**Big Data Analysis Report File**

**Topic: World Happiness**

**Team Oreo**

**Participants:**

Komiljon Yuldashev - 12230144

Begon Eliott - 12230078

Coste Felix - 12230200

Jafar Saidov - 12194931

**Introduction**

In today's data-driven environment, big data analysis has become a potent tool for gaining insightful information and making wise judgments. Organizations are aggressively investing in initiatives that tap into the possibilities of Big Data as a result of the ever-increasing volume, velocity, and diversity of data. This report thoroughly studies our project, which attempts to identify the major factors and aspects that impact the high ranks of World Happiness by utilizing Big Data study approaches.

Our research aims to reveal hidden patterns, trends, and correlations inside massive datasets by utilizing cutting-edge technology and complex analytics algorithms using Tableau and ChatGPT. We believe that the knowledge gathered via this research has the ability to influence strategic choices, improve operations, and bring to light lucrative commercial prospects.

Throughout this report, we will examine the project's methodology, the datasets used, the analytical strategies used, and the important conclusions that resulted from our study. We will also go through the ramifications and possible applications of these results, as well as the difficulties faced and the methods used to get beyond them.

This study intends to add to the increasing body of knowledge in the field of Big Data Analysis by sharing our experiences, lessons learned, and noteworthy results. It also offers helpful insights for scholars, practitioners, and companies starting similar initiatives.

Now let's travel into the world of big data analysis and investigate the complexities and discoveries that our project has made.

**Methodology**

We will describe the approach used to carry out the analysis for the project based on Big Data Analysis in this part. The major components of the methodology are enumerated in the following sentences:

**Tools Applied:**

Tableau: With Tableau, users can quickly analyze, display, and share data in an engaging and dynamic way. Tableau is a sophisticated data visualization and business intelligence application. Users may connect to different data sources, alter and clean data, and produce meaningful dashboards, charts, and reports with only a few clicks.

ChatGPT: ChatGPT is a potent data analysis tool that makes use of its sophisticated natural language processing skills to enable dialog and offer insights. ChatGPT can help data analysts with database queries, dataset exploration, and difficult analytical jobs because of its capacity to comprehend and produce human-like replies.

Microsoft Excel: Users may effectively organize, modify, and analyze data with Microsoft Excel, a strong tool for data analysis. Excel's extensive set of functions and formulae make it possible for users to do computations, make charts and graphs, and draw conclusions from huge datasets. Users may easily analyze and summarize data with the use of its capabilities, which include sorting, filtering, and pivot tables.

Python. NumPy library: Data analysis depends heavily on the robust Python module NumPy. With its effective and adaptable tools, NumPy improves data analysis operations by enabling quicker computations and offering crucial capabilities for managing numerical data.

**Goals and Objectives Identified:**

Patterns in Overall Happiness: Our goal was to discover and examine patterns in overall happiness over the given time frame. This involves looking at how satisfaction levels changed over time and identifying any notable trends or shifts.

Factors Affecting Happiness: We looked at the variables that affected people's degrees of happiness. This meant examining many elements to see how they affected happiness, including things like income, employment, health, and social factors.

Correlations: To find connections and links between various variables and happiness, we used correlation analysis. To find any important insights, we also looked for patterns in the data.

We performed a longitudinal analysis to track alterations in happiness levels across time. We were able to recognize long-term trends and patterns thanks to this investigation.

We looked at and studied differences in happiness levels between multiple countries with different happiness levels. Potential disparities in happiness were clarified by this investigation.

\

**Data gathering and preparation:**

We were able to collect a large dataset with important variables for studying happiness.

To tidy up and get the dataset ready for analysis, data preparation operations were carried out. For instance, in Tableau we connected datasets for every year among each other so that it would be convenient for us to conduct analysis based on all years simultaneously.

**Data Evaluation**

To evaluate the dataset, we used a variety of statistical and data mining approaches. This comprised, among other things, time series analysis, clustering, and descriptive statistics.

In order to provide insightful conclusions, the analysis was carried out iteratively while experimenting with various combinations of variables and methodologies.

**Visualization and Analysis:**

To properly communicate the results, we developed interactive visuals in Tableau and reflected them in Microsoft Excel using its charts. These visualizations assisted in finding patterns, correlations, and trends in the data.

To make meaningful inferences regarding happiness trends, factors impacting happiness, correlations, and the obtained insights were interpreted and examined.

Our investigation was built on the methods described above, which enabled us to get important insights into patterns in happiness and associated factors. We were able to properly visualize and analyze the data using Tableau and ChatGPT, which contributed to a thorough grasp of the project's goals.

**Data Sources**

Kaggle is a well-known online site for sharing and discovering datasets, and it is where the datasets utilized in this study were found. The dataset includes a thorough assortment of World Happiness Reports from 2015 to 2019. These studies provide important insights into happiness levels and related aspects, serving as a comprehensive assessment of the state of happiness in the world.

**Regarding the Dataset:**

A well-known publication that offers a thorough review of happiness levels in various countries is The World Happiness Report. It is becoming acknowledged as a trustworthy resource for comprehending well-being and guiding policy choices. The reports released in 2015, 2016, 2017, 2018, and 2019 are included in the dataset used for this research.

The dataset includes a number of variables that represent various facets of happiness. In the analysis, the following variables were used:

*Country:* Displays the name of the nation under study.

*Region:* Describes the region of the world the country is a part of.

*Happiness Rank:* Indicates a country's position in relation to other nations based on its happiness score.

Based on answers to the Cantril ladder question, the *happiness score* represents the general degree of happiness in a nation.

The estimated standard error of *the happiness score* is indicated by the term *"standard error."*

*Economy (GDP per Capita):* This indicator gauges a nation's economic health by looking at the extent its Gross Domestic Product (GDP) per capita contributed to the Happiness Score.

*Family:* Captures the vitality and assistance of social ties within a nation by looking at the extent it contributed to the Happiness Score.

*Health (Life Expectancy*): The extent of the population's average life expectancy contributed to the Happiness Score, which reflects general health conditions, is represented.

*Freedom:* Indicates the extent at which personal freedom and self-determination contributed to the Happiness Score

*Trust (Government Corruption):* Measures the extent at which the degree of trust that people have in the government and their opinions about corruption contributed to the Happiness Score.

*Generosity:* The extent at which a measure of how generous and selfless a nation contributed to the Happiness Score.

The dataset makes use of information from the Gallup World Poll, which measures peoples' levels of life satisfaction using the Cantril ladder question. Utilizing Gallup weights to make sure the estimates are representative, the scores and rankings are calculated from samples that are nationally representative for the years 2015–2019.

We used this dataset to investigate and get insights into the trends, variances, and factors affecting happiness levels around the world in various nations and areas. The dataset's extensive coverage and wealth of data allow for a thorough investigation of phenomena related to happiness.

**Analytical Strategies**

This section will describe the analytical approaches used in the project to generate insights from the data by using Tableau and ChatGPT.

**Tableau**

Our analysis heavily incorporated Tableau, a potent tool for data visualization. We were able to study and display the dataset successfully thanks to its interactive and user-friendly interface. Tableau was used to implement the following analytic techniques:

*Comprehensive exploratory data analysis* was carried out in order to better comprehend the dataset. We were able to spot trends, and patterns in the data using Tableau's visualizations, which included multiple analytical charts.

Tableau's side-by-side representations and interactive dashboards made *comparison analysis* easier. To find differences and insights, we analyzed happiness ratings, rankings, and other factors across nations, regions, and years.

**ChatGPT**

By offering text analysis and producing insightful results, ChatGPT, an AI-powered language model, significantly contributed to our analysis. With the use of ChatGPT, the following analytical techniques were used:

*Sentiment Analysis:* We were able to conduct happiness-related sentiment analysis using ChatGPT. We learned more about the overall positive or negative correlations with happiness in various nations or locations by analyzing the textual content's sentiment.

*Topic Modeling:* ChatGPT assisted in locating hidden themes or subjects in textual data by using topic modeling approaches. We found underlying themes connected to happiness characteristics or regional differences by clustering and categorizing text based on similar content.

*Text Generation:* Using the data analysis as a foundation, ChatGPT produced textual insights. We were able to construct summaries, descriptions, and interpretations of the results by utilizing its language generating abilities, which increased the analysis's breadth.

We were able to combine the advantages of data visualization and text analysis methods by merging Tableau with ChatGPT. We were able to spot important links, trends, and insights within the dataset thanks to this thorough methodology, which gave us a thorough grasp of the project's goals and outcomes.

**Microsoft Excel**

It was utilized to create specific charts of overall trends depending on the tables obtained from Tableau.

**NumPy**

With the aid of NumPy, we conducted a *correlation analysis* to find patterns in the correlations between the variables. We looked at the correlation coefficients obtained using specific functions of NumPy like “corrcoef()”, which helped us understand how happiness interacts with other variables.

**Findings and Conclusions**

**Overall Happiness Trends**

Through a comparative analysis of happiness ranks from 2015 to 2019, we identified notable overall trends in happiness levels. The examination of happiness ranks across these years allowed us to discern patterns and variations in global happiness. Here are the key findings regarding the overall trends.

For instance, we can highlight here the top-10 performing countries that almost always during the period of 2015-2019 were in top-10 based on happiness ranks:

From this diagram (check the file “Happiness Trends of Top-10 Countries.xlsx”), we can clearly identify 3 types of countries:

1. Countries that are being ranked relatively stable over the years (New Zealand, Australia, Sweden)
2. Countries where the level of happiness is overall decreasing (Switzerland, Canada, Iceland)
3. Countries where the level of happiness is overall increasing (Denmark)

It is worth saying that these types also could be applied to all other left countries (check the file “Happiness Trends of All Countries.xlsx”) which will be also discussed. We selected top-10 countries because first of all it is easy to demonstrate a sample of 10 countries in the diagram and secondly this sample should be enough diverse, and these demonstrated countries are enough diverse.

**Factors Impacting Happiness**

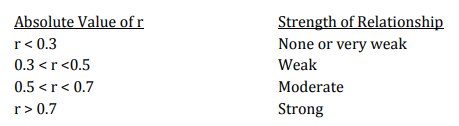
Now let’s understand the factors (that are measured as extents to the Happiness Score) that were a reason why there are such 3 types by going deeply into the analysis of the countries.

Let’s discuss each type one by one:

1. In order to discuss the overall factors that affect the countries and make them stay stable we should consider a variety of example countries (check the file “Happiness Trends for the 1-st Type Countries.xlsx”), like Afghanistan, Armenia, Burundi, Madagascar, Mauritania, Palestinian Territories, United Arab Emirates, Uruguay:

Now let’s find out correlations between the happiness rank and all possible factors that may impact it.

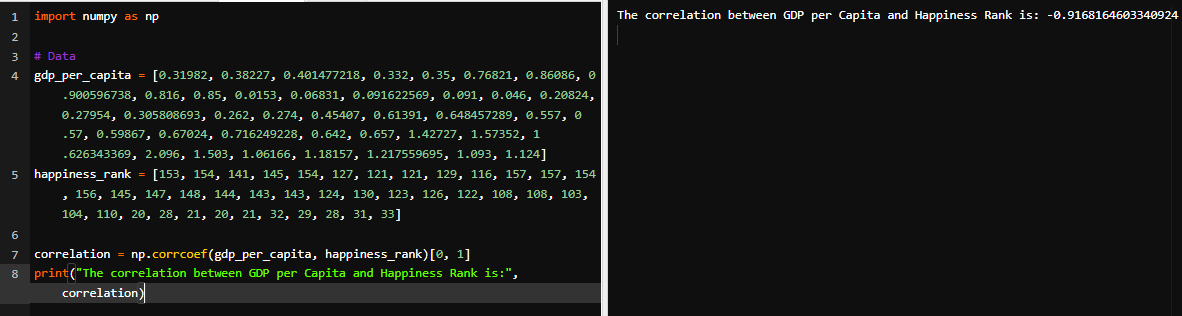
In order to identify if the correlation is strong/weak/moderate we will use the following table:



Firstly let’s consider the GDP per Capita across the years and for all the 1-st type countries:

In the above diagram, you can observe the values of the GDP per capita for those countries.

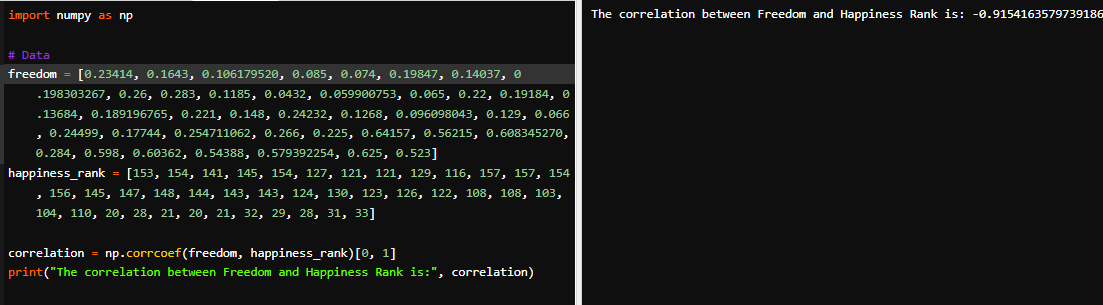
Now, considering the happiness rank and the GDP per capita, let’s calculate the correlation between them:



As you see, the correlation is negative (-0.917) which means that when one value increases the other value decreases. For instance, if the GDP per capita increases then the happiness rank decreases (which means that the happiness level increases). And as the absolute correlation value 0.917 > 0.7 that means that this is a strong negative correlation. So, the GDP per Capita is strongly related to the happiness rank.

Now, the next factor is Freedom to make life choices:

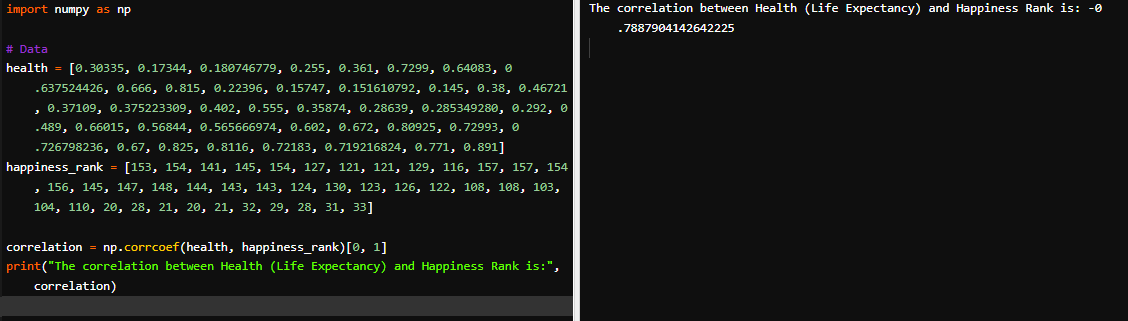
So, the correlation is:



There is also a strong negative correlation (-0.915) which means that this factor also strongly impacts on the happiness ranking.

Another factor is Health (Life Expectancy):

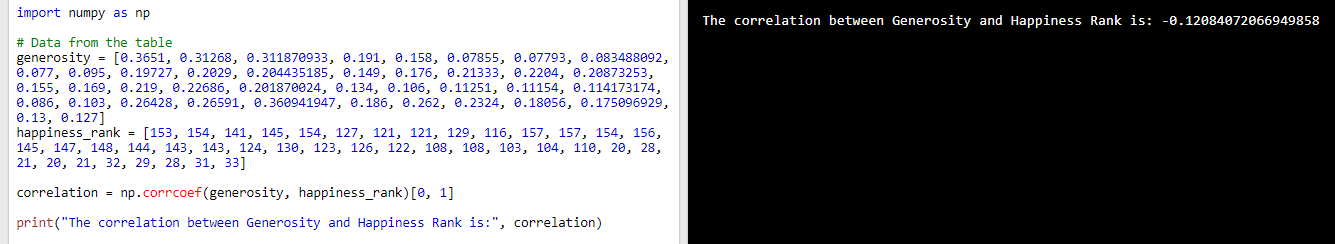
The correlation is:



Analyzing this correlation we can observe that definitely for this relationship the correlation is less than for Freedom and GDP per Capita, but it is still a strong negative (0.78 > 0.7) one.

The following factor is Generosity:

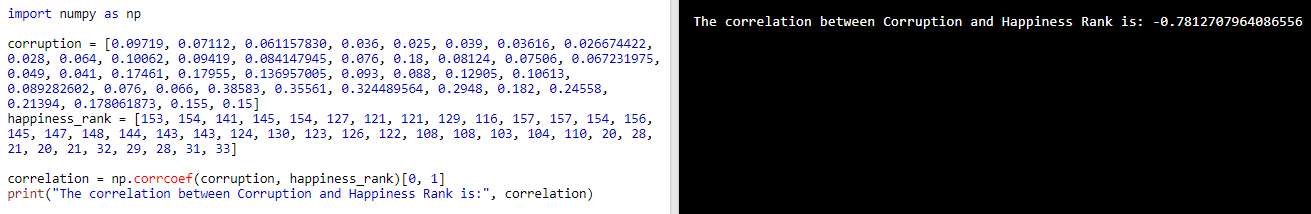
And its correlation can be calculated using the following code:



As you see the absolute value of correlation is less than: 0.12 < 0.3 which means that there is too weak correlation or almost no correlation. So, the impact of generosity is too weak on the happiness level of the 1-st type of country.

The next factor is Trust (Perception of Government Corruption):

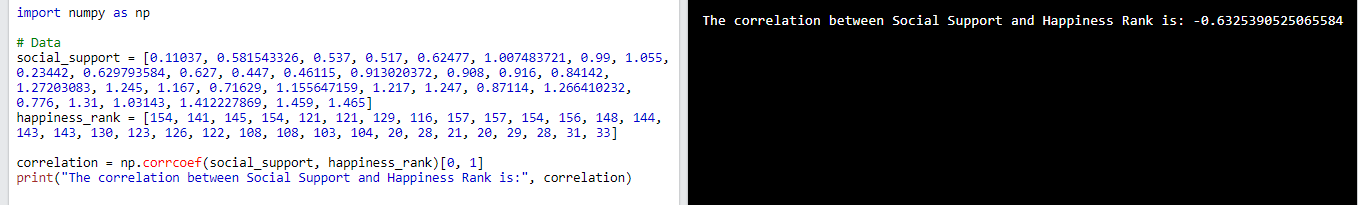
The correlation is:



0.78 > 0.7, so this is another strong negative correlation between Perception of Corruption and Happiness.

The last factor is Social Support:

And its correlation:



The absolute value is less than 0.7 => 0.5 > 0.63 < 0.7 which means that this is a moderate negative correlation.

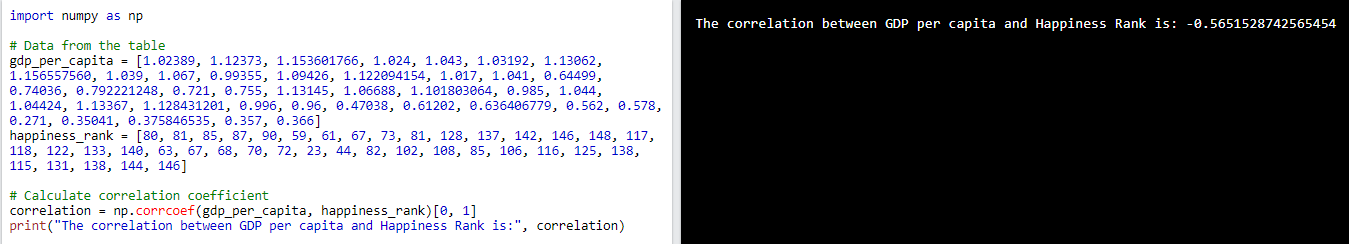
In conclusion, we can mention that the factors like GDP per capita, Freedom, Health, and Corruption strongly affect the happiness rank for this type of countries, Social Support moderately affects and Generosity almost doesn’t affect the happiness rank.

Now let’s move on to the 2-nd type of countries (Azerbaijan, Belarus, Botswana, India, Libya, Zambia, Zimbabwe, Venezuela):

We will do the same analysis that we did for the 1-st type to identify the most important factors here:

GDP per Capita:

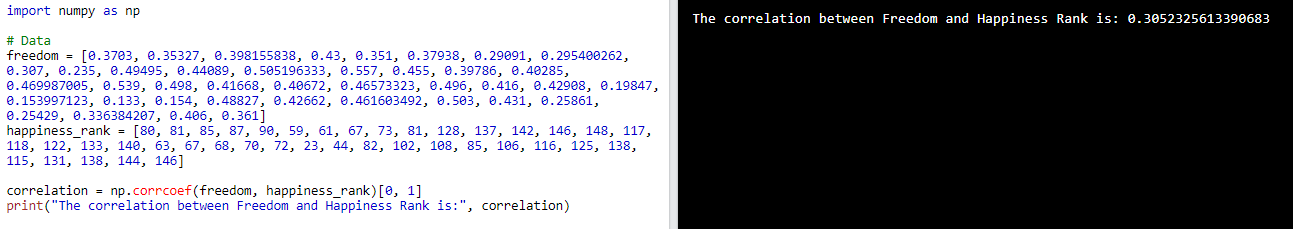
The correlation is:



0.5 < 0.565 < 0.7. That means that there is a moderate negative correlation between GDP per capita and Happiness Rank.

Freedom to make life choices:

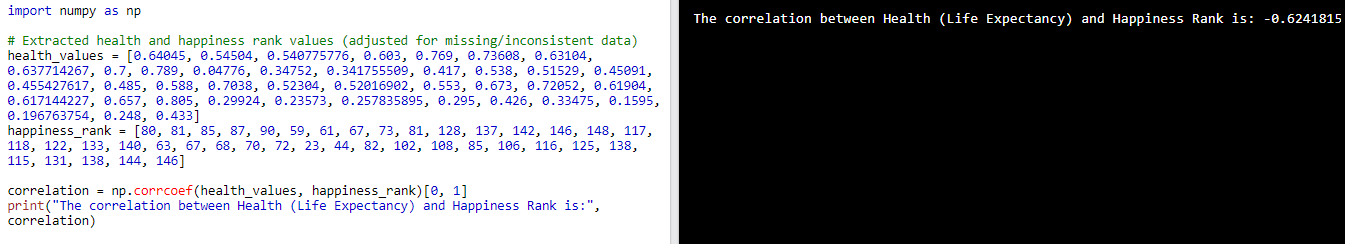
The correlation is:



0.3 < 0.305 < 0.5. However, it should be noticed that there is a positive weak correlation, which is unusual and means that when freedom increased then the happiness rank also increased (the happiness level decreased). But the correlation rate is near to the very weak correlation so its effect is near to very weak which means that this factor barely affects the happiness level. Nevertheless, we got the result we got, so now let’s proceed further.

Health (Life Expectancy):

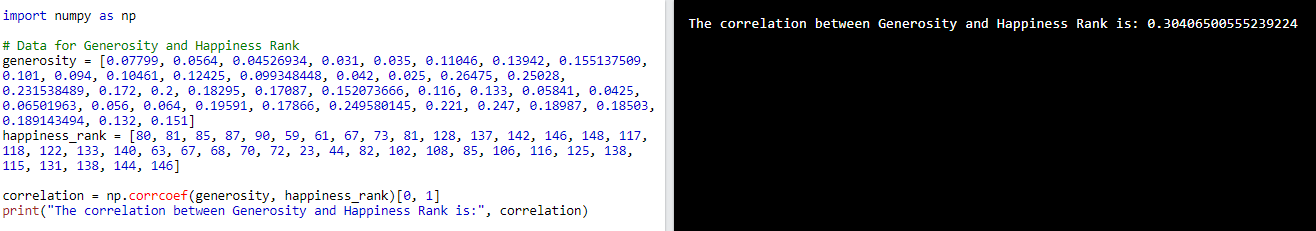
The correlation is:



That is 0.5 < 0.624 < 0.7 a moderate negative correlation.

Generosity:

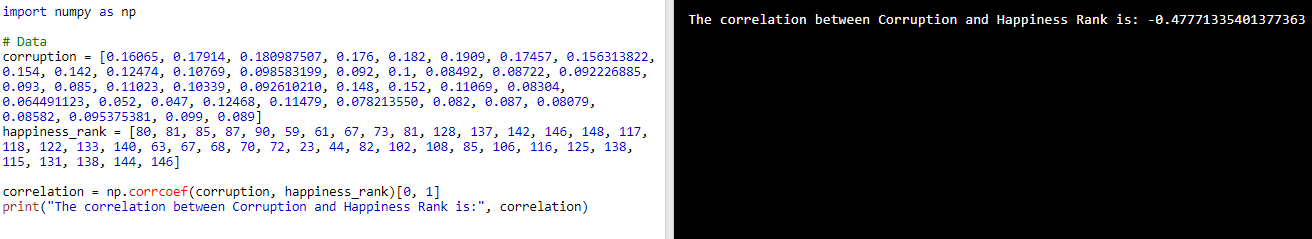
The correlation is:



0.3 < 0.304 < 0.5. Again a weak positive correlation means this factor barely affects the happiness level.

Perception of Corruption:

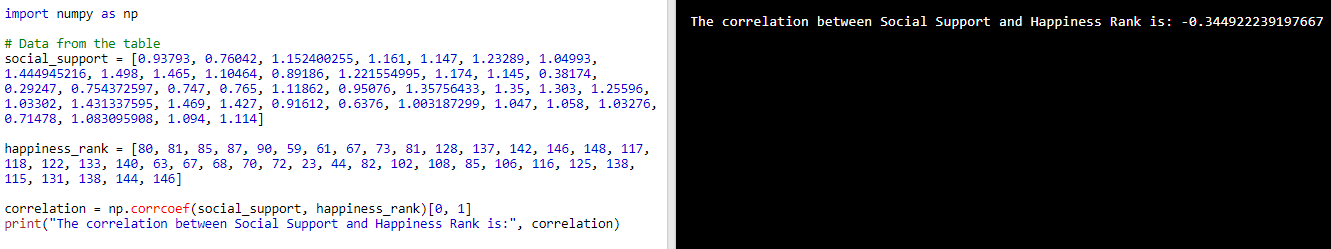
The correlation is:



0.3 < 0.47 < 0.5, a weak negative correlation.

Social Support:

The correlation is:



0.3 < 0.345 < 0.5, a weak negative correlation.

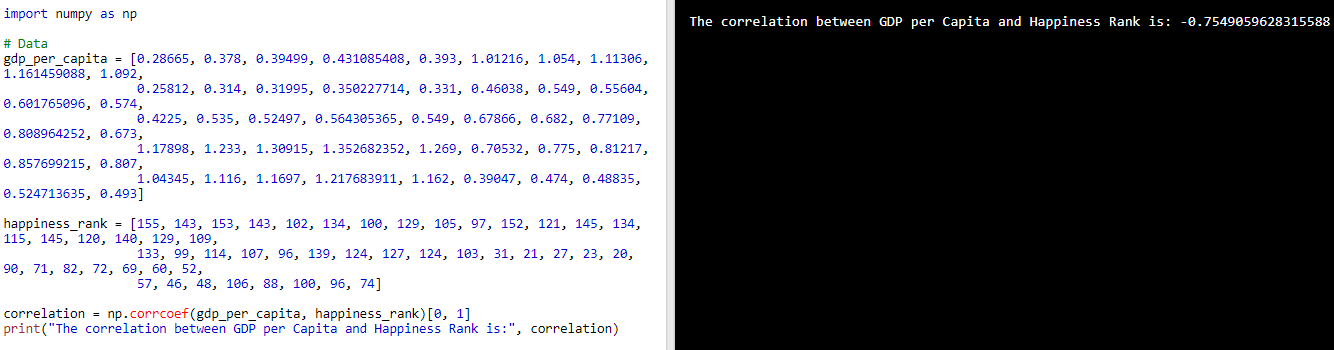
In conclusion, our analysis shows that parameters like GDP per capita, health, and corruption have a negative correlation with Happiness Rank, indicating that they may lower levels of happiness in general. The positive correlation between freedom, generosity, and social support, on the other hand, shows that these factors do, albeit little, contribute to happiness. It is crucial to remember that while these correlations shed light on the connections between various variables and Happiness Rank, they do not prove causation. Happiness levels within a nation can also be influenced by other variables including personal circumstances.

Now, let’s analyze the 3-rd type countries (Benin, Bulgaria, Burkina Faso, Cambodia, Cameroon, Congo (Brazzaville), Czech Republic, Poland, Philippines, Romania, Tajikistan):

We will do the same analysis that we did for the 1-st and 2-nd types to identify the most important factors here:

GDP per Capita:

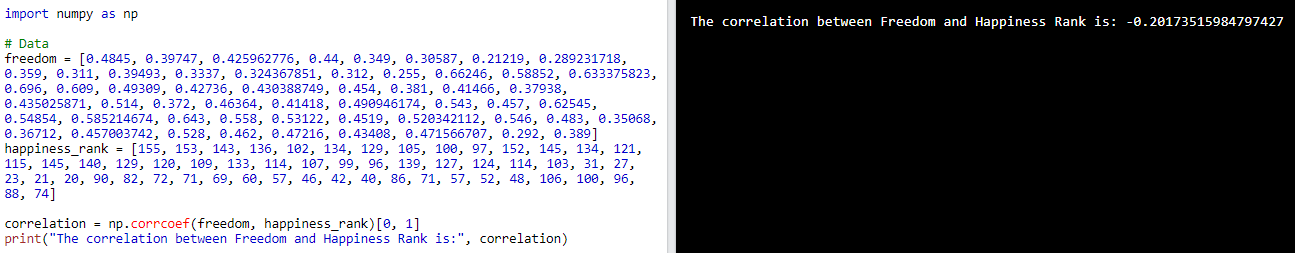
The correlation is:



0.755 > 0.7, a strong negative correlation.

Freedom to make life choices:

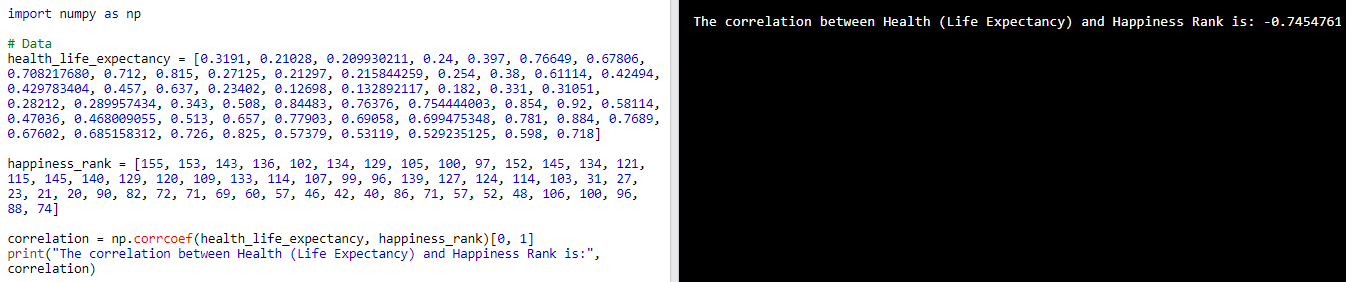
The correlation is:



0.2017 < 0.3, a very weak/no negative correlation.

Health (Life Expectancy):

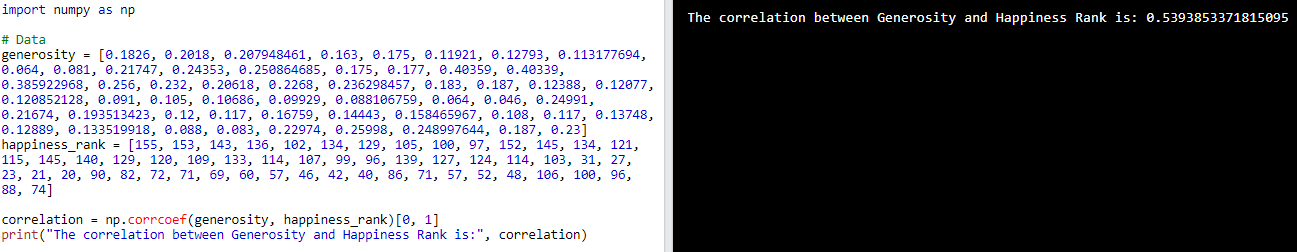
The correlation is:



0.745 > 0.7, a strong negative correlation.

Generosity:

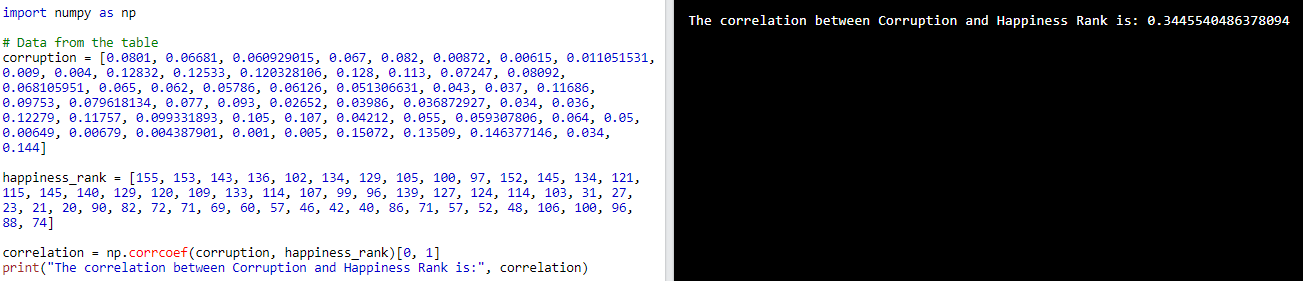
The correlation is:



0.5 < 0.54 < 0.7, a moderate positive correlation. That means that this factor negatively affects the happiness level for this type countries at a moderate rate.

Perception of Corruption

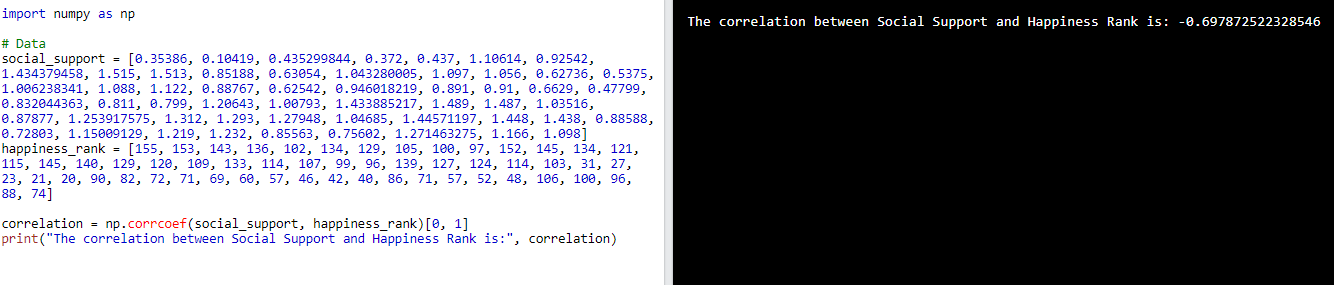
The correlation is:



0.3 < 0.344 < 0.5, a weak positive correlation.

Social Support:

The correlation is:



0.5 < 0.697 < 0.7, a moderate negative correlation.

The analysis's findings imply that social support, health, and indicators like GDP per capita are significantly correlated with how happy a nation is. The connections between these elements and happiness, nevertheless, are intricate and occasionally difficult to understand. It is important to keep in mind that correlation does not imply causality and that there can be additional underlying causes and cultural variations that affect the overall happiness score. To fully comprehend the challenges of comparing and evaluating happiness levels across nations, more investigation and analysis are required.

**Conclusion**

By discerning overall trends in happiness ranks, we aimed to understand the broader patterns and changes in global happiness levels during the years 2015-2019. This comparative analysis provided valuable insights into the general direction of happiness, stability or fluctuations in ranks, regional variations, and long-term trajectories. The identification of overall trends contributes to our understanding of the dynamics of happiness at a global scale and forms the foundation for further analysis and exploration in the project.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | GDP per Capita | Freedom | Health | Generosity | Corruption | Social Support |
| 1-st Type | -3 | -3 | -3 | (-)0 | -3 | -2 |
| 2-nd Type | -2 | 1 | -2 | 1 | -1 | -1 |
| 3-rd Type | -3 | (-)0 | -3 | 2 | 1 | -2 |

0 – very weak/no correlation 1 – weak correlation 2 – moderate correlation 3 – strong correlation

Positive sign 🡪 positive correlation (negatively affects the happiness level);

Negative sign 🡪 negative correlation (positively affects the happiness level)

It is crucial to remember that these results offer a general picture and should be understood in light of the distinctive socioeconomic and cultural aspects of each country. However, the investigation demonstrates the intricate interactions between distinct variables and their effects on the happiness rank, providing important insights for comprehending and advancing wellbeing in diverse nations.

**Can the countries shift from one type to another? Observations:**

*Strongly correlated variables:* The GDP per capita, health, corruption, and social support all exhibit negative correlations with countries in the first group, which are distinguished by stable rankings. A country in this category could transition to another kind if any of these characteristics significantly improves or deteriorates. For instance, a country in the first category may move into the third category of nations with rising happiness overall if its GDP per capita and level of health both significantly rise.

*Factors that have little or no correlation:* Countries, where the level of happiness is overall decreasing, have factors that show weak or no correlations with the happiness rank. While factors like freedom, generosity, corruption, and social support may have some influence, their correlations are not significant enough to consistently impact the happiness rank in these countries. Therefore, shifts from this type to another may occur due to changes in other factors or external circumstances that have a stronger impact on overall happiness levels.

*Multiple factors interaction:* The correlation coefficients sometimes may show the intricate interplay between several components. If numerous factors work together to determine a nation's happiness ranking and change to any one of these criteria may result in changes to another or to both. For instance, a country of the second type (overall decreasing) that simultaneously improves its GDP per capita, social support, and health while also seeing an increase in perception of corruption may move up to the first type of stable rankings.

*External variables and context:* When determining a country's happiness ranking, it's vital to take into account both external influences and particular contextual features. Political stability, environmental circumstances, cultural values, and social dynamics are only a few examples of the variables that might have a big impact. A country's type may change as a result of changes to these external elements since they have an impact on the underlying factors.

Overall, it is important to note that these explanations are conjectural based on the discovered correlation coefficients and that in actuality, there are many factors that affect a country's happiness ranking. Shifts between types could also be influenced by additional, unmeasured elements and how they interact (like external variables and context).

**Applications**

There might be various multiple applications of the provided analysis of the World Happiness Report across the years 2015-2019, but the significant ones are:

Governments and policymakers can use these findings to identify the crucial traits that promote happiness and well-being in various types of countries. They can use this information to create programs and regulations aimed at increasing happiness across the board in their own nations. To buck the trend, for instance, officials can focus on addressing the specific causes, such as declining life expectancy or increased corruption, if a country is categorized as having decreasing happiness.

Comparative research. This strategy can be used as a starting point by academics and researchers to conduct studies on happiness and well-being in other countries. By categorizing countries according to their happiness rankings, researchers can better understand the dynamics and factors that contribute to variations in happiness levels. This analysis could serve as a springboard for further research into the social, economic, and cultural variables that affect happiness.

Worldwide advancement. International organizations and development organizations can utilize this information to guide their projects and programs in diverse countries. These companies can construct targeted campaigns by understanding the factors that influence happiness and the patterns of development in order to address specific issues and improve well-being. For example, activities aimed at boosting social support, reducing corruption, or enhancing healthcare systems can be prioritized in countries where happiness levels have been discovered to be dropping.

**Faced Challenges**

During our team's initial foray into using Tableau for data analysis, we encountered several noteworthy challenges. One of the most prominent hurdles was the steep learning curve associated with the platform. Understanding the intricacies of Tableau's interface, terminology, and workflow required dedicated time and effort. We found ourselves investing in external learning resources, such as tutorials and online courses, to gain proficiency and navigate through the various features effectively.

Connecting and integrating different data sources within Tableau also presented difficulties. Establishing connections to databases from Kaggle required troubleshooting and occasional assistance. As there are no predefined connectors for Kaggle in Tableau (like for Google BigQuery). Furthermore, integrating multiple data sources and performing complex data blending demanded a deeper understanding of Tableau's capabilities.

Data preparation posed another significant challenge. Ensuring our datasets were properly structured and cleaned for analysis demanded meticulous attention to detail. Merging disparate data sources (for example over one parameter like country), addressing missing values, and rectifying data quality issues proved to be time-consuming tasks that required careful consideration.

Designing effective visualizations proved to be another noteworthy challenge. While Tableau offers a wide range of visualization options, selecting the most suitable chart types, arranging visual elements, and conveying insights clearly through visuals required a firm grasp of data visualization principles. It took time and experimentation to develop a sense of effective visualization design and apply it consistently across our analyses.

Performance optimization emerged as a challenge as our datasets and analyses grew in size and complexity. Generating visualizations, executing calculations, and interacting with dashboards sometimes led to slower performance or unresponsiveness. We had to delve into Tableau's performance optimization techniques, such as data aggregation and filtering, to address these issues effectively.

Sharing and collaboration within Tableau posed additional challenges. Understanding the various options for sharing visualizations, managing permissions, and deploying dashboards to different platforms or audiences required a learning curve. We encountered some difficulties while trying to streamline the sharing and collaboration process within our team.

Despite these challenges, we were able to overcome them through perseverance, utilizing Tableau's extensive documentation and community support, and seeking guidance from experienced Tableau users. Over time, our team gained proficiency in using Tableau, enabling us to analyze datasets more efficiently and produce impactful visualizations.

**Each member’s documented investment of their effort into the project:**

1. Yuldashev Komiljon - 12230144 🡪 came up with the technique of using Tableau for analyzing the overall trends of the countries and then identifying specific correlations among factors. Supported everyone with the understanding of Tableau and its functionality, therefore among the team he has the most awareness of Tableau and how to use it (consequently, documented the faced challenges while using Tableau). He did research on the stable ranking countries.
2. Begon Eliott – 12230078 🡪 came up with the code that uses NumPy to calculate the correlation between a specific factor and a happiness rank. He did the research on the 2-nd type of countries; documented the Applications section
3. Coste Felix – 12230200 🡪 came up with the idea of categorizing the countries into 3 types based on the overall trends. He did the research on the last type of countries (3-rd type); documented the observations of the Findings and Conclusions section, which is the section about the shifting of countries from one type to another.
4. Jafar Saidov – 12194931 🡪 investigated the nature of correlations, and appropriately categorized each correlation by its strength and positiveness/negativeness. Offered the most suitable data source obtained from Kaggle with a variety of parameters and multiple records, therefore contributed mostly to the section about Data Source in the report file.

Each of the team members significantly contributed to the writing and design of the report file (mostly 🡪 Introduction, Methodology and Analytical Strategies sections), and making the final presentation file. They collaboratively did that work by supporting each other in creating the report file and the presentation.