Can we predict education levels based on livability?

Alec Brooks

2024-11-17

We will examine whether four specific livability factors can help predict education levels. using five different datasets on various countries worldwide: ArablePerCapita, CurrentReadiness, HealthRisks, FreshWater, and HigherEdRankings.

ArablePerCapita: Available land area per capita.

CurrentReadiness: Proportion of each government's readiness to support people and businesses.

HealthRisks: Proportion of health risk factors.

FreshWater: Available fresh water resources.

HigherEdRankings: Higher education ranking, based on US News.

1) Combine datasets.

Table 1: Cleaned and Merged Data

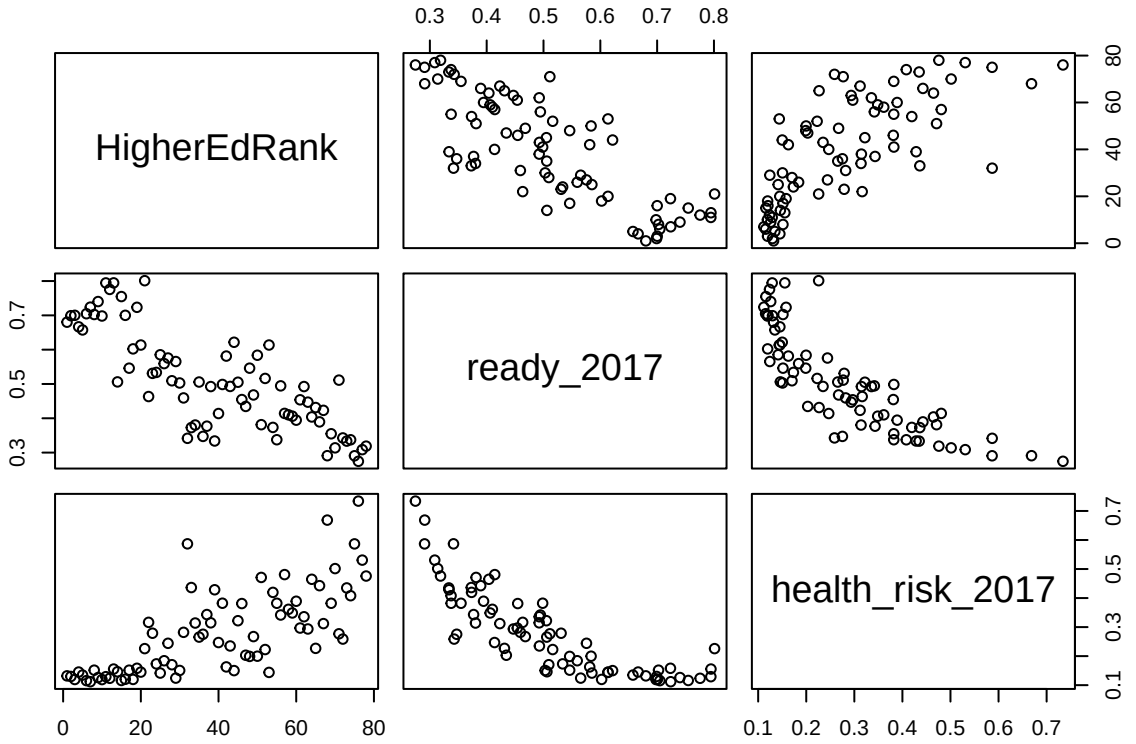
Abbreviation	HigherEdRank	Arable_2017	FWPC_2017	ready_2017	health_risk_2017
ARE	27	0.0046905	15.81077	0.5756336	0.2441548
ARG	34	0.8900027	6629.61183	0.3803120	0.3139794
AUS	8	1.2499868	19998.48792	0.7024480	0.1515935

2) Check correlations

Table 2: Correlation Table

dependent	Correlation_with_HigherEdRank
Arable_2017	-0.0776629
FWPC_2017	-0.1174615
ready_2017	-0.8025978
health_risk_2017	0.7357899

Arable_2017 and FWPC_2017 appear to have very low correlation, while Ready_2017 and Health_Risk_2017 have a much stronger relationship. We will perform a full correlation test to examine these values further.



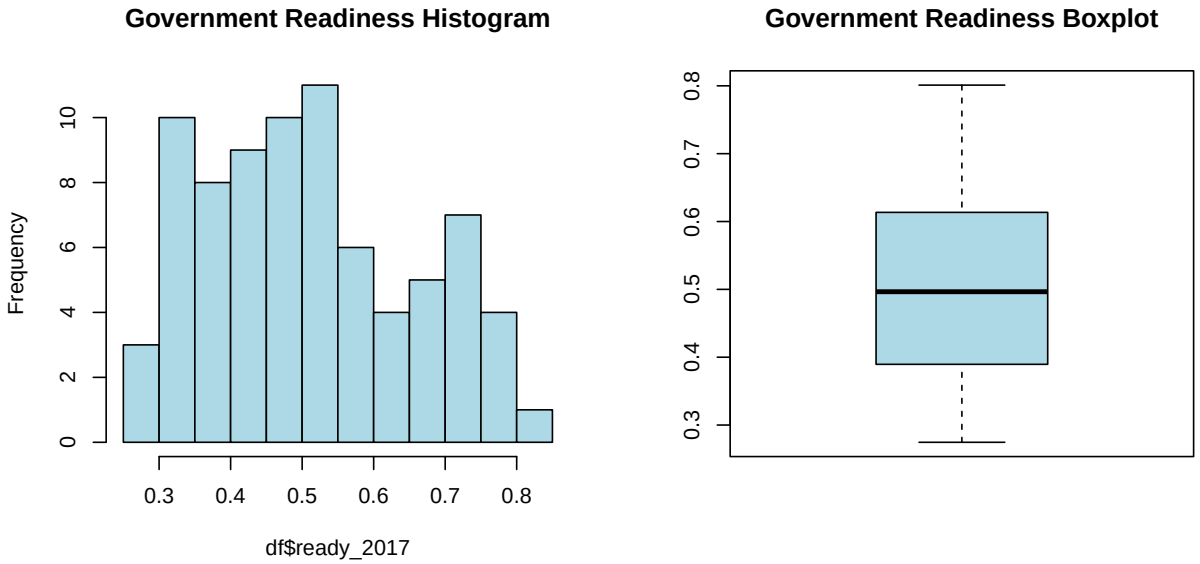

```
##
##  Pearson's product-moment correlation
##
## data:  df$HigherEdRank and df$ready_2017
## t = -11.729, df = 76, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8697830 -0.7061954
## sample estimates:
##        cor
## -0.8025978
```

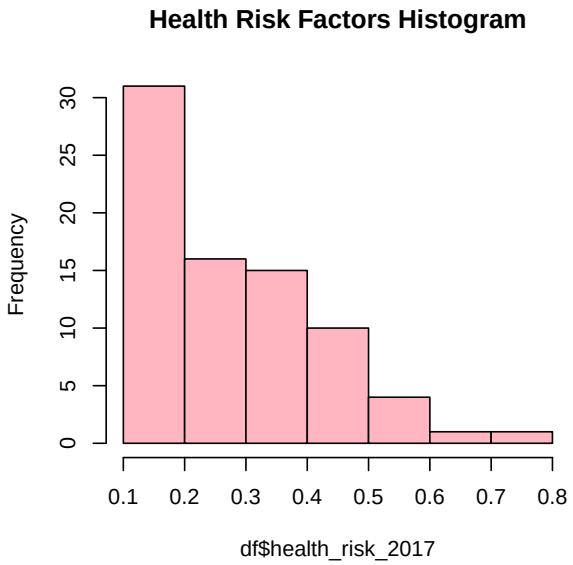
```
##
##  Pearson's product-moment correlation
##
## data:  df$HigherEdRank and df$health_risk_2017
## t = 9.4719, df = 76, p-value = 1.674e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6137525 0.8234864
## sample estimates:
##       cor
## 0.7357899
```

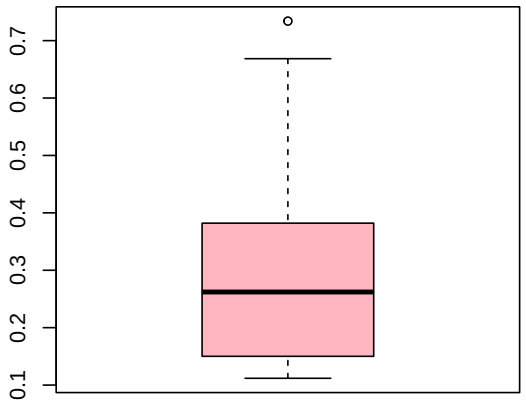
Visually, we can see that Ready_2017 appears to have a negative correlation with HigherEdRank, while Health_Risk_2017 is positively correlated. This is further supported by the findings of our Pearson's

product-moment test. The correlation coefficient of Ready_2017 is approximately -0.80, indicating a very strong negative correlation, while the correlation coefficient of Health_Risk_2017 is approximately 0.73, indicating a strong positive correlation. Additionally, the p-value of both tests is approximately zero. As we can see from our correlation tests, both variables correlate, and this correlation is statistically significant and reliable.

3) Fit a linear model







Looking at the Government Readiness histogram and box plot, we see the data is fairly evenly distributed with no clear skewness. The largest number of observations falls within the 0.4 to 0.7 range, with significant peaks around 0.5 and 0.6. The distribution is roughly uniform with no significant skewness or outliers. This suggests that the Government Readiness variable is a good candidate for linear regression since its distribution is balanced and free of irregularities.

The Risk Factors histogram shows a right-skewed distribution, meaning most nations have low health risk factors (close to 0.1 to 0.3). Looking at the box plot, the median is close to 0.2, with at least one extreme outlier in the upper range. The skewness and outlier violate the assumption of normality, making this variable less ideal for linear regression. For this reason, moving forward, we will use the Government Readiness variable as the most reasonable candidate for linear regression.

```
##
## Call:
## lm(formula = ready_2017 ~ HigherEdRank, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20524 -0.04635 -0.00164  0.05376  0.19828
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7095768  0.0197110   36.00   <2e-16 ***
## HigherEdRank -0.0050851  0.0004335  -11.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08621 on 76 degrees of freedom
## Multiple R-squared:  0.6442, Adjusted R-squared:  0.6395
## F-statistic: 137.6 on 1 and 76 DF,  p-value: < 2.2e-16
```

4) Write the regression equation

$$\hat{y} = 0.7095768 - 0.0050851x$$

Where $\hat{y}$ is the Government Readiness Score, and x is the Higher Education Rank.

---

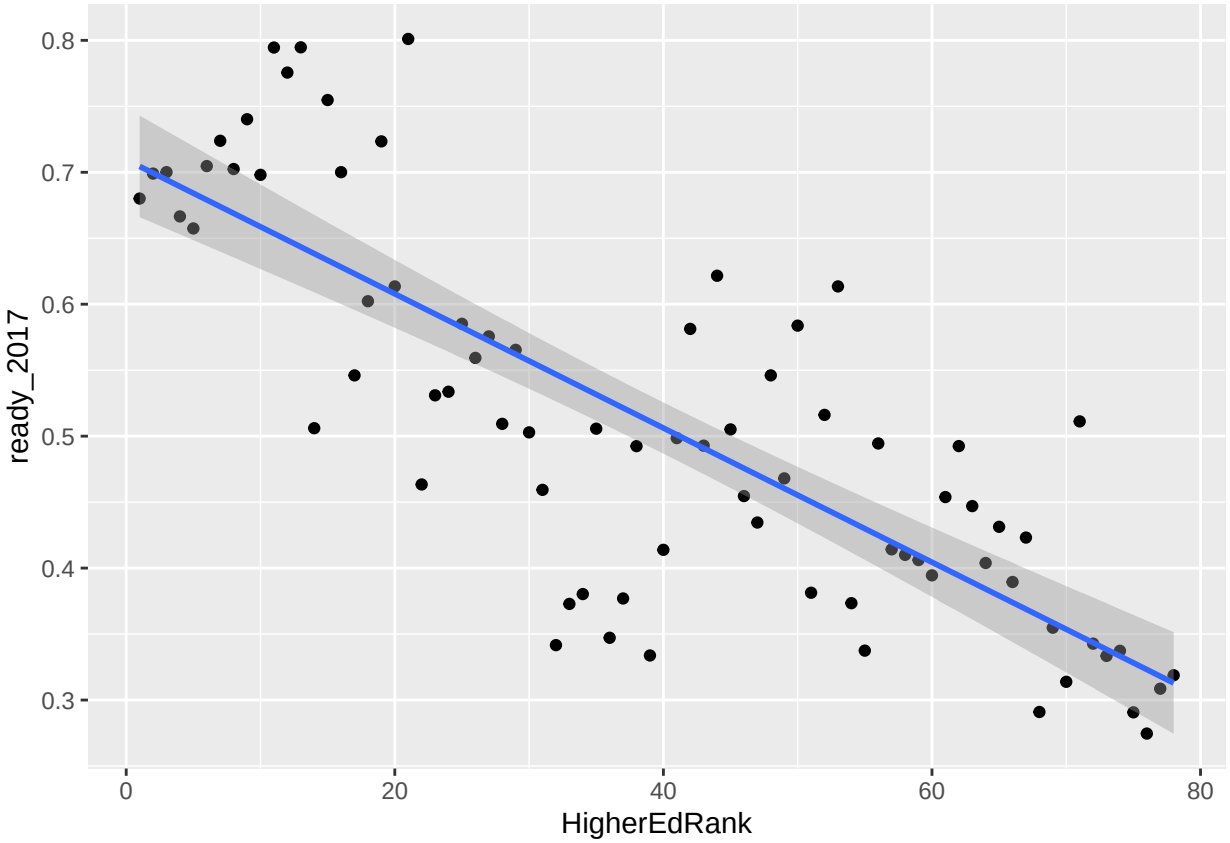
5) Interpret the slope and intercept values.

This formula tells us that if the Higher Education Rank was set to 0, we would expect a Government Readiness Score of approximately 0.7. Additionally, for each increase in the Higher Education Rank, the Government Readiness Score would decrease by approximately 0.005. It is important to note that the Higher Education Rank is assigned with the lowest value being the best. This means that, although our data is negatively correlated, it does not imply that nations with better higher educational attainment have lower Government Readiness Scores. In fact, it's the opposite. Because the best nations have the lowest ranks, this negative correlation actually reflects a positive relationship between Education and Government Readiness.

6) Evaluate the R² value and explain its meaning.

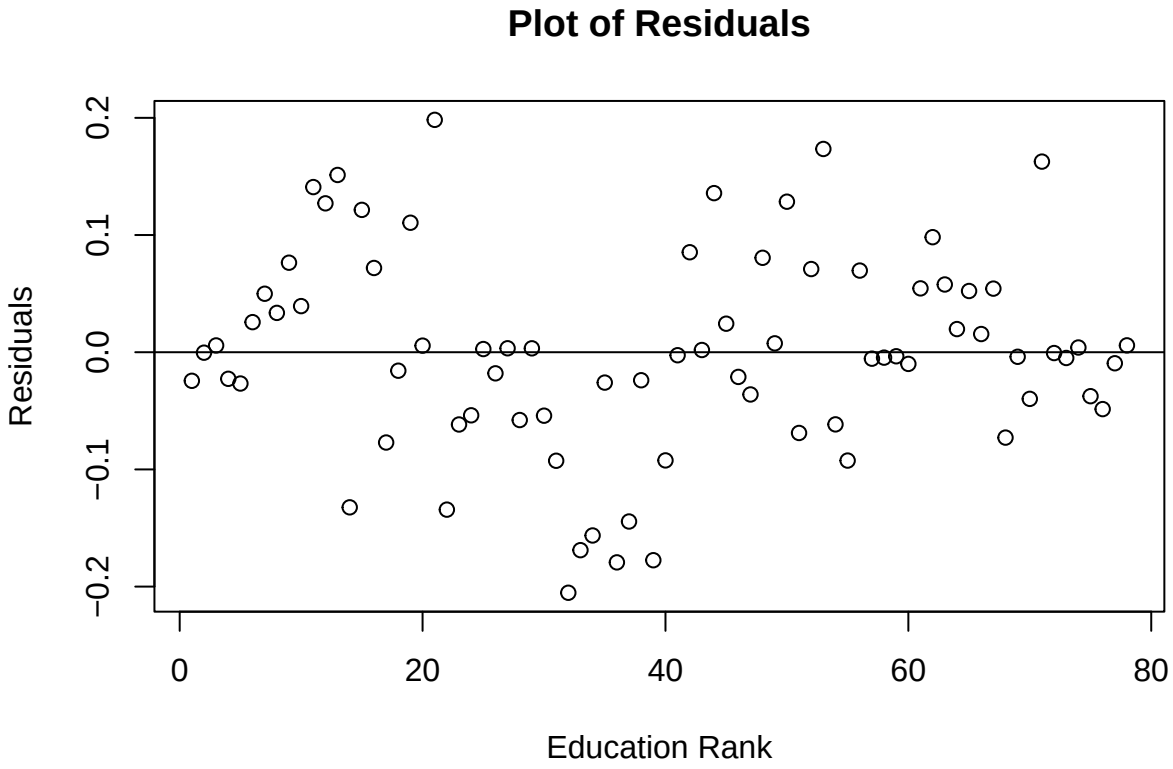
The R^2 value of 0.6442 indicates that 64% of the variation in the Government Readiness Score is explained by the Higher Education Rank. This suggests that the model has a moderate level of explanatory power, although 36% of the variation is attributable to other factors not included in the model. While this level is sufficient to continue with predictions (assuming suitability), it is important to acknowledge that the model's accuracy will be limited, and its predictions should be interpreted with caution.

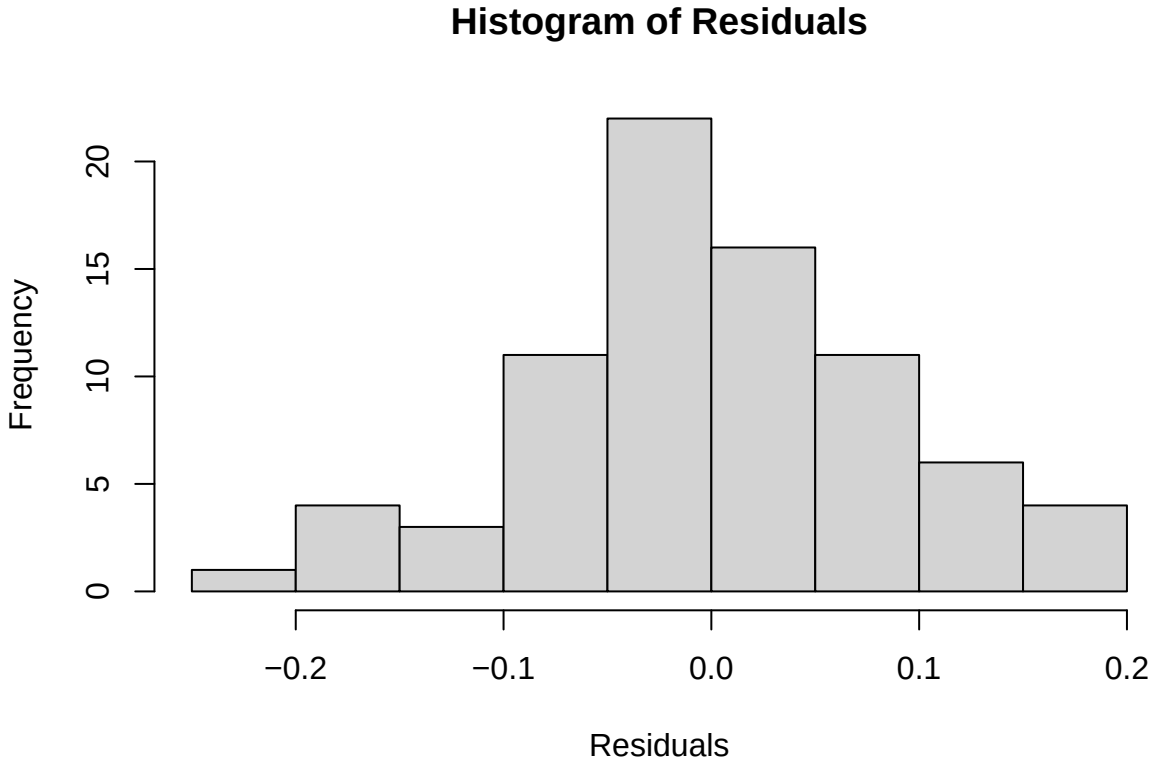
7) Plot the data



Looking at the plot of our data, we can see the previous information displayed graphically. The negative correlation is evident, with the line sloping downward toward higher education rankings. The cloud of data is fairly spread out, as observed in our histogram, and the education rankings are organized along the x-axis in ascending order. This correlation indicates a positive relationship, given that lower ranks represent better educational attainment.

8) Assess suitability





Looking at a plot of our residuals, we can see that they are evenly distributed above and below the mean, as well as relatively even across the x-axis. Additionally, the histogram of our data is slightly skewed but otherwise normal. This suggests that our model could be a reliable tool for prediction, as long as we keep in mind the somewhat low R^2 value.

9) Predict values at three chosen points.

To determine what rankings we will use for our prediction, we will consider the number of countries currently in our dataset, which is 78, compared to the 193 United Nations member nations. This discrepancy is likely due to limited information on the educational systems of some nations. However, if we imagine a scenario where this data becomes available, we can use this number to select three additional points. Let's take the remaining rankings for the unaccounted 115 nations at even intervals: 116, 154, and 192.

Table 3: Predict Values

HigherEdRank	fit	lwr	upr
116	0.1197044	0.0508489	0.1885600
154	-0.0735296	-0.1742882	0.0272290
192	-0.2667637	-0.3998676	-0.1336597

10) Interpret predictions based on the model.

When analyzing the predictions, several observations and limitations emerge regarding the relationship between educational ranking and Government Readiness.

At an educational ranking of 116, the model predicts a Government Readiness value of 0.11, with a confidence interval of [0.05, 0.18]. This suggests that nations with mid-level educational rankings tend to have notably low Government Readiness scores, which aligns with expectations for weaker preparedness in this range. However, as we examine rankings of 154 and 192, the predicted Government Readiness values become negative. While the confidence intervals for these predictions include negative values, it is uncertain whether the Government Readiness metric can logically fall below zero. If the score is non-negative (like as a percentage or index starting at zero), these examples show a limitation in our model. This analysis highlights the importance of aligning model assumptions with the actual behavior and nature of the data to ensure predictions remain realistic and meaningful.

Full Code

```
library(readr)
library(ggplot2)
ArablePerCapita <- read_csv("ArablePerCapita.csv")
current_readiness <- read_csv("current_readiness.csv")
FreshWater <- read_csv("FreshWater.csv")
HealthRisks <- read_csv("HealthRisks.csv")
HigherEdRankings <- read_csv("HigherEdRankings.csv")
#1
df <- merge(HigherEdRankings, ArablePerCapita[-c(1,3,4)],
            by.x = "Abbreviation", by.y = "Country Code")
df <- merge(df, FreshWater[-c(1,3,4)], by.x = "Abbreviation", by.y = "Country Code")
df <- merge(df, current_readiness[-c(2,3,4)], by.x = "Abbreviation", by.y = "ISO3")
df <- merge(df, HealthRisks[-c(2,3,4)], by.x = "Abbreviation", by.y = "ISO3")
knitr::kable(head(df[-2], n = 3), caption = "Cleaned and Merged Data")
#2
cor_table <- data.frame(
  dependent = c("Arable_2017", "FWPC_2017", "ready_2017", "health_risk_2017"),
  Correlation_with_HigherEdRank =
    c(cor(df$HigherEdRank, df$Arable_2017), cor(df$HigherEdRank, df$FWPC_2017),
      cor(df$HigherEdRank, df$ready_2017), cor(df$HigherEdRank, df$health_risk_2017)))
knitr::kable(cor_table, caption = "Correlation Table")
plot(df[-c(1,2,4,5)])
cor.test(df$HigherEdRank,df$ready_2017)
cor.test(df$HigherEdRank,df$health_risk_2017)
#3
par(mfrow = c(2, 2))
hist(df$ready_2017, main = "Government Readiness Histogram", xlab = "", col = "#ADD8E6")
boxplot(df$ready_2017, main = "Government Readiness Boxplot", col = "#ADD8E6")
hist(df$health_risk_2017, main = "Health Risk Factors Histogram", xlab = "",
     col = "#FFB6C1")
boxplot(df$health_risk_2017, main = "Health Risk Factors Boxplot", col = "#FFB6C1")

lmMod <- lm(ready_2017~HigherEdRank,data=df)
summary(lmMod)
#7
ggplot(df, aes(HigherEdRank,ready_2017)) + geom_point() + stat_smooth(method=lm)
#8
plot(df$HigherEdRank, lmMod$residuals, main = "Plot of Residuals", ylab = "Residuals",
     xlab = "Education Rank")
abline(0,0)
hist(lmMod$residuals, main = "Histogram of Residuals", xlab = "Residuals")
#9
pred <- data.frame(HigherEdRank = c(116, 154, 192))
predicted_values <- predict(lmMod, pred, interval = "confidence")
pred <- cbind(pred, predicted_values)
knitr::kable(pred, caption = "Predict Values")
```