Gender Recognition by Voice EDA Report

Jorge Agustín Erosa Herrera

October 16th, 2019

1 Introduction

I decided to analyze the data set called "Gender Recognition by Voice" from *kaggle*, which can be found here. I think it's a really interesting data set to study, because we can see the development of Artificial Intelligence and where to focus our training models based on acoustic properties of the human voice/speech. The data set consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples are pre-processed by acoustic analysis in *R* using the *seewave* and *tuneR* packages, with an analyzed frequency range of 0hz-280hz (human vocal range).
According to the original analysis of the data set, which can be found here, when the samples are applied to an artificial intelligence/machine learning algorithm to learn gender-specific traits, the resulting program achieved a best accuracy of up to 99/100 percent.
This data set gathers the following already calculated features:

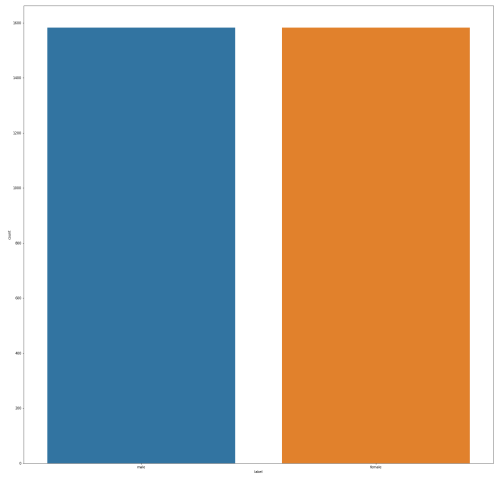
- meanfreq: mean frequency (in kHz)
- sd: standard deviation of frequency
- median: median frequency (in kHz)
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- IQR: interquantile range (in kHz)
- skew: skewness
- kurt: kurtosis
- sp.ent: spectral entropy
- sfm: spectral flatness
- mode: mode frequency
- centroid: frequency centroid

- meanfun: average of fundamental frequency measured across acoustic signal
- minfun: minimum fundamental frequency measured across acoustic signal
- maxfun: maximum fundamental frequency measured across acoustic signal
- meandom: average of dominant frequency measured across acoustic signal
- mindom: minimum of dominant frequency measured across acoustic signal
- maxdom: maximum of dominant frequency measured across acoustic signal
- dfrange: range of dominant frequency measured across acoustic signal
- modindx: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
- label: male or female

In this case we have 20 features and a column specifying if the studied sample corresponds to a male or a female.
The purpose of this Exploratory Data Analysis (EDA) is to study the variables and how they interact with each other and how they affect the end result of a voice analysis. Also to find out which one(s) of these variables/features we can use to rudely infer if the person speaking is female or male.

2 Univariate Analysis

We start our EDA by counting the number of samples and features we have to work with, giving us a result of 3168 samples and 20 features per sample. After that, we have to check for duplicated values, and we find 2 duplicated values, 1 male and 1 female. In this case we eliminate both and we are left with 3166 samples. After that, we count how many male and how many female we have, and we



illustrate it with the following plot:
Here we can see how there's an even sample size between the targets (in this case male and female), which is great, because it makes our analysis more accurate. Then we'll make more graphical approach which will show us a considerable amount of outliers in most of the variables/features. We will also make a non-graphical analysis using the following line of code:

```
data.describe().T
```

```
count       mean          std       min       25%       50%       75%      max
```

And this will show us the number, mean, standard deviation, min, max and quartiles of each variable/feature.
Here we can see how most variables have decimal values except kurt, skew, maxdom and dfrange.
We will also use the following lines of code:

```
data.skew().T
```

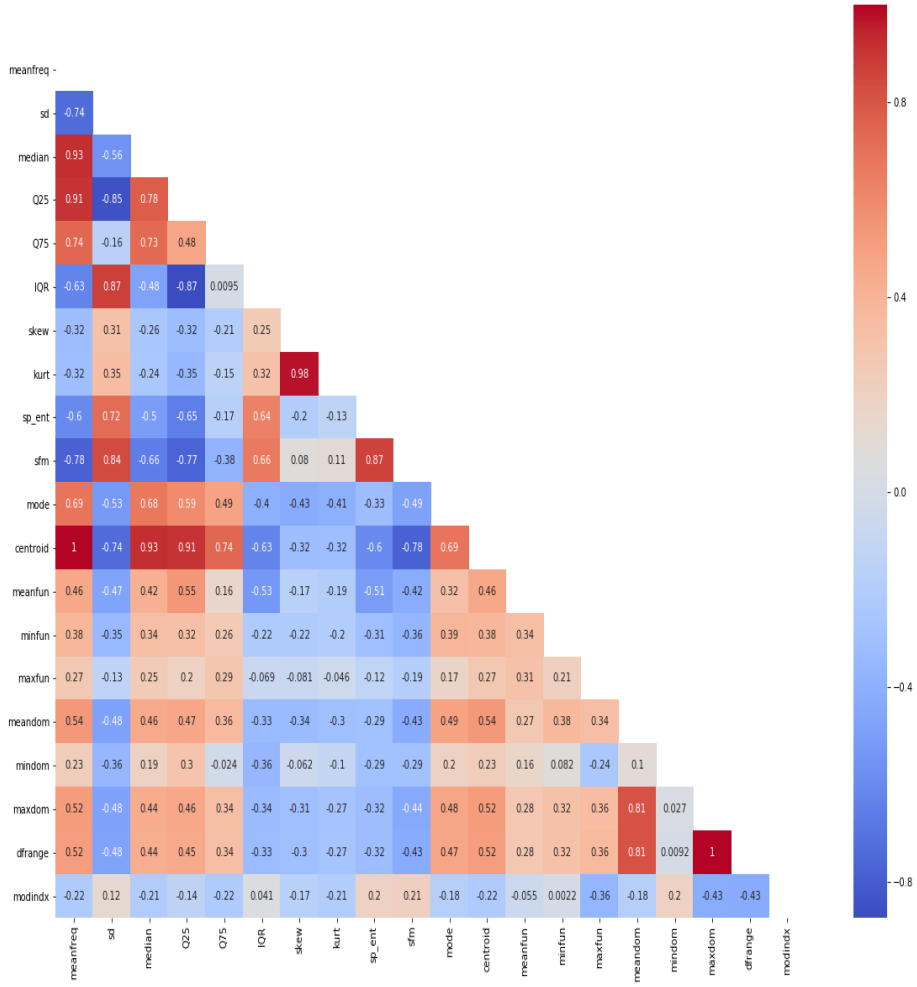
```
data.kurt().T
```

to find out each feature skewness and kurtosis to have an idea of the graphs to follow.

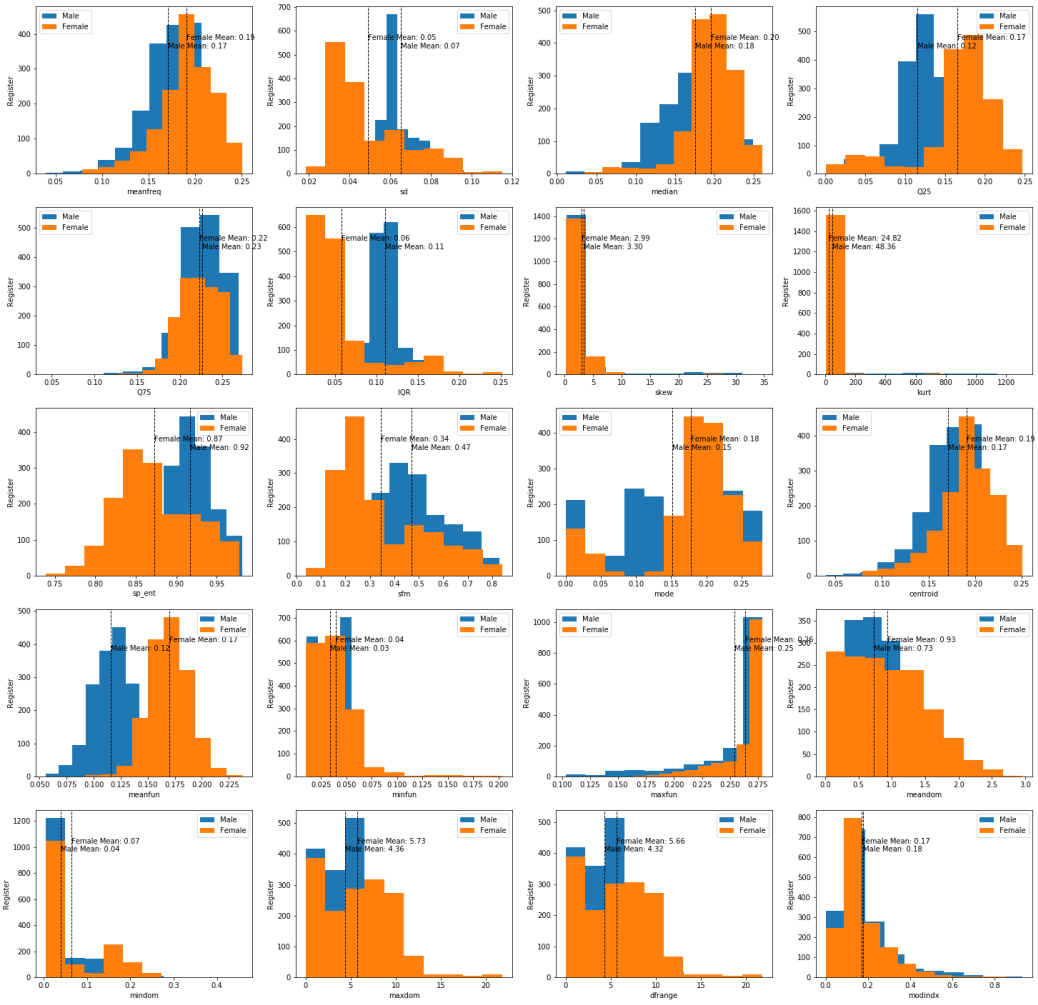
3 Multivariate Analysis

Now we'll start with the Multivariate Analysis which will make more sense, since we have divided data between male and female. First of all we'll plot a matrix to see the correlation between numeric variables. We have the following

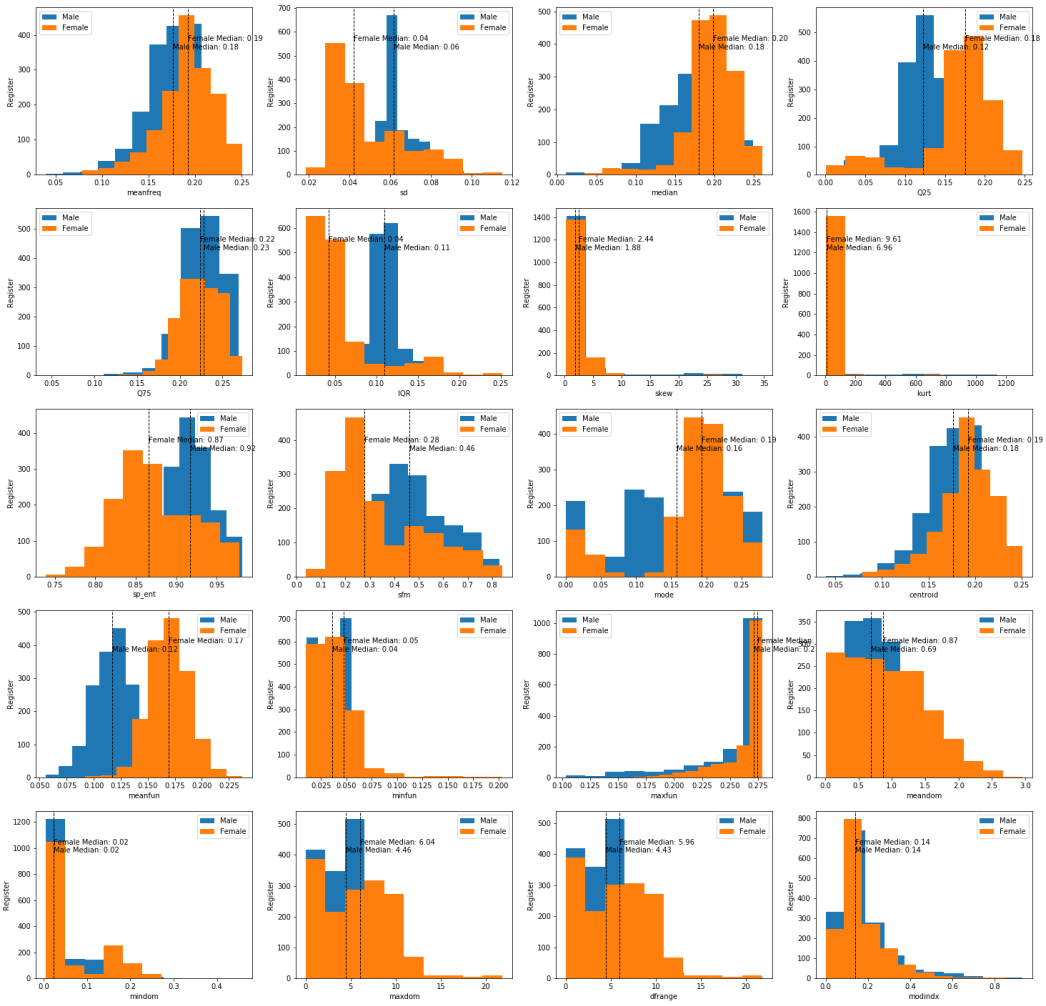
graph:



From this we conclude that the variables centroid and meanfreq as well as dfrange and maxdom behave really similar, which tells us we can discriminate 1 of each to make our data set even simpler. After doing that we'll plot a the histograms of each variable/feature, with different colors depending on their label (male/female) and we'll draw a line exactly where the mean is located to see how it behaves in contrast to the graphic.



Now we'll do the same but showcasing the meadian instead of mean and we'll see it's much more accurate in contrast to the plot.

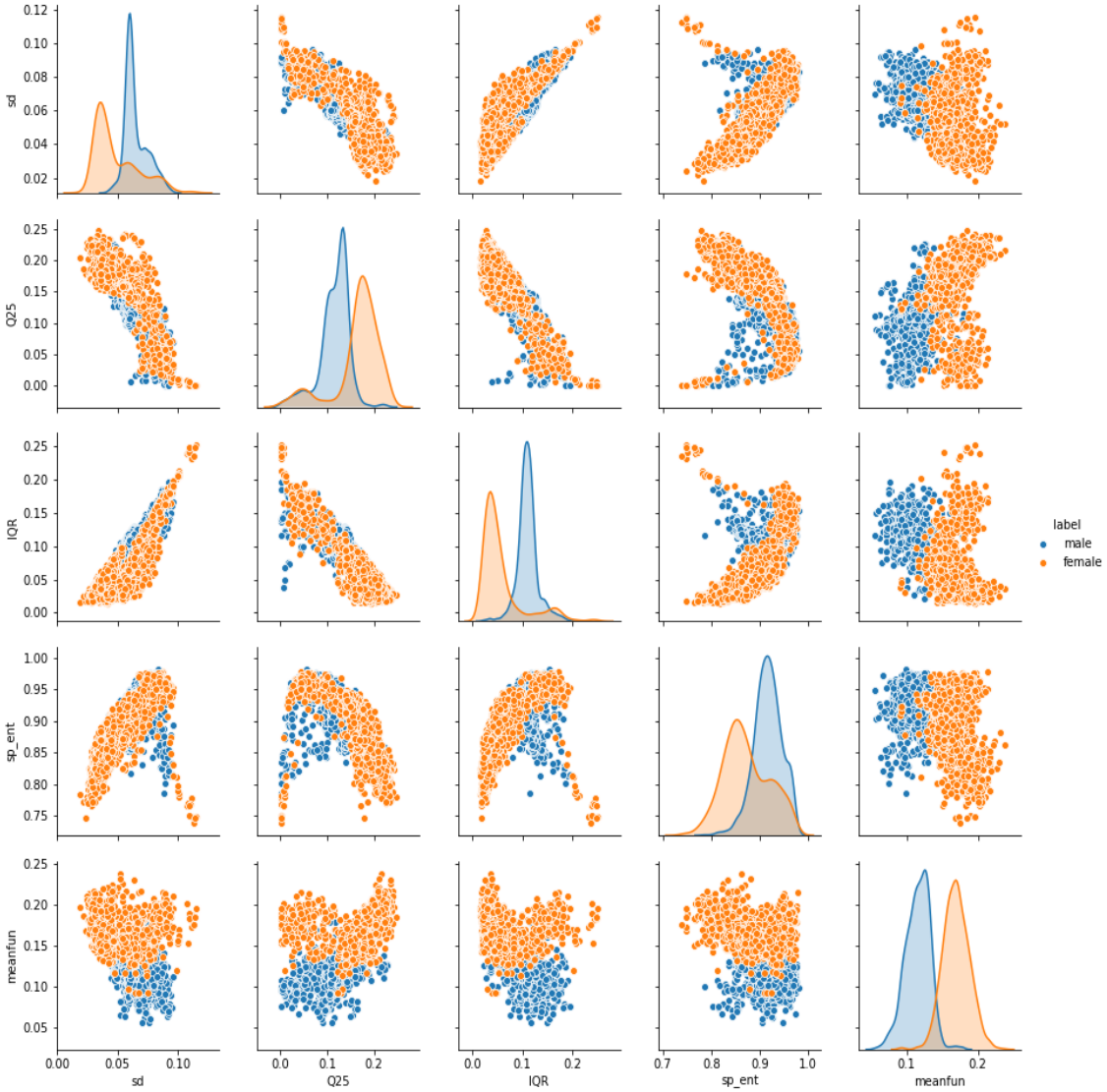


With this information, we can clearly notice how 5 variables/features stand out and clearly establish a difference between each label:

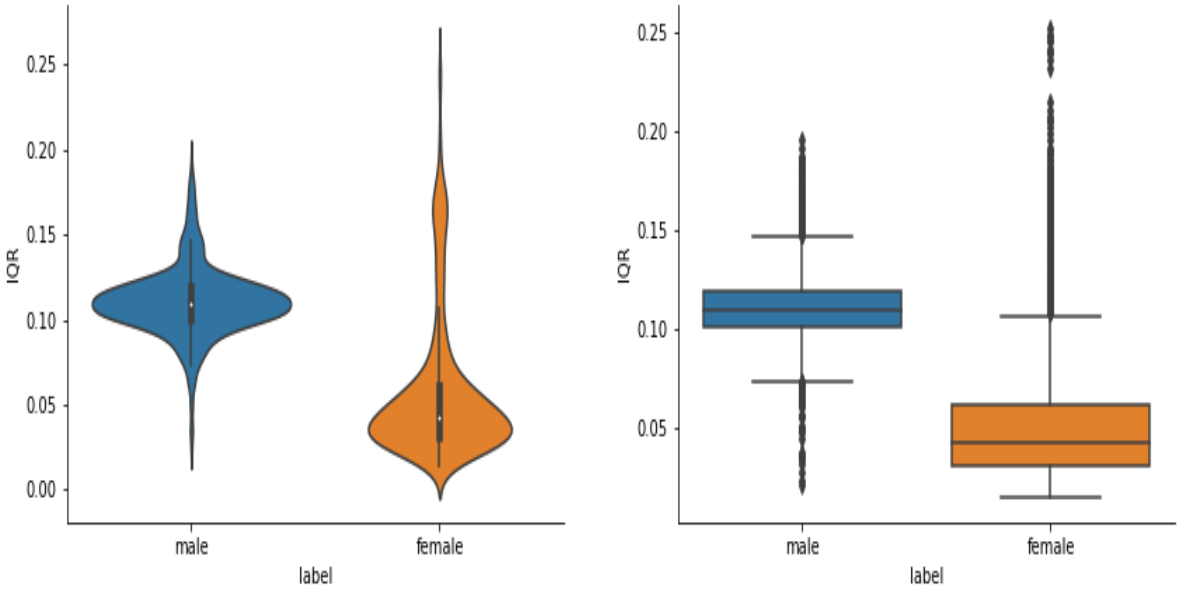
- sd: standard deviation of frequency
- Q25: first quantile (in kHz)

- IQR: interquantile range (in kHz)
- sp ent: spectral entropy
- meanfun: average of fundamental frequency measured across acoustic signal

We will show if there's any relation between these variables using a pairplot:



For these variables we'll do box and violin plots to clearly showcase the difference between one another. The plot will look something like this:



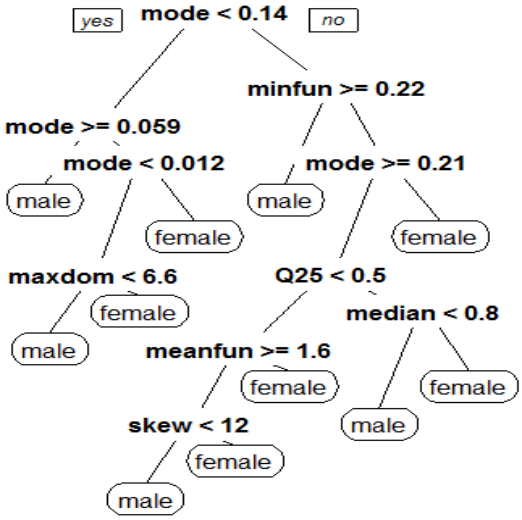
4 Conclusion

From the EDA we can conclude that to quickly make an assumption we can take the variables standard deviation(sd), first quantile(Q25), interquantile range(IQR), spectral entropy(sp ent) and average of fundamental frequency(meanfun) and observe their behaviour. For example females tend to have a larger first quantile (tending to .18) and a larger mean of fundamental frequency (tending to .175) yet tend to have lower numbers on the other 3 variables. Contrary to men who have a small first quantile (tending to .12). We can clearly see the differences aren't much but they're considerable when you see we're talking about small decimal numbers.

We can also see how spectral entropy (complexity, error, order) of a system in this case the acoustic waves, play a major role and as an abstract concept allow us to visualize differences that without this level of abstraction we wouldn't be able to graphically understand the differences and "guess" a person's sex by its voice.

5 Classification And Regression Tree

Still an obscure technique to us, but this is what Kory Becker's trained model does. We can see how the mode plays a major role when training the model, and just like our conclusion states, the first quantile (Q25) and the fundamental frequency mean (meanfun) play a major role between the branches.



6 References

http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-le
https://www.kaggle.com/primaryobjects/voicegender
https://www.mathworks.com/help/signal/ref/pentropy.html#mw_a57f549d-996c-47d9-8d45-e80cb
https://github.com/mrc03/Gender-Recognition-by-Voice-Val.-Acc.-0.
9908-