



I made a few very simple plots but there is a lot of data cleaning that I need to do to make this dataset more manageable. I am thinking about only including yarns that are still being produced, and also only natural fibers. This would make the data set a lot more manageable. An issue with the data is that the categorical variables have almost 10,000 categories for many of them, which is an issue in predicting at this level of machine learning models at least and for visualizing. I think including the highest rated fiber types however would be really interesting. I think I have too many ideas for the scope of this project and what it actually needs to be. I have started cleaning the data but definitely need to spend a lot more time working on it, in order to do a proper EDA.