Wojciech Artichowicz JDSZ1

**mtcars** dataset analysis for linear regression

The dataset contains the following information on car properties:

|  |  |
| --- | --- |
| mpg | Miles/(US) gallon |
| cyl | Number of cylinders |
| disp | Displacement (cu.in.) |
| hp | Gross horsepower |
| drat | Rear axle ratio |
| wt | Weight (1000 lbs) |
| qsec | 1/4 mile time |
| vs | V/S |
| am | Transmission (0 = automatic, 1 = manual) |
| gear | Number of forward gears |
| carb | Number of carburetors |

Entities (cars) that appeared in the dataset are:

|  |  |  |
| --- | --- | --- |
| Mazda RX4 | Merc 450SE | AMC Javelin |
| Mazda RX4 Wag | Merc 450SL | Camaro Z28 |
| Datsun 710 | Merc 450SLC | Pontiac Firebird |
| Hornet 4 Drive | Cadillac Fleetwood | Fiat X1-9 |
| Hornet Sportabout | Lincoln Continental | Porsche 914-2 |
| Valiant | Chrysler Imperial | Lotus Europa |
| Duster 360 | Fiat 128 | Ford Pantera L |
| Merc 240D | Honda Civic | Ferrari Dino |
| Merc 230 | Toyota Corolla | Maserati Bora |
| Merc 280 | Toyota Corona | Volvo 142E |
| Merc 280C | Dodge Challenger |  |

Dataset contains 32 records. One record per car (which are identical to abovementioned unique entities).

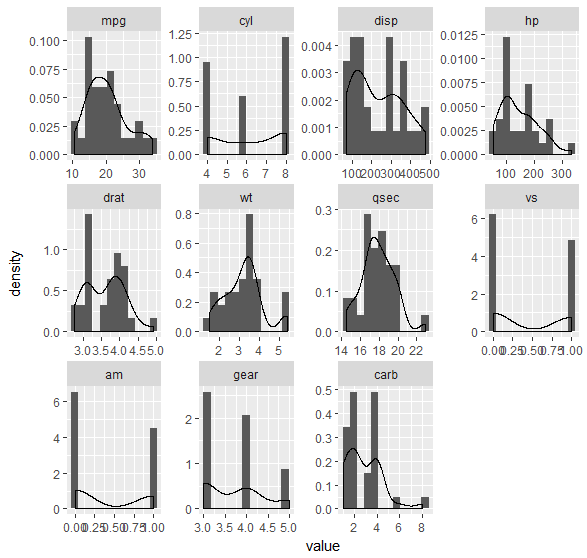


Fig.1. Distribution plots of dataset variables (exploratory analysis)

Distributions of the variables are depicted in Fig. 1. Basing on the Fig. 1. it can be stated that variables *cy*l, *vs*, *am*, *gear*, *carb* can be treated as ordinal or nominal variables. They should be taken into special consideration during clustering procedure and pattern identifications.

In order to determine the possible linear correlations between pairs of variables the correlation matrix has been constructed. It is presented below in Tab. 1.

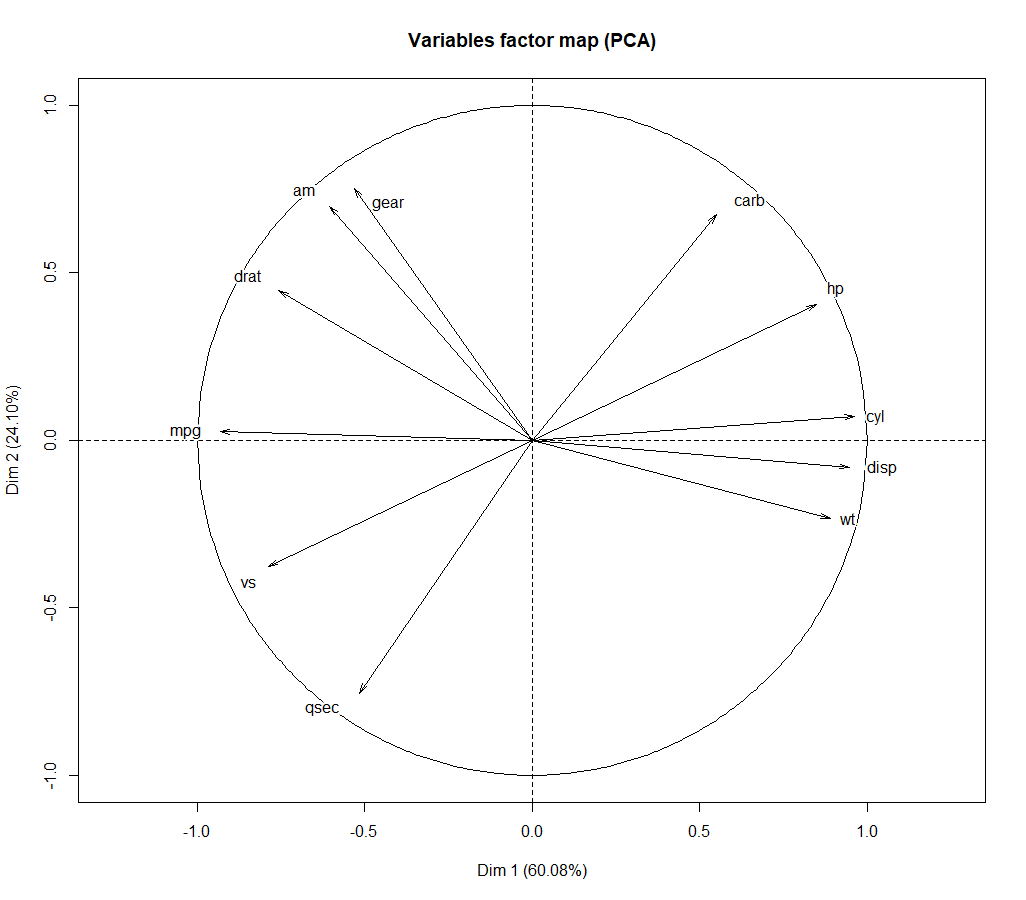
Tab. 1. Correlation matrix of the *mtcars* dataset

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **mpg** | **cyl** | **disp** | **hp** | **drat** | **wt** | **qsec** | **vs** | **am** | **gear** | **carb** |
| **mpg** | 1 | **-0.85** | **-0.85** | **-0.78** | **0.681** | **-0.87** | 0.419 | **0.664** | 0.6 | 0.48 | -0.55 |
| **cyl** |  | 1 | **0.902** | **0.832** | **-0.7** | **0.782** | -0.59 | **-0.81** | -0.52 | -0.49 | 0.527 |
| **disp** |  |  | 1 | **0.791** | **-0.71** | **0.888** | -0.43 | **-0.71** | -0.59 | -0.56 | 0.395 |
| **hp** |  |  |  | 1 | -0.45 | **0.659** | **-0.71** | **-0.72** | -0.24 | -0.13 | **0.75** |
| **drat** |  |  |  |  | 1 | **-0.71** | 0.091 | 0.44 | **0.713** | **0.7** | -0.09 |
| **wt** |  |  |  |  |  | 1 | -0.18 | -0.56 | **-0.69** | -0.58 | 0.428 |
| **qsec** |  |  |  |  |  |  | 1 | **0.745** | -0.23 | -0.21 | **-0.66** |
| **vs** |  |  |  |  |  |  |  | 1 | 0.168 | 0.206 | -0.57 |
| **am** |  |  |  |  |  |  |  |  | 1 | **0.794** | 0.058 |
| **gear** |  |  |  |  |  |  |  |  |  | 1 | 0.274 |
| **carb** |  |  |  |  |  |  |  |  |  |  | 1 |

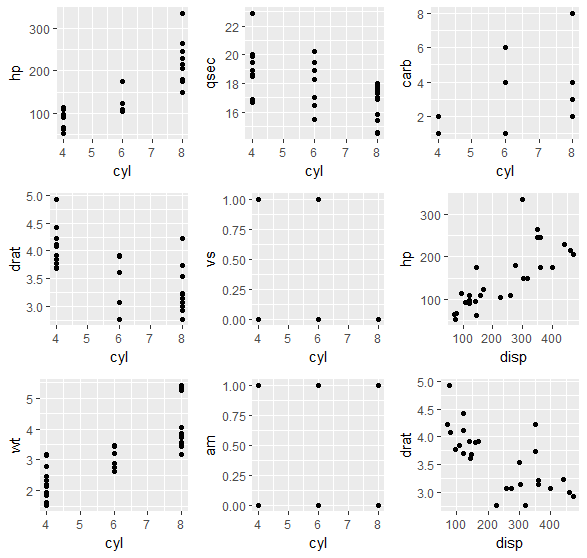
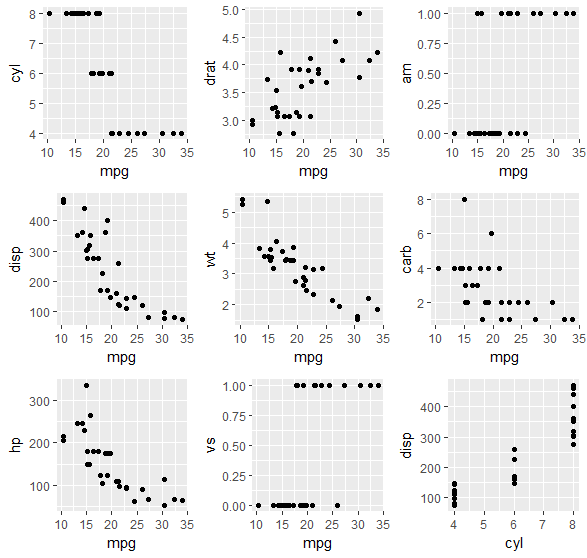
In Fig. 2. a scatterplot of all variables pairs is presented. Further graphical exploration of the data has been driven in order to choose the proper variables from the viewpoint of this exercise.

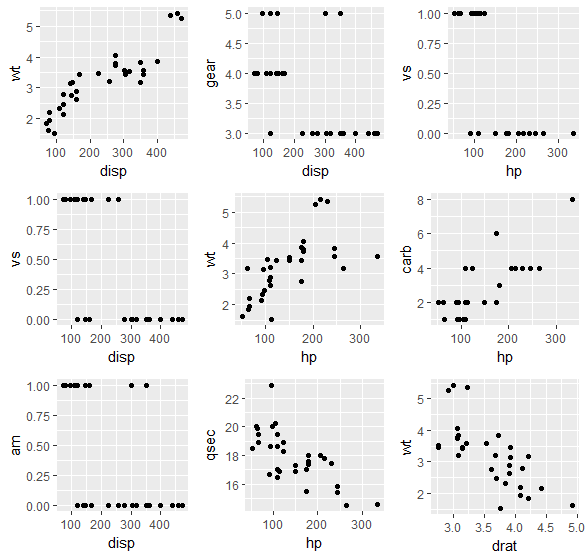
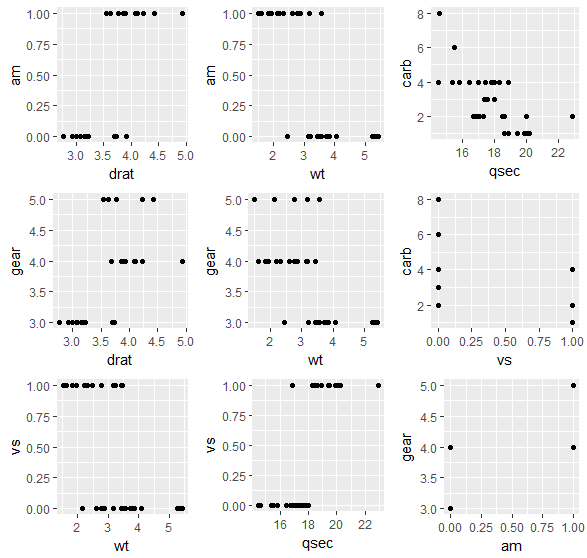


Fig.2. Scatter plots of dataset variables (exploratory analysis)



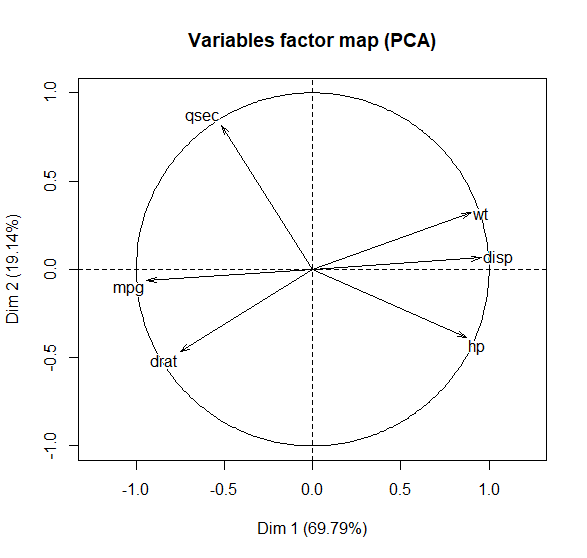
The brief look at linear correlations can be taken from PCA circle plot. Based





PCA of the reduced dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | Eigenvalue | Percentage of variance | Cumulative percentage of variance |
| PCA1 | 4.18739648 | 69.7899413 | 69.78994 |
| PCA2 | 1.14811212 | 19.135202 | 88.92514 |
| PCA3 | 0.33335666 | 5.5559444 | 94.48109 |
| PCA4 | 0.15436054 | 2.5726757 | 97.05376 |
| PCA5 | 0.12479601 | 2.0799335 | 99.1337 |
| PCA6 | 0.05197818 | 0.8663031 | 100 |



kPCA