Hello, in this notebook I'll try to explore and analyze the data.

I am quite a beginner on plots and visualization, but would love to learn more. Any feedback could help me to learn and experience more :)

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt           # for the following plots I import the matplotlib library
```

From sidebar, it's visible that we have 3 different CSV files for the years 2015, 2016, 2017.

So, I will analyze one of them first then the other. Then I will try to make a conclusion to come up with a result.

```
data_15 = pd.read_csv("../input/2015.csv");
data_16 = pd.read_csv("../input/2016.csv");    # I store them separately to work easily w/o confusions
data_17 = pd.read_csv("../input/2017.csv");
```

```
data_15.info()  #checking values and the types
```

```
data_15.head()      # what I see first is our data is sorted by Happiness Rank,
```

```
data_15.tail()  # most important value seems to be "Happiness Score"
```

```
data_15.describe()  # getting more involved with the statistical side of data
```

```
data_15.corr()    # this is probably the most important part of the kernel. We may see which factors directly affect happiness.
                  # numerical output might be enough to work, but as a data scientist candidate I should make a visible result.
```

To make a cool looking heatmap, I will use seaborn

```
import seaborn as sns
```

```
plt.figure(figsize=(16,10))  # on this line I just set the size of figure to 12 by 10.
sns.heatmap(data_15.corr(), annot=True, linewidths=1, cmap="ocean_r", fmt=".2f")  # seaborn has very simple solution for heatmap

plt.suptitle("Correlation Map", fontsize=18)

plt.show()   # whitest and greenest are most correlated
```

Usable correlation columns for Happiness Score are:

Economoy, Family, Health

Let's work on them to see clearly

```
plt.clf()  # to clear our plots before re-creating them.

plt.figure(figsize=(16,15));  # to make it easy to divide and see

plt.subplot(3,1,1);      # 1st row
plt.scatter(data_15['Happiness Score'], data_15['Economy (GDP per Capita)'], color='g');
plt.xlabel("Happiness Score");
plt.ylabel("Economy");

plt.subplot(3,1,2);     # 2nd row
plt.scatter(data_15['Happiness Score'], data_15['Family'], color='b');
plt.xlabel("Happiness Score");
plt.ylabel("Family");

plt.subplot(3,1,3);     # 3rd row
plt.scatter(data_15['Happiness Score'], data_15['Health (Life Expectancy)'], color='r');
plt.xlabel("Happiness Score");
plt.ylabel("Health");


plt.suptitle("FAMILY / ECONOMY / HEALTH",fontsize=18)
plt.tight_layout()
plt.show()
```

Graphs above shows that Health, Economy and Family does have big impact on Happiness.

Let's explore the data from 2016

and see how region affects the Happiness

```
data_16.head()
```

```
data_16.tail()
```

```
data_16.info()
```

```
data_16.describe()
```

```
data_16.corr()   # since our Region is stored as object value, we cannot see the correlation of it for happiness score.

# so the table below is not super helpful for the task I need to do.
# but we may find a good correlation to use on our plot
```

```
plt.figure(figsize=(16,10))
sns.heatmap(data_16.corr(),annot=True,cmap="YlGnBu",fmt=".2f");
plt.show()
```

The heatmap above shows that Health is very corralated with Economy

```
all_regions = data_16.Region
unique_regions = set(all_regions)    #in here I am getting the names of each possible region to put them on dictionary.
print(unique_regions)

print(len(unique_regions))
```

I will show the affect of region on the Happiness score,

to do this I will create a dictionary that would match the Region keys with the given colors.

there is 10 regions on our data, there should be 10 keys in our dictionary

```
region_colors = {'Middle East and Northern Africa':'red',
                 'Latin America and Caribbean':'green',
                 'Eastern Asia':'aqua',
                 'Sub-Saharan Africa':'blue',
                 'Southeastern Asia':'grey',
                 'Western Europe':'pink',
                 'North America':'yellow',
                 'Australia and New Zealand':'orange',
                 'Central and Eastern Europe':'purple',
                 'Southern Asia':'olive'};

type(region_colors)


colors = []
for i in data_16['Region']:

    colors.append(region_colors[i])




plt.clf()
plt.figure(figsize=(20,10))
plt.scatter(data_16['Economy (GDP per Capita)'], data_16['Health (Life Expectancy)'], s=(data_16['Happiness Score']**4), alpha=0.5, c=colors)
plt.grid(True)

plt.xlabel("Economy")
plt.ylabel("Health")

plt.suptitle("Health Economy graph with sizes as Happiness score and colors as Region", fontsize=18)

plt.show()
```

The chart above clearly shows Regions by color, Happiness scores by the size of each scatter.

What we see clearly is, Sub-Saharan countries has very low Economy and Health. An by this, they have lowest scores of happiness most of the time.

Conclusion

My analysis and graphs on the data shows that, there are some really big factors on Happiness such as Health, Family and Economy.

It's also visible that some Regions on world has averagely higher Scores than the others. For example, Eastern Europe and Sub-Saharan Africa has dramatic difference.

```
# thanks for reading, Your Comments and Votes are important for me. Best Regards, Efe.
```