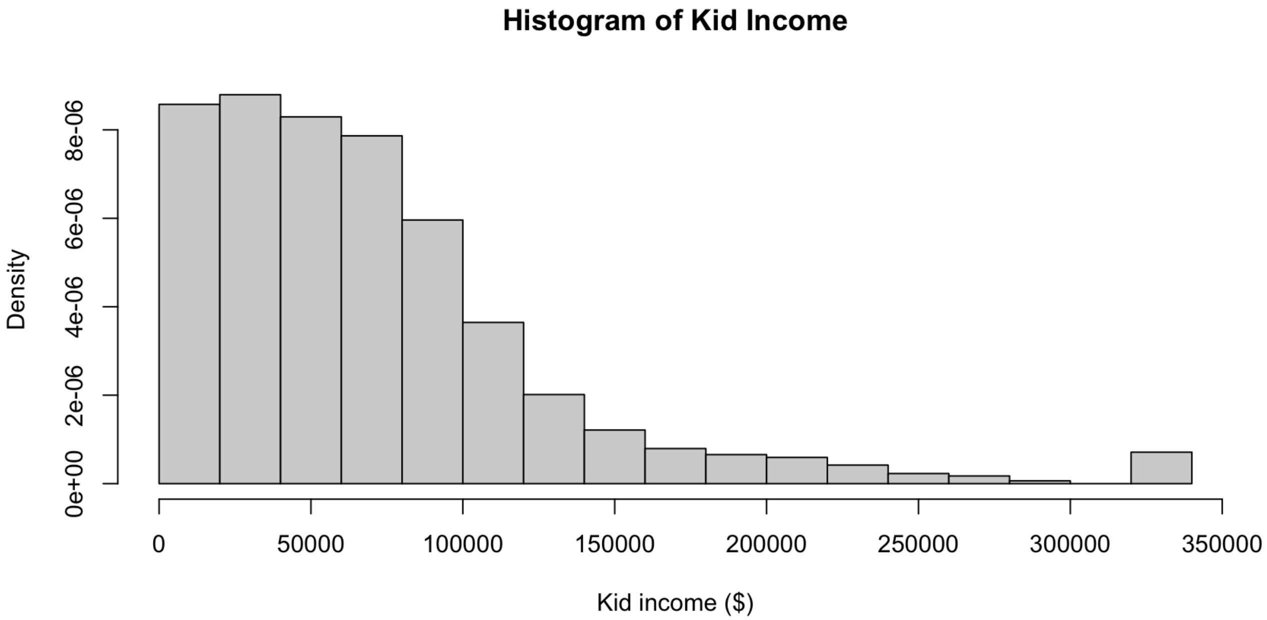
Jared Ni
Econ 50 Lab 1

1. Histogram of kid_income:



2. Sample mean kid income of the sample is $70499.94

3. What fraction of observations have kid_income below its mean?
    a. New variable called below_mean that equals 1 if kid_income is strictly less than its mean, and 0 if it's greater than or equal to its mean
    b. Sample mean of below_mean is 0.596.
    c. The histogram is greatly skewed toward the left. Only a small percentage of kids have income above the sample mean, and their wealth has driven up the sample mean. A much higher percentage of kids (59.6%) have income below mean.

4. The median kid income of the sample is $58750.

5. The sample standard deviation of kid income is $59552.

6. Indicator variables
    a. The fraction of observations are within one SD of the mean kid_income is 0.787.
    b. The fraction of observations are within two SD of the mean kid_income is 0.949.

7. Percentile rank transformation.
    a. I used the rank() function to generate a new variable called kid_income_rank that equals each observation's rank based on kid_income.

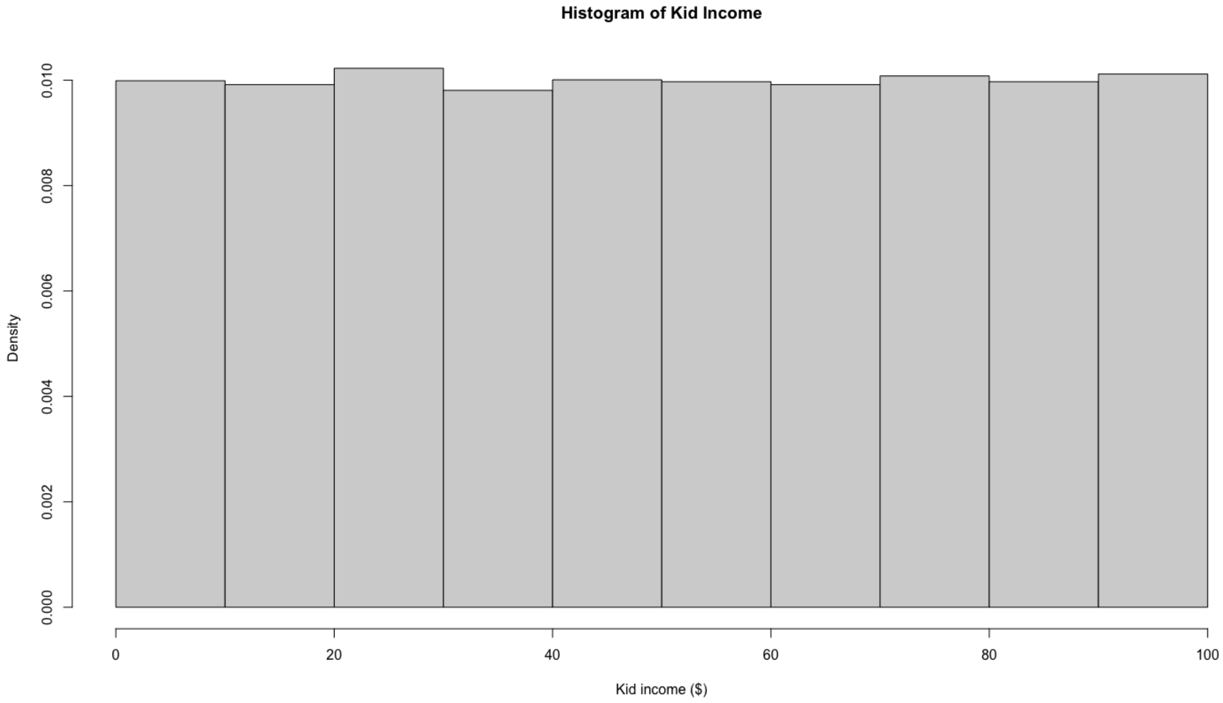
b. I sorted the data by kid_income. The top (beginning) of the data frame corresponds to the very left of the kid income ranking, where the ranking is dismally low. The right (end) of the data frame corresponds to the very right of the kid income ranking, where the rankings are high. The bottom 10 ranks in the sorted (ascending) kid_income_rank are 74.5, which corresponds to the least child earning data points in kid_income, kids who earn $0 (very left of the original histogram of kid_income) and the top 10 ranks in the sorted kid_income_rank is 5477.5, which corresponds to the the greatest child earning data points in kid_income, kids who earn $329331 (very right of the histogram of kid_income).

c. I normalized the ranks and stored the data in kid_inc_rank.

d. After normalization of the kid income rank, the 74.5 (very bottom of the ranking) in the original kid_income_rank variable became 1.36011, corresponding to an earning of $0 in kid_income; and the 5477.5 (very top of the ranking) became 100, corresponding to an earning of $329331 in kid_income.

8. Plot histogram of kid_inc_rank.

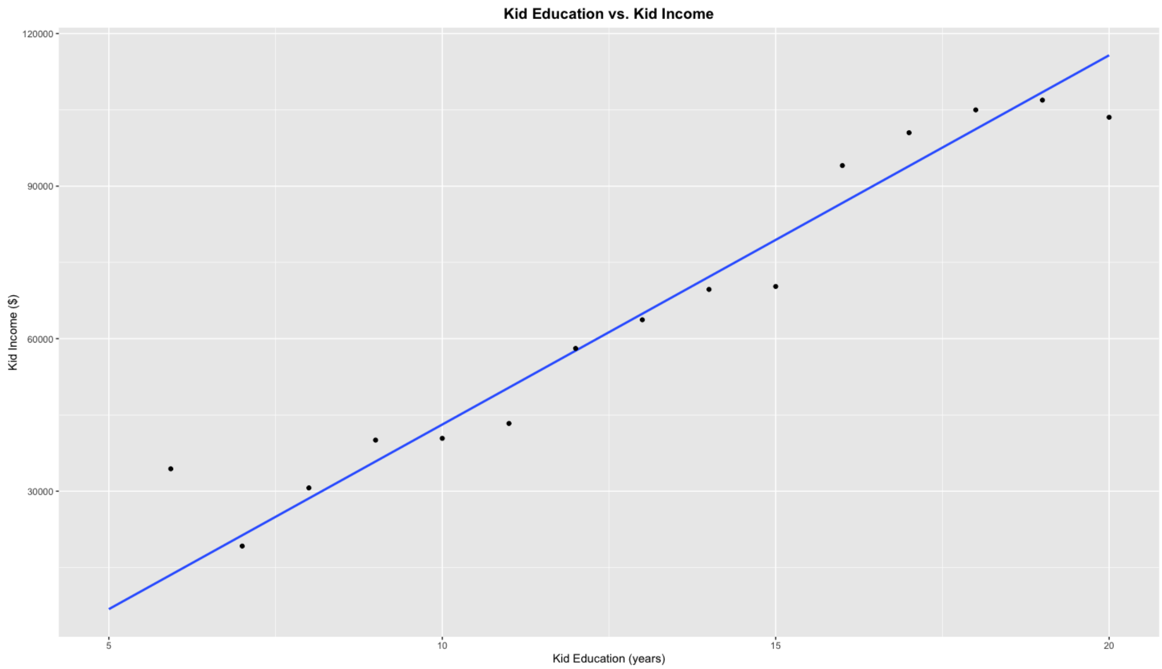
   a. Histogram of rank:



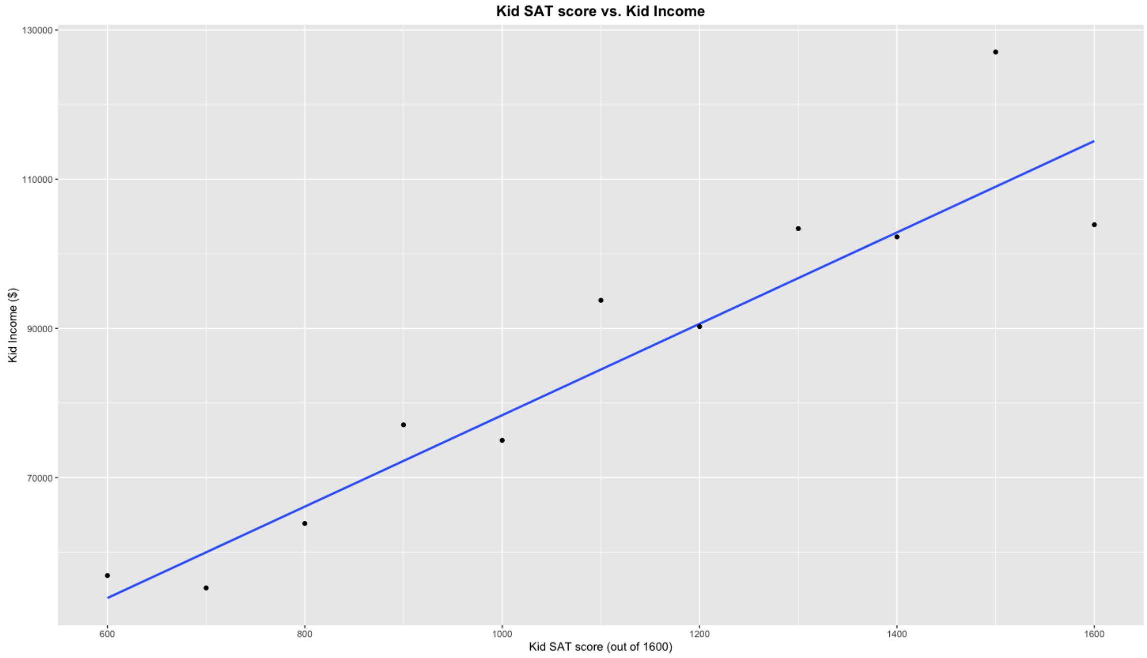
   b. Sample mean of kid_inc_rank is 50.087, and sample median is 50.1.

9. Binned scatter plots

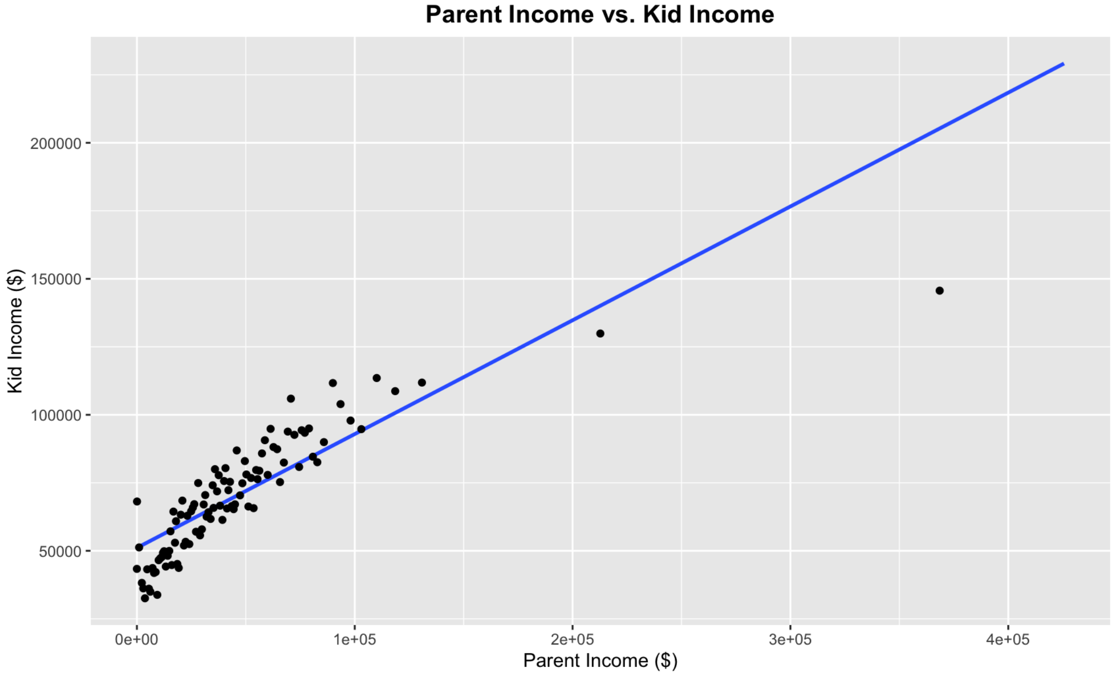
A convincing linear relationship I found is between the kid's years of education and the kid's income. Higher years of education that a kid receives is directly correlated with high kid income, while low years of education is directly correlated with lower income.



Another convincing linear relationship I found is between kid SAT score and kid income. High SAT score a kid receives is correlated with high income in the future, while lower SAT score is correlated with lower income in the future.



A non-linear relationship I found among this dataset is between parent income and kid income. On the binned scatter plot divided into 100 percentiles (n = 100), we can see that for lower levels of parent income, kid income grew at a faster rate than it does for higher levels of parent income. Toward the end, the growth of kid income dips below the linear growth of parent income of the linear regression by large margins. This graph is more resembling a logarithmic relationship, thus I argue that the relationship between parent and child income in this data set is non-linear.



10. Randomized Experiment
    a. Entered seed number
    b. Generated a new variable treatment_group. There are 2739 observations in my controlled group, and 2747 observations in my treatment group.
    c. Computed the mean and sample standard deviation for all variables separately for observations in treatment and control groups.
    d. Filled google form.
    e. Randomized experiments help us eliminate bias within the dataset and establish the proper relationship between the two sets of data we are trying to measure the effect of, or how one variable affects another. Randomized experiment eliminates the causal effects in the differences between the different observations caused by their conditions, such as the neighborhood the people of the experiment were observed in. Randomization eliminates all such differences and makes the two groups perfectly comparable. On the other hand, human judgements are flawed more often than not, even when we don’t realize. Random assignment is not perfect, but it is fair because it is randomized and decided by a randomized algorithm, so it's reliable from time to time, whereas humans are subject to implicit bias from all kinds of sources, even if we like to think that we are not. With human judgements, observations studied are more prone to causal effects. Since we don’t want human judgment to impair the validity of the data we are testing, we turn to randomized experiments. Using random assignment ensures the fairness of our experiment and simulates a truly random outcome that mimics the real world outcome, thus we should use random assignments.