Project Report

Working with Data/Communicating with Data

| **Happiness Data Analysis 2015-2019** |
| --- |

# Project description

| Dataset used from Kaggle- Happiness Data (years - 2015, 2016, 2017, 2018, 2019) using a set of surveys from 158 countries . The surveys aim to analyze data to investigate the ways in which elements of people’s working lives drive their wellbeing.  Interesting that each country is given a happiness score that is the sum of scores from the following 5 criteria are:   * Economy: real GDP per capita * Health: healthy life expectancy * Freedom: freedom to make life choices * Trust: perceptions of corruption * Generosity: perceptions of generosity   The countries are then ranked from most happy to least happy.  This ranking could be an indication of a country's progress, the operative word is 'could'.  The Happiness Score is derived simply by interviewing a random population sample from each country in a poll that “asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale.”  Using these 5 variables, we can construct a linear regression model which may help us predict the happiness score in these countries. We can then compare the predicted score to the actual score to observe how accurate our model is.  In this notebook, I will carry out an exploratory data analysis of the World Happiness Reports for years 2015-2019. I will also use the linear regression model to predict the happiness score for a selected set of data.  I will attempt to answer this question; What factors influence the happiness of citizens the most?  The steps include:  1.Data Understanding and Cleaning  2.Data Visualization and Analysis  3.Modeling and Prediction  4.Correlation between criteria’s  5.Graph Evaluation |
| --- |

# Accessing the data

| Data stored in Git hub -  <https://github.com/Nikitatamotia/Winter-Project->  Want to understand the variations in the happiness score and each of the 5 considered features across time. Have 5 years of data, and want to understand how world happiness has changed across time. With more research in healthcare, it makes sense that the life expectancy across the world should increase over the years.  Want to find interesting patterns and trends in the temporal data for different countries which can challenge our known facts about the world.  In doing so we will understand if the world's richest countries are also the world's happiest countries. We will also be able to dive deeper and understand countries that have the most/least freedom of choice as well as corruption. |
| --- |

# Describe the data

| Some of the columns are filtered across excel, as comparing data sets in python the column names have to be the same. Else, I couldn't run the codes in pythons to work with 5 years of data.  Removed headers are -  -Standard Error  -Lower Confidence Interval  -Upper Confidence Interval  -Whisker.high  -Whisker.low  -Family and Social support  -Region - This column was only in excel’s 2015,2016,2017 , and it wouldn’t match with 2018, 2019 data therefore it's removed too to match the columns heading across the excel’s.  -Dystopia Residual column filtered from excel 2015,2016,2017 as there no data available in 2018,2019  To match the headers across excel’s, changes have been made in 2018 & 2019 because those are odd ones out -  “Country or region” to “country” as region wasn’t available with countries data  “Overall Rank” to “Happiness Rank”  “Score” to “Happiness Score”  "GDP per Capita" to “Economy (GDP per Capita)”  "Freedom to make life choices" to "Freedom"  "Perceptions of corruption" and "Trust (Government Corruption)"  9 Data records consist columns across excels are-  Country, Happiness Rank, Happiness Score, Economy (GDP per Capita), Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, year  And there data type is object  Summarizing how data frame worked in python as -  -Downloaded the data set data.csv from the Resources tab. (Kaggle) and saved as a combined excel sheet.  <https://github.com/Nikitatamotia/Winter-Project-/blob/main/Combine_happiness.xlsx>  -Data Import as a pandas DataFrame  -Checked the number of observations  -Obtain the column headings and size of data  -Check the data type for each column.  -List the 5 happiest countries in five years  -List the 5 least happy countries in five years  -Checked if there are any missing values  -If necessary remove any observations to ensure that there are no missing values and the values in -each column are of the same data type.  -Saved new data frame with no “NULL” Value  <https://github.com/Nikitatamotia/Winter-Project-/blob/main/Combine_happiness%202.xlsx>  Obtain the mean, minimum and maximum value for each column containing numerical data.  <https://github.com/Nikitatamotia/Winter-Project-/blob/main/Project_activity_using_Python.ipynb>  In this analysis, I want to create visualizations that will help us explain happiness in different countries of the world. Number of countries in the report for each year is different. Want to create a unified set of countries for which have data across 5 years to perform analysis. Hence, would not perform analysis on countries with missing/incomplete information.  There is criticism of the World Happiness Report that the 5 features considered in formulating the happiness index are not created equally and are biased towards GDP of the countries. I want to analyze the data to prove/disprove this domain expert. |
| --- |

# Summarize the data

| Sharing the Data calculation in Python and Excel  Link:  <https://drive.google.com/drive/folders/1ygU3n14PO1lwkU7c5peBxsacDJXtsllY?usp=sharing> |
| --- |

# Describe the analysis you will do

| The strongest correlation in correlation matrix (other than the 1.00 for variables with themselves) is between GDP, and life expectancy. A larger GDP per capita leads to a larger health life expectancy, and vice versa.  But overall correlation analysis revealed a strong relationship among Happiness, Economy (GDP per Capita), Health (Life Expectancy), and Freedom.  Freedonm is also the third major factor affecting Happiness Score as per correlation matrix.  Used function “corr” to get the consolidated data in python   |  | **Happiness Rank** | **Happiness Score** | **Economy (GDP per Capita)** | **Health (Life Expectancy)** | **Freedom** | **Trust (Government Corruption)** | **Generosity** | **year** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **Happiness Rank** | 1 | -0.992052555 | -0.795110368 | -0.744099444 | -0.540150609 | -0.372780578 | -0.118290294 | -0.006594466 | | **Happiness Score** | -0.992052555 | 1 | 0.789719483 | 0.742842791 | 0.553364921 | 0.398417857 | 0.138142081 | 0.005946129 | | **Economy (GDP per Capita)** | -0.795110368 | 0.789719483 | 1 | 0.787751996 | 0.345614605 | 0.306307428 | -0.01364567 | 0.017229894 | | **Health (Life Expectancy)** | -0.744099444 | 0.742842791 | 0.787751996 | 1 | 0.341155211 | 0.250512013 | 0.010718147 | 0.130137697 | | **Freedom** | -0.540150609 | 0.553364921 | 0.345614605 | 0.341155211 | 1 | 0.459593278 | 0.29056363 | 0.011118432 | | **Trust (Government Corruption)** | -0.372780578 | 0.398417857 | 0.306307428 | 0.250512013 | 0.459593278 | 1 | 0.31891961 | -0.122264439 | | **Generosity** | -0.118290294 | 0.138142081 | -0.01364567 | 0.010718147 | 0.29056363 | 0.31891961 | 1 | -0.192415943 | | **year** | -0.006594466 | 0.005946129 | 0.017229894 | 0.130137697 | 0.011118432 | -0.122264439 | -0.192415943 | 1 | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

# Evidence of the analysis

| I have used Python , Graphs in excel to analysis the data of ?Happiness year 2015 to 2019 and saved in colab notebook.  Excel well explained the calculations with graphs and Data and its also saved in Colab notebook.  https://drive.google.com/drive/folders/1ygU3n14PO1lwkU7c5peBxsacDJXtsllY?usp=sharing |
| --- |

# Summary and conclusion

| <https://github.com/Nikitatamotia/Winter-Project-/blob/main/Project_activity_using_Python.ipynb>  Observations¶  1. GDP( 0.789) and Life Expectancy (0.742) are strongly correlated with the Happiness score.  2. Freedom correlates quite well with the Happiness score, however, Freedom correlates quite well with all data.  3. Government Trust still has a mediocre correlation with the Happiness score.  4. If trust is high, the distribution is all over the place. It seems to be just a negative indicator on a threshold.  5. It seems like the common criticism for "The World Happiness Report" is quite valid. A high focus on GDP and strongly correlated features such as life expectancy.  6. As common wisdom dictates - money makes you happy up to a certain threshold. However, having a good social network is important and family + friends tend to provide that. High life expectancy and health make you worry less about how you'll survive thus making you more happy.  GDP per capita, Freedom to make life choices and Life expectancy are great determinants of Happiness score and can be used to predict the future scores. However, this is not conclusive because unforeseen occurrences like pandemic, natural disasters and economic meltdown happen, even to the most stable countries so these scores can actually change.  Hence, verified the criticism of 'World Happiness Report' because happiness score is strongly correlated with GDP and Life Expectancy and to some extent with Freedom. It will be interesting to get data after 2019, clearly "Trust (Government Corruption)" demonstrates the effect of COVID-19 pandemic which has taken a serious toll on human life. This also impacted the normal life style and had a great impact on 'Trust in Government' factor and influenced Happiness score after years 2019.Therefore there are many factors who effect Happiness score from time to time and only "GDP", "Life expectancy" and "Freedom" can't be the influence factors only. |
| --- |