FOUNDATION FOR ORGANISATIONAL RESEARCH AND EDUCATION

NEW DELHI

Academic Session 2023-2025

Project-1

Vehicle/Car Sales Trends and Pricing Insights

Machine Learning for Managers

FMG 32 Section A

Submitted to:

Prof. Amarnath Mitra

Submitted by:

321032 – Nisha Arora

1. Project Objectives

1.1 Segmentation of Consumer Data using Unsupervised Machine Learning Clustering Algorithms like K-Means clustering

1.2 Number of appropriate clusters using performance matrix Silhouette score

1.3 To determine the segment and Characteristics of each cluster (to sell the product/service)

2. Description of Data

2.1. Data Source, Size, Shape

2.1.1. Data Source –
https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

2.1.2. Data Size (in KB | MB | GB …) – **88 MB**

2.1.3. Data Shape | Dimension:
Number of Variables - **16**
Number of Records – **558837**

2.2. Description of Variables

2.2.1. Index Variable(s): Car Id

2.2.2. Variables or Features having Categories | Categorical Variables or Features (CV)

2.2.2.1. Variables or Features having Nominal Categories | Categorical Variables or Features - **Nominal Type**:
make, model, trim, body, transmission, state, colour, interior, seller

2.2.2.2. Variables or Features having Ordinal Categories | Categorical Variables or Features - **Ordinal Type:** Condition

2.2.3. Non-Categorical Variables or Features: vin, odometer, mmr, selling price, sale date

Car ID: Unique identifier for each car
Year: Numeric representation of manufacturing year
Make: Brand or manufacturer of the car
Model: Specific model name of the car
Trim: Variant or version of the model
Body: Type of body style (e.g., sedan, SUV)
Transmission: Type of transmission system (e.g., automatic, manual)
VIN: Vehicle Identification Number, unique to each car
State: State where the car is located
Condition: Condition of the car, possibly ordinal categorical data
Odometer: Numeric representation of mileage
Color: Color of the car
Interior: Color or material of the interior
Seller: Entity selling the car
MMR: Market value of the car, likely non-categorical data
Selling Price: Price at which the car is sold
Sale Date: Date and time of sale

2.3. Descriptive Statistics

2.3.1. Descriptive Statistics: Categorical Variables or Features

2.3.1.1. Count | Frequency Statistics

Color

Row ID	count
black	22203
white	21649
silver	16729
gray	16352
blue	10163

Model

Row ID	count
Altima	6063
F-150	2992
Fusion	2604
Camry	2460
Escape	2247

Make

Row ID	count
Ford	20837
Chevrolet	12069
Nissan	10809
Toyota	8033
Dodge	6191

Body

Row ID	count
Sedan	42596
SUV	23537
sedan	8328
suv	4968
Minivan	4348

Transmission

Row ID	count
Sedan	2
automatic	108246
manual	3514
sedan	5

2.3.2. Descriptive Statistics: Non-Categorical Variables or Features

2.3.2.1. Measures of Central Tendency

Row ID	Column	Min	Max	Mean	Std. devi...	Variance	Skewness	Kurtosis	Overall s...	No. missi...	No. NaNs	No. +oos	No. -oos	Median	Row count	Hi
condition	condition	1	49	30.574	13.314	177.254	-0.83	-0.197	3,417,183.716	0	0	0	0	?	111767	1
odometer	odometer	1	999,999	68,363.626	53,249.21	2,835,478,413....	1.802	12.954	7,640,797,387....	0	0	0	0	?	111767	1
mmr	mmr	25	178,000	13,782.935	9,718.146	94,442,361.104	2.026	11.693	1,540,477,346....	0	0	0	0	?	111767	25
sellingprice	sellingprice	1	171,500	13,626.721	9,787.374	95,792,682.357	1.959	10.783	1,523,017,736....	0	0	0	0	?	111767	1

2.3.2.2. Measures of Dispersion

Statistics

Rows: 4 | Columns: 12

Name	Type	# Missing val...	# Unique val...	Minimum	Maximum	25% Quantile	50% Quantile...	75% Quantile	Standard
condition	Number (dou...	0	42	1	49	24	34	41	13.314
odometer	Number (dou...	0	78138	1	999,999	28,408	52,407	99,088	53,249.21
mmr	Number (dou...	0	1066	25	178,000	7,100	12,250	18,350	9,718.146
sellingprice	Number (dou...	0	1222	1	171,500	6,900	12,100	18,250	9,787.374

Source of data-

https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

3. Analysis of Data

3.1. Data Pre-Processing

3.1.1. Missing Data Statistics and Treatment

3.1.1.1.1. Missing Data Statistics: 16

3.1.1.1.2. Missing Data Treatment: make, model, trim, body, transmission, state, colour, interior, seller, condition, vin, odometer, mmr, selling price, sale date

3.1.1.1.2.1. Removal of Records with More Than 50% Missing Data

3.1.1.2.1. Missing Data Statistics: Categorical Variables or Features

Name	# Missing values
year	0
make	2141
model	2170
trim	2203
body	2688
transmission	13241
state	0
color	163
interior	163
seller	0

3.1.1.2.2. Missing Data Treatment: Categorical Variables or Features - 10

3.1.1.2.2.1. Removal of Variables or Features with More Than 50% Missing Data: make, model, trim, body, transmission, state, colour, interior, seller, condition

3.1.1.2.2.2. Imputation of Missing Data using Descriptive Statistics: Mode

3.1.1.3.1. Missing Data Statistics: Non-Categorical Variables or Features

Name	# Missing values
vin	2
condition	2342
odometer	21
mmr	9
sellingprice	2
saledate	2

3.1.1.3.2. Missing Data Treatment: Non-Categorical Variables or Features - 6
3.1.1.3.2.1. Removal of Variables or Features with More Than 50% Missing Data: vin, odometer, mmr, selling price, sale date
3.1.1.3.2.2. Imputation of Missing Data using Descriptive Statistics: Mean

3.1.2. Numerical Encoding of Categorical Variables or Features (Encoding Schema - Alphanumeric Order)
- In this case, category to number node will be used to encode the categorical variables.

Color-
8 – black
9 – blue
14 – gray
22 – silver
24 – white

Model
30-Altima
91- F-150
90- Fusion
62- Camry
75- Escape

Make
19-Ford
0-Chevrolet
5-Nissan
17-Toyota
22-Dodge

Body
0- Sedan
1- SUV
28-sedan
44-suv
9-Minivan

Transmission

0 – Sedan
1- Automatic
2 – Manual
3 – sedan

3.1.3. Outlier Statistics and Treatment (Scaling | Transformation)

3.1.3.1.1. Outlier Statistics: Non-Categorical Variables or Features

Row ID	Outlier ...	Membe...	Outlier ...	Lower ...	Upper ...
Row0	condition	111767	0	-1.5	66.5
Row1	odometer	111767	2066	-77,611	205,105
Row2	mmr	111767	3244	-9,775	35,225
Row3	sellingprice	111767	3222	-10,125	35,275

3.1.3.1.2. Outlier Treatment: Non-Categorical Variables or Features

3.1.3.1.2.1. Standardization

3.1.3.1.2.2. Normalization using Min-Max Scaler:

Min-max normalization, also known as feature scaling, is a technique used in data preprocessing to scale numerical features to a specific range, typically between 0 and 1.
The formula for min-max normalization is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

3.1.3.1.2.3. Log Transformation

3.1.4. Data Bifurcation: Training & Testing Sets
The training and testing data have been bifurcated into 70% and 30% respectively.

3.2 Data Analysis
3.2.1 Unsupervised Machine Learning Algorithm

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into a predefined number of non-overlapping clusters. The algorithm aims to group data points into clusters in such a way that the similarity (or distance) between data points within the same cluster is maximized, while the similarity between data points in different clusters is minimized. In this project, K-means will be the clustering algorithm used for unsupervised learning. The metrics used in k-means is Euclidean distance.

K=2 (This represents the total number of clusters that will be formed are 2)

Row ID	year	Car id (...	make (t...	model (...	trim (to...	body (t...	transmi...	state (t...	color (t...	interior ...	seller (t...	condition	odometer	mmr	sellingp...
cluster_0	2,009.797	27,931.815	14.638	137.198	127.906	2.74	0.031	11.869	3.649	1.402	948.234	-0.029	0.037	-0.051	-0.057
cluster_1	2,010.32	83,818.236	14.212	134.301	132.128	15.118	0.032	12.436	3.582	1.303	1,361.663	0.034	-0.046	0.059	0.064

K=3(This represents the total number of clusters that will be formed are 3)

Row ID	year	Car id (...	make (t...	model (...	trim (to...	body (t...	transmi...	state (t...	color (t...	interior ...	seller (t...	condition	odometer	mmr	sellingp...
cluster_0	2,009.555	18,610.21	14.912	141.186	130.099	2.7	0.033	11.769	3.653	1.433	926.687	-0.055	0.083	-0.086	-0.096
cluster_1	2,010.104	55,871.868	14.289	132.26	127.484	2.935	0.03	11.878	3.618	1.349	1,125.701	0.018	-0.017	0.005	0.018
cluster_2	2,010.517	93,093.027	14.074	133.79	132.453	21.12	0.031	12.809	3.575	1.275	1,411.948	0.044	-0.079	0.093	0.088

K=4 (This represents the total number of clusters that will be formed are 4)

Row ID	year	Car id (...	make (t...	model (...	trim (to...	body (t...	transmi...	state (t...	color (t...	interior ...	seller (t...	condition	odometer	mmr	sellingp...
cluster_0	2,009.458	14,069.339	15.088	142.801	131.521	2.691	0.032	11.613	3.638	1.451	856.947	-0.067	0.1	-0.104	-0.117
cluster_1	2,010.097	42,135.444	14.19	131.809	124.457	2.791	0.03	12.122	3.66	1.357	1,046.45	0.004	-0.019	-0.004	-0.002
cluster_2	2,010.135	70,014.327	14.307	134.287	130.679	3.032	0.033	11.972	3.6	1.33	1,257.999	0.019	-0.018	0.024	0.038
cluster_3	2,010.55	97,796.725	14.107	134.024	133.402	27.266	0.03	12.908	3.563	1.27	1,462.085	0.053	-0.083	0.101	0.096

K=5 (This represents the total number of clusters that will be formed are 5)

Row ID	year	Car id (...	make (t...	model (...	trim (to...	body (t...	transmi...	state (t...	color (t...	interior ...	seller (t...	condition	odometer	mmr	sellingp...
cluster_0	2,009.393	11,216.451	15.092	142.808	130.496	2.678	0.032	11.601	3.619	1.467	807.013	-0.075	0.11	-0.12	-0.136
cluster_1	2,010.059	33,528.515	14.324	134.162	126.374	2.709	0.032	11.974	3.704	1.352	1,054.256	-0.001	-0.02	0.005	0.003
cluster_2	2,009.961	55,943.041	14.494	134.536	128.512	3.011	0.03	11.879	3.596	1.368	1,153.602	0.003	0.025	-0.03	-0.016
cluster_3	2,010.507	78,397.787	14.015	131.271	130.314	3.13	0.031	12.243	3.602	1.283	1,204.597	0.04	-0.094	0.098	0.097
cluster_4	2,010.377	100,637.457	14.196	135.956	134.408	33.356	0.032	13.075	3.555	1.29	1,559.081	0.045	-0.044	0.068	0.07

3.2.2 Clustering Model Performance Evaluation

The silhouette score is a metric used to evaluate the quality of clustering in unsupervised learning. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A silhouette score ranges from -1 to 1, where a higher score indicates better clustering:

- Silhouette Score of 1 indicates that clusters are well-separated.
- Silhouette Score of 0 indicates overlapping clusters.
- Silhouette Score close to -1 indicates that samples have been assigned to the wrong clusters.

K= 2

Row ID	Mean Si...
cluster_0	0.23
cluster_1	-0.208
Overall	0.011

K=3

Row ID	Mean Si...
cluster_0	0.622
cluster_1	0.509
cluster_2	0.614
Overall	0.582

K= 4

Row ID	Mean Si...
cluster_0	0.62
cluster_1	0.502
cluster_2	0.498
cluster_3	0.604
Overall	0.556

K= 5

Row ID	Mean Si...
cluster_0	0.617
cluster_1	0.497
cluster_2	0.495
cluster_3	0.491
cluster_4	0.595
Overall	0.539

Since the overall value of Mean Silhouette Coefficient of K=3 is maximum and closest to 1 , so this will be choice for us. We will go with total of three number of clusters.

Now for K=3, Clusters 0,1 and 2 have various different characteristics.

Cluster 0

Name of Variable	Characteristics
year	2012
make	Ford
model	Altima
trim	Base
body	Sedan
transmission	automatic
state	fl
condition	19
color	black
interior	black
seller	ford motor credit company

Cluster 1

Name of Variable	Characteristics
year	2012
make	Honda
model	Camry
trim	LE
body	SUV
transmission	automatic
state	ca
condition	19
color	black
interior	black
seller	ford motor credit company

Cluster 2

Name of Variable	Characteristics

year	2013
make	Chevrolet
model	F-150
trim	Base
body	Sedan
transmission	automatic
state	fl
condition	21
color	gray
interior	black
seller	the hertz corporation

3.2.3 Cluster Analysis using Base Model as K-Means

3.2.3.1 Cluster Analysis with Categorical Variables

The Kruskal-Wallis test is a non-parametric statistical test used to determine whether there are statistically significant differences between the medians of two or more independent groups. The test is appropriate when the data do not meet the assumptions required for parametric tests like ANOVA. In KNIME, Kruskal-Wallis Test is used to analyse the categorical variable. The variables that have $p < 0.05$, those variables will be significant in the analysis of clusters.

Year

Row ID	H-Value	p-value	Mean R...	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	874.871	0.0	36,150.457	34,066.5	39,263.779	44,661	41,944.894	44,661

Make

Row ID	H-Value	p-value	Mean R...	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	40.171	1.8924198874614717E-9	39,779.732	40,859	39,033.123	40,859	38,541.674	40,859

Model

Row ID	H-Value	p-value	Mean R...	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	122.94	0.0	40,379.249	40,659	38,404.096	36,763	38,568.612	38,373.5

Transmission

Row ID	H-Value	p-value	Mean R...	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	6.191	0.04525202659744...	39,203.13	37,897.5	39,065.718	37,897.5	39,086.399	37,897.5

Colour

Row ID	H-Value	p-value	Mean R...	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	5.763	0.05605650587541...	39,365.064	43,728.5	39,091.768	43,728.5	38,898.322	43,728.5

State

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	H-Value	p-value	Mean R...	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	234.26	0.0	38,014.166	37,623.5	38,505.297	37,623.5	40,835.341	46,538

Interior

Row ID	H-Value	p-value	Mean R...	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	133.82	0.0	40,208.479	46,853	39,082.565	46,853	38,063.177	46,853

Seller

Table "default" - Rows: 1 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	H-Value	p-value	Mean R...	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	40.414	1.6760195453713322E-9	38,908.538	39,642	38,620.898	37,698	39,824.765	40,161.5

The p-value associated with the Kruskal-Wallis test is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there are statistically significant differences between the medians of cluster 0 and cluster 1.

The mean and median ranks of each cluster indicate the average and middle positions of the observations within each group. The differences in these values between the two clusters suggest variations in the distribution of data points, contributing to the rejection of the null hypothesis.

We see that all the categorical variables have p-value less than 0.05 indicating that there are significant differences in the distributions of the data between cluster 0 and cluster 1 as indicated by the Kruskal-Wallis test results.

Year, Make, Model, Transmission, State, Interior, Seller are significant. Colour is not significant.

3.2.3.2 Cluster analysis with Non-Categorical Variables

In KNIME, ANOVA is used to analyse the non-categorical variables. The variables that have $p < 0.05$, those variables are significant in the analysis of clusters.

For K = 3

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
odometer	Between Groups	349.2027	2	174.6014	176.3981	0.0
odometer	Within Groups	77,436.1284	78233	0.9898		
odometer	Total	77,785.3311	78235			
mmr	Between Groups	418.9728	2	209.4864	210.489	0.0
mmr	Within Groups	77,860.3601	78233	0.9952		
mmr	Total	78,279.3329	78235			
sellingprice	Between Groups	449.663	2	224.8315	226.0394	0.0
sellingprice	Within Groups	77,814.9481	78233	0.9947		
sellingprice	Total	78,264.6111	78235			

Descriptive Statistics :

Row ID	Test Co...	Group	N	Missing ...	Missing ...	Mean	Standa...	Standa...	Confide...	Confide...	Confide...	Minimum	Maximum
Row0	odometer	cluster_0	26124	0	0	0.083	1.038	0.006	0.95	0.07	0.095	-1.349	2.743
Row1	odometer	cluster_1	26016	0	0	-0.017	0.991	0.006	0.95	-0.029	-0.005	-1.349	2.743
Row2	odometer	cluster_2	26096	0	0	-0.079	0.953	0.006	0.95	-0.091	-0.068	-1.349	2.743
Row3	odometer	Total	78236	0	0	-0.005	0.997	0.004	0.95	-0.012	0.002	-1.349	2.743
Row4	mmr	cluster_0	26124	0	0	-0.086	0.991	0.006	0.95	-0.098	-0.074	-1.586	2.567
Row5	mmr	cluster_1	26016	0	0	0.005	0.989	0.006	0.95	-0.007	0.017	-1.586	2.567
Row6	mmr	cluster_2	26096	0	0	0.093	1.013	0.006	0.95	0.08	0.105	-1.586	2.567
Row7	mmr	Total	78236	0	0	0.004	1	0.004	0.95	-0.003	0.011	-1.586	2.567
Row8	sellingprice	cluster_0	26124	0	0	-0.096	0.99	0.006	0.95	-0.108	-0.084	-1.537	2.562
Row9	sellingprice	cluster_1	26016	0	0	0.018	0.989	0.006	0.95	0.006	0.03	-1.554	2.562
Row10	sellingprice	cluster_2	26096	0	0	0.088	1.013	0.006	0.95	0.076	0.1	-1.542	2.562
Row11	sellingprice	Total	78236	0	0	0.003	1	0.004	0.95	-0.004	0.01	-1.554	2.562

The null hypothesis is rejected in all of the variables since the p-value is less than 0.05, indicating that there are significant differences in odometer, mmr, selling price between the groups.

Odometer, mmr and Selling Price are Significant.

4. Results and observation

4.1 Appropriate Number of Segments or Clusters

Cluster No.	Clusters	Silhouette Score	Mean
2	Cluster 0	0.23	0.011
	Cluster 1	-0.208	
3	Cluster 0	0.622	0.582
	Cluster 1	0.509	
	Cluster 2	0.614	
4	Cluster 0	0.62	0.556
	Cluster 1	0.502	
	Cluster 2	0.498	
	Cluster 3	0.604	
5	Cluster 0	0.617	0.539
	Cluster 1	0.497	
	Cluster 2	0.495	

	Cluster 3	0.491	
	Cluster 4	0.595	

The silhouette score for all the clusters is present. The analysis of the table will be done on 2 factors: -

1) Higher the silhouette score i.e. close to 1 more are the clusters separated and close to 0 indicates the clusters are overlapping

2) Sometimes having a smaller number of clusters can be very simplistic and the service provider may take simple decisions according it which will eventually hamper their market penetration and having simplified services/products may forgone the people who are the potential customers. Having more services will give the service provider a unique value proposition to attract customers.

4.2 Cluster analysis

4.2.1 Categorical Variables

It has been observed that all the variables except color are contributing to the cluster for making the service or product. This is because the p-value is less than 0.05 (confidence level at 95% for the model) which in turn tells that all other categorical variables are significant for the process of making the clusters.

4.2.2 Non-Categorical Variables

It has been observed that half the variables are contributing to the cluster for making the service or product. This is because the p-value is less than 0.05 (confidence level at 95% for the model) which in turn tells that odometer, mmr and selling price are significant

are significant for the process of making the clusters.

5. Managerial Insights

5.1 The managerial insights that can concluded by doing the k-means clustering as well as selecting the appropriate number of clusters as 3 are: -

→ Insights for cluster 0 which represent car with maker Ford, model Altima and sedan body

1. Tailored Marketing: Direct marketing towards Ford enthusiasts, highlighting the unique advantages and features of Ford automobiles.
2. Personalized Deals: Customize promotions and offers to showcase Ford's key attributes, like its resilience and performance.
3. Sedan-Centric Approaches: Craft marketing strategies around sedan models, such as the Altima, to engage consumers attracted to this specific body type.
4. Loyalty Initiatives: Introduce loyalty programs or rewards structures to encourage repeat purchases and cultivate brand allegiance within the Ford community.
5. Upselling Possibilities: Recognize chances to propose supplementary Ford products or services that complement sedan ownership, like service packages or extended warranties.

→ Insights for cluster 1 which represent car with maker Honda, model Camry and SUV body

1. Honda SUV Emphasis: Highlight the reliability, safety features, and versatility of Honda SUVs to appeal to customers in this cluster.
2. Family-Oriented Marketing: Create marketing campaigns that emphasize the spaciousness and family-friendly aspects of Honda SUVs like the Camry.
3. Adventure and Lifestyle: Emphasize the outdoor and adventure capabilities of Honda SUVs to resonate with customers seeking an active lifestyle.
4. Convenience Services: Offer convenience services such as home delivery for test drives or vehicle maintenance to cater to busy SUV owners.
5. Community Engagement: Engage with local communities and events to showcase Honda SUVs and build rapport with potential customers.

→ Insights for cluster 2 which represent customers who prefer Chevrolet's car, F-150 model and sedan body

1. Chevrolet Sedan Focus: Highlight the comfort, fuel efficiency, and affordability of Chevrolet sedan models like the F-150 to attract customers in this cluster.
2. Urban Lifestyle Appeal: Position Chevrolet sedans as ideal vehicles for city living, emphasizing features like compact size and easy manoeuvrability.
3. Value Proposition: Emphasize the value proposition of Chevrolet sedans, offering competitive pricing and cost-effective ownership experiences.
4. Technology Integration: Showcase the latest technology and infotainment features available in Chevrolet sedans to appeal to tech-savvy customers.
5. After-Sales Services: Provide excellent after-sales services such as maintenance packages and roadside assistance to enhance customer satisfaction and loyalty.

5.2 Cluster (Heterogenous) Identity

Identity of cluster 1: Customers who value reliability, affordability, comfort, fuel efficiency, and practicality.

Identity of cluster 2: Customers who are family-oriented, seeking vehicles that offer ample space and versatility for various activities and lifestyles, also have interest in features that enhance convenience and comfort.

Identity of cluster 3: Customers who may prioritize a balance of performance, affordability, and style for everyday purpose. May have potential interest in advanced technology features.