ClusterProject

September 21, 2023

[1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:

```
data = pd.read_csv("happyscore_income.csv")
data
```

[2]:

```
          country  adjusted_satisfaction  avg_satisfaction  std_satisfaction  \
0         Armenia                   37.0               4.9              2.42
1          Angola                   26.0               4.3              3.19
2       Argentina                   60.0               7.1              1.91
3         Austria                   59.0               7.2              2.11
4       Australia                   65.0               7.6              1.80
..            …                    …               …               …
106       Uruguay                   58.0               7.0              2.13
107       Vietnam                   53.0               6.1              1.77
108  South Africa                   49.0               6.3              2.47
109        Zambia                   37.0               5.0              2.61
110      Zimbabwe                   33.0               4.1              2.16

     avg_income  median_income  income_inequality  \
0       2096.76    1731.506667          31.445556
1       1448.88    1044.240000          42.720000
2       7101.12    5109.400000          45.475556
3      19457.04   16879.620000          30.296250
4      19917.00   15846.060000          35.285000
..          …            …                …
106     7544.40    5269.226667          45.014444
107     2231.40    1643.580000          39.242500
108     3889.32    1506.400000          63.726667
109      956.76     510.060000          55.120000
110     1768.56    1230.600000          43.150000

                            region  happyScore      GDP     country.1
0     'Central and Eastern Europe'       4.350  0.76821       Armenia
1             'Sub-Saharan Africa'       4.033  0.75778        Angola
2    'Latin America and Caribbean'       6.574  1.05351     Argentina
```

```
3                 'Western Europe'       7.200  1.33723       Austria
4      'Australia and New Zealand'       7.284  1.33358     Australia
..                              …         …      …           …
106  'Latin America and Caribbean'       6.485  1.06166       Uruguay
107            'Southeastern Asia'       5.360  0.63216       Vietnam
108           'Sub-Saharan Africa'       4.642  0.92049  South Africa
109           'Sub-Saharan Africa'       5.129  0.47038        Zambia
110           'Sub-Saharan Africa'       4.610  0.27100      Zimbabwe

[111 rows x 11 columns]
```

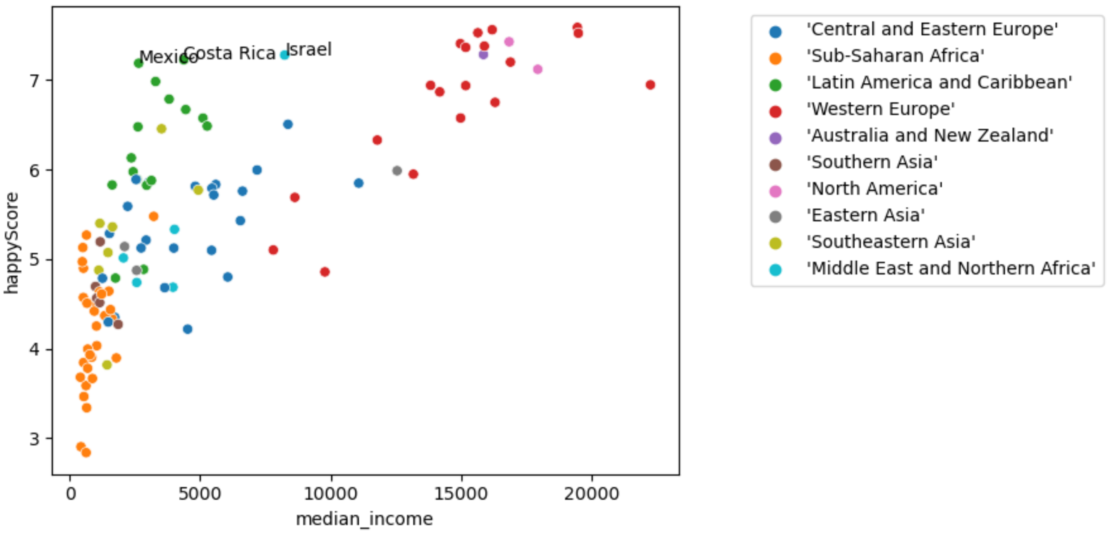
```
[6]: outlier_countries = np.where((data['happyScore'] >= 7) &
                          ((data['region'] == "'Latin America and Caribbean'")
                           | (data['region'] == "'Middle East and Northern␣
      ↪Africa'")))
     happiness = pd.DataFrame(display(data.loc[outlier_countries]))
```

```
       country  adjusted_satisfaction  avg_satisfaction  std_satisfaction  \
22  Costa Rica                   73.0               8.5              1.71
45      Israel                   61.0               7.3              2.09
70      Mexico                   69.0               8.3              2.02

      avg_income  median_income  income_inequality  \
22   6901.466667    4373.520000          49.018889
45  10645.240000    8234.680000          41.940000
70   4148.000000    2646.973333          48.974444

                               region  happyScore      GDP   country.1
22      'Latin America and Caribbean'       7.226  0.95578  Costa Rica
45  'Middle East and Northern Africa'       7.278  1.22857      Israel
70      'Latin America and Caribbean'       7.187  1.02054      Mexico
```

```
[4]: sns.scatterplot(x = data['median_income'],y= data['happyScore'], hue =␣
      ↪data['region'])
     plt.legend(bbox_to_anchor=(1.1, 1), loc='upper left')
     plt.text(x=data.loc[22]['median_income'], y=data.loc[22]['happyScore'], s=data.
      ↪loc[22]['country'])
     plt.text(x=data.loc[45]['median_income'], y=data.loc[45]['happyScore'], s=data.
      ↪loc[45]['country'])
     plt.text(x=data.loc[70]['median_income'], y=data.loc[70]['happyScore'], s=data.
      ↪loc[70]['country'])
     plt.show()
```



I decided to follow a similar pattern of scatterplot as shown in the course after observing that the columns in fact point to the same trends, with most of the numeric columns indicating income with a different measure. Although I used the median income, as preferred, it could very well be done with average as presented in the course.

I chose to use as the real difference, however, the inclusion of a hue according to the regions. Thanks to the hue, we can similar trends being presented by regions. While the lowest scores in income and happiness as easily seen as belonging to African countries, the highest ones are also easily seen belonging to European and North American countries.

As it was noticed before, some countries with low income have higher happyscores. Due to the hue, it is possible to see that this pattern belongs to Latin American countries. I chose to identify just the countries that presented themselves with a very high happiness index, above 7, that appeared as outliers and special cases to me. Those were Mexico and Costa Rica in Latin American, and also Israel to the Middle East and Northern Africa. Israel is a very special case in the Middle East, and in the plot can be seen extremely far from other countries from the same region. Although the focus was in these 3 countries with a very high score on happiness, other points that present themselves away from their region could also be of interest in a different analysis.

To conclude this, sorting was not exactly needed for this analysis, except for the creating of a hue in the graph, which groups the data by regions. Equally the filter used was to identify my points of interest in order to label them in the scatterplot.