Name(s): _____

DATA 101 Assignment 2: ANSWER KEY

Work in teams of 2

Instructions. Answer the following questions using the linked datasets and activities. Always note the source of your information when applicable. Write your responses in the spaces provided and show your work.

1. Complete this brief survey about your interests: https://forms.gle/ZG3U1yeCWdbbbNCe6. **My answers:** I majored in Economics; I'm from Türkiye; I play guitar with the faculty band and play soccer; My favorite movie is Big Fish; My MBTI type is ENFJ; My dream job is being a professor!

2. Complete the activity about Johann Sebastian Bach's works from Week 2's course slides on Github, page 6. Write your answers in the space provided below.

   We could measure each piece's key, note density, number of instruments, among other characteristics. This information needs to come from recordings. Today, we have many streaming services that include Bach's catalogue so those recordings can be downloaded and fed into computer for analysis. Some measurement errors could center around differences in interpretations of Bach's original work.

3. Go to `tidy tuesday`'s repository. Find a dataset you are interested in. Write or paste below the dataset in a spreadsheet format, including the first 5 rows. Why did you pick this dataset, and what makes it interesting to you? What are the units of the dataset? (Recall: the unit of a dataset is the object or subject we took measurements from.)

	id	pokemon	species_id	height	weight	base_experience	type_1	type_2	hp	attack	defense	special_attack	special_defense	speed	color_1	color_2	color_f	egg_group_1	egg_group_2	url_icon	generation_id
1	id	pokemon	species_id	height	weight	base_experience	type_1	type_2	hp	attack	defense	special_attack	special_defense	speed	color_1	color_2	color_f	egg_group_1	egg_group_2	url_icon	generation_id
2	1	bulbasaur	1	0.7	6.9	64	grass	poison	45	49	49	65	65	45	#78C850	#A040A0	#81A763	monster	plant	//archives.bulbagarden.net/media/upload/7/7b/001MS6.png	1
3	2	ivysaur	2	1	13	142	grass	poison	60	62	63	80	80	60	#78C850	#A040A0	#81A763	monster	plant	//archives.bulbagarden.net/media/upload/a/a0/002MS6.png	1
4	3	venusaur	3	2	100	236	grass	poison	80	82	83	100	100	80	#78C850	#A040A0	#81A763	monster	plant	//archives.bulbagarden.net/media/upload/0/07/003MS6.png	1
5	4	charmander	4	0.6	8.5	62	fire	NA	39	52	43	60	50	65	#F08030	NA	NA	monster	dragon	//archives.bulbagarden.net/media/upload/7/7d/004MS6.png	1
6	5	charmeleon	5	1.1	19	142	fire	NA	58	64	58	80	65	80	#F08030	NA	NA	monster	dragon	//archives.bulbagarden.net/media/upload/b/be/005MS6.png	1

   I chose the Pokémon dataset. It is interesting how and who specified Pokémon characteristics which as you can see take on numerical form. The units of the dataset are Pokémon. One can analyze how Pokémon strengh is distributed, if monotonically increase with evolution among other possibilities.

4. Go to https://enrollment.streamlit.app. Find the information below.

   (a) What was the course with the highest enrollment in Althouse 201 and when? It's a three way tie between Fundamentals of Business (Fall 2021) and Introduction to Microeconomics sections 1 and 2 (Spring 2024) with 40 students enrolled.

   (b) What was the course with the highest enrollment in the Althouse building and when? It was Introduction to Macroeconomics with 41 students (Spring 2023).

(c) What was the course with the highest enrollment that I taught and when? It was Elementary Statistics with 24 students (Fall 2021).

5. In each case determine whether the variable is categorical or numerical. If the variable is numerical, state whether it is discrete or continuous.

    (a) The number of people entering the post office between 1pm and 3 pm on Monday Discrete numerical.

    (b) The numbers on the uniforms of a soccer team. Categorical.

    (c) The length of time to swim 50 yards. Discrete continuous.

    (d) The brand of phone purchased by a customer. Categorical.

    (e) Credit card numbers. Categorical.

6. Below are the numbers of billionaires from five different states (consider this a sample of states) in the Northeast (according to *Forbes.com*).

State	Billionaires
New York	67
Connecticut	11
Pennsylvania	7
Massachusetts	7
New Jersey	5

    (a) Compute the mean number of billionaires. Show your work.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{67 + 11 + 7 + 7 + 5}{5} = \frac{97}{5} = 19.4$$

    (b) Compute the standard deviation (by hand and show your work) of the number of billionaires (round to the nearest tenth).
    Formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

Using the values ($n = 5$, $\bar{x} = 19.4$):

$$s = \sqrt{\frac{(67 - 19.4)^2 + (11 - 19.4)^2 + (7 - 19.4)^2 + (7 - 19.4)^2 + (5 - 19.4)^2}{5 - 1}}$$

Compute each term:
$$(67 - 19.4)^2 = 47.6^2 = 2265.76$$
$$(11 - 19.4)^2 = (-8.4)^2 = 70.56$$
$$(7 - 19.4)^2 = (-12.4)^2 = 153.76$$
$$(7 - 19.4)^2 = (-12.4)^2 = 153.76$$
$$(5 - 19.4)^2 = (-14.4)^2 = 207.36$$

$$s = \sqrt{\frac{2265.76 + 70.56 + 153.76 + 153.76 + 207.36}{4}} = \sqrt{\frac{2851.2}{4}} = \sqrt{712.8} \approx 26.7$$

(c) Which of the given numbers contributes the most to the standard deviation? Why?
The largest squared deviation is

$$(67 - 19.4)^2 = 2265.76$$

from New York. Because this value is much larger than the others, as an outlier, New York contributes the most to the standard deviation.