DSCI 2 project

When manipulating the different data, I used multiple datasets to effectively apply the methods we have learned in class such as pandas and NumPy to visualize and manipulate different datasets. The Netflix data set includes different movies and shows along with the years they were added to Netflix. The Airbnb data shows different factors such as location of where Airbnb commonly are. The student status set includes the scores of students on a standardized test, whether they studied or not, to which I added a column called total score in which I added their separate scores together and divided them to get their total score. Along with the Spotify data set containing 2 numerical columns called dance ability and popularity which I graphed and performed a simple linear regression with to see if they are correlated. From the EDA that I performed, I was able to make certain conclusions about each data set. For example, in the student's data set after separating them into female and male I observed how the males got a higher math score on average compared to the females. Along with finding that the females scored higher on both reading and writing compared to the males. Overall, by use of bar plots and histograms, I was able to observe how the students who studied tended to score higher than those who did not. As for the Netflix data set, I found that the earliest release date of a movie is from 1942, And that many more movies are released than shows. For the Airbnb data I found that the distribution of Airbnb's is much higher in cities such as Brooklyn and Manhattan than in others.

The simple linear regression model I chose to make was for the Spotify data set. I decided to compare the danceability of a song versus its popularity. To see if there was any correlation I applied a multitude of techniques learned throughout the course and was able to add a trend line to the graph which suggests that dance ability does have a slightly significant correlation to a song's popularity. The slope of the line was not very steep and was also positive.

Some challenges that I faced while performing these functions include having trouble with the string extract or replace functions. It was difficult trying to reach into the data set and manipulate the text strings or values to display them in a different way. Thankfully, there are a number of resources including the textbooks that were very helpful and efficient in helping. Another challenge is that some of the ideas of how to further manipulate the data cannot be carried out because of lack of coding knowledge, but that can easily be fixed by using the readily available sources.