Exam 01

COMP/STAT112 (Spring 2025)

Name: Pablo Jiménez Section: 01

1 Background

You will work with the Food Consumption and CO2 Emissions dataset used in Week 8 of year 2020 (Feb 2, 2020) in the TidyTuesday project. The dataset contains one file as described in the Data Dictionary below.

1.1 Data Dictionary

`food_consumption.csv`

variable	class	description
country	character	Country Name
food_category	character	Food Category
consumption	double	Consumption (kg/person/year)
co2_emmission	double	Co2 Emission (Kg CO2/person/year)

1.2 Grand Research Question

Using an appropriate viz, you need to answer the following grand research question:

~~What does the consumption of each food category in each country look like~~?

To answer this question, a data scientist need to form a good understanding of the dataset first. Follow the steps below and answer any prompt that each might have.

2 Install Packages

```
install.packages("tidytuesdayR")    # to download dataset from TidyTuesday project
install.packages("tidyverse")       # for visualization
```

Working with this dataset required the packages listed in the code chunk above. Including the above code chunk in the Quarto file is not appropriate. Why? What should be done instead?

Because it would be redundant for someone who already has the packages installed, we can instead just load them by using library(...)

3 Load Packages

```
library(tidytuesdayR)
```

```
Warning: package 'tidytuesdayR' was built under R version 4.3.3
```

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.3

Warning: package 'readr' was built under R version 4.3.3

Warning: package 'dplyr' was built under R version 4.3.3

Warning: package 'stringr' was built under R version 4.3.3

Warning: package 'lubridate' was built under R version 4.3.3

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Inspect the warning message shown as a result of running the code chunk above. How many packages were loaded when loading the `tidyverse` package? 6 Circle them in the output.

4 Get Data

```
tuesdata <- tt_load('2020-02-18')
fc <- tuesdata$food_consumption
```

What does the above code chunk do?

it stores the data of tt_load('2020-02-18') in the value tuesdata which later on is stored in "fc" without the food consumption title on the columns.

5 Understand Data

List a minimum of three initial steps that should be carried after loading the above dataset and the corresponding R functions to accomplish each.

#	Step	R function
1	View the data	View(...)
2	check the types of data	str(...)
3	check the number of rows	nrow(...)
4		
5		

6 Explore Data

6.1 Top Observations

	country	food_category	consumption	co2_emmission
1	Argentina	Pork	10.51	37.20
2	Argentina	Poultry	38.66	41.53
3	Argentina	Beef	55.48	1712.00
4	Argentina	Lamb & Goat	1.56	54.63
5	Argentina	Fish	4.36	6.96
6	Argentina	Eggs	11.39	10.46
7	Argentina	Milk - inc. cheese	195.08	277.87

8	Argentina	Wheat and Wheat Products	103.11	19.66
9	Argentina	Rice	8.77	11.22
10	Argentina	Soybeans	0.00	0.00
11	Argentina	Nuts inc. Peanut Butter	0.49	0.87
12	Australia	Pork	24.14	85.44
13	Australia	Poultry	46.12	49.54
14	Australia	Beef	33.86	1044.85
15	Australia	Lamb & Goat	9.87	345.65
16	Australia	Fish	17.69	28.25
17	Australia	Eggs	8.51	7.82
18	Australia	Milk - inc. cheese	234.49	324.01
19	Australia	Wheat and Wheat Products	70.46	13.44
20	Australia	Rice	11.03	14.12
21	Australia	Soybeans	0.19	0.09
22	Australia	Nuts inc. Peanut Butter	8.73	15.45

6.2 Bottom Observations

	country	food_category	consumption	co2_emission
1409	Liberia	Pork	4.01	14.19
1410	Liberia	Poultry	8.91	9.57
1411	Liberia	Beef	0.78	24.07
1412	Liberia	Lamb & Goat	0.48	16.81
1413	Liberia	Fish	4.13	6.59
1414	Liberia	Eggs	2.05	1.82
1415	Liberia	Milk - inc. cheese	3.04	4.33
1416	Liberia	Wheat and Wheat Products	10.95	2.09
1417	Liberia	Rice	94.75	121.25
1418	Liberia	Soybeans	0.63	0.28
1419	Liberia	Nuts inc. Peanut Butter	1.31	2.32
1420	Bangladesh	Pork	0.00	0.00
1421	Bangladesh	Poultry	1.40	1.50
1422	Bangladesh	Beef	1.28	39.50
1423	Bangladesh	Lamb & Goat	1.33	46.58
1424	Bangladesh	Fish	18.07	28.85
1425	Bangladesh	Eggs	2.08	1.91
1426	Bangladesh	Milk - inc. cheese	21.91	31.21
1427	Bangladesh	Wheat and Wheat Products	17.47	3.33
1428	Bangladesh	Rice	171.73	219.76
1429	Bangladesh	Soybeans	0.61	0.27
1430	Bangladesh	Nuts inc. Peanut Butter	0.72	1.27

6.3 Observations

Look at the top and bottom 22 observations from the dataset printed above. What are the units of observations?

Each row represents the consumption, food category and CO_2 emissions from each specific country

How many food categories are there? 11 n_distinct(fc$'food_category')

How many countries are there? 24

7 Understand Variables Individually

How many variables does the grand research question involve? 2

Before answering the grand research question, a data scientist needs to understand the distribution of each involved variable. List all the involved variables in the table below with **one** appropriate plot type that can be used to visualize it without worrying about the R code details.

#	Variable	Plot Type
1	Food category	Barchart
2	Consumption	Histogram
3		
4		
5		

8 Understand Consumption

Let us also try to understand the **overall** food consumption for (1) each food category (2) each country. List one appropriate plot for each bivariate viz and what should goes into their aesthetic without worrying about the R code details.

Bivariate Viz	Plot Type	Aesthetic Details
Overall Food Consumption / Food Category	Boxplot	x-axis Food category y-axis Food consumption
Overall Food Consumption / Country	Boxplot	x-axis country y-axis ~~food~~ Food consumption

9 Answering Grand RQ

List as many plot types (consider also their varieties) that can be used to answer the grand research question then list what should goes into their aesthetic (without worrying about its R code details) and what are some of the potential challenges you might face.

#	Plot Type	Aesthetic Details	Potential Challenges
1	Bar chart	~~Food category~~ = x = County Food ~~Color~~ = y = Food ~~category~~ Consumption color = food type	Wouldn't be as easy to visualize other values (mean)
2	Histograms	//	//
3	Point plot	x = county y = Food consumption color = food type	Would be able to distinguish the number
4	Boxplot	x county = y = food consumption with food type	don't see challenges but how to use a boxplot
5	~~Choropleth map~~	~~color = food type~~,	
6			

Which of these plots is the most appropriate one? Why?

I would say that a bar chart would be the best plot to visualize the data

10 Beyond Viz

10.1 Effectiveness

List a minimum of five concepts that you should apply to your final viz to make it more effective?

1. Make it color blind friendly for further analysis
2. Could add captions for the author, the date or source
3. Could have a title and labels
4. Needs to have a purpose and not be misleading
5. Doesn't have to be misleading
6.
7.

10.2 Additional Questions

List two additional questions, new or follow-up, that you would like to answer based on the this dataset.

1. What would be the average and outliers in the data
2. ~~Where would the~~ How would the data ~~to~~ look like in a choropleth map
3. How does the consumption looks like in each country
4.