Problem Set 2: Intro to Data

1. **Marvel Cinematic Universe films.** The data frame below contains information on Marvel Cinematic Universe films through the Infinity saga (a movie storyline spanning from Ironman in 2008 to Endgame in 2019). Box office totals are given in millions of US Dollars.

		Length				Gross	
	Title	Hrs	Mins	Release Date	Opening Wknd US	US	World
1	Iron Man	2	6	5/2/2008	98.62	319.03	585.8
2	The Incredible Hulk	1	52	6/12/2008	55.41	134.81	264.77
3	Iron Man 2	2	4	5/7/2010	128.12	312.43	623.93
4	Thor	1	55	5/6/2011	65.72	181.03	449.33
5	Captain America: The First Avenger	2	4	7/22/2011	65.06	176.65	370.57
...	...	...	...	...	...	...	...
23	Spiderman: Far from Home	2	9	7/2/2019	92.58	390.53	1131.93

   a. How many observations and how many variables does this data frame have? What is the observational unit (what each row corresponds to)?

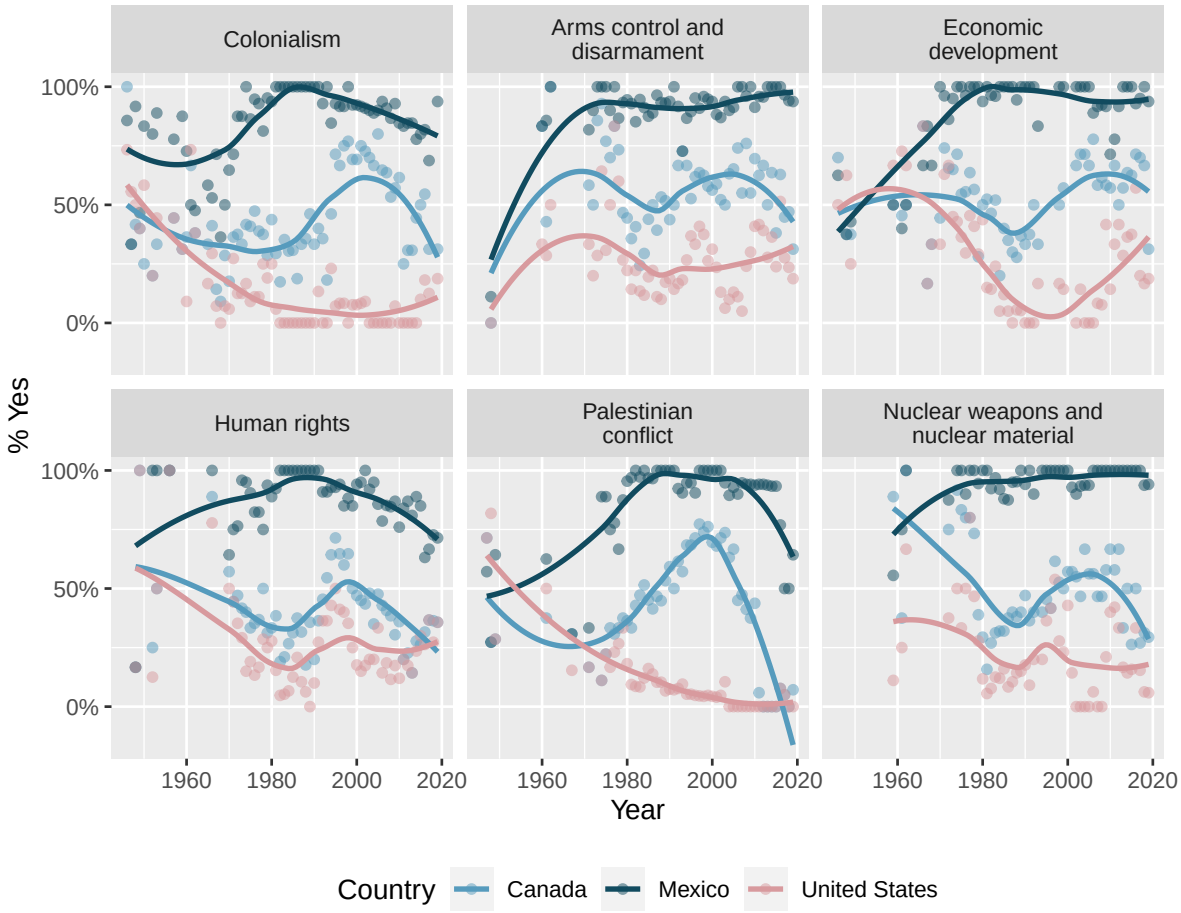
2. **Cherry Blossom Run.** The data frame below contains information on runners in the 2017 Cherry Blossom Run, which is an annual road race that takes place in Washington, DC. Most runners participate in a 10-mile run while a smaller fraction take part in a 5k run or walk.

						Time			
	Bib	Name	Sex	Age	City / Country	Net	Clock	Pace	Event
1	6	Hiwot G.	F	21	Ethiopia	3217	3217	321	10 Mile
2	22	Buze D.	F	22	Ethiopia	3232	3232	323	10 Mile
3	16	Gladys K.	F	31	Kenya	3276	3276	327	10 Mile
4	4	Mamitu D.	F	33	Ethiopia	3285	3285	328	10 Mile
5	20	Karolina N.	F	35	Poland	3288	3288	328	10 Mile
...	...	...	...	...	...	...	...	...	...
19961	25153	Andres E.	M	33	Woodbridge, VA	5287	5334	1700	5K

   a. How many observations and how many variables does this data frame have? What is the observational unit?

   b. What type of data is captured by each of the variables in the data frame? Are there any that you find ambiguous?

3. **UN Votes.** The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2019, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.



a. List the variables used in creating this visualization.

b. What appears to be the observational unit?

c. Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

4. *__Space launches.__ The following summary table shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success or failure).[1]

	1957 - 1999		2000-2018	
	Failure	Success	Failure	Success
Private	13	295	10	562
State	281	3751	33	711
Startup	0	0	5	65

a. What variables were collected on each launch in order to create to the summary table above?

b. State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

c. Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

[1]The data used in this exercise comes from the JSR Launch Vehicle Database, 2019 Feb 10 Edition.

d. What kind of a graphic would you use to describe the success rate of launches broken down by agencies and time period? Describe how each variable would be mapped to an aesthetic cue of the graphic.

5. ***Views on immigration.** Nine-hundred and ten (910) randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.

Response	Conservative	Liberal	Moderate
Apply for citizenship	57	101	120
Guest worker	121	28	113
Leave the country	179	45	126
Not sure	15	1	4

a. What percent of these Tampa, FL voters identify themselves as conservatives?

b. What percent of these Tampa, FL voters are in favor of the citizenship option?

c. What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

d. What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

e. Do political ideology and views on immigration appear to be associated? Explain your reasoning.

f. Conjecture other possible variables that might explain the potential relationship between these two variables.