**Is Open Access to Information a Predictor of Happiness?**

Short answer: Maybe….

**Hypothesis**

In pursuit of my Masters of Library and Information Science the topic of accessible, transparent, and accurate information is very often a topic of concern and conversation. Most information professionals would agree that data openness and accessibility is not just important, but is a human right. But, after that… so what?

The question I present to explore is “Does having access to information make people happy?”.

**Data Origin Description**

To explore if happiness is related to having access to open data, I used R-Studio to run a correlation. I used the variables of happiness (x=Happiness) and access to open data (y=Openness) from two different datasets and linked them by a common value of “Country”.

The **Happiness** measurements came from “The World Happiness Report” for 2016. The happiness score is compiled by a number of factors related to government, prosocial behavior, and information technology. The World Happiness Report 2016 Update, which ranks 156 countries by their happiness levels is among the annual surveys that have been produced by the United Nations Sustainable Development Solutions Network and found on <https://worldhappiness.report/>. The raw data that I imported can be found on Kaggle <https://www.kaggle.com/unsdsn/world-happiness>

The Global **Open Data** Index is an annual compilation by Open Knowledge International, that measures the state of open government data around the world. The crowdsourced survey is designed to assess the openness of specific government datasets using the variables of context, data, use and impact. The 2016 dataset that I used was found at the World Wide Web Foundation, Open Data Barometer Global Report (Fourth Edition) <http://www.opendatabarometer.org>

**Dataset Limitations/Missing Data**

In order to join these two datasets, I used common countries. Some areas were not considered as countries or measured in both datasets. Scrubbing those out left a universe of 108 countries. Additionally, in order focus on the relationship, I only included the scaled happiness and openness scores. This leaves out the other variables that were used to develop each score.

**Correlation**

* Correlation coefficient (Cor.coeff = 0.6780585)
* t-test statistic value (t = 9.5426)
* Degrees of freedom (df= 107)
* p-value = 5.542e-16
* 95 percent confidence interval: 0.5615834 to 0.7681831

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Happiness  Score | Openness Score | fitted | se.fit | resid | hat | sigma | cooksd |
| 7.09 | 19.9 | 5.14 | 0.0920 | 1.94 | 0.0124 | 0.808 | 0.0353 |
| 7.50 | 38.9 | 5.75 | 0.0809 | 1.75 | 0.00959 | 0.812 | 0.0219 |
| 6.70 | 17.9 | 5.08 | 0.0956 | 1.62 | 0.0134 | 0.815 | 0.0265 |
| 7.27 | 46.3 | 5.99 | 0.0895 | 1.28 | 0.0117 | 0.821 | 0.0144 |
| 6.38 | 18.7 | 5.10 | 0.0942 | 1.28 | 0.0130 | 0.821 | 0.0159 |

Residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -1.70378 | -0.51451 | 0.08753 | 0.54500 | 1.94478 |

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.504496 0.138638 32.491 < 2e-16 \*\*\*

Openness 0.032095 0.003363 9.543 5.54e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

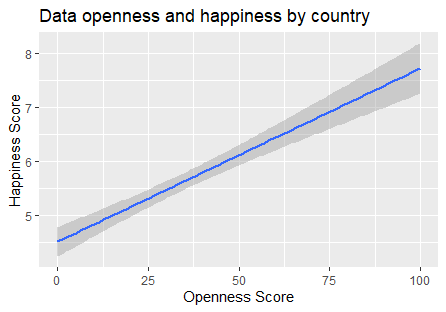
Residual standard error: 0.8262 on 107 degrees of freedom

Multiple R-squared: 0.4598, Adjusted R-squared: 0.4547

F-statistic: 91.06 on 1 and 107 DF, p-value: 5.542e-16

**Explanation**

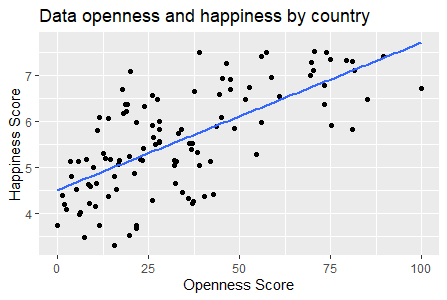
A correlation score of 0.6780585 shows a moderate to strong relationship between access to open data and happiness!



So, can we say that if governments provide more transparency and access to data, then all of the people will be happy?

Not quite… When we dig in further, we realize that we aren’t too confident that these results did not occur by chance. This is found by looking at the p-value is the significance level of the t-test (p-value = 5.542e-16). Based on the p-value of 5.54, it’s somewhat likely that the relationship could have occurred by chance.

The confidence interval of the correlation coefficient at 95% is between 0.5615834 & 0.7681831



Additionally, there are so many other factors that could influence happiness and perhaps relate to the type of areas where governments allow open access. However, given the strength of the coefficient, it would be worth looking further into the relationship.

**Appendix - Code**

Happy16 <- read.csv("School/Data Analysis/PredictDataAssign/Happy2016.csv", header = TRUE)

class(Happy16)

dim(Happy16)

names(Happy16)

str(Happy16)

glimpse(Happy16)

summary(Happy16)

head(Happy16)

tail(Happy16)

Opendata <- read.csv("School/Data Analysis/PredictDataAssign/ODB16.csv", header = TRUE)

class(Opendata)

dim(Opendata)

names(Opendata)

str(Opendata)

glimpse(Opendata)

summary(Opendata)

head(Opendata)

tail(Opendata)

#combining

open\_data\_happiness <- Opendata %>%

left\_join(Happy16, by = "Country") %>%

select(Country,Happiness.Score,ODB.Score.Scaled)

head(open\_data\_happiness)

glimpse(open\_data\_happiness)

class(open\_data\_happiness)

dim(open\_data\_happiness)

names(open\_data\_happiness)

str(open\_data\_happiness)

summary(open\_data\_happiness)

ggplot(open\_data\_happiness,aes(x=open\_data\_happiness$ODB.Score.Scaled, y=open\_data\_happiness$Happiness.Score )) + geom\_point()+

stat\_smooth(method = "lm",se = FALSE)+

labs(x="Openness Score", y="Happiness Score",

title = "Data openness and happiness by country")

lm\_HappyData <- lm(Happiness.Score ~ ODB.Score.Scaled, data = open\_data\_happiness)

lm\_HappyData

summary(lm\_HappyData)

coef(lm\_HappyData)

#what the model predicted

fitted\_HappyData <- fitted.values(lm\_HappyData)

fitted\_HappyData

# difference between actual and predicted

residual\_HappyData <- residuals(lm\_HappyData)

residual\_HappyData

# joins the origional data set with the linear model and the residual

lm\_matrix\_HappyData <- broom::augment(lm\_HappyData)

head(lm\_matrix\_HappyData)

lm\_matrix\_HappyData %>%

arrange(desc(.resid)) %>%

head()

#absolute value

lm\_matrix\_HappyData$.resid\_abs <- abs(lm\_matrix\_HappyData$.resid)

lm\_matrix\_HappyData %>%

arrange(desc(.resid\_abs)) %>%

head()

ggplot(open\_data\_happiness,aes(x=open\_data\_happiness$ODB.Score.Scaled,

y=open\_data\_happiness$Happiness.Score )) +

geom\_smooth(method = "lm") +

labs(x="Openness Score", y="Happiness Score",

title = "Data openness and happiness by country")

cov(open\_data\_happiness$ODB.Score.Scaled, open\_data\_happiness$Happiness.Score)

cor(open\_data\_happiness$ODB.Score.Scaled, open\_data\_happiness$Happiness.Score)

res <- cor.test (open\_data\_happiness$ODB.Score.Scaled, open\_data\_happiness$Happiness.Score, method = "pearson")

res