Lab 1

*Instructions*

You can use the following data set for your lab. It is already in Rstudio in the mosaicData package (look under packages in lower right window).

HELPrct

Click on the data set in Rstudio and learn about the data set. Be sure to watch Lab 1 video for details.

For the dataset that you choose, you need to determine the following.

1. Install the following packages. Refer to the video for instructions.
2. Mosaic (this package will install several others, it will take a few minutes)
3. Openbwbar a new Rmarkdown file. Name it Lab1\_yourlastname
4. Click on the MosaicData package.
5. Click on the HELPrct data set description.
6. Summarize the description of the data set from RStudio. Make sure that you use complete sentences.
7. Pick **three** variables of interest to you. Two should be **quantitative** and one should be **categorical**. State in your text what variables that you chose and the types.
8. Construct an appropriate graphical summary for all three variables of interest listed above. (Boxplot or histogram for quantitative and a bargraph for categorical).
9. You should have three plots.
10. Remark on the shape, skewness, and spread for the graphs of the quantitative variables. Are there any outliers?

You can use the following commands:

histogram(~yourquantitativevariable, data = HELPrct)

bwplot(~yourquantitativevariable, data = HELPrct)

bargraph(~yourcategoricalvariable, data=HELPrct)

1. Construct comparative boxplots by using one quantitative variable (choose only 1) and one categorical variable.
2. Describe what you see in the context of the variables and using complete sentences. Compare with respect to center, spread, and outliers.

Use the command:

bwplot(~yourquantitativevariable THe yourcategoricalvariable, data = HELPrct)

1. Use your two quantitative variables of interest and construct a scatterplot.

Use the command:

xyplot(yourquantitativevariable1~yourquantitativevariable2, data =HELPrct)

Does there appear to be a relationship between your two variables? Explain using complete sentences.

1. What questions have you formed regarding your chosen dataset that you may want to determine an answer, but are unable to now? All comments should be within your RMarkdown file.

Make sure that you include your Rcode inside of R chunks within a Markdown file as well as the graphs and explanations and knit to a Word document.

---

title: "Untitled"

author: "Matt Massey"

date: "5/20/2021"

output:

html\_document:

df\_print: paged

word\_document: default

pdf\_document: default

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

Question 5.

The HELPrct dataset is based on the HELP study, which was a clinical trial of recruited adult inpatients admitted for detoxification and who had no primary care physician. These patients were randomly selected to receive a multidisciplinary assessment and a brief motivational intervention or usual care. The goal of this study was to link treatment to primary medical care.

Question 6.

Three variables were chosen for this lab: substance, which is a categorical variable of primay abused substance by the patient; avg\_drinks, which is a quantitative variable of the average number of alcoholic drinks per day by the patient; daysanysub, which is a quantitative variable of the number of days to the first post-detox. substance use.

Question 7.

```{r}

histogram(~daysanysub, data=HELPrct)

bwplot(~avg\_drinks, data=HELPrct)

bargraph(~substance, data=HELPrct)

```

Question 7b.

The histogram of daysanysub is bimodal with a pimary mode at 0 days and a subordinate mode at ~180 days. Overall the data is skewed to the right (positive skew) and shows about ~280 days spread.

The modified boxplot of avg\_drinks shows a spread from 0 drinks to almost 150 drinks, a relatively narrow IQR ~25, a median of ~10, and a number of outliers beyond ~60 drinks. Overall the data is skewed to the right (positive skew).

Question 8.

```{r}

bwplot(~daysanysub | substance, data=HELPrct)

```

Question 8a.

The three comparative modified boxplots above show daysanysub by substance. The median and IQR for both alcohol and cocaine are very similar at ~50 days and ~150 days, respectively, while the total spread is slightly lower for cocaine (~225 days) vs. alcohol (~255 days). Neither alocohol nor cocaine show any outliers. In contrast, heroin shows a much lower median (~5 days), IQR (~50 days), and total spread (~140 days); heroin also shows a number of outliers beyond 150 days. Alcohol, cocain, and heroin are all skewed to the right to varying degrees, with cocaine the least (close to symmetrical), followed by alcohol, and then heroin.

Question 9.

```{r}

xyplot(avg\_drinks~daysanysub, data=HELPrct)

```

There doesn't appear to be any obvious relationship between avg\_drinks and daysanysub. People with avg\_drinks above ~60 seem have mostly lower daysanysub, however, this is such a small number of samples I don't think this is conclusive. There are also two weak clusters of data points at/near 0 daysanysub and 180 daysanysub, which is also shown in the histogram above.

Question 10.

Several questions I am interested in from the data above:

What is the signficance of the bimodal distribution of daysanysub at 0 and 180? Is this real or from the sampling method?

What is different about the heroin outliers (boxplot)? Are these truly outliers? If so, what is different about these patients that allowed them to not use any substances for so long compared to the other heroin patients?