Great Learning
POWER AHEAD

A Regression Study of Price Determinants of used cars in US Automobile markets

BY: GROUP 5	
ANKUR KUMAR	MUSKAAN PASSI
LAKSHYA SHARMA	KRITAGYA KASHYAP
ADITYA KAPUR	ABHINAV TYAGI
JASWANT SINGH	

MENTORED BY:
MS. VIBHA SANTHANAM

- **PROBLEM STATEMENT**

  A Study of Price Determinants of used cars in US Automobile markets

- **BUSINESS IMPACT**

  A used car is a pre-owned or second hand vehicle, which is repaired and refurbished to regain working conditions before the sale. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features.

DATA OVERVIEW

Used Cars Dataset Vehicles Listing From Craigslist.org

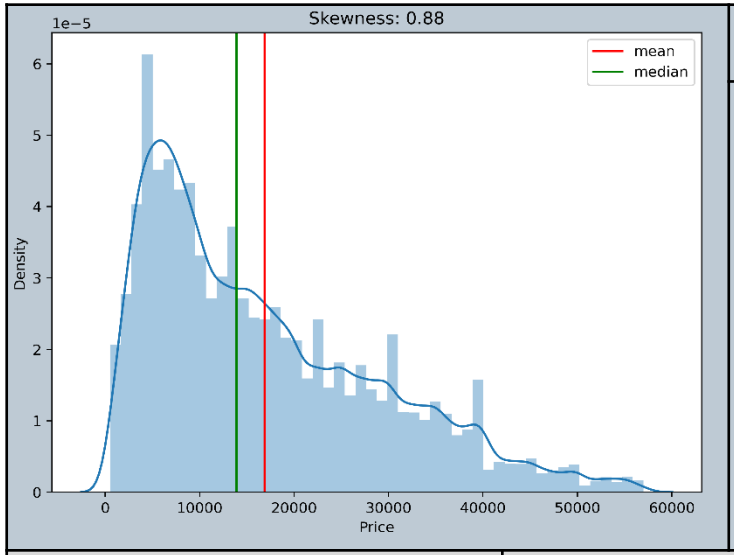
Categorical Columns		Numerical Columns	Unique	Date and Time
• Region • Manufacturer • Model • Condition • Cylinders • Fuel • Title Status • Region URL	• Transmission • Drive • Size • Type • Paint Color • State	• Price • Lat • Long • Odometer • County • Year	• I.D. • URL • Image URL • VIN • Description	• Posting Date

Data Pre-Processing and Feature Engineering

Removing Redundant Columns	Columns with too-many Sub categories
➢ I.D. ➢ Region Url ➢ Image Url ➢ Lat ➢ Long ➢ County	➢ Region (404) ➢ State (51) ➢ Model (29,668) ➢ Region URL (413) ➢ Manufacturer (43)

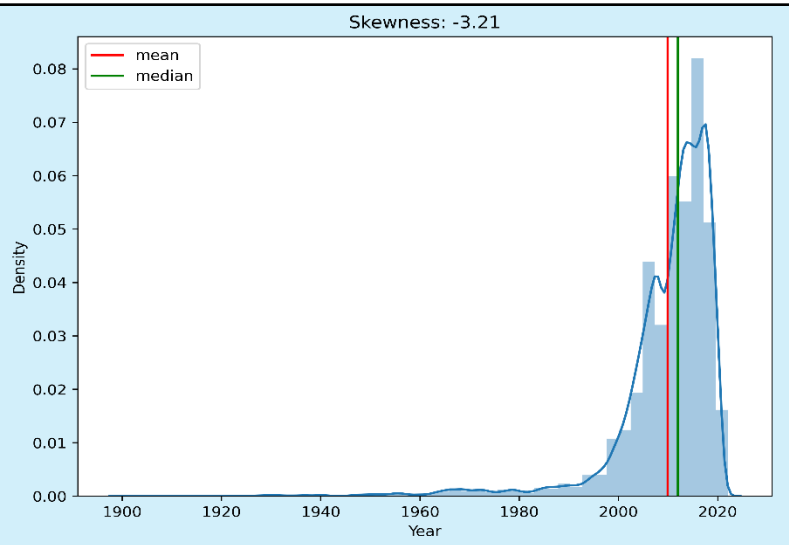
Introduced new columns	Columns that have Outliers	Columns which are converted into bins
➢ Age ➢ State Zone ➢ Is Tax ➢ Vin Car	➢ Odometer ➢ Price ➢ Year	➢ State (into State Zone) ➢ Paint colour ➢ Year (into Vint car)

EXPLORATORY DATA ANALYSIS – Univariate Analysis



Price(Target)

- Minimum value in price is 500 and highest is approx. 57,000 $.
- Mean price of cars is approx. 16,000.

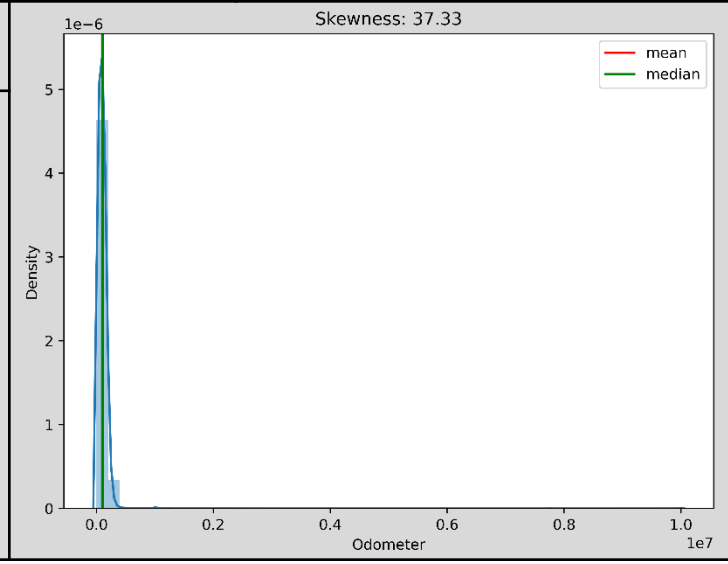


Year

- Maximum cars purchased between year 2010 to 2020.
- There are some Vintage and Antique cars from 1900 to 1970.

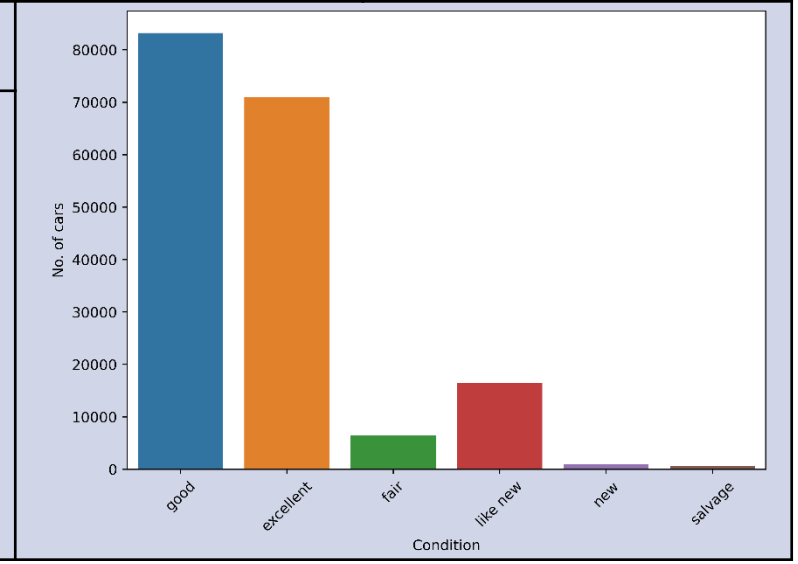
Odometer

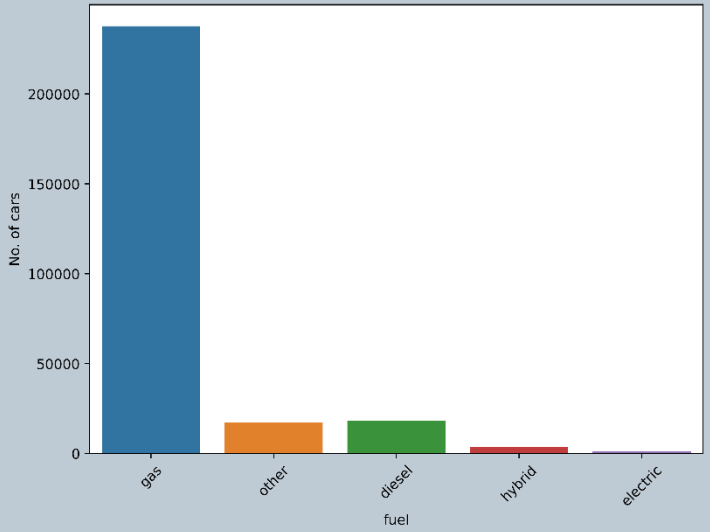
- There are some 0 values present as well as some values above 10 lakhs , which as per Domain Knowledge is not practically possible.



Condition

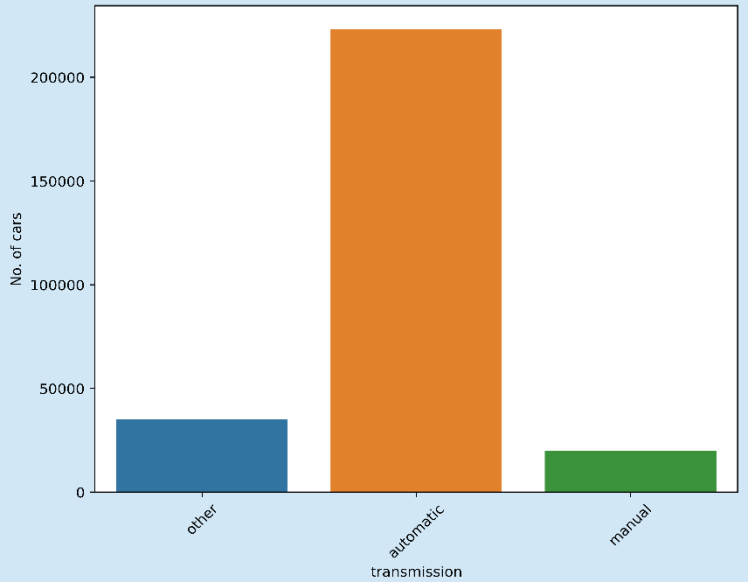
- 30% used cars are in good condition and 24% in excellent condition.
- Like new, fair and new, salvage are in very few.





Fuel

- Most of the cars have fuel type of Gas , whereas there are very few Electric cars.

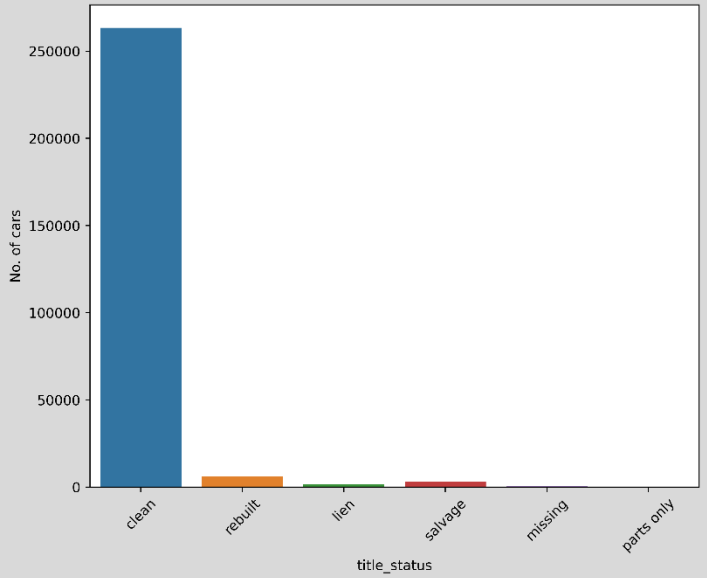


Transmission

- Most used cars are automatic and the least are those with manual transmission.

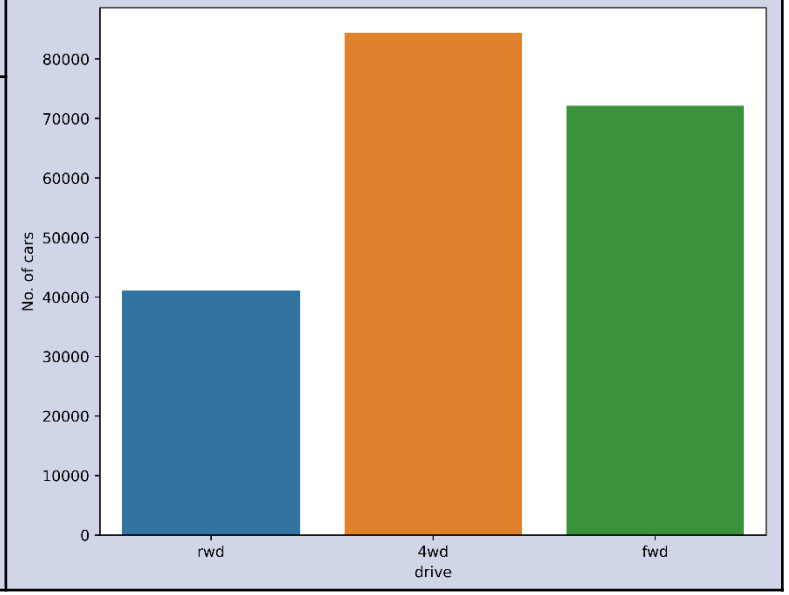
Title Status

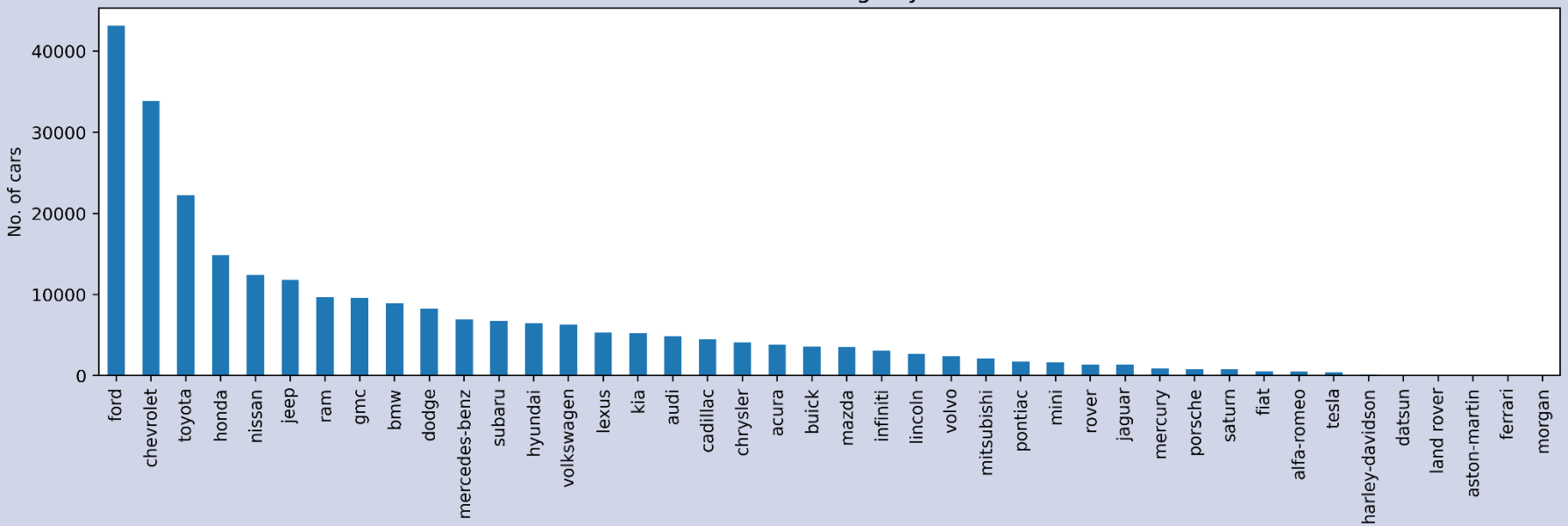
- 90% of the used cars are clean.
- And very few cars having rebuilt, salvage, lien, missing as their title status.



Drive

- 4 wheel drive is the most commonly found car and rear wheel drive is the least found.



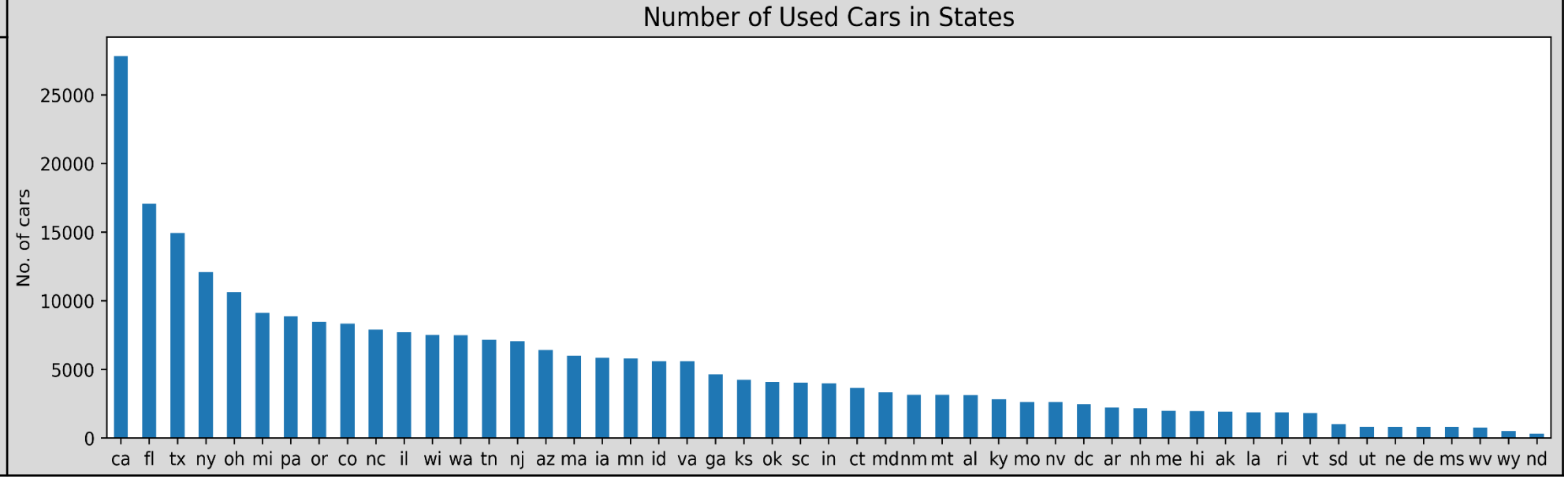


Manufacturer

- The top 3 most popular listed used car manufacturers are ford, chevrolet, and toyota while the least 3 are morgon, ferrari and aston-martin.

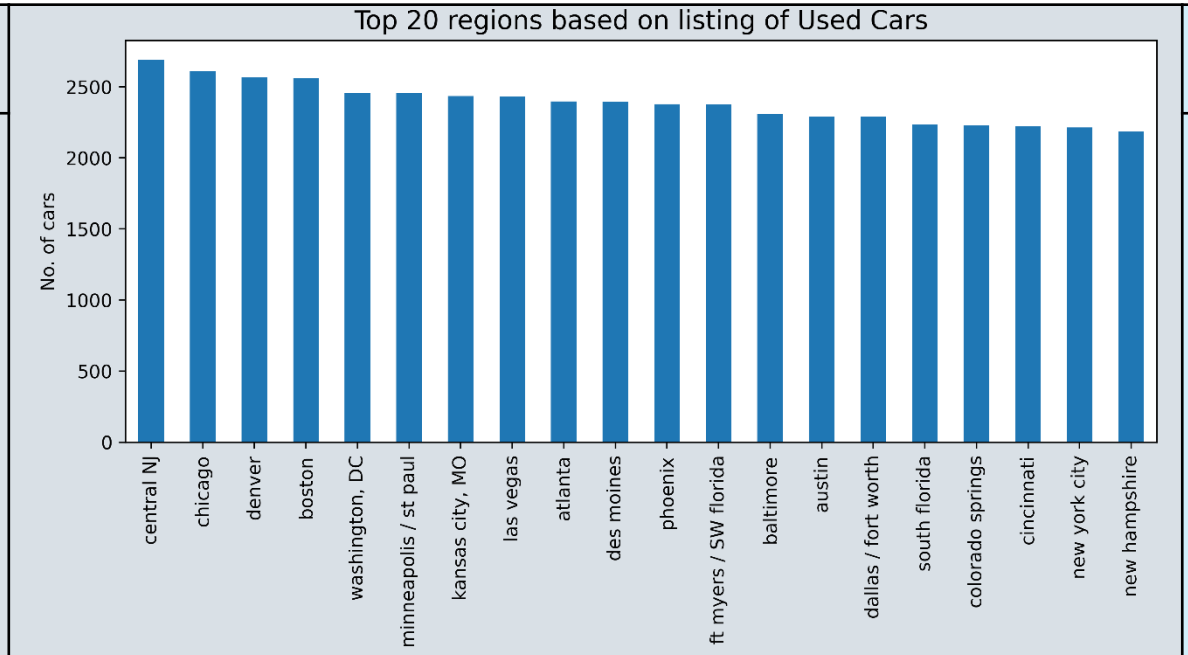
State

- In used car market most of the cars are from california.
- Top 5 states are California, Florida, Texas, New York, Ohio in used car market.



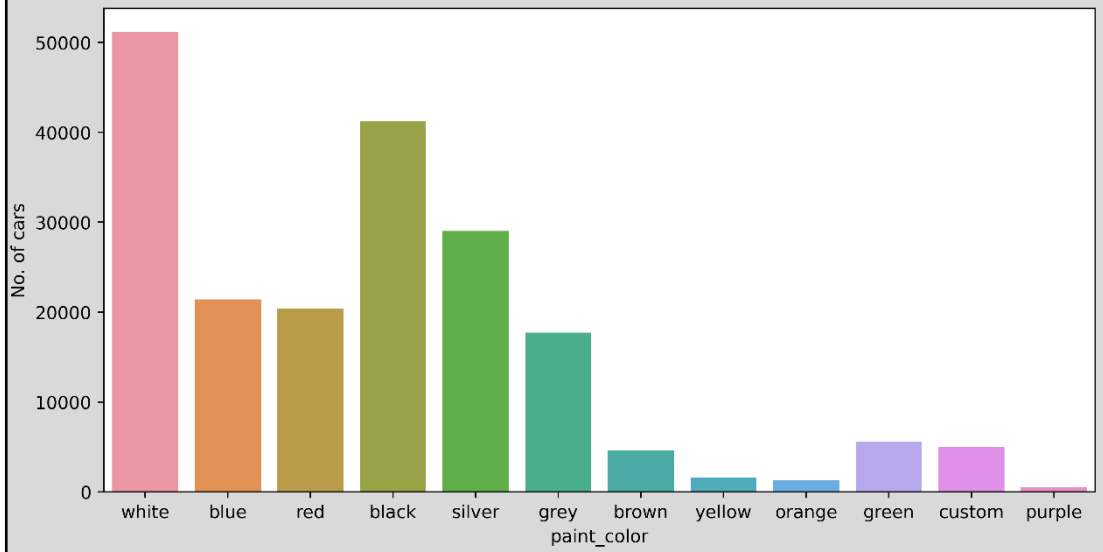
Region

- Central Nj have highest number of listed used cars.



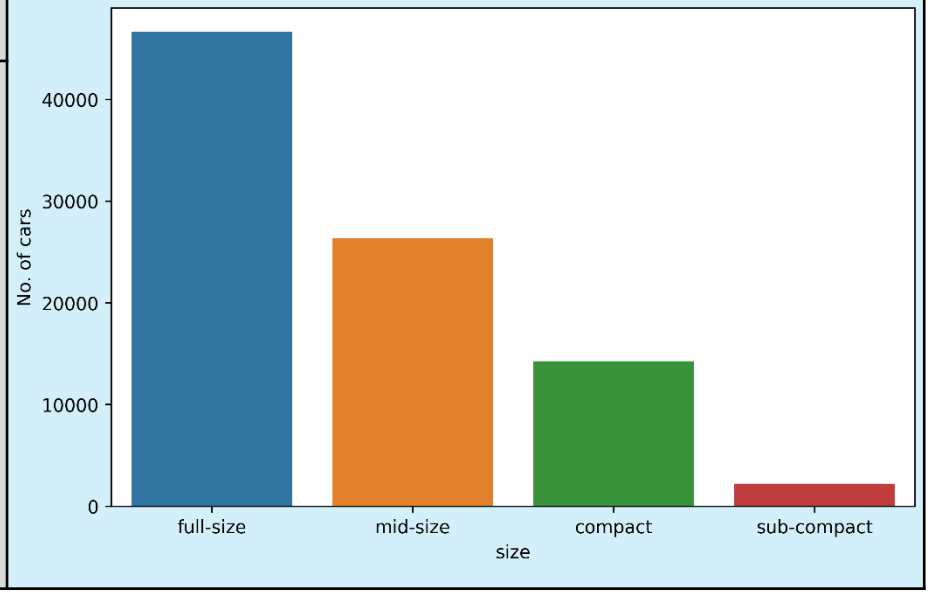
Size

- Full size cars have the most count followed by mid size and sub-compact size are found the least.



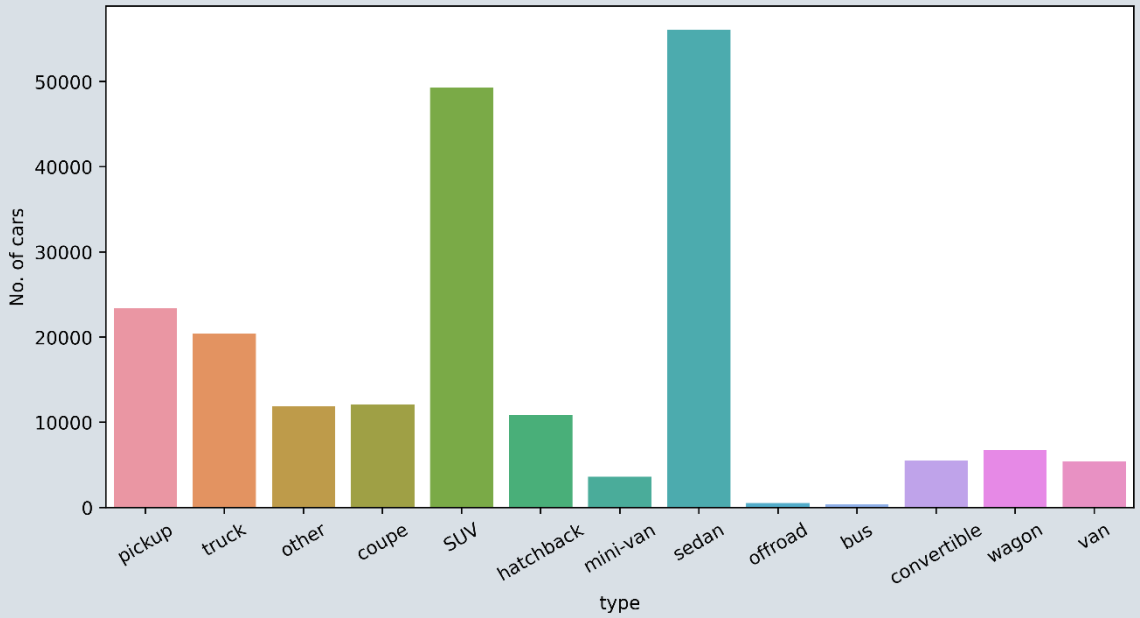
Paint Colour

- White and Black are most popular colours in used cars.
- Where Purple, Orange & yellow are least popular.



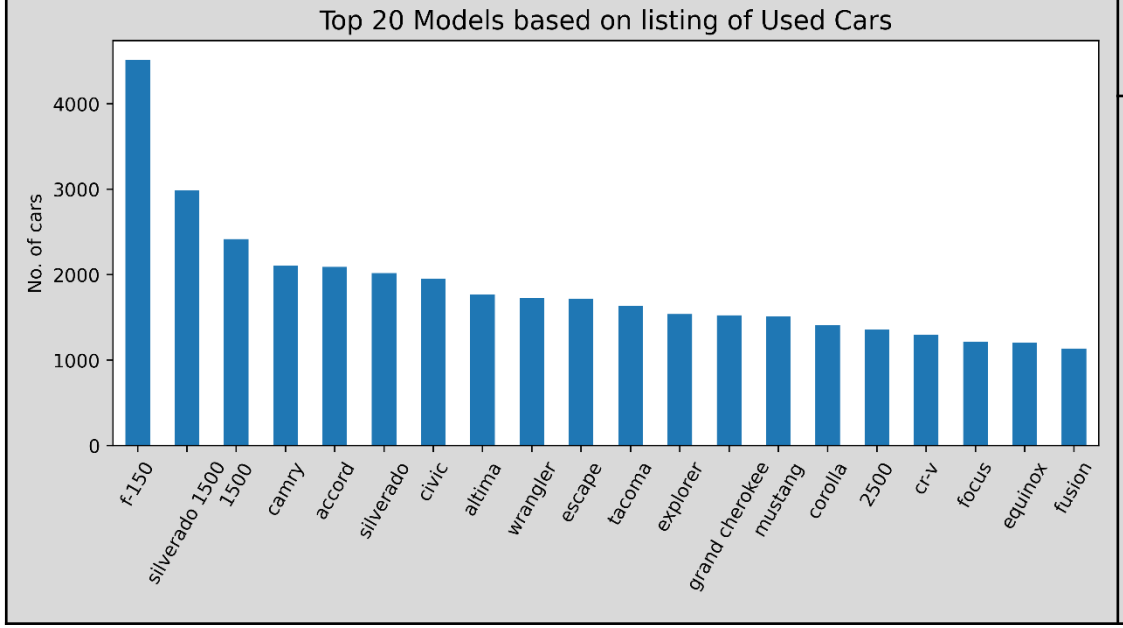
Type

- Sedan and SUVs are the most found cars in Used Car market.
- Where offroad and bus are the least in used car market.



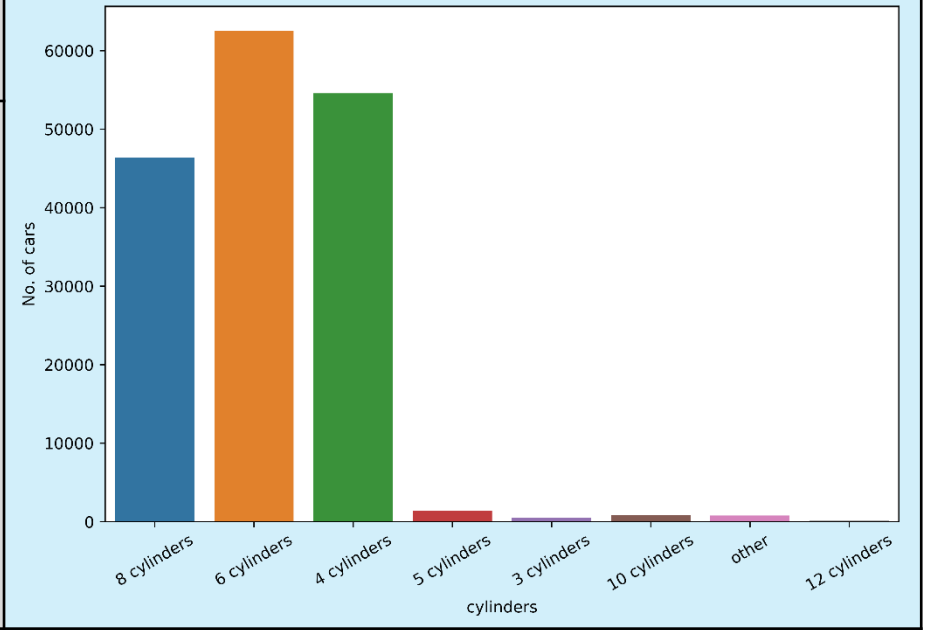
Cylinders

- Cars having 6, 4 and 8 cylinders occupy the most used car market shares.
- Cars having 3, 5, 10, 12 are very less car market shares.

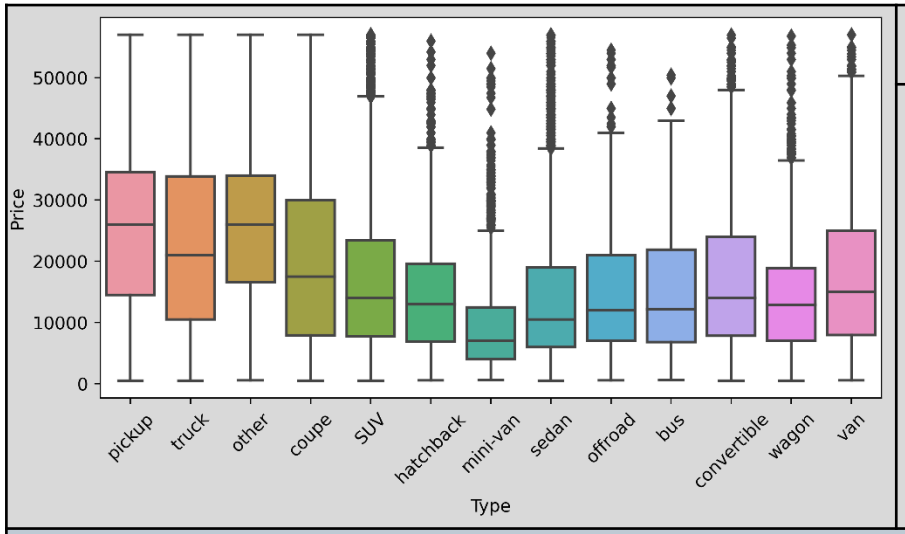


Models

- F150 car model is the most found model in used car market and it belongs to Ford.



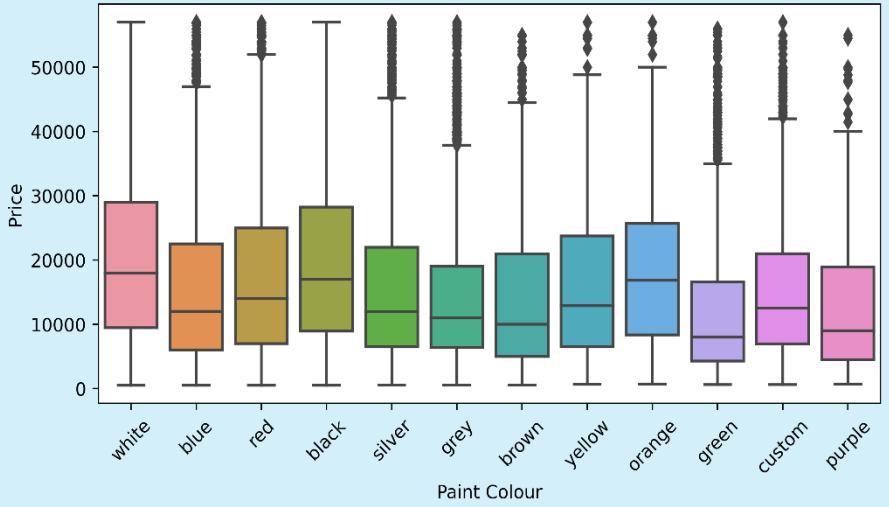
Bivariate Analysis - With Respect to Target (Price)



Type vs Price

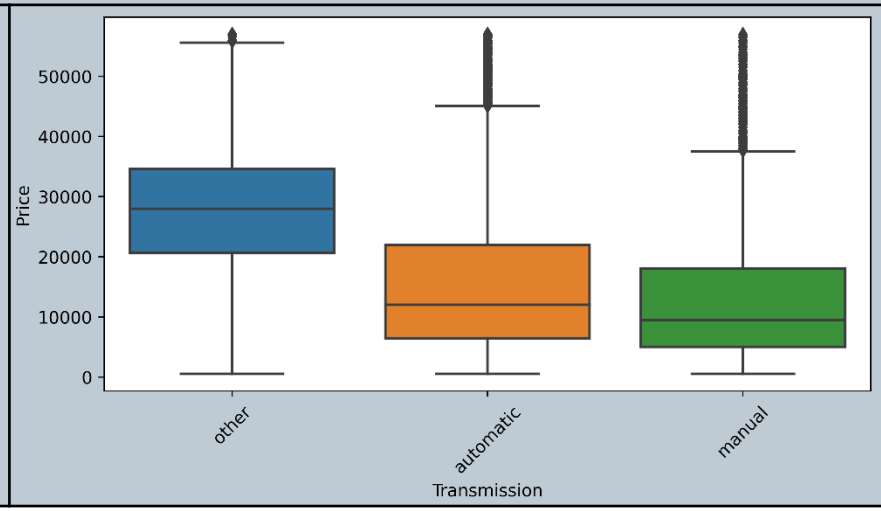
- Cars type pickup and other have the highest average price.
- Mini-van has the least average price.

Paint Colour vs Price

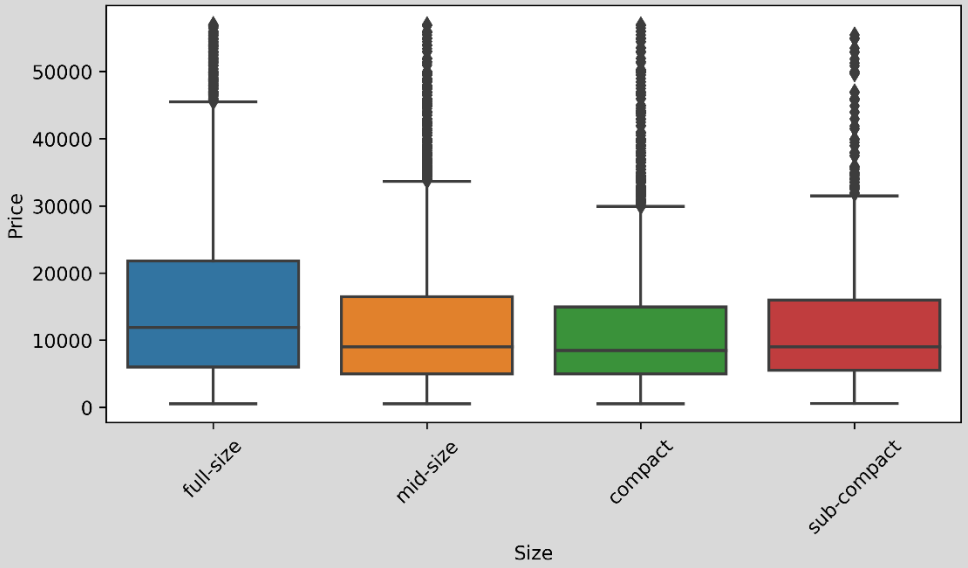


Transmission vs Price

- Automatic has higher average price than manual cars.
- In transmission category 'other' has the highest average price.



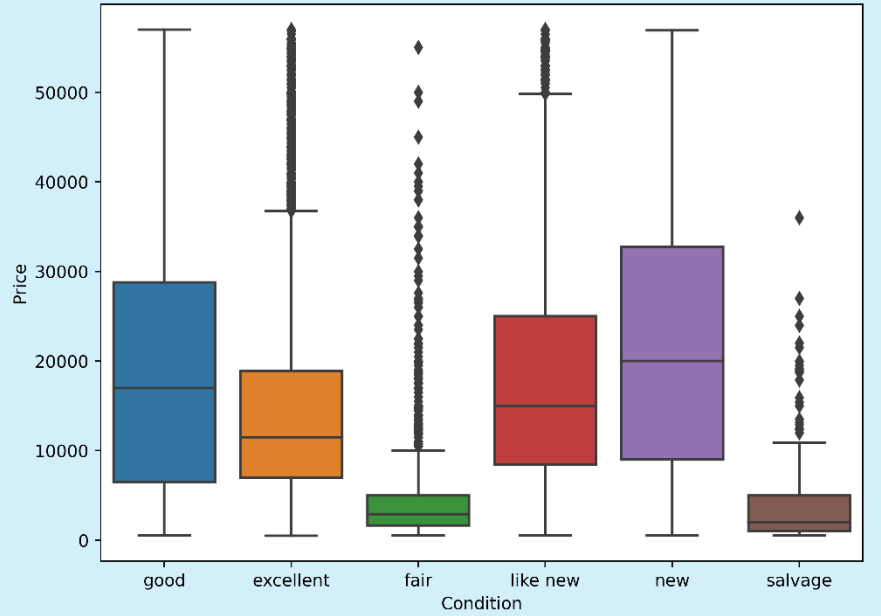
- Cars with 'paint_color' white and black have the highest average price.
- Cars with paint-color green has the least average price.



Size vs Price

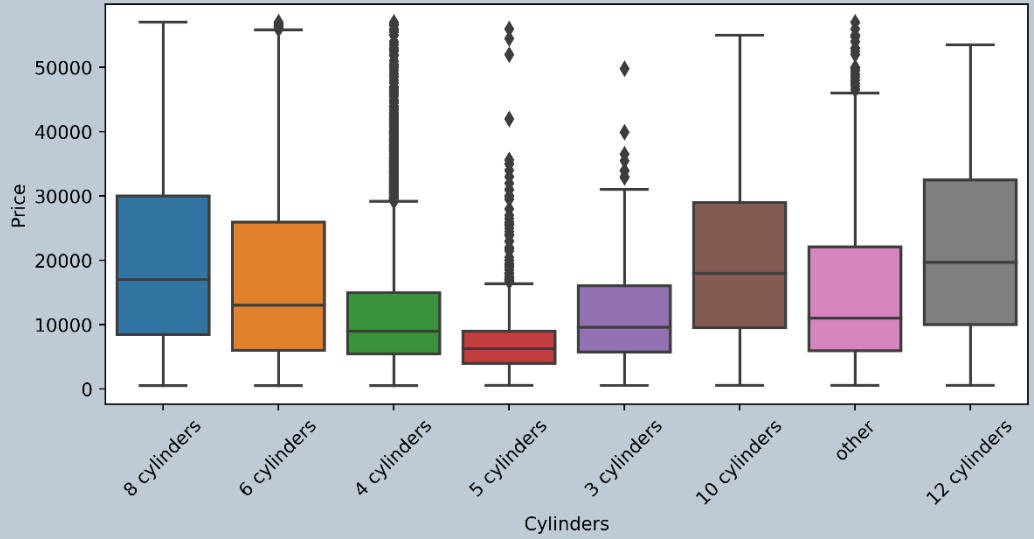
➢ All the categories have approximately equal average price.

Condition vs Price

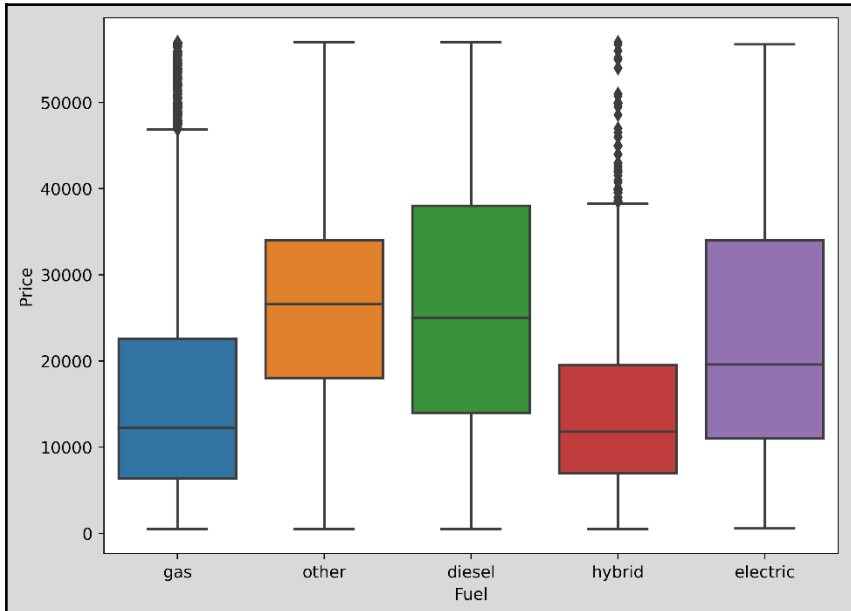


Cylinders vs Price

➢ Higher the number of cylinders higher the price except 5 cylinders.



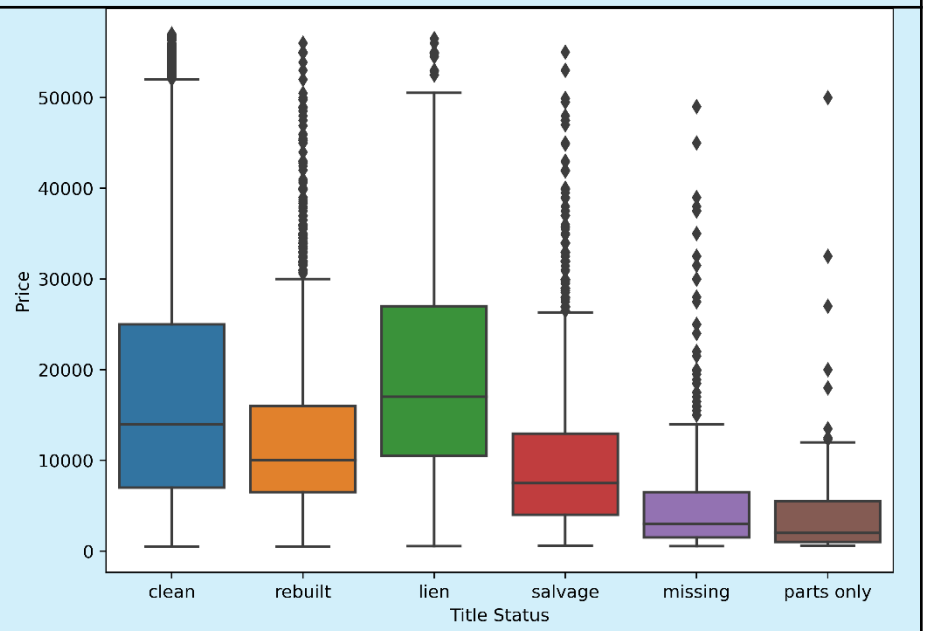
➢ Better the condition higher the price except fair condition.



Fuel vs Price

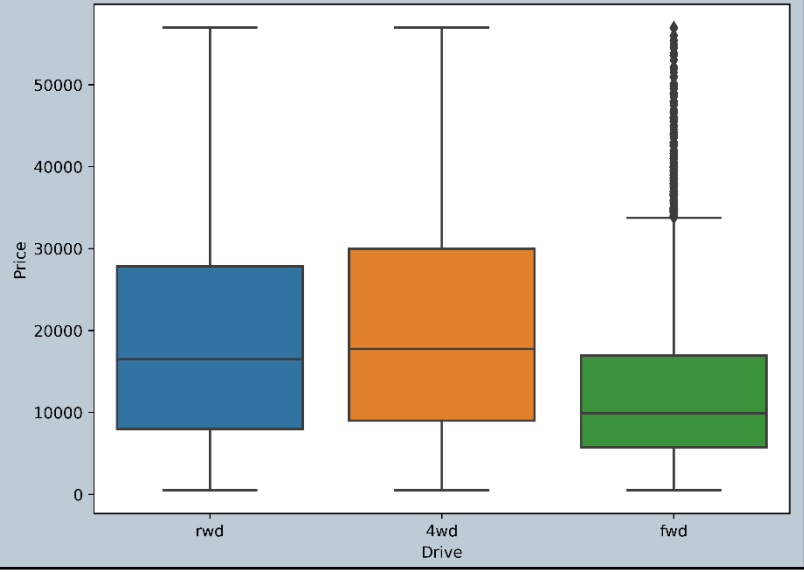
- Average cost of 'other' category fuels is the highest.
- The fuel types with Least Average price are gas and hybrid.

Title Status vs Price

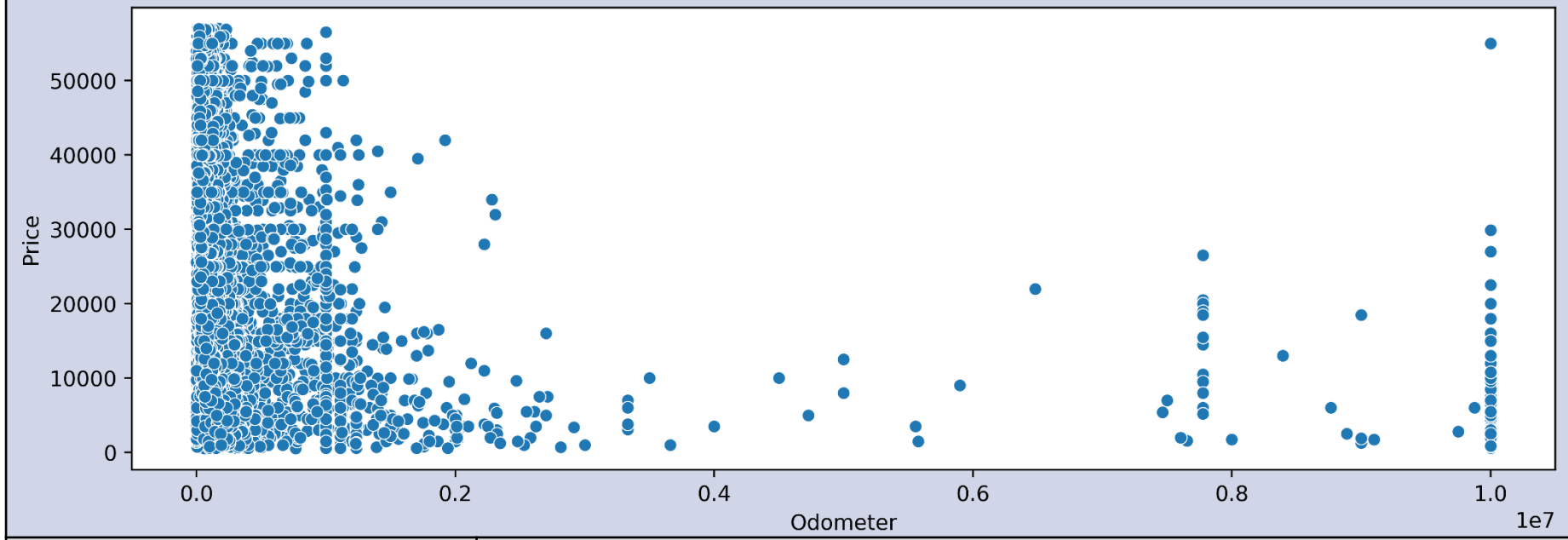


Drive vs Price

- 4wd (4 wheel drive) has the highest average price. And rwd approximately same as 4wd.



- Cars with title status 'lien' has the highest average price.
- Missing and parts only of cars are the least costly.

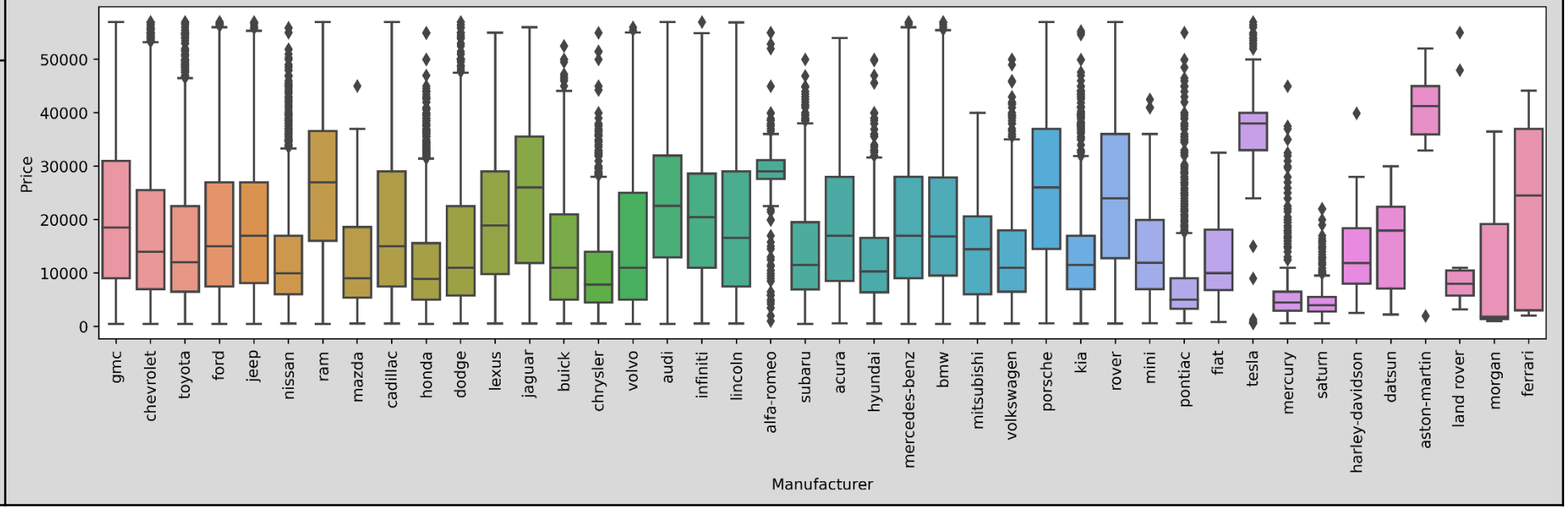


Odometer vs Price

- A lot of Odometer values seem to be concentrated around the value 0, which is very misleading.
- Cars with least distance travelled have higher prices.

Manufacturer vs Price

- Aston Martin, tesla, alfa-romeo manufacturer have the highest average price.
- Saturn, Mercury and pontiac manufacturer have the least average price.



Missing Values

	Total	Percentage of Missing Values
county	426880	100.000000
size	306361	71.767476
cylinders	177678	41.622470
condition	174104	40.785232
VIN	161042	37.725356
drive	130567	30.586347
paint_color	130203	30.501078
type	92858	21.752717
manufacturer	17646	4.133714
title_status	8242	1.930753
lat	6549	1.534155
long	6549	1.534155
model	5277	1.236179

	Total	Percentage of Missing Values
odometer	4400	1.030735
fuel	3013	0.705819
transmission	2556	0.598763
year	1205	0.282281
image_url	68	0.015930
posting_date	68	0.015930
url	0	0.000000
price	0	0.000000
state	0	0.000000
region_url	0	0.000000
region	0	0.000000
id	0	0.000000

MISSING VALUES TREATMENT

- **Year -** The values were filled using →
  1) URL
  2) VIN (using vin function in python)
  3) Remaining values imputed with mode year values.
- **Manufacturer -** Filled using→
  1) VIN
  2) Model
  3) URL
  4) Remaining Values Binned as Unknown
- **Condition -** Filled using→
  1) Title Status
  2) Rest using python Forward Fill function.
- **Odometer** – Filled using→
  1) All values greater than 999999 imputed as 999999.
  2) Null and 0 values filled with mean values basis the model .
  3) Remaining values filled using median.

- Imputed mode values on basis Model & Manufacturer are: **Transmission, Type, Fuel**.
- Imputed mode values on the basis Model, Manufacturer & Type are: **Drive, Size, Paint colour**.
- **Cylinders** - The values were filled using →
  1) Electric cars with other.
  2) Description.
  3) Model.
  4) Manufacturer.

Statistical Tests (Predictor vs Target)

Significant Variables			Non significant variables		
Features	**Test**	**P value**	**Features**	**Test**	**P value**
Year	Mann-Whitney U	0.0	Region	Anova	1.0
Odometer	Mann-Whitney U	0.0	Manufacturer	Anova	1.0
Condition	Anova	0.000001	Cylinder	Anova	0.390706
State	Anova	0.032514	Fuel	Anova	0.964539
Is tax	T-test independent	0.009575	Transmission	Anova	0.270099
Age	T-test independent	0.0005157	Drive	Anova	0.057264
Vint_car	Anova	0.000009	Size	Anova	0.208265
Model	Due to high categories we can't run statistical test and from domain knowledge we keep this variable for modelling.		Type	Anova	0.388538
			Paint colour	Anova	0.294331
			State Zone	Anova	0.588757

Correlation (Predictor vs Target)

Feature	Correlation
Region	-0.072851
Price	1.000000
Year	0.344641
Manufacturer	-0.078222
Model	-0.004239
Condition	0.028708
Cylinders	0.244160
Odometer	-0.412041
Size	0.184555
Type	0.035677

Feature	Correlation
State	-0.000448
State_Zone	0.068220
Age	-0.344641
is_tax	0.001162
fuel_diesel	0.184503
fuel_electric	0.027906
fuel_gas	-0.252713
fuel_hybrid	-0.021164
fuel_other	0.184598
transmission_automatic	-0.227737

Feature	Correlation
transmission_manual	-0.080006
transmission_other	0.333662
drive_4wd	0.212183
drive_fwd	-0.287314
drive_rwd	0.083461
vint_car_antique	-0.002434
vint_car_classic	-0.207867
vint_car_modern	0.186869
vint_car_vintage	-0.000414

DATA MODELLING

- **BASE MODEL: Decision Tree Regressor**

- Let's see final comparison for the performance of all the models that we have tried to make the final conclusion through Data Frame:

	Model Name	Train R2 Score	Test R2 Score	Train RMSE	Test RMSE	MAPE
0	Decision Tree	0.999604	0.696382	247.414093	6816.527200	40.538615
1	Ridge	0.551710	0.550717	8323.137291	8291.999291	65.752941
2	Lasso	0.551695	0.550725	8323.279008	8291.927331	65.720329
3	Elastic Net	0.420750	0.422053	9461.076248	9404.668752	62.097979
4	Random Forest	0.978557	0.851884	1820.349500	4761.028578	43.168481
5	AdaBoost	0.635130	0.635568	7508.906508	7468.054537	60.193155
6	Gradient Boost	0.781405	0.783581	5812.027572	5755.016932	50.773069
7	XGBoost	0.856421	0.837678	4710.345448	4984.118219	42.128192
8	CatBoost	0.851534	0.841018	4789.844196	4932.564387	42.989694
9	Tuned CatBoost	0.882055	0.851753	4269.212741	4763.129450	41.563495

Final Insights

- ✓ From this table, we can observe that CatBoost model is our best fit model.
- ✓ After parameter tuning of CatBoost our Tuned **CatBoost** Model is our Final Model.
- ✓ Tuned CatBoost has given best R-Squared i.e. 85.17%.

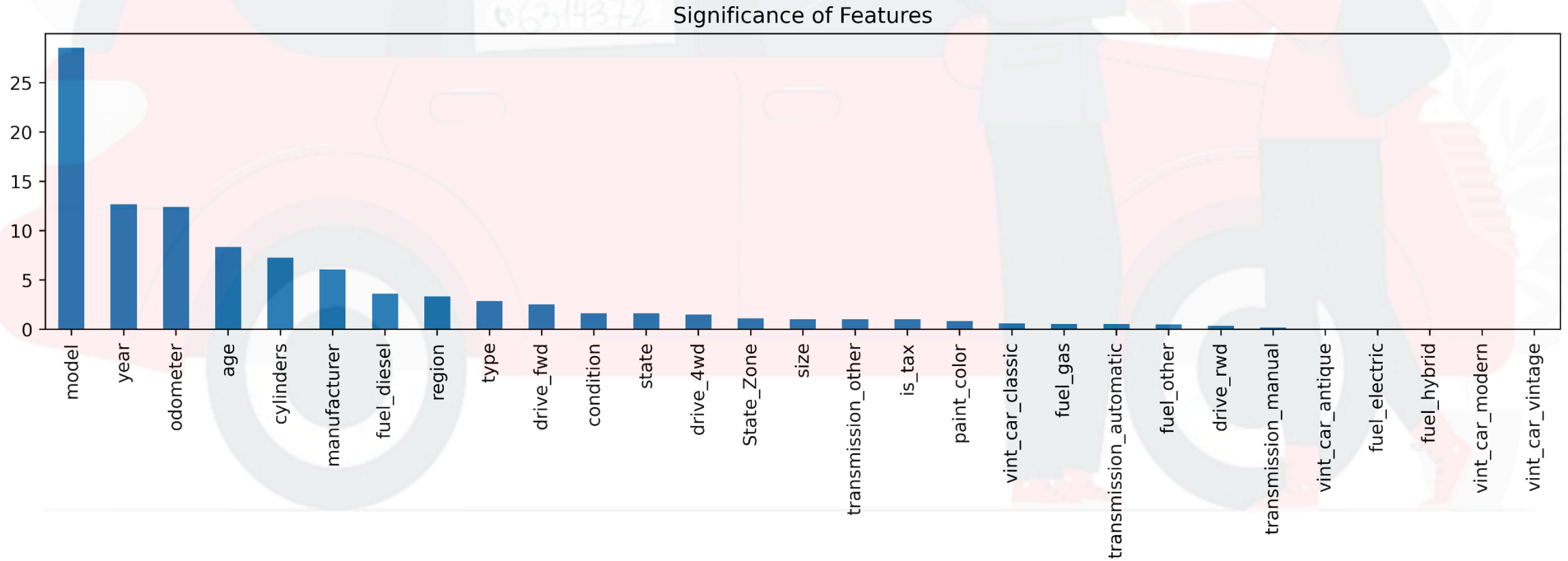
Regression

Feature Importance

Bar plot to check the importance of the features.

1st most significant feature : Model

2nd most significant feature : Year



Conclusion

1. Used cars with high Mileage are cheaper.
2. Used cars with better Appearance are expensive.
3. Used cars which come from big manufacturer are cost more.
4. Used cars with pickup or truck or coupe or convertible type are cost more.
5. Used cars with white or black paint-color are expensive.
6. Used cars with more cylinders are expensive.
7. Used cars with higher vehicle age are cost less.

Thank You!