# Predictors of Income

Michael Hazboun

Department of Psychology, University of Minnesota - Twin Cities

PSY 8012: Data Science in Psychology

Dr. Richard Landers

May 5th, 2024

# Predictors of Income

This project is meant to analyze the General Social Survey (GSS) 2018 data set, which can be found [here](https://gss.norc.org/get-the-data/spss). The analyses are primarily focused on the relationship between family income and various other variables, such as age and number of pets owned. Additionally, the data set is used in an attempt to predict income through the utilization of machine learning on a large number of the variables in the data set. Various machine learning models were utilized and a comparison document, of the models, is provided within the larger repository. A [shiny app](https://michaelhazboun.shinyapps.io/shiny_final/) is also created that can be utilized to learn more about our reduced model. A fully functional web-based binder, with the version of R and the packages utilized when the project was first made is also provided, and can be accessed [here](https://mybinder.org/v2/gh/MichaelHazboun/psy8712-final/HEAD?urlpath=rstudio).

## **Hypotheses and Research Questions**

Hypothesis 1:

Are income and age correlated? (H0: , H1:)

Hypothesis 2:

Is income different based on how scientific you believe astrology to be, and if yes, which levels show a significant difference? (given j is 1 to 3, one being “Very scientific”, two being “Sort of scientific”, and three being “Not at all scientific”: H0: , H1: .

Research Question 1: (this might technically be two research question)

How well can we predict family income when utilizing 4 variables to predict it (age, perceived standard of living compared to parents at the same age, the number of pets owned and how scientific you believe astrology to be), across three different machine learning algorithms/models (elastic net, random forest and eXtreme Gradient Boosting [LM was deemed too simplistic and thus was not utilized]) and how do those predictions compare to a model that utilized all of the variables in the data set with at most 75% missingness?

# Methods

### Open Science Materials

A binder was created for this project and can be accessed [here](https://mybinder.org/v2/gh/MichaelHazboun/psy8712-final/HEAD?urlpath=rstudio). A binder is a low-ability time-capsule of sorts. It is a re-creation of the specific version of R and the packages that have been utilized, with access to copies of all the files within a github repository, at the time the project was first created and ran. This is utilized to enable and ease the reproducibility process, and that is exactly why I have made one. It is so one can easily open up my binder and run my analyses and achieve near identical results to what I had found. It can be found by following the hyperlink in the project overview, or above.

Additionally, my github repository can be found [here](https://github.com/MichaelHazboun/psy8712-final) which contains instructions on how to run the project in the README.md file. The README.md file can be accessed by scrolling down on the home page of the repository or by clicking on the README.md file.

### Participants

The participants are 2,348 U.S. residents who were interviewed in 2018 for the General Social Survey.

### Measures

Family Income, coded as INCOME, is a dichotomized measure of family income that asks people to select which income group their family fell in before taxes. (1: under 1k, 2: 1k to 2.999k, 3: 3k to 3.999k, 4: 4k to 4.999k, 5: 5k to 5.999k, 6: 6k to 6.999k, 7: 7k to 7.999k, 8: 8k to 9.999k, 9: 10k to 14.999k, 10: 15k to 19.999k, 11: 20k to 24.999k and 15: more than 25k.)

Perceived standard of living compared to parents at the same age, coded as PARSOL, is a dichotomous variable in which people state if they believe their standard of living is “much better” (1) to “much worse” (5) than that of their parents when they were the same age

How scientific you believe astrology to be, coded as ASTROSCI, is a dichotomous variable in which people state if they believe astrology is “very scientific” (1), “sort of scientific” (2), or “not at all scientific” (3).

Age, coded as AGE, is a continuous variable stating the respondents age.

Number of pets, coded as NUMPETS, is a continuous variable of how many pets the family owns

A large number of other variables were utilized in answering the research question, details of those can be found in the GSS codebook, found within the docs folder on the github repository.

### Procedure

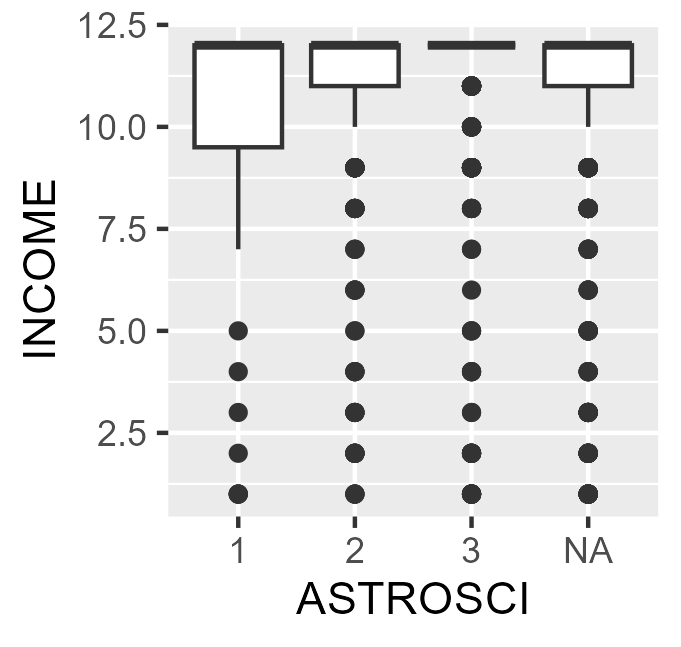
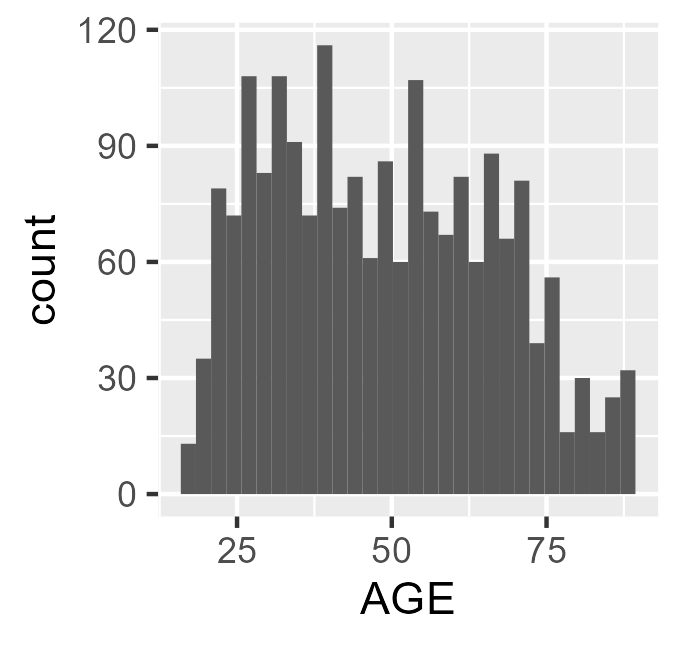
This data set was created by interviewing a relatively large sample of the US population during the year 2018 for an average interview time of roughly one hour and a half.

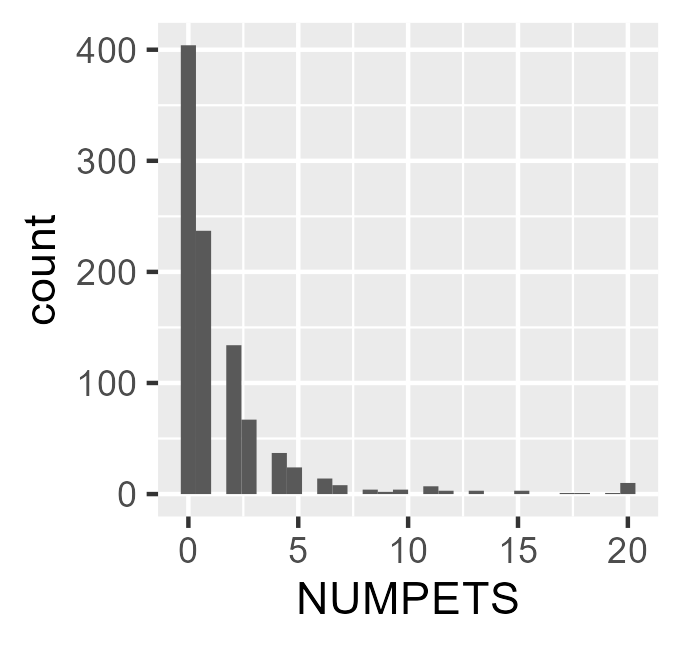
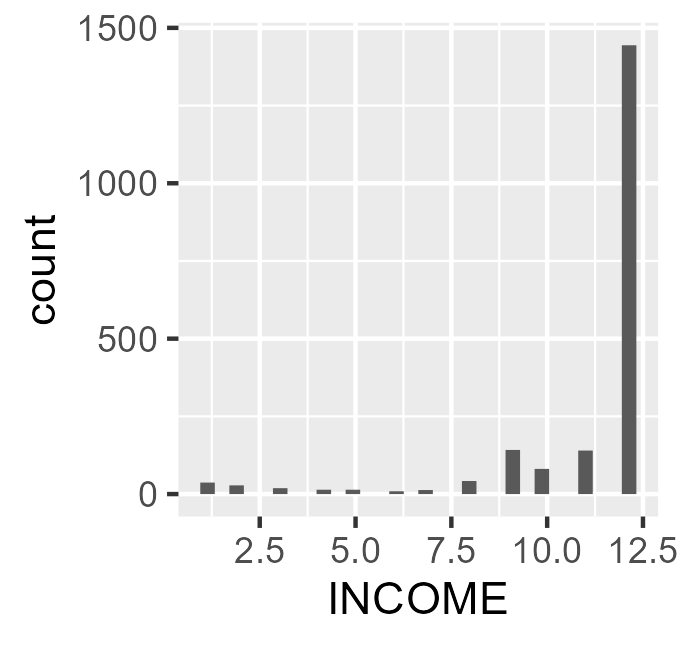
# Analyses

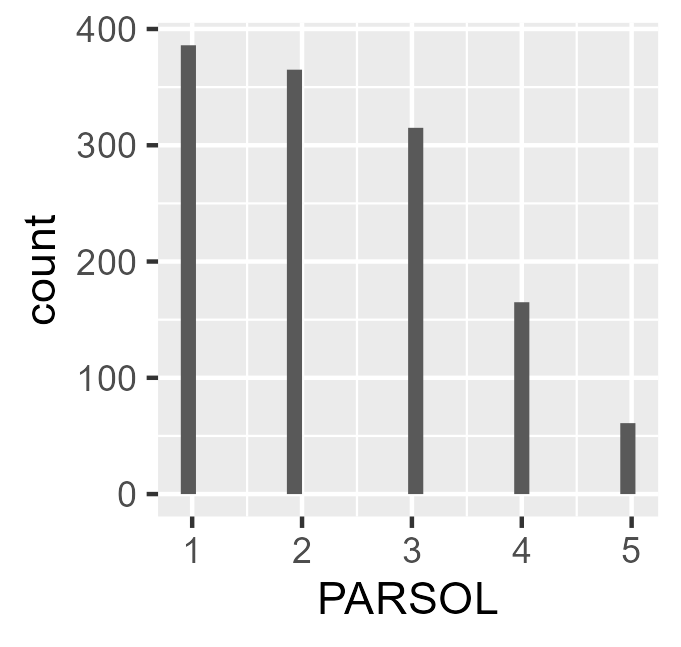
### Descriptive Statistics and Static Visualizations

| Variable | Mode | Mean | SD |
| --- | --- | --- | --- |
| INCOME | 12 | 10.94 | 2.38 |
| AGE | 34 | 48.60 | 18.06 |
| NUMPETS | 0 | 1.71 | 3.03 |
| PARSOL | 1 | 2.34 | 1.16 |

| Option | N |
| --- | --- |
| 1 | 70 |
| 2 | 335 |
| 3 | 573 |
| NAs | 1005 |







### Interactive Visualization

This project also comes with a shinyapp, the app is meant to allow one to better visualize the relationship between any two variables of choice from the reduced model. The app also provides the correlation between each of the selected variables, other than the combination of NUMPETS and ASTROSCI because there wasn’t any data between the two variables that overlapped. The available option is a tool that allows the control of the size of the points on the scatter plot. The next two options allow the specification of the x and y values of the scatter plot. The fourth option allows the removal of individuals who are under the age of 25. The final option allows the alteration of the points color.

### **Data** **Cleaning**

For the reduced model, the cleaning process started with the selection of all of the desired variables and the removal of the remaining variables. That was then followed by the removal of NA values from the INCOME variable, this was done to allow the aov function and the machine learning models to work properly. All of the variables were then made numeric to allow for proper processing. The ASTROSCI variable was turned into a factor because it was a factor, previously everything was turned numeric first to reduce the total number of lines written. The full model also went through minor cleaning, by first removing variables with more than 75% missingness, removing the other income variables, removing NA from our desired INCOME variable and turning everything into a numeric variable.

### Analysis

#### **H1:** Are income and age correlated? (H0: , H1:)

The correlation between age and family income was 0.02 and the p value was 0.32 so the correlation was not statistically significant. (I reformatted this to normal text but I did copy and paste it from R)

#### H2: Is income different based on how scientific you believe astrology to be, and if yes, which levels show a significant difference? (given j is 1 to 3, one being “Very scientific”, two being “Sort of scientific”, and three being “Not at all scientific”: H0: , H1: .

The p value from the anova was 0.0029, so there was a statistically significant difference across groups. The groups that had a statistically significant difference between their incomes were 3 vs 1 and 3 vs 2, the Tukey output can be found below. (I reformatted this to normal text but I did copy and paste it from R)

| Comparison | diff | lwr | upr | p.adj |
| --- | --- | --- | --- | --- |
| Two-One | 0.39 | -0.36 | 1.14 | 0.44 |
| Three-One | 0.83 | 0.11 | 1.54 | 0.02 |
| Three-Two | 0.44 | 0.04 | 0.83 | 0.02 |

#### Research Question 1: (this might technically be two research question): How well can we predict family income when utilizing 4 variables to predict it (age, perceived standard of living compared to parents at the same age, the number of pets owned and how scientific you believe astrology to be), across three different machine learning algorithms/models (elastic net, random forest and eXtreme Gradient Boosting [LM was deemed too simplistic and thus was not utilized]) and how do those predictions compare to a model that utilized all of the variables in the data set with at most 75% missingness?

When only utilizing 4 variables we do quite a poor job predicting family income (both \_reduced columns, cv\_rqs is the k-fold validation r squared while ho\_rqs with the holdout validation r squared), with honestly quite terrible holdout r squares. However, utilizing the full model (\_full) we are able to predict family income quite well, with nearly perfect prediction with the k-fold validation using the xgboost model, and random forest models. Those two models also found incredibly large r squares with the holdout validation too (~0.95 and ~0.94). So we can predict income quite well by utilizing the full model with either the random forest or eXtreme gradient boosting models/algorithms.

| algo | cv\_rqs\_reduced | ho\_rqs\_reduced | cv\_rqs\_full | ho\_rqs\_full |
| --- | --- | --- | --- | --- |
| elastic net | 0.0183 | 0.02144 | 0.6637 | 0.1800 |
| random forest | 0.3853 | 0.0141 | 0.9850 | 0.9472 |
| xgboost | 0.4862 | 0.0066 | 0.9998 | 0.9434 |

# Reflection

I have yet to actually do any coding for my research up to this point and I don’t think I will for a bit of a while at this point, so I can’t tell what I’ll be taking back other than everything. However, I definitely have benefited a lot from this class simply through my improved understanding of R, the functions and how to code R. For example, this class has both amplified what I learned in 8814, but also made learning the code in 8815 much easier; troubleshooting was much more manageable, writing the code was much easier and everything simply looked cleaner. I honestly feel like that was the best part of the class, simply gaining a better understanding of how R works is most probably the most valuable aspect of this class. But one other thing that I will forever utilize after this class is checking documentation and the help files, simply they are god sent.