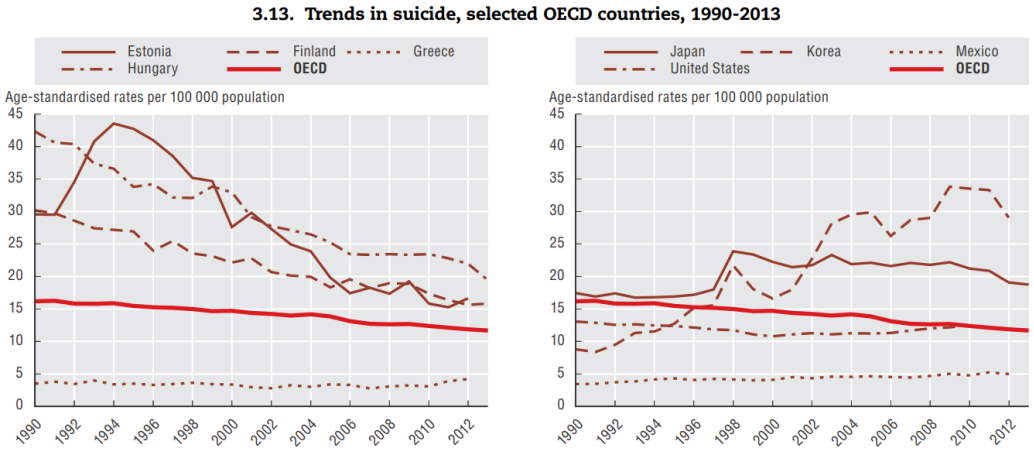
Final Report

Lu Lily
Jeon Yeeun
Son Kanghee

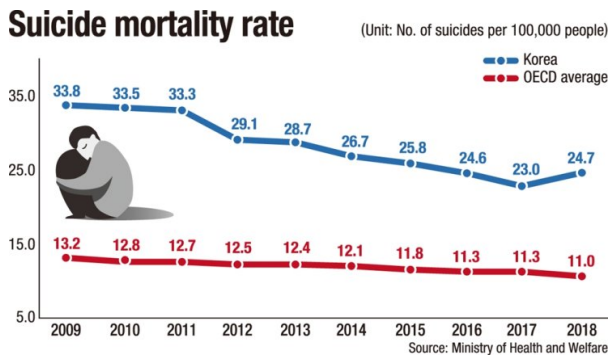
Introduction

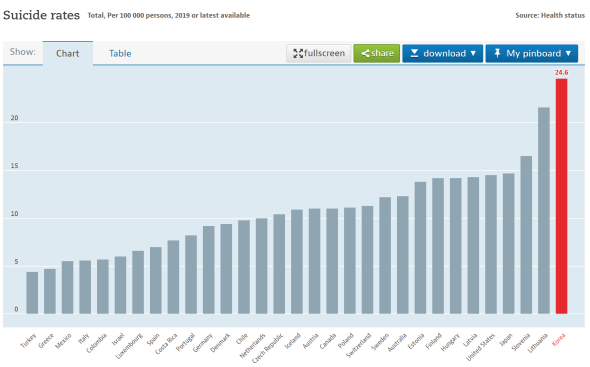
Korea has one of the highest suicide rates per capita of any developed country in the world. Since 2003, Korea has topped the suicide rate of all OECD (Organisation for Economic Co-operation and Development) countries almost uncontested except for 2016 and 2017 in which Korea's suicide rate was the second highest. The following graphs show Korea's suicide rates from 1990-2013 compared to the OECD average.



Source: OECD Health Statistics 2015, http://dx.doi.org/10.1787/health-data-en.

Although the suicide rate had been decreasing since 2009, the rate has gone up again as of 2018. The most recent data from 2019 also shows Korea topping the suicide rates with a rate that is very similar to the rate in 2018. The following two graphs show the suicide rate of Korea compared to the OECD average from 2009 to 2018, and the suicide rates of OECD countries in 2019.





There is much existing research on suicide trends based on demographic information and societal factors, but little research has been done on environmental and other external factors. Current existing research also does not often utilize Python for data analysis. Based on the assumption that increased happiness will lead to decreased suicide rates, we propose to use Python libraries and machine learning methods to find correlations between Happiness Scores and environmental and other external factors with the hope that more insight can be obtained into external risk factors for suicide that are currently going unnoticed. These results may be used to spur further research on ways to mitigate the effects of such external risk factors and to build a better understanding of what factors are most important in the happiness of a nation.

Literature Review

Our project is to find out the factors that can affect happiness, and what kind of effort it takes to make people “happier”. Based on our analysis, we aim to provide a direction for the future by finding ways for people to live happier lives in South Korean society.

Before starting our study, we had to find out about the current situation related to Korean people's happiness. Among them, we looked at a study [1] that analyzed the suicide rate of Koreans. In this study, as of 2017, suicidal thoughts and suicide attempts decreased, but the number of suicide deaths and suicide rates were rising again, and it was confirmed that the numbers were still high. However, this study did not study the factors that could potentially affect the suicide rate. It should be borne in mind that factors that can influence suicide can also affect individual happiness, so we should keep in mind that we should focus on the underlying cause rather than the change in people's happiness level itself. Next, we investigated various environmental [2], governmental [3][4], and economic factors [5][6] that can affect happiness.

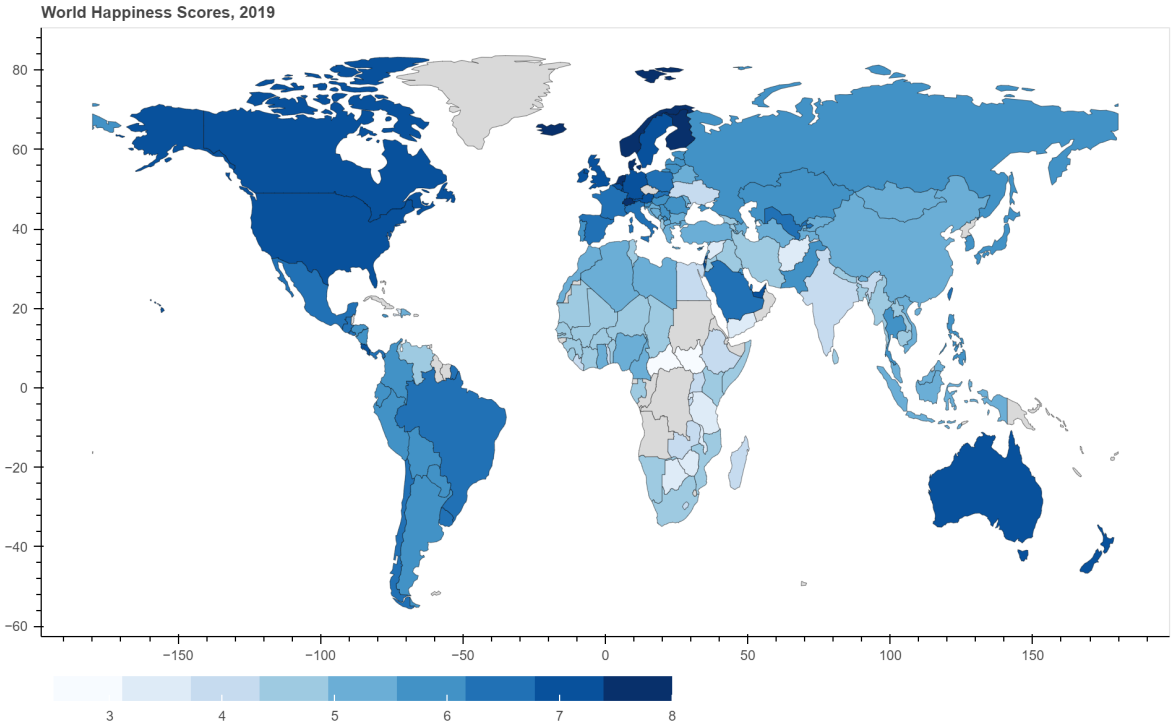
A study [2] focused on environmental factors that explained whether personal and environmental factors affect happiness. The purpose of the study was to clarify the relationship between overall happiness and life satisfaction. It argues that happiness consists of three related components: positive influence, absence of negative influence, and satisfaction with life as a whole. Therefore, we will be able to use factors that can affect life satisfaction and quality of life for analysis. And based on the claims of the study, we will be able to create a model that predicts the happiness index using positive and negative factors and life satisfaction in the analysis process we will proceed later.

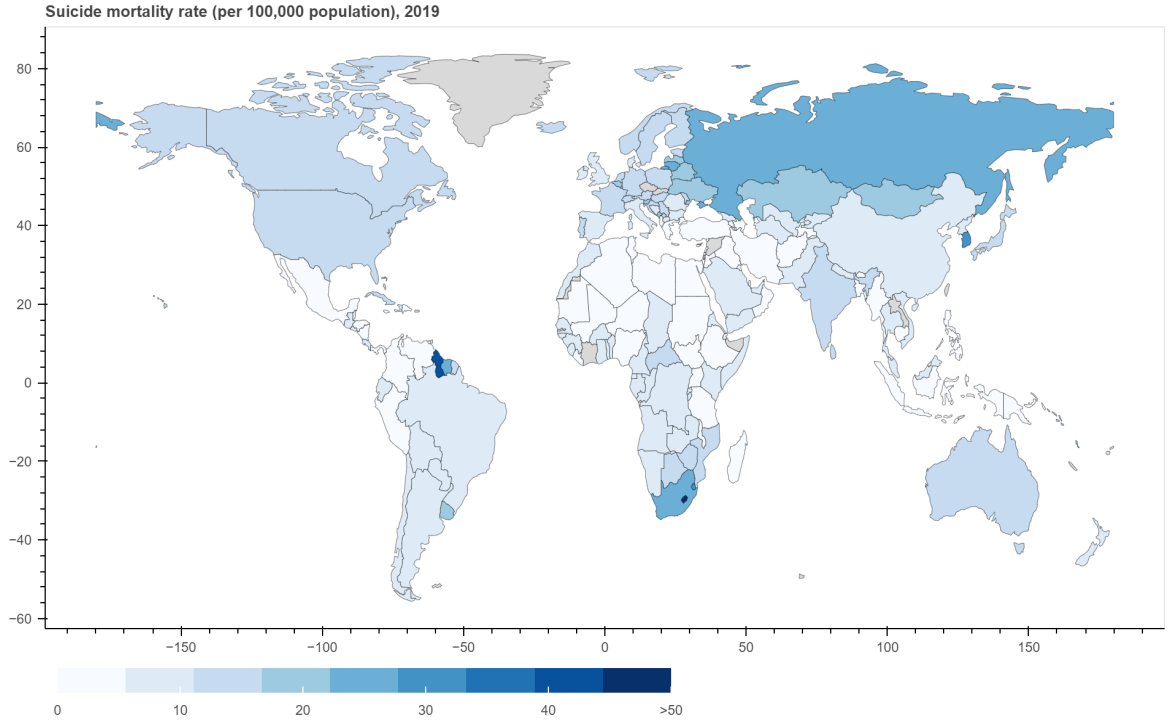
Another study [3] confirms that close relationships with people and good social infrastructure (public sports facilities, cultural facilities, green space creation, etc.) increase people's satisfaction. This could enter the realm of social help, and the content of relationships could be considered as a factor in quality of life. Above all, we can see that the government's support must be guaranteed to some extent in improving the quality of life, and this is important for people to be happy. In addition, we need to consider policy factors that the government can influence on individuals [4], such as living conditions in the country, social support(citizen security, social foundation for human relationships, public services, etc), and freedom. In this regard, one study [4] found that the effects of government actions on well-being act strongly on individuals, but are often difficult to separate from the effects of other events that occur simultaneously. Therefore, it is recommended to measure citizen satisfaction with the government in various areas of life. It is worth thinking about how the excellence of government can be measured and how the influence of government on happiness can be determined.

Finally, in several other studies [5] [6], the analysis of the effects of economic factors on happiness could be found. One of them [5] analyzed whether economic and cultural factors could be involved in happiness. In the needs theory, it was concluded that good living conditions for economic factors are the prerequisites for satisfying low and high needs. This is actually consistent with the empirical result that differences in happiness between countries are related to economic and cultural differences. However, in the previous comparative theory, a significant relationship between economic and cultural factors and happiness could not be found. Therefore, a follow-up study is needed to find out what factors are responsible for this difference. We need to confirm this as we conduct research. Another study [6] investigated the effect of per capita GDP (gross domestic product) on happiness considering the role of air pollution. The study found that even with an increase in GDP per capita, too high a level of air pollution can lead to a decrease in happiness, but a low level of air pollution can increase happiness. Therefore, we need to check whether the income fraction or GDP can have a sufficient effect on happiness when it is influenced by other factors.

Data Description

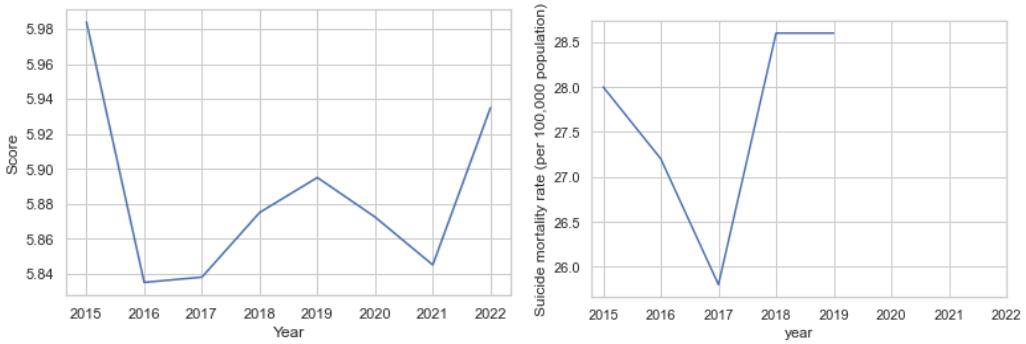
World map of happiness score & suicide rate





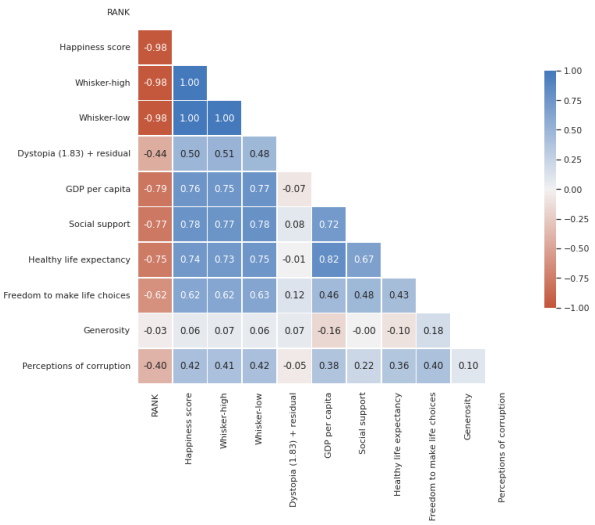
Using the World Happiness Report datasets ‘2015.csv’-’2022.csv’, we geographically mapped the happiness scores for each year. From the ‘suicide homicide gdp.csv’ dataset, we also geographically mapped suicide rates from 2015-2019. Based on the maps, the happiest countries are generally in the Northern America, Northern Europe, and Central Europe regions. Australia and New Zealand are also very happy. The most suicidal countries appear to be in North Asia and East Europe but there are some major outliers in South America and Africa. The suicide map looks somewhat correlated with the happiness map where some of the happier countries are also more suicidal.

Trends of Koreans' happiness scores and suicide rates



Using the World Happiness Report datasets '2015.csv'-'2022.csv' and the 'suicide homicide gdp.csv' dataset, we selected out only the data for South Korea and graphed the trend from 2015 to 2022. The 'suicide homicide gdp.csv' dataset ends in year 2019. From the data, South Korea was happiest in 2015 and had a downward trend from then to 2016. From 2016-2019 there was an upward trend in happiness score. From 2019-2021 the happiness score trended down again. Since 2021, the happiness score has been on the rise but it still falls short of the 2015 score. The suicide rates trended downward from 2015-2017 and rose again from 2017 onward. This trend is quite similar to the happiness score trend, though the suicide rates shot up much higher after 2017 compared to the happiness score.

Scale of Pearson correlation coefficient	Interpretation
$0.00 \leq r \leq 0.19$	Very low correlation
$0.20 \leq r \leq 0.39$	Low correlation
$0.40 \leq r \leq 0.59$	Moderate correlation
$0.60 \leq r \leq 0.79$	High correlation
$0.80 \leq r \leq 1.00$	Very high correlation



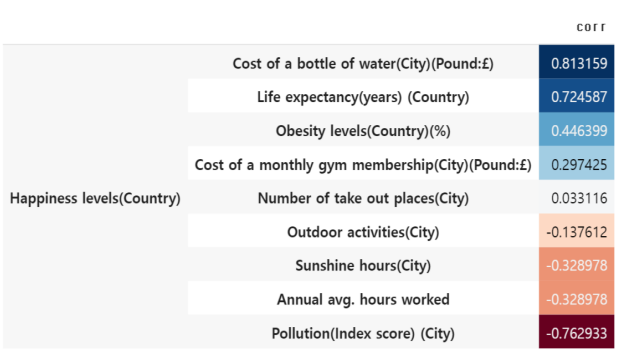
Correlation between Happiness and Three Main Factors

Using the '2022.csv' dataset, we looked at the overall correlation between happiness score and other features. Dystopia, Freedom, Perceptions of corruption are elements of national policy. Besides, GDP per capita is an economic factor, and healthy life expectancy, generosity, and social support(relationship) can be considered as an environmental factor.

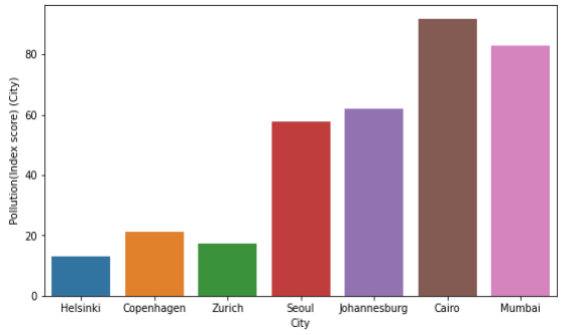
GDP, social support, and life expectancy have a high correlation of more than 0.7. Freedom is 0.62, which is highly correlated with happiness score. Through this result, we can see a certain amount of economic stability is required for happiness, welfare factors for basic living, and health have an important influence on happiness. Therefore, in the next analysis, it is necessary to look at the healthy living environment and economic aspects in more detail. Also, since the perception of corruption is subjective, it is necessary to confirm it by examining other data on corruption.

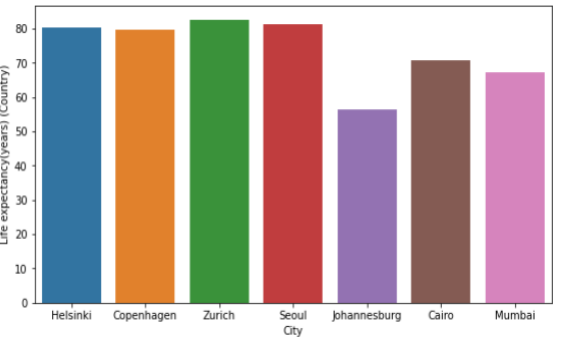
Correlation between Happiness and Environmental Factors

Using the 'healty_lifestyle_city_2021.csv' dataset, we tried to examine the influence of personal lifestyle and environment on happiness. The cost of a bottle of water is an indicator of the price of daily necessities.



Cost of a bottle of water has a 0.81 correlation and Life expectancy has a 0.72 correlation. These two elements have very high positive correlations with happiness. Pollution correlation is -0.76, which has a very high negative correlation. This suggests that an increase in life expectancy has a positive effect on happiness. It shows that an individual's healthy lifestyle does not always have a positive effect on happiness. However, factors that can be fatal to health and society, such as environmental pollution, are negative to happiness. And it is necessary to find out what changes have occurred in economic growth, standard of living, or income and to examine whether such changes affect happiness in relation to economic factors.





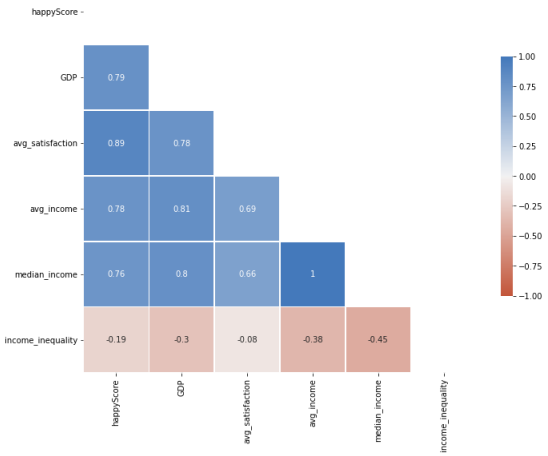
City	Happiness levels(Country)
Helsinki	7.80
Copenhagen	7.64
Zurich	7.56

City	Happiness levels(Country)
Seoul	5.87

City	Happiness levels(Country)
Mumbai	3.57
Cairo	4.15
Johannesburg	4.81

Then we selected the top 3 countries, the bottom 3 countries, and Korea for the happiness index within the data we have. It shows whether high and low happiness index is related to pollution and high life expectancy, and what kind of relationship it has in Korea. The difference between the top 3 countries and the bottom 3 countries is evident in the degree of pollution, and there is a difference in life expectancy, but it is not as clear a difference as pollution. We can see that Seoul has high pollution and life expectancy.

Correlation between Happiness and Economic Factors & Satisfaction



corr

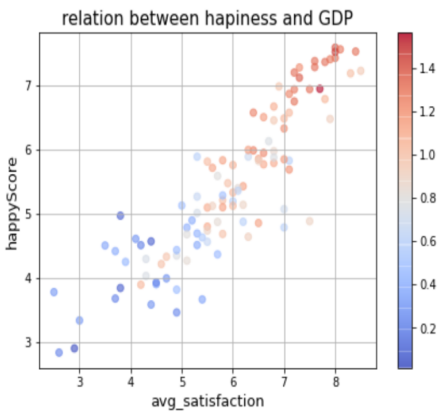
happyScore	avg_satisfaction	0.890000
	GDP	0.790000
	avg_income	0.780000
	median_income	0.760000
	income_inequality	-0.190000

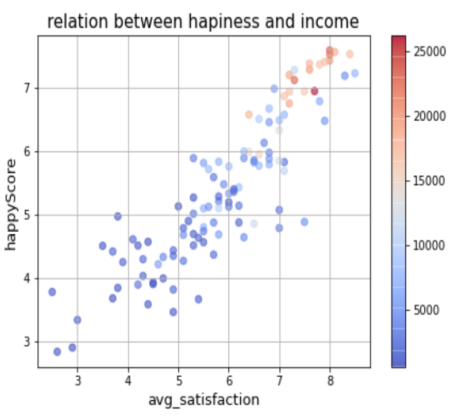
Using the 'happyscore_income.csv' dataset(2021), we did another EDA to see the correlation between happiness and economic factors. We select avg_satisfaction, avg_income, median_income, income_inequality, happyScore and GDP.

Satisfaction has 0.89 correlation, which has the highest positive correlation with the happiness score, and GDP and income factors also have high positive correlations with the happiness score. Also, there is a close correlation between each element from the left corr plot.

Relationship Between Satisfaction and Happiness & Distribution of GDP and Income

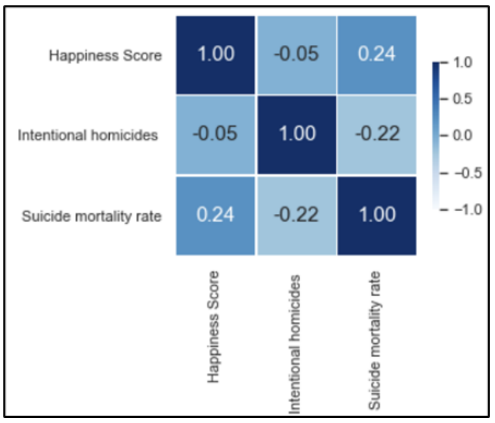
In the GDP scatter plot, people with a high index generally feel happier than those with a low index, but the degree of distribution is not extreme. However, in the income scatter plot, when the average income is 15000 or more, people feel significantly higher happiness than when the average income is lower than that. From the above two graphs, it can be seen that the higher the GDP and income, the higher the happiness score and satisfaction, and there is a linear relationship between them.



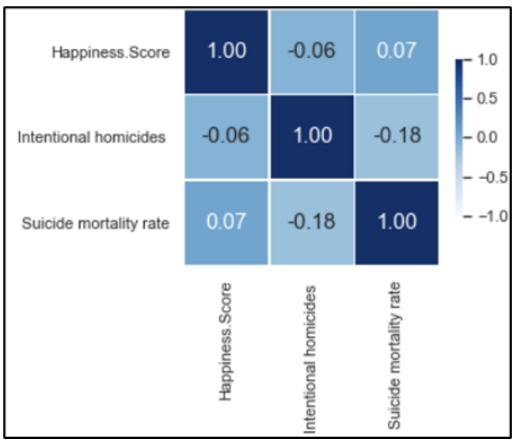


Correlation between Happiness and humicide, suicide rate

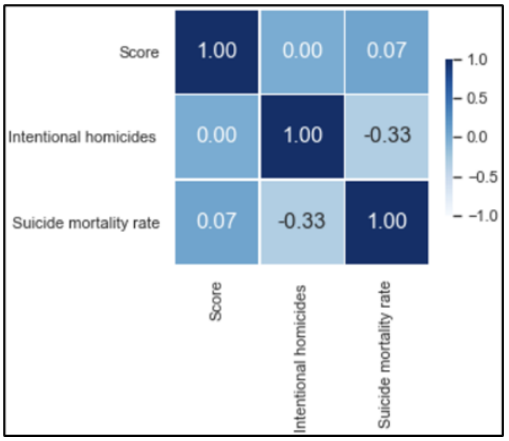
<2015>



<2016>



<2017>



<2018>

In the 'suicide.csv' dataset, there are information such as homicide rate (Intentional homicides per 100,000 people) and suicide rate(Suicide mortality rate per 100,000 population) of each country. Based on rough assumption that suicide rate will be related with happiness, we looked for relationship between Happiness Score and homicide rate and suicide rate. 4 years(2015-2018) of dataset were overlapping between ‘suicide.csv’ dataset and dataset with happiness scores. In 4 years, the correlation between happiness score and international homicide rate was in the range-0.06~0, so there was very low negative correlation. The correlation between happiness score and suicide mortality rate was in range 0.07~0.24, so we could also find low correlation. Before doing EDA, we thought that suicide rate will have high correlation to happiness, but it was interesting to find out that there’s low correlation, different from our initial thoughts.

Correlation between Happiness and CPI Score

In the ‘history.csv’ dataset, there is information of CPI(Corruption Perceptions Index) Scores of each country. CPI Scores are based on how corrupt a country’s public sector is perceived to be. (100 is very clean and 0 is highly corrupt.)

2 years(2015-2016) of dataset were overlapping between ‘history.csv’ dataset and dataset with happiness scores. The correlation between CPI Scores and happiness scores were 0.670004 in 2015 and 0.67599 in 2016. So we can say that there is high correlation between CPI.

We can say that as the country’s public sector is perceived to be more cleaner, then more happiness people feel.

Analysis & Results

Regression model and classification model

```
--ML Model Output--

Decision Tree RMSE:: 0.96 (+/- 0.21)
Decision Tree Expl Var: -13.69 (+/- 27.46)
CV Runtime: 0.019981861114501953


Random Forest RMSE:: 0.75 (+/- 0.56)
Random Forest Expl Var: -4.41 (+/- 10.80)
CV Runtime: 0.6092503070831299


Gradient Boosting RMSE:: 0.74 (+/- 0.43)
Gradient Boosting Expl Var: -6.14 (+/- 16.61)
CV Runtime: 0.2599925994873047


Ada Boosting RMSE:: 0.78 (+/- 0.55)
Ada Boosting Expl Var: -5.71 (+/- 16.33)
CV Runtime: 0.704174280166626


Neural Network RMSE:: 1.11 (+/- 1.39)
Neural Network Expl Var: 0.01 (+/- 0.05)
CV Runtime: 0.02612781524658203
```

```
--ML Model Output--

Decision Tree Acc: 0.80 (+/- 0.13)
Decision Tree AUC: 0.84 (+/- 0.13)
Decision Tree F1:0.78 (+/- 0.15)
CV Runtime: 0.14917874336242676

Random Forest Acc: 0.84 (+/- 0.10)
Random Forest AUC: 0.95 (+/- 0.07)
Random Forest F1:0.84 (+/- 0.08)
CV Runtime: 1.9644887447357178

Gradient Boosting Acc: 0.82 (+/- 0.13)
Gradient Boosting AUC: 0.91 (+/- 0.09)
Gradient Boosting F1:0.81 (+/- 0.16)
CV Runtime: 2.9386110305786133

Ada Boosting Acc: 0.72 (+/- 0.26)
Ada Boosting AUC: 0.92 (+/- 0.05)
Ada Boosting F1:0.68 (+/- 0.37)
CV Runtime: 2.261735677719116

Neural Network Acc: 0.56 (+/- 0.26)
Neural Network AUC: 0.76 (+/- 0.22)
Neural Network F1:0.37 (+/- 0.29)
CV Runtime: 2.0405333042144775
```

Based on what we have learned in class, we selected features by wrapper method and made regression models of Decision tree, random forest, gradient boosting, ada boosting, and neural network. Also, we discretized the target into bins and then treated it as a classification problem and then built various classification models.

For all of the models of regression, we thought that RMSE is high and explained variance is low, which means it is not a good model. In classification models, we calculated the accuracy, AUC, and F1 Score(the harmonic mean of recall and precision) of each model. Both Acc (except in Neural Network) and AUC were high. F1 Score is high in the Decision Tree, Random Forest, and Gradient Boosting but low in the Ada Boosting and Neural Network.

Therefore, we can see that the classification model obtained better results than the regression model. And we confirmed that random forest is better among classification models.Because we were not satisfied with the results of 5 regression models, we decided to build a new model, a multiple linear regression model.

Multiple linear regression model

In EDA, we calculated the correlation between happiness score and some features. Based on Pearson correlation coefficient, we selected all the features that are highly(Pearson correlation coefficient: 0.6~0.79) or very highly (Pearson correlation coefficient: 0.8~1.0) correlated with happiness score.

Scale of Pearson correlation coefficient	Interpretation
$0.00 \leq r \leq 0.19$	Very low correlation
$0.20 \leq r \leq 0.39$	Low correlation
$0.40 \leq r \leq 0.59$	Moderate correlation
$0.60 \leq r \leq 0.79$	High correlation
$0.80 \leq r \leq 1.00$	Very high correlation

Correlation	Features
Very high correlation 0.80<r<1.00	Whisker-high, whisker-low, Cost of a bottle of water, average satisfaction
High correlation 0.60<r<0.79	GDP, Social Support, Freedom, CPI Score, life expectancy, pollution, GDP, average income
Moderate correlation 0.40<r<0.59	Dystopia+residual, obesity level
Low correlation 0.20<r<0.39	Cost of a monthly gym membership, sunshine hours, annual average hours work
Very low correlation 0.00<r<0.19	Generosity, number of take-out places, outdoor activities, income inequality

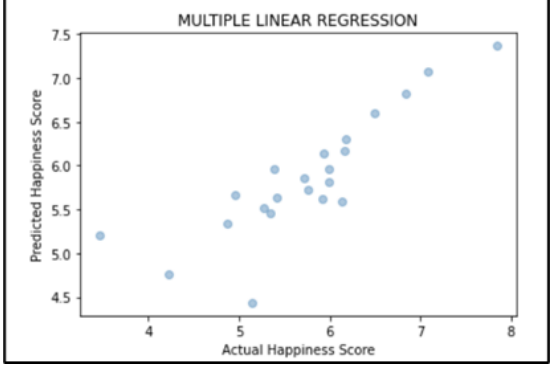
Features ‘average satisfaction’ and ‘cost of a bottle of water’ were very highly correlated, and ‘GDP’, ‘average income’, ‘social support’, ‘life expectancy’, and ‘freedom to make choice’ were highly correlated with happiness scores.When building a model, we did not consider ‘whisker-high’ and ‘whisker-low’ because they were not features that affects happiness scores. Also, we couldn’t consider ‘cost of a bottle of water’, and ‘life expectancy’ because they were in the dataset of cities (other features were in dataset of countries), and the dataset was relatively small compared to other datasets. We merged all the other features into a single dataset so that we could easily build a multiple linear regression model.With these features, we made a multiple linear regression model, so that if we know the features, then we can predict the happiness score.

```
In [45]:
start = time.time()

x = df6[['Freedom to make life choices', 'Social support','CPI  2021 Score','avg_satisfaction', 'GDP', 'avg_income']]
y = df6[['Happiness score']]

x_train, x_test, y_train, y_test = train_test_split(x, y, train_size=0.8, test_size=0.2)

mlr = LinearRegression()
mlr.fit(x_train, y_train)

print(mlr.score(x_train, y_train))
print("runtime :", time.time() - start)
```

```
0.8441995656240189
runtime : 0.019191741943359375
```

<multiple linear regression model with feature selection>

```
y_predict = mlr.predict(x_test)
import matplotlib.pyplot as plt

plt.scatter(y_test, y_predict, alpha=0.4)
plt.xlabel("Actual Happiness Score")
plt.ylabel("Predicted Happiness Score")
plt.title("MULTIPLE LINEAR REGRESSION")
plt.show()
```



We built the model, and with the .score() method, we calculated the accuracy. We were able to get a coefficient of determination(R^2): 0.8442, and generally over 0.7 is good enough. The runtime took 0.019

Also we could find out that the model was inaccurate with low happiness scores.

To see if feature selection was meaningful, we made another model considering all the features regardless of correlation with happiness score.

The accuracy of feature selected model was lower (0.844<0.999), but also the runtime was shorter(0.019<0.027), so that we could see that we made parsimonious model with feature selection.

```
In [51]:
start = time.time()

x1 = df8[['Standard error of ladder score','upperwhisker', 'lowerwhisker', 'Logged GDP per capita','Social support',
          'Healthy life expectancy','Freedom to make life choices','Generosity','Perceptions of corruption',
          'Ladder score in Dystopia','Explained by: Log GDP per capita','Explained by: Social support',
          'Explained by: Healthy life expectancy','Explained by: Freedom to make life choices','Explained by: Generosity',
          'Explained by: Perceptions of corruption','Dystopia + residual','CPI  2021 Score','adjusted_satisfaction',
          'avg_satisfaction','std_satisfaction','avg_income','median_income','income_inequality','GDP']]
y1 = df8[['Happiness score']]

x1_train, x1_test, y1_train, y1_test = train_test_split(x1, y1, train_size=0.8, test_size=0.2)

mlr = LinearRegression()
mlr.fit(x1_train, y1_train)

print(mlr.score(x1_train, y1_train))
print("runtime :", time.time() - start)
```

```
0.9999999126914741
runtime : 0.027480363845825195
```

<multiple linear regression model considering all the features>

```
feature_values = [[0.95, 0.95, 90, 9, 1.5,20000]]
predict = mlr.predict(feature_values)
print(predict)
```

```
[[7.7779057]]
```

With this model, if we know the value of features, we can predict the happiness score. For example if freedom to make life choices is 0.95, social support is 0.95, CPI Score is 90, average satisfaction is 9, GDP is 1.5, and average income is 20000, then we can predict that the happiness score is 7.7779057. So, we can use this model to see how the happiness score will increase or decrease as the values of other features change.

Conclusion

The results show that the features we selected are highly positively correlated and fairly accurate predictors for happiness score. Although we wished to find ways to decrease suicides, we had the false premise of assuming increased happiness would decrease suicides. Even so, Korea is around the middle range of happiness scores. Therefore, we want to switch our focus to improving these factors in order to improve overall happiness and life satisfaction in Korea, and to save solving the suicide problem for the future. We thought about some ways to increase happiness in Korea based on the features that are very highly or highly correlated with the happiness.

How to increase CPI score, average satisfaction
The Corruption Perceptions Index is an indicator for scoring and ranking countries/regions on corruption perceptions. The score represents the perceived level of public sector corruption on a scale from 0 to 100, where 0 means very corrupt and 100 means very clean.
Corruption can be mainly caused by socio-cultural factors such as kinshipism and return culture (customs). In order to reduce corruption, the improvement of the bribery culture should be prioritized, and the awareness of fairness should be universal.
In Korea, there are currently laws related to bribery and the Anti-Corruption and Civil Rights Commission, an anti-corruption authority. As a national effort, it is necessary to clearly define the range of acceptable gifts, to continuously check whether there are any problems with the current law, and to revise the deficient parts. The Anti-Corruption and Civil Rights Commission should be politically independent and neutral, and should be empowered to conduct investigations. In addition, it is necessary to promote fairness and anti-corruption to the public nationwide.

Education and promotion on this should be conducted regularly within a company or organization. In addition, a ‘multiple regression analysis model’ can be created to examine the relative importance of factors that determine job satisfaction by periodically conducting a survey on employee satisfaction. Based on the results, it is possible to identify factors that cause dissatisfaction and stress in the organization and job, and come up with a plan to improve and increase satisfaction. Satisfaction with the work environment can be increased, and corruption caused by dissatisfaction and stress can be reduced.

How to increase freedom, social support
According to the dataset, social support is the national average of the binary responses (0=no, 1=yes) to the Gallup World Poll (GWP) 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 to make life choices is the national average of binary responses (0=no, 1=yes) to the GWP question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”
To increase social support, the government should make more social assistance programs and individuals should try to build more healthy relationships with others. Also, people should try to live altruistic lives rather than egoistic lives.

According to Human Freedom Index, in South Korea, freedom indexes such as ‘Freedom of Foreign Movement’, ‘Media Freedom’, ‘Criminal Justice’, ‘Legal Gender’ were low. To increase the freedom of foreign movement, the Korean government should increase freedom to leave the country. To increase media freedom, we should widen the extent of censorship and self-censorship among journalists and the press and put less political pressure on media. To increase freedom of ‘criminal justice’, the criminal justice system should be less corrupt. To increase the freedom of legal gender, people should be more free to legally change their sex and gender.
0.89,0.5

How to increase GDP, avg_income
GDP is a measure of the market value of all the final goods and services produced in a country during a specified year.
Average Income is the mean income earned per person in a given country in a specified year.

Both the GDP and average income can be increased if the Korean government focuses on producing and promoting exports. The government should invest in infrastructure that allows Korea to be more productive. High quality education, training, and the enforcement of strong labor laws and workers’ rights will also increase the average income.

Future Work

One of the initial purpose of our project was to find a way to lower suicide rate in Korea by making people more happy. However, after we did EDA, we could find out that suicide has low correlation with happiness, which is different from what we have thought. Therefore, a follow-up study and data are needed to find out what kind of effort is needed to lower the suicide rate in Koreans. We want to look into datasets with information about individuals who have attempted or committed suicide rather than generalized information about a country.
Also, there can be more features that are related to happiness. So, we can follow-up studies of individual factors that may be related to happiness. We can also try to find more features in new datasets that may be related to happiness.

Reference

[1] Kim, S. H., Lee, D. W., Kwon, J., Yang, J., Park, E. C., & Jang, S. I. (2021). Suicide Related Indicators and Trends in Korea in 2019.
https://www.koreascience.or.kr/article/JAKO202119061986694.page
[2] Lu, L. (1999). Personal or environmental causes of happiness: A longitudinal analysis. *The Journal of Social Psychology*, *139*(1)
https://www.proquest.com/docview/199795937?pq-origsite=gscholar&fromopenview=true
[3] Kim, M., & Im, H. N. (2020). The Relationship between Social Infrastructure and Happiness. *Journal of Korea Planning Association-Vol*
http://kpaj.or.kr/_common/do.php?a=current&bidx=2091&aidx=25223
[4] John F. Helliwell, Richard Layard and Jeffrey D. Sachs (2019). World Happiness Report
https://media-01.imu.nl/storage/heart4happiness.nl/2097/wp/2019/03/WHR19.pdf#page=13
[5] Schyns, P. (1998). Crossnational differences in happiness: Economic and cultural factors explored.
https://www.proquest.com/docview/1308072168/fulltextPDF/7B85670C20EE4658PQ/1?accountid=11283
[6] Fotourehchi, Z., & Ebrahimpour, H. (2019). Happiness, economic growth and air pollution: an empirical investigation.
https://www.inderscienceonline.com/doi/abs/10.1504/IJHD.2019.098047

Google drive link of our work(EDA, Analysis)
You can see the details of the rest of our work that are not in this final report from this link.
https://drive.google.com/drive/folders/1caPxSJWFMuutcrrFTLMiV4ANYiv_MaUM?usp=sharing