

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

Lesson 7: How to Perform PCA in R

In this section, you will try a PCA using a simple and easily understood data set. You will use the mtcars data set, which is built into R. This data set consists of data on 32 models of cars, taken from an American motoring magazine (Motor Trend, 1974). For each car, there are 11 features, expressed in varying units (U.S. units). They are as follows:

- mpg: Fuel consumption (miles per (US) gallon)—more powerful and heavier cars tend to consume more fuel
- cyl: Number of cylinders—more powerful cars often have more cylinders
- disp: Displacement (cu.in.) is the combined volume of the engine's cylinders.
- hp: Gross horsepower is a measure of the power generated by the car.
- drat: Rear axle ratio describes how a turn of the drive shaft corresponds to a turn of the wheels. Higher values will decrease fuel efficiency.
- wt: Weight (1,000 lb) is self-explanatory!
- qsec: 1/4 mile time describes the car's speed and acceleration
- vs: Engine block—this denotes whether the vehicle's engine is shaped like a V or is a more common straight shape.
- am: Transmission—this denotes whether the car's transmission is automatic (0) or manual (1).
- gear: Number of forward gears—sports cars tend to have more gears.
- carb: Number of carburetors—this is associated with more powerful engines.

Note that the units used vary and occupy different scales.



Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

```
c(1:7,10,11)
```

Again, think of the `c` like it means the "container" of your data. If you simply type `mtcars` in R, it will pull up this data because it is built in. This command would give the subset of data you need:

```
mtcars[,c(1:7,10,11)]
```

Type that in R and examine the data:

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	4	2

A table displaying a small subset of mtcars.

To understand how the subsetting works, try typing in mtcars by itself in R and hit Enter. You might wonder why the comma is there. Whenever you subset data in R, you can do so by rows or columns. If you were to subset the data by rows, then that goes first. If you subset by columns, that will go second. So it is possible to do both!

c(subset of rows , subset of columns)



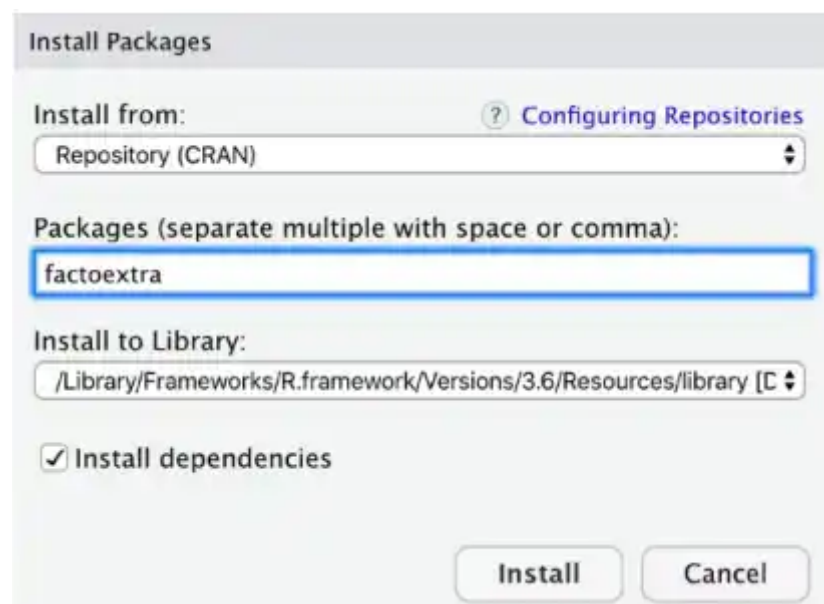
Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

In this code, you are running the PCA (`prcomp`) on the `mtcars` subset data. There are two options provided: `center`, and `scale`. You will learn much more about the details of PCA in a later course, so for now these options are not explored. Finally, the results are assigned to a designated name: `mtcars.pca`. You could name this differently if you like.

Now create a scree plot of this data. To do this, install a package: `factoextra`. If you have not installed it yet, you can run the Install Packages... under Tools in RStudio:



Installing the factoextra package within RStudio.

It is usually a good habit to select the "Install dependencies" when you install packages, as this will also install any other packages needed. Now that you have it installed, load it and then run the scree plot:

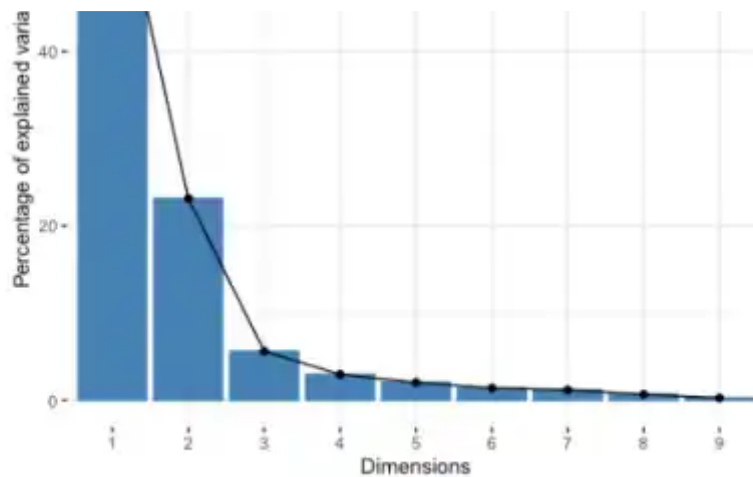


Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)



Scree plot.

To get the graph to give you an eigenvalue perspective, run this version of the code (including the `addlabels` option gives you the eigenvalue to easily find when it drops below 1!):

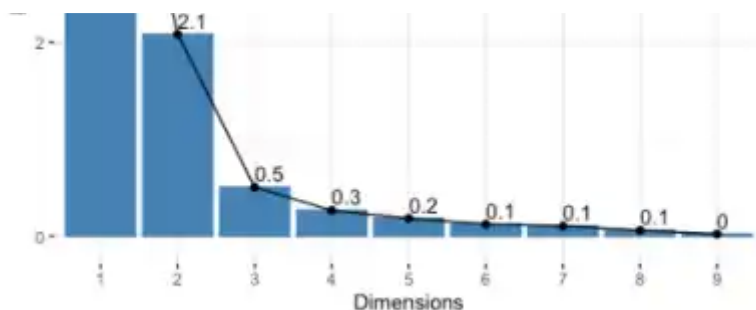
```
fviz_eig(mtcars.pca, choice = "eigenvalue", addlabels=TRUE)
```

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)



Scree plot with eigenvalues.

This graph indicates that there is no real benefit in using more than two components because all components beyond that have eigenvalues less than 1.

Finally, output the loadings for the components:

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

```

      PC7      PC8      PC9
mpg -0.38138068 -0.12465987  0.11492862
cyl -0.15893251  0.81032177  0.16266295
disp -0.18233095 -0.06416707 -0.66190812
hp   0.69620751 -0.16573993  0.25177306
drat  0.04767957  0.13505066  0.03809096
wt   -0.42777608 -0.19839375  0.56918844
qsec  0.27622581  0.35613350 -0.16873731
gear -0.08555707  0.31636479  0.04719694
carb -0.20604210 -0.10832772 -0.32045892

```

Table output of components.

You should only be interested in the first two components based on the scree plot (PC1 PC2). When you compare the values, you should be interested in which values are the highest for each one. For PC1, there is more weight for mpg, cyl, disp, hp, wt. For PC2, there is more weight for drat, qsec, gear, carb. You can save these combined variables (i.e., components) and use them instead of the original values. This will be explored further in a future class.

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

© 2020 Western Governors University – WGU. All rights reserved.

[Privacy Policy](#) | [Terms of Service](#)

[Honor Code](#)

