DATA 606 Data Final Project

Amber Ferger

Part 1 - Introduction

My research question is: **Is there a relationship between age, income, and region and an individual's comfort level with public religious displays?** I care about this particular subject because in an ever increasingly polar political world, it is important to understand the general beliefs and viewpoints of individuals from all facets of life. Understanding whether demographics play a role in religious comfort level can help us better empathize with individuals from all walks of life. This in turn can help to guide discussions and aid in decision making.

Part 2 - Data

- **Data collection:** This data was collected using a SurveyMonkey poll, conducted between July 29 and August 1, 2016. The survey asked 661 respondents questions about public displays of religion. Source data can be found here: https://github.com/fivethirtyeight/data/tree/master/religion-survey
- **Cases:** Each case represents a respondent's survey answers. There are a total of 979 cases in the dataset.
- **Variables:** The dependent (response) variable is comfort level and it is numeric. The independent variables are Age (quantitative), Income (quantitative), and Region (qualitative).
- **Type of study:** This is a survey, which is a type of observational study.
- **Scope of inference - generalizability:** The population of interest is all adult individuals within the US. I don't believe this information can be generalized to the population as a whole because it inherently eliminates individuals that do not have access to a computer, which could be entire geographic regions. Additionally, although we know that the SurveyMonkey Audience service was used to generate survey participants, we don't know what the actual methodology was for selecting the individuals that the link was sent to, so it might not represent the population as a whole. Finally, the survey results could contain bias as a result of non-response. It is possible that those that chose not to respond to the survey differ drastically from those that did respond.
- **Scope of inference - causality:** By nature, surveys cannot be used to establish causal links between the variables of interest. We can identify whether there is a relationship between the two variables, but correlation does not imply causation.

Part 3 - Exploratory data analysis

Data Cleansing

There are a number of fields within this dataset, so I am going to subset it only to general demographics and survey responses related to the comfort of seeing religious actions outside of the respondent's religion.

```
# rename columns
colNames <- c('RELIGION','RELIGION2', 'EVANGELICAL', 'RELIGIOUS_SERVICES', 'FREQ_PRAY_WITH_MOTIONS', 'F
'FREQ_PRAY_FOR_OTHERS', 'FREQ_ASK_TO_PRAY_WITH_SOMEONE', 'FREQ_BRING_UP_RELIGION',
'FREQ_ASK_ABOUT_RELIGION', 'FREQ_DECLINE_FOOD_FOR_RELIGION', 'FREQ_WEAR_RELIGIOUS_CLOTHING', 'FREQ_PART

'COMFORT_OWN_PRAY_WITH_MOTIONS',
'COMFORT_OWN_PRAY_WITH_OBJECTS',
'COMFORT_OWN_PRAY_BEFORE_MEALS',
'COMFORT_OWN_PRAY_FOR_OTHERS',
'COMFORT_OWN_ASK_TO_PRAY_WITH_SOMEONE',
```

```
'COMFORT_OWN_BRING_UP_RELIGION',
'COMFORT_OWN_ASK_ABOUT_RELIGION',
'COMFORT_OWN_DECLINE_FOOD_FOR_RELIGION',
'COMFORT_OWN_WEAR_RELIGIOUS_CLOTHING',
'COMFORT_OWN_PARTICIPATE_IN_PUBLIC_RELIGIOUS_EVENT',

'COMFORT_OTHER_PRAY_WITH_MOTIONS', 'COMFORT_OTHER_PRAY_WITH_OBJECTS', 'COMFORT_OTHER_PRAY_BEFORE_MEALS'
'COMFORT_OTHER_WEAR_RELIGIOUS_CLOTHING', 'COMFORT_OTHER_PARTICIPATE_IN_PUBLIC_RELIGIOUS_EVENT', 'COMFORT
'COMFORT_SEE_OTHER_ASK_ABOUT_RELIGION', 'COMFORT_SEE_OTHER_DECLINE_FOOD_FOR_RELIGION', 'COMFORT_SEE_OTH

names(religionData) <- colNames

colsToKeep <- c('COMFORT_SEE_OTHER_PRAY_WITH_MOTIONS', 'COMFORT_SEE_OTHER_PRAY_WITH_OBJECTS', 'COMFORT_S
'COMFORT_SEE_OTHER_ASK_ABOUT_RELIGION', 'COMFORT_SEE_OTHER_DECLINE_FOOD_FOR_RELIGION', 'COMFORT_SEE_OTH

religionData <- religionData %>%
  filter(RELIGION != 'Response' & HOUSEHOLD_SALARY != 'Prefer not to answer' & US_REGION != '') %>%
  select(colsToKeep)
```

In order to quantify each individual's comfort with public religious displays, I will convert the categorical features to numeric. First, I will rank each survey response in order of comfort (with Not at all comfortable having the lowest ranking and Extremely comfortable having the highest ranking). Since the columns are already ordered, the numerical values should maintain the hierarchy. Next, I will compute an average ranking across survey responses. The new variable called *AVERAGE_RANKING* contains this information.

```
comfortColumns <- seq(1:10)

for (i in comfortColumns) {
  religionData[[i]] <- factor(religionData[[i]], levels = c("", "Response", "Not at all comfortable", "

}

# eliminate records where any of the survey responses are blank
toBeRemoved<-which(religionData$COMFORT_SEE_OTHER_PRAY_WITH_MOTIONS==""|religionData$COMFORT_SEE_OTHER_

religionData<-religionData[-toBeRemoved,]

religionData$AVERAGE_RATING <- 0


# loop through all survey questions and convert each response to a number
# add number to the AVERAGE_RATING column
for (i in comfortColumns) {
  religionData[[i]]<-as.numeric(religionData[[i]])
  religionData$AVERAGE_RATING <- religionData$AVERAGE_RATING + religionData[[i]]
}

# final average rating
religionData$AVERAGE_RATING <- religionData$AVERAGE_RATING/10
```

The last thing we need to do is convert our age and household income variables to numeric. We will use the lower bound in each of the variables to represent the number.

```
lookup1 <- data.frame(AGE=c('18 - 29','30 - 44','45 - 59','60+'), NUMERIC_AGE=c(18,30,45,60))
lookup2 <- data.frame(HOUSEHOLD_SALARY=c('$0 to $9,999','$10,000 to $24,999','$25,000 to $49,999','$50,

lookupData <- merge(religionData, lookup1, by="AGE")
lookupData2 <- merge(lookupData, lookup2, by ="HOUSEHOLD_SALARY")
lookupData2$AGE <- lookupData2$NUMERIC_AGE
lookupData2$HOUSEHOLD_SALARY <- lookupData2$NUMERIC_INCOME
religionData <- lookupData2
```

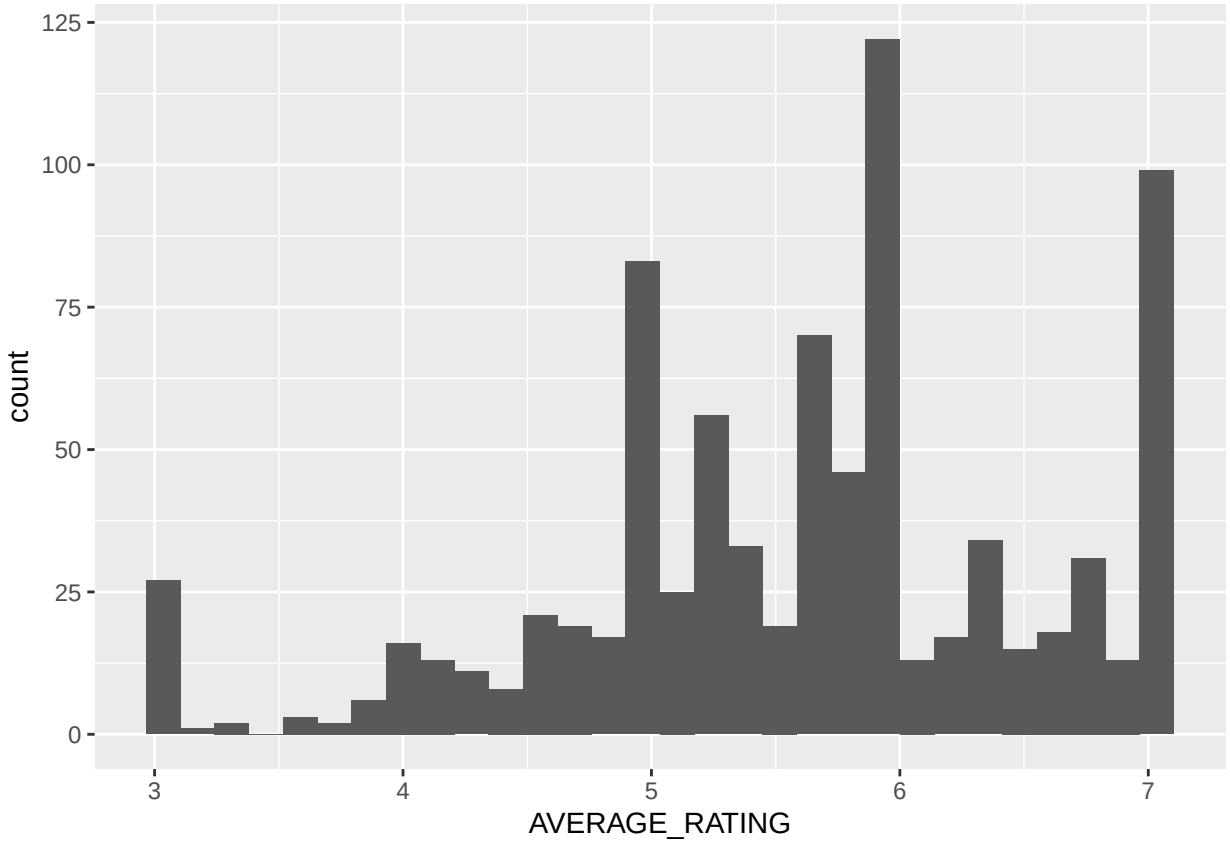
Summary Statistics

The summary statistics show that the average comfort with religious displays is ~5.613, with a min value of 3 and a max value of 7. We can see from the histogram of average ratings that our data is relatively normal with a left-skew.

```
summary(religionData$AVERAGE_RATING)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   5.700   5.613   6.225   7.000
```

```
ggplot(religionData, aes(x=AVERAGE_RATING))+ geom_histogram()
```

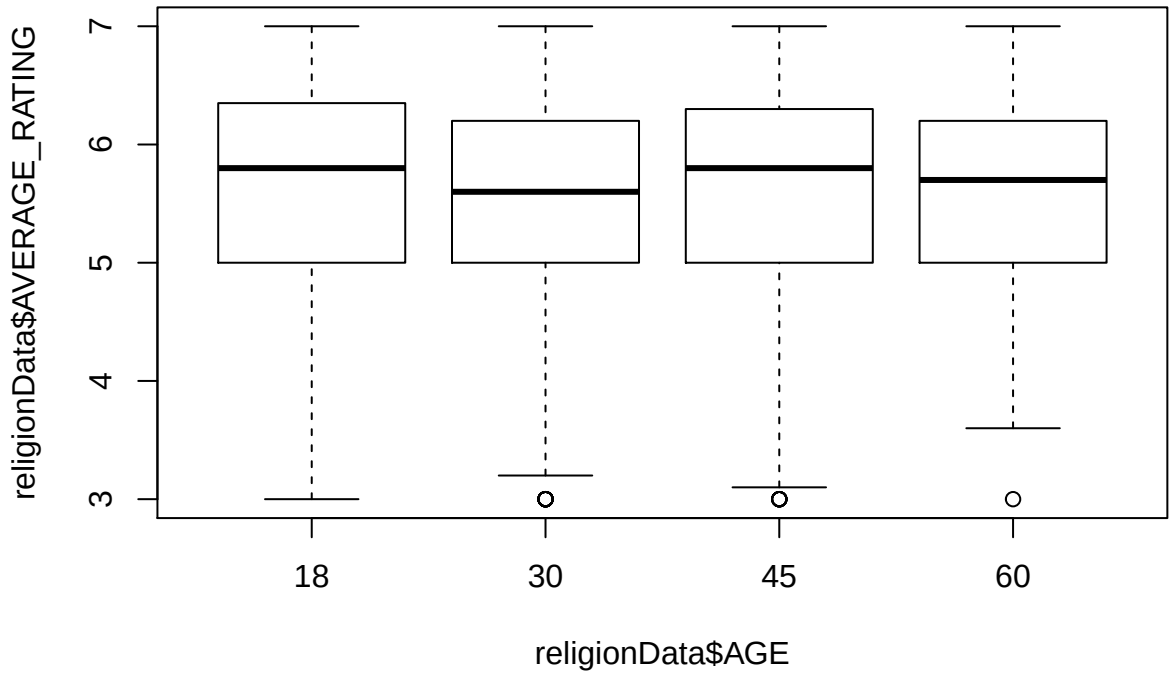


AGE: A first glance at the Age variable shows that there isn't much variability in comfort rating between age groups.

```
religionData %>%
  group_by(AGE) %>%
  summarise(MEAN_BY_AGE = mean(AVERAGE_RATING),
            MEDIAN_BY_AGE = median(AVERAGE_RATING),
            STDEV_BY_AGE = sd(AVERAGE_RATING))
```

```
## # A tibble: 4 x 4
##     AGE MEAN_BY_AGE MEDIAN_BY_AGE STDEV_BY_AGE
##   <dbl>       <dbl>         <dbl>        <dbl>
## 1    18        5.68           5.8        0.979
## 2    30        5.56           5.6        1.03
## 3    45        5.62           5.8        0.968
## 4    60        5.60           5.7        0.835
```

```
boxplot(religionData$AVERAGE_RATING~religionData$AGE)
```



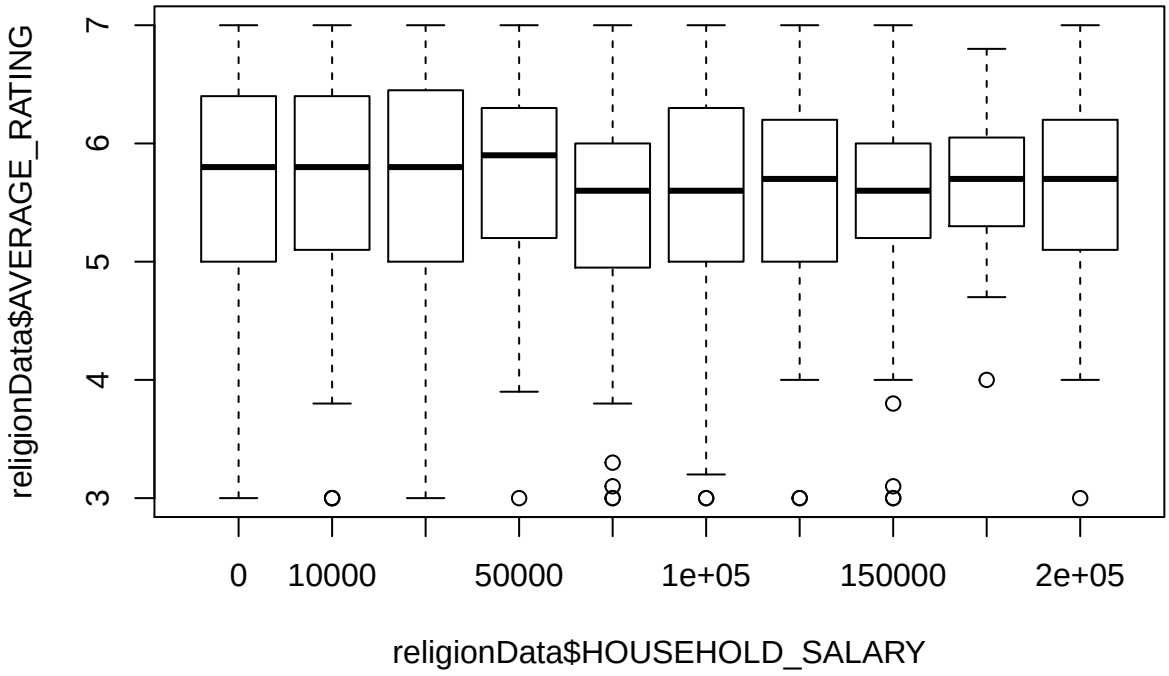
INCOME: A first glance at the Household Salary variable shows a similar output – there doesn't appear to be much variability in average comfort ratings between groups. However, we do see a slightly higher comfort level from salary groups between $10,000 and $74,999.

```
religionData %>%
  group_by(HOUSEHOLD_SALARY) %>%
  summarise(MEAN_BY_SALARY = mean(AVERAGE_RATING),
            MEDIAN_BY_SALARY = median(AVERAGE_RATING),
            STDEV_BY_SALARY = sd(AVERAGE_RATING))  %>%
```

```
  ungroup %>%
  arrange(-MEAN_BY_SALARY)
```

```
## # A tibble: 10 x 4
##    HOUSEHOLD_SALARY MEAN_BY_SALARY MEDIAN_BY_SALARY STDEV_BY_SALARY
##               <dbl>          <dbl>            <dbl>           <dbl>
##  1            50000           5.76              5.9           0.832
##  2            10000           5.68              5.8           0.946
##  3            25000           5.63              5.8           1.01
##  4           175000           5.61              5.7           0.728
##  5           200000           5.60              5.7           0.855
##  6           125000           5.58              5.7           0.983
##  7                0           5.57              5.8           1.14
##  8           100000           5.54              5.6           0.939
##  9            75000           5.49              5.6           0.930
## 10           150000           5.45              5.6           1.01
```

```
boxplot(religionData$AVERAGE_RATING~religionData$HOUSEHOLD_SALARY)
```



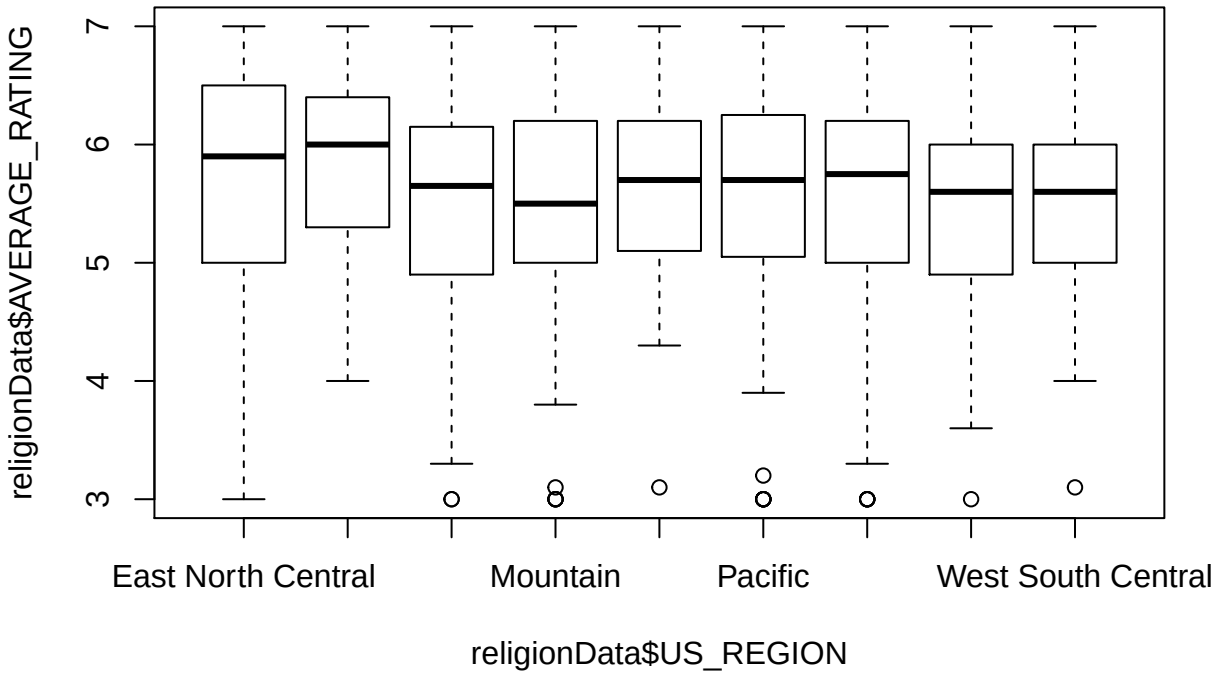
REGION: And here's where it starts to get more interesting. It looks like there might be a true difference in comfort rating due to geographic region.

```
religionData %>%
  group_by(US_REGION) %>%
  summarise(MEAN_BY_REGION = mean(AVERAGE_RATING),
```

```
            MEDIAN_BY_REGION = median(AVERAGE_RATING),
            STDEV_BY_REGION = sd(AVERAGE_RATING)) %>%
  ungroup %>%
  arrange(-MEAN_BY_REGION)
```

```
## # A tibble: 9 x 4
##   US_REGION          MEAN_BY_REGION MEDIAN_BY_REGION STDEV_BY_REGION
##   <chr>                       <dbl>            <dbl>           <dbl>
## 1 East South Central           5.88             6              0.781
## 2 East North Central           5.71             5.9            1.03
## 3 New England                  5.68             5.7            0.817
## 4 South Atlantic               5.62             5.75           0.966
## 5 West South Central           5.61             5.6            0.840
## 6 Pacific                      5.59             5.7            1.000
## 7 Middle Atlantic              5.55             5.65           0.929
## 8 West North Central           5.48             5.6            0.880
## 9 Mountain                     5.43             5.5            1.09
```

```
boxplot(religionData$AVERAGE_RATING~religionData$US_REGION)
```



Part 4 - Inference

I will be doing a multiple regression in order to identify if there is a relationship between salary, region, age, and comfort level.

```r
model <- lm(AVERAGE_RATING ~ HOUSEHOLD_SALARY + US_REGION + AGE, data = religionData)
summary(model)
```

```
##
## Call:
## lm(formula = AVERAGE_RATING ~ HOUSEHOLD_SALARY + US_REGION +
##     AGE, data = religionData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7656 -0.5932  0.1102  0.6344  1.6005
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  5.784e+00  1.179e-01  49.047   <2e-16 ***
## HOUSEHOLD_SALARY            -5.474e-07  6.169e-07  -0.887   0.3752
## US_REGIONEast South Central  1.667e-01  1.664e-01   1.002   0.3167
## US_REGIONMiddle Atlantic    -1.592e-01  1.223e-01  -1.302   0.1932
## US_REGIONMountain           -2.840e-01  1.464e-01  -1.940   0.0527 .
## US_REGIONNew England        -2.219e-02  1.560e-01  -0.142   0.8869
## US_REGIONPacific            -1.084e-01  1.169e-01  -0.927   0.3540
## US_REGIONSouth Atlantic     -9.284e-02  1.085e-01  -0.856   0.3923
## US_REGIONWest North Central -2.296e-01  1.473e-01  -1.559   0.1194
## US_REGIONWest South Central -1.123e-01  1.357e-01  -0.828   0.4080
## AGE                         -9.972e-04  2.239e-03  -0.445   0.6562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9545 on 829 degrees of freedom
## Multiple R-squared:  0.01208,    Adjusted R-squared:  0.0001675
## F-statistic: 1.014 on 10 and 829 DF,  p-value: 0.4294
```

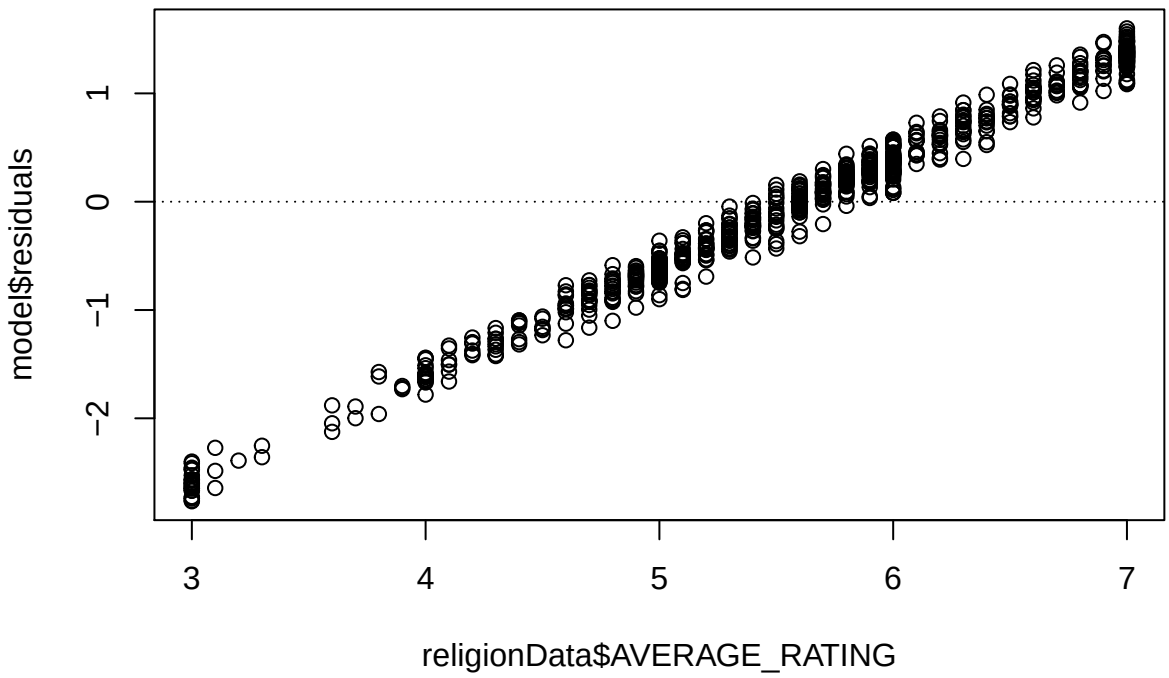
The coefficients of each variable tell us the amount of increase in comfort as a result of a particular value. For example, if a person is from the West South Central region, they have an average comfort level that is 0.1123 points lower than someone who is not.

Our final model shows that none of the variables are significant enough to reject the null hypothesis (p-value for each variable is > 0.05). There is one that comes close - The US Mountain region, but it is not low enough to accept the alternative hypothesis.

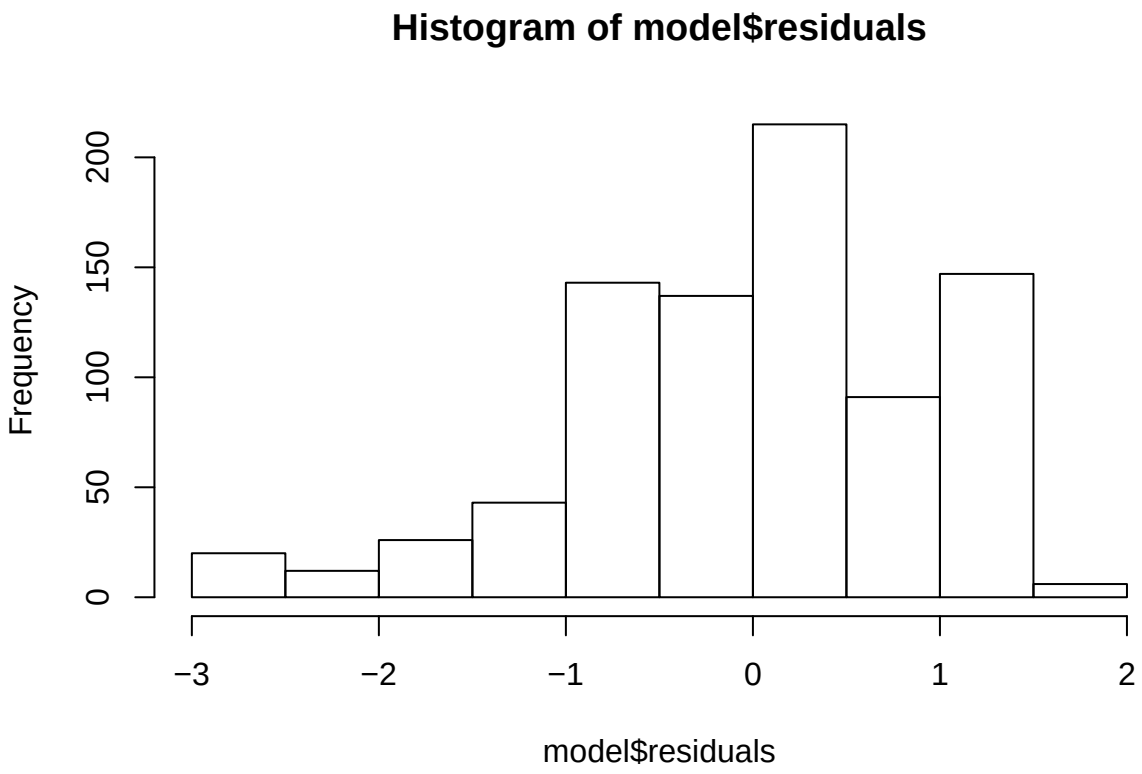
Additionally, we can see that the R^2 value is 0.01208, which tells us that only 12% of the variability in the model is explained by these factors.

Checking conditions

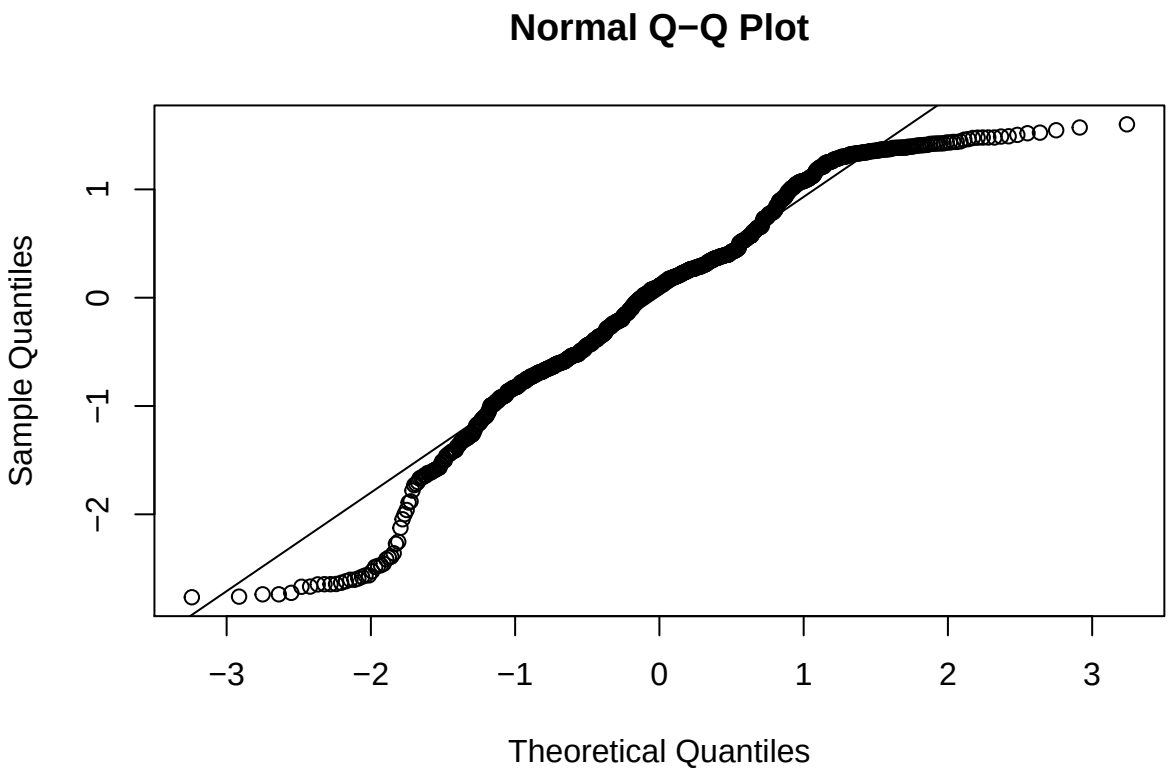
```r
plot(model$residuals ~ religionData$AVERAGE_RATING)
abline(h = 0, lty = 3)
```




```
hist(model$residuals)
```




```
qqnorm(model$residuals)
qqline(model$residuals)
```



The residuals show us:

- **Linearity:** Is not met - there appears to be a pattern in the residuals (increasing trend)
- **Nearly Normal Residuals:** Aside from a slight left-skew, the residuals appear to be normally distributed
- **Constant Variability:** This also does not pass. The tails of the residuals stray from the line.

Part 5 - Conclusion

From this analysis, we learned that there is not a statistically significant relationship between age, region, income, and average comfort with religious displays. Additionally, we learned that the raw data that was provided does not support proper inference, which means that we should re-evaluate the variables that were chosen to analyze. For future work, I would like to examine the relationships between some of the other variables included in this dataset (ex: gender, ethnicity).

References

FiveThirtyEight Data - https://github.com/fivethirtyeight/data/tree/master/religion-survey