Tidyverse dplyr and ggplot2: How to Manipulate and Visualize Data Easily in R

Emily Lin, Digital Projects Studio at University of Michigan Libraries

**Project Description:** R is a popular programming language and free software environment for statistical computing and visualization. The language and software is widely used among statisticians and data miners for developing statistical software and data analysis. This tutorial is designed to give the reader a quick start on their journey with R. The intended audience is someone with a basic understanding of data analysis and programming languages. The tutorial is mainly divided into two parts: Data manipulation and visualization. The data manipulation portion explains how to use base R functions and the dplyr package to clean, reformat, subset, and summarize the data in various ways. The visualization portion explains how to use the ggplot2 package to create interesting visualizations of the data that was manipulated.

**Applications:** The tutorial clearly explains the common uses of each function by applying them to a focus dataset. Thus, the code from this tutorial can be adapted for data manipulation and visualization for any data set.

**Problems & Solutions:**

- Missing values were represented with a question mark “?” in the dataset used.

- **Solution:** Used stringr library functions and regular expression to replace missing values with NA.

- Too much data causing scatterplot with fitted line not showing any meaningful relationship.

- **Solution**: Filtered down data by race and sex to visualize smaller dataset, showing more visible trend.