Intermediate Analysis

Final Project Presentation

Final Project Presentation.

Intermediate Analysis

(ALY6015)

Guided by:
Prof. Yvonne Leung

Submitted by:

Pratikkumar Indravadan Malaviya (002963548)

Dhruvil Sunilbhai Patel (002955193)

Taiye Murtala (002965496)

Date of submission :

18th May’2022

AGENDA

DATASET

- Information of Dataset
- Data Understanding

DATA ANALYSIS

- Identify the questions
- Explanation of methods

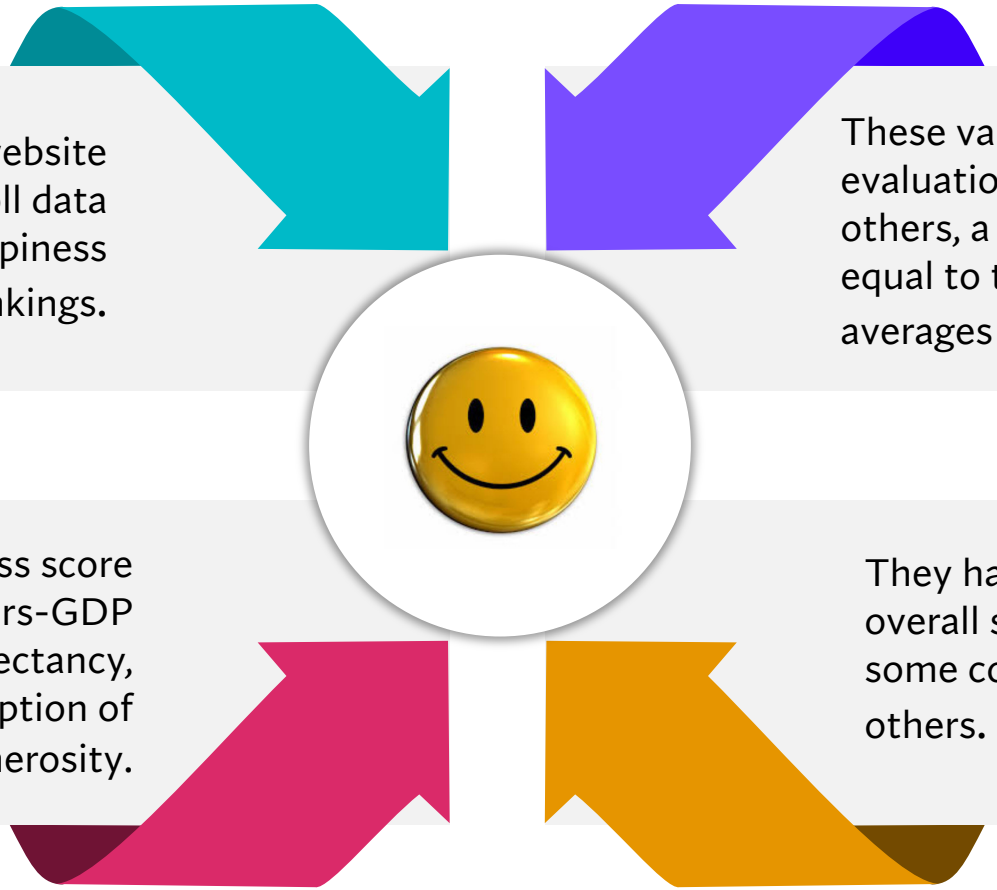
SUMMARY

- Conclusion

DATASET

The dataset is owned by the website kagle.com. The Gallup World Poll data was used to calculate the happiness scores and rankings.

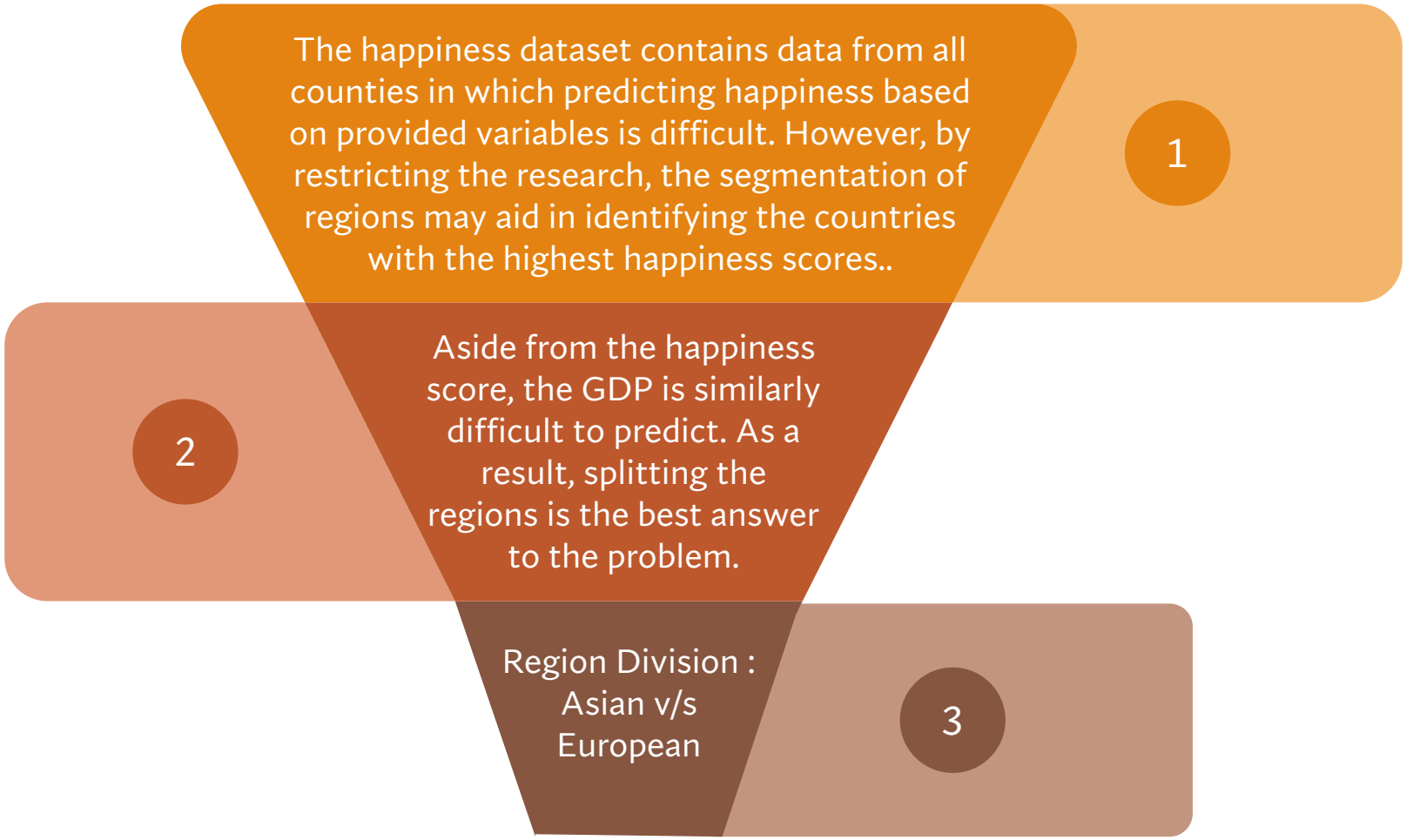
These variable contribute to making life evaluations higher in each country than others, a hypothetical country with values equal to the world's lowest national averages for each of the six factors.



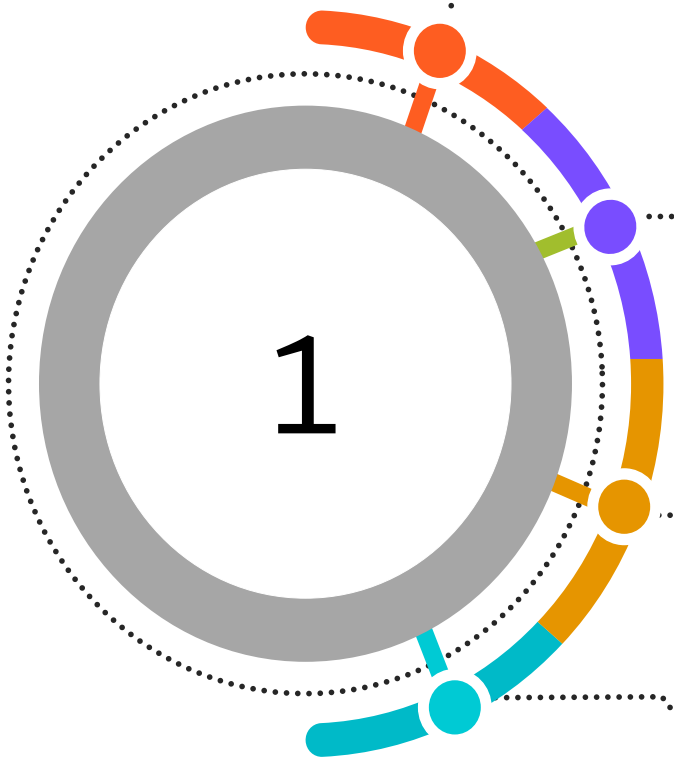
The columns following the happiness score calculate how much each of six factors-GDP per capita, social support, life expectancy, freedom of age life choices, perception of corruption, and generosity.

They have no bearing on each country's overall score, but they do explain why some countries are ranked higher than others.

STRATEGY


The happiness dataset contains data from all counties in which predicting happiness based on provided variables is difficult. However, by restricting the research, the segmentation of regions may aid in identifying the countries with the highest happiness scores..
1
Aside from the happiness score, the GDP is similarly difficult to predict. As a result, splitting the regions is the best answer to the problem.
2
Region Division : Asian v/s European
3

IDENTIFY THE QUESTIONS



Goal

Our goal is to compare Happiness Score.

Question

What predicts Happiness Score ?

Variables

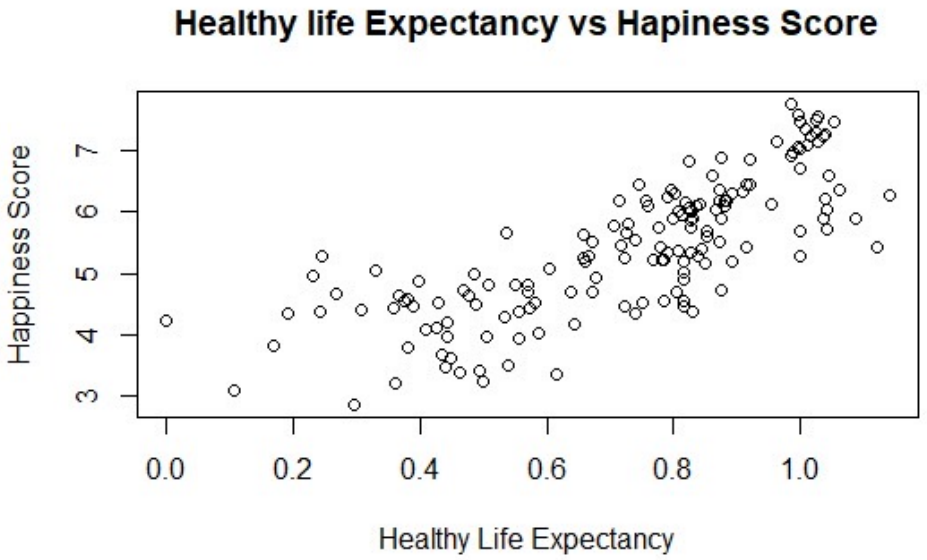
Dependent : Happiness Score
Independent : Healthy Life Expectancy , Freedom of make life choices.

Method

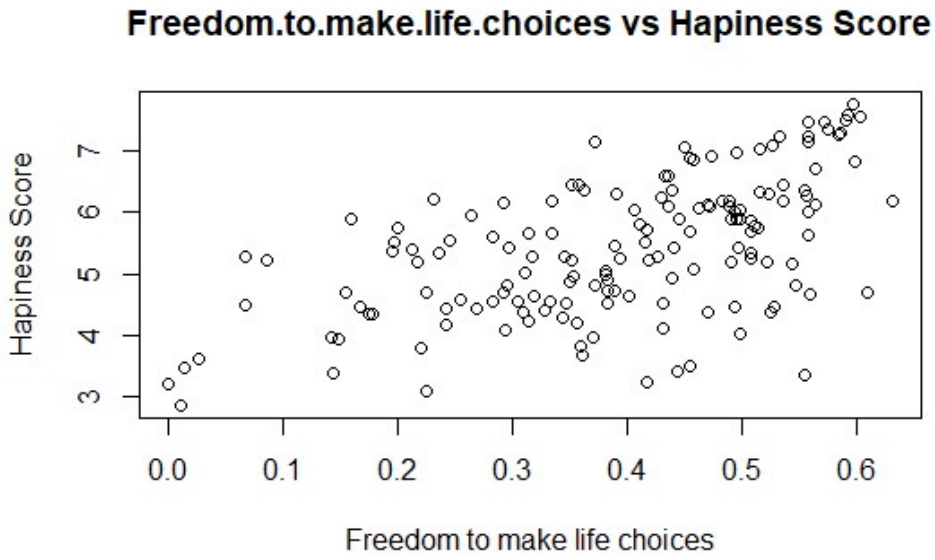
Simple Linear Regression

Simple linear regression is used to estimate the relationship between two quantitative variables. we use simple linear regression when we want to know: How strong the relationship is between two variables.

EDA : Question-1

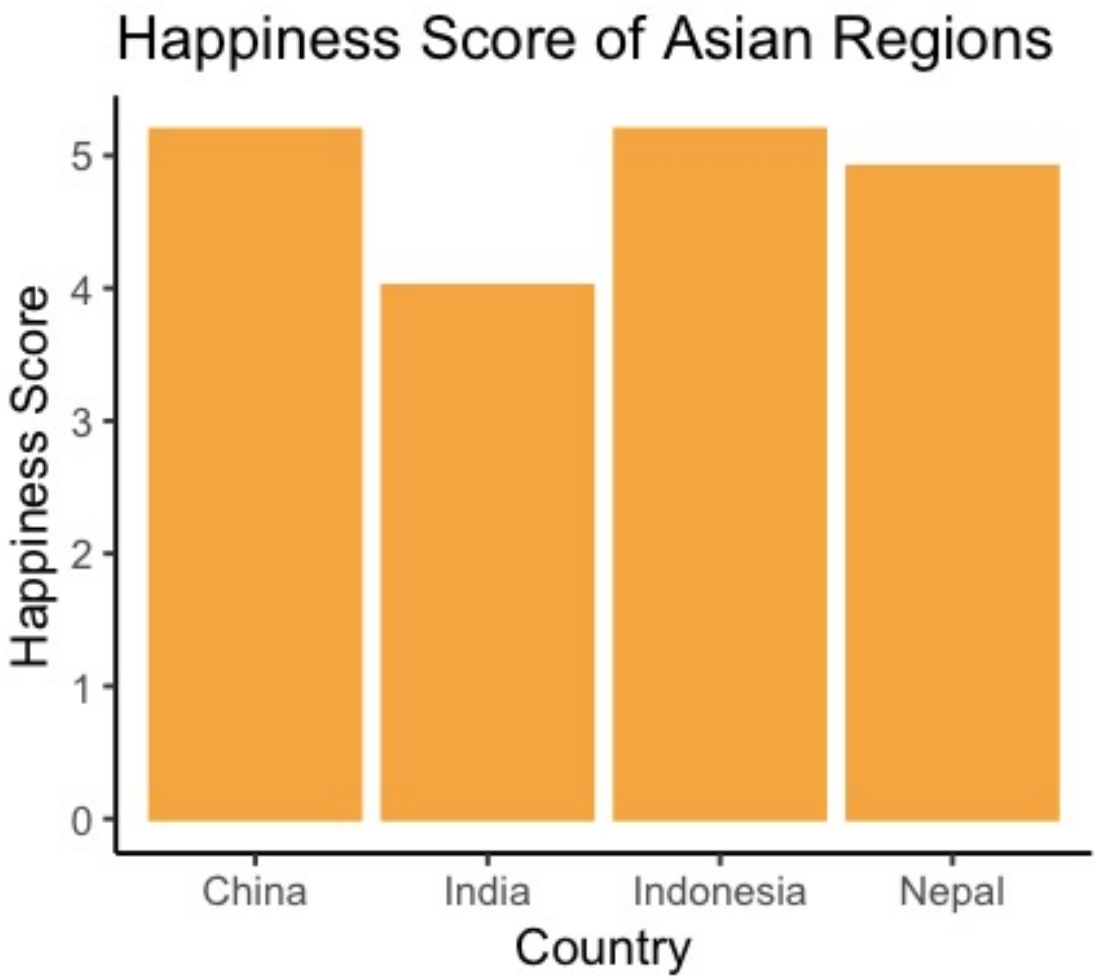


- The plot shows that as the Healthy Life Expectancy rises, so does the Happiness Score



- Similarly, as people's freedom to make life decisions grows, so does their happiness score.

DIVISON OF REGIONS



INTERPRETATION FOR ASIAN REGION

```
> model1<-lm(Score~Healthy.life.expectancy+Freedom.to.make.life.choices,data = Asia)

> summary(model1)

Call:
lm(formula = Score ~ Healthy.life.expectancy + Freedom.to.make.life.choices,
    data = Asia)

Residuals:
       1        2        3        4
 0.54214 -0.09806 -0.10255 -0.34154

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.230      5.363   0.975    0.508
Healthy.life.expectancy         3.501      3.245   1.079    0.476
Freedom.to.make.life.choices   -5.888     12.318  -0.478    0.716

Residual standard error: 0.6563 on 1 degrees of freedom
Multiple R-squared:  0.5381,    Adjusted R-squared:  -0.3857
F-statistic: 0.5825 on 2 and 1 DF,  p-value: 0.6796
```

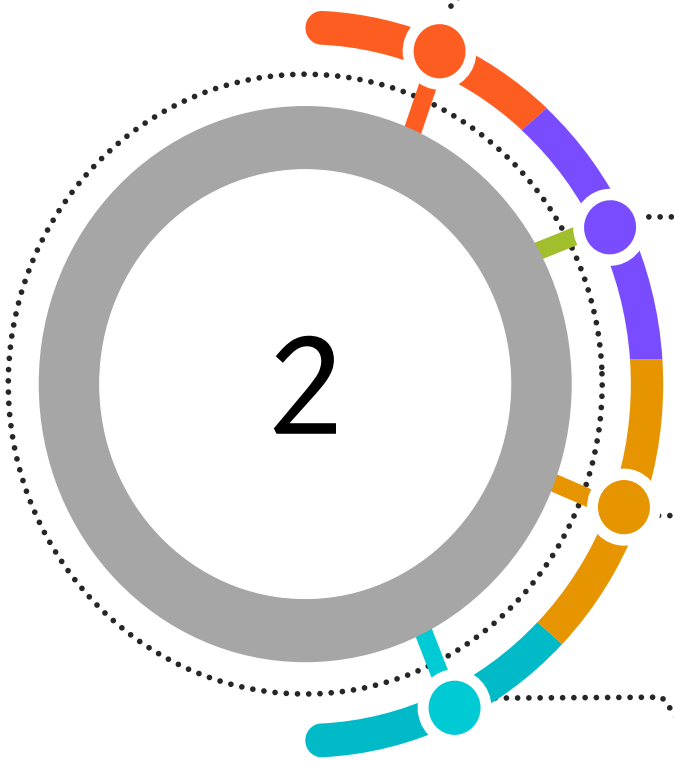
- From the following we can say that healthy life expectancy and freedom to make life choices are not significant because they are both greater than alpha.
- The coefficient value indicates the mean change in the dependent variable given a one unit change in the independent variable.
- We can therefore say that as healthy life expectancy increases by a unit so also happiness score increases by **3.501**. The same happens to freedom to make life choices but in this case, as the freedom to make life choices increases by a unit there is also a decrease by **– 5.888** in happiness score.

INTERPRETATION FOR EUROPEAN REGION

```
> model2<-lm(Score~Healthy.life.expectancy+Freedom.to.make.life.choices,data =Europe)
> summary(model2)

Call:
lm(formula = Score ~ Healthy.life.expectancy + Freedom.to.make.life.choices,
    data = Europe)

Residuals:
       1        2        3        4
 0.15722 -0.17036 -0.02352  0.03666

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.7736     1.2359   1.435   0.3874
Healthy.life.expectancy        3.1756     1.3519   2.349   0.2562
Freedom.to.make.life.choices   4.5420     0.5764   7.879   0.0804 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2359 on 1 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9665
F-statistic: 44.33 on 2 and 1 DF,  p-value: 0.1056
```

- From the following we can say that healthy life expectancy and freedom to make life choices are not significant because they are both greater than alpha.
- The coefficient value indicates the mean change in the dependent variable given a one unit change in the independent variable.
- We can therefore say that as healthy life expectancy increases by a unit so also happiness score increases by **3.1756** The same applies to freedom to make life choice which also make happiness score increases by **4.5420**

- **By comparing the statistics from both the regions, we can conclude that the European regions are having higher happiness score than that of the Asian regions.**



Goal

Our objective is to identify countries with the greatest GDP.

Question

If Freedom to make life choices predicts GDP ?

Variables

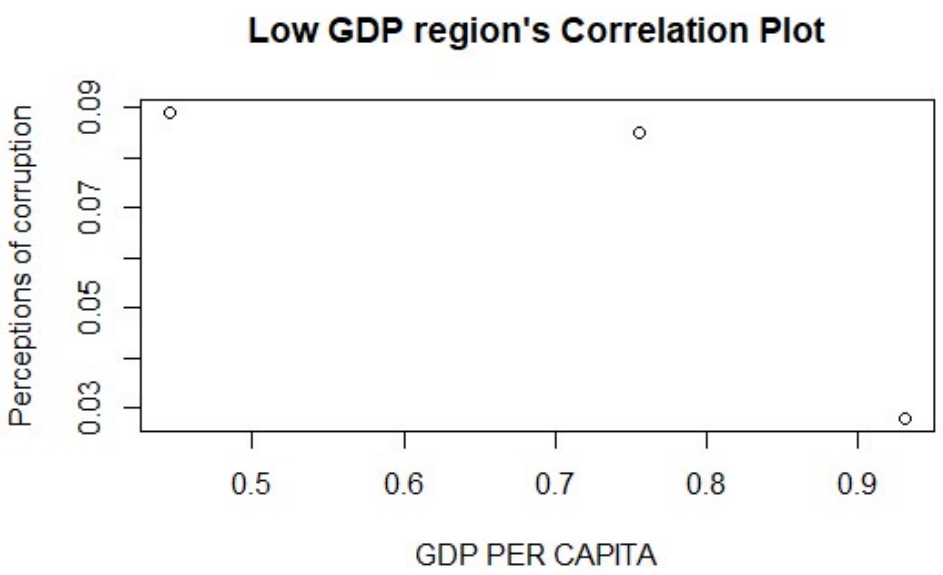
Dependent : Perception of corruption.
Independent : GDP per capital.

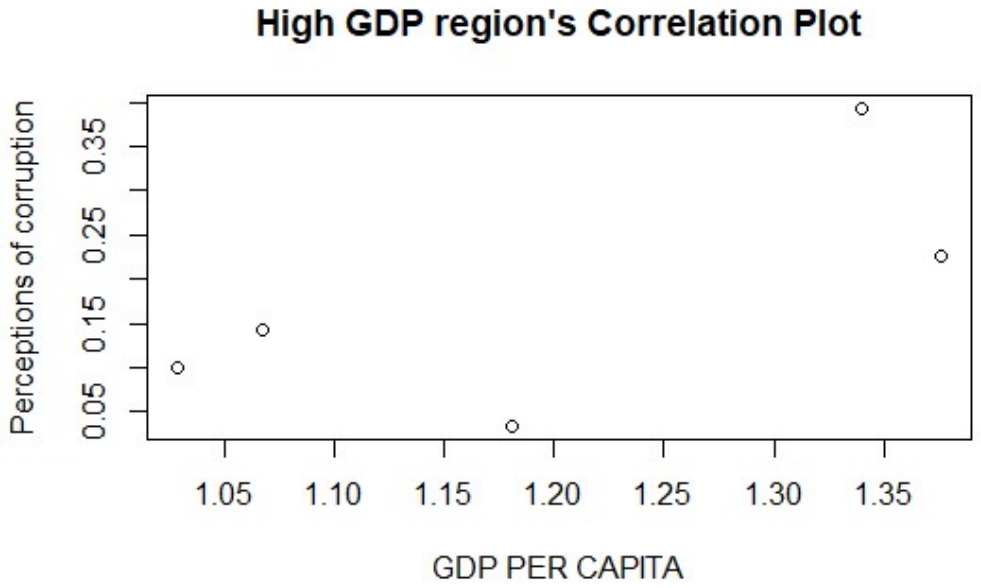
Method

Correlation

The correlation matrix is a matrix that shows the correlation between variables. it gives the correlation between all the possible pairs of values in a matrix format. We can use a correlation matrix to summarize a large dataset and to identify the patterns and make a decision accordingly.

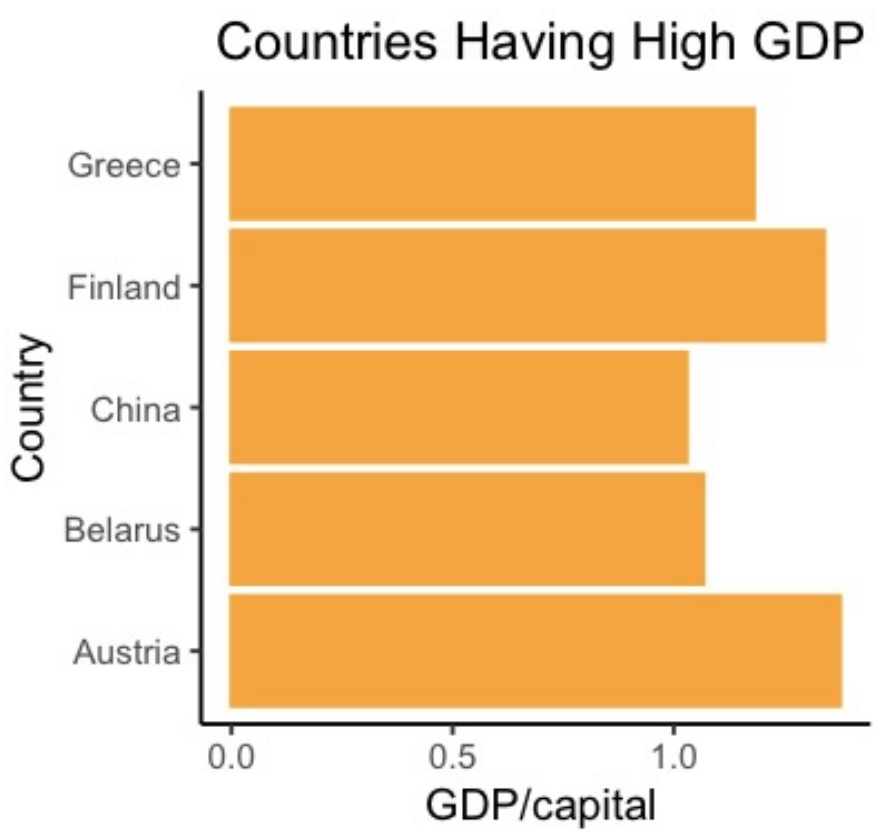
EDA : Question-2





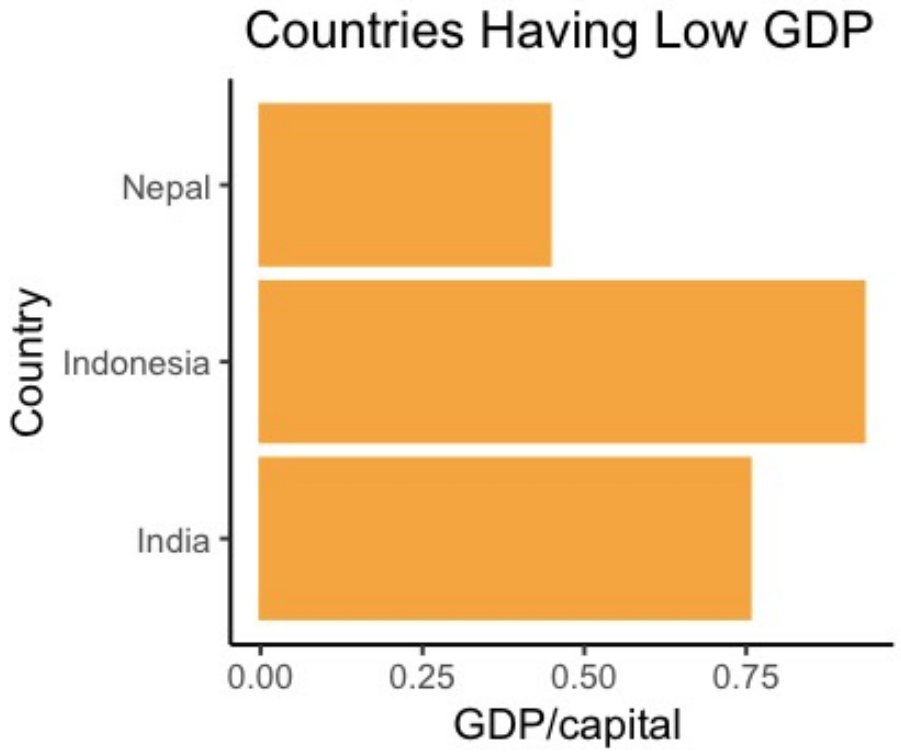
- A negative correlation is a relationship between two variables that move in opposite directions.
- That means the corruption increases and the GDP decreases and vice versa (in regions with low GDP).

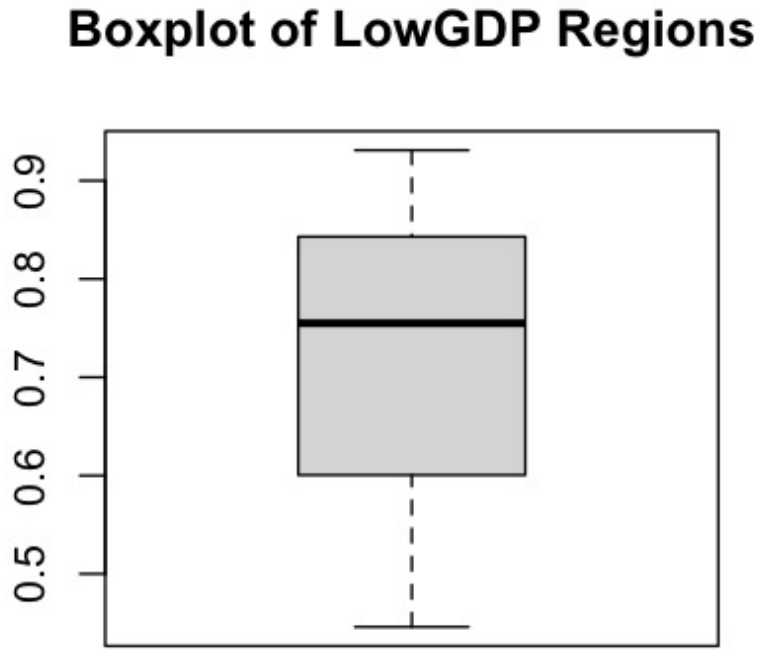
- Here there is a stronger correlation of between the High GDP region and perception of corruption



- The graph depicts the countries with the highest GDP.
- The output is derived from the constraint that these filtered counties in the dataset have a GDP greater than one.

- As can be seen, the countries with the highest GDP scores exceed 1.35, while the median values are about 1.7.





- The graph depicts the countries with the lowest GDP.
- The noteworthy change is that all of the countries with low GDP are in Asia.

- The boxplot of LowGDP locations shows that all GDP scores are less than 1, with the median being somewhere around 0.8.

INTERPRETATION

High GDP Regions

Country	GDP per Capital
Finland	1.340
Austria	1.376
Belarus	1.067
Greece	1.181
China	1.029
Correlation	0.6860

Low GDP Regions

Country	GDP per Capital
Indonesia	0.931
Nepal	0.446
India	0.755
Correlation	-0.8127

- When two variables have a negative correlation, **one variable increases while the other falls**, and vice versa. This link may or may not indicate causation between the two variables, but it does reflect a pattern that may be observed.

❑ **So, from the above happiness plots and the GDP plots we can conclude that the European region have higher GDP and they are having higher Happiness score. Moreover, the Asian regions have the low happiness score as well as the low GDP.**

CONCLUSION

We noticed that changing one variable affects the other by utilising the lm function in a happiness comparison question.

As a result, we discovered that European regions have a higher happiness score than Asian ones.

Similarly, when comparing high and low GDP regions, we determined that there is a negative link between LowGDP regions.

As a result, we may conclude that the European region has a greater GDP and a higher Happiness score. Furthermore, Asian regions have both a low happiness score and a low GDP.

BIBLIOGRAPHY

1. Frost, Jim. “How to Interpret P-Values and Coefficients in Regression Analysis - Statistics By Jim.” *Statistics By Jim*, statisticsbyjim.com, 12 Apr. 2017, https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression

2. “Visualize Correlation Matrix Using Correlogram - Easy Guides - Wiki - STHDA.” *Visualize Correlation Matrix Using Correlogram - Easy Guides - Wiki - STHDA*, www.sthda.com, http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram. Accessed 16 May 2022.

3. Holtz, Yan. “The R Graph Gallery – Help and Inspiration for R Charts.” *The R Graph Gallery*, r-graph-gallery.com, https://r-graph-gallery.com/. Accessed 16 May 2022

4. *Negative Correlation Explained*. (2021, September 21). Investopedia; www.investopedia.com. https://www.investopedia.com/ask/answers/040815/how-should-i-interpret-negative-correlation.asp