Data Set Description – For this lab, I have chosen the Vehicle dataset from Kaggle, which contains information about new and old used cars. This dataset can be used for a variety of purposes, but in our coursework, we will focus on using it to solve a regression problem - predicting the selling price of a car.

The dataset includes the following features with corresponding descriptions:

- Name: categorical feature that indicates the make and model of the car (e.g., Maruti Suzuki, Hyundai Verna)
- Year: categorical feature that represents the year when the car was last purchased
- Selling_price: continuous feature that represents the price at which the car is being offered for sale; this is the predictor variable we will be using
- Km_driven: continuous feature that represents the number of kilometers the car has been driven
- Fuel: categorical feature that indicates the type of fuel the car uses (e.g., petrol, diesel, CNG, LPG, electric)
- Seller_type: categorical feature that indicates whether the seller is an individual or a dealership
- Transmission: categorical feature that indicates whether the car has automatic or manual gear transmission
- Owner: categorical feature that indicates whether the owner is the first, second, or subsequent owner
- Mileage: feature that represents the mileage of the car
- Engine: feature that indicates the engine type of the car
- Max_power: feature that represents the maximum power output of the car
- Torque: feature that represents the torque output of the car
- Seats: categorical feature that indicates the number of seats in the car (e.g., 4, 5, etc.)

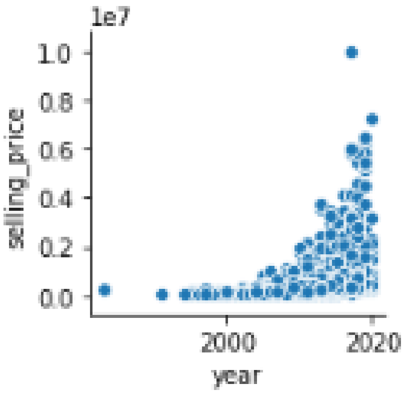
Initial plan for data exploration – Since we have around 10 categorical features, we will need to explore these variables to determine how many categories they have and try to find if there are any important information in them that we can try and use for predicting the selling price of the cars.

Along with that the two continuous variables that we have which are Selling Price and Kilometers driven, I would want to look at their underlying distributions as well.

Also we would want to check for Null values and if there are any imputations required for the same.

Let us first look scatter plot for all the variables against selling Price

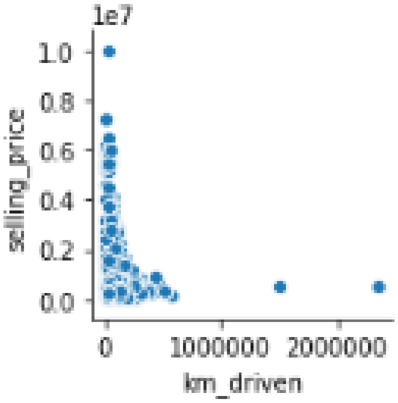
Year:



For year it seems like these are the years from when the cars are from, and also newer cars seems to have more value than older cars which makes sense.

Take away: Rather than keeping all the year information of when the cars are from we can create buckets or categories to keep the year information and reduce the number of categorical variables. We will address this in the feature engineering section

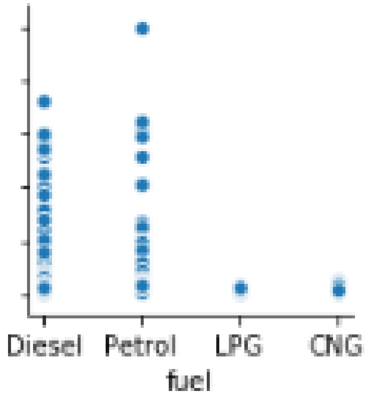
Km_driven:



Km_driven is a continuous variable and prices are higher for less driven vehicles, which makes sense. There seems to be two outliers in the km_driven variable

Take away: we will be treating the outliers as well as looking at the underlying distribution of the cars as well

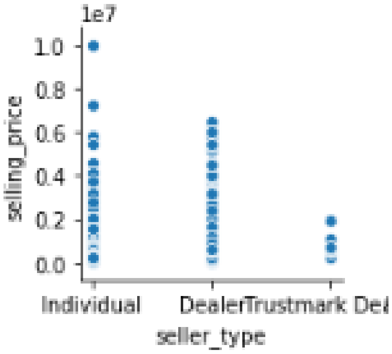
Fuel:



Fuel is again a categorical variable, where the categories are Diesel, Petrol, LPG and CNG. These are all the different type of fuel powered cars that are available in the data set. Diesel and petrol powered cars are costlier compared to Gas powered cars

Take Away: We can combine LPG and CNG categories together as Gas powered cars; Diesel and Petrol as oil powered cars.

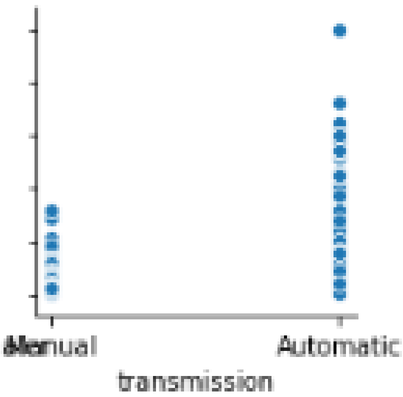
Seller_type:



There are 3 different types of sellers, Individual, Dealer, and Trustmark Dealers;

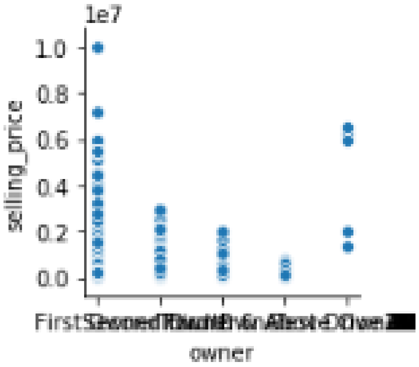
Take Away: We can merge trust mark dealers with dealers and trust mark dealers are just a type of dealers who have some sort of certification

Transmission:



Automatic and Manual are the two kinds of transmission, Automatic car selling price is more than Manual as expected.

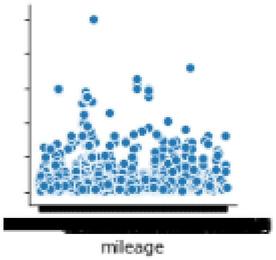
<u>Owner</u>:

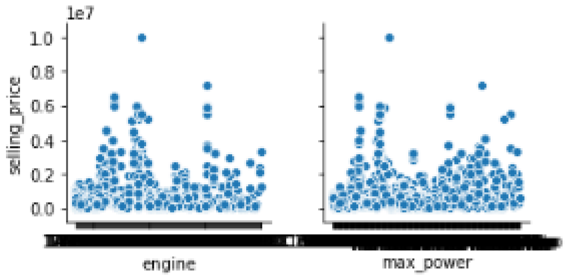


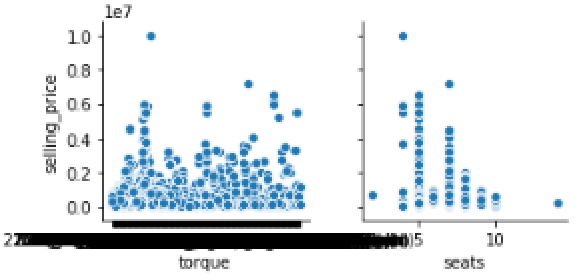
There are multiple categories in owner, First, second, third etc times owner.

Take Away: We can look into it if we want to reduce the number of categories here

Mileage/Engine/Max Power/Torque:



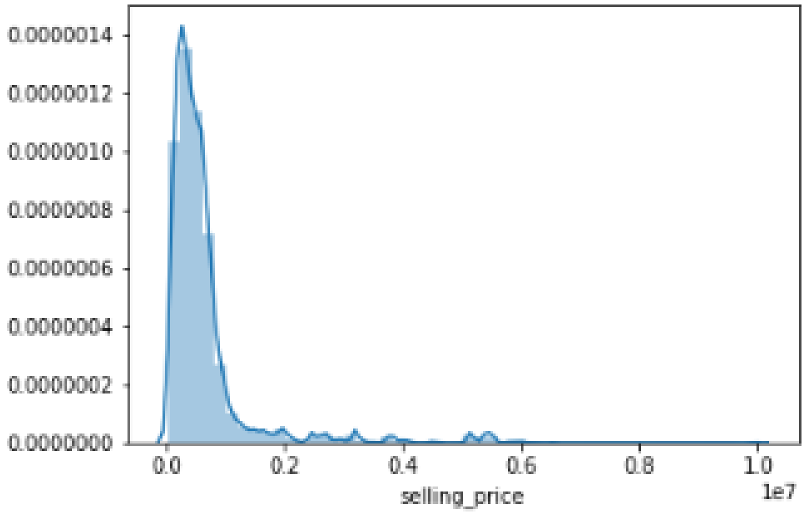




All these variables have more than 10 categories, we can look into them in the feature engineering stage if there is a way to reduce the number of variables in them

Now let's look at the distribution for some of the continuous variables

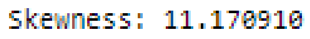
<u>Selling_Price</u>:



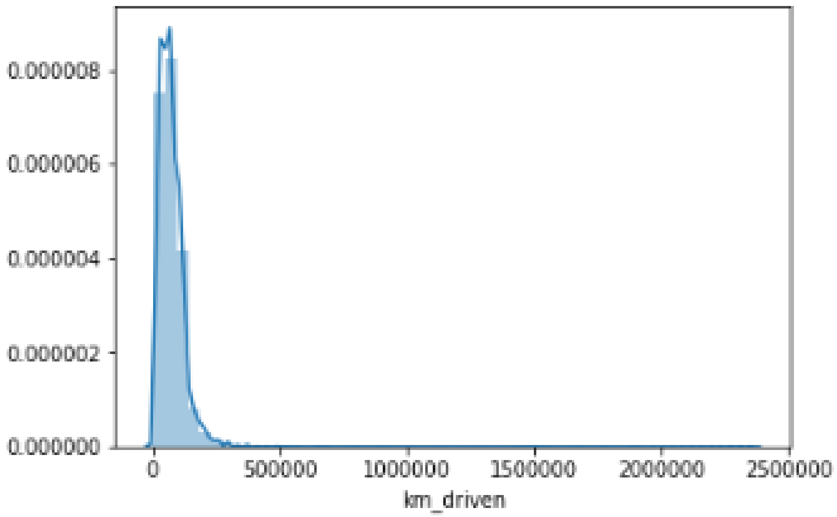
As it can be observed from the diagram above that the underlying distribution for selling price is a normal distribution and highly positively skewed, with the skewness being 4.19

Take Away: We would want to transform the variable into less skewed normal distribution we can try with log transformation as it is the most widely used transformation technique

Km_driven:



```
Skewness: 11.170910
```



It can be observed from the plot above that underlying distribution for KM_driven is an even more highly positive skewed normal distribution with a skewness parameter of 11.17.

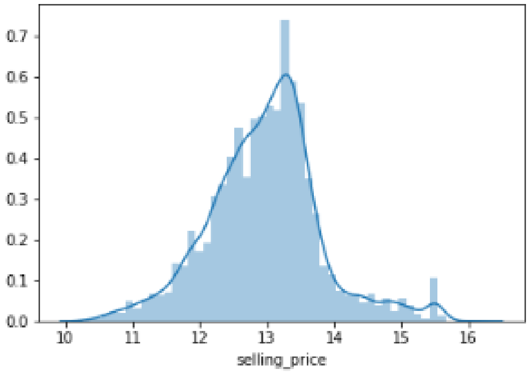
Take Away: We would again want to perform some transformation on this distribution to see if we can improve the distribution to make it more symmetric

Data Cleaning and Feature Engineering –

Let's first start with dealing with the transformation of the continuous variables

Selling_Price:

Log transformation is one of the most widely used transformation techniques on skewed normal distributions let's try with that.



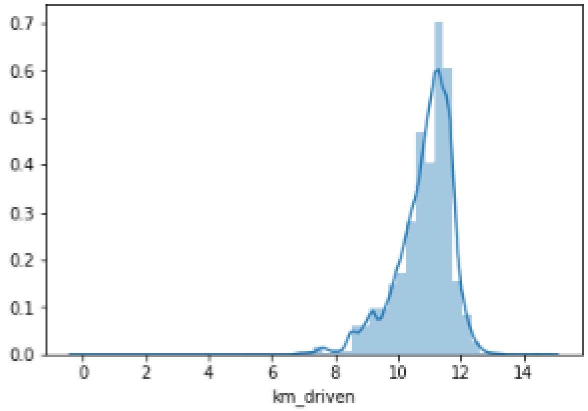
With log transformation on the selling price we are able to get a more symmetric normal distribution, with a skewness of 0.22 which lies in the range of fairly symmetric bell curve -0.5 to 0.5

<u>Km_Driven:</u>

Again for KM_Driven let’s start with log transformation, if it does not work we will check with sqrt and box cox transformation as well

Log transformation:

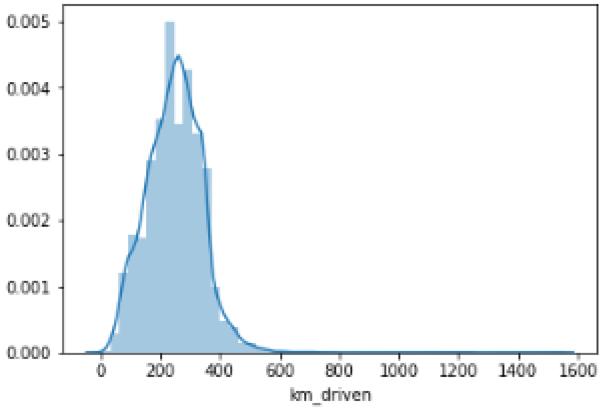
```
Skewness: -1.285045
```



Log transformation reduced the skewness but not its more negatively skewed normal distribution

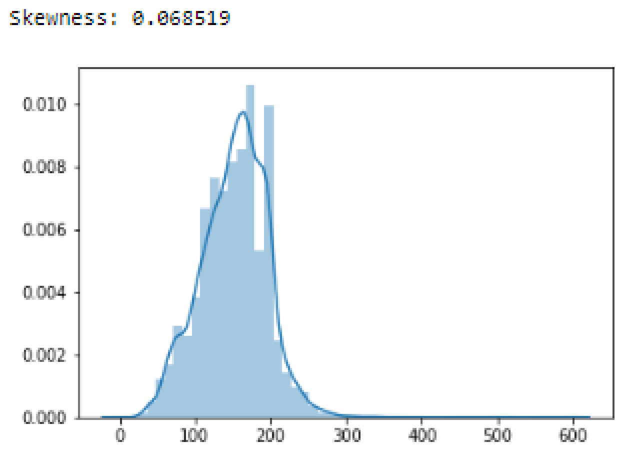
Square Root transformation:

```
Skewness: 0.662873
```



Sqrt transformation does a better job than log transformation but the data is still a little bit skewed

Box Cox Transformation:

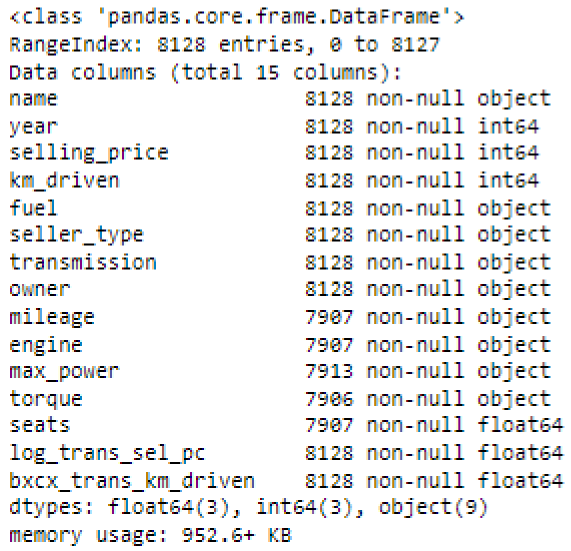


```
Skewness: 0.068519
```

Box cox transformation was able to reduce the skewness to 0.06 which is within the range of 0.5 to -0.5

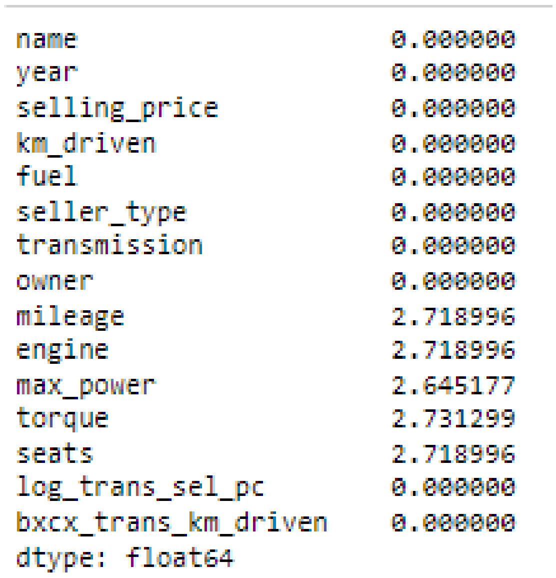
Handling Missing Values:

From the Data frame info, we were able to notice there are some null values within the columns mileage, engine, max_power, torque, seats



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8128 entries, 0 to 8127
Data columns (total 15 columns):
name                    8128 non-null object
year                    8128 non-null int64
selling_price           8128 non-null int64
km_driven               8128 non-null int64
fuel                    8128 non-null object
seller_type             8128 non-null object
transmission            8128 non-null object
owner                   8128 non-null object
mileage                 7907 non-null object
engine                  7907 non-null object
max_power               7913 non-null object
torque                  7906 non-null object
seats                   7907 non-null float64
log_trans_sel_pc        8128 non-null float64
bxcx_trans_km_driven    8128 non-null float64
dtypes: float64(3), int64(3), object(9)
memory usage: 952.6+ KB
```

Let’s look at the percentage of null in these variables



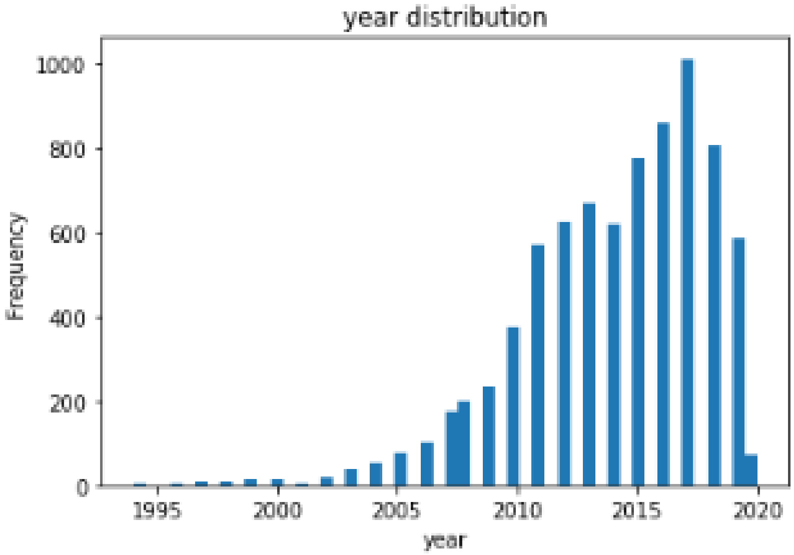
```
name                    0.000000
year                    0.000000
selling_price           0.000000
km_driven               0.000000
fuel                    0.000000
seller_type             0.000000
transmission            0.000000
owner                   0.000000
mileage                 2.718996
engine                  2.718996
max_power               2.645177
torque                  2.731299
seats                   2.718996
log_trans_sel_pc        0.000000
bxcx_trans_km_driven    0.000000
dtype: float64
```

We can see from the above metrics that the percentage of null is less than 3 percent in these variables, then we can very easily get rid of these rows from the data frame.

Now let’s look into feature engineering for some of these variables:

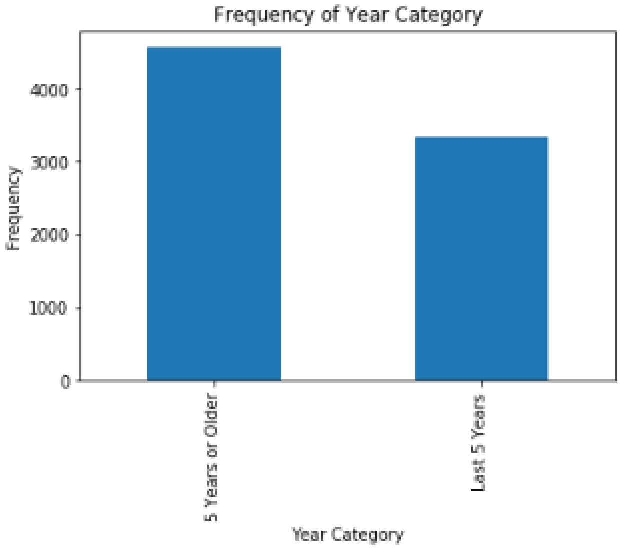
Year:

As already observed that in our dataset we have cars ranging from 1995 till 2020; and since this is a categorical variable these will create a lot of columns for the data when trying to fit into a model

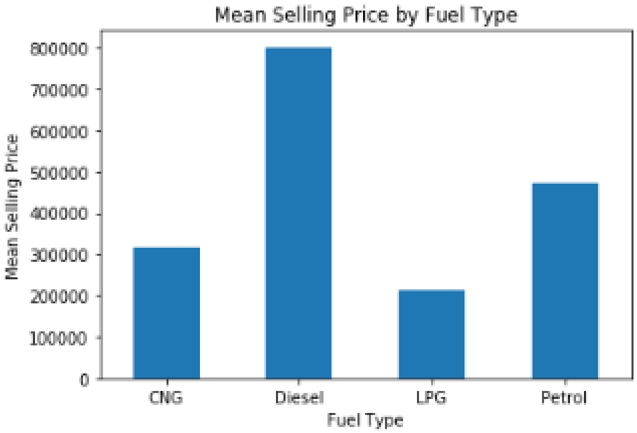


We can create two buckets for the data, we can create it to last 5 years and 5 years and above. So 2015 to 2020 will become last 5 years and any of the cars older than 2015 will become 5 years and older

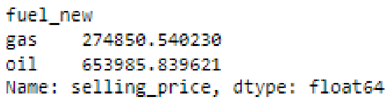
After transformation this is how it looks like



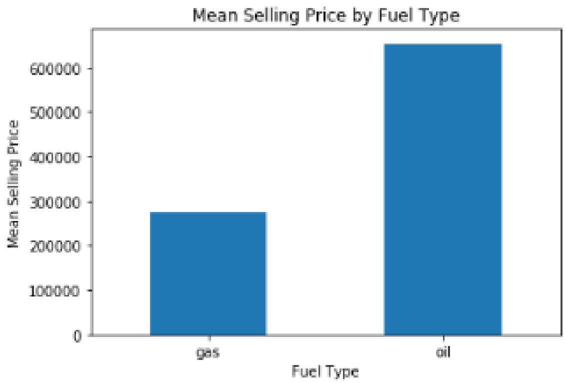
Fuel:



As we have seen before there are 4 type of fuel and we can convert it to two types Gas and Oil cars, above plot shows that the mean price of Diesel cars is the most followed by petrol and the gas cars

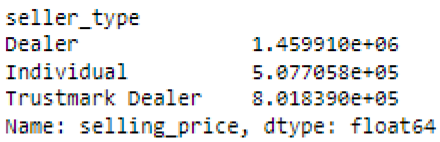


```
fuel_new
gas    274850.540230
oil    653985.839621
Name: selling_price, dtype: float64
```

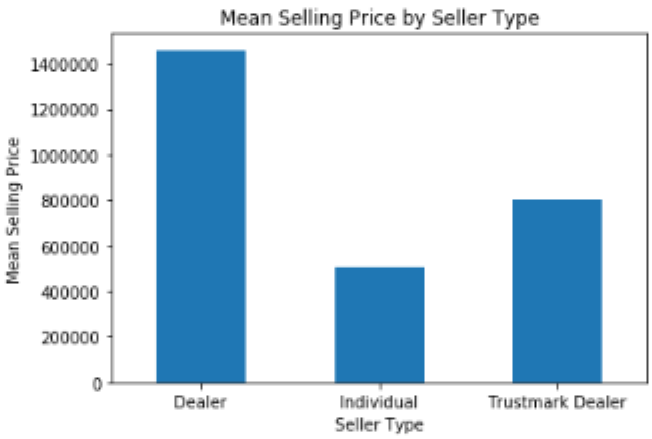


After transformation the overall story still remain the same, we have just reduced the number of categories

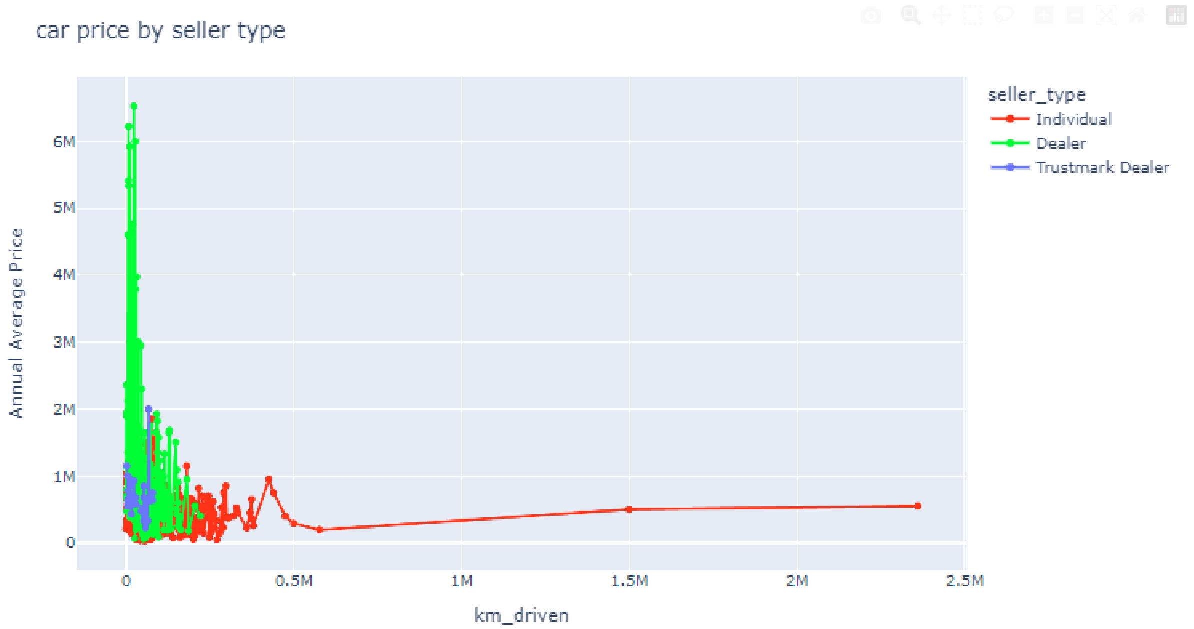
Seller Type: As observed before we were having three types of sellers and I would want to convert Trustmark Dealer to Dealer as well



```
seller_type
Dealer              1.459910e+06
Individual          5.077058e+05
Trustmark Dealer    8.018390e+05
Name: selling_price, dtype: float64
```

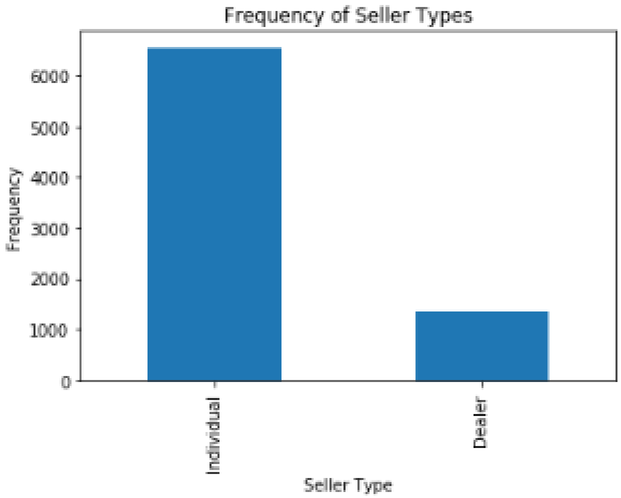


Overall it can be observed that the price of cars sold by dealership is higher than individual sellers, this can be due to the fact that dealerships tend to be able to attract more customers and tend to sell cars that are newer.



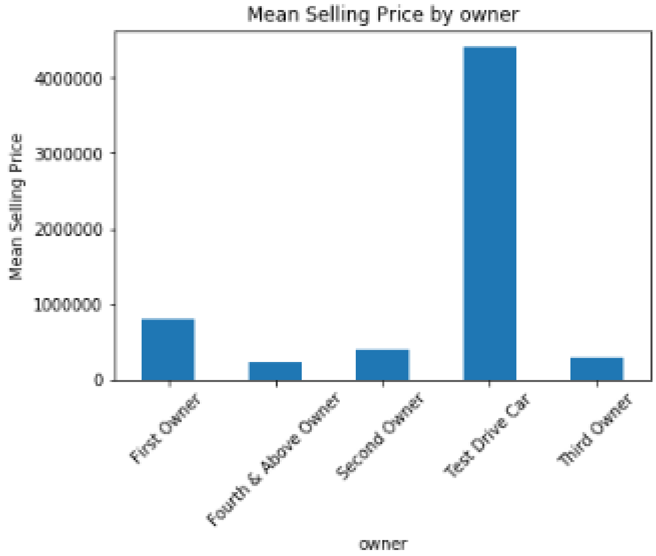
As we can observe that price for cars where the amount of km driven are nearly the same tend to have more price when sold by dealership compared to when they are sold by individuals. Also there are more cars driven more than 500K km driven sold by individuals compared to Dealerships.

After transformation



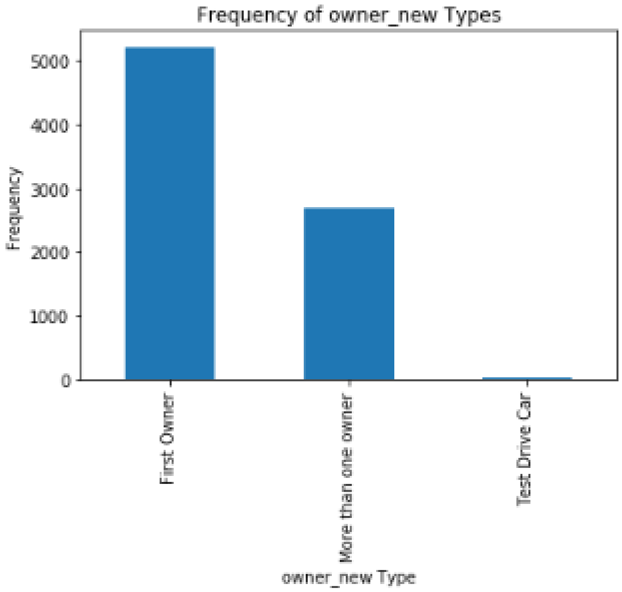
Owner:

As we had observed before, that there are first hand owners followed by second, third and Fourth and above as well. Below is the mean selling price based on owner types



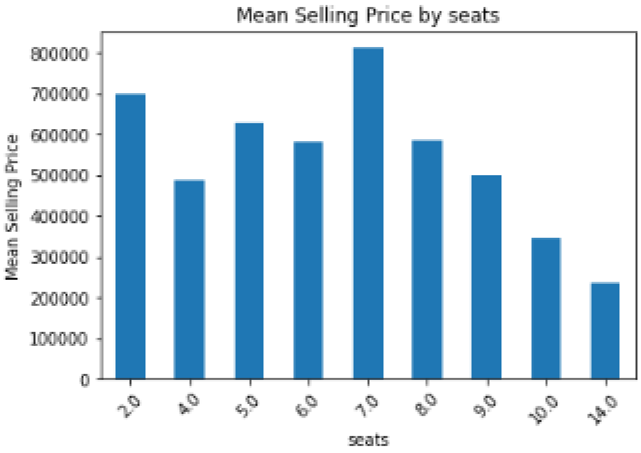
From the above plot we can see that there are new cars or test drive cars that are having the most prices followed by the used cars. We know that used cars will have less price compared to New cars. Rather than keeping so many categories, we can convert it to New, First Owner, More than One owner

After transformation



Seats:

We had observed that there are cars with multiple different kinds of seat options available. Below is the mean selling price distribution based on number of seats in the car



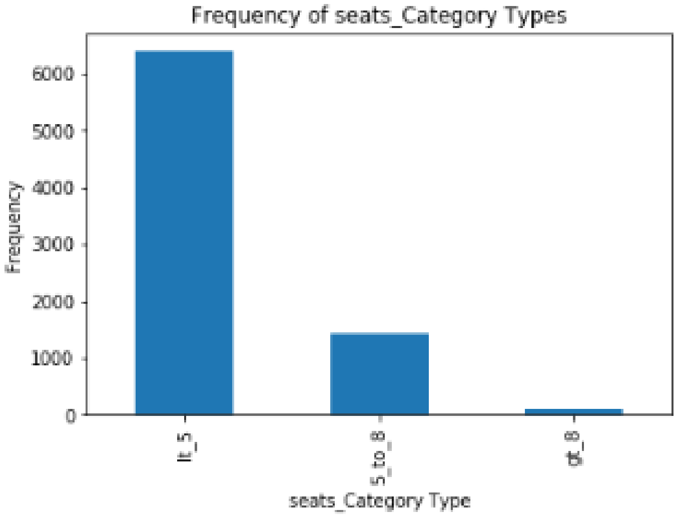
It can be observed that the selling price of 7 seater cars are the most followed by 2 seaters and 4 seaters

Also there are cars with 9, 10 and 14 seaters as well

Rather than keeping so many categories we can again bucket them

Since greater than 8 seats seem to be a different kind of car altogether we can keep them as a different bucket, normally most cars are 5 seaters so that can be a category on its own and remaining 5 to 8 seats can be a category as well

After transformation

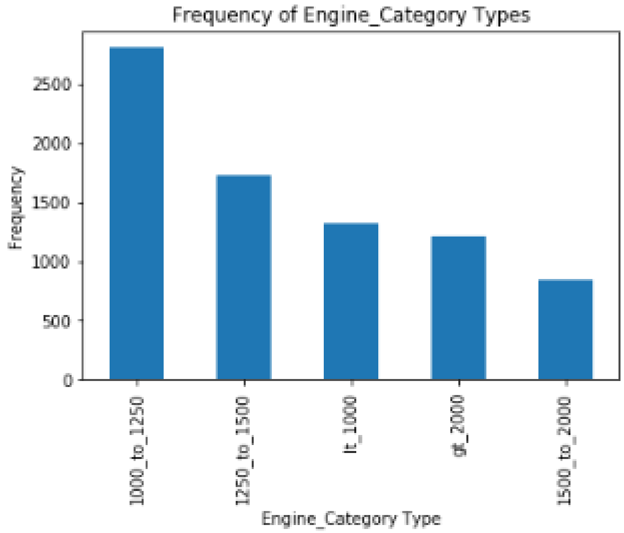


Engine:

Engine is one of those variables that seem to have a lot of categories, to be exact there are around 121 categories of engine

Based on our understanding about engines and different variation of engines we have bucketed engines into 5 categories; Less than 1000, 1000 to 1250, 1250 to 1500, 1500 to 2000 and greater than 2000 cc cars

After transformation

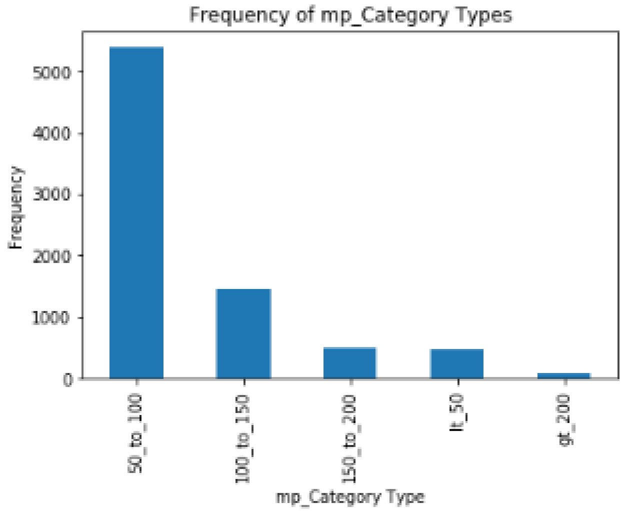


Max Power:

Similar to engine there are multiple categories of max power as well, we were able to bucket them into

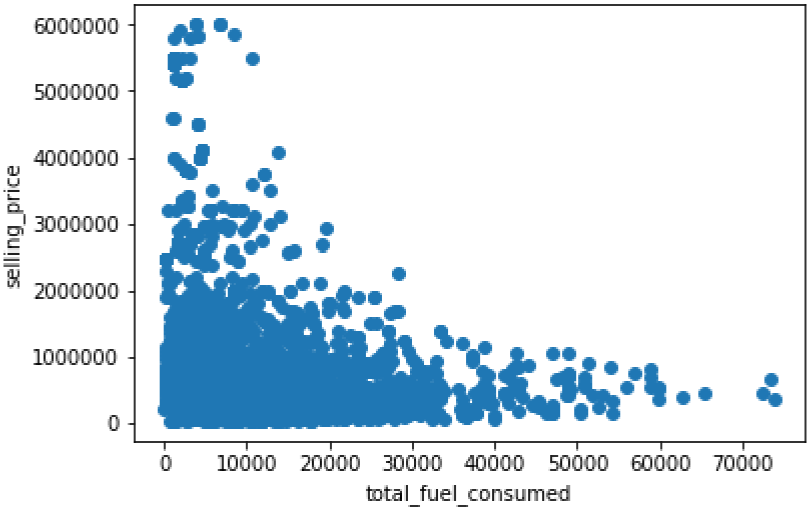
Less than 50, 50 to 100, 100 to 150, 150 to 200 and greater than 200

After transformation



Mileage:

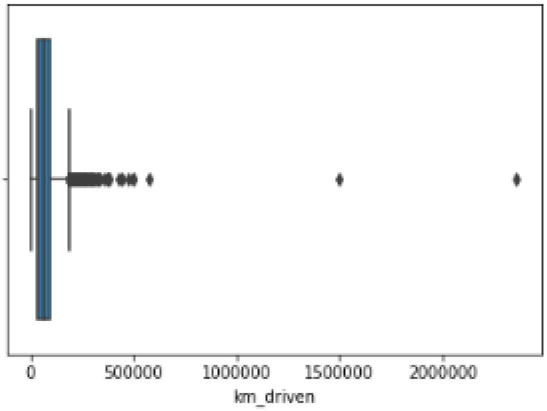
Mileage is another variable again with a lot of categories, but rather than bucketing them what we have done is using km_driven and mileage we have created a new variable called total fuel consumed
we have utilized km_driven/ mpg(converted to km per litre)

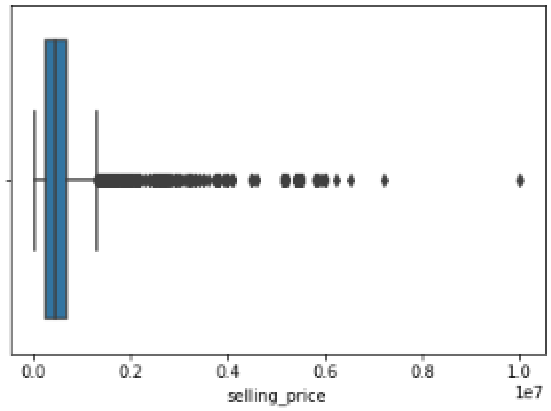


Outlier removal:

We have looked into km driven and selling price which are our two continuous variables to look for outliers

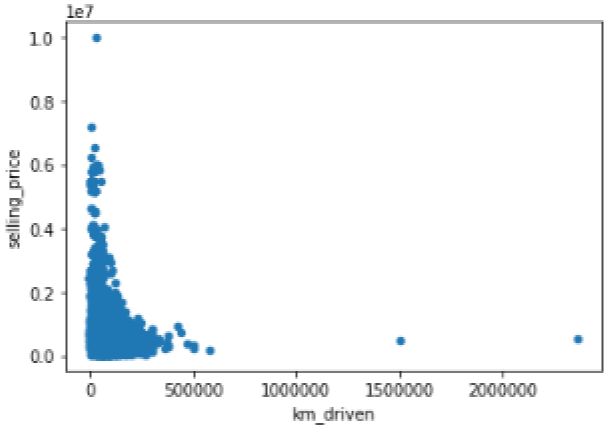
Box plot



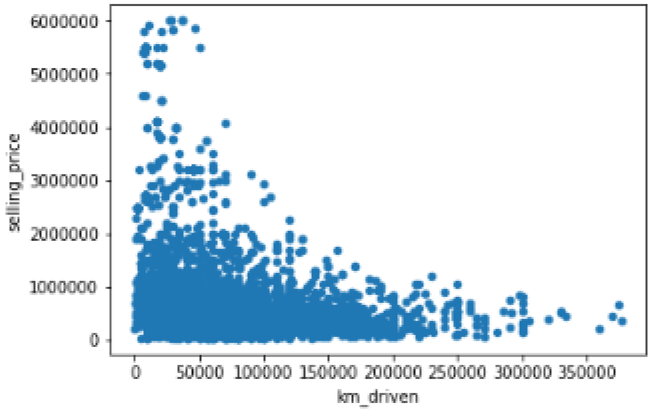


From the above two plots we are able to identify some outliers in both the columns

Another look at the scatter plot between KM driven and Selling Price



After deleting outliers

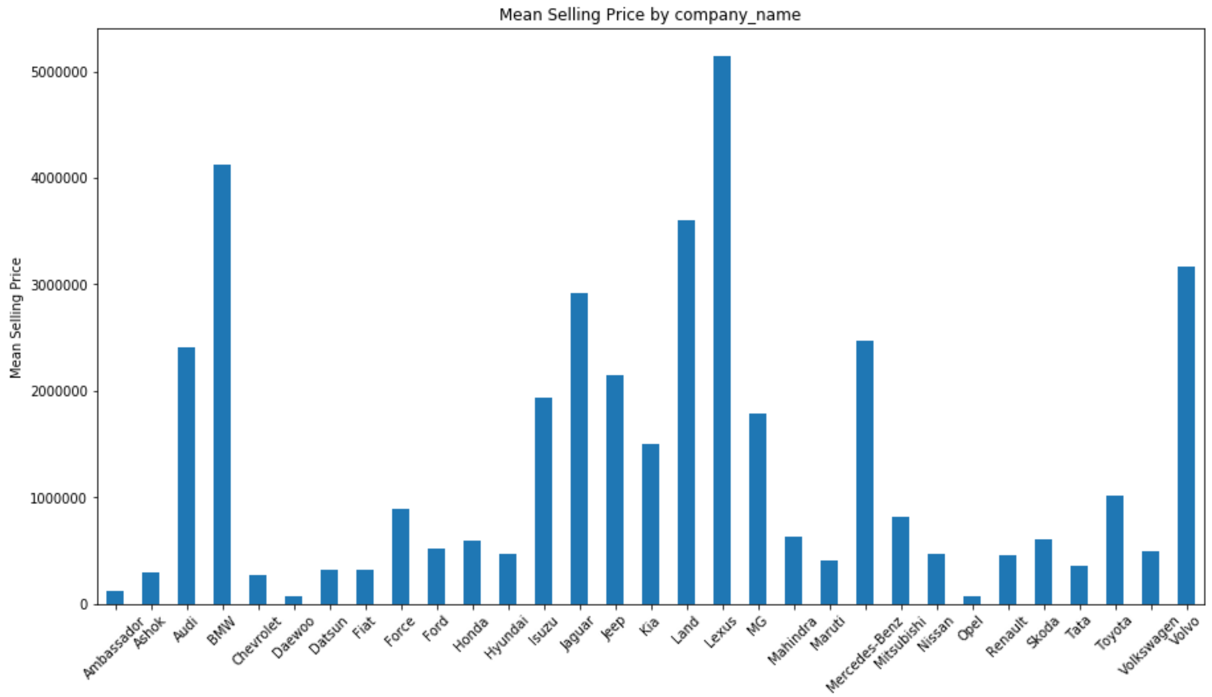


Key Findings and Insights:

Based on our insights in the EDA and the feature engineering section there are several columns were we identified pretty dense categorical features and we have reduced the number of categories by grouping them. We were also able to remove null values and treat outliers.

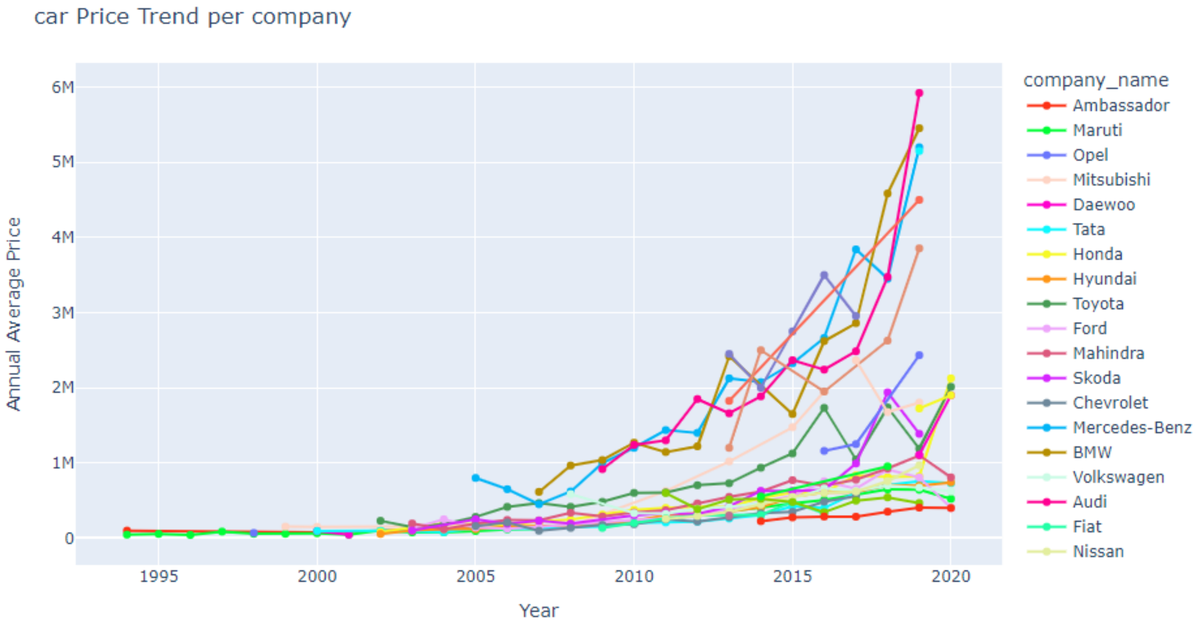
In our data set we also have one feature called name, which has the name of the car and by doing some string manipulation we were able to retrieve the name of the company of the car.

Below are some charts that show the mean selling price for the cars belonging to all the different companies



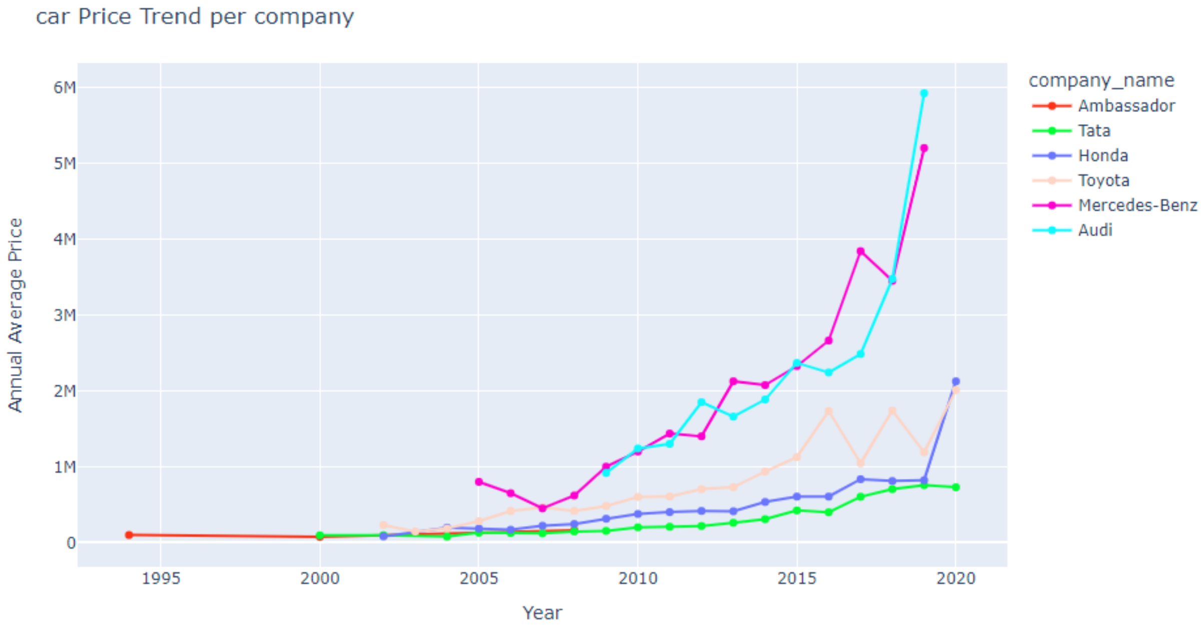
From the above chart we can tell that cars by lexus have the highest selling price followed by BMW and Volvo

Another look at the variation of the selling r=price per company based on the year of manufacture of the car



It’s too hard to tell much from this graph but we can easily see there is a difference in the price tends between the car companies

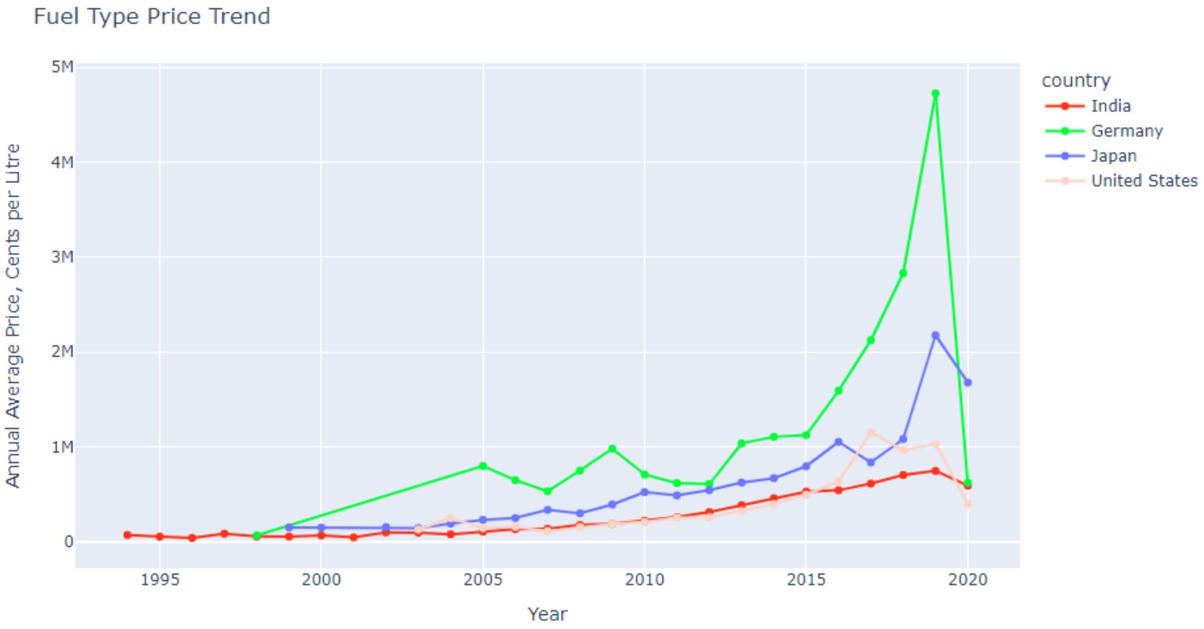
Similar example can be seen from the below graph as well with less number of car companies



But we know this from our own experiences and also from this graph that there is a significant price difference between luxury cars, premium cars and economical cars.

Based on this assumption again we divided the cars based on the company into these three categories

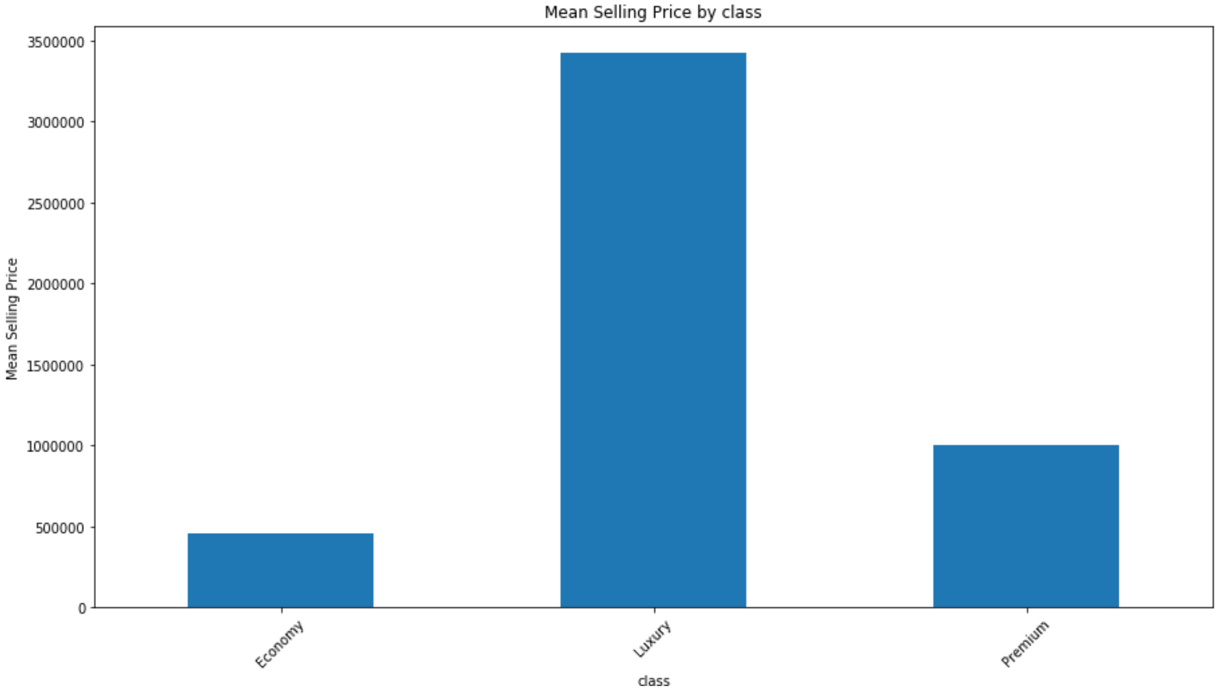
Also along with this from the name of the car companies we can also find which country these companies belong too and see if there is a difference in selling price based on the country of origin of the car companies



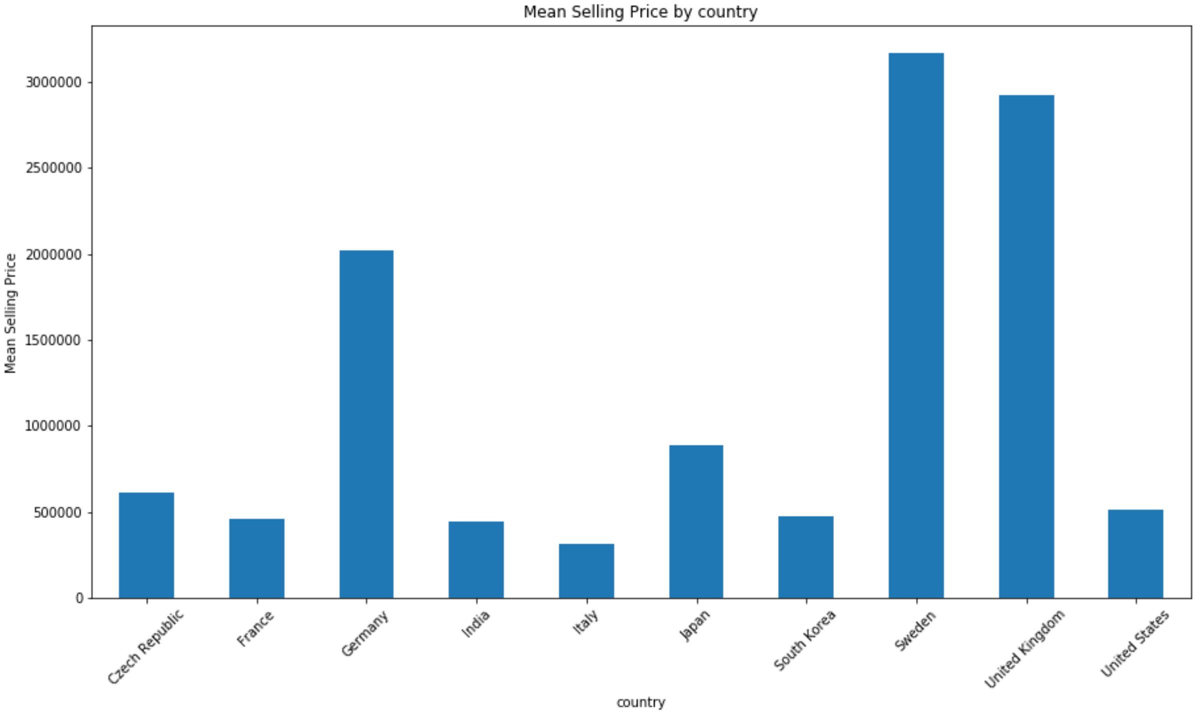
Again here we see that German cars then to have higher price followed by Japanese cars and by Indian cars

Using these we created two new columns Country and class

Below are the charts showing mean selling price based on country and class

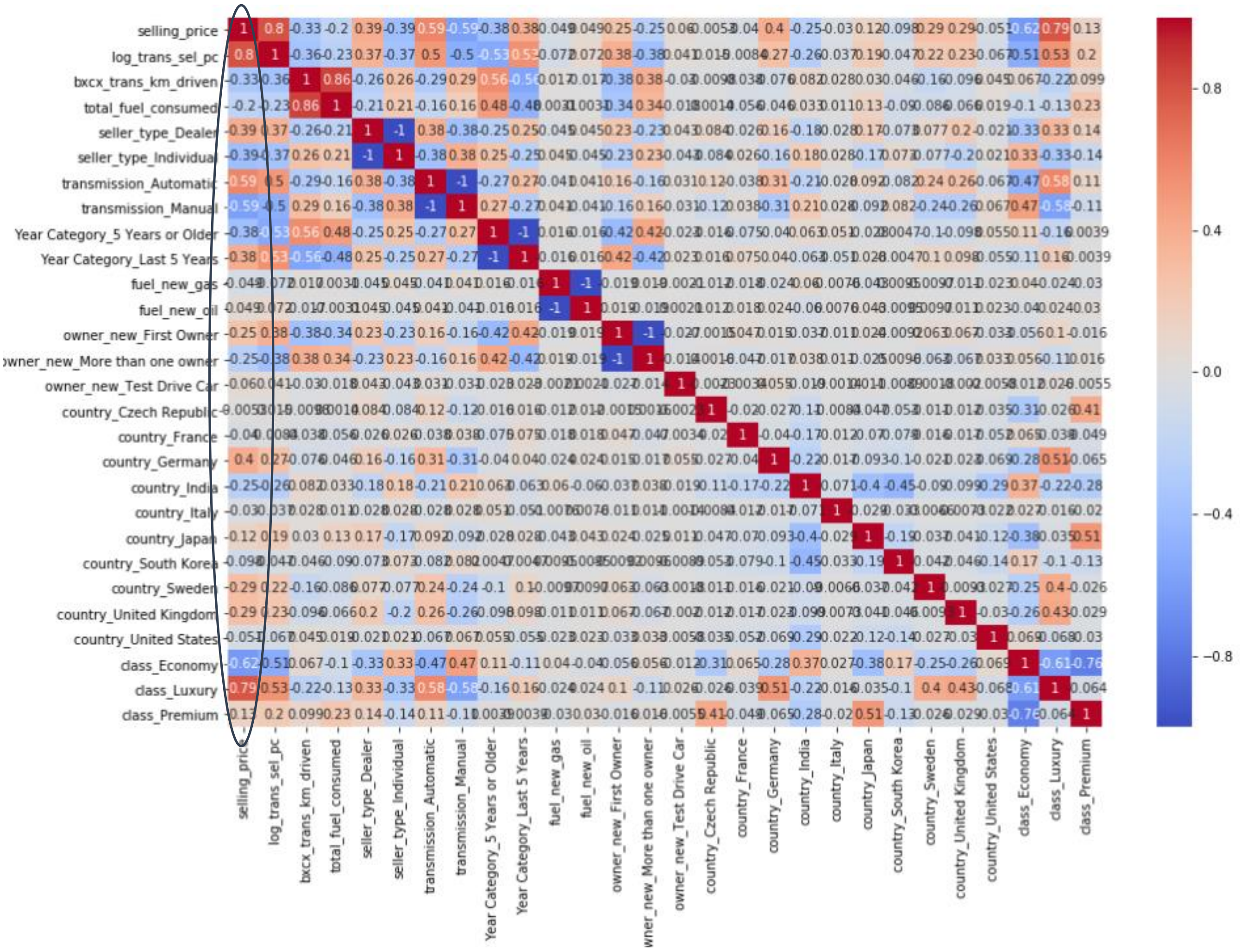


As expected Luxury cars have highest selling price followed by premium and economy cars



Again from this chart we can tell European cars have higher selling price than cars made in USA

Finally, after one hot encoding the rest of the categorical columns below is the correlation plot



Class variables are highly correlated with selling price, transmission variables, seller type

Formulating 3 hypothesis about this data:

1. More the KM driven for the car less is the selling price for the car this is the alternate hypothesis, whereas null hypothesis is km driven has nothing to do with selling price

2. Cars sold by dealers have higher Selling Price than the ones sold by Individuals; this is the alternate hypothesis, whereas null hypothesis is that dealers don't really have an impact on the car selling price

3. Less than 5 seat cars have higher selling Price than cars with more seats, null hypothesis number of seats in the car does not impact that selling price of the car

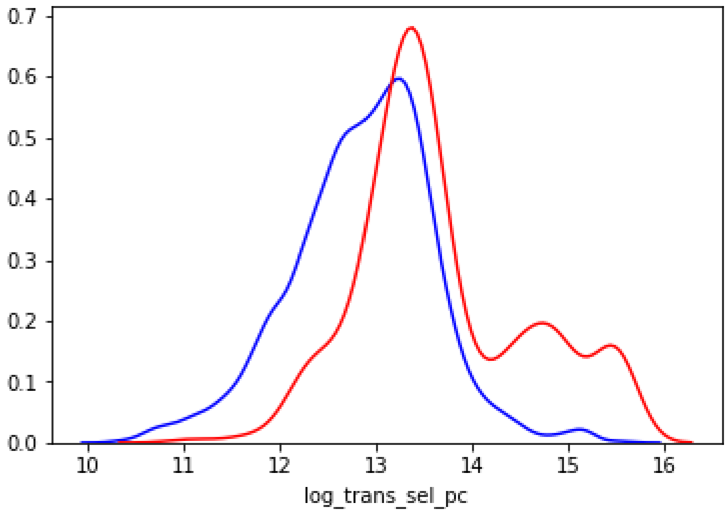
Threshold for the p value is 0.05 to reject the null hypothesis

Conducting a formal significance test:

We conducted a couple of hypothesis testing from the three hypothesis that we had about the data

Cars sold by dealership tend to sell for more price compared to cars sold by individual sellers. We had already seen that the mean selling price of cars sold by dealership were higher than the individual sellers. But let us confirm if this is actually true or by chance

We looked at the log transformed selling price distribution for the Dealership vs the Individual sellers



From the above graph we can see that there is quite a bit of overlap between the two, now we need to look if the probability of selling price for a dealership is higher than that of an individual seller given that the null hypothesis i.e. there is no difference in the selling price is greater than 0.05 %

From our code we find that the p value is a very small value, much smaller compared to 0.05.

Therefore, we can with confidence confirm that our hypothesis is true and we can reject the null hypothesis.

Next Steps:

At this point we have cleaned our dataset, removed nulls, removed outliers we have feature engineered most of the categorical columns, created some new derived columns as well and we see that some of these features have a high correlation with our predictor variable Selling Price.

After this step we will probably look more closely between the relation of log transformed selling price and rest of the feature set and look if there are some features that are collinear and can be removed.

We can also remove some features where the correlation is less than 0.2 as a threshold.

After that we look into some basic models like Ordinary least square model or Tree based models to see how the prediction is working for us.

Quality of this data set and a request for additional data if needed:

Overall the quality of the dataset was pretty good for looking into the relation between car selling prices and rest of the feature set. The features provided were really informative in understanding about the car. The usual specifications that a buyer looks into before buying a car were also provided. These were all very helpful information and shows a pretty high correlation with the selling price of the car.

Some more information that can be helpful if provided will be where are these cars being sold around which locations. So that we can study more about the selling price and also understand if cities and towns have some relationship with the price of the car. Maybe some more information about the seller can also be helpful. Like if the seller has previously sold cars, if so are there any ratings or metrics to understand how trustworthy the seller is. Along with this other information about the car that can be useful is if there was any damage to the car, color of the car, was there ever any accident that occurred with the car. Some information about the brake pads, 0-20s acceleration etc would also help us improve determining the price of the car.

Overall I enjoyed working with this dataset and would like to work on the next steps in the upcoming labs.