Conducting Data Analysis on US Used Cars Dataset from September 2020

Authors: Brian Bao, Karen Alvarez, Natalie Sanchez, Shadman Sayef
Department of Information Systems, California State University
CIS 45601-01 Introduction to Big Data
bbao3@calstatela.edu, kalvar155@calstatela.edu, nsanch184@calstatela.edu, ssayef@calstatela.edu

Abstract: The used car market in the U.S. underwent significant shifts during the COVID-19 pandemic, with disruptions in supply chains and changing consumer behavior affecting sales patterns and inventory composition. This project focuses on analyzing a large-scale dataset of over 3 million used car listings collected from Kaggle in September 2020. Using Hadoop and Hive for data storage and processing, and visualization tools like Excel and Power BI, we investigated trends in pricing, popular brands, exterior colors, and regional patterns. Our pipeline—from raw data ingestion to query execution and export—was designed to handle big data efficiently and produce actionable insights for understanding post-pandemic market dynamics.

1. Introduction

The COVID-19 pandemic brought unexpected shifts to many industries, and the used car market was no exception. With new car production slowed due to supply chain issues, more buyers turned to used vehicles—pushing up demand and altering market dynamics.

To explore these changes, we worked with a large dataset from Kaggle containing over 3 million used car listings from across the U.S. The dataset included key details like make, model, fuel type, mileage, pricing, and accident history, giving us a comprehensive view of the market during this period.

We used Hadoop and Hive to manage and process the data, then built visualizations in Excel and Power BI to highlight trends in pricing, popular brands, colors, and regional inventory. This project helped us understand how real-world events impact consumer behavior and how big data tools can uncover those patterns.

2. Related Work

The COVID-19 pandemic significantly impacted various sectors, and the automotive industry's used car market was no exception. Several studies have examined the pandemic's effects on transportation and related markets that provide a foundation to understand the dynamics observed in the 2020 US used car dataset.

The pandemic's disruption to supply chains, a key factor in the used car market as it affects new car production and thus used car availability, is discussed in the MIT paper, Strategic Transformation Trends within Automobile Supply Chains in the Post-Pandemic Era. This study focuses on how automobile original equipment manufacturers (OEMs) adapted their supply chain strategies to enhance resilience.[1] The paper's insights into increased safety stock, dual-sourcing, and supplier collaboration are relevant to understanding the supply-side constraints that influenced used car pricing and inventory, factors that our analysis of the 2020 dataset will address.

Furthermore, the UK-focused study, Less is more: Changing travel in a post-pandemic society, provides valuable context on how travel behavior changed during the pandemic. The findings on reduced overall travel, decreased car traffic, and the rise of online shopping are relevant to understanding the demand-side dynamics of the used car market.[2]

The Joint Research Centre (JRC) conducted a study, Post-pandemic trends in urban mobility, which offers a European perspective on post-pandemic urban mobility trends. The report's findings on increased car usage, the slow recovery of public transport, and the potential for increased car dependency are particularly relevant to our investigation of changing consumer preferences in the used car market. The JRC report uses data to analyze the pandemic's impact on car usage.[3] Their findings about increased car usage align with the need to investigate inventory trends of different car types (fuel-efficient vs. SUVs/trucks).
In summary, these studies provide a multifaceted understanding of the pandemic's impact on transportation, supply chains, and consumer behavior. Our analysis of the 2020 used car dataset builds upon this existing research by providing a focused examination of how these broader trends manifested in a specific market segment, contributing to a deeper understanding of the automotive industry's resilience and adaptation during this unprecedented period. Our research will extend this body of work by specifically examining the interplay of pricing, sales, and inventory composition within the used car market in 2020, with a particular focus on regional variations and the influence of vehicle characteristics.

3. Specifications

The dataset used in this analysis covers real-world listings for over 3 million used cars from September 2020 and prior. It was obtained from Kaggle and the raw file measures 9.3GB in size uncompressed. The data comes stored in .csv (comma-separated values) format, allowing for easy processing in the Hadoop environment. The data comprises 66 columns of various information about the vehicles, and approximately 3 million rows of data.

The Hadoop cluster used rests on Hadoop version 3.1.2, and the hardware reads as follows: 31GB of RAM and 6 virtual CPUs on an AMD EPYC 7763 64-core prcoessor ruinning at 2.45 GHz. At time of processing and analysis, the HDFS cluster showed an available storage of 31GB of a maximum 390 GB, showing 357GB in use across 5 nodes (2 master, 3 worker).

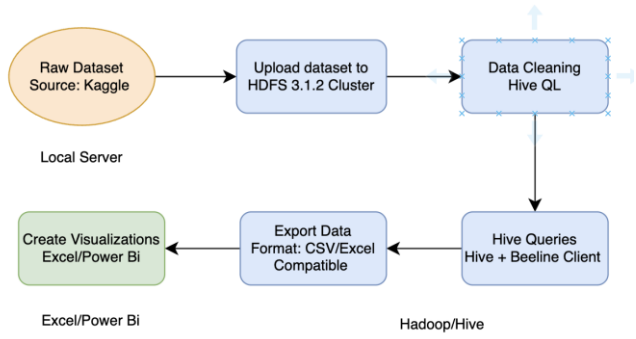
4. Implementation of Flowchart

To handle such a large dataset, we followed a structured workflow that helped us move from raw files to clean, usable data. It began with downloading the dataset from Kaggle—about 2GB zipped. We used scp to securely transfer it into our Linux environment, where it was unzipped into a 9.3GB CSV file containing over 3 million listings.

Once unzipped, we uploaded the data to HDFS and set up a Hive schema. We created two tables: one that kept all 66 columns from the original dataset (used_cars_full) and a second, streamlined version (used_cars_simple) focused on price, make, model, mileage, fuel type, and accident history. This made our queries more efficient and easier to manage.

From there, we used HiveQL to run our analysis. When we were ready to visualize the results, we exported the simplified table using INSERT OVERWRITE DIRECTORY, merged the output files using getmerge, and brought the final CSV into Excel and Power BI.

Each step in the flowchart helped us stay organized and made sure nothing slipped through the cracks during processing.



5. Data Engineering

The section has been renamed to “Data Engineering” from “Data Cleaning” as per the professor’s recommendation; this is on account of the dataset being from a pre-cleaned source. Kaggle provides pre-cleaned .csv (comma-separated values) files with consistent column headers. This prerequisite allowed this project to more heavily prioritize schema construction, table formatting, querying, and analysis rather than the more traditional approach of data cleaning.

The raw dataset file was first downloaded to a local machine as a compressed 2GB .zip file, then uploaded to HDFS. A directory was created in HDFS to house it, and it was then unzipped into a 9.3GB .csv file.

Seeing as Hive tables cannot skip columns in a “pick-and-choose" style, the “master” external table was first created using Beeline. “used_cars_full” comprises the full 66 columns of the dataset, and its fields cover a mixture of string, integer, float, and boolean data types.

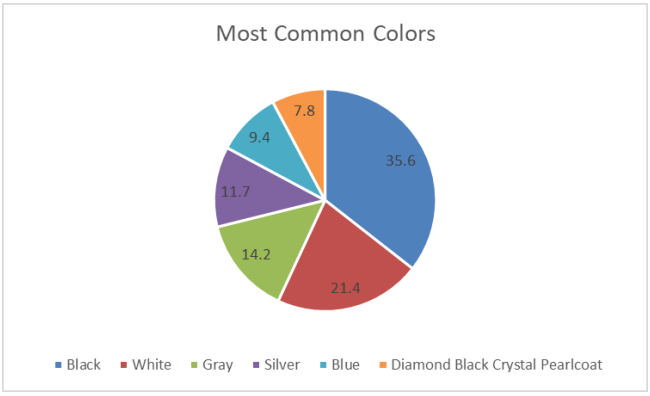
Then, the simplified table schema was derived from our external table for simplified and targeted analysis. “used_cars_simple” removes fields irrelevant to analysis, keeping only columns most relevant to vehicle price and characteristics. This lighter-weight table allows for faster queries, as it was built for performance. From there, various queries were run for surface-level analysis before the stripped data was prepared for deeper analysis.

The Hive command INSERT OVERWRITE DIRECTORY was used to write the columns found in “used_cars_simple” to a new HDFS directory in partitioned output format, creating files in the format of “000xxx_0”, with x being any combination of numbers in ascending order.

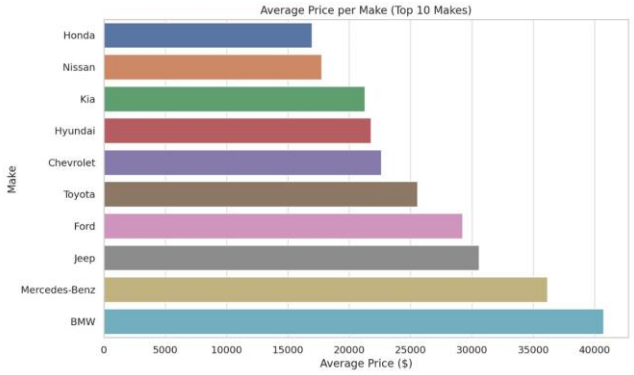
From there, a simple merge command joined the partitioned outputs into a single .csv file, which was first copied from HDFS to Linux file system, from which it was downloaded to the local machine for analysis using the scp command.

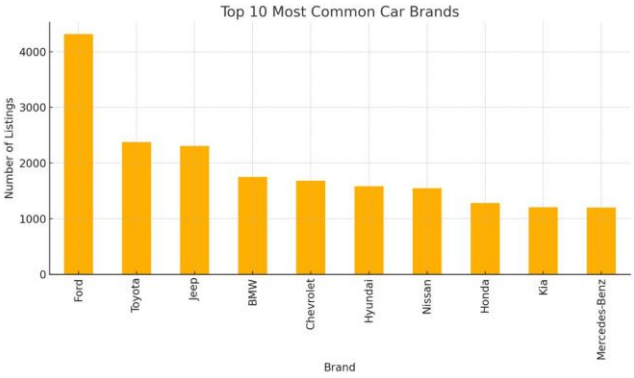
6. Analysis and Visualization

We developed a number of visualizations using the dataset to gain a deeper understanding of the 2020 used car market's dynamics. The most popular exterior color for cars, according to a pie chart, was black, which accounted for more than 35% of all listings. It was followed by Diamond Black Crystal Pearlcoat, white, gray, silver, and blue, demonstrating a definite consumer preference for dark and neutral hues.

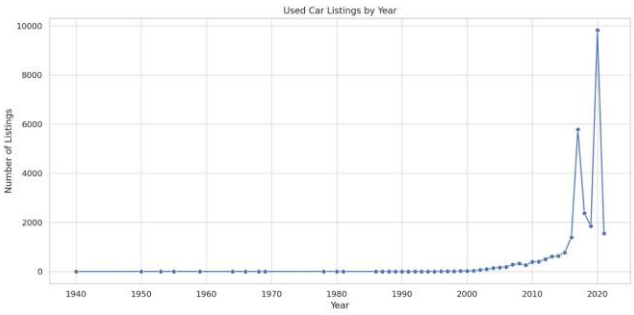


Dealerships that are planning their inventory or resale strategies may find this color trend helpful. Mercedes-Benz and BMW were at the top of another chart that showed the average price by car make, indicating the premium value attached to luxury brands. Conversely, the average price of more reasonably priced brands like Honda and Nissan was much lower.





A year-by-year line graph of used car listings showed a notable increase in listings around 2020, most likely as a result of the COVID-19 pandemic's impact on supply chains and new car manufacturing.



This spike highlights the market's brief move toward used cars as a result of the scarcity of new ones. Together, these graphics help us analyze changes in supply and demand, customer preferences, and pricing trends. They support the paper's findings regarding market adaptation and resilience during the pandemic by offering verifiable proof of how regional and economic factors affected used car inventory and sales behavior in 2020.

References

[1] T.A. Jones, “Writing a good paper,” *IEEE Trans. on General Writing*, Vol. 1, no. 2, pp.1-10, May 2002.
[2] K. Hwang, *Computer Arithmetic*, John Wiley, 1997.