STA130H1S – Winter 2024

Week 2 Problem Set - Sample Answers

N. Moon, J. Speagle, and [ADD YOUR NAME HERE]

Instructions

How do I hand in my solutions and how do I check my work

You will submit your solutions (.Rmd and .pdf) on MarkUs at the following link: https://markus4.teach.cs.toronto.edu/2024-01/courses/1 Submissions are due at 5pm on Thursdays; see Quercus page for the specific deadline for each problem set.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them. Some of these parts will also be graded in some weeks.

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in the name of the variable we're looking for - re-read the question carefully and make sure you're following the instructions.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message).
- Come to TA or instructor office hours with your issue.

[Question 1]

The `artwork_sample.csv` file contains data for a sample of pieces of art owned by the Tate Art Museum. While more variables are available on the Tate Art Museum's site (github.com/tategallery/collection), you will only be working with the variables featured in the `artwork_sample.csv` file (see below). Save this data object in an R object called `artwork`.

- `id`: Unique ID for each piece of artwork
- `artist`: Name of the artist
- `title`: Title of the artwork
- `type`: Medium used
- `year`: Year the artwork was created
- `width`: width of the artwork, in mm
- `height`: height of the artwork, in mm
- `units`: measurement units for width and height of the artwork
- `area`: surface area (in squared cm)

```
library(tidyverse)  # load the tidyverse package so it is available to use
artwork <- read_csv("artwork_sample.csv")  # read in the data
```

(a) Use the `glimpse()` function to view properties of the `artwork` data set. How many observations does it include? How many variables are measured for each observation? Replace `NULL` with your answers in the code chunk below to save your answers in `Q1a_number_of_variables` and `Q1a_number_of_observations`.

```
Q1a_number_of_variables <- NULL
Q1a_number_of_observations <- NULL
```

Write 1-2 sentences describing the sample using this information and the context.

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

(b) Create 3 histograms to explore the distribution of years of creation for this sample of pieces of art: (i) one with 3 bins, (ii) one with 30 bins, and (iii) one with 150 bins; make sure to specify meaningful axis labels where appropriate. Which of these histograms is most appropriate to describe the distribution of the artworks' years of creation? Why? Write a few sentences describing the distribution based on the histogram you chose as most appropriate.

```
# Create your plots below
hist1 <- NULL
hist2 <- NULL
hist3 <- NULL

# the gridExtra package allows for plots to be
# arranged in a grid layout - we'll load it here
# you are NOT REQUIRED to know or use this function
# we're just showing you how we set things up
library(gridExtra)
grid.arrange(hist1, hist2, hist3, nrow=1, ncol=3)
```

Which histogram is most appropriate to visualize these data?

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

Justify your answer above in a few sentences. Hint: Don't forget to refer to the 3 aspects of quantitive distributions and comment on how each plot lets you visualize each aspect.

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

(c) Construct a plot to visualize the distribution of the `type` variable. Make sure to specify meaningful axis labels where appropriate. Hint: If you choose a categorical variable with many different categories, you may find it useful to use `coord_flip()` to flip the bars horizontally and/or change the options in the R code chunk to make the plot large (ex: `{r, fig.height=15, fig.width=5}`). From the choices below, select the best description for the distribution of the `type` variable.

```
# Create your plot below


# Among the four descriptions below, decide which one is most accurate
# and replace NULL with "A", "B", "C" or "D" accordingly
Q1c <- NULL
```

A: There are about twice as many line engravings than photographs, and almost twice as many photographs than watercolours. The most common type of artwork in this sample of 1000 is watercolour.

B: The distribution of artwork type is not symmetrical. The center of the distribution is photograph because it appears in the middle.

C: There are about twice as many photographs than line engravings, and almost twice as many watercolours than photographs. The most common type of artwork in this sample of 1000 is watercolour.

D: The distribution of artwork type is left skewed because there are fewer line engravings than photographs and watercolours. The most common type of artwork in this sample is waterclour.

(d) Construct a set of three boxplots showing visual summaries of the distribution of surface area (`area`) for each type of artwork (`type`); make sure to specify meaningful axis labels where appropriate.

```
# Create your plots below
```

Write 3-4 sentences comparing these distributions.

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

(e) Make a scatterplot showing the `width` versus `height` for each type of artwork (`type`) using the `colour` and/or `facet_wrap()` option; make sure to specify meaningful axis labels where appropriate.

```
# Create your plot below
```

Write 1-3 sentences describing any trends you see.

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

[Question 2] The `ncbirths` data set is part of the `openintro` package. It consists of observations for a sample of 1000 births in North Carolina in 2004. Type `?ncbirths` in the R console for more information about the data and to see the definition of each variable. The code below loads the required libraries for this question and provides a glimpse of the `ncbirths` data frame.

```
glimpse(births14)
```

```
## Rows: 1,000
## Columns: 13
## $ fage          <int> 34, 36, 37, NA, 32, 32, 37, 29, 30, 29, 30, 34, 28, 28,~
## $ mage          <dbl> 34, 31, 36, 16, 31, 26, 36, 24, 32, 26, 34, 27, 22, 31,~
## $ mature        <chr> "younger mom", "younger mom", "mature mom", "younger mo~
## $ weeks         <dbl> 37, 41, 37, 38, 36, 39, 36, 40, 39, 39, 42, 40, 40, 39,~
## $ premie        <chr> "full term", "full term", "full term", "full term", "pr~
## $ visits        <dbl> 14, 12, 10, NA, 12, 14, 10, 13, 15, 11, 14, 16, 20, 15,~
## $ gained        <dbl> 28, 41, 28, 29, 48, 45, 20, 65, 25, 22, 40, 30, 31, NA,~
## $ weight        <dbl> 6.96, 8.86, 7.51, 6.19, 6.75, 6.69, 6.13, 6.74, 8.94, 9~
```

```
## $ lowbirthweight <chr> "not low", "not low", "not low", "not low", "not low", ~
## $ sex            <chr> "male", "female", "female", "male", "female", "female",~
## $ habit          <chr> "nonsmoker", "nonsmoker", "nonsmoker", "nonsmoker", "no~
## $ marital        <chr> "married", "married", "married", "not married", "marrie~
## $ whitemom       <chr> "white", "white", "not white", "white", "white", "white~
```

(a) Type `?births14` in the R console to answer the questions below. Make sure not to change the variable names, and replace each `NULL` with your answer.

```
# In what year were these data collected?
Q2a_year <- NULL

# What is the name of the variable with the following definition:
# "Weight gained by mother during pregnancy in pounds."
# Make sure that your answer is in quotation marks here, and
# remember that R is case sensitive!
Q2a_variable <- NULL
```

(b) Before even doing any analysis, it is good to consider whether the data in question might have any strong bias that could impact any conclusions you may draw. Based on the information contained documentation (particularly in `Description` and `Source`), please write 3-4 sentences either (1) arguing the dataset should be generally unbiased and representative of all births in North Carolina or, if not, (2) what potential issues there might be with the data.

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

(c) Choose two categorical variables and plot their distributions. Identify whether each of these variables is a nominal or ordinal categorical variable. Write one or two sentences interpreting each plot.

```
# Create your plots below
```

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

(d) Consider the variable `weight`. Replace NULL below with the type of this variable (either "continuous numerical", "discrete numerical", "nominal categorical", "nominal ordinal", or "binary"). Create a plot to visualize the distribution and write 2-3 sentences describing the distribution.

```
# What is the type of this variable?
# Hint: Replace NULL below with the correct option above, making sure to copy exactly and to include qu
Q2d <- NULL

# Create your plot below
```

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

(e) Construct a plot that shows the relationship between birth weight (`weight`) and mother's smoking status (`habit`); make sure to specify meaningful axis labels where appropriate.

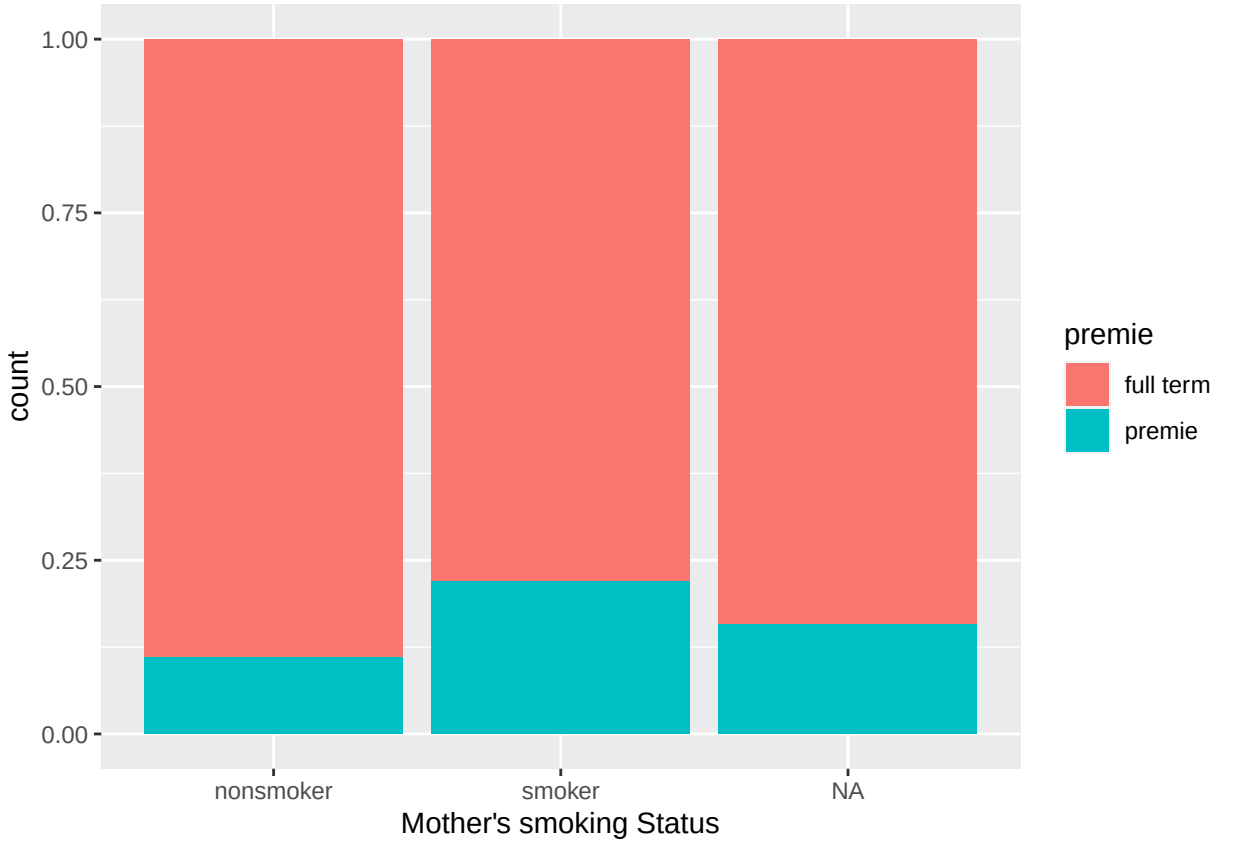
```
# Create your plot below
```

(f) Based on your plot, would you conclude that there is a strong association, weak association, or no association between birth weight (`weight`) and mother's smoking status (`habit`)? Why or why not?

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

(g) Sometimes, a potential association could be due factors that affect both variables. These are called *confounding variables* (we will discuss these in more detail later in the course). Below, we have included a plot showing the *proportion* of births that were classified as premature or full-term (`premie`) versus smoking status (`habit`).

```
# Proportion of premature births vs mother's smoking status
births14 %>% ggplot(aes(x = habit, fill=premie)) +
  geom_bar(position="fill") +
  labs(x="Mother's smoking Status")
```



Does the above plot suggest that a mother's smoking habit could lead to premature births? Why or why not?

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

(h) In general, commomn medical advice for women who are pregnant or expect to become pregnant is to stop smoking to protect the health of their baby. Based on your plot and answer above, pleased write 2-4 sentences discussing whether this advice seems justified or not.

<Type your answer to the written question here (note that R code does not run outside of the grey R code chunks).>

We hope this exercise helped to emphasize the importance of the ethical considerations we discussed in class when it comes to making, interpreting, and communicating data visualizations. Don't forget to review sample solutions once they are posted on Quercus.