Final Project: Happiness Prediction

ECE 225A Probability and Statistics in Data Science
A12620672, Kalvin Goode
A53317226, Sheng-Wei Chang

1. INTRODUCTION

In this modern society, many people feel unhappy. We would like to use data find the factors which affect happiness. Also, we would like to find the relationship between these factors and the factor, and by the factor of the following year to predictor happiness. Therefore, this project explores the relationship between happiness variables including GDP per capita, social support, healthy life expectancy at birth, etc. We choose the linear regression method to analysis. The dataset is from Kaggle: 'World Happiness Report.' [1] It collects data from 165 countries. The aim of the project is to understand the most important factors for people to be happy in their lives.

2. VARIABLE DEFINITION

(1) Happiness Score(Happy) [2]

The survey measure is based on January, 2019 release of the Gallup World Poll (GWP). Years range from 2005 to 2018. The question is 'Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?' Responses are calculated as the average of 0-or-1 response in each country each year.

(2) Log GDP per Capita (GDP) [2]

This variable, GDP per capita, in purchasing power parity (PPP) at constant 2011 international dollar prices are from the World Development Indicators (WDI). This variable is in a log scale.

(3) Social Support (Social) [2]

Same as the variable Happiness Score, the survey measure is based on the Gallup World Poll (GWP). The question is 'If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?' The answers are binary, 0 for No and 1 for Yes. Responses are calculated as the average of 0-or-1 response in each country each year.

(4) Freedom to Make Life Choices (Freedom) [2]

Same as the variable Happiness Score, the survey measure is based on the GWP. The question is 'Are you satisfied or dissatisfied with your freedom to choose what you do with your life?' The answers are binary, 0 for No and 1 for Yes. Responses are calculated as the average of 0-or-1 response in each country each year.

(5) Perceptions of Corruption (Corruption) [2]

This variable is based on two question in the GWP. The questions include ‘Is corruption widespread throughout the government or not’ and ‘Is corruption widespread within businesses or not?’ The answers are binary, 0 for No and 1 for Yes. Responses are calculated as the average of 0-or-1 response in each country each year.

(6) Democratic Quality (Democracy) [3]

This variable is taken from a country-year panel of governance indicators from the Worldwide Governance Indicators (WGI) project (Kaufmann, Kraay and Mastruzzi; last update: September 21, 2018)

(7) Positive Affect [2]

This variable is defined as the average of three measure in the GWP: happiness, laugh and enjoyment. The questions are ‘Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Happiness?’, ‘Did you smile or laugh a lot yesterday?’, and ‘Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Enjoyment?’ The answers are binary, 0 for No and 1 for Yes. Responses are calculated as the average of 0-or-1 response in each country each year.

(8) Negative Affect [2]

This variable is defined as the average of three measure in the GWP: worry, sadness and anger. The questions are ‘Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Worry?’, ‘Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Sadness?’, and ‘Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Anger?’ The answers are binary, 0 for No and 1 for Yes. Responses are calculated as the average of 0-or-1 response in each country each year.

3. METHOD

(1) In the original dataset, it includes 24 columns. Some variables lack data in specific countries. Thus, we choose 8 variables that we are most interested in the analysis.

(2) First, we used a boxplot to see the general statistics for each variable and to examine significant outliers. Then, we also check the general relationship between variables using the scatterplot matrix (Figure 1).

(3) At the same time fit a simple first-order linear regression model that uses all the variables we have in the data set. The overall p-value, in summary, reflects the overall fitness of the first-order model. Using a hypothesis test on a slope from the predictors, we can check if there is a linear relationship between individual predictor and response. We set $p=0.01$ to determine the significance of the result in all hypothesis tests.

(4) Using ANOVA (Analysis of variance) function, not only we can confirm the results are significant for the hypothesis test, but also examining how much of the Total Sum of Square is due to the variation of the individual predictor. In other words, it provides information on the significance of

the sequential Sum of Square by showing how much error had the predict reduce when adding to the linear regression model.

(5) To decide the final model, we perform a stepwise regression using forward addition and backward elimination. Stepwise regression is a method to find the best model from the subset of predictors in the original model. Even though models with more predictors will explain the outcome better and reduce errors, smaller models are easier to interpret. Therefore, stepwise regression can find the least complex model that can sufficiently explain the outcome. One way to measure the complexity is by calculating Akaike's Information Criteria (AIC). It is based on the Residual Sum of Square (SSE), sample size and the number of model parameters [4]:

$$\mathrm{AIC}_p = \mathrm{n} \times \ln(SSE_p) - \mathrm{n} \times \ln(n) + 2\mathrm{p}$$

The model with the lowest AIC value indicates the best model among other possible model selection.

(6) Finally, we checked the overall fit of the final model with summary function and examine potential violations of assumption in linear regression by using plotting residuals and normal Q-Q plot.

4. RESULT

Below is a scatterplot matrix (Figure 1) for all the variables in the data set, we see that there is a positive strong linear relationship between Happy and GDP, between Happy and Social, a weak linear relationship between Happy and Freedom, between Happy and Positive Affect, between Happy and Democracy, no clear relationship between Happy and Corruption, between Happy and Negative.

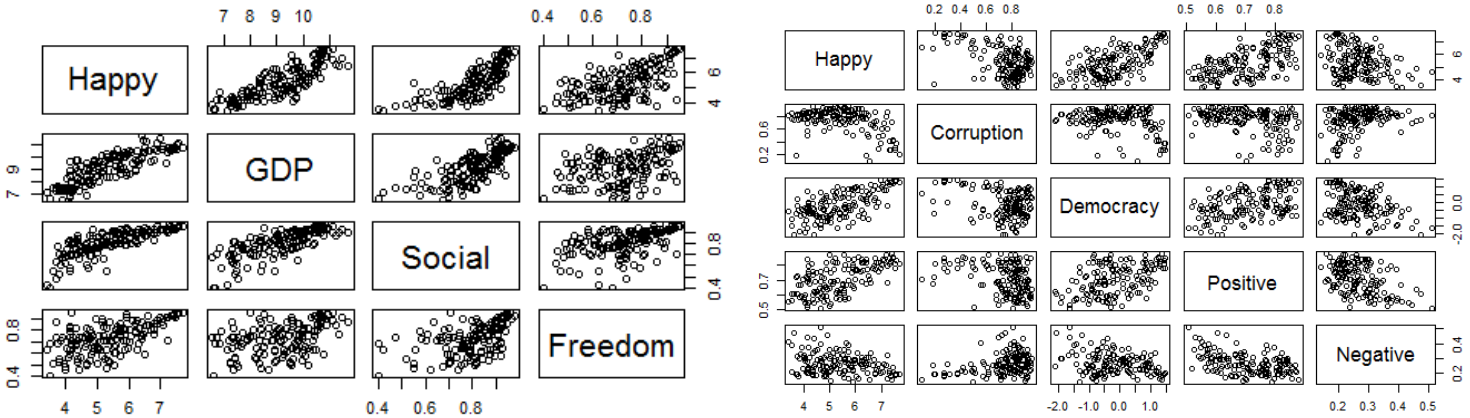

Figure 1: Scatterplot matrix

First, we performed the hypothesis testing of the overall regression model to check the general relationship in the data with

H_0: the slopes of all variables to 'Happy' are 0, $b_{GDP} = b_{Social} = \cdots = b_{Negative} = 0$, and

H_a: not all slopes are 0.

From the summary function of the first-order model, we see that 'Democracy' and 'Negative' have a p-value larger than 0.01, and other variables have a p-value smaller than 0.01. This indicates 'Democracy'

and ‘Negative’ do not show the relationship to outcome ‘Happy’ when also considering other variables in the models. Therefore, we removed ‘Democracy’ and ‘Negative’ to update the models and resulted in an overall p-value smaller than 0.01. Thus, we can reject H_0 and conclude that not all slopes are 0, and all predictors in the model have some relationship to outcome ‘Happy’. The initial model we get is

$$\text{Happy} = -2.14 + 0.49 \times \text{GDP} + 1.84 \times \text{Social} + 0.5 \times \text{Freedom} - 0.73 \times \text{Corruption} + 2.47 \times \text{Positive}$$

Using ANOVA function to perform F-test, it confirms that all predictors have p-values smaller than 0.01. Thus, all predictors have significant partial Sum of Square in this model, which can interpret as ‘GDP’ is significant given nothing in the model, ‘Social’ is significant given ‘GDP’ is in the model, and ‘Freedom’ is also significant given ‘GDP’ and ‘Social’ are in the model, etc.

To determine the model has only the necessary predictors, we perform stepwise regression with the forward addition and the backward elimination. Comparing several AIC values between models, we see that the model without ‘Freedom’ is as good as the original model. Therefore, we removed ‘Freedom’ to update the model and the summary function of linear regression had confirmed all the remaining predictors are significant to the response ‘Happy’. Lastly, using Residuals plot and Normal Q-Q plot (Figure 2), we checked if the residuals of the model are normally distributed with constant variance, which is an important assumption of linear regression. From all available tests of the linear regression model, we concluded the final model is:

$$\text{Happy} = -1.97 + 0.49 \times \text{GDP} + 1.93 \times \text{Social} - 0.85 \times \text{Corruption} + 2.77 \times \text{Positive}$$

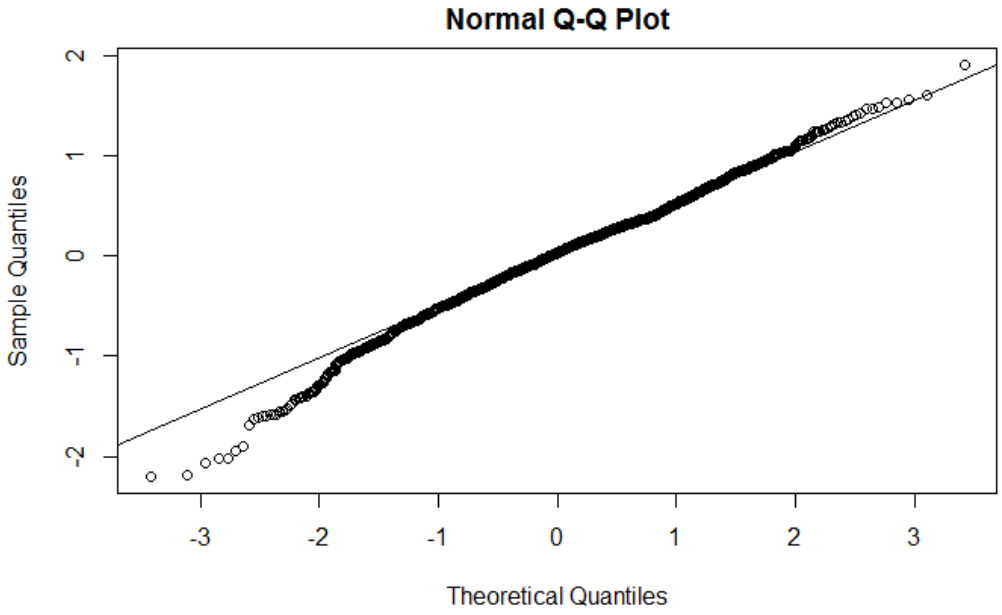

Figure 2: Q-Q Plot

5. DISCUSSION

We can interpret the final model as follows: The happiness of a person is 2.54 is the person does not have social support and positive experience, doesn't believe in corruption and lives in the country that has average GDP (9.22) in the world. Happiness is affected the most by having a positive recent experience. Having happiness experience increases happiness by 2.77 while other variables held constant. Having social support increases happiness by 1.93 while other variables held constant. Believing in widespread corruption in the country decreases happiness by 0.85 while other variables held constant.

Moreover, for each additional unit of GDP of a country, happiness increases by 0.49 while other variables held constant. We also see that multiple R-squared is 0.76, so 76% of the variance in the predictor 'happy' can be explained by knowing 'GDP', 'Social', 'Corruption' and 'Positive Affect'. From the residuals plot (Figure 3) of the final model, we see that the final model is identically distributed with constant variance and no outlier. From the Q-Q plot and histogram of residuals, we find that the residuals are normally distributed, therefore, no truncated distribution problem and aligned to all the assumptions of the linear regression model.

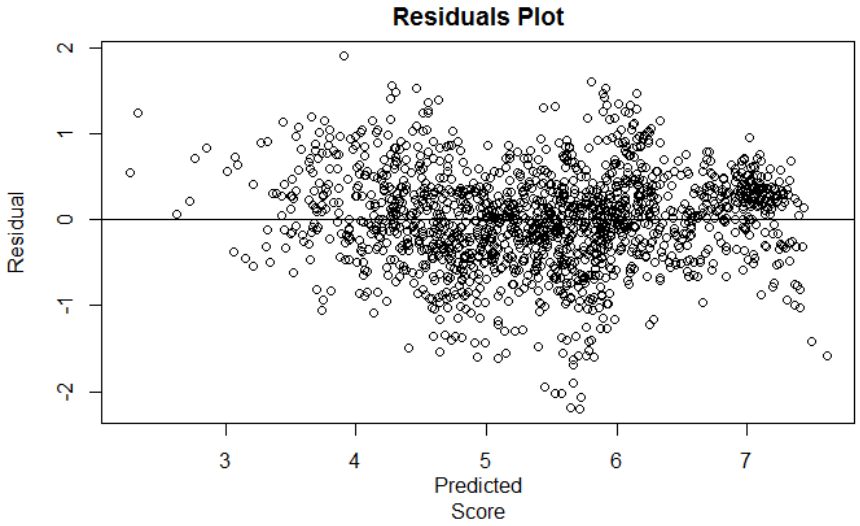

Figure 3: Residuals Plot

Using the final model, we make a prediction on happiness for the year 2019 and 2020 and plotted the regression line categorize by continent (Figure 4). We see that there are few outliers that have relative low happiness since some countries have missing data in some years result in irrational prediction. This shows the limitation of linear regression: missing data and outliers can affect the performance on prediction. Since 2005, we see that Oceania consistently ranks the highest in overall happiness, we believe the reason is Oceania only consists of 15 countries and thus the sample size is small compared to other Continent. Americas and Europe have similar happiness throughout the year. In addition, it seems to us that happiness does not increase nor decrease from year to year. This is surprising because as technology advances, people should have more time to enjoy life.

With curiosity, we plotted the linear regression line in GDP vs Happy (Figure 5). We see that most continents have a positive correlation between GDP and Happy. Again, Oceania shows a slightly negative correlation due to small samples resulted in a large variance.

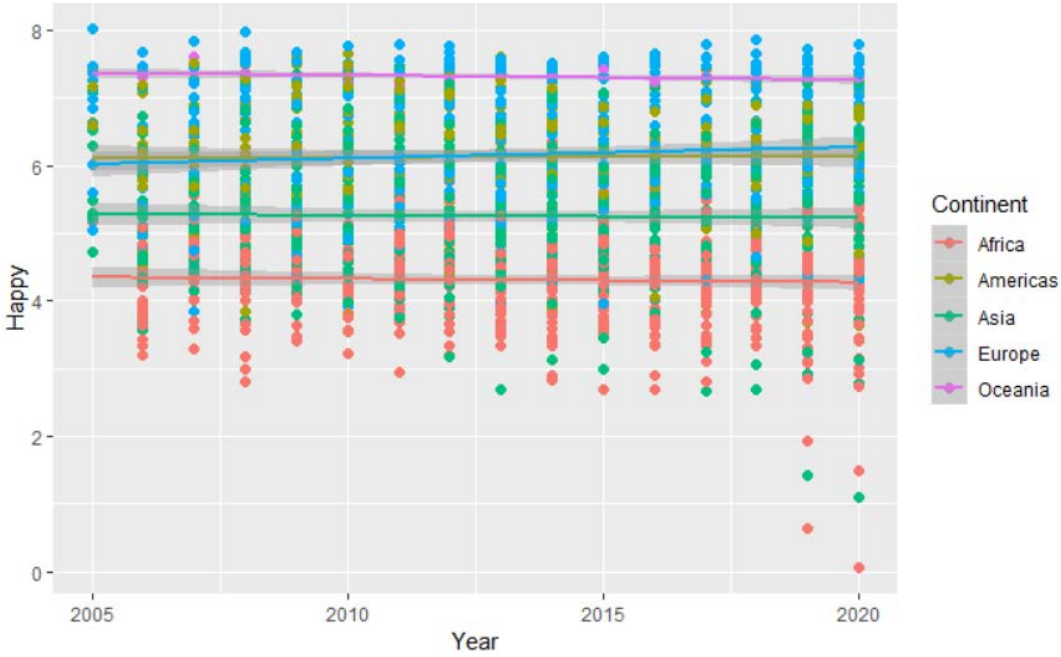

Figure 4: Happiness from 2005 to 2020 (2019 & 2020 are predictions)

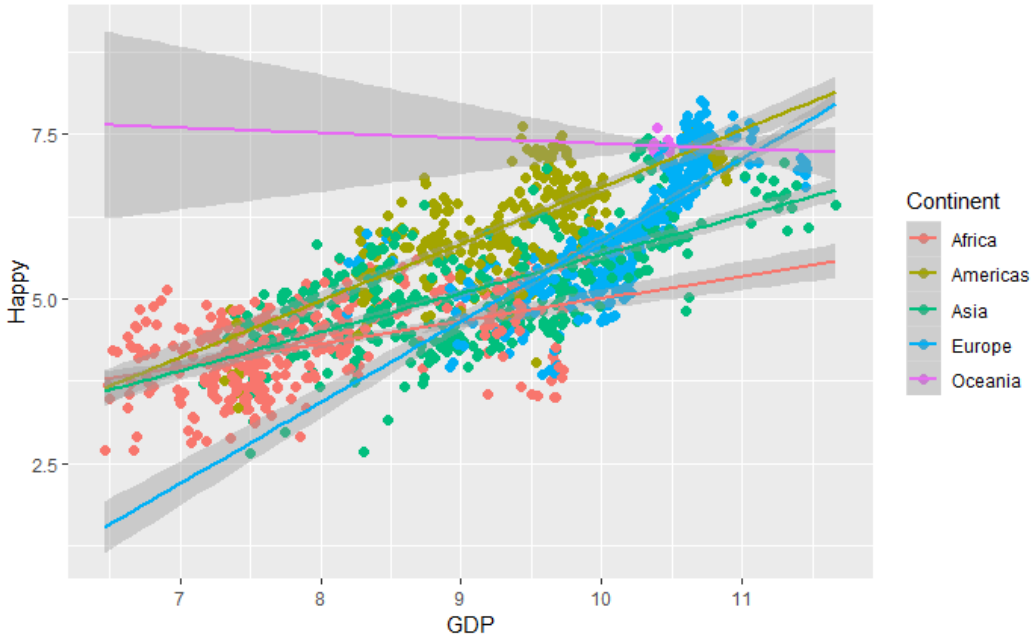

Figure 5: GDP vs Happiness

Furthermore, some of the variables we used in this project are subjective. For example, 'Social support', 'Freedom to Make Life Choices', 'Perceptions of Corruption', 'Positive Affect', and 'Negative Affect' are based on some simple questions. Also, the answers to these questions are just binary. We think that binary answers may not be accurate enough even though the survey has been done to many people.

6. FUTURE RESEARCH

In this project, we only use eight variables to approach the prediction model. However, we believe that there are still lots of factors that affect happiness. For example, in this prediction model, we didn't include variables like the unemployment rate, average work hour, and others which are also important to people's happiness. We think that in further research, we may collect more data about these variables of each year and each country, and consider their effect.

Also, due to the last part of the discussion. Some variable may be too subject. We think we could consider more kinds of datasets, comparing the quality of those datasets and choose more accuracy variables to process and analyze.

7. REFERENCE

[1] I Putu Angga K (2019). World Happiness Report 2019. From Kaggle: https://www.kaggle.com/jojoker/world-happiness-report-2019

[2] John F. Helliwell, Haifang Huang and Shun Wang (2019). Statistical Appendix 1 for Chapter 2 of World Happiness Report 2019. From Kaggle: https://www.kaggle.com/jojoker/world-happiness-report-2019

[3] John F. Helliwell, Haifang Huang and Shun Wang (2019). Statistical Appendix 2 for Chapter 2 of World Happiness Report 2019. From Kaggle: https://www.kaggle.com/jojoker/world-happiness-report-2019

[4] Wikipedia. Akaike information criterion. Wikipedia, The Free Encyclopedia. From: https://en.wikipedia.org/wiki/Akaike_information_criterion

8. APPENDIX

We use R language to complete this project. The codes are shown below.

```
summary(happiness)
```

```
##       Year          Happy           GDP             Social
##  Min.   :2005   Min.   :2.662   Min.   : 6.466   Min.   :0.2902
##  1st Qu.:2009   1st Qu.:4.602   1st Qu.: 8.270   1st Qu.:0.7474
##  Median :2012   Median :5.343   Median : 9.415   Median :0.8370
##  Mean   :2012   Mean   :5.446   Mean   : 9.218   Mean   :0.8116
##  3rd Qu.:2015   3rd Qu.:6.295   3rd Qu.:10.177   3rd Qu.:0.9055
```

```
##  Max.   :2018   Max.   :8.019   Max.   :11.770   Max.   :0.9873
##
##     Freedom         Corruption       Democracy          Positive
##  Min.   :0.2601   Min.   :0.0352   Min.   :-2.4482   Min.   :0.3625
##  1st Qu.:0.6382   1st Qu.:0.6868   1st Qu.:-0.7516   1st Qu.:0.6228
##  Median :0.7521   Median :0.7995   Median :-0.2128   Median :0.7207
##  Mean   :0.7351   Mean   :0.7426   Mean   :-0.1169   Mean   :0.7092
##  3rd Qu.:0.8503   3rd Qu.:0.8690   3rd Qu.: 0.6481   3rd Qu.:0.7995
##  Max.   :0.9852   Max.   :0.9833   Max.   : 1.5750   Max.   :0.9436
##
##     Negative            Country        Continent
##  Min.   :0.09549   Argentina :  13   Africa  :389
##  1st Qu.:0.20813   Armenia   :  13   Americas:279
##  Median :0.25776   Azerbaijan:  13   Asia    :462
##  Mean   :0.26873   Bangladesh:  13   Europe  :423
##  3rd Qu.:0.31900   Belarus   :  13   Oceania : 24
##  Max.   :0.70459   Bolivia   :  13
##                   (Other)   :1499
```

```
model1=lm(Happy~GDP+Social+Freedom+Corruption+Positive+Democracy+Negative, data=hap
piness)
summary(model1)
```

```
##
## Call:
## lm(formula = Happy ~ GDP + Social + Freedom + Corruption + Positive +
##     Democracy + Negative, data = happiness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2698 -0.3288  0.0319  0.3284  1.7555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.80781    0.22362  -8.084 1.24e-15 ***
## GDP          0.46075    0.01907  24.159  < 2e-16 ***
## Social       1.83722    0.18361  10.006  < 2e-16 ***
## Freedom      0.41003    0.14180   2.892 0.003885 **
## Corruption  -0.74985    0.09170  -8.177 5.93e-16 ***
```

```
## Positive     2.44030    0.17659  13.819  < 2e-16 ***
## Democracy    0.07861    0.04635   1.696 0.047958 *
## Negative     0.18291    0.19469   0.939 0.347622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5595 on 1569 degrees of freedom
## Multiple R-squared:  0.7613, Adjusted R-squared:  0.7602
## F-statistic: 714.8 on 7 and 1569 DF,  p-value: < 2.2e-16
```

```
model2=lm(Happy~GDP+Social+Freedom+Corruption+Positive, data=happiness)
summary(model2)
```

```
##
## Call:
## lm(formula = Happy ~ GDP + Social + Freedom + Corruption + Positive,
##     data = happiness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23208 -0.34332  0.03005  0.34079  1.87259
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.14985    0.18309 -11.742  < 2e-16 ***
## GDP          0.49000    0.01707  28.708  < 2e-16 ***
## Social       1.84518    0.17420  10.592  < 2e-16 ***
## Freedom      0.50652    0.13962   3.628 0.000295 ***
## Corruption  -0.73225    0.09014  -8.123 9.09e-16 ***
## Positive     2.47117    0.17398  14.203  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5614 on 1571 degrees of freedom
## Multiple R-squared:  0.7594, Adjusted R-squared:  0.7586
## F-statistic: 991.6 on 5 and 1571 DF,  p-value: < 2.2e-16
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: Happy
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## GDP          1 1264.02 1264.02 4011.07 < 2.2e-16 ***
## Social       1  116.91  116.91  370.98 < 2.2e-16 ***
## Freedom      1   96.04   96.04  304.76 < 2.2e-16 ***
## Corruption   1   21.85   21.85   69.33 < 2.2e-16 ***
## Positive     1   63.57   63.57  201.73 < 2.2e-16 ***
## Residuals 1571  495.07    0.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#temp=na.omit(happiness)
temp=happiness%>%select(.,-c("Continent","Country","Year","Negative","Democracy"))

null=lm(Happy~1,data=temp)
full=lm(Happy~., data=temp)

step(null,scope=list(lower=null,upper=full), direction="forward")
```

```
## Start:  AIC=421.4
## Happy ~ 1
##
##             Df Sum of Sq     RSS      AIC
## + GDP        1   1264.02  793.44 -1079.24
## + Social     1   1024.89 1032.57  -663.81
## + Positive   1    635.28 1422.18  -158.95
## + Freedom    1    591.42 1466.04  -111.06
## + Corruption 1    453.28 1604.18    30.95
## <none>                  2057.46   421.40
##
## Step:  AIC=-1079.24
## Happy ~ GDP
##
##             Df Sum of Sq    RSS     AIC
## + Positive   1   229.583 563.86 -1615.9
## + Freedom    1   153.709 639.73 -1416.8
## + Social     1   116.908 676.53 -1328.6
```

```
## + Corruption  1    65.984 727.46 -1214.2
## <none>                  793.44 -1079.2
##
## Step:  AIC=-1615.91
## Happy ~ GDP + Positive
##
##             Df Sum of Sq    RSS     AIC
## + Social     1    31.489 532.37 -1704.5
## + Corruption  1    25.065 538.79 -1685.6
## + Freedom    1    20.374 543.48 -1671.9
## <none>                  563.86 -1615.9
##
## Step:  AIC=-1704.53
## Happy ~ GDP + Positive + Social
##
##             Df Sum of Sq    RSS     AIC
## + Corruption  1    33.149 499.22 -1803.9
## + Freedom    1    16.502 515.87 -1752.2
## <none>                  532.37 -1704.5
##
## Step:  AIC=-1803.91
## Happy ~ GDP + Positive + Social + Corruption
##
##          Df Sum of Sq    RSS     AIC
## + Freedom  1    4.1477 495.07 -1815.1
## <none>                499.22 -1803.9
##
## Step:  AIC=-1815.07
## Happy ~ GDP + Positive + Social + Corruption + Freedom

##
## Call:
## lm(formula = Happy ~ GDP + Positive + Social + Corruption + Freedom,
##     data = temp)
##
## Coefficients:
## (Intercept)         GDP     Positive       Social   Corruption
##     -2.1498      0.4900       2.4712       1.8452      -0.7323
```

```
##     Freedom
##      0.5065
```

```
step(full, direction="backward")
```

```
## Start:  AIC=-1815.07
## Happy ~ GDP + Social + Freedom + Corruption + Positive
##
##             Df Sum of Sq    RSS     AIC
## <none>                   495.07 -1815.1
## - Freedom     1     4.148 499.22 -1803.9
## - Corruption  1    20.795 515.87 -1752.2
## - Social      1    35.355 530.43 -1708.3
## - Positive    1    63.573 558.65 -1626.5
## - GDP         1   259.711 754.78 -1152.0
```

```
##
## Call:
## lm(formula = Happy ~ GDP + Social + Freedom + Corruption + Positive,
##     data = temp)
##
## Coefficients:
## (Intercept)          GDP       Social      Freedom   Corruption
##     -2.1498       0.4900       1.8452       0.5065      -0.7323
##    Positive
##      2.4712
```

```
temp=happiness%>%group_by(Country)%>%summarize_all(.,mean)
```

```
pairs(temp[3:6])
```

```
pairs(temp[c(3,7:10)])
```

```
plot(fitted(model1),residuals(model1),xlab="Predicted
Score",ylab="Residual",main="Residuals Plot")
abline(h=0)
```

```
qqnorm(residuals(model2))
qqline(residuals(model2))
```

```
hist(residuals(model2))
```

#predicting 2019

```
#replace na with mean by country
happiness_update=summarize_at(happiness,c(2:9),funs(mean(.,na.rm=TRUE)))

## Warning: funs() is soft deprecated as of dplyr 0.8.0
## please use list() instead
##
##   # Before:
##   funs(name = f(.))
##
##   # After:
##   list(name = ~ f(.))
## This warning is displayed once per session.

for (r in 1:nrow(happiness))
{
  for(co in 2:(ncol(happiness)-2))
  {
    if(is.na(happiness[r,co])[1,1])
    {
      happiness[r,co]=happiness_update%>%filter(Country==happiness$Country[r])%>%se
lect(c(co))
    }
  }
}

countries=happiness%>%select(Country,Continent)%>%unique()
continents=happiness%>%select(Country,Continent)%>%unique()
features=list("GDP","Social","Freedom","Corruption","Democracy","Positive","Negativ
e")
init=TRUE
a=happiness
n=nrow(a)+1
for(yr in 2019:2020)
{
  for(c in 1:dim(countries)[1])
  {
    for(f in features)
```

```
    {
      if(init)
      {
        a[n,"Year"]=yr
        a[n,"Country"]=countries$Country[c]
        temp=continents%>%filter(Country==countries$Country[c])
        a[n,"Continent"]=temp$Continent
        init=FALSE
      }
      b=happiness%>%filter(Country==countries$Country[c])%>%lm(get(f)~Year,data=.)%
>%predict(.,data.frame(Year=yr),interval="none")
      a[n,f[[1]]]=b[[1]]
    }
    b=happiness%>%filter(Country==countries$Country[c])%>%
              lm(Happy~GDP+Social+Freedom+Positive,data=.)%>%
              predict(.,data.frame(GDP=a$GDP[n],Social=a$Social[n],
                 Freedom=a$Freedom[n],Positive=a$Positive[n]),interval="none")
    a$Happy[n]=b[[1]]
    init=TRUE
    n=n+1
  }
}

model_gdp=lm(GDP~Year, data=happiness)
model_soc=lm(Social~I(Year),data=happiness)
model_fre=lm(Freedom~I(Year),data=happiness)
model_cor=lm(Corruption~I(Year),data=happiness)
model_pos=lm(Positive~I(Year),data=happiness)

ggplot(data=happiness, aes(x=GDP,y=Happy,colour=Continent))+xlab("GDP")+geom_point
(size=2)+
  stat_smooth(method = 'lm',fullrange=TRUE)

ggplot(data=a, aes(x=Year,y=Happy,colour=Continent))+xlab("Year")+geom_point(size=
2)+stat_smooth(method = 'lm',fullrange=TRUE)
```