# US Used Cars Data Analysis

CIS 5200: System Analysis and Design

Authors: Aakriti Rupal, Anish Omprakash Pandey, Phue Thant, Vatsal Pancholi
**Department of Information Systems, California State University, Los Angeles**
arupal@calstatela.edu, apandey9@calstatela.edu, pthant@calstatela.edu, vpancho@calstatela.edu

## 1. Abstract

US is one of the largest markets for Used cars. Price of a used car is a complex function of many factors/variables (car attributes). Historically, the price of a used car has been a subjective evaluation of different car attributes. The work aims to utilize a mathematical approach to understand the effect of different car attributes (make/model, mileage, fuel economy etc.). The paper covers the analysis performed to understand the effect of different factors that affect the price of used cars. The dataset for the analysis is for the US used cars was obtained by screen scraping the data available in a popular website CarGurus that offer services to compare the price of used or new cars in the US. The dataset (scraped content) is available for download on Kaggle[1]
Large dataset (3M rows, 66 car attributes) was used as they are more resistant to bias for certain factors and helpful for complex analysis. Descriptive and statistical analysis using HiveQL was conducted for the different factors/variables (car attributes) using HiveQL. Patterns and trends in data was extracted using visual analytics conducted using Excel and Power BI visualization tools.

## 2. Introduction

The project is focused on analyzing different factors that affect prices of used cars. The project was particularly interesting as it involved working with an extremely large data set. Project was executed by utilizing the concepts of Hadoop, HIVE, and data analytics to gain insights as the traditional excel/ tableau tools would not have scaled for the large dataset we used. The existing analysis can be extended to predict/ validate car prices using various parameters by utilizing unsupervised and supervised machine learning concepts
First step in the analysis included identifying the fields from the dataset that can be critical for the analysis e.g., (mileage, car_make, car_franchise, vehicle_year, days_on_market, etc.). Thereafter, the analysis was segmented into data validation, descriptive analysis and analysis using a statistical approach. Furthermore, visualizations using excel 3D map and power BI were also included for a quick and easy insight into patterns and trends from the analysis.

## 3. Background and Existing work

Used car data is one of the most common datasets that has been analyzed using a variety of algorithms. Our work is primarily based and builds on 3 existing works detailed in the next section. Our analysis primarily focused on utilizing big data technologies to overcome the drawbacks of analysis performed in the existing work

## 4. Related Work

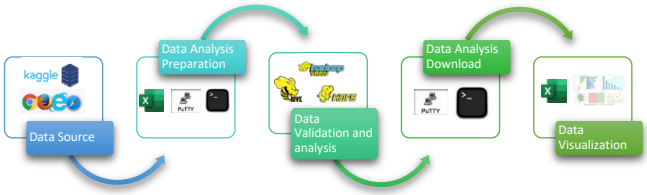
One of the most popular related work is from "Jovian – Data Science and Machine Learning (Sep 8, 2021) – Exploratory Data Analysis of Used Cars in the United States, using Python, Pandas, Seaborn and Plotly" [1] In this study, they have the same dataset as our project and the insights of the used cars sale by region, brand, days on market and average prices of used cars sale are similar to what we obtained during our analysis. However, they have used only 1/3rd of the original data size. The major difference between our analysis and this project is that they have used python visualization libraries to visualize the data while we have utilized big data technologies like HDFS/Hive for our analysis.
Another related work is from "Exploring and Analyzing Used Car Data Set by Irtasam Ali Wains (Nov 19, 2020)" [2]. This study also uses the same dataset and focuses mainly on the different used cars sale trends including car manufactures, type of vehicles, fuel, sales by year and prices with the different car conditions. For the visualization of the data, study utilizes logistic models, violin plots and box plots. The major difference from our analysis is the use of machine learning to predict used car prices
Another related work for used cars analysis is from "Used Cars Price Prediction and Valuation using Data Mining Techniques by Rochester Institute of Technology, Abdulla Alshared (12-2021)." [3] In this study, they have focused on prediction of price for used cars in Dubai. The work uses similar analysis attributes such as car body type, car brand, year of the used car etc. The analysis is done using python. However, the major difference from our project is that while the work uses data mining and machine learning approach this work utilizes Hadoop and HiveQL.

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

## 5. Project and Analysis Workflow

The figure below shows the workflow of our project. Overall following steps were performed for the analysis:

1) Data Source – Data set downloaded from Kaggle and split into small manageable CSVs (100,000 line each) files
2) Data Analysis Preparation - Standardized the data types for consistency across all columns and loaded them to Hadoop file system (HDFS)
3) Data Validation and Analysis – Validated data using select (*) and count(*) HiveQL commands. Group by and Hive mathematical functions (Avg, count etc.) was used for performing analysis,
4) Data analysis download – Download hive results to local computer for data visualization
5) Data Visualization – Utilized bar charts, line graphs and excel 3D maps to represent the analysis of our project.




## 6. Data and System Specifications

The dataset was generated in September 2020 from CarGurus website. It is a real word data with ~ 3 million records in 66 columns.

*Table 1: Data Specifications*

| Metric | Value |
|---|---|
| Data size | 9.98 GB |
| No. of columns | 66 columns |
| No. of records | 2,974,151 records |

System specifications of the AWS cluster that was used for the analysis is as below:

*Table 2 System Specifications*

| Metric | Value |
|---|---|
| Cluster version | Hive 3.1.2-amazn-4 |
| No. of nodes in the cluster | 3 |
| Memory size | 58.58 GB |
| CPU speed | 2.0GHz |
| CPU core | 8 |

## 7. Data Analysis

Used car data was analyzed to determine the effect of different car attributes (car type, engine type, etc.). Following steps were used for the analysis:

**7.1 Data Validation:** This was required to ensure that the data loaded on HDFS is accurate and complete. Basic select commands were used to display few fields from the dataset for data validation. Below screenshot shows results from one of the select commands.

| vin | body_type | make_name | maximum_seating | wheel_system | price |
|---|---|---|---|---|---|
| JF1VA2M67G9829723 | Sedan | Subaru | 5 | AWD | 46995 |
| SALRR2RV0L2433391 | SUV / Crossover | Land Rover | 7 | AWD | 67430 |
| SALCJ2FXXLH862327 | SUV / Crossover | Land Rover | 7 | AWD | 48880 |
| SALYK2EX1LA261711 | SUV / Crossover | Land Rover | 5 | AWD | 66903 |
| 3MZBPABL6KM107908 | Sedan | Mazda | 5 | FWD | 23695 |
| SALYK2EX5LA275434 | SUV / Crossover | Land Rover | 5 | AWD | 68520 |
| SALCJ2FX6LH858128 | SUV / Crossover | Land Rover | 7 | AWD | 51245 |
| SALZL2GX4LH007593 | SUV / Crossover | Land Rover | 5 | AWD | 84399 |
| ZARBAAC41FM129303 | Coupe | Alfa Romeo | 2 | RWD | 97579 |
| SALZJ2FX0LH081763 | SUV / Crossover | Land Rover | 5 | AWD | 51885 |

*Figure 1: Data Validation*

**7.2 Descriptive Data Analysis:** By definition, a descriptive analysis describes the characteristics/ features of data. This analysis was performed as it closely illustrates the relationship between different factors (fuel economy, engine size, car model) that are associated with used cars. Following analysis was performed:

i. Effect of car size engine on fuel economy
   a) Analysis Variables
      - Input variable: Car engine type
      - Output variable: Average fuel economy of the car
   b) Analysis Method: Hive and bar graph visualization
   c) Observations/Findings:
      - Cars with smaller engines have better fuel economy.
      - Hybrid engines have better fuel economy when compared to regular engines.
      - In the used car dataset, I3 hybrid has the most fuel economy.

ii. Number of cars sold based on the car type
   a) Analysis Variables
      - Input variable: car type
      - Output variable: Number of cars sold (vin number)
   b) Analysis Method: Hive and bar graph visualization
   c) Observations/Findings:
      - SUVs followed by sedans are the most preferred used car type.

iii. Number of cars sold based on the car model
   a) Analysis Variables
      - Input variable: car model
      - Output variable: Number of cars sold (vin number)
   b) Analysis Method: Hive, Bar graph
   c) Observations/Findings:
      - Ford F-150 is the highest selling car model.

iv. Average price of the car for different vehicle years.

v. Since the price of the car varies with the car model, we have selected Ford 150 as the car model to determine the effect of vehicle year on average price of the car.
   a) Analysis Variables
      - Input variables: vehicle model (Ford F-150), vehicle year
      - Output variables: Number of cars sold (vin number)
   b) Analysis Method: Hive and Line graph
   c) Observation/Findings:
      - Average price of the car decreases for older car models. Example, for Ford F-150, a 2021 model will cost more a 2015 model.

vi. Distribution of used cars sales across different cities in the US
   a) Analysis Variables
      - Input variable: city
      - Output variable: Number of cars sold (vin number)
   b) Analysis Method: Hive, pie-chart and an excel 3D map
   c) Observations/Findings:
      - Used car sales is highest in Houston, Texas.
      - The geo spatial map shows the relative used car sales in different cities.

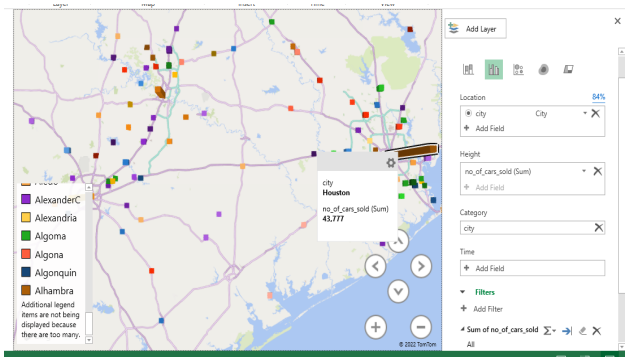



*Figure 2: Distribution of used cars sales across different cities in the US (Excel 3D Map)*

vii. Effect of price of the used car based on car accidents
   a) Analysis Variables
      - Input variable: has accidents
      - Output variable: discount percentage on the car price
   b) Analysis Method: Hive
   c) Observations/Findings:
      - Used cars with an accident history have greater discounts over the cars with no accident records.

viii. Effect of car accidents on days on market
   a) Analysis variables
      - Input variables: has accidents
      - Output variables: average of days on market
   b) Analysis Method: Hive
   c) Observation/Findings:
      - If the car has had an accident, it stays in the market for a longer period before it is sold.

ix. Car sales distribution across the US west coast cities
   a) Analysis Variables
      - Input variables: city
      - Output variables: Number of cars sold (vin number)
   b) Analysis Method: Hive and Power BI filled map
   c) Observation/ Findings:
      - The maximum sales of car are around the SFO and Los Angeles areas.
      - The below geo spatial visualization shows the result.

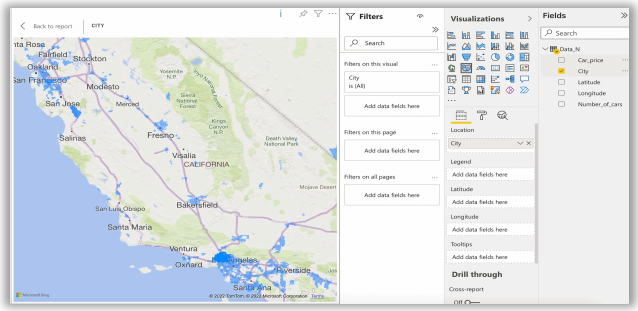



*Figure 3: Car sales distribution across the US west coast cities (Excel Filled Map)*

x. Price distribution of used cars across US cities
   a) Analysis Variables
      - Input variables: city, car price
      - Output variables: Car price distribution across cities
   b) Analysis Method: Hive and Power BI map
   c) Observation/ Findings:
      - The price distribution in a pie-chart across different US cities is shown in a geo spatial map as attached below
      - Different colors depicting different car price ranges can be seen in the pie chart
      - As Houston is one of the highest selling car cities, the price distribution pie-chart at this location is larger than other cities

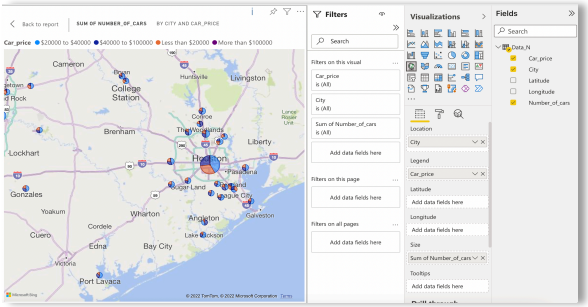



*Figure 4: Price distribution of used cars across US cities (Power BI)*

7.3 **Analysis using statistical approach:** This methodology was followed to select potential attributes affecting car price e.g., car type, has accidents, owner

count, days on market, vehicle year and mileage. Buckets or range of values was created for continuous variables/ factors. To analyze effect of one attribute, all other factors were kept constant and the effect of that attribute on car price was determined. Following are the two factors discussed in detail.

i. Effect of average price of a used car with respect to number of owners
   a) Analysis Variables
      - Input variables: car model (Ford F-150), has accidents (False), mileage (<10000 miles), days on market (0-30 days), vehicle year (5-10 years) and number of owners
      - Output variables: price of the car
   b) Analysis Method: Hive and Bar graph
   c) Observation/Findings:
      - Average price of a used car decreases if the if it has had more owners.
      - Below is the visual bar graph representation of the result.

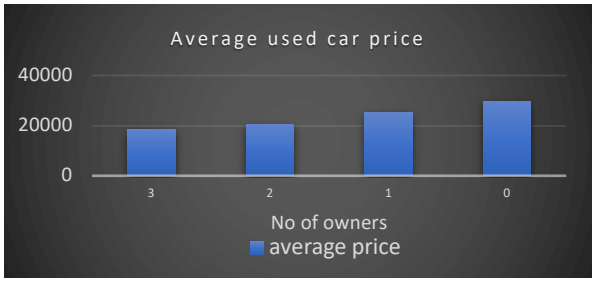



*Figure 5: Effect of average price of a used car with respect to number of owners (Excel Bar Chart)*

i. Effect of average price of a used car with respect to vehicle years
   a) Analysis variables
      - Input variables: car model (Ford F-150), has accidents (False), mileage (<10000 miles), days on market (0-30 days), number of owners (0) and vehicle years[2]
      - Output variable: price of the car
   b) Analysis Method: Hive and Bar graph
   c) Observation/Findings:
      - Average price of a used car decreases for older models.
      - Below is the visual representation of the finding.

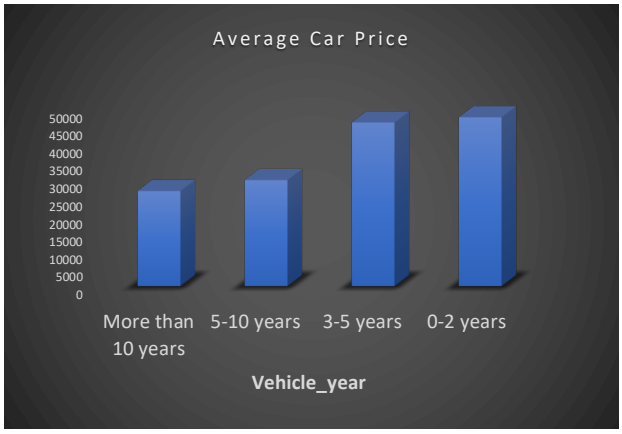



*Figure 6: Effect of average price of a used car with respect to vehicle years (Excel Bar Chart)*

## 8. Conclusion:

In summary, the analysis provides insights into important factors that determine the used cars prices in the US. The work uses a very large dataset (~3 million records) for analysis which would not have been possible without HDFS and HIVE. It also showcases interesting patterns in used cars sales data as it is both extensive and comprehensive (66 attributes analyzed). The analysis used to further develop predictive models using supervised and unsupervised machine learning. All the codes and documentation have been uploaded to GitHub[2] and is open for anyone who wants to utilize it to further their analysis. The project was extremely useful for the team as everyone got hands on experience with working with a large dataset, writing HIVEQL, utilizing visual analytics, and extracting insights from a complex set of attributes.

## 9. References:

[1] H. Gupta, "Exploratory Data Analysis of Used Cars in the United States," 08 09 2021. [Online]. Available: https://blog.jovian.ai/understanding-used-cars-market-in-usa-52489e10d551. [Accessed 22 05 2022].

[2] I. A. Wains, "Exploring and Analyzing Used Car Data Set," 19 11 2020. [Online]. Available: https://medium.com/swlh/exploring-and-analyzing-used-car-data-set-2e2bf1f24d52. [Accessed 22 05 2022].

[3] A. AlShared, "Used Cars Price Prediction and Valuation using Data Mining Techniques," 12 2021. [Online]. Available: https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220&context=theses. [Accessed 22 05 2022].

[2]GitHub URL for code (https://github.com/arupal23/CIS-5200-Project-Files)