CASE STUDY 051
[Python]
Car Prices Linear Regression Model

Here are some clues in case you are stuck with the case study:

1. The dataset has no column names. You have to use the function read_csv passing the following parameters:
   ```
   header=-1, names = [list_of_column_names]
   ```
   The column names you can discover from the info file describing the dataset (automobile.txt)

2. The missing data in this dataset has the symbol '?'. So you have to specify the parameter na_values when you call the read_csv function, like:
   ```
   na_values = '?'
   ```

3. To check if a column contains a null value, you can use:
   ```
   df.num_of_doors.isnull().any()
   ```
   To check all the dataset at once:
   ```
   df.isnull().any()
   ```

4. To check all the levels of a categorical variable, you can use the function unique, like the example below:
   ```
   df.fuel_type.unique()
   ```

5. To remove a column, you should use the drop() method of a pandas dataframe, specifying the axis = 1.

6. To change the value of a specific information in a cell (cell = row;column), you can use the function loc, like the example below:
   ```
   df.loc[index,'column_name'] = new_value
   ```

7. To encode the categorical variables, you can use the pandas.get_dummies, like:
   ```
   df = pd.get_dummies(df, columns = cat_columns, drop_first=True)
   ```

8. To split the data between train and test, you have to use the module train_test_split from scikit learn
   ```
   from sklearn.model_selection import train_test_split
   ```

9. To split, the code is:
   ```
   train, test = train_test_split(df, test_size=0.2)
   ```

10. To import the linear model module from scikit learn
    ```
    from sklearn import linear_model
    ```

11. To import the modules mean_squared_error and r2_score from scikit learn

```
from sklearn.metrics import mean_squared_error, r2_score
```