Happiness Data EDA

Horvath, Laurent, Smith

2023-07-23

Introductory Statements:

While many analytical efforts are best served by a method- and results-agnostic approach to data analysis, formulating clear research questions in advance of data exploration can provide a focused environment in which these questions can be answered accurately and in-depth. In pursuit of this, two questions describe our approach to characterizing the relationships among a variety of national happiness indicators from the 2019 UN World Happiness Report, as well as several metrics published on the national level by the World Bank, also in 2019. Though in our case the following research questions were formulated ex post facto, one could certainly seek answers to them as an original motivation:

1. Which metric(s) from the UN Report was *most* meaningful in predicting a nation's Happiness Score?
2. Can we identify variables from an entirely different source of data (World Bank metrics) which are also meaningful predictors of Happiness Scores?

The following variables from the UN and World Bank are the subject of this paper:

```r
(country.info<-variable.names(data)[1:3])
```

```
## [1] "country_name" "country_code" "region"
```

```r
happy.info<-variable.names(data)[4:11]
wb.info<-variable.names(data)[13:18]
```

For additional clarity, below are "full" names and explanations of variables from 'happy.info' and 'wb.info':

- "happy_rank" - Ranking of happiness score
- "happy_score" - Aggregate happiness score calculated from all other factors
- "happy_gdpc" - GDP per capita
- "happy_supp" - Sense of social support
- "happy_health" - Healthy life expectancy at birth
- "happy_free" - Happiness with level of personal freedom
- "happy_gen" - How often people contribute to charitable causes
- "happy_trust" - Trust level that own national government is not corrupt
- "wb_pov" - % of population below UN international poverty rate
- "wb_unemp" - % of able-bodied labor force unemployed
- "wb_elec" - % of population with access to electricity
- "wb_renew" - % of final energy use from renewable sources
- "wb_hom" - Homicide rate per 100,000 people
- "wb_debt" - National debt as % of government GDP

Variable Assessments:

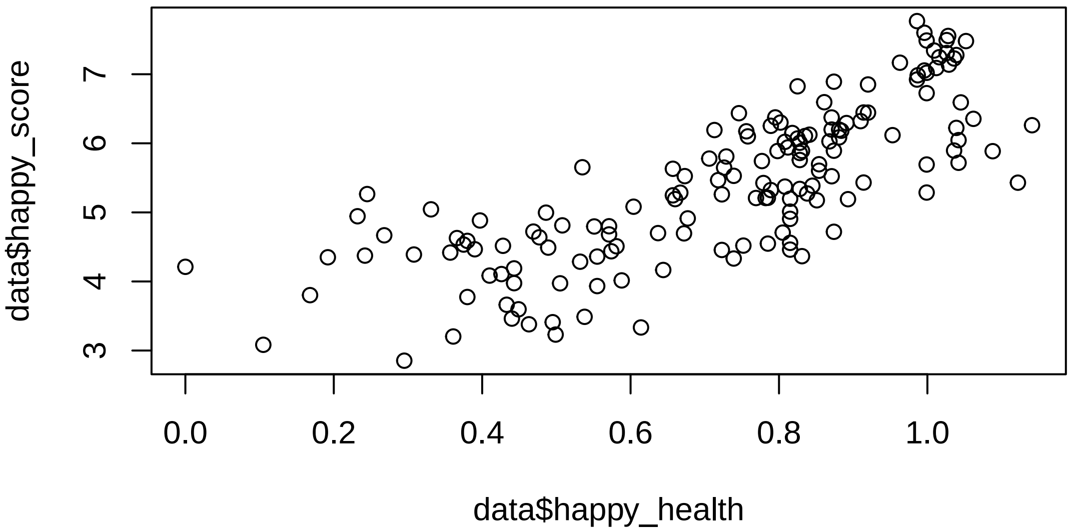
We began with tests of both homoscedasticity and normality for our many variables in order to determine their validity for use in more advanced analyses. A quick review of these ideas:

1. Homo/Heteroscedasticity refers to the degree to which the variability of the data points is roughly constant across the range. To evaluate this, we plot the variable against the primary dependent variable 'happy__score'. If the resulting plot appears consistent in shape across the range of the chart, without "fan-like" shapes representing a change in the variance of the data, it suggests that a given predictor variable has homoscedasticity. This is one of the core assumptions of Ordinary Least Squares-based regression, hence the need to test for this it. We also evaluated more formally by constructing a bivariate linear regression model for each predictor variable and applying the Studentized Breusch-Pagan test, with the null hypothesis that homoscedasticity is present.

2. Normality refers to the assumption that the data follows a normal distribution when plotted on a histogram or Q-Q Plot and is a precondition for many parametric statistical tests like t-tests. It is also a core assumption of Ordinary Least Squares-based linear regression. Here we use the common Shapiro-Wilk test based on the correlation between the data and the expected values for a normal distribution. It calculates a 'W' statistic with the null hypothesis that the data *are* normally distributed.

Here we show testing for a few different variables in order to demonstrate some greater trends we later identified:

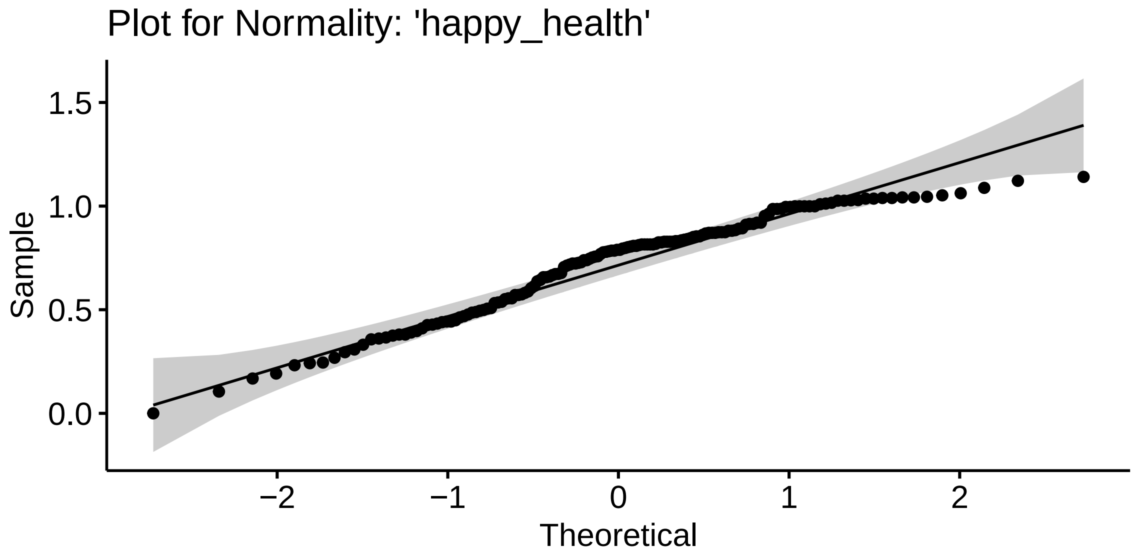
```
# 'happy_health': Healthy life expectancy at birth
plot(x=data$happy_health, y=data$happy_score)  # Scatterplot: homoscedasticity
```



```
bptest(health_lin_model)  # Breusch-Pagan - significant p-value indicates heteroscedasticity
```

```
##
##  studentized Breusch-Pagan test
##
## data:  health_lin_model
## BP = 0.41888, df = 1, p-value = 0.5175
```

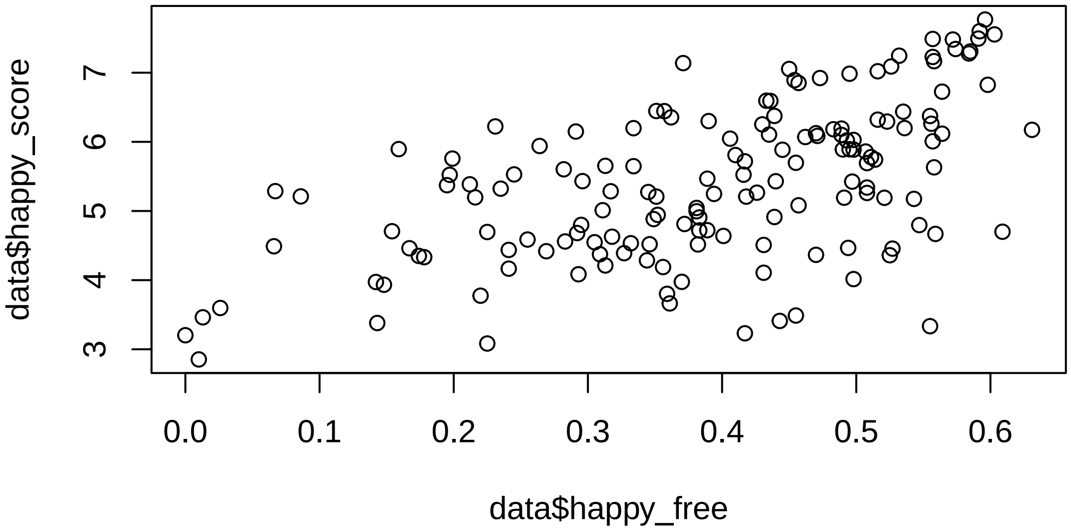
```
###
### NOTE: THESE CODEBLOCKS ARE REPRESENTATIVE OF METHODS FOR EACH VARIABLE
### As such, we omit the code for successive tests as it is near-identical
###
# Q-Q plot: normality
ggqqplot(data$happy_health, title = "Plot for Normality: 'happy_health'") # Q-Q plot: normality
```



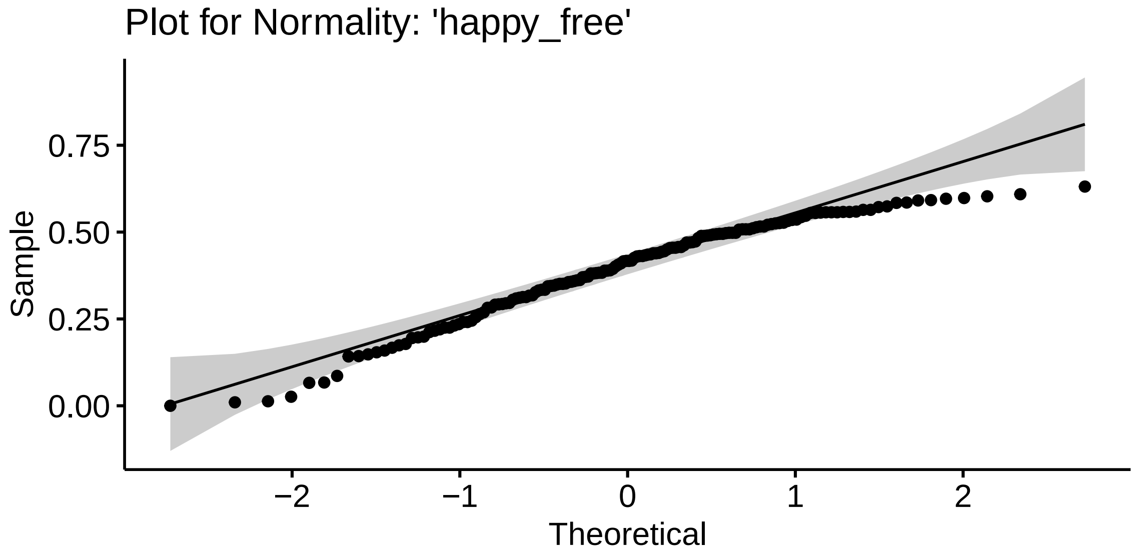
```
(shapiro_health <- shapiro.test(data$happy_health)) # Shapiro-Wilk test: normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$happy_health
## W = 0.95341, p-value = 4.498e-05
```

- **Conclusion:** ‘happy_health’ is **homoscedastic**, but is not normally distributed.

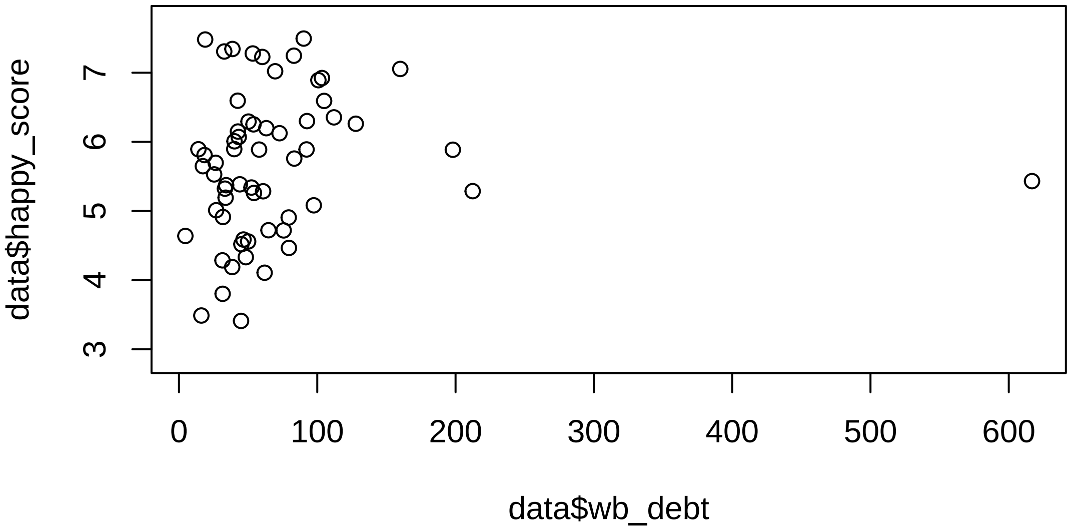


```
##
##  studentized Breusch-Pagan test
##
## data:  free_lin_model
## BP = 3.9933, df = 1, p-value = 0.04568
```

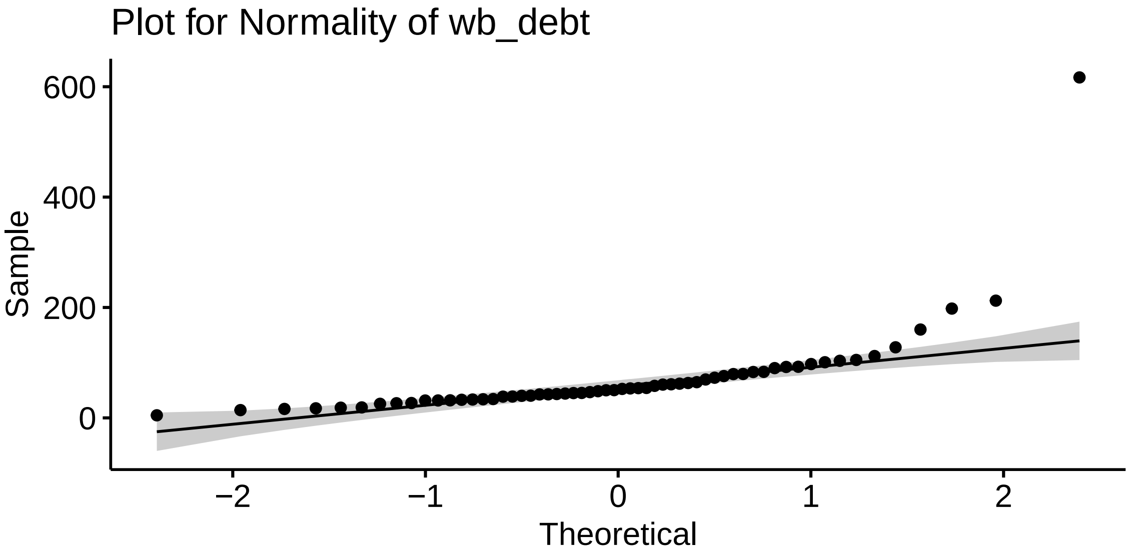

Plot for Normality: 'happy_free'

```
##
##  Shapiro-Wilk normality test
##
## data:  data$happy_free
## W = 0.9543, p-value = 5.386e-05
```

- **Conclusion:** 'happy_free' is **heteroscedastic**, and is not normally distributed.



```
##
##  studentized Breusch-Pagan test
##
## data:  debt_lin_model
## BP = 0.47173, df = 1, p-value = 0.4922
```



```
##
##  Shapiro-Wilk normality test
##
## data:  data$wb_debt
## W = 0.49949, p-value = 5.378e-13
```

- **Conclusion:** ‘wb_debt’ is homoscedastic and not normally distributed. More on this soon

Through application of these methods to **all** available variables, we found we have 3 "categories" of predictors in terms of our primary dependent variable 'happy_score':

1. **Significant predictors of Happiness** which were *Homoscedastic*
2. **Significant predictors of Happiness** which were *Heteroscedastic*
3. **Non-significant predictors** of 'happy_score'

Our analysis also revealed that *none* of our predictor variables are normally distributed, and several are heteroscedastic. Additionally, Q-Q Plots showed that several had notable outliers. As a result, we understand that methods relying on Ordinary Least Squares-based calculations were not appropriate choices, and chose to employ Robust Regression and Principal Component Analysis for more advanced analyses .

Correlation Table and Discussion:

Here we assess *all* of the predictor variables in the context of 'happy_score' to check for pairwise significance and linearity. The following table of correlation coefficients and corresponding p-values (generated using the cor() method) allowed us to assess whether our pairwise-significant predictors appeared to possess meaningful linearity, as indicated by high absolute values of correlation coefficients.

Finally, we need to consider the assumption of independence, which (at least initially) is a more qualitative discussion.

Correlation Matrix:

```r
# First, we do some decile-level imputation of 'wb_pov', as this was our original method,
# and we follow with a different method for the remaining variables in later analyses.
wb_pov_index <- which(names(data) == "wb_pov")
data <- data.frame(data[, 1:wb_pov_index], wb_pov_imp = NA, data[, (wb_pov_index + 1):ncol(data)])
data$wb_pov_imp <- data$wb_pov
decile_medians <- data %>% group_by(happy_dec) %>%
  summarize(decile_median_wb_pov = median(wb_pov, na.rm = TRUE))

for (i in 1:nrow(decile_medians)) {  # For each row in  the decile_medians frame just built
  decile <- decile_medians$happy_dec[i]  # Fetch the decile value from decile_medians

  # Then, fetch the actual decile-level median *value* from decile_medians and store in median_val
  median_val <- decile_medians$decile_median_wb_pov[i]

  # Fill in missing values in 'wb_pov_imp' for rows with a matching 'happy_dec' value
  # with the calculated median
  data$wb_pov_imp[data$happy_dec == decile & is.na(data$wb_pov_imp)] <- median_val
}

# Start with a list of fields we want to correlate with happy_score
field_list <- c("happy_gdpc", "gdpc_change", "happy_supp", "happy_health", "happy_free",
                "happy_gen", "happy_trust", "wb_pov_imp", "wb_unemp", "wb_elec", "wb_renew",
                "wb_hom", "wb_debt")

# Create an empty matrix to store the calculated correlation coefficients
cor_matrix <- matrix(NA, nrow = length(field_list), ncol = 2,
                     dimnames = list(field_list, c("Correlation", "p-value")))

# Loop through the variables and calculate correlation coefficients with happy_score
for (i in 1:length(field_list)) {
  correlation_result <- cor.test(data$happy_score, data[[field_list[i]]])
  cor_matrix[i, "Correlation"] <- correlation_result$estimate
  cor_matrix[i, "p-value"] <- correlation_result$p.value
}
# Result is a matrix (cor_matrix) sorted by field order in the dataframe from left to right
# Now, sort p-values in descending order *of significance* (most significant precdictors on top)
# and store in a new matrix
(sorted_cor_matrix <- cor_matrix[order(cor_matrix[, "p-value"], decreasing = FALSE),])
```

```
##              Correlation      p-value
## happy_gdpc    0.79388287 4.315481e-35
## happy_health  0.77988315 3.785454e-33
## happy_supp    0.77705779 8.975120e-33
## wb_pov_imp   -0.62074645 5.443402e-18
## wb_elec       0.59809642 2.604329e-16
## happy_free    0.56674183 1.237924e-14
## wb_renew     -0.40336896 2.342823e-07
## happy_trust   0.38418210 7.376013e-07
## wb_unemp     -0.26102388 1.076600e-03
## gdpc_change   0.13799002 8.789017e-02
## happy_gen     0.07582369 3.468195e-01
## wb_debt       0.09531356 4.688100e-01
## wb_hom       -0.06665641 4.930814e-01
```

From these results, we see that Significant pairwise predictors of 'happy_score' are:

1. happy_gpdc
2. happy_health
3. happy_supp
4. wb_pov_imp
5. wb_elec
6. happy_free
7. wb_renew
8. happy_trust
9. wb_unemp

All the above predictors exhibit meaningful linearity. So, on a *bivariate* level, these correlations are meaningful to us in terms of strength of relationship with 'happy_score'.

Among the significant pairwise predictors, the following *do* exhibit homoscedasticity:

1. **happy_health**
2. happy_gpdc
3. happy_supp
4. wb_elec

Among the significant pairwise predictors, the following *do not* exhibit homoscedasticity:

1. **happy_free**
2. wb_pov_imp
3. wb_renew
4. happy_trust
5. wb_unemp

The following are *not* significant pairwise predictors:

1. **wb_debt**
2. gdpc_change
3. happy_gen
4. wb_hom

Note, once again, that the "model" variables continue to display results consistent with initial observations. Additionally, on the advice of our supervisor, we've included and will continue to analyze 'gpdc_change' - a new variable of our own construction - which was built from comparison of 2018 and 2019 GDPC data.

The Assumption of Independence

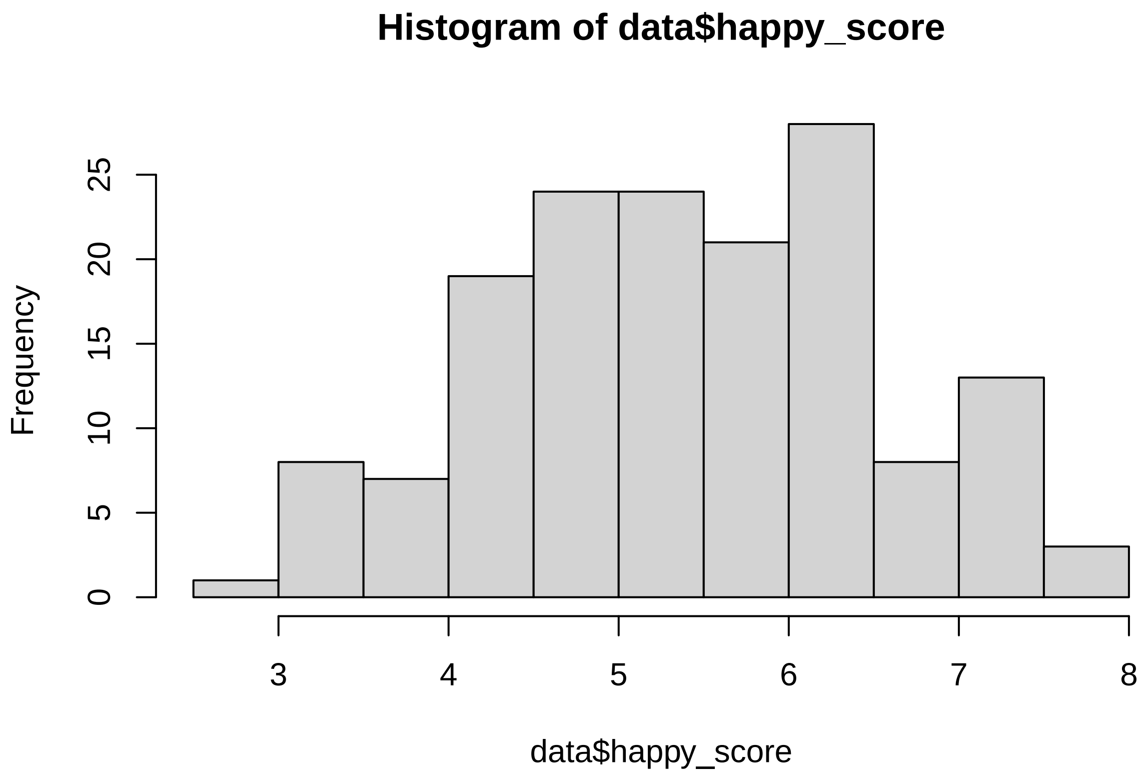
Because many analytical methods depend upon the independence of observations, it was incumbent upon us to assess that assumption with regard to these data. As commonly understood, the assumption of independence is such that the value of one observation neither affects nor provides any information about the value of another observation. In the context of our data, we have *two* perspectives on independence to consider.

We begin with the question: "Do we have reason to believe the happiness score of one country is not dependent on the happiness score of another?" Given the globalized and economically intertwined nature of the human race, it seems reasonable that if a country suddenly became very unhappy, its geographical or financial neighbors might soon follow as a result. There are meaningful dependencies between countries, such as the USA's large-scale importation of manufactured consumer goods from China, or the shuffling of agricultural products throughout European countries. Additionally, problems from one country have a history of "spilling over" into others. The Chernobyl disaster and even the modern Russo-Ukranian war have had tangible effects across other regions through radiation, agricultural disruption, and refugee movement.

We can, in fact, see some statistically provable mean-differences that support this idea. Consider the following ANOVA-based analysis:

First we assess 'happy_score' for legitimacy in one-way ANOVA: it must be approximately normal and have homogeneity of variance as defined by Levene's test.

```
hist(data$happy_score)  # Appears approximately normal
```



```
(shapiro_happy_score <- shapiro.test(data$happy_score)) # Is normal by Shapiro-Wilk
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$happy_score
## W = 0.9872, p-value = 0.1633
```

```
leveneTest(happy_score ~ region, data = data)  # p-value is 0.057. Close enough to proceed
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   9  1.8906 0.05761 .
##       146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Results indicate 'happy_score' is normal and the Levene's p-value is 0.057. Close enough to proceed, so let's look at happiness in the context of the 'region' field:

```
# Now a one-way ANOVA of happiness score against world region
summary(region_anova_model <- aov(happy_score ~ region, data = data))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## region        9  98.78  10.976   17.18 <2e-16 ***
## Residuals   146  93.27   0.639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ANOVA indicates a significant difference in happiness scores among the world regions

tukey_region_result <- TukeyHSD(region_anova_model) # Post-hoc testing via Tukey HSD
tukey_df <- as.data.frame(tukey_region_result$region) # Convert Tukey results to dataframe

sorted_tukey_df <- tukey_df[order(tukey_df$`p adj`), ] # Sort by descending adjusted p-value strength

# Filter to only statistically significant adjusted p-values
filtered_tukey_results <- sorted_tukey_df[sorted_tukey_df$`p adj` < 0.05, ]

# Check number of statistically significant relationships under Tukey
print(nrow(filtered_tukey_results))  # There are 14 significant mean differences
```

```
## [1] 14
```

```
head(filtered_tukey_results, 3) # Some examples to demonstrate:
```

```
##                                                     diff        lwr       upr
## Western Europe-Sub-Saharan Africa               2.375477  1.6840532  3.066900
## Western Europe-Middle East and Northern Africa  1.638957  0.8554486  2.422465
## Sub-Saharan Africa-Latin America and Caribbean -1.464710 -2.1663818 -0.763039
##                                                       p adj
## Western Europe-Sub-Saharan Africa              6.294965e-14
## Western Europe-Middle East and Northern Africa 1.685637e-08
## Sub-Saharan Africa-Latin America and Caribbean 1.813789e-08
```

Tukey HSD testing shows significant mean happiness differences at the World region level, and we can make some reasonable inferences about these. It makes sense that wealthy, peaceful regions (Australia/NewZealand, Western Europe) are happier than poor, war-torn regions (most of Africa, the Middle East). Therefore it's logical to assume that the very existence of these "happiness clusters" indicates nations *do* have some effect on each other's happiness levels.

Therefore, on a "macro" level, we have a strong argument *against* the assumption of independence for *nations* in terms of 'happy_score'. However, if we focus our construction on the metric 'happy_score' itself, a different question arises: "Were the happiness scores submitted by individual respondents in different countries directly affected by scores submitted elsewhere?" In other words, on a more individualistic level, "Is the response of a person in Kansas *directly* affected by the response of a person in London?" In this context, we would argue that, no, they were not. It seems quite unlikely that our Kansas respondent would have considered "Before I answer, how happy did the Laotians say they were?"

The fact that scores are aggregated at the nation-level makes it easy to overlook that respondents are not, themselves, nations, but the citizens of these nations. The data is still derived from individuals and their responses, so, taking into account the context of the study and our specific interest in the data, we conclude the following: Given that 'happy_score' is calculated from ratings provided by individual citizens, we can assume that individuals did not collude across national boundaries before responding. Hence, these individual observations comprise the nation-level scores and we can assume nation-level scores are independent. While it is certain that the overall happiness of individuals in a given country can be affected by the happiness, environments, or actions of other nations, such interdependence is an inevitable quality of modern globalized society. Allowing this to stymie analysis of the international community would preclude the existence of the field of international social research.

In sum, the assumption of independence with respect to 'happy_score' is a contextually complex one - conditional on a deeper understanding of how we can study "national happiness" at all, and an acceptance that our predictor variables are inevitably interrelated, but no less useful to us.

Results as they relate to Variable Assessment:

Let's now look again at the 3 categories of relations that we've been illustrating, and attempt some explanations for the behavior of variables they contain:

1. Significant predictor of Happiness AND Homoscedastic

'happy_supp's relationship with overall happiness is quite straightforward and logical. Broadly, and uniformly, it is reasonable to conclude that how socially supported a person feels would affect their overall happiness. We can likely all relate to this sentiment as the whole world has experienced COVID by now. Extended isolation, thought not a factor in this data, has well-documented impact on mental health and, thereby, happiness.

2. Significant predictor of Happiness AND Heteroscedastic

'wb_unemp' is an interesting case in that it is certainly a significant predictor, but certain countries exhibit results that are quite varied from "logical expectations". For example, Costa Rica (CRI) and Brazil (BRA) have relatively high happiness rankings at 12 and 32, respectively, but have high unemployment percentages of 10.81 and 12.05. Though higher unemployment is broadly a great indicator of lower happiness score, Costa Rica is known for its high happiness levels, due in part to cultural and geographical considerations such as connection to nature and an orientation toward community well-described by the local term "pura vida".

3. NOT Significant predictor of Happiness

The non-significance of, for example, 'wb_debt' is also logical; consider the countries of Iceland (ISL) and Mozambique (MOZ). Both report quite similar 'wb_debt' values of 90.18 and 79.51, respectively, but have 'happy_rank's of 4 and 123 (out of 156). Additionally, our own country is well-known for vast sums of debt, yet sits at a 'happy_rank' of 20. Beyond these specific examples, national debt is also something we intuitively understand to be a stronger influence on the relationship between countries more than the relationship a country has with its people. Some countries are deeply in debt because they are, in fact, quite poor; others are deeply in debt because they are wealthy, advanced economies using debt as an economic "lever."

Robust Regression Testing:

Maximum likelihood-type M-estimation

Assessments show that the assumptions of OLS-based regression are too badly violated by the variables to proceed with that methodology. Additionally, several of them have notable outliers.This leads us to a new analytical method: Maximum likelihood-type M-estimation. MM-Estimation is an extension of Huber's M-estimation (which is a generalization of the ordinary least squares (OLS) method such that the sum of absolute differences is minimized rather than the sum of squares). **MM-Estimation has high breakdown point properties, meaning it can handle a larger number of outliers**. It involves an initial step to obtain robust first estimates (via S-estimation or LTS-estimation), and then iteratively refines these estimates via standard M-estimation. This method is available in R via the 'robustbase' package. Note that we need to scale all of the predictor variables, because interpreting the strength of predictors based on their coefficients or t-values can be misleading, especially as predictors vary in scale or have units of measurement.

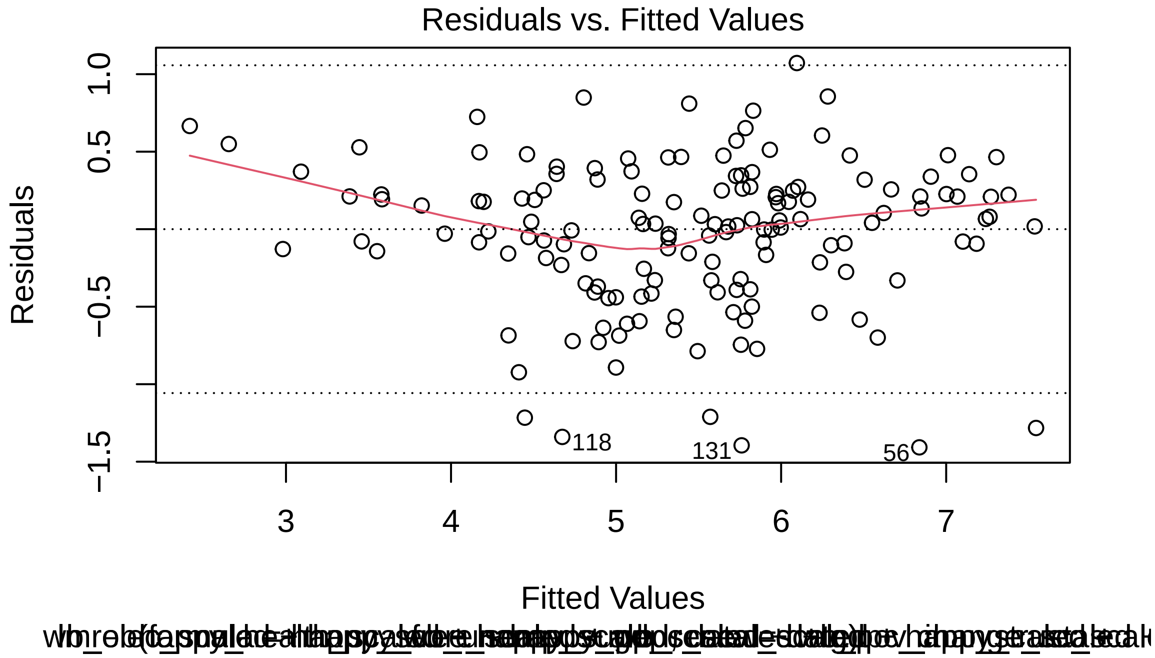
```r
# Create scaled versions of the predictors
data$happy_gdpc_scaled <- scale(data$happy_gdpc)
data$gdpc_change_scaled <- scale(data$gdpc_change)
data$happy_health_scaled <- scale(data$happy_health)
data$happy_supp_scaled <- scale(data$happy_supp)
data$wb_pov_imp_scaled <- scale(data$wb_pov_imp)
data$wb_elec_scaled <- scale(data$wb_elec)
data$happy_free_scaled <- scale(data$happy_free)
data$wb_renew_scaled <- scale(data$wb_renew)
data$happy_trust_scaled <- scale(data$happy_trust)
data$wb_unemp_scaled <- scale(data$wb_unemp)
# Running the robust regression with scaled predictors and summarizing
summary(mm_est_model <- lmrob(happy_score ~ happy_gdpc_scaled + gdpc_change_scaled + happy_health_scale
```

```
##
## Call:
## lmrob(formula = happy_score ~ happy_gdpc_scaled + gdpc_change_scaled + happy_health_scaled +
##     happy_supp_scaled + wb_pov_imp_scaled + wb_elec_scaled + happy_free_scaled +
##     wb_renew_scaled + happy_trust_scaled + wb_unemp_scaled, data = data)
##  \--> method = "MM"
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40755 -0.32969  0.02509  0.25358  1.07178
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.447369   0.044358 122.804  < 2e-16 ***
## happy_gdpc_scaled    0.360418   0.082552   4.366 2.44e-05 ***
## gdpc_change_scaled  -0.007474   0.041355  -0.181 0.856841
## happy_health_scaled  0.297762   0.070358   4.232 4.17e-05 ***
## happy_supp_scaled    0.263960   0.073883   3.573 0.000485 ***
## wb_pov_imp_scaled   -0.203876   0.080785  -2.524 0.012730 *
## wb_elec_scaled      -0.054946   0.089771  -0.612 0.541489
## happy_free_scaled    0.107358   0.055463   1.936 0.054925 .
## wb_renew_scaled      0.135611   0.061810   2.194 0.029886 *
## happy_trust_scaled   0.183402   0.060459   3.034 0.002882 **
## wb_unemp_scaled     -0.123790   0.042842  -2.889 0.004474 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 0.4229
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.8437, Adjusted R-squared:  0.8325
## Convergence in 19 IRWLS iterations
##
## Robustness weights:
##  16 weights are ~= 1. The remaining 135 ones are summarized as
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2455  0.8547  0.9424  0.8857  0.9829  0.9988
## Algorithmic parameters:
##        tuning.chi                 bb        tuning.psi        refine.tol
##         1.548e+00          5.000e-01         4.685e+00         1.000e-07
##           rel.tol          scale.tol         solve.tol          zero.tol
##         1.000e-07          1.000e-10         1.000e-07         1.000e-10
##       eps.outlier              eps.x warn.limit.reject warn.limit.meanrw
##         6.623e-04          1.233e-11         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s          k.max
##            500             50              2              1            200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0           2000
##                   psi           subsampling                 cov
##            "bisquare"         "nonsingular"         ".vcov.avar1"
## compute.outlier.stats
##                  "SM"
## seed : int(0)
```

```
plot(mm_est_model) # Plotting the model
```

```
## recomputing robust Mahalanobis distances
```

```
## saving the robust distances 'MD' as part of 'mm_est_model'
```



- Some very interesting results. In this model, which by adjusted R-squared explains about 83% of the variance in 'happy_score', the three variables 'gdpc_change', 'wb_elec', and 'happy_free' are not significant in predicting happy_score. This is quite a change from the results of the correlation table and the PCA.

Based on the absolute value of the standardized coefficients in the model, the three most important predictors are:

1. happy_gdpc_scaled (0.3621)
2. happy_health_scaled (0.2957)
3. happy_supp_scaled (0.2633)

This means that changes in GDP per capita, health, and social support (when these predictors are measured in standard deviation units) are associated with the largest changes in happiness score.

Based on the p-values in the model, the three most significant predictors are:

1. happy_gdpc_scaled (2.55e-05)
2. happy_health_scaled (4.31e-05)
3. happy_supp_scaled (0.000505)

Again we have GDP per capita, health, and social support. The fact that these variables emerge as important using both criteria strengthens the evidence for their primacy.

- We also see a meaningful role for 'wb_pov_imp' and 'happy_trust', which is borne out by the biplot of the upcoming PCA.

Principal Component Analysis:

Beyond our tests of Robust Regression, we chose to explore a Principal Component Analysis of our predictor variables as a means of dimensionality reduction and exploration of the relationships among said variables.

Principal Component Analysis (PCA) is a technique used to emphasize variation and highlight strong patterns in a data set and can be performed in R using the 'prcomp()' method. 'prcomp()' is preferred over the alternate 'princomp()' method because 'prcomp()' uses singular value decomposition, which is more numerically accurate.

Additionally, we need to standardize the predictor variables so that they're all on the same scale, mostly because PCA is sensitive to the variances of predictor variables. The 'scale()' method standardizes variables to have a mean of 0 and standard deviation of 1, supporting our approach.

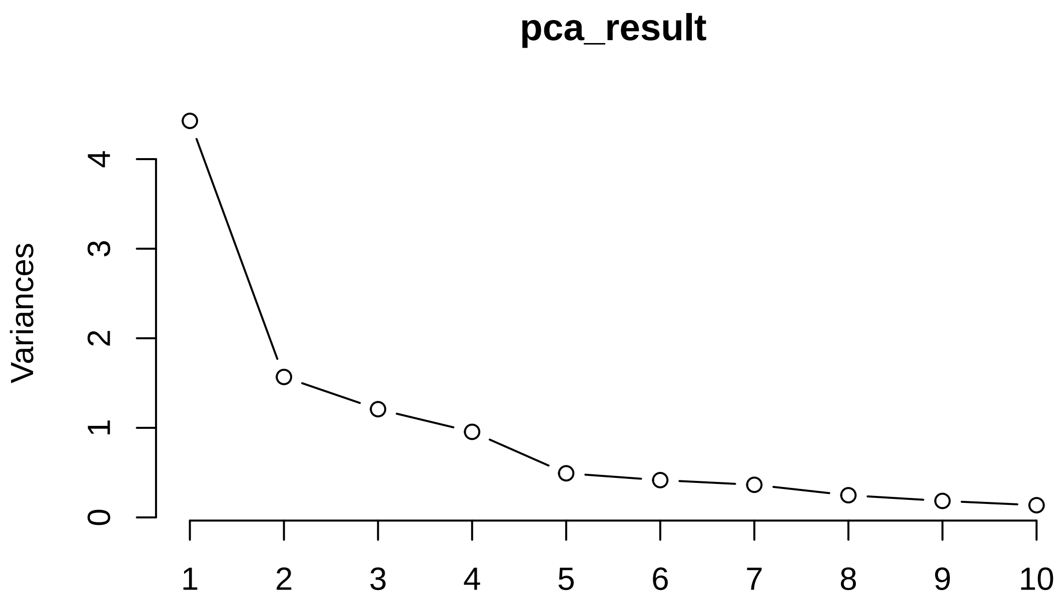
Note that PCA cannot handle missing or infinite values. So, rather than doing row-deletion, we impute the *very* small number of missing values for the significant World Bank predictors ('wb_elec', 'wb_renew', and 'wb_unemp') using the Multivariate Imputation by Chained Equations imputation method from the "mice" package.

So, let's begin:

```r
# Start by subsetting the data frame to include only the significant predictor variables
predictor_subset <- data[,c('happy_gdpc', 'gdpc_change', 'happy_health', 'happy_supp',
                             'wb_pov_imp', 'wb_elec', 'happy_free', 'wb_renew',
                             'happy_trust', 'wb_unemp')]

# Perform multivariate imputation
imputed_data <- mice(predictor_subset, m=5, maxit = 50, method = 'pmm', seed = 500)
# Syntax notes:
# 'm' is the # of multiple imputations to be created (5)
# 'maxit' is the # of iterations to be performed by the EM algorithm while creating imputations (50)
# 'method' selects method for imputation ('pmm'= Predictive Mean Matching, preferred for numeric data)
# 'seed' sets the seed for the random number generator, which makes the results reproducible
```

```r
complete_data <- complete(imputed_data, 1)  # Create a fully-imputed dataset
scaled_predictors <- scale(complete_data)   # Scale the predictors
pca_result <- prcomp(scaled_predictors)     # Perform the PCA
summary(pca_result)                         # Summarize results
plot(pca_result, type="l")                  # Generate and output Scree plot
```

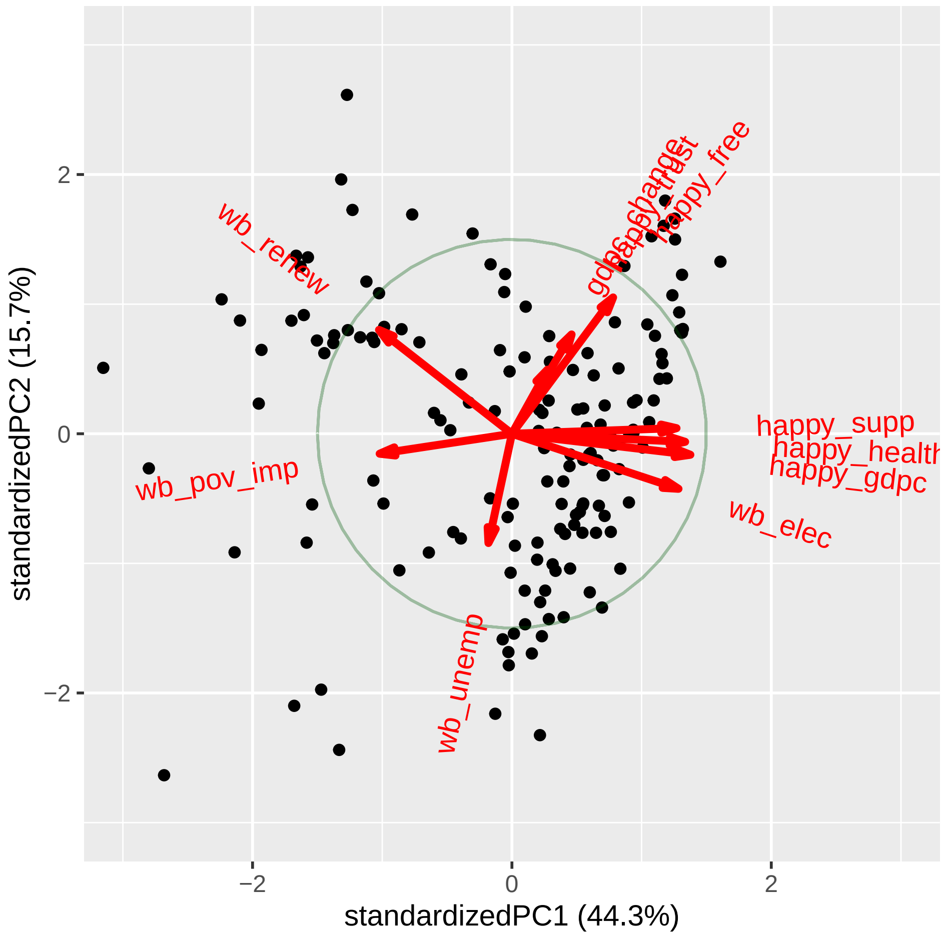


```r
##  Generate a Biplot

# Note that a Biplot displays the scores of the samples on the principal components
# as points and the loadings of the variables as vectors.
#
# The direction and length of the vectors indicate how each variable influences the
# principal components.
#
# Variables that are close together on the plot are positively correlated, variables that are
# orthogonal are uncorrelated, and variables that are far apart (180 degrees from each other)
# are negatively correlated.

ggbiplot(pca_result, circle=TRUE, varname.size=4, varname.color="red", varname.adjust=2) +
          coord_fixed(xlim=c(-3, 3), ylim=c(-3, 3))
```

```
## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.
```



```r
# In case of difficulty reading plot in RStudio, we set parameters for an output image file
# Open 'png' device and recreate ggbiplot
png(filename="PCAbiplot.png", width=2000, height=2000, res=300)
ggbiplot(pca_result, circle=TRUE, varname.size=4, varname.color="red", varname.adjust=2) +
          coord_fixed(xlim=c(-3, 3), ylim=c(-3, 3))
```

```
## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.
```

```r
dev.off() # Close the device
```

PCA Results and Discussion:

Rule of thumb is to keep enough components to explain at least 70% of the total variance. So we retain the first three components, which explain a total of 72.1% of the variance. The upcoming table shows the loadings of the original variables on each component and can be interpreted in terms of correlations.

For example, for PC1:

- The variable 'happy__gdpc' has a loading of 0.44, which means it is quite strongly positively correlated with PC1.
- The variable 'wb__pov__imp' has a loading of -0.32, meaning it is negatively correlated with PC1.
- The variables 'happy__health', 'happy__supp', and 'wb__elec' also have strong positive loadings on PC1, suggesting that these variables move together. When one increases, the other ones tend to increase as well.

These loadings help us interpret the "meaning" of each component. For example, if PC1 is heavily influenced by variables related to happiness and well-being (happy_gdpc, happy_health, happy_supp, wb_elec), we could interpret PC1 as a measure of "general well-being".

Similarly, we would interpret the other components based on the variables with high loadings for those components. For PC2, 'happy_free', 'wb_unemp', and 'wb_renew' have high loadings, so we could say PC2 is a measure of "freedom, development, and employment."

And for PC3, it's basically 'gdpc_change' with some contribution from 'wb_pov_imp', 'happy_trust', and 'wb_unemp'. So we might look at this as a "recent events" component insofar as if something had suddenly "gone bad" for a country, you might see meaningful shifts in GDPC change, trust in government, unemployment, and poverty rate.

```r
# PCA loadings
print(pca_result$rotation)
```

```
##                     PC1         PC2         PC3         PC4         PC5
## happy_gdpc    0.43601492 -0.08631106  0.11584279  0.07152610 -0.03446961
## gdpc_change   0.08684690  0.26340989 -0.75740868  0.14654875  0.46860188
## happy_health  0.42329530 -0.03310911  0.15226476  0.01080276  0.08031692
## happy_supp    0.40220922  0.02294171 -0.02278338 -0.06405957 -0.47203250
## wb_pov_imp   -0.32327653 -0.08196608  0.39207721  0.25302362  0.29279470
## wb_elec       0.40705934 -0.22611930 -0.07980974  0.01363160  0.23937737
## happy_free    0.24721993  0.56005395  0.02527734  0.04297318 -0.21945020
## wb_renew     -0.32483526  0.42674206 -0.14204302  0.03667218 -0.40088274
## happy_trust   0.14581717  0.40800171  0.33726173  0.64884076  0.19534201
## wb_unemp     -0.05739449 -0.44884309 -0.30749655  0.69339347 -0.39988061
##                      PC6        PC7           PC8         PC9         PC10
## happy_gdpc    0.003834684  0.1759008 -0.0009646932 -0.48520635  0.718565077
## gdpc_change  -0.178659199  0.1948489  0.1688089744 -0.09967509 -0.004884293
## happy_health -0.160858876  0.3553862 -0.4835106264 -0.29034217 -0.565435058
## happy_supp   -0.258848541  0.3600125  0.5321324282  0.34503761 -0.106941963
## wb_pov_imp   -0.737683107  0.1353102  0.1052888971  0.03091244  0.103793392
## wb_elec      -0.125965075 -0.1924577 -0.4012985602  0.66683968  0.246359428
## happy_free   -0.367688974 -0.6414226 -0.0552296030 -0.15131035 -0.044884291
## wb_renew     -0.079546896  0.4117075 -0.5056658046  0.17143858  0.263107103
## happy_trust   0.410793681  0.1109847  0.1221699387  0.20434592 -0.040272202
## wb_unemp     -0.069786360 -0.1630706 -0.0935689975 -0.10777825 -0.090328023
```

General Conclusions and Research Questions Answered

- Through multiple methods of assessment, it is clear that the most meaningful predictors of a nation's happiness were GDPC, healthy life expectancy, and sense of social support.
- Variables brought in from the World Bank data which were significant predictors of happiness were all conceptually and logically related to those three primary predictors (unemployment, electricity access, poverty rate, renewable fuels, etc.)
- While the results of robust regression and PCA were not identical, they produced very similar *conceptual* results.
- Comparisons of happiness by region (not discussed) further reinforced the primacy of economy, safety, support, and health as predictors of national happiness.