TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

The aim of this report is to conduct an exploratory data analysis (EDA) of the 'automobile' dataset, which contains information on 205 vehicles. Through thorough data exploration and visualization methods, we endeavour to uncover trends and factors influencing the performance and characteristics of automobiles.

By leveraging the extensive attributes provided in the dataset (fuel type, body style, drive wheels, wheel base, length, width, height, curb weight, engine type, number of cylinders, engine size, bore, stroke, horsepower, peak rpm, city mpg, highway mpg and price) we seek to unveil the underlying narratives within the automotive landscape. From identifying correlations between engine size and horsepower to understanding the most common vehicle make, this report aims to provide a comprehensive understanding of the multifaceted world of automobiles.

Through our data-driven approach, we aim to illuminate evolving consumer preferences, manufacturing strategies employed by automobile companies, and overarching trends shaping the automotive industry.

DATA CLEANING

To clean the ‘automobile’ dataset, the following methods were employed to enhance its relevance and accuracy for analysis:

1. Replacement of Question Marks with NaN Values

Question marks in the dataset were replaced with NaN (Not a Number) values. This ensures that any missing data is recognized and can be dealt with appropriately during analysis, facilitating a user-friendly approach to handling missing values.

2. Data Formatting

Columns stored as 'objects' were converted to float or numeric data types to facilitate data manipulation for exploratory data analysis (EDA) and visualization purposes. Specifically, the columns for bore, stroke, horsepower, peak rpm, and price underwent this transformation.

By implementing these cleaning methods, the dataset was refined to ensure relevance, accuracy, and optimal performance for subsequent analysis and modeling tasks.

MISSING DATA

The missing data in the 'bore', 'stroke', 'horsepower', 'peak-rpm', and 'price' columns were identified. These missing values can be classified as MCAR (Missing Completely At Random), indicating that their absence is consistent across all observations and unrelated to any observed or unobserved factors within the dataset (HyperionDev, 2021).

To address the missing data:

1. Fill in Missing 'Bore' and 'Stroke' Values:

The missing 'bore' and 'stroke' values were replaced with the respective means calculated from vehicles with similar horsepower and peak-rpm attributes. This approach acknowledges the relationship between engine specifications and power output (Road&Track, n.d.).

2. Fill in Missing 'Horsepower' and 'Peak-rpm' Values:

Similarly, the missing 'horsepower' and 'peak-rpm' values were filled using the respective means derived from vehicles with similar bore and stroke specifications. This method accounts for the interdependence between engine parameters and performance metrics.

3. Fill in Missing 'Price' Values:

For the missing 'price' values, they were replaced with the mean price of vehicles from the same make. This strategy leverages the manufacturer's pricing trends to estimate missing values.

By employing these approaches, the missing data in the 'bore', 'stroke', 'horsepower', 'peak-rpm', and 'price' columns were effectively addressed, enhancing the completeness and accuracy of the dataset for further analysis.

DATA STORIES AND VISUALISATIONS

Valuable insights were gained from the data by creating and analyzing visual representations. Below are the various data stories along with their visualizations:

Top 5 most expensive vehicles

Figure 1 displays a bar plot comparing the prices of the five most expensive vehicles and the five cheapest vehicles in the dataset. This visualization offers a clear illustration of the price variation within the dataset, highlighting both the high-end and low-end vehicles. It provides valuable insights into the significant price disparity between luxury vehicles and non-luxury ones in the market, while also considering other attributes that influence vehicle pricing, such as wheel drive, horsepower and body style etc.

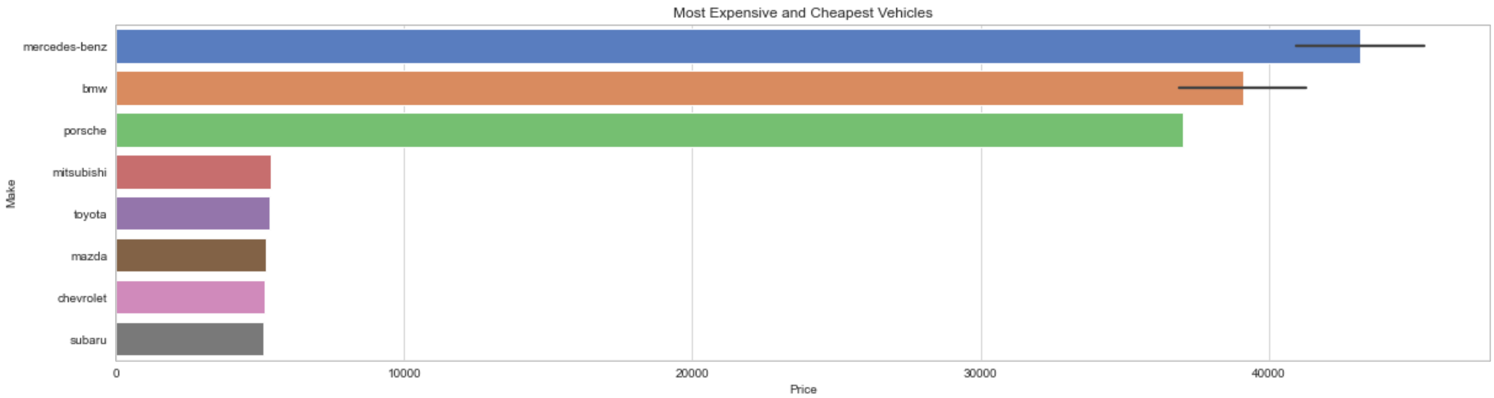

Figure 1

Top 5 horsepower ratings

Figure 2 presents a bar plot comparing vehicles with the top five horsepower ratings to those with the five lowest horsepower ratings. This comparison allows us to discern the varying performance levels of each vehicle, providing insights into their respective power capabilities. Additionally, the comparison sheds light on and individuals consideration of price differences associated with vehicles of different horsepower ratings.

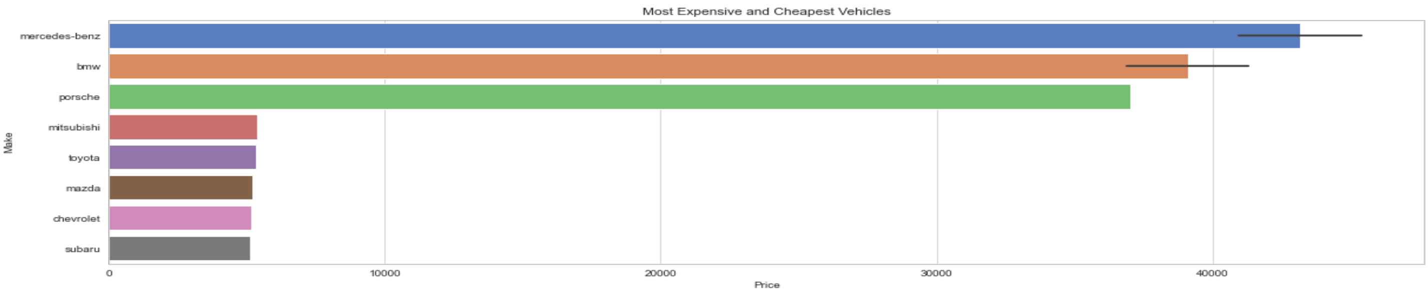

Figure 2

MPG in the city by vehicle make

In *Figure 3*, the city fuel efficiency (mpg) of various vehicle brands is depicted, providing valuable insights into which manufacturers excel in ensuring economical fuel consumption for urban driving. This data aids potential buyers, particularly those who predominantly drive in city environments, in making informed purchase decisions. The analysis reveals that Jaguars exhibit the highest fuel efficiency in city conditions, whereas Chevrolets demonstrate the lowest performance in this aspect.

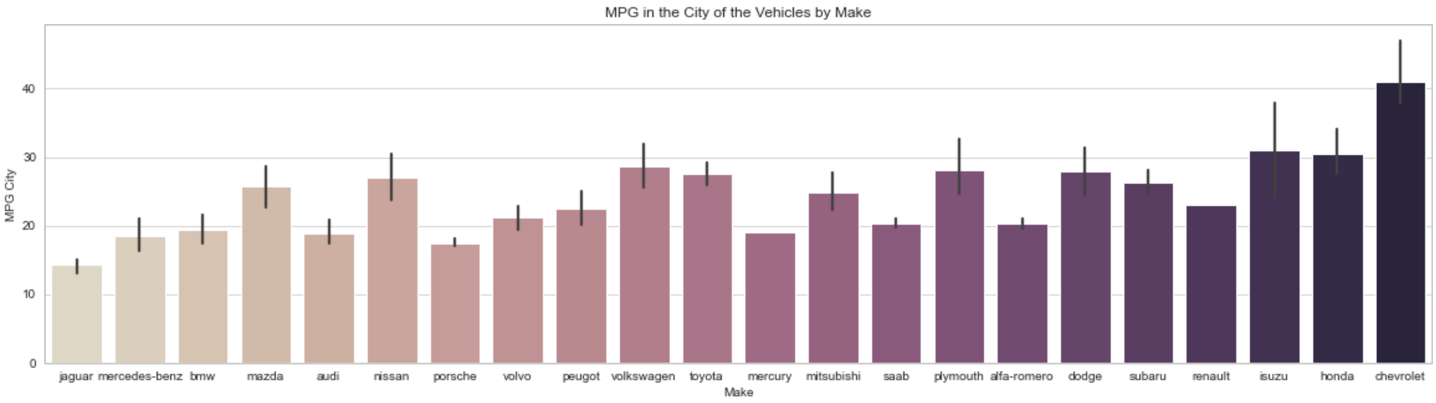

Figure 3

MPG on the highway by vehicle make

Figure 4 displays the highway fuel efficiency (mpg) of various vehicle makes, offering valuable insights into which manufacturers excel in delivering cost-effective fuel consumption for highway driving. This information is crucial for prospective buyers, especially those who frequently travel on highways, to make informed purchasing decisions. The analysis indicates that, similar to city mpg, Jaguars exhibit the highest fuel efficiency on the highway, while Chevrolets demonstrate the lowest performance in this regard.

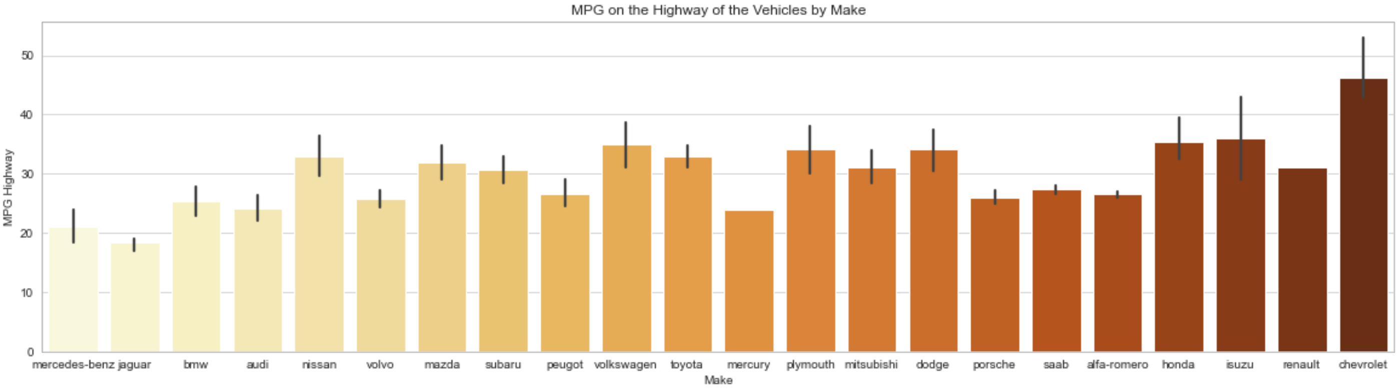

Figure 4

Most common vehicle make

Figure 5 presents a bar plot illustrating the frequency of vehicles for each make, which can offer valuable insights for consumers seeking reliable and affordable cars. This data may aid in making informed decisions regarding maintenance costs and overall affordability. It is noteworthy that Toyota emerges as the vehicle make with the highest frequency in the dataset.

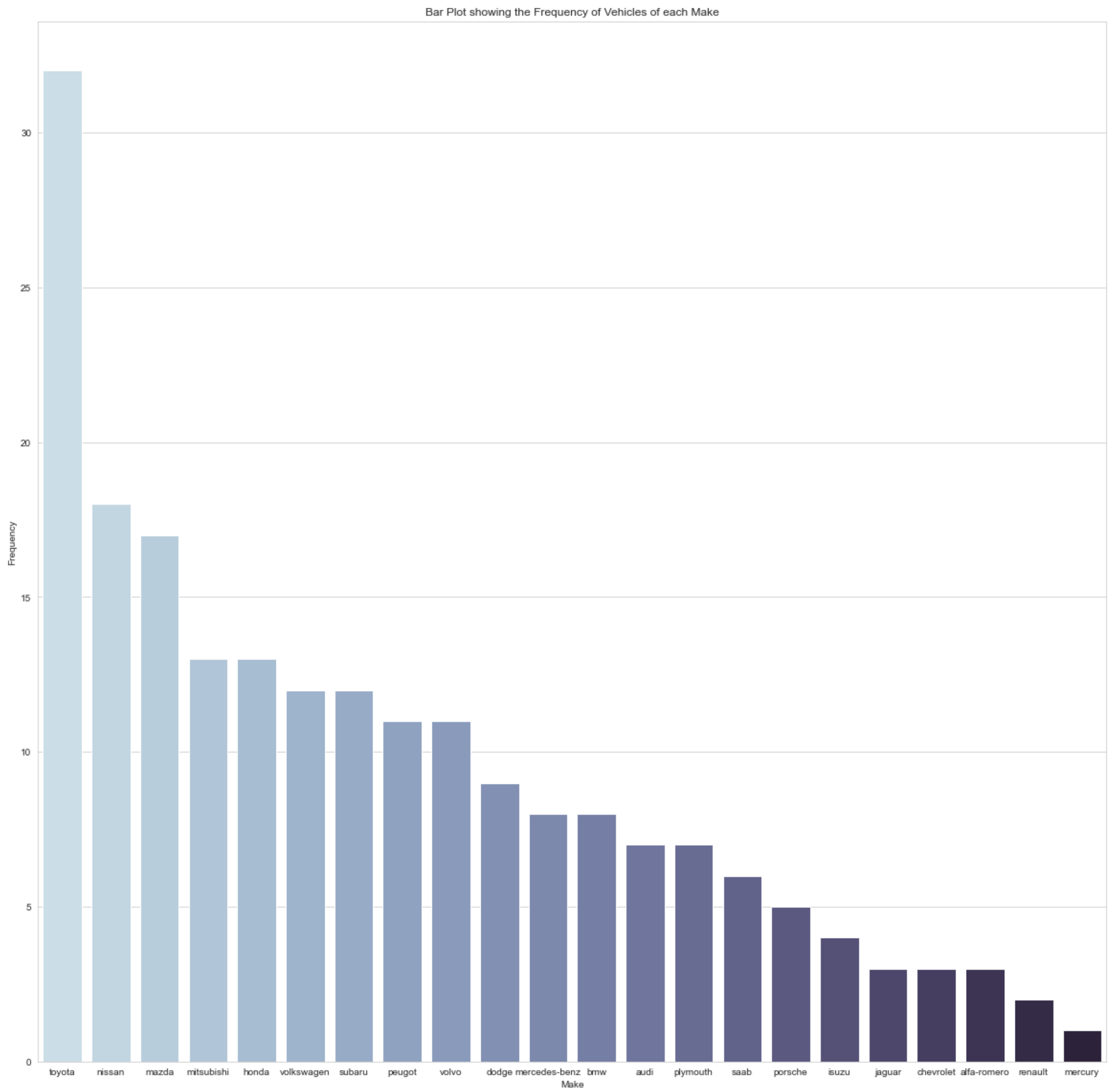

Figure 5

Engine Size vs Horsepower

In *Figure 6*, a distinct positive correlation between engine size and horsepower is evident. This trend suggests that, generally, as the engine size of a vehicle increases, its horsepower rating also tends to rise. This data could be valuable for individuals seeking vehicles with particular performance characteristics.

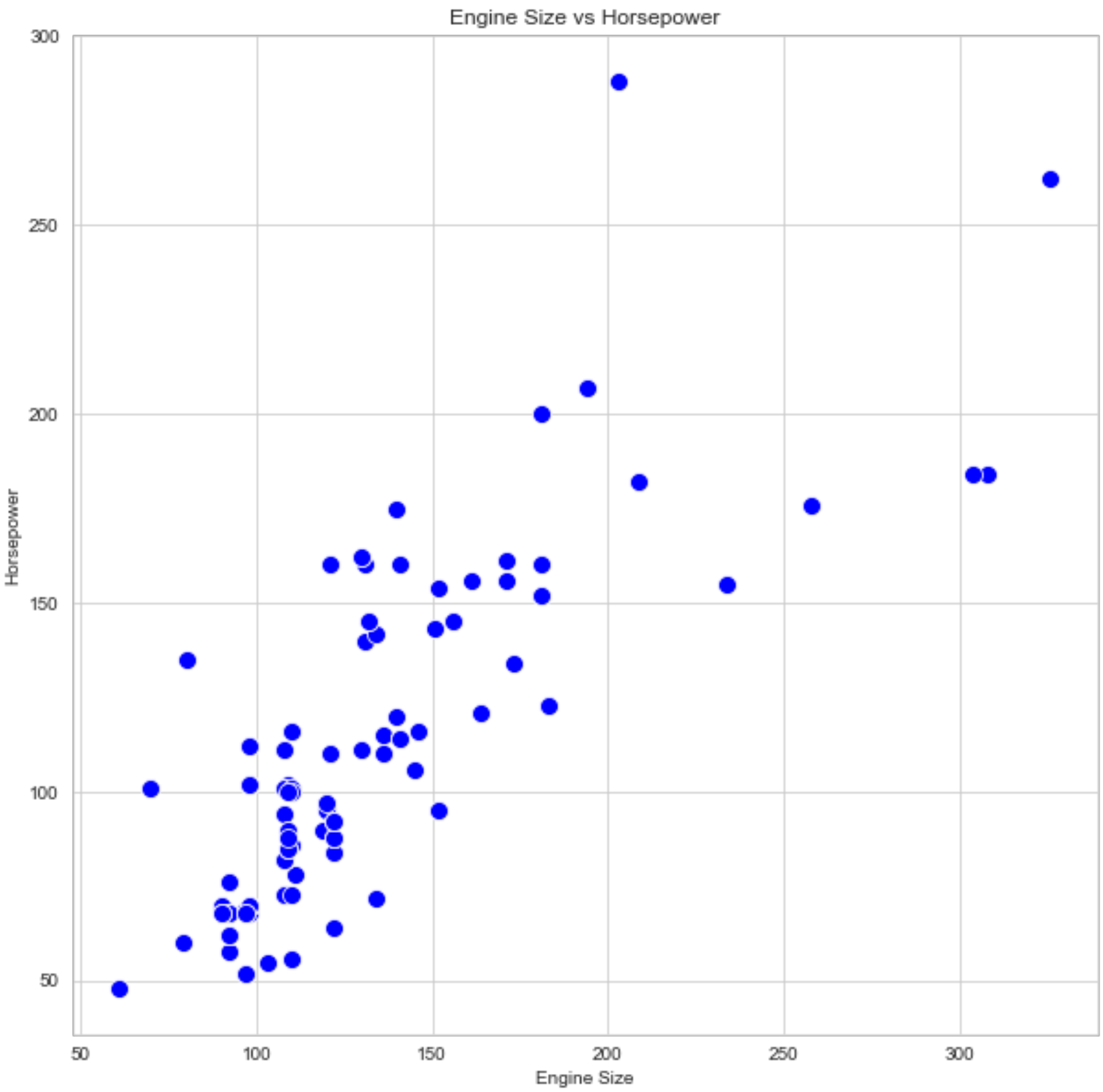

Figure 6

Percentage of cars of each body style

Figure 7 illustrates the distribution of vehicle body styles, with sedans comprising the largest portion at 46.8%, followed by hatchbacks at 34.1%. Wagon and hardtop body styles represent 12.2% and 3.9%, respectively, while convertibles account for 2.9% of the overall vehicle distribution.

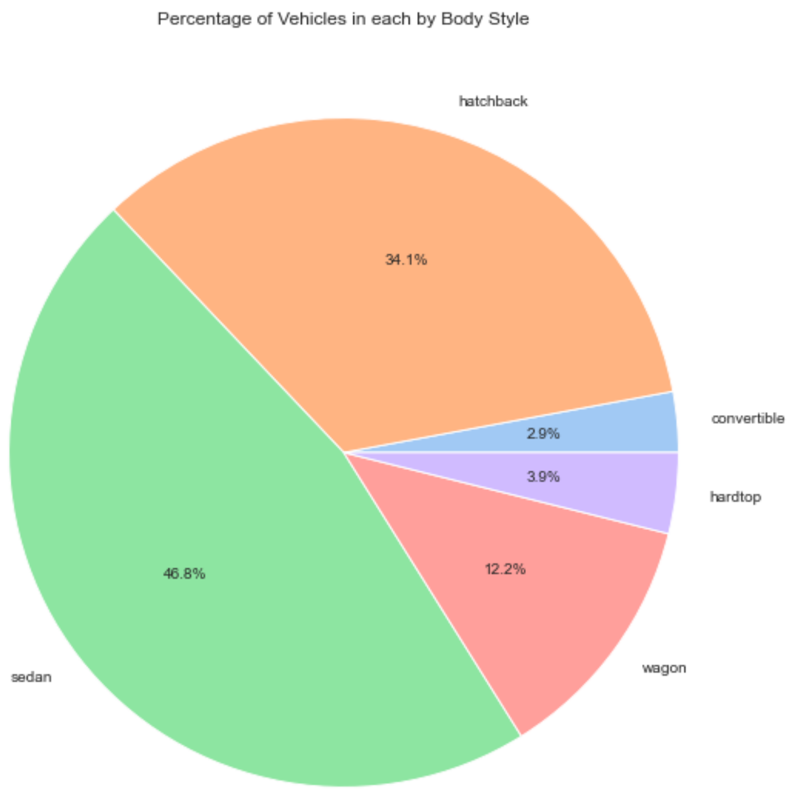

Figure 7

Price comparison of drive wheel categories

In *Figure 8* it is clear that rear-wheel-drive vehicles hold the highest prices, followed by 4-wheel-drive vehicles, while front-wheel-drive vehicles are the most budget-friendly option. This insight can assist customers in making informed decisions when selecting a vehicle based on its drive wheel type and their own unique driving requirements.

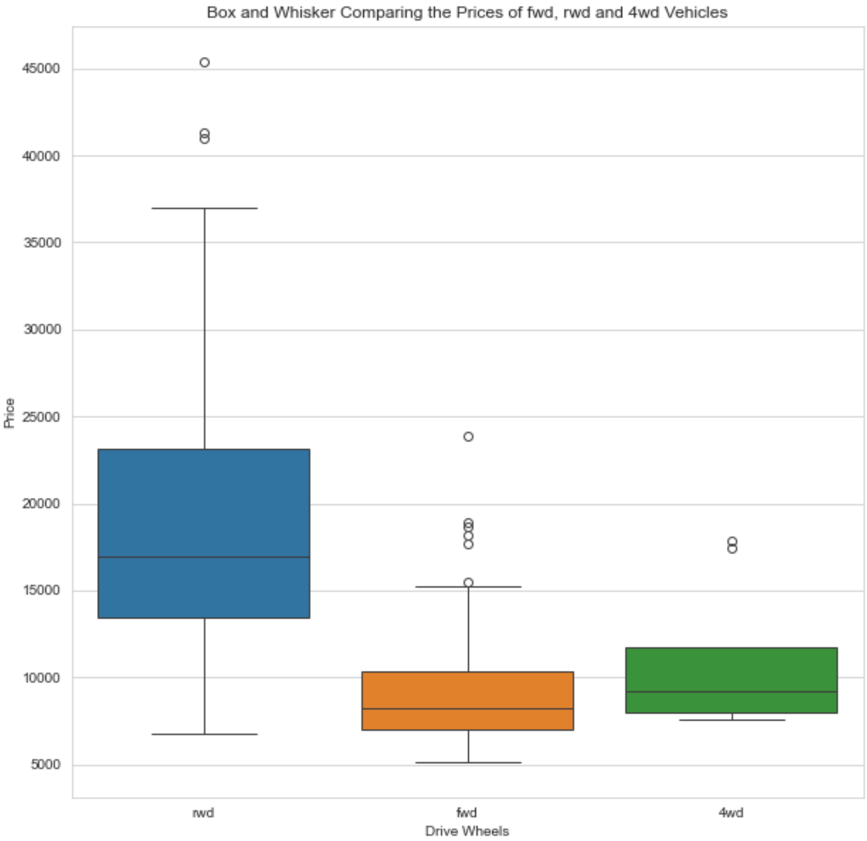
Box and Whisker Comparing the Prices of fwd, rwd and 4wd Vehicles
Price
45000
40000
35000
30000
25000
20000
15000
10000
5000
rwd
fwd
4wd
Drive Wheels

Figure 8

References

HyperionDev. (2021). Data Analysis - Preprocessing. [Educational notes]. Retrieved from Dropbox-NK23110009394.

Road & Track. (n.d.). Engine Stroke vs. Bore Explained. Retrieved from https://www.roadandtrack.com/car-culture/a30443334/engine-stroke-vs-bore-explained/

THIS REPORT WAS WRITTEN BY : Nagitta Kasirye-Koikanyang