Εθνικό Μετσόβιο Πολυτεχνείο

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Πρόγραμμα Μεταπτυχιακών Σπουδών Επιστήμης Δεδομένων και Μηχανικής Μάθησης

Διερευνητική Ανάλυση Δεδομένων μέσω R

Εργασία Μαθήματος:
«Προγραμματιστικά Εργαλεία και Τεχνολογίες για την Επιστήμη Δεδομένων»

Βόγκας Ιωάννης, Α.Μ.: 03400206

Μεταπτυχιακός φοιτητής προγράμματος Επιστήμης Δεδομένων και Μηχανικής Μάθησης

e-mail: ioannisvogkas@mail.ntua.gr

Ιανουάριος 2024

Περιεχόμενα

Σχήματα

Πίνακες

1 Εισαγωγή

Σκοπός της παρούσας εργασίας είναι η διερευνητική ανάλυση δεδομένων που αφορούν στις επιδόσεις μαθητών σε τρία διαφορετικά επιστημονικά πεδία. Πιο συγκεκριμένα, αξιοποιήθηκε ένα σύνολο δεδομένων του Προγράμματος Διεθνούς Αξιολόγησης Μαθητών PISA που περιλάμβανε μετρήσεις για τις επιδόσεις μαθητών ηλικίας 15 ετών, το έτος 2015, στους τομείς των μαθηματικών, της ανάγνωσης και των επιστημών. Εξετάστηκαν χαρακτηριστικά των μαθητών που επηρεάζουν την απόδοση τους, όπως το φύλο, η χώρα και η γεωγραφική ήπειρος στις οποίες ανήκουν, ενώ διερευνήθηκαν και πιθανές συσχετίσεις μεταξύ των επιδόσεων στους διαφορετικούς τομείς. Τέλος, επεξηγήθηκαν και ερμηνεύτηκαν τα συμπεράσματα που προέκυψαν σε κάθε μέρος της ανάλυσης. Η υλοποίηση της ανάλυσης έγινε μέσω της γλώσσας προγραμματισμού R, χρησιμοποιώντας τις βιβλιοθήκες data.table και ggplot2 για την επεξεργασία και την οπτικοποίηση των δεδομένων αντίστοιχα.

2 Προεπεξεργασία Δεδομένων

Το αρχικό σύνολο δεδομένων αποτελούταν από 1161 γραμμές και 5 στήλες. Θεωρήθηκε σημαντικό να γίνουν κάποιες αλλαγές στα δεδομένα ώστε να υλοποιηθεί η ανάλυση στο επόμενο βήμα.

Αρχικά, χρησιμοποιώντας την βιβλιοθήκη data.table, έγινε η εισαγωγή των δεδομένων στο RStudio και πραγματοποιήθηκε έλεγχος για τον τύπο της κάθε μεταβλητής. Όλες οι μεταβλητές ήταν τύπου χαρακτήρα και η μόνη που χρειάστηκε να μετατραπεί ήταν η «2015», που έδειχνε την απόδοση και έπρεπε να γίνει αριθμητική.

Στη συνέχεια, διαγράφηκαν μεταβλητές, δημιουργήθηκαν νέες και μετονομάστηκαν κάποιες στήλες. Η μεταβλητή Series Code παρείχε πληροφορία για το μάθημα (Math, Reading, Science) και το φύλο και θα ήταν χρήσιμο για το στάδιο της ανάλυσης να χωριστεί σε δύο διαφορετικές στήλες.Έτσι, δημιουργήθηκαν οι καινούργιες μεταβλητές Discipline, που δέχεται τις τιμές [Math, Reading, Science], και Gender, που δέχεται τις τιμές [All, FE, MA]. Ακόμα, αφαιρέθηκαν οι στήλες Series Name και Series Code, αφού δεν προσέφεραν κάποια πληροφορία για να αξιοποιηθεί στην ανάλυση. Προστέθηκε μία ακόμα μεταβλητή, η Region, που δείχνει σε ποια γεωγραφική ήπειρο ανήκει η κάθε χώρα και δέχεται τις τιμές [EU, AM, AF, AS, OC], που αντιστοιχούν στην Ευρώπη, στην Αμερική, στην Αφρική, στην Ασία και στην Ωκεανία. Τέλος, μετονομάστηκε η μεταβλητή «2015», που έδειχνε την απόδοση των μαθητών, σε Performance.

Σημαντικό ήταν να γίνει έλεγχος για την ύπαρξη κενών τιμών. Η μεταβλητή Performance ήταν η μόνη στην οποία υπήρχαν κενές τιμές. Σε κάθε χώρα που υπήρχαν NULL τιμές, αυτές βρισκόντουσαν σε όλα τα πεδία της, ανεξαρτήτως μαθήματος και φύλου. Επομένως η αντικατάσταση τους με κάποια τιμή, όπως η διάμεσος, δεν προσέφερε καμία

ουσιαστική πληροφορία για την συγκεκριμένη χώρα. Άρα, διαγράφηκαν οι γραμμές στις οποίες η μεταβλητή Performance ήταν κενή.

Το τελικό σύνολο δεδομένων το οποίο αναλύθηκε αποτελούταν από 612 γραμμές και τις εξής 6 στήλες:

- Country Name (character): Ονομασία χώρας
- Country Code(character): Κωδική ονομασία χώρας
- Performance (numeric): Μετρική απόδοσης για το 2015
- Discipline (character): Επιστημονικό πεδίο (Math, Reading, Science)
- Gender(character): Φύλο (All, FE, MA)
- Region (character): Γεωγραφική ήπειρος (EU, AM, AF, AS, OC)

Η προγραμματιστική υλοποίηση της προεπεξερασίας των δεδομένων είναι η εξής:

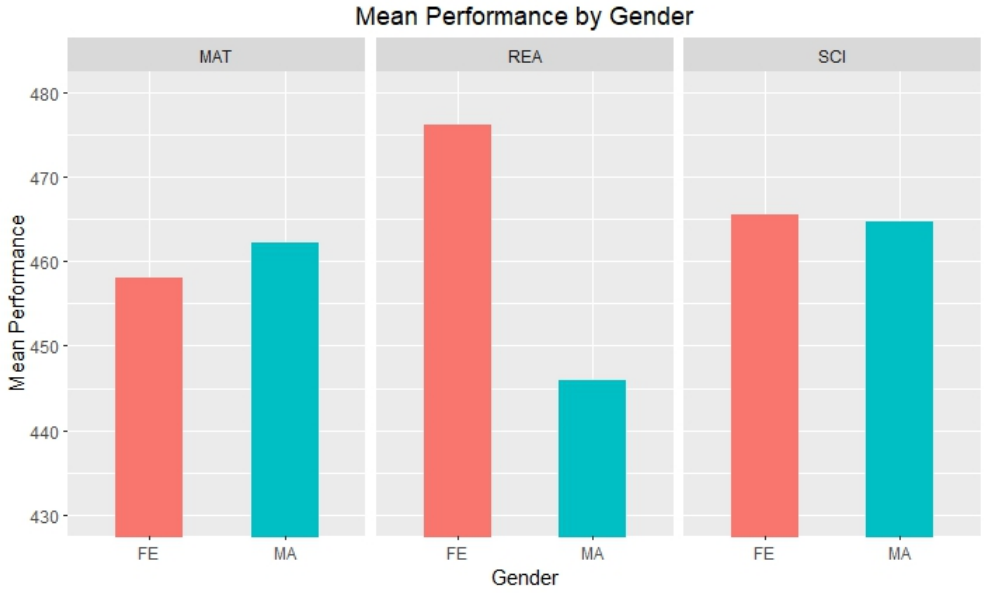
```
library(data.table);library(ggplot2)
# import data set
scores = fread("Pisa mean perfomance scores 2015 Data.csv", header = TRUE)
View(scores)
attach(scores)

# check types
classes = sapply(scores, class)
classes

# create new columns, rename and remove columns
series_code = scores[, `Series Code`]
scores[, `:=` (Discipline = substr(series_code,9,11),
               Gender = substr(series_code,13,14))]
setnames(scores,"2015","Performance")
scores = scores[, -c("Series Name", "Series Code")]

# fill missing values, remove rows, change column type
scores[Gender == "",Gender := 'All']
scores[Performance == "..", .N] #check all columns
scores = scores[Performance != "..",]
scores[, Performance:=as.numeric(Performance)]

# create column Region
unique(scores$`Country Code`)
amr = c("ARG", "BRA", "CAN", "CHL", "COL", "CRI",
        "DOM","MEX", "PER", "TTO", "USA", "URY")
afr = c("DZA", "TUN")
```

```
ocn = c("AUS", "NZL")
asia = c("HKG", "IDN", "ISR", "JPN", "JOR", "KAZ" , "KOR",
         "LBN", "MAC", "MYS","QAT", "SGP",  "THA", "ARE", "VNM")

scores[, `:=` (Region = "EU")]
scores[`Country Code` %in% amr,  Region := "AM"]
scores[`Country Code` %in% afr,  Region := "AF"]
scores[`Country Code` %in% ocn,  Region := "OC"]
scores[`Country Code` %in% asia,  Region := "AS"]
View(scores)
```

Απόσπασμα Κώδικα 1: Προεπεξεργασία δεδομένων

3 Ανάλυση Δεδομένων

Στην ενότητα 3 παρουσιάζεται το στάδιο της ανάλυσης. Αρχικά, εξετάστηκε η απόδοση των μαθητών, σε κάθε επστημονικό πεδίο, ανάλογα με το φύλο τους.



Σχήμα 1: Μέση επίδοση μαθητών ανάλογα με το φύλο τους

Από το Σχήμα 1 φαίνεται ότι το φύλο των μαθητών είναι ένας παράγοντας που επηρεάζει σημαντικά την απόδοση τους. Πιο συγκεκριμένα, στο πεδίο των μαθηματικών τα αγόρια πετυχαίνουν, κατά μέσο όρο, καλύτερες επιδόσεις από αυτές των κοριτσιών. Το αντίθετο συμβαίνει στο πεδίο της ανάγνωσης, όπου τα κορίτσια σημειώνουν σημαντικά καλύτερες επιδόσεις. Σε αυτό το πεδίο τα αγόρια πετυχαίνουν την χαμηλότερη τους απόδοση, ενώ τα κορίτσια την μέγιστη τους απόδοση, συγκριτικά με τα άλλα δυο επιστημονικά πεδία. Τέλος, στο πεδίο των επιστημών παρουσιάζεται μία πολύ εξισορροπημένη κατάσταση μεταξύ των αποδόσεων των δύο φύλων, με μία πολύ μικρή υπεροχή

των κοριτσιών.

Η προγραμματιστική υλοποίηση του Σχήματος 1 είναι η εξής:

```
# create table [Gender, Discipline, Mean Performance]
perf_gender = scores[Gender != "All", .(`Mean Performance` = mean(
    Performance)), by = .(Discipline, Gender)]

# plot bar chart
ggplot(perf_gender, aes(x = Gender,  y = `Mean Performance`, fill = Gender)
    ) +
  geom_bar(stat = "identity",  width = 0.5) +
  facet_grid(. ~ Discipline) +
  coord_cartesian(ylim = c(430, 480))  +
  ggtitle("Mean Performance by Gender")+
  theme(plot.title = element_text(hjust = 0.5), legend.position="none")
```

Απόσπασμα Κώδικα 2: Διαφορές μεταξύ των δύο φύλων

Για την εκτενέστερη ανάλυση της επίδρασης του φύλου στην απόδοση των μαθητών, εξετάστηκε το πως αυτή διαφοροποιείται για κάθε χώρα. Για να επιτευχθεί αυτό, βρέθηκε η διαφορά που υπάρχει στην απόδοση μεταξύ των δύο φύλων, σε κάθε χώρα και επιστημονικό πεδίο. Στο Σχήμα 2 απεικονίζονται αυτά τα αποτελέσματα.

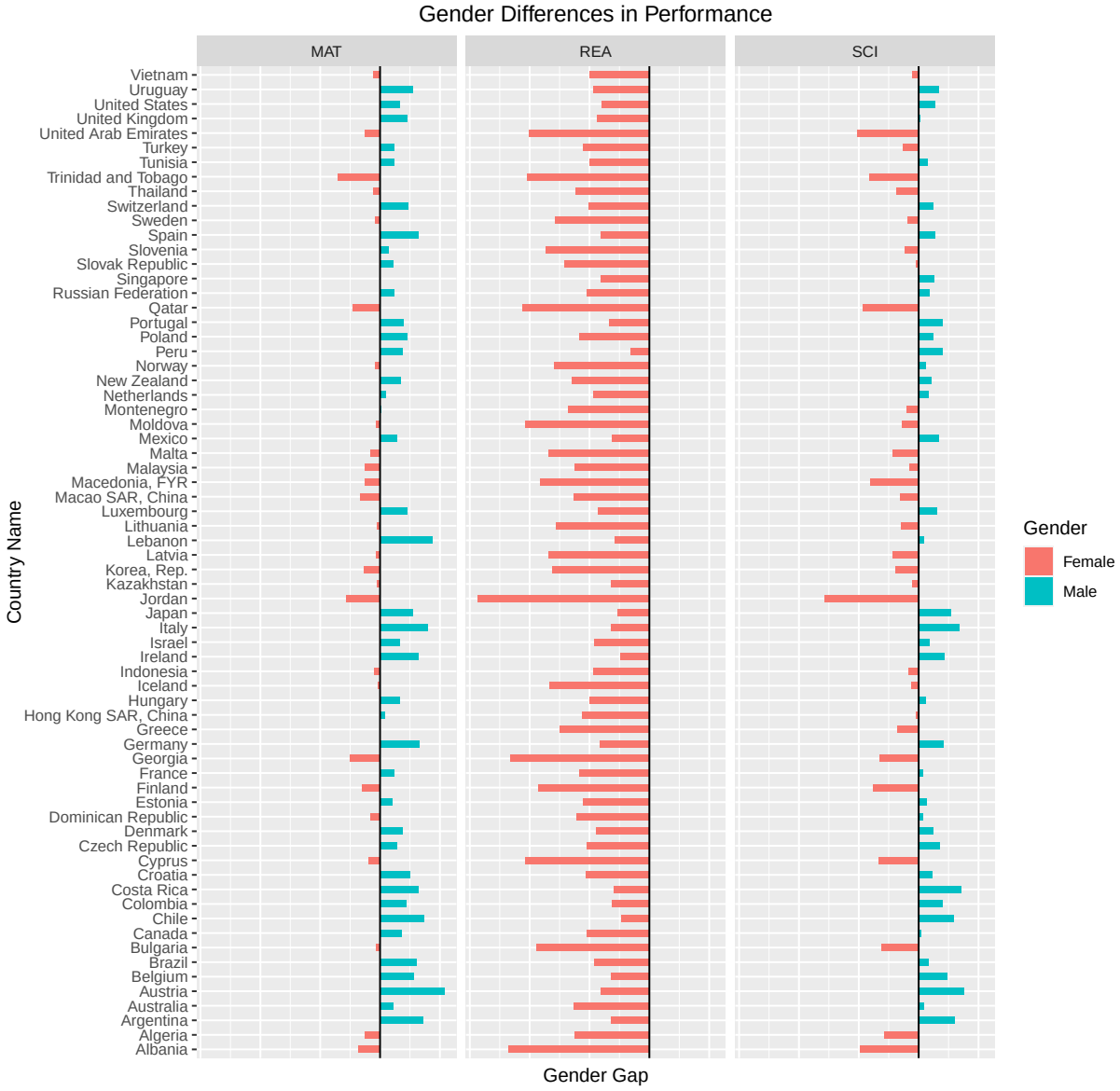
Από το Σχήμα 2, αρχικά επιβεβαιώνεται η μεγάλη απόκλιση των επιδόσεων των δύο φύλων στο πεδίο της ανάγνωσης. Σε καμία χώρα η απόδοση των αγοριών δεν είναι μεγαλύτερη από αυτή των κοριτσιών. Στο πεδίο των μαθηματικών, στις περισσότερες χώρες τα αγόρια σημειώνουν καλύτερες επιδόσεις, επαληθεύοντας τα όσα αναφέρθηκαν προηγουμένως. Στο πεδίο των επιστημών παρατηρείται ότι στις περισσότερες χώρες τα αγόρια πετυχαίνουν υψηλότερες επιδόσεις από αυτές των κοριτσιών, κάτι που έρχεται σε αντίθεση με τα προηγούμενα συμπεράσματα. Όμως, σε αρκετές χώρες στις οποίες τα κορίτσια υπερτερούν, οι επιδόσεις τους είναι με μεγάλη διαφορά καλύτερες από αυτές των αγοριών. Άρα, παρότι στις περισσότερες χώρες τα αγόρια υπερτερούν σε αυτόν τον τομέα, αυτές οι διαφορές εξισορροπούν και ανατρέπουν την κατάσταση όταν υπολογίζεται ο συνολικός μέσος όρος των αποδόσεων των δύο φύλων στο πεδίο των επιστημών.

Η προγραμματιστική υλοποίηση για την διαφορά των επιδόσεων είναι η εξής:

```
# create table [Country Name, Discipline, Gender Gap]
gender_dif = scores[Gender != "All", .(`Gender Gap` = diff(Performance)),
    by = .(`Country Name`, Discipline)]
gender_dif[`Gender Gap` > 0, Gender := "Male" ]
gender_dif[`Gender Gap` < 0, Gender := "Female" ]
gender_dif[`Gender Gap` > 0 & Discipline == "SCI", .N ] # comparsion in
    science

# plot figure 2
ggplot(gender_dif, aes(x=`Country Name`, y=`Gender Gap`, fill=Gender)) +
    geom_bar(stat='identity', width=.5) + facet_grid(. ~ Discipline) +
```

```
  geom_hline(yintercept = 0) +
  ggtitle("Gender Differences in Performance")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  coord_flip()
```

Απόσπασμα Κώδικα 3: Διαφορές μεταξύ των δύο φύλων ανά πεδίο και χώρα

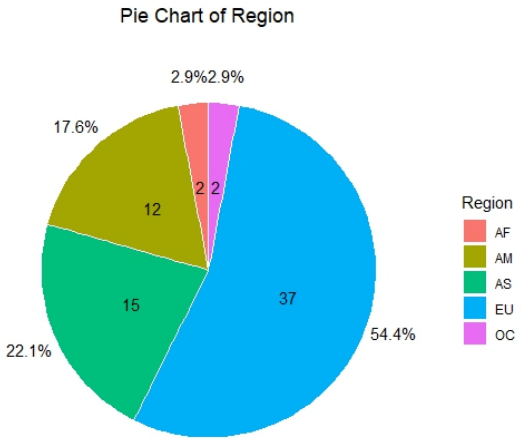


Σχήμα 2: Διαφορές στις επιδόσεις των μαθητών ανάλογα με το φύλο και την χώρα τους

Στη συνέχεια της ανάλυσης, συγκρίθηκαν οι επιδόσεις μεταξύ των χωρών που βρίσκονται σε διαφορετική γεωγραφική ήπειρο. Οι χώρες χωρίστηκαν στις εξής 5 ηπείρους: Ευρώπη, Αμερική, Αφρική, Ασία και Ωκεανία. Θεωρήθηκε απαραίτητο να βρεθεί ο αριθμός των χωρών από τις οποίες αποτελείται η κάθε ήπειρος.

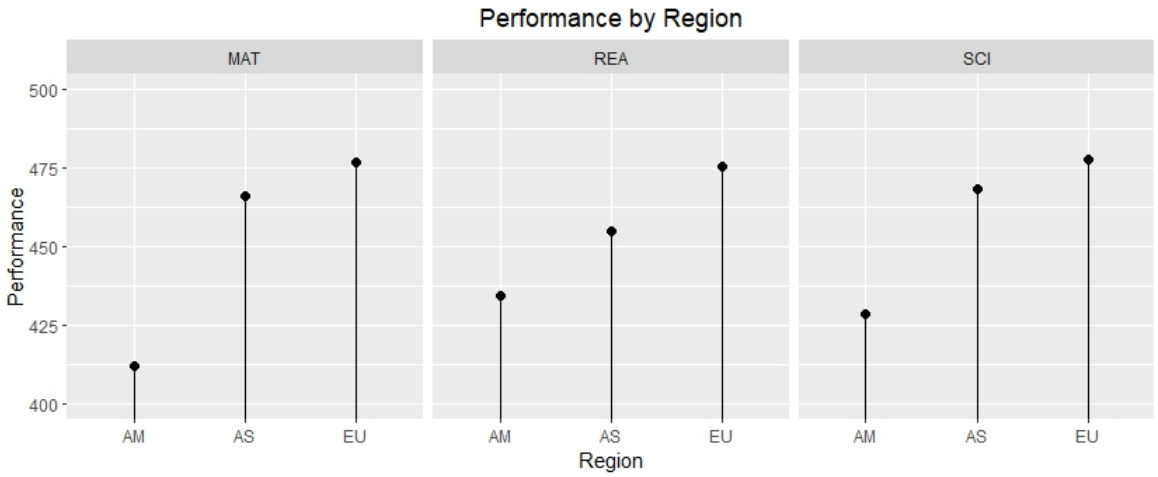
Όπως φαίνεται από το Σχήμα 3, η πλειονότητα των χωρών ανήκει στην Ευρώπη, ενώ

σημαντικό πλήθος χωρών υπάρχει στην Αμερική και στην Ασία. Δεν υπάρχουν αρκετά δεδομένα για χώρες που βρίσκονται στην Αφρική και στην Ωκεανία, αφού μόνο το 6% του συνολικού αριθμού των χωρών ανήκει εκεί. Επομένως, για την συνέχεια της ανάλυσης και όπου πραγματοποιούνται συγκρίσεις μόνο μεταξύ των ηπείρων, λήφθηκαν υπόψιν οι τρεις με το μεγαλύτερο ποσοστό. Έτσι, τα συμπεράσματα είναι πιο ισχυρά, ενώ παράλληλα συνεχίστηκε να αξιοποιείται μεγάλο ποσοστό των δεδομένων, περίπου το 94%.



Σχήμα 3: Πλήθος χωρών ανά ήπειρο

Στο διάγραμμα του Σχήματος 4 φαίνεται η μεταβολή της απόδοσης των μαθητών ανά ήπειρο σε κάθε επιστημονικό πεδίο. Σύμφωνα με το συγκεκριμένο σύνολο δεδομένων, οι Ευρωπαίοι μαθητές πετυχαίνουν κατά μέσο όρο τις υψηλότερες βαθμολογίες και στα τρία πεδία. Αντίθετα, οι μαθητές από την Αμερική καταλαμβάνουν την 3η θέση, μεταξύ των τριών, σε όλους τους εξεταζόμενους τομείς. Οι χώρες της Ασίας βρίσκονται ενδιάμεσα, ενώ στα πεδία των μαθηματικών και των επιστημών προσεγγίζουν σε μεγάλο βαθμό τις επιδόσεις αυτών της Ευρώπης.



Σχήμα 4: Επιδόσεις μαθητών ανά ήπειρο για κάθε επιστημονικό πεδίο

Η υλοποίηση των Σχημάτων 3 και 4 μέσω της R είναι η εξής:

```
#create table [Region, number_of_countries, percentage]
region_chart = scores[, .(nmb_of_countries = .N/9), by = .(Region)]
region_chart[, `:=` (freq = nmb_of_countries / sum(nmb_of_countries))]
```

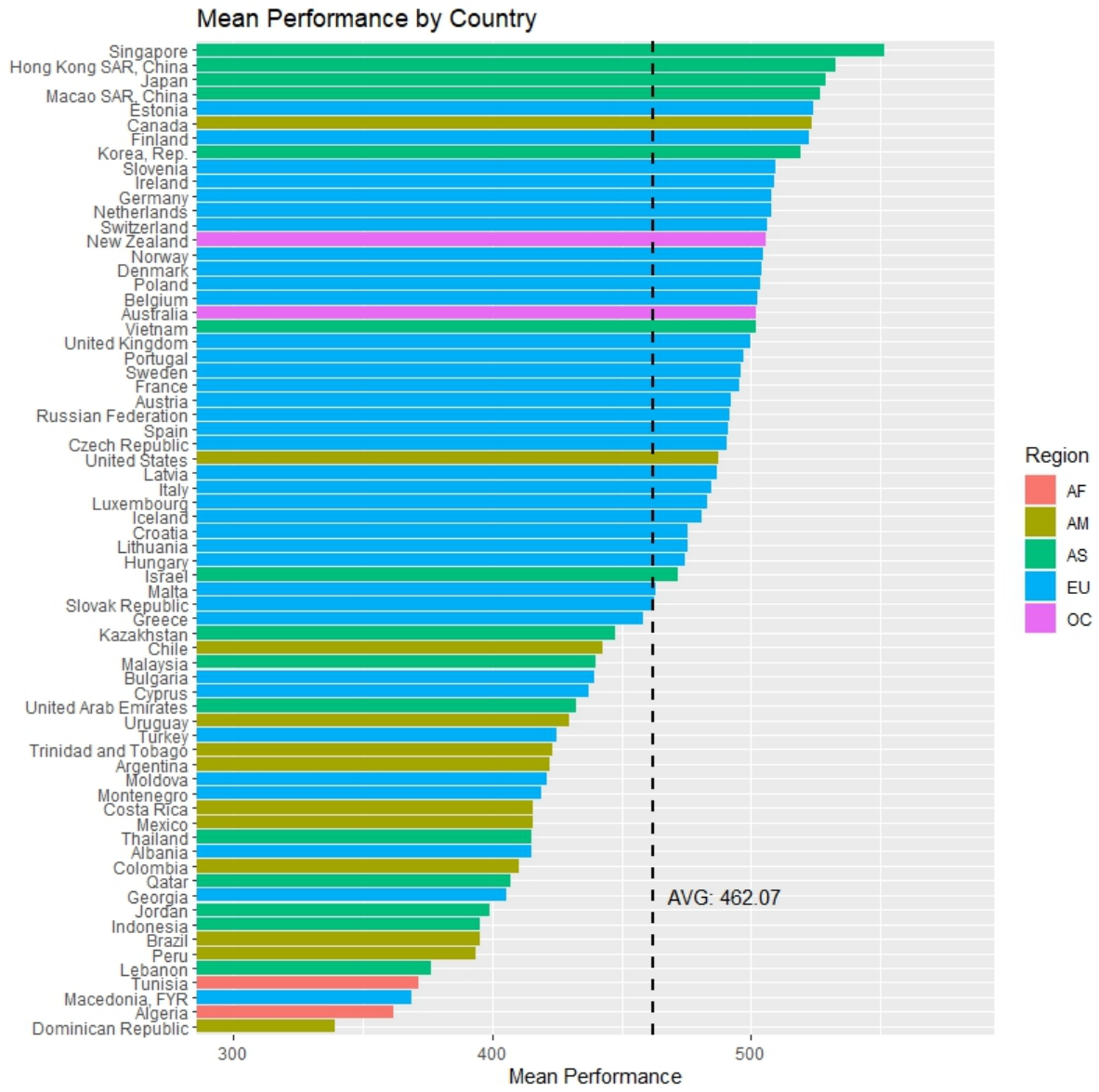
```
#plot pie chart
ggplot(region_chart, aes(x = "", y=nmb_of_countries, fill = Region)) +
  geom_bar(color = "white", stat = "identity") +
  geom_text(aes(x = 1.6, label = scales::percent(freq, accuracy = .1)),
            position = position_stack(vjust = .5)) +
  theme_void() +
  geom_text(aes(label = nmb_of_countries),
            position = position_stack(vjust = .5)) +
  theme(axis.line = element_blank(),
        plot.title = element_text(hjust = .5)) +
  ggtitle("Pie Chart of Region") +
  coord_polar("y")

#create table [Region, Discipline, Performance]
perf_region = scores[Gender == "All" & Region != "AF" & Region != "OC", .(
    Performance = mean(Performance)), by = .(Region, Discipline)]
perf_region
#plot lollipop chart
ggplot(perf_region, aes(x=Region, y=Performance)) +
  geom_point(size=2) +
  geom_segment(aes(x=Region,
                   xend=Region,
                   y=0,
                   yend=Performance)) +
  facet_grid(. ~ Discipline) +
  coord_cartesian(ylim = c(400, 500)) +
  labs(title="Performance by Region") +
  theme(plot.title = element_text(hjust = 0.5))
```

Απόσπασμα Κώδικα 4: Πλήθος χωρών και επίδοση ανά ήπειρο

Αφού έχει παρουσιαστεί η γενική κατάσταση που επικρατεί μεταξύ των ηπείρων, είναι σημαντικό να αναλυθεί εκτενέστερα ο τρόπος με τον οποίο επηρεάζει κάθε χώρα ξεχωριστά την συνολική επίδοση της ηπείρου στην οποία ανήκει. Στο Σχήμα 5, φαίνεται η κατάταξη τους βάσει του συνολικού μέσου όρου και στα τρία πεδία, δηλαδή δεν εξετάζεται το κάθε επιστημονικό πεδίο ξεχωριστά. Έτσι, ορίζεται ως μέτρο απόδοσης ο μέσος όρος των τριών τομέων για κάθε χώρα και εξάγεται ένα γενικό συμπέρασμα για τις βαθμολογίες των μαθητών.

Παρότι στο Σχήμα 4 φαίνεται οι χώρες της Ευρώπης να έχουν τις υψηλότερες, κατά μέσο όρο, βαθμολογίες, δεν είναι αυτές που καταλαμβάνουν τις πρώτες θέσεις στην κατάταξη του Σχήματος 5. Όμως, το γεγονός ότι οι περισσότερες Ευρωπαϊκές χώρες πετυχαίνουν επιδόσεις πάνω από το μέσο όρο διαμορφώνει τη συγκεκριμένη κατάσταση. Μεγάλες αποκλίσεις υπάρχουν στις αποδόσεις των χωρών της Ασίας, με την Σιγκαπούρη, την Κίνα, την Ιαπωνία και την Κορέα να πετυχαίνουν τις καλύτερες βαθμολογίες και τις υπόλοιπες να βρίσκονται σε χαμηλότερα στρώματα της κατάταξης. Από την άλλη, οι μα-

θητές της Αμερικής που σημειώνουν υψηλές βαθμολογίες βρίσκονται κυρίως στον Καναδά και στις ΗΠΑ, ενώ η πλειοψηφία των χωρών της συγκεκριμένης ηπείρου πετυχαίνει πολύ χαμηλότερες επιδόσεις. Τέλος, οι δύο χώρες της Ωκεανίας βρίσκοναι κοντά στο μέσο όρο των επιδόσεων, ενώ η Αλγερία και η Τυνησία από την Αφρική βρίσκονται στο τέλος της κατάταξης.



Σχήμα 5: Μέση επίδοση μαθητών ανά χώρα

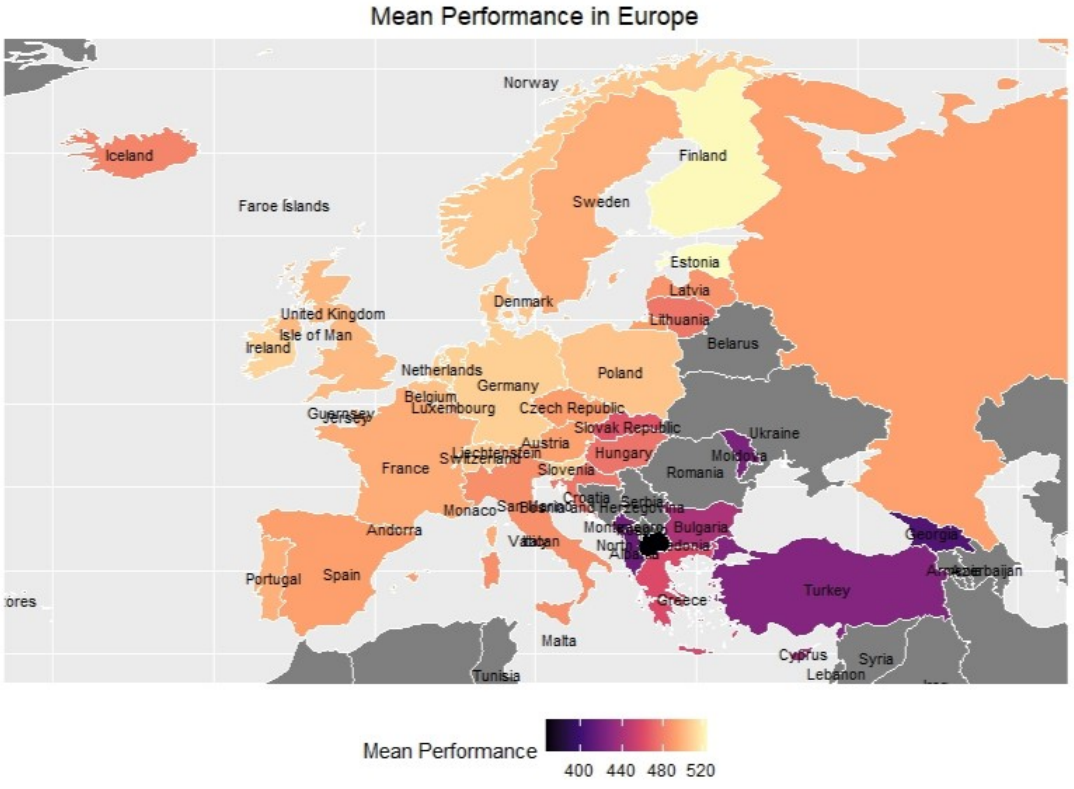
Η υλοποίηση του Σχήματος 5 μέσω της *R* είναι η εξής:

```
# create table [Country Name, Region, order(Mean Performance)]
avg_country = scores[Gender == "All", .(`Mean Performance` = mean(
    Performance)), by = .(`Country Name`, Region)]
avg_country = avg_country[order(avg_country$`Mean Performance`), ] # sort
```

```r
avg_country$`Country Name` = factor(avg_country$`Country Name`, levels =
    avg_country$`Country Name`)

# plot bar chart
ggplot(avg_country, aes(x =`Country Name`, y = `Mean Performance`, fill =
    Region)) +
  geom_bar(stat = "identity") +
  geom_hline(yintercept=mean(avg_country$`Mean Performance`), linetype="
    dashed", color = "black", size=1) +
  annotate("text", x=10, y=490, label=paste("AVG:", as.character(round(mean
    (avg_country$`Mean Performance`),2))) ) +
  ggtitle("Mean Performance by Country") +
  theme(axis.title.y=element_blank()) +
  coord_flip(ylim = c(300, 580))
```

Απόσπασμα Κώδικα 5: Μέση επίδοση ανά χώρα

Όπως αναφέρθηκε προηγουμένως, οι χώρες της Ευρώπης αποτελούν πάνω από το μισό του συνόλου των δεδομένων, περίπου το 55%. Έτσι, λοιπόν, ιδιαίτερη σημασία έχει η εκτενέστερη ανάλυση στο συγκεκριμένο υποσύνολο για την εξαγωγή συμπερασμάτων. Με την χρήση χάρτη (Σχήμα 6) παρουσιάζονται οι επιδόσεις ανά χώρα, για να γίνουν επιπλέον συγκρίσεις σχετικά με την γεωγραφική τους θέση. Εξετάζεται η συνολική απόδοση των μαθητών στους τρεις επιστημονικούς τομείς χωρίς να αναλύεται ξεχωριστά ο καθένας.



Σχήμα 6: Μέση επίδοση μαθητών στις χώρες της Ευρώπης

Από το Σχήμα 6 διαπιστώνεται ότι οι χώρες που σημειώνουν τις, κατά μέσο όρο, χαμηλότερες επιδόσεις βρίσονται σε κοντινή απόσταση μεταξύ τους. Με αύξουσα σειρά κατάταξης, η Βόρεια Μακεδονία, η Γεωργία, η Αλβανία, το Μαυροβούνιο, η Μοδαβία και η Τουρκία πετυχαίνουν το χαμηλότερο μέσο όρο επιδόσεων στην Ευρώπη. Αντίθετα, η Εσθονία και η Φινλανδία βρίσκονται πρώτες με μεγάλη διαφορά στη σχετική κατάταξη. Σημαντικές επιδόσεις πετυχαίνουν και οι μαθητές στη Σλοβενία, στην Ιρλανδία, στη Γερμανία, στην Ολλανδία και στην Ελβετία. Τέλος, η Ελλάδα βρίσκεται στη θέση 29 της σχετικής κατάταξης, στο σύνολο των 37 χωρών που υπάρχουν στα συγκεκριμένα δεδομένα.

Ο κώδικας R, και ορισμένες απαραίτητες επεξηγήσεις, για την κατασκευή του χάρτη:

```r
#create table [Country Name, Mean Performance] for European countries
avg_eu_perf = scores[Gender == "All" & Region == "EU", .(`Mean Performance` = mean(Performance)), by = .(`Country Name`)]

#ggplot package for map
map = map_data("world")
setDT(map)
attach(map)

# match some names between avg_eu_perf and map
map[region == "Russia", region := "Russian Federation"]
map[region == "UK", region := "United Kingdom"]
cmap[region == "Slovakia", region := "Slovak Republic"]
map[region == "North Macedonia", region := "Macedonia, FYR"]

# merge data tables
map_eu_perf = avg_eu_perf[map, on = .(`Country Name` == region), allow.cartesian=TRUE ]
c_names = aggregate(cbind(long, lat, group) ~ `Country Name`, data=map_eu_perf, FUN=function(x)mean(range(x)))

ggplot(map_eu_perf, aes(long, lat, group = group))+
  geom_polygon(aes(fill = `Mean Performance` ), color = "white")+
  geom_text(data=c_names, aes(long, lat, label = `Country Name`), size=3) +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        plot.title = element_text(hjust = 0.5),
        legend.position="bottom") +
  labs(title="Mean Performance in Europe")+
  scale_fill_viridis_c(option = "magma", trans = "sqrt") +
```

```
  coord_fixed(xlim = c(-25,50), ylim = c(35,70), ratio = 1.3)
# mean performance ranking
avg_eu_perf[order(`Mean Performance`)]
```

Απόσπασμα Κώδικα 6: Κατασκευή χάρτη Ευρώπης

Στο συγκεκριμένο κομμάτι κώδικα, έγινε χρήση του πακέτου map για να γίνει η οπτικοποίηση του χάρτη του Σχήματος 5. Ακόμα, χρειάστηκε να διορθωθούν κάποια ονόματα χωρών στο data table map για να είναι ίδια με τα αντίστοιχα του data table των βασικών δεδομένων και να γίνει σωστά η ένωση των δύο πινάκων (π.χ. τα ονόματα Slovakia και Slovakia Republic αναφέρονται στην ίδια χώρα). Τέλος, οι χώρες για τις οποίες δεν υπάρχουν δεδομένα απεικονίζονται με γκρίζο χρώμα.

Για να διαμορφωθεί καλύτερο συμπέρασμα σχετικά με τις αποδόσεις των χωρών, θα γίνει αναφορά και στις αποδόσεις τους σε κάθε επιστημονικό πεδίο ξεχωριστά. Ο Πίνακας 1 δείχνει σε ποιους τομείς υπερτερεί η κάθε χώρα και το Σχήμα 7 οπτικοποιεί αυτά τα δεδομένα προσφέροντας ορισμένα σημαντικά συμπεράσματα.

	Country Name	MAT	REA	SCI	Mean
1	Singapore	564.19	535.10	555.57	551.62
2	Hong Kong SAR, China	547.93	526.68	523.28	532.63
3	Japan	532.44	515.96	538.39	528.93
4	Macao SAR, China	543.81	508.69	528.55	527.02
5	Estonia	519.53	519.14	534.19	524.29
6	Canada	515.65	526.67	527.70	523.34
7	Finland	511.08	526.42	530.66	522.72
8	Korea, Rep.	524.11	517.44	515.81	519.12
9	Slovenia	509.92	505.22	512.86	509.33
10	Ireland	503.72	520.81	502.58	509.04
11	Germany	505.97	509.10	509.14	508.07
12	Netherlands	512.25	502.96	508.57	507.93
13	Switzerland	521.25	492.20	505.51	506.32
14	New Zealand	495.22	509.27	513.30	505.93
15	Norway	501.73	513.19	498.48	504.47
16	Denmark	511.09	499.81	501.94	504.28
17	Poland	504.47	505.70	501.44	503.87
18	Belgium	506.98	498.52	502.00	502.50
19	Australia	493.90	502.90	509.99	502.26
20	Vietnam	494.52	486.77	524.64	501.98
21	United Kingdom	492.48	497.97	509.22	499.89
22	Portugal	491.63	498.13	501.10	496.95
23	Sweden	493.92	500.16	493.42	495.83
24	France	492.92	499.31	494.98	495.73
25	Austria	496.74	484.87	495.04	492.22
26	Russian Federation	494.06	494.63	486.63	491.77
27	Spain	485.84	495.58	492.79	491.40
28	Czech Republic	492.33	487.25	492.83	490.80
29	United States	469.63	496.94	496.24	487.60
30	Latvia	482.31	487.76	490.23	486.76
31	Italy	489.73	484.76	480.55	485.01
32	Luxembourg	485.77	481.44	482.81	483.34
33	Iceland	488.03	481.53	473.23	480.93
34	Croatia	464.04	486.86	475.39	475.43

	Country Name	MAT	REA	SCI	Mean
35	Lithuania	478.38	472.41	475.41	475.40
36	Hungary	476.83	469.52	476.75	474.37
37	Israel	469.67	478.96	466.55	471.73
38	Malta	478.64	446.67	464.78	463.36
39	Slovak Republic	475.23	452.51	460.77	462.84
40	Greece	453.63	467.04	454.83	458.50
41	Kazakhstan	459.82	427.14	456.48	447.81
42	Chile	422.67	458.57	446.96	442.73
43	Malaysia	446.11	430.58	442.95	439.88
44	Bulgaria	441.19	431.72	445.77	439.56
45	Cyprus	437.14	442.84	432.60	437.53
46	United Arab Emirates	427.48	433.54	436.73	432.59
47	Uruguay	417.99	436.57	435.36	429.98
48	Turkey	420.45	428.34	425.49	424.76
49	Trinidad and Tobago	417.24	427.27	424.59	423.04
50	Argentina	409.03	425.30	432.23	422.19
51	Moldova	419.66	416.23	428.00	421.30
52	Montenegro	417.93	426.88	411.31	418.71
53	Costa Rica	400.25	427.49	419.61	415.78
54	Mexico	408.02	423.28	415.71	415.67
55	Thailand	415.46	409.13	421.34	415.31
56	Albania	413.16	405.26	427.23	415.21
57	Colombia	389.64	424.91	415.73	410.09
58	Qatar	402.40	401.89	417.61	407.30
59	Georgia	403.83	401.29	411.13	405.42
60	Jordan	380.26	408.10	408.67	399.01
61	Indonesia	386.11	397.26	403.10	395.49
62	Brazil	377.07	407.35	400.68	395.03
63	Peru	386.56	397.54	396.68	393.60
64	Lebanon	396.25	346.55	386.49	376.43
65	Tunisia	366.82	361.06	386.40	371.43
66	Macedonia, FYR	371.31	351.74	383.68	368.91
67	Algeria	359.61	349.86	375.75	361.74
68	Dominican Republic	327.70	357.74	331.64	339.03

Πίνακας 1: Επιδόσεις χωρών ανά επιστημονικό πεδίο

Από τον Πίνακα 1 φαίνεται ότι η κατάταξη του μέσου όρου που χρησιμοποιήθηκε σε προγούμενες συγκρίσεις δεν διαφέρει σημαντικά από τις επιμέρους επιδόσεις των

χωρών σε κάθε επιστημονικό τομέα. Για το πεδίο των μαθηματικών οι 5 πρώτες θέσεις απαρτίζονται από χώρες της Ασίας, με την Σιγκαπούρη να πετυχαίνει την υψηλότερη βαθμολογία μεταξύ όλων των επιστημονικών πεδίων και χωρών. Η Κίνα, η Ιαπωνία και η Κορέα ακολουθούν με υψηλές βαθμολογίες στον κλάδο των μαθηματικών. Από τις χώρες της Ευρώπης η Ελβετία σημειώνει την καλύτερη επίδοση, παρότι, όπως έχει αναφερθεί, δεν βρίσκεται στις υψηλότερες θέσεις βάσει των συνολικών αποδόσεων. Όσον αφορά στο πεδίο της ανάγνωσης, παρατηρείται μία μικρή πτώση των χωρών της Ασίας και μία άνοδο από εκείνες της Αμερικής. Η Σιγκαπούρη και το Hong Kong στην Κίνα διατηρούν τις δύο πρώτες θέσεις και με ελάχιστη διαφορά στην απόδοση ακολουθούν ο Καναδάς, η Φινλανδία και η Ιρλανδία. Τέλος, για τον τομέα των επιστημών υπάρχει βελτίωση στις χώρες της Ευρώπης, με την Εσθονία και την Φινλανδία να βρίσκονται μόνο πίσω από την Σιγκαπούρη και την Ιαπωνία. Μη αναμενόμενη είναι η επίδοση των μαθητών από το Βιετνάμ, που πετυχαίνουν την 6η καλύτερη στο συγκεκιμένο πεδίο.
Ακολουθεί η προγραματιστική υλοποίηση με ορισμένες επεξηγήσεις για τον Πίνακα 1:

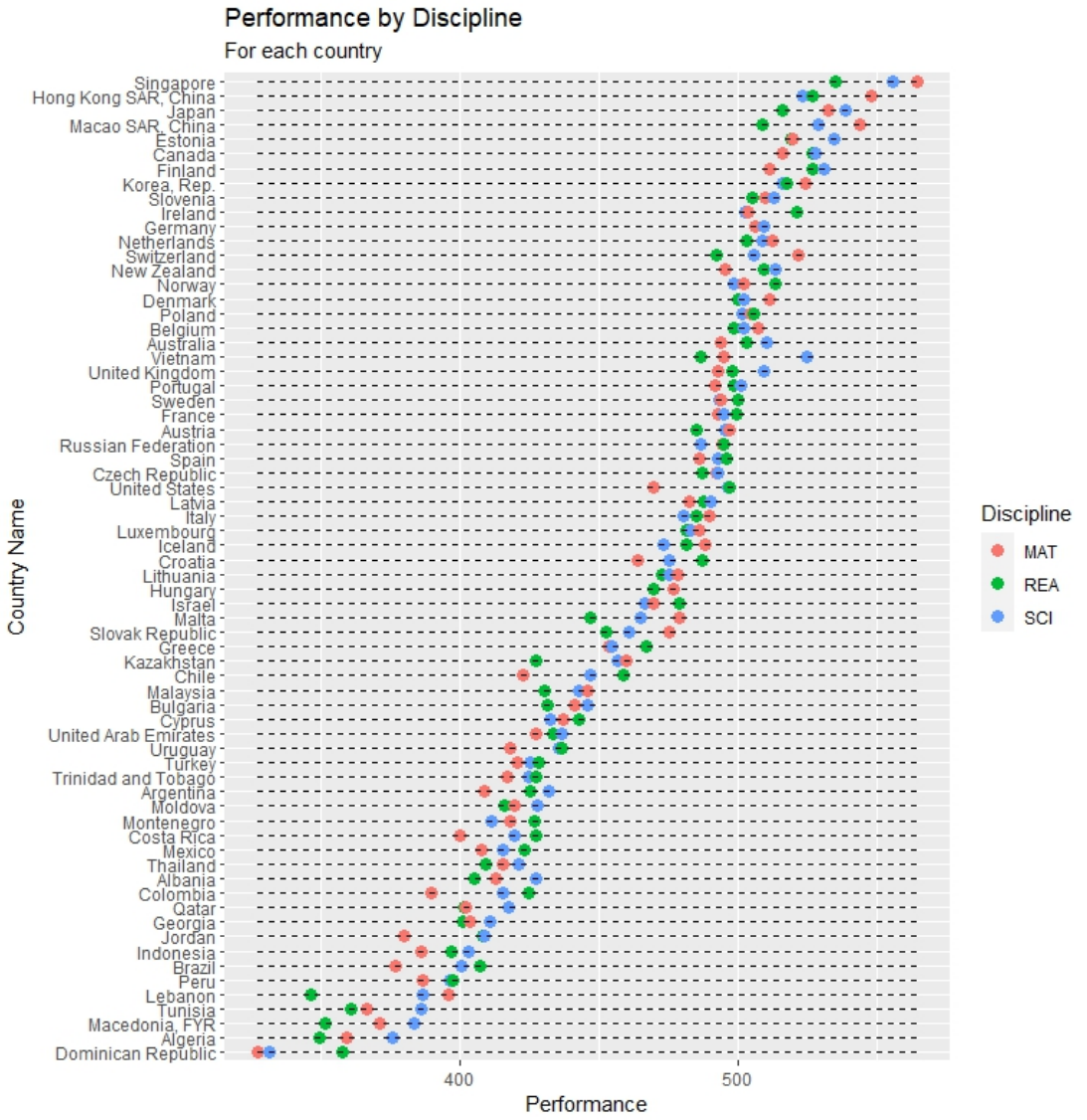
```
# create 3 tables [Country Names, Performance by Discipline]
mat_country = scores[Gender == "All" & Discipline == "MAT", .(`Country Name
    `, Performance)]
setnames(mat_country,"Performance","MAT")
rea_country = scores[Gender == "All" & Discipline == "REA", .(`Country Name
    `, Performance)]
setnames(rea_country,"Performance","REA")
sci_country = scores[Gender == "All" & Discipline == "SCI", .(`Country Name
    `, Performance)]
setnames(sci_country,"Performance","SCI")
avg_discipline = mat_country[rea_country, on = .(`Country Name`) ][sci_
    country, on = . (`Country Name`)]

avg_discipline[,`:=` (`Mean Performance` = (avg_discipline$MAT + avg_
    discipline$REA + avg_discipline$SCI)/3)]
avg_discipline = avg_discipline[order(-avg_discipline$`Mean Performance`),]
     # sort by mean performance
library(xtable)
print(xtable(avg_discipline, type = 'latex')) # create latex text
```

Απόσπασμα Κώδικα 7: Κατασκευή πίνακα χωρών

Για την υλοποίηση στην R, κατασκευάστηκαν τρεις πίνακες (ένας για κάθε τομέα) και ενώθηκαν, ώστε σε κάθε γραμμή να υπάρχουν οι επιδόσεις των χωρών στα διαφορετικά πεδία. Ακόμα, με την χρήση της βιβλιοθήκης xtable δημιουργήθηκε το κείμενο latex για τον Πίνακα 1.

Από το Σχήμα 7 φαίνεται ότι ο κλάδος των μαθηματικών συγκεντρώνει την χαμηλότερη βαθμολογία για τις περισσότερες χώρες. Κυρίως για τις χώρες που βρίσκονται στις

χαμηλότερες θέσεις της κατάταξης το φαινόμενο αυτό εμφανίζεται εντονότερα, αφού οι κόκκινες κουκίδες βρίσκονται αρκετά πιο αριστερά από τις άλλες δύο. Άλλη μία διαπίστωση είναι ότι για τις χώρες με υψηλότερες επιδόσεις οι βαθμολογίες μεταξύ των πεδίων δεν έχουν μεγάλες αποκλίσεις, ενώ για όσες βρίσκονται χαμηλά υπάρχουν πιο μεγάλες διαφορές στις επιδόσεις ανά επιστημονικό τομέα. Μία πιθανή ερμηνεία για το φαινόμενο αυτό είναι ότι οι μαθητές με υψηλές επιδόσεις καταβάλουν την αντίστοιχη προσπάθεια που απαιτείται σε κάθε τομέα για να μεγιστοποιήσουν την απόδοση τους, ενώ οι μαθητές με χαμηλότερες επιδόσεις δεν επιδιώκουν να βελτιώσουν πιθανές αδυναμίες τους και έτσι προκύπτουν αυτές οι διαφορές.



Σχήμα 7: Επιδόσεις μαθητών ανά επιστημονικό πεδίο για κάθε χώρα

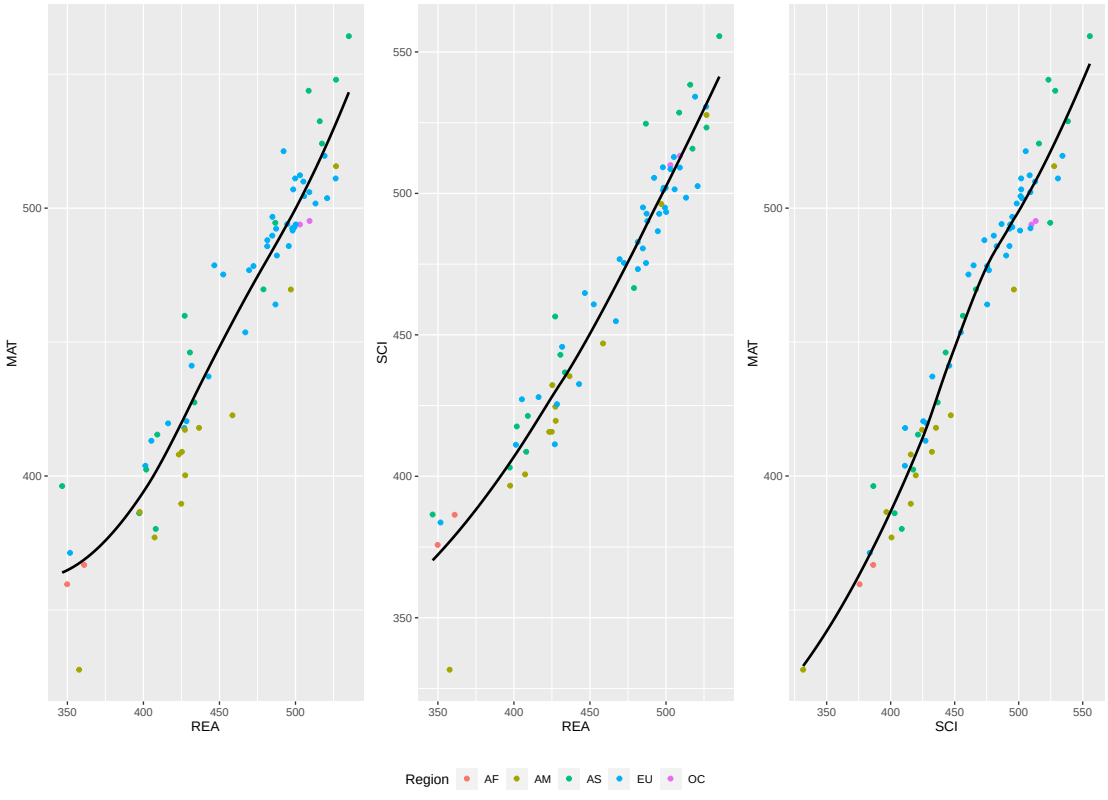
Η προγραμματιστική υλοποίηση του Σχήματος 7 είναι η εξής:

```
perf_disc = scores[Gender == "All",.(Performance, Discipline),by=.(`Country
    Name`)]
```

```
perf_disc[, `:=` (Mean=mean(Performance)), by =.(`Country Name`)]
perf_disc = perf_disc[order(-perf_disc$Mean, perf_disc$Performance), ] #
    sort
perf_disc$`Country Name` = factor(perf_disc$`Country Name`, levels =rev(
    unique(perf_disc$`Country Name`)), ordered=TRUE)

ggplot(perf_disc, aes(x=`Country Name`, y=Performance)) +
  geom_point(aes(col=Discipline), size=3) +
  geom_segment(aes(x=`Country Name`,
                   xend=`Country Name`,
                   y=min(Performance),
                   yend=max(Performance)),
                   linetype="dashed",
                   size=0.1) +
  labs(title="Performance by Discipline",
       subtitle="For each country") +
  coord_flip()
```

Απόσπασμα Κώδικα 8: Επιδόσεις ανά πεδίο και χώρα

Κοιτάζοντας τον Πίνακα 1 και το Σχήμα 7 δημιουργούνται ερωτήματα γύρω από την ύπαρξη συσχετίσεων μεταξύ των επιδόσεων στα διαφορετικά επιστημονικά πεδία. Στο επόμενο μέρος της ανάλυσης θα εξεταστεί λεπτομερώς η ύπαρξη τέτοιων συσχετίσεων, βάσει του Σχήματος 8.



Σχήμα 8: Συσχετίσεις μεταξύ επιστημονικών πεδίων

Από το Σχήμα 8 είναι προφανές ότι τα τρία επιστημονικά πεδία διατηρούν υψηλές συσχετίσεις. Οι τομείς των μαθηματικών και των επιστημών παρουσιάζουν την μεγαλύτερη συσχέτιση (συντελεστής Pearson 0.974), ενώ τα πεδία των μαθηματικών και την ανάγνωσης την χαμηλότερη (συντελεστής Pearson 0.939). Σημαντική συσχέτιση φαίνεται να διατηρούν και τα πεδία της ανάγνωσης με αυτά των επιστημών (συντελεστής Pearson 0.96). Ως προς την ερμηνεία των παραπάνω συσχετίσεων, είναι φυσικό τα πεδία των μαθηματικών και της ανάγνωσης να έχουν την χαμηλότερη συσχέτιση, αφού τα αντικείμενα που εξετάζουν δεν είναι παρεμφερή. Ανίθετα, τα μαθηματικά με το πεδίο των επιστημών έχουν πολλά κοινά στα περιεχόμενα τους και για αυτό η απόδοση των μαθητών παρουσιάζει ισχυρή συσχέτιση.

Ακολουθεί η προγραμματιστική υλοποίηση του Σχήματος 8:

```
# create 3 tables [Country Names, Performance by Discipline, Region]
mat_country = scores[Gender == "All" & Discipline == "MAT", .(`Country Name
    `, Performance)]
setnames(mat_country,"Performance","MAT")
rea_country = scores[Gender == "All" & Discipline == "REA", .(`Country Name
    `, Performance)]
setnames(rea_country,"Performance","REA")
sci_country = scores[Gender == "All" & Discipline == "SCI", .(`Country Name
    `, Performance, Region)]
setnames(sci_country,"Performance","SCI")
corr_disc = mat_country[rea_country, on = .(`Country Name`) ][sci_country,
    on = . (`Country Name`)]

# 3 plots
p1 = ggplot(corr_disc, aes(x = REA, y = MAT)) +
  geom_point(aes(color = Region)) +
  geom_smooth(method = 'loess', color = "black", se=F) +
  theme(legend.position="none")
p2 = ggplot(corr_disc, aes(x = REA, y = SCI)) +
  geom_point(aes(color = Region)) +
  geom_smooth(method = 'loess', color = "black", se=F) +
  theme(legend.position="none")
p3 = ggplot(corr_disc, aes(x = SCI, y = MAT)) +
  geom_point(aes(color = Region)) +
  geom_smooth(method = 'loess', color = "black", se=F)+
  theme(legend.position="none")
p=plot_grid(p1, p2, p3, ncol=3)

# create one legend
legend = get_legend(p1 + guides(color = guide_legend(nrow = 1)) +
                      theme(legend.position = "bottom") )
library(cowplot)
plot_grid(p, legend, ncol=1, rel_heights = c(1, .1))
```

```
30
31 # find correlation
32 cor(corr_disc$SCI, corr_disc$REA)
```

Απόσπασμα Κώδικα 9: Συσχετίσεις επιστημονικών πεδίων

4 Σύνοψη Παρατηρήσεων και Συμπερασμάτων

Καταλήγοντας, από την παρούσα ανάλυση διαπιστώθηκε ότι το φύλο και η χώρα των μαθητών είναι παράγοντες που επηρεάζουν σημαντικά την επίδοσή τους. Αρχικά, τα κορίτσια υπερτερούν στον τομέα της ανάγνωσης, ενώ το ανίθετο συμβαίνει στα μαθηματικά όπου τα αγόρια σημειώνουν καλύτερες βαθμολογίες. Οι μαθητές Ευρωπαϊκών χωρών φάνηκε να καταγράφουν, κατά μέσο όρο, υψηλότερες αποδόσεις συγκριτικά με τις υπόλοιπες ηπείρους και στα τρία πεδία. Στην Ασία παρατηρήθηκαν μεγάλες αποκλίσεις στις επιδόσεις των χωρών, αφού οριμένες κατέγραψαν τις κορυφαίες επιδόσεις παγκοσμίως ανά επιστημονικό πεδίο και οι υπόλοιπες βρέθηκαν στα χαμηλότερα στρώματα των σχετικών κατατάξεων. Τέλος, βρέθηκαν ισχυρές συσχετίσεις μεταξύ των επιδόσεων στα διαφορετικά επιστημονικά πεδία.

Ο εμπλουτισμός των δεδομένων είναι πιθανόν να οδηγήσει σε βαθύτερα συμπεράσματα και σημαντικότερες ανακαλύψεις. Πιο συγκεκριμένα, η μελέτη της μεταβολής των επιδόσεων στον χρόνο, που δεν εξετάστηκε στην παρούσα εργασία, και η αξιοποίηση δεδομένων σχετικά με την οικονομική κατάσταση και το εκπαιδευτικό σύστημα της κάθε χώρας είναι κατευθύνσεις στις οποίες μπορεί να επεκταθεί η ανάλυση. Τέλος, η σύγκριση μαθητών διαφορετικών ηλικιών αποτελεί ακόμα μία πρόταση για να συνεχιστεί η παραπάνω ανάλυση.