

Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try <https://etherpad.wikimedia.org>).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

All content is publicly available under the Creative Commons Attribution License:
<https://creativecommons.org/licenses/by/4.0/>

```
-----

# creating this so my dir aligns with leaners

library(tidyverse)
library(ggplot2)

gapminder_data <- read.csv("data/gapminder_data.csv")

#gapminder_data <- read_csv("~/Desktop/un-report/gapminder_data.csv")

ggplot(data = gapminder_data)+
  aes(x = log(gdpPercap), y = lifeExp, size = pop/1000000, color = continent)+
  geom_point(alpha = 0.5) + # we made our points slightly transparent, because it makes it easier to see
overlapping points
  scale_color_brewer(palette = "Set1") +
  labs(x = "GDP Per Capita", y = "Life Expectancy", color= "Continent", size="Population (in millions)")
+
  theme_classic()
ggsave("gdp_percap.pdf")

#####
##day 2 work: data manipulation and cleaning
#####

gapminder_data <- read_csv("data/gapminder_data.csv")
glimpse(gapminder_data)
view(gapminder_data)

summarize(gapminder_data, avgLifeExp = mean(lifeExp))

# what is the maximum lifeExp

summarize(gapminder_data, maxLifeExp = max(lifeExp))
```

```

#pipes
# %>%
# means 'and then do this thing'

#option 1 without a pipe
summarize(gapminder_data, maxLifeExp = max(lifeExp))

#option 2 with a pipe
gapminder_data %>% summarize(aveLifeExp = mean(lifeExp))

# save the summary table
gapminder_data_summary <- gapminder_data %>% summarize(aveLifeExp = mean(lifeExp))

#filter command

gapminder_data %>% filter(year==2007) %>%
  • summarize(avgLifeExp07 = mean(lifeExp))

# find the earliest year in the data using summarize and min
gapminder_data %>% summarize(startpoint = min(year))

# filter the data to 1952 and then find the avg GDP per capita

gapminder_data %>% filter(year==1952) %>% summarize(avgGDPpercap==mean(gdpPercap))

# calculate life expectancy by year

gapminder_data %>% group_by(year) %>%
  • summarize(avg = mean(lifeExp))

# calculate the avg life expectancy by continent
gapminder_data %>% group_by(continent) %>%
  • summarize(avg = mean(lifeExp),
  • min=min(lifeExp),
  • max=ma(lifeExp))

#adding columns to dataset using mutate

gapminder_data %>% mutate(gdp = gdpPercap * pop)

# we have a column called pop
# use mutate to create a col for popInMillions

gapminder_data %>% mutate (popInMillions = pop/1000000)

# select to select cols

```

```
gapminder_data %>% select(pop, year)
```

```
gapminder_data %>% select(-continent)
```

```
#print a dataframe with the country, continent, year, and lifeExp
```

```
gapminder_data %>% select(country, continent, year, lifeExp)
```

```
#using helper functions
```

```
gapminder_data %>% select(year, starts_with("c"))
```

```
#print dataframe of all cols that end in the letter "p"
```

```
gapminder_data %>% select(ends_with('p'))
```

```
gapminder_data %>% select(country, continent, year, lifeExp) %>%
```

- pivot_wider(names_from = year, values_from = lifeExp)

```
#save a dataframe that contains only the americas in 2007
```

```
gapminder_data_2007 <- gapminder_data %>% filter(year==2007 & continent=='Americas') %>%
```

- select(-year, -continent)

```
# cleaning messy data
```

```
co2_emission_dirty <- read_csv('data/co2-un-data.csv', skip =2,  
col_names = c('region', 'country', 'year', 'series', 'value', 'footnotes', 'source'))
```

```
glimpse(co2_emissions_dirty)
```

```
co2_emissions_dirty %>% select(country, years, series, value) %>%
```

- print(n=50)

```
co2_emissions_dirty %>% select(country, year, series, value) %>%
```

- mutate(series = recode(series, 'Emissions (thousand metric tons of carbon dioxide)', =
'total_emissions',
 - 'Emissions per capita (metric tons of carbon dioxide)' = 'per_capita_emissions')) %>%
 - pivot_wider(names_from = series, values_from = value)

```
# filter down to 2005 and drop the year column
```

```
o2_emissions_dirty %>% select(country, year, series, value) %>%
```

- mutate(series = recode(series, 'Emissions (thousand metric tons of carbon dioxide)', =
'total_emissions',
 - 'Emissions per capita (metric tons of carbon dioxide)' = 'per_capita_emissions')) %>%
 - pivot_wider(names_from = series, values_from = value)
 - filter(year == 2005) %>%
 - select(-year)

```
#previously we created gapminder for 2007
```

```
# now we have co2 emissions for 2005
```

```
inner_join(gapminder_data, co2_emissions, by='country')
```

```
anti_join(gapminder_data, co2_emissions, by='country')
```

```
o2_emissions_dirty %>% select(country, year, series, value) %>%
```

- mutate(series = recode(series, 'Emissions (thousand metric tons of carbon dioxide)', =
'total_emissions',
 - 'Emissions per capita (metric tons of carbon dioxide)' = 'per_capita_emissions')) %>%
 - pivot_wider(names_from = series, values_from = value)
 - filter(year == 2005) %>%
 - select(-year)
 - mutate(country = recode(country,
 - " Bolivia (Plurin. State of)" = "Bolivia",
 - "United States of America" = "United States",
 - "Venezuela (Boliv. Rep. of)" = "Venezuela"))

```
anti_join(gapminder, co2_emissions, by='country')
```

```
#address Puerto Rico
```

```
gapminder_data <- gapminder_data %>% mutate(country = recode(country,
```

- "Puerto Rico" = "United States")

```
gapminder_co2 <- inner_join(gapminder_data, co2_emissions, by='country')
```

```
glimpse(gapminder_co2)
```

```
gapminder_co2 %>% mutate(region = if_else(country == "Canada" | country ==  
"United States" | country == "Mexico", "north", "south"))
```

```
write_csv(gapminder_co2, "data/gapminder_co2.csv")
```

```
#####
```

```
#### R Markdown ####
```

```
#####
```

```
#in unix shell
```

```
pwd
```

```
#go to un-report folder
```

```
cd Desktop/un-report
```

```
mkdir reports
```

```
#open R markdown file in R studio
```

```
#file -> new file -> R markdown, then rename and save in reports folder
```

```
```${r packages}  
library(tidyverse)
```
```

```
this is a report and it was prepared for this workshop I'm in
```

```
```${r}  
gapminder_1997 <- read_csv("data/gapminder_1997.csv")
```
```

```
```${r gdp_lifeexp_1997_plot}  
ggplot(data = gapminder_1997)+
 • aes(x=gdpPercap, y = lifeExp, color = continent, size = pop/1000000) +
 • geom_point()+
 • labs(x="GDP Per Capita", y='Life Expectancy', title = Do people in wealthy countires live
 longer?", size="Population (in millions)")
```
```

```
to remove library(tidyverse) from report:
```

```
```${r, include=False}  
```
```

```
to include ouput but not code
```

```
```${r, echo=False}  
```
```

```
```${r}  
nCountries <- gapminder_1997 %>%
 • select(country) %>%
 • n_distinct()
```

```
minGDP <- gapminder_1997 %>%
 • summarise(round(min(gdpPercap))) %>%
 • pull()
```

```
maxGDP <- gapminder_1997 %>%
 • summarise(round(max(gdpPercap))) %>%
 • pull()
```
```

the above plot shows the relationship between GDP per capita and life expectancy for the total of `r nCountries` countries. Also min GDP `r minGDP` and maxGDP `maxGDP`

|Summary of Data|

|-----|-----|

|Number of Countries| `r nCountries`|

|Min of GDP per capita| `r minGDP`|

|Max of GDP per capita| `r maxGDP`|

```
``{r}
```

```
#load library
```

```
library(knitr)
```

```
#print kable
```

```
gapminder %>%
```

- filter(continent == "Oceania") %>%
- kable()

```
``
```

<https://umcarpentries.org/intro-curriculum-r/05-r-markdown/index.html#why-use-r-markdown>