Biweekly Challenge #4 - Tell a Story!

Due Thursday, December 21th at 5PM

Please show all output but no code, warnings, or messages, and make sure your assignment is no more than one page. Submit the assignment as a PDF file along with the .RMD used to create it in your Google Drive folder.

The goal of this challenge is practice your data storytelling skills on a dataset of your choice. We recommend looking through R's included datasets using the command `library(help = "datasets")` to find one that seems interesting to you, then loading it using the command `data(dataName)`.

Exercise 1. For this assignment, you should pick a dataset of your choice and create 1-2 elements (such as visualizations, models, or tables) that help illustrate something about the data you found interesting i.e. the 'story' you are trying to tell about it. Your elements do not have to be complex, but should instead highlight interpretability - the easier we can understand your story, the better! Submissions should include descriptive text that explains why you chose the dataset and illuminates your story. *20 pts*

10 points will be awarded for creating a single element, with the other 10 coming from your explanation. If you choose to create any additional elements, you can get up to 10 additional points of extra credit on the assignment.

Example Submission

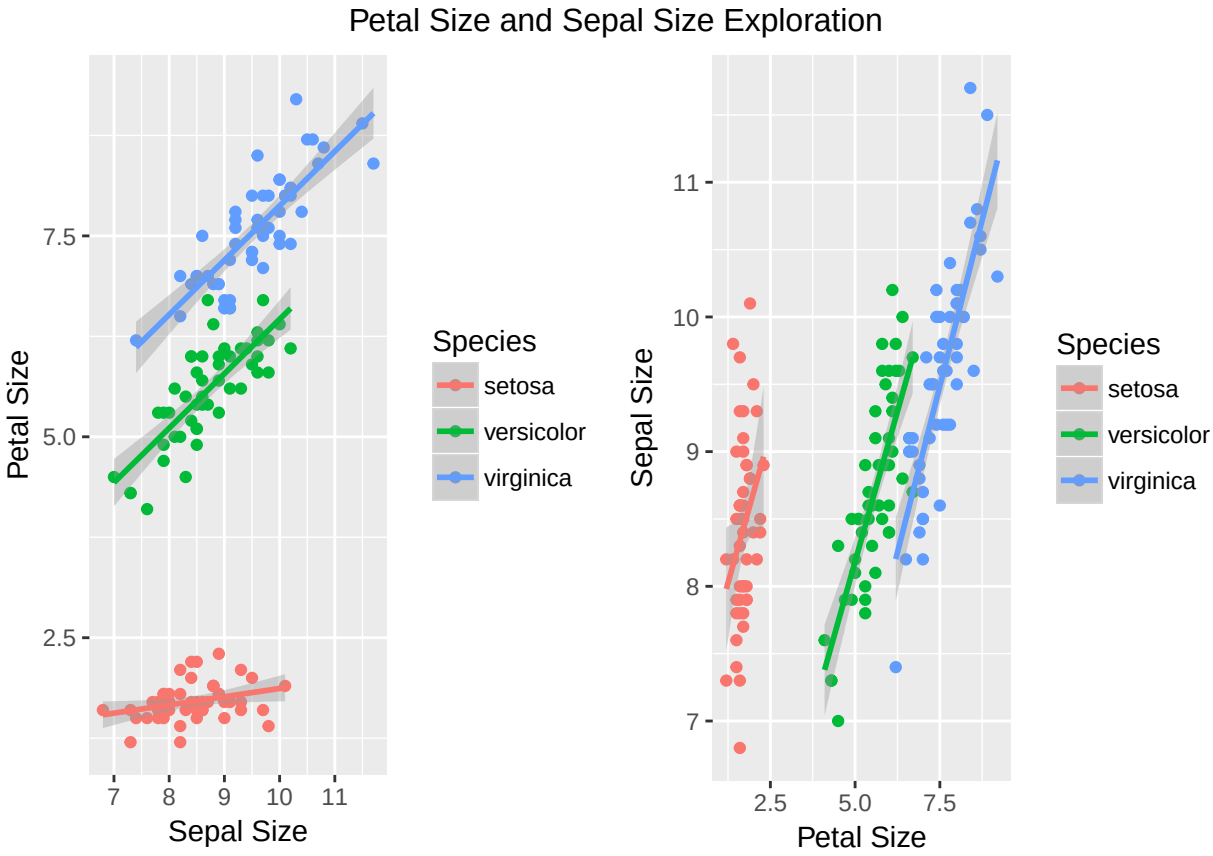
I have chosen to work with the `iris` dataset, a classic dataset used in elementary machine learning and statistical analysis for multiclass classification tasks. My story will involve describing differences in sepal and petal measurements by species, as well as determining the correlation between these variables. Let's view the first few observations of the dataset.

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

First, we can quantify average differences across species through the use of a table, noticing that each species appears to increase in size measurements from setosa to versicolor to virginica, with each being strictly greater than the one before it. Setosas also have a noticably lower standard deviation in petal size compared to the other species.

```
## # A tibble: 3 x 6
##      Species avg.sepal.size  sepal.sd avg.petal.size  petal.sd total.size
##       <fctr>          <dbl>     <dbl>          <dbl>     <dbl>      <dbl>
## 1  virginica          9.562 0.8341903          7.578 0.6911363     17.140
## 2 versicolor          8.706 0.7316211          5.586 0.6372806     14.292
## 3     setosa          8.434 0.6829139          1.708 0.2310932     10.142
```

We can use a graph to highlight these relationships further and get a sense of how sepal size influences petal size by species, and vice-versa.



Based on the plots, it is clear that sepal size and petal size have a positive linear relationship. We can test the strength of this relationship through its correlation.

```
## [1] 0.6005739
```

Conclusion

From the results of this analysis, we can conclude that each species exhibits clear differences between petal and sepal size, with virginica being strictly greater than the other two and versicolor being strictly greater than setosa. Sepal and petal size themselves are also positively correlated. In the future, clustering techniques like k-nearest neighbors could be effective in predicting species given the strong distinctions each species has in terms size measurements.