World Happiness Report 2021

HarvardX PH125.9x

Data Science: Capstone

Introduction:

Context

The World Happiness Report is a landmark survey of the state of global happiness . The first World Happiness Report was released on April 1, 2012 as a foundational text for the UN High Level Meeting: Well-being and Happiness: Defining a New Economic Paradigm, drawing international attention. The first report outlined the state of world happiness, causes of happiness and misery, and policy implications highlighted by case studies. In 2013, the second World Happiness Report was issued, and in 2015 the third. Since 2016, it has been issued on an annual basis on the 20th of March, to coincide with the UN's International Day of Happiness. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

Content

The 2021 World Happiness Report, released on March 20, 2021, ranks 156 countries based on an average of three years of surveys between 2017 and 2019. The 2020 report especially focuses on the environment – social, urban, and natural, and includes links between happiness and sustainable development.

The happiness scores and rankings use data from the Gallup World Poll . The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

Project Objective

- The top 10 and bottom 10 countries in 2021 happiness report.
- Factors that contributes most/least to happiness.
- Can happiness changes according to the quality of the society in which people live.

Data Source

Data was downloaded from a CSV file from Kaggle

Importing the library & dataset

```
library(tidyverse)
library(data.table)
library(corrplot)
library(GGally)
library(rworldmap)
library(caret)
library(data.table)
library(ggplot2)
library(car)
library(pscl)

dh_2021 <- read.csv("world-happiness-report-2021.csv")
```

Overview

```
glimpse(dh_2021)
```

```
## Rows: 149
## Columns: 20
## $ ï..Country.name                            <chr> "Finland", "Denmark", "Swit~
## $ Regional.indicator                         <chr> "Western Europe", "Western ~
## $ Ladder.score                               <dbl> 7.842, 7.620, 7.571, 7.554,~
## $ Standard.error.of.ladder.score             <dbl> 0.032, 0.035, 0.036, 0.059,~
## $ upperwhisker                               <dbl> 7.904, 7.687, 7.643, 7.670,~
## $ lowerwhisker                               <dbl> 7.780, 7.552, 7.500, 7.438,~
## $ Logged.GDP.per.capita                      <dbl> 10.775, 10.933, 11.117, 10.~
## $ Social.support                             <dbl> 0.954, 0.954, 0.942, 0.983,~
## $ Healthy.life.expectancy                    <dbl> 72.000, 72.700, 74.400, 73.~
## $ Freedom.to.make.life.choices               <dbl> 0.949, 0.946, 0.919, 0.955,~
## $ Generosity                                 <dbl> -0.098, 0.030, 0.025, 0.160~
## $ Perceptions.of.corruption                  <dbl> 0.186, 0.179, 0.292, 0.673,~
## $ Ladder.score.in.Dystopia                   <dbl> 2.43, 2.43, 2.43, 2.43, 2.4~
## $ Explained.by..Log.GDP.per.capita           <dbl> 1.446, 1.502, 1.566, 1.482,~
## $ Explained.by..Social.support               <dbl> 1.106, 1.108, 1.079, 1.172,~
## $ Explained.by..Healthy.life.expectancy      <dbl> 0.741, 0.763, 0.816, 0.772,~
## $ Explained.by..Freedom.to.make.life.choices <dbl> 0.691, 0.686, 0.653, 0.698,~
## $ Explained.by..Generosity                   <dbl> 0.124, 0.208, 0.204, 0.293,~
## $ Explained.by..Perceptions.of.corruption    <dbl> 0.481, 0.485, 0.413, 0.170,~
## $ Dystopia...residual                        <dbl> 3.253, 2.868, 2.839, 2.967,~
```

The data consists of 149 rows and 20 columns.

Cleaning the dataset

We first check Null values in each data frame and found there is no null values in the data frame 2021. Then we renamed the columns and removed some of them for ease of access.

```
dh_2021_new <- dh_2021[c(1:3,7:12)]
```

```
dh_2021_new <- dh_2021_new %>%
  rename("Country" = "ï..Country.name",
         "region"="Regional.indicator",
         "GDP"= "Logged.GDP.per.capita",
         "score" ="Ladder.score",
         "support" = "Social.support",
         "Life.exp" = "Healthy.life.expectancy",
         "Freedom" ="Freedom.to.make.life.choices",
         "corruption"="Perceptions.of.corruption")
```

Checking for missing values & the class of every columns

```
colSums(is.na(dh_2021_new))
```

```
##    Country     region      score        GDP    support   Life.exp    Freedom
##          0          0          0          0          0          0          0
## Generosity corruption
##          0          0
```

```
sapply(dh_2021_new, class)
```

```
##     Country      region       score         GDP     support    Life.exp
## "character" "character"   "numeric"   "numeric"   "numeric"   "numeric"
##     Freedom  Generosity  corruption
##   "numeric"   "numeric"   "numeric"
```

All the columns except country name & regional indicator have numeric datatypes. Therefore, the current data types of the columns are fine for our analysis.

Exploratory Data Analysis

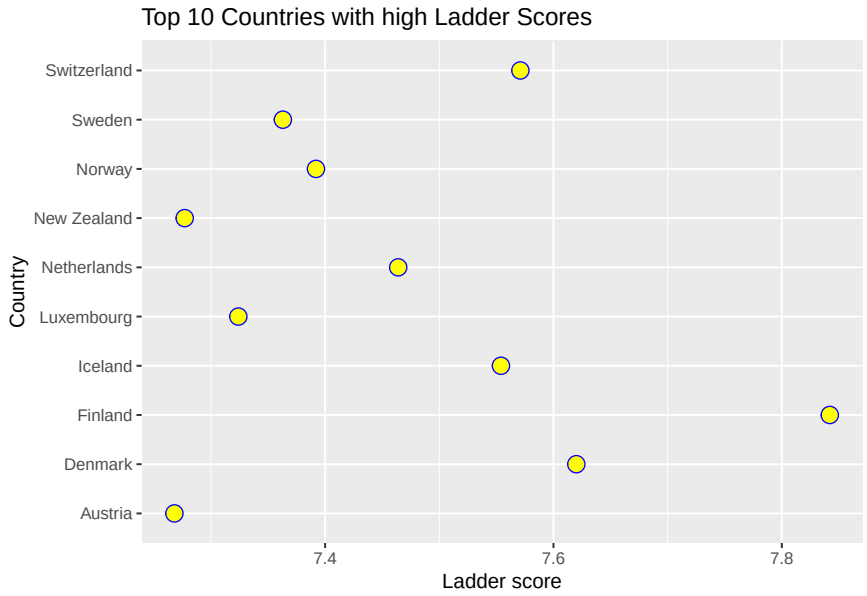
Top 10 and bottom 10 countries

Top 10

```
top10 <- dh_2021_new %>%
  head(10) %>%
  mutate(Level = "Top10")
```

```
ggplot(top10, aes(x= Country,  y=score)) +
  geom_point( shape = 21,
              fill = "yellow",
              color = "blue",
```

```
              size = 4) +
  ylab("Ladder score") +
  ggtitle("Top 10 Countries with high Ladder Scores" ) +
  coord_flip()
```



We can see that the top 10 happy countries are Finland, Denmark, Switzerland, Iceland, Netherlands, Norway, Sweden, Luxembourg, New Zealand & Austria.

Bottom 10

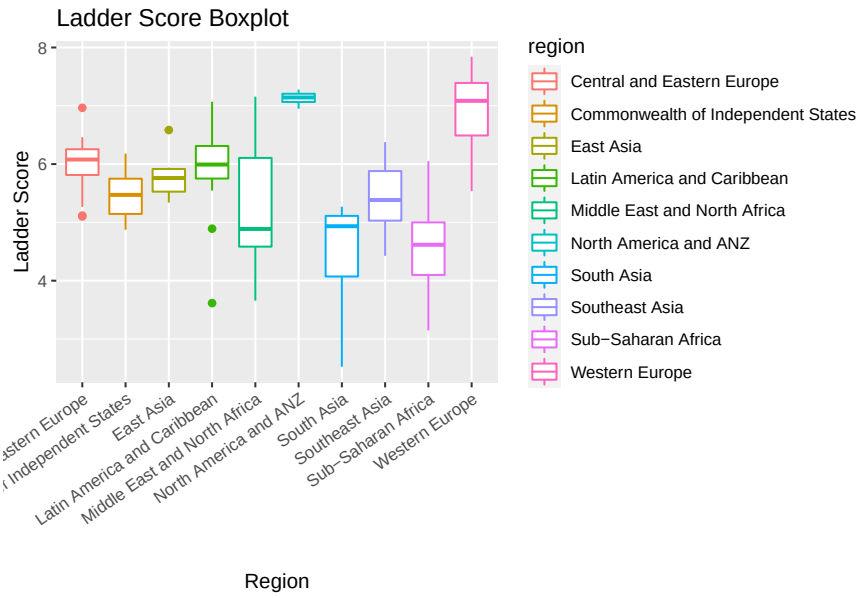
```
bottom10 <- dh_2021_new %>%
  tail(10) %>%
  mutate(Level = "Bottom10")
```

```
ggplot(bottom10, aes(x= Country,  y=score)) +
  geom_point( shape = 21,
              fill = "yellow",
              color = "blue",
              size = 4) +
  ylab("Ladder score") +
  ggtitle("Bottom 10 Countries with high Ladder Scores" ) +
  coord_flip()
```

Bottom 10 countries are Afghanistan, Zimbabwe, Rwanda, Botswana, Lesotho, Malawi, Haiti, Tanzania, Yemen & Burundi.

Plotting happiness by region using box plot.

```
ggplot(dh_2021_new, aes(x=region,
                        y= score,
                        colour = region)) +
  geom_boxplot() +
  labs(title = "Ladder Score Boxplot",
       x = "Region",
       y = "Ladder Score") +
  theme(axis.text.x = element_text(angle = 35,
                                   vjust = 1,
                                   hjust = 1))
```



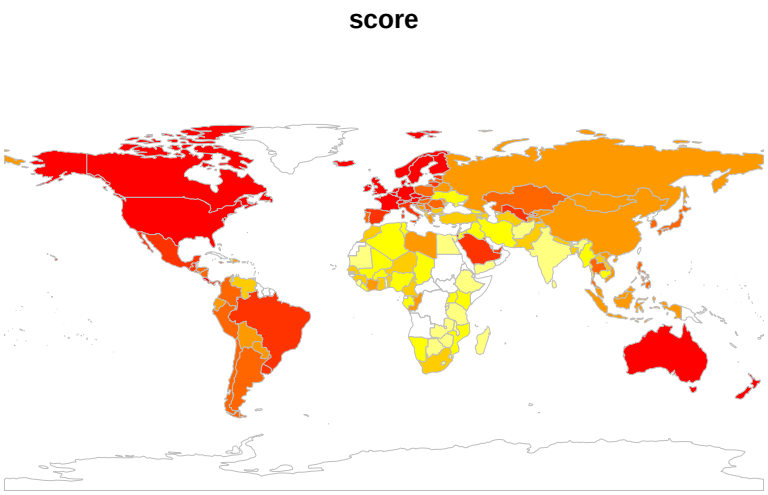
We can clearly see the trend that Western European countries are happier while the South Asian and Sub-Saharan Africa countries seems to be saddest.

Visualizing happiness on the world map

```
world2021 <- joinCountryData2Map(dh_2021_new,
                                 joinCode = "NAME",
                                 nameJoinColumn = "Country")
```

```
## 144 codes from your data successfully matched countries in the map
## 5 codes from your data failed to match with a country code in the map
## 99 codes from the map weren't represented in your data
```

```
par(mar = c(1, 1, 1, 1))
map_2021 <- mapCountryData( world2021,
                            nameColumnToPlot="score",
                            addLegend = FALSE)
```

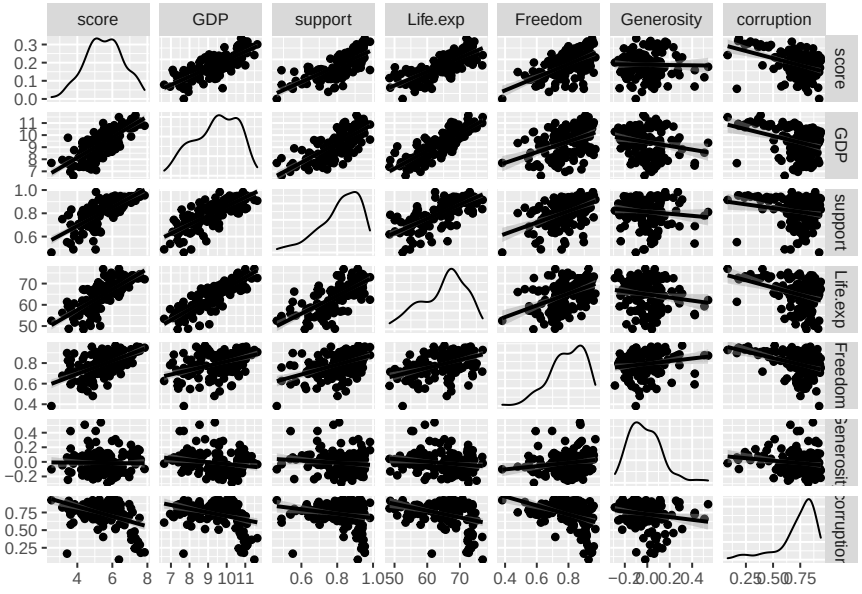


Vizualizing how the factors are related to the Ladder score:

Scatter Plot:

Scatter plot matrix of correlations between different factors. Scatterplots are great at showing relationships between two variables, even when there are many potentially overlapping data points.

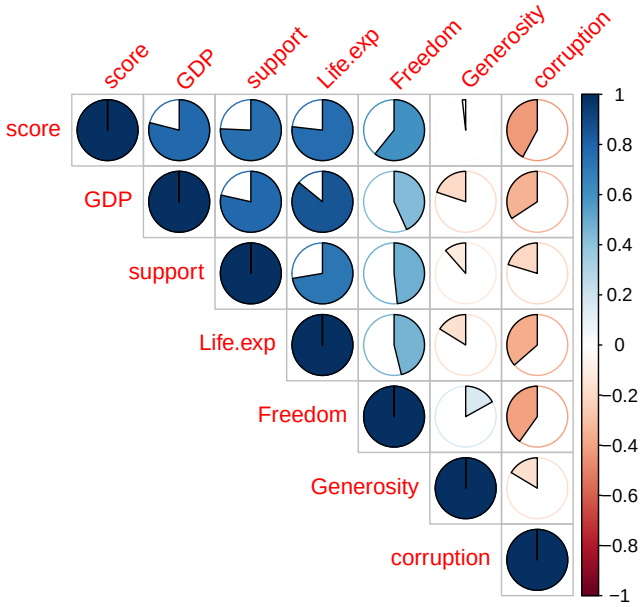
```
ggpairs(dh_2021_new,
        columns=c(3:9),
        upper=list(continuous="smooth"),
        lower=list(continuous="smooth"))
```



Correlation matrix:

GDP per capita, social support, healthy life expectancy, freedom to make choices, generosity, and corruption, all these factors may have significant impacts on happiness score and to know how these factors impact any countries we introduce a correlation matrix to formulate this relationship.Thus, governments may be advised to focus on these significant factors to improve the life there.

```
corr_dh <- dh_2021_new[c(3:9)]
M <- cor(corr_dh)
corrplot(M,
         method = "pie",
         tl.srt=45,
         type="upper")
```



Clearly from both the visualizations we understand that ladder score has the strongest positive correlations with GDP, social support, healthy life expectancy & freedom to make life choices but perceptions of corruption & generosity had weakest to no correlations with the ladder score & other factors.

Analysis

We take the attributes and use a regression to predict the output of the happiness, we chose the "Ladder score" as the output variable.

```
# Model : Creating Test and Train set

set.seed(1, sample.kind="Rounding")
```

```
## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
test_index <- createDataPartition(y=dh_2021_new$score,
                                  times = 1,
                                  p = 0.2,
                                  list = FALSE)

train_set <- dh_2021_new[-test_index,]
test_set <- dh_2021_new[test_index,]
```

Linear Regression

Here we use Linear Regression to model the dependence of Ladder score on a set of predictors - GDP, social support, healthy life expectancy, freedom to make life choices, perceptions of corruption, generosity.

```
model1 <- lm(score ~ GDP +
               support +
               Life.exp +
               Freedom +
               Generosity +
               corruption,
             data = train_set)

summary(model1)
```

```
##
## Call:
## lm(formula = score ~ GDP + support + Life.exp + Freedom + Generosity +
##     corruption, data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7667 -0.2666  0.0792  0.2932  1.0006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.89540    0.68697  -2.759 0.006791 **
## GDP          0.25111    0.09118   2.754 0.006889 **
## support      1.97837    0.72266   2.738 0.007219 **
## Life.exp     0.03892    0.01376   2.828 0.005565 **
## Freedom      1.93475    0.54070   3.578 0.000516 ***
## Generosity   0.73705    0.40792   1.807 0.073519 .
```

```
## corruption  -0.74663    0.33170  -2.251 0.026377 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5177 on 110 degrees of freedom
## Multiple R-squared:  0.772,  Adjusted R-squared:  0.7595
## F-statistic: 62.06 on 6 and 110 DF,  p-value: < 2.2e-16
```

A P value greater than 0.05 means that no effect was observed. So we ignore the factor, Generosity and recreate our model.

```
model1_re <-lm(score ~ GDP +
     support +
     Life.exp +
     Freedom +
     corruption,
   data = train_set)

summary(model1_re)
```

```
##
## Call:
## lm(formula = score ~ GDP + support + Life.exp + Freedom + corruption,
##     data = train_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92112 -0.29554  0.04199  0.31810  0.99970
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.73192    0.68790  -2.518  0.01324 *
## GDP          0.21679    0.09008   2.407  0.01776 *
## support      2.10525    0.72654   2.898  0.00453 **
## Life.exp     0.03757    0.01388   2.706  0.00788 **
## Freedom      2.19640    0.52623   4.174 5.98e-05 ***
## corruption  -0.85828    0.32920  -2.607  0.01038 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5229 on 111 degrees of freedom
## Multiple R-squared:  0.7652, Adjusted R-squared:  0.7546
## F-statistic: 72.34 on 5 and 111 DF,  p-value: < 2.2e-16
```

Looking at the adjusted R-Squared value we see that its a good value for showing accuracy and can explain the variation data from the Ladder Score

```
# Prediction:

predict_model1 <- predict(model1_re,
                          newdata = test_set)

# RSME
```

```
RMSE(predict_model1, test_set$score)
```

```
## [1] 0.6288289
```

We can see that GDP, Social support, Healthy life expectancy and freedom to make choices are the strongest predictors.

Model 2: Logistics Regression

```
model2 <- glm(score ~ GDP +
                support +
                Life.exp +
                Freedom +
                corruption,
              data = train_set)
```

```
# disable scientific notation for model summary
options(scipen=999)

summary(model2)
```

```
##
## Call:
## glm(formula = score ~ GDP + support + Life.exp + Freedom + corruption,
##     data = train_set)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.92112  -0.29554   0.04199   0.31810   0.99970
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -1.73192    0.68790  -2.518   0.01324 *
## GDP          0.21679    0.09008   2.407   0.01776 *
## support      2.10525    0.72654   2.898   0.00453 **
## Life.exp     0.03757    0.01388   2.706   0.00788 **
## Freedom      2.19640    0.52623   4.174 0.0000598 ***
## corruption  -0.85828    0.32920  -2.607   0.01038 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2734655)
##
##     Null deviance: 129.271  on 116  degrees of freedom
## Residual deviance:  30.355  on 111  degrees of freedom
## AIC: 188.17
##
## Number of Fisher Scoring iterations: 2
```

In typical linear regression, we use R^2 as a way to assess how well a model fits the data. Values close to 0 indicate that the model has no predictive power. In practice, values over 0.40 indicate that a model fits the data very well.

Assessing the model fit

```
pR2(model2)["McFadden"]
```

```
## fitting null model for pseudo-r2
```

```
##  McFadden
## 0.4932444
```

The value 0.4932444 indicates that our model fits the data very well and has high predictive power.

Checking for multicollinearity

```
vif(model2)
```

```
##        GDP    support   Life.exp    Freedom corruption
##   4.649561   2.998361   3.888673   1.497897   1.336003
```

Values of VIF exceeding 10 are often regarded as indicating multicollinearity. Since none of the predictors in our models have a VIF over 10, multicollinearity is not an issue in our model.

```
# Prediction:

predict_model2 <- predict(model2, test_set, type="response")

# RMSE:

RMSE(predict_model2, test_set$score)
```

```
## [1] 0.6288289
```

Comparing to the results, we see that logistic regression performs similarly to the linear regression model. GDP, Social support, Healthy life expectancy and freedom to make choices are the strongest predictors.

Conclution

We clearly saw the trend that Western European countries are happier while the South Asian and Sub-Saharan Africa countries seems to be saddest. Finland being the happiest country and Afghanistan in the bottom of the list.

From the linear regression & logistics regression model we see that GDP per capita, Social support, freedom to make choices and Life expectancy are great predictors of Happiness score and can be used to predict the future scores.

By looking at the happiness report and analyzing them, we are able to understand what makes countries and their citizens happier, thus allowing us to focus on prioritizing and improving these aspects of each nation.

However, this is not conclusive because unforeseen problems like pandemic, natural disasters and economic problems can happen, even to the most stable countries so these scores can actually change.