DAV_Assignment

May 6, 2018

1 DAV Assignment -- Report on World Happiness Data

Exploration of the world happiness report data and analyzing it.
Name : ***Saiteja Talluri***
Roll no : ***160050098***
Department : ***Computer Science***

1.1 Importing Packages, Loading and normalizing data

Importing all the important packages such as pandas, numpy, seaborn, sklearn, matplotlib and plotly.

Extracting the data from the csv file into a dataframe.

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import matplotlib
        import seaborn as sns
        from scipy import stats
        from sklearn.tree import DecisionTreeClassifier
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.cluster import KMeans
        from sklearn.metrics import classification_report, confusion_matrix
        from sklearn.datasets import fetch_20newsgroups_vectorized
        from sklearn.feature_selection import chi2
        from sklearn.feature_selection import RFE
        from sklearn.ensemble import ExtraTreesClassifier
        from sklearn import datasets
        from sklearn import metrics
        import cartopy
        import cartopy.io.shapereader as shpreader
        import cartopy.crs as ccrs
        import types
        from sklearn.manifold import TSNE
        import plotly.plotly as py
        import plotly.graph_objs as go
        from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
        init_notebook_mode(connected=True)
        %matplotlib inline


        import seaborn as sns
        sns.set(style="whitegrid", palette="muted")
        current_palette = sns.color_palette()
        df = pd.read_csv("WorldHappinessIndex.csv")
        df.head()
```

```
Out[1]:        Country          Region  Happiness Rank  Happiness Score  \
        0  Switzerland  Western Europe               1            7.587
        1      Iceland  Western Europe               2            7.561
        2      Denmark  Western Europe               3            7.527
        3       Norway  Western Europe               4            7.522
        4       Canada   North America               5            7.427

           Standard Error  Economy (GDP per Capita)   Family  \
        0         0.03411                   1.39651  1.34951
        1         0.04884                   1.30232  1.40223
        2         0.03328                   1.32548  1.36058
        3         0.03880                   1.45900  1.33095
        4         0.03553                   1.32629  1.32261

           Health (Life Expectancy)  Freedom  Trust (Government Corruption)  \
        0                   0.94143  0.66557                        0.41978
        1                   0.94784  0.62877                        0.14145
        2                   0.87464  0.64938                        0.48357
        3                   0.88521  0.66973                        0.36503
        4                   0.90563  0.63297                        0.32957

           Generosity  Dystopia Residual
        0     0.29678            2.51738
        1     0.43630            2.70201
        2     0.34139            2.49204
        3     0.34699            2.46531
        4     0.45811            2.45176
```

1.2 Initial Data Visualization

1.2.1 World Map of Happiness Score

Pictorially displaying the happiness score distribution across the globe.

```
In [2]: data = dict(type = 'choropleth',
                   locations = df['Country'],
                   locationmode = 'country names',
                   z = df['Happiness Score'],
                   text = df['Country'],
                   colorbar = {'title':'Happiness Score'},)
        layout = dict(title = 'Global Happiness Score',
                     geo = dict(showframe = False,
                               projection = {'type': 'Mercator'}))
        choromap3 = go.Figure(data = [data], layout=layout)
        iplot(choromap3)
```

1.2.2 World Map of Happiness Rank

Pictorially displaying the happiness rank distribution across the globe.

```
In [3]: data = dict(type = 'choropleth',
                   locations = df['Country'],
                   locationmode = 'country names',
                   z = df['Happiness Rank'],
                   text = df['Country'],
                   colorbar = {'title':'Happiness Rank'},)
        layout = dict(title = 'Global Happiness Rank',
                     geo = dict(showframe = False,
                               projection = {'type': 'Mercator'}))
        choromap3 = go.Figure(data = [data], layout=layout)
        iplot(choromap3)
```

1.2.3 Inferences :

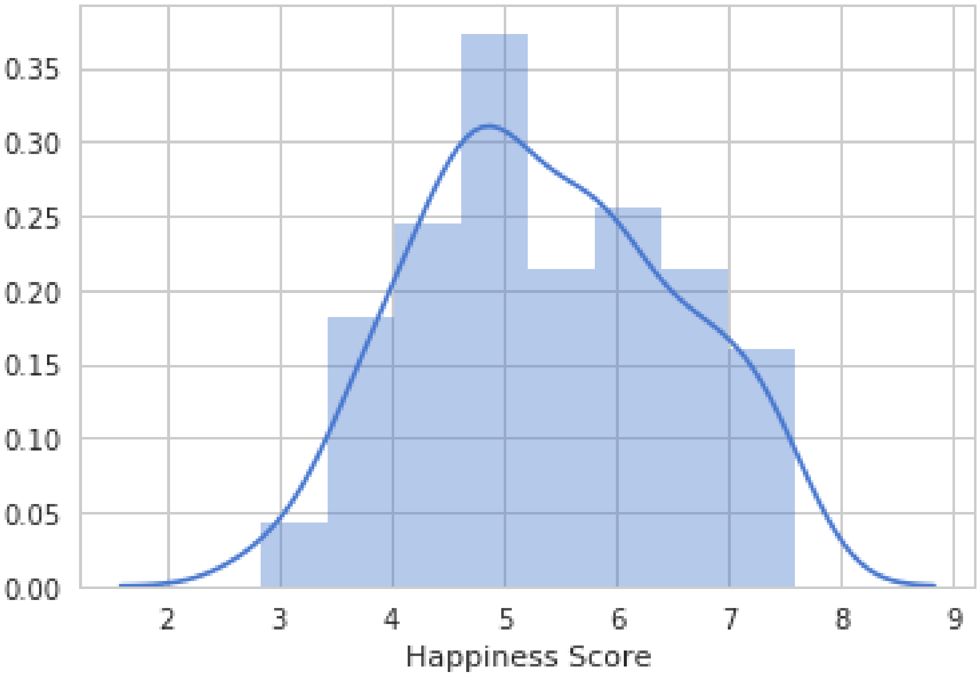
1) All the countries in North America, South America, Australia and Western Europe have very high Happiness Score
2) All the countries in Africa, Eatern Europe and Southern Asia have low Happiness Score
3) All the countries in Northern Asia have moderate Happiness Score

1.2.4 Kernel Density Estimates of Happiness Score and the six factors

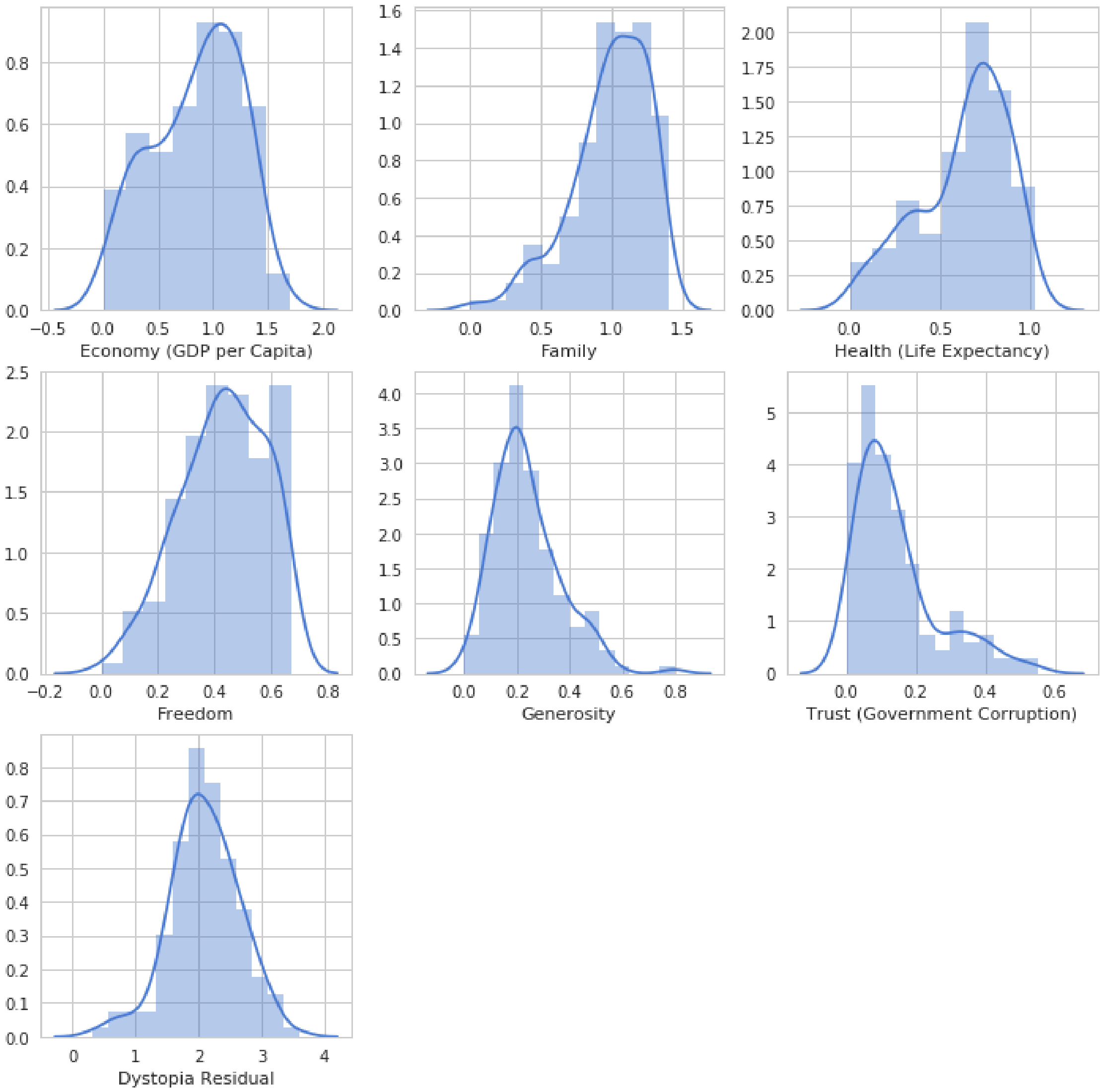
The following is the default plot with a kernel density estimate and histogram with bin size determined automatically. (Y-axis : Density, X-axis : Happiness Score or the 6 factors)

```
In [4]: sns.distplot(df['Happiness Score'])
```

```
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff29893e290>
```



```
In [5]: happiness_factors = ['Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)',
                     'Freedom', 'Generosity', 'Trust (Government Corruption)',
                     'Dystopia Residual']

        def plot_columns_on_grid(data, columns, grid):
            for i, column in enumerate(columns):
                plt.subplot(grid[0], grid[1], i+1)
                sns.distplot(data[column])

        plt.figure(figsize=(12,12))
        plot_columns_on_grid(df, happiness_factors, (3, 3))
```



1.2.5 Inferences :

1) Some of the distributions look like we have at least two distinct groups of countries. For instance the Health data has the majority clustered around 0.7 but also a second group of countries around 0.3.

2) Some of the distributions look like we have only one group of countries. For instance the Dystopia Residual has the majority clustered around 2 and the rest are spread out and didn't form a cluster anywhere.

1.2.6 Linear fitting of the Happiness score in terms of the 6 factors contributing to it

We all know that Happiness Score is calculated from the 6 features and the residual i.e., Economy (GDP per Capita), family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity and Dystopia Residual. We will explore the linear relation using coefficients obtained from Linear Regression by splitting into training and testing data set.

```
In [6]: Y = df['Happiness Score']
        X = df.drop(['Happiness Score', 'Happiness Rank', 'Country', 'Region'], axis=1)
        from sklearn.model_selection import train_test_split
        X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.3, random_stat
        from sklearn.linear_model import LinearRegression
        lm = LinearRegression()
        lm.fit(X_train,Y_train)
```

```
Out[6]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

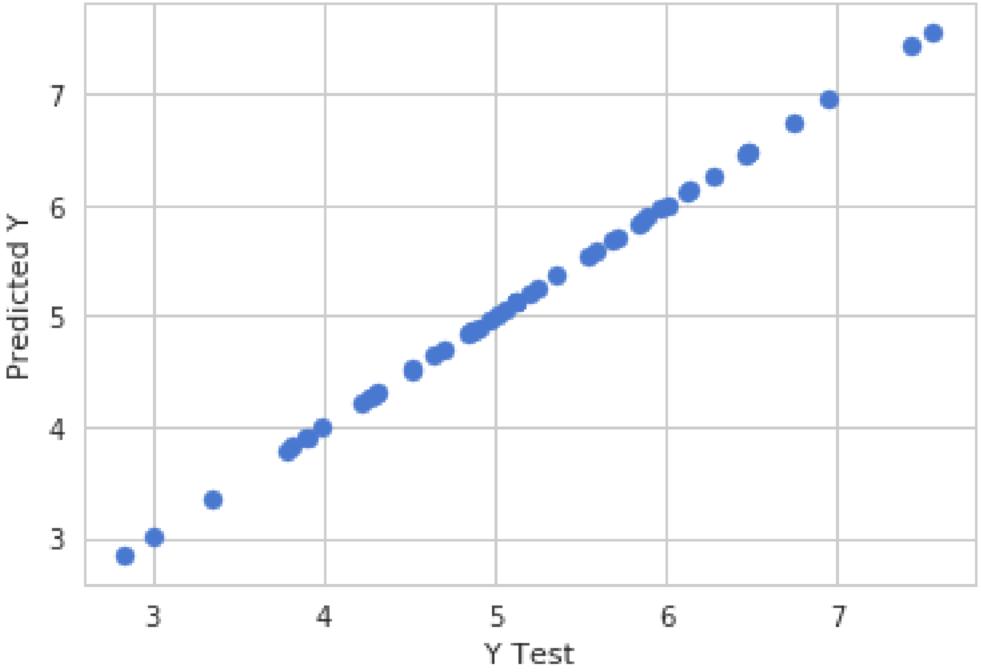
```
In [7]: print('Coefficients:',lm.coef_)
```

```
('Coefficients:', array([ -6.62640689e-04,   1.00012756e+00,   9.99809853e-01,
         9.99984279e-01,   9.99719976e-01,   9.99885249e-01,
         9.99747287e-01,   9.99955045e-01]))
```

```
In [8]: predictions = lm.predict( X_test)
```

```
In [9]: plt.scatter(Y_test,predictions)
        plt.xlabel('Y Test')
        plt.ylabel('Predicted Y')
```

```
Out[9]: <matplotlib.text.Text at 0x7ff29817c610>
```



```
In [10]: from sklearn import metrics

         print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, predictions))
         print('Mean Squared Error:', metrics.mean_squared_error(Y_test, predictions))
         print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test, predicti
```

```
('Mean Absolute Error:', 0.00026861910100900444)
('Mean Squared Error:', 9.548227095661695 6e-08)
('Root Mean Squared Error:', 0.00030900205655726138)
```

```
In [11]: coeffecients = pd.DataFrame(lm.coef_,X.columns)
         coeffecients.columns = ['Coeffecient']
         coeffecients
```

```
Out[11]:                                 Coeffecient
         Standard Error                    -0.000663
         Economy (GDP per Capita)           1.000128
         Family                             0.999810
         Health (Life Expectancy)           0.999984
         Freedom                            0.999720
         Trust (Government Corruption)      0.999885
         Generosity                         0.999747
         Dystopia Residual                  0.999955
```

1.2.7 Inferences :

As expected the happiness score is a perfect linear plot of the factors with the coefficients given in the table above.

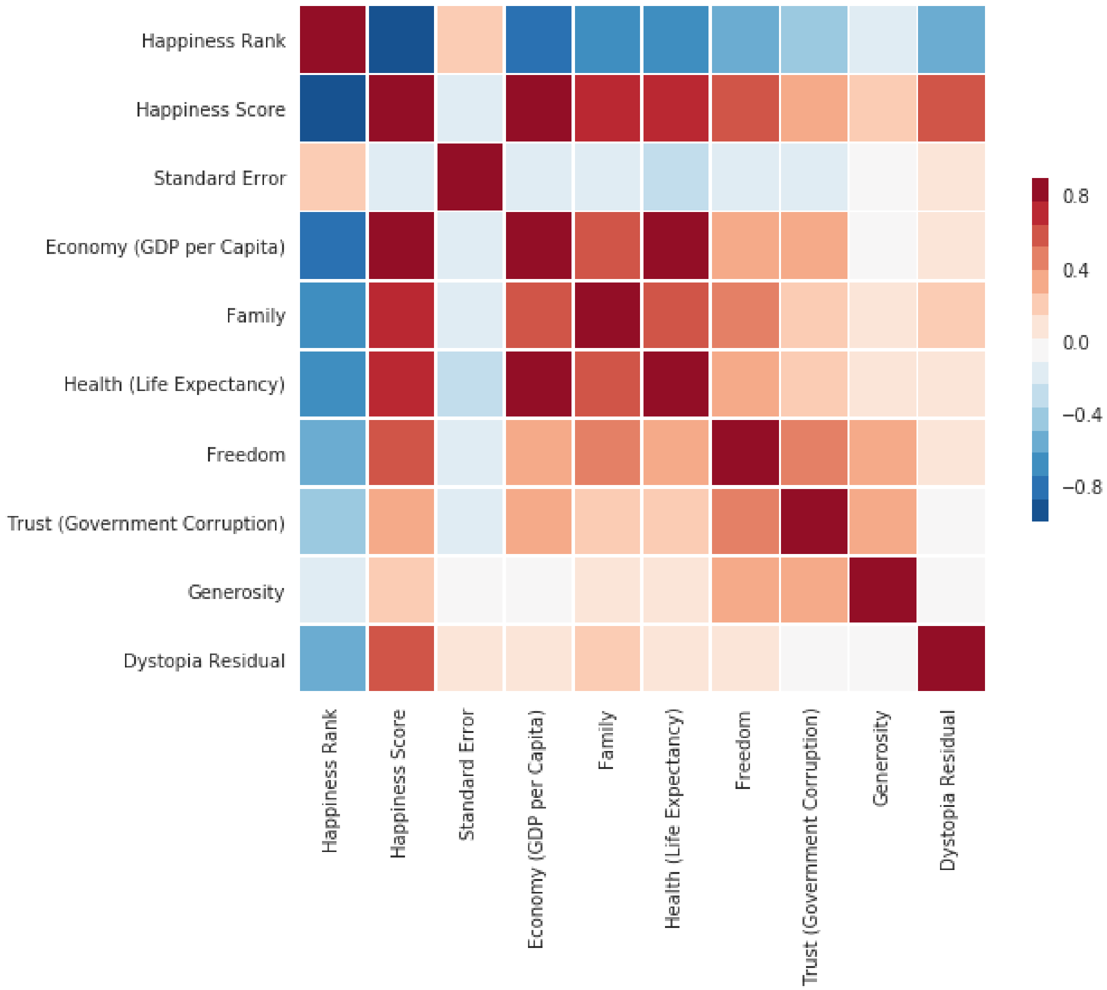
1.3 Factors Contributing to Happiness

1.3.1 Correlation Matrix as a heat map

The following is a heat map describing the correlation matrix (correlation coefficient of the corresponding co-ordinates) in terms of color encoded matrix.

```
In [12]: correlation_mat = df.corr()
         f, ax = plt.subplots(figsize=(9, 7))
         sns.heatmap(correlation_mat, vmax=.9,cmap=sns.color_palette("RdBu_r", 15),
                     square=True,linewidths=.5, cbar_kws={"shrink": .5})
```

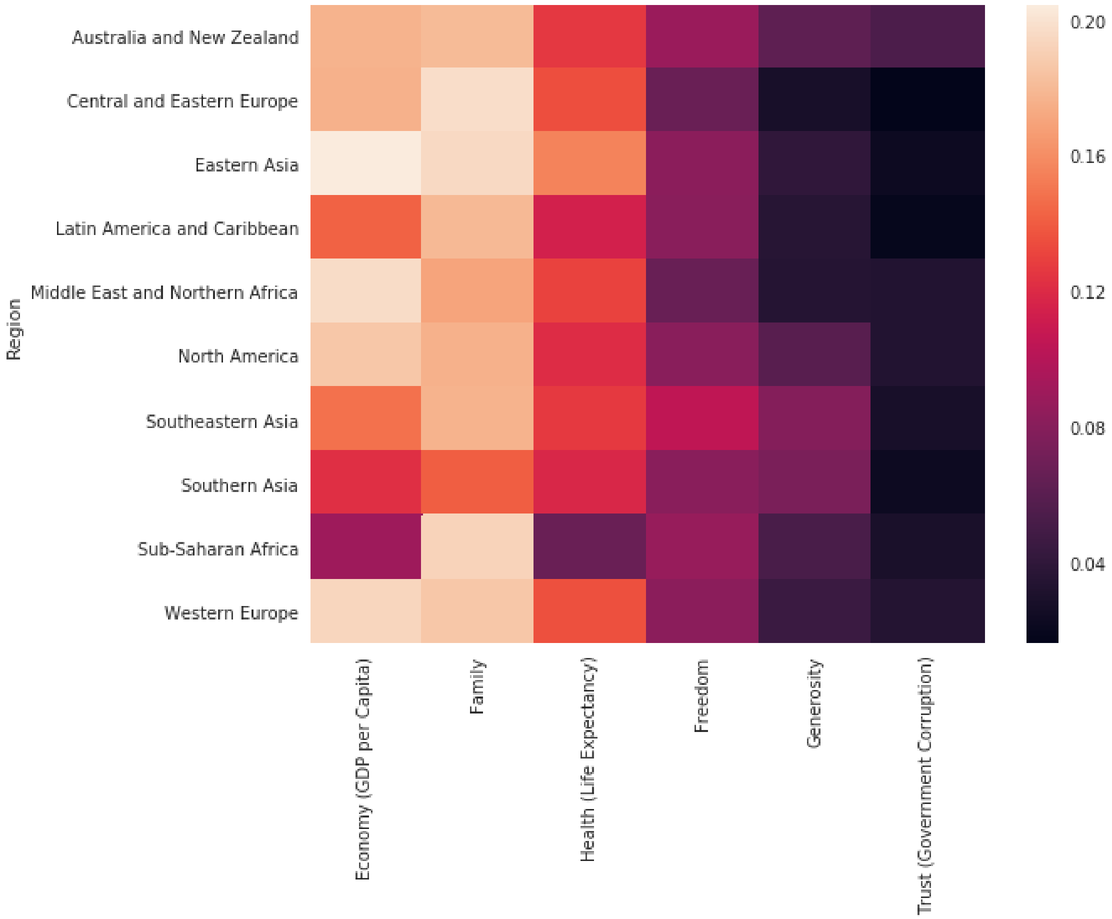
```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff298167f90>
```



1.3.2 Regional Influence of factors as a heat map

Influence of the 6 factors Economy, Family, etc. on happiness depending on regions. Nomalize the factors to the total happiness score.

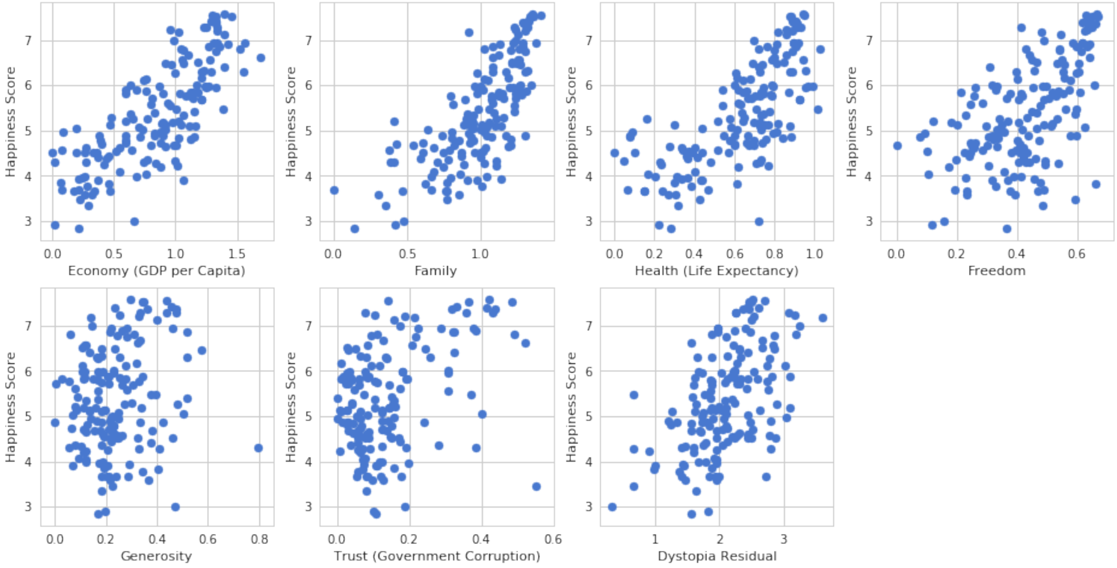
```
In [13]: by_region = df.groupby('Region')

In [14]: f, ax = plt.subplots(figsize=(9, 7))
         sns.heatmap(by_region[happiness_factors[:-1]].mean().div(by_region['Happiness Score']

Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff2985dc310>
```



1.3.3 Scatter Plot of Happiness Score Vs Factors Corresponding to it

Influence of the 6 factors Economy, Family, etc. on happiness can be visualised as a scatter plot so as to get how they are correlated.

```
In [15]: happiness_factors = ['Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)'
                      'Freedom', 'Generosity', 'Trust (Government Corruption)',
                      'Dystopia Residual']

         def plot_columns_on_grid(data, columns, grid):
             for i, column in enumerate(columns):
                 plt.subplot(grid[0], grid[1], i+1)
                 plt.scatter(data[column],df['Happiness Score'])
                 plt.xlabel(column);
                 plt.ylabel('Happiness Score');

         plt.figure(figsize=(16,8))
         plot_columns_on_grid(df, happiness_factors, (2, 4))
```



1.3.4 Inferences :

1)The economy and family are by far the most important contributors to the total happiness score. Generosity and Trust are the least important factors. Freedom and Life Expectancy are moderate factors

2) Order of dependency of Happiness Score : Economy >= Family >> Health > Freedom >> Generosity > Trust.

1.4 Happiness by region

1.4.1 Tabular representation of Happiness Score Vs Region

The following is a tabular representation of mean happiness factors and happiness score of the regions across the globe.

```
In [16]: by_region[['Happiness Score'] + happiness_factors].mean().sort_values(by='Happiness S
```

```
Out[16]:                                   Happiness Score  Economy (GDP per Capita)  \
         Region
         Australia and New Zealand                7.285000                  1.291880
         North America                            7.273000                  1.360400
         Western Europe                           6.689619                  1.298596
         Latin America and Caribbean              6.144682                  0.876815
         Eastern Asia                             5.626167                  1.151780
         Middle East and Northern Africa          5.406900                  1.066973
         Central and Eastern Europe               5.332931                  0.942438
         Southeastern Asia                        5.317444                  0.789054
         Southern Asia                            4.580857                  0.560486
         Sub-Saharan Africa                       4.202800                  0.380473
```

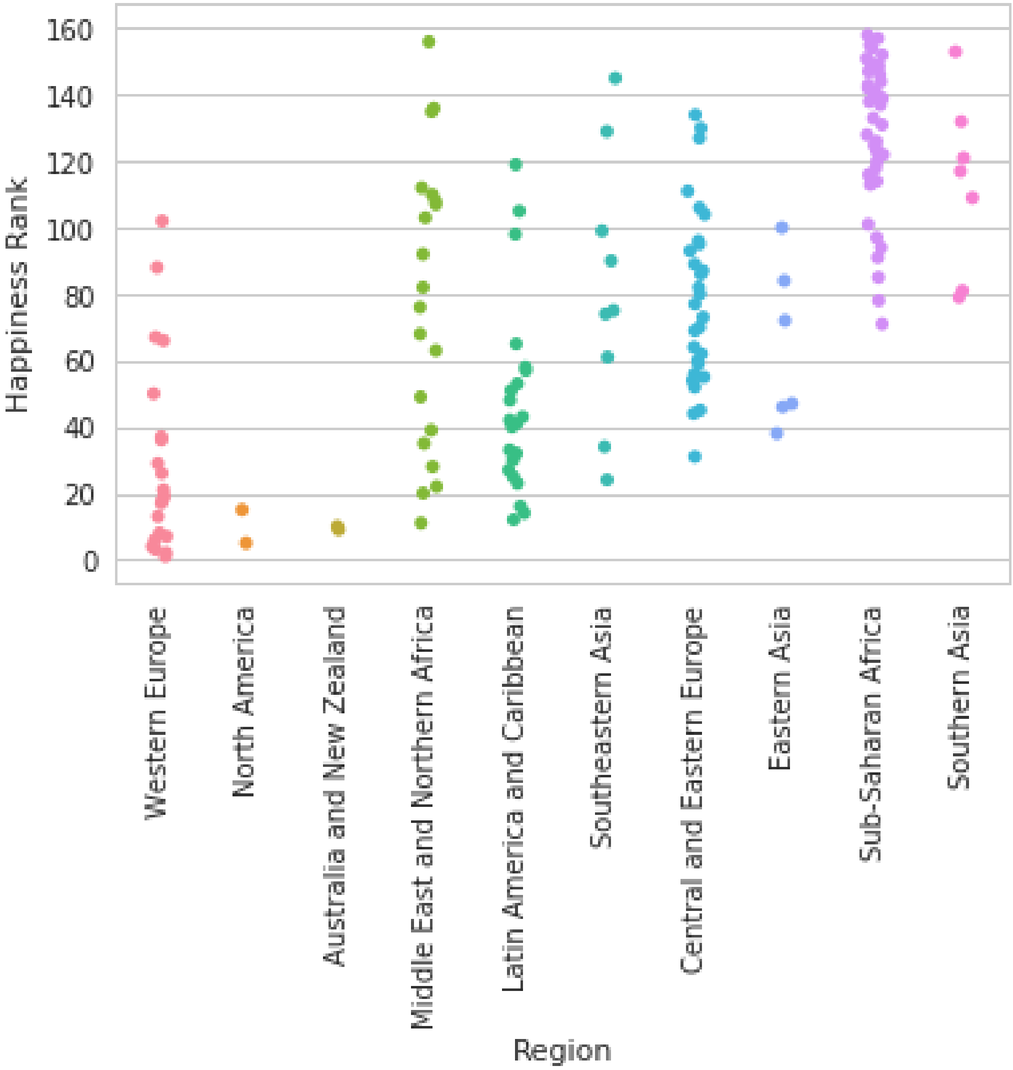
```
                                   Family  Health (Life Expectancy)   Freedom  \
Region
Australia and New Zealand        1.314450                  0.919965  0.645310
North America                    1.284860                  0.883710  0.589505
Western Europe                   1.247302                  0.909148  0.549926
Latin America and Caribbean      1.104720                  0.703870  0.501740
Eastern Asia                     1.099427                  0.877388  0.462490
Middle East and Northern Africa  0.920490                  0.705616  0.361751
Central and Eastern Europe       1.053042                  0.718774  0.358269
Southeastern Asia                0.940468                  0.677357  0.557104
Southern Asia                    0.645321                  0.540830  0.373337
Sub-Saharan Africa               0.809085                  0.282332  0.365944

                                 Generosity  Trust (Government Corruption)  \
Region
Australia and New Zealand          0.455315                       0.392795
North America                      0.429580                       0.244235
Western Europe                     0.302109                       0.231463
Latin America and Caribbean        0.217788                       0.117172
Eastern Asia                       0.225885                       0.127695
Middle East and Northern Africa    0.190375                       0.181702
Central and Eastern Europe         0.152264                       0.086674
Southeastern Asia                  0.419261                       0.151276
Southern Asia                      0.341429                       0.102536
Sub-Saharan Africa                 0.221137                       0.123878

                                 Dystopia Residual
Region
Australia and New Zealand                 2.265355
North America                             2.480935
Western Europe                            2.151185
Latin America and Caribbean               2.622577
Eastern Asia                              1.681607
Middle East and Northern Africa           1.980009
Central and Eastern Europe                2.021400
Southeastern Asia                         1.783020
Southern Asia                             2.016769
Sub-Saharan Africa                        2.019980
```

1.4.2 Strip plot of Happiness Rank Vs Region

The following is a strip plot with the regions on X-axis and their happiness score jittered on the Y-axis. It can be used to get an estimate of happiness scores in the regions.

```
In [17]: g = sns.stripplot(x = "Region", y = "Happiness Rank", data = df, jitter = True)
         plt.xticks(rotation = 90)
```

```
Out[17]: (array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]), <a list of 10 Text xticklabel objects>)
```



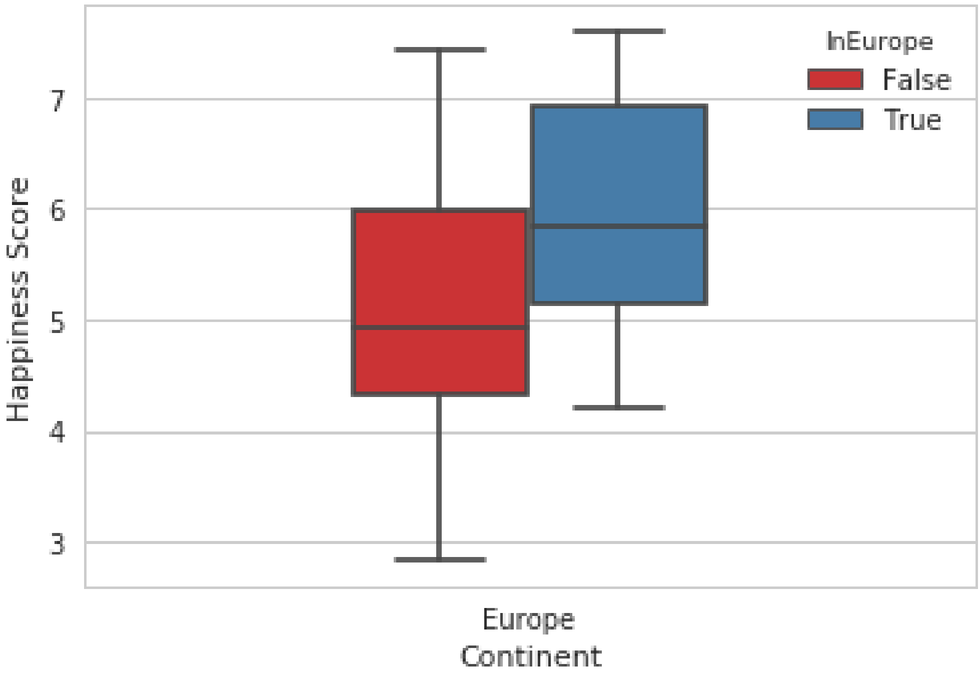
1.4.3 Box Plot of Happiness Rank of Europe Vs Non Europe

The following is a box plot to compare Happiness Score ranges and distribution of European Vs Non European countries and can be extended to other continents as well.

```
In [18]: europe=['Switzerland','Iceland','Denmark','Norway','Finland','Netherlands',
                 'Sweden','Austria','Luxembourg','Ireland','Belgium','United Kingdom','Germany
                 'France','Czech Republic','Spain','Malta','Slovakia','Italy','Moldova','Sloven
                 'Lithuania','Belarus','Poland','Croatia','Russia','North Cyprus','Cyprus','Kos
                 'Turkey','Montenegro','Romania','Serbia','Portugal','Latvia','Macedonia','Alba
                 'Bosnia and Herzegovina','Greece','Hungary','Ukraine','Bulgaria']
```

```
In [19]: df['InEurope']=(df['Country'].isin(europe))
         df['Continent']= 'Europe'
```

```
sns.boxplot(x='Continent',y='Happiness Score',hue='InEurope',width = 0.4,
            data=pd.concat([df[['Continent','Happiness Score','InEurope']]]),palette=
```



1.4.4 Inferences :

1) Australia and New Zealand is the region with the most happy people, closely followed by North America.

2) The least happy people are living in Sub-Saharan Africa followed by Southern & South-eastern Asia.

1.5 Clustering Analysis

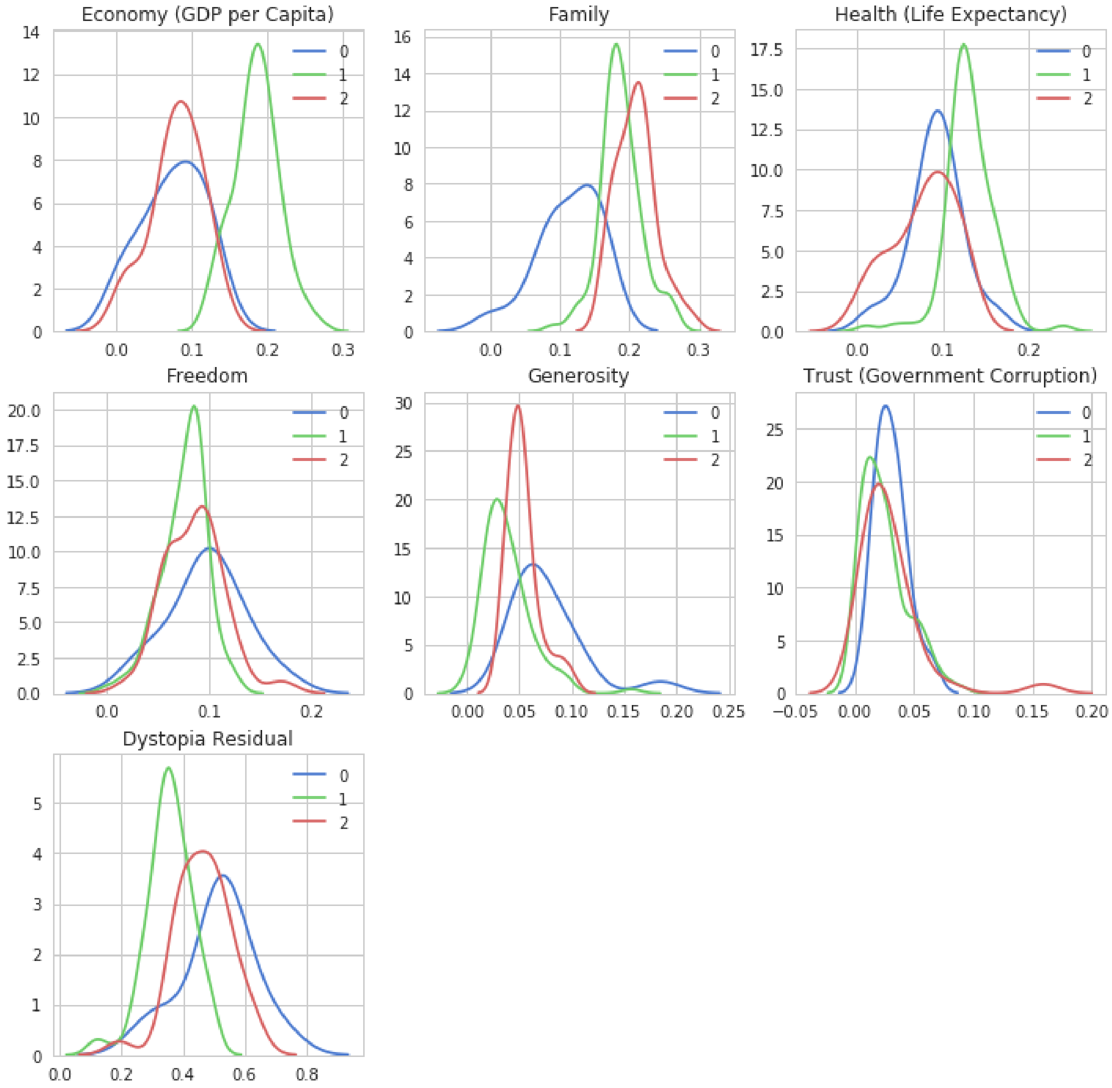
1.5.1 Using K-mean clustering

In the original data, the happiness factors such as Economy, Family, etc. sum up to the happiness Score. Consequently, a country with high happiness score also tend to have high factors. To analyze how the influence of economy on happiness varies between countries, we first normalize the factors using the total happiness score.

```
In [20]: df_norm = df
         df_norm[happiness_factors] = df_norm[happiness_factors].div(df['Happiness Score'].valu

In [21]: cluster_n = 3
         k_means = KMeans(init='k-means++', n_clusters=cluster_n, n_init=10)
         cluster_labels = k_means.fit_predict(df_norm[happiness_factors[:-1]])
```

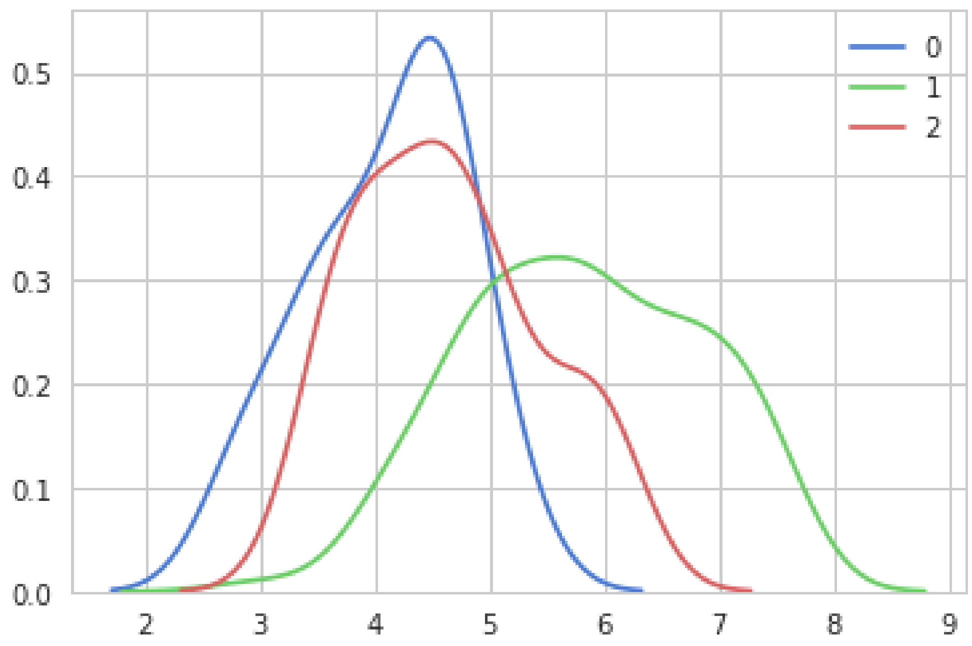
1.5.2 Plotting distributions of the factors for each cluster:

```
In [22]: plt.figure(figsize=(12,12))
         for i, factor in enumerate(happiness_factors):
             ax = plt.subplot(3, 3, i+1)
             for cluster in range(cluster_n):
                 sns.kdeplot(df_norm.loc[cluster_labels == cluster, factor], label=cluster)
                 ax.set_title(factor)
```



1.5.3 Comparing the happiness score distribution for the clusters:

```
In [23]: for cluster in range(cluster_n):
             sns.kdeplot(df.loc[cluster_labels == cluster, 'Happiness Score'], label=cluster)
```



1.5.4 Inferences :

1) There is a big difference between the happiness score distributions of the clusters
2) It can be plotted on globe to get more information about the clusters .