Homework 6

Mahdi El Amin

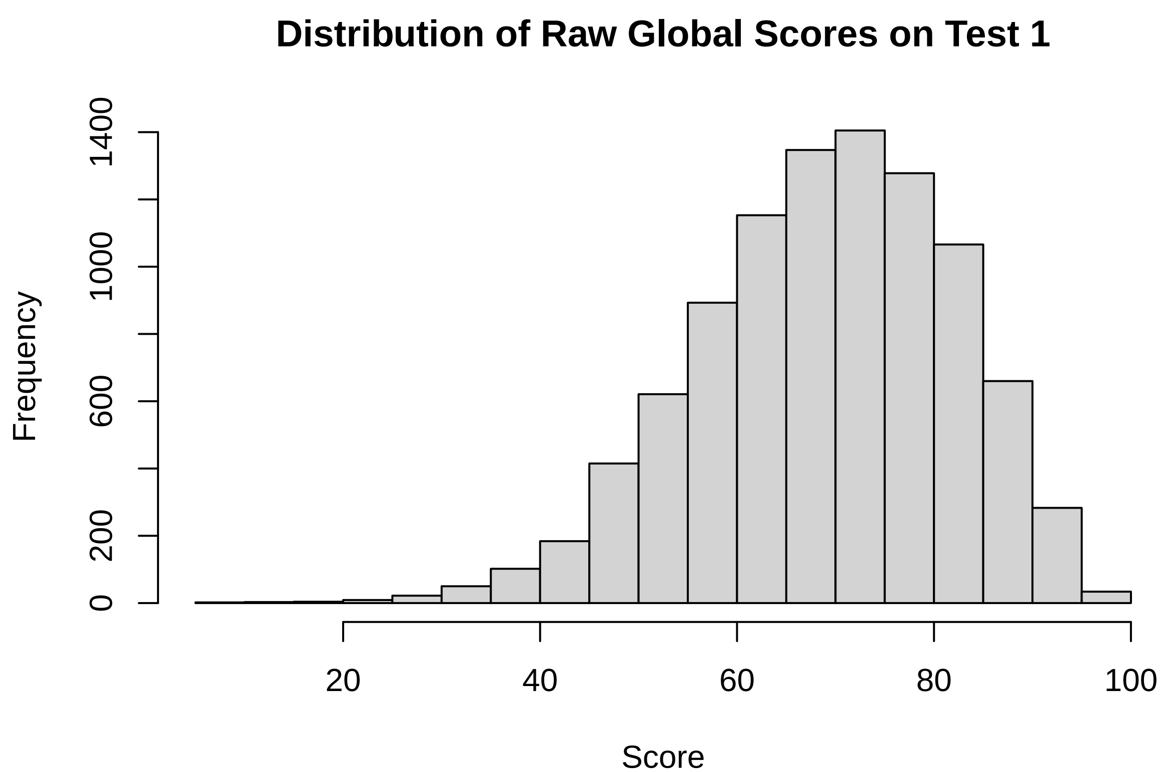
30/11/2021

Education in France

```r
panel97 <- read_dta("C:/Users/mahdi/Desktop/PSE/Measurements/Homework 6/panel97.dta")
```

(a)

```r
hist(panel97$gscore_t1, xlab = "Score", main = "Distribution of Raw Global Scores on Test 1")
```

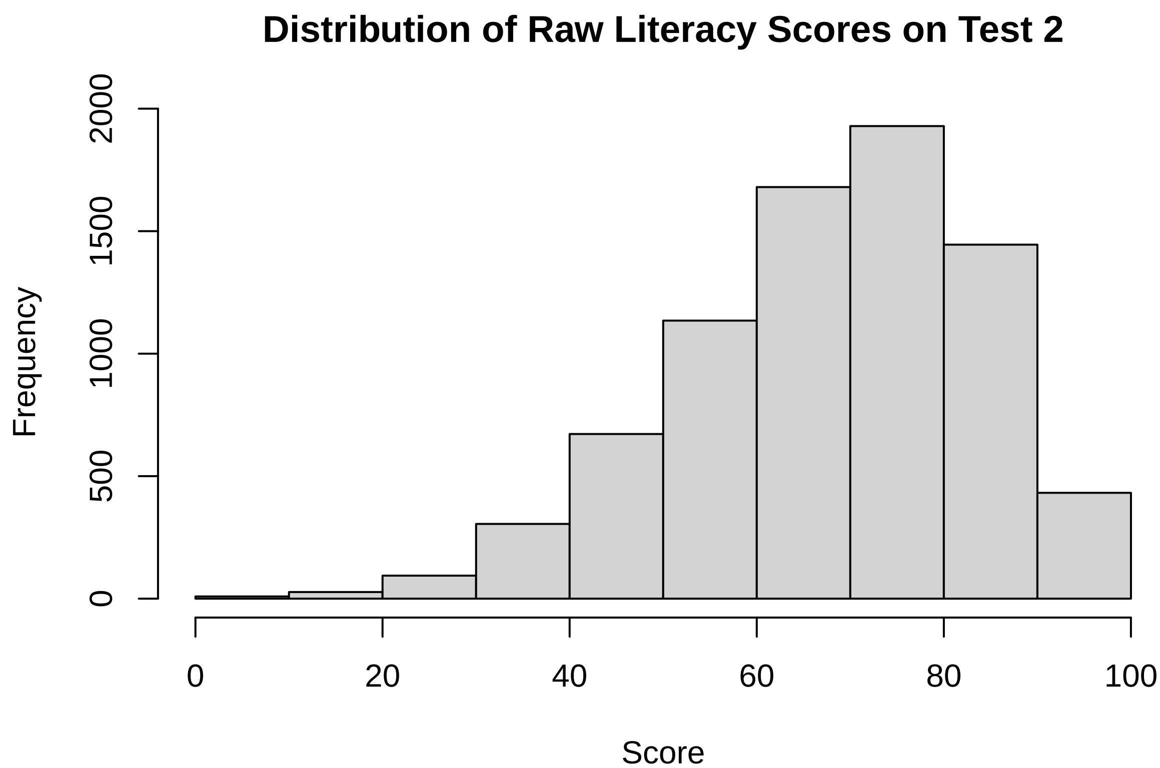


```
summary(panel97$gscore_t1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  9.091  60.494  69.753  68.995  78.395  98.980     110 
```

At the entry tests to primary school, the pupils scored an average of 69.753% with the lowest grade being 9.091% and the highest being 98.98%.

```
hist(panel97$fscore_t2, xlab = "Score", main = "Distribution of Raw Literacy Scores on Test 2")
```

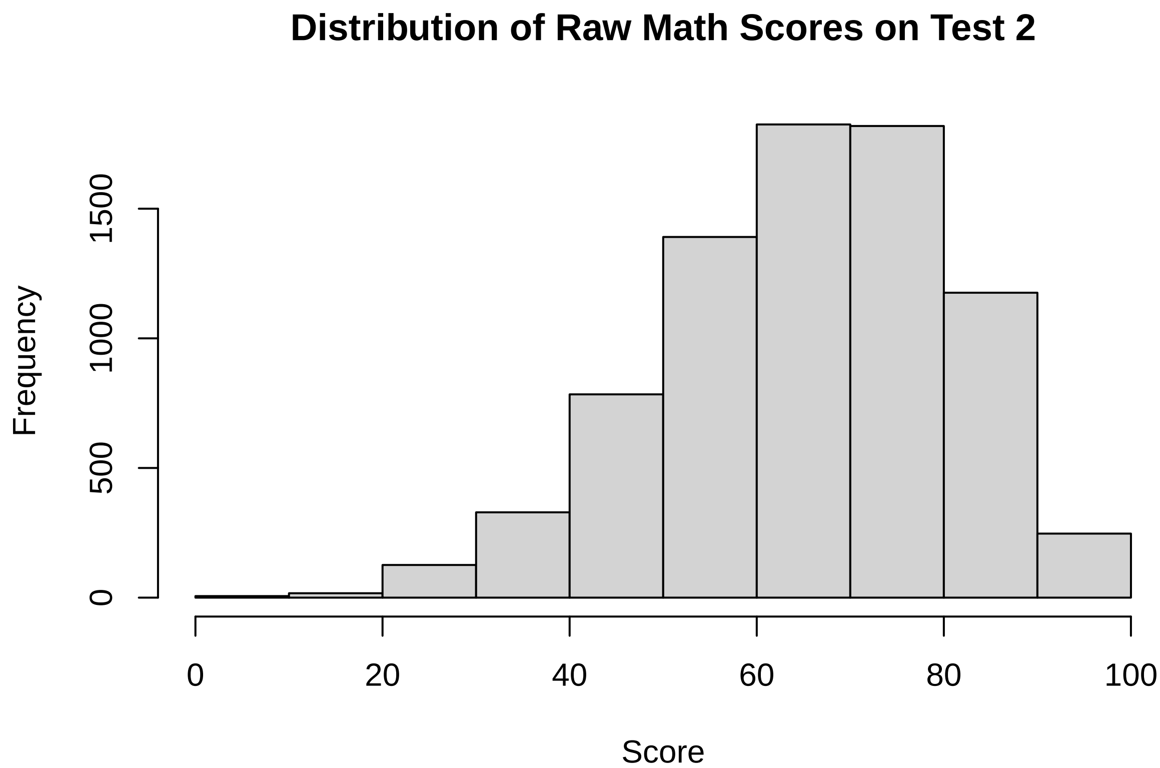



```
summary(panel97$fscore_t2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   57.45   69.23   67.63   79.12  100.00    1913 
```

In the literacy test at the beginning of the 3rd grade of primary school (second test) the pupils scored an average of 69.23% with the lowest grade being 0% and the highest being 100%.

```
hist(panel97$mscore_t2, xlab = "Score", main = "Distribution of Raw Math Scores on Test 2")
```

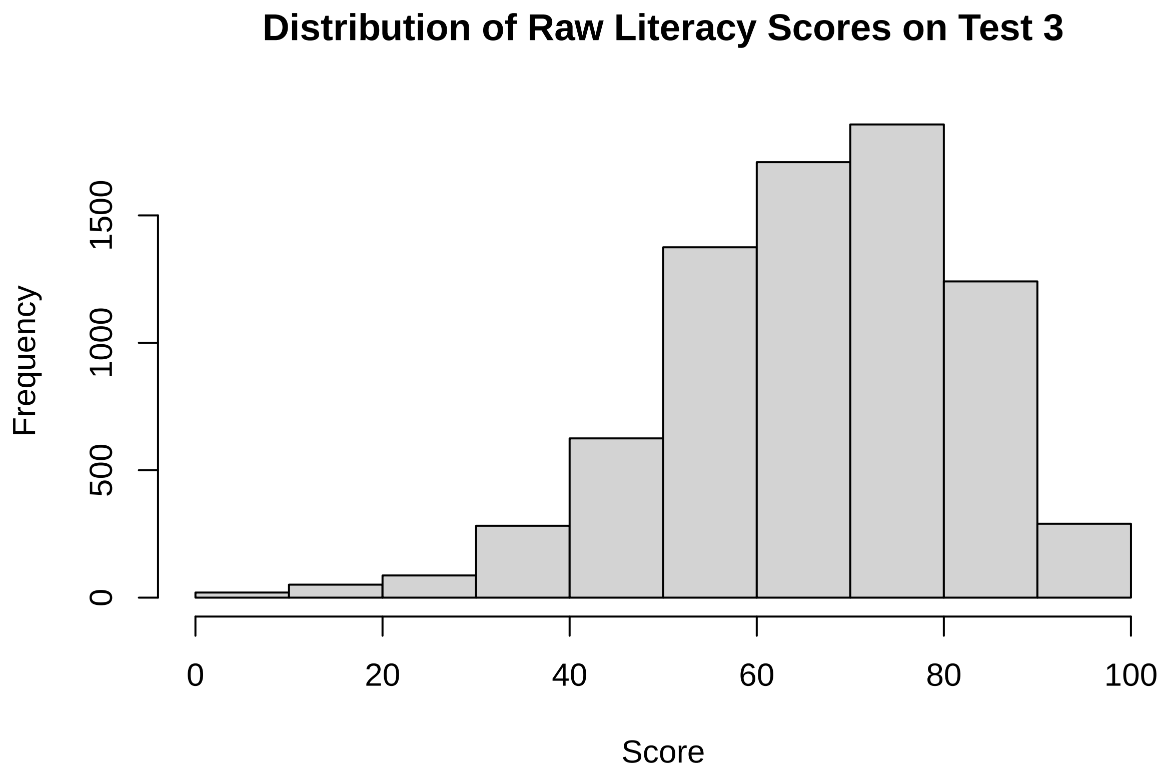



```
summary(panel97$mscore_t2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   56.25   67.50   66.04   77.50  100.00    1921
```

In the math test at the beginning of the 3rd grade of primary school (second test) the pupils scored an average of 67.5% with the lowest grade being 0% and the highest being 100%.

```
hist(panel97$fscore_t3, xlab = "Score", main = "Distribution of Raw Literacy Scores on Test 3")
```

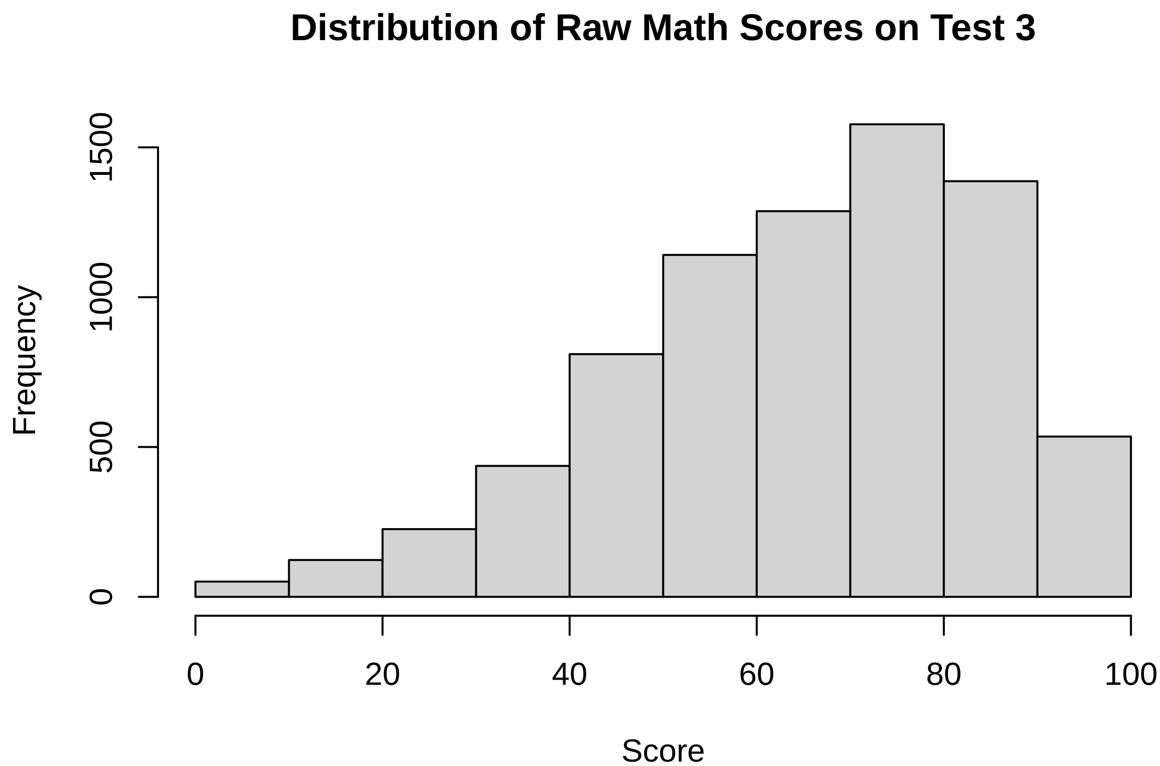



```
summary(panel97$fscore_t3)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.149  56.471  67.816  66.356  78.161 100.000    2104
```

In the literacy test at the end of primary school (third test) the pupils scored an average of 67.816% with the lowest grade being 1.149% and the highest being 100%.

```
hist(panel97$mscore_t3, xlab = "Score", main = "Distribution of Raw Math Scores on Test 3")
```




```
summary(panel97$mscore_t3)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   52.56   67.53   65.03   80.52  100.00    2067
```

In the math test at the end of primary school (third test) the pupils scored an average of 67.53% with the lowest grade being 0% and the highest being 100%.
Since we can get the global scores of the second and third tests using the average of the literacy and math scores in each year, we notice that global scores decrease with each test. We also notice that the literacy scores are always higher than the math scores but the difference never exceeds 2 percentage points. Finally, looking at the histograms, all the distributions are right skewed, meaning the majority of pupils pass their tests.

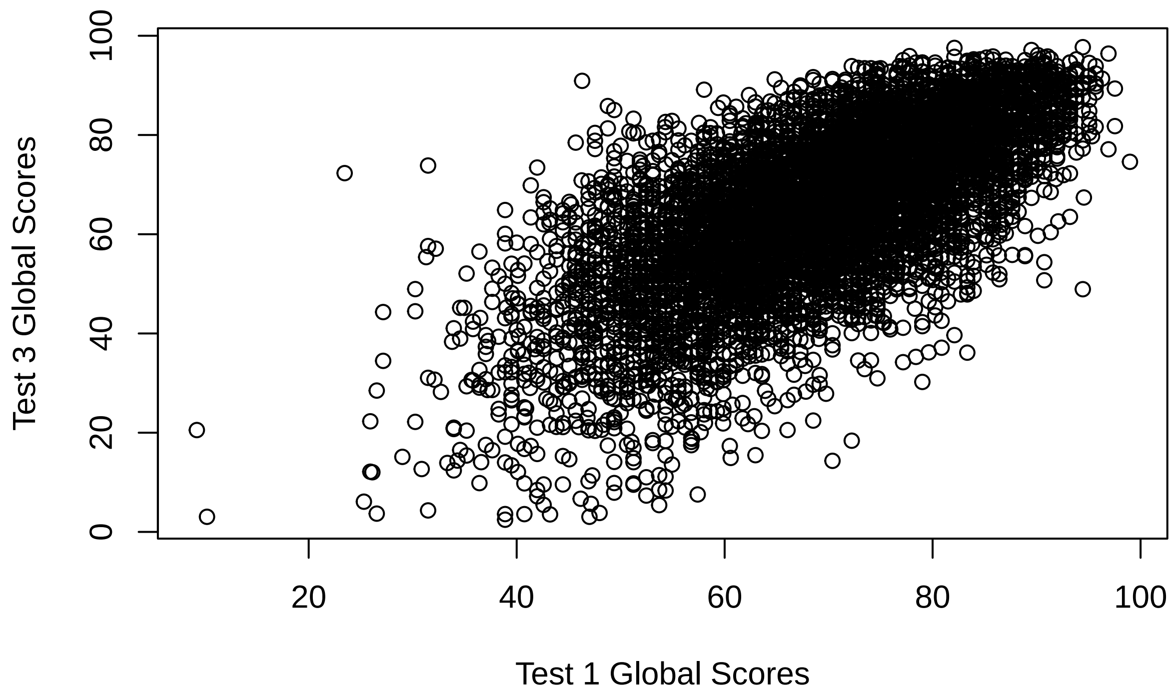
(b)

We won't check to see if there are missing values in the global scores columns for test 2 and 3, because in case there is but the pupil's scores for math and literacy are present in the dataset then we can calculate the global scores ourselves.
There are 110 missing values in the global scores of test 1.
There are 1913 missing values in the literacy scores of test 2.
There are 1921 missing values in the math scores of test 2.
There are 2104 missing values in the literacy scores of test 3.
There are 2067 missing values in the math scores of test 3.
These missing values would prevent us from studying the trajectories or progress of their respective students,

since they represent lack of information. Moreover, since around a 5th of the data has missing values in at least one of the test scores, this could lead to imprecise measurements in our experiment. Also, depending on the occupation of the parents of the pupils with missing values, this might also lead to biased measurements. A possible explanation for the increase of missing values with each test is the pupils failing more than once or dropping out of school.
An approach we can take to potential lower the amount of missing values is check for which entries we have the global scores of test 2 and one of either the literacy or math scores of that test, that way we can use the formula for the global score to calculate the missing score. This method also works for test 3.

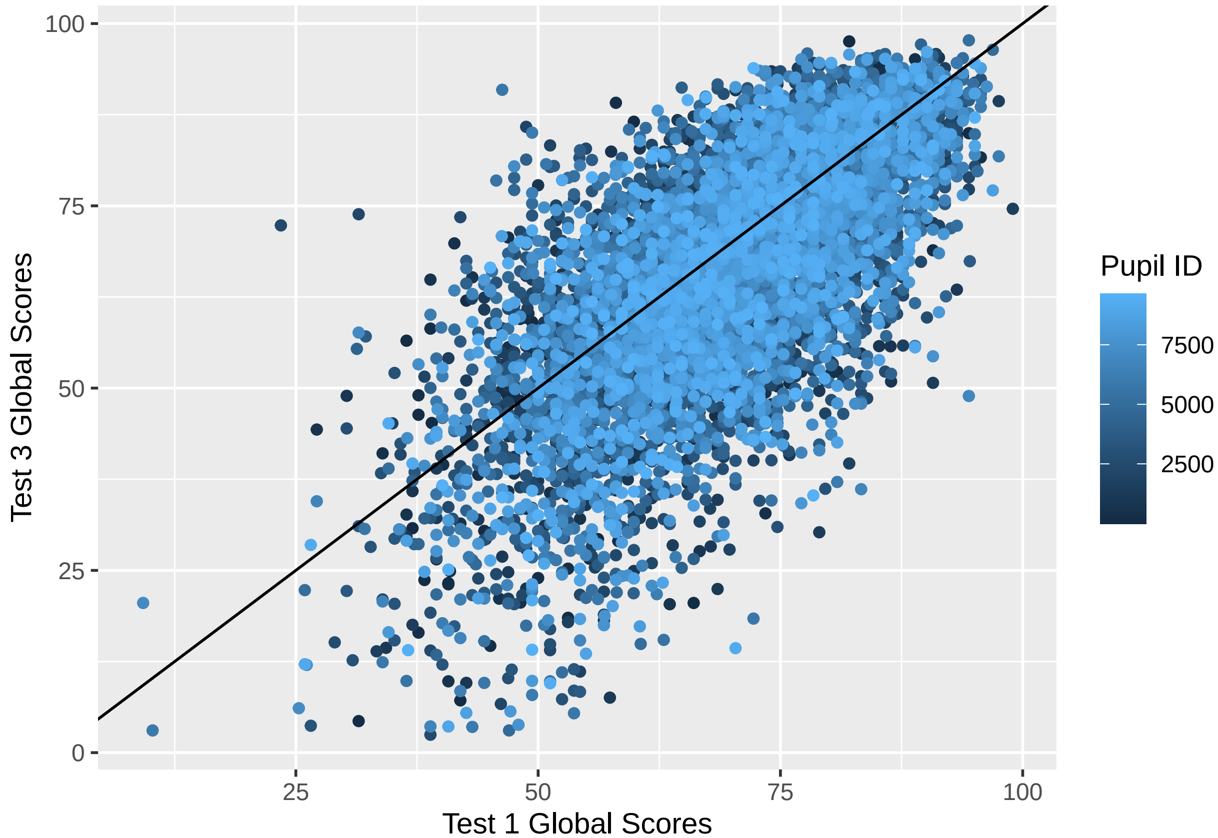
(c)

```
plot(panel97$gscore_t1, panel97$gscore_t3,
     xlab = "Test 1 Global Scores", ylab = "Test 3 Global Scores")
```



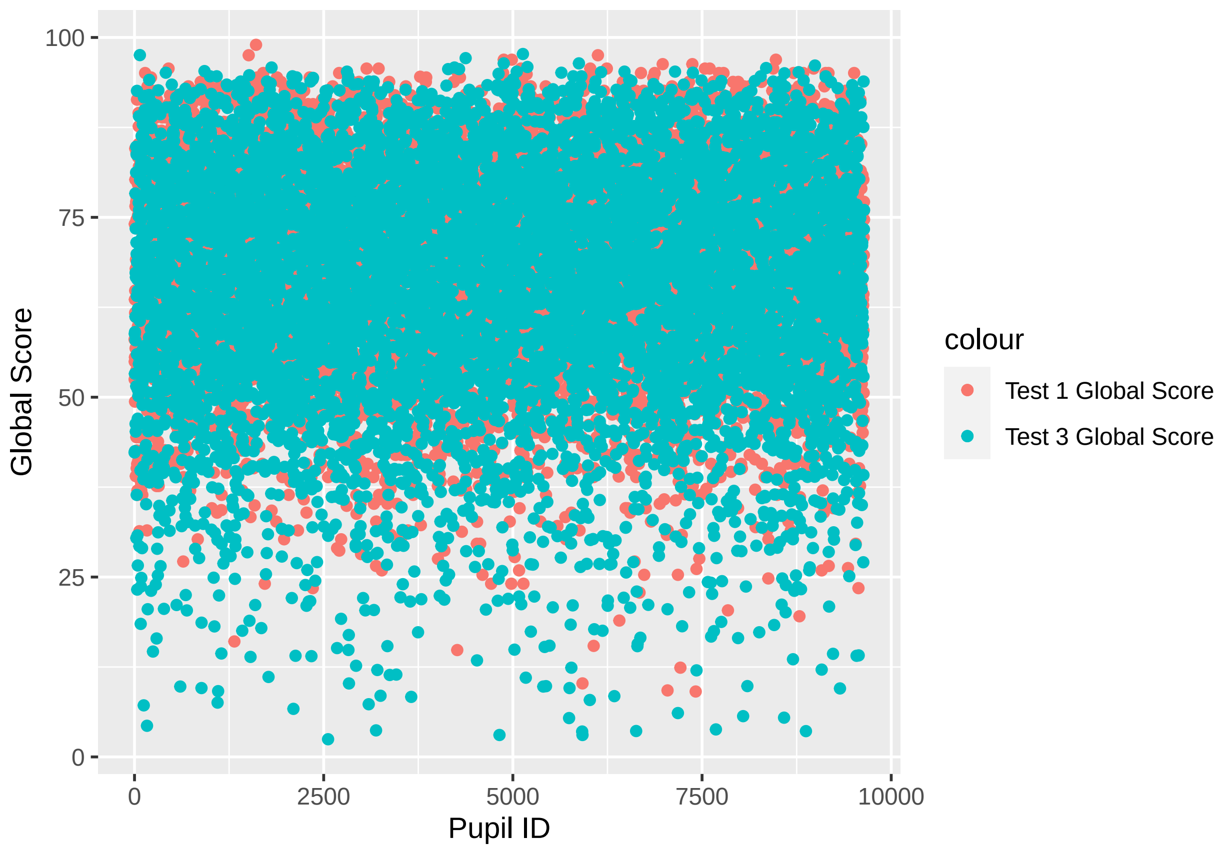
If we're trying to compare the test results to see which had higher grades scored in them we can simply add a straight line with slope 1 passing through the origin.

```
ggplot(panel97, aes(x = panel97$gscore_t1, y = panel97$gscore_t3, color = panel97$pupil_id)) +
  geom_point() +
  labs(x = "Test 1 Global Scores", y = "Test 3 Global Scores", color = "Pupil ID") +
  geom_abline(slope = 1)
```



This way, every point symbolizes a different pupil and if the point lies below the straight line then that means their test 1 global score was higher than their test 3 global score, and vice versa.
We notice a larger amount of points lying below the line than above it, therefore deducing what we proved in part (a) which is that Test 3 global scores are lower than those of Test 1.
Alternatively we can also plot both scores against the pupils' IDs

```
ggplot(panel97, aes(pupil_id)) +
  geom_point(aes(y=gscore_t1, colour="Test 1 Global Score")) +
  geom_point(aes(y=gscore_t3, colour="Test 3 Global Score")) +
  labs(y = "Global Score", x = "Pupil ID")
```



While this too isn't the best, we can see, especially when looking at the bottom of the plot, what we deduced in part (a) which is that Test 3 global scores are lower than those of Test 1.

(d)

In order to measure progress, we see if the global score of the third year of primary school is higher than that of the first year. If yes, we can say the pupil is progressing positively. If no, they are progressing negatively. And we use the same method to measure progress between the third year and the final year of primary school.
However, grades alone aren't enough to measure progress. As we saw earlier, the average global scores decrease with each test. Therefore, it is possible for a pupil to score the same or even slightly lower in a subsequent test but their percentile or decile rank to increase. Table 1 shows the proportion of students whose grades, percentiles or deciles have increased.
Note: between the first and third year we can only compare the global scores, but between the third and final years we can compare global, literacy, and math.

```
#progress between the first and the third year of primary school
#make sure we have a value for every gscore possible
panel97$gscore_t2[is.na(panel97$gscore_t2) & !is.na(panel97$fscore_t2) & !is.na(panel97$mscore_t2)] <-
  (panel97$fscore_t2[is.na(panel97$gscore_t2) & !is.na(panel97$fscore_t2) &
                       !is.na(panel97$mscore_t2)] +
     panel97$mscore_t2[is.na(panel97$gscore_t2) & !is.na(panel97$fscore_t2) &
                         !is.na(panel97$mscore_t2)])/2
panel97$progress13 <- ifelse(panel97$gscore_t2 > panel97$gscore_t1, 1, 0)
#progress between the third year and the final year of primary school
#make sure we have a value for every gscore possible
```

```r
panel97$gscore_t3[is.na(panel97$gscore_t3) & !is.na(panel97$fscore_t3) & !is.na(panel97$mscore_t3)] <-
  (panel97$fscore_t3[is.na(panel97$gscore_t3) & !is.na(panel97$fscore_t3) &
                       !is.na(panel97$mscore_t3)] +
     panel97$mscore_t3[is.na(panel97$gscore_t3) & !is.na(panel97$fscore_t3) &
                         !is.na(panel97$mscore_t3)])/2
panel97$progress35 <- ifelse(panel97$gscore_t3 > panel97$gscore_t2, 1, 0)

#progress between the first and third year of primary school looking at percentiles and deciles
panel97$progress13p <- ifelse(panel97$p_gscore_t2 > panel97$p_gscore_t1, 1, 0)
panel97$progress13d <- ifelse(panel97$d_gscore_t2 > panel97$d_gscore_t1, 1, 0)
#progress between the third and final year of primary school looking at percentiles and deciles
#global scores
panel97$progress35p <- ifelse(panel97$p_gscore_t3 > panel97$p_gscore_t2, 1, 0)
panel97$progress35d <- ifelse(panel97$d_gscore_t3 > panel97$d_gscore_t2, 1, 0)
#literacy scores
panel97$progress35pl <- ifelse(panel97$p_fscore_t3 > panel97$p_fscore_t2, 1, 0)
panel97$progress35dl <- ifelse(panel97$d_fscore_t3 > panel97$d_fscore_t2, 1, 0)
#math scores
panel97$progress35pm <- ifelse(panel97$p_mscore_t3 > panel97$p_mscore_t2, 1, 0)
panel97$progress35dm <- ifelse(panel97$d_mscore_t3 > panel97$d_mscore_t2, 1, 0)

df <- data.frame(Method = c("Global Scores Increased", "Percentile of Global Score Increased",
                            "Percentile of Literacy Score Increased", "Percentile of Math Score Increase
                            "Decile of Global Score Increased", "Decile of Literacy Score Increased",
                            "Decile of Math Score Increased"),
                 B13 = c(length(which(panel97$progress13 == 1))/length(which(!is.na(panel97$progress13)
                         length(which(panel97$progress13p == 1))/
                           length(which(!is.na(panel97$progress13p))), NA, NA,
                         length(which(panel97$progress13d == 1))/
                           length(which(!is.na(panel97$progress13d))), NA, NA),
                 B35 = c(length(which(panel97$progress35 == 1))/length(which(!is.na(panel97$progress35)
                         length(which(panel97$progress35p == 1))/
                           length(which(!is.na(panel97$progress35p))),
                         length(which(panel97$progress35pl == 1))/
                           length(which(!is.na(panel97$progress35pl))),
                         length(which(panel97$progress35pm == 1))/
                           length(which(!is.na(panel97$progress35pm))),
                         length(which(panel97$progress35d == 1))/
                           length(which(!is.na(panel97$progress35d))),
                         length(which(panel97$progress35dl == 1))/
                           length(which(!is.na(panel97$progress35dl))),
                         length(which(panel97$progress35dm == 1))/
                           length(which(!is.na(panel97$progress35dm)))))
options(knitr.kable.NA = '-')
kable(df, col.names = c("Criteria", "Between 1st and 3rd Year", "Between 3rd and Final Year"),
      caption = "Number of Students that Progressed based on Different Criteria") %>%
  row_spec(0,bold=TRUE)
```

Regardless of the criteria used, we notice that less than half the students progressed between the first and third year of primary school. However the global scores and percentiles of around half of the students have progressed between the third and final year, and the deciles of around two fifths of them progressed as well.

Table 1: Number of Students that Progressed based on Different Criteria

Criteria	Between 1st and 3rd Year	Between 3rd and Final Year
Global Scores Increased	0.3741595	0.5168977
Percentile of Global Score Increased	0.4275544	0.5274879
Percentile of Literacy Score Increased	-	0.5300798
Percentile of Math Score Increased	-	0.5100763
Decile of Global Score Increased	0.3357943	0.4116181
Decile of Literacy Score Increased	-	0.4178944
Decile of Math Score Increased	-	0.4096183

(e)

```
#global scores between first and third year
p13 <- matrix(nrow = 10, ncol = 10)
for (i in 1:10) {
  for (j in 1:10) {
    p13[i,j] <- (sum(panel97$d_gscore_t1 == i & panel97$d_gscore_t2 == j, na.rm = TRUE)/
                   sum(!is.na(panel97$d_gscore_t1) & !is.na(panel97$d_gscore_t2), na.rm = TRUE))/
      (sum(panel97$d_gscore_t1 == i, na.rm = TRUE)/
         sum(!is.na(panel97$d_gscore_t1), na.rm = TRUE))
  }
}
#global scores between third and final year
p35 <- matrix(nrow = 10, ncol = 10)
for (i in 1:10) {
  for (j in 1:10) {
    p35[i,j] <- (sum(panel97$d_gscore_t2 == i & panel97$d_gscore_t3 == j, na.rm = TRUE)/
                   sum(!is.na(panel97$d_gscore_t3) & !is.na(panel97$d_gscore_t2), na.rm = TRUE))/
      (sum(panel97$d_gscore_t2 == i, na.rm = TRUE)/
         sum(!is.na(panel97$d_gscore_t2), na.rm = TRUE))
  }
}

kable(round(p13*100),
      caption = "Transition matrix across deciles of the global score distribution between test 1 and te
  add_footnote("The columns represent the decile rank in Test 1 while the rows represent the decile rank
```

```
kable(round(p35*100),
      caption = "Transition matrix across deciles of the global score distribution between test 2 and te
  add_footnote("The columns represent the decile rank in Test 2 while the rows represent the decile rank
```

Looking at Table 2, the highest probabilities based on a student's test 1 score are always either remaining in the same decile or moving only one decile spot either higher or lower. If the initial test score was in the first decile, then the highest probability is for the student to remain in the first decile in the second decile. If the initial score was in the second decile, the highest probability is also to remain in the same decile. If the intial score was in the third decile, the highest probability is actually to drop by one decile spot. And so on, and so forth.

Table 3 follows a similar logic and pattern as Table 2. We notice that student likelihood of staying in the first and last deciles between tests 2 and 3 are however higher than those between tests 1 and 2.

Table 2: Transition matrix across deciles of the global score distribution between test 1 and test 2

29	13	10	5	3	2	2	0	1	0
28	20	13	10	6	4	3	1	1	0
17	19	15	16	10	8	6	3	2	0
11	16	19	18	13	11	9	5	3	1
8	12	15	14	14	14	12	9	5	2
3	9	12	14	16	16	14	11	9	5
2	5	8	11	15	12	16	18	12	9
1	3	5	7	10	17	16	16	18	16
1	2	2	3	8	12	13	21	26	22
0	1	2	2	4	5	10	15	24	44

[a] The columns represent the decile rank in Test 1 while the rows represent the decile rank in Test 2

Table 3: Transition matrix across deciles of the global score distribution between test 2 and test 3

41	22	16	10	6	3	1	1	0	0
15	23	20	16	11	6	4	2	1	0
9	21	19	19	14	9	5	4	1	0
4	12	14	18	15	15	13	5	2	1
3	8	12	14	15	17	13	11	6	1
1	4	9	12	17	15	16	17	8	3
1	2	6	8	12	18	16	16	15	5
0	2	2	5	7	11	19	22	21	12
0	1	1	1	4	7	11	18	29	26
0	0	0	0	1	2	4	11	24	57

[a] The columns represent the decile rank in Test 2 while the rows represent the decile rank in Test 3

(f)

(a)

```r
#Test 1
diff_t1 <- mean(panel97$gscore_t1[panel97$birthm == 1], na.rm = TRUE) -
  mean(panel97$gscore_t1[panel97$birthm == 12], na.rm = TRUE)
t1t <- t.test(panel97$gscore_t1[panel97$birthm == 1], panel97$gscore_t1[panel97$birthm == 12])
```

In Test 1, people born in January scored 7.3426237 percentage points more than those born in December with a [6.1201113,8.565136] 95% confidence interval.

```r
#Test 2
diff_t2 <- mean(panel97$gscore_t2[panel97$birthm == 1], na.rm = TRUE) -
  mean(panel97$gscore_t2[panel97$birthm == 12], na.rm = TRUE)
t2t <- t.test(panel97$gscore_t2[panel97$birthm == 1], panel97$gscore_t2[panel97$birthm == 12])
```

In Test 2, people born in January scored 6.0373013 percentage points more than those born in December with a [4.4224327,7.6521699] 95% confidence interval.

```r
#Test 3
diff_t3 <- mean(panel97$gscore_t3[panel97$birthm == 1], na.rm = TRUE) -
  mean(panel97$gscore_t3[panel97$birthm == 12], na.rm = TRUE)
t3t <- t.test(panel97$gscore_t3[panel97$birthm == 1], panel97$gscore_t3[panel97$birthm == 12])
```

In Test 3, people born in January scored 4.7438612 percentage points more than those born in December with a [2.9204792,6.5672432] 95% confidence interval.
We notice the difference in scores between the two groups decreases with each test, and the confidence intervals get wider.

(b)

```r
#Test 1
diff_occ_t1 <- mean(panel97$gscore_t1[panel97$parent_occ == 3], na.rm = TRUE) -
  mean(panel97$gscore_t1[panel97$parent_occ == 6], na.rm = TRUE)
t1t_occ <- t.test(panel97$gscore_t1[panel97$parent_occ == 3],
                  panel97$gscore_t1[panel97$parent_occ == 6])
```

In Test 1, children of executives/professionals scored 10.3562316 percentage points more than those of blue-collar workers with a [9.6453483,11.0671149] 95% confidence interval.

```r
#Test 2
diff_occ_t2 <- mean(panel97$gscore_t2[panel97$parent_occ == 3], na.rm = TRUE) -
  mean(panel97$gscore_t2[panel97$parent_occ == 6], na.rm = TRUE)
t2t_occ <- t.test(panel97$gscore_t2[panel97$parent_occ == 3],
                  panel97$gscore_t2[panel97$parent_occ == 6])
```

In Test 2, children of executives/professionals scored 11.6250714 percentage points more than those of blue-collar workers with a [10.7389214,12.5112214] 95% confidence interval.

```r
#Test 3
diff_occ_t3 <- mean(panel97$gscore_t3[panel97$parent_occ == 3], na.rm = TRUE) -
  mean(panel97$gscore_t3[panel97$parent_occ == 6], na.rm = TRUE)
t3t_occ <- t.test(panel97$gscore_t3[panel97$parent_occ == 3],
                  panel97$gscore_t3[panel97$parent_occ == 6])
```

In Test 3, children of executives/professionals scored 16.2057454 percentage points more than those of blue-collar workers with a [15.2485003,17.1629905] 95% confidence interval.
Unlike the evolution of the January vs. December global test score gap in the different years, the difference in scores increases in the evolution of the global test score gap between children of executives/professionals and children of blue-collars. Meanwhile, the confidence intervals in the latter evolution don't get wider nor thinner.

(c) The comparison of test scores of children born in January vs. December is likely to underestimate the true effect of relative age differences on test scores, since in evaluation children born in either of these months are put within the same age group, or year, and their scores are compared to those of children of other age/year groups. So the difference we calculated in part (a) will lead to an imprecise comparison. It may be better to compare children born within the same 6 months than those born in the same year.

Human Development Index: Old vs. New Definition

```r
hdi_2019 <- read_dta('C:/Users/mahdi/Desktop/PSE/Measurements/Homework 6/hdi_2019.dta')
```

Question 1

(a) The old HDI averages the three indices while the new one takes the cubic root of the product of the indices. The health index is the only one that remained unchanged between the two HDIs.

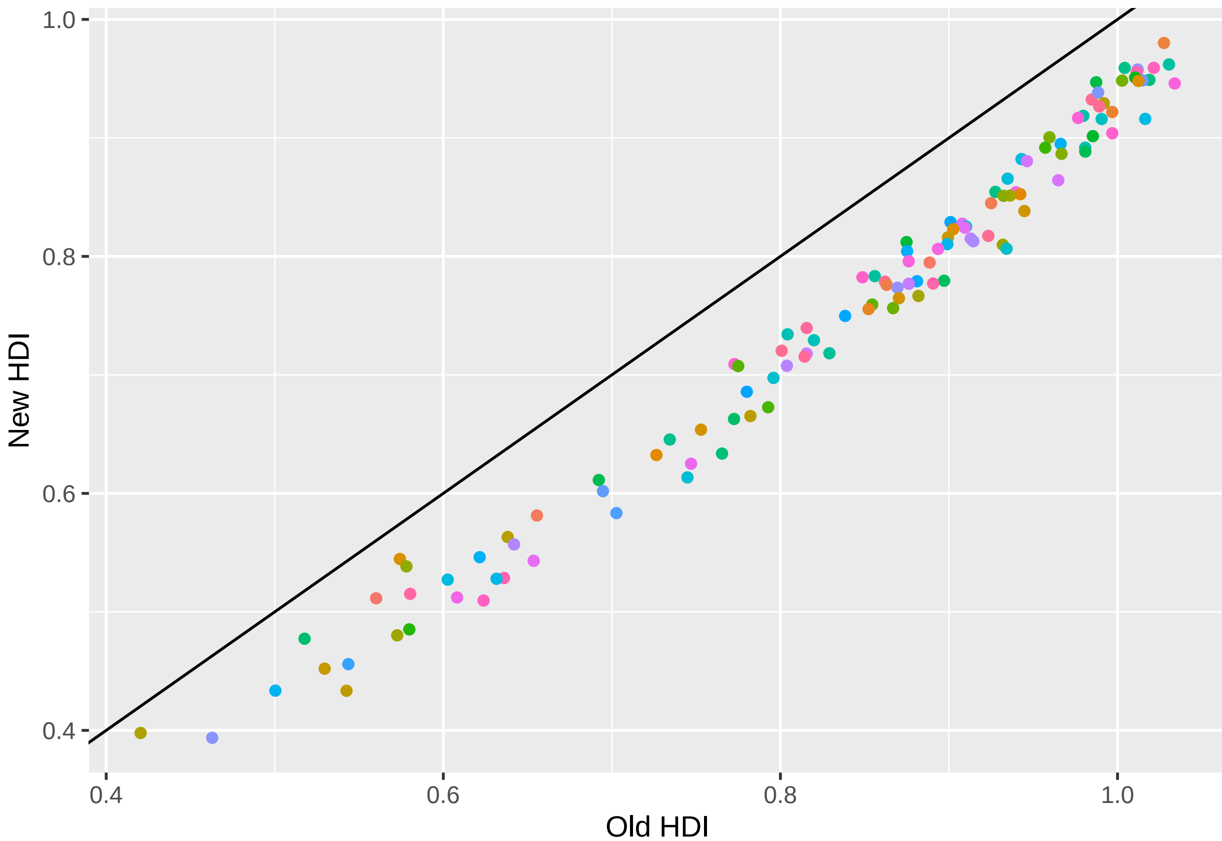
(b) The education indices of both HDIs depend on two independent variables each. In the old HDI, the education index was a weighted average of the adult literacy rate with the gross enrollment rate in primary, secondary and tertiary education (each divided by 100). While, in the new HDI, the education index is the average (not weighted) of the average years of schooling for individuals aged 25 (divided by 15) or older with the expected years of schooling (divided by 18).

(c) The income indices of both HDIs depend on the logarithm function (ln(.)). The difference is that the old HDI uses the GDP per capita of a country, while the new HDI uses the GNI per capita. In addition to a difference in the denominator value.

Question 2

```r
hdi_2019$oldhdi <- ((hdi_2019$le - 20)/(85-20))/3 +
  (2*(hdi_2019$al/100)/3 + (hdi_2019$ge/100)/3)/3 +
  ((ln(hdi_2019$gdp) - ln(100))/(ln(40000) - ln(100)))/3
hdi_2019$newhdi <- (((hdi_2019$le - 20)/(85-20))*
                      ((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2)*
                      ((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(100))))^(1/3)
ggplot(hdi_2019, aes(x = oldhdi, y = newhdi, color = country)) +
```

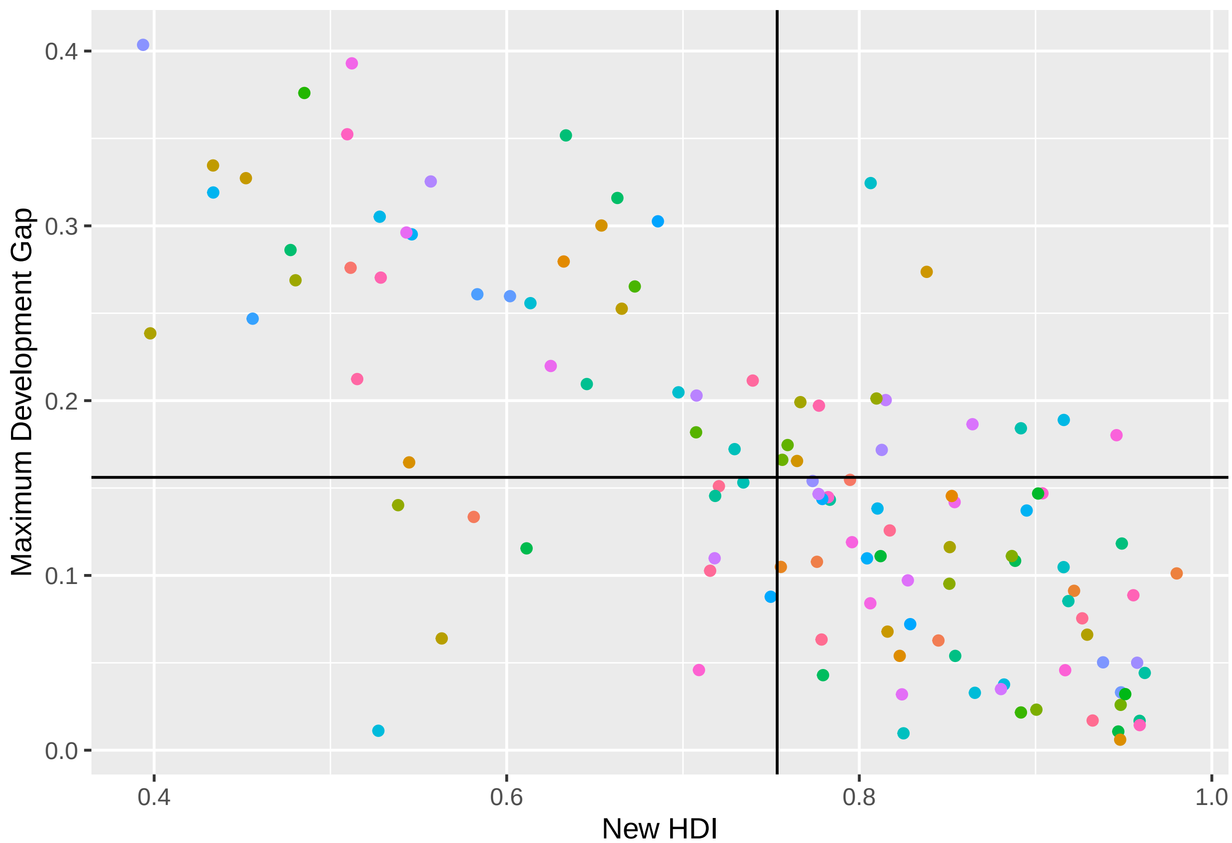
```r
  geom_point() + labs(x = "Old HDI", y = "New HDI") +
  geom_abline(slope = 1) + theme(legend.position = "none")
```



Each color/point in the above graph represents a different country. Looking at the above scatter plot, we can deduce from the fact that all points are below the straight line of slope 1 passing through the origin that the old HDI always yielded a greater result than the new one. Moreover, we notice that all values of the new HDI are below 1 while the old HDI had results slightly greater than 1.

Question 3

```r
hdi_2019$maxdiff <- apply(data.frame(
  a = abs(((hdi_2019$le - 20)/(85-20)) - (((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(100))))),
  b = abs(((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2) - ((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(1
  c = abs(((hdi_2019$le - 20)/(85-20)) - ((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2))), 1, max)
ggplot(hdi_2019, aes(x = newhdi, y = maxdiff, color = country)) +
  geom_point() + labs(x = "New HDI", y = "Maximum Development Gap") +
  theme(legend.position = "none") + geom_hline(yintercept = mean(hdi_2019$maxdiff)) +
  geom_vline(xintercept = mean(hdi_2019$newhdi))
```



In order to measure whether a country has uneven development we look at the maximum difference in any two of the three indices used to calculate the new version of the HDI for that country. We then calculate the average of these differences and any country that's maximum difference is greater than average is considered to have uneven development.
In the above graph, we plotted the maximum development gap against the new HDI for each country. The horizontal line is the average of the maximum development gaps and the verticle line is the average of the new HDIs. If a country, i.e. a point, is above the horizontal line and to the left of the vertical line, that means that they have uneven development and a low HDI. 36 countries lie there. A country that is below the horizontal line and to the right of the vertical line does not have uneven development and has a high HDI. 58 countries lie there. The 26 remaining countries lie in one of the two remains sections of the graph. The small number of countries in the remaining sections compared to the other countries can be used to claim that the new version of the HDI penalizes countries with uneven development. However, more observations are needed in order to definitively make that claim.

Question 4

(a)

```
#calculate percentiles
hdi_2019$gni_perc <- ecdf(hdi_2019$gni)(hdi_2019$gni)
#increase life expectancy by 3 years in 50% poorest countries
hdi_2019$oldhdi_p <- ifelse(hdi_2019$gni_perc <= 0.5,
                            ((hdi_2019$le + 3 - 20)/(85-20))/3 +
                              (2*(hdi_2019$al/100)/3 + (hdi_2019$ge/100)/3)/3 +
                              ((ln(hdi_2019$gdp) - ln(100))/(ln(40000) - ln(100)))/3,
                            ((hdi_2019$le - 20)/(85-20))/3 +
```

```r
                              (2*(hdi_2019$al/100)/3 + (hdi_2019$ge/100)/3)/3 +
                              ((ln(hdi_2019$gdp) - ln(100))/(ln(40000) - ln(100)))/3)
#increase life expectancy by 3 years in 50% richest countries
hdi_2019$oldhdi_r <- ifelse(hdi_2019$gni_perc > 0.5,
                            ((hdi_2019$le + 3 - 20)/(85-20))/3 +
                              (2*(hdi_2019$al/100)/3 + (hdi_2019$ge/100)/3)/3 +
                              ((ln(hdi_2019$gdp) - ln(100))/(ln(40000) - ln(100)))/3,
                            ((hdi_2019$le - 20)/(85-20))/3 +
                              (2*(hdi_2019$al/100)/3 + (hdi_2019$ge/100)/3)/3 +
                              ((ln(hdi_2019$gdp) - ln(100))/(ln(40000) - ln(100)))/3)
poorest_avrg <- mean(hdi_2019$oldhdi_p)
richest_avrg <- mean(hdi_2019$oldhdi_r)
```

Increasing life expectancy by 3 years in the 50% poorest countries leads to an average of 0.8444133 under the old version of the index, and increasing life expectancy by 3 years in the 50 % richest countries leads to an average of 0.8444133. Hence they have the same effect on the average of the old version of HDI. So no policy option is favored over the other.

(b)

```r
#increase life expectancy by 3 years in 50% poorest countries
hdi_2019$newhdi_p <- ifelse(hdi_2019$gni_perc <= 0.5,
                            (((hdi_2019$le + 3 - 20)/(85-20))*
                               ((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2)*
                               ((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(100))))^(1/3),
                            (((hdi_2019$le - 20)/(85-20))*
                               ((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2)*
                               ((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(100))))^(1/3))
#increase life expectancy by 3 years in 50% richest countries
hdi_2019$newhdi_r <- ifelse(hdi_2019$gni_perc > 0.5,
                            (((hdi_2019$le + 3 - 20)/(85-20))*
                               ((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2)*
                               ((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(100))))^(1/3),
                            (((hdi_2019$le - 20)/(85-20))*
                               ((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2)*
                               ((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(100))))^(1/3))
new_poorest_avrg <- mean(hdi_2019$newhdi_p)
new_richest_avrg <- mean(hdi_2019$newhdi_r)
```

Increasing life expectancy by 3 years in the 50% poorest countries leads to an average of 0.7597211 under the new version of the index, and increasing life expectancy by 3 years in the 50 % richest countries leads to an average of 0.7607029. Since the latter average is higher, the second policy option would be favored.

(c)

```r
gdp_inc <- (ln(40000)-ln(100))/65
hdi_2019$oldhdi_le <- ((hdi_2019$le + 1 - 20)/(85-20))/3 +
  (2*(hdi_2019$al/100)/3 + (hdi_2019$ge/100)/3)/3 +
  ((ln(hdi_2019$gdp) - ln(100))/(ln(40000) - ln(100)))/3
hdi_2019$oldhdi_gdp <- ((hdi_2019$le - 20)/(85-20))/3 +
  (2*(hdi_2019$al/100)/3 + (hdi_2019$ge/100)/3)/3 +
  ((ln(hdi_2019$gdp) + gdp_inc - ln(100))/(ln(40000) - ln(100)))/3
```

We run the above code to obtain the values of the old version of the HDI after adding one year to life expectancy and the values after increasing GDP by $\frac{ln40000-ln100}{65}$. We see that all but 38 values are equal, however the maximum difference between a country's calculated HDI when life expectancy is increased and when GDP is increased is 2.220446×10^{-16}, which is negligible. So we can say the two effects are equivalent. Moreover we can show equivalence as follows:

$$\frac{\partial HDI_i^{old}}{\partial LE_i} = \frac{1}{3}(\frac{1}{65})$$

$$\frac{\partial HDI_i^{old}}{\partial GDP_i} = \frac{1}{3}(\frac{\Delta lnGDP_i^*}{ln40000 - ln100})$$

$$\frac{\partial HDI_i^{old}}{\partial LE_i} = \frac{\partial HDI_i^{old}}{\partial GDP_i} \Rightarrow \Delta lnGDP_i^* = \frac{ln40000 - ln100}{65}$$

(d)

$$\frac{\partial HDI_i^{new}}{\partial LE_i} = \frac{1}{3}(\frac{1}{65})(\frac{LE_i - 20}{85 - 20})^{-\frac{2}{3}}[\frac{1}{2}(\frac{MYS_i}{15}) + \frac{1}{2}(\frac{EYS_i}{18})]^{\frac{1}{3}}(\frac{lnGNI_i - ln100}{ln75000 - ln100})^{\frac{1}{3}}$$

$$\frac{\partial HDI_i^{new}}{\partial GNI_i} = (\frac{LE_i - 20}{85 - 20})^{\frac{1}{3}}[\frac{1}{2}(\frac{MYS_i}{15}) + \frac{1}{2}(\frac{EYS_i}{18})]^{\frac{1}{3}}\frac{1}{3}\frac{\Delta GNI_i^*}{ln75000 - ln100}(\frac{lnGNI_i - ln100}{ln75000 - ln100})^{-\frac{2}{3}}$$

$$\frac{\partial HDI_i^{new}}{\partial LE_i} = \frac{\partial HDI_i^{new}}{\partial GNI_i} \Rightarrow \Delta GNI_i^* = \frac{ln75000 - ln100}{65}\frac{(\frac{lnGNI_i - ln100}{ln75000 - ln100})}{(\frac{LE_i - 20}{85 - 20})}$$

To test this empirically:

```
gni_inc <- ((ln(75000)-ln(100))/65)*
  ((((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(100))))/
     (((hdi_2019$le - 20)/(85-20))))
hdi_2019$newhdi_le <- (((hdi_2019$le + 1 - 20)/(85-20))*
                        ((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2)*
                        ((ln(hdi_2019$gni) - ln(100))/(ln(75000) - ln(100))))^(1/3)
hdi_2019$newhdi_gni <- (((hdi_2019$le - 20)/(85-20))*
                         ((hdi_2019$mys/15)/2 + (hdi_2019$eys/18)/2)*
                         ((ln(hdi_2019$gni) + gni_inc - ln(100))/(ln(75000) - ln(100))))^(1/3)
```

Similarly as before, we only get 40 cases that don't match but the maximum difference is 1.110223×10^{-16} which is negligible, so we can say the two effects are equivalent.

(e) From what we've seen from parts (a) and (b) of this exercise, the new version of the HDI index can increase in value even if only the richer countries develop while poorer ones remain the same. In fact, it increases more when richer countries develop than when poorer ones do. That would give an inaccurate perspective of human development; leading us to believe there is more development than there actually is.