**Class Assignment 1 – PCA Analysis**

**a. How did you clean, explore the data?**

The data was cleaned using the “Missing Data Handling” under TRANSFORM tab. The number of rows deleted at this stage were zero.

**Were there any categorical variables in the dataset?**

There were two categorical variables – fuel type and color. These were handled by creating dummies by doing the following:

TRANSFORM tab 🡪 Transform Categorical data 🡪 Create dummies

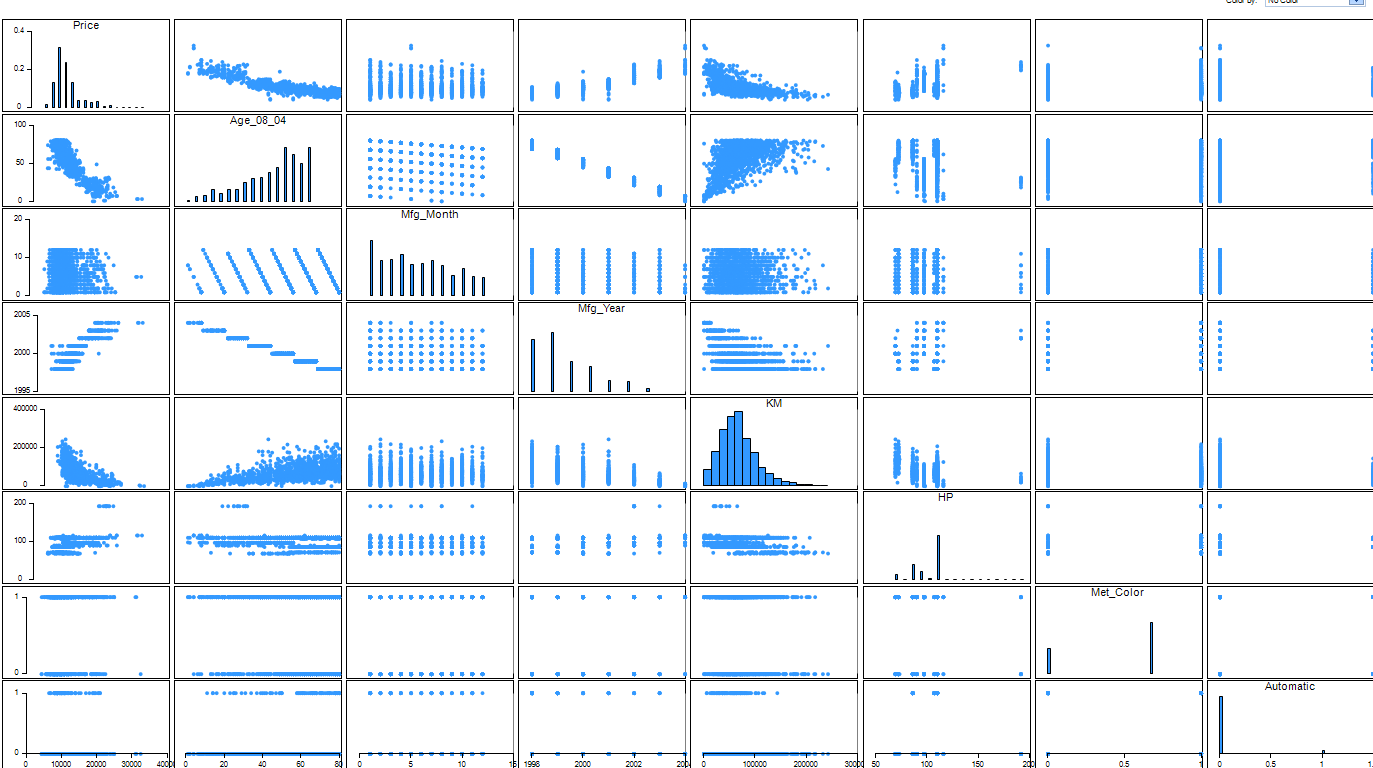
This step was followed for both “Original Data” and on “New Data”.

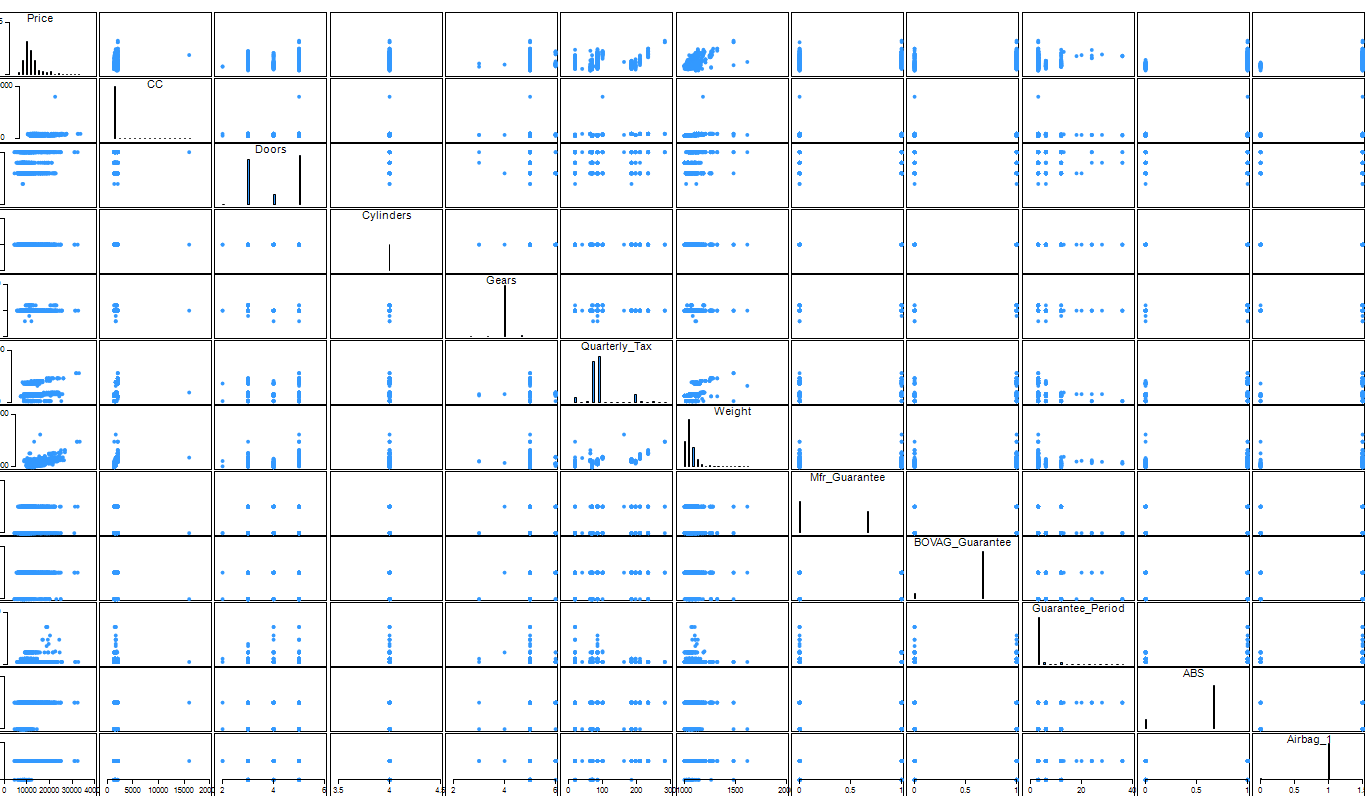
**Explain the scatter plot matrix results in terms of correlation amongst predictors**

**and with the outcome. Provide screen shot of the matrix.**

Outcome Vs Numerical Predictors: As seen in the top row of the two scatterplot snapshots, price has a positive correlation with age of the cars. Also some kind of relationship exists between the price and the accumulated kilometers on the odometer known as the “KM” and the weight of the car. The other variables are classified in nature and so the scatterplots do not reveal any kind of relationship even though there could be one.

Relationship between predictors: No strong correlation is to be seen between any two predictors which means that there is no reason to delete any one variable.

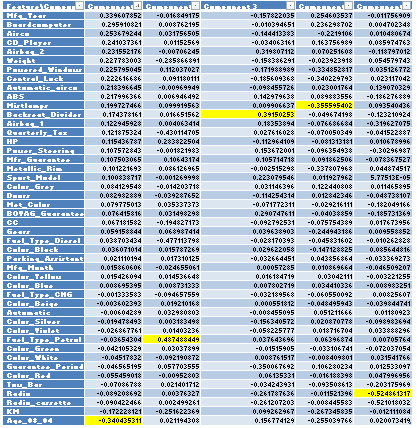




**b. Which top 5 variables did you select using PCA technique?**

The top 5 variables as found by the PCA technique below were :

1. AGE\_08\_04
2. Fuel type – Petrol
3. Backseat Divider
4. Mistlamps
5. Radio



**c. How did you select the best model - explain the ‘full selection criteria’ you**

**followed for comparison between the two models? (Refer to the Lecture 2 slides**

**for ‘full model selection criteria’ for Linear Regression)**

Best model – Model # 3

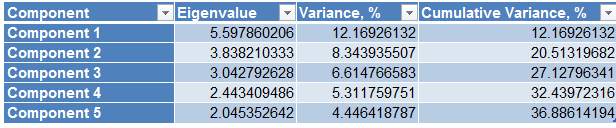
Criteria: The basis of selection of the model was the highest R2 validation score with the lowest RMSE validation score. Amongst the four models, model 3 gave the highest R2 in validation scores. Also the RMSE for validation is greater than RMSE for training.

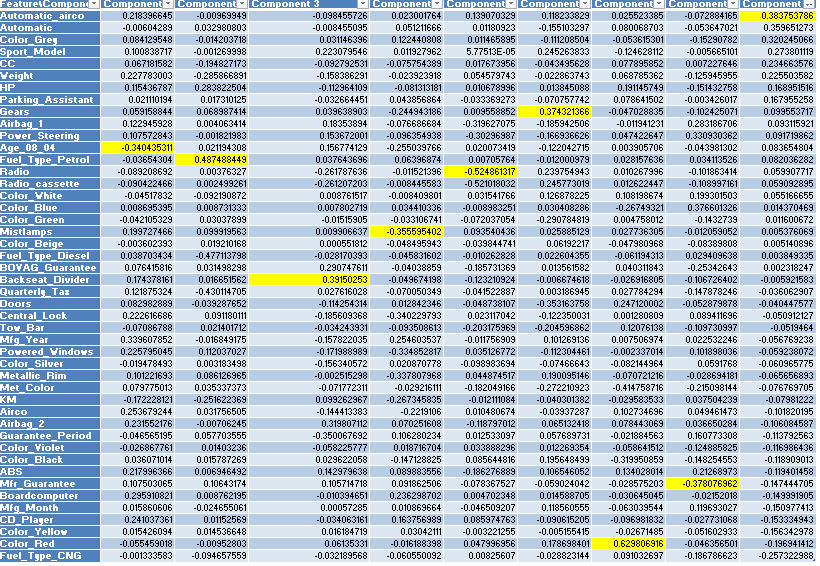
The number of components considered in model 3 was 9 which increased the data considered from 37% to 49% as shown in the snapshots below.

Also, comparing the predicted prices to the original of similar models revealed that they were within an acceptable range.

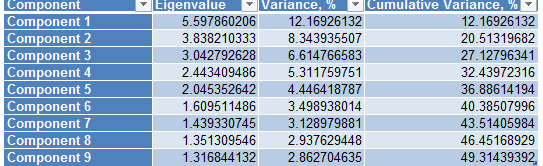
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model #** | **R2** | **RMSE** | **Partition** | **# of PCA components** |
| # 1 | Training – 0.80  Validation – 0.80 | Training - 1690  Validation - 1695 | 60-40 | 5 |
| # 2 | Training – 0.80  Validation – 0.73 | Training - 1187  Validation - 1264 | 75-25 | 5 |
| # 3 | Training – 0.84  Validation – 0.81 | Training – 1478  Validation - 1520 | 60-40 | 9 |
| # 4 | Training – 0.84  Validation – 0.79 | Training – 1463  Validation - 1571 | 75-25 | 9 |

**5 components – 37 % data considered**





**9 components – 49 % data considered**



Output

|  |  |
| --- | --- |
| **Record ID** | **Prediction: Price** |
| **Record 1** | 15672.61832 |
| **Record 2** | 15218.04529 |
| **Record 3** | 15763.82035 |
| **Record 4** | 15006.19863 |
| **Record 5** | 14650.01669 |
| **Record 6** | 13952.71728 |
| **Record 7** | 14611.9479 |
| **Record 8** | 14400.10124 |
| **Record 9** | 14403.78061 |