10-import

Describe the data and the problem

I am interested in predicting the housing Values in Suburbs of Boston.

We have the some interesting features of the house, for instance,

black 1000(Bk - 0.63)ˆ2 where Bk is the proportion of blacks by town.

ptratio pupil-teacher ratio by town.

How would these enviornment variables affect a house's value?

I found this data set from our class repo, more information could be found at https://www.kaggle.com/c/boston-housing/overview/description

loading pkgs

```
library(tidyverse)
library(ggplot2)
library(fs)
library(purrr)
library(stringr)
library(assertr)
```

load data

```
df <- read_csv("./data/BostonHousing.csv")
```

some functions

```
#for check how many unique values for each var.
col_uni <- function(df){
  map(df,unique)
}
```

##overview

```
dim(df)
```

```
## [1] 506  14
```

```
str(df)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : num  1 2 2 3 3 3 5 5 5 5 ...
```

```
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   crim = col_double(),
##   ..   zn = col_double(),
##   ..   indus = col_double(),
##   ..   chas = col_double(),
##   ..   nox = col_double(),
##   ..   rm = col_double(),
##   ..   age = col_double(),
##   ..   dis = col_double(),
##   ..   rad = col_double(),
##   ..   tax = col_double(),
##   ..   ptratio = col_double(),
##   ..   b = col_double(),
##   ..   lstat = col_double(),
##   ..   medv = col_double()
##   .. )
```

```
summary(df)
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio            b        
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

check the dataset

```
# check how many missing values in our dataset
sum(is.na(df))
```

```
## [1] 0
```

```
# check the limit (based on description on the kaggle) for each variable
df %>%
  assert(in_set(c(1, 0)), chas)%>%
  assert(in_set(c(0:24)), rad)%>%
  assert(within_bounds(0,1), nox) %>%
  assert(within_bounds(0,100), zn)%>%
  assert(within_bounds(0,100), crim)%>%
  assert(within_bounds(0,Inf),indus)%>%
  assert(within_bounds(0,120), age) %>%
  assert(within_bounds(0,Inf), dis) %>%
  assert(within_bounds(0,100), ptratio) %>%
  assert(within_bounds(0,Inf), tax) %>%
  assert(within_bounds(0,Inf), medv) %>%
  assert(within_bounds(0,Inf), b) %>%
  assert(within_bounds(0,100), lstat)
```

```
## # A tibble: 506 x 14
##       crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio
##      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
##  1 0.00632  18    2.31     0 0.538  6.58  65.2  4.09     1   296    15.3
##  2 0.0273    0    7.07     0 0.469  6.42  78.9  4.97     2   242    17.8
##  3 0.0273    0    7.07     0 0.469  7.18  61.1  4.97     2   242    17.8
##  4 0.0324    0    2.18     0 0.458  7.00  45.8  6.06     3   222    18.7
##  5 0.0690    0    2.18     0 0.458  7.15  54.2  6.06     3   222    18.7
##  6 0.0298    0    2.18     0 0.458  6.43  58.7  6.06     3   222    18.7
##  7 0.0883   12.5  7.87     0 0.524  6.01  66.6  5.56     5   311    15.2
##  8 0.145    12.5  7.87     0 0.524  6.17  96.1  5.95     5   311    15.2
##  9 0.211    12.5  7.87     0 0.524  5.63 100    6.08     5   311    15.2
## 10 0.170    12.5  7.87     0 0.524  6.00  85.9  6.59     5   311    15.2
## # ... with 496 more rows, and 3 more variables: b <dbl>, lstat <dbl>,
## #   medv <dbl>
```

check each var

From the result below, we can find that there are some variables only have few unique values. they may be factor instead of numeric.

"rad" & "chas" should be factor variables.

```
map(col_uni(df),length)
```

```
## $crim
## [1] 504
##
## $zn
## [1] 26
##
## $indus
## [1] 76
```

```
##
## $chas
## [1] 2
##
## $nox
## [1] 81
##
## $rm
## [1] 446
##
## $age
## [1] 356
##
## $dis
## [1] 412
##
## $rad
## [1] 9
##
## $tax
## [1] 66
##
## $ptratio
## [1] 46
##
## $b
## [1] 357
##
## $lstat
## [1] 455
##
## $medv
## [1] 229
```