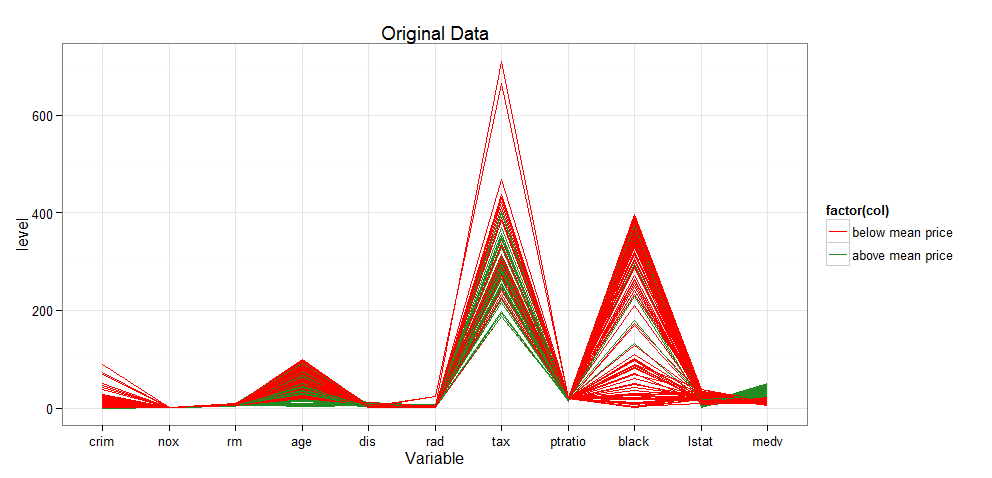
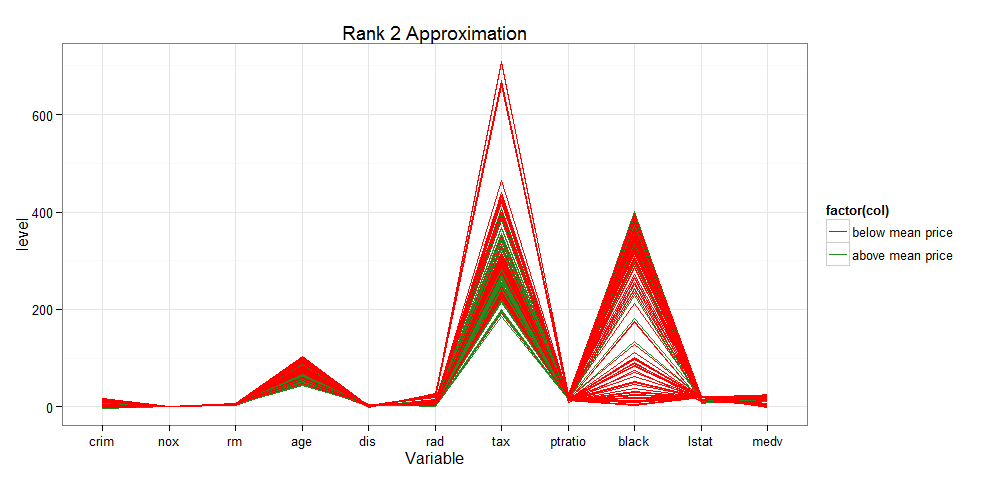
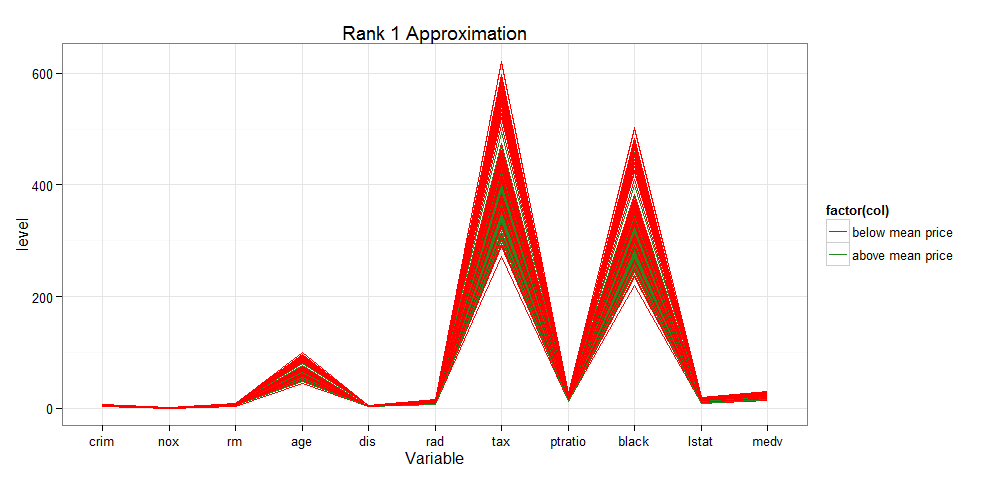
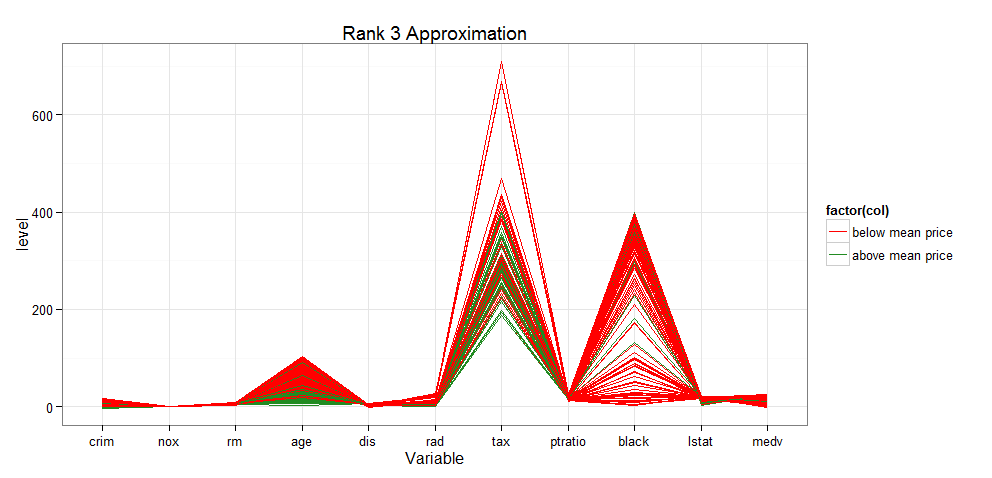
1. a) Parallel coordinate plot of original data



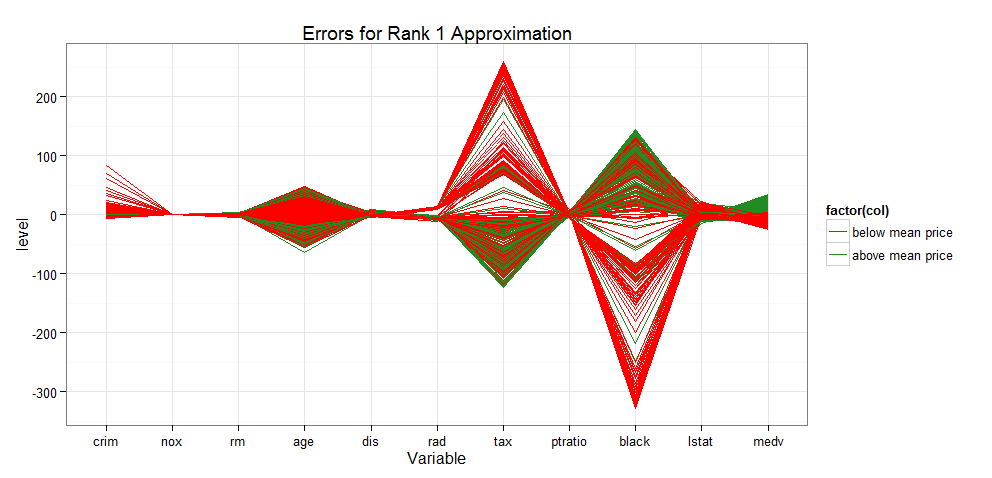
This plot is color coded to show the properties that are above (green) and below (red) the mean price value. There appears to be a negative correlation of proportion of older units that are occupied by owners (“age”), per capital crime rate to value. Nitric oxide concentration, number of rooms per dwelling, distance to employment center and parent-teacher ratio all seem to move in the same way.

(b) Parallel coordinate plots of rank *k* approximations

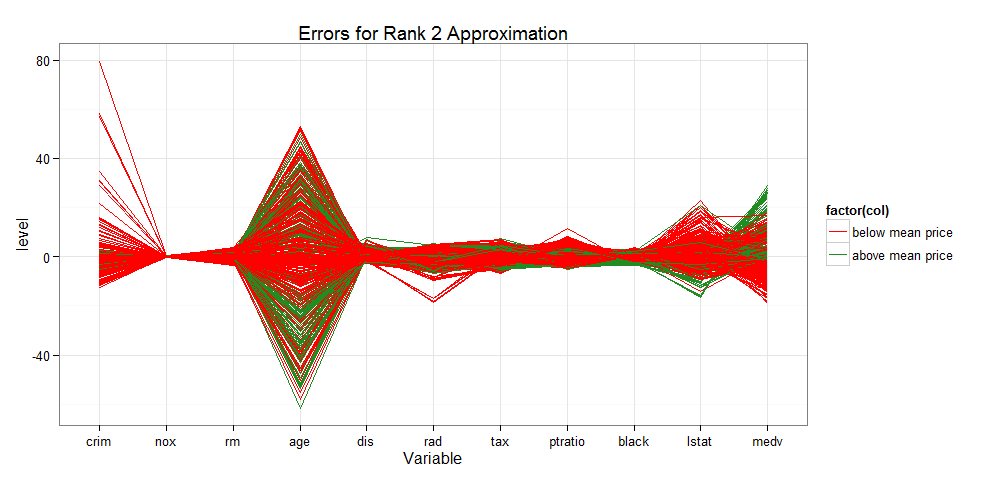




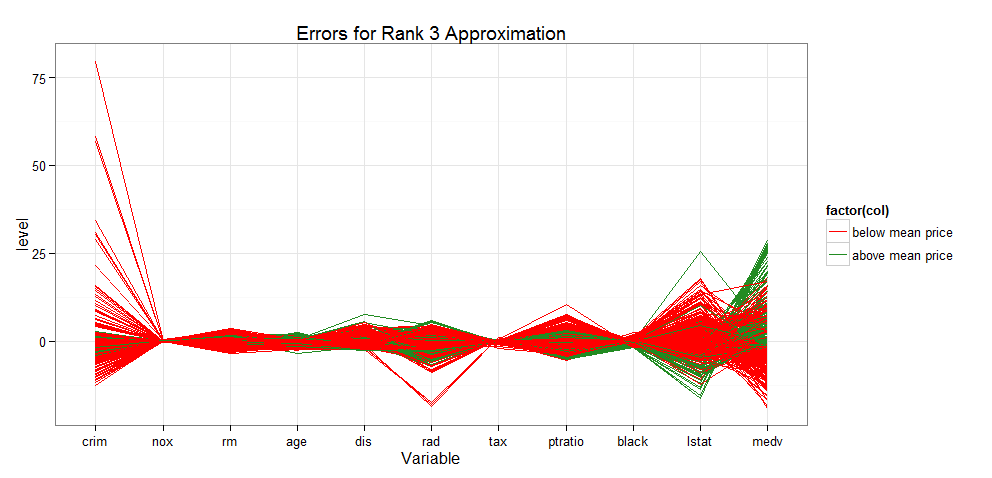
c) Parallel coordinate plots of residual matrices for rank 1, 2, and 3 approximations. Scales are different for each plot, with the plot for rank 1 approximation showing the largest range of errors.



In Rank 1 approximation the largest errors occur in the full-value property and proportion of African-Americans variables

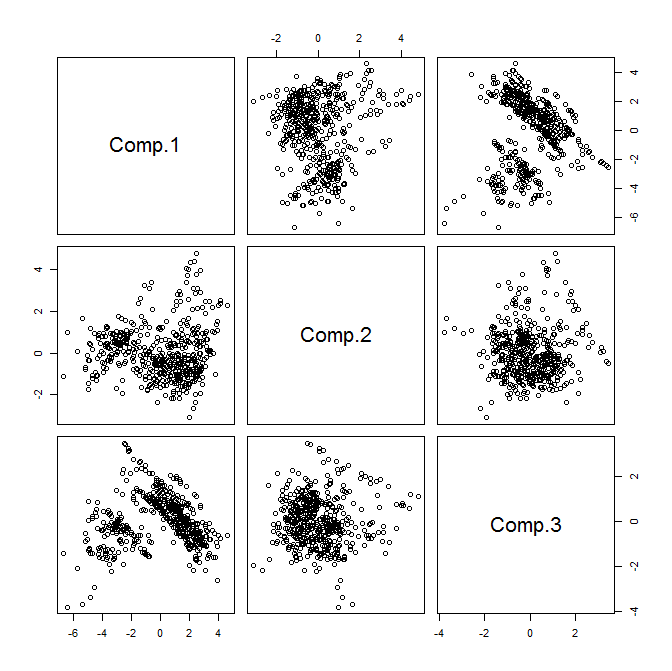
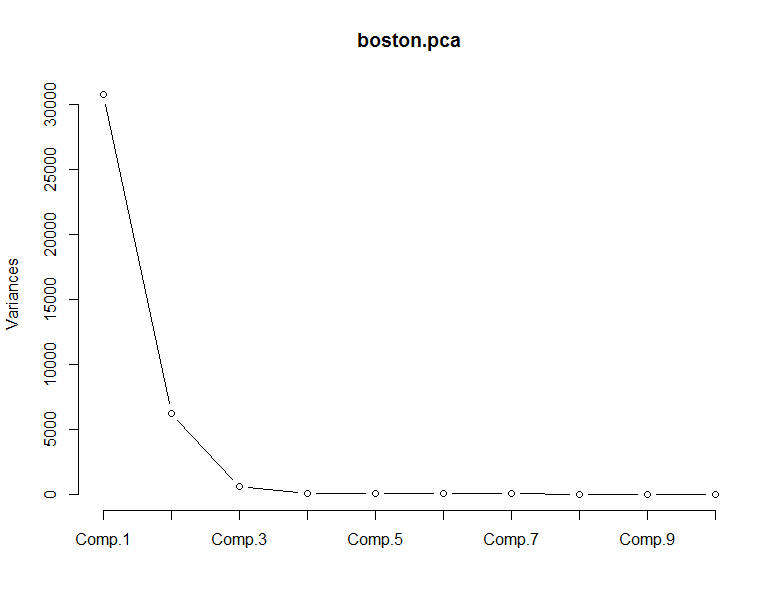


Overall errors are smaller for Rank 2 approximation, with the largest errors occurring in the crime rate and full-value property variables.



The overall range of errors is smaller in Rank 3 approximation than in Rank 2 approximation.

d) PCA of centered and scaled data

 1) PCA scatter plot. There appears to be a separation of groups when looking at PC 1 versus PC 3. When colored according to median value, these groups did not represent a higher or lower price (but perhaps this is because we are using X14 in the analysis itself).

2) The scree plot indicates that the majority of variation is explained by the first three Principal Components, with the remainder adding very little additional information.

2.

a) Eigenvector for second largest eigenvalue

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cement | Slag | Ash | Water | Plasticizer | Coarse | Fine | Age |
| 0.11373709 | -0.68605290 | 0.14294751 | -0.05325628 | -0.28292960 | 0.62994342 | 0.01939111 | 0.12598089 |

Coarse aggregate and blast furnace slag contribute approximately equal amounts to PC 2, but in opposite directions, and each contribution twice times the next highest factor, which is superplasticizer. Each of cement, fly ash, and aging contribute about half of the superplasticizer contribution, and in the opposite direction. Water and fine aggregate are not important contributors to PC 2.

b)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cement | Slag | Ash | Water | Plasticizer | Coarse | Fine | Age |
| -0.44616267 | -0.43738376 | -0.38188581 | -0.38874117 | -0.05174995 | -0.34931986 | -0.43336994 | -0.01288097 |

All but two of the variables contribute approximately equal weights to the last PC. Superplasticizer and aging have very little contribution to this PC. The last PC explains 0.4% of the variability.

c) Correlation matrix

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cement | Slag | Ash | Water | Plasticizer | Coarse | Fine | Age |
| Cement | 1.00 | 0.28 | -0.40 | -0.08 | 0.09 | -0.11 | -0.22 | 0.08 |
| Slag | 0.28 | 1.00 | -0.32 | 0.11 | 0.04 | -0.28 | -0.28 | -0.04 |
| Ash | -0.40 | -0.32 | 1.00 | -0.26 | 0.38 | -0.01 | 0.08 | -0.15 |
| Water | -0.08 | 0.11 | -0.26 | 1.00 | -0.66 | -0.18 | -0.45 | 0.28 |
| Plasticizer | 0.09 | 0.04 | 0.38 | -0.66 | 1.00 | -0.27 | 0.22 | -0.19 |
| Coarse | -0.11 | -0.28 | -0.01 | -0.18 | -0.27 | 1.00 | -0.18 | 0.00 |
| Fine | -0.22 | -0.28 | 0.08 | -0.45 | 0.22 | -0.18 | 1.00 | -0.16 |
| Age | 0.08 | -0.04 | -0.15 | 0.28 | -0.19 | 0.00 | -0.16 | 1.00 |

The strongest correlations shown are the negative correlations between fly ash and cement, superplasticizer and water , and fine aggregate and water. Superplasticizer and fly ash also show a moderate positive correlation.

d) There is some multicollinearity in the explanatory variables, particularly among fly ash, cement, water, and superplasticizer. Correlation in the explanatory variables results in non-unique coefficient estimates in ordinary least squares analysis. The variances and covariances are therefore large.

e) For a regression model, I would use at least the first 5 principal components. These PCs explain 87.5% of the variance.