Εθνικό Μετσόβιο Πολυτεχνείο

Προγραμματιστικά Εργαλεία και Τεχνολογίες
στην Επιστήμη των Δεδομένων

Διερευνητική Ανάλυση Δεδομένων με τη χρήση της R

Σταύρος Κετσέας
stavrosketseas@mail.ntua.gr
ΑΜ 03400084
Μεταπτυχιακός Φοιτητής ΕΔΕΜΜ

22 Ιανουαρίου 2022

1 Εισαγωγή

1.1 Covid-19

Ο Covid-19 είναι ιός του αναπνευστικού συστήματος που εντοπίστηκε για πρώτη φορά στα τέλη του 2019 στην πόλη Γουχάν της Κίνας. Λόγω της τεράστιας μεταδοτικότητάς του, απλώθηκε με γρήγορο ρυθμό στον υπόλοιπο κόσμο και μέσα σε λίγους μήνες κατηγοριοποιήθηκε ως πανδημία. Από την έναρξη της πανδημίας εως σήμερα, 5.5 εκατομμύρια άνθρωποι παγκοσμίως έχουν χάσει τη ζωή τους από τον ιό. Εκτός από το τίμημα των ανθρώπινων ζωών, ο Covid-19 είχε δραματικές επιπτώσεις και στις υπόλοιπες εκφάνσεις της καθημερινότητας, από ατομικό επίπεδο εως επίπεδο χωρών, οικονομίας κλπ.

Όπως ήταν αναμενόμενο, η μελέτη της πανδημίας προσέλκυσε ένα μεγάλο μέρος της επιστημονικής έρευνας τα τελευταία χρόνια. Ειδικότερα, ο κλάδος της επιστήμης των δεδομένων αντιμετώπισε μια τεράστια πρόκληση, καθώς ένα πλήθος ερωτημάτων για την πανδημία μπορούσε, και ήταν απαραίτητο, να προσεγγιστεί με μεγάλη προσοχή μέσω εργαλείων στατιστικής.

Ίσως τα κυριότερο όπλα εναντίον του Covid-19 ήταν μια σειρά από εμβόλια που παρήχθησαν ανεξάρτητα και δόθηκαν στην κυκλοφορία στα τέλη του 2020. Η πορεία των εμβολιασμών ανά χώρα αποτελεί και το βασικό περιεχόμενο του συνόλου δεδομένων μας. Συγκεκριμένα, το dataset μας περιλαμβάνει στατιστικά για το (αθροιστικό) πλήθος των εμβολιασμών που έχει πραγματοποιήσει κάθε χώρα, για κάθε ημέρα της πανδημίας. Για τα μονοδοσικά εμβόλια, ένας εμβολιασμός καθιστά ένα άτομο πλήρως εμβολιασμένο. Για τα πολυδοσικά, ο πρώτος εμβολιασμός το καθιστά μερικώς εμβολιασμένο, μέχρι την τελευταία δόση η οποία το καθιστά πλήρως εμβολιασμένο. Το σύνολο δεδομένων μας περιέχει τόσο το πλήθος των πλήρως εμβολιασμένων, όσο και αυτό των μερικώς εμβολιασμένων.

1.2 Σκοπός της εργασίας

Ο σκοπός αυτής της εργασίας είναι καταρχήν η επεξεργασία των δεδομένων με τη χρήση του στατιστικού πακέτου R, και εν συνεχεία η πραγματοποίηση διερευνητικής ανάλυσης, δηλαδή στατιστικής ανάλυσης που εστιάζει περισσότερο στην οπτική παρουσίαση των αποτελεσμάτων μέσω κατάλληλων γραφημάτων, και λιγότερο σε αυστηρό μαθηματικό φορμαλισμό. Μια τέτοια προσέγγιση είναι χρήσιμη, αφενός διότι πολλές φορές εποπτικά διακρίνονται πιο εύκολα τάσεις που ίσως να είναι κρυμμένες πίσω από τα δεδομένα, δίνοντας έτσι σωστές κατευθύνσεις στον ερευνητή για το τι εργαλεία και μοντέλα να χρησιμοποιήσει, και αφετέρου επειδή είναι ιδιαίτερα αποτελεσματική στη μετάδοση της πληροφορίας σε άτομα που δεν έχουν το απαραίτητο υπόβαθρο να κατανοήσουν μια πιο τεχνική μελέτη.

Κατά τη συγγραφή αυτής της εργασίας, και πραγματοποιώντας κάποιους τυπικούς συγκριτικούς ελέγχους ανάμεσα σε χώρες, διαπιστώσαμε τις τεράστιες διαφορές που δεν οφείλονταν σε απλή τυχαιότητα. Αποφασίσαμε το κύριο αντικείμενό μας να είναι η ανάδειξη αυτών των ανισοτήτων και η προσπάθεια ερμηνείας τους.

Όπως κάθε αγαθό για το οποίο η ζήτηση είναι αντιστρόφως ανάλογη της ποσότητάς του, έτσι για τα εν λόγω εμβόλια η αρχική τους διάθεση έγινε με αρκετά μεροληπτικό τρόπο. Αν και οι λεπτομέρειες για τη διαδικασία σύμφωνα με την οποία πουλήθηκαν στις χώρες είναι ακόμα απόρρητες, τα δεδομένα που έχουμε στα χέρια μας επιτρέπουν να εξάγουμε κάποια έμμεσα συμπεράσματα.

Σύμφωνα με τη γνωστή Αρχή του Pareto, γνωστή και ως νόμο του 80-20, για ένα μεγάλο μέρος δραστηριοτήτων, το 80% των αποτελεσμάτων οφείλεται στο 20% των προσπαθειών. Προσαρμοσμένη στο πρόβλημά μας, η αρχή αυτή θα έλεγε ότι το 80% των εμβολίων έχει καταλήξει στο 20% των χωρών. Στην πραγματικότητα, τα νούμερα αυτά είναι αρκετά πιο ενδιαφέροντα. Θα δούμε ότι το 80% των εμβολίων έχει καταλήξει στο 9% των χωρών, στο 50.3% του συνολικού πληθυσμού της γης, αλλά και στο 66.5% του συνολικού ΑΕΠ. Θα γίνει σαφές ότι η διάθεση των εμβολίων δεν έγινε ποτέ με ανθρωπιστικό γνώμονα.

Εκτός από το κομμάτι της στατιστικής ανάλυσης των δεδομένων, δόθηκε ιδιαίτερη βαρύτητα και στη σωστή χρήση της R για την πραγματοποίηση αυτού. Έτσι, ο κώδικας της R, τόσο για την επεξεργασία των δεδομένων όσο και για το σχεδιασμό των διαφόρων διαγραμμάτων, περιλαμβάνεται στο κυρίως κείμενο και καταλαμβάνει ένα σημαντικό μέρος της εργασίας μας.

2 Προεπεξεργασία

2.1 Εγκατάσταση πακέτων

Για τη σωστή φόρτωση του πακέτου `coronavirus`, πρώτα από όλα απαιτήθηκε η έκδοση της R με την οποία δουλεύαμε να είναι πλήρως ενημερωμένη, καθώς επίσης και το βοηθητικό πακέτο `readr`. Εγκαταστήσαμε τα εν λόγω πακέτα, μαζί με κάποια ακόμα που θα χρειαζόμασταν αργότερα στα πλαίσια πλέον της στατιστικής ανάλυσης, με τις παρακάτω εντολές:

Εγκατάσταση πακέτων

```
# Packages required to read the dataset.
install.packages('coronavirus')
install.packages('readr')

# Packages for data manipulation and visualization.
install.packages('data.table')
install.packages('ggplot2')
install.packages('directlabels')
install.packages('RColorBrewer')
install.packages('ggrepel')

# Packages for drawing maps.
install.packages(c('rgeos', 'cowplot', 'googleway', 'ggspatial', 'libwgeom',
  'sf', 'rnaturalearth', 'rnaturalearthdata'))

# Package containing various country metrics, including the GDP.
install.packages('wbstats')
```

Η εγκατάσταση των πακέτων πραγματοποιήθηκε μία μόνο φορά. Στην έναρξη κάθε καινούργιας συνεδρίας, φορτώναμε τα πακέτα ως εξής:

Φόρτωση πακέτων

```
# We load the packages.
library('data.table')     # For working with tables
library('ggplot2')        # For data visualization
library('directlabels')   # For ggplot labelling
library('coronavirus')    # The covid-19 dataset
```

```
library('readr')          # Required by the coronavirus package
library('scales')         # For changing the diagram scales
library('ggrepel')        # For labeling diagrams
library('wbstats')        # GDPs
library('RColorBrewer')   # Color palettes
library('sf')             # Map drawing packages
library('rnaturalearth')
library('rnaturalearthdata')
library('rgeos')
```

2.2 Φόρτωση και προεπεξεργασία δεδομένων

Τα δεδομένα που θα χρησιμοποιήσουμε περιέχονται στο πακέτο της R με το όνομα coronavirus. Το εν λόγω πακέτο, περιέχει δύο μεγάλα datasets, ένα με στατιστικά κρουσμάτων/θανάτων/αναρρώσεων που ονομάζεται `coronavirus`, και ένα με στατιστικά για τον αριθμό των εμβολιασμένων ατόμων να χώρα, που ονομάζεται `covid19_vaccine` και στο οποίο θα εστιάσουμε σε αυτή την εργασία.

Να σημειώσουμε σε αυτό το σημείο ότι ενώ το πρώτο dataset ενημερώνεται καθημερινά, *δεν* ισχύει το ίδιο για το dataset με το οποίο θα δουλέψουμε εμείς. Για ένα μεγάλο διάστημα τα πιο πρόσφατα δεδομένα του `covid19_vaccine` έφταναν μέχρι τον Οκτώβρη του 2020. Παρ'όλα αυτά, κατά τη συγγραφή της εργασίας πραγματοποιήθηκε εκ νέου ανανέωση και πλέον τα δεδομένα μας φτάνουν μέχρι και την 1/1/2022.

Ο `covid19_vaccine` είναι ένας πίνακας με συνολικά 18 μεταβλητές (στήλες). Από αυτές μας ενδιαφέρουν κυρίως οι ακόλουθες:

country_region: Το όνομα της χώρας/περιοχής.
date: Η ημερομηνία της παρατήρησης.
doses_admin: Αθροιστικό πλήθος χορηγηθείσων δόσεων μέχρι την τρέχουσα ημερομηνία. Για εμβόλια στα οποία απαιτούνται περισσότερες της μιας δόσης, κάθε δόση προσμετράται ξεχωριστά.
people_partially_vaccinated: Αθροιστικό πλήθος μερικώς εμβολιασμένων μέχρι την τρέχουσα ημερομηνία.
people_fully_vaccinated: Αθροιστικό πλήθος πλήρως εμβολιασμένων μέχρι την τρέχουσα ημερομηνία.
population: Ο πληθυσμός της χώρας/περιοχής.

2.2.1 Βασική επεξεργασία

Πρώτη επαφή με το vaccine dataset

```
# Load and update the vaccine dataset.
vac = read_csv("https://raw.githubusercontent.com/RamiKrispin/coronavirus/
  master/csv/covid19_vaccine.csv",
  col_types = cols(date=col_date(format="%Y-%m-%d")))

# Convert them in a data table format.
setDT(vac)

# Add the World population.
vac[country_region=='World']$population = 7753000000
```

```
# Create a column with the relative frequency of the fully vaccinated people.
vac$fully_vaccinated_ratio = vac$people_fully_vaccinated / vac$population
vac$fratio = vac$fully_vaccinated_ratio # Keep the numeric ratio.

# Convert the relative frequencies into a percentage format.
vac$fully_vaccinated_ratio = percent(vac$fully_vaccinated_ratio, .1)

# Repeat for the partially vaccinated people.
vac$partially_vaccinated_ratio = vac$people_partially_vaccinated /
  vac$population
vac$pratio = vac$partially_vaccinated_ratio

vac$partially_vaccinated_ratio = percent(vac$partially_vaccinated_ratio, .1)
```

Το επόμενο βήμα είναι να κρατήσουμε τα δεδομένα σε επίπεδο χωρών. Παρατηρήσαμε ότι οι καταχωρίσεις που προέρχονται από επαρχίες, έχουν στη στήλη `combined_key` ένα αναγνωριστικό της μορφής "Province, Country". Αντίθετα, για τις καταχωρίσεις που αφορούν τα αθροιστικά στατιστικά σε επίπεδο χωρών, το `combined_key` τους είναι της μορφής "Country". Αυτό το γεγονός καθιστά την εύρεση των γραμμών που αφορούν σε στατιστικά επιπέδου χώρας αρκετά εύκολη, απλά ελέγχουμε κατά πόσον το combined key τους ταυτίζεται με το όνομα της χώρας (country region).

Στατιστικά σε επίπεδο χωρών

```
# The provinces have "province, country" as a combined key.
# The aggregated country stats have a plain "country" instead.

# We also want to keep the world data, so we make sure the World
# entries also satisfy the previous requirement.
vac[country_region=='World']$combined_key = 'World'
vac_cl = vac[country_region==combined_key]
```

2.2.2 Περαιτέρω επεξεργασία

Το σύνολο δεδομένων μας πλέον έχει μια πιο εύχρηστη μορφή. Παρ'όλα αυτά, θα συνεχίσουμε με επιπλέον επεξεργασία, είτε για να διορθώσουμε πιθανές παραλήψεις του, είτε για να το φέρουμε σε μορφή που θα εξυπηρετεί καλύτερα τα ερωτήματα που σκοπεύουμε να μελετήσουμε. Ένας διαγνωστικός έλεγχος μέσω της εντολής

```
unique(vac_cl$continent_name)
[1] "North America" NA "Asia" "Europe" "South America" "Africa"  "Oceania"
```

αποκάλυψε την ύπαρξη χωρών που δεν έχουν αντιστοιχηθεί σε κάποια ήπειρο (εμφάνιση NA). Ενώ αυτό αναμένετο για τις καταχωρίσεις του "World", τα παγκόσμια data δεν ήταν η μοναδική πηγή προβλημάτων:

Χώρες δίχως ήπειρο

```
# The indices of the countries which are missing their continents.
prob_ind = which(is.na(vac_cl$continent_name))

# The countries that correspond to them.
prob_countries = unique(vac_cl[prob_ind]$country_region)
```

```
prob_countries
[1] "World"  "Sudan"  "Kosovo"

# We fill their continents by ourselves.
vac_cl[vac_cl$country_region=='Sudan']$continent_name = 'Africa'
vac_cl[vac_cl$country_region=='Kosovo']$continent_name = 'Europe'
```

Επιλύσαμε το πρόβλημα εισάγοντας με το χέρι της ηπείρους για το Κόσοβο και το Σουδάν.

3 Διερευνητική ανάλυση

3.1 Ποσοστά πλήρως εμβολιασμένων

Σε αυτή την παράγραφο η κύρια μεταβλητή ενδιαφέροντος αποτελεί το ποσοστό των πλήρως εμβολιασμένων ατόμων για κάθε χώρα. Μας ενδιαφέρει να δούμε ποιες χώρες ξεχωρίζουν, ποιες υπολείπονται, καθώς και το αν υπάρχουν σημαντικές διαφοροποιήσεις ανάμεσα σε ηπείρους ή ολόκληρες περιοχές εντός της ίδιας ηπείρου.

Καθώς τα ποσοστά των εμβολιασμένων ατόμων αλλάζουν μέρα με τη μέρα, χρειαζόμαστε μια μετρική που θα είναι ανεξάρτητη του χρόνου και η οποία θα είναι αντιπροσωπευτική για τη γενική εικόνα της υπό μελέτης χώρας. Επιλέξαμε αυτή η μετρική να είναι το μέγιστο ποσοστό των πλήρως εμβολιασμένων που πέτυχε η κάθε χώρα καθόλη τη διάρκεια της πανδημίας, δηλαδή υπολογίζουμε το ποσοστό των πλήρως εμβολιασμένων για μια συγκεκριμένη χώρα καθημερινά (δηλ. την πληροφορία της στήλης `fratio`) και κρατάμε τη μέγιστη τιμή του.

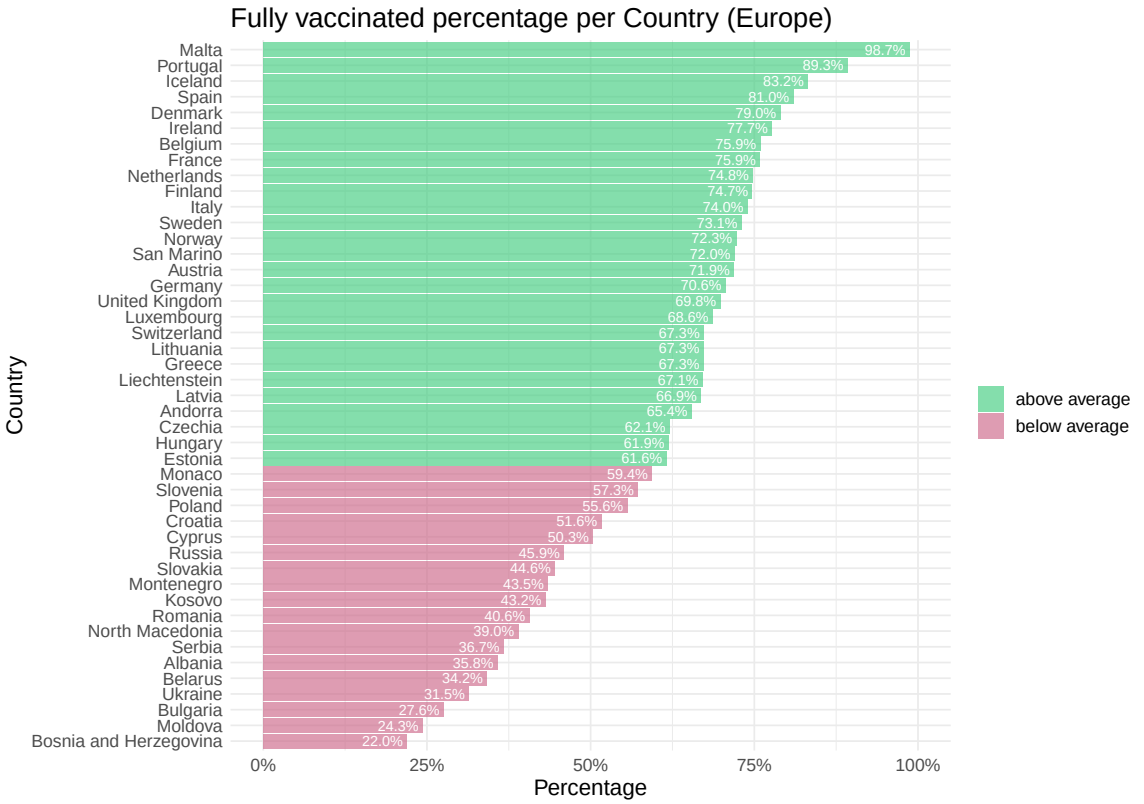
Μέγιστο ποσοστό πλήρως εμβολιασμένων

```
# For each country, compute the maximum full vaccination ratio.
max_full = vac_cl[vac_cl[, .I[which.max(fratio)], by=country_region]$V1]
```

Ο πίνακας `max_full` περιέχει πλέον μία γραμμή για κάθε χώρα: την καταχώριση της χώρας στον αρχικό πίνακα η οποία αντιστοιχεί στην ημέρα όπου επετεύχθη το μέγιστο fratio.

Ευρωπαϊκά ποσοστά (Σχήμα 1)

```
# Keep the European countries.
max_full_eu = max_full[continent_name=='Europe']

# Sort them in a descending order w.r.t. the vaccination percentage.
eu_sorted = max_full_eu[order(-fratio)]

# Compute the mean vaccination percentage within Europe.
eu_mean = mean(eu_sorted$fratio)

# Find which countries lie above and below the mean. Included for
# illustration purposes.
eu_sorted$above_average = ifelse(eu_sorted$fratio>=eu_mean,
  'above average', 'below average')
eu_sorted$above_average = as.factor(eu_sorted$above_average)

# Plot the diagram.
```



Σχήμα 1: Μέγιστα ποσοστά πλήρως εμβολιασμένων ανά χώρα για τις χώρες τις Ευρώπης. Οι χώρες χρωματίζονται ανάλογα με το αν υπερτερούν ή υπολείπονται του μέσου όρου ολόκληρης της Ευρωπαϊκής ηπείρου (60.2%).

```
ggplot(data=eu_sorted, aes(x=reorder(country_region, fratio), y=fratio,
    fill=above_average)) + theme_minimal() + coord_flip() +
  geom_bar(stat="identity", alpha=0.6) +
  labs(x="Country", y="Percentage",
       title="Fully vaccinated percentage per Country (Europe)") +
  scale_fill_manual(' ', values=c("seagreen3","palevioletred3")) +
  scale_y_continuous(labels=scales::percent, limits=c(0,1)) +
  geom_text(aes(label=fully_vaccinated_ratio, x=country_region, y=fratio),
  position=position_dodge(width=0.8), vjust=0.5, col='white', size=2, hjust=1)
```

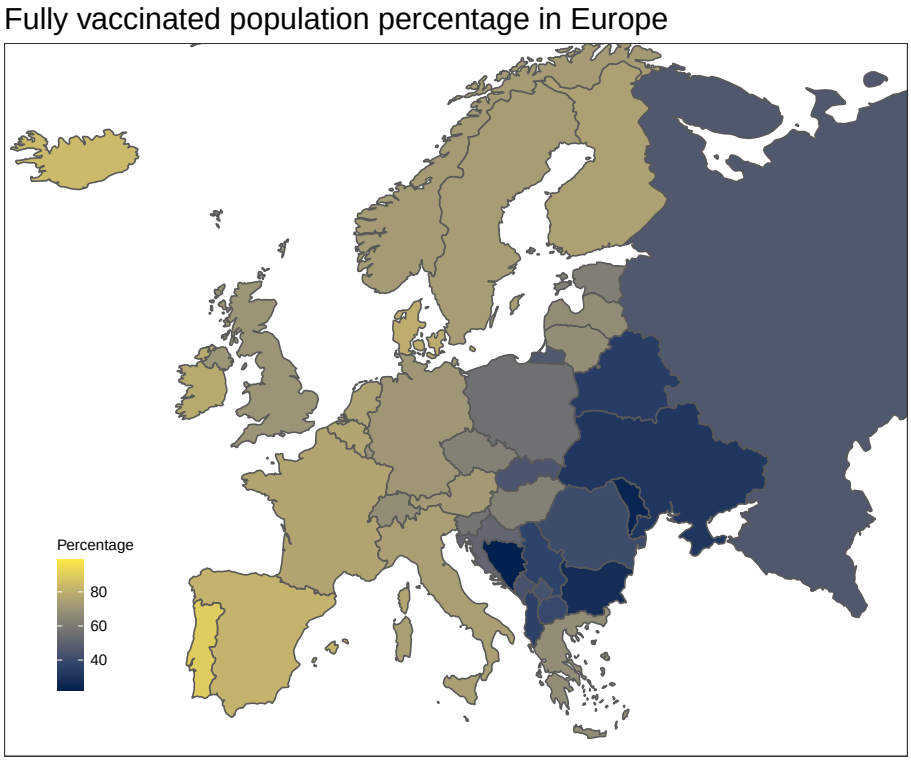
Ως μια προκαταρκτική ανάλυση, απομονώσαμε τα ποσοστά εμβολιασμών για την Ευρωπαϊκή ήπειρο και τα απεικονίσαμε στο Σχήμα 1. Παρατηρούμε ότι η Μάλτα υπερτερεί με διαφορά των υπολοίπων με ποσοστό εμβολιασμού στο 98.7%. Μια πιο ενδιαφέρουσα παρατήρηση είναι ότι οι χώρες οι οποίες τα πηγαίνουν χειρότερα, δείχνουν να είναι οι Βαλκανικές χώρες, καθώς και οι χώρες του πρώην ανατολικού μπλοκ. Μια περαιτέρω ανάλυση με ταυτόχρονη απεικόνιση των δεδομένων στον Ευρωπαϊκό χάρτη θα αποσαφηνίσει αν όντως υπάρχει ή όχι το μοτίβο αυτό.

3.2 Γεωχωρική ανάλυση

Ευρωπαϊκός χάρτης (Σχήμα 2)

```
# Load the world map.
world = ne_countries(scale="medium", returnclass="sf")

# Keep the European countries.
Europe = world[which(world$continent=="Europe"), ]

# Add a column in the Europe data frame with the vaccination ratios
# for the countries that have the  same names in both tables.
for (i in 1:nrow(max_full_eu)){
  row = which(Europe$name==max_full_eu[i, "country_region"]$country_region)
  Europe[row, "my_ratio"] = max_full_eu[i, "fratio"]
}

# Identify the countries which have been left out.
Europe$name[which(is.na(Europe$my_ratio))]
[1] "Aland"  "Bosnia and Herz."  "Czech Rep."  "Faeroe Is."
[5] "Guernsey"  "Isle of Man"  "Jersey"  "Macedonia"  "Vatican"

# We match the countries which have different names in the two tables.

# Bosnia
row = which(Europe$name=='Bosnia and Herz.')
i = which(max_full_eu$country_region=='Bosnia and Herzegovina')
Europe[row, "my_ratio"] = max_full_eu[i, "fratio"]

# Czechia
row = which(Europe$name=='Czech Rep.')
i = which(max_full_eu$country_region=='Czechia')
Europe[row, "my_ratio"] = max_full_eu[i, "fratio"]

# North Macedonia
row = which(Europe$name=='Macedonia')
i = which(max_full_eu$country_region=='North Macedonia')
Europe[row, "my_ratio"] = max_full_eu[i, "fratio"]

# We ignore the remaining missing countries, "Aland",
# "Faeroe Is.", "Guernsey", "Isle of Man", "Jersey", "Vatican".

# Plot the European map.
ggplot(Europe) + geom_sf(aes(fill=my_ratio*100)) +
  coord_sf(xlim=c(-25,52), ylim=c(34,71), expand=FALSE) +
  theme_bw() + scale_fill_viridis_c(option='cividis') +
  labs(title="Fully vaccinated population percentage in Europe",
    fill='Percentage') + theme(legend.position=c(0.1, 0.2))
```

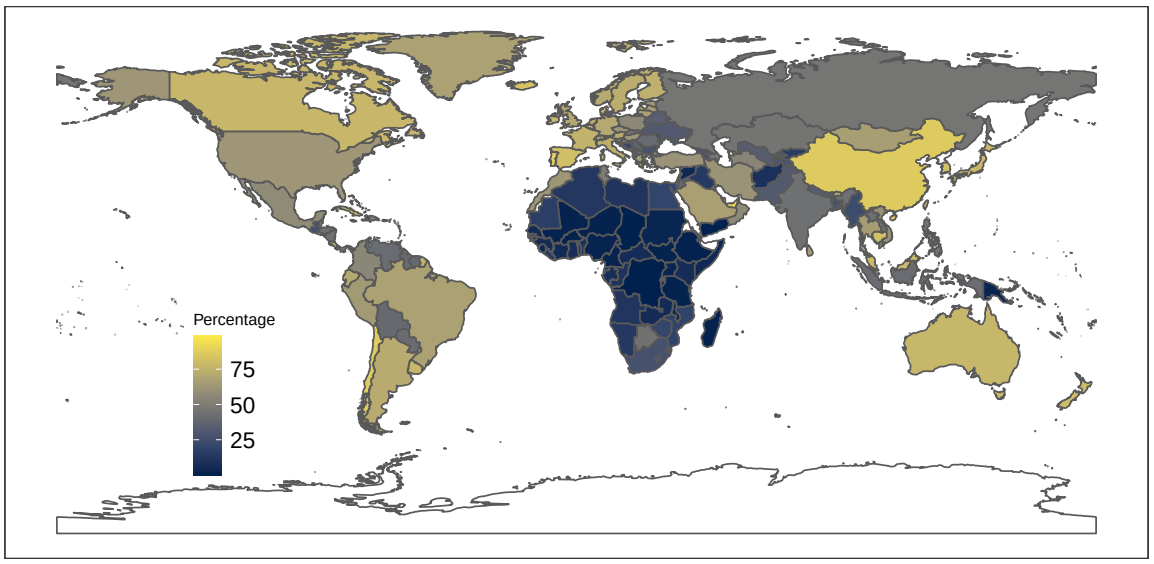
Από το χάρτη του Σχήματος 2 διαπιστώνουμε πράγματι ότι οι περισσότερες χώρες της ανατολικής Ευρώπης βρίσκονται σημαντικά πίσω σε σχέση με αυτές της δυτικής Ευρώπης. Το ζήτημα αυτό είναι σημαντικό και αξίζει μια περαιτέρω διερεύνηση. Καταρχήν, ζωγραφίζουμε ολόκληρο τον παγκόσμιο χάρτη για να δούμε αν υπάρχουν γενικευμένες αντίστοιχες τάσεις, δηλαδή γειτονικές ομάδες χωρών που παρουσιάζουν παρόμοια υψηλά ή χαμηλά νούμερα.



Σχήμα 2: Μέγιστα ποσοστά πλήρως εμβολιασμένων ανά χώρα για τις χώρες τις Ευρώπης. Με εξαίρεση την Ελλάδα, οι περισσότερες χώρες της ανατολικής Ευρώπης υπολείπονται των υπολοίπων.

Παγκόσμιος χάρτης (Σχήμα 3)

```
# Load the world map.
world = ne_countries(scale="medium", returnclass="sf")

# Add a column in the World data frame with the vaccination ratios for
# the countries that have the  same names in both data tables.
for (i in 1:nrow(max_full)){
  row = which(world$name==max_full[i,"country_region"]$country_region)
  world[row,"my_ratio2"] = max_full[i,"fratio"]
}

# Include the countries with different names.
row = which(world$name_long=='United States')
i = which(max_full$country_region=='US')
world[row, "my_ratio2"] = max_full[i, "fratio"]
# Bosnia, Czechia and North Macedonia were also added this way.

# Countries missing altogether had to be added on our own
# from the official WHO site.
mis_coun = list('Libya', 'Aruba', 'Anguilla', 'Burundi', 'Benin',
  'Burkina Faso', 'Bermuda', 'Central African Rep.', 'Dem. Rep. Congo',
  'Congo', 'Cape Verde', 'Cuba', 'Curaçao', 'Djibouti', 'Hong Kong', 'Haiti',
  'Kiribati', 'Liberia', 'Lesotho', 'Madagascar', 'Niger', 'Palestine',
  'Somalia', 'Turkmenistan', 'Tajikistan', 'Chad', 'Tanzania',  'Samoa',
  'Yemen', 'Korea', 'Vanuatu', 'British Virgin Is.', 'Tonga',  'Saint Helena',
  'Fr. Polynesia', 'Puerto Rico', 'Nauru', 'New Caledonia',
  'Turks and Caicos Is.', 'Greenland', 'Micronesia')
```



Σχήμα 3: Μέγιστα ποσοστά πλήρως εμβολιασμένων ανά χώρα. Η αφρικανική ήπειρος ξεχωρίζει για τα χαμηλά ποσοστά της, ενώ η αμερικανική και η ευρωπαϊκή ήπειρος προπορεύονται στο θέμα του εμβολιασμού. Χώρες για τις οποίες δεν έχουμε δεδομένα εμφανίζονται με λευκό χρώμα.

```
mis_rat = list(.115, .739, .611, .0002, .099, .016, .768, .071, .001,
  .1, .46, .848, .623, .026, .62, .006, .19, .149, .3, .02, .019, .308,
  .048, .532, .29, .005, .015, .614, .012, .819, .164, .558, .536, .624,
  .566, .768, .669, .637, .701, .675, .374 )

# Add the previous countries' ratios to the 'World' data frame.
for (i in 1:length(mis_coun)){
  row = which(world$name==mis_coun[[i]])
  world[row, "my_ratio2"] = mis_rat[[i]]
}

# Plot the global map.
ggplot(world) + geom_sf(aes(fill=my_ratio2*100)) + theme_bw() +
  scale_fill_viridis_c(option='cividis', na.value="white") +
  scale_color_viridis_c(option='cividis') +
  labs(title = "Fully vaccinated population percentage", fill='Percentage') +
  theme(legend.position = c(0.2, 0.3))
```

Ο χάρτης του Σχήματος 3 αναδεικνύει το πρόβλημα. Η συντριπτική πλειοψηφία των Αφρικανικών χωρών έχουν μείνει δραματικά πίσω στη μάχη του εμβολιασμού. Το ίδιο ισχύει και για αρκετές χώρες της ασιατικής ηπείρου, με τα υψηλά ποσοστά της Κίνας να τονίζουν την αντίθεση με τις γειτονικές της χώρες.

3.3 Ποσοστά εμβολιασμού και ΑΕΠ

Το πρόβλημα, ενώ με μια πρώτη ματιά ίσως να φαίνεται γεωχωρικό, στην πραγματικότητα είναι αρκετά πιο πολυσύνθετο. Η πιο λογική υπόθεση είναι ότι οι διαφαινόμενες ανισότητες οφείλονται σε οικονομικούς λόγους, και πως το γεωχωρικό μοτίβο εξηγείται

απλά από το ότι η οικονομική κατάσταση δε διαφέρει σημαντικά ανάμεσα σε γειτονικές χώρες. Ελέγχουμε την υπόθεση αυτή, εισάγοντας στη συζήτηση το Εθνικό Ακαθάριστο Προϊόν (ΑΕΠ) της κάθε χώρας, μέσω του πακέτου `wbstats`.

ΑΕΠ

```
# Load world country metrics, only keep their 2020 GDP.
wb_dat = wb(indicator = "NY.GDP.PCAP.KD")
setDT(wb_dat)
gdp = wb_dat[date==2020]
```

Ο σκοπός μας δεν είναι απλά να συγκρίνουμε το μέγιστο fratio με το ΑΕΠ της κάθε χώρας. Η ανισότητα στο θέμα του εμβολιασμού δεν εκδηλώθηκε μόνο μέσω αυτού του στατιστικού, αλλά με την πολιτική διάθεσης των εμβολίων καθ'όλη τη διάρκεια της πανδημίας, και ιδιαίτερα κατά το ξεκίνημα, όπου ο ανταγωνισμός μεταξύ των χωρών ήταν εντονότερος. Υπήρχαν περιπτώσεις πολλών χωρών (πχ. Ηνωμένο Βασίλειο, ΗΠΑ, Καναδάς, Ιαπωνία κλπ) οι οποίες μονοπώλησαν την αγορά εμβολίων, κατορθώνοντας να αγοράσουν περισσότερα εμβόλια από όσα χρειάζονταν βάσει του πληθυσμού τους.

Αντί λοιπόν να ασχοληθούμε με το μέγιστο fratio της κάθε χώρας, μας ενδιαφέρει ολόκληρη η πορεία που ακολούθησε αυτό το στατιστικό, και ιδιαίτερα κατά τους πρώτους μήνες του λανσαρίσματος των εμβολίων. Για το λόγο αυτό, εισάγαμε κάποιες ημερομηνίες $d_1, \dots, d_7$, τις οποίες ονομάσαμε checkpoints, κατά τις οποίες κοιτάζουμε το fratio κάθε χώρας. Καθώς το fratio δε μεταβάλλεται σημαντικά από μέρα σε μέρα, επιλέξαμε λίγα σχετικά checkpoints, τα οποία φροντίσαμε να είναι πυκνότερα κατά την έναρξη της διάθεσης των εμβολίων για τους προαναφερθέντες λόγους.

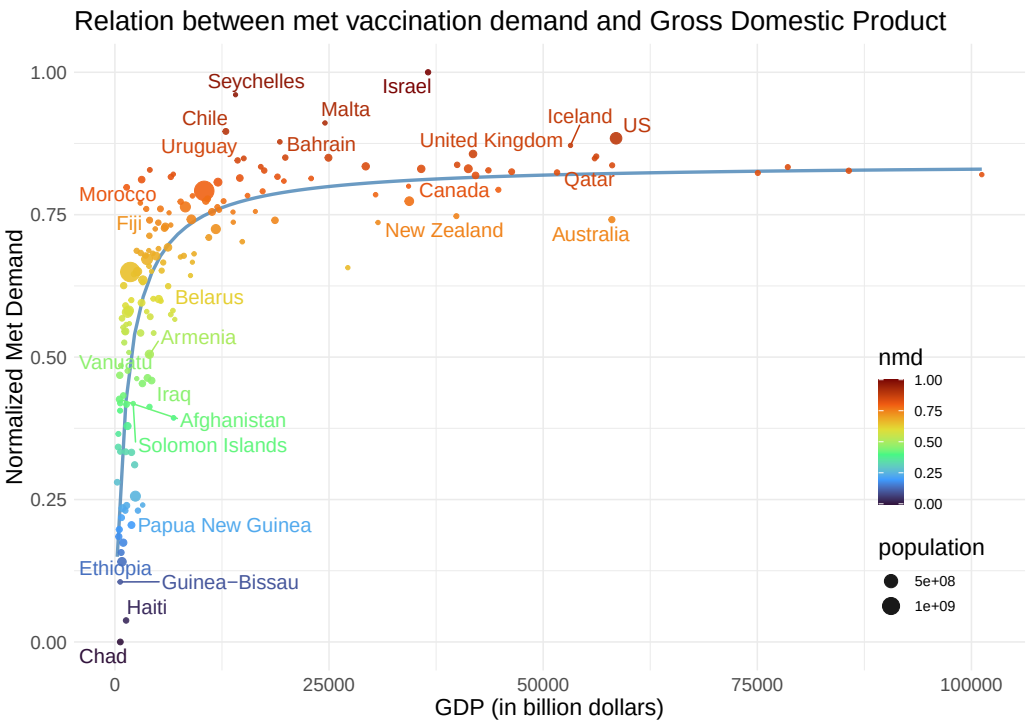
Στη συνέχεια ορίζουμε τη μετρική μας να είναι η $\mathrm{MD} = \sum w_i d_i$, όπου $(w_1, \dots, w_7)$ είναι ένα διάνυσμα βαρών, που δίνει μεγαλύτερο βάρος όσο πιο πίσω πηγαίνουν οι ημερομηνίες. Το στατιστικό μας το ονομάσαμε MD από το Met Demand. Αφού υπολογίσαμε το MD κάθε χώρας, στη συνέχεια πήραμε το λογάριθμό του και το κανονικοποιήσαμε, ορίζοντας την ποσότητα αυτή ως Normalized Met Demand (NMD). Αυτό έγινε καθαρά για λόγους παρουσίασης και δεν επηρέασε καθόλου τα συμπεράσματά μας.

ΑΕΠ vs NMD (Σχήμα 4)

```
# Keep only the entries of vac_cl at the dates of interest.
checkpoints = c('2021-02-01', '2021-03-01', '2021-04-01', '2021-05-01',
                '2021-07-01', '2021-09-01', '2021-12-01')
vac_cl$date_str = as.character(vac_cl$date)
new_dt = vac_cl[date_str %in% checkpoints]

# Compute the met demand metric using the following function.

compute_metric = function(vector, n=7, weights=c(5, 4, 3, 2, 1, 1, 1)){
  # Given a vector of length <=7, we augment its first coordinates
  # with zeros until it has length equal to 7, and then compute
  # the logarithm of the inner product <vector, weights>.

  k = length(vector)
  filler = rep(0, n-k)
  new_vec = c(filler, vector)
  return(log(sum(new_vec * weights)))
  }

# A function that normalizes a given vector.
```

```
min_max_norm = function(x) {(x - min(x)) / (max(x) - min(x))}

# Initialize an empty data frame which will store the country
# info as well as the met demand metric.
new_data = data.frame(country=character(), continent=character(),
                      metric=double(), gdp=double(), population=double(),
                      doses=double())

# Compute the metric for each country. Save the results into new_data.
country_list = unique(new_dt$country_region)

for (c in country_list){
  data = new_dt[country_region==c]  # Submatrix of the specific country
  metric = compute_metric(data$fratio) # Met demand metric value
  my_iso = data$iso2[1] # Find the country in the gdp table
  row = which(gdp$iso2c==my_iso)
  my_gdp = gdp$value[row] # Find its gdp value
  if (length(my_gdp>0)){ # Add it in the new_data table
    x = data.frame(data$country_region[1], data$continent_name[1], metric,
    my_gdp, data$population[1], data$doses_admin[1])
    names(x) = c("country", "continent", "metric", "gdp", "population","doses")
    new_data = rbind(new_data, x)}
}

setDT(new_data)

# Remove NAns (also Burundi which skews our results).
new2 = na.omit(new_data, "metric")
new2 = na.omit(new2, "gdp")
new2 = new2[country!='Burundi']

# Normalize our metric.
new2$normalized = min_max_norm(new2$metric)

# Add labels for countries with extreme values of our metric.
new2$repel_label = ifelse((new2$normalized>0.88 | new2$normalized<0.14),
                          new2$country, '')

# Fit a nonlinear least square model to our data.
m = nls(normalized ~ a * gdp / (b + gdp), data=new2)

# Fitted function.
nls_fun = function(x){coef(m)[1] * x / (coef(m)[2] + x)}

ggplot(new2, aes(x=gdp, y=normalized)) +
  stat_function(fun=nls_fun, col='steelblue') +
  geom_point(aes(color=normalized, size=population)) +
  geom_text_repel(size=5, aes(label=repel_label, color=normalized)) +
  theme_minimal() + scale_color_viridis_c(option='turbo') +
  labs(x="GDP (in billion dollars)", y="Met demand", title="Relation between
    met vaccination demand and Gross Domestic Product") +
  theme(legend.position=c(0.9, 0.3))
```

Ενώ το NMD δεν έχει κάποια φυσική ερμηνεία, είναι σημαντικό να κρατήσουμε στο μυαλό μας ότι χώρες με NMD κοντά στο 1 κατάφεραν να αποκτήσουν πρόσβαση στα εμβόλια που χρειάζονταν βάσει του πληθυσμού τους πολύ συντομότερα σε σχέση με



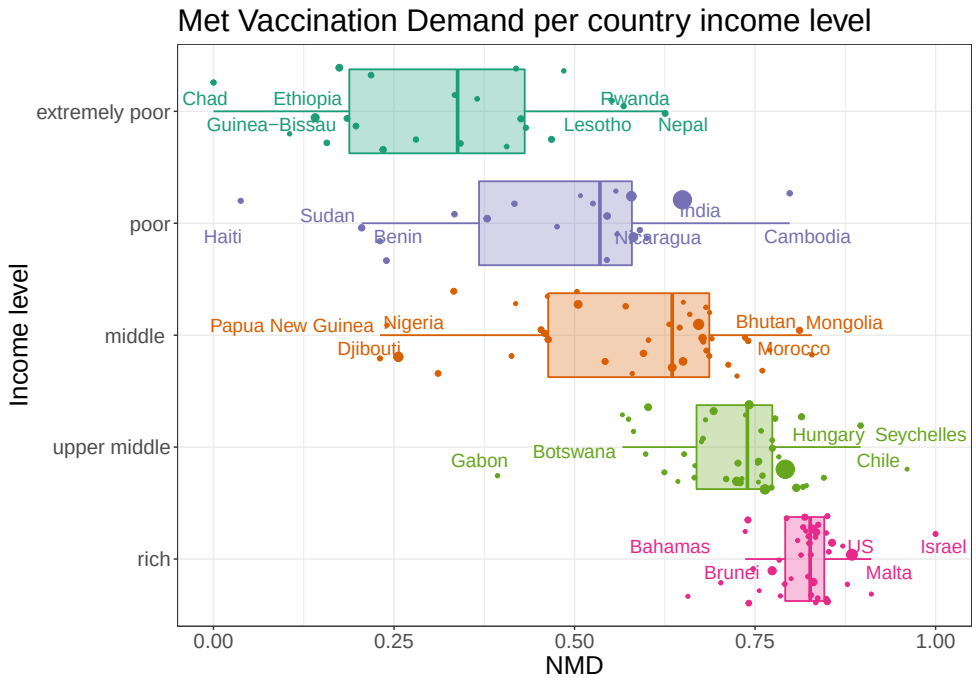
Σχήμα 4: Εξάρτηση μεταξύ ΑΕΠ και normalized met demand. Διακρίνουμε μια προφανή σχέση εξάρτησης, καθώς άυξηση του ΑΕΠ οδηγεί σε αύξηση της εμβολιαστικής δυνατότητας. Προσαρμόστηκε επιπλέον ένα μη γραμμικό μοντέλο στα δεδομένα μας (μπλε καμπύλη), υποδεικνύοντας μια σχέση της μορφής NMD = $\frac{0.8\text{gdp}}{1279+\text{gdp}}$. Στο σχήμα, επιπλέον, ονοματίζουμε τις καλύτερες και χειρότερες χώρες με βάση το NMD ανά ήπειρο.

τις υπόλοιπες, οι οποίες ίσως και να μην τα απέκτησαν και καθόλου. Αυτό μπορεί να οφείλεται στον πολύ μικρό πληθυσμό τους, άρα και στον μικρό αριθμό εμβολίων που χρειάστηκε να αγοράσουν, ή στην οικονομική τους δυνατότητα να πραγματοποιήσουν την απαιτούμενη αγορά. Στο Σχήμα 4 μπορούμε να ελέγξουμε και τις δύο αυτές υποθέσεις.

Σχετικά με τη δεύτερη, διαπιστώνουμε καταρχήν ότι το NMD αυξάνει καθώς μεγαλώνει το ΑΕΠ. Παρατηρούμε επίσης την ύπαρξη μιας τιμής κατωφλίου για το ΑΕΠ, περίπου για ΑΕΠ = 12.000 δις δολλάρια, πριν και μετά από την οποία αλλάζει εντελώς η συμπεριφορά του NMD. Πριν από αυτό το κατώφλι, ακόμα και μικρές αλλαγές στο ΑΕΠ έχουν θεαματική επίδραση στην εμβολιαστική δυνατότητα της χώρας, ουσιαστικά εντείνοντας τις εμβολιαστικές ανισότητες μεταξύ των φτωχών χωρών. Αντίθετα, πάνω από αυτό το κατώφλι, οι αλλαγές στο ΑΕΠ δε φαίνεται να έχουν καμία επίδραση στο NMD. Ουσιαστικά, από έναν πλούτο και πάνω, οι χώρες είναι σε θέση να καλύψουν τις ανάγκες τους με την ίδια αποτελεσματικότητα.

Το ίδιο σχήμα μας δίνει τη δυνατότητα να διακρίνουμε πιθανή σχέση εξάρτησης ανάμεσα στο NMD και τον πληθυσμό της εκάστοτε χώρας, καθώς το μέγεθος του σημείου κάθε χώρας έχει σχεδιαστεί ώστε να είναι ανάλογο του πληθυσμού της. Βάσει του σχήματος, δε φαίνεται να υπάρχει κάποιο μοτίβο σχετικά με τον πληθυσμό και το NMD.

Τέλος, στο Σχήμα 5 απεικονίζουμε με έναν ακόμη τρόπο τις εμβολιαστικές ανισότητες. Καταρχήν, δημιουργούμε πέντε κατηγορίες χωρών με βάση το ΑΕΠ τους, και στη συνέχεια τις απεικονίζουμε, ομαδοποιώντας τες πρώτα ανά εισοδηματική κατηγορία.



Σχήμα 5: Εμβολιαστική κάλυψη των χωρών για τα διάφορα επίπεδα εισοδήματος (κατηγοριοποιημένα βάσει του ΑΕΠ της εκάστοτε χώρας). Το μέγεθος του κάθε σημείου είναι ανάλογο του πληθυσμού της χώρας που αντιπροσωπεύει.

Γίνεται σαφές και από αυτό το σχήμα ότι η εμβολιαστική κάλυψη βελτιώνεται όσο αυξάνει η εισοδηματική ικανότητα της χώρας. Παράλληλα, όσο αυξάνει το εισόδημα, τόσο μειώνεται η διασπορά του NMD, πράγμα που συμβαδίζει με το κατώφλι και την όλη συμπεριφορά της μπλε καμπύλης που παρατηρήθηκε στο Σχήμα 4.

Ραβδόγραμμα βάσει ΑΕΠ (Σχήμα 5)

```
# Label each country according to its income level.
new2_sort_doses$country_lvl = ifelse(new2_sort_doses$gdp <= 1940.3075, 'poor',
  'not poor')
for (i in 1:dim(new2_sort_doses)[1]){
  if (new2_sort_doses[i, ]$gdp <= 1040.3075){
    new2_sort_doses[i, ]$country_lvl = 'extremely poor'}
  else if (new2_sort_doses[i, ]$gdp <= 1940.3075){
    new2_sort_doses[i, ]$country_lvl = 'poor'}
  else if (new2_sort_doses[i, ]$gdp <= 5121.0446){
    new2_sort_doses[i, ]$country_lvl = 'middle '}
  else if (new2_sort_doses[i, ]$gdp <= 14732.0415 ){
    new2_sort_doses[i, ]$country_lvl = 'upper middle'}
  else {new2_sort_doses[i, ]$country_lvl = 'rich'}
  }

# Add proper labels to the countries that stand out within each income level.
bottom_list_full = new2_sort_doses[order(-normalized), tail(country, 3),
by=country_lvl]$V1 # Bottom three countries for each income level

top_list_full = new2_sort_doses[order(-normalized),
head(country, 3), by=country_lvl]$V1 # Top three for each income level
```

```
full_list=c(bottom_list_full,top_list_full) # All together

# We add a column with the appropriate labels. The label is the country's
# name if it belongs to the previous list, and the empty string otherwise.
new2_sort_doses$repel_label=ifelse(new2_sort_doses$country%in%full_list,
new2_sort_doses$country, '')


ggplot(new2_sort_doses, aes(x=reorder(country_lvl, -normalized), y=normalized,
    fill=country_lvl)) +
  geom_boxplot(aes(color=country_lvl), outlier.shape=NA, alpha=0.5) +
  geom_jitter(aes(color=country_lvl, size=population)) +
  geom_text_repel(size=5, aes(color=country_lvl, label=repel_label)) +
  labs(title='Met Vaccination Demand per country income level',
    x='Income level', y='NMD') + theme_bw() +
  scale_fill_brewer(palette="Dark2") + scale_color_brewer(palette="Dark2") +
  coord_flip()
```

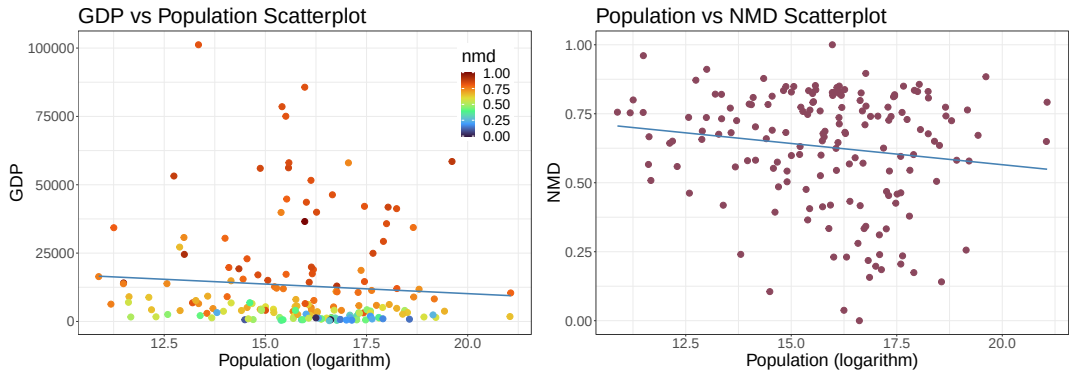
3.4 Εμβολιασμοί και Pareto

Αναφέραμε στην εισαγωγή της εργασίας κάποια στατιστικά για το ποιο ποσοστό των χωρών/ΑΕΠ/πληθυσμού απέκτησε το 80% των εμβολίων. Αυτά υπολογίστηκαν ως εξής:

Pareto

```
# Given a vector df, the pareto function estimates the
# percentage of the entries that contribute to the 100*a percent
# of the accumulated values in it.
pareto = function(df, variable, a=0.8){
  d = length(df)
  my_index = 1
  quant = 0
  while (quant<a){
    quant = sum(df[1:my_index]) / sum(df) # Current quantile
    my_index = my_index + 1 # Current index
  }
  return (my_index / d)
}

# Order the table in a descending order of doses administered.
new2_sort_doses = new2[order(-doses)]

# The world population percentage that acquired 80% of the vaccines.
pareto(new2_sort_doses$population)
[1] 0.502994

# The gdp percentage that acquired 80% of the vaccines.
pareto(new2_sort_doses$gdp)
[1] 0.6646707

# The country percentage that acquired 80% of the vaccines.
pareto(new2_sort_doses$doses)
[1] 0.08982036
```

Ουσιαστικά, τα παραπάνω νούμερα αποτελούν ακόμα μια ισχυρή ένδειξη ότι οι οικονομικοί λόγοι διαδραμάτισαν σημαντικότερο ρόλο κατά τη διάθεση των εμβολίων, από ότι



Σχήμα 6: Διαγράμματα σκέδασης για να διερευνηθεί πιθανή επιρροή του πληθυσμού, ως κρυφή μεταβλητή επίδρασης, στο φαινόμενο των ετεροβαρών εμβολιασμών που παρατηρήθηκε στην προηγούμενη παράγραφο. Και στα δύο διαγράμματα δεν παρατηρήθηκε σημαντική συσχέτιση του πληθυσμού ούτε με το ΑΕΠ, ούτε με την κύρια μεταβλητή ενδιαφέροντος, το NMD.

λόγοι που θα αφορούσαν την πληθυσμιακή κάλυψη που θα επιτύγχανε αυτή: Τα εμβόλια κάλυψαν μεγαλύτερο κομμάτι του παγκόσμιου ΑΕΠ (66.5%) από ότι του παγκόσμιου πληθυσμού (50.3%).

Διαγράμματα Σκέδασης (Σχήμα 6)

```
ggplot(new2, aes(x=log(population), y=gdp, col=nmd)) + geom_point(size=4) +
  geom_smooth(method='lm', se=F) + scale_color_viridis_c(option='turbo') +
  theme_bw() + labs(title = "GDP vs Population Scatterplot") +
  xlab('Population (logarithm)') + ylab('GDP')
```

4 Επίλογος

Αρχικά, επεξεργαστήκαμε το σύνολο δεδομένων μας σχολαστικά, ώστε να είναι σε θέση να εξυπηρετήσει την εξαγωγή μη τετριμμένων συμπερασμάτων για την πορεία του εμβολιασμού έναντι του covid-19 για τις διάφορες χώρες, ηπείρους κλπ.

Κατά τη διάρκεια κάποιων βασικών αναλύσεων διαπιστώσαμε την ύπαρξη μοτίβων που δε μπορούσαν να εξηγηθούν απλά μέσω γεωχωρικών σχέσεων. Εμπλουτίζοντας το σύνολο δεδομένων μας με τα οικονομικά στοιχεία των υπό μελέτη χωρών, η εικόνα άρχισε να γίνεται πιο ξεκάθαρη: Παρατηρήσαμε μια εμφανή συσχέτιση μεταξύ της εμβολιαστικής κάλυψης μιας χώρας και της οικονομικής της ευρωστίας.

Όπως είναι γνωστό, συσχέτιση δε συνεπάγεται απαραίτητα αιτιατότητα. Για παράδειγμα, αν η πραγματική αίτια της ετεροβαρούς εμβολιαστικής δυνατότητας που παρατηρήσαμε τύγχανε να είναι ισχυρά συσχετισμένη με το ΑΕΠ, τότε το ΑΕΠ θα ήταν και αυτό ισχυρά συσχετισμένο με την εμβολιαστική δυνατότητα, χωρίς όμως να αποτελεί απαραίτητα αιτία της.

Για αυτό το λόγο προσπαθήσαμε να εξαλείψουμε την κύρια μεταβλητή που θα μπορούσε να μας παραπλανήσει, δηλαδή τον πληθυσμό. Τόσο μέσω διαγραμμάτων (Σχήματα 4, 6), όσο και μέσω του επιχειρήματος τύπου Pareto, ο πληθυσμός δε φάνηκε να συσχετίζεται ούτε με το ΑΕΠ, ούτε με τα ποσοστά εμβολιασμού. Φυσικά, περαιτέρω μελέτη είναι απαραίτητη ώστε να εξαληφθούν και άλλες μεταβλητές ενδιαφέροντος εκτός του πληθυσμού.