7PAM2000-0105-2021

Applied Data Science1

Title : Factors affecting the happiness Index

Name : Vijay Kumar Ramavath
STUDENT ID : 21031708

Factors affecting the happiness Index

Abstract

In the following assignment, data concerning the happiness index of various countries are being taken. The scores are calculated by using data from Gallup Poll. There six factors, namely production, freedom of people, generosity of people, corruption, life expectancy and finally social support are being observed in each of the countries, and finally, the relationship has been established among them in this report. It has been accessed that how some of the factors affect the happiness index. For that, Histogram is used to calculate the distribution of countries with respect to their happiness index. The correlation matrix is used to access the relationship between the factors. Finally, a regression model is used to calculate the effect of GDP on happiness. Jupiter notebook of python has been used to visualize the data.

Histogram

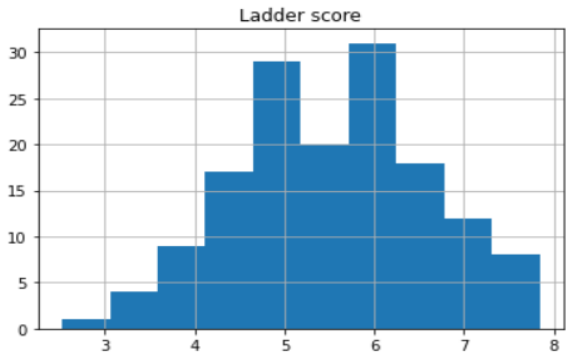
Histogram is used so that it can found out that how many countries belong to the higher happiness index category. From the dataset only ladder score has been taken and normal histogram has been plotted

Code

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sn

In [6]: dframe = pd.read_csv("world-happiness-report-2021.csv")

In [9]: dframe.hist(column='Ladder score')

Out[9]: array([[<AxesSubplot:title={'center':'Ladder score'}>]], dtype=object)
```

Output

It can be observed that in the following histogram there are there are around 31 countries having highest ladder score of 5.8-6.5. Additionally, it can also be seen that the lowest ladder score is 2-3 with only country in it.



Correlation Matrix

A correlation matrix determines the relationship between two variables. It generally ranges from -1 to +1, where any correlation value which is less than zero means that both the variables are negatively correlated. That is, with the increase in the value of one variable, there is a decrease in the value of the other and vice versa.

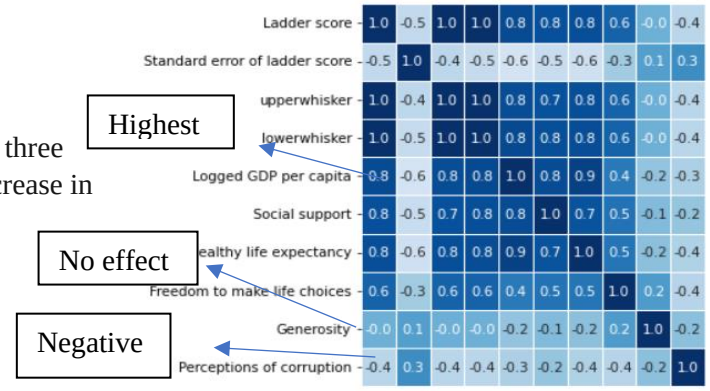
In the assignment, we have checked the relationship between the ladder score and the six factors to see how each of the factors affects the happiness index of the country.

Code

```
In [26]: theta_0 = lr_model.intercept_
         theta_1 = lr_model.coef_
         theta_0, theta_1

Out[26]: (-1.3719060741319824, array([0.73203909]))

In [27]: dframe.columns

Out[27]: Index(['Country name', 'Regional indicator', 'Ladder score',
                'Standard error of ladder score', 'upperwhisker', 'lowerwhisker',
                'Logged GDP per capita', 'Social support', 'Healthy life expectancy',
                'Freedom to make life choices', 'Generosity',
                'Perceptions of corruption', 'Ladder score in Dystopia',
                'Explained by: Log GDP per capita', 'Explained by: Social support',
                'Explained by: Healthy life expectancy',
                'Explained by: Freedom to make life choices',
                'Explained by: Generosity', 'Explained by: Perceptions of corruption',
                'Dystopia + residual'],
               dtype='object')

In [28]: f, axx = plt.subplots(figsize=(10,10))
         sn.heatmap(dframe.iloc[:,1:].corr(), linewidths=0.5, cmap="Blues", annot=True,fmt=".1f", ax=axx)
         plt.show()
```

Solution

In the following solution, it has been observed that The relationship between the ladder score and the first three factors is the highest. This we can say that with the increase in GDP, social support, and healthy life expectancy, the ladder score also increases, whereas generosity has no effect on the ladder score. And also, with the increase in corruption happiness Index reduces.



Scatter diagram and regression line

Finally a scatter plot is used to find the relationship between the GDP of a country and happiness index. Additional a regression line has been drawn and an equation has been calculated to predict the value of the happiness index as a dependent variable and GDP as independent variable.

Code

```
In [11]: data = pd.read_csv('world-happiness-report-2021.csv')
         x = data ['Logged GDP per capita']
         y = data['Ladder score']

In [16]: plt.scatter(x, y)

Out[16]: <matplotlib.collections.PathCollection at 0x1fb39f92580>
```

```
In [23]: from sklearn.linear_model import LinearRegression

In [25]: x = data['Logged GDP per capita'].values.reshape(-1,1)
         y = data['Ladder score']
         lr_model = LinearRegression()
         lr_model.fit(x, y)
         y_pred = lr_model.predict(x)
         plt.scatter(x, y)
         plt.xlabel('Logged GDP per capita')
         plt.ylabel('Ladder score')
         plt.plot(x, y_pred)

Out[25]: [<matplotlib.lines.Line2D at 0x1fb3a77ccd0>]
```

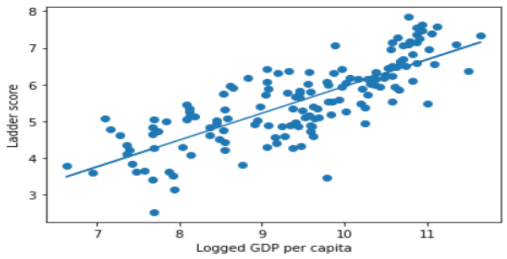
Solution

In the following diagram a positive relationship between the GDP and the happiness score is being observed. The value of θ1 and θ2are -1.37 and 0.73. Through which the values of the happiness index can be calculated.

Regression equation

Y= -1.37X + 0.73

```
Out[25]: [<matplotlib.lines.Line2D at 0x1fb3a77ccd0>]
```




```
In [26]: theta_0 = lr_model.intercept_
         theta_1 = lr_model.coef_
         theta_0, theta_1

Out[26]: (-1.3719060741319824, array([0.73203909]))
```

Conclusion

Thus it can be concluded that the happiness of people around the world varies on various factors ranging from GDP , social support, healthy life expectancy etc. There are various countries that still needs a lot of development to increase the happiness of their people.

References

kaggle.com. (n.d.). World Happiness Report 2021. [online] Available at: https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021.