CARS4U

EXPLORATORY DATA ANALYSIS AND PRICING MODEL

by **JAKE EIDE**

BACKGROUND

Cars4U is a tech start-up operating in the Indian pre-owned car market. The business would like to come up with a pricing model that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing. For example, if the business knows the market price, it will never sell anything below it.

OBJECTIVE

1) Explore and visualize the Cars4U dataset.

2) Build a linear regression model to predict the prices of used cars.

3) Generate a set of insights and recommendations that will help the business.

DATA INFORMATION

The data contains information about 7253 used cars for sale by Cars4U.

VARIABLE	DESCRIPTION
S.No.	Serial Number.
Name	Name of the car which includes Brand name and Model name.
Location	The location in which the car is being sold or is available for purchase Cities.
Year	Manufacturing year of the car.
Kilometers_driven	The total kilometers driven in the car by the previous owner(s) in KM.
Fuel_Type	The type of fuel used by the car.
Transmission	The type of transmission used by the car.
Owner	Type of ownership.
Mileage	The standard mileage offered by the car company in kmpl or km/kg.
Engine	The displacement volume of the engine in CC.
Power	The maximum power of the engine in bhp.
Seats	The number of seats in the car.
New_Price	The price of a new car of the same model in INR Lakhs.
Price	The price of the used car in INR Lakhs.

PROBLEM DEFINITION

What are the main factors that determine the price of a used car in India?

The main problem to tackle is to create a pricing model that effectively predicts the price of used cars in the India market. Machine Learning algorithms will be used to build an effective model. Building an effective model is paramount, as Cars4U can use the model as a tool to maximize profit. A well-built model can have positive financial implications for the business.

To build the best possible model, our analysis will be centered on the “Price” statistic. What data affects price, and to what degree? I will show graphs that heavily impact price. The following exploratory data analysis will uncover meaningful relationships between variables as they relate to car prices.

EXPLORATORY DATA ANALYSIS - UNIVARIATE: **PRICE**

Cars4U prices have a large spread in values, but are still clustered around the median value.

Observations:

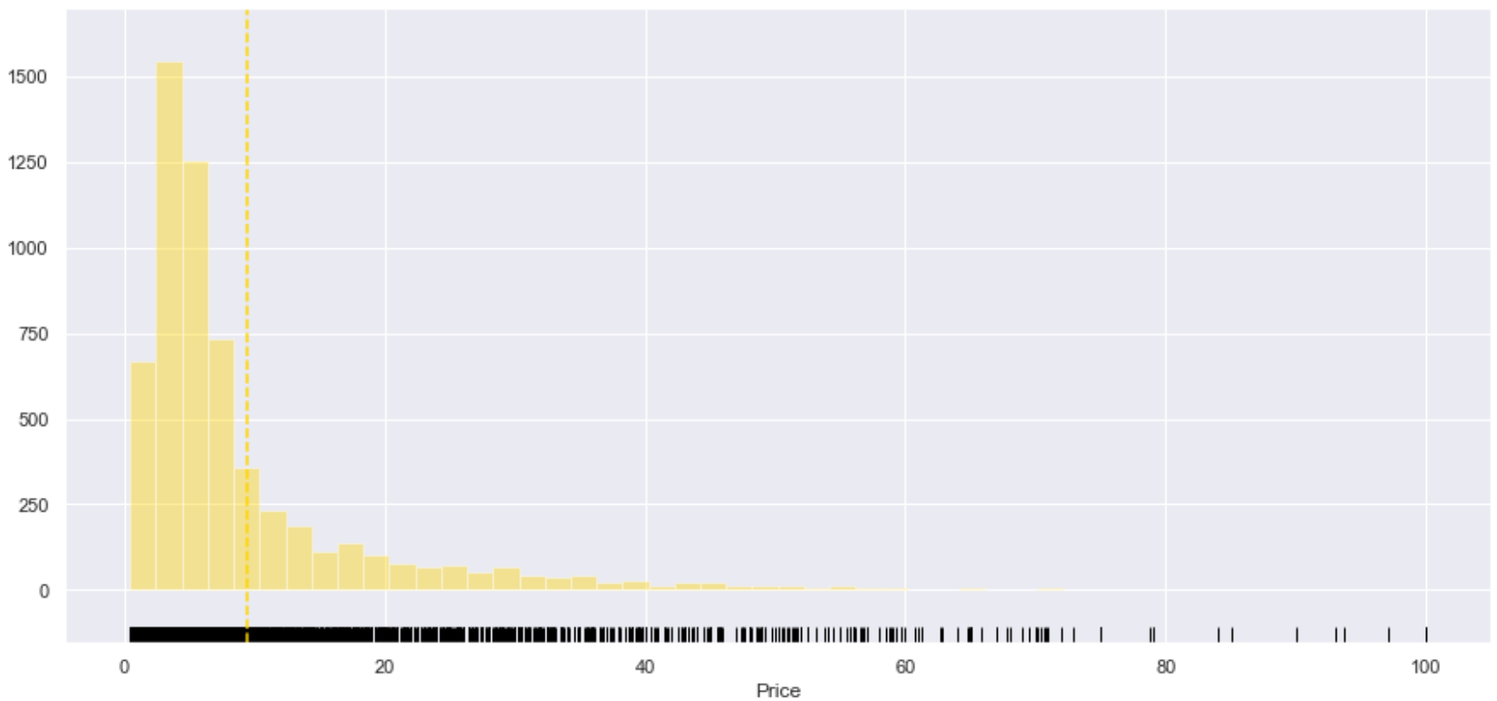
• Looking at the histogram and box plots to the right, we see that there is a large right skew to the data. These expensive vehicles make the mean price of cars much higher than the median price..

Observations on Central Tendency:

• Mean Price is 9.5 Lakhs
• Median Price is 5.6 Lakhs
• The middle 50% of cars are between 3.5 and 10 Lakhs

Insight for the business:

• The majority of your used cars will be priced in the 3 to 11 Lakhs range. These are the price points you can expect.
• Vehicles that are more expensive than 20 Lakhs are much more rare, but can have a large range of potential values.



EXPLORATORY DATA ANALYSIS - UNIVARIATE: **NEW PRICE**

Here we look at the price of a new car of the same model (priced in INR Laks).

Observations:

• We can see that the prices of new vehicles have a right skew that is similar to the prices of used vehicles. These expensive vehicles should be viewed as the exception, not the rule.

Observations on Central Tendency:

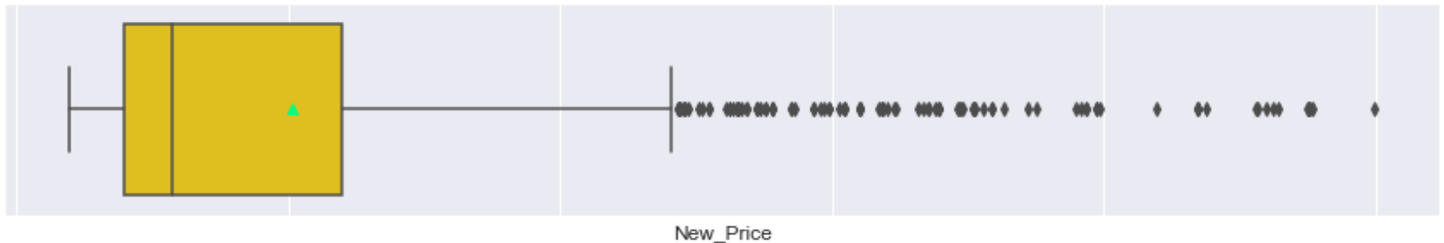
• Mean New Price is 20 Lakhs

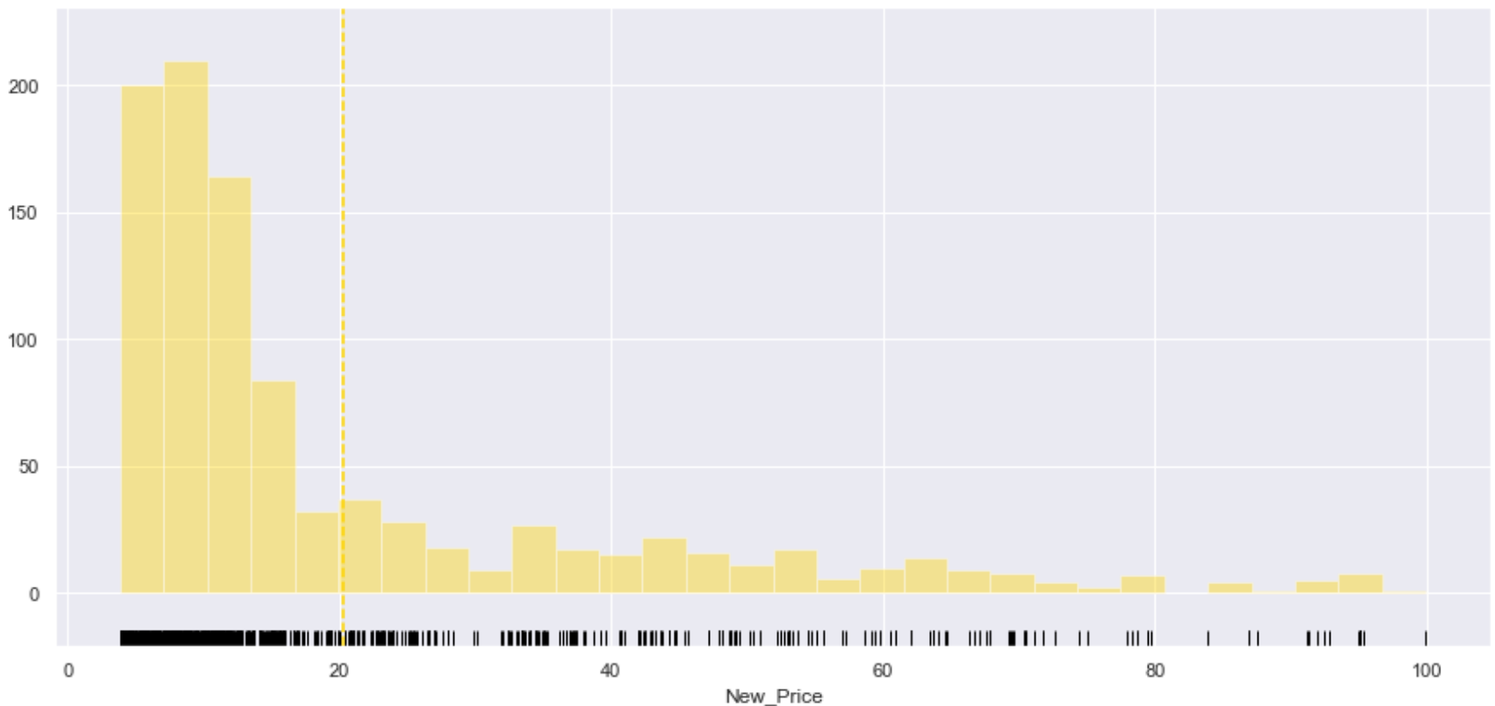
• Median New Price is 11.5 Lakhs

• The middle 50% of cars are between 7.9 and 24 Lakhs

Insight for the business:

• The majority of cars in this data set had a new price between 7 and 25 Lakhs. If you look at new car models that are priced within this range, these are the types of cars you can expect to fill your inventory.





EXPLORATORY DATA ANALYSIS - UNIVARIATE: KILOMETERS DRIVEN

Observations:

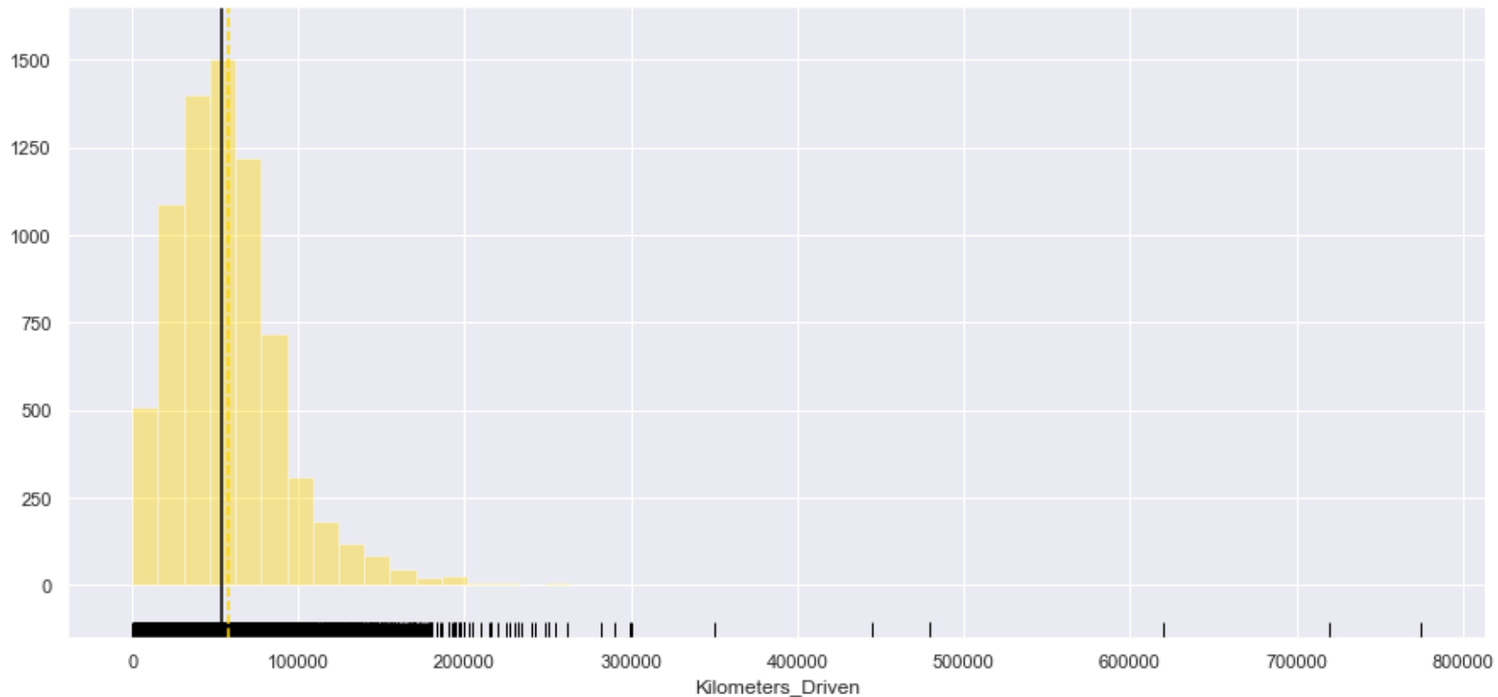
• There is a large amount of right skew, due to some cars having been driven a large number of kilometers.

Observations on Central Tendency:

• Mean is 58,000 km
• Median BMI is roughly 53,000 km
• The middle 50% of cars have been driven between 34,000 and 73,000 km

Insight for the business:

• The majority of your used cars will have between 30,000 and 80,000 km. These are the kinds of cars you can expect to see in your inventory most often.
• Vehicles that have more than 130,000 km are much more rare.



EXPLORATORY DATA ANALYSIS - UNIVARIATE: **ENGINE**

This is a measure of the displacement volume of the engine, measured in cc. This is an indication of the engine's size.

Observations:

• There is a very large range of engines among the inventory of Cars4U. The smallest engine has 72 cc whereas the largest has 5998 cc.

• The histogram on the right shows that there are some common sizes that reappear: 1197 cc, 1248 cc, and 1498 cc are the three most common engines.

Observations on Central Tendency:

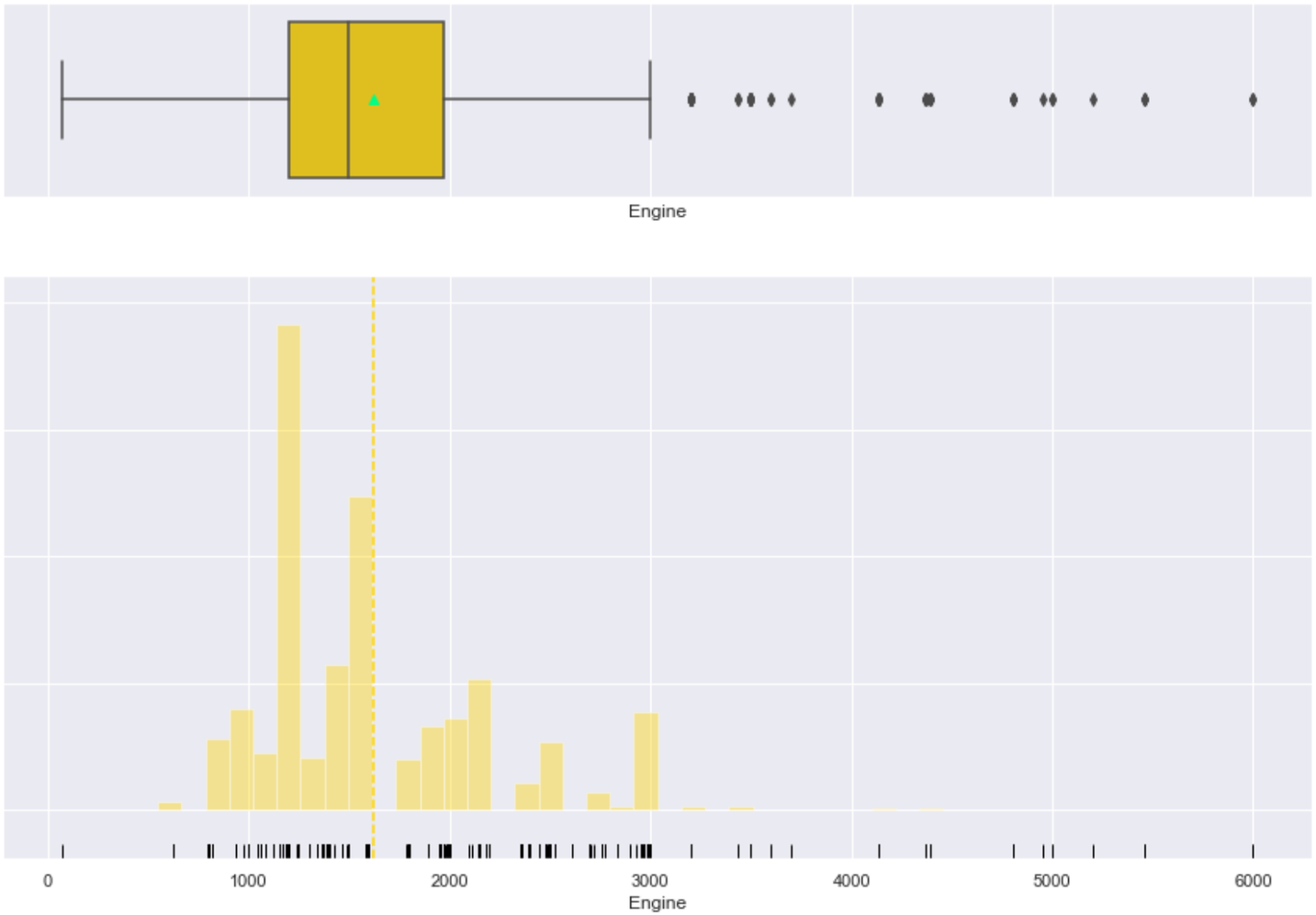
• Mean is roughly 1600 cc

• Median is 1500 cc

• The middle 50% of engines have displacement volumes between 1200 and 2000 cc

Insight for the business:

• The majority of your used cars will have between 1,000 and 2,000 cc. These are the kinds of cars you can expect to see in your inventory most often.



EXPLORATORY DATA ANALYSIS - UNIVARIATE: **POWER**

This is a measure of the maximum power of the engine, measured in bhp.

Observations:

• There are only five cars in the data set with bhp above 500. These five cars are made by Porsche, Jaguar, Bentley, and Lamborghini.

• There are some cars above 220 bph, which represent high-end luxury and sports cars. These kinds of cars are much less common.

Observations on Central Tendency:

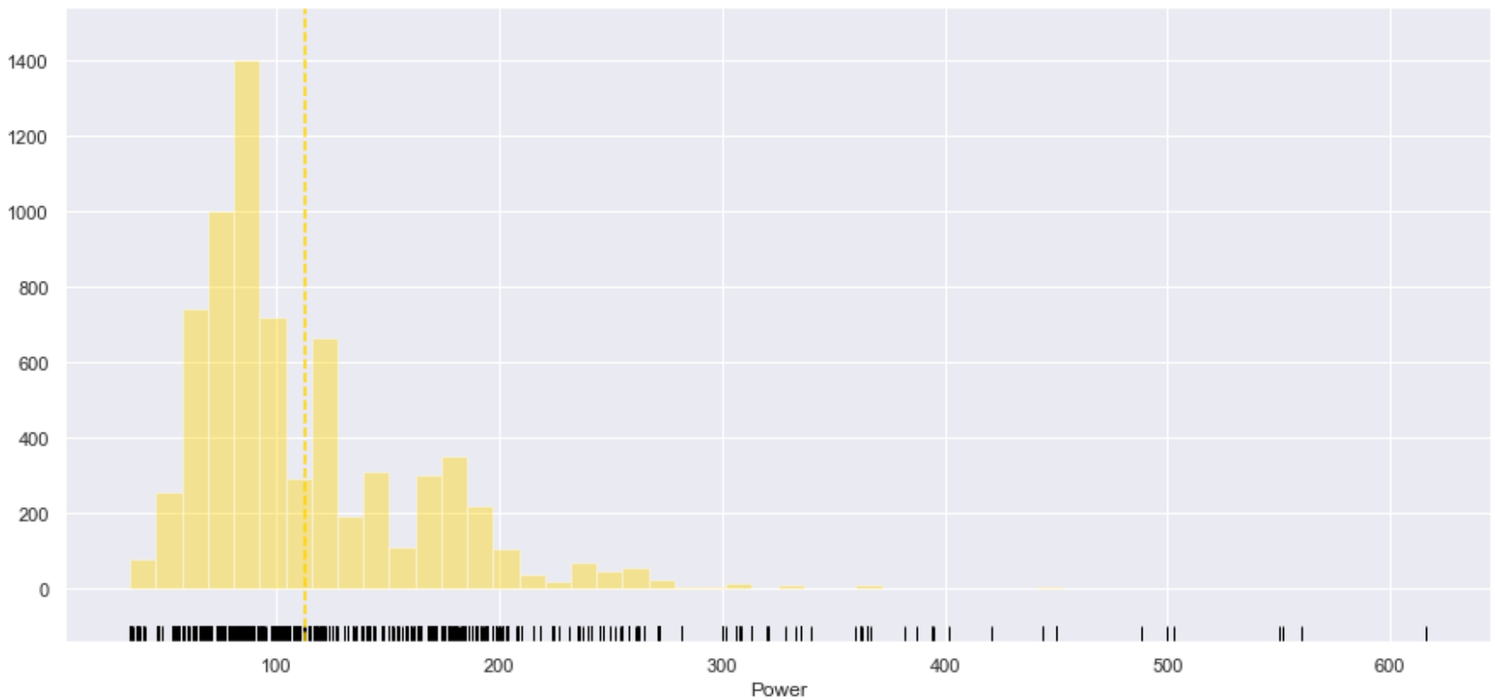
• Mean is 113 bhp

• Median is 94

• The middle 50% of engines have bhp values between 75 and 138

Insight for the business:

• The majority of your used cars will have between 65 and 190 bhp. These are the kinds of cars you can expect to see in your inventory most often.



EXPLORATORY DATA ANALYSIS - UNIVARIATE: **FUEL EFFICIENCY**

This is the standard mileage offered by the car.

Observations:

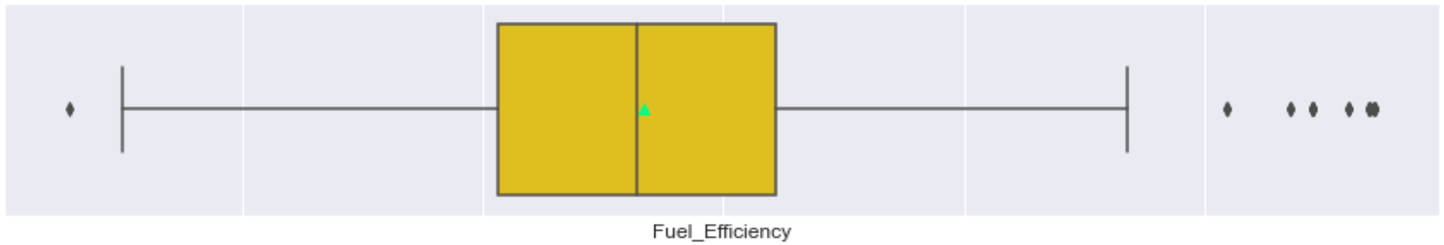
• The distribution of the data is fairly symmetrical and mound-shaped.
• There are a couple of outliers on either side.
• The mean and median are very close, meaning the outliers are not significantly skewing the data.

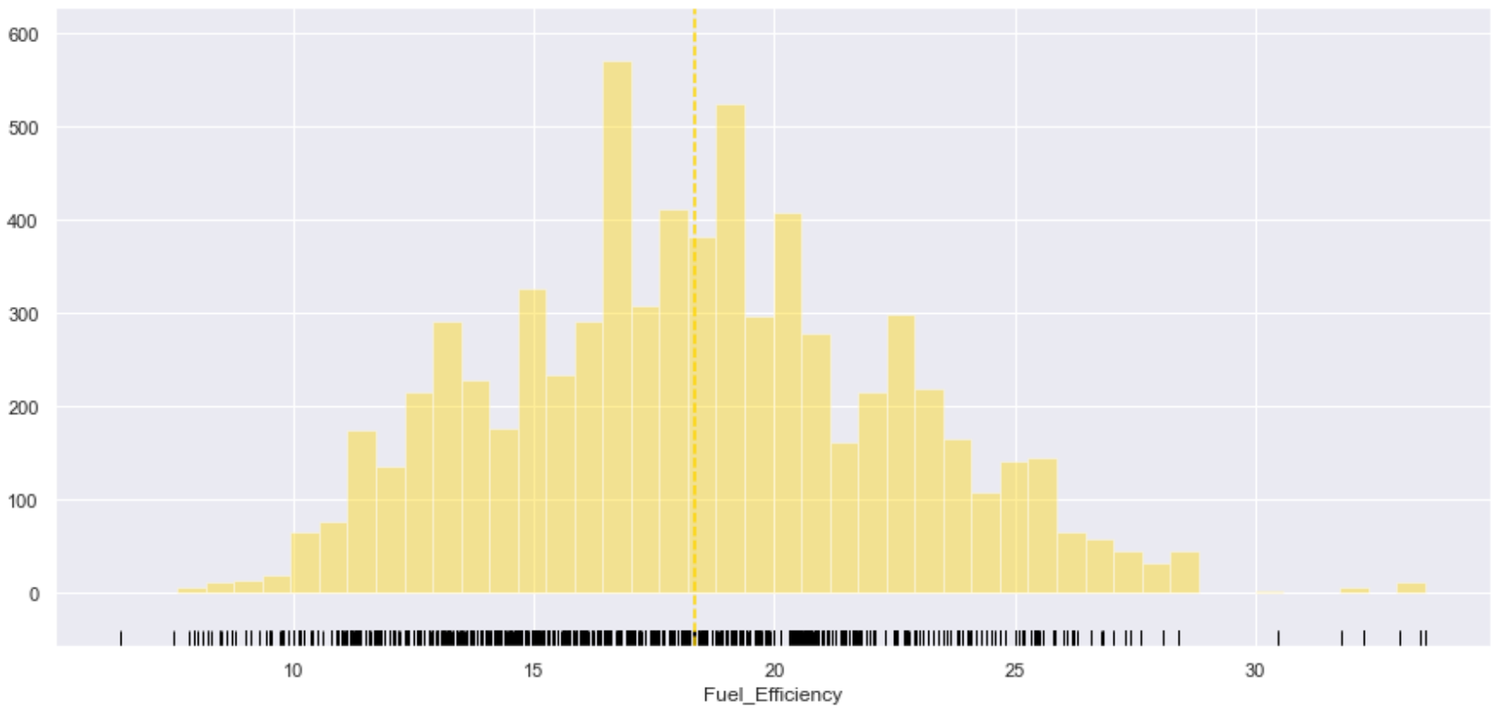
Observations on Central Tendency:

• Mean is 18 km per unit
• Median is 18 km per unit
• The middle 50% of cars have values between 15 and 21 km per unit

Insight for the business:

• The majority of your used cars will have between 13 and 23 km per unit. These are the kinds of cars you can expect to see in your inventory most often.





EXPLORATORY DATA ANALYSIS - UNIVARIATE: **LOCATION**

This is the location in which the car is being sold.

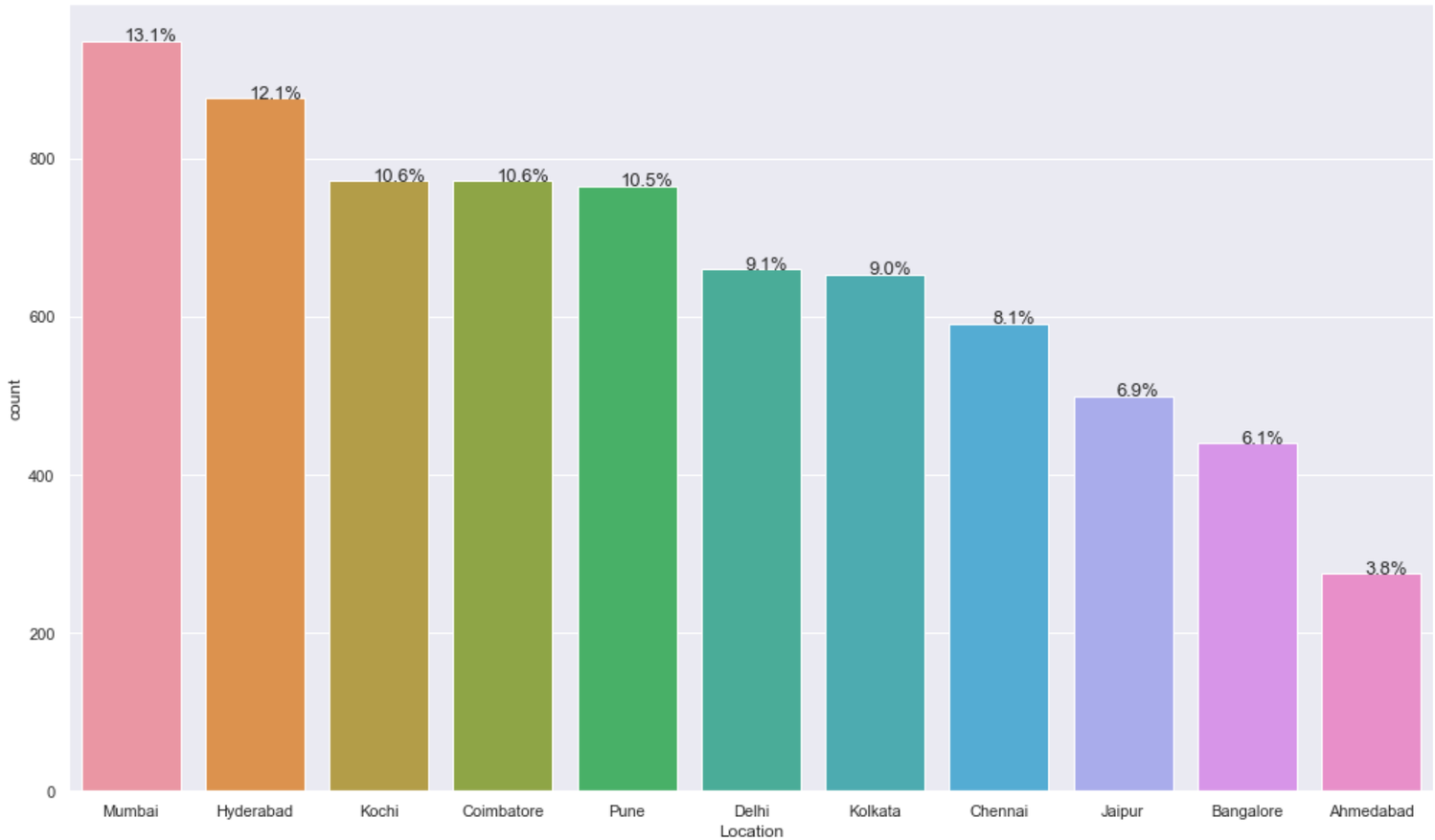
Observations:

• This data set has cars sold in eleven cities.

Insight for the business:

• Mumbai and Hyderabad will likely be the two cities with the largest inventory of cars for sale

• Of the 11 cities in the data set, Ahmedabad has the least amount of cars for sale.



EXPLORATORY DATA ANALYSIS - UNIVARIATE: **YEAR**

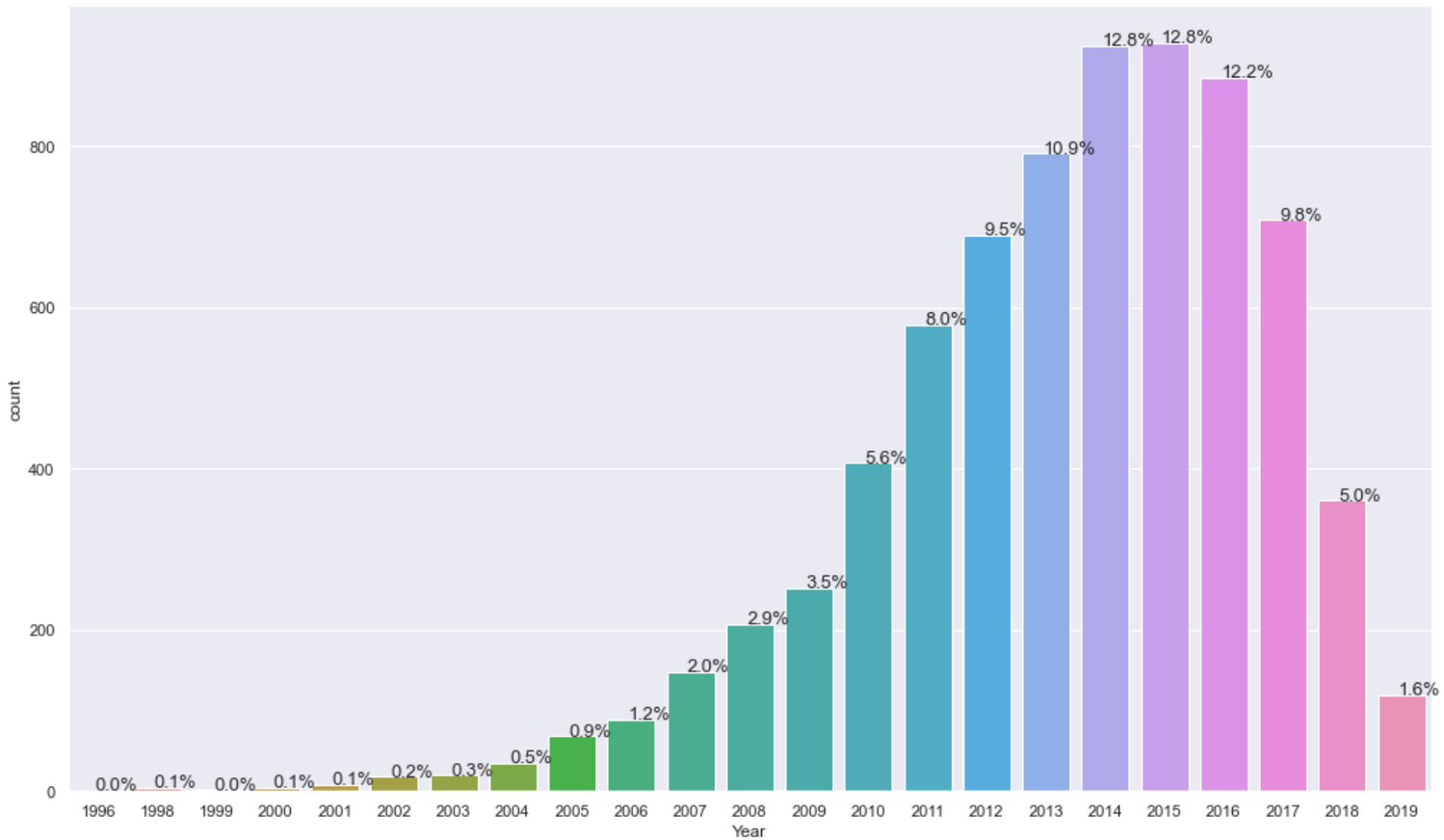
This is the location in which the car is being sold.

Observations:

• The most recent year in this data set is 2019, which indicates that the data was pulled in 2019 or 2020.
• There is a left skew, due to the smaller number of older vehicles that are for sale.
• The range of years available is about 25 years.

Insight for the business:

• The peak of your inventory will likely be cars that are about 5 years old.
• The majority of your cars will likely be between 3 and 9 years old.



EXPLORATORY DATA ANALYSIS - UNIVARIATE: **FUEL TYPE**

Observations:

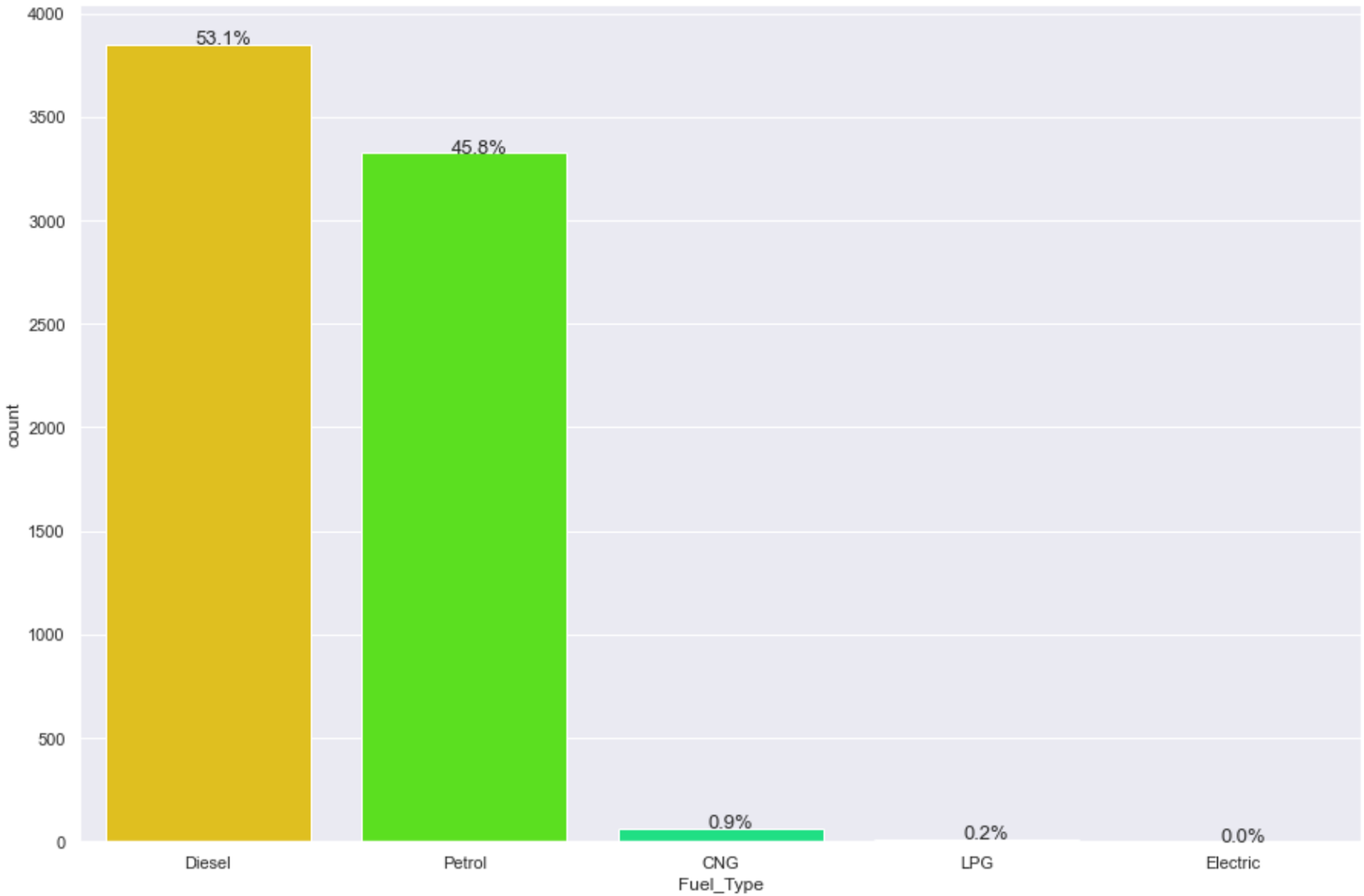
• The vast majority of cars are either Diesel or Petrol.

• CNG, LPG, and Electric vehicles all together only account for slightly more than 1% of cars.

Insight for the business:

• As a very rough guideline, Cars4U can expect approximately half of its cars to be Diesel, and the other half to be Petrol.

• Diesel vehicles make up the majority of the inventory.



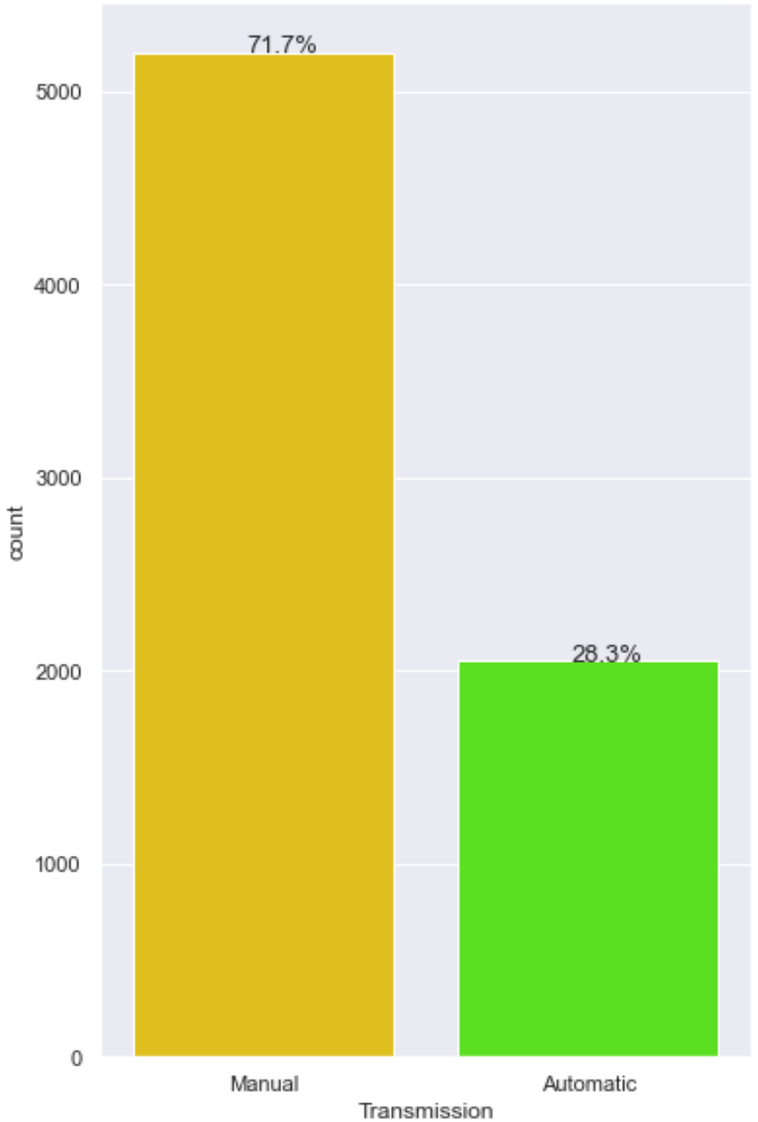
EXPLORATORY DATA ANALYSIS - UNIVARIATE: **TRANSMISSION**

Observations:

• More that two thirds of the vehicles have a manual transmission.

Insight for the business:

• Cars4U has a higher inventory of vehicles with a manual transmission, but it still has a sizeable amount of automatic cars for sale.



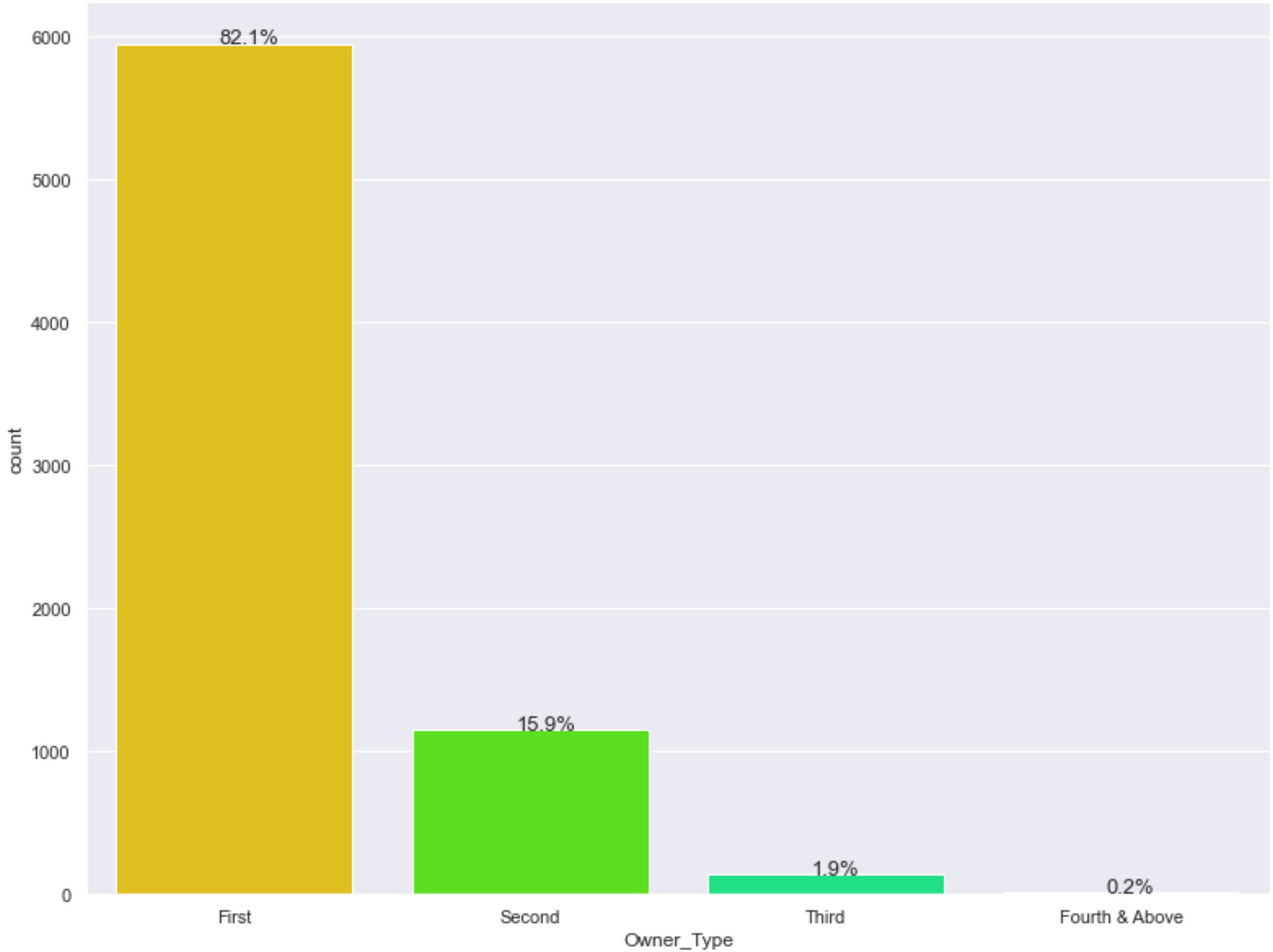
EXPLORATORY DATA ANALYSIS - UNIVARIATE: **OWNER TYPE**

Observations:

• The vast majority of cars have only had one owner.

Insight for the business:

• Cars4U can assume the majority of its business will involve buying and selling cars that have only had one previous owner.



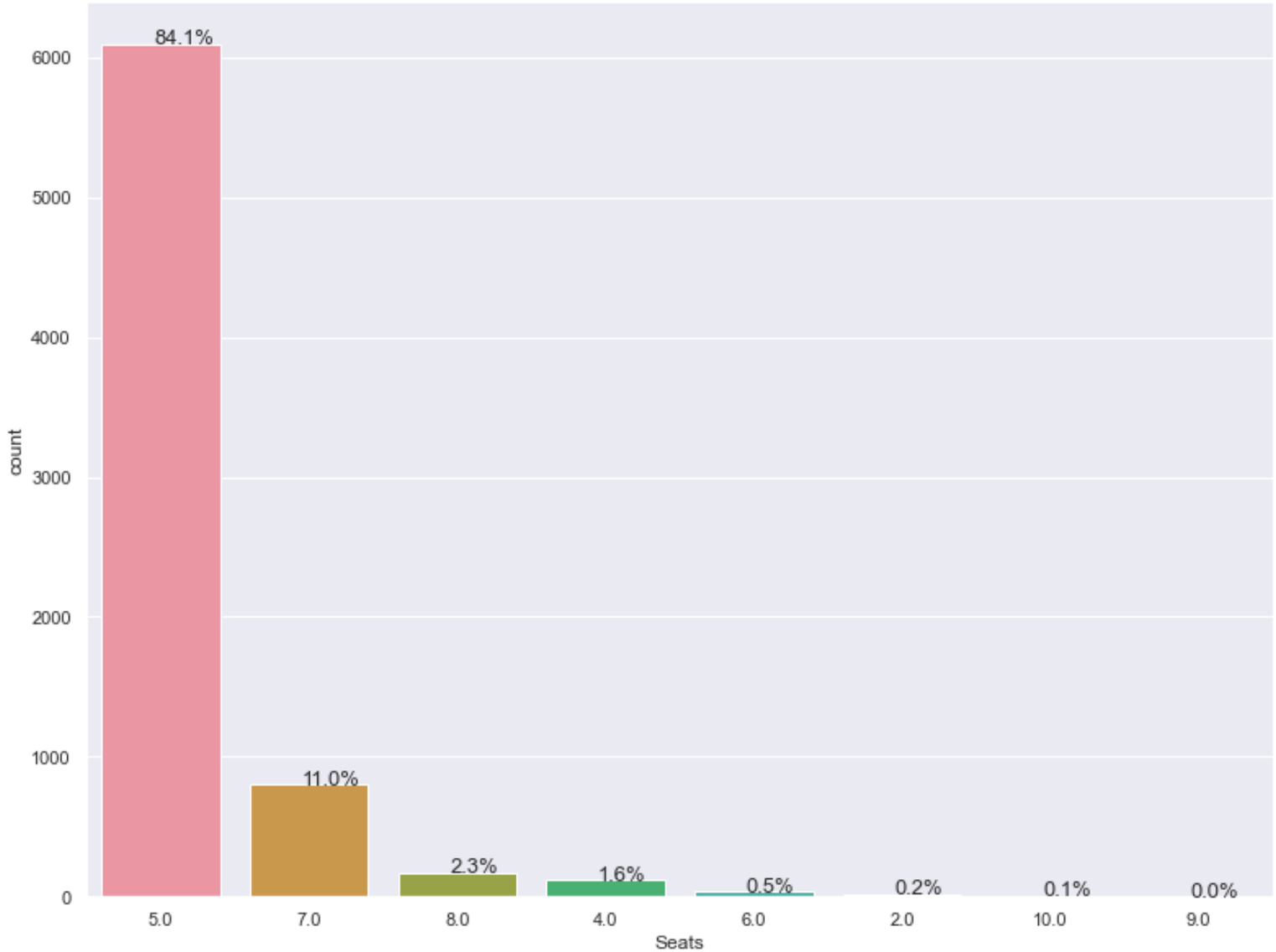
EXPLORATORY DATA ANALYSIS - UNIVARIATE: **SEATS**

Observations:

• The vast majority of cars have 5 seats.

Insight for the business:

• Cars4U can assume the majority of its business will involve buying and selling cars that have 5 seats.



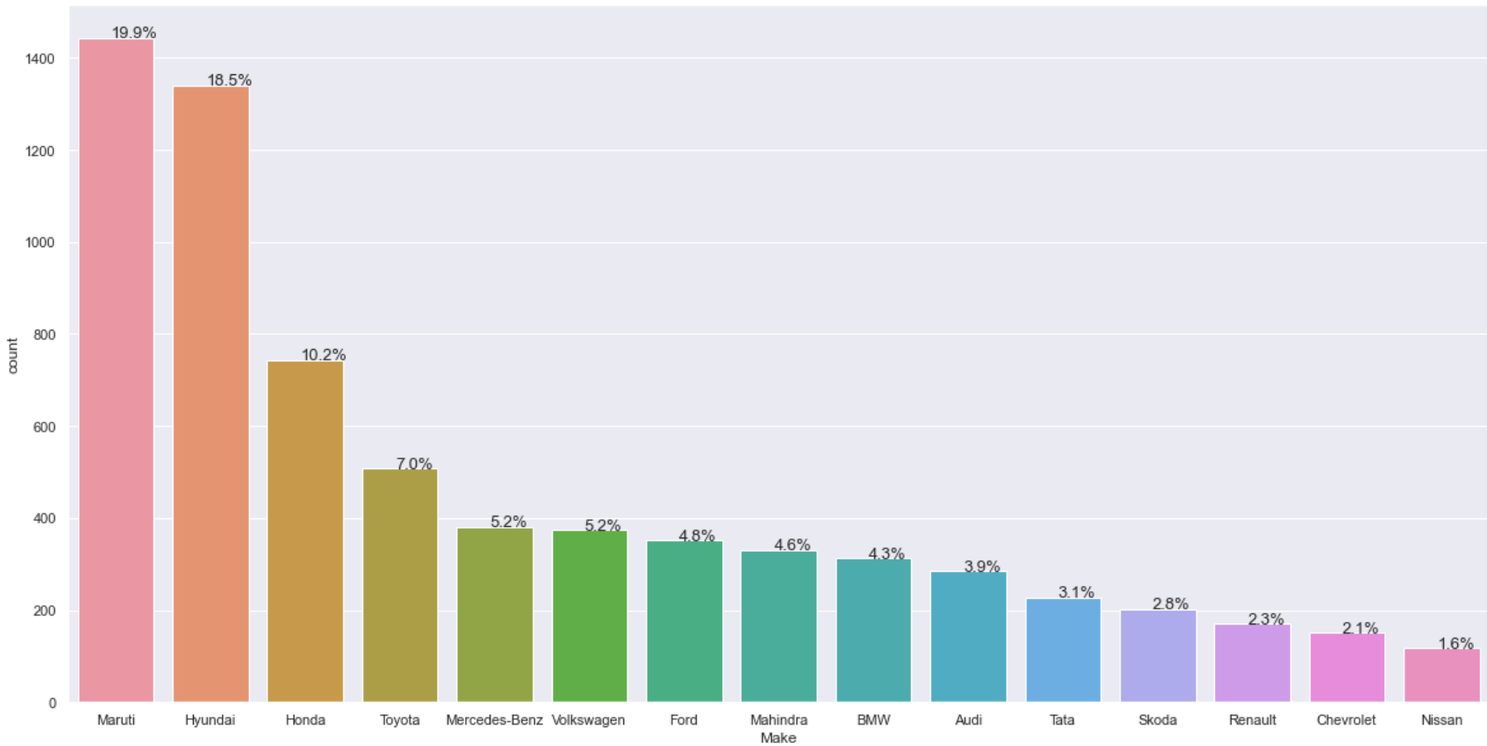
EXPLORATORY DATA ANALYSIS - UNIVARIATE: **MAKE**

Observations:

• The charts on the rights show the top 15 automakers represented in the data set. More brands exists in the data, but they represent smaller portions of the data.

Insight for the business:

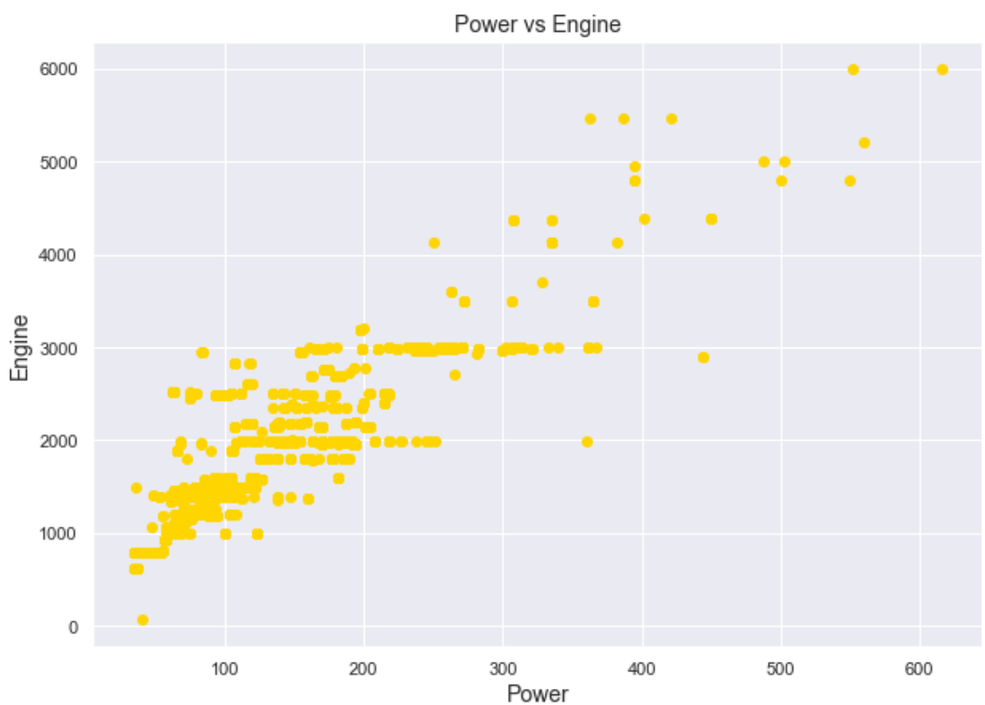
• Maruti and Hyundai are very important to the Cars4U business. These two brands alone make up almost 40% of the inventory in this data set.



EXPLORATORY DATA ANALYSIS:
BIVARIATE CORRELATION

Observations:

- We are mainly intereted in variables that correlate with price
- Price correlates highly with New_Price and Power.
- Engine and Power have a high correlation.
- Engine and Fuel_Efficiency have a noticeable negative correlation.



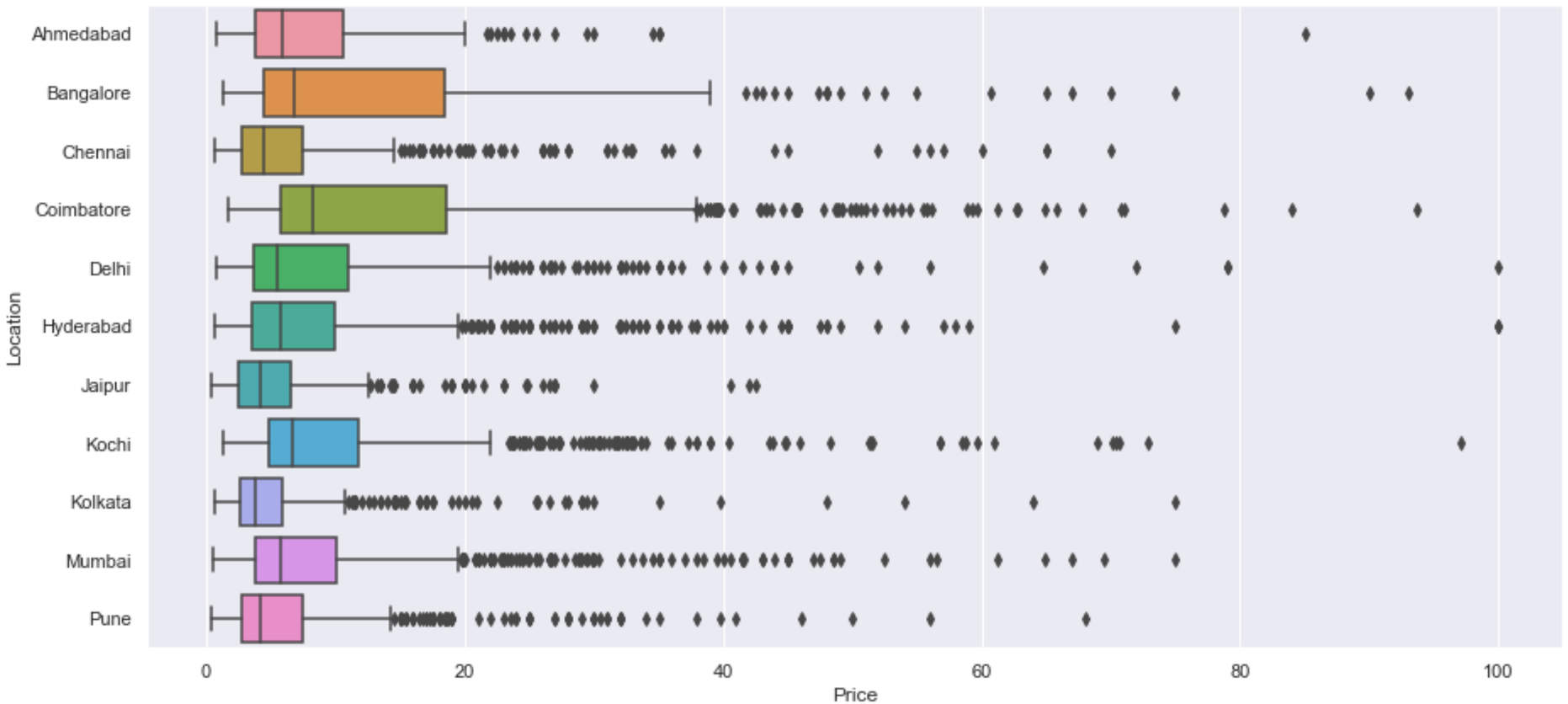
EXPLORATORY DATA ANALYSIS - BIVARIATE: **PRICE VS. LOCATION**

Key Question:

• Do cars cost different amounts in different cities?

Observations:

• It looks like there is some variability of pricing in different cities.
• Kolkata, Pune, and Jaipur appear to have lower prices.
• Bangalore and Coimbatore have higher prices.

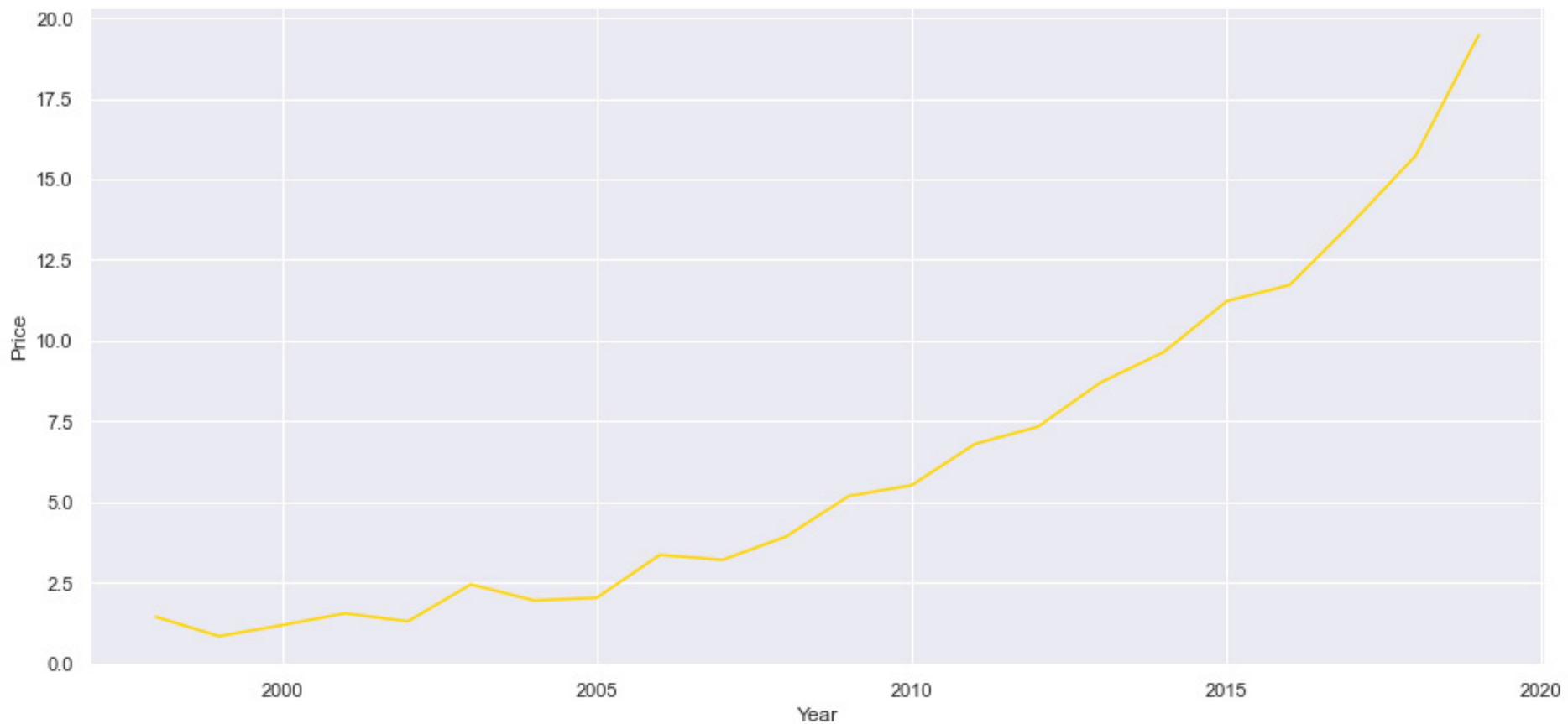


EXPLORATORY DATA ANALYSIS - BIVARIATE: **PRICE VS. YEAR**

Observations:

• As expected, the price of used cars gets lower as they get older.

• As a rough guide, the price of a car goes down by about 1.3 Lakhs every year it gets older.

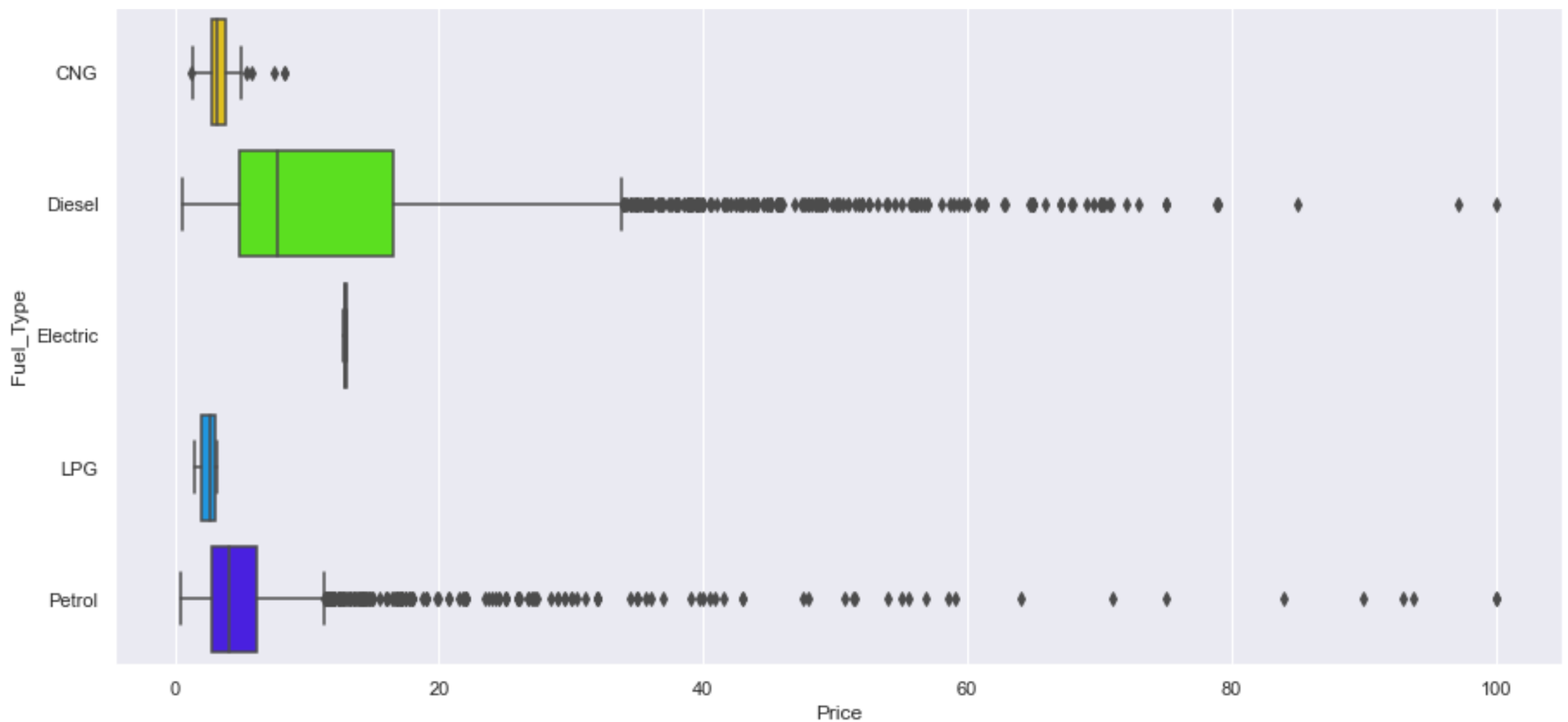


EXPLORATORY DATA ANALYSIS - BIVARIATE: **PRICE VS. FUEL TYPE**

Observations:

• As we've already noted, there are many more Diesel and Petrol cars in the data set. This leads to a larger spread of values in Diesel and Petrol vehicles.

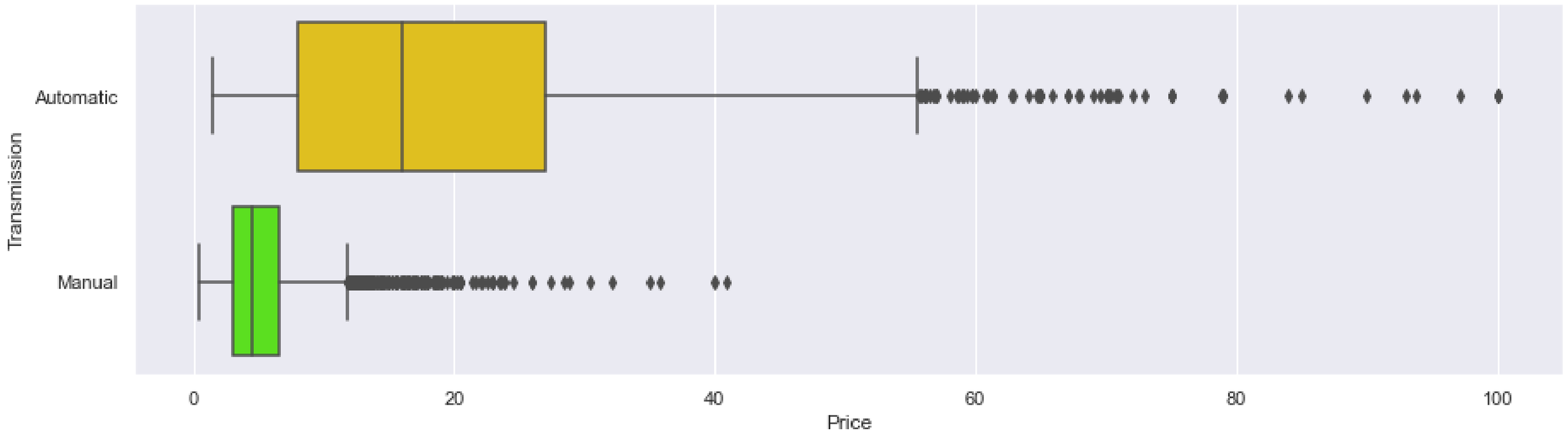
• In general, Diesel cars are more expensive than Petrol cars.



EXPLORATORY DATA ANALYSIS - BIVARIATE: **PRICE VS. TRANSMISSION**

Observations:

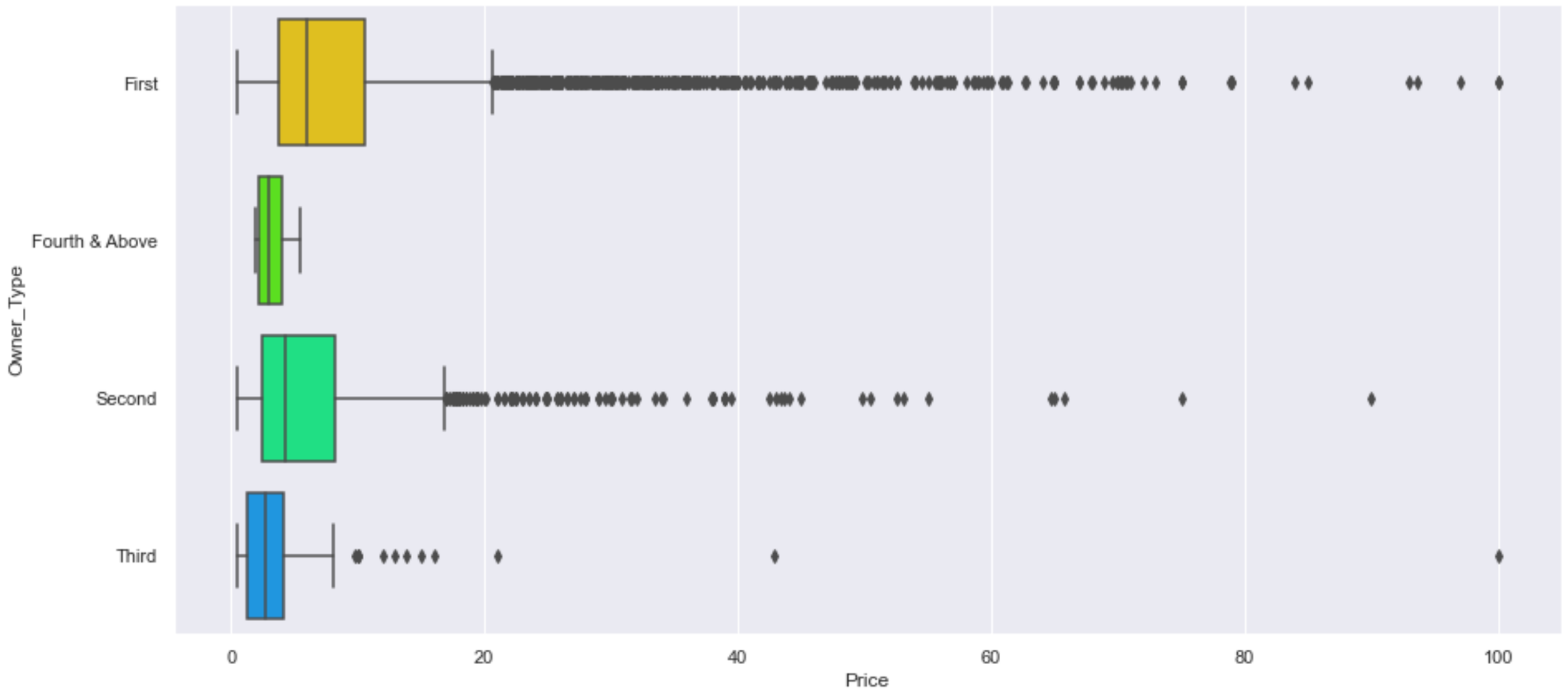
• Automatic cars have higher prices than cars with a manual transmission.



EXPLORATORY DATA ANALYSIS - BIVARIATE: **PRICE VS. OWNER TYPE**

Observations:

• Cars with more owners generally have lower prices.

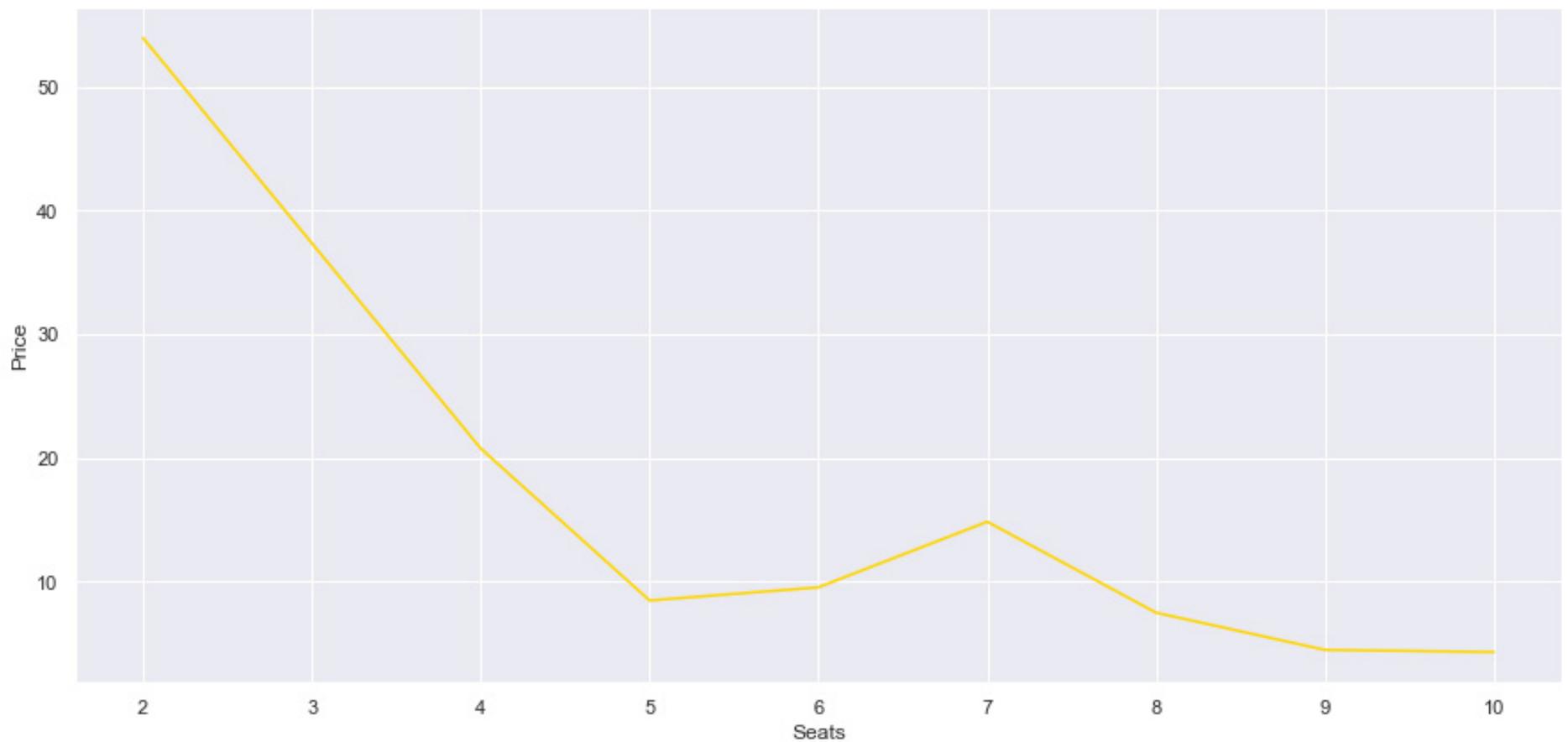


EXPLORATORY DATA ANALYSIS - BIVARIATE: **PRICE VS. SEATS**

Observations:

• Cars with only two seats are the most expensive, likely because they are sports cars.

• Vehicles with 9 or 10 seats have the lowest prices.



MODEL SUMMARY

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.873
Model:	OLS	Adj. R-squared:	0.873
Method:	Least Squares	F-statistic:	1161.
Date:	Fri, 07 May 2021	Prob (F-statistic):	0.00
Time:	23:11:00	Log-Likelihood:	-10772.
No. Observations:	5077	AIC:	2.161e+04
Df Residuals:	5046	BIC:	2.181e+04
Df Model:	30		
Covariance Type:	nonrobust		
Omnibus:	199.818	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	591.510
Skew:	-0.096	Prob(JB):	3.59e-129
Kurtosis:	4.661	Cond. No.	5.45e+07

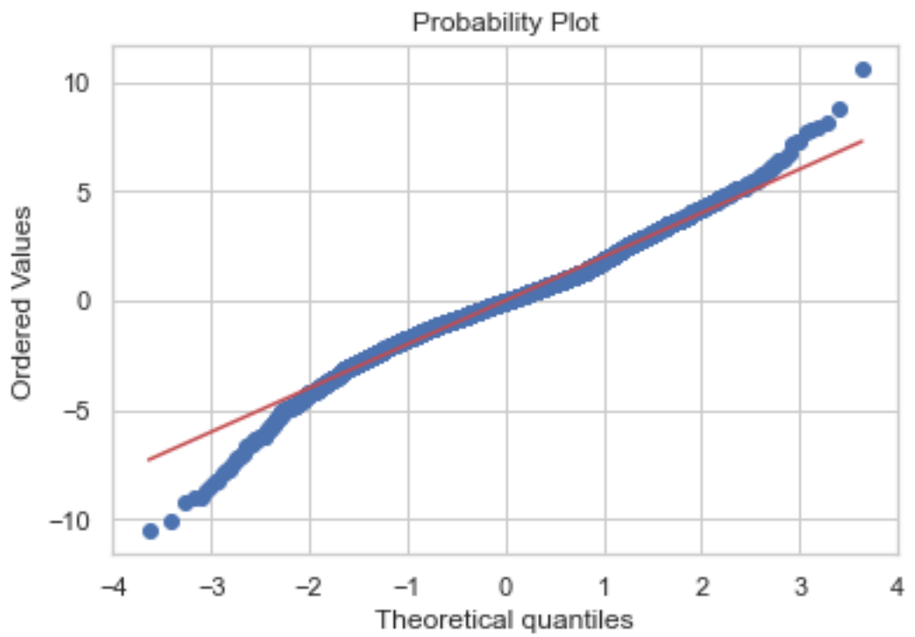
MODEL SUMMARY

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1005.4748	24.308	-41.363	0.000	-1053.130	-957.820
Year	0.5022	0.012	41.410	0.000	0.478	0.526
Kilometers_Driven	-1.448e-05	1.29e-06	-11.224	0.000	-1.7e-05	-1.2e-05
Power	0.0566	0.001	47.386	0.000	0.054	0.059
Seats	0.1460	0.050	2.905	0.004	0.047	0.245
Fuel_Efficiency	-0.1265	0.013	-10.119	0.000	-0.151	-0.102
Bangalore	0.7637	0.126	6.083	0.000	0.518	1.010
Coimbatore	0.4710	0.100	4.720	0.000	0.275	0.667
Delhi	-0.4650	0.105	-4.449	0.000	-0.670	-0.260
Hyderabad	0.7051	0.096	7.371	0.000	0.518	0.893
Kolkata	-1.1634	0.110	-10.594	0.000	-1.379	-0.948
Mumbai	-0.3545	0.094	-3.773	0.000	-0.539	-0.170
Diesel	1.9229	0.087	22.004	0.000	1.752	2.094
Electric	9.5106	1.448	6.568	0.000	6.672	12.350
Manual	-1.5821	0.093	-17.080	0.000	-1.764	-1.401
Second	-0.3068	0.083	-3.699	0.000	-0.469	-0.144
Ford	-2.4587	0.158	-15.573	0.000	-2.768	-2.149
Honda	-2.8902	0.124	-23.266	0.000	-3.134	-2.647
Hyundai	-2.2379	0.113	-19.796	0.000	-2.459	-2.016
Jaguar	1.1735	0.370	3.167	0.002	0.447	1.900
Jeep	2.3705	0.535	4.430	0.000	1.321	3.420
Land_Rover	2.3977	0.321	7.465	0.000	1.768	3.027
Mahindra	-3.2200	0.160	-20.069	0.000	-3.535	-2.905
Maruti	-1.7471	0.121	-14.391	0.000	-1.985	-1.509
Mercedes-Benz	1.0782	0.144	7.509	0.000	0.797	1.360
Mini	5.7339	0.474	12.096	0.000	4.805	6.663
Nissan	-2.9408	0.243	-12.122	0.000	-3.416	-2.465
Renault	-2.8878	0.212	-13.639	0.000	-3.303	-2.473
Skoda	-2.5949	0.182	-14.242	0.000	-2.952	-2.238
Tata	-3.7763	0.191	-19.790	0.000	-4.150	-3.402
Volkswagen	-3.0109	0.152	-19.784	0.000	-3.309	-2.713

MODEL SUMMARY

Observations:

- Adjusted R-squared is 0.873, Our model is able to explain 87.3% of variance that shows model is good.
- We have low test and train errors (2.019 and 2.119).
- Both the errors are comparable, so our model does not suffer from overfitting.



BUSINESS INSIGHTS AND RECOMMENDATIONS

1. It is noticable that Cars4U does not have a significant presense in some of India larger cities. It is worth exploring other markets.

2. The pricing model could be much more usable to Cars4U if data was included on the profit margin for each vehicle. This would allow our model to calculate which models Cars4U should aggressively attempt to purchase for resale.

3. Further car data might help improve our pricing model. I think it may have been valuable to know some of the following:

- color of car
- condition of car (body damage, etc.)
- history of accidents

4. Cars4U could benefit from an inventory:

- a model that predicts car inventory by time, location, or other factors; turnover of different makes and models may be faster or slower according to market, month, or other market demands

THANK YOU

EXPLORATORY DATA ANALYSIS AND PRICING MODEL

by **JAKE EIDE**