# **DATA COLLECTION AND CURATION**

# **USED CARS ANALYSIS**

Name- Yugal Desai

Student ID-200493459

Date- February 26, 2022

**Contents**

[INTRODUCTION 3](#_Toc96768096)

[RESEARCH QUESTIONS 4](#_Toc96768097)

[DATA CLEANING 4](#_Toc96768098)

[ANALYSIS AND VISUALIZATIONS 5](#_Toc96768099)

[CONCLUSION 24](#_Toc96768100)

[APPENDIX 25](#_Toc96768101)

## 

## **INTRODUCTION**

The dataset has been retrieved from Kaggle.com and this dataset has over 3.5 million records and 16 columns. This dataset has the following columns-

* Maker – Company of cars
* Model – Different car models
* Mileage – The total Kilometers driven by the car
* Manufacture year – The year of manufacturing of the car
* Engine displacement – Engine CC (Size of piston cylinder)
* Engine power – Power of the engine in kW
* Body type – Type of the vehicle
* Color slug – Color of the body
* Stk year – year of last emission control
* Transmission – Automatic or manual transmission
* Door count – Number of doors
* Seat count – Number of seats
* Fuel type – Gasoline, Diesel, CNG, LPG and electric fuel types
* Date created – When the ad was scraped
* Date last seen – When the ad was last seen.
* Price in euros- List price converted to EUR

This dataset is raw and needs to be cleaned to gain utmost accuracy in the analysis. Some of the columns which require cleaning are maker, model, manufacturing year and price in euros, along with this, this dataset needs to use dimensionality reduction techniques to reduce and remove any insignificant columns present in the dataset.

The tools used for this analysis is spark data frames for analyzing the data and PowerBI for visualizing the analysis. This analysis targets 3 categories of groups based on the price in euros which are economic segment, intermediate segment, and luxury segment.

Economic segment targets the users who are willing to invest 3000 to 20000 euros in cars.

Intermediate segment targets the users who are willing to invest 20000 to 300000 euros in cars.

Lastly, Luxury segment targets the users who are willing to invest 300000 to 2000000 euros in cars.

Along with this, the analysis focus on manufacturing year and mileage as these are some of the important features which determines the price of these cars.

## **RESEARCH QUESTIONS**

These are the research questions that were taken into consideration for analyzing the data and produce various results-

1. Which type of transmission do most of the users prefer?
2. Which type of cars do users prefer?
3. What are the door and seat count preferred by users?
4. Who are the top 10 maker and model based on the highest average price?
5. Who are the top 10 maker and model based on the lowest average price?
6. To find the top 5 manufacturer based on the price range - 3000 to 20,000 euros.
7. To find the bottom 5 manufacturer based on the price range - 3000 to 20,000 euros.
8. To find the top 5 manufacturer based on the price range - 20,000 to 300,000 euros.
9. To find the bottom 5 manufacturer based on the price range - 20,000 to 300,000 euros.
10. To find the top 5 manufacturer based on the price range - 300,000 to 2000,000 euros.
11. Based on the 3 categories (economic, intermediate and luxury), suggest cars based on the manufacturing years.

## **DATA CLEANING**

There are few columns that needed to be cleaned to receive higher accuracy of analysis.

Below are some of the data cleaning methods executed-

1. Dropping any columns that have more than 50% missing values or NULLS.

* There were 3 columns from the dataset that had more than 50% missing values.
* These columns are- color\_slug, stk\_year and fuel\_type.

1. The manufacture year must be between 2000 and 2017.

* The manufacture year had multiple null values.
* Also, this column has values which can be considered as potential outliers such as 1949, etc.

1. Both maker and model must have no null values.

* There were about 50k rows in both maker and model which had null values.
* These null values are eliminated for better accuracy of the analysis.

1. The price range is between 3000 to 2000,000.

* Some of the values in price column are higher than 2 trillion euros which are outliers. Also, some rows have 0 as price which can be removed.
* These outliers are eliminated from the dataset.

## **ANALYSIS AND VISUALIZATIONS**

Transmission:

Chart, pie chart

Description automatically generated

* The graph represents that manual transmission are more common than automatic transmission.
* The count of cars with manual transmission is more than 900k and the count of cars with automatic transmission is around 310k.
* The percentage of manual transmission is more than 75% whereas the percentage of automatic transmission is 25%.

Type of cars:

Chart

Description automatically generated

Chart

Description automatically generated

* The graph represents the different body types which are preferred by the users.
* Compact cars are highly liked by the users, the count and percentage of compact cars are over 40k and around 85%.
* The second most preferred body type is van and the least preferred is sedan which is less than 0.05%.

Door and seat count:

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

* From pie chart 1, door counts 4 and 5 are most popular amongst the users whereas 2 and 3 are the least preferred.
* Cars with 4 and 5 doors combined contribute to 90% of the overall door count.
* From pie chart 2, cars with 5 seats are highly preferred when compared to cars with 4 and 7 seats. Around 80% prefer 5-seater cars and only 1.3% prefer 2-seater cars.

Top 10 maker based on highest average price:

Graphical user interface

Description automatically generated

* The graph illustrates the top 10 makers and models based on the highest average price.
* Lamborghini aventador is the most expensive car and the average price of the car is around 366k euros.
* Whereas Bentley continental-GT is the least expensive car from the graph and the average price of the car is around 105k euros.
* These cars might be considered as luxury cars as these cars are either hyper cars or sports cars which usually have 2 doors and are 2-seater cars.
* Although cars like tesla roadster and Bentley are 4-seater cars.

Top 10 maker based on lowest average price:

Graphical user interface, application

Description automatically generated

* The graph illustrates the top 10 makers and models based on the lowest average price.
* Skoda Galaxy is the least expensive car in the dataset and the average price of the car is around 3k euros.
* Whereas Nissan Frontier is the most expensive car from the graph and the average price of the car is around 3.5k euros.
* These cars might be considered as Economic cars as these cars are either compact cars or sedan cars which usually have 4 doors and are 5-seater cars.
* One of the reason these cars are less pricey than their counterparts is the manufacturing year. These cars might have been manufactured in early 2000.
* Also, the mileage or the amount of KMs travelled by these cars might be significantly higher than other cars.

Top 5 manufacturer based on the price range - 3000 to 20,000 euros:

Chart, bar chart

Description automatically generated

* This graph represents the cars which come under the Economic segment and the price of these cars is based on the average price.
* Although the difference between the price of these cars is minimal, Volvo v40 is the most expensive manufacturer in this segment and the average cost of the car is a little less 20k euros.
* On the other hand, Volkswagen beetle is the least expensive manufacturer in this graph and the average cost of this car is around 19.5k euros.

Seat and Door count of the cars:

Text

Description automatically generated with medium confidence

Bottom 5 manufacturer based on the price range - 3000 to 20,000 euros:

Chart, bar chart

Description automatically generated

* This graph represents the cars which come under the Economic segment and the price of these cars is based on the average price.
* Although the difference between the price of these cars is minimal, Skoda Galaxy is the least expensive manufacturer in this segment and the average cost of the car is a little less than 3.1k euros.
* On the other hand, Chevrolet Alero is the most expensive manufacturer in this graph and the average cost of this car is around 3.3k euros.

Seat and Door count of the cars:

Table

Description automatically generated with medium confidence

Top 5 manufacturer based on the price range - 20,000 to 300,000 euros:

Graphical user interface, application

Description automatically generated

* This graph represents the cars which come under the Intermediate segment and the price of these cars is based on the average price.
* BMW z8 is the most expensive manufacturer in this segment and the average cost of the car is around 245k euros.
* On the other hand, Rolls-Royce Wraith is the least expensive manufacturer in this graph and the average cost of this car is around 137k euros.

Seat and Door count of the cars:

Text

Description automatically generated

Bottom 5 manufacturer based on the price range - 20,000 to 300,000 euros:

Chart, bar chart

Description automatically generated

* This graph represents the cars which come under the Intermediate segment and the price of these cars is based on the average price.
* Although the difference between the price of these cars is minimal, Fiat 500x is the least expensive manufacturer in this segment and the average cost of the car is around 20k euros.
* On the other hand, Mazda cx-9 is the most expensive manufacturer in this graph and the average cost of this car is around 20.5k euros.

Seat and Door count of the cars:

Table

Description automatically generated

Top 5 manufacturer based on the price range - 300,000 to 2000,000 euros.

Chart, pie chart

Description automatically generated

* This pie chart represents the cars which come under the Luxury segment and the price of these cars is based on the average price.
* There are only 2 cars in the luxury segment. Lamborghini Aventador is the most expensive manufacturer in this segment and the average cost of the car is around 365k euros.
* Whereas Porsche Carrera-GT is the least expensive car in this segment and average cost of the car is around 300k euros.

Seat and Door count of the cars:

Graphical user interface, text

Description automatically generated

Economic Segment based on manufacturing year:

For this analysis, the manufacturing year has been divided into 2 categories-

* Old segment (manufacture\_year > 1999 AND manufacture\_year < 2010)
* New segment (manufacture\_year > 2013)

Old Segment

Chart, bar chart

Description automatically generated

Table

Description automatically generated

* This graph illustrates the top 5 maker, model based on average price and the manufacture between 2000 and 2009.
* This is classified as old segment and aims the cars that have been manufactured in early 2000.
* Lexus SC is the most expensive car in this segment and the manufacture year of this car is 2009.
* Honda s2000 is the least expensive car from this graph and the manufacture year is 2006.

Old Segment

Graphical user interface

Description automatically generated

Table

Description automatically generated

* This graph illustrates the bottom 5 maker, model based on average price and the manufacture between 2000 and 2009.
* Nissan Maxima is the least expensive car in this segment and the manufacture year of this car is 2000.
* Hyundai xg is the most expensive car from this graph and the manufacture year is 2002.

New Segment

Chart, bar chart

Description automatically generated

Graphical user interface, table

Description automatically generated

* This graph illustrates the top 5 maker, model based on average price and the manufacture greater than 2013.
* This is classified as new segment and aims the cars that have been manufactured in after 2013.
* Volkswagen Touran is the most expensive car in this segment and the manufacture year of this car is 2014.
* Volkswagen Jetta is the least expensive car from this graph and the manufacture year is 2015.

New Segment

Graphical user interface, chart, bar chart

Description automatically generated

Table

Description automatically generated

* This graph illustrates the bottom 5 maker, model based on average price and the manufacture greater than 2013.
* Volvo v50 is the cheapest car in this segment and the manufacture year of this car is 2014.
* Peugeot 107 is the most expensive car from this graph and the manufacture year is 2016.

Intermediate Segment based on manufacturing year:

Old Segment

Chart, bar chart

Description automatically generated

Table

Description automatically generated

* This graph illustrates the top 5 maker, model based on average price and the manufacture between 2000 and 2009.
* This is classified as old segment and aims the cars that have been manufactured in early 2000.
* Ford GT is the most expensive car in this segment and the manufacture year of this car is 2005.
* Bentley Brooklands is the least expensive car from this graph and the manufacture year is 2008.

Old Segment

Chart, bar chart

Description automatically generated

Table

Description automatically generated

* This graph illustrates the bottom 5 maker, model based on average price and the manufacture between 2000 and 2009.
* Lexus SC is the cheapest car in this segment and the manufacture year of this car is 2009.
* Lexus LS-460 is the most expensive car from this graph and the manufacture year is 2008.

New Segment

Chart, bar chart

Description automatically generated

Table

Description automatically generated

* This graph illustrates the top 5 maker, model based on average price and the manufacture greater than 2013.
* This is classified as new segment and aims the cars that have been manufactured in after 2013.
* Bentley Continental-Flying spur is the most expensive car in this segment and the manufacture year of this car is 2016.
* Bentley Continental GTC is the least expensive car from this graph and the manufacture year is 2015.

New Segment

A picture containing graphical user interface

Description automatically generated

Graphical user interface, table

Description automatically generated

* This graph illustrates the top 5 maker, model based on average price and the manufacture greater than 2013.
* Ford Mondeo is the cheapest car in this segment and the manufacture year of this car is 2014.
* Nissan 370-Z is the most expensive car from this graph and the manufacture year is 2016.

## **CONCLUSION**

These are the conclusions based on the analysis done and the research questions.

1. Manual transmission is the most used transmission type amongst the users.
2. The preferred body type is compact cars.
3. Door counts 4 and 5 are most popular amongst the users whereas 2 and 3 are the least preferred.
4. Cars with 5 seats are highly preferred when compared to cars with 4 and 7 seats. Around 80% prefer 5-seater cars and only 1.3% prefer 2-seater cars.
5. Top 10 maker based on highest and lowest average price: Lamborghini aventador is the most expensive car in the dataset and skoda galaxy is the cheapest car in this dataset.
6. Top and bottom 5 manufacturers based on the price range - 3000 to 20,000 euros: Volvo V40 is the most expensive car in the economic segment whereas skoda galaxy is the cheapest car in this segment.
7. Top and bottom 5 manufacturers based on the price range - 20,000 to 300,000 euros: BMW z8 is the most expensive car in the intermediate segment whereas Fiat 500x is the cheapest car in this segment.
8. There are only 2 cars present in the luxury segment based on the price range - 300,000 to 2000,000 euros: Lamborghini aventador is the most expensive car in this segment and Porsche Carrera GTx is the cheapest car.
9. Two categories of manufacturing year were established:

* Old segment – manufacturing year between 2000 and 2009.
* New segment - manufacturing year greater than 2013.

1. For Economic Segment:

* Old Segment – Lexus SC was the most expensive vehicle in this segment and Nissan maxima was the cheapest amongst their counterparts.
* New Segment – Volkswagen Touran was the most expensive vehicle in this segment and Volvo v50 was the cheapest amongst their counterparts.

1. For Intermediate Segment:

* Old Segment –Ford GT was the most expensive vehicle in this segment and Lexus SC was the cheapest amongst their counterparts.
* New Segment – Bentley Continental-Flying spur was the most expensive vehicle in this segment and Ford Mondeo was the cheapest amongst their counterparts.

## **APPENDIX**

-Importing sql types and describing the schema:

Text

Description automatically generated

The data type of date created, and date last seen is DateType in this schema.

-Creating Data frame to load the dataset from HDFS to Scala and displaying the table using show ():

Graphical user interface, text

Description automatically generated

Table:

Graphical user interface

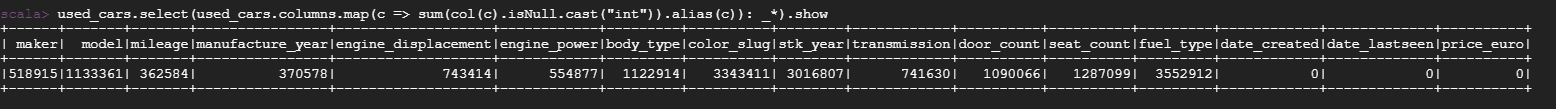
Description automatically generated with low confidence

-Changing the datatype to timestamp for date created and date last seen:

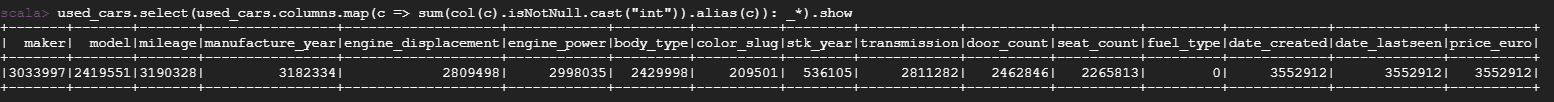
Calendar

Description automatically generated with medium confidence

-Missing values in the columns:



-Non missing values in the columns:



The columns which have more than 50% missing values are color\_slug, stk\_year and fuel\_type.

-Eliminating these columns from the dataset:

A picture containing calendar

Description automatically generated

-Checking for any repeated price across the ads:

Text

Description automatically generated

1295.34 is the price which will be eliminated from the dataset as it has been repeated 67k times.

-Cleaning the dataset:

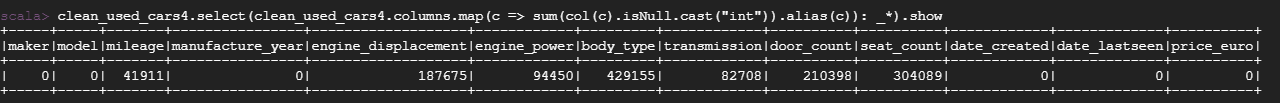
For cleaning the dataset,

* manufacture year must be between 2000 and 2017
* Removing all the null values from maker and model
* The price range is from 3000 to 2000,000 (3000 ≤ price ≤ 2000,000)
* Removing 1295.34 price from price\_eur column

A picture containing graphical user interface

Description automatically generated

-Checking for null values after cleaning the dataset:



The maker and model have no null values.

-Counting the number of records remaining after cleaning the dataset:



Over 1.3 M records are present in the dataset after cleaning it.

-Finding the top 10 maker and model based on highest average price:

Text

Description automatically generated

-Finding the 10 maker and model based on lowest average price:

Text

Description automatically generated

-Finding the top 5 Economic segment cars:

Text

Description automatically generated

-Finding the bottom 5 Economic segment cars:

Text

Description automatically generated

-Finding the top 5 Intermediate segment cars:

Text

Description automatically generated

-Finding the bottom 5 Intermediate segment cars:

Text

Description automatically generated

-Finding the cars in the luxury segment:

Text

Description automatically generated

- Finding the top 5 cars in economic segment which has manufacturing year between 2000 and 2009

Text

Description automatically generated

-Finding the bottom 5 cars in economic segment which has manufacturing year between 2000 and 2009

Text

Description automatically generated

-Finding the top 5 cars in intermediate segment which has manufacturing year between 2000 and 2009

Text

Description automatically generated

-Finding the bottom 5 cars in intermediate segment which has manufacturing year between 2000 and 2009

Text

Description automatically generated

-Finding the top 5 cars in economic segment which has manufacturing year greater than 2013.

Text

Description automatically generated

-Finding the bottom 5 cars in economic segment which has manufacturing year greater than 2013.

Text

Description automatically generated

-Finding the top 5 cars in intermediate segment which has manufacturing year greater than 2013.

Text

Description automatically generated

-Finding the bottom 5 cars in intermediate segment which has manufacturing year greater than 2013.

Text

Description automatically generated