Karen W. Wells

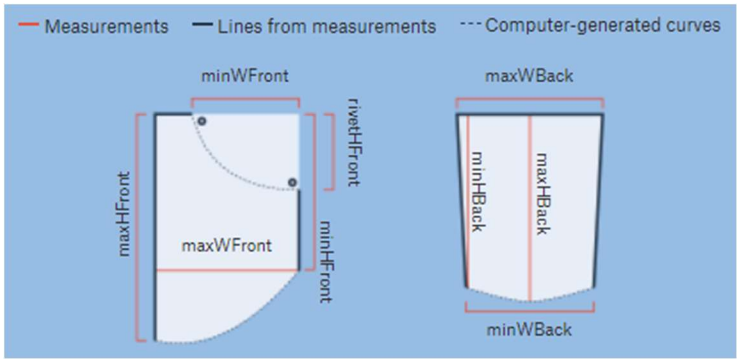
Python for Data Analysis: Spring 2022 Final Project

https://github.com/KarenWWells/Can-Code-Class/blob/main/KWW_Final_Project.ipynb

While searching for a dataset for the final project, I came upon an interesting article in The Pudding. (The Pudding is a digital publication that creates visual essays with data. (https://pudding.cool/)). I discovered the article, "Someone Clever Once said Women Were Not Allowed Pockets" by Jan Diehm and Amber Thomas (https://pudding.cool/2018/08/pockets/) in which the pocket dimensions of 40 pairs of men's jeans and 40 pairs of women's jeans were analyzed. The overall findings document "just how inferior women's pockets are to men's" – which came as no surprise to me as a women. Diehm and Amber report that "on average, the pockets in women's jeans are 48% shorter and 6.5% narrower than men's pockets". They further explore "fashion over function: what items can actually fit in pockets?" The authors incorporate amazing creativity when displaying the pocket size dimensions. The graphics are clever and effective in portraying the results of the data analysis through scroll-driven animation. The reader is also able to explore the functionality of women's pockets, compared to men's pockets, by electronically placing pre-defined objects, like a cell phone or a hand, in an average women's front pocket and an average men's front pocket. It is truly a brilliant display of data. For example, only 40% of women's front pockets can fit an I-Phone X whereas all of the men's front pockets can fit the I-Phone X. These results were consistent whether discussing skinny jeans or straight jeans. The methods used to obtain the dataset are presented along with how measurements were taken, etc. It is a compelling research paper with a clear and deliberate message - the functionality of women's pockets are inferior.

The Puddle article did a terrific job presenting the data. So why would I choose to re-examine this topic with my new found rudimentary Python skills? Since I will be teaching a new year-long data science course to high school students in the fall, the answer is straightforward. I can illustrate to my students the differences between innovative types of displays, as presented in the article, with traditional data displays. I can also empathize, as their teacher (and a newbie coder!), that it **is** hard to write code. However, it can be exciting and fun to create the many different data displays. Finally, I can demonstrate that the field of data science is new, emerging, creative and fun – and they might want to continue their studies so they too can create such innovative and thought-provoking graphics.

I was able to upload the dataset from Github, (https://github.com/KarenWWells/FinalProject_DA_2022/blob/main/Pocket%20Measurements.csv). The pocket measurements presented are defined as shown at the right.



Data cleaning was required after plotting a pie chart showing the different brands. "Guess" had two entries and "7 For All Mankind" had two entries. I was easily able to find the errors in the raw data file and fix them (one an extra space before the comma and in the other "All" versus "all").

Price in the original data set showed no difference between men's jeans and women's jeans as confirmed with a boxplot and by comparing the price summary statistics for both men (M) and women (W) (mean: M - $81, W- $80). However, there were several high price outliers for all the True Religion Jeans and 7 for All Mankind Jeans (the prices were above $130) which I felt required further examination. After sorting the

data in the .CSV file, I confirmed that these jeans were not associated with any measurement outliers. Therefore, I chose to not eliminate these price outlier data points.

It should be noted that all data were collected on the same size jeans – a waist size of 32 inches. Also, while all data presented in the .CSV file is in centimeters, the data presented in the Pudding article is in inches. This is presumably to work in the metric system for taking the measurements but to use the standard measurement for the United States of inches when presenting the findings. For purposes of this analysis, the original units (centimeters) will be analyzed.

I began my analysis by creating histograms of the various measurements to identify variables of interest. The histograms for maximum height front and maximum width back were clearly bimodal which I assumed to be associated with the differences between the men's jeans and the women's jeans. The shapes of the other graphs had distributions which were more difficult to discern. Price clearly had some outliers and they were examined further as described above.

Boxplots allowed identification of measurement outliers. I specifically created the boxplots on the same scale so the relative differences could be examined. To allow me to further understand the differences, I created tables of summary statistical data for all jeans, then for men's jeans and women's jeans separately. I utilized two different techniques so I could expand my knowledge of dfstatsum and df.describe(). Since back pockets on all jeans have a stylistic look, it was not surprising to me that the variation in maximum and minimum back pocket dimensions between men's and women's jeans showed unremarkable differences. The standard deviations for all back pocket measurements, both men and women, were all less than an inch. This contrasted sharply with the spread in maximum and minimum front pocket heights and widths. The standard deviations of each front pocket dimension overall, as well as the break down for men (M) and women(W) jeans separately, is as follows: minimum width of front pockets 1.05 cm (M – 0.99 cm, W – 0.79 cm,); maximum front pocket width 1.45 cm (M – 1.31 cm, W – 1.39 cm); minimum height of the front pockets 3.50 cm (M – 3.19 cm, W – 1.62); maximum height of the front pockets 4.89 cm (M – 2.06 cm, W – 2.10 cm). To better understand what these measures of spread are telling us, I created a boxplot for each dimension depicting the results for both men and women on the same set of axes using a "for loop" and color coding the results with the traditional colors of blue for men's jeans and pink for women's jeans. The only category where there is no overlap in dimensions between the men's jeans and the women's jeans is the maximum height of the front pockets. This is not visible when looking at a pair of jeans and really begs the question (since they are the same size jeans) why do manufactures feel it is necessary to limit the room provided to women in their front pockets? The minimum height of the front pocket and the width of the front pocket are visible on the outside suggesting stylistic consideration. However, men had a much larger spread in measurements for minimum height of front pockets than for women, suggesting the opening size for a men's pocket can vary much more than for women's pockets. (The remaining plots were created in the python learning process. They can be ignored!)

Although my analysis was not as "dynamic" as the one provided in The Pudding article, my sense is Diehm and Amber went through a similar "old school" analysis first before deciding which variables to focus on and why. The graphs I created coupled with the summary statistics yield valuable information in working through the process of analyzing data and drawing conclusions. I believe a deep understanding of the data is required to create a presentation but the updated dynamic visuals cannot be beat in presenting the results to the general population.