# What Makes Countries Happy?

**Shantanu Gupta**

**Chandana Kandari**

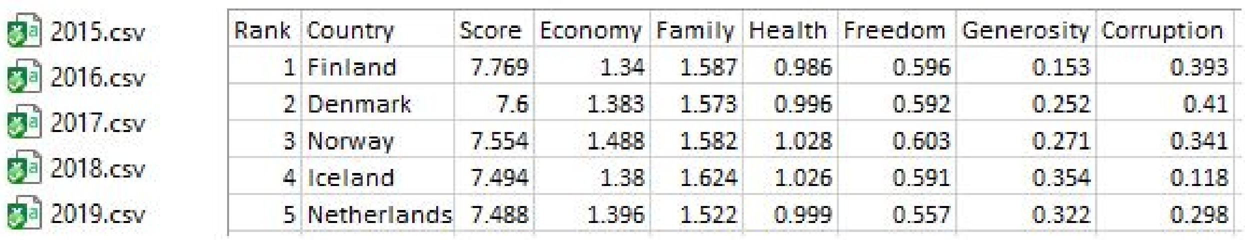
**12/05/2020**

## IFT 598: Analyzing Big Data

## Project Summary

**Goal:** Analyze the World Happiness Report for trends and insights.

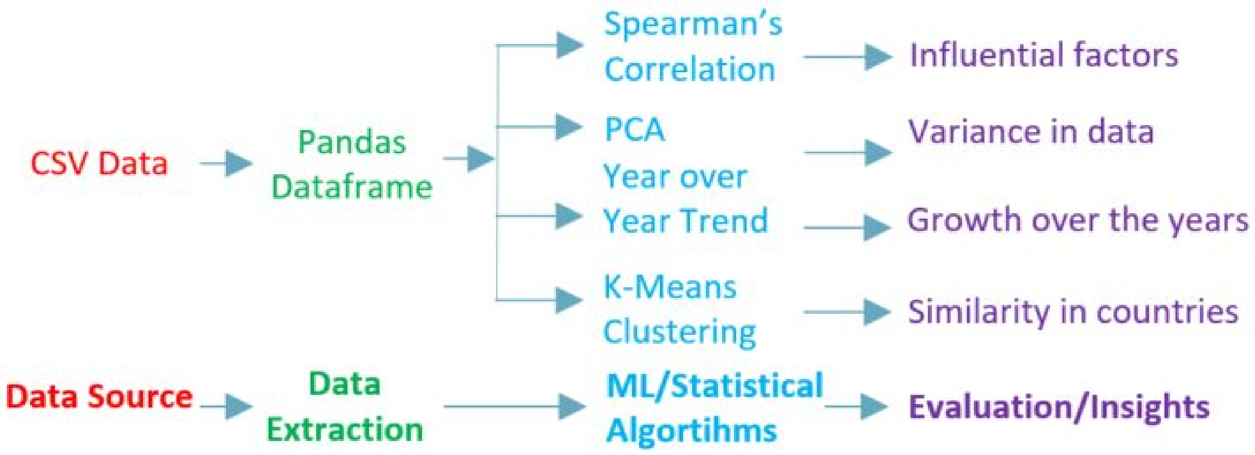
**Dataset:** 5 CSV files for dataset, containing countries with a final score and features as shown below.



2015.csv
2016.csv
2017.csv
2018.csv
2019.csv

| Rank | Country | Score | Economy | Family | Health | Freedom | Generosity | Corruption |
|---|---|---|---|---|---|---|---|---|
| 1 | Finland | 7.769 | 1.34 | 1.587 | 0.986 | 0.596 | 0.153 | 0.393 |
| 2 | Denmark | 7.6 | 1.383 | 1.573 | 0.996 | 0.592 | 0.252 | 0.41 |
| 3 | Norway | 7.554 | 1.488 | 1.582 | 1.028 | 0.603 | 0.271 | 0.341 |
| 4 | Iceland | 7.494 | 1.38 | 1.624 | 1.026 | 0.591 | 0.354 | 0.118 |
| 5 | Netherlands | 7.488 | 1.396 | 1.522 | 0.999 | 0.557 | 0.322 | 0.298 |

**Methods:**



### Technologies

Python 3, pandas, Matplotlib, seaborn, scikit-learn, Jupyter Notebook.

Code: https://github.com/shantanu-93/world-happiness-data-analysis

**We did not use an external database since our dataset was comparatively small (only 5 CSV files) and did not warrant a SQL/NoSQL database for our analysis.**

### Results

- Financial well being, Health and Family are the most important contributors to an individual's happiness.
- The world is seeing a positive trend over the years.
- The drastic decline in some countries can be attributed to government upheavals and war situations.

### Introduction

The World Happiness Report is an annual publication of the United Nations Sustainable Development Solutions Network. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

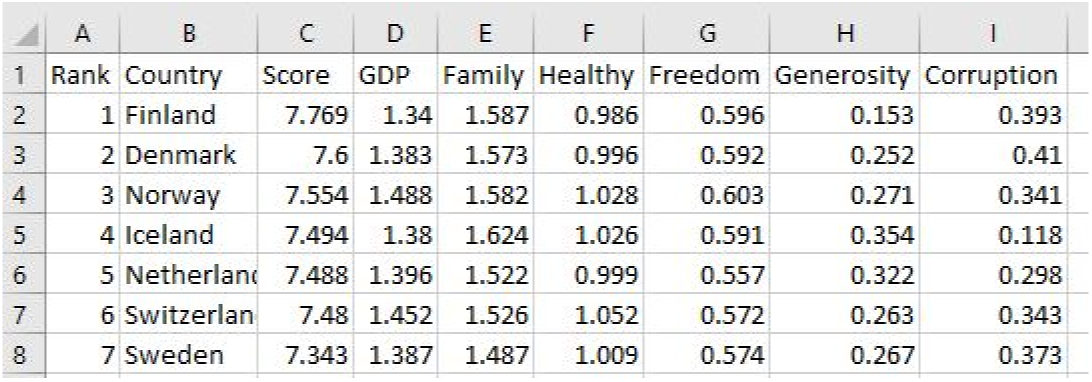
Six measurements are taken per country that describe the extent to which these factors contribute in evaluating the happiness in each country.

- GDP per Capita
- Family
- Life Expectancy
- Freedom
- Generosity
- Trust: Government Corruption

### Action Information

We want to know which of the six factors contribute to happiness and did any country experience a significant increase or decrease in happiness year over year. We are also going to find the correlation between each variable and the happiness score.

### Dataset



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Rank | Country | Score | GDP | Family | Healthy | Freedom | Generosity | Corruption |
| 2 | 1 | Finland | 7.769 | 1.34 | 1.587 | 0.986 | 0.596 | 0.153 | 0.393 |
| 3 | 2 | Denmark | 7.6 | 1.383 | 1.573 | 0.996 | 0.592 | 0.252 | 0.41 |
| 4 | 3 | Norway | 7.554 | 1.488 | 1.582 | 1.028 | 0.603 | 0.271 | 0.341 |
| 5 | 4 | Iceland | 7.494 | 1.38 | 1.624 | 1.026 | 0.591 | 0.354 | 0.118 |
| 6 | 5 | Netherlan | 7.488 | 1.396 | 1.522 | 0.999 | 0.557 | 0.322 | 0.298 |
| 7 | 6 | Switzerlan | 7.48 | 1.452 | 1.526 | 1.052 | 0.572 | 0.263 | 0.343 |
| 8 | 7 | Sweden | 7.343 | 1.387 | 1.487 | 1.009 | 0.574 | 0.267 | 0.373 |

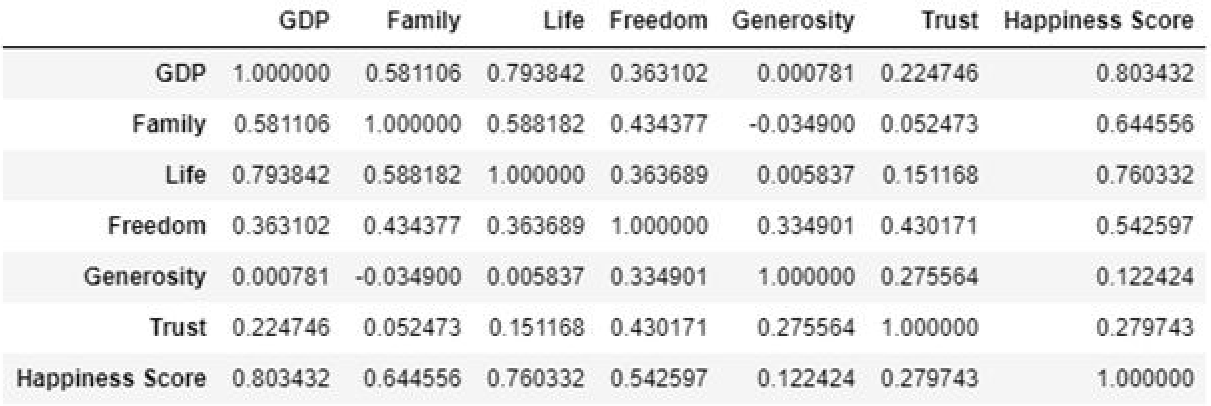
## Data preprocessing

We first wanted to understand how much of our data was missing. We went through both files and determined that we had no missing values. Next, we compared the dataset files we were given to ensure that they data was consistent. We believe some of the variable names are not clear enough and we decided to change the name of several of them a little bit. Also, we will remove whisker low and whisker high variables from the dataset because these variables give only the lower and upper confidence interval of happiness score and there is no need to use them for visualization and prediction.

The next step is adding another column to the dataset which is the Region. We want to work on different continents to discover whether there are different trends for them regarding which factors play a significant role in gaining a higher happiness score. Asia, Africa, North America, South America, Europe, and Australia are our six continents in this dataset.

We have chosen a target, as a variable that classifies the range of the feature variables into Top, Top-Mid, Low-Mid, Low.

**Since the data from the years have a bit of a different naming convention, so we have abstract these to a common name.**



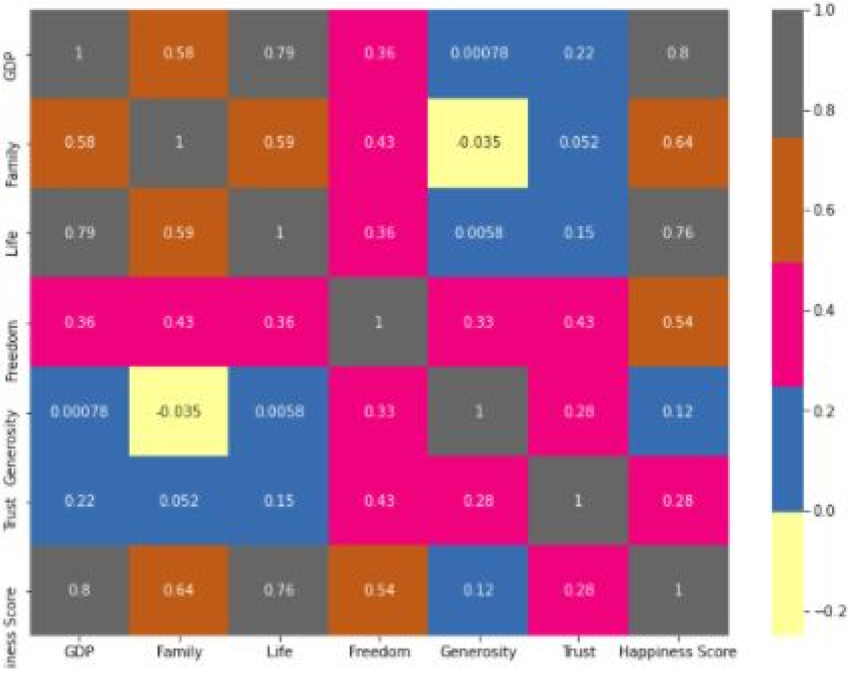
| | GDP | Family | Life | Freedom | Generosity | Trust | Happiness Score |
|---|---|---|---|---|---|---|---|
| GDP | 1.000000 | 0.581106 | 0.793842 | 0.363102 | 0.000781 | 0.224746 | 0.803432 |
| Family | 0.581106 | 1.000000 | 0.588182 | 0.434377 | -0.034900 | 0.052473 | 0.644556 |
| Life | 0.793842 | 0.588182 | 1.000000 | 0.363689 | 0.005837 | 0.151168 | 0.760332 |
| Freedom | 0.363102 | 0.434377 | 0.363689 | 1.000000 | 0.334901 | 0.430171 | 0.542597 |
| Generosity | 0.000781 | -0.034900 | 0.005837 | 0.334901 | 1.000000 | 0.275564 | 0.122424 |
| Trust | 0.224746 | 0.052473 | 0.151168 | 0.430171 | 0.275564 | 1.000000 | 0.279743 |
| Happiness Score | 0.803432 | 0.644556 | 0.760332 | 0.542597 | 0.122424 | 0.279743 | 1.000000 |

## Our Questions Answered

1. **We wanted to know what were the major factors of overall happiness?**

We analyzed what factors most influenced a country. By analyzing the data from 2015 to 2019, we analyze the major factors that contribute to an increase or decrease in happiness scores. Through this dataset, we have discovered many relationships across happiness and our informative variables. Here we show the spearman's correlation matrix between each variable and the happiness score.

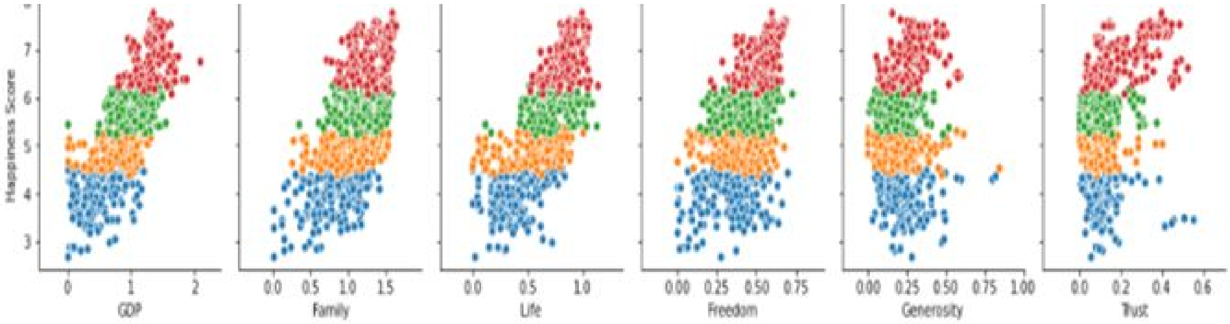
**We find that GDP(Financial well being), Life Expectancy (Health) and Family are the most contributing factors to a person's happiness.**



2. **We wanted to see how the independent variables are correlated with the Happiness Score?**

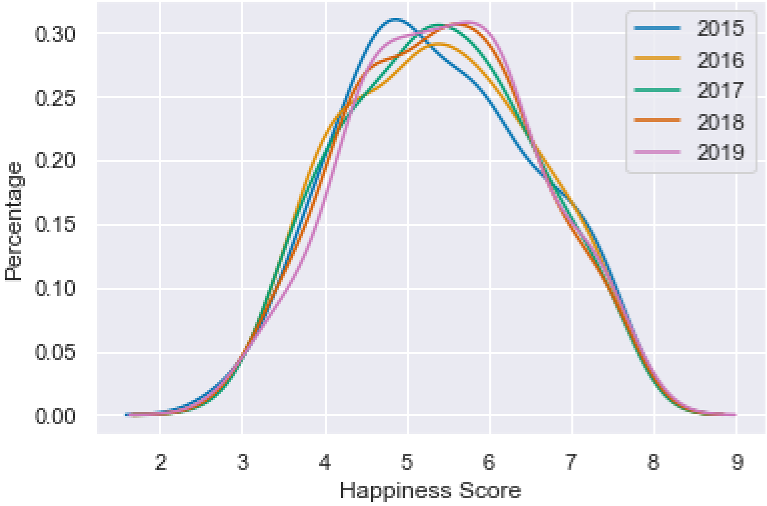
Features: GDP, Family, Life Expectancy, Freedom, Generosity, Trust: Government Corruption

We found that as certain variables increased such as GDP(Economy), Family, and Life Expectancy, overall rank of happiness increased. This was substantial as it gave us an understanding of happiness on a macro level. While plotting certain graphs we realized that there is a linear relationship with some of these informative variables. Although there are some outliers, the majority of correlations behaved the same. **When a country's happiness score and rank increased, their three most informative variables also increased.**



### 3. How has the happiness score changed year-over-year?

We find there is a gradual increase in happiness over the years, which is a good sign! The gradual shift in tilt near the maximum density changed from average happiness score of 4.5 in 2015 to 6 in 2019.

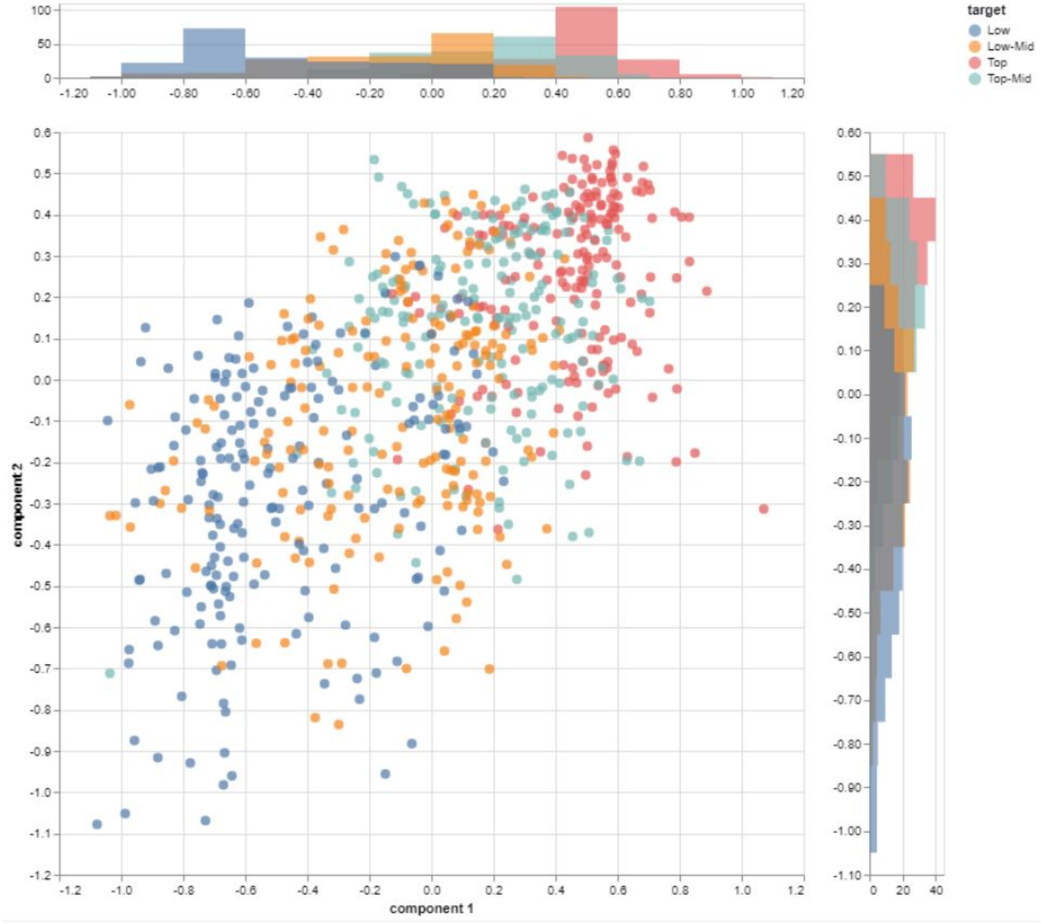


### 4. Explaining the variance in the data with PCA

Principal Component Analysis (PCA) is one of many dimensionality reduction techniques. PCA transforms the input data by projecting it into a lower number of dimensions called components without losing much information.

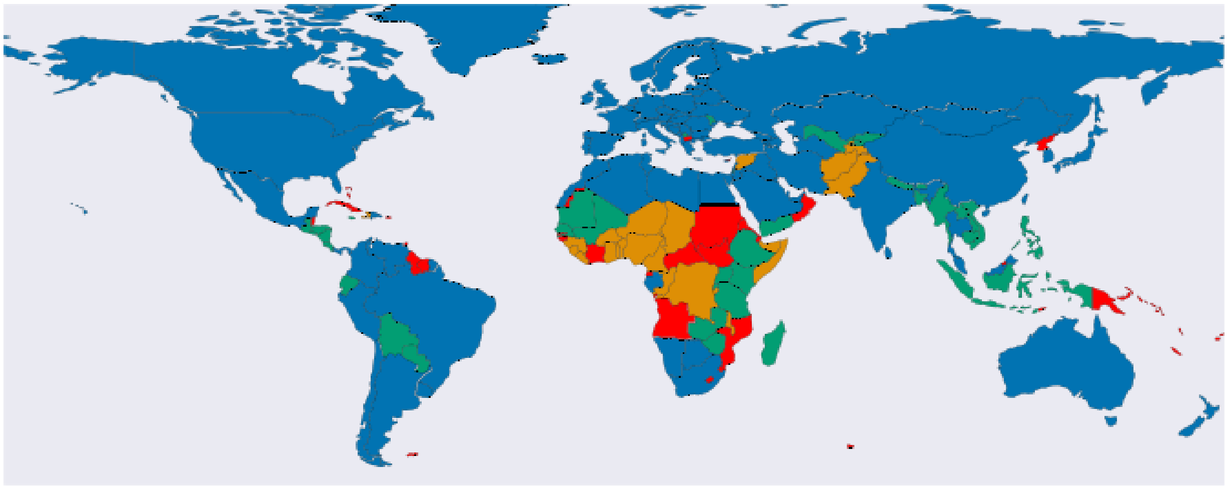
Dimensionality reduction using PCA is done because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

After dimensionality reduction, there usually isn't a particular meaning assigned to each principal component. The new components are just the two main dimensions of variation. Together, the two components contain 95.80% of the information.



5. **How does the clusters in influential contributing factors map across the globe?**

Performing k-means clustering over the data shows that most of the world considers the common factors like GDP, Health and Family as the most influential factors. But below map also shows that there is a different view in some regions around the world which could be attributed to the government stability in those countries.



## Technologies Used

We implemented our project in Python 3.x since it supports a variety of libraries for data manipulation, visualization and data science algorithms.

We performed our analysis in Jupyter Notebooks as they provide a great interface for experimentations and in-place visualizations.

We used pandas and numpy libraries for data engineering, manipulations and computations.

We used matplotlib, seaborn and altair libraries for the various visualizations shown in the report.

We used scikit-learn library to perform our data science experiments.

Code: https://github.com/shantanu-93/world-happiness-data-analysis

## Learnings

During the course of the project, we learned how to collect, transform and use raw data to generate insights into some interesting questions we sought to answer. We learned a lot about various visualizations and how to plot graphs, barcharts and other statistical figures. We learned about popular statistical and data science algorithms like spearman’s correlation, PCA and k-means clustering and how to use them in practice.

## Conclusion

We provided an in-depth analysis and explanation of the most influential factors of the happiness score of a country based on the data provided. Our analysis explains the correlation among the independent features and overall happiness score of a country. This analysis can potentially help governments plan policies to boost the long-term happiness of its citizens. It can also help organisations identify countries where they can do business with higher success rates. This can also serve as a marker to find where you might want to live next!

## References

[1] https://worldhappiness.report
[2] https://en.wikipedia.org/wiki/World_Happiness_Report
[3] www.kaggle.com/unsdsn/world-happiness
[4] www.builtin.com/data-science/step-step-explanation-principal-component-analysis
[5] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm