Data Exploration

After exploring the data from the NLSY79 database, we were able to find many different features related to the participants' lives that we were interested in exploring, as well as choose a column to act as our income data. We decided to go with the most recent income data that was available to us, which was in 2018. This data does not include military income, but does include all other forms of wages and salary the respondent had earned in the past year. After removing people who did not enter a valid income for 2018, we found that we had 6,571 remaining participants. While the reasons behind people not providing their 2018 income could be an indication of what their income is or their general status, we will be excluding any analysis of these possible reasons from our model.

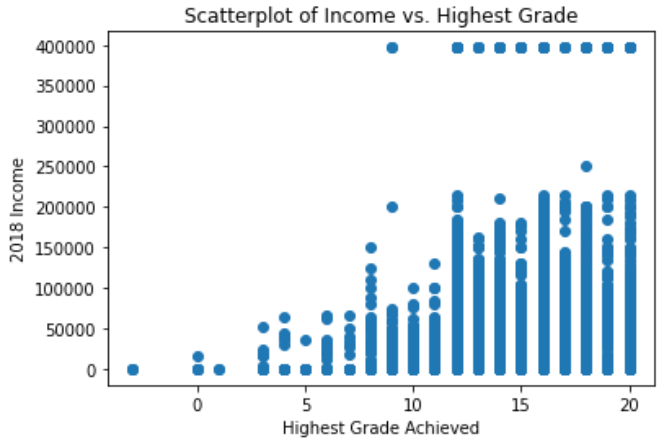
For the features that we want to explore, we picked 15 different columns that represented a variety of different aspects of the participants' lives. Below are the feature names that we are exploring:

ID Code	Sample ID Code	Parent 1 Highest Grade	Race	Sex	Self Esteem Score	AFQT Percentile Score	Type of Residence 1983	Average Drinks 2008	Highest Grade	Par 1 Alive	Par 2 Alive	Par 1 Death Age	Par 2 Death Age	Fam Size 1979
1	5	12	3	2	-5	-4	-5	-5	12	-4	-4	-4	-4	5
2	5	8	3	2	16	6841	-4	-4	12	0	1	-4	-4	5
3	5	10	3	2	20	49444	11	2	12	1	1	-4	-4	5
4	5	10	3	2	-5	55761	11	-5	14	-4	-4	-4	-4	5
5	1	13	3	1	23	96772	3	-5	18	-4	-4	-4	-4	4

For many of these features, there are a variety of types of missing data. There are a range of negative values possible for each of these features, which indicate situations such as the respondent was never asked the question or it's not applicable to the respondent. How we are managing the missing data depends on the feature. For example, for the highest grade achieved, all negative values are being replaced by the average value. Many of the features, such as race, sex, and family size, contain no missing values across the respondents that we have 2018 income data for.

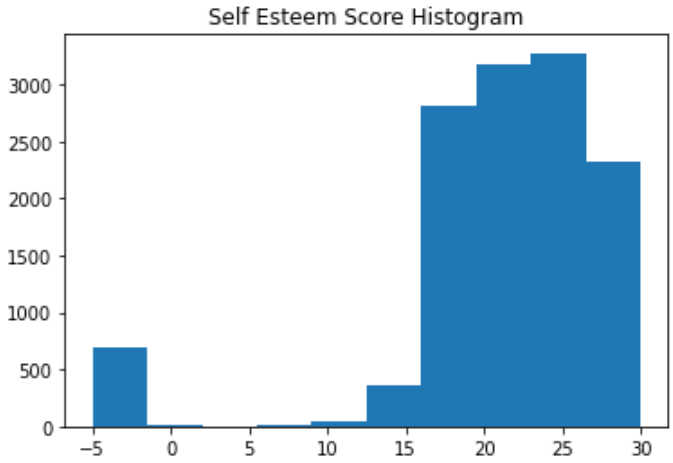
Descriptive Statistics

Some preliminary exploration we did on our dataset includes many graphs and histograms. For example, we graphed 2018 Income vs Highest grade achieved as seen below:

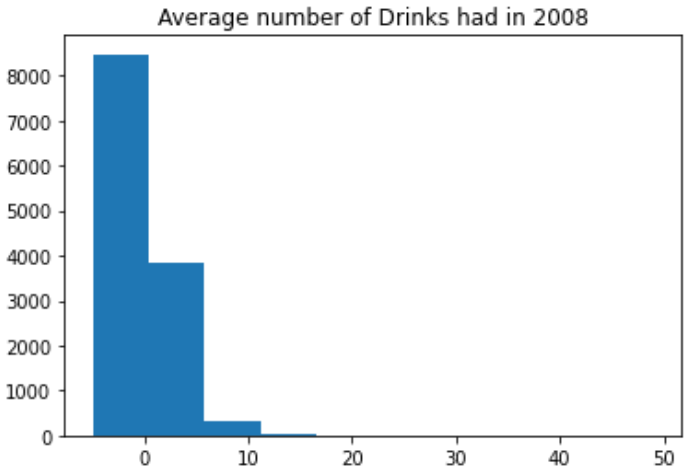


As you can see in the graph, there does seem to be a trend of higher grades relating to higher incomes. This supports us using the feature "highest grade achieved" in our model.

The Histogram below shows the data we have on a self esteem score metric in our data. As you can see, the score seems to go between 0 and 30, with many scores in the high range. You can also see that there are a few data points of -4, which in this data set means that they did not answer this question. This is something we need to take into account when creating our model because -4 isn't an actual score, it is just a lack of score.



This Histogram below shows the data we have on the average number of drinks a person might have. You can see that the distribution of this data leans heavily toward the low side. There is some missing data in this value as well, you can see that many of the values are -4, which are actually no answer.



Model Effectivity

In order to create the best possible final model, we will be taking a variety of steps to make sure our model would predict well on an out of sample population. First we will split our data into a training set and a test set. This will allow us to test any models we develop on the test set to make sure it fits well on data that was not used to make the model. In order to specifically avoid over and underfitting, we will make sure to only use the features that offer the greatest impact on our predictions, and will likely drop any features that we deem we either do not have enough data to be valuable and could easily lead to overfitting, or that are causing us to have a significantly higher error on the test set than the training set.

In order to determine whether our model is effective, we will be calculating the least square error of our final model on the test population. We will be looking at this error throughout the process of making our model, so that we can determine which enhancements we are making are beneficial and which are either unproductive or counterproductive for fitting our test samples. Since there are distinct subgroups in our data (race, sex, etc), we will also perform a test to make sure our model works well on any given subgroup, and model issues fitting minority populations are not overshadowed by the overall error. This should give us a high confidence that our final model can predict on any given out of sample population.

Preliminary Analysis and Going Forward

Per our goal, we used 15 features to predict the participants' income of the year 2018; The features we used include "highest grade achieved", "residence type in year 19xx", and so on. All these features are already numbers (i.e. not "strings") when the dataset has been downloaded, so there are no typical transformations needed at this point; some transformations might be needed as we move forward towards more complicated models.

However, many values are negative, meaning either the participant refused/failed to answer the questions, or the question was simply not applicable. If we simply dropped all rows which contain any such values, we will be left with only 36 rows; therefore, we have to manage negative values in a more meaningful way. Our current approach is to use average values to replace negative values, while dropping all rows where the income (our y variable) is not available. This leaves us with 6571 rows, which is sufficient to make a fairly complicated model.

We also split the dataset into training and testing sets. We used the 20% testing, and 80% training rule from class. This split helps reduce overfitting/underfitting, and will give us an impression/expectation on how our model will perform on new data.

With all these preparations, we finally fit a basic linear regression model to predict the effects of "highest grade achieved" by the participant, on their income of the year 2018. We added an offset and trained our model to minimize squared error. The result, as expected, was: the higher grade achieved, the higher expected income for the participant. This concludes our preliminary analysis, which consisted of three steps: cleaning up data (managing negative N/A values), splitting training/testing dataset (8-2 rule), and fitting a prototype linear model.

Our next step is to continue analyzing each feature and evaluate their individual effectiveness on the prediction. After this, we will select what features to continue working on, and probably add in different features to further complicate our model. Data transformation might be required once more features take part, because some features might be related, or cannot be directly used linearly.