PS 1

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
6	Marvel's The Avengers	2	23	5/4/2012	207.44	623.36	1518.82
7	Iron Man 3	2	10	5/3/2013	174.14	409.01	1214.81
8	Thor: The Dark World	1	52	11/8/2013	85.74	206.36	644.78
9	Captain America: The Winder Soldier	2	16	4/4/2014	95.02	259.77	714.42
10	Guardians of the Galaxy	2	1	8/1/2014	94.32	333.72	773.34
...	...	...	...	...	...	...	...
22	Avengers: Endgame	3	1	4/26/2019	357.12	858.37	2797.8
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)?
   b. (Required for Marvel fans) Which movie on the list made the most money during the opening week? Which made the least? Does this outcome surprise you? What could be the possible reasons?

2. **Smoking habits of UK residents.** A survey was conducted to study the smoking habits of 1,691 UK residents. Below is a data frame displaying a portion of the data collected in this survey. A blank cell indicates that data for that variable was not available for a given respondent.[1]

						amount	
	sex	age	marital_status	gross_income	smoke	weekend	weekday
1	Female	61	Married	2,600 to 5,200	No	NA	NA
2	Female	61	Divorced	10,400 to 15,600	Yes	5	4
3	Female	69	Widowed	5,200 to 10,400	No	NA	NA
4	Female	50	Married	5,200 to 10,400	No	NA	NA
5	Male	31	Single	10,400 to 15,600	Yes	10	20
...	...	...	...	...	...	NA	NA
1691	Male	49	Divorced	Above 36,400	Yes	15	10

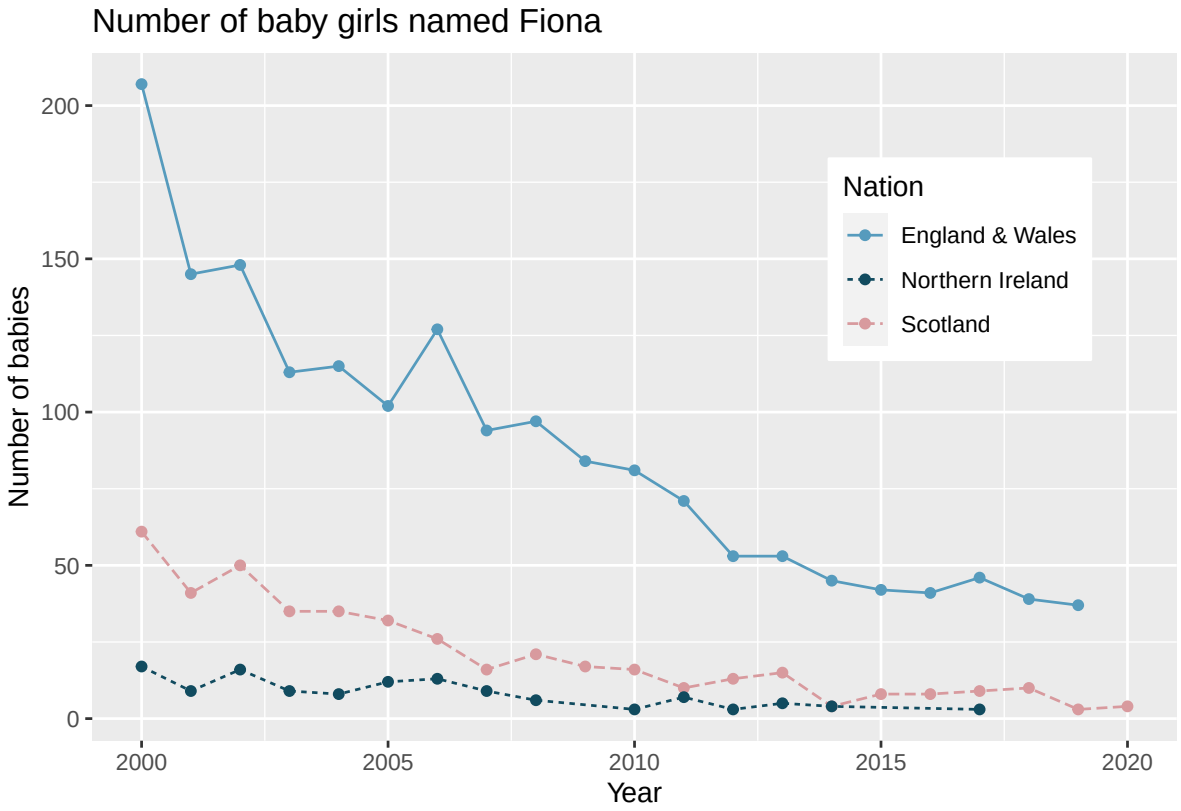
[1] The `smoking` data used in this exercise can be found in the **openintro** R package.

a. What does each row of the data frame represent?

b. How many participants were included in the survey?

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.

3. **UK baby names.** The visualization below shows the number of baby girls born in the United Kingdom (comprised of England & Wales, Northern Ireland, and Scotland) who were given the name "Fiona" over the years.[2]



a. List the variables you believe were necessary to create this visualization.

b. 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).[3]

	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?

[2]The `ukbabynames` data used in this exercise can be found in the **ukbabynames** R package.
[3]The data used in this exercise comes from the JSR Launch Vehicle Database, 2019 Feb 10 Edition.

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?

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.