Data Exploration - R

November 8, 2020

1 ZEWK - Hands On Datenvisualisierung, Explorative Datenanalyse in R

Seminar von Letty und Karen

Dieses Notebook dient als Beispiel für die Implementation einer Datenvisualisierungspipeline in der Sprache R. Es kann sowohl in Jupyter als auch in Jupyter Lab ausgeführt werden, jedoch können sich einzelne Shortcuts unterscheiden.

1.1 Benutzung von Jupyter (Lab)

Hier ein paar praktische und wichtige Kommandos und Tastenkombinationen die ihr kennen solltet:
Außerhalb einer Zelle:

- ENTER - Zelle editieren
- strg + ENTER - Zelle ausführen
- shift + ENTER - Zelle ausführen und zur nächsten gehen

Innerhalb einer Zelle (Editiermodus der Zelle):

- ESC - Zelle verlassen
- D, D - Zelle löschen
- A - leere Zelle oberhalb (above) einfügen
- B - leere Zelle unterhalb (below) einfügen

1.2 Explorative Datenanalyse

Wir haben einen unbekannten Datensatz (https://github.com/owid/covid-19-data/tree/master/public/data) und wollen herausfinden welche Daten sich darin verbergen um Arbeitshypothesen und Fragestelltungen für Visualisierungen zu entwickeln.

1.2.1 Setup von hilfreichen Python Packages

[1]:
```
library(dplyr)     ## Hadley Wickham's grammar of data manipulation package
library(ggplot2)   ## Hadley Wickham's grammar of graphics plotting package
library(GGally)    ## ggplot version of pairs plot
library(MASS)
library(psych)
library(tidyverse)
```

```
Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2


Attaching package: 'MASS'


The following object is masked from 'package:dplyr':

    select



Attaching package: 'psych'


The following objects are masked from 'package:ggplot2':

    %+%, alpha


-- Attaching packages
--------------------------------------------------------------------------------
tidyverse 1.3.0 --
```

```
v tibble  3.0.4     v purrr   0.3.4
v tidyr   1.1.2     v stringr 1.4.0
v readr   1.4.0     v forcats 0.5.0

-- Conflicts ---------------------------------------------------------
----------------------- tidyverse_conflicts() --
x psych::%+%()    masks ggplot2::%+%()
x psych::alpha()  masks ggplot2::alpha()
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x MASS::select()  masks dplyr::select()
```

1.2.2 Import der Daten

[2]:
```
my_data <- read.csv("owid-covid-data.csv", header=TRUE, sep = "," )
```

1.2.3 Übersicht über die Daten bekommen

Welche Parameter haben wir? Zusammenfassung der Daten erzeugen

[3]:
```
# Informationen über die verwendeten Datentypen
glimpse(my_data)
```

```
Rows: 55,247
Columns: 49
$ iso_code                          <chr> "AFG", "AFG",
"AFG", "AFG", "AFG…
$ continent                         <chr> "Asia", "Asia",
"Asia", "Asia", …
$ location                          <chr> "Afghanistan",
"Afghanistan", "A…
$ date                              <chr> "2019-12-31",
"2020-01-01", "202…
$ total_cases                       <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ new_cases                         <dbl> 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0,…
$ new_cases_smoothed                <dbl> NA, NA, NA, NA,
NA, NA, 0, 0, 0,…
$ total_deaths                      <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ new_deaths                        <dbl> 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0,…
$ new_deaths_smoothed               <dbl> NA, NA, NA, NA,
NA, NA, 0, 0, 0,…
$ total_cases_per_million           <dbl> NA, NA, NA, NA,
```

```
NA, NA, NA, NA, …
$ new_cases_per_million              <dbl> 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0,…
$ new_cases_smoothed_per_million     <dbl> NA, NA, NA, NA,
NA, NA, 0, 0, 0,…
$ total_deaths_per_million           <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ new_deaths_per_million             <dbl> 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0,…
$ new_deaths_smoothed_per_million    <dbl> NA, NA, NA, NA,
NA, NA, 0, 0, 0,…
$ icu_patients                       <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ icu_patients_per_million           <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ hosp_patients                      <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ hosp_patients_per_million          <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ weekly_icu_admissions              <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ weekly_icu_admissions_per_million  <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ weekly_hosp_admissions             <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ weekly_hosp_admissions_per_million <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ total_tests                        <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ new_tests                          <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ total_tests_per_thousand           <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ new_tests_per_thousand             <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ new_tests_smoothed                 <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ new_tests_smoothed_per_thousand    <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ tests_per_case                     <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ positive_rate                      <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ tests_units                        <chr> "", "", "", "",
"", "", "", "", …
$ stringency_index                   <dbl> NA, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0…
$ population                         <dbl> 38928341,
```

```
38928341, 38928341, 38…
$ population_density                <dbl> 54.422, 54.422,
54.422, 54.422, …
$ median_age                        <dbl> 18.6, 18.6, 18.6,
18.6, 18.6, 18…
$ aged_65_older                     <dbl> 2.581, 2.581,
2.581, 2.581, 2.58…
$ aged_70_older                     <dbl> 1.337, 1.337,
1.337, 1.337, 1.33…
$ gdp_per_capita                    <dbl> 1803.987,
1803.987, 1803.987, 18…
$ extreme_poverty                   <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ cardiovasc_death_rate             <dbl> 597.029, 597.029,
597.029, 597.0…
$ diabetes_prevalence               <dbl> 9.59, 9.59, 9.59,
9.59, 9.59, 9…
$ female_smokers                    <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ male_smokers                      <dbl> NA, NA, NA, NA,
NA, NA, NA, NA, …
$ handwashing_facilities            <dbl> 37.746, 37.746,
37.746, 37.746, …
$ hospital_beds_per_thousand        <dbl> 0.5, 0.5, 0.5,
0.5, 0.5, 0.5, 0…
$ life_expectancy                   <dbl> 64.83, 64.83,
64.83, 64.83, 64.8…
$ human_development_index           <dbl> 0.498, 0.498,
0.498, 0.498, 0.49…
```

[4]:
```
# Parameterübersicht
names(my_data)
```

1. 'iso_code' 2. 'continent' 3. 'location' 4. 'date' 5. 'total_cases' 6. 'new_cases' 7. 'new_cases_smoothed' 8. 'total_deaths' 9. 'new_deaths' 10. 'new_deaths_smoothed' 11. 'total_cases_per_million' 12. 'new_cases_per_million' 13. 'new_cases_smoothed_per_million' 14. 'total_deaths_per_million' 15. 'new_deaths_per_million' 16. 'new_deaths_smoothed_per_million' 17. 'icu_patients' 18. 'icu_patients_per_million' 19. 'hosp_patients' 20. 'hosp_patients_per_million' 21. 'weekly_icu_admissions' 22. 'weekly_icu_admissions_per_million' 23. 'weekly_hosp_admissions' 24. 'weekly_hosp_admissions_per_million' 25. 'total_tests' 26. 'new_tests' 27. 'total_tests_per_thousand' 28. 'new_tests_per_thousand' 29. 'new_tests_smoothed' 30. 'new_tests_smoothed_per_thousand' 31. 'tests_per_case' 32. 'positive_rate' 33. 'tests_units' 34. 'stringency_index' 35. 'population' 36. 'population_density' 37. 'median_age' 38. 'aged_65_older' 39. 'aged_70_older' 40. 'gdp_per_capita' 41. 'extreme_poverty' 42. 'cardiovasc_death_rate' 43. 'diabetes_prevalence' 44. 'female_smokers' 45. 'male_smokers' 46. 'handwashing_facilities' 47. 'hospital_beds_per_thousand' 48. 'life_expectancy' 49. 'human_development_index'

[5]:
```
# Zusammenfassung der Datentabelle
summary(my_data) # .transpose()  # für bessere konsumierbarkeit
```

```
  iso_code          continent           location            date
Length:55247       Length:55247       Length:55247       Length:55247
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character



 total_cases         new_cases      new_cases_smoothed  total_deaths
Min.   :       1   Min.   : -8261   Min.   :  -552.0   Min.   :      1
1st Qu.:     164   1st Qu.:     0   1st Qu.:     0.9   1st Qu.:     12
Median :    1947   Median :    13   Median :    18.4   Median :     81
Mean   :  154401   Mean   :  1818   Mean   :  1782.3   Mean   :   6563
3rd Qu.:   20040   3rd Qu.:   218   3rd Qu.:   228.1   3rd Qu.:    677
Max.   :49373235   Max.   :584128   Max.   :520994.1   Max.   :1243083
NA's   :3632       NA's   :923      NA's   :1723       NA's   :12808
  new_deaths       new_deaths_smoothed total_cases_per_million
Min.   :-1918.00   Min.   :-232.143    Min.   :    0.00
1st Qu.:    0.00   1st Qu.:   0.000    1st Qu.:   83.52
Median :    0.00   Median :   0.286    Median :  547.83
Mean   :   45.76   Mean   :  45.514    Mean   : 2862.20
3rd Qu.:    4.00   3rd Qu.:   3.857    3rd Qu.: 3178.61
Max.   :10491.00   Max.   :7565.000    Max.   :66459.59
NA's   :923        NA's   :1723        NA's   :3909
new_cases_per_million new_cases_smoothed_per_million total_deaths_per_million
Min.   :-2212.545     Min.   :-269.978               Min.   :   0.00
1st Qu.:    0.000     1st Qu.:   0.237               1st Qu.:   3.76
Median :    2.067     Median :   3.703               Median :  19.41
Mean   :   35.801     Mean   :  34.522               Mean   :  91.06
3rd Qu.:   24.336     3rd Qu.:  27.509               3rd Qu.:  81.37
Max.   : 8652.658     Max.   :2472.188               Max.   :1237.55
NA's   :987           NA's   :1788                   NA's   :13070
new_deaths_per_million new_deaths_smoothed_per_million  icu_patients
Min.   :-67.9010       Min.   :-9.6780                  Min.   :   0.0
1st Qu.:  0.0000       1st Qu.: 0.0000                  1st Qu.:   6.0
Median :  0.0000       Median : 0.0260                  Median :  35.0
Mean   :  0.6358       Mean   : 0.6253                  Mean   : 255.1
3rd Qu.:  0.2750       3rd Qu.: 0.3880                  3rd Qu.: 149.2
Max.   :215.3820       Max.   :63.1400                  Max.   :7019.0
NA's   :987            NA's   :1788                     NA's   :51375
icu_patients_per_million hosp_patients     hosp_patients_per_million
Min.   :  0.00           Min.   :    0.0   Min.   :  0.00
1st Qu.:  1.51           1st Qu.:   31.0   1st Qu.:  9.20
Median :  4.89           Median :  160.0   Median : 30.94
Mean   : 11.56           Mean   : 1784.4   Mean   : 64.72
```

```
3rd Qu.: 12.63           3rd Qu.:  827.5   3rd Qu.: 73.83
Max.   :110.88           Max.   :33004.0   Max.   :716.50
NA's   :51375            NA's   :50684     NA's   :50684
weekly_icu_admissions weekly_icu_admissions_per_million weekly_hosp_admissions
Min.   :   0.00       Min.   : 0.00                     Min.   :    0.00
1st Qu.:   2.00       1st Qu.: 0.37                     1st Qu.:   13.92
Median :  10.38       Median : 1.37                     Median :  100.71
Mean   : 148.97       Mean   : 4.74                     Mean   : 1374.19
3rd Qu.:  81.33       3rd Qu.: 4.48                     3rd Qu.:  639.51
Max.   :4375.41       Max.   :67.03                     Max.   :33314.87
NA's   :54919         NA's   :54919                     NA's   :54680
weekly_hosp_admissions_per_million  total_tests          new_tests
Min.   :   0.00                    Min.   :        1   Min.   : -28671
1st Qu.:   3.27                    1st Qu.:    59889   1st Qu.:   1042
Median :  12.11                    Median :   251832   Median :   3636
Mean   :  48.98                    Mean   :  2145159   Mean   :  27462
3rd Qu.:  36.91                    3rd Qu.:  1015154   3rd Qu.:  13610
Max.   :1731.75                    Max.   :160000000   Max.   :1492409
NA's   :54680                      NA's   :34203       NA's   :34466
total_tests_per_thousand new_tests_per_thousand new_tests_smoothed
Min.   :   0.00          Min.   :-2.46          Min.   :      0
1st Qu.:   3.83          1st Qu.: 0.07          1st Qu.:   1105
Median :  19.72          Median : 0.34          Median :   3952
Mean   :  74.00          Mean   : 0.88          Mean   :  26401
3rd Qu.:  80.81          3rd Qu.: 1.03          3rd Qu.:  14518
Max.   :1787.74          Max.   :26.04          Max.   :1209474
NA's   :34203            NA's   :34466          NA's   :31753
new_tests_smoothed_per_thousand tests_per_case     positive_rate
Min.   : 0.00                   Min.   :    1.53   Min.   :0.00
1st Qu.: 0.07                   1st Qu.:   10.96   1st Qu.:0.01
Median : 0.34                   Median :   30.52   Median :0.03
Mean   : 0.87                   Mean   :  185.60   Mean   :0.07
3rd Qu.: 1.03                   3rd Qu.:   98.21   3rd Qu.:0.09
Max.   :19.15                   Max.   :45864.00   Max.   :0.65
NA's   :31753                   NA's   :33495      NA's   :33129
tests_units        stringency_index   population        population_density
Length:55247       Min.   :  0.00   Min.   :8.090e+02   Min.   :    0.137
Class :character   1st Qu.: 39.23   1st Qu.:1.327e+06   1st Qu.:   37.728
Mode  :character   Median : 61.11   Median :8.279e+06   Median :   88.125
                   Mean   : 56.93   Mean   :8.610e+07   Mean   :  360.961
                   3rd Qu.: 78.24   3rd Qu.:2.983e+07   3rd Qu.:  214.243
                   Max.   :100.00   Max.   :7.795e+09   Max.   :19347.500
                   NA's   :9739     NA's   :313         NA's   :2893
  median_age    aged_65_older    aged_70_older    gdp_per_capita
Min.   :15.10   Min.   : 1.144   Min.   : 0.526   Min.   :   661.2
1st Qu.:23.20   1st Qu.: 3.552   1st Qu.: 2.085   1st Qu.:  5321.4
Median :31.10   Median : 6.981   Median : 4.393   Median : 14048.9
Mean   :31.22   Mean   : 9.196   Mean   : 5.812   Mean   : 20678.5
```

```
3rd Qu.:39.70   3rd Qu.:14.762   3rd Qu.: 9.395   3rd Qu.: 31400.8
Max.   :48.20   Max.   :27.049   Max.   :18.493   Max.   :116935.6
NA's   :6090    NA's   :6829     NA's   :6346     NA's   :6737
extreme_poverty cardiovasc_death_rate diabetes_prevalence female_smokers
Min.   : 0.10   Min.   : 79.37        Min.   : 0.990      Min.   : 0.10
1st Qu.: 0.50   1st Qu.:156.14        1st Qu.: 5.310      1st Qu.: 1.90
Median : 2.00   Median :238.34        Median : 7.110      Median : 6.40
Mean   :12.38   Mean   :252.36        Mean   : 8.063      Mean   :10.76
3rd Qu.:18.10   3rd Qu.:318.99        3rd Qu.:10.390      3rd Qu.:19.60
Max.   :77.60   Max.   :724.42        Max.   :30.530      Max.   :44.00
NA's   :22910   NA's   :6114          NA's   :4323        NA's   :16988
 male_smokers   handwashing_facilities hospital_beds_per_thousand
Min.   : 7.70   Min.   : 1.19          Min.   : 0.100
1st Qu.:21.40   1st Qu.:21.22          1st Qu.: 1.300
Median :31.40   Median :52.23          Median : 2.500
Mean   :32.64   Mean   :52.16          Mean   : 3.093
3rd Qu.:40.90   3rd Qu.:83.74          3rd Qu.: 4.200
Max.   :78.10   Max.   :99.00          Max.   :13.800
NA's   :17481   NA's   :32018          NA's   :10974
life_expectancy human_development_index
Min.   :53.28   Min.   :0.354
1st Qu.:69.87   1st Qu.:0.601
Median :75.40   Median :0.752
Mean   :73.95   Mean   :0.723
3rd Qu.:79.38   3rd Qu.:0.847
Max.   :86.75   Max.   :0.953
NA's   :1018    NA's   :7817
```

[6]:
```
describe(my_data) # taken from psych package
```

A psych: 49 × 13

	vars	n	mean	sd	me
	<int>	<dbl>	<dbl>	<dbl>	<d
iso_code*	1	55247	1.063311e+02	6.154047e+01	106
continent*	2	55247	3.668778e+00	1.421770e+00	4.0
location*	3	55247	1.070614e+02	6.177495e+01	107
date*	4	55247	1.803795e+02	8.015154e+01	183
total_cases	5	51615	1.544013e+05	1.551656e+06	194
new_cases	6	54324	1.817666e+03	1.672939e+04	13.
new_cases_smoothed	7	53524	1.782345e+03	1.621142e+04	18.
total_deaths	8	42439	6.562907e+03	5.325832e+04	81.
new_deaths	9	54324	4.576235e+01	3.800177e+02	0.0
new_deaths_smoothed	10	53524	4.551382e+01	3.671447e+02	0.2
total_cases_per_million	11	51338	2.862199e+03	5.629671e+03	547
new_cases_per_million	12	54260	3.580082e+01	1.259550e+02	2.0
new_cases_smoothed_per_million	13	53459	3.452198e+01	9.116582e+01	3.7
total_deaths_per_million	14	42177	9.105843e+01	1.757604e+02	19.
new_deaths_per_million	15	54260	6.358497e-01	2.964449e+00	0.0
new_deaths_smoothed_per_million	16	53459	6.253297e-01	1.898826e+00	0.0
icu_patients	17	3872	2.551433e+02	7.259169e+02	35.
icu_patients_per_million	18	3872	1.156340e+01	1.788121e+01	4.8
hosp_patients	19	4563	1.784365e+03	4.740088e+03	160
hosp_patients_per_million	20	4563	6.472256e+01	9.578770e+01	30.
weekly_icu_admissions	21	328	1.489701e+02	4.685762e+02	10.
weekly_icu_admissions_per_million	22	328	4.743933e+00	8.964706e+00	1.3
weekly_hosp_admissions	23	567	1.374186e+03	3.716430e+03	100
weekly_hosp_admissions_per_million	24	567	4.897691e+01	1.346709e+02	12.
total_tests	25	21044	2.145159e+06	9.756498e+06	251
new_tests	26	20781	2.746223e+04	1.113352e+05	363
total_tests_per_thousand	27	21044	7.400298e+01	1.476119e+02	19.
new_tests_per_thousand	28	20781	8.831059e-01	1.636691e+00	0.3
new_tests_smoothed	29	23494	2.640139e+04	1.025570e+05	395
new_tests_smoothed_per_thousand	30	23494	8.657664e-01	1.514974e+00	0.3
tests_per_case	31	21752	1.856019e+02	8.894804e+02	30.
positive_rate	32	22118	6.597550e-02	8.723641e-02	0.0
tests_units*	33	55247	2.441436e+00	1.788981e+00	1.0
stringency_index	34	45508	5.692708e+01	2.625599e+01	61.
population	35	54934	8.609906e+07	6.037173e+08	827
population_density	36	52354	3.609614e+02	1.643391e+03	88.
median_age	37	49157	3.121666e+01	9.042700e+00	31.
aged_65_older	38	48418	9.195618e+00	6.302385e+00	6.9
aged_70_older	39	48901	5.811629e+00	4.300141e+00	4.3
gdp_per_capita	40	48510	2.067854e+04	2.033443e+04	140
extreme_poverty	41	32337	1.237764e+01	1.939279e+01	2.0
cardiovasc_death_rate	42	49133	2.523565e+02	1.174421e+02	238
diabetes_prevalence	43	50924	8.063169e+00	4.174354e+00	7.1
female_smokers	44	38259	1.075674e+01	1.047286e+01	6.4
male_smokers	45	37766	3.264068e+01	1.344849e+01	31.
handwashing_facilities	46	23229	5.216055e+01	3.163940e+01	52.
hospital_beds_per_thousand	47	44273	3.093203e+00	2.515747e+00	2.5
life_expectancy	48	54229	7.395480e+01	7.393933e+00	75.
human_development_index	49	47430	7.227389e-01	1.532037e-01	0.7

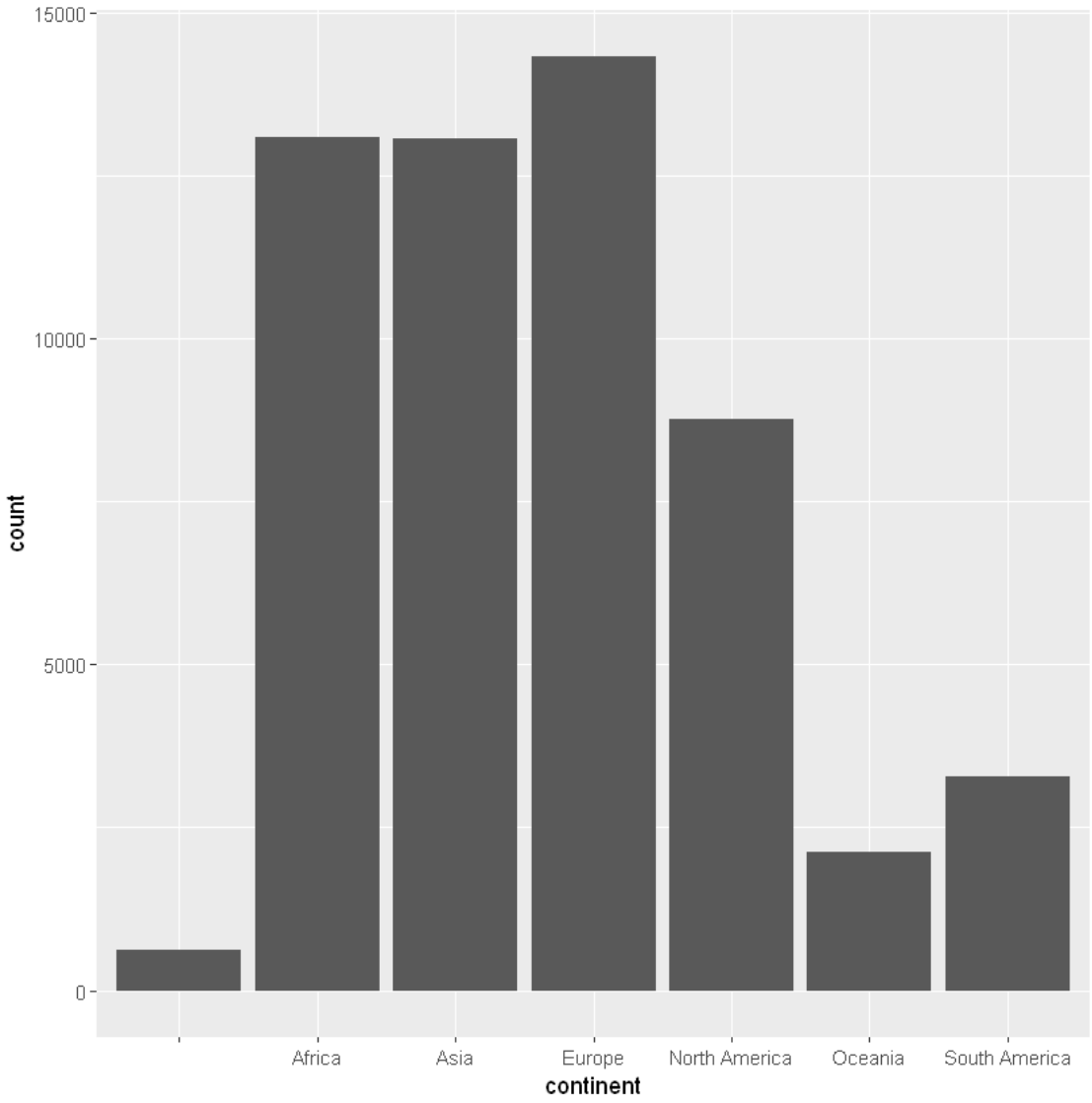
[7]:
```
# Übersicht der vorhandenen Kontinente
unique(my_data$continent)
```

1. 'Asia' 2. 'Europe' 3. 'Africa' 4. 'North America' 5. 'South America' 6. 'Oceania' 7. ''

[8]:
```
# wie viele einträge haben wir pro kontinent?
table(my_data$continent)
```

```
                Africa           Asia         Europe North America
           626    13087          13061          14325           8756
       Oceania South America
          2118           3274
```

[9]:
```
# wie viele einträge haben wir pro kontinent? als Graphik
ggplot(data = my_data) + geom_bar(mapping = aes(x = continent))
```



[10]:
```
# Ausgabe eines bestimmten parameters (hier total_cases) gruppiert nach
 ↪continent
describeBy(my_data$total_cases,my_data$continent)
```

```
 Descriptive statistics by group
group:
   vars   n    mean       sd median trimmed    mad min      max    range skew
X1    1 590 6756146 12156164    705 3779983 957.76  10 49373235 49373225 1.85
   kurtosis       se
X1     2.31 500461.4
------------------------------------------------------------
group: Africa
   vars     n    mean       sd median trimmed     mad min    max  range skew
X1    1 12784 13220.2 61321.84   1378 3243.69 1995.58   1 734175 734174  9.3
   kurtosis     se
X1    92.95 542.35
------------------------------------------------------------
group: Asia
   vars     n     mean       sd median  trimmed     mad min     max   range
X1    1 11740 94046.59 521040.7   5567 25531.95 8225.46   1 8462080 8462079
    skew kurtosis      se
X1 12.11   159.25 4808.81
------------------------------------------------------------
group: Europe
   vars     n     mean       sd median  trimmed     mad min     max   range
X1    1 12838 57378.79 160655.4   4474 18722.87 6556.06   1 1733440 1733439
   skew kurtosis     se
X1 5.27    33.77 1417.9
------------------------------------------------------------
group: North America
   vars    n     mean       sd median trimmed    mad min     max   range skew
X1    1 8559 127955.4 791484.8    140 5459.11 198.67   1 9739545 9739544 8.34
   kurtosis      se
X1    74.05 8555.22
------------------------------------------------------------
group: Oceania
   vars    n   mean      sd median trimmed   mad min   max range skew kurtosis
X1    1 1975 2204.8 5803.84     61  623.88 78.58   1 27645 27644 3.46    11.27
      se
X1 130.6
------------------------------------------------------------
group: South America
   vars    n     mean       sd median trimmed      mad min     max   range skew
X1    1 3129 279332.2 813340.2   8225 86055.2 12179.56   1 5590025 5590024 4.65
```

```
   kurtosis       se
X1    22.67 14540.17
```

[11]:

```
# Ausgabe eines bestimmten parameters (hier total_cases) gruppoert nach␣
 ↪continent
describeBy(my_data$weekly_icu_admissions_per_million,my_data$continent)
```

```
Warning message in min(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für min; gebe Inf zurück"
Warning message in max(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für max; gebe -Inf zurück"
Warning message in min(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für min; gebe Inf zurück"
Warning message in max(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für max; gebe -Inf zurück"
Warning message in min(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für min; gebe Inf zurück"
Warning message in max(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für max; gebe -Inf zurück"
Warning message in min(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für min; gebe Inf zurück"
Warning message in max(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für max; gebe -Inf zurück"
Warning message in min(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für min; gebe Inf zurück"
Warning message in max(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für max; gebe -Inf zurück"
Warning message in min(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für min; gebe Inf zurück"
Warning message in max(x, na.rm = na.rm):
"kein nicht-fehlendes Argument für max; gebe -Inf zurück"
```

```

 Descriptive statistics by group
group: 
   vars n mean sd median trimmed mad min  max range skew kurtosis se
X1    1 0  NaN NA     NA     NaN  NA Inf -Inf  -Inf   NA       NA NA
------------------------------------------------------------
group: Africa
   vars n mean sd median trimmed mad min  max range skew kurtosis se
X1    1 0  NaN NA     NA     NaN  NA Inf -Inf  -Inf   NA       NA NA
------------------------------------------------------------
group: Asia
   vars n mean sd median trimmed mad min  max range skew kurtosis se
X1    1 0  NaN NA     NA     NaN  NA Inf -Inf  -Inf   NA       NA NA
------------------------------------------------------------
group: Europe
   vars    n mean    sd median trimmed   mad min    max range skew kurtosis    se
```

```
X1    1 328 4.74 8.96   1.37    2.59 1.99   0 67.03 67.03 3.64    16.13 0.49
------------------------------------------------------------
group: North America
   vars n mean sd median trimmed mad min  max range skew kurtosis se
X1    1 0  NaN NA     NA      NaN  NA Inf -Inf  -Inf   NA       NA NA
------------------------------------------------------------
group: Oceania
   vars n mean sd median trimmed mad min  max range skew kurtosis se
X1    1 0  NaN NA     NA      NaN  NA Inf -Inf  -Inf   NA       NA NA
------------------------------------------------------------
group: South America
   vars n mean sd median trimmed mad min  max range skew kurtosis se
X1    1 0  NaN NA     NA      NaN  NA Inf -Inf  -Inf   NA       NA NA
```

[12]:
```
# Auswahl von Daten
# Aggregation der GDPR Daten pro Land
# (die Daten sind für jeden zeitlichen Eintrag gleich, dadurch brauchen wir nur␣
 ↪jeweils den ersten Eintrag)
data_gdp <- my_data[!duplicated(my_data$location), ]

# Nur daten von Deutschland
data_DE <- my_data[my_data$location == "Germany", ]
```

1.2.4 Analyse fehlender Werte

[13]:
```
# wie viel prozent der Daten fehlen pro Parameter?
apply(my_data, 2, function(col)sum(is.na(col))/length(col))
```

iso_code 0 **continent** 0 **location** 0 **date** 0 **total_cases** 0.0657411262149981 **new_cases** 0.0167067895089326 **new_cases_smoothed** 0.0311872137853639 **total_deaths** 0.231831592665665 **new_deaths** 0.0167067895089326 **new_deaths_smoothed** 0.0311872137853639 **total_cases_per_million** 0.0707549731207124 **new_cases_per_million** 0.0178652234510471 **new_cases_smoothed_per_million** 0.032363748257824 **total_deaths_per_million** 0.236573931616196 **new_deaths_per_million** 0.0178652234510471 **new_deaths_smoothed_per_million** 0.032363748257824 **icu_patients** 0.929914746502073 **icu_patients_per_million** 0.929914746502073 **hosp_patients** 0.917407280033305 **hosp_patients_per_million** 0.917407280033305 **weekly_icu_admissions** 0.994063026046663 **weekly_icu_admissions_per_million** 0.994063026046663 **weekly_hosp_admissions** 0.989736999294079 **weekly_hosp_admissions_per_million** 0.989736999294079 **total_tests** 0.619092439408475 **new_tests** 0.623852878889351 **total_tests_per_thousand** 0.619092439408475 **new_tests_per_thousand** 0.623852878889351 **new_tests_smoothed** 0.574746140061904 **new_tests_smoothed_per_thousand** 0.574746140061904 **tests_per_case** 0.606277263923833 **positive_rate** 0.599652469817366 **tests_units** 0 **stringency_index** 0.176281065035206 **population** 0.00566546599815375 **population_density** 0.0523648342896447 **median_age** 0.110232229804333

aged_65_older 0.123608521729687 **aged_70_older** 0.114865965572791
gdp_per_capita 0.121943272937897 **extreme_poverty** 0.414683150216301
cardiovasc_death_rate 0.110666642532626 **diabetes_prevalence** 0.0782485926837656
female_smokers 0.307491809510019 **male_smokers** 0.316415370970369
handwashing_facilities 0.579542780603472 **hospital_beds_per_thousand** 0.198635220011946 **life_expectancy** 0.0184263398917588 **human_development_index** 0.141491845711079

[14]:
```
# wie viel fehlende werte (in prozent) haben wir in den Daten zu Europa?
data_EU  <- my_data[my_data$continent == "Europe", ]
apply(data_EU, 2, function(col)sum(is.na(col))/length(col))
```

iso_code 0 **continent** 0 **location** 0 **date** 0 **total_cases** 0.103804537521815 **new_cases** 0.0104712041884817 **new_cases_smoothed** 0.0231762652705061 **total_deaths** 0.217102966841187 **new_deaths** 0.0104712041884817 **new_deaths_smoothed** 0.0231762652705061 **total_cases_per_million** 0.103804537521815
new_cases_per_million 0.0104712041884817
new_cases_smoothed_per_million 0.0231762652705061
total_deaths_per_million 0.217102966841187 **new_deaths_per_million** 0.0104712041884817 **new_deaths_smoothed_per_million** 0.0231762652705061
icu_patients 0.729703315881326 **icu_patients_per_million** 0.729703315881326
hosp_patients 0.681465968586387 **hosp_patients_per_million** 0.681465968586387
weekly_icu_admissions 0.977102966841187
weekly_icu_admissions_per_million 0.977102966841187
weekly_hosp_admissions 0.960418848167539
weekly_hosp_admissions_per_million 0.960418848167539 **total_tests** 0.489633507853403 **new_tests** 0.491169284467714 **total_tests_per_thousand** 0.489633507853403 **new_tests_per_thousand** 0.491169284467714
new_tests_smoothed 0.42282722513089 **new_tests_smoothed_per_thousand** 0.42282722513089 **tests_per_case** 0.435532286212915 **positive_rate** 0.434973821989529
tests_units 0 **stringency_index** 0.20369982547993 **population** 0 **population_density** 0.0493542757417103 **median_age** 0.193507853403141 **aged_65_older** 0.193507853403141
aged_70_older 0.211378708551483 **gdp_per_capita** 0.154973821989529
extreme_poverty 0.399301919720768 **cardiovasc_death_rate** 0.176404886561955
diabetes_prevalence 0.115741710296684 **female_smokers** 0.19825479930192
male_smokers 0.19825479930192 **handwashing_facilities** 0.9482722513089
hospital_beds_per_thousand 0.132844677137871 **life_expectancy** 0.0492146596858639 **human_development_index** 0.159441535776614

1.2.5 Ausreißer

[15]:
```
average_gdp = mean(data_gdp$gdp_per_capita, na.rm=TRUE)
std_gdp = sd(data_gdp$gdp_per_capita, na.rm=TRUE)
data_gdp[which(data_gdp$gdp_per_capita > (average_gdp + 2*std_gdp)), ]
# average_gdp + 2*std_gdp
```

A data.frame: 8 × 49

	iso_code <chr>	continent <chr>	location <chr>	date <chr>	total_cases <dbl>	new_ca <dbl>
7601	BRN	Asia	Brunei	2020-03-10	1	1
24361	IRL	Europe	Ireland	2019-12-31	NA	0
27290	KWT	Asia	Kuwait	2019-12-31	NA	0
29853	LUX	Europe	Luxembourg	2019-12-31	NA	0
37196	NOR	Europe	Norway	2019-12-31	NA	0
40398	QAT	Asia	Qatar	2019-12-31	NA	0
44021	SGP	Asia	Singapore	2019-12-31	NA	0
51291	ARE	Asia	United Arab Emirates	2019-12-31	NA	0

[16]:
```
data_gdp[which(data_gdp$gdp_per_capita < (average_gdp - 2*std_gdp)), ]
```

```
Warning message in cbind(parts$left, ellip_h, parts$right, deparse.level = 0L):
"number of rows of result is not a multiple of vector length (arg 2)"
Warning message in cbind(parts$left, ellip_h, parts$right, deparse.level = 0L):
"number of rows of result is not a multiple of vector length (arg 2)"
Warning message in cbind(parts$left, ellip_h, parts$right, deparse.level = 0L):
"number of rows of result is not a multiple of vector length (arg 2)"
Warning message in cbind(parts$left, ellip_h, parts$right, deparse.level = 0L):
"number of rows of result is not a multiple of vector length (arg 2)"
```

A data.frame: 0 × 49

iso_code <chr>	continent <chr>	location <chr>	date <chr>	total_cases <dbl>	new_cases <dbl>	new_cases_smoothe <dbl>

1.2.6 Einfache visualisierungen - Verteilungsanalyse

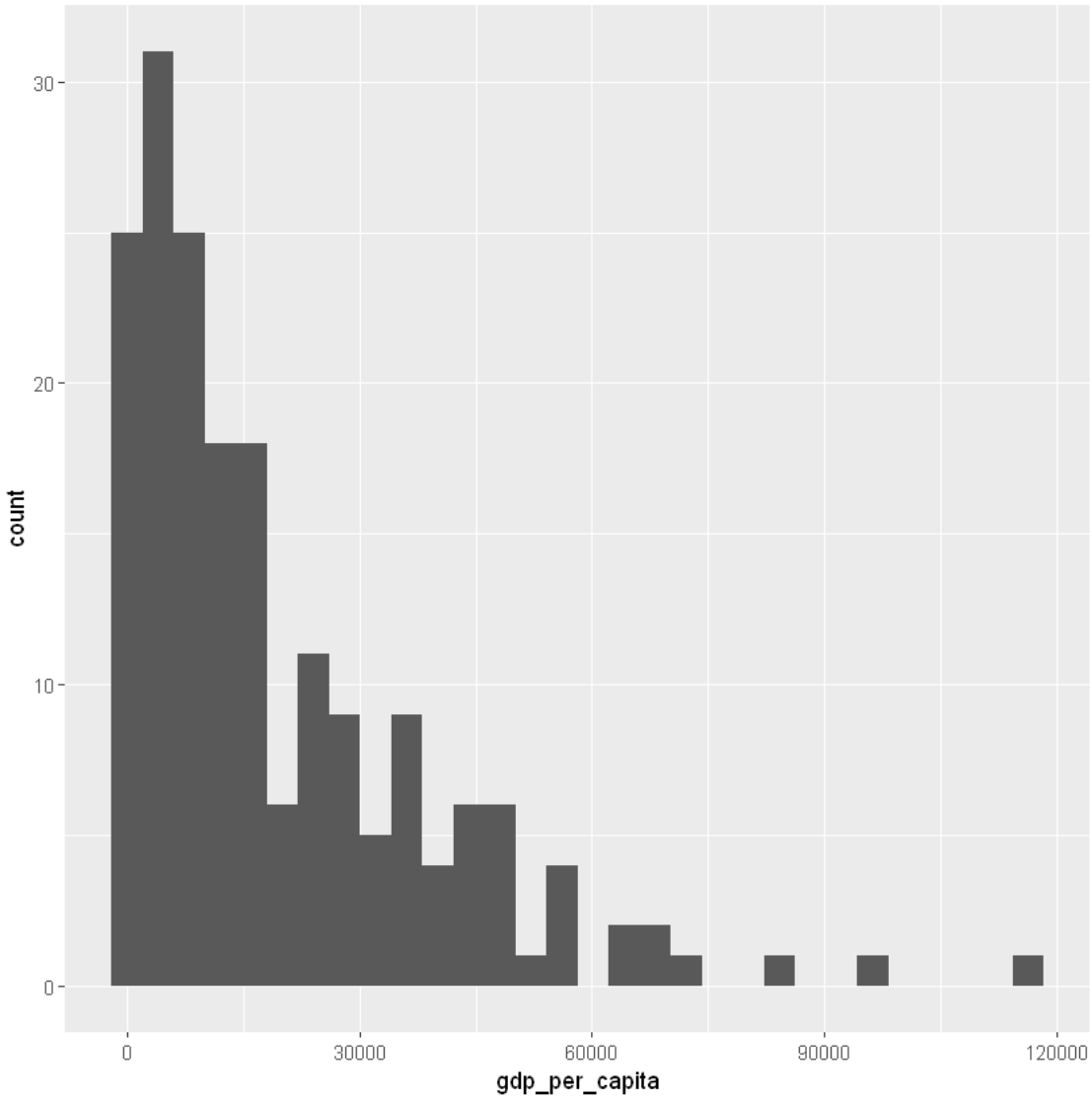
[17]:
```
ggplot(data = data_gdp, aes(x=gdp_per_capita)) + geom_histogram()
ggsave(file="test.svg")
```

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

Warning message:
"Removed 29 rows containing non-finite values (stat_bin)."
Saving 6.67 x 6.67 in image

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

Warning message:
"Removed 29 rows containing non-finite values (stat_bin)."
```


30
20
10
0
count
0
30000
60000
90000
120000
gdp_per_capita

[ ]: