The data used for this analysis is sourced from the General Social Survey (GSS), containing information on education, prestige, and perceived social class. The dataset comprises a broad range of sociodemographic and attitudinal variables collected from a nationally representative sample of individuals. It replicates these questions yearly– providing a collection of data on social trends overtime.

This wealth of information includes details about individuals' educational backgrounds, their career and job-related aspects, and their subjective assessments of their own social status. The dataset is, therefore, a powerful tool for researchers and social scientists to explore and analyze various social trends and phenomena over time.

This analysis focuses on three variables: *degree*, which is the highest degree obtained by the respondent, *prestige*, which is a score assigned to professions based on a rating system developed by NORC (National Opinion Research Center). Using a nine-step ladder, the prestige scores in the GSS studies were formed by asking respondents what the social standing of various occupations were to them. The last variable involved is class; it is the respondent's self perceived social standings.

For *degree*, before any data cleaning or filtering, the original dataset consisted of a total of 72,230 records, with respondents providing their educational qualifications. Within this dataset, there are six unique options that respondents could select as their highest level of education: "bachelor's," "less than high school," "high school," "graduate," "associate/junior college," and the presence of missing values ("nan"). Among these options, "high school" emerged as the most frequently chosen degree level. It was selected by 36,446 respondents, and the overall data indicates that most respondents did not continue school after getting a high school diploma.

Prestige level captures the perceived social prestige or status of the respondent's occupation. Before any data cleaning or filtering, this variable contained a total count of 24,303. In this dataset, the "prestige" variable employs a standardized prestige score, calculated as a straightforward mean value of the prestige ratings assigned to various occupation categories. These ratings are then converted to a scale ranging from 0 (representing the lowest prestige level) to 100 (representing the highest prestige level). The "prestige" variable exhibits diversity, encompassing 63 distinct numeric values assigned to the participants' jobs. However, the most prevalent prestige score among the respondents was 50.0. This score was selected by 1,913 individuals, signifying the prominence of this middle-of-the-road prestige level within the dataset. Additionally, it's worth noting that there were instances of missing data ("nan") in this variable.

The self-perceived social class is divided into six distinct categories, each offering a unique perspective on how individuals assess their societal status. These include “middle class,” “working class,” “'upper class,” “lower class, “no class,” and instances of missing data denoted as ("nan”). Prior to any data cleaning or filtering, the *class* variable encompassed a total of 68,894 records. Within these six social class categories, "middle class" stands out as the most prominent. To be specific, it was the most frequently chosen category, with 31,014 respondents identifying themselves as part of the middle class.

During the process of handling and preparing the data, several challenges were encountered. First, dealing with a large dataset required reading it in chunks, making it more manageable. To begin, the dataset's substantial size necessitated a chunked approach to reading the data. This method allowed for more efficient management of the data's scale. While this solved the issue of size, it brought about the challenge of consolidating these fragments into a coherent dataset. To address this, a new dataframe was created, enhancing the organization and overall cohesion of the data.

Another challenge emerged in the form of the "prestige" variable. The lack of comprehensive documentation regarding its meaning and calculation required a deeper dive into past documents and discussions to gain insight. This step was pivotal in fully comprehending the variable's significance within the dataset.

Initially, the consideration was given to replacing missing values with "Unknown" to maintain consistency across the dataset. However, upon further evaluation, a different approach was adopted. It was decided to drop rows with missing values and also exclude any instances where "prestige," "educ," and "class" were included as unique values. This choice was made because these variables were not essential for the analysis and were not part of the options provided in the CSS codebook for their respective questions. This streamlined the dataset and ensured that only the relevant data points were retained for analysis.

Data consistency was also a vital concern. An inconsistency was detected in the "prestige" variable, where certain records included the term "prestige" as a category. To maintain uniformity, I standardized the variable by eliminating such anomalies

Lastly, to facilitate exploratory data analysis (EDA) and visualization, "no_na" versions of variables were created. These versions excluded records with missing values, yielding cleaner data for EDA, enabling a more focused analysis of patterns and relationships.