DAT-4253 LM 9 - Lab 1: Regression with Interaction

AUTHOR
Aaron Younger

Business Understanding

This analysis explores the relationship between income, education, and social connections and its potential influence of on health. This analysis will be used as evidence to support the validity of a campaign promoting social connectedness as a way to improve health. The campaign is sponsored by Jack Person, executive director of a public policy organization, who seeks to understand whether social connectedness impacts health while controlling factors of income and education.

Data Understanding

R Version

```
suppressWarnings(RNGversion("3.5.3"))
```

Libraries

```
library(readxl)
library(DataExplorer)
library(tidyverse)
library(dplyr)
library(e1071)
library(dlookr)
library(psych)
library(moments)
library(ggplot2)
library(caret)
library(Metrics)
library(car)
library(auditor)
```

Import Dataset

```
library(readxl)
health_data <- read_excel("jaggia_ba_2e_ch08_data.xlsx",
    sheet = "Health_Factors")
View(health_data)
```

Explore Dataset

```
health_data %>% head()
```

```
# A tibble: 6 × 4
  Health Social Income College
   <dbl>  <dbl>  <dbl>   <dbl>
1     52     58     80       0
2     55     68     43       0
3     80     66    284       0
4     93     67    159       1
5     55     94     70       0
6     92     65    270       1
```

```
health_data %>% tail()
```

```
# A tibble: 6 × 4
  Health Social Income College
   <dbl>  <dbl>  <dbl>   <dbl>
1     73     89     38       1
2     60     74     78       0
3     96     79    296       1
4     96     77    178       1
5     66     67     90       1
6     58     90     35       0
```

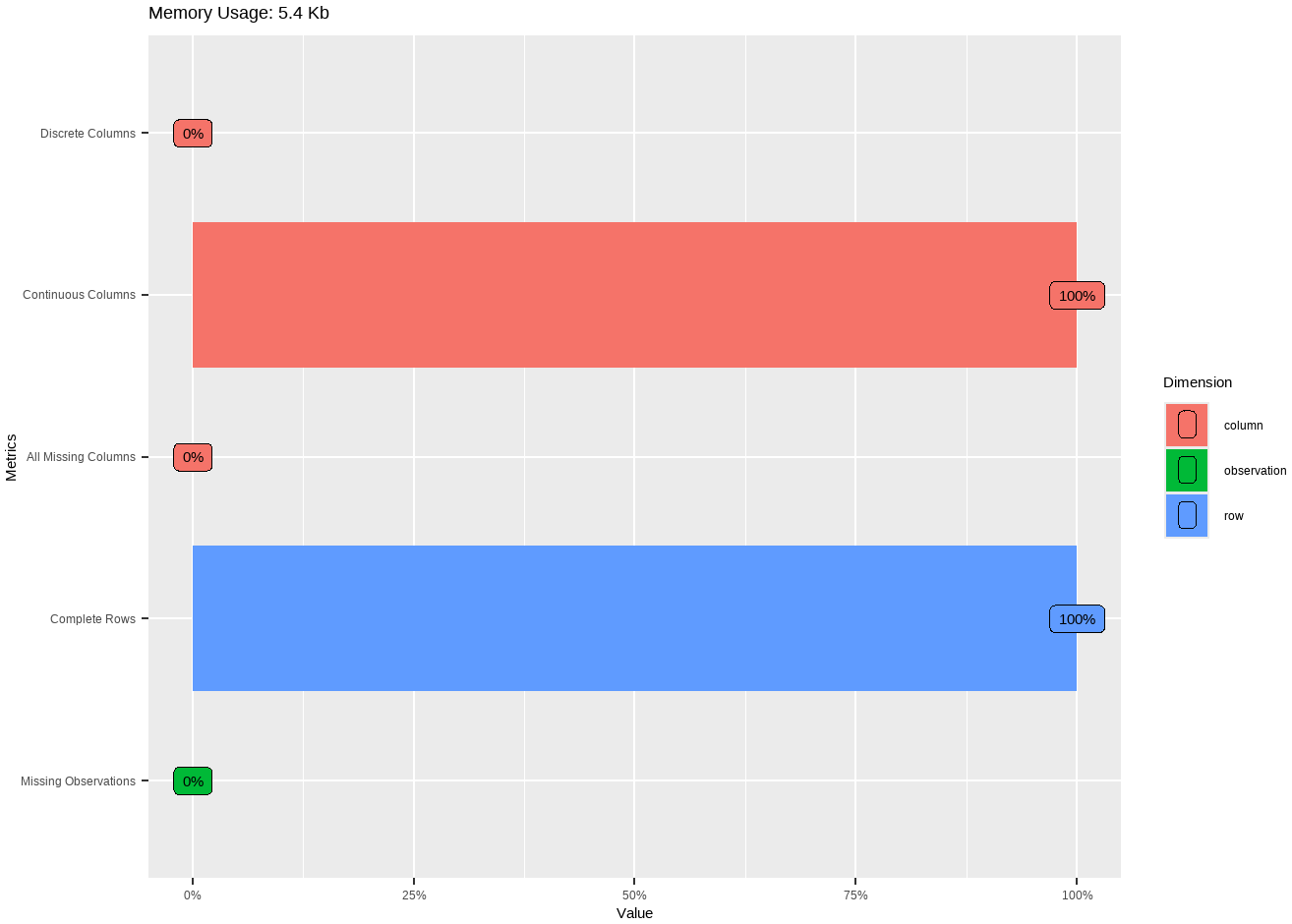
```
health_data %>% nrow()
```

```
[1] 120
```

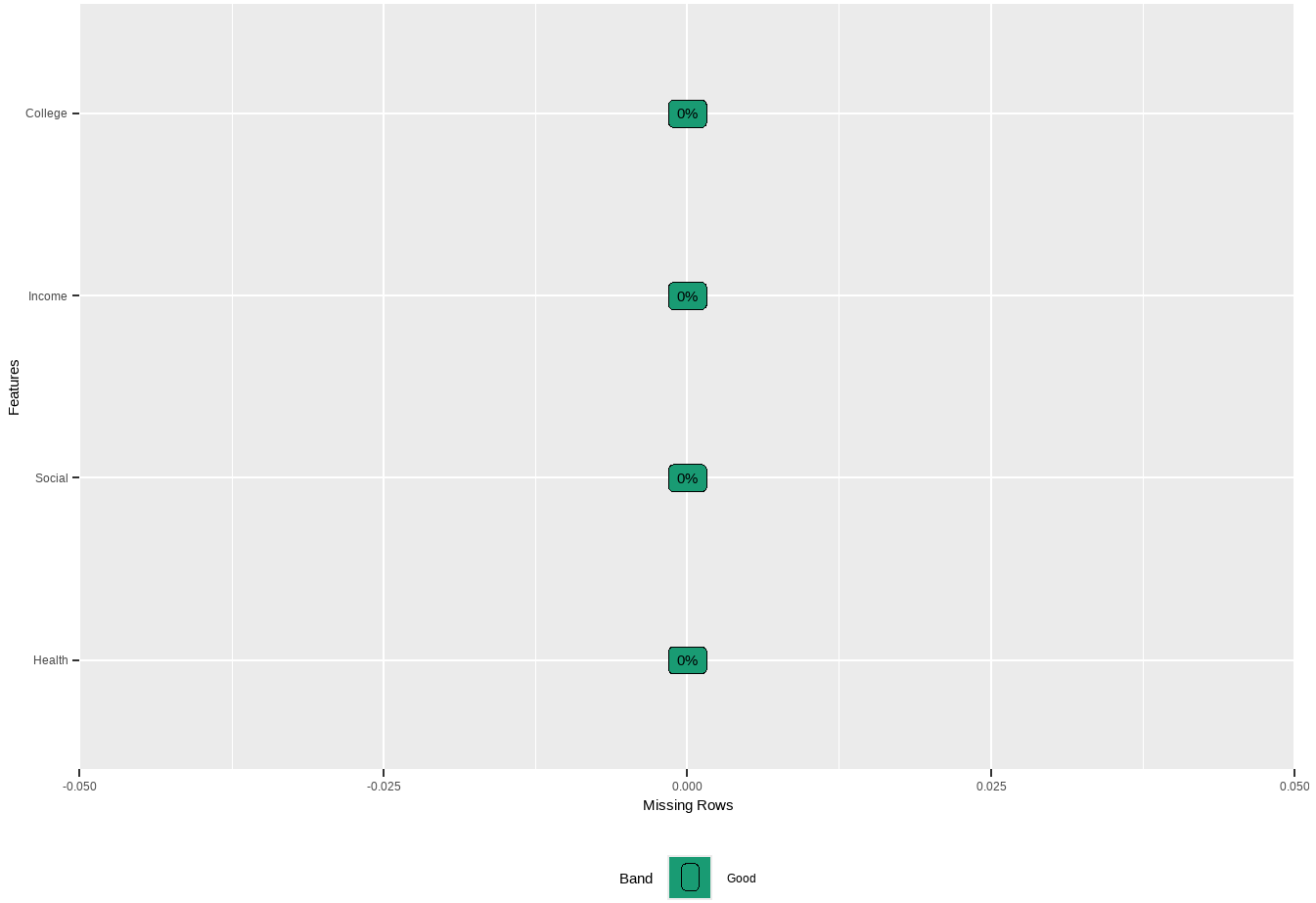
```
health_data %>% ncol()
```

```
[1] 4
```

```
health_data %>% plot_intro()
```




```
health_data %>% plot_missing()
```




```
health_data %>% str()
```

```
tibble [120 × 4] (S3: tbl_df/tbl/data.frame)
 $ Health : num [1:120] 52 55 80 93 55 92 94 81 60 97 ...
 $ Social : num [1:120] 58 68 66 67 94 65 59 96 96 85 ...
 $ Income : num [1:120] 80 43 284 159 70 270 288 57 124 219 ...
 $ College: num [1:120] 0 0 0 1 0 1 1 1 0 1 ...
```

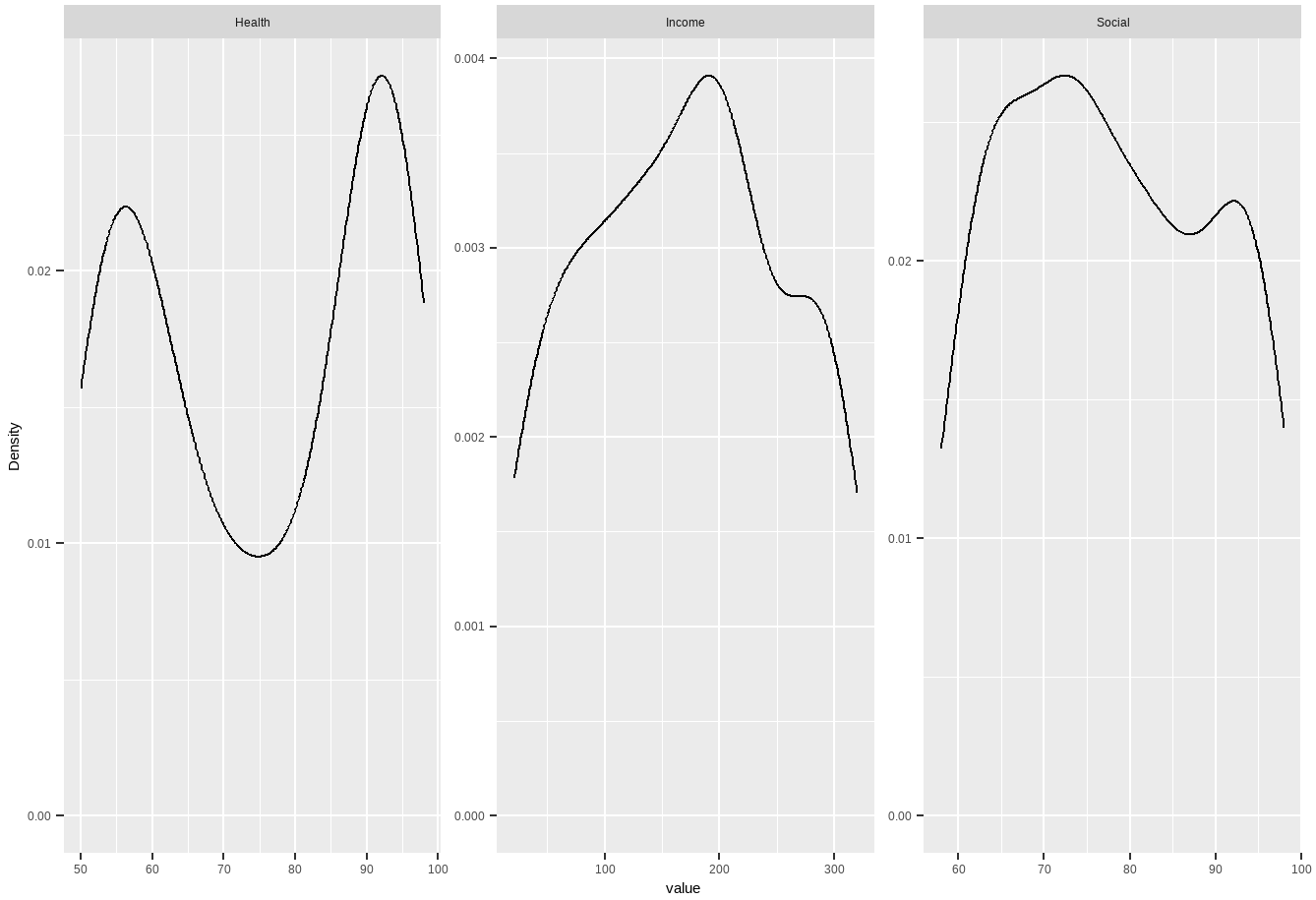
Comments about Dataset:
This is survey data. This data has 120 observations and four columns, health, income, social, and college. Every variable in this dataset is continuous/numeric, with the college variable being a 1/0 binary. This dataset has no missing values.

Variable Key:
- Social: Measured on a scale of 1-100 represents the social activity of a person.
- College: A binary variable (yes/no variable) representing if a person went to college or not.
- Income: Represents a persons income, variable in 1000's.
- Health: Measured on a scale of 1-100 represents the health of a person. This variable is the dependent variable.

EDA

Distribution of Numeric Variables

```
health_data %>% plot_density()
```



Comments on distribution of numeric variables:
The distribution of Income and Social look relatively normal with no clear skewness. Health has a bimodal distribution indicating two distinct clusters within the data, lower and higher health.

Skewness of Numeric Variables

```
apply(health_data[, 1:3], 2, skewness)
```

```
      Health       Social       Income 
-0.118983998  0.100112731 -0.004555104 
```

Comments on skewness of numeric variables:
The skewness values (-0.12 for Health, -0.005 for Income, and 0.10 for social) are all close to zero showing that skewness is not an issue in this data and normality is reasonable.

Kurtosis

```
apply(health_data[, 1:3], 2, kurtosis)
```

```
  Health   Social   Income 
1.330800 1.787957 1.927356 
```

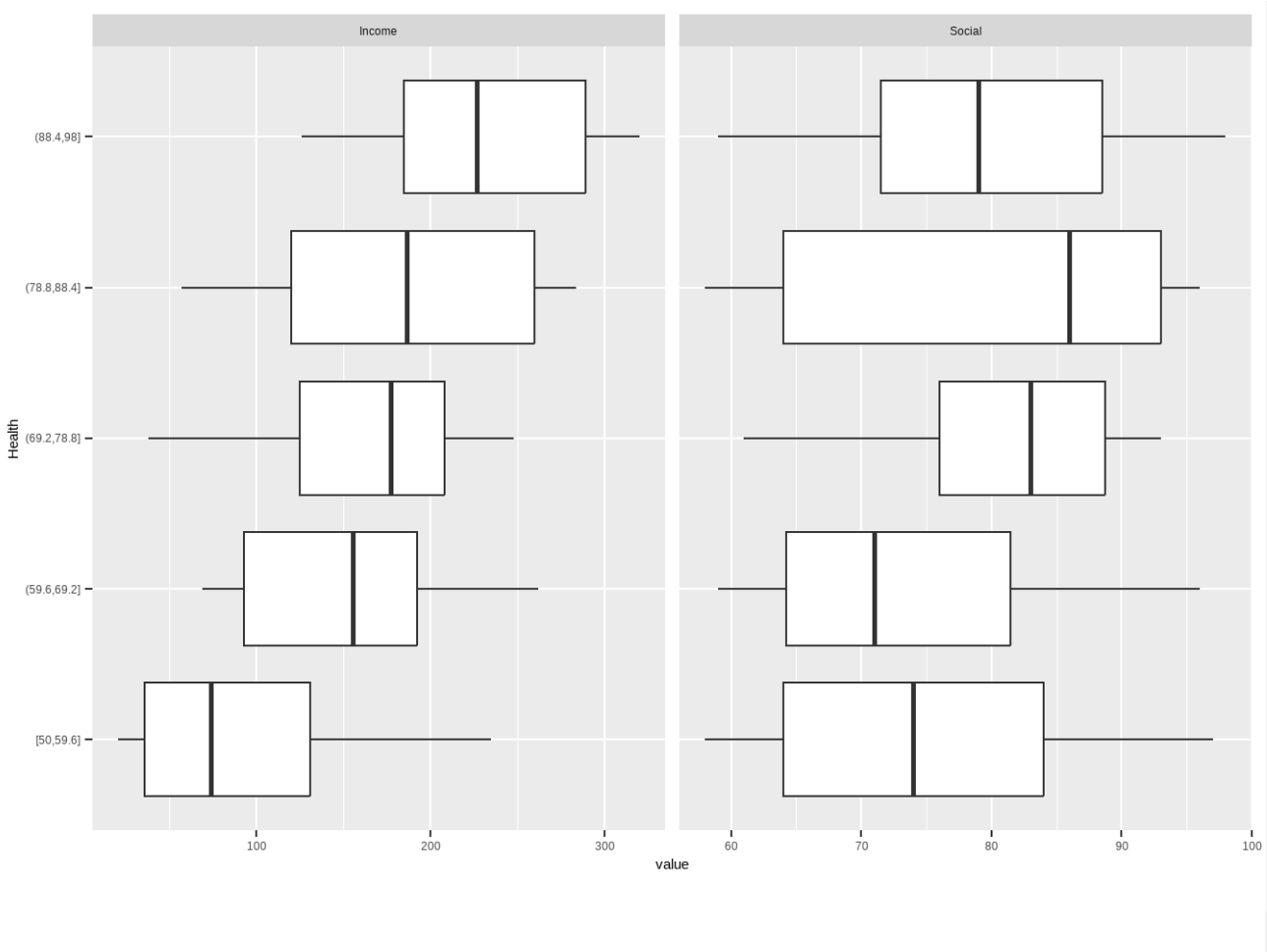
Comments on Kurtosis of numeric variables:
Although the kurtosis values indicate that distributions are slightly playkurtic, they are not extreme enough to assume non-normality.

Look for Potential Outliers

```
diagnose_outlier(health_data)
```

```
# A tibble: 4 × 6
  variables outliers_cnt outliers_ratio outliers_mean with_mean without_mean
  <chr>            <int>          <dbl>         <dbl>     <dbl>        <dbl>
1 Health               0              0           NaN    75.2         75.2  
2 Social               0              0           NaN    77.6         77.6  
3 Income               0              0           NaN   169.         169.   
4 College              0              0           NaN     0.558        0.558
```

```
health_data %>% plot_boxplot(by="Health")
```



Comments on outliers:
There are no outliers present in this data.

Look for Interaction Effects

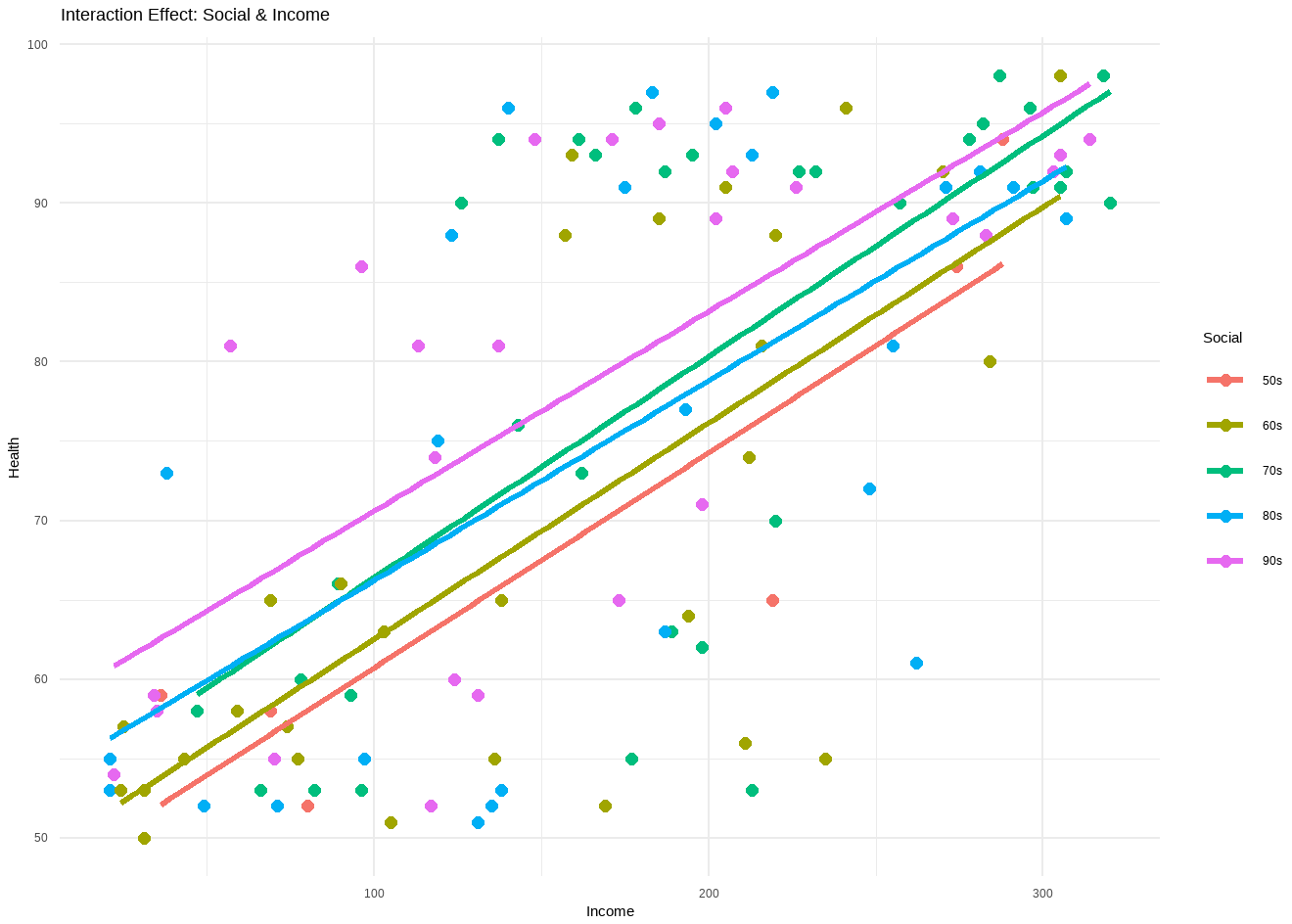
Social and Income

```
range(health_data$Social)
```

```
[1] 58 98
```

```
# Graph to find interaction effect
health_data$Socialgroups <- cut(
  health_data$Social,
  breaks = c(49,59,69,79,89,99),
  labels = c("50s", "60s", "70s", "80s", "90s")
)

ggplot(health_data, aes(x = Income, y = Health, color = Socialgroups)) +
  geom_point(size = 2)+
  geom_smooth(method = "lm", se = FALSE, linewidth = 1.2)+
  labs(
    title = "Interaction Effect: Social & Income",
    x = "Income",
    y = "Health",
    color = "Social"
  ) +
  theme_minimal()
```




```
# Use statistical test to find interaction effect

lrm_is <- lm(Health ~ Income + Social, data = health_data)
lrm_iis <- lm(Health ~ Income*Social, data = health_data)
anova(lrm_is, lrm_iis)
```

```
Analysis of Variance Table

Model 1: Health ~ Income + Social
Model 2: Health ~ Income * Social
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    117 17541
2    116 17490  1    50.449 0.3346 0.5641
```

Comments on interaction effect between Social and Income:

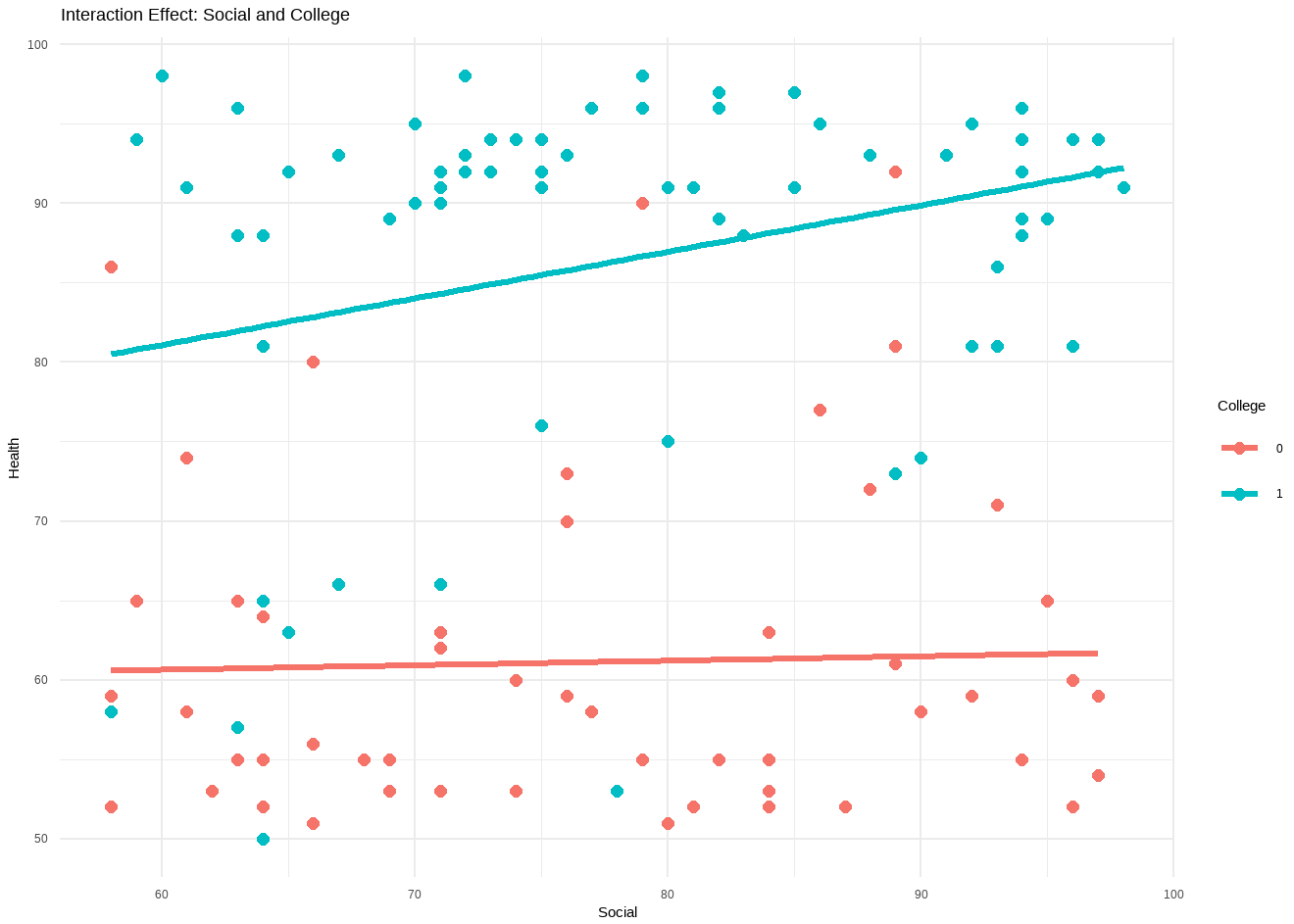
Based on the scatterplot, the regression lines for each Social group have similar slopes with very little intersection. This indicates that the relationship between Income and Health is consistent across levels of Social groups. Therefore, it is not reasonable to conclude that an interaction effect exists between Social and Income based solely on the graph. To confirm this statistically, a two-model comparison was conducted. The first model included Health as the dependent variable and Income and Social as independent variables. The second model included the same predictors but also added the interaction term (Income × Social). An ANOVA was then performed to determine whether including the interaction term significantly improved model fit. Because the p-value from the ANOVA test was greater than 0.05 it was not

statistically significant, adding the interaction term did not improve the model. Therefore, there is no evidence of an interaction effect between Income and Social.

Social and College

```
health_data$College <- as.factor(health_data$College)

ggplot(health_data, aes(x = Social, y = Health, color = College)) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1.2) +
  labs(
    title = "Interaction Effect: Social and College",
    x = "Social",
    y = "Health",
    color = "College"
  ) +
  theme_minimal()
```




```
# ANOVA Test

lrm_sc <- lm(Health ~ Social + College, data = health_data)
```

```
lrm_ssc <- lm(Health ~ Social*College, data = health_data)
anova(lrm_sc, lrm_ssc) ## Greater than 0.05
```

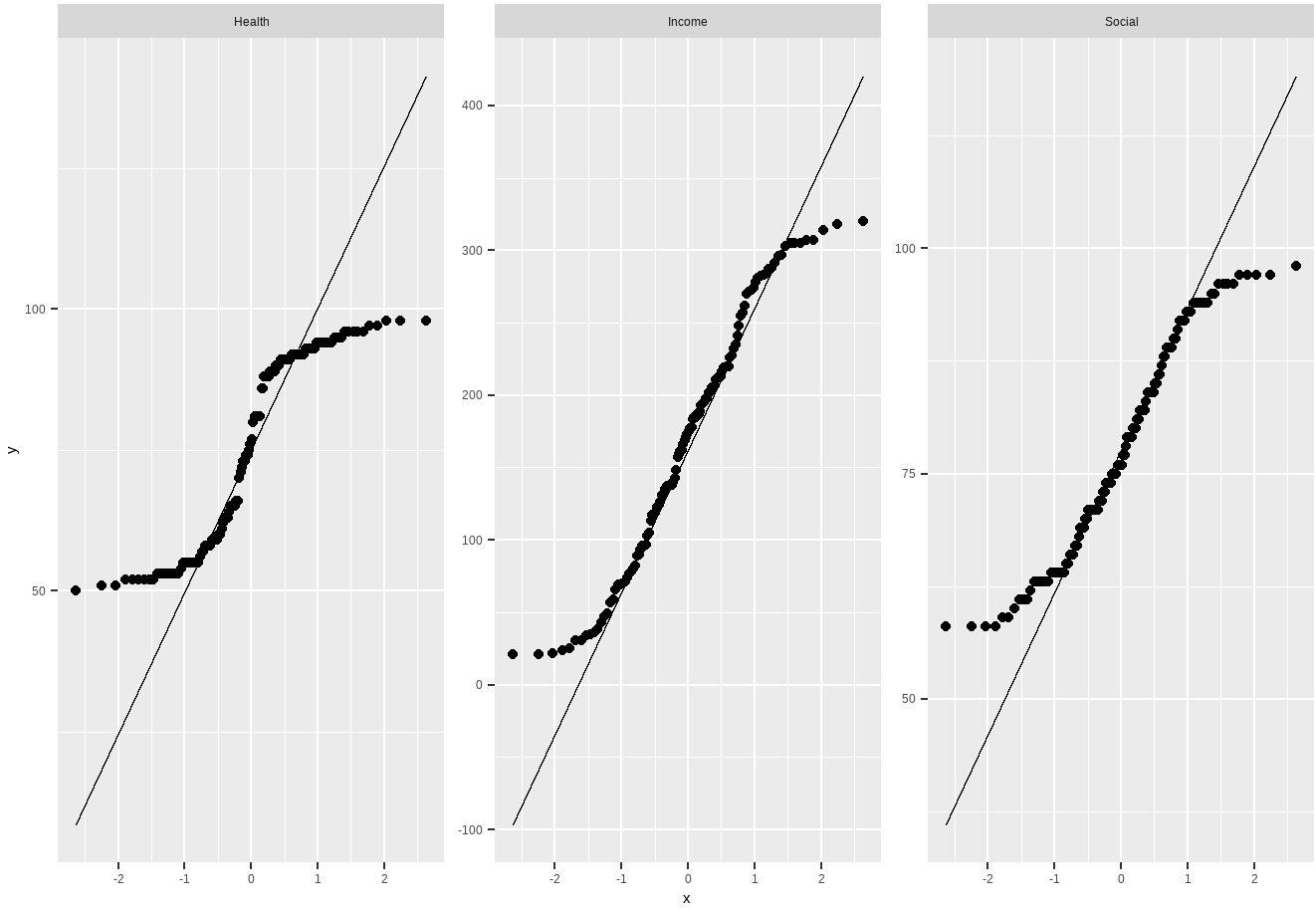
```
Analysis of Variance Table

Model 1: Health ~ Social + College
Model 2: Health ~ Social * College
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    117 15024
2    116 14732  1    291.57 2.2958 0.1324
```

Comments on Interaction Effect Between Social and College:
Based on the scatterplot, the regression lines for binary variable college show different slope with individuals with college education having a steeper positive slope than those without college education. The graph visually suggests there is an interaction effect. To verify this I ran a two model test, one model included health as the dependent variable and Social and College as the predictor variables, the second model has the same variables but Social and College were multiplied to assume an interaction affect. These models were then put into an ANOVA test, to test if the interaction effect gave better model fit and significance. The ANOVA test had a p-value greater than 0.05 indicating the interaction effect did not significantly improve model fit. So even though the scatterplot suggests an interaction effect occurs, the ANOVA test showed interaction between social and college was no significant.

Plot-QQ

```
health_data$College <- as.factor(health_data$College)
health_data %>% plot_qq()
```

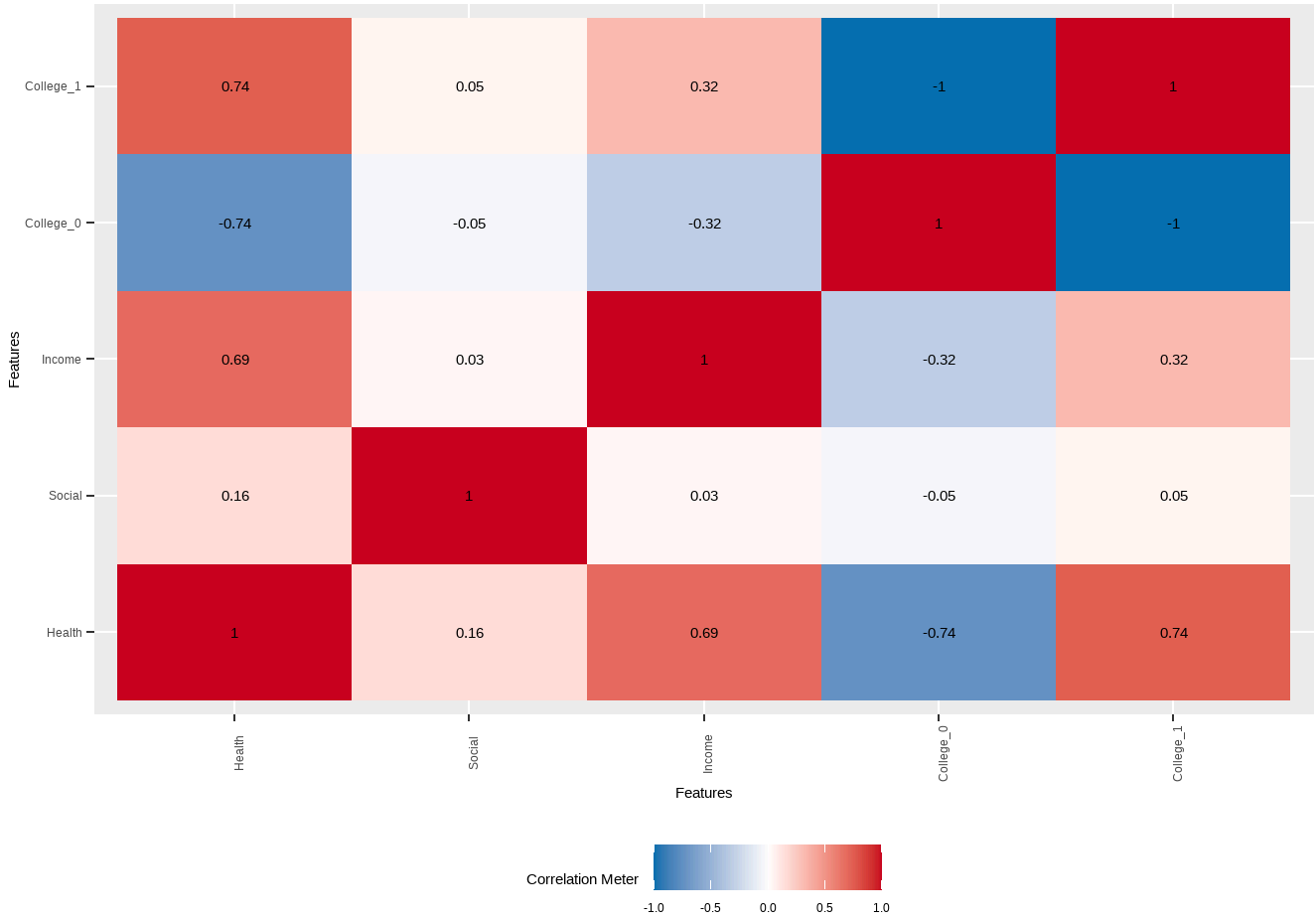


Comments on QQ-Plots:
The Q-Q Plots show that all three variables generally follow a normal distribution, there is some slight deviation at the tails but the mid points do fall along the middle line. Health does deviate the most at the tails but normality can be assumed for these three variables.

Correlation Matrix

```
health_data <- health_data %>%
  select(-Socialgroups)

DataExplorer::plot_correlation(health_data)
```



Comments on Correlation Matrix:
The correlation Matrix shows that all predictor variables have a positive relationship with the dependent variable health. Income and individuals who have gone to college show strong positive correlation with the dependent variable. Social shows weak positive correlation with Health. Also the Social variable shows almost no correlation with income or college possibly supporting the fact of no interaction effect between income and college.

Data Preparation

Partition

```
set.seed(1)

health_data$College <- as.factor(health_data$College)

my_index <- createDataPartition(health_data$Health, p = 0.8, list = FALSE)
trainset <- health_data[my_index, ]
testset <- health_data[-my_index, ]
```

```
## Means of variables are all very close
mean(health_data$Health)
```

```
[1] 75.225
```

```
mean(trainset$Health)
```

```
[1] 75.18557
```

```
mean(testset$Health)
```

```
[1] 75.3913
```

Comments on Data Partition:
Before modeling the dataset is partitioned into a 80/20 split so the data can be trained then tested. A set.seed of 1 was also given for reproducibility of the model results. The mean of the dependent variable across train and test set was very close.

Model - Assume a 10% Level of Significance

First Linear Regression Model

```
lr_ctrl <- trainControl(method = "cv", number = 10)

model1 <- train(Health~., data = trainset, method = "lm", trControl = lr_ctrl)
summary(model1)
```

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.0524  -4.9491   0.4468   5.1155  15.7155 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.541630   5.250000   6.389 6.55e-09 ***
Social       0.196554   0.064811   3.033  0.00314 ** 
Income       0.089856   0.009414   9.545 1.87e-15 ***
College1    20.249077   1.676365  12.079  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.55 on 93 degrees of freedom
```

```
Multiple R-squared:  0.8108,    Adjusted R-squared:  0.8047
F-statistic: 132.9 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
car::vif(model1$finalModel)
```

```
  Social   Income College1
1.004545 1.171930 1.173952
```

Assessing Coefficient Significance and Interpreting Each Significant Coefficient, Model 1:
For this model which uses no interaction effect and assumes independence among all predictor variables, every predictor variable is significant.
- Social is a significant predictor of health because its p-value is less than 0.10. On average for every one unit increase in Social, there is a 0.197 unit increase in health ceteris paribus.
- Income is a significant predictor of health because its p-value is less than 0.10. On average for every one unit increase in Income, there is a 0.09 unit increase in health ceteris paribus.
- Individuals attending college is a significant predictor of health for health because its p-value is less than 0.10. On average individuals that attend college have a health scores 20.25 units higher than those who have not attented college ceteris paribus.

There is also no multicollinearity between the variables.

Second Linear Regression Model

```
model2 <- train(Health~ Social*Income + Social*College, data = trainset, method = "lm",
         trControl = lr_ctrl)
summary(model2)
```

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0669  -4.7055   0.2419   4.4065  18.8887

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       28.1140494 10.5014791   2.677  0.00881 **
Social             0.2660099  0.1339642   1.986  0.05008 .
Income             0.1870529  0.0598664   3.125  0.00239 **
College1           1.7295130 10.9215156   0.158  0.87453
`Social:Income`   -0.0012491  0.0007595  -1.645  0.10350
`Social:College1`  0.2379433  0.1385894   1.717  0.08940 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.465 on 91 degrees of freedom
```

```
Multiple R-squared:  0.8191,    Adjusted R-squared:  0.8091
F-statistic: 82.38 on 5 and 91 DF,  p-value: < 2.2e-16
```

Assessing coefficient Significance and interpreting each significant coefficient, Model 2:
This Model assumes interaction effect between social and income and social and college. Due to the inclusion of interaction effects in this model not every variable is significant.
- Social is a significant predictor of health because its p-value is less than 0.10. On average for every one unit increase in social, there is a 0.266 unit increase in health ceteris paribus.
- Income is a significant predictor of health because its p-value is less than 0.10. On average for every one unit increase in income, there is a 0.187 unit increase in health ceteris paribus.
- College1 is not a significant predictor of health because its p-value is greater than 0.10. - Social:Income the interaction effect of Social and Income is not a significant predictor of Health because its p-value is greater than 0.10.
- Social:College1 the interaction effect of social and individuals is statistically significant. This means that the effect of Social on Health differs between college-educated and non-college individuals. Specifically, for individuals who have attended college, each one-unit increase in Social is associated with an additional 0.238-unit increase in Health compared to those without college education.

Compare Models Using Adjusted R2

```
m1 <- model1$finalModel
m2 <- model2$finalModel

adjR2_m1 <- summary(m1)$adj.r.squared
adjR2_m2 <- summary(m2)$adj.r.squared
c(adjR2_m1, adjR2_m2)
```

```
[1] 0.8047339 0.8091078
```

```
anova(m1, m2)
```

```
Analysis of Variance Table

Model 1: .outcome ~ Social + Income + College1
Model 2: .outcome ~ Social + Income + College1 + `Social:Income` + `Social:College1`
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     93 5301.6
2     91 5071.4  2    230.22 2.0655 0.1327
```

Comments on Model Fit using Adjusted R2 and ANOVA Test:
The adjusted R^2 values were 0.8047 for model 1 and 0.8091 for model 2, which is a very small increase. However, Model 2 explains slightly more variation within the dataset than model 1. The ANOVA test shows that the interaction effects in model 2 does not significantly improve model fit due to the p-value (0.1327) being greater than 0.10.

Evaluation

Predict Model 1 and Model 2

```
pred_m1 <- predict(model1, newdata = testset)
pred_m2 <- predict(model2, newdata = testset)

test_results <- cbind(
  testset,
  Pred_m1 = pred_m1,
  Pred_m2 = pred_m2
)
View(test_results)
```

Evaluation Metrics For Both Models

```
## RMSE Metrics
RMSE_m1 <- rmse(testset$Health, pred_m1)
RMSE_m2 <- rmse(testset$Health, pred_m2)

## MAE Metrics
MAE_m1 <- mae(testset$Health, pred_m1)
MAE_m2 <- mae(testset$Health, pred_m2)

## MAD Metrics
MAD_m1 <- mad(testset$Health, pred_m1)
MAD_m2 <- mad(testset$Health, pred_m2)

## MAPE Metrics
MAPE_m1 <- mape(testset$Health, pred_m1)
MAPE_m2 <- mape(testset$Health, pred_m2)

## Make table with values
metrics_table <- data.frame(
  Model = c("Model 1", "Model 2"),
   RMSE = c(RMSE_m1, RMSE_m2),
   MAE = c(MAE_m1, MAE_m2),
   MAD = c(MAD_m1, MAD_m2),
   MAPE = c(MAPE_m1, MAPE_m2)
)
metrics_table
```

```
    Model      RMSE      MAE      MAD      MAPE
1 Model 1 9.316092 8.144582 11.81140 0.1149739
2 Model 2 8.673423 7.527340 11.40091 0.1061629
```

REC Plot and Values

```
m1_audit <- audit(model1, data = testset, y = testset$Health)
```

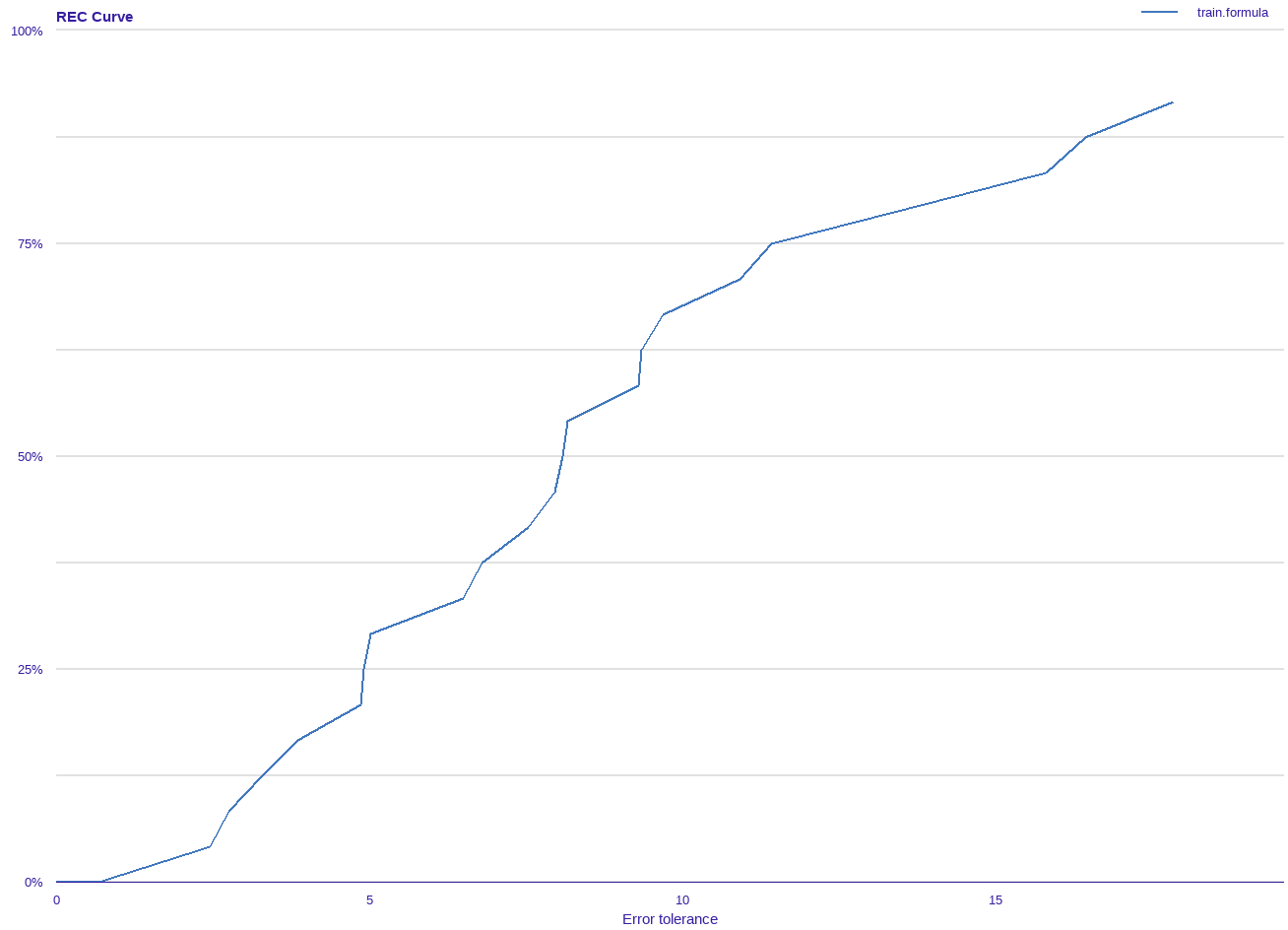
```
Preparation of a new explainer is initiated
  -> model label       :  train.formula  (  default  )
  -> data              :  23  rows  4  cols
  -> data              :  tibble converted into a data.frame
  -> target variable   :  23  values
  -> predict function  :  yhat.train  will be used (  default  )
  -> predicted values  :  No value for predict function target column. (  default  )
  -> model_info        :  package caret , ver. 6.0.94 , task regression (  default  )
  -> predicted values  :  numerical, min =  55.0954 , mean =  75.60388 , max =  97.69605
  -> residual function :  difference between y and yhat (  default  )
  -> residuals         :  numerical, min =  -15.82296 , mean =  -0.2125714 , max =  17.83759
  A new explainer has been created!
```

```
m2_audit <- audit(model2, data = testset, y = testset$Health)
```

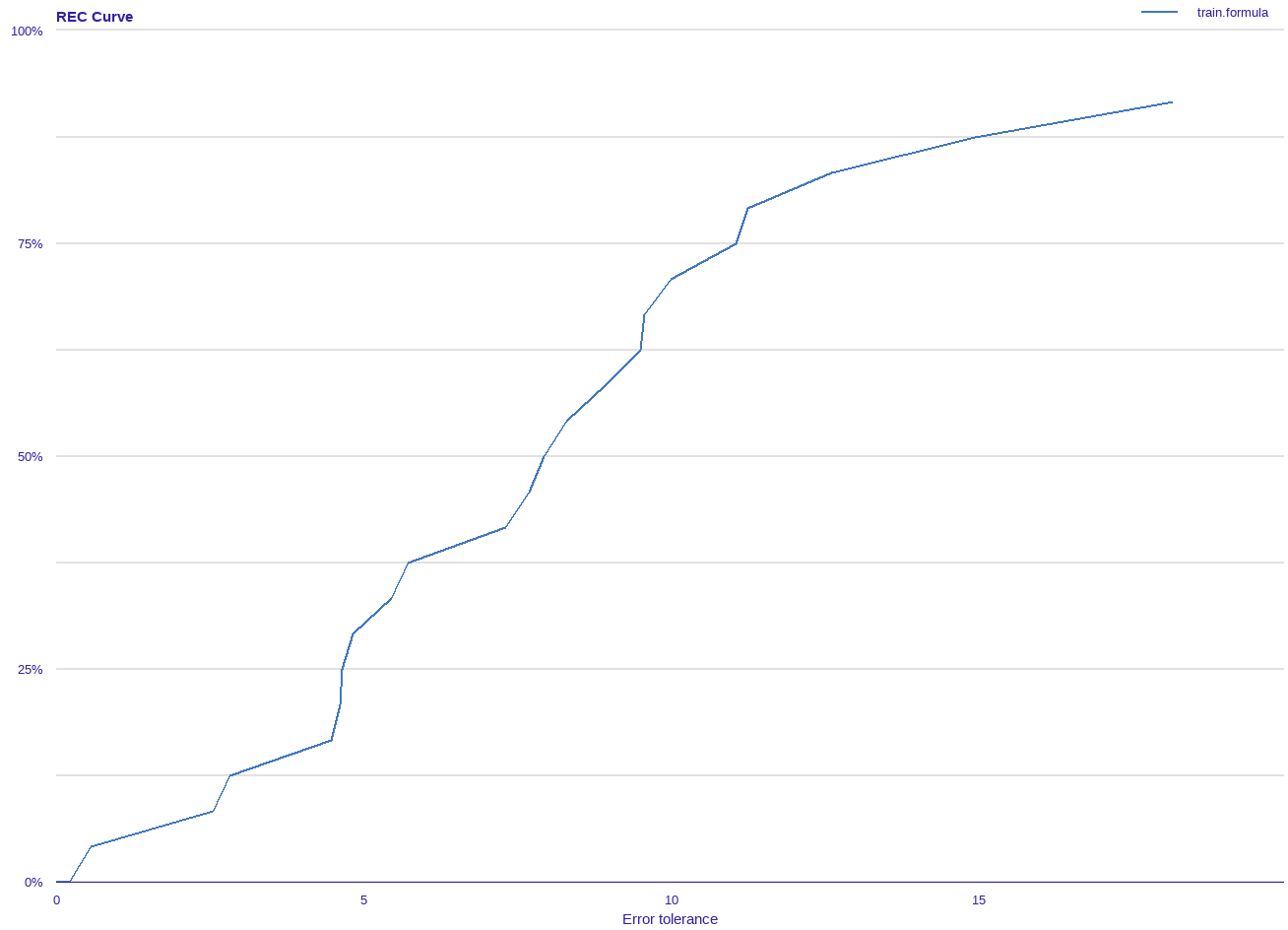
```
Preparation of a new explainer is initiated
  -> model label       :  train.formula  (  default  )
  -> data              :  23  rows  4  cols
  -> data              :  tibble converted into a data.frame
  -> target variable   :  23  values
  -> predict function  :  yhat.train  will be used (  default  )
  -> predicted values  :  No value for predict function target column. (  default  )
  -> model_info        :  package caret , ver. 6.0.94 , task regression (  default  )
  -> predicted values  :  numerical, min =  55.17894 , mean =  75.62555 , max =  96.92181
  -> residual function :  difference between y and yhat (  default  )
  -> residuals         :  numerical, min =  -14.95245 , mean =  -0.2342505 , max =  18.15963
  A new explainer has been created!
```

```
mr_m1 <- model_residual(m1_audit)
mr_m2 <- model_residual(m2_audit)

plot_rec(mr_m1)
```




```
plot_rec(mr_m2)
```




```
score_rec(m1_audit)
```

```
rec: 7.419352
```

```
score_rec(m2_audit)
```

```
rec: 6.830881
```

Comments on Evaluation Metrics and REC Values and Plots: Which Model performs better:
Model two has slightly better Numerical metrics of RMSE, MAE, MAD, and MAPE which means model 2 achieves lower prediction errors and thus has better predictive accuracy compared to model 1. However, the REC values and curves suggest that model 1's predictions achieve a higher proportion of low error predictions which means model 1 has fewer larger wrong predictions than model 2.

Overall, both models perform well with differences of accuracy and errors being marginal. Although model 2 did not provide significantly better model fit, Model 2 has better numerical metrics of predictive accuracy, a higher adjusted R2 and an interaction effect that was significant. I think Mr. Person would like to know about the information the interaction effect brings.

Deployment

Using Model 2, the model with interaction effects.

Insight for Mr. Person

Based off Analysis I can say that the variables of Income, Education, and Social connections affect Health. Two models were tested one assuming these variables were independent of each other and the second looking for any interaction effects between them. The Model selected for deployment and evidence support of the potential Social Connectedness campaign was the model that took into account interaction effects.

Question 1

How does Social Connectedness influence health?
Social Connectedness does influence health. While holding Income and Education type constant it was found that for every one unit increase in Social connectedness there was on average a 0.266 unit increase in Health. Meaning that the more a person is socially connected the healthier he will be. It is also important to note that Social connectedness and its impact on health is affected by a persons college education. If a person attended college then each one-unit increase in Social connectedness is associated with an additional 0.238-unit increase in Health compared to those without a college education.

Question 2

Does it make sense for his organization to promote Social Connectedness?
Yes, I believe it makes sense for Mr.Person's organization to promote the social connectedness event. Since this analysis found that social connectedness and health are positively related I think it is important to inform people of this.

Question 3

What would he need to consider in the marketing campaign to communicate that message?
To present this message he should first start off by reminding people how important their health is. He should be informative and show that this campaign is for both people struggling and not struggling with health. He should then show that social connectedness positively influences a persons health. He should also be aware of the anxiety that is around being social but encourage that social interaction means everyone can help everyone get healthier. For people struggling with depression and emotional health it can help them push through it. And for people that are not struggling with health issues social engagement can be a preventative measure to sickness.