**Web Scraping (Data Mining) with EDA Project**

1. **Problem Statement**

Analysis of data of Used Cars from [*https://cardekho.com/*](https://cardekho.com/)

Domain : Website has Used Car Sales data

Related websites : cars24.com, carwale.com, cartrade.com

Reason for choosing cardekho.com :

Provided more number of features for web scraping

Website allows for web scraping without any denial of 404 error

Data can be scrapped using wide range of filters available in it

1. **Data Collection**

BeautifulSoup was used to collect the data scrapped from the website. While scraping itself, few variables were handled and cleaned as per the requirement.

The for loop is run on the url by considering different filters from the above cell.

Since the data is huge if the cities (12 cities) are considered on whole, they were chosen individually in different notebooks to extract the data from the cities individually.

Later, they were put into a data frame (df)

Number of rows and columns in the data frame: 12954 rows and 11 columns

Column names in the raw file of scraped data :

* 1. carname = name of the car along with year of registration and brand-variant
  2. city - place where car being sold['ahmedabad','bangalore','chennai','delhi-ncr','gurgaon','hyderabad','jaipur','kolkata','mumbai','new-delhi','noida','pune']
  3. car\_price = price of the car quoted
  4. body\_type=['sedan','Hatchback','SUV','Luxury','MUV','Minivan','Convertible','Coupe','Hybrid','Wagon','super-luxury','hybrid']
  5. tranmission= ['manual','automatic']
  6. seller\_type= ['dealer','individual']
  7. owner\_type= ['unregistered','first-owner','second-owner','third-owner-or-more']
  8. color= ['White','Silver','Red','Blue','Black','Brown','Orange','Green','violet','Yellow','Gold','Other']
  9. Year = data taken from the car name using regex
  10. Fuel\_type = Fuel used, km\_run and locality of the car sold

1. **Data Understanding and Cleaning**

Regex Usage to clean data :

The column Fuel\_type had to be further cleaned to extract the needed values of km\_driven, fuel\_used and locality into separate columns. Similarly, column CarName had to be further cleaned and processed to extract Brand column in the data frame. Both these requirements were fulfilled using regex.

Duplicate Data: The cities new-delhi,Gurgaon,Noida and Delhi-NCR had same data mentioned across all the cities. Hence deleted 3 cities out of the above four. Only data pertaining to Delhi-NCR were retained.

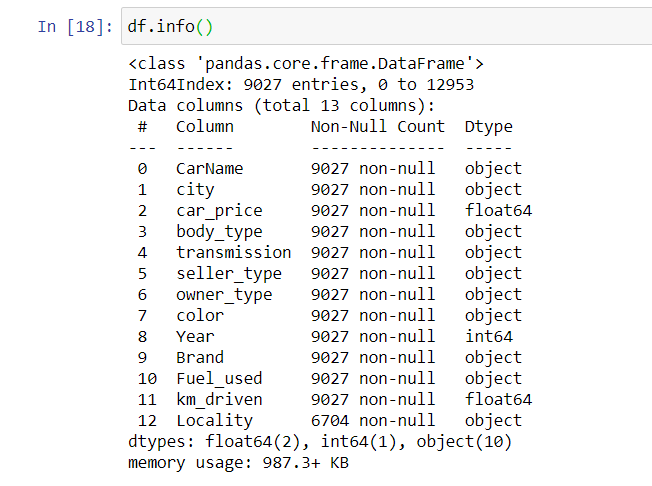
Junk Data: The column owner\_type had four options : first,second, third & more and unregistered.

Unregistered Car meant its New Car, where the problem statement deals with Used Cars. Hence, we had to delete all data pertaining to this junk unregistered value in owner\_type column.

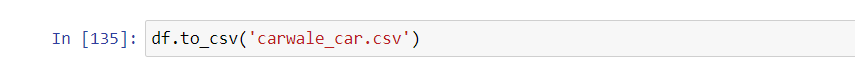
After cleaning and removing (duplicate & junk) data and also modifying the columns, the following are the properties of the table derived:

Rows: 9027

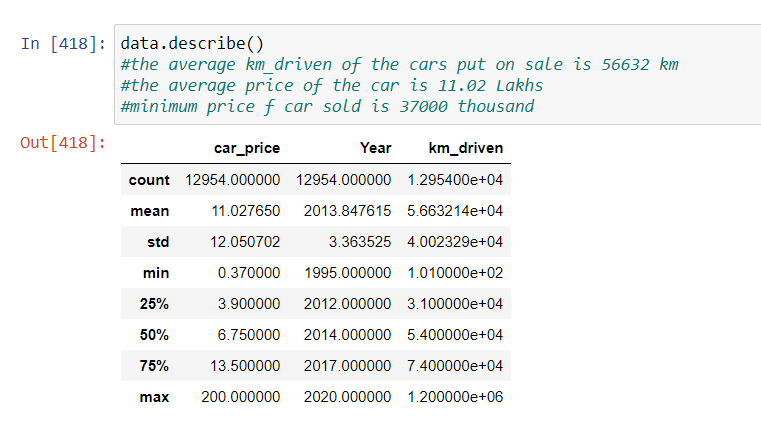
Columns: 13



1. **Converting the cleaned and usable data into csv file**



1. **Data Investigation & Analysis**



**Conclusion :**

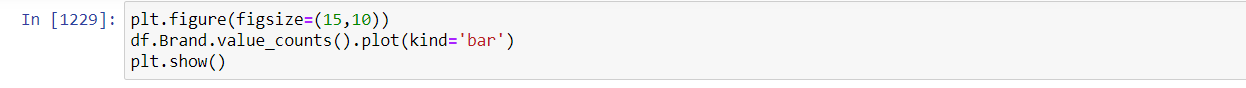
The average km\_driven of the cars put on sale is 56632 km

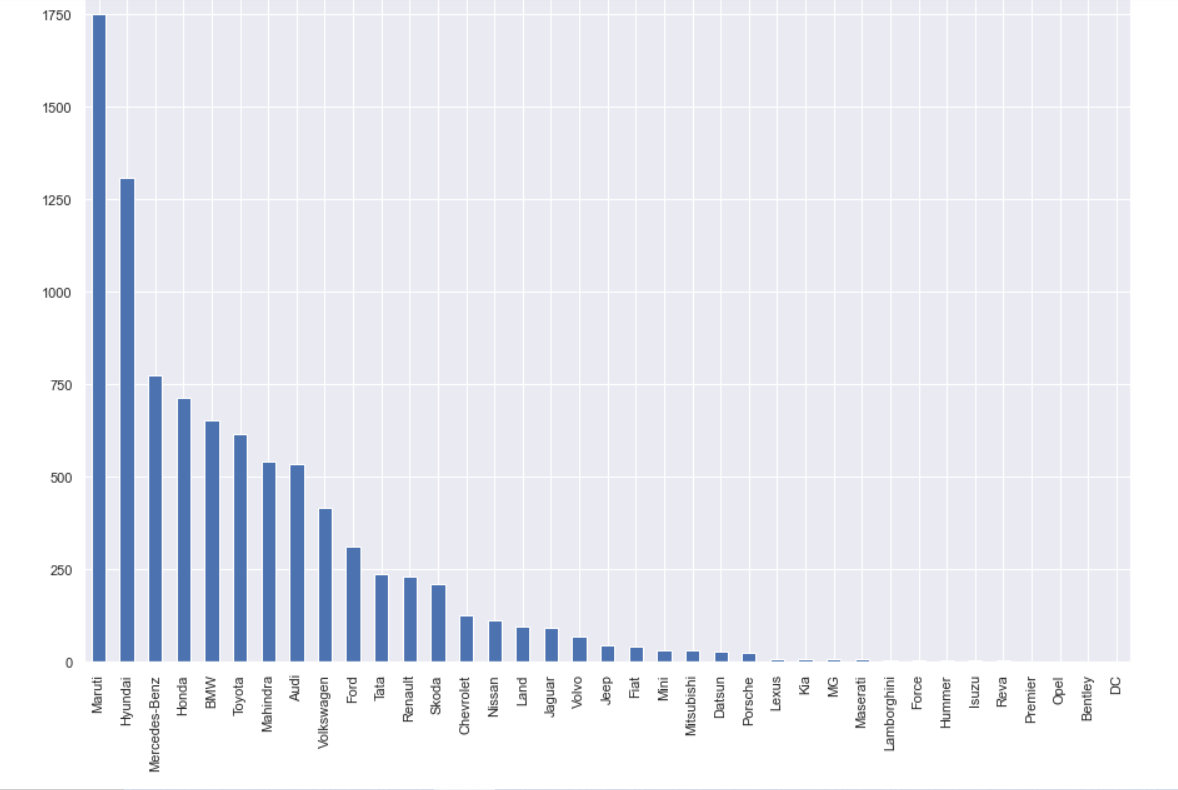
The average price of the car is 11.02 Lakhs

The minimum price of car sold is 37000 thousand

* + 1. **Univariate Analysis**

1. The count of cars according to brands.



****

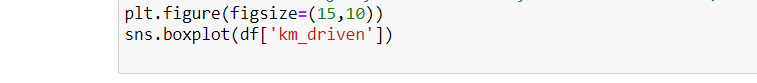
**Conclusion**

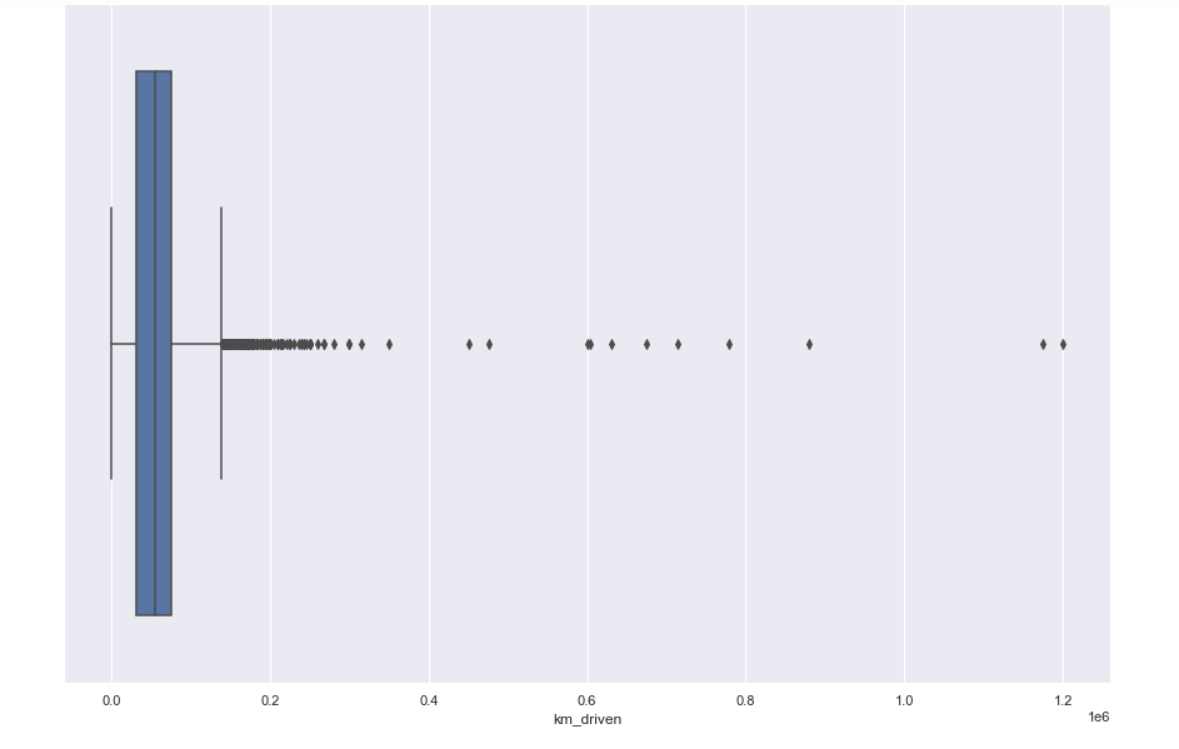
Maruti cars were present the most in non-luxury variants

Hyundai were second

Surprisingly, Mercedes and BMW were also sold comparatively to other brands.

1. The range of kilometres driven of the cars that were being sold



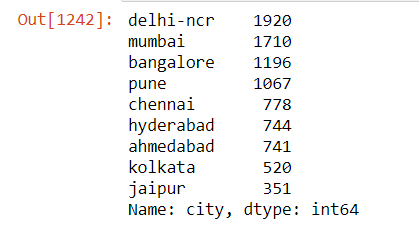


**Conclusion**

The range of kilometers is from 30000 to 70000 , rest of all are outliers

* + - 1. Cars sorted as per city with delhi-ncr having the highest no. of cars and jaipur the least.

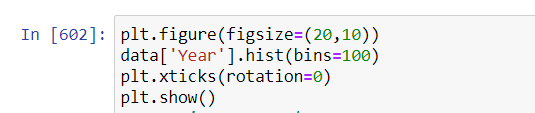


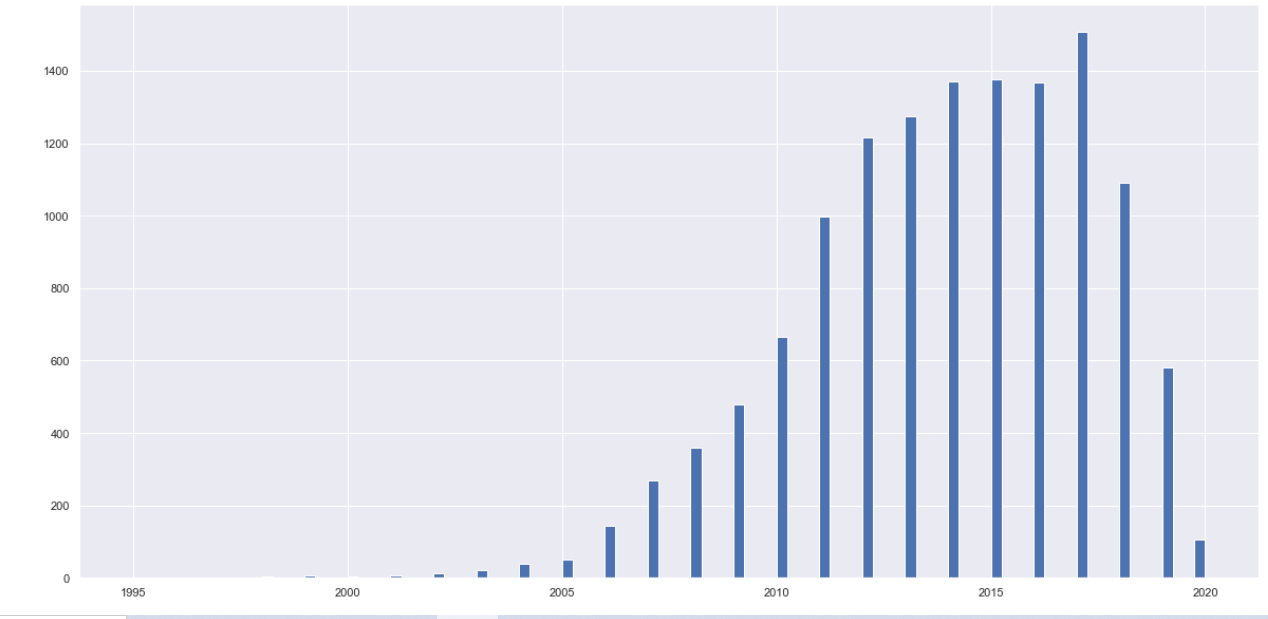


**Conclusion**

Cities like kolkata , chennai , hyderabad, ahmedabad and jaipur may be relying on public transport whereas the metros like delhi mumbai and bangalore and pune are more towards own-transport

* + - 1. Registration year-wise cars sale

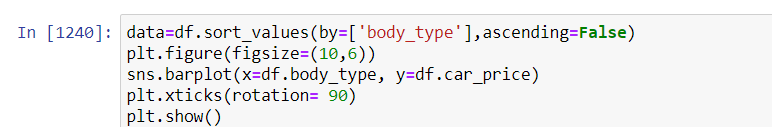


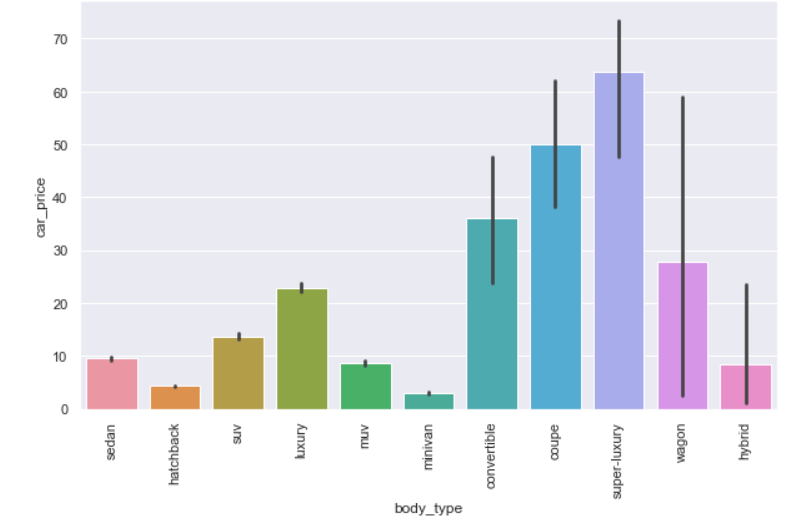


**Conclusion**

2017 registration year cars saw the highest number of sales. which is periodically declining towards 2020

* + 1. **Bivariate Analysis** 
       1. Minimum, max and mean of the car price according to the body type

****

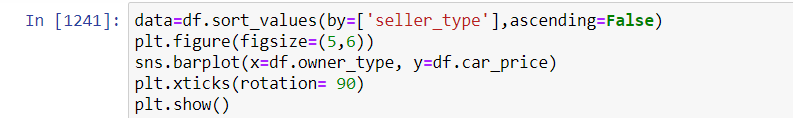
****

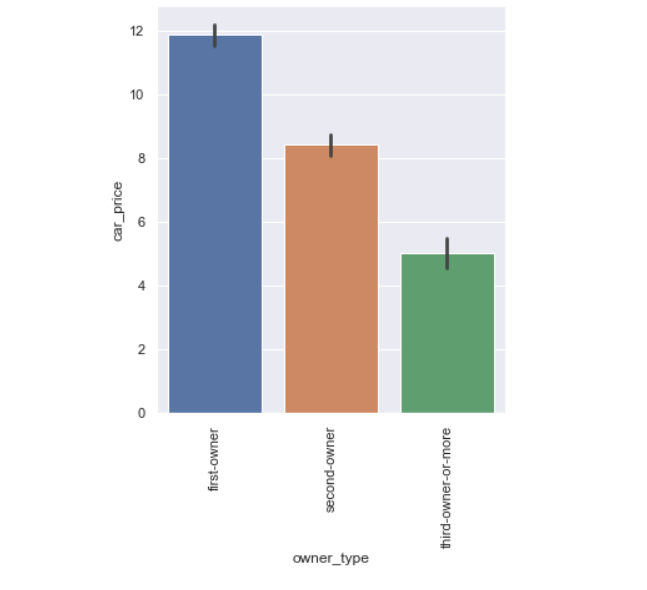
**Conclusion**

The costliest ones are the super luxury cars and minimum are hatch-back and minivan types.

Luxury and wagon types have costs in mid-range.

* + - 1. Car Price for a given type of owner

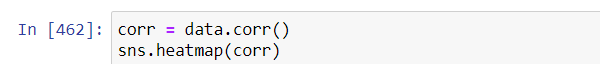


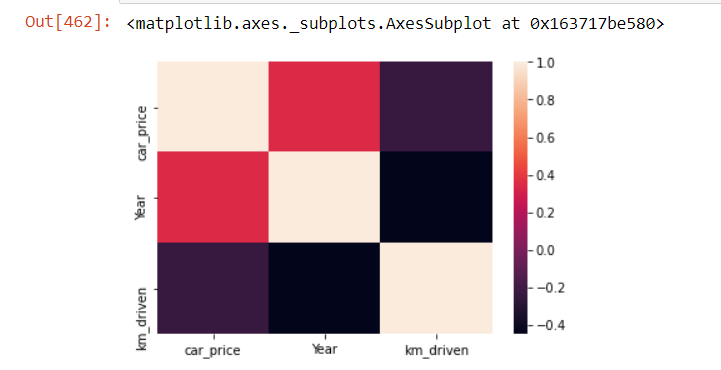


**Conclusion :**

As it is obvious, first owner cars costs higher followed by second and third owner cars

* + 1. **Multivariate Analysis**
       1. Correlation (via heatmap) between car\_price, Year and Km\_driven columns





**Conclusion**

The older the car the more is the no. of km\_driven

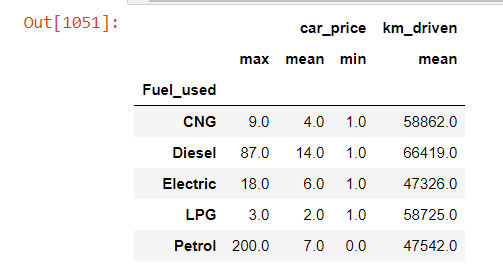
The more the number of km\_driven lower is the price.

More than age of the car , it is the km\_driven that decides the price of the car.

* + - 1. Car statistics (price, km-driven) for a given owner-type as per the fuel used

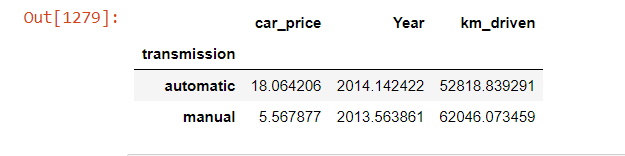
Code :

pd.pivot\_table(index =['Fuel\_used', 'owner\_type'],data=df, aggfunc={'km\_driven': np.mean,'car\_price': [min,max,np.mean]}).round()



* + - 1. Car price and km driven as per transmission type



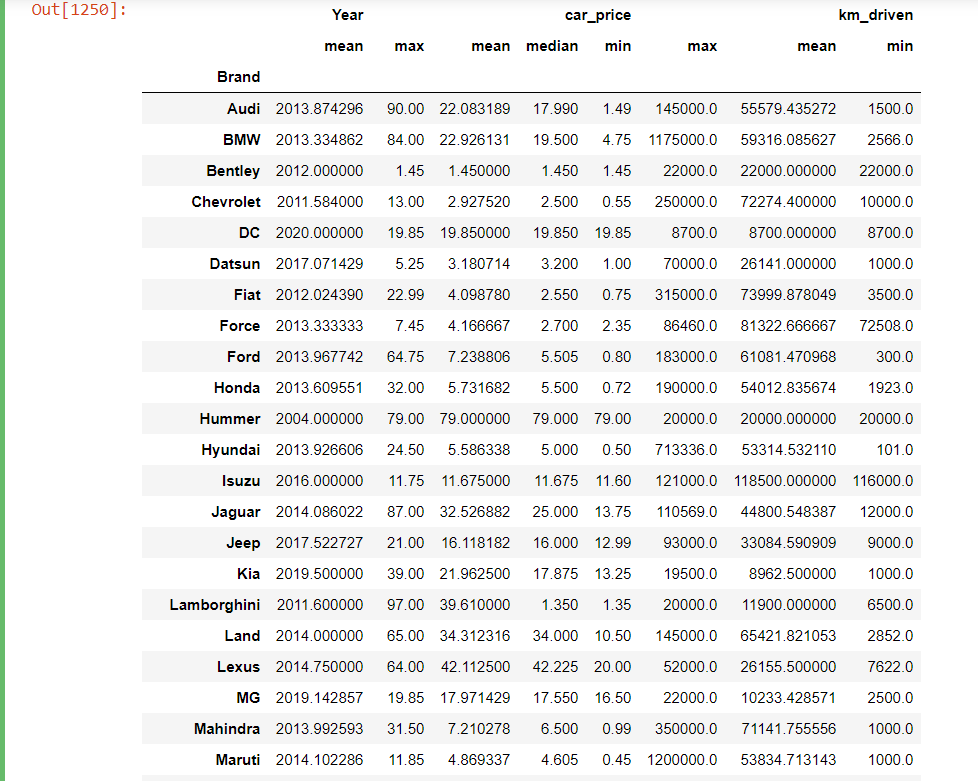


**Conclusion :** The cost of automatic cars is thrice the cost of manual cars.

1. **Client Requirement**
   * + 1. The statistics of car\_price and km\_driven as per the brand

Code:

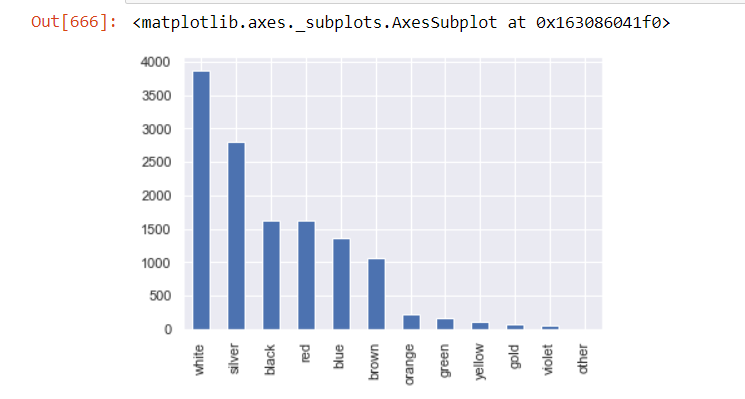
pd.pivot\_table(data, index =['Brand'], aggfunc={'km\_driven': [min , max, np.mean],'car\_price': [min , max, np.mean, np.median],'Year': np.mean})



**Conclusion :** For the brand maruti- the minimum and maximum car price is 45000 and 11.85 lakh respectively with an average of 4.86 lakhs and for the same brand maruti - the minimum km\_driven is 1000 and maximum is 12 lakh kms and an average of 53800 kms

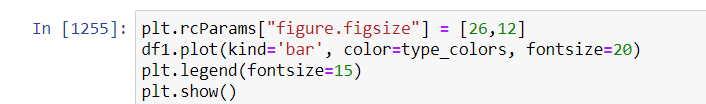
* + - 1. The cars sold according to color

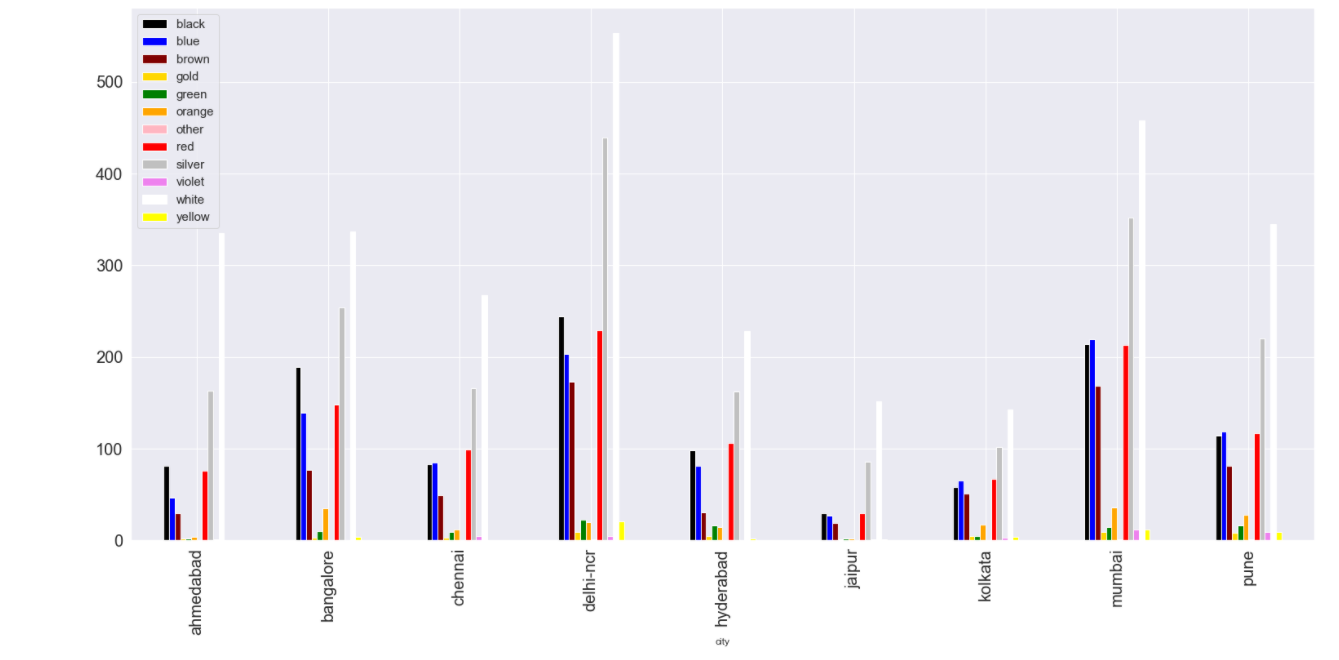




**Conclusion** : White cars are the maximum to be sold followed by silver and black where as rare colors of violet , gold , yellow , etc are the least to be sold

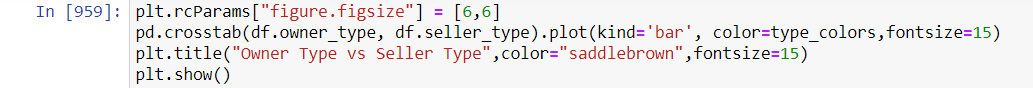
* + - 1. The cars available in a particular city according to color

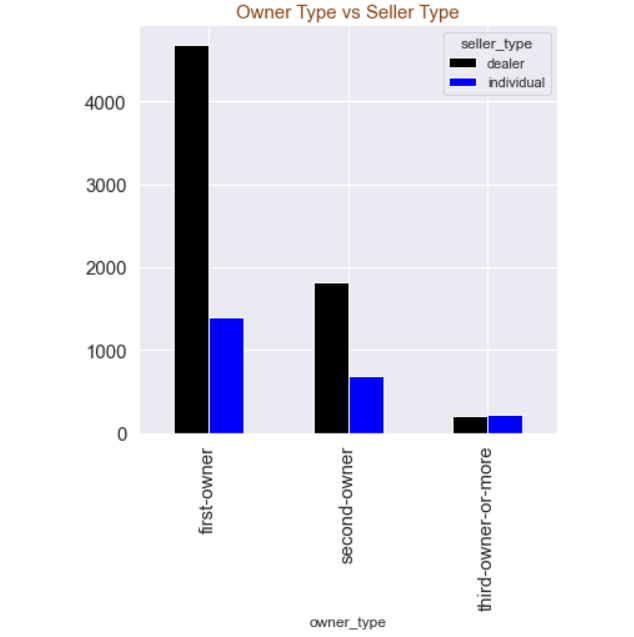




**Conclusion :** white cars are most sold cars across the city

* + - 1. The number of cars sold for different owner types based on whether the seller is a dealer or individual

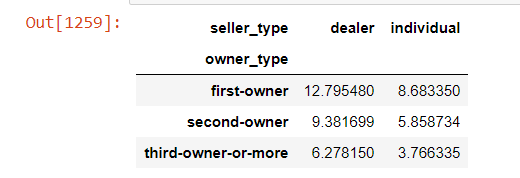




**Conclusion :** People are preferring to sell the car via dealers

* + - 1. The average price of cars being sold for a given owner\_type and seller type

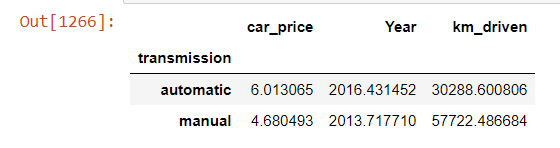




**Conclusion** : First owner cars sold by a dealer are costly whereas third owner cars sold by individually cost the least. For a customer who is willing to sell a car, irrespective of the owner-type, its preferable that he sells it via a dealer

* + - 1. Statistical analysis for the most sold car maruthi



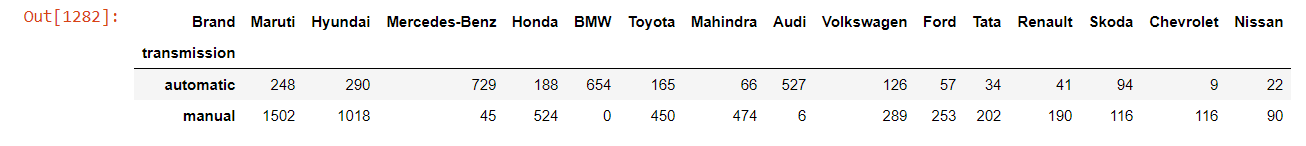


**Conclusion** : Manual maruthi cars are used to the maximum extent than automatic cars.

The older the better , manual cars have overtook automatic cars in terms of usage

* + - 1. Brands sold according to transmission type

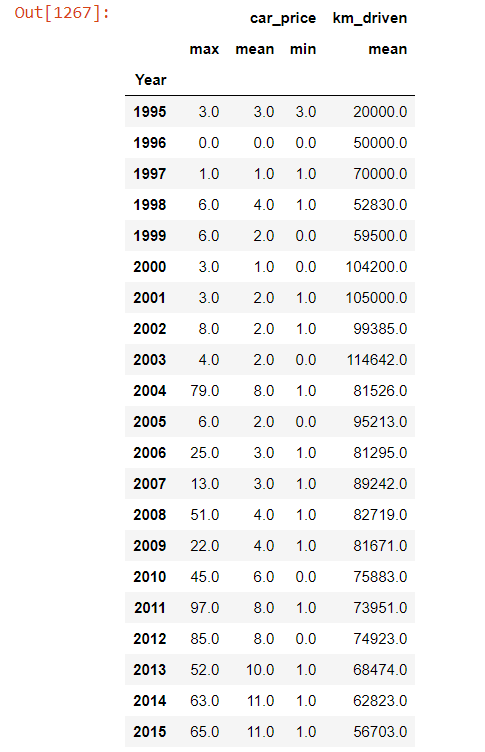




**Conclusion :** The luxury cars like mercedes and bmw are mostly having automatic transmission. Audi car has the least manual cars for sale

* + - 1. Average rate of cars being sold according to the year



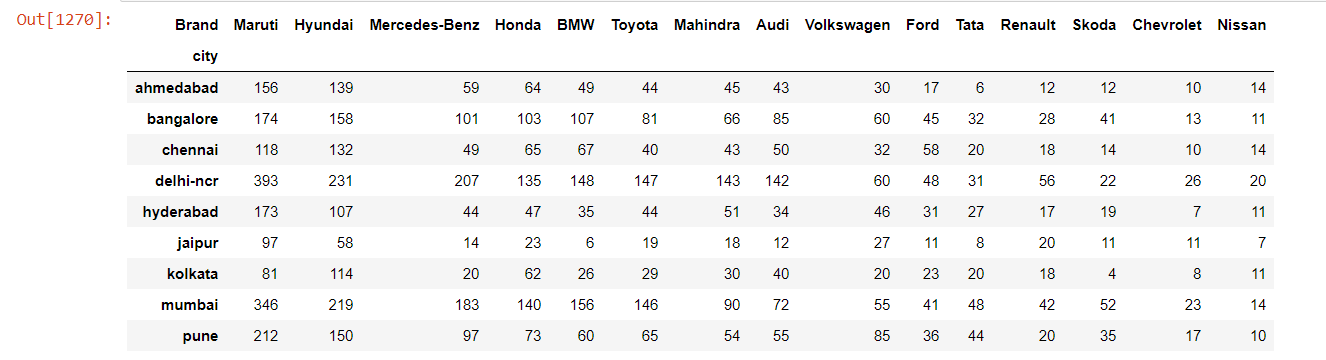




**Conclusion :** The average rates of car are steadily increasing in the modern times

* + - 1. Preferred brands sold in cities

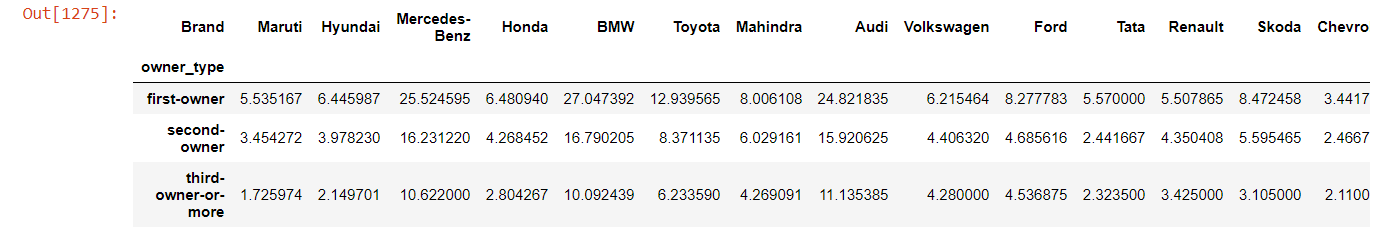




**Conclusion** : Maruti was sold the highest in all the cities except for chennai and kolkata where the hyundai was sold the highest

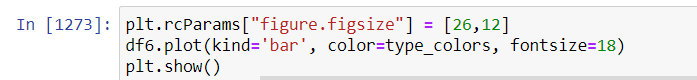
* + - 1. The average car rates for a given brand according to owner type

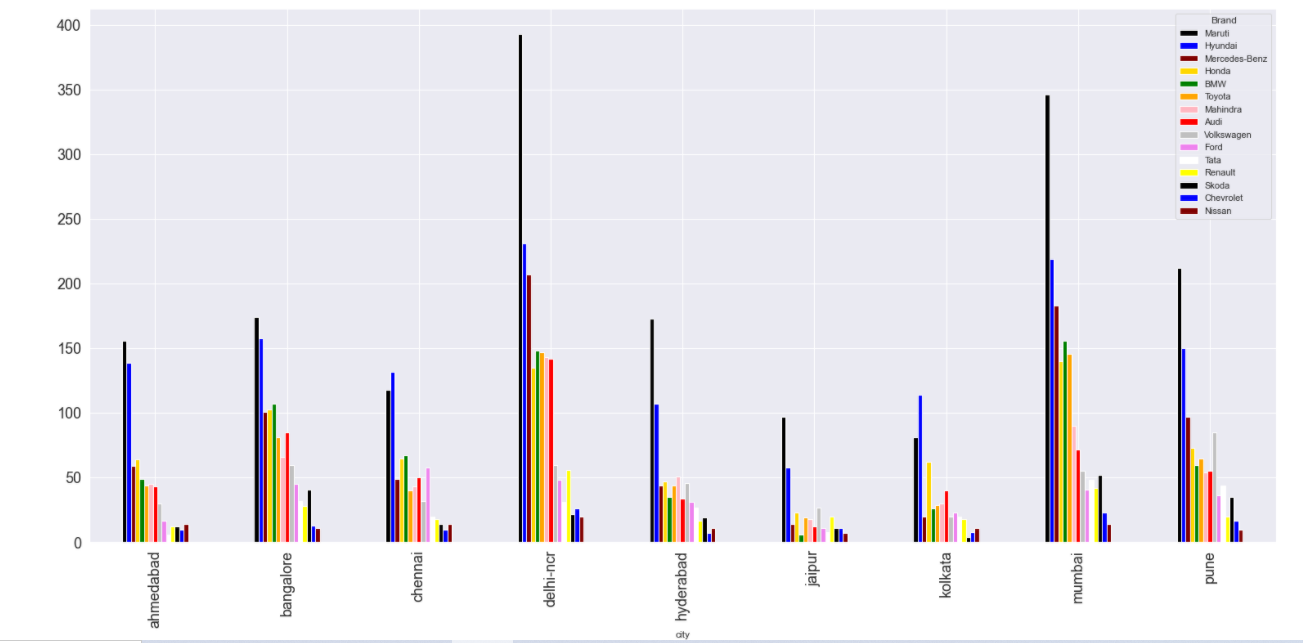




**Conclusion** : The average rates of car are steadily increasing in the modern times

* + - 1. Depiction of car brands, sold the maximum according to cities





**Conclusion :** Maruti is the highest sold brand followed by Hyundai across cities

**Submitted By :**

Adarsh Challa

Sandhya CSR