STAT7026 Assignment 1 Report

Rui Qiu

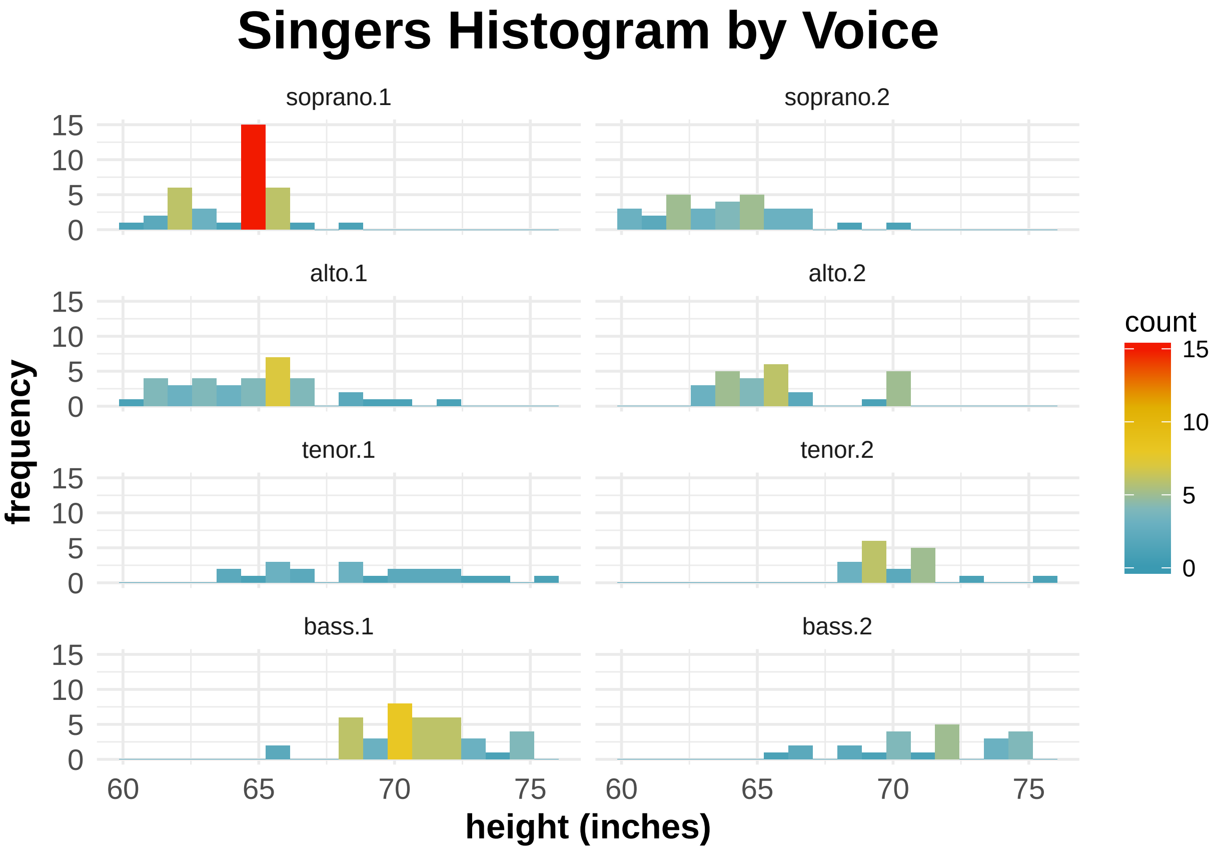
2017-08-18

Background

The heights in inches of the singers in the New York Choral Society in 1979 are contained in a list called singers with components `soprano.1`, `soprano.2`, `alto.1`, `alto.2`, `tenor.1`, `tenor.2`, `bass.1` and `bass.2`. These components are listed in order of decreasing pitch. The first four components are female voices and the last four male. We are interested in the comparison of the height distributions, and what factors appear to be important in describing height. We transformed the ragged list into a 3-column dataframe. Those 3 columns represent the height (in inches) of the singer, the voice/pitch of the singer and the gender of the singer.

Histogram by Voice

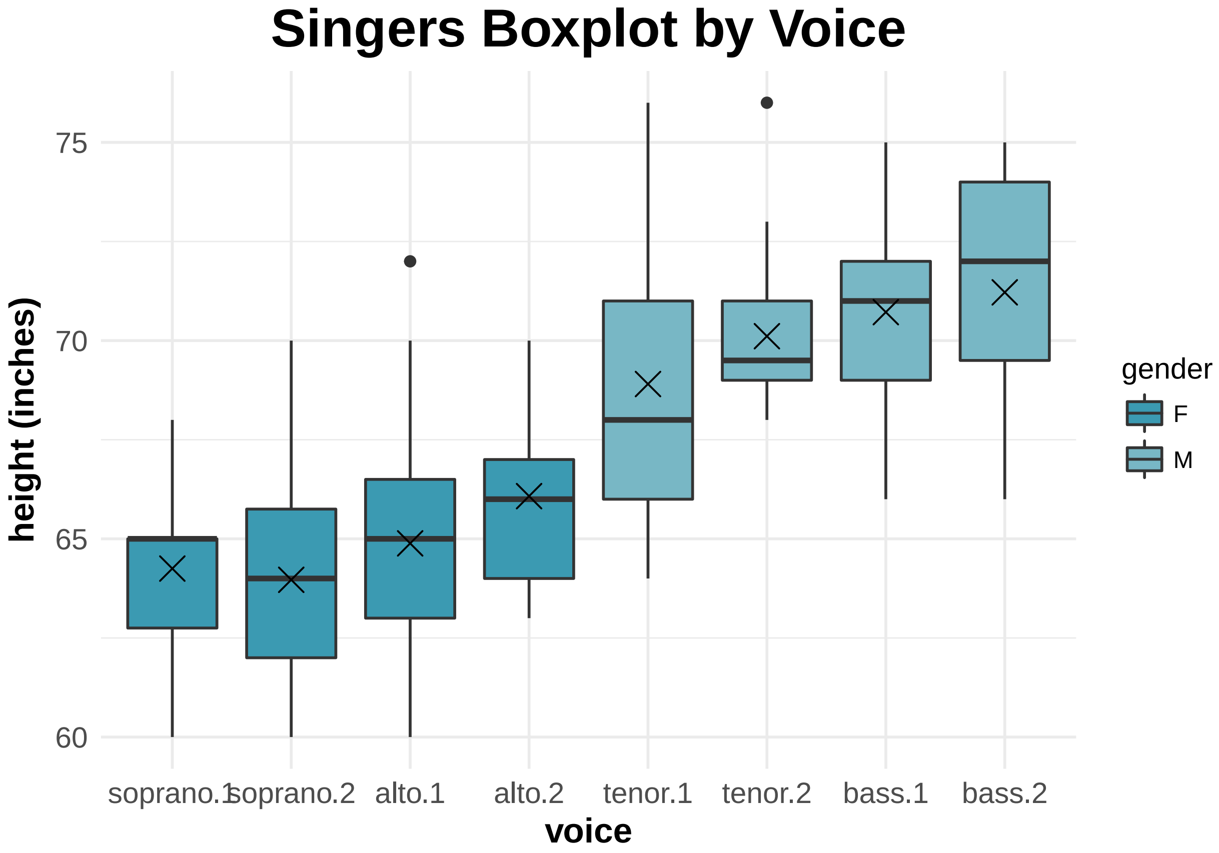
Since our data only contains 2 variables, *height* and *voice* (and a hidden variable *gender*, which will be talked about later), the first idea comes to our mind is to visualize the data in a histogram. A histogram will be obvious to present how many singers have the same height for each voice. While constructing the histgram, we rearrange the data by 8 voices ordered from the highest pitch (`soprano.1`) to the lowest pitch (`bass.2`). The main histogram is divided into a 4-by-2 gird, so that each pair of similar voices, e.g. `soprano.1` and `soprano.2`, will be on the same row. We also include a spectrum legend, using different colors to indicate differrent quantities.



So far, we roughly get images of distributions of those 8 voices. However, we would like to investigate further. Typically, we need more statistical information and side-by-side comparison.

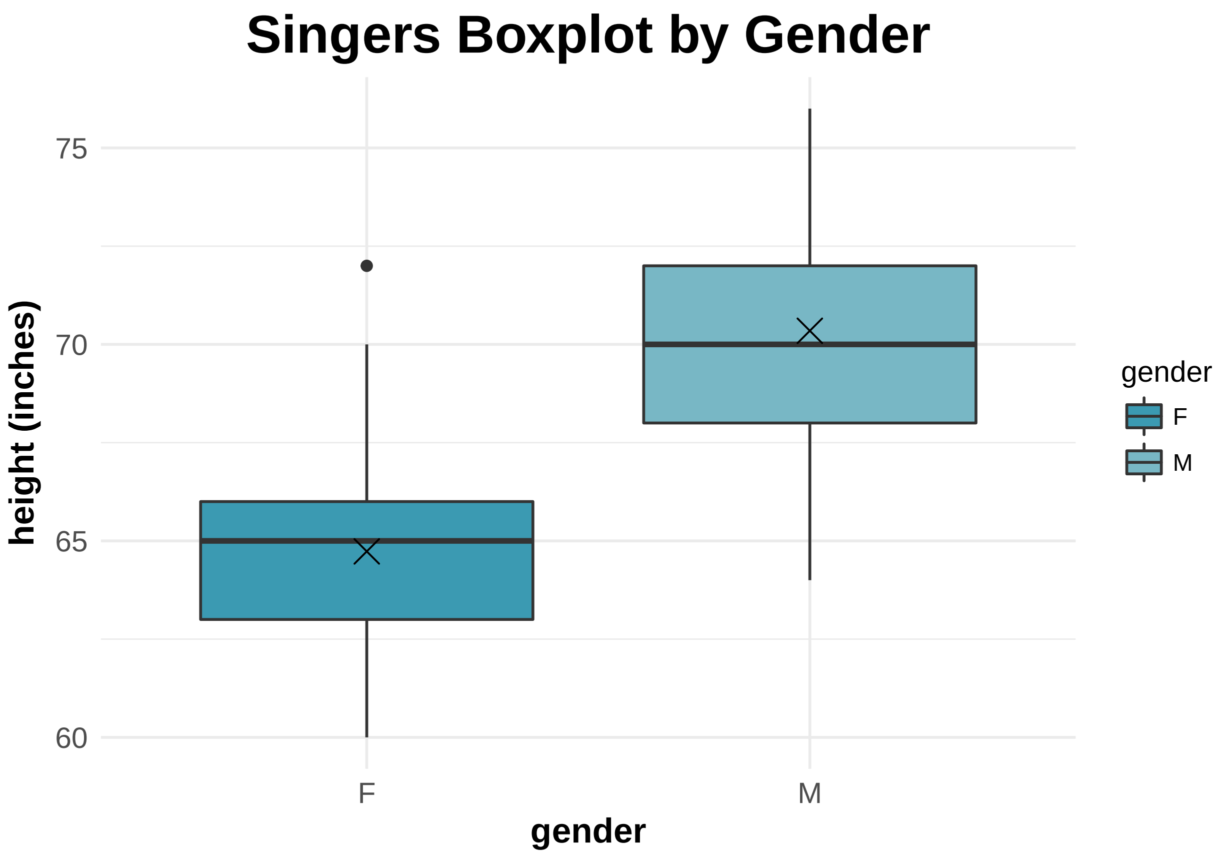
Boxplot by Voice

Therefore, we next plot our data in a boxplot. On x-axis direction, we arrange the 8 voices by decreasing pitch order again. Also, note that the little cross inside each box represents the mean of that group of height.



Boxplot by Gender

Additioanlly, we know that those 8 voices can be divided into two hidden but easily understandable groups, which are male and female. Similarly, we plot the boxplot by gender.



Conclusions

- **Conclusion from the plots:**
  - Generally, singers with higher pitch have relatively smaller height, singers with lower pitch have relatively larger height. We can rephrase this discovery with more statistical terminology, that is "**the pitch of a singer and the height of his/her seems negatively correlated**". However, this is not 100% true, since `soprano.1` has both higher mean and median than `soprano.2` although it represents a higher pitch itself.
  - And **male singers tend to have lower pitch and larger height than female singers**.
  - Simialr groups of singers tend to have similar height distributions. (We don't consider `alto.2` and `tenor.1` as "similar groups" since they are one group of female singers and a one group of male singers.)
- **Intuitive explanation:** One explanation is that we can consider human as a piccolo, the shorter air column it has, the frequency of vibration is higher, so that the pitch is higher. This is why male singers with `bass` voice are always taller than female `soprano`s.
- **Potential improvement:**
  - In visualziation, since we use histograms to "see" what each distribution looks like, and boxplot to check the statistics of distributions, maybe a violin plot that combines the features of these two would be more compact.
  - In investigation, if we are able to quantify the definition of "pitch", instead of just leaving it as a categorical variable, then we can fit in a regression to further study the relationship between variable height and pitch. Because although people within the same voice group are tagged with the same pitch, there might still exist some variations.
  - Due to the limitation of sample size, some patterns within might be still in the mist. A larger sample definitely would help.

References

- R for Data Science by Garrett Grolemund and Hadley Wickham (http://r4ds.had.co.nz/).
- Wes Anderson Palettes (https://github.com/karthik/wesanderson).