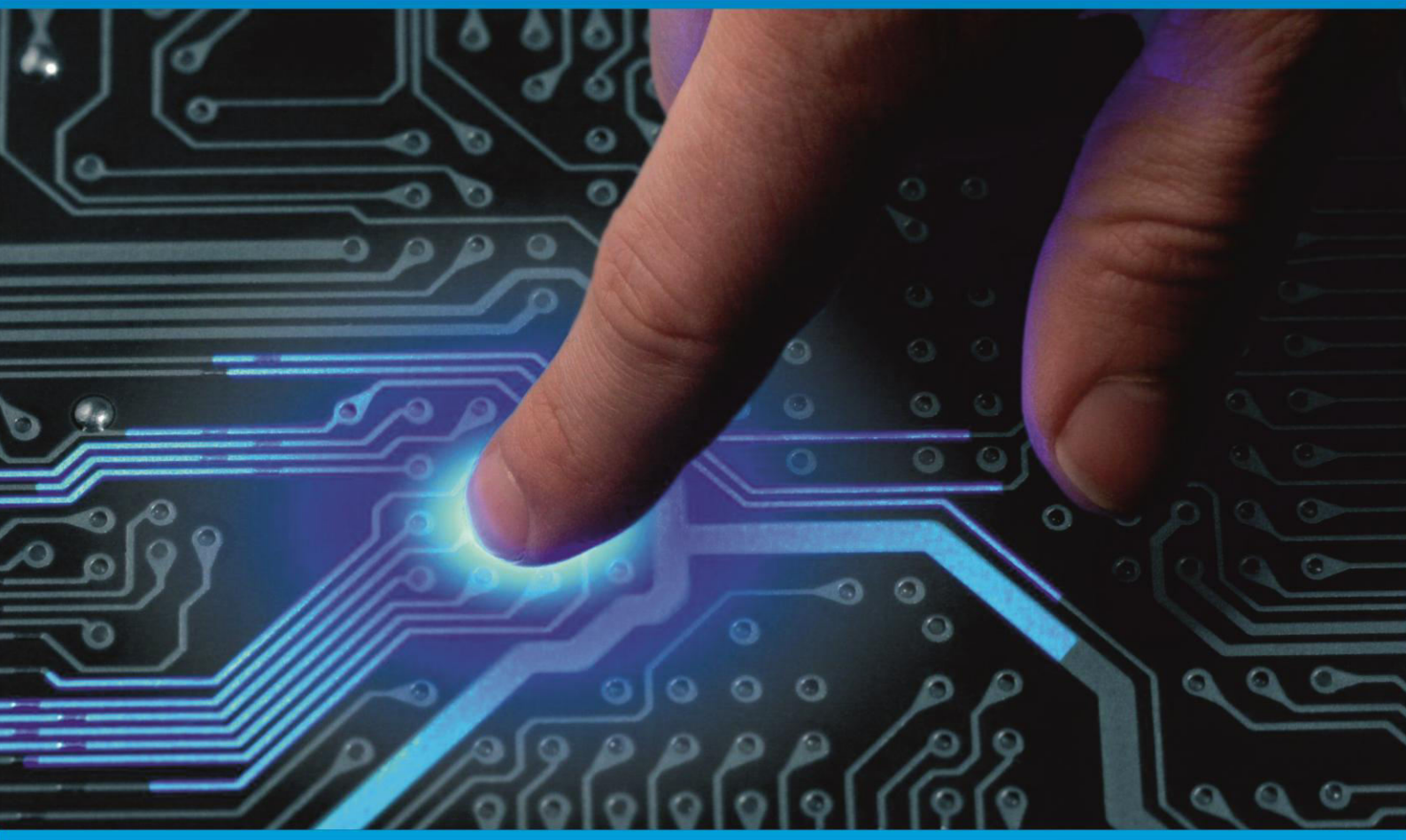
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 5, May 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

A Novel Study and Analysis on Pricing of Automobile Industry

Nikhil Mishra[1], Siddharth Nanda[2]

U.G Student, School of Engineering, Ajeenkya DY Patil University, Pune, Maharashtra, India [1]

Faculty, School of Engineering, Ajeenkya DY Patil University, Pune, Maharashtra, India [2]

ABSTRACT: The automobile industry is a booming sector with nonstop growth in terms of product sales and new model launches in the market. The Indian automobile sector has a high chance of success, being the key reason why companies like Kia, Jeep, and Citroen are entering the market. The major question is what are the key factors that the companies entering this market should focus upon, as each price segment is highly flooded with well-established companies like Maruti Suzuki, what key areas should the new companies focus upon? In this paper there is detailed analysis through visualizations on understanding the pricing criteria and what factors does the market demand, ensuring that the new companies meet them also ensuring the maximization of profit.

KEYWORDS: Stratified Sampling, Data Visualization, Student's T-Test, Variance Inflation Factor

I. INTRODUCTION

A total of daily 73,632 vehicles alone are sold in the Indian subcontinent alone, making it one of the fastest-growing industries. The automobile industry is ranked in the top 10 industries contributing towards the GDP of the country, let aside the after-sales service like servicing and maintenance bumping up the profits graph. So, to get a piece of this pie of profits there is cut-throat competition in this market with existing companies and also new companies entering the market every quarter of the year. Some of the new companies ended up gaining a huge share of this "profit-pie" like Kia and some companies were thrown out of the same market like Chevrolet.

The key question that pops up is that what does it take to enter or even stay in this market? In this paper, we try to answer that. First, we pull up the data about the vehicle currently being sold and various parameters like engine type, wheelbase length, etc figuring out the important parameters that play a key role in deciding the price of the vehicle and how well those parameters describe the price of the vehicle, Second, we apply analytics on the data plotting various plots to understand the different parameters and their relationship with the price of the car. After that we apply stratified shuffling to the dataset shuffling them in accordance with the *fuel type,* that is diesel or gas ensuring the training and testing data are divided in the same proportion of the fuel. Thirdly, we find out the important parameters for deciding the price of the vehicle by checking its p-value and Variance Inflation Factor(VIF) majorly and other statistic values such as F-statistic, coefficient, standard error, etc.

The statistical findings through VIF and p-value play a major role in figuring out the key parameters which help us finding the values which affect the price of the vehicle the most.

1.1 Data

The following columns are present in the dataset which we are analyzingcontains the following attributes:

Column Name	Description
car_ID	Unique ID of each car
symboling	Insurance risk rating
CarName	Car Make and Model Name
fueltype	Diesel or Gas
aspiration	Standard or Turbocharged
doornumber	Number of doors in the vehicle

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

carbody	Type of body of the car
drivewheel	Wheels where engine power is distributed
enginelocation	Location of the engine in the vehicle
wheelbase	Length from the front axle to the rear axle
carlength	Length from the front bumper to rear bumper
carwidth	Width of the car
carheight	Length from the floor of the car to the roof
curbweight	The total mass of the vehicle with standard equipment
enginetype	Type of the engine
cylindernumber	Number of valves in the car
enginesize	Size of the engine
fuelsystem	The method through which fuel is injected into the engine
boreratio	The ratio of the distance traveled by a piston in a cylinder to the diameter of the cylinder.
stroke	An internal-combustion engine goes through four strokes: intake, compression, combustion (power), and exhaust
compressionratio	the ratio of the maximum to minimum volume in the cylinder of an internal combustion engine.
horsepower	Horsepower refers to the power an engine produces.
peakrpm	peak power is produced in the upper-speed range where there's both high torque and high RPM.
citympg	Distance covered (in miles) per gallon in a city
highwaympg	Distance covered (in miles) per gallon on a highway
price	Costing of the car

Table 1 : Attributes of the Data

II. LITERATURE SURVEY

Sudhir K [1] tracks down that homegrown firms cost forcefully in passage level fragments, in which the Japanese have acquired a more prominent piece of the pie however are more helpful in bigger vehicle portions. It is notable that multimarket contact gives extra procedures to firms to improve collaboration (Bernheim and Whinston 1990). Investigating what serious conduct one section means for the conduct of firms in different fragments would be a productive territory for future examination

TF Brenshan[2] et. al shows that as costs of domestic vehicles keep on dropping, purchasers may in the long run fume advantage of the serious limitations forced during the development of this infant industry. They also predict that each Chinese family will claim a vehicle by 2050. This increment in car possession is massively critical for a country in which bicycles represented perhaps the most well-known types of transportation under two decades back.

Busse[3] et al demonstrated in their paper that obstacles in other markets may include frequent changes for quality and highlights or the presence of clients who need adequate mastery to discover or decipher value data. The requirement that clients need value data likewise makes it more uncertain that the advancements will be powerful in business sectors for extravagant consumables that are bought oftentimes. Second, the idea of "representative limits" might be innately more trustworthy in enterprises that are known to offer profound limits to workers. This may incorporate aircraft and retail chains, the two of which offer liberal representative limits. At last, we can estimate that the EDP advancements are additionally bound to succeed if the item quality is by and large known.

Roger[4] et al shows that the theory of blemished rivalry was utilized to clarify the obvious absence of connection Amon's base and double TFP gauges in U.S. fabricating. The exact outcomes presented above show the significant informative force of this hypothesis. The speculation likewise plainly overwhelms elective clarifications as, for instance, those dependent on work storing and abundance limit. As a side-effect of the investigation, I likewise give an elective strategy to assessing markup proportions that don't need the solid distinguishing suspicions as found in Hall's examination

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

Rothengatter[5] et al show in their paper that when one loosens up the severe suspicions of neo-old style government assistance hypothesis the "primary best" – rules like negligible expense evaluating breakdown. In reality, it is a significant issue of financial guidance to think about the unique motivating force designs, the agreeableness, and the institutional results of an estimating plan. When these angles are brought bit by bit into the examination the valuing of transport foundation on the basis of minor expenses is not, at this point ideal it can prompt genuine aggravations of long haul motivations. It can without much of a stretch be shown that the presentation of a spending limitation drives as of now to the outcome that nonlinear, non-uniform evaluating, for example, multi-part levies is Pareto-prevalent.

SN Teli[6] et al researched that improving the bottom line is the goal while preventive expenses may expand, generally speaking, operational expense will decrease through the decrease of disappointments, and An appropriately comprehended and oversaw quality expense system will help associations in acknowledging cost reserve funds while avoiding a portion of the genuine entanglements that can accompany cost-cutting; diminishes in the item or administration quality, increased client disappointment, added adjust costs, or straightforward movements in costs starting with one territory then onto the next.

Mahendran[7] et al show that lean assembling is effectively executed in Rane motor valves restricted, Tiruchirappalli, Tamilnadu, India. The non-esteem added time is diminished from 794 min to 566 min, 28.71% improved. The worth added time is diminished from 1,602 sec to 1,156 sec, 27.84% improved. The all-out stock is diminished from 1,268 to 950, as 25.07%. The rate esteem expansion is expanded from 3.25% to 3.29. From the man-machine outline, the consolidated working of pounding and machine machines, the working time expanded from 24 hours to 29 hours. The inactive time is decreased from 24 hours to 19 hours. The general proficiency of the business is improved.

The results of the analysis by Ashgar[8] et al demonstrate that Audi's positive survey rate was bigger than other contenders, by a level of 87%. More-finished, the negative extremity of Audi Organization is more limited than other contender companies, by a level of 18%. From the consequences of this exploration, we can presume that the customers of Audi Organization have bigger satisfaction when contrasted with Honda, Toyota BMW, and Mercedes's clients. The consequences of this exploration will help the customers that wanting to buy a vehicle to dissect among these three organizations on the premise of previous customer reviews.

Yu[9] et al shows in their paper that the relevant policies efficiently direct the new energy car industry to grow overwhelmingly from the parts of thorough advancement, energy-saving format, cost, charging gear and batteries, appropriations, charge exclusion, advances, and investment. The execution of the help strategy in the important zones of the improvement of the new energy car industry will colossally affect the ecological insurance, tax collection, work, logical exploration, and related businesses in the applicable zones, in this way carrying the territorial monetary advancement to another peak.

The plan of this examination by Anuragi[10] et al was to comprehend Kia's methodology that permitted it to eclipse its rivals being another contestant in a market that is in the grips of a delayed log jam. The exploration features the variables that control the client's purchasing choice. The reactions by the directors of brands overviewed were contrasted with recognizing the recognized practices by Kia that gave it the edge. This exploration will assist the brands with arranging the plunge by zeroing in on the significant angles influencing the purchasing choice. Today, the clients want the best highlights at the best cost because of the accessibility of numerous choices. Consequently, the brands need to offer the most recent items and administrations at a reasonable cost. According to market flow requests, broad statistical surveying and ideal progressions in innovation speed up client securing. To expand the deals and serve a bigger market fragment, the brands ought to likewise investigate online circulation channel which is at present on the ascent.

III. CONCEPTS IMPLEMENTED

3.1 Stratified Sampling

Stratified sampling is a method of sampling from a population that can be divided into subpopulations in statistics.When subpopulations within an overall population differ, it may be helpful in statistical surveys to sample each subpopulation (stratum) separately.Before sampling, stratification is the process of separating members of a population into homogeneous subgroups.The strata should be used to divide the population.That is, it must be collectively exhaustive and mutually exclusive: each element of the population must be assigned to one and only one

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

stratum.Then, within each stratum, simple random sampling is used. The goal is to increase sample accuracy by minimizing sampling error.It can generate a weighted mean with less variability than a simple random sample of the population's arithmetic mean.When Monte Carlo methods are used to estimate population statistics from a known population, stratified sampling is a method of reducing variance in computational statistics.

1. Proportionate allocation employs a sampling fraction proportional to the total population in each stratum. For example, if the population is made up of n total individuals, m of whom are male and f of whom are female (and where m + f = n), the relative size of the two samples (x1 = m/n males, x2 = f/n females) should represent this proportion.
2. Optimal (or disproportionate) allocation - The sampling fraction of each stratum is proportional to both the percentage (as stated above) and the standard deviation of the variable's distribution. To produce the least amount of overall sampling variance, larger samples are taken in strata with the highest variability. The mean and variance of stratified random sampling are given by:

$$\bar{x} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{x_h}$$

$$s_{\bar{x}}^2 = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

where,
L = number of strata
N = the sum of all stratum sizes
N_h = size of stratum h
x_h= sample mean of stratum h
n_h=number of observations in stratum h
s_h= sample standard deviation of stratum h

3.2 Student's t-test

When the population standard deviation is unknown, the Student's t-test is used to test hypotheses about the mean of a small sample drawn from a normally distributed population. Typically, a null hypothesis is first developed, which states that there is no effective difference between the observed sample mean and the hypothesized or stated population mean—that any measured difference is due solely to chance. In an agricultural study, for example, the null hypothesis might be that fertilizer application had no effect on crop yield, and an experiment would be conducted to see if it had. In general, a t-test may be two-sided (also known as two-tailed), stating merely that the means are not equivalent, or one-sided, stating whether the observed mean is greater or less than the hypothesized mean.Using the formula for the *t*-statistic,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

3.3 Variance Inflation Factor

In regression analysis, a variance inflation factor (VIF) is used to eliminate multicollinearity. Multicollinearity occurs when there is a correlation between predictors (i.e. independent variables) in a model; its presence can have a negative impact on your regression results. The VIF calculates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.
VIFs are typically calculated by software as part of a regression analysis. A VIF column would appear in the output. VIFs are computed by regressing a predictor against each other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula. "i" is the predictor you're looking at (e.g. x_1 or x_2):

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

IV. RESULTS

In this section, we will be focusing upon the findings through the dataset where we first visualize the target variable with all the other parameters, then split the dataset into training and testing with the aid of stratified shuffling after

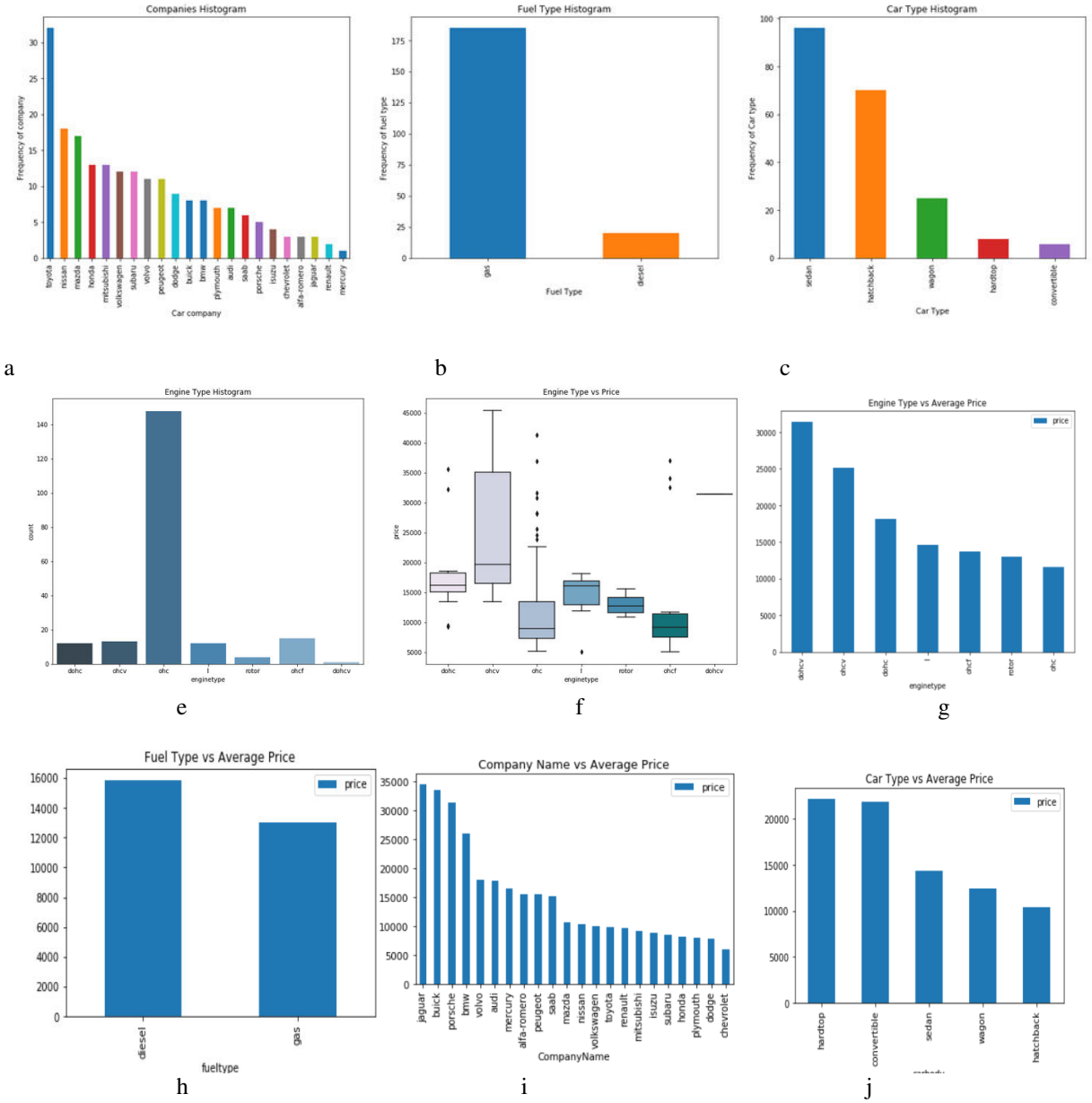
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

which we will apply statistical techniques to find out the independent variables which influence the dependent variable the most.
We start the analysis by visualizing a few categorical variables plotting them against the dependent variable (price) and noting down our inference in the end and doing the same for numerical data.

4.1 Visualizing Categorical Data



Plot 1. Categorical data v/s price

From plot 1, after thorough analysis we have encountered various key inferences, we can jot down those inferences in very simple terms:from plot 1a we can decipher thatToyota seemed to be the favored car company, from plot 1b it can be understood that the number of gas-fueled cars is more than diesel, from plot 1c it is deciphered that the sedan is the top car type preferred by the buyers, from plot 1e we note that the ohcengine type seems to be the most favored type, from plot 1f we decipher that ohcv has the highest price range (while dohcv has only one row), ohc and ohcf have the low price range, plot 1g expresses that Jaguar and Buick seem to have the highest average price, from plot 1h it is deciphered that diesel has a higher average price than gas and plot 1i and plot 1j signifies that hardtop and convertible have higher average price.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

4.2 Visualizing numerical data

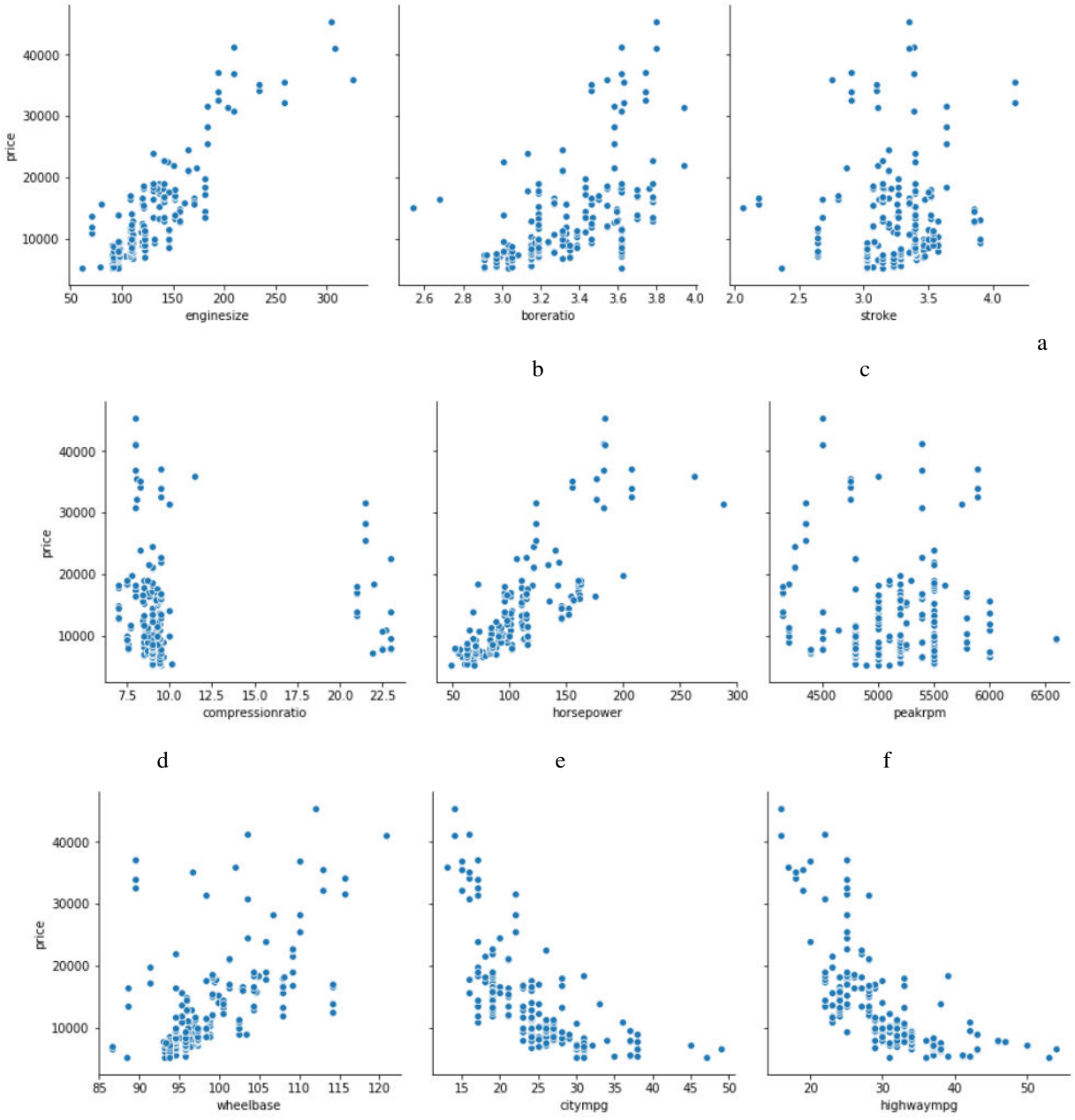


Image 2. Plots of Numerical data v/s price

From plot 2, after thorough analysis we have encountered various key inferences, we can jot down those inferences in very simple terms: from plot 2a, 2b, 2e, 2g we notice that enginesize, boreratio, horsepower, wheelbase - seem to have a significant positive correlation with price and from plot 2h, 2i it is deciphered that citympg, highwaympg - seem to have a significant negative correlation with price.

4.3 Statistical Findings

Using stratified sampling, we first split the population into training and testing datasets, meaning that all fuel types are included in the training and testing datasets. This is accomplished by the use of the sklearn kit, which provides train/test indices for splitting data into train/test sets. This cross-validation object is a combination of StratifiedKFold and Shuffle

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

Split, and it produces stratified randomised folds. The folds are generated by keeping the percentage of samples for each class constant.

After dividing the dataset, into stratified samples we apply multiple OLS Regression on the model to find out the variables that influence the dependent variable significantly, deciding the significance by looking at the t value and comparing with the **confidence (p-value) which is 95% significance.**

OLS Regression Results

Dep. Variable:	price	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.923
Method:	Least Squares	F-statistic:	172.1
Date:	Tue, 23 Mar 2021	Prob (F-statistic):	1.29e-70
Time:	09:41:49	Log-Likelihood:	205.85
No. Observations:	143	AIC:	-389.7
Df Residuals:	132	BIC:	-357.1
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0947	0.042	-2.243	0.027	-0.178	-0.011
curbweight	0.2657	0.069	3.870	0.000	0.130	0.402
horsepower	0.4499	0.074	6.099	0.000	0.304	0.596
fueleconomy	0.0933	0.052	1.792	0.075	-0.010	0.196
carwidth	0.2609	0.062	4.216	0.000	0.138	0.383
hatchback	-0.0929	0.025	-3.707	0.000	-0.143	-0.043
sedan	-0.0704	0.025	-2.833	0.005	-0.120	-0.021
wagon	-0.0997	0.028	-3.565	0.001	-0.155	-0.044
dohcv	-0.2676	0.079	-3.391	0.001	-0.424	-0.112
twelve	-0.1192	0.067	-1.769	0.079	-0.253	0.014
Highend	0.2586	0.020	12.929	0.000	0.219	0.298

Omnibus:	43.093	Durbin-Watson:	1.867
Prob(Omnibus):	0.000	Jarque-Bera (JB):	130.648
Skew:	1.128	Prob(JB):	4.27e-29
Kurtosis:	7.103	Cond. No.	32.0

Image 3. OLS Summary during first run

Through this OLS summary report, it is quite evident that multiple columns(like twelve having p value 0.079) which have a p value greater that 0.005 can be removed as their significance is not much evident in determining the price of the vehicle.

This step is repeated **7 times**along with checking the VIF of the data leaving us with the top 5 variables that influence the price of the car the most.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	308.0
Date:	Tue, 23 Mar 2021	Prob (F-statistic):	1.04e-67
Time:	09:41:50	Log-Likelihood:	181.06
No. Observations:	143	AIC:	-352.1
Df Residuals:	138	BIC:	-337.3
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0824	0.018	-4.480	0.000	-0.119	-0.046
horsepower	0.4402	0.052	8.390	0.000	0.336	0.544
carwidth	0.3957	0.046	8.677	0.000	0.306	0.486
hatchback	-0.0414	0.013	-3.219	0.002	-0.067	-0.016
Highend	0.2794	0.022	12.591	0.000	0.236	0.323

Omnibus:	29.385	Durbin-Watson:	1.955
Prob(Omnibus):	0.000	Jarque-Bera (JB):	98.010
Skew:	0.692	Prob(JB):	5.22e-22
Kurtosis:	6.812	Cond. No.	12.9

Image 4. OLS Summary at the last run

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

And the VIF at the 7th step:

	Features	VIF
0	const	10.04
1	horsepower	2.22
2	carwidth	2.08
4	Highend	1.53
3	hatchback	1.10

Image 5. VIF

Post all the independent variable findings we perform a residual analysis on the model checking the Error terms.

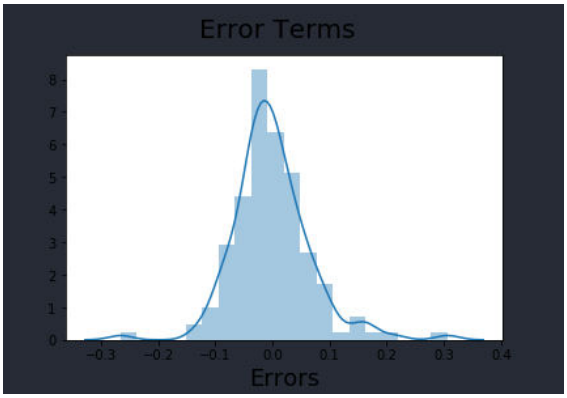


Image 6. Error distribution

The error terms seem to be approximately normally distributed, so the assumption on the linear modeling seems to be fulfilled.

V. FUTURE SCOPE AND DISCUSSION

With the changes in the market, new data about the vehicles selling will keep on adding bringing more findings and more independent factors affecting the price changes of the car. More factors like the demographics of the location can also be considered in the price affection factor, bringing a wider perspective and ensuring that the market is entered with proper preparation. Demographic will play a key role in understanding the needs of the market and the vehicle to be designed for the same. A company selling SUVs will be more profitable in the terrain area and sand areas as a bigger tire size is required similarly a sedan selling company will be more profitable in cities and suburban.

VI. CONCLUSION

This study demonstrates that the data in the columns horsepower, vehicle-width, hatchback, and high-end influence car pricing to a greater degree than the other factors, suggesting that anytime a new enterprise wants to enter this market, it must bear these factors in mind to ensure optimum profitability. Both of these variables can only be measured using different statistical approaches such as the Variance Inflation Factor and the effects of t-tests after stratifying the sample, showing that statistical findings can be a shortcut for a company's path to performance.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905051 |

REFERENCES

1. Sudhir, K., 2001. Competitive pricing behavior in the auto market: A structural analysis. *Marketing Science*, *20*(1), pp.42-60.
2. Bresnahan, T.F., 1981. Departures from marginal-cost pricing in the American automobile industry: Estimates for 1977–1978. *Journal of Econometrics*, *17*(2), pp.201-227.
3. Busse, M.R., Simester, D.I. and Zettelmeyer, F., 2010. "The best price you'll ever get": The 2005 employee discount pricing promotions in the US automobile industry. *Marketing science*, *29*(2), pp.268-290.
4. Roeger, W., 1995. Can imperfect competition explain the difference between primal and dual productivity measures? Estimates for US manufacturing. *Journal of political Economy*, *103*(2), pp.316-330.
5. Rothengatter, W., 2003. How good is first best? Marginal cost and other pricing principles for user charging in transport. *Transport policy*, *10*(2), pp.121-130.
6. Teli, S.N., Majali, V.S., Bhushi, U.M., Gaikwad, L.M. and Surange, V.G., 2013. Cost of poor quality analysis for automobile industry: A case study. *Journal of The Institution of Engineers (India): Series C*, *94*(4), pp.373-384.
7. Mahendran, S., Senthilkumar, A. and Jeyapaul, R., 2018. Analysis of lean manufacturing in an automobile industry-a case study. *International Journal of Enterprise Network Management*, *9*(2), pp.129-142.
8. Asghar, Z., Ali, T., Ahmad, I., Tharanidharan, S., Nazar, S.K.A. and Kamal, S., 2018, October. Sentiment Analysis on Automobile Brands Using Twitter Data. In *International Conference on Intelligent Technologies and Applications* (pp. 76-85). Springer, Singapore.
9. Yu, Y. and Jiang, J., 2020, March. Analysis on the impact of new energy automobile industry support policy on regional economy. In *IOP Conference Series: Earth and Environmental Science* (Vol. 467, No. 1, p. 012209). IOP Publishing.
10. Anuragi, K., Raj, A. and Bajpai, S., 2021. Analysis of Kia Motors' Booming Market Penetration amidst Downturn of India's Automobile Industry. *IJRAR-International Journal of Research and Analytical Reviews (IJRAR)*, *8*(1), pp.20-25.
11. https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html accessed on 20-03-2021 at 3 pm.
12. https://www.investopedia.com/terms/stratified_random_sampling.asp accessed on 18-03-2021 at 1 pm.
13. https://www.investopedia.com/terms/v/variance-inflation-factor.asp accessed on 19-03-2021 at 4 pm.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details