Task 1

You've got about 20 years of individual level case data, and you'd like to generate a table of annual case counts and rates to use in a presentation.

See: https://www.washtenaw.org/2617/Tuberculosis-TB-Information

Washtenaw County TB Data

Year	# Active TB Cases	# Pulmonary Cases	# Multi-drug Resistant Cases	Washtenaw Active TB rate (all types) per 100,000
2024	8	5	0	2.15
2023	9	8	0	2.4
2022	8	6	1	2.1
2021	4	3	0	1.0
2020	6	6	0 (1 INH resistant)	1.6
2019	6	5	0	1.6
2018	9	5	1	2.4
2017	8	5	0	2.2
2016	10	7	0	2.8
2015	6	5	0	1.7
2014	7	2	0	2.0
2013	4	3	0	1.2
2012	5	4	0	1.5
2011	8	5	0	2.3
2010	9	8	0	2.6
2009	6	5	0	1.9
2008	10	7	0	3.1
2007	6	5	1	1.9
2006	7	5	0	2.2
2005	10	3	0	3.1

Steps:

1. Load the libraries you think you might need! In this case primarily `tidyverse` to start (plus maybe more later).

Consider using `library()`, `install.packages()`

2. Identify where your data file is located on your computer, and load the data file into R as a data frame.

Consider using `read.csv()`, `paste0()`

3. Make a new column in the data frame that is the year of each `Onset_Date` entry

Consider subsetting a vector using `[#]`, `class()`, `as.POSIXct()` or `as.Date()`, `substr()`, `as.character()`, `year()`

4. Make a new dataframe of Active TB cases per year, by grouping by year and counting the number of cases in each group.

Consider using `summarize()`, `length()`, `unique()`, `group_by()`

5. Make a second dataframe of Pulmonary TB cases per year, by filtering to pulmonary cases only, grouping by year, and counting the number of cases in each group.

Consider using `filter()`, `summarize()`, `length()`, `unique()`, `group_by()`

6. Merge the two dataframes you created in steps 3 and 4 together into one dataframe with three columns: year, active TB cases, and pulmonary TB cases.

Consider using `merge()`

7. Sort the new dataframe you made in step 5 so that the most recent year is at the top and the oldest year is at the bottom.

Consider using `arrange()`, `desc()`

8. Load in the county_pop2.csv file into R as a data frame.

Consider using `read.csv()`, `paste0()`

9. Merge the population data onto your TB case dataframe you made in step 6.

Consider using `merge()`

10. Calculate the case rate per 100,000 population, and round that value to the nearest 1 decimal place, as a new column in your dataframe from step 8.

Consider using `mutate()`, `round()`

11. Select only the four columns of interest from the dataframe you made in step 9 as a new data frame: The year, the active case rate, the pulmonary case rate, and the case rate per 100K population.

Consider using `select()`

12. Check your work, and compare your table with the original reference image/table.

13. Stretch question! If you're comfortable with this, try formatting your table to look fancy, for example using the package gt or the `kable()` function in Quarto/Rmarkdown

Task 2

You want to visualize whether cases are increasing or decreasing over the years. In addition, you'd like to quantify the extent of the increase or decrease.

1. Make a scatter plot with year on the x-axis and case rate per 100,000 population on the y-axis.

Consider using `ggplot()`, `geom_point()`, `theme_bw()`, `coord_cartesian()`, `labs()`

2. Make a new scatter plot with year on the x-axis and case rate per 100,000 population on the y-axis, and a line fitted to the points over time.

Consider using `ggplot()`, `geom_point()`, `theme_bw()`, `coord_cartesian()`, `labs()`, `geom_smooth()`

3. Break down the linear model. Is the linear relationship significant? Calculate the case rate from 2005 and the case rate from 2024. How much and in what direction have TB cases changed?

Consider using `lm()`, `summary()`

Task 3

You'd like to compare your county data to another data source. For example purposes, we'll compare to nationwide TB case rates, but the method here could be the same for any other data source (such as another county, or the state of Michigan).

1. Load in the national_tb.csv file into R as a data frame.

Consider using `read.csv()`, `paste0()`

2. Make a scatter plot with year on the x-axis and case rate per 100,000 population on the y-axis. Both the Washtenaw County data and the Nationwide data should appear on the chart, with different colored points.

Consider using `ggplot()`, `geom_point()`, `theme_bw()`, `coord_cartesian()`, `labs()`

3. Re-format the county data and the nationwide data into three columns: Year, Rate, and Area, with Area being either "Washtenaw County" or "United States". Combine these two datasets into one long dataframe.

Consider using `select()`, `colnames()`, `rbind()`

4. Make a new scatterplot using the dataframe you made in Step 3. Year should be on the x-axis and Rate should be on the y-axis. Make sure to assign “group” and “color” to Area. Then use geom_smooth() to add lines of best fit to the data. What do you notice about the two geographic trends?

Consider using `ggplot()`, `geom_point()`, `theme_bw()`, `coord_cartesian()`, `geom_smooth()`, `labs()`

5. Make a new scatterplot using the dataframe you made in Step 3 filtered to only data from 2020 to current. Year should be on the x-axis and Rate should be on the y-axis. Make sure to assign “group” and “color” to Area. Then use geom_smooth() to add lines of best fit to the data. What do you notice about the two geographic trends now?

Consider using `ggplot()`, `geom_point()`, `theme_bw()`, `coord_cartesian()`, `geom_smooth()`, `labs()`, `filter()`

Extras

You’d like to identify the multi-drug resistant set of TB cases in the dataset. (Note: There are a lot of ways to solve this particular problem!)

1. Using the original case data frame, make a smaller dataset consisting of only the InvestigationID, and all 6 included drug resistance columns.

Consider using `read.csv()`, `paste0()`*, subsetting columns in a dataframe by number*

2. In the new smaller dataframe you made in Step 1, replace all instances of “Susceptible” with 0, all instances of “Resistant” with 1, and all instances of “” with 0. (This will allow us to do some math! In each case row, if the sum of our antibiotic testing columns is greater than 1, then the individual displayed a multi-drug resistant profile)

Consider using subsetting to assign value: `df[df == “”] <- #`

3. Sum across the antibiotic testing columns, making a new column with that value.

Consider using `mutate_all()`, `rowSums()`

4. Make a new column that conditionally applies the value 1 if the column from Step 3 is > 1, and applies 0 if it is not.

Consider using `mutate()`, `case_when()`

5. Make a new data frame of only the InvestigationID and the column created in Step 4.

Consider using `select()`

6. Merge the new data frame you created in Step 5 to the original case dataframe.

Consider using `merge()`

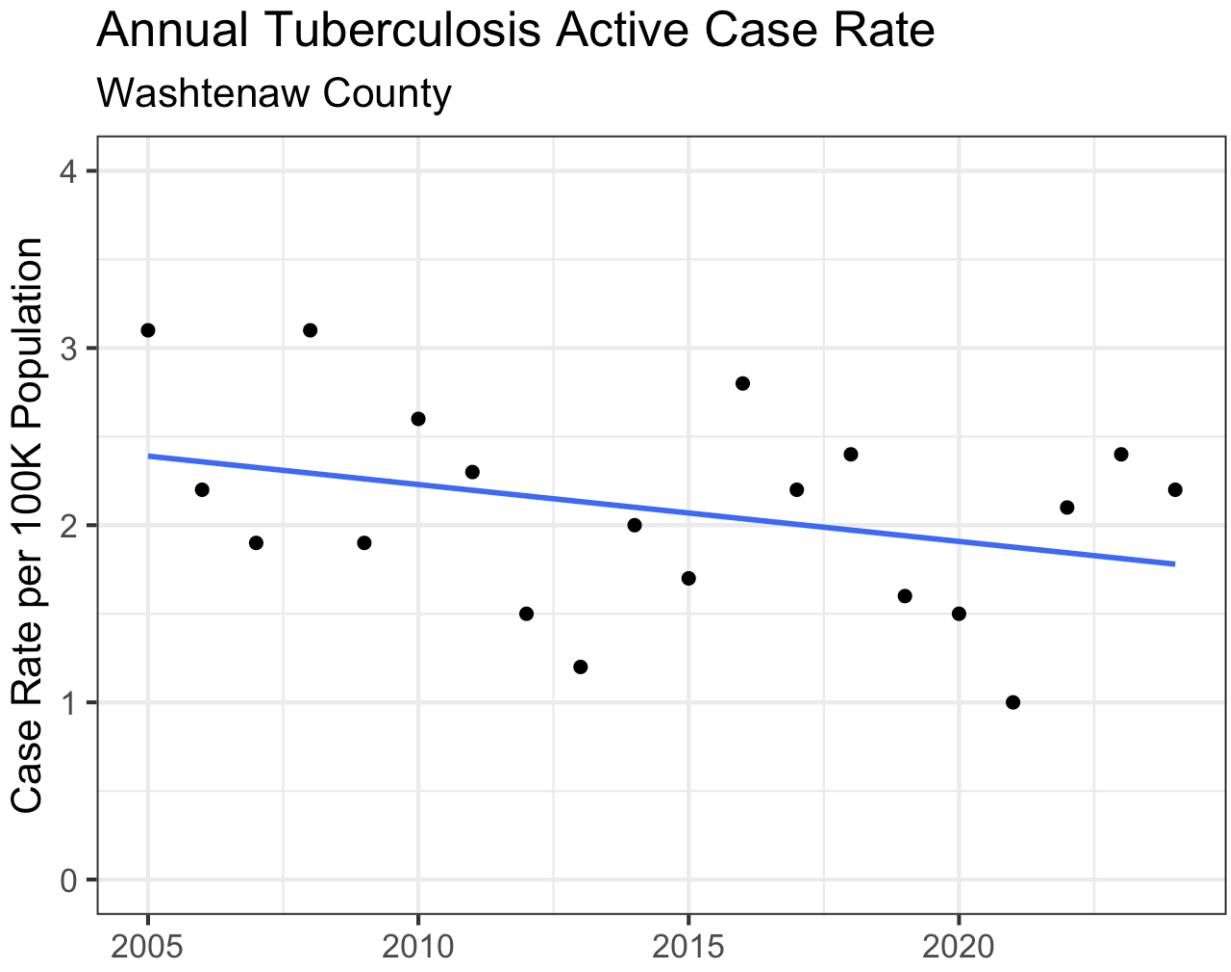
What should my output look like for each task?

If you find yourself having a tough time visualizing what you're aiming for making based on the task, below are examples of the kind of thing we were aiming for in each task—but your output doesn't have to look exactly like this! You may choose different ways to format or plot your output, this is just an example!

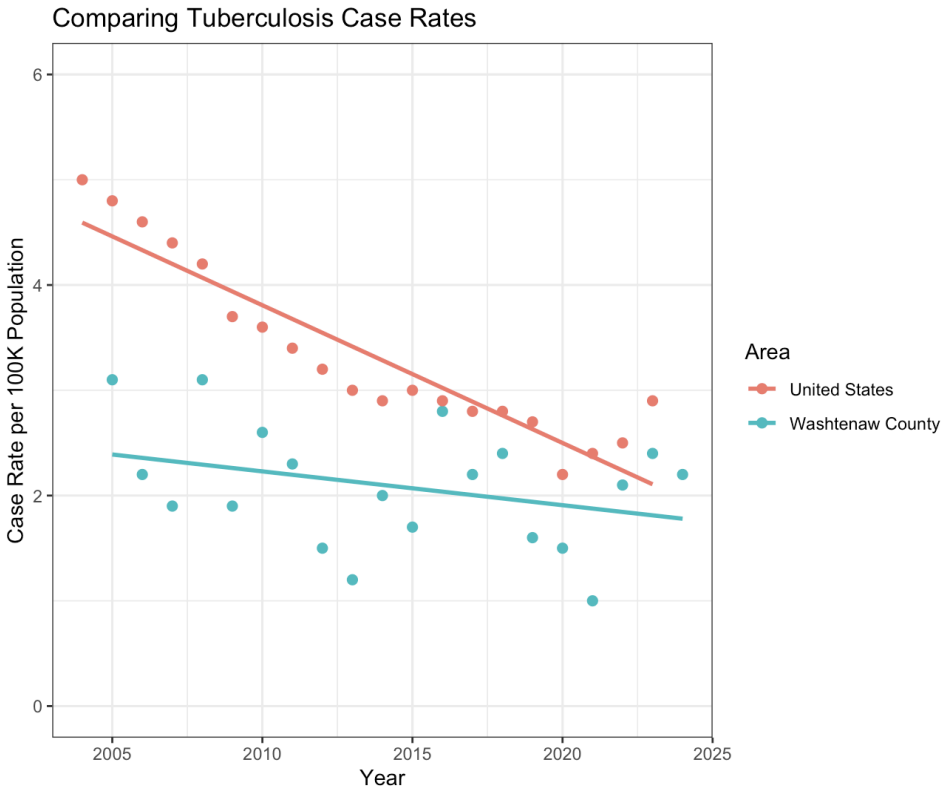
Task 1

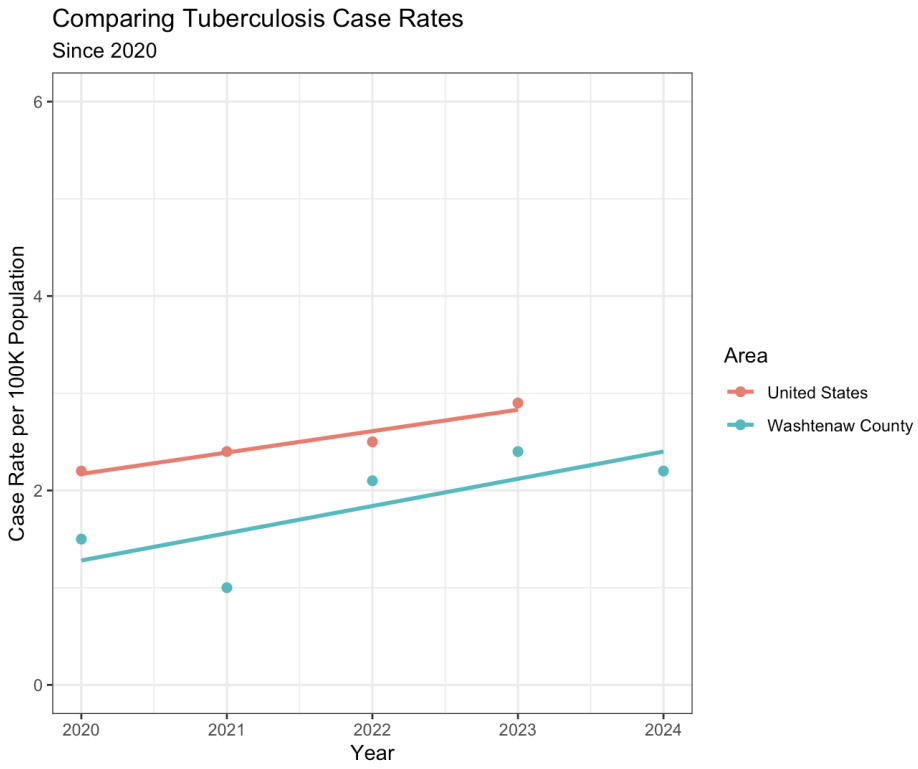
	year_onset_date	active_case_count	pulm_case_count	case_rate_per100k
1	2024	8	5	2.2
2	2023	9	8	2.4
3	2022	8	6	2.1
4	2021	4	3	1.0
5	2020	6	6	1.5
6	2019	6	5	1.6
7	2018	9	5	2.4
8	2017	8	5	2.2
9	2016	10	7	2.8
10	2015	6	5	1.7
11	2014	7	2	2.0
12	2013	4	3	1.2
13	2012	5	4	1.5
14	2011	8	5	2.3
15	2010	9	8	2.6
16	2009	6	5	1.9
17	2008	10	7	3.1
18	2007	6	5	1.9
19	2006	7	5	2.2
20	2005	10	3	3.1

Task 2


Annual Tuberculosis Active Case Rate
Washtenaw County
Case Rate per 100K Population
0
1
2
3
4
2005
2010
2015
2020

Task 3


Comparing Tuberculosis Case Rates
Case Rate per 100K Population
Year
Area
United States
Washtenaw County
0
2
4
6
2005
2010
2015
2020
2025


Comparing Tuberculosis Case Rates
Since 2020
Case Rate per 100K Population
Year
Area
United States
Washtenaw County
0
2
4
6
2020
2021
2022
2023
2024