CS-333 Assignment-2 Exploratory Data Analysis

Darshit Amit Pandya

Used cars are a goto option for people targeting big brand cars within a limited budget. I was interested in analyzing the used car sales data in order to understand what specific type of car pulls off the highest selling price. Accordingly, my initial question was: **What specific type of (used) car pulls off the highest sales value based upon certain automobile attributes?**

Dataset:

Through a brief online search on the major dataset websites, I found the following dataset to be quite effective to answer my initial question: **Used Car Prices (test_data.csv)** (https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction?select=train-data.csv).

Metrics:

- **Rows/Values:** 6019
- **Columns/Attributes:** 14

Attribute Description:

Attribute Name	**Attribute Type**	**Attribute Definition**
F1	Ordinal	Sr. No. / Index of the Entry
Name	Nominal	Model Name
Location	Nominal	Location
Year	Interval	Year of Manufacture
Kilometers Driven	Quantitative	Odometer measurement
Fuel type	Nominal	Type of Fuel (Petrol/Diesel/CNG/LPG/Electric)
Transmission	Nominal	Automatic/Manual Transmission
Owner type	Nominal	First/Second/Third/Fourth Owner
Mileage	Quantitative	Mileage
Engine	Quantitative	Engine capacity (cc)
Power	Quantitative	Engine Power (bhp)
Seats	Nominal	No. of Seats
New Price	Quantitative	Price of the same model if purchased brand new
Price	Quantitative	Price of the used car

Glossary:

- **1 Lakh INR (Indian Rupees) = 100,000 INR (Indian Rupees)**
- **INR = Indian Rupees (1 US $ = 83 INR Appx.)**
- **Km = Kilometer (1 Mile = 1.6 Km Appx.)**

Data Preprocessing & Data Cleaning:

To begin with analyzing the data, I checked for the completeness of the data for each attribute in the dataset. Following are the observations and the respective actions taken to ensure data integrity:

- **Attribute - "F1**": This attribute represents the Sr. no. or the index of the entries in the dataset. Hence, this complete attribute has been ignored throughout the data analysis and visualization phase as it doesn't impact or contribute to the questions asked throughout this process.

- **Attribute - "Name"**: This attribute contains a string that includes the name of the manufacturer and the model name of the used car. It doesn't significantly contribute to the initial question and hence, it has been ignored for the purpose of visualization. Had it been just the manufacturer name, I would have considered exploring it in detail.

- **Attribute - "New Price**": This attribute has 86% missing/null values. This is due to the fact that a model which might be available in 2012, might not be available in the market and could have been discontinued by the manufacturer. Hence, a New Price entry could not be created for the same. Accordingly, this attribute has been ignored throughout the analysis and visualization phase as 86% values are missing.

- **Attribute - "Engine"**: Around 1% of the values are missing for this attribute. As the amount of missing values are quite negligible, only the missing values (1%) have been ignored throughout the analysis and visualization process and the rest 99% values have been considered.

- **Attribute - "Power"**: Around 1% of the values are missing for this attribute. As the amount of missing values are quite negligible, only the missing values (1%) have been ignored throughout the analysis and visualization process and the rest 99% values have been considered.

- **Attribute - "Seats"**: Around 1% of the values are missing for this attribute. As the amount of missing values are quite negligible, only the missing values (1%) have been ignored throughout the analysis and visualization process and the rest 99% values have been considered.

- **Attribute - "Mileage"**: This attribute cannot be used as we cannot compare mileage for CNG/LPG (per Kg) to Petrol/Diesel (per Liter) as their costs for Kg and Liter are on a different scale.

Exploratory Data Analysis:

To begin with, I plotted the distribution of the sale price of the used cars from the dataset in order to understand the bandwidth of the price, i.e., to understand the overall price-range and to identify outliers through the distribution of data. In order to get a complete idea of the distribution of the price of the used cars, I used the Box Plot as well as the Histogram, as found in Fig. 1(a) and Fig. 1(b) below.

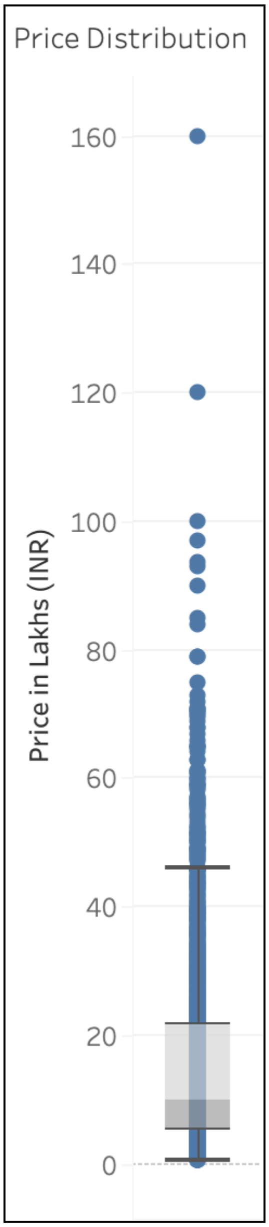

Fig. 1(a)

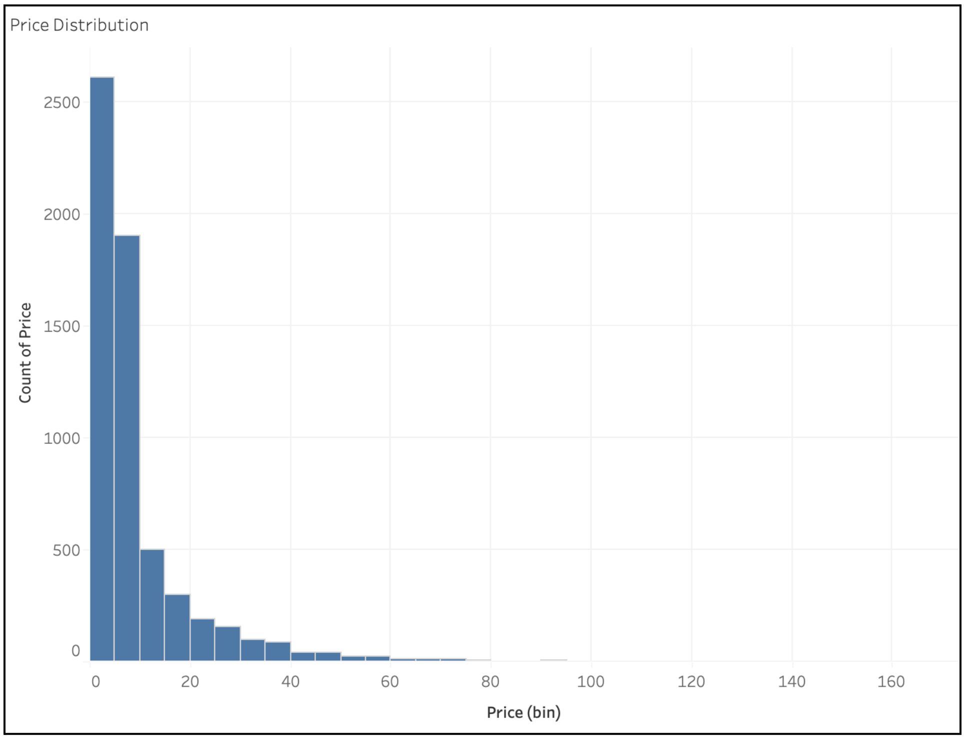

Fig. 1(b)

This plot clearly demonstrated that the distribution is skewed towards the lower-end of the values. Selling Price of most used cars is less than 45 Lakhs (INR). Also, it is quite evident that certain cars were also sold as high as 160 Lakhs (INR). Accordingly, I understood that there exists some quantity of outliers within this data. Moreover, on observing carefully, the minimum selling price was 0.44 Lakh INR. Also, the section from 0.44 Lakh to around 10 Lakhs is quite dense. However, it would be impossible to know how the price varies without considering certain major factors related to the car's valuation.

Now, one major factor contributing to the selling price can be the odometer reading. As a general perception, the greater the odometer reading, the lesser the price as it indicates that the car's parts have been used more and hence, they may have deteriorated over time and would need replacement. To check the correlation between the selling price and the odometer reading I plotted the following scatter-plot.

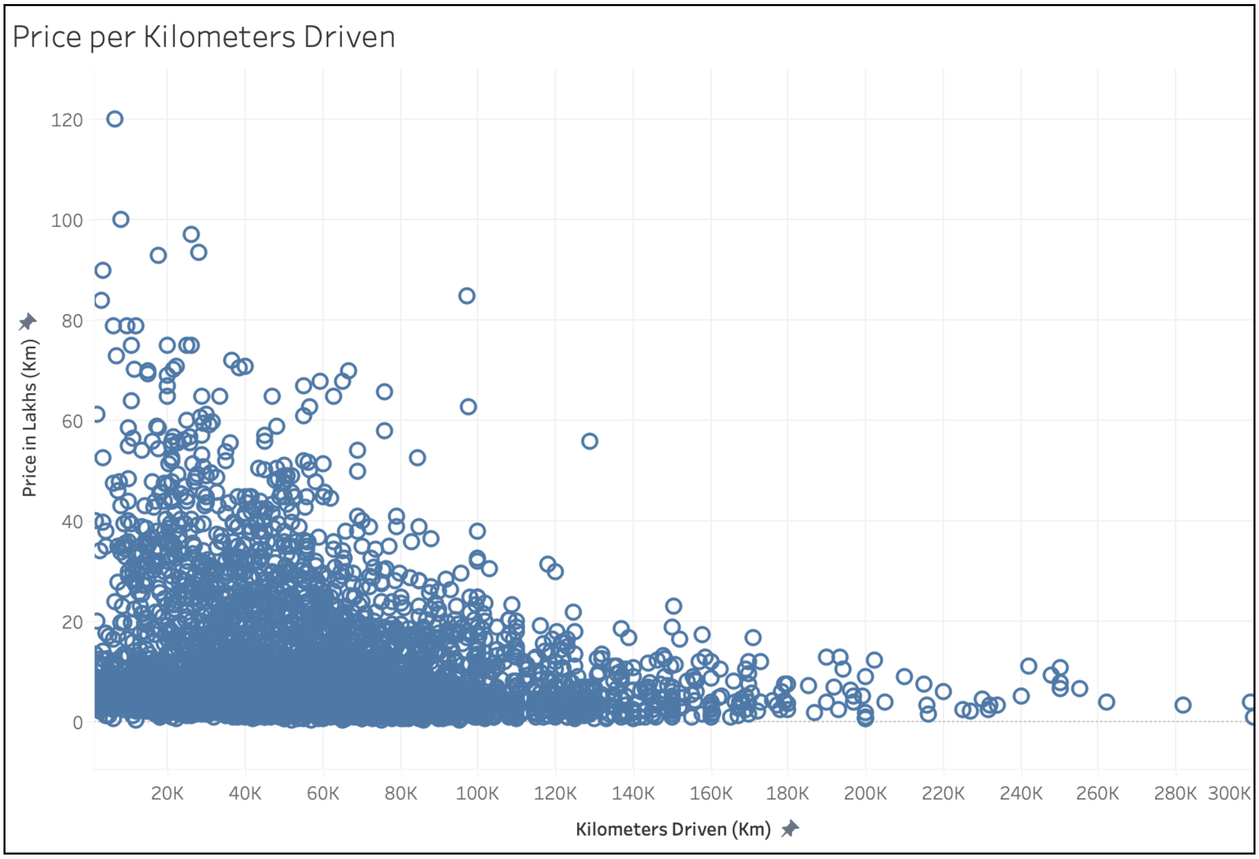

Fig. 2

As expected, the selling price is decreasing consistently, as seen in Fig. 2 above. The highest range of prices occurs where the kilometers driven are less than 20K and the lowest range of prices occurs as we move towards the right-side of the plot. Accordingly, it is evident that this attribute significantly contributes to the selling price of the used cars.

However, there are certain low price values even on the left-side of the plot (less kilometers driven). One reason that I can think of is the other factors that might contribute to the price. For example, a car which has been driven for less than 10K kilometers can have a lower price if it is a second-hand car or due to less seats (small size) or due to being manufactured 30 years back or the combination of any such examples. Accordingly, this correlation cannot be fully relied upon for decision-making. Hence, I started to explore other major correlations affecting the sale price. One such fact might be the ownership category. For example, as a general perception, the first-owner's selling price would be much higher and would decrease as the number of owners increase pertaining to the fact that the car might be used more in terms of kilometers driven and also the condition of the car. Based upon this inference, I decide to rephrase my initial question: **In addition to the odometer reading, can the ownership type of the car contribute to the selling price?** To decode this, I plotted the average sale price against the ownership category.

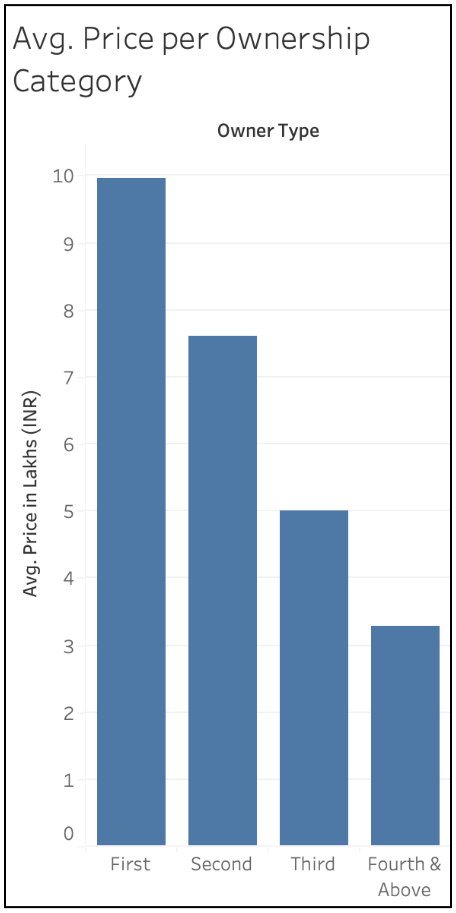

Fig. 3(a)

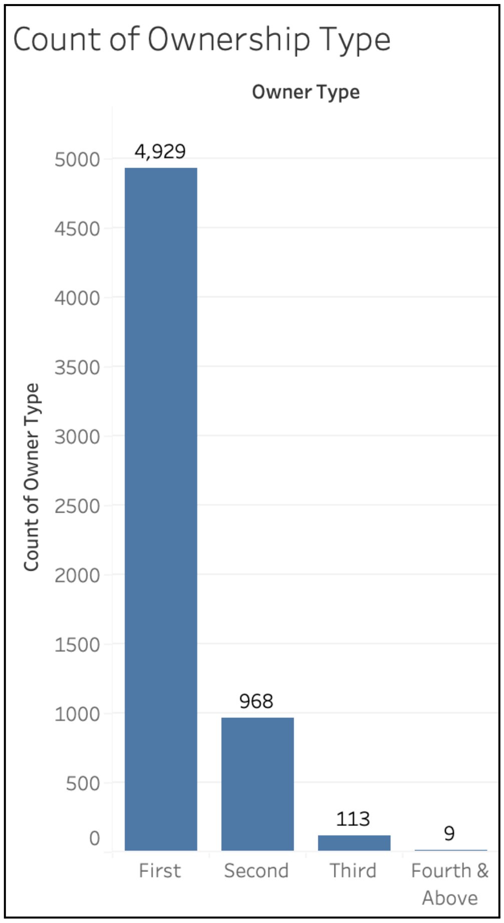

Fig. 3(b)

As seen in Fig. 3(a) above, the trend is inline with my prediction. Hence, this feature is a significant contributing factor to the final sale price. However, on observing the histogram of the ownership records (Fig. 3(b)), it is evident that the categories “Fourth & Above” and “Third” have negligible values and don't contribute significantly to generalize the price trend. Accordingly, I considered only “First” and “Second” categories for the final visualization.

Now, another significant (and obvious) contributor to the selling price is the size of the car, basically, it is understandable that a SUV will cost more than a Hatchback. Based upon this inference, I decided to rephrase my question: **In addition to the odometer reading and the type of ownership, can the size of the car (no. of seats) contribute to the selling price?** Accordingly, in order to analyze this pattern, I plotted the following plots of avg. price against the number of seats.

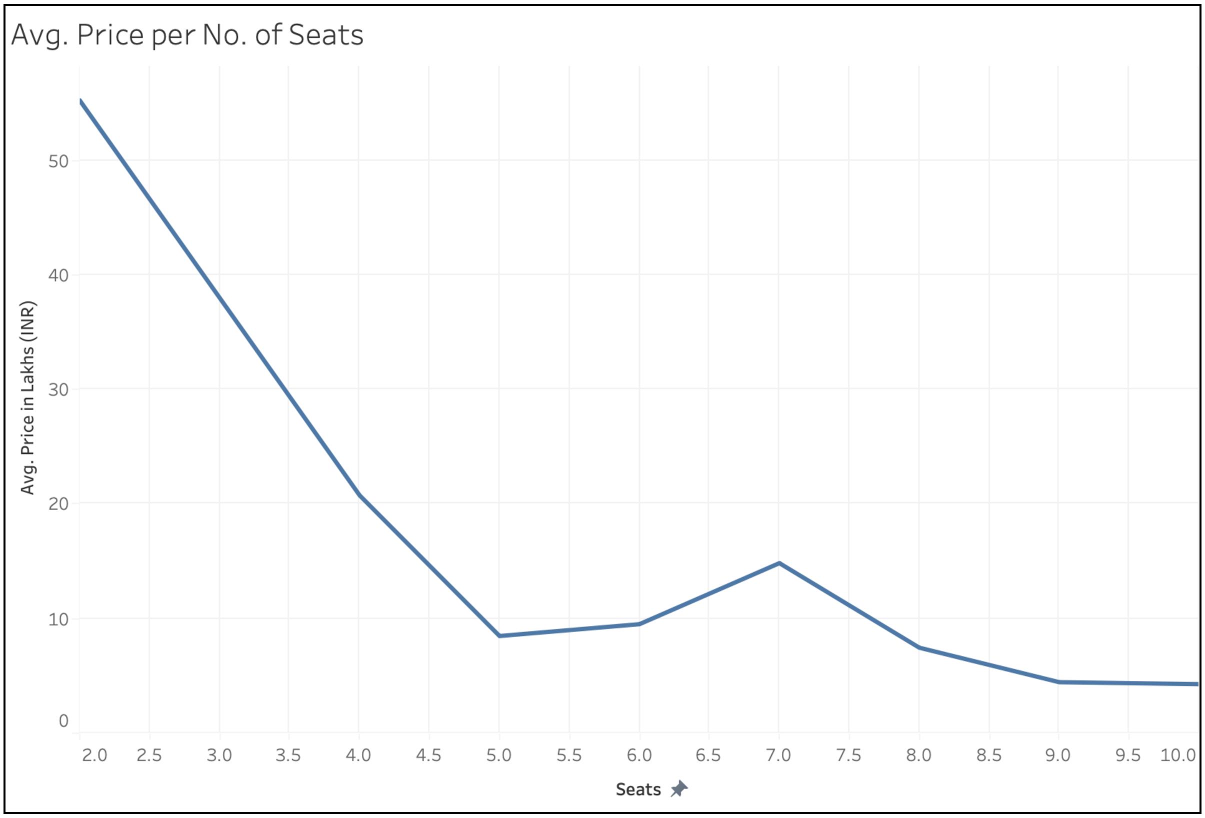

Fig. 4(a)

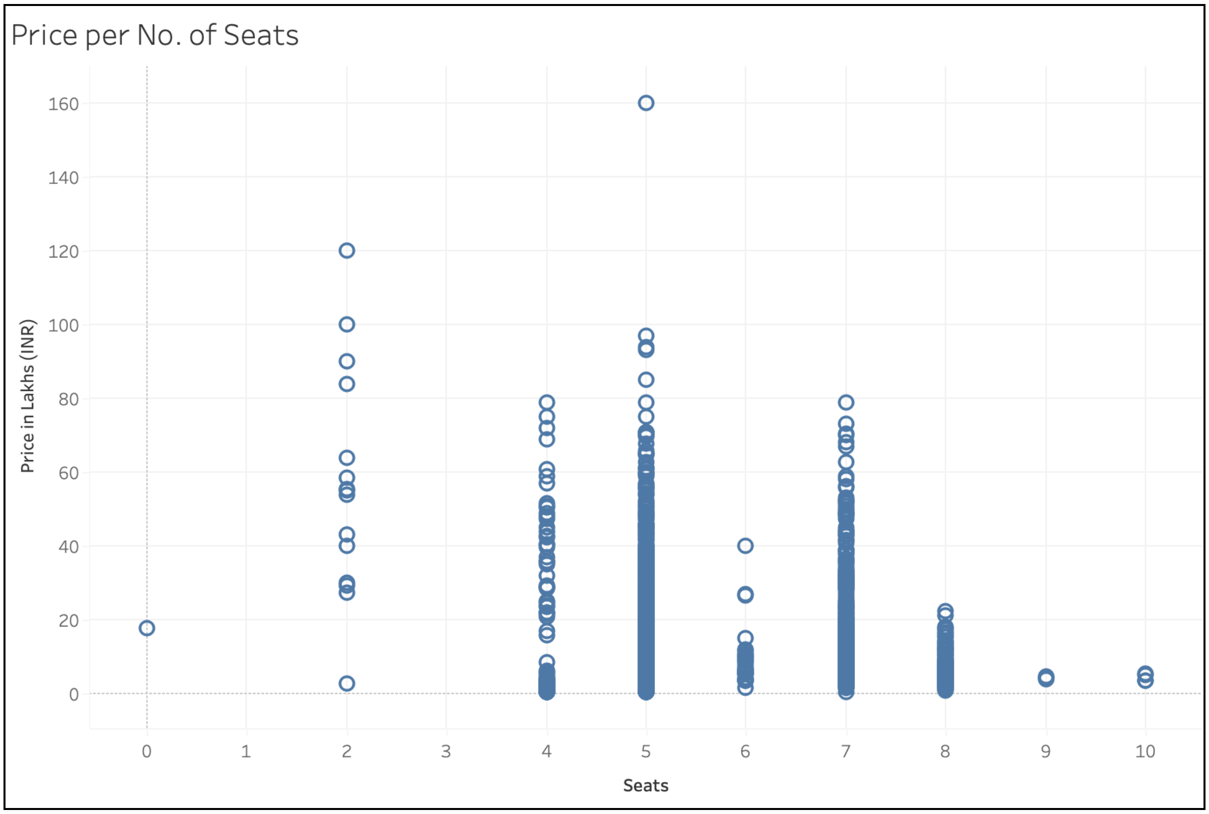

Fig. 4(b)

As seen in Fig. 4(a), there is a reverse trend - the average price decreases as the number of seats (or size of vehicles) increases. Also, as seen in Fig. 4(b), there's no significant correlation between the size of the car and the price. One reason for this is that the cars having more seats might be old and might be of fourth-ownership or beyond, which might significantly decrease the value of the vehicle as seen in the Ownership plots (Fig. 3(a) and Fig. 3(b)) before. Accordingly, I generated a plot for this purpose.

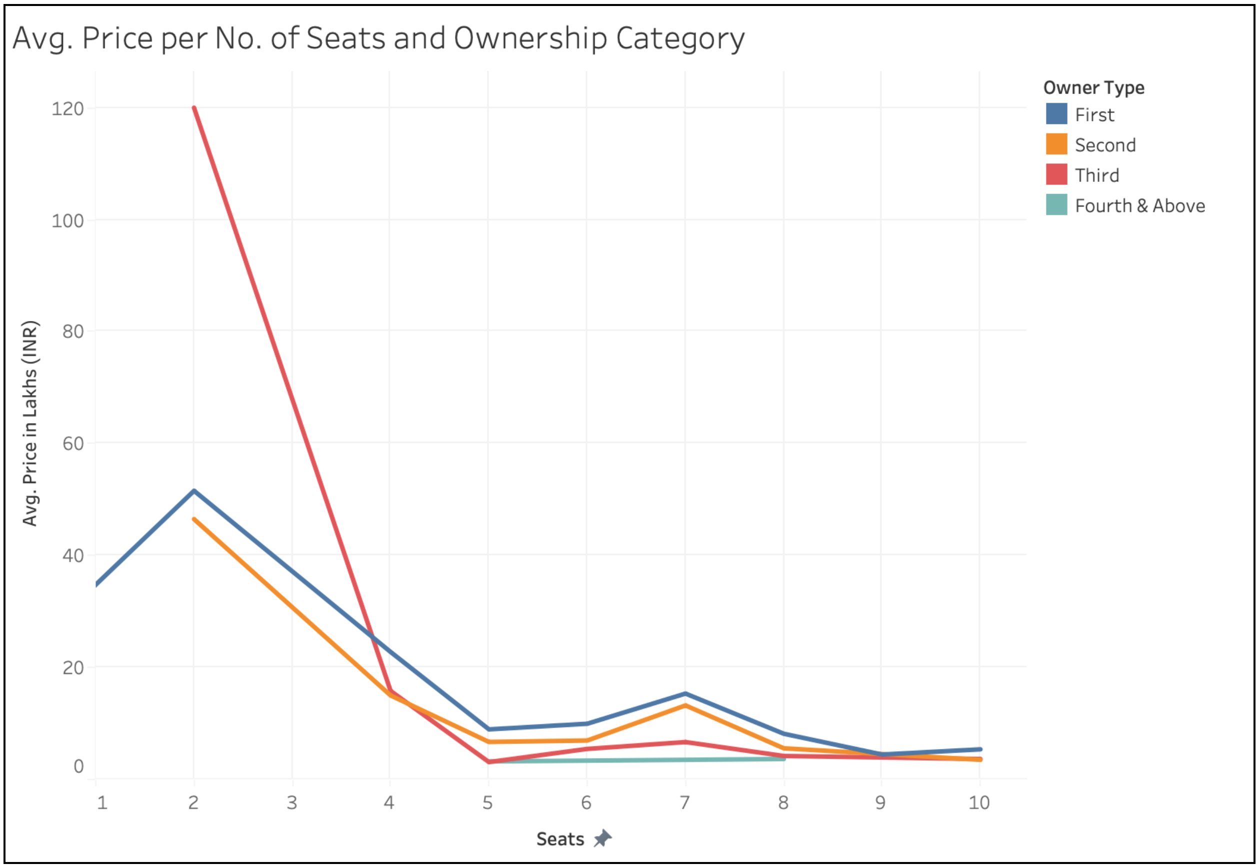

Fig. 5

However, as seen in Fig. 5 above, the Ownership Category doesn't affect the trend, as the trend is still decreasing. Accordingly, I reached a conclusion that the no. of seats is not a correct indicator of Price as I am unable to find a substantial reason to back this reverse trend. Hence, it's better to ignore the seats attribute.Furthermore, one major factor that can impact the selling price is the Year of manufacture, which signifies how old a car is. As a general perception, the older the car the lesser the price. Based upon this inference and the fact that size(no. of seats) isn't a major contributing factor to the selling price, I decided to **revert** to my previous question and transform it to: **In addition to the odometer reading and the type of ownership, can the manufacturing year of the car contribute to the selling price?** Following this preposition, I plotted the relevant scatter-plots as seen below. (I am considering the average selling price, as the price distribution over the years doesn't contain any significant number of outliers.)

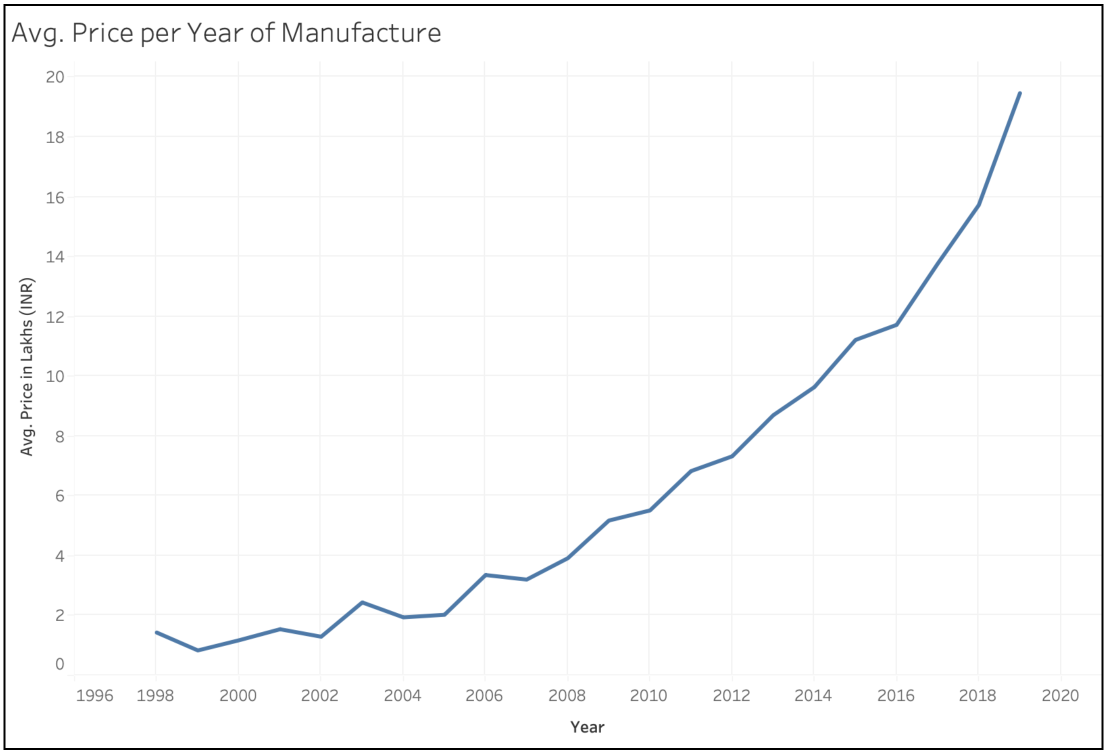
Avg. Price per Year of Manufacture
Avg. Price in Lakhs (INR)
20
18
16
14
12
10
8
6
4
2
0
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year

Fig. 6(a)

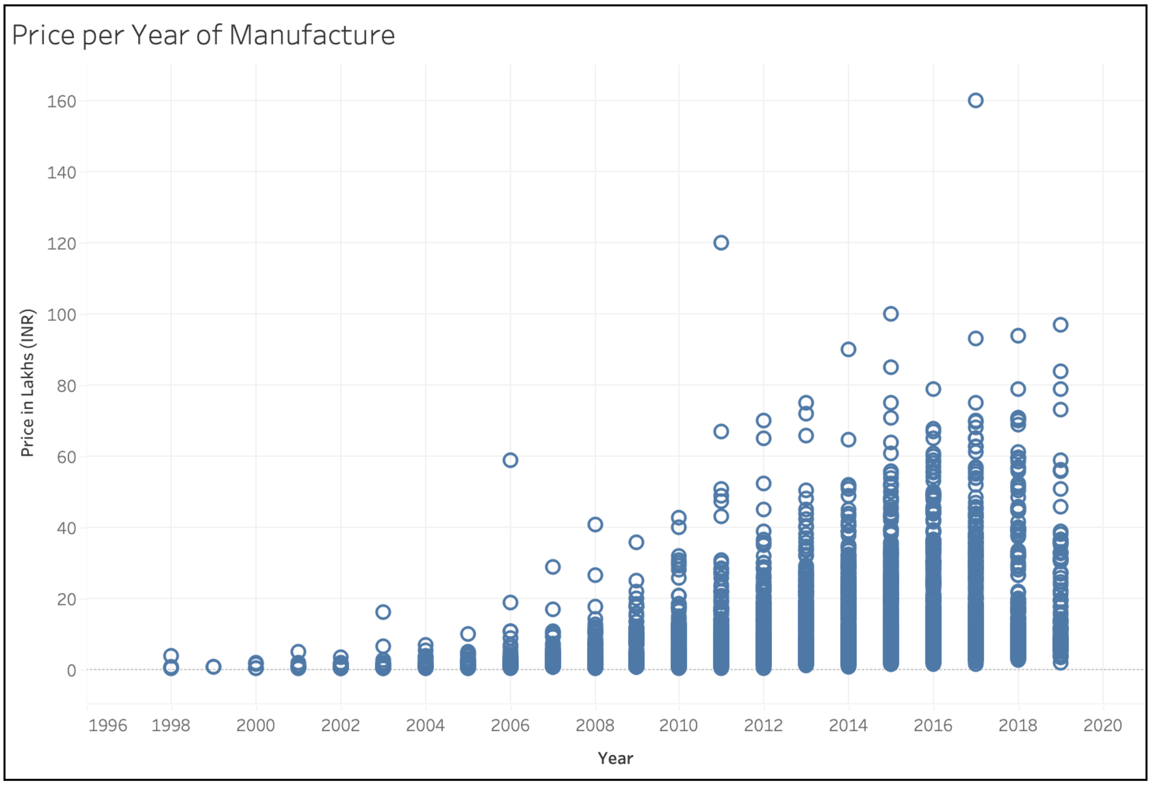
Price per Year of Manufacture
Price in Lakhs (INR)
160
140
120
100
80
60
40
20
0
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year

Fig. 6(b)

As seen in Fig. 6(a) and Fig. 6(b) above, the higher prices occur as the Year of Manufacture approaches. However, there are certain points on the right-side of the Fig. 6(b) where the price is low. This can be due to other impacting factors including Ownership Type or others. After confirming Year as the significant contributor to the selling price, another significant factor is the Fuel Type. Based upon this inference, I decided to rephrase my question: **In addition to the odometer reading, the type of ownership, and the year of manufacture, can the car's fuel type contribute to the selling price?**

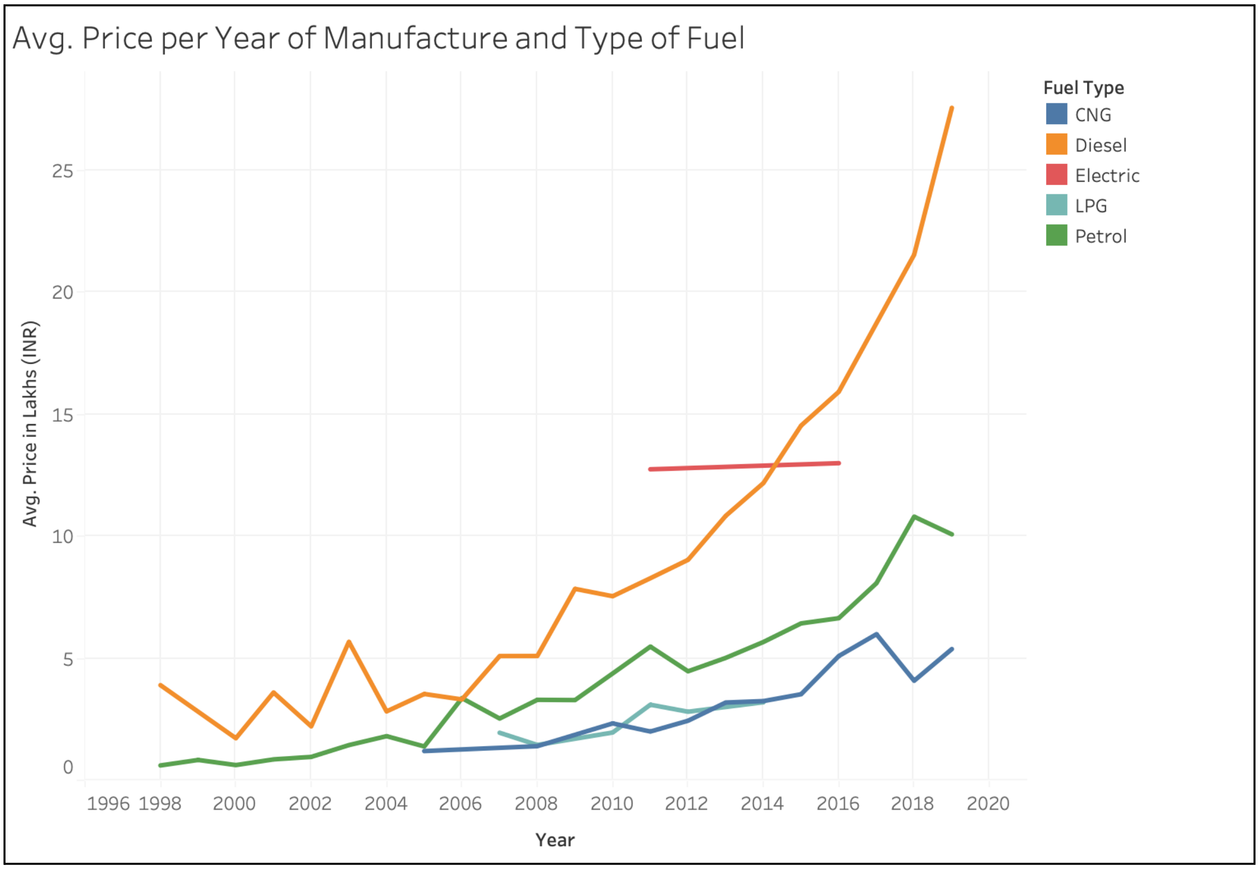

Fig. 7

The above plot (Fig. 7) adds to the fact that the Fuel Type also follows the similar avg. price trend as seen in the previous plot. As it confirms the trend, Fuel Type can be considered a significant contributing factor to the selling price. Moreover, Transmission can play a major role in impacting the Selling Price. For Example, in the USA, people prefer Automatic Transmission; however, in India, people prefer to choose Manual Transmission. As this dataset is from India, I believe Manual Transmission sale would be higher as people sometimes do consider the fact that the maintenance/repair cost of an Automatic Transmission car might cost more as it is not fixable by the local garages and needs the manufacturer's involvement which increases the cost of service of car. Based upon this inference, I decided to rephrase my question: **In addition to the odometer reading, the type of ownership, the year of manufacture, and the car's fuel type, can the car's transmission type contribute to the selling price?** The following plot extends the Fuel Type plot (Fig. 7) above.

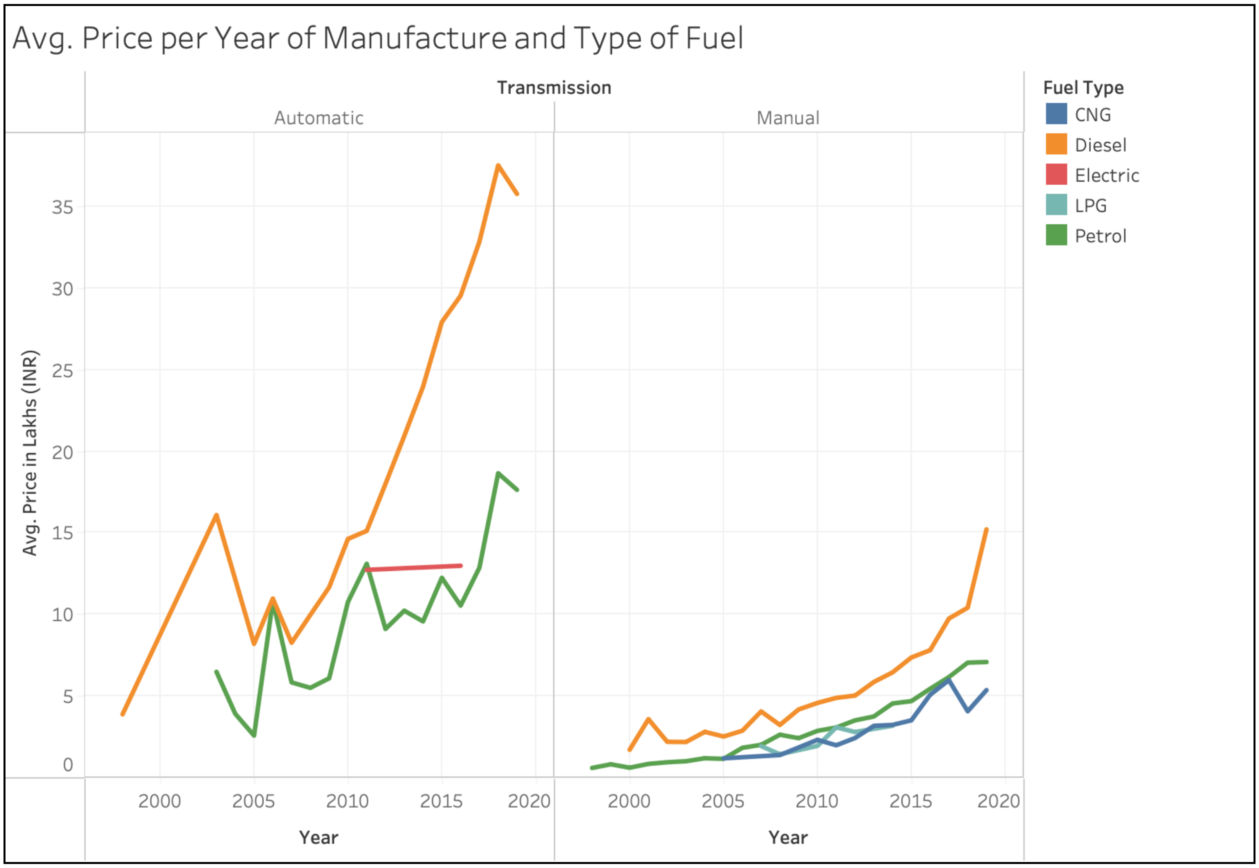
Avg. Price per Year of Manufacture and Type of Fuel
Transmission
Automatic
Manual
Avg. Price in Lakhs (INR)
35
30
25
20
15
10
5
0
2000
2005
2010
2015
2020
Year
Fuel Type
CNG
Diesel
Electric
LPG
Petrol

Fig. 8(a)

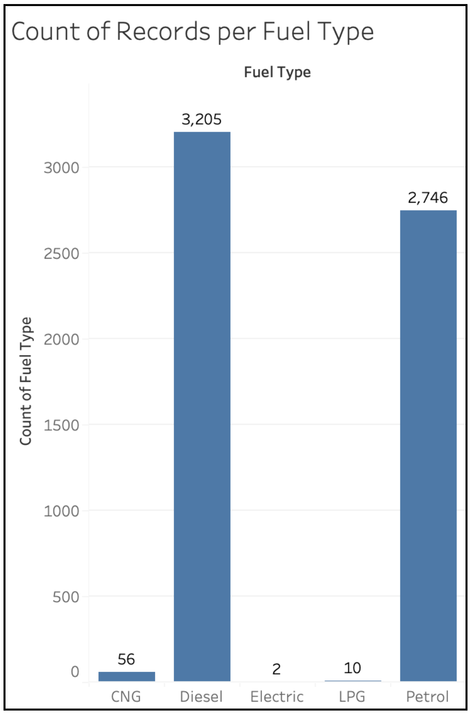
Count of Records per Fuel Type
Fuel Type
Count of Fuel Type
3000
2500
2000
1500
1000
500
0
56
3,205
2
10
2,746
CNG
Diesel
Electric
LPG
Petrol

Fig. 8(b)

It is evident from Fig. 8(s) that the trend still continues in accordance to the previous plots, i.e., the increase in price as the year of manufacture nears the present time. As seen in Fig. 8(b), there is significantly less quantity data for CNG, LPG, and Electric Fuel Type. It can be safely concluded that the major focus should be concentrated upon the Petrol and Diesel Fuel types in order to answer the question.

Other major factors that can contribute to the pricing are the Engine cc and the Power Rate. However, as the Power Rate (in bhp) technically represents the actual/practical output of the theoretical Engine cc values, it is understandable that a buyer would focus more on the Power Rate (bhp) part than the Engine cc aspect. Using both attributes creates redundancy. Accordingly, I have selected Power Rate among the two measures. As a general perception, as the Power of the vehicle increases, the price should increase. Based upon this inference, I decided to rephrase my question: **In addition to the odometer reading, the type of ownership, the year of manufacture, the car's fuel type, and the car's transmission type, can the car's power rating contribute to the selling price?** Accordingly, two plots have been plotted below to understand the contribution of Power Rate to the selling price.

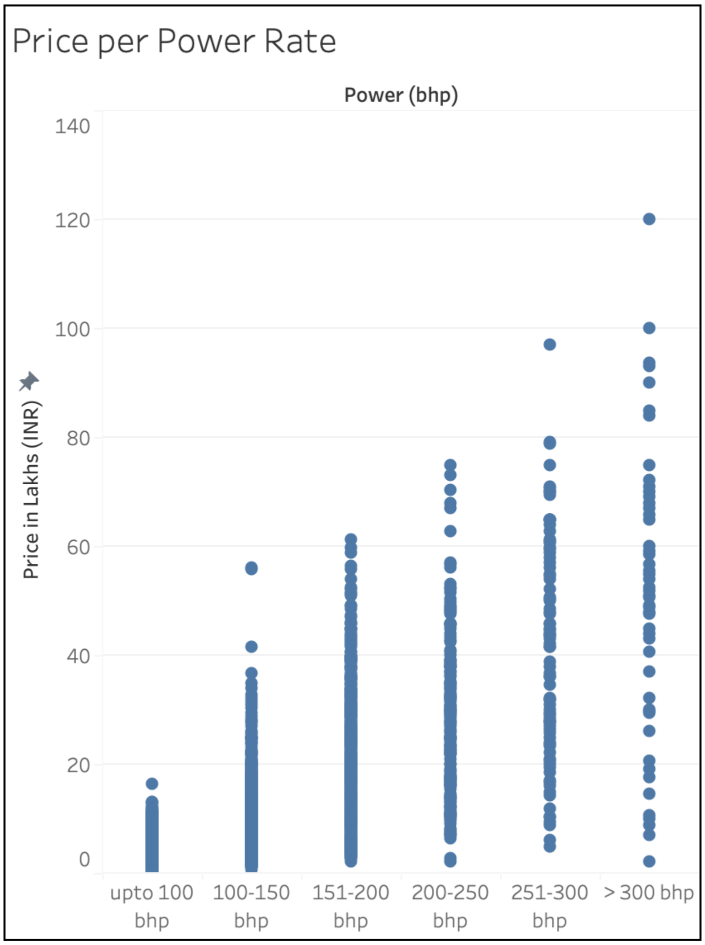

Fig. 9(a)

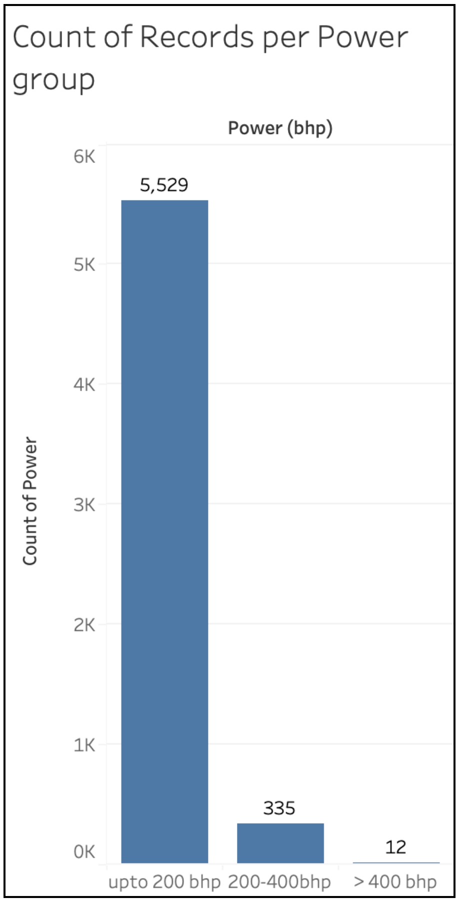

Fig. 9(b)

As evident from the above plot (Fig. 9(a)), higher price values occur for higher Power Rates (in bhp). However, there are certain lower values on the right pertaining to the combination of various other factors. Still, as per the trend in Fig. 9(a), it can be safely said that the Power Rate is a good factor to include. Moreover, in Fig. 9(b), it can be seen that the amount of values available for >400 bhp Power Rate is quite low and would not contribute significantly to the selling price. Accordingly, I have considered bhp values for Power Rates up to 400 only.

Finally, location is the factor that is left to be explored. Based upon this inference, I decided to rephrase my question: **In addition to the odometer reading, the type of ownership, the year of manufacture, the car's fuel type, and the car's transmission type, and the car's power rating, can location of sale be a contributing factor to the selling price?**

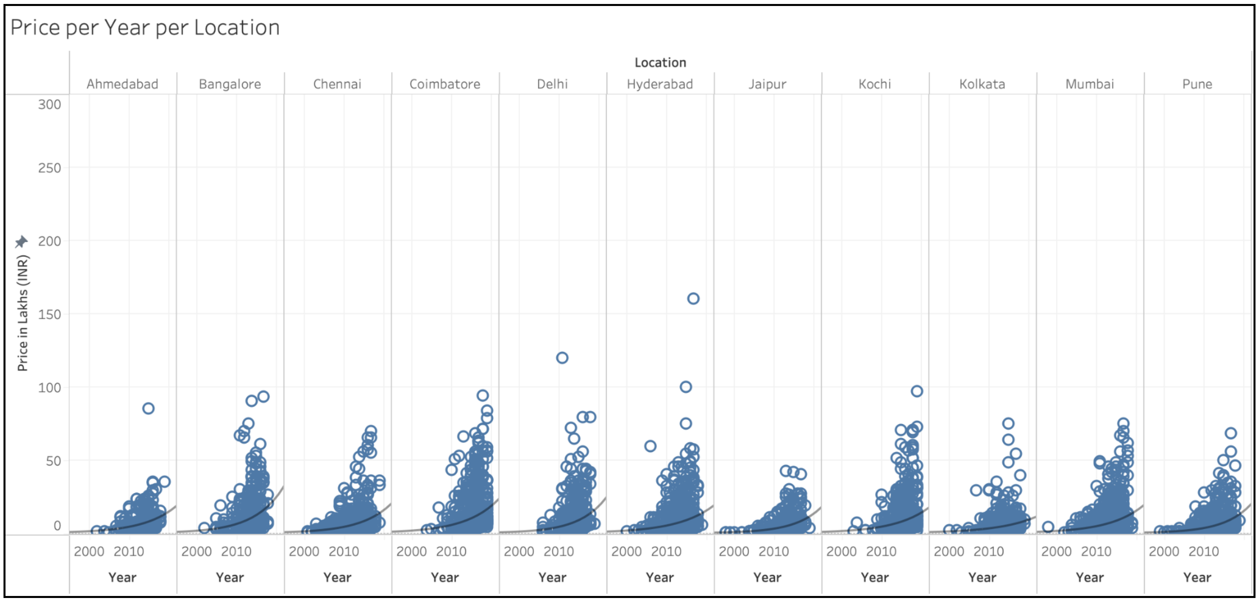

Fig. 10

As seen in the above plot (Fig. 10), apart from some outliers, the prices are similarly spread across the years and increase in all cities as the Year of Manufacture approaches. As there's a common Sales Tax (GST) for sales across states and cities in India, there generally isn't a huge difference in sale prices. Moreover, people wouldn't order a delivery of the used vehicle from other states if there's no significant price difference as the delivery cost would add up to the total price. Also, the trend lines across the sub-plots in Fig. 2 are similar. Accordingly, using location as a factor would increase redundancy and significantly reduce the effectiveness of the visualization. With this, I conclude my Exploratory Data Analysis (EDA).

Finally, to answer the initial question regarding the visualization of the dataset, a final visualization can be presented as below considering the major factors discussed above, namely: Price, Odometer Reading, Year of Manufacture, Ownership Type, Fuel Type, Transmission Type, Power Type. I merged the above mentioned attributes into the Trellis plots generated previously. The final plot is presented below:

Final Visualization Plot:

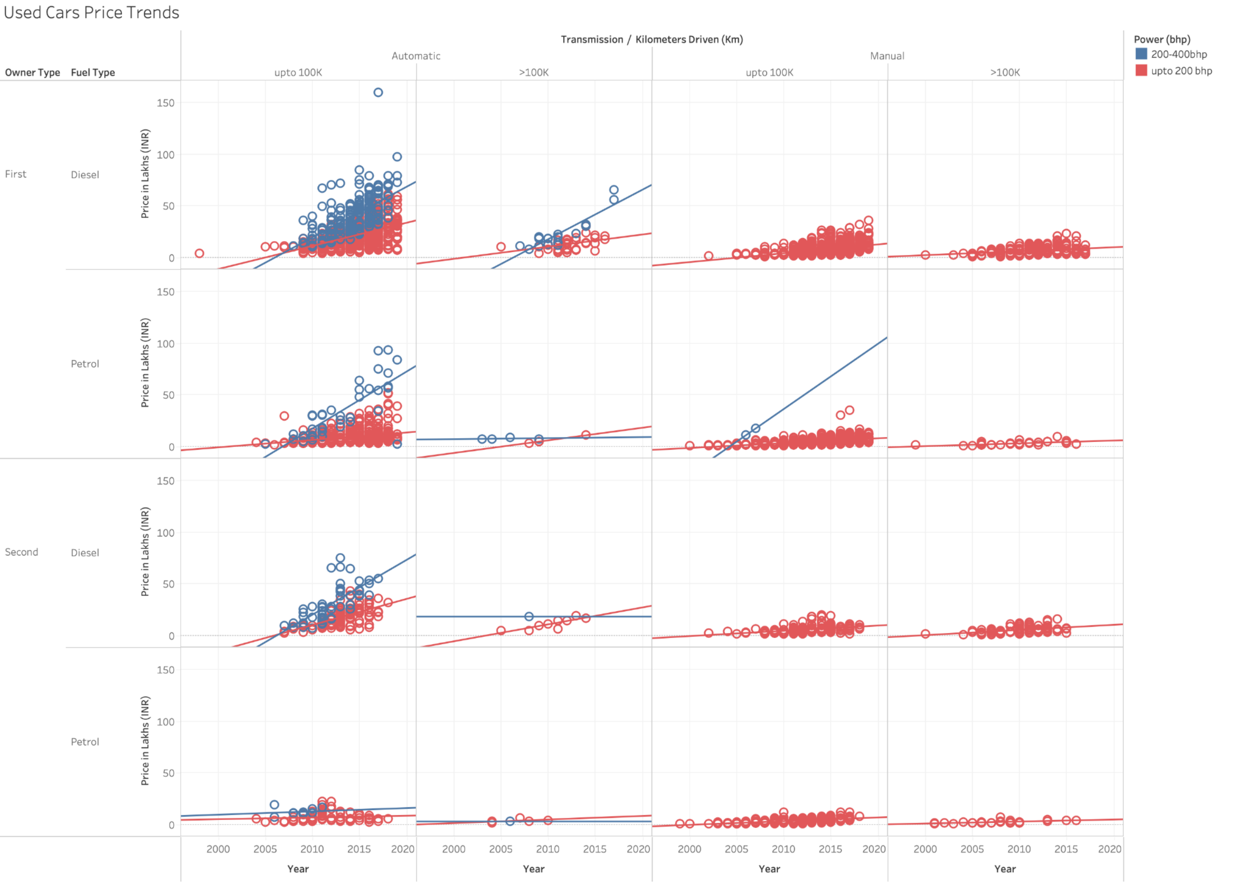

Fig. 11 Analyzing trends in Used Car Sale Prices. This visualization breaks down the sale prices of used cars, considering key factors such as year of manufacture, transmission type, fuel type, power type, ownership type, odometer reading, and their impact on the final sale price of the used car. Color shows details about Power type (bhp).

(Image Caption is mentioned at the bottom of the image.)

Final Plot Description: This plot (Fig. 11) visualizes all the factors contributing significantly to a used car's selling price in the form of a Trellis plot structure. It uses Year of Manufacturing as the X-axis, faceted by the Transmission type and the Odometer reading. It uses the Price as the Y-axis, faceted by the Ownership type and the Fuel type. The trend lines allow us to summarize the trends based on the particular attributes of a used car. **To answer the initial question (What specific type of (used) car pulls off the highest sales value based upon certain automobile attributes?),** it is evident from the plot that a used car having 200-400 bph power rate, a diesel engine, up to 100K odometer reading, first ownership, automatic transmission, and manufactured after 2018, would pull-off the highest selling value

Limitations: It is assumed that the manufacturer name does not affect the sale price. It might be the case that the manufacturer name, if given as a separate attribute, might affect the sale price of a used car as sometimes brand values determine the pre-owned sale prices. However, in this dataset, it was given as combined with the model name, which makes it difficult to infer from.