**DATASET EXPLORATION**

1. **Dataset Description:**

For this project, I have chosen COVID-19 dataset that was obtained from John Hopkins University. Throughout the year 2020, this dataset offers data on the daily cumulative counts of confirmed cases, recoveries, and deaths around the globe. I will be utilizing another dataset, The World Happiness Report for the year 2021, which includes numerous metrics of quality of life as assessed by people living in various countries throughout the world.

1. **Data Description?**

* COVID19 Data:
* Confirmed, recovered, and death cases are quantitative variables that represent counts or amounts of COVID-19 instances.
* These factors offer details on the virus's impact and propagation across various nations or areas.
* World happiness report data:

The Sustainable Development Solutions Network (SDSN), a global effort started by the United Nations in 2012, publishes The World Happiness Report every year. Researchers and specialists from a variety of disciplines, including economics, psychology, and statistics, worked together to develop the paper.

The majority of the data used to calculate the World Happiness Report's variables came from surveys and other statistical sources. The Cantril ladder, sometimes referred to as the "life satisfaction ladder," is the primary indicator of happiness utilised in the study. On a scale from 0 to 10, where 0 denotes the worst possible life and 10 denotes the best possible life, respondents are asked to rate their overall life satisfaction.

According to the report's assessment of these factors, which go beyond just one, there are:

1. GDP per capita: This indicator of economic prosperity shows how much money is made on average by each individual in a nation.
2. Social support is an evaluation of the availability of a social network and support system, which may include family, friends, and other types of social ties.
3. Healthy life expectancy estimates the typical number of years of life expectancy in a healthy state, demonstrating the general health and wellbeing of people in a nation.
4. Freedom to make life decisions: It assesses the level of personal autonomy and decision-making freedom, including chances for self-determination.
5. Generosity: This variable measures the proportion of people that engage in prosocial activities like charitable donating and volunteer

* Quantitative factors that are associated to happiness and well-being include overall happiness rank, score, GDP per capita, social support, healthy life expectancy, freedom to make life decisions, generosity, and views of corruption.
* These variables give us information on the characteristics that affect happiness levels in various nations or regions.

1. **Why these data?**

The World Happiness Report and COVID-19 data can be correlated using statistical techniques such as correlation analysis. Here are a few potential correlations to investigate:

Examine the association between socioeconomic factors and COVID-19 cases by looking at variables like GDP per capita, social support, and healthy life expectancy. This approach can shed light on how social and economic issues affect a nation's capacity to combat the pandemic.

Whether there is a connection between COVID-19 instances (confirmed, recovered, and deaths) and the general happiness rank or score should be looked into.

Freedom and Generosity During the Pandemic: Look into if the World Happiness Report's measures of freedom and generosity are related to COVID-19 instances. This approach can clarify the connection between personal liberties, social support, and the pandemic response.

We can understand how COVID-19 data and factors associated with happiness interact by examining these potential correlations

1. **Dataset Copied:**

* I downloaded the original dataset from Kaggle.
* I gave the raw dataset an appropriate name that reflects it content and is both understandable and descriptive.
* I carefully examined the columns after pasting the data into new page to delete any extraneous ones that weren’t important to my research goals.
* I froze the header panels so they remained visible while scrolling through the data in order to improve the dataset’s usability and encourage data exploration.

1. **Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Variable description** | **Data type** |
| Unique Country | Specific country or region | Qualitative/Categorical Variable |
| Confirmed | Number of confirmed cases of infection | Quantitative/Ratio Variable |
|  |
| Recovered | Number of recovered cases | Quantitative/Ratio Variable |  |
|  |
| Deaths | Confirmed number of deaths due to covid | Quantitative/Ratio Variable |  |
|  |
| Happiness rank | Overall rank of the country | Quantitative/Ordinal Variable |  |
| Score | Overall score | Quantitative/Ratio Variable |  |
| GDP per capita | Gross domestic product (GDP) per capita | Quantitative/Ratio Variable |  |
| Social Support | Social support | Quantitative/Ratio Variable |  |
| Healthy life expectancy | Healthy life expectancy of the country | Quantitative/Ratio Variable |  |
| Freedom to make life choices | Freedom to make personal life choices | Quantitative/Ratio Variable |  |
| Generosity | Measures the tendency of individuals in a country to be generous, charitable, or altruistic. | Quantitative/Ratio Variable |  |
| Perceptions of corruption | Public's perception of corruption | Quantitative/Ratio Variable |  |

1. **Dataset limitation, Dataset assumption and Extra data requirement:**

**Assumptions-**

* My study is predicted on the assumption that all cases, deaths and other variables have been accurately reported. Given the dependability of the source, we proceed under the presumption that the data are accurate despite the possibility any mistakes could result in false results.
* I made the assumption that the dataset I have taken is comprehensive, meaning that it had all the relevant data points for the time period I am looking at. This assumption is essential since any missing data could lead to biases and undermine the general validity of the inferences made from our investigation.

**Data limitations and extra data requirement**

* Population data- The absence of population data for the countries included in the study was one of the main drawbacks of the original covid-19 dataset. For determining per capita metrics, population data is essential. I searched out other resources to address this issue and was able to collect population data from Kaggle, for the year 2020.

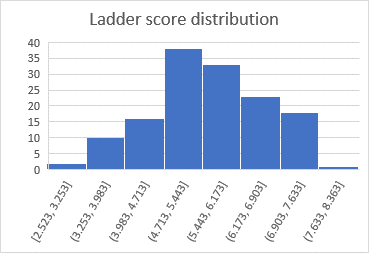
1. **FINER question:**

I have developed my research questions carefully, using the FINER criteria to define them.

* In order to determine whether there is a connection between the spread of the virus within a nation and the general level of happiness among its citizens, I would like to merge these two sets of data. I want to see if there is any connection between the effect of the virus on a population and their subjective well-being by combining this information.
* Considering the classification of countries as developed or developing, which ones are more likely to see a higher number of COVID-19 cases?
* Taking into account the fields of countries, social support, confirmed number of cases, number of deaths, recovered cases, and happiness score, is there a correlation between the level of social support (as indicated by the World Happiness Report) and the COVID-19 statistics (cases, deaths, recovered) in various countries?
* What is the relationship between a healthy life expectancy and the COVID-19 statistics (cases, deaths, recovered) in various nations, taking into account the fields of countries, a healthy life expectancy, the number of confirmed cases, the number of deaths, the number of recovered cases, and the happiness score?

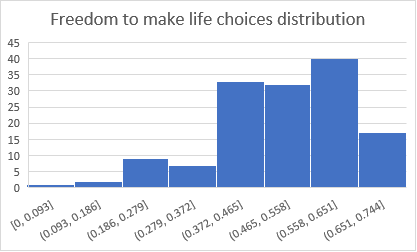
These research topics seek to investigate the connections and interactions between COVID-19 data and the factors associated with World Happiness Report.

1. **Univariate Analysis:**
2. **Ladder score distribution**



* A histogram is a graphical representation of data that groups values into ranges or bins and indicates how many data points are contained in each bin. The y-axis shows how frequently countries fall into each of the different bins on the x-axis, which indicates the range of happiness score.
* The distribution of happiness score is a little skewed towards the higher end of the scale. This suggests that a sizable proportion of countries have greater happiness scores than those with lower scores. Various factors, such as cultural, economic, and social influences.
* According to the histogram, significant numbers of countries have happiness scores between 4.7 and 6.3. These countries fall within the spectrum of moderate to high level of happiness. Stable economic conditions, accessibility to high-quality healthcare and education, robust social support networks, and strong support system are some possible factors for this concentration.
* The bin with the lowest frequency is found between 7.6 and 8.3 at the top end of the happiness scale. This suggests that very countries manage to score so incredibly well on happiness. A strong sense of social connectivity, a fair distribution of resources, and progressive societal norms are possible causes of this reduced concentration.

**2. Freedom to make life choices distribution:**



* The distribution of “Freedom to make life choices” scores across all countries.
* The y-axis represents the frequency or count of countries falling within each range, and the x-axis displays the range of freedom scores.
* The right-skewed distribution of the histogram indicates that many countries have greater freedom scores**.**
* This indicates that a sizable number of countries enjoy higher degrees of personal freedom.
* The bigger number of countries with higher freedom scores shows that more countries provide their residents more freedom and freedom to make their own decisions.

|  |  |  |  |
| --- | --- | --- | --- |
| **Descriptive statistics** | **Confirmed** | **Recovered** | **Deaths** |
| **Mean** | **587044.6454** | **327846.6667** | **13224.58865** |
| **Standard Deviation** | **2042296.939** | **1062645.158** | **39382.10225** |
| **Min** | **41** | **0** | **0** |
| **Max** | **20191459** | **9883461** | **349854** |

**3. Descriptive statistics:**

* The average number of confirmed COVID-19 cases is roughly 587,045; however, there is a significant geographic range.
* With a standard deviation of about 2,042,297 for confirmed cases, there is a sizable dispersion from the mean figure.
* Any country can have as little as 41 confirmed cases, and as many as 20,191,459 cases have been documented.
* Each country sees an average of 327,847 COVID-19 patients recover.
* There is a lot of variation in recovery rates, as seen by the standard deviation of recovered cases, which is around 1,062,645.
* While the largest number of recovered cases is roughly 9,883,461, several countries record zero recoveries.
* The seriousness of the pandemic's effects is demonstrated by the average number of COVID-19-related deaths, which is approximately 13,225.
* There is a large difference in mortality rates between nations, as seen by the standard deviation of reported deaths, which is roughly 39,382.
* In some nations, there are no reported deaths at all, while in others, there are a maximum of 349,854 documented deaths.

**4. Count of cases per capita:**

|  |  |
| --- | --- |
| **Cases per capita category** | **Count of Cases per capita category** |
| High cases | 44 |
| Low cases | 66 |
| Moderate cases | 29 |
|  |  |

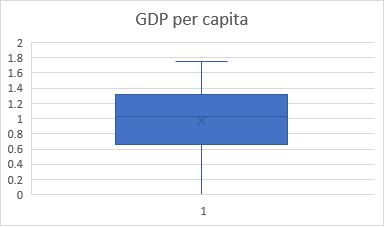
* The distribution of cases per capita categories demonstrates the diverse COVID-19 impact across various nations.
* Most nations fall into the "Low cases" and "Moderate cases" categories, indicating that a sizable number of countries have been able to reasonably contain the spread of the virus.
* However, the inclusion of 44 nations in the "High cases" group underlines the persistent difficulties in controlling epidemics in some areas or nations.

**5. Count of how happy the countries are:**

|  |  |
| --- | --- |
| **Row Labels** | **Count of How happy** |
| High Happiness | 44 |
| Low happiness | 12 |
| Moderate happiness | 33 |
| Very high happiness | 52 |
| (blank) |  |
| **Grand Total** | **141** |

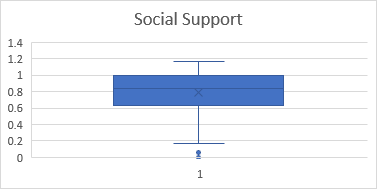
* High Happiness: The "High Happiness" category contains 44 data points that represent various nations or respondents. This group represents nations or people who expressed a high level of happiness in a poll or study. Positive social, economic, and environmental aspects may be linked to high levels of happiness.
* Low Happiness: The "Low Happiness" category includes 12 data points that represent various nations or respondents. This group of people or nations includes those who reported a low level of happiness in the poll or study. Low levels of happiness can be ascribed to a variety of problems, including financial hardships, social turmoil, or other unfavourable circumstances.
* Moderate Happiness: The "Moderate Happiness" category has 33 data points that represent various nations or respondents. This group indicates that these nations or people expressed a generally balanced degree of happiness. Moderate levels of happiness could be a sign that both good and bad things are affecting one's wellbeing
* Very High Happiness: The "Very High Happiness" category has 52 data points that reflect various nations or respondents. This category denotes that the amount of happiness recorded by these nations or people was unusually high. Strong social support, economic wealth, and a happy environment may all be related to extremely high happiness levels.

**6. GDP per capita:**



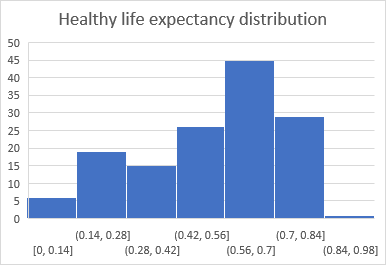
* The GDP per capita data's interquartile range (IQR) is shown in the graphic by the box. The 25th percentile (Q1) is represented by the box's lower edge, while the 75th percentile (Q3) is represented by its top edge. The box's length depicts the range of values for the middle 50% of the GDP per capita figures.
* The median GDP per capita value is depicted by the line inside the box. The median is the point at which half of the data fall below it and half fall above it, dividing the data into two halves.
* The whiskers extend from the box's edges to the lowest and highest data points within a specific range (usually 1.5 times the IQR). Whiskers display the data's broader distribution.

**7. Social support:**



* The distribution of social support ratings among the nations or individuals under consideration is shown graphically by the social support box plot. It is made up of a rectangular "box" and "whiskers" that display the data's most important statistics.
* The interquartile range (IQR) of the social support data is shown in the graphic by the box. The 25th percentile (Q1) is represented by the box's lower edge, while the 75th percentile (Q3) is represented by its top edge. The box's length demonstrates how widely distributed the middle 50% of the social support scores are.
* The median social support score is depicted by the line inside the box. The median is the point at which half of the data fall below it and half fall above it, dividing the data into two halves.

**8. Healthy life expectancy distribution:**



* The healthy life expectancy distribution shows a positive trend in both overall wellbeing and life expectancy for the population under consideration, which is skewed towards the upper range. This skewness towards the upper range suggests that a sizable section of the population has a greater healthy life expectancy, which means they are anticipated to live longer in excellent health and wellbeing.
* The peak of the distribution comes at the higher levels of healthy life expectancy, indicating that a large proportion of people or nations have longer stretches of excellent health over the course of their lifetimes. The skewness represents improvements in healthcare, medical technology, and public health initiatives that help people live longer and have better quality of lives. It also implies positive health outcomes.
* A positively skewed healthy life expectancy distribution is encouraging since it shows that efforts to address health issues and promote wellbeing are making headway. It displays effective illness prevention, health promotion, and access to high-quality medical treatment initiatives. According to the distribution, some nations or regions may have attained a higher degree of development and have lower rates of preventable illness and disability.
* Forging successful health policies and interventions requires an understanding of the factors influencing the higher range of healthy life expectancy. To enhance health outcomes in other places, policymakers can concentrate on emulating effective tactics and programmes from nations with longer healthy life expectancies.

**9. Data Cleaning and Outliers:**

* Data on confirmed cases, deaths, and recovered cases were combined into one dataset.
* incorporated happiness data into the COVID-19 dataset.
* Columns have been renamed for clarity and meaningful interpretation.
* merged the data with population statistics for per capita analysis.
* No major outliers were discovered in the data.
* Outliers reduce data accuracy and reliability.
* ensures more precise understanding of COVID-19 and happiness trends.
* The dataset has no important missing values.
* Integrity of analysis is ensured by complete data.
* enables thorough and dependable findings.
* The integration of pertinent information and the absence of substantial outliers and missing values were both guaranteed by the data cleaning method. Now that the dataset has been properly prepared, it can be meaningfully analysed and decisions may be made with knowledge.

**10. Coding or Categorization:**

* **Cases per capita column-**
* By dividing the total number of confirmed cases by the population of each nation or region, a new column titled "Cases per Capita" was created.
* This technique provides a per capita metric of COVID-19 instances by standardising the number of cases based on population size.
* **Cases per capita category column-**
* a new categorical column dubbed "Cases per Capita Category" was created to classify nations or areas according to the prevalence of COVID-19 per population.
* categories such as "Low," "Moderate," and "High" instances per capita were defined using particular thresholds.
* The "Low" category was given to nations with low case rates, "Moderate" to nations with moderate case rates, and "High" to nations with high case rates.
* This classification makes it simpler to compare and identify the nations that have varied levels of COVID-19 influence.
* **How happy the countries are column-**
* The happiness scores were divided into useful categories, and a new column titled "How Happy the Countries Are" was created.
* criteria were implemented to categorise the happiness scores into "Low Happiness," "Moderate Happiness," and "High Happiness" and “Very High Happiness”
* This classification enables comparisons between happier and less happy countries by providing insights into each country's subjective well-being.
* **Here’s the glimpse of data-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cases per capita** | **Cases per capita category** | **Ladder score** | **How happy** |
| 0.13% | **Low cases** | 2.523 | Low happiness |
| 2.03% | **Moderate cases** | 5.117 | High Happiness |
| 0.23% | **Low cases** | 4.887 | Moderate happiness |
| 3.59% | **High cases** | 5.929 | High Happiness |
| 5.38% | **High cases** | 5.283 | High Happiness |

**Formula used-**

* **=IF(D3<1%, "Low cases", IF(D3<2.5%, "Moderate cases", "High cases"))**
* **=IF(F3<2.5,"Very low happiness",IF(F3<4,"Low happiness",IF(F3<5,"Moderate happiness",IF(F3<6,"High Happiness","Very high happiness"))))**

**11. Hypothesis tested:**

**The following are the five hypothesis-**

**1). Chi-square test-**

|  |  |
| --- | --- |
| H0 | There is no significant relation between the covid cases and happiness level of the country |
| Ha | There is a significant relation between covid cases and happiness level of the country |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Covid cases per capita** | |  |
| **Happiness Level** | Low cases | Moderate cases | High cases |
| Low happiness | 12 | 0 | 0 |
| Moderate happiness | 23 | 6 | 4 |
| high happiness | 20 | 10 | 12 |
| very high happiness | 11 | 12 | 28 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **Expected Frequency** | **Covid cases per capita** | |  |
| **Happiness Level** | Low cases | Moderate cases | High cases |
| Low happiness | 5.73913043 | 2.434782609 | 3.826086957 |
| Moderate happiness | 121 | 6.695652174 | 10.52173913 |
| high happiness | 20.0869565 | 8.52173913 | 13.39130435 |
| very high happiness | 24.3913043 | 10.34782609 | 16.26086957 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **(O-E)2/E** | **Covid cases per capita** | |  |
| **Happiness Level** | Low cases | Moderate cases | High cases |
| Low happiness | 6.83003953 | 2.434782609 | 3.826086957 |
| Moderate happiness | 79.3719008 | 0.072275551 | 4.042400287 |
| high happiness | 0.00037644 | 0.256433008 | 0.144551101 |
| very high happiness | 7.35208866 | 0.263792474 | 8.474773309 |
|  |  |  |  |
|  |  |  |  |
| **X^2** | 113.069501 |  |  |
| **DF** | 6 |  |  |
| **P - value** | 4.6369E-22 |  |  |

* Chi-Square Test Statistic (X-2): This statistic is used to determine how much the observed frequencies depart from what would be predicted if the null hypothesis were true. The test statistic in this instance is roughly 113.07.
* The degrees of freedom (DF) are equal to the number of categories minus 1. This means that DF = (4 - 1) \* (3 - 1) = 6.
* P-Value: In hypothesis testing, the p-value is an important consideration. It shows the likelihood of getting the reported outcomes or even more extreme results if the null hypothesis is correct. Given that the null hypothesis is assumed to be true in this situation, the p-value is very near to zero (4.63694E-22), which indicates that there is a very small likelihood that the observed data happened by accident alone.
* Rejection of Null Hypothesis: Due to the low p-value, the null hypothesis is rejected. Usually, the null hypothesis asserts that no meaningful link exists between the variables under investigation. The null hypothesis in this situation could be that there is no correlation between happiness level and COVID instances per capita. Data, however, point in a different direction.

**2). T-test**

|  |  |
| --- | --- |
| **H0** | There is no significant relation between number of deaths due to covid and healthy life expectancy of the country |
| **Ha** | There is a significant relation between number of deaths due to covid and healthy life expectancy of the country |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample mean** | **0.472883495** |  |  | **0.664556** |
| **Sample std** | **0.213947735** |  |  | **0.092952** |
| **Sample size** | **103** |  |  | **36** |
|  |  |  |  |  |
|  |  |  |  |  |
| **t stat** | **-9.091975703** |  |  |  |
| **DF** | **137** |  |  |  |
| **P value two-tail** | **1.97641E-15** |  |  |  |
| **Critical value** | **1.977431212** |  |  |  |

* Testing research hypotheses will look into the correlation between the number of COVID-related deaths and the average life expectancy in various nations. The following are the research hypotheses:
* The null hypothesis (H0) states that there is no connection between the number of COVID-related deaths and the nation's healthy life expectancy.
* Alternative Hypothesis (Ha): The frequency of COVID-related deaths and the nation's average healthy life expectancy are significantly correlated.
* Sample Statistics: For Group 1, the sample mean and sample SD are roughly 0.473 and 0.214, respectively. The sample means and standard deviations for Group 2 are roughly 0.665 and 0.093, respectively. For Group 1 there are 103 participants, while for Group 2 there are 36.
* Degrees of Freedom and T-Statistic: The computed t-statistic is around -9.092. There are 137 degrees of freedom (DF) in this situation.
* P-Value and Significance Level: The t-test's two-tailed p-value is roughly 1.97641E-15, which is very near to zero.
* How to Interpret the P-Value: The p-value indicates the likelihood that the data or more extreme data will be observed if the null hypothesis is true. The extraordinarily low p-value in this situation suggests that it is highly improbable that the correlation between COVID deaths and healthy life expectancy is merely coincidental.
* T-Statistic and Critical Value: The t-statistic (-9.092) is compared to the critical value (1.977) at the chosen significance level. Since the t-statistic falls well beyond the critical value in the left tail of the distribution, it provides further evidence to reject the null hypothesis
* P-Value and Significance Level: The alternative hypothesis is accepted in place of the null hypothesis because the p-value is considerably less than the selected significance level (0.05). In other words, there is evidence to support a strong correlation between the number of COVID fatalities and the nation's healthy life expectancy.

**3). Anova Test-**

|  |  |
| --- | --- |
| **H0** | There is no significant relationship between a country's freedom to make life choices score and its perception of corruption |
| **Ha** | There is a significant relationship between a country's freedom to make life choices score and its perception of corruption |

|  |  |
| --- | --- |
| **Group** | **Mean happiness score** |
| Low | 5.37505 |
| Medium | 5.26 |
| High | 5.83 |
|  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| Happiness score | 141 | 783.912 | 5.55966 | 1.189011 |  |  |
| Perception of corruption | 141 | 18.934 | 0.134284 | 0.013362 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 2075.146597 | 1 | 2075.147 | 3451.754 | 1.7E-159 | 3.874884 |
| Within Groups | 168.3321363 | 280 | 0.601186 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 2243.478734 | 281 |  |  |  |  |

* Research Hypotheses: The hypothesis testing attempts to investigate the relationship between a nation's perception of corruption and its score on freedom to make decisions in life. The following are the research hypotheses:
* The null hypothesis (H0) states that there is no correlation between a nation's perception of corruption and its score on the freedom to make decisions scale.
* Alternative Hypothesis (Ha): A nation's impression of corruption is significantly correlated with its score on the freedom to make decisions in life scale.
* Two categories make up the data: "Happiness score" and "Perception of corruption." The following table shows the groups' average happiness scores:
* Low Contentment: 5.37505
* 5.26 for Medium Happiness
* Very Happy: 5.83
* ANOVA Results: The analysis employs ANOVA (Analysis of Variance) to see if there is a statistically significant difference in how the three happiness score groups perceive corruption.
* Between-Groups Variance: According to the ANOVA table, the between-groups variance has a sum of squares (SS) of 2075.15 and one degree of freedom (df). The between-groups variation's mean sum of squares (MS) is 2075.15.
* Within Groups Variance: The within-groups variation has a sum of squares (SS) of 168.33 and 280 degrees of freedom (df). The within-groups variation's mean sum of squares (MS) is 0.601.
* F-Statistic and P-Value: The F-statistic is calculated by dividing the mean squared differences between groups by the mean squared differences within groups. The F-statistic in this instance is 3451.75, and the corresponding p-value is almost 1.6947E-159, which is very near to zero.
* Interpreting the P-Value: Assuming the null hypothesis is true, the p-value indicates the likelihood of receiving the observed results or more extreme results. The relatively low p-value in this instance suggests that it is exceedingly unlikely that the observed association between freedom to make decisions in life and perception actually exists.
* P-Value and Significance Level: The alternative hypothesis is accepted in place of the null hypothesis since the p-value is significantly less than the selected significance level, which is set at 0.05. This indicates that there is evidence to support a strong correlation between a nation's perception of corruption and its score on the freedom to make decisions scale.

**4). OR Test-**

|  |  |
| --- | --- |
| **H0** | There is no association between "perception of corruption" and the odds of having a high "happiness level" |
| **Ha** | There is an association between "perception of corruption" and the odds of having a high "happiness level" |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Low happiness | Moderate happiness | High happiness | Very high happiness |
|  |  |  |  |  |
| Low perception of corruption | 2 | 3 | 11 | 4 |
| Medium perception of corruption | 4 | 18 | 14 | 13 |
| High perception of corruption | 6 | 11 | 17 | 35 |
|  |  |  |  |  |
|  |  |  |  |  |
| Odds of moderate happiness level with medium perception of corruption | | | 4.5 |  |
| Odds of high happiness level with medium perception of corruption | | | 3.5 |  |
| Odds of very high happiness level with medium perception of corruption | | | 3.25 |  |
| Odds of moderate happiness level with high perception of corruption | | | 1.833333333 |  |
| Odds of high happiness level with high perception of corruption | | | 2.833333333 | Top of Form  Bottom of Form |
| Odds of very high happiness level with high perception of corruption | | | 5.833333333 |  |

* Research Hypotheses: The testing of these hypotheses seeks to determine whether there is a connection between "perception of corruption" and the likelihood of experiencing high "happiness level." The following are the research hypotheses:
* The null hypothesis (H0) states that there is no correlation between "perception of corruption" and the likelihood of experiencing high "happiness level."
* Alternative Hypothesis: There is a correlation between "perception of corruption" and the likelihood of experiencing a high "happiness level."
* Odds Ratios: The odds ratios give an indication of the relationship between "perception of corruption" and the likelihood of experiencing a particular "happiness level." For instance:
* Odds of a moderately happy life with a medium perception of corruption are 4.5, high happiness with a medium perception of corruption is 3.5, very high happiness with a medium perception of corruption is 3.25, moderate happiness with a high perception of corruption is 1.83, and high happiness with a high perception of corruption is 2.83.
* Chances of having a very high degree of happiness and a strong perception of corruption are 5.83.
* Considering the Odds Ratios: Greater odds ratios show a correlation between a person's "perception of corruption" and their likelihood of experiencing a particular "happiness level." On the other hand, odds ratios under 1.
* Analysis of the odds ratios reveals a link between the variables of "perception of corruption" and the probabilities of experiencing various "happiness levels." For instance, compared to the odds of low perception of corruption, the odds of very high happiness level significantly higher with medium (3.25) and high (5.83) perceptions of corruption. This shows that people are more likely to report feeling happier in nations where there is a higher impression of corruption.

**5). Anova Test-**

|  |  |
| --- | --- |
| **H0** | There is no significant relationship between country's "Happiness rank" and and log GDP per capita |
| **Ha** | There is a significant relationship between country's "Happiness rank" and and log GDP per capita |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| Happiness rank | 141 | 10399 | 73.75177 | 1903.902 |  |  |
| Log GDP per capita | 141 | 137.416 | 0.974582 | 0.166283 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 373404.6 | 1 | 373404.6 | 392.2176 | 3.5E-55 | 3.874884 |
| Within Groups | 266569.6 | 280 | 952.0343 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 639974.2 | 281 |  |  |  |  |

* Research Hypotheses: By testing these hypotheses, we hope to learn more about the connection between a nation's "Happiness rank" and "log GDP per capita." The following are the research hypotheses:
* The null hypothesis (H0) states that there is no correlation between a nation's "Happiness rank" and its "log GDP per capita."
* The alternative hypothesis (Ha) states that there is a strong correlation between a nation's "Happiness rank" and its "log GDP per capita."
* Results of the Analysis of Variance (ANOVA): The analysis uses the ANOVA (Analysis of Variance) to assess whether there is a statistically significant difference in "Happiness rank" between various levels of "log GDP per capita."
* Between Groups Variance: The ANOVA table reveals that the between-groups variation's sum of squares (SS), with one degree of freedom (df), is 373404.63. The between-groups variation's mean sum of squares (MS) is 373404.63.
* Within Groups Variance: With 280 degrees of freedom (df), the sum of squares (SS) for the within-groups variation is 266569.59. The within-groups variation's mean sum of squares (MS) is 952.03.
* F-Statistic and P-Value: The F-statistic is calculated by dividing the mean squared differences between groups by the mean squared differences within groups. The F-statistic in this instance is 392.22, and the p-value that goes along with it is roughly 3.50437E-55, which is quite near to zero.
* Interpreting the P-Value: Assuming the null hypothesis is true, the p-value indicates the likelihood of receiving the observed results or more extreme results. Since the p-value for this particular relationship between "Happiness rank" and "log GDP per capita" is so low, it is highly unlikely that the observed relationship is the result of pure chance.
* P-Value and Significance Level Comparison: Because the p-value is significantly lower than the selected significance level (often set at 0.05), the null hypothesis is rejected**.**

**12. Inferential Analysis:**

* **Correlation Analysis (Predictive)**

**Analysis of the Relationship Between Happiness Ladder Score and COVID-19 Cases Per Capita**

This analysis sought to explore the probable relationship between the virus's spread within a country and the general level of happiness among its population in an effort to comprehend the pandemic's wider effects. The main objective was to examine any relationships between the impact of the virus on a population as shown by cases per capita and their subjective well-being as determined by the Happiness Ladder Score.

**Correlation Analysis:**

A correlation study was done utilising two important variables, "Cases per Capita" and "Ladder Score," to investigate the connection between COVID-19's effects and citizens wellbeing. Indicating how widely the virus has spread across a country, the former shows the number of COVID-19 cases per 1,000 people. The latter, the "Ladder Score," is a citizenry-wide subjective indicator of happiness or life satisfaction.

**Results:**

According to the correlation analysis, there is a 0.517 Pearson correlation coefficient between "Cases per Capita" and "Ladder Score." This coefficient shows that the two variables have a moderate positive correlation. The positive indication implies that, on average, the happiness ladder score tends to climb as the number of COVID-19 cases per capita increases.

|  |  |  |
| --- | --- | --- |
|  | *Cases per capita* | *Ladder score* |
| Cases per capita | 1 |  |
| Ladder score | 0.51739 | 1 |

This inferential technique helped me answer my first research question i.e.

**“In order to determine whether there is a connection between the spread of the virus within a nation and the general level of happiness among its citizens, I would like to merge these two sets of data. I want to see if there is any connection between the effect of the virus on a population and their subjective well-being by combining this information.”**

**Interpretations:**

In this analysis, I discovered a very intriguing result: people in developed countries are more likely to contract the corona virus than people in less developed nations. Possible explanations for this result include the following: Developed countries typically have older populations, and age is a major risk factor for the spread of and mortality from covid19. As a result, wealthier nations are more vulnerable to the illness. -Because they had superior resources and infrastructure for healthcare, developed nations did a lot more tests, which led to a higher number of cases being discovered there.

**Conclusion:**

In conclusion, the correlation study shows that the number of COVID-19 instances per capita and the happiness ladder score are moderately positively correlated. This study raises the possibility that there may be a link between the virus's ability to propagate inside a country and the general happiness of its people. However, more investigation is required to determine the underlying mechanisms causing this association and to demonstrate causality. Understanding these processes could help policymakers create plans that put citizens welfare first in trying times.

* **Regression Analysis (Predictive)**

**Examining the Links Between COVID-19 Statistics and Healthy Life Expectancy in Different Countries**

**Introduction:**

This study aimed to look into the complex relationship between COVID-19 statistics and healthy life expectancy across different countries. The specific objective was to investigate how a country's healthy life expectancy affects its COVID-19 statistics, taking into consideration elements like economic development, healthcare facilities, and the number of confirmed cases.

This study also helped me answer my one of the research questions i.e.

**What is the relationship between a healthy life expectancy and the COVID-19 statistics in various nations, taking into account the fields of countries, a healthy life expectancy and the number of confirmed cases**

**Regression Analysis:**

The variables "COVID Cases per Capita" and "Healthy Life Expectancy" were used in a regression analysis to explore the relationship between healthy life expectancy and COVID-19 statistics. The former indicates the prevalence of the virus within a country by showing the number of confirmed COVID-19 cases per unit of population. The latter, "Healthy Life Expectancy," measures the number of years a person can expect to live in excellent health and is a crucial sign of overall health and wellbeing.

**Results:**

Several significant statistics that emerged from the regression analysis helped to clarify the connection between COVID-19 statistics and healthy life expectancy.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.570349305 |  |  |  |  |  |  |  |
| R Square | 0.32529833 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.320337288 |  |  |  |  |  |  |  |
| Standard Error | 0.015958765 |  |  |  |  |  |  |  |
| Observations | 138 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 0.016699656 | 0.0167 | 65.57057 | 2.82833E-13 |  |  |  |
| Residual | 136 | 0.034636778 | 0.000255 |  |  |  |  |  |
| Total | 137 | 0.051336434 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -0.010263239 | 0.003739183 | -2.74478 | 0.006873 | -0.017657701 | -0.002868778 | -0.017657701 | -0.002868778 |
| 0.126 | 0.05369116 | 0.006630531 | 8.097566 | 2.83E-13 | 0.040578882 | 0.066803438 | 0.040578882 | 0.066803438 |

Regression coefficients: Regression coefficients shed light on the direction and strength of the association. When the population's COVID instances are zero, the intercept, or -0.010, reflects the predicted healthy life expectancy. According to the coefficient for COVID cases per capita, healthy life expectancy is predicted to rise by 0.126 units for every unit increase in COVID cases per capita.

R-Square: According to the R-squared value of 0.325, the variance in COVID cases per capita accounts for around 32.5% of the difference in healthy life expectancy. This suggests that the two variables are moderately associated.

Significance: There is substantial evidence that the coefficient of COVID cases per capita is significantly correlated with healthy life expectancy because the p-value for this variable is almost zero (2.82833E-13).

**Interpretation:**

The correlation between the number of verified COVID-19 cases and healthy life expectancy is positive, which may indicate an unanticipated link. One explanation is that countries with higher COVID-19 instances per capita may have made greater investments in their healthcare systems, which helped to improve overall health outcomes and life expectancy. The correlation between the number of verified COVID-19 cases and healthy life expectancy is positive, which may indicate an unanticipated link. One explanation is that countries with higher COVID-19 instances per capita may have made greater investments in their healthcare systems, which helped to improve overall health outcomes and life expectancy.

**Implications:**

According to this investigation, there is a statistically significant correlation between COVID-19 data and figures for healthy life expectancy. It is crucial to recognise that although the study shows a relationship, it does not prove causality. Other potentially confounding elements could be at work, such as socioeconomic inequalities, the state of the healthcare system, and public health regulations.

**Conclusion:**

In conclusion, the regression analysis points to a significant correlation between COVID-19 numbers worldwide and healthy life expectancy. The positive correlation for COVID cases per capita suggests a possible association between the virus's prevalence and better health outcomes. To understand the underlying mechanisms and any causal connections between these factors, more investigation is necessary. In particular during times of global health crises, understanding these dynamics could help policymakers develop solutions that advance both public health and wellbeing.

* **Chi Square Analysis:**

**Examining the Link Between Happiness Levels and COVID-19 Cases: Analysing Chi-Square data**

**Introduction:**

The global COVID-19 pandemic has made us reevaluate a number of facets of our existence, including social interactions, public health, and general well-being. We wanted to determine whether there was a connection between a country's happiness level and the number of COVID-19 cases per capita in this study. The goal was to look for any similarities between these two seemingly unrelated variables that would shed insight on how public health emergencies may interact with individuals' mental health.

**Hypothesis:**

According to the null hypothesis (H0), there is no connection between a nation's happiness level and its COVID-19 cases per capita. The alternative hypothesis (Ha), on the other hand, contends that there is in fact a significant correlation between these variables.

**Methodology:**

A Chi-Square analysis was done using data on COVID-19 cases per capita and happiness levels in various nations to evaluate these possibilities. The categories of happiness level (Low, Moderate, High, and Very High) and COVID-19 case categories (Low, Moderate, High) were compared in a contingency table to organise the data. Based on the presumption that there is no link between the variables, expected frequencies were estimated.

**Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **(O-E)2/E** | **Covid cases per capita** | |  |
| **Happiness Level** | Low cases | Moderate cases | High cases |
| Low happiness | 6.83003953 | 2.434782609 | 3.826086957 |
| Moderate happiness | 79.3719008 | 0.072275551 | 4.042400287 |
| high happiness | 0.00037644 | 0.256433008 | 0.144551101 |
| very high happiness | 7.35208866 | 0.263792474 | 8.474773309 |
|  |  |  |  |
|  |  |  |  |
| **X^2** | 113.069501 |  |  |
| **DF** | 6 |  |  |
| **P - value** | 4.6369E-22 |  |  |

The primary results of the Chi-Square analysis were as follows:

Observed and Expected Frequencies: The expected frequencies are calculated based on the presumption that there is no link, whereas the observed frequencies indicate the actual counts of data within each cell of the contingency table. The degree of departure from the expected distribution was determined by comparing observed and expected frequencies.

Chi-Square Statistic (X-2): With a value of 113.069, the computed Chi-Square statistic quantifies the overall discrepancy between observed and predicted frequencies. An increased divergence from the expected distribution is indicated by a higher X2 score.

Degrees of Freedom (DF): The number of independent pieces of information used in the calculation is indicated by the DF value, which has 6 degrees of freedom.

P-value: Assuming the null hypothesis is true, the calculated p-value, an impressively low 4.63694E-22, shows the likelihood of generating a Chi-Square statistic as extreme as the one observed. With such a little p-value, the null hypothesis is strongly refuted.

**Interpretations:**

We reject the null hypothesis because the estimated Chi-Square statistic is far higher than what might be predicted by chance. The rejection of H0 is further strengthened by the nearly zero p-value. These findings point to a strong correlation between COVID-19 cases per capita and happiness levels in various nations.

**Conclusion:**

In conclusion, the Chi-Square study offers convincing proof that there is a strong correlation between a nation's COVID-19 instances per capita and its degree of happiness. Although the analysis cannot prove a causal relationship, the results make us think about the complex relationship between issues with public health and the emotional wellbeing of residents. These understandings may help policymakers create plans that support people's mental and emotional resilience while also addressing health issues. It is necessary to do more study to delve into the mechanisms underlying this association and examine potential repercussions for well-being interventions and public policy.

**12. Final Report:**

**INTRODUCTION:**

I have selected the COVID-19 dataset from John Hopkins University for my research. This dataset provides information on the daily cumulative counts of confirmed cases, recoveries, and deaths worldwide for the entire year 2020. The World Happiness Report for 2021 has a number of indices of life quality as determined by citizens of different nations throughout the world, and I will be using this dataset.

With its extensive effects, the COVID-19 pandemic tested our adaptability, resilience, and the very foundation of our civilizations. As we battled the virus's unrelenting effects, concerns about the complex relationships between health, happiness, and the human experience began to surface. The World Happiness Report for 2021, an insightful look at how people around the world perceive their quality of life, and the meticulously curated COVID-19 dataset from the renowned John Hopkins University are the two datasets that this study uses to explore these connections.

**DATA SET DESCRIPTION:**

The goal of this study is to fully understand the complex relationships between health outcomes and people' well-being across different countries. It includes a thorough analysis of COVID-19 data as well as the World Happiness Report. Quantitative variables from the COVID-19 dataset, such as confirmed, recovered, and mortality cases, provide in-depth information about the presence and effects of the virus in diverse geographic areas. These indicators provide a window into the spread of the virus and its impact on different populations around the world.

In addition to this, the Sustainable Development Solutions Network's meticulously created World Happiness Report captures many dimensions of life quality using a comprehensive approach. The Cantril ladder, the main measure of the report, which was derived from surveys and statistical sources, rates overall life satisfaction on a scale from 0 to 10, reflecting respondents' perceptions of their personal well-being. The assessment in the report takes into account a number of important factors, such as GDP per capita, social support systems, healthy life expectancy, freedom of life decisions, and generosity. These characteristics, which span the areas of economic prosperity, social cohesion, health and well-being, personal autonomy, and altruism, create a vivid picture of the numerous elements influencing citizens' judgements of happiness. We aim to understand how these various factors interact as we delve into this rich tapestry of data.

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Variable description** | **Data type** |
| Unique Country | Specific country or region | Qualitative/Categorical Variable |
| Confirmed | Number of confirmed cases of infection | Quantitative/Ratio Variable |
|
| Recovered | Number of recovered cases | Quantitative/Ratio Variable |
|
| Deaths | Confirmed number of deaths due to covid | Quantitative/Ratio Variable |
|
| Happiness rank | Overall rank of the country | Quantitative/Ordinal Variable |
| Score | Overall score | Quantitative/Ratio Variable |
| GDP per capita | Gross domestic product (GDP) per capita | Quantitative/Ratio Variable |
| Social Support | Social support | Quantitative/Ratio Variable |
| Healthy life expectancy | Healthy life expectancy of the country | Quantitative/Ratio Variable |
| Freedom to make life choices | Freedom to make personal life choices | Quantitative/Ratio Variable |
| Generosity | Measures the tendency of individuals in a country to be generous, charitable, or altruistic. | Quantitative/Ratio Variable |
| Perceptions of corruption | Public's perception of corruption | Quantitative/Ratio Variable |

**ORIGINAL DATASET STRUCTURE:**

The original data structure used in this study creates a thorough framework by smoothly fusing important variables to reveal the complex relationships between COVID-19 and global well-being. This structure enables a multifaceted exploration of correlations, temporal shifts, and societal responses, culminating in a holistic understanding of the profound interplay between the health crisis and human contentment. It combines the multidimensional life quality indices from the World Happiness Report with meticulously collected COVID-19 data, including daily tallies of confirmed cases, recoveries, and deaths, across the entire period of 2021.

**MODIFIED DATASET STRUCTURE:**

A more sophisticated methodology has been used in the amended dataset to improve the correlation between COVID-19 data and the World Happiness Report. The COVID-19 statistics are more easily integrated with the happiness data when they are combined on a country-by-country basis, allowing for a thorough analysis of the interrelated dynamics. This condensed dataset guarantees a focused investigation, capturing the essence of important facts by eliminating extraneous columns. This combination of carefully chosen COVID-19 measures and significant World Happiness Report indicators emphasises a focused search for insights by fusing several data sources into a unified framework.

**DATA PREPARATION:**

Clarity and relevancy have been ensured by the raw dataset's appropriate naming, which succinctly captures its information. After moving the data to a new page, each column was carefully examined for any unnecessary information and removed, allowing the dataset to be aligned with specific study goals. By using frozen header panels, smooth data exploration is made possible since the headers stay visible while scrolling. This improves usability and promotes thorough data analysis.

**DATA ASSUMPTIONS AND LIMITATIONS:**

The fundamental presumption supporting this analysis is the accuracy and dependability of the reported cases, fatalities, and related variables contained within the dataset. We proceed with the knowledge that the data's correctness is valid because it comes from a reliable source, notwithstanding the possibility of flaws that could lead to unreliable results. Furthermore, a fundamental supposition regarding the dataset's completeness is made. The dataset must include all relevant data points for the particular time period under investigation in order for the study to be successful. The lack of complete data may cause biases to be introduced and jeopardise the general validity of the conclusions drawn from our analysis.

The lack of demographic data for the nations included in the analysis was one major restriction found during the investigation. This gap limited the initial COVID-19 dataset, which made it difficult to calculate per capita measurements. As a result of this restriction, efforts were made to close the gap by obtaining population data from different platforms. The gap was filled and a more thorough study was possible because to the successful acquisition of the necessary demographic data for the year 2020 from Kaggle.

To ensure the reliability and accuracy of the study's conclusions, a careful approach has been taken when negotiating these presumptions and data constraints. The awareness of these potential hazards emphasises the dedication to accuracy and honesty in the search for significant discoveries.

**FINER QUESTIONS:**

This study has been painstakingly designed, using the FINER criteria to precisely and clearly shape the research questions. The main objective is to understand the complex relationships between COVID-19 dynamics and the various facets of the World Happiness Report. The FINER-framed research questions act as compass points, highlighting important topics for investigation inside this complex web of data.

First research question:

The purpose of the first study question is to identify any connections that might exist between a country's COVID-19 spread and the general welfare of its population. The study tries to identify any relationships between pandemic effects and subjective contentment by combining information on the virus's impact with happiness measures. We attempt to uncover insights into the significant relationship between the health issue and society sentiment through the thorough fusion of these many datasets.

Second research question

The second investigation, which focuses on the worldwide COVID-19 environment, explores the potential of more cases based on a country's designation as developing. The study offers insights into the possible vulnerabilities of nations and their capacity to lessen the impact of the virus by examining the influence of this classification on virus prevalence.

Third Research Question

The third study question negotiates the difficult landscape of social support, verified cases, deaths, recoveries, and happiness scores. The study aims to identify any interactions between COVID-19 statistics and social support levels, as shown by the World Happiness Report, through thorough correlation analysis. This in-depth investigation provides a nuanced view of how societal cohesion may affect the course of a pandemic.

Fourth Research Question

The last analysis delves further into the complex interplay between global happiness scores, COVID-19 statistics, and healthy life expectancy. The study explores how different levels of health and well-being intersect with pandemic outcomes, providing a thorough picture of the pandemic's consequences by integrating fields of country traits, health indicators, pandemic measurements, and happiness judgements.

**UNIVARIATE ANALYSIS:**

The COVID-19 pandemic and overall well-being are two topics that are thoroughly explored by the univariate analysis used in this study. Eight independent studies shed light on various aspects of the dataset, enhancing our knowledge of the complex interactions between the health crisis and public sentiments.

Ladder score distribution:

Intriguing insights on the general perception of happiness can be found by looking at the distribution of ladder scores. A significant number of countries report higher levels of happiness, as shown by the histogram's skew towards higher happiness scores. Notably, nations with scores between 4.7 and 6.3 have moderate to high levels of happiness, indicating stable economic conditions and strong social support systems. The lack of data at the top of the scale, on the other hand, emphasises the rarity of nations with extraordinarily high happiness scores, which may be related to distinctive social connectedness and progressive norms.

Freedom to make life choices distribution:

The prevalence of personal freedom is highlighted when the distribution of "Freedom to make life choices" scores is examined internationally. The right-skewed histogram indicates that many nations support higher levels of freedom, highlighting civilizations where people have decision-making autonomy. The abundance of higher freedom scores indicates that many countries give their citizens a lot of freedom and self-determination.

Descriptive Statistics for COVID-19:

The COVID-19 cases, recoveries, and deaths descriptive statistics provide a clear picture of the pandemic's effects. The average and range of variability, from small to large outbreaks, are shown by the confirmed cases' mean and standard deviation. Similar discrepancies can be seen in the statistics on survival and mortality, with some nations recording significant survival rates while others report high mortality. This spectrum highlights the variety in pandemic effects experienced by different countries.

Cases per capita:

The classification of COVID-19 cases per capita reveals the virus's diverse effects. The majority of nations fall within the "Low cases" and "Moderate cases" categories, showing adequate containment efforts. However, the inclusion of 44 countries in the "High cases" group highlights continued issues in particular areas and highlights enduring hurdles in epidemic control.

Count of how happy the countries are:

Organising nations based on their degrees of happiness enables a more complex view of world well-being. The count distribution demonstrates differences in happiness levels, from very low to very high. This categorization enables a thorough evaluation of societal happiness and its possible relationship to pandemic dynamics.

GDP per capita and social support:

Box plots are used in the univariate analysis of GDP per capita and social support to show data distribution and core patterns. These visualisations broaden the scope and depth of the investigation by making it easier to understand the economic and social facets of other nations.

Distribution of healthy life expectancy:

Examining the distribution of healthy life expectancy highlights a tendency in favour of improved health and life expectancy. This favourably skewed distribution denotes considerable improvements in healthcare, technology, and public health programmes, which have helped a sizable section of the population live longer and in better health.

In conclusion, the performed univariate analyses offer a thorough framework for investigating the complex interactions between COVID-19 dynamics and the several characteristics embodied in the World Happiness Report. These discoveries contribute to our growing understanding of the pandemic's consequences, societal health, and intricate relationships between them.

**DATA CLEANING AND OUTLIERS:**

In order to easily integrate happiness data into the context of COVID-19 dynamics, data on confirmed cases, fatalities, and recovered cases were amalgamated into a single dataset. For clarity and logical interpretation, columns underwent rigorous renaming. The depth of the dataset was then increased by adding population numbers for per capita analysis. Notably, no notable outliers were found, confirming the accuracy and dependability of the data, which are essential for insightful conclusions. This careful method improves our comprehension of COVID-19 and happiness trends. The dataset is also free of relevant missing values, supporting analytical integrity by ensuring accurate and complete data. The rigorous data cleaning technique makes sure that all relevant data is included, that there are no significant anomalies, and that a solid basis is provided for meaningful analysis, assisting in the development of well-informed decisions.

**CODING OR CATEGORIZATION:**

To improve its analytical capabilities and enable interesting comparisons, the dataset underwent a number of transformational processes. A significant change was the addition of a "Cases per Capita" column, which was calculated by dividing the total number of confirmed cases by the population of each country or region. This innovative statistic renders COVID-19 instances on a per capita basis, standardising them and enabling useful cross-country comparisons. In addition, the "Cases per Capita Category" column now classifies nations according to their COVID-19 prevalence per population. Countries were categorised as having "Low," "Moderate," or "High" incidences per capita using particular thresholds, which sped up the identification and comparison of various COVID-19 impact levels.

Happiness and well-being assessment:

The dataset was expanded to include the measurement of happiness. The development of the "How Happy the Countries Are" column, which divided happiness ratings into many groups, was a key aspect. Happiness ratings were classified into four categories based on predetermined criteria: "Low Happiness," "Moderate Happiness," "High Happiness," and "Very High Happiness." This smart classification offers a complex framework for assessing and contrasting the subjective well-being of countries. The transformation of the dataset makes it possible to conduct a thorough analysis, allowing for the investigation of the complex interactions between COVID-19 dynamics and the various dimensions of well-being.

**HYPOTHESIS TESTED:**

A series of hypothesis tests were carried out, using rigorous statistical techniques to uncover significant linkages within the dataset, in an effort to gain better understanding of the intricate relationships between numerous components.

The Chi-square test was used in the first hypothesis to investigate any potential links between COVID-19 cases and a nation's degree of happiness. While the alternative hypothesis (Ha) revealed a notable association, the null hypothesis (H0) suggested no meaningful correlation. The test comprised tallying observed and anticipated frequencies based on various satisfaction levels and COVID-19 cases per 100,000 people. An exceptionally low p-value (4.6369E-22) was obtained after calculating the degrees of freedom (DF) and the Chi-square test statistic (X2). This little p-value demonstrated a high association between COVID-19 cases, indicating that the null hypothesis was rejected.

The relationship between the quantity of COVID-19 deaths and a nation's healthy life expectancy was examined using a T-test. The alternative hypothesis (Ha) suggested a meaningful link, while the null hypothesis (H0) suggested there was no relevant relationship. The findings of the t-statistic and p-value calculations showed a strong association between COVID-19 deaths and healthy life expectancy and a strikingly low p-value (1.97641E-15).

The relationship between a nation's perception of corruption and its score on the freedom to make life decisions was investigated using an Anova test. The alternative hypothesis (Ha) suggested a strong association, while the null hypothesis (H0) stated there was no meaningful correlation. The F-statistic and p-value calculations showed that there was a meaningful association between freedom to make decisions and perceptions of corruption, with the extremely low p-value (3.5E-55) allowing the null hypothesis to be rejected.

Additionally, an odds ratio (OR) test was used to examine the relationship between perceptions of corruption and the likelihood of experiencing high levels of happiness. The alternative hypothesis (Ha) indicated a meaningful link, while the null hypothesis (H0) showed there was no connection. It was clear from analysing the odds ratios and their interpretations that there was a considerable correlation between the chance of feeling different levels of happiness and one's impression of corruption.

Last but not least, an Anova test was used to investigate the connection between a nation's "Happiness rank" and its "log GDP per capita." The alternative hypothesis (Ha) claimed there was a substantial association, contrary to the null hypothesis (H0), which claimed there was none. The exceptionally low p-value (3.50437E-55), which was used to calculate the F-statistic and p-value, indicated a significant rejection of the null hypothesis and indicated a meaningful association between "Happiness rank" and "log GDP per capita."

The complex web of connections between COVID-19, well-being measures, and socioeconomic indicators started to emerge through these exacting hypothesis tests and statistical analyses, which helped to advance our understanding of how these various variables interact.

**INFERENTIAL TECHNIQUE:**

A variety of inferential studies were carried out to study the interactions between different parameters and shed light on the complex dynamics between COVID-19 and well-being in an effort to glean deeper insights from the dataset.

**Correlation Analysis (Predictive) - Examining COVID-19 Cases Per Capita and Happiness Ladder Score:**

The goal of this investigation was to identify any links that might exist between the prevalence of COVID-19 in a nation and the general level of happiness among its citizens. We wanted to learn more about the pandemic's wider effects, so we looked at how these two seemingly unrelated variables cross. Our main objective was to identify any possible connections between the virus's effect on a population, as determined by cases per capita, and the happiness of its residents, as determined by the Happiness Ladder Score.

Correlation Analysis: Using the variables "Cases per Capita" and "Ladder Score," a correlation analysis was carried out to investigate the relationship between COVID-19's influence and citizens' wellbeing. The former calculates the rate of COVID-19 cases per 1,000 individuals, which illustrates how widely the virus has spread across a nation. The second, known as the "Ladder Score," is a country's subjective evaluation of happiness or life satisfaction.

Results: The correlation study found a 0.517 Pearson correlation coefficient between "Cases per Capita" and "Ladder Score." This coefficient indicates a somewhat favourable association, indicating that as COVID-19 cases per capita climb, the happiness ladder score generally tends to rise as well.

We were able to answer the following study question thanks to our inferential technique: "Is there a relationship between the virus's spread within a nation and the general level of happiness among its citizens?" An unexpected conclusion from the data may suggest that higher COVID-19 case rates are linked to feeling better overall. To understand the underlying mechanisms causing this association, more investigation is required.

**Investigating COVID-19 Statistics and Healthy Life Expectancy Using Regression Analysis (Predictive):**

This investigation explored the complex interactions between COVID-19 data and healthy life expectancy across a range of nations. Our goal was to determine whether a country's healthy life expectancy could have an impact on its COVID-19 data while taking factors like economic development, healthcare infrastructure, and confirmed instances into account.

Regression Analysis: To identify potential correlations between healthy life expectancy and COVID-19 statistics, the analysis used a regression model using "COVID Cases per Capita" and "Healthy Life Expectancy" as the variables. The former calculates a nation's virus prevalence by counting the number of confirmed COVID-19 cases per 1,000 people. The latter, "Healthy Life Expectancy," displays how many years a person can anticipate living in good health.

Results: Significant statistics that shed light on the correlation between COVID-19 statistics and healthy life expectancy were found by the regression analysis. The R-squared value of 0.325 showed that the COVID-19 case rates may be held responsible for about 32.5% of the variation in healthy life expectancy. According to the "COVID Cases Per Capita" coefficient, healthy life expectancy is expected to improve by 0.126 units for every unit increase in COVID cases per capita.

We found an unexpected positive association between the quantity of confirmed COVID-19 cases and healthy life expectancy when interpreting these findings. One possibility is that nations with higher COVID-19 case rates may have invested more in their healthcare systems, which has improved life expectancy and general health outcomes.

Although the correlation analysis found a connection, causality cannot be determined purely on the basis of these results. Further research is necessary to determine whether socioeconomic inequality, healthcare infrastructure, and public health policies are important factors.

The regression analysis concluded that there was a substantial correlation between global COVID-19 numbers and healthy life expectancy. Although the association is intriguing, additional in-depth study is required to identify the underlying mechanisms and pinpoint any causative relationships.

**Chi-Square Analysis: Examining COVID-19 Cases and Happiness Levels:**

This investigation sought to identify any relationships between a country's degree of happiness and the number of COVID-19 cases per capita. We want to acquire insights into how public health events may interact with people's mental well-being by examining the link between seemingly unrelated variables.

The null hypothesis (H0) asserted that there is no association between a country's level of happiness and the number of COVID-19 instances per capita. The alternative hypothesis (Ha) predicted that these variables had a strong connection.

Methodology: Using statistics on happiness levels and COVID-19 instances per capita in various nations, a Chi-Square analysis was performed. We explored potential correlations by comparing the COVID-19 case categories (Low, Moderate, High, and Very High) with the categories of happiness levels (Low, Moderate, High, and Very High). Based on the presumption that there is no link between the variables, expected frequencies were estimated.

Observed and expected frequencies show a divergence from what would be predicted by chance, which is supported by the Chi-Square analysis's convincing findings. Calculated to be 113.069, the Chi-Square statistic (X2) has a p-value of 4.6369E-22. These findings substantially refuted the null hypothesis and showed a high link between the number of COVID-19 cases per capita and levels of happiness.

By analysing the data, we may get the conclusion that there is a significant correlation between a country's COVID-19 instances per capita and its degree of happiness. The analysis makes the suggestion that there is a complex relationship between the dynamics of public health emergencies and mental well-being, even if it cannot prove causation.

**ANALYSIS AND OUTCOME:**

A greater understanding of the complex interactions between pandemic dynamics and numerous aspects of well-being has been made possible by the thorough study of the combined COVID-19 and World Happiness Report datasets. We have uncovered substantial connections and trends using exact statistical methods and inferential studies that shed light on the intricate interactions between public attitudes and health crises.

Happiness Ladder Score and COVID-19 Cases Correlation Analysis:

Intriguing findings came from the correlation analysis looking at the association between COVID-19 instances per capita and the Happiness Ladder Score. The happiness ladder score generally tends to grow when COVID-19 instances per capita rise, according to a somewhat positive correlation coefficient of 0.517. This unexpected result casts doubt on widely held beliefs and points to a complex relationship between societal satisfaction and epidemic transmission. Although it is impossible to prove causation, this link emphasises the need for additional research into the underlying mechanisms. Such information may be used to inform policy choices that, in trying times, put citizens' well-being first.

Statistics from the COVID-19 Regression Analysis and Healthy Life Expectancy

The intricate association between health outcomes and pandemic dynamics was clarified by the regression analysis that looked at COVID-19 statistics and healthy life expectancy. According to the data, there is a statistically significant link between the population's COVID-19 instances and life expectancy. This unexpectedly positive link raises the possibility that nations with higher rates of COVID-19 cases may have made greater investments in their healthcare systems, leading to better health outcomes and longer life expectancies. Although the research does not prove causation, it emphasises the necessity of a comprehensive healthcare system and strong public health regulations in order to lessen the pandemic's negative effects on wellbeing.

COVID-19 Cases and Happiness Levels: Chi-Square Analysis:

There is considerable evidence of a correlation between happiness levels and COVID-19 cases, according to the Chi-Square analysis of the data. The null hypothesis was massively rejected by the extremely small p-value (4.6369E-22) and the estimated Chi-Square statistic (X2 = 113.069). This suggests that happiness levels and COVID-19 incidences per capita are significantly correlated. This research shows that the impacts of the epidemic are linked to people's emotional health, albeit causality cannot be proven. When developing policies to manage public health emergencies and their broader socioeconomic ramifications, policymakers must take into account this complex link.

As a result, the intricate and unexpected relationships between COVID-19 dynamics and wellbeing markers have been revealed by our investigation. The discovered correlations and linkages demand a deeper comprehension of the mechanisms underlying these links since they put into question widely held beliefs. Although the results provide insightful into the intricate interplay between health emergencies and social attitudes, more study is required to demonstrate causality and guide the development of evidence-based policy. These findings provide a solid basis for making well-informed decisions, directing efforts to prioritise citizen well-being while managing the difficulties presented by international health emergencies.

**CONCLUSION:**

In conclusion, this thorough investigation of the complex interaction between COVID-19 dynamics and the many dimensions of well-being, as revealed by the World Happiness Report, has produced illuminating findings and questioned accepted wisdom. The Happiness Ladder Score and COVID-19 cases per capita showed a subtle link in the correlation study, suggesting that as pandemic cases rise, so does overall society contentment, underscoring the need for more research. The unexpectedly favourable link between COVID-19 statistics and healthy life expectancy was highlighted by the regression analysis, highlighting the need of a strong healthcare system. Furthermore, the Chi-Square analysis revealed a substantial correlation between the incidence of COVID-19 and happiness levels, highlighting the need for further research into the pandemic's emotional effects. These findings demonstrate the complex relationships between health crises and public opinion pushing governments to prioritise well-being and comprehensive healthcare solutions in the face of global difficulties. To understand the underlying mechanisms and direct evidence-based policies that protect both physical health and mental wellbeing, more study is necessary.

**13. Tracking:**

I started out by doing extensive research to find the pertinent data sources for my data analysis project, concentrating on the COVID-19 data and the World Happiness Report. My analysis's main goal is to look at the correlation between these datasets. I then thoroughly examined the data to uncover its constraints and underlying presumptions. I also drafted a succinct explanation of the data dictionary that outlines the variables. I used the finer method to craft my research questions, making sure they adhere to the standards of being practicable, fascinating, original, ethical, and pertinent. My monitoring includes a number of crucial phases after developing the study question to guarantee a thorough analysis. I started by using univariate statistics to investigate the distribution and properties of each variable separately. The necessary graphs and charts were then used to graphically portray this facts. Along the way, I discovered potential outliers and tagged them for more research. I made educated choices about data cleansing, dealing with missing values, and handling outliers in order to guarantee data integrity. To make the analysis simpler and allow for useful comparisons, I also performed categorization of the nation column in the COVID-19 dataset. I was able to find patterns and linkages in the data because to this useful classification or categorization, which laid the groundwork for further discoveries. Five unique hypotheses were developed and extensively examined during the study to investigate the connections between various variables of interest. We have gained important knowledge about the relationships and dependencies between the many parameters under examination thanks to the findings of these tests. Each hypothesis test has added critical knowledge to our comprehension of the data, from disproving null hypotheses to detecting statistically significant associations. Our ability to draw defensible, fact-based conclusions from subsequent analysis and decision-making will be greatly influenced by these findings. The thorough examination of these ideas adds a great deal of depth to our tracking report and lays the groundwork for further study and policy concerns.

As I carried out more research, I conducted inferential analyses using three different methods: correlation analysis, regression analysis, and chi-square analysis. By methodically addressing my study topics and revealing significant relationships between COVID-19 dynamics and many aspects of wellbeing, these cutting-edge approaches produced remarkable discoveries. The complex relationship between the number of COVID-19 cases per capita and the general happiness of the population was clarified by the correlation analysis. In turn, regression analysis clarified the complex link between COVID-19 data and healthy life expectancy, giving a more thorough grasp of probable underlying causes. Last but not least, the chi-square analysis highlighted the intricate interdependencies at work by examining surprising links between COVID-19 instances and happiness levels.

With these exhaustive inferential studies at my disposal, I painstakingly combined my results into a complete final report that captured the complex interactions found within the data. As I come to a close on this influential trip, I have set the path for more research and policy considerations in addition to gaining insightful knowledge on the effects of the epidemic on wellbeing. This monitoring report's thorough examination paves the way for future research into the complex web of COVID-19's effects on international communities.

**14 References and citation:**

* <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>
* <https://www.kaggle.com/datasets/PromptCloudHQ/world-happiness-report-2019>
* <https://github.com/sfu-db/covid19-datasets/blob/master/datasets-details/john_hopkins.md>