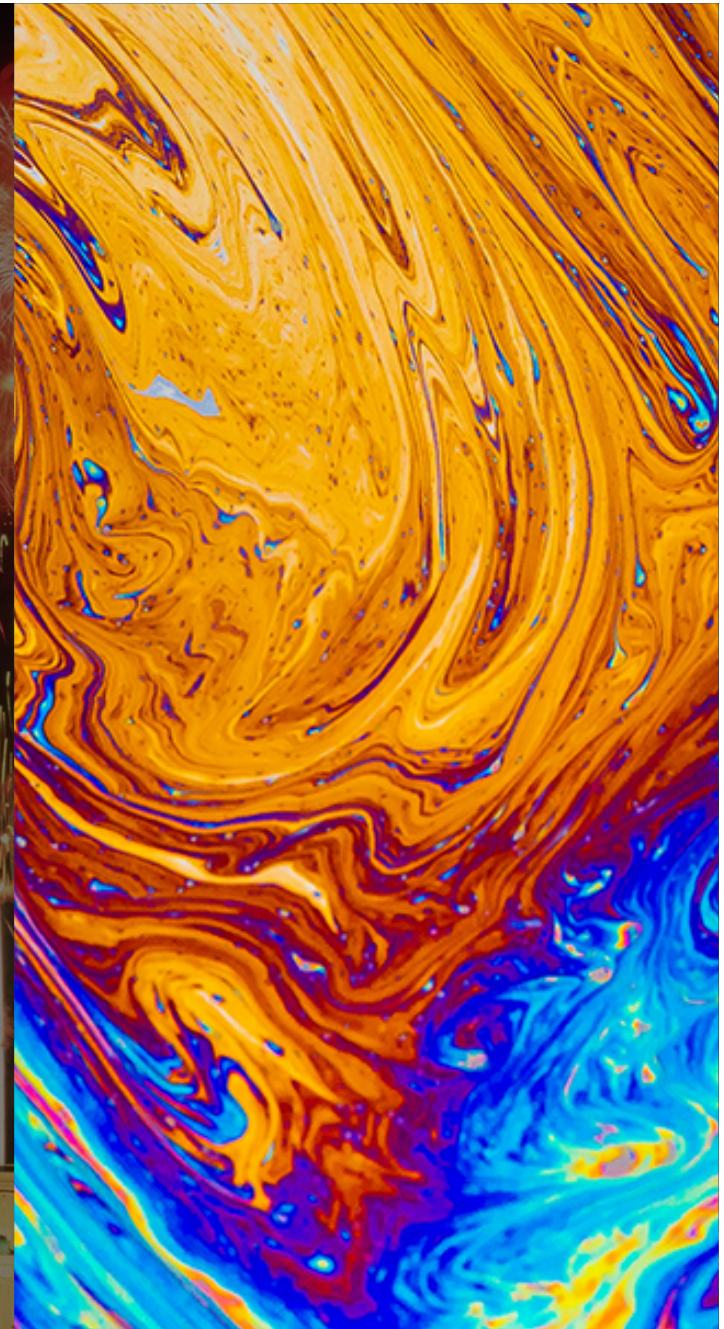
Report for Happiness around the world

A story of 2015-2022 Happiness reports



Group 07 ETC5513

Department of Econometrics and Business Statistics

zyan0056@student.monash.edu

10th April 2022

Background

Abstract and Data Source

Motivation

Do you know what contributes to happiness? Our motivation of analyzing the world happiness index is to determine relevant factors and then try to improve countries' happiness scores based on our results

Data Source

Our data is the 2015-2022 world happiness report obtained from the **Kaggle** website. It contains factors that are related to the happiness scores in each country at a world wide level.

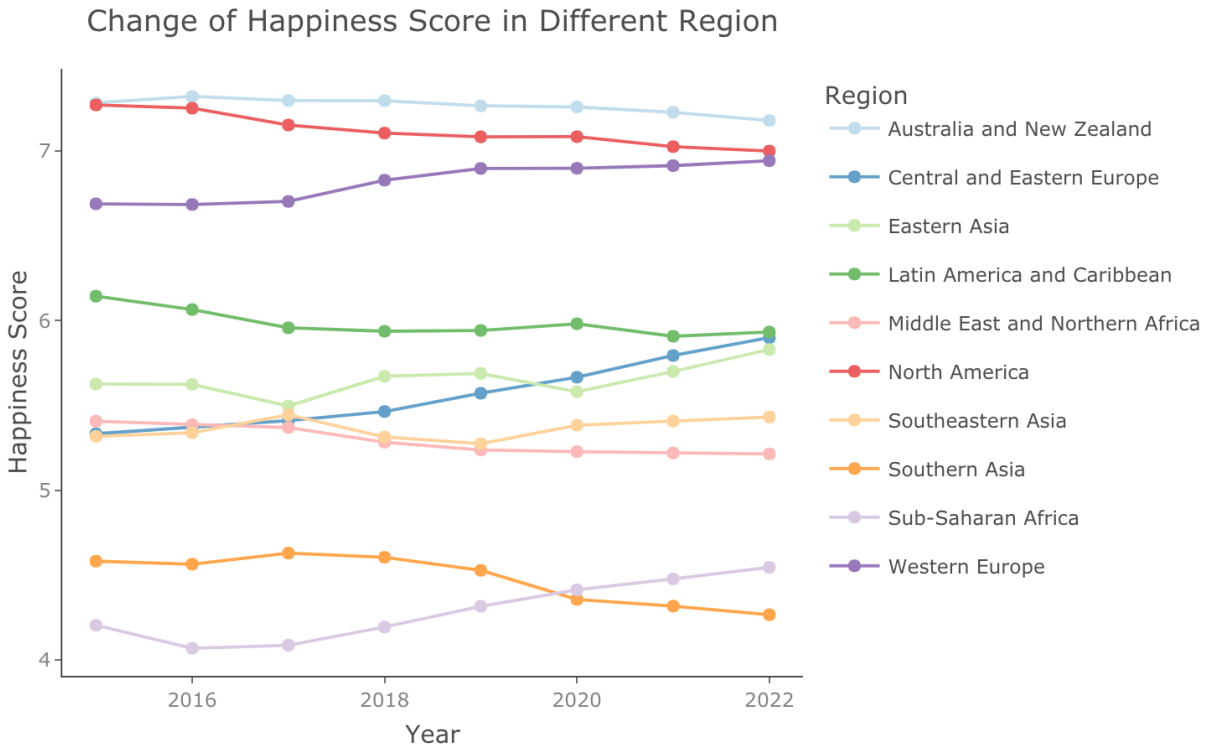
Research Aim

Our research aim to solve the following questions

- The influences of COVID-19 on the world happiness score and the correlations between happiness in 2021.
- The changes of happiness between 2015 and 2022 in different regions.
- The impact of economic situation and the health status on happiness.
- The important variables in explaining happiness scores via different models and discover their marginal effects in a linear model.

Exploratory Data Analysis PART I

How will happiness trends change between 2015 and 2022 in different regions?



🧐**What is the relationship between economic situation and health status with the happiness?**

Exploratory Data Analysis

From 2015 to 2022, What is the relationship between economic situation and health status with the happiness?

Exploratory Data Analysis

From 2015 to 2022, What is the relationship between economic situation and health status with the happiness?

Year <chr>	(Intercept) <dbl>	Economy <dbl>	Health <dbl>
2015	3.250	1.616	1.203
2016	2.980	1.516	1.662
2017	2.968	1.504	1.632
2018	3.085	1.511	1.565
2019	2.892	1.359	1.775
2020	3.117	1.305	1.791
2021	3.298	1.316	1.854
2022	2.370	1.406	2.085

🧐**What we can see?**

📊 Based on the figure above and previous analysis. We can conclude that there is a positive correlation for both economic situation and health status on Happiness score between 2015 and 2022. The idea behind this is the better economic situation and health status people have, they will feel more happy.

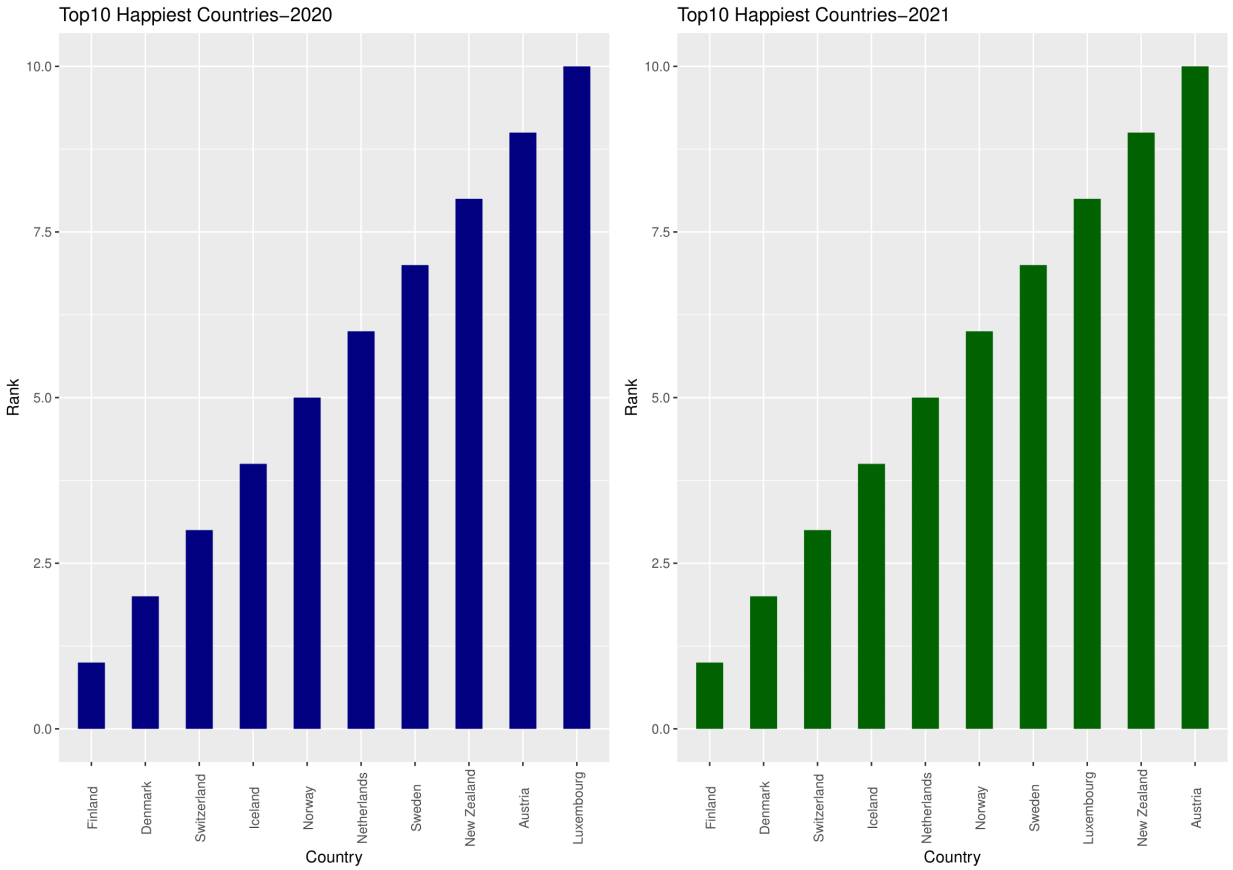
Exploratory Data Analysis Part II

Research Question

- The top 10 countries in happiness score since COVID-19.
- The distribution of the top 10 countries on the world map in 2021.
- Relation between happiness score and other 6 attributes in 2021.

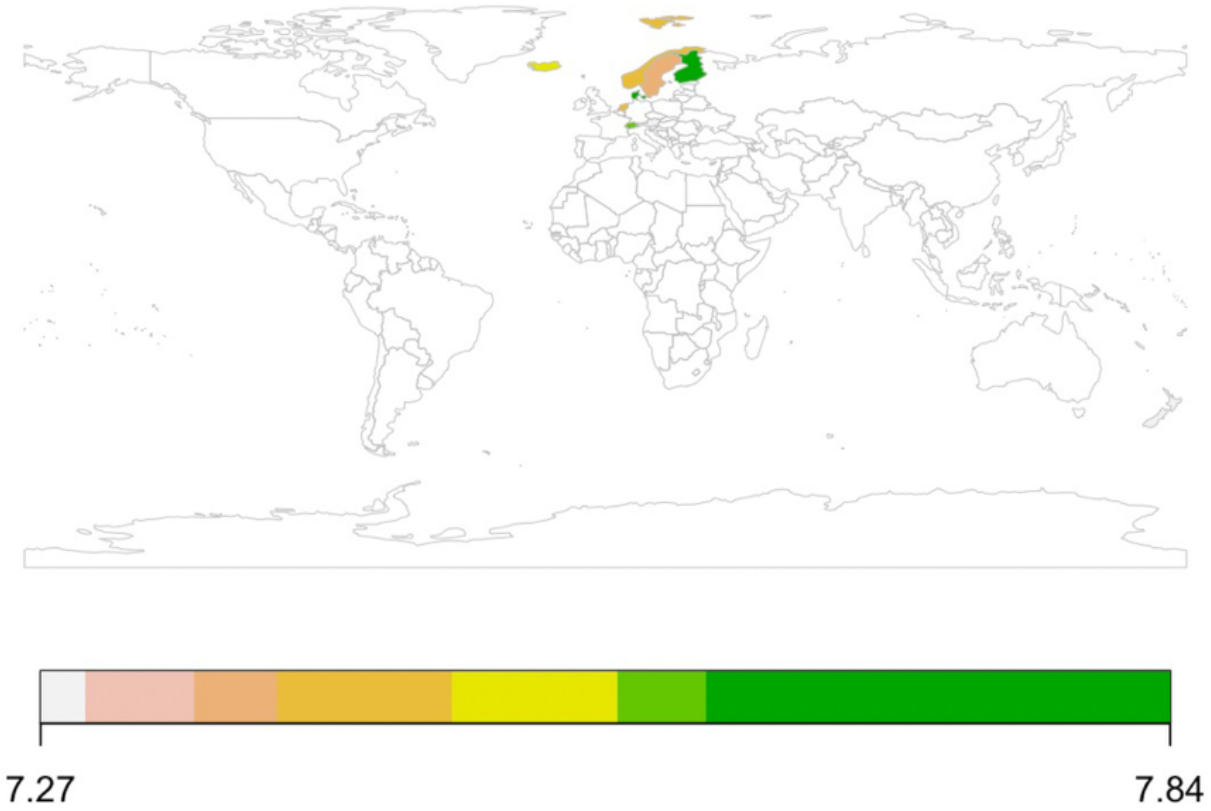
Exploratory Data Analysis

The top 10 countries in happiness score since COVID-19



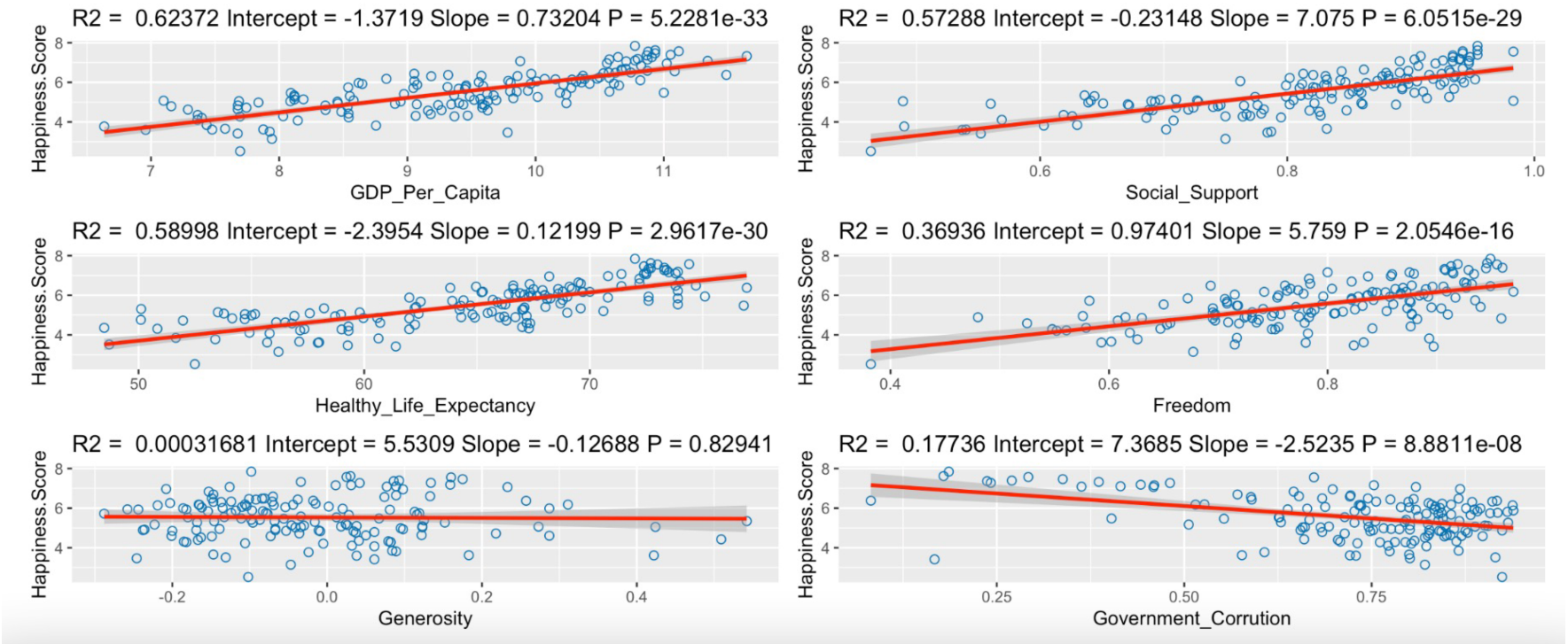
Exploratory Data Analysis

The distribution of top 10 countries on the world map in 2021



Exploratory Data Analysis

Relation between Happiness score and other indicators in 2021



Modelling

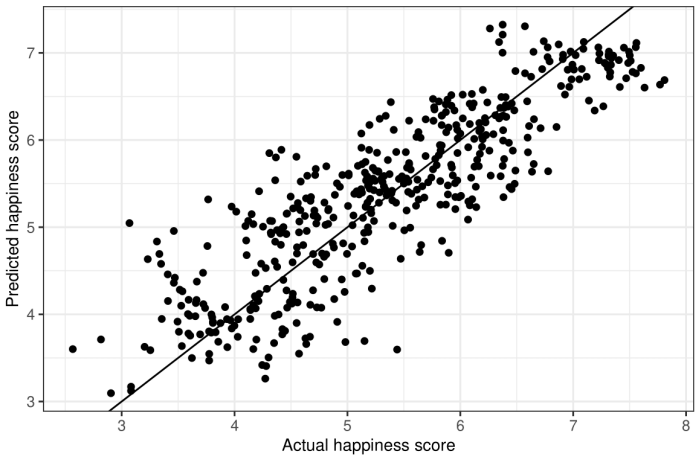
We will investigate the important variables in explaining the happiness scores via fitting various models including:

- **Multivariate Linear Model** : Simple Linear regression with multiple variables.
- **Support Vector Machine Model**: Supervised Machine learning algorithm to fit regressions.
- **Decision Tree Model**: Binary tree model have control statement.
- **Random Forest Model**: Use multiple learning algorithms (resampling and tree) to give us better results.

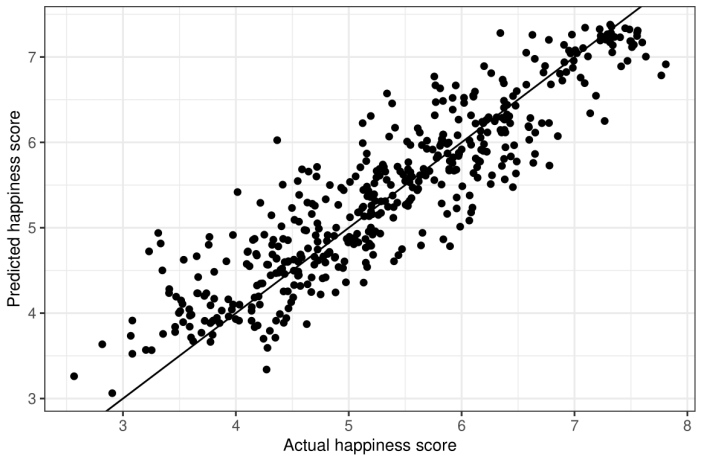
Modelling (Training)

Training and Test Split Ratio = 0.4

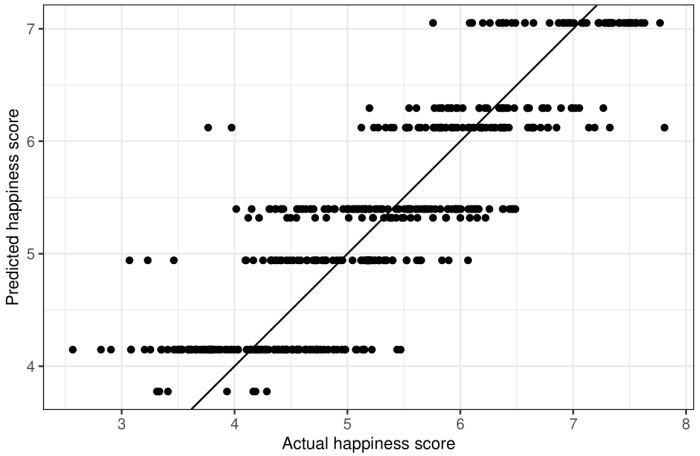
Multiple Linear Regression



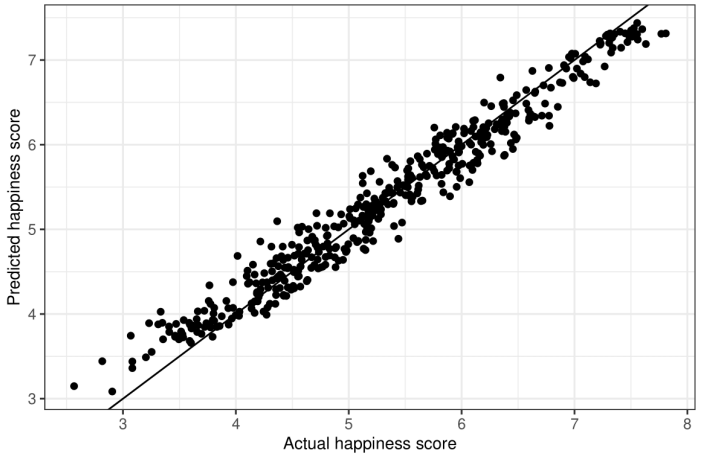
Support Vector Machine



Decision Tree Regression



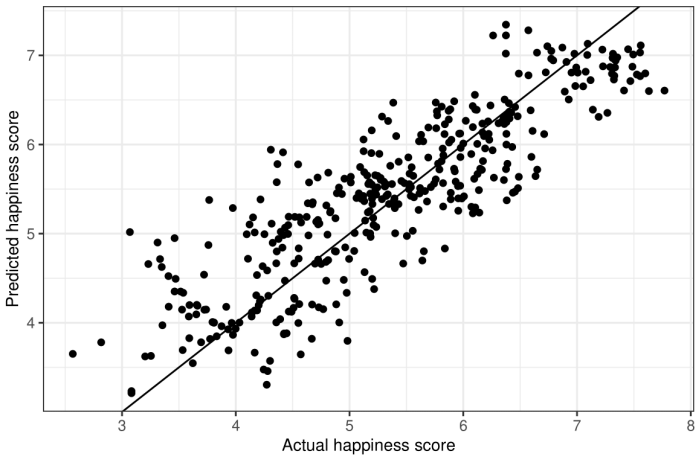
Random Forest Regression



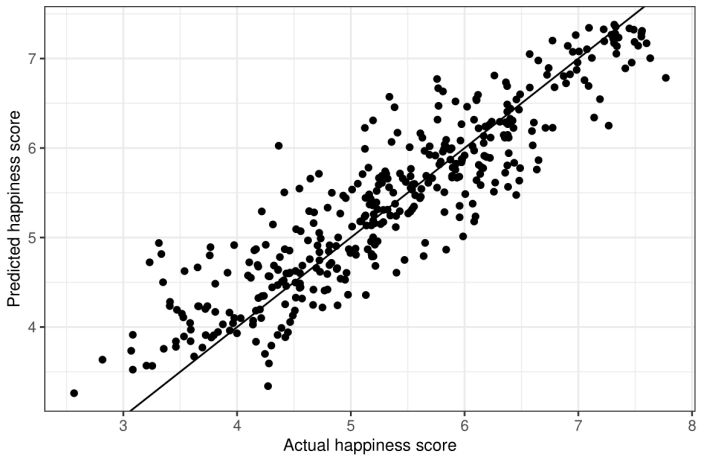
Modelling (Training)

Training and Test Split Ratio = 0.5

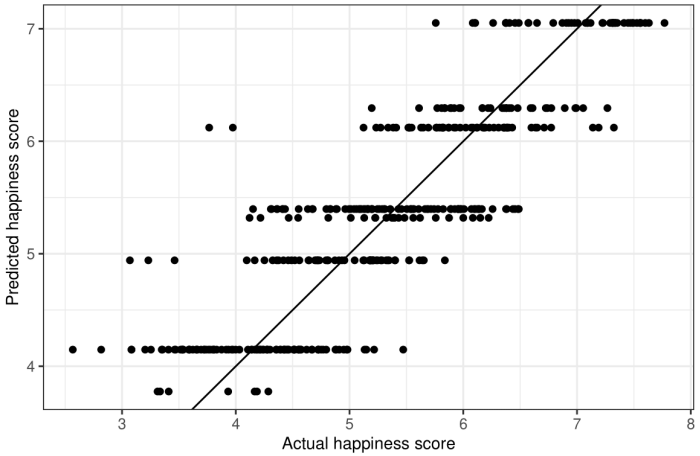
Multiple Linear Regression



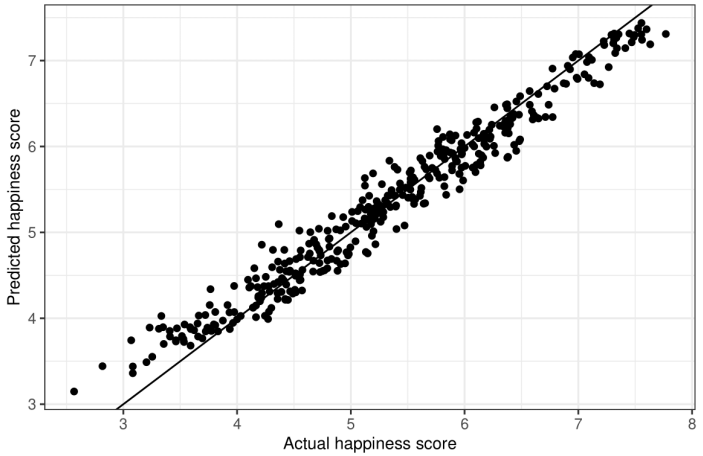
Support Vector Machine



Decision Tree Regression



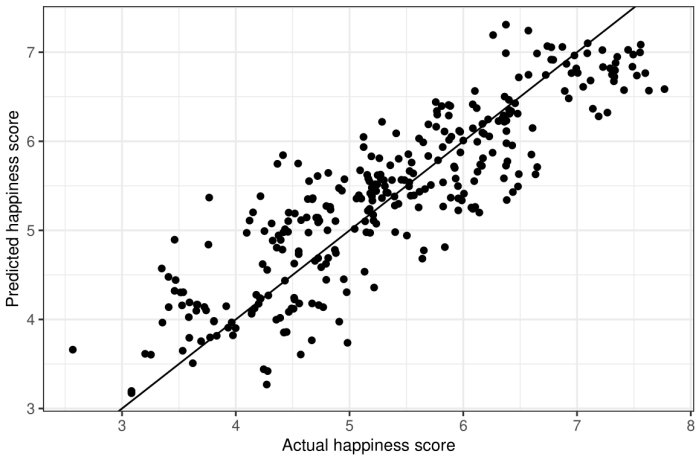
Random Forest Regression



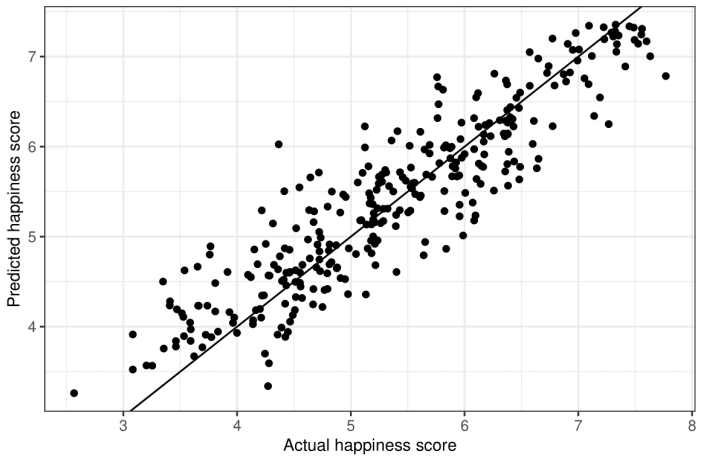
Modelling (Training)

Training and Test Split Ratio = 0.6

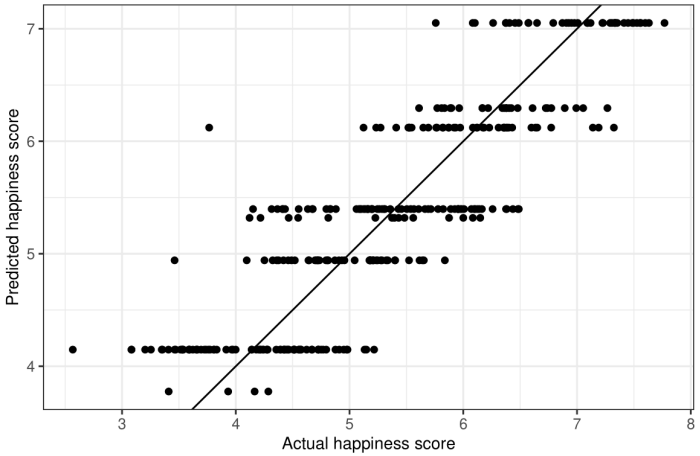
Multiple Linear Regression



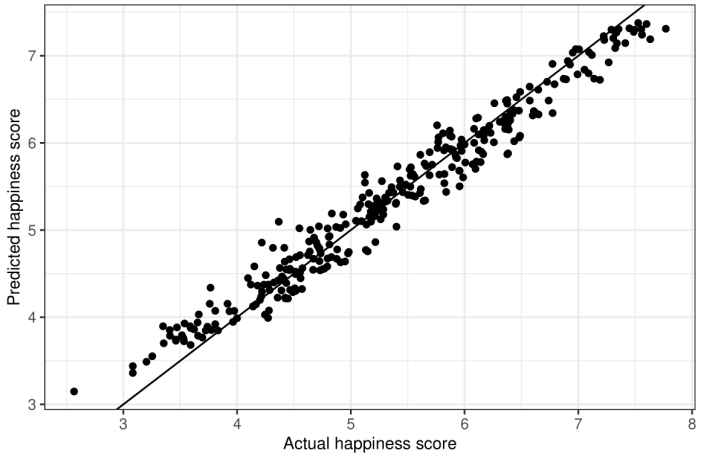
Support Vector Machine



Decision Tree Regression



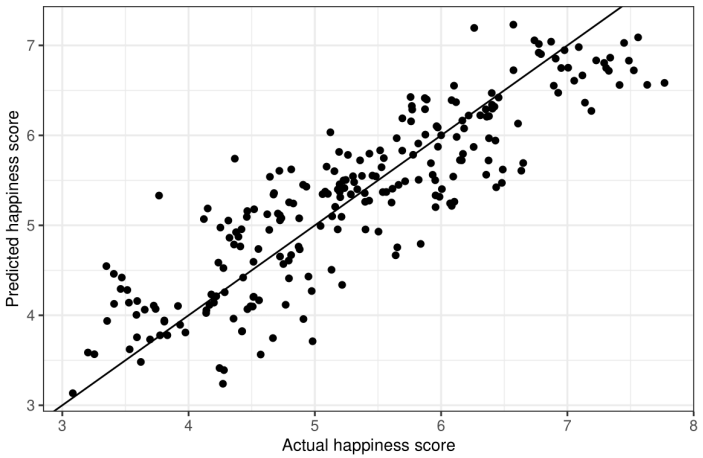
Random Forest Regression



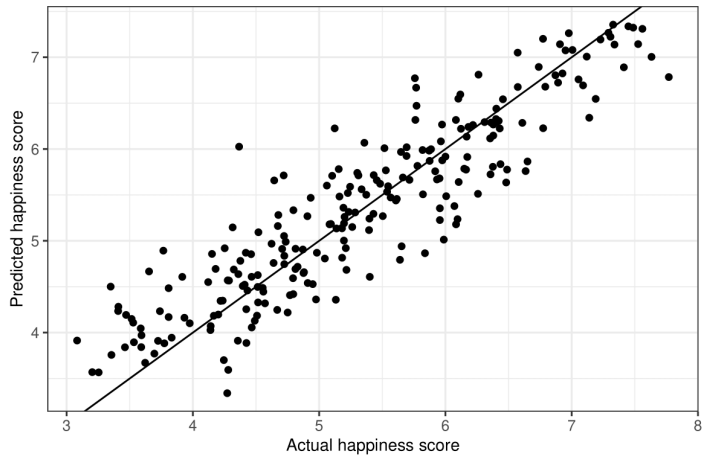
Modelling (Training)

Training and Test Split Ratio = 0.7

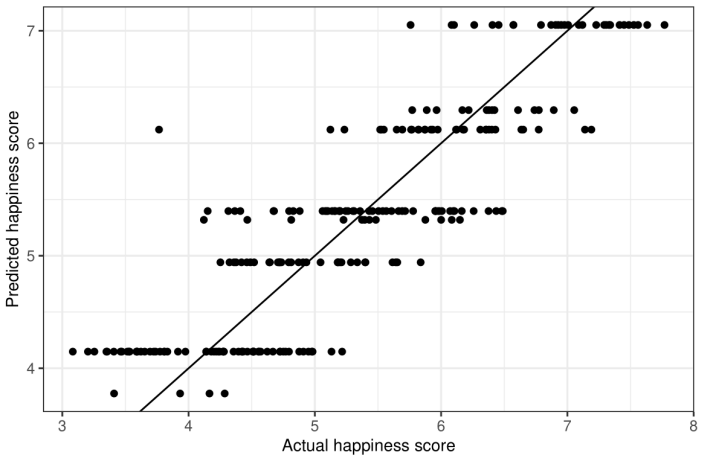
Multiple Linear Regression



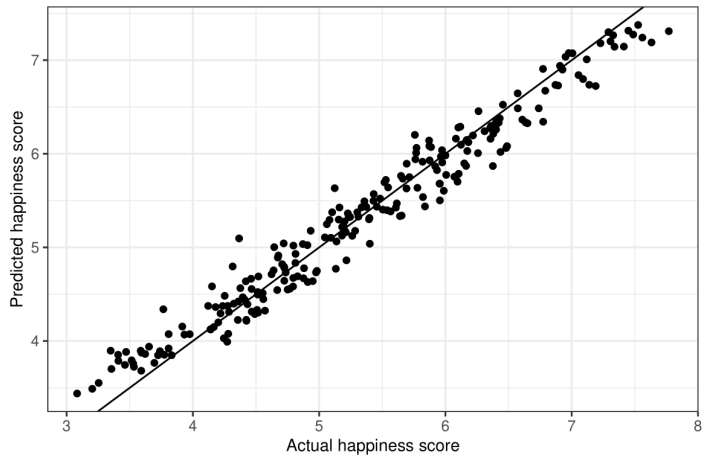
Support Vector Machine



Decision Tree Regression



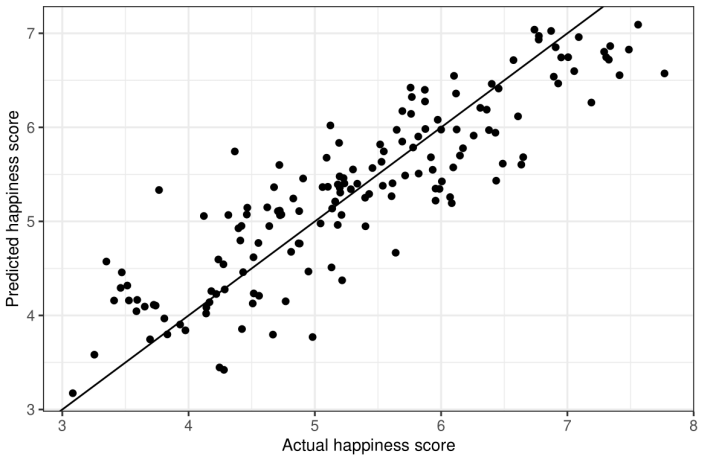
Random Forest Regression



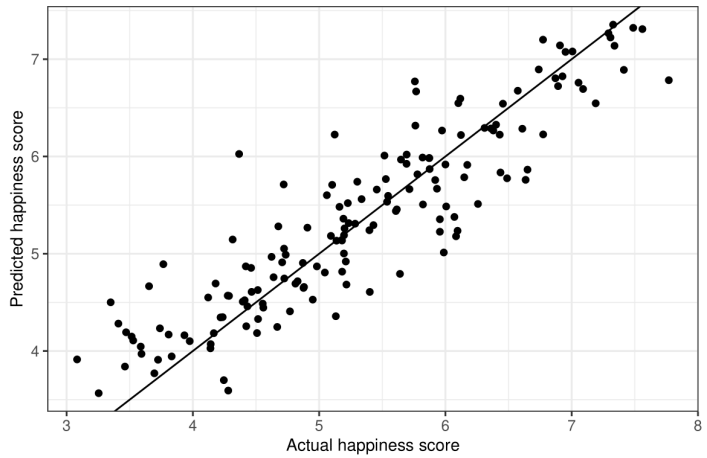
Modelling (Training)

Training and Test Split Ratio = 0.8

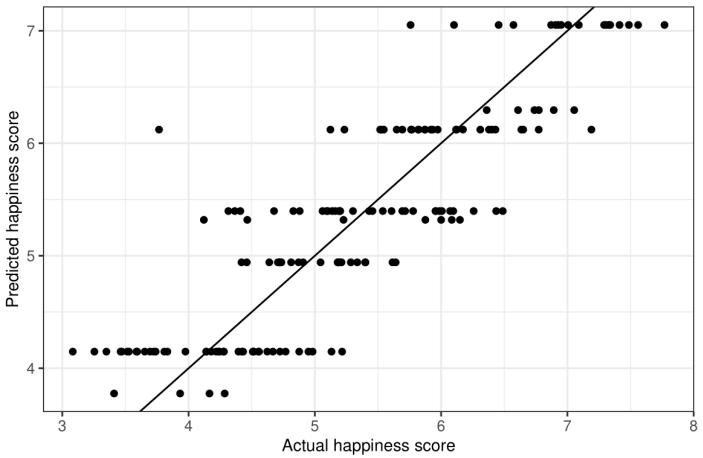
Multiple Linear Regression



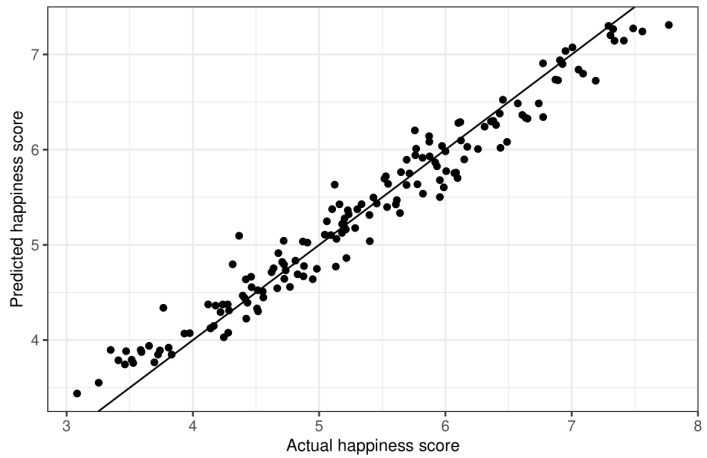
Support Vector Machine



Decision Tree Regression



Random Forest Regression



Modelling

🥳**Overall we can see the Random Forest model performs the best of fitting this data. Then we can sort our variables by their importance.**

Variable importance order

The right table indicates that the most important variables here are the economy (in GDP per capita) and Health.

Limitation: The random forest model cannot provides a better scope on the relationships of variables.

Variable importance for Random Forest model

	IncNodePurity
Economy	361.79333
Health	319.73556
Freedom	168.45474
Generosity	94.87697

Little Improvement

To better discover the happiness score, two new variables will be added🥳.

- We cannot tell the order of importance because our target is to improve the happiness index efficiently.
- The feeling of Freedom and Generosity is too abstract.
- People have different standards towards to them.
- We build up the model with the Economy, Health, CPI(Consumer Price Index), year, and Population.

Linear regression model for happiness scores without new data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.437442	0.0752984	32.370444	0
Health	1.223204	0.1392052	8.787054	0
Economy	1.454316	0.0841520	17.282015	0
Freedom	1.372180	0.1095802	12.522156	0
Generosity	1.170207	0.1396671	8.378545	0

Multivariate Linear model

To analyse the relationships between these variables, we constructed a Multivariate Linear Model in 2016 to 2020.

$$log(score) = -4.6000 - 0.0008\ cpi + 0.2591\ log(economy)$$
$$-0.0026\ log(population) + 0.0114\ log(health) + 0.0032\ year$$

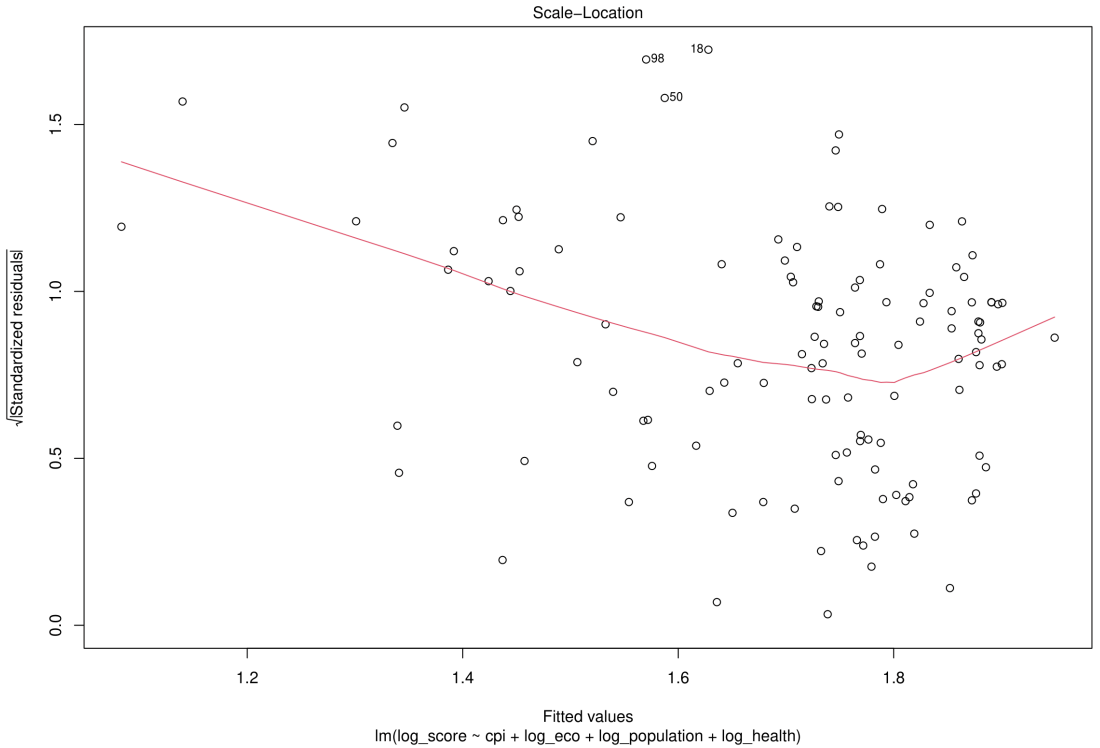
We then take the logarithm to analyse the percentage change in these variables and eliminate the heteroscedasticity. Then all the coefficients are the percentage change ratio can be written $\frac{\%\Delta Happiness}{\%\Delta Variables}$.

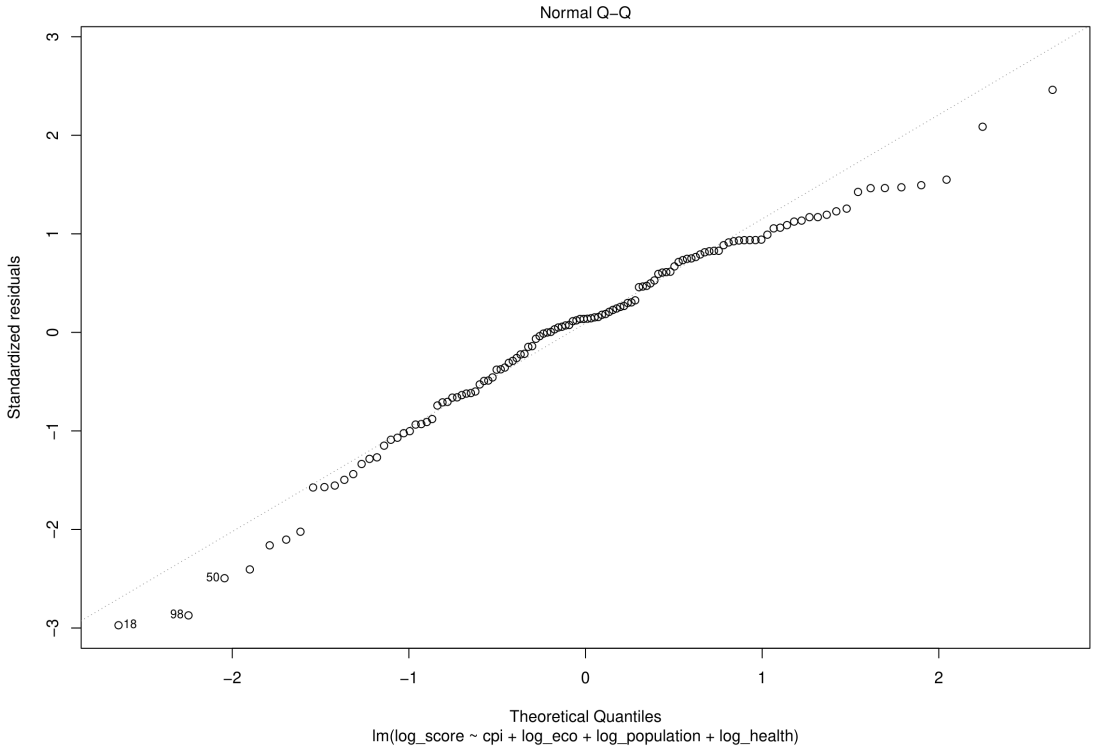
We can see over 2016 to 2020 , a one percent increase in economy will increase the percentage change of happiness score by 25.91%.

We can also see that the percentage change of health status will also lead a positive increase in happiness score. However, percentage changes on both Consumer Price index and population will lead to a negative percentage change in happiness score.

Multivariate Linear Model

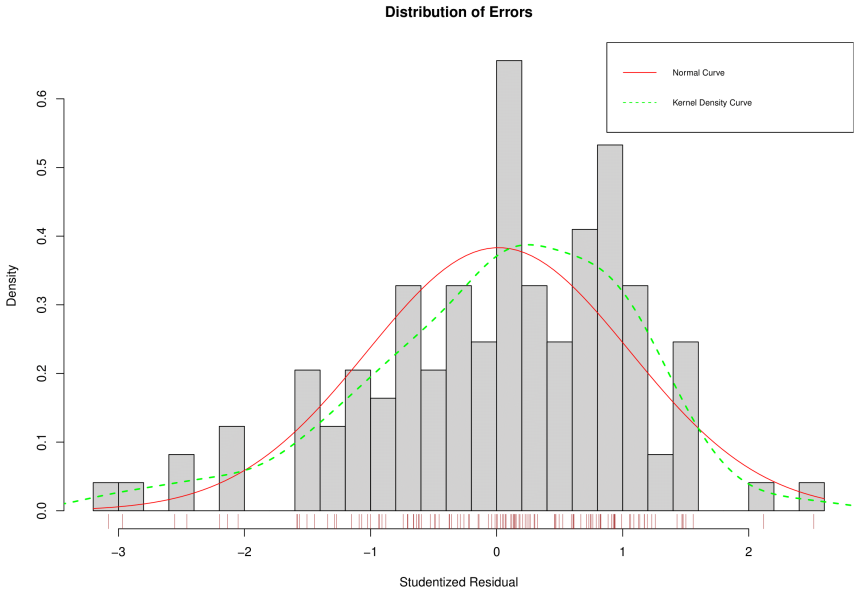
Analysis of Residuals

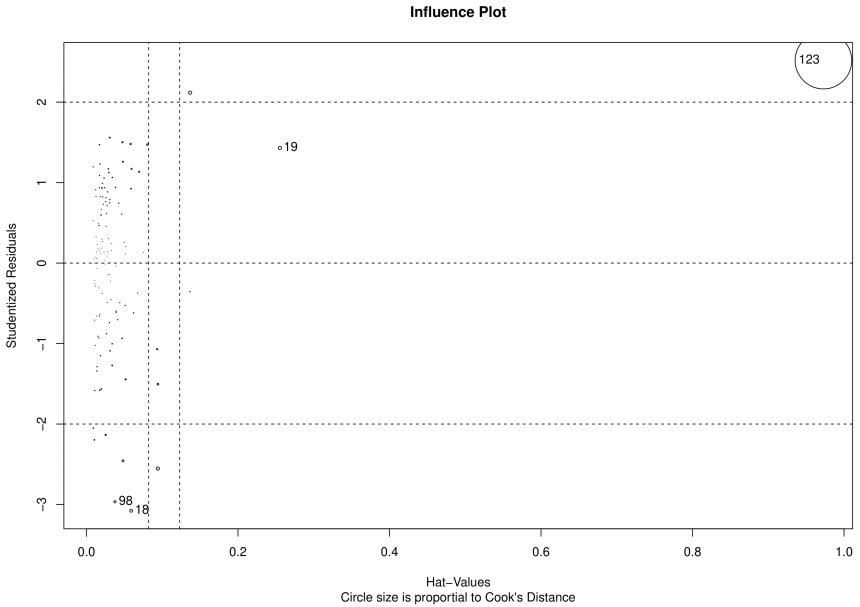




Multivariate Linear Model

Analysis of Residuals





Multivariate Linear Model

👀 Insights on Our Multivariate Linear Model Model

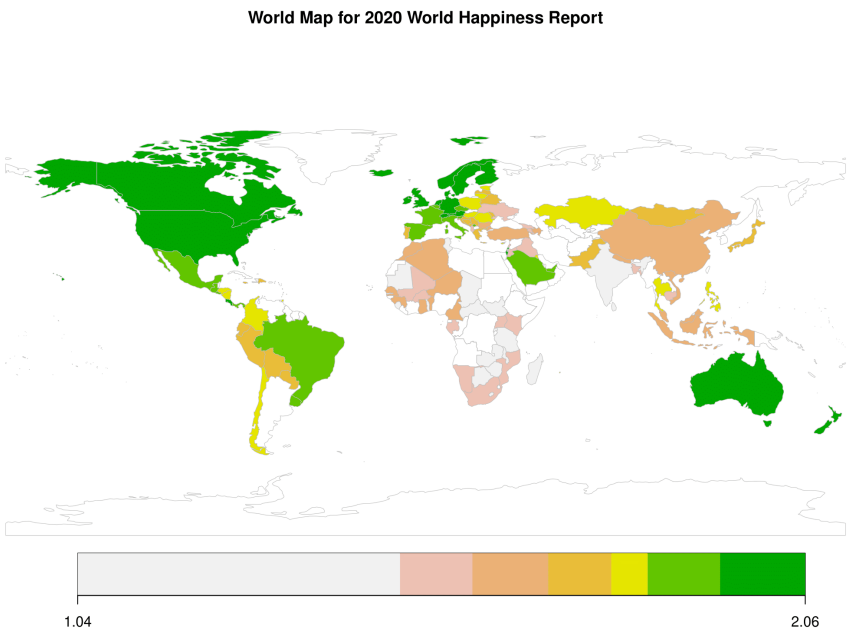
Limitations of our analysis:

Endogenous Bias: Another latent variable in the error term still correlate with the variables inside the model, which influence our model's accuracy.

Sample Selection Bias:

1. We only use the most recent data available (up to year 2020 data).
2. The structure of happiness ranks may also change after the pandemic.
3. NAs may lead to truncation bias.

Significant Outliers: From the analysis before we can get that there are few significant outliers may also influence the modelling accuracy too.

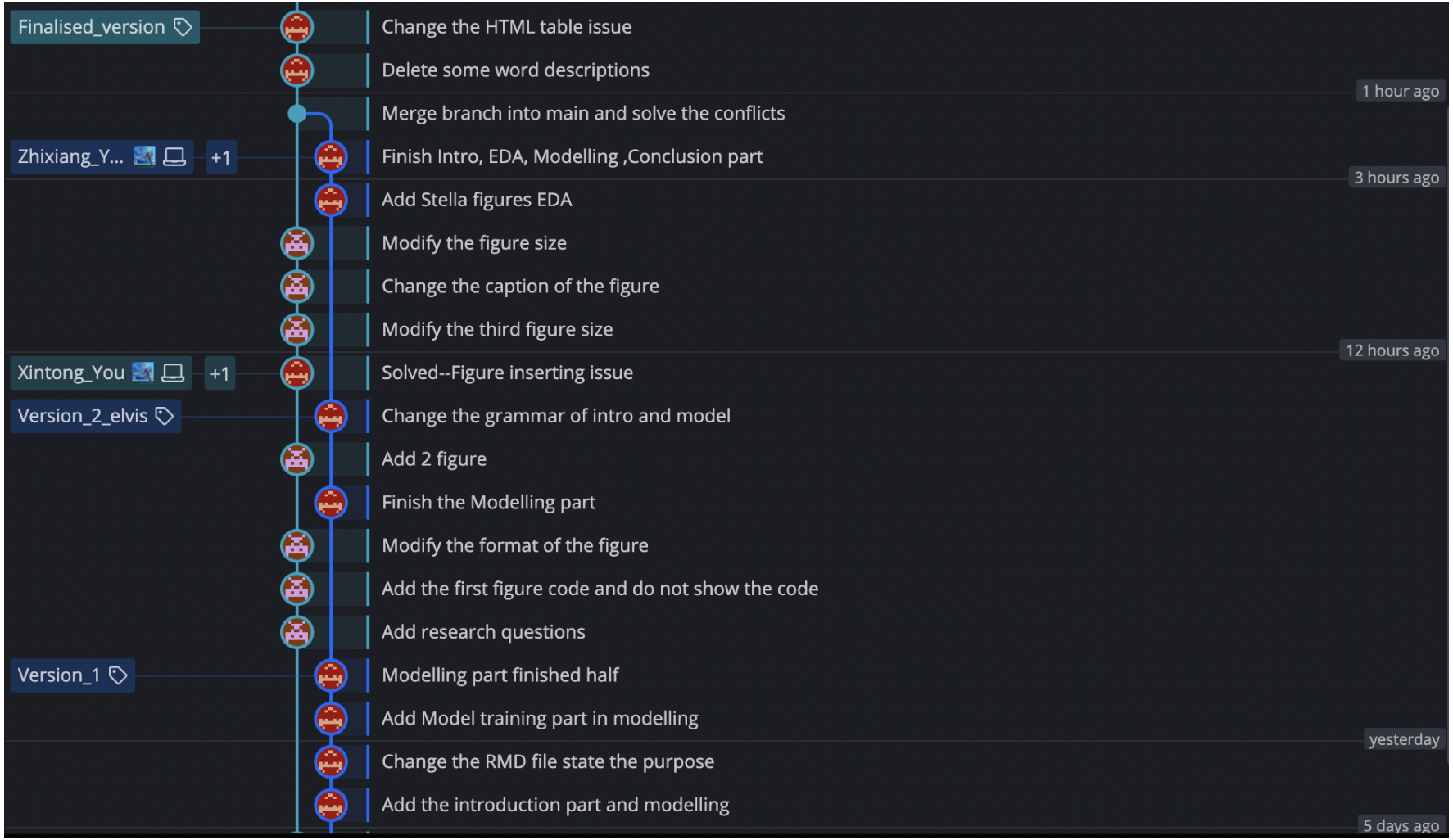


Conclusion

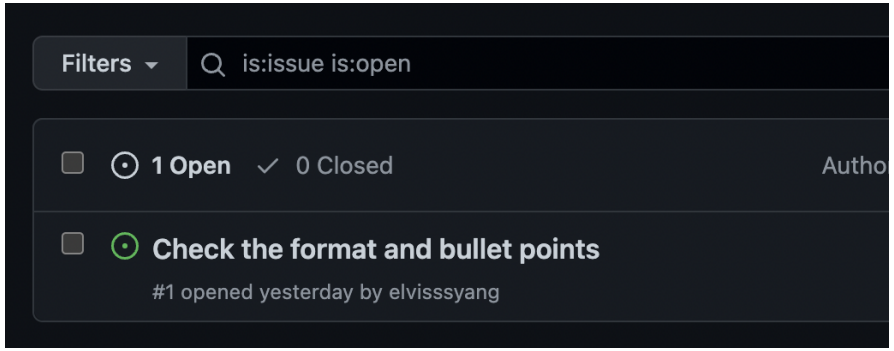
Based on above analysis, we explored few ideas behind the World Happiness Report from 2015 to 2022. Then we can conclude:

- Since the outbreak of COVID-19, the countries with highest happiness score are mainly from Northern and Western Europe. Finland has been the happiest country during the COVID-19 period.
- People living in better economic situations often feel happier and their scores don't change much than those who live in rural areas. Moreover, health status will also help to increase the happiness, no matter where they live.
- The most important variables to explain the happiness score are the economy situation (measured in GDP per capita) and health statues. Moreover, in our Linear Model, they both contribute an positive elasticity to the happiness score. Nevertheless, we discover that two relevant variables (Population, Consumer Price Index) lead to a negative elasticity to the happiness score.

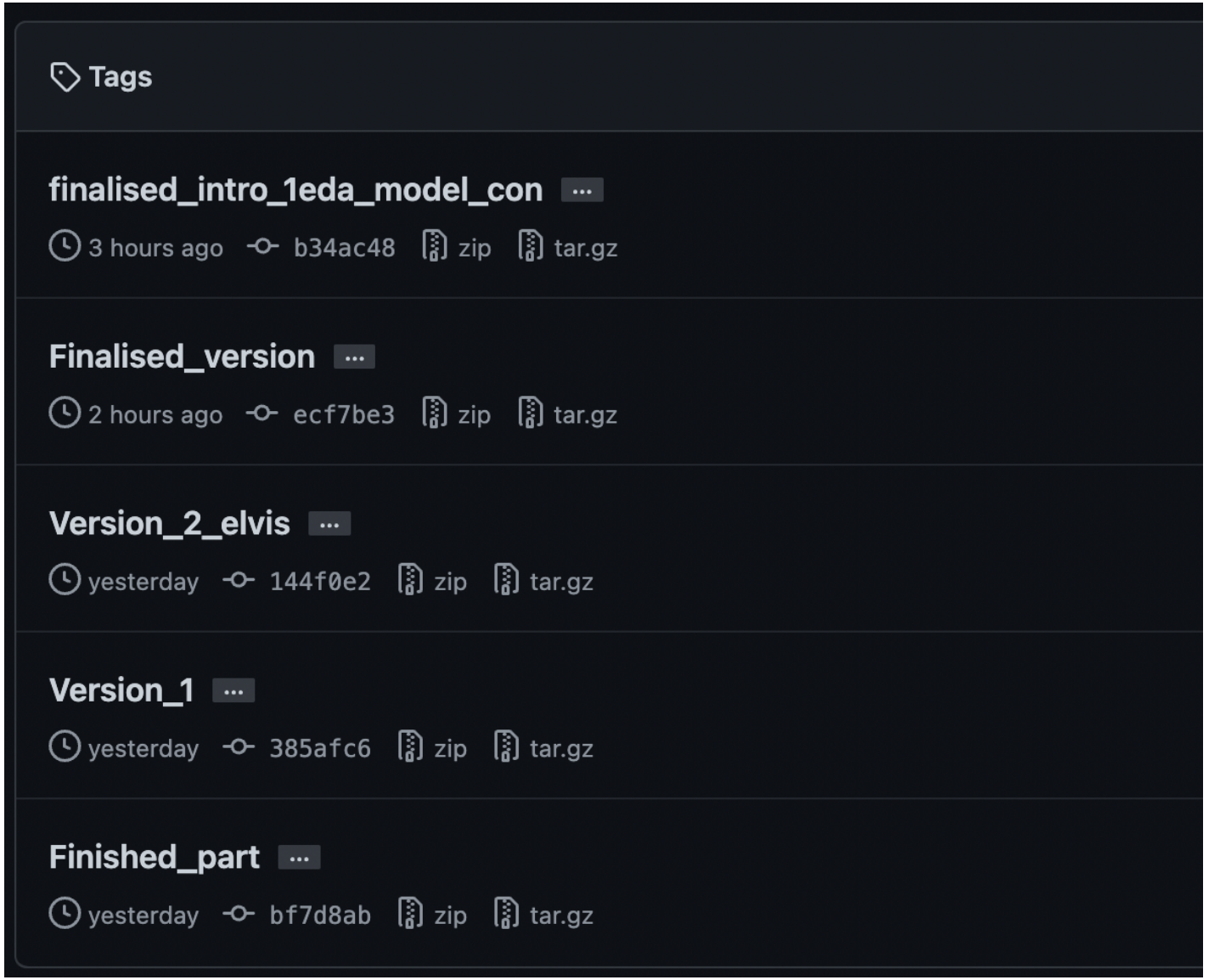
Collaboration through git

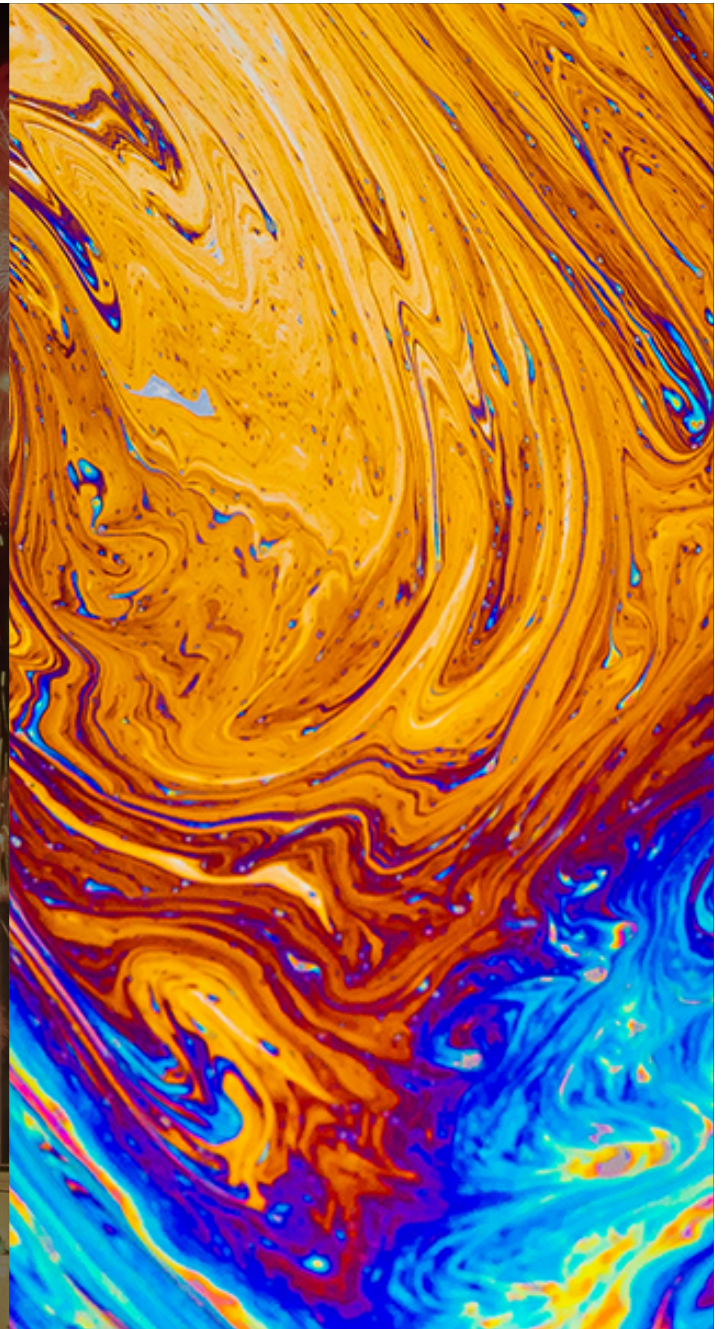


Collaboration through git



Collaboration through git





Acknowledgements

Slides produced using Rmarkdown with xaringan styling.

Thanks for Listening
Any Questions?

Author: Group 07 ETC5513

Department of Econometrics and Business Statistics

zyan0056@student.monash.edu