hw02_dplyr

Exercise 1: Basic dplyr

Loading packages

```r
library(dplyr)
```

```
##
## 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
```

```r
library(gapminder)
```

1.1

```r
gapminder %>%
  filter(year < 1980 & year > 1969,
         country %in% c("Canada", "Spain", "Portugal")) %>%
  knitr::kable()
```

country	continent	year	lifeExp	pop	gdpPercap
Canada	Americas	1972	72.88	22284500	18970.571
Canada	Americas	1977	74.21	23796400	22090.883
Portugal	Europe	1972	69.26	8970450	9022.247
Portugal	Europe	1977	70.41	9662600	10172.486
Spain	Europe	1972	73.06	34513161	10638.751
Spain	Europe	1977	74.39	36439000	13236.921

1.2

```r
gapminder %>%
  filter(year < 1980 & year > 1969,
         country %in% c("Canada", "Spain", "Portugal")) %>%
  select(country, gdpPercap) %>%
  knitr::kable()
```

country	gdpPercap
Canada	18970.571
Canada	22090.883
Portugal	9022.247
Portugal	10172.486
Spain	10638.751
Spain	13236.921

1.3

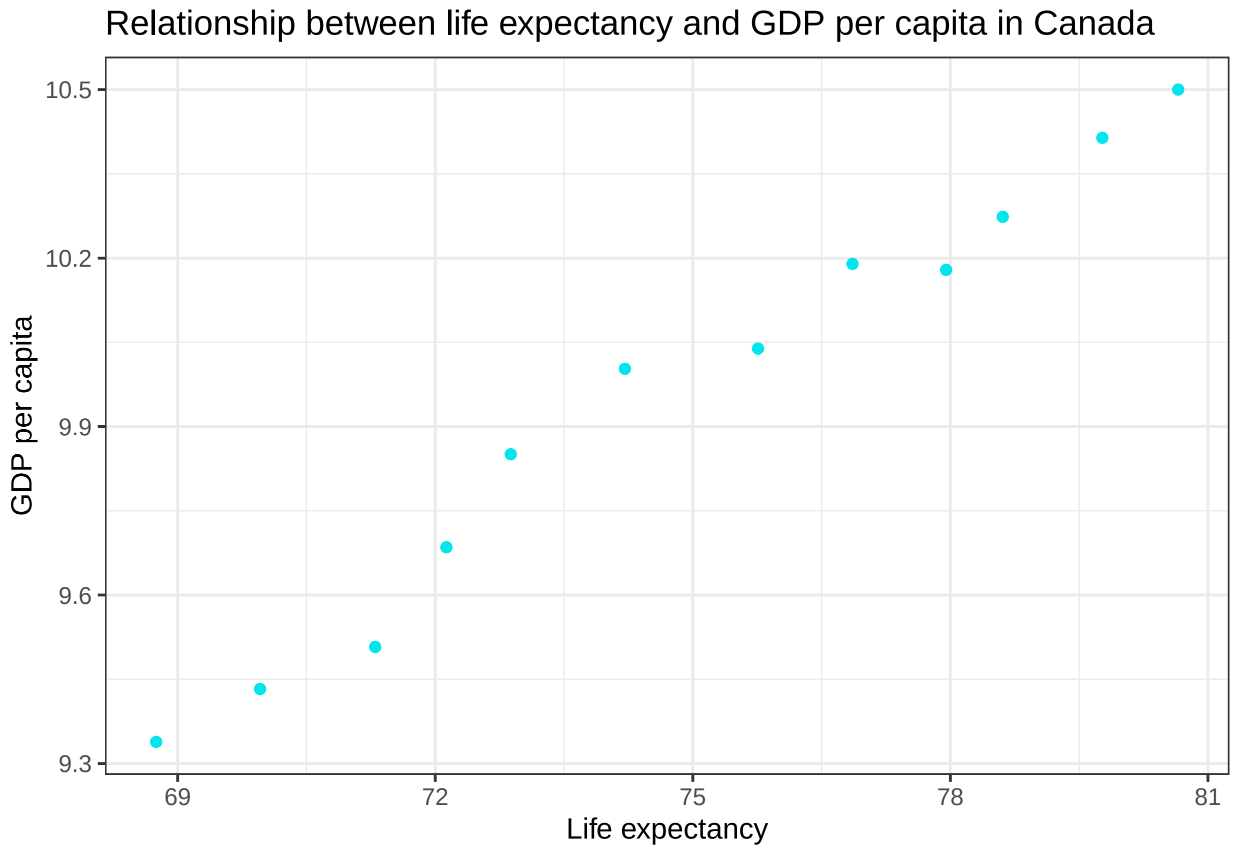
```
gapminder %>%
  group_by(country) %>%
  select(country, year, lifeExp) %>%
  mutate(lifeExp_increase = lifeExp - lag(lifeExp)) %>%
  arrange(lifeExp_increase) %>%
  filter(lifeExp_increase < 0) %>%
  DT::datatable()
```

1.4

```
gapminder %>%
  group_by(country) %>%
  summarize(maxGdpPercap = max(gdpPercap)) %>%
  DT::datatable()
```

1.5

```
library(ggplot2)

gapminder %>%
  filter(country == "Canada",
         lifeExp, gdpPercap) %>%
  mutate(gdpPercap_log = log(gdpPercap)) %>%
  ggplot(aes(lifeExp, gdpPercap_log)) +
  geom_point(colour="turquoise2") +
  xlab("Life expectancy") +
  ylab("GDP per capita") +
  ggtitle("Relationship between life expectancy and GDP per capita in Canada") +
  theme_bw()
```



Exercise 2: Explore individual variables with dplyr

Pick one categorical variable and one quantitative variable to explore. Answer the following questions in whichever way you think is appropriate, using dplyr:

What are possible values (or range, whichever is appropriate) of each variable? What values are typical? What's the spread? What's the distribution? Etc., tailored to the variable at hand.

Feel free to use summary stats, tables, figures.

```
library(tidyverse)
```

```
## -- Attaching packages -------------------------------------------------------------------------

## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#exploring categorical variable: type of cars
mtcars %>%
  rownames_to_column("type") %>%
  distinct(type) %>%
  filter(stringr::str_detect(type, "Toyota")) %>% #looking for Toyota models in the dataset
  knitr::kable()
```

type
Toyota Corolla
Toyota Corona

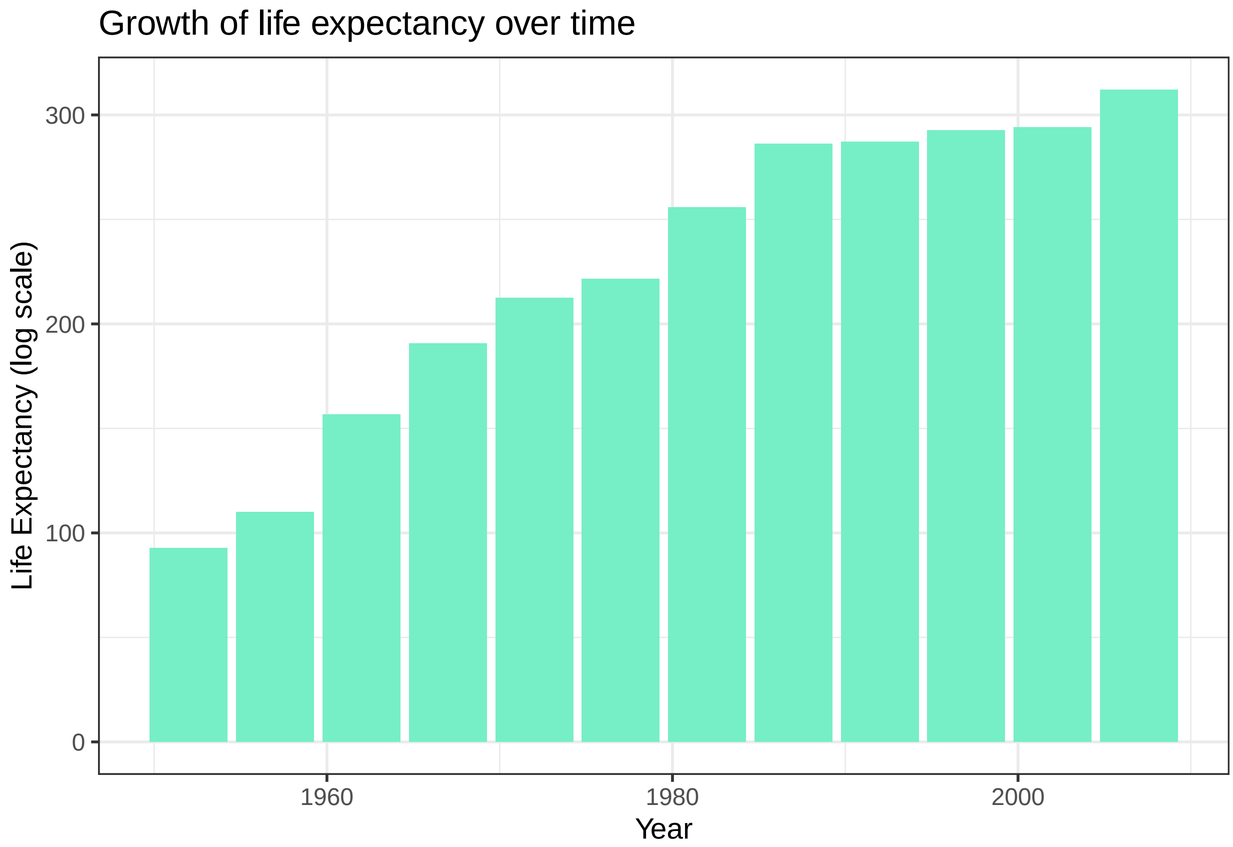
```
mtcars %>%
  rownames_to_column("type") %>%
  group_by("type") %>%
  filter(mpg > 30 & hp > 50) %>% #finding all cars with mpg above 30 and hp above 50
  arrange(desc(mpg, hp)) %>%
  mutate(relationship_mpg_hp = mpg * hp) %>%
  knitr::kable()
```

type	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	“type”	relationship_mpg_hp
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1	type	2203.5
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1	type	2138.4
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2	type	1580.8
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2	type	3435.2

```
#exploring numerical variable: year
gapminder %>%
  distinct(year) %>% #all the values for this variable
  knitr::kable()
```

year
1952
1957
1962
1967
1972
1977
1982
1987
1992
1997
2002
2007

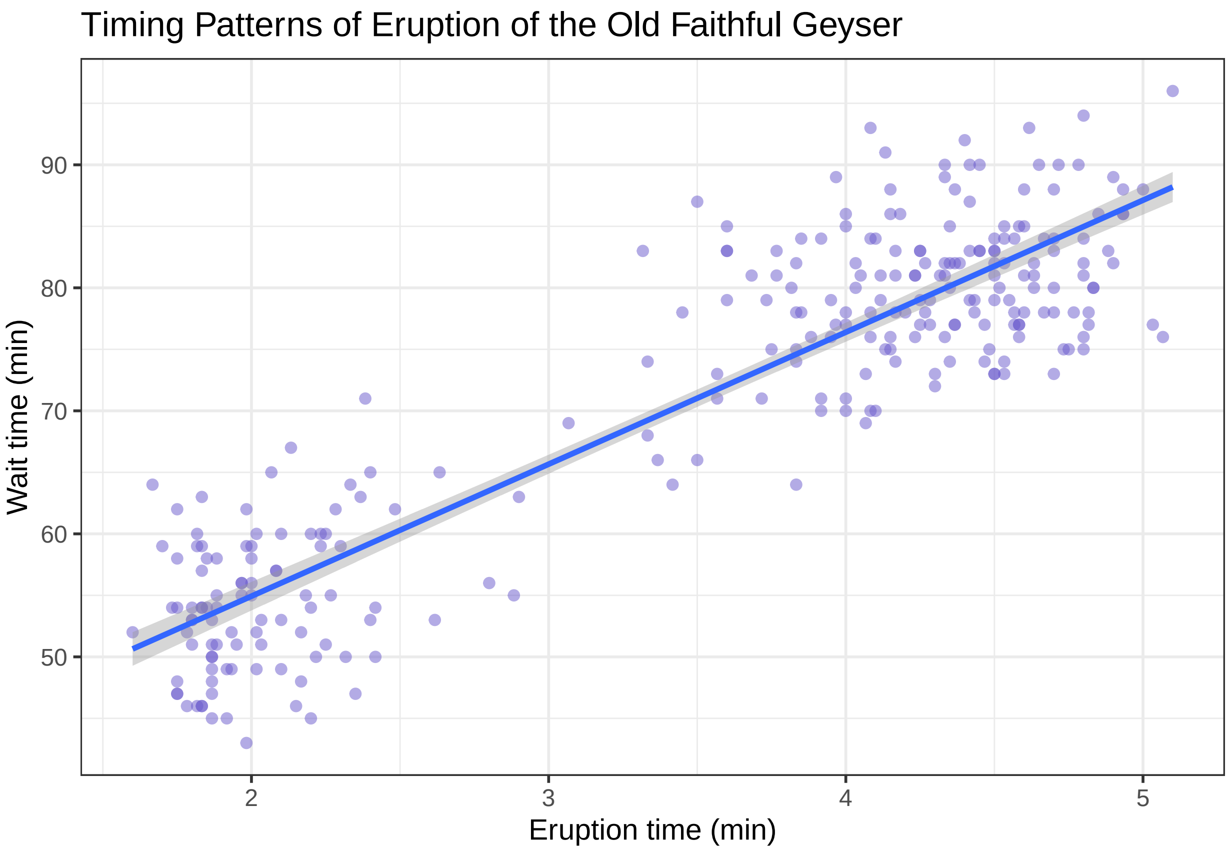
```
gapminder %>%
  filter(lifeExp > 60, gdpPercap > 5000) %>% #interested in countries with lifeExp > 60 and gdpPercap >
    group_by(year) %>%
  arrange(desc(year)) %>%
    select(-c(country, pop, gdpPercap)) %>% #getting rid of columns that I am not interested in
  ggplot(aes(year, log(lifeExp))) +
  geom_bar(stat="identity", fill = "aquamarine2") +
  xlab("Year") +
  ylab("Life Expectancy (log scale)") +
  ggtitle("Growth of life expectancy over time") +
  theme_bw()
```



Exercise 3: Explore various plot types

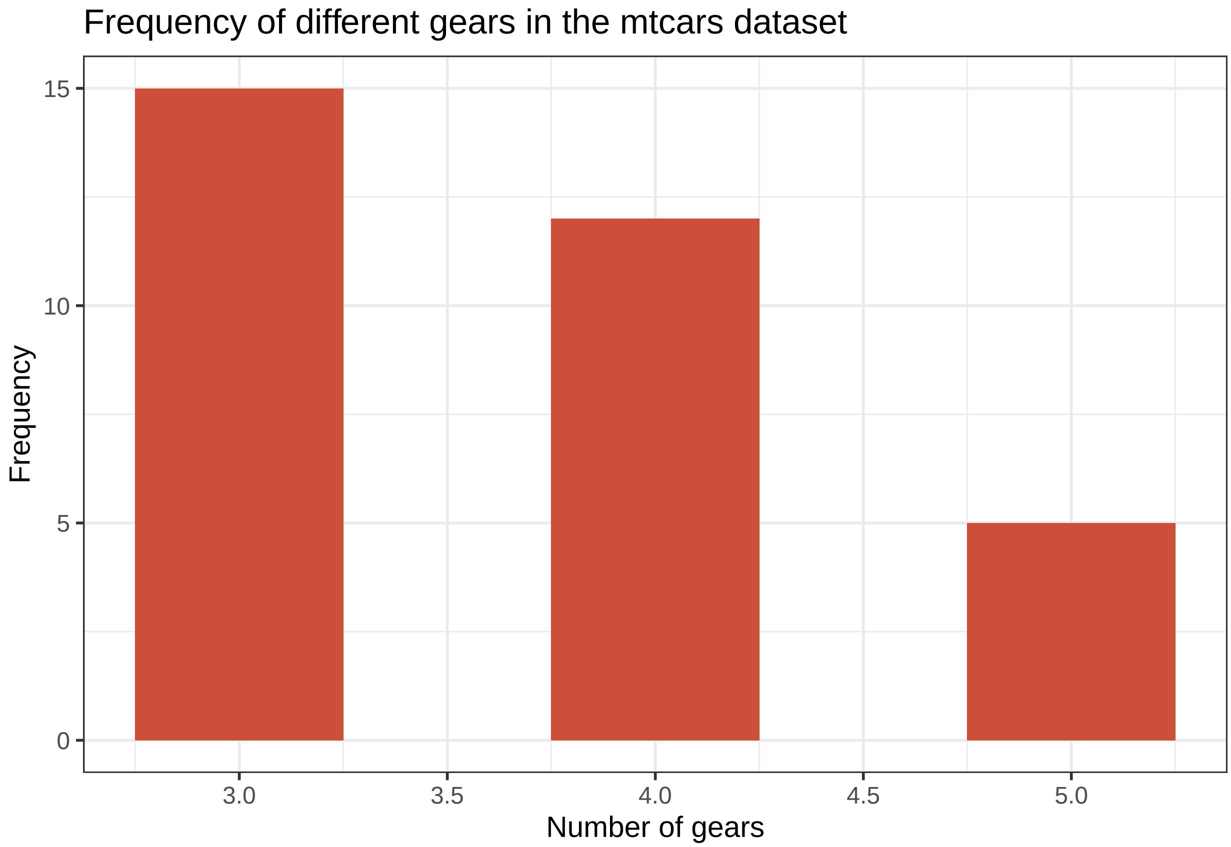
```r
library(ggplot2)

DT::datatable(datasets::faithful)
```

```r
faithful %>%
  ggplot(aes(eruptions, waiting)) +
  xlab("Eruption time (min)") +
  ylab("Wait time (min)") +
  ggtitle("Timing Patterns of Eruption of the Old Faithful Geyser") +
  geom_point(alpha = 0.5, colour = "slateblue3") +
  geom_smooth(method = "lm") +
  theme_bw()
```



```
DT::datatable(datasets::mtcars)
```

```
mtcars %>%
  ggplot(aes(gear,)) +
  geom_bar(width = 0.5, fill = "tomato3") +
  xlab("Number of gears") +
  ylab("Frequency") +
  ggtitle("Frequency of different gears in the mtcars dataset") +
  theme_bw()
```



Recycling

For people who want to take things further.

Evaluate this code and describe the result. Presumably the analyst's intent was to get the data for Rwanda and Afghanistan. Did they succeed? Why or why not? If not, what is the correct way to do this?

filter(gapminder, country == c("Rwanda", "Afghanistan")) Read What I do when I get a new data set as told through tweets from SimplyStatistics to get some ideas!

```
DT::datatable(filter(gapminder, country == c("Rwanda", "Afghanistan"))) #This method is successful in r
```

```
#Alternatives
##Way 1
gapminder %>%
  filter(country == "Rwanda" | country == "Afghanistan") %>%
  DT::datatable()
```

```
##Way 2
gapminder %>%
  filter(country %in% c("Rwanda", "Afghanistan")) %>%
  DT::datatable()
```

Tibble display

Present numerical tables in a more attractive form using knitr::kable() for small tibbles (say, up to 10 rows), and DT::datatable() for larger tibbles.