Data Cleaning:

* The dataset includes some features that don’t necessarily provide any useful information, or have very less amount of data (with a lot of missing values). ‘notRepairedDamage’ had many NaN values (~20%) and ‘nrOfPictures’ column had zero variance and all of the values in this column were zeroes. Hence both these columns were dropped.
* After dropping the columns, all the records that have null values in any of the columns were dropped.
* Since there is a discrepancy with the ‘yearOfRegistration’ column, which has values greater than the current year and before 1900s. All such rows were dropped. If the horsepower (powerPS) was less than 1 and greater than 2400, those records were dropped since it is not possible to have 0hp and over 2400hp for a vehicle. And if the price was below 500 dollars and above 4,00,000 dollars, those rows were dropped as well.
* Some rows had monthOfRegistration = 0, and these values were replaced with the most frequent month in this dataset.
* Then all the duplicate rows were deleted and there were 280557 rows and 18 columns that are useful and accurate for the analysis.

Analysis 1:

1. The resultant pandas dataframe is saved as csv file as ‘Ger\_autos\_cleaned.csv’.
2. The 'year of registration' feature seems to roughly follow normal distribution. With more data it would ideally follow normal distribution.
3. When the rows are grouped with respect to ‘vehicleType’ column and their average prices are listed, SUVs and coupe are expensive and kleinwagen is the most affordable type compared to other vehicle types
4. The availability of limousine vehicles is at a maximum whereas the number of coupes and SUV vehicle types are comparatively less
5. The correlation coefficient between price and kilometers is -0.33.
   * The negative sign indicates that there is negative correlation i.e., as kilometers increases price decreases (inversely proportional)
   * But the strength of correlation is weak, meaning the effect on price as kilometers increases is quite low

Analysis 2:

* Volkswagen cars are predominantly available on the site with about 59,000 listings.
* Automatic gear cars are more expensive than manual cars and SUVs and coupes are the most expensive in both categories.
* Most of the sellers are private (about 99%)

Analysis 3:

* To reduce the memory occupied by the dataset, we could do the following:

1. downcasting- find the range of values in numerical columns and reduce int64 and float64 to sub-datatypes that occupy lesser space
2. drop columns that have zero variance or are unnecessary for our analysis.

* Among all fuel types, automatic cars are more expensive than manual cars except for hybrid fuel type. Hybrid fuel cars also happen to be more expensive than other fuel types. LPG cars are more affordable.
* Automatic cars also have higher horsepower compared to manual cars. Kleinwagens have least horsepower in both automatic and manual types and coupes have the highest hp
* Most of the cars in all brands lie below 10000 dollars and Porshe’s coupe is the most expensive in this dataset. Jaguar and Porshe have some of the most expensive cars in the dataset.