**Data Science for Social Scientists**

PSYC 546, Spring 2023

Homework Assignment 4

**Due Date**: February 22nd (by 8:15 PM)

**Reminder**: See the assigned readings and data visualization lecture slides for a tutorial on how to use R to perform the various functions included in the homework assignment below. **Once completed, you should submit a completed version of this document and your final R script file to the Homework Assignment 4 – Submission Portal on Canvas**.

**R/RStudio**

Your submitted R script file should contain code to answer all of the questions below. Please use comments (e.g., #Question 1) to label the code for each question. For Questions 1-4, use the traditional plotting functions in R to complete the problems (i.e., the functions covered during the Data Visualization Pt. 1 week). ggplot2 will be utilized later in Questions 5-7.

1. Import the **survey.csv** file from Canvas. Make a histogram of the life satisfaction scale scores (Mlifesat). Have the horizontal x axis labelled as “Life Satisfaction Scores”, the chart title as “Histogram of Life Satisfaction Scores”, have the y-axis limit be 0 to 100, and have the color of the bars be turquoise. Paste the figure below. [1 point]

Chart, histogram

Description automatically generated

1. Using the same data set, make a boxplot of the continuous age variable. Have the boxplot broken down by sex, have the x axis labelled as “Biological Sex”, have the y axis labelled as “Age (in Years)”, and have the label names be “Males” and “Females” instead of 0 and 1, respectively. Paste the figure below. [1 point]

Chart, box and whisker chart

Description automatically generated

1. Using the same data set, use the plot() function to create a scatter plot with Positive Affect (Mposaff) on the x axis and Negative Affect (Mnegaff) on the y axis. Make the x axis and y axis labels “Positive Affect” and “Negative Affect”, respectively, have the limit of the y axis be 1 to 5 so it matches the x axis, and use the point shape to make the points little black circles. Paste the figure below. [1 point]

Chart, scatter chart

Description automatically generated

1. Using the same data set, create a bar plot based on the number of non-smokers and smokers in the data set. Non-smokers have a value of 0 and smokers have a value of 1 on the binary smoking status variable. Have the bar plot display horizontally (i.e., the bars should be going from left to right), have the x axis labelled as “Count”, have the label names be “Non-Smokers” and Smokers” instead of 0 and 1, respectively. Paste the figure below. [1 point]

Chart, waterfall chart

Description automatically generated

1. In contrast to the above problems, use ggplot2 and other tidyverse-related packages to complete the rest of the problems. Using the **midwest** data set from ggplot2, create a scatter plot with percent of adults with a college degree in the county (percollege) on the x-axis and percent of residents living under the poverty line (percbelowpoverty) on the y-axis. Paste your final figure below. Make sure the figure includes all the following elements [2 points overall]:
   1. Colors of the points based on midwestern state
   2. The size of the points equaling 1.5
   3. Have a smooth line added to the scatter plot
   4. The label of the x-axis being “Percent with College Degree”
   5. The label of the y-axis being “Percent below Poverty Line”
   6. The legend displayed on top of the figure

Chart, scatter chart, bubble chart

Description automatically generated

1. Using the **gapminder** data set from the dslabs package, create a time series line graph with year on the x axis and life expectancy on the y axis. Paste your final figure below. Make sure the figure includes all the following elements [2 points overall]:
   1. Only the following four countries are included: United States, Japan, Egypt, Romania
   2. The y axis should be labelled “Life Expectancy”
   3. The x axis should be labelled “Year”
   4. The figure should be faceted based on country
   5. The faceted figure should be a 2x2 matrix (i.e., 2 rows and 2 columns)

Chart, line chart

Description automatically generated

1. Using the **diamonds** data set from ggplot2, first create a scatter plot that only includes diamonds with a color of “I”, that has carat on the x axis, price on the y axis, size of .75 for the points, x-axis label of “Carat”, and a y-axis label of “Price (in dollars)”. Assign this entire ggplot figure to an object called P1. Then create a new object called P2, that includes the original P1 figure with one following addition: have the color of the points be based on the cut of the diamond. Finally, use the grid.arrange() function with the **gridExtra** library to combine these two objects within a single figure. Have the number of columns be 2 so that the two figures are side-by-side horizontally. Paste your final figure below. [2 points]

Chart, scatter chart

Description automatically generated