

## 1. Data Selection

Dataset: Cross-Sectional Child Income Statistics by College Tier and Parent Income Percentile

Dataset description: The data consists of 1500 rows divided into 100 blocks of 15 rows. Each block represents a parent's mean income percentile and is broken down into 15 different selectivity and type combinations. In total, there are 4 variables that represents parents' information and 11 variables that give information about their kids. All the data consists of numerical values and the selectivity and type combinations are coded into 15 classes (e.g. Ivy Plus, Highly selective public, Never attending college, etc).

Motivation: We chose the dataset "Cross-Sectional Child Income Statistics by College Tier and Parent Income Percentile" with the aim to explore how one generation's education and income impacts the next one's in the U.S. The conclusions drawn from this analysis will enable us to acquire a deeper view of cross-sectional mobility. Also, we believe that this dataset can lead us to further investigations about the U.S. educational system and if college opportunities are fair or biased. Finally, we chose this dataset as it provides more information on both generations simultaneously than the other dataset in the assignment 1.

## 2. Data Preparation

After exploring the data, we found that there were very few missing values and only numerical values (see appendix). Moreover, the data is structured by ascending percentiles and tier names. The tier names are encoded by numerical values and finally since the income represents averages, it is complicated to find outliers in the aggregated data. However, comparing the mean with the more robust median, it enables us to guess the presence of outliers in the sample. We concluded that this dataset is already suitable for visualization and does not need any data pre-processing.

As a first sanity check, we started by understanding what represent the data and each column of the dataset. We noticed that some data are fixed in every block, i.e. the `tot_count` column, the tier name and the tier values. After concluding that the dataset was clean and there was no need for data pre-processing, we will then perform the next sanity check by displaying graphs, such as bivariate scatter plot to see if there is any relationship between the columns or histograms to see the distributions of the income. With our final goal in mind, to use the relevant features to predict income variability of the second generation, we want to select the relevant features by looking at the linear dependencies among them. We notice the last column *density features* is calculated by the *count* over the *total count* feature and therefore it is linearly dependent on the count, which makes it unnecessary. All in all, reducing the number of variables enables us to visualize the data more clearly, illustrate and our sanity check, as it reduces the dimensionality.

As mentioned in the first assignment, the dataset about Cross-Sectional Statistics on Children's Income Distributions by College Tier found on the same [data repository platform](#), can provide further details about the future of the kids in our chosen dataset. It is a subset of the first dataset from the kids' perspective. It

enables us to complete our analysis of our chosen dataset. In particular, it helps us explore the relationship between the college type and income and to compare with the previous generation. However, we also notice that the latter dataset starts from the 9<sup>th</sup> percentile whereas our chosen dataset starts from the 1<sup>st</sup> one. Therefore, we need to explore if joining these two datasets is feasible while keeping consistency and cleanliness.

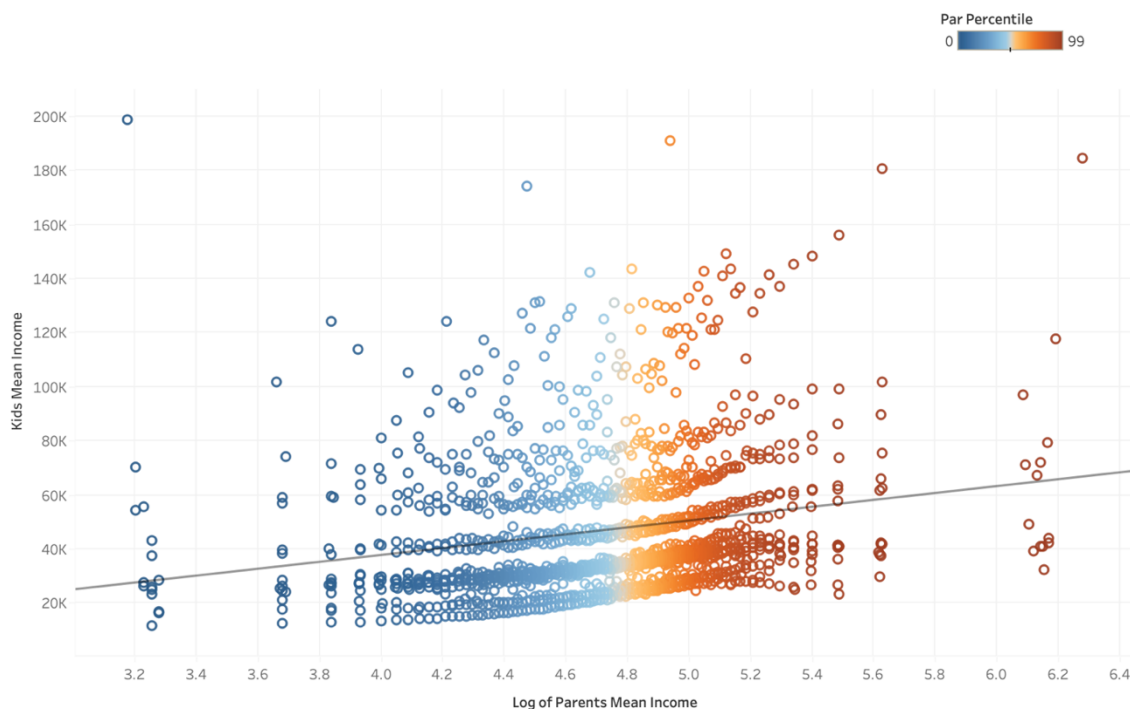
### 3. Research Questions

We are considering some of the follow Research Questions:

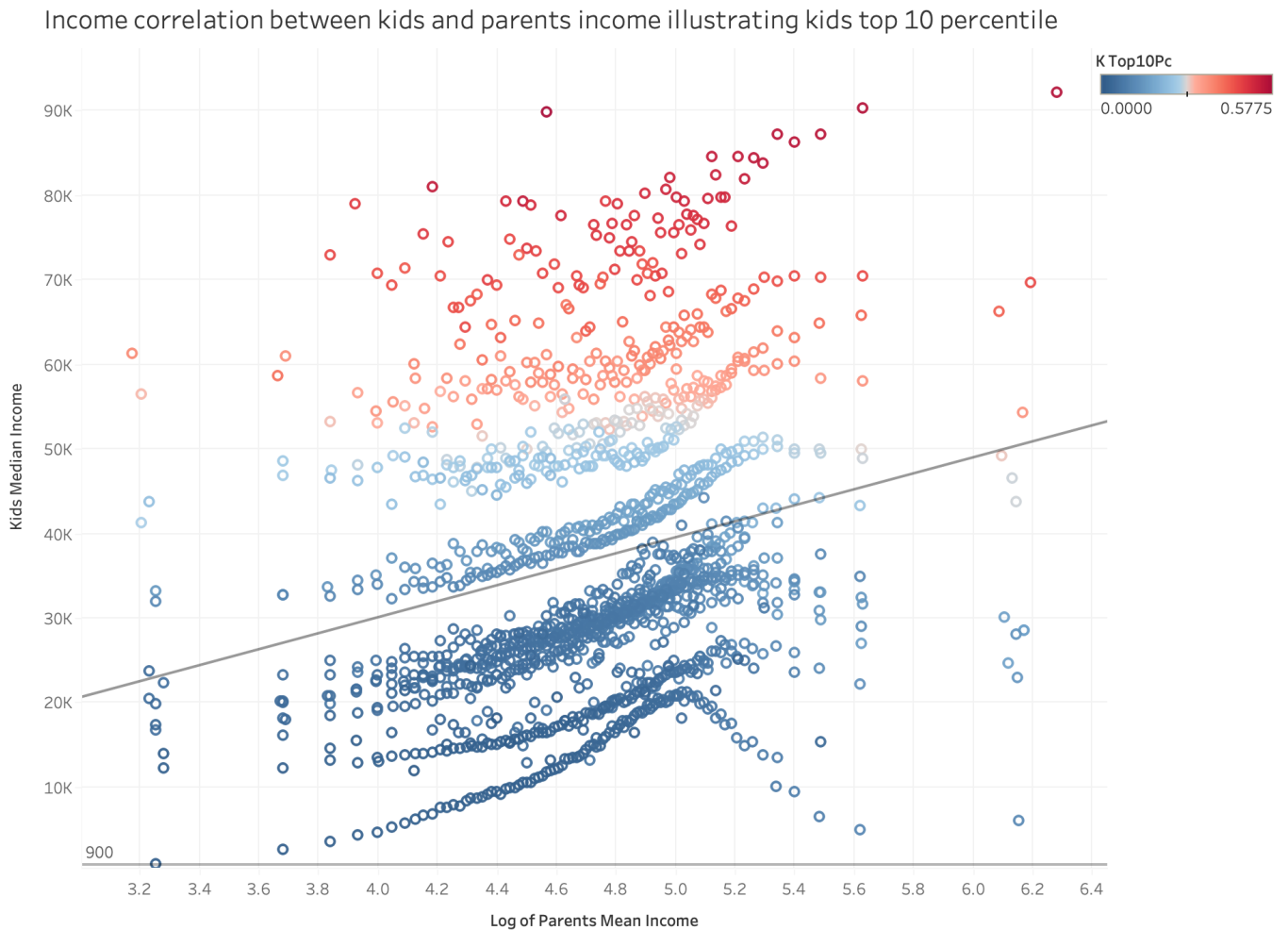
- What is the link a generation's income and education in their offspring's future income in the U.S.?
- What is the role of different colleges in inter-generational and economic mobility in U.S.?
- What may be the confounding factors and causalities of one generation on their offspring's financial situation?
- How fair are U.S. educational system and college opportunities based on one offspring's previous generation's income and education?
- What is the impact of “new money” vs “old money” in the offspring's educational and financial situation?

### 4. Data Visualization

Attending college with unsufficient data: Relationship between Parents Mean Income and Kids Median Income

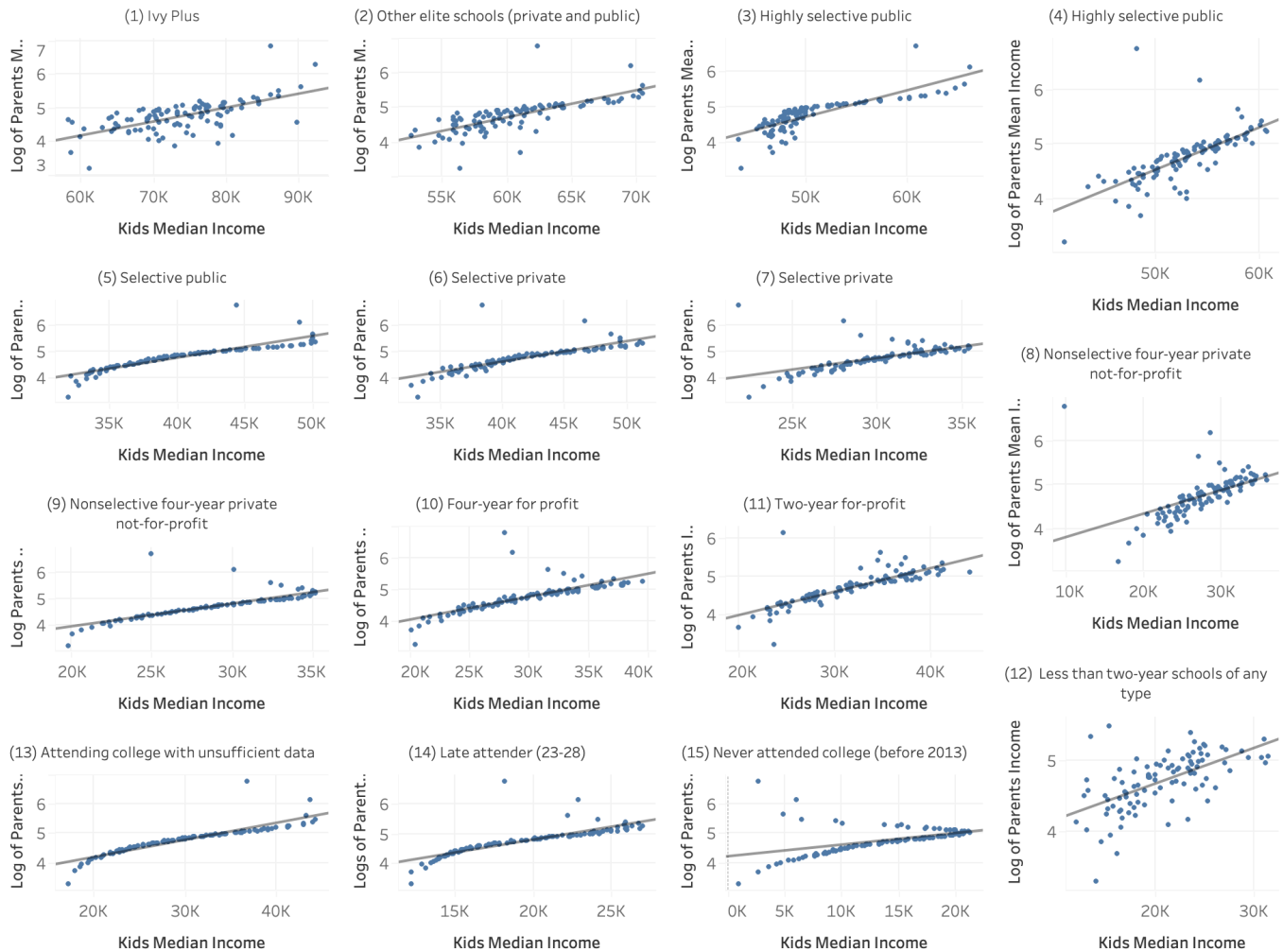


There seems to be a weak positive relationship between the parents and kids income mean. This is strongly influenced by the presence of outliers. Indeed, a possible explanation would be that the kids whose parents went to Ivy League schools, tend to earn a significantly higher salary than the kids whose parents are in the same earning's percentile.



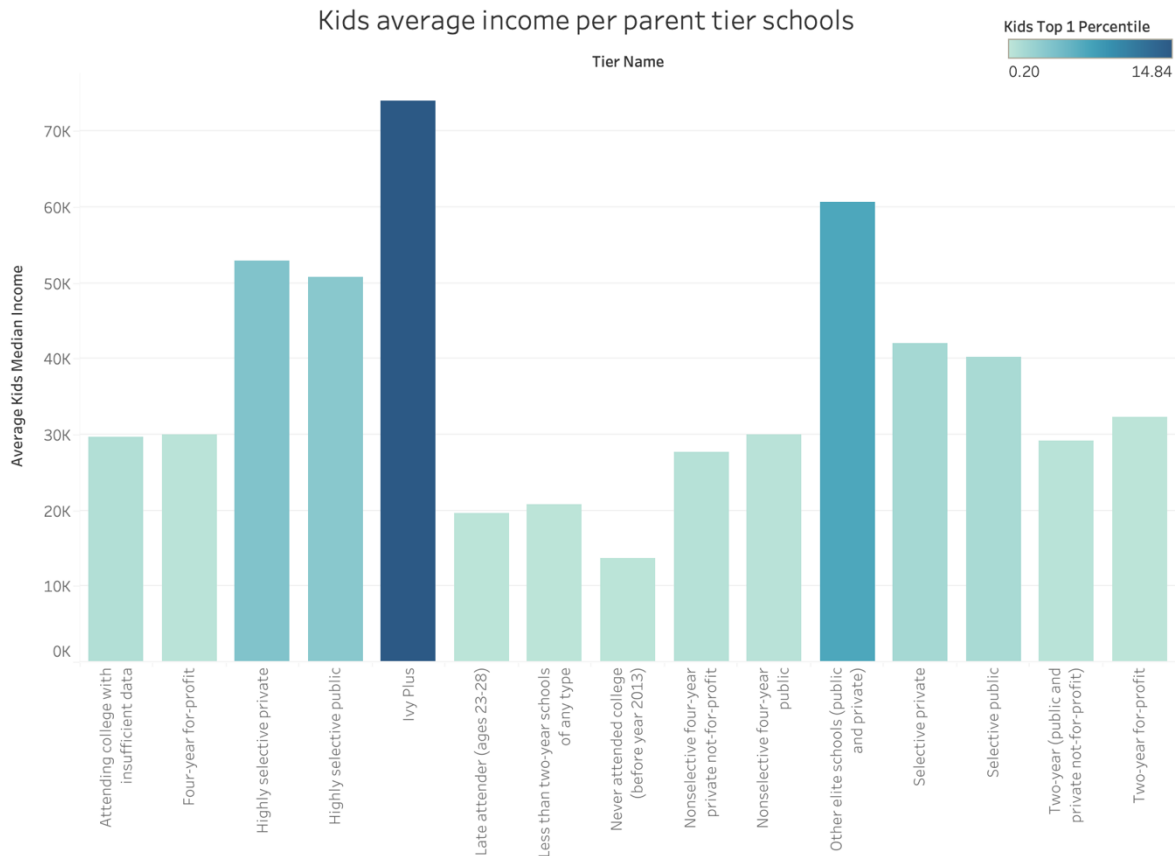
Here we highlight the presence of outliers among kids with high income whose parents earn the least. We further reduced the impact of outliers by using the median measure instead of the mean.

## Relationship between Parents Mean Income and Kids Median Income

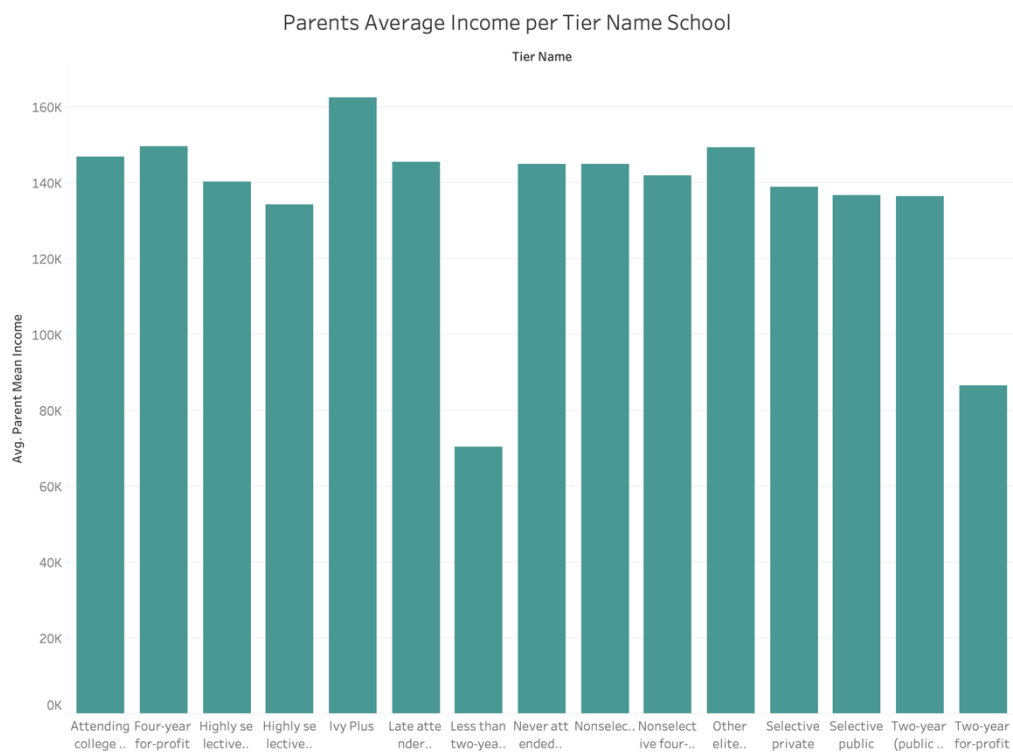


The scatter plots for each tier schools show a significantly stronger relationship between the parent and kid's income compared to the first scatter plot that did not take into account the parent's education background. This means that regardless of the parent's income, their education play a more important role in their kids future financial situation.

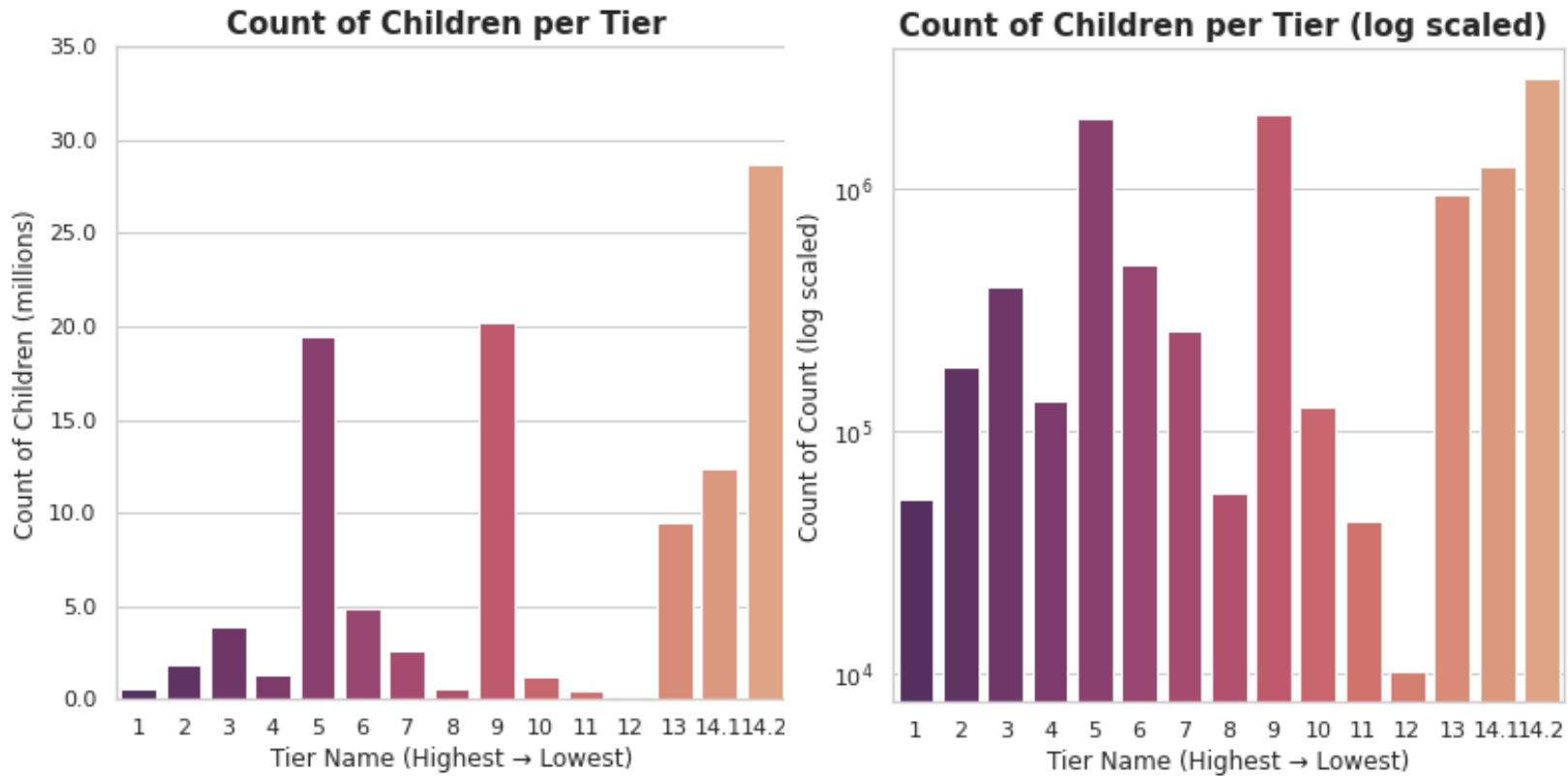
Assignment 2 – Exploratory Data Analysis  
HAJJI Alia, TAO Marinie and WILMET Vincent



There is an apparent link between kids financial situation and parents' education. This is well captured by the gap between the two highest earnings group whose parent went to elite schools and Ivy leagues school and two least earnings group whose parents never attended college or were late attender

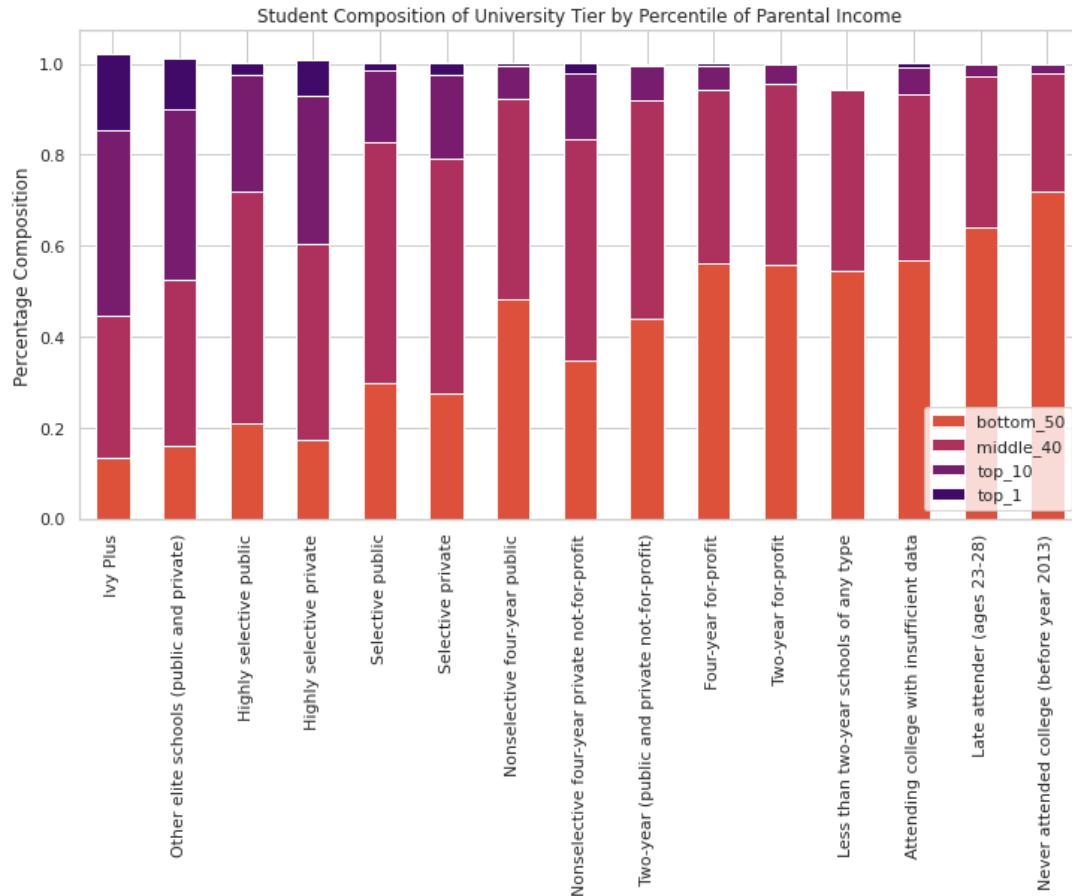


On average, we notice a constant amount of parent's income for each tier schools name. However, due to the unavailable data on the number of the parents, we cannot conclude our statement.



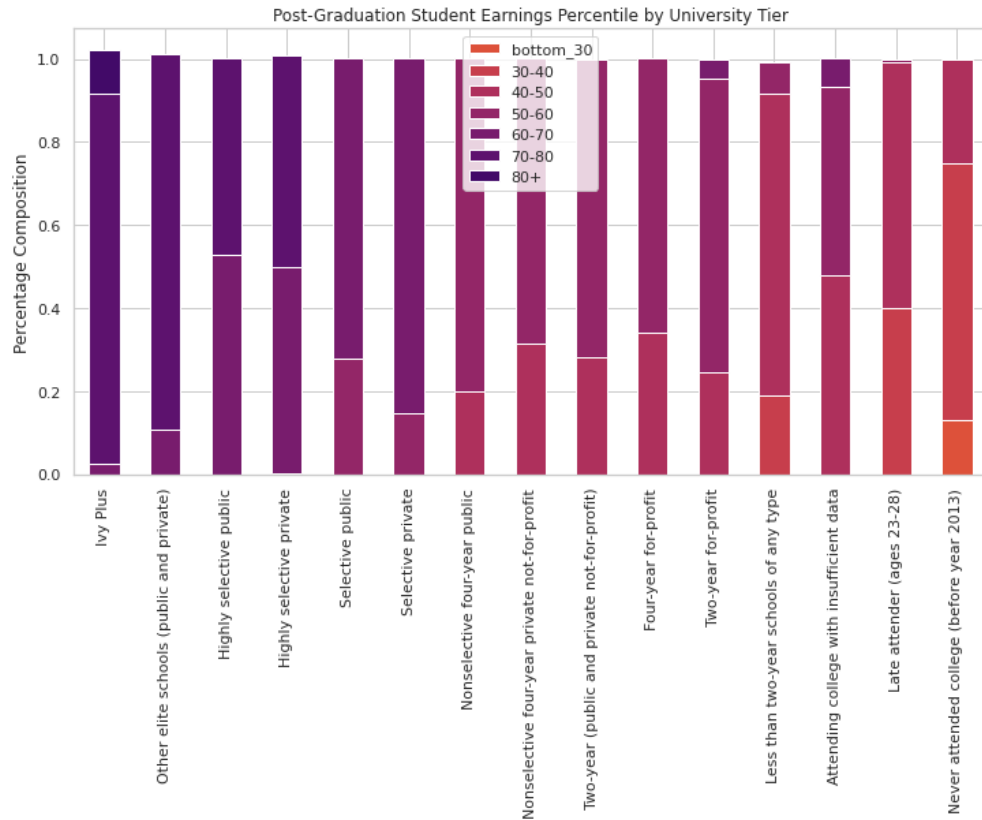
These bar charts look at the distribution of children/students/offspring in each tier of school. The first chart shows the data is heavily skewed towards tier 5, 9 and 14.2 (Selective Public, Two-year public and private not-for-profit, never attending college, respectively). This makes sense because public schools can host more students due to their government funding. It also makes sense that the highly selective and elite schools (1-4) have lower student counts as it is their brand to do so.

We wanted to look at this student distribution because it gives us an idea of what the structure of our data looks like when doing the essential exploratory data analysis (EDA) first steps. We can see that log scaling the data provides a better distribution for possible future analyses.



In this graph we see that nearly 60% of students in Ivy Plus schools come from parents in the top 10%. Parents with the middle 40% of income percentiles dominate the medium selective schools. The children of parents in the bottom 50% dominate in two or less years of higher education, or never attend college at all.

We can infer from this graphic that the percentile of parental income may be strongly linked to the quality of school the children go to. It also can explain the school's preference in students.





After graduating, we can see that 95% of Ivy Plus students earn in the top 70%+ income percentile. Juxtaposed to the other extreme, everyone who never attended college will never make more than the bottom 50% of income percentiles.

This further validates the criticism of wealth inequality where those with higher educations (especially highly selective schools) will earn more money.

However, it can be said that if you are in the 15% of students from bottom 50% backgrounds who make it to the Ivy League, there is clear upward mobility.



## 5. Appendix

 dataset.isna().sum()	[221] dataset.info(verbose=True)
 par_pctile 0 tier 0 tier_name 0 par_mean 4 k_mean 4 k_rank 4 k_top1pc 4 k_top5pc 4 k_top10pc 4 k_q5 4 k_q4 4 k_q3 4 k_q2 4 k_q1 4 k_nowork 4 married 4 k_median 4 k_median_nozero 10 count 4 tot_count 4 density 4 dtype: int64	<class 'pandas.core.frame.DataFrame'> RangeIndex: 1515 entries, 0 to 1514 Data columns (total 21 columns): # Column Non-Null Count Dtype --- 0 par_pctile 1515 non-null float64 1 tier 1515 non-null float64 2 tier_name 1515 non-null object 3 par_mean 1511 non-null float64 4 k_mean 1511 non-null float64 5 k_rank 1511 non-null float64 6 k_top1pc 1511 non-null float64 7 k_top5pc 1511 non-null float64 8 k_top10pc 1511 non-null float64 9 k_q5 1511 non-null float64 10 k_q4 1511 non-null float64 11 k_q3 1511 non-null float64 12 k_q2 1511 non-null float64 13 k_q1 1511 non-null float64 14 k_nowork 1511 non-null float64 15 married 1511 non-null float64 16 k_median 1511 non-null float64 17 k_median_nozero 1505 non-null float64 18 count 1511 non-null float64 19 tot_count 1511 non-null float64 20 density 1511 non-null float64 dtypes: float64(20), object(1) memory usage: 248.7+ KB