Final Project

Pavankumar Pendela

12/12/2021

Problem

This case study involved things like evaluating prices and determining which car specifications the buyer will receive. We will aim to obtain information for the cheapest, average, high, and most costly cars in this case study.

What are the specifications that clients will receive based on these four price ranges?

About dataset

Kaggle provided the data for this project. There are 205 rows and 26 columns in the car dataset. The dataset provides information on all automotive specifications, including body style, driving wheel, engine position, wheel-base, length, width, and height, as well as a few more.

Let's look at dataset details

symboling	normalized	make	fuel-type	aspiration	num-of-d	body-sty	drive-wheels	engine-locati	wheel-base	length	width	height	curb-weig	engine-type	num-of-cy	engine-siz	fuel-syst	bore	stroke	compres	horsepov	peak-rpn	city-mpg	highway-mpg	price
3	?	alfa-rom	gas	std	two	convertib	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	27	13495
3	?	alfa-rom	gas	std	two	convertib	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	27	16500
1	?	alfa-rom	gas	std	two	hatchbac	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152	mpfi	2.68	3.47	9	154	5000	19	26	16500
2	164	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109	mpfi	3.19	3.4	10	102	5500	24	30	13950
2	164	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136	mpfi	3.19	3.4	8	115	5500	18	22	17450
2	?	audi	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	15250
1	158	audi	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	17710
1	?	audi	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	18920
1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131	mpfi	3.13	3.4	8.3	140	5500	17	20	23875
0	?	audi	gas	turbo	two	hatchbac	4wd	front	99.5	178.2	67.9	52	3053	ohc	five	131	mpfi	3.13	3.4	7	160	5500	16	22	?
2	192	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	29	16430
0	192	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	29	16925
0	188	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164	mpfi	3.31	3.19	9	121	4250	21	28	20970
0	188	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc	six	164	mpfi	3.31	3.19	9	121	4250	21	28	21105
1	?	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3055	ohc	six	164	mpfi	3.31	3.19	9	121	4250	20	25	24565
0	?	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3230	ohc	six	209	mpfi	3.62	3.39	8	182	5400	16	22	30760
0	?	bmw	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc	six	209	mpfi	3.62	3.39	8	182	5400	16	22	41315
0	?	bmw	gas	std	four	sedan	rwd	front	110	197	70.9	56.3	3505	ohc	six	209	mpfi	3.62	3.39	8	182	5400	15	20	36880
2	121	chevrolet	gas	std	two	hatchbac	fwd	front	88.4	141.1	60.3	53.2	1488	l	three	61	2bbl	2.91	3.03	9.5	48	5100	47	53	5151
1	98	chevrolet	gas	std	two	hatchbac	fwd	front	94.5	155.9	63.6	52	1874	ohc	four	90	2bbl	3.03	3.11	9.6	70	5400	38	43	6295
0	81	chevrolet	gas	std	four	sedan	fwd	front	94.5	158.8	63.6	52	1909	ohc	four	90	2bbl	3.03	3.11	9.6	70	5400	38	43	6575
1	118	dodge	gas	std	two	hatchbac	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	9.41	68	5500	37	41	5572
1	118	dodge	gas	std	two	hatchbac	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	38	6377
1	118	dodge	gas	turbo	two	hatchbac	fwd	front	93.7	157.3	63.8	50.8	2128	ohc	four	98	mpfi	3.03	3.39	7.6	102	5500	24	30	7957
1	148	dodge	gas	std	four	hatchbac	fwd	front	93.7	157.3	63.8	50.6	1967	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	38	6229
1	148	dodge	gas	std	four	sedan	fwd	front	93.7	157.3	63.8	50.6	1989	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	38	6692

Here is the link for dataset https://www.kaggle.com/natigmamishov/eda-and-regression-on-automobile-dataset/data

Approach

I will attempt clustering the cars by specification. For this, I will drop the symbolling and normalized-losses variables from the data set and work with just the specification variables which are both categorical and numerical.

One of the most widely used clustering algorithms is the K-means approach. Simply put, K is a set of clusters into which the data can be separated based on their attribute similarities and differences. Each cluster has a core that is more similar to nearby observatories. The amount of similarities to be used to cluster can be expressed as a distance between two observations in K-means clustering. The calculation is then done using this distance to determine which cluster each member of the observation belongs to. New cluster centers are determined with each fresh observation, and new observations are assigned to the appropriate cluster.

Why K-means

The simplest is K-means. To put in place and run. All you have to do is select "k" and run it several times. Most smart algorithms are far more difficult to develop and have a lot more parameters to set. Furthermore, the majority of people do not require high-quality clusters. They are content with anything that can be done remotely for them. Plus, when they have more complex clusters, they aren't sure what to do. They require K-means, which models clusters using the simplest model ever - a centroid: a huge reduction of data to centroids.

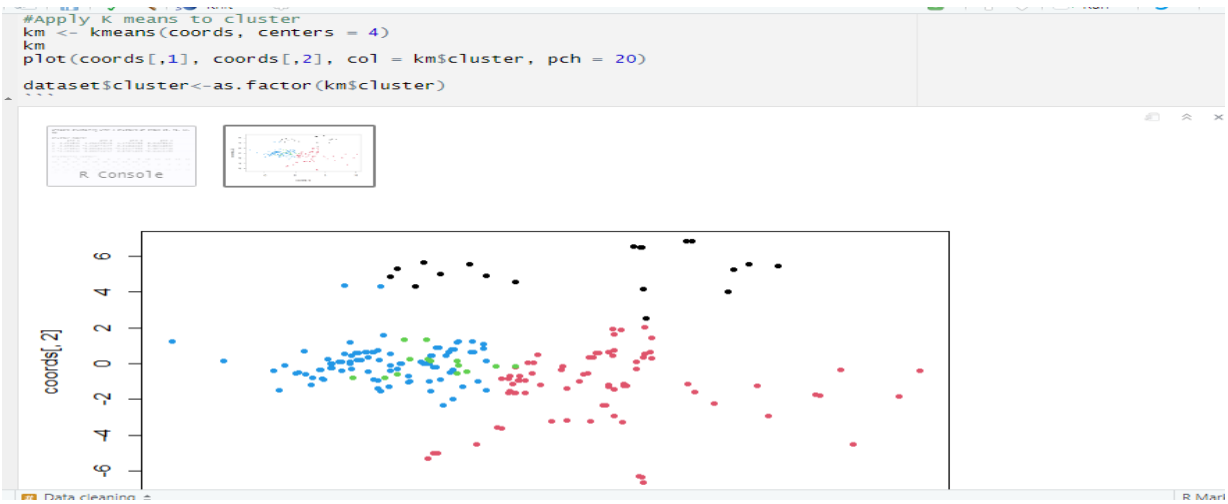
I decided to use the following steps to get the results

- **Run the required libraries**
- **Import the data**
- **Data cleaning**
- **Factorial analysis**
- **Clustering**

The following are the conclusions I reached based on my research.

By using Factorial analysis of mixed data (PCAmix) to analyse a data table where observations are described both by quantitative variables and qualitative variables method I choose 4 centres's to apply K-means algorithm.

I found the results accordingly



By using these analyses the four clusters results are accordingly.

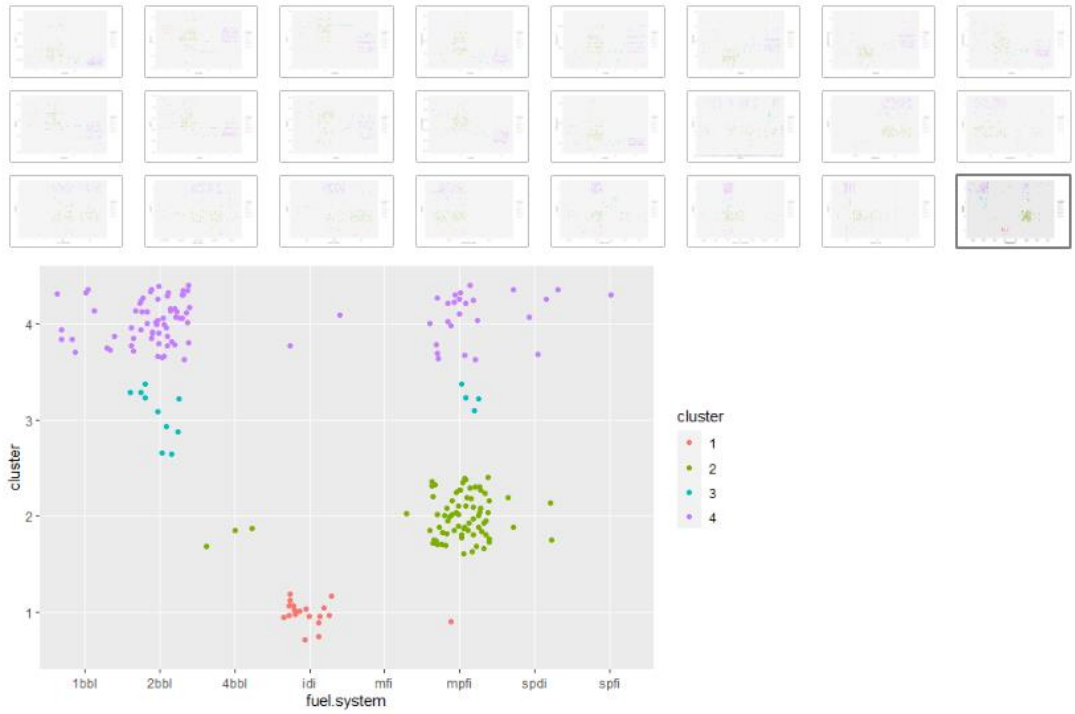
Cluster 1: Cheapest, most economical, low horsepower, small, small engine cars Cluster

Cluster 2: Standard average priced cars with average specs across the board

Cluster 3: Mid high price, diesel, 4 door sedan/wagons, pretty big & heave in dimension with low RPM

Cluster 4: Most expensive, biggest horsepower, biggest engine cars Cluster

The results below are based on clustering techniques for each specification individually.



Conclusion

People can choose their cars directly based on the findings, according to their desired features and budget.

Companies frequently advise their consumers on which characteristics they should obtain based on their budget.