Exercice Spark 2

Nous recevons un fichier CSV d'un concessionnaire automobile contenant des informations sur les achats de voitures. Le fichier contenant des données historiques sur plusieurs années, il est possible qu'un client apparaisse plusieurs fois dans le fichier.
Les colonnes dans ce fichier CSV, séparées par des virgules, sont, dans l'ordre :

id_client, producteur_voiture, couleur_voiture, modele_voiture, prix_voiture

Par exemple, la ligne suivante :

1234567,Renault,rouge,Clio,15000

signifie que le client 1234567 a acheté une Renault Clio rouge qu'il a payé 15000 euros.

Nous souhaitons connaître la liste des modèles de voitures Renault rouges dont le prix est supérieur à 15000 euros achetés par chaque client.

Question 1 : Proposez un algorithme en Spark pour résoudre le problème donné. La première ligne du fichier CSV contient les entêtes. Précisez si les transformations que vous utilisez sont *wide* ou *narrow*.

Coup de pouce : une fois que vous avez créé un RDD *r* à partir du fichier d'entrée, vous pouvez obtenir les entêtes par l'instruction suivante : *r.first()*. On pourra donc utiliser cette information pour enlever les entêtes du RDD.

Question 2 : Résolvez le même problème en SparkSQL. La première ligne du fichier CSV contient les entêtes.

Coup de pouce : l'instruction pour charger les données dans un *Dataframe*, en incluant les entêtes, est la suivante :

```
cars =
sql_ctx.read.format("csv").option("header","true").load("./cars.csv")
```