Graphing Bivariate Relationships

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)

Before you start,

1. 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.
2. Recode variables as needed.
    - If your response variable is categorical with many levels, you must collapse the levels 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]. Replace the `template` in the file name with your username.

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.

5. Submit your knitted PDF to the **05 Bivariate Graphics** folder in Google Drive

Example

In these examples I use `ggplot2` for plotting, `dplyr` for data management and to create summary statistics, `knitr` to create nice tables, `gridExtra` to put plots side by side, and `scales` to format the y axis with percentages. (You do not need to report this information in your assignment. This is for your knowledge only.) I also explain what I'm doing in code here for you to learn. you do not need to explain your own code.

```
library(ggplot2); library(dplyr); library(knitr); library(scales); library(gridExtra)
knitr::opts_chunk$set(echo = TRUE, warning=FALSE, message=FALSE)
```

C ~ C Association

For this example I am using the `NC Births` data set, data on 1000 births in 2004 from North Carolina. This example explores the association between the smoking status of the mother (`habit`) and whether or not the baby was born prematurely (`premie`).

```
nc <- read.csv("https://norcalbiostat.netlify.com/data/NCbirths.csv", header=TRUE)
```

Frequency tables to get counts, and *row* percents because I specifically want to compare of premie babies within the smoking and non-smoking groups.

```
kable(addmargins(table(nc$habit, nc$premie)))
```

	full term	premie	Sum
nonsmoker	739	133	872
smoker	107	19	126
Sum	846	152	998

```
kable(prop.table(table(nc$habit, nc$premie), margin=1)*100, digits=1)
```

	full term	premie
nonsmoker	84.7	15.3
smoker	84.9	15.1

I then create a new data set `nc.nomiss` that contains only the variables I want to plot, and then call `na.omit()` on it to remove any records with missing data.

```
nc.nomiss <- nc %>% select(habit, premie) %>% na.omit() # remove missing data before plotting
```

And I use `dplyr` to create the grouped percentages seen above as output from `prop.table`, so that I can create a barchart with proportions on it.

```
calc.props <- nc.nomiss %>% group_by(habit, premie) %>%
              summarise(count=n()) %>%
              mutate(pct=round(count/sum(count),3))
calc.props
```

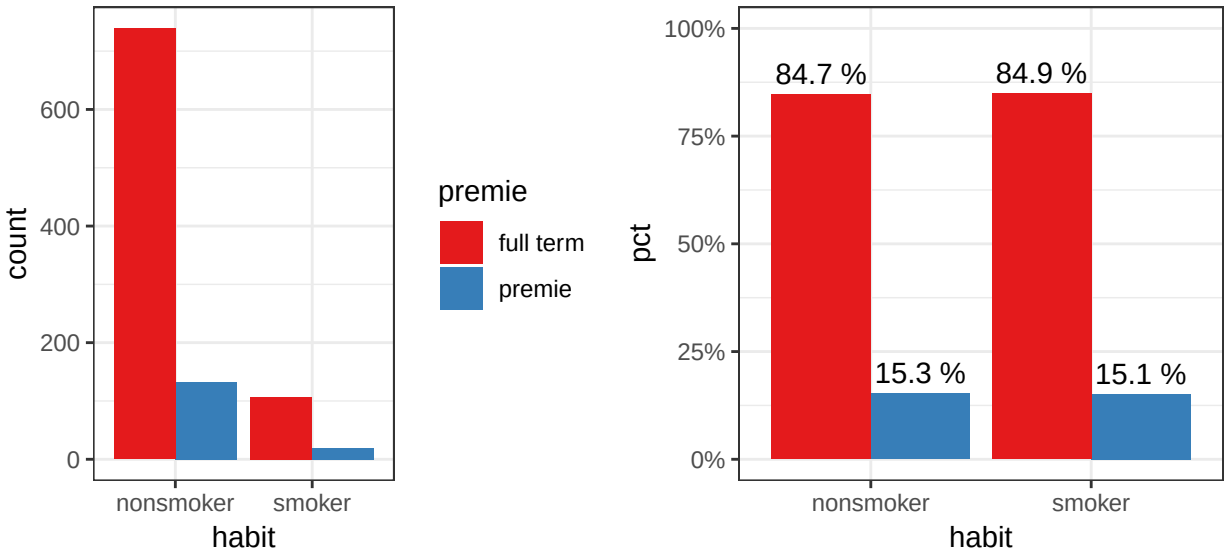
```
## # A tibble: 4 x 4
## # Groups:   habit [2]
##   habit     premie    count   pct
##   <fct>     <fct>     <int> <dbl>
## 1 nonsmoker full term   739 0.847
## 2 nonsmoker premie      133 0.153
## 3 smoker    full term   107 0.849
## 4 smoker    premie       19 0.151
```

Now I will create two bar charts, one for counts and one for percentages. I can put them side by side using the `grid.arrange()` function contained with the `gridExtra` package.

Notice for the barplot of percents I'm using `geom_col()`, and specifying `y=pct`, which is the variable from the data set above that contains the desired percentages.

```
barplot.counts <- ggplot(nc.nomiss, aes(x=habit, fill=premie)) +
                  geom_bar(position="dodge") + theme_bw() +
                  scale_fill_brewer(palette="Set1")
barplot.pcts   <- ggplot(calc.props, aes(x=habit, fill=premie, y=pct)) +
                  geom_col(position="dodge") + theme_bw() +
                  scale_fill_brewer(palette="Set1", guide=FALSE) +
                  geom_text(aes(label = paste(pct*100, "%")),
                                position = position_dodge(0.9), vjust=-.5) +
                  scale_y_continuous(limits=c(0,1), labels=percent)

grid.arrange(barplot.counts, barplot.pcts, ncol=2)
```



Contrary to what I was expecting, there is equal proportion of prematurely born babies to non-smokers compared to babies born to smokers (about 15% for each group). There is no association between the smoking status of the mother and the likelihood of the baby being born prematurely.

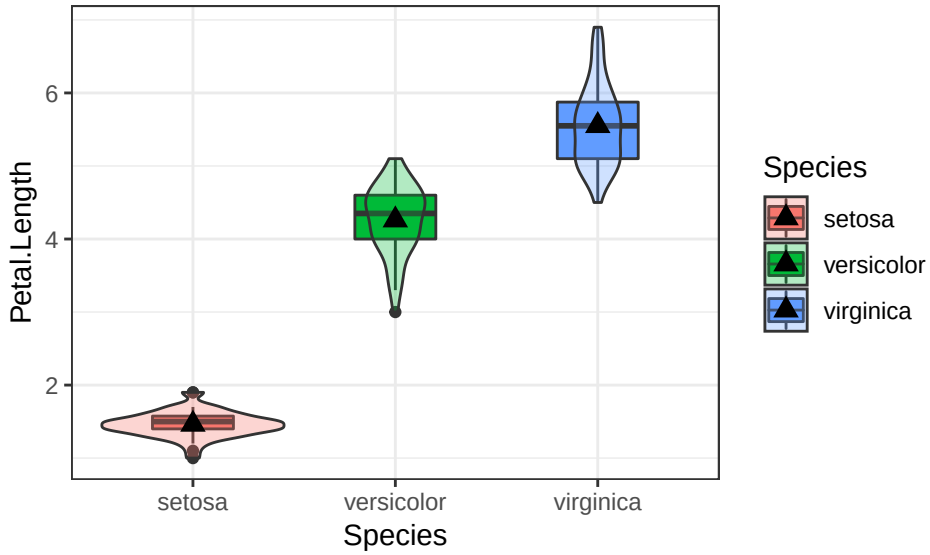
Q ~ C Association

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`).

```
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 species. *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.

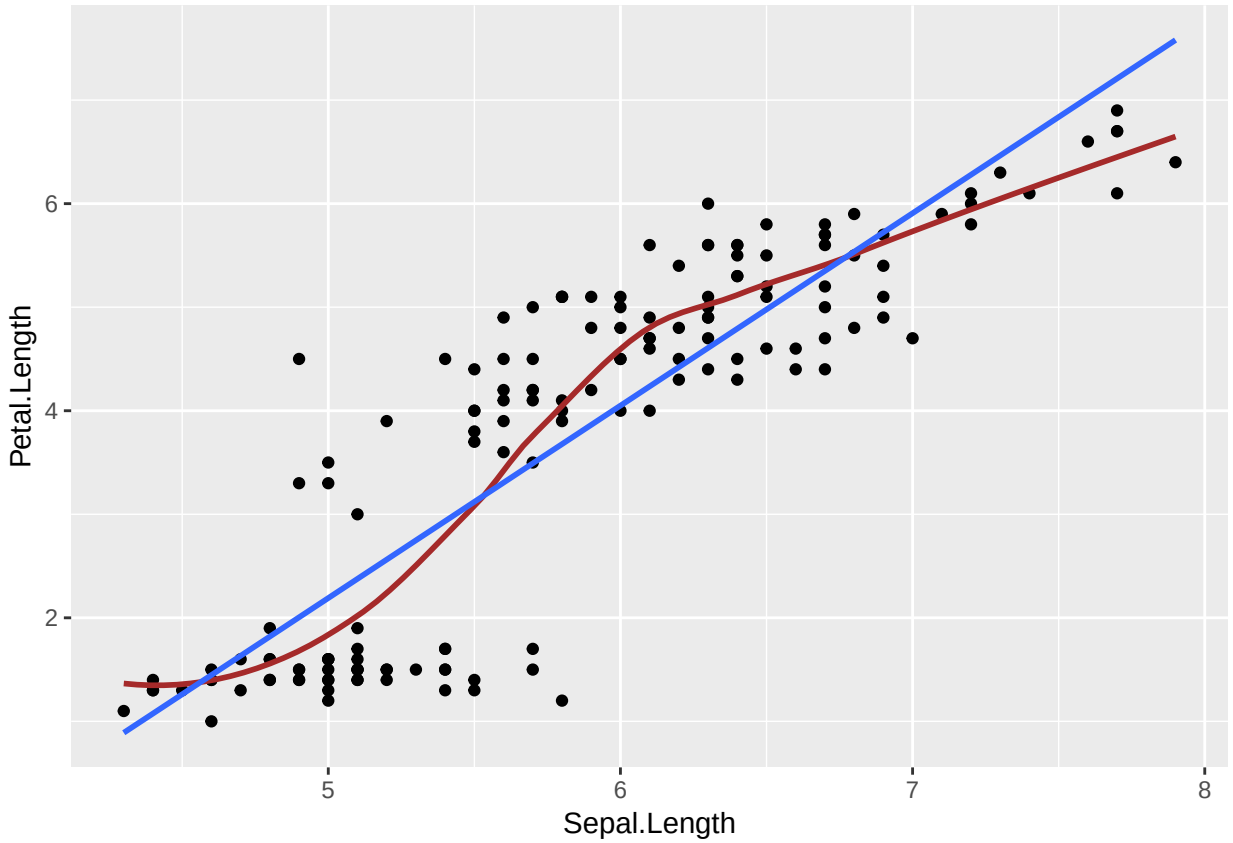
Q ~ Q Association

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

```
cor(iris$Petal.Length, iris$Sepal.Length) # calculate the correlation
```

```
## [1] 0.8717538
```

```
ggplot(iris, aes(x=Sepal.Length, y=Petal.Length)) + geom_point() +
  geom_smooth(se=FALSE, col="brown") + geom_smooth(se=FALSE, method="lm")
```



There is a positive association between sepal and petal length of an iris. The correlation coefficient is 0.87, but the form of the data may not be linear. There is a cluster of values below 2cm petal length which are separated away from the rest of the data. It is hard to assess the linearity of this relationship due to the separate clusters.