Question 3

The exploratory analysis conducted on the given dataset included utilizing the various methods of pandas, numpy, sklearn, matplotlib, and so on.

I first initialized the columns with date values like dateCrawled as 0 as they can’t be correlated or compared with.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_csv("autos.csv")
df['dateCreated'] = 0
df['dateCrawled']=0
df['lastSeen']=0
```

Alternatively, the unnecessary columns could simply be dropped.

```
df.drop(['dateCrawled'])
```

I then performed a correlation in the CSV file itself and determined that price, registration and kilometer columns have an almost negligible correlation value. I essentially carried out linear regression in this segment.

I also checked the correlation values using pandas and determined the same values:

```
print(df['price'].corr(df['yearOfRegistration']))
Output: -0.00018233084415959184
```

As the ‘yearOfRegistration’, ‘price’, and ‘kilometer’ did not have a substantial value for correlation, it can be assumed that the values are independent of one another. Therefore, these columns and their associated values can be utilized for training the particular model. Other column values that could potentially play a part in determining the value of the used cars are ‘notRepaired’ and ‘powerPS’

During the feature selection process, there were some challenges encountered. I had anticipated the presence of multiple cells corresponding to a particluar column being empty such as ‘gearbox’ and ‘notRepaired’.

One challenge I had encountered was the ValueError with respect to the date and time values. I found a temporary solution by initializing the column values to 0, as aforementioned.

In addition, I had not anticipated but eventually encountered the presence of special characters within the ‘name’ file such as underscore(_), ellipsis(…), and so on.

Bonus: As the objective is to provide accurate quotations to customers on the price to offer for the purchase of their used cars, we need to make use of a regression algorithm. The target value (the price) is continuous in nature.

Thus, the model can be utilized for training is the Lasso Regression Model.

Lasso Regression Model is essentially Divide and Conquer but for datasets (only on the basis of dividing, not conquering per se).

It works on the principle of shrinkage and also penalizes the cost function to mitigate more extreme values.

Lastly, this model is particularly useful when there are only a small amount of inputs that should be considered for the model training.