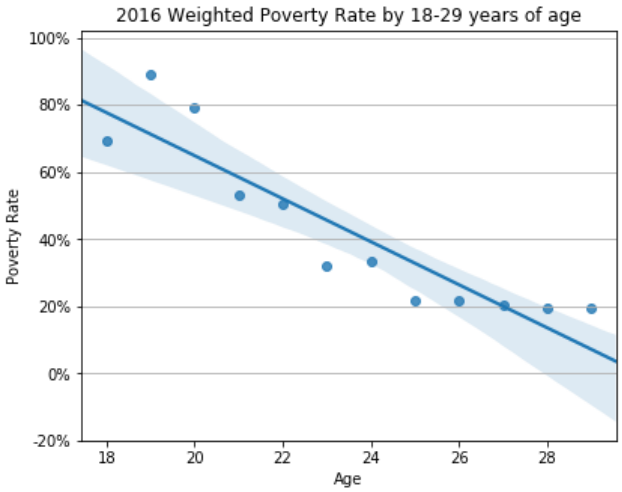
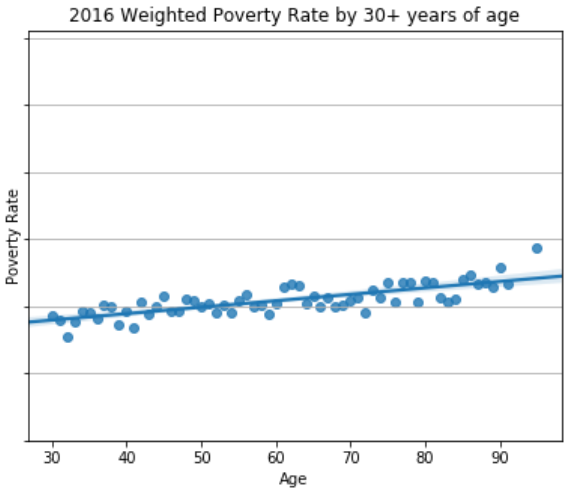
Predicting and Combating Poverty in NYC

Capstone Project 1 Milestone Report

Craig Paulette

c74@protonmail.com





Contents

1. Problem statement

With a population of over 8.5 million and a 2016 poverty rate of 19.5%, New York City is home to over 1.5 million people living in poverty. This project aims to predict poverty and analyze how to combat it in New York City.

The project has two primary ‘clients’:

- New York City government, in regards to which policies to consider and the impacts of these policies on the overall poverty rate and the lives of its citizens.
- Individual citizens, insofar as individual choices can impact individual poverty status.

While other possible clients may benefit from this research, they would be strong candidates for further research and are not the focus of this effort:

- Federal and state governments can have powerful impacts on poverty rates. Given the richness of the existing NYC data set, federal and state initiatives are not a primary focus of this project.
- Many non-governmental organizations address poverty and related issues. However, given the number of different organizations, goals, and perspectives, non-governmental actors are not considered a primary client here.

2. Data Wrangling Summary

A. The Dataset

- Data from https://data.cityofnewyork.us/browse?q=poverty
- 12 annual data files, from 2005 to 2016 inclusive (e.g. NYCgov_Poverty_Measure_Data__2016_.csv)
- CSV files with ~80 columns and ~60,000 rows each
- Each file had essentially the same format and contained (mostly) the same information
- Data was taken from US Census Bureau American Community Survey data, augmented by NYC Poverty Research Team
    - https://www.census.gov/programs-surveys/acs/
    - https://www1.nyc.gov/site/opportunity/poverty-in-nyc/poverty-measure.page
- The data includes fields such as annual income, education level, age, ethnicity, number of people in household, household-level costs, the official US poverty threshold, and custom NYC poverty statistics taken from imputed taxation levels along with income and costs
- Data types:
    - Classification types encoded as integers (e.g. 1 if in poverty, 2 if not in poverty)
    - Floats for financial data (e.g. wages for the calendar year)

B. Data Wrangling

Process Overview:
Since each file was expected to have roughly the same information, the process was as follows:

1. Examine/clean one file
    - Investigate and clean any NaN values
    - Create a function to check for outliers outside of allowable values
        - For example, for '1 if in poverty, 2 if not'; 3 is not an allowable value
2. Assuming a reasonable level of success in step 1, create a function to automate the steps taken and run it on all data files
3. Fix any inconsistencies between the datasets and merge them into a single dataframe
4. Investigate and fix any remaining outliers

Details on Step 1:
The 2016 data file had 79 columns and only 9 columns containing NaNs ('Not a Number' values):

- One column had NaNs representing *either* a person who is less than five years old, or only speaks English. I created two new categories: one for anyone who only speaks English, and one for people less than five years old.
- Seven columns had NaNs meaning that the question wasn't relevant (e.g. a question about employment status asked to an eight-year-old). I changed these to a new category 0 for each.
- The only other NaNs were two rows without an official poverty threshold. I removed these two rows.

Sanity checks:

- I checked the data against the data dictionary, and for outliers. A few discrepancies led to more data cleaning; for example, when I noticed a value of 10 on a 9-point scale (only 100 rows out of 60,000+), I changed it to a zero (no information).
- To check my edits, I subtracted a copy of the original dataframe from the edited dataframe; the number of differences by column matched exactly the number of NaNs in the original dataframe.

Details on Step 2:

I created a function to automate the process and ran it against all files, with output of any anomalies to the notebook. I discovered a few anomalies in the 2005-2007 files:

- Coding for the ‘weeks worked’ column changed from 2005 to 2016. I wrote and applied a function to recode the older files to be in line with the newer format.
- Some columns had different names and simply needed to be updated.
- Some less-important columns are missing in the older files.
- One column was coded as NaN rather than 0 when the person was less than five.

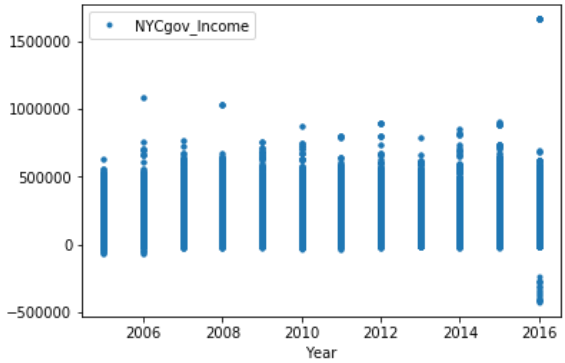
Details on Step 3:

I added a 'Year' column to each dataset and merged them into a single dataframe. I fixed a few inconsistencies in column naming that I'd previously missed. Along the way, I noticed that the data dictionary for the 2016 data is missing a column, so I emailed NYC Open Data to make them aware.

Details on Step 4:

I'd already reviewed the classification columns, so I was most interested in checking visually for any outliers that were technically legal but questionable. I created a list of columns of interest and ran quick-and-dirty charts by year for these columns.

One column stood out:



The NYCgov_Income -- the primary value in calculating NYC poverty status -- has 23 rows in 2016 where the person had both income and taxes in the six-figure range, and their resulting income after taxes was less than negative-100,000 dollars. Other years did have some values of NYCgov_Income that were less than zero; but none as egregious as this. I reached out to the NYC Poverty Research Team for clarification, but have gotten no response.

To retain any meaningful information in below-zero values without having the data skewed by enormous incomes and tax values (that only existed in one year), I updated the combined all-years dataframe to only include rows with NYCgov_Income of more than -50,000 (negative fifty-thousand dollars).

This removed a total of 31 rows (23 from 2016, and eight from all other years) from the combined dataframe.

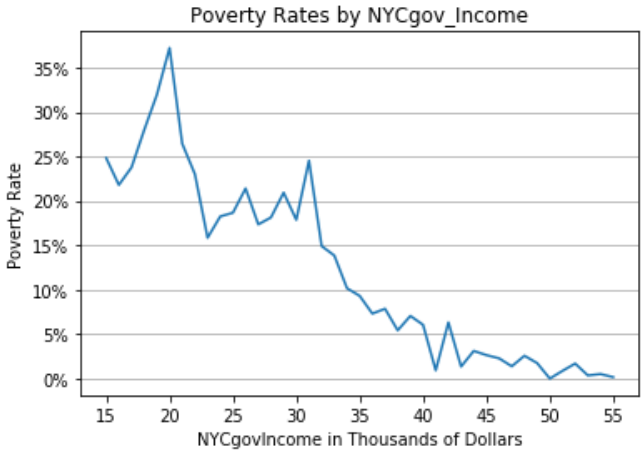
3. Initial Findings From Exploratory Data Analysis

NYC calculates poverty by ‘poverty units’; think of them like households, but consider a single parent renting space in an older couple’s home. In this instance, there are two poverty units, not one. I’ll colloquially use ‘household’ to mean ‘poverty unit’ here for readability.

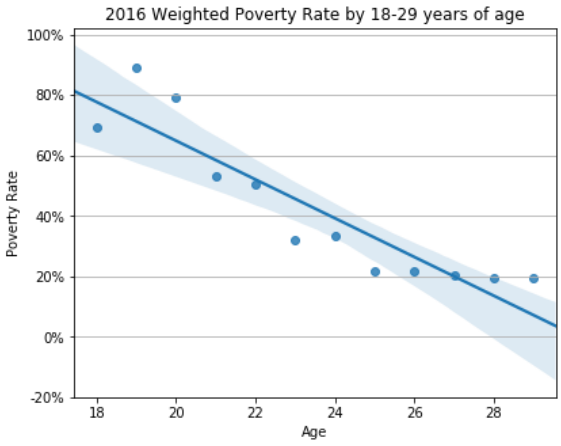
The NYC poverty calculation is **NYCgov_Income versus NYCgov_Threshold.** NYCgov_Income is total income net of taxes and fixed costs (medical, rent, childcare) for the poverty unit. NYCgov_Threshold is the NYC poverty threshold, which is the official US poverty threshold modified by a NYC housing-cost adjustment.

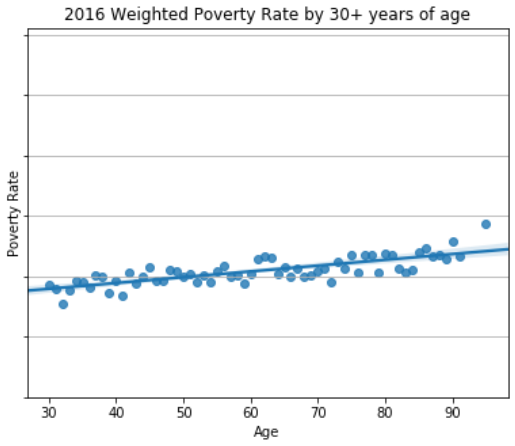
Thus, both NYCgov_Income and NYCgov_Threshold rely on some computed values that extrapolate from the US Census American Community Survey data from which they’re drawn.

Poverty rates generally trend down with increasing income, and drop below 10% for households (poverty units) with total income of more than $35,000.

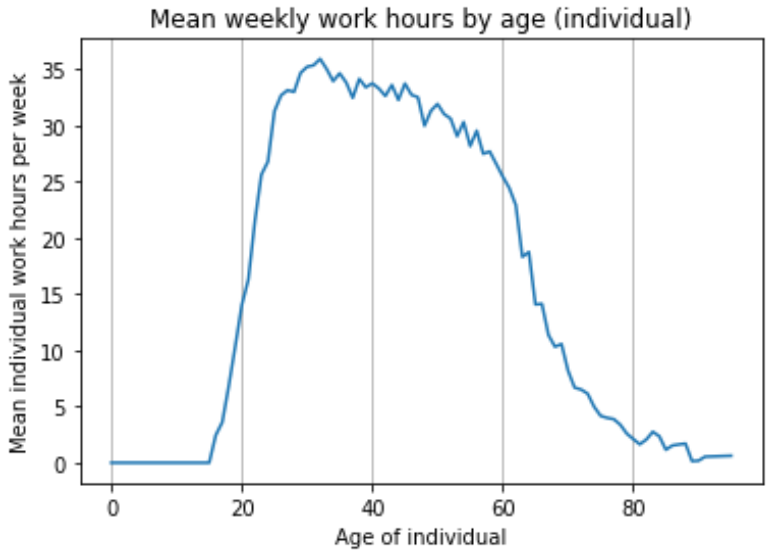


Poverty rates are especially high among younger households, and generally trend upward after 30 years of age.





Individual work hours are lowest where poverty rates are highest.

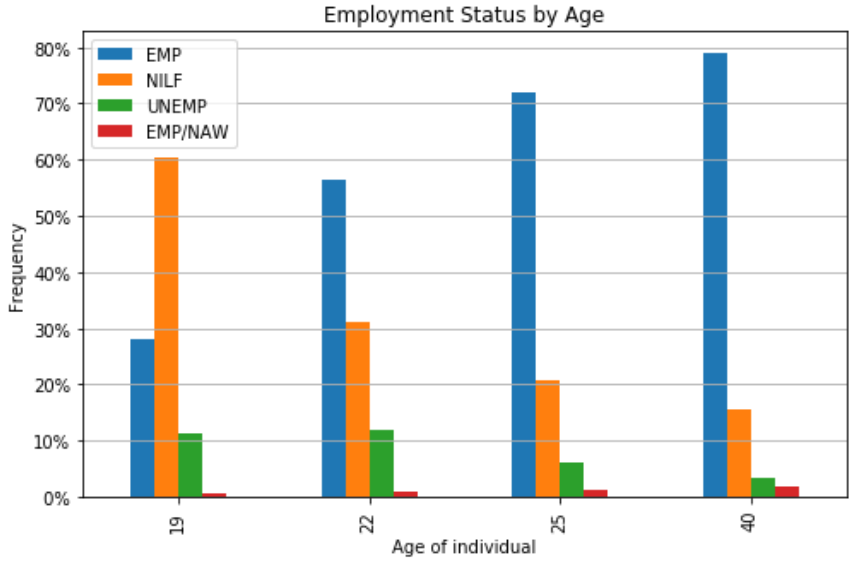


Mean NYCgov_Income levels are generally significantly different across age groups, with a few exceptions.

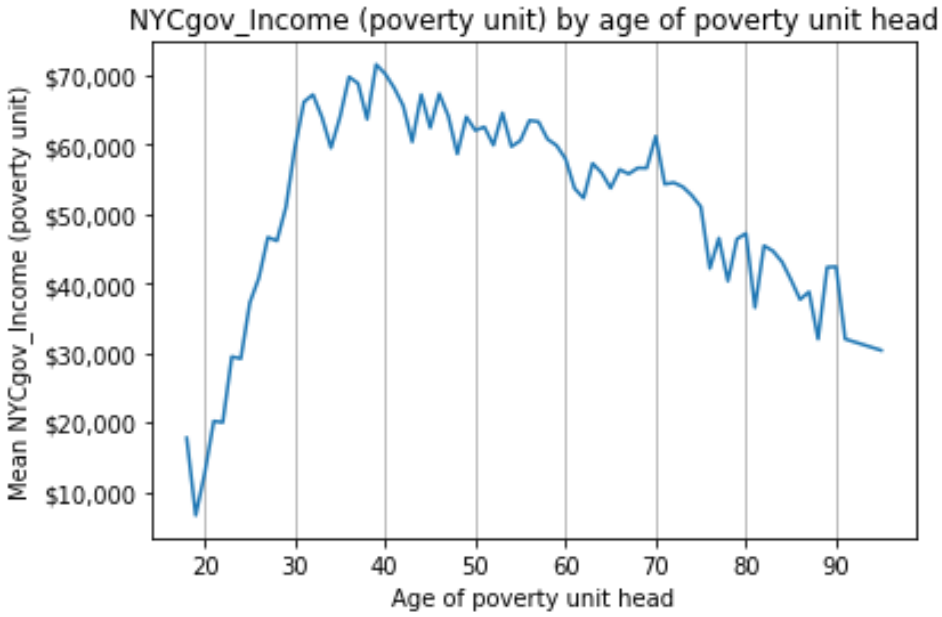
As seen below, households headed by 24-year-olds don't have significantly different mean NYCgov_Income than households headed by 23-year-olds, (p=0.136); but this is the exception, rather than the rule.

```
==============================================================================
Comparing age 24 to age 23:                          Test for equality of means
==============================================================================
                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
subset #1   -278.1040    186.483     -1.491      0.136    -643.610      87.403
==============================================================================
Comparing age 25 to age 24:                          Test for equality of means
==============================================================================
                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
subset #1   8072.7908    175.281     46.056      0.000    7729.242    8416.339
==============================================================================
Comparing age 26 to age 25:                          Test for equality of means
==============================================================================
                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
subset #1   3552.0452    177.841     19.973      0.000    3203.479    3900.611
```

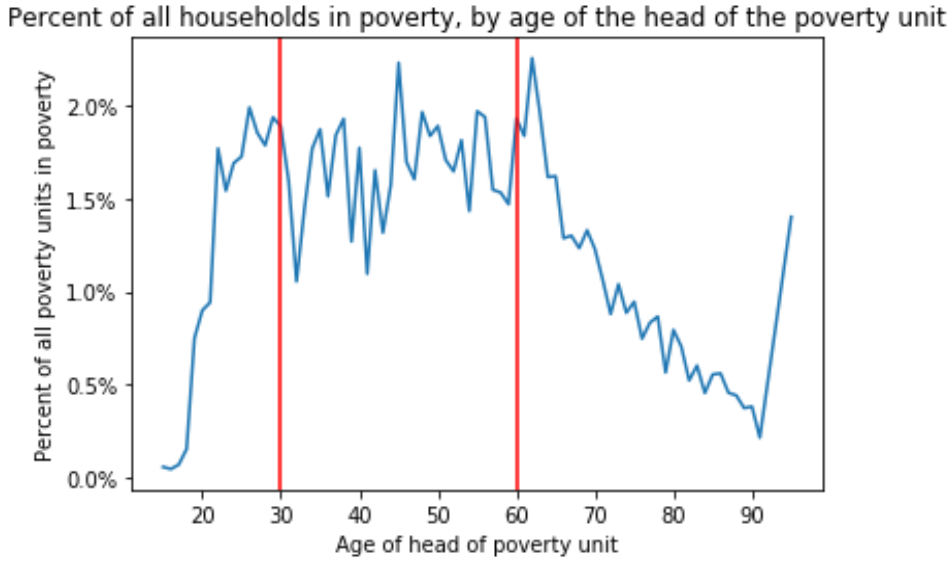
Young people's work hours are often lower due to school. Below, 60% of 19-year-olds are not in the labor force ('NILF'), but this quickly drops to 20% by age 25. (Age 40 is the rightmost column, for comparison to an older age group.)



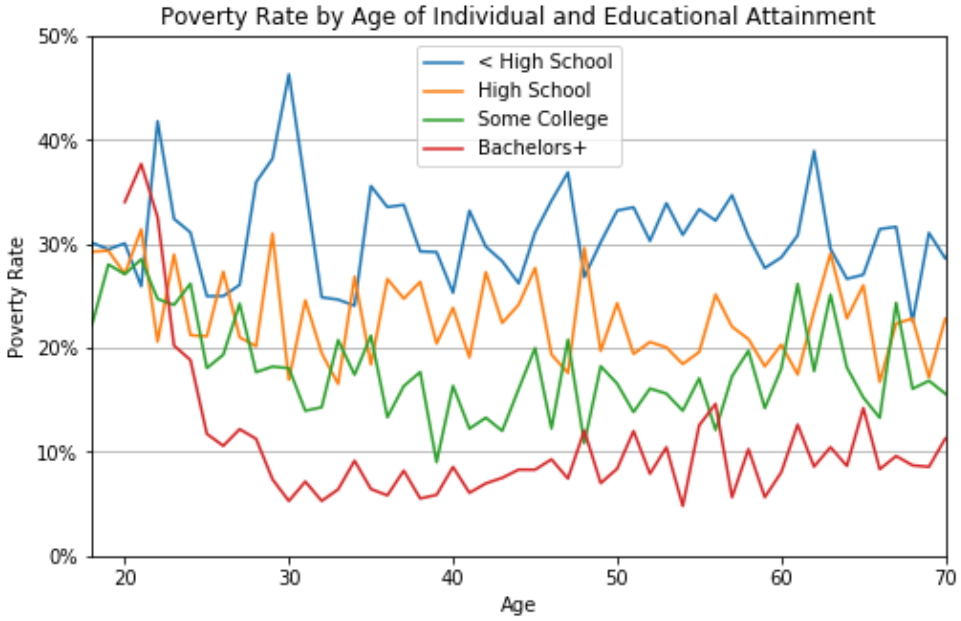
As older people reduce their work hours, their retirement and other benefits aren't enough to offset the shortfall in wages.



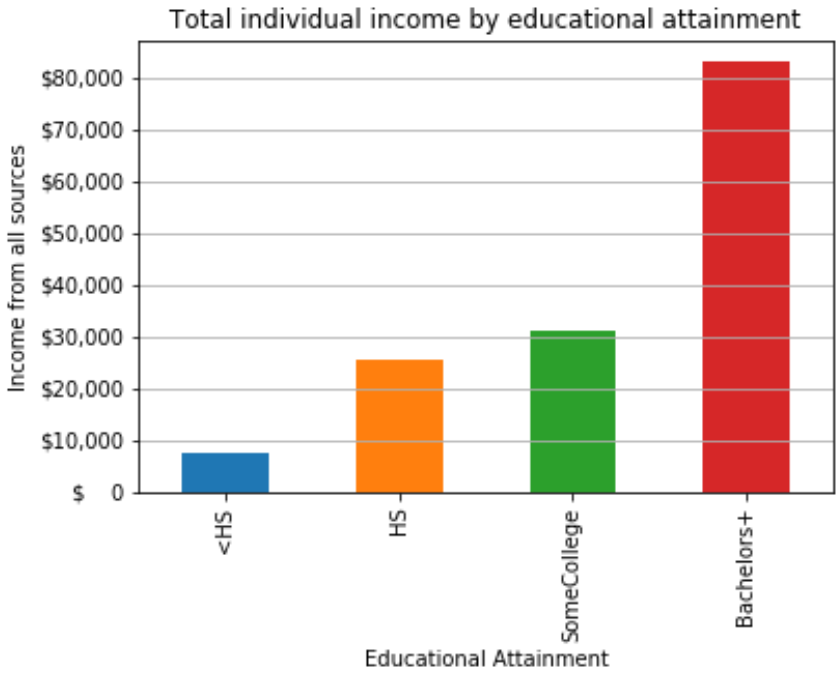
Although younger and older households have higher rates of poverty, they are smaller groups. Thus, 60% of households in poverty are headed by someone between the ages of 30 and 60:



Education levels have a strong impact on poverty rates:



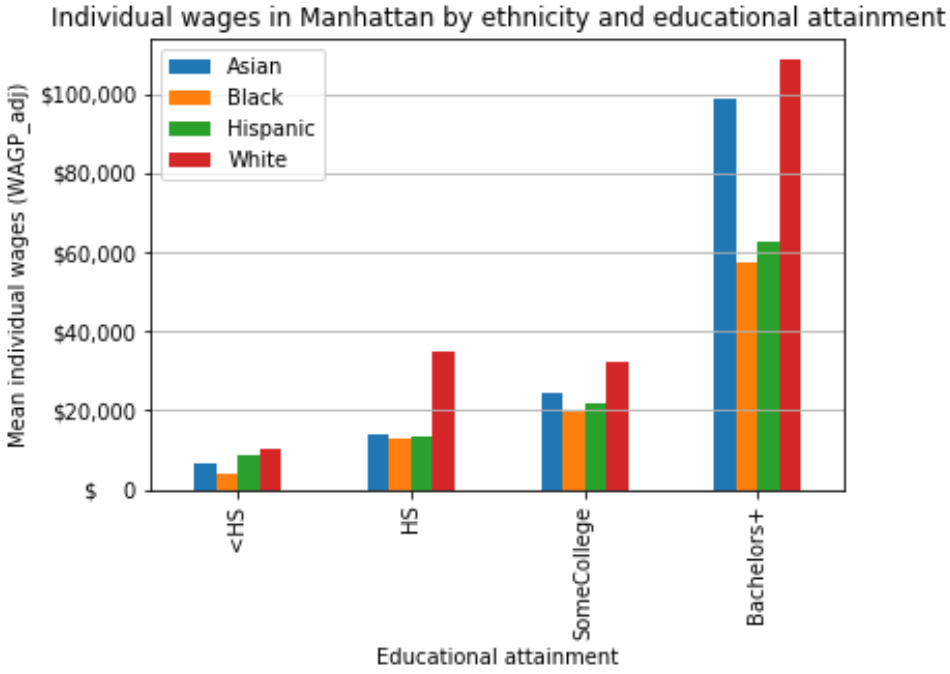
Income is strongly correlated with higher education.



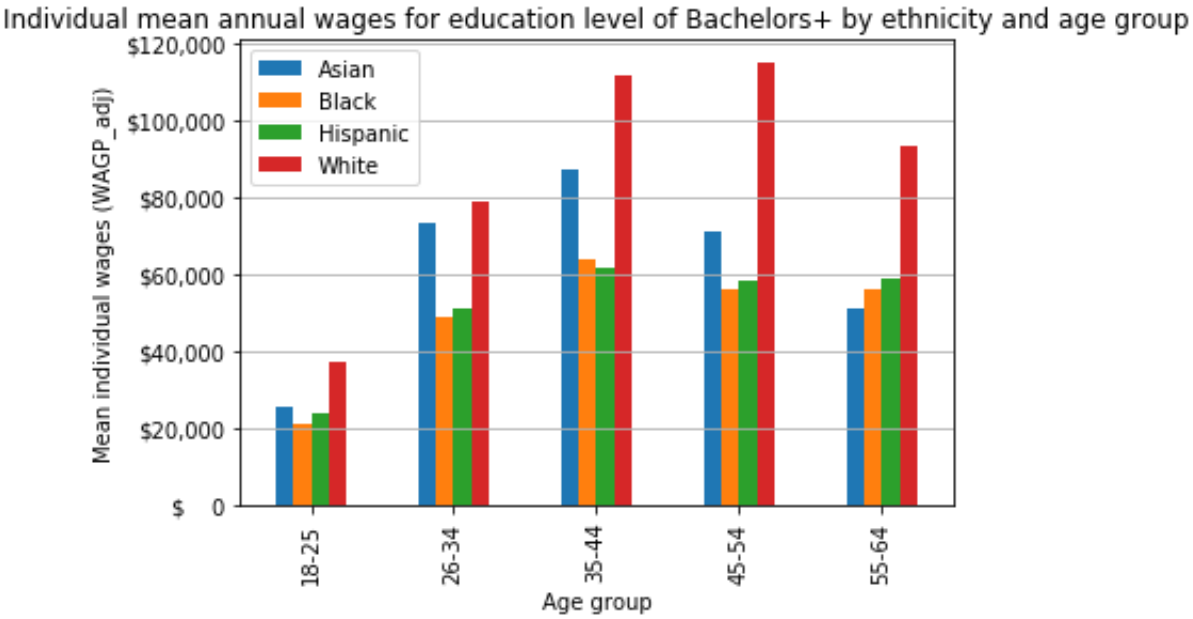
A test of means confirms that even the smallest difference above is significant (p-value of zero to three digits).

```
Comparing mean individual income of high school graduates to those with some college:
                          Test for equality of means
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
subset #1  -5932.0293     44.287   -133.945      0.000   -6018.830   -5845.228
==============================================================================
```

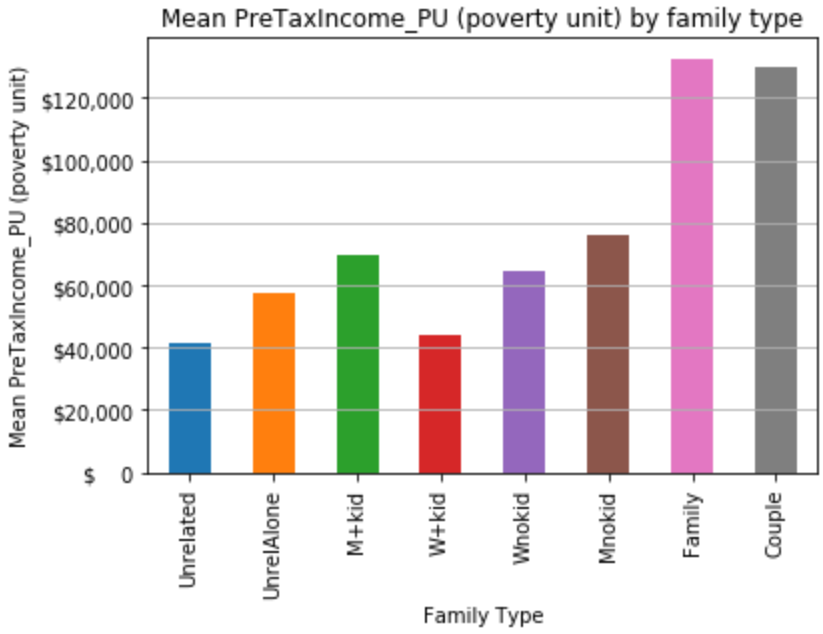
Education, ethnicity, and location correlate with individual wages in possibly-surprising ways.



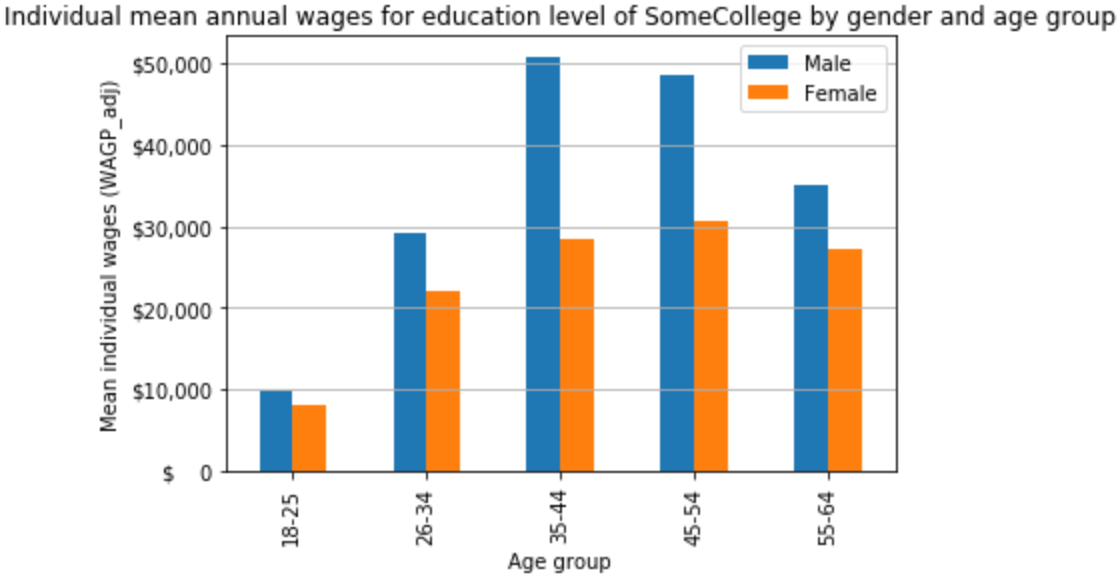
Even among college graduates, some apparently-large ethnic disparities exist across age groups (on the order of 25% or $20,000 annual salary). There are many possible confounding variables, many of which do not appear in the dataset - for example, what type of job an individual has.



Family type has a strong correlation with poverty, and even on pretax income. Interestingly, comparing single men with and without children, to single women with and without children, we see that single fathers average $6,000 less than single men without children; but single mothers average almost $22,000 less than single women without children.



At every education level, and in each working-age decade, men get paid more than women.

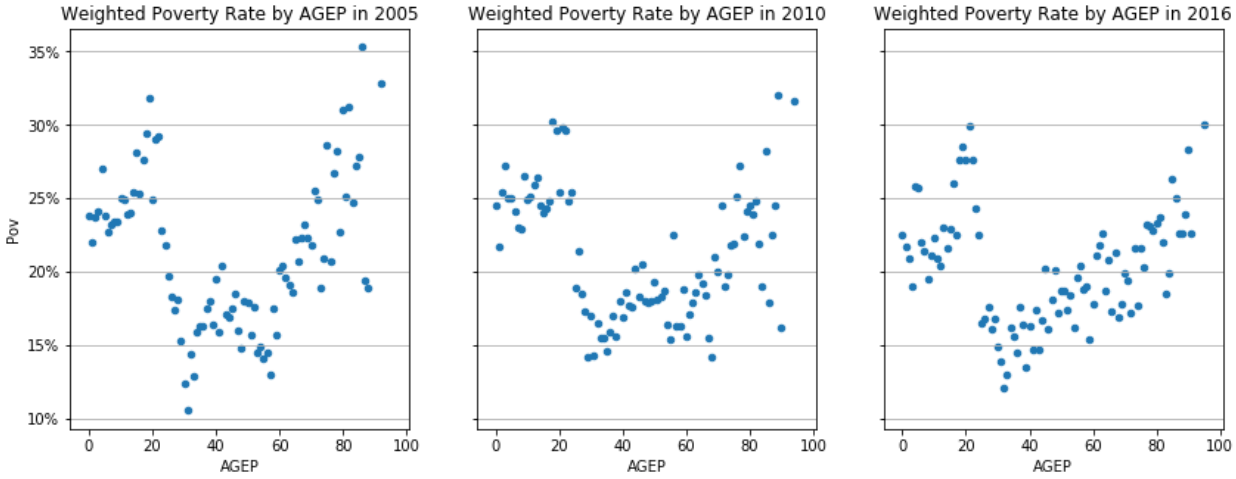


Even the smallest difference shown above between men's and women's salaries is significant, with p-value of zero to three digits. (Here 'statistical significance' means in the absence of confounding variables, which we cannot assume given the limits of our dataset.)

Comparing male to female aged 18-25 with some college:
Test for equality of means

	coef	std err	t	P>|t|	[0.025	0.975]
subset #1	5.6002	0.032	175.778	0.000	5.538	5.663

Poverty rates have changed from 2005 to 2016, but the overall structure and rates have stayed similar. Across the key measures, the structural similarities outweigh the differences, as evidenced by the poverty rate itself: 20.3% in 2005, and 19.5% in 2016.



4. Next Steps

This is an iterative process; as we continue with the analysis, these findings and highlights will be updated.