World Happiness and Democracy

Introduction

Everyone is looking for happiness. Then what factors influence man's feeling about happiness? Great income, healthy body, perfect family, enough freedom, generous neighbourhood and community, incorruptible governance, democratic freedom... It looks every individual factor does not define happiness directly but to some extent they play their roles in the process of how to feel happiness.

Thank the United Nations Sustainable Development Solutions Network for providing World Happiness Report annually. Let's have this opportunity of Quantitative Analysis on World Happiness.

The UN World Happiness Reports have an index called as national happiness score as well as six correlative life factors, which comprise a system of indicators to rank happiness across the world. The rankings of national happiness are based on a Cantril ladder survey. Nationally representative samples of respondents are asked to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale.

Based on explorary data analysis of world Happiness Reports we would like to answer a question: Is the world a happy place for people? Moreover, in the past half year, the people of Hong Kong have shown a strong spirit of pursuit for more democratic freedoms through widespread protests. These events led us to wonder if there's a strong relationship between democracy and happiness.

1. Data Cleaning

4E Rules

When facing the five World Happiness Report datasets, I remembered the three interrelated rules which make a dataset tidy: 1. each variable must have its own column; 2. each observation must have its own row; 3. each value must have its own cell.(https://r4ds.had.co.nz/tidy-data.html#fig:tidy-structure)

Before starting on this project These rules were just what I followed when cleaning dataset. But now after the excited jouney of data cleaning for this project, I add the fourth rule: each tidy dataset must keep original information as accurate and complete as possible.

Decide tidy dataset columns

Firstly I found there were 9 common variables in five datasets: rank, country, score, gdp per capita, social support, life expectancy, freedom, generosity, trust. First three of them delivered informations about countries' names and their happiness rankings, the other were six key explanatory factors. Nevermind they had differnet names in different reports sometimes, I could easily find they were pointing to the same variables. For example, in 2015 report dataset there was a variable "Trust (Government Corruption)",then in 2018 and 2019 report datasets a variable "Perceptions of corruption" has taken the place. Because these variables had necessary informations to answer our questions I'd like to use the nine variables to create new tidy data frame.

However in 2015,2016 and 2017 reports datasets variable "family" came out instead of "social support". I seached the official website for this variable(https://worldhappiness.report/). There was only social support in the documentations throughout five years reports. So I treated "family" as input error and replaced it with "social support".

To analyze trend of happiness among countries in the five years, I added time dimension with mutating column "year" .

When we have year and score columns, it's easy to caculate the ranking sequence for each country per year. Then I dropped rank column off. Now nine columns again!

Missing values and input errors

After importing and reviewing the five original datasets in .csv type, I did NOT see any N/A value at all. When joining the datasets, I met a problem . A warning message came out like this : "Error: Column `trust` can't be converted from numeric to factor." I had to check "trust" column in every dataset. Only 2018.csv had "trust" column(origin was "Perceptions.of.corruption") treated as factor. Other "trust" columns in .csv files had dbl attribute. Keeping forward, I carefully reviewed any element of this column until meeting a N/A. This observation has country name as United Arab Emirates with happiness rank 20 in 2018. An opinion jumped to my mind that please drop the observation off as it was a missing value.

I did not follow it, as I knew, if doing like that, other four observations marching this country name from other four datasets would be cut off as well(reason will be followed soon). Moreover there were two results followed: 1. this analysis report would be based on incomplete data, which means the possibility of wrong analysis conclusion goes up; 2. If some readers of this report were people from United Arab Emirates, I am absalutely sure they would be strongly unsatisfied. So I dicided to refer to original data. Below are two observations: Observation A20(from 2018.csv): 20,United Arab Emirates,6.774,2.096,0.776,0.670,0.284,0.186,N/A Observation B20(from https://s3.amazonaws.com/happiness-report/2018/WHR2018Chapter2OnlineData.xls): 20,United Arab Emirates,6.774,1.467,1.296,0.776,0.670,0.284,0.186

We can see observation A20 added value of "2.096" at the positon of "1.467,1.296", then again "N/A" at the end.

Story just began. In 2018.csv, observation A19(happiness rank 19 in 2018) was :19, Israel, 6.814, 1.301, 1.559, 0.883, 0.533, 0.354, 0.272

However B19(happiness rank 19 in 2018, from https://s3.amazonaws.com/happiness-report/2018/WHR2018Chapter2OnlineData.xls) was: 19,United Kingdom,6.814,1.301,1.559,0.883,0.533,0.354,0.272 The different country names were so obvious that everyone would realize them when checking A20 and B20. After comparing the values of happiness ranks and scores from 2017, 2018 and 2019, I exchanged the names of two countries in 2018.csv.

I found some 0's presenting the values of a couple of key explanatory factors. They might be missing or not. According to the documentations, 0 represented the worst possible life in the Cantril ladder survey. It was an option and had possibility. Because I did not catch up with a better way, I kept those 0's. Luckily I did not find any happiness score with the value of 0. Just a joke!

Dropped those countries which did not existed in all five years' datasets

For keeping our analysis in a continous time dimension, I took out of those observations whose countries' names did not have five years' records. However there were a couple of exceptions. "Hong Kong" and "Taiwan" existed in 2015, 2016,2018 and 2019 reports except 2017, because the names of "Hong Kong S.A.R., China" and "Taiwan Province of China" have taken them in this year report. I restored the names of "Hong Kong" and "Taiwan" to 2017 dataset.

Because having the background of Asian Cultrue, I easily recognized the situation of missing Hong Kong and Taiwan. Forward same things might happened to other countries whose names in those reports changed from 2017 to 2019. For verifying my hypothesis, I compared 2017 dataset to 2015 and 2019 datasets seperately, then two countries had been found as "North Cyprus" and "Trinidad and Tobago".

2. Is the World happy?

It is hard to say yes or no for this question. However we can figour out that a country feels happy when he has happiness scores over 5.

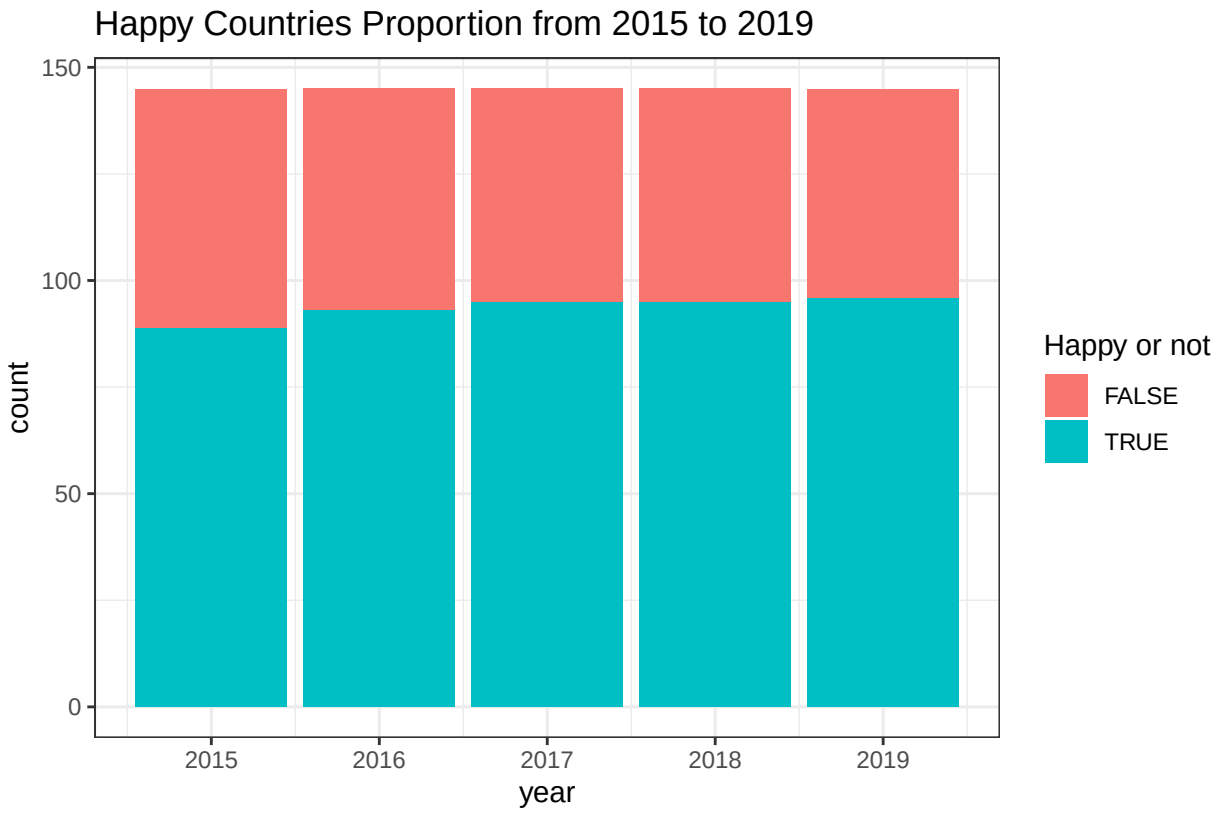

Figure 2.1

Figure 2.1 tells us around 2/3 countries felt happy in last five years while the number of happy countries increased a little bit. It looks like that we had "yes" to this question. However the limit of this analysis was number of countries, and not number of people. If adding the weight of population among different countries, we might see different answer.

For analysis of relationship between happiness score and year, we use linear model to get the data of predictions and residuals.

```
## # A tibble: 725 x 11
##    country score   gdp socsupport  lexp freedom trust generosity year
##    <chr>   <dbl> <dbl>      <dbl> <dbl>   <dbl> <dbl>      <dbl> <chr>
##  1 Switze~  7.59  1.40       1.35 0.941   0.666 0.420      0.297 2015
##  2 Switze~  7.51  1.53       1.15 0.863   0.586 0.412      0.281 2016
##  3 Switze~  7.49  1.56       1.52 0.858   0.620 0.367      0.291 2017
##  4 Switze~  7.49  1.42       1.55 0.927   0.66  0.357      0.256 2018
##  5 Switze~  7.48  1.45       1.53 1.05    0.572 0.343      0.263 2019
##  6 Iceland  7.56  1.30       1.40 0.948   0.629 0.141      0.436 2015
##  7 Iceland  7.50  1.43       1.18 0.867   0.566 0.150      0.477 2016
##  8 Iceland  7.50  1.48       1.61 0.834   0.627 0.154      0.476 2017
##  9 Iceland  7.50  1.34       1.64 0.914   0.677 0.138      0.353 2018
## 10 Iceland  7.49  1.38       1.62 1.03    0.591 0.118      0.354 2019
## # ... with 715 more rows, and 2 more variables: region <chr>, resid <dbl>
```

Table 2.2

Then I use boxplot to draw happiness score. 2017 is the lowest, but the overall trend is an upward tendency.

The graph of residuals. The years of 2018 and 2019 is above 0.

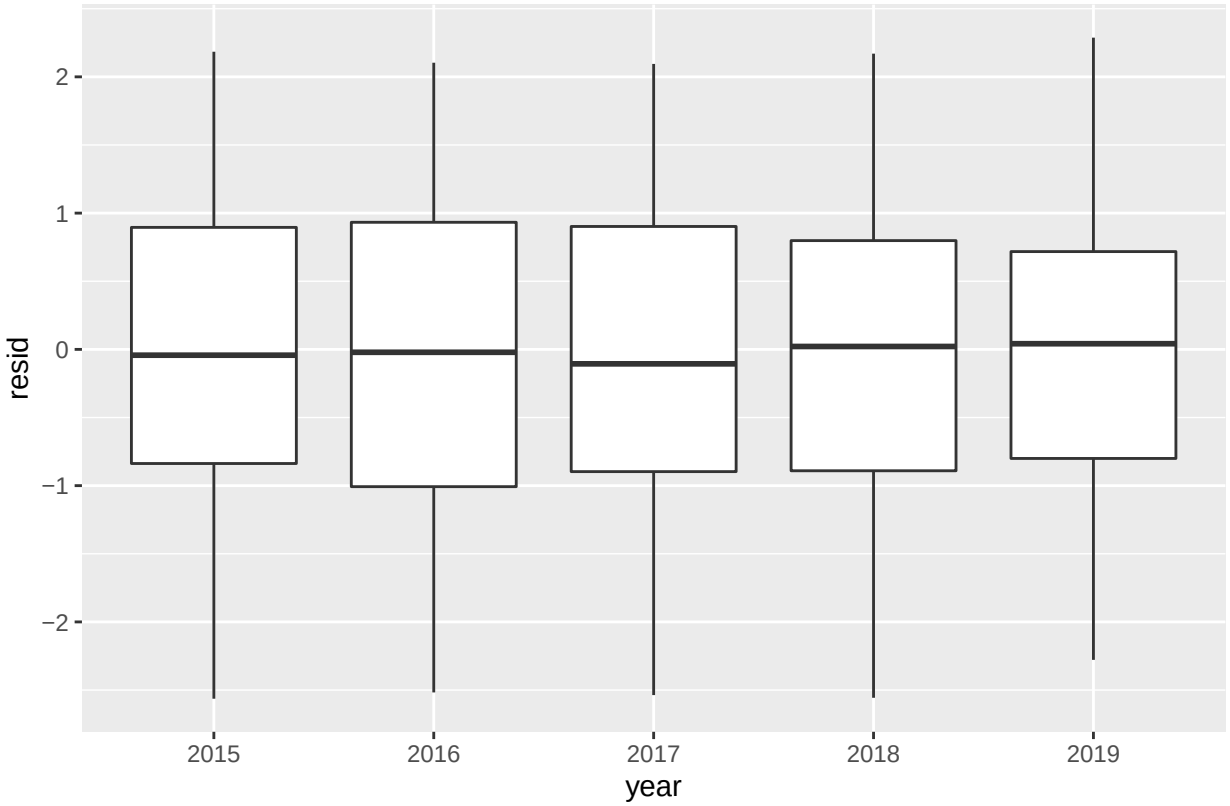

Figure 2.3

Then I separate the countries into different Regions to check each Continent's happiness score Continent. North Amrica, Australia and NZ are generally high, but their score drop down a little bit during the 5 years.Western Europe and Sub-Saharan Africa was increasing. Other countries score average did not change.

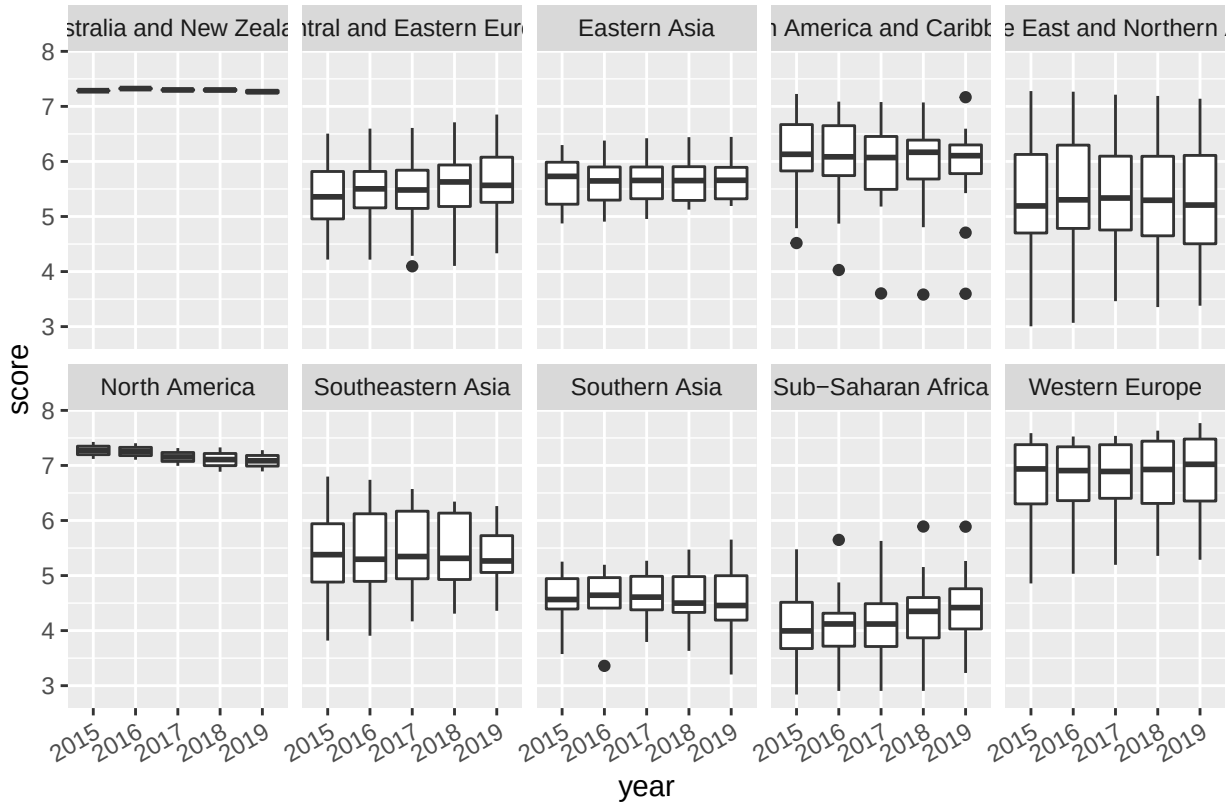

Figure 2.4

The geom_line of predictions, we can see from 2015 to 2019 the prediction is increasing.

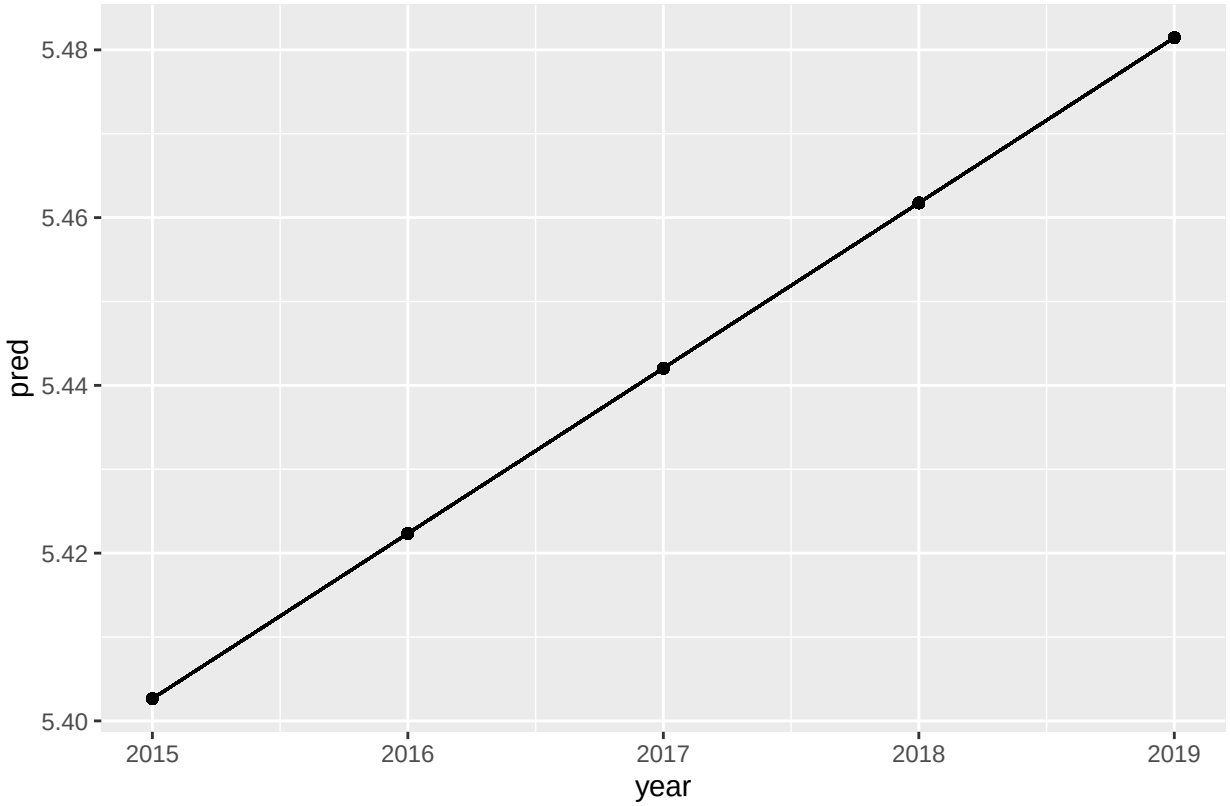

Figure 2.5

3. Democracy vs happiness

Hong Kong's spirit encouraged us to explore the relationship between democracy and happiness. Is democracy strongly relative to happiness?

To measure a country's amount of democratic freedoms, we'll use the data from https://www.gapminder.org/data/documentation/democracy-index/, which assigns each country a democracy index. The more democratic freedoms a country has, the higher the democracy index.

```
## Observations: 664
## Variables: 3
## $ year           <chr> "2015", "2016", "2017", "2018", "2015", "2016",...
## $ country        <chr> "Afghanistan", "Afghanistan", "Afghanistan", "A...
## $ DemocracyIndex <dbl> 27.7, 25.5, 25.5, 29.7, 59.1, 59.1, 59.8, 59.8,...
```

Figure 3.1, The Democracy table has 3 variables, and DemocracyIndex is a scale from 0-100 that gets higher the more democratic freedoms a country has.

This dataset only provides data up to year 2018, so we'll exclude the year 2019 from our analysis.

We must also rename the countries so that they'll have the same name in both datasets, and get rid of any countries that aren't in both datasets.

We'll do this by first gathering a list of countries that aren't included in both datasets.

```
## # A tibble: 36 x 1
```

```
## # Groups:   country [9]
##    country
##    <chr>
##  1 United States
##  2 North Cyprus
##  3 Kosovo
##  4 Bosnia and Herzegovina
##  5 Palestinian Territories
##  6 Congo (Kinshasa)
##  7 Georgia
##  8 Congo (Brazzaville)
##  9 Ivory Coast
## 10 United States
## # ... with 26 more rows
```

Table 3.2 Countries that are either not present in both datasets, or are named differently

Then we'll look through each of them using `democracy$country` and determine if the countries should be renamed or removed

Our 2 datasets are tidied and ready for merging.

```
## Observations: 560
## Variables: 11
## Groups: country [140]
## $ country        <chr> "Switzerland", "Iceland", "Denmark", "Norway", ...
## $ score          <dbl> 7.587, 7.561, 7.527, 7.522, 7.427, 7.406, 7.378...
## $ gdp            <dbl> 1.39651, 1.30232, 1.32548, 1.45900, 1.32629, 1....
## $ socsupport     <dbl> 1.34951, 1.40223, 1.36058, 1.33095, 1.32261, 1....
## $ lexp           <dbl> 0.94143, 0.94784, 0.87464, 0.88521, 0.90563, 0....
## $ freedom        <dbl> 0.66557, 0.62877, 0.64938, 0.66973, 0.63297, 0....
## $ trust          <dbl> 0.41978, 0.14145, 0.48357, 0.36503, 0.32957, 0....
## $ generosity     <dbl> 0.29678, 0.43630, 0.34139, 0.34699, 0.45811, 0....
## $ year           <chr> "2015", "2015", "2015", "2015", "2015", "2015",...
## $ region         <chr> "Western Europe", "Western Europe", "Western Eu...
## $ DemocracyIndex <dbl> 90.9, 95.8, 91.1, 99.3, 90.8, 90.3, 89.2, 94.5,...
```

Figure 3.3 A glimpse of what our final dataset looks like when the democracy index is added in.

Let's visualize the relationship between the democracy index and the happiness score to spot any patterns.

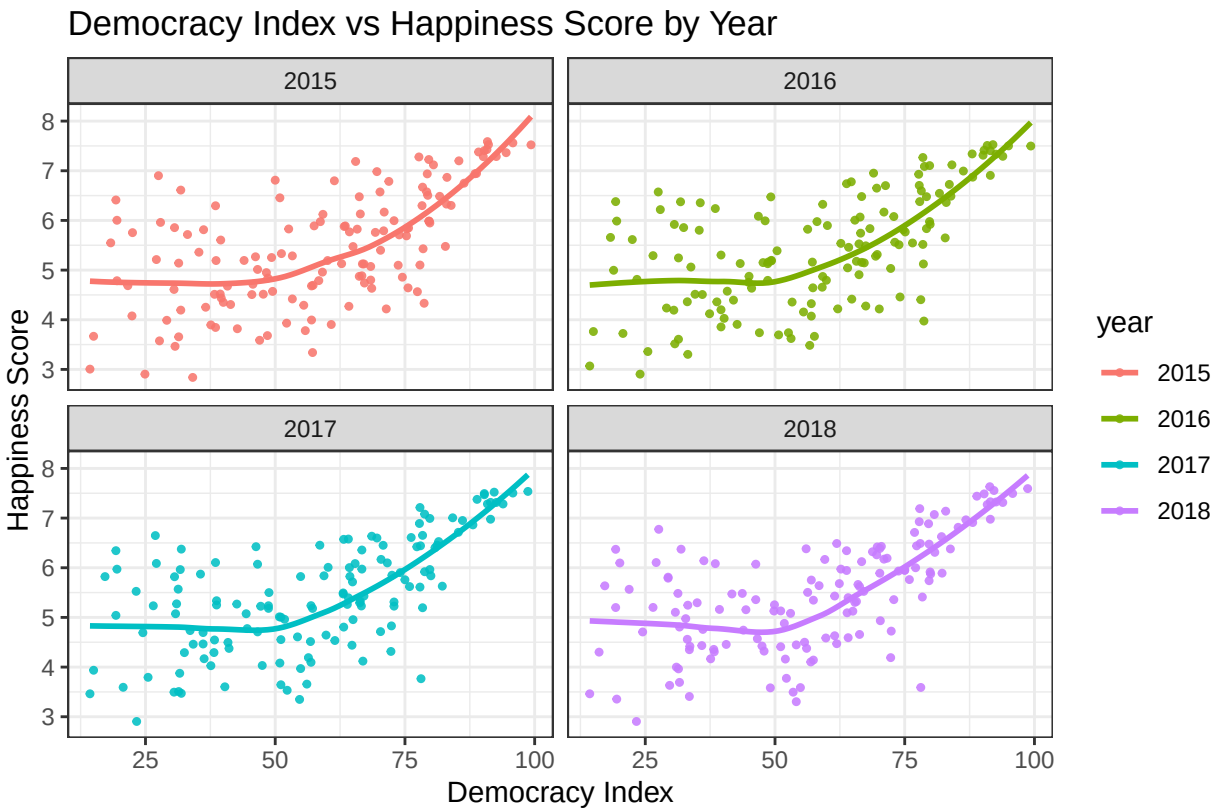

Figure 3.4 Visualizing the trend between democracy index and happiness score.

There is a interesting trend in Figure 3.4. When a country had Democracy Index more than 50, democracy had strongly positive relationship with happiness. But when the country has Democracy Index below 50, it appears that democracy has no relationship with a country's happiness.

For each of the years, we can see a mediocre positive quadratic relationship between the democracy index and their happiness score. Although there's more variation near the lower end of the democracy index, there's a strong correlation near the higher end of the demoracy index.

We will build a quadratic model that best fits the dataset using mean-square residuals.

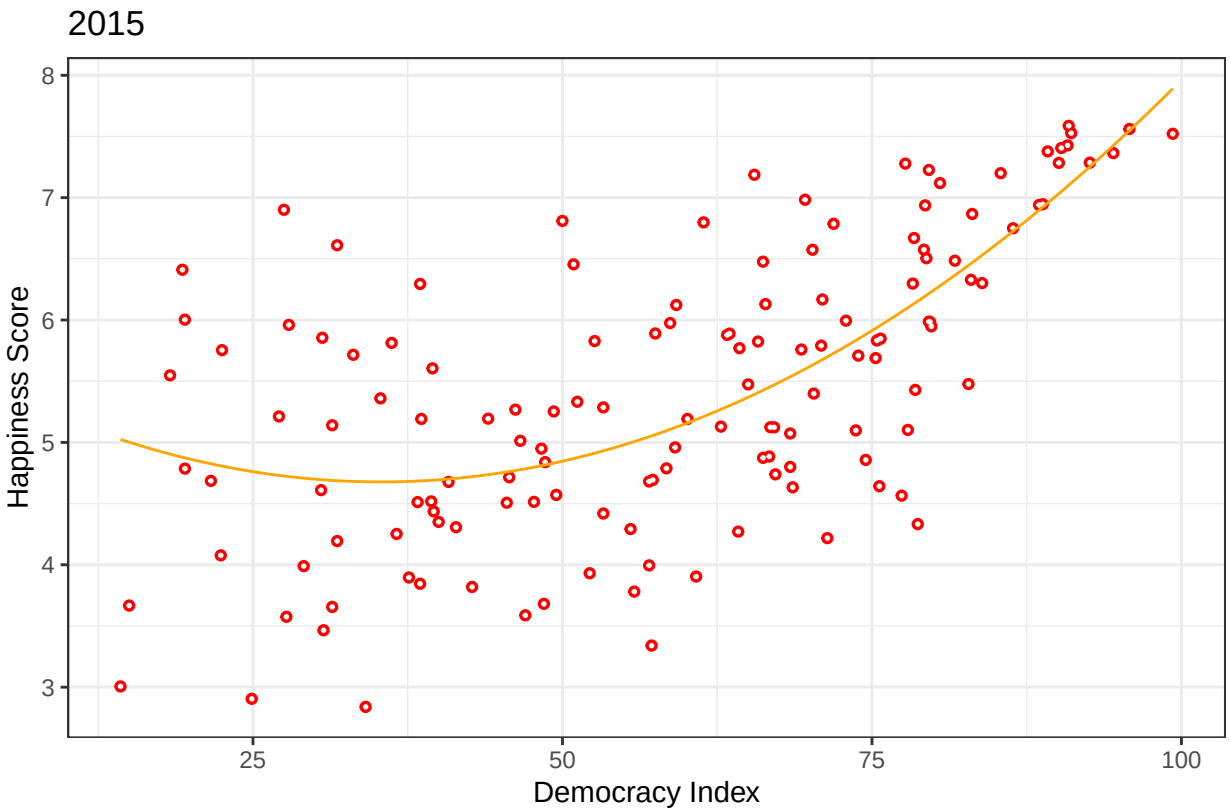

Figure 3.5 Qudratic Regression for 2015

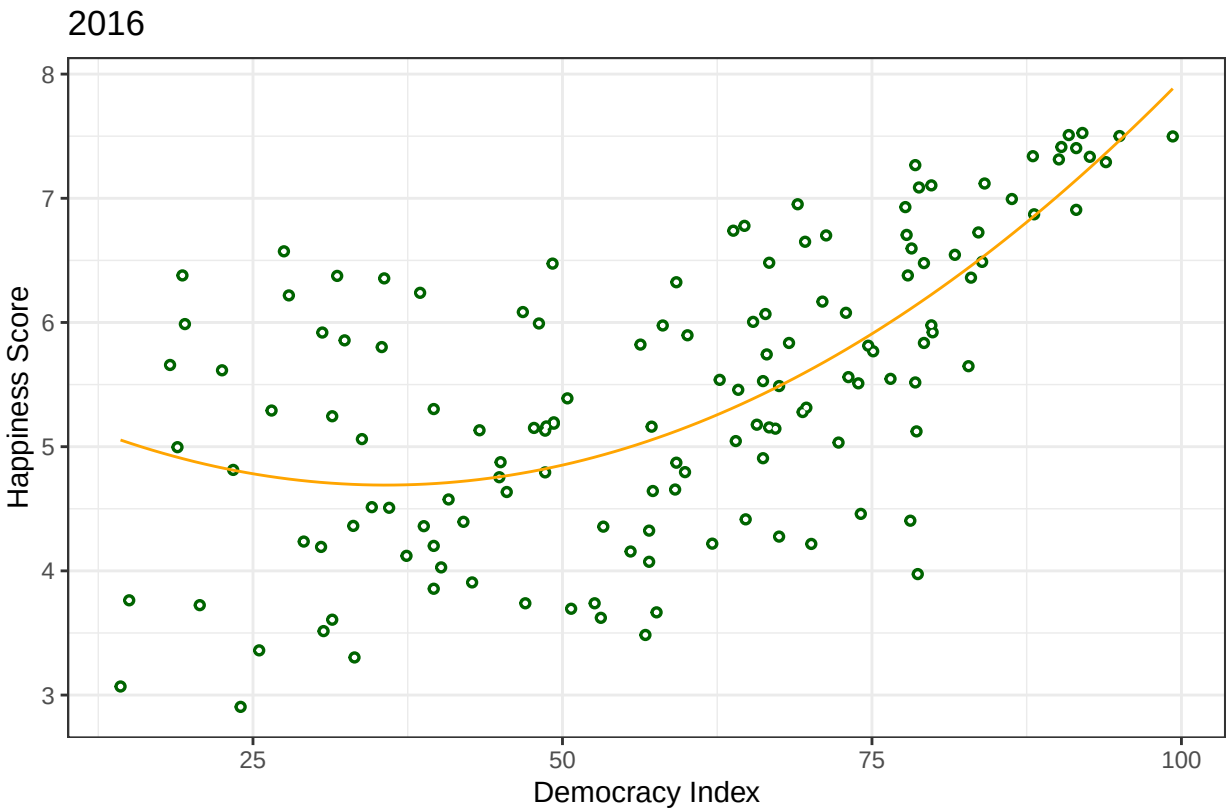

Figure 3.6 Qudratic Regression for 2016

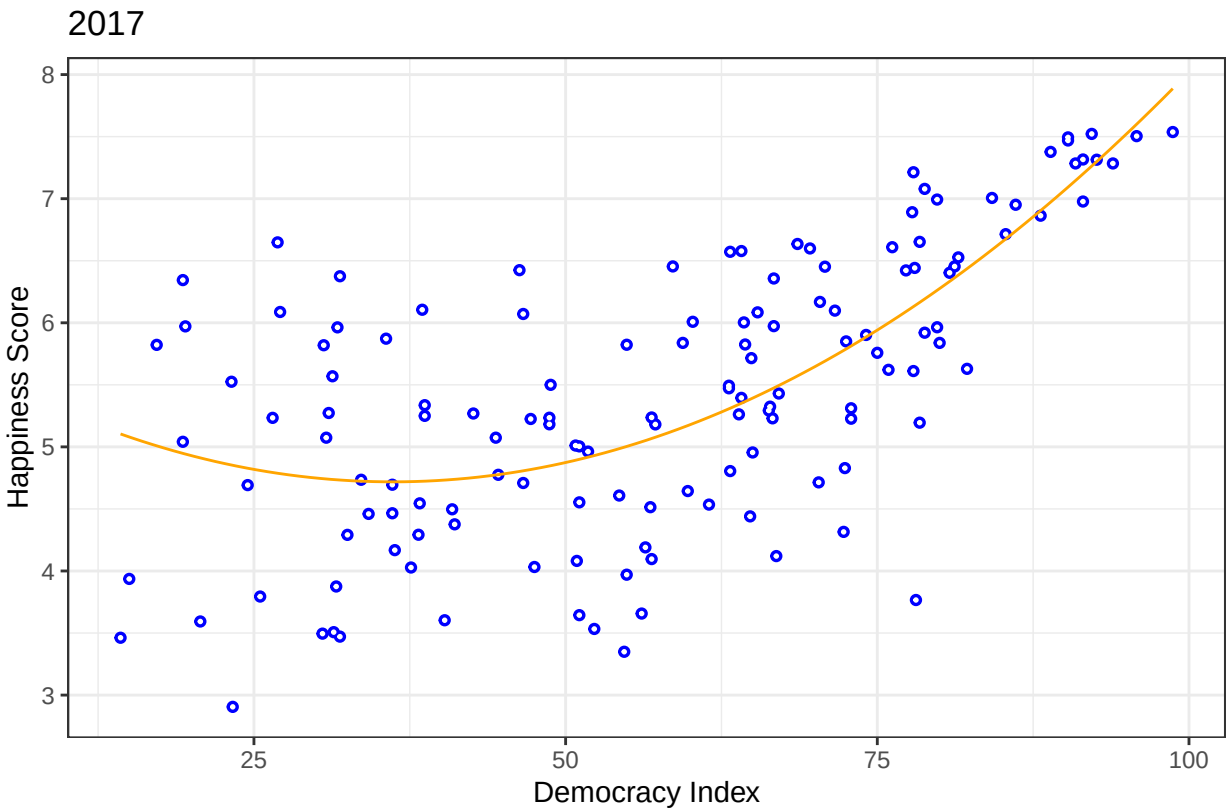

Figure 3.7 Qudratic Regression for 2017

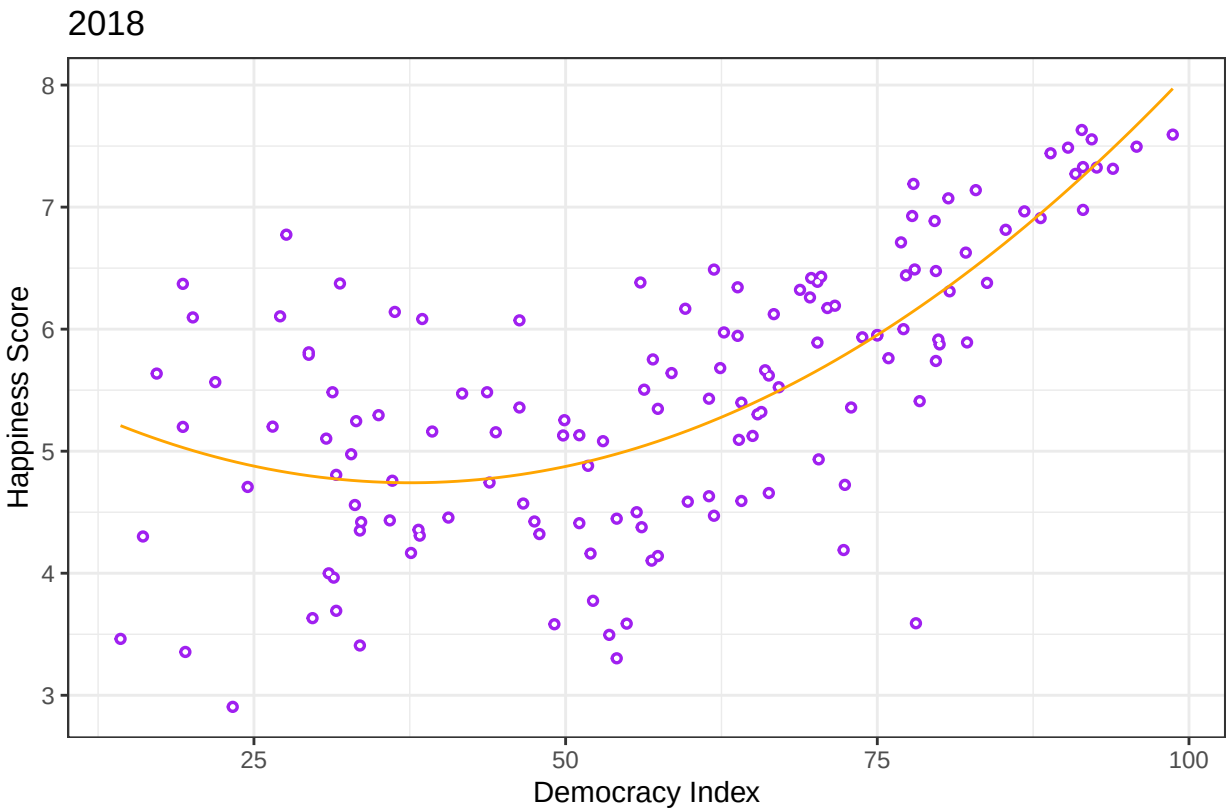

Figure 3.8 Qudratic Regression for 2018

The quadratic regression appears to fit greatly, but we must make sure that our residual plots don't exhibit any patterns, and the residuals should be normally distributed around 0. The formula for a residual is (actual - expected).

Let's make sure our residuals are normally distributed around 0.

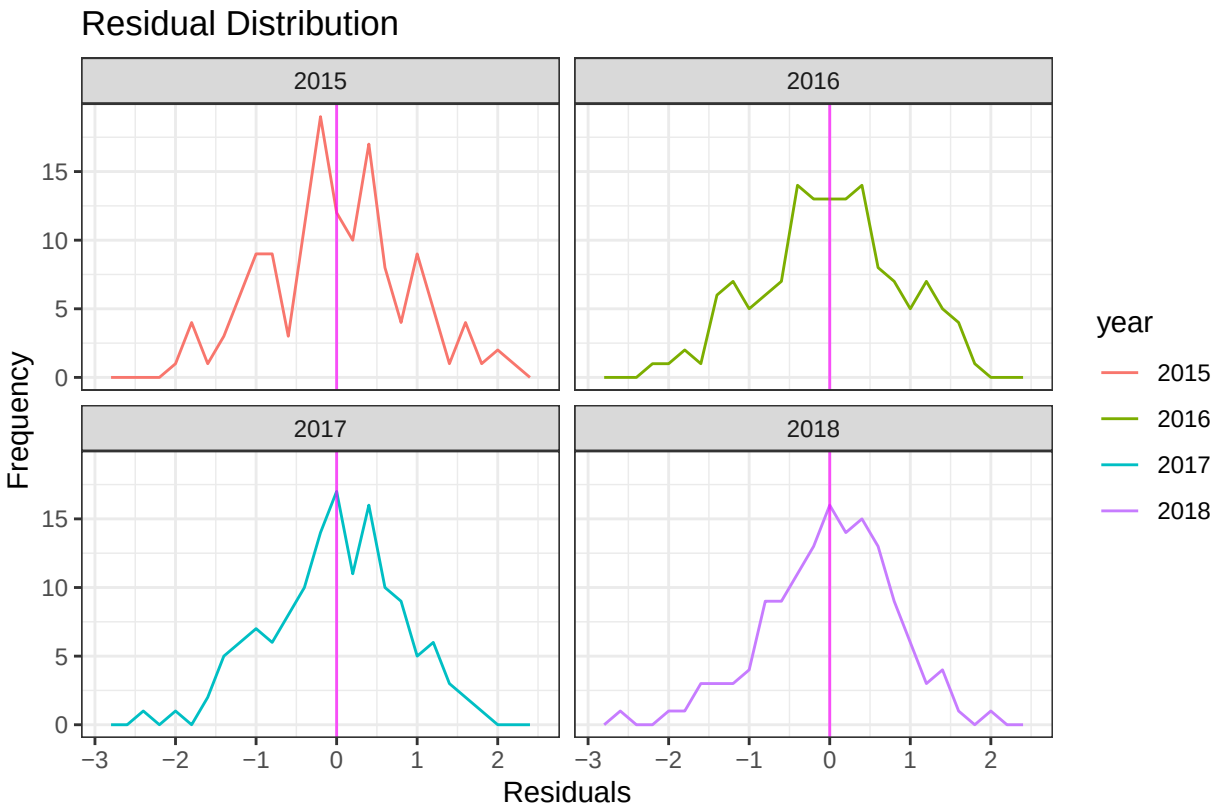

Figure 3.9 Distribution of Residuals

Every year's residual distribution is approximately normal and centered around 0. Next we'll look for any patterns in our residual plot.

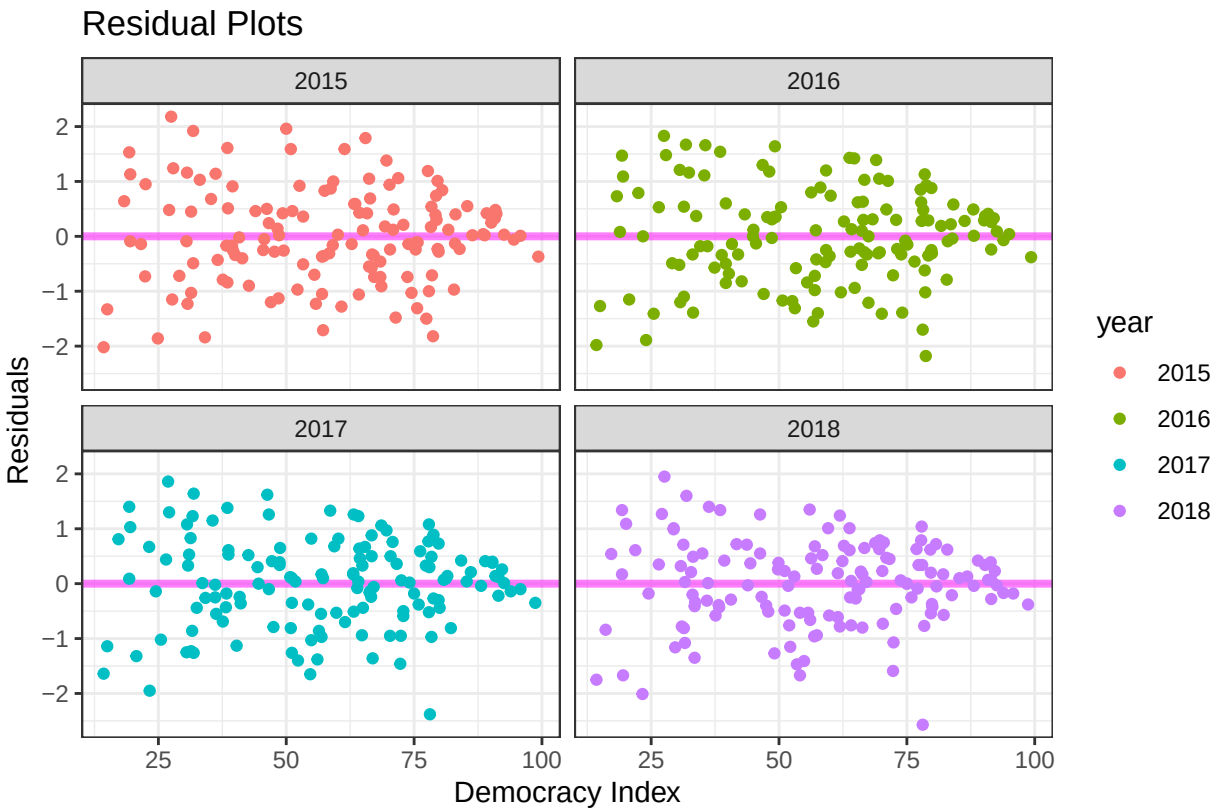

Figure 3.10 Residual Plots

The residual plots show no pattern, so our models are pretty good estimators.

We'll calculate a coefficient of determination: r-sqaured. This determines the percentage of the data that can be explained by our model.

The formula for r-squared is sum((predicted_y - mean(y))^2) / sum((observed_y - mean(y))^2)

```
## # A tibble: 4 x 2
##   year  rSquared
##   <chr>    <dbl>
## 1 2015     0.450
## 2 2016     0.447
## 3 2017     0.480
## 4 2018     0.504
```

Table 3.11 Coefficient of determiniation for each model

About 45%-50% of the data can be explained by our models. This isn't very good, but it's also not too bad, and it may still be useful when predicting a country's happiness score based on its democracy index.

We may test this out against a country that we didn't use for our model, such as Angola for the year 2015.

```
##   country year DemocracyIndex score predicted_score
## 1  Angola 2015           33.5 4.033        4.679784
```

Table 3.12 Predicted happiness score for Angola in the year 2015

Our model predicted that Angola would have a happiness score around 4.7 in the year 2015, while the actual happiness score is around 4.0. This is a moderately close estimate.

Conclusion

We all know that a complete and accurate data source is important and is fundamental to data analysis. That's why we have accepted the 4E rules to do the data cleaning job. Especially appling the rules played a critical role in exploring some important input errors. That was absolutely a fruitful experience. In the exciting exploratory analysis journey, we could see that the world is mostly a happy place for people, but the results were only based on a survey of a sample of 1000 people for each country. Since every country has a different amount of population, it would be more accurate if a weight was assigned for each country for calculating the world's happiness score. For future research, someone could assign an appropriate weight for each country, and re-calculate the world's happiness score. From our analysis on the relationship between a country's democracy and its happiness score, a country tends to be happier when it has more democracy, especially for countries that have democracy index more than 50.