Date_a_Scientist Capstone

Sjoerd Kelderman
2/11/2018

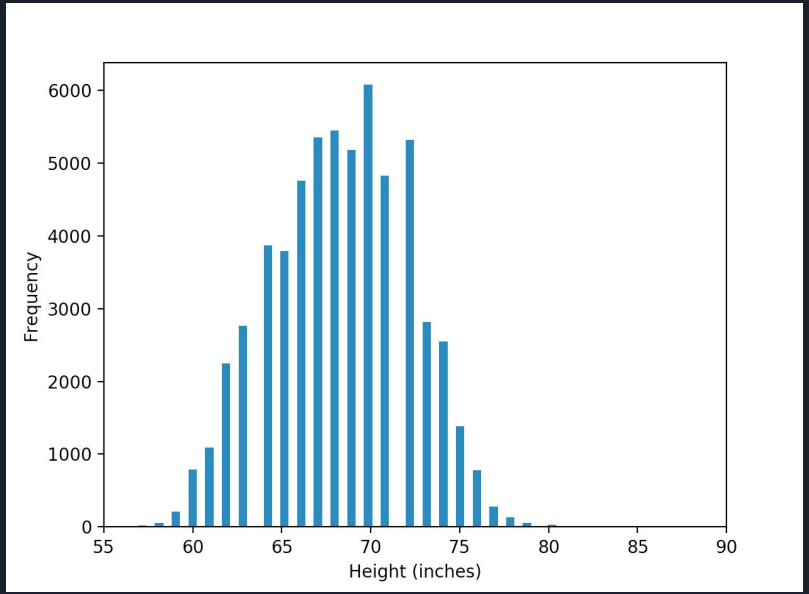
Contents

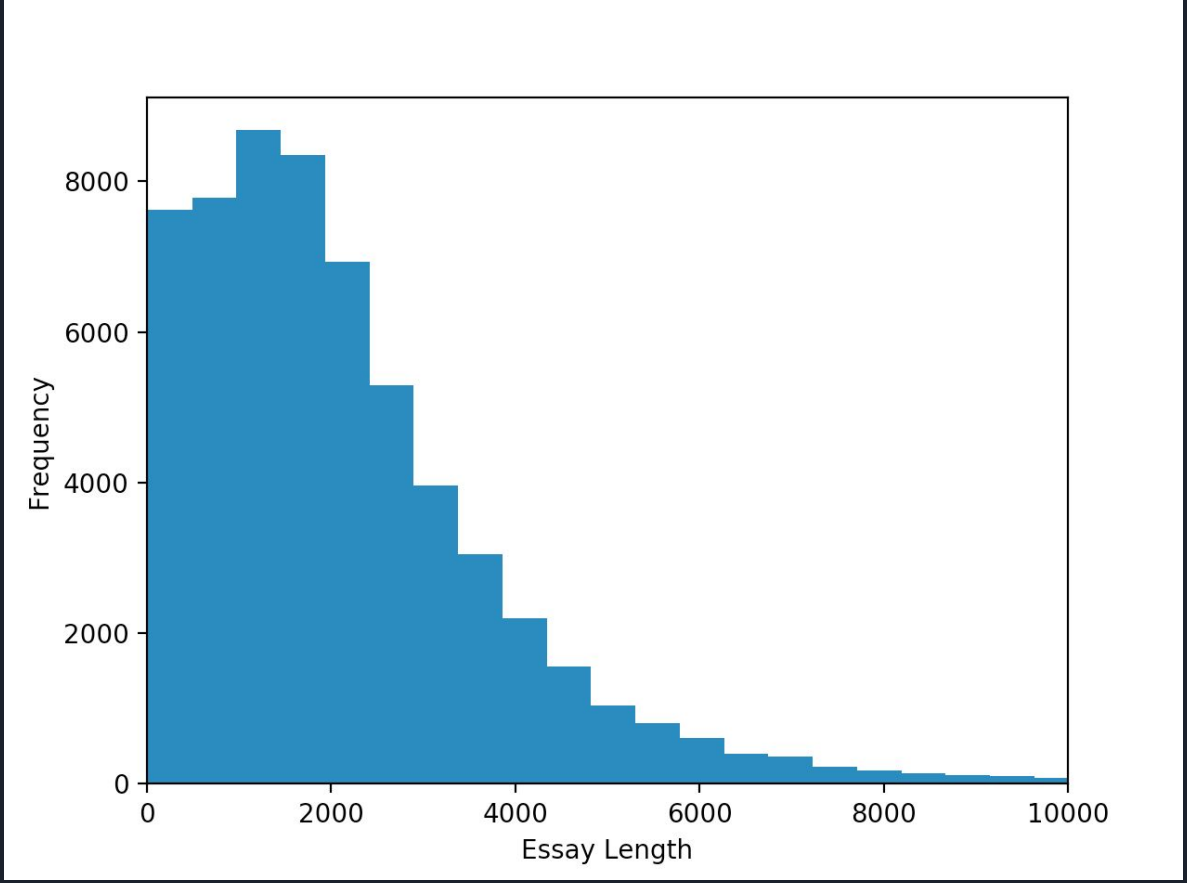
Data Exploration

I first printed out a few different data heads as well as value counts to see what some of the data types had for options.

Then I graphed a histogram of the frequency of different heights. (shown to the right.

Finally, I graphed the frequency of different essay lengths. (shown on the next slide.)




8000
6000
4000
2000
0
Frequency
0
2000
4000
6000
8000
10000
Essay Length

Questions

After exploring, I settled on two different questions.

The first was given alcohol, drugs, and diet, can I predict body type? This I was going to approach with a classifier.

The second was given sex and education, can I guess essay length? I wanted to use reggression for this.

Data Augmentation

I chose to create an essay_len section that was the total essay length in char count for each person.

I also changed a lot of variable sections to be number codes instead to work with the data.

Classification Approaches

I implemented a K-Means classifier and a Support Vector Machine for my two classifiers.

Both classifiers had pretty long runtimes, and the support vector machine performed slightly better in terms of accuracy.

I think I would have been able to find a K that had better accuracy as it was around 23 % but I ran out of time to test more k values since the runtime was so long.

More accuracy scores are printed/listed in the code.

Regression Approaches

For regression, I implemented multiple linear regression and a K-Means regressor.

Neither of these had good accuracies. Both were around 1% with the accuracy of the K-Means for the test data even being around 0.1%

I think part of the difficulty was the ordering of the data. I hadn't organized it beforehand and realize now that maybe that affected the regression.

Conclusion

I think with more time I could have had better accuracy with both classification and regression. I wanted to be able to print out graphs for the regression as well, but since I didn't organize the data sequentially initially, I didn’t have the time.

I ran out of time at the end because I had another class I was taking as well, and the given time estimate for this project led me to believe that I would be able to do it in a day.

Overall, it was a great learning experience still.