Jacob Perrone

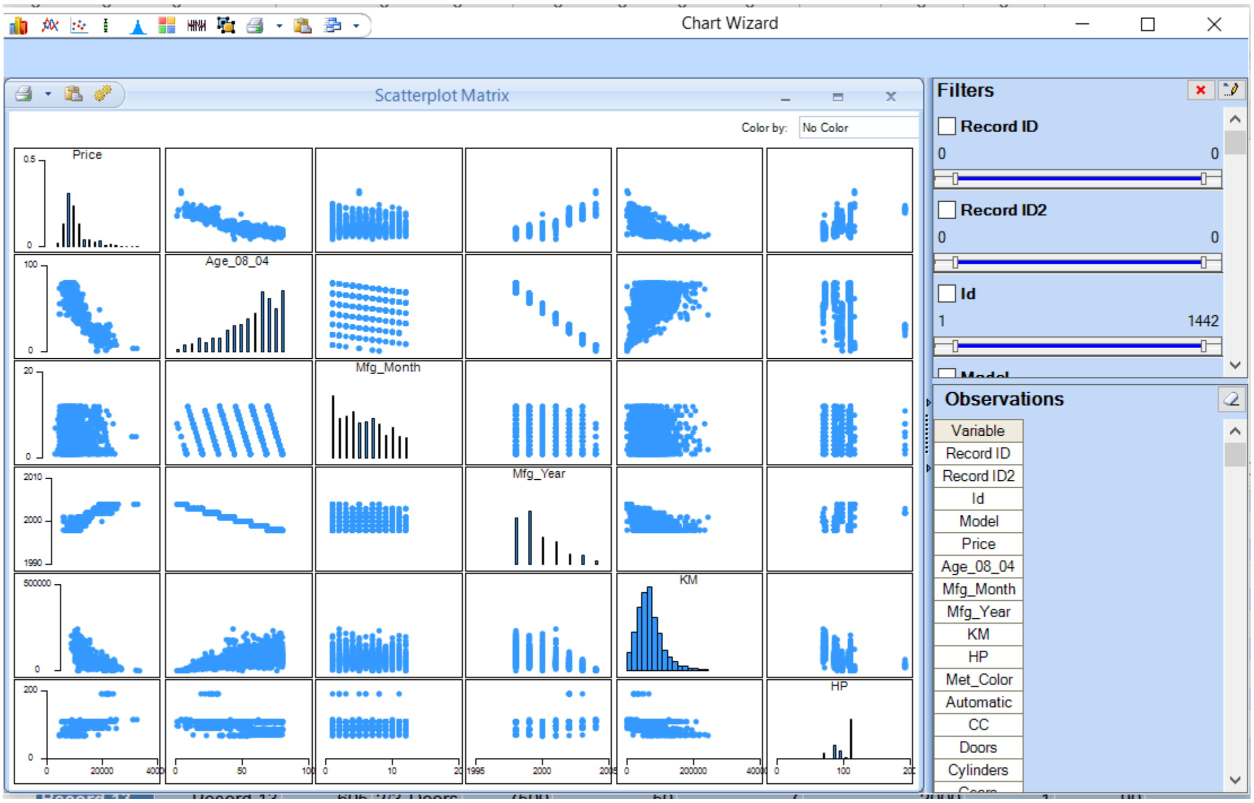
Graded Assignment 1 Report

* on the excel workbook the LInRegOutput6 has the best model and the tab is colored yellow.

** the tab top5 and 95% variance correspond to the tables containing the summary statistics of the models for the respective PCA method.

1. Data handling/ Missing data/ 11a

Going through each column of predictors and sorting in descending order to find invalid or missing cells. There were none. Then I created dummy values for the features fuel_type and color. Once that was done the following screenshots are of the scatterplot matrices. The first is using the first five predictors and the second uses predictors I thought could have lead to dimension reduction as they are highly correlated. In the second screenshot, those variables seem correlated but PCA did not have any of them selected as a top predictor (up to top 15).



Color by: No Color

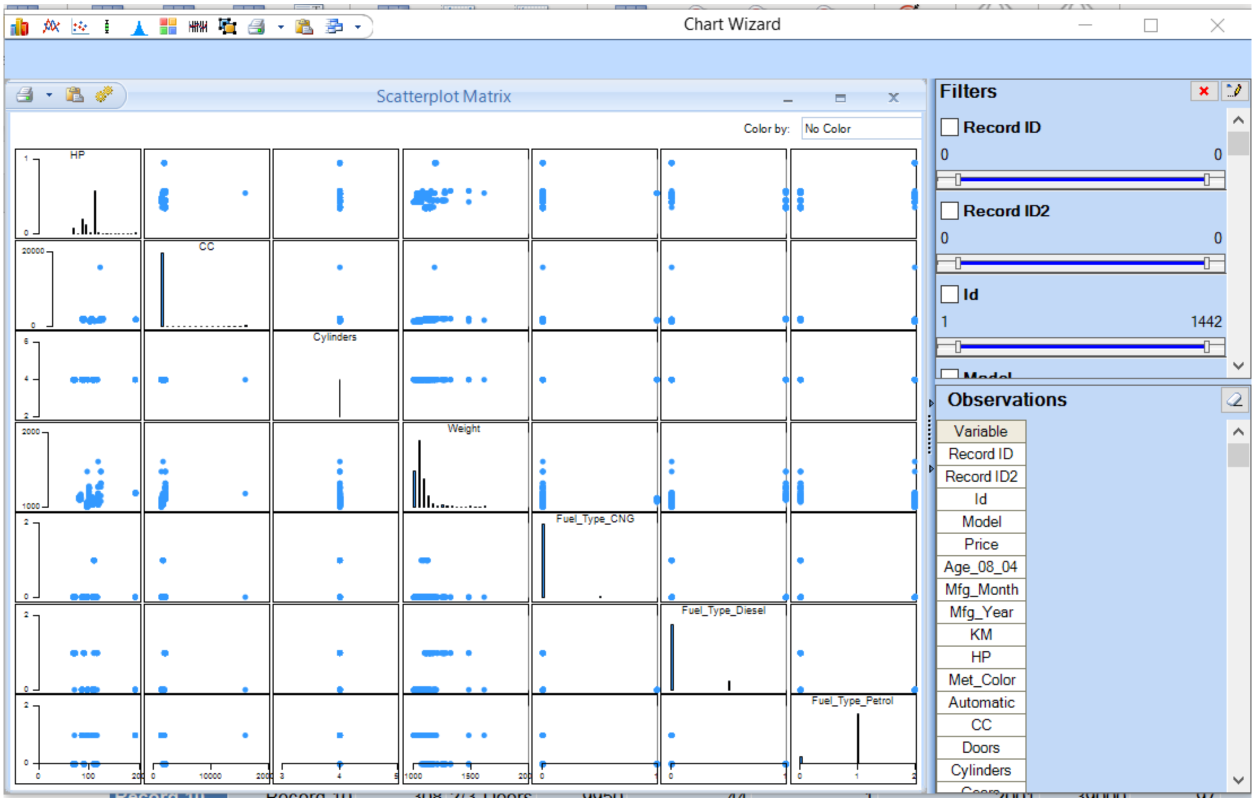
☐ Record ID
0 0

☐ Record ID2
0 0

☐ Id
1 1442

☐ Model

Variable
Record ID
Record ID2
Id
Model
Price
Age_08_04
Mfg_Month
Mfg_Year
KM
HP
Met_Color
Automatic
CC
Doors
Cylinders
Gears

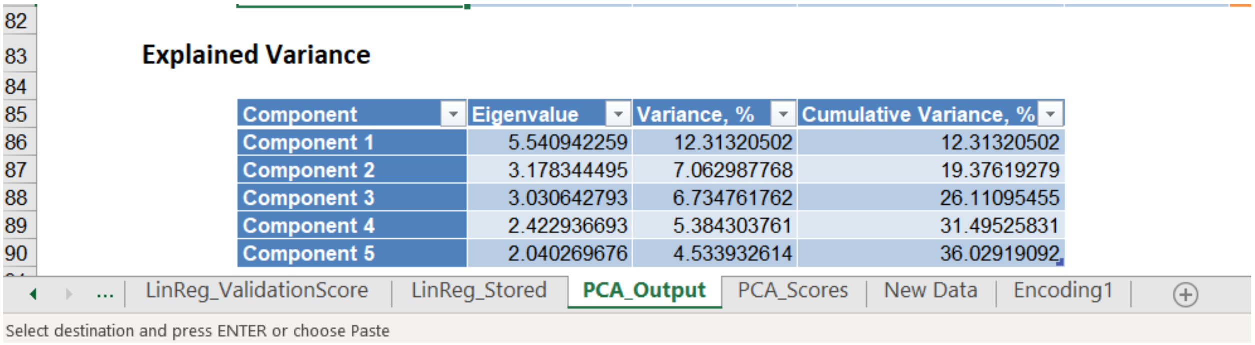


11b.

The top 5 predictors were:

Age_08_04
fuel_type_petro
l
backseat divider
mistlamps
radio

These 5 covered 36% of the variance.



Explained Variance

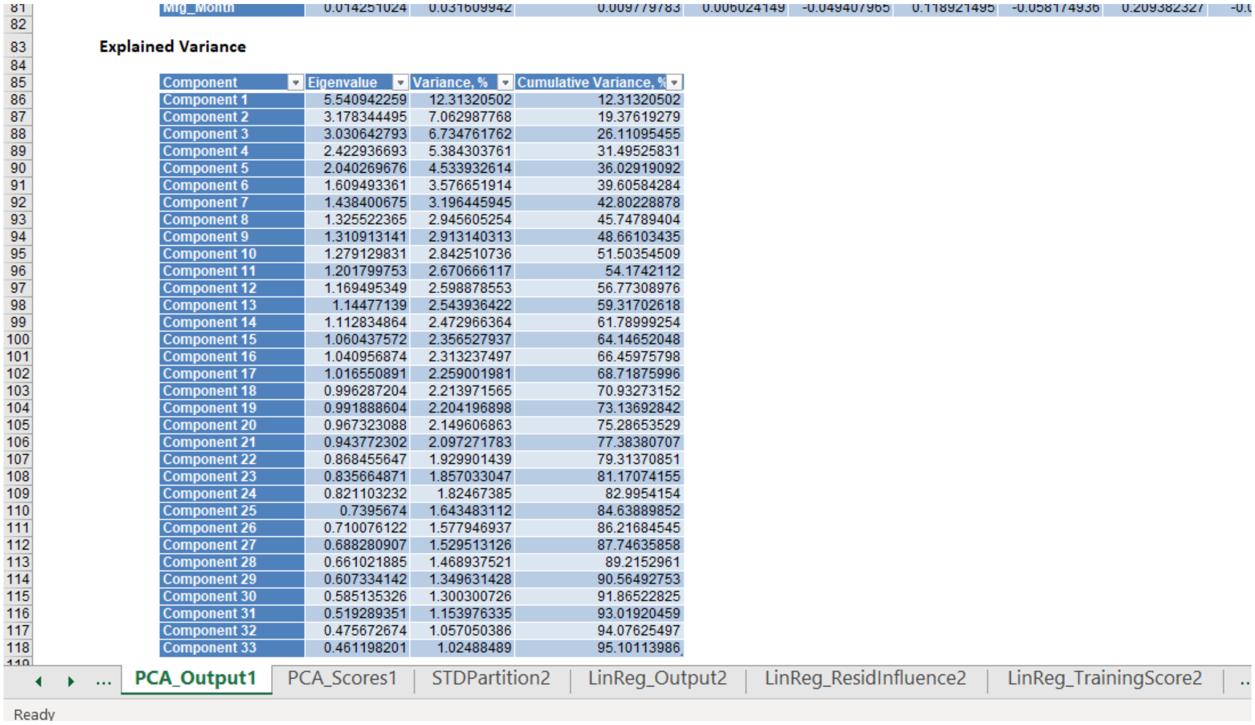
Component	Eigenvalue	Variance, %	Cumulative Variance, %
Component 1	5.540942259	12.31320502	12.31320502
Component 2	3.178344495	7.062987768	19.37619279
Component 3	3.030642793	6.734761762	26.11095455
Component 4	2.422936693	5.384303761	31.49525831
Component 5	2.040269676	4.533932614	36.02919092

Select destination and press ENTER or choose Paste

The top 15 predictors were:

Age_08_04
fuel_type_petrol
backseat divider
mistlamps
radio
gears
color red
color blue
color grey
color green
color black
color white
color violet
* two of them were used for multiple components

The top 15 covered 64% of the variance.



Mfg_Month 0.014251024 0.031609942 0.009779783 0.006024149 -0.049407965 0.118921495 -0.058174936 0.209382327 -0.0

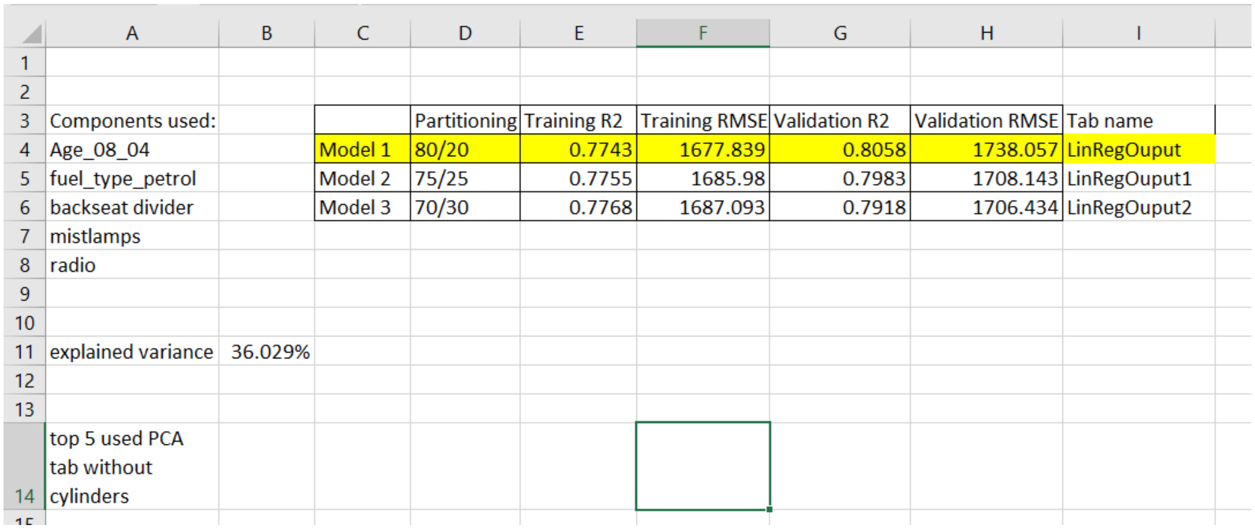
Explained Variance

Component	Eigenvalue	Variance, %	Cumulative Variance, %
Component 1	5.540942259	12.31320502	12.31320502
Component 2	3.178344495	7.062987768	19.37619279
Component 3	3.030642793	6.734761762	26.11095455
Component 4	2.422936693	5.384303761	31.49525831
Component 5	2.040269676	4.533932614	36.02919092
Component 6	1.609493361	3.576651914	39.60584284
Component 7	1.438400675	3.196445945	42.80228878
Component 8	1.325522365	2.945605254	45.74789404
Component 9	1.310913141	2.913140313	48.66103435
Component 10	1.279129831	2.842510736	51.50354509
Component 11	1.201799753	2.670666117	54.1742112
Component 12	1.169495349	2.598878553	56.77308976
Component 13	1.14477139	2.543936422	59.31702618
Component 14	1.112834864	2.472966364	61.78999254
Component 15	1.060437572	2.356527937	64.14652048
Component 16	1.040956874	2.313237497	66.45975798
Component 17	1.016550891	2.259001981	68.71875996
Component 18	0.996287204	2.213971565	70.93273152
Component 19	0.991888604	2.204196898	73.13692842
Component 20	0.967323088	2.149606863	75.28653529
Component 21	0.943772302	2.097271783	77.38380707
Component 22	0.868455647	1.929901439	79.31370851
Component 23	0.835664871	1.857033047	81.17074155
Component 24	0.821103232	1.82467385	82.9954154
Component 25	0.7395674	1.643483112	84.63889852
Component 26	0.710076122	1.577946937	86.21684545
Component 27	0.688280907	1.529513126	87.74635858
Component 28	0.661021885	1.468937521	89.2152961
Component 29	0.607334142	1.349631428	90.56492753
Component 30	0.585135326	1.300300726	91.86522825
Component 31	0.519289351	1.153976335	93.01920459
Component 32	0.475672674	1.057050386	94.07625497
Component 33	0.461198201	1.02488489	95.10113986

2. Partitioning and PCA

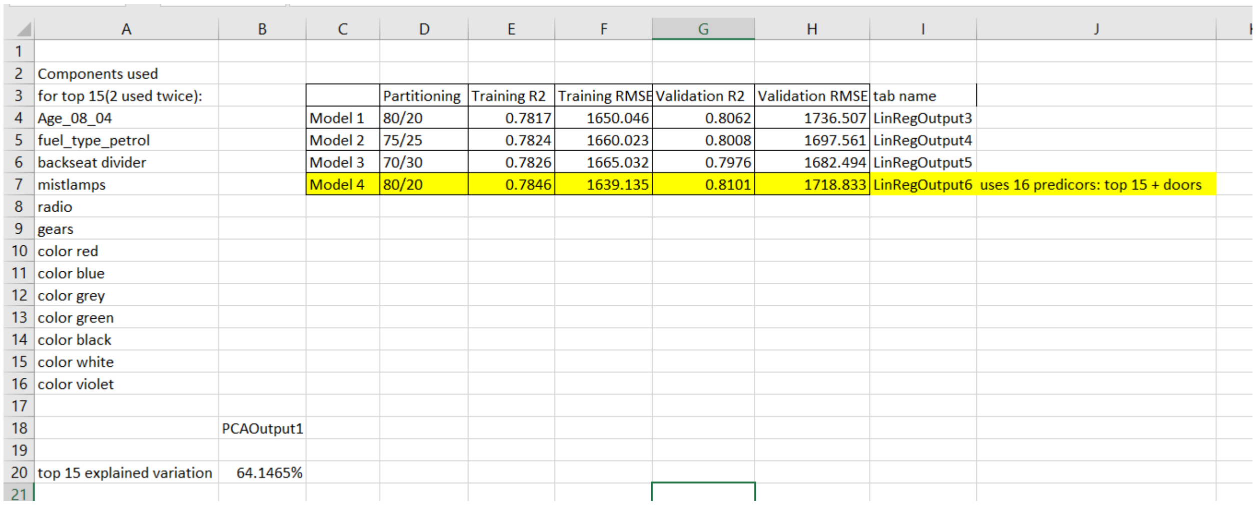
For PCA, I first ran using the top 5 predictors which ended up explaining 36% of the variance. I highlighted the weights in the PCAOutput tab and recorded the components to use for partitioning. I used partitioning with 80/20, 75/25, 70/30 splits and recorded the R2 and RMSE values (in file “Tables for graded assignment 1” as to avoid clutter in the main workbook). Once the results were tabulated I then went back to the encoding tab and ran PCA for 95% of the variance which gave 33 predictors. Since that was a lot I took the top 15 as instructed in the assignment instruction document. As before I highlighted the components weights and recorded the top 15 to pass to partitioning. The model 4 in this table had an extra predictor (doors) which led to better validation R2. As a note, when attempting to run PCA with the cylinders predictor it would not allow me to add it, upon looking at the column all the values were the same so I decided to omit that predictor from PCA.

Top 5 table:



	A	B	C	D	E	F	G	H	I
1									
2									
3	Components used:			Partitioning	Training R2	Training RMSE	Validation R2	Validation RMSE	Tab name
4	Age_08_04		Model 1	80/20	0.7743	1677.839	0.8058	1738.057	LinRegOuput
5	fuel_type_petrol		Model 2	75/25	0.7755	1685.98	0.7983	1708.143	LinRegOuput1
6	backseat divider		Model 3	70/30	0.7768	1687.093	0.7918	1706.434	LinRegOuput2
7	mistlamps								
8	radio								
9									
10									
11	explained variance	36.029%							
12									
13									
14	top 5 used PCA tab without cylinders								

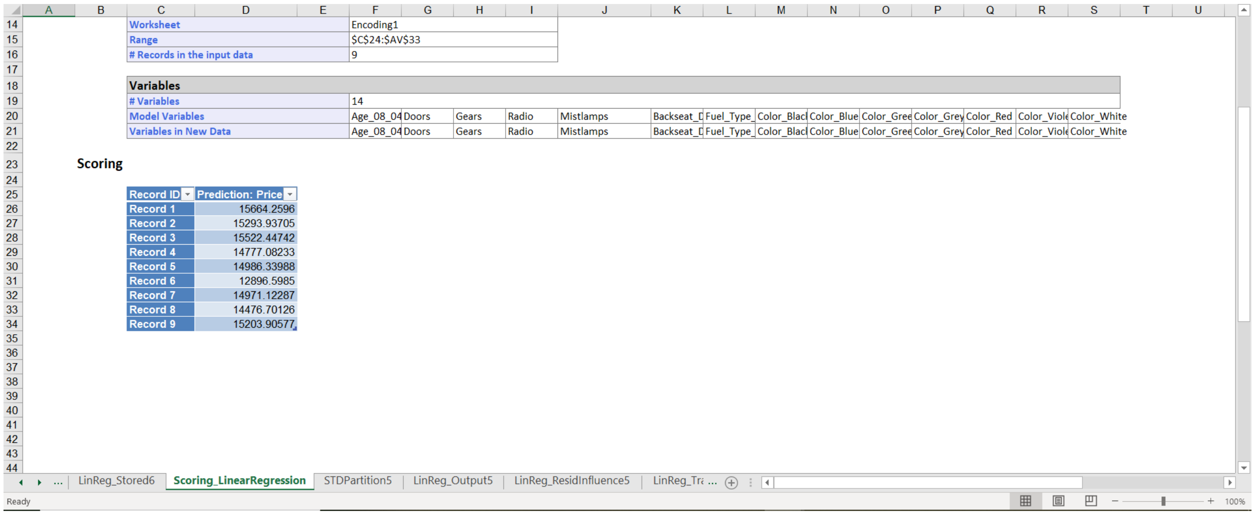
95% variance table:



	A	B	C	D	E	F	G	H	I	J
1										
2	Components used									
3	for top 15(2 used twice):			Partitioning	Training R2	Training RMSE	Validation R2	Validation RMSE	tab name	
4	Age_08_04		Model 1	80/20	0.7817	1650.046	0.8062	1736.507	LinRegOutput3	
5	fuel_type_petrol		Model 2	75/25	0.7824	1660.023	0.8008	1697.561	LinRegOutput4	
6	backseat divider		Model 3	70/30	0.7826	1665.032	0.7976	1682.494	LinRegOutput5	
7	mistlamps		Model 4	80/20	0.7846	1639.135	0.8101	1718.833	LinRegOutput6	uses 16 predicors: top 15 + doors
8	radio									
9	gears									
10	color red									
11	color blue									
12	color grey									
13	color green									
14	color black									
15	color white									
16	color violet									
17										
18		PCAOutput1								
19										
20	top 15 explained variation	64.1465%								
21										

3. Scoring

For choosing the best model out of the 6, first I looked at if the R2 value for both training and validation were > 70% (they all were) and if the two R2 values were close to each other. Then the model with the highest validation R2 value was chosen as the best model for each (top 5 and 95% variation). Then out of the two best I again chose the one with the highest R2 to deploy on the new data and predict the price (model 4). Below is a screenshot of the predicted values on the new data, which I believe to be accurate as the predicted values fall in the range of the prices of the original data.



Worksheet	Encoding1
Range	$C$24:$AV$33
# Records in the input data	9

Variables														
# Variables	14													
Model Variables	Age_08_04	Doors	Gears	Radio	Mistlamps	Backseat_D	Fuel_Type_	Color_Blacl	Color_Blue	Color_Gree	Color_Grey	Color_Red	Color_Viole	Color_White
Variables in New Data	Age_08_04	Doors	Gears	Radio	Mistlamps	Backseat_D	Fuel_Type_	Color_Blacl	Color_Blue	Color_Gree	Color_Grey	Color_Red	Color_Viole	Color_White

Scoring

Record ID	Prediction: Price
Record 1	15664.2596
Record 2	15293.93705
Record 3	15522.44742
Record 4	14777.08233
Record 5	14986.33988
Record 6	12896.5985
Record 7	14971.12287
Record 8	14476.70126
Record 9	15203.90577