398E_Final_Project

Eric Li, Roshan Shet, Kevin Liang

2022-12-02

Introduction: In this project, we analyzed data from the WHO in order to investigate the relationship between a country's average BMI and Adult Mortality Rate. We decided to use linear regression to fit a model on this relationship.

The dataset can be found here: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Load required packages.

```
library(tidyverse)
library(RMySQL)
library(pacman)
p_load(rpart,tidymodels)
```

Read the data in.

```
mysqlconnection = dbConnect(RMySQL::MySQL(),
                            dbname='Class_Data',
                            host='localhost',
                            port=3306,
                            user='root',
                            password='cmsc398e')

query = "WITH table1 AS (
            WITH table2 AS (
                SELECT * FROM Class_Data.`life expectancy data` WHERE `BMI` IS NOT NULL AND
                `Adult Mortality` IS NOT NULL
            )
            SELECT *, UPPER(Country) AS `Country Name`, AVG(`BMI`) OVER (PARTITION BY `Country`)
            `Average BMI`, AVG(`Adult Mortality`) OVER (PARTITION BY `Country`)
            `Average Adult Mortality` FROM table2 GROUP BY `Country`
        )
        SELECT *, DENSE_RANK() OVER (ORDER BY `Average Adult Mortality` DESC) AS
        `Adult Mortality Rank` FROM table1"

result = dbSendQuery(mysqlconnection, query)
# Stores resulting table as dataframe
df = fetch(result)
```

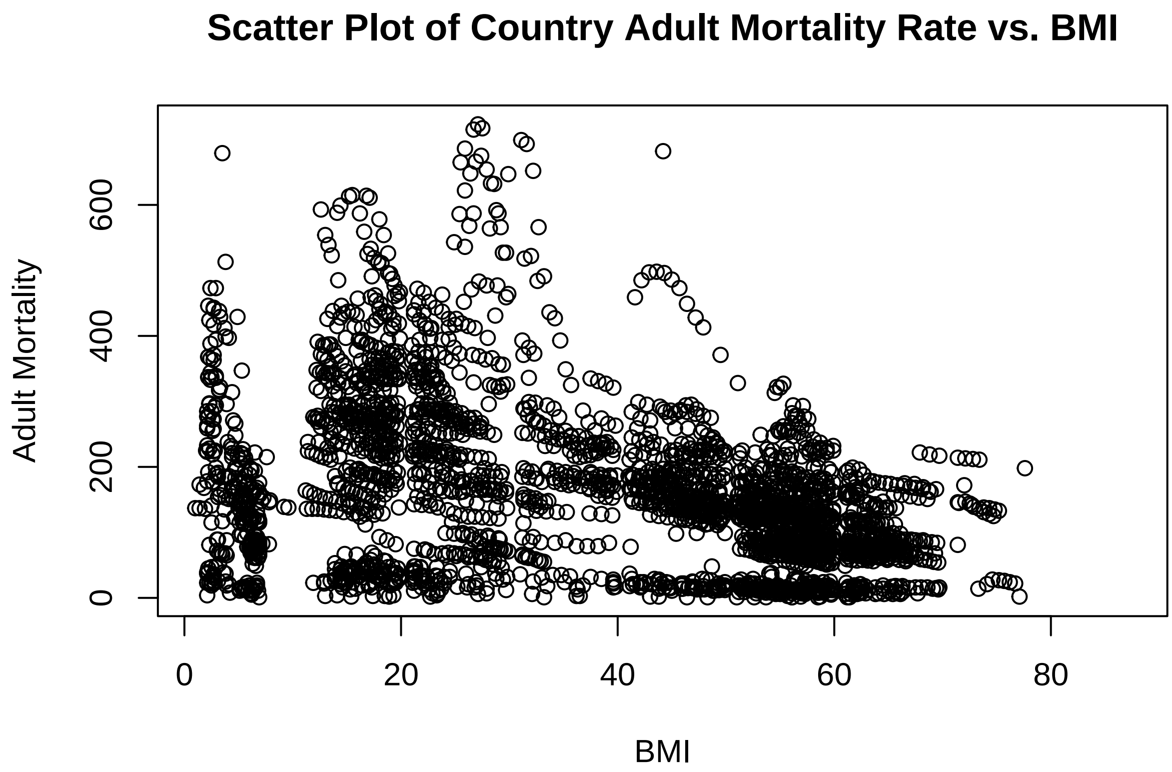
```
head(df)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
```

```
## 2 Afghanistan 2014 Developing            59.9             271             64
## 3 Afghanistan 2013 Developing            59.9             268             66
## 4 Afghanistan 2012 Developing            59.5             272             69
## 5 Afghanistan 2011 Developing            59.2             275             71
## 6 Afghanistan 2010 Developing            58.8             279             74
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                 83
## 2    0.01              73.523582          62     492 18.6                 86
## 3    0.01              73.219243          64     430 18.1                 89
## 4    0.01              78.184215          67    2787 17.6                 93
## 5    0.01               7.097109          68    3013 17.2                 97
## 6    0.01              79.679367          66    1989 16.7                102
##   Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 1     6              8.16         65      0.1 584.25921   33736494
## 2    58              8.18         62      0.1 612.69651     327582
## 3    62              8.13         64      0.1 631.74498   31731688
## 4    67              8.52         67      0.1 669.95900    3696958
## 5    68              7.87         68      0.1  63.53723    2978599
## 6    66              9.20         66      0.1 553.32894    2883167
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                 17.2               17.3                           0.479
## 2                 17.5               17.5                           0.476
## 3                 17.7               17.7                           0.470
## 4                 17.9               18.0                           0.463
## 5                 18.2               18.2                           0.454
## 6                 18.4               18.4                           0.448
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
## 6       9.2
```

Let's plot Adult Mortality Rate vs. BMI:

```
plot(df$BMI, df$`Adult.Mortality`, xlab = "BMI", ylab = "Adult Mortality", main = "Scatter Plot of Coun
```



The data appears to be somewhat negatively correlated. Let's try fitting a linear regression model to it.

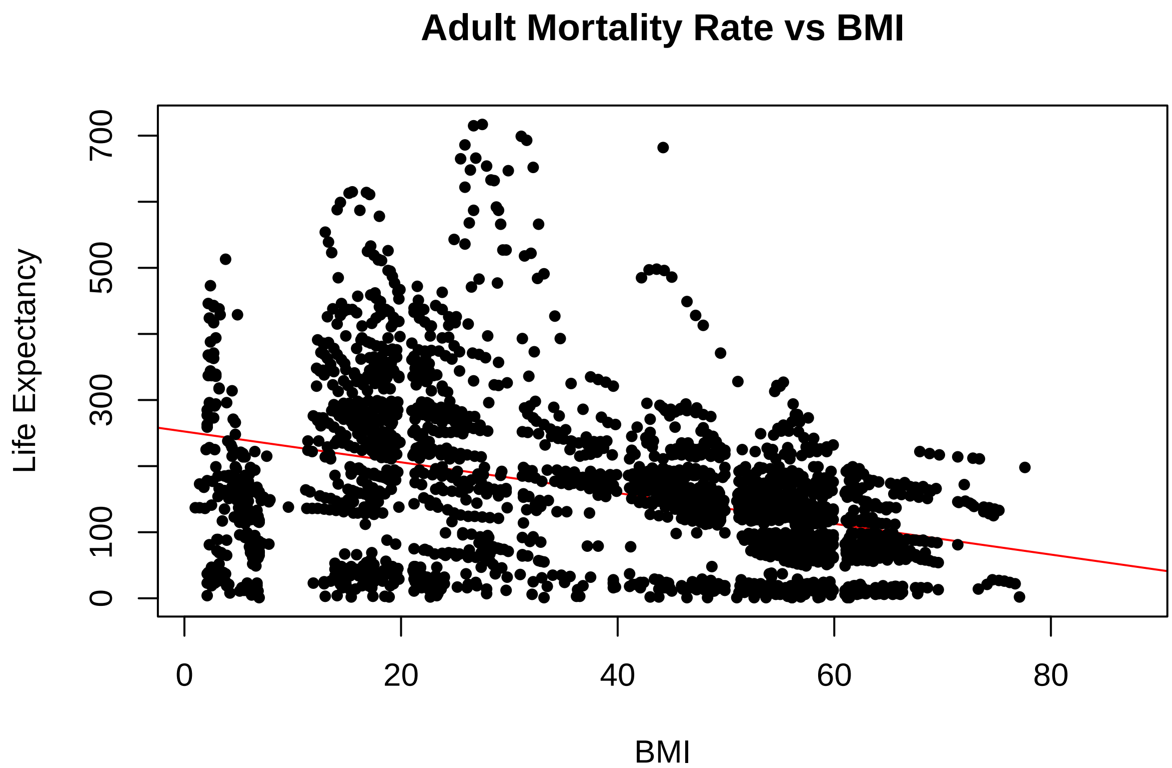
Split the data into a training and testing set.

```
data_split <-
  df %>%  rsample::initial_split(
    data = ,
    prop = 0.8)
train_data <- training(data_split)
test_data <- testing(data_split)
```

We will fit a linear regression model onto the data in order to compare BMI and Adult Mortality Rate.

Null Hypothesis: There is no relationship between BMI and Adult Mortality Rate (i.e. beta1 = 0).

```
x <- train_data$BMI
y <- train_data$Adult.Mortality
model <- lm(Adult.Mortality ~ BMI, data = train_data)
plot(x,y,main = "Adult Mortality Rate vs BMI",abline(model,col="red"),cex = 0.8,pch = 16,xlab = "BMI",y
```




```r
summary(model)
```

```
##
## Call:
## lm(formula = Adult.Mortality ~ BMI, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -243.45  -71.14   -4.01   58.59  532.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 252.3411     5.1687   48.82   <2e-16 ***
## BMI          -2.3274     0.1192  -19.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114.8 on 2315 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.1415, Adjusted R-squared:  0.1411
## F-statistic: 381.5 on 1 and 2315 DF,  p-value: < 2.2e-16
```

Linear Equation: Adult Mortality Rate = -2.39*BMI + 255.11

Since the p-value for the coefficient of x is <2e-16, we can reject it at the 5% significance level since 2e-16 < 0.05. This means that we can reject the null hypothesis. In this situation, the null hypothesis is H0 = 0, which suggests there is no correlation between the predicted value and the observed value (test vs predicted). Since we reject the null hypothesis, we can conclude that a correlation does exist between the 2 values.

Test the model on the testing set.

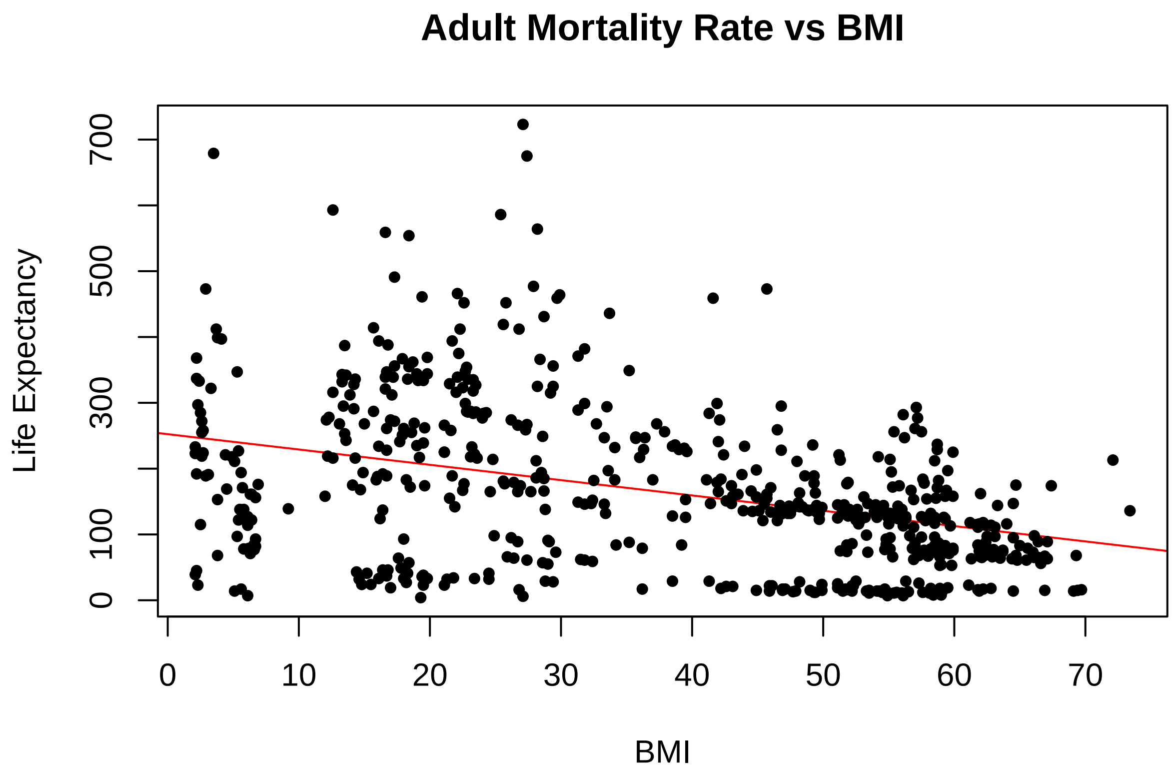
```
predictions = predict(model, newdata = test_data)
summary(lm(test_data$`Adult.Mortality`~predictions))
```

```
##
## Call:
## lm(formula = test_data$Adult.Mortality ~ predictions)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -243.24  -80.71   -2.59   62.76  529.71
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.2683    17.4645  -1.561    0.119
## predictions   1.1653     0.1016  11.468   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112 on 577 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.1856, Adjusted R-squared:  0.1842
## F-statistic: 131.5 on 1 and 577 DF,  p-value: < 2.2e-16
```

The p-value for the coefficient of x is low again (<2e-16), but the p-value for the intercept is quite high, so it may be wise to drop it. However, we can still reject our null hypothesis.

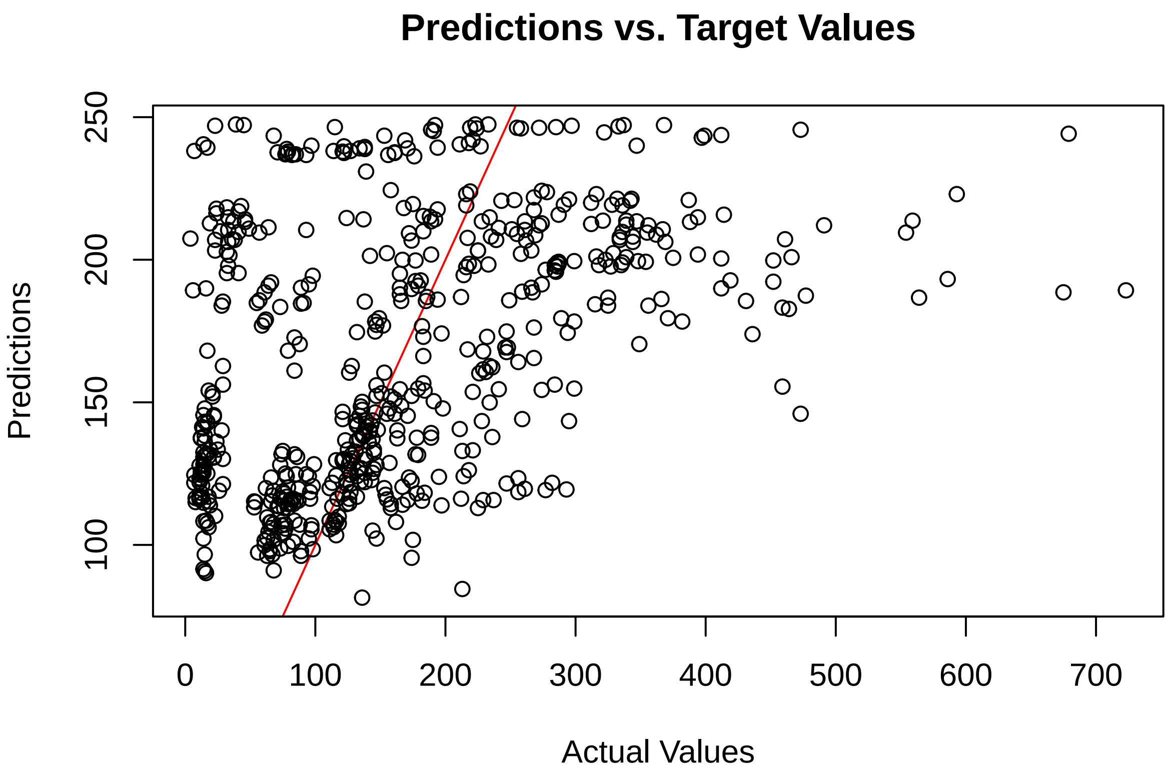
Here is how our linear model looks with the testing data:

```
x <- test_data$BMI
y <- test_data$Adult.Mortality
plot(x,y,main = "Adult Mortality Rate vs BMI",abline(model,col="red"),cex = 0.8,pch = 16,xlab = "BMI",y
```



Let's see how good our predictions are by plotting them vs. the actual values:

```
pred_test = data.frame(target = test_data$Adult.Mortality, pred = predictions)
plot(pred_test$target, pred_test$pred, xlab="Actual Values", ylab="Predictions",
     main="Predictions vs. Target Values", abline(a=0, b=1, col="red"))
```



In general, we want the points to follow the line y=x, as we want out predictions to be close to the actual values. In this case, there appears to be about an equal number of predictions above and below the line.

Let's also calculate the RMSE for each prediction:

```
pred_test_rmse = pred_test %>% mutate(dif = (target - pred)^2, rmse = (dif - nrow(pred_test))^0.5)
head(pred_test_rmse)
```

```
##    target     pred       dif      rmse
## 8     287 215.8016  5069.209  66.94183
## 11    291 219.2927  5141.943  67.48291
## 13    295 221.1545  5453.152  69.75064
## 15    316 223.0164  8645.945  89.76606
## 27     15 136.2061 14690.908 118.75567
## 30     15 143.1881 16432.195 125.87373
```

Calculate the average error.

```
100*mean(!sapply(pred_test_rmse$rmse, is.na))/(max(!sapply(pred_test_rmse$target, is.na))-min(!sapply(p
```

```
## [1] 75.17007
```

The average error is not too large, indicating that the linear regression model may be a good predictor for Adult Mortality Rate using BMI as input when presented with unseen data.