**DATA ANALYTICS REPORT**

**On**

**Used Cars Investment Viability**

**Prepared by**

**Emeka Thompson**

**Emeka.thompson@mygeorgian.ca**

**July 04, 2021**

## Executive Summary

This report presents a data-driven analysis in the used-cars market to influence the firms' investment decisions. Data collected from classified adverts of used cars for sale across Germany and the Czech Republic have been carefully analyzed. This report also examines the relationship between the make, model, fuel type and price of used cars to serve the middle-class market of the industry. The findings in this report show the influence of the make and type of fuel the vehicle uses on its price and recommends the most viable vehicle manufacturers.

**Contents**

[Executive Summary 2](#_Toc92104336)

[Background 2](#_Toc92104337)

[Purpose 3](#_Toc92104338)

[Questions 3](#_Toc92104339)

[Methodology 3](#_Toc92104340)

[Data Description 4](#_Toc92104341)

[Report Findings 4](#_Toc92104342)

[Findings 1 4](#_Toc92104343)

[Findings 2 6](#_Toc92104344)

[Findings 3 7](#_Toc92104345)

[Findings 4 8](#_Toc92104346)

[Recommendations 9](#_Toc92104347)

[Conclusion 9](#_Toc92104348)

[TECHNICAL ANALYTICS 10](#_Toc92104349)

[Data Cleaning Steps 10](#_Toc92104350)

[Data Analysis 11](#_Toc92104351)

## Background

The used vehicle market has grown in importance in the past decade and is positioned to continue to do so in the future. Any motor vehicle that has previously been registered is considered used. The demand for used cars has increased as many people are buying or planning to buy their car to avoid public transport due to the ongoing covid-19 pandemic. The used car market is very active in many countries, and the business generates considerable economic income. Even as new car sales have slowed down in the recent past, the used car market has continued to grow over the past year and is larger than the new car market now.

This firm is considering investing in a used car business, which is viable but very risky and may impact its reputation and brand if not carefully understood. Investing in used car businesses entails much more than just putting some cars on a lot and waiting for customers to come along. It requires understanding how the vehicles' features, models, and availability influence their demand and prices.

## Purpose

The purpose of this report is to provide an intuitive guide to stakeholders that will enable them to make sound investment and business decisions towards investing in used cars. While it is important to ensure that a wide range of used cars brands and models from leading manufacturers are available for sale, it is more important to ensure that most of these cars will be sold. The target market includes students, middle-class families, low earners, and competitive advantage. The analysis shows some features of used cars that influence their prices and demand.

## Questions

The following question will be addressed in this report:

1. What is the relationship between car makes, models and price?

2. What are the top five vehicle manufacturers you would recommend? Why?

3. Does fuel type have any impact on the car price?

4. What age of used cars will make the most sales?

## **Methodology**

The dataset used, Classified Ads for Cars – a collection of used cars for sale in the Czech Republic and Germany, was downloaded from kaggle.com [[1]](#footnote-1) in a .csv format. The dataset was uploaded and last updated in 2017. The dataset was cleaned and analyzed using Spark in the Google Cloud Platform shell. The conditions used to filter the dataset to fit the purpose of this report were cars:

1. Worth between £1,000 and £60,000
2. Manufactured after the year 2004
3. With mileage less than 190,000
4. With known manufacturer names and models

The codes used for the cleaning and analyses are presented in the data cleaning section of this report, with screenshots in the appendix. Visualization of the results of the analysis was done with Microsoft Excel and Tableau. The findings from the analysis have been presented in subsequent sections of this report.

## Data Description

The dataset used had about 3.5 million rows and 16 columns. The dataset was quite dirty and filled with missing data, almost 50% of the entire dataset. Below is a descriptive summary of the columns in the dataset.

1. maker – The make of the car, 46 cars were sampled
2. model – The model of the car
3. mileage – In KM
4. manufacture\_year – Year of manufacture
5. engine\_displacement – In ccm
6. engine\_power – In kW
7. body\_type – The body/shape of the vehicle
8. color\_slug – The colour of the car
9. stk\_year – Year of the last emission control
10. transmission – Automatic or manual
11. door\_count – Number of doors
12. seat\_count – Driver and passenger sitting capacity
13. fuel\_type – Gasoline, diesel, CNG, LPG, electric
14. date\_created – When the ad was scraped
15. date*last*seen – When the ad was last seen. The data did not contain cars older than 60 days
16. price\_eur – List price converted to EUR

## Report Findings

### Findings 1

There were 46 cars makers listed for sale in the dataset, and they varied differently in their prices. The analysis shows a strong relationship between the car makes and their prices. The visualization below shows that Maserati has the highest price average, followed by Porsche, Lotus, Infinity and Jaguar. It is not unknown that some cars do not have a second-hand value while some others retain their value for a very long time.

The analysis results also showed that many of the most expensive cars are foreign cars (Maserati – Italy, Lotus – UK, Jaguar – India, Jeep – US etc.). However, many other factors influence the prices of used cars, such as use-condition, mileage, popularity, and design.

### Findings 2

Some of the vehicles are rare to find, while some are always in demand cars according to what is available. A visualization of the analysis of the car manufacturers that control the market is presented below.

We recommend the top 10 car manufacturers for this investment: Volkswagen, Skoda, Opel, Audi, Ford, Citroen, Fiat, Peugeot, Renault, and BMW. These ten cars make up over 60% of all the vehicles listed in the used car market. The reason why these ten cars are topping in availability is that they are German, Czech and French cars. This implies that folks will more readily buy vehicles that are local to their country because of the availability of parts and mechanic garages.

Further analysis of the Top brands and their most available models are presented below. The visualization below also shows the top 25 models by count. The most available model is Volkswagen – Golf, Skoda – Fabia, Opel – Astra, and Ford – focus.

Chart, bar chart

Description automatically generated

**Top 10 Brands and Their Top Models**

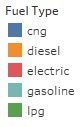
### Findings 3

A significant factor to consider when choosing a car is the type of fuel they use. The average fuel price for each fuel type in the dataset has been calculated, and the percentages are represented below.

Diesel and Gasoline-fueled cars, at 58.21% and 41.68%, respectively, are the most expensive cars, while electric, CNG and LPG-powered cars are the least expensive at 0.03 and 0.02%, respectively. The fuel type used influences the average price of used cars in the market. One reason for this price gap in most countries around the world, diesel is relatively cheaper than premium gasoline[[2]](#footnote-2). Another reason is that diesel-powered cars are generally more efficient on the highway than gas-powered cars[[3]](#footnote-3). Therefore, the demand influences the price.

Chart, bubble chart

Description automatically generated



**Influence Fuel Type of on Price**

### Findings 4

The year a car was manufactured significantly influences the price at which the vehicle may be sold. The dataset was filtered to show the increase in availability and average prices of the different car brands and models. From the analysis, there was a constant increase in the number of available used cars from 2005 to 2015. There was a significant and sharp drop in the average price of used cars in the years 2016 and 2017.

The reason for the sharp drop has more to do with the data collection than with the availability of the cars. The data was first uploaded in 2016 and updated in 2017. Therefore, not enough ads were recorded compared to the previous years. The most expensive used cars were recorded in 2015. This data shows that the more recent the used car is, the higher its demand and hence, its value.

Chart, line chart

Description automatically generated

**Influence Year of Manufacture on Price**



## Recommendations

Before taking further steps to invest, we recommend that:

1. Cars manufactured by Volkswagen, Skoda, Opel, Audi, Ford, and Citroen should be considered first for the business investment.
2. Cars from more recent years will provide better returns because they have good value and are in high demand and should therefore be considered.
3. Additional research and analysis will have to be conducted to determine if factors like car age, the average income of people in the business location and local competition will affect the investment's profitability.

Conclusion   
Understanding that the makes and model influence the price of used vehicles suggests that cars' choices in the dealership should be selected based on their models. Given the target market, investing in vehicles that use gasoline, CNG, LPG, or electricity may be a better option, to begin with, than vehicles powered by diesel due to their high cost.

## TECHNICAL ANALYTICS (APPENDIX)

## Data Cleaning Steps

The data was downloaded from Kaggle.com as a .csv file and imported to Google Cloud Platform via Dropbox. Before carrying out analysis the data was loaded in Spark dataframe called *carsData* using the following code:

val carsData = spark

.read.format("csv")

.option("header", "true")

.load("hdfs://10.128.0.2:8020/BigData/cars.csv")

To check the columns and schema by running these commands:

carsData.schema

carsData.printSchema()

All schemas were implicitly cast to string type. For this analysis, only columns that can answer our business questions were used, and their schemas were adjusted to match the datatype. The columns of the dataset were reduced, renamed, and assigned new schema using the following codes.

val cars\_cols = carsData.select(col("maker"),

col("model"),

col("mileage").cast("int"),

col("manufacture\_year").alias("year").cast("int"),

col("engine\_displacement").alias("engine\_ccm").cast("int"),

col("transmission"),

col("seat\_count").alias("seats").cast("int"),

col("fuel\_type"),

col("price\_eur").alias("price").cast("Float"))

The following steps were taken to clean the data:

1. The range of car prices in the price\_eur column had ridiculously high and low values. To use a price range that fits the business purpose (between £1000 and £60000), the price column was filtered with the following code.

val cars1 = cars\_cols.filter("price > 1000 AND price < 60000")

1. To select only cars manufactured after the year 2004:

val cars2 = cars1.filter("year > 2005")

1. To select cars with mileage less than 190,000:

val cars3 = cars2.filter("mileage < 190000")

1. To remove missing data in the maker and model columns:

val cars4 = cars3.filter(col("maker")

.isNotNull && col("model").isNotNull)

1. To remove duplicate rows in the dataset.

val cars5 = cars4.distinct

## Data Analysis

The new number of rows in the clean dataset was 1,102,758. This was determined using the statement – cars5.count. The following analyses were carried out.

1. The influence of the manufacturer name on the price of the used vehicles was analyzed by pivoting the makers column and the average car price using the following statement in Hive.

val cars\_p = cars5.groupBy(col("maker"))

.agg(round(avg(col("price")),2).alias("price")).orderBy(col("price").desc);

Text

Description automatically generated with medium confidence

1. The most available used vehicles in the market were analyzed by looking at the frequency of each car manufacturer in the dataset. The following statement was used.

val cars\_count = cars5.groupBy(col("maker")).count()

.orderBy(col("count").desc).limit(20)

Below is the screenshot of data obtained from Spark:

Text

Description automatically generated

1. To find out if fuel type had an influence on the car price, the average fuel prices for each of the samples were calculated and grouped by the fuel\_type column. The following statement was used.

val cars\_f = cars5.groupBy(col("fuel\_type"))

.agg(round(avg(col("price")),2).alias("price"))

.orderBy(col("price").desc)

The results indicate that diesel cars are more expensive than electric and LPG-powered vehicles.

Text

Description automatically generated

1. To understand how the manufacturing year of used vehicles are influencing their prices, the following code was used.

val cars\_mod\_yp = cars5.groupBy(col("maker"), col("model"), col("year"))

.agg(round(avg(col("price")),2).alias("price"))

.orderBy(col("price").desc)

The screenshot is presented below.

Text

Description automatically generated

1. https://www.kaggle.com/mirosval/personal-cars-classifieds [↑](#footnote-ref-1)
2. https://www.globalpetrolprices.com/articles/4/ [↑](#footnote-ref-2)
3. https://driving.ca/auto-news/news/to-diesel-or-not-to-diesel-that-is-the-question [↑](#footnote-ref-3)