FallFinalProject

December 12, 2022

#

Fall Semester Final Project

[1]:
```
# Run this code to load the required packages
suppressMessages(suppressWarnings(suppressPackageStartupMessages({
  library(coursekata)
})))

wcmatches <- read.csv('https://raw.githubusercontent.com/rfordatascience/
 ↪tidytuesday/master/data/2022/2022-11-29/wcmatches.csv')
worldcups <- read.csv('https://raw.githubusercontent.com/rfordatascience/
 ↪tidytuesday/master/data/2022/2022-11-29/worldcups.csv')
```

Here are six different data sets to get you started (they are all preloaded with coursekata library). Take a look at them, choose one that interests you for your report.

Option 1: `gapminder`

- This data frame contains health and income outcomes for 184 countries from 1960 to 2016.
- Click here for a description of all the variables.

Option 2: `temp_carbon`

- This data frame has data on the annual mean global temperature anomaly on land, sea and combined, 1880-2018. Annual global carbon emissions, 1751-2014.
- Click here for a description of all the variables.

Option 3: `admissions`

- The data set contains the admission data for six majors for the fall of 1973 for UC Berkeley.
- Click here for a description of all the variables.

Option 4: `murders`

- The data set contains gun murder data from FBI reports. Also contains the population of each state.
- Click here for a description of all the variables.

Option 5: `NutritionStudy`

- The data set contains data related to nutrition and health for 315 individuals.
- Click here for a description of all the variables.

Option 6: `wcmatches` or `worldcups` - The `worldcups` dataset contains every single World Cup match played in its history from Uruguay in 1930 to Russia in 2018.
- The `wcmatches` dataset contains every football match played in the World Cup. - Click here for a description of all the variables.

If there's a data frame you're interested in using that's not listed here, then discuss with your teacher about using it for this report.

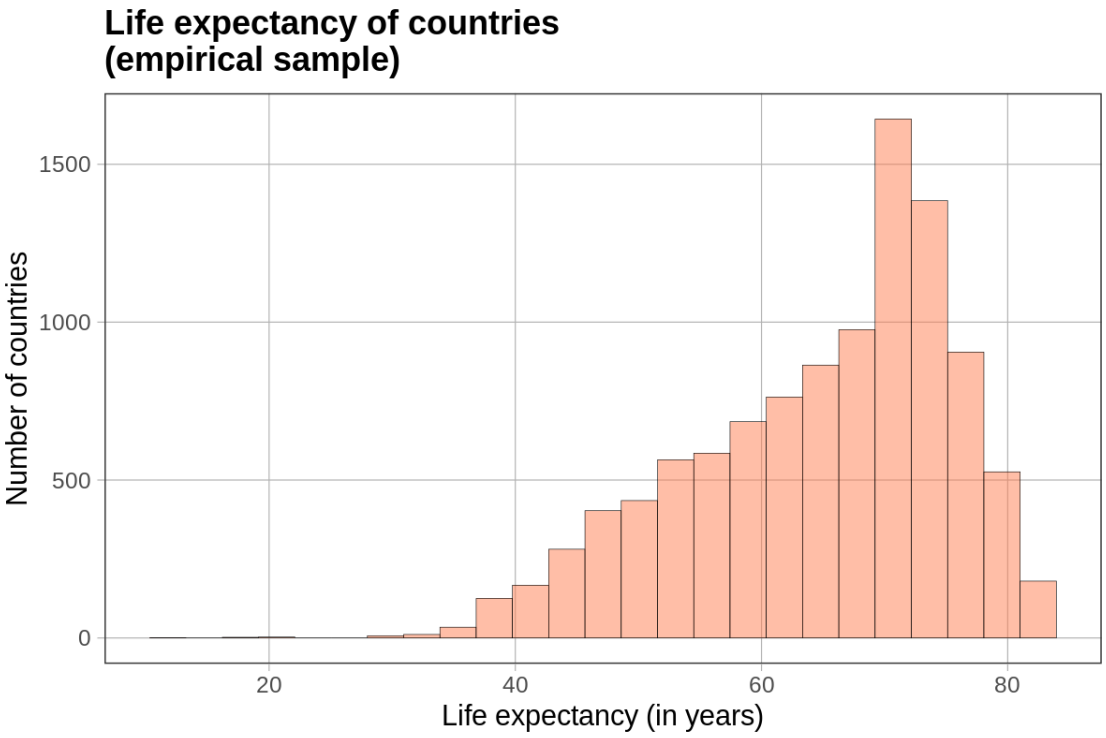
> For my data analysis report, I'm going to be using the gapminder data frame, which contains health and income outcomes for 184 countries from 1960 to 2016. The information in this data frame was sourced from the Gapminder Foundation, a non-profit organization dedicated to educating people about the world and correcting their worldview misconceptions. The variables in the data frame are the country, year, infant mortality rate (deaths per 1000 infants), life expectancy (in years), fertility (average children per woman), population, gross domestic product, continent, and geographical region.

[2]:
```
glimpse(gapminder)
```

```
Rows: 10,545
Columns: 9
$ country          <fct> "Albania", "Algeria", "Angola",
"Antigua and Barbuda"…
$ year             <int> 1960, 1960, 1960, 1960, 1960, 1960,
1960, 1960, 1960,…
$ infant_mortality <dbl> 115.40, 148.20, 208.00, NA, 59.87,
NA, NA, 20.30, 37.…
$ life_expectancy  <dbl> 62.87, 47.50, 35.98, 62.97, 65.39,
66.86, 65.66, 70.8…
$ fertility        <dbl> 6.19, 7.65, 7.32, 4.43, 3.11, 4.55,
4.82, 3.45, 2.70,…
$ population       <dbl> 1636054, 11124892, 5270844, 54681,
20619075, 1867396,…
$ gdp              <dbl> NA, 13828152297, NA, NA,
108322326649, NA, NA, 966778…
$ continent        <fct> Europe, Africa, Africa, Americas,
Americas, Asia, Ame…
$ region           <fct> Southern Europe, Northern Africa,
Middle Africa, Cari…
```

> In my report, I will be exploring the variable `life_expectancy`, which contains the annual life expectancies of each country in the data frame. The distribution of `life_expectancy` has been visualized in the frequency histogram below:

[3]:
```
gf_histogram(~life_expectancy, data = gapminder, fill = "coral")%>%
gf_labs(title = "Life expectancy of countries\n(empirical sample)", x = "Life␣
  ↪expectancy (in years)", y = "Number of countries")
```



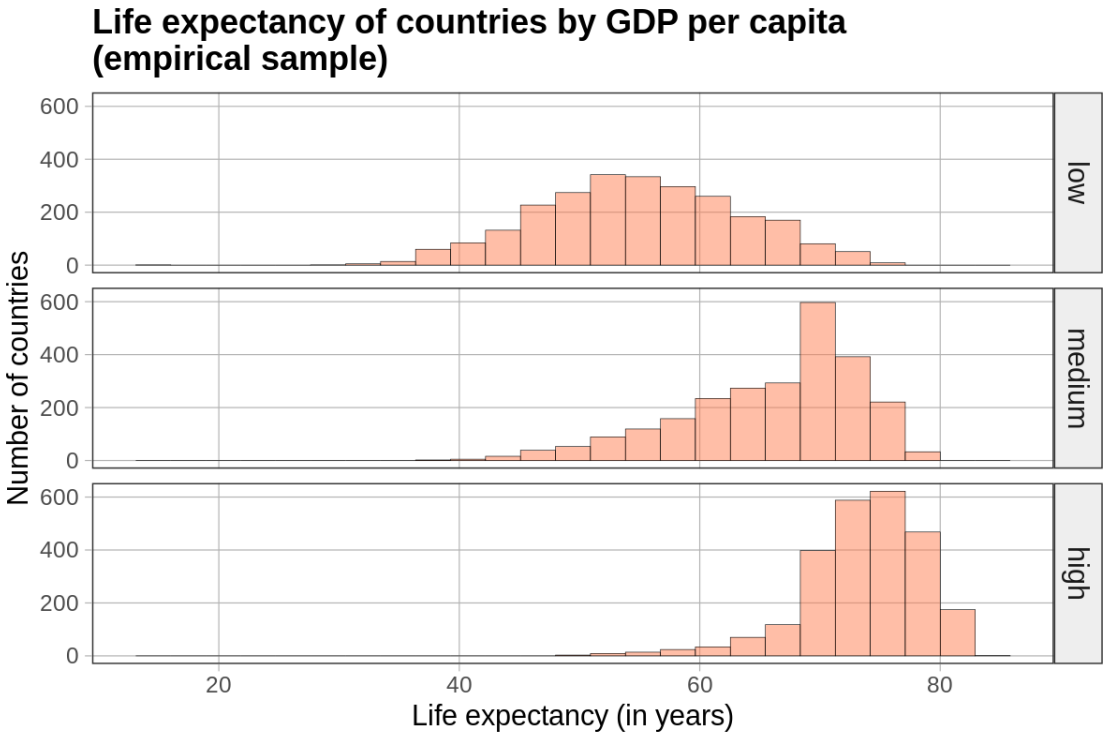
The distribution of `life_expectancy` is skewed to the left and unimodal with a peak at around 70 years. There are several outlier countiers at the lower end of the distribution, where life expectancy is around 15 to 20 years. The distribution has a relatively wide spread, with a range of around 70 years.

Specifically, I will be exploring the relationship between the gross domestic product (GDP) per capita and life expectancy of a country. GDP is defined as the market value of the final goods and services produced in a country, and is used as a measure of a country's economic activity and wealth. GDP per capita is a country's GDP divided by its population, which allows countries with differing populations to be compared fairly. GDP per capita can be used as a measure of wealth per person. I predict that GDP per capita and life expectancy have a direct relationship, because GDP per capita can be interpreted as a country's capability to support its public health.

My hypothesis can be represented by the word equation: life expectancy = GDP per capita + other stuff

Many of the countries in the data frame didn't have a recorded GDP for certain years, so I excluded those data points from the visualization before doing anything else. I created the quantitative variable `gdpPerCapita` (GDP per capita) by dividing `gdp` by `population`. I then turned the variable `gdpPerCapita` into a categorical variable so I could use it to create a faceted histogram. I did that by ordering the values of `gdpPerCapita` from lowest to highest and then dividing the values into three equal-sized groups: low, medium, and high. I saved what group each observation belonged to under the variable `gdpPerCapita.group`.

[4]:
```
gapminder <- gapminder[!(is.na(gapminder$gdp)), ]
gapminder$gdpPerCapita <- gapminder$gdp / gapminder$population
gapminder$gdpPerCapita.group <- ntile(gapminder$gdpPerCapita, 3)
gapminder$gdpPerCapita.group <- factor(gapminder$gdpPerCapita.group, levels =␣
↪c(1,2,3), labels = c("low", "medium", "high"))

gf_histogram(~life_expectancy, data = gapminder, fill = "coral")%>%
gf_facet_grid(gdpPerCapita.group ~ .)%>%
gf_labs(title = "Life expectancy of countries by GDP per capita\n(empirical␣
↪sample)", x = "Life expectancy (in years)", y = "Number of countries")
```



The distribution for the low GDP per capita group is rougly symmetrical and centered around 55 years with a range of about 50 years. The medium and high GDP per capita groups are both skewed to the left with a range of about 40 years. The medium GDP per capita group is centered around 65 years and the high GDP per capita group is centered around 75 years. The overall distribution of life expectancy appears greatest for the high GDP per capita group and lowest for the low GDP per capita group. The distribution for the high GDP per capita group is more strongly skewed than the distributions for the medium and low GDP per capita groups.

The differences between the distributions of life expectancy suggest that countries with higher GDP per capita have greater life expectancies. However, the differences might not be statistically significant and may be just the result of randomness in the data. Overall, I believe that the visualizations support the plausibility of my hypothesis that a

country's GDP per capita has a direct relationship with life expectancy in that country.

[5]:
```
mean(~life_expectancy, data = subset(gapminder, gdpPerCapita.group == "low"))
mean(~life_expectancy, data = subset(gapminder, gdpPerCapita.group == "medium"))
mean(~life_expectancy, data = subset(gapminder, gdpPerCapita.group == "high"))
```

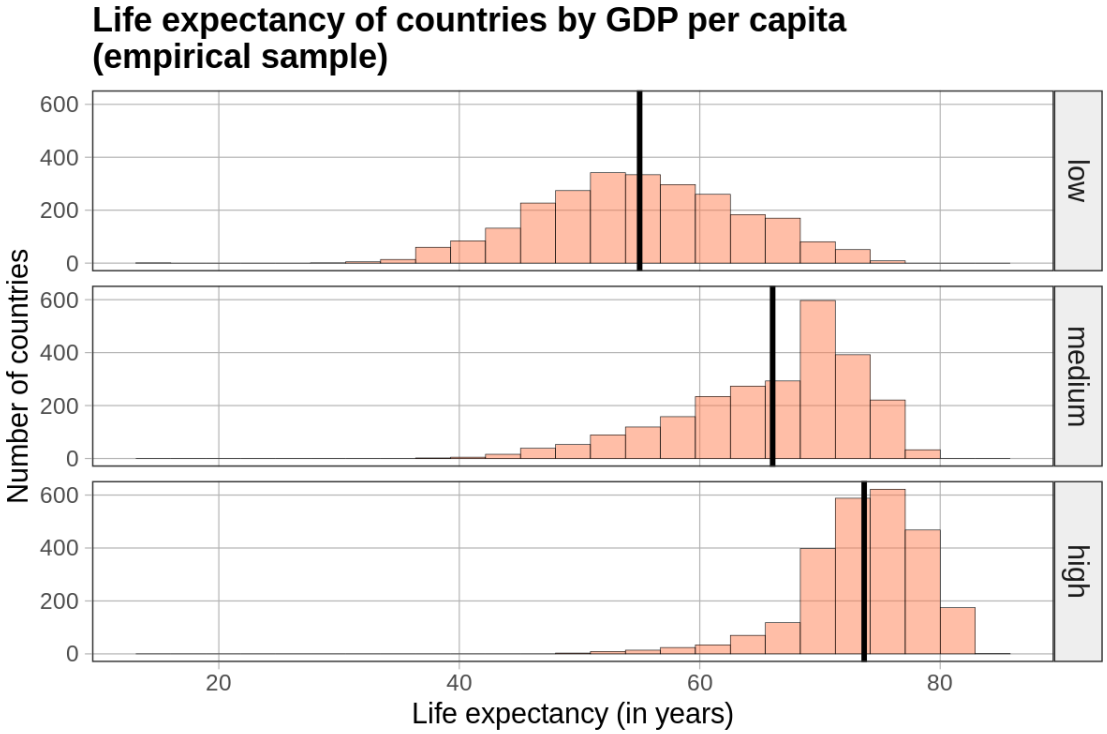
54.9996831683168

66.0623890649762

73.6749366085578

The empty model for a distribution is the mean of the distribution. For the low GDP per capita group, the mean life expectancy is about 55.000 years. For the medium GDP per capita group, the mean life expectancy is about 66.062. For the high per capita GDP group, the mean life expectancy is about 73.675 years. The empty models have been added to the faceted histogram below:

[6]:
```
data.frame_1 <- gapminder%>%
group_by(gdpPerCapita.group)%>%
summarize(mean = mean(life_expectancy))

gf_histogram(~life_expectancy, data = gapminder, fill = "coral")%>%
gf_facet_grid(gdpPerCapita.group ~ .)%>%
gf_vline(xintercept = ~mean, data = data.frame_1, color = "black")%>%
gf_labs(title = "Life expectancy of countries by GDP per capita\n(empirical␣
 ↪sample)", x = "Life expectancy (in years)", y = "Number of countries")
```



The GLM notation for the model is as follows: $Y_i = b_0 + e_i$

In the distributions of `life_expectancy`, Y_i represents a country's life expectancy for a certain year, b_0 represents the mean life expectancy of that country's GDP per capita group, and e_i represents the deviation of Y_i from b_0.

In order to determine whether or not GDP per capita is related to a country's life expectancy, I simulated a random data generating process for `life_expectancy`. I used the `shuffle` function to shuffle the values of `gdpPerCapita.group` and saved them under the new variable `gdpPerCapita.group_shuffled.1`. I then created a faceted histogram, faceting by `gdpPerCapita.group_shuffled.1` instead of `gdpPerCapita.group` to compare distributions of `life_expectancy` between the simulated sample to the emprical sample. I created two more simulated samples (`gdpPerCapita.group_shuffled.2`, `gdpPerCapita.group_shuffled.3`) to confirm that the differences between the simulated sample and the empirical sample were a result of the random DGP.

[7]:
```
gapminder$gdpPerCapita.group_shuffled.1 <- shuffle(gapminder$gdpPerCapita.group)
gapminder$gdpPerCapita.group_shuffled.2 <- shuffle(gapminder$gdpPerCapita.group)
gapminder$gdpPerCapita.group_shuffled.3 <- shuffle(gapminder$gdpPerCapita.group)
# separated otherwise right half of data frame is cut off in PDF
head(select(gapminder, country, life_expectancy, gdpPerCapita, gdpPerCapita.
 ↪group), 5)
head(select(gapminder, gdpPerCapita.group_shuffled.1, gdpPerCapita.
 ↪group_shuffled.2, gdpPerCapita.group_shuffled.3), 5)
```

A data.frame: 5 × 4

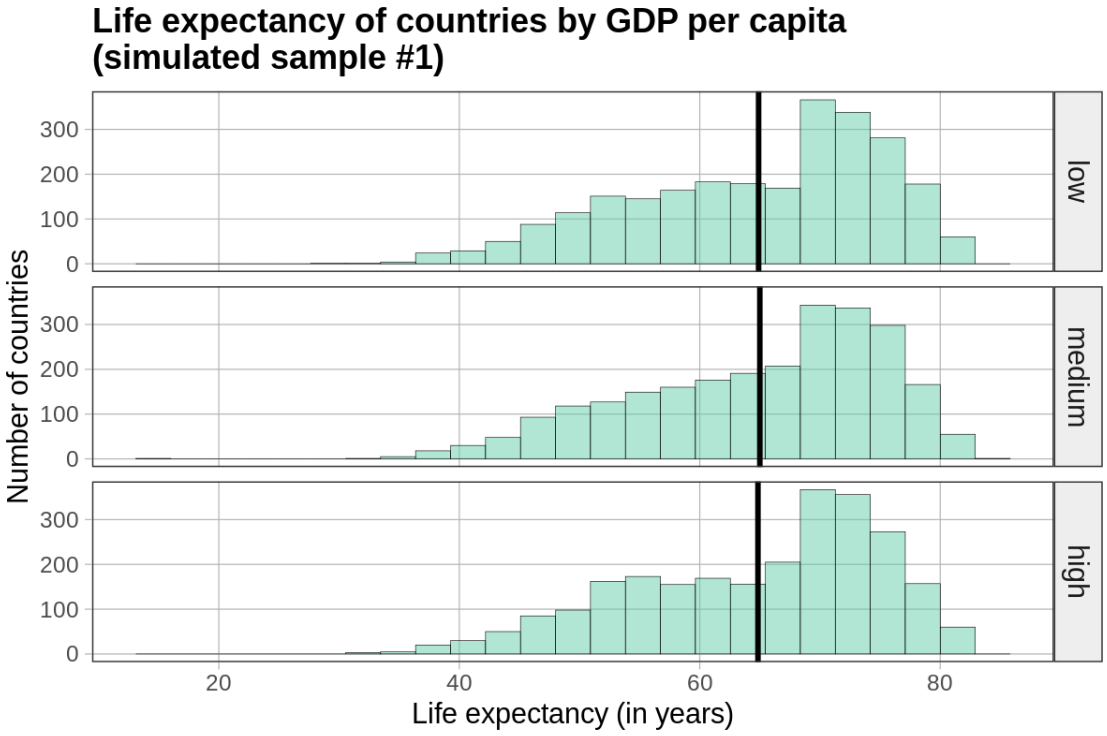
	country <fct>	life_expectancy <dbl>	gdpPerCapita <dbl>	gdpPerCapita.group <fct>
2	Algeria	47.50	1242.992	medium
5	Argentina	65.39	5253.501	high
8	Australia	70.87	9393.197	high
9	Austria	68.75	7415.259	high
11	Bahamas	62.00	11926.570	high

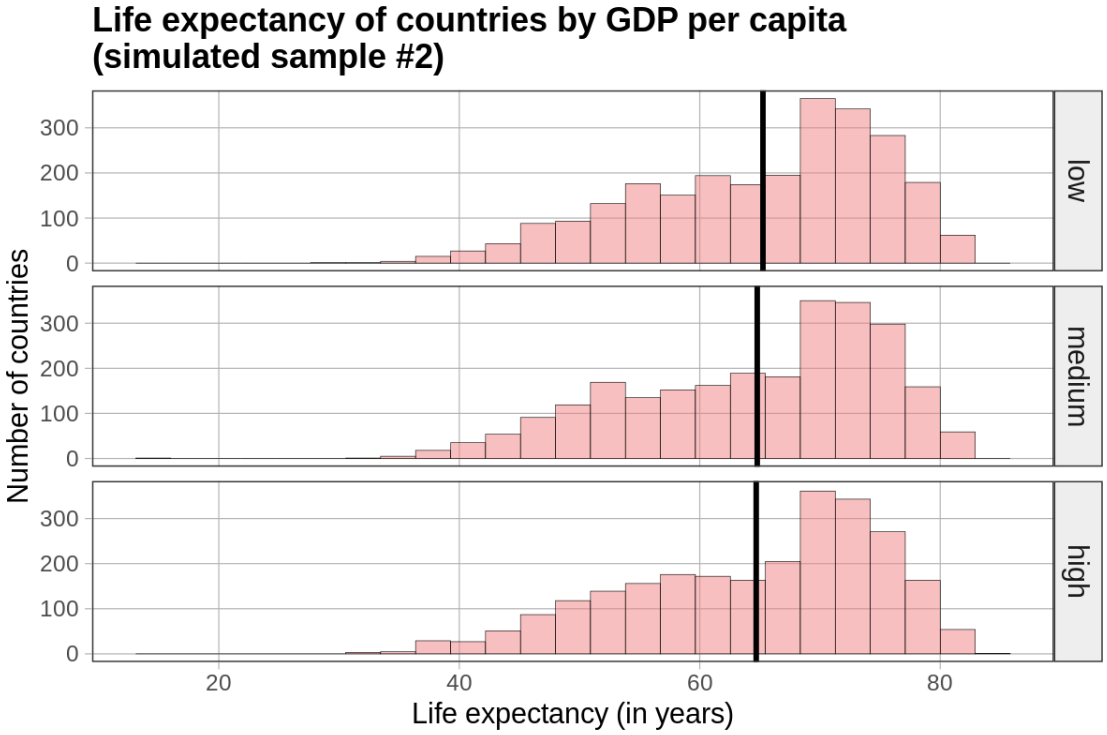
A data.frame: 5 × 3

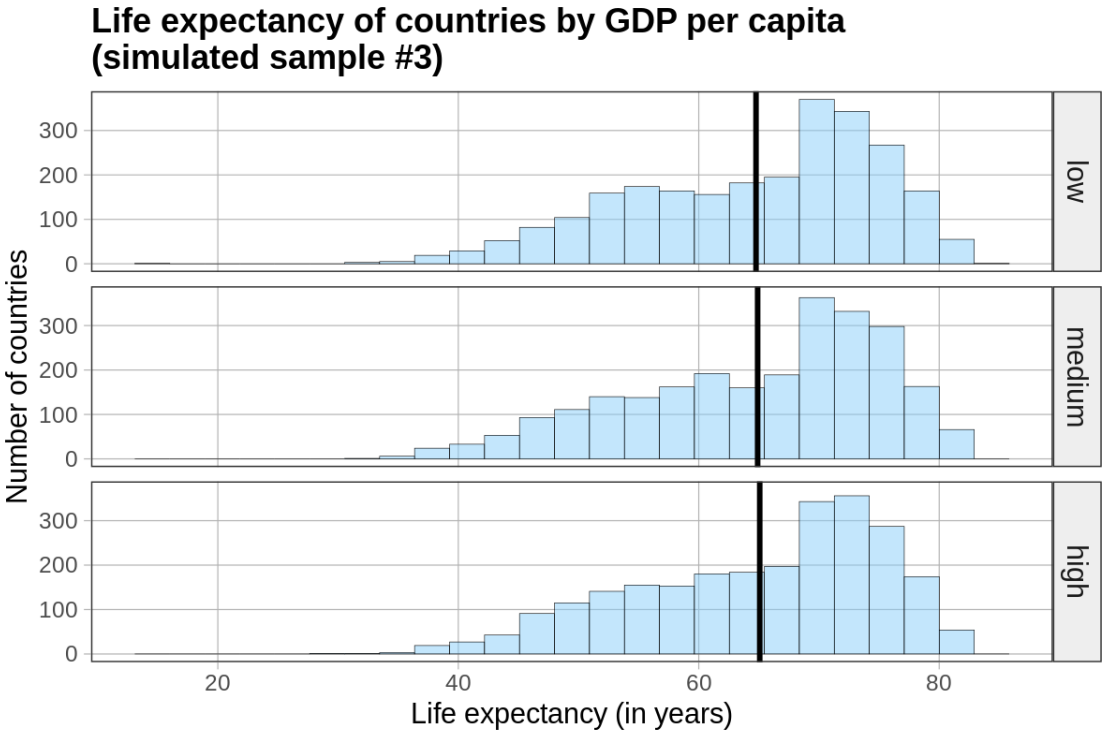
	gdpPerCapita.group_shuffled.1 <fct>	gdpPerCapita.group_shuffled.2 <fct>	gdpPerCapita.group <fct>
2	low	low	medium
5	medium	medium	high
8	low	low	high
9	medium	high	low
11	medium	medium	medium

[9]:
```
# simulated sample 1
data.frame_2 <- gapminder%>%
group_by(gdpPerCapita.group_shuffled.1)%>%
summarize(mean = mean(life_expectancy))
```

```r
gf_histogram(~life_expectancy, data = gapminder, fill = "aquamarine3")%>%
gf_facet_grid(gdpPerCapita.group_shuffled.1 ~ .)%>%
gf_vline(xintercept = ~mean, data = data.frame_2, color = "black")%>%
gf_labs(title = "Life expectancy of countries by GDP per capita\n(simulated␣
 ↪sample #1)", x = "Life expectancy (in years)", y = "Number of countries")

# simulated sample 2
data.frame_3 <- gapminder%>%
group_by(gdpPerCapita.group_shuffled.2)%>%
summarize(mean = mean(life_expectancy))

gf_histogram(~life_expectancy, data = gapminder, fill = "lightcoral")%>%
gf_facet_grid(gdpPerCapita.group_shuffled.2 ~ .)%>%
gf_vline(xintercept = ~mean, data = data.frame_3, color = "black")%>%
gf_labs(title = "Life expectancy of countries by GDP per capita\n(simulated␣
 ↪sample #2)", x = "Life expectancy (in years)", y = "Number of countries")

# simulated sample 3
data.frame_4 <- gapminder%>%
group_by(gdpPerCapita.group_shuffled.3)%>%
summarize(mean = mean(life_expectancy))

gf_histogram(~life_expectancy, data = gapminder, fill = "lightskyblue")%>%
gf_facet_grid(gdpPerCapita.group_shuffled.3 ~ .)%>%
gf_vline(xintercept = ~mean, data = data.frame_4, color = "black")%>%
gf_labs(title = "Life expectancy of countries by GDP per capita\n(simulated␣
 ↪sample #3)", x = "Life expectancy (in years)", y = "Number of countries")
```


Life expectancy of countries by GDP per capita
(simulated sample #1)
Number of countries
300
200
100
0
low
medium
high
20
40
60
80
Life expectancy (in years)


Life expectancy of countries by GDP per capita
(simulated sample #2)
Number of countries
300
200
100
0
low
medium
high
20
40
60
80
Life expectancy (in years)



The simulated random DGP produced data that differs heavily from the empirical sample. In the empirical sample, each of the different GDP per capita groups' distribution of `life_expectancy` has a different shape, spread, and mean. In the simulated samples, all three GDP per capita groups' distributions of `life_expectancy` have nearly the same shape, spread, and center.

In the empirical sample, there is a clear relationship between GDP per capita (`gdpPerCapita.group`) and life expectancy (`life_expectancy`). However, in the simulated samples, where each GDP per capita is randomly reassigned to a different observation, there is no apparent correlation between the two variables. Therefore, due to the similarities between each of the simulated samples, and the differences between the simulated samples and the empirical sample, it can be inferred that a country's GDP per capita has a direct relationship with the life expectancy of its people.

The results of my analysis imply that a country's GDP per capita is directly related its residents' life expectancy. In other words, the wealthier a country is, the better it can promote the health of its people, including how long they generally live. This means that, in countries more strongly afflicted by poverty, not only do people live more difficult lives, they also live for less time, which shows how unfair life can be for less privileged people born into unfortunate circumstances.

0.1 Instructions

There are four main components to this data analysis report: 1. Introduction 2. Explore Variation 3. Modeling a Random DGP 4. Conclusion

Make sure that your report reads seamlesslessly. If you were to delete all the blue boxes, your report should still make sense! Your evidence should be embedded into your report, not just added at the very end. If you need grammar and spelling help, type into a Word or Google Doc first and then copy the text back in. See the rubric in Canvas for more details on expectation.

0.2 Reminders

- This is an individual project.
- The work you produce must be your own (do not claim other work as your own).
- If you borrowed ideas, make sure to reference your sources.
- You may work with others for help (but make sure your data sets are different).
- Receving help does not mean you have others do the work for you.

0.3 1.0 - Intro/ Overview of the Problem or Question

The goal of this section is to provide an overview of the **context, situation,** or **problem.** Below are some questions for you to consider while writing the report.

Introduce the **context** - Which data frame are you using and why? - Who has provided the data (as much as you know)? In general, why do you think they collected this data? - What are the cases and variables in the data frame? Showcase your data frame as evidence. - What outcome variable do you want to explore? *Showcase and describe that distribution.* (Remember that statistics is the study of *variation* – so be sure to pick a variable that has a lot of variation!)

Introduce the beginnings of your **situation** or **problem**: - What research question do you want to explore with the data? In other words, what quantitative outcome variable will you explore and what variable will you use to predict it? - What are your initial predictions about this situation? - What word equation might represent your hypothesis?

0.4 2.0 - Exploring Variation

The goal of this section is to explore variation in your explanatory and outcome variables. That exploration will almost certainly include visual displays of your data. Below are some questions for you to consider.

Start your exploration:

- What visualizations would best convey the variability you're trying to explain? (You can include multiple visualizations if you think they will be helpful to your readers.)
- If needed, would it help to create new variables based on the existing ones to better see any differences?
- Are there any mistakes in the data that you found that need to be fixed? For example, maybe there are some cases that need to be excluded? Or maybe you had some missing data that you needed to handle?

Start your analysis:

- What do you see in your visualization?

- What are some reasons (from the data) for suspecting that this explanatory variable explains some of the variation in the outcome? How about reasons for suspecting that this explanatory variable does **NOT** explain some of the variation in the outcome?
- So far, just based on these visualizations, what do you think about your research question or hypothesis?

0.5 3.0 - Modeling a Random DGP

The goal of this section is to determine whether the pattern you're observing in your data is coming from a random DGP or not.

- What is the empty model? Add your model to your visualization.
- What is the GLM notation for your model?
- What do all parts of your GLM mean in terms of your context?
- Try mimicking a random data generating process using the shuffle function. Shuffle one (or both) of your variables in your visualization. Does a random DGP generally produce data similar to your empirical sample? Or do you see patterns that are very different? Describe your results. Save several of your shuffle examples as evidence.
- From what we have done so far, are we able to determine whether your explanatory variable causes changes in the outcome variable? Why or why not? Explain completely.

0.6 4.0 Conclusion

The goal of this section is to help your audience understand what can be learned from your data analysis. You are reminding them of how all this work relates back to your initial word equation.

Questions to consider: - Why did you choose to pursue this research question and what were you predicting? - How would you summarize all that you did, what you found, and how it relates to the motivating question? - What are the implications of the results and what it means for the audience or the world? - How well do you think your model represents the DGP? Are there any other factors that did not previously consider that you are now curious about?

0.6.1 Submit your first draft via Canvas here

- There are two submission dates for the first draft (the sooner you submit the sooner you can get your feedback and work on the final draft).

- Wednesday 12/07/2022
- Friday 12/09/2022

0.6.2 Submit your final draft via Canvas here

- The final draft is due by Wednesday 12/14/2022
- You can submit sooner that the due date

0.6.3 Prepare for your presentation on the day of the final (Wednesday 12/14/2022)

- You will need to prepare slides to go with your presentation.
- Go here to read instructions and expectations for your slides and presentation.