



# World Happiness Report

## Mini Project

CZ1015 Introduction to Data Science and Artificial Intelligence

Gupta Jay • Nguyen Duy Khanh • Tieu Phat Dat



# Problem Statements

Data Exploration &  
Visualisation



What are the top  
factors which affect  
happiness?

Are there any wrong  
perceived factors about  
happiness?

Classification &  
Regression

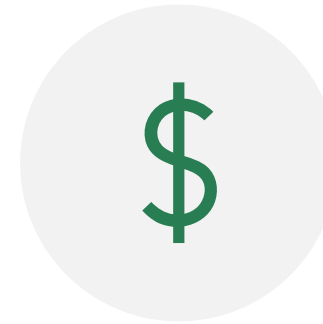


Create your own  
Country

Predict it's Happiness  
Category

Find the most similar  
Country as yours

Data Exploration &  
Visualisation



Economic Analysis  
of Singapore over  
the Years

Democratic Quality

Social Support

Life Satisfaction

# Exploratory Data Analysis



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 15 columns):
country                                156 non-null object
Region indicator                       156 non-null object
Life ladder, 2015-2017                 156 non-null float64
Standard error, life ladder, 2015-2017 156 non-null float64
Log of GDP per person, 2015-2017       152 non-null float64
GDP per person, 2015-2017              152 non-null float64
Healthy life expectancy, 2015-2017     153 non-null float64
Social support, 2015-2017              155 non-null float64
Standard error, social support, 2015-2017 155 non-null float64
Freedom to make life choices, 2015-2017 155 non-null float64
Standard error, freedom to make life choices, 2015-2017 155 non-null float64
Generosity, 2015-2017, without adjustment for GDP per person 155 non-null float64
Standard error, generosity, 2015-2017  155 non-null float64
Perceptions of corruption, 2015-2017   148 non-null float64
Standard error, perceptions of corruption, 2015-2017 148 non-null float64
dtypes: float64(13), object(2)
memory usage: 18.4+ KB
```

- The dataset had **different number** of non-null observations for various columns.
- The Column names have a hard readability due to long names, inconsistent formatting (Use of special characters (,-))

# Data Cleaning

```
In [73]: # Sort the dataset by Countries instead of Happiness Score
data1.sort_values(by = ["Country"], inplace = True, ascending = True)
data1 = data1.reset_index(drop=True)
data1 = pd.DataFrame(data1[:156])
data = pd.concat([data1,data2],axis=1)

# Convert all Variable Names to UPPERCASE
data.columns = data.columns.str.upper()

# Remove all spaces and dots from Variable Names
data.columns = data.columns.str.replace(".", "")
data.columns = data.columns.str.replace(" ", "_")
data.columns = data.columns.str.replace(":", "_")
data.columns = data.columns.str.replace("+", "_")
data.columns = data.columns.str.replace("-", "_")
```

Sort the dataset by Countries  
& Join the two datasets

Change to UPPERCASE

```
In [74]: # Extracting the required variables from the dataframe
extracts = ["COUNTRY", "HAPPINESS_SCORE", "LOG_OF_GDP_PER_PERSON_2015-2017", "GDP_PER_PERSON_2015-2017", "HEALTHY_LIFE_EXPECTANCY",
            "SOCIAL_SUPPORT_2015-2017", "FREEDOM_TO_MAKE_LIFE_CHOICES_2015-2017", \
            "GENEROSITY_2015-2017_WITHOUT_ADJUSTMENT_FOR_GDP_PER_PERSON", "PERCEPTIONS_OF_CORRUPTION_2015-2017"]

data = pd.DataFrame(data[extracts])
```

Renaming and Reformatting the  
Dataset for easier readability and  
Data processing

```
In [75]: # Rename the columns of the dataframe for easier readability
data.rename(columns = {'GDP_PER_PERSON_2015-2017': 'GDP_PER_PERSON'}, inplace = True)
data.rename(columns = {'LOG_OF_GDP_PER_PERSON_2015-2017': 'LOG_OF_GDP_PER_PERSON'}, inplace = True)
data.rename(columns = {'FREEDOM_TO_MAKE_LIFE_CHOICES_2015-2017': 'FREEDOM'}, inplace = True)
data.rename(columns = {'HEALTHY_LIFE_EXPECTANCY_2015-2017': 'HEALTHY_LIFE_EXPECTANCY'}, inplace = True)
data.rename(columns = {'SOCIAL_SUPPORT_2015-2017': 'SOCIAL_SUPPORT'}, inplace = True)
data.rename(columns = {'GENEROSITY_2015-2017_WITHOUT_ADJUSTMENT_FOR_GDP_PER_PERSON': 'GENEROSITY'}, inplace = True)
data.rename(columns = {'PERCEPTIONS_OF_CORRUPTION_2015-2017': 'PERCEPTIONS_OF_CORRUPTION'}, inplace = True)
```

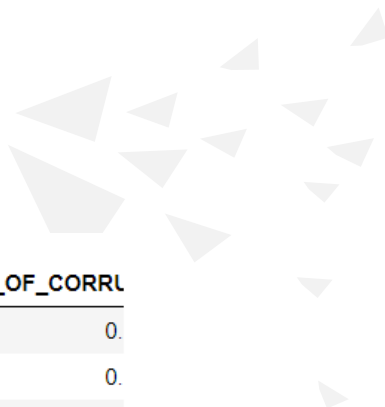
```
In [78]: # Remove rows with missing values
data = data.dropna(how='any',axis=0)

# After removing missing values, reset the index
data = data.reset_index(drop=True)
```

Remove NaN Values

# Data Cleaning

## Project Dataframe



Out[9]:

	COUNTRY	HAPPINESS_SCORE	GDP_PER_PERSON	HEALTHY_LIFE_EXPECTANCY	SOCIAL_SUPPORT	FREEDOM	GENEROSITY	PERCEPTIONS_OF_CORRU
0	Afghanistan	3.632	1741.687500	52.013329	0.525075	0.445294	0.179054	0.
1	Albania	4.586	11363.095700	68.871552	0.639576	0.726340	0.259975	0.
2	Algeria	5.295	13914.723630	65.604858	0.776977	0.439177	0.128988	0.
3	Angola	3.795	6260.132813	52.460709	0.765275	0.374173	0.106829	0.
4	Argentina	6.388	18807.310550	67.398483	0.905565	0.853390	0.163174	0.

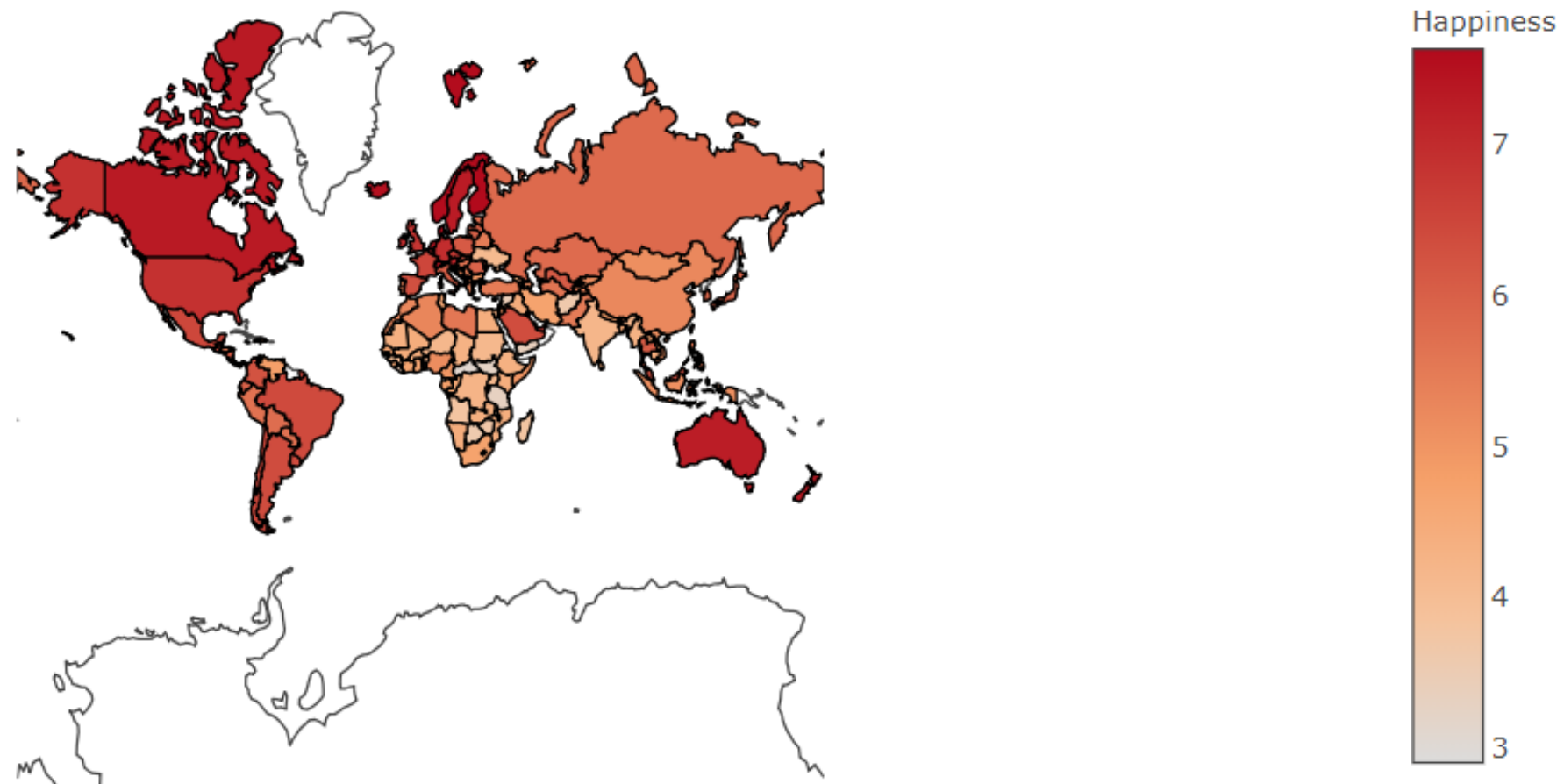
Description of the dataset

- COUNTRY** : Name of each Country
- HAPPINESS\_SCORE** :A metric measured between 2015 to 2017 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."
- GDP\_PER\_PERSON** : GDP per Capita of each Country in terms of Purchasing Power Parity (PPP) (in USD)
- HEALTHY\_LIFE\_EXPECTANCY** : Healthy Life Expectancy at birth are constructed based on data from the World Health Organization (WHO) and WDI.
- SOCIAL\_SUPPORT** : National average of the binary responses (either 0 or 1) to the question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
- FREEDOM** : National average of binary responses to the question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
- GENEROSITY** : Generosity is the residual of regressing the national average of responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.
- PERCEPTIONS\_OF\_CORRUPTION** : Perceptions of corruption are the average of binary answers to two GWP questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?"

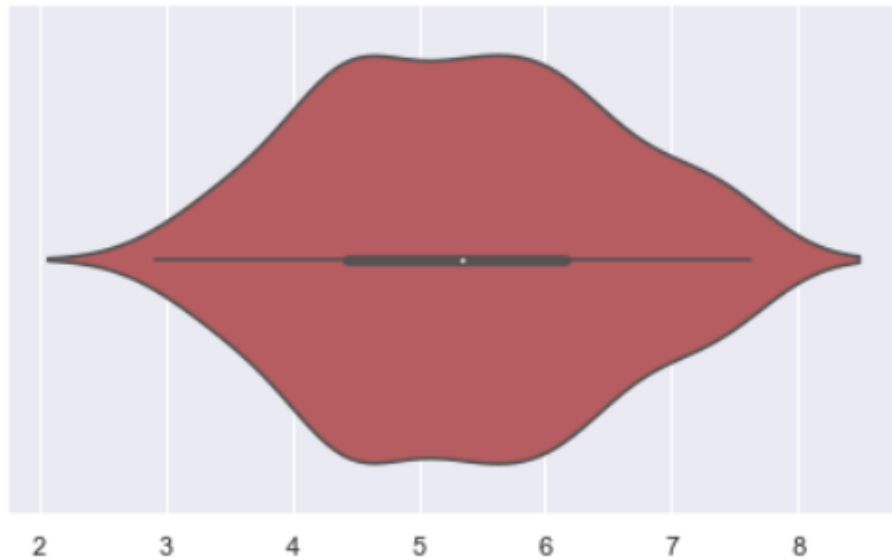
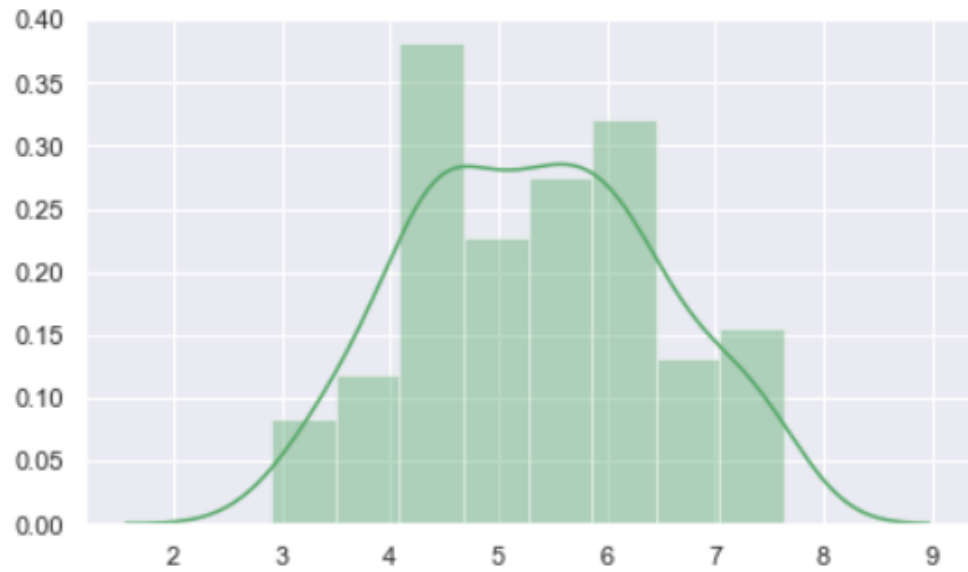
Data Exploration / Visualisation

# Choropleth Map

Happiness Index 2018



```
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x14279bfb588>
```



## Data Exploration & Visualisation

# Happiness Categorisation



### Happy

Happiness Score > 6



### Likely Happy / Normal

4 < Happiness Score < 6

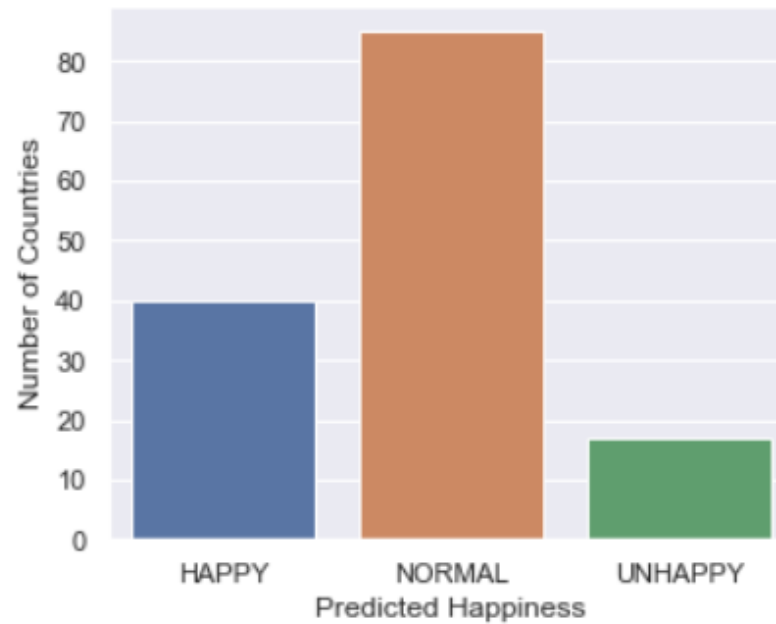


### Unhappy

Happiness Score < 4

# Visualising Happiness

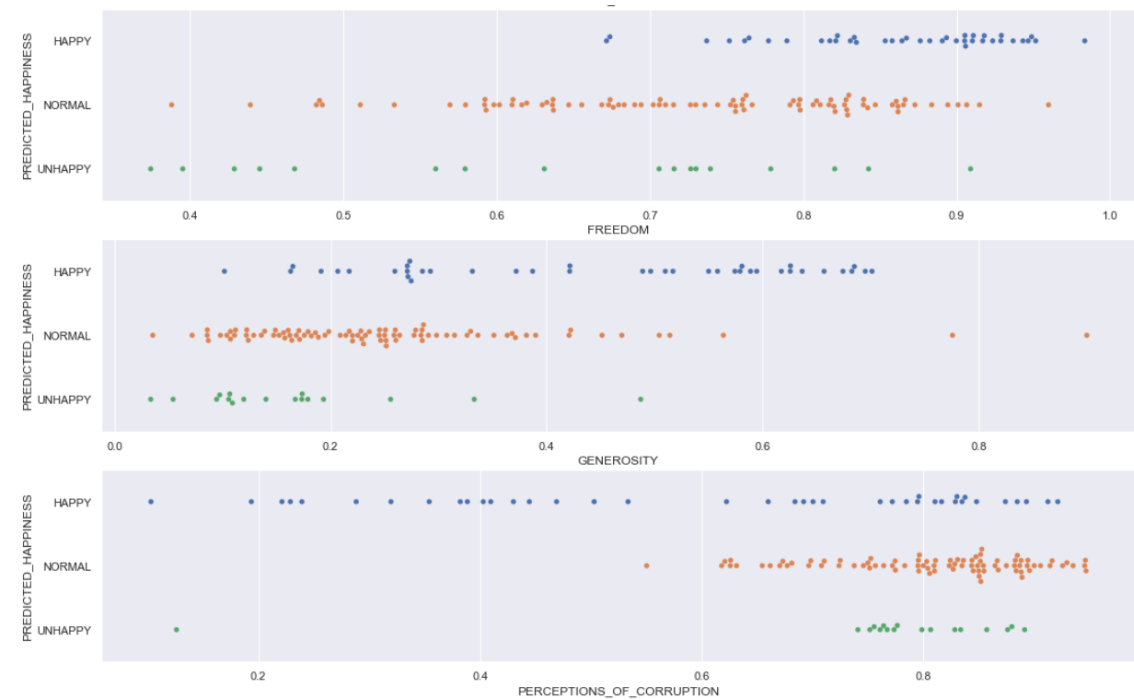
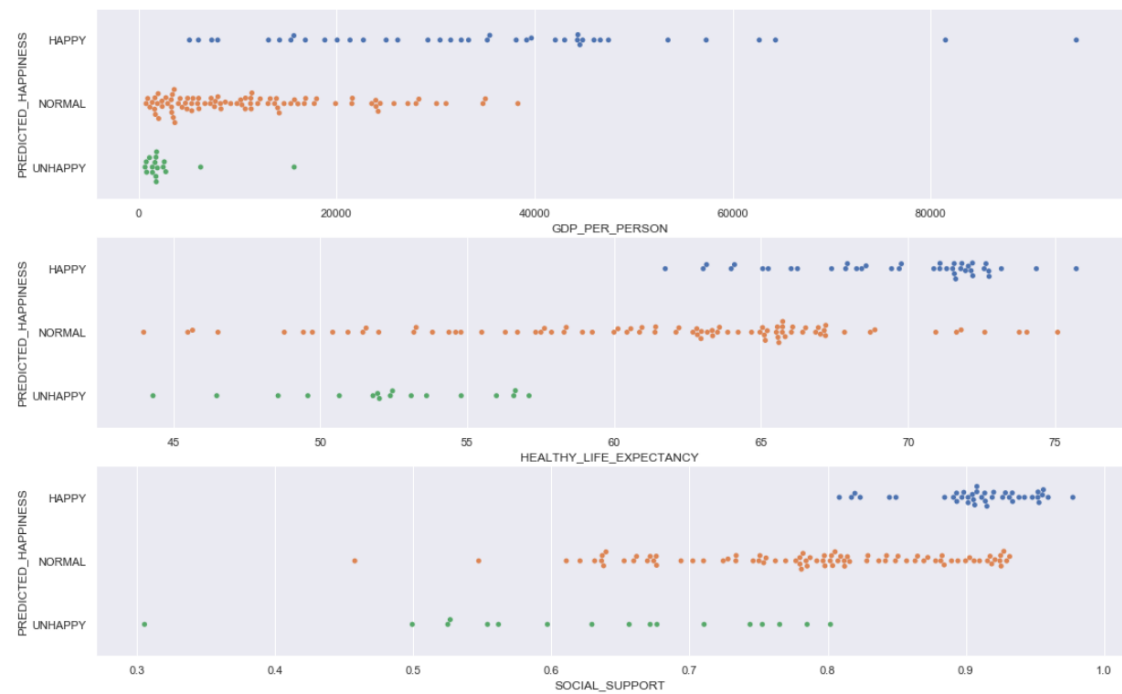
```
Out[18]: Text(0,0.5,'Number of Countries')
```





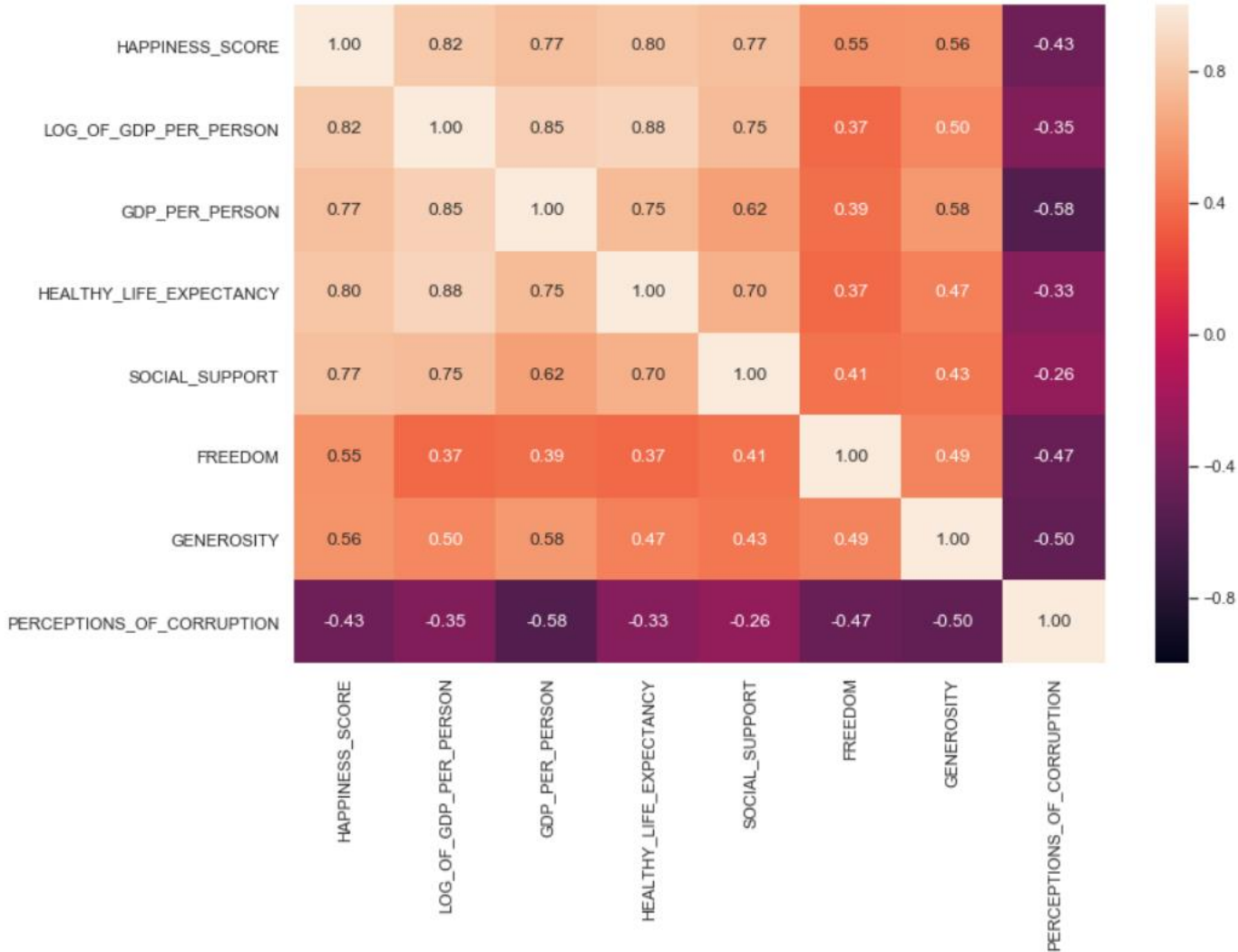
# Data Exploration / Visualisation

## Swarm Plots



# Data Exploration / Visualisation

## Visualising Happiness



### Top factors affecting happiness:

1. GDP Per Capita
2. Life Expectancy
3. Social Support
4. Generosity

### Perception about Happiness:

#### Corruption [Correlation: |0.4|]

Corruption is often seen as an important factor for a country's growth and economy, but it matters the least at an individual level to citizens of a Country as a factor to their happiness.


# Create your own Country using Data Science

Predict Happiness Category and the most similar  
Country to our Virtual Country









Enter your Country Name: Zootopia  User Input


-----  
Let us analyse the feasabiliy [Happiness Category/Score] of Zootopia by entering some important features:

-----  
Reference: Average GDP Per Capita of all Countries: USD 17000, Singapore's GDP Per Capita: USD 81000


Enter the Average GDP Per Capita for your Country (in USD): 40000  User Input


-----  
Enter the Average Life Expectancy of your Citizens: (Age) 71.5  User Input

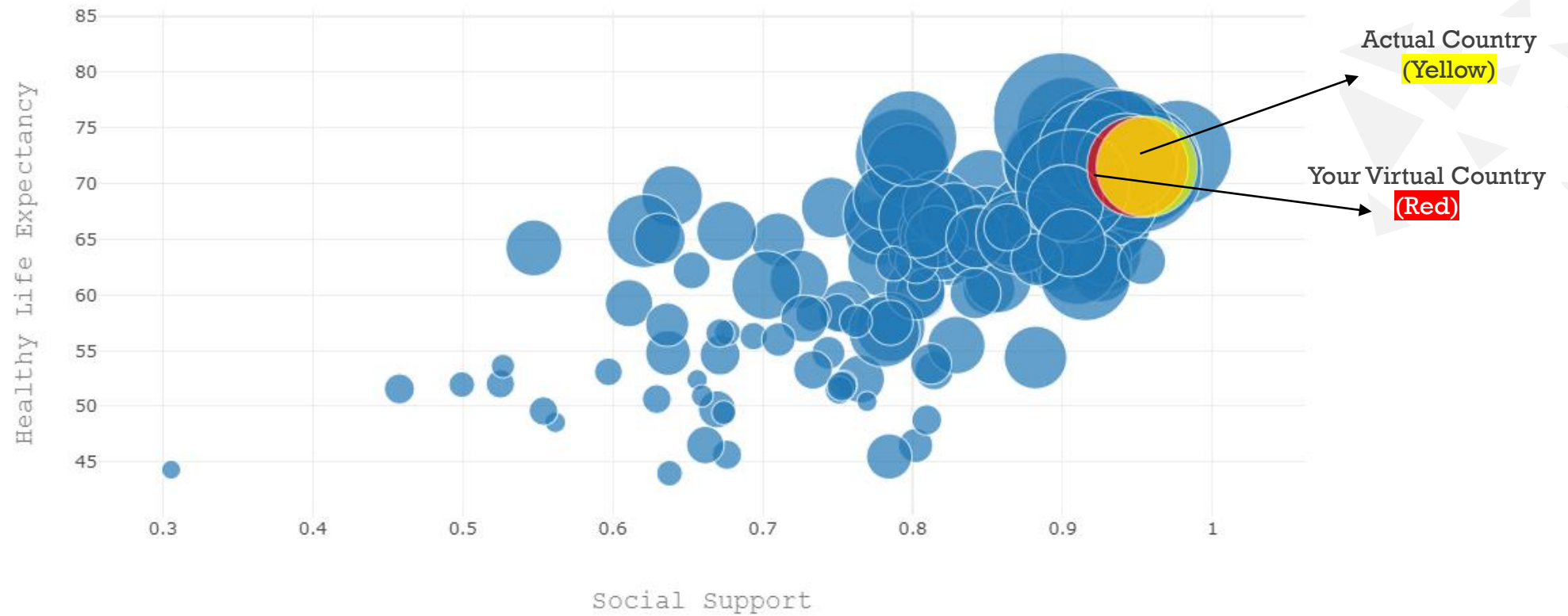
-----  
Reference Question: Social Support - If you were in trouble, do you have relatives or friends you can count on to help you when ever you need them, or not?

Enter the Social Support Index (Choose between 0 and 1, 0 - Lowest : 1 - Highest): 0.95  User Input

-----  
Reference Question: Generosity - Have you donated money to a charity in the past month?

How generous are people in your Country? (Choose between 0 and 1, 0 - Lowest : 1 - Highest): 0.94  User Input

-----  
Congrats! We can say that the citizens of Zootopia will be HAPPY with a probability of 98.47 %  Output



-----

Legend:

Red: Your Country

Yellow: The Country Most Similar to your Country

-----

-----

Finland is the country that has the most similar features to yours ← Output

-----

# Machine Learning

## Predicting Happiness



### Classification – Random Forest

- Random Forest is used to predict the Happiness Category i.e. Happy, Normal and Unhappy using **GDP per Capita, Healthy Life Expectancy, Social Support and Generosity**.
- We can get an accuracy of up to **75%** by using the Random Forest Classification Algorithm.



### Regression – Linear Regression

- Linear Regression is used to predict which Country is the most similar to our virtual country in terms of **GDP per Capita, Healthy Life Expectancy and Social Support**.
- We can get an accuracy of up to **70%** by using the Linear Regression algorithm.





# Classification – Random Forest

```
# Create Random Forest Classifier
forest = RandomForestClassifier(n_estimators=80, random_state=42, max_depth=10, min_samples_split=0.1, min_samples_leaf=0.001)
forest.fit(x_train, y_train)
```

- **n-estimators:** The number of trees in the forest
- **random-state:** It is the random number generator
- **Max-depth:** The maximum depth of the tree.
- **Min\_samples\_split:** The minimum number of samples required to split an internal node
- **Min\_samples\_leaf :** The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least *min\_samples\_leaf* training samples in each of the left and right branches.



# Regression – Linear Regression

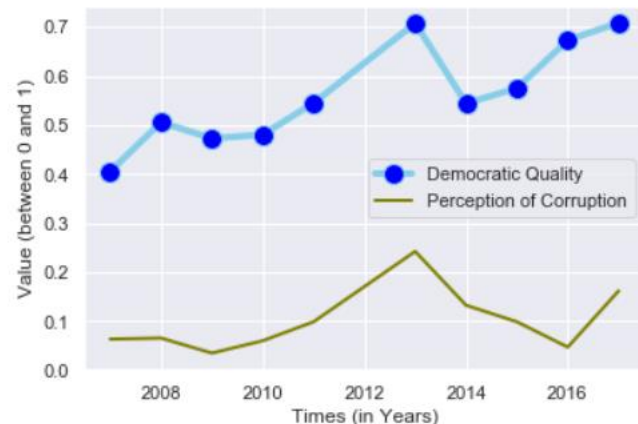
Several regression models have been trained and tested on our **Training Set** and **Test Set**. Below is there performance:

Model	R <sup>2</sup> Score (Training set – Test set)
Linear Regression	0.7483 - 0.7414
Ridge Regression	0.7482 - 0.7414
Lasso Regression	0.5361 - 0.5003

- Ridge and Lasso regression have been trained with GridSearchCV to find out the best *alpha* parameter that returned the highest R<sup>2</sup> score.
- In the end, Linear Regression was chosen due to highest R<sup>2</sup> score and fast computing time.

# Economic Analysis of Singapore

```
Out[51]: <matplotlib.legend.Legend at 0x28e2acfc630>
```

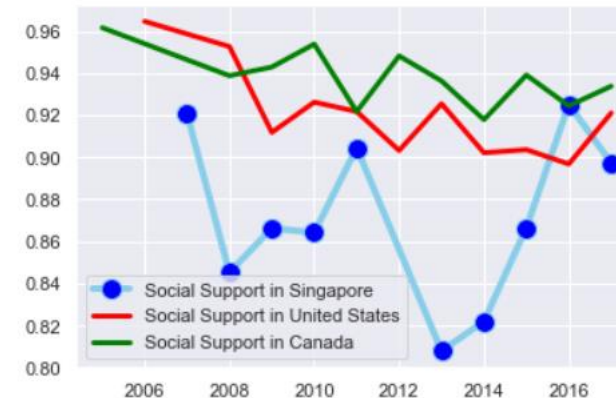


## Democratic Quality

Corruption in Singapore is generally perceived as low. However, a few minor scandals led to an sharp dip in Democratic Quality of Singapore.

We can infer from the following observations that an increase in the Perception of Corruption led to a decrease in the Democratic Quality in the subsequent years.

```
Out[17]: <matplotlib.legend.Legend at 0x21dd382fd68>
```



## Social Support

The government has provided schemes and subsidies to lower-income groups. For example, the Workfare Income Supplement Scheme which helps them build up their retirement savings.

Due to these policies, we can see an increase in Social Support of Singapore from 2013 onwards.



# Economic Analysis of Singapore

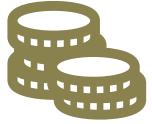
## Life Satisfaction



We can observe a sharp downtrend in life satisfaction from 2014 to 2016.

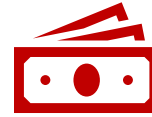
This can be attributed to the fact that the finance & insurance and health-care sectors didn't perform well in the following years and registered a 1.44 point decline from the previous year.

# Interesting Observations



## Corruption and Democracy

- We infer from our observations that an increase in the **Perception of Corruption** led to a **decrease** in the **Democratic Quality** in the subsequent years of Singapore.



## Money != Happiness

- Out of the top factors identified which contribute to happiness, GDP per Capita is the only factor which relates to money.
- All other factors such as Life Expectancy, Generosity and Social Support contribute significantly in happiness of a Country.

# Project Distribution



Gupta Jay	Tieu Phat Dat	Nguyen Duy Khanh
Problem Statements	Problem Statements	Problem Statements
Data Preparation and Cleaning	Data Preparation and Cleaning	Data Preparation and Cleaning
Basic Statistics & Exploratory Data Analysis / Visualisation	Section 2: Create your own Country - Regression Models (Linear Regression, Lasso Regression, Ridge Regression and GridSearchCV)	Section 1: What are the top factors which affect happiness?
Section 1: What are the top factors which affect happiness? Are there any wrong perceived factors about happiness? Section 2: Create your own Country - Classification Model (Random Forest)	Exception Handling	Section 2: Create your own Country - Classification Model (Random Forest)
Economic Analysis of Singapore		





Thank You