World Happiness report 2021

Parnian Jahangiri Rad

2/14/2022

```
library(tidyverse)
library(janitor)
library(ggcorrplot)
library(caTools)
```

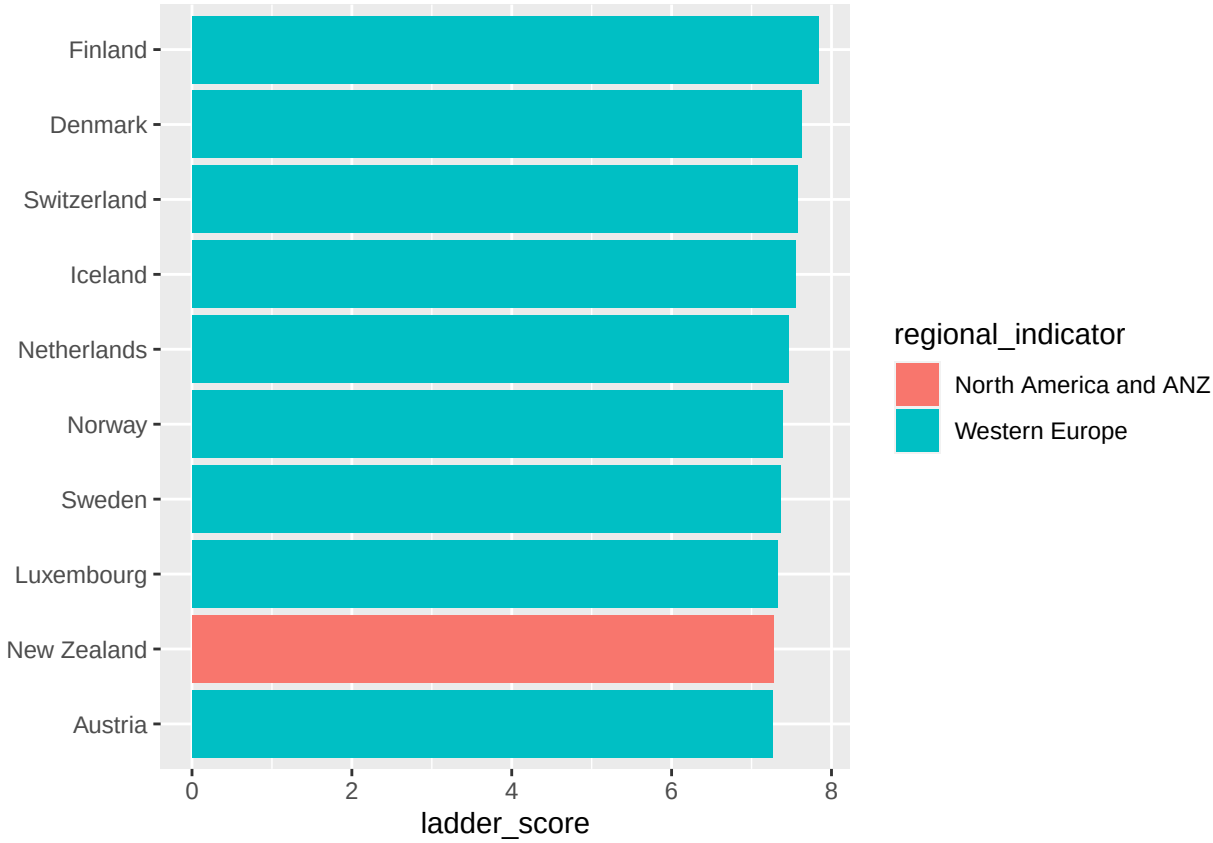
```
data1 <- read_csv("world-happiness-report-2021.csv")
```

```
## Rows: 149 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Country name, Regional indicator
## dbl (18): Ladder score, Standard error of ladder score, upperwhisker, lowerw...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data1 <- data1 %>%
  clean_names()
```

```
#First 10 happiest countries(based on ladder_score):
data2 <- data1 %>%
  select(-starts_with("explained_by"))
#View(data2)
top_10 <- data2 %>%
  arrange(desc(ladder_score)) %>%
  head(10)
bottom_10 <- data2 %>%
  arrange(ladder_score) %>%
  head(10)
```

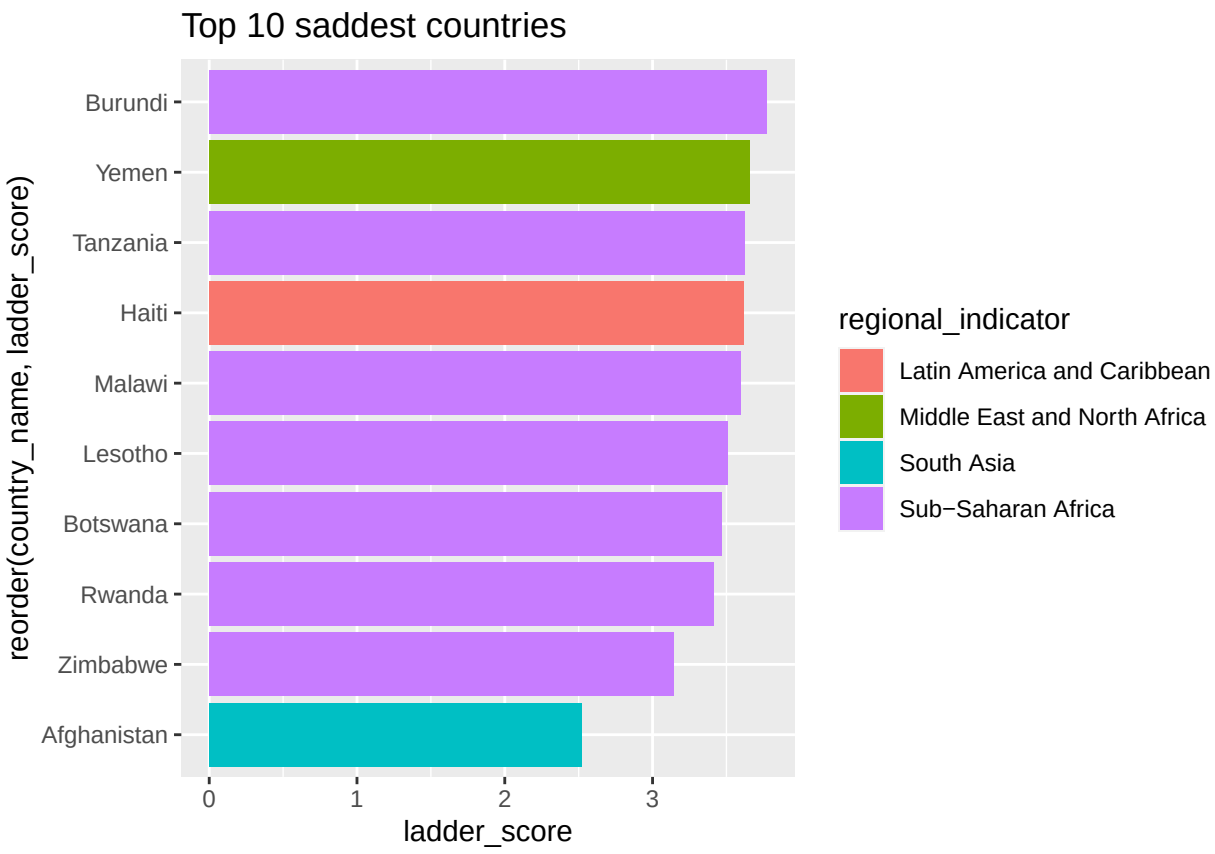
```
ggplot(data = top_10,aes(x = reorder(country_name,ladder_score),
                         y = ladder_score,
                         fill = regional_indicator)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme(axis.title.y = element_blank())
```




```
  ggtitle("Top 10 happiest countries")
```

```
## $title
## [1] "Top 10 happiest countries"
##
## attr(,"class")
## [1] "labels"
```

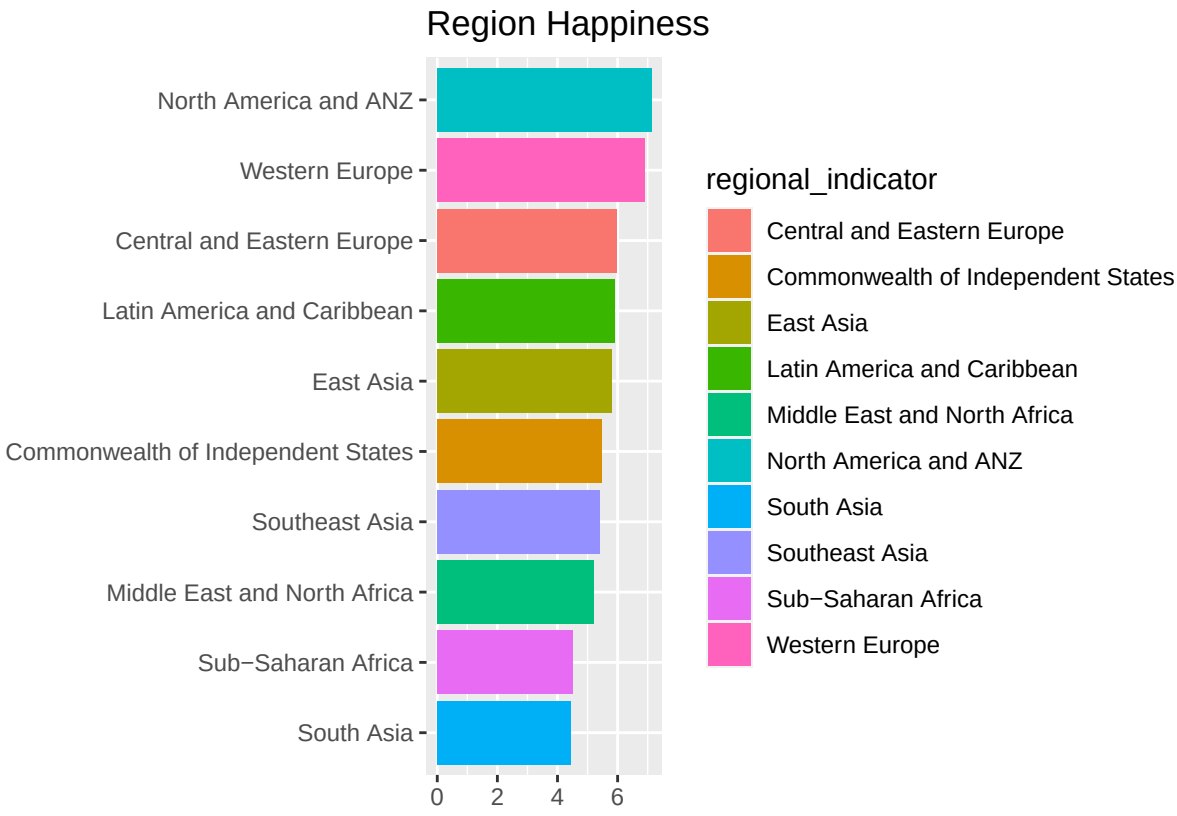
```
ggplot(data = bottom_10,aes(x = reorder(country_name,ladder_score),
                            y = ladder_score,
                            fill = regional_indicator)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  ggtitle("Top 10 saddest countries")
```



Find happiest regions

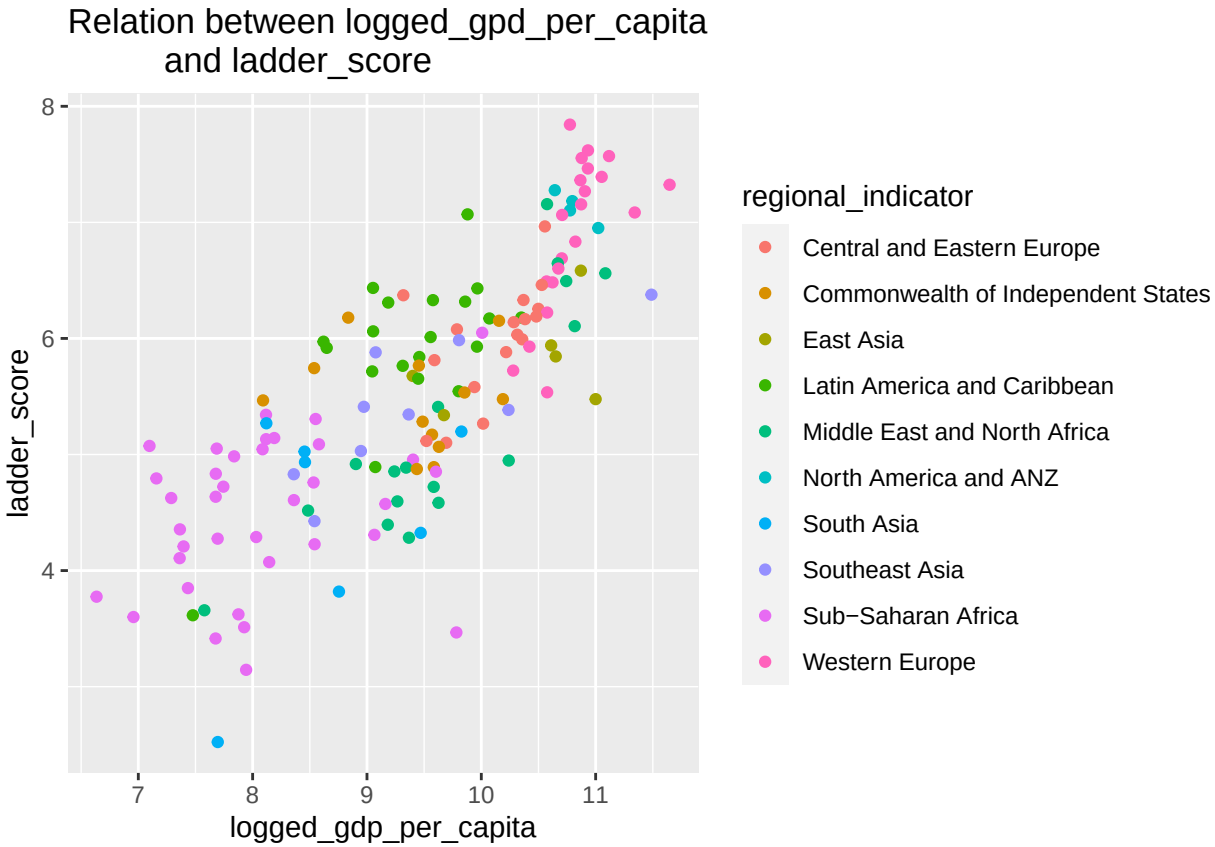
```
region_happiness <- data2 %>%
  group_by(regional_indicator) %>%
  summarise(mean(ladder_score))
```

```
ggplot(data = region_happiness,
       aes(x = reorder(regional_indicator, `mean(ladder_score)`),
           y = `mean(ladder_score)`,
       fill = regional_indicator)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  xlab("") +
  ylab("") +
  ggtitle("Region Happiness")
```



Relation between logged_gpd_per_capita and ladder_score

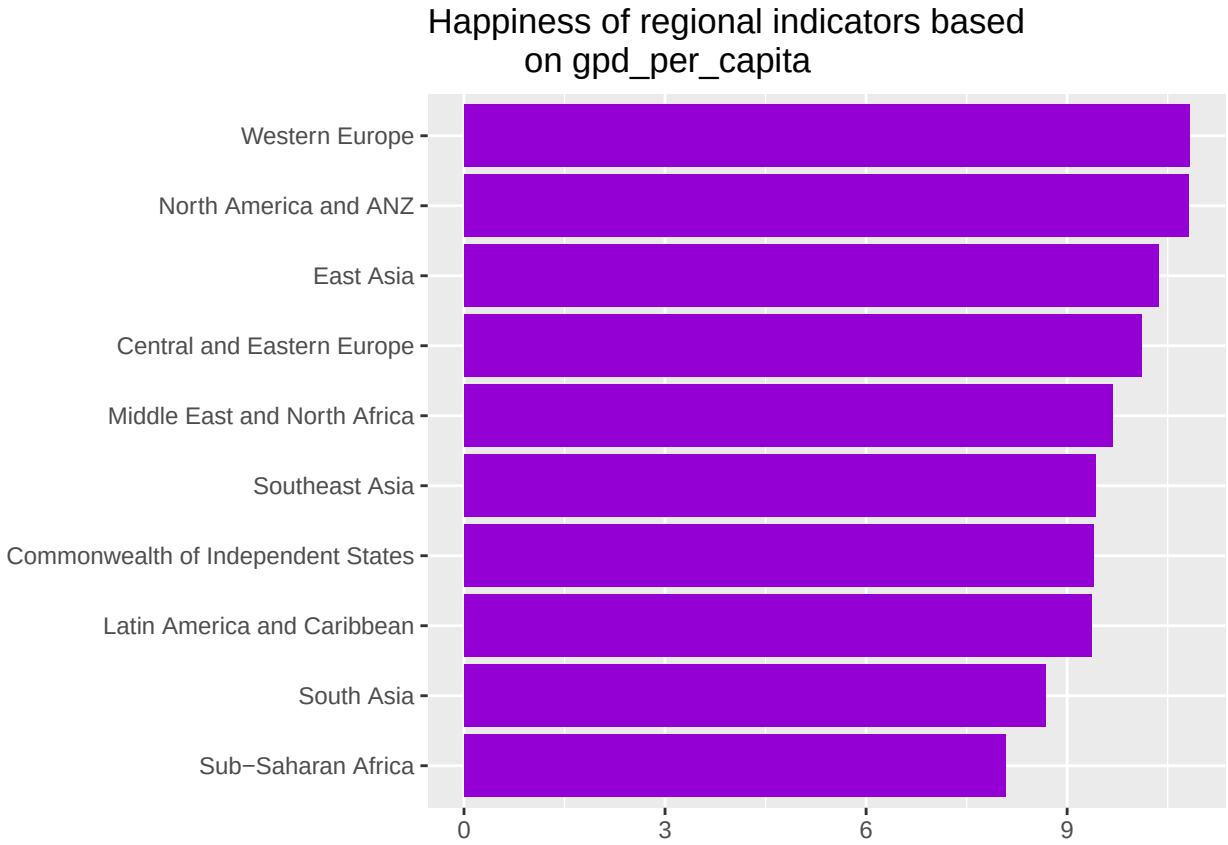
```
ggplot(data = data2 ,
       aes(x = logged_gdp_per_capita ,
           y = ladder_score)) +
  geom_point(aes(color = regional_indicator)) +
  ggtitle("Relation between logged_gpd_per_capita
          and ladder_score")
```



increase in `logged_gpd_per_capita` increases `ladder_score`.

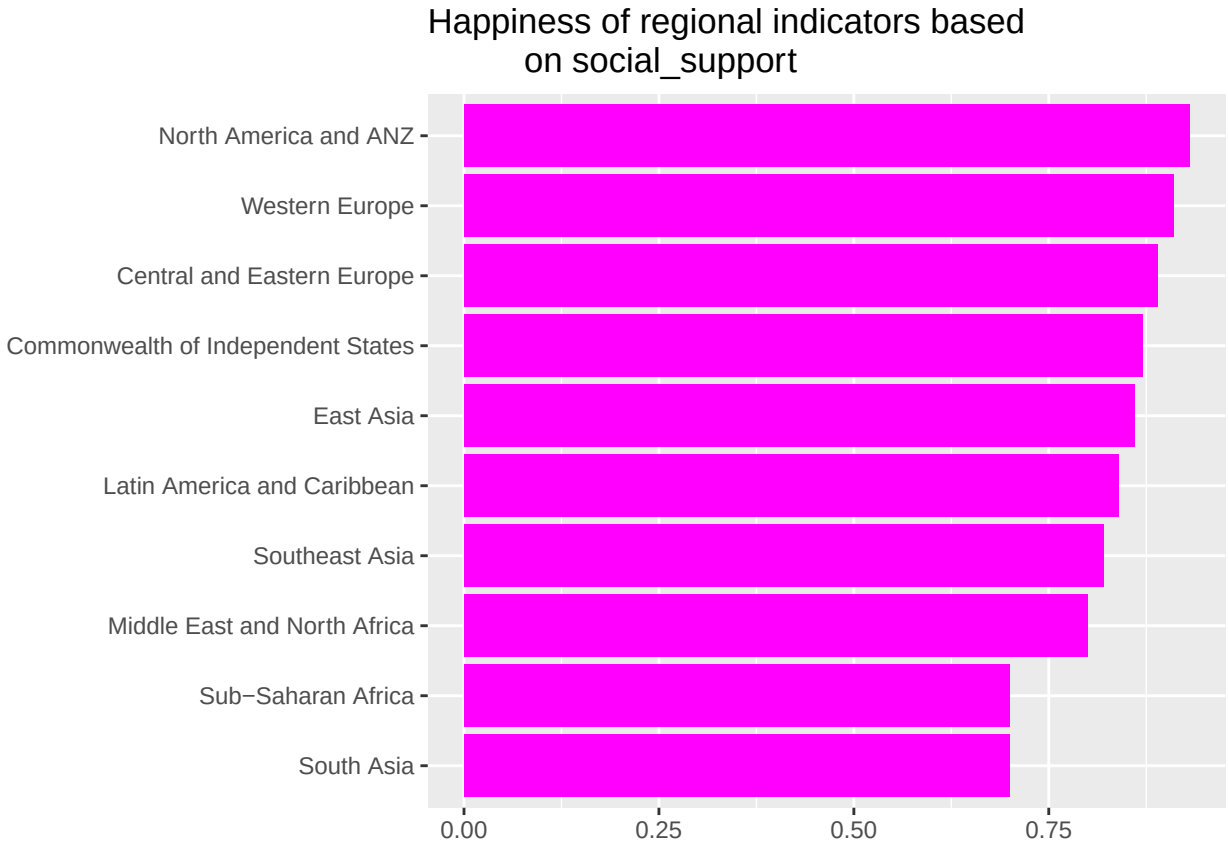
There are 7 features which affects happiness based on `ladder_score`:
- `log_gpd_per_capita`
- `social_support`
- `healthy_life_expendancy`
- `freedom_to_make_life_choices`
- `generosity`
- `perceptions_of_corruption`
- `dystopia_residual`

```
#regional_data based on logged_gpd_per_capita
regional_data_gpd <- data2 %>%
  group_by(regional_indicator) %>%
  summarise(mean(logged_gdp_per_capita))
is.num <- sapply(regional_data_gpd,is.numeric)
regional_data_gpd[is.num] <- lapply(regional_data_gpd[is.num], round, 2)
ggplot(regional_data_gpd,aes(
        x = reorder(regional_indicator,`mean(logged_gdp_per_capita)`),
        y = `mean(logged_gdp_per_capita)`))+
        coord_flip()+
  geom_bar(stat="identity",fill="darkviolet") +
  theme(axis.title.x = element_blank(),
          axis.title.y = element_blank()) +
  ggtitle("Happiness of regional indicators based
          on gpd_per_capita")
```



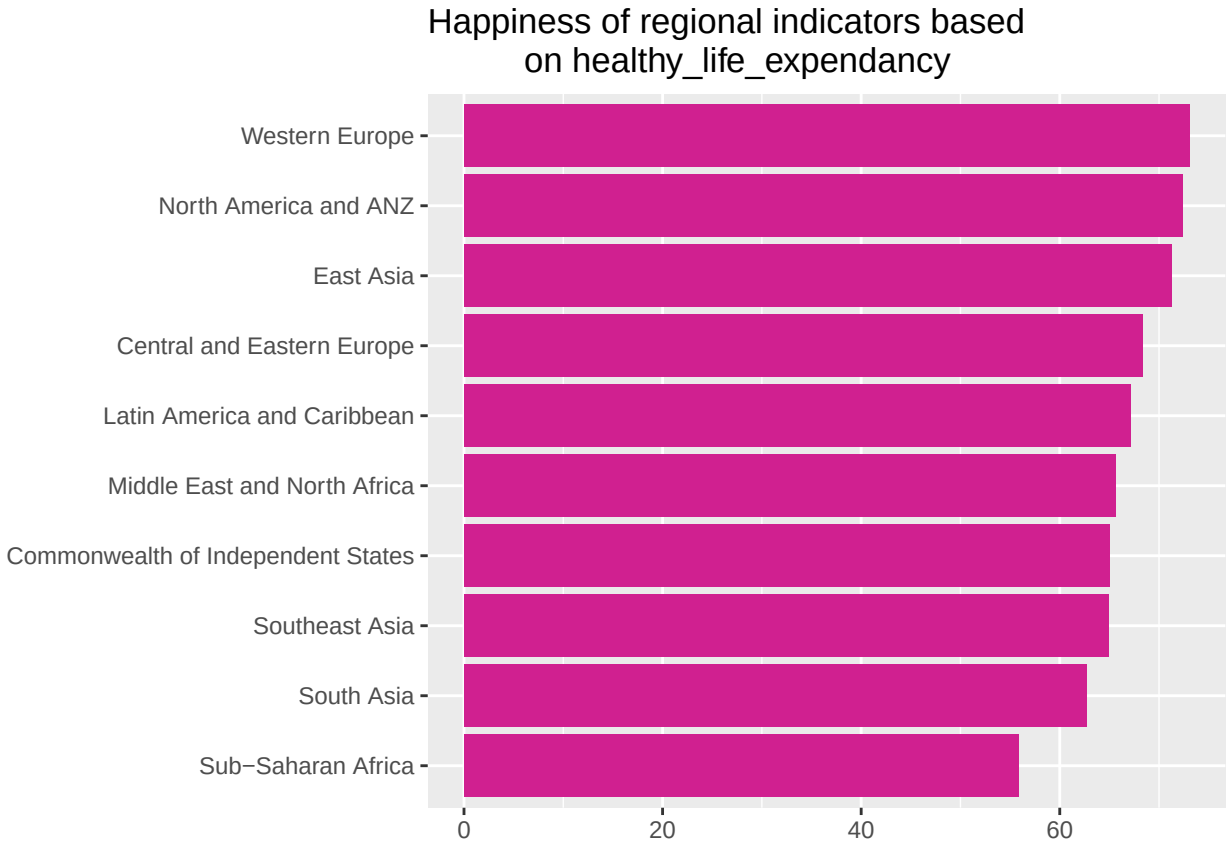
It seems that `logged_gdp_per_capita` has a high impact on ladder_score.(important feature).

```
#regional_data based on social_support
regional_data_social_support <- data2 %>%
  group_by(regional_indicator) %>%
  summarise(mean(social_support))
is.num <- sapply(regional_data_social_support,is.numeric)
regional_data_social_support[is.num] <- lapply(regional_data_social_support[is.num], round, 2)
ggplot(regional_data_social_support,aes(
        x = reorder(regional_indicator,`mean(social_support)`),
        y = `mean(social_support)`))+
        coord_flip()+
  geom_bar(stat="identity",fill="magenta") +
  theme(axis.title.x = element_blank(),
          axis.title.y = element_blank()) +
  ggtitle("Happiness of regional indicators based
          on social_support")
```



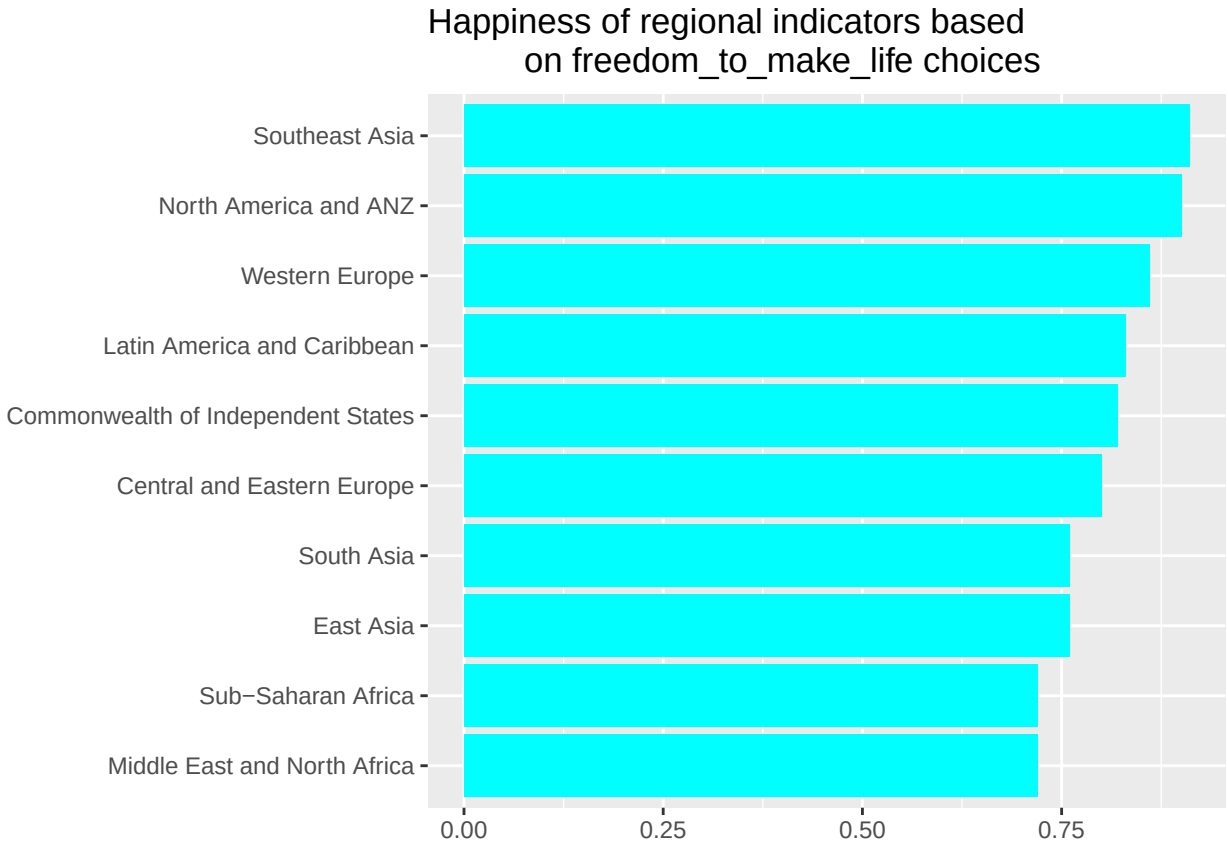
`social_support` seems to be an important feature as well.

```
#regional_data based on healthy_life_expendancy
regional_data_healthy <- data2 %>%
  group_by(regional_indicator) %>%
  summarise(mean(healthy_life_expectancy))
is.num <- sapply(regional_data_healthy,is.numeric)
regional_data_healthy[is.num] <- lapply(regional_data_healthy[is.num], round, 2)
ggplot(regional_data_healthy,aes(
        x = reorder(regional_indicator,`mean(healthy_life_expectancy)`),
        y = `mean(healthy_life_expectancy)`))+
        coord_flip()+
  geom_bar(stat="identity",fill="violetred") +
  theme(axis.title.x = element_blank(),
          axis.title.y = element_blank()) +
  ggtitle("Happiness of regional indicators based
          on healthy_life_expendancy")
```



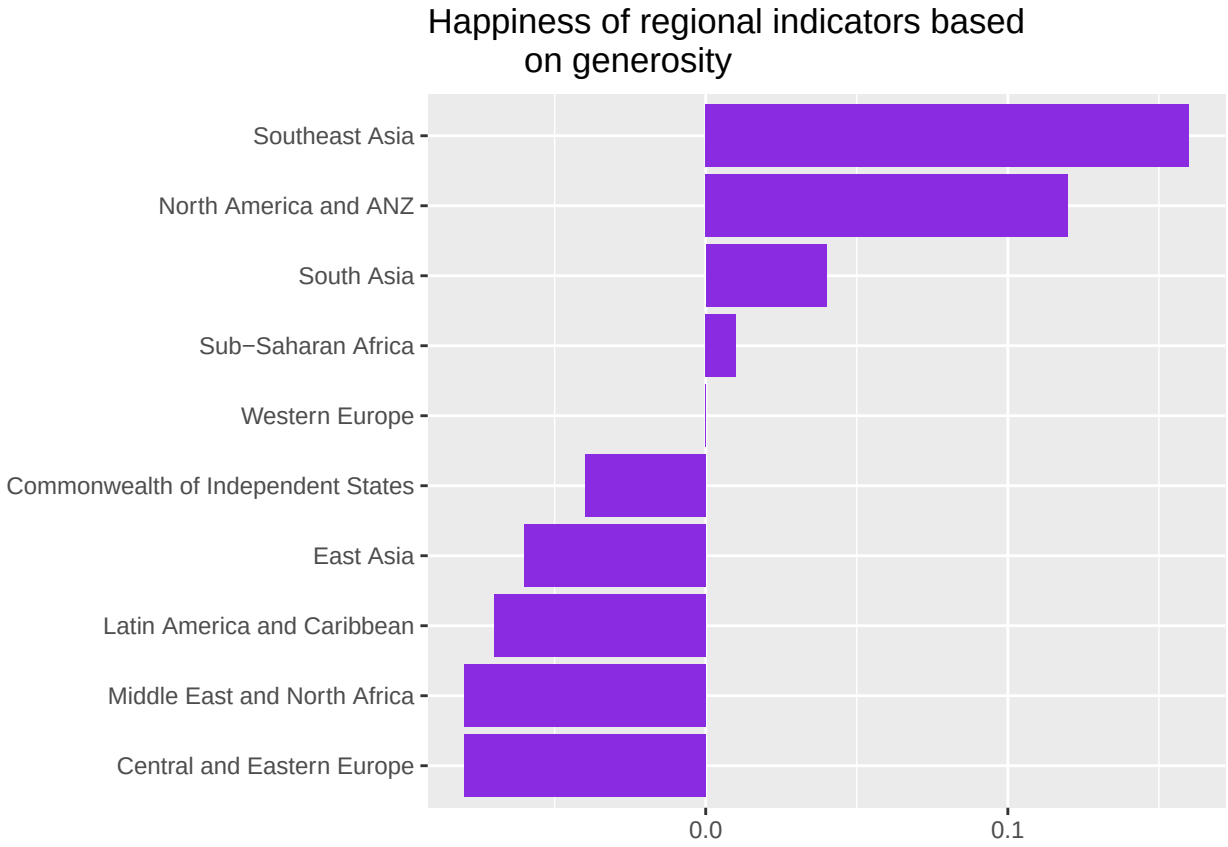
Let's check out freedom_to_make_life_choices:

```
#regional_data based on freedom_to_make_life choices
regional_data_freedom <- data2 %>%
  group_by(regional_indicator) %>%
  summarise(mean(freedom_to_make_life_choices))
is.num <- sapply(regional_data_freedom,is.numeric)
regional_data_freedom[is.num] <- lapply(regional_data_freedom[is.num], round, 2)
ggplot(regional_data_freedom,aes(
        x = reorder(regional_indicator,`mean(freedom_to_make_life_choices)`),
        y = `mean(freedom_to_make_life_choices)`))+
        coord_flip()+
  geom_bar(stat="identity",fill="cyan") +
  theme(axis.title.x = element_blank(),
          axis.title.y = element_blank()) +
  ggtitle("Happiness of regional indicators based
          on freedom_to_make_life choices")
```

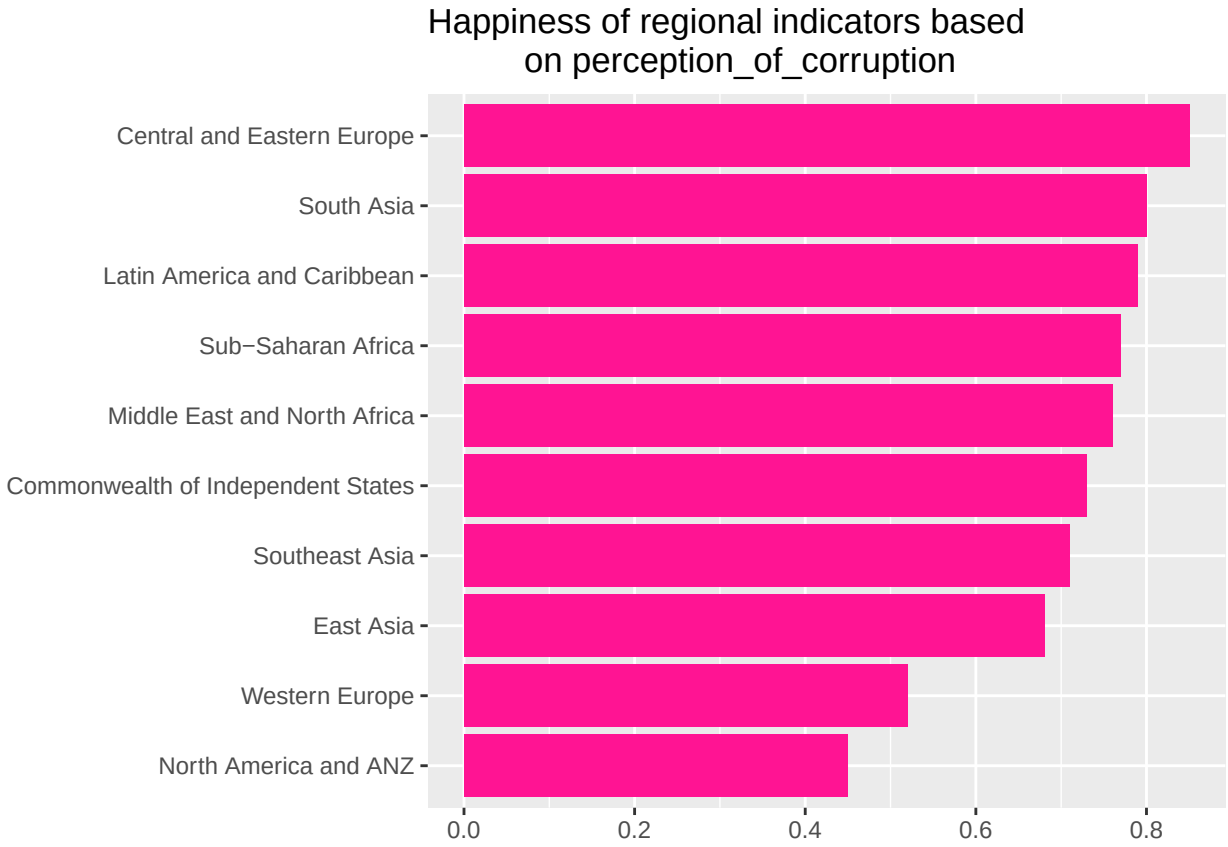


Interesting!!

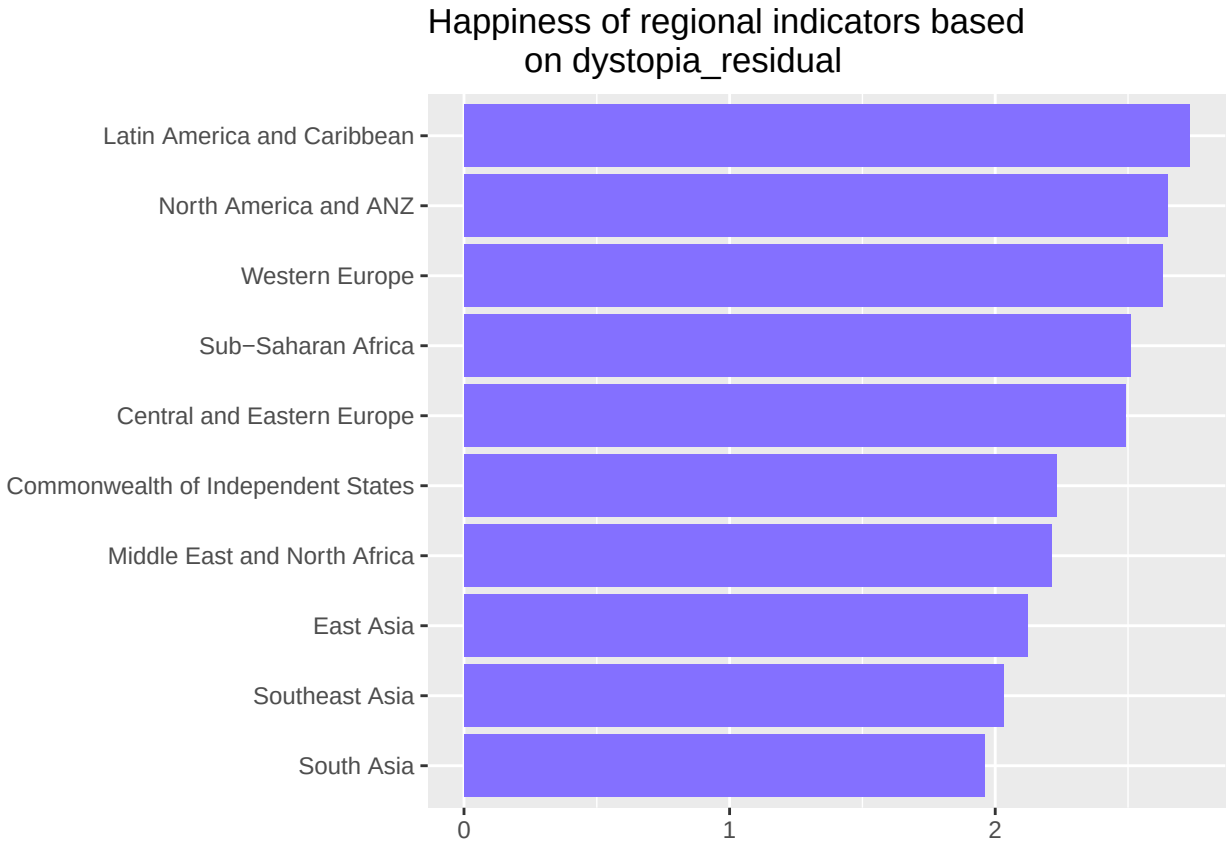
```
#regional_data based on generosity
regional_data_generosity <- data2 %>%
  group_by(regional_indicator) %>%
  summarise(mean(generosity))
is.num <- sapply(regional_data_generosity,is.numeric)
regional_data_generosity[is.num] <- lapply(regional_data_generosity[is.num], round, 2)
ggplot(regional_data_generosity,aes(
        x = reorder(regional_indicator,`mean(generosity)`),
        y = `mean(generosity)`))+
        coord_flip()+
  geom_bar(stat="identity",fill="blueviolet") +
  theme(axis.title.x = element_blank(),
          axis.title.y = element_blank()) +
  ggtitle("Happiness of regional indicators based
          on generosity")
```




```
#regional_data based on perception_of_corruption
regional_data_perception <- data2 %>%
  group_by(regional_indicator) %>%
  summarise(mean(perceptions_of_corruption))
is.num <- sapply(regional_data_perception,is.numeric)
regional_data_perception[is.num] <- lapply(regional_data_perception[is.num], round, 2)
ggplot(regional_data_perception,aes(
        x = reorder(regional_indicator,`mean(perceptions_of_corruption)`),
        y = `mean(perceptions_of_corruption)`))+
        coord_flip()+
  geom_bar(stat="identity",fill="deeppink") +
  theme(axis.title.x = element_blank(),
          axis.title.y = element_blank()) +
  ggtitle("Happiness of regional indicators based
          on perception_of_corruption")
```




```
#regional_data based on dystopia_residual
regional_data_dystopia <- data2 %>%
  group_by(regional_indicator) %>%
  summarise(mean(dystopia_residual))
is.num <- sapply(regional_data_dystopia,is.numeric)
regional_data_dystopia[is.num] <- lapply(regional_data_dystopia[is.num], round, 2)
ggplot(regional_data_dystopia,aes(
        x = reorder(regional_indicator,`mean(dystopia_residual)`),
        y = `mean(dystopia_residual)`))+
        coord_flip()+
  geom_bar(stat="identity",fill="lightslateblue") +
  theme(axis.title.x = element_blank(),
          axis.title.y = element_blank()) +
  ggtitle("Happiness of regional indicators based
          on dystopia_residual")
```




```
regional_mean_data <- inner_join(region_happiness,
  inner_join(regional_data_gpd,
  inner_join(regional_data_social_support,
  inner_join(regional_data_healthy,
  inner_join(regional_data_freedom,
  inner_join(regional_data_generosity,
  inner_join(regional_data_perception,
                                regional_data_dystopia,
                                by="regional_indicator"),
           by="regional_indicator"),
         by="regional_indicator"),
       by="regional_indicator"),
     by="regional_indicator"),
   by="regional_indicator"),
 by="regional_indicator")
```

Feature importance using **correlation** (we will also plot it):

```
cor(regional_mean_data$`mean(ladder_score)`,
    regional_mean_data$`mean(logged_gdp_per_capita)`)
```

```
## [1] 0.9173778
```

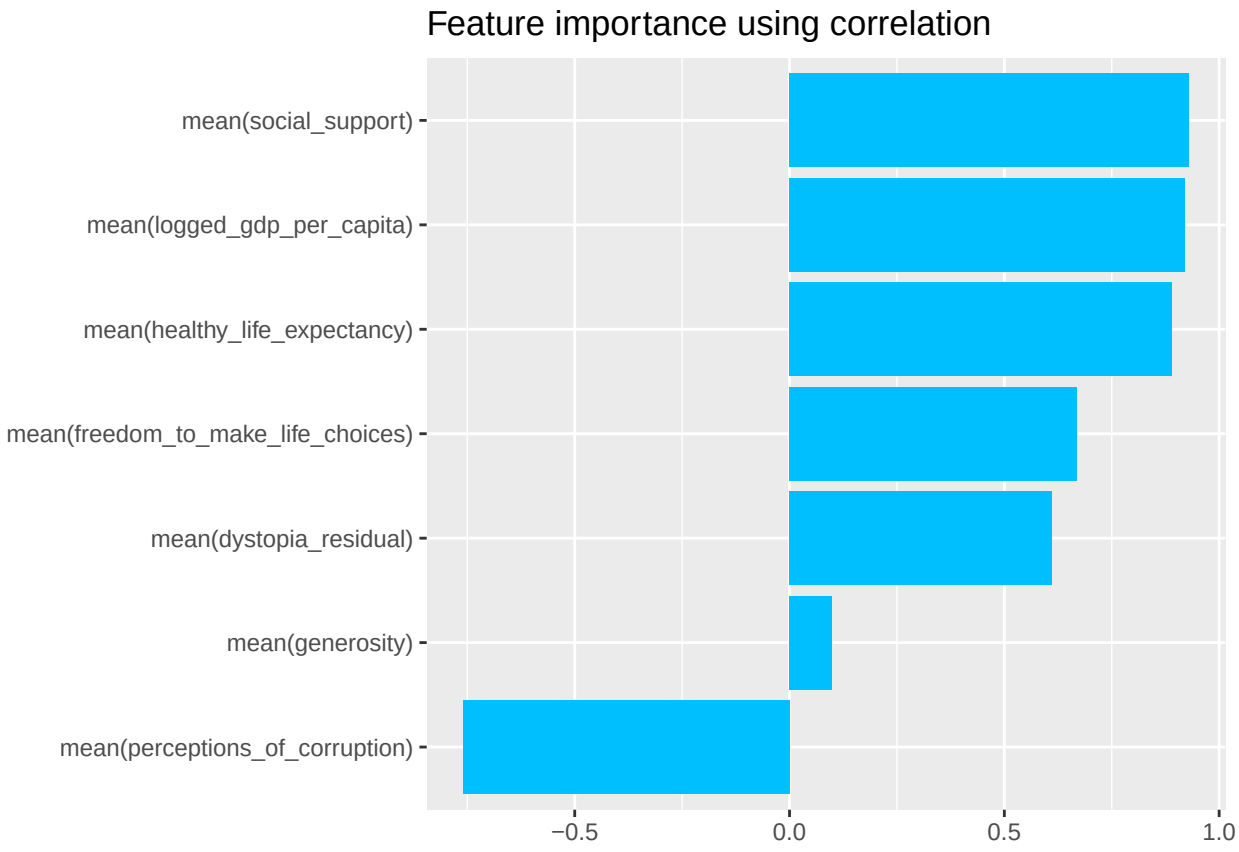
```
cor(regional_mean_data$`mean(ladder_score)`,
    regional_mean_data$`mean(social_support)`)
```

```
## [1] 0.9294476
```

```
names <- names(regional_mean_data)
names <- names[-c(1,2)]
y <- vector()
for(i in names){
  y <- append(y,
              cor(regional_mean_data$`mean(ladder_score)`,
                  regional_mean_data[[i]]))
}
y
```

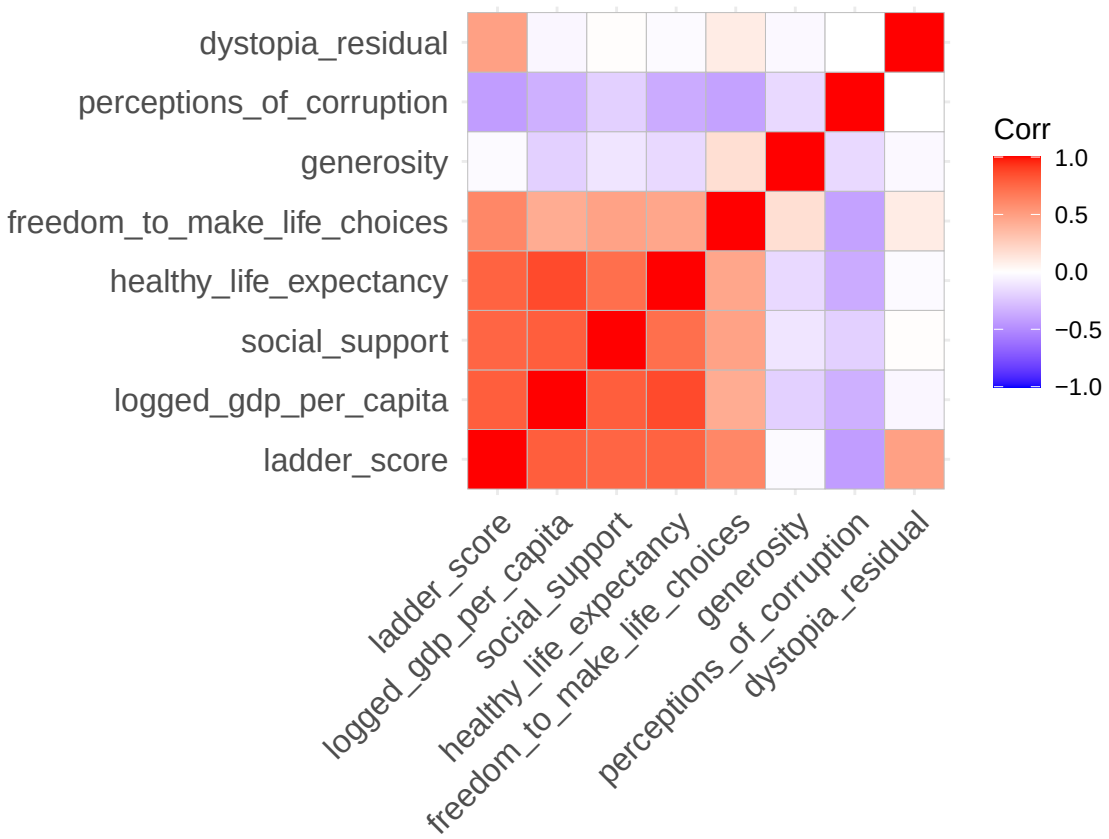
```
## [1]  0.91737784  0.92944760  0.89038409  0.67264629  0.09779774 -0.75544218
## [7]  0.60798861
```

```
df1 <- data.frame(names,y)
is.num <- sapply(regional_data_social_support,is.numeric)
df1[is.num] <- lapply(df1[is.num], round, 2)
ggplot(df1,aes(
        x = reorder(names,y),
        y = y)) +
        coord_flip() +
  geom_bar(stat="identity",fill="deepskyblue") +
  theme(axis.title.x = element_blank(),
          axis.title.y = element_blank()) +
  ggtitle("Feature importance using correlation")
```



Pearson Correlation Matrix

```
cdf <- data2 %>%
  select(ladder_score,
         logged_gdp_per_capita,
         social_support,
         healthy_life_expectancy,
         freedom_to_make_life_choices,
         generosity,
         perceptions_of_corruption,
         dystopia_residual)
cor_df <- data.frame(cor(cdf))
ggcorrplot(cor_df, method ="square")
```



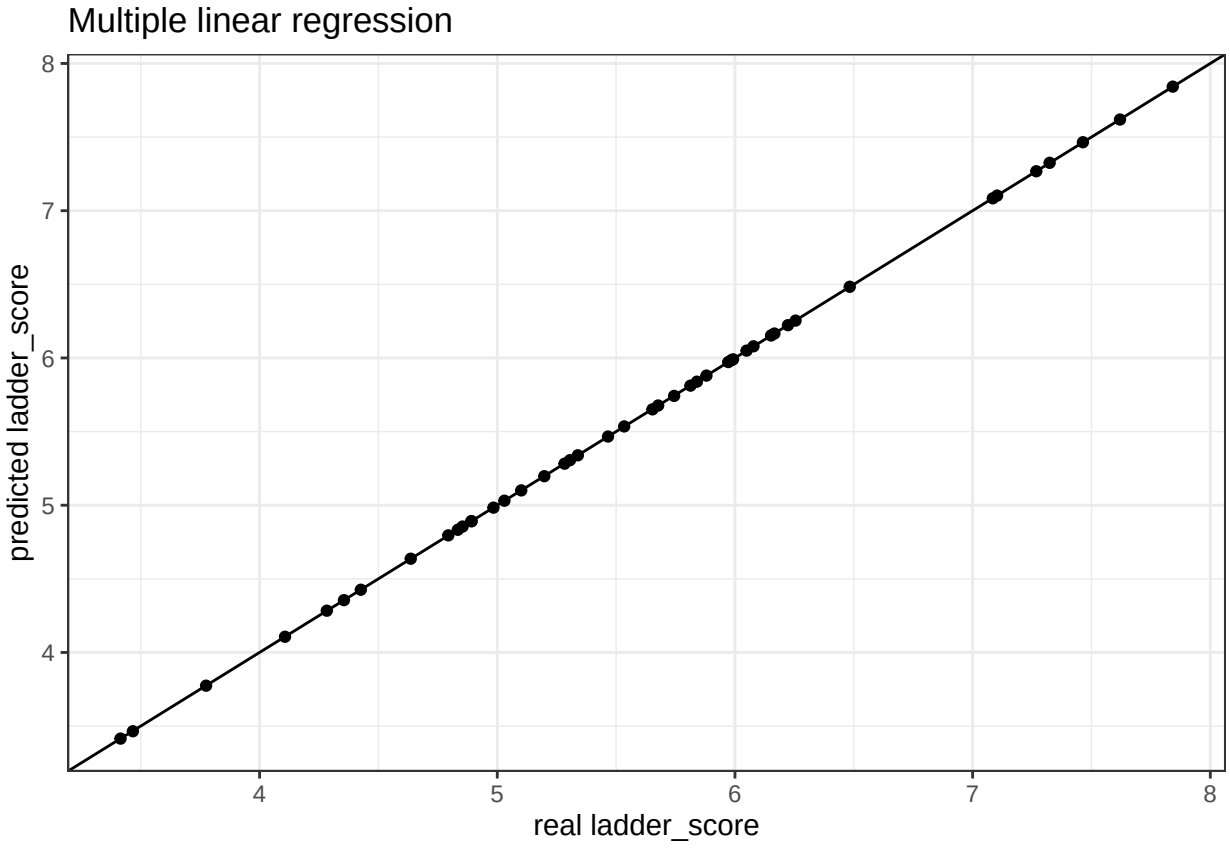
Linear regression

```
set.seed(66)
split = sample.split(cdf$ladder_score, SplitRatio = 0.7)
training_set = subset(cdf, split == TRUE)
test_set = subset(cdf, split == FALSE)
lm_model <- lm(formula = ladder_score ~ ., data = training_set)
summary(lm_model)
```

```
##
## Call:
## lm(formula = ladder_score ~ ., data = training_set)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0019883 -0.0005445 -0.0001218  0.0005847  0.0021011
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.563e+00  1.357e-03 -3363.9   <2e-16 ***
## logged_gdp_per_capita          3.494e-01  1.799e-04  1941.6   <2e-16 ***
## social_support                 2.252e+00  1.336e-03  1685.5   <2e-16 ***
## healthy_life_expectancy        3.149e-02  2.621e-05  1201.4   <2e-16 ***
## freedom_to_make_life_choices   1.217e+00  9.705e-04  1254.0   <2e-16 ***
```

```
## generosity                    6.521e-01  6.433e-04  1013.7   <2e-16 ***
## perceptions_of_corruption    -6.390e-01  6.405e-04  -997.6   <2e-16 ***
## dystopia_residual             1.000e+00  1.741e-04  5745.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0009039 on 96 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.093e+07 on 7 and 96 DF,  p-value: < 2.2e-16
```

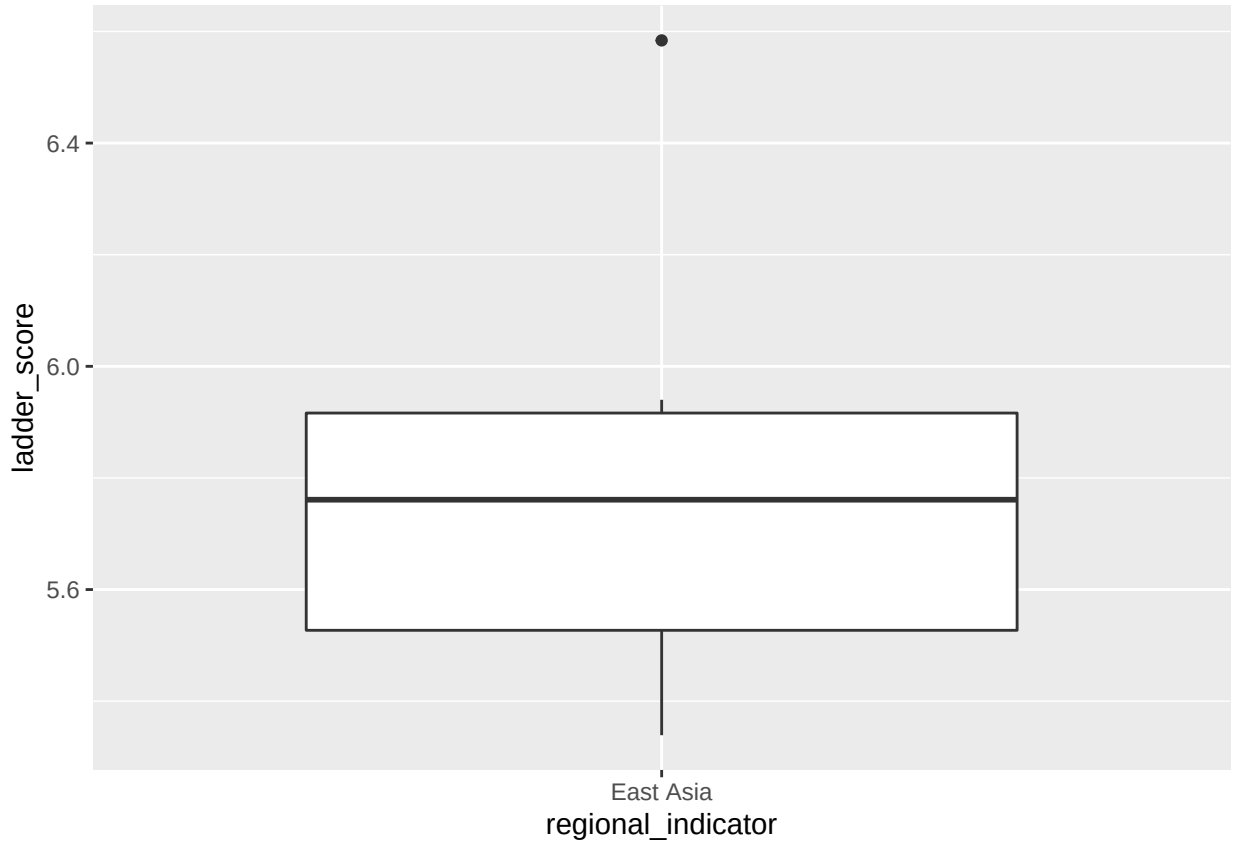
```
y_pred <- predict(lm_model, newdata = test_set)
pred_df <- as.data.frame(cbind(Prediction = y_pred , Actual = test_set$ladder_score))
ggplot(pred_df, aes(Actual, Prediction )) +
  geom_point() +
  theme_bw() +
  geom_abline() +
  labs(title = "Multiple linear regression" ,
       x = "real ladder_score",
       y = "predicted ladder_score")
```



One-sample T-test

East Asia as our sample

```r
ea <- data2 %>%
  filter(regional_indicator == "East Asia")
ggplot(data = ea,aes(
  x = regional_indicator,
  y = ladder_score
      )) +
  geom_boxplot()
```




```r
t.test(ea$ladder_score, mu = 5)
```

```
##
##  One Sample t-test
##
## data:  ea$ladder_score
## t = 4.512, df = 5, p-value = 0.006329
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  5.348672 6.271994
## sample estimates:
## mean of x
##  5.810333
```