**CSDA 1020**

**Big Data Analytics Tools**

**Project 2 – Apache Spark SQL**

**Group 1:**

Syed Multani, Student ID: 306282

Alvin Lee, Student ID: 306271

Andrew Dugal, Student ID: 305561

**Introduction and Creating table**

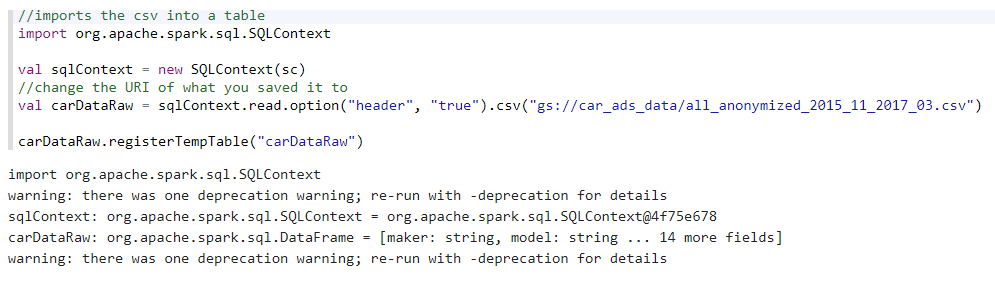
***Data and initial inspection:***

Similar to our first project, we accessed our data from <https://www.kaggle.com/mirosval/personal-cars-classifieds>. Data for this set was compiled from several websites across Czech Republic and Germany. It contains 16 columns and approximately 3.5 million rows. Data is uncleaned - showing some null values and irregularities that we will need to clean out when we create our tables in Apache Hive in order to present more accurate analysis.

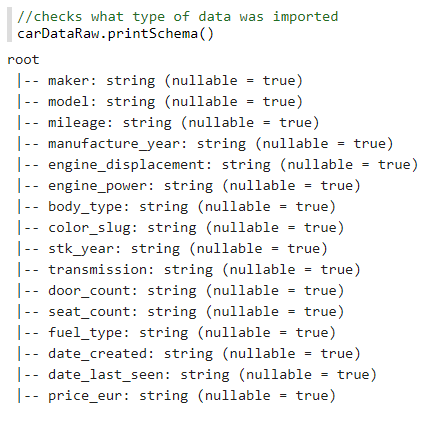
For this project we will be doing analysis using Apache Spark SQL – utilizing Apache Zeppelin interpreter notebook.

***Loading data to Zeppelin and cleaning data:***

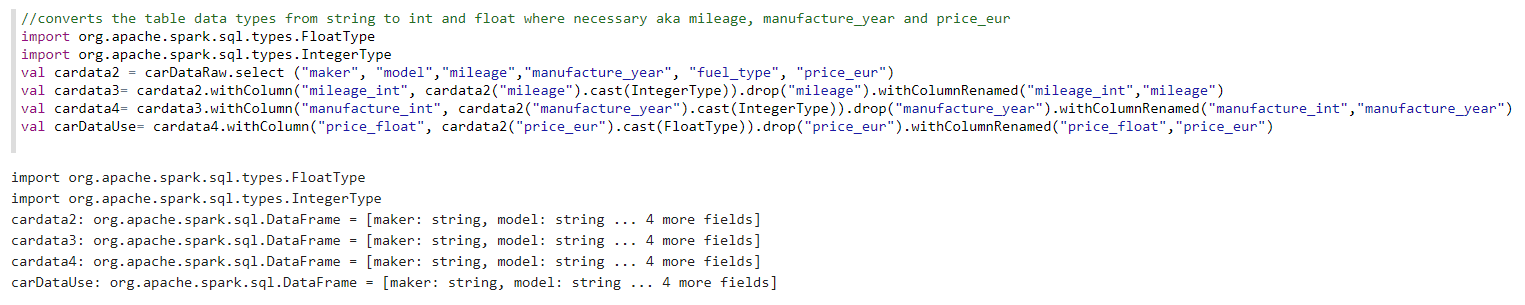
Our first step was to import the dataset into Zeppelin from our bucket by using the following code:



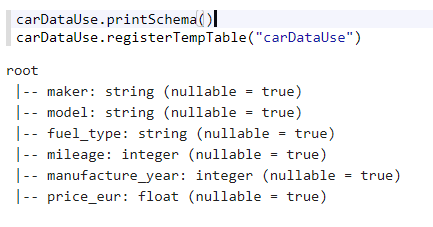
As you can see in the above question- a warning was issued showing us that datatype for all columns was imported as string. We ran the following code to check the types of data that was imported:



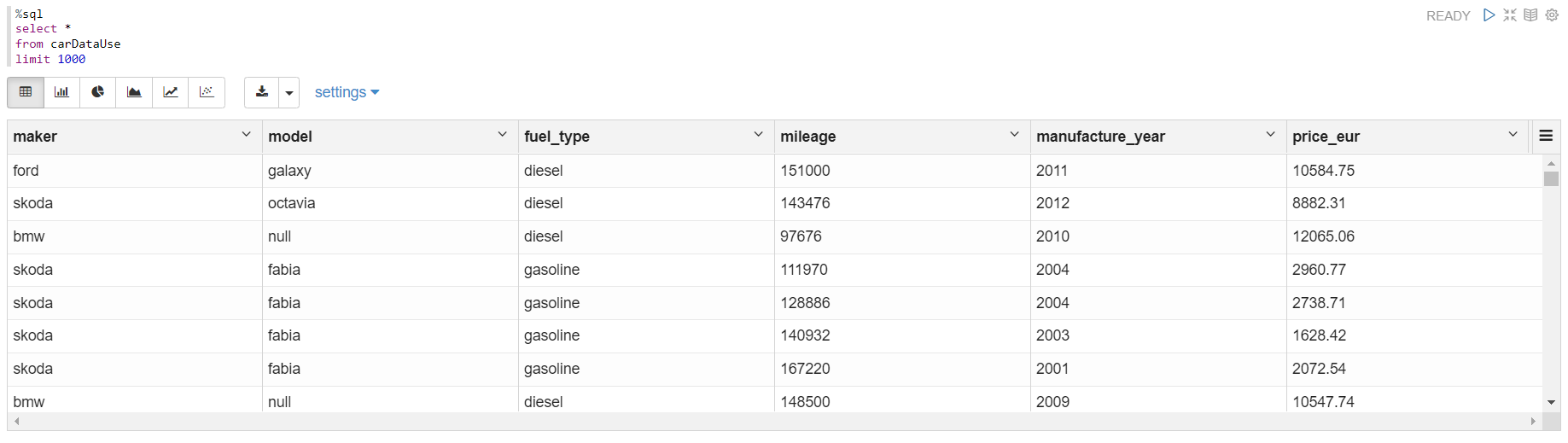
As you can see- all data was indeed imported as string type. We ran the following code to convert columns that weren’t string type, to their actual data type (integer, float):

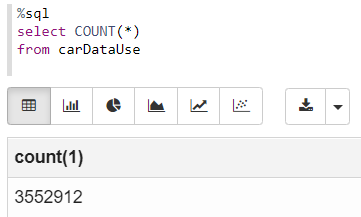


Our next step was to run the following code to check and make sure that the data types were properly converted. As you can see, column data types were effectively converted to their proper types.



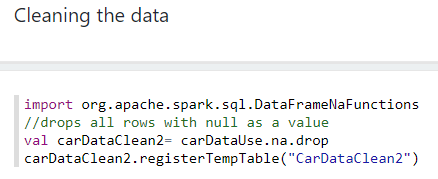
Since we were able to convert and check our data- we ran our table with the top 1000 entries. Below is a screen shot of the code and first several rows and another showing that all dataset entries were pulled into our table:



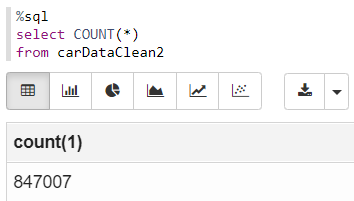


***Cleaning data:***

After we had finished importing and converting our data- we cleaned our data by removing all null variables (code and screenshot below). Once our data was cleaned and ready to go- with 84707 entries (as you will see below), we were ready to answer our analysis questions.



Entries left after cleaning:



**Questions and Analysis**

***Question 1: What is the relationship between car makers, models and price?***

For this question we used the maker, model, manufacturer\_year, and avg(price\_eur) columns to explore the relationship between makers, models and price. As you will see in our screenshot (below), we isolated the years 2005-2018 as we believed that would be the most likely the age range most people would be looking at when shopping for a used vehicle. We also isolated the price from 8000 to 100000 Euros. This price range cleaned out some of the outlier and irregular numbers that we were seeing on the high and low end of our dataset. For instance, there were multiple thousands of entries at 1295 Euros that would have skewed our results, and there were some entries in the millions of dollars at the top that would have skewed our results. We determined as a group that none of those low or high entries would be relevant to the majority of individuals and families looking for a reliable used vehicle, and most were likely incorrect anyway.

As far as the relationship between makers, models and price is concerned, each manufacturer has multiple models with obviously no models being used or made by multiple manufacturers. Initially when we looked at the dataset, price looked to be partially dependent on multiple factors with fuel type being a main factor. The reason we did not include fuel type, however, in this question is because many of the values for price that we all listed at 1295 Euros were for vehicles that ran on electricity, propane, or natural gas. They would have severely skewed the results for not only this question but for further analysis questions

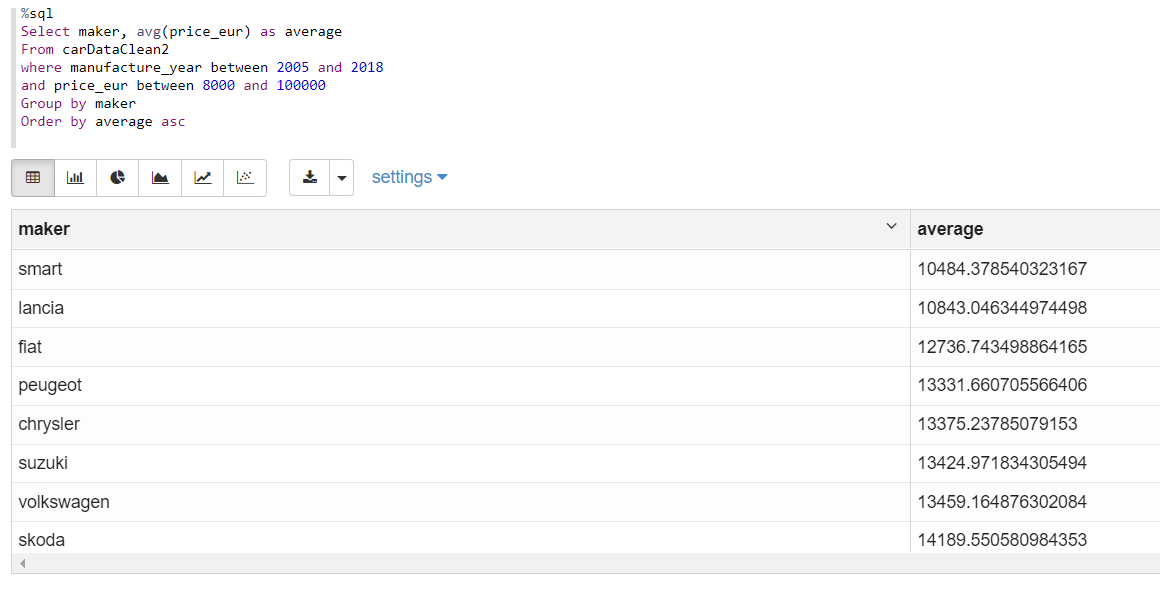
Higher end makers (such as BMW, Bentley, Porsche), larger vehicles (trucks, vans, full sized SUVs, and newer cars all more expensive. Older, smaller compact vehicles, (and initially when we explored the data prior to cleaning which removed a lot of the irregular prices) vehicles that use compressed natural gas, propane, or electricity for fuel are cheaper on average- outside of Tesla, which fit more into the higher end luxury class, which is an outlier compared to other electric vehicles.

See below for our code and table related to this question:



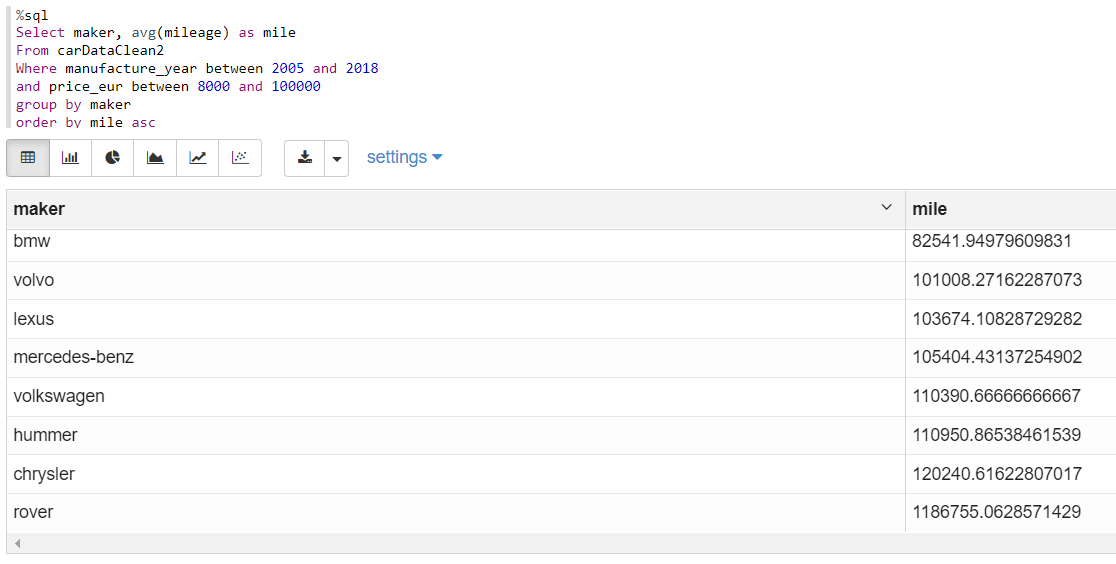
***Question 2: What are the top five vehicle manufacturers would you recommend? Why?***

Below we will present a couple of different options for manufacturer recommendation, each using different variables. The first option (code and results pictured below) lists the top manufacturers by the average price across all of their vehicles manufactured between 2005-2018 with cost above 8000 Euro and below 100000 Euro. As you can see the top 5 vehicle manufactures by average price would be Smart, Lancia, Fiat, Peugeot, and Chrysler. This gives good insight into which brands could provide the best value depending on the customers preferred price point- and further evaluation could be done to look at things like mileage, maintenance records, etc.



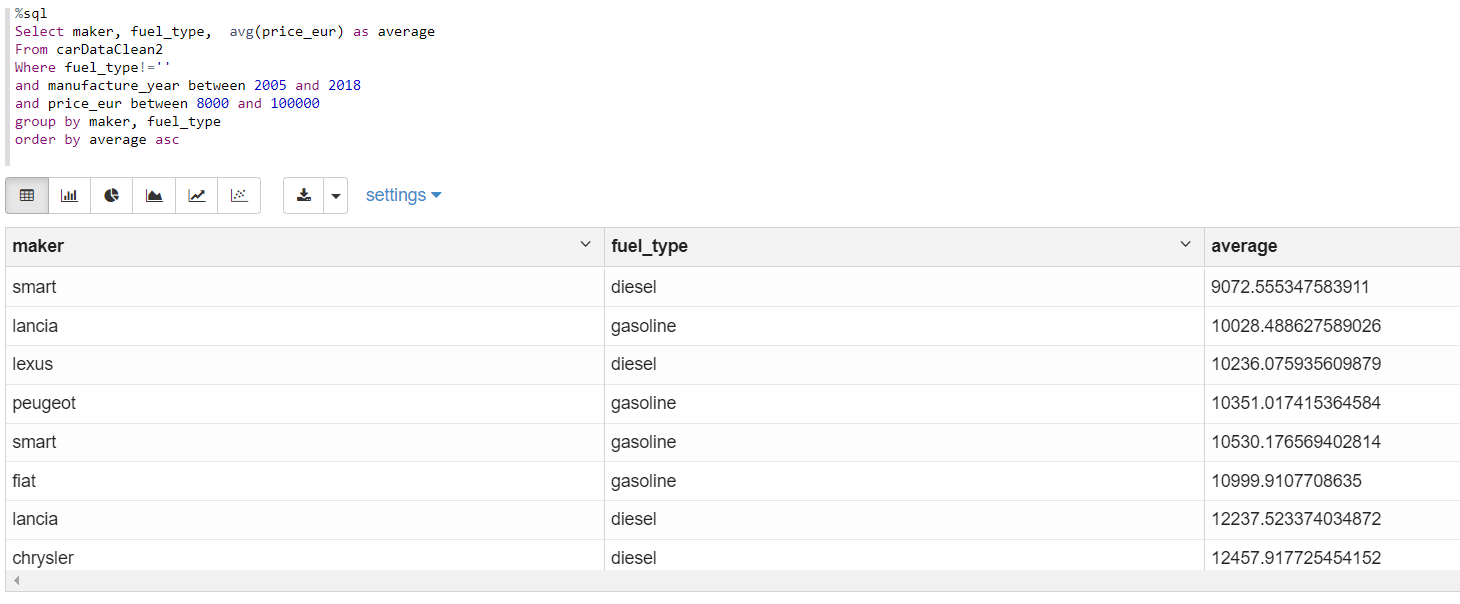
The second option we looked at (code and screenshot below) was the average mileage broken down by the maker across their entire line in the dataset. Once again we looked at years 2005-2018, and price above 8000 Euro and below 100000 Euro. Mileage and longevity is something we all agreed would be something of particular value to customers when buying a used car for several reasons- they would want to know how much mileage the car has to indicate how much wear and tear is on the vehicle. They would also use mileage to value how much usage they would still get out of a the vehicle for the money they are investing into it. For this reason we looked at the bottom of the results for the brands that do the most mileage. It’s no surprise that the top 5 brands are Rover, Chrysler, Hummer, VolksWagen, and Mercedes-Benz. These are all higher end luxury type brands that make bigger vehicles (sedans, wagons, trucks, SUVs) and tend to have more longevity.

One interesting observation we noticed was that Chrysler is in top 5 for both average price, and mileage. Due to that fact we would definitely be recommending customers take particular look at Chrysler as they seem to be providing good value across multiple points of analysis.



***Question 3: Does fuel type have any impact on the car price? Explain.***

For this question we grouped by maker (which includes all models across their line-up) and fuel type, and ordered that by the average price of the vehicle. Code and results are shown below. Once again, we broke it down by the same year and price outline that we did for all other queries. As you will notice by looking at the top results – comparing the makers’ gas and diesel models doesn’t provide us with a consistent trend. For instance, on average, Smart is cheaper as diesel powered than gasoline and Lancia is cheaper as gasoline than diesel. For this reason, we decided to take a different simplified approach to this question (continued on next page).



Since our first query resulted in an inconsistent picture- we simplified our approach and using the same year and price guidelines, and explored the entire dataset grouping the average price of both fuel types (see below). As you can see, when looking across our dataset- on average, diesel is more expensive than gasoline by a fairly significant margin. As a reminder, since the other fuel types- electric, propane, and natural gas all had irregularities and inconsistencies in their data points those values were basically cleaned out. It would have been interesting to have had reliable numbers to work with for those fuel types.

As our results indicate, the answer here is that yes, on average fuel type does have an impact on the price of the car with diesel being more expensive on compared to gasoline powered vehicles. Some potential observational reasons behind this would be that many of the diesel powered vehicles are larger, heavier duty types such as trucks and SUVs. Another potential reason is that many higher end luxury cars offer a turbo diesel option on some of their larger models, this is common amongst Mercedes, BMW, and VW, which would drive the cost of the car up. Diesel powered engines tend to handle more miles and hold value better than gasoline as well. So those are just a few potential reasons for the difference.

