Math231: Statistics **Project Report**

World Happiness Report

Hareem Raza	**22100277**
Momin Mehmood Butt	**21100286**
Muhammad Raahim Khan	**21100157**

Table of Contents

Abstract

This report aims to explore various *social, urban and national factors* that may affect the happiness of citizens around the globe. The layout of the report constitutes of the description of the methodology employed to analyze the data including dataset selection, data cleaning and preprocessing, exploratory data analysis and data modeling (using multiple linear regression) followed by a detailed analysis of the findings of each phase. The analysis focuses on determining the significance of the various factors in contributing to the prediction of happiness score and the quality of our fitted model. Multiple sub-questions are answered at each phase as we go along the way. However, the most important and interesting results of the analysis are summarized in the end along with plausible recommendations for the way forward.

Introduction

With the advent of COVID, a sense of despair and uncertainty has prevailed all around. The well-being and mental health of people is severely affected throughout the world. This hints towards the idea that as the world is progressing day by day, our focus is shifting more towards the never-ending advancements instead of the things that really matter. One of such things that we are overlooking is happiness. It is very important to know about the determining factors of happiness and how they affect the well-being of individuals so that we, as a society, can work together to improve them in order to form a happier environment. Therefore, in this report, we set out to look at the bigger picture and to ascertain the answer to the following research question:

"How does the happiness of citizens depend on various social, urban and national factors?"

Data Set

The World Happiness Report is a landmark survey of the state of global happiness. The survey is one of its kind which was initiated in 2012 and has been conducted annually across 155 countries since then. It measures several social, urban, national and personal factors which may affect happiness in one way or the other. Countries with a higher happiness score are considered to be happier and better for the well-being of their citizens while the one with lower scores are encouraged to work on citizen-centric policies. Overall, the report has gained immense global recognition because of its completeness and utility in interdisciplinary studies.

The data set that we will be using is The World Happiness Report from years 2015 through 2019. It is based on five CSV files i.e., one for each year which were combined to form a consolidated dataset. We intend to determine the extent to which GDP, family, health and life expectancy, freedom, absence of corruption and generosity contribute to the well-being of citizens across the years[1]. Background research about the data collection process for this survey revealed that the respondents were asked to take a Gallup World Poll. They had to rate the effect of each of the above-mentioned factors on their life out of ten with the worst possible score being 0 (i.e., the factor does not lead to happiness) and the highest score being 10 (i.e., the factor contributes heavily in their happiness)

[1] The details of feature selection and data integration are mentioned in the upcoming sections.

Method

In order to systematically find the answer to our research question, we divided our methodology into three phases namely data cleaning (and preprocessing), exploratory data analysis (EDA) and finally regression analysis.

Data Cleaning

Before diving into the analysis, it was important to understand and preprocess the data. We first imported relevant libraries that we need to use for our analysis. Next, we read the datasets and checked their dimensions to get an overview of the number of rows and columns we are dealing with.

```
df_2015 <- read.csv("Dataset/2015.csv")
df_2016 <- read.csv("Dataset/2016.csv")
df_2017 <- read.csv("Dataset/2017.csv")
df_2018 <- read.csv("Dataset/2018.csv")
df_2019 <- read.csv("Dataset/2019.csv")

# Printing dimensions (rows x cols)
cat("Dimensions of 2015 dataset: ", dim(df_2015), "\n")
cat("Dimensions of 2016 dataset: ", dim(df_2016), "\n")
cat("Dimensions of 2017 dataset: ", dim(df_2017), "\n")
cat("Dimensions of 2018 dataset: ", dim(df_2018), "\n")
cat("Dimensions of 2019 dataset: ", dim(df_2019), "\n")
```

```
Dimensions of 2015 dataset:  158 12
Dimensions of 2016 dataset:  157 13
Dimensions of 2017 dataset:  155 12
Dimensions of 2018 dataset:  156 9
Dimensions of 2019 dataset:  156 9
```

Fig. 1: R-output of dimensions of datasets

As we can see, the number of columns (features) vary in each dataset. We chose the ones which were common in all of them to later form a single bigger dataset and dropped the one's which were different. We then cleaned each dataset separately. A similar pattern was followed for each of them. We first check Null values in each data frame and dealt with them accordingly. Luckily, we found little to no null values in the data frames. In case we did, we imputed them appropriately. We then standardized the columns i.e., renamed them for ease of access, converted the names to lowercase and checked for appropriate datatypes. To confirm if the data cleaning was successful, we took a sample of 10 rows from each dataset and checked if the changes that we made were applied successfully.

After cleaning the data frames separately, we merged them into a single data frame called *df*. It consists of 782 rows and 10 columns as shown in the R output below:

```
# Merging all dataframes into one
df <- rbind(df_2015, df_2016, df_2017, df_2018, df_2019)
cat("Dimensions of combined dataset: ", dim(df), "\n")

cat("Columns of combined dataset (df):", colnames(df))
```

```
Dimensions of combined dataset:  782 10
Columns of combined dataset (df): country rank score gdp family health freedom corruption generosity year
```

Fig. 2: R-output of dimension and columns of combined dataset

We then checked if the dataset has any null values, we found a single null value in the corruption column. We had two options, either to drop the row or to fill the null value. We decided to do the latter and imputed the missing value with the mean of the corruption column.

```
cat('Null values in combined dataset (Before Cleaning):\n')
print(sapply(df, function(x) sum(is.na(x))))

# Imputing na value with mean
df$corruption[is.na(df$corruption)] <- mean(df$corruption, na.rm=TRUE)

cat('Null values in combined dataset (After Cleaning):\n')
print(sapply(df, function(x) sum(is.na(x))))

cat("Are there any null values in the dataframe?", is.null(df))
# No missing/na values are present in dataset now
```

```
Null values in combined dataset (Before Cleaning):
   country       rank      score        gdp     family     health    freedom
         0          0          0          0          0          0          0
corruption generosity       year
         1          0          0
Null values in combined dataset (After Cleaning):
   country       rank      score        gdp     family     health    freedom
         0          0          0          0          0          0          0
corruption generosity       year
         0          0          0
Are there any null values in the dataframe? FALSE
```

Fig. 3: R-output of Null values before and after cleaning

The False in the R output above indicates that there are no null or missing values in the dataset now.

Finally, we check if the datatypes of each column are appropriate or if we need to change anything. Let's have a look. (Fig.4 on the next page)

```
sapply(df, class)
```

```
    country        rank       score         gdp      family      health
"character"   "integer"   "numeric"   "numeric"   "numeric"   "numeric"
    freedom  corruption  generosity        year
  "numeric"   "numeric"   "numeric"   "numeric"
```

Fig. 4: R-output of datatypes of columns

All the columns except country have either numeric or integral datatypes. In fact, these are the columns (or contributing factors as mentioned earlier) which will be used in our analysis. Therefore, the current data types of the columns are fine for our subsequent analysis and no labeling or encoding is needed as such. This brings us to the second phase of our methodology i.e., exploratory data analysis (EDA).

Exploratory Data Analysis

With a dataset of 782 rows and 10 columns, it was important to start by analyzing what the data can reveal beyond formal modeling, regression analysis and other tests. Therefore, we performed Exploratory Data Analysis to visualize data and to summarize its characteristics. We also wanted to see if there are any evident trends so that we could structure our research question accordingly.

Features Analyzed

By now, we know that we had 10 features. One of them is happiness score, which is the dependent variable. Background research from the data source that we used and other websites enabled us to find out the exact meaning and purpose of the independent variables as well. The details of each one of them are given below:

Feature	Description
Score (i.e., happiness score)	A metric measured by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."
Country	Name of the country of each sampled person.
Rank	Rank of the country based on the happiness score.
GDP	The measure of a country's economic output.
Family	The level of social support that an individual has in the form of family, friends and relatives.
Freedom	The measure of the level of autonomy an individual has to make decisions or perform tasks.
Health	Life expectancy
Corruption	The level of corruption according to Corruption Perceptions Index (CPI). In this case, corruption refers to people's perception of whether lack of corruption leads to higher happiness.
Generosity	The amount of charity donated by an individual.
Year	The year in which the sampled people were surveyed.

Table 1: Descriptions of the features (columns) in the dataset

Correlation

Now that we knew what each feature represented, we wanted to see how these features were correlated. We calculated the Pearson Correlation (r) between different features in the form of a correlation matrix. We went through each one of them checked if the variables were positively linearly correlated, negatively linearly correlated or none.

```
# Correlation of all variables with score (happiness score)
corr <- cor(df_corr[, -1], method="pearson")
corr
```

```
                score         gdp      family     health   freedom corruption
score      1.0000000  0.78928400  0.64879934 0.74245574 0.5512580  0.3980267
gdp        0.7892840  1.00000000  0.58596553 0.78433757 0.3405110  0.3046554
family     0.6487993  0.58596553  1.00000000 0.57265026 0.4203608  0.1263331
health     0.7424557  0.78433757  0.57265026 1.00000000 0.3407451  0.2505034
freedom    0.5512580  0.34051099  0.42036084 0.34074513 1.0000000  0.4593896
corruption 0.3980267  0.30465539  0.12633312 0.25050338 0.4593896  1.0000000
generosity 0.1375777 -0.01456048 -0.03726161 0.01063811 0.2907055  0.3189051
            generosity
score       0.13757773
gdp        -0.01456048
family     -0.03726161
health      0.01063811
freedom     0.29070551
corruption  0.31890511
generosity  1.00000000
```

Fig. 5: Pearson correlation between features

For ease of interpretation, we then plotted the correlations in a color coded heatmap along with a reference scale to understand it. The resulting plot was as follows:

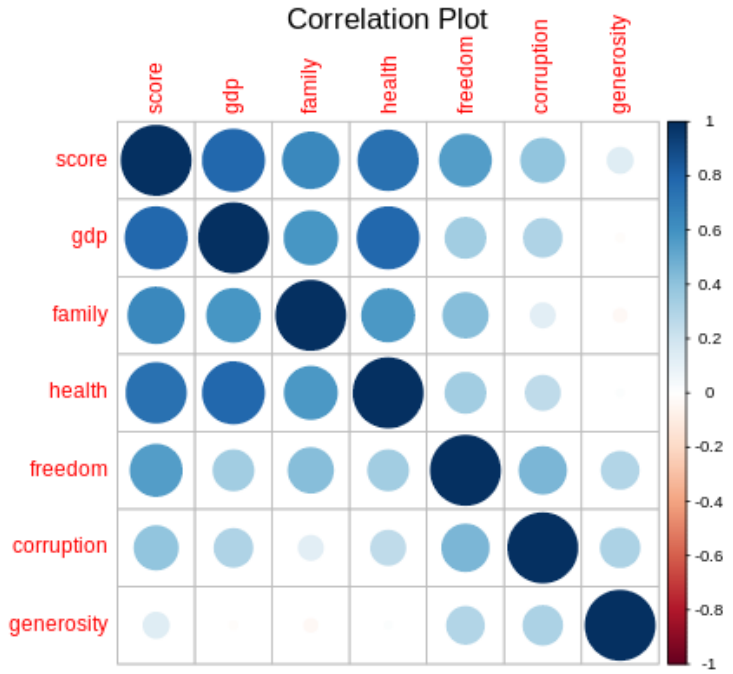

Fig. 6: Heatmap of correlation (r)

Evident from the scale on the right, darker plots between the features represented strong linear correlation (r closer to +1). Blue represented positive linear correlation (r > 0) while red represented negative linear relationship (r < 0). Obviously, the correlation of a feature with itself is 1 i.e., perfect correlation. Surprisingly, none of the features is negatively correlated with the other. This hints towards the fact that since all the predictor variables measuring happiness in one form of the other, increasing one of them does not result in a decrease in the other contributing factor. Score, our response variable, has the strongest positive correlations with the remaining features. GDP, family and health also have medium correlations with the remaining variables but generosity had weakest to no correlations with the other variables.

Since our primary goal was to predict happiness score and the factors affecting them, we decided to have a deeper look at the correlations of each possible contributing factor with the happiness score (the dependent variable) to have an idea about which factors are more likely to affect happiness score strongly. This time around we intended to look at the exact values of Pearson Correlation to make as accurate judgement as possible:

```
subset(corr, select = c(score))
```

```
                score
score      1.0000000
gdp        0.7892840
family     0.6487993
health     0.7424557
freedom    0.5512580
corruption 0.3980267
generosity 0.1375777
```

Fig. 7: Pearson correlation of features with score

Happiness score was positively linearly correlated with all the contributing factors under consideration. It was most strongly correlated with GDP followed by health and family. It had moderate correlation with freedom and a weak positive correlation with corruption and generosity.

Happiness Score

In the first two steps of Exploratory Data Analysis, we emphasized a lot on our response variable, score. Let's have a deeper look at it to understand its distribution in the dataset. This will make the interpretation of our results from regression modeling easier. As mentioned earlier, happiness score was the rating given by people to their happiness at the time of the survey. The maximum possible score was 10 (which means that the respondent is the happiest that he/she could be) and the minimum possible score was 0 (which means that the respondent was not happy at all).

The five-number summary of happiness score shows that happiness score is in the range 2.693 and 7.769. The mean happiness score is 5.379 and the median happiness score is 5.322. The minimum and maximum scores are reasonable enough i.e., they are neither too low nor too high. Since there are no outliers (as evident in the box plot below), it is clear that there are no countries with exceptionally high or low happiness score which need to be taken into account.

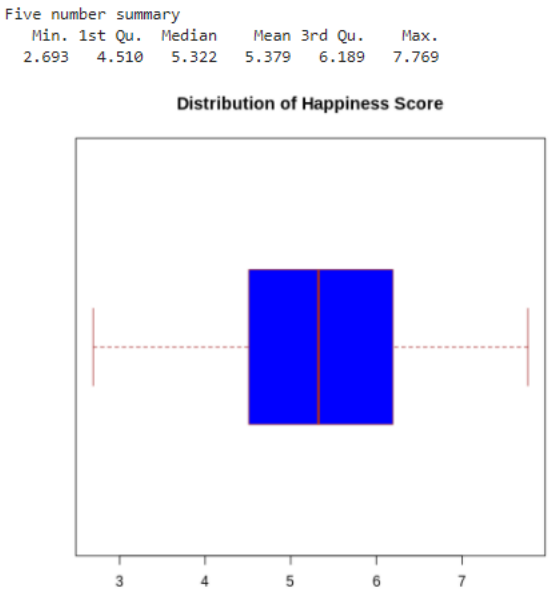

Fig. 8: Distribution of happiness score (Box plot)

Similarly, the distribution of happiness score below also shows that there is not even a single country which has a perfect happiness score. In fact, all the countries have a happiness score lower than 8. Most countries have a happiness score around 5 to 6 (which includes both the mean and median score). Therefore, overall, the distribution of happiness score is symmetrical.

Fig. 9: Distribution of happiness score (Density plot)

Country Wise Happiness Score

Although country-wise analysis of happiness score is beyond the scope of our research question, it was interesting to note the variation of happiness scores across different countries. Therefore, we decided to have a bird's eye view of the happiness scores across countries.

The world map (plotted in R) below shows color coded happiness scores of 156 countries in our data set. The colors on the left side of the spectrum i.e., shades of blue are less happy while the ones towards the right side of the spectrum i.e., shades of red and yellow are happier. Looking at the map, developed countries like Brazil, Russia United States and Canada are clearly happier than underdeveloped countries.

Fig. 10: Happiness score across the world with relevant scale

To have a more specific picture, the top 20 happiest countries and their scores were plotted which are as follows:

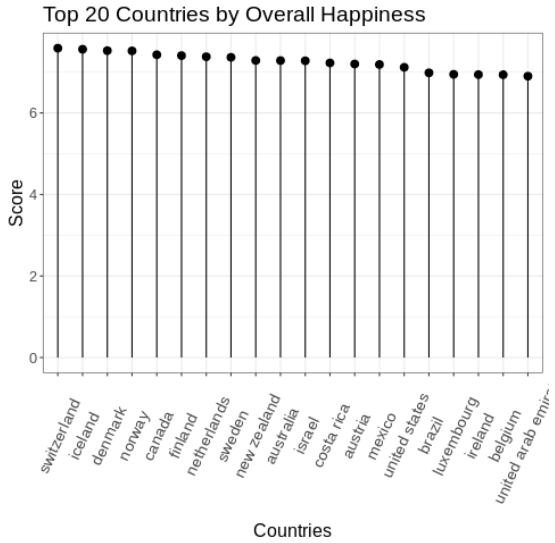

Fig. 11: Top 20 happiest countries (by happiness score)

Regression Analysis

After understanding the data thoroughly using visualizations and summaries of both the predictor and response variables, we moved on to understand the effect of the selected predictor variables on the happiness score through a model.

We used Multiple Linear Regression of the contributing factors against happiness score with an assumption that our predictor variables have a linear relationship with the response variable (as deduced from the Pearson correlation values calculated earlier).

The basic idea of the modeling was to find a linear combination of the levels of GDP, family, health, freedom, corruption, and generosity that best predicts happiness score. We identified significant variables (with p-value < 0.05) in the prediction of score. Important diagnostics for the model included determining coefficient of determination (R-squared and Adjusted R-squared) and making and analyzing ANOVA table. We also visualized our model and the residuals using Residual Plots and Q-Q plots the interpretations of which are discussed in the latter part of the report. Lastly, to make sure our model was not faulty (i.e., it neither underfitted nor overfitted), we also employed several measures – tested multicollinearity, data distribution and then drew final conclusions from the results.[2]

[2] The details of regression and its analysis are in the upcoming section.

Analysis

Multiple Linear Regression

As mentioned earlier, we employed Multiple Linear Regression to model the dependence of response variable (happiness score) on a set of predictor variables (GDP, family, health, freedom, corruption, and generosity). The relevant R code for this is shown in the excerpt below alongside the summary table of the regression model as the output.

```
# Multi-linear regression
multiple.regression <- lm(score ~ gdp + family + health + freedom
                          + corruption + generosity, data=df[, c(2:9)])

# Printing summary of variance table of the results
summary(multiple.regression)
```

```
Call:
lm(formula = score ~ gdp + family + health + freedom + corruption +
    generosity, data = df[, c(2:9)])

Residuals:
     Min       1Q   Median       3Q      Max
-1.93063 -0.32722  0.01712  0.35756  1.66431

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.17749    0.07981  27.283  < 2e-16 ***
gdp          1.14675    0.08276  13.857  < 2e-16 ***
family       0.64109    0.08062   7.952 6.47e-15 ***
health       1.00394    0.13140   7.641 6.39e-14 ***
freedom      1.47913    0.16338   9.053  < 2e-16 ***
corruption   0.85366    0.22328   3.823 0.000142 ***
generosity   0.59359    0.17564   3.380 0.000762 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5498 on 775 degrees of freedom
Multiple R-squared:  0.764,     Adjusted R-squared:  0.7622
F-statistic: 418.2 on 6 and 775 DF,  p-value: < 2.2e-16
```

Fig. 12: R output of the summary table of regression model

Regression Equation

From the R-output above, we constructed the following regression equation for the prediction of happiness score:

$$\text{score} = \beta o + \text{gdp}*\beta 1 + \text{family}*\beta 2 + \text{health}*\beta 3 + \text{freedom}*\beta 4 + \text{corruption}*\beta 5 + \text{generosity}*\beta 6$$

$$\text{score} = 2.17749 + 1.14675*\text{gdp} + 0.64109*\text{family} + 1.00394*\text{health} + 1.47913*\text{freedom} + 0.85366*\text{corruption} + 0.59359*\text{generosity}$$

These coefficients of the above equation are interpreted as the marginal increase in the happiness score when a variable changes by 1 unit and all the other variables remained fixed.

- **βo**: In case where all other predictors are absent from the model, the happiness score will be 2.17749
- **gdp**: If all other features are kept constant, then a unit increase in the gdp will result in an increase of 1.14675 units in the happiness score.
- **family**: If all other features are kept constant, then a unit increase in the level of family (social support) will result in an increase of 0.64109 units in the happiness score.
- **health**: If all other features are kept constant, then a unit increase in the health (life expectancy) will result in an increase of 1.00394 units in the happiness score.
- **freedom**: If all other features are kept constant, then a unit increase in the freedom will result in an increase of 1.47913 units in the happiness score.
- **corruption**: If all other features are kept constant, then a unit increase in the corruption (in our context, trust in government) will result in an increase of 0.85366 units in the happiness score.
- **generosity**: If all other features are kept constant, then a unit increase in the generosity will result in an increase of 0.59359 units in the happiness score.

Feature Significance (p-value)

We can gauge the significance of each of these predictor variables in explaining the response variable by analyzing the p-values given in the Pr(>|t|) column in the summary table of the regression model. A p-value is a measure of the probability that an observed difference could have occurred just by random chance and it can also be referred to as the measure of the strength of the evidence against the null hypothesis. For our model, null hypothesis (Ho) states that there is no relationship between happiness score and the respective predictor variable. If p-value for a predictor variable is less than 0.05 (the standard threshold), we will consider it to be significant in estimating the response variable.

It can be observed from the summary table that all values in the Pr(>|t|) column are less than 0.05. This indicates to the fact that all the features amongst Bo, gdp, family, health, freedom, corruption, and generosity play a significant role in estimating the happiness score. The overall p-value of our regression model is also lesser than 0.05, thus, it is statistically significant and can be used to predict happiness score from the given predictor variables.

Following table summarizes the p-values of each predictor variable along with whether it is significant or not:

Variable	P-Value	Significance (✔/✗)
Intercept	< 2e-16	✔
GDP	< 2e-16	✔
Family	6.47e-15	✔
Health	6.39e-14	✔
Freedom	< 2e-16	✔
Corruption	0.000142	✔
Generosity	0.000762	✔

Table 2: Significance of features (columns) as indicated by their respective p-value

Coefficient of Determination

R-squared is a statistical measure that represents the proportion of the variance for a dependent variable explained by independent variable(s) in a regression model. In this case, the Multiple R-squared value represents the variability in happiness score which is explained by the 6 features. However, in a multivariate regression setting, Multiple R-squared is not a reliable measure because it keeps increasing as we add more features to the model even though one or more variables among these may be insignificant (which then leads to an overfitted model). It can lead to inaccurate estimations regarding our model's prediction power.

Therefore, in such a regression model, Adjusted R-squared is employed and it ensures its value decreases when a non-significant variable is added to the model. In this case, the value of Adjusted R-squared is 0.7622 or 76.22% which depicts that roughly **76%** of the variability in happiness score is estimated by the model using the features gdp, family, health, freedom, corruption, and generosity.

Diagnostics

Regression models are based on a set of assumptions and it is imperative to check those assumptions before drawing conclusions regarding the relationship of the response variable to the set of predictors. One the preliminary model is fitted; different techniques can be employed for this purpose.

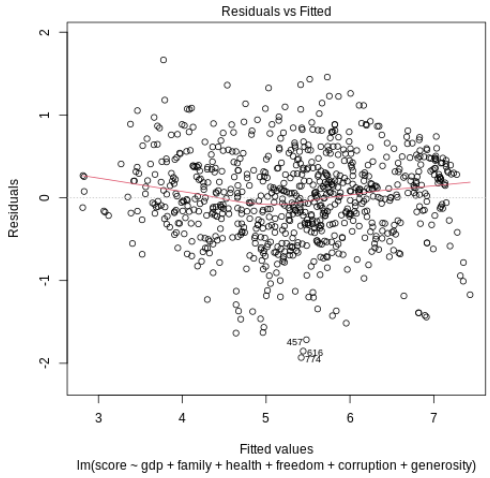

Fig. 13: Residual vs fitted plot of regression model

The above figure illustrates residual values (predicted values of the unknown errors) on the y-axis and fitted values on the x-axis. Ideally, the plot should look like a random scatter about the line residuals = 0 with constant variance because the sum of residuals is 0. It can be seen in the above scatterplot that the points are scattered randomly and there is no evident pattern that could, in turn, violate one or more assumptions of the regression model. Points numbered 457, 616, and 774 are represented as outliers in this figure because they are far away from the residuals = 0 line but the number is not significant as compared to the amount of data we have.

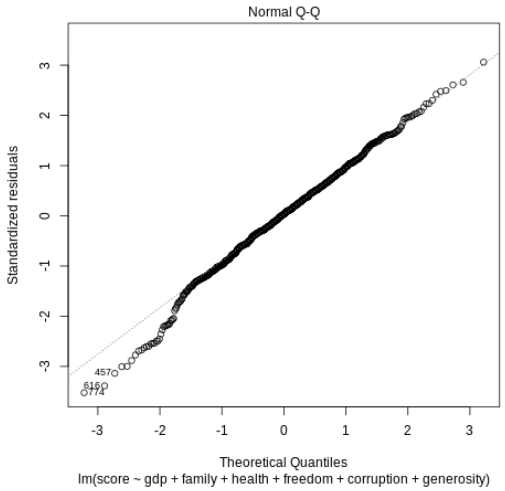

Fig. 14: Quantile plot of regression model

The above figure illustrates a normal quantile-quantile plot between the standardized residuals and the theoretical quantiles. It is used to evaluate the regression assumption that residuals are normally distributed. The points lie in a straight, diagonal line on this plot if the data comes from a normal distribution. Therefore, for the normality assumption to hold, the points should lie on or close to the y = x line. The curvature in the tails of this plotted line indicate towards the presence of outliers in the data. For instance, the points numbered 457, 616, and 774 are represented as outliers in this figure (just like the residual plot) because they are far away from the y = x line. Despite some deviance from the y = x line, a hefty majority of the data points lie on it and, hence, hold the normality assumption valid.

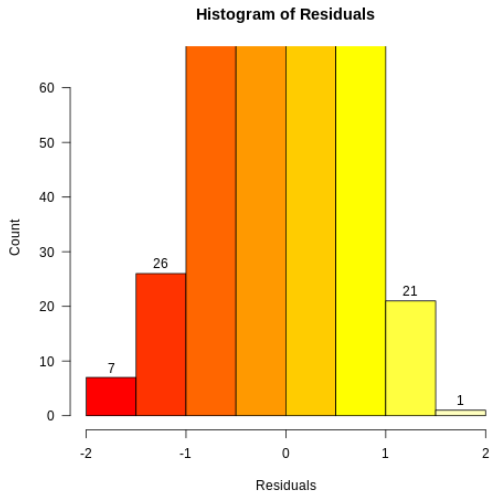

Fig. 15: Histogram of residuals of regression model

To reaffirm this notion of residuals being normally distributed, we also plot their histogram to showcase that the normality assumption in fact holds valid. The plot above also depicts that the mean is centered around 0 which is the case for data that is normally distributed.

```
%%R

cat("Skewness of model: ", round(skewness(multiple.regression$residuals), 3))

# We can see model is not skwewed
```

```
Skewness of model:  -0.286
```

Fig. 16: Skewness of regression model

We also tested the skewness of the model and, as shown above in the code snippet, the model is only slightly skewed to the left by –0.286 which is not much and can be accounted for as negligible. Therefore, we conclude that our model holds the set of assumptions that need to be considered in regression.

ANOVA Analysis

Next, we moved on to an integral part of regression i.e., ANOVA analysis. Following is the R-output of the ANOVA table we constructed:

```
anova(multiple.regression)
```

```
Analysis of Variance Table

Response: score
            Df Sum Sq Mean Sq  F value    Pr(>F)
gdp          1 618.47  618.47 2046.102 < 2.2e-16 ***
family       1  52.48   52.48  173.614 < 2.2e-16 ***
health       1  22.66   22.66   74.960 < 2.2e-16 ***
freedom      1  54.78   54.78  181.237 < 2.2e-16 ***
corruption   1   6.68    6.68   22.101 3.062e-06 ***
generosity   1   3.45    3.45   11.422 0.0007622 ***
Residuals  775 234.26    0.30
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 29: ANOVA table of regression model

The ANOVA (Analysis of Variance) test allows to draw comparisons between more than two groups simultaneously to determine whether there is any relationship between them. We will have a look at the main results of the output one by one:

- The degrees of freedom of our data are calculated to be N-(P+1) = 782 – (6 + 1) = 775. These are the number of data values from the dataset that are free to vary. Degrees of freedom will be used multiple time to calculate statistical measure like F-value as shown below.
- The result of ANOVA formula, F statistic (also called the ANOVA coefficient), can be calculated as

  F = Regression Mean Square / Residual Mean Square.

  The F statistic allows for the analysis of multiple groups of data to analyze and determine the variability between samples and within samples. If no true variance exists between the groups, that is, there is no evident difference between the groups, the F statistic is close to 1. However, in our model the high F-values indicate that each predictor variable plays a significant role predicting happiness score. This was indicated by the calculation of p-values as well but F-values verify our prior claim. Therefore, as shown in the ANOVA table above, all of the differences between the means are statistically significant which leads us to the conclusion that the regression model we have fitted, as a collection of the predictor variables, is highly significant.

Analysis of Predictor Variables

In this section, we will analyze each predictor variable one by one to understand the results of the model better. It includes its distribution and correlation and effect on happiness score along with plausible reasons.

GDP

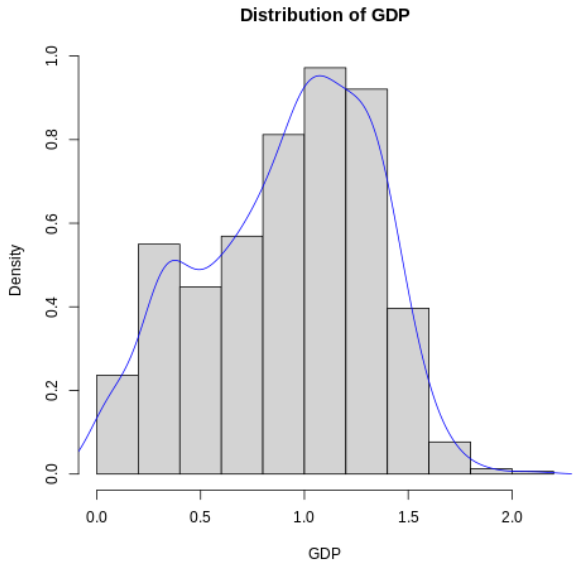

Fig. 17: Distribution of GDP - Density Plot

- The minimum GDP per capita is 0.000 while the maximum GDP per capita is 2.096 which gives a range of 2.096. Interquartile range for GDP per capita is 0.6297.
- GDP per capita has a strong positive correlation with happiness score as illustrated in the correlation plot, too.
- As our regression model identified, a unit increase in GDP per capita also leads to the highest increase in units of happiness score.
- Intuitively, countries which have a better economy have less inflation and more ease for the citizens. An example here would be a comparison between any developed and developing country. Countries with a better economy have a higher living standard which could possibly lead to citizens' happiness.

```
Five number summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.6065  0.9822  0.9160  1.2362  2.0960
```

Fig. 18: Five number summary of GDP

Family

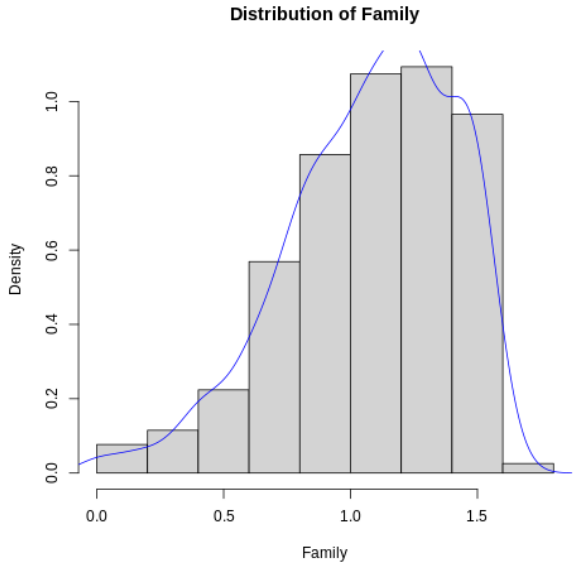

Fig. 19: Distribution of Family - Density Plot

- The minimum level of social support that an individual has in the form of family, friends and relatives is 0.000 while the maximum is 1.6440 which gives a range of 1.6440. Interquartile range for family / social support is 0.4579.
- The level of social support an individual receives has a strong positive correlation with happiness score as illustrated in the correlation plot, too.
- As our regression model identified, a unit increase in family (social support) also leads to an increase in units of happiness score.
- Logically, it makes sense that people require a certain level of social support from relatives and friends to be happy and this relationship between the two things is embedded into the societal contexts. A lot of people rely on their support system for their happiness. Therefore, family (social support) is a plausible measure to predict happiness.

```
Five number summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.8694  1.1247  1.0784  1.3273  1.6440
```

Fig. 20: Five number summary of family

Health

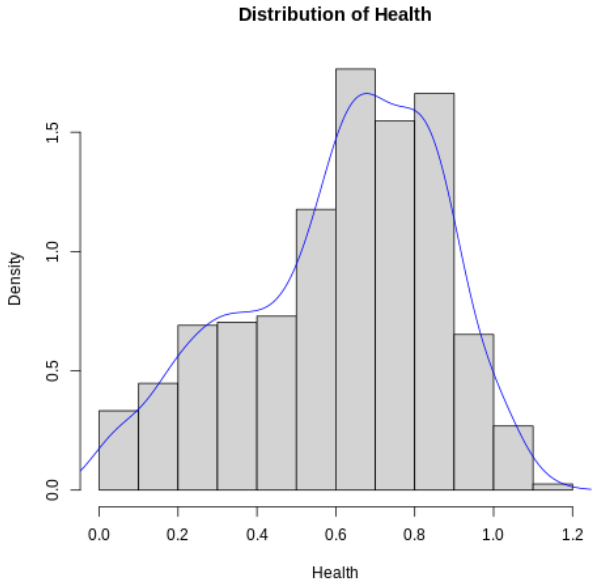

Fig. 21: Distribution of Health - Density Plot

- The minimum level of healthy life expectancy is 0.000 while the maximum is 1.1410 which gives a range of 1.1410. Interquartile range for healthy life expectancy is 0.3678.
- The healthy life expectancy has a strong positive correlation with happiness score as illustrated in the correlation plot, too.
- As our regression model identified, a unit increase in healthy life expectancy also leads to an increase in units of happiness score.
- Intuitively, it makes sense the other way around, too, given that being happier improves the quality of life and consequently leads to a healthier life expectancy. In these unprecedented times of the pandemic, the need for healthy life, high immunity and greater life expectancy is highlighted where our happiness is based on our health and the health of the people around us.

```
Five number summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.4402  0.6473  0.6124  0.8080  1.1410
```

Fig. 22: Five number summary of healthy life expectancy

Freedom

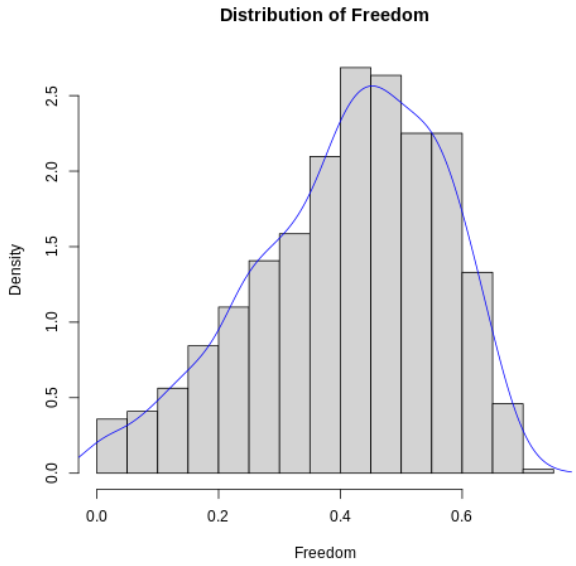

Fig. 23: Distribution of Freedom - Density Plot

- The minimum level of freedom to make life choices is 0.000 while the maximum is 0.7240 which gives a range of 0.7240. Interquartile range for freedom is 0.2212.
- Liberty to make life choices has a positive correlation with happiness score as illustrated in the correlation plot, too.
- As our regression model identified, a unit increase in freedom also leads to an increase in units of happiness score.
- Again, this makes intuitive sense since self-centered autonomy and choices are essential factors impact happiness, for many people.

```
Five number summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.3098  0.4310  0.4111  0.5310  0.7240
```

Fig. 24: Five number summary of freedom

Corruption

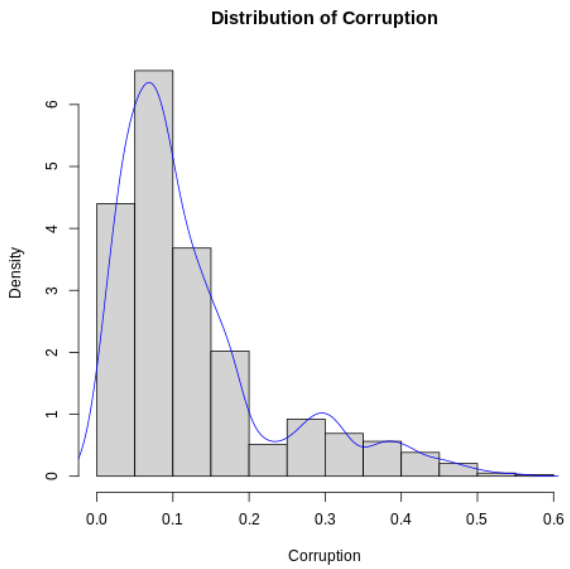

Fig. 25: Distribution of Corruption - Density Plot

- The minimum level of perceptions of corruption is 0.000 while the maximum is 0.55191 which gives a range of 0.55191. Interquartile range for perceptions of corruption is 0.10161.
- The low level of corruption and increased trust in government of a particular country has a fairly weak positive correlation with happiness score as illustrated in the correlation plot as well.
- As our regression model identified, a unit increase in perceptions of corruption (in our case the perception of trust in government) leads to the increase in units of happiness score but this increase is not as profound as that contributed by predictor variables like GPA per capita or healthy life expectancy.
- Perceptions of corruption, as shown in the histogram, are rightly skewed which means that very countries have high perceptions of corruption, that is, more countries have corruption problems.

```
Five number summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.05425 0.09103 0.12544 0.15586 0.55191
```

Fig. 26: Five number summary of corruption

Generosity

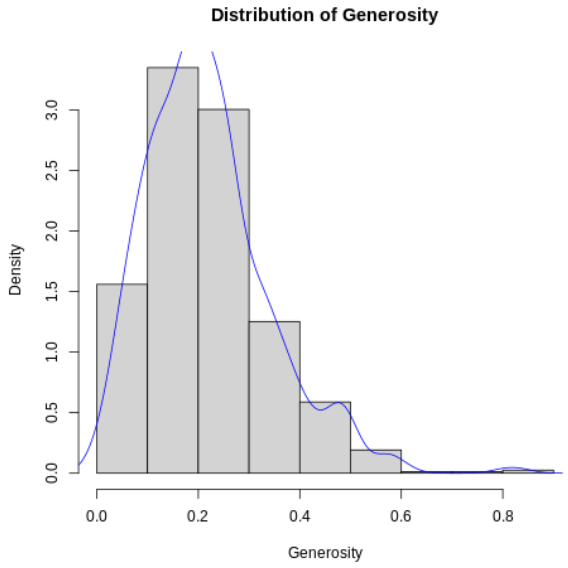

Fig. 27: Distribution of Generosity - Density Plot

- The minimum level of generosity is 0.000 while the maximum is 0.8381 which gives a range of 0.55191. Interquartile range for generosity is 0.1488.
- Generosity has a weak positive correlation with happiness score as illustrated in the correlation plot, too.
- As our regression model identified, a unit increase in generosity leads to the increase in units of happiness score but this increase is not as profound as that contributed by predictor variables like GPA per capita or healthy life expectancy.

```
Five number summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1300  0.2020  0.2186  0.2788  0.8381
```

Fig. 28: Five number summary of generosity

Multicollinearity Analysis

Multicollinearity, a phenomenon in which high intercorrelations between two or more independent variables in a multiple regression model exist, can lead to skewed and misleading results and makes the regression model estimates unreliable and unstable. While we had checked correlations between the variables before starting the analysis, we cross-checked once again in order (this time with a perspective of multicollinearity) to check if the variables used in the model are have little to no multi-collinearity or not.

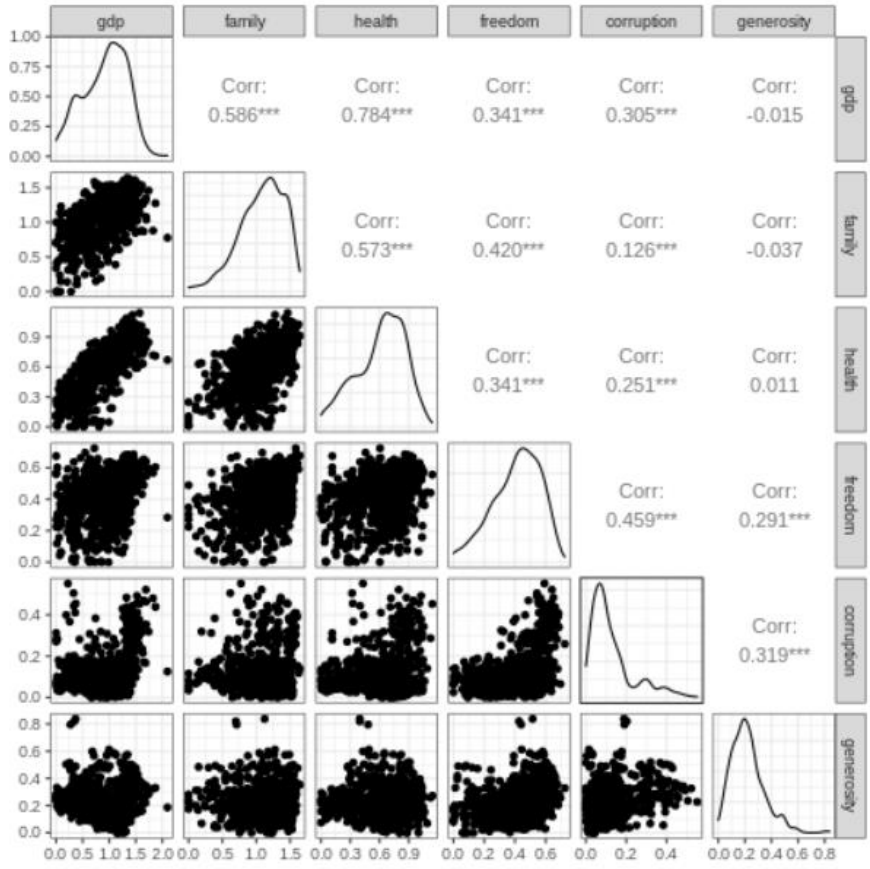

Fig. 30: Visualization of multicollinearity (if any)

The above plot illustrates the scatterplot of each independent variable with every other predictor to visually look for multicollinearity. The diagonal represents the density plots for gdp, family, health, freedom, corruption, and generosity. As can be noted from the figure, only health and gdp are high correlated with each other whereas family-gdp and health-family are moderately correlated with each other. In general, it can be concluded that there is very slight multicollinearity in the overall collection of the independent variables which does not affect the reliability of our statistical inferences.

We further affirm this notion by using a Variance inflation factor (VIF) model below.

```
vif_score <- car::vif(multiple.regression)
vif_score
```

```
      gdp     family     health    freedom corruption generosity
 2.936310   1.823643   2.750506   1.611972   1.440463   1.192625
```

Fig. 31: VIF scores of predictor variables

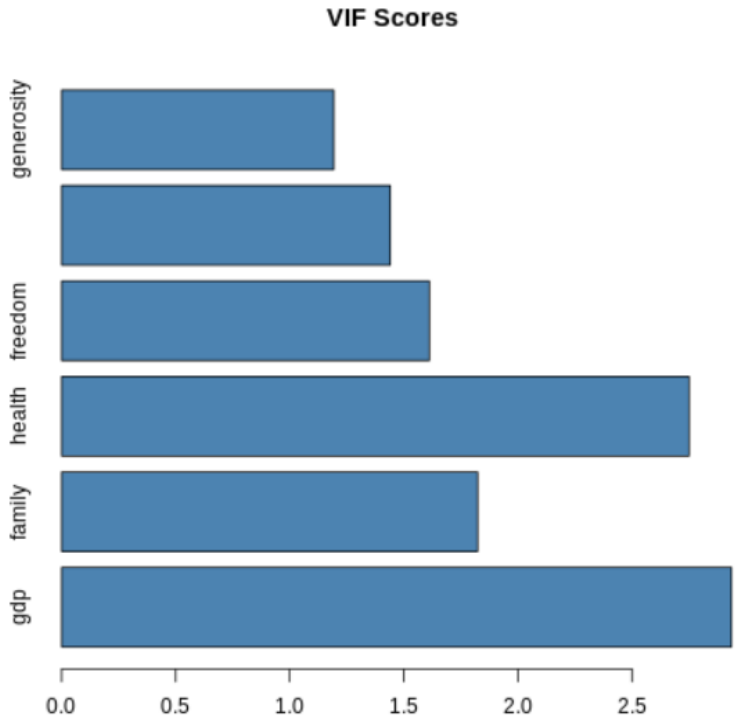

Fig. 32: Bar chart of VIF scores

The measure used here, VIF, quantifies the amount of multicollinearity within a group of independent variables in a multivariate regression setting (i.e., the regression technique that we have used). In numeric terms, it is the proportion of the variance of the overall model and the variance of the respective independent variable. A large value of VIF, thus, indicates that an independent variable has a highly collinearity to the other multiple regression variables which should be adjusted for.

In our regression model's case, none of the independent variables have VIF values which exceed 3 (i.e., VIF <= 3), as depicted in the horizontal bar chart above. In general, a VIF above 10 indicates high correlation. Since the values of our predictor variables are less than the threshold, there is little to no multicollinearity which could have made the model unreliable. Overall, no further adjustments for multicollinearity are required.

Results and Conclusion

In addition to the multiple sub questions which were answered at each stage of the analysis, following are the key results that can be drawn from the extensive analysis, visualizations, and elaboration of statistical numbers above:

- Factors like GDP, family, health, freedom, corruption, and generosity play a significant role in determining the happiness score. Increase in any of these features, keeping the others constant, leads to an increase in the happiness (as explained by the regression equation and its interpretation). Out of all the factors, freedom has the greatest impact on the happiness score followed by GDP per capita and health (life expectancy) of the citizens.

- All the contributing factors had p-value < 0.05 (I.e., our threshold to reject the null hypothesis). Therefore, all of them were statistically significant to predict happiness score. However, p-value alone is not sufficient to confirm our findings. Therefore, ANOVA table and calculation of F-statistic leads us to the conclusion that the regression model we have fitted, as a collection of the predictor variables, is indeed highly significant.

- Upon exploratory data analysis, we found little to no outliers and very less skewness in the overall data which made it easier to model the data and interpret relevant statistics. The multiple linear regression sufficed to model and describe our data effectively.

- Overall, the data selection, data consolidation, data cleaning, feature selection, exploratory data analysis, regression modelling and subsequent analysis enables us to find the answer to the research question posed in the beginning i.e., various social, urban and national factors do contribute in determining the happiness of citizens.

Way Forward

With such interesting results and conclusions from our analysis, we are hopeful that subsequent happiness reports and measurements of well-being can be used effectively to assess the progress and happiness of nations. Our analysis covered the happiness reports from the year 2015 to 2019. However, after COVID-19, the focus of the happiness report requires a slight shift from typical predictions to understanding what effect the pandemic has had on subjective well-being and vice versa.

More than two million people have died worldwide and the threat of variants and uneven policy decisions on how to respond has created uncertainty in what the future holds. But despite this, the need of the hour is to make things easier for each other by understanding the well-being of individuals and taking necessary steps. This is where happiness report and its analysis come into play. Even if the contributing factors in the report of upcoming years remain the same, it is important to evaluate that what factors affect the happiness of individuals and states more than the other in context of the pandemic. This will not only enable better citizen-centric policy formulations by the government but will also allow us, as individuals to contribute to a happier environment for each other.

References

[1] Network, Sustainable Development Solutions. "World Happiness Report." *Kaggle*, 27 Nov. 2019, www.kaggle.com/unsdsn/world-happiness. (Dataset)

[2] "World Happiness Report." *Home*, Sustainable Development Solutions Network, worldhappiness.report.

[3] "R Packages on CRAN and Bioconductor." *RDocumentation*, DataCamp, www.rdocumentation.org