Extra Spark Questions

Dataset:

All the required datasets are available in the same folder

Questions

Question No	Question	Dataset Name
1	Assume that you are reading a book and finding out the number of unique words in the book. Yes we guessed it right, you are trying to do the same word count problem using Spark. Need to deal with unstructured data and I have specified the way to deal with unstructured data in class using RDD. Can you just think if you could perform the same task using RDD? Note : you can make use of SQL functions like explode, split and lower	Book.txt
2	Find the minimum and maximum temperature of the given dataset using Dataframes and we have done the same using RDD earlier. Note: Explore doing this using sparksession.read	1800.csv
3	Find out the total spent by the customer with dataframes. Note: load the data as a dataframe with a schema and then try to obtain the result by performing groupby and summing	Customer orders.csv
4	Let us try something little complicated and interesting in this exercise. The objective is to find the to find the most popular superhero, in the Marvel Superhero Universe. So the input data format looks like this.There's a Marvel-graph.txt file,that is just a list of numbers. And these numbers represent IDs associated with each superhero. So the first number in the graph is the superhero we're talking about. And it's followed by the superhero IDs of every superhero, that hero has appeared	Marvel graph.txt MarvelNames.txt

	with, in other comic books. Now, one little nitpicky detail here, is that a hero may actually span multiple lines. So for really popular heroes like Spider-Man or something, their data might span multiple lines. So you might see Spider-Man superhero ID appearing first on more than one line, so our code will have to take that into consideration. And remember that each superhero ID in a line is not necessarily unique, we might need to combine together multiple lines to get the complete picture for a given hero. We also have a Marvel-names.txt file that maps the superhero IDs to human readable names. So for example, Spider-Man slash Peter Parker, is superhero ID 5306.	
5.	Objective to is to predict revenue based on page speed using a linear model	Regression.txt