PSZB17-210 - Seminar_4

Zoltan Kekecs

March 2, 2021

4. Ora - Adatexploracio

Az ora celja az adatexploracios modszerek elsajatitasa.

Package-ek betoltese

A kovetkezo package-ekre lesz szuksegunk

```
if (!require("gridExtra")) install.packages("gridExtra")
library(gridExtra) # for grid.arrange
if (!require("psych")) install.packages("psych")
library(psych) # for describe
if (!require("tidyverse")) install.packages("tidyverse")
library(tidyverse) # for dplyr and ggplot2
```

Adatok betoltese

Beolvassuk a WHO altal 2020.09.28-an feltoltott COVID-19 adatokat a read_csv() funkcioval, es elmentjuk egy COVID_data nevu objektumba. A **read_csv()** funkcio a tidyverse resze, es egybol tibble formatumban menti el az adatainkat.

```
COVID_data_raw <- read_csv("https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owi
```

Adatok attekintese

Mindig erdemes azzal kezdeni, hogy **megismerkedunk az adat** szerkezetevel es tartalmaval.

A **tibble objektum** meghivasaval kapthatunk nemi informaciot az adattabla szerkezeterol. Lathatjuk hany sor es hany oszlop van az adattablaban, es lathatjuk milyen class-ba tartoznak (chr, dbl ...)

```
COVID_data_raw
```

```
## # A tibble: 72,096 x 59
##    iso_code continent location date       total_cases new_cases new_cases_smoot~
##    <chr>    <chr>     <chr>    <date>           <dbl>     <dbl>            <dbl>
##  1 AFG      Asia      Afghani~ 2020-02-24           1         1           NA
##  2 AFG      Asia      Afghani~ 2020-02-25           1         0           NA
##  3 AFG      Asia      Afghani~ 2020-02-26           1         0           NA
##  4 AFG      Asia      Afghani~ 2020-02-27           1         0           NA
##  5 AFG      Asia      Afghani~ 2020-02-28           1         0           NA
##  6 AFG      Asia      Afghani~ 2020-02-29           1         0            0.143
##  7 AFG      Asia      Afghani~ 2020-03-01           1         0            0.143
##  8 AFG      Asia      Afghani~ 2020-03-02           1         0            0
##  9 AFG      Asia      Afghani~ 2020-03-03           2         1            0.143
```

```
## 10 AFG      Asia      Afghani~ 2020-03-04           4         2            0.429
## # ... with 72,086 more rows, and 52 more variables: total_deaths <dbl>,
## #   new_deaths <dbl>, new_deaths_smoothed <dbl>, total_cases_per_million <dbl>,
## #   new_cases_per_million <dbl>, new_cases_smoothed_per_million <dbl>,
## #   total_deaths_per_million <dbl>, new_deaths_per_million <dbl>,
## #   new_deaths_smoothed_per_million <dbl>, reproduction_rate <dbl>,
## #   icu_patients <lgl>, icu_patients_per_million <lgl>, hosp_patients <lgl>,
## #   hosp_patients_per_million <lgl>, weekly_icu_admissions <lgl>,
## #   weekly_icu_admissions_per_million <lgl>, weekly_hosp_admissions <lgl>,
## #   weekly_hosp_admissions_per_million <lgl>, new_tests <dbl>,
## #   total_tests <dbl>, total_tests_per_thousand <dbl>,
## #   new_tests_per_thousand <dbl>, new_tests_smoothed <dbl>,
## #   new_tests_smoothed_per_thousand <dbl>, positive_rate <dbl>,
## #   tests_per_case <dbl>, tests_units <chr>, total_vaccinations <lgl>,
## #   people_vaccinated <lgl>, people_fully_vaccinated <lgl>,
## #   new_vaccinations <lgl>, new_vaccinations_smoothed <lgl>,
## #   total_vaccinations_per_hundred <lgl>, people_vaccinated_per_hundred <lgl>,
## #   people_fully_vaccinated_per_hundred <lgl>,
## #   new_vaccinations_smoothed_per_million <lgl>, stringency_index <dbl>,
## #   population <dbl>, population_density <dbl>, median_age <dbl>,
## #   aged_65_older <dbl>, aged_70_older <dbl>, gdp_per_capita <dbl>,
## #   extreme_poverty <dbl>, cardiovasc_death_rate <dbl>,
## #   diabetes_prevalence <dbl>, female_smokers <dbl>, male_smokers <dbl>,
## #   handwashing_facilities <dbl>, hospital_beds_per_thousand <dbl>,
## #   life_expectancy <dbl>, human_development_index <dbl>
```

Leiro statisztikak

Ha az egyes valtozok **leiro statisztikaira** (descriptive statistics) vagyunk kivancsiak, kerhetjuk ezt a mar tanult modon.

Peldaul lekerhetjuk a valtozo alapveto legalacsonyabb es legmagasabb erteket, atlagat, medianjat, a kvartiliseket, es hogy hany hianyzo adat van (ha van) a **summary()** funkcioval (miutan a select funkcioval kivalasztottuk, melyik valtozora vagyunk kivancsiak)

```
COVID_data_raw %>%
  select(total_cases) %>%
  summary()
```

```
##   total_cases
##  Min.   :        1
##  1st Qu.:      706
##  Median :     7577
##  Mean   :   610304
##  3rd Qu.:    82603
##  Max.   :114442646
##  NA's   :964
```

Vagy megkapthatjuk ugyanezt az osszes valtozora, ha ugyanezt az egesz adattablara futtatjuk le. Persze a karakter osztalyba tartozo valtozoknal mindezeknek a leiro statisztikaknak nincs ertelme, ott csak a class informaciot kaptjuk az output-ban.

```
COVID_data_raw %>%
  summary()
```

Az exploracio megmutatta hogy van nehany irrealisztikus adat. Ennek az az oka hogy kontinensekre es regiokra lebontott osszefoglalo adatokat is tartalmaz a tablazat. Ezeket ugy tudjuk legkonnyebben kivenni

hogy kivesszuk azokat a sorokat, ahol a continent valtozo NA erteket vesz fel. (Vedd eszre hogy ezt "!" es az is.na() funkciok kombinaciojaval oldjuk meg. A ! jelentese "NOT".)

```
COVID_data <- COVID_data_raw %>%
  filter(!is.na(continent))


COVID_data %>%
  select(total_cases) %>%
  summary()
```

```
##   total_cases
##  Min.   :       1
##  1st Qu.:     628
##  Median :    6565
##  Mean   :  203025
##  3rd Qu.:   67334
##  Max.   :28664481
##  NA's   :950
```

```
COVID_data_raw %>%
  select(total_cases) %>%
  summary()
```

```
##   total_cases
##  Min.   :        1
##  1st Qu.:      706
##  Median :     7577
##  Mean   :   610304
##  3rd Qu.:    82603
##  Max.   :114442646
##  NA's   :964
```

Gyakorlas

- Hany regisztralt eset volt osszesen Magyarorszagon a tegnapi napig (*total_cases*)?
- Mi volt a legmagasabb uj eset-szam Magyarorszagon (*new_cases*)?

Megtobb leiro statisztika

A **Psych** package segitsegevel a **describe()** funkcio megtobb hasznos informaciot adhat. Ez a funkcio elsosorban szam-valtozok leirasara szolgal, es karakter tipusu kategorikus valtozok eseten sok warning message-et ad, ezert erdemes a funciot csak a szam-valtozokra lefuttatni (ezt alabb a select() funkcioval erem el.)

```
COVID_data %>%
  select(-date, -iso_code, -continent, -location, -contains("tests"), -positive_rate) %>%
  describe()
```

```
##                                vars     n        mean          sd
## total_cases                       1 67566   203024.59   1139284.95
## new_cases                         2 67559     1681.76      9159.12
## new_cases_smoothed                3 66611     1689.43      9008.48
## total_deaths                      4 58643     6075.65     25639.29
```

```
## new_deaths                                5 58642       43.29       185.58
## new_deaths_smoothed                       6 66611       37.76       163.03
## total_cases_per_million                   7 67566     7153.58     14314.81
## new_cases_per_million                     8 67559       64.78       168.79
## new_cases_smoothed_per_million            9 66611       64.77       141.33
## total_deaths_per_million                 10 58643      168.05       306.43
## new_deaths_per_million                   11 58642        1.36         3.93
## new_deaths_smoothed_per_million          12 66611        1.19         2.80
## reproduction_rate                        13 57602        1.02         0.35
## icu_patients                             14     0         NaN           NA
## icu_patients_per_million                 15     0         NaN           NA
## hosp_patients                            16     0         NaN           NA
## hosp_patients_per_million                17     0         NaN           NA
## weekly_icu_admissions                    18     0         NaN           NA
## weekly_icu_admissions_per_million        19     0         NaN           NA
## weekly_hosp_admissions                   20     0         NaN           NA
## weekly_hosp_admissions_per_million       21     0         NaN           NA
## total_vaccinations                       22     0         NaN           NA
## people_vaccinated                        23     0         NaN           NA
## people_fully_vaccinated                  24     0         NaN           NA
## new_vaccinations                         25     0         NaN           NA
## new_vaccinations_smoothed                26     0         NaN           NA
## total_vaccinations_per_hundred           27     0         NaN           NA
## people_vaccinated_per_hundred            28     0         NaN           NA
## people_fully_vaccinated_per_hundred      29     0         NaN           NA
## new_vaccinations_smoothed_per_million    30     0         NaN           NA
## stringency_index                         31 61934       58.92        22.15
## population                               32 68516 43599575.60 156432943.82
## population_density                       33 66980      328.83      1596.52
## median_age                               34 65184       30.57         9.15
## aged_65_older                            35 64428        8.81         6.27
## aged_70_older                            36 64814        5.59         4.28
## gdp_per_capita                           37 65362    19156.64     19740.78
## extreme_poverty                          38 44519       13.28        20.00
## cardiovasc_death_rate                    39 65973      257.23       118.73
## diabetes_prevalence                      40 66800        7.79         3.94
## female_smokers                           41 51834       10.59        10.45
## male_smokers                             42 51117       32.63        13.51
## handwashing_facilities                   43 32902       50.91        31.95
## hospital_beds_per_thousand               44 60413        3.04         2.48
## life_expectancy                          45 68103       73.14         7.57
## human_development_index                  46 65951        0.73         0.15
##                                          median     trimmed         mad
## total_cases                            6565.00    35972.93     9699.17
## new_cases                                50.00      289.35       74.13
## new_cases_smoothed                       60.71      306.87       89.80
## total_deaths                            202.00     1118.22      292.07
## new_deaths                                1.00        7.54        1.48
## new_deaths_smoothed                       0.86        6.03        1.27
## total_cases_per_million                 996.97     3495.34     1463.16
## new_cases_per_million                     5.55       26.75        8.23
## new_cases_smoothed_per_million            7.62       29.95       11.27
## total_deaths_per_million                 28.54       89.88       41.08
## new_deaths_per_million                    0.09        0.54        0.13
```

```
## new_deaths_smoothed_per_million             0.10        0.50        0.15
## reproduction_rate                           1.03        1.02        0.24
## icu_patients                                  NA         NaN          NA
## icu_patients_per_million                      NA         NaN          NA
## hosp_patients                                 NA         NaN          NA
## hosp_patients_per_million                     NA         NaN          NA
## weekly_icu_admissions                         NA         NaN          NA
## weekly_icu_admissions_per_million             NA         NaN          NA
## weekly_hosp_admissions                        NA         NaN          NA
## weekly_hosp_admissions_per_million            NA         NaN          NA
## total_vaccinations                            NA         NaN          NA
## people_vaccinated                             NA         NaN          NA
## people_fully_vaccinated                       NA         NaN          NA
## new_vaccinations                              NA         NaN          NA
## new_vaccinations_smoothed                     NA         NaN          NA
## total_vaccinations_per_hundred                NA         NaN          NA
## people_vaccinated_per_hundred                 NA         NaN          NA
## people_fully_vaccinated_per_hundred           NA         NaN          NA
## new_vaccinations_smoothed_per_million         NA         NaN          NA
## stringency_index                           61.11       60.24       23.34
## population                            9660350.00 16582763.24 13023827.05
## population_density                         83.48      113.57       88.35
## median_age                                 29.70       30.44       12.75
## aged_65_older                               6.38        8.14        5.06
## aged_70_older                               3.86        5.04        3.12
## gdp_per_capita                          12951.84    15810.25    14552.01
## extreme_poverty                             2.00        8.89        2.67
## cardiovasc_death_rate                     242.65      246.67      121.48
## diabetes_prevalence                         7.11        7.42        3.45
## female_smokers                              6.30        9.21        7.86
## male_smokers                               31.40       32.00       14.53
## handwashing_facilities                     49.84       51.14       45.66
## hospital_beds_per_thousand                  2.40        2.67        1.93
## life_expectancy                            74.62       73.69        7.06
## human_development_index                     0.75        0.74        0.17
##                                              min         max        range  skew
## total_cases                                 1.00 2.866448e+07 2.866448e+07 14.41
## new_cases                                -46076.00 2.997860e+05 3.458620e+05 15.23
## new_cases_smoothed                       -1121.71 2.497266e+05 2.508483e+05 15.12
## total_deaths                                1.00 5.146570e+05 5.146560e+05  9.49
## new_deaths                              -1918.00 4.398000e+03 6.316000e+03 10.01
## new_deaths_smoothed                      -232.14 3.354140e+03 3.586290e+03  9.64
## total_cases_per_million                     0.00 1.409306e+05 1.409306e+05  3.30
## new_cases_per_million                   -2153.44 8.652660e+03 1.080609e+04  9.15
## new_cases_smoothed_per_million           -276.82 2.648770e+03 2.925600e+03  4.07
## total_deaths_per_million                    0.00 2.180450e+03 2.180450e+03  2.61
## new_deaths_per_million                    -76.44 2.183300e+02 2.947700e+02 12.66
## new_deaths_smoothed_per_million           -10.92 6.314000e+01 7.406000e+01  4.79
## reproduction_rate                           0.00 6.740000e+00 6.740000e+00  0.96
## icu_patients                                 Inf        -Inf        -Inf    NA
## icu_patients_per_million                     Inf        -Inf        -Inf    NA
## hosp_patients                                Inf        -Inf        -Inf    NA
## hosp_patients_per_million                    Inf        -Inf        -Inf    NA
## weekly_icu_admissions                        Inf        -Inf        -Inf    NA
```

```
## weekly_icu_admissions_per_million           Inf         -Inf         -Inf    NA
## weekly_hosp_admissions                      Inf         -Inf         -Inf    NA
## weekly_hosp_admissions_per_million          Inf         -Inf         -Inf    NA
## total_vaccinations                          Inf         -Inf         -Inf    NA
## people_vaccinated                           Inf         -Inf         -Inf    NA
## people_fully_vaccinated                     Inf         -Inf         -Inf    NA
## new_vaccinations                            Inf         -Inf         -Inf    NA
## new_vaccinations_smoothed                   Inf         -Inf         -Inf    NA
## total_vaccinations_per_hundred              Inf         -Inf         -Inf    NA
## people_vaccinated_per_hundred               Inf         -Inf         -Inf    NA
## people_fully_vaccinated_per_hundred         Inf         -Inf         -Inf    NA
## new_vaccinations_smoothed_per_million       Inf         -Inf         -Inf    NA
## stringency_index                           0.00 1.000000e+02 1.000000e+02 -0.48
## population                               809.00 1.439324e+09 1.439323e+09  7.88
## population_density                         0.14 2.054677e+04 2.054663e+04 10.35
## median_age                                15.10 4.820000e+01 3.310000e+01  0.10
## aged_65_older                              1.14 2.705000e+01 2.591000e+01  0.78
## aged_70_older                              0.53 1.849000e+01 1.797000e+01  0.91
## gdp_per_capita                           661.24 1.169356e+05 1.162744e+05  1.81
## extreme_poverty                            0.10 7.760000e+01 7.750000e+01  1.65
## cardiovasc_death_rate                     79.37 7.244200e+02 6.450500e+02  0.87
## diabetes_prevalence                        0.99 3.053000e+01 2.954000e+01  1.16
## female_smokers                             0.10 4.400000e+01 4.390000e+01  0.95
## male_smokers                               7.70 7.810000e+01 7.040000e+01  0.54
## handwashing_facilities                     1.19 9.900000e+01 9.781000e+01 -0.05
## hospital_beds_per_thousand                 0.10 1.380000e+01 1.370000e+01  1.75
## life_expectancy                           53.28 8.675000e+01 3.347000e+01 -0.59
## human_development_index                    0.39 9.600000e-01 5.600000e-01 -0.38
##                                         kurtosis        se
## total_cases                               271.15   4382.97
## new_cases                                 314.39     35.24
## new_cases_smoothed                        303.00     34.90
## total_deaths                              123.39    105.88
## new_deaths                                145.38      0.77
## new_deaths_smoothed                       130.45      0.63
## total_cases_per_million                    13.84     55.07
## new_cases_per_million                     250.56      0.65
## new_cases_smoothed_per_million             23.55      0.55
## total_deaths_per_million                    7.15      1.27
## new_deaths_per_million                    400.14      0.02
## new_deaths_smoothed_per_million            38.91      0.01
## reproduction_rate                          11.41      0.00
## icu_patients                                  NA        NA
## icu_patients_per_million                      NA        NA
## hosp_patients                                 NA        NA
## hosp_patients_per_million                     NA        NA
## weekly_icu_admissions                         NA        NA
## weekly_icu_admissions_per_million             NA        NA
## weekly_hosp_admissions                        NA        NA
## weekly_hosp_admissions_per_million            NA        NA
## total_vaccinations                            NA        NA
## people_vaccinated                             NA        NA
## people_fully_vaccinated                       NA        NA
## new_vaccinations                              NA        NA
```

```
## new_vaccinations_smoothed                    NA        NA
## total_vaccinations_per_hundred               NA        NA
## people_vaccinated_per_hundred                NA        NA
## people_fully_vaccinated_per_hundred          NA        NA
## new_vaccinations_smoothed_per_million        NA        NA
## stringency_index                          -0.42      0.09
## population                                65.44 597629.76
## population_density                       115.04      6.17
## median_age                                -1.25      0.04
## aged_65_older                             -0.69      0.02
## aged_70_older                             -0.36      0.02
## gdp_per_capita                             4.14     77.22
## extreme_poverty                            1.68      0.09
## cardiovasc_death_rate                      0.77      0.46
## diabetes_prevalence                        2.42      0.02
## female_smokers                            -0.17      0.05
## male_smokers                               0.26      0.06
## handwashing_facilities                    -1.50      0.18
## hospital_beds_per_thousand                 3.97      0.01
## life_expectancy                           -0.40      0.03
## human_development_index                   -0.89      0.00
```

Gyakorlas

- Mi az egy millio fore eso uj esetek (*new_cases_per_million*) ferdesegi mutatoja (skew/skewness)?
- Hany valid (nem NA) adat szerepel az adatbazisban az egy fore eso gdp-rol (*gdp_per_capita*)?

Faktorok

Nehany karaktervaltozonak csak **korlatozott mennyisegu eleme** lehet, mint peldaul a continent (North America, Asia, Africa, Europe, South America, Oceania). Ezeket megjelolhetjuk faktor (factor) osztalyu valtozokent, es akkor az R tobb informaciot fog adni rola.

```
COVID_data <- COVID_data %>%
              mutate(continent = factor(continent),
                     location = factor(location))

levels(COVID_data$continent)
```

```
## [1] "Africa"        "Asia"          "Europe"        "North America"
## [5] "Oceania"       "South America"
```

```
table(COVID_data$continent)
```

```
##
##        Africa          Asia        Europe North America       Oceania
##         18950         17244         17439          8444          2035
## South America
##          4404
```

```
COVID_data <- COVID_data %>%
              mutate(continent = factor(continent))
```

A **levels()** funkcio megmutatja mik a faktorunk szintjei, de lathato ez akkor is ha csak meghivjuk a valtozot magat.

A **table()** funkcio pedig tablazatot keszit arrol, hogy az egyes csoportokban hany megfigyeles talalhato

Amikor kilistazzuk a faktor valtozot, akkor is kiirja az R a lista aljara, hogy milyen faktorszintek vannak.

```
levels(COVID_data$continent)

table(COVID_data$continent)

COVID_data$continent
```

Alabb csinalunk egy COVID_data_latest valtozot, amivel csak az adatbazisban szereplo legutolso napra vonatkozo adatok szerepelnek, hogy kisebb legyen az adattabla amivel dolgozunk.

```
COVID_data_latest = COVID_data %>%
  filter(date == max(COVID_data$date))
```

Miutan egy valtozot faktorkent azonositottunk, bizonyos funkciok kepesek felhasznalni ezt az informaciot. Peldaul a summary() function igy mar a fenti **summary()** funkcio is kiadja az **egyes faktorszintekrol** hogy hany megfigyeles tartozik az egyes kategoriakba (faktorszintekbe).

```
COVID_data %>%
 mutate(continent = as.character(continent)) %>%
  select(continent) %>%
  summary()
```

```
##   continent
##  Length:68516
##  Class :character
##  Mode  :character
```

```
# continent is already recognized as a factor variable
COVID_data_latest %>%
  select(continent) %>%
  summary()
```

```
##          continent
##  Africa       :54
##  Asia         :46
##  Europe       :47
##  North America:23
##  Oceania      : 9
##  South America:12
```

Van, hogy szeretnenk **kizarni** bizonyos **faktorszinteket** az elemzesbol. Pl. ha valamelyik faktor szintbol nagyon keves megfigyeles van, mondjuk Oceaniat, mondjuk mert ugy gondoljuk hogy az tulsagosan "elszigetelt" a vilag tobbi reszetol, oket lehet hogy szeretnenk kizarni a kesobbi elemzesekbol hogy egyszerusitsuk az eredmenyeink ertelmezeset. Ezt a mar korabban tanult **filter()** funkcio segitsegevel konnyeden megtehetjuk, azonban arra figyelnunk kell, hogy az R megjegyzi a faktorszinteket, es azt azt kovetoen is a **valtozohoz rendelve tartja**. A **faktorszintek meg akkor is megmaradnak ha nem marad egy megfigyeles sem** az adott faktorszinten az adattablaban.

```
COVID_data_latest %>%
  filter(continent != "Oceania") %>%
  select(total_cases, continent) %>%
  summary()
```

```
##   total_cases               continent
##  Min.   :      27   Africa       :54
##  1st Qu.:    8595   Asia         :46
##  Median :   79002   Europe       :47
##  Mean   :  632095   North America:23
##  3rd Qu.:  288267   Oceania      : 0
##  Max.   :28664481   South America:12
##  NA's   :1
```

Igy ezeket a szinteket ejthetjuk a **droplevels()** funkcioval.

```
COVID_data_latest_noOceania = COVID_data_latest %>%
  filter(continent != "Oceania") %>%
  mutate(continent = droplevels(continent))



COVID_data_latest_noOceania %>%
  select(continent) %>%
  summary()
```

```
##          continent
##  Africa       :54
##  Asia         :46
##  Europe       :47
##  North America:23
##  South America:12
```

Faktorszintek egymashoz viszonyitott erteke

Legtobbszor a faktorszintek kozott nincs "ertekbeli" kulonbseg, egyszeruen csoportnevekrol van szo, de neha egy meghatarozott relacio van kozottuk, pl. a legmagasabb iskolai vegzettsge lehet vegzettsge nelkuli < altalanos iskolai < kozepiskolai < felsofoku . . . Ittfaktorszinteknek van egy meghatarozott hierarchiaja, vagy sorrendje. Ilyen valtozo nincs ebben az adatbazisban, de konnyeden csinalhatunk ilyen faktor valtozot.

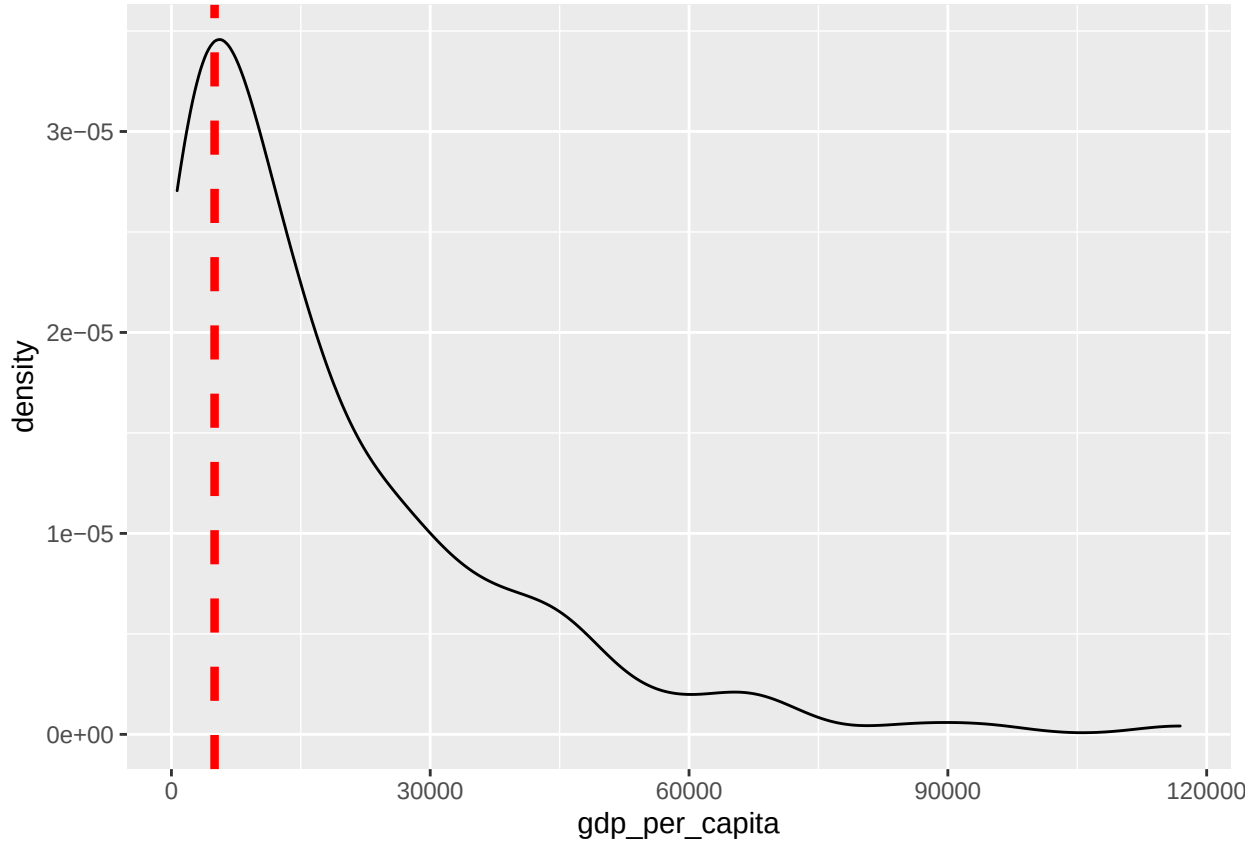
Ehhez arra van szuksegunk, hogy egy **numerikus valtozot alakitsunk faktorra**, pl. elkepzelheto hogy ossze akarjuk hasonlitani azokat az orszagokat ahol 5000 alatti a gdp_per_capita azokkal akinel e feletti, hogy hogyan kulonboznek a COVID adatok.

```
COVID_data_latest %>%
  select(gdp_per_capita, continent) %>%
  drop_na() %>%
  group_by(continent) %>%
  summarize(mean_gdp = mean(gdp_per_capita))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 2
##   continent     mean_gdp
##   <fct>            <dbl>
## 1 Africa           5444.
## 2 Asia            22185.
## 3 Europe          33361.
## 4 North America   17126.
## 5 Oceania         12392.
## 6 South America   13841.
```

```
COVID_data_latest %>%
  select(gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
  aes(x = gdp_per_capita) +
  geom_density() +
  geom_vline(xintercept = 5000, linetype="dashed",
                color = "red", size=1.5)
```



Ilyenkor hasznalhatjuk a **mutate()** es **case_when()** funkciok kombinaciojat hogy csinaljunk egy uj valtozot. Ebbe a kodba beleepitettem a **factor()** funkciot is, hogy azonnal meghatarozzuk, hogy ez az uj valtozo egy faktor, es nem egy egyszeru karaktervektor. A factor() funkcio nelkul is lefut a kod, de akkor meg kellene egy kulon sor ahol megadjuk hogy ez egy faktorvaltozo.

```
COVID_data = COVID_data %>%
  mutate(gdp_per_capita_kat = factor(
                                    case_when(gdp_per_capita < 5000 ~ "small",
                                              gdp_per_capita >= 5000 & gdp_per_capita < 10000 ~ "medi
                                              gdp_per_capita > 10000 ~ "large")))
levels(COVID_data$gdp_per_capita_kat)
```

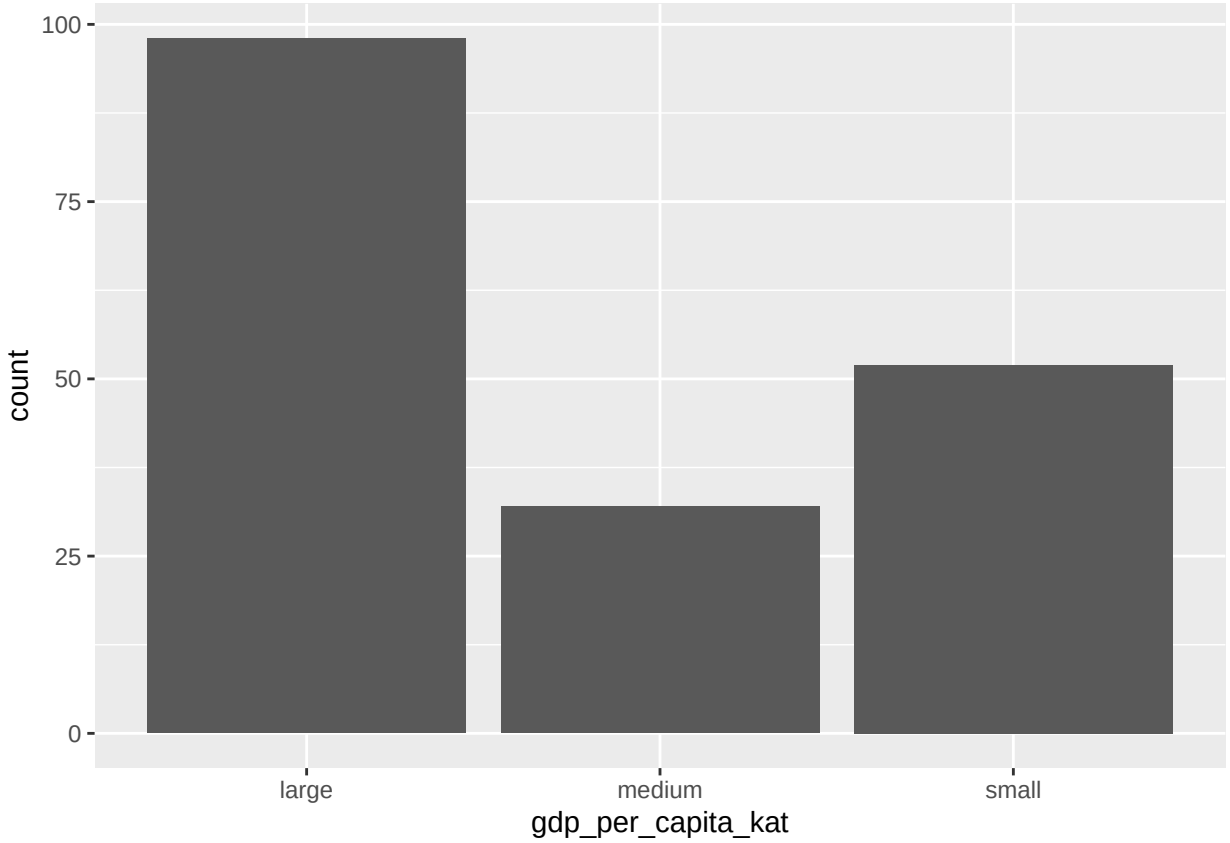
```
## [1] "large"  "medium" "small"
```

```
# ugyanez a COVID_data_latest -al

COVID_data_latest = COVID_data_latest %>%
  mutate(gdp_per_capita_kat = factor(
```

```
                                    case_when(gdp_per_capita < 5000 ~ "small",
                                              gdp_per_capita >= 5000 & gdp_per_capita < 10000 ~ "medi
                                              gdp_per_capita > 10000 ~ "large")))
```

Amikor abrat rajzolunk erreol a valtozorol, lathatjuk hogy a faktorszintek sorrendje “large”, “medium”, es “small” az abran.

```
COVID_data_latest %>%
  select(gdp_per_capita_kat) %>%
  drop_na() %>%
  ggplot() +
  aes(x = gdp_per_capita_kat) +
  geom_bar()
```

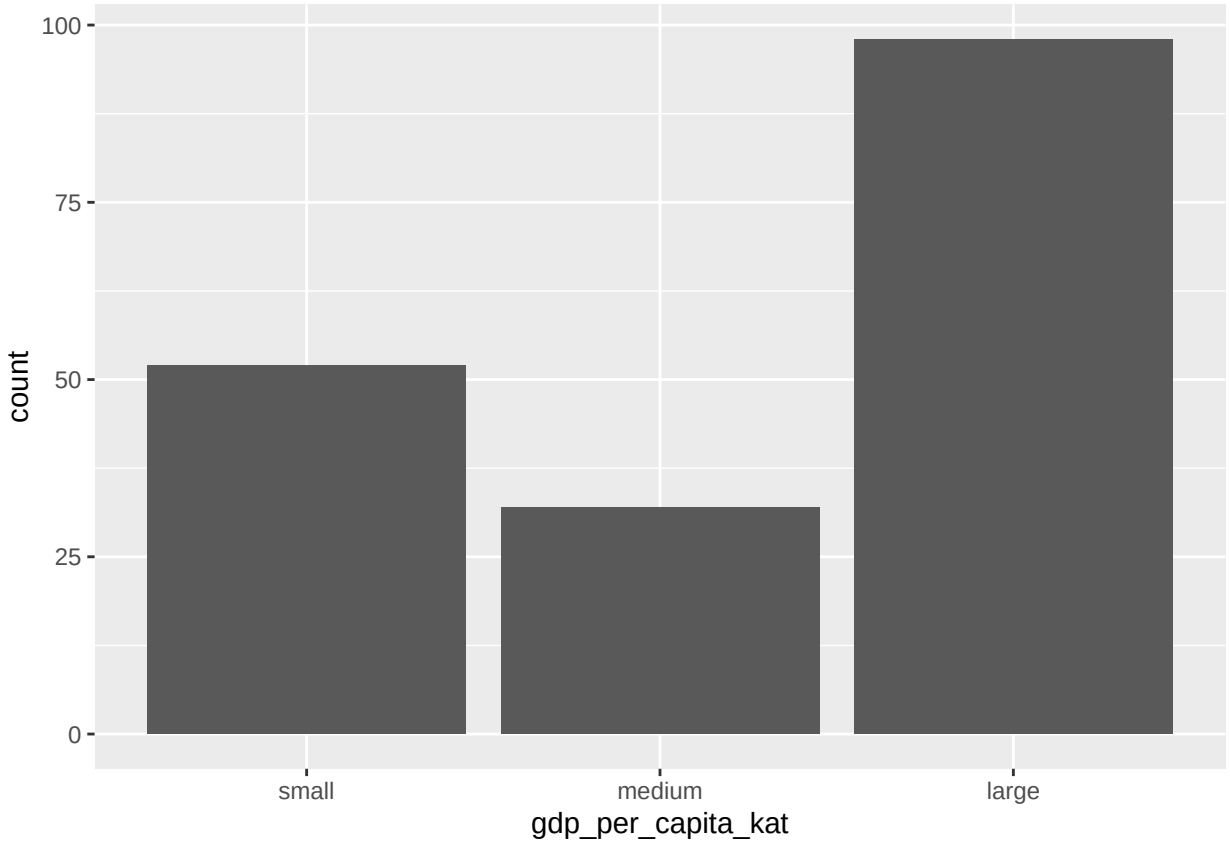


Ez nem feltetlenul legikus abrazolas, hiszen altalaban a kisebbtol a nagyobbig szoktunk haladni balrol jobbra. De az R nem tudja mit jelentenek a faktorszintek nevei. A faktorszintek sorrendjenek meghatarozasanal ezert alapertelmezett modon **abc sorrendet hasznal**.

Specifikalhatjuk maskepp is a faktroszintek sorrendjet a factor funkcioban a **levels = c()** parameteren keresztul egy vektorban megadva.

```
COVID_data_latest = COVID_data_latest %>%
mutate(gdp_per_capita_kat = factor(gdp_per_capita_kat, levels = c(
                                        "small",
                                         "medium",
                                         "large")))

COVID_data_latest %>%
```

```
  select(gdp_per_capita_kat) %>%
  drop_na() %>%
  ggplot() +
  aes(x = gdp_per_capita_kat) +
  geom_bar()
```



Attol meg hogy megadjuk a levels-el a faktorszintek listazasi sorrendjet, az R meg mindig egyenrangukent kezeli a faktorszinteket. Ha azt szeretnenk ha az R ugy ertekelne hogy a faktorszintek valamilyen hierarchikus sorrendben van, vagyis **ordinalis valtozokent**, akkor ezt a factor() funkcion belul az **ordered = T** parameter beallitasaval tehetjuk meg.

Ha ezt teszuk, a faktor valtozo kilistazasakor relacio-jelek kerulnek a faktorszintek koze, es mas funkciok is fel tudjak majd hasznalni ezt az informaciot.

```
COVID_data_latest = COVID_data_latest %>%
mutate(gdp_per_capita_kat = factor(gdp_per_capita_kat, ordered = T, levels = c(
                                                "small",
                                                 "medium",
                                                 "large")))
COVID_data_latest$gdp_per_capita_kat
```

```
##   [1] small  large  large  <NA>   medium large  large  medium large  large
##  [11] large  large  large  small  large  large  large  medium small  medium
##  [21] medium large  large  large  large  large  small  small  small  small
##  [31] large  medium small  small  large  large  large  small  small  large
##  [41] small  large  <NA>   large  large  small  large  small  medium large
##  [51] large  large  medium large  small  large  medium small  medium large
```

```
##  [61] large  large  small  medium large  small  large  large  medium small
##  [71] small  medium small  small  large  large  medium large  large  large
##  [81] large  <NA>   large  large  medium large  medium large  small  medium
##  [91] large  small  medium large  large  small  small  large  <NA>   large
## [101] large  small  small  large  large  small  large  small  small  large
## [111] large  small  medium <NA>   large  large  medium small  medium medium
## [121] small  large  large  medium small  medium large  large  large  medium
## [131] small  large  small  medium large  medium large  large  large  large
## [141] large  small  large  large  large  medium large  small  large  small
## [151] large  large  small  large  large  large  small  <NA>   large  large
## [161] small  large  large  small  large  large  large  <NA>   <NA>   small
## [171] small  large  medium small  large  large  large  small  medium large
## [181] large  large  large  medium small  <NA>   large  medium small  small
## [191] small
## Levels: small < medium < large
```

Kategorikus valtozo ujrakodolasa

Egy masik funkcio amivel manipulalhatjuk a faktorszinteket, a **recode()**. Ha kategorikus valtozokat szeretnenk atkodolni, mondjuk ha szeretnenk a deli felteket az eszaki feltekevel osszehasonlitani, ezt a kovetkezokeppen tehetjuk:

```
COVID_data = COVID_data %>%
  mutate(continent_south_north = factor(recode(continent,
                                               "Oceania" = "South",
                                               "South America" = "South",
                                               "Africa" = "South",
                                               "Asia" = "North",
                                               "Europe" = "North",
                                               "North America" = "North")))

levels(COVID_data$continent_south_north)
```

```
## [1] "South" "North"
```

```
COVID_data_latest = COVID_data_latest %>%
  mutate(continent_south_north = factor(recode(continent,
                                               "Oceania" = "South",
                                               "South America" = "South",
                                               "Africa" = "South",
                                               "Asia" = "North",
                                               "Europe" = "North",
                                               "North America" = "North")))
```

Gyakorlas

- szurd az adatokat ugy hogy csak a tegnapi adatokkal dolgozzunk.
- csinalj egy uj kategorikus valtozot (nevezzuk ezt *new_cases_per_million_kat*-nak) a mutate() funkcio hasznalataval amiben azok az orszagok ahol a *new_cases_per_million* valtozo 20 alatt van “small”, ahol 20 vagy a felett van “large” kategoriaba keruljenek.
- figyelj oda hogy faktorkent jelold meg ezt az uj valtozot (Ezt lehet az elozo lepesben a mutate() funkcion belul, vagy egy kulon lepesben, de mindenkeppen a factor() vagy az as.factor() funkciokat erdemes hozza hasznalni)
- mentsd el ezt a valtozot az eredeti adatobjektumban ugy hogy kesobb is lehessen vele dolgozni

- keszits egy tablazatot arrol, hogy hanyan esnek a *new_cases_per_million_kat* egyes kategoriaiba.
- Add meg a faktorszintek helyes sorrendjet: small, large (Ird felul a *new_cases_per_million_kat* korabbi valtozatat ezzel a valtozattal ahol a szintek mar helyes sorrendben vannak, vagy ezt a sorrendezest is bele vonhatod az eredeti funkcioba, amivel a valtozot generaltad)
- Ellenorizd, hogy valoban helyes sorrendben szerepelnek-e a faktor szintjei.

Exploracio vizualizacion keresztul

Az egyes valtozok vizualizacioja es a leiro statisztikak atvizsgalasa elengedhetetlen hogy azonositsuk az esetleges adatbeviteli **hibakat es egyeb nemvart furcsasagokat** az adataink kozott.

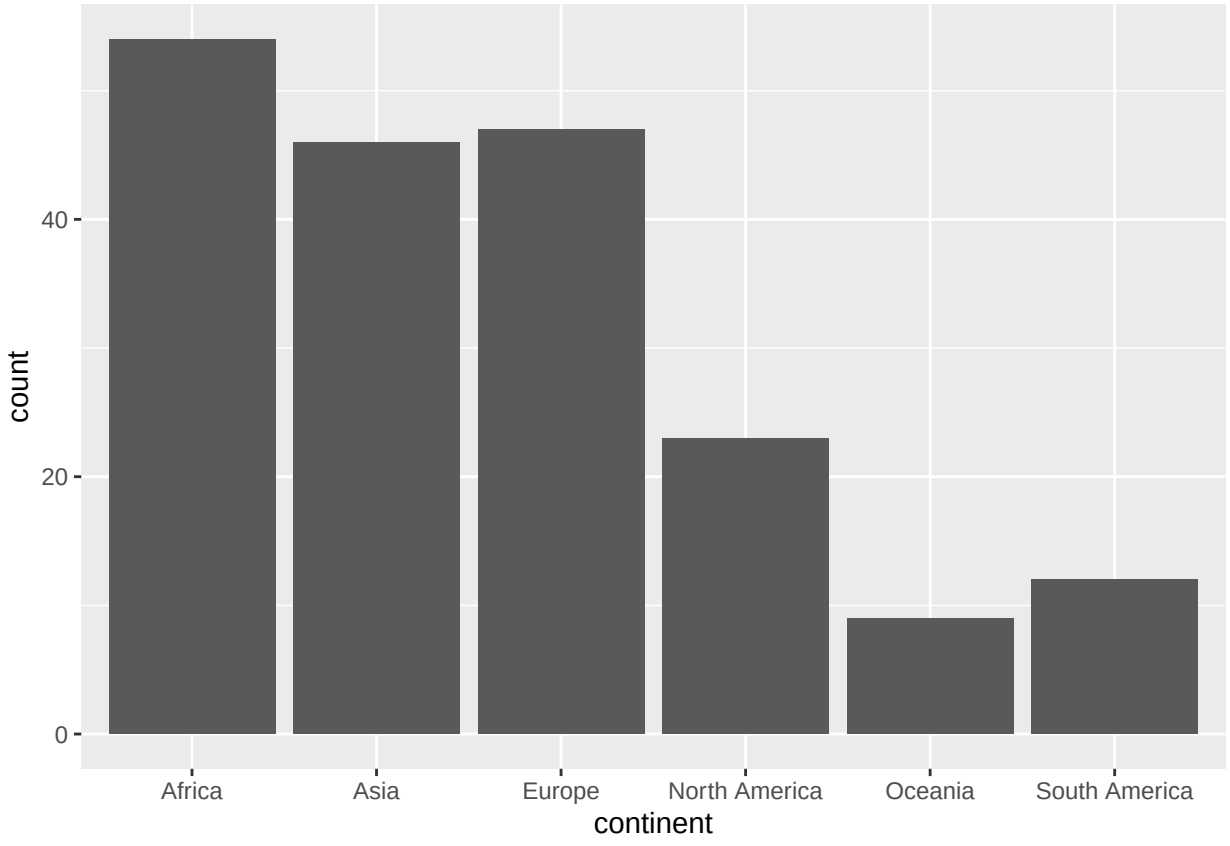
MINDING ellenorizd az adataidat ezekkel a modszerekkel mielott komolyabb adatelemzesbe kezdesz, hogy meggyozodj rola, hogy az adatok tisztak es megfelenek az elvarasaidnak.

Egyes valtozok vizualizacioja

Az egyes valtozok peldaul **abrak** (plot) segitsegevel megvizsgalhatok.

A **kategorikus** valtozokat gyakran oszlopdiagrammal (**geom_bar**) abrazoljuk,

```
COVID_data_latest %>%
ggplot() +
  aes(x = continent) +
  geom_bar()
```

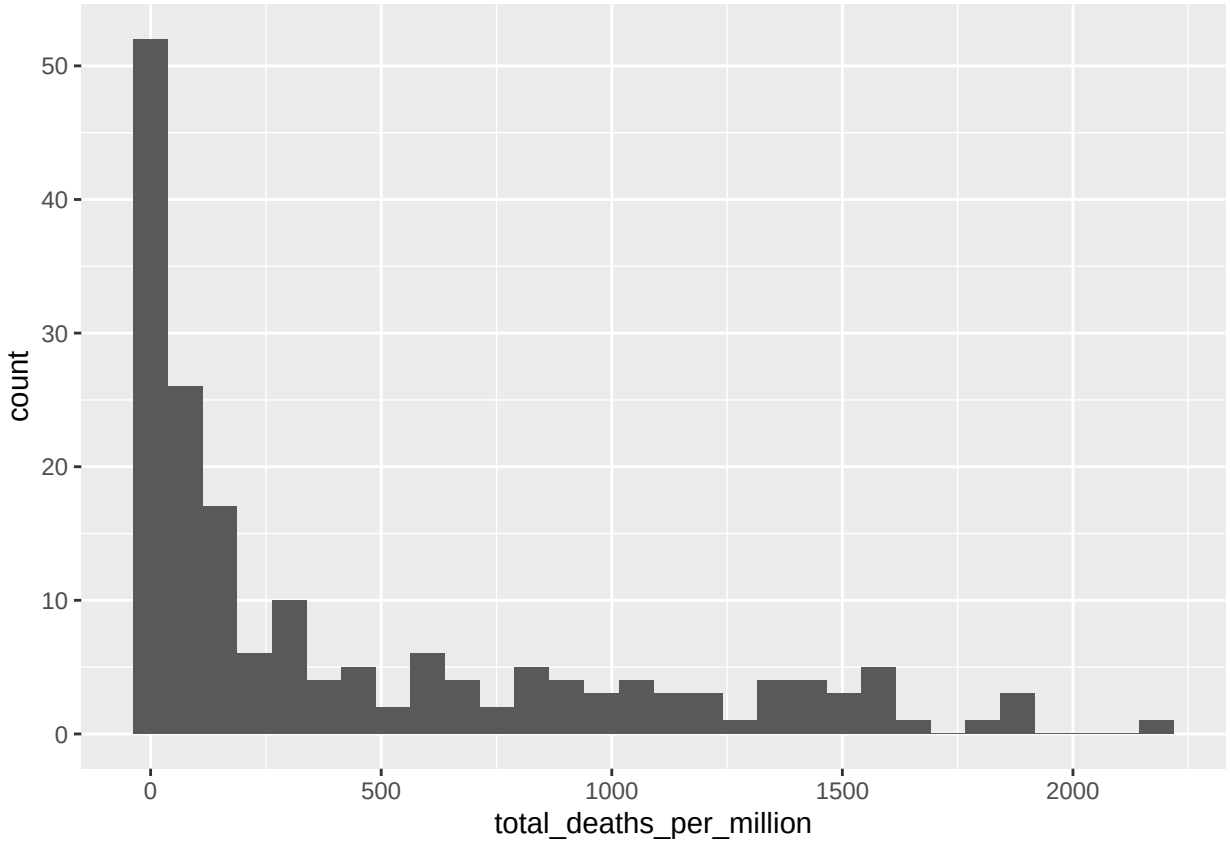



```
COVID_data_latest %>%
ggplot() +
```

```
  aes(x = total_deaths_per_million) +
  geom_histogram()
```

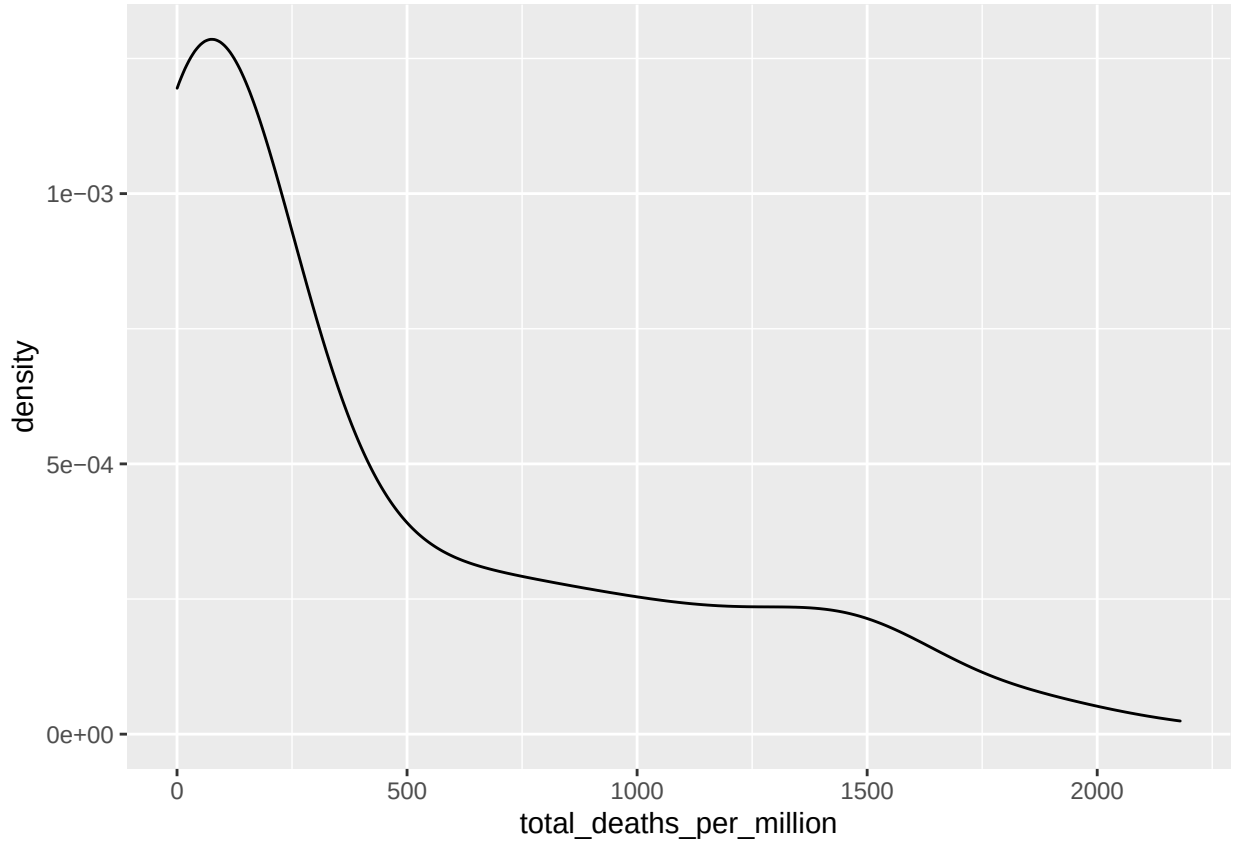
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 12 rows containing non-finite values (stat_bin).
```




```
COVID_data_latest %>%
ggplot() +
  aes(x = total_deaths_per_million) +
  geom_density()
```

```
## Warning: Removed 12 rows containing non-finite values (stat_density).
```



Gyakorlas

Szurd az adatokat ugy hogy csak a 2020-09-07-en jeletett adatokkal dolgozzunk

Hasznald a fent tanult modszereket, hogy **azonositsd az COVID_data adattablaban levo hibakat** vagy nem vart furcsasagokat.

- A vizualizacion tul a View(), describe(), es summary() funciokat erdemes hasznalni az adatok elso attekintesere
- A numerikus (vagy eppen folytonos) valtozoknal vizsgald meg a minimum es maximum erteket es a hianyzo adatok mennyiseget, valamint az eloszlast.
- A kategorikus valtozoknal vizsgald meg az osszes faktorszintet es az egyes szintekhez tartozo megfigyelesek mennyiseget.

A hibakat a kovetkezokeppen javithatjuk.

A **mutate()** es a **replace()** funkciok hasznalataval **cserelhetunk ki** ertekeket mas ertekekre. Azt, hogy ilyenkor hianyzo adatra (NA), vagy egy masik, valoszinu ertekre kell megvaltoztatni az erteket, a szituaciotol fogg. Altalaban a biztosabb megoldas ha hianyzo adatnak jeloljuk a kerdeses erteket (NA), de ez sok adatveszteshez vezethet. Ha eleg valoszinu hogy mi a helyes valasz, beirhatjuk, DE **minden javitast fel kell tuntetni** a kutatasi jelentesben (es a ZH soran is), hogy az olvaso szamara tiszta legyen, hogy itt egy adathelyettesites vagy kizaras tortent!

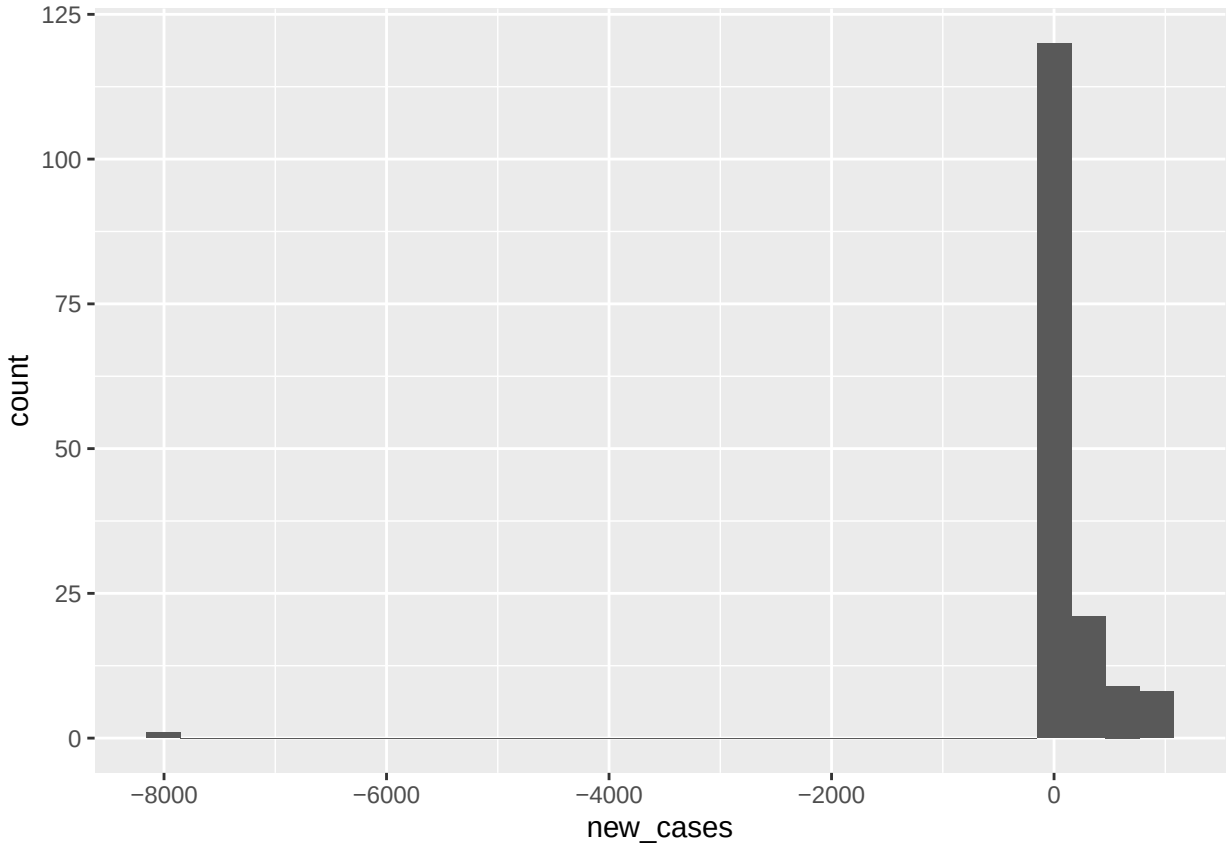
Mindig erdemes a javitott adatokat **uj adattablaba** elmenteni. A mi esetunkben az COVID_data_corrected nevet adtuk a javitott objektumnak. Igy a nyers adataink megmaradnak, ami hasznos lehet kesobbi

muveleteknel.

```
COVID_data %>%
  filter(date == "2020-09-07") %>%
  select(new_cases) %>%
  summary()
```

```
##    new_cases
##  Min.   :-7953
##  1st Qu.:    4
##  Median :   62
##  Mean   : 1177
##  3rd Qu.:  400
##  Max.   :75809
##  NA's   :1
```

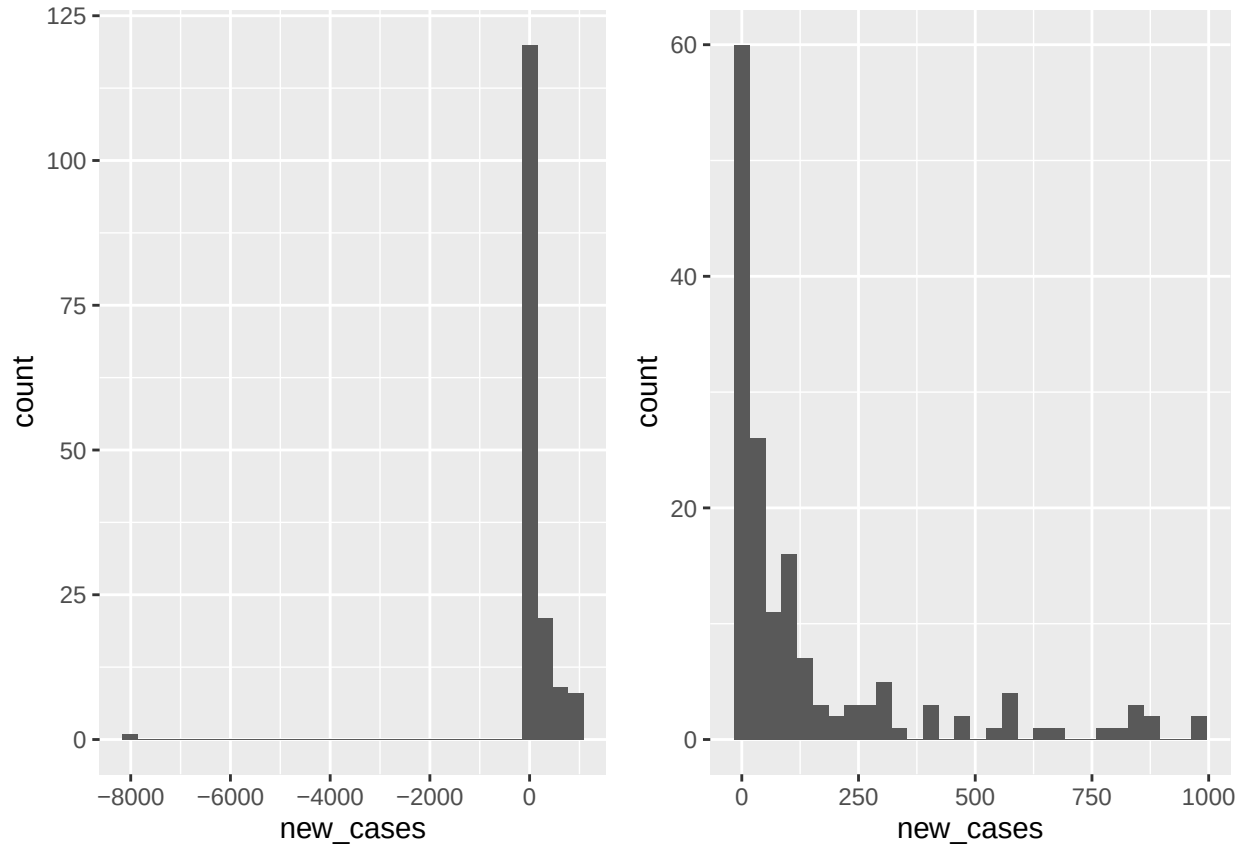
```
COVID_data %>%
  filter(date == "2020-09-07", new_cases < 1000) %>%
  ggplot()+
    aes(x = new_cases)+
    geom_histogram()
```




```
COVID_data_corrected <- COVID_data %>%
  mutate(new_cases = replace(new_cases,  new_cases=="-7953", NA))
```

Erdemes **megbizonyosodni rola**, hogy az adatcsere sikeres volt. Alabb az adatok vizualizaciojaval gyozodunk meg errol, de az adatok megjelenitesevel, vagy a leiro statisztikak lekerdezesevel is megteheto ez, ha az informativ.

```r
# hasznalhatnak meg az alabbiakat is arra,
# hogy megbizonyosodjunk abban, hogy sikeres volt a csere
# View(COVID_data_corrected)
# describe(COVID_data_corrected)
# summary(COVID_data_corrected$szocmedia_3)
# COVID_data_corrected$szocmedia_3

old_plot <-
  COVID_data %>%
  filter(date == "2020-09-07", new_cases < 1000) %>%
  ggplot()+
    aes(x = new_cases)+
    geom_histogram()

new_plot <-
  COVID_data_corrected %>%
  filter(date == "2020-09-07", new_cases < 1000) %>%
  ggplot()+
    aes(x = new_cases)+
    geom_histogram()


grid.arrange(old_plot, new_plot, ncol=2)
```



Tobb valtozo kapcsolatanak felterkepezese

Tobb valtozo kapcsolatat is felterkepezhetjuk tablazatok es abrak segitsegevel.

Ket kategorikus (csoportosito) valtozo kapcsolatanak felterkepezese

Feltaro elemzes

Most vizsgaljuk meg azt, hogy 2020-09-28-an mi az osszefuggese a gdp kategorianak (*gdp_per_capita_kat*) a kontinenssel (*continent*) ahol az orszag elhelyezkedik.

A legegyszerubb modja ket csoportosito valtozo kapcsolatanak megvizsgalasara a ket valtozo **kereszttablazatanak (crosstab)** elkezsitese a **table()** funkcioval.

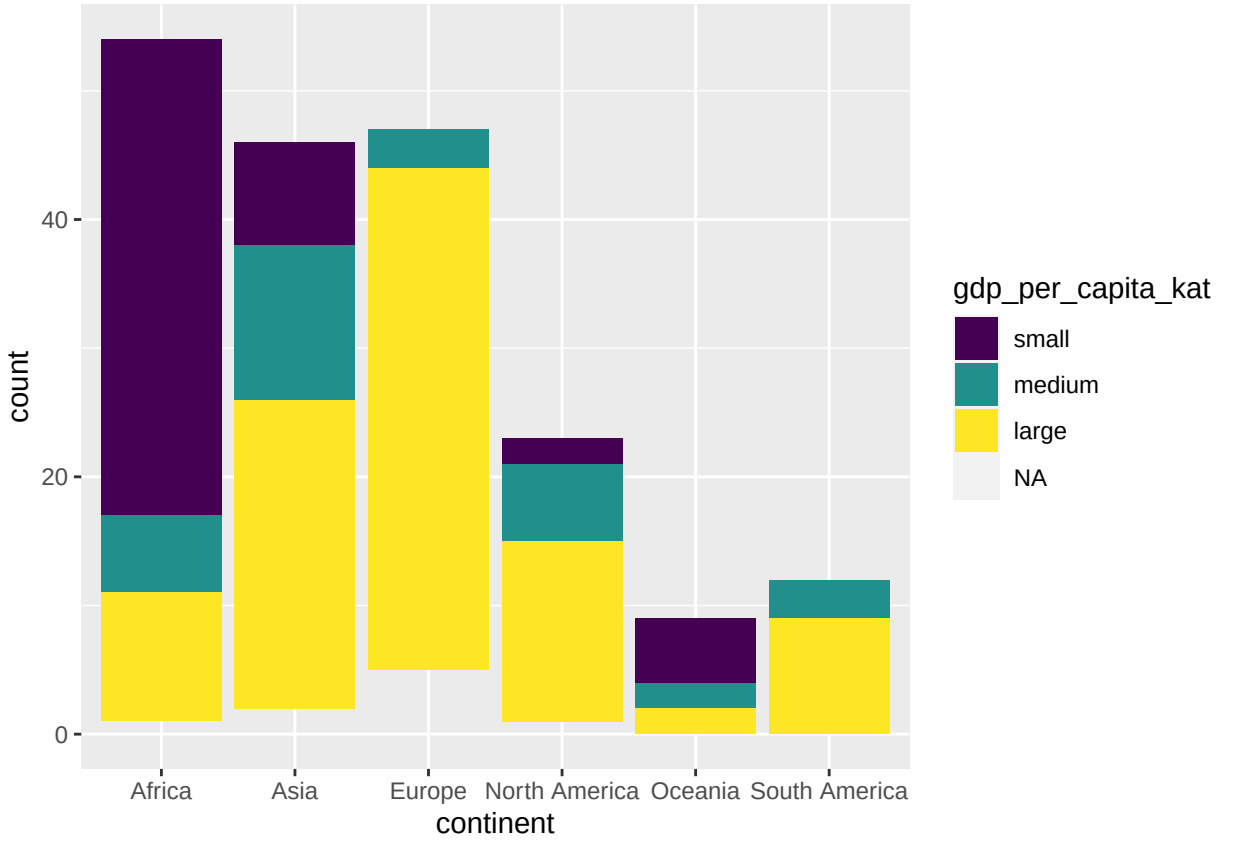
```
table(COVID_data_latest$gdp_per_capita_kat, COVID_data_latest$continent)
```

```
##
##          Africa Asia Europe North America Oceania South America
##   small      37    8      0             2       5             0
##   medium      6   12      3             6       2             3
##   large      10   24     39            14       2             9
```

Sokszor ennel sokkal **szemleletesebb az abrak** (plot) hasznalata.

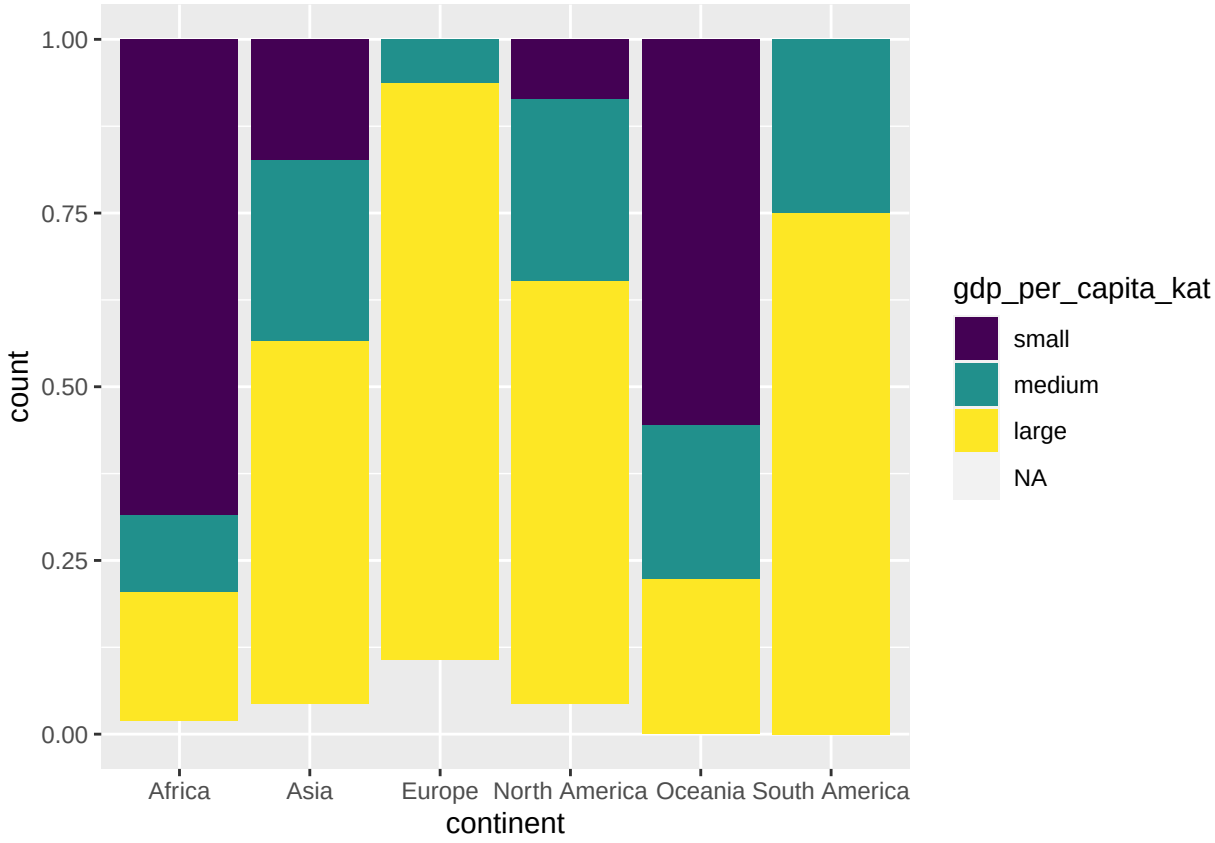
Erre az egyik lehetoseg a **stacked bar chart** (egymasra tornyozott oszlopdiagram, a **geom_bar()** geomot hasznaljuk) hasznalata. Itt az egyik valtozo kategoriai adjak meg hany oszlop lesz (ez a valtozo lesz az x tengelyen reprezentalva, igy ezt az "x =" reszen adhatjuk meg), a masik valtozo az oszlopokat szinekkel szegmentalja, ezt pedig a "**fill =**" reszen adhatjuk meg.

```
COVID_data_latest %>%
ggplot() +
  aes(x = continent, fill = gdp_per_capita_kat) +
  geom_bar()
```



Ha az egyes faktorszinteken nagyon **kulonbozo mennyisegu megfigyeles** van, ez a megjelenites neha felrevezeto kovetkeztetesekhez vezethet, igy neha hasznosabb ha az oszlopok nem szamossagot (count), hanem **reszaranyt (proportion)** jelolnek. Ha ezt szeretnenk, ahelyett hogy uresen hagynank a geom_bar() funkciot, a kovetkezot adjuk meg: **geom_bar(position = "fill")**. Vagy hasznalhatjuk az eltolt oszlopdiagramot (dodged barchart) (a **position = "dodge"** parameter megadasaval a geom_bar() -on belul)

```
COVID_data_latest %>%
ggplot() +
  aes(x = continent, fill = gdp_per_capita_kat) +
  geom_bar(position = "fill")
```



Gyakorlas

Hasznald a fent tanult modszereket, hogy megvizsgald a COVID_data_latest adatbazisban a **new_cases_per_million_kat** es a **continent** valtozok kozotti osszefuggest. - hasznalj **geom_bar()** geomot a megjeleniteshez - probald meg mind a **szamossagot**, mind a **reszaranyt** kifejezo abrat megvizsgalni geom_bar(position = "fill") - milyen **kovetkeztetest** tudsz levonni az abrakrol?

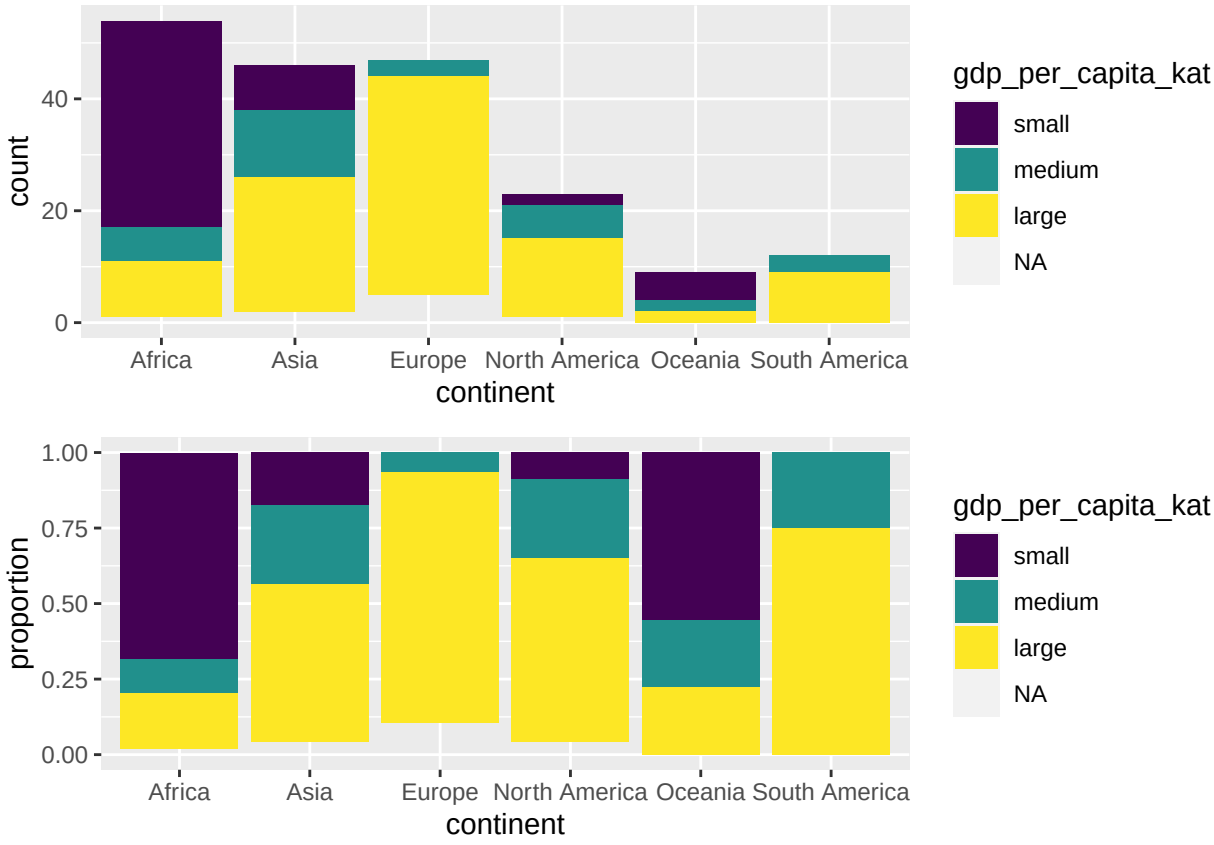
```
# a fenti gyakorlashoz a new_cases_per_million_kat valtozot igy lehet legeneralni:

COVID_data = COVID_data %>%
  mutate(new_cases_per_million_kat = factor(
                                      case_when(new_cases_per_million < 20 ~ "small",
                                                new_cases_per_million >= 20 ~ "large"), ordered = T, lev

levels(COVID_data$new_cases_per_million_kat)
```

```
## [1] "small" "large"
```

```
# ugyanez a COVID_data_latest -al

COVID_data_latest = COVID_data_latest %>%
  mutate(new_cases_per_million_kat = factor(
                                      case_when(new_cases_per_million < 20 ~ "small",
                                                new_cases_per_million >= 20 ~ "large"), ordered = T, lev
```

geom_bar() megjelenitesnel fontos hogy ha az egyes megfigyelesek **keves megfigyelesbol allnak**, az abra megteveszto lehet, mert az abra nem jelzi a megfigyelesek szamat es igy azt, hogy milyen biztosak lehetunk az eredmenyben. Ilyen esetekben az egyik kategoriat ki lehet venni az abrarol, vagy a **szamossagot es a reszaranyt abrazolo abrakat egymas mellet** lehet bemutatni, hogy igy kiegeszitsek egymast. Ehhez hasznalhatjuk a **grid.arrange()** funkciot.

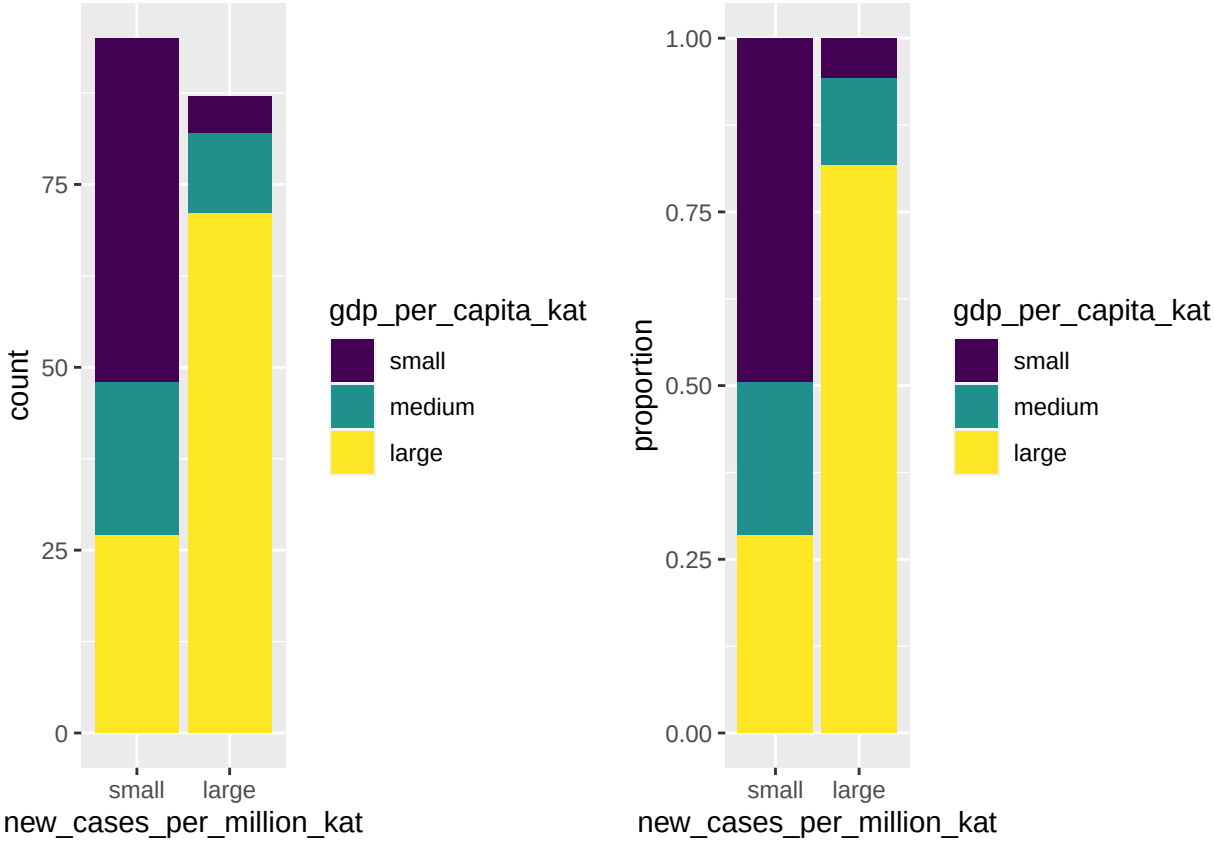
```
szamossag_plot <-
COVID_data_latest %>%
ggplot() +
  aes(x = continent, fill = gdp_per_capita_kat) +
  geom_bar()

reszarany_plot <-
COVID_data_latest %>%
ggplot() +
  aes(x = continent, fill = gdp_per_capita_kat) +
  geom_bar(position = "fill") +
  ylab("proportion")

grid.arrange(szamossag_plot, reszarany_plot, nrow=2)
```



A theme(legend.position) es a guides() funciok hasznalataval kontrollalhatjuk hogy hol es hogyan jelenjen meg a **jelmagyarazat** az abran. Az abra **interpretalhatosaga** attol fuggoen is **valtozhat**, hogy melyik valtozot tesszuk az x-tengelyre es melyiket szinkent abrazolva.

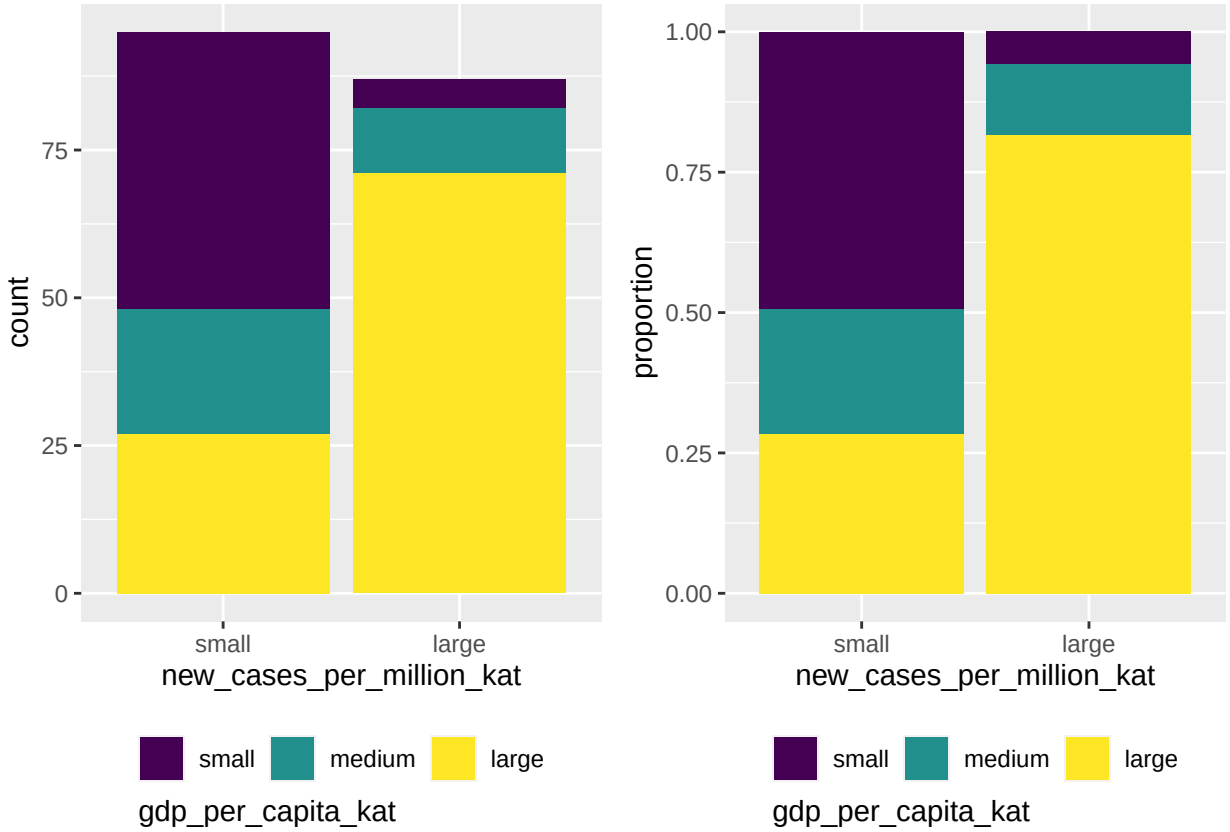
Az alabbi abrakon az egymillio fore vetitett uj esetek szamanak kapcsolatat nezzuk meg a gdp-vel. Mindket valtozo eseten a csoportositott valtozot (_kat) hasznaljuk.

```r
barchart_plot_3 <-
COVID_data_latest %>%
  select(new_cases_per_million_kat, gdp_per_capita_kat) %>%
  drop_na() %>%
ggplot() +
  aes(x = new_cases_per_million_kat, fill = gdp_per_capita_kat) +
  geom_bar()


barchart_plot_4 <-
COVID_data_latest %>%
  select(new_cases_per_million_kat, gdp_per_capita_kat) %>%
  drop_na() %>%
ggplot() +
  aes(x = new_cases_per_million_kat, fill = gdp_per_capita_kat) +
  geom_bar(position = "fill") +
  ylab("proportion")

grid.arrange(barchart_plot_3, barchart_plot_4, ncol=2)
```



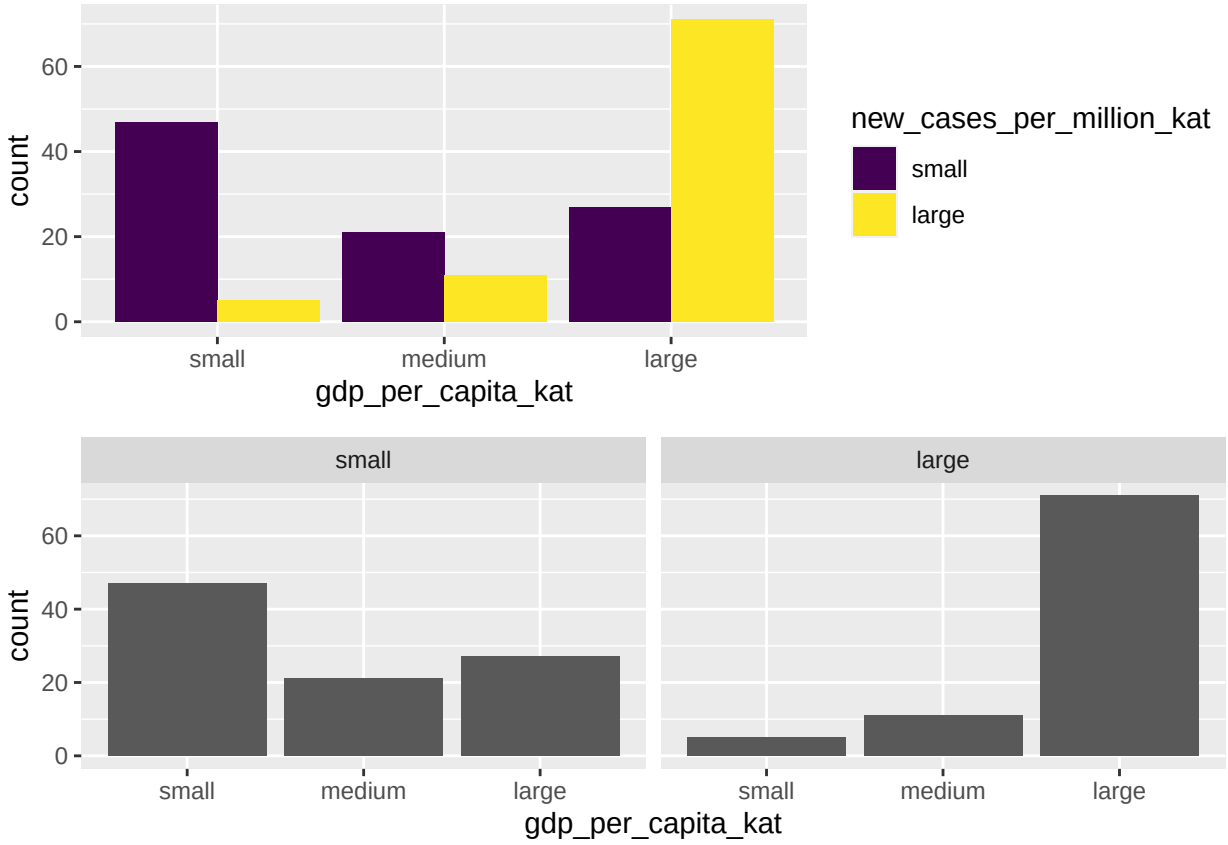

```r
# a theme(legend.position) es a guides() funciok
# hasznalataval kontrollalhatjuk hogy hol es hogyan
# jelenjen meg a jelmagyarazat az abran

barchart_plot_3 <-
COVID_data_latest %>%
```

```r
  select(new_cases_per_million_kat, gdp_per_capita_kat) %>%
  drop_na() %>%
ggplot() +
    aes(x = new_cases_per_million_kat, fill = gdp_per_capita_kat) +
    geom_bar() +
    theme(legend.position="bottom") +
    guides(fill = guide_legend(title.position = "bottom"))


barchart_plot_4 <-
COVID_data_latest %>%
  select(new_cases_per_million_kat, gdp_per_capita_kat) %>%
  drop_na() %>%
ggplot() +
  aes(x = new_cases_per_million_kat, fill = gdp_per_capita_kat) +
  geom_bar(position = "fill") +
  theme(legend.position="bottom") +
  guides(fill = guide_legend(title.position = "bottom")) +
  ylab("proportion")

grid.arrange(barchart_plot_3, barchart_plot_4, ncol=2)
```



Ujabb modja a barchart segitsegevel valo megjelenitesnek ha az oszlopok nem egymasra tornyozva, hanem **egymas mellett** jelennek meg, vagy ha a masodik valtozo szerint **kulon paneleken (facet)** jelennek meg.

```
barchart_plot_5 <-
COVID_data_latest %>%
  select(new_cases_per_million_kat, gdp_per_capita_kat) %>%
  drop_na() %>%
ggplot() +
  aes(x = gdp_per_capita_kat, fill = new_cases_per_million_kat) +
  geom_bar(position = "dodge")

barchart_plot_6 <-
COVID_data_latest %>%
  select(new_cases_per_million_kat, gdp_per_capita_kat) %>%
  drop_na() %>%
ggplot() +
  aes(x = gdp_per_capita_kat) +
  geom_bar() +
  facet_wrap(~ new_cases_per_million_kat)

grid.arrange(barchart_plot_5, barchart_plot_6, nrow=2)
```



Egy kategorikus es egy numerikus valtozo kapcsolata

Vizsgaljuk meg hogy hogyan alakul az egy fore juto GDP kontinensenkent. A GDP ebben az esetben egy folytonos valtozó (gdp_per_capita), es ennek az osszefuggeset szeretnenk megvizsgalni egy kategorikus valtozoval (continent).

Az exploraciot kezdhetjuk leiro statisztikak lekerdezesevel csoportonkent. Peldaul ha arra vagyunk kivancsiak,

milyen a GDP atlaga es szorasa kontinensenkent, ezt megvizsgalhatjuk a **group_by()** es a **summarize()** segitsegevel.

```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  group_by(continent) %>%
    summarize(mean = mean(gdp_per_capita),
              sd = sd(gdp_per_capita))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 3
##   continent       mean     sd
##   <fct>          <dbl>  <dbl>
## 1 Africa         5444.  6183.
## 2 Asia          22185. 25406.
## 3 Europe        33361. 18030.
## 4 North America 17126. 12871.
## 5 Oceania       12392. 16121.
## 6 South America 13841.  5110.
```

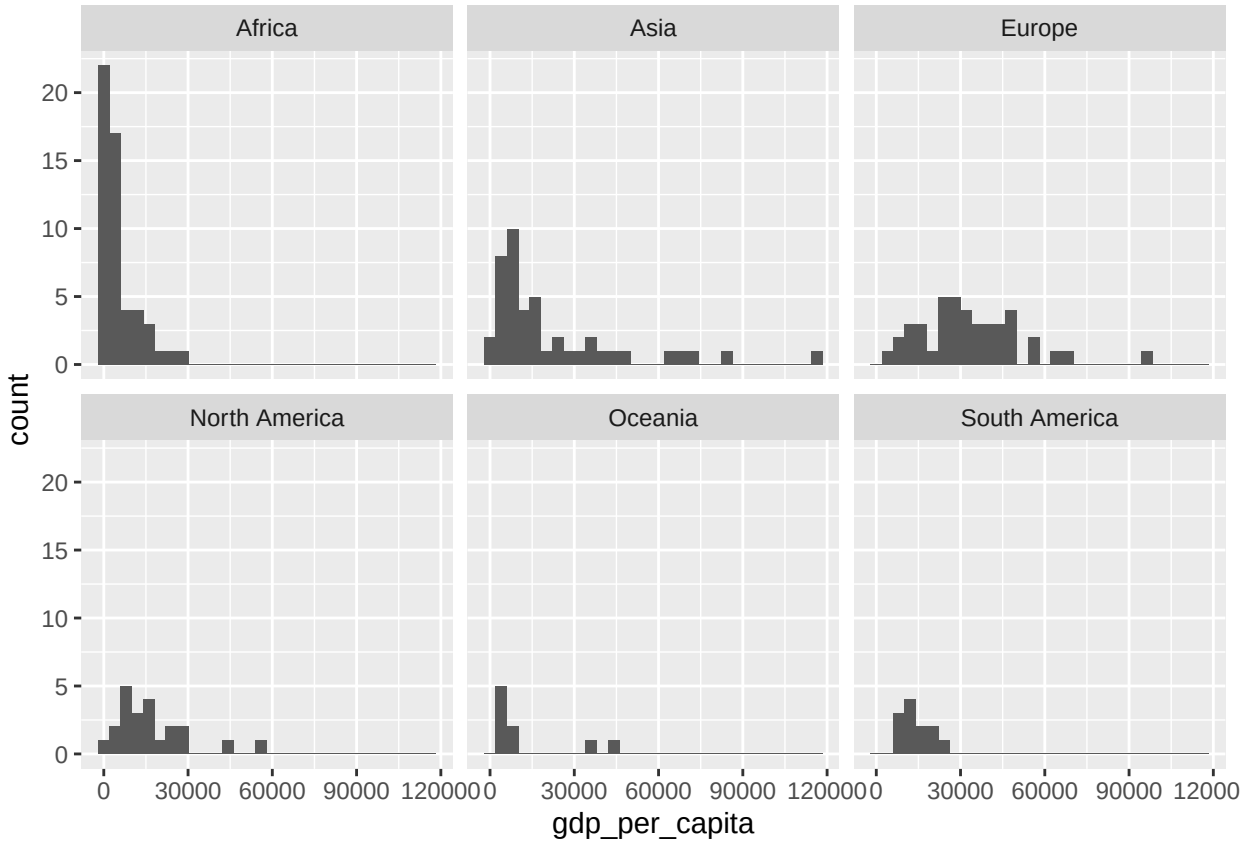
A ket valtozo kapcsolatat megvizsgalhatjuk **abrakkal** is. Pl. hasznalhatjuk a

- **facet_wrap()** fuggvenyt egy **geom_histogram()**-al kobinalva
- a **geom_boxplot()** -ot
- esetleg hasznalhatunk egy egymasra illesztett **geom_density()** plot-ot ahol a kategoriak mas mas szinnel vannak jelolve.
- talan ebben az esetben a legtisztabb kepet a **geom_violin()** mutatja, ami a geom_boxplot() es a geom_density() keverekenek tekintheto. Ezt kiegeszithetunk egy **geom_point()** -al, hogy pontosan latsszon, hany megfigyelesen alapulnak az abra adatai.
- az egyik kedvencem a **geom_violin()** a **geom_jitter()**-el valo kombinacioban

Mindig erdemes **tobb megkozelitest** is hasznalni az adat-exploracio kozben, hogy minel reszletesebb kepet kaphassunk, es csokkentsuk a valoszinuseget hogy egyik vagy masik megkozelites hianyossagai felrevezetnek minket.

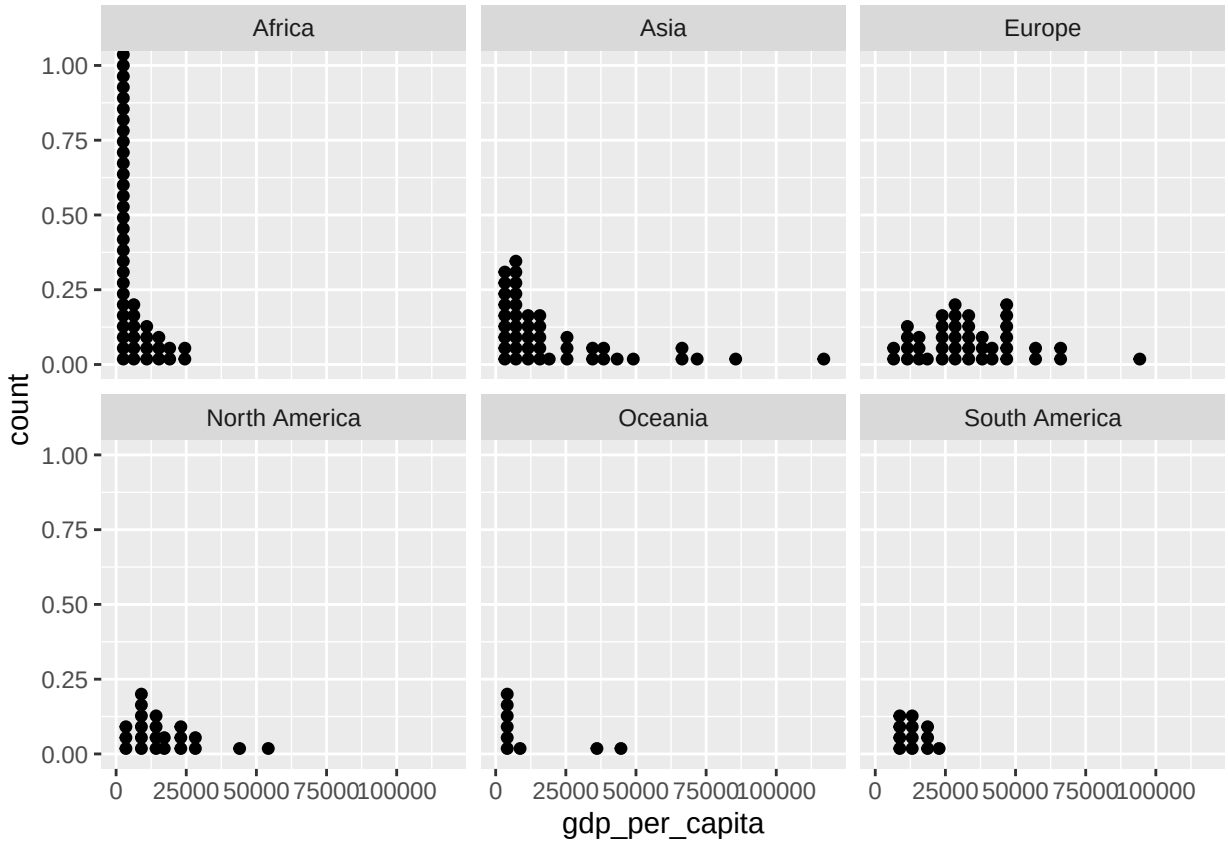
```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
    aes(x = gdp_per_capita) +
    geom_histogram() +
    facet_wrap(~ continent)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

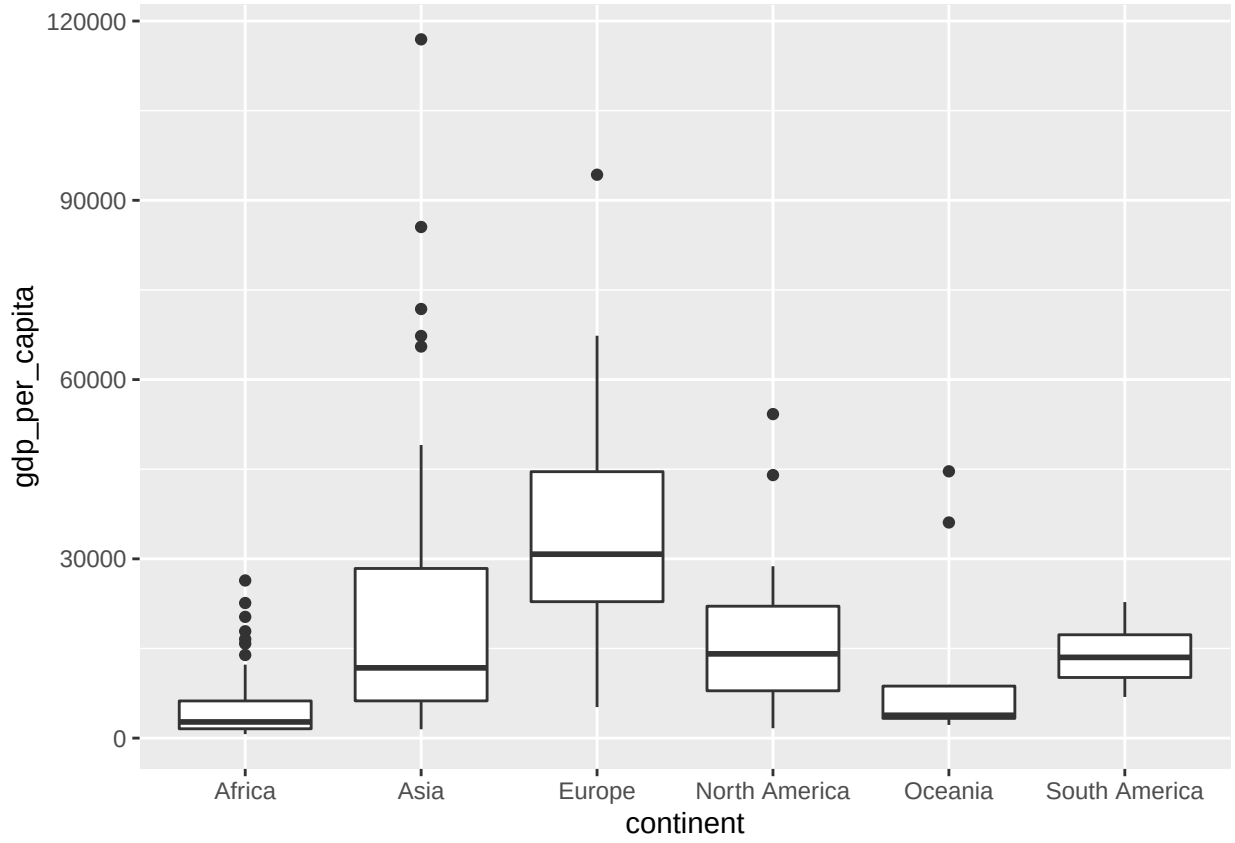



```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
    aes(x = gdp_per_capita) +
    geom_dotplot() +
    facet_wrap(~ continent)
```

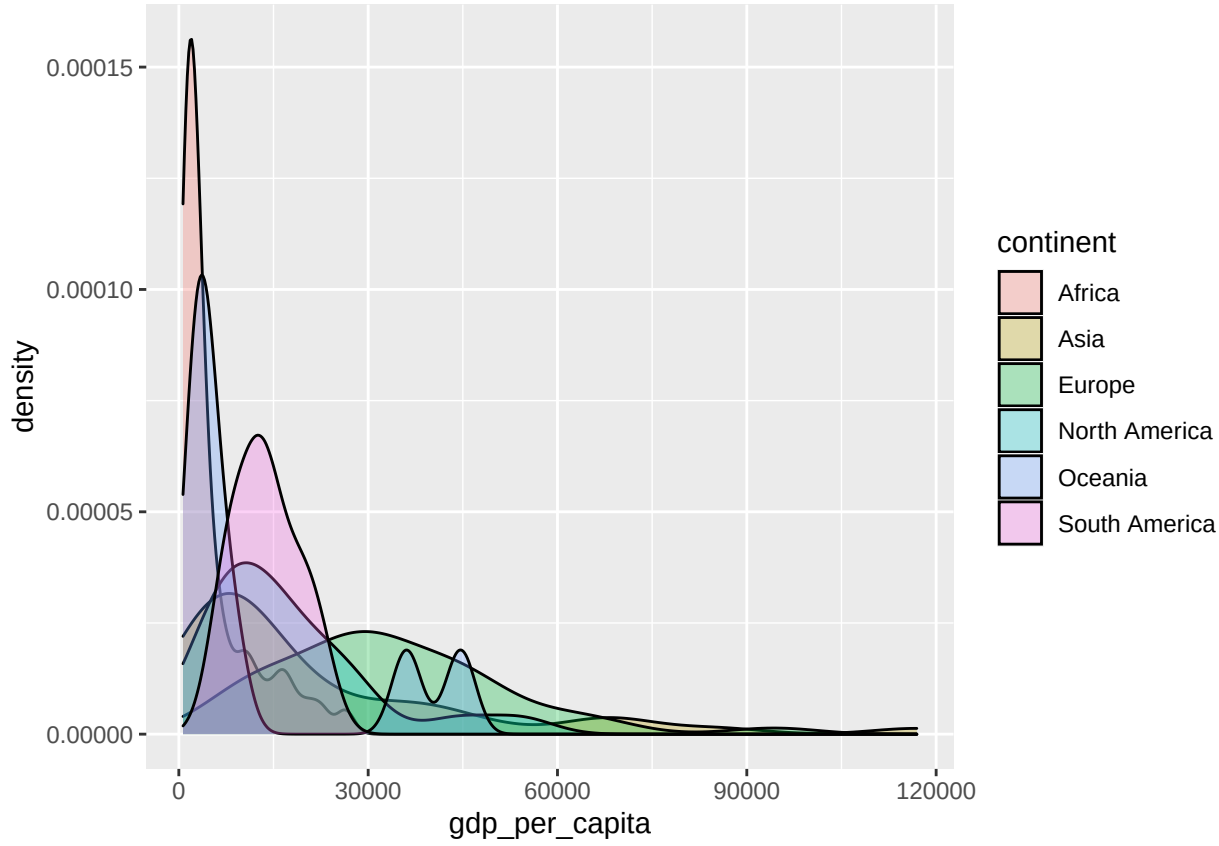
```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



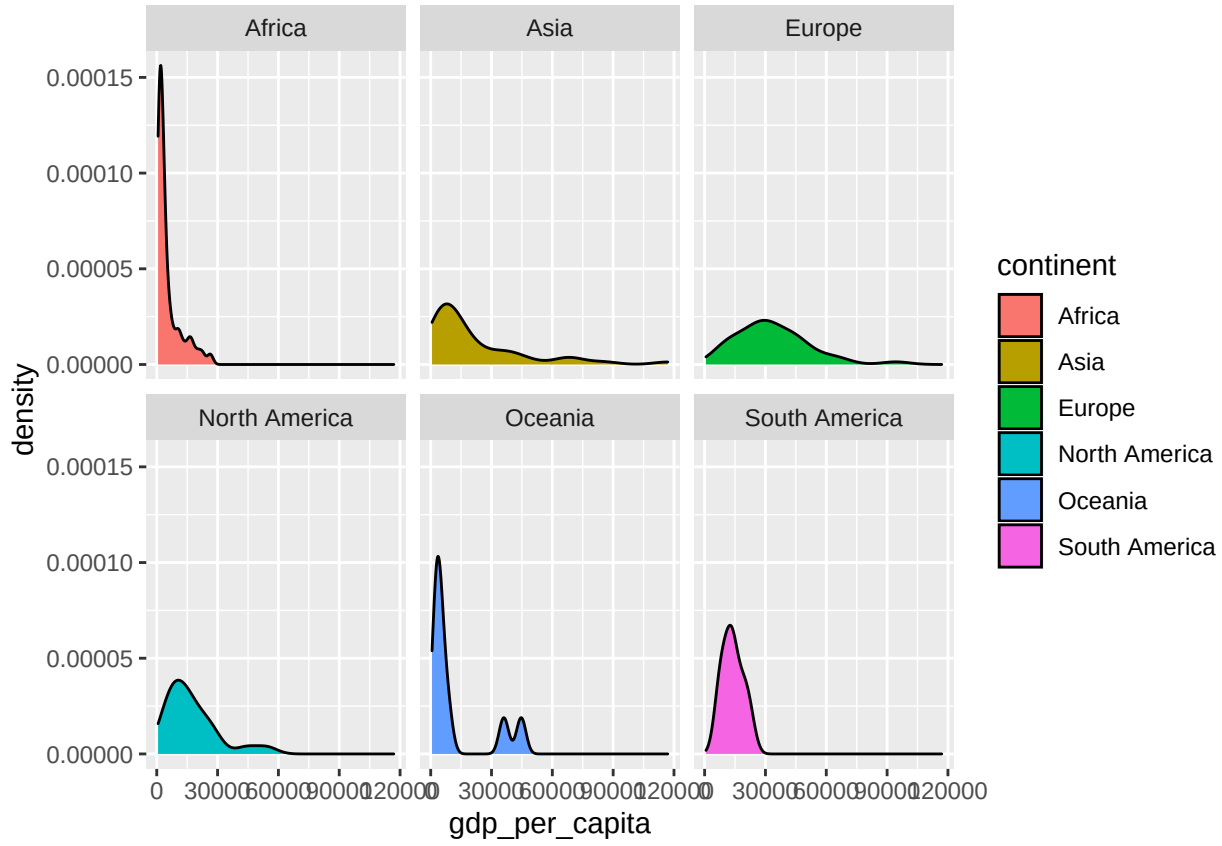

```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
    aes(x = continent, y = gdp_per_capita) +
    geom_boxplot()
```



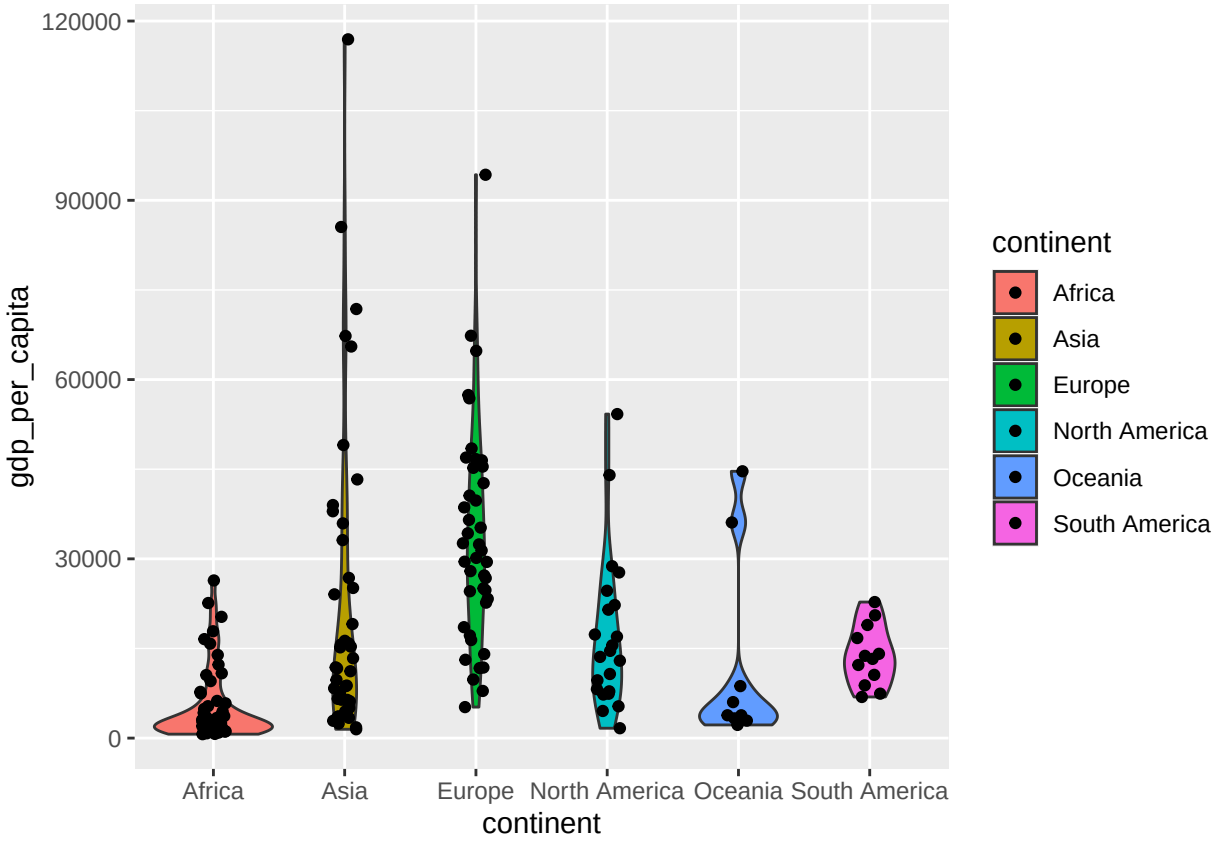

```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
    aes(x = gdp_per_capita, fill = continent) +
    geom_density(alpha = 0.3)
```




```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
    aes(x = gdp_per_capita, fill = continent) +
    geom_density()+
  facet_wrap(~continent)
```



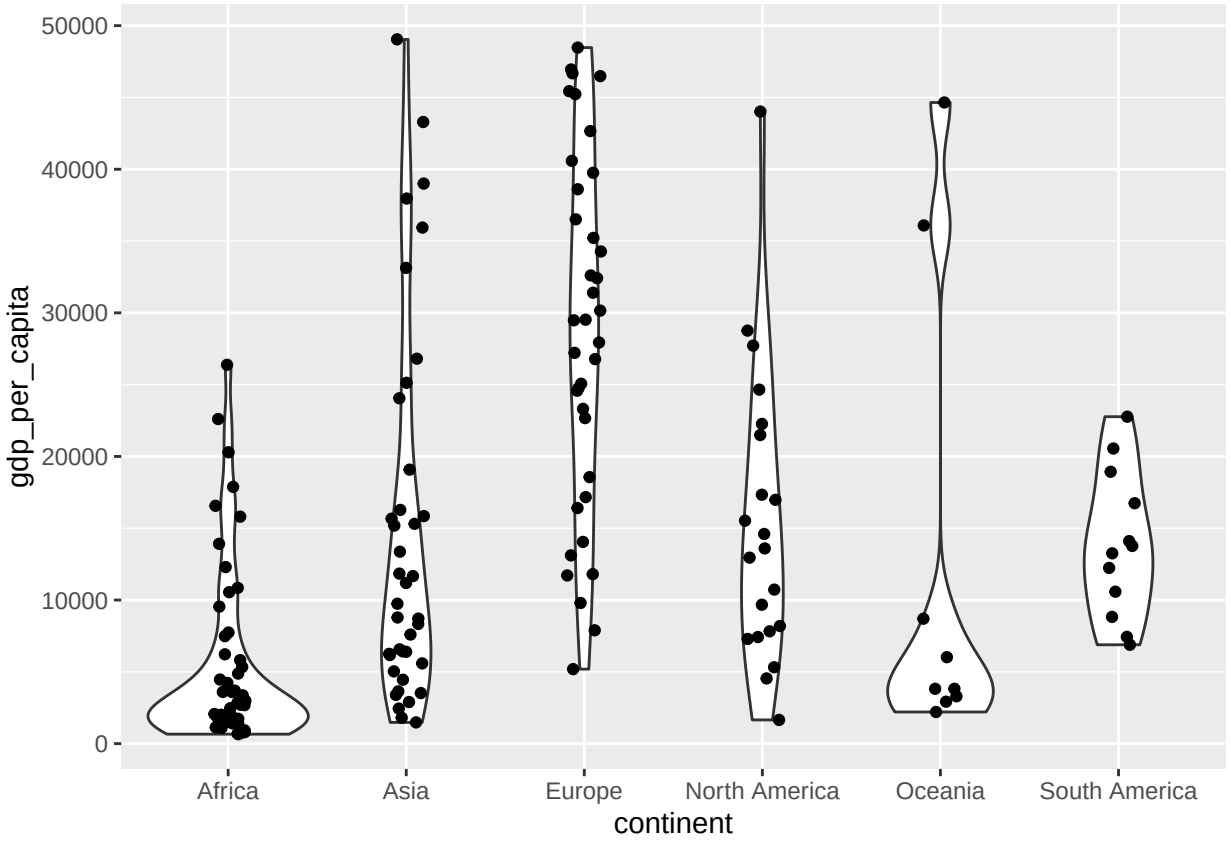

```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
    aes(x = continent, y = gdp_per_capita, fill = continent) +
    geom_violin() +
    geom_jitter(width = 0.1)
```



A fenti abran latszik, hogy Azsiaban a legtobb orszagban viszonylag alacsony a gdp, viszont van nehany **kiurgo ertek**, az atlagot felhuzza ebben a csoportban.

Ha szeretnenk **kizarni az elemzesunkbol** az extrem ertekekt, a **filter()** funkcio beekelesevel a pipe-ba megepithetjuk a fenti abrankat es tablazatokat ugy, hogy csak a 50000-nel alancsonyabb GDP-ju orszagok keruljenek az abrara.

```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  filter(gdp_per_capita < 50000) %>%
    ggplot() +
      aes(x = continent, y = gdp_per_capita) +
      geom_violin() +
      geom_jitter(width = 0.1)
```




```
COVID_data_latest %>%
  select(continent, gdp_per_capita) %>%
  drop_na() %>%
  filter(gdp_per_capita < 50000) %>%
    group_by(continent) %>%
      summarize(mean = mean(gdp_per_capita),
                sd = sd(gdp_per_capita))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 3
##   continent       mean     sd
##   <fct>          <dbl>  <dbl>
## 1 Africa         5444.  6183.
## 2 Asia          14591. 12710.
## 3 Europe        28661. 12390.
## 4 North America 15359. 10092.
## 5 Oceania       12392. 16121.
## 6 South America 13841.  5110.
```

Ha szeretnenk latni hogy a kisebb vagy nagyobb uj esetszammal jellemezheto orszagok (new_cases_per_million_kat) hogyan kulonboznek a GDP tekinteteben kontinensenkent akkor mar **harom valtozo** kapcsolatat kell abrazolnunk. Ehhez a facet_grid() funkciot lehet hasznalni, vagy kulonbozo esztetikai elemeket (aes()) lehet a kulonbozo valtozokhoz rendelni.

Gyakorlas

Hasznald a fent tanult modszereket, hogy megvizsgald a **total_cases_per_million** es a **gdp_per_capita_kat** valtozok kozotti osszefuggest.

- hasznald a fenti geomokat, es keszits legalabb ket kulonbozo abrat mas-mas geomokkal

---

Ket numerikus valtozo kapcsolata

Ket numerikus valtozo kozotti kapcsolat jellemzesere altalaban a korrelacios egyutthatot szoktuk hasznalni (cor()). A **cor()** funkciot akar tobb mint ket valtozo paronkenti korrelaciojanak meghatarozasara is lehet hasznalni.

A **drop_na()** funkcioval kiejthetjuk azokat a megfigyeleseket, ahol a valtozok barmelyikeben hianyzo adat (NA) van. Ha ezt nem tesszuk meg, a cor() fuggveny NA eredmenyt adhna ha valamelyik valtozoban NA-val talalkozik.

```
COVID_data_latest %>%
  select(new_cases_per_million, gdp_per_capita) %>%
  drop_na() %>%
      cor()
```

```
##                       new_cases_per_million gdp_per_capita
## new_cases_per_million             1.0000000      0.3296953
## gdp_per_capita                    0.3296953      1.0000000
```

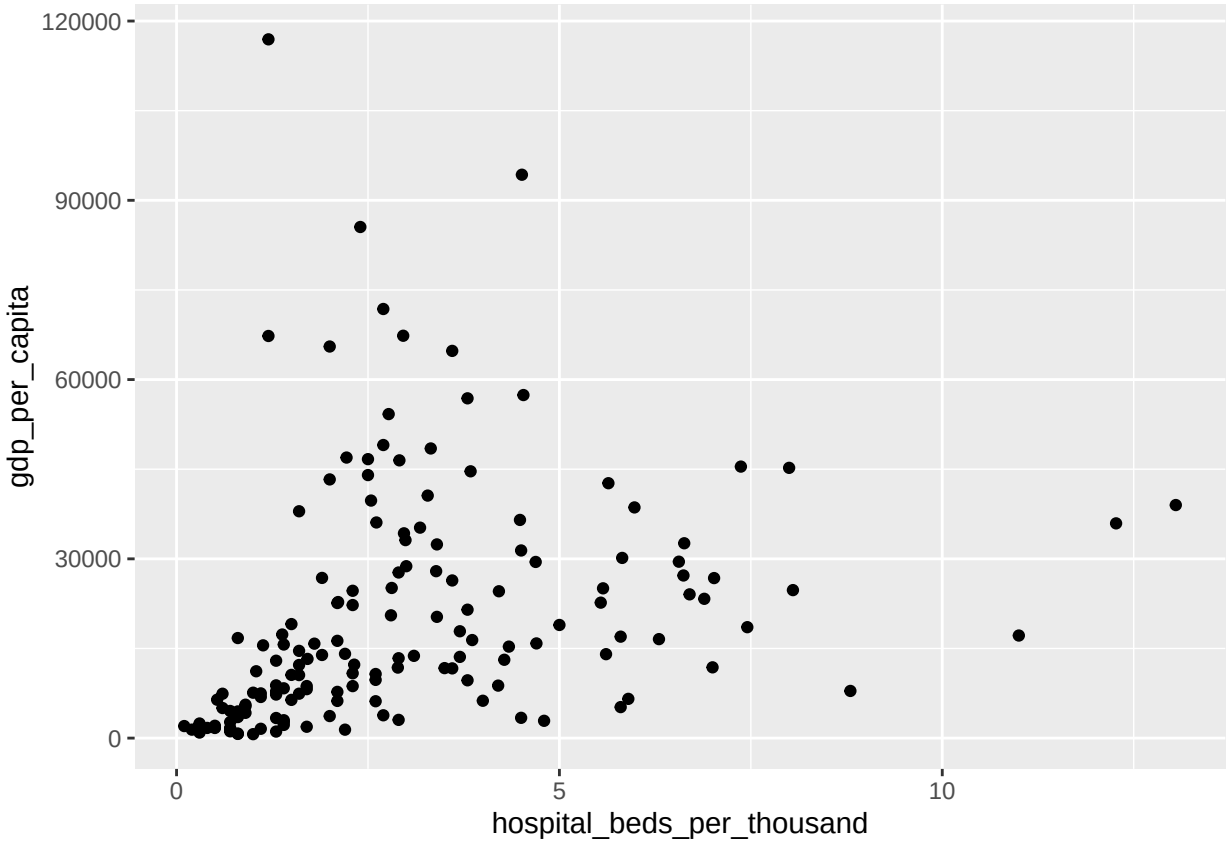
```
COVID_data_latest %>%
  select(new_cases_per_million, gdp_per_capita, hospital_beds_per_thousand) %>%
  drop_na() %>%
      cor()
```

```
##                            new_cases_per_million gdp_per_capita
## new_cases_per_million                  1.0000000      0.3082168
## gdp_per_capita                         0.3082168      1.0000000
## hospital_beds_per_thousand             0.2026815      0.2995055
##                            hospital_beds_per_thousand
## new_cases_per_million                       0.2026815
## gdp_per_capita                              0.2995055
## hospital_beds_per_thousand                  1.0000000
```

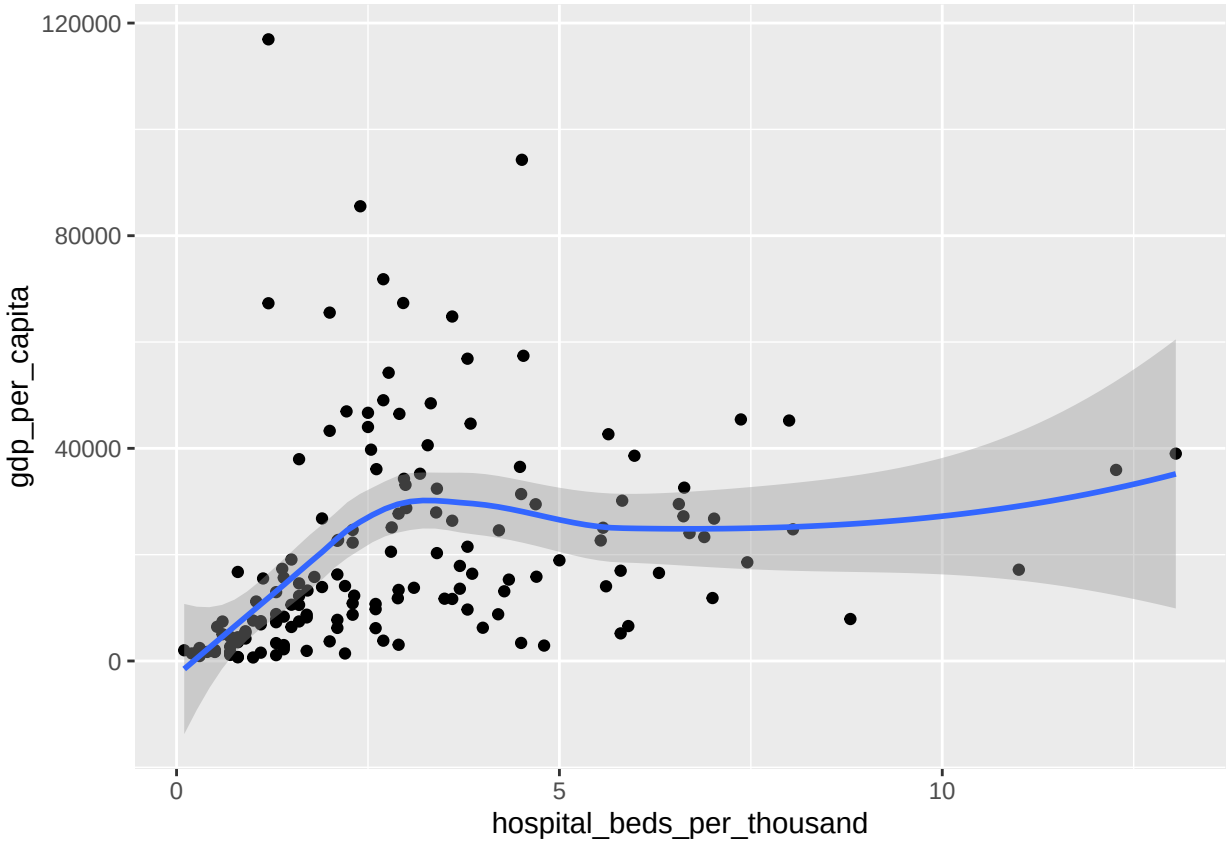
A numerikus valtozok kozotti kapcsolatot altalaban pont diagrammal szoktuk abrazolni (**geom_point()**)

A **geom_smooth()** layer hozzaadasaval kaphatunk a pontok kozott meghuzodo trendrol egy kepet. A kek vonal az ugyevezett trendvonal, a szurke sav a konfidencia intervallum. Ezekrol kesobb meg reszletesebben beszelunk majd

```
COVID_data_latest %>%
  select(hospital_beds_per_thousand, gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
    aes(x = hospital_beds_per_thousand, y = gdp_per_capita) +
    geom_point()
```




```
COVID_data_latest %>%
  select(hospital_beds_per_thousand, gdp_per_capita) %>%
  drop_na() %>%
  ggplot() +
    aes(x = hospital_beds_per_thousand, y = gdp_per_capita) +
    geom_point() +
    geom_smooth()
```



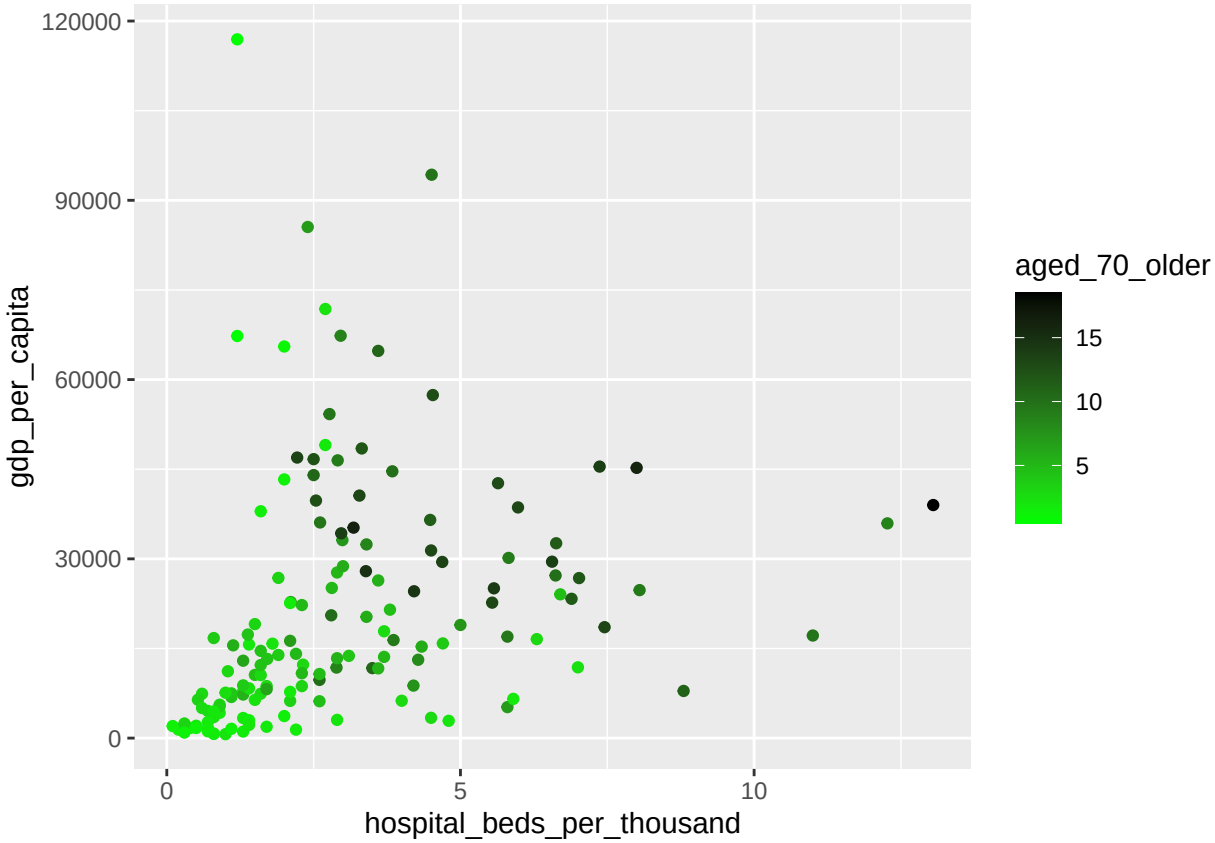
Gyakorlas

Milyen eros a kapcsolat a aged_70_older es a gdp_per_capita kozott?

- hatarozd meg a korrelacios egyutthatot a valtozok kozott
- abrazold a valtozok kapcsolatat

Tobb folytonos valtozo kapcsolata megjelenitheto peldaul ugy, hogy az egyik valtozot egy szinskalahoz rendeljuk az alabbi modon.

```
COVID_data_latest %>%
  select(hospital_beds_per_thousand, gdp_per_capita, aged_70_older) %>%
  drop_na() %>%
  ggplot() +
    aes(x = hospital_beds_per_thousand, y = gdp_per_capita, col = aged_70_older) +
    geom_point()+
  scale_colour_gradientn(colours=c("green","black"))
```


120000
90000
60000
30000
0
gdp_per_capita
0
5
10
hospital_beds_per_thousand
aged_70_older
15
10
5