## **Michal Sobierajski, Vivek Sriram, and Matthew Kim**

## **Statistics 571/701 Final Project**

**May 9th, 2021**

**The Greatest Wealth is Health: Investigating Connections between Health Statistics and Happiness Index across the Globe**

## **Overview**

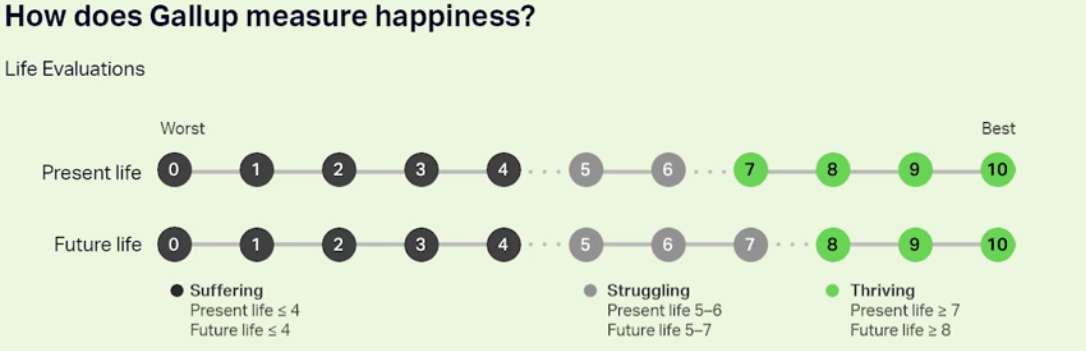
We are looking to explore the relationship between country happiness metrics and country healthcare indicators. Traditionally thought of us as being defined by wealth, we are looking to explain the intra-country happiness variance through healthcare indicators. The aim is to extend studies documented on Kaggle with new variables. Our motivation starts with Maslow’s hierarchy of needs, where we can find Physiological and Safety needs ranking higher than Psychological and Self-actualisation needs.

Our research finds several key variables that impact country level happiness, including substance abuse (tobacco and alcohol), infrastructure concerns (road traffic & clean fuel/technology) and healthcare quality-affordability (% spent vs probability of death from disease). We believe these findings should motivate discussions around addressing these concerns, particularly given how directly government controls can impact most of them.

## **Project Motivation**

**Happiness Definition**

The independent variable we are looking to explain is country happiness. A qualitative metric, happiness can either be proxied through various variables or estimated through surveys. As part of our research, we focus on the Gallup Happiness metric. Conducting an annual poll, Gallup asks respondents rate their current lives on a 0 to 10 scale, 10 being the best possible life and 0 being the worst possible life.

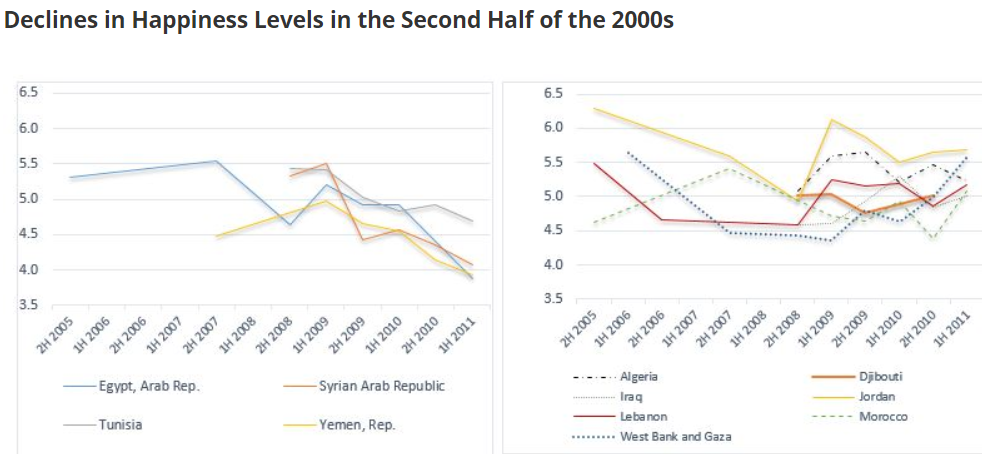


Source: Gallup, <https://news.gallup.com/poll/266057/gallup-global-happiness-center.aspx>

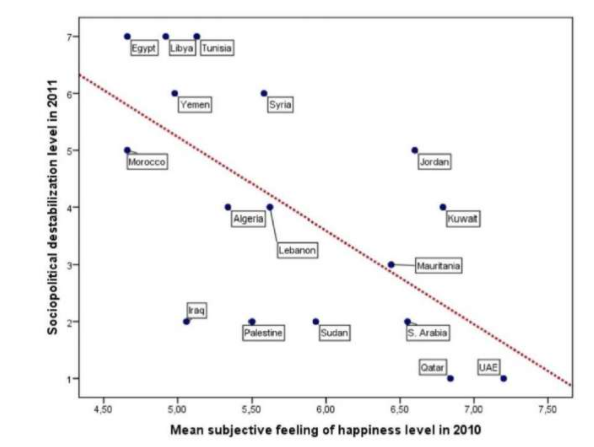
This data set is available through Kaggle (<https://www.kaggle.com/unsdsn/world-happiness>) and as part of annual reports prepared by Gallup and the United Nations Sustainable Development Solutions Network (“SDSN”, <https://worldhappiness.report/>). We make use of happiness data for 150+ countries, choosing their 2020 value over earlier data points.

**Importance of measuring happiness to predict population discontent**

Although difficult to observe, population happiness is a key variable for governments. As evidenced by World Bank research[[1]](#footnote-1), negative happiness trends have pre-empted major protests in the Arab world known as the ‘Arab Spring’. Similarly, HSE research has indicated 2010 happiness levels provided a much more accurate forecast of the Arab Spring than macro indices such as employment of GDP/capita data.[[2]](#footnote-2)



Source: Gallup, World Bank



Source: National Research University Higher School of Economics

**Trends in National Happiness over 2015-2020**

Looking over 2015-2020, we find an overall increase in world happiness, calculated as the equal weighted mean across 150+ countries.

Source: Gallup

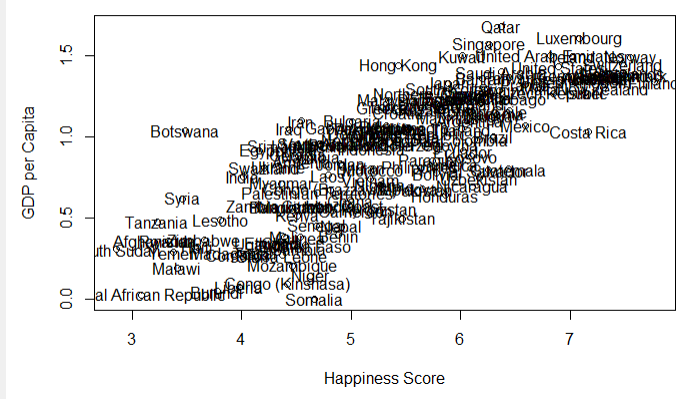
We find that this trend is a balance of several regions’ deterioration and improvement:

* South Asia, Commonwealth of Independent States (ex-Soviet Union countries) and Southeast Asia look largely unchanged over the time period
* Central and Eastern Europe, Western Europe, Sub-Saharan Africa, and East Asia look to have recorded significant gains over 2015-2020
* North America and ANZ, Middle East and North Africa & Latin America and Caribbean look to have deteriorated over 2015-2020

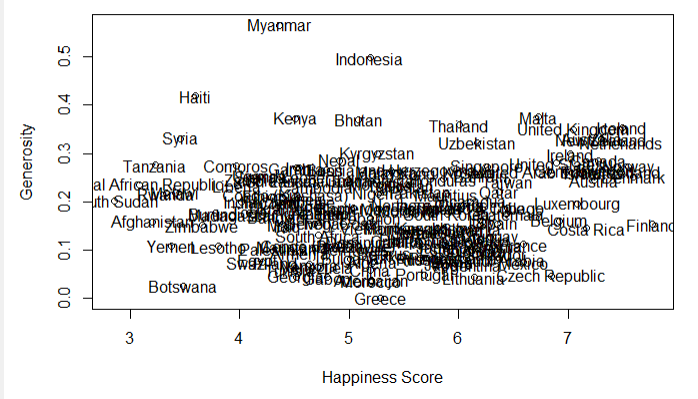
**Covariate universe selection**

We continue by defining our universe of covariates. We begin by utilizing the broad dataset of country healthcare metrics provided by Kaggle which we then narrow down to a subset of countries and variables for which we have significant data. Variable definitions are presented in the appendix, including the Kaggle definitions, with removed variables highlighted in red. Previous research investigated the Happiness score’s relationship with several key families of variables – wealth (GDP), health (life expectancy), freedom, family, generosity and trust (perception of corruption).

As outlined in the chart below we can see significant positive linear relationship between happiness and GDP per Capita. Clearly, people living in Qatar have a ‘better’ life than those living in Syria or South Sudan. We are more interested in explaining the health-happiness relationship to focus on metrics that governments can change in the short to medium term to improve their positioning relative to their peers and not unique cases. For example, in the chart below you can see Botswana and Costa Rica exhibiting similar transformed GDP per capita but very different Happiness scores.



Similar linear relationships are found for Family, Life Expectancy and Freedom. Government Corruption and Generosity look to have less of an impact compared to other variables and/or non-linear relationships. Generosity’s relationship with Happiness is outlined below:



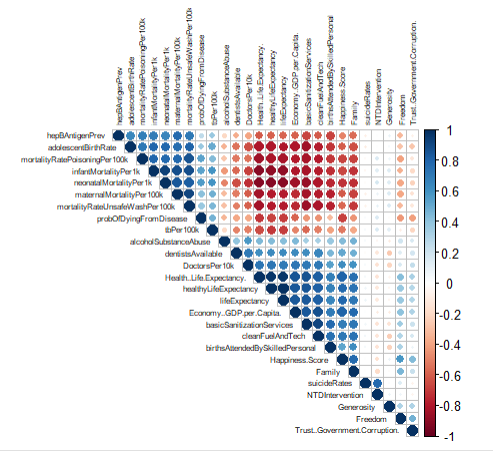
## **Data Review & Transformation**

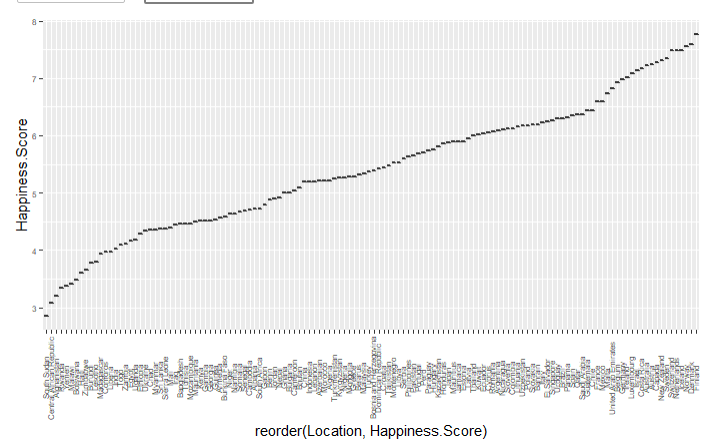
**Data Transformation**

Having reviewed our list of covariates, we find a universe of 150+ countries that needs to be narrowed down to between 80-90 if we are to make use of the healthcare dataset. Similarly, our initial universe of 30+ covariates is also reduced to reflect the lack of data available in certain countries. We remove fields such as basic handwashing at home %, malaria incidence per 1k, Modern Family Planning %, Sanitation Services per population and new H Infections per 1k. Data timeliness is addressed by utilizing the latest available date for each covariate with mean-inputting for variables for which we cannot find data.

**Exploratory Data Analysis**

Conducting an EDA, we investigate our broad data set to find a range of approximately 3-8 on the happiness scale for our country universe with extremes defined by South Sudan and Finland. We then investigate the cross-correlation between variables, including those mentioned in the initial





Our dataset exhibits some degree of inter variable relationships (>0.5 or <-0.5 correlation) which we appreciate as unavoidable within a broader healthcare dataset. Clearly, life expectancy will rise as neonatal mortality rates fall.

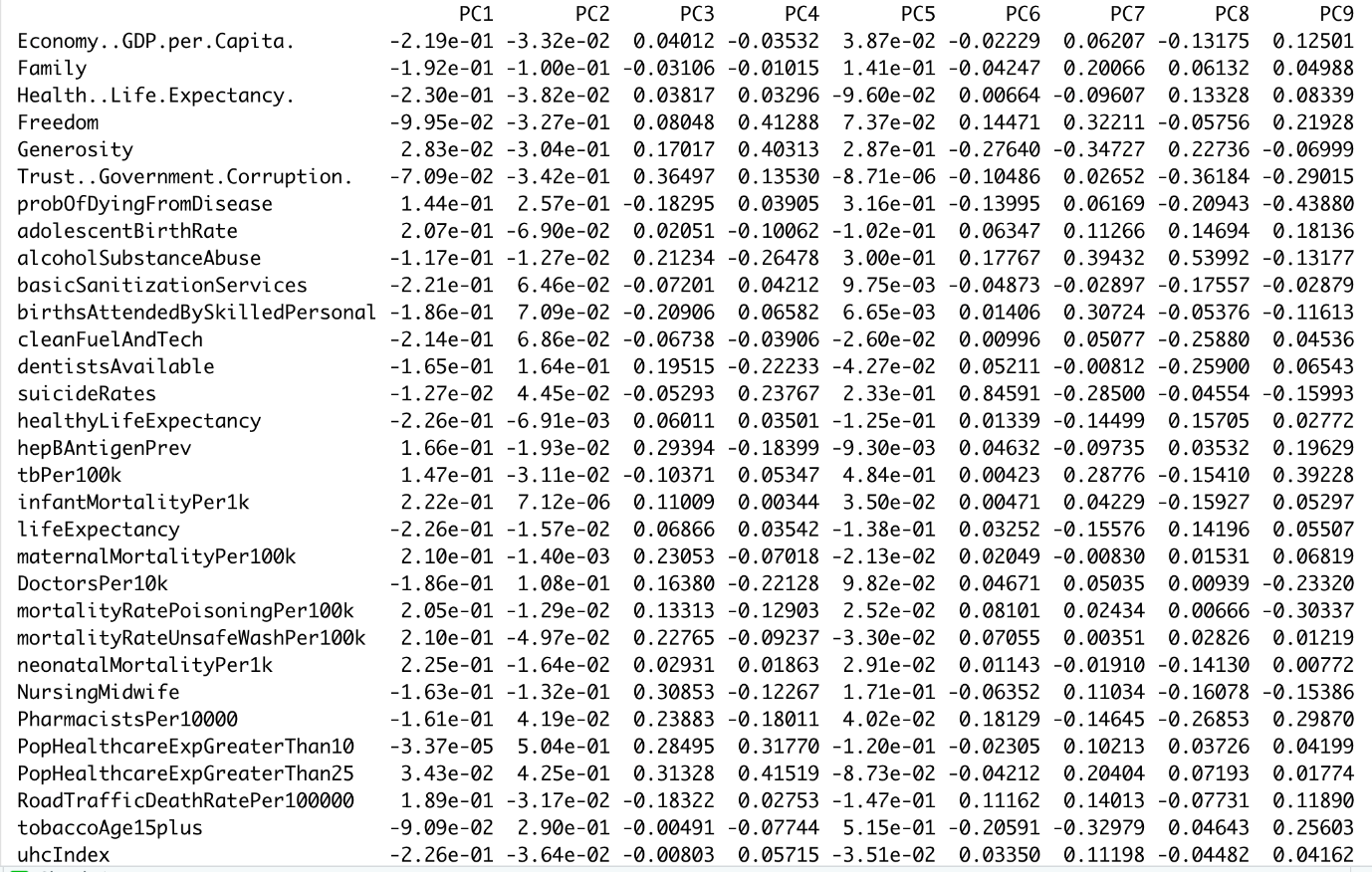
**PCA & K-means**

We perform Principal Components Analysis on our data (including the previous study’s variables, only excluding one variable NTD interventions which had extreme outliers) to investigate how countries separate from one another. Given PCA and k-means clustering, we see that the optimal number of clusters for our data is 3. When visualizing these clusters over the PCA plot (PC 2 vs. PC 1), we see that our data separate into the three clusters roughly along the first principal component.

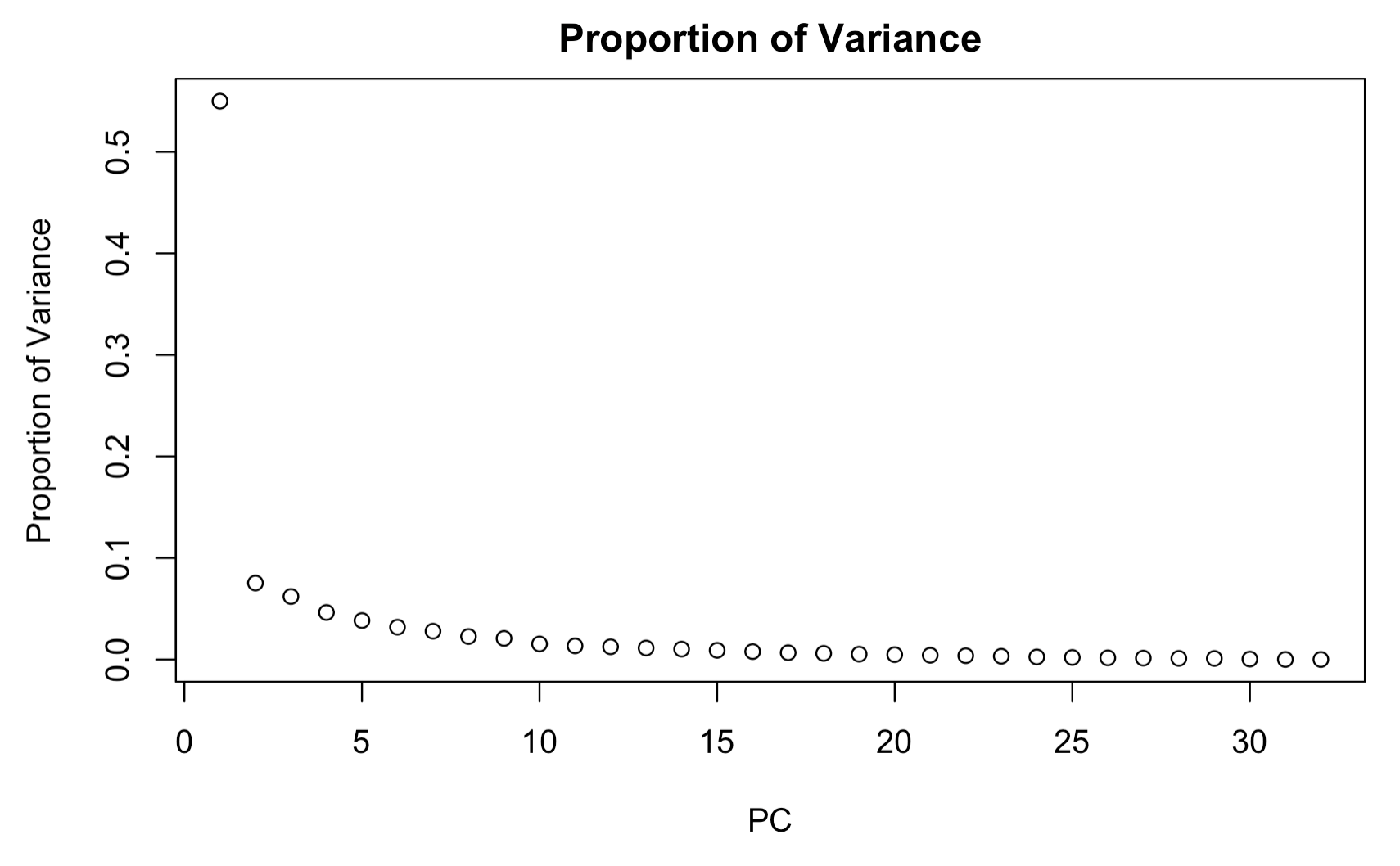
Timeline

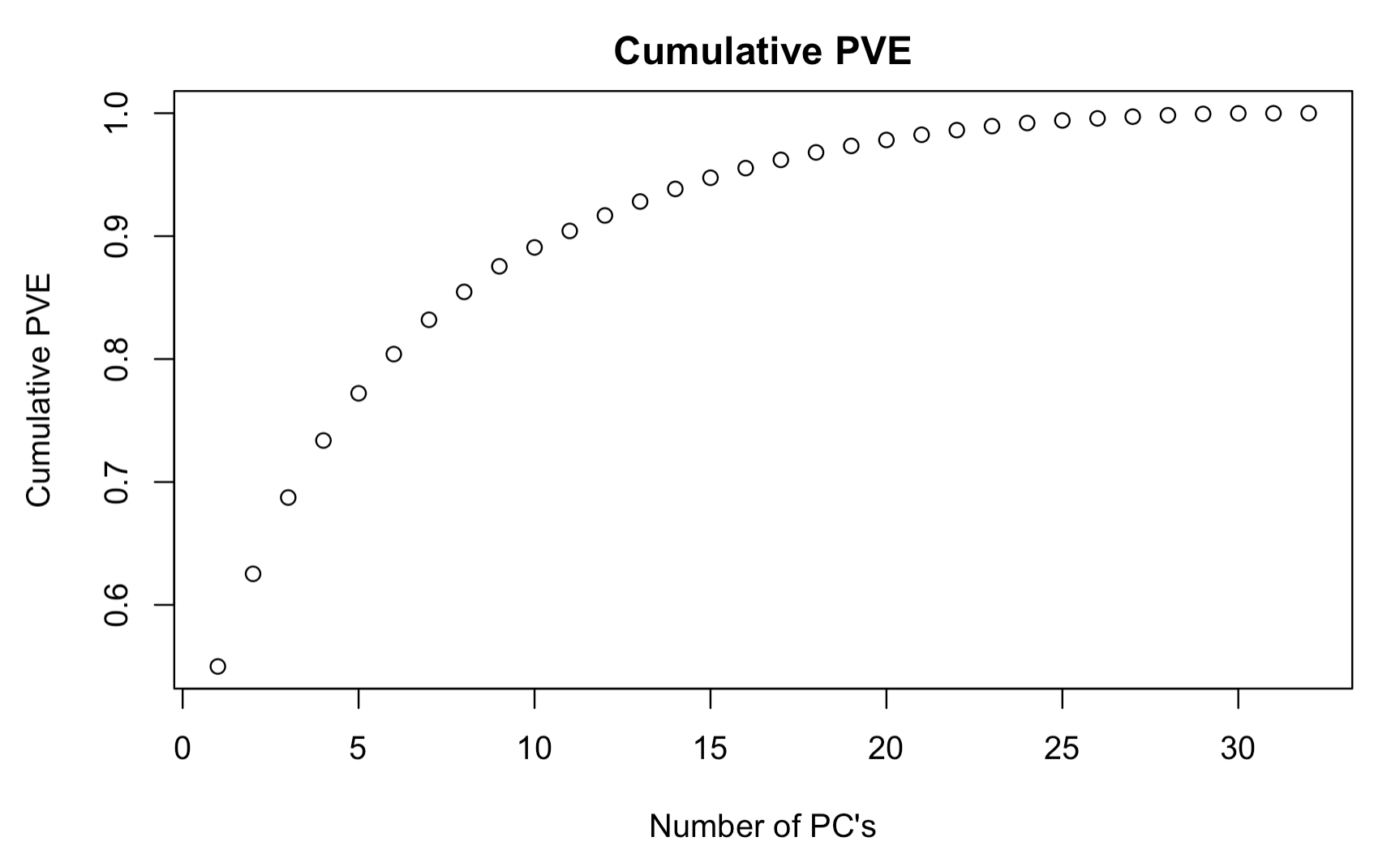
Description automatically generated with low confidence

Given the large number of variables involved, it is difficult to make overarching interpretations of the PCA weights, but we take comfort in observing relative similarity of components (absolute weights) in PC1 indicating some merit in retaining them for further analysis. As exhibited in the table below, weights range from -0.2 to 0.2 for most covariates in PC1.



This is all the more positive given the significant proportion of variance explained by PC1, which alone explains more than half the variance in our happiness dataset.





## **Regression Analysis**

**Cleaning Data**

To run Lasso, we need to further clean the dataset by getting rid of countries with any missing values for any of our variables. We also exclude the variables from the World Happiness Report (e.g., GDP, corruption) since these were already examined in the report. These eliminated variables do not fit into our study’s goal of examine the relationship between country happiness and country healthcare indicators and would most likely not reveal any new takeaways.

**Lasso Model**

For our Lasso model, we use glmnet to output results for different λ values. We can examine the default plot to explore possible values of λ.

Chart, histogram

Description automatically generated

We can see from this plot that lambda.min, which gives the smallest cvm, is ~0.07 and outputs ~8 nonzero β’s. Lambda.1se, the largest λ whose cvm is within the cvsd bar for the lambda.min value is ~0.12 and also outputs ~8 nonzero β’s. There is minimal difference here in terms of model size, and we select lambda.1se to use in our model.

We output variables from lambda.1se and run a linear model with the 8 selected variables. Our initial model using these variables produces the following output:

Call:

lm(formula = lm.input, data = data.happiness)

Residuals:

Min 1Q Median 3Q Max

-1.39938 -0.32213 0.05825 0.40495 0.88986

Coefficients:

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

(Intercept) 5.232608 1.820656 2.874 0.005236 \*\*

probOfDyingFromDisease -0.033839 0.018657 -1.814 0.073605 .

alcoholSubstanceAbuse 0.045335 0.019447 2.331 0.022362 \*

cleanFuelAndTech 0.009349 0.004082 2.290 0.024753 \*

lifeExpectancy 0.019218 0.023368 0.822 0.413387

PopHealthcareExpGreaterThan10 -0.020303 0.009223 -2.201 0.030703 \*

RoadTrafficDeathRatePer100000 -0.041663 0.011996 -3.473 0.000848 \*\*\*

tobaccoAge15plus -0.025841 0.007707 -3.353 0.001243 \*\*

uhcIndex -0.001042 0.011233 -0.093 0.926357

---

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

Residual standard error: 0.541 on 77 degrees of freedom

Multiple R-squared: 0.767, Adjusted R-squared: 0.7428

F-statistic: 31.68 on 8 and 77 DF, p-value: < 2.2e-16

While this is a fairly strong model, we see that a number of the selected variables are not significant at the .05 level. We need to take additional steps to fine tune the model and eliminate some variables which are not significant.

**Fine-tuning**

First, we use regsubsets to fine tune our model, taking the predictors from our LASSO output and lowering the dimension. We then use Cp to cut the dimension further.

Chart, scatter chart

Description automatically generated

We choose to use the model that minimizes Cp, resulting in a model size of 6. We find the set of variables with min Cp and use these to fit the next model.

Call:

lm(formula = Happiness.Score ~ ., data = data.happiness[, c("Happiness.Score",

cp.var)])

Residuals:

Min 1Q Median 3Q Max

-1.49571 -0.33069 0.05584 0.42266 0.93652

Coefficients:

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

(Intercept) 6.736043 0.399032 16.881 < 2e-16 \*\*\*

probOfDyingFromDisease -0.044038 0.013626 -3.232 0.00179 \*\*

alcoholSubstanceAbuse 0.045474 0.018308 2.484 0.01511 \*

cleanFuelAndTech 0.010324 0.002504 4.123 9.17e-05 \*\*\*

PopHealthcareExpGreaterThan10 -0.018938 0.008987 -2.107 0.03826 \*

RoadTrafficDeathRatePer100000 -0.046325 0.010532 -4.399 3.37e-05 \*\*\*

tobaccoAge15plus -0.024630 0.007489 -3.289 0.00150 \*\*

---

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

Residual standard error: 0.5366 on 79 degrees of freedom

Multiple R-squared: 0.7648, Adjusted R-squared: 0.7469

F-statistic: 42.82 on 6 and 79 DF, p-value: < 2.2e-16

Our new model includes 6 variables, which are all significant at the .05 level. As a result, we do not need to do any additional tuning to further eliminate insignificant variables. If we still had insignificant variables at this stage, we would reduce the model further using a method such as backwards selection until all variables were significant. Since all variables are significant, we can take this as our final model.

We next do a quick model diagnosis to test our linear model assumptions.

Chart, scatter chart

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

Our residual plot does not show any clear pattern, and our normal Q-Q plot is for the most part linear, other than a slight deviation at the high end. While the linear model assumptions are not perfect, we can say from these plots that they are reasonably satisfied.

## **Model Interpretation**

Our final model for predicting happiness based on health factors includes six variables:

* Probability of dying from common diseases between the ages of 30 and 70
* Alcohol substance abuse per capita
* Proportion of population with primary reliance on clean fuels and technologies
* Proportion of population with household expenditures on health greater than 10% of total household income
* Road traffic death rate
* Prevalence of tobacco use among people 15 and older

For the most part, these variables all impact happiness in our model in the way which we would expect. An increase in the use of clean fuel and technologies results in an increase in happiness. While there are several reasons why this could be the case, use of clean fuel and technologies could have an impact on overall health, as it most likely impacts pollution levels. It is possible that countries that use more clean fuel and technologies have lower levels of respiratory diseases. Countries that use clean technologies may also be more concerned with factors like the environment and citizen health, which could be related to higher happiness.

Increases in probability of dying from diseases, proportion of population with healthcare expenditures greater than 10% of total household income, road traffic death rate, and prevalence of tobacco use all resulted in decrease happiness. These are all fairly predictable, as decreasing any of these variables would generally lead to an expected increase in quality of life. There may also be some connection between these variables. For example, we might expect countries with higher tobacco use to spend more on healthcare. The largest negative coefficient was road traffic death rate. It is possible that a higher road traffic death rate could be linked to worse overall infrastructure/development in a country, which we would expect in less happy countries.

The one interesting relationship we observe is that alcohol substance abuse has a positive relationship with happiness. We would generally expect alcoholism to be more prevalent in countries that are unhappy, so this is surprising to observe. It is difficult to pinpoint the exact reason for this. This could be because of the interaction with the other variables, or it could be that richer countries generally consume more alcohol overall, and richer countries tend to be happier.

## **Appendix**

|  |  |  |
| --- | --- | --- |
| Subset | Variable | Definition |
| **Life expectancy and Healthy life expectancy** | lifeExpectancyAtBirth | Life expectancy at birth, country wise mentioned in age (years). |
| HALElifeExpectancyAtBirth | Healthy life expectancy (HALE) at birth, country wise mentioned in age(years) |
| WHOregionLifeExpectancAtBirth | Life expectancy at birth, Region wise mentioned in age (years). |
| HAleWHOregionLifeExpectancy | Healthy life expectancy at birth, region wise mentioned in age(years) |
| %HaleInLifeExpectancy | Healthy life and life expectancy at birth with the % of HALE in life expectancy |
| **Maternal, newborn and child mortality** | maternalMortalityRatio | Maternal mortality ratio per 100,000 births |
| birthAttendedBySkilledPersonal | Births attended by skilled personals (percentile) |
| infantMortalityRate | Probability of dying between birth and age 1 per 1000 live births. |
| neonatalMortalityRate | Probability of children dying in the first 28 days of life. |
| under5MortalityRate | Probability of children dying below the age of 5 per 1000 live births. |
| **Communicable Diseases** | incedenceOfMalaria | Malaria incidence per 1000 population at risk |
| incedenceOfTuberculosis | Incidence of TB per 100,000 population per year. |
| hepatitusBsurfaceAntigen | Hepatitis B surface antigen (HBsAg) prevalence among children under 5 years) |
| interventionAgianstNTD's | Reported number of people requiring interventions against NTDs. |
| newHivInfections | New HIV infections per 1000 uninfected population |
| **Noncommunicable diseases and mental health** | 30-70cancerChdEtc | Probability of dying between the age of 30 and exact age of 70 from any of the cardiovascular disease, cancer, diabetes, or chronic respiratory disease. |
| crudeSuicideRates | Crude suicide rates per 100,000 population |
| **Substance abuse** | AlcoholSubstanceAbuse | Total (recorded + unrecorded) alcohol per capita (15 +) consumption’s |
| **Road Traffic Injuries** | roadTrafficDeaths | Estimated road traffic death rate per 100,000 population |
| **Sexual and reproductive health** | reproductiveAgeWomen | Married or in-union women of reproductive age who have their need for family planning satisfied with modern methods (%) |
| adolescentBirthRate | Adolescent birth rate per 1000 women aged 15-19 years |
| **Achieve universal health coverage (UHC) including financial risk protection** | uhcCoverage | UHC index of service coverage (SCI) |
| dataAvailibilityForUhc | Data availability of UHC index of essential service coverage (%) |
| population10%SDG3.8.2 | Population with household expenditures on health greater than 10% of total household expenditure or income (SDG indicator 3.8.2) (%) |
| population25%SDG3.8.2 | Population with household expenditures on health greater than 25% of total household expenditure or income (SDG indicator 3.8.2) (%) |
| **Mortality from environment pollution** | airPollutionDeathRate | Ambient and household air pollution attributable death rate per 100,00 population and the same data with age-standardized. |
| mortalityRateUnsafeWash | Mortality rate attributed to exposure to unsafe WASH services per 100,000 population SDG3.9.2 |
| mortalityRatePoisoning | Mortality rate attributed to unintentional poisoning per 100,000 population |
| **Tobacco control** | tobaccoAge15 | Prevalence of current tobacco use among persons aged 15 years and older (age- standardized rate) |
| **Health Workforce** | medicalDoctors | Medical doctors per 10,000 population. |
| nursingAndMidwife | Nursing and midwifery personnel per 10,000 population. |
| Dentists | Dentists available per 10,000 population |
| Pharmacists | Pharmacists per 10,000 population |
| **Eliminate violence Against women and girls** | eliminateViolenceAgainstWomen | Proportion of ever-partnered women and girls aged 15-49 years subjected to physical and/or sexual violence by a current or former intimate partner in previous 12 months. |
| **Drinking Water** | basicDrinkingWaterServices | Population using at least basic drinking water services (%) |
| **Sanitation and Hygiene** | atLeastBasicSanitizationServices | Population using at least basic sanitation services (%) |
|  | safelySanitization | Population using safe sanitation services (%) |
|  | basicHandWashing | Population with basic handwashing facilities at home (%) |
| **Clean household energy** | cleanFuelAndTech | Proportion of population with primary reliance on clean fuels and technologies (%) |

Source: Kaggle & WHO

1. World Bank, <https://blogs.worldbank.org/arabvoices/developing-growing-less-happy-what-explains-paradox-arab-world> [↑](#footnote-ref-1)
2. National Research University Higher School of Economics, https://phys.org/news/2019-12-happiness-arab.html [↑](#footnote-ref-2)