**Samantha Sloate**

**Data Visualization in R**

**Final Project Proposal**

**1. Introduction­­­**

My general research question is:

*How are components of personal income and demographics within the North Carolina population projected to change over time? How will these changes affect North Carolina’s income tax revenue in 2040?*

This question is part of my larger Master’s thesis for the Public Policy program, where I am working for the Office of State Budget and Management (OSBM) to forecast state income tax liabilities twenty years in the future. That analysis involves creating a static microsimulation model using income and tax data from the Current Population Survey (CPS), population predictions from state demographers and additional data from OSBM.

The model and analysis will be done outside of R, but importing, cleaning, and manipulating the dataset to produce understandable visualizations are equally essential components of this project.

I aim to produce four main visualizations:

1. trends in demographics (race, age, and sex) over time;
2. trends in components of income over time;
3. differences in total state adjusted gross income between 2019 and 2040 ; and
4. differences in tax liability between 2019 and 2040.[[1]](#footnote-1)

**2. Data**

*Main Dataset*

The main dataset that will be used combines data from three sources: state demographers, North Carolina’s OSBM, and the CPS. Each component is explained in more detail below. See attached files for a comprehensive codebook and summary statistics.

*CPS Data*

The Current Population Survey, administered by the Bureau of Labor Statistics and the Census, collects relevant income, tax, and household characteristic data. Specifically, I will be using the Annual Social and Economic Supplement portion of the CPS. These data form the basis of the main dataset.

*OSBM Data*

Certain necessary components of personal income, such as capital gains, are not included in the CPS data. Thus, I impute additional data provided by OSBM into the main dataset.

*State Demographic Data*

Demographic data from OSBM includes population predictions for all combinations of age, race, sex, and ethnicity. The dataset that will be used in this R analysis utilizes these demographic data to adjust CPS’ weights, making observations more indicative of the North Carolina population now and in the future.

**3. Data Analysis Plan**

**Key variables** to answer my question include

1. taxable sources of income (wage, business, farm, retirement, etc.) over time;
2. age, sex, race, and ethnicity of people in North Carolina over time;
3. adjusted gross income; and
4. personal income tax liability.

**Preliminary exploratory data analysis** shows that North Carolina will grow by about 2,168,000 people from 2019 to 2040, which represents a 21% increase. Currently, the largest source of personal income in North Carolina comes from wage incomes (Figure 1).

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

The **methods** I will use to answer my question include descriptive graphical analysis, calculation and visualization of state AGI (to the best of my availability, given the components of data that I have), and visualization of tax liability.

Because my client is not looking for causal answers and instead looking for trends, the descriptive analysis above will be enough to support my “answer.” I do not believe any statistical tests need to be run.

1. My client, OSBM, has requested that their tax calculator (coded in SAS) be used for the tax liability analysis. Thus, I will feed my dataset into their tax calculator and produce a single number for 2020 tax liability. This number will necessarily not be part of the main dataset and will instead be read in to RStudio separately. [↑](#footnote-ref-1)