ECON 21300 Assignment 1, Yixin Hu

Question 1:

It seems that the data may need to be cleaned. Before doing this, though, estimate the treatment effect of attending LLS in two ways. First, use a simple difference in test dist means of those who won or lost the lottery. Second, include covariates as controls (think OLS) and re-estimate the treatment effect. Report your estimates and standard errors from each technique. Are the results plausible?

Step 1: Importing and getting to know the data

First, I will load the required libraries and check some basic charactoristics of the data. Specifically, checking the data type (chr,dbl,... etc), as well as noting the variables at hand and number of observations (rows). Until otherwise stated, I will assume that those who entered the lab school were indeed enrolled through a 'valid lottery', meaning that it is a random sample of the population in question.

Note that there are multiple items that jumps out, (take the -99 min income for example). While we should usually take a more detailed look at the data as well as clean the data *before* doing any regressions, we will follow the direction of the question and reserve these processes to the next question (Question 2).

Step 2: Estimating treatment of attending lab school

Step 2.1: Estimating treatment without covariates Now to address the question: estimating the treatment effect of attending lab school (variable: *lab*) on current scores (variable: *score*). According to the question, the following directly measures the treatment effect on the raw data without considering covariates. I will give a summary output directly following each model and at the end give a verbal discussion on both models.

```
##
## Call:
## lm(formula = score ~ lab, data = raw_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -37.67  -29.60  -27.94  -26.35 1274.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.539      1.729  20.550   <2e-16 ***
## lab            2.507      2.707   0.926    0.355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 152 on 9998 degrees of freedom
## Multiple R-squared:  8.572e-05,  Adjusted R-squared:  -1.429e-05
## F-statistic: 0.8571 on 1 and 9998 DF,  p-value: 0.3546
```

Step 2.2: Adding covariates as controls In adding more covariates, I believe that past performance on tests will give information on the current performance on tests. Therefore, I will be adding the past scores data (variable: *past_score*) into the model as a covariate to control over. The justification is that, even under the current assumption that the lottery is valid, one's previous ability to take an exam will contain information about one's ability to take an exam right now. Therefore, I believe that controlling for past scores is necessary to grasp at a more sensible interpretation of the regression by removing the information given by past scores that might by captured within the lab variable.

```
##
## Call:
## lm(formula = score ~ lab + past_score, data = raw_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -50.38  -32.26  -27.66  -23.19 1258.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.959      8.964  -1.111    0.267
## lab            4.363      2.728   1.599    0.110
## past_score     5.614      1.085   5.172 2.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151.8 on 9997 degrees of freedom
## Multiple R-squared:  0.002754,   Adjusted R-squared:  0.002555
## F-statistic: 13.81 on 2 and 9997 DF,  p-value: 1.029e-06
```

Looking at the summary of the models above along with the descriptive statistics of the *lab* and *score* variables, it is extremely difficult to come up with a good explanation as to what is going on. Not only is the coefficient on the *lab* variable statistically insignificant (p-value at 0.355 for the first model, and 0.110 for the second), but when taking the summary statistics of *lab* and *score*, it also lacks a clear economic interpretation and significance.

In particular, notice the scores range from 1.88 to 1312.939 yet the model reports the coefficient on the *lab* regressor to be 2.507 for the first model, and 4.363 for the second, as well as the extremely low R-squareds and very high residual standard errors. This is a very insignificant amount when looking at the large range of test scores and does not lead to any sound economic interpretations as to the efficacy of attending the lab school (not to mention that the *lab* variable itself has a max value of *3*, where if it were only a dummy indicating whether or not an individual is attending lab school, we would expect *{0,1}*). Due to all these complications, the results here lack any statistical and economic significance to draw any plausible conclusion.

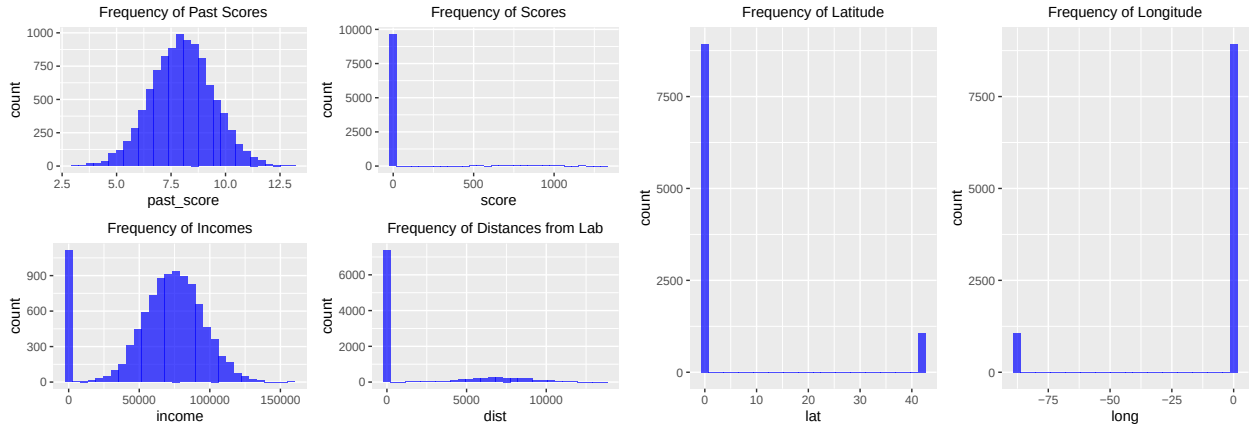
Question 2:

Clean the data. This may or may not require you to delete, transform, standardize, and scale observations or columns. Whatever you do, describe what you did with specificity and why you did it.

Step 1: A deeper dive into the raw data:

Step 1.1: Looking at continuous variables First, I will generate graphs for each of the variables to get a high-level understanding of the data we are dealing with. Here, I will take a look at scores, past scores, incomes, and distances from the lab school. For each of the four graphs, it will be a simple histogram to show the frequency of each observation. The x-axis will be the actual values found in the data, and the

y-axis will be the count of particular observations. Showing the graphs aims to visually identify what might be glaringly off about the data and thus give us a direction to proceed with cleaning.



I will summarize the information for each variable and potential next steps in data wrangling:

(First, rejoice that we didn't find any NA values.)

1. Past Scores:

- I believe this shows a very plausible and convincing distribution. This is because, assuming standardized testing (such as the SAT) will often times normalize/curve their scores according to a distribution, the visualization seems to indicate no outstanding features that we do not otherwise expect. Thus this gives us a very convincing collection of data because it fits a seemingly normal distribution. Furthermore, there are no extreme outliers, negative values, NAs, or anything out of the ordinary that we need to wrangle with.

2. (Current) Scores:

- Recall that I did not restrict the width of the histogram, and thus the histogram in question will include all instances of scores. This visual aid indicates that, while most entries are situated around 0, there are extreme outliers in this data (such as the max value of 1312.939) since the median is situated at 8.251. Given the cleaner nature of past scores, I will make the assumption that the scoring scheme for this standardized testing has not changed overtime (as there is still quite a lot of values situated near 0), and use **past scores as a reference to clean the current scores data.**

3. Income:

- The graph for income shows a relatively normal distribution similar in shape to that of the past scores (which seems more or less clean). However, there are more than 900 values situated **below zero**. This requires further investigation because we see from the summary statistics that the minimum value is *-99*, which like the infamus *-999* will require careful cleaning.

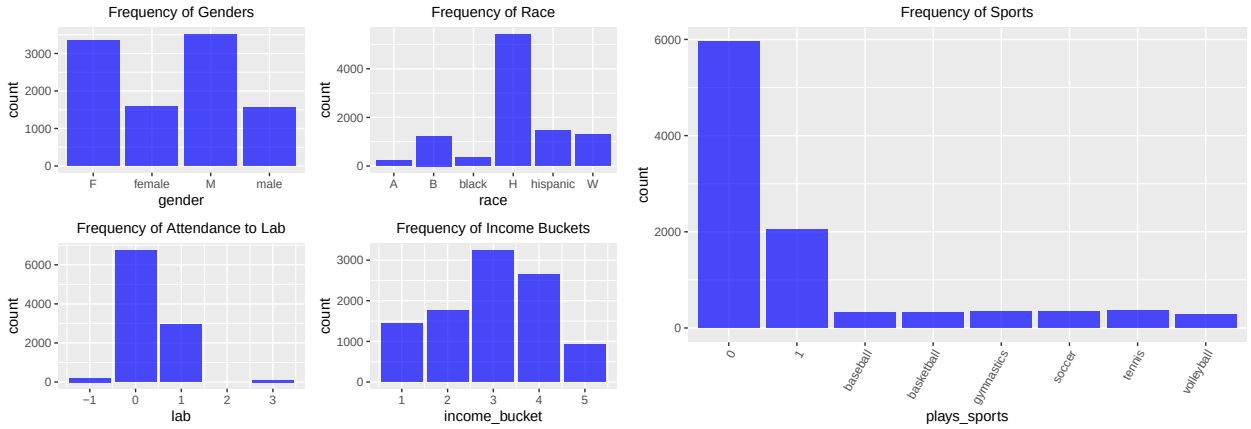
4. Frequency in Distance from Lab:

- Similar to the current scores, the visual representation of this graph shows a strong concentration of observations near zero, and outlying values which stretch to 13609.374. However, it could very well be a common error regarding units (i.e. some inputs kilometres while others inputs metres). To justify my thinking (assuming that the difference in units is just a two-way difference such as km and m), **I will see if scaling down the outlying values (starting from around 2500 units) will give us a more sensible result**.

5. Longitude and Latitude:

- These two variables are talked about together, as I do not believe that one can truly make sense of these variables. The reason for my claim is that for both variables, the variable's 3rd quartile is 0, which I will interpret as no information gathered (rather than 7000+ people living somewhere in the Atlantic ocean). I will concede that for the non zero values, there could be some argument that can be made there. However, even in that case, I still believe that it will not be useful for our analysis on the treatment effect of lab school education. This is because the only reasonable and pertinent argument that can be made is the location between on the individual's home and the lab school may have some importance. However, this boils down to distance to the lab school, which is already covered by the *dist* variable. Therefore, I will disregard these two variables for now.

Step 1.2: Looking at categorical variables Now I will repeat the visualization and summarizing process as shown above, but this time for the categorical variables. Here, I will be using bar charts to visualize the variables gender, race, sports and lab school attendance. This is meant to see the unique values of each categorical variable and their frequency. In doing so, we may catch repeating values (such as 'male' and 'M') among other inconsistencies.



Like before, I will summarize the findings:

Again, rejoice in that we do not have NA values.

1. Gender, Race, and Sports:

- I am grouping these three variables to talk about together because they share very similar issues. In particular, while they do not exhibit any NA values, these three variables have **different categories which are, in actuality, the same category**. For example, in the gender data, there is "M" and "male", where if I make a safe assumption that M really *means* male, then these two categories point to the exact same category, yet they are seperatly counted. Same situation occurs with "H" and "hispanic" for the race categorical variable. For the sports variable, we see first a set of $0, 1$, where I interpret as {No Sports, Plays Sports}. However, we see then specific sports such as basketball, baseball etc. Therefore, I believe that these specific sports separated from the variable 1 can actually be combined into the value '1'. Therefore, the next step is to combine redundant categories in these variables

2. Attendance to Lab:

- Assuming that this variable indicates whether or not the individual goes to the lab school, it is completely reasonable to expect a $0, 1$ dummy variable. However, we see values such as "3" and "-1". At

first glance, I could possibly make the (very shaky) argument that "-1" is an input error (typo) of "1", but "3" really cannot be explained with information at hand. Because attendance to the lab school is critical for our analysis, I believe that the best course of action is to drop these columns for the fear of making an incorrect assumption and skewing the results.

3. Income Bucket:

- While there are no explicit/glaring issues with this data, it is important to recall from the previous analysis on continuous variables that income itself is quite problematic (with -99s). If we make the assumption that these income buckets are directly constructed via the income column, then we are bound to have errors in analysis if we use this variable. In all, I believe that it would be ideal to clean the income variable first, then perhaps reconstruct our own income buckets by using the same bins if necessary.

Step 1.3: ID I have reserved the ID variable to a section of its own because it deserves special attention. This is because identification should be representative of an unique observation (which means that each row, which represents a unit of observation, should have their own ID). Therefore, this will be the final "check" to ground all the aforementioned data cleaning. The only item to check is to see whether or not the IDs are unique. In the following function, I will return all the ID values that are unique.

```
## [1] 5000
```

Beyond this number, it seems that there indeed many observations to clean out. Therefore, the next course of action is threefold: 1. to delete completely identical rows 2. to delete rows with duplicated ids yet extremely faulty information in one specific category (i.e. for two ID entries, all else the same, one has a -99 in the income category; then delete the one with the -99 in the income category) 3. to carefully consider the rows with duplicated ids, but both has very 'believable' data.

Step 2: Cleaning the data accoring to our observational conclusions

In the following sections, I will clean each of the indicated variables which require cleaning, provide visualizations and summaries along the way, and track my modifications

Step 2.1: Cleaning continuous data: First, I will deal the -99s identified within income. I will follow the direction indicated in our class: I will first set the all the -99 to zero. However, recall that we have double counted IDs, and thus there is a possibility where we can reasonably delete the rows with faulty data (i.e. if all else equal, two identical rows only differ in income, where one has -99 and the other has a more believeable value, I believe it would be sensible to delete the row with the faulty income data). With that in mind, I will change -99s to 0s in case of a regression, and identify rows where we might delete the rows outright by placing a dummy variable where 1 represents that the income observation was a -99 and 0 otherwise. The following print out gives the amount of rows with income below zero, which happens to all be -99s.

```
## [1] 1115
```

Now we will look at scores. As previously noted, past scores seems to have a fairly reasonable distribution, meaning no NAs, extreme outliers, and follows a normal distribution as we expect from a standardized test. Judging by what I found previously with the past scores and current scores, I will use the past scores as a reference point to determine what the 'reasonable' score range should be for the current scores. The first set of summary statistics is about past scores, the second one concerns current scores.

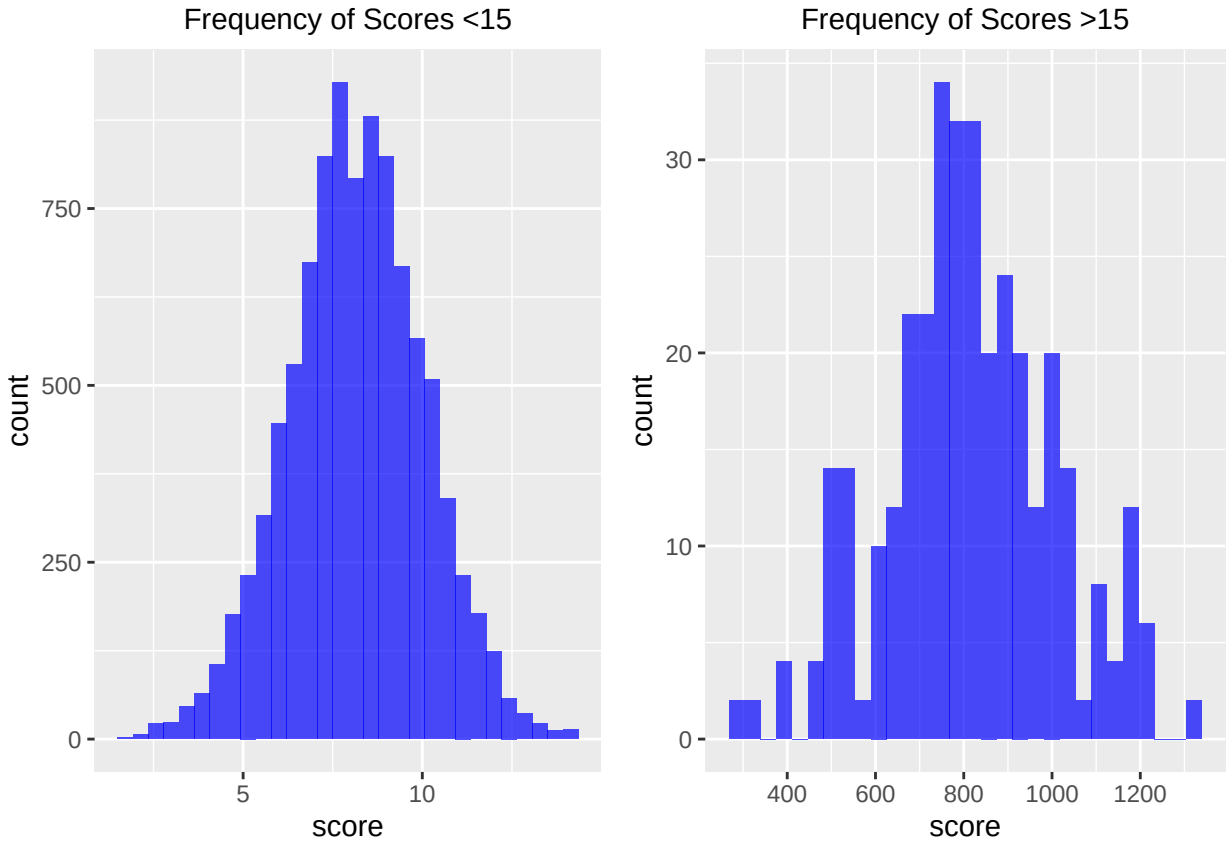
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.239   7.046   8.008   8.004   8.929  13.190
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    1.877    6.989    8.251   36.302    9.594 1312.939
```

Judging by the the summary stats and visualizations made previously, I believe that the maximum score on the standardized test, assuming that how the test calculates scores is consistent between the two times, the tests scores should be upwardly bounded around 15. First, I will do a very "hardcoded" operation to tease out the local upper boundaries of the group of scores near 0 on the visualization. In particular, I simply placed bins and found a steady value where the number of observations do not change for a long period of time. (See code for reference)

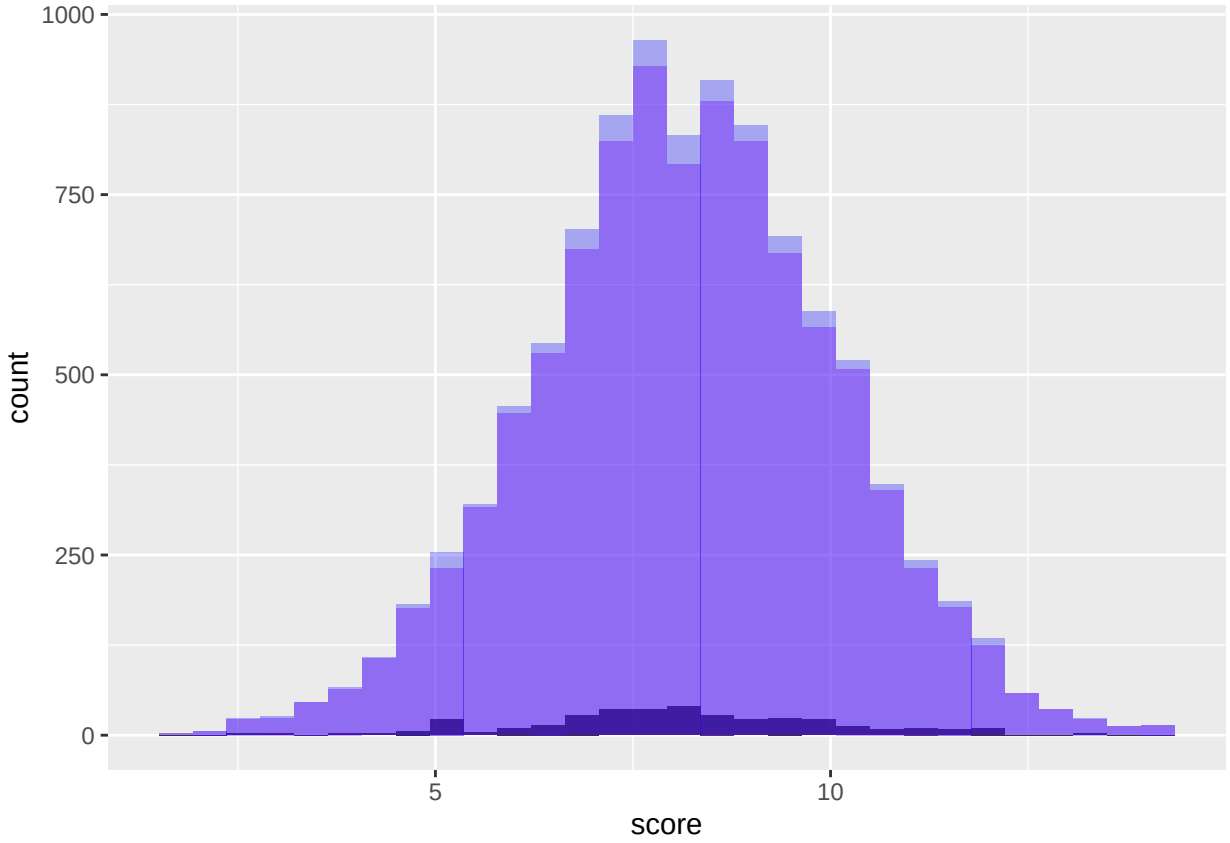
Even in this very "hardcoded" or "theoryless" operation, we are able to tease out how many scores are above a certain cutoff. In doing so, we can more clearly see what the cutoff could be. Now, making the assumption for now that 15 is our cut off, I would like to see what is the **maximum** value that is **below** 15 and the **minimum** value that is **above** 15. In doing so, I wish to show that, under the assumption that scores should be roughly normal, we have no reason to expect a very large gap in scores.

The result is that max(below15) = 14.3, and min(above15) = 266.25. Notice that there is a big jump from the cutoff below15 and above15! From here, I will split the score data from 15 and check the distributions for those scores below and those scores above:



Here, we get a very interesting result! Once we subdivide the score variable to those above 15 and below 15, we find that each looks somewhat normally distributed, which is exactly what we expect from a standardized exam score. Now looking at the score ranges of the two series, I suspect that the scores above 15 are created due to a missing decimal point which caused all data to be scaled up by 100 times.

To show this, I will divide the scores from above 15 by 100 and plot it together with the scores below 15 on the same graph. The following histogram shows how a rescaled version of scores above 15 fits into the overall distribution. If we do not find any outliers here, I reasonably conclude that the rescaling was successful, and that my hypothesis about the resealing factor of 100 is a reasonable argument.

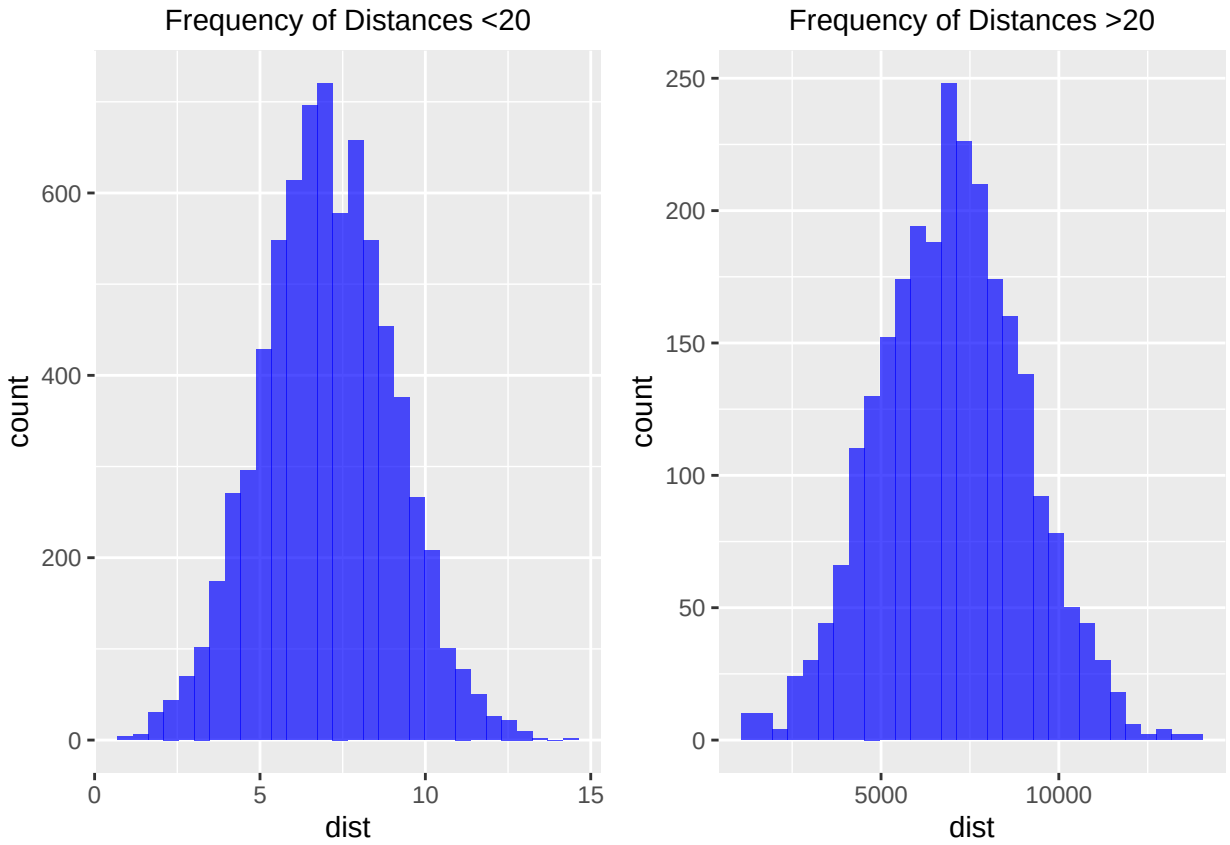


Here, the dark blue component of the histogram at the very bottom of the graph shows the rescaled counts above 15. The purple component of the histogram shows the original scores of scores below 15. The light blue component which sticks out at the top shows the result of the compounding of the two aformentioned subcomponents. Nevertheless, this shows that there are no significant outliers, and that the rescaling fits within the normal distirbution which we expect from a standardized test. Therefore, my conclusion is that I will move forward and clean the score data by dividing all scores > 15 by the factor of 100.

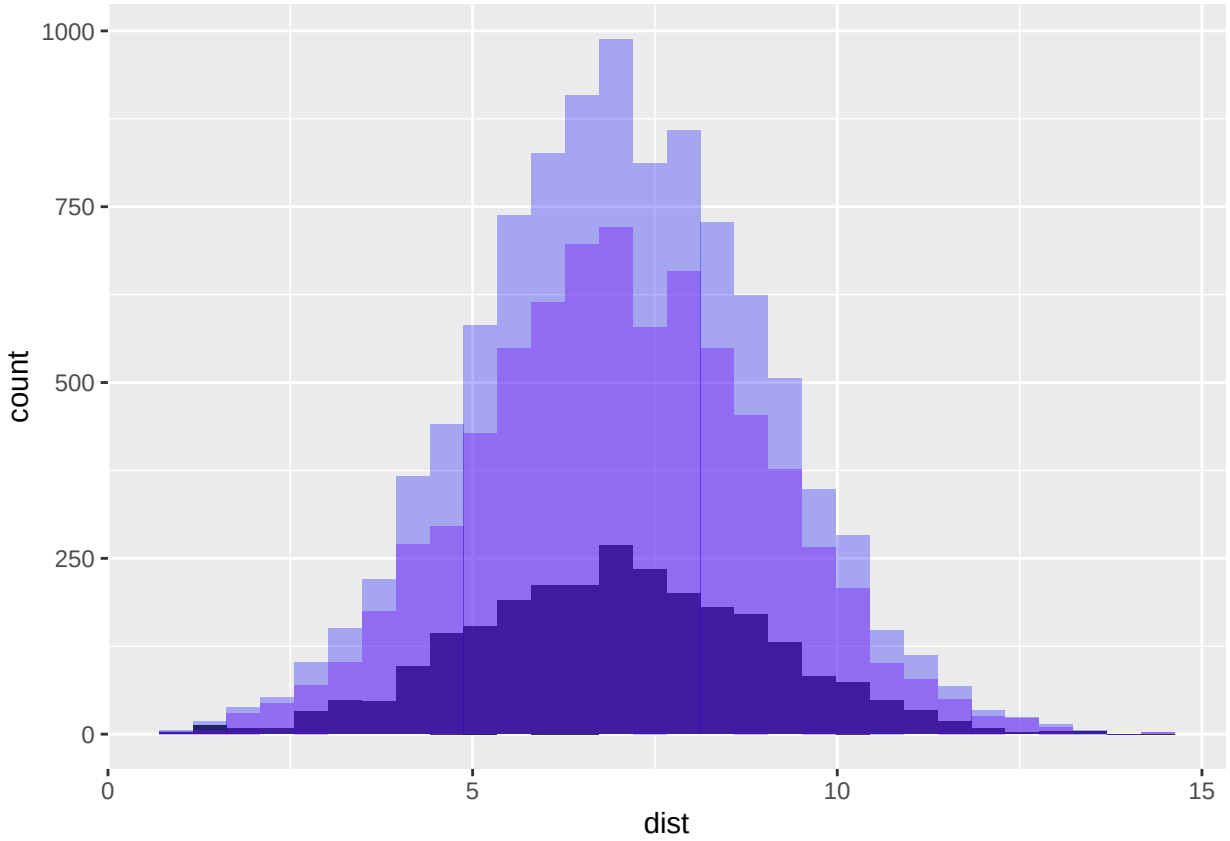
Now I will deal with distance the same way I have dealt with scores, which is checking for a scaling factor which we can use to bring the outliers back into the distribution we are expecting. Here, we are not explicitly expecting a normal distribution for distance (and maybe at most even say that we expect a distribution skewed to zero, since we expect people who goes to the school lives close to the school, but even that expectation is not completely reasonable, since they could very well just commute a longer distance). This time, we dont have a reference, but judging by the visualization, there could be a cut at around 2000 units. (see code for reference)

Again, I would like to concede the brute-force nature of this analysis, but this nevertheless indicates at least a few cutoffs we can use, since we have found multiple 'stable' counts at different values, the smallest which is 20. For the sake of argument, let us take 20 as the cutoff and find the max below 20 and the min above 20, and see if this is a valid cutoff.

After a quick operation, we find that max(below20) = 14.21, where as min(above20) is 1084.052. Hence we have identified the structural break here. As a supporting visualization, I will also show the graphs of the two:



Notice that neither distributions have glaring issues such as extreme outliers. As there is no theoretical reason why I believe the distances should be normally distributed, it is more important that the two subsets of data are visually extremely similar. As previously, I suspect that there is a scaling issue and thus divide the distances > 20 by the factor. Here, judging by the indications on the x-axis, I will choose 1000 as the factor, as there is not only empirical evidance to do so (x-axis), there is also a theoretical reason where I suspect the different sets of data here is operating in kilometres and metres (1km = 1000m, just in case this information is needed for the reader). The full justification will be given after the graph.



Here, the dark blue component is the rescaled data from above 20 units. The middle shaded purple is the data from below 20 units. The light blue component is the aggregation of the two aforementioned components. Because the visual graph does not show extreme outliers or anything out of the ordinary, it seems that the rescaling was able to capture the outliers back into the distribution that we expected. The theoretical explaination for this has been mentioned in passing, but to restate, I beleive that there was an entry error where individuals confused the units *km* and *m*, and thus explains the discrepency in the factor of 1000. (i.e. it could be that one researcher knows that the unit is *km*, while the other one thought they were using *m* instead).

Now moving to logitudal and latitudal data. Since I cannot reasonably find any use for the longitude and latitude data (*see full justification in Q2, section 1.1, summary point 5.*) I will proceed to delete these dataframes from our discussion to streamline the dataframe.

Step 2.2 Cleaning categorecal data: Recall the unique values in the different categorical variables race, gender, lab, and sports:

Notice that race has entries "hispanic" and "black" along with "H" and "B". As seen in the output above, full names are not consistent with the naming scheme of the other categories (A and W). I will make the assumption that "hispanic" points to the same category as "H", and thus categorize "hispanic" into "H", and same logic with "black" and "B".

Moving to gender, I believe that a similar story is occuring based on our previous analysis. Specifically, "male" and "M" are pointing to the same category, and "female" and "F" are pointing to the same category. Therefore, in the same line of thought as before, I will recode the data accordingly.

In a similar line of thought, I believe that for the sports data, "0" means does not play sports, while "1" indicates the individual plays sports. Therefore, I believe that all the names of sports (basketball etc.) can be categorized as "1", since they do indeed play sports, its just that they have indicated a particular sport,

and thus it is not counted as the category "1" – play sports. Therefore, I will recode the data accordingly. Notice however, that the data type of the sports column is "chr", which means that I will have to recode into the charactor "1" first, and the change the column to a more convenient data type such as integer.

For the lab variable, recall from the previous section that, assuming that this variable indicates whether or not the individual goes to the lab school, it is completely reasonable to expect a $0, 1$ dummy variable. However, we see multiple -1 values and 3 values. Therefore, I will find how many -1 and 3s there are and determine their importance on our data.

```
##
##   -1    0    1    3
##  206 6752 2938  104
```

```
## the number of non {0,1} values are:  310
```

```
## the number of {0,1} values are:  9690
```

```
## 0.1018397 is the percent of all the lab data which contain non {0,1} values
```

Because I cannot make sense of these -1 and 3s and would not like to introduce bias into our analysis, I will remove these rows. I recongnize that I am trading off robustness in the data for a more manageable interpretation, but I believe that removing these values (which might get dropped anyways due to the duplicated IDs) will give us a significantly more interpretable result when doing regressions.

Moving onto income bucket, recall from before that we have -99 values in our income data. If this income bucket was created using the income data, then there would be no real point in cleaning this data as it is inherently faulty. If we were to ever use this data, I will create a new income bucket using the original bins. Therefore, I will remove this incorrect data.

Step 2.3: Dealing with IDs Now I will deal with the ID column, which will be important because we do not want to double count, assuming each ID should be unique (since we are not dealing with panel data accross time). As mentioned in our data exploration phase, we note that there are duplicates in the data. Based on this understanding of the data, I will do the following: 1. delete completely identical rows, since they represent double counted data. 2. delete rows with duplicated ids yet extremely faulty information in one specific category (i.e. for two ID entries, all else the same, one has a -99 in the income category; then delete the one with the -99 in the income category). This is because the non-faulty row contains all the relevant information in these very narrow cases, where as the row which **only** differs with a -99 value will not be useful for our analysis. 3. to carefully the rows with duplicated ids, but both has very 'believable' data. This is the most tricky part, and I will deal with these on a case-by-case basis should they arise.

First, note what we are aiming for, ie, the true count of unique ids and proceed with the cleaning.

```
## [1] 5000
```

While cleaning, we have identified 3885 completely identical rows (ID or otherwise), and removed them accordingly.

Then, I found all the "distinct id with negative_99" pairs. This is to see if we can eliminate the rest of the repeated rows by considering instances where we had garbage data (ie. -99s). As a result, by filtering out garbage data, we were able to directly arrive at our desired state – where we have all unique IDs! Therefore, I will not need to do the third step in our operation (looking at individual cases), and I am also able to remove the was_neg99 column, since it no longer has any meaning in our analysis (since all rows originally with -99s are gone, and if we keep them, we would just have a column with only one value). Now that we have a dataframe with no duplicate observations, I will now remove columns which do not make sense.

Specifically, I will remove all rows which have either -1 or 3 for the lab attendance variable. I realize that there could be the interpretation where the -1s are typos of 1s etc, but I believe that sacrificing these rows so no bias is introduced will be more constructive for the task at hand.

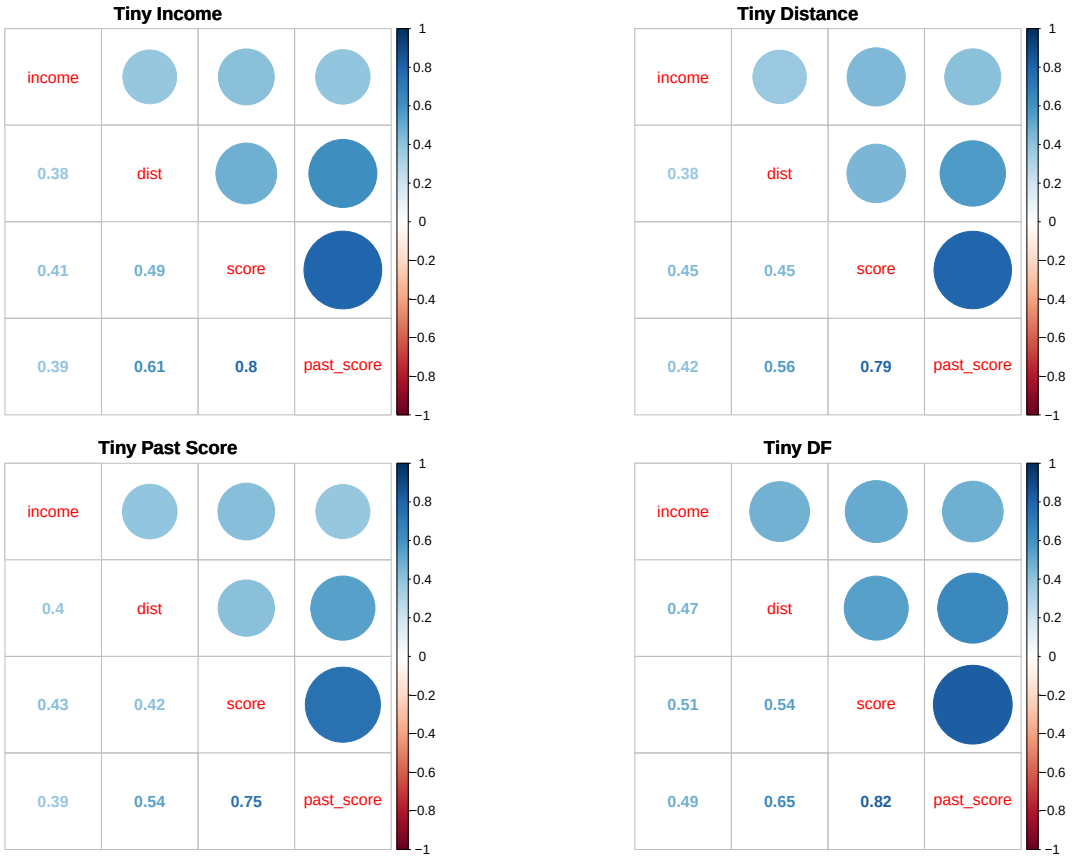
From here, I will move forward with Q3.

Question 3:

With your cleaned data, re-estimate and report the treatment effects for the two methods in question 1, assuming the randomization is valid. What are the differences in results between the two methods? Do the estimates imply a big or small impact of the school? Positive or negative? Is one estimate more convincing than the other?

Step 1: Allocating for robustness check and exploring correlations Before doing regressions, lets consider the outliers which might affect the results of estimating the effect on the treatment effect on grades by lab school education. Recall that all the continuous variables which might seek to explain current scores (so: income, distance, past scores) are normally distirbuted. Here, I will cut off the tails of each distribution and save them in a separate dataframe to and run a regression on each to determine the initial results. Then, I will use each resultant regression on the full data set to check for the robustness of the result and interpret from there.

Now I will use a correlation matrix to check the correlations between each of the continuous variables from each of the tiny dataframes created above. In doing so, I can get a rough idea as to each variables correlate with each other in the cleaned dataframe, and discuss potential regressions from there.



Notice that each of the tiny dataframes gives very slightly different results. It is quite interesting that removing outliers for any of the aforementioned variables in the dataframe will generally decrease the

correlations between the variables. However, this is a trade-off that I believe will be helpful in producing a more concrete analysis because we would not want outliers to skew our results. Furthermore, all variables are positively correlated with each other with varying degrees of correlation, and that past scores are the most highly correlated item with present scores, regardless of the different With everything ready, I will move forward with regressions.

Step 2: Producing Regressions

Step 2.1: Replicating the no-covariate regression with different tinys.

```
##
## Call:
## lm(formula = score ~ lab, data = tiny_inc_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8559 -1.1446  0.0152  1.1874  5.9160
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.09770    0.03218 251.610   <2e-16 ***
## lab          0.14440    0.05914   2.442   0.0147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.783 on 4357 degrees of freedom
## Multiple R-squared:  0.001367,   Adjusted R-squared:  0.001137
## F-statistic: 5.962 on 1 and 4357 DF,  p-value: 0.01465
```

```
##
## Call:
## lm(formula = score ~ lab, data = tiny_dist_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8424 -1.1593  0.0147  1.1850  6.0103
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.08421    0.03172 254.901  < 2e-16 ***
## lab          0.20757    0.05915   3.509 0.000454 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.768 on 4357 degrees of freedom
## Multiple R-squared:  0.002818,   Adjusted R-squared:  0.002589
## F-statistic: 12.31 on 1 and 4357 DF,  p-value: 0.0004544
```

```
##
## Call:
## lm(formula = score ~ lab, data = tiny_pscore_df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -5.3817 -1.0973  0.0126  1.1387  5.9859
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.08629    0.02932 275.830  < 2e-16 ***
## lab          0.22982    0.05366   4.283 1.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.621 on 4357 degrees of freedom
## Multiple R-squared:  0.004192,   Adjusted R-squared:  0.003964
## F-statistic: 18.34 on 1 and 4357 DF,  p-value: 1.885e-05


##
## Call:
## lm(formula = score ~ lab, data = working_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2484 -1.2247  0.0026  1.2588  6.0746
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.12549    0.03250 249.979   <2e-16 ***
## lab          0.10195    0.05903   1.727   0.0842 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.889 on 4843 degrees of freedom
## Multiple R-squared:  0.0006154,  Adjusted R-squared:  0.0004091
## F-statistic: 2.982 on 1 and 4843 DF,  p-value: 0.08424
```

Here, we have produced 4 regressions, the first three consists of removing outliers of particular variables, where the last is the full clean dataframe. For the sake of completeness, I have showed all results in the write-up. In all of the new regressions, we have much more interpretable and believeable results relative to those in question 1. It is very interesting that removing outliers in our data (regardless of category), strengthens both the economic significance (larger magnitude of coefficient) as well as statistical significance (p-values) of our regression. However, this regression is effectively saying *"not a single other characteristic of an individual under the current assumptions could explain a variation in the future scores."* However, I do not think it is the case, even with the current valid lottery assumptions. I will move forward with the model with covariates and compare the simple model above with the final model with cerntain covariates.

Step 2.2: Replicating the model with covariates Now adding more covariates into the model, we must consider what should be the proper variables to consider. I believe the first variable to consider is past scores. This is because of the following reasons: 1. Theoretically, I believe that one's ability to take previous exams should serve as an indication for their ability to take future exams. This is because their overall 'competency' for taking the same exam over time should be more or less consistent, i.e. someone not that great at taking the standardized test wouldn't be an awesome exam taker out of the blue. On the other hand, a great exam taker has no reason to suddenly become terrible with standardized tests. Therefore, not accounting for these situations in our model would skew our conclusion on the effectiveness of attending the lab school. 2. Emperically, the correlation between the two scores are quite high, and this high correlation

which indicates how the two columns are interrelated should be taken into account when constructing a model. Else the error term on the regression would be quite high.

It should also be noted that I am **not** going to consider (for now) other variables such as distance and income, even though they do show some interesting correlations. This is because I am still following the assumption that the lottery is valid, which results in the following interpretation. Under the valid lottery assumption, there is no reason to believe that income and distance has any affect on current scores for two reasons: 1. There is no reason to believe that somehow living closer to a school, or one's parents having more income, directly explains one's ability to take exams. 2. More critically, one could argue that individuals with higher income or live closer to the school are more likely to get in, and thus we need to control for these variables. However, remember that the current assumption is a valid lottery, which means that regardless of any factor (income, distance etc.) we are assuming that anyone is equally likely to get picked into the lab school. Therefore, while these are valid arguments, it is under under the current assumptions that these controls are currently unnecessary.

With our intuition written, the regression is ran with each of the data frames: The first regression is ran upon the no-outlier distance data, the second on the no-outlier income data, the third on the no-outlier past scores data, and the last on the full data frame.

```
##
## Call:
## lm(formula = score ~ lab + past_score, data = tiny_dist_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0875 -0.7103  0.0035  0.7186  3.3702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89071    0.10103  -8.817   <2e-16 ***
## lab          0.57890    0.03513  16.481   <2e-16 ***
## past_score   1.10866    0.01226  90.399   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 4356 degrees of freedom
## Multiple R-squared:  0.6533, Adjusted R-squared:  0.6531
## F-statistic:  4104 on 2 and 4356 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = score ~ lab + past_score, data = tiny_inc_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0474 -0.7092  0.0027  0.7040  3.4011
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.79127    0.09808  -8.068 9.17e-16 ***
## lab          0.59378    0.03474  17.094  < 2e-16 ***
## past_score   1.09402    0.01185  92.329  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.037 on 4356 degrees of freedom
## Multiple R-squared:  0.6623, Adjusted R-squared:  0.6621
## F-statistic:  4271 on 2 and 4356 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = score ~ lab + past_score, data = tiny_pscore_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1048 -0.7222  0.0055  0.7169  3.3597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.97908    0.11848  -8.264   <2e-16 ***
## lab          0.57411    0.03507  16.369   <2e-16 ***
## past_score   1.11975    0.01444  77.520   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.051 on 4356 degrees of freedom
## Multiple R-squared:  0.5815, Adjusted R-squared:  0.5813
## F-statistic:  3027 on 2 and 4356 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = score ~ lab + past_score, data = working_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1024 -0.7207  0.0090  0.7133  3.3669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.05423    0.08924  -11.81   <2e-16 ***
## lab          0.61962    0.03298   18.79   <2e-16 ***
## past_score   1.12739    0.01074  105.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.043 on 4842 degrees of freedom
## Multiple R-squared:  0.6951, Adjusted R-squared:  0.695
## F-statistic:  5519 on 2 and 4842 DF,  p-value: < 2.2e-16
```

Here we have a slightly different result from before. Between the regressions which control for past scores, we see varying magnitudes of the coefficients, standard errors, and statistical significance, but these values all have the same sign, and their magnitudes have small enough differences to where I would believe that they tell the same story: the lab school positively affects exam outcomes by around 0.57 to 0.61 points. The economic significance will vary by people, (if someone believes that an increase in 0.57 to 0.61 is necessary for, say college admissions, then it might be economically significant to them), but in general, these values are statistically significant and one could argue for economic significance.

After controlling for past scores, we see the coefficient on lab has increased by about 0.3 points while retaining statistical significance, decreasing standard error, and *vastly* increasing the R-squared value (even

if it is not exactly a fair comparison between different regressions, it is still worth noting that the R-squared has increased by a lot).

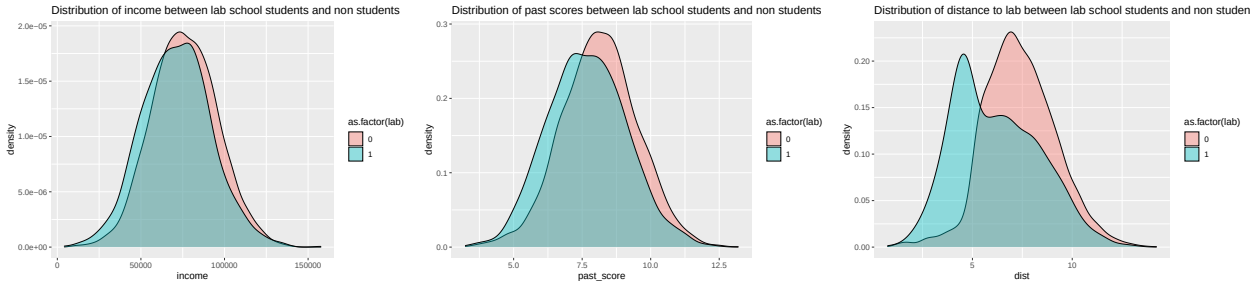
One explaination/interpretation of the regressions above is the following: After controlling for past scores, we are effectively controlling for those individuals who might not excel at exam taking *even if* they attend lab school. In doing so, we are able to make a stronger argument about the actual effectiveness of attending the lab school, and that the lab school is in effect "not to blame" for not being able to increase scores if certain individuals do not excel at exam taking. In fact, one could interpret the regression as that one's previous scores matters much more when looking at future scores, and the lab school might not be that economically significant after all (but people will have different arguments). For the aforementioned reasons, I believe that the second regression type (controlling for past scores), is more economically interpretable under the valid lottery assumption, which shows a small, positive effect of attending lab school on future scores.

Question 4:

Professor Levitt sometimes makes mistakes. What evidence would you present to convince him that acceptance to LLS is not determined by a valid lottery?

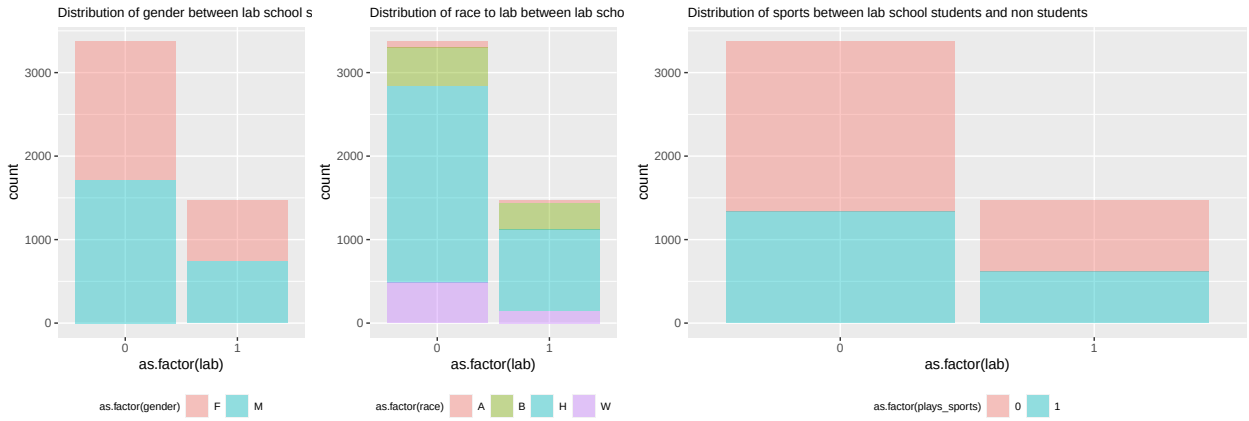
Notice that every analysis above is made under the assumption that the lottery is valid, and I have gone to great lengths at reinforcing that the previous kind of thinking only works under the valid lottery assumption. Now I seek to use simple visualizations in the cleaned data to show that the lottery is indeed invalid.

First off, I will operate on the definition that a "valid lottery" means "a random sampling of the population", where every single individual in the population has equal chance of being selected by said lottery. If that is indeed the case, I would expect that individuals' characteristics who go to lab school to be extremely similar to those who do not go to lab school, and that their only difference is lab school attendence. However, notice the following graphs: These are comparison between past scores, income, distance from the school of those who attend the lab school and those who don't. For each graphic, I used the full clean data as well as the corresponding tiny data to that variable.



Here, we see that the visual analysis tells us that there is indeed characteristic difference between lab students and non-lab students. In particular, the most significant aspect of this uneven characteristic distribution is the distance. We see a sharp spike of individuals who live very close to lab relative to those who do not go to lab. If this is indeed a valid lottery, this discrepancy should never occur. Therefore, even just looking at distance shows a characteristic difference between those who go to lab and those who don't. Of course, it is also worth noting that the distributions of past scores and income are slightly different, but distance is nevertheless the most significant issue at hand. Conducting T-tests, we see that the difference between those who go to lab and those who don't is significant for all three continuous variables. (Check code for the actual tests.)

More over, we can look at the distributions of categorical data. I will seek to show that there is indeed difference between the characteristics of those who go to lab school and those who don't. I will use bar graphs to indicate this difference. Specifically, we would expect that, under a valid lottery, that there should be no disproportionate data (i.e. a certain ethnicity or gender is disproportionately more likely to attend lab school relative to their proportion in the population). This will be indicated by a split evenly between the two bars. Otherwise, it will be a description of disproportional data between the two groups.



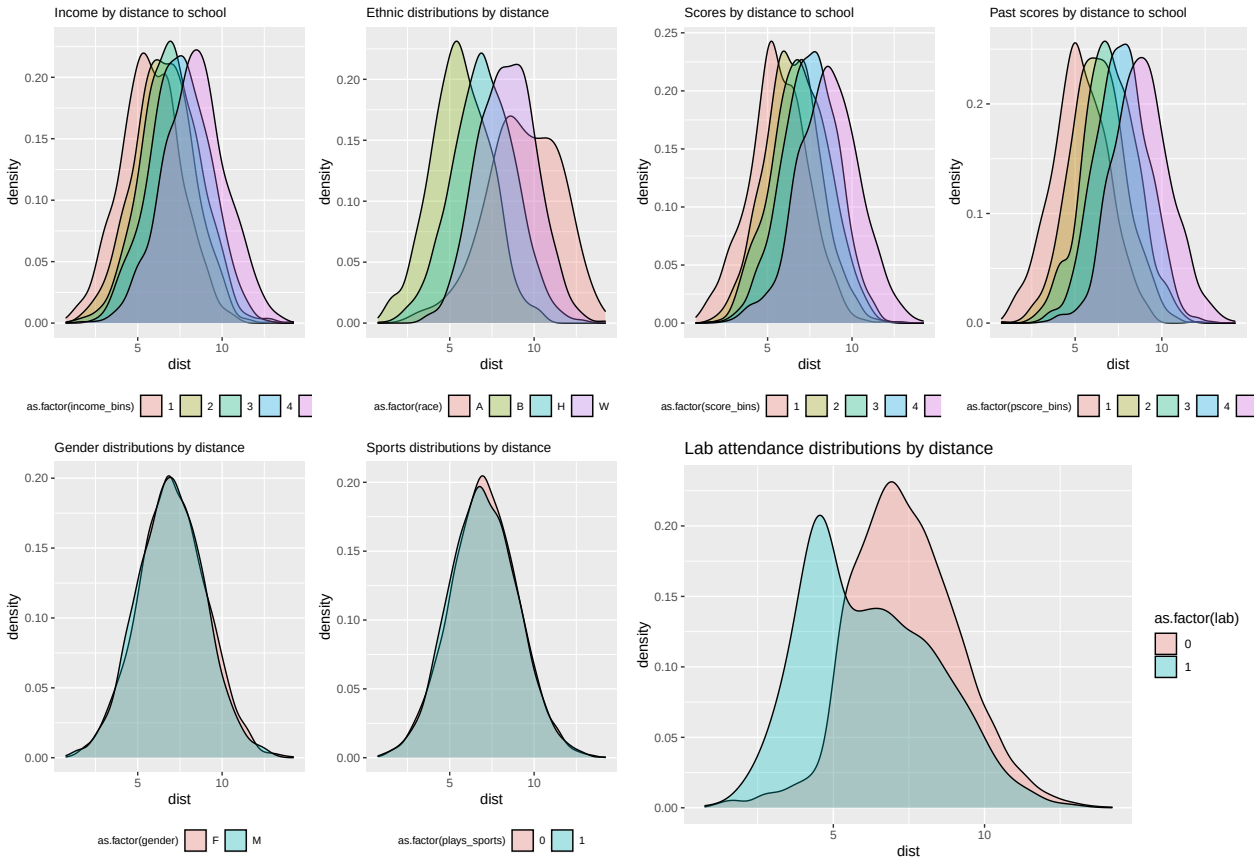
Among the categorical variables, it seems that gender does not contain significant disproportionate data, and sports show small discrepancies. However, race data does show some discrepancy between those who do enter lab school and those who don't. In particular, individuals labeled "H" and "W" are proportionally less likely to be attending lab, where as individuals labeled "B" are more likely to be attending lab. Therefore, it is shown visually that there is discrepancy between the two groups, and thus on the level of the categorical data, it seems that race data goes against our intuition of a "valid lottery". As a result there must be some implicit bias which is causing this discrepancy, and I will get to my hypothesis in the next section.

Combining the analyses above, we see that the so called "valid lottery" seems to have a bias towards individuals who live close to the school. If we set the definition of "valid lottery" as a "random sampling" of the population, then this evidence goes against the definitions, which indicates that **the "valid lottery" is not exactly valid after all.**

Question 5:

Now explain how you think students actually get into the school, based on your analysis of the data. Assuming your hypothesis is correct, is it possible to estimate a causal impact of attending the school on test scores? Do your best providing an estimate and standard error.

Following directly what we have found in Question 4, I will move forward to argue that distance might serve to point at a causal relation between the effectiveness of lab education on exam scores. First off, I would like to follow up from Q4: judging by the distribution on distance between those who attend and those who do not attend the lab school, I believe what is really happening is that the so called 'valid lottery' is conducted with a kind of bias involved: Recognizing we have data of all those who actually registered for the lottery (as mentioned in the Question prompt in Q1), I believe what is really happening is that there might be an implicit bias towards those who live closer, i.e. those who live closer might be treated more favorably by the so called 'lottery' system. To that end, it would be helpful check the characteristic of individuals who live close to the school vs those who live far. Here, I will present some figures which outline the different characteristics of individuals as we move further away from the lab school. This is to show that distance does indeed outline different characteristics, and thus identity what variables we want to control for in our regression. First, I will bin past and current scores as well as re-bin the cleaned income variable into 5 bins (20th,40th,60th,80th percentiles as cutoffs respectively). Then, I will directly show the graphs which outline the varying characteristics of individuals as we move further away from the lab school. This will aim to highlight certain characteristics as being associated with distance from the lab school.



From the many graphs above, we seek to show the story that individual characteristics varies with distance from the lab school, and that they are not homogenous with respect to distance. We see that income, race, and past scores all vary as we move further away from the lab school. However, gender and sports participation does not vary with closeness to the school. All being said, we may want to use these variables which have significant difference as a method to control for how the school actually impacts scores. But first, I would like to make a new variable called "closeness", defined by $\frac{max-dist_i}{max-min}$, where i represents a single observation. I recognize this is a normalization process, and requires new interpretations of the coefficients. However, since those who live close to the school generally go to the school, I believe that this is a more intuitive measure of how distance may play a role in determining whether someone attends lab or not. This is because we can directly see what is the effect of *"being 1% closer to the school relative to everyone else"*.

Now we will use this to describe how closeness to the school may serve as a way to understand the causal effectiveness of attending lab school. In particular, there is no reason to believe that living close to the lab school has a direct effect on one's ability to improve their exam scores. Furthermore, unlike variables such as income, which may be used to improve one's score by purchasing tutoring services, distance to a school does not have such a characteristic where it can vary one's testing improvement level other than attending the school itself.

Of course, one could validly argue that we could have confounding factors if there is a school that is far from the lab school that excels at improving standardized test scores, and thus being further from the school might also have a positive effect on grades. However that would be extremely difficult to argue for. Intuitively, this is because distance is a radius measure: unless somehow the lab school is surrounded (concentrically) by other prep-schools, it is not necessarily the case that distance inherently contains structural flaws for it to be a instrument of measurement. One could also argue that variables such as income might be the determining factor rather than distance. However, as mentioned previously, income might affect scores through other means, where as distance does not inherently have the direct effect on scores.

Furthermore, we have also identified that income, past_scores, and race could have interactions with distance

which skew our final regression result. Therefore, I will add these as a control to control for the variation due to this characteristic. With all that said, I will now measure how closeness to the school can affect one's scores through attending lab school.

In a word: individuals closer to the school tend to enroll in the school more often due to the bias in the lottery, but this distance to a school might not otherwise directly affect one's ability to improve test scores. Therefore distance, once controlled for, should be a good instrument for measuring the true effectiveness of attending the lab school.

Now to actually run the regression with our proposed instrument on each of the tiny data frames (outliers removed). The first regression is ran upon the no-outlier distance data, the second on the no-outlier income data, the last on the no-outlier past scores data. (see code for all regressions)

Summarizing the regressions above, the true significance of attending the lab school is reported in each of the regressions without outliers. Amongst all three, we see that the value ranges between 0.327 to 0.386, and each report a standard error ranging between 0.14 to 0.16. Now that we have this foundation to build upon, I will run the regression with our full data frame to check for the robustness of these results once we include all the outliers.

```
##
## Call:
## ivreg(formula = score ~ lab + past_score + income + as.factor(race) |
##     closeness + past_score + income + as.factor(race), data = working_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -4.022353 -0.692941  0.005912  0.684590  3.355921
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.006e+00  1.742e-01  -5.776 8.14e-09 ***
## lab               3.320e-01  1.393e-01   2.384   0.0172 *
## past_score        1.019e+00  1.318e-02  77.294  < 2e-16 ***
## income            1.340e-05  9.067e-07  14.783  < 2e-16 ***
## as.factor(race)B -1.098e-01  1.096e-01  -1.002   0.3164
## as.factor(race)H -9.306e-02  1.018e-01  -0.914   0.3609
## as.factor(race)W -8.209e-02  1.070e-01  -0.767   0.4431
##
## Diagnostic tests:
##                   df1  df2 statistic p-value
## Weak instruments    1 4838   278.570  <2e-16 ***
## Wu-Hausman          1 4837     5.509   0.019 *
## Sargan              0   NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.024 on 4838 degrees of freedom
## Multiple R-Squared: 0.7064,  Adjusted R-squared: 0.706
## Wald test:  1890 on 6 and 4838 DF,  p-value: < 2.2e-16
```

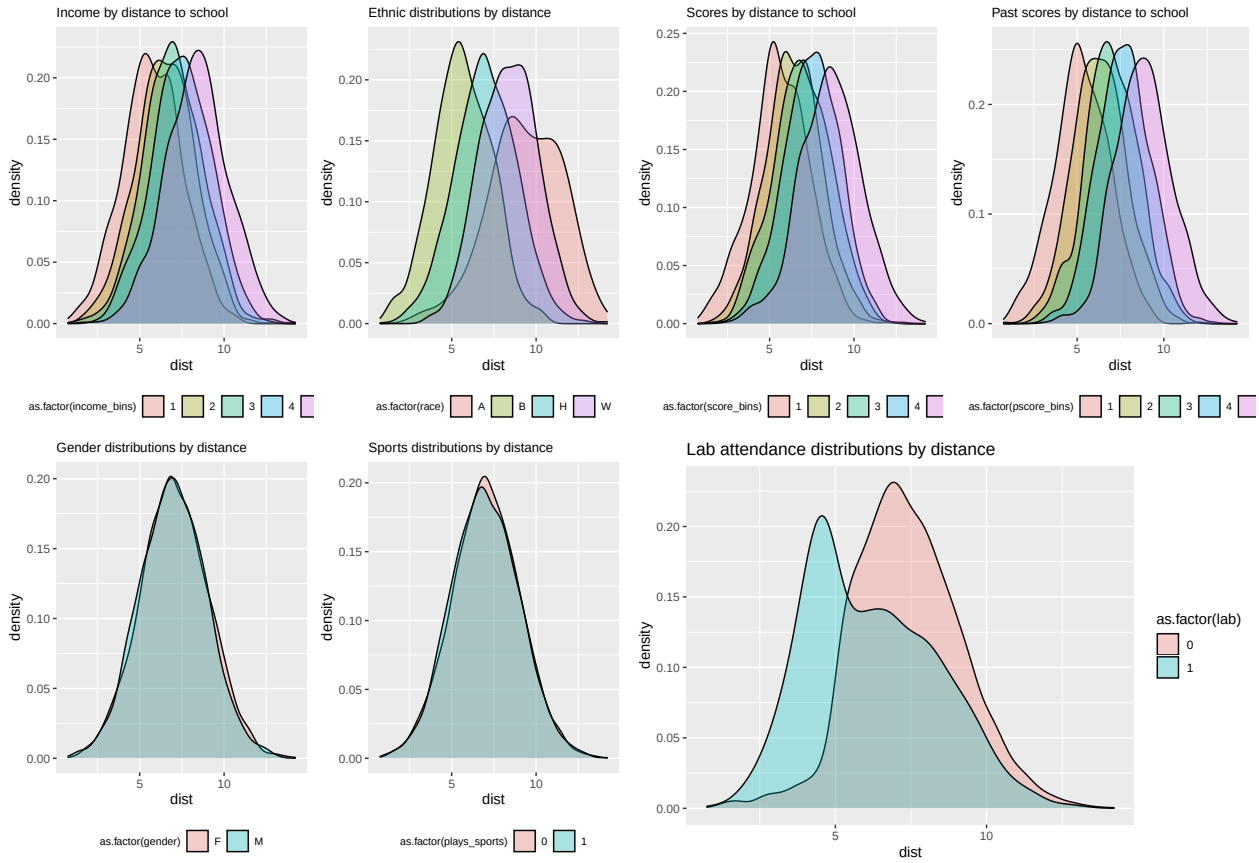
Notice then that even when we include outliers, the results (in terms of coefficient and statistical significance) are still very close to what we had when we had removed such outliers. Notice that this is indeed a (relatively) strong instrument statistically as well, as noted by the Weak Instruments test. Since we have argued that distance (or closeness) is a strong instrument economically, and that this also holds statistically, I believe this is representative and the true effect of attending the lab school.

Therefore, using the robust results, I conclude that the **true effectiveness of attending the lab school is (Using the IV) an increase by 0.332** points to one's standardized test score with a **standard error of 0.139**, and his result is **significant to the 0.0172 percentile.**

Question 6:

The same data set can be used for multiple purposes. Instead of estimating the treatment effect of LLS, suppose we want to know about the community/location surrounding LLS. What can you tell us? Be as specific as possible.

After our analysis in part 4 and part 5, we have found multiple descriptive items that one maybe interested in while using this data set.



In particular, we find that individuals who live close to the lab school are much more likely (in proportions) to attend the lab school relative to those who live far. Furthermore, we find that those who live close the lab school, on average and relative to those who live far, have less standardized testing scores (past and present), less income, and more likely to be of the ethnic group "B" and less likely to be ethnic group "W", "A", and "H". These are all very interesting findings, suggesting that the community around the lab school might have quite different composition relative to the rest population. Therefore, these characteristics of the surrounding community around the lab school may serve as motivation (not necessarily justification) for policy prescriptions such as education opportunity/financial aid for education for individuals who reside in this community (based on the evidence to suggest that income is on average lower in the surrounding community).