HU ML Problem Set 1

Aaron

September 20, 2016

```
library(tidyverse)
library(magrittr)
library(car)
library(knitr)
```

Homework 1

Complete the following problems from the Kelleher book (Chapter 3) 5, 8, 10

Problem 5

The table below shows the scores achieved by a group of students on an exam.

```
data_frame(id = 1:20, score = c(42, 47, 59, 27, 84, 49, 72, 43, 73, 59,
                                58, 82, 50, 79, 89, 75, 70, 59, 67, 35)) -> df

head(df)
```

```
## # A tibble: 6 × 2
##      id score
##   <int> <dbl>
## 1     1    42
## 2     2    47
## 3     3    59
## 4     4    27
## 5     5    84
## 6     6    49
```

Using this data, perform the following tasks on the SCORE feature:

1. A **range normalization** that generates data in the range (0, 1)

```
range_normalization = function(x, high, low) {

    ((x - min(x)) / (max(x) - min(x))) * (high - low) + low

}

# ranged normalized score to (0,1)

df %$% range_normalization(score, 1, 0)
```

```
##  [1] 0.2419355 0.3225806 0.5161290 0.0000000 0.9193548 0.3548387 0.7258065
##  [8] 0.2580645 0.7419355 0.5161290 0.5000000 0.8870968 0.3709677 0.8387097
## [15] 1.0000000 0.7741935 0.6935484 0.5161290 0.6451613 0.1290323
```

2. A **range normalization** that generates data in the range (− 1, 1)

```r
# ranged normalized score to (0,1)

df %$% range_normalization(score, 1, -1)
```

```
##  [1] -0.51612903 -0.35483871  0.03225806 -1.00000000  0.83870968
##  [6] -0.29032258  0.45161290 -0.48387097  0.48387097  0.03225806
## [11]  0.00000000  0.77419355 -0.25806452  0.67741935  1.00000000
## [16]  0.54838710  0.38709677  0.03225806  0.29032258 -0.74193548
```

3. A **standardization** of the data

```r
# using built-in function like a classy cat

df %>%
    mutate(std_score = round(scale(score), 2)) %>%
    select(std_score) %>% as_vector()
```

```
##  std_score1  std_score2  std_score3  std_score4  std_score5  std_score6
##       -1.10       -0.81       -0.11       -1.97        1.34       -0.69
##  std_score7  std_score8  std_score9 std_score10 std_score11 std_score12
##        0.64       -1.04        0.70       -0.11       -0.17        1.22
## std_score13 std_score14 std_score15 std_score16 std_score17 std_score18
##       -0.63        1.05        1.63        0.81        0.52       -0.11
## std_score19 std_score20
##        0.35       -1.50
```

Problem 8

The table below shows socio-economic data for a selection of countries for the year 2009, using the following features:

- COUNTRY: The name of the country
- LIFEEXPECTANCY: The average life expectancy (in years)
- INFANTMORTALITY: The infant mortality rate (per 1,000 live births)
- EDUCATION: Spending per primary student as a percentage of GDP
- HEALTH: Health spending as a percentage of GDP
- HEALTHUSD: Health spending per person converted into US dollars

```r
country = read_csv("../data/country.csv")

str(country)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    15 obs. of  6 variables:
##  $ country         : chr  "Argentina" "Can1eroon" "Chile" "Colombia" ...
##  $ life_expectancy : num  75.6 53.3 78.9 73.2 78.6 ...
##  $ infant_mortality: num  13.5 67.7 7.8 16.5 4.8 52.5 31.2 8.5 7.1 85.5 ...
##  $ education       : num  16.84 7.14 17.36 15.59 44.17 ...
##  $ health          : num  9.53 4.92 8.4 7.6 12.1 ...
##  $ health_usd      : num  734.1 60.4 801.9 391.9 672.2 ...
```

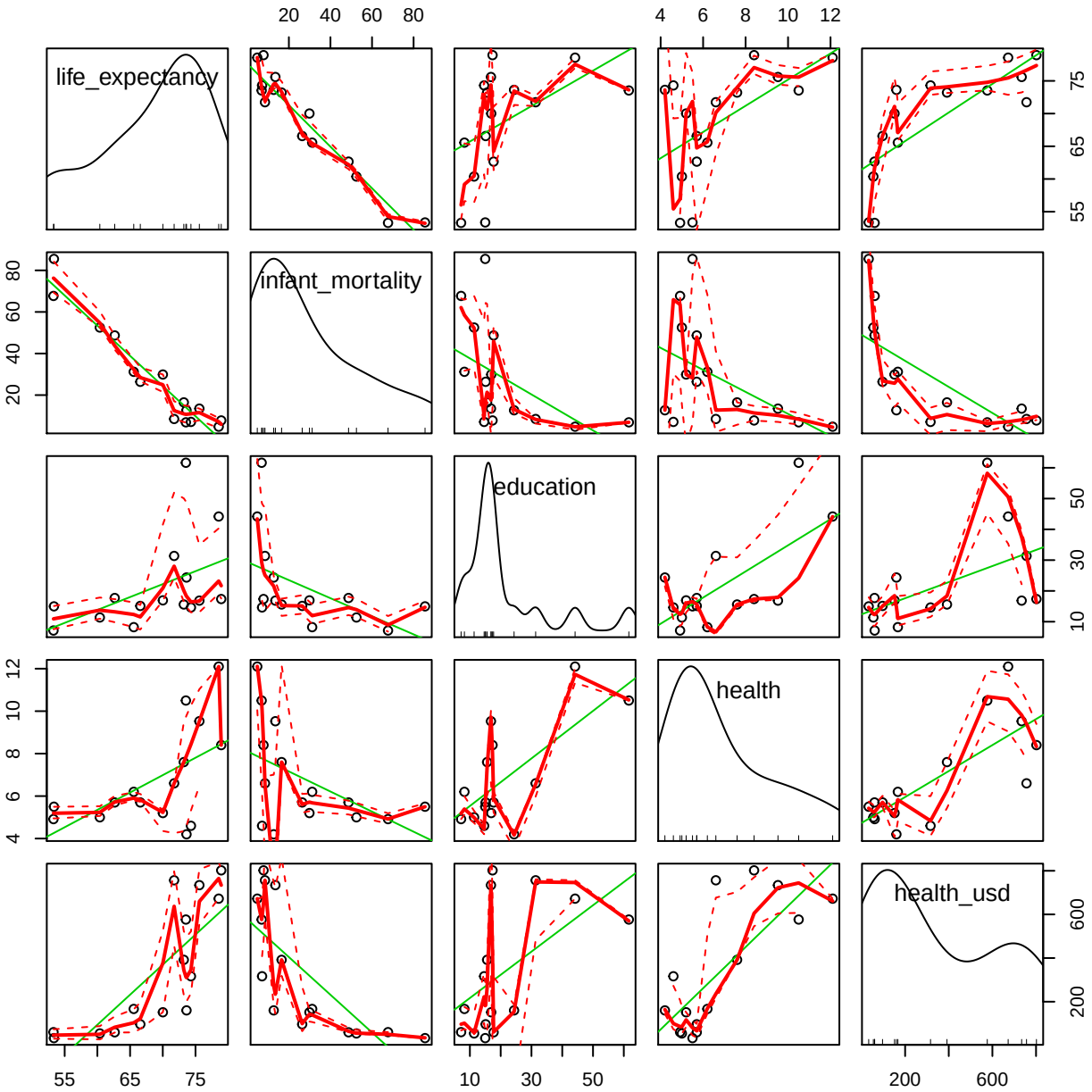
1. Calculate the correlation between the LIFEEXPECTANCY and INFANT-MORTALITY features.

```r
country %$%
    cor(life_expectancy, infant_mortality)
```

```
## [1] -0.960733
```

2. The image below shows a scatter plot matrix of the continuous features from this dataset (the correlation between LIFEEXPECTANCY and INFANTMORTALITY has been omitted). Discuss the relationships between the features in the dataset that this scatter plot highlights.

```
country[,2:6] %>%
    scatterplotMatrix()
```



- Life Expectancy and Infant Mortality are negatively correlated, the relationship is almost linear.
- After Life Expectancy exceeds 65, it is positively correlated with Health (and Health USD)
- Infant Mortality is negatively correlated with Health (and Health USD) before 40.

Problem 10

The following data visualizations are based on the tachycardia prediction dataset from Question 9 (after the instances with missing TACHYCARDIA values have been removed and all outliers have been handled). Each visualization illustrates the relationship between a descriptive feature and the target feature, TACHY-

CARDIA and is composed of three plots: a plot of the distribution of the descriptive feature values in the full dataset, and plots showing the distribution of the descriptive feature values for each level of the target. Discuss the relationships shown in each visualizations.

1. The visualization below illustrates the relationship between the continuous feature DIA. B.P. and the target feature, TACHYCARDIA.

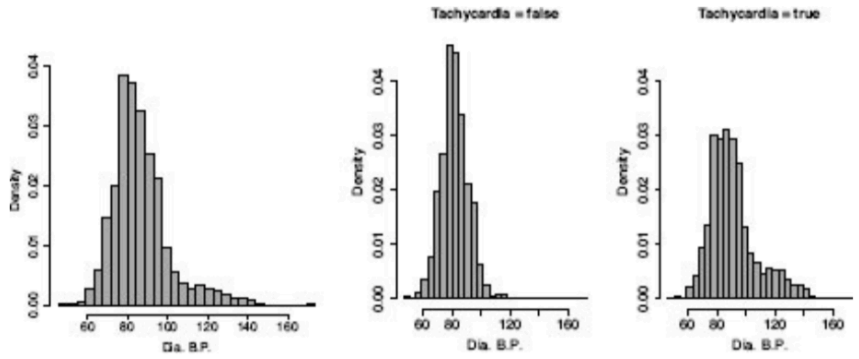


Figure 1:

The false target variable is more correlated with DIA B.P at range of 60 to 90.

The true target variable is more correlated with DIA B.P at range of 100 to 160.

2. The visualization below illustrates the relationship between the continuous HEIGHT feature and the target feature TACHYCARDIA.

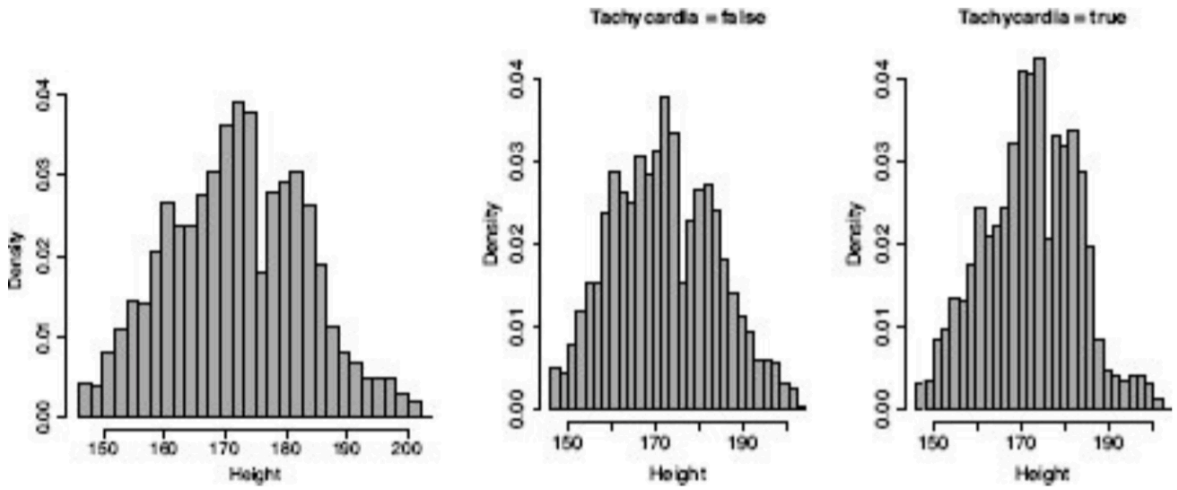


Figure 2:

Target variable is equally distributed in the full range of height, expect that the false target variable is more represented in height range above 190.

3. The visualization below illustrates the relationship between the categorical feature PREV. TACHY. and the target feature, TACHYCARDIA.

The target variable is strongly correlated with true Prev Tachy variable.

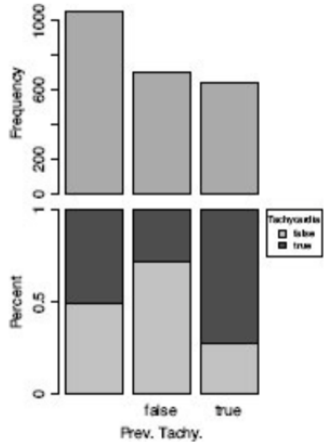

Frequency
1000
600
200
0
1
0.5
0
Percent
Tachycardia
false
true
false
true
Prev. Tachy.

Figure 3: