**Data Exploration**

**Activity**

The goal of this activity is to interactively explore data and become acquainted with ggplot code.

In this activity, you will be interactively exploring data (NCHS, which stands for National Center for Health Statistics, from the {DataComputing} package) using the esquisser() function in the {esquisse} package.

Before you begin, you will need to load the following packages and data (Note: You can copy and paste the following into an R script):

###Loading packages

library(devtools)

devtools::install\_github("DataComputing/DataComputing")

library(DataComputing)

library(esquisse)

##Load the dataset

data("NCHS")

When you first load a dataset into R, it’s good practice to examine the data to get a sense of its structure (e.g., the variables (columns), the cases (rows)).

1. Examine the dataset using at least three different methods.
2. How many variables are in the dataset?
3. How many categorical variables are there in the dataset? How many quantitative variables are there? How are the type of these variables labeled in R?
4. How many cases are in the dataset? What do the cases represent?

Now that you have a sense of the data, let’s visualize it to get a sense as to what the data have to say.

esquisser(NCHS, viewer = "browser")

You are going to be provided with an example before you are asked to make your own. Before you get to it, [FlowingData](https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/) provides some nice recommendations for data viz workflow:

* **Ask the data questions.** Simple research questions will guide the types of visualizations that you should construct.

* **Start with the basics and work incrementally.** Before constructing complicated or multivariate or interactive graphics, start with simple visualizations. An understanding of the simple patterns provides a foundation upon which to build more advanced analyses and visualizations. This incremental process works particularly well with the layered grammar of graphics in ggplot.
* **Focus.** Reporting a large number of visualizations can overwhelm the audience and obscure your conclusions. Instead, pick out a focused yet comprehensive set of visualizations. [Here](https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/) is an example of one dataset visualized 25 different ways, each with a different focus and interpretation, and what can happen if you let the data ramble on without a focus.

**EXPLORING THE DATA**

* Drag and drop the *sex* variable into the box labeled *X*.

1. What type of plot was created? What is displayed on the x-axis? What is displayed on the y-axis?
2. What question might someone have asked if they were interested in this plot?
3. Copy and paste the code (in the *Export & code* tool in the lower bar) into the space below. Label as much of the code as you can (e.g., Where is the data set name? Where is the variable name? What do you think fill = does?).

By default, the interactive plotting tool will create a plot for you. You can change this by clicking the square button (with a circle icon in it) in the upper left under the word Data.

Now let’s add another variable to the plot.

* Drag and drop the *smoker* variable into the box labeled *Fill.*

1. What changed between the first plot (with only the *sex* variable) and the second plot (when the smoker *variable* was added)?
2. What question might someone have asked if they were interested in this plot?
3. Copy and paste the code (in the *Export & code* tool in the lower bar) into the space below. Label as much of the code as you can.

Let’s add titles and labels to the plot.

* Provide an appropriate Title, X label, Y label, and Fill label to the plot.

1. Copy and paste the code (in the *Export & code* tool in the lower bar) into the space below. Label as much of the new code as you can.

Let’s omit the missing data.

* In the *Data* tool in the lower bar, click on the blue dot next to NA for the smoker variable to remove those that had NA for their smoking status. [Note: The *Data* tool allows you to filter data if you want to examine a subset of the data.]

1. Copy and paste the code (in the *Export & code* tool in the lower bar) into the space below. Label as much of the new code as you can.

One more thing before you create your own. Let’s change the position of the bars.

* In the *Plot options* tool in the lower bar, change the position to dodge and fill. After doing do, examine the code to see how the code changed. [Note: The *Plot options* tool allows you to modify the plot features, such as colors, theme, and legend position.]

**YOUR TURN!**

Now it is your turn to tell a story with the data!

Tips (or should they be called Rules? :) ):

* Be curious!
* Explore the tool and data!
* Have fun!

1. In your small group, come up with a question you want to ask the data and create a visual of that (as best as you can) with this interactive plot. There are some things you will not be able to do with a tool, so do the best you can. After you have created your data visualization, designate one person from your group to share it with the class and what story it tells about the data.