Final Project Report

Analysis of the Indian Car Market

Quantitative & Qualitative Analysis of Variable Impact on Vehicle Selling Price

By: Joyce Teng (jt939) and Alissa Eng (ace77)

Executive Summary

The Indian car market presents a unique opportunity for automakers and sellers due to its interesting blend of rapid growth and low market saturation. Despite having a low penetration rate of just 57 cars per 1,000 people, India ranks as the third-largest car market behind China and the United States. Additionally, its rapidly expanding middle-class population, harsh environment, and wide range of transportation alternatives, including motorcycles, scooters, and three-wheelers, contribute towards a market where affordability, fuel efficiency, and durability are the key driving factors behind consumer behavior over size or comfort.

This report analyzes the variables that influence vehicle selling price in India through a variety of statistical analysis methods, including distribution analysis, chi-square dependency tests, and regression models. The data suggest that selling prices in India are strongly right-skewed, signaling a market heavily concentrated in the budget/ affordability segment. Newer vehicles have higher prices, and price dispersion has increased in recent years, likely due to the presence of luxury outliers. Chi-square tests reveal that fuel type is dependent on transmission, and seller type is dependent on selling price. These align with market behavior.

Regression analysis reveals that incorporating nonlinear terms significantly improves predictive analysis. While a basic linear model explains ~64% of price variation, adding quadratic terms increases this to ~75%. A model using squared year, engine size, and cubic horsepower captures ~76.95% of price variation (measured using adjusted R^2). These findings indicate that variables such as year and engine size follow quadratic relationships, while maximum horsepower follows a cubic relationship, showing that returns to certain vehicle attributes accelerate rather than increase proportionally.

Key insights include:

- Prioritize newer manufacturing years due to an exponential depreciation pattern
- Invest in larger engines as power output increases exponentially
- Horsepower generates increasing returns up to a point, after which there are diminishing returns
- Seller type affects pricing, with dealers and certified sellers commanding premium prices

Overall, our analysis emphasizes that automakers and sellers that wish to operate in India's car market must prioritize affordability and performance efficiency to best position themselves for successful market penetration.

Table of Contents

Proposal & Background:

The Indian car market is especially unique in its growth potential in the near future. The country's rapidly growing middle class, combined with noticeably high tariffs, makes the automobile industry an interesting market to navigate. In stark contrast to America, which operates primarily on automobiles aside from a select few metropolitan cities that rely on public transport, including buses and trains, India has far more alternatives, including motorcycles, three-wheelers, and scooters. As a result, cars are less of a necessity due to the various alternatives possible that cost less and are more efficient, in terms of space, maintenance, as well as other factors. However, recently, India has presented as one of the fast-growing major auto markets globally, falling only behind China and the United States. On the other hand, despite this recent rapid growth, as of 2023, India has a very low car penetration rate of just 57 cars per 1,000 people, compared to China, which sits around 322 per 1,000 people. As a result, unlike other countries where the need has been mostly fulfilled, and the market is saturated, India's car market still has space for automakers to enter and take up market share.

While the Indian car market has great potential for growth, there are many factors that make it difficult to navigate and for an automaker to become profitable. Difference in consumer taste, with a strong preference for affordability and fuel efficiency over size and comfort, in addition to India's harsh operating environment, forces automakers and sellers to strategically position their offerings.

This report will highlight the most significant variables that impact the selling price of vehicles in the Indian market.

Breakdown of the variables:

- **Car model name**
- **Year**:
    - Year indicates the vehicle's manufacturing date. A larger value corresponds to a more recent model.
- **Km driven:**
    - The number of kilometers already driven on the vehicle - this is especially significant for used vehicles.
- **Fuel type:**
    - Fuel type indicates the type of fuel the vehicle uses. Different fuels come with different cost structures, efficiency levels, and consumer demand patterns. Our dataset includes petrol, diesel, CNG, and LPG.
- **Seller type:**
    - Identifies whether the vehicle is sold by an individual, a dealer, or a trustmark dealer. This affects the selling price because dealers often offer newer or certified vehicles, while individuals typically sell older or more used cars.
- **Transmission:**
    - Specifies whether the vehicle has a manual or automatic gearbox. Manual transmissions are more common in lower-priced cars, while automatic transmissions are associated with newer and higher-valued vehicles.
- **Count of previous owners**

- **Mileage**:
    - Mileage is measured in km/ltr/kg, which describes kilometers per liter per kilogram, which describes fuel efficiency. Kilometers per liter is often used when the fuel is a liquid - petrol, diesel, etc., kilometers per kilogram is often used when the fuel is a gas - CNG. Our analysis includes CNG, diesel, LPG, and petrol.
- **Engine size:**
    - Refers to the total volume of all the cylinders within an engine. This dataset measures this in cubic centimeters. The larger the engine size, the more fuel it is able to carry, and it is often correlated to more power, faster acceleration, and more fuel consumption.
- **Maximum Horsepower:**
    - Horsepower measures how powerful the engine is and the rate at which an engine can do work. The higher this value, the faster the acceleration and the maximum speed the vehicle is able to reach.
- **Number of seats**

Objectives:

- Analysis of selling price distribution
- How variables, including year, engine size, and maximum horsepower, correlate with selling price
- Dependency of fuel type and transmission
- Dependency of seller type and selling price
- Correlation of quantitative variables and analysis of non-linear relationships

Question 1: What is the shape of the distribution of selling price, and what does this tell us about what consumers have a preference for?

Our data suggests that the majority of cars are priced between 0 and 1,170,000 rupees, indicating that the distribution is strongly right-skewed with a strong preference for more affordable vehicles. The data also includes a long right tail, signalling that there does exist a number of high-priced vehicles that appear at a much lower frequency in comparison. Around 50% of vehicles are priced under 600,000 rupees (which translates to $6660), and 90% of vehicles are priced under 1,000,000 rupees ($11,000). A steep increase at low prices in our cumulative distribution function (Figure a2) signifies that many cars are clustered in the lower price range. Additionally, the overall shape confirms the right-skewed distribution pattern indicated in our probability distribution function. This supports that our data suggests most cars are low-priced, with a few expensive outliers.

Figure a1: Probability Distribution Function

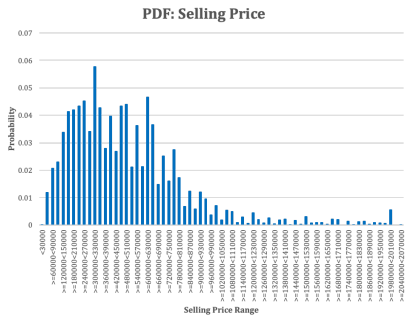
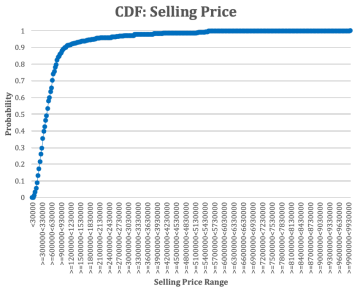
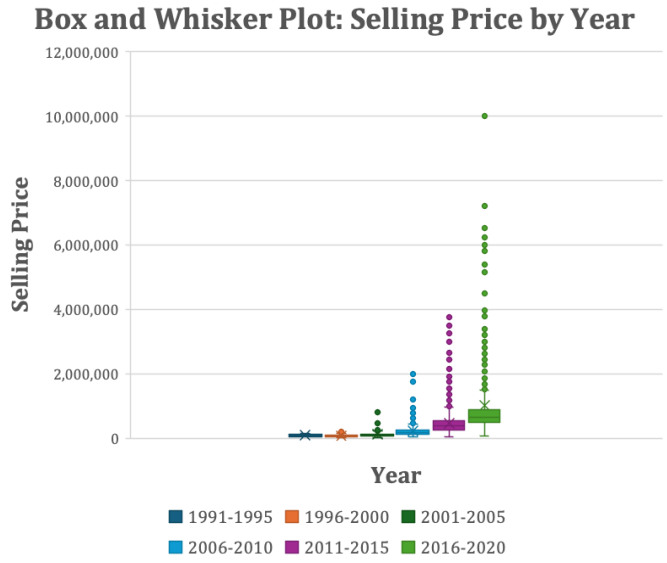


Figure a2: Cumulative Distribution Function



Taking this data into consideration, we continued to analyze our data distribution by year. Utilizing a box-plot model, we identified that the median selling price increases significantly over time. At the same time, there is also an increase in the number of outliers present. This makes sense as the closer the manufacturer year is to the present, cars have the higher the tendency to have differences in other variables such as packages, mileage, and condition. As a result, there are far more extreme values that would be classified as outliers. Furthermore, high-end luxury models are relatively new and have only been manufactured recently, distorting the data, and are naturally labeled as outliers. Older cars are typically less diverse in those same features and, in general, uniformly depreciate. This can be seen with older vehicles having the lowest medians and the smallest interquantile ranges, as well as the fewest number of outliers, as the selling price range is fairly inclusive. 2016-2020 has the highest median as well as the largest spread, representing both high median selling prices as well as extreme luxury outliers. Overall, this data and analysis suggest a strong positive relationship between year and selling price.

Figure a3: Box and Whisker Plot



<u>Question 2:</u> Do factors such as fuel type and seller type influence the selling price, and are these variables dependent on each other?

Figure b1: Chi-Square Test for Fuel Type and Transmission

Actual		Transmission		
		Automatic	Manual	Total
Fuel	CNG	0	57	57
	Diesel	534	3868	4402
	LPG	0	38	38
	Petrol	516	3115	3631
	Total	1050	7078	8128
Expected		Transmission		
		Automatic	Manual	Total
Fuel	CNG	7.363435	49.63656	57
	Diesel	568.66388	3833.336	4402
	LPG	4.9089567	33.09104	38
	Petrol	469.06373	3161.936	3631
	Total	1050	7078	8128
Expected		Transmission		
		Automatic	Manual	Total
Fuel	CNG	7.363435	1.092343	8.455778
	Diesel	2.1129959	0.313457	2.426452
	LPG	4.9089567	0.728229	5.637186
	Petrol	4.6966185	0.696729	5.393348
	Total	19.082006	2.830758	21.91276
df	3			
p-value	6.80E-05			

To evaluate whether fuel type and transmission are independent and whether this relationship may influence selling price, we conducted a chi-square test using the observed and expected values. The chi-square output produced a p-value of 6.80×10^{-5}, which is below our significance level of 0.05. As a result, we reject the null hypothesis and conclude that fuel type and transmission type are dependent variables. This finding suggests that certain fuel types tend to be paired with specific transmission systems. Diesel and petrol vehicles appear more frequently with manual transmissions, while automatic transmissions are more common among higher-priced models. If petrol vehicles are more likely to offer automatic transmission options, that combination may contribute to higher average selling prices within the petrol segment.

Figure b2: Chi-Square Test for Seller Type and Selling Price

Seller Type	0–100k	100–200k	200–300k	300–400k	400–500k	500–600k	600–700k	700–800k	800–900k	900–1M	>1M	Total
Individual	323	880	1080	1037	790	751	613	418	249	167	418	7847
Dealer	5	30	59	71	154	95	155	86	52	44	370	1353
Trustmark Dealer	0	0	29	2	2	35	64	62	2	3	33	8311
Total	328	910	1168	1110	946	881	832	566	303	214	821	8079

Seller Type	0–100k	100–200k	200–300k	300–400k	400–500k	500–600k	600–700k	700–800k	800–900k	900–1M	>1M	Total
Individual	318.5810125	883.868053	1134.459215	1078.124768	918.834262	855.700829	808.107934	549.746503	294.298923	207.854685	797.423815	7847
Dealer	54.93056071	152.3988117	195.6063869	185.8930561	158.427776	147.542146	139.336056	94.7887115	50.7437802	35.8388414	137.493873	1353
Trustmark Dealer	337.4189875	936.131947	1201.540785	1141.875232	973.165738	906.299171	855.892066	582.253497	311.701077	220.145315	844.576185	8311
Total	710.9305607	1972.398812	2531.606387	2405.893056	2050.42778	1909.54215	1803.33606	1226.78871	656.74378	463.838841	1779.49387	17511

Seller Type	0–100k	100–200k	200–300k	300–400k	400–500k	500–600k	600–700k	700–800k	800–900k	900–1M	>1M	Total
Individual	0.061295086	0.016927678	2.614290656	1.568692777	18.0644842	12.8108602	47.1064623	31.5729905	6.97247688	8.03015475	180.534402	309.353037
Dealer	45.38568077	98.30436959	95.40232936	71.01079843	0.12374849	18.7111087	1.76091626	0.81488025	0.03109915	1.85844481	393.174604	726.57798
Trustmark Dealer	337.4189875	936.131947	1144.240719	1137.878735	969.169849	837.650821	732.677716	464.855432	307.71391	214.186197	779.865589	7861.7899
Total	382.8659634	1034.453244	1242.257339	1210.458226	987.358081	869.17279	781.545095	497.243303	314.717486	224.074797	1353.5746	8897.72092

df	20
p-value	0.000000000000

In addition to examining fuel type, we also evaluated whether seller type (individual, dealer, and trustmark dealer) and selling price are independent. Using a chi-square test, we observed that the p-value was close to zero and below our significance level of 0.05. This indicates that seller type and selling price are dependent variables. These results make sense because dealers and trustmark dealers typically sell newer, better-maintained, or certified vehicles that often come with inspection services or warranties, which support higher prices. Individual sellers tend to list older or more heavily used vehicles and operate in the lower-priced segment of the market. Selling price depends a lot on the quality of cars each seller offers and how much trust and value buyers feel they are getting.

Question 3: What is the relationship between the qualitative variables, and what does this tell automakers interested in entering the market or resellers considering reselling their vehicle?

Preliminary Analysis – Model 1:

Figure c1: Model 1 Analysis

Summary of Fit	
RSquare	0.640136
RSquare Adj	**0.639863**

Summary of Fit	
Root Mean Square Error	488242.5
Mean of Response	649813.7
Observations (or Sum Wgts)	7906

Preliminary analysis of the data utilizing JMP software shows that, based on linear regression modeling, our data has an adjusted R^2 value of 0.63,9, indicating that approximately 63.9% of the variation in the selling price is explained by the variables in the model. This data is depicted in Figure c1.

The regression model suggests that the variables currently provide a fair fit for the data. The estimates inform us that for each year the vehicle is newer, the price increases by around ₹41 thousand. Each additional kilometer driven decreases the selling price by ₹1.41. Each unit increase in fuel efficiency increases the price by ~₹10,514. Each unit increase in engine size increases the price by ~₹114.65. Each additional horsepower increases the price by ~₹15,654. This is indicated in Figure c2.

Figure c2: Model 1 Term Breakdown

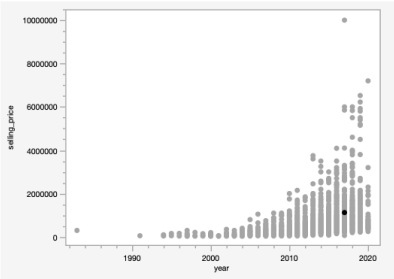
Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-83456276	3689853	-22.62	<.0001*	-90689364	-76223188
year	41110.173	1846.794	22.26	<.0001*	37489.969	44730.378
km_driven	-1.406987	0.113081	-12.44	<.0001*	-1.628656	-1.185319
mileage(km/ltr/kg)	10514.36	1902.622	5.53	<.0001*	6784.7187	14244.002
engine	114.65358	22.42028	5.11	<.0001*	70.703902	158.60326
max_power	15654.539	256.6877	60.99	<.0001*	15151.363	16157.714
seats	-73292.96	8285.675	-8.85	<.0001*	-89535.07	-57050.84

One variable that shows an interesting pattern is the fact that the number of seats estimated suggests that as the number of seats increases, there is a negative relationship with the selling price. Taking a broader analysis of the automobile industry, this may be due to the fact that luxury cars typically have fewer seats in comparison to budget options. As a result, this creates an interesting correlation between the two variables.

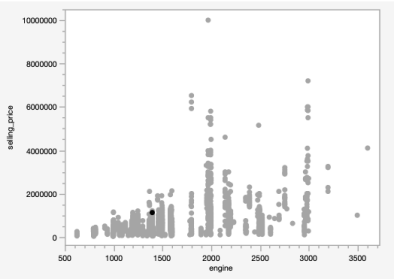
Non-Linear Variables:

In our analysis, there are three variables that indicate potential non-linear relationships: year, engine size, and maximum horsepower (Figure d1). From a glance, it appears that that year, engine size, and maximum horsepower may potentially better represent a quadratic relationship to the power of two.

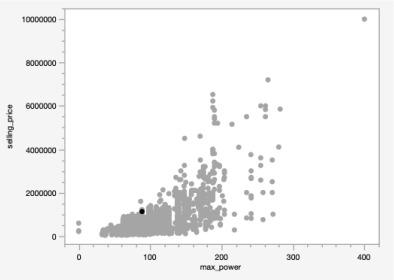
Selling Price by Year



Selling Price by Engine



Selling Price by Max Power



Model 2: Model including squared year, engine size, and maximum horsepower variables.

By including the squared regression of the three variables into model 2, we can see a significant increase in our adjusted R^2 value - from 0.639863 to 0.747948. This indicates that this representation of the relationship between variables and selling price accounts for around 74.79% of the variation in the selling price is explained by the variables in the model.

Figure e1: Model 2 – Summary of Fit

Summary of Fit	
RSquare	0.748235
RSquare Adj	0.747948
Root Mean Square Error	408457.4
Mean of Response	649813.7
Observations (or Sum Wgts)	7906

Model 2 shows an increase in the adjusted R^2 value compared to Model 1. This suggests that Model 2 is a better representation of the variation present in the data than Model 2.

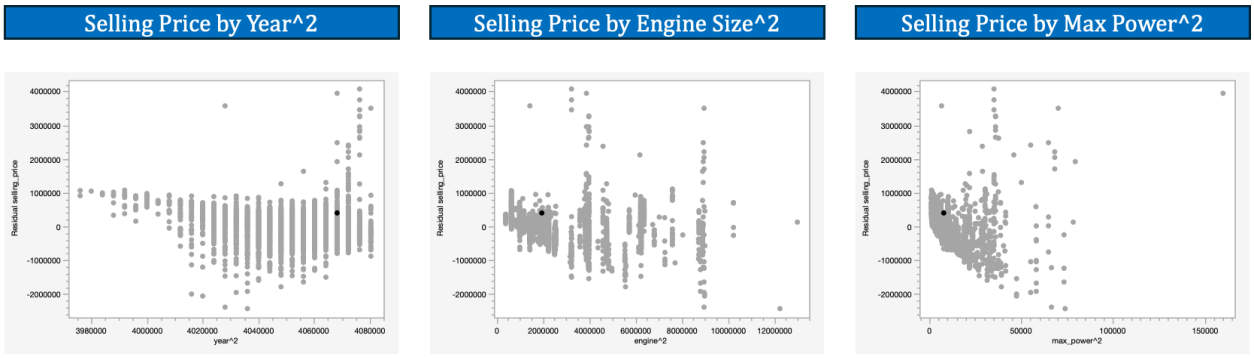
Figure e2: Model 2 – Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.606e+8	3633639	-44.21	<.0001*
year	79485.065	1812.81	43.85	<.0001*
km_driven	-0.489341	0.097201	-5.03	<.0001*
mileage(km/ltr/kg)	4876.3377	1613.363	3.02	0.0025*
engine	405.30613	24.37788	16.63	<.0001*
max_power	5504.3291	324.3186	16.97	<.0001*

seats	-22355.09	7010.41	-3.19	0.0014*
(year-2013.98)*(year-2013.98)	5279.2724	212.2866	24.87	<.0001*
(engine-1458.71)*(engine-1458.71)	-0.349286	0.021654	-16.13	<.0001*
(max_power-91.5874)*(max_power-91.5874)	132.96658	2.903083	45.8	<.0001*

This model tells us that for each year the vehicle is newer, the price increases by around ₹79,485. Each additional kilometer driven decreases the selling price by approximately ₹0.49. Each unit increase in fuel efficiency increases the price by ~ ₹4,878. Each unit increase in engine size increases the price by ~ ₹405.31. Each additional horsepower increases the price by ~ ₹5,504. Each additional seat decreases the price by ~ ₹29,255. The model includes squared terms for year, engine, and horsepower that capture nonlinear relationships in how these variables affect price. Although mileage and seats have a slightly higher p-value compared to the other variables, they remain statistically significant.

Figure e3: Model 2 - Residual Plots



The residual plots of these three variables following quadratic regression indicate that a majority of the patterns present have been extracted, especially for year and engine size. However, the plot for maximum horsepower indicates that there may still be information left to extract. This will be further broken down and analyzed in model 3.

Model 3: Squared year & engine size, cubic maximum horsepower

Figure f1: Model 3 – Summary of Fit

Summary of Fit	
RSquare	0.769827
RSquare Adj	0.769535
Root Mean Square Error	390574.8

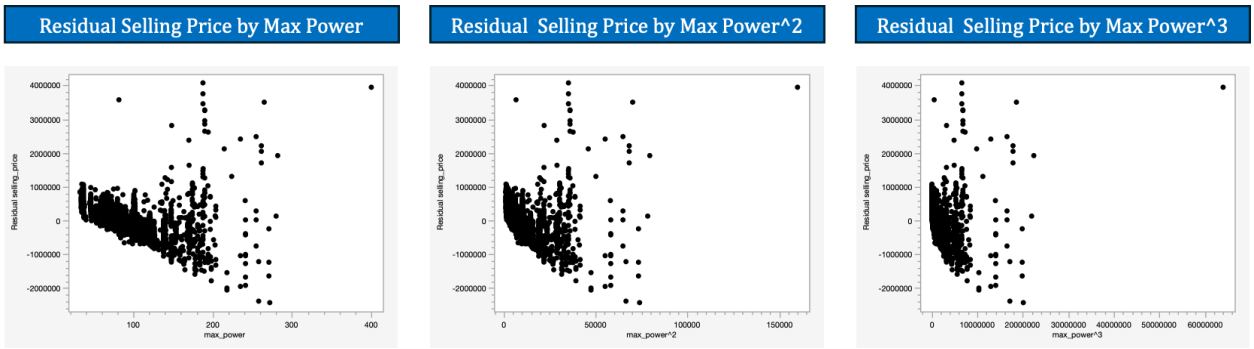
Model 3 shows an increase in the adjusted R^2 value compared to both Model 1 and Model 2. This suggests that Model 3 is a better representation of the variation present in the data

Figure f2: Model 3 – Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.634e+8	3476097	-47.02	<.0001*
year	80784.74	1734.101	46.59	<.0001*
km_driven	-0.411265	0.09299	-4.42	<.0001*
mileage(km/ltr/kg)	8737.2701	1549.238	5.64	<.0001*
engine	564.19283	24.03063	23.48	<.0001*
max_power	3353.7645	320.0298	10.48	<.0001*
seats	-13508.93	6711.364	-2.01	0.0442*
(year-2013.98)*(year-2013.98)	4686.4954	204.1579	22.96	<.0001*
(engine-1458.71)*(engine-1458.71)	-0.538916	0.021847	-24.67	<.0001*
(max_power-91.5874)*(max_power-91.5874)	238.3109	4.763456	50.03	<.0001*
(max_power-91.5874)*(max_power-91.5874)*(max_power-91.5874)	-0.550525	0.02023	-27.21	<.0001*

This model tells us that for each year the vehicle is newer, the price increases by around ₹80,785. Each additional kilometer driven decreases the selling price by approximately ₹0.41. Each unit increase in fuel efficiency increases the price by ~ ₹8,737. Each unit increase in engine size increases the price by ~ ₹564.19. Each additional horsepower increases the price by ~ ₹3,354. Each additional seat decreases the price by ~ ₹13,509.

Figure f3: Model 1-3 – Residual Plot of Maximum Horsepower



Comparison of the 3 Models:

	Model 1	Model 2	Model 3
RSquare	0.640136	0.748235	0.769827
RSquare Adj	0.639863	0.747948	**0.769535**
Root Mean Square Error	488242.5	408457.4	390574.8
Mean of Response	649813.7	649813.7	649813.7
Observations (or Sum Wgts)	7906	7906	7906

Out of the three models, Model 3 has the highest adjusted R^2 value as well as the lowest root mean square error, suggesting that Model 3 offers the strongest predictive performance and most accurately captures the relationship between variables in the data.

Singularity & Multicollinearity Tests

Figure g1:

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-1.634e+8	3476097	-47.02	<.0001*	.
year	80784.74	1734.101	46.59	<.0001*	2.3262056
km_driven	-0.411265	0.09299	-4.42	<.0001*	1.4452539
mileage(km/ltr/kg)	8737.2701	1549.238	5.64	<.0001*	2.0262293
engine	564.19283	24.03063	23.48	<.0001*	7.5980228
max_power	3353.7645	320.0298	10.48	<.0001*	6.7820137
seats	-13508.93	6711.364	-2.01	0.0442*	2.1475381
(year-2013.98)*(year-2013.98)	4686.4954	204.1579	22.96	<.0001*	1.5752795
(engine-1458.71)*(engine-1458.71)	-0.538916	0.021847	-24.67	<.0001*	4.3547183
(max_power-91.5874)*(max_power-91.5874)	238.3109	4.763456	50.03	<.0001*	11.131764
(max_power-91.5874)*(max_power-91.5874)*(max_power-91.5874)	-0.550525	0.02023	-27.21	<.0001*	5.8642632

Model 3 does not have singularity issues; however, there may exist concerning VIF values (*highlighted in blue in Figure g1)*.

Taking a closer look at the terms that indicate alarmingly high VIF values (around 4 and above), we notice that they are the squared and cubed versions of other terms already present in the data. This would make sense for why there would be a correlation. Additionally, engine and max power VIFs are also high.

This makes sense since a more efficient engine would likely result in or be connected to higher horsepower. Overall, our VIF values do **NOT** serve as a concern.

Out of the quantitative variables, kilometers driven, fuel efficiency, and number of seats follow a linear relationship, while year and engine size follow a squared quadratic relationship, and maximum horsepower best fits a cubic relationship.

Key Takeaways:

Sellers within and those hoping to enter the Indian car market should take into consideration the following variables: year of manufacture, kilometers driven, fuel efficiency, engine size, maximum horsepower, fuel type, transmission, and seller type. The number of seats a vehicle has also has an impact on selling price; however, out of all the variables, it is relatively less significant as a determinant of selling price. Sellers should take into account the patterns each variable follows and determine if it is worth the investment. For instance, if the automaker is aiming to target the budget automobile industry, they should focus on a specific range of maximum horsepower rather than produce vehicles that fall into the middle range, where we see a slowed increase in selling price per horsepower. Additionally, sellers should ensure that their vehicles are sold as close to the manufacturer's date as possible, since after that, there is an exponential decrease in the selling price. Similarly, with engine size, the larger the engine, the higher its compression ratios, and so its power output increases exponentially with size. A producer should invest in engine size as it results in exponential price returns. Taking these into account, the automaker will have a more data-driven decision-making process regarding which sector of the market they want to target and how they can best fulfill their needs.