Part I

1. $\mathbb{P}(A) = 0.3$, $\mathbb{P}(B) = 0.7$

   (a) Can you compute $\mathbb{P}(A \text{ and } B)$ if you only know $\mathbb{P}(A)$ and $\mathbb{P}(B)$?

   (b) Assuming that events A and B arise from independent random processes,

      (i) what is $\mathbb{P}(A \text{ and } B)$?

      (ii) what is $\mathbb{P}(A \text{ or } B)$?

      (iii) what is $\mathbb{P}(A \mid B)$?

   (c) If we are given that $\mathbb{P}(A \text{ and } B) = 0.1$, are the random variables giving rise to events A and B independent?

   (d) If we are given that $\mathbb{P}(A \text{ and } B) = 0.1$, what is $\mathbb{P}(A \mid B)$?

2. A 2010 Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, where the counts have been replaced with relative frequencies.

		Response			
		Earth is warming	Not warming	Don't Know Refuse	Total
	Conservative Republican	0.11	0.20	0.02	0.33
Party and Ideology	Mod/Lib Republican	0.06	0.06	0.01	0.13
	Mod/Cons Democrat	0.25	0.07	0.02	0.34
	Liberal Democrat	0.18	0.01	0.01	0.20
	Total	0.60	0.34	0.06	1.00

   (a) Are believing that the earth is warming and being a liberal Democrat mutually exclusive?

   (b) What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat?

   (c) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a liberal Democrat?

   (d) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a conservative Republican?

   (e) Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Explain your reasoning.

   (f) What is the probability that a randomly chosen respondent is a moderate/liberal Republican given that he does not believe that the earth is warming?

3. Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease. The test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from the Fox television show *House* that is often used after a patient tests positive for lupus: "It's never lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer.

4. At a university, 13% of students smoke.

   (a) Calculate the expected number of smokers in a random sample of 100 students from this university.

(b) The university gym opens at 9 am on Saturday mornings. One Saturday morning at 8:55 am there are 27 students outside the gym waiting for it to open. Should you use the same approach from part (a) to calculate the expected number of smokers among these 27 students?

5. Consider the following card game with a well-shuffled deck of cards. If you draw a red card, you win nothing. If you get a spade, you win $5. For any club, you win $10 plus an extra $20 for the ace of clubs.

   (a) Define a random variable that describes the amount you win at this game, with the possible values that it can take along with their probabilities. Also, find the expected winnings for a single game and the standard deviation of the winnings.

   (b) What is the maximum amount you would be willing to pay to play this game? Explain your reasoning.

6. An airline charges the following baggage fees: $25 for the first bag and $35 for the second. Suppose 54% of passengers have no checked luggage, 34% have one piece of checked luggage and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

   (a) Define a random variable that describes the baggage fee revenue for a single passenger, with the possible values that it can take along with their probabilities. The compute the average revenue per passenger, and compute the corresponding standard deviation.

   (b) About how much revenue should the airline expect for a flight of 120 passengers? With what standard deviation? Note any assumptions you make and if you think they are justified.

Optional Challenge Problem: A *chord* of a circle is a straight line segment whose endpoints both lie on the circle. For a fixed circle, what is the probability that the length of a randomly drawn chord will exceed that circle's radius?

Part II

Inside the `tidyverse` package is a package called `dplyr` that has powerful tools for data wrangling. Here you'll get a bit of practice with it on the data that the class collected from the thesis tower last week. First, load the packages and data.

```
library(tidyverse)
library(oilabs)
data(theses)
```

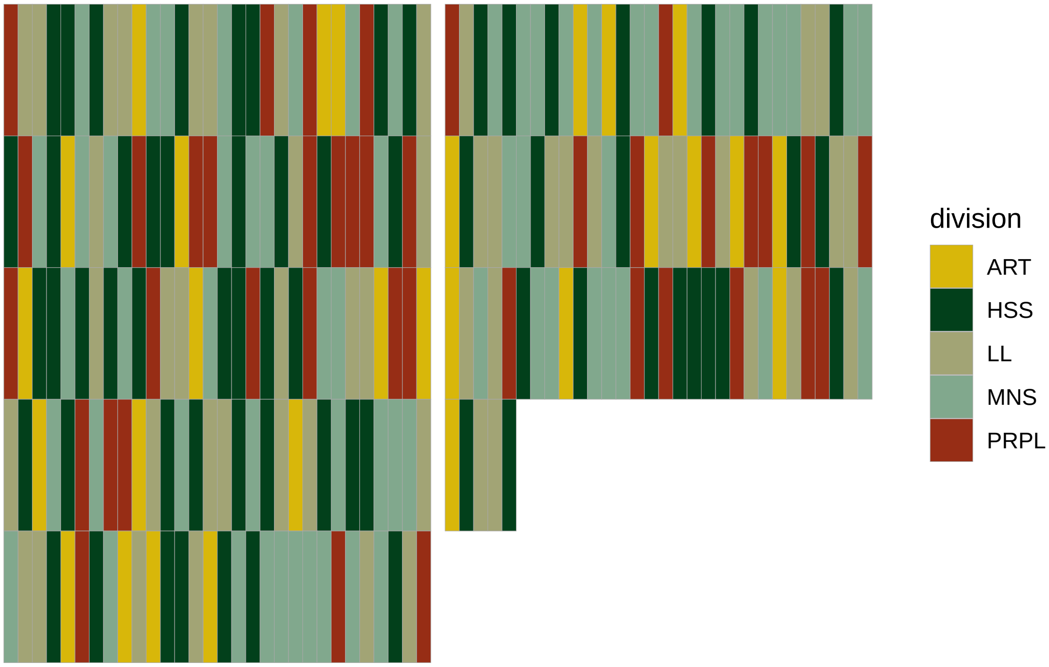
7. `arrange()` is a function used to sort a dataframe either numerically or alphabetically. In general, if you want to arrange `data` by the ascending order of the `var` variable, the usage is

```
data %>%
  arrange(var)
```

To reverse the ordering, you simply change `var` to `desc(var)`.

Arrange the `theses` data in descending order based on checkouts (print out both the code and the output, which shows the first 10 rows). Who is this "best-selling" author?

8. In the plot below, each colored box represents a thesis and they are laid out on a "bookshelf" alphabetically by the author's last name.



Use `arrange()` to print out the first 10 rows of the dataframe, sorted alphabetically by the author's last name. Use this to verify that the first 10 theses' divisions are the same divisions shown in the upper left corner of the plot.

9. `filter()` is a function that is used to subset the rows of a data set using logical operators. Create a new data set that contains only the students that identify as scientists.

10. `mutate()` is a function that creates new variables based on the existing variables in the dataframe. For example, if you wanted to create a new variable called `sumvars` that was the sum of two of the existing variables, you would use

```
data %>%
  mutate(sumvars = var1 + var2)
```

Create a new column that stores the number of checkouts per year (from the date of publication until 2020).

11. `summarize()` is a function used to summarize a column of data with a single number. You can do many at once, as in:

```
data %>%
  summarize(avg = mean(var1),
            med = median(var1),
            sd  = mean(var2))
```

Use this function compute the mean and median number of pages.

12. A final non-wrangling task: visualize the distribution of the number of pages using the geom of your choice.