Worksheet - MPPHC 2025
Hands-on Workshop: Using R for Epidemiologic Analysis
2025-10-27

This workshop will loosely follow the following steps. For coding steps, there may be boxes with functions included in orange, which may be helpful for completing the step. However, there is often more than one way to complete a task in R! There are also BONUS sections included, for an extra challenge. This format is meant to accommodate multiple levels of R skill for the duration of this workshop.

1. Copy resources from https://micom-hub.github.io/r_workshop_page.html. Place data files in a folder on your computer, to work from.

2. Open RStudio, and open a new Rscript file (File > New File > R Script)

3. Load in two R libraries: tidyverse and gt

```
library()
```

4. Create a new variable, and set the value of it to the character string path to the folder that has your data files. Then, use that to load in the faux_c.csv data file, which has our case data included in it. View the dataset, to make sure it looks correct.

```
paste0()
read.csv()
view()
```

Hint: Pay attention to the direction of slashes!

5. BONUS Section 1: Checking Data Quality with Code
    a. Check that all Investigation IDs are unique

```
length()
unique()
nrow()
if(){}
print()
stop()
```

b. Check that Case Status is all confirmed

```
table()
```

c. Check the minimum and maximum Referral Date included in this data set, and check that all values have a year of 2024

```
class()
as.POSIXct()
sample()
seq()
table()
year()
min()
max()
```

6. Visualization #1: ggplot()
    a. https://miottawa.org/wp-content/uploads/Final-2024-STI-Annual-Report.pdf (most of) Figure 3; Write down what categories and data you will need to know in order to make a chart like this:

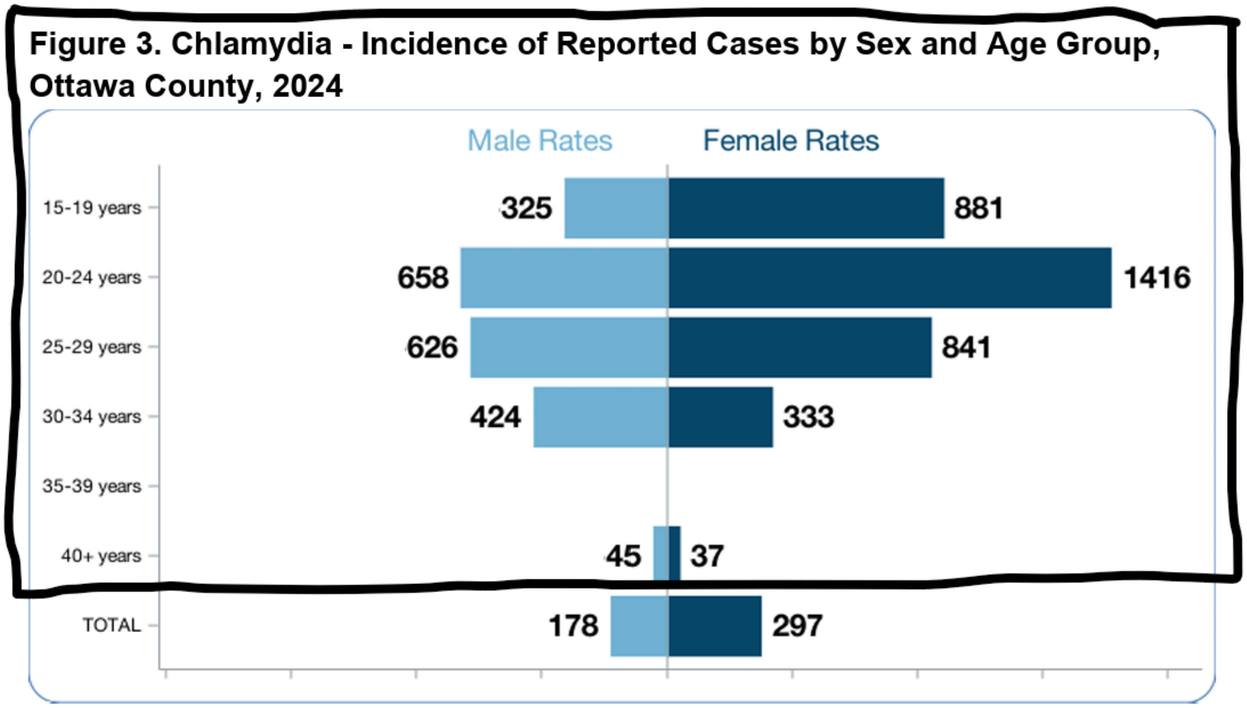


Figure 3. Chlamydia - Incidence of Reported Cases by Sex and Age Group, Ottawa County, 2024

Population Data Source: 2020 Decennial Census[2]

b. Make a new variable in your data frame called “age_group” that matches the age groups to show in the visualization.

```
mutate()
case_when()
```

c. Make a new data frame, grouped by age_group and Sex_at_Birth, with a new column of the count of cases that fall within each group.

```
group_by()
summarize()
length()
unique()
```

d. Read in the population data, which we’ll need to calculate cases per 100K population. Then, since we’ll need to combine the two datasets, make sure the age and sex variables match between the two.

```
read.csv()
paste0()
table()
```

e. Edit the age and sex columns in the population data frame to match the case data frame. Careful - You might have to “re-make” the population data frame a bit!

```
mutate()
case_when()
group_by()
summarize()
```

Hint: Look at the age and sex columns for both dataframes, using `table()`*. Make a new column in the population dataframe called ‘Sex_at_Birth’ that changes the category column (‘Female population’, ‘Male population’, ...) to ‘Female’, ‘Male’, ‘Total’. Make another new column that turns the ‘age_full’ column into a column that has a 40+ age category, then group and rename the population age categories to match the case data.*

f. Merge the population and case data frames by the age and sex variables.

```
merge()
```

g. Add a new variable into the merged data frame you created, for cases per 100k population.

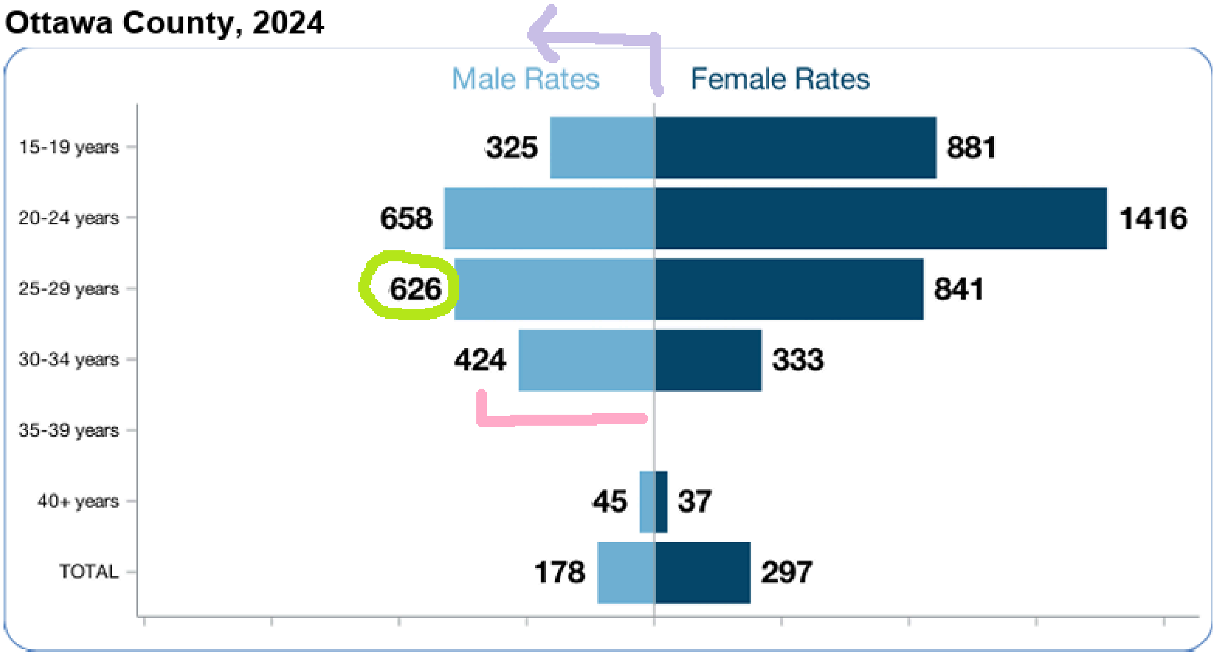
```
mutate()
```

h. One last data-wrangling step! We need to make three additional columns to help with formatting this age-pyramid style bar-chart: (1) cases_per100k with the “Male” values negative, and the “Female” values positive; (2) text_position, with the negative “Male” value of cases_per100k - 60, and the positive “Female” value of cases_per100k + 60; (3) text_label, with all cases_per100k values positive and rounded to the nearest integer.

```
mutate()
```

Hint:

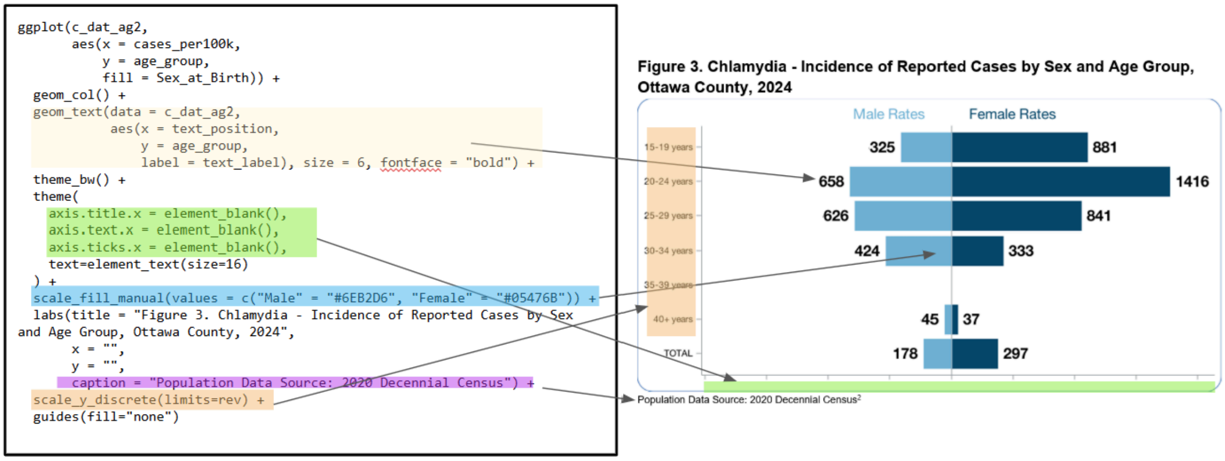
Figure 3. Chlamydia - Incidence of Reported Cases by Sex and Age Group, Ottawa County, 2024



Population Data Source: 2020 Decennial Census[2]

i. Make the ggplot!

```
ggplot()
aes()
geom_col()
geom_text()
theme_bw()
theme()
element_blank()
element_text()
scale_fill_manual()
labs()
scale_y_discrete()
guides()
```



```
ggplot(c_dat_ag2,
       aes(x = cases_per100k,
           y = age_group,
           fill = Sex_at_Birth)) +
  geom_col() +
  geom_text(data = c_dat_ag2,
            aes(x = text_position,
                y = age_group,
                label = text_label), size = 6, fontface = "bold") +
  theme_bw() +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    text=element_text(size=16)
  ) +
  scale_fill_manual(values = c("Male" = "#6EB2D6", "Female" = "#05476B")) +
  labs(title = "Figure 3. Chlamydia - Incidence of Reported Cases by Sex
and Age Group, Ottawa County, 2024",
       x = "",
       y = "",
     caption = "Population Data Source: 2020 Decennial Census") +
  scale_y_discrete(limits=rev) +
  guides(fill="none")
```

Figure 3. Chlamydia - Incidence of Reported Cases by Sex and Age Group, Ottawa County, 2024

Population Data Source: 2020 Decennial Census[2]

j. Save the ggplot image as a 13in x 8in .png file, in the same folder location that your other data files are located in. Add today's date to the file name.

```
ggsave()
paste0()
gsub()
Sys.Date()
```

7. BONUS Section 2: Go back and add on the "Total" column to the bottom of the chart

8. Visualization #2: gt()
   a. https://miottawa.org/wp-content/uploads/Final-2024-STI-Annual-Report.pdf (half of) Table 3; Write down what categories and data you will need to know in order to make a chart like this:

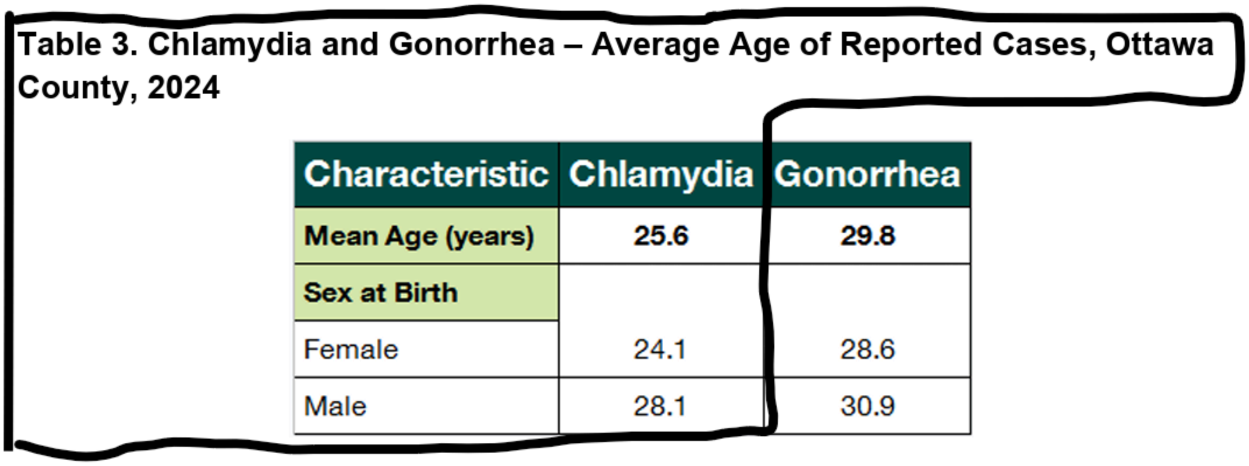


Table 3. Chlamydia and Gonorrhea – Average Age of Reported Cases, Ottawa County, 2024

Characteristic	Chlamydia	Gonorrhea
Mean Age (years)	**25.6**	**29.8**
Sex at Birth		
Female	24.1	28.6
Male	28.1	30.9

   b. Calculate the total average age of all the chlamydia cases in 2024, rounded to 1 decimal point, and assign that value to a variable

```
round()
mean()
```

   c. Make a new small dataframe, with data grouped by sex at birth, that calculates the average age, rounded to 1 decimal point, of each category.

```
round()
mean()
group_by()
summarize()
```

   d. Make a series of dataframes to combine into the final set
      i. One dataframe with a column called "Characteristic" and a value in that column of "Mean age (years)"; and another column called "Chlamydia" and a value in that column of your value from 8b.

```
data.frame()
```

ii. One dataframe with a column called “Characteristic” and a value in that column of “Sex at Birth”; and another column called “Chlamydia” and a blank value in that column.

```
data.frame()
```

iii. Rename the columns of the dataframe you made in 8c to “Characteristic” and “Chlamydia”

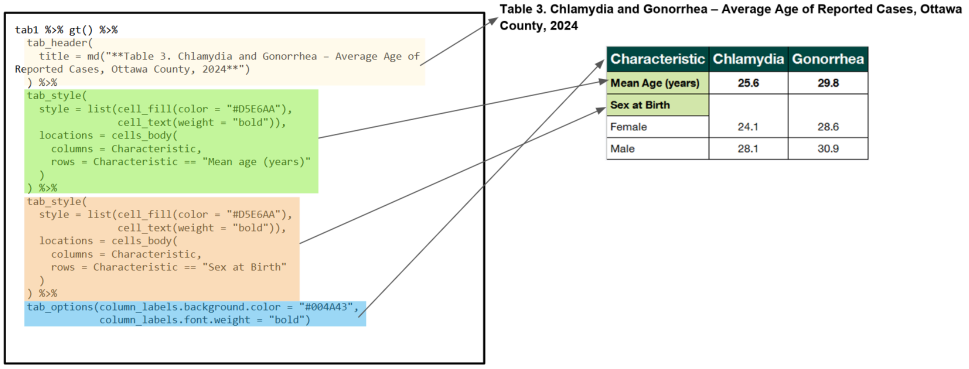
```
colnames()
c()
```

e. Stack / combine the three dataframes made in 8d in the order you made them.

```
rbind()
```

f. Apply gt() to the final dataframe made in 8e. Add a title, coloring, and bolded text to match the formatting of the original table.

```
gt()
tab_header()
md()
tab_style()
cell_fill()
cell_text()
cells_body()
tab_options()
```



```
tab1 %>% gt() %>%
  tab_header(
    title = md("**Table 3. Chlamydia and Gonorrhea – Average Age of
Reported Cases, Ottawa County, 2024**")
  ) %>%
  tab_style(
    style = list(cell_fill(color = "#D5E6AA"),
                 cell_text(weight = "bold")),
    locations = cells_body(
      columns = Characteristic,
      rows = Characteristic == "Mean age (years)"
    )
  ) %>%
  tab_style(
    style = list(cell_fill(color = "#D5E6AA"),
                 cell_text(weight = "bold")),
    locations = cells_body(
      columns = Characteristic,
      rows = Characteristic == "Sex at Birth"
    )
  ) %>%
  tab_options(column_labels.background.color = "#004A43",
              column_labels.font.weight = "bold")
```

Table 3. Chlamydia and Gonorrhea – Average Age of Reported Cases, Ottawa County, 2024

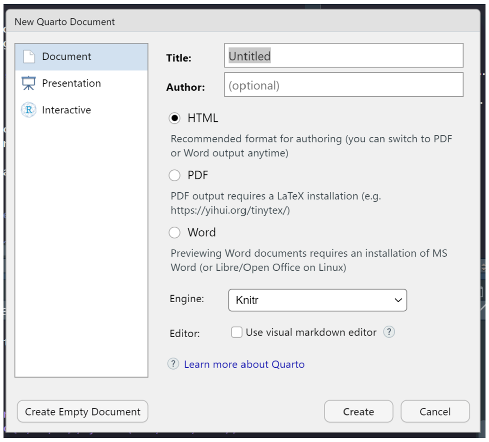
Characteristic	Chlamydia	Gonorrhea
Mean Age (years)	**25.6**	**29.8**
Sex at Birth		
Female	24.1	28.6
Male	28.1	30.9

    g. Save the table as a 200px x 400px png image, in the same folder location that your other data files are located in. Add today’s date to the file name.

```
gtsave()
paste0()
gsub()
Sys.Date()
```

9. While making dated image files that write out to a known location does streamline some code and processes, you might prefer the files to all be included in one report. Let’s move the code we just wrote into a [Quarto report](), to accomplish that!
    a. In RStudio, in the upper left hand corner, click File > New File > Quarto Document…

    b. Add a title and click “Create”; Note, you can make PDF or Word documents from Quarto as well, as you can see from the selection options. For now, we’ll stick with HTML.



Title: Untitled
Author: (optional)

● HTML
Recommended format for authoring (you can switch to PDF or Word output anytime)
○ PDF
PDF output requires a LaTeX installation (e.g. https://yihui.org/tinytex/)
○ Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux)

Engine: Knitr
Editor: ☐ Use visual markdown editor

    c. Things to note:
        i. If you’re making an html document, you need to “embed-resources” so that other people can open the html file if you send it to them
        ii. You can add in notes and regular text to the body of the report, and format it using markdown. Code blocks are noted with ```{r} ```. (You can use other languages too!)
        iii. Using [execution options]() inside each code block controls if the code, warning messages, or other information appears or doesn’t appear.

iv. For ggplots, you can control the size of the figure output with fig.height and fig.width inside the curly brackets marking the code block. (You can also do this in the header portion of the quarto document, if you want all figures to be the same size and you don’t want to repeat that in every code block)

v. For gt() tables, you have to control the size of the table inside the tab_options() choice, with table.width = "40%"