STAT8008 PCA Project

Paul Christopher, R00207143

12/17/2021

Introduction

Cars were selected at random from among 93 American passenger car models that were listed in both the Consumer Reports issue and the PACE Buying Guide magazines. Pickup trucks and Sport/Utility vehicles were eliminated due to incomplete information and duplicate models were listed at most once. The variables relevant to this project are defined as follows:

Variable Name	Description
Origin	Manufacturing company origin: 1=”non-USA” or 2=”USA”
Price	Price (in $1,000)
Type	Level: 1=“Small” 2=“Sporty” 3=“Compact” 4=“Midsize” 5=“Large” 6=“Van”
Engine Size	Engine size (litres).
RPM	RPM (revs per minute at maximum horsepower)
Fuel.Tank.Capactity	Fuel tank capacity (US gallons)
MPG.city	City MPG (miles per US gallon by EPA rating)
Weight	Weight (pounds)
Horsepower	Horsepower of the car

Preliminary cleaning and analysis

The original dataset was transformed so as to extract the make and model and use that for the row names. Some variables were removed, such as the number of passengers, insurance category and the length of the car, which were deemed to be uninformative. One row was deleted due to the fact that its cylinder was classed as ‘rotary’, rather than being a numeric value. That brought the total number of observations down to 92 from 93. A further problem with the ‘cylinders’ variable was found in that it was coded as a factor. This would have prevented various EDA and PCA methods being amenable to this variable. This was transformed into a numeric variable using the `as.numeric` method.

A preliminary summary of the dataset is shown at Table 2.

A pairs plot of the remaining variables using the `pairs` function is shown at Figure 1. This shows that all of the variables, apart from RPM, are highly correlated with each other. A slight exception is that RPM seems to be reasonably correlated with fuel tank capacity, and, to a lesser extent, weight.

A correlation matrix is shown at Figure 2.[1]

The relationship between three pairs of variables are examined a little more closely below.

Firstly, the relationship between MPG and Horsepower is examined and a plot produced at Figure 3 (a).

[1] Using the `ggcorrplot` package - see Kassambara (2019).

Table 2: Summary description of the dataset

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Price	92	19.368	9.615	7.400	12.175	22.850	61.900
EngineSize	92	2.887	0.926	1	2.2	3.4	6
Fuel.tank.capacity	92	16.628	3.279	9.200	14.500	18.575	27.000
RPM	92	5,267.391	586.076	3,800	4,800	5,712.5	6,500
Weight	92	3,074.837	592.832	1,695	2,608.8	3,533.8	4,105
MPG.city	92	21.784	11.034	7.900	14.575	25.000	80.000
Cylinders	92	4.967	1.305	3	4	6	8
Horsepower	92	142.620	51.341	55	102.8	170	300

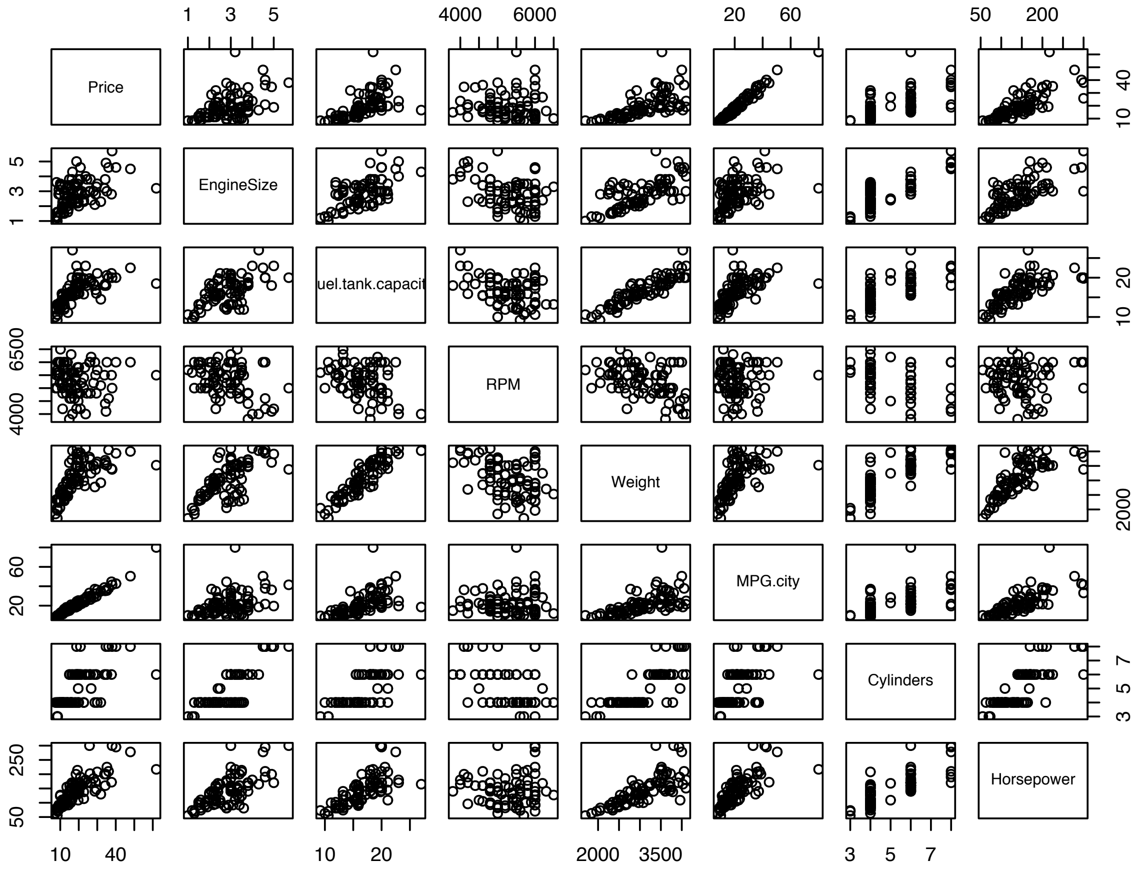


Figure 1: Pairs plot for cars data.

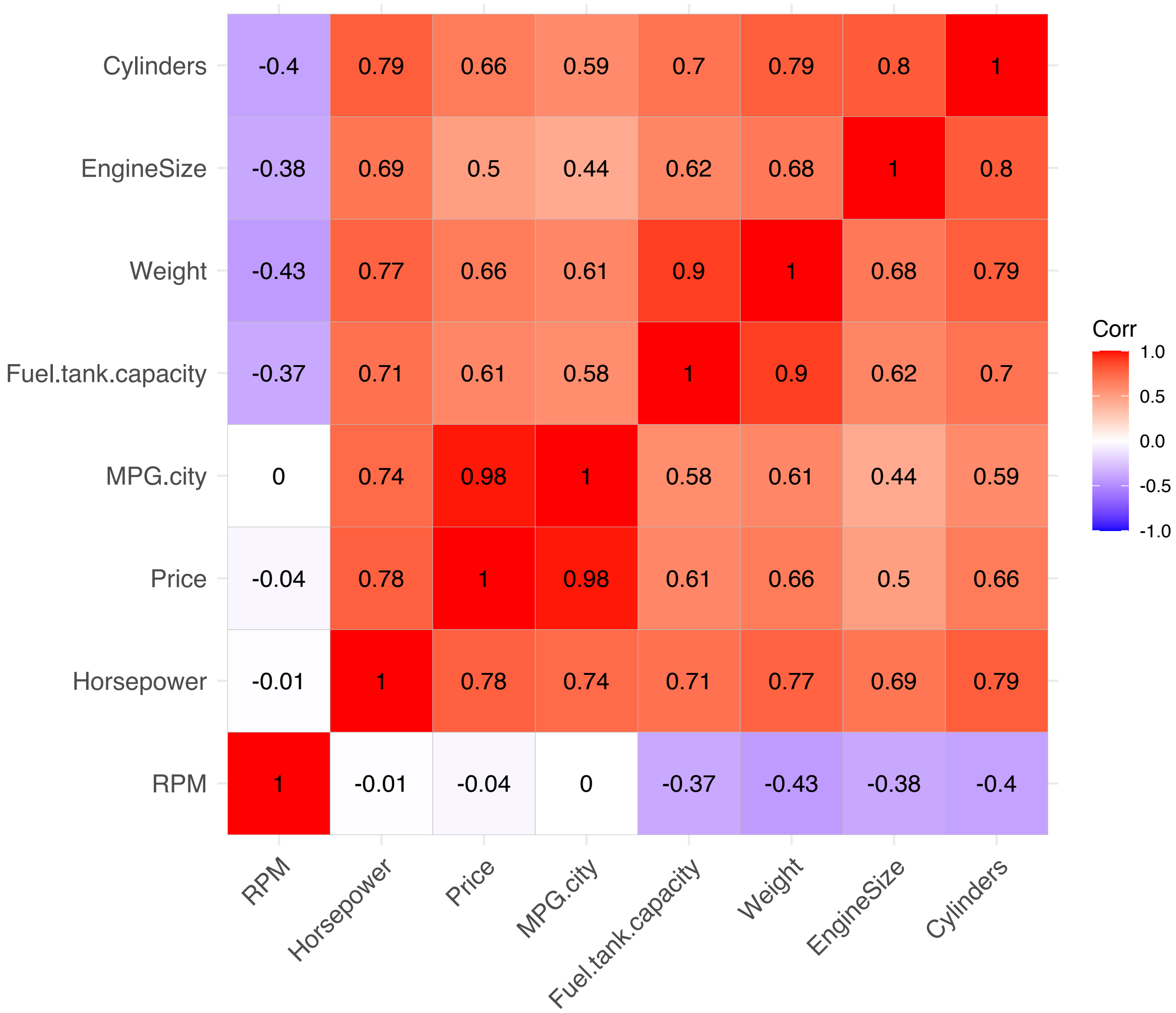
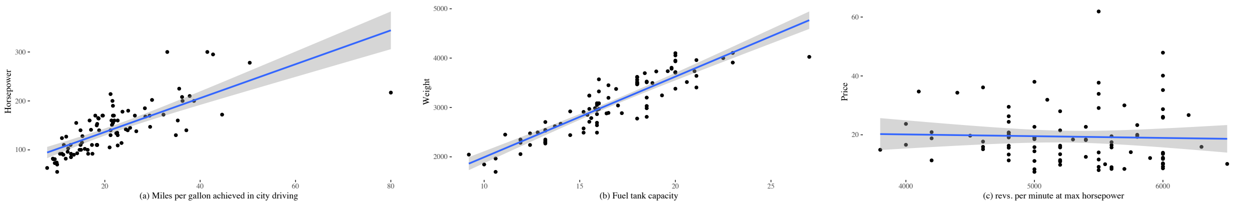


Figure 2: Correlation matrix for car dataset.

Secondly, the relationship between Fuel tank capacity and weight is examined and a plot produced at (b). Finally, the relationship between RPM and Price is examined and a plot produced at (c).



(a) Miles per gallon achieved in city driving

(b) Fuel tank capacity

(c) revs. per minute at max horsepower

Figure 3: Relationship between: MPG v Horsepower; fuel tank capacity v weight; RPM v Price.

3D Plots

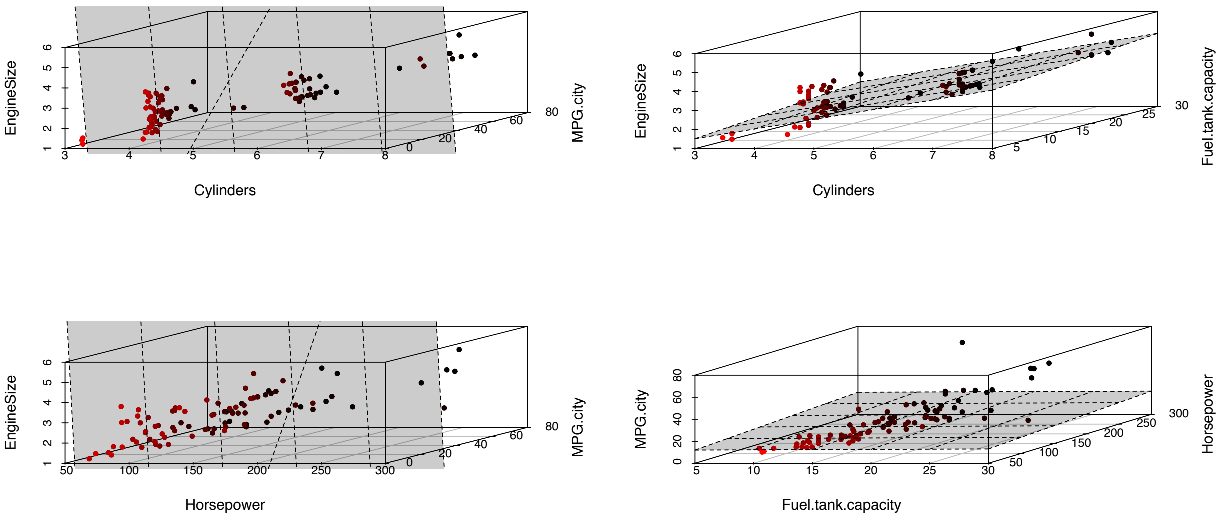


Figure 4: 3D Plots with 2D regression planes

Is the data suitable for data reduction techniques?

To answer this question, a Bartlett's Sphericity test was conducted. The results, shown below, confirm that there is sufficient correlation amongst the variables to carry out data reduction techniques such as PCA or factor analysis, with a χ^2 value of 904 and an associated p-value of less than 0.001.[2]

```
# Test of Sphericity

Bartlett's test of sphericity suggests that there is sufficient significant correlation in the data for
```

[2] Bartlett's (1951) test of sphericity tests whether a matrix (of correlations) is significantly different from an identity matrix. The test provides probability that the correlation matrix has significant correlations among at least some of the variables in a dataset: Lüdecke et al. (2020).

Principal Components Analysis

In the first instance, PCA was performed using the `prcomp` function from base R. The variables were standardised before performing the PCA, using the `scale = TRUE` parameter. In parallel, the same PCA was done using the `PCA` function from the `FactoMineR` package. Results from the two models are examined and compared under various headings below.

(a) Eigenvalues and eigenvectors

Using the `prcomp` function the eigenvalues for each component were as follows:

```
[1] 5.27778839 1.35892942 0.55444575 0.43561618 0.18550692 0.09872687 0.07540567 0.01358078
```

The eigenvectors are printed below at Table 3.

Table 3: Eigenvectors as calculated by `prcomp`.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Price	0.37	0.37	0.32	-0.28	0.08	0.08	0.03	-0.73
EngineSize	0.34	-0.24	-0.57	-0.32	0.63	0.03	0.08	0.01
Fuel.tank.capacity	0.37	-0.14	0.16	0.64	0.25	0.36	-0.46	-0.01
RPM	-0.14	0.72	-0.47	0.33	0.06	0.28	0.21	0.01
Weight	0.40	-0.17	0.11	0.40	-0.06	-0.22	0.77	0.01
MPG.city	0.35	0.41	0.39	-0.27	0.16	0.03	0.01	0.68
Cylinders	0.39	-0.15	-0.26	-0.23	-0.65	0.53	0.02	0.08
Horsepower	0.39	0.21	-0.32	0.12	-0.28	-0.67	-0.39	0.00

Using the `FactoMineR` package, the eigenvalues are reported at Table 4. This output also includes the percentage of variance and cumulative percentage of variance.

Table 4: Eigenvalues and percentage of variance from FactoMineR

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	5.278	65.972	65.972
comp 2	1.359	16.987	82.959
comp 3	0.554	6.931	89.890
comp 4	0.436	5.445	95.335
comp 5	0.186	2.319	97.654
comp 6	0.099	1.234	98.888
comp 7	0.075	0.943	99.830
comp 8	0.014	0.170	100.000

(b) Amount of information explained by components

The amount of information explained by the components as well as the cumulative variance from the model created by the `prcomp` function can be seen in Table 5. The same information, as calculated by `FactoMineR` can be seen at Table 4.

(c) Factor loadings

The factor loadings calculated by `prcomp` for each original variable are displayed at Table 3. They are equivalent to the eigenvectors for each component. They are computed from the correlation matrix[3] of the

[3] Usually denoted S.

Table 5: Information and variance explained by components (as calculated by ‘prcomp‘)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.297	1.166	0.745	0.660	0.431	0.314	0.275	0.117
Proportion of Variance	0.660	0.170	0.069	0.054	0.023	0.012	0.009	0.002
Cumulative Proportion	0.660	0.830	0.899	0.953	0.977	0.989	0.998	1.000

original m x n data matrix[4], where m are the number of observations and n are the number of variables[5] in the data matrix and where $\lambda_i, i \in \{1, ..., n\}$, are the eigenvalues of S. It can be computed using the spectral decomposition theorem which states as follows: *For any real, symmetric N x N matrix S, there exists an orthogonal matrix U such that*:

$$B = \begin{pmatrix} \lambda_1 & 0 & ... \\ 0 & \lambda_2 & ... \\ ... & & \\ ... & 0 & \lambda_n \end{pmatrix} = U^{-1}SU$$

is a diagonal matrix.

The entries λ_i in the diagonal matrix are the eigenvalues of matrix S and the column vectors of U are the eigenvectors.[6] The eigenvectors represent the principal components of S and the elements of the eigenvectors of S are the ‘coefficients’ or ‘loadings’ of the principal components. The equivalent output from `FactoMineR` can be seen at Table 7.

However, most software packages, such as R, do not use this method, instead opting for the more computationally efficient singular value decomposition (SVD) method. This is the method used by `prcomp`.[7]

The scores for the first five cars calculated by `prcomp` are shown at Table 6. The scores can be interpreted as being the co-ordinates of the observations on the new, calculated, principal components or dimensions.

Table 6: Scores of the rotated data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Acura Integra	-1.387	1.353	-1.138	-0.039	0.379	-0.135	0.380	0.037
Acura Legend	2.379	1.320	0.287	-0.269	-0.170	-0.082	0.187	0.004
Audi 90	1.381	0.983	0.237	-0.236	-0.486	0.165	0.258	-0.026
Audi 100	2.600	1.580	1.162	0.046	0.088	0.718	-0.256	0.063
BMW 535i	2.083	1.349	0.243	0.936	1.166	-0.614	-0.148	0.021

(d) Correlation between original variables and principal components

The correlation between the original components and each principal component is provided, in `prcomp`, by the eigenvector associated with that principal component. So, from an examination of Table 3, one can see that the first principal component is positively correlated, roughly to the same extent, with all of the original variables, apart from RPM.

Similarly, the second principal component is most significantly correlated with RPM and only weakly so with all the other original variables. This suggests that the first principal component represents all of the variables apart from RPM, whilst the second PC represents RPM.

A correlation matrix between the original variables and the first two PCs produced by `FactoMineR` is at Table 7.

[4] Here denoted D.

[5] Sometimes called attributes.

[6] from Janert (2011), p.328 *et. seq.*

[7] However, the other base R function, `princomp`, uses the spectral decomposition approach. Hartmann, Krois, and Waske (2018). Despite the different algorithms, both `prcomp` and `princomp` produced the same results on this dataset. The results from `printcomp` are not produced here for reasons of brevity but can be checked on a separate R script file on request.

Table 7: Correlation matrix between variables and PCs produced by FacotMineR

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Price	0.847	0.429	-0.237	-0.184	0.035
EngineSize	0.792	-0.276	0.421	-0.213	0.270
Fuel.tank.capacity	0.861	-0.166	-0.118	0.421	0.109
RPM	-0.329	0.843	0.348	0.218	0.026
Weight	0.913	-0.196	-0.083	0.266	-0.027
MPG.city	0.802	0.480	-0.288	-0.181	0.070
Cylinders	0.897	-0.171	0.195	-0.149	-0.279
Horsepower	0.896	0.250	0.240	0.078	-0.122

A biplot is shown at Figure 5. The biplot combines the principal component scores and the loadings from the eigenvectors in a single plot. As can be seen, most of the eigenvectors associated with the original variables (apart from RPM) are roughly in the direction of the first principal component, whilst the eigenvector associated with the RPM variable is almost orthogonal to those other eigenvectors, in the direction of the second principal component. This reflects what we saw earlier in the correlation plots and matrices.

Amongst the eigenvectors that are aligned in the direction of the first PC, they can be split into two sub-groups, those that are positive in the direction of the second PC (MPG, horsepower and price) and, those that are in the negative direction of the second PC (engine size, weight, fuel tank capacity).

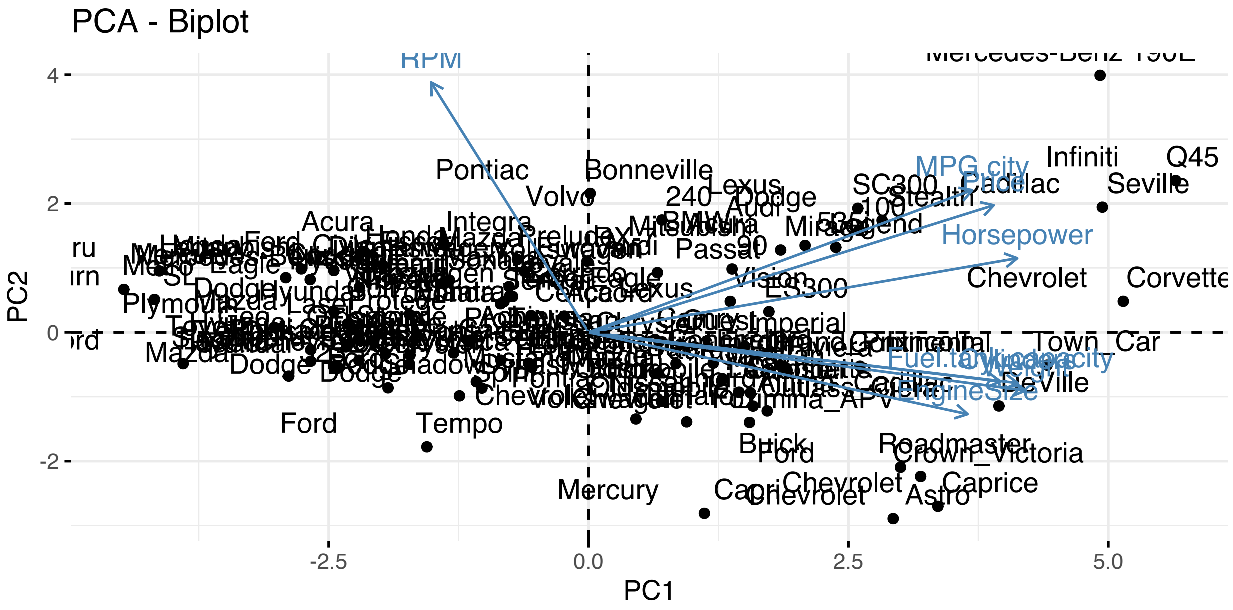


Figure 5: Biplot of the data

The quality of representation of the variables on a factor map is called cos2 (square cosine, squared coordinates). A matrix of variables cos^2 values from the `factoMineR` model is shown at Table 8. Note that:

- A high cos^2 indicates a good representation of the variable on the principal component. In this case the variable is positioned close to the circumference of the correlation circle;
- A low cos^2 indicates that the variable is not perfectly represented by the PCs. In this case the variable is close to the center of the circle;

- For a given variable, the sum of the cos^2 on all the principal components is equal to one.[8]

Table 8: Cos^2 of variables from `FactoMineR`

	PC1	PC2	PC3	PC4	PC5
Price	0.717	0.184	0.056	0.034	0.001
EngineSize	0.627	0.076	0.177	0.045	0.073
Fuel.tank.capacity	0.741	0.027	0.014	0.177	0.012
RPM	0.108	0.711	0.121	0.048	0.001
Weight	0.834	0.039	0.007	0.071	0.001
MPG.city	0.643	0.230	0.083	0.033	0.005
Cylinders	0.805	0.029	0.038	0.022	0.078
Horsepower	0.803	0.062	0.058	0.006	0.015

(e) How many components to retain?

A screeplot, as produced by `prcomp`, is shown at Figure 6. It suggests that two components should be retained. This is unsurprising, given that all of the variables appeared to be explained by the first two principal components: the second component representing the RPM variable and; the first component representing all of the other variables (which, as was seen earlier, were all highly correlated with each other.) Morevoer, as can be seen from Table 4, the cumulative percentage of variance explained by the first two components is 83%. This only increases by an additional 7% with the addition of a third component. It is deemed that the extra 7% is not worth reducing the parsimony or simplicity of the model.

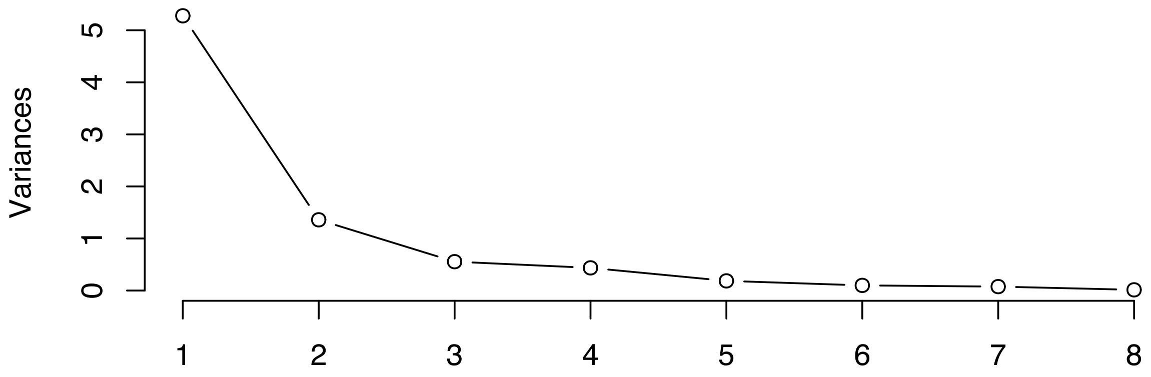


Figure 6: Screeplot of components produced by 'prcomp'

(f) Relation between observations and each of the components

PC scores show where, on the dimensions of the principal components, each of the observations lie. When examined in conjunction with the biplot, it can illustrate how each observation is explained by each component.

[8] Kassambara (2017).

To take one stand-out observation as an example, the Mercedes 190E, from the biplot, it can be seen that it is located far out in the upper right quadrant. The score output from `prcomp` for this observation confirms this:

```
        PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8
 4.92111907  3.99060930  2.57401510 -1.99417000  0.61978746  0.08807697  0.06370263  0.39901858
```

being located 4.9 along the PC1 axis and 4 along the PC2 axis.

(g) Cars by origin

The library `factoextra` was used to create a screeplot and correlation circle of the dataset at Figure 7.[9]

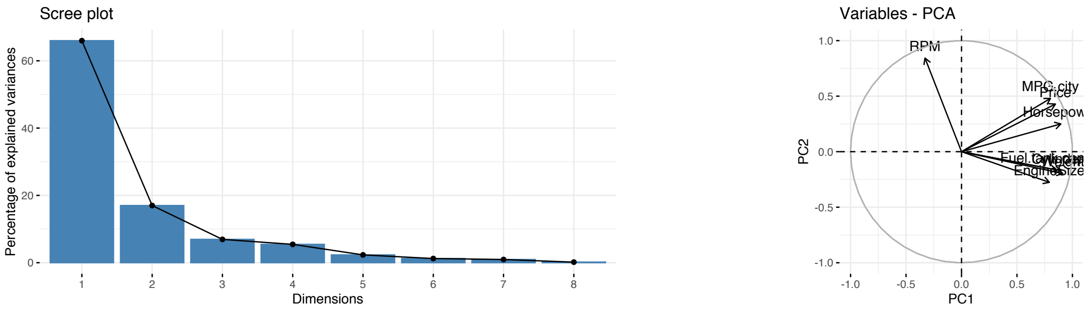


Figure 7: Screeplot and correlation circle from factoextra

Figure 8 shows a biplot representing the cars by origin: USA and non-USA.

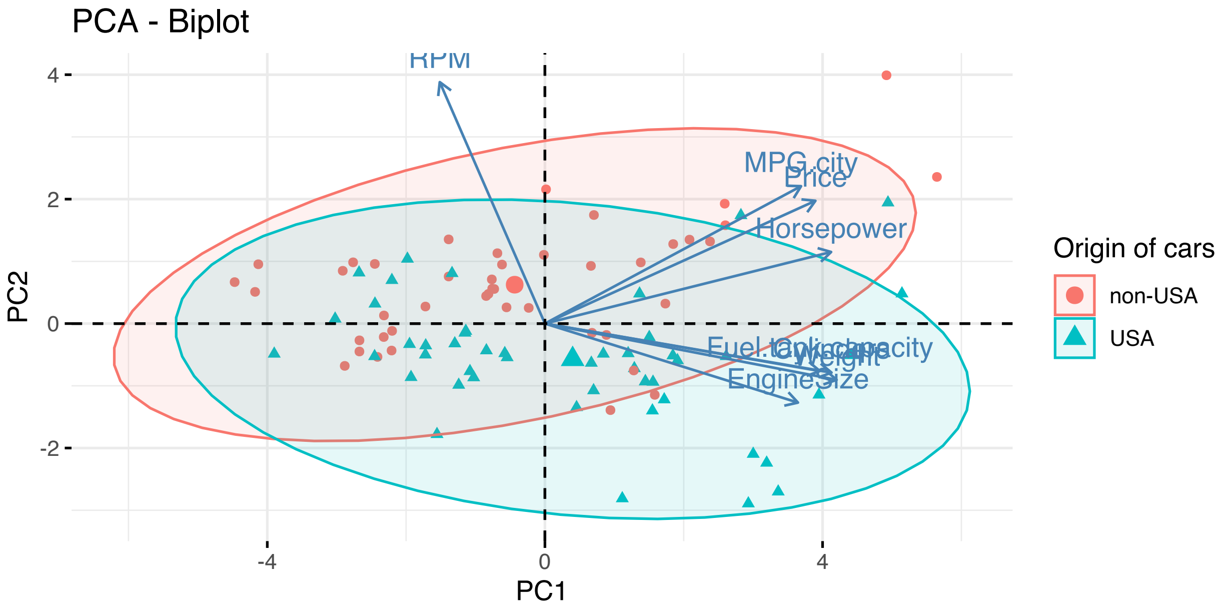


Figure 8: Biplot showing cars by origin

What is striking about this plot, is how the split of the dataset into two groups (USA and non-USA) conforms to the patterns and clustering we saw earlier in the biplot.

USA cars generally conform to the variability represented along the engine size, weight and fuel tank capacity eigenvectors, whereas non-USA cars to that along the MPG, horsepower and price eigenvectors. RPM is marginally more cosely aligned with non-USA cars than with USA cars as an explanatory factor.

[9] Kassambara (2017).

Table 11: USA Information and variance explained by components (as calculated by ‘prcomp‘)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.375	1.091	0.714	0.552	0.415	0.319	0.250	0.134
Proportion of Variance	0.705	0.149	0.064	0.038	0.022	0.013	0.008	0.002
Cumulative Proportion	0.705	0.854	0.918	0.956	0.977	0.990	0.998	1.000

Second analysis

In the second analysis, the dataset is split into two groups comprising:

- USA cars, and;
- non-USA cars and, the PCA is repeated using only those cars from the USA.

A summary of the second PCA, using cars from the USA only, is provided in the following tables and plots.

(a_2) Eigenvalues and eigenvectors

The eigenvalues for the dataset containing only the USA cars is below:

```
[1] 5.64155799 1.19045200 0.50920414 0.30450464 0.17235207 0.10163958 0.06230889 0.01798070
```

Table 9: USA Eigenvectors as calculated by `prcomp`.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Price	0.37	-0.26	0.44	0.18	-0.21	0.10	-0.06	-0.71
EngineSize	0.37	0.03	-0.16	-0.70	-0.52	-0.06	-0.26	0.08
Fuel.tank.capacity	0.35	0.32	-0.50	0.37	-0.34	-0.08	0.51	-0.11
RPM	-0.17	-0.77	-0.49	0.04	-0.13	0.35	0.05	-0.01
Weight	0.39	0.20	-0.25	0.33	0.21	0.41	-0.64	0.11
MPG.city	0.37	-0.31	0.41	0.28	-0.21	-0.02	0.14	0.68
Cylinders	0.39	0.01	0.04	-0.40	0.51	0.45	0.47	0.00
Horsepower	0.37	-0.33	-0.24	0.00	0.45	-0.70	-0.10	-0.06

Table 10: USA Eigenvalues and percentage of variance from FactoMineR

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	5.642	70.519	70.519
comp 2	1.190	14.881	85.400
comp 3	0.509	6.365	91.765
comp 4	0.305	3.806	95.571
comp 5	0.172	2.154	97.726
comp 6	0.102	1.270	98.996
comp 7	0.062	0.779	99.775
comp 8	0.018	0.225	100.000

(b) Amount of information explained by components

(c) Factor loadings

Table 12: USA Scores of the rotated data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Buick Century	-1.268	-0.012	-0.131	-0.032	-0.720	-0.069	-0.143	-0.018
Buick LeSabre	1.065	0.207	-0.111	-0.266	0.070	-0.018	-0.147	-0.062
Buick Roadmaster	2.696	1.680	-0.287	0.670	-0.252	-0.258	-0.173	-0.084
Buick Riviera	1.639	-0.061	0.292	0.122	-0.274	0.039	-0.010	-0.205
Cadillac DeVille	3.877	0.009	1.659	-0.780	-0.081	0.017	0.100	-0.016

(d) Correlation between original variables and principal components

Table 13: USA Correlation matrix between variables and PCs produced by FacotMineR

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Price	0.890	0.288	-0.313	0.098	0.088
EngineSize	0.887	-0.028	0.113	-0.385	0.215
Fuel.tank.capacity	0.821	-0.345	0.358	0.204	0.142
RPM	-0.401	0.838	0.348	0.021	0.054
Weight	0.915	-0.218	0.176	0.183	-0.087
MPG.city	0.872	0.333	-0.295	0.153	0.087
Cylinders	0.934	-0.016	-0.032	-0.218	-0.213
Horsepower	0.871	0.356	0.172	-0.002	-0.186

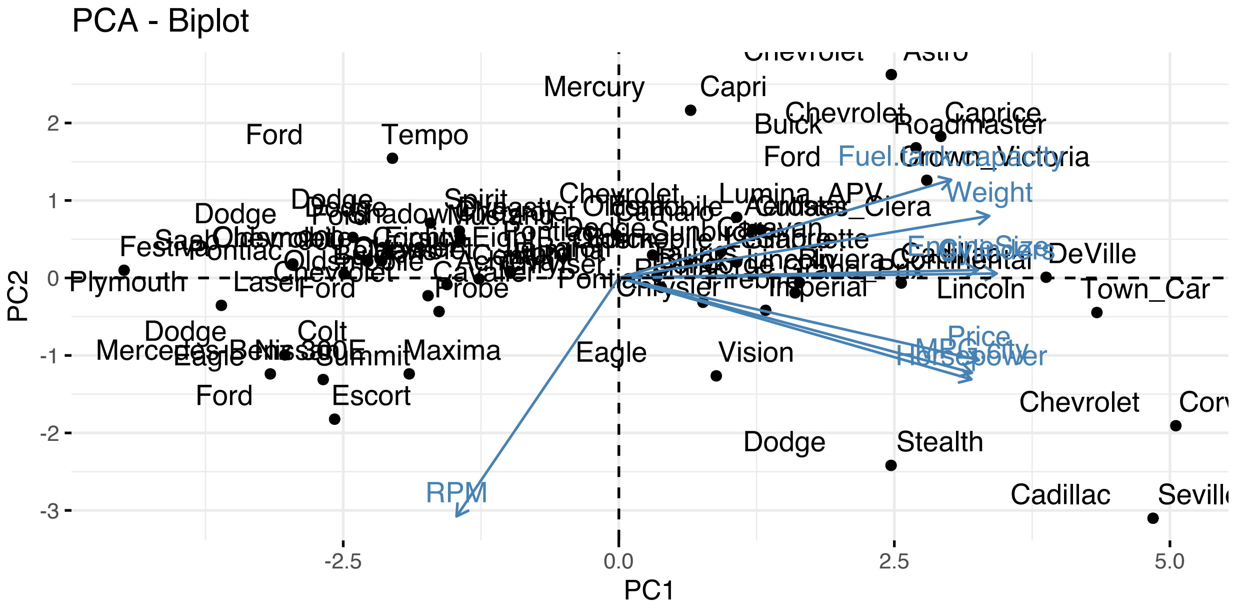


Figure 9: USA Biplot of the data

Table 14: USA Cos^2 of variables from `FactoMineR`

	PC1	PC2	PC3	PC4	PC5
Price	0.791	0.083	0.098	0.010	0.008
EngineSize	0.787	0.001	0.013	0.148	0.046
Fuel.tank.capacity	0.674	0.119	0.128	0.042	0.020
RPM	0.161	0.702	0.121	0.000	0.003
Weight	0.837	0.047	0.031	0.033	0.008
MPG.city	0.761	0.111	0.087	0.023	0.008
Cylinders	0.872	0.000	0.001	0.048	0.045
Horsepower	0.759	0.127	0.030	0.000	0.035

(e) How many components to retain?

As before, it would seem that retaining 2 components would be optimal, given the look of the screeplot and that they account for over 85% of the cumulative percentage of variance.

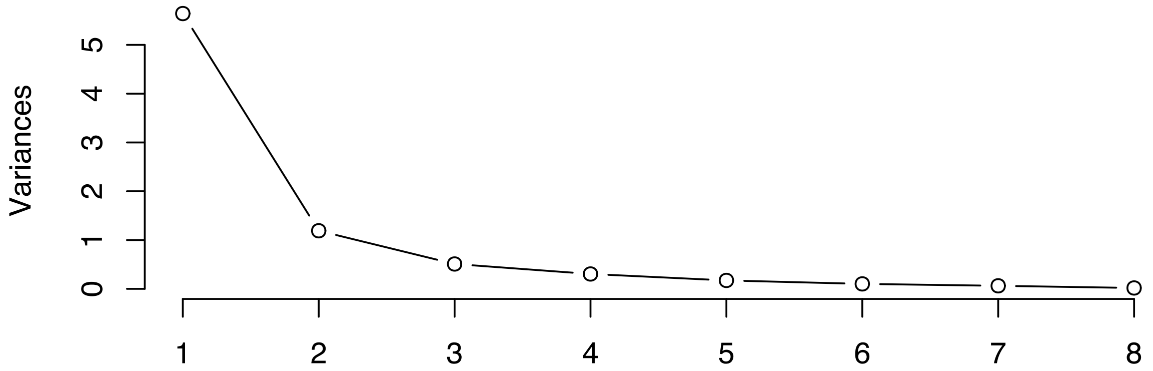


Figure 10: USA Screeplot of components produced by ‘prcomp‘

(f) Relation between observations and each of the components

To take one stand-out observation as an example, the Ford Crown Victoria, from the biplot, it can be seen that it is located in the upper right quadrant.

```
        PC1          PC2          PC3          PC4          PC5          PC6          PC7          PC8
 2.794205886  1.260599789 -0.301910420 -0.935378003  0.629228205  0.272281273  0.006411545  0.016631108
```

Conclusion

The cumulative percentage of variance accounted for by 2 components is about 2% higher in respect of the dataset comprising only USA cars, as opposed to the full dataset.

When the data is limited to USA cars only, the cylinders and engine size variables become almost completely parallel to the first component, whereas with the full dataset, they had been at a greater positive angle from that axis. Morevoer, the RPM loading is also much closer to the axis of the second component when the dataset is restricted to USA cars only.

The cos^2 of the variables for the USA only dataset is also slightly higher in respect of the first two components when compared to the full dataset.

In summary, it can be stated that the first two components 'fit' the data better when the dataset is limited to USA cars only compared to the full dataset.

Declaration

All work contained in the submission is the student's own except for others work which is clearly referenced.

References

Hartmann, K., J. Krois, and B. Waske. 2018. Berlin: E-Learning Project SOGA: Statistics; Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin.

Janert, Philipp K. 2011. *Data Analysis with Open Source Tools: A Hands-on Guide for Programmers and Data Scientists*. 1. ed. Bejing: O'Reilly.

Kassambara, Alboukadel. 2017. *Practical Guide to Principal Component Methods in R*. Edition 1. United States}: CreateSpace Independent Publishing Platform.

———. 2019. *Ggcorrplot: Visualization of a Correlation Matrix Using 'Ggplot2'*. https://CRAN.R-project.org/package=ggcorrplot.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, and Dominique Makowski. 2020. "Extracting, Computing and Exploring the Parameters of Statistical Models Using R." *Journal of Open Source Software* 5 (53): 2445. https://doi.org/10.21105/joss.02445.