Predicting world happiness

- Authors: Jessica Guesman and Zachary VonStein
- DS-300 METHODS OF DATA SCIENCE & ANALYTICS
- May 4th 2021

Predicting World Happiness

- We will be using multiple linear regression and polynomial regressions to build a ladder score in order to predict world happiness.
- since we are using multiple linear regression we will also be using backwards elimination.

Looking at the Data set

- used two data sets from a collection of data sets from Kaggle.
- the data is from the years of 2020 and 2021
- uncleaned, the data sets start off with 20 variables.

Cleaning the Data Sets

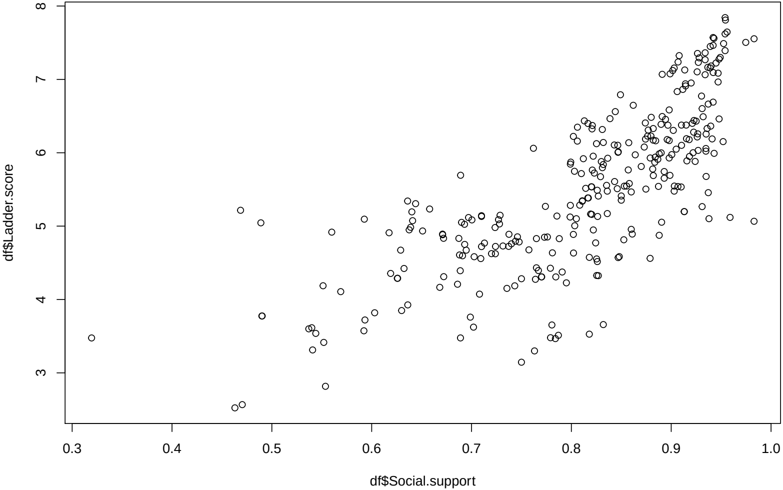
- Began with renaming the columns to match the variables.
- Then columns were removed. These inculde Standard.error.of.the.ladder.score, upperwhisker, lowerwhisker, Ladder.score.in.dystopia, Explained.by.Logged.GDP.Per.Capita, Explained.by.Social.support, Explained.by.Healthy.life.expectancy, Explained.by.Freedom.to.make.life.choices, Explained.by.Generosity, Explained.by.Perceptions.of.corruption, Dystopia.Residual, Country.name, and Regional Indicator.
- Then finally we bind the datasets together.

Why these were removed

- Country.name was removed because it holds about 150 unique variables.
- Regional.indicator removed because it holds 9 different variables
- All variables starting with explained, and Dystopia.Residual were removed because there was no information on what they were.
- Removed Standard.error.of.the.ladder.score, upperwhisker, and lower whisker because they were not relevant variables for our research question.

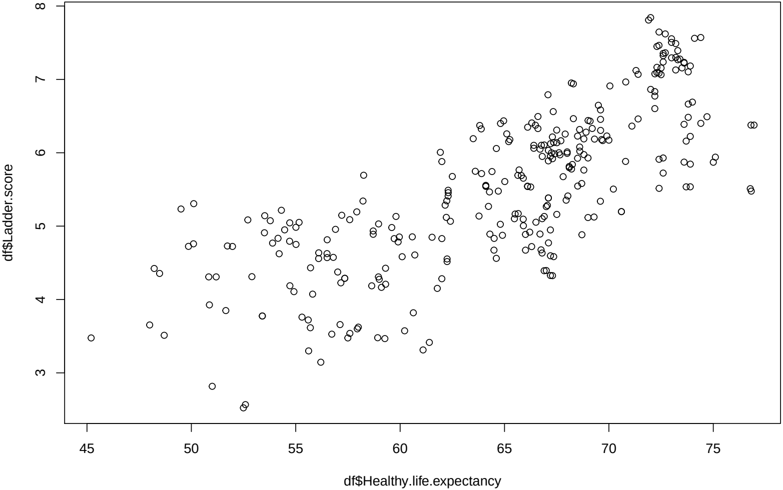
EDA

```r
plot(x = df$Social.support, y = df$Ladder.score)
```



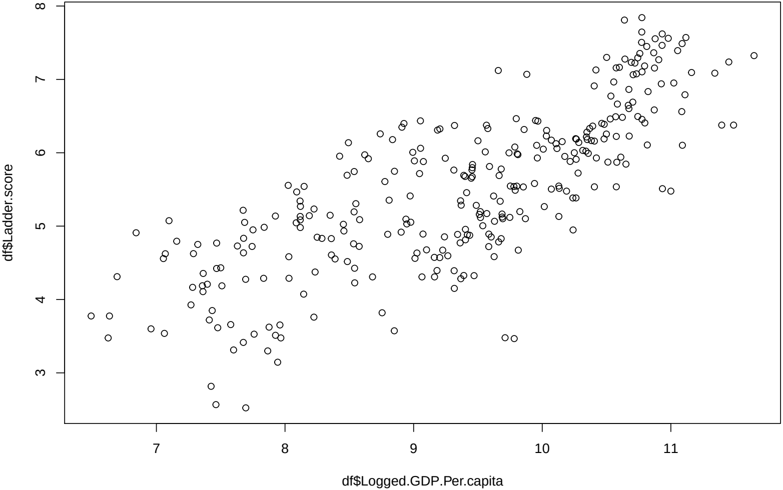
EDA Continued

```r
plot(x = df$Healthy.life.expectancy, y = df$Ladder.score)
```



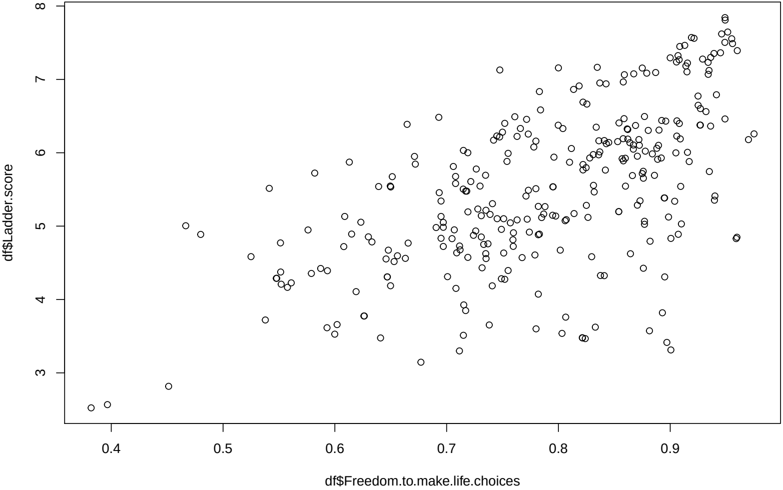
EDA Continued

```r
plot(x = df$Logged.GDP.Per.capita, y = df$Ladder.score)
```



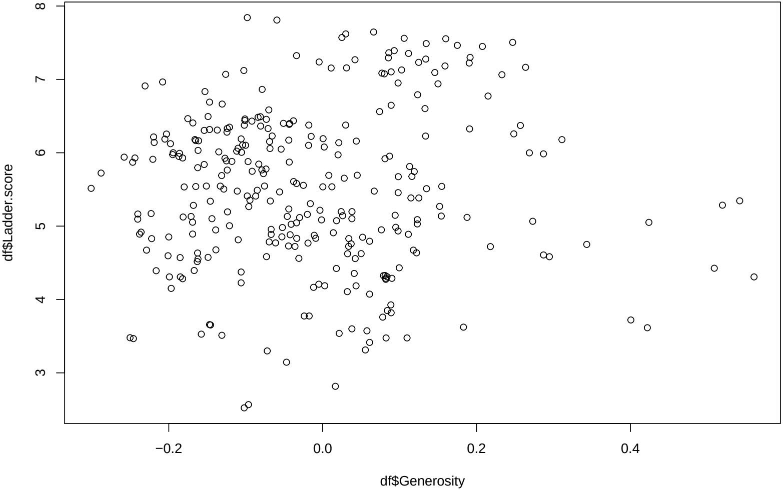
EDA Continued

```
plot(x = df$Freedom.to.make.life.choices, y = df$Ladder.sco
```



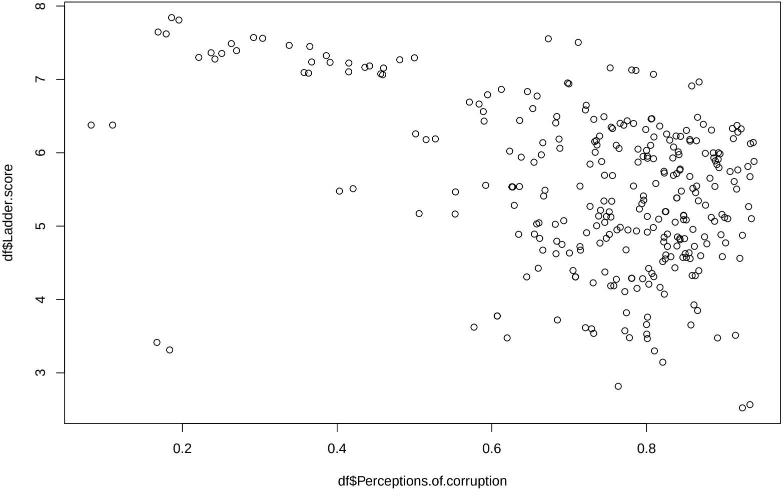
EDA Continued

```r
plot(x = df$Generosity, y = df$Ladder.score)
```



EDA Continued

```r
plot(x = df$Perceptions.of.corruption, y = df$Ladder.score)
```



Multiple Linear Regression

- All variables were made to be numeric if they weren't already.
- split the data into training and test sets
- All remaining variables were written into the formula for multiple regression

Backwards elimination for the multiple Regression

- We chose a p value of .05 for backwards elimination, and only one variable exceeded that p value
- Generosity exceeded the p value and was removed. And no other variables exceeded the p value after Generosity was removed, so that was the end of backwards elimination

Results of Multiple Regression

- Looking at the adjusted R value of .7388 from the summary of our regression, we can say that this is a decent model to predicting Happiness. If we could increase the adjusted R value our model would be an even better fit.

Polynomial Regression

- We decided to do polynomial regression as well since some of the scatter plots in our Exploratory Data Analysis look like the could follow the line of a parabola.
- The variables that followed that shape were Social.support, and Healthy.life.expectancy.
- So we fit those two to polynomial regression.

Backwards Elimination for Polynomial Regression.

- We began by removing Perceptions.of.corruption because it exceeded the p value by almost 20 percent.
- We then had to do backwards elimination on Generosity because it was also greater than the p value, although it was very close to it. We could have kept it since it was close .05 but we decided to get rid of it.

Results of the Polynomial Regression

- We then pulled the summary of the polynomial regression and looked at the reported adjusted r value, which is .763.
- Before backwards elimination our adjusted R squared was .764.
- Ultimately not much change occured when removing the variables Perception.of.corruption, and Generosity.

Comparing the Two Regressions

- When we compare the adjusted R squares of each of the regressions we can see that the polynomial regression would be better at predicting world happiness than the Multiple regression would be.