title: 'Lab #2: Exploratory Data Analysis '
date: "AP Stats - Fall 2020"

output: html_document

NOTICE: PRINT STATEMENTS MAY BE MISSING,REFER TO COMPILED Jupyter Notebook for reference

Name: Parth Iyer

Skills

- Identify the population, sample, and variables of interest in a study
- Identify variable types
- Use R to generate and interpret descriptive statistics and graphics for numeric variables
- Editing graphics in R using arguments
- Learning and applying Markdown language for text editing
- Learning and applying R Markdown documents for R code and output generation

  Please make sure to show all R code and output after each question so that we can see your work. Write a sentence for each numerical value produced describing its meaning **in context with the proper units**. Be sure to submit your full project as a .zip file that includes your edited, renamed .Rmd file and the .Rproj file to receive credit. If you don't recall how to do something, you should first refer to you notes, the primer from Lab #1 and #2, the internet, then of course just ask Abhi or Chris.

  For this lab, you should begin by reading the accompanying lab 2 Primer, completing online tutorials.Remembver to keep previous labs always handy so you can reference how you did something in case you forget

Introduction

In this lab, we're going to be learning both Exploratory Data Analysis (Descriptiive Statistics) for single quantitative (numeric) variables, and also R Markdown. R Markdown allows us to create documents with embedded code and outputs. You should read the accompanying lab Primer for a guide to R Markdown.

Part 1: Data Collection: Basal Metabolic Rate

As scientists, we need to not only be able to identify the research question, population, sample, and variables of interest for our own questions, but for other studies as well. The following is an abstract of a recent publication. Read it carefully, and then answer the questions about it.

Abstract
Basal metabolic rate (BMR) is posited to be a fundamental control on the structure and dynamics of ecological networks, influencing organism resource use and rates of senescence. Differences in the maintenance energy requirements of individual species therefore potentially predict extinction likelihood. If validated, this would comprise an important link between organismic

ecology and macro-evolutionary dynamics. To test this hypothesis, the BMRs of organisms within fossil species were determined using body size and temperature data, and considered in the light of species' survival and extinction through time. Our analysis focused on the high-resolution record of Pliocene to recent mollusks (bivalves and gastropods) from the Western Atlantic. Species-specific BMRs were calculated by measuring the size range of specimens from museum collections, determining ocean temperature using the HadCM3 global climate model, and deriving values based on relevant equations. Intriguingly, a statistically significant difference in metabolic rate exists between those bivalve and gastropod taxa that went extinct and those that survived throughout the course of the Neogene. This indicates that there is a scaling up from organismic properties to species survival for these communities. Metabolic rate could therefore represent an important metric for predicting future extinction patterns, with changes in global climate potentially affecting the lifespan of individuals, ultimately leading to the extinction of the species they are contained within. We also find that, at the assemblage level, there are no significant differences in metabolic rates for different time intervals throughout the entire study period. This may suggest that Neogene mollusk communities have remained energetically stable, despite many extinctions.

Indicate the correct answer by bolding your chosen response(s) around the answer (by surrounding your answer with ** or __), but not including the corresponding letter. For example: a. **this answer**.

1a: What is the research question in this study?

a. What is the metabolic energy rate of molluscs?

b. Can differences in metabolic requirements predict extinction likelihood?

c. Does basal metabolic rate influences resource use?

1b: What is the population about which the researchers want to make inferences?

a. All organisms in the fossil record
b. All bivalves and gastropods
c. Pliocene to recent bivalves and gastropods

d. Pliocene to recent molluscs from the Western Atlantic

1c: What is the sample that the researchers collected to make inferences about the population?

a. Bivalves and gastropods
b. Pliocene to recent bivalves and gastropods

c. Pliocene to recent molluscs from museum collections

d. Organisms in the fossil record

1d: What is/are the variable(s) that the researchers *directly* measured for each individual? (choose all that apply)

a. Basal metabolic rate (BMR)
b. Ocean Temperature

c. Specimen Size

1e: What type of study was this?

a. Observational

b. Experimental

Part 2: Exploratory Data Analysis: Government Shutdowns

In United States, a government shutdown occurs when Congress fails to pass a bill or a continuing resolution to fund federal government operations and agencies or when the President refuses to sign such a bill or resolution. The duration of a shutdown is random depending, and it prolongs when a conflict between Congress and the President cannot find a compromising solution. `shutdown_all.csv` includes records of all US federal government shutdowns. It includes the year of the shutdown, the sitting President, the number of days the shutdown lasted, and the month the shutdown started.

2a: Modify the code below to read in the data from `shutdown_all.csv` and store it as the object `shutdown` by replacing the blanks/underscores ***() (some hints provided as: ___hint___ , or blanks simply as _***).

```
shutdown = pd.read_csv("shutdown_all.csv")
```

2b: We might want to know more about the data that is inside `shutdown`. We can use the `names()` function to display the variable names to use in your code, while `str()` will display the type of variable R thinks the variables are (numeric: "int", or categorical: "Factor") plus the first few observations in the vector. You should always check this or open your dataframe to inspect how the data is recoreded for a variable in case it's been recoreded differently (or with errors like letters in a numerical variable) than you think.

Modify the code by supplying the dataframe object name to both codes in place of the blanks to inspect your variables. List each variable in the dataset below and indicate its full variable type (e.g. categorical nominal), based on what you see in the code output. You should provide a justification for the variable type.

```
shutdown.columns
vartype(shutdown.values)
```

> Variable: Type -- Justification
> (remember to put two spaces after each line to force a line break. You will also need to start each new line with a `>` to denote a continued answer.)

```
The Year and the length are both quantitative, but President & Month Start are both quantitative.
```

2c: Modify the code below to calculate the mean and standard deviation of the number of days federal shutdowns last by replacing any blanks with the appropriate code. Remember you need to load the `mosaic` package first before we calculate these statistics (you should get in the habit of just doing this once each time you start a lab at the top as well as ggformula). You can install this package (which you have to do before requiring it) either through the console using install.packages or by using the packages tab on the right side of rstudio. *Hint: we always put the numeric variable we're computing stats on after the tilda (~) and the name of the dataframe you're using that has the variables and data in it after data =*

```
mean(shutdown.Length)
stdev(shutdown.Length)
```

2d: Modify the code below to calculate a measure of center (Mean or Median) and of spread (IQR or SD) that are **resistant** for the length of government shutdowns (we'll soon see why we're using **resistant** variables (hint: look @ the shape of the distribution and where the mean is compared to the median)).

```
median(shutdown.Length)
IQR(shutdown.Length)
```

2e: How many federal shutdowns have there been? Add code below to determine the `length()` of your variable (the number of cases in the dataset). `length()` requires the **dataframe$variable** input, not the formula template.

```
len(shutdowns)
```

> Insert your answer below.
> 21

2f: Compare your two measures of center. What can you determine about the shape of the data from this comparison? Make sure to include your justification.

> Insert your answer below.
> The mean is much higher than the median and the Standard deviation is higher than the IQR indicating the presence of high outliers.

2g: Let's create a histogram of the length of government shutdowns by modifying the code below to see if it agrees with what you said above.

Notice that this code is spaced out differently than we've seen before. It is just to make commenting easier -- R will keep reading the function until it reaches the end `)`. Comment on each line of code to indicate what it does and modify the code to add labels to the graph axes.

You must install ggformula before requiring it. You can install ggformula either through the console using install.packages or by using the packages tab on the right side of rstudio. **Don't forget to label your axes (using specific wording), Title, appropriate axis scaling, and key if appropriate!**

Note: gf_histogram is the function we'll be using to make histograms. All the values inside the function in between the commas are called arguments. Look up the function gf_histogram in the help menu to see all the possible arguments (scroll down) you can use for a given function, there's usually a lot more availabe to you than are required in these labs.

```
shutdown.hist(column="Length")
```

2h: Reproduce your code below to create a histogram again, but this time include an additional argument, `binwidth = 3`. What changed in your graphic?

Hint: If you don't tell R what size to make your bins it will automatically choose a size that will work sometimes and not others, so you should always decide for yourself what's best and tell it.

```
shutdown.hist(column="Length",bins=7)
```

> Insert answer below.
> The number of columns was reduced

2i: We want our graphic to be easy to read, including the bin ranges. Reproduce your code from 2h below to create another histogram, but this time include an additional argument: `boundary = 0`. What changed in your graphic?

```

```

> Insert answer below.

```
Not Necessary, Pandas always uses 0 as a base unless a value undercuts it.
```

2j: Describe your histogram's distribution in terms of the four characteristics we discussed in class. *Hint: shape, center, spread, outliers*

Note: you can use the stats you'll compute in the next problem to help you calculate your fences to determine outliers

> Insert your answer below.

```
The Graph is right skewed with a median of 1 and an IQR of 1.5
```

2k: Look at the frequency distribution in the histogram. What does this tell us about the length of government shutdowns in general? How does this compare to our measures of center you discussed in 2f?

> Insert your answer below.

```
Most government shutdowns are short, however there are a few long ones.
```

2l: Create a boxplot for the length of government shutdowns. Modify the code below.

```
shutdown.boxplot(column="Length")
```

2m: Modify the code below to use the function that will produce the 5 number summary 'favstats' on which a boxplot is based. Then use the function 'quantile' to see what it gives you too. Lastly use the function 'IQR'

Note: you should be able to use these stats to compute your upper and lower fences to determine if there are any outliers now! Check out your boxplot and see how it represents outliers graphically

```
favstats(shutdown.Length)
qunatile(shutdown.Length)
```

2n: Describe the distribution of length of government shutdowns based on the boxplot, being sure to address all four characteristics we discussed in class.

> Insert your answer below.
> The distribution is Right Skewed with one upper outlier.

2o: When we calculated mean, standard deviation, etc., what symbols should we be using for those calculations and why? Use the math notation ($ signs) to write each symbol.

Hint: think about our dataframe, does it represent a popultion of all government shutdowns or a sample of government shutdowns in history?

> Insert your answer below.
> N/A in python

Part 3: Curry

Stephen Curry is one of the most prolific scorers currently in the NBA. We can look at the number of points he scored during games in 2015.

3a: Read in the data `curry2015.csv` and store it as the object `curry`.

```
curry = pd.read_csv("curry2015.csv")
```

3b: Let's say that your data is a sample. What is the population and sample for this scenario? We always want to be very specific when answering this question!

> Insert your answer & justification below.
> Population = All the games played by Curry's team in 2015
> Sample = All the games played by Curry in 2015

3c: Create a histogram of the number of points Curry scores in a game by adding in arguments to the function below, including any arguments needed to follow the guidelines for good graphics.

```
curry.hist(columns='Points',bins=10)
```

> Note: Ignore the Error with the histogram, Pandas has some interface issues with matplotlib

3d: Create a boxplot of the number of points Curry scores in a game by adding in arguments to the function below. Do not add additional arguments other than the three asked for here.

```
curry.boxplot(column="Points")
```

3e: Describe the distribution of the number of points Curry scores in a game, using both the histogram and boxplot, being sure to address all four characteristics of a distribution.

> Insert your answer below.
> The distribution is approximatley Uniform with a minimum less than 10 and a maximum greater than 50. There are no outliers in this set

3f: Based on your description, what are the most appropriate measures of center and spread for these data? Justify your answer.

> Insert your answer below.
> Either measure of center and spread would be accurate due to the uniform distribution of the points.

3g: We can use `length()` to find the number of cases in our dataset, but we know we aren't actually graphing that many data points, as Curry didn't play every game. Use `favstats()` to determine the number of games that Curry did and didn't play.

```
len(curry)
```

> How many games did and didn't Curry play? (Insert your answer below).

81-79 = 2

3h: Calculate the appropriate summary statistics (this means now that you now the shape of the distribution which of the summary stats are most relevant (resistant or non-resistant variables?), for the number of points Curry scores in a game. Write them below using mathematical notation, for example $s = 10$

Note: careful to think if you're calculating numeric summaries of a sample or a population here (Greek or non-Greek letter!).

```
mean(curry.Points)
stdev(curry.Points)
```

> Insert your answers below.
> 30.063291139240505
> 9.76432776599821

3i: Using in-line code, write a sentence that describes the center and spread of the dataset. You may need to insert a code chunk here, or you may not, depending on how you choose to perform this task.

> Insert your answer below.
> The mean is approximatly 30 points per gam with a standard deviation of 9.76 points per game.

3j: Using in-line code, write a sentence about the maximum number of points Curry scored in 2015.

> Insert your answer below.
> The most points scored by curry in 2015 was 53 points

Part 4: Championship Droughts

The Super Bowl is a yearly competition between the best two American football teams in the NFL. Abhi enjoyed watching the playoff games, and has heard announcers talking about the number of years it has been since different teams have won the Super Bowl. He was interested about the number of seasons since any football team (not just NFL) have won a championship. He researched the year each team in the league last won the Super Bowl and recorded the number of years (equivalent to a season) it has been since they won, based on the year the team was created. The data is stored in `SuperBowl_lastwin.csv`.

4a: What is the population and sample for this scenario?

> Provide an answer and justification below.
> All the football teams, the sample is the football teams in the NFL.

4b: Was this study observational or experimental? *Hint: there is only one specific manipulation that causes a study to be experimental: Randomly assigning the explanatory variable*

> Provide an answer and justification below.
> This study is observational since the data available is all that is there. The experimenters are not able to influence the teams results since the games have passed.

4c: Describe the sampling method Abhi used to select his sample.

> Provide an answer and justification below.
> He selected a specific subset of the sample (NFL teams).

4d: What type of variable is recorded in `SuperBowl_lastwin`?

> Provide an answer and justification below.
> There is 1 categorical variable, the team name, and there are two quantitative variables: the year won and years from the last win.

4e: Insert a code chunk that will read in your data and then create an appropriate graphic of your sample data. It should include any modifier as necessary to label the graphic.

```
champs.boxplot(column="YearsSinceWon")
```

4f: Calculate the relevant summary statistics. After your code chunk, describe each statistic using mathematical ($'s) and verbal notation (good ol' words).

```
median(champs.YearsSinceWon)
IQR(champs.YearsSinceWon)
favstats(champs.YearsSinceWon)
```

> Provide mathematical (e.g. $equation$) and verbal interpretation of each value calculated.
> The median is a good center beacuse the data is skewed left. The IQR helps us understand the spread better.

4g: Describe the distribution of the number of years since winning in Abhi's sample.

> Insert your answer below.
> The distribution is unimodal and left skewed. Many teams have never won with an approximatly normal distribution for the rest of the teams in terms of years since last win.

Before you finish -- check your knitted document. Is it easy to read? Do you need to add spaces anywhere to create line breaks (Remember 2 spaces are needed to make a line break)? Did you accidentally delete any of the answer formatting (the stuff in green after the >)?