Presented To

Dr Ammar Aftab Raja

Presented By

Khalil ul Rehman Mirza (Leader)

F2020313010

Shahzada Farhan Mehmood

F2020313004

Farhan Ahmad

F20203130023

Links

https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/linear-regression/assignment-2/reading-test-scores/pisa2009train.csv

http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011038

Problem Statement

Load the given data file

Clean the data

Apply suitable operations on it

Get the presentable data from the data file

Data Summary

- Data Summary

```
     grade            male          raceeth            preschool      expectBachelors
Min.   : 8.00   Min.   :0.0000   Length:3663        Min.   :0.0000   Min.   :0.0000
1st Qu.:10.00   1st Qu.:0.0000   Class :character   1st Qu.:0.0000   1st Qu.:1.0000
Median :10.00   Median :1.0000   Mode  :character   Median :1.0000   Median :1.0000
Mean   :10.09   Mean   :0.5111                      Mean   :0.7228   Mean   :0.7859
3rd Qu.:10.00   3rd Qu.:1.0000                      3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :12.00   Max.   :1.0000                      Max.   :1.0000   Max.   :1.0000
                                                    NA's   :56       NA's   :62
   motherHS    motherBachelors    motherWork        fatherHS      fatherBachelors
Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:1.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
Median :1.00   Median :0.0000   Median :1.0000   Median :1.0000   Median :0.0000
Mean   :0.88   Mean   :0.3481   Mean   :0.7345   Mean   :0.8593   Mean   :0.3319
3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :1.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
  fatherWork       selfBornUS      motherBornUS     fatherBornUS    englishAtHome
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000
Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000
Mean   :0.8531   Mean   :0.9313   Mean   :0.7725   Mean   :0.7668   Mean   :0.8717
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
NA's   :233      NA's   :69       NA's   :71       NA's   :113      NA's   :71
computerForSchoolwork read30MinsADay   minutesPerWeekEnglish studentsInEnglish schoolHasLibrary
Min.   :0.0000        Min.   :0.0000   Min.   :   0.0        Min.   : 1.0      Min.   :0.0000
1st Qu.:1.0000        1st Qu.:0.0000   1st Qu.: 225.0        1st Qu.:20.0      1st Qu.:1.0000
Median :1.0000        Median :0.0000   Median : 250.0        Median :25.0      Median :1.0000
Mean   :0.8994        Mean   :0.2899   Mean   : 266.2        Mean   :24.5      Mean   :0.9676
3rd Qu.:1.0000        3rd Qu.:1.0000   3rd Qu.: 300.0        3rd Qu.:30.0      3rd Qu.:1.0000
Max.   :1.0000        Max.   :1.0000   Max.   :2400.0        Max.   :75.0      Max.   :1.0000
NA's   :65            NA's   :34       NA's   :186           NA's   :249       NA's   :143
 publicSchool       urban          schoolSize    readingScore
Min.   :0.0000   Min.   :0.0000   Min.   : 100   Min.   :168.6
1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.: 712   1st Qu.:431.7
Median :1.0000   Median :0.0000   Median :1212   Median :499.7
Mean   :0.9339   Mean   :0.3849   Mean   :1369   Mean   :497.9
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1900   3rd Qu.:566.2
Max.   :1.0000   Max.   :1.0000   Max.   :6694   Max.   :746.0
                                  NA's   :162
```

Colum names

- **Colnames**

```
> colnames(df)
 [1] "grade"                "male"                 "raceeth"
 [4] "preschool"            "expectBachelors"      "motherHS"
 [7] "motherBachelors"      "motherWork"           "fatherHS"
[10] "fatherBachelors"      "fatherWork"           "selfBornUS"
[13] "motherBornUS"         "fatherBornUS"         "englishAtHome"
[16] "computerForSchoolwork" "read30MinsADay"       "minutesPerWeekEnglish"
[19] "studentsInEnglish"    "schoolHasLibrary"     "publicSchool"
[22] "urban"                "schoolSize"           "readingScore"
```

Count the NA values

```
na_values = sum(is.na(df))
na_values
```

```
> na_values
[1] 2950
```

```
# here is head of the data frame and sum of rows
na_values = sum(is.na(df))
na_values
#2nd way
table(is.na(df))
```

```
> na_values
[1] 2950
> #2nd way
> table(is.na(df))

FALSE   TRUE
84962   2950
```

Getting rid of NA Values

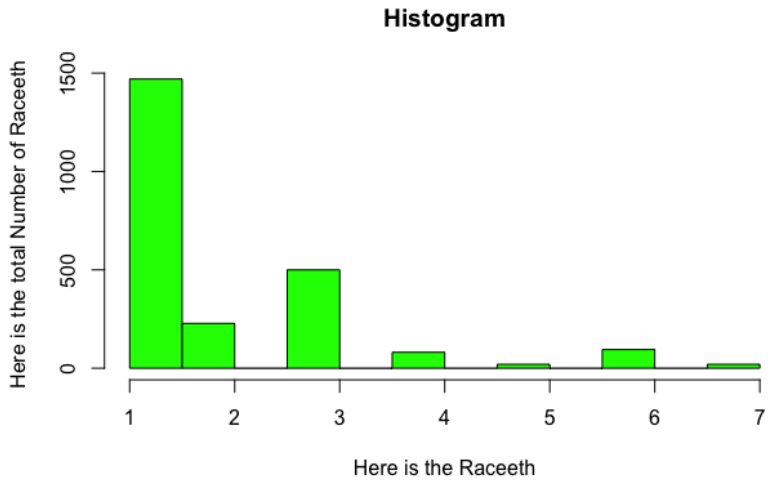
```
# getting rid of the empty value by using omit
df1=na.omit(df)
sum(is.na(df1))
```

```
> df1=na.omit(df)
> sum(is.na(df1))
[1] 0
```

Histogram

- Generation Histogram of raceeth
- BY converting string into integers

```
#here i am getting the raceeth data and setting it onto numbers from 1,7 and then ploat a histogra
sample_data = c(df1$raceeth)
sample_data
raceethdf = as.numeric(gsub("White", 1, gsub("Black", 2, gsub("Hispanic", 3,gsub("More than one ra
is.numeric(raceethdf)
hist(raceethdf, main = 'Histogram', xlab = 'Here is the Raceeth', ylab = 'Here is the total Number
```



Conclussion

- Total Number of Rows
- NA number of Rows
- Total male Students
- Total Number of female Students
- Total Number of Male Students

- Checking the reding30mint student Frequency
- Draw the raceeth column a histogram
- Table function for the total number of female and male students