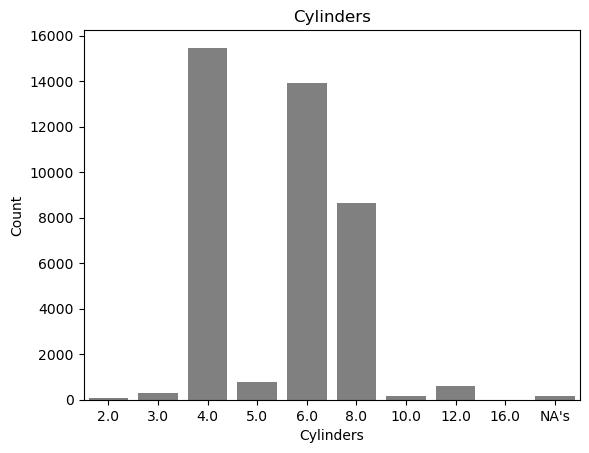
1. In replicating the data analysis as provided in the specimen and to enhance the reliability of plots and correlations, the data was cleaned that involved excluding the values that were not available. For instance, consider the numeric variable atvType, which represents the type of alternative fuel or advanced technology vehicle. In this case, 38248 (91%) of the vehicles were categorized as “Not Available”. These values were removed to observe the influence of various fuel types on vehicles over period.

An interesting aspect that was observed during the data analysis revealed a negative correlation between cylinders and UCity, unadjusted MPG. The observed -0.67 correlation underscores that the bigger vehicles tend to achieve fewer miles per gallon, highlighting an inverse relationship between number of cylinders and fuel efficiency. However, when compared different cylinders over period, vehicles with 4, 6, and 8 cylinders were getting slightly better on UCity.



1. The choice for a regression model is based on the continuous nature of the dependent variable, UCity. Regression models are preferred in this given context because they are designed to predict numeric values, making them well suited for these scenarios where the outcome is a quantity that can vary over a range.
2. The variables that shouldn’t be considered as inputs to the model are:
   1. **chargeXXXX:** Data is absent for all three charge variables—charge120, charge240, and charge240b, and no insights were provided with UCity.
   2. **range:** A total of 39913 vehicles don’t have any range information. While the remaining 168 vehicles have a strong correlation of 0.60 with UCity, it isn’t reliable because the distribution is heteroskedastic.
   3. **engId:** It is just an index variable and has no use for prediction.
   4. **evMotor:** A total of 38345 vehicles don’t have any evMotor information, and only one specific motor is outperforming UCity measure. Thus, this variable in the model can lead to overfitting.
3. To avoid underfitting/overfitting variables such as evMotor should be avoided because one specific motor is outperforming UCity measure, and this could lead to overfitting for the model. Additionally, to evaluate the model on underfitting/overfitting a few techniques can be employed such as train-test split, cross validation, early stopping, etc.