Data activity – Car emissions data set

The following data activity uses the car emissions data set emissions.csv, which you will have recently received in a separate email. This data set is a cut-down (and slightly modified) version of the csv file available at https://datahub.io/dataset/car-fuel-consumptions-and-emissions; please note that all anomalies and errors in the original data set have deliberately been retained.

Please attempt all of the questions below, and neatly record all of the code you use to answer each question. Feel free to write your answers either as comments within your code, or in a separate text file. At the end of the activity, you will need to promptly email back all of your work (including your answers, code and plots) for subsequent discussion.

There are a total of 14 questions for you to consider, followed by an extension. To help get you started, the numerical answers to questions 1 and 2 have been provided. Good luck!

Question 1
How many rows and columns are there in the data set? (Answer: 45511 rows and 23 columns.)

Question 2
How many rows of the data set feature cars manufactured by Lamborghini? (Answer: 103 rows.)

Question 3
Which name appears most often in the *manufacturer* column of the data?

Question 4
Which manufacturer's name appears in the data with two different spellings (differing only by case)? Explain how you might fix this problem, giving a suitable command.

Question 5
Four values in the *noise_level* column are clearly erroneous. What row numbers do these errors occur on?

Question 6
Generate a histogram showing the data in the *noise_level* column, excluding the outliers you identified above, and output this histogram as a pdf file.

Question 7
The *extra_urban_metric* column also contains two highly anomalous values. What is the *model* of car associated with these outliers?

Question 8
Estimate (roughly) a more reasonable value of *extra_urban_metric* for the anomalous data points in question 7, by examining the readings for other cars of the same model.

Question 9
Based on the data, what is the relationship between the columns *urban_metric*, *extra_urban_metric* and *combined_metric*? Using this relationship, can you improve upon your answer to question 8?

Question 10
Describe the relationship between *urban_metric* and *urban_imperial*. Can you explain this relationship?

Question 11
How many entries in the *description* column contain the text "sunroof"? (Hint: The answer is a number between 30 and 40.)

Question 12
Which manufacturer's cars have the highest average noise level, looking across the whole data set?

Question 13
Analyse the relationship between engine capacity and CO_2 emissions. Generate two or more plots to illustrate this relationship, and store them within a single pdf file. How much CO_2 would you expect to be emitted by a car with an engine capacity of 7500? How confident are you in this estimate?

Question 14
What problem do you notice in the final column of the data (labelled *particulates_emissions*), and how would you go about fixing it?

Extension
(If you reach this extension, please record the total amount of time spent on questions 1-14 above, and the total amount of time you spend on this extension.)
Following on from question 13, which other variables in the data set are most relevant in predicting a car's CO_2 emissions? How about other kinds of emissions? (When answering these questions, feel free to apply any mathematical or statistical techniques you deem to be appropriate.)
