DATA ANALYSIS

# WORLD HAPPINESS REPORT

## EDA

# Introduction

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th.  I choose the report in 2016 to do my EDA project. The data in 2016 is relatively new and tidy. The reason The World Happiness Report interests me is data can bring humanity messages through the wave of numbers. Also, the reports continue to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions.

## Source

I pick up this resource on Kaggle. Many unique and professional open data there. I look through the list. Some are about blockchain, some are Apps usage rate, but rarely of them hold clean data and credible. Most data that I give up for limited variables. It will be hard to do visualization for further steps. The limitation of this data is that each country has a unique name, even it has a column 'region'. It's hard to manipulate them as a group.

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

## Choosing Data

The data originally has 12 columns. I pick up 7 of them which I think are more related to happiness evaluation. By looking through those data, I find it is arranged by happiness rank instead of countries, rank 1 is the happiest country in the world and the data for this country are in the row. The interesting find is, countries are categorized by region which I would like to use for visualization exploration. I use Apple Numbers to clean my data delete the needless columns. After I finish the data cleaning, I export it as a CSV file to the desktop.

### Columns

#Country
#Region
#Happiness Rank
#Happiness Score
#Economy (GDP per Capita)
#Family Health (Life Expectancy)
#Freedom

| Country | Region | Happiness Rank | Happiness Score | Economy (GDP per Capita) | Health (Life Expectancy) | Freedom |
|---|---|---|---|---|---|---|
| **Denmark** | Western Europe | 1 | 7.526 | 1.44178 | 0.79504 | 0.57941 |
| **Switzerland** | Western Europe | 2 | 7.509 | 1.52733 | 0.86303 | 0.58557 |
| **Iceland** | Western Europe | 3 | 7.501 | 1.42666 | 0.86733 | 0.56624 |
| **Norway** | Western Europe | 4 | 7.498 | 1.57744 | 0.79579 | 0.59609 |
| **Finland** | Western Europe | 5 | 7.413 | 1.40598 | 0.81091 | 0.57104 |
| **Canada** | North America | 6 | 7.404 | 1.44015 | 0.8276 | 0.5737 |
| **Netherlands** | Western Europe | 7 | 7.339 | 1.46468 | 0.81231 | 0.55211 |
| **New Zealand** | Australia and New Zealand | 8 | 7.334 | 1.36066 | 0.83096 | 0.58147 |
| **Australia** | Australia and New Zealand | 9 | 7.313 | 1.44443 | 0.8512 | 0.56837 |
| **Sweden** | Western Europe | 10 | 7.291 | 1.45181 | 0.83121 | 0.58218 |
| **Israel** | Middle East and Northern Africa | 11 | 7.267 | 1.33766 | 0.84917 | 0.36432 |
| **Austria** | Western Europe | 12 | 7.119 | 1.45038 | 0.80565 | 0.54355 |
| **United States** | North America | 13 | 7.104 | 1.50796 | 0.779 | 0.48163 |
| **Costa Rica** | Latin America and Caribbean | 14 | 7.087 | 1.06879 | 0.76146 | 0.55225 |
| **Puerto Rico** | Latin America and Caribbean | 15 | 7.039 | 1.35943 | 0.77758 | 0.46823 |

I install the package and choose some tools to use and import my CSV file. Before I start to make the 'ggplot' chart, I need to check the data resource has no "NA" .

```
> sum(is.na(happyrate))
[1] 0
```

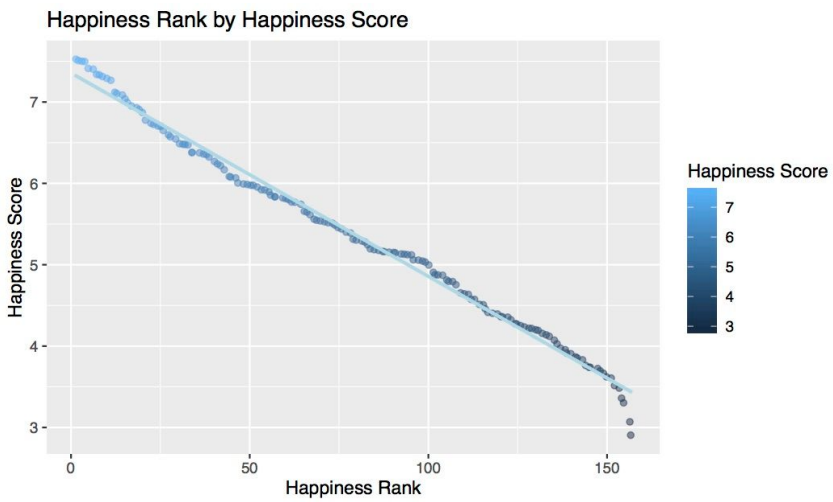
First, I need to use 'glimpse' to overview my data. I got 157 observations which are my rows and 7 variables as columns. Rstudio is so convenient it lists the heads of all columns, country; region; happiness rank; happiness score; economy; health; freedom. However, the 'happiness rank' column is special to others because of its ordinal variables belonging. I will work that point out to make a comparison between 'happiness rank' with other variables.

```
> glimpse(happyrate)
Observations: 157
Variables: 7
$ Country                    <fct> Denmark, Switzerland, Iceland, Norway, Finland, C…
$ Region                     <fct> Western Europe, Western Europe, Western Europe, W…
$ `Happiness Rank`           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15…
$ `Happiness Score`          <dbl> 7.526, 7.509, 7.501, 7.498, 7.413, 7.404, 7.339, …
$ `Economy (GDP per Capita)` <dbl> 1.44178, 1.52733, 1.42666, 1.57744, 1.40598, 1.44…
$ `Health (Life Expectancy)` <dbl> 0.79504, 0.86303, 0.86733, 0.79579, 0.81091, 0.82…
$ Freedom                    <dbl> 0.57941, 0.58557, 0.56624, 0.59609, 0.57104, 0.57…
```

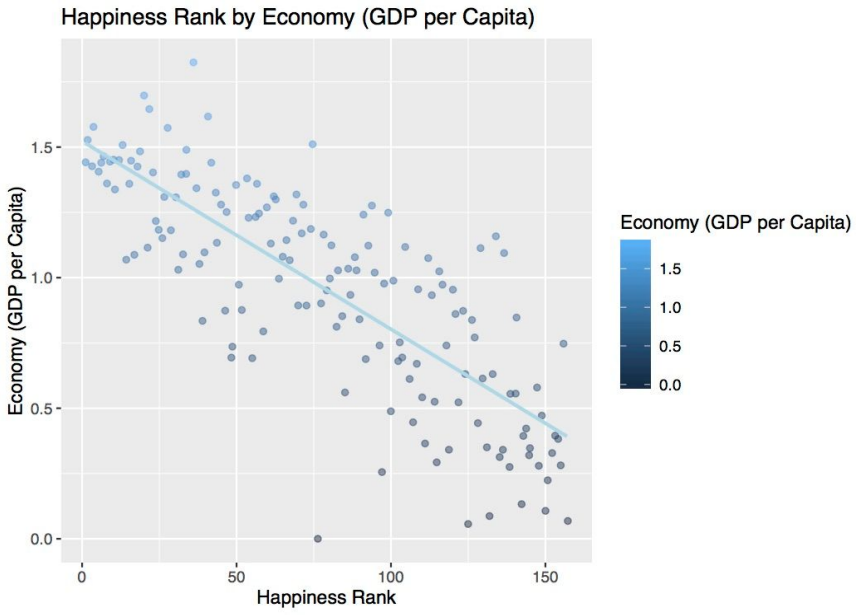
Then I use 'summary' to have an overall consideration of my dataset. I get Min, 1st Qu, Median, Mean, 3rd Qu, Max, of each variable column. After doing that preparation, I start to use 'ggplot' to make EDA.

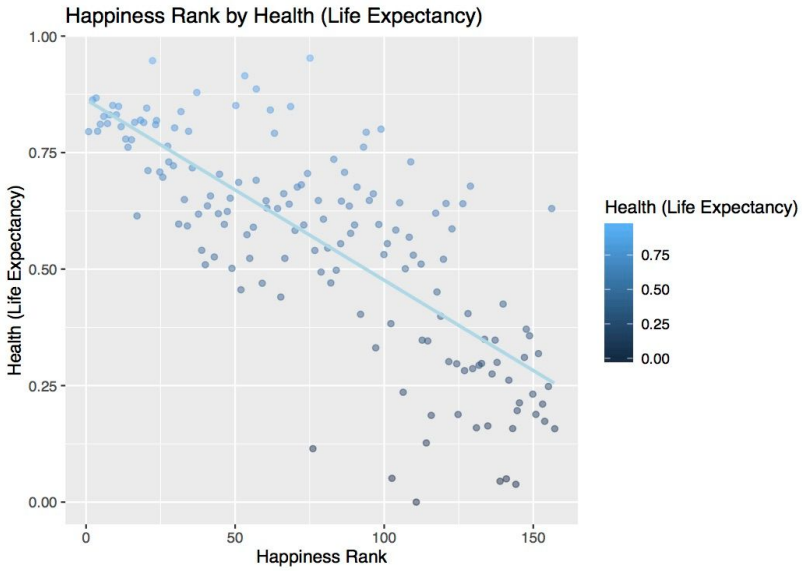
```
> summary(happyrate)
       Country                                 Region   Happiness Rank
 Afghanistan:  1   Sub-Saharan Africa              :38   Min.   :  1.00
 Albania    :  1   Central and Eastern Europe      :29   1st Qu.: 40.00
 Algeria    :  1   Latin America and Caribbean     :24   Median : 79.00
 Angola     :  1   Western Europe                  :21   Mean   : 78.98
 Argentina  :  1   Middle East and Northern Africa:19   3rd Qu.:118.00
 Armenia    :  1   Southeastern Asia               : 9   Max.   :157.00
 (Other)    :151   (Other)                         :17
 Happiness Score Economy (GDP per Capita) Health (Life Expectancy)    Freedom
 Min.   :2.905   Min.   :0.0000           Min.   :0.0000           Min.   :0.0000
 1st Qu.:4.404   1st Qu.:0.6702           1st Qu.:0.3829           1st Qu.:0.2575
 Median :5.314   Median :1.0278           Median :0.5966           Median :0.3975
 Mean   :5.382   Mean   :0.9539           Mean   :0.5576           Mean   :0.3710
 3rd Qu.:6.269   3rd Qu.:1.2796           3rd Qu.:0.7299           3rd Qu.:0.4845
 Max.   :7.526   Max.   :1.8243           Max.   :0.9528           Max.   :0.6085
```

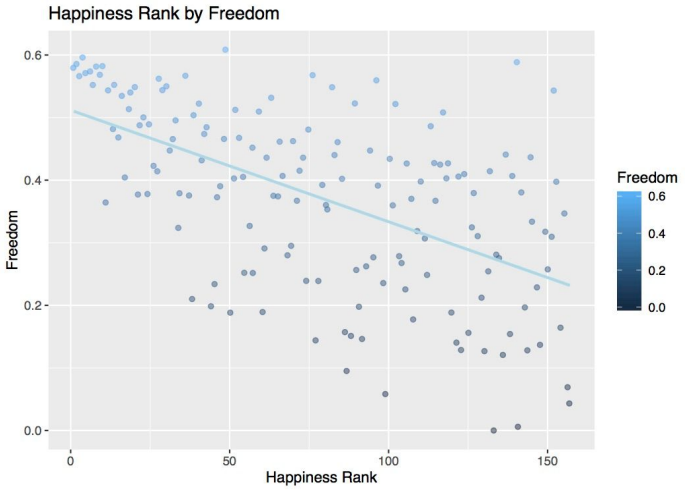
## RESULTS



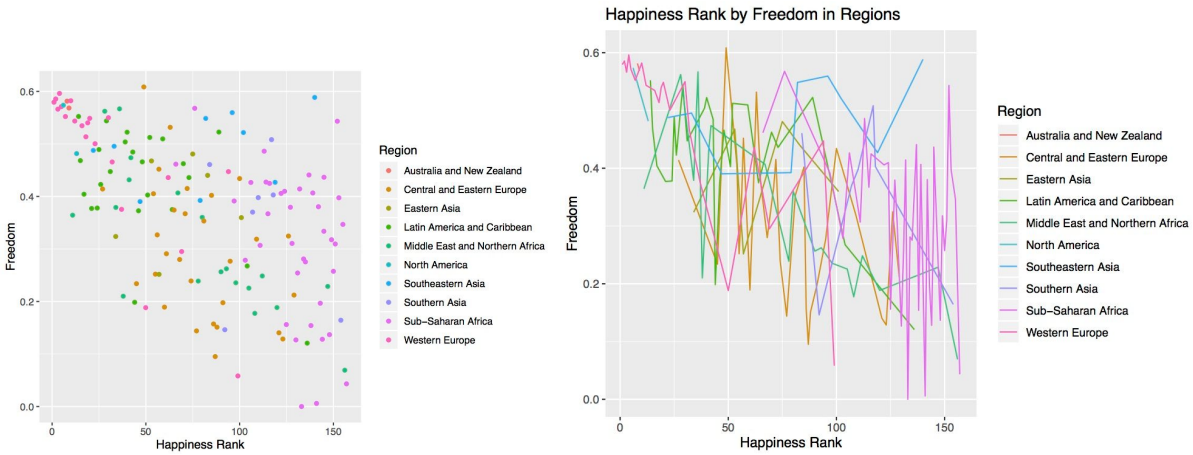

I make the dots transparent with color and add a line for the data trend. The relationship between happiness score and rank is very tight. It's the correlation coefficient, very positive. The higher level of the happiness rank is, the more happiness score the country has.



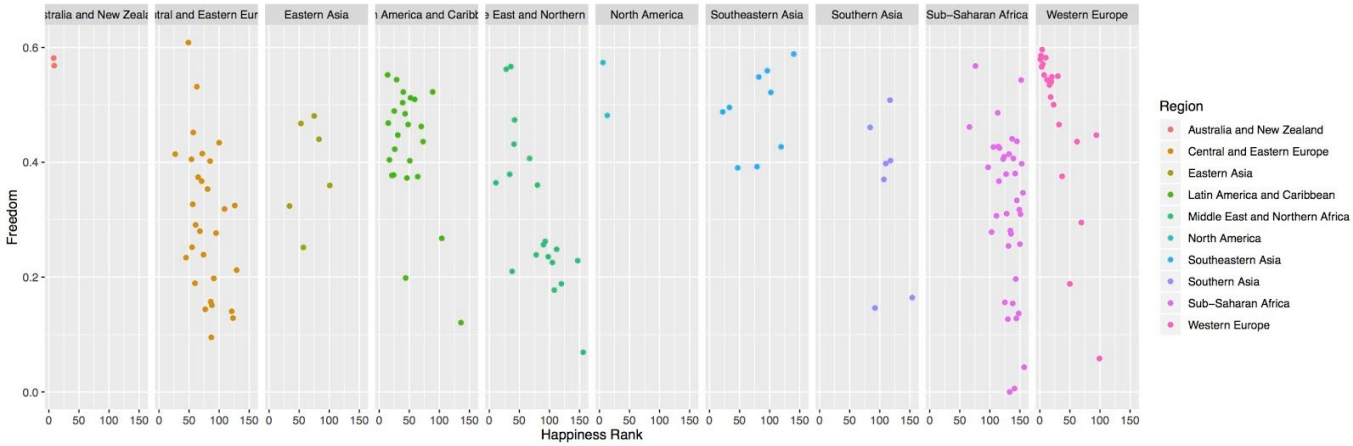


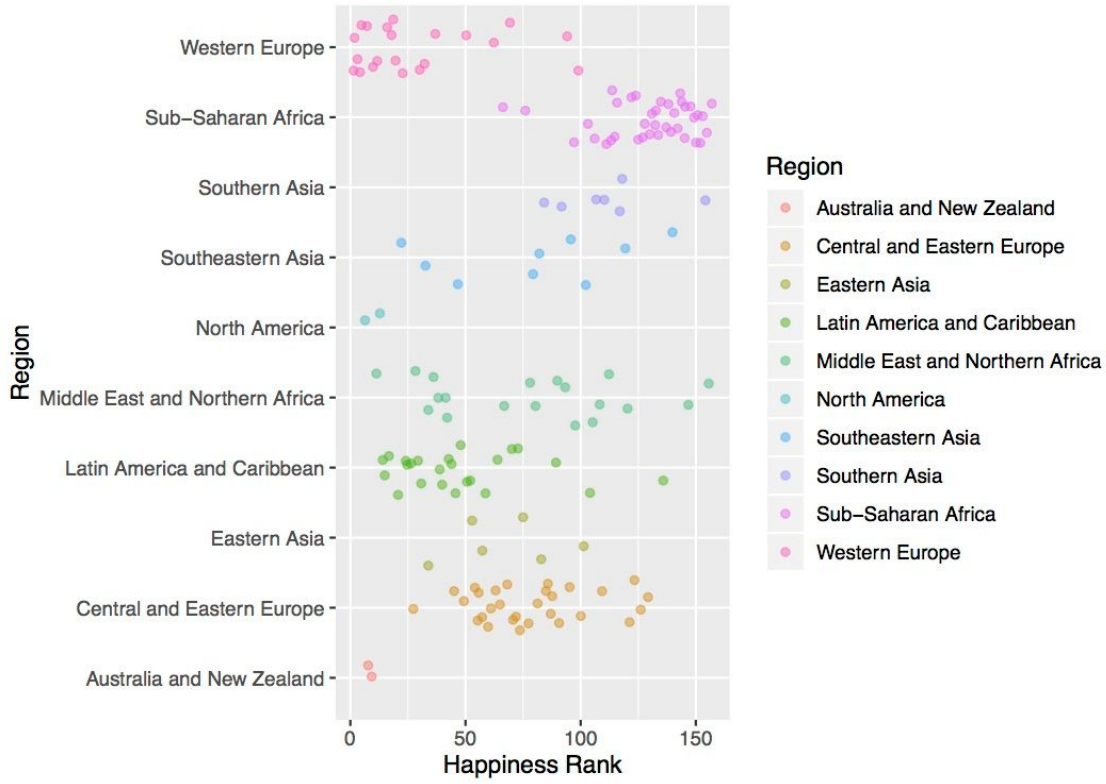

The other columns, economy, health, freedom, their relationship with happiness rank aren't as tight as happiness score. However, the line trends of the data, they are all positive relationships. What attracts my attention is the country has the best economic situation (GDP), however, people's happiness rank isn't the No.1. The same appearance in health&rank and freedom&rank data. From that observation, I come up with the conclusion that happiness is influenced by multiple aspects instead of a single one. Moreover, I find the unique cluster points in the freedon&rank between 0 to 50. The other points are kind of scattered. That cluster points show a very positive relationship in 2 variables. More freedom can help people feel happier.




Region column group different countries into several areas. I try to add more variables into one chart to make a difference. I even change the geometry. Unfortunately, I finally get a mess points and lines which are hard for me to read.




In order to make the chart more clear. I use ‘facet_grid’ to divide the different regions into different charts. Then, I can easily observe the relationship between freedom&happiness_rank in each region. I find people in Western Europe have more freedom and they live happier than others. People from North America also in good condition. Whereas regions like Central and Eastern Europe, Sub-Saharan Africa, are in the low range of happiness rank and they don’t have much freedom.




To be more specific, I delete one variable. From the new chart, countries from Western Europe, points are clustered around 0 to 25 of happiness rank. However, Sub-Saharan Africa points are around 100 to 150. The chart gives the viewer a strong comparison with each other. I also calculate the mean number of variables for each region.

```
> happyrate_summary
                          Region Country Region Happiness Rank Happiness Score
1        Australia and New Zealand      NA     NA        8.50000        7.323500
2       Central and Eastern Europe      NA     NA       78.44828        5.370690
3                   Eastern Asia      NA     NA       67.16667        5.624167
4      Latin America and Caribbean      NA     NA       48.33333        6.101750
5  Middle East and Northern Africa      NA     NA       78.10526        5.386053
6                   North America      NA     NA        9.50000        7.254000
7               Southeastern Asia      NA     NA       80.00000        5.338889
8                   Southern Asia      NA     NA      111.71429        4.563286
9               Sub-Saharan Africa      NA     NA      129.65789        4.136421
10                  Western Europe      NA     NA       29.19048        6.685667
   Economy (GDP per Capita) Health (Life Expectancy)   Freedom
1                 1.4025450               0.8410800 0.5749200
2                 1.0475369               0.6315921 0.3005283
3                 1.2773117               0.8066800 0.3872817
4                 0.9934100               0.6127025 0.4266463
5                 1.1393232               0.6164379 0.3097689
6                 1.4740550               0.8033000 0.5276650
7                 0.8963811               0.5613689 0.4901233
8                 0.6606714               0.4536700 0.3500257
9                 0.4743213               0.2398724 0.3154242
10                1.4170557               0.8257538 0.4775871
```

**Conclusion**

Finding appropriate data is the key to EDA. If data isn't suitable for visualization and unreliable, data explorers can't do further work on that dataset. The preparation is necessary, checking 'NA', glimpsing data frame, etc.

If putting too many variables in one chart, the viewer will feel hard to generate findings since there are too many comparable data near each other.  The trend line is a good way to see the data-relationship is positive or negative. Also, Exploring data with an unbias attitude is important. Data reflects phenomena which show the truth to some extent. When I see the most wealthy country is not the happiest place. I find happiness can't be controlled by a single aspect.  The more I navigate through the data, the more findings that surprise me. Be aware of data structure is also helpful when using tools to manipulate them. Otherwise, Rstudio will come up with errors even the tool is right.

The number itself can't tell things unless people use it to make a comparison. Rstudio is a very useful tool to help me go through data which is a very powerful way to reveal the truth. Exploring a dataset with an objective attitude teaches me a lot.

## APPENDIX

```
install.packages("ggthemes")
#install.packages("dplyr")

library(ggthemes)
library(tidyverse)
library(dbplyr)
library(lubridate)

happyrate <- read.csv("Desktop/2016worldhappiness.csv", header=TRUE, check.names = FALSE)

class(happyrate)
dim(happyrate)
names(happyrate)
str(happyrate)
glimpse(happyrate)
summary(happyrate)

head(happyrate)

sum(is.na(happyrate))

ggplot(happyrate, aes(x =`Happiness Rank`, y =`Happiness Score`, color = `Happiness Score`)) +
  geom_jitter(alpha=0.5)+
  stat_smooth(method="lm", se=FALSE, col="light blue") +
  labs(title = "Happiness Rank by Happiness Score")

ggplot(happyrate, aes(x =`Happiness Rank`, y =`Economy (GDP per Capita)`, color = `Economy (GDP per Capita)`)) +
  geom_jitter(alpha=0.5)+
  stat_smooth(method="lm", se=FALSE, col="light blue") +
  labs(title = "Happiness Rank by Economy (GDP per Capita)")

ggplot(happyrate, aes(x =`Happiness Rank`, y =`Health (Life Expectancy)`, color = `Health (Life Expectancy)`)) +
  geom_jitter(alpha=0.5)+
  stat_smooth(method="lm", se=FALSE, col="light blue") +
  labs(title = "Happiness Rank by Health (Life Expectancy)")

```

```r
ggplot(happyrate, aes(x =`Happiness Rank`, y =`Freedom`, color = `Freedom`)) +
  geom_jitter(alpha=0.5)+
  stat_smooth(method="lm", se=FALSE, col="light blue") +
  labs(title = "Happiness Rank by Freedom")

ggplot(happyrate, aes(x =`Happiness Rank`, y =`Region`, color = `Region`)) +
  geom_jitter(alpha=0.5)+
  stat_smooth(method="lm", se=FALSE, col="light blue") +
  labs(title = "Happiness Rank by Region")

happyrate_more <- ggplot(happyrate, aes(x =`Happiness Rank`, y =`Freedom`, color = `Region`))

happyrate_more + geom_point()
happyrate_more + geom_jitter()
happyrate_more + geom_line()+
  labs(title = "Happiness Rank by Freedom in Regions")

ggplot(happyrate, aes(x =`Happiness Rank`, y =`Freedom`, color=`Freedom`)) +
  geom_jitter(alpha=0.5) +
  stat_smooth(method="lm", se=FALSE, col="blue") +
  labs(title = "Happiness Rank by Freedom")

ggplot(happyrate, aes(x =`Happiness Rank`, y =`Region`, color = `Region`)) +
  geom_jitter(alpha=0.5)

`happyrate`$`Region`[`happyrate`$`Region`==""] <- "NA"

posn <- position_jitter(width = .1)
happyrate_more + geom_point(position=posn) +
  facet_grid(.~Region)

posn <- position_jitter(width = .1)
happyrate_more + geom_point(position=posn)

happyrate_summary <- aggregate(happyrate[1:7], list(happyrate$Region), mean)
names(happyrate_summary)[1] <-"Region"
happyrate_summary
warnings()
```