num           -0.048577   0.105685  -0.460 0.645824
## age_bin(28,38]        -0.092992   0.099436  -0.935 0.349790
## age_bin(38,48]        -0.040847   0.114705  -0.356 0.721798
## age_bin(48,58]        -0.099296   0.132716  -0.748 0.454433
## age_bin(58,68]        -0.057530   0.161045  -0.357 0.720955
## age_bin(68,78]        -0.397004   0.193486  -2.052 0.040305 *
## age_bin(78,88]        -0.199815   0.317420  -0.629 0.529091
## age_bin(88,98]         1.425715   0.828681   1.720 0.085493 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.399 on 2137 degrees of freedom
##   (47 observations deleted due to missingness)
## Multiple R-squared:  0.2251, Adjusted R-squared:  0.2182
## F-statistic: 32.67 on 19 and 2137 DF,  p-value: < 2.2e-16
```

From the results of the regression, we found the vast majority of our factors to be statistically significant. Identifying ethnically and residency duration were the only explanatory variables of interest found not to be statistically significant:

$$\beta_{\text{Eth\_Identity}} \approx 0.035$$

$$\text{SE}_{\text{Eth\_Identity}} \approx 0.052$$

$$\text{P} > 0.10$$

. Our estimator contains zero within a single standard error, giving us confidence in the conclusion of insignificance, and allowing us to disregard strength of ethnical identification as a factor in further analysis. Our control variables were the opposite; most were found to be statistically insignificant. Of the set, we find an individuals income bin to be by far the most statistically relevant ($\beta_{\text{Income}} = 0.098,\ \text{P} < 0.001$). There's also a weak relationship between gender and QoL significant at a 5% alpha level ($\beta_{\text{Gender}} = -0.133,\ \text{P} < 0.05$). We find no significant relationship between retirement/student status with QoL, implying that there is no inherent differences in average happiness between those who are students, in the workforce, or retired. We find similar results for a raw numeric age variable or age bins. Taking this into account, our control variable set is trimmed down to income level and gender.

To account for our factor/control removals, we re-append the rows previously removed based on N/A values in the relevant rows.

```
# Isolate the rows we need to add back in
rows_to_add <- social_columns |>
  filter(is.na(eth_identity_num) | is.na(student_num) | is.na(retired_num)) |>
  drop_na(difficulties_num, speaking_num, us_familiarity_num, belonging_num, us_born_num, Duration.of.Residency,
Close.Friends, Quality.of.Life)

# Combine Dataframes by rows
social_wd <- bind_rows(social_wd, rows_to_add)

# Drop insignificant variables
social_wd <- subset(social_wd, select = c(-Identify.Ethnically, -eth_identity_num, -Retired, -retired_num, -Stude
nt, -student_num, -age_bin, -Age, -Close.Friends, -Duration.of.Residency))
social_factors <- social_wd |>
  select(speaking_num, difficulties_num, us_familiarity_num, belonging_num, us_born_num,  income_num, male_num, Q
uality.of.Life,)

social_model <- lm(formula = Quality.of.Life ~ ., data=social_factors)
summary(social_model)
```

```
##
## Call:
## lm(formula = Quality.of.Life ~ ., data = social_factors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9325 -0.8075  0.0684  0.9540  3.9589
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.70305    0.12473  45.724  < 2e-16 ***
## speaking_num        0.40261    0.04183   9.624  < 2e-16 ***
## difficulties_num   -0.13407    0.02844  -4.714 2.58e-06 ***
## us_familiarity_num  0.33357    0.04737   7.042 2.51e-12 ***
## belonging_num       0.16901    0.03930   4.300 1.78e-05 ***
## us_born_num        -0.29670    0.11045  -2.686  0.00728 **
## income_num          0.10384    0.01246   8.334  < 2e-16 ***
## male_num           -0.13893    0.06073  -2.288  0.02226 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.42 on 2232 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.2093, Adjusted R-squared:  0.2069
## F-statistic: 84.42 on 7 and 2232 DF,  p-value: < 2.2e-16
```

```
# Get a count of rows added back in
nrow(rows_to_add)
```

```
## [1] 42
```

We also further explore the relationship between the factors `English.Difficulties` and `English.Speaking` to ensure the two variables aren't measuring the same effect.

```
# Examine correlation between speaking ability and difficulty levels
# We find a significant inverse relationship


# cat(
#   "`English.Difficulties` categories:", format_vec(unique(social_wd$English.Difficulties)),
#   "\n`English.Speaking` categories:", format_vec(unique(social_wd$English.Speaking))
# )

# Examine regression for significant relationship between English speaking ability
# and difficulties while speaking English.
aux_reg <- lm(formula = difficulties_num ~ speaking_num, data=social_wd)
summary(aux_reg)
```

```
##
## Call:
## lm(formula = difficulties_num ~ speaking_num, data = social_wd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7639 -1.0554 -0.2915  0.7085  1.9446
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.76387    0.05645  31.246   <2e-16 ***
## speaking_num -0.23616    0.02512  -9.403   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.085 on 2244 degrees of freedom
## Multiple R-squared:  0.03791,    Adjusted R-squared:  0.03748
## F-statistic: 88.41 on 1 and 2244 DF,  p-value: < 2.2e-16
```

```
social_wd |>
  select(speaking_num, difficulties_num) |>
  cor()
```

```
##                  speaking_num difficulties_num
## speaking_num        1.0000000       -0.1946955
## difficulties_num   -0.1946955        1.0000000
```

We find a small but significant relationship, English speaking ability explains about 4% of variation in frequency of English difficulties. We find similar evidence in the correlation matrix, showing us an weak inverse relationship. Even though the relationship is significant, they seem independent enough that we will leave both variables in as explanitories.

As a final check for the presence of endogeneity between speaking ability and frequency of English difficulties, we compare the results of a long regression containing both as factors and a short regression excluding frequency of difficulties. We find around a $.35\mathrm{SE}$ difference in speaking ability's β estimation, a marginal change in magnitude, further discrediting the idea of of speaking ability being endogenous. We also see that by including frequency of difficulties, our model's explanatory power increases by around 1%, which convinced us to include difficulties frequency as a factor in the final set.

$$\mathrm{R}_{\mathrm{S}}^{2} = 0.201$$

$$\mathrm{R}_{\mathrm{L}}^{2} = 0.209$$

We also use the long model to confirm the existence of a strong and significant relationship with Quality of Life for all other explanitories and controls.

```
short_model <- lm(formula = Quality.of.Life ~. - difficulties_num, data=social_factors)
long_model <- lm(formula = Quality.of.Life ~ ., data=social_factors)
stargazer(short_model, long_model, type = "text")
```

```
## 
## =====================================================================
##                                    Dependent variable:
##                     -------------------------------------------------
##                                      Quality.of.Life
##                               (1)                      (2)
## ---------------------------------------------------------------------
## speaking_num                0.418***                 0.403***
##                             (0.042)                  (0.042)
## 
## difficulties_num                                    -0.134***
##                                                      (0.028)
## 
## us_familiarity_num          0.361***                 0.334***
##                             (0.047)                  (0.047)
## 
## belonging_num               0.155***                 0.169***
##                             (0.039)                  (0.039)
## 
## us_born_num                 -0.221**                -0.297***
##                             (0.110)                  (0.110)
## 
## income_num                  0.103***                 0.104***
##                             (0.013)                  (0.012)
## 
## male_num                    -0.157**                 -0.139**
##                             (0.061)                  (0.061)
## 
## Constant                    5.482***                 5.703***
##                             (0.116)                  (0.125)
## 
## ---------------------------------------------------------------------
## Observations                 2,240                    2,240
## R2                           0.201                    0.209
## Adjusted R2                  0.199                    0.207
## Residual Std. Error    1.426 (df = 2233)        1.420 (df = 2232)
## F Statistic         93.898*** (df = 6; 2233) 84.423*** (df = 7; 2232)
## =====================================================================
## Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

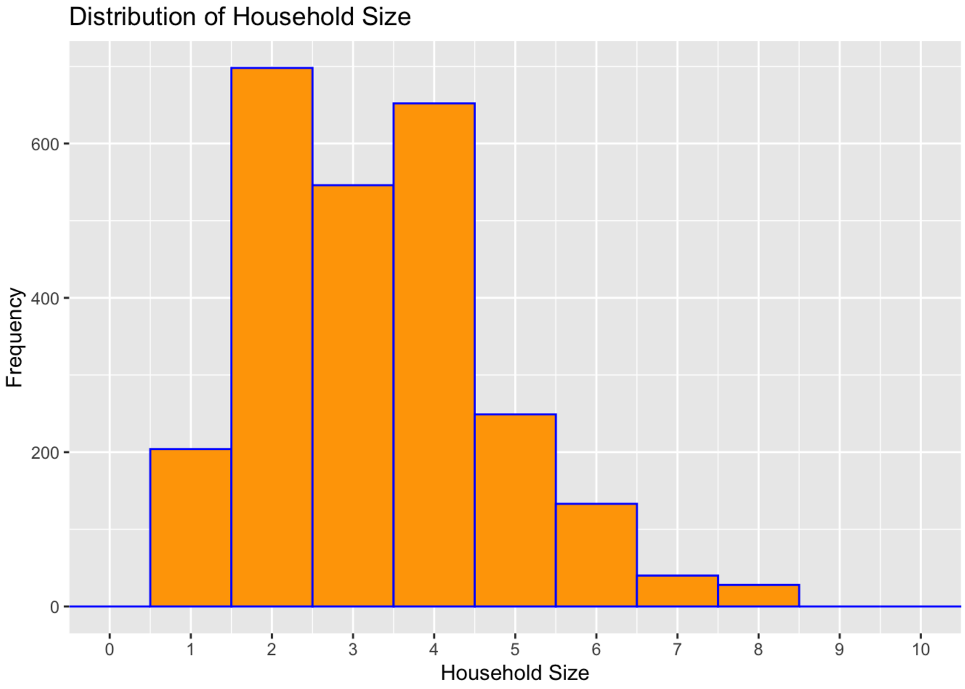
We find

3. Results

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

```
# Numeric univariate visualization
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 tick marks of
the x-axis
  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
```

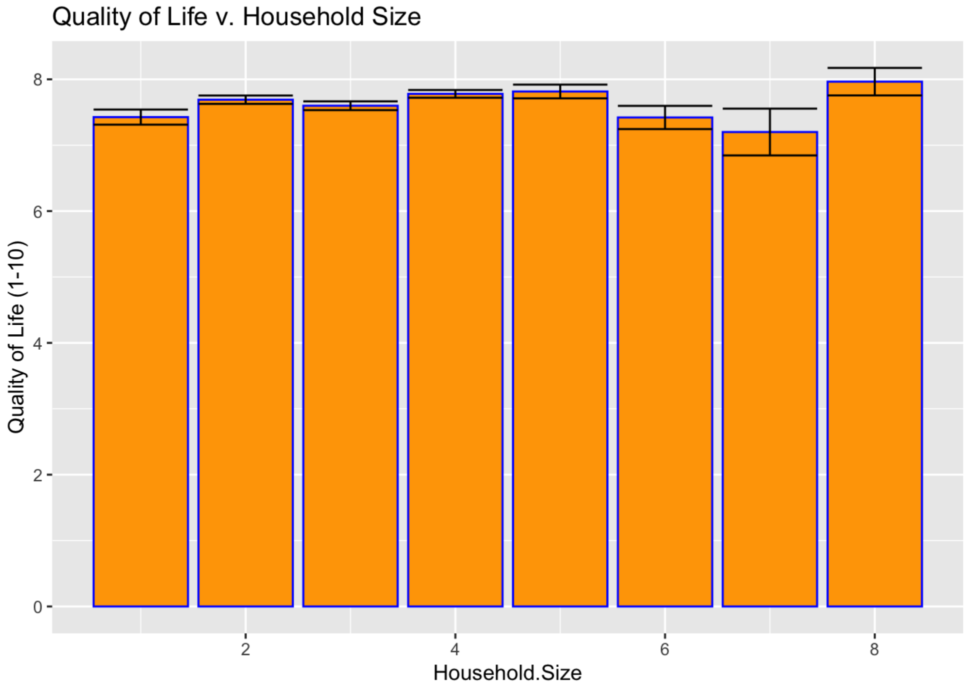
```
IQR(working_dataset$Age)
```

```
## [1] 26
```

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()`).
```

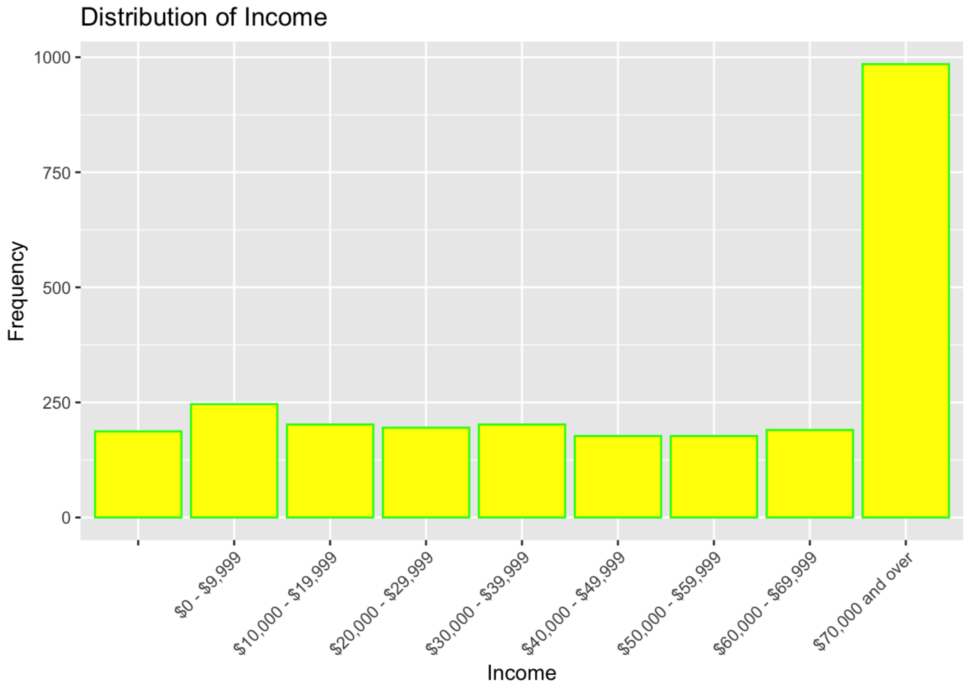


Using the difference between the error bars for each household size as comparison, it seems that there isn't any correlation between household size and quality of life. The only significant difference can be seen with household size of 7 vs. 2, 4, 5, and 8 which may just be by random chance.

Question 2: How does income affect the quality of life for Asian Americans?

```
ggplot(working_dataset) +
  geom_bar(aes(x = Income), color = "green", fill = "yellow",
               bins= 10, binwidth = 1, center = 1) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of Income",
       x = "Income",
       y = "Frequency")
```

```
## Warning in geom_bar(aes(x = Income), color = "green", fill = "yellow", bins =
## 10, : Ignoring unknown parameters: `bins`, `binwidth`, and `center`
```

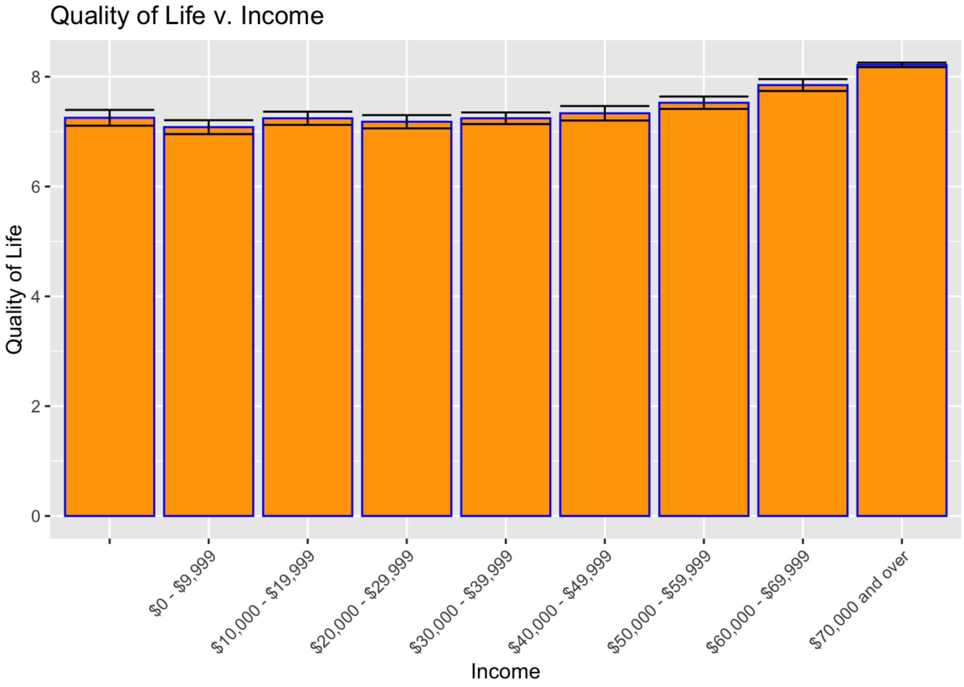


```
median(working_dataset$Income, na.rm =  T)
```

```
## [1] "$50,000 - $59,999"
```

From the data given, there are more subjects with an income as high as $70,000 or above. The median of the data is nearly around $60,000-$69,999.

```
working_dataset |>
ggplot(aes(x = Income, 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") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Income",
       y = "Quality of Life",
       title = "Quality of Life v. Income") +
  # Add error bars
  geom_errorbar(stat = "summary", fun.data = "mean_se")
```



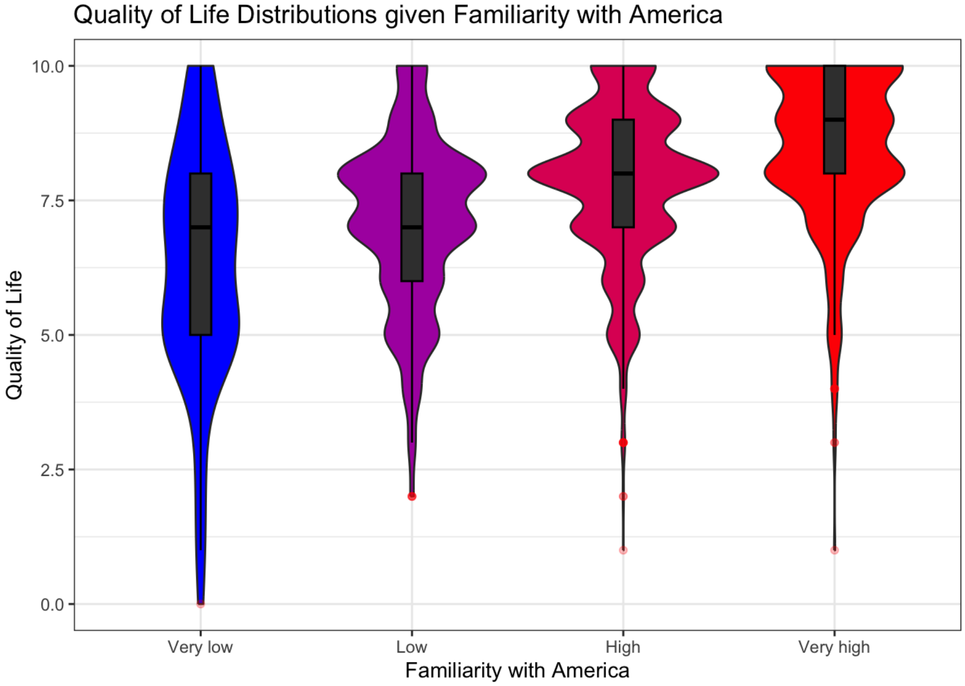
The visualization of the relationship between income and the quality of life provided that the quality of life slowly increases from increasing income. From the data, those that have at least $70,000 are more likely to experience a higher quality of life, proving it has statistical difference from those that have less than $70,000 in their deposits.

Question #3: How do social and cultural dynamics affect the quality of life for Asian Americans?

The first thing we notice is that those born in the US are significantly less happy on average at any alpha level than those who immigrate by about 5%. ($\beta_{\text{born\_us}} = -0.476, P \approx 2.09 \times 10^{-}5$).

```
# Ordered Label Switch Vectors
desired_order <- c("Very low", "Low", "High", "Very high")
desired_order_num <- c(0, 1, 2, 3)

# Distribution of QoL depicted w/ Violin Plot for each familiarity level
# Box Plot included for easy comparison of means between groups
social_wd |>
  ggplot(aes(x = as.factor(us_familiarity_num), y = Quality.of.Life)) +
  geom_violin(aes(fill = us_familiarity_num)) +
  geom_boxplot(fill="#3d3d3d", color="black", width = 0.1, outlier.alpha = 0.25, 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_discrete(breaks = desired_order_num, labels = desired_order) +
  scale_fill_gradient(low = "blue", high = "red") +  # Optional: Color gradient for fill
  theme_bw() +
  theme(legend.position = "none")
```



Model?

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

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 2: How does income affect the quality of life for Asian Americans?

Overall, the visualization provided the bar graph of each income level in relation to the quality of life that Asian Americans have experienced. The graph displayed a somewhat increase in the quality of life due to the increased income level, though gradually. The data satisfies my expectation that those with high income or profit are more likely to be lively and experience enjoyment in life. The implication states that Asian Americans would be motivated to obtain a higher profit at their possession as a way to live a better life. Therefore, income can have an effect on how Asian Americans are thriving.

Question 3: How do cultural identity and social integration dynamics affect the quality of life for Asian Americans?

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 (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
(https://www.austintexas.gov/sites/default/files/files/Boards_and_Commissions/Asian%20American%20Quality%20of%20Life%20Study%20(201**