Denise Dodd

Predicting Used Car Prices

10 Questions and Answers

Question #1: Is there a way to increase the model's accuracy?

In my initial ideation of this project, I used a grid search to find the best hyperparameters for the Random Forest model. The grid search took extensive amounts of time and tied up much of my computer's memory which slowed down processing. When the hyperparameters were returned I used them to retrain my model and the R-Squared value only increased by .001 indicating that they hypertuned model which paused all other processing on my computer for a great length of time was only able to explain .1% more of the variance in selling price. I made the decision that the time taken to complete the grid search was not worth the minimally improved performance. If you choose to run this project, you can weigh the pros and cons and make a decision that is right for you regarding if you would like to hyper-tune the model to improve performance.

Question #2: What is the benefit of using a pipeline?

In addition to making the code more streamlined and readable, using a pipeline makes it more efficient to repeat multiple steps. For example, my pipeline scales numerical columns and makes dummies of categorical columns. I could have added additional tasks in my pipeline such as a grid search for hyperparameters (as referenced in previous question). Rather than repeating these steps multiple times, I just had to pass my data through the pipeline when selecting my data via my loop, when creating an instance of my model outside of the loop, and when making predictions on my own dataset at the end of the project. This makes the code "cleaner" than constantly repeating the same steps or introducing additional variables into the code.

Question #3: What if I only know some of the variables in my predictive dataframe?

There are two options to accommodate this. The option I would suggest is that you enter the missing data to the best of your ability. For example, if you are unsure what kind of transmission you are looking for, take your best guess of the transmission that you are leaning towards. Another way to rectify this with the predictive dataframe is to make two entries: one with an automatic transmission and one with a manual transmission.

The other option is to remove the transmission column from the original features variable and retrain the model. However, this will be more time consuming, and the loss of training variables will likely result in lower efficiency in the model.

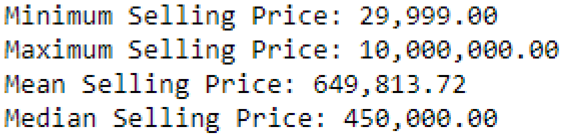
Question #4: This data has a variable titled km_driven, I would like to review this variable in terms of miles.

This can be done by creating a new column which multiples the km_driven column by 0.621371 to convert it to miles and changing the name of the column to mi_driven. Once the mi_driven column has been created, I would suggest dropping the km_driven column to prevent multicollinearity.

The mileage(km/ltr/kg) column is also measured in km. If the km_driven column is converted to miles I would recommend also changing the mileage(km/ltr/kg) column to miles so your distance units are consistent across the dataset.

Question #5: I want a general idea of how much a used car costs so I understand the volume of money that is involved.

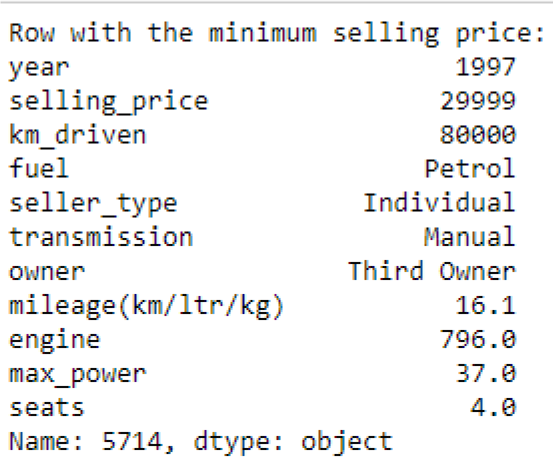
The cost of a car varies based on different variables, but below are the overall stats from this dataset. Bear in mind that the unit of currency for this dataset is unknown. If this project is run using a dataset of your local use cars, you will know the unit of currency.



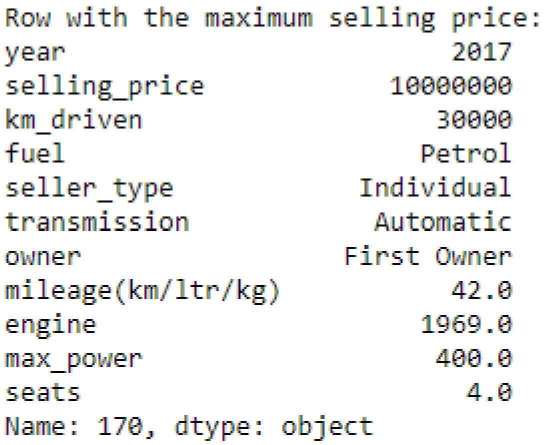
```
Minimum Selling Price: 29,999.00
Maximum Selling Price: 10,000,000.00
Mean Selling Price: 649,813.72
Median Selling Price: 450,000.00
```

Question #6: There is a huge difference between the min and max values! Can I see the details for the cars with the min and max values so I can understand the wide spread in selling price?

You can see below that the car with the minimum selling price is much older, has many more kms driven, is a manual transmission, and on its third owner. Additionally, the minimum price car has much lower mileage and engine power than the maximum priced car.



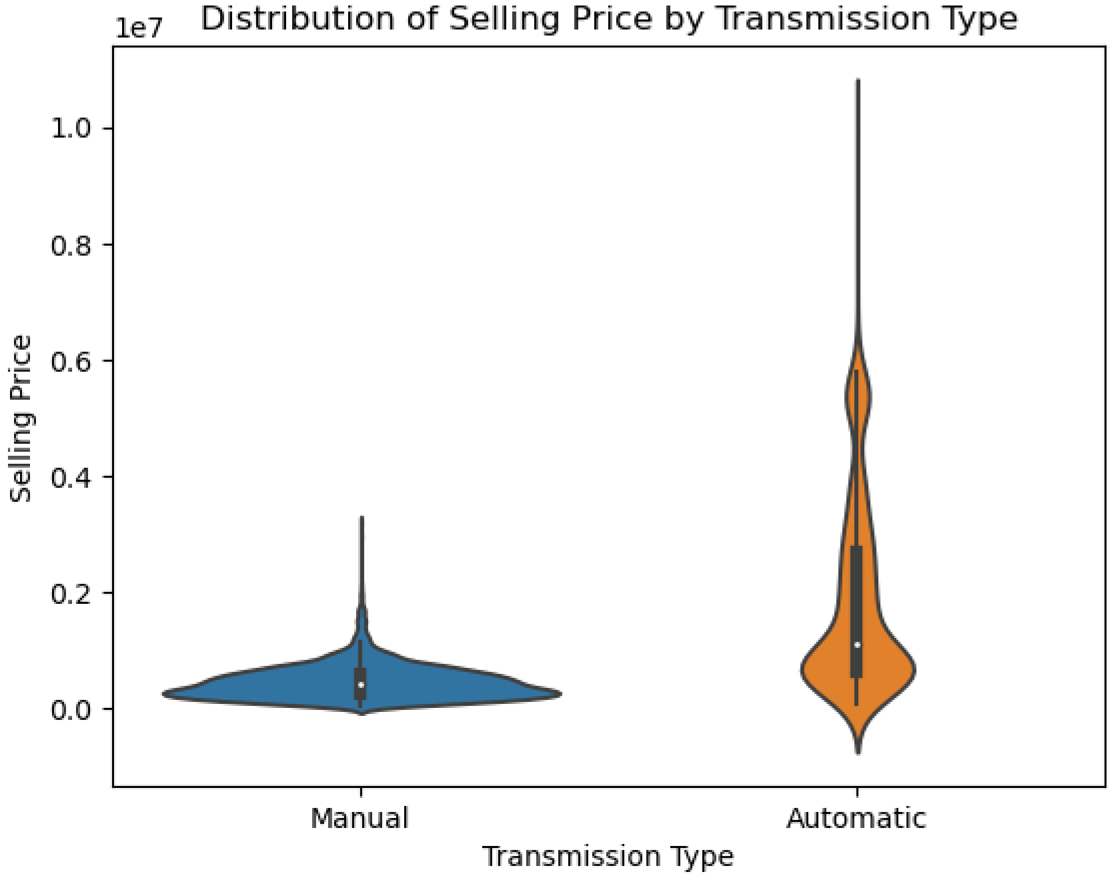
```
Row with the minimum selling price:
year                               1997
selling_price                     29999
km_driven                         80000
fuel                             Petrol
seller_type                  Individual
transmission                     Manual
owner                       Third Owner
mileage(km/ltr/kg)                 16.1
engine                            796.0
max_power                          37.0
seats                               4.0
Name: 5714, dtype: object
```



```
Row with the maximum selling price:
year                               2017
selling_price                  10000000
km_driven                         30000
fuel                             Petrol
seller_type                  Individual
transmission                  Automatic
owner                       First Owner
mileage(km/ltr/kg)                 42.0
engine                           1969.0
max_power                         400.0
seats                               4.0
Name: 170, dtype: object
```

Question #7: Transmission is by far the most important variable in the model based on the Features Importance graph. How does selling price differ between the two transmission types?

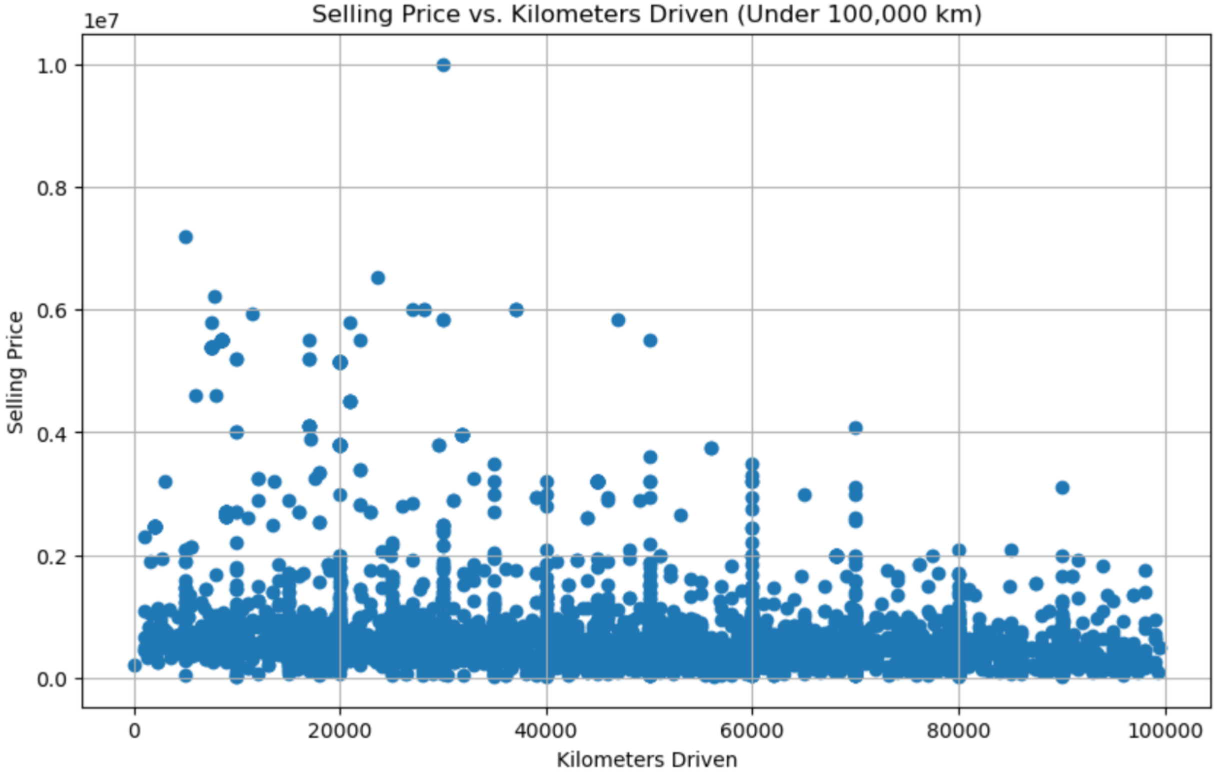
Manual transmission has a smaller range of selling price with prices ranging from slightly above 0 to outliers close to 400,000. Most of the selling prices are below 200,000. Automatic transmissions have a larger range with outliers as high as 10,000,000. Overall, automatic transmissions are more expensive than manual transmissions.



Question #8: How does km_driven affect selling_price?

Km_driven is included in the top 5 most important features for the model, but it still has a low importance score. Additionally, the correlation matrix shows that km_driven has a -.22 correlation with sale price. This is not a very strong correlation, but it does show that the variables move in opposite directions (while one variable increases the other decreases.)

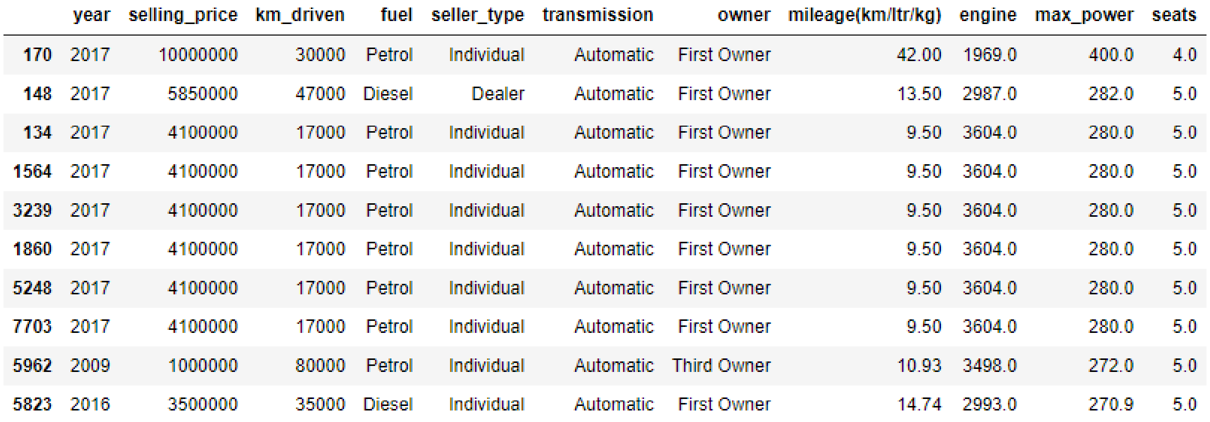
The below scatterplot was created from a filtered dataframe only showing entries with km_driven less than 100,000. There are outliers with km_driven far greater than 100,000 which caused the plots to gather and be difficult to read. This scatterplot shows that for the most part, regardless of the number of kilometers driven, most cars have a selling price between 0 and 200,000. The lower km_driven cars have outliers with higher selling prices which explains the -.22 correlation.



Question #9: I want a car with a powerful engine. Can I see the 10 cars with the most max_power?

The car with the selling_price outlier also appears to be an outlier in terms of max_power. The data seems to be accurate for this entry, but a user might want to consider removing this entry prior to fitting the model due to its outlier status.

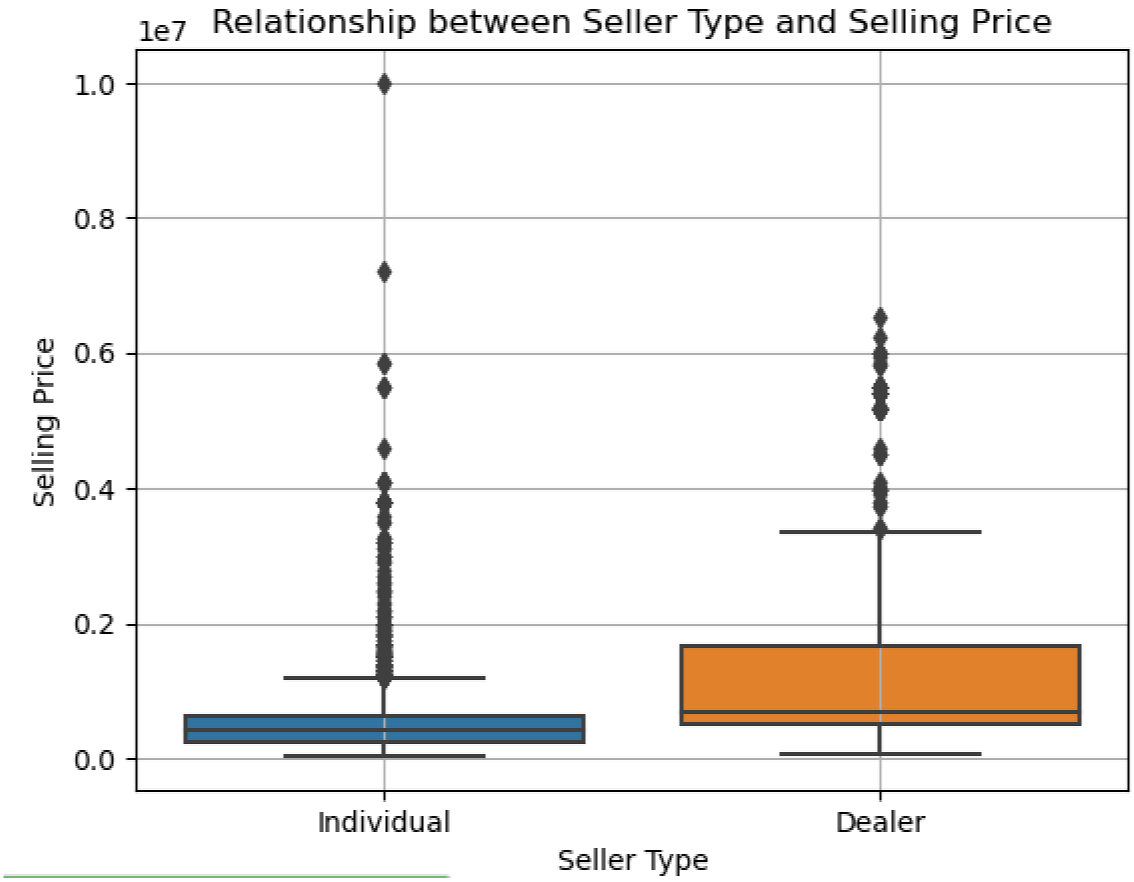
Overall, the cars with the more powerful engines tend to be sold by first owner individuals with petrol fuel and 5 seats. 410,000 is the mode selling price for these cars.



	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage(km/ltr/kg)	engine	max_power	seats
170	2017	10000000	30000	Petrol	Individual	Automatic	First Owner	42.00	1969.0	400.0	4.0
148	2017	5850000	47000	Diesel	Dealer	Automatic	First Owner	13.50	2987.0	282.0	5.0
134	2017	4100000	17000	Petrol	Individual	Automatic	First Owner	9.50	3604.0	280.0	5.0
1564	2017	4100000	17000	Petrol	Individual	Automatic	First Owner	9.50	3604.0	280.0	5.0
3239	2017	4100000	17000	Petrol	Individual	Automatic	First Owner	9.50	3604.0	280.0	5.0
1860	2017	4100000	17000	Petrol	Individual	Automatic	First Owner	9.50	3604.0	280.0	5.0
5248	2017	4100000	17000	Petrol	Individual	Automatic	First Owner	9.50	3604.0	280.0	5.0
7703	2017	4100000	17000	Petrol	Individual	Automatic	First Owner	9.50	3604.0	280.0	5.0
5962	2009	1000000	80000	Petrol	Individual	Automatic	Third Owner	10.93	3498.0	272.0	5.0
5823	2016	3500000	35000	Diesel	Individual	Automatic	First Owner	14.74	2993.0	270.9	5.0

Question #10: How does seller_type affect selling_price?

Both Individual sellers and Dealers have roughly the same median selling price with the Dealer’s median just slightly higher than the Individual’s selling price. Individuals have a wider range of selling prices and many outliers indicating that there is not as much consistency in how these cars are priced. Dealer prices have more consistency, but the box (interquartile) and upper whisker show that higher prices are more common when purchasing from a dealer.


Relationship between Seller Type and Selling Price
1e7
Selling Price
Individual
Dealer
Seller Type