CARS4U PROJECT SUPERVISED LEARNING

Project 3

Authored by C. Kamakani Kahunahana

BUSINESS OVERVIEW
PROBLEM AND SOLUTION APPROACH

PROBLEM

Demand for used cars continues to increase. Management needs an effective way to accurately price their used car inventory to maximize returns and minimize risk.

There is a huge demand for used cars in the Indian Market today. As sales of new cars have slowed down in the recent past, the pre-owned car market has continued to grow over the past years and is larger than the new car market now.

Unlike new cars, where price and supply are fairly deterministic and managed by OEMs, used cars are very different beasts with huge uncertainty in both pricing and supply. Keeping this in mind, the pricing scheme of these used cars becomes important in order to grow in the market.

SOLUTION APPROACH

Build a predictive pricing model for used cars using linear regression. Leverage this model to help managers maximize profit and minimize risk associated with used car sales.

Using data provided, we will identify key variables from market characteristics, vehicle features, and ownership history which have a statistically significant influence on the value of used cars.

Based on this analysis I will build and test a linear regression model which can assist management in more accurately determining the value and sales price of their used car inventory.

BASE DATASET OVERVIEW

Details	Amount
Observations	7253
Variables	14
Null	New_Price: 6247, Price: 1234, Seats: 53, Power: 46, Engine: 46, Mileage: 2

This is a description of the base dataset however, much transformation will need to be done prior to performing any analysis or model construction

Variable	Type	Description
S.No.	Integer	Serial number
Name	Category	Name of model and brand
Location	Category	City available for purchase
Year	Integer	Manufacture year
Kilometers_Driven	Integer	Total km driven by previous owner
Fuel_Type	Category	Petrol, Diesel, Electric, CNG, LPG
Transmission	Category	Manual or Automatic
Owner	Category	Type of owner (first, second, third, 4th or more)
Mileage	Category	Mileage rating in km
Engine	Category	Displacement in CC
Power	Category	Max bhp
Seats	Float	Number of seats in car
New_Price	Category	Price of a new car of same model
Price	Float	Price of used car

01

DATA PREPROCESSING

DATA PRE-PROCESSING

PROCESSING OBJECTS AND UNNECESSARY DATA

Variable	Pre-Processing Strategy
S.No.	Dropped because redundant with DF Index
Name	Split name into Brand_Name and Model_Name
Location	Convert to category
Year	none
Kilometers_Driven	none
Mileage	Removed unnecessary descriptive values; convert to float
Fuel_Type	Convert to category
Transmission	Convert to category
Owner_Type	Transformed to numerical value to reduce columns
Power	Removed “bhp” and converted to float
Brand_Name	Car make category
Model_Name	Car model name category

- Dropped S.No.
- Split and remove unnecessary data
- Convert objects to categories
- Split Name into two variables Brand_Name and Model_Name

BASE DATASET OVERVIEW
PROCESSING NULL VALUES

PROCESSING NULL VALUES

Variable	Total Null	Pre-Processing Strategy
New_Price	6247	Drop row < 50% values
Price	1234	Fill with median values grouped by Brand_Name
Seats	53	Fill with median values grouped by Brand_Name
Power	46	Fill with median values grouped by Brand_Name
Engine	46	Fill with median values grouped by Brand_Name
Mileage	2	ill with median values grouped by Brand_Name

- Dropped New_Price because missing > 50% of values
- Use the median value base on Brand_Name to fill in null values by column

NEW DATASET OVERVIEW
NUMERICAL VALUES

NUMERICAL VARIABLE DESCRIPTION

	Year	Kilometers_Driven	Seats	Price (INR Lakh)	Owner_Type	Engine	Power	Mileage
Count	7,253	7,253	7,253	7,253	7,253	7,253	7,253	7,253
Mean	2,013.37	58,699.06	5.28	9.25	1.20	1616.12	112.46	18.14
Standard Deviation	3.25	84,427.00	0.81	10.63	0.46	594.43	53.21	4.56
Minimum	1996.00	171.00	0.00	0.44	1.00	72.00	34.20	0.00
25%	2011.00	34,000.00	5.00	3.75	1.00	1198.00	75.00	15.17
50%	2014.00	53,416.00	5.00	5.35	1.00	1494.00	93.70	18.16
75%	2016.00	73,000.00	5.00	9.89	1.00	1968.00	138.03	21.10
Maximum	2019.00	6,500,000,00	10.00	160.00	4.00	5998.00	616.00	33.54

- Mean year is 2013, min 1996, max 2019
- Kilometers_Driven has wide variability and appears to have a few large outliers (6,500,000)
- Seats has a mean of 5.28 but a min of 0.00 is odd since a car must have seats
- Price has many outliers at both ends of the range
- Owner_Type indicates most used cars have just one owner before being resold
- Engine has outliers which must be examined
- Power has outliers but IQR range seems reasonable
- Mileage seems normally distributed

NEW DATASET OVERVIEW
CATEGORICAL VALUES

CATEGORICAL VARIABLE DESCRIPTION

	Location	Fuel_Type	Transmission	Brand_Name	Model_Name
Count	7,253	7,253	7,253	7,253	7,253
Unique	11	5	2	32	2041
Top	Mumbai	Diesel	Manual	Maruti	XUV500 W8 2WD
Frequency	949	3852	5204	1444	155

- Location of Mumbai is the most common. There are 11 different locations
- There are 5 different Fuel_Types with Diesel being most common
- There are only two kind of Transmissions,. Manual is most common
- Brand_Name has 32 different types. Maruti is the most common
- Model Name has 2041 unique values making it problematic for linear regression and too large to create dummies

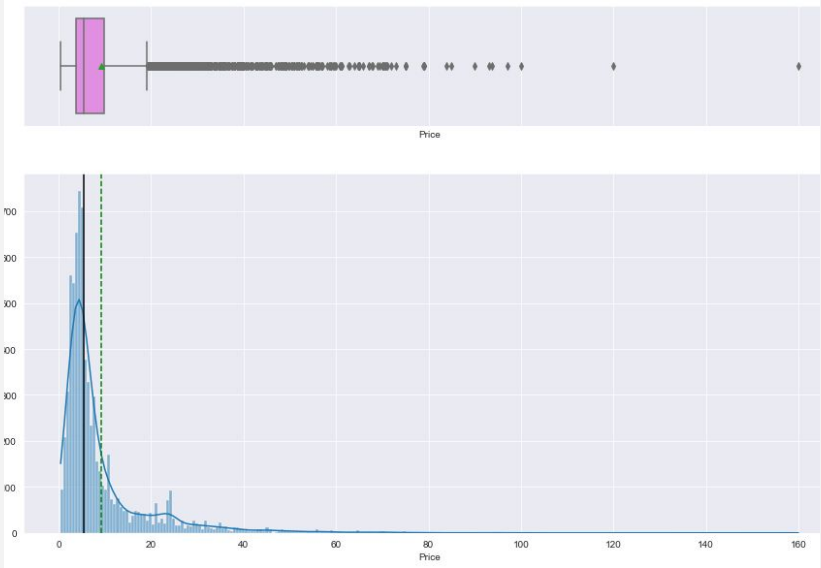
02

EXPLORATORY DATA ANALYSIS

VEHICLE INVENTORY ANALYSIS

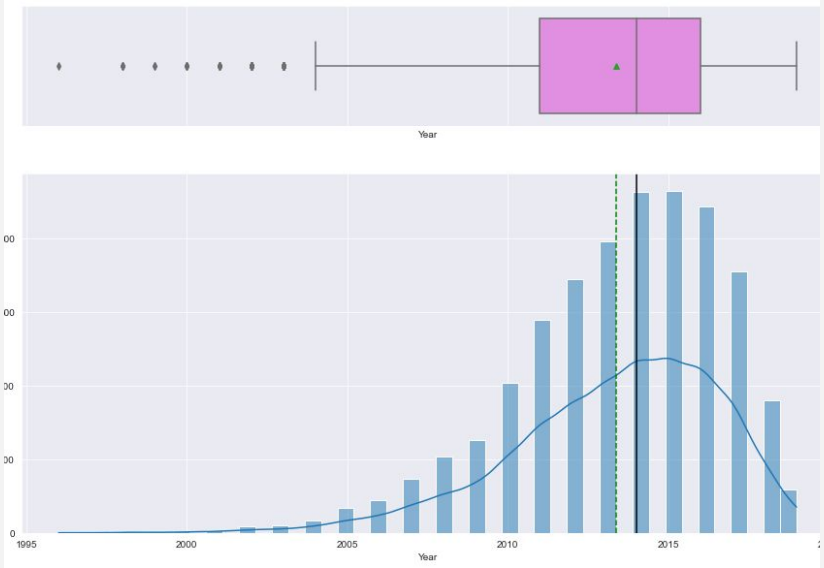
VEHICLE PRICE AND YEAR OF MANUFACTURE

VEHICLE PRICE



- High variability in price with long tail and many outliers to higher end
- Extreme values at 120 and 160 may need to be trimmed or capped for linear regression
- Mean price is just 9.25

YEAR MANUFACTURED

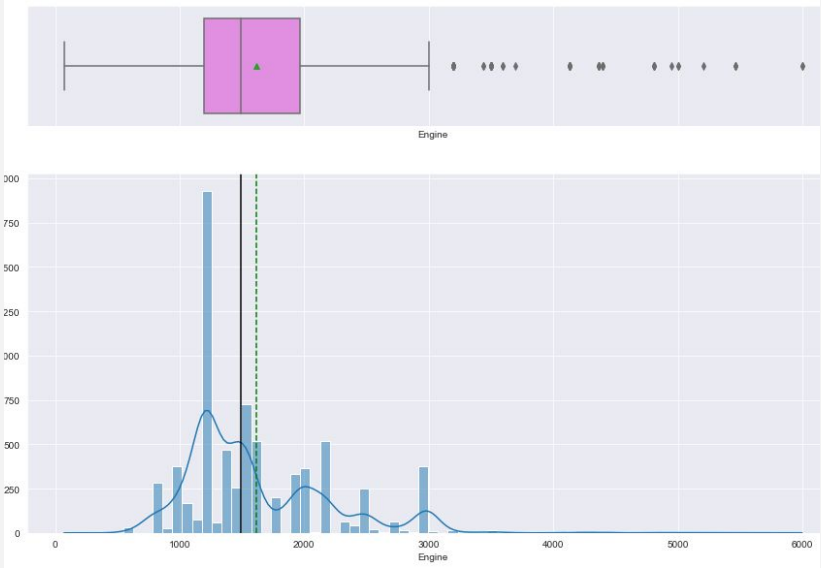


- Long left skew with outliers in the lower range
- Mean is 2013
- Most vehicles were made between 2011-2016

VEHICLE INVENTORY ANALYSIS

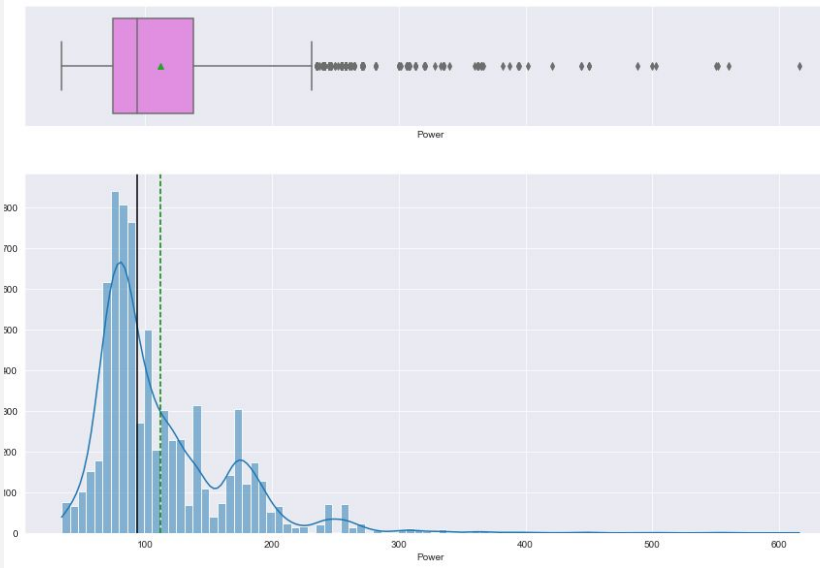
ENGINE SIZE AND POWER BHP

ENGINE SIZE



- Right skew with many outliers in the upper range
- May need to be trimmed or capped for linear regression
- Average engine size 1616.12

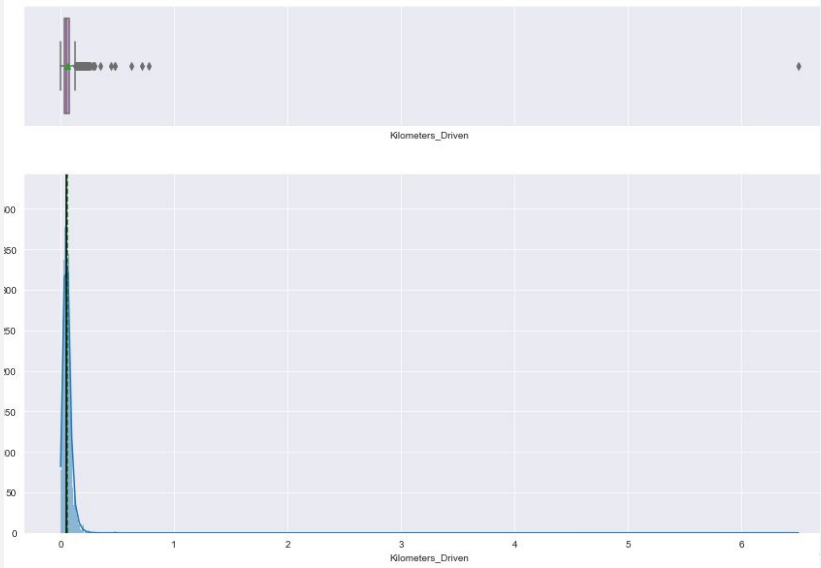
POWER BHP



- Many outliers in the upper end may need to be trimmed or capped
- Average is 112.46
- There appears to be a few high-powered vehicles (sports cars) which skew the data

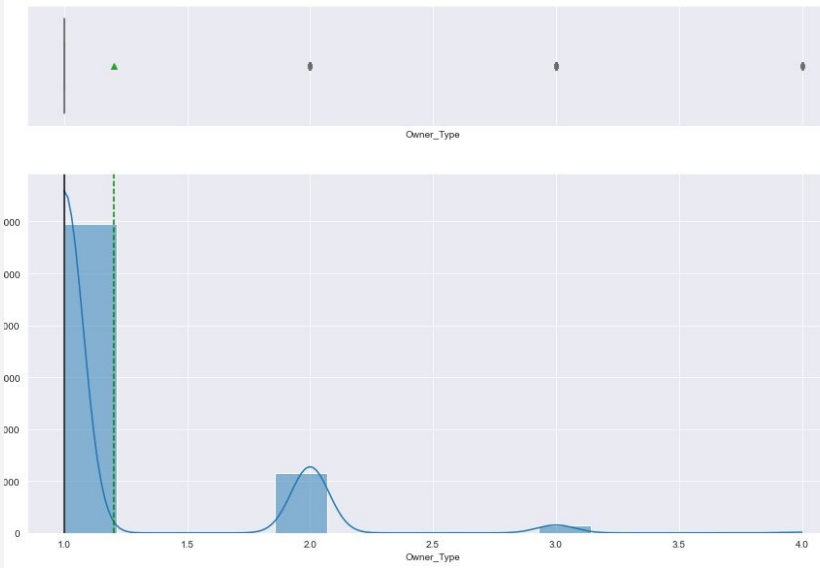
VEHICLE INVENTORY ANALYSIS
KILOMETERS_DRIVEN AND OWNER_TYPE

KILOMETERS_DRIVEN



- There is an extreme value at 6,500,000 which is having an outsized impact on the distribution
- Will need to be trimmed or capped
- Average distance driven is 58,699

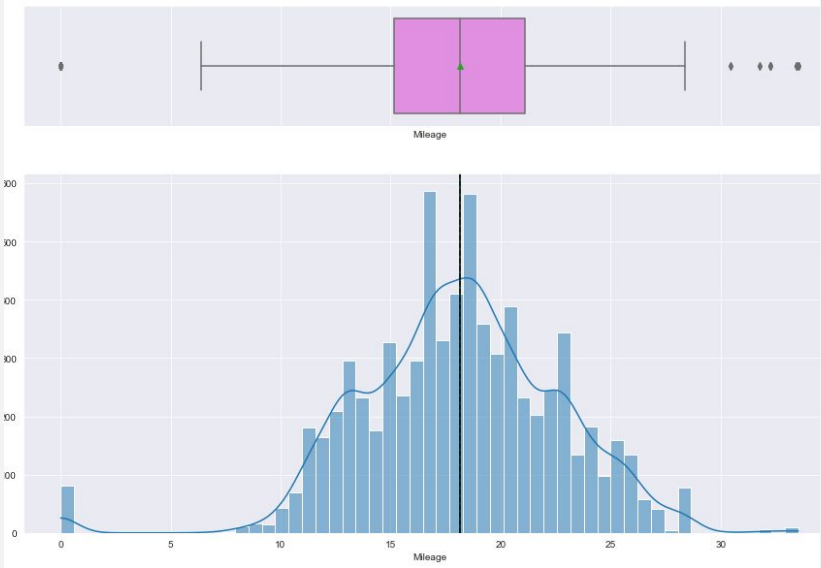
OWNER_TYPE



- Most vehicles are being sold after just one owner

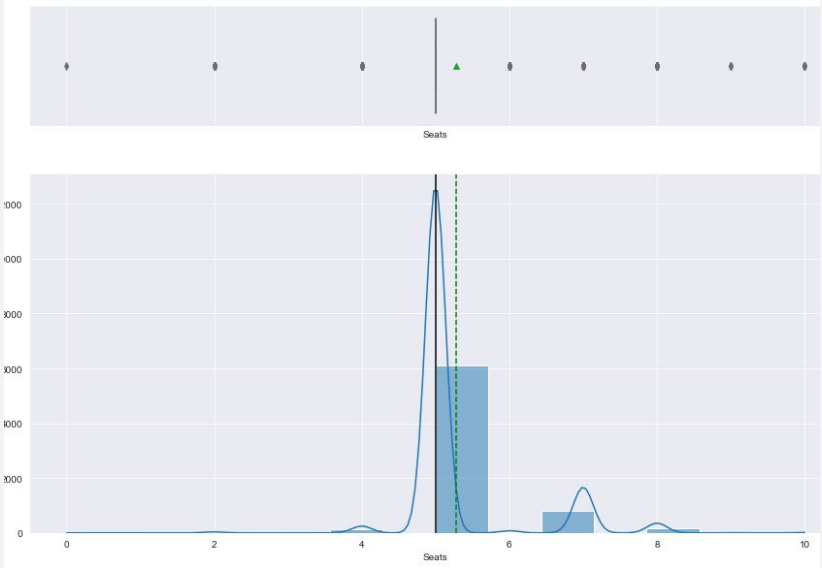
VEHICLE INVENTORY ANALYSIS
MILEAGE AND NO. OF SEATS

MILEAGE



- The amount of mileage reported for each vehicle seems to have a normal distribution with a few outliers
- Average reported mileage is 18.14

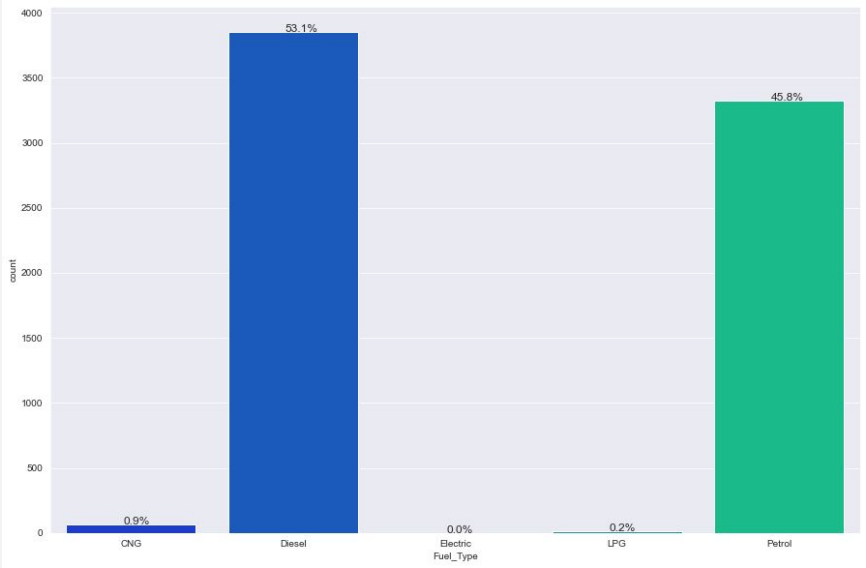
NO. OF SEATS



- The average car has 5 seats
- There are some unusual values of 0 seats and 10 seats which need to be examined

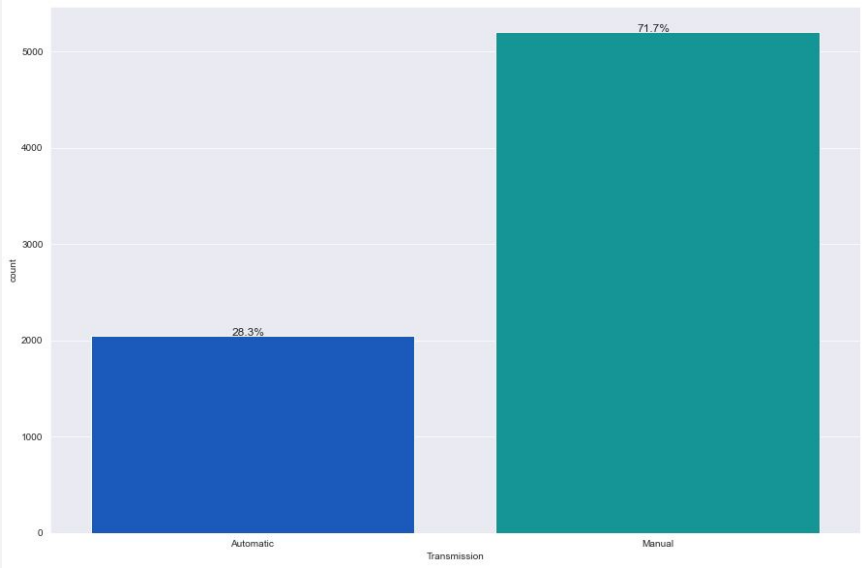
VEHICLE INVENTORY ANALYSIS
FUEL TYPE AND TRANSMISSION

FUEL TYPE



- Diesel comprises 53.1% and Petrol 45.8%
- CNG 0.9%, LPG 0.2%, and Electric < 0.0% (2 count) are minimal

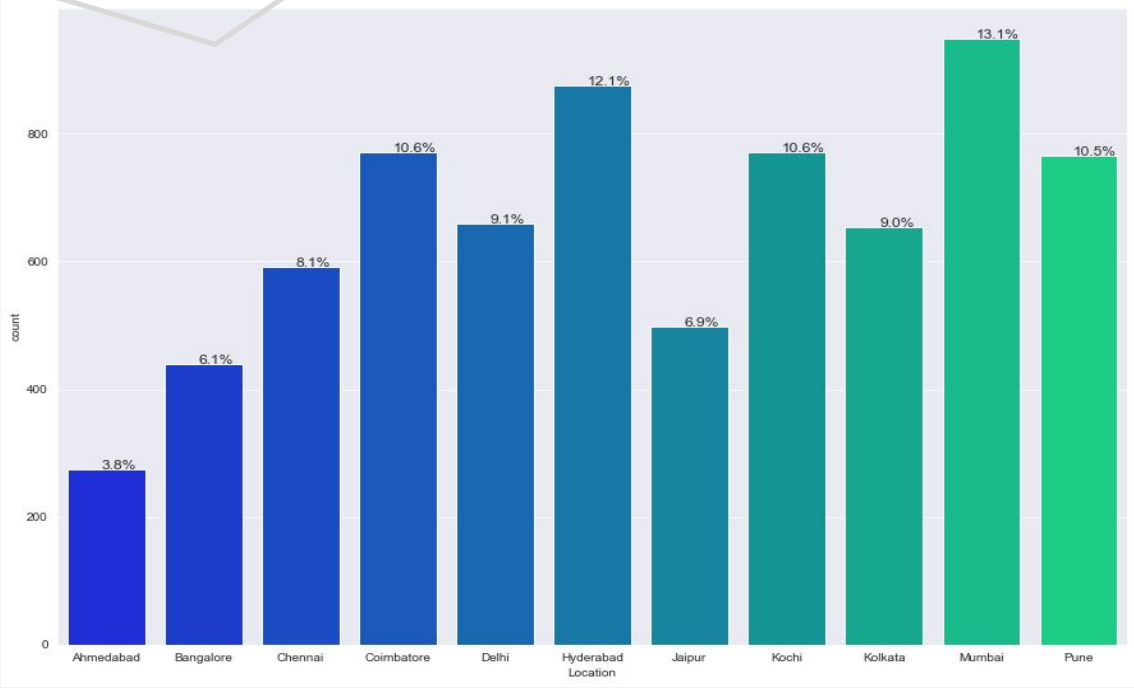
TRANSMISSION



- Most vehicles have manual transmissions accounting for 71.7%

VEHICLE INVENTORY ANALYSIS
VEHICLE LOCATION

LOCATION

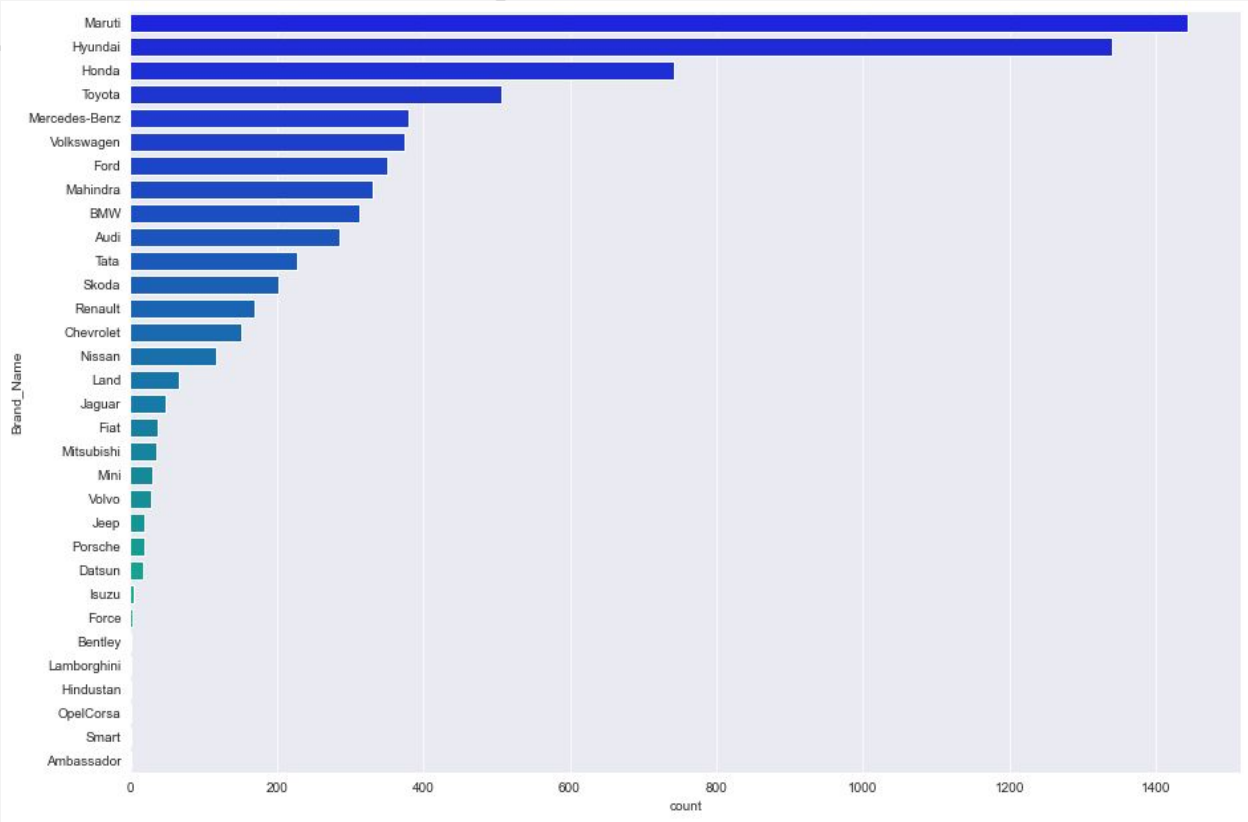


- Mumbai (13.1%) is the most common location followed closely by Hyderabad (12.1%)
- Ahmedabad (3.8%) has the least number of occurrences

VEHICLE INVENTORY ANALYSIS

BRAND NAMES

BRAND NAMES



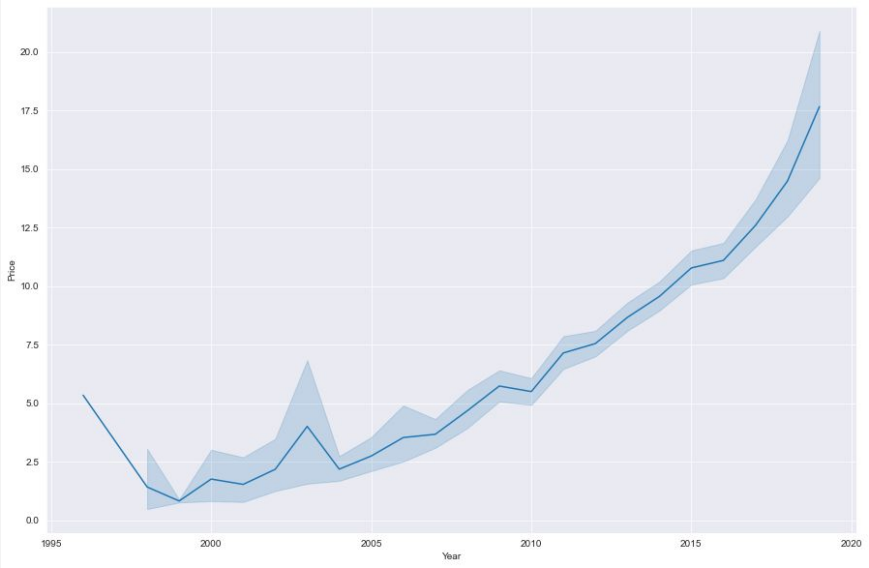
- There are 32 unique brand names
- Maruti (20%) and Hyundai (18%) are the most common vehicle brands in this study
- The top 10 brands account for 83% of all the vehicles in this study

03

BIVARIATE & MULTIVARIATE ANALYSIS

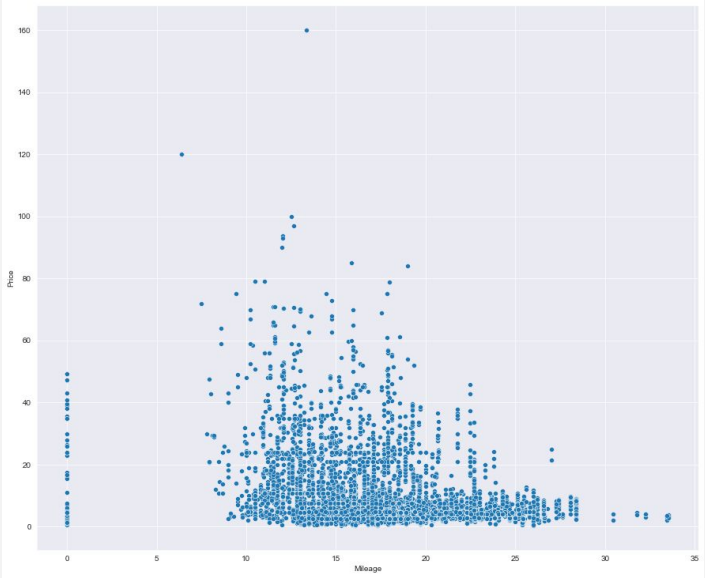
VEHICLE PRICE ANALYSIS
YEAR AND MILEAGE

YEAR VS PRICE



- Newer models command higher prices

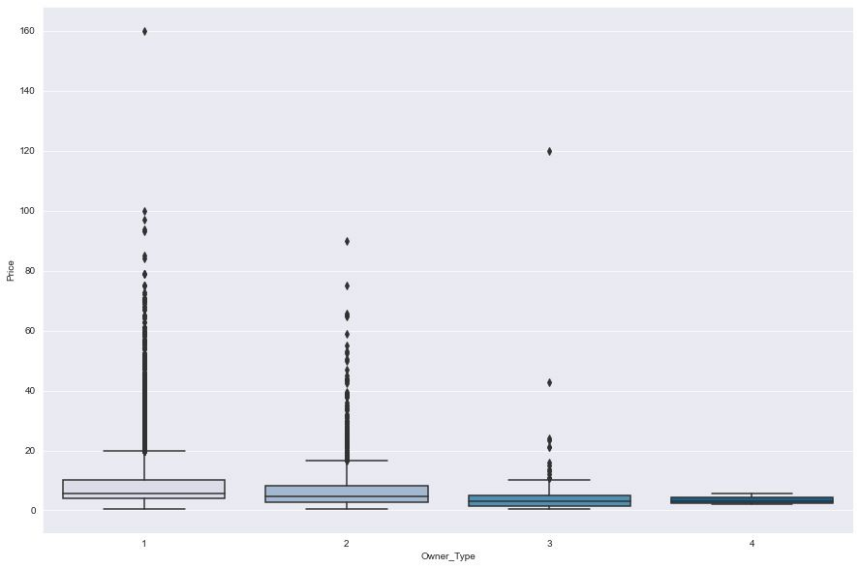
MILEAGE VS PRICE



- The amount of mileage a vehicle gets does not seem to be correlated with price.
- There are some model reporting zero mileage which needs to be treated for errors
- There are two outliers with high prices 160 and 120 both with low mileage. These may be high-end sports cars.

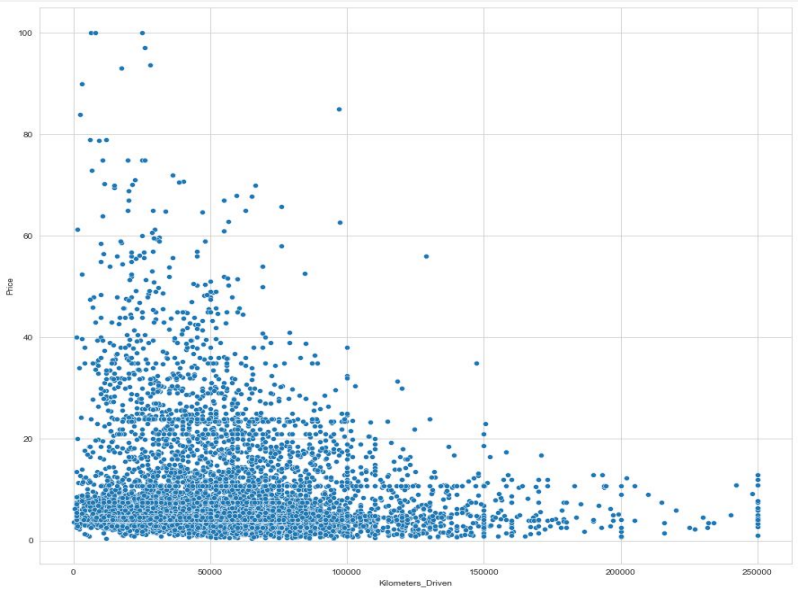
VEHICLE PRICE ANALYSIS
OWNER TYPE AND KILOMETERS DRIVEN

OWNER TYPE VS PRICE



- Cars with only one owner have higher prices
- There is an outlier in the one owner plot

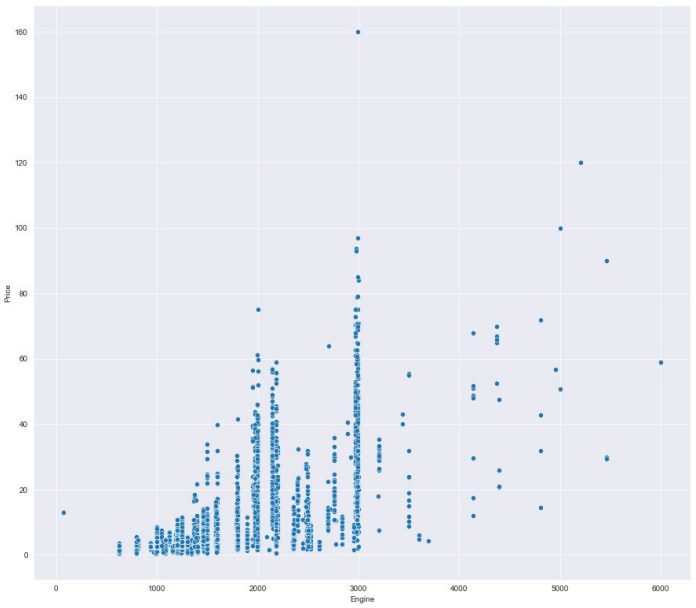
KILOMETERS DRIVEN VS PRICE



- There seems to be a negative relationship between kilometers driven and price

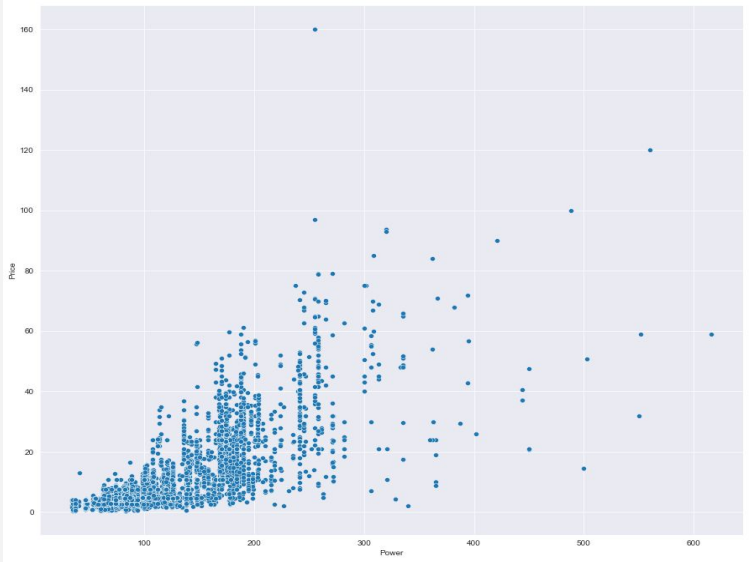
VEHICLE PRICE ANALYSIS
ENGINE SIZE AND POWER BHP

ENGINE SIZE VS PRICE



- There is a positive relationship between engine size and price

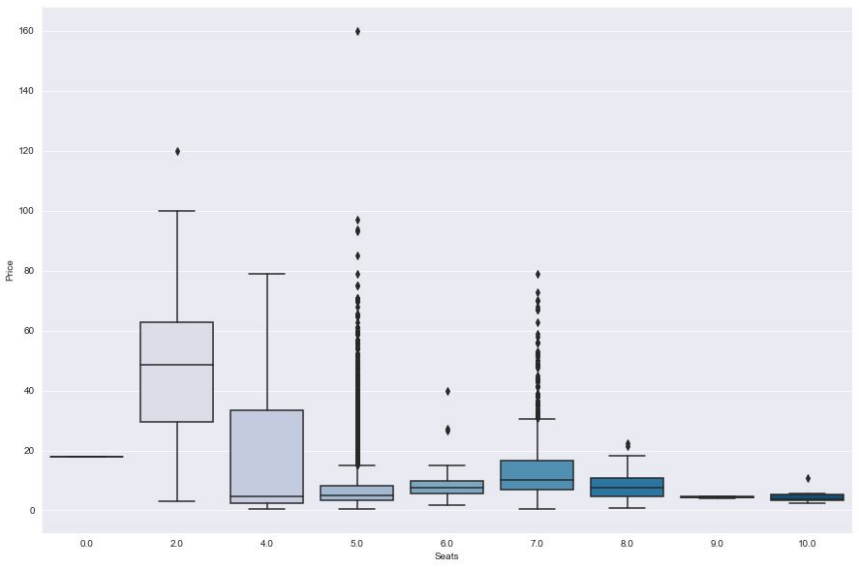
POWER BHP VS PRICE



- There is a positive relationship between power and price
- Engine and Power are positively correlated (0.90)

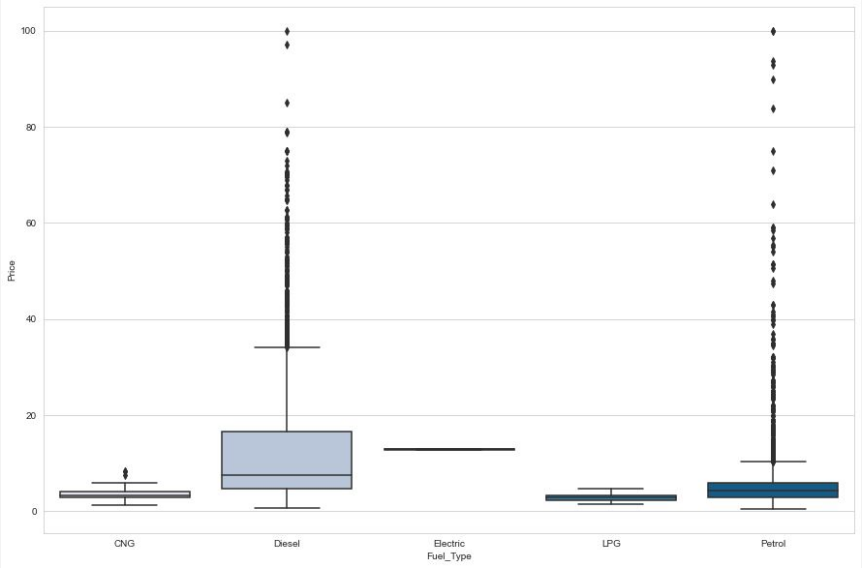
VEHICLE PRICE ANALYSIS
SEATS AND FUEL TYPE

NO. SEATS VS PRICE



- Cars with 2 seats have the highest average and median price
- 5 seats is the most common. There is one big outlier.

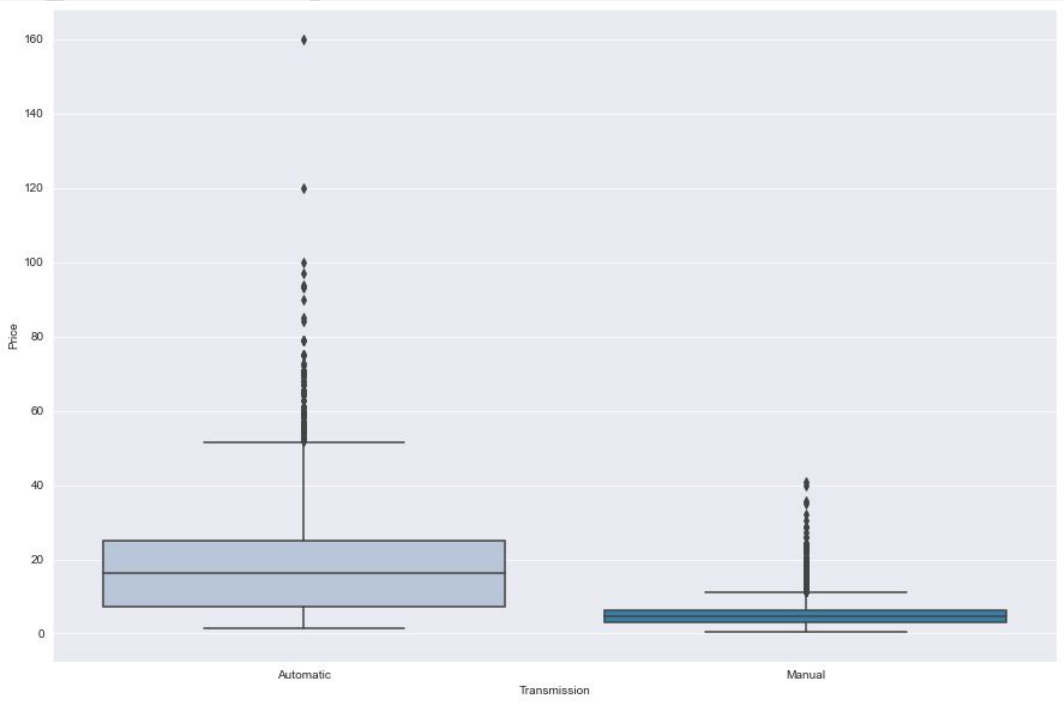
FUEL TYPE VS PRICE



- Diesel cars have the highest median price

VEHICLE PRICE ANALYSIS
TRANSMISSION TYPE

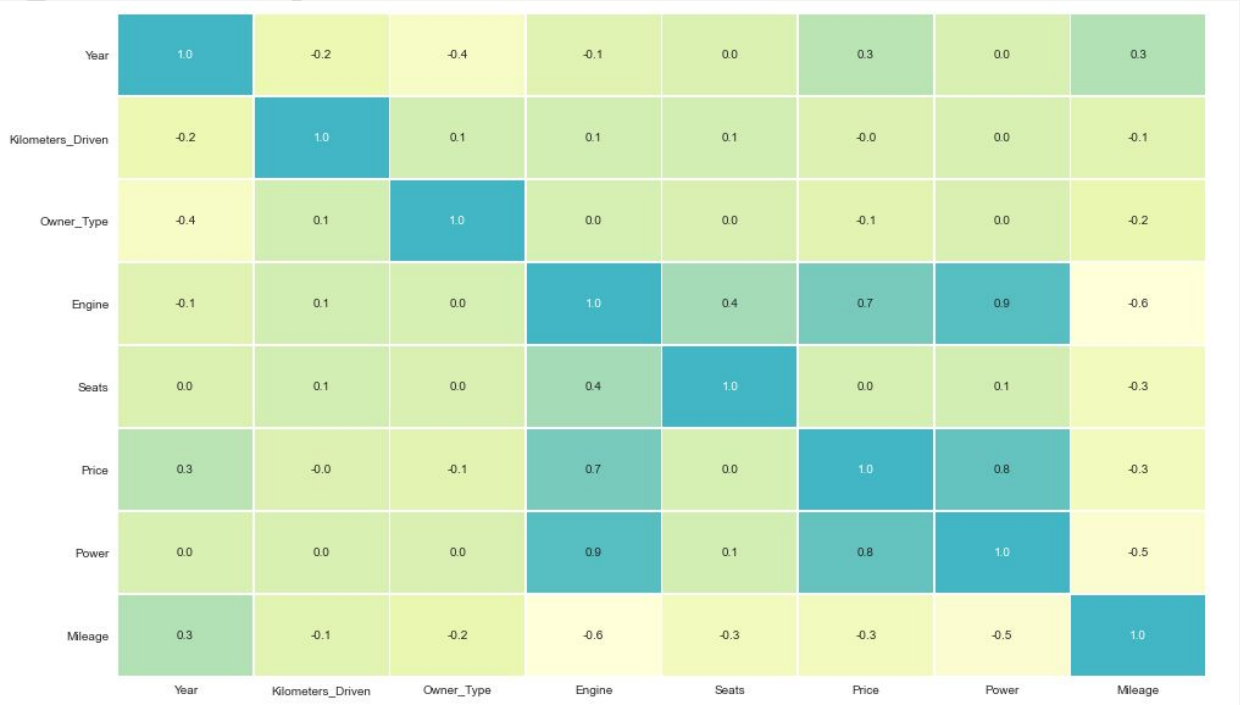
PRICE VS TRANSMISSION TYPE



- Automatic transmission vehicles command higher prices than do manual cars

VEHICLE PRICE ANALYSIS
VARIABLE CORRELATION ANALYSIS

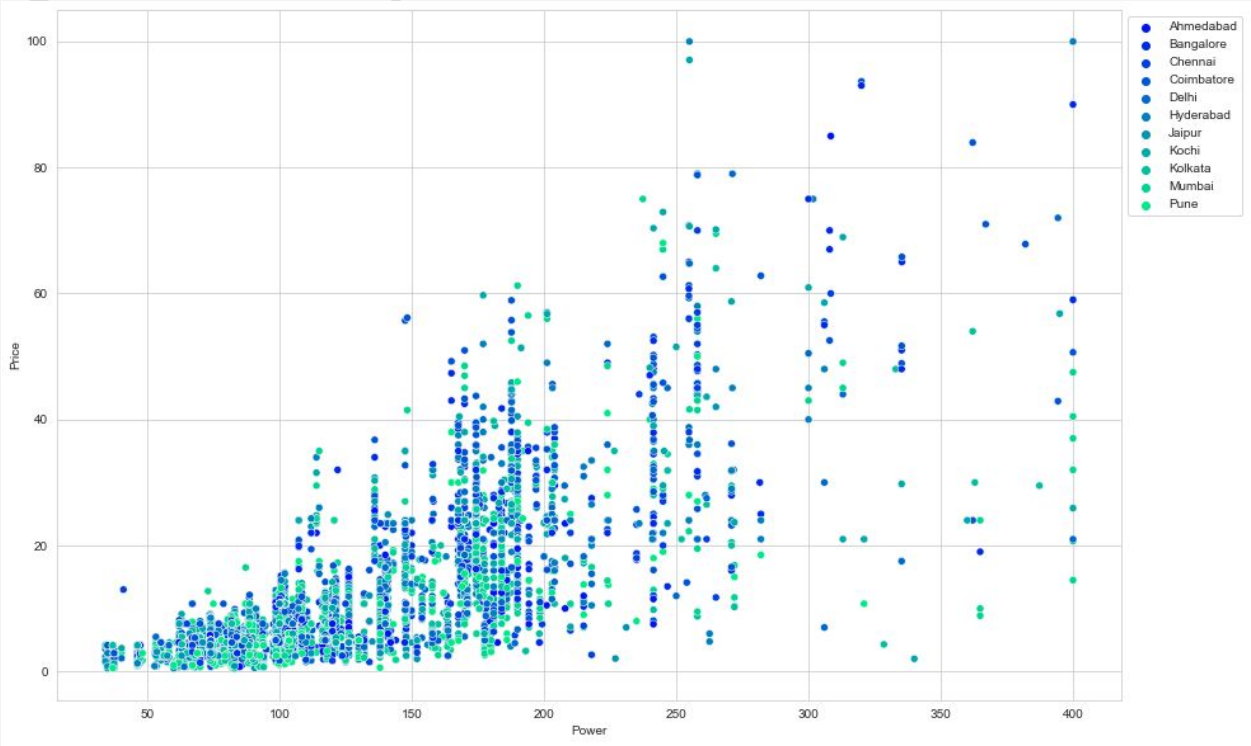
CORRELATION HEAT MAP



- Price has strong correlation to Engine (0.7), Power (0.8)
- Price has a negative relationship with Kilometers_Driven (-0.2) and Mileage (-0.3)
- Engine and Power are highly correlated (0.9)

VEHICLE PRICE ANALYSIS
MULTIVARIATE

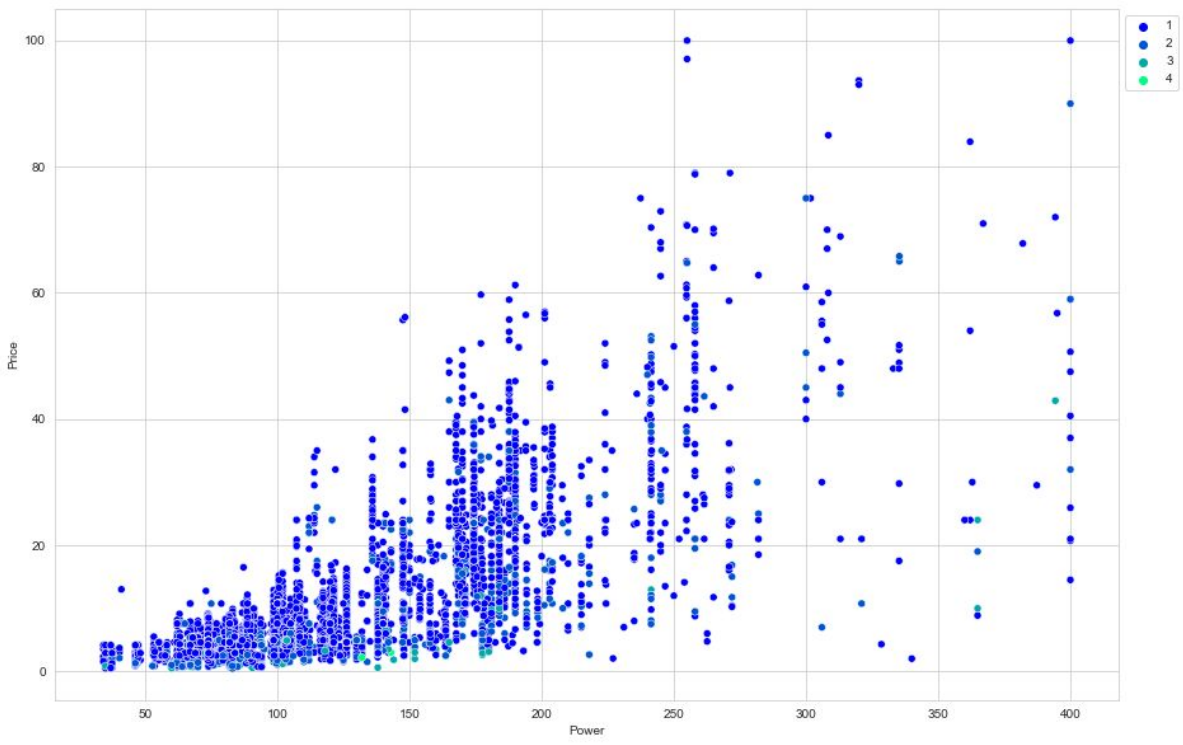
PRICE VS POWER VS LOCATION



- Cars with more powerful engines in bigger cities have higher prices

VEHICLE PRICE ANALYSIS
MULTIVARIATE

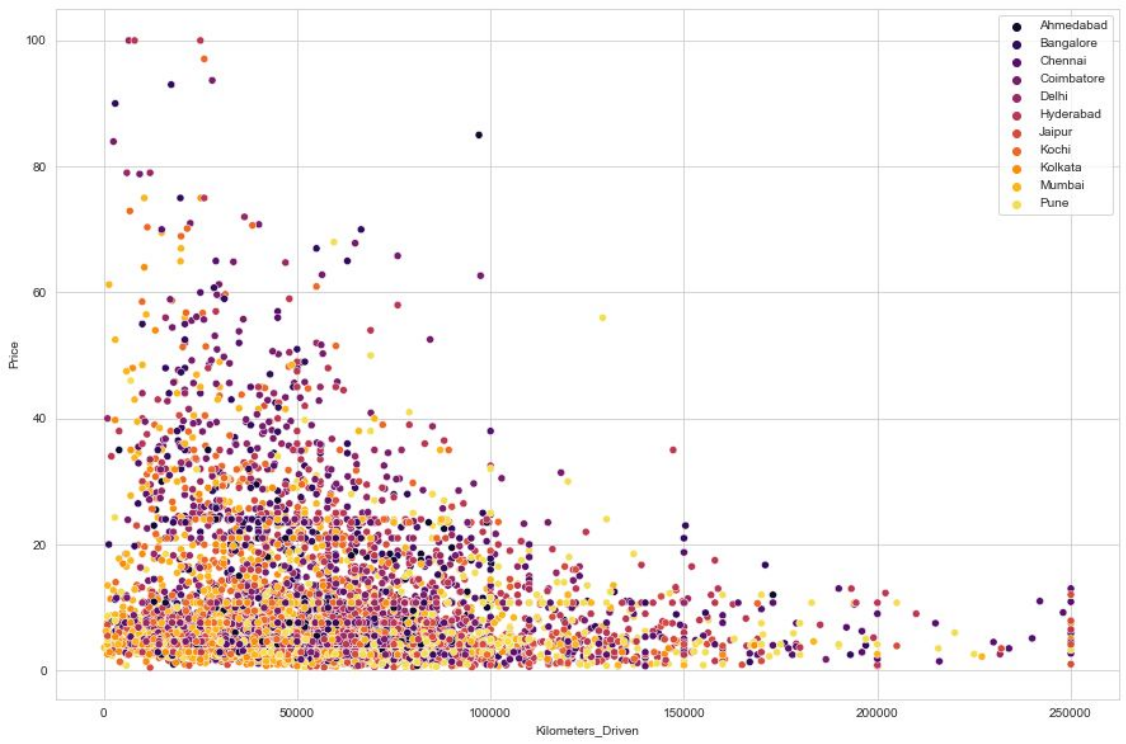
PRICE VS OWNER TYPE VS LOCATION



- Cars with more powerful engines and first owners have higher prices

VEHICLE PRICE ANALYSIS
MULTIVARIATE

PRICE VS KM DRIVEN VS LOCATION



- Cars with less kilometers driven located in the biggest cities have the highest prices

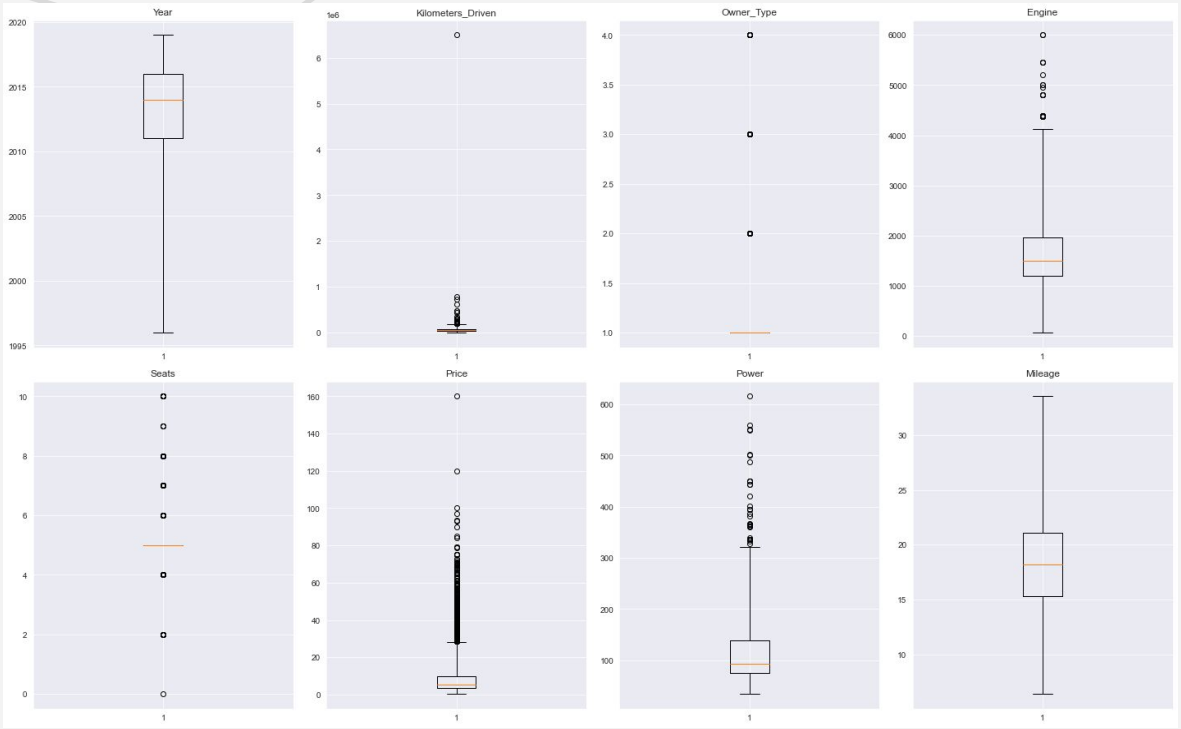
04

MODEL BUILDING

OUTLIER TREATMENT
LINEAR REGRESSION

OUTLIERS TREATMENT
OBSERVE OUTLIERS

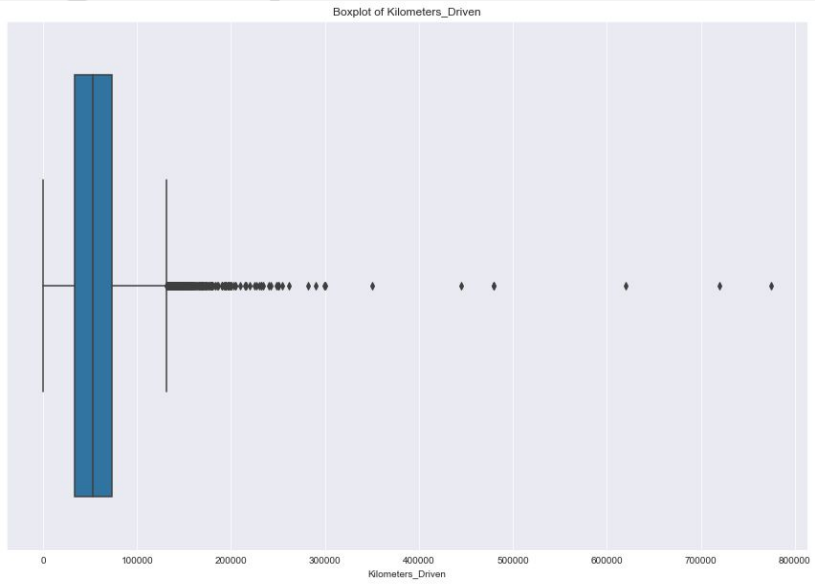
OUTLIERS



- Large outliers are heavily skewing the data
- Variables to treat are Kilomters_Driven, Engine, Power, and Price
- I will take a conservative stance and carefully cap extreme outliers which can skew the results without affecting the overall power of the variable

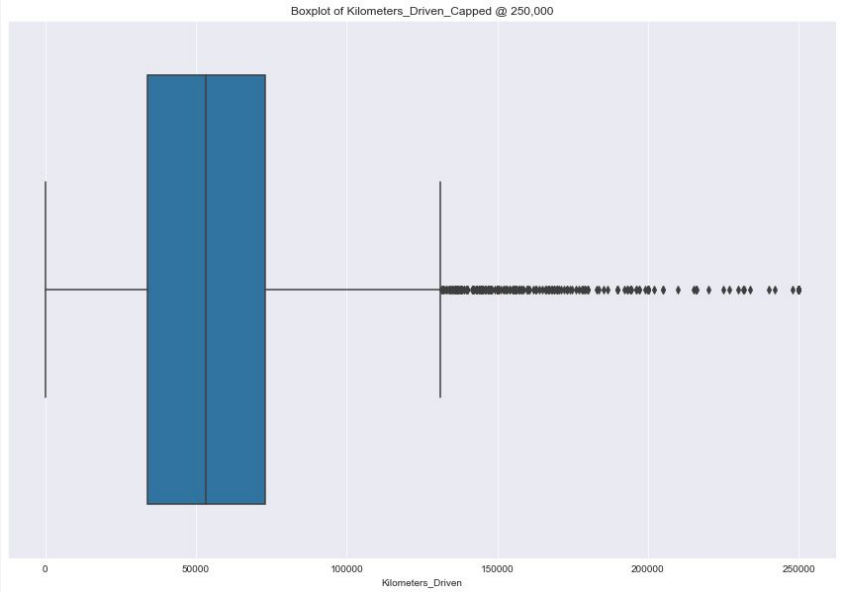
OUTLIERS TREATMENT
KILOMETERS_DRIVEN

KILOMETERS_DRIVEN BEFORE



- Large outliers are heavily skewing the data

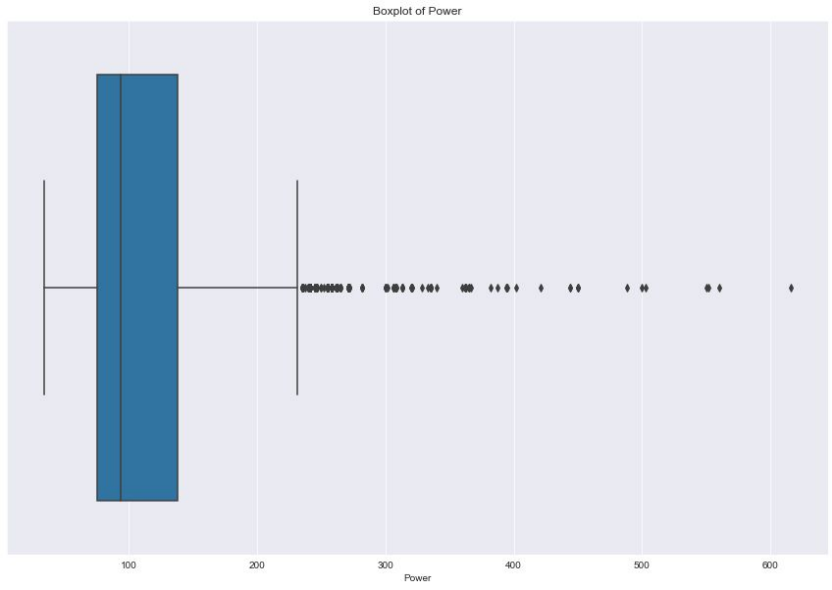
KILOMETERS_DRIVEN AFTER CAPPED @ 250000



- 16 outliers have been capped at 250,000
- Before mean = 58,699.06
- After cap mean = 57476.96

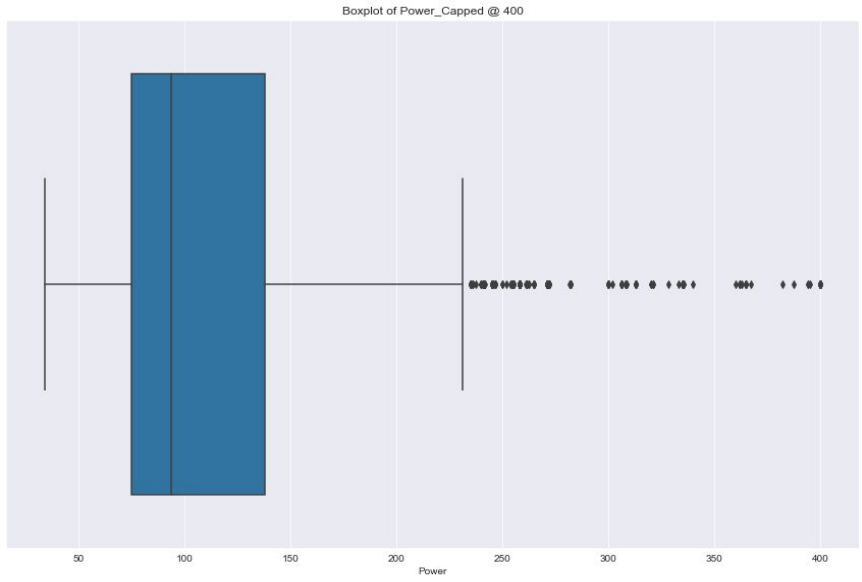
OUTLIERS TREATMENT
POWER

POWER BEFORE



- Large outliers are heavily skewing the data

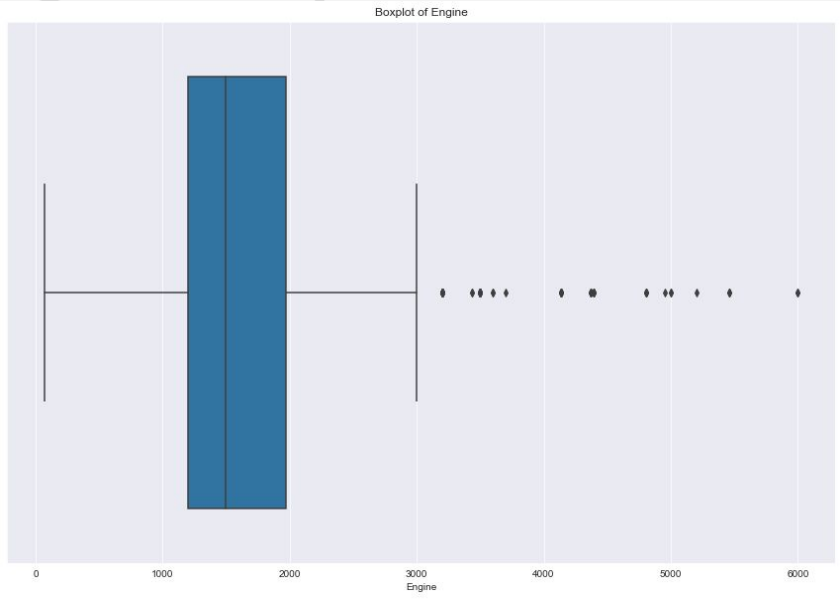
POWER AFTER CAPPED @ 400



- 14 outliers have been capped at 400
- Before mean = 112.46
- After cap mean = 112.26

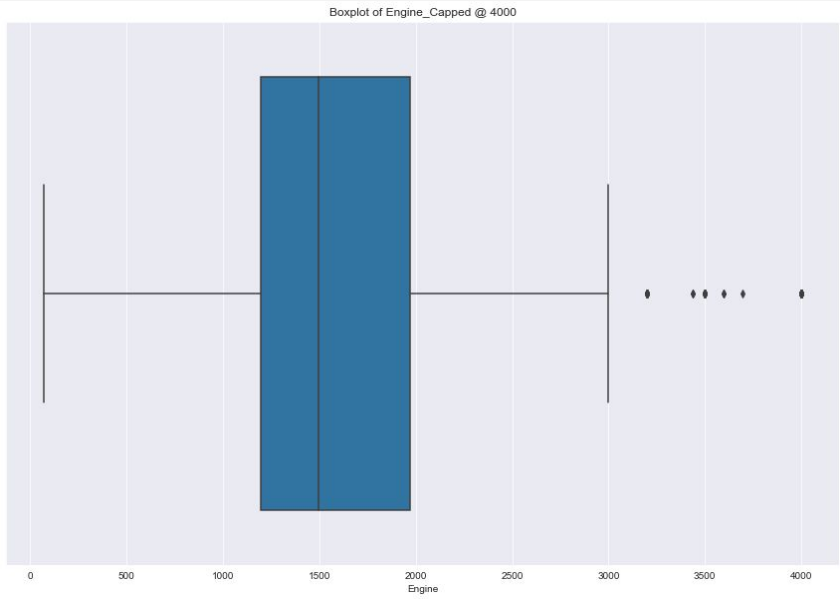
OUTLIERS TREATMENT
ENGINE

ENGINE BEFORE



- Large outliers are heavily skewing the data

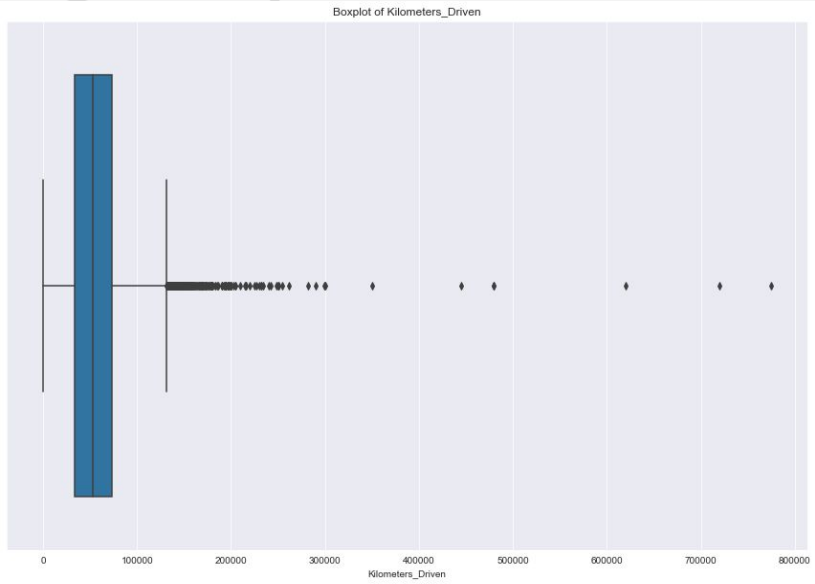
ENGINE AFTER CAPPED @ 4000



- 16 outliers have been capped at 4,000
- Before mean = 1615.82
- After cap mean = 1613.03

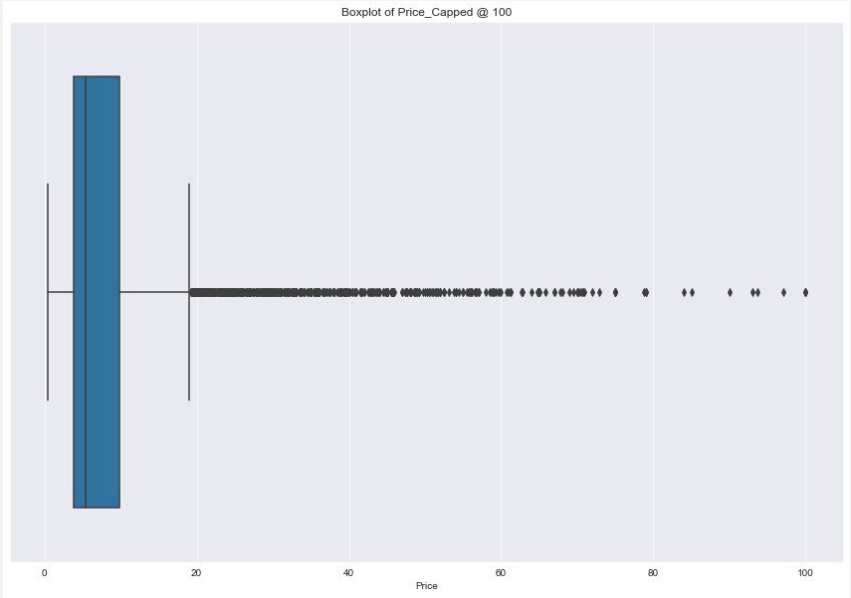
OUTLIERS TREATMENT
PRICE

PRICE BEFORE



- Large outliers are heavily skewing the data

PRICE AFTER CAPPED @ 100



- 3 outliers have been capped at 100
- Before mean = 9.24
- After cap mean = 9.23

OUTLIERS TREATMENT
FINAL DATASET BEFORE DUMMY VARIABLES

VARIABLE DESCRIPTION

	Year	Kilometers_Driven	Seats	Price (INR Lakh)	Owner_Type	Engine	Power	Mileage	Kilometers_ Driven_LOG	Power_LOG
Count	7251	7251	7251	7251	7251	7251	7251	7251	7251	7251
Mean	2,013.37	57503.51	5.28	9.24	1.20	1613.22	112.29	18.33	10.76	4.63
Standard Deviation	3.25	34328.87	0.81	10.51	0.46	580.13	52.08	4.15	0.71	0.41
Minimum	1996.00	171.00	2.00	0.44	1.00	72.00	34.20	6.40	5.14	3.53
25%	2011.00	34,000.00	5.00	3.75	1.00	1198.00	75.00	15.30	10.43	4.32
50%	2014.00	53392.00	5.00	5.35	1.00	1493.00	93.70	18.20	10.89	4.54
75%	2016.00	73,000.00	5.00	9.89	1.00	1968.00	138.03	21.10	11.20	4.93
Maximum	2019.00	250000.00	10.00	100.00	4.00	4000.00	400.00	33.54	12.43	5.90

- Outliers have been successfully treated and we are ready to build a model
- Added natural logarithm transformations for Kilometers_Driven and Power

05

KEY INSIGHTS & RECOMMENDATIONS

LINEAR REGRESSION MODEL
PERFORMANCE SUMMARY

Key Question
Can we build a linear regression model to help managers accurately predict used car values?

Objective
Build a predictive pricing model for used cars using linear regression. Leverage this model to help managers maximize profit and minimize risk associated with used car sales.

Training Details
Dataset shape (7251, 53)
Log transformations did not impact performance

Independent Variable = Price
Dependent Variables = Year, Kilometers_Driven, No. of Seats, Power, Mileage, Location, Fuel_Type, Transmission, Brand_Name

Performance Details
Adjusted R^2 = 0.776 (Training data R2 = 0.7796) - The model explains 77.6% of the variability in the independent variable (Price)

Train error: Residual Mean Squared Error = 5.006
Test error: Residual Mean Squared Error = 5.065

Conclusion
This model can help managers predict the price of used cars in India by explaining 77% of the price using the provided data.
Key insights and business recommendation on following pages

KEY INSIGHTS

PREDICTIVE STATISTICS - Observations

- The year the car was made is a significant factor in determining price (0.7275 * year). Newer cars have higher prices.
- The amount of kilometers driven reduces value by small amount
- Cars with more seats command higher prices in general with a coefficient of 0.4462
- The amount of Power a vehicle has is small positive factor in price at 0.0840
- Higher mileage cars have lower prices with a coefficient factor of -0.1584
- The location of the vehicle can have a powerful impact on vehicle price.
- Bangalore has the most expensive cars at 1.62 * price followed by Coimbatore at 1.47 and Chenai at 1.35
- Kolkata (-1.00), Delhi (-0.90) and Mumbai (-0.8382) had the biggest negative impact on price
- Fuel Type can impact price. Petrol cars were cheaper at (-1.65)
- In contrast to Electric vehicles which command premium prices at + 8.67
- Brand Names can also impact price with most Brands having a negative impact in price except for Lamborghini which can increase price by 14.7685

BUSINESS RECOMMENDATIONS

1. **Key Insight: Year of Manufacture has positive impact on price**
   Recommendation: The company should look for late model vehicles manufactured in the last 2-3 years

2. **Key Insight - The number of seats (i.e. car size) is positive determinant of price**
   Recommendation: Look for larger vehicles to have higher resale values

3. **Key Insight - The location of the vehicle is the most powerful determining factor for price**
   Recommendation: The company should look to acquire used vehicles in lower priced market such as Delhi and Kolkata and move them to higher priced market such as Bangalore, Chennai, and Mumbai

4. **Key Insight - Brand is not an important factor but electric vehicles have 8.7x higher resale prices**
   Recommendation: Create an electric vehicle strategy as they may offer the biggest future potential for growth