PSTAT100- Final

LAUREN RABIN (414463-0)

TASHA LEE (9881905)

Bravo Choi (5810288)

Ivy Li (6671747)

Abstract

This analysis delves into the 2023 World Happiness Report, which offers an overview of well-being across various countries over time. The dataset includes variables like subjective well-being (Life Ladder), economic indicators (Log GDP per capita), social support, health (Healthy life expectancy at birth), personal freedom, generosity, perceptions of corruption, and emotional states (Positive affect and Negative affect).

We aim to understanding the interplay between socio-economic and cultural factors and their impact on happiness. We employ a combination of graphical methods and linear regression analyses to explore these relationships and how they have evolved over time. Scatter plots and bar charts visually inspect relationships between variables, while linear regression models quantify these relationships. Additionally, a Principal Component Analysis (PCA) is conducted to compare major happiness components across regions.

The analysis shows a strong positive relationship between Log GDP per capita and Life Ladder scores, with a regression coefficient of 0.76337 and a significant p-value (<2e-16). This indicates that higher economic prosperity is associated with higher happiness levels. Social support also has a significant positive impact on happiness, explaining 60.89% of the variance in Life Ladder scores. Conversely, perceptions of corruption have a significant negative relationship with Life Ladder scores, with a coefficient of -2.63881 and a p-value (<2e-16). Temporal trends reveal that happiness varies across regions, with Oceania maintaining the highest average Life Ladder scores over time and Africa the lowest.

The findings underscore the critical role of economic and social factors in determining happiness levels. Higher GDP per capita and robust social support networks positively influence subjective well-being, while higher perceptions of corruption have a detrimental effect. These insights highlight the importance of addressing economic disparities and governance issues to enhance global well-being. The regional analysis reveals significant differences in happiness trends, suggesting that tailored approaches are necessary to address the unique challenges and opportunities within different regions.

Introduction

In this analysis, we explore the World Happiness Report from 2023 which captures various aspects of well-being across different countries over several years. The dataset includes variables such as subjective well-being (Life.Ladder), economic indicators (Log.GDP.per.capita), social support, health (Healthy.life.expectancy.at.birth), personal freedom, generosity, perceptions of corruption, and emotional states (Positive.affect and Negative.affect).

Our primary research question is: **“How do various socio-economic and cultural factors interact to influence overall happiness across different countries, and how have these relationships evolved over time?”**

To answer this, we focus on several key inquiries using a combination of graphical and regression methods:

How does GDP per capita affect happiness?
- We will use a scatter plot of Life.Ladder versus Log.GDP.per.capita to visually inspect the relationship between economic prosperity and happiness, and apply a linear regression model to quantify this relationship.

What is the relationship between social support and happiness across different countries?
- A bar chart showing average Social.support and Life.Ladder scores by country will highlight the variation across nations. A linear regression model will assess the impact of social connections on happiness.

Is there a correlation between perceptions of corruption and happiness?
- We will create a scatter plot of Life.Ladder versus Perceptions.of.corruption to explore the association between perceived corruption levels and well-being, and use a linear regression model to determine the statistical relationship.

Are there significant differences in happiness trends over time between different regions?
- A line plot showing trends in Life.Ladder scores over time for different regions will illustrate how happiness has evolved. We will use a time series analysis or panel data regression model to analyze these temporal trends and regional differences.

Additionally, we will perform a Principal Component Analysis (PCA) to compare the major happiness components across regions, identifying key factors that contribute to happiness and their regional variations.

The dataset provides a rich source of information, capturing both objective and subjective measures, which allows for a multifaceted analysis of global well-being. By analyzing these variables, we seek to uncover the intricate relationships and key determinants of well-being, contributing to a deeper understanding of what makes life fulfilling and happy across different cultures and nations.

How do economic, social, and health factors collectively influence the overall happiness and life satisfaction in different regions of the world?

We will perform PCA to identify the principal components and visualize the relationships between variables and regions.

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WB
a
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ purrr     1.0.2     ✔ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ dplyr::select() masks MASS::select()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
## corrplot 0.92 loaded
```

EDA

```
# read in dataset
whr <- read.csv('whr-2023.csv')
#whr-2023.csv
head(whr)
```

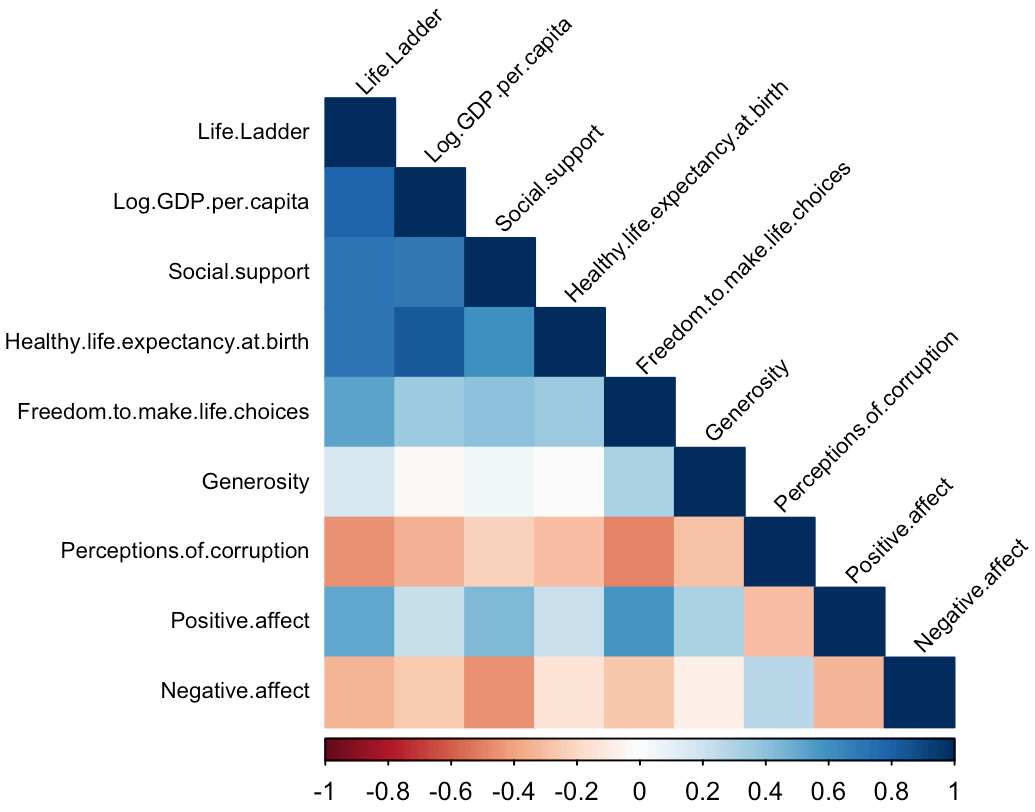
```
##   Country.name year Life.Ladder Log.GDP.per.capita Social.support
## 1  Afghanistan 2008       3.724              7.350          0.451
## 2  Afghanistan 2009       4.402              7.509          0.552
## 3  Afghanistan 2010       4.758              7.614          0.539
## 4  Afghanistan 2011       3.832              7.581          0.521
## 5  Afghanistan 2012       3.783              7.661          0.521
## 6  Afghanistan 2013       3.572              7.680          0.484
##   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices Generosity
## 1                             50.5                         0.718      0.168
## 2                             50.8                         0.679      0.191
## 3                             51.1                         0.600      0.121
## 4                             51.4                         0.496      0.164
## 5                             51.7                         0.531      0.238
## 6                             52.0                         0.578      0.063
##   Perceptions.of.corruption Positive.affect Negative.affect
## 1                     0.882           0.414           0.258
## 2                     0.850           0.481           0.237
## 3                     0.707           0.517           0.275
## 4                     0.731           0.480           0.267
## 5                     0.776           0.614           0.268
## 6                     0.823           0.547           0.273
```

For our exploratory data analysis we examine several key variables:

- **`Country.name`** : Name of the country.
- **`year`** : Year of data collection.
- **`Life.Ladder`** : Subjective well-being scale (0-10).
- **`Log.GDP.per.capita`** : Logarithm of GDP per capita.
- **`Social.support`** : Availability of social connections.
- **`Healthy.life.expectancy.at.birth`** : Expected healthy lifespan at birth.
- **`Freedom.to.make.life.choices`** : Personal autonomy in life choices.
- **`Generosity`** : Financial altruism and charitable behavior.
- **`Perceptions.of.corruption`** : Perceived level of public corruption.
- **`Positive.affect`** : Prevalence of positive emotions.
- **`Negative.affect`** : Prevalence of negative emotions.

It would make sense for a lot of these variables to be correlated. Lets investigate if this is true so we can get a better idea of these relationships and overall identify which have the most significant impact on overall happiness across different countries and time. We will calculate a correlation matrix and plot the correlation heatmap to visualize these potential relationships.

```
whr_cleaned <- na.omit(whr)

# selecting numeric variables
numeric_vars <- whr_cleaned[, c("Life.Ladder", "Log.GDP.per.capita", "Social.support",
                                "Healthy.life.expectancy.at.birth",
                                "Freedom.to.make.life.choices",
                                "Generosity", "Perceptions.of.corruption", "Positive.affec
t",
                                "Negative.affect")]

# calc corr matrix
correlation_matrix <- cor(numeric_vars)

# plot heatmap
corrplot(correlation_matrix, method = "color", type = "lower",
         tl.col = "black", tl.srt = 45, tl.cex = 0.7)
```



We will walk you through the main takeaways from this heatmap, finding which variables have positive/negative relationships with Life Ladder:

Strong positive relationship:
- **Log GDP per capita AND Life ladder**: This make sense as typically higher GDP per capita tends to be associated with higher levels of happiness due to increased access to resources and overall higher standards of living.

- **Social support AND Life ladder**: Strong social connections and support networks tend to contribute to overall happiness and well-being.
- **Health AND Life ladder**: Better health outcomes and longer life expectancy are typically associated with higher levels of happiness, as good health contributes to overall quality of life.

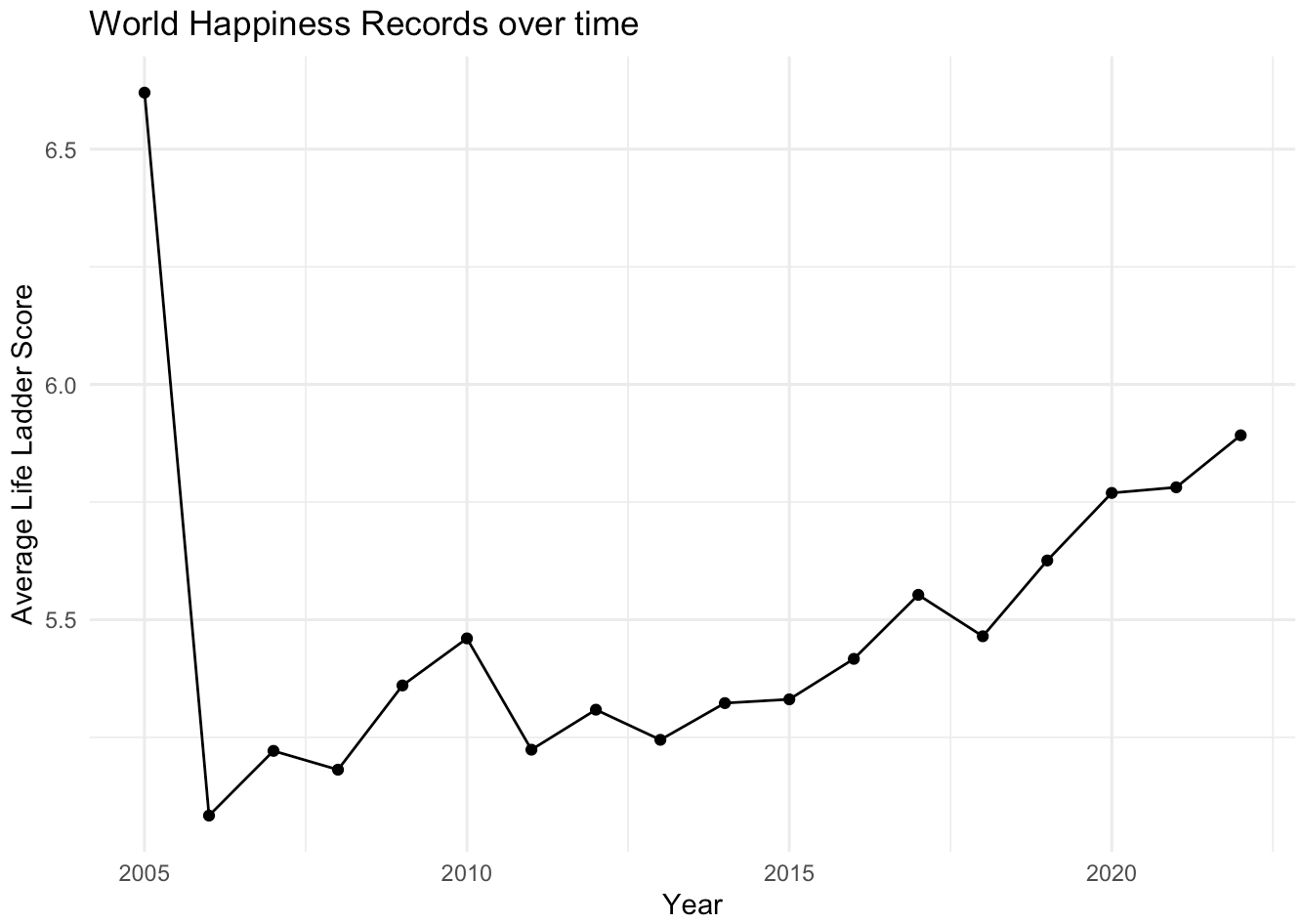
Moderate negative relationship:
- **Perceptions of corruption AND life ladder**: Higher levels of perceived corruption can potentially lead to decreased trust in institutions/government which could influence overall happiness levels.
Correlation within variables: Positive and negative affect obviously have a corresponding positive and negative affect with positive/negative aspects. Generosity seems to have no significant correlation with other variables.
Overall, perceptions of corruption have a moderate negative relationship with the other variables, where freedom to make life choices had a moderate positive one.

Investigate!

Now that we have an understanding of what our dataset looks like and what our goals our, let's continue with a series of explorations and analyses.

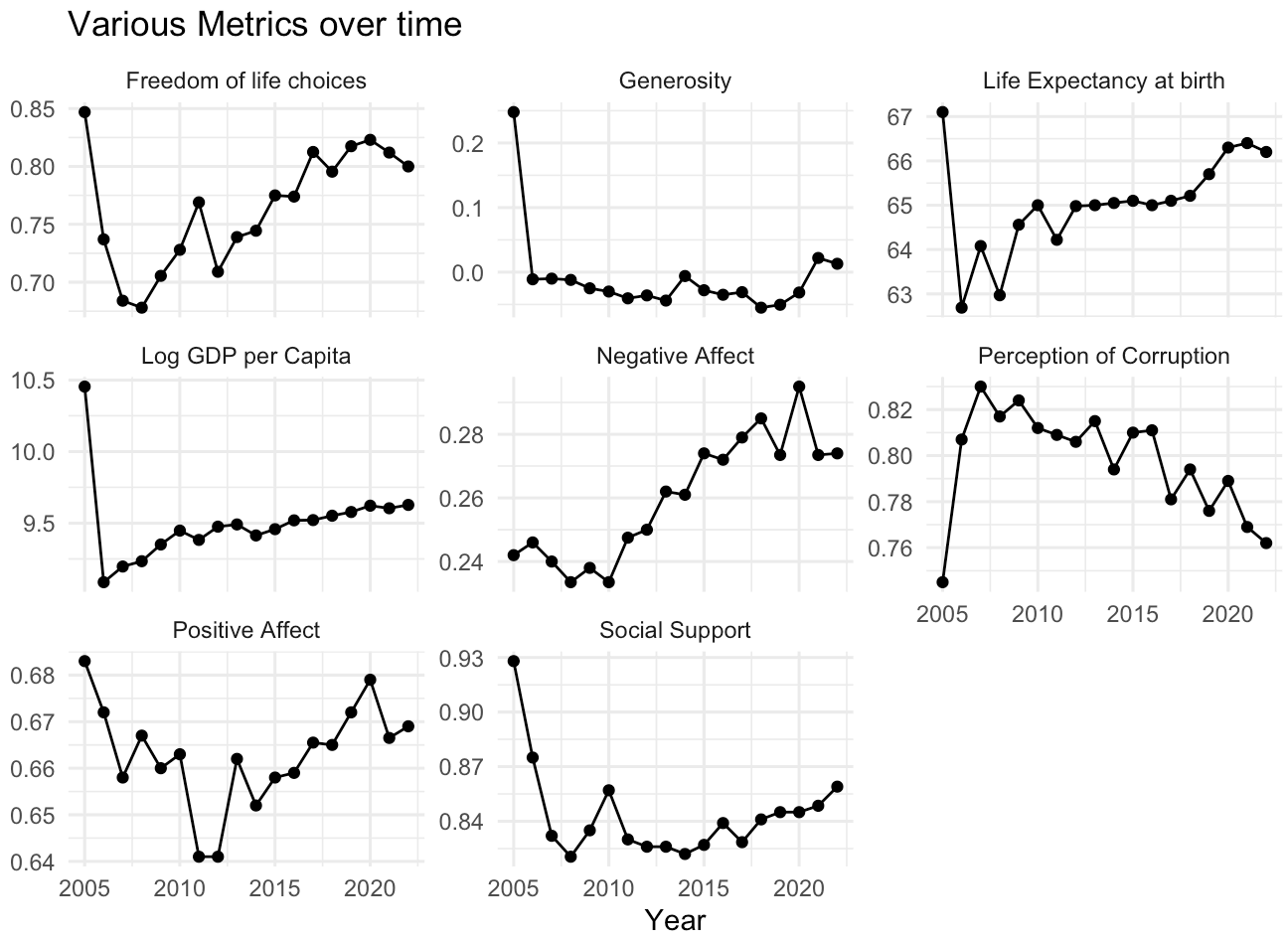
First, we can investigate World Happiness Records over time.

```
whr %>%
  group_by(year) %>%
  summarise(avg_lifeLadder = median(Life.Ladder)) %>%
  ggplot(aes(
    x = year,
    y = avg_lifeLadder
  )) +
  geom_point() +
  geom_line() +
  labs(
    x = 'Year',
    y = 'Average Life Ladder Score',
    title = 'World Happiness Records over time'
  ) +
  theme_minimal()
```



The above distribution of Life Ladder scores shows that there has been an overall increase in happiness since 2006. However, in 2005, the average (median) life ladder score was 6.62, the highest it has been for all years since then. We will investigate this sudden drop in life ladder scores.

Next, we can look at the distributions for each metric over time to see if any of them also saw a drastic change between 2005 and 2006.



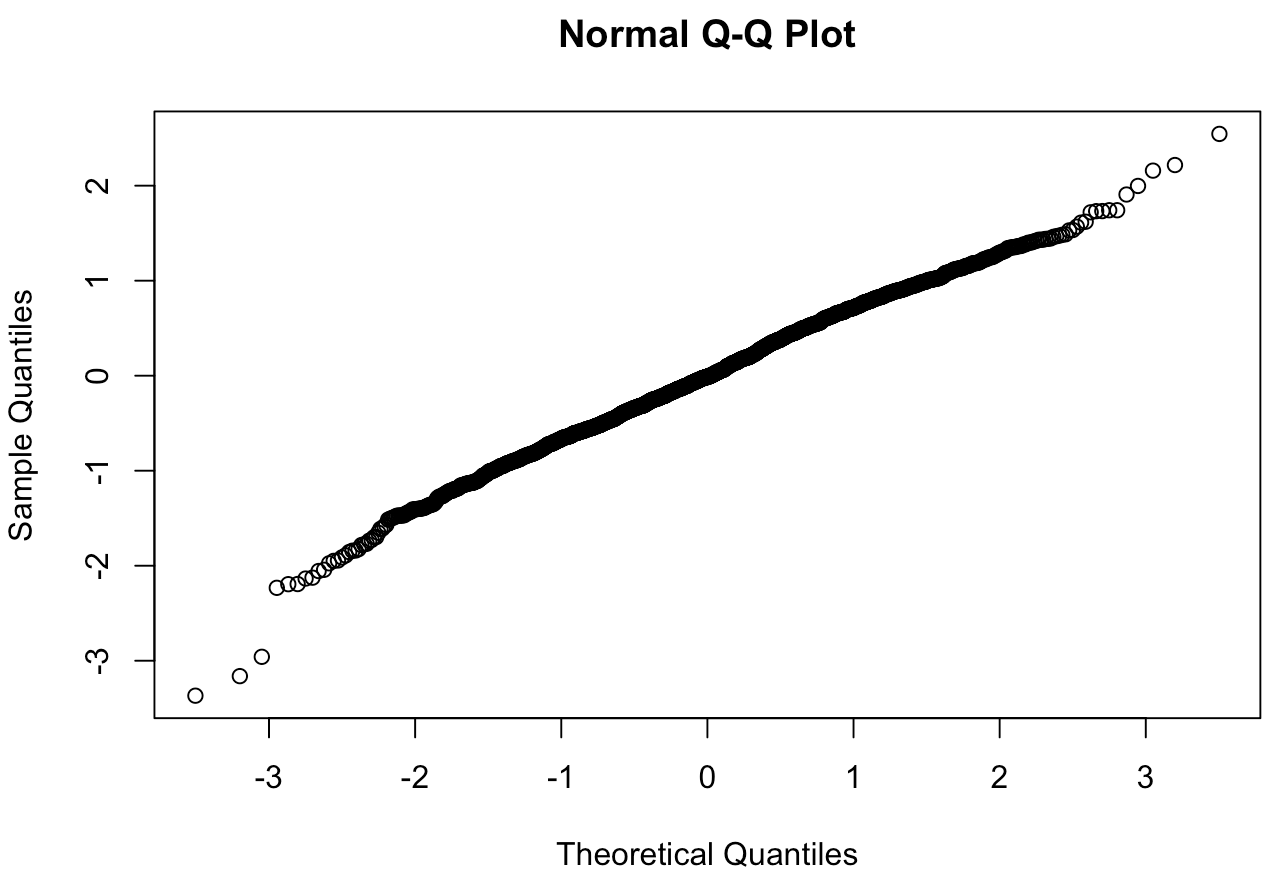
The above distributions show that almost all of the given metrics faced a drastic change between 2005-2006. The metrics that we would expect to have a positive relationship with happiness show a sharp decline in 2006, while those we would expect to have a negative relationship (such as perception of corruption) show a sharp increase.

Next, we can test the correlation of these supposed relationships by regressing them onto Life Ladder scores. We will choose a few key metrics to examine closely, starting with log GDP per capita.

How does GDP per capita affect happiness?

First, we will investigate the relationship between log GDP per capita and life ladder score. Before performing our regression, we need to check that the conditions to perform linear regression are satisfied. First, we can check normality using a QQ-plot of the residuals.

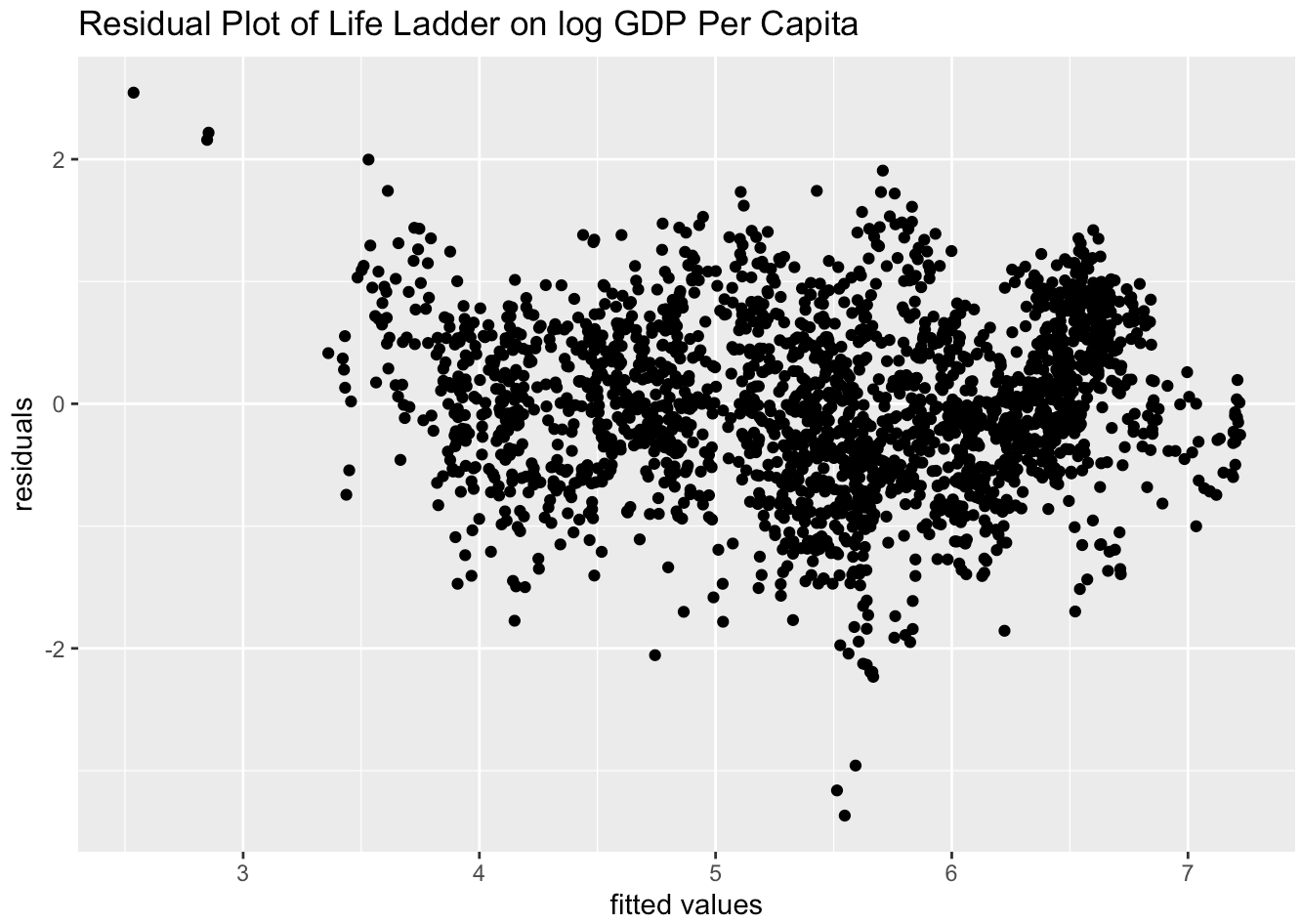
```
lm <- lm(whr$Life.Ladder ~ whr$Log.GDP.per.capita)

qqnorm(lm$residuals)
```



The QQ-plot above shows an overall linear trend with some deviation in the lower tail. As simple linear regression assumes the 'noise' follows a normal distribution, we will proceed with caution when performing our linear regression.

Next, we need to check that our data is homoskedastic using a residual plot.

```
ggplot(lm,
       aes(
         x = lm$fitted.values,
         y = lm$residuals)) +
  geom_point() +
  labs(
    x = 'fitted values',
    y = 'residuals',
    title = 'Residual Plot of Life Ladder on log GDP Per Capita'
  )
```



As there is no prominent relationship within our residual plot and an approximately constant range, we will proceed with our linear regression.

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \varepsilon_i$$

Let

$$\widehat{y}_i$$

be life ladder score and

$$x_i$$

be log GDP per capita.

```
##
## Call:
## lm(formula = whr$Life.Ladder ~ whr$Log.GDP.per.capita)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3679 -0.4745 -0.0117  0.5071  2.5449
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.68302    0.12220  -13.77   <2e-16 ***
## whr$Log.GDP.per.capita  0.76337    0.01292   59.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6953 on 2177 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.616,  Adjusted R-squared:  0.6158
## F-statistic:  3493 on 1 and 2177 DF,  p-value: < 2.2e-16
```

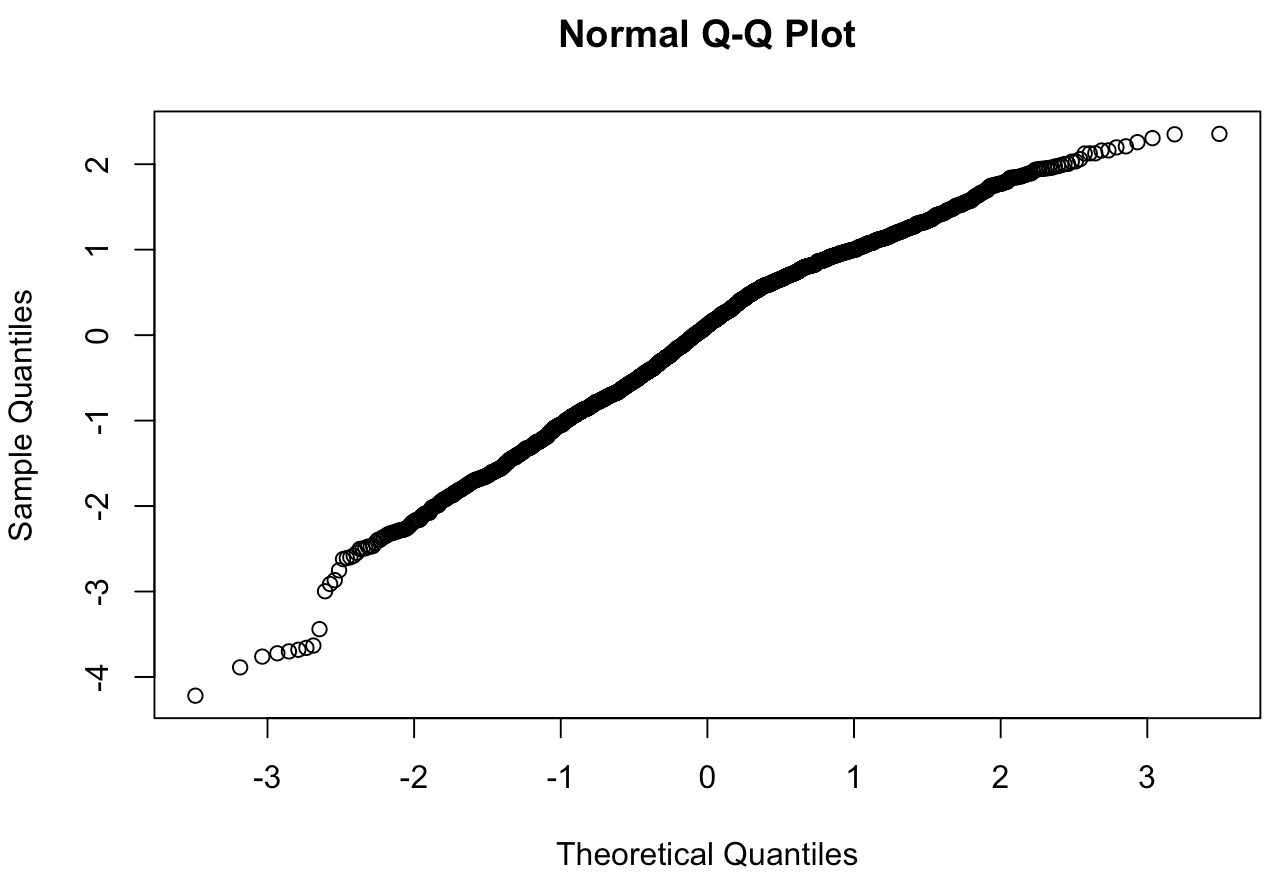
The intercept of this linear model is -1.68302 suggesting that when the log GDP per capita of a country is zero, the predicted life ladder score is -1.68302. We should note that the lowest observed log GDP per capita in our dataset is 5.527, so our predicted intercept value is susceptible to the dangers of extrapolation. Thus, as life ladder scores only exist on the scale of 0-10, we can see that our predicted intercept of -1.683027 is not reasonable.

The coefficent for log GDP per capita is 0.76337, meaning that for each additional unit increase in log GDP per capita, the predicted life ladder score will increase by 0.76337, holding all else constant. The corresponding p-value of <2e-16 is lower than 0.05, indicating that at a 95% significance level there is strong evidence to reject the null hypothesis that there is no linear relationship between log GDP per capita and life ladder score.

Next, we can follow the same procedure to investigate the relationship between perceptions of corruption and life ladder score.

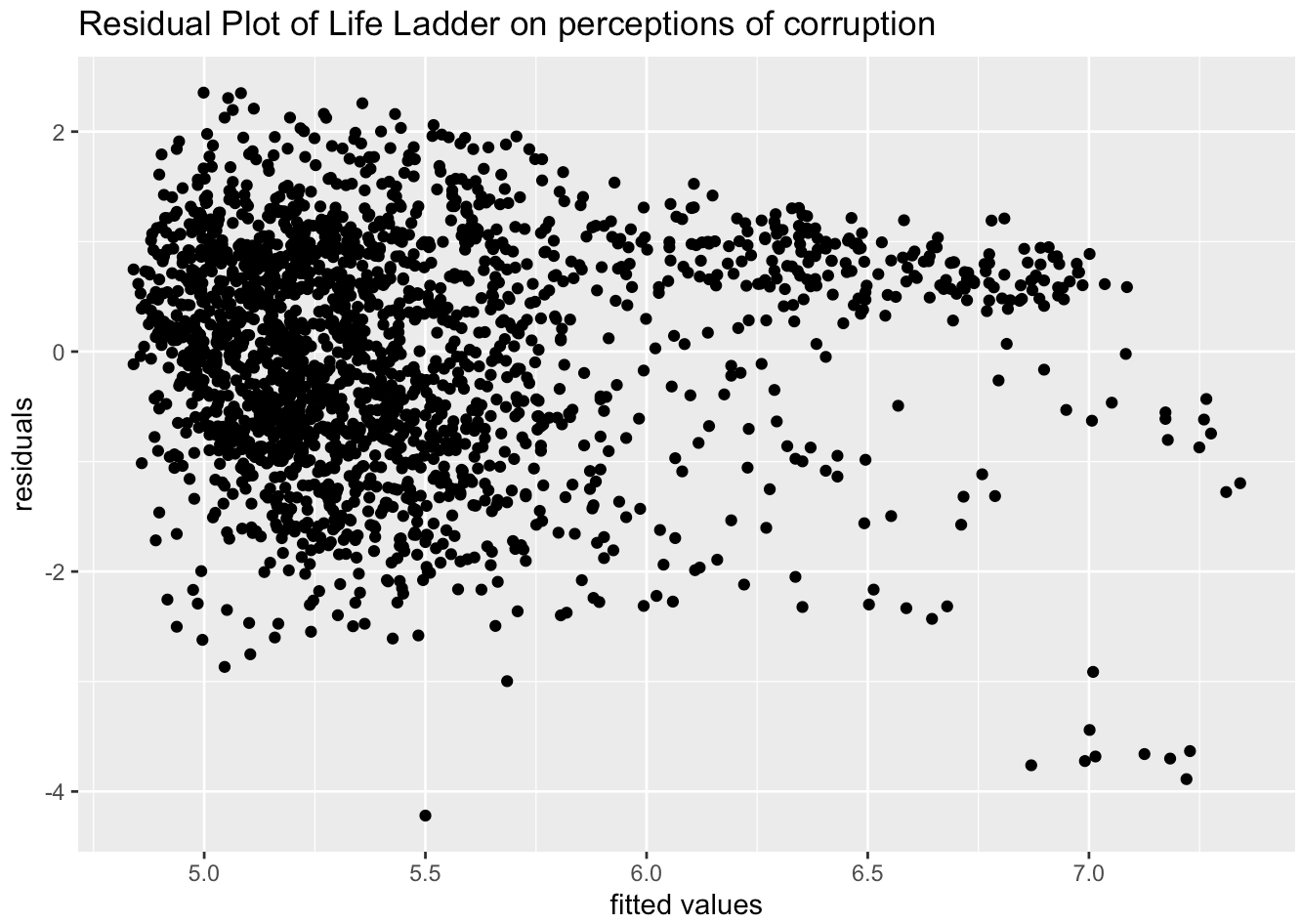
Similarly, we will check normality using a QQ-plot of the residuals and homoskedasticity using a residual plot before proceeding.

```
lm <- lm(whr$Life.Ladder ~ whr$Perceptions.of.corruption)

qqnorm(lm$residuals)
```



The QQ-plot above shows an overall linear trend with some deviation in the lower tail. As simple linear regression assumes the 'noise' follows a normal distribution, we will proceed with caution when performing our linear regression.

```
ggplot(lm,
       aes(
         x = lm$fitted.values,
         y = lm$residuals)) +
  geom_point() +
  labs(
    x = 'fitted values',
    y = 'residuals',
    title = 'Residual Plot of Life Ladder on perceptions of corruption'
  )
```



As there is no prominent relationship within our residual plot and an approximately constant range, we will proceed with our linear regression.

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \varepsilon_i$$

Let

$$\widehat{y}_i$$

be life ladder score and

$$x_i$$

be perceptions of corruption.

```
##
## Call:
## lm(formula = whr$Life.Ladder ~ whr$Perceptions.of.corruption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2190 -0.7156  0.1136  0.8007  2.3544
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    7.43424    0.09288   80.04   <2e-16 ***
## whr$Perceptions.of.corruption -2.63881    0.12094  -21.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.025 on 2081 degrees of freedom
##   (116 observations deleted due to missingness)
## Multiple R-squared:  0.1862, Adjusted R-squared:  0.1858
## F-statistic: 476.1 on 1 and 2081 DF,  p-value: < 2.2e-16
```

The intercept of this linear model is 7.43424 suggesting that when the perceptions of corruption level is zero, the predicted life ladder score is 7.43424. The lowest observed perception of corruption in our dataset is 0.035, so our predicted intercept value could be susceptible to extrapolation.

The coefficent for perception of corruption is -2.63881, meaning that for each additional unit increase in perception of corruption, the predicted life ladder score will decrease by -2.63881, holding all else constant. The corresponding p-value of <2e-16 is lower than 0.05, indicating that at a 95% significance level there is strong evidence to reject the null hypothesis that there is no linear relationship between perceptions of corruption and life ladder score.

What is the relationship between social support and happiness across different countries?

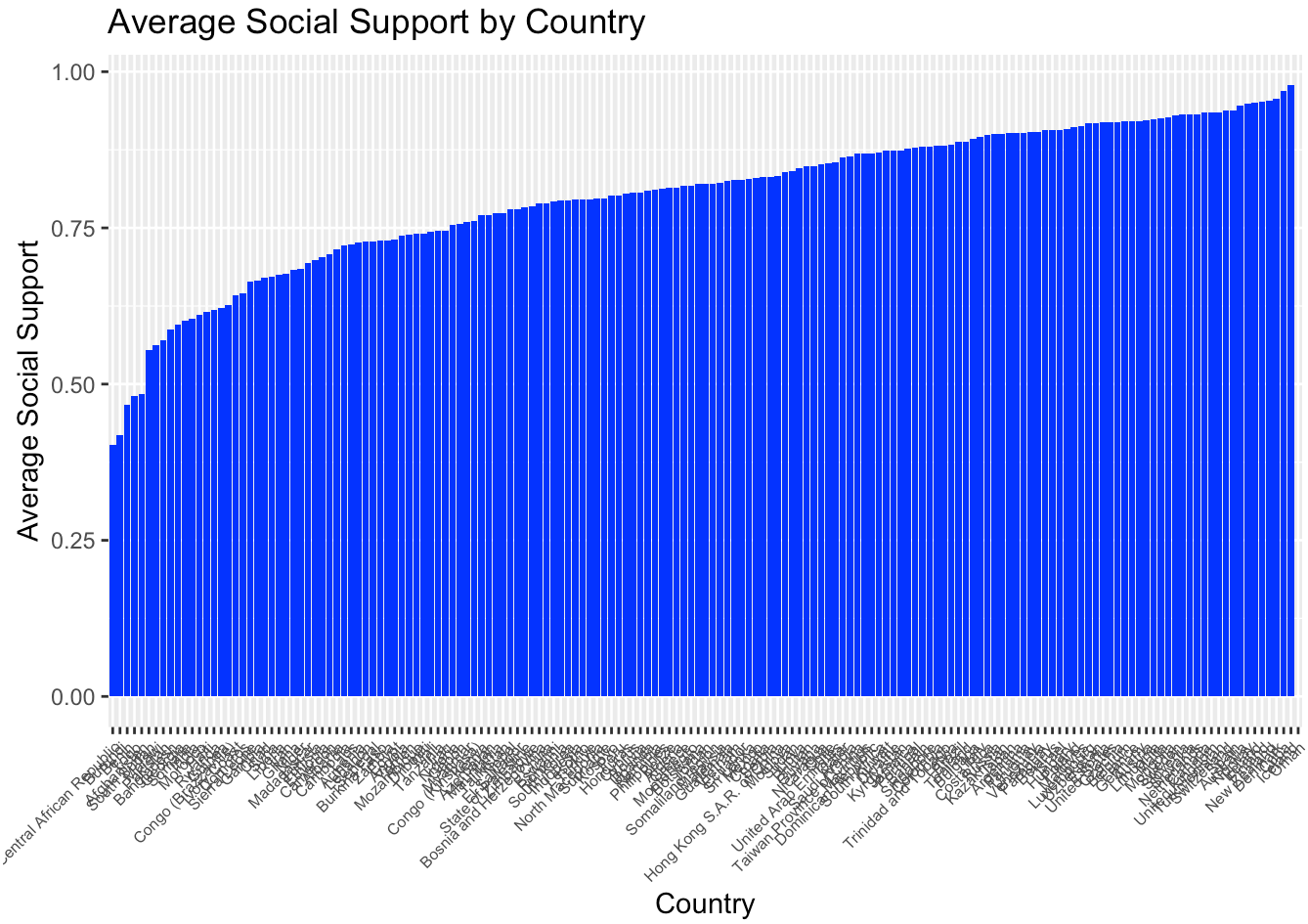
First, we calculate average social support and life ladder.

```
average_scores <- whr %>%
  group_by(Country.name) %>%
  summarize(
    avg_social_support = mean(Social.support, na.rm = TRUE),
    avg_life_ladder = mean(Life.Ladder, na.rm = TRUE)
  )
```

Now plot the bar chart for average Social support by country.

```
ggplot(average_scores, aes(x = reorder(Country.name, avg_social_support), y = avg_social
_support)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Average Social Support by Country",
       x = "Country",
       y = "Average Social Support") +
  theme(axis.text.x = element_text(size = 6, angle = 45, hjust = 1))
```

```
## Warning: Removed 1 rows containing missing values (`position_stack()`).
```



There are so many countries which makes it difficult to investigate, however we can look at overall trend with the bar plot. It seems about 1/3 of the countries have an average social support score of 0.75 or below. The majority of the data has above a 0.75 score. Let's take a closer look and see what countries are the bottom 3, mid 3, and top 3 based on their average social support scores.

```
# sort scores
sorted_scores <- average_scores %>%
  filter(!is.na(avg_social_support)) %>%
  arrange(avg_social_support)

# bottom 3 and top 3 countries
bottom_countries <- head(sorted_scores, 3)
top_countries <- tail(sorted_scores, 3)

# mid 3 countries
if (nrow(sorted_scores) %% 2 == 0) {
  mid_index <- nrow(sorted_scores) / 2
  mid_countries <- sorted_scores[(mid_index - 1):(mid_index + 1), ]
} else {
  mid_index <- (nrow(sorted_scores) + 1) / 2
  mid_countries <- sorted_scores[(mid_index - 1):(mid_index + 1), ]
}
```

The 3 countries that have the lowest average social support score and their average life ladder:

```
bottom_countries
```

```
## # A tibble: 3 × 3
##   Country.name             avg_social_support avg_life_ladder
##   <chr>                                 <dbl>           <dbl>
## 1 Central African Republic              0.402            3.52
## 2 Burundi                               0.418            3.55
## 3 Benin                                 0.466            4.09
```

The middle three countries:

```
mid_countries
```

```
## # A tibble: 3 × 3
##   Country.name avg_social_support avg_life_ladder
##   <chr>                     <dbl>           <dbl>
## 1 Montenegro                0.818            5.31
## 2 Botswana                  0.820            3.95
## 3 Jordan                    0.821            5.09
```

The top three countries:

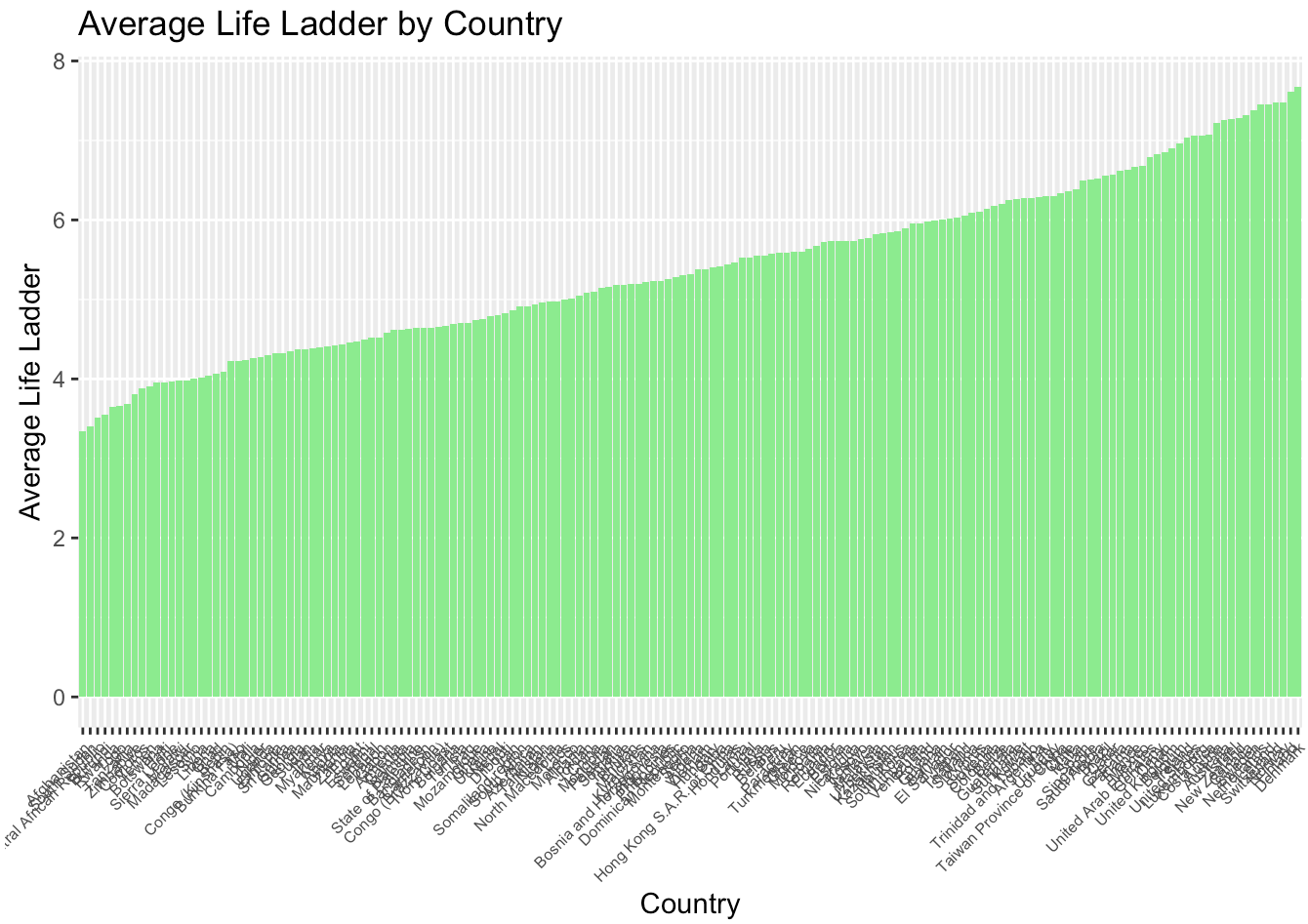
```
top_countries
```

```
## # A tibble: 3 × 3
##   Country.name avg_social_support avg_life_ladder
##   <chr>                     <dbl>           <dbl>
## 1 Denmark                   0.957            7.67
## 2 Cuba                      0.97             5.42
## 3 Iceland                   0.978            7.46
```

Interesting! It is also interesting to note that overall as the average social support increases, so does average life ladder. We will investigate this further later!

Bar chart for average Life Ladder by country

```
ggplot(average_scores, aes(x = reorder(Country.name, avg_life_ladder), y = avg_life_ladd
er)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Average Life Ladder by Country",
       x = "Country",
       y = "Average Life Ladder") +
  theme(axis.text.x = element_text(size = 6, angle = 45, hjust = 1))
```



Same thing for this bar plot, we'll need to take a closer look.

```
# Sort scores
sorted_scores_2 <- average_scores %>%
  filter(!is.na(avg_life_ladder)) %>%
  arrange(avg_life_ladder)

# Bottom 3 and top 3 countries
bottom_countries_2 <- head(sorted_scores_2, 3)
top_countries_2 <- tail(sorted_scores_2, 3)

# Mid 3 countries
if (nrow(sorted_scores_2) %% 2 == 0) {
  mid_index <- nrow(sorted_scores_2) / 2
  mid_countries_2 <- sorted_scores_2[(mid_index - 1):(mid_index + 1), ]
} else {
  mid_index <- (nrow(sorted_scores_2) + 1) / 2
  mid_countries_2 <- sorted_scores_2[(mid_index - 1):(mid_index + 1), ]
}
```

The 3 countries that have the lowest average life ladder score and their average social support:

```
bottom_countries_2
```

```
## # A tibble: 3 × 3
##   Country.name            avg_social_support avg_life_ladder
##   <chr>                                <dbl>           <dbl>
## 1 Afghanistan                          0.485            3.35
## 2 South Sudan                          0.555            3.40
## 3 Central African Republic             0.402            3.52
```

The middle three countries:

```
mid_countries_2
```

```
## # A tibble: 3 × 3
##   Country.name avg_social_support avg_life_ladder
##   <chr>                     <dbl>           <dbl>
## 1 Montenegro                0.818            5.31
## 2 Serbia                    0.831            5.32
## 3 Algeria                   0.815            5.38
```

The top three countries:

```
top_countries_2
```

```
## # A tibble: 3 × 3
##   Country.name avg_social_support avg_life_ladder
##   <chr>                     <dbl>           <dbl>
## 1 Norway                    0.948            7.48
## 2 Finland                   0.952            7.62
## 3 Denmark                   0.957            7.67
```

The same pattern follows when investigating life ladder, which leads us to believe there is a relationship between these variables. Denmark is in the top 3 of average social support and life ladder! Central African Republic is in the bottom 3 of average social support and life ladder.

Lets investigate if there is a relationship between these variables using a linear regression model…

Now let's build a regression model which will predict the average Life Ladder score using the average Social Support score.

```
regression_model <- lm(avg_life_ladder ~ avg_social_support, data = average_scores)
summary(regression_model)
```

```
##
## Call:
## lm(formula = avg_life_ladder ~ avg_social_support, data = average_scores)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56733 -0.44222 -0.07881  0.49263  1.38328
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.2196     0.3573  -0.615     0.54
## avg_social_support   6.9974     0.4406  15.880   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6616 on 162 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6089, Adjusted R-squared:  0.6064
## F-statistic: 252.2 on 1 and 162 DF,  p-value: < 2.2e-16
```

Overall, there is a strong, statistically significant positive relationship between Social Support and Life Ladder scores across different countries. The model explains a substantial portion, 60.89%, of the variance in happiness (Life Ladder scores) based on social support. The coefficient for Social Support of 6.9974 suggests that increases in social support are strongly associated with increases in perceived happiness.

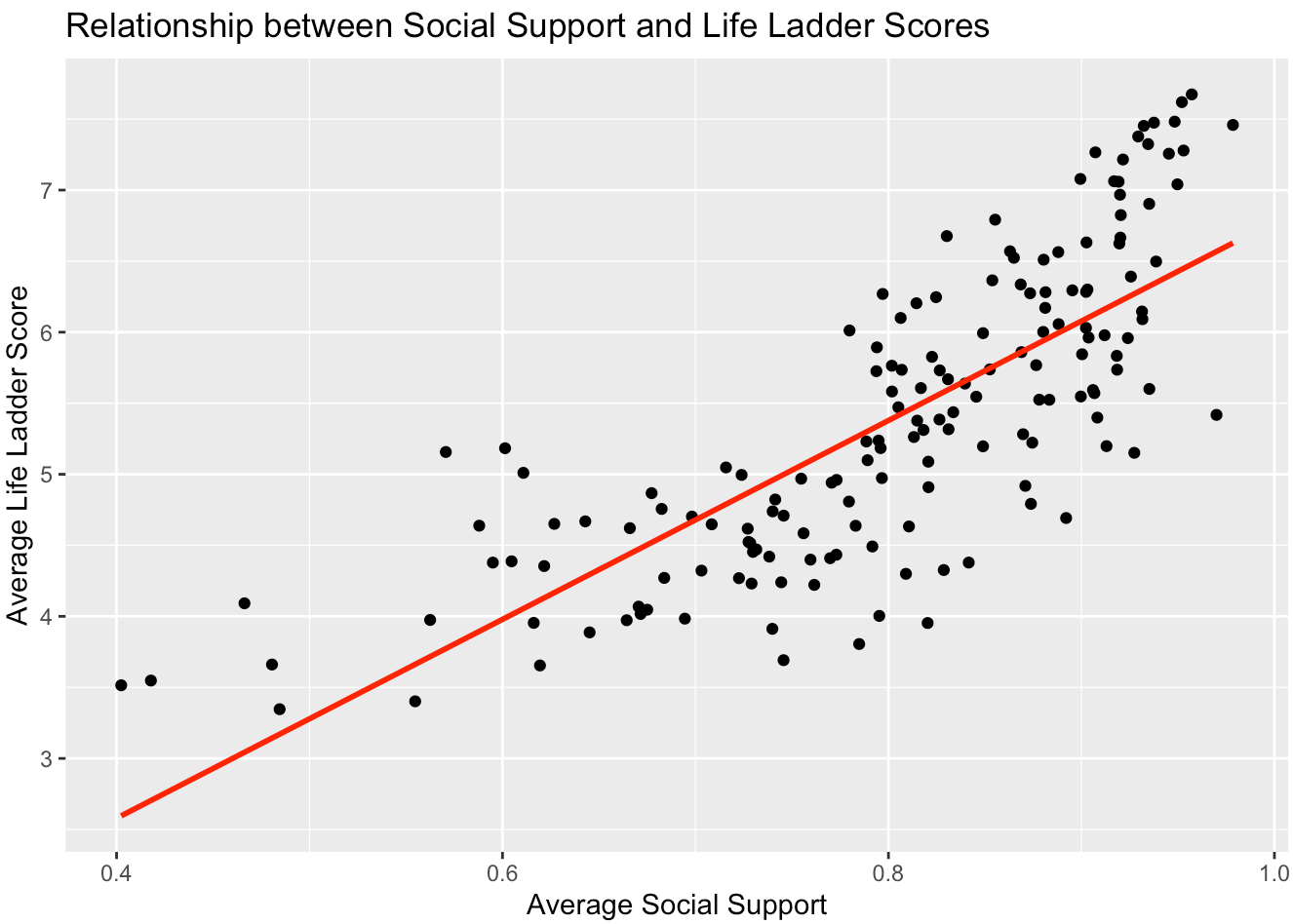
Now let's create a scatter plot with regression line.

```
ggplot(average_scores, aes(x = avg_social_support, y = avg_life_ladder)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relationship between Social Support and Life Ladder Scores",
       x = "Average Social Support",
       y = "Average Life Ladder Score")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Here, each point represents a country, and its position shows the relationship between Social Support and Life Ladder scores for that country. Overall, the model seems to be a good fit for the data, as it stays relatively close to the regression line. The regression line slopes upwards in the positive direction along with the data points, indicating a positive relationship between Social Support and Life Ladder.

Are there significant differences in happiness trends over time between different regions?

We will use the package country code to categorize the countries into regions!

```
#install.packages("countrycode")
library(countrycode)

whr_with_continent <- mutate(whr, Continent = countrycode(Country.name, "country.name",
"continent"))
```

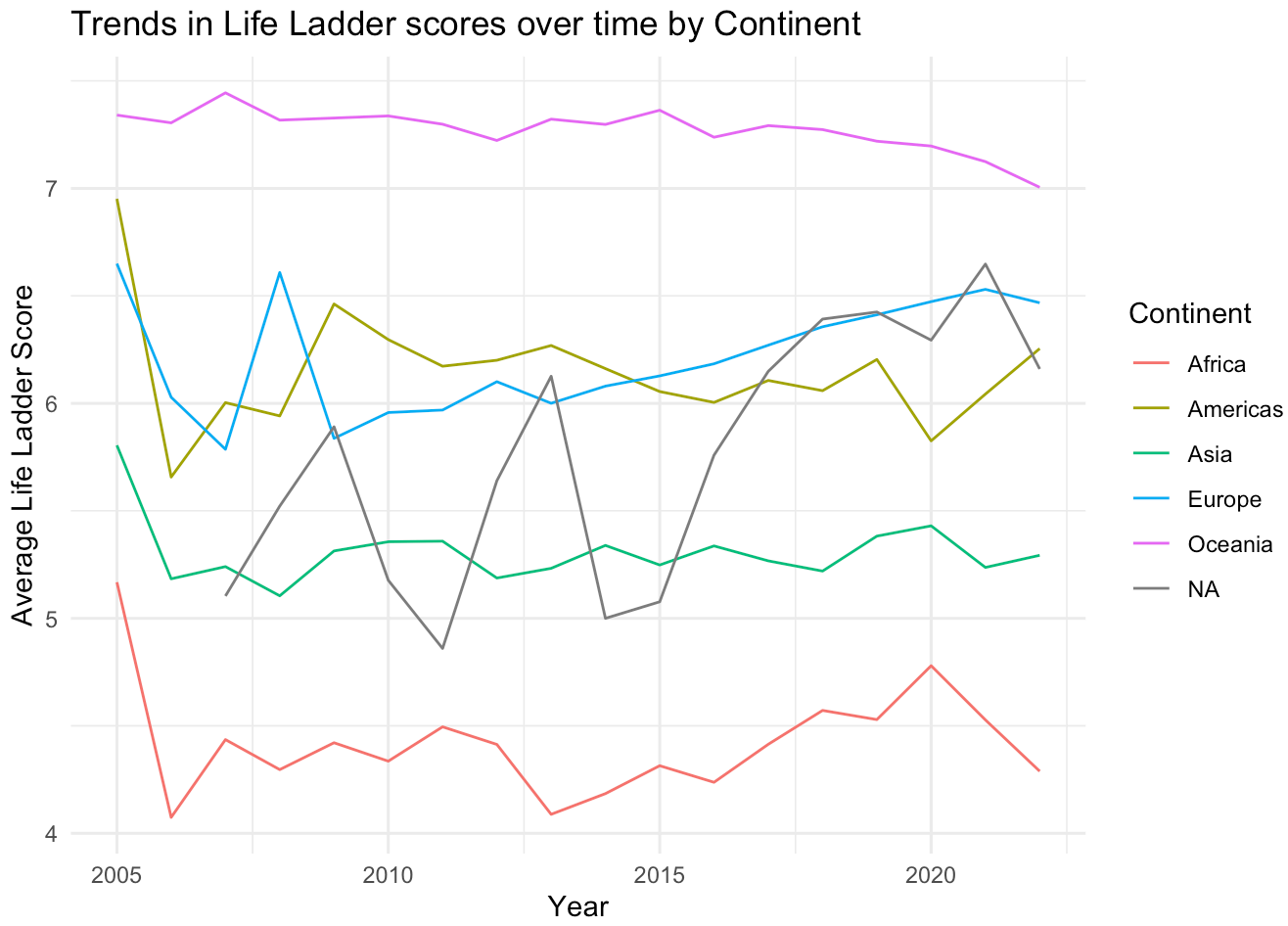
```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `Continent = countrycode(Country.name, "country.name",
##   "continent")`.
## Caused by warning:
## ! Some values were not matched unambiguously: Kosovo
```

Now we will calculate continental average Life Ladder scores over time and plot them.

```
continental_avg_scores <- whr_with_continent %>%
  group_by(Continent, year) %>%
  summarize(avg_life_ladder = mean(Life.Ladder, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'Continent'. You can override using the
## `.groups` argument.
```

```
# plot
ggplot(continental_avg_scores, aes(x = year, y = avg_life_ladder, color = Continent)) +
  geom_line() +
  labs(title = "Trends in Life Ladder scores over time by Continent",
       x = "Year",
       y = "Average Life Ladder Score") +
  theme_minimal()
```



From this plot we can see that happiness varies a lot across different regions. Overall, Oceania has the highest average Life Ladder over time and Africa has the lowest. Oceania and Asia's Life Ladder remains relatively constant over time. Oceania maintains around a 7 and Asia 5.5. The Americas has a lot of variation and at 2005 seemed to have its all time high of 7, however, over time has decreased. The same goes for Asia, at 2005 with almost a 6 but dropped significantly after and has maintained around an average score of 5.25. For Europe, around the early 2000's, Life Ladder score saw a lot of stark increases and decreases but has over time seemed to level out and slowly increase to 6.5.

Princpal Component Analysis

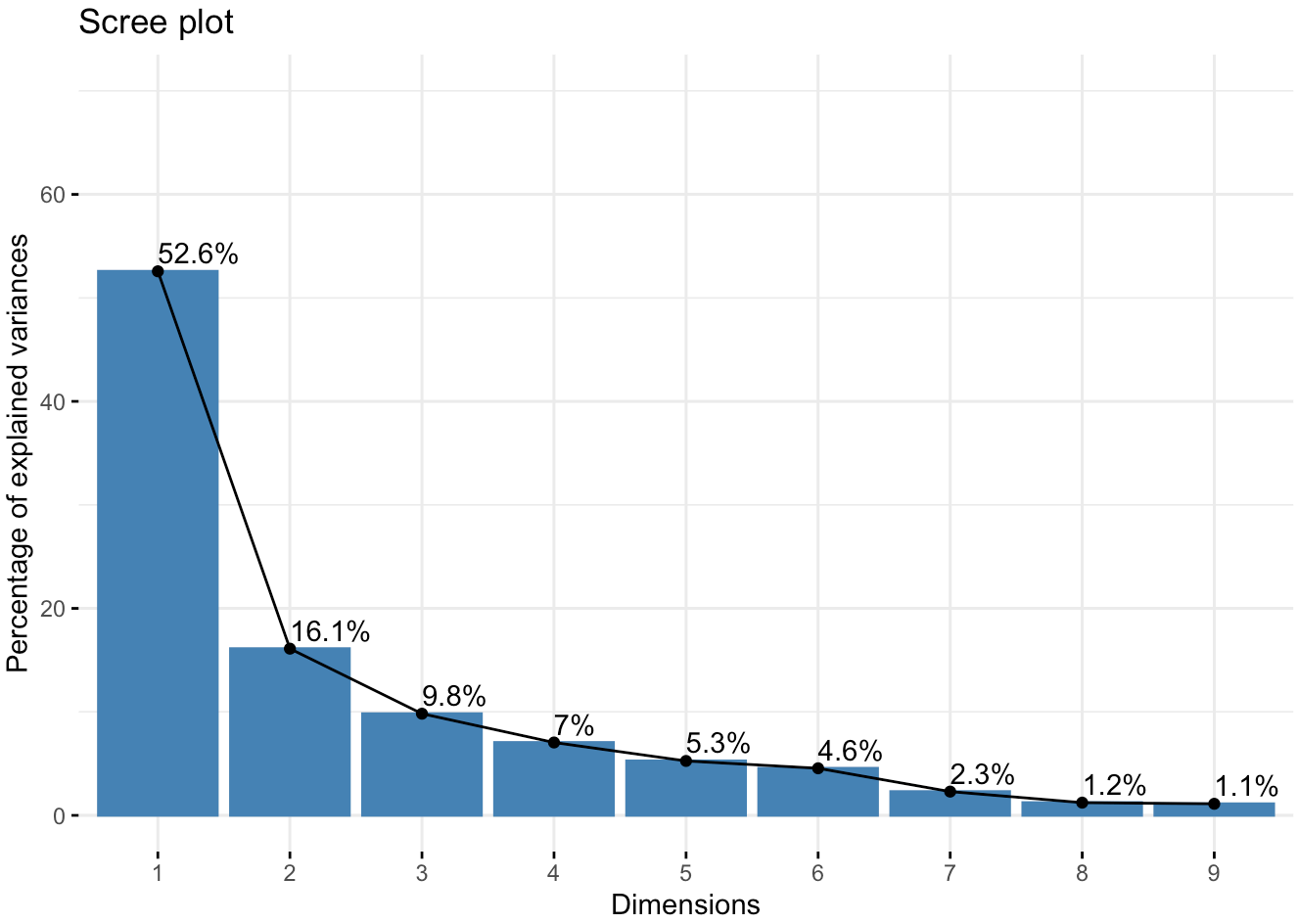
The objective of this study is to analyze the factors influencing overall happiness and life satisfaction across different regions of the world. By utilizing the World Happiness Report dataset, we aim to identify key determinants such as economic, social, and health factors that contribute to happiness. To achieve this, we employ Principal Component Analysis (PCA) to reduce the dimensionality of the data and visualize the relationships between these variables. PCA helps in understanding how these variables interact and influence happiness across various regions, providing valuable insights for policymakers and researchers.

Principal Component Analysis was performed on the standardized dataset to identify the main components that explain the variance in the data. The PCA results were visualized using a Scree plot and a biplot. The Scree plot helped determine the number of principal components to retain, while the biplot provided insights into the relationships between the variables and the clustering of countries by region.

```
# Summary of Analysis
summary(whr_pca)
```

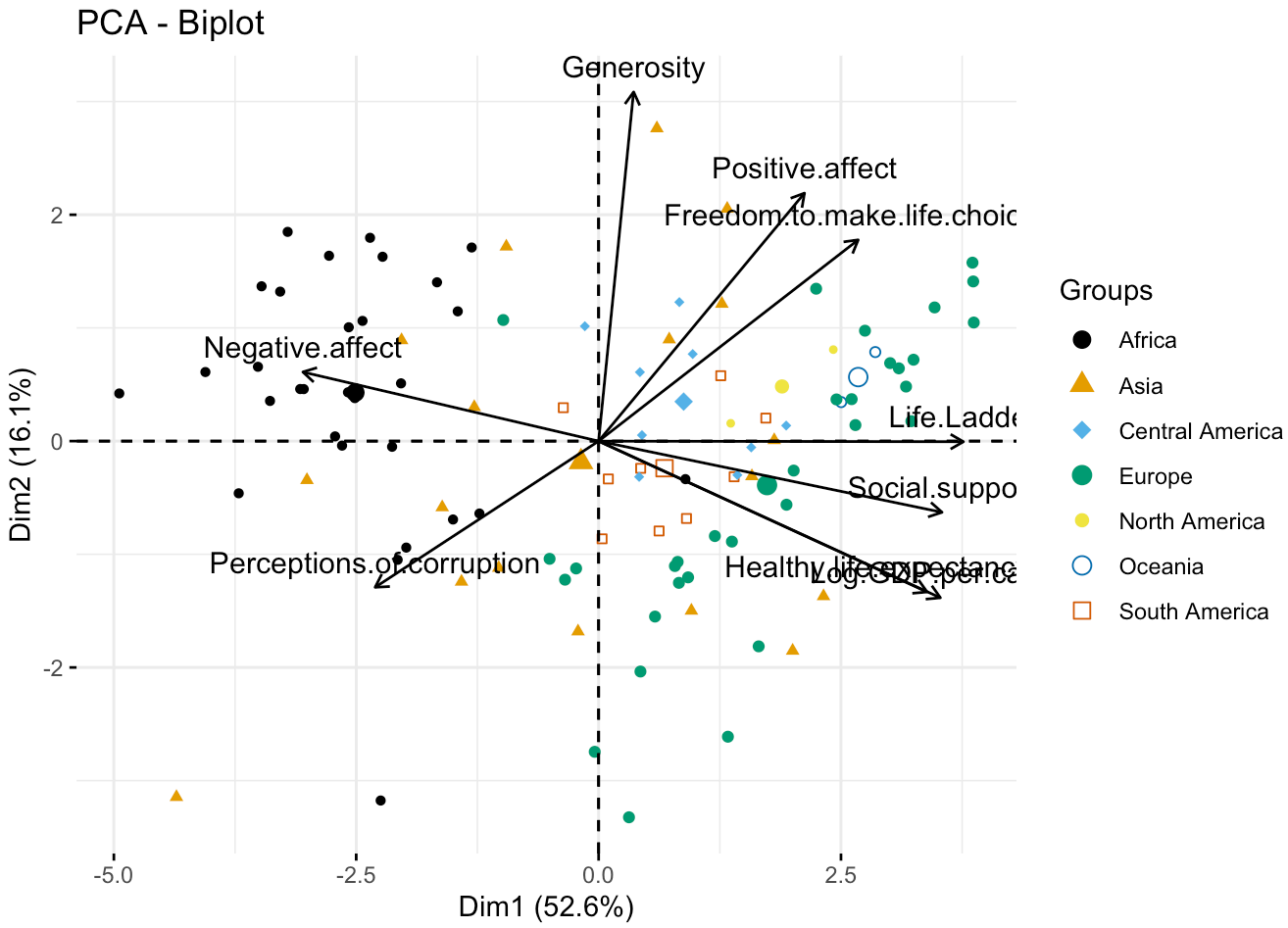
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5    PC6     PC7
## Standard deviation     2.1751 1.2041 0.94028 0.79610 0.68800 0.6399 0.45485
## Proportion of Variance 0.5257 0.1611 0.09824 0.07042 0.05259 0.0455 0.02299
## Cumulative Proportion  0.5257 0.6868 0.78499 0.85541 0.90801 0.9535 0.97649
##                            PC8     PC9
## Standard deviation     0.33289 0.31738
## Proportion of Variance 0.01231 0.01119
## Cumulative Proportion  0.98881 1.00000
```

```
# Scree Plot of Variance
fviz_eig(whr_pca,
         addlabels = T,
         ylim = c(0,70))
```



The Scree plot and table of Importance of components illustrate the percentage of variance explained by each principal component. In our analysis, the first principal component (PC1) explained 52.6% of the variance, and the second principal component (PC2) explained 16.1% of the variance. Together, these two components accounted for 68.7% of the total variance in the dataset. The Scree plot showed a sharp drop in variance explained after the first component, indicating that the first two components capture the most significant variation in the data. This suggests that focusing on PC1 and PC2 will provide a comprehensive understanding of the key factors influencing happiness.

```
## Scale for shape is already present.
## Adding another scale for shape, which will replace the existing scale.
```



The biplot visualizes the first two principal components, with arrows representing the original variables and points representing the countries, color-coded by region. The direction and length of the arrows indicate the influence of each variable on the principal components. In our biplot, the arrows for freedom to make life choices, life ladder, social support, GDP per capita, and healthy life expectancy pointed in similar directions, indicating that these variables are positively correlated. Countries clustering around these arrows, particularly European, North American, and Oceania countries, exhibited high values for these variables, suggesting strong positive societal factors and overall well-being.

On the other hand, arrows for negative affect and perceptions of corruption pointed in opposite directions to the positive factors, indicating a negative correlation. African countries, clustering around these arrows, exhibited higher levels of negative emotions and perceptions of corruption, indicating significant challenges in achieving higher overall happiness. The biplot effectively highlighted the differences between regions, with Asian countries showing a mix of high positive factors and some negative factors, while Central and Southern American countries center around the origin, indicating they have more average happiness profiles.

Conclusion

In summary, we began by examining the relationships between key variables such as GDP per capita, social support, healthy life expectancy, personal freedom, generosity, perceptions of corruption, and emotional states (positive and negative affect). Our exploratory data analysis, including correlation matrices and heatmaps,

revealed that GDP per capita, social support, and healthy life expectancy were strongly positively correlated with life satisfaction (Life Ladder), while perceptions of corruption had a moderate negative relationship.

Through scatter plots and linear regression models, we quantified these relationships. Our findings showed a strong positive relationship between GDP per capita and Life Ladder scores, with higher economic prosperity being associated with higher happiness levels. Similarly, social support had a significant positive impact on happiness, explaining 60.89% of the variance in Life Ladder scores. In contrast, perceptions of corruption had a significant negative impact on happiness, highlighting the detrimental effect of governance issues on well-being.

We also explored temporal trends in happiness across different regions. Our analysis revealed that Oceania maintained the highest average Life Ladder scores over time, while Africa had the lowest. These trends underscore the regional disparities in happiness and the need for tailored approaches to address the unique challenges faced by different regions.

The Principal Component Analysis (PCA) provided further insights by reducing the dimensionality of the data and visualizing the relationships between variables and the clustering of countries by region. The Scree plot indicated that the first two principal components captured 68.7% of the total variance, making them significant for our analysis. The biplot revealed that European countries generally clustered around positive factors such as GDP per capita, social support, life ladder, and healthy life expectancy, indicating strong positive societal factors. In contrast, African countries were more influenced by negative factors like negative affect and perceptions of corruption.