What is the relationship between income level and perceptions of life and others?

Sarah Nakamoto

University of Minnesota

PSY 8712 Data Science in Psychology

Dr. Richard Landers

May 3, 2024

What is the relationship between income level and perceptions of life and others?

The main purpose of this project is to explore the relationship of income level with other variables in the General Social Survey 2018 dataset. A Shiny application was created to explore how various factors relating to demographic factors of respondents (sex, race), life perceptions (life excitement, happiness) and perceptions of others (fairness, trust in others) change the distribution of income levels in the dataset. Based on the visualization, three variables were selected for further analysis: fairness, happiness, and race. Additionally, three machine learning algorithms (elastic net, random forest, eXtreme gradient boosting) were compared and used to predict income from the data with less than 75% missingness.

## Hypotheses and Research Questions

RQ1: How effectively can the variables in the GSS 2018 dataset predict income?

RQ2: Are there significant differences in income level for different perceptions of fairness?

RQ3: Are there significant differences in income level for different happiness levels?

RQ4:Are there significant differences in income level for different races?

# Method

### Open Science Materials

For this project, I created a binder, which converts a Github repository into interactive notebooks. Binder allows for the notebooks to be run in an executable environment, making code reproducible for anyone, even those with different virtual environments (i.e. a different version of R). My binder for the project is linked here: <https://mybinder.org/v2/gh/nakam087/psy8712-final/HEAD>

The study materials can be accessed via the psy8712-final repository via the Github ID: nakam087. The readme, placed below the folders and files, details the contents of the folders (where the code and data is placed) and contains links to both the Shiny app/binder.

### Participants

There were 1983 total participants in the dataset, who were recruited by the National Opinion Research Center to explain trends in contemporary American society. This is a retrospective analysis on data that is publicly available.

### Measures

RQ1 explores how effectively the variables in the GSS dataset can predict income, so for this process, 3 machine learning algorithms were used (elastic net, random forest, eXtreme gradient boosting) and assessed for model performance. For RQ2-RQ4, which explore specific variables related to income, an interactive Shiny app was created, which allows the user to filter an income level visualization based on sex, happiness level, race, trust level, fairness, and life excitement. Specific questions for the fairness, happiness and race variables were created based on filtering results on the visualizations, which were statistically investigated through linear regression analyses. Significant results were further explored through a Tukey’s HSD test.

### Procedure

A nationally representative sample of American non-institutionalized adults were recruited for the General Social Survey. Participants were interviewed (median=1.5 hours) based on items that have been selected by the Board of Overseers, and data were processed according to NORC procedures. The survey includes items related to demographics, political and religious attitudes, and psychological evaluation.

# Analyses

### Descriptive Statistics and Static Visualizations

Below is a histogram of total family income level, just to get a sense of the GSS 2018 dataset. It seems that the majority of respondents are in the higher income level.

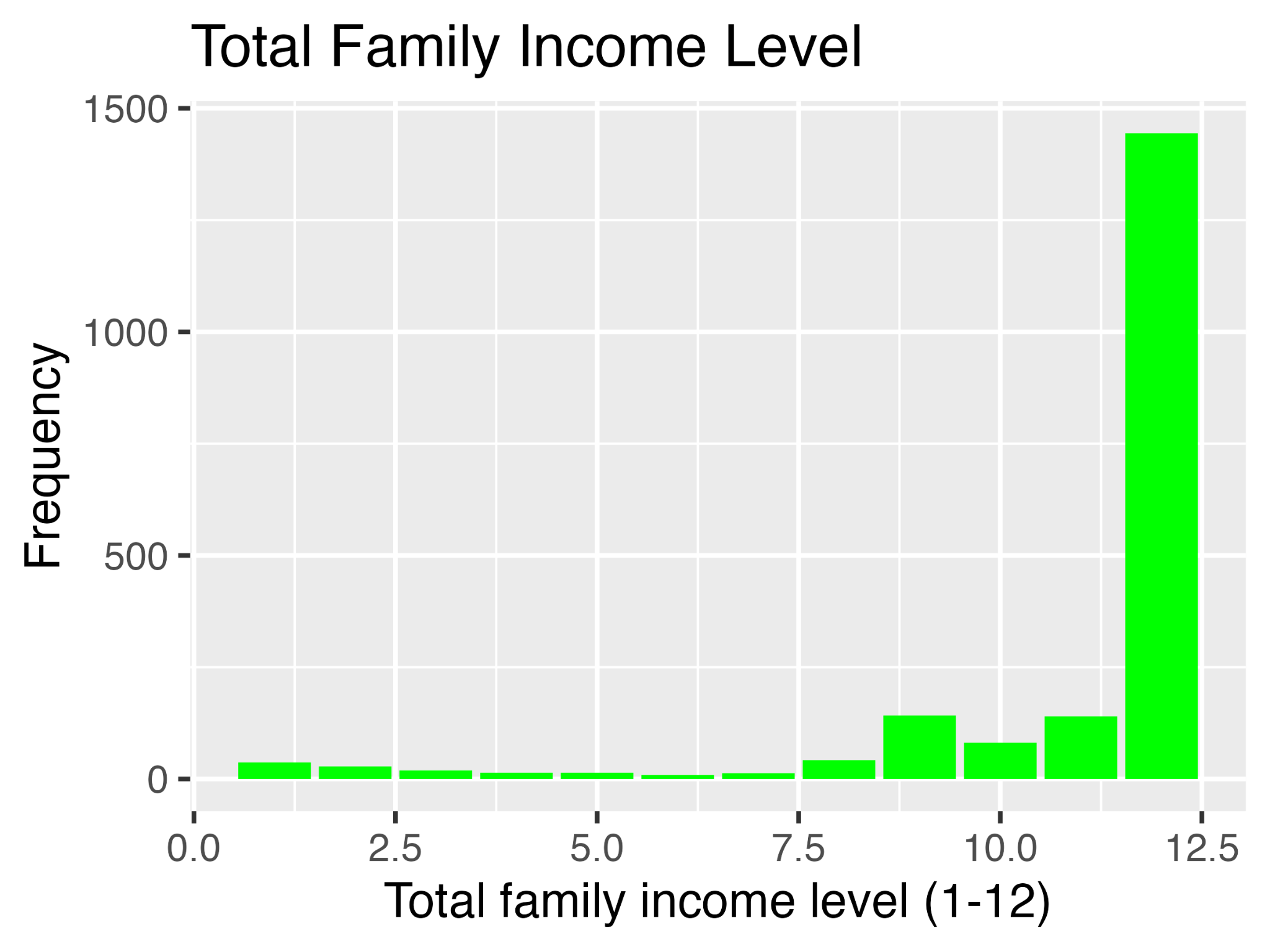
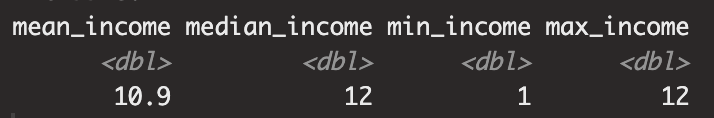


Table of summary statistics for the machine learning data



### Interactive Visualization

The interactive visualization was used to display the effects of filtering variables on the income histogram. The Shiny application explored how various factors relating to the demographics of respondents (sex, race), life perceptions (life excitement, happiness) and perceptions of others (fairness, trust in others) change the distribution of income levels in the dataset. Participants have the options to select between various levels of these factors (radio buttons) to explore how choosing different options changes the display of income. In total, there are 6 different sets of radio buttons, each with 2-3 choices.

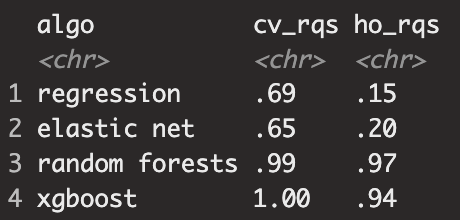
### Data Cleaning

The data was cleaned to exclude other factors related to income, as this could cloud the results (other income results, like personal income, may relate almost perfectly to family income). This removed 4 other variables. Additionally, only variables with less than 75% missingness were included to ensure better model fit, as those variables may not provide enough information to accurately predict any variables. All variables were additionally changed to numeric, as alternative formats will cause errors downstream. Thus, the machine learning data was reduced to 1983 observations of 834 variables, which is still a large enough dataset to employ machine learning. For the Shiny web app, the dataset was reduced further to include only income and variables of interest. These variables were perceptions of fairness, trust in others, perceptions of life, race, sex, and age. Additionally, to avoid errors in visualization, all empty rows were also taken out. For ease of use, the factor levels of the variables were converted to their respective answers from the codebook. This reduced the dataset further to 653 observations of 14 variables.

### Analysis

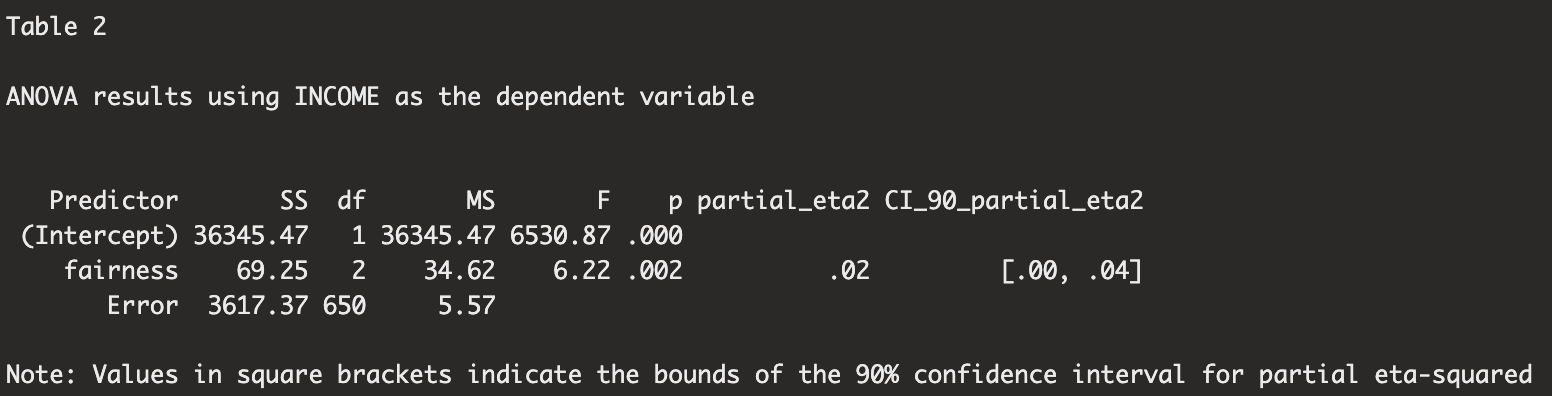
RQ1: How effectively can the variables in the GSS 2018 dataset predict income?

This is a table of the performance of the machine learning algorithms.

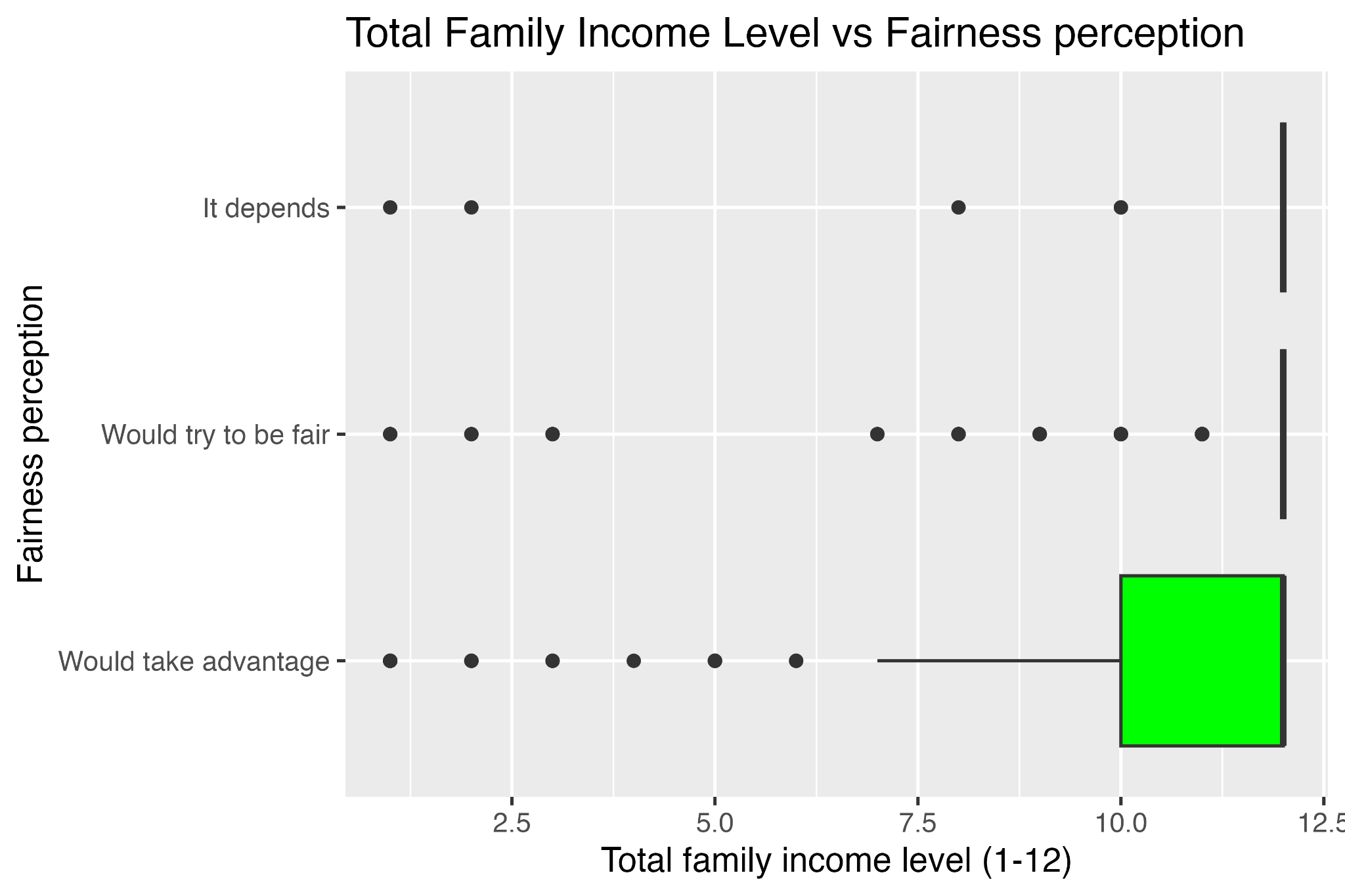


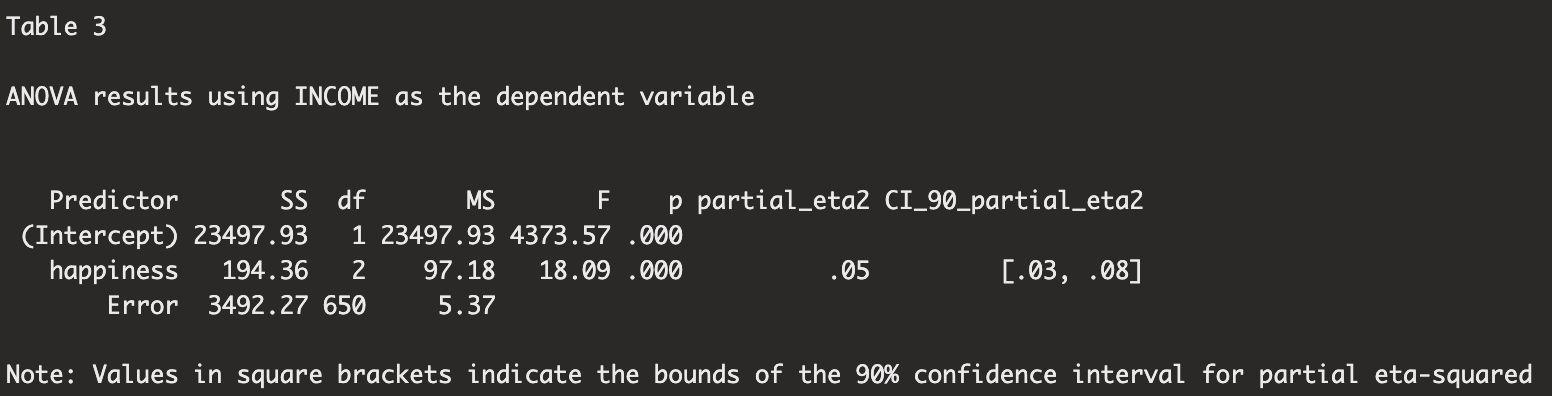
From this table, it seems that random forest was the best performing algorithm, as it was able to perform well on both training and novel data (a good balance of bias and variance). The other models did not perform as well generally, as it seems some had difficulty performing on novel data (high bias), or just performed worse overall. It seems that income can be predicted pretty well from the GSS 2018 data as approximately 97% of the variation in income the variable being predicted) can be explained by the independent variable(s) in the regression model.

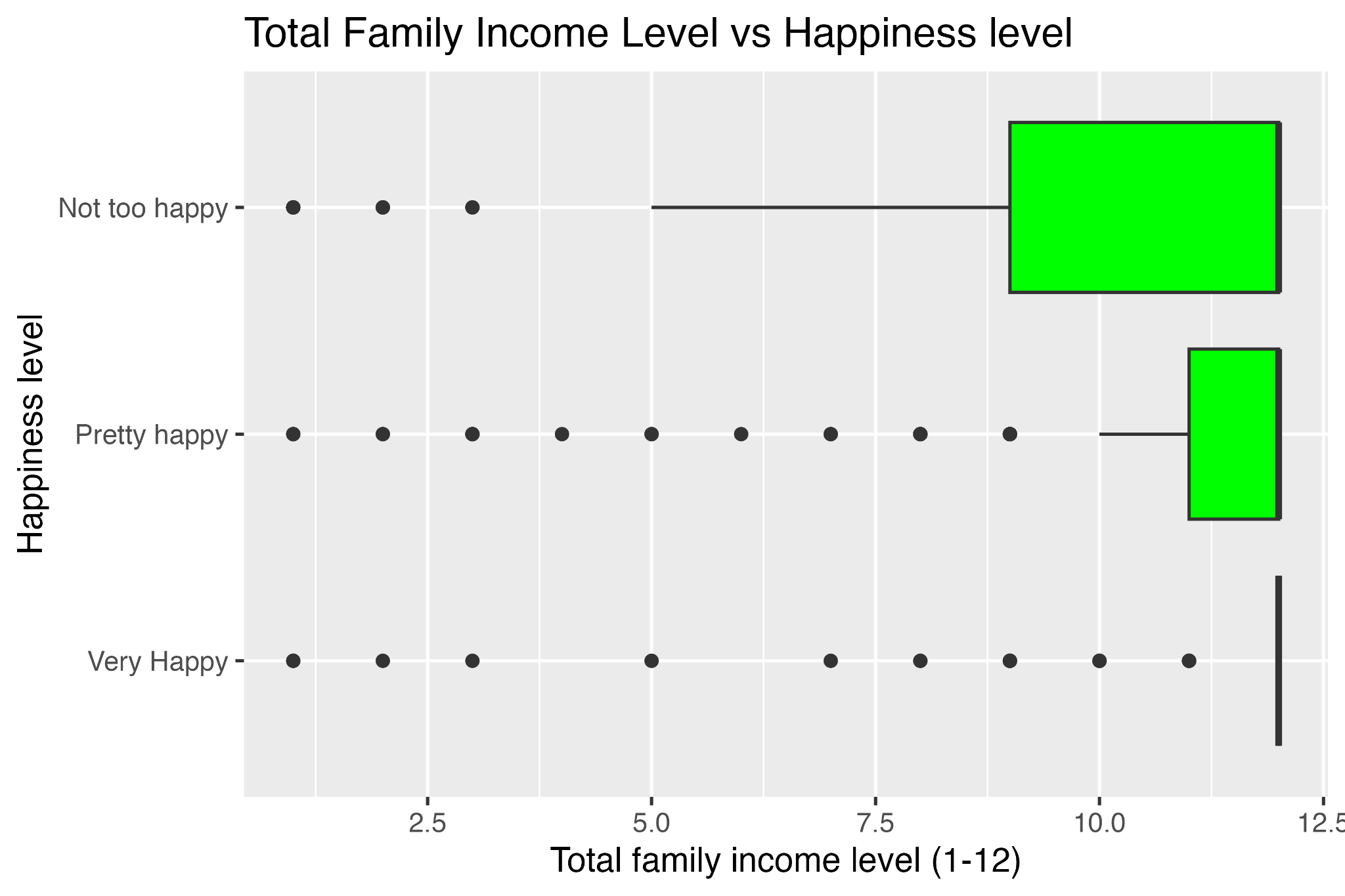
RQ2: Are there significant differences in income level for different perceptions of fairness?

Below are the ANOVA results and boxplots relating to research questions 2-4.

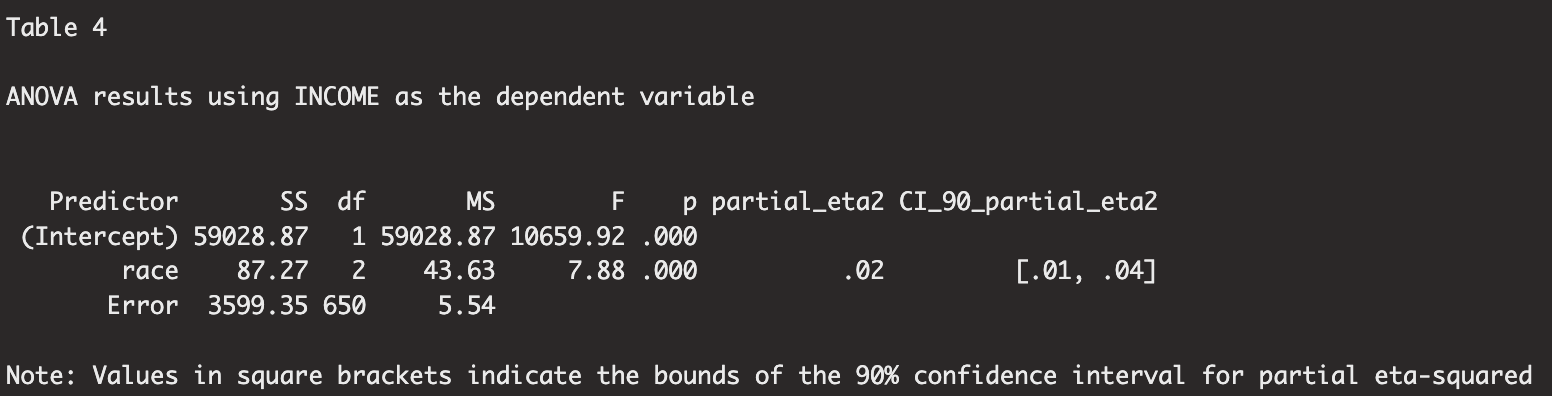
Yes, there are significant differences in income level for different perceptions of fairness (alpha=0.05). A Tukey’s HSD test indicates that the respondents who answered “Would try to be fair” vs those who answered“Would try to take advantage” significantly differed by around 0.658 in level. Having more money seems to relate to perceiving others as more fair.

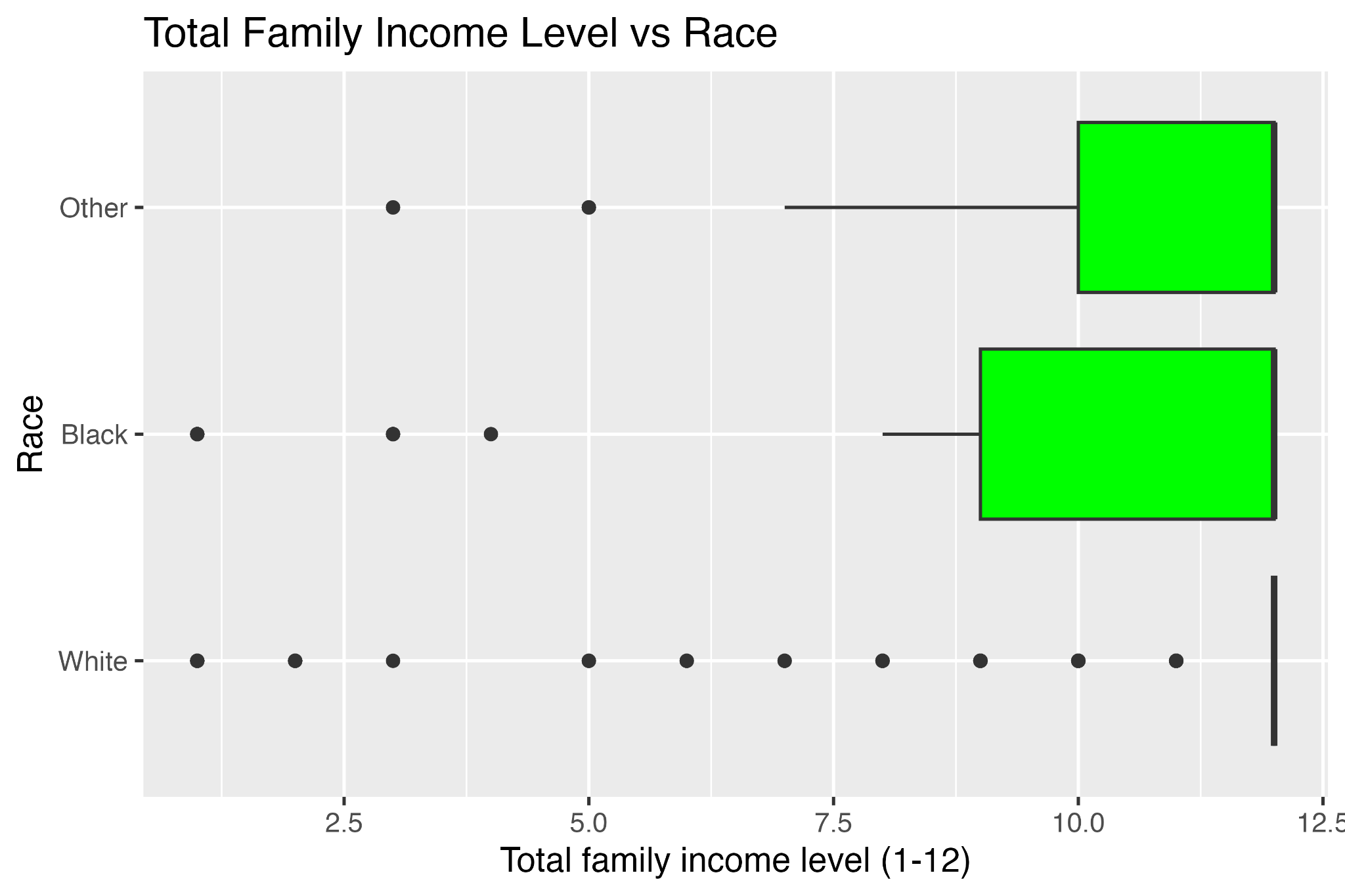


RQ3: Are there significant differences in income level for different happiness levels?There are significant differences in income level for different happiness levels(alpha=0.05). A Tukey’s HSD test indicates that the respondents who answered “Not too happy” vs those who answered “Would try to take advantage” significantly differed by around -1.45 in level. Additionally, respondents who answered “Not too happy” vs “Pretty happy’ differed around -1.25 in level. Maybe money does buy happiness!



RQ4:Are there significant differences in income level for different races?

There are significant differences in income level for different races (alpha=0.05). A Tukey’s HSD test indicates that the respondents who identified as Black differed from those who identified as white by -0.93. It should be noted that there were only 3 categories for race in this dataset: white, black, and other.



# Reflection

I have learned a lot of new techniques in Data Science. While I was able to review some R over the summer via datacamp, I did not get to explore more complicated data visualization tools or practice “good coding practices.” I think the most valuable “idea” I got out of this course was the idea of engineering my approach as well as good troubleshooting practices. This is still something I am trying to figure out (it took me 12 hours and arduous recoding to figure out my errors were emerging from my out-of-date R studio), but something that will be invaluable to my general PhD education. In terms of actual subject matter, I think the NLP and machine learning sections really helped to solidify the concepts of Elena’s project. While reading documentation and articles helped me to build base knowledge, actually being able to write my own code helped me to internalize what she was actually doing in her project. I’m interested to see how what I have learned in this section also applies to the Social Media project as well. After this class, I will definitely be trying to be as efficacious as possible with my debugging approaches, and will be using tidyverse for my data cleaning/manipulation practices!