Problem Set 1: An Introduction to Predictive Analytics

Shreyansh Kumar Das

2026-01-21

1. Report the "class" of the data set. How many rows and columns are in this data set? What do the rows and columns represent?
Ans.

```
library(MASS)
class(Boston)
```

```
## [1] "data.frame"
```

```
dim(Boston)
```

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

The data set is a "data.frame". The data set contains 506 rows and 14 columns. each row represents a specific suburb in the Boston area and each column represents a specific socio-economic or environment variable(e.g. crime rate, number of rooms, property tax rates etc.).

---

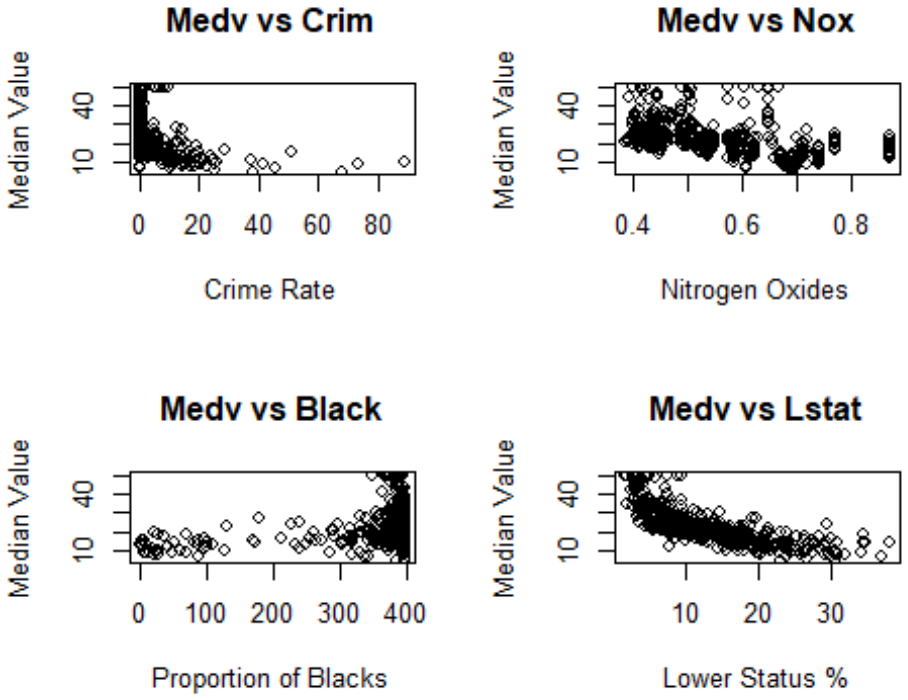
2. Create a smaller data set with the variables median value of owner-occupied homes, per capita crime rate, nitrogen oxides concentration, proportion of blacks and percentage of lower status of the population. Choosing median value owner occupied homes as the response and the rest as the predictors, make scatter plots of the response vesus each predictor. Present the scatter plots in different panels of the same graph. Comment on your finding.
Ans.

```
small_Boston=Boston[,c("medv","crim","nox","black","lstat")]
par(mfrow=c(2,2))
plot(small_Boston$crim, small_Boston$medv, xlab="Crime Rate", ylab="Median
Value", main="Medv vs Crim")

plot(small_Boston$nox, small_Boston$medv, xlab="Nitrogen Oxides",
ylab="Median Value", main="Medv vs Nox")

plot(small_Boston$black, small_Boston$medv, xlab="Proportion of Blacks",
ylab="Median Value", main="Medv vs Black")
```

```
plot(small_Boston$lstat, small_Boston$medv, xlab="Lower Status %",
ylab="Median Value", main="Medv vs Lstat")
```



Findings: There is a strong negative between *lstat* and *medv*, suggesting that as percentage of lower-status population increases, ome prices drops significantly. *crim* and *nox* also show negative relationships with home prices. The *black* variable shows a more complex distribution, with many high values accross all price points.

---

3. Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors mentioned in (2), for that suburb. How do these values compare to the overall ranges for those predictors? Comment on your findings.
Hint: Mention which percentile these values belong to.
Ans.

```
min_med=which.min(Boston$medv)
Boston[min_med, c("medv", "crim", "nox", "black", "lstat")]
```

```
##     medv    crim   nox black lstat
## 399    5 38.3518 0.693 396.9 30.59
```

```
vars=c("crim", "nox", "black", "lstat")
```

```
percentiles=sapply(vars, function(v) {ecdf(Boston[[v]])(Boston[min_med, v])})
percentiles
```

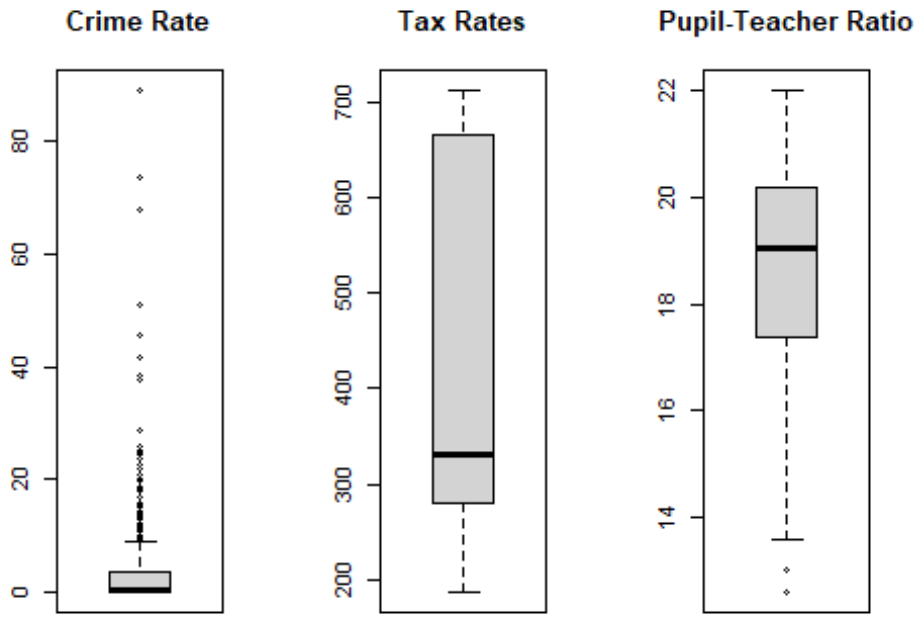
```
##      crim       nox     black     lstat
## 0.9881423 0.8577075 1.0000000 0.9782609
```

Findings: The suburbs with the lowest home values ($5,000) are extreme outliers characterized by severe social and environmental distress. These areas sit in the 98th percentile for crime and the top 10% for “lower status” population, while also having a 100th percentile “black” index, indicating high demographic concentration. They face heavy pollution with nitrogen oxide levels in the 82nd percentile, likely due to their proximity to industrial zones. Ultimately, these suburbs stand out because their crime rates are nearly 100 times higher than the Boston average.

---

4. Does any suburb of Boston stand out for having notably high crime rates,tax rates, or pupil–teacher ratios? Hint: Use a boxplot to detect any outliers. If so, identify the suburbs that show the outlier values.
Ans.

```
par(mfrow = c(1, 3))
boxplot(Boston$crim, main="Crime Rate")
boxplot(Boston$tax, main="Tax Rates")
boxplot(Boston$ptratio, main="Pupil-Teacher Ratio")
```



```r
outlier_crime = which(Boston$crim > 20);outlier_crime
```

```
##  [1] 379 381 385 387 388 399 401 404 405 406 407 411 414 415 418 419 428
441
```

```r
outlier_tax = which(Boston$tax > 600);outlier_tax
```

```
##   [1] 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373
374
##  [19] 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391
392
##  [37] 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409
410
##  [55] 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427
428
##  [73] 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445
446
##  [91] 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463
464
## [109] 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481
482
## [127] 483 484 485 486 487 488 489 490 491 492 493
```

```r
outlier_ptratio = which(Boston$ptratio < 14);outlier_ptratio
```

```
##  [1] 197 198 199 258 259 260 261 262 263 264 265 266 267 268 269 284
```

Findings: There are many outliers at the high end of crime rate. While there are no individual dots as outliers, there is a large cluster of suburbs with a high tax rate of 666 that stands far apart from the median. There are several outliers on the low end(suburbs with very small class sizes), but the high end is uniform.
