Graphing Bivariate Relationships

Assignment Overview

To fully explore the relationship between two variables both summary statistics and visualizations are important. For this assignment you will describe the relationship between these four specific combinations of data types:

- Categorical explanatory and categorical explanatory variable. (C ~ C)
- Quantitative explanatory and categorical explanatory variable. (Q ~ C)
- Any combination of the above with a binary variable (B ~ C, C ~ B, or Q ~ B)
- Quantitative response and quantitative explanatory variable. (Q ~ Q)

Setup

I. Determine what variables you want to graph based on your research topic.
   - You will need a mixture of categorical and quantitative variables for this assignment.
   - You should use variables that are relevant to your research topic.
   - If you have not yet identified both a quantitative, a binary, and a categorical variable that you are interested in, now is the time to go back to the codebook and figure this out.

II. Recode variables as needed.
   - If your response variable is categorical with many levels, you may want to collapse it down to fewer than 5 levels.
   - It is perfectly acceptable to recode variables temporarily for exploratory purposes and not put it in your data management file.

Instructions

0. Use the template provided: [RMD] for R users, and [Word] for SPSS users.

For each bivariate relationship under consideration you will do the following:

1. Name and explain the two variables under consideration.
2. Create the appropriate graphic for bivariate relationship under consideration. For these plots binary variables are treated as categorical variables with only 2 levels.
   - C ~ C: Side by side barplot
   - Q ~ C: Paneled histogram with density overlaid, or a grouped boxplot with overlaid violin plot.
   - Q ~ Q: Scatterplot. Add both lowess and linear trend lines.
3. Calculate appropriate grouped summary statistics
   - For continuous outcomes you'll want to describe measures including the sample size, mean, median, range and variance for each level of the categorical variable.
   - For categorical outcomes you'll want to calculate %'s of your outcome measurement across levels of your covariate.
     - i.e. proportion of males who are smokers compared to proportion of females who are smokers
     - or proportion of smokers who are male, compared to proportion of non-smokers who are male.
4. Explain the relationship or trends you see in the data in a summary paragraph. Put this paragraph below the graphic.

- Use summary statistics in your text explanation.
- Use specific features of the graphic in your text explanation.
  - i.e. are there outliers only in one group?
  - Do the data seem clumped or clustered in one region of the scatterplot?
  - Is there a linear or non-linear pattern?
  - Does one combination of categorical levels (C~C) seem to hold most the data?
  - Are there any outlying data points? Don't list off each one, just state if there is and where approximately it's at.

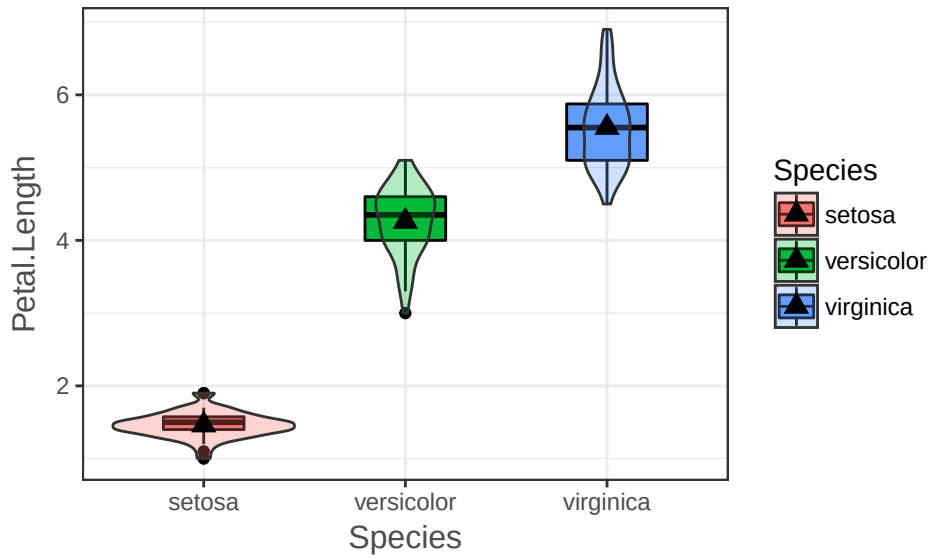
Example

This example explores the association between the length of an iris petal and the species of iris. The quantitative response variable is petal length (`Petal.Length`) and the categorical explanatory variable is species (`Species`).

```
library(tidyverse); library(knitr)
iris %>% group_by(Species) %>%
         summarise(mean=mean(Petal.Length),
                   sd=sd(Petal.Length),
                   n=n()) %>%
         kable(digits=2)
```

Species	mean	sd	n
setosa	1.46	0.17	50
versicolor	4.26	0.47	50
virginica	5.55	0.55	50

```
ggplot(iris, aes(x=Species, y=Petal.Length, fill=Species)) +
  geom_boxplot(width=.4) + geom_violin(alpha=.3) +
  stat_summary(fun.y="mean", geom="point", size=3, pch=17,
    position=position_dodge(width=0.75)) + theme_bw()
```



There are 50 iris plants within each species. There is clear difference in the average Petal length across the pecies. *Setosa* has the smallest average petal length of 1.46 cm and the smallest variation with a standard

deviation of 0.17cm. *Veriscolor* has an average petal length of 4.26cm with SD of 0.47cm, and *Virginica* has the largest average petal length of 5.55cm and the largest variation with a standard deviation of 0.55cm.

Submission

- Upload the final PDF to `hw03 Bivariate Graphics/Incoming` folder in Google Drive with the file name: *userid_biv_graph.pdf*
- This assignment will be peer reviewed.