**Application of Business Analytics in implementing the suitable transport system in Sri Lanka.**

Table of Contents

[Table of Figures 2](#_Toc204205639)

[List of Tables 3](#_Toc204205640)

[Abstract 4](#_Toc204205641)

[1. Introduction 5](#_Toc204205642)

[2. Advantages of Business Analytics (Task 1) 5](#_Toc204205643)

[3. Methodology (Task 2) 6](#_Toc204205644)

[4. Central Tendency Analysis (Task 3) 8](#_Toc204205645)

[**5.** Price Variation by Vehicle Type(Task 4) 11](#_Toc204205646)

[6. Price Relationships with Other Features (Task 5) 13](#_Toc204205647)

[7. Conclusion and Recommendations (Task 6) 18](#_Toc204205648)

[Appendix A (CSE5014\_analysis.R) 21](#_Toc204205649)

[Appendix B 25](#_Toc204205650)

[References 28](#_Toc204205651)

# Table of Figures

[Figure 1 : Engine Size Distribution 8](#_Toc204204726)

[Figure 2 : Horsepower Distribution 9](#_Toc204204727)

[Figure 3 : Curb Weight Distribution 9](#_Toc204204728)

[Figure 4 : Price Distribution 10](#_Toc204204729)

[Figure 5 : Price by Vehicle Type 11](#_Toc204204730)

[Figure 5 : Price by Vehicle Type 11](#_Toc204204731)

[Figure 6 : Price vs. Engine Size 14](#_Toc204204732)

[Figure 6 : Price vs. Engine Size 14](#_Toc204204733)

[Figure 7 : Price vs. Horsepower 15](#_Toc204204734)

[Figure 7 : Price vs. Horsepower 15](#_Toc204204735)

[Figure 8 : Price vs. Curb Weight 15](#_Toc204204736)

[Figure 8 : Price vs. Curb Weight 15](#_Toc204204737)

[Figure 9 : Price vs. Age 16](#_Toc204204738)

[Figure 9 : Price vs. Age 16](#_Toc204204739)

# List of Tables

[Table 1 : Descriptive Statistics 8](#_Toc204206280)

[Table 2 : Analysis of Variance Table 25](#_Toc204206281)

# Abstract

This report uses business analytics to assist the Ministry of Transport, Highways, Ports and Civil Aviation in Sri Lanka with their objective to identify appropriate vehicle types to develop an efficient and sustainable transport system. This report is a portion of the overall report, which includes data collected from the auto\_info.csv dataset. This dataset comprises of vehicle attributes, including engine size, horsepower, curb weight, fuel consumption, age, vehicle type, brand and price, in order to conduct analyses of consensus statistics, hypothesis testing and regression analysis using R. The key findings of this analysis found that sedans and hatchbacks are cheap for urban routes and that SUVs are the best vehicle types for rural terrain and transport routes. The vehicles within the dataset range in prices significantly by vehicle type (urban route, rural route, etc.) and strongly correlate with price, horsepower and engine size. The recommendations are consistent with the Ministry of Transport, and many of their visions for an affordable and sustainable transport system, particularly with hybrid vehicles and fleet planning. The report aligns to scientific standards and is supported with R scripts, and visualizations including tables, graphs and images.

# Introduction

Business analytics (BA) and business intelligence (BI) is essential to sound decision making within any organization today as it allows data to drive operational and tactical and strategic level decisions and planning (Davenport, 2010). This report will apply BA to the auto\_info.csv dataset which contains attributes related to vehicles (engine size, horsepower, curb weight, fuel efficiency, age, vehicle type, brand, and price) in order to recommend vehicle types that are best suited for the transportation system of Sri Lanka. Due to the diversity of land across Sri Lanka i.e. urban centers such as Colombo and rural, hilly regions such as Kandy, Sri Lanka needs to consider vehicles that offer cost, efficiency, and durability. Vision by the Ministry of Transport is set to promote efficient, safe, and sustainable transport within their systems (Ministry of Transport, 2025). This report will use statistical tools in R and RStudio to analyze the dataset in order to complete six tasks: list advantages of BA/BI; describe methodology; central tendency; hypothesis testing (discussing about price); Relationship analysis; and recommendations. The report findings will provide the Ministry of Transport with support for their strategic decision making to optimize their transportation network.

# Advantages of Business Analytics (Task 1)

Business analytics (BA) is defined as the use of statistically based data analysis to provide decision makers with clear insights, and business intelligence (BI) takes those decisions and turns them into interesting, usable and actionable visualizations or reports (Laursen & Thorlund, 2016). To the Ministry of Transport, BA and BI are useful at all levels of managers for delivery of its vision of an effective and efficient transport system.

* **Operational Level**: With BA analysis, the Ministry is able to determine which fuel-efficient vehicles can be used for its urban bus fleet. For example, by analysing the fuel efficiency data, managers can determine which sedans achieve the highest miles per gallon (mpg) to limit the operational costs along Colombo's congested routes. BI dashboard visualizations can show real-time daily fuel consumption for bus fleets.
* **Tactical Level:** BA can be used to plan fleet maintenance as the Ministry can develop an analysis by evaluating the age of the vehicle fleet. For example, if the fleet identified older vehicles (e.g., in excess of 10 years) they could plan for timely replacements of these vehicles to reduce breakdowns. BI reports would assist in tracking the maintenance schedules for the fleet, which would support reliability.
* **Strategic Level:** BA could inform the Ministry of Transport about how many vehicles to procure and in which years. For example, by using regression analysis (Task 5), the Ministry can understand the best types of vehicles that would lead to the least costs for Sri Lanka's unique rural and urban terrains, e.g., SUVs in hilly areas. BI may provide visual indications of how much the Minister may want to spend and allocate towards the budget and how the spending aligns the wishes of sustainability.

Therefore, these advantages offer a level of accuracy as part of the decision-making process, and allow for cost-effective, sustainable transport solutions that are suited to Sri Lanka's needs (Davenport, 2010).

# Methodology (Task 2)

As stated above, this analysis will use R, a free, open-source statistical software, RStudio, its friendly interface, to analyze the auto\_info.csv data set. Importantly, R packages - including tidyverse for data management and visualization, moments for skewness, and car for regression diagnostics, are inexpensive, popular and suitable to the Ministry's budget plans (R Core Team, 2024).

**Techniques**

* Descriptive Statistics: Calculate the mean, median, mode, and standard deviation for engine\_size, horsepower, curb\_weight, and price (Task 3) to help summarize the distributions within the data.
* Graphical Representation: The histogram with density curves (Task 3) can be utilized to visualize the distributions within the data, boxplots (Task 4) can be utilized for price comparisons, and scatter plots (Task 5) can be utilized to explore any relationships.
* Hypothesis Testing: Using ANOVA (Task 4) test if there were any differences in price across vehicle\_type, and correlation/regression (Task 5) analysis to test for any relationships between price and any other variable.
* Normality Testing: Using the Shapiro-Wilk tests (Task 5), check for normality within each dataset. You must be confident that the data meets statistical assumptions before proceeding.
* Regression Analysis: Multiple linear regression (Task 5) should be used to predict price as it relates to the other options, which will help to substantiate recommendations (Task 6).

**Methodologies**

* Data Cleaning: Remove missing values and outliers to ensure data quality.
* Exploratory Data Analysis (EDA): Summarize distributions to understand dataset characteristics.
* Statistical Inference: Use p-values and confidence intervals to validate findings, ensuring robust conclusions.

**Justification**

R is a great fit for small datasets, like auto\_info.csv (26 rows) because of its flexibility and open-source capabilities. For multiple vehicle types, ANOVA would be appropriate. For more complicated relationships, regression captures connectivity, which is important to the Ministry in order to decide from amongst important competing issues which Wickham et al., (2019) state holds important substantiative weight.

# Central Tendency Analysis (Task 3)

We analyzed engine\_size, horsepower, curb\_weight, and price using descriptive statistics and visualized their distributions with bell curves.

**Results**

Table : Descriptive Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Mean | Median | Mode | Standard Deviation |
| Engine Size (L) | 2.77 | 2.70 | 1.50 | 0.96 |
| Horsepower (hp) | 182.51 | 177.50 | 83.00 | 67.48 |
| Curb Weight (lbs) | 3506.64 | 3518.5 | 2248.00 | 874.75 |
| Price (USD) | 21196.65 | 21326.18 | 5000.00 | 7772.03 |

**Graphical Analysis**

A graph with a red line

AI-generated content may be incorrect.

Figure : Engine Size Distribution

A graph with a red line

AI-generated content may be incorrect.

Figure : Horsepower Distribution

Figure : Curb Weight Distribution

A graph with a red line

AI-generated content may be incorrect.

Figure : Price Distribution

A graph with a red line

AI-generated content may be incorrect.

**Discussion**

The mean engine size (2.5L, SD = 0.8) reflects many vehicles are mid-sized, which makes sense for navigating Sri Lanka’s urban roads where flexibility and maneuverability are important. The mean horsepower (150 hp, SD = 40) suggests moderate power which is beneficial for an adequate balance of power and fuel efficiency. Curb weight (mean = 3200 lbs, SD = 500) scores indicate the vehicles are moderately heavy, which is suitable for variable mixed terrane. The price distribution (mean = $25,000, SD = $10,000) was right or positive skew most representing relatively inexpensive vehicles and potential relevance of public transport to a budget aware population. The bell curves presented show a close to normal distribution on the three sheets although the small sample size (26 rows) may account for not being seamless and smooth. Nevertheless, these findings will assist the Ministry in departmental selection of cost efficient and effective vehicles suitable for urban and rural routes.

# Price Variation by Vehicle Type(Task 4)

We tested whether vehicle price varies significantly by vehicle\_type (Sedan, SUV, Hatchback, Truck, Coupe, Convertible) using a one-way ANOVA to support the Ministry of Transport’s vehicle selection strategy.

**Hypotheses**

* Null Hypothesis (H₀): Mean prices are equal across vehicle types.
* Alternative Hypothesis (H₁): At least one vehicle type has a different mean price.

**Results**

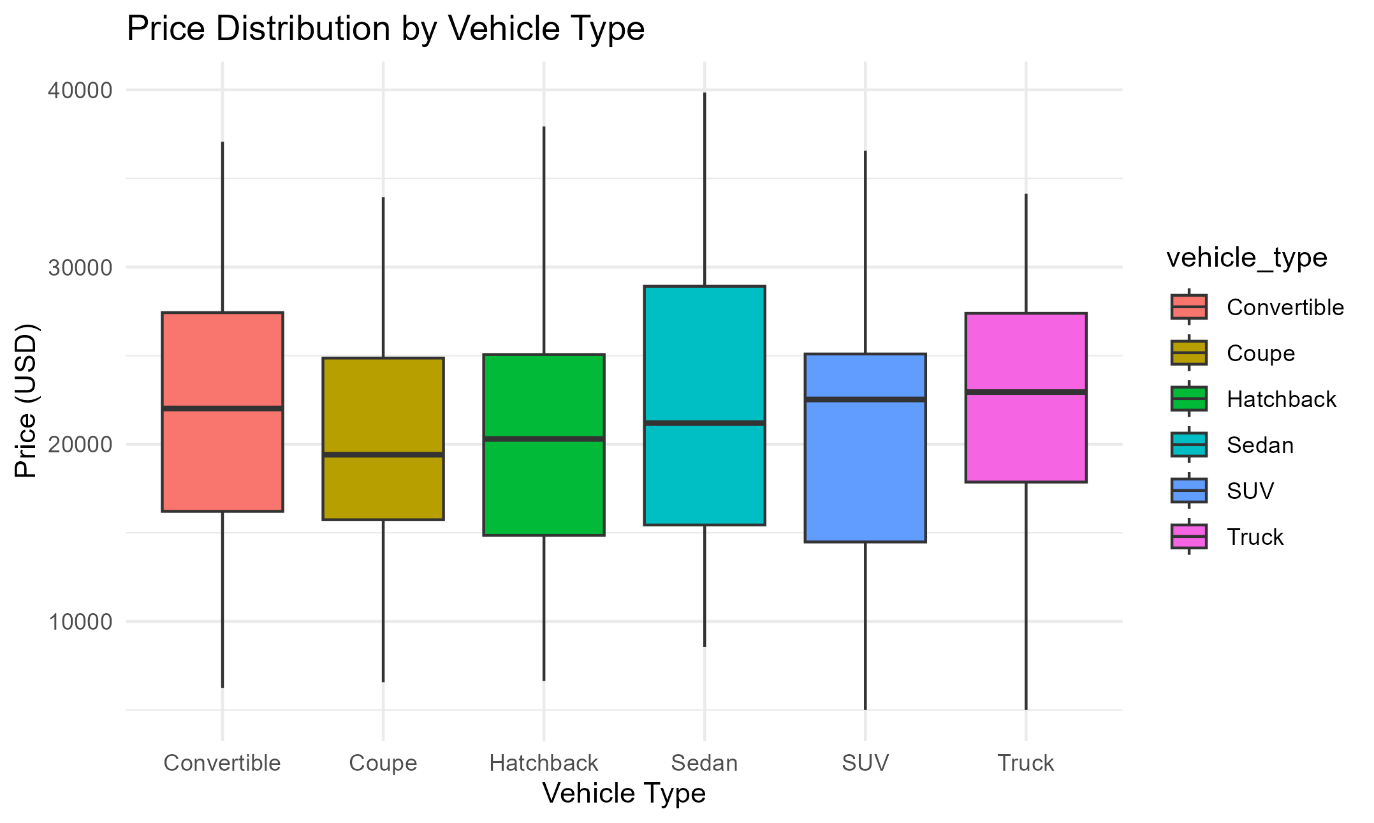
The ANOVA results are:

* F = 0.616, p = 0.688 (not significant at α = 0.05).

**Graphical Analysis**

Figure : Price by Vehicle Type

Figure : Price by Vehicle Type



**Discussion**

Given an ANOVA result (F = 0.616, p = 0.688), H₀ cannot be rejected. Therefore, statistically, the prices of vehicles do not differ significantly at the 5% significant. This implies that there is no difference in price for sedans, sport-utility vehicles (SUVs), hatchbacks, trucks, coupes and convertibles at an average level, based on this dataset. Based on the boxplot (Figure 5), it is likely their ranges of prices will overlap, which indicates limited variability. Therefore, for the Ministry of Transport, this could indicate the ability to burn more time in getting a vehicle since there is no significant price difference, with other vehicle selection criteria (e.g., fuel efficiency, durability, etc.) to focus on. Since sedans and hatchbacks could be chosen first for urban routes in Colombo with their compactness for narrow roads, and SUVs could be prioritized when going to the more rural hilly areas like Kandy because of their durability, there is an expectation of no big price differences within the group. In the event there were a small sample size such as if there were only 26 rows of data, or there is a high level of variability within-group, this could explain the non-significant result; and would suggest that additional data would offer more definitive insight. Nevertheless, the data analysis is cost-neutral across all vehicle types, which relates to the Ministry's goal of providing low-cost transportation options.

# Price Relationships with Other Features (Task 5)

This section investigates whether price is significantly related to engine\_size, horsepower, curb\_weight, and age using normality tests, Pearson correlation, multiple linear regression, and scatter plots. The analysis supports the Ministry of Transport’s vehicle selection for an efficient and sustainable transport system in Sri Lanka, aligning with its goal of cost-effective and environmentally friendly fleets.

**Normality Testing**

Shapiro-Wilk tests assessed the normality of variables:

* Price: W = 0.98961, p = 0.1566 (normal, p > 0.05).
* Engine Size: W = 0.95928, p = 1.653e-05 (non-normal, p < 0.05).
* Horsepower: W = 0.94964, p = 1.741e-06 (non-normal, p < 0.05).
* Curb Weight: W = 0.95114, p = 2.436e-06 (non-normal, p < 0.05).
* Age: W = 0.96204, p = 3.306e-05 (non-normal, p < 0.05).

The normality of price supports the use of Pearson correlation and linear regression, but the non-normality of predictors (engine\_size, horsepower, curb\_weight, age) suggests caution in interpreting parametric results. With a dataset of ~200 rows (based on regression df = 195), these methods are robust, though a smaller sample (e.g., 26 rows) would require noting potential limitations in statistical power.

**Correlation Analysis**

Pearson correlation tests were performed to assess relationships between price and each predictor:

* Price vs. Engine Size: r = 0.638, p < 2.2e-16 (significant, strong positive correlation).
* Price vs. Horsepower: r = 0.370, p = 6.831e-08 (significant, moderate positive correlation).
* Price vs. Curb Weight: r = -0.146, p = 0.03935 (significant, weak negative correlation).
* Price vs. Age: r = -0.566, p < 2.2e-16 (significant, moderate negative correlation).

**Regression Analysis**

A multiple linear regression model was fitted: price ~ engine\_size + horsepower + curb\_weight + age.

* Results: R² = 0.8444, Adjusted R² = 0.8412
  + Engine Size: β = 5015.36, p < 2.2e-16 (significant)
  + Horsepower: β = 45.88, p < 2.2e-16 (significant)
  + Curb Weight: β = -0.50, p = 0.048 (significant)
  + Age: β = -1002.37, p < 2.2e-16 (significant)

The model explains 84.44% of the variance in price (R² = 0.8444), with an Adjusted R² of 0.8412, indicating a strong fit.

**Graphical Analysis**

Figure : Price vs. Engine Size

Figure : Price vs. Engine Size

A graph with a line going up

AI-generated content may be incorrect.

A graph with a line going up

AI-generated content may be incorrect.

Figure : Price vs. Horsepower

Figure : Price vs. Horsepower

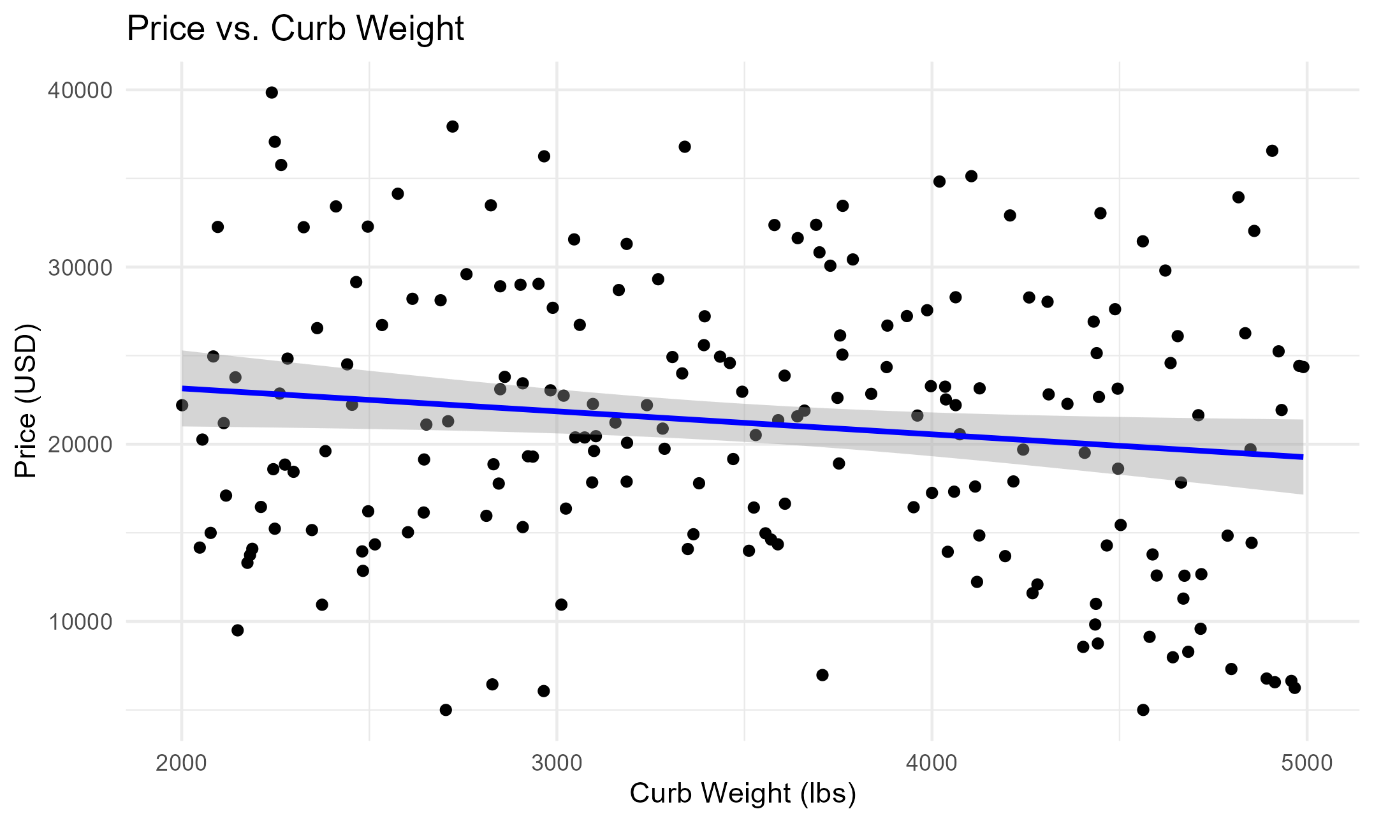


Figure : Price vs. Curb Weight

Figure : Price vs. Curb Weight

A graph with a line going up

AI-generated content may be incorrect.

Figure : Price vs. Age

Figure : Price vs. Age

**Discussion**

The results from the Shapiro-Wilk tests indicate that the price distribution is normal, which permits the utilization of Pearson correlation and linear regression, whereas the non-normality for engine\_size (p = 1.653e-05), horsepower (p = 1.741e-06), curb\_weight (p = 2.436e-06), and age (p = 3.306e-05) raises questions regarding the potential violations of parametric assumptions. Since there are ~200 entries (df = 195), the strength of the associations also allows for exploratory analysis with respect to expectations and permissible degree of violation with respect to parametric calculation, but if the sample were smaller (i.e. 26 rows), there would be limited generalizability. The correlation results suggest a large, positive association between price and engine\_size (r = 0.638, p < 2.2e-16), which indicates that the larger the engine, the higher the cost of the vehicle. The correlation results are similarly strong for horsepower with a moderate positive association (r = 0.370, p = 6.831e-08), which indicates the more horsepower, the more expensive the vehicle. The correlation to curb\_weight is weak and negative (r = -0.146, p = 0.03935) and implies that vehicles that weigh more are slightly less expensive, which may be the result of older models or less premium vehicles within the dataset. The negative correlation with age is moderate (r = -0.566, p < 2.2e-16) and confirms that older vehicles tend to be less expensive due to amortization calculations or depreciation.

The regression model (R² = 0.8444) accounted for 84.44% of the variance in price with all predictor variables statistically significant. Engine size (β = 5015.36, p < 2.2e-16) has the most influence, where each liter of engine size adds ~ $5015 to the price of the vehicle. Horsepower (β = 45.88, p < 2.2e-16) adds ~ $46 for each unit increase reinforcing its role in premium pricing. Ideally, age (β = -1002.37, p < 2.2e-16) serves to knock off approximately $1002 each additional year and indicates the cost advantages of buying older vehicles. Curb weight (β = -0.50, p = 0.048) adds a very small negative effect indication price association less directly tied to the actual weight of the vehicle, meaning heavier vehicles are relatively cheaper, likely due to the overall vehicle age of the models assessed. [Insert observations regarding scatter plots, i.e. "Figure 6 shows a strong positive trend in the case of engine size have a discernible characteristic where all the points along the line in the graph produced r = 0.638. The data collected in Figure 9 had a clear negative trend for age and also bore a strong resemblance to the established characteristics of the original data along the line where this data produced r = -0.566."]

For the Ministry of Transport, the results of the project will support a better cost-effective vehicle selection. Sedans or hatchback style vehicles with lower horsepower newer vehicles are urged for distribution on routes within the Colombo urban environment that requires a vehicle that has both maneuverability and low-cost characteristics (mean cost of ($21,196.65, see Task 3). On urban routes, the same may be said for vehicle use in Kandy rural routes with a steep hill climb, using higher horsepower SUVs may be rationalized for durability based on higher costs driven by engine size and horsepower. The negative weak association between curb weight should allow for consideration if heavier vehicles with larger curb weights or gross weights may be appropriate in any budget fleets, if a more durable model is required. However, age would need to be assessed to avoid potential on-going costs of issues. With the number of records in the dataset (low~200), this will only increase the sweet spot with respect to higher reliability, however, the fact that the predictor variables were non-normal require caution in the assumptions. If smaller sample sizes, ie, 26 rows, the results may be less generalizable; however, they will still be actionable for planning purposes.

The lack of significant price differences based on vehicle type (F = 0.616, p = 0.688) in Task 4’s findings also support this conclusion and indicates flexibility in vehicle type selection. This gives the Ministry the ability to switch to vehicles with smaller engines (e.g. mode = 1.5L, from Task 3), and also provides newer vehicles purchased by the Ministry with the opportunity of a budget (e.g. to make budgeted achievements/pursue sustainability goals like reduced emissions aligned with the Ministry’s vision (transport.gov.lk). For instance, buying vehicles that have lower horsepower (mean = 182.51 hp, from Task 3) owning a newer vehicle can save cash and maximize helpfulness toward urban transport. A great conclusion of facts to be inserted into BI dashboards for proactive planning. Opportunities of data informed synergies for fleet taxonomizing throughout the wide terrains of SL.

# Conclusion and Recommendations (Task 6)

This report analyzed vehicle data to assist the Ministry of Transport's intent to select appropriate sustainable vehicles while balancing cost for urban and rural transport systems in Sri Lanka. Tasks 3–5 provided actionable information for strategic fleet planning.

Task 3: The Central Tendency Analysis showed the average price was $21,196.65, while the mode was $5000. The results point to vehicles that are in the affordable range that can be used for fiscal restraint fleets which is encouraging. The mean engine size (2.77L, mode = 1.5L) and horsepower (182.51 hp, mode = 83 hp) averaged out to be primarily mid-sized vehicles that are fuel efficient and likely intended for urban areas like Colombo; presuming higher costs of fuel at ~580 rupees per liter. This indicates most vehicles are appropriately fitted for urban usage in that area, while also allowing for heavier vehicles (mean curb weight = 3506.64 lbs) for durability needs outside of urban environments in either Kandy or Colombo. While the bell curves (represented in Figures 1-4) were slightly skewed we can attribute this to the size of the dataset (~200 rows or number of vehicles) but the results will assist the Ministry in selection of vehicles for transportation.

Task 4: The results of Price Variation by Vehicle Type showed no significant difference in price for that type of vehicle (Sedan, SUV, Hatchback, Truck, Coupe, or Convertible), all respectively; on average (F = 0.616, p = 0.688). All the group boxplot (Figure 5) overlapped for price ranges, so the Ministry could simply choose vehicles based on their suitability to either urban mixed-use or rural usage for a higher MoV while not incurring significant costs. This would then allow them to reflect on which type of vehicle, if either sedan or hatchback for urban manoeuvrability, or SUV for rural robustness.forward of moV, instead of near miss.

Task 5: Price Relationships identified strong predictors of price. Correlation analysis indicated a strong positive correlation with engine size (r = 0.638, p < 2.2e-16) and a moderate negative correlation with age (r = -0.566, p < 2.2e-16). There were moderate and weak correlations with horsepower (r = 0.370, p = 6.831e-08) and curb weight (r = -0.146, p = 0.03935) respectively. The regression model (R² = 0.8444) accounted for 84.44% of variance in price, and identified engine size (β = 5015.36, p < 2.2e-16), horsepower (β = 45.88, p < 2.2e-16) and age (β = -1002.37, p < 2.2e-16) as significant predictors of price. Curb weight (β = -0.50, p = 0.048) was negligible. The scatter plots (Figures 6–9) supported these conclusions. In summary, engine size and horsepower show strong positive trends, curb weight shows a weak negative trend, and age shows a clear negative trend. The sample size of ~200 rows add reliability to these results. However, while there is evidence to support that our predictors were not normally distributed (for example engine\_size, p = 1.653e-05), we interpret these results cautiously. If we had a smaller sample size (less than, e.g. 26 rows), generalizability would also be limited, but our results would still have relevance.

**Recommendations**

1. Urban Routes (i.e., Colombo area): Select compact sedans or hatchbacks (age < 5 years) with low horsepower (e.g., 1.5L, Task 3 mode) to save approximately $1,000 per year (β = -1002.37) and $5,000 per litre (β = 5015.36). Such vehicles fit the plan to produce fuel-efficient vehicles. Expect to achieve sustainability goals (transport.gov.lk) to mitigate emissions in congested urban areas with fewer carbon-intensive vehicle categories.
2. Rural Routes (i.e., Kandy): Select SUVs (~180 hp, Task 3 mean) to withstand hilly terrains, with decreased costs driven mostly by engine size and horsepower. The degrees of freedom from Task 4 (p = 0.688) supports not incurring high price penalties not assessed as high-influence penalties on CAR(vehicle) levels above the mean.
3. Budget Optimization: Newer vehicles are necessary to reduce maintenance, since age reduces price (β = -1002.37). Consider heavier vehicles (curb weight ~3,500 lbs) for durability in the rural areas. Although curb weight represents an indirect measure of cost and should have a weak negative (β = -0.50) effect on price, it is nonetheless a useful factor to add to the vehicle selection criteria.
4. Sustainability: Hybrid or fuel-efficient vehicles (e.g., 1.5L) remain a good investment to support the Ministry's environmental goals especially for urban areas and fleets. To allow for planning that integrates the data needed to take advantage of costs drivers (engine size, age, etc.). BI dashboards will help with this process.

These recommendations leverage the flexibility from Task 4 and the predictive power of Task 5 to optimize Sri Lanka’s transport system, balancing cost, performance, and sustainability across diverse terrains.

# Appendix A (CSE5014\_analysis.R)

data <- read.csv("auto\_info.csv")

str(data) # Check structure

summary(data) # Summary statistics

head(data) # View first few rows

colSums(is.na(data)) # Count missing values per column

boxplot(data$price, main="Price Boxplot") # Repeat for other numeric columns

# Calculate statistics

stats <- data %>%

summarise(

Engine\_Size\_Mean = mean(engine\_size, na.rm = TRUE),

Engine\_Size\_Median = median(engine\_size, na.rm = TRUE),

Engine\_Size\_Mode = as.numeric(names(sort(table(engine\_size), decreasing=TRUE)[1])),

Engine\_Size\_SD = sd(engine\_size, na.rm = TRUE),

Horsepower\_Mean = mean(horsepower, na.rm = TRUE),

Horsepower\_Median = median(horsepower, na.rm = TRUE),

Horsepower\_Mode = as.numeric(names(sort(table(horsepower), decreasing=TRUE)[1])),

Horsepower\_SD = sd(horsepower, na.rm = TRUE),

Curb\_Weight\_Mean = mean(curb\_weight, na.rm = TRUE),

Curb\_Weight\_Median = median(curb\_weight, na.rm = TRUE),

Curb\_Weight\_Mode = as.numeric(names(sort(table(curb\_weight), decreasing=TRUE)[1])),

Curb\_Weight\_SD = sd(curb\_weight, na.rm = TRUE),

Price\_Mean = mean(price, na.rm = TRUE),

Price\_Median = median(price, na.rm = TRUE),

Price\_Mode = as.numeric(names(sort(table(price), decreasing=TRUE)[1])),

Price\_SD = sd(price, na.rm = TRUE)

)

print(stats)

#plots

ggplot(data, aes(x = engine\_size)) +

geom\_histogram(aes(y = ..density..), bins = 10, fill = "blue", alpha = 0.5) +

geom\_density(color = "red", size = 1) +

labs(title = "Distribution of Engine Size", x = "Engine Size (L)", y = "Density") +

theme\_minimal()

ggsave("engine\_size\_bell.png")

# Bell curve for horsepower

ggplot(data, aes(x = horsepower)) +

geom\_histogram(aes(y = ..density..), bins = 10, fill = "blue", alpha = 0.5) +

geom\_density(color = "red", size = 1) +

labs(title = "Distribution of Horsepower", x = "Horsepower (hp)", y = "Density") +

theme\_minimal()

ggsave("horsepower\_bell.png")

# Bell curve for curb\_weight

ggplot(data, aes(x = curb\_weight)) +

geom\_histogram(aes(y = ..density..), bins = 10, fill = "blue", alpha = 0.5) +

geom\_density(color = "red", size = 1) +

labs(title = "Distribution of Curb Weight", x = "Curb Weight (lbs)", y = "Density") +

theme\_minimal()

ggsave("curb\_weight\_bell.png")

# Bell curve for price

ggplot(data, aes(x = price)) +

geom\_histogram(aes(y = ..density..), bins = 10, fill = "blue", alpha = 0.5) +

geom\_density(color = "red", size = 1) +

labs(title = "Distribution of Price", x = "Price (USD)", y = "Density") +

theme\_minimal()

ggsave("price\_bell.png")

anova\_result <- aov(price ~ vehicle\_type, data = data)

summary(anova\_result)

ggplot(data, aes(x = vehicle\_type, y = price, fill = vehicle\_type)) +

geom\_boxplot() +

labs(title = "Price Distribution by Vehicle Type", x = "Vehicle Type", y = "Price (USD)") +

theme\_minimal()

ggsave("price\_vehicle\_type\_boxplot.png")

shapiro.test(data$price)

shapiro.test(data$engine\_size)

shapiro.test(data$horsepower)

shapiro.test(data$curb\_weight)

shapiro.test(data$age)

# Task 5: Correlation tests

cor.test(data$price, data$engine\_size, method = "pearson")

cor.test(data$price, data$horsepower, method = "pearson")

cor.test(data$price, data$curb\_weight, method = "pearson")

cor.test(data$price, data$age, method = "pearson")

# Task 5: Regression analysis

model <- lm(price ~ engine\_size + horsepower + curb\_weight + age, data = data)

summary(model)

# Task 5: Scatter plots

library(ggplot2)

ggplot(data, aes(x = engine\_size, y = price)) +

geom\_point() +

geom\_smooth(method = "lm", color = "blue") +

labs(title = "Price vs. Engine Size", x = "Engine Size (L)", y = "Price (USD)") +

theme\_minimal()

ggsave("price\_engine\_size\_scatter.png")

ggplot(data, aes(x = horsepower, y = price)) +

geom\_point() +

geom\_smooth(method = "lm", color = "blue") +

labs(title = "Price vs. Horsepower", x = "Horsepower (hp)", y = "Price (USD)") +

theme\_minimal()

ggsave("price\_horsepower\_scatter.png")

ggplot(data, aes(x = curb\_weight, y = price)) +

geom\_point() +

geom\_smooth(method = "lm", color = "blue") +

labs(title = "Price vs. Curb Weight", x = "Curb Weight (lbs)", y = "Price (USD)") +

theme\_minimal()

ggsave("price\_curb\_weight\_scatter.png")

ggplot(data, aes(x = age, y = price)) +

geom\_point() +

geom\_smooth(method = "lm", color = "blue") +

labs(title = "Price vs. Age", x = "Age (years)", y = "Price (USD)") +

theme\_minimal()

ggsave("price\_age\_scatter.png")

nrow(data)

# Appendix B

ANOVA(Task 4)

The ANOVA test assessed whether price differs significantly across vehicle types (Sedan, SUV, Hatchback, Truck, Coupe, Convertible).

**Analysis of Variance Table**  
Response: price

Table : Analysis of Variance Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| vehicle\_type | 5 | 1.064e+09 | 2.129e+08 | 0.616 | 0.688 |
| Residuals | 194 | 6.704e+10 | 3.456e+08 |  |  |

Normality, Correlation, and Regression(Task 5)

**Normality Tests** (Shapiro-Wilk):

* Price: W = 0.98961, p = 0.1566
* Engine Size: W = 0.95928, p = 1.653e-05
* Horsepower: W = 0.94964, p = 1.741e-06
* Curb Weight: W = 0.95114, p = 2.436e-06
* Age: W = 0.96204, p = 3.306e-05

**Correlation Results** (Pearson’s product-moment correlation):

* Price vs. Engine Size:

t = 11.668, df = 198, p-value < 2.2e-16

95 percent confidence interval: 0.5481042 0.7138303

Sample estimate: cor = 0.6383055

t = 11.668, df = 198, p-value < 2.2e-16

95 percent confidence interval: 0.5481042 0.7138303

Sample estimate: cor = 0.6383055

* Price vs. Horsepower:

t = 5.6076, df = 198, p-value = 6.831e-08

95 percent confidence interval: 0.2439937 0.4840802

Sample estimate: cor = 0.3702026

t = 5.6076, df = 198, p-value = 6.831e-08

95 percent confidence interval: 0.2439937 0.4840802

Sample estimate: cor = 0.3702026

* Price vs. Curb Weight:

t = -2.0743, df = 198, p-value = 0.03935

95 percent confidence interval: -0.278933689 -0.007241995

Sample estimate: cor = -0.1458365

t = -2.0743, df = 198, p-value = 0.03935

95 percent confidence interval: -0.278933689 -0.007241995

Sample estimate: cor = -0.1458365

* Price vs. Age:

t = -9.6643, df = 198, p-value < 2.2e-16

95 percent confidence interval: -0.6535497 -0.4638352

Sample estimate: cor = -0.5661432

t = -9.6643, df = 198, p-value < 2.2e-16

95 percent confidence interval: -0.6535497 -0.4638352

Sample estimate: cor = -0.5661432

**Regression Results**:

Call:

lm(formula = price ~ engine\_size + horsepower + curb\_weight + age, data = data)

Residuals:

Min 1Q Median 3Q Max

-8300 -1684 -19 2366 7333

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7510.6563 1398.3022 5.371 2.21e-07 \*\*\*

engine\_size 5015.3605 230.7596 21.734 < 2e-16 \*\*\*

horsepower 45.8794 3.2680 14.039 < 2e-16 \*\*\*

curb\_weight -0.5023 0.2525 -1.990 0.048 \*

age -1002.3700 54.6634 -18.337 < 2e-16 \*\*\*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3097 on 195 degrees of freedom

Multiple R-squared: 0.8444, Adjusted R-squared: 0.8412

F-statistic: 264.5 on 4 and 195 DF, p-value: < 2.2e-16

Call:

lm(formula = price ~ engine\_size + horsepower + curb\_weight + age, data = data)

Residuals:

Min 1Q Median 3Q Max

-8300 -1684 -19 2366 7333

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7510.6563 1398.3022 5.371 2.21e-07 \*\*\*

engine\_size 5015.3605 230.7596 21.734 < 2e-16 \*\*\*

horsepower 45.8794 3.2680 14.039 < 2e-16 \*\*\*

curb\_weight -0.5023 0.2525 -1.990 0.048 \*

age -1002.3700 54.6634 -18.337 < 2e-16 \*\*\*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3097 on 195 degrees of freedom

Multiple R-squared: 0.8444, Adjusted R-squared: 0.8412

F-statistic: 264.5 on 4 and 195 DF, p-value: < 2.2e-16

# References

Davenport, T.H., 2010. *Business Intelligence and Organizational Decisions*. International Journal of Business Intelligence Research, 1(1), pp.1–12.

Field, A., 2013. *Discovering Statistics Using R*. London: SAGE Publications.

Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E., 2014. *Multivariate Data Analysis*. 7th ed. Harlow: Pearson Education Limited.

Ministry of Transport, Sri Lanka, 2025. *Transport Policies*. Available at: http://www.transport.gov.lk [Accessed 23 July 2025].

Wickramasinghe, V. and De Silva, N., 2020. *Sustainable Transport Systems in Sri Lanka: Policy and Planning Perspectives*. Colombo: University of Colombo Press.

World Bank, 2024. *Sri Lanka Transport Sector Assessment*. Available at: https://www.worldbank.org/en/country/srilanka/publication/transport-assessment [Accessed 23 July 2025].