

A photograph of two people wearing blue overalls and gloves, working in a field of green plants. The background is blurred, showing more of the field and some yellow and blue structures.

The Best Things In Life Are Free, Including Volunteers

By: Jhaelle Payne

Table of Contents

●	The Setting
●	Research Question
●	Dataset Description
●	About the Response Variable
●	Feature Engineering
●	Exploratory Data Analysis
●	Feature Selection
●	Principal Component Analysis
●	Decision Tree with Hyperparameters
●	Results
●	Accuracy
●	Conclusion

Let's get started!

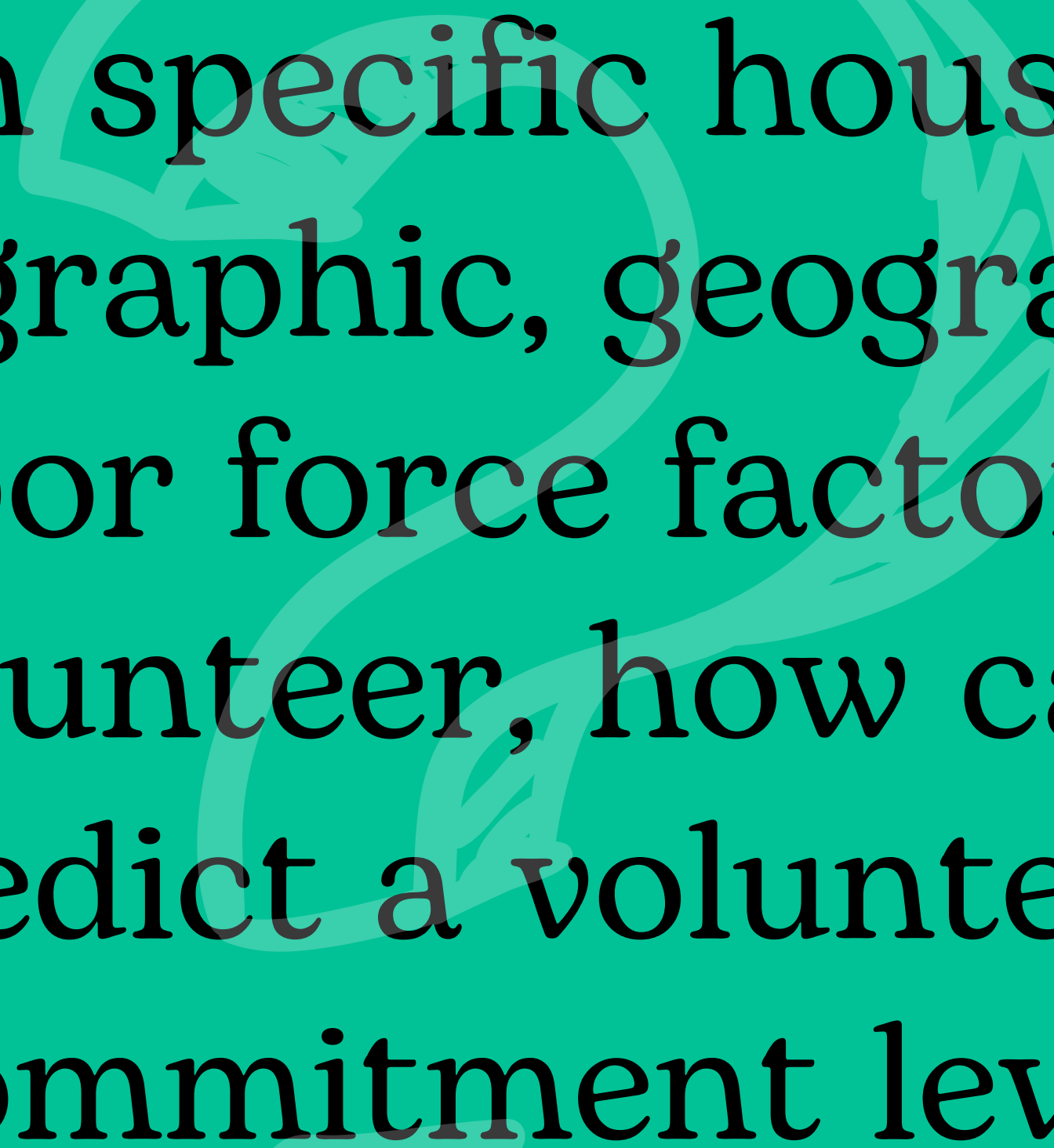
The Setting



My best friend walked into my apartment and slumped over the couch with a sigh. She told me the four words that shocked me to my core, "I want to quit." She was referring to her job as a nurse assistant at a nonprofit. But, for the longest, I have only known her to want her career to surround something that gives back to a community. She absolutely loves nonprofits.

As I made her a cup of green tea, she explained that her job was draining her. She had to work two other jobs to make ends meet because their stipend was not covering enough. As much as she loved working at a nonprofit, she felt her cup was always half full. Empty by the end of the day.

That caused my mind to wonder. Can we target specific volunteers who are more at risk of feeling burnt out? Nonprofit organizations can identify these volunteers and offer some extra support. Now for our specific question ->



Given specific household,
demographic, geographical,
and labor force factors about
a volunteer, how can we
predict a volunteer's
commitment level?

CURRENT POPULATION SURVEY, September 2019
Volunteering and Civic Life Supplement

TECHNICAL DOCUMENTATION
CPS—19

This file documentation consists of the following materials:

Attachment 1	Abstract
Attachment 2	Overview - Current Population Survey
Attachment 3	Overview – September 2019 Volunteering and Civic Life Supplement
Attachment 4	Glossary
Attachment 5	How to Use the Record Layout
Attachment 6	Basic CPS Record Layout
Attachment 7	Current Population Survey, September 2019 Volunteering and Civic Life Supplement Record Layout
Attachment 8	Current Population Survey, September 2019 Volunteering and Civic Life Supplement Questionnaire
Attachment 9	Industry Classification Codes
Attachment 10	Occupation Classification Codes
Attachment 11	Specific Metropolitan Identifiers
Attachment 12	Topcoding of Usual Hourly Earnings
Attachment 13	Tallies of Unweighted Counts
Attachment 14	Countries and Areas of the World
Attachment 15	Allocation Flags
Attachment 16	Source and Accuracy of the September 2019 Volunteering and Civic Life Supplement Data
Attachment 17	User Notes

NOTE

Questions about accompanying **documentation** should be directed to Center for New Media and Promotions Division, Promotions Branch, Bureau of the Census, Washington, D.C. 20233. Phone: (301) 763-4400.

Questions about the **subject matter** should be directed to the CPS Branch, U.S. Census Bureau, Washington, D.C. Phone: (301) 763-3806, or dsd.cps@census.gov.

Dataset Description

To answer this question, I used the 2019 population survey with volunteering questions ("CURRENT POPULATION SURVEY, September 2019 Volunteering and Civic Life Supplement") from the NYC open data website. This dataset consisted of 139217 rows and 419 columns. This dataset consisted of household, geographic, demographic, labor force, and lifestyle data from randomly selected household members 16 years or older. For my predictors, I only chose 44 variables that best represent the predictors I wanted. Due to time complexity, I could not use all 419 variables. As my response variable, I chose "How Often Did You Volunteer?" Let's dive deeper into the response variable first.

Response Variable:

How often did [you/[NAME]] volunteer?

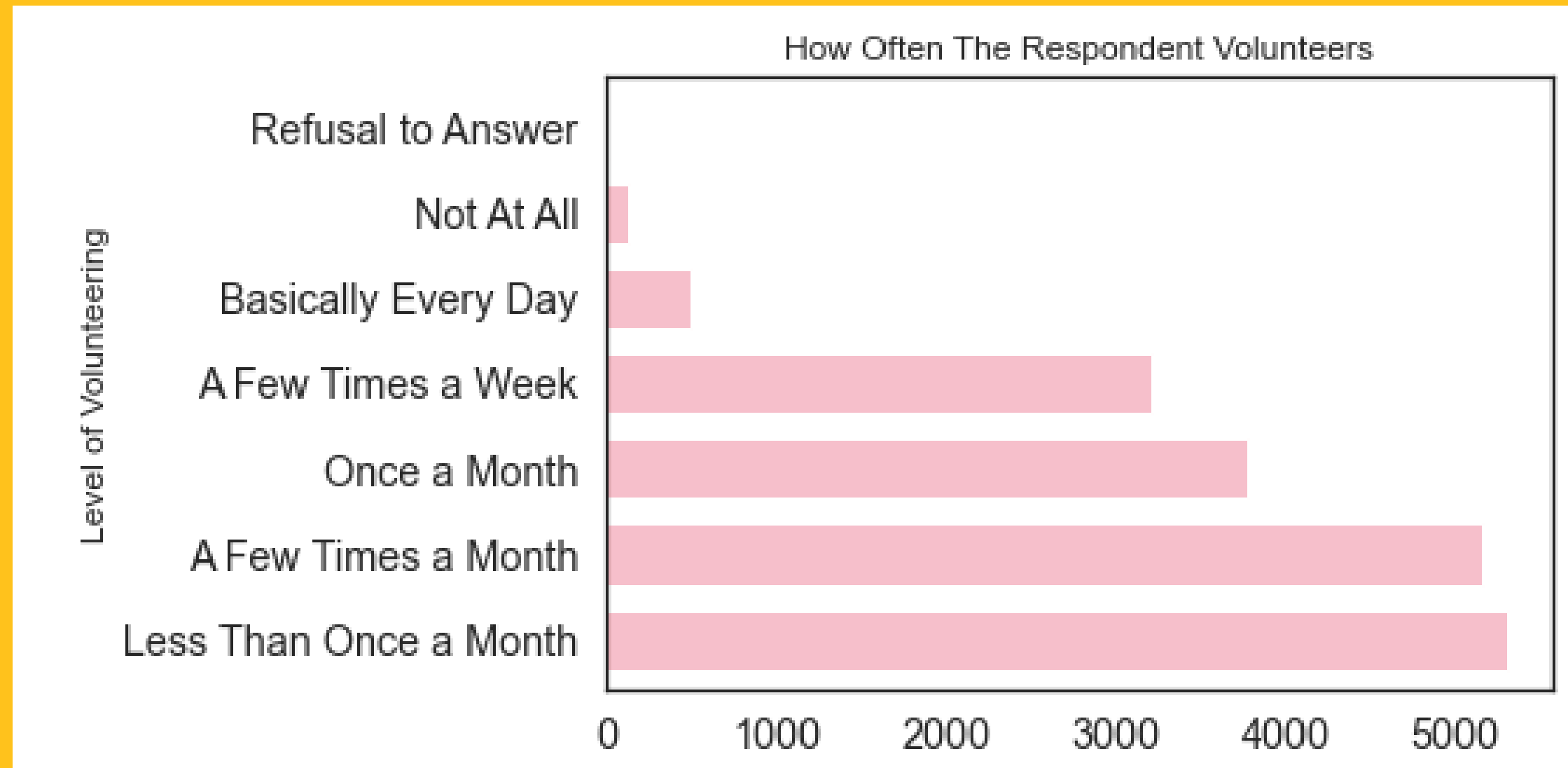
Possible Responses:

- 9 No answer
- 3 Refusal
- 2 Do not know
- 1 Not in universe
- 1 = Basically every day
- 2 = A few times a week
- 3 = A few times a month
- 4 = Once a month
- 5 = Less than once a month
- 6 = Not at all

Notes:

- The dataset only allowed numerical responses
- I removed those rows with -9, -2, or -1 in the response column, as I was not interested in predicting them.
- The only "non-response" answer I kept was "Refusal" because I thought it would be interesting to see what type of person refused to answer. However, I only kept -3 for my Exploratory Data Analysis and removed it when predicting

Distribution of Response Variable After Removal of "No Answer", "Do Not Know" and "Not In Universe"



Takeaways:

- My dataset is imbalanced, meaning some levels of volunteering are more represented than others. I will have to take this into account when building my models.
- It is clear that most people volunteer to some extent.
- Barely anyone refused to answer this question.
- Most respondents volunteer less than once a month. I interpreted this response as the respondent volunteers every few months and sometimes never during other months.

Feature Engineering: Introduction

Due to the size of my dataset, I implemented four feature engineering methods that would minimize my dataset to features with the most considerable variability on my response variable.

i.e., I only kept features that had the most significant influence on the outcome of my response variable.

I performed the removal of null values, discretization, variable transformation, and scaling for these reasons:

Irrelevant Features	This dataset had many variables that placed weight on some responses for their personal report. These variables were unrelated to my purposes, as well as redundant variables.
Time Complexity	With many features, training our dataset would have taken too long, and we always want to be as efficient as possible.
Explainability	My main concern was losing my model's explainability and the inability to use specific models. Too many features make it hard to explain your final model, and particular models can not be used with null values.
Data-Model Compatibility	Some models I used did not allow null values or unscaled datasets. Therefore, I removed null values and scaled my features before I proceeded.
Overfitting	We want to avoid overfitting at all costs. With overfitting, there is only a slight difference between our actual and predicted values (low bias). However, we can not implement our model on new data (high variance). Data with more features are more prone to overfitting than data with fewer features.

Feature Engineering: Methods

Null Values

There were null values in the dataset; however, this was only in one column labeled "FILLER." In other columns, if respondents did not respond, there was a value for that (-9). I left those in because it would be interesting to see how no answers in the features affected the response variable. Therefore, I took a subset of the relevant columns to my question—mainly the "non Filler" columns.

Discretization

There were multiple values in my response variable that I did not deem necessary to answer my question. This includes -9 (No answer), -2 (Do not know), and -1 (Not in universe). I do not know what "Not in universe" means for the respondent. I am only interested in people who have volunteered and their occurrence.

Variable Transformations

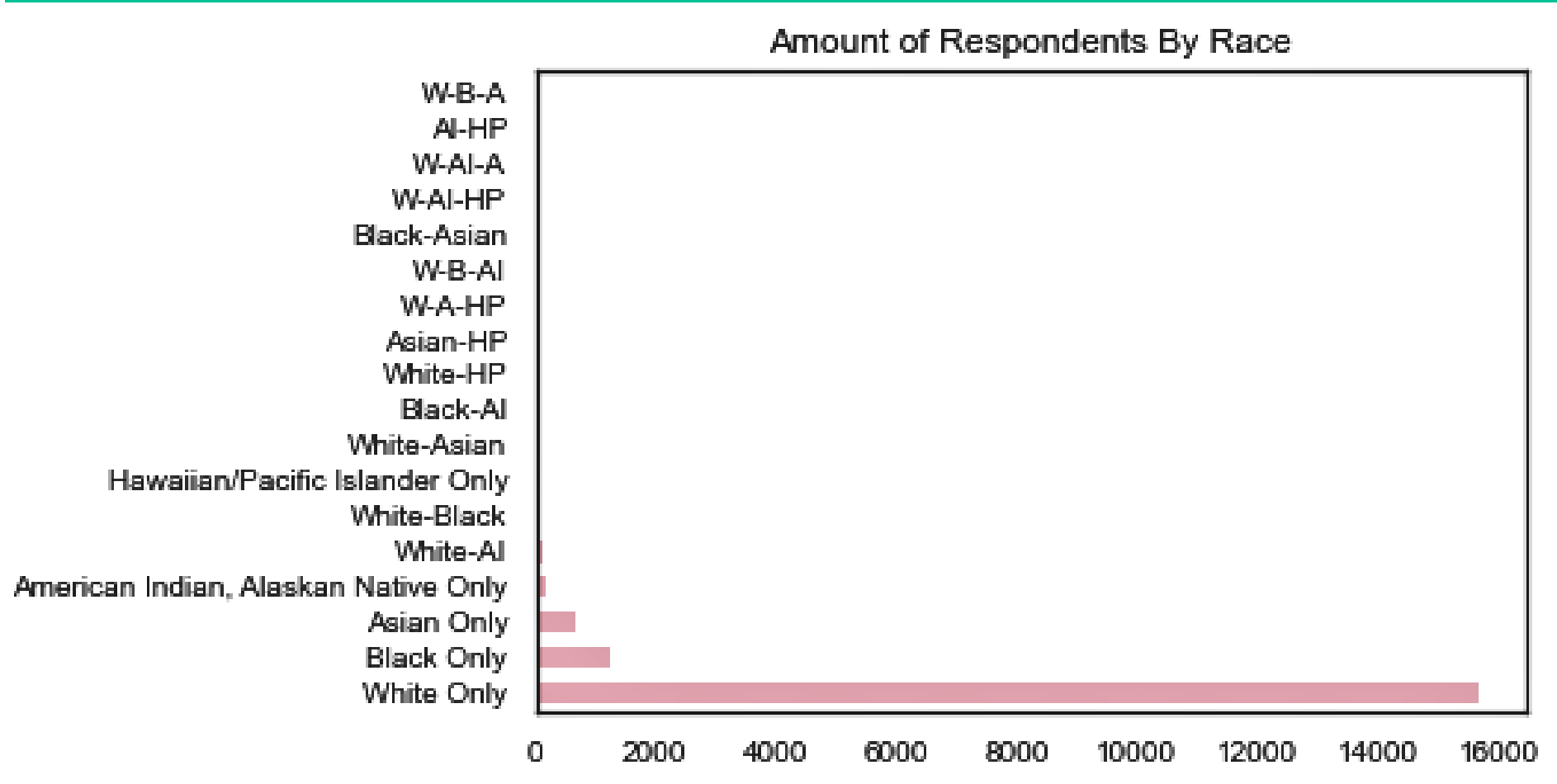
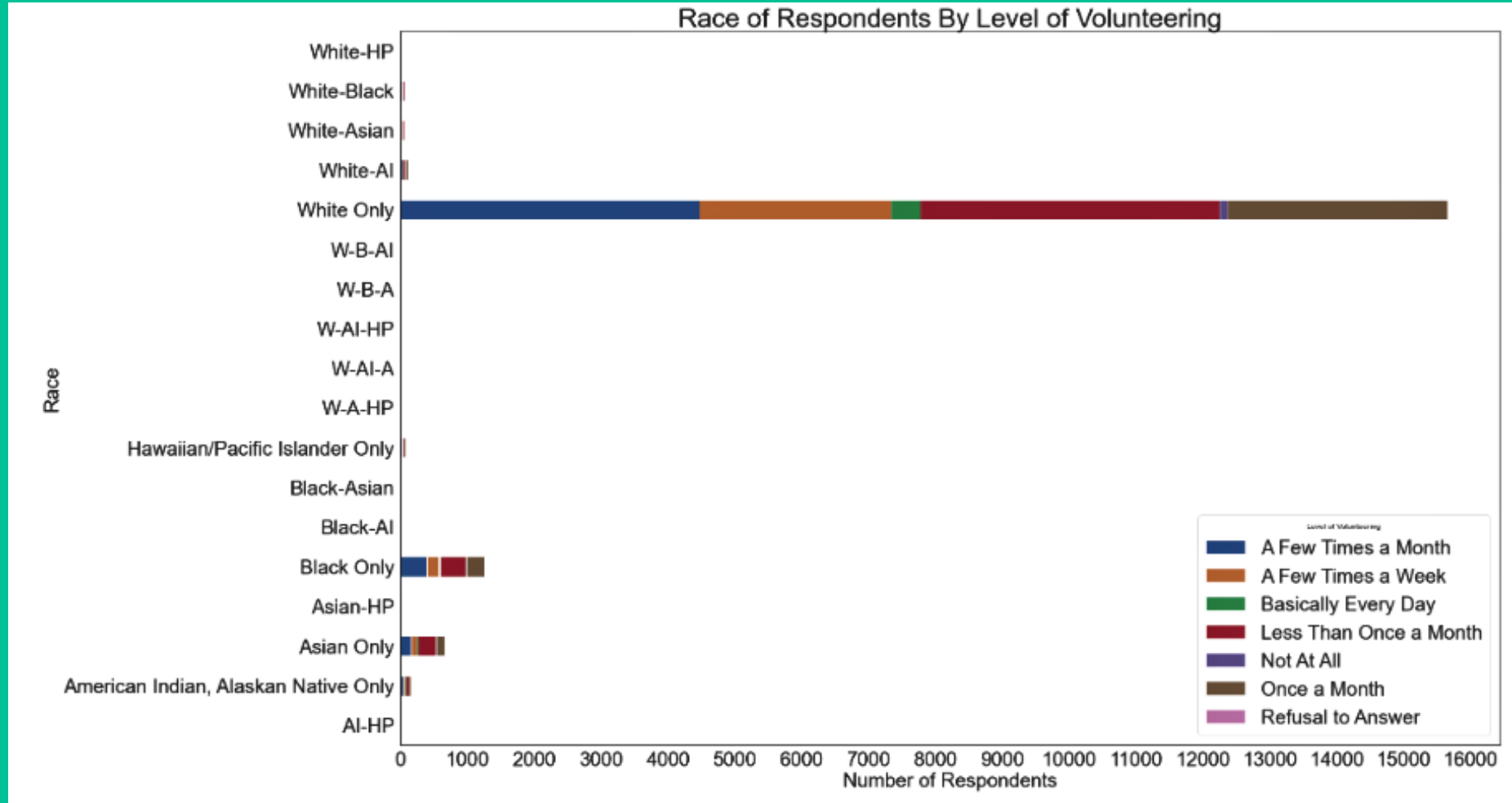
I transformed my variables to understand my feature's variability on my response variable. I used Principal Component Analysis to complete this for easy data exploration, to understand essential variables, and to spot outliers.

Scaling

I only scaled my features, not the response variable, to prepare my dataset for various machine-learning techniques. The goal was not to allow features with a larger range to take importance over features with a smaller range. I used Min-Max Scaling, which rescales all values in a feature so it falls between 0 and 1, with 0 replacing the minimum value and 1 replacing the highest value.

Now, we're ready to get our hands dirty with some Exploratory Data Analysis (EDA)! But before we do, let's get a better understanding of our response variable and who these respondents are.

EDA of Race



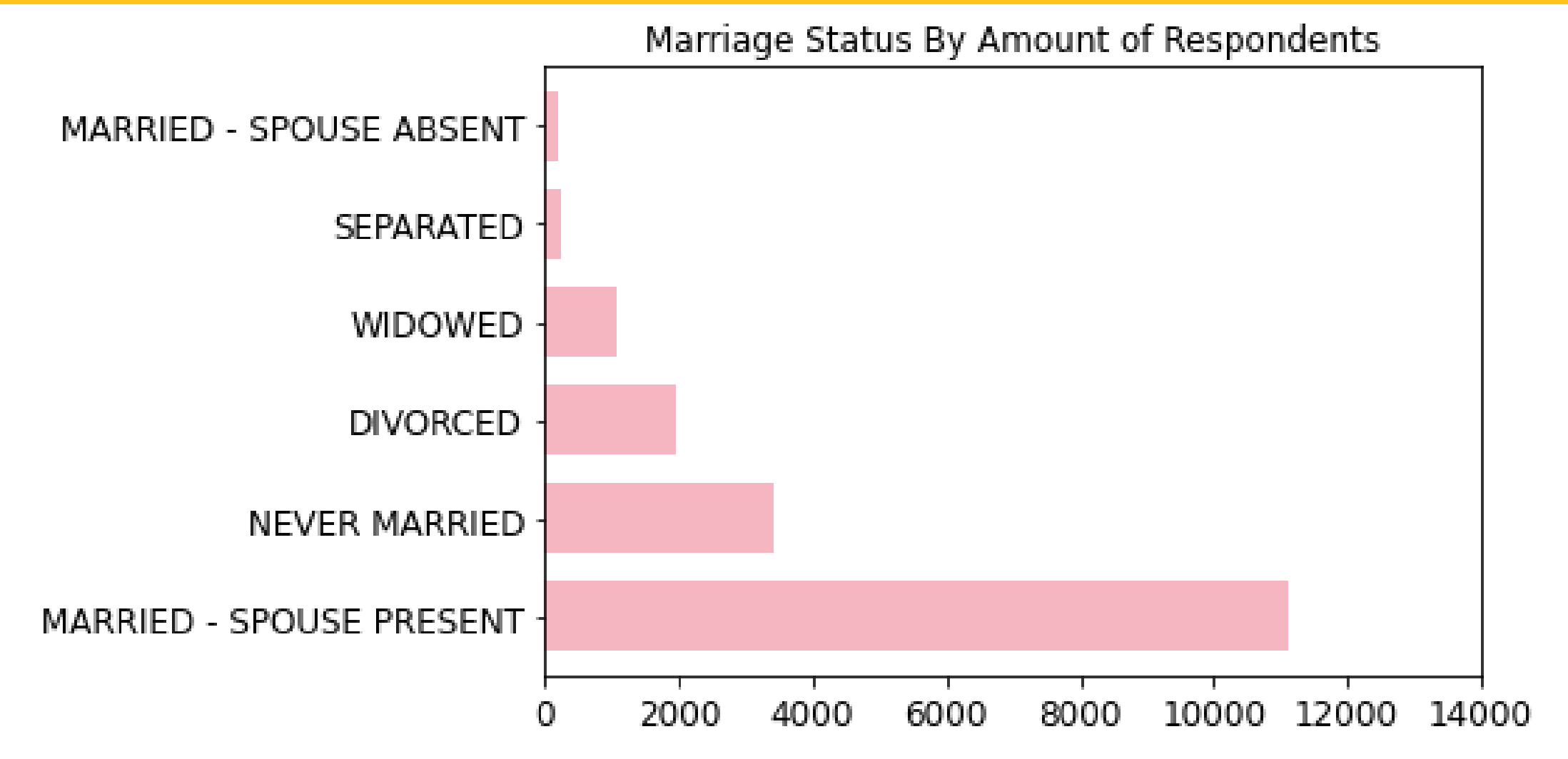
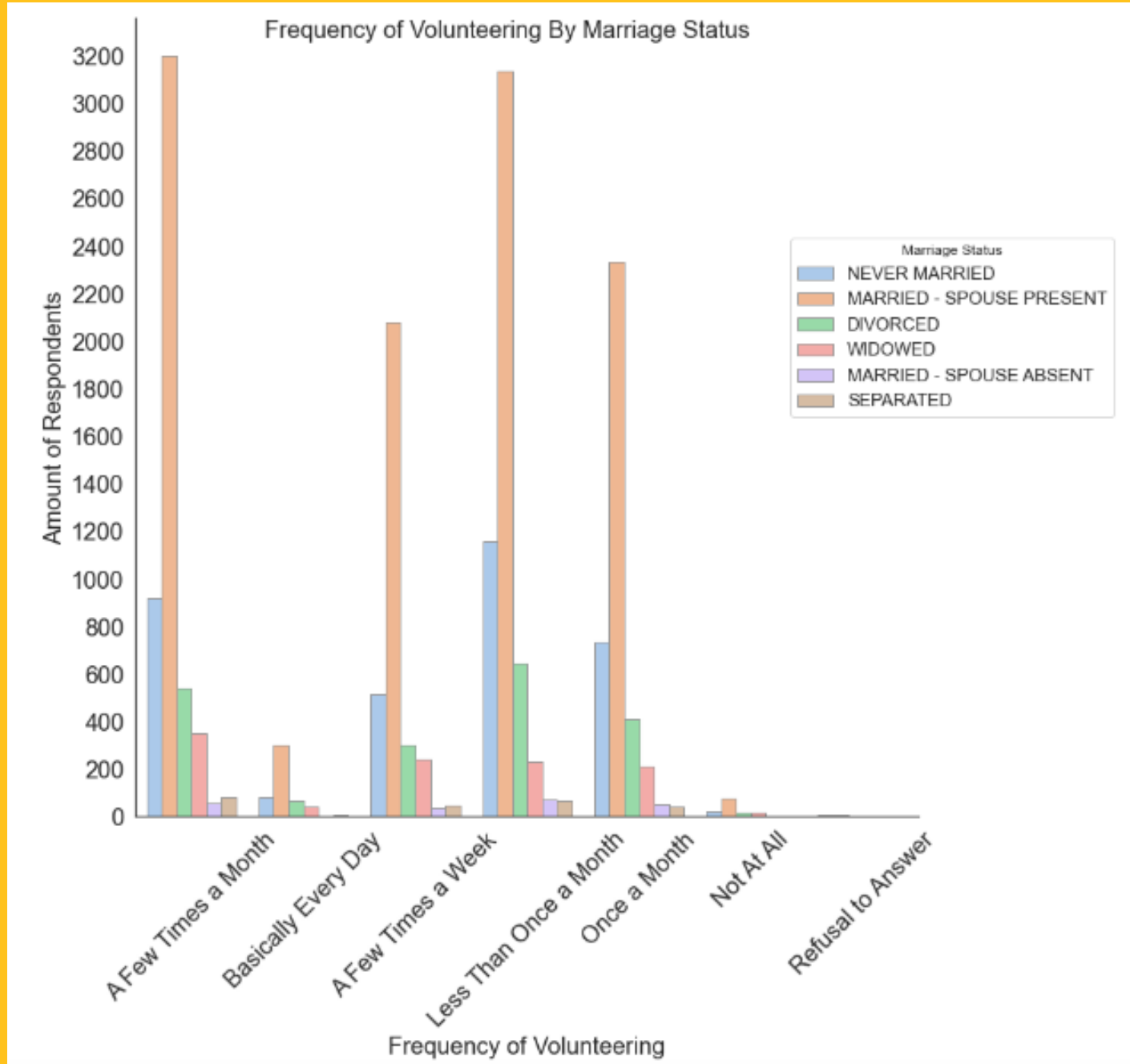
Takeaways:

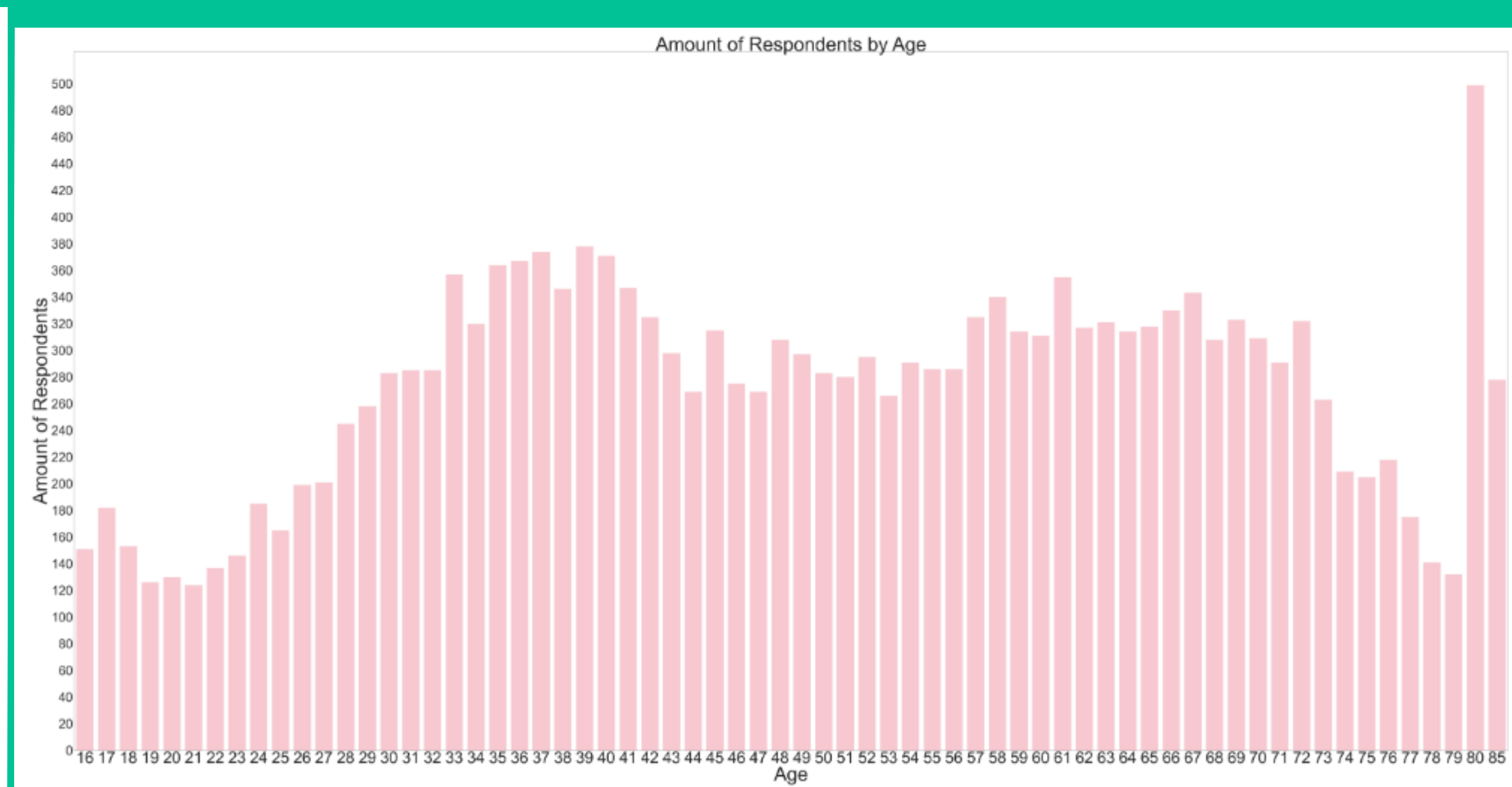
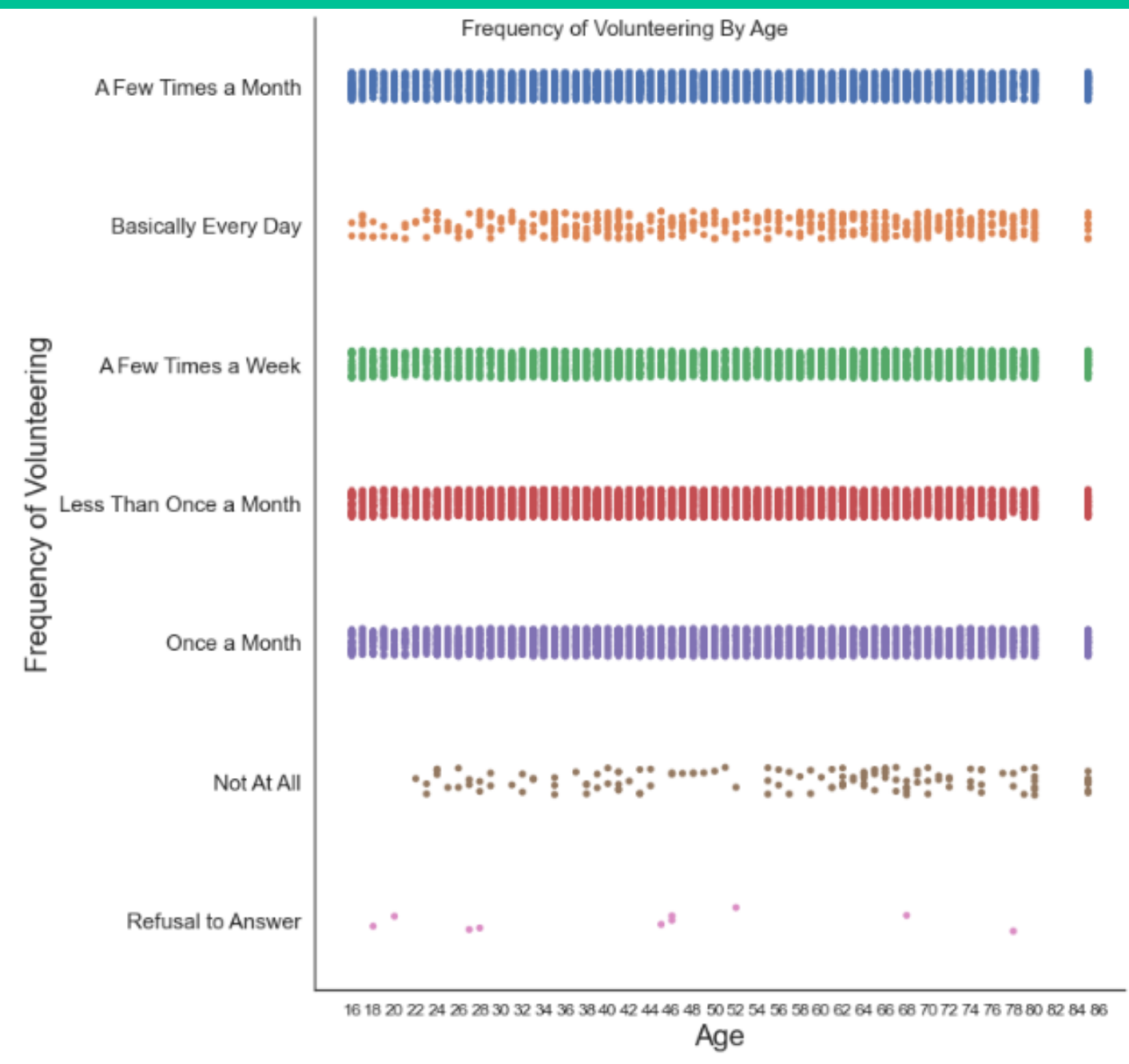
- Respondents who identified as White had the most significant representation in this study, while respondents who identified as multiracial had the lowest. However, according to the U.S. Census, this is proportional to the U.S. population.
- Most respondents who identified as White or Black volunteered less than one month to a few times a month.
- Most respondents who identified as Asian volunteered for less than one month.
- Most respondents who identified as American Indian, Alaskan Native volunteered less than one month to a few times a week.

EDA of Marriage Status

Takeaways:

- Respondents who were married with a present spouse had the most significant representation in this study, while respondents who were married with an absent spouse had the lowest.
- Respondents who were primarily married volunteered a few times a month. However, respondents that were married with an absent spouse or were separated volunteered but rarely more than a few times a week.





EDA of Age

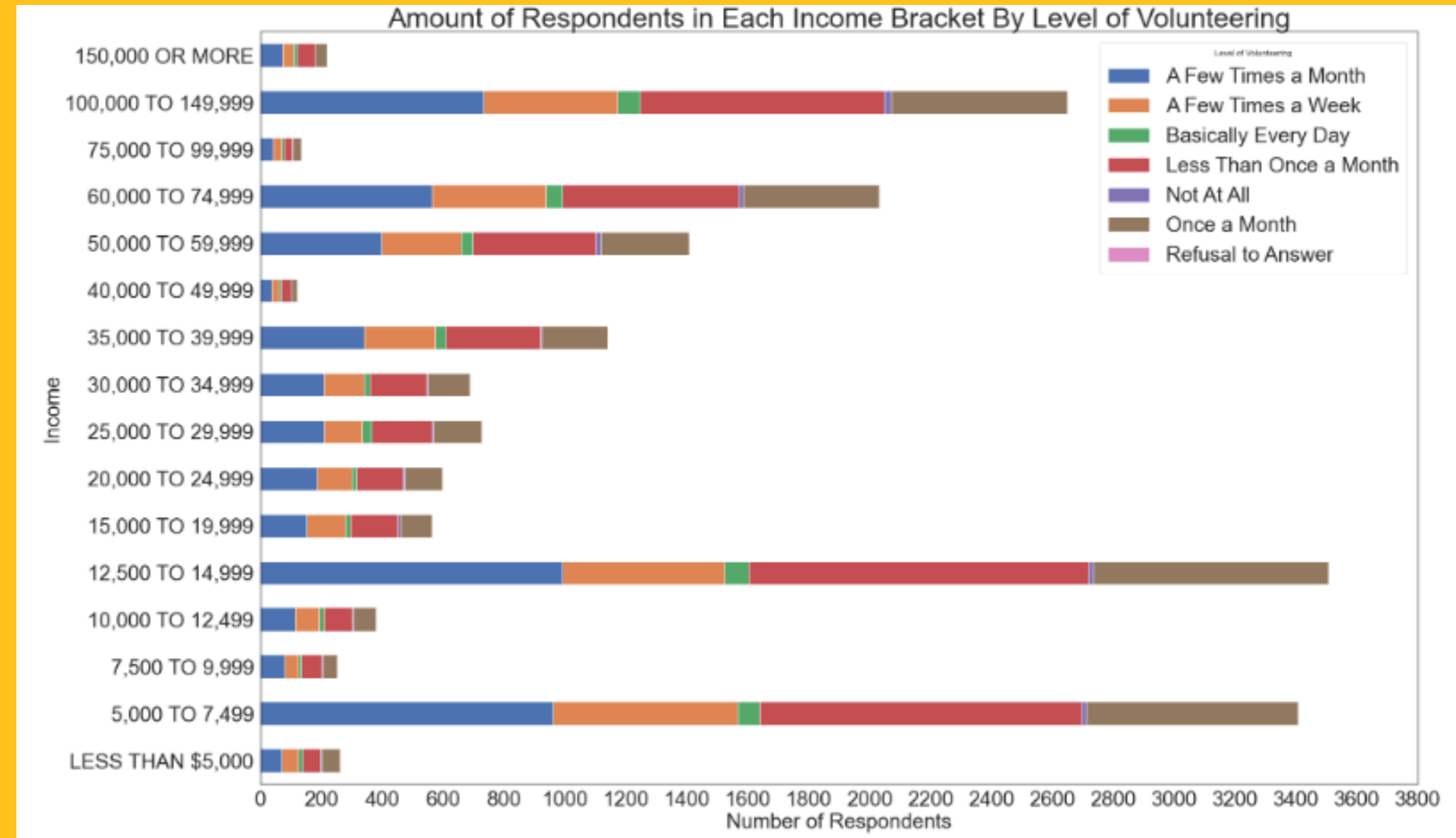
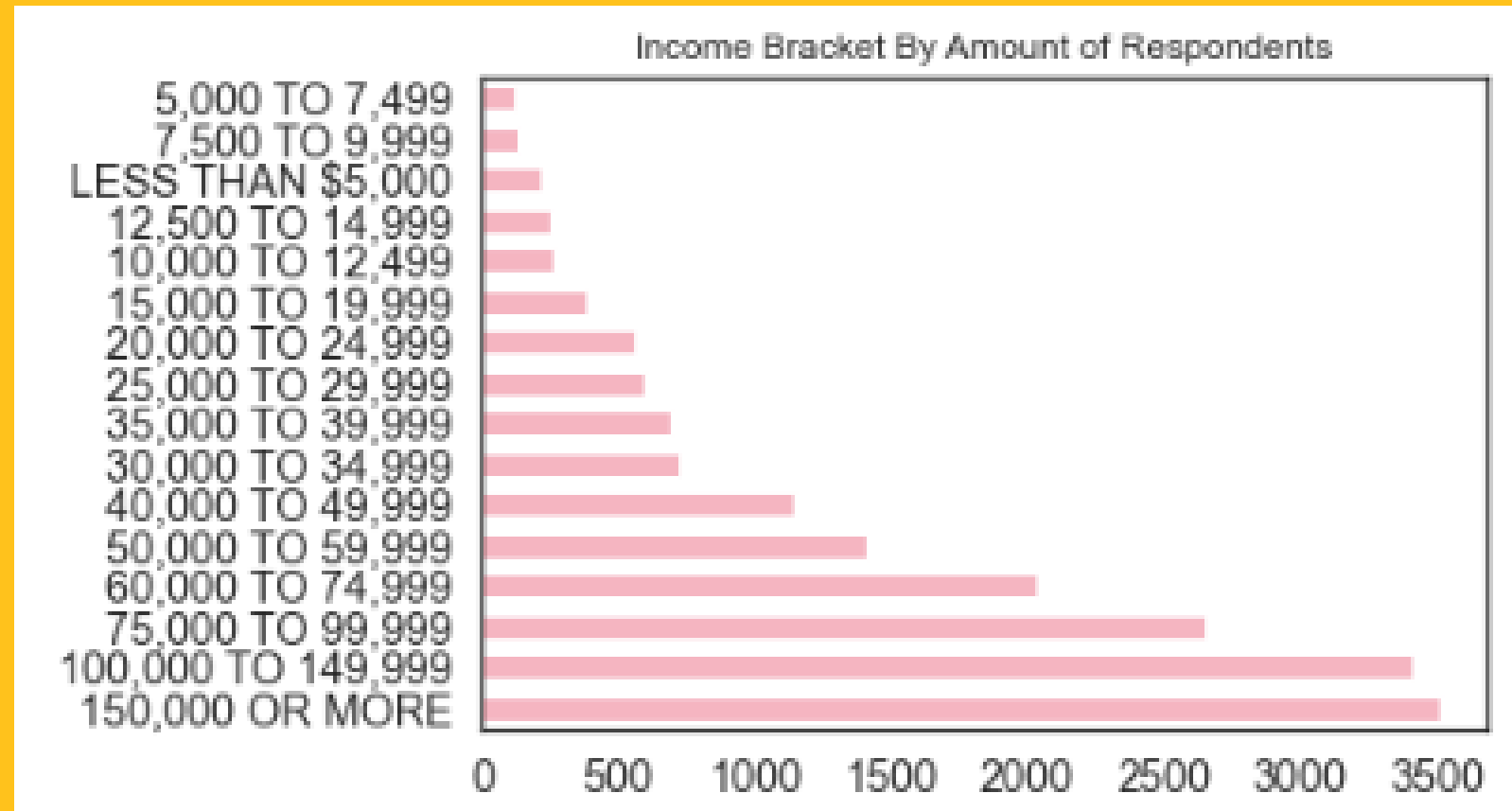
Takeaways:

- To Note: According to the Technical Documentation, 81-84 and 86+ were not used. Age 80 = 80-84, Age 85 = 85+.
- Respondents in their 30s, 60s, and 80+ had the most significant representation in this study, while respondents under 30 and in their late 70s had the lowest.
- Although a small amount, more respondents do not volunteer at all the older they become.

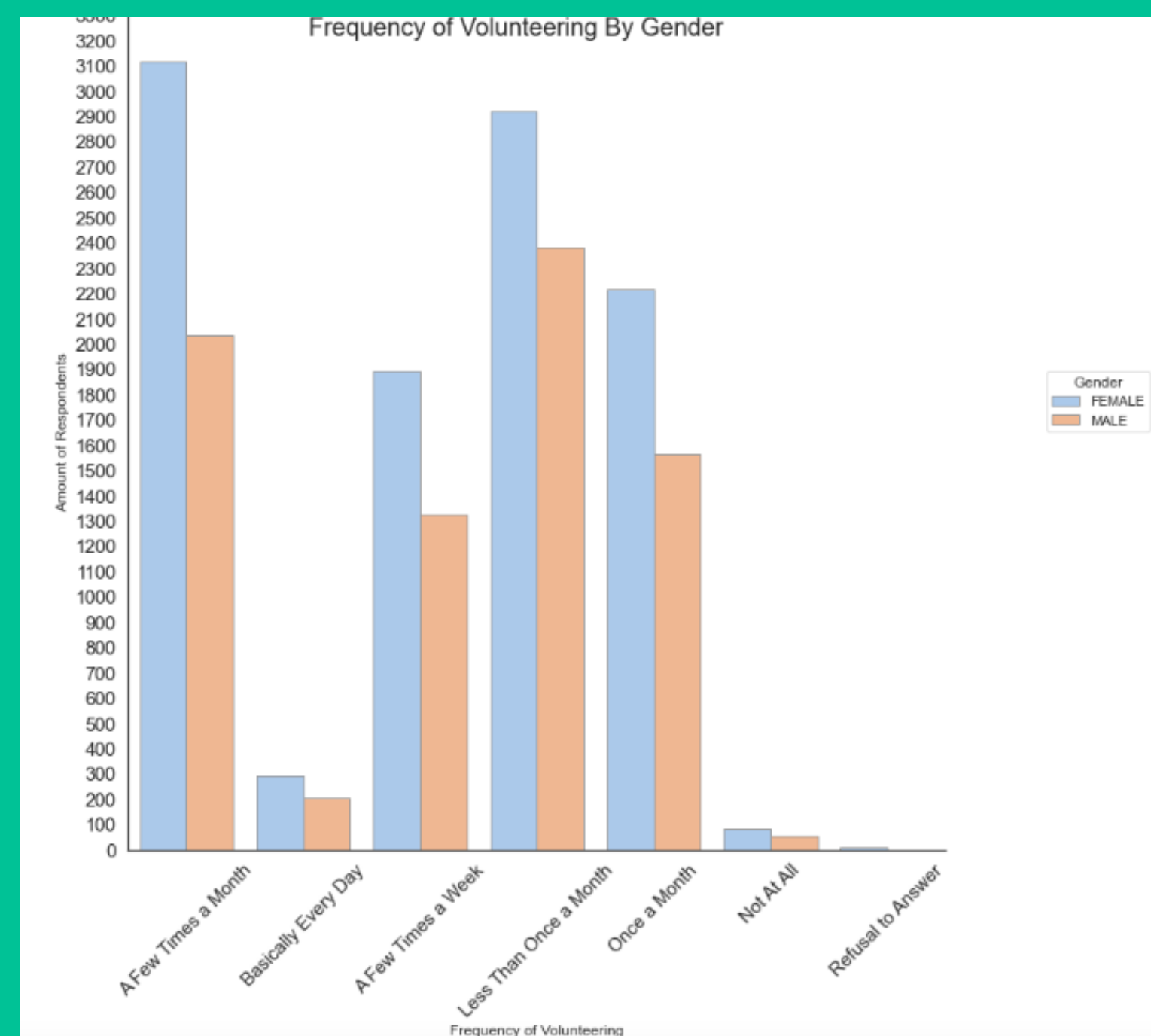
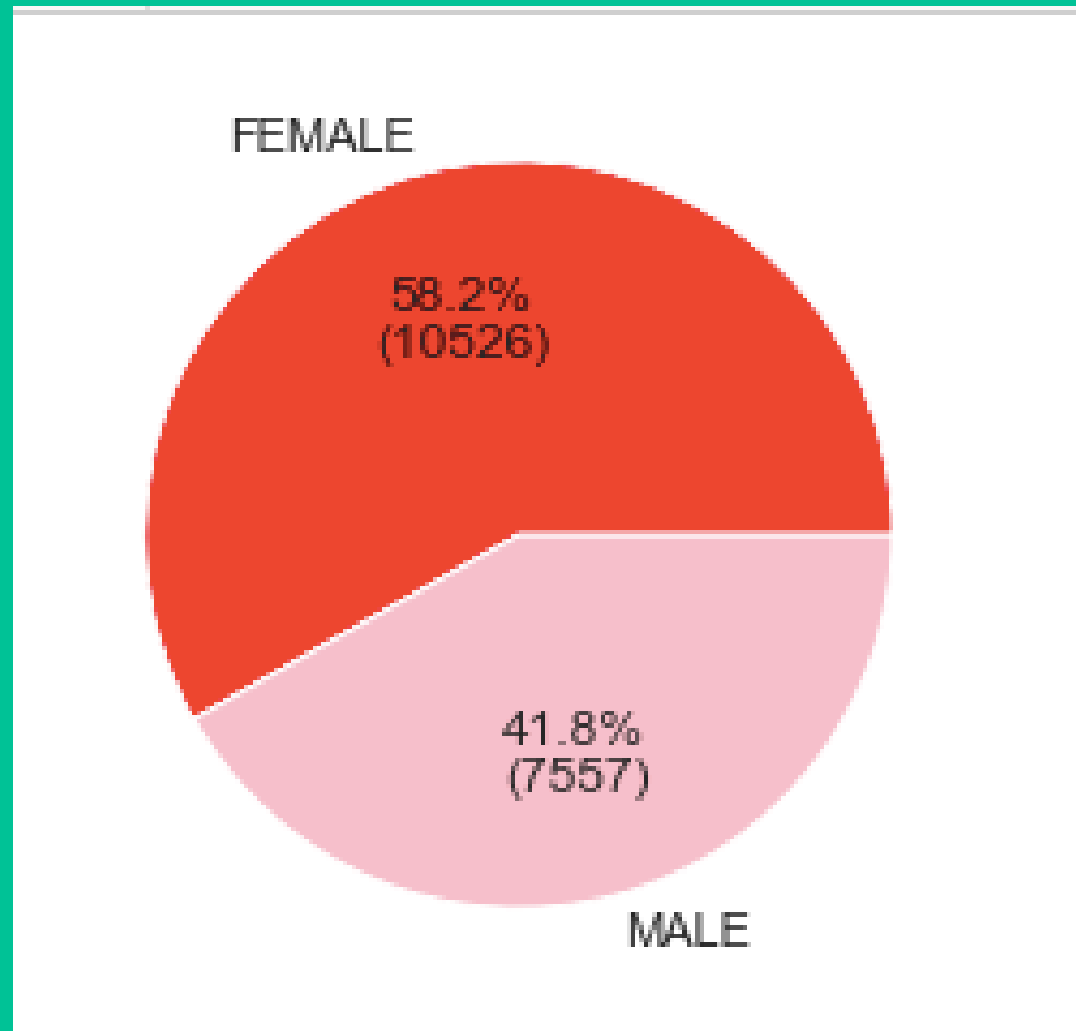
EDA of Family Income Bracket

Takeaways:

- Respondents whose household earns more than \$150,000 a year had the largest representation in this study, while respondents whose household earned \$5,000-\$7,499 had the lowest.
- For most respondents that do not volunteer at all, their households earned between \$100,000-\$149,999
- For most respondents that volunteer basically everyday, their households earned between \$12,500-\$14,999



EDA of Gender



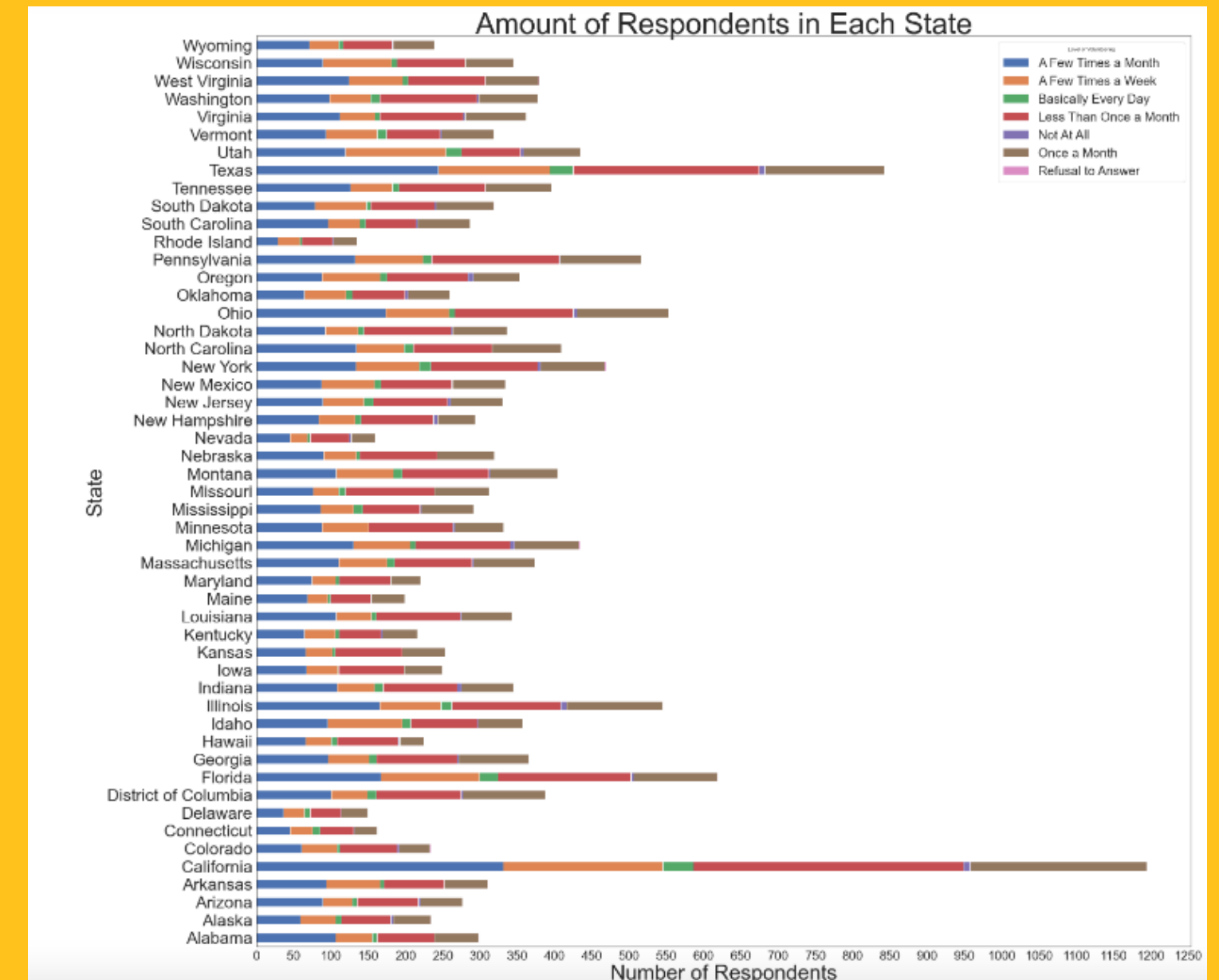
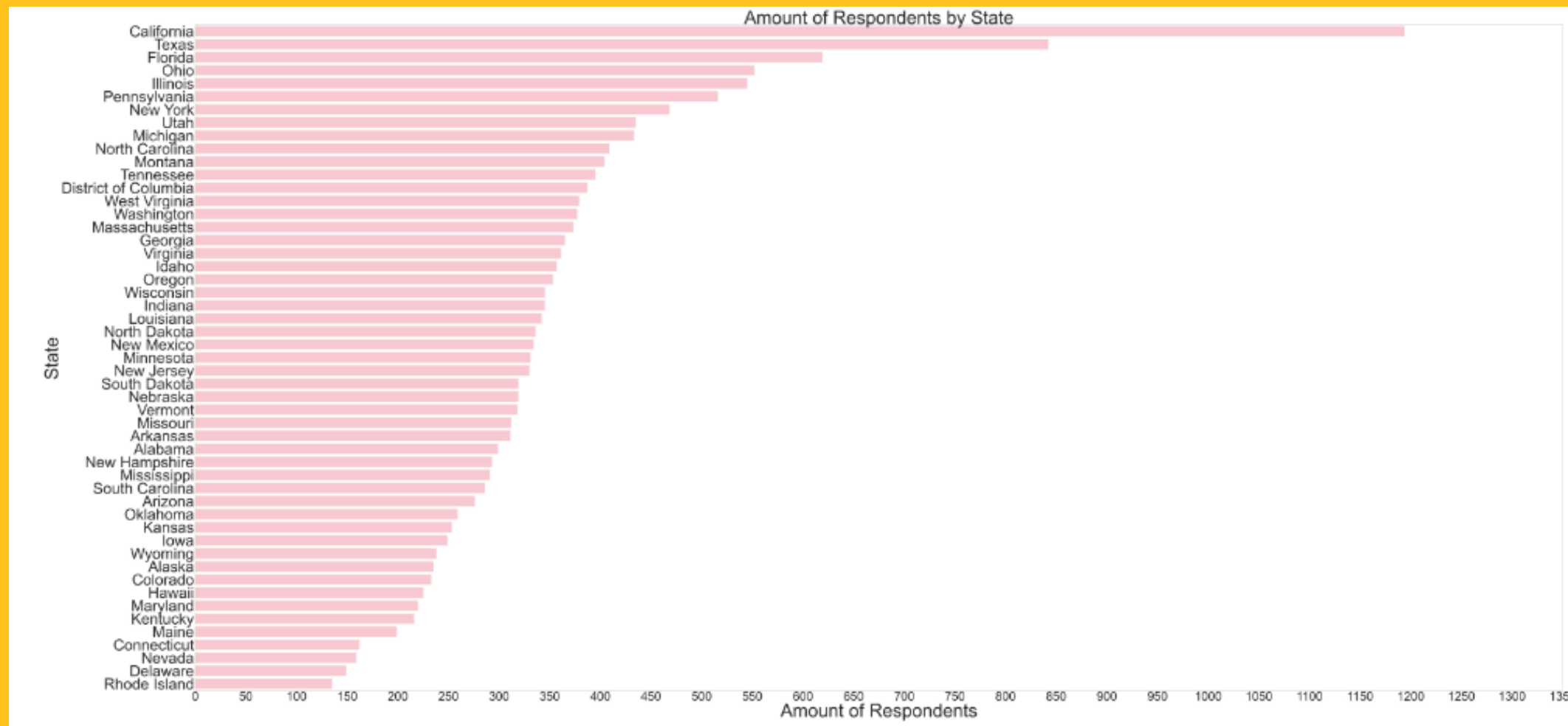
Takeaways:

- Respondents who identified as Female had the largest representation in this study, while respondents who identified as Male had the lowest.
- Most respondents who identified as Female volunteered a few times a month
- Most respondents who identified as Male volunteered less than a month

EDA of State

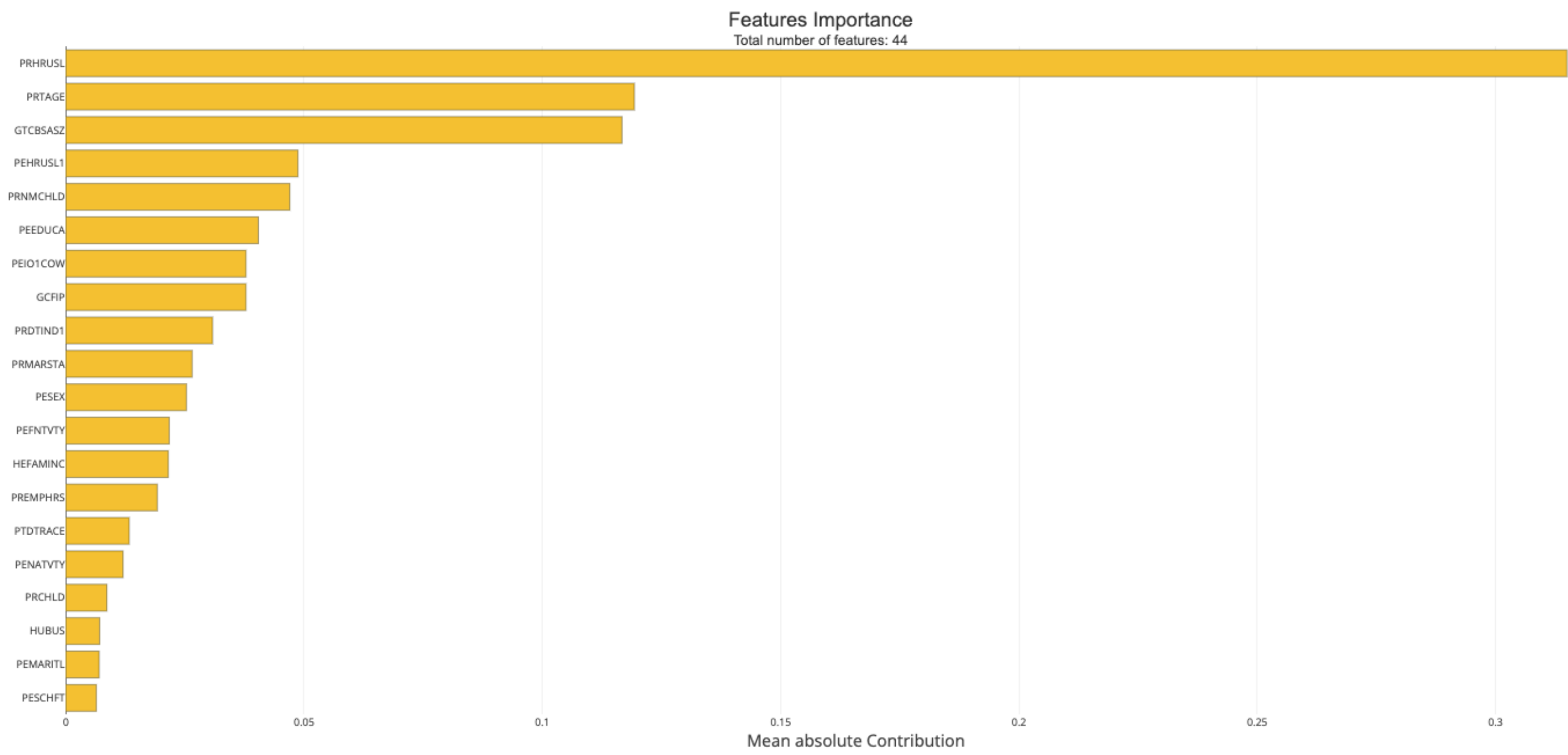
Takeaways:

- Respondents who lived in California had the most significant representation in this study, while respondents in Rhode Island and Delaware had the lowest.
- Most respondents who lived in California and Rhode Island volunteered for less than one month.
- Most respondents who lived in New York (I'm biased and wanted to make a note of NY!) volunteered less than one month to a few times a month.



Feature Selection

I wanted to narrow down my features from 44 to 15. There was not an exact method I used to choose 15, however, I wanted to avoid overfitting and be able to explain my model better. Using the Shapash package, I got the top 20 most important features. These were:



Codes:

- PRHRUSL: USUAL HOURS WORKED WEEKLY
- PRTAGE: PERSONS AGE
- GTCBSASZ: Metropolitan Area (CBSA) SIZE
- PEHRUSL1: HOW MANY HOURS PER WEEK DO YOU USUALLY WORK AT YOUR MAIN JOB?
- PRNMCHLD: Number of own children <18 years of age
- PEEDUCA: HIGHEST LEVEL OF SCHOOL COMPLETED OR DEGREE RECEIVED
- PEIO1COW: INDIVIDUAL CLASS OF WORKER CODE ON FIRST JOB (Type of Job)
- GCFIP:FEDERAL INFORMATION PROCESSING STANDARDS (FIPS) STATE CODE
- PRDTIND1: DETAILED INDUSTRY RECODE - JOB 1
- PRMARSTA: MARITAL STATUS BASED ON ARMED FORCES PARTICIPATION
- PESEX: SEX
- PEFNTVTY: FATHER'S COUNTRY OF BIRTH
- HEFAMINC: FAMILY INCOME
- PREMPHRS: REASON NOT AT WORK OR HOURS AT WORK
- PTDTRACE: RACE
- PENATVTY: COUNTRY OF BIRTH
- PRCHLD: PRESENCE OF OWN CHILDREN <18 YEARS OF AGE BY SELECTED AGE GROUP
- HUBUS: DOES ANYONE IN THIS HOUSEHOLD HAVE A BUSINESS OR A FARM?
- PEMARITL: MARITAL STATUS
- PESCHFT: ARE YOU ENROLLED IN SCHOOL AS A FULL-TIME OR PART-TIME STUDENT?

Now, let's see how well these newly selected features can explain our response variable!

Principal Component Analysis (PCA)

What is this?

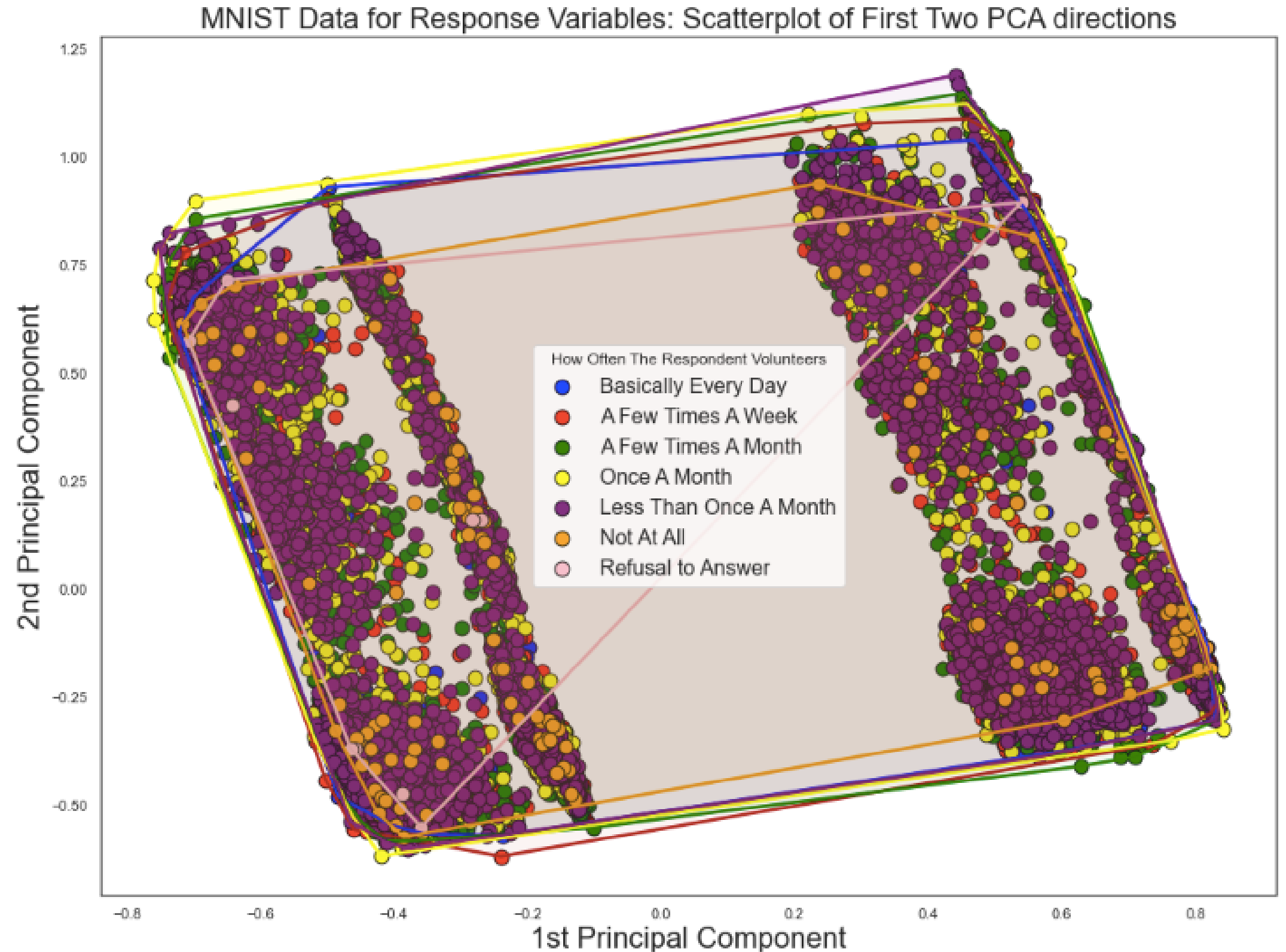
PCA transformed our columns into new features called Principal Components (PC). Each PC has information on the amount of variance it contains. We want to represent all of our features in the smallest amount of columns possible. However, this does not make it a feature selection technique.

Why did you use this?

I wanted to understand better how much our features influence our response variable.

What was the result?

The first 2 PCs could explain the majority of our data's variability. According to the visual on the right, very few points belonging to the same category are distinctly clustered and region bound. For the most part, they overlap. Unfortunately, this may mean that the data captured in the first two PCs must be more informative to discriminate the categories. This may also mean that our data can not explain our response variable, lessening our chance of a clear pattern. Although PCA did not go as planned, we will see if K-Nearest Neighbors can point out a clear pattern.



Decision Tree with Hyperparameters

What is this?

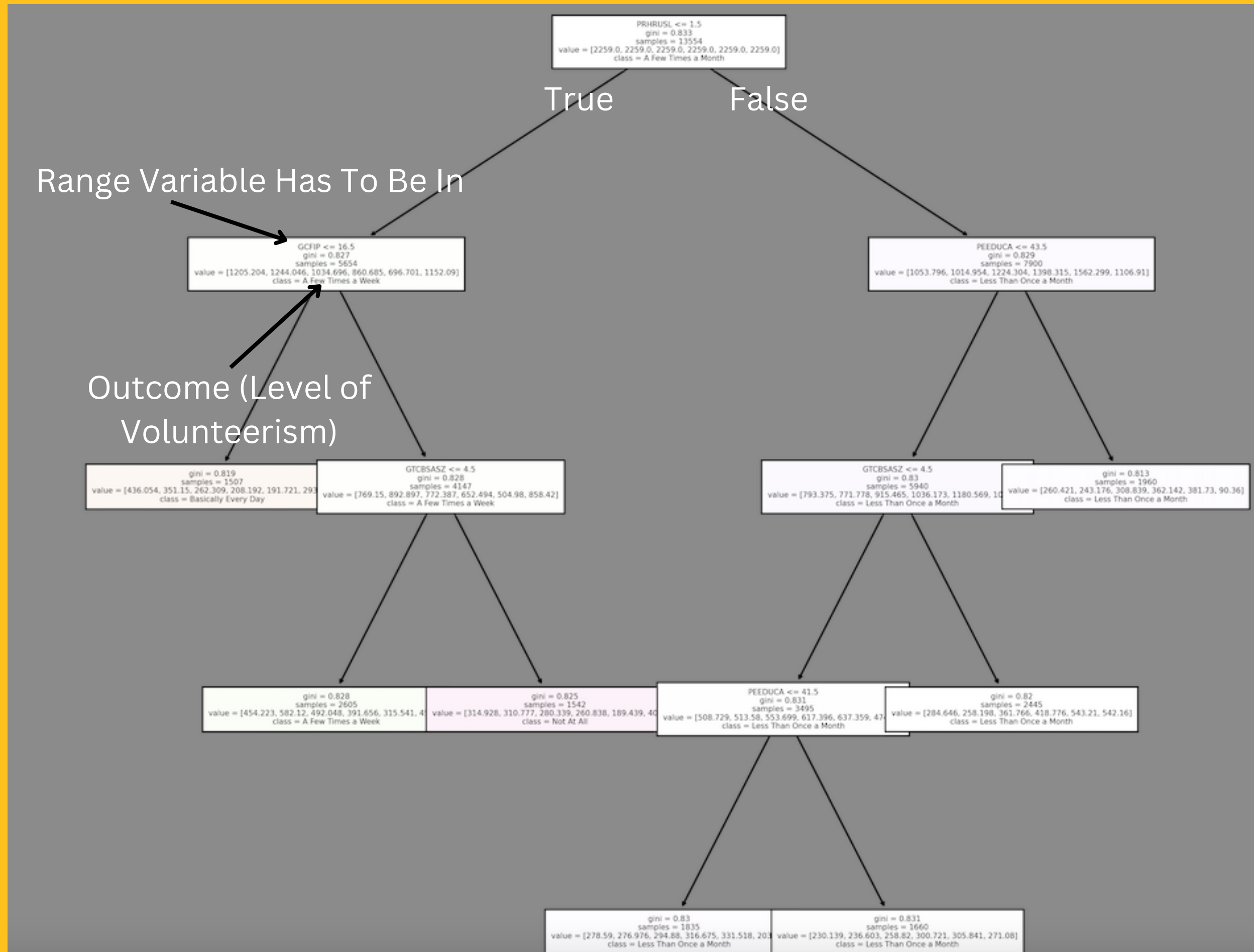
Decision Tree in classification asked a series of questions about my dataset to partition my data and find the most likely factors/criteria for each outcome.

Why did you use this?

I wanted to know what factors would result in a certain level of volunteerism, so I needed a classification model. Some requirements that I had for my model were speed and explainability. I specifically tuned the hyperparameters to avoid overfitting.

What was the result?

Using the results from my general model without hyperparameters, I decreased my features from 43 to 15 and obtained the best hyperparameters for my dataset. After including the hyperparameters, I obtained the following graph that predicted all outcomes except for Once a Month and Few Times a Month.



Decision Tree with Hyperparameters: Text Version

What is this?

This is the text version of the previous visual. It shows the results on the decision tree in a clearer but much less glamorous way.

```
|--- PRHRUSL <= 1.50
|   |--- GCFIP <= 16.50
|   |   |--- class: 1
|   |--- GCFIP > 16.50
|   |   |--- GTCBSASZ <= 4.50
|   |   |   |--- class: 2
|   |   |--- GTCBSASZ > 4.50
|   |   |   |--- class: 6
|--- PRHRUSL > 1.50
|   |--- PEEDUCA <= 43.50
|   |   |--- GTCBSASZ <= 4.50
|   |   |   |--- PEEDUCA <= 41.50
|   |   |   |   |--- class: 5
|   |   |   |--- PEEDUCA > 41.50
|   |   |   |   |--- class: 5
|   |   |--- GTCBSASZ > 4.50
|   |   |   |--- class: 5
|   |--- PEEDUCA > 43.50
|   |   |--- class: 5
```

I Volunteer Basically Everyday

Who Am I?

This volunteer likely works between 0-20 hours a week and lives in one of the following states: Alabama, Arkansas, Arizona, California, Colorado, Connecticut, Delaware, Washington D.C, Florida, Georgia, Hawaii, or Idaho.

I Volunteer A Few Times A Week

Who Am I?

This volunteer likely works between 0-20 hours a week and does not live in one of the following states: Alabama, Arkansas, Arizona, California, Colorado, Connecticut, Delaware, Washington D.C, Florida, Georgia, Hawaii, or Idaho. However, they also do not live in a metropolitan area; if they do, the population is less than 999,999.

I Volunteer Less Than Once A Month

Who Am I?

These volunteers likely had one of the three scenarios:

- They likely worked more than 21 hours a week and obtained a Master's degree or higher.
- They likely worked more than 21 hours a week, obtained a Bachelor's degree or lower, and lived in a metropolitan area of at least 1,000,000 people.
- They likely worked more than 21 hours a week, obtained an Associate's or Bachelor's degree, and lived in a metropolitan area of less than 999,999 people or a nonmetropolitan area.

I Don't Volunteer At All

Who Am I?

This volunteer likely works between 0-20 hours a week and does not live in one of the following states: Alabama, Arkansas, Arizona, California, Colorado, Connecticut, Delaware, Washington D.C, Florida, Georgia, Hawaii, or Idaho. However, they live in a metropolitan area with a population of at least 1,000,000.

Accuracy

It is extremely important to check how accurate our model is. Here is how this model measured up:

MAE = 1.14

MSE = 2.31

RMSE = 1.52

What does this mean?

Our model performed well on our test set!

Mean Absolute Error (MAE)

This is the average of the absolute differences between our true and predicted values. The lower this number, the better. It is less biased for higher values, so it may not 100% reflect the performance when dealing with higher errors.

Mean Squared Error (MSE)

This is the average of the squared differences between our true values and our predicted values. It ends up penalizing larger errors. We want this number as low as possible as well.

Root Mean Square Error (RMSE)

This indicates the spread of the residual errors, so it takes the square root of MSE. We want this number as low as possible too.

Conclusion

What does this all mean?

While I was surprised that our classifier did not use demographic factors to determine volunteer frequency, the factors chosen were just as interesting. Volunteers less likely to frequently volunteer tended to already work many hours (a category my best friend fit into) and live in metropolitan areas. While urban areas may provide a large pool of volunteers, they may need extra support and a clear plan from the organization on how they can dedicate their time to work and volunteer.

This also meant that my best friend was not alone! She fit into both categories, living in a major city and working a lot of hours in one week.



Further Scope

Thank you for coming on this journey with me! Here are some further questions/notes I had.

Further Question: How do people's commitment to volunteerism change over their lives, and what factors influence that?

Limitations: Self-reported surveys are only as accurate as the people that are answering them are. Everyone's definition of a level of volunteerism could be different. Also, while conducting EDA, some populations did not align with the U.S Census percentages (such as marriage status), which could indicate that this is not the best sample of the U.S. population.

PCA Model: Unfortunately, my PCA model did not go as planned. The possibility of random data made me doubt my later results. In a future follow-up project, I might choose another dataset.

Citations

<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.shutterstock.com%2Fimage-vector%2Ftired-african-woman-sitting-on-sofa-2126681324&psig=AOvVaw2DkEWH3Pwp-nOWpBblOeBq&ust=1683939722620000&source=images&cd=vfe&ved=OCBIQjhqxqFwoTCLCDy7rK7v4CFQAAAAAdAAAAABAF>

<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.istockphoto.com%2Fillustrations%2Ffriends-talking&psig=AOvVaw13nf2e8QecHrAkplDolVo4&ust=1683939886116000&source=images&cd=vfe&ved=OCBIQjhqxqFwoTCLiJ44nL7v4CFQAAAAAdAAAAABAE>

<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.istockphoto.com%2Fillustrations%2Fcartoon-of-a-black-girl-thinking&psig=AOvVawOPcfZDHbL4zMJLZvFmEb5w&ust=1683939963483000&source=images&cd=vfe&ved=OCBAQjRxqFwoTCJCBia7L7v4CFQAAAAAdAAAAABAF>

<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.istockphoto.com%2Fillustrations%2Fgirl-waving-goodbye&psig=AOvVaw2ETtKquQsh8FwlubtX03jp&ust=1683940027291000&source=images&cd=vfe&ved=OCBAQjRxqFwoTCMifgMzL7v4CFQAAAAAdAAAAABAE>

<https://neptune.ai/blog/feature-selection-methods>

<https://www.machinelearningplus.com/machine-learning/principal-components-analysis-pca-better-explained/>

<https://towardsdatascience.com/hyperparameters-of-decision-trees-explained-with-visualizations-1a6ef2f67edf>

<https://www.census.gov/>

<https://akhilendra.com/evaluation-metrics-regression-mae-mse-rmse-rmsle/#:~:text=MAE%20is%20less%20biased%20for,dealing%20with%20large%20error%20values.>

<https://github.com/MeghanaKshirsagar/Perspectives-on-Data-science/blob/main/Imbalanced%20Dataset%20Approaches/Algorithm%20Level%20Methods.ipynb>