R Coursework – College Distance Dataset

Dataset used was “College Distance” [https://vincentarelbundock.github.io/Rdatasets/doc/AER/CollegeDistance.html] which includes cross-section data from the High School and Beyond survey that roughly surveyed 1,100 high schools, accumulating 4,739 students, observed across 14 variables.

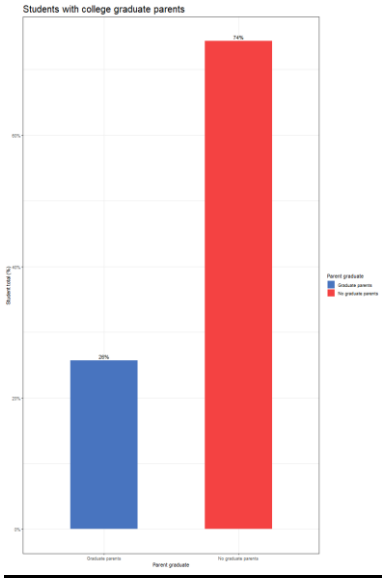
I cleansed the data by:

1. Dropping potential NA values.
2. Dropping the first column “rownames” as values were inconsistent when accessing the CSV file, e.g., row number beyond 13,000 which is impossible because the dataset has less than 5,000 entries. I then
3. Reformatted each string value entry for every variable to provide my visuals with formally presented labels.
4. Some visuals had removed outliers

Visuals generated with R including an interactable version in Observable with alternative visuals. [https://observablehq.com/d/3ffc0cc05fc8d1fc]

Q1. How many of the students have college graduate parents in comparison to those that do not?

I want to first understand how many students have parents that are college graduates. The data provides graduate parents individually as either mother or father. I combined the two using a conditional to differentiate between students that have and do not have any graduate parents. This will provide convenience for later research. Thus, a mutation was applied with a new column titled ‘pgraduate’ with two variable values: “Graduate parents” and “No graduate parents”. A bar plot was suitable to visualise the difference.



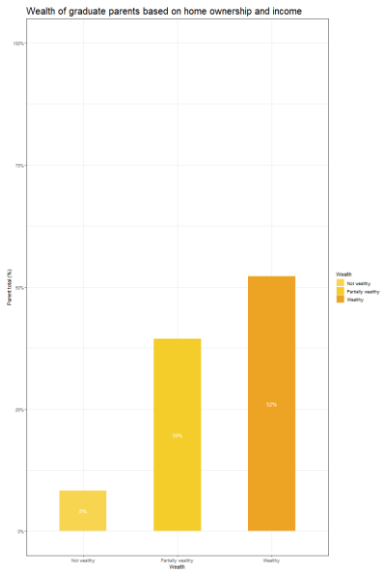
I was able to visualise the proportional differences between students with and without graduate parents, establishing that only 26% of students have either a mother, father or both parents that are college graduates in contrast to the 74% that do not. There is a large sample size difference between graduates and non-graduates, and so to build off of this visualisation, I would like to investigate

students belonging to each value to determine any notable differences between them, as to then understand whether a majority of students, those without graduate parents, are at an advantage, disadvantage or on par to the minority.

A bar chart was suitable given its ability to compare different magnitudes between different values with the X axis and fill value being the parent graduate and Y axis being the total count as this visualisation is only concerned with univariate data, providing a good understanding of their contrasting sample size. I used and will continue to use percentage values to standardise the data. They act as a suitable format because they can assist with comprehension of large sample sizes in proportion to other samples of data, as opposed to using raw values. They appear as labels on the bars which I will maintain in bar plot visuals to enhance the viewing experience as they provide the exact percentages of samples rather than relying on imagery to understand their proportions. Blue and red colour encodings were used because they contrast well, establishing further individuality and clarity between non-graduates and graduates, something I will reuse when reassessing the two for consistency. “Theme_bw” will continue to be used for all my visuals because it is simple and clear with an outline of the graph as well as its axes’ points and labels.

Q2. How many families with college graduate parents have an income higher than $25,000 and own their homes?

I want to assess their distribution of wealth as a hypothesis I carry is that most graduate parents should be wealthy given their prospects, which is a factor that could affect scores. I can get a rough idea through two variables the dataset provides which is their income status that is either less than or greater than $25,000 and whether they own their home, meaning that someone can be considered wealthy if they can afford it and have a high salary. I formed three values under a “wealth” category: “Not wealthy” for those that do not own their homes and have less than $25,000, “Wealthy” which is for families with both and “Partially wealthy” as a mediator for either case being true. A bar plot was again suitable.



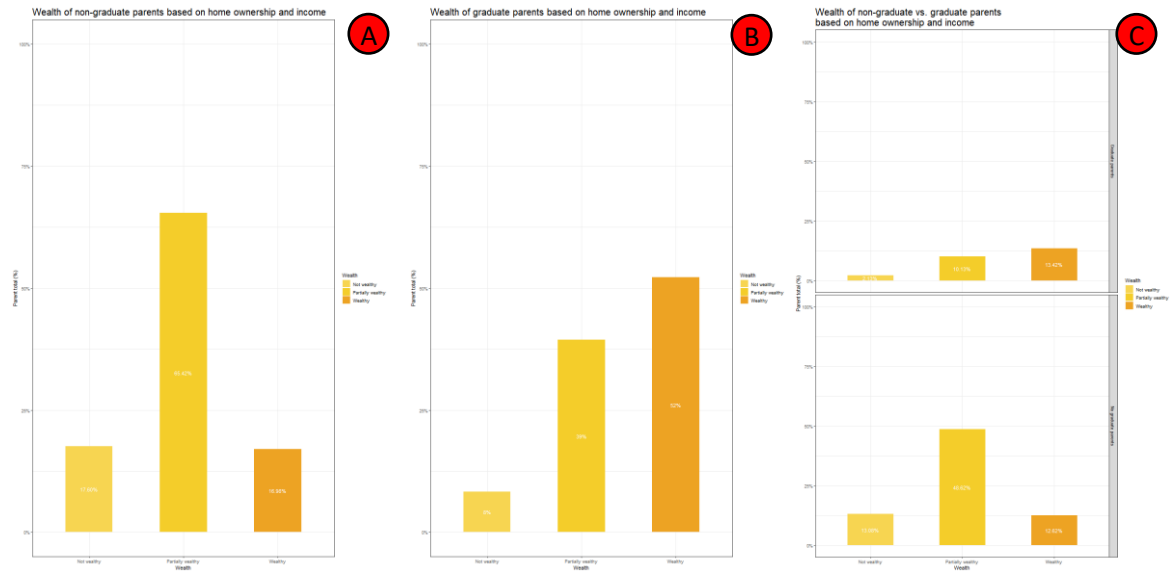
There is a clear display of how well-off graduates are given that only 8% have less than $25,000 and do not own their homes while a majority have some form of wealth, more particularly half of them being wealthy. Slight disparity occurs between the wealthy and partially wealthy, opposed to the

larger difference to the not wealthy. The differences lead into a rise within the data, accentuating the density of graduates that have some form of wealth.

A bar plot with percentage values was suitable to compare the sample sizes of the statuses, like the visualisation of graduates vs. non-graduates. Univariate visualisation was required again as the X axis and fill were the statuses and the Y axis the total count. The colour encoding was based on the common colour associated with wealth, gold, as humans connotate the two, enhancing clarity. A rising gradient is used to represent a sequence of the statuses, following Gestalt's principle of similarity, enabling easier comparisons as they relate but are individualised with a hierarchy.

Q3. How do families with no graduate parents compare? Is there a discrepancy in wealth?

This comparison will help to highlight the differences between families with and without graduate parents by carrying out a side-by-side analysis of each of their wealth status distributions. Both plots will be present, alongside a plot with faceting to show a direct comparison and how the sample size of each status appears against the entire dataset.



Notably, wealthy non-graduates account for less than a fifth of their sample size, whereas more than half of the sample of graduates in graph B are wealthy, insinuating how less likely it is for non-graduates to be wealthy. They also have more cases distributed within the not wealthy status. A majority of non-graduates, 65.42% to be precise, are partially wealthy in contrast to the 39% distribution for graduates.

This emphasises how different the wealth is distributed amongst graduates and non-graduates as a majority of graduates are likelier to be partially wealthy or wealthy as opposed to non-graduates where their distribution leans more towards being partially wealthy with an almost identical distribution between the not wealthy and wealthy, suggesting there are similar chances to be of either status but are still relatively low.

Assessing the facetted variant, the smallest group in the entire dataset are the graduate parents that are not wealthy whereas the biggest is the partially wealthy non-graduates, and that wealthy graduates and non-graduates have equal distributions. These differences highlight how sparse the

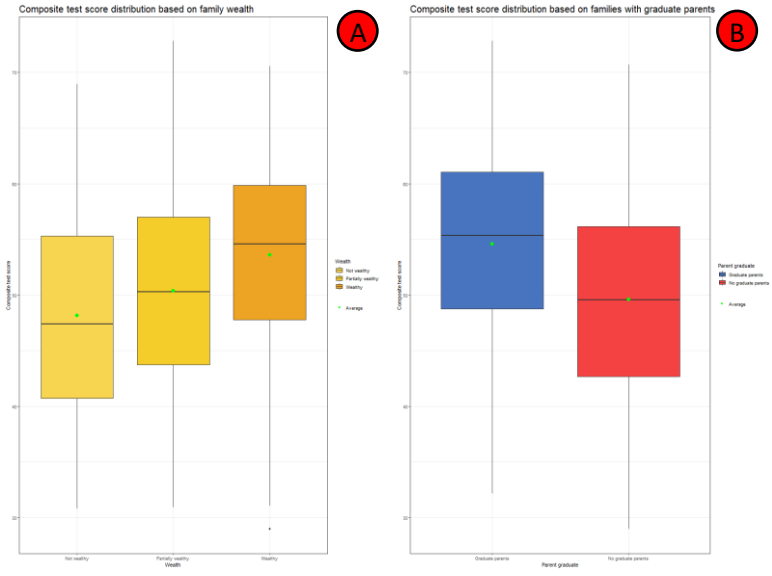
variables are among the samples as a very small percentage account for the wealthy and/or have graduate parents. This will be considered later.

To conclude, having graduate parents enables some leverage over students without them. They are expected to have stabler households as it is common for them to own their homes or have an income greater than $25,000, but even more common to have both. In contrast, families with non-graduate parents commonly have one or the other, while having a smaller, yet equal chance of being not wealthy or wealthy. It can be inferred that students with graduate parents will have easier access to resources and opportunities due to the likelihood of having wealth, meaning greater chances in achieving higher test scores. This insight would help to procure my next question.

The same encoding philosophies were used to create the non-graduate graph, keeping my visuals consistent such as the max percentage for equal comparison. A facetted version of the graphs was included to assess sample sizes against the whole dataset with maintained encodings for consistency.

Q4. To what extent does this difference in parent education and family wealth affect their children's composite scores?

Continuing from the previous research, I want to finally assess the extent to which wealth and graduate statuses affect composite test scores, as my hypothesis believed these to be factors that could affect it. This can be verified through a box plot and using the "score" variable to test against what I previously collected.



Graph A shows clear differences in scores across the different statuses with each half of the status' score distribution and median rising in tandem with wealth, implying that wealth has a subtle effect on scores. Further cementing this idea, their averages were taken, providing clearer indications of wealth affecting scores as each average rises with the status. Interestingly, partially wealthy students have a median and average score that are almost similar including a student(s) that achieved the highest score, suggesting that scores are evenly distributed, and that additional wealth would not improve performance by a substantial amount. Non-wealthy students have fewer outliers towards the upper whisker, contrasting wealthy students as it is the opposite, implying that wealth may not affect those students. These could be students that are naturally gifted, study harder or have

graduate parents that could help them with their education. Distributions overall are quite similar, but the wealthy status has a slightly smaller box box and a shorter upper whisker with an average lower than its median, implying students commonly achieve lower than average, insinuating students with wealthy backgrounds will still have to put in effort towards the resources and opportunities they have access to.

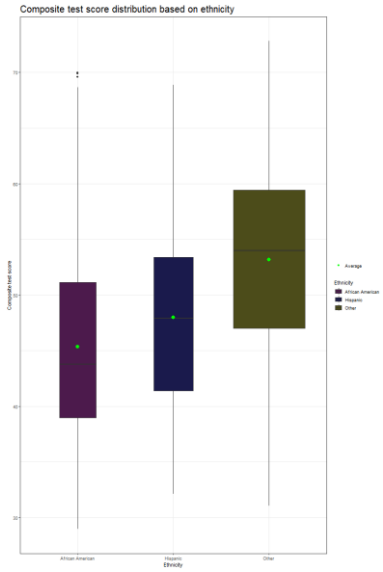
For graph B, students with graduate parents usually score higher than those who do not, with distribution similar to the wealthy and partially wealthy statuses, suggesting that graduate families with wealth have a greater distribution of students that achieve higher scores. For families without graduates, their median and average align with no skewing, with distribution falling around the partially and not wealthy statuses, implying that scores are more stable but usually lower. Students of graduates however have slight skewing, meaning there are small outliers with lower scores than usual. Notably, the graduate's distribution has a student(s) that achieved the highest score whereas the opposite applies to the distribution of non-graduates, reinforcing the imbalance that parent education may bring to one's studies.

In conclusion, it is evident that graduate parents and wealth statuses can influence education, with students that have graduate parents and/or greater wealth being likelier to achieve higher scores. Based on my facetted visualisation, this means that a small number of students have an advantage over the majority. However, the differences are not very extreme, given their overall distributions being close, although less fortunate students are still more likely to achieve less than students that are more fortunate based on the 50% distribution and averages. Students that have neither may struggle when attempting to achieve higher scores as the visuals support the idea that having graduate parents correlates with wealth and vice versa.

Colour encodings were reused from previous visuals to maintain consistency and immediate comprehension of variables. A box plot was ideal for comparisons of their distribution, averages, medians, outliers etc. and identify any trends or significant points of interest. I used stat_summary() to create a mean, coloured green to make it stand out clearly, to support the identification of aforementioned trends and comparisons to the median to understand distributions more in-depth. My data needed a bivariate graph which it was suitable for, with the Y axis now being the variable "score".

Q5. How do test composite scores vary amongst the different ethnicity groups?

Continuing to assess factors affecting composite test scores, I decided to study the ethnicities as the dataset provides two underrepresented ethnicities, "Hispanic" and "African American", with other ethnicities labelled "Other". Ethnic minorities struggle because of the issues they face like racism, accessibility, discrimination etc., so I would like to assess if there are additional issues regarding education by assessing score distributions. If there are notable differences between them, then I will expand upon this. I will use a boxplot again to assess score distributions.

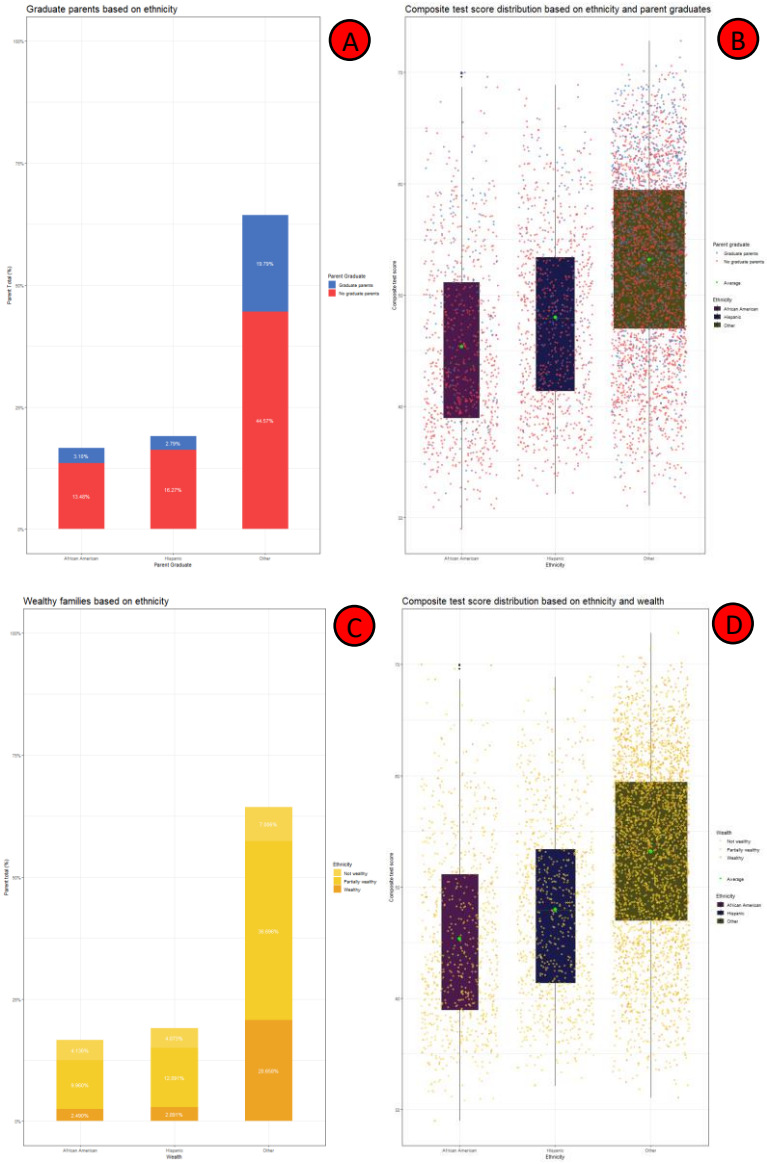


It is evidently clear that ethnic minorities do not achieve equal scores compared to other ethnicity groups. The distribution outlines how the 50% of both minorities perform which is roughly between 40-50 marks, with slight variance between their median and average but suggesting positive skewness and consistent scores overall. Other ethnicities have a 50% that achieves between 45 to nearly 60 marks with an average around 54, albeit with a median higher than the average suggesting most students achieve slightly lower. However, this has a small impact on the score discrepancy because their lower quartile accounts for the minorities' upper quartiles. Additionally, they have outliers achieving the highest marks in the dataset. This visual proposes that ethnic minorities struggle with education in contrast to represented ethnicities. Their sample size based on the widths emphasises how underrepresented they are in the dataset as well because the other ethnicity is roughly double their widths.

A box plot was reused for its ability to assess and compare distributions of bivariate data in-depth, similar to what I mentioned in my previous box plot visualisation, with the ethnicity along the X axis and scores along the Y. I now linked the sample size to the width of the boxes as to assess the extent of how underrepresented they were in the dataset. Unique dark hued colour encodings were used for each ethnicity to help differentiate them which will benefit me later. The average colour remained the same as it still stands out.

Q6. Could this imply that underrepresented ethnicities are frequently disadvantaged within education compared to other ethnicities?

I want to now explore further and give reason as to why some ethnic minority students may be underperforming to cause such differences. I will use my previous research on wealth and parent education to assess to what extent ethnic minorities are privileged, given that it was implied that these factors can affect scores. The assessment will include stacked bar charts and an altered variant of the previous boxplot.



Nearly a third of the other ethnicity sample has graduate parents contrasting both minority samples which have less than a fifth of them having graduate parents individually. These statistics highlights the imbalance, suggesting ethnic minority students are less likely to have graduate parents. Graph B shows students with graduate parents are very prevalent in and beyond the upper quartile across all groups. There is an abundance of graduate parents than non-graduate parents in the other ethnicity sample for its upper quartile and whisker compared to the other groups, reiterating the fact that that graduate parents correlate with higher scores and that there is a greater distribution of them for other ethnicities.

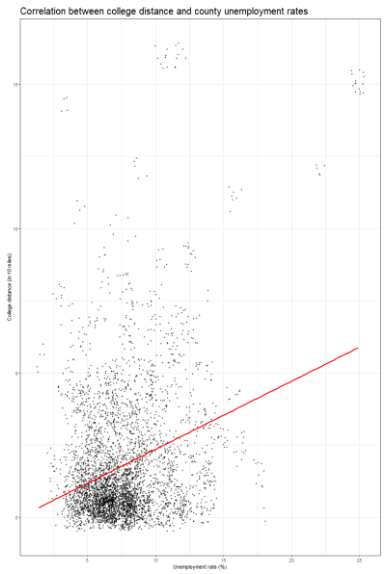
All ethnicities have partially wealthy as the most dominant status based on graph C. However, the proportion of each ethnic minorities' wealthy status is very similar where they are an extremely small percentage in contrast to the wealthy of other ethnicities. Ethnic minorities' not wealthy status is also more prevalent against it whereas the opposite applies for other ethnicities. Based on graph D, evidence supports the wealthy being more common in the upper end of the box plots, again correlating score with wealth and how much prominence the wealthy have against the not wealthy in the other ethnicity boxplot.

By taking previous research into account of wealth and parent education, I was able to deduce that ethnic minorities are less privileged though this visualisation, giving potential reason on why they may perform less adequately than other ethnicities.

I needed a bivariate graph that could still represent a total. A stacked bar plot was a candidate because I could set the ethnicities to the X axis, wealth and pgraduate to a fill, whilst getting their totals. Colour encodings were applied to the fill variables, distinguishing the graphs and the variable values within each ethnicity. Box plots have points enabled with appropriate variables applied as the fill value to each point with jittering and alpha manipulation to help identify them. Colour encodings were maintained for consistency, and by already having them in a darker hue, I was able to contrast the ethnicity boxes against the points to help them stand out.

Q7. Given the extent to which the economy can impact an individual's education, how much can it also affect the availability of education?

Researching on how else the economy could impact education, it can be assessed through its availability. The variable "distance", the distance between a person and a college in 10 miles, can be used to see how it varies based on the economy. The county unemployment rate variable "unemp" will be used to ascertain as to whether the economy is failing or succeeding. My hypothesis is that if unemployment rates are high, then college distance is likely greater because the rates would suggest that the economy of the county is low, therefore affecting the availability of education because they cannot build a college, thus affecting whether a student may enrol for college. I will visualise the two variables against each other within a scatter plot to view the correlation.



It can be inferred that the economy has an impact on the availability of education to a substantial degree. The visualisation depicts density around the lower rates of unemployment between 5-8%. As the unemployment rate increases, it becomes less dense and small groups form at the upper end of college distance, signifying the rate at which it is affected. Despite outliers around lower rates of unemployment, the data is not skewed enough as the line of best fit evidently shows less how both variables scale together with a positive correlation, verifying my hypothesis.

Distances greater than 16 were removed because some were abnormally high. Now using non-discrete variables, a scatter plot was suitable to study correlation for a range of values, also enabling bivariate data which was the distance on the Y axis and unemployment rate on the X. Jittering and alpha manipulation was used to prevent points getting lost amongst each other including a red line of best fit, enhancing visual clarity.