Lan Nguyen
Individual Project
Intro to Data Analytics
Prof. Jack Follis

Individual Project Summary

1. **Introduction:**
   - Topic: Happiness Ranking Across Countries.
   - Purpose: It is important to take into account of which indicators can determine and contribute to the Happiness Index. Therefore, it would give more confidence and aspect to whoever want to experience a new horizon.
2. **Data Collection:**
   - Data was collected from website Kaggle. https://www.kaggle.com/unsdsn/world-happiness

     There are 5 datasets record detail to the change in happiness score and ranking among countries over the years from 2015 to 2019.
   - Each dataset mainly contains features: Ranking, Country Name, Region, Happiness Score, Health Expectancy, Freedom, GDP per Capita, Generosity, Corruption.
   - By exploring the dataset, it may help us have more understanding of how happy countries are ranked.
3. **Methods:**
   - Four datasets 2016,2017,2018 and 2019 will be main datasets to be explored. Since each dataset has the same structure, I create another column name "Year" in 2019 datasets then I stack all the value from 2016,2017,2018 to it and label with Year accordingly. I name the new dataset "combineddata"
   - Datasets 2016 has "Region" features but the other 3. Thus I have to left_join combineddata with 2016, which is selected only 2 columns "Country" and "Region", by "Country".
   - Due to different way in naming some countries between datasets, new countries listed that are not in the 2016 dataset, joining dataset has more than 20 counts N/A under "Region". Thus, I use dataset 2016 to benchmark with the joining dataset to see the mismatch and thus reduce the N/A into 0 and label those value to the correct region.
   - I will use graphs and RainForest to exanimate the dataset to discover the main factor influences the Happiness Score. Since then, I will factor Happiness Score into 3 level: Low, Medium and High base on descriptive statistic of Score (Min, 1stQ, 3rdQ and Max)

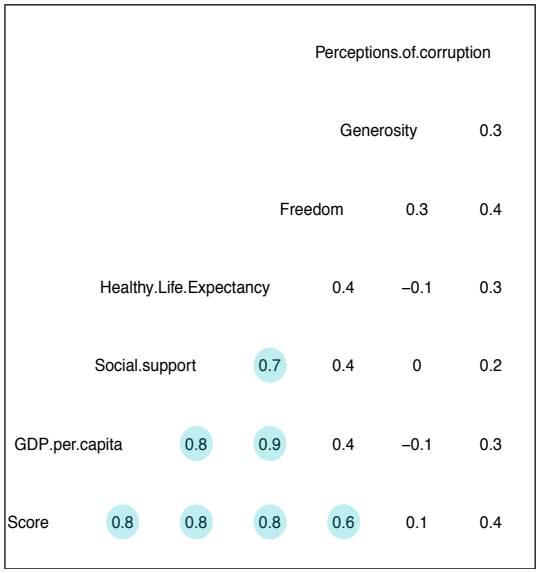
4. Result & Discussion:

Check Correlation

CORRELATION-2017

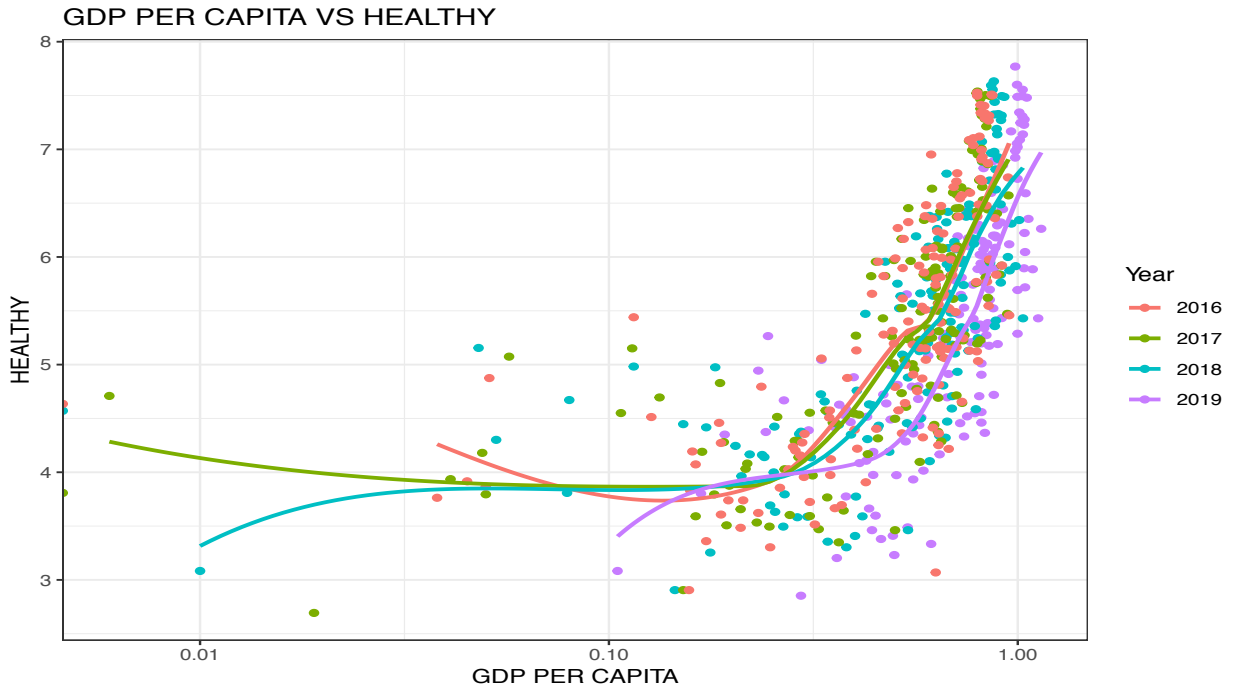
CORRELATION–2018

CORRELATION–2019

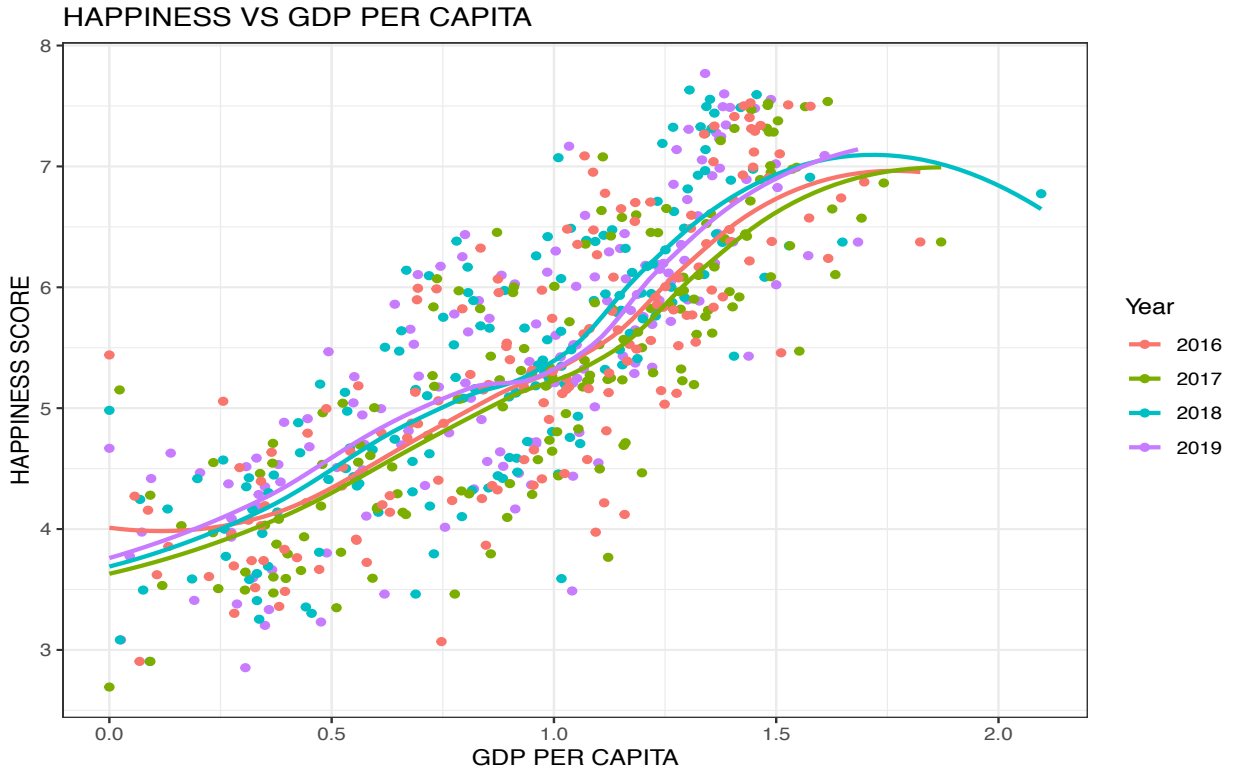


Those figures show us that there are 4 main factor, GDP per capita, Social Support, Healthy Life, and Freedom that have positive correlations with Happiness score. While the other factors that has very weak correlation with Happiness Score.

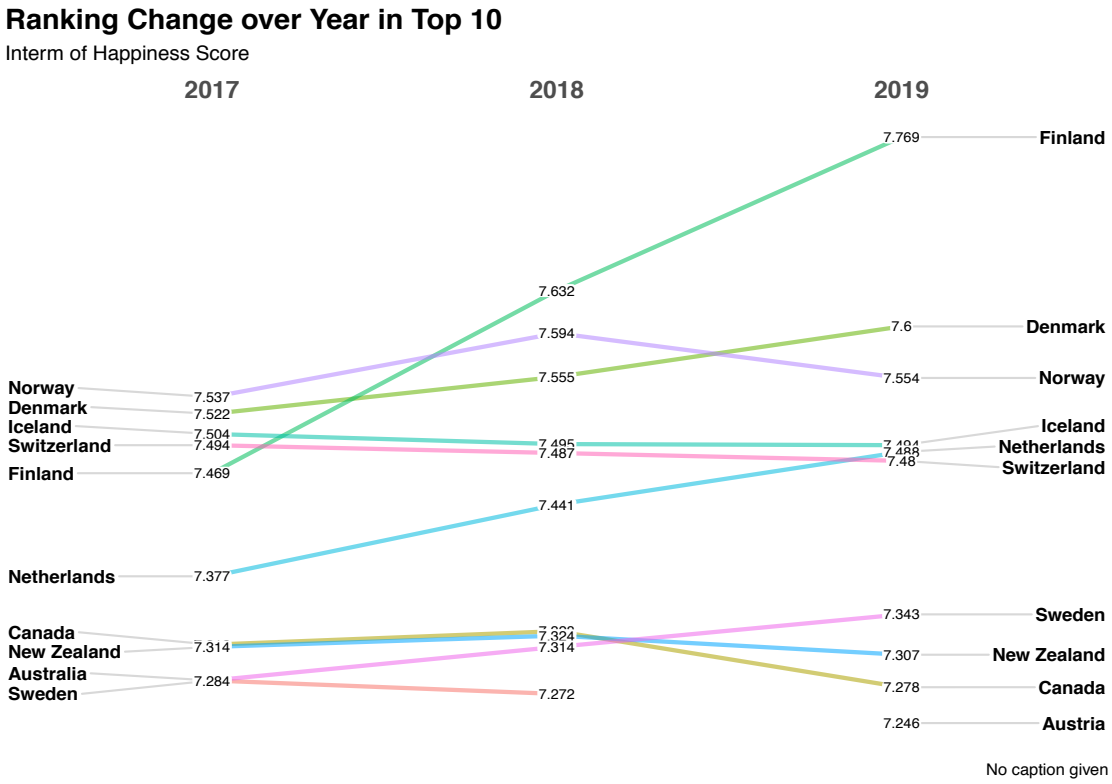
Also GDP per capita has strong correlation with Social Support and Healthy Life Expectancy. It would mean that citizens have more support from government and more access to medical utilities.



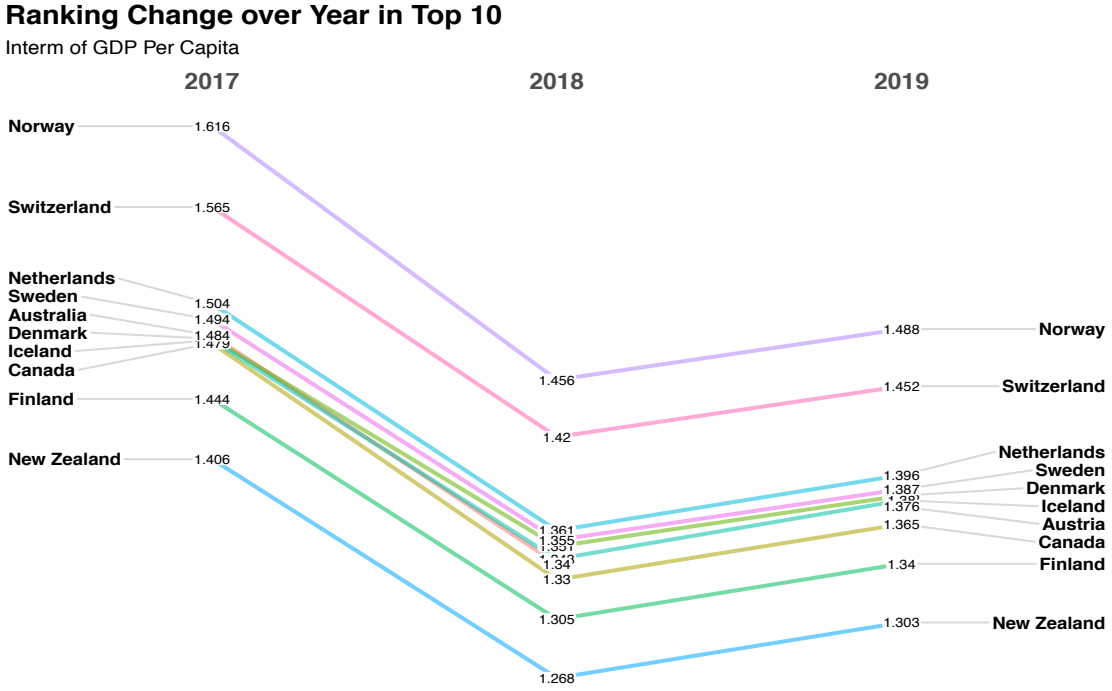
The graph bellows also give us a clear signal that the increase in GDP per capita will lead to the increase of Happiness Score



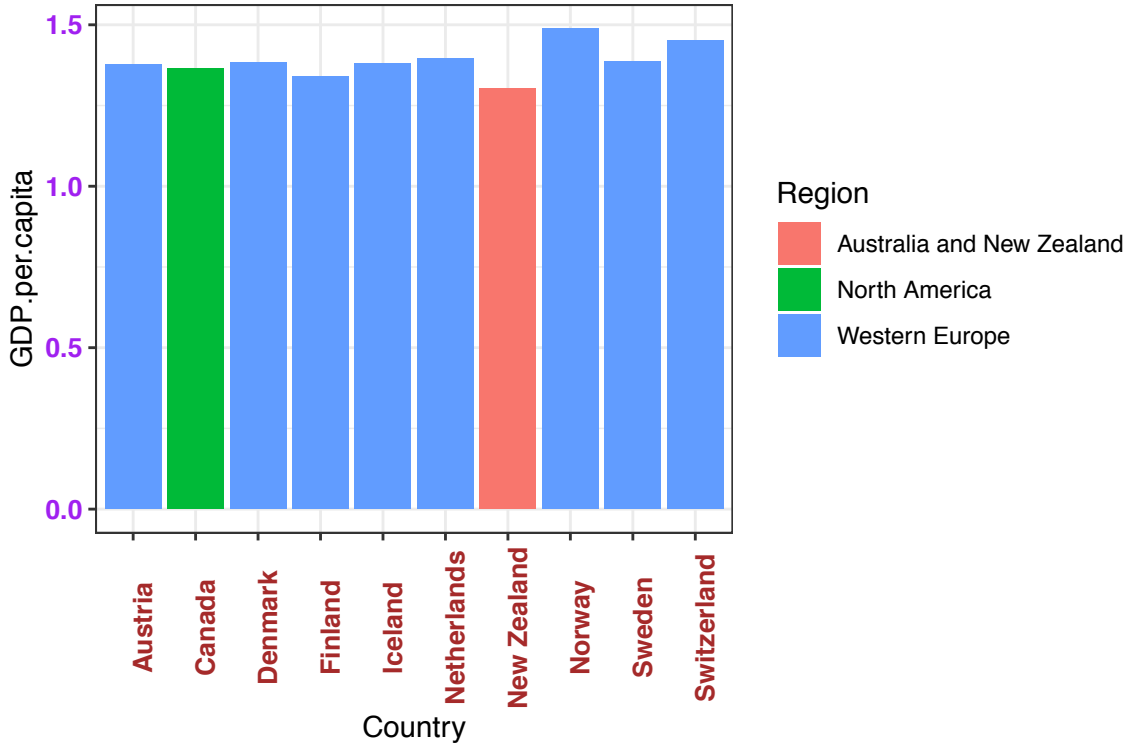
Check the change in top 10 in highest happiness ranking from 2017-2019



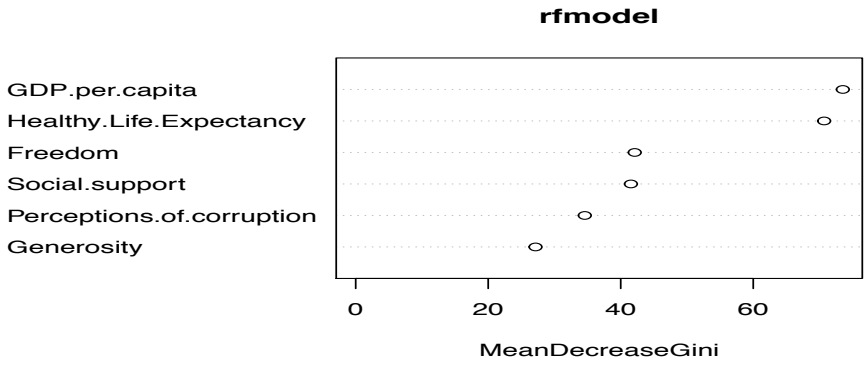
Finland had a very stable trend that moved from 5th from 2017 to the 1st position in 2019.
Australia was out of top 10 in 2019 and replaced by Austria.
Now we look at how GDP per Capita changing from 2017-2019



During the year 2017 to 2018, there were similar decline all over Europe and recovery starting from the beginning of 2018.
Again, familiar countries in the top 10 are still from Europe (8/10) even there are some countries with higher GDP are still not in the list of top 10 happiest countries such as: Qatar (1.684), Singapore (1.572) and United Arab (1.503). However, the majority of happiness countries are located in Europe



Thus, we build model to determine that GDP per capita really determine happiness using Rain Forest model. As we can see, the GDP per Capita is the most important factor



```
                Type of random forest: classification
                      Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 23.77%
Confusion matrix:
       Low Medium High class.error
Low     67     32    0   0.3232323
Medium  24    190   20   0.1880342
High     0     35   99   0.2611940
> confusionMatrix(test$Score,predict)
Confusion Matrix and Statistics

          Reference
Prediction Low Medium High
    Low     25      8    0
    Medium   6     66    6
    High     0     16   28

Overall Statistics

               Accuracy : 0.7677
                 95% CI : (0.6932, 0.8317)
    No Information Rate : 0.5806
    P-Value [Acc > NIR] : 8.088e-07

                  Kappa : 0.6148

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Low Class: Medium Class: High
Sensitivity              0.8065         0.7333       0.8235
Specificity              0.9355         0.8154       0.8678
Pos Pred Value           0.7576         0.8462       0.6364
Neg Pred Value           0.9508         0.6883       0.9459
Prevalence               0.2000         0.5806       0.2194
Detection Rate           0.1613         0.4258       0.1806
Detection Prevalence     0.2129         0.5032       0.2839
Balanced Accuracy        0.8710         0.7744       0.8456
> |
```

As the Accuracy is almost 77%, for instance, we can see in the prediction that: It predicts 25 countries with low Happiness Score but actually there are 8 countries are Medium, also they predict 28 High Happiness but only 12 in fact.

** Conclusion: GPD per capita is a leading factor that does not only influence the Happiness Score but also the other factors such as Healthy Expectancy, Social Support as well. And we can see that most of Europe countries have such high indexes. Thus, Europe would be a promising destination.