ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΜΑΘΗΜΑ:Προγραμματιστικά εργαλεία και τεχνολογιές για
την επιστήμη δεδομένων
ΤΙΤΛΟΣ ΕΡΓΑΣΙΑΣ:Exploratory Data Analysis using R
ΜΕΤΑΠΤΥΧΙΑΚΟΣ ΦΟΙΤΗΤΗΣ: Αντώνιος Προμπονάς
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ:Επιστήμη Δεδομένων και
Μηχανική Μάθηση
ΑΜ: 03400232
EMAIL: antonisprompo@gmail.com

ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ: 2023-2024

1 ΕΙΣΑΓΩΓΗ

Μερικά χρόνια πριν, ένα από τα πιο καταξιωμένα και διακεκριμένα πανεπιστήμια της Ιταλίας, πραγματοποίησε μία πολύ ενδιαφέρουσα μελέτη. Η μελέτη αυτή εστίασε στην επίδοση 1161 δεκαπεντάχρονων μαθητών, από διάφορα μέρη του κόσμου, σε 3 βασικούς κλάδους εκπαίδευσης. Συγκεκριμένα, το έτος 2015, το πανεπιστήμιο της Πίζας διεξήγε έρευνα , η οποία αποσκοπούσε στο να βγάλει χρήσιμα συμπερασματα ,για το κατά πόσο το φύλο του μαθητή, η χώρα προέλευσης του ή και η περιοχή στην οποία μεγάλωσε, επηρεάζουν την μέση επίδοση του σε μαθήματα, όπως τα μαθηματικά($Mathematics$), η ανάγνωση κειμένου ($Reading$) και οι επιστήμες ($Science$).
Τα συμπεράσματα που μπορούν να προκύψουν από αυτή την έρευνα δεν είναι συγκεκριμένα και ποικίλουν ανάλογα με την οπτική και το γνωστικό επίπεδο του κάθε αναγνώστη.Μέσα ,λοιπόν, από αυτή την αναφορά πραγματοποιώ τη δικιά μου ανάλυση και καταθέτω τις δικιές μου απόψεις σχετικά με το κατά πόσο το κοινωνικό υπόβαθρο του κάθε μαθητή επηρεάζει τη μέση επίδοση του στους προαναφερόμενους τομείς εκπαίδευσης.

Η ανάλυση μου, εκπονήθηκε στη γλώσσα προγραμματισμού (R) και πραγματοποιήθηκε με βάση τη χρήση των *data tables* και *ggplots*.

2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Τα αποτελέσματα της παραπάνω έρευνας αποτυπώνονται σε 2 σύνολα δεδομένων. Το 1ο *dataset* (*'Pisa mean performance scores* 2015 *Data.csv'*) περιέχει για κάθε 1 από τους 1161 μαθητές 5 μεταβλητές. Η 1η μεταβλητή ονομάζεται *Country Name* και περιέχει το όνομα της χώρας από την οποία πρόερχεται ο μαθητής, η 2η μεταβλητή ονομάζεται *Country Code* και περιέχει το κωδικό της χώρας , η 3η μεταβλητή ονομάζεται *Series Name* και περιέχει έναν σχολιασμό σχετικά με το τι ασχολείται ο μαθήτής, καθώς και το φύλο του , σε πολλές περιπτώσεις. Η 4η μεταβλητή ονομάζεται *Series Code* και περιέχει το κωδικό του σχολιασμού που μόλις αναφέραμε. Η 5η μεταβλητή ονομάζεται ¨2015' και περίχει τη μέση επίδοση του μαθητή για έναν από τους 3 εκπαιδευτικούς κλάδους για το έτος 2015.
Το 2ο *dataset* (*'Pisa mean performance scores* 2013 - 2015 *Definition and Source.csv'*), στην ουσία αποτελεί επεξήγη του 1ου ,χωρίς να προσθέτει κάποια επιπλέον πληροφορία ,η οποία μπορεί να φανεί χρήσιμη στην πορεία για την τελική ανάλυση του αποτελέσματος.

Για να μπορέσει ωστόσο, να ξεκινήσει η διαδικάσια της ανάλυσης, πρέπει πρώτα να διαβάσουμε τα δεδομένα και να τα αποθηκέυσουμε στους κατάλληλους πίνακες. Η διαδικασία αυτή υλοποιείται μέσω του ακόλουθου συνόλου εντολών:

Listing 1: R code

```
file_path <- file.choose()
data <- fread(file_path)
file_path2 <- file.choose()
data2 <- fread(file_path2)
country_name <- data$V1
country_name <- country_name[-1]
country_code <- data$V2
country_code <- country_code[-1]
series_name <- data$V3
series_name <- series_name[-1]
series_code <- data$V4
```

```
series_code <- series_code[-1]
twenty_fifteen <- data$V5
twenty_fifteen <- twenty_fifteen[-1]
Table = data.table(
    Country_Name=country_name,
    Country_Code=country_code,
    Series_Name=series_name,
    Series_Code=series_code,
    '2015'=twenty_fifteen
)
```

Στο πίνακα που μόλις δημιουργήθηκε παρατηρείται έντονα η απώλεια τιμών στη μεταβλητή ¨2015¨. Όπως είναι εύκολα αντιληπτό, το συγκεκριμένο ζήτημα είναι πολύ κομβικό, διότι χωρίς την επίδοση των μαθήτων δεν μπορούμε να εξάγουμε ασφαλή και έγκυρα συμπεράσματα για την επίδραση των υπόλοιπων χαρακτηριστικών στο τελικό αποτέλσμα.Για το λόγο αυτό, πρέπει να αφαιρέσουμε όχι μόνο τα *null values*, τα οποία εμφανίζονται με τη μορφή ¨..¨, αλλά και τις αντίστοιχες γραμμές του πίνακα των άλλων μεταβλητών που αντιστοιχούν σε *null value* στη μεταβλητή ¨2015' . Η συγκεκριμένη διαδικασία υλοποιείται με τις ακόλουθες εντολές:

Listing 2: R code

```
remove_null<- function(x){
x[x != ".."]
}
dt <- lapply(Table, remove_null)
twenty_fifteen <- dt$'2015'
```

Μετά την εκτέλεση της συγκεκριμένης διαδικασίας,το πλήθος των τιμών που περιέχει η μεταβλητή ¨2015¨, έπεσε από το 1161 στο 612. Με εύκολα μαθηματικά , η συγκεκριμένη απόκλιση ανέρχεται σε 549 *null* τιμές.

Ωστόσο, πριν προχωρήσω στη διαδικάσια αφαίρεσης των τιμών των άλλων μεταβλητών που αντιστοιχούν σε *null value*, πρέπει πρώτα να προχωρήσω σε κάποιες άλλες διαδικασίες. Αρχικά, ελέγχω τον τύπο των μεταβλητών. Έπειτα, από τον έλεγχο αυτό, διαπιστώνω ότι η μεταβλήτη ¨2015' , δεν βρίσκεται σε *numeric form*, αλλά σε *character*. Στη συνέχεια, δημιουργώ τις μεταβλητές *gender* και *discipline*. Η μεταβλητή *gender* αναφέρεται στο φύλο του μαθητή, ενώ η μεταβλητή

discipline αναφέρεται στο κομμάτι εκπαίδευσης που έχει την αντίστοιχη επίδοση ο μαθητής. Και οι 2 μεταβλητές δημιουργούνται με βάση τις τιμές που περιέχει η μεταβλητή *Series Name*. Πιο συγκεκριμένα, κάθε τιμή της μεταβλητής *Series Name* έχει την ακόλουθη μορφή:

```
"PISA: Mean performance on the science scale. Male"
```

Σε κάθε τέτοια περίπτωση μπορούμε να συλλέξουμε χρήσιμες πληροφορίες. Οι πληροφορίες αυτές σχετίζονται με το αν ο μαθητής είναι αρσενικού ή θυληκού γένους καθώς και με το αν ο εκπαιδευτικός κλάδος που ασχολείται είναι *Mathematics*, *Reading* ή *Science*. Στο συγκεκριμένο παράδειγμα, πρόκειται για έναν μαθητή αρσενικού γένους που μελετά '*Science*¨. Στις περιπτώσεις που δεν προσδιορίζεται το φύλο του μαθητή, η τιμή που θα καταχωρείται είνα ΄ΝΑ΄.Τέλος,δημιουργώ τη μεταβλητή $Disc_Code$,με τιμές (*MAT*,*REA*,*SCI*) και συνηδειτοποιώ ότι μετά την δημιουργία των μεταβλητών *gender* και *discipline*, οι μεταβλητές *Series Name* και *Series Code* , δεν μου χρειάζονται πια, οπότε τις ¨πετάω¨.
Όλη η παραπάνω διαδικασία υλοποιείται με τις εξής εντολές:

Listing 3: R code

```
class(country_name)
class(country_code)
class(series_name)
class(series_code)
class(twenty_fifteen)
twenty_fifteen <- as.numeric(twenty_fifteen)
class(twenty_fifteen)
na.omit(twenty_fifteen)
cn<- list()
cc<- list()
sn<-list()
sc<-list()
performance<- list()

#Creating lists that doesn't correspond to NA value
N<- length(twenty_fifteen)
counter<-1
for (i in 1:N){
if (!is.na(twenty_fifteen[i])){
cn[counter]<-country_name[i]
```

```
cc[counter]<-country_code[i]
sn[counter]<- series_name[i]
sc[counter]<-series_code[i]
performance[counter]<-twenty_fifteen[i]
counter<- counter + 1
}
}
country_name <- unlist(cn)
country_code <- unlist(cc)
series_name <- unlist(sn)
series_code <- unlist(sc)
performance <- unlist(performance)

library(stringr)
N=length(series_name)
print(N)
gender=list()
discipline=list()
for (i in 1:N){
words <- strsplit(series_name[i], "-")[[1]]
last_word <- tolower(words[length(words)])
discipline[i]<- str_extract(series_name[i], "(?<=the\\s)(\\w+)(?=\\s
if (last_word == "female") {
    gender[i]='F'
  } else if (last_word == "male") {
    gender[i]='M'
  } else {
    gender[i]='NA'
  }
}
gender <- unlist(gender)
discipline<- unlist(discipline)
series_code <- sub(".*PISA\\.(.*)\\..*", "\\1", series_code)
series_code <- sub(".*PISA\\.(.*)(\\..*)?", "\\1",
series_code)
Final_Table = data.table(
Country_Name=country_name,
Country_Code=country_code,
Disc_Code=series_code,
```

```
Gender=gender ,
Discipline=discipline ,
Performance=performance
)
```

Και ενώ σιγά σιγά πλησιάζουμε στο σημείο που έχουμε συμπληρώσει όλες τις απαιραίτητες προϋποθέσεις , προκείμενου να ξεκινήσουμε τη τελική φάση της ανάλυσης μας, πραγματοποιούμε έναν επιπλέον έλεγχο στα δεδομένα που έχουμε δημιουργήσει, χρησιμοποιώντας τη τεχνική των *aggregations*.

Listing 4: R code

```
Final_Table[, .N, by = Country_Name]
Final_Table[, .N, by = Discipline]
Final_Table[, .N, by = Gender]
```

Μέσα από τη χρήση των παραπάνω εντολών παίρνουμε ένα αποτέλεσμα που δείχνει ομοιομορφία στη κατανομή των δεδομένων . Συγκεκριμένα, από τις 612 πλείαδες, η κατανομή των τιμών στην εκάστοστε μεταβλητή είναι ισόποση. Συγκεκριμένα, τα 612 δεδομένα της μεταβλητής *Disc Code* αποτελούνται 204 τιμές MAT, REA και SCI, τα 612 δεδομένα της μεταβλητής *Country* απότελούνται από 68 χώρες, όπου η κάθε μία έχει πλήθος 9 και τα 612 δεδομένα της μεταβλήτης *Gender* αποτελούνται από 204 τιμές F, M και NA.
Η τιμή NA της μεταβλητής *Gender*, αποτελεί πρόβλημα στη διαδικασία της ανάλυσης μας.Δεν μας επιτρέπει να είμαστε 100% σίγουροι για το κατά πόσο το φύλο του μαθητή σε συνδυασμό με τους άλλους παράγοντες επηρεάζουν την επίδοση του και για το λόγο αυτό, δημιουργούμε ένα τελικό *datatable*, που θα περιέχει 408 γραμμές. Μέσα από αυτόν, το τελικό πίνακα θα μπορέσουμε να αναλύσουμε εξουνουχιστικά την επιρροή της κάθε μεταβλήτης στο τελικό αποτέλεσμα.
Ο κώδικας της τελικής μετατροπής είναι ο εξής:

Listing 5: R code

```
COUNTRY_NAME=list()
COUNTRY_CODE=list()
DISCIPLINE=list()
DISC_CODE=list()
GENDER=list()
```

```
PERFORMANCE=list()
counter<-0
N<- length(performance)
for (i in 1:N){
if (gender[i]!='NA'){
counter=counter+1
GENDER[counter]=gender[i]
COUNTRY_NAME[counter]=country_name[i]
COUNTRY_CODE[counter]=country_code[i]
DISCIPLINE[counter]=DISCIPLINE[i]
DISC_CODE[counter]=series_code[i]
PERFORMANCE[counter]=performance[i]
}
}
COUNTRY_NAME<-unlist(COUNTRY_NAME)
COUNTRY_CODE<- unlist(COUNTRY_CODE)
DISC_CODE<- unlist(SERIES_CODE)
GENDER<- unlist(GENDER)
PERFORMANCE<- unlist(PERFORMANCE)
DISCIPLINE<- unlist(REGION)
Data_Table = data.table(
Country_Name=COUNTRY_NAME,
Country_Code=COUNTRY_CODE,
Disc_Code=DISC_CODE,
Gender=GENDER,
Discipline=DISCIPLINE,
Performance=PERFORMANCE
)
Data_Table[, .N, by = Country_Name]
Data_Table[, .N, by = .(REGION)]
Data_Table[, .N, by = .(DISC_CODE)]
Data_Table[, .N, by = .(GENDER)]
```

3 ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Στην παρακάτω εικόνα απεικονίζεται ο πίνακας, από τον οποίο θα βγάλουμε τα τελικά μας συμπεράσματα.

```
> Data_Table[, .N, by = .(Disc_Code)]
   Disc_Code   N
1:       MAT 136
2:       REA 136
3:       SCI 136
> Data_Table[, .N, by = .(GENDER)]
   GENDER   N
1:      F 204
2:      M 204
> print(Data_Table)
     Country_Name Country_Code Disc_Code Gender  Discipline Performance
  1:      Albania          ALB       MAT      F mathematics    417.7500
  2:      Albania          ALB       MAT      M mathematics    408.5455
  3:      Albania          ALB       REA      F     reading    434.6396
  4:      Albania          ALB       REA      M     reading    375.7592
  5:      Albania          ALB       SCI      F     science    439.4430
 ---
404:    Lithuania          LTU       MAT      M mathematics    492.9591
405:    Lithuania          LTU       REA      F     reading    499.0386
406:    Lithuania          LTU       REA      M     reading    473.9191
407:    Lithuania          LTU       SCI      F     science    525.9139
408:    Lithuania          LTU       SCI      M     science    523.3141
```

Κάθε μεταβλητή περιέχει 408 τιμές και η ομοιόμορφη κατανομή των τιμών εξακολουθεί να υφίσταται. Όπως παρατηρούμε και στην εικόνα, έχουμε τη μεταβλητή $Disc\ Code$ που αποτελείται από 3 τιμές με πλήθος 136 η κάθε μία, έχουμε τη μεταβλητή $Gender$ που αποτελείται από 2 τιμές με πλήθος 204 η κάθε μία και έχουμε τη μεταβλητή $Country\ Name$ που αποτελείται από 68 τιμές με πλήθος 6 η κάθε μία.

Για να μπορέσουμε να βγάλουμε ένα έγκυρο συμπέρασμα για την απόδοση των μαθήτων θα χρειαστεί να υπολογίσουμε πρώτα των κανόνα των 5 τιμών για τη μεταβλητή $(Performance)$. Συμφώνα με τον κανόνα αυτό, αφού ταξινομήσουμε τις τιμές της μεταβλητής, βρίσκουμε γι΄ αυτήν 5 τιμές. Οι τιμές αυτές είναι: $min(Performance)$, $max(Performance)$, $median(Performance)$,$Q1(Performance)$ και $Q3(Performance)$. Η $Q1$ αναφέρεται στην ενδίαμεση τιμή των min και $median$ και η $Q3$ αναφέρεται στην ενδιάμεση τιμή των $median$ και max.Ωστόσο, επειδή θεωρώ τη μέση τιμή $(mean(Performance)$) καλύτερο μέτρο σύγκρισης από την ενδίαμεση τιμή του συνόλου δεδομένων , από τις τιμές του κανόνα η $median$ τιμή , δεν θα συμπεριληφθει στην ανάλυση μου.
Η μέση επίδοση των 408 μαθητών είναι 462.117, η max είναι 564.2545 και η min 325.5866. Οι τιμές $Q1$ και $Q3$ είναι 419.7179 και 501.864 αντίστοιχα.
Από το σύνολο των 408 μαθητών, έκεινοι που είχαν μέση επίδοση πάνω από το μέσο όρο είναι 228, εκ των οποίων οι 110 ήταν αγόρια και οι 118 κορίτσια. Τα παιδιά που είχαν επίδοση μεγαλύτερη από το

μέσο όρο στα μαθηματικά είναι 77 , εκ των οποίων τα 39 είναι αγόρια και τα 38 κορίτσια. Τα παιδιά που είχαν επίδοση μεγαλύτερη από το μέσο όρο στο *reading* είναι 76 , εκ των οποίων τα 34 είναι αγόρια και τα 42 κορίτσια. Τα παιδιά που είχαν επίδοση μεγαλύτερη από το μέσο όρο στο *science* είναι 75, εκ των οποίων τα 37 είναι αγόρια και τα 38 κορίτσια.

Σε πρώτη φάση , λοιπόν, συμπεραίνουμε ότι το φύλο δεν επηρεάζει ιδιαίτερα την απόδοση των μαθητών σε $Mathematics$ και $Science$. Ωστόσο, παρατηρείται ότι στο $Reading$, τα κορίτσια έχουν μία μεγαλύτερη έφεση. Για πιο ασφαλή συμπεράσματα ελέγχουμε το ποια παιδιά πετύχανε επίδοση που ανήκει στο υψηλότερο 25% αλλά και στο χαμηλότερο 25% των τιμών.

Από το σύνολο των 408 μαθητών, έκεινοι που είχαν μέση επίδοση μεγαλύτερη από $Q3$ είναι 102, εκ των οποίων οι 46 είναι αγόρια και οι 56 κορίτσια. Από αυτά τα 102 παιδιά, εκείνα που είχαν επίδοση μεγαλύτερη από $Q3$ στα μαθηματικά είναι 30 , εκ των οποίων τα 17 είναι αγόρια και τα 13 κορίτσια.Ακόμη, τα παιδιά που είχαν επίδοση μεγαλύτερη από $Q3$ στο *reading* είναι 34 , εκ των οποίων τα 7 είναι αγόρια και τα 27 κορίτσια.Τέλος, τα παιδιά που είχαν επίδοση μεγαλύτερη από $Q3$ στο *science* είναι 38, εκ των οποίων τα 22 είναι αγόρια και τα 16 κορίτσια.

Από την άλλη πλευρά, από το σύνολο των 408 μαθητών, έκεινοι που είχαν μέση επίδοση μικρότερη από $Q1$ είναι 102, εκ των οποίων οι 60 είναι αγόρια και οι 42 κορίτσια.Από τα παιδιά αυτά, εκείνα που είχαν επίδοση μικρότερη από $Q1$ στα μαθηματικά είναι 41 , εκ των οποίων τα 20 είναι αγόρια και τα 21 κορίτσια. Τα παιδιά που είχαν επίδοση μικρότερη από $Q1$ στο *reading* είναι 33 , εκ των οποίων τα 25 είναι αγόρια και τα 8 κορίτσια, ενώ κλείνοντας τα παιδιά που είχαν επίδοση μικρότερη από $Q1$ στο *science* είναι 28, εκ των οποίων τα 15 είναι αγόρια και τα 13 κορίτσια.

Μέσα από τα παραπάνω στοίχεια μπορούμε να βγάλουμε μερικά χρήσιμα συμπεράσματα.Πρώτα από όλα είναι οφθαλμοφανές ότι τα κορίτσια έχουν πολύ μεγαλύτερη έφεση στο κομμάτι του $Reading$.Αυτό δικαιολογείται όχι μόνο από το γεγονός ότι είναι η πλειοψηφία των καλύτερων επιδόσεων άλλα και από το γεγονός ότι είναι η μειοψηφία των χειρότερων επιδόσεων στο συγκεκριμένο κλάδο.Ακόμη στους τομείς $Mathematics$ και $Science$ παρατηρείται μία μικρή υπεροχή των αγοριών ως προς τις καλύτερες επιδόσεις και μια ισορροπία των δύο φύλων στους τομείς αυτούς , ως προς τις χειρότερες επιδόσεις.

Στο συγκεκριμένο σημείο, και αφού ελέγξαμε μεμονομένα την επιρροή που έχει το φύλο στην απόδοση του μαθητή , θα ελέγξουμε για τον κάθε εκπαιδευτικό τομέα ξεχωριστά , το πόσο κομβικό ρόλο παίζει το μέρος από το οποίο προέρχεται ο μαθητής.
Στο τομέα των μαθηματικών, όπως ανεφέραμε και πιο πάνω υπάρχουν 30 παιδιά που ανήκουν στο 25% των μαθητών με υψηλή επίδοση και 41 παιδιά που ανήκουν στο 25% των μαθητών με χαμηλή επίδοση.
Τα περισσότερα από τα παιδιά με υψηλή επίδοση προέρχονται από την Ευρώπη (16 στο σύνολο), χωρις όμως να υπάρχει κάποια χώρα, η οποία να ξεχωρίζει σε πλήθος. Ωστόσο, υπάρχουν και 15 παιδιά από άλλα μέρη του κόσμου. Συγκεκριμένα, υπάρχουν 6 παιδιά από την Αφρική, τα οποία προέρχονται από χώρες όπως το Μπουρουντί, οι Κομόρες και το Καμερούν, 6 παιδιά από την Βόρεια και λατινική Αμερική , 1 από την Αυστραλία και 2 από την Ασία.
Από την άλλη πλευρά, τα πράγματα είναι ισορροπημένα. Υπάρχουν 13 μαθητές από την Αφρική, 11 μαθητές από την Ευρώπη, 10 από χώρες της λατινικής και Βόρειας Αμερικής και 8 από την Ασία.Και στις 2 περιπτώσεις δεν υπάρχει ήπειρος που να έχει χώρα με πλήθος μαθητών πάνω από 2.

Στο κομμάτι του *Reading*, τα περισσότερα από τα παιδιά με υψηλή επίδοση προέρχονται πάλι από την Ευρώπη. Συγκεκριμένα η Ευρώπη κατέχει 13 μαθητές, ενώ υπάρχουν ακόμη 2 μαθητές από την Ασιά, 6 από χώρες της Αμερικής και 7 μαθητές από την Αφρική. Αξίζει να σημειωθεί ότι και σε αυτή την κατηγορία οι χώρες Μπουρουντί, Κομόρες και Καμερούν διαθέτουν μαθητές που μπόρεσαν να ξεχωρίσουν. Από την άλλη πλευρά,οι μαθητές με τις χειρότερες επιδόσεις στο *Reading* είναι κυρίως από την Αφρική και την Ασία. Συγκεκριμένα,οι μαθητές από την Αφρική είναι 12 και από την Ασία 10. Ακόμη, υπάρχουν 4 μαθητές από χώρες της Αμερικής και 7 από την Ευρώπη.Και στις 2 περιπτώσεις δεν υπάρχει ήπειρος που να έχει χώρα με πλήθος μαθητών πάνω από 2.

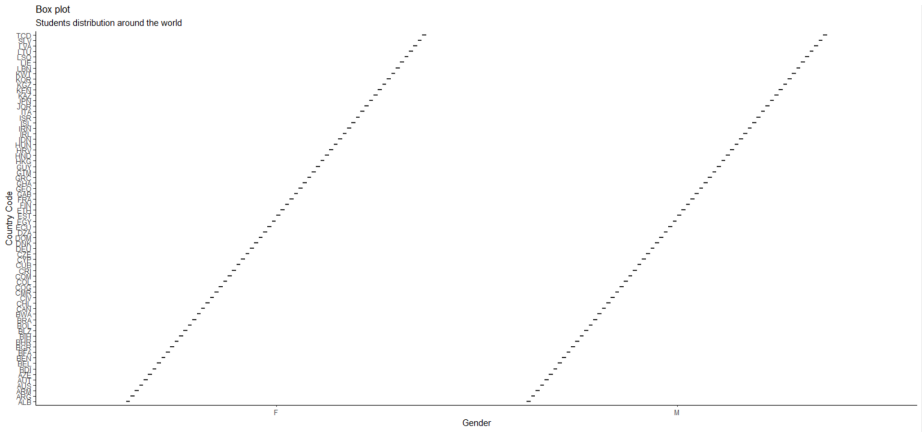
Στο τομέα *Science*, τα περισσότερα από τα παιδιά που κατάφεραν να πετύχουν υψηλή απόδοση προέρχονται από την Ευρώπη (17 στο σύνολο). Επίσης, υπάρχουν 6 μαθητές από την Αφρική, με τις χώρες Καμερούν, Μπουρουντί και Κομόρες να δείχνουν συνέπεια και σε αυτόν τον επιστημονικό τομέα. Ακόμα, υπάρχουν 8 μαθητές από χώρες της Αμερικής,1 από την Αυστραλία και 6 από την Ασία.

Όσον αφορά την άλλη μεριά, υπάρχουν 10 μαθητές από την Αφρική, 7 μαθητές από την Ευρώπη, 3 από την Αμερική και 8 από την Ασία.Και στις 2 περιπτώσεις δεν υπάρχει ήπειρος που να έχει χώρα με πλήθος μαθητών πάνω από 2.

Μέσα από τα παραπάνω δεδομένα παρατηρούμε ότι η Ευρώπη διαθέτει μαθητές που ξεχωρίζουν τόσο με την υψηλή τους απόδοση , όσο και με τη χαμηλή τους και στις 3 επιστημονικές κατηγορίες.Οι Ευρωπαϊκές χώρες με σταθερή συνέπεια και υψηλές επιδόσεις και στους 3 κλάδους είναι η Ιρλανδία,η Ιταλία ,η Κροατία, η Αυστρία, το Βέλγιο, η Βουλγαρία και η Ουγγαρία, ένω χώρες με σταθερή συνέπεια στις χαμηλές επιδόσεις όλων των κλάδων είναι η Ελλάδα, η Αλβανία, η Γερμανία, η Δανία και η Γαλλία. Επίσης, παρ΄όλο που η Αφρική έχει το μεγαλύτερο πλήθος μαθητών με χαμηλή επίδοση και στις 3 κατηγορίες, με μαθητές από χώρες όπως το Τσαντ, η Μπουρκίνα Φάσο, η Γκάνα,η Αλγερία και η Μποτσουάνα, αξίζει να επισημανθεί για άλλη μια φορά ότι έχει 3 χώρες(Μπουρουντί, Καμερούν και Κομόρες) με σταθερή συνέπεια στις υψηλές επιδόσεις σε όλους τους εκπαιδευτικούς κλάδους. Από την Ασία, ξεχωρίσαν με τις υψηλές του επιδόσεις, επί των πλήστων μαθητές από το Καζακστάν αλλά και από την Αρμενία. Ώστόσο, ξεχώρισαν ως προς τις χαμηλές επιδόσεις και πολλοί μαθητές από αρκετές διαφορετικές χώρες της Ηπείρου. Συγκεκριμένα, μαθητές από το Χονγκ Κονγκ, τη Κορέα , την Ινδονησία, το Αζερμπαϊτζαν και το Κουβέιτ εμφάνισαν με συνέπεια και στους 3 κλάδους χαμηλές επιδόσεις.Τέλος, από τις χώρες της Βόρειας και Λατινικής Αμερικής, οι χώρες με μαθητές που ξεχώρισαν για την υψηλή τους επίδοση και στους 3 κλάδους είναι η Γουατεμάλα, η Χιλή και το Εκουαδόρ. Από την άλλη μεριά, χαμηλή επιδόση και στους 3 εκπαιδευτικούς κλάδους έκαναν μαθητές από τη Βολιβία, την Αργεντινή, το Ελ Σαλβαδορ και το Μπενίν.
Ο συνδυασμός φύλου και χώρας προέλευσης του μαθήτη , δεν προσδίδει κάποια περαιτέρω πληροφορία στην ανάλυση μας. Το μόνο που ίσως , χρείαζεται να τονίσουμε είναι ότι και σε αυτή τη περιπτώση το πλήθος των αγοριών και των κοριτσιών είναι ισορροπημένο για κάθε χώρα και συνεπώς δεν υπάρχει κάποιο μέρος που να έχει περισσότερους επιτυχόντες αρσενικού γένους από ότι θυληκού γένους.Το παρακάτω διάγραμμα επιβεβαίωνει τη συγκεκριμένη θέση.

Listing 6: R code

```
theme_set(theme_classic())
g <- ggplot(Data_Table, aes(Gender, Country_Code))
g + geom_boxplot(varwidth=T, fill="plum") +
labs(title="Box plot",
subtitle="Students' gender distribution around the world",
caption="Source: mpg",
x="Gender",
y="Country Code")
```



Σχήμα 1: Το συγκεκριμένο *boxplot* αποτυπώνει τη ισορροπία που επικρατεί στη κατανομή των μαθητών ως προς τη χώρα και το φύλο τους.Ο άξονας y περιέχει τις χώρες, ενώ ο άξονας x περιέχει τις τιμές 'F' και 'M'

Οι παραπάνω έλγχοι κωδικοποιήθηκαν στην R, ως εξής:

Listing 7: R code

```
max(PERFORMANCE)
min(PERFORMANCE)
Q1 <- quantile(PERFORMANCE, 0.25)
mean(PERFORMANCE)
median(PERFORMANCE)
Q3 <- quantile(PERFORMANCE, 0.75)
mean(PERFORMANCE)

Data_Table[, .N, by = .(PERFORMANCE>mean(PERFORMANCE))]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & GENDER == "M", .N]
```

```
Data_Table[PERFORMANCE > mean(PERFORMANCE) & GENDER == "F",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "MAT",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "SCI",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "REA",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "MAT"
& GENDER == "M",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "MAT"
& GENDER == "F",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "SCI"
& GENDER == "M",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "SCI"
& GENDER == "F",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "REA"
& GENDER == "M",.N]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "REA"
& GENDER == "F",.N]


Data_Table[, .N, by = .(PERFORMANCE>Q3)]
Data_Table[(PERFORMANCE>Q3)]
Data_Table[PERFORMANCE > Q3 & GENDER == "M",.N]
Data_Table[PERFORMANCE > Q3 & GENDER == "F",.N]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "MAT",.N]
Data_Table[PERFORMANCE > Q3 & SDISC_CODE == "SCI",.N]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "REA",.N]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "MAT"
& GENDER == "M",.N]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "MAT"
& GENDER == "F",.N]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "SCI"
& GENDER == "M",.N]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "SCI"
& GENDER == "F",.N]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "REA"
& GENDER == "M",.N]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "REA" & GENDER == "F",.N]

Data_Table[, .N, by = .(PERFORMANCE<Q1)]
Data_Table[PERFORMANCE <Q1 & GENDER == "M",.N]
```

```
Data_Table[PERFORMANCE <Q1 & GENDER == "F" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "MAT" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "SCI" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "REA" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "MAT"& GENDER == "M" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "MAT" & GENDER == "F" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "SCI"& GENDER == "M" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "SCI"& GENDER == "F" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "REA"& GENDER == "M" ,.N]
Data_Table[PERFORMANCE <Q1 & DISC_CODE == "REA"& GENDER == "F" ,.N]

Data_Table[PERFORMANCE > Q3 & DISC_CODE == "MAT" ,  .N,
.(Country_Name)]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "REA" ,  .N,
.(Country_Name)]
Data_Table[PERFORMANCE > Q3 & DISC_CODE == "SCI" ,  .N,
.(Country_Name)]

Data_Table[PERFORMANCE < Q1 & DISC_CODE == "MAT" ,  .N,
.(Country_Name)]
Data_Table[PERFORMANCE < Q1 & DISC_CODE == "REA" ,  .N,
.(Country_Name)]
Data_Table[PERFORMANCE < Q1 & DISC_CODE == "SCI" ,  .N,
.(Country_Name)]

Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "MAT" ,  .N,
.(Country_Name)]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "REA" ,  .N,
.(Country_Name)]
Data_Table[PERFORMANCE > mean(PERFORMANCE) & DISC_CODE == "SCI" ,  .N,
.(Country_Name)]


Data_Table[PERFORMANCE < Q1 & DISC_CODE == "SCI"
& GENDER=="M" ,  .N,.(Country_Name)]
Data_Table[PERFORMANCE < Q1 & DISC_CODE == "SCI"
& GENDER=="F" ,  .N,.(Country_Name)]

Data_Table[PERFORMANCE < Q1 & DISC_CODE == "REA"
```

```
& GENDER=="M", .N, .(Country_Name)]
Data_Table[PERFORMANCE < Q1 & DISC_CODE == "REA"
& GENDER=="F", .N, .(Country_Name)]

Data_Table[PERFORMANCE < Q1 & DISC_CODE == "MAT"
& GENDER=="M", .N, .(Country_Name)]
Data_Table[PERFORMANCE < Q1 & DISC_CODE == "MAT"
& GENDER=="F", .N, .(Country_Name)]

Data_Table[, .N, .(Performance>Q3, Disc_Code=="MAT")]
Data_Table[, .N, .(Performance>Q3, Series_Code=="SCI")]
Data_Table[, .N, .(Performance>Q3, Series_Code=="REA")]
```

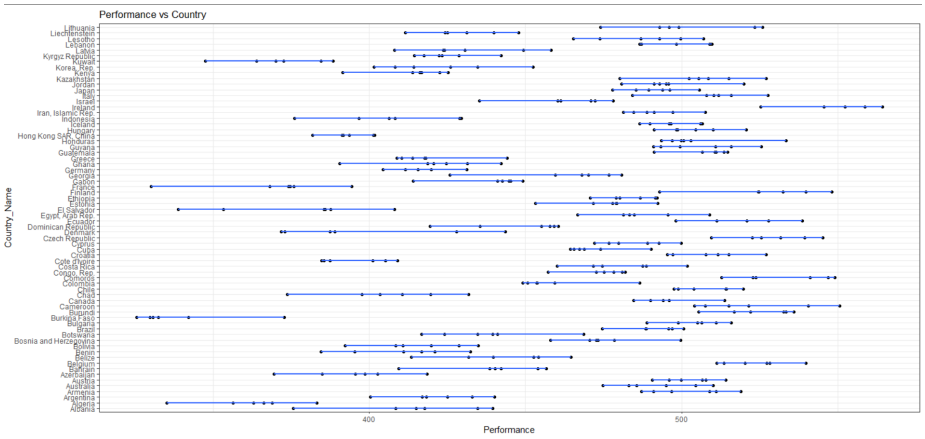
4 ΣΥΜΠΕΡΑΣΜΑ

Ως ένα γενικό συμπέρασμα της ανάλυσης μας μπορύμε να πούμε ότι τελικά η χώρα προέλευσης και το φύλο του μαθητή, επηρεάζουν σε σημαντικό βαθμό την απόδοση του, σε κάθε τομέα.Ωστόσο, αξίζει να σημειωθεί ότι δεν προκύπτει από κάπου ότι ο συνδυασμός των 2 αυτών παραγόντων επηρεάζει την επίδοση κάποιου παιδιού στα μαθηματικά, το *reading*, ή το *science*.Η επίδοση επηρεάζεται μεμονωμένα.
Τα κορίτσια έχουν έφεση στο *reading*, χωρίς να εξαρτάται από ποια χώρα είναι.Ακόμη για παράδειγμα, τα παιδιά από την Ιρλανδία, έχουν επιτυχίες σε όλους του κλάδους ανεξαρτήτως φύλου, διότι και τα αγόρια και τα κορίτσια από αυτή τη χώρα σημειώνουν υψηλή απόδοση.Φυσικά, το ίδιο ισχύει και για παιδιά από χώρες που βάση την ανάλυση μας, σημείωσαν χαμηλή επίδοση.Τέλος, είναι απαραίτητο να επισημάνουμε ότι η απόδοση ενός μαθητή δεν εξαρτάται από το αντικείμενο από το οποίο μελετάει. Και στους 3 κλάδους, σημειώθηκαν πολύ υψηλές επιδόσεις , ενώ όσο αφορά τις χαμηλές επιδόσεις,τα παιδιά που τις πραγματοποίησαν ήταν από συγκεκριμένες χώρες , γεγονός που αποδεικνύει προβλημα στο δικό τους εκπαιδευτικό σύστημα.
Παρακάτω, ακολουθούν μερικά διαγράμματα, τα οποία επιβεβαιώνουν τα όσα ισχυρίστηκα, μαζί με τον κώδικα υλοποίησης τους.

5 ΔΙΑΓΡΑΜΜΑΤΑ

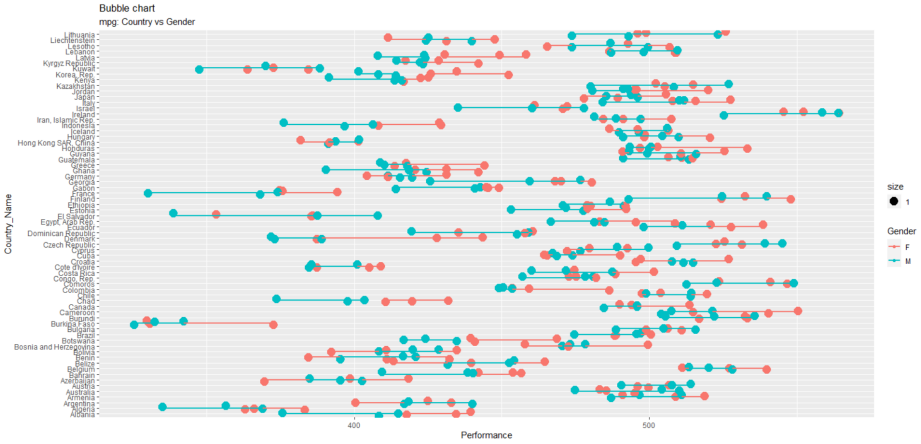
Listing 8: R code

```
#Code for diagram 2
g <- ggplot(Data_Table, aes(x=Performance, y=Country_Name))+
```

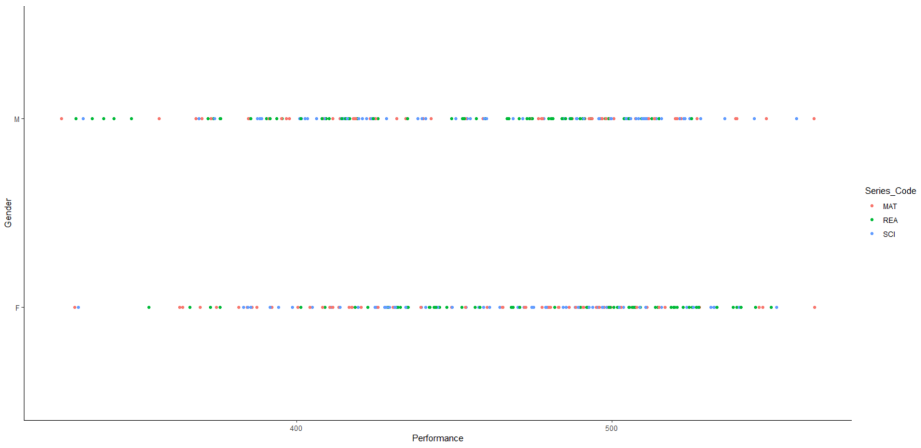


Σχήμα 2: Το συγκεκριμένο διάγραμμα απεικονίζει την απόδοση των μαθητών από διάφορα μέρη του κόσμου.Επιβεβαιώνει τον πλουραλισμό των μαθητών σε υψηλές και χαμηλές κλίμακες επιδόσεων από όλες τις χώρες. Κάθε κουκίδα είναι ένας μαθητής, που αντιστοιχεί σε μία τιμή(χώρα) του άξονα y και σε μία τιμή(επίδοση) του άξονα x.

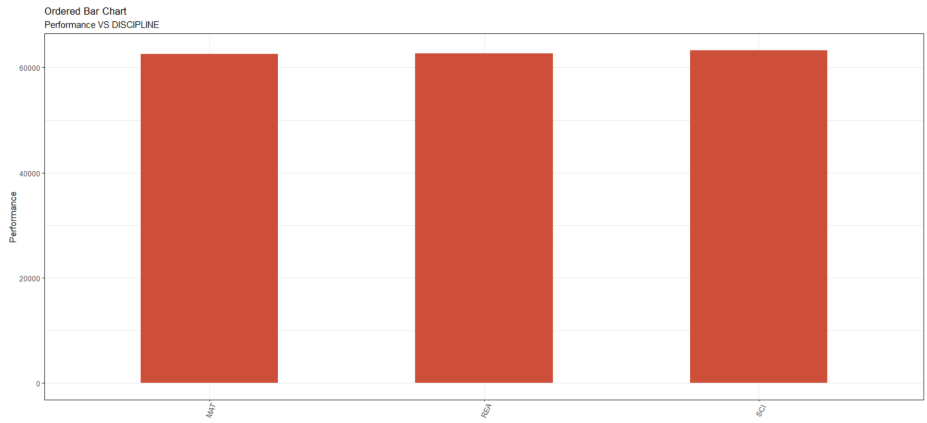
```
geom_point() +labs(title="Performance vs Country",
caption = "Source: mpg") +
geom_smooth(method="lm", se=FALSE) +theme_bw()

#Code for diagram 3
g <- ggplot(Data_Table, aes(Performance, Country_Name)) +
labs(subtitle="mpg: Country vs Gender",
title="Bubble chart")
g + geom_jitter(aes(col=Gender, size=1)) +
geom_smooth(aes(col=Gender), method="lm", se=F)

#Code for diagram 4
ggplot(Data_Table, aes(x = Performance, y = Country_Code, color = Se
geom_point()
#Code for diagram 5
theme_set(theme_bw())
ggplot(Data_Table, aes(x=Disc_Code, y=Performance)) +
geom_bar(stat="identity", width=.5, fill="tomato3") +
labs(title="Ordered Bar Chart",
subtitle="Performance VS DISCIPLINE",caption="source: mpg") +theme(a
```



Σχήμα 3: Σε συνέχεια του προηγούμενο διαγράμματος (Σχήμα2), το συγκεκριμένο διάγραμμα προσθέτει επιπλέον πληροφορία καθώς αποτυπώνει και το φύλο του μαθητή ανάλογα με το χρώμα της κουκίδας. Συγκεκριμένα, οι μπλε κουκίδες αντιπροσωπεύουν τα αγόρια ενώ οι κόκκινες κουκίδες αντιπροσωπεύουν τα κορίτσια.Είναι ένα εξειδικευμένο διάγραμμα,που δείχνει την μέση απόδοση για το κάθε μαθητή του κάθε φύλου και της κάθε χώρας.



Σχήμα 4: Το συγκεκριμένο διάγραμμα απεικονίζει την απόδοση των μαθητών με βάση το φύλο τους και σε συνδυασμό με τον εκπαιδευτικό τομέα με τον οποίο ασχολούνται. Κάθε κουκίδα είναι ένας μαθητής, που αντιστοιχεί σε μία τιμή(φύλο) του άξονα y και σε μία τιμή(επίδοση) του άξονα x.Το χρώμα της κουκίδας είναι ανάλογο του μαθήματος που ασχολείται ο μαθητής.Συγκεκριμένα, η κόκκινη κουκίδα αντιπροσωπεύει τα Μαθηματικά, η πράσινη το *Reading* και η μπλε το *Science*.Σκοπός του διαγράμματος είναι αναδείξει με τον πιο απλοϊκο τρόπο , το τι απόδοση είχε το κάθε φύλο για εκπαιδευτικό κλάδο.



Σχήμα 5: Το συγκεκριμένο διάγραμμα έχει στο άξονα των y μία κλίμακα τιμών της μεταβλητής $Performance$ και στον άξονα των x τις κωδικοποιημένες τιμές της μεταβλητής $Discipline$. Η χρήση του αποδεικνύει ότι οι 3 κλάδοι περιλμβάνουν μαθητές που μπορούν να πετύχουν και υψηλές άλλα και χαμηλές αποδόσεις. Συνεπώς, είναι εύκολα κατανοητό, πως η επίτυχια ενός μαθητή δεν εξαρτάται από το αντικείμενο που μελετάει.