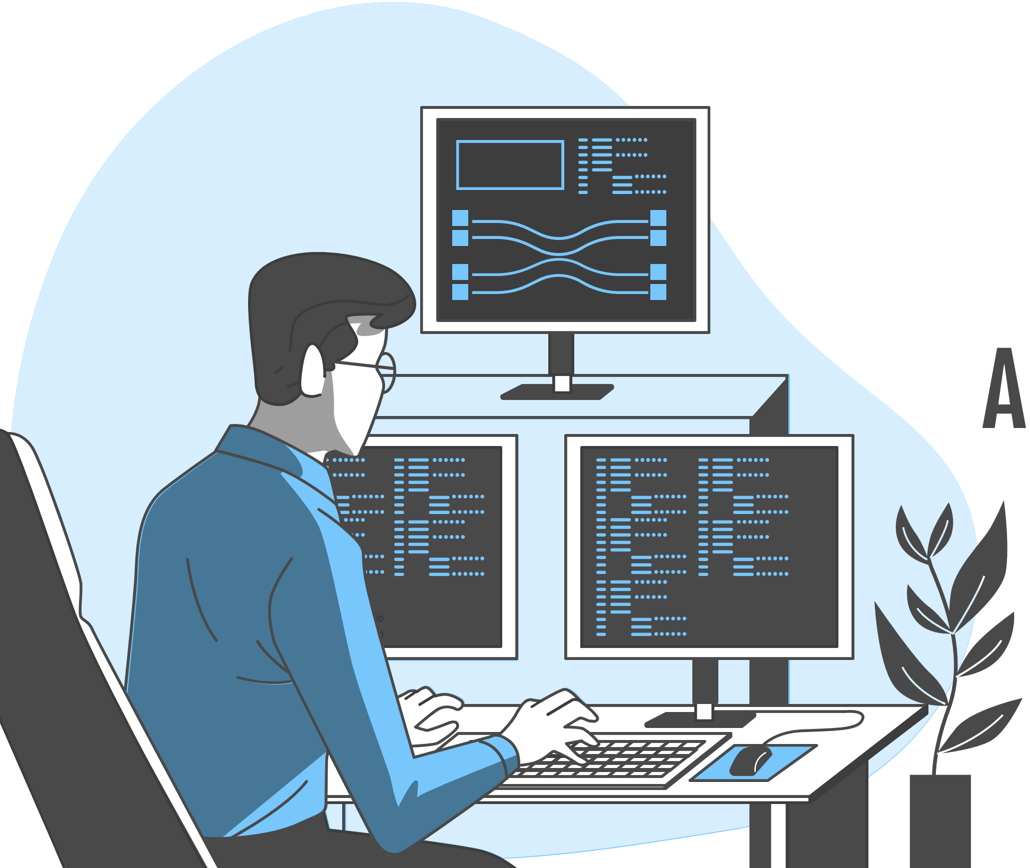


SC1015 Data Science & Artificial Intelligence Mini-Project

Group Members:
Teng Song Heng (U2122030K)
Priscilla Celine Setiawan (U2123732G)
Sim Shi Jie, Keith (U2121044B)

Practical Motivation

Happiness Index

"Singapore is ranked 27th"

Factors

What factors affect a country's happiness?

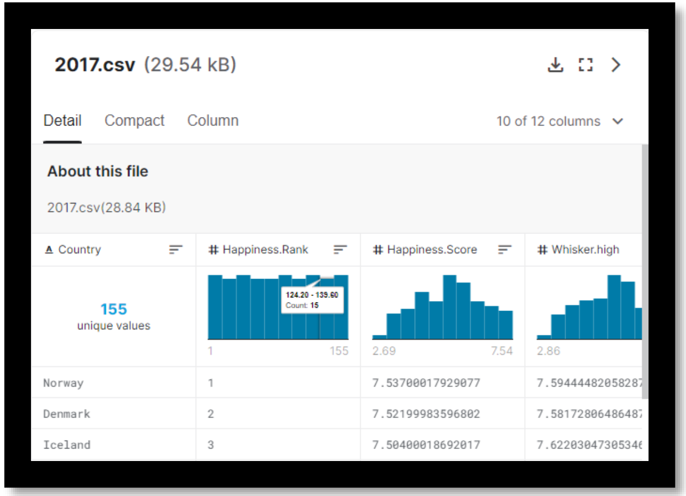
Improvements

What can countries do to increase happiness?

Problem Definition

"To predict the happiness score of a country"

Chosen Dataset



2017 World Happiness Dataset

Usage :

Predict the happiness score of a country based on several predictors

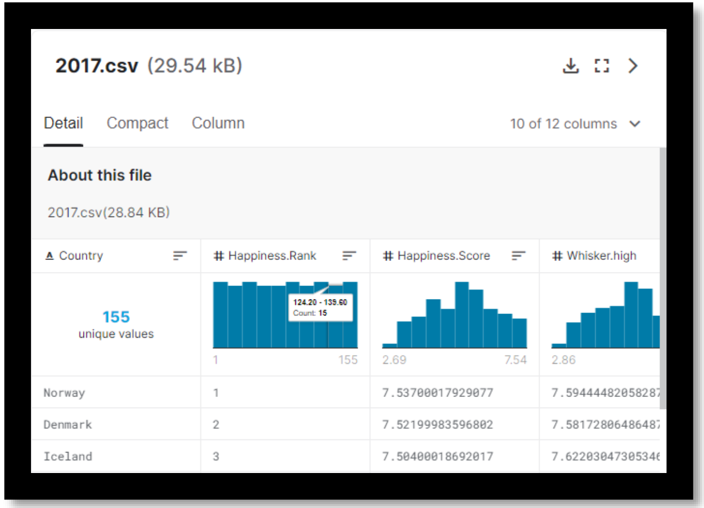
E.g., GDP, Social Support, etc.

Source :

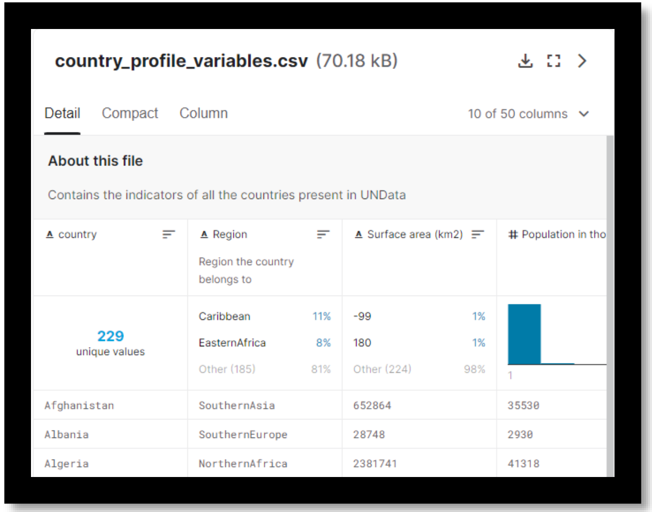
World Happiness Report (Kaggle)

https://www.kaggle.com/datasets/unsdsn/world-happiness

Creation of Dataset



2017 World Happiness Dataset



2017 UN Country Statistic Dataset

Problem :

Country data contained within are very generalised and lack specificity.

Solution :

Introduce detailed country data for data analysis and happiness score prediction

E.g., Employment rate, Mortality rate, etc.

Dataset Creation

Problem :

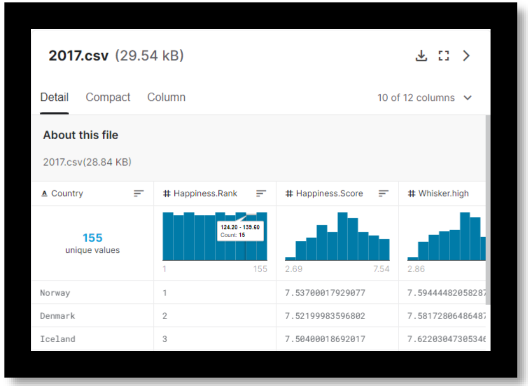
Sample size too small

(less than 200 countries)

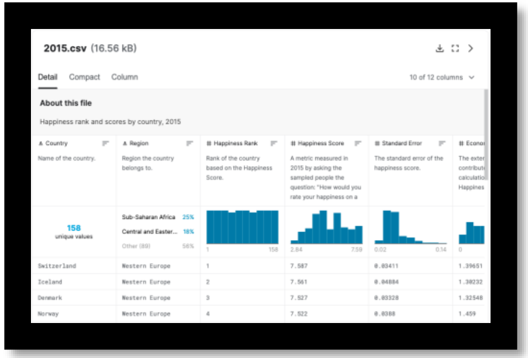
Solution :

Add more samples from other years

E.g., Adding merged data of 2015 world happiness with 2015 statistics



2017 World Happiness Dataset



2015 World Happiness Dataset

Dataset Creation

Problem :

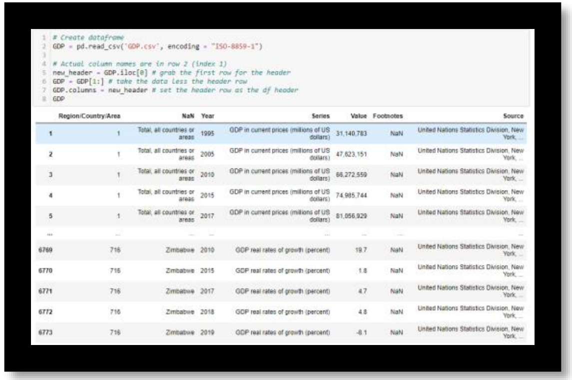
No readily available country statistics dataset as each variable is separated into multiple CSV files.

UN website with CSV

Solution :

Manually download the CSV files on UN website and reformat it to fit usages.

E.g., Filter by 2015 data, etc.



```
# Create dataframe
GDP = pd.read_csv('GDP.csv', encoding = "ISO-8859-1")

# Actual column names are in row 2 (index 1)
new_header = GDP.iloc[0] # grab the first row for the header
GDP = GDP[1:] # take the data less the header row
GDP.columns = new_header # set the header row as the df header
GDP
```

	Region/Country/Area	NaN	Year	Series	Value	Footnotes	Source
1	1	Total, all countries or areas	1995	GDP in current prices (millions of US dollars)	31,140,783	NaN	United Nations Statistics Division, New York, ...
2	1	Total, all countries or areas	2005	GDP in current prices (millions of US dollars)	47,623,151	NaN	United Nations Statistics Division, New York, ...
3	1	Total, all countries or areas	2010	GDP in current prices (millions of US dollars)	66,272,559	NaN	United Nations Statistics Division, New York, ...
4	1	Total, all countries or areas	2015	GDP in current prices (millions of US dollars)	74,985,744	NaN	United Nations Statistics Division, New York, ...
5	1	Total, all countries or areas	2017	GDP in current prices (millions of US dollars)	81,056,929	NaN	United Nations Statistics Division, New York, ...
...	...	...	...	...	...	...	...
6769	716	Zimbabwe	2010	GDP real rates of growth (percent)	19.7	NaN	United Nations Statistics Division, New York, ...
6770	716	Zimbabwe	2015	GDP real rates of growth (percent)	1.8	NaN	United Nations Statistics Division, New York, ...
6771	716	Zimbabwe	2017	GDP real rates of growth (percent)	4.7	NaN	United Nations Statistics Division, New York, ...
6772	716	Zimbabwe	2018	GDP real rates of growth (percent)	4.8	NaN	United Nations Statistics Division, New York, ...
6773	716	Zimbabwe	2019	GDP real rates of growth (percent)	-8.1	NaN	United Nations Statistics Division, New York, ...

Reformatting data from UN CSV

Dataset Cleaning

- **2017 UN Country Statistic Dataset:**
  - Replace '-99', '…', '…/...' With NaN
  - Replace '~0', '~0.0', '-~0.0' with 0
  - Split columns with combined data into two columns.

- **Fill in NA values with <u>kNN Imputation</u>**
  - Replace missing values with mean value of the nearest neighbours using the Euclidean distance metric

- **Merging UN Data with Happiness Report**
  - Based on country name
  - Ensure both dataframes use the same country name
  - E.g., 'United States' vs 'United States of America'

- **Adding extra columns**
  - Longitude and Latitude for EDA

- **Export final dataset.csv file**
  - For ease of use in EDA and ML

Final Dataset

```
1 final_dataset
```

4]:

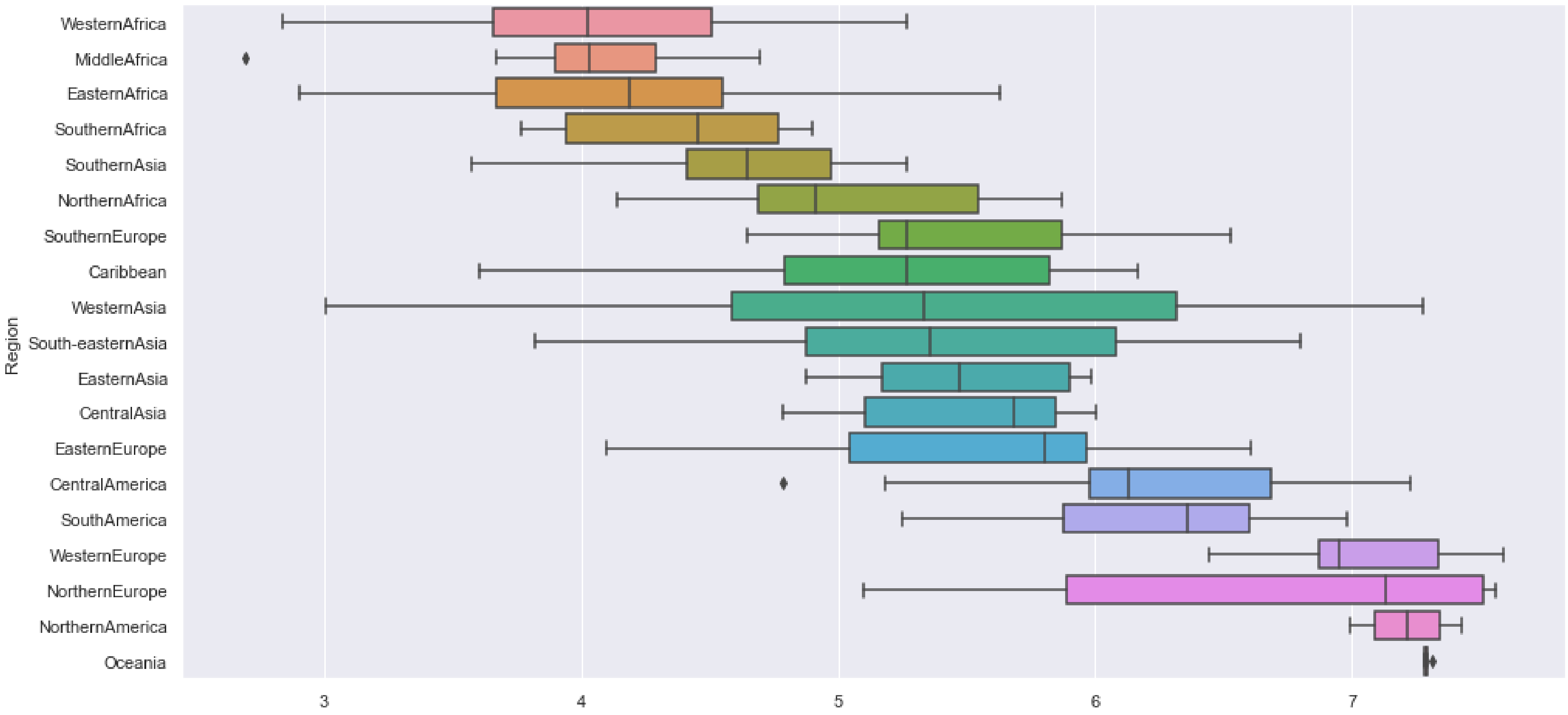
	country	Region	Employment: Industry (% of employed)	Education: Primary gross enrol. ratio (male per 100 pop.)	Education: Primary gross enrol. ratio (female per 100 pop.)	Population age distribution (60+ years, %)	Population age distribution (0-14 years, %)	Pop. using improved sanitation facilities (rural, %)	Pop. using improved sanitation facilities (urban, %)	Pop. using improved drinking water (rural, %)	...	GDP per capita (current US$)	Employment Agricultur (% o employed
0	Afghanistan	SouthernAsia	17.0	122.7	83.5	4.0	44.9	24.005882	33.2	20.400000	...	544.0	47.
1	Albania	SouthernEurope	18.6	104.0	107.3	17.9	18.7	50.100000	43.5	58.870588	...	3939.0	41.
2	Algeria	NorthernAfrica	31.0	118.7	113.0	8.9	28.7	23.100000	17.0	65.900000	...	4178.0	10.
3	Angola	MiddleAfrica	8.6	121.1	105.9	3.6	47.1	17.488235	33.4	37.011765	...	4167.0	50.
4	Argentina	SouthAmerica	23.7	111.4	111.2	14.9	25.2	65.994118	46.7	77.464706	...	14971.0	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	.
297	Venezuela	SouthAmerica	26.8	101.3	98.6	9.9	27.6	69.900000	97.5	77.900000	...	11068.9	11.
298	Vietnam	South-easternAsia	22.9	109.3	108.4	11.1	23.1	69.700000	94.4	96.900000	...	2067.9	41.
299	Yemen	WesternAsia	17.9	105.7	88.9	4.6	39.9	34.100000	92.5	46.500000	...	1106.4	32.
300	Zambia	EasternAfrica	9.9	103.3	104.0	3.7	44.8	35.700000	55.6	51.300000	...	1311.1	54.
301	Zimbabwe	EasternAfrica	7.3	100.8	99.1	4.2	41.2	30.800000	49.3	67.300000	...	890.4	67.

302 rows × 28 columns

```
1 final_dataset.to_csv("Dataset.csv", index=False)
```

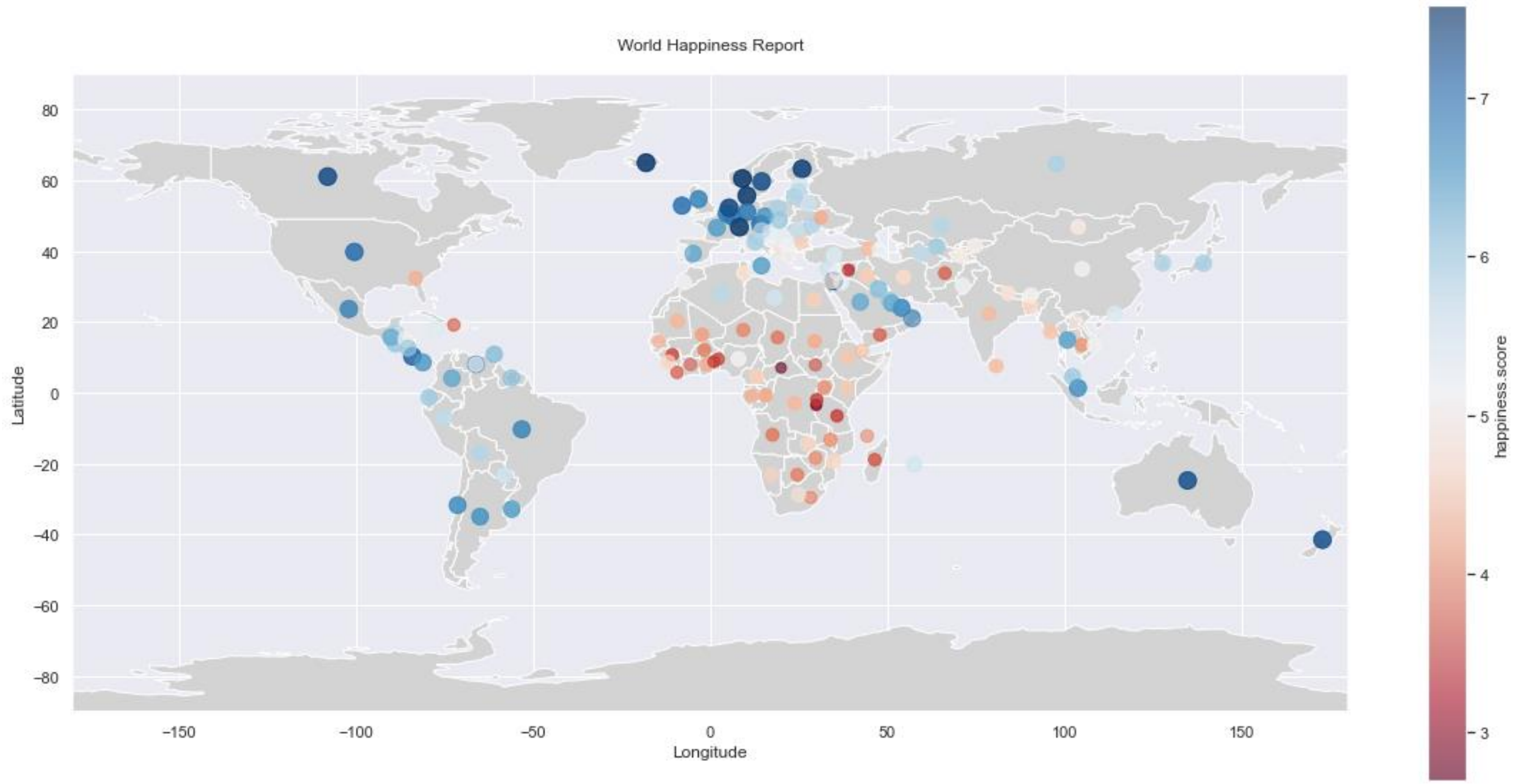
Final Cleaned dataset

Exploratory Data Analysis



Happiness score to Regions

Exploratory Data Analysis

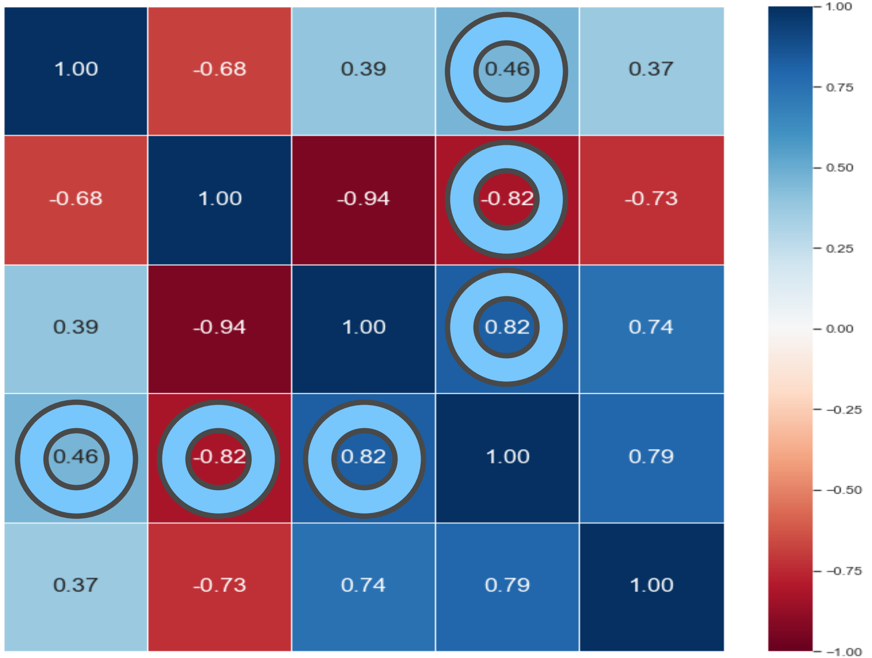


Geographical representation of Happiness score

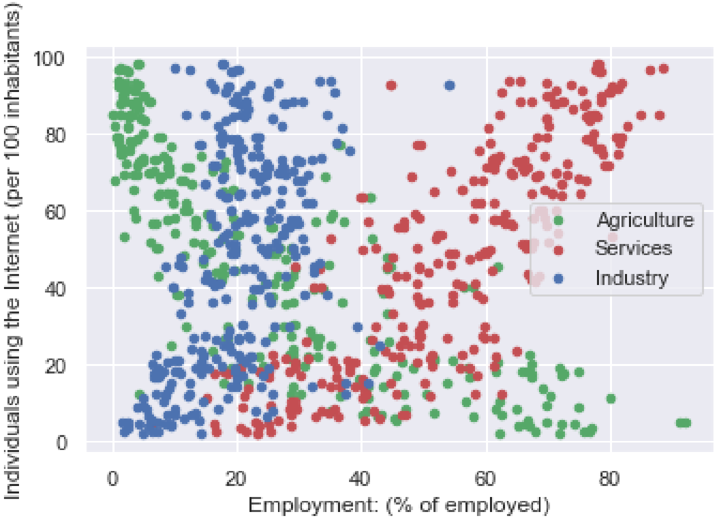
Exploratory Data Analysis

Correlation of **Internet Usage** and **Employment**

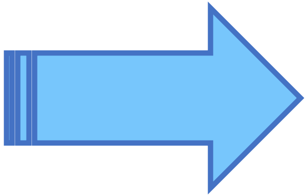
Employment: Agriculture (% of employed)	-0.8247
Employment: Industry (% of employed)	0.4637
Employment: Services (% of employed)	0.8212
Individuals using the Internet (per 100 inhabitants)	1.0000



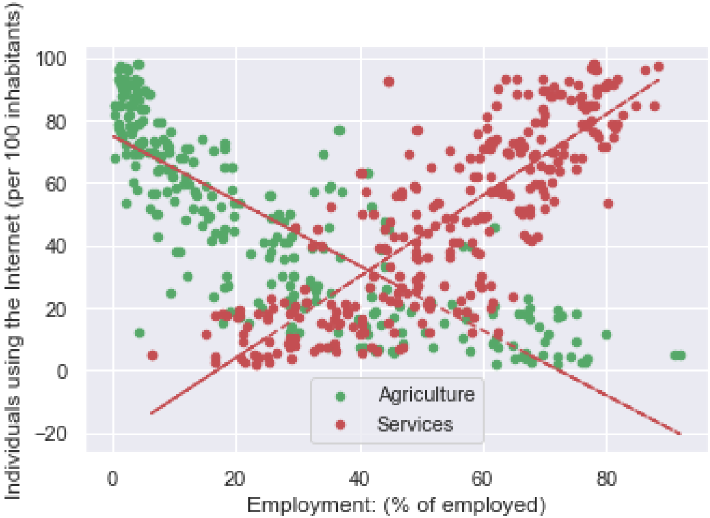
Exploratory Data Analysis



Internet Usage and Employment

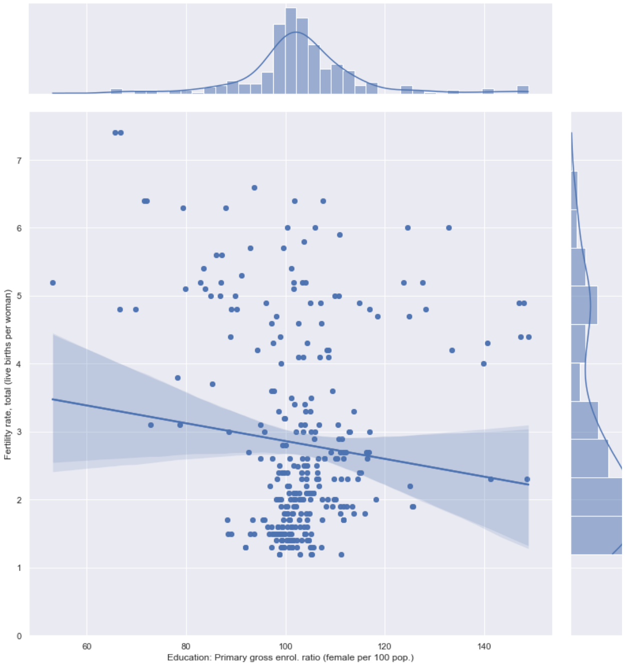


Removed Industry with low correlation and added trend line

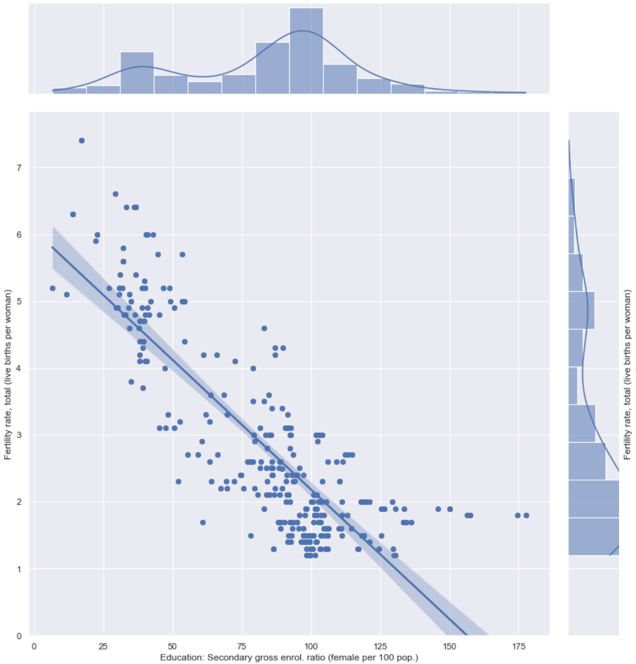


Internet Usage and Employment Trend line

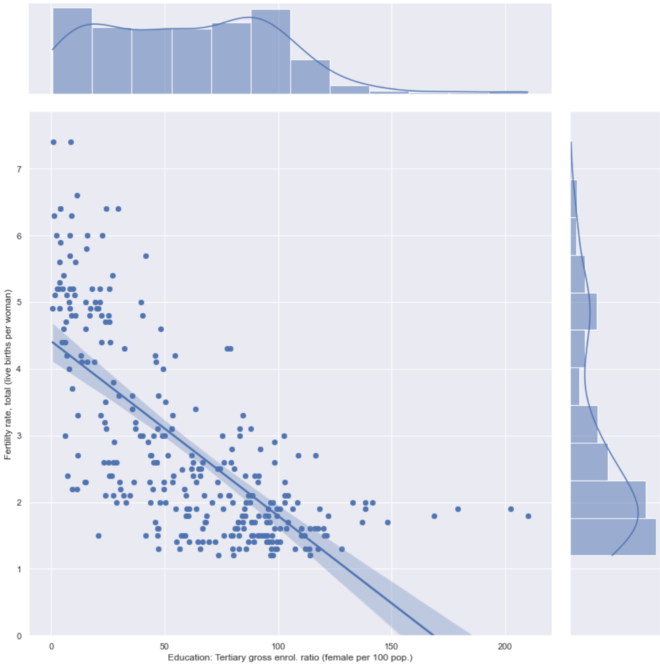
Exploratory Data Analysis



Education: Primary (Female) to Fertility Rate



Education: Secondary (Female) to Fertility Rate

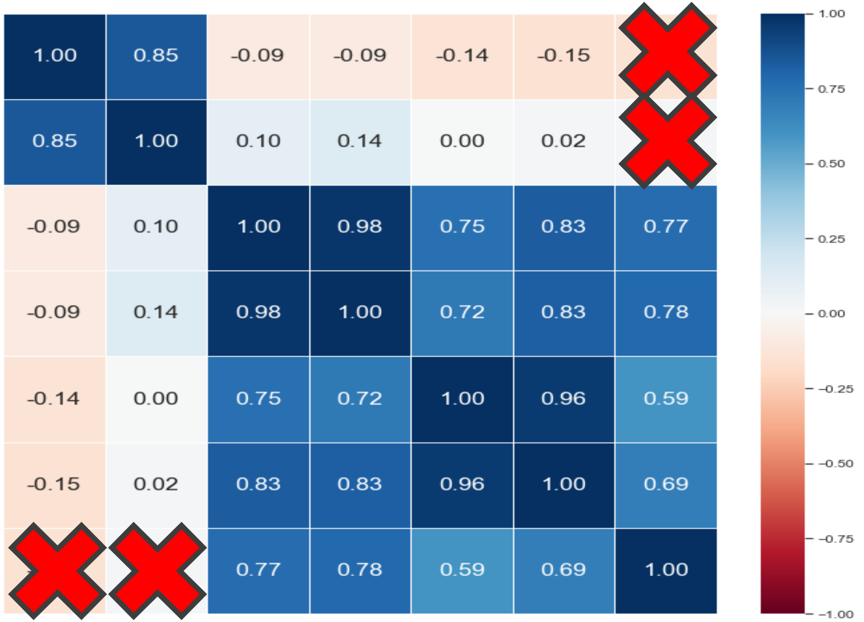


Education: Tertiary (Female) to Fertility Rate

Exploratory Data Analysis

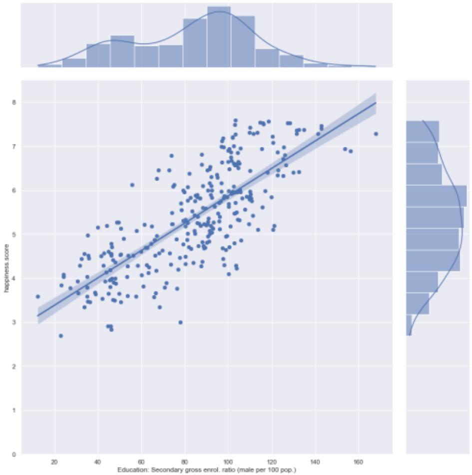
No Correlation between **Primary Education** and **Happiness Score**

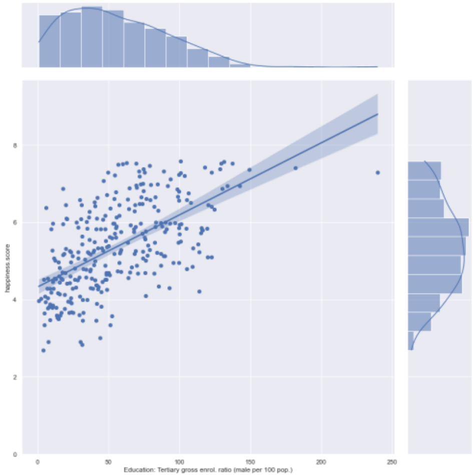
```
Education: Primary gross enrol. ratio (male per 100 pop.)        -0.14
Education: Primary gross enrol. ratio (female per 100 pop.)       0.01
Education: Tertiary gross enrol. ratio (male per 100 pop.)        0.59
Education: Tertiary gross enrol. ratio (female per 100 pop.)      0.68
Education: Secondary gross enrol. ratio (male per 100 pop.)       0.76
Education: Secondary gross enrol. ratio (female per 100 pop.)     0.78
```



Exploratory Data Analysis

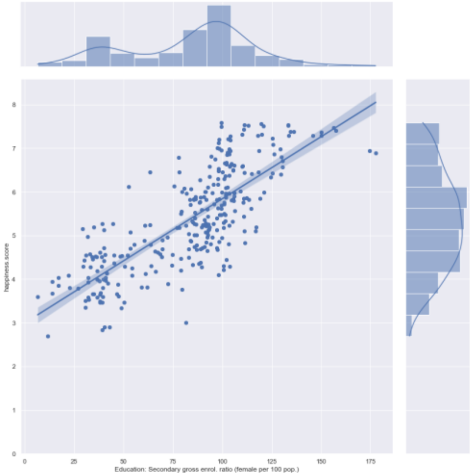
Education :
Secondary (Male)
To Happiness Score

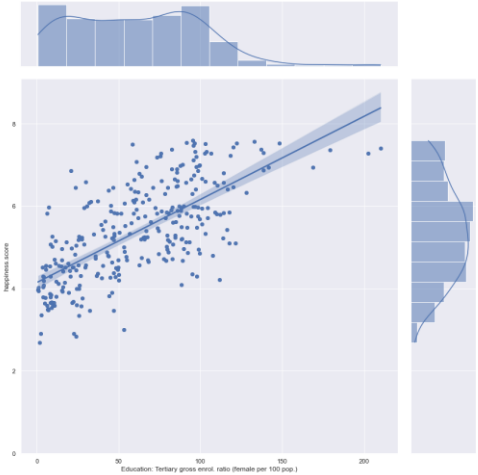




Education :
Tertiary (Male)
To Happiness Score

Education :
Secondary (Female)
To Happiness Score



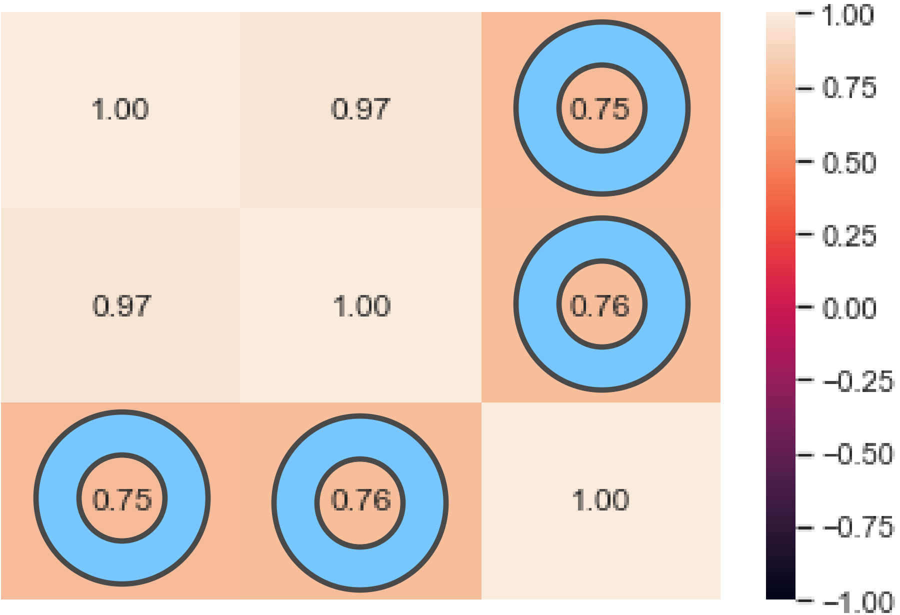


Education :
Tertiary (Female)
To Happiness Score

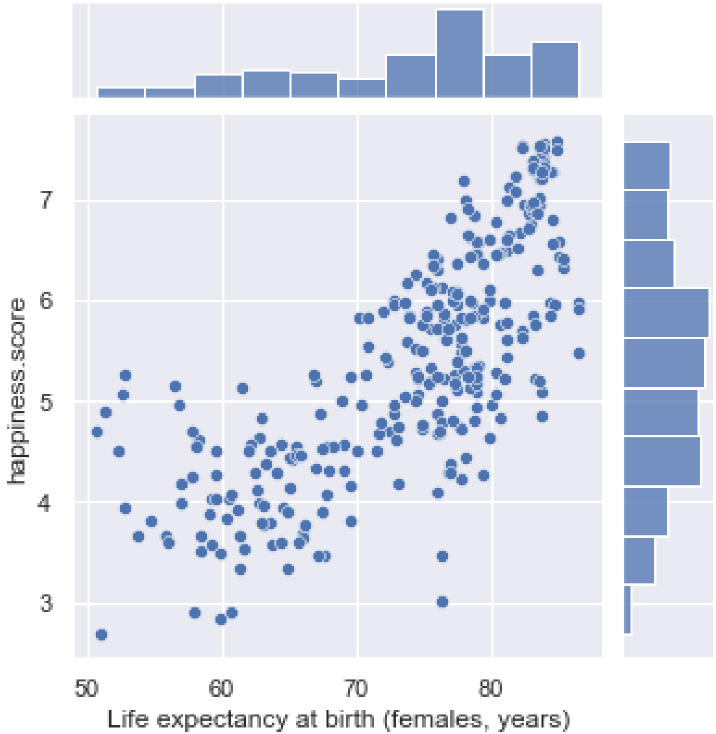
Exploratory Data Analysis

Correlation of **Life Expectancy** and **Happiness Score**

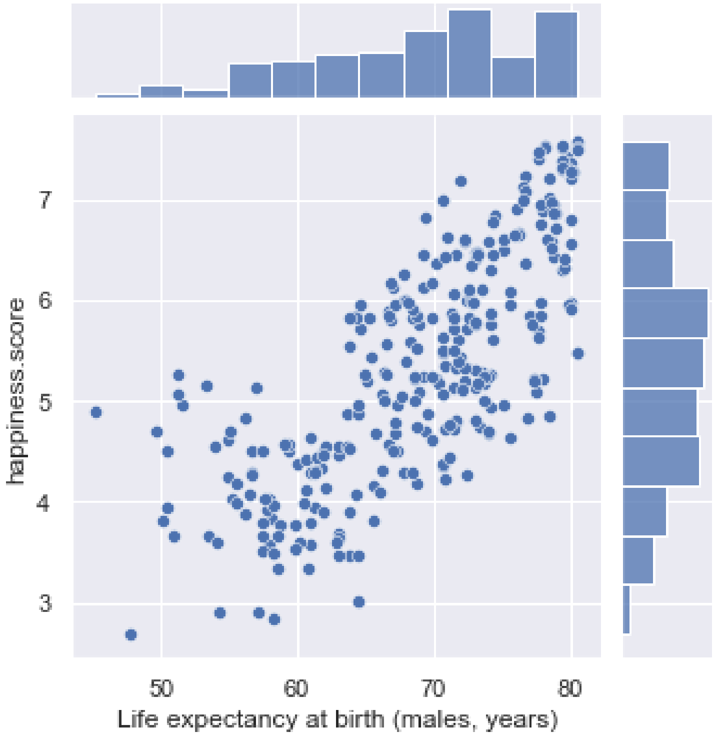
```
Life expectancy at birth (females, years)     0.75
Life expectancy at birth (males, years)       0.76
```



Exploratory Data Analysis

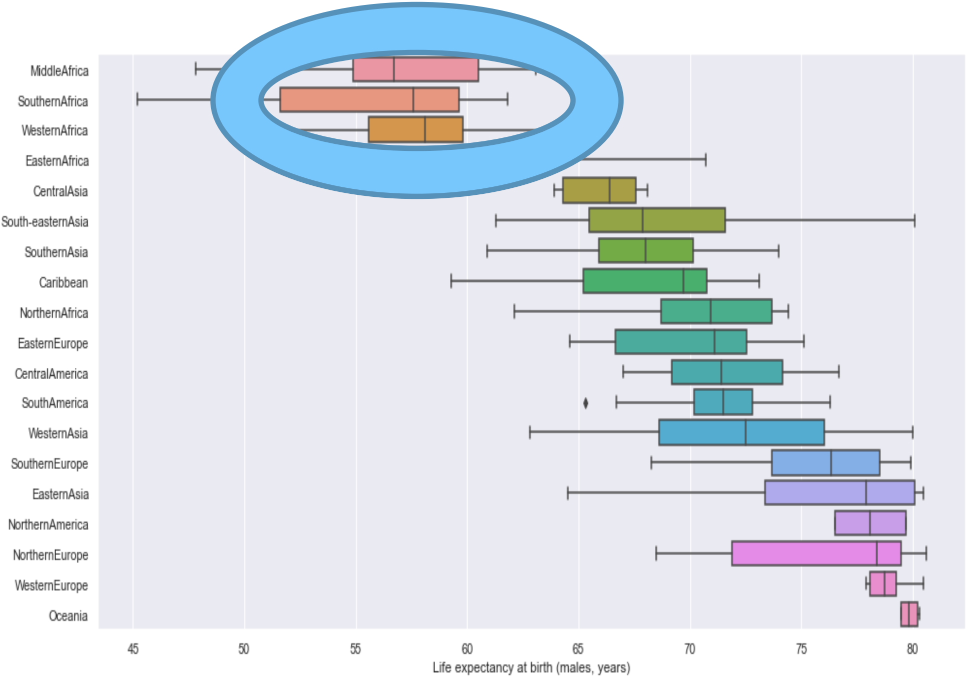


Life Expectancy at Birth (Female)
to Happiness Score

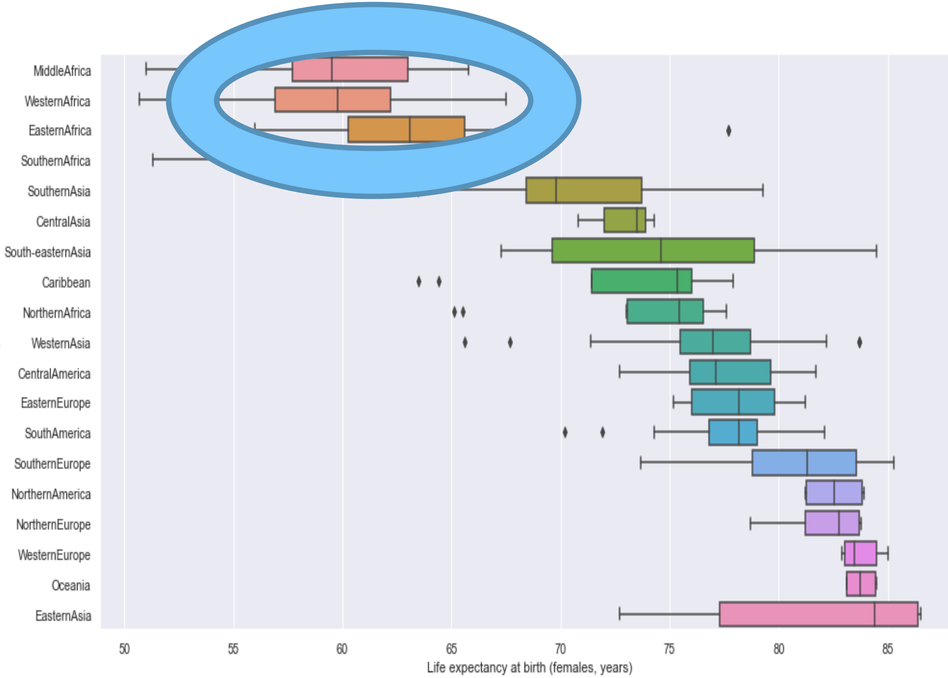


Life Expectancy at Birth (Male)
to Happiness Score

Exploratory Data Analysis

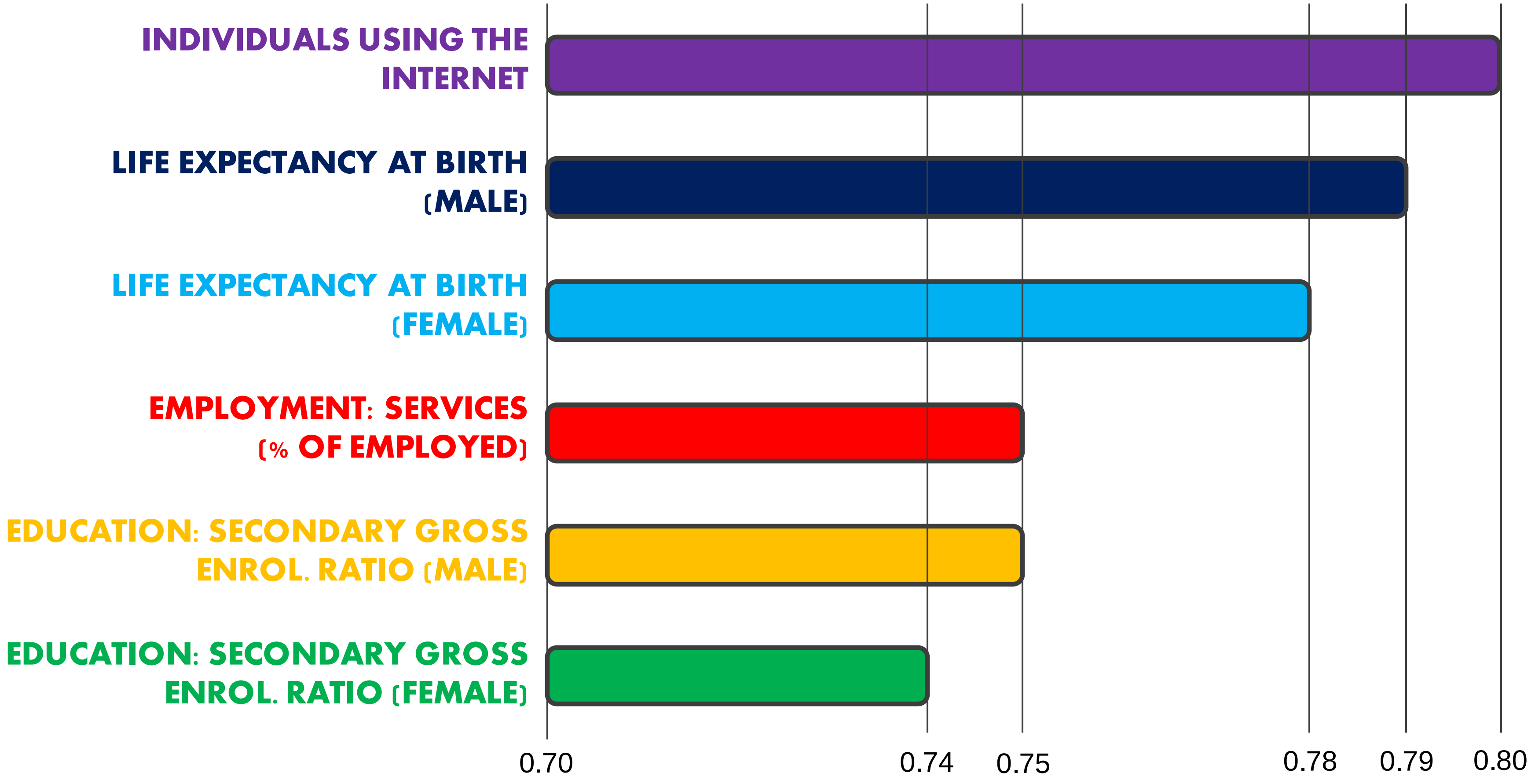


Life Expectancy at Birth (Males) to Region



Life Expectancy at Birth (Female) to Region

Exploratory Data Analysis



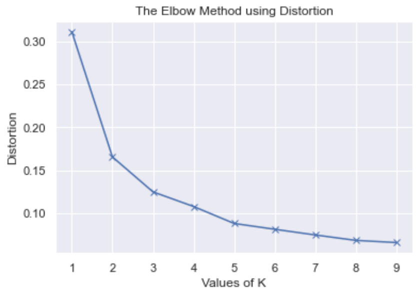
Choosing Machine Learning Model

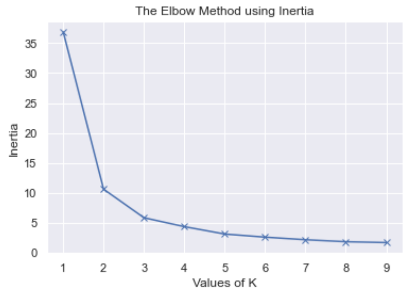
	DESCRIPTION	REASON
K Means Clustering	A form of clustering that aims to partition the samples into K number of clusters with the nearest mean from the chosen centroid	• Clustering gives the user the ability to understand how happy a country is at a simple glance of their group category
ElasticNet Linear Regression	Combination of Least Squares and the regression penalty of both Lasso and Ridge Regression	• A small sample size of ~300. • Only a few features that are highly correlated were chosen as predictors.
Random Forest Regression	Supervised learning algorithm that uses ensemble learning method for regression	• Robust to outliers. • Lower risk of overfitting.

K Means Clustering

Techniques / What is it for?

- **MinMax scaling**, this is required as K Means clustering clusters samples into group based on the **distance** from the chosen **centroid**.
- **Elbow method** is used to determine the **number of clusters**, K.
  - **Distortion** method
  - **Inertia** method

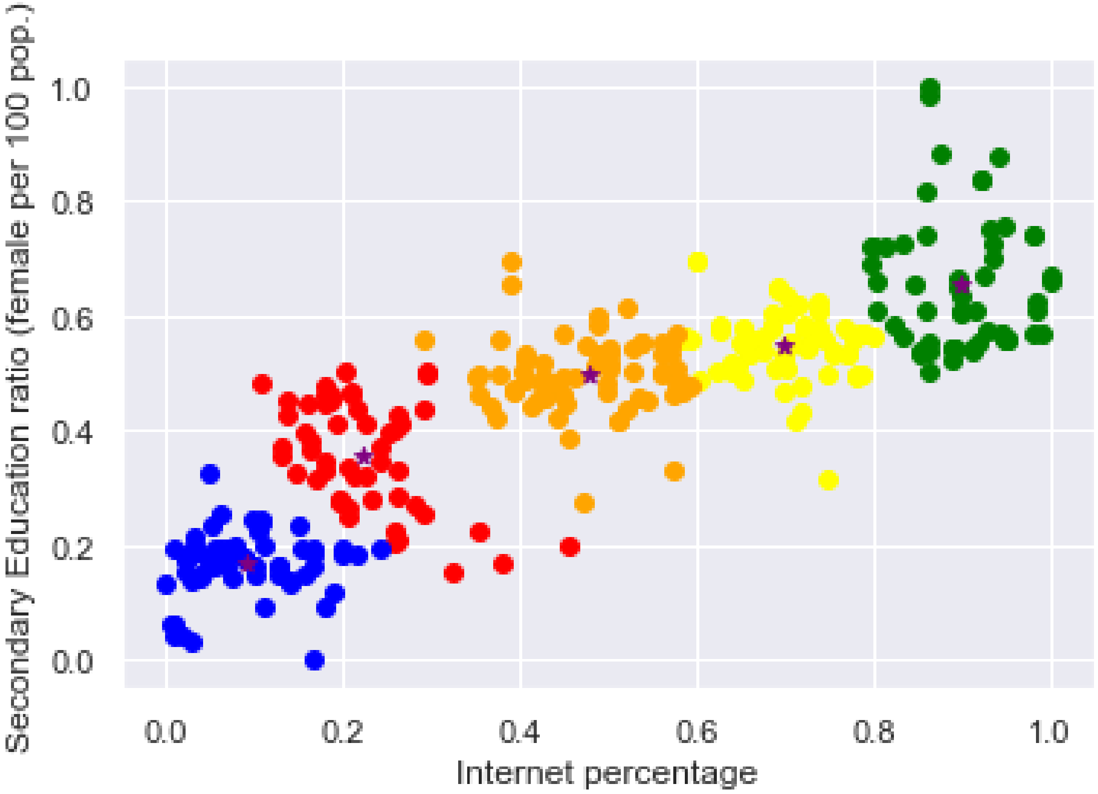




How does it work :

1. Find number of K centroids
2. Group samples based on K distance.

K Means Clustering



ElasticNet Linear Regression

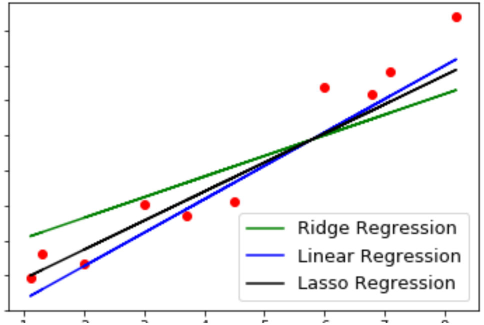
Techniques:

1. Combine the strengths of **lasso** and **ridge regression** into one.
2. **Cross validation** is used to tune the hyperparameters to make more accurate predictions.

How it works:

1. **Least Squares** is the basic linear regression model.
2. **Lasso** (L1 Regularization) and **Ridge Regression** (L2 Regularization) are very similar as they both introduces a small amount of bias to reduce the variance.
3. The only difference is that **Lasso Regression** can shrink less important features coefficient all the way to zero, which helps with feature selection.
4. Minimizes the objective function :

```
1 / (2 * n_samples) * ||y - Xw||^2_2
+ alpha * l1_ratio * ||w||_1
+ 0.5 * alpha * (1 - l1_ratio) * ||w||^2_2
```



Random Forest Regression

Techniques:

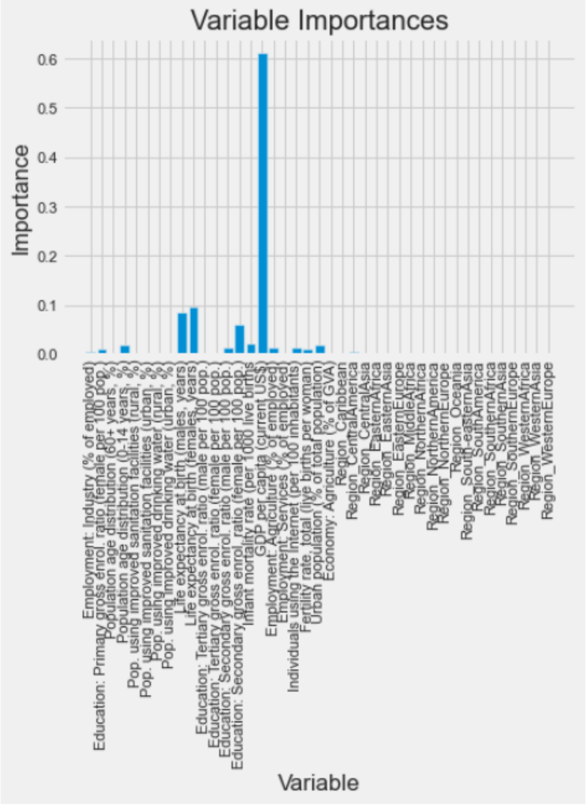
1. Ensemble learning method : combine prediction from multiple ML algorithms to make more accurate prediction than a single model.
2. Bootstrap (bagging) : random sampling with replacement of a small subset of data from data set

How does it work :

1. Construct many decision trees
2. Each tree is created from a different sample of data and features.
3. Each tree makes its own individual prediction.
4. Averaged the predictions to produce a single result.

Variable Importance

1. GDP per capita (current US$) : 0.61
2. Life expectancy at birth (males, years) : 0.1
3. Life expectancy at birth (females, years) :0.08
4. Education: Secondary gross enrol. ratio (female per 100 pop.) :0.06
5. Population age distribution (0-14 years, %) :0.02
6. Infant mortality rate (per 1000 live births :0.02
7. Urban population (% of total population) :0.02



Comparison of Accuracy and RMSE

Accuracy (R^2)	Train	Test
ElasticNet	71.4%	59.3%
Random Forest (max depth =20)	97.5%	80.0%
Random Forest (max depth = 3)	84.8%	76.2%
Random Forest (with most important variables)	97.4%	80.8%
Random Forest (variables with high coefficient)	96.9%	67.4%

Root Mean Squared Error (RMSE)	Train	Test
ElasticNet	0.6244	0.6434
Random Forest (max depth =20)	0.1820	0.4780
Random Forest (max depth = 3)	0.4523	0.5221
Random Forest (with most important variables)	0.1886	0.4688
Random Forest (variables with high coefficient)	0.2049	0.6112

OUTCOME

K Means Clustering

- Easy to implement and visualize
  - Data is labelled and unnecessary for unsupervised learning
  - Accuracy is low

Elastic Net Regression

- Combine the strengths of lasso and ridge regression into one.
  - Low Accuracy and High RMSE value
  - Computationally more expensive than Lasso or Ridge regression.

Random Forest

- Highest Accuracy and lowest RMSE value
- Able to find variables that are significant
- Reduce overfitting
  - Slow training time

Insights and Recommendation

<u>Insights</u>

1. Through EDA and correlation matrix, we found out that these variables are important in determining happiness scores of countries.
    - Employment: Services (% of employed) - 0.744378
    - Life expectancy at birth (females, years) - 0.754074
    - Life expectancy at birth (males, years) - 0.763078
    - Education: Secondary gross enrol. ratio (male per 100 pop.) - 0.768661
    - Education: Secondary gross enrol. ratio (female per 100 pop.) - 0.782680
    - Individuals using the Internet (per 100 inhabitants) - 0.789634
2. Exploration of other machine learning techniques compared to those learnt in class
3. Choosing the right machine learning technique for a specific use case.
4. New methods to cleaning and processing data
5. Importance of normalizing data before machine learning

<u>Recommendations</u>

1. Import other years to increase the sample size
2. Tuning of hyperparameters to improve accuracy in machine learning (e.g., GridSearchCV)

References

1. Htoon, K. S. (2020, July 3). A guide to KNN imputation. Retrieved April 22, 2022, from https://medium.com/@kyawsawhtoon/a-guide-to-knn-imputation-95e2dc496e
2. Nagpal, A. (2017, October 14). *L1 and L2 regularization methods*. L1 and L2 Regularization Methods. Retrieved April 22, 2022, from https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c
3. StatQuest with Josh Starmer. (2018, September 25). *Regularization Part 1: Ridge (L2) Regression* [Video]. Youtube. Retrieved April 22, 2022, from https://www.youtube.com/watch?v=Q81RR3yKn30
4. StatQuest with Josh Starmer. (2018, October 1). *Regularization Part 2: Lasso (L1) Regression* [Video]. Youtube. Retrieved April 22, 2022, from https://www.youtube.com/watch?v=NGf0voTMlcs
5. StatQuest with Josh Starmer. (2018, October 8). *Regularization Part 3: Elastic Net Regression* [Video]. Youtube. Retrieved April 22, 2022, from https://www.youtube.com/watch?v=1dKRdX9bfIo
6. Koehrsen, W. (2017, December 27). *Random Forest in Python*. Medium; Towards Data Science. https://towardsdatascience.com/random-forest-in-python-24d0893d51c0
7. *Random Forest Regression: When Does It Fail and Why?* (2020, May 22). Neptune.ai. https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why
8. *k-Means Advantages and Disadvantages | Clustering in Machine Learning | Google Developers*. (2019, May 6). Google Developers. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages
9. codebasics. (2019, February 4). *Machine Learning Tutorial Python - 13: K Means Clustering Algorithm* [Video]. Youtube. Retrieved April 8, 2022, from https://www.youtube.com/watch?v=EItlUEPCIzM&t=1229s&ab_channel=codebasics
10. "K-Means Pros & Cons." *HolyPython.com*. (29, June 2021). https://holypython.com/k-means/k-means-pros-cons/.

References

1. "Elbow Method for Optimal Value of K in Kmeans." *GeeksforGeeks*, 9 Feb. 2021, https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/.
2. "K-Means Clustering with Scikit-Learn." *DataCamp Community*, https://www.datacamp.com/community/tutorials/k-means-clustering-python.