Data Engineering Project
World Happiness Report

Authors:

George-Vlad Manolache 1718986

Petru Balan 1719379

1.Introduction

In this report we aim to analize the World Hapiness dataset from years 2015-2019. This dataset gives the happines rank and happiness score from the countries all around the world based on multiple factors like GDP per capita, Social support, Healthy life expectancy, Freedom, Generosity, Perceptions of corruption, etc. Sum of these values of these values gives us the happines score and the higher the happiness score, the lower the happines rank. We can define the meaning of these factors as the extent to which these factors lead to happines.

Preview in 2019 data:

	Overall rank	Country	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1	Finland	7.769	1.34	1.587	0.986	0.596	0.153	0.393
1	2	Denmark	7.6	1.383	1.573	0.996	0.592	0.252	0.41
2	3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
3	4	Iceland	7.494	1.38	1.624	1.026	0.591	0.354	0.118
4	5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298

The purpose of the report is to find which factors are more important in order to achive a higher level of happines and to describe the relations between them.

This report can be divided in three parts as follows: Cleaning the data, Visualisation and analysis of the data, Making predictions.

2. Cleaning the data

We start by checking the datatype of our features, search for duplicates, search for NA values and to remove unecessary columns.

It turns out that Datasets are clean in terms of datatypes and there is no need to convert one datatype to another. We also didn't find any duplicates.

During this process, it turned out there is one missing value in happy_2018 dateset, so we dropped it. Also, Overall rank is a unneccessary column and it can be removed.

3.Visualisation and analysis

3.1 Explanatory Statistics

We calculated the mean, standard deviatoin, median, minimum and maximum value of every feature of the dataset for every year.

2015:

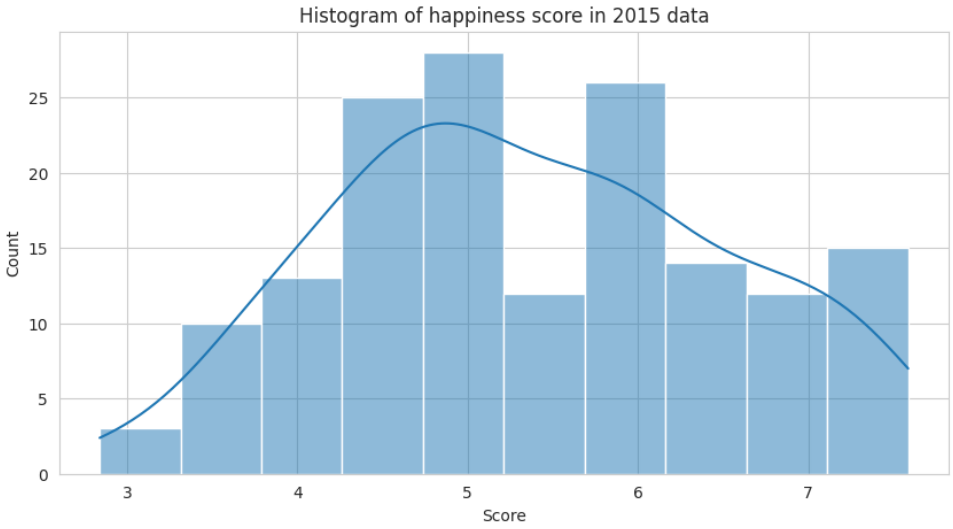
	Score	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Generosity	Dystopia Residual
count	158	158	158	158	158	158	158
mean	5.375734	0.846137	0.991046	0.630259	0.428615	0.237296	2.098977
std	1.14501	0.403121	0.272369	0.247078	0.150693	0.126685	0.55355
min	2.839	0	0	0	0	0	0.32858
25%	4.526	0.545808	0.856823	0.439185	0.32833	0.150553	1.75941
50%	5.2325	0.910245	1.02951	0.696705	0.435515	0.21613	2.095415
75%	6.24375	1.158448	1.214405	0.811013	0.549092	0.309883	2.462415
max	7.587	1.69042	1.40223	1.02525	0.66973	0.79588	3.60214

2019:

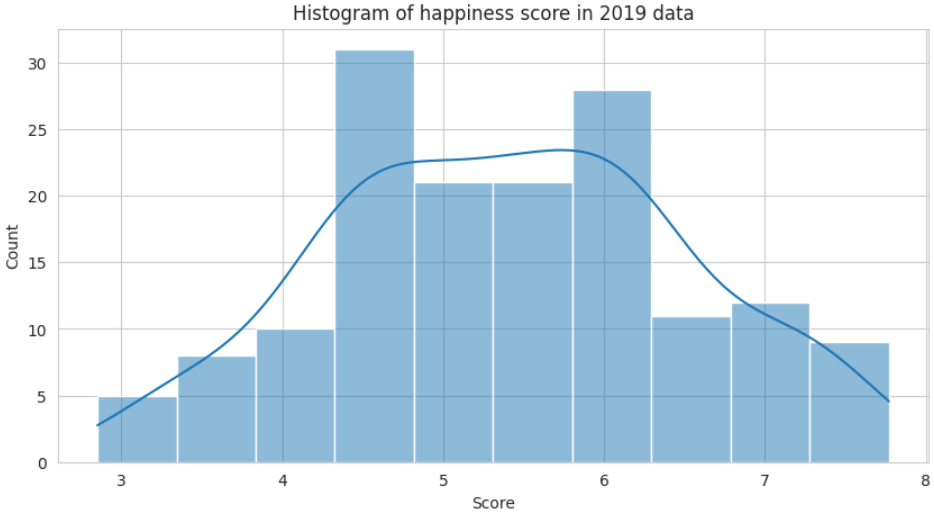
	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
count	156	156	156	156	156	156	156
mean	5.407096	0.905147	1.208814	0.725244	0.392571	0.184846	0.110603
std	1.11312	0.398389	0.299191	0.242124	0.143289	0.095254	0.094538
min	2.853	0	0	0	0	0	0
25%	4.5445	0.60275	1.05575	0.54775	0.308	0.10875	0.047
50%	5.3795	0.96	1.2715	0.789	0.417	0.1775	0.0855
75%	6.1845	1.2325	1.4525	0.88175	0.50725	0.24825	0.14125
max	7.769	1.684	1.624	1.141	0.631	0.566	0.453

Also, we plotted a histogram with number of countries in each happiness score range for every year

2015:

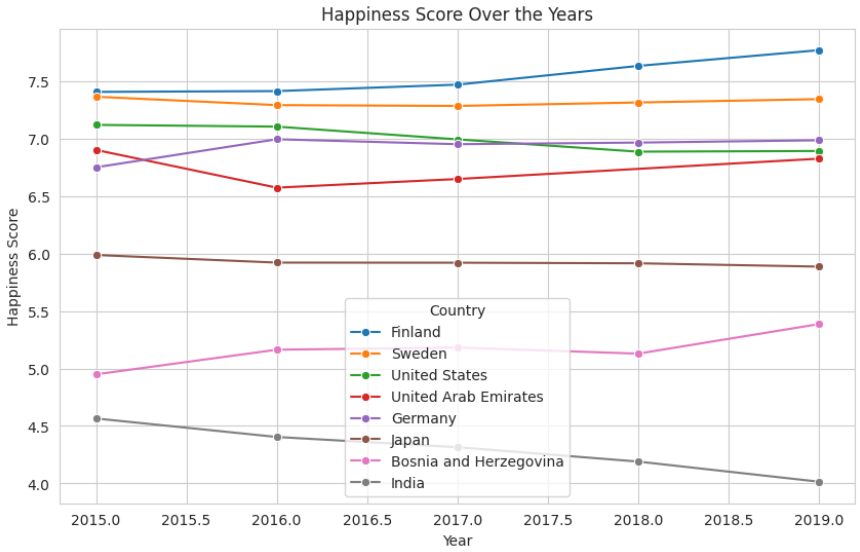

Histogram of happiness score in 2015 data
Count
Score
0
5
10
15
20
25
3
4
5
6
7

2019:


Histogram of happiness score in 2019 data
Count
Score
0
5
10
15
20
25
30
3
4
5
6
7
8

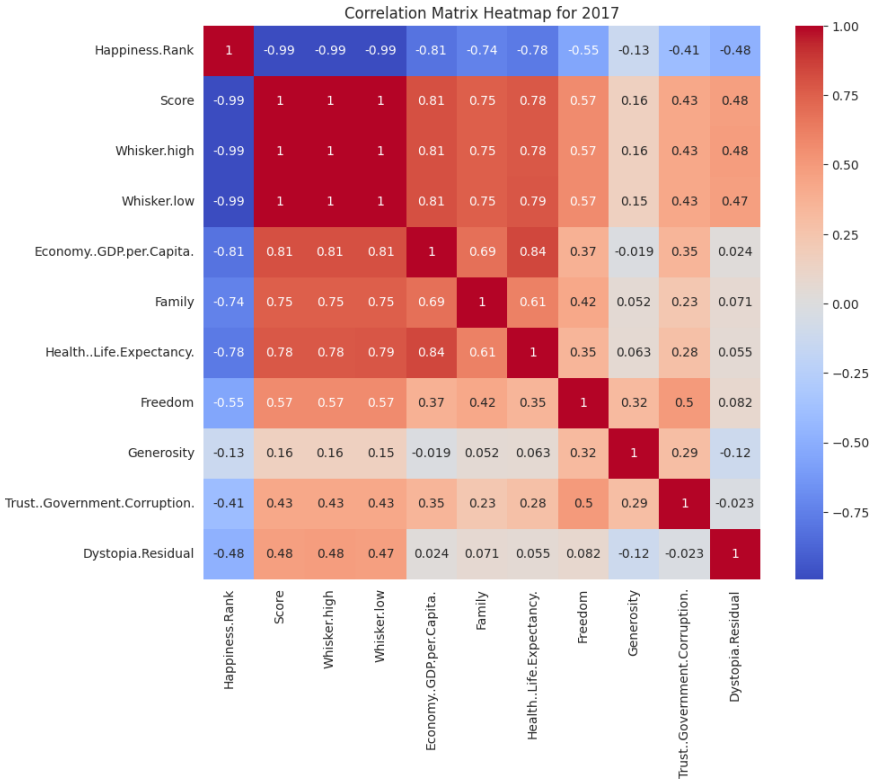
3.2 Visualization of trends throughout the years

We combined data from different years into a single DataFrame to be able to see the evolution of the Happines Score in time. The result is the following:



We can observe notable improvements for countries like Bosnia and Finland, which is also the country with the best score. Instead, India is constantly decreasing in happiness in every year. But there are also countries like Japan or Sweden who maintaint their level of happines.

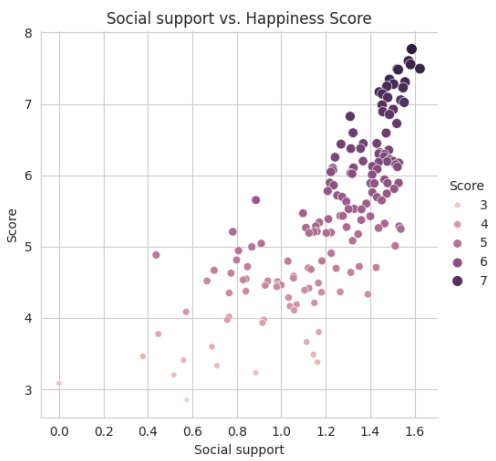
3.3 Correlation between features

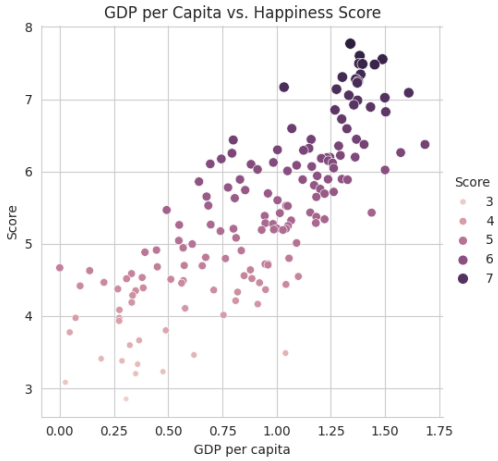


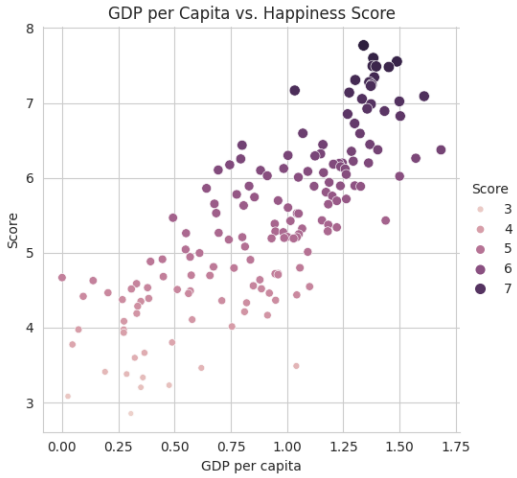
Looking at the correlation matrix we can see that factors as Social support, GDP per capita, Healthy life expectancy, Familiy have strong positive correlation and have a big contribution to the overall Happines Score. On the other side, we can't tell more about Generosity, which has the correlation scores near 0. **3.4**

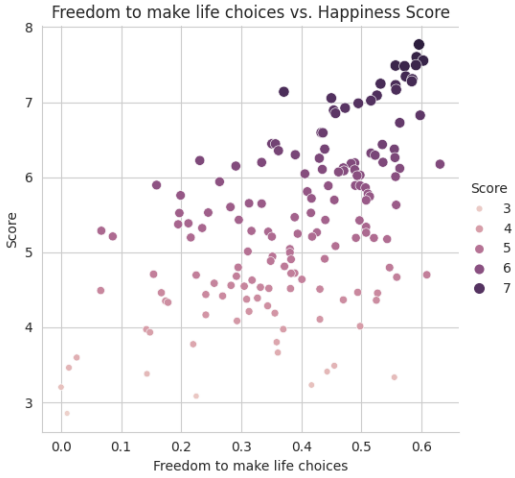
Relationship between core parameters and Happiness Score

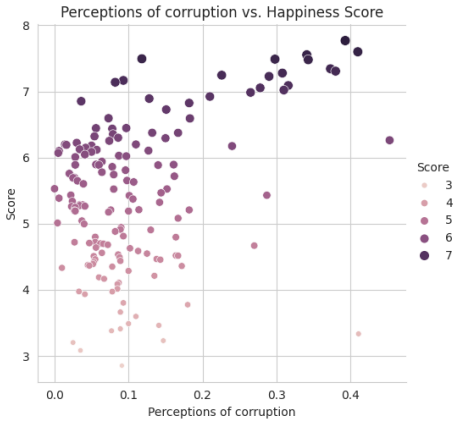
Below we explore facotors like Social support, GDP per Capita, Freedom, Perceptions of corruption in relation with the Happines Score.







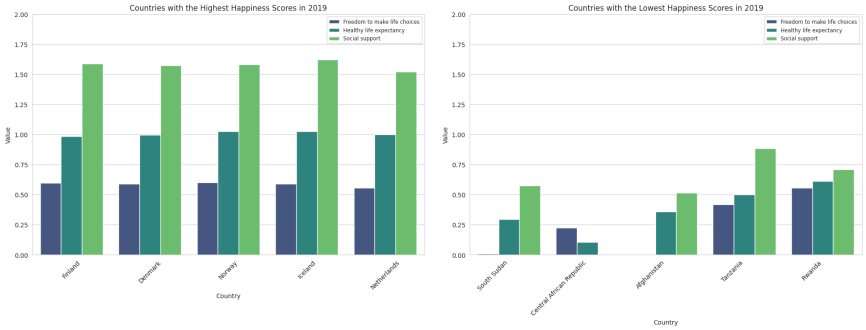




These plots give us the confirmation about the scores we alreday obtained from the confusion matrix, meaning the fact that between the Happiness Score and Social support there is a strong positive correlation, the same for GDP per Capita, while for Perceptions of corrruptions we can identify a weak positive correlation and for Freedom we can say that the correlation is somewhere in the middle.

3.5 Highest and Lowest Happines Scores

We explore here the top five countries with the Highest Happines Scores vs the ones with the Lowest Score of Happinens for the year 2019 in terms of Freedom to make life choices, Healthy life and Social support.



We can see that Social support and Healthy life can be important facts for a higher score of happines, all of the happiest countries having the Social Support score above 1.5 and a score of 1 for Healthy life. As for the least happies countries, we can observe that these scores are pretty low. Instead, it's interesting to see that Rwanda has almost the same score of Freedom as the happiest countries.

3.6 PCA and Clustering

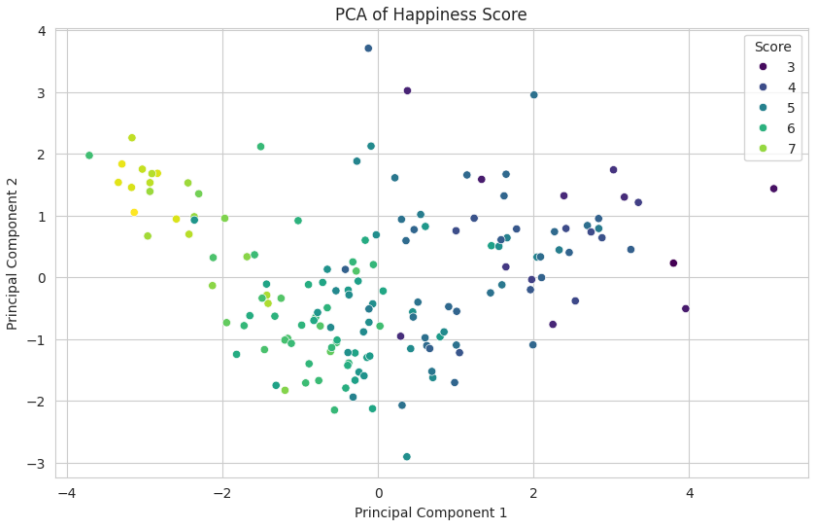
Next, we decide to explore our data features of the dataset using 2 unsupervised techniques: Principal Component Analysis for dimensionality reduction of the data and K-means for clustering the data described by PCA features.

PCA

We normalized the data and described it using 4 principal components with the following Explained Variance:

Ratio: 50%, 23%, 10% and 9%.

Sum: 0.93



We can see that by applying PCA to reduce the number of features to 4 it explains more than 90% of the variance in the data.

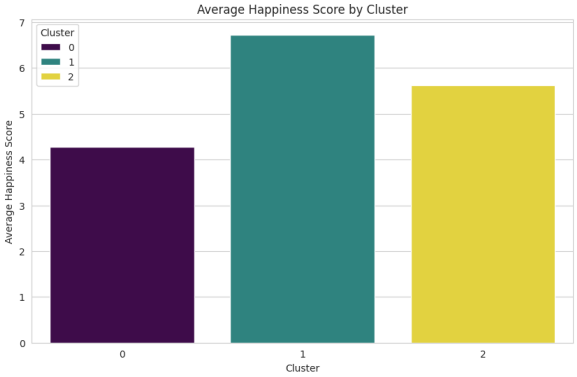
Also we can see that the happiness score is highly correlated with the first principal component(negative correlation).

Clustering

We found that the optimal number of clusters for groupping the data by PCA features is 3, using the Elbow Method and

calculating WCSS (Within-Cluster Sum of Squares distance). We also calculated the Average Happines Score by Cluster.

Results can be seen below:



3.7 Recommend similar countries

We calculated a similarity matrix of countries based on the cosine distance between the features used to describe the Happiness Score. We use this matrix to recommend for each country the top 5 most similar countries in terms of GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity, Perceptions of corruption

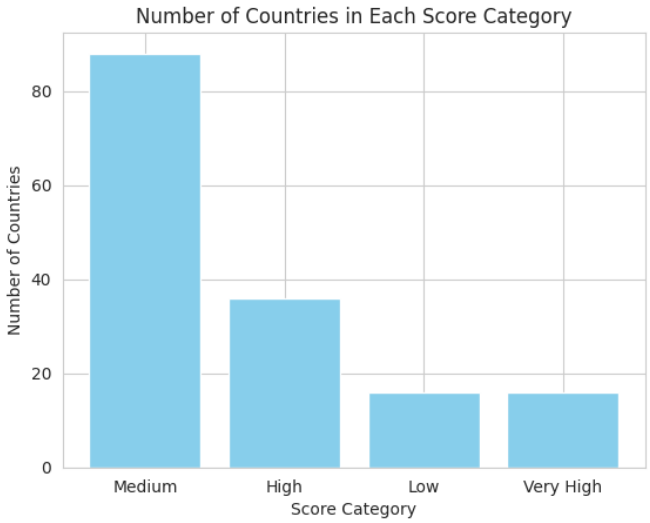
For example, the most 5 similar countries to Spain are: Chile, Italy, Slovakia, Japan and Cyprus

4. Predictions

We try to predict the data using 2 supervised algorithms: KNN and Linear Regression

4.1 KNN

The first step for this approch is to divide the Happines Score in 4 categories: Low (2-4), Medium (4-6), High (6-7) and Very High (7-8).



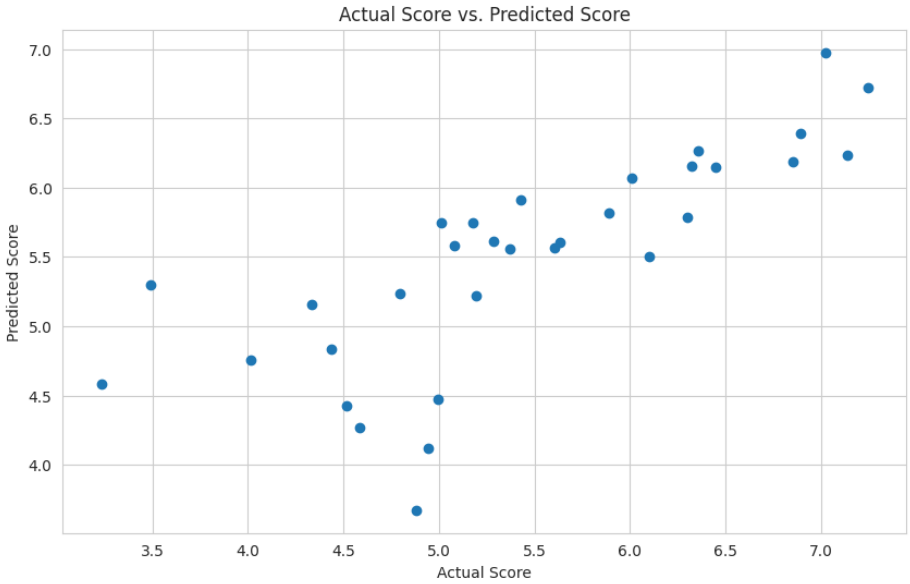
Second, we split our data into train and validation and apply the KNN algorithm using as features the ones mentionated in the previous sections and Score Categories as target. We calculated the overall accuracy and precision, recall and F1-score for each category. The results can be seen below:

Accuracy: 0.70

	precision	recall	f1-score	support
Low	1	0.5	0.67	2
Medium	0.88	0.72	0.79	29
High	0.35	0.75	0.48	8
Very High	1	0.62	0.77	8
accuracy			0.7	47
macro avg	0.81	0.65	0.68	47
weighted avg	0.81	0.7	0.73	47

4.2 Linear Regression

Again we split our data into train and validation and use a Linear Regression model to make predictions. We used Mean Squared Error as a criterion. The result is the following:



Mean Squared Error: 0.414464138352835

We can see that the happiness score can be predicted, with decent results.

5.Conclusions

In this notebook, we have analyzed the World Happiness Report dataset. We have seen that the happiness score is highly correlated with GDP per capita, social support, healthy life expectancy, and freedom to make life choices.

We have also seen that the happiness score can be predicted using these parameters by a Linear Regression method or categorizing the score by labels and using KNN, obtaining decent results.

We have also seen that the data can be clustered into 3 clusters in PCA space. Another thing we tried was to recommend similar countries give a country based on a similarity matix computed with cosine distance between features.