Volunteer Tax Return Analysis

Description

For this project, I wanted to understand if a person's adjusted gross income played a factor in whether they will have their taxes done by a volunteer in North Carolina. I also looked at additional 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, and student filers. This is one of the first projects I worked on in my Master's program.

Navigation

- README File - this file that provides and overview of the project
- Code Repository - this folder contains all of the Python code as well as necessary data files for running the Python code
- Volunteer Tax Return Presentation - presentation summarizing the findings of my analysis

Tools

For this project I used Python to import/clean the data, perform the exploratory data analysis, and run a regression model with the data to predict whether a person would use a volunteer tax return or not. I also used Google Docs and Google Slides to create my documentation.

Data

I used data from the IRS.gov website that included all of the information on North Carolina tax returns for the 2017 tax year (taxes filed in 2018).

https://www.irs.gov/pub/irs-soi/17in54cmcsv.csv

Methods/Evaluation/Techniques

The majority of this project focused on the exploratory data analysis portion of the data science project lifecycle. During EDA, I reviewed histograms of the data to see if there was a skew in the data, inspected the data for outliers, created additional features, and reviewed the correlation coefficient to determine linear direction and strength. I also perform a regression and multiple regression analysis on the data to determine the ability to predict if someone would use a volunteer tax return preparer or not.

Conclusion

When performing a regression analysis on the effect of income grouping on volunteer tax return, the R squared resulted in 0.11 with a p-value less than 0.05. This indicates that the model is statistically significant with the income grouping predicting 11% of variance in the amount of volunteer tax returns filed. This is not a great R squared and in the future I would like to include more variables to increase this value without over complicating the model.