Predictive Analytics: Assignment 1.2

Joshua Greenert

DSC630-T301 Predictive Analytics

11/30/2022

Write a summary of your data and identify at least two questions to explore visually with your data.

The dataset selected contains information over cars that were sold from 1992 to 2020. In addition to the year sold, the data contains information about the fuel type, killometers driven, selling price, transmission type, seller type, and owner (whether it was the first owner or beyond). This dataset was selected from Kaggle on 11/30/2022 from the URL linked below.
https://www.kaggle.com/datasets/akshaydattatraykhare/car-details-dataset?resource=download The two questions that we'll explore in this exploratory data analysis are:

What car manufacturer (otherwise known as the make) typically sells at the highest price for first owners?

How many average miles are used up on the car for first owners?

Prepare the data

In [1]:
```
# Import the required libaries.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Pull in the data from the GitHub Repo.
df_cars = pd.read_csv("car_details.csv")
df_cars.head(5)
```

Out[1]:

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner

In [2]:
```
# Split the make from the front of the car name; Use every first word as the indicator o
car_make = []

# loop over all of the cars, split them by space, and add the first word to the array.
for row in df_cars.name:
    carSplit = row.split(" ")
    car_make.append(carSplit[0])

# Add the new column to the dataframe.
```

```
df_cars['make'] = car_make

# Show that the operation was successful.
df_cars.head(5)
```

Out[2]:

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	make
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner	Maruti
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner	Maruti
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner	Hyundai
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner	Datsun
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner	Honda

Create a histogram or bar graph from your data.

In [25]:
```
# Group by name, then by owner.
grouped_df = df_cars.groupby(['name', 'owner']).mean().sort_values(by=['selling_price'],

# Show the top 10 based on name
grouped_df.head(10)
```

Out[25]:

		year	selling_price	km_driven
name	**owner**			
Audi RS7 2015-2019 Sportback Performance	**First Owner**	2016.0	8900000.0	13000.0
Mercedes-Benz S-Class S 350d Connoisseurs Edition	**First Owner**	2017.0	8150000.0	6500.0
Mercedes-Benz GLS 2016-2020 350d 4MATIC	**First Owner**	2016.0	5500000.0	77350.0
BMW X5 xDrive 30d xLine	**First Owner**	2019.0	4950000.0	30000.0
Audi A5 Sportback	**First Owner**	2020.0	4700000.0	1500.0
Volvo XC 90 D5 Inscription BSIV	**First Owner**	2017.0	4500000.0	80000.0
Mercedes-Benz E-Class Exclusive E 200 BSIV	**First Owner**	2018.0	4500000.0	9800.0
Mercedes-Benz GL-Class 350 CDI Blue Efficiency	**Second Owner**	2014.0	4400000.0	100000.0
Land Rover Range Rover 4.4 Diesel LWB Vogue SE	**First Owner**	2010.0	4200000.0	100000.0
BMW 7 Series Signature 730Ld	**First Owner**	2014.0	4000000.0	47000.0

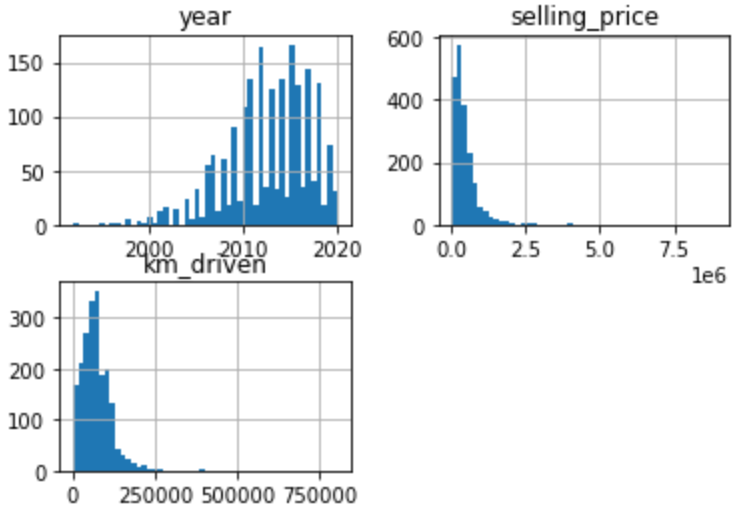
In [46]:
```
# Group by name, then by owner.
grouped_df_driven = df_cars.groupby(['owner']).mean().sort_values(by=['km_driven'], asce

# Show the top 10 based on name
grouped_df_driven.head(5)
```

Out[46]:

	year	selling_price	km_driven
owner			
Third Owner	2009.338816	269474.003289	99304.506579

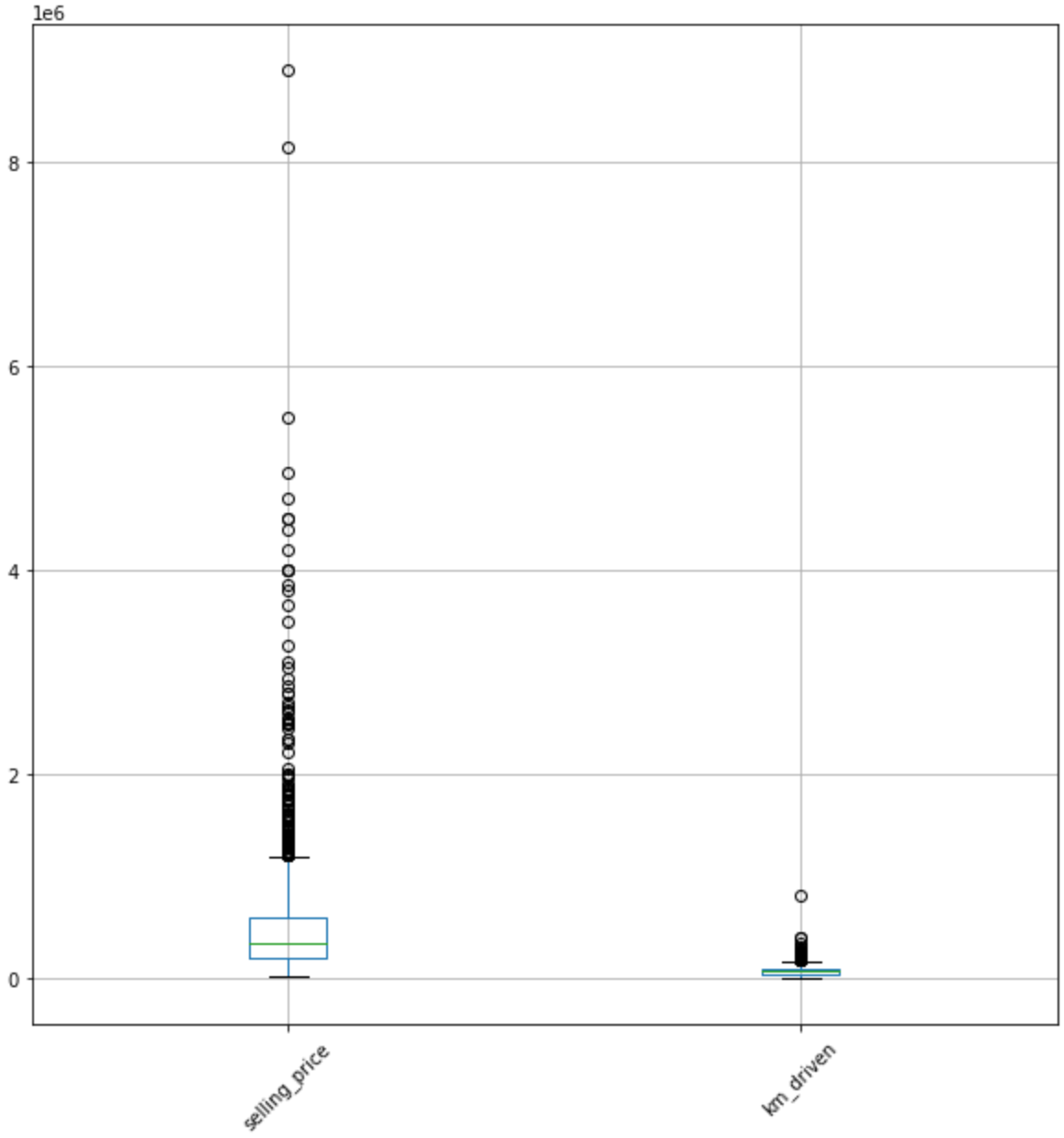
Fourth & Above Owner	2007.395062	173901.197531	99138.135802
Second Owner	2010.983725	343891.088608	81783.518987
First Owner	2014.440678	598636.969633	56015.009887
Test Drive Car	2019.529412	954293.941176	4155.000000

In [27]:
```
# Plot a histogram.
hist = grouped_df.hist(bins=50)
```

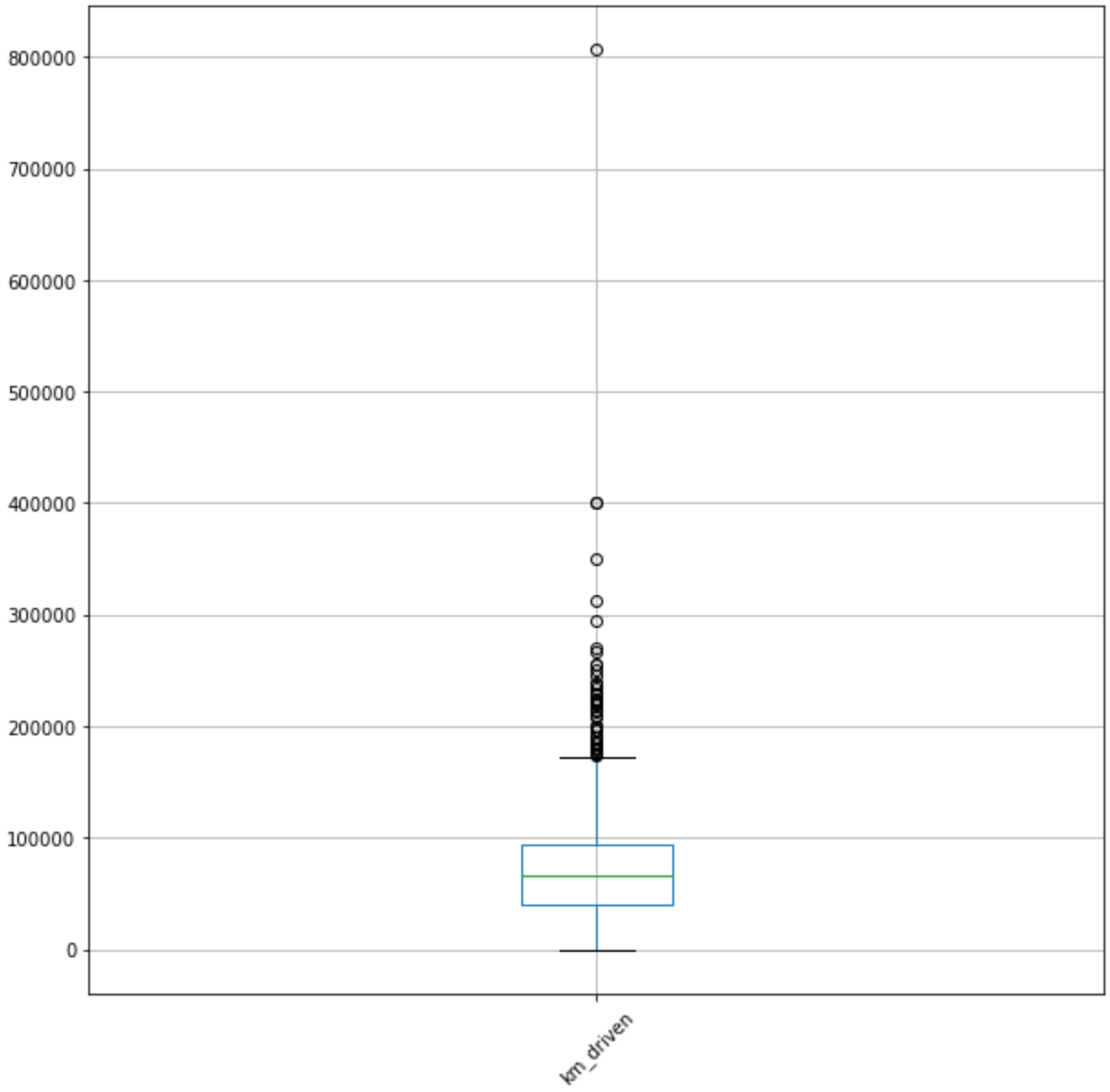


Create a boxplot from your data.

In [37]:
```
# Create a boxplot using the selling price.
boxplot = grouped_df.boxplot(column=['selling_price'],rot=45, fontsize=10, figsize= (10,
```

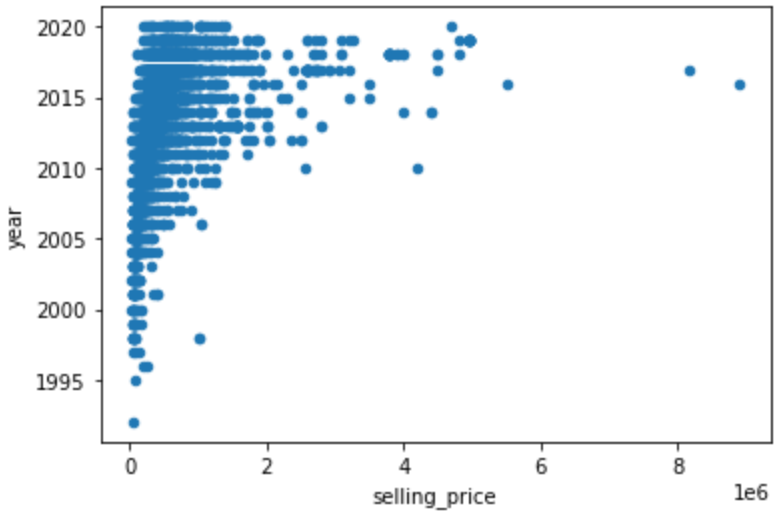


In [44]:
```
# Create a boxplot using the km driven
boxplot = grouped_df.boxplot(column=['km_driven'],rot=45, fontsize=10, figsize= (10, 10)
```



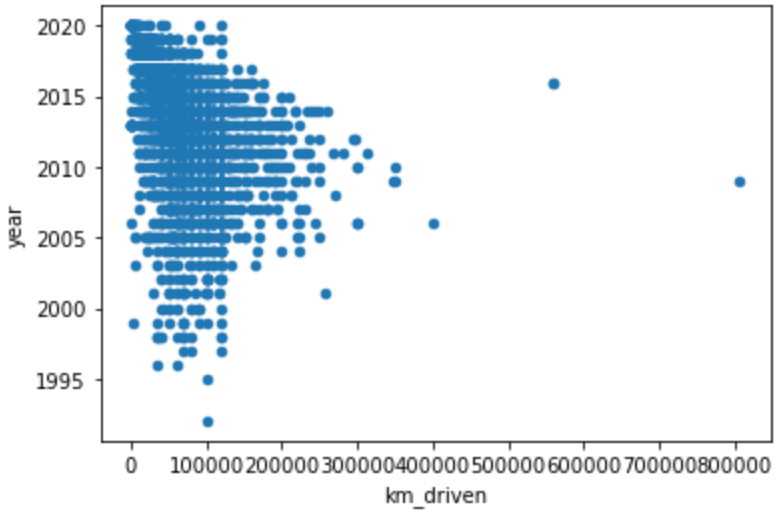
Create a bivariate plot from your data.

In [41]:
```
# Plot a bivariate plot for selling price
df_cars.plot(x='selling_price', y='year', kind='scatter')
plt.show()
```



In [45]:
```
# Plot a bivariate plot for killometers driven
df_cars.plot(x='km_driven', y='year', kind='scatter')
```

```
plt.show()
```



Conclusion

To reiterate, there were two questions that we were attempting to solve while producing the visualizations above.

1. What car manufacturer (otherwise known as the make) typically sells at the highest price for first owners?
2. How many average miles are used up on the car for first owners?

The scatter plots and box plots undoubtedly express that there are outliers within the dataset. Regardless, it's apparent that first owners of Audi, Mercedes-Benz, and BMW vehicles typically sell at the highest price. Notably, these vehicles tend to be higher-priced vehicles from the offset anyways, so this was somewhat anticipated for the end result. In terms of mileage, the average amount of miles used up prior to selling for first owners is approximately 56,015 kilometers (or 34,806.11 miles).