

Exploratory Data Analysis on the Cleaned Car Data

Here we will perform:

1. Univariate analysis
2. Bivariate analysis
3. Multivariate analysis

```
In [1]: # first let's import the necessary libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
pd.set_option("display.max_columns",None)
```

```
In [2]: # first let's import the cleaned car csv data
df = pd.read_csv("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\used_car_price_prediction.csv")
```

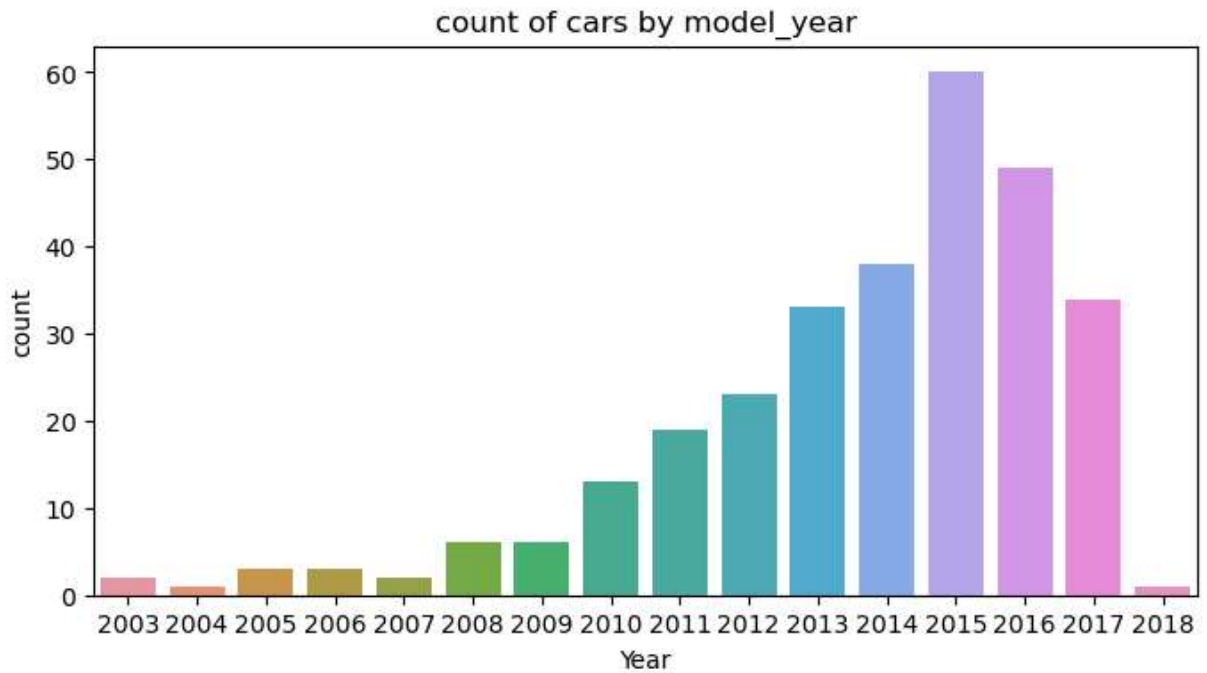
```
In [3]: df.sample(3)
```

```
Out[3]:
```

	Car_Name	Year	Age	Selling_Price(\$)	Present_Price(\$)	Kms_Driven	Fuel_Type	Seller_Type
230	Verna	2013	11	4.80	9.40	60241	Petrol	Individual
244	Brio	2015	9	5.40	6.82	31427	Petrol	Individual
248	City	2016	8	10.25	13.60	49562	Petrol	Individual

Univariate Analysis

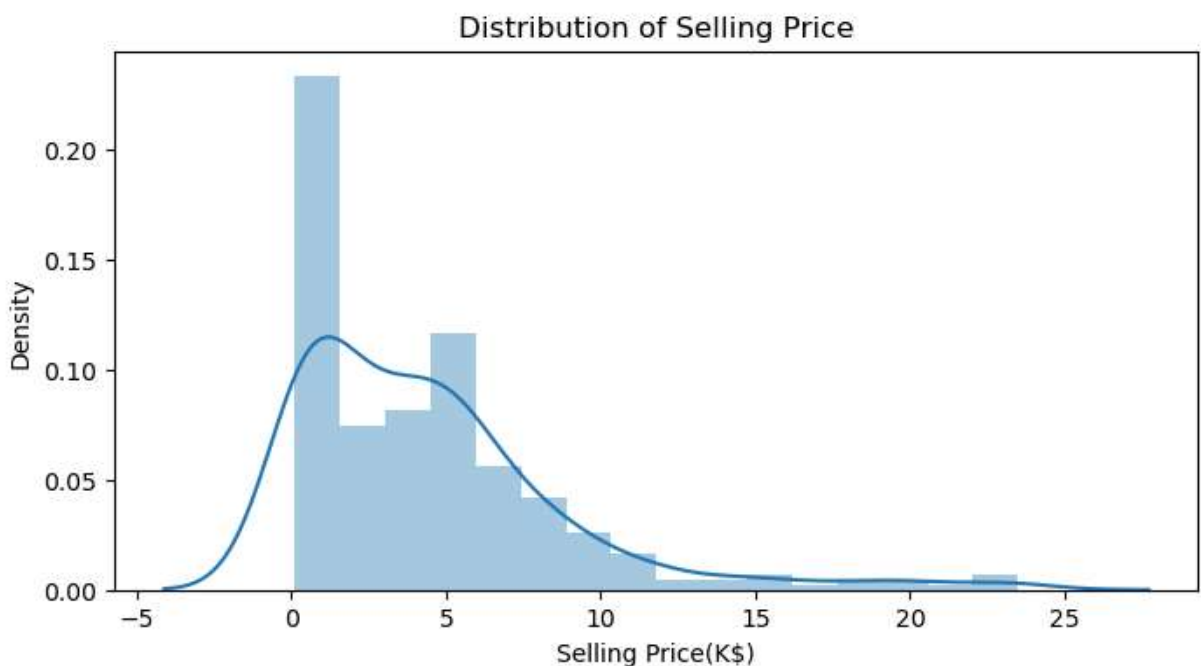
```
In [4]: # univariate analysis on year
plt.figure(figsize = (8,4))
sns.countplot(data = df,x = "Year")
plt.title("count of cars by model_year")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\used_car_price_prediction\\univariate_analysis\\count_of_cars_by_model_year.png")
plt.show()
```



From the above visualization we can clearly see that

1. Most of the cars model_year in the dataset falls between 2013 and 2017.
2. Cars with the model year 2015 have the highest count.

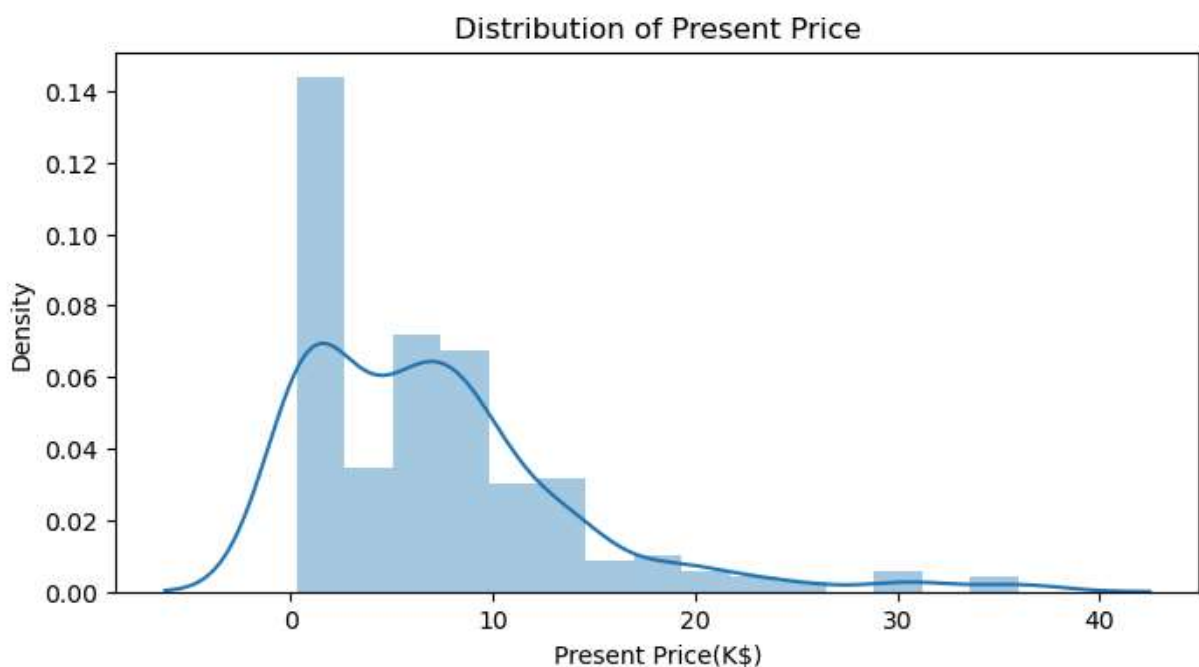
```
In [5]: # univariate analysis on Selling Price
plt.figure(figsize = (8,4))
sns.distplot(df["Selling_Price($)"])
plt.title("Distribution of Selling Price")
plt.xlabel("Selling Price(K$)")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



From the above Selling price distribution plot we can see that:

1. The plot is right skewed, means most of the cars have selling price which is less than the mean selling price.
2. most of the cars have selling price between 0 and 10k dollars.
3. The mean horsepower is greater than the median horsepower.

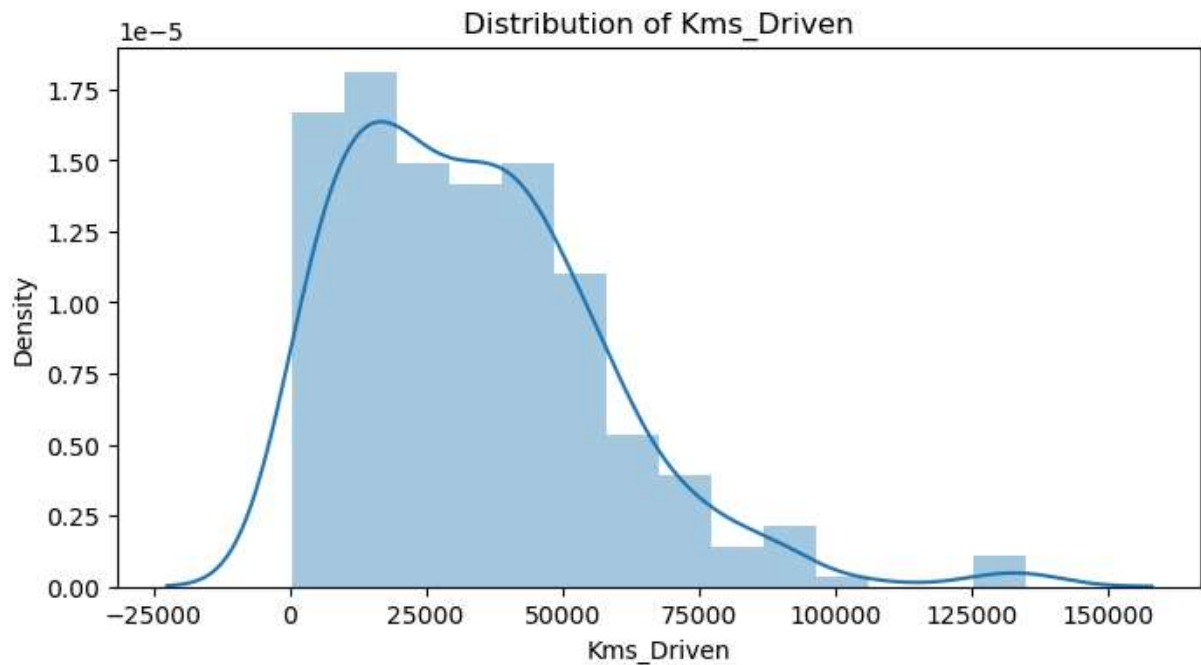
```
In [6]: # univariate analysis on Present Price
plt.figure(figsize = (8,4))
sns.distplot(df["Present_Price($)"])
plt.title("Distribution of Present Price")
plt.xlabel("Present Price(K$)")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



From the above plot we can see that

1. the plot is right skewed, that the most of the cars have the present price less than the mean present price.
2. the mean present price is less than that of the median present price.
3. most of the cars have present price between 0 and 15k dollars.

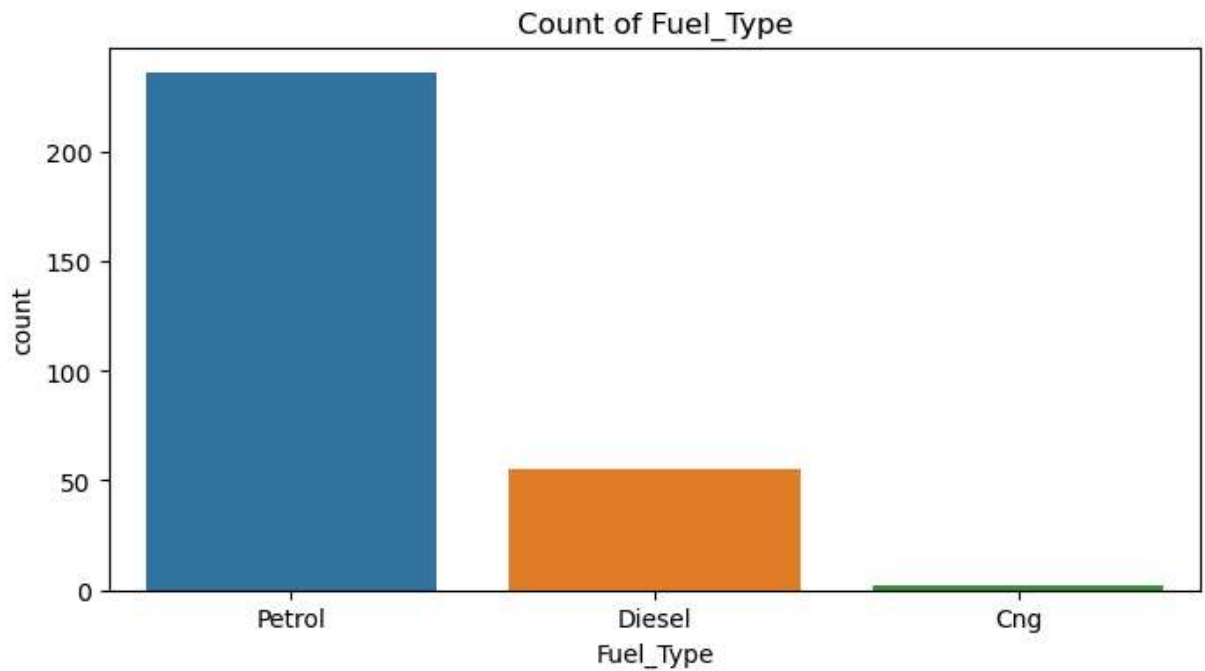
```
In [7]: # univariate analysis on Kms_Driven
plt.figure(figsize = (8,4))
sns.distplot(df["Kms_Driven"])
plt.title("Distribution of Kms_Driven")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



From the above distribution plot we can see that:

1. The plot is right skewed, means most of the cars have a value of kms_driven less than that of the mean kms_driven.
2. the mean value of kms_driven is greater than that of the median of kms_driven.
3. most of the cars have the value of kms_driven between 0 and 75000kms.

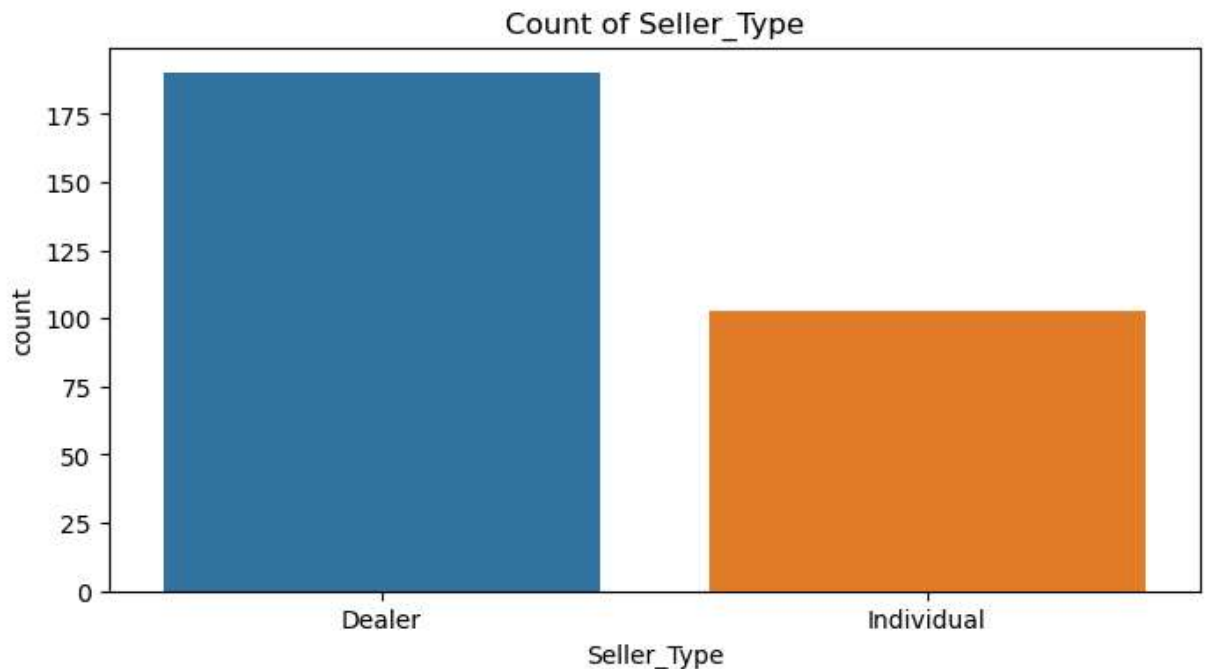
```
In [8]: # univariate analysis on Fuel_Type
plt.figure(figsize = (8,4))
sns.countplot(data = df, x = "Fuel_Type")
plt.title("Count of Fuel_Type")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



From the above plot we can see that:

1. The majority of cars in the dataset use petrol as their fuel.
2. The cars that use Cng(compressed neutral gas) as their fuel are too low.

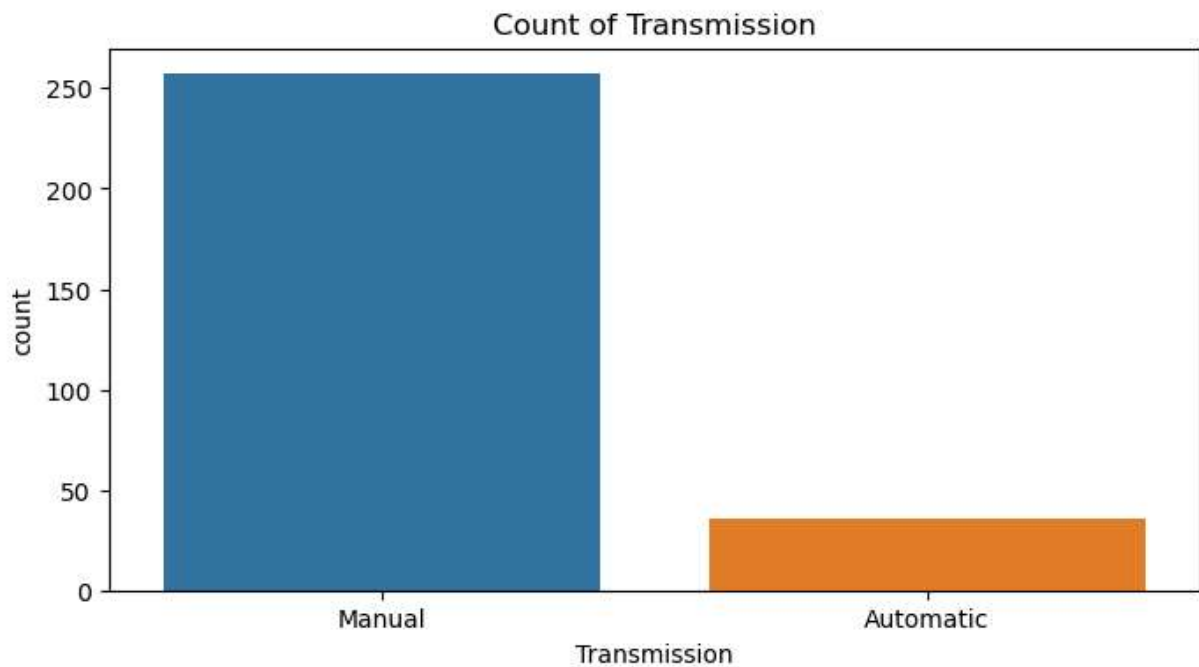
```
In [9]: # univariate analysis on Seller_Type
plt.figure(figsize = (8,4))
sns.countplot(data = df, x = "Seller_Type")
plt.title("Count of Seller_Type")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



From the above plot we can see that:

1. Most of the cars are sold by Dealers.

```
In [10]: # univariate analysis on Transmission
plt.figure(figsize = (8,4))
sns.countplot(data = df, x = "Transmission")
plt.title("Count of Transmission")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



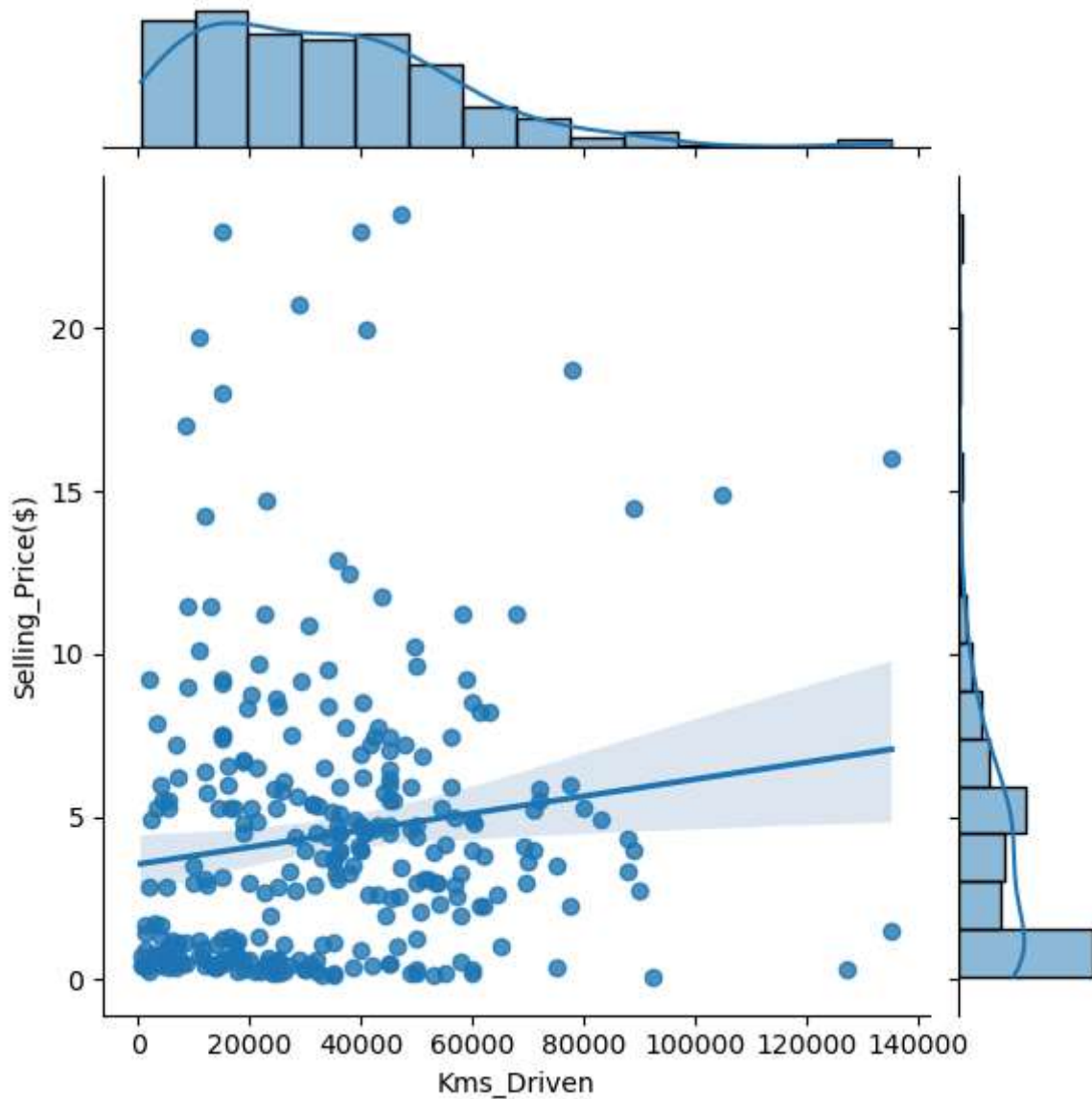
From the above plot we can see that:

1. Majority of the cars have a manual Transmission.

Bivariate Analysis

```
In [11]: # Bivariate Analysis between Selling_Price($) and Kms_Driven
plt.figure(figsize = (8,4))
sns.jointplot(data = df, x = "Kms_Driven", y = "Selling_Price($)", kind = "reg")
#plt.title("Selling_Price($) vs Kms_Driven")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```

<Figure size 800x400 with 0 Axes>



As we can see from the above plot:

1. most of the cars in the dataset have lower selling price and lower amount of driven kms.
2. from the plot we can see that kms_driven and selling price have slightly positive correlation, means as the kms_driven increases the selling price of the car also increases.

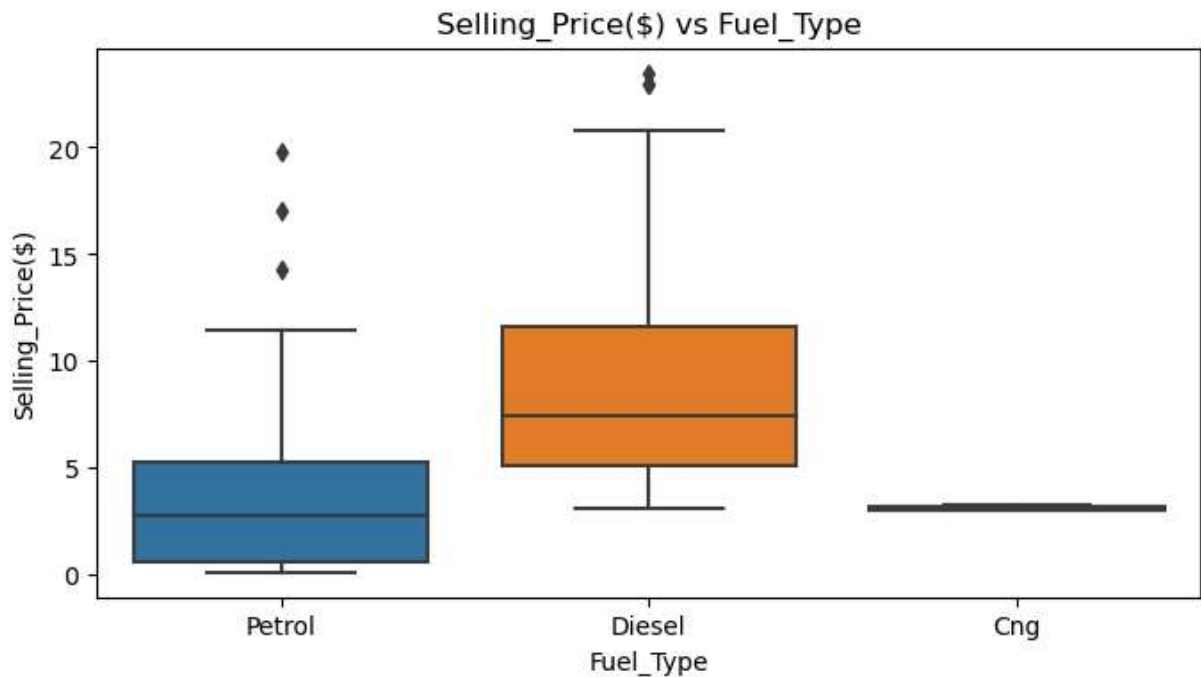
```
In [12]: # Bivariate Analysis between Selling_Price($) and Transmission
plt.figure(figsize = (8,4))
sns.boxplot(data = df, x = "Transmission", y = "Selling_Price($)")
plt.title("Selling_Price($) vs Transmission")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



From the above plot we can see that:

1. the median of selling price for automatic transmission vehicles is greater than that of the median selling price of the manual transmission vehicles.
2. the maximum selling price belongs to the automatic transmission vehicles.
3. there are still some selling price outliers in the manual transmission vehicles.
4. the selling price of manual transmission vehicles falls around the (mean of selling price of manual transmission vehicles) compared to that of the automatic transmission vehicles.

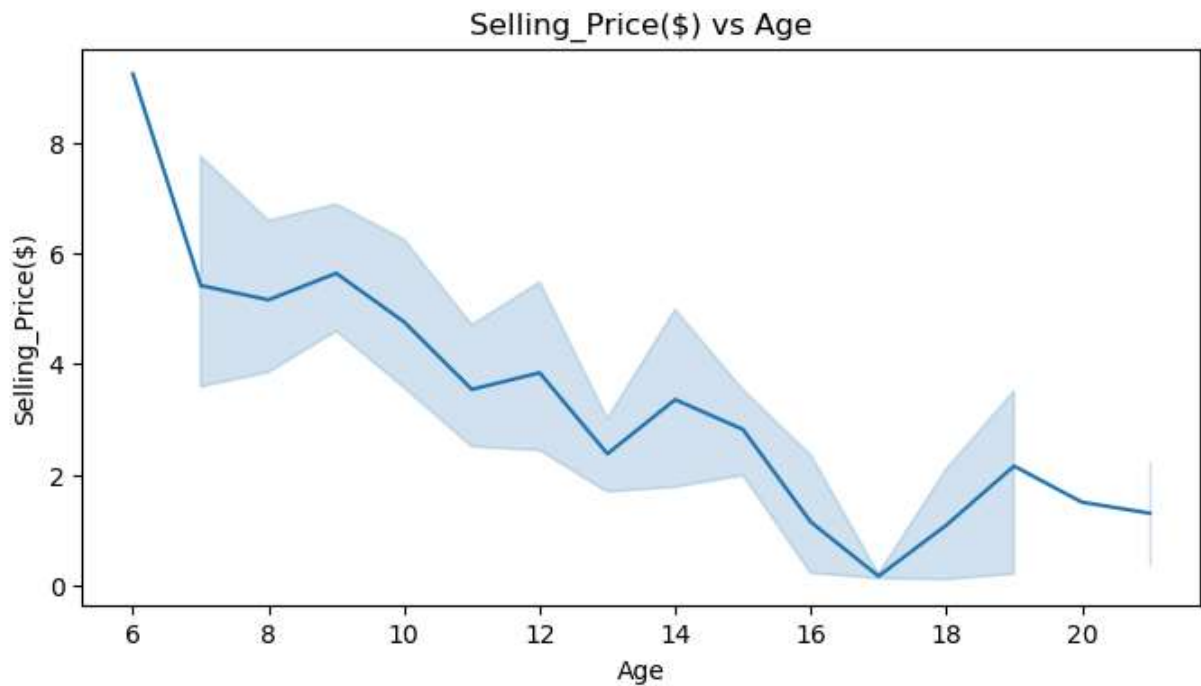
```
In [13]: # Bivariate Analysis between Selling_Price($) and Fuel_Type
plt.figure(figsize = (8,4))
sns.boxplot(data = df, x = "Fuel_Type", y = "Selling_Price($)")
plt.title("Selling_Price($) vs Fuel_Type")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```

From the above plot we can see that:

1. The median selling price for the Diesel cars is higher than the selling price of other fuel type cars.
2. There are selling price outliers in the Petrol and Diesel type cars.
3. The number of cars with Cng(compressed neutral gas) is too small.
4. The maximum selling price belongs to the Diesel fuel type cars.

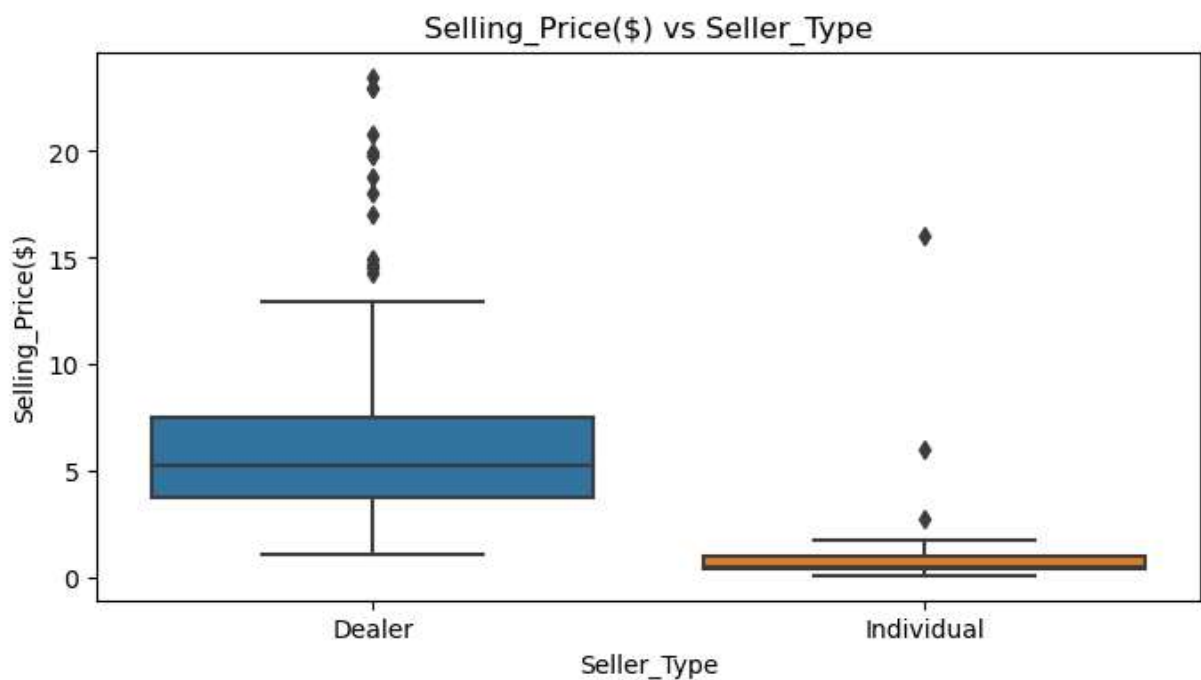
```
In [14]: # Bivariate Analysis between Selling_Price($) and Age
plt.figure(figsize = (8,4))
sns.lineplot(data = df, x = "Age", y = "Selling_Price($)")
plt.title("Selling_Price($) vs Age")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



As we can see From the above plot:

1. The selling price of cars and Age variable have reverse correlation, as the age increases the selling price of the car decreases.

```
In [15]: # Bivariate Analysis between Selling_Price($) and Seller_Type
plt.figure(figsize = (8,4))
sns.boxplot(data = df, x = "Seller_Type", y = "Selling_Price($)")
plt.title("Selling_Price($) vs Seller_Type")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



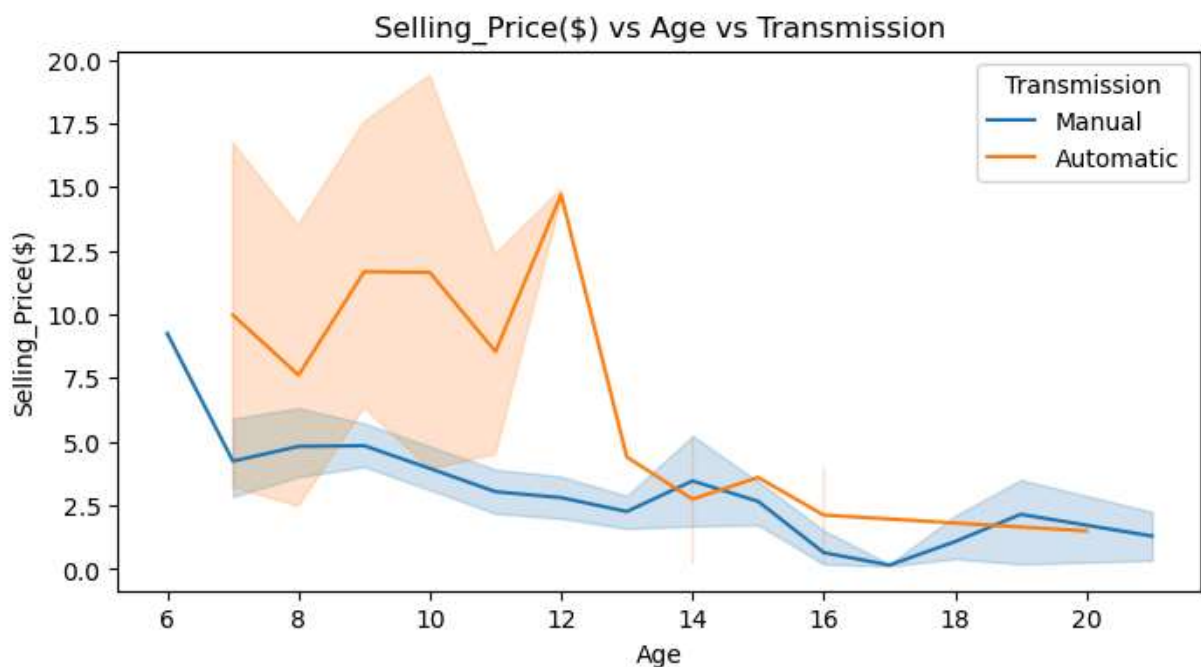
As we can see from the above plot:

1. The median selling price of cars sold by Dealer is greater than that of the median selling price of cars sold by individuals.
2. The highest amount of selling price belongs to cars sold by Dealer.
3. Most of the selling prices for cars sold by individual falls around the mean selling price of cars sold by individual.
4. More selling price outliers found in cars Sold by Dealer.

Multivariate Analysis

```
In [16]: # multivariate analysis on Selling_Price($) vs Age vs Transmission

plt.figure(figsize = (8,4))
sns.lineplot(data = df, x = "Age", y = "Selling_Price($)", hue = "Transmission")
plt.title("Selling_Price($) vs Age vs Transmission")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```



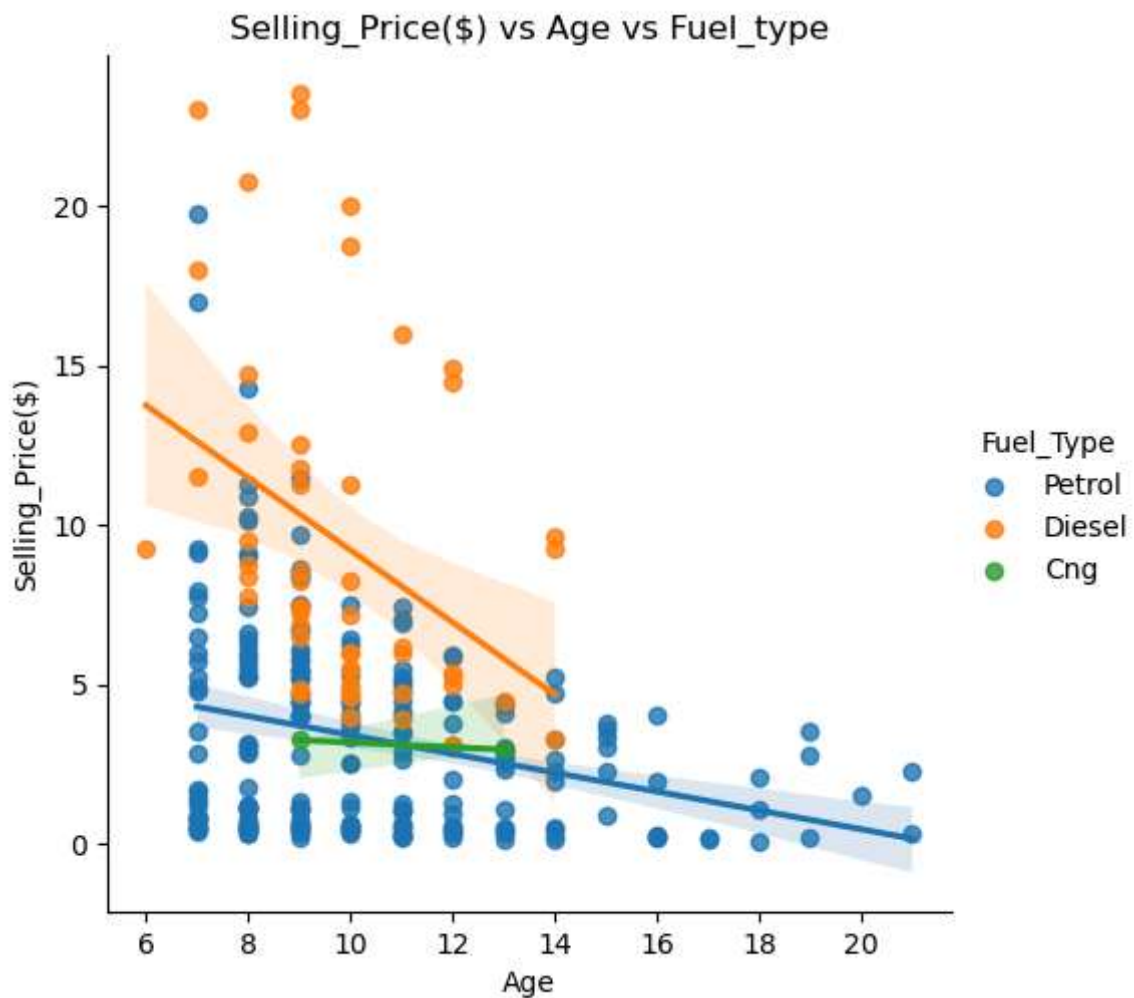
As we can see from the above plot:

1. As the age increases the selling price for both manual transmission and automatic transmission vehicles decreases.

```
In [17]: # multivariate analysis on Selling_Price($) vs Age vs Fuel_Type

plt.figure(figsize = (8,4))
sns.lmplot(data = df, x = "Age", y = "Selling_Price($)", hue = "Fuel_Type")
plt.title("Selling_Price($) vs Age vs Fuel_type")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```

<Figure size 800x400 with 0 Axes>



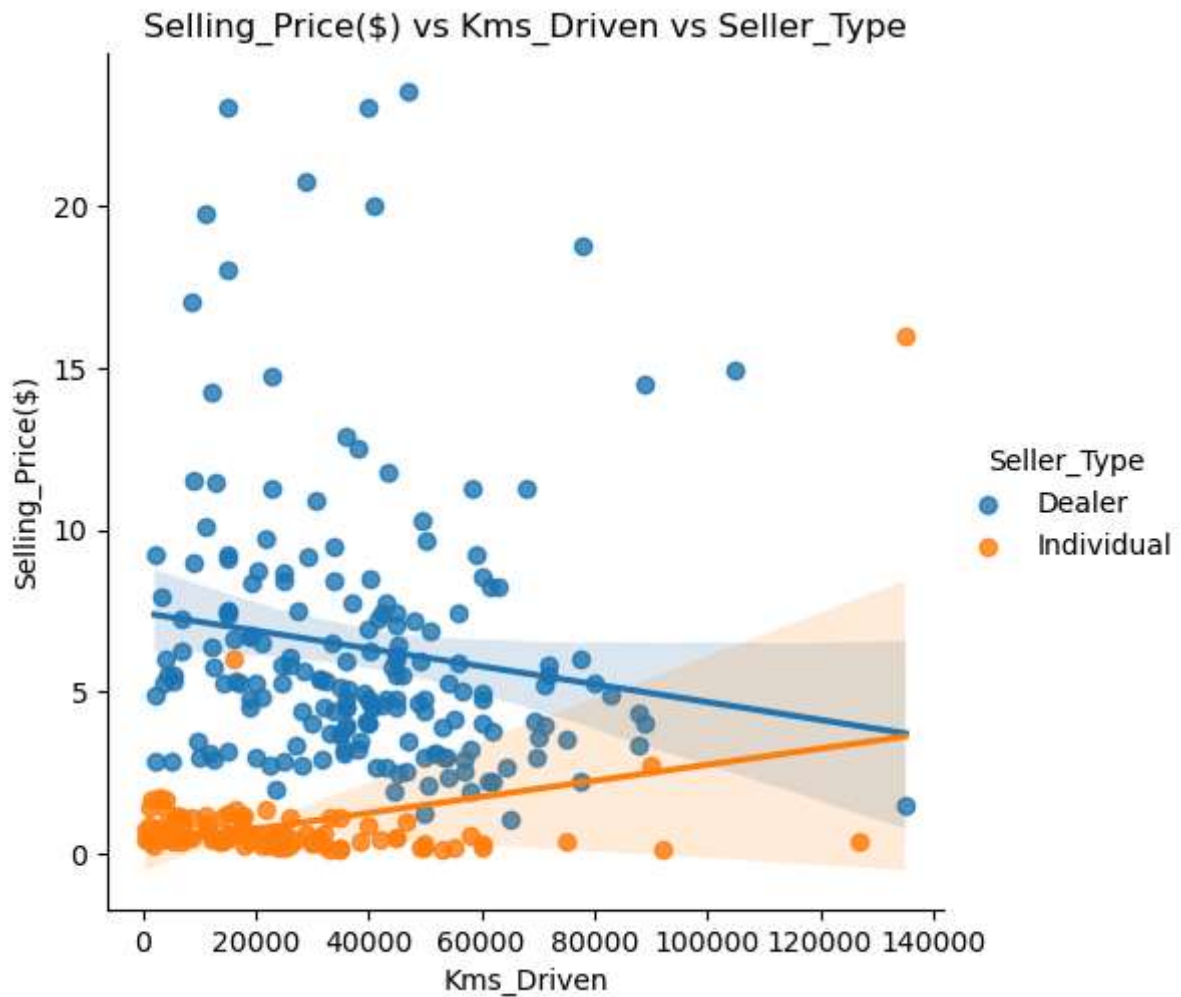
As we can see from the above plot:

1. As the age increases the selling price decreases for all fuel_type cars.
2. The decrease in selling price as the age increase is higher for the cars with Diesel Fuel type.
3. Relatively the Decrease in selling price for as the Age increases is negligible for the Cng(compressed neutral gas) cars, this is caused by small amount of data points for the Cng fuel type cars in the dataset.

```
In [18]: # multivariate analysis on Selling_Price($) vs Kms_Driven vs Seller_Type

plt.figure(figsize = (8,4))
sns.lmplot(data = df, x = "Kms_Driven", y = "Selling_Price($)", hue = "Seller_Type")
plt.title("Selling_Price($) vs Kms_Driven vs Seller_Type")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```

<Figure size 800x400 with 0 Axes>

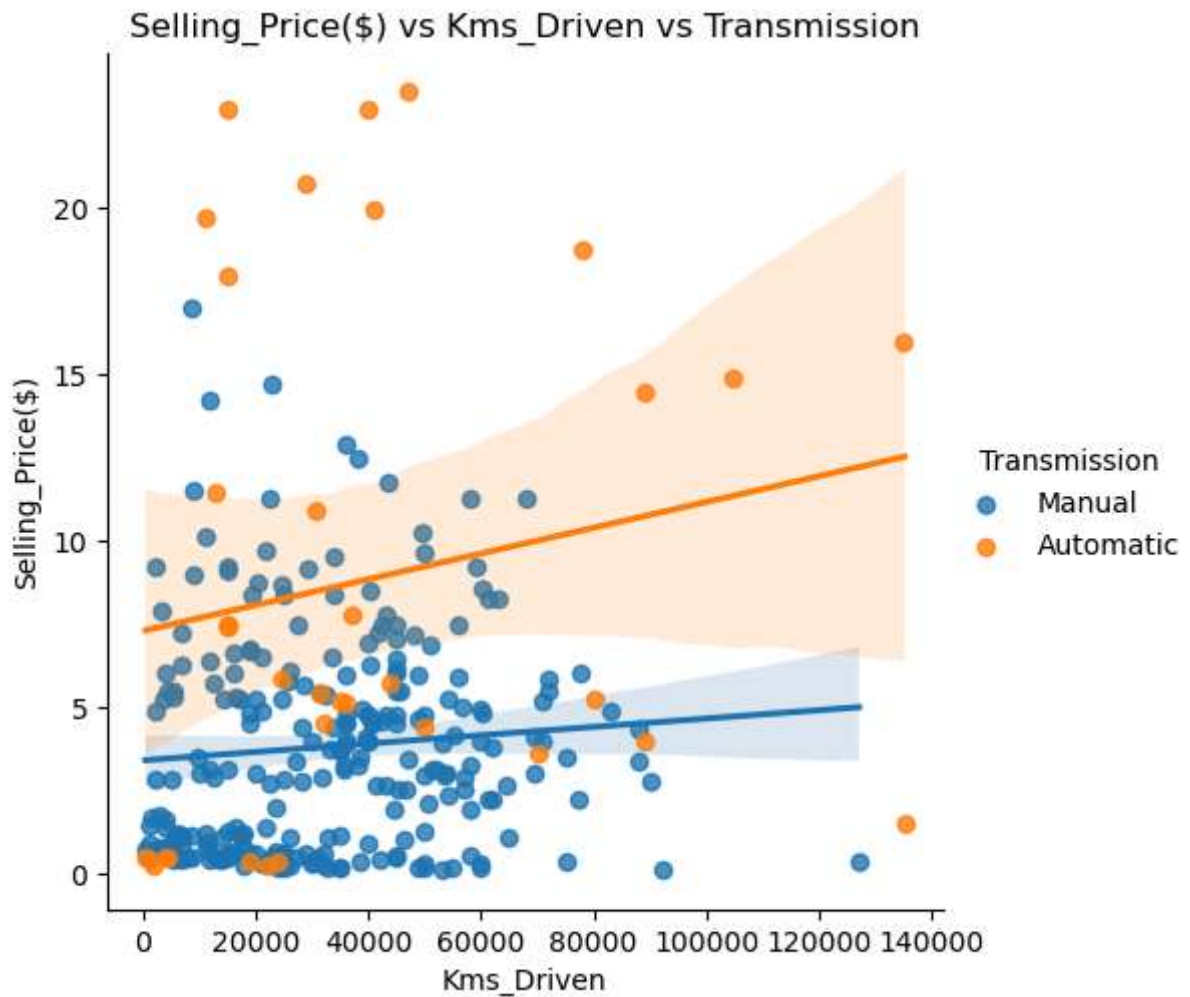


As we can see from the above plot:

1. As the Kms Driven increases the selling price for cars sold by Dealer decreases.
2. As the Kms Driven increases the selling price for cars sold by Individual increases.

```
In [19]: # multivariate analysis on Selling_Price($) vs Kms_Driven vs Transmission
plt.figure(figsize = (8,4))
sns.lmplot(data = df, x = "Kms_Driven", y = "Selling_Price($)", hue = "Transmission")
plt.title("Selling_Price($) vs Kms_Driven vs Transmission")
plt.savefig("C:\\Users\\yozil\\Desktop\\My projects\\used car price prediction\\use
plt.show()
```

<Figure size 800x400 with 0 Axes>



As we can see from the above plot:

1. As the Kms driven increases the selling price also increases for both manual and automatic transmission cars.
2. The increase in selling price as the kms_driven is higher for the automatic transmission cars than that of the manual transmission vehicles.

In []:

In []: