Main idea: **Why** and **How**, not **What** (your "what" is your code)

1. What is your question?

I want to know if gender can be predicted by speech.

* How are you operationalizing your dependent variable (if any?)

*preparing the variable for the model:*

*Operationalization is the process of strictly defining variables into measurable factors. The process defines fuzzy concepts and allows them to be measured, empirically and quantitatively.*

My dependent variable (y), is the gender of each Simpson’s character. The top 324 characters have their gender classified. This information was contained in a separate data frame that had to be joined to my main data frame, which contained information that will be used as independent variables (x). I used an inner join from Pandas, but first I had to rename the character column within my gender data frame so it matched the column name on my main data frame that I inner joined it on. After the join, I dropped any rows that had a null for the gender.

* What are your potential independent variables, why, and how will you operationalize them?

Starting out I have a column containing the spoken lines by character that has been normalized. I decided that right away I just wanted to get an ‘out of the box’ model ran to get a baseline score and then I could go back and reassess variables, preprocessing, and hyper parameters.

First thing I did to my normalized text column was using tfidf to preprocess the text. Stop words were removed, text was vectorized, and I used ngrams= 1,3.

Depending on my initial models, I may go back and do a little more processing to my x variables. There are a few more columns I could add into my model. Location, for example, is one of the variables that I could go back and vectorize to add to my model.

2) What is your data?

* Where did you find it and how?
* What are the major transformations you made and why?

I did not have to do any major transformations on the main data frame. It was pretty well organized and already contained a normalized text column, which would have been something I would have had to do.

The main transformation that I did was preparing the gender data frame and joining it to my main data frame to make one cohesive set of data to work off of. Once that was completed I was able to drop a lot of columns and really had a clean and organized dataset to work from.

* Did you do any form of dimensionality reduction? Why?

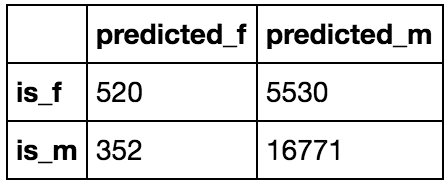
Other than dropping quite a few columns as variables I have not further performed dimensionality reduction. I began modeling off of only one x column (normalized text), so there was no need initially for dimensionality reduction.

Depending on the results of my model I may choose to run KNN or PCA to reduce dimensionality (assuming that I go back and word vectorize the normalized text column).

3) Potential Models?

* Looking at your data and your question, what are the potential techniques you might use? Why?

I started with an ‘out of the box’ Decision Tree Regression (not sure if that was even correct model). It showed over prediction of male characters. Next, I ran a Random Forest Classification model with a few parameters and returned a model score 0.746170 and a confusion matrix showed:



I am now running a GridSearchCV to determine hyper parameters for the Random Forest Classification model.

As you develop your research further, please add in sections for the following two points as you come across them:

4) Results

* For each of your potential models, what did you find? How well does it score on holdout data? If you chose hyperparameters, why did you choose the ones you did / did you use gridsearch and what was the outcome there?

5) Answer your original question!

* What did you find out about the world? Are there models that seem well or not well suited to the task?
* Given more code / more time / other resources, what are changes you would make to your analysis?