DATA ANALYSIS REPORT
ON
WORLD HAPPINESS

BY

RRAF

RUTH NDUTA

ROY KIMANI

ALLAN MUTISYA

FIONA CHERUTO

Business Understanding

Business Overview

Most people probably believe that happiness is in the eye of the beholder, an individual's choice, something to be pursued individually rather than as a matter of national policy. Happiness seems far too subjective, too vague, to serve as a touchstone for a nation's goals, much less its policy content. That indeed has been the traditional view.

Yet the evidence is changing this view rapidly. A generation of studies by psychologists, economists, pollsters, sociologists, and others has shown that happiness, though indeed a subjective experience, can be objectively measured, assessed, correlated with observable brain functions, and related to the characteristics of an individual and the society.

Asking people whether they are happy, or satisfied with their lives, offers important information about the society. It can signal underlying crises or hidden strengths. It can suggest the need for change. Such is the idea of the emerging scientific study of happiness, whether of individuals and the choices they make, or of entire societies and the reports of the citizenry regarding life satisfaction. This analysis summarizes the fascinating and emerging story of these studies.

Business Objective

Research question: Can Freedom to make life choices/arrangements and Gdp per capita improve The Happiness Score?

Null Hypothesis: People with great freedom of making life choices and a higher Gdp per capita will report greater happiness score than people with less freedom of making life choices and less Gdp per capita

Alternative Hypothesis: People with great freedom of making life choices and a higher Gdp per capita will not report greater happiness score than people with less freedom of making life choices and less Gdp per capita.

Assessing the situation

1. Resource Inventory: Datasets:World Happiness Report 2015 to 2020- [Link]
2. Software: Github, Google Collaboratory, Trello, Tableau.
3. Assumptions: The data provided is correct and up to date
4. Constraints: There are no constraints

Data Mining Goals

Our data mining goals for this project are as follows:

- To determine the major factor for the happiness around the world.

- To determine the relationship between Gdp per capita and the happiness score.
- To determine the relationship of happiness score to life expectancy.
- To determine the impact of freedom to make choices on the happiness score.

Data Mining Success Criteria

Our success criteria will be measured by the following criteria;

- We give an objective account of the data analysis, with insights majorly coming from the dataset.

Target audience

This project is aimed at informing all groups of people who love, want or need to travel to foreign countries eg government expatriates, international tourists, pilgrims, asylum seekers e.t.c about the happiness status of other countries and factors influencing happiness rate in these countries.

We also aim to inform governments in the respective countries on factors influencing the happiness rate in not only their countries but others too. This will help in better decision making, better policy formulation and benchmarking.

Data Understanding

Data Understanding Overview

The data for the project is retrieved from Kaggle.

- World Happiness Report Dataset - the primary report was distributed in 2015, the second in 2016, the third in 2017, and the fourth within the 2018 upgrade. The World Joy 2019, and the most current 2020 report which positions 155 nations by their bliss levels, was discharged.

Data Description

For the dataset provided, here is a brief description.

We had a total of 6 datasets for the years between 2015-2020.

The data sets were merged into 1 common dataset using the common fields available for all. After merging, we had a total of 935 rows and 10 columns which we proceeded to clean and then perform further analysis..

We performed exploratory data analysis and employed dimension reduction techniques in a bid to understand the data.

In order to perform hypothesis testing, we employed sampling, statistical distributions techniques that will be discussed below.

.

Verifying Data Quality

Only one column had missing data and was filled with 0. Outliers were detected and removed from the dataset using the interquartile range technique.

Data Preparation

These are the steps followed in preparing the data

Loading Data

The libraries to be used throughout the analysis were imported and the datasets loaded and merged using the columns common to all.

Cleaning Data

The merged dataset had no duplicates. We found only one column with missing data which was filled with 0. Outliers were detected and removed from the dataset using the interquartile range technique. After thorough cleaning of the data set we were left with 759 rows and 8 columns. Which was exported for further analysis. The columns are as described below.

Column name	Description
Country	Different states in the world
Happiness score	The national average of the responses to the main life evaluation question
Gdp_per_capita	gross domestic product acts as a metric for determining a country's economic output per each person living there.
Family	a group consisting of parents and children living together in a household.
Life expectancy	statistical measure of the average time a person is expected to live
Freedom	the power or right to act, speak, or think as one wants without hindrance or restraint,
Government Corruption,	dishonesty or criminal offense undertaken by a person or organization entrusted with a position of authority,
Generosity	the quality or fact of being plentiful or large.

Analysis

Univariate analysis

This involved value counts for each of the columns by plotting univariate distributions using kernel density estimation. A kernel density estimate (KDE) plot is a method for visualizing the distribution of observations in a dataset, analogous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions. From the graphs we can observe that the variables don't have a normal distribution.since for a normal distribution has a bell-shaped figure around the mean.

We also performed measures for central tendency and looked for skewness in the data. It was established that happiness score and generosity have a symmetrical distribution. Gdp per capita, family/social support, life expectancy, freedom to make choices they have a negative skew while the government corruption has a positive skew.

Bivariate Analysis

Our main interest is the Happiness score and how it varies when compared with different factors. We used a jointplot which draws multiple bivariate plots with univariate marginal distributions to compare the relationship between the happiness score and the Gdp per capita in the different countries. We found that as the Gdp per capita increases the happiness score increases.

We also did a kernel density estimate (KDE) plot which is a method for visualizing the distribution of observations in a dataset, analogous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions. The government corruption was clustered at a low point between -1 and 0.2, the happiness score is at a higher position between 4 and 8. Thus indicating that, when the government corruption is at a lower level, the happiness score for the citizens is high.

Our final visualization was a pairplot. This was used to observe the relationship between the happiness score and all the variables stated and observed a linear relationship in some variables such as the happiness score and the freedom to make choices, with family and with life expectancy.

Multivariate Analysis

This involves observation and analysis of more than one statistical outcome variable at a time. We decided to use the Factor Analysis technique which groups similar variables into dimensions so as to make a large dataset more manageable and more understandable by extracting maximum common variance from all variables and putting them into a common score. We chose factor analysis technique as our reduction technique because our dependent variable(happiness score) is a continuous variable.

We performing factor analysis, we make the following assumptions:

- There are no outliers in data.
- Sample size should be greater than the factor.
- There should not be perfect multicollinearity.
- There should not be homoscedasticity between the variables.

To evaluate the factorability or sampling adequacy of our dataset, we used Bartlett's test of sphericity which checks whether or not the observed variables intercorrelated at all using the observed correlation matrix against the identity matrix. We found the p-value equal to 0; proving that the test was statistically significant, indicating that the observed correlation matrix is not an identity matrix.

To confirm the factorability of our data we also performed the Kaiser-Meyer-Olkin (KMO) Test which determines the adequacy for each observed variable and for the complete model. The overall KMO for our data is 0.69, indicating that we can proceed with our planned factor analysis.

To choose the number factors, we employed the Kaiser criterion. We found that only 2-factors eigen values are greater than one. Hence, we only chose 2 factors (or unobserved variables).
After performing Factor analysis for these 2 factors, we observed that factor 1 has a high factor loading for : gdp per capita, life expectancy and family whereas factor 2 has a high factor loading for : government corruption and freedom. It is worth noting that 53% cumulative Variance is explained by the 2 factors.

Sampling

This is the process of selecting certain members or a subset of the population to make statistical inferences from them and to estimate characteristics of the whole population. The samplity methods can either be Probability Sampling Methods or Non Probability Sampling Methods. In our analysis, we will employ the probability sampling method which is a sampling technique where a sample from a larger population is chosen using a method based on the theory of probability.

First, due to its simplicity and lack of bias, we implemented the Simple Random Sampling technique which is a technique used to pick the desired sample size and for selecting observations from a population in such a way that each observation has an equal chance of selection until the desired sample size is achieved. We used a sample size (n=100) to randomly select 100 countries that was to be used to test the hypothesis. Upon selecting random samples and plotting a relationship between happiness score and the GDP per capita, the results agree with the general population. That is, as GDP per capita increases so does the happiness score.

Using the selected sample size, we went ahead and implemented Cluster Sampling. We wanted to cluster our dataset depending on Gdp per capita and freedom to make choices. Thus, we extracted the Gdp per capita and freedom to make choices and used them as our input while clustering using the python sklearn library. We obtained the predicted clusters for each observation by using the fit_predict method from sklearn, created a copy and a new series for the clustered points. We then went ahead and plotted a scatter

plot to display the Gdp per capita and freedom to make choices clustered points. Our clusters were along point zero on the cartesian plane. Hence, we did not use this cluster sample.

Statistical Distribution

Statistical distributions are functions that describe the relationship between observations in a sample space. In our case, we implemented this by trying to test the data for normality. The normality tests are statistical processes used to determine if a sample or any group of data fits a standard normal distribution. We did the Shapiro-Wilk test which evaluates a data sample and quantifies how likely it is that it was drawn from a Gaussian distribution. The function returns both the W-statistic calculated by the test and the p-value which were stat-0.98, p-0.000000000487 showing that the sample distribution was not Gaussian.

To confirm this we used a Q-Q plot which generates its own sample of the idealized distribution that we are comparing with, in this case the Gaussian distribution. Deviations by the dots from the line showed a deviation from the expected distribution, confirming that the distribution is not normal.

Hypothesis Testing

This is a systematic way to select samples from a group or population with the intent of making a determination about the expected behavior of the entire group.

Null Hypothesis: People with great freedom of making life choices and a higher Gdp per capita will report greater happiness score than people with less freedom of making life choices and less Gdp per capita

Alternative Hypothesis: People with great freedom of making life choices and a higher Gdp per capita will not report greater happiness score than people with less freedom of making life choices and less Gdp per capita.

Correlation Test

Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. Checking the assumptions for the test of independence which are:

- The two samples are independent
- The variables were collected independently of each other, i.e. the answer from one variable was not dependent on the answer of the other

The p-value is the likelihood of an event happening if the null hypothesis is true. If the p-value is less than the significance level, then we reject the null hypothesis whereas if the p-value is greater than the significance level then we fail to reject the null hypothesis.We used a significance level of 0.05 to test our hypothesis. Implementing Pearson's correlation to our analysis, gives a Pearson's correlation of 0.552. This proves that there is a correlation between the happiness score and the GDP per capita and also

between the happiness score and freedom since the p-value for both of these variables is 0.552 which is > .05 and therefore, we accept the null hypothesis.

Evaluation

For evaluation we used parametric and non-parametric tests to explain other supporting hypotheses formulated to help understand and evaluate the data in depth.

One-Sample Z-test

Ho: The GDP per capita and freedom to make choices population mean is equal to the GDP per capita and freedom to make choices sample mean

Ha: The GDP per capita and freedom to make choices population mean is not equal to the GDP per capita in freedom to make choices sample mean.

The Z-test statistic will tell us whether means of the sample and the population are different or equal. After getting the mean of the sample and the mean of the population, using the one sample z-test we looked for the p-value which we got as 0.0 and since 0.0 < .05, we rejected the null hypothesis ni. In conclusion, the mean of the sample and population is not the same.

Two-sample Z-test

The two sample Z-Test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different.

Ho = The GDP per capita and Freedom to make choices population mean have no statistical difference with GDP per capita and Freedom to make choices population means

Ha = The GDP per capita and Freedom to make choices population mean heve statistical difference with GDP per capita and Freedom to make choices population means

From the two-sample z test we get a p-value of 1.752944330811864e-236 which is less than the significance level of .05. Hence, we reject the null hypothesis and conclude that the sample mean and the population mean are different. Therefore, we reject the null hypothesis.

Point Estimation

Point estimation is the process of finding an approximate value of some parameter of a population from a random sample of the population. The population parameter can be the mean (average) of the population. Simply put, estimating a population parameter by using sample data.

The assumptions made are:

- The population follows a normal distribution
- Assign a population mean. For us to be able to understand the point estimator better, we are going to arbitrarily so that we can see how accurate a point estimator is.

Using our data, we created a sample (n = 100). We then looked for the mean of this sample and compared it to the true sample mean. We observed that based on a sample of 100 GDP per capita our estimator underestimates the true mean by -0.23476889738845907.

We can conclude that we can get a fairly accurate estimate of a large population from a fairly small subset. However, it is important to note that point estimates are never perfect. There will always be an error associated with the estimates.

Chi-squared goodness of fit

This type of test is used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.

H0: A variable follows a hypothesized distribution.

H1: A variable does not follow a hypothesized distribution.

Since the p-value (0.0) is less than 0.05, we alternative hypothesis. This means we have sufficient evidence to say that the true distribution of GDP per capita is different from the distribution of the Freedom to make choices.

Recommendations

1. Country governments should push for freedom to make choices since this influences the people's' will to give i.e. generosity.
2. For civil society, they need to keep the government in check to reduce corruption and focus on development to improve the citizens' life expectancy.

3. For the government to provide better social services to improve the citizens well being since this influences a positive growth on the gdp per capita and the life expectancy goes up as a result there of.

Summary

The above analysis was done using a python colab notebook and Tableau for visualization. For project management, we used Trello for the project management.

The links are as follows:

Python notebook [Link]
Trello [Link]
Tableau [Link]