An Investigation on the influencing Factors of the Price of Used Cars in the US

Carolin Poschen - Project - 736-03 Big Data Engineering

Introduction

When buying a first car, the preferred option often is to buy a used car. This is particularly the case for students and graduates that may not have the driving experience or financial backing to buy a new car. But even when buying a used car, many things should be considered, with the price being only one of them. This project looks into the influencing factors of the price of used cars in the US, utilizing Big Data technologies like *PySpark* and *AWS EMR*[1].

Section 2 describes the data used in this project, as well as the data preparation and what will be analyzed. Section 3 provides analysis results and this paper finishes with a reflection of the project and the methods used in section 4.

Method

Data Set, PySpark, AWS EMR

The data set used in this project is the *US Used Cars Data set*[2] from the Data Science Platform *Kaggle*. The data set was created by automatically collecting real data of used cars for sell in the US by the Kaggle User *AnanayMital*. It contains information on 66 different variables for 3,000,000 used cars. The variables include information about the car type, size and color, the engine type and power and when and where the car was listed at what price.

Out of the three V's of Big Data, *Volume*, *Velocity* and *Variety*, this data set satisfies the criteria for *Volume*, as it is a rather large data set. While this data set is mostly

[1] https://aws.amazon.com/emr/

[2] https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset

structured data, provided in a *.csv*-format, the data also satisfies the criteria for *Variety*, as it contains structured data like numeric or categorical values, and semi-structured and unstructured data in form of lists and text.

The data preparation and analysis is done using *PySpark* with an *AWS EMR* cluster. The EMR cluster is deployed using a *AWS CloudFormation*[3] template. The data set in its unprocessed form is single 10GB *.csv*-file, so the AWS EMR cluster is a sensible pick for a computing cluster, as it was designed to process large files. Similarly, PySpark is a collaboration between Apache Spark and Python, providing the functionality of SQL, Dataframes and datasets, which makes it a good tool to process large amounts of data. The visualization is done on a subset of the data using Databricks.

Data Collection and Preparation

The data set used in this project is stored in an *AWS S3*[4] bucket, from where it is read to the EMR cluster with PySpark.

When reading the data as a DataFrame, all data is read in a *String*-format, so the first step in the data preparation is to cast all non-String columns to the correct data type and ensure that only valid values are in each column, using regular expressions or lists of valid values. If a column has too many missing values (here ¿70%), the column is dropped, as it does not contain enough data. Additionally, for some numeric columns, categorical columns are introduced to better compare cars to each other. Afterwards, a new DataFrame is created, containing only the columns needed to answer a specific research question. This allows to select data more targeted towards a particular question and to filter records per question based on their values. This filtering includes dropping rows with more than a certain number of missing values (here 15%) and dropping rows where the main predicting values are null. From the initial 3 million records with 66 predictors, this leads to five DataFrames with the sizes listed in Table 1.

[3] https://aws.amazon.com/cloudformation

[4] https://aws.amazon.com/s3/

Table 1: DataFrame Sizes

	Mileage	City	Time	Color	Accident or Fleet
Records	1720802	1777625	1751024	1774922	1001791
Columns	8	9	9	10	10

Results

The project analyses different variables in a used car data set and if there is an influence between the variables value and the price of a used car. To analyze that, part of the data preprocessing was to create a DataFrame for each analysis question and select a subset of variables of interest for this DataFrame.

For each analysis question, like looking at the relationship between the mileage of a car and the price, or in which city the car is up for sell, first, all rows are analyzed and an overall result is produced and then this project looks at different categories for each question like the body type of a car, the horsepower or the car size. This provides information about all available data as well as if there is a difference for different car types.

Figure 1 shows the average price of a subset of all used cars over the time that the data was collected. There is a large increase in the average price just for the year 2017, which should be invested more in the future. It might be just the random subset of data that this graph is based on or an actual price increase in this year.

Figure 2 shows the average price for used cars if they have or have not had an accident before or if they used to be part of a fleet. While it was expected, that cars that have not been in an accident before are more expensive, cars that were not part of a fleet also are on average more expensive than cars that were not part of a fleet.

Discussion and Reflection

This project utilizes Big Data Technologies to analyze a data set on used cars in the US and investigate different factors influencing the price.

Challenges encountered in this project include the data collection. The current project reads the data from an S3 bucket, an improvement would be to read the data

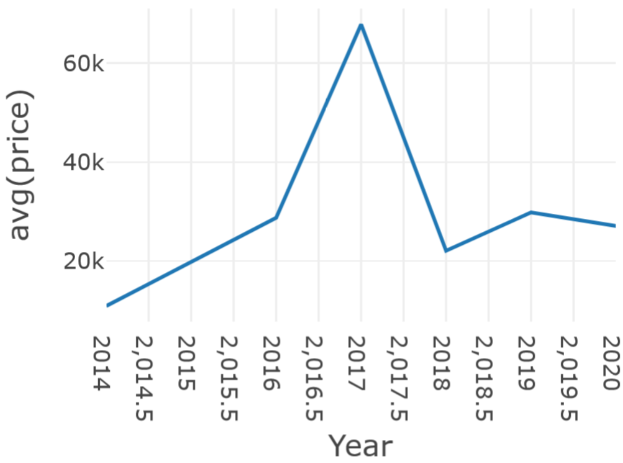

Figure 1: Average Price of Used Cars over Time

from Kaggle directly instead of requiring a user to upload it to S3 first. When reading the data, there seems to be a problem with some rows, resulting in many more invalid or null values than the Kaggle statistics let to believe, resulting in a lot more data preparation necessary to ensure the data is valid and clean. Once the data preparation was complete, the analysis itself was rather straight forward, but presenting the analysis results was a struggle, because there are so many different categories to present and so much data has to be presented in a conscious way.

For better comparison and analysis, the different categories should be better defined and more clearly separated based on more variables to get better results. The project right now only analyses few variables with few categories, but for better details, the analysis should include more variables and finer categories.

For the programming aspect of the project, while right now the code is in separate files by topic and contains comments to explain the code, each file needs to be run individually and the code only contains function calls instead of user defined functions for repeated actions. An improvement to the project would be to have an additional script that takes care of running all the preprocessing and analyzing scripts instead of

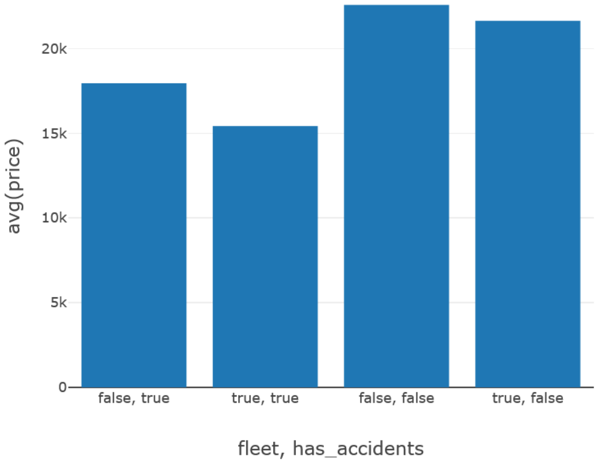

Figure 2: Average Price of Used Cars With/ Without and Accident or belonging to a Fleet

having the user run everything manually.

The biggest challenge, that I did not master was to create the visualizations from the complete data set. My idea was to connect the S3 bucket to Tableau, as I worked with Tableau before and connecting S3 as a data source seemed like a doable challenge. The solution I used for this project to create the visualizations is to use a subset of the data and create the visualizations in Databricks. The complete file was too big to use Databricks, so this was not an option, but in the future, I would use a different approach to visualize my data.