Kenny Waite
DSC 530 Term Project

Effect of Income on Volunteer Tax Returns in North Carolina

Question/Hypothesis

Does a person's adjusted gross income play a factor in whether they will have their taxes done by a volunteer in North Carolina? That is the central question I am attempting to answer with this project.

My hypothesis is yes, people in lower income tax groupings take advantage of volunteer tax return services more than people in higher income tax groupings.

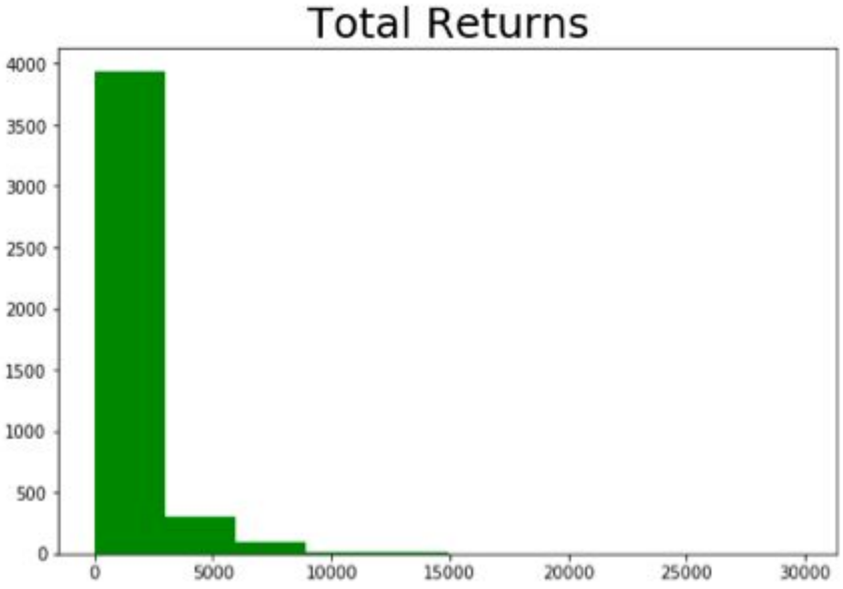
I also wanted to look at other variables to see if there were any trends in volunteer prepared returns and the people who are more likely to use this service including:

- The number of dependents for tax filer
- City
- Farm Returns
- Students

Variables

- **Size of adjusted gross income:** six subcategories based on adjusted gross income grouping
- **Number of volunteer prepared returns - Total:** Total number of volunteer prepared tax returns filed for each zip code in each gross income grouping
- **ZIP code**: The city zip code associated with the return.
- **Number of Returns**: The total number of returns filed for each zip code in each gross income grouping
- **Number of farm returns**: The total number of farm returns filed for each zip code in each gross income grouping
- **Child tax credit:** The total amount of returns files with child tax credit deduction for each zip code in each gross income grouping
- **Student loan interest deduction:**The total amount of returns files with student loan interest deduction for each zip code in each gross income grouping

Total Returns Histogram and Outliers

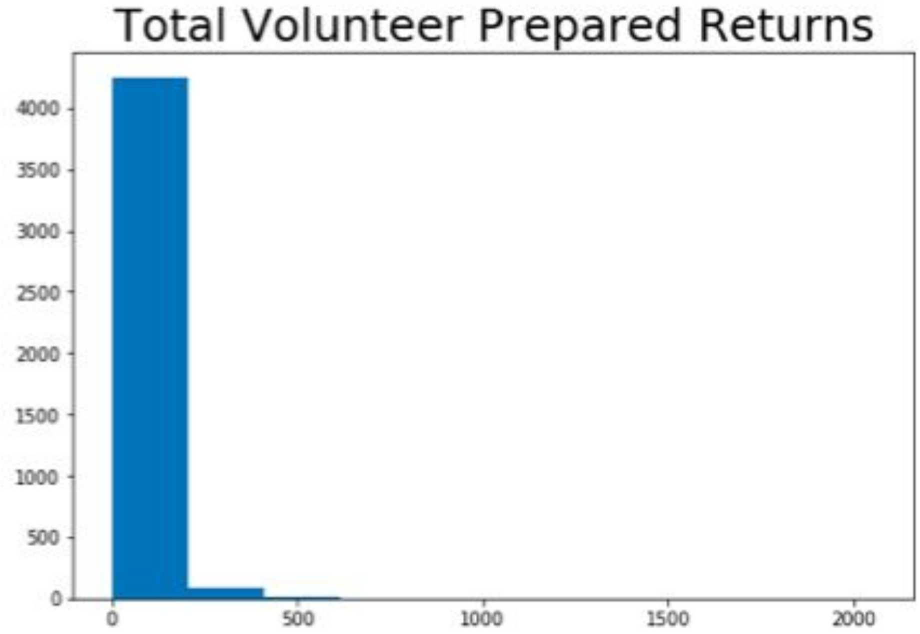


The mean of the total returns variable is 1,041.04
The mode of the total returns variable is 0.0
The spread of the total returns variable is 2,806,884.16
The Q3 and Q1 tails of total returns variable is 1,260.0 and 110.0

Outlier Analysis

- No outliers detected for total returns variable
- Reviewed top returns close to 15,000 and all zip codes belonged to a major city in NC (Charlotte, Raleigh, etc.)
- Reviewed all zip codes with 0 returns. This was for income groupings in the higher range for smaller cities
- No outliers removed

Total Volunteer Returns Histogram and Outliers

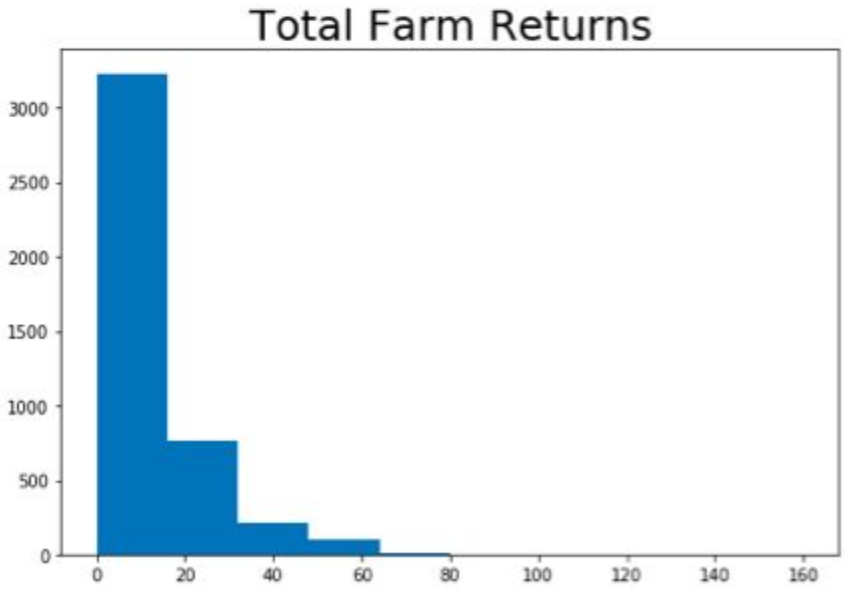


The mean of the total volunteer returns variable is 21.22
The mode of the total volunteer returns variable is 0.0
The spread of the total volunteer returns variable is 41,27.67
The Q3 and Q1 tails of total volunteer returns variable is 0.0 and 0.0

Outlier Analysis

- No outliers detected for total volunteer returns variable
- Reviewed top volunteer returns and one stood out with about 250 more than the next highest.
- Determined not to be an outlier
- Reviewed all zip codes with 0 returns. This was for income groupings in the higher range for smaller cities
- No outliers removed

Total Farm Returns Histogram and Outliers

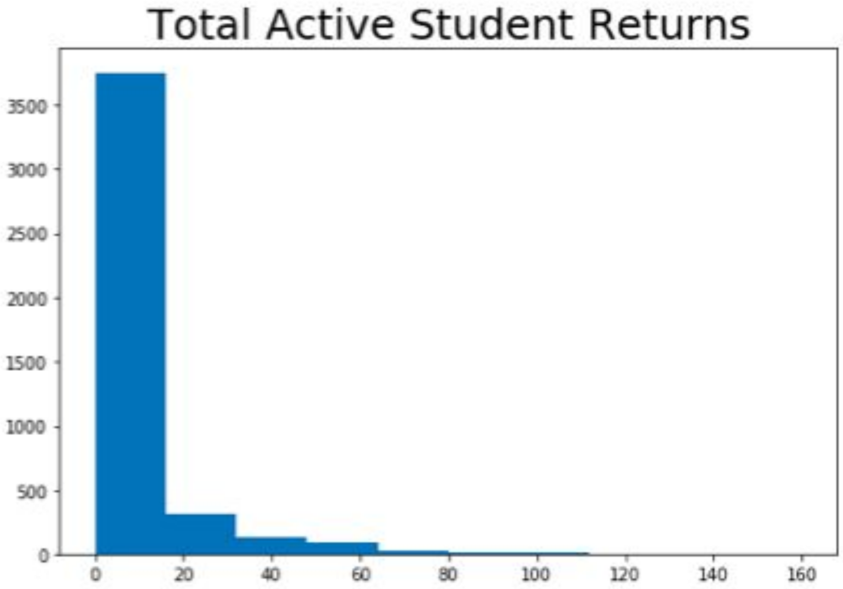


The mean of the total farm returns variable is 8.28
The mode of the total farm returns variable is 0.0
The spread of the total farm returns variable is 244.74
The Q3 and Q1 tails of total farm returns variable is 20.0 and 0.0

Outlier Analysis

- No outliers detected for total farm returns variable
- Reviewed top returns and all zip codes belonged to a rural cities in NC
- No outliers removed

Total Active Student Returns Histogram and Outliers

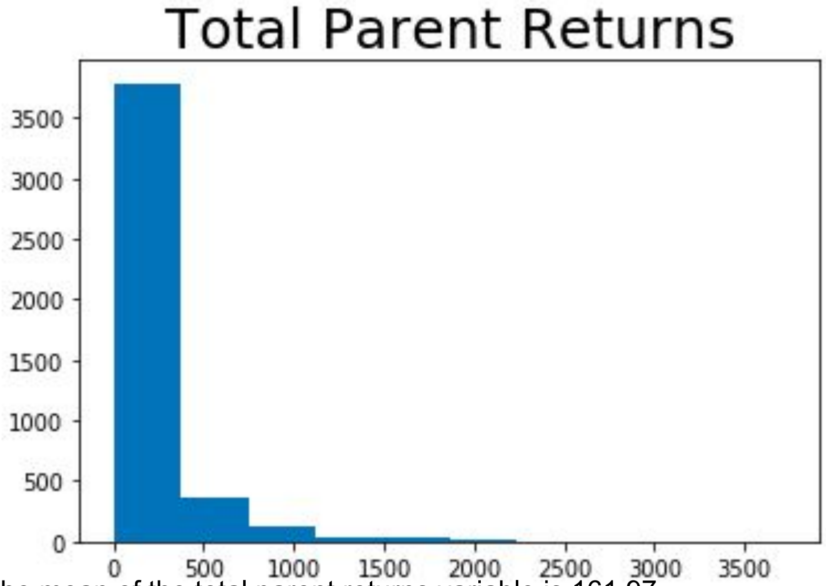


The mean of the total active student returns variable is 5.36
The mode of the total active student returns variable is 0.0
The spread of the total active student returns variable is 244.22
The Q3 and Q1 tails of total active student returns variable is 0.0 and 0.0

Outlier Analysis

- No outliers detected for total active student returns variable
- No outliers removed

Total Parent Returns Histogram and Outliers

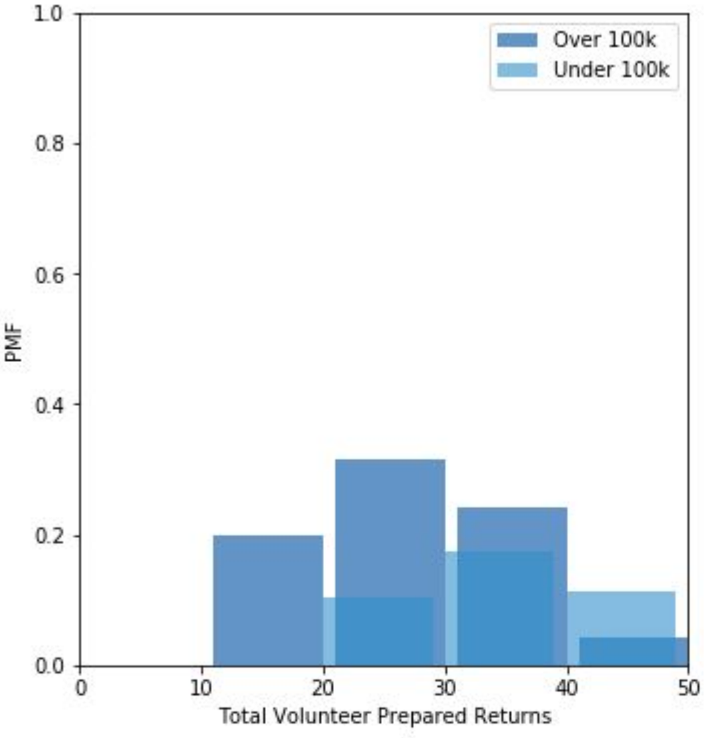


The mean of the total parent returns variable is 161.97
The mode of the total parent returns variable is 0.0
The spread of the total parent returns variable is 77,150.47
The Q3 and Q1 tails of total parent returns variable is 190.0 and 0.0

Outlier Analysis

- No outliers detected for total parent returns variable
- No outliers removed

Total Volunteer Tax Returns PMF Compared

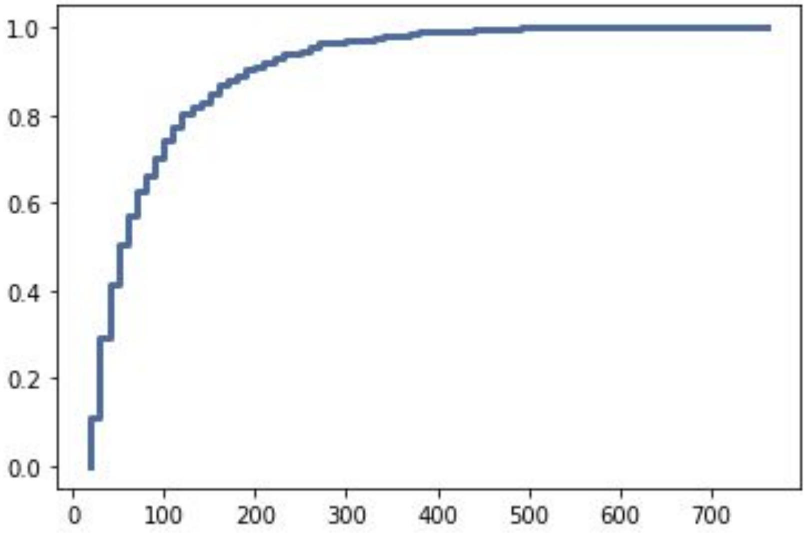


For this PMF comparison scenario I compared the PMFs for volunteer tax returns filed for people who have incomes over 100k and under 100k.

This graph shows that the highest probability a zip code with filers under 100k will file between 30-40 volunteer prepared tax returns as about 20%.

Compared to filers over 100k, this number drops to the 20-30 range at about 35%.

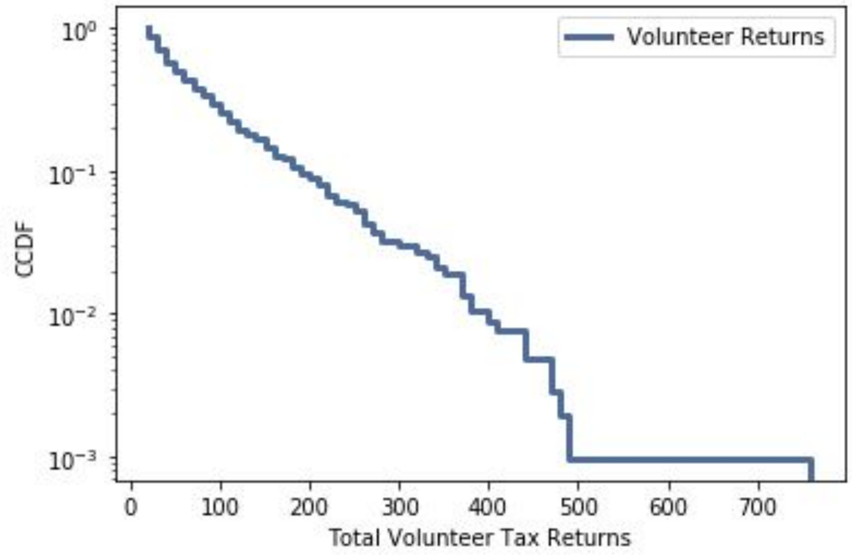
Total Volunteer Tax Returns CDF



This CDF tells us that the most common values are are between 20 and 100. This means the most common amount of volunteer tax returns by zip code is between 20 and 100.

About 10% of volunteer tax returns are between 0-10 and about 90% are less than 300 volunteer filed tax returns

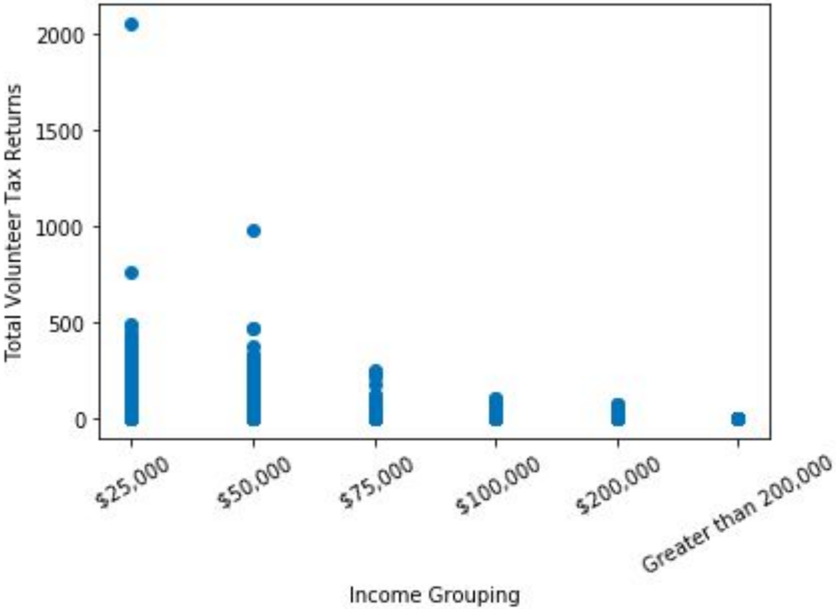
Analytical Distribution - Complementary CDF



This complementary CDF is pretty straight, which indicates an exponential distribution is a good model for this data.

This indicates the underlying assumption that total volunteer tax returns are affected by income level

Scatterplot 1 Volunteer Tax Returns and Income

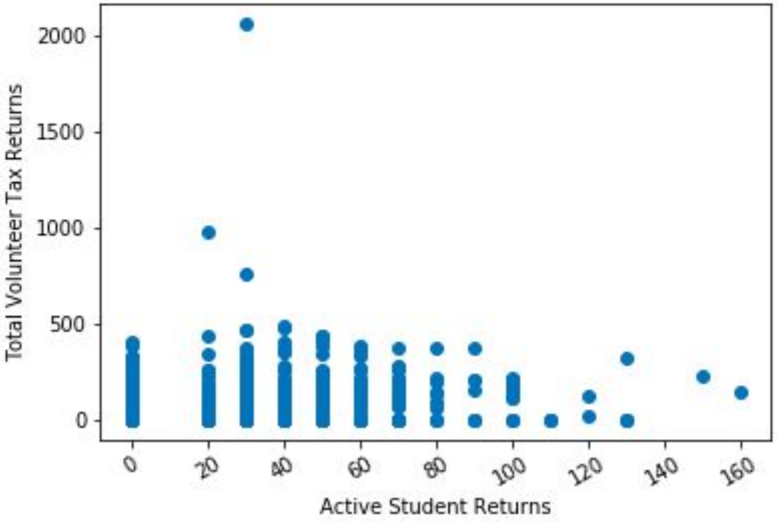


The first scatter plot I created plotted the relationship between the total number of volunteer tax returns and the income grouping of the tax filers.

The scatter plot appears to have a negative linear relationship. This is confirmed by calculating the Spearman’s Rank Correlation of -0.49. I chose Spearman’s Rank over Pearson’s due to the left skewness of the data. This calculation tells us that as income increases, the number of volunteer tax returns decrease.

This correlation doesn't imply causation. The correlation is only telling us which direction the two variables move together

Scatterplot 2 Volunteer Tax Returns and Student Filers



The second scatter plot I created plotted the relationship between the total number of volunteer tax returns and the total number of student tax filers.

The scatter plot appears to have a positive linear relationship. This is confirmed by calculating the Spearman’s Rank Correlation of 0.44. I chose Spearman’s Rank over Pearson’s due to the left skewness of the data. This calculation tells us that as number student tax return filers increases, the number of volunteer tax returns increases.

This correlation does not imply causation. The correlation is only telling us which direction the two variables move together

Testing a Difference in Means

Given a samples and an apparent effect, what is the probability of the amount of volunteer tax returns decreasing as income increases by chance?

Null Hypothesis: The volume of volunteer prepared tax returns is not affected by the income grouping

The difference in means test results in about 0.07 with a p-value of 0.02, which means that we expect to see a difference as big as the observed effect about 7% of the time. So this effect is not statistically significant.

Regression and Multiple Regression Analysis

Regression Analysis

- When performing a regression analysis on the effect of income grouping on volunteer tax return, the R squared resulted in 0.11.
- This indicates that the model is not statistically significant and the model should not be used to predict the number of volunteer tax returns by zip code based on incoming grouping

Multiple Regression Analysis

- When performing a regression analysis on the effect of income grouping and active student filers on volunteer tax return, the R squared resulted in 0.11.
- This indicates that the model is not statistically significant and the model should not be used to predict the number of volunteer tax returns by zip code based on incoming grouping and active student filers