Final Project

Julia Cuellar

2020-11-19

Instructions

Part 1

Throughout the semester you will incrementally create your individual final project. By the end of this week you will need to:

1. Identify at least 3 different datasets and perform some initial exploration.

Potential Sources for Datasets

a. Kaggle

b. Open Data Network

c. American Community Survey

d. Bureau of Labor Statistics

e. Bureau of Economic Analysis

f. Open Data Cincinnati

g. Data.gov

h. Healthdata.gov

i. Amazon Web Services Datasets

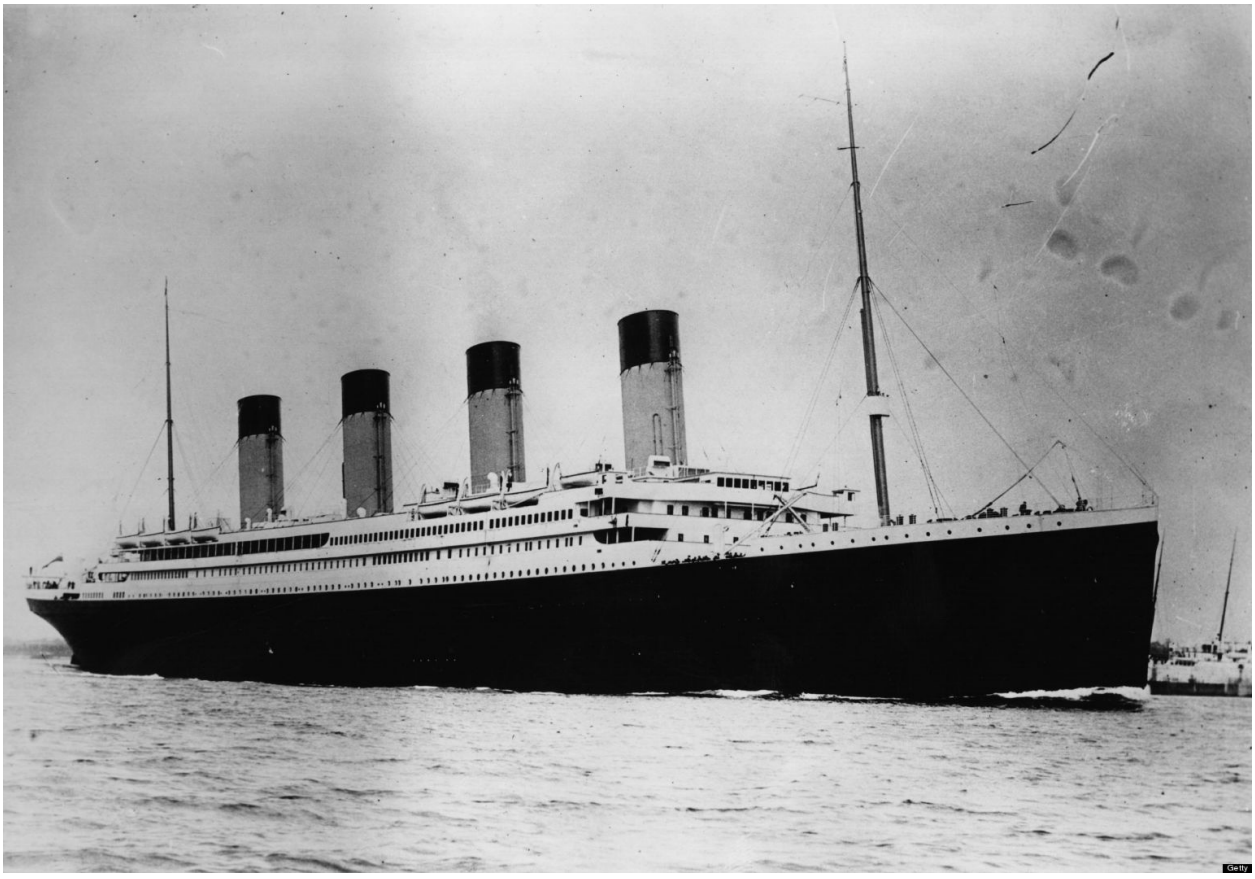
j. The General Society Survey

2. Put into practice what you have learned so far in the course.
3. Complete Section 1 of the Final Project Template and submit by the end of the week. Once you have narrowed down the 3 data set candidates for your final project you need to start thinking about what type of questions you would want to ask and answer of these data sets. You will draft research questions you want to use to guide your data analysis in a later week of the course as well as generally describe what you'll be doing for the next few weeks.

Titanic, Skulls, & Possum Part 1

Titanic

- The first data set I would like to explore is the Titanic data set which contains the variables:
  - Name
  - PClass
  - Age
  - Sex
  - Survived
  - SexCode



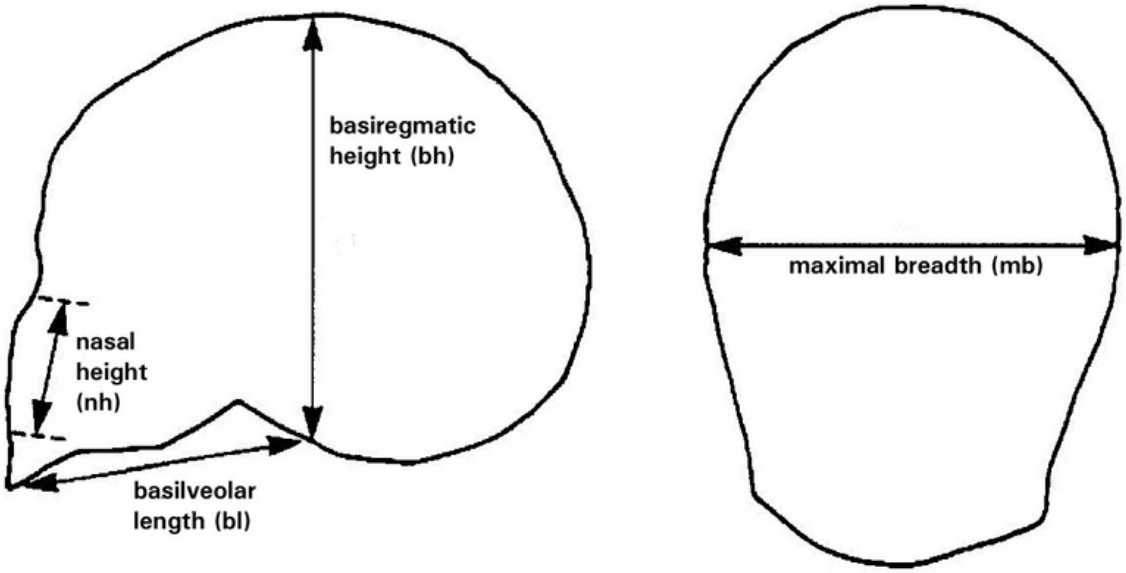
- As most should know about the tragic history about the Titanic (as well as the classic movie based off of the 1912 disaster), this is positively a data set I wanted to explore with an analysis of it. There are a few questions I propose from this data set:

1. What passengers survived more (1st, 2nd, or 3rd) or was reported as missing?
2. Which gender survived more (male or female)?
3. What was the survival rate versus death rate?

At the moment, I just know that the ggplot2 and cars packages will be utilized for histograms or scatterplots, but when plunging more into this analysis I probably will operate with more.

Skulls

- The second data set I would like to make an analysis of is called the skulls data set which encompasses the variables:
  - epoch
  - mb
  - bh
  - b1
  - nh



- I've always been fascinated with Egyptian mythology, so I definitively I have to make an analysis of this data set. There are a few questions I suggest from this data set:

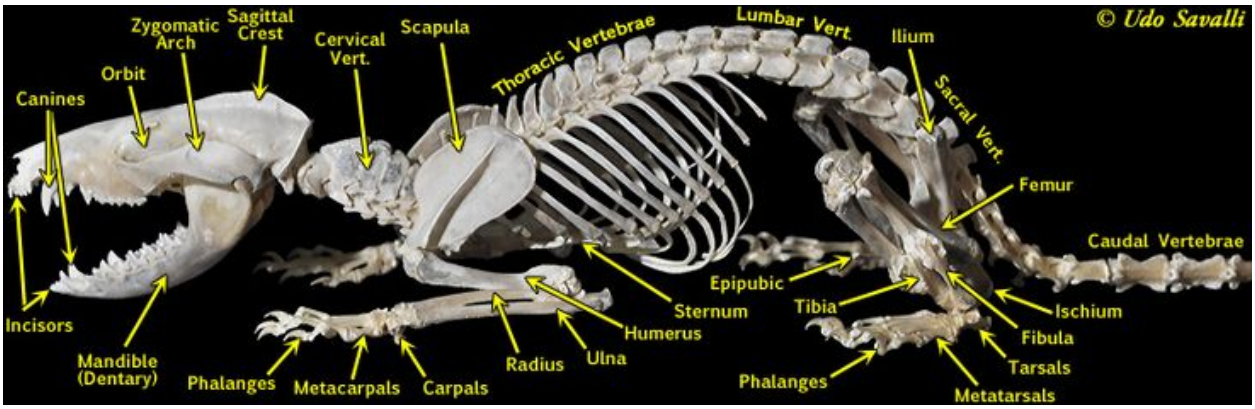
1. What is the significance of the epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) of where the skulls originated from?
2. Is there a relationship between basiregmatic height of the skull versus the epoch it was founded in?
3. Is there a relationship between basilveolar length of the skull versus the epoch it was founded in?
4. Is there a relationship between nasal height of the skull versus the epoch it was founded in?

At the moment, I just know that the ggplot2 and cars packages will be utilized for histograms or scatterplots, but when submersing more into this analysis I probably will harness more.

Possum

- The last data set I would like to dive in will be the possum data set which accommodates the variables:

- case
- site
- Pop
- sex
- age
- hdlngth
- skullw
- totlngth
- taill
- footlgth
- earconch
- eye
- chest
- belly



- Possums aren't necessarily cute, but they are the only marsupial in North America, plus this has many variables that I can prod with many analysis techniques. There are a few questions I frame from this data set:

1. What is the significance of the gender (male or female)?
2. Is there a relationship between total length versus the gender?
3. Is there a relationship between tail length versus the gender?
4. Is there a relationship between chest girth versus the gender?
5. Is there a relationship between belly girth versus the gender?

At the moment, I just know that the ggplot2 and cars packages will be utilized for histograms or scatterplots, but when sinking more into this analysis I probably will exploit more.

Part 2

Post the second step of your final project. Include the following:

Data importing and cleaning steps are explained in the text and in the Github exercises. (Tell me why you are doing the data cleaning activities that you perform). Follow a logical process.

With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

What do you not know how to do right now that you need to learn to import and cleanup your dataset?

Discuss how you plan to uncover new information in the data that is not self-evident.

What are different ways you could look at this data to answer the questions you want to answer?

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

How could you summarize your data to answer key questions?

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

What do you not know how to do right now that you need to learn to answer your questions?

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Titanic, Skulls, & Possum Part 2

Titanic

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##                                             Name PClass    Sex Survived SexCode
## 1                   Allen, Miss Elisabeth Walton    1st female        1       1
## 2                    Allison, Miss Helen Loraine    1st female        0       1
## 3           Allison, Mr Hudson Joshua Creighton    1st   male        0       0
## 4 Allison, Mrs Hudson JC (Bessie Waldo Daniels)    1st female        0       1
## 5                 Allison, Master Hudson Trevor    1st   male        1       0
## 6                            Anderson, Mr Harry    1st   male        1       0
```

Cleaning up the Titanic data required the dplyr package with select as well as subset to remove the first column (numbering the data set) and the Age column (missing data NA).

Since my first question to answer is which passengers survived more (1st, 2nd, or 3rd) or was reported as missing, the easiest way to answer is just subsetting the data by passenger class and then seeing which passengers survived more by the survived column. I will showcase a histogram for final results.

My second question to answer is in a similar fashion by subsetting the data by sex to observed which gender, male or female, survived more and again by using the survived column. I will showcase a histogram for final results.

My last question is the survival rate versus the death rate from the survived column and once again just subsetting the data. I will probably present a scatterplot for final results.

Skulls

```
##     epoch  mb  bh  bl nh
## 1 c4000BC 131 138  89 49
## 2 c4000BC 125 131  92 48
## 3 c4000BC 131 132  99 50
## 4 c4000BC 119 132  96 44
## 5 c4000BC 136 143 100 54
## 6 c4000BC 138 137  89 56
```

Cleaning up the skulls data was just a matter of removing the first column with [-1].

After cleaning up the data, to answer my first question for the significance of the epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) of where the skulls originated from, I will subset the data in the epoch column and then perform some form of regression analysis to determine the significance of each epoch. I will probably present a scatterplot for final results as well as a histogram.

To answer my second question about basiregmatic height of the skull versus the epoch it was founded in can be evaluated in a similar fashion to the previous question since I would already have the epoch column subsetted, I can then perform some form of regression analysis to determine the significance for the relationship between the two. I will probably present a scatterplot for final results.

The third question I propose was about the basilveolar length of the skull versus the epoch it was founded in can once again be sought out in a same fashion as the previous question by performing some form of regression analysis once again to determine the significance for the relationship between the two. I will probably present a scatterplot for final results.

My final question is about the nasal height of the skull versus the epoch it was founded in can once more be seek out in a like manner fashion to the previous questions by some form of regression analysis as well to determine the significance for the relationship between the two. I will probably present a scatterplot for final results.

Possum

```
##   sex hdlngth skullw totlngth taill earconch chest belly
## 1   m    94.1   60.4     89.0  36.0     54.5  28.0    36
## 2   f    92.5   57.6     91.5  36.5     51.2  28.5    33
## 3   f    94.0   60.0     95.5  39.0     51.9  30.0    34
## 4   f    93.2   57.1     92.0  38.0     52.2  28.0    34
## 5   f    91.5   56.3     85.5  36.0     53.2  28.5    33
## 6   f    93.1   54.8     90.5  35.5     53.6  30.0    32
```

Cleaning up the possum data was all by just harnessing data.frame and creating a new data set with the following variables: sex, hdlngth, skullw, totlngth, taill, earconch, chest, & belly. The other variables (the first column was just random letters and numbers, the second column was about the case number, the third and fourth columns were about the site and population, and lastly, the age and footlgth both had missing values NA) were unneeded and removed from the final cleaning of the data set.

Since my first question is determining the significance between the relationship between male and female, I will perform some form of regression analysis after subsetting the data. I will probably present a scatterplot for final results as well as a histogram.

My next question applies to the relationship between total length versus the gender and since my sex column is already subsetted, I can perform some form of regression analysis as well to determine the significance between the two. I will probably present a scatterplot for final results.

My following question about the relationship between tail length versus the gender succeeds with the same fashion as the prior question, so I can perform some form of regression analysis too to determine the significance between the two. I will probably present a scatterplot for final results.

The fourth question is about the relationship between chest girth versus the gender flows in a like fashion as the previous question, therefore, I can perform some form of regression analysis once again to determine the significance between the two. I will probably present a scatterplot for final results.

The last question is about the relationship between belly girth versus the gender and once more I can perform some form of regression analysis once again to determine the significance between the two. I will probably present a scatterplot for final results.

At the moment, I believe I will not implement any machine learning techniques, but rather, just some form of regression analysis.

Part 3

Post the last step of your Final Project. This should be an attached file that contains each step in the final project. Include the following:

Overall, write a coherent narrative that tells a story with the data as you complete this section.

Summarize the problem statement you addressed.

Summarize how you addressed this problem statement (the data used and the methodology employed).

Summarize the interesting insights that your analysis provided.

Summarize the implications to the consumer (target audience) of your analysis.

Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.

Titanic, Skulls, & Possum Part 3

Titanic

```
##     n
## 1 193
```

```
##     n
## 1 129
```

```
##     n
## 1 119
```

```
##     n
## 1 160
```

```
##     n
## 1 138
```

```
##     n
## 1 573
```

Since the count for each passenger class sorted by who survived and who died did not have the same dimensions, I was unable to present any plot (histogram or scatter or bar). Therefore, to answer my question about which passengers survived more (1st, 2nd, or 3rd) or was reported as missing (none in the data set), I will present a table below.

Passenger Class	Survived	Died
1st	193	129
2nd	119	160
3rd	138	573

Clearly seen from the table above, the 1st passenger class survived the most compared to the other two classes and the 3rd passenger class was the most who perished.

```
##     n
## 1 308
```

```
##     n
## 1 154
```

```
##     n
## 1 142
```

```
##     n
## 1 709
```

Since the count for the sex sorted by who survived and who died did not have the same dimensions, I was unable to present any plot (histogram or scatter or bar). Therefore, to answer my second question about which gender, male or female, survived more, I will present a table below.

Sex	Survived	Died
Female	308	154
Male	142	709

Clearly seen from the table above, females survived more compared to males who perished more.

```
##     n
## 1 450
```

```
##     n
## 1 863
```

Since the count for survival sorted by who survived and who died did not have the same dimensions, I was

unable to present any plot (histogram or scatter or bar). Therefore, to answer my third question about the survival rate versus the death rate, I will present a table below.

Survived	Died
450	863

Clearly seen from the table above, the survival rate was outweighed by the death rate.

Summary

The Titanic was a luxury British steamship that sank in the early hours of April 15, 1912 after striking an iceberg, leading to the deaths of more than 1,500 passengers and crew.

While exploring this data set, I wanted to know which passenger class survived more, which gender survived more, and what was the survival rate versus the death rate.

The only package I utilized for this data set was the dplyr which allowed me to harness select. I also operated with the subset and count functions.

Since I was unable to provide any form of plot (histogram, scatter, or bar), I presented 3 different tables categorizing passenger class versus survival rate, sex versus survival rate, and survival rate versus death rate.

As I mentioned earlier, the 1st passenger class survived more while the 3rd passenger class perished more. The conclusions I came to for this analysis was due to the wealthy having more influence compared to lower class individuals.

A similar deduction can be said about the survival rate for sex. As seen by other tragedies in history, women (and children) are more spared (or saved) compared to their male counterparts.

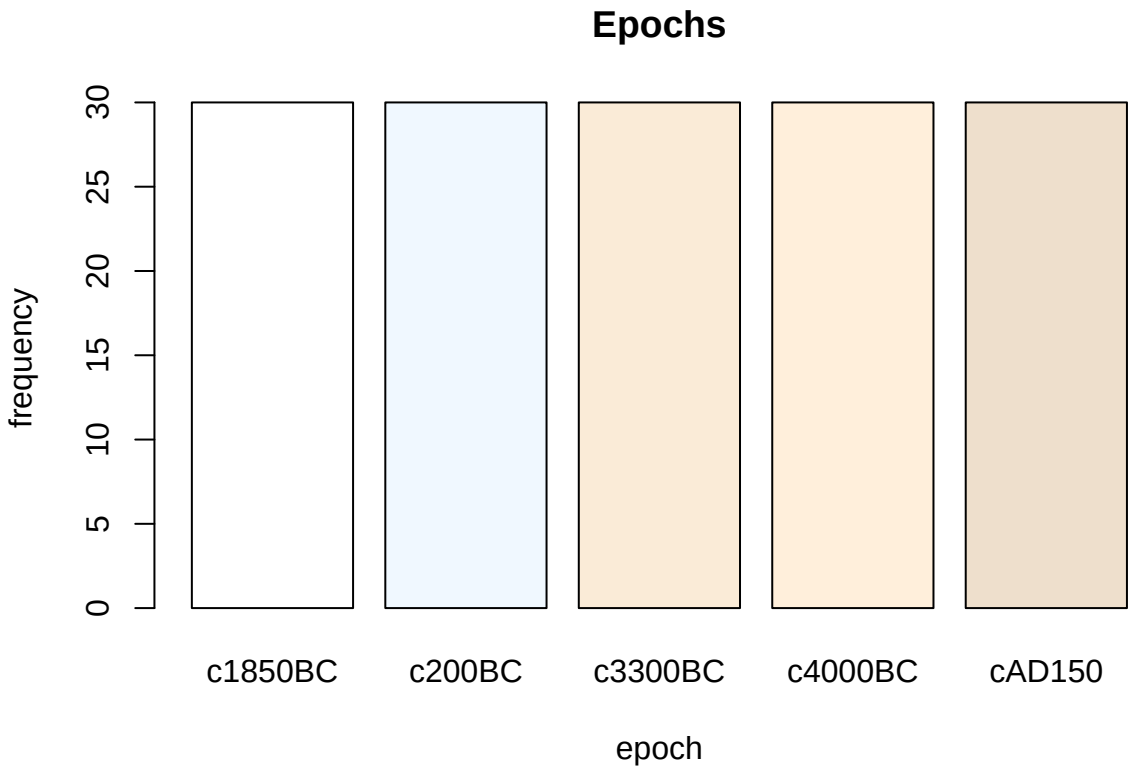
Lastly, since the Titanic is known as a catastrophe, the death rate would surpass the survival rate by a huge margin.

After doing this analysis upon this data set, a target audience would be anyone interested in an analysis of this data set or people who are history buffs.

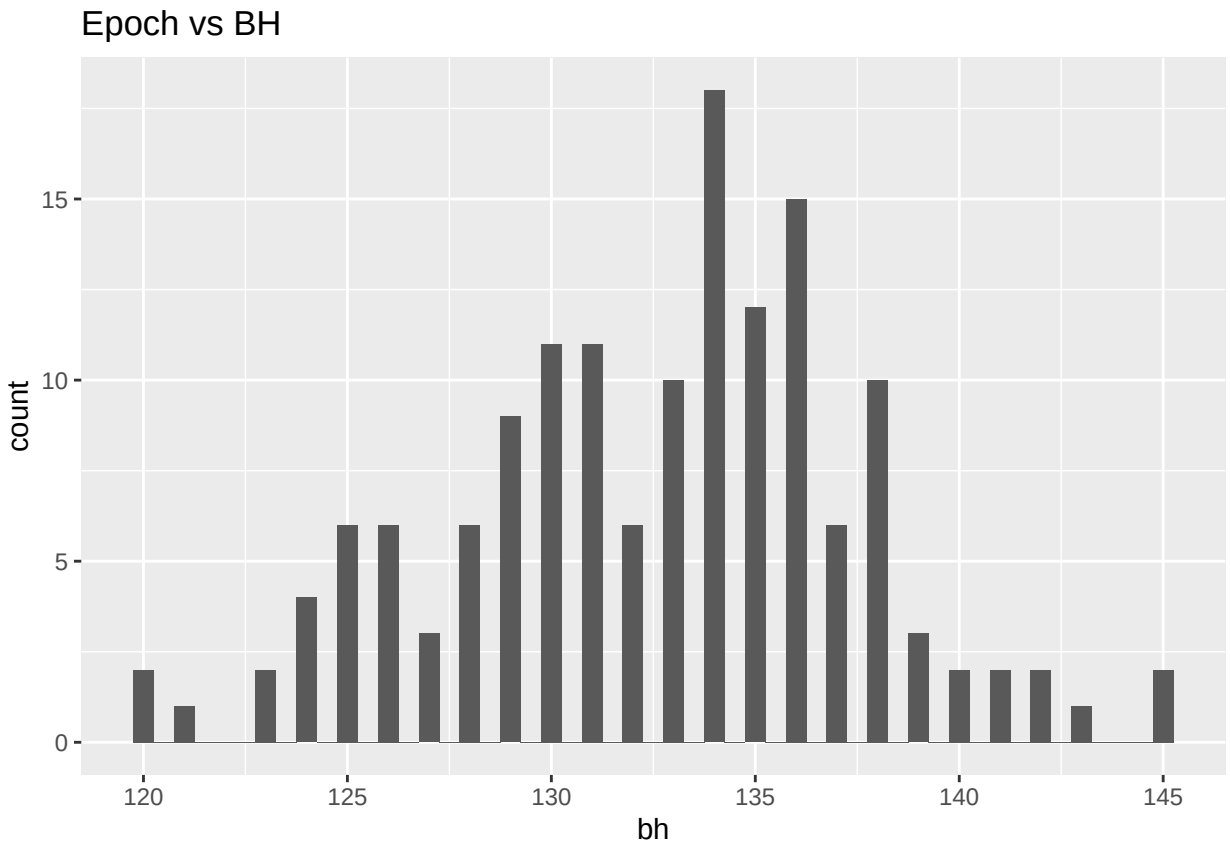
The main limitation that came from my analysis of this data set is that I was not able to showcase a plot (histogram, scatter, or bar) to display the passenger class versus survival rate, sex versus survival rate, and survival rate versus death rate, but I was able to present in a simple tabular format.

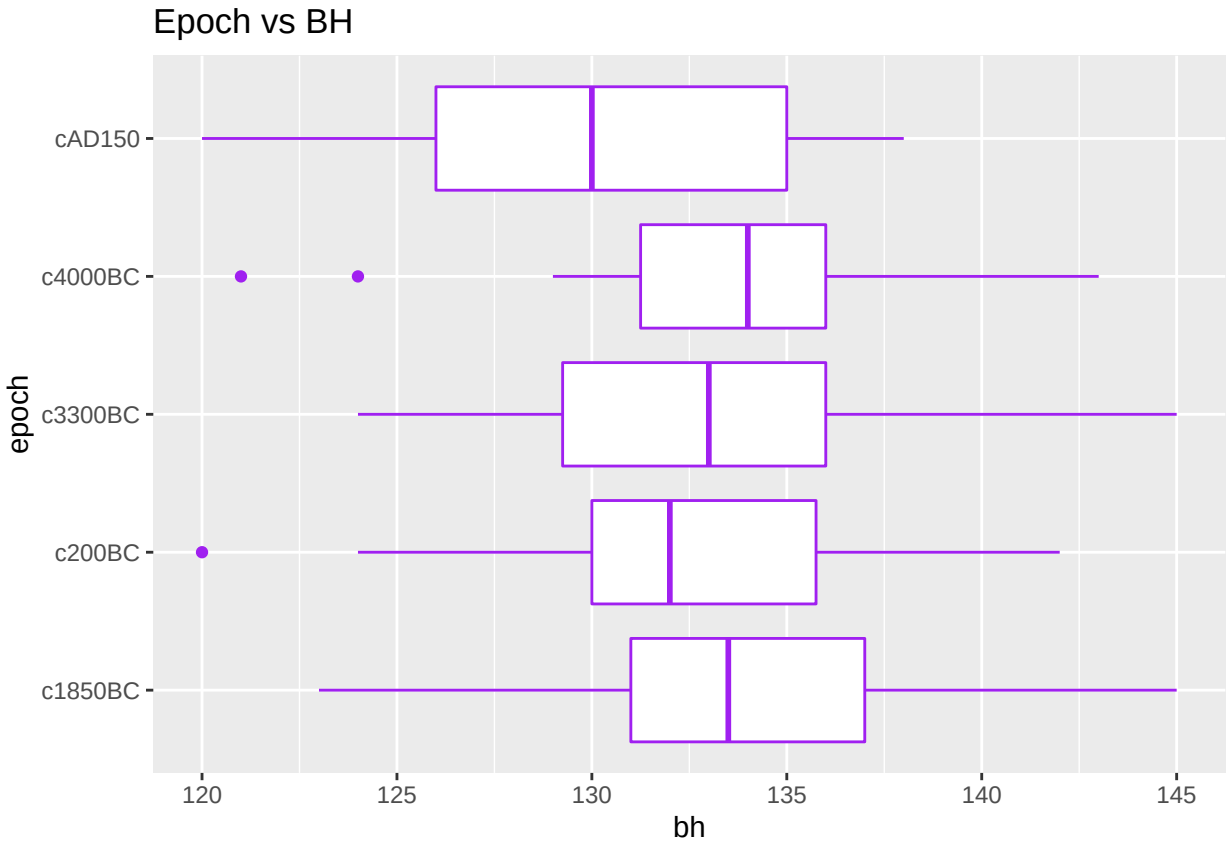
If someone were to analyze this data set, I hope they could provide some form of plot (histogram, scatter, or bar) besides a table for passenger class versus survival rate, sex versus survival rate, and survival rate versus death rate.

Skulls

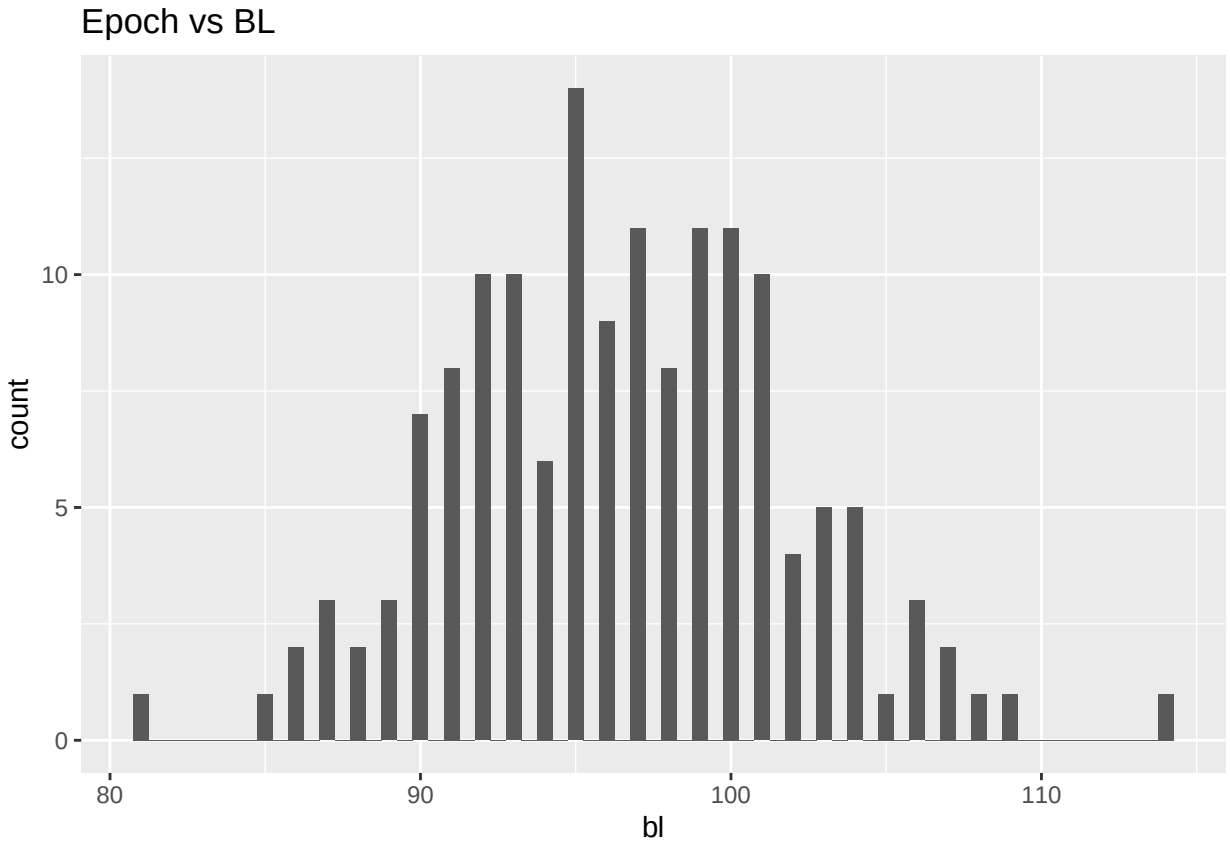


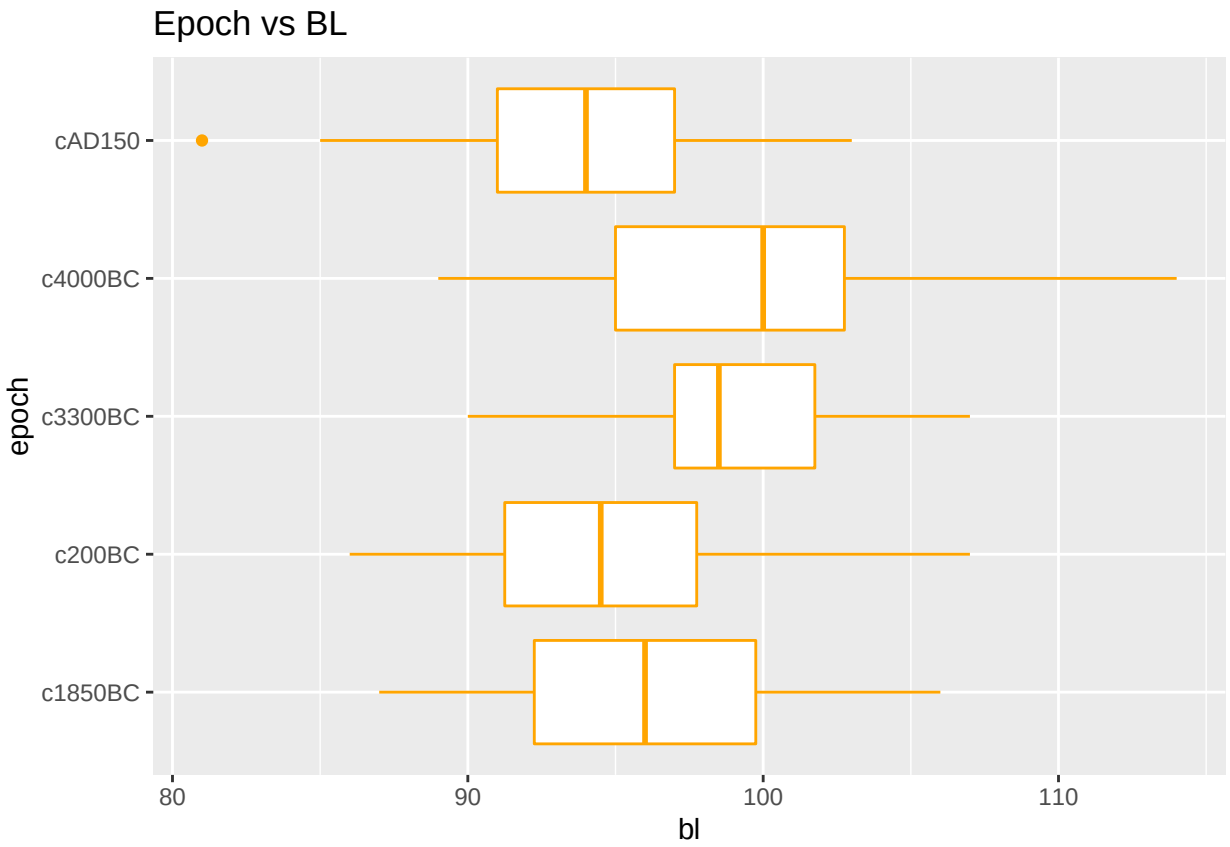
After observing the plot (a bar chart), the epochs were divided evenly, thus I conclude that there really isn't any significance between the various epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150).


Epoch vs BH
count
15
10
5
0
120
125
130
135
140
145
bh

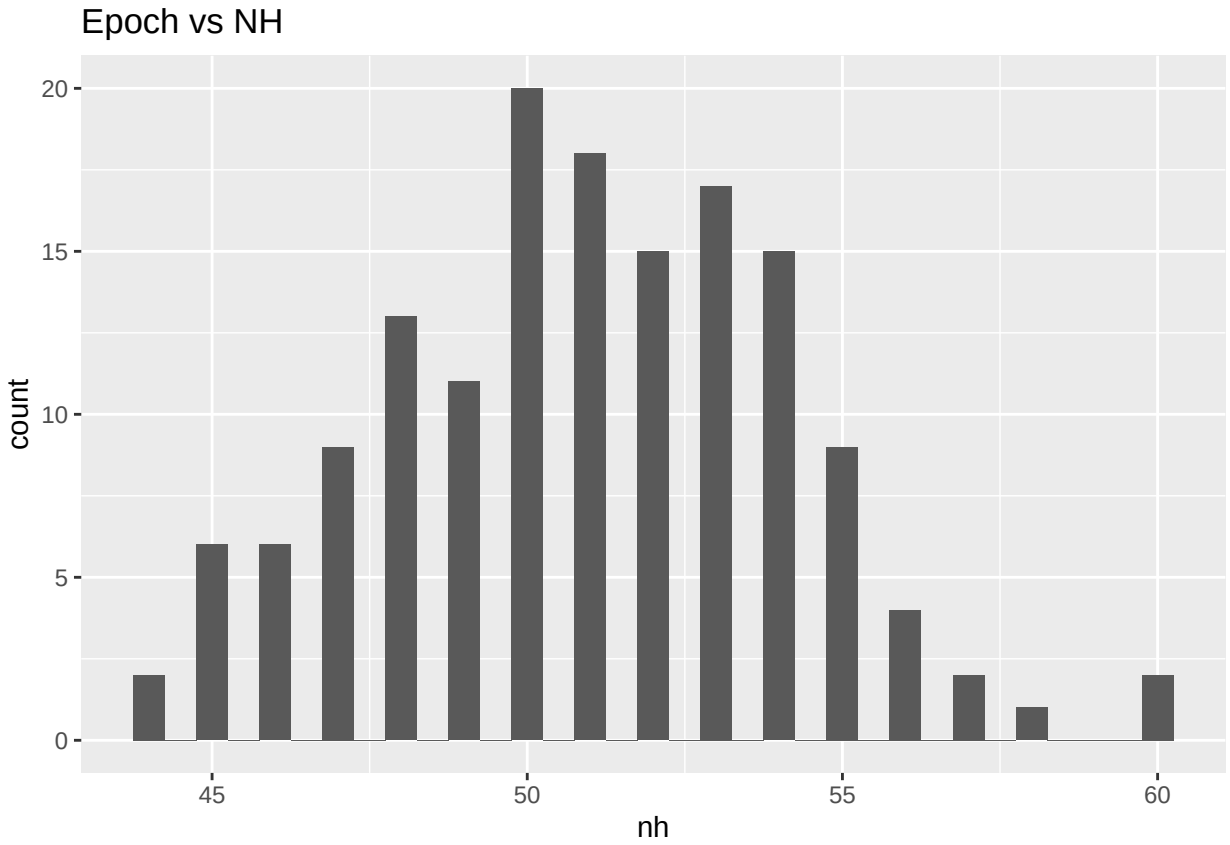


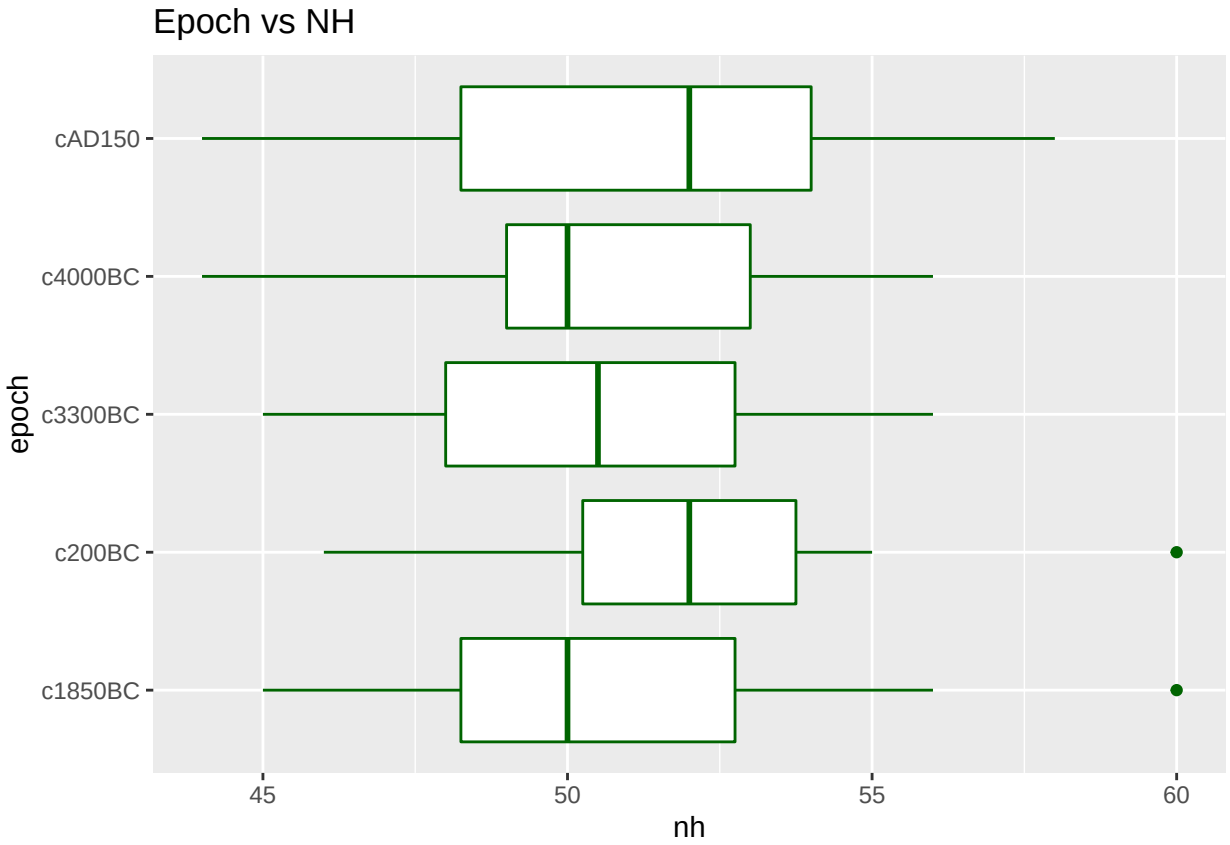
After observing both plots (histogram and boxplot), I have determined that there isn't really a significance between the epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) where the skull was founded in versus the bh (basiregmatic height of the skull).


Epoch vs BL
count
10
5
0
80
90
100
110
bl



After observing both plots (histogram and boxplot), I have determined that there isn't really a significance between the epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) where the skull was founded in versus the bl (basilveolar length of the skull).


Epoch vs NH
count
20
15
10
5
0
45
50
55
60
nh



After observing both plots (histogram and boxplot), I have determined that there isn't really a significance between the epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) where the skull was founded in versus the bl (nasal height of the skull).

Summary

An epoch is described as a period of time in history, typically one marked by notable events or particular characteristics. This data set contains a set of 5 different epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) for Egyptian skulls rendered by the characteristic measurements of maximal breadth, basiregmatic height, basilveolar length, and nasal height.

While exploring this data set, I wanted to know the significance of the epochs of where the skulls originated from, the relationship between basiregmatic height of the skull versus the epoch it was founded in, the relationship between basilveolar length of the skull versus the epoch it was founded in, and the relationship between nasal height of the skull versus the epoch it was founded in.

The only package I utilized for this data set was the ggplot2 which allowed me to plot histograms and boxplots. I also operated with the as.factor function.

As I mentioned earlier, there isn't really a significance between the various epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) due to the division of the epochs being the same.

A similar deduction can be said about there being no significance between the epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) where the skull was founded in versus the bh (basiregmatic height of the skull).

Another alike conclusion can be said about there being no significance between the epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) where the skull was founded in versus the bl (basilveolar length of the skull).

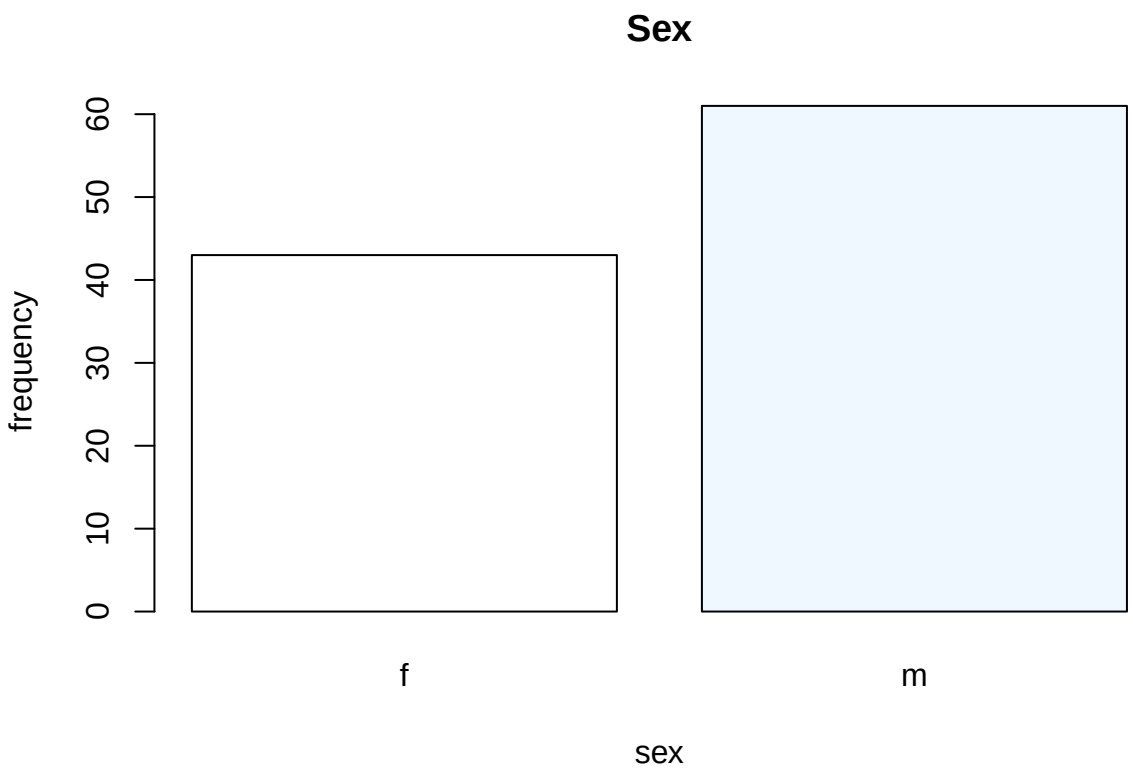
Lastly, the same diagnosis can be said about there being no significance between the epochs (c4000BC c3300BC, c1850BC, c200BC, or cAD150) where the skull was founded in versus the nh (nasal height of the skull).

After doing this analysis upon this data set, a target audience would be anyone interested in an analysis of this data set or people who are geologists, archaeologists, or any scientist that handles Egyptian historical artifacts.

The main limitation that came from my analysis of this data set was my limited knowledge (and time) of a regression model of categorical data versus quantitative data.

If someone were to analyze this data set, I hope they could perform a regression model of categorical data versus quantitative data.

Possum



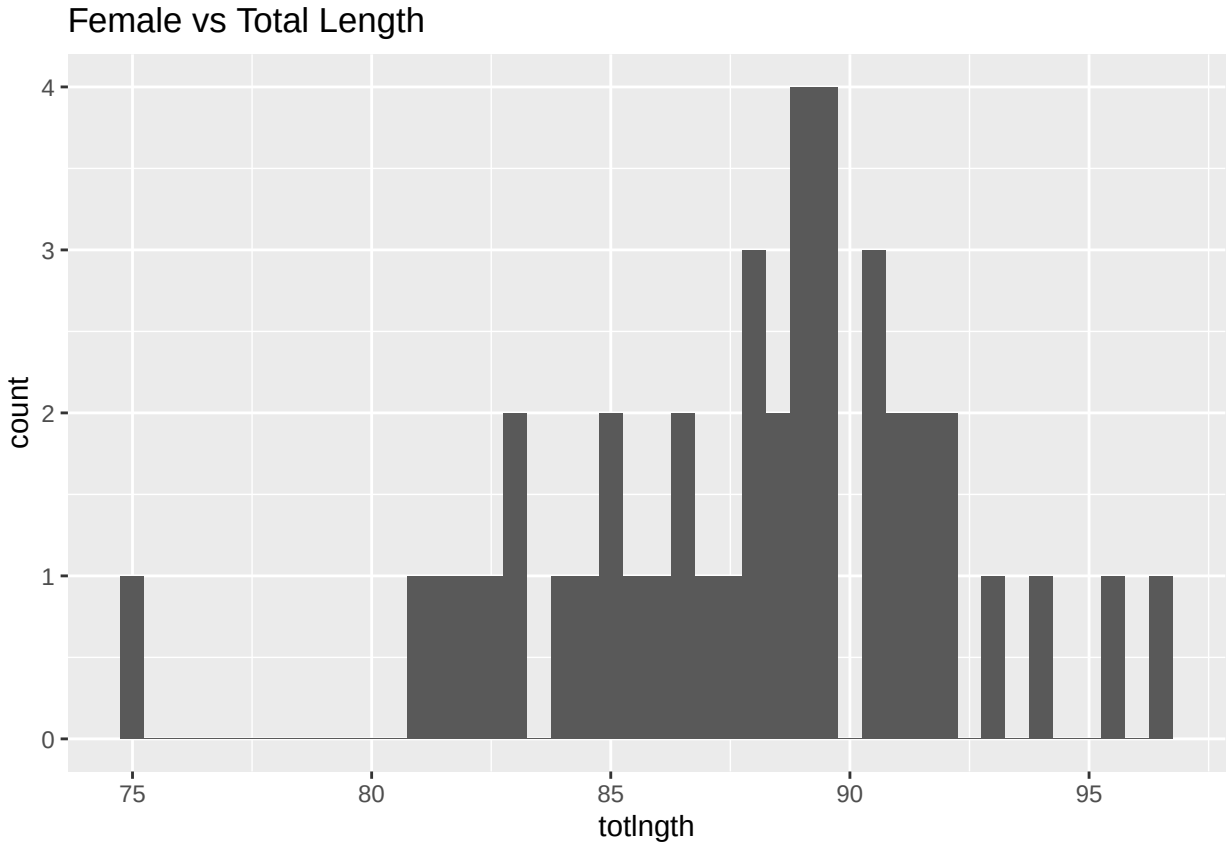

```
##    n
## 1 43
```

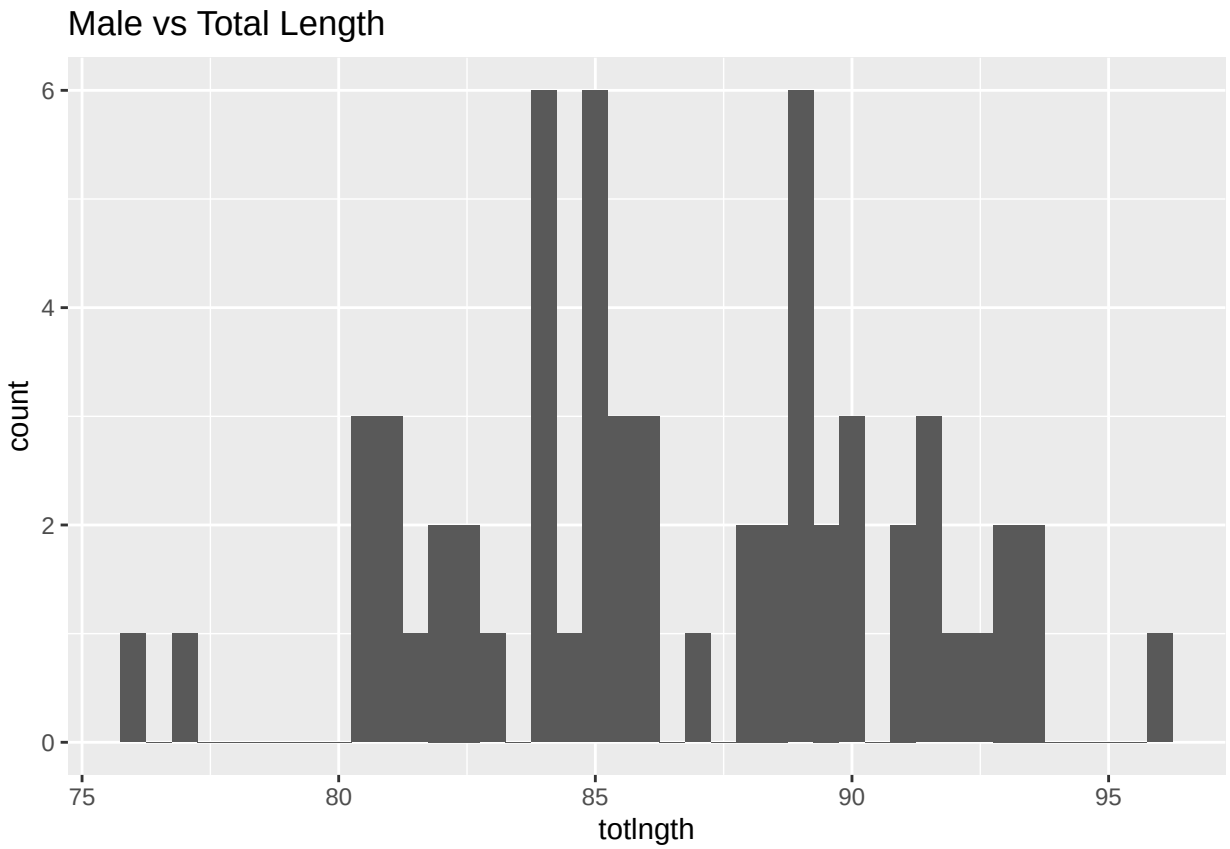
```
##    n
## 1 61
```

```
## [1] 0.704918
```

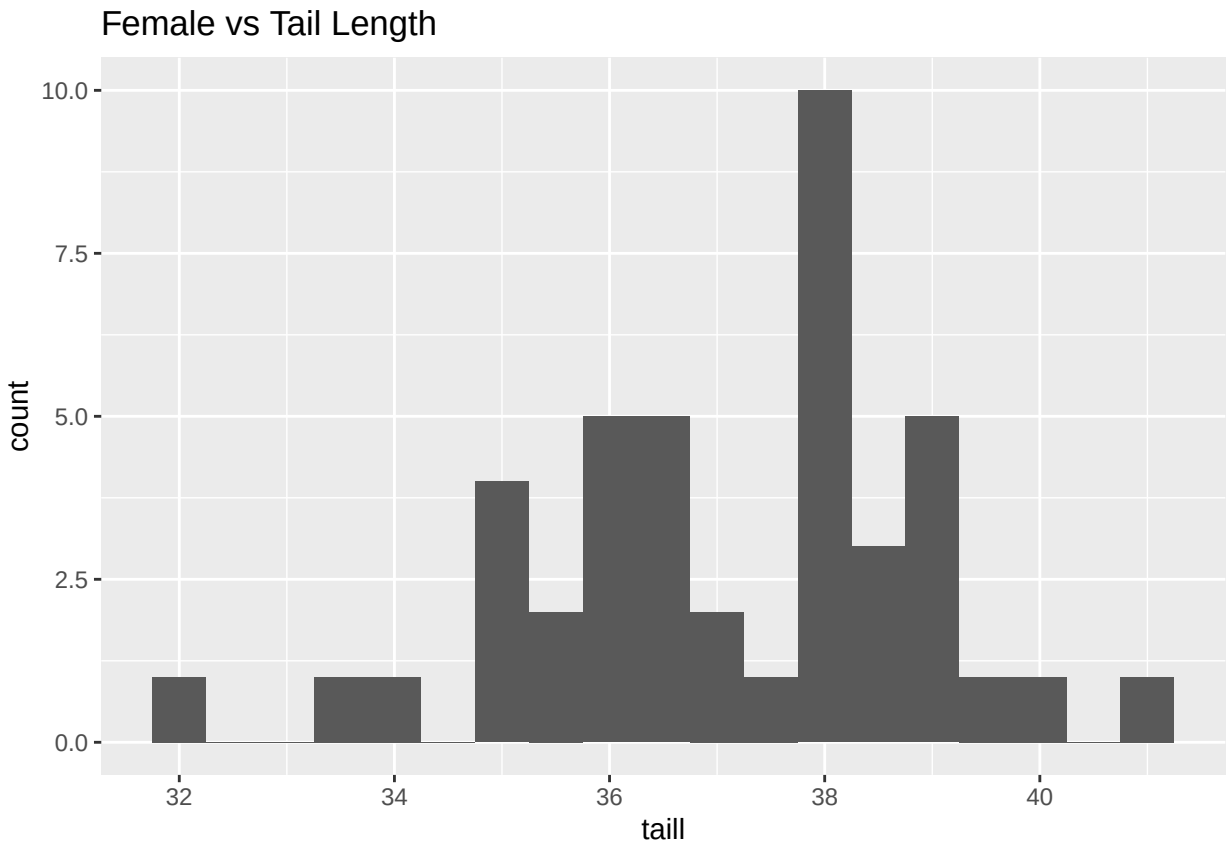
Sex	Frequency	Odds
Female	43	
Male	61	0.7

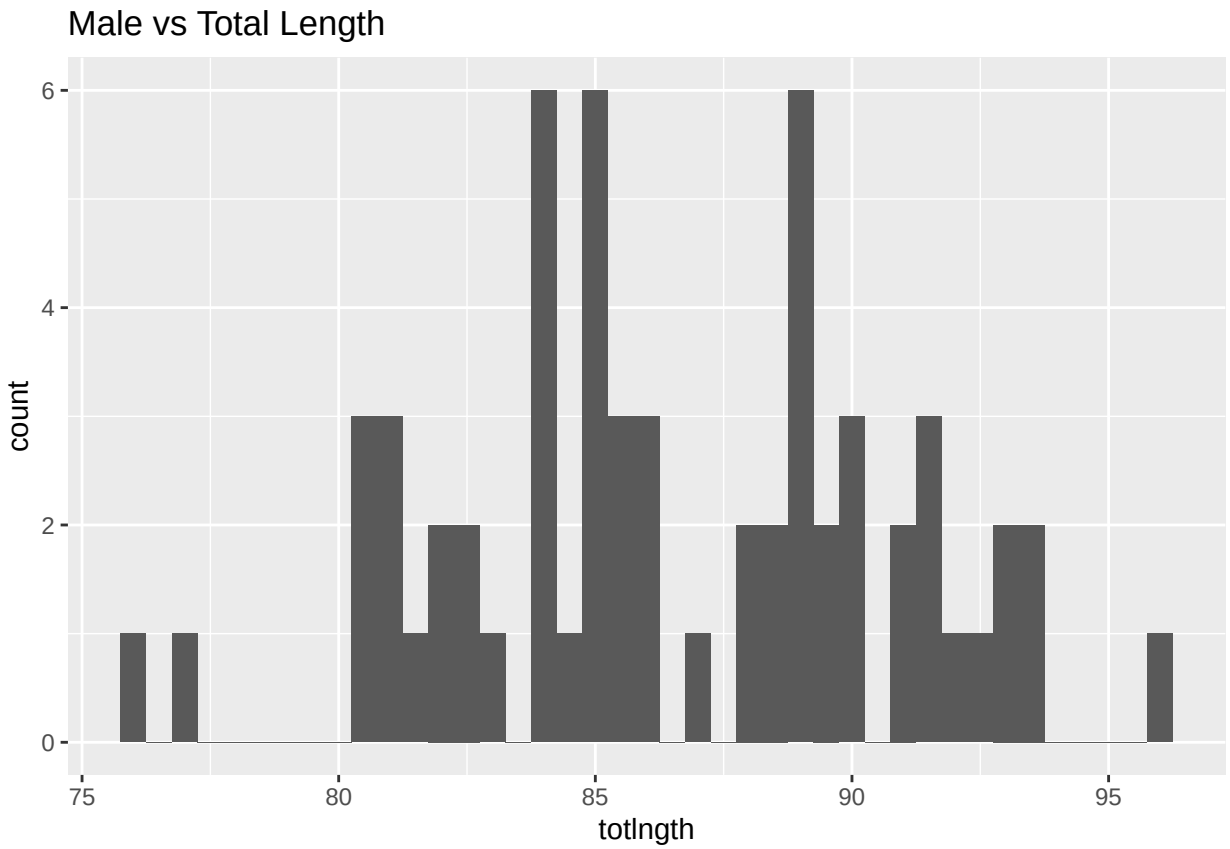
As seen from the plot (bar chart) and the table above, there is a significance between the sex (male or female) by a margin of about 70%.



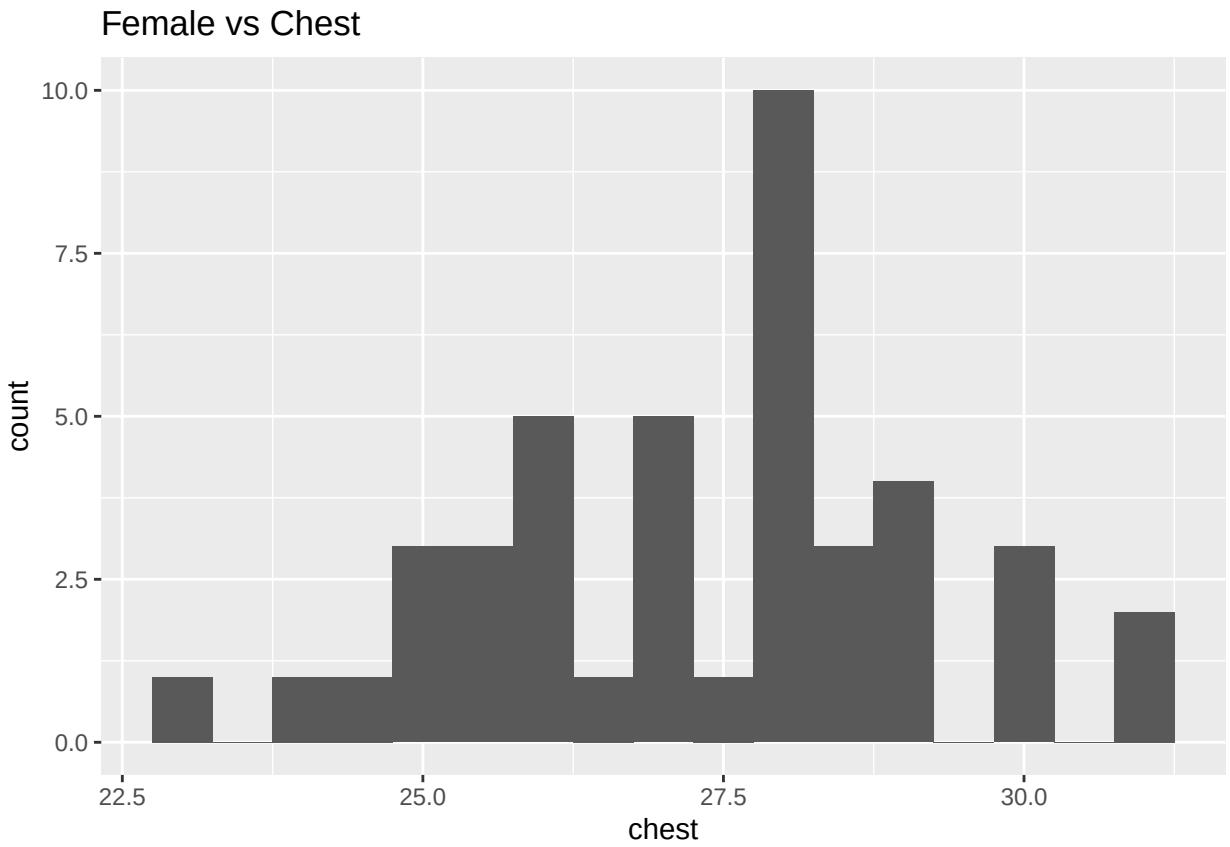


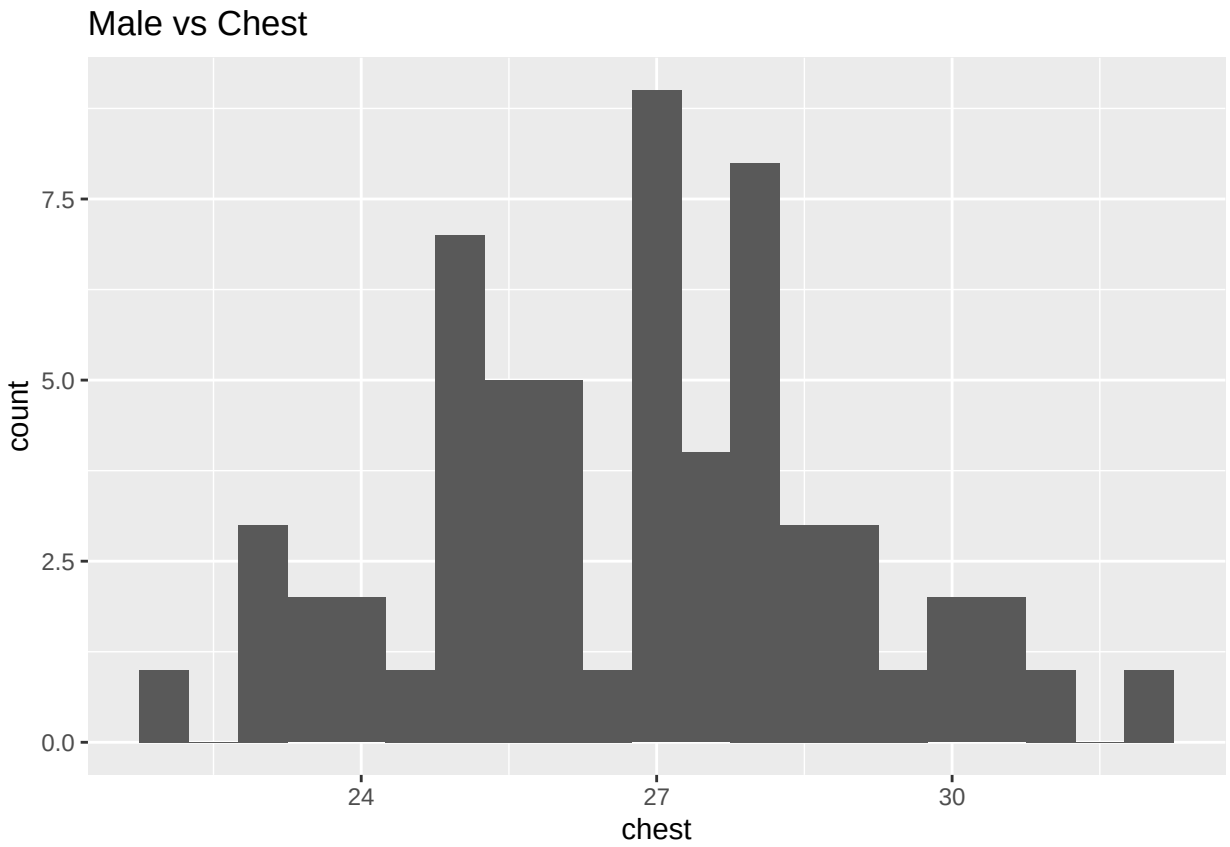
Comparing the 2 histograms of female versus total length and male versus total length, there seems to be a relationship between the sex and total length of the possum.


Female vs Tail Length
10.0
7.5
5.0
2.5
0.0
count
32
34
36
38
40
taill

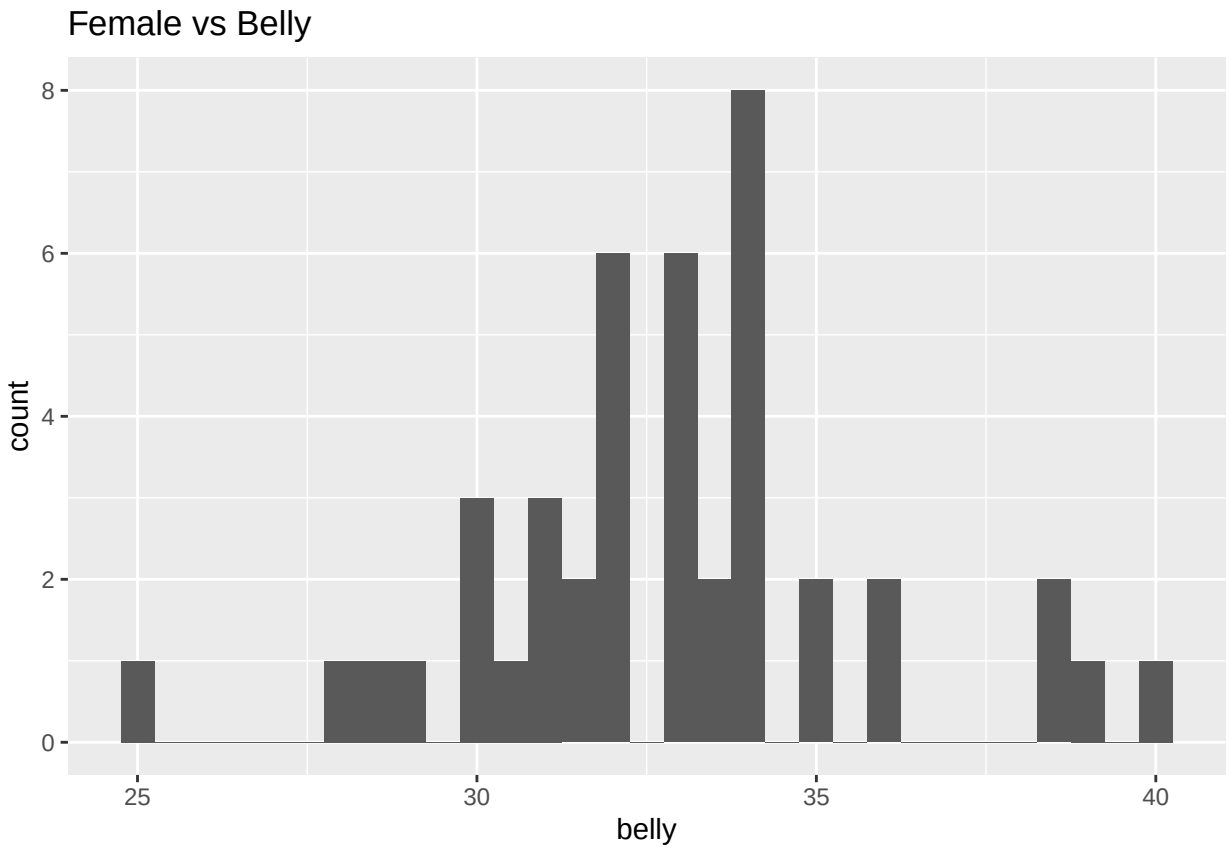


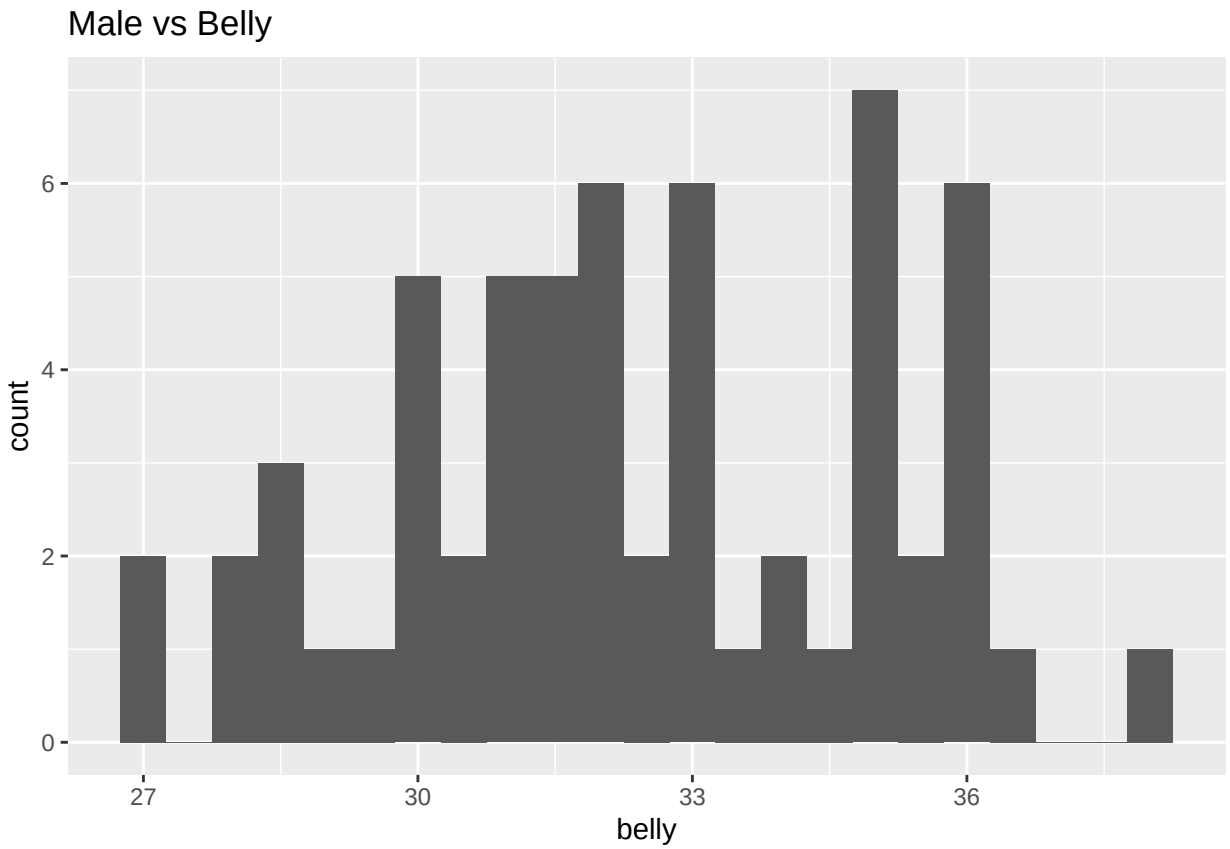
Comparing the 2 histograms of female versus tail length and male versus tail length, there seems to be a relationship between the sex and tail length of the possum.


Female vs Chest
10.0
7.5
5.0
2.5
0.0
count
22.5
25.0
27.5
30.0
chest



Comparing the 2 histograms of female versus chest girth and male versus chest girth, there seems to be a relationship between the sex and chest girth of the possum.


Female vs Belly
count
8
6
4
2
0
25
30
35
40
belly



Comparing the 2 histograms of female versus belly girth and male versus belly girth, there seems to be a relationship between the sex and belly girth of the possum.

Summary

The possum is the only marsupial located in North America who is an omnivore by nature to where their species has the male counterparts being slightly significantly larger than their female counterparts.

While exploring this data set, I wanted to know the significance of the gender (male or female), the relationship between total length versus the gender, the relationship between tail length versus the gender, the relationship between chest girth versus the gender, and the relationship between belly girth versus the gender.

The only packages I utilized for this data set was the dplyr and ggplot2 which allowed me to harness select and plot histograms. I also operated with the as.factor, subset, and count functions.

As I mentioned earlier, there is a significance between the sex (male or female) due to the male dominance over their female counterparts by about 70%.

A similar deduction can be said about there being a significant relationship between the sex (male or female) versus the total length due to the males being slightly larger than their female counterparts.

Another alike conclusion can be said about there being a significant relationship between the sex (male or female) versus the tail length due to the males being slightly larger than their female counterparts.

Once again, same diagnosis can be said about there being a significant relationship between the sex (male or female) versus the chest girth due to the males being slightly larger than their female counterparts.

Lastly, equivalent results can be said about there being a significant relationship between the sex (male or female) versus the belly girth due to the males being slightly larger than their female counterparts.

After doing this analysis upon this data set, a target audience would be anyone interested in an analysis of this data set or people who are veterinarians, possum lovers, zoologists, or any scientist/doctor who studies the biology/anatomy of a possum.

The main limitation that came from my analysis of this data set was my limited knowledge (and time) of a multivariate regression model of categorical data versus quantitative data.

If someone were to analyze this data set, I hope they could perform a multivariate regression model of categorical data versus quantitative data.

Conclusion

Frankly speaking, if I had more time to work on a single data set (rather than 3), I probably would have explored more with the possum data set and even performed a multivariate regression model. Alas, I procrastinated, but still provided a decent EDA of all of 3 of my data sets (Titanic, skulls, and possums). However, I hope that either me or someone with an extensive knowledge of regression models (uni to multi) could further expand on these data sets and be able to provide better plots.

I hope you enjoyed my EDA of the data sets Titanic, skulls, and possum.

References

https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/datasets.csv

https://vincentarelbundock.github.io/Rdatasets/doc/Stat2Data/Titanic.html

https://vincentarelbundock.github.io/Rdatasets/doc/HSAUR/skulls.html

https://vincentarelbundock.github.io/Rdatasets/doc/DAAG/possum.html