CMSE 201 Final Project

✅ Caitlyn Gerhard

✅ Section_003

✅ 4/25/22

How does a country's human development index, or HDI, correlate with its Netflix subscription fee?

Background and Motivation

The motivation for this project is to see if the cost for a Netflix subscription for a country correlates with its human development index, or HDI, which helps to determine whether a country is developed or developing. I am also interested in seeing if the Netflix library size is dependent upon a country's HDI. My hypothesis is that countries with a higher HDI have a higher cost for a Netflix subscription and also have a larger Netflix library size. First, I had to gather my data from the data sets. I chose to use the Netflix data set from 2021 because it was recent and relevant data and I thought it would be neat to try and find a correlation using this data. I used HDI data from 2019 because it was the most accurate and easily accessible recent data I could find for HDI. I wanted to look into this correlation because after finding the Netflix data, I wondered if developing and developed countries vary in their prices for Netflix subscriptions. Although the data represent different years, they are still comparable because the difference in years would not have varied the prices or HDI values to a significant amount that would make them vastly different and incapable of being correlated.

Methodology

I started by loading in the data that I needed into numpy arrays. From there, I defined a linear equation and had the best fit lines plotted with the data. I then used these lines of best fit to calculate the r squared values for all of the plots. I realized that the HDI to Netflix library size had a significantly lower correlation than the others based on the r squared value, so I did not include that in my final plot. I plotted the other lines all together on one final plot to show how the different Netflix subscription types related to each other and to see how the correlation of HDI to cost increased with the increase of subscription type. I then found the slope of all of the lines of best fit to show a quantitative correlation between the HDI values for the countries and the cost for the three types of Netflix subscriptions.

In [1]:
```
#Loading in all of the necessary modules
import pandas as pd
import numpy as np
from sklearn.metrics import r2_score
```

```
from scipy.optimize import curve_fit
import matplotlib.pyplot as plt
%matplotlib inline

##Netflix data:
#country = np.loadtxt("Netflix_subscription_fee_Dec2021.csv", usecols = [0], unpack=Tru
library = np.loadtxt("Netflix_subscription_fee_Dec2021.csv", usecols = [1], unpack=True
costB = np.loadtxt("Netflix_subscription_fee_Dec2021.csv", usecols = [4], unpack=True,
costS = np.loadtxt("Netflix_subscription_fee_Dec2021.csv", usecols = [5], unpack=True,
costP = np.loadtxt("Netflix_subscription_fee_Dec2021.csv", usecols = [6], unpack=True,
```

In [2]:

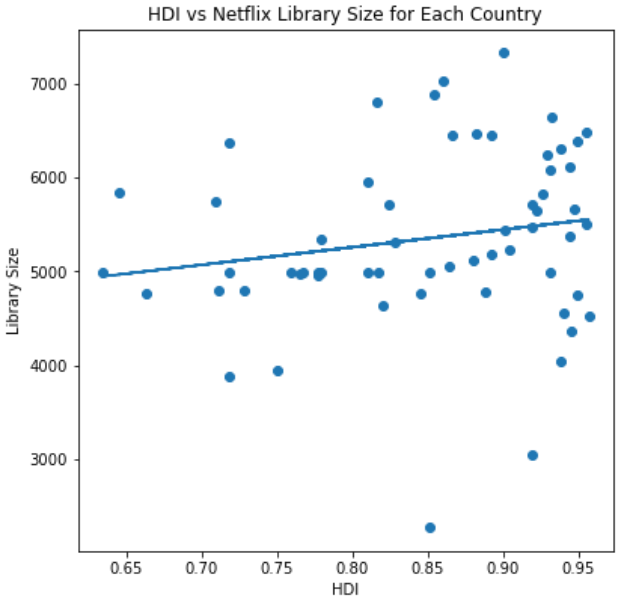
```
##HDI data
HDI = np.loadtxt("HDI_Index.csv", usecols = [1], unpack=True, delimiter=',',skiprows =
```

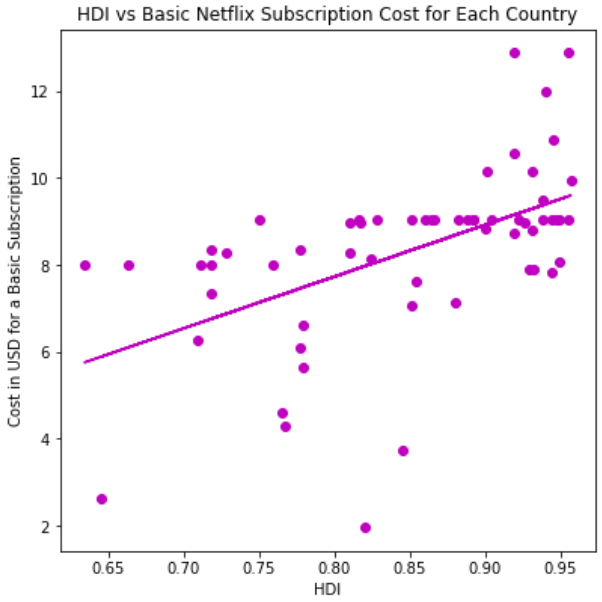
In [3]:

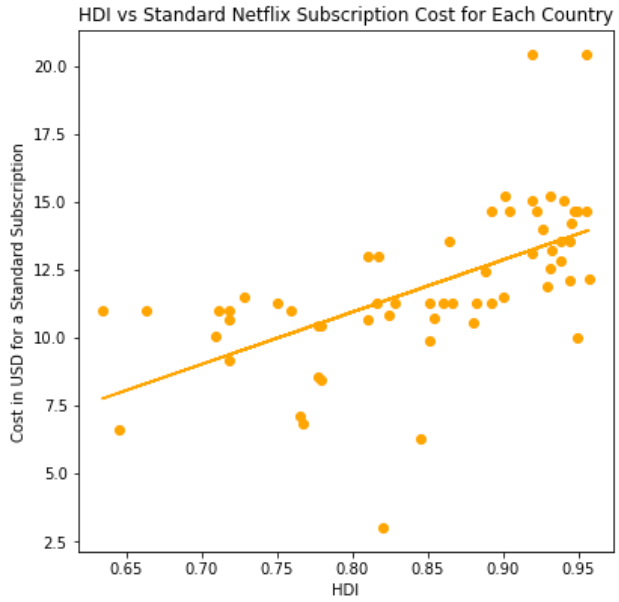
```
#defining a linear function for the lines of best fit to follow for each plot
def lin_fun(x,A,B):
    return(A*x) + B

#creating the line of best fit for my HDI vs library size data
linliba, linlibb = curve_fit(lin_fun,HDI,library)

A = linliba[0]
B = linliba[1]
y = lin_fun(HDI,A,B)

#creating the line of best fit for my HDI vs cost for a basic Netflix subscription data
line, linf = curve_fit(lin_fun,HDI,costB)

A_expected = line[0]
B_expected = line[1]
y_expected = lin_fun(HDI,A_expected,B_expected)

##creating the line of best fit for my HDI vs cost for a standard Netflix subscription
linc, lind = curve_fit(lin_fun,HDI,costS)

A_expect = linc[0]
B_expect = linc[1]
y_expect = lin_fun(HDI,A_expect,B_expect)

##creating the line of best fit for my HDI vs cost for a premium Netflix subscription d
lina, linb = curve_fit(lin_fun,HDI,costP)

A_expec = lina[0]
B_expec = lina[1]
y_expec = lin_fun(HDI,A_expec,B_expec)

#plotting HDI vs Netflix library size with the line of best fit
plt.figure(figsize = [14,14])
plt.subplot(2,2,1)
plt.scatter(HDI,library)
plt.plot(HDI,y)
plt.xlabel("HDI")
plt.ylabel("Library Size")
plt.title("HDI vs Netflix Library Size for Each Country")

#plotting HDI vs cost for a basic Netflix subscription with the line of best fit
plt.subplot(2,2,2)
```

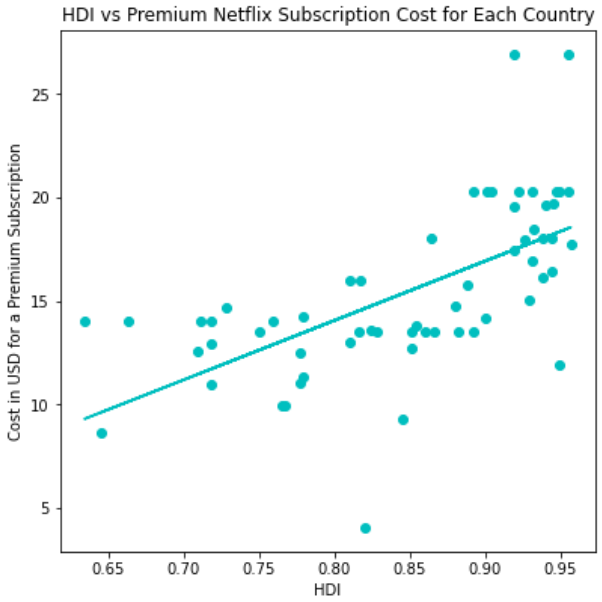
```
plt.scatter(HDI,costB,color = 'm')
plt.plot(HDI,y_expected,color = "m")
plt.xlabel("HDI")
plt.ylabel("Cost in USD for a Basic Subscription")
plt.title("HDI vs Basic Netflix Subscription Cost for Each Country")

##plotting HDI vs cost for a standard Netflix subscription with the line of best fit
plt.subplot(2,2,3)
plt.scatter(HDI,costS,color = 'orange')
plt.plot(HDI,y_expect,color = "orange")
plt.xlabel("HDI")
plt.ylabel("Cost in USD for a Standard Subscription")
plt.title("HDI vs Standard Netflix Subscription Cost for Each Country")

##plotting HDI vs cost for a premium Netflix subscription with the line of best fit
plt.subplot(2,2,4)
plt.scatter(HDI,costP,color = 'c')
plt.plot(HDI,y_expec,color = "c")
plt.xlabel("HDI")
plt.ylabel("Cost in USD for a Premium Subscription")
plt.title("HDI vs Premium Netflix Subscription Cost for Each Country")
```

Out[3]:
```
Text(0.5, 1.0, 'HDI vs Premium Netflix Subscription Cost for Each Country')
```









In [4]:

```
#r squared calculation information found from: https://www.youtube.com/watch?v=J-Us_Ez9

#calculating the r squared value for the basic Netflix subscription data
r_squaredB = r2_score(costB,y_expected)
print("The r squared value for the data in the basic Netflix subscription to its line o

#calculating the r squared value for the standard Netflix subscription data
r_squaredS = r2_score(costS,y_expect)
print("The r squared value for the data in the standard Netflix subscription to its lin

#calculating the r squared value for the premium Netflix subscription data
r_squaredP = r2_score(costP,y_expec)
print("The r squared value for the data in the premium Netflix subscription to its line

#calculating the r squared value for the Netflix library size data
r_squaredlib = r2_score(library,y)
print("The r squared value for the data in the Netflix library size to its line of best
```

```
The r squared value for the data in the basic Netflix subscription to its line of best f
it is 0.28128835182878864
```

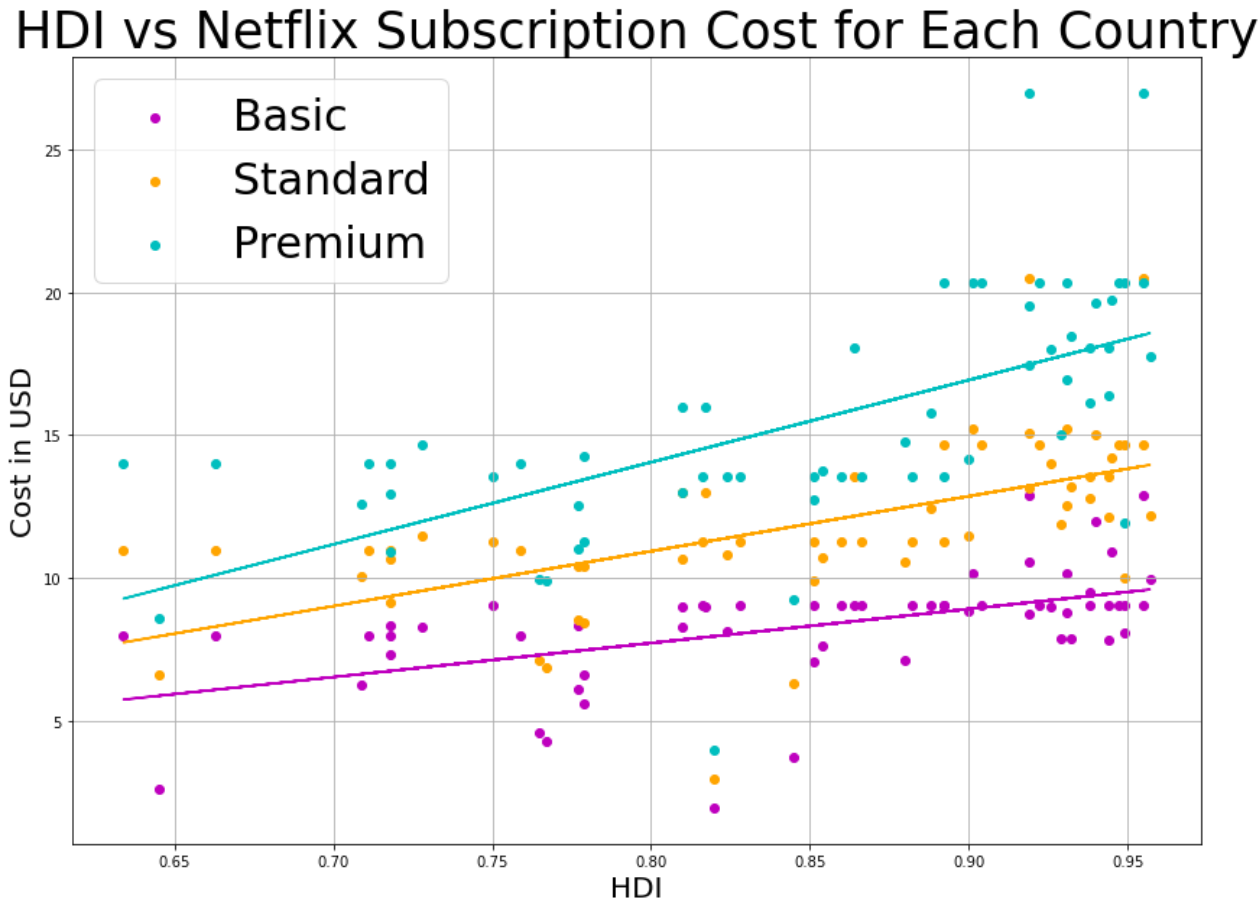
```
The r squared value for the data in the standard Netflix subscription to its line of bes
t fit is 0.34209974181591396
The r squared value for the data in the premium Netflix subscription to its line of best
fit is 0.39194431404379815
The r squared value for the data in the Netflix library size to its line of best fit is
0.032253617143145896
```

Results

In [5]:

```
#creating the final plot with all of the data together on 1 plot
plt.figure(figsize = (14,10))

#plotting HDI vs cost for a basic Netflix subscription with the line of best fit
plt.scatter(HDI,costB,color = 'm',label = 'Basic')
plt.plot(HDI,y_expected,color = "m")

#plotting HDI vs cost for a standard Netflix subscription with the line of best fit
plt.scatter(HDI,costS, color = 'orange',label = 'Standard')
plt.plot(HDI,y_expect,color = "orange")

#plotting HDI vs cost for a premium Netflix subscription with the line of best fit
plt.scatter(HDI,costP,color = 'c',label = 'Premium')
plt.plot(HDI,y_expec,color = "c")

#setting x and y labels, a title, a legend, and adding a grid to the final plot
plt.xlabel("HDI",fontsize = 20)
plt.ylabel("Cost in USD", fontsize = 20)
plt.title("HDI vs Netflix Subscription Cost for Each Country",fontsize = 35)
plt.legend(fontsize = 30)
plt.grid()

#calculating the slopes for all of the lines of best fit so they can be printed below
basicslope = ((y_expected[5] - y_expected[3])/(HDI[5]-HDI[3]))
standslope = ((y_expect[5] - y_expect[3])/(HDI[5]-HDI[3]))
premslope = ((y_expec[5] - y_expec[3])/(HDI[5]-HDI[3]))

print("Basic Netflix Subscription")
print("The slope of the best fit line for a basic subscription is",basicslope)
print("The r squared value for the data in the basic Netflix subscription to its line o
print("Standard Netflix Subscription")
print("The slope of the best fit line for a standard subscription is",standslope)
print("The r squared value for the data in the standard Netflix subscription to its lin
print("Premium Netflix Subscription")
print("The slope of the best fit line for a premium subscription is",premslope)
print("The r squared value for the data in the premium Netflix subscription to its line
```

```
Basic Netflix Subscription
The slope of the best fit line for a basic subscription is 11.904122913510035
The r squared value for the data in the basic Netflix subscription to its line of best f
it is 0.28128835182878864
Standard Netflix Subscription
The slope of the best fit line for a standard subscription is 19.21025853828562
The r squared value for the data in the standard Netflix subscription to its line of bes
t fit is 0.34209974181591396
Premium Netflix Subscription
The slope of the best fit line for a premium subscription is 28.718906331863213
The r squared value for the data in the premium Netflix subscription to its line of best
fit is 0.39194431404379815
```



Discussion and Conclusion

From my results, I learned that there really is no correlation between a country's HDI and its Netflix library size. I determined this based on it's extremely small r squared value. Since the r squared value represents how close the data is to the line of best fit, the Netflix library size having an r squared value of about 0.03 made it clear to me that there wasn't a correlation among those variables. I did, however, find that there was a slight correlation between a country's HDI and its cost for Netflix subscriptions. The r squared value for the basic Netflix subscription was about 0.28. The r squared value for the standard Netflix subscription was about 0.34. The r squared value for the premium Netflix subscription was about 0.39. Based off of these numbers, I can conclude that as the subscription type's price increases, the HDI value for a country and the Netflix subscription cost become increasingly correlated.

Although there are some outliers, countries with a higher HDI tend to have a larger cost for a Netflix subscription. As the HDI of a country increases, the basic Netflix subscription cost for that country increases. Their correlation occurs a rate of about 11.9 USD per HDI value. As the HDI of a country increases, the standard Netflix subscription cost for that country increases. Their correlation occurs a rate of about 19.2 USD per HDI value. As the HDI of a country increases, the premium Netflix subscription cost for that country increases. Their correlation occurs a rate of about 28.7 USD per HDI value. I found these values based on the slopes of the lines of best fit for each of the plots. The slope represents the average rate of change of the data which is why the HDI value and cost can be correlated in such a way.

References

Netflix dataset: *Kanawattanachai, Prasert. "Netflix Subscription Fee in Different Countries." Kaggle, 15 Jan. 2022, https://www.kaggle.com/prasertk/netflix-subscription-price-in-different-countries.*

HDI dataset: "Human Development Reports." | *Human Development Reports*, United Nations Development Programme, https://hdr.undp.org/en/composite/HDI.

R squared tutorial: Stevenson, Robert. "Calculate R squared in Python." *Youtube,* 25 Feb.2021, https://www.youtube.com/watch?v=J-Us_Ez9PDU&t=100s