Assignment 3

Kine Maakestad Susann Sivertsen

Svar på spørsmål

Spørsmål 1

Filen ddf_concepts.csv inneholder ingen verdier. Det den derimot inneholder er prosentvis av arbeidsledighet, hvor mange barn som har dødd av en alder av 1-59 måneder og nyfødte barn som har dødd.

Spørsmål 2

I denne filen inneholder det også ingen verdier, men filen inneholder land, og hvilken av disse landene som har høy inntekt, middels inntekt og lav inntekt. Den har også med hvor landene ligger i verden, for eksempel så ligger Afghanistan i Asia, og de spesifiserer også at det landet ligger i Sør Asia.

Spørsmål 3

Ddf–entities–geo–un_sdg_region.csv inneholder de forskjellige regionene og om de er TRUE eller FALSE.

Spørsmål 4

Gapminder pakken inneholder 6 variabler. Disse er:

- Country: faktor med 142 nivåer
- Continent: faktor med 5 nivåer
- Year: områder fra 1952 til 2007 med trinn på 5 år
- Pop: populasjon
- gdpPercap: BNP per innbygger (US$, inflasjon-justert

Australia og New Zealand ligger i kontinentet Asia.

Spørsmål 5

Her laster vi inn et nytt datasett, og skal deretter flytte Australia og New Zealand fra Asia til Oseania.

```
g_c <- read_csv("data/ddf--entities--geo--country.csv")
```

```
## Rows: 273 Columns: 22

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): country, g77_and_oecd_countries, income_3groups, income_groups, is...
## dbl  (3): iso3166_1_numeric, latitude, longitude
## lgl  (2): is--country, un_state

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
print(g_c)
```

```
## # A tibble: 273 x 22
##    country   g77_and_oecd_countries income_3groups income_groups   `is--country`
##    <chr>     <chr>                  <chr>          <chr>           <lgl>
##  1 abkh      others                 <NA>           <NA>            TRUE
##  2 abw       others                 high_income    high_income     TRUE
##  3 afg       g77                    low_income     low_income      TRUE
##  4 ago       g77                    middle_income  lower_middle_i~ TRUE
##  5 aia       others                 <NA>           <NA>            TRUE
##  6 akr_a_dhe others                 <NA>           <NA>            TRUE
##  7 ala       others                 <NA>           <NA>            TRUE
##  8 alb       others                 middle_income  upper_middle_i~ TRUE
##  9 and       others                 high_income    high_income     TRUE
## 10 ant       others                 <NA>           <NA>            TRUE
## # ... with 263 more rows, and 17 more variables: iso3166_1_alpha2 <chr>,
## #   iso3166_1_alpha3 <chr>, iso3166_1_numeric <dbl>, iso3166_2 <chr>,
## #   landlocked <chr>, latitude <dbl>, longitude <dbl>,
## #   main_religion_2008 <chr>, name <chr>, un_sdg_ldc <chr>,
## #   un_sdg_region <chr>, un_state <lgl>, unhcr_region <chr>,
## #   unicef_region <chr>, unicode_region_subtag <chr>, world_4region <chr>,
## #   world_6region <chr>
```

```
spec(g_c)
```

```
## cols(
##   country = col_character(),
##   g77_and_oecd_countries = col_character(),
##   income_3groups = col_character(),
##   income_groups = col_character(),
##   `is--country` = col_logical(),
##   iso3166_1_alpha2 = col_character(),
##   iso3166_1_alpha3 = col_character(),
##   iso3166_1_numeric = col_double(),
##   iso3166_2 = col_character(),
##   landlocked = col_character(),
##   latitude = col_double(),
##   longitude = col_double(),
##   main_religion_2008 = col_character(),
##   name = col_character(),
##   un_sdg_ldc = col_character(),
##   un_sdg_region = col_character(),
##   un_state = col_logical(),
##   unhcr_region = col_character(),
##   unicef_region = col_character(),
##   unicode_region_subtag = col_character(),
##   world_4region = col_character(),
##   world_6region = col_character()
## )
```

I denne har vi flyttet Australia og New Zealand til Oseania, og bare inkludert landene som har iso3166_1_alpha3 koden.

```
g_c <- g_c%>%
  mutate(continent = case_when(
    world_4region == "asia" & un_sdg_region %in% c("un_australia_and_new_zealand","un_oceania_exc_austra
    world_4region == "asia" & !(un_sdg_region %in% c("un_australia_and_new_zealand", "un_oceania_exc_au
    world_4region == "africa" ~ "Africa",
    world_4region == "americas" ~ "Americas",
    world_4region == "europe" ~ "Europe")
  ) %>%
  filter(!is.na(iso3166_1_alpha3))
```

Spørsmål 6

a

I dette nye datasettet finner vi ut hvor mange land som er der nå.

```
length(unique(g_c$country))
```

```
## [1] 247
```

b

Her ser man hvor mange land det nå er i hver kontinent.

```
g_c %>%
  group_by(continent) %>%
  summarise(countries = length(unique(country)))
```

```
## # A tibble: 5 x 2
##   continent countries
##   <chr>         <int>
## 1 Africa           59
## 2 Americas         55
## 3 Asia             47
## 4 Europe           58
## 5 Oceania          28
```

Spørsmål 7

Vi laster enda ett nytt datasett, og endrer **time** variabelen til **date**. Videre endrer vi **time** til **year**.

```
lifeExp <- read_csv("data/countries-etc-datapoints/ddf--datapoints--life_expectancy_years--by--geo--time
col_types = cols(time = col_date(format = "%Y")))
lifeExp <- lifeExp %>%
  rename(year = time)
names(lifeExp)
```

```
## [1] "geo"                  "year"                 "life_expectancy_years"
```

```
length(unique(lifeExp$geo))
```

```
## [1] 195
```

Spørsmål 8

```
length(unique(lifeExp$geo))
```

```
## [1] 195
```

Det er 195 land som har informasjon om LifeExp.

Spørsmål 9

Her reduserer vi variablene til **country, name, iso3166_1_alpha3, un_sdg_region, world_4region, continent, world_6region**.

```
g_c <- g_c %>%
  select(country, name, iso3166_1_alpha3, un_sdg_region, world_4region, continent, world_6region) %>%
  left_join(lifeExp, by = c("country" = "geo"))
names(g_c)
```

```
## [1] "country"              "name"                 "iso3166_1_alpha3"
## [4] "un_sdg_region"        "world_4region"        "continent"
## [7] "world_6region"        "year"                 "life_expectancy_years"
```

Spørsmål 10

```
g_c_min <- g_c %>%
  group_by(country) %>%
  summarise(min_year = min(year))
table(g_c_min$min_year)
```

```
##
## 1800-01-01 1950-01-01
##        186          9
```

Den første observasjonen av lifeExp til de forskjellige landene er 186.

Spørsmål 11

Under kan man se at de 9 landene som bare har **life expentancy** data fra 1950.

```
g_c_min %>%
  filter(min_year == "1950-01-01")
```

```
## # A tibble: 9 x 2
##   country min_year
##   <chr>   <date>
## 1 and     1950-01-01
## 2 dma     1950-01-01
## 3 kna     1950-01-01
## 4 mco     1950-01-01
## 5 mhl     1950-01-01
## 6 nru     1950-01-01
## 7 plw     1950-01-01
## 8 smr     1950-01-01
## 9 tuv     1950-01-01
```

Spørsmål 12

Leser her inn et nytt datasett, og endrer til dato.

```
pop <- read_csv("data/countries-etc-datapoints/ddf--datapoints--population_total--by--geo--time.csv",
                col_types = cols(time = col_date(format = "%Y")))
```

```
g_c <- g_c %>%
  left_join(pop, by = c("country" = "geo", "year" ="time"))

rm(pop)
```

Spørsmål 13

Leser inn nytt datasett.

```
gdp_pc <- read_csv("data/countries-etc-datapoints/ddf--datapoints--gdppercapita_us_inflation_adjusted--
col_types = cols(time = col_date(format = "%Y")))
```

```
g_c <- g_c %>%
  left_join(gdp_pc, by = c("country" = "geo", "year" = "time"))
```

Endrer her variabel navnene.

```
g_c <- g_c %>%
  rename("lifeExp" = "life_expectancy_years") %>%
  rename("pop" = "population_total") %>%
  rename("gdpPercap" = "gdppercapita_us_inflation_adjusted")
names(g_c)
```

```
##  [1] "country"          "name"             "iso3166_1_alpha3" "un_sdg_region"
##  [5] "world_4region"    "continent"        "world_6region"    "year"
##  [9] "lifeExp"          "pop"              "gdpPercap"
```

Spørsmål 14

Her bruker vi dataene fra hver femte år.

```
t1 <- paste(c(seq(1800, 2015, by = 5), 2019), "01-01", sep = "-") %>%
  parse_date(format = "%Y-%m-%d")

g_c_5year <- g_c %>%
  filter(year %in% t1) %>%
  select(country, name, continent, year, lifeExp, pop, gdpPercap)

dim(g_c_5year)

g_c_min_yr_gdp <- g_c_5year %>%
  group_by(gdpPercap) %>%
  summarise (min_year = min(year))

g_c_min_yr_gdp %>%
  count(min_year = g_c_min_yr_gdp$min_year)
```

Spørsmål 15

Chunken under brukes for å vise hvilket år hvert land har innhentet BNP.

```
tmp <- g_c %>%
  filter (!is.na(gdpPercap)) %>%
  group_by(country) %>%
  summarise (nr=n()) %>%
  arrange((country))
```

Videre filtreres det ned til landene som har rapportert BNP over en lengre periode. I dette tilfelle de siste 60 årene:

```
g_c_5year <- tmp %>%
  filter(nr > 60)
```

Etter dette kan vi se at det er 84 land som har rapportert BNP de siste 60 årene.

Spørsmål 16

Først i denne oppgaven lager vi ett nytt datasett for å finne observasjonene for å finne antall land med verdier i tidsperioden 1960-2019. Her finner vi 191 land.

```
c_min_y <- g_c %>%
  filter (!is.na(gdpPercap)) %>%
  group_by(country) %>%
  summarise(min_year = min(year))

dim(c_min_y)
```

```
## [1] 191   2
```

Her ser vi at i my_gapminder_1960 datasettet er det 25886 observasjoner og 11 variabler.

```
c_min_y_60 <- c_min_y$country[c_min_y$min_year == "1960-01-01"]
my_gapminder_1960 <- g_c %>%
  filter(country %in% c_min_y_60)

dim(my_gapminder_1960)
```

```
## [1] 25886    11
```

Videre ser man at det er 86 land med registrert data:

```
length(unique(my_gapminder_1960$country))
```

```
## [1] 86
```

Videre finner vi oversikt over NA verdier:

```
(num_NA <- my_gapminder_1960[is.na(my_gapminder_1960$gdpPercap) == TRUE, ])
```

```
## # A tibble: 20,647 x 11
##    country name      iso3166_1_alpha3 un_sdg_region      world_4region continent
##    <chr>   <chr>     <chr>            <chr>              <chr>         <chr>
##  1 arg     Argentina ARG              un_latin_america_~ americas      Americas
##  2 arg     Argentina ARG              un_latin_america_~ americas      Americas
##  3 arg     Argentina ARG              un_latin_america_~ americas      Americas
##  4 arg     Argentina ARG              un_latin_america_~ americas      Americas
##  5 arg     Argentina ARG              un_latin_america_~ americas      Americas
##  6 arg     Argentina ARG              un_latin_america_~ americas      Americas
##  7 arg     Argentina ARG              un_latin_america_~ americas      Americas
##  8 arg     Argentina ARG              un_latin_america_~ americas      Americas
##  9 arg     Argentina ARG              un_latin_america_~ americas      Americas
## 10 arg     Argentina ARG              un_latin_america_~ americas      Americas
## # ... with 20,637 more rows, and 5 more variables: world_6region <chr>,
## #   year <date>, lifeExp <dbl>, pop <dbl>, gdpPercap <dbl>
```

Til slutt finner hvor mange land det er fra hvert kontinent i datasettet:

```
my_gapminder_1960 %>%
  distinct(country, continent) %>%
  group_by(continent) %>%
  count () %>%
  kable ()
```

continent	n
Africa	29
Americas	25
Asia	14
Europe	15

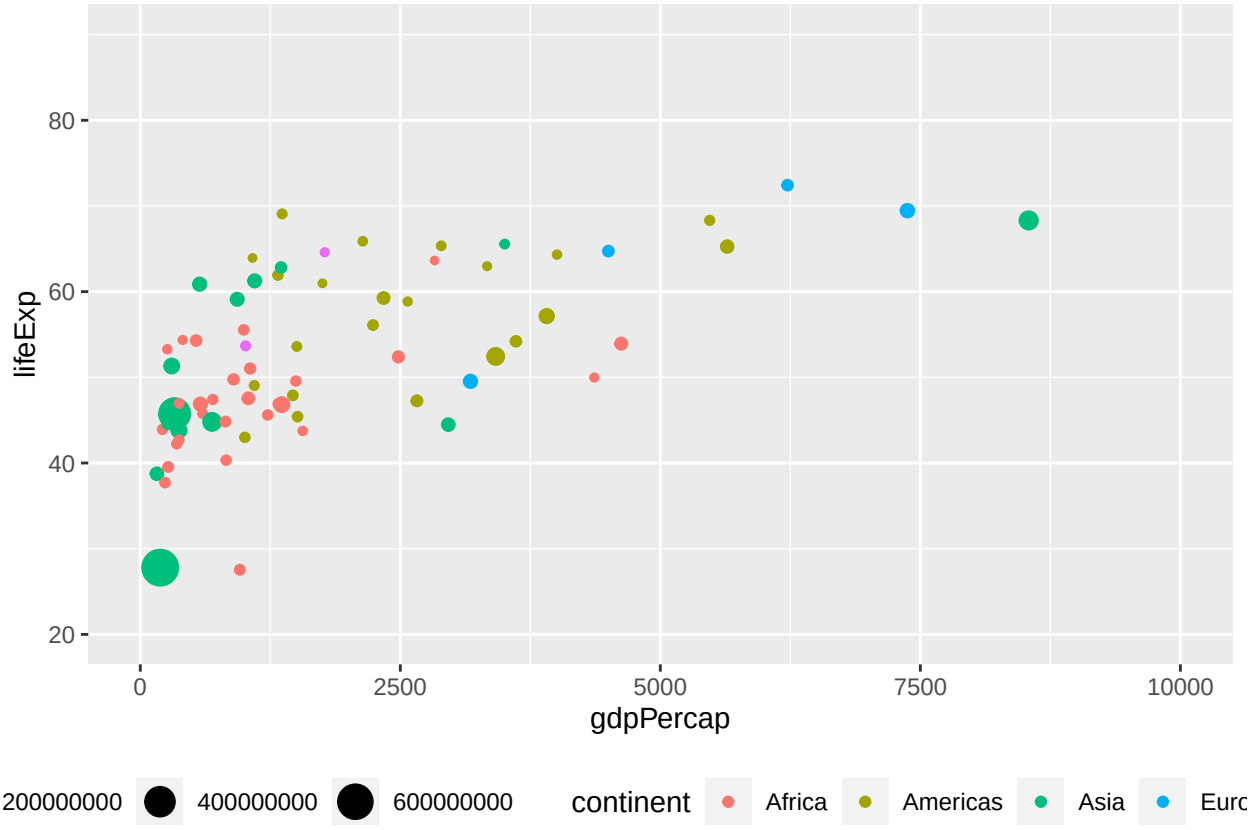
continent	n
Oceania	3

Spørsmål 17

Her bruker vi **ggplot()** for årene 1960, 1980, 2000 og 2019.

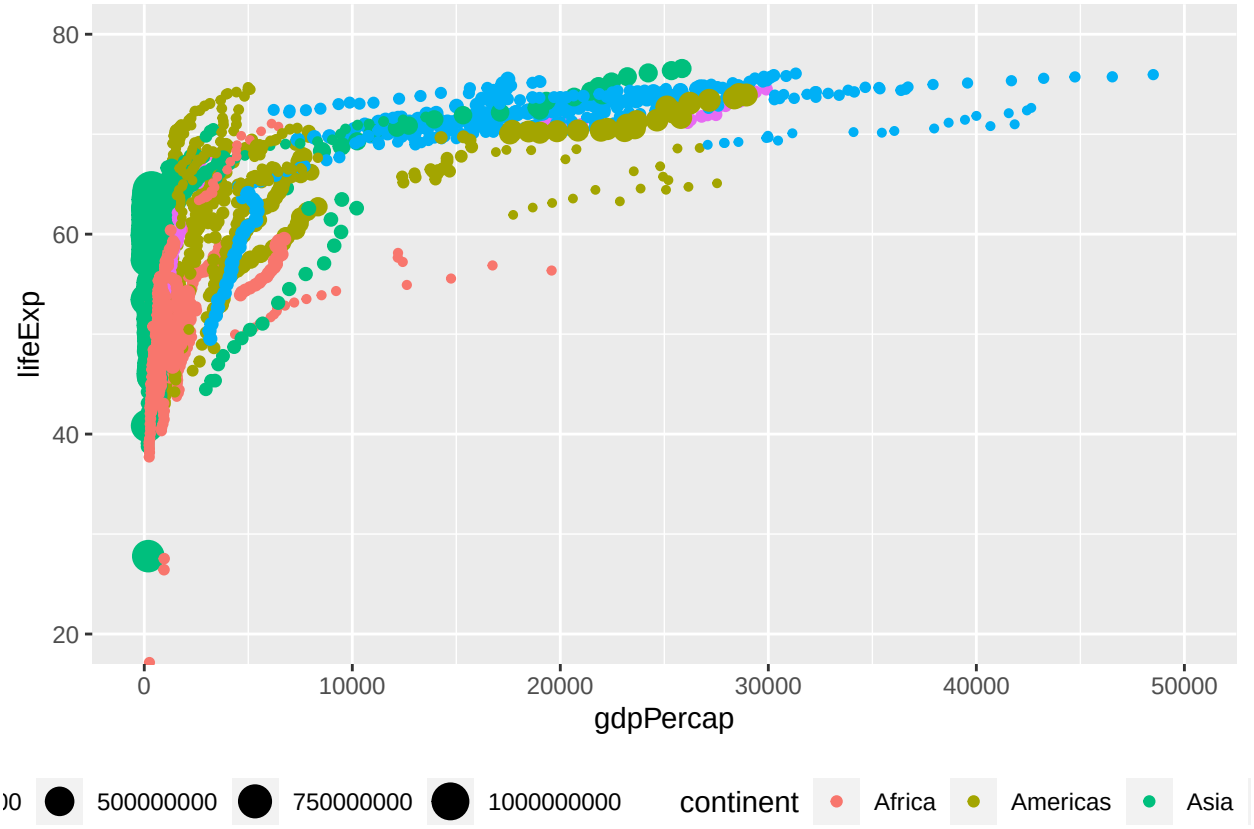
```
my_gapminder_1960 %>%
filter(year <= "1960-01-01") %>%
  ggplot(mapping = aes(x = gdpPercap, y = lifeExp, size = pop, colour = continent)) +
  geom_point() +
  coord_cartesian(ylim = c(20, 90), xlim = c(0,10000)) +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 13760 rows containing missing values (geom_point).
```



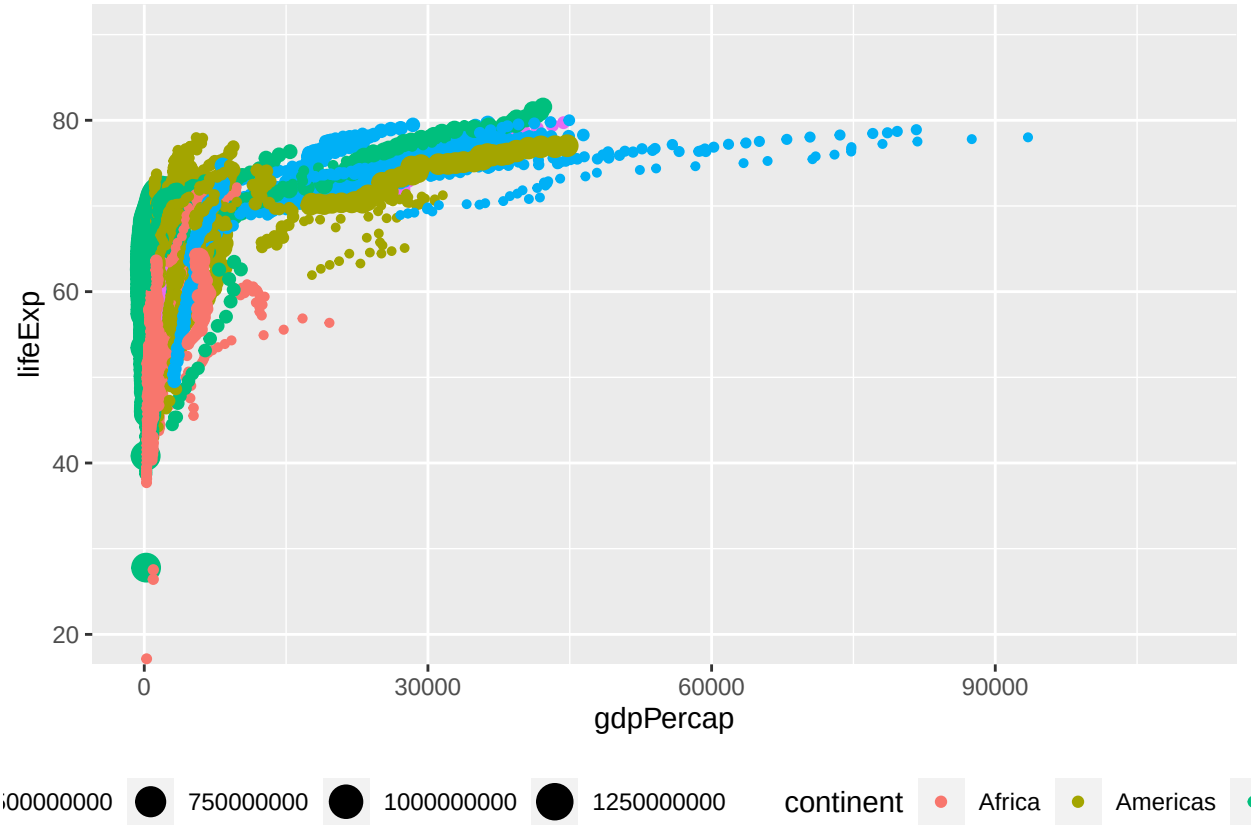

```
my_gapminder_1960 %>%
  filter(year <= "1980-01-01") %>%
  ggplot(mapping = aes(x = gdpPercap, y = lifeExp, size = pop, colour = continent)) +
  geom_point() +
  coord_cartesian(ylim = c(20, 80), xlim = c(0,50000)) +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 13760 rows containing missing values (geom_point).
```




```
my_gapminder_1960 %>%
  filter(year <= "2000-01-01") %>%
  ggplot(mapping = aes(x = gdpPercap, y = lifeExp, size = pop, colour = continent)) +
  geom_point() +
  coord_cartesian(ylim = c(20, 90), xlim = c(0,110000)) +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 13760 rows containing missing values (geom_point).
```

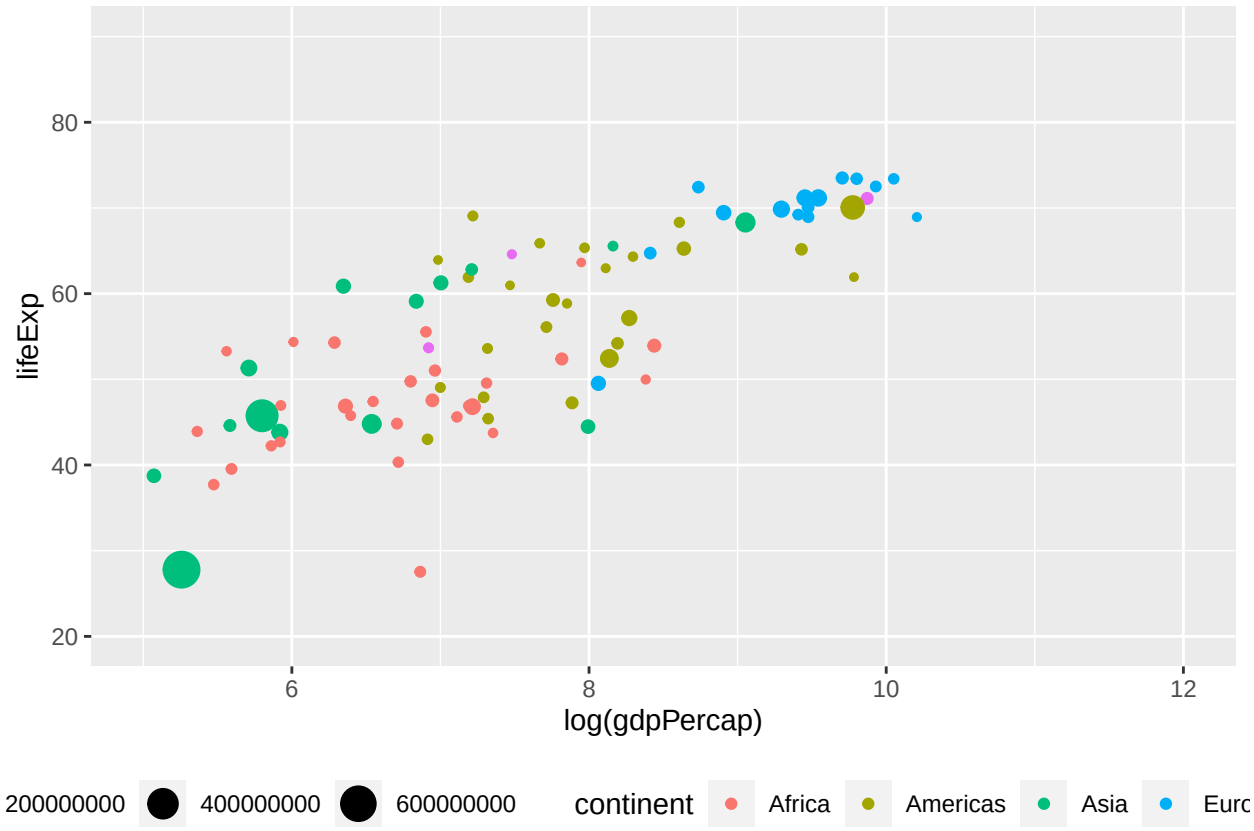


Spørsmål 18

Her bruker vi log i ggplottene.

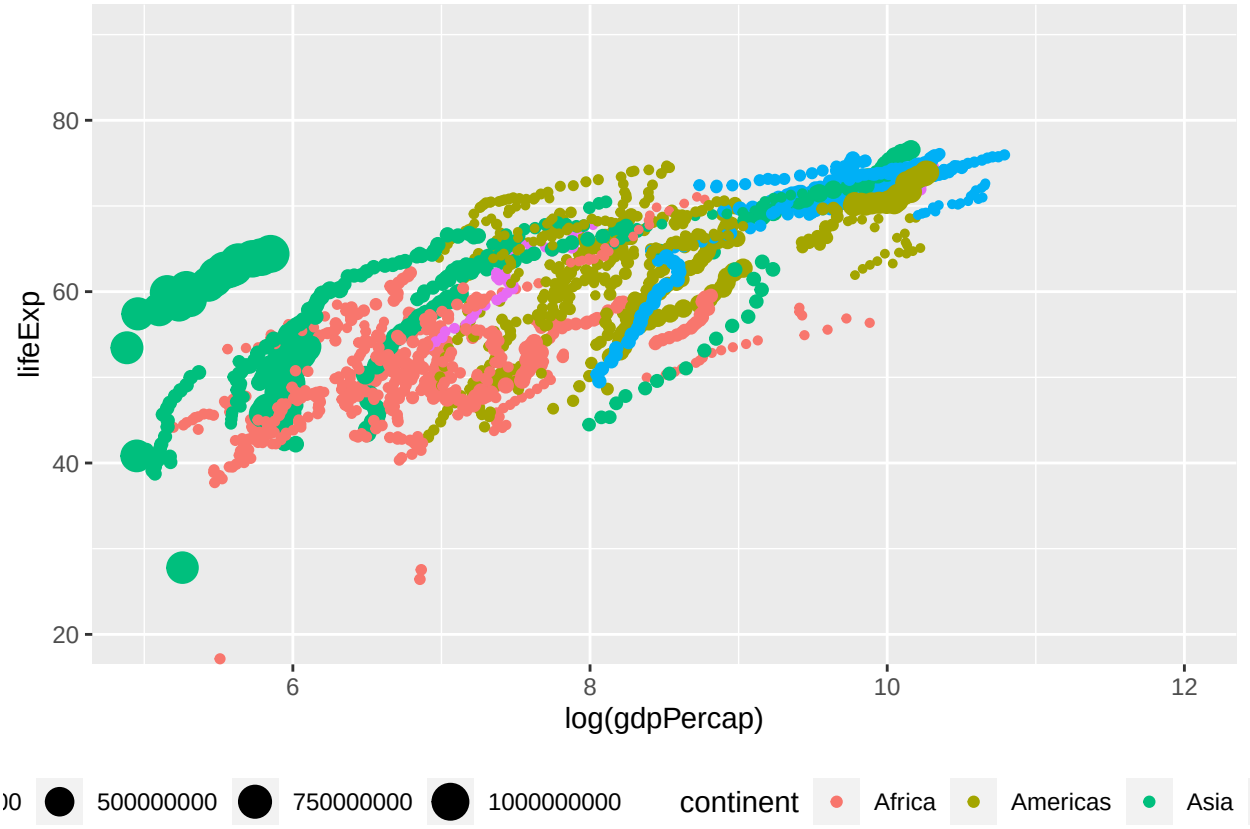
```
my_gapminder_1960 %>%
  filter(year <= "1960-01-01") %>%
  ggplot(mapping = aes(x = log(gdpPercap), y = lifeExp, size = pop, colour = continent)) +
  geom_point() +
  coord_cartesian(ylim = c(20, 90), xlim = c(5, 12)) +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 13760 rows containing missing values (geom_point).
```



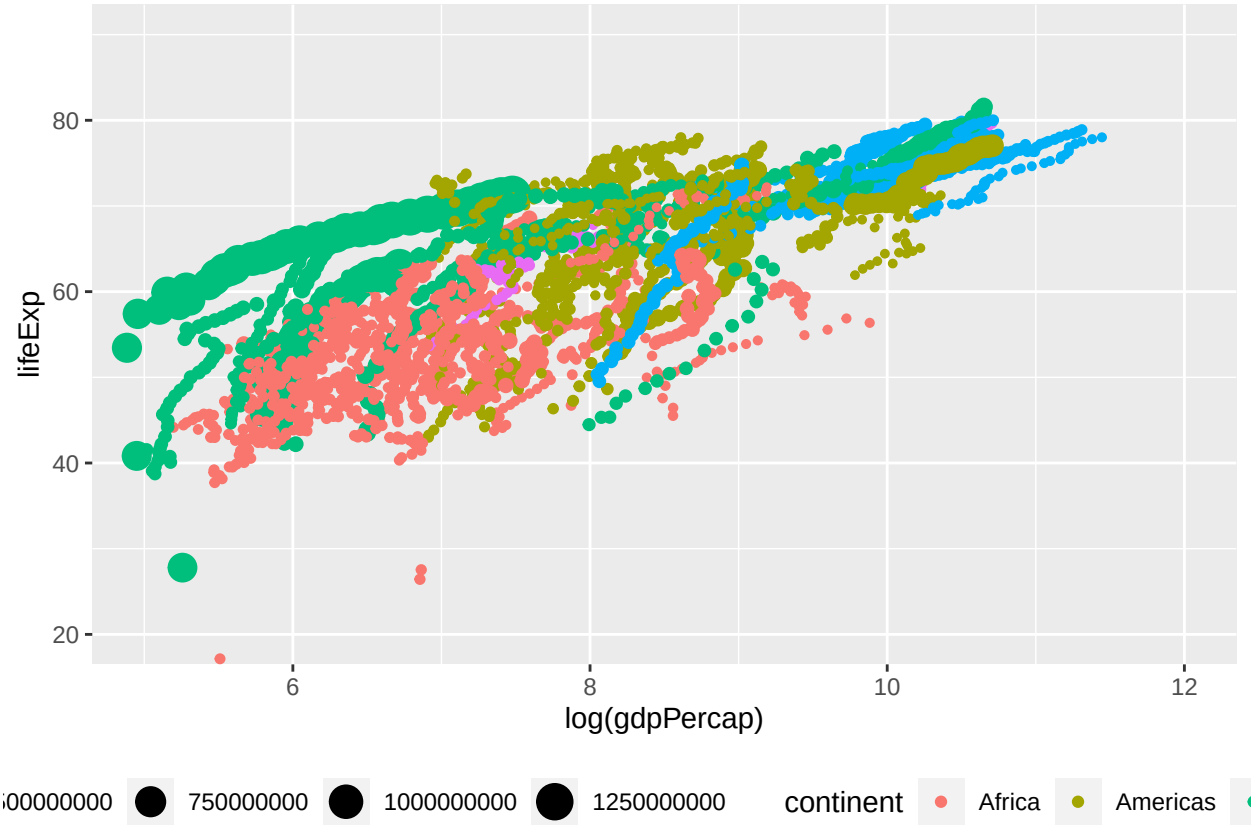

```
my_gapminder_1960 %>%
  filter(year <= "1980-01-01") %>%
  ggplot(mapping = aes(x = log(gdpPercap), y = lifeExp, size = pop, colour = continent)) +
  geom_point() +
  coord_cartesian(ylim = c(20, 90), xlim = c(5, 12)) +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 13760 rows containing missing values (geom_point).
```



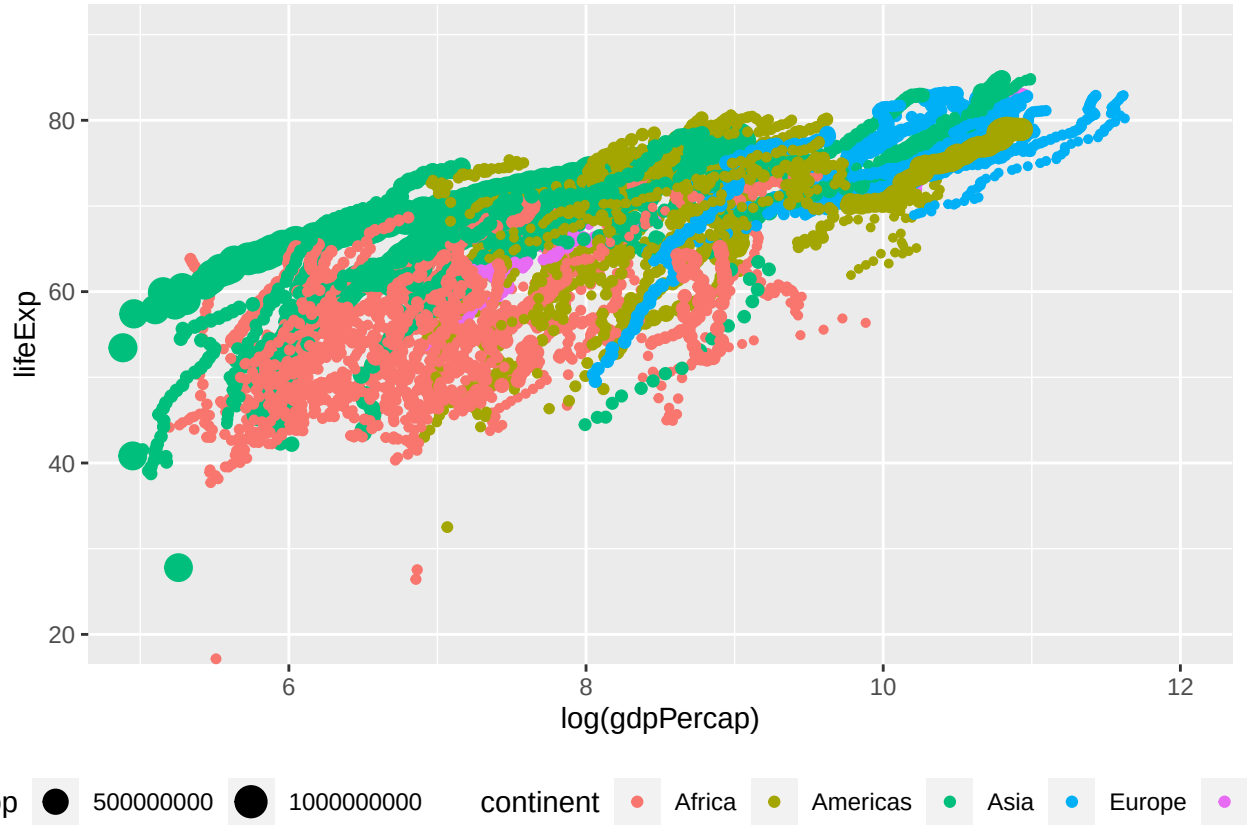

```
my_gapminder_1960 %>%
  filter(year <= "2000-01-01") %>%
  ggplot(mapping = aes(x = log(gdpPercap), y = lifeExp, size = pop, colour = continent)) +
  geom_point() +
  coord_cartesian(ylim = c(20, 90), xlim = c(5, 12)) +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 13760 rows containing missing values (geom_point).
```




```
my_gapminder_1960 %>%
  filter(year <= "2019-01-01") %>%
  ggplot(mapping = aes(x = log(gdpPercap), y = lifeExp, size = pop, colour = continent)) +
  geom_point() +
  coord_cartesian(ylim = c(20, 90), xlim = c(5, 12)) +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 13765 rows containing missing values (geom_point).
```



Spørsmål 19

I de siste 59 årene har det vært noen store observasjoner som man kan se. Man kan se at det er en økning i antall land som rapporterer BNP per innbygger. Man ser god ut en stor utvikling i landene og kontinentene som rapporterer BNP per innbygger.

Spørsmål 20

```
write.table(g_c, file = "my_gapminder.csv", sep = ",")
write.table(g_c_5year, file = "my_gapminder_red.csv", sep = ",")
```