AP1_Assignment 2_Do Thi Ngoc Nhi_22106334

Executive Summary According to the examination of car maintenance data, the majority of vehicles repaired have lower to mid-range horsepower, primarily between 50 and 125 hp suggesting an emphasis on fuel efficiency for daily use. Common difficulties include cylinder, chassis concerns; however, there are different troubles related to fuel types with gas vehicles exhibit a greater range of issues than diesel vehicles. Replacements and modifications are the main maintenance procedures for problems, especially for sedans and hatchbacks from gas fueled-type. The findings also highlight body styles and fuel types to analyze maintenance requirements and suitable methods.

Task 1. Overall Horsepower Distribution

Automobile data

```
## 'data.frame':    204 obs. of  13 variables:
##  $ PlateNumber   : chr  "53N-001" "53N-002" "53N-003" "53N-004" ...
##  $ Manufactures  : chr  "Alfa-romero" "Alfa-romero" "Audi" "Audi" ...
##  $ BodyStyles    : chr  "convertible" "hatchback" "sedan" "sedan" ...
##  $ DriveWheels   : chr  "rwd" "rwd" "fwd" "4wd" ...
##  $ EngineLocation: chr  "front" "front" "front" "front" ...
##  $ WheelBase     : num  88.6 94.5 99.8 99.4 99.8 ...
##  $ Length        : num  169 171 177 177 177 ...
##  $ Width         : num  64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 64.8 ...
##  $ Height        : num  48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 54.3 ...
##  $ CurbWeight    : int  2548 2823 2337 2824 2507 2844 2954 3086 3053 2395 ...
##  $ EngineModel   : chr  "E-0001" "E-0002" "E-0003" "E-0004" ...
##  $ CityMpg       : int  21 19 24 18 19 19 19 17 16 23 ...
##  $ HighwayMpg    : int  27 26 30 22 25 25 25 20 22 29 ...
```

Engine data

```
## 'data.frame':    88 obs. of  8 variables:
##  $ EngineModel : chr  "E-0001" "E-0002" "E-0003" "E-0004" ...
##  $ EngineType  : chr  "dohc" "ohcv" "ohc" "ohc" ...
##  $ NumCylinders: chr  "four" "six" "four" "five" ...
##  $ EngineSize  : int  130 152 109 136 136 131 131 108 164 164 ...
##  $ FuelSystem  : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
##  $ Horsepower  : chr  "111" "154" "102" "115" ...
##  $ FuelTypes   : chr  "gas" "gas" "gas" "gas" ...
##  $ Aspiration  : chr  "std" "std" "std" "std" ...
```

Maintenance data

```
## 'data.frame':    374 obs. of  7 variables:
##  $ ID         : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
##  $ PlateNumber: chr  "53N-001" "53N-001" "53N-001" "53N-001" ...
##  $ Date       : chr  "15/02/2024" "16/03/2024" "15/04/2024" "15/05/2024" ...
##  $ Troubles   : chr  "Break system" "Transmission" "Suspected clutch" "Ignition (finding)" ...
##  $ ErrorCodes : int  -1 -1 -1 1 -1 1 1 0 -1 -1 ...
##  $ Price      : int  110 175 175 180 85 1000 180 0 180 180 ...
##  $ Methods    : chr  "Replacement" "Replacement" "Adjustment" "Adjustment" ...
```

```
#Replace the "?" with NA to handle missing data and can be used to analyze with NA function
Automobile[Automobile == "?"] <- NA
Engine[Engine == "?"] <- NA
Maintenance[Maintenance == "?"] <- NA
```

```
#Convert categorical variables BodyStyles, FuelTypes,  ErrorCodes to factors
#These datas are used in plotting charts
Automobile$BodyStyles <- as.factor(Automobile$BodyStyles)
Engine$FuelTypes <- as.factor(Engine$FuelTypes)
Maintenance$ErrorCodes <- as.factor(Maintenance$ErrorCodes)
```

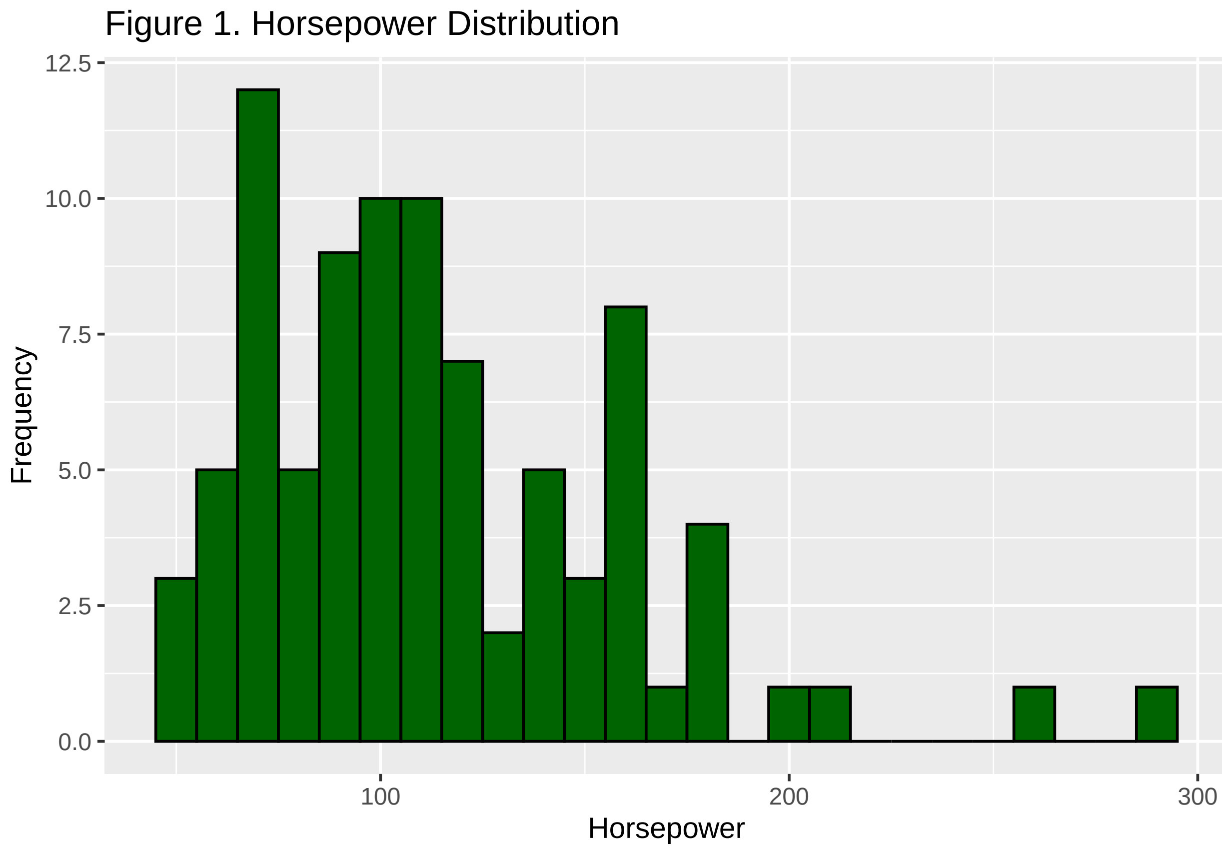
```
#Replace the missing values in column Horsepower with the mean horsepower
Engine$Horsepower <- as.numeric(Engine$Horsepower)
#Convert categorical variable Horsepower to a numeric because mean() needs the numeric input

Engine$Horsepower[is.na(Engine$Horsepower)] <- mean(Engine$Horsepower, na.rm = TRUE)
#is.na(Engine$Horsepower): Indicate the positions of NA values in the Horsepower column
#na.rm = TRUE: ignore NA values when calculating the mean
#mean(): continuous data
```

```
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
ggplot(Engine, aes(x = Horsepower)) +
  geom_histogram(binwidth = 10, fill = "darkgreen", color = "black") +
  labs(title = "Figure 1. Horsepower Distribution", x = "Horsepower", y = "Frequency")
```



Due to its suitability for presenting continuous data, a histogram was selected to show the horsepower distribution from the Engine dataset. It can be seen that the majority of automobiles entering the maintenance center have lower to mid-range horsepower, ranging from 50 to 125, while a few have high horsepower engines from 200hp and above. This indicates that these might be exceptions or represents luxury brands or imported cars. Overall, most vehicles come to the maintenance are typically designed for daily and family use, which prioritize fuel efficiency over high performance.

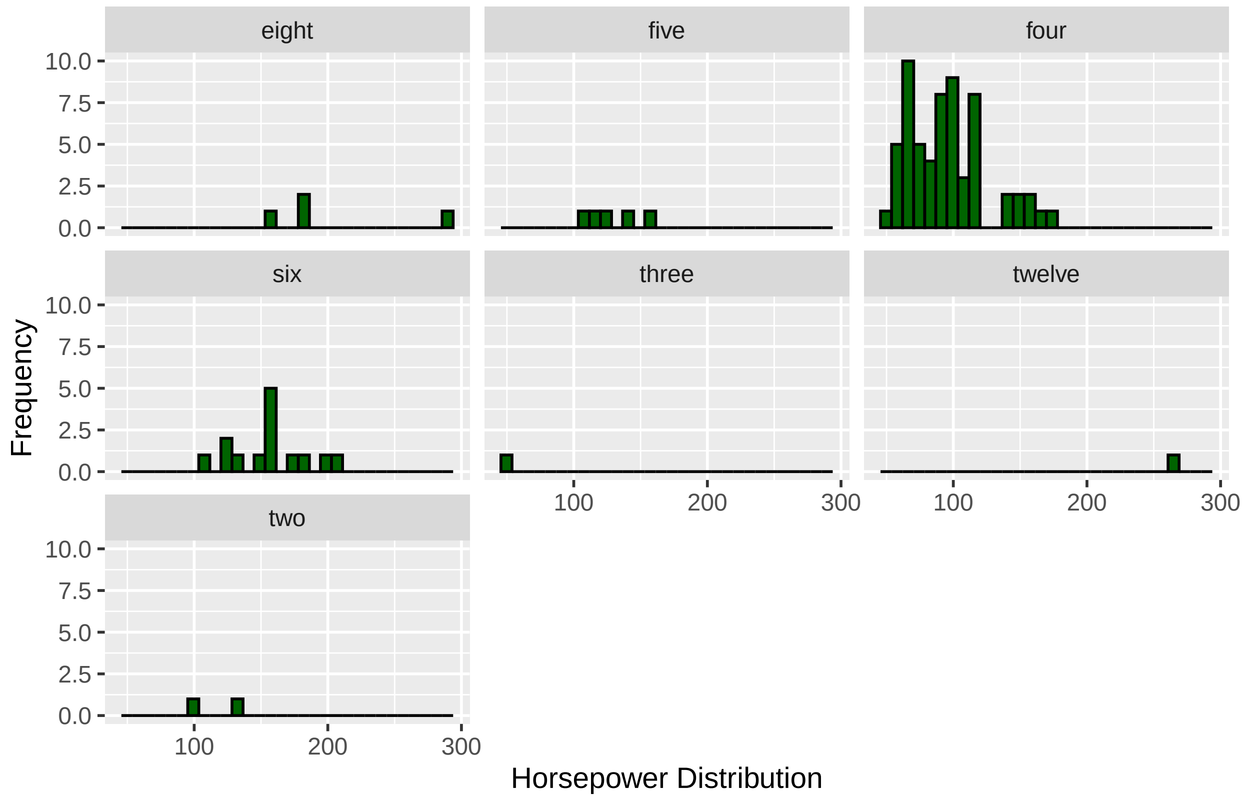
Task 2. Horsepower Distribution

2.1. By Number of Cylinders

```
library("ggplot2")
ggplot(Engine, aes(x = Horsepower)) +
  geom_histogram(fill = "darkgreen", color = "black") +
  labs(title = "Figure 2. Horsepower Distribution by Number of Cylinders",
       x = "Horsepower Distribution", y = "Frequency") +
  facet_wrap(~ NumCylinders)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 2. Horsepower Distribution by Number of Cylinders



The histogram shows the relationship between horsepower and engine cylinder count. There is a large number of four-cylinder engines, which have an average output of 75–125 horsepower. In contrast, five and six-cylinder engines with higher horsepower value are less common, indicating a concentration on efficiency or certain car models. Eight and twelve cylinders vehicles are uncommon, some special diagnostic and maintenance techniques are required.

2.2. By Engine Size

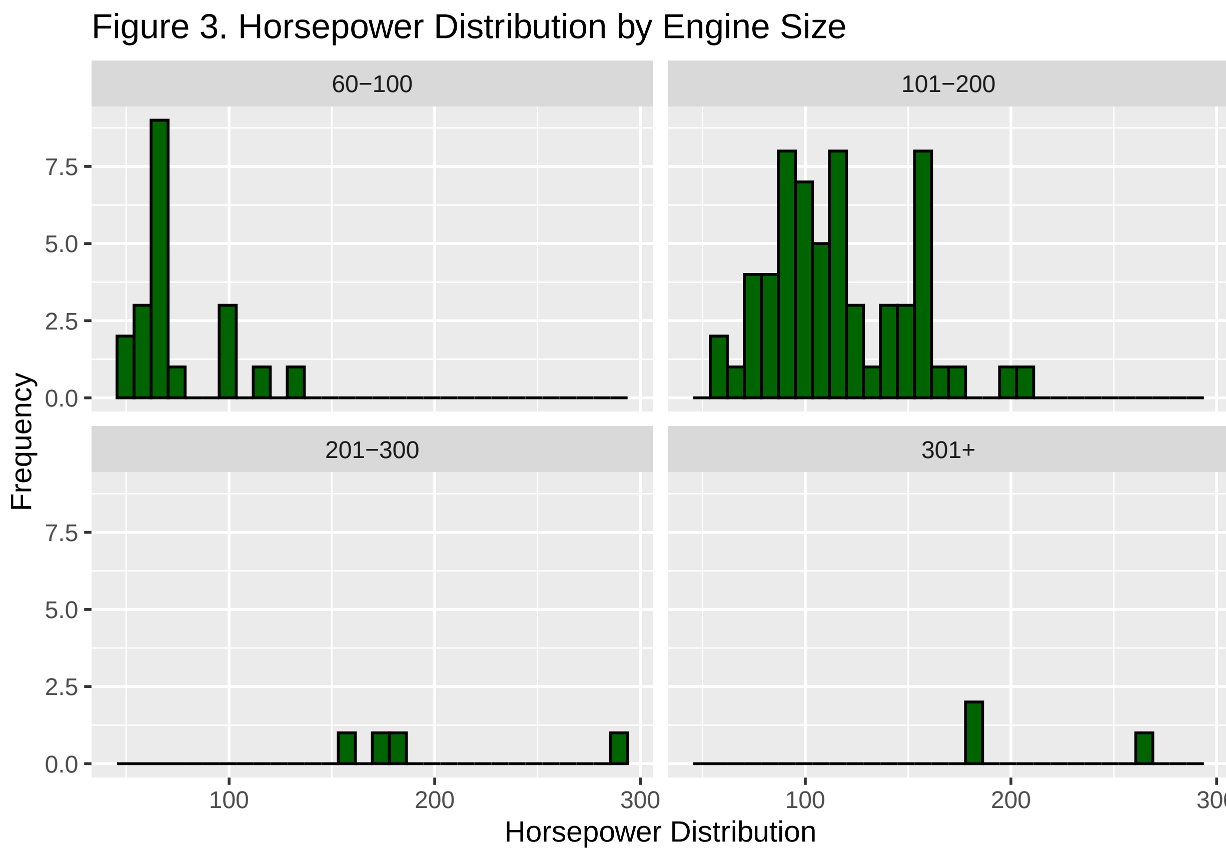
```
#Create a new categorical variable 'EngineSize'
# cut(): sort the engine sizes into selected ranges and assign the labels
Engine$EngineSize <- cut(Engine$EngineSize,
                         breaks = c(60, 100, 200, 300, Inf),
                         labels = c("60-100", "101-200", "201-300", "301+"),
                         right = FALSE)
print(table(Engine$EngineSize)) #Test the distribution of categories
```

```
##
##  60-100 101-200 201-300    301+
##      20      61       4       3
```

```
library("ggplot2")
ggplot(Engine, aes(x = Horsepower)) +
  geom_histogram(fill = "darkgreen", color = "black") +
  facet_wrap( ~ EngineSize) +
```

```r
  labs(title = "Figure 3. Horsepower Distribution by Engine Size",
       x = "Horsepower Distribution", y = "Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Through the division of engine sizes into several ranges (e.g., 60-100, 101-200, 201-300, and 301+), the histogram provides an illustration of the relationship between horsepower and engine size. The 60-100 engine size group shows a distribution at the lower horsepower range, with a noticeable peak for around 50hp cars. A more varied section of automobiles appears by the 101–200 range's broader range, which has a central tendency around 75–150 horsepower. The more uncommon 201-300 and 301+ engine sizes, which are linked to moderate to high horsepower, demonstrate specialized automobiles that would need special diagnostic tools and highly skilled expertise.

Generally, the majority are four-cylinder or less vehicles with horsepower between 75 and 150 and engine sizes between 101 and 200, common for daily driving in city and highway. Higher horsepower levels with bigger sizes are typically seen in engines with eight and twelve cylinders, which results in improved performance.

Task 3. Common Troubles in Engines

3.1. Top 5 Troubles

```r
Maintenance_troubles = unique(Maintenance$Troubles)
#Create a list of different troubles recorded in the Maintenance data
```

```r
exTroubles <- function(Maintenance_troubles, Maintenance){
  df <- subset(Maintenance,
               subset = (Maintenance_troubles == Maintenance$Troubles
                         & Maintenance$ErrorCodes != 0))
#Create a new data frame with troubles column that only have troubles or suspected of having troubles
  return(df)
}
Maintenance_troubles <- lapply(Maintenance$Troubles, exTroubles, Maintenance)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
OnlyTroubles <- Maintenance %>%
  filter(Troubles != "No error" & !is.na(Troubles))
  #Filter out the rows with no trouble
TopTroubles <- OnlyTroubles %>%
  group_by(Troubles, ErrorCodes) %>%
  #Group the unique values in Troubles and ErrorCodes columns
  summarise(count = n()) %>%
  #Count trouble type and creates a new column named count
  arrange(desc(count)) %>%
  #Arrange count column in descending order
  head(5) #Select the first 5 rows
```

```
## `summarise()` has grouped output by 'Troubles'. You can override using the
## `.groups` argument.
```

```r
library("kableExtra")
```

```
## Warning: package 'kableExtra' was built under R version 4.4.3

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Troubles	ErrorCodes	count
Cylinders	1	38
Chassis	-1	25
Ignition (finding)	1	22
Noise (finding)	1	19
Worn tires	-1	16

```
kbl(TopTroubles) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

The top five most frequently reported car problems are shown in the table. With a significant count of 38, "Cylinders" is the most common issue, indicating possible engine problems. "Chassis" and "Ignition (finding)" account for 25 and 22 respectively, possibly for component and electrical issues, emphasizing the importance of durable frameworks and stable ignition systems. "Noise (finding)" could be caused by damage or breakdown, and "Worn tires" emphasizes the need for regular maintenance when there is degradation of long usage.

3.2. Top 5 Troubles by Fuel Types

```
library(dplyr)
AutomobileEngine <- Automobile %>%
  left_join(Engine, by = "EngineModel", relationship = "many-to-many")
#merge Automobile and Engine based on the "EngineModel"
TroubleFuels <- AutomobileEngine %>%
  left_join(Maintenance, by = "PlateNumber", relationship = "many-to-many")
#merge Maintenance and the new dataset based on the "PlateNumber"
#relationship = "many-to-many": Specify multiple matching rows in the join columns

#seperate the fuel types into different tables
dieseldf <- TroubleFuels %>%
  filter(FuelTypes == "diesel") #Filter out diesel rows
gasdf <- TroubleFuels %>%
  filter(FuelTypes == "gas") #Filter out gas rows

CommonDieselTroubles <- dieseldf %>%
  filter(!is.na(FuelTypes) & !is.na(Troubles) & Troubles != "No error") %>%
  #Filter out NA values and rows where Troubles is "No error"
  group_by(FuelTypes, Troubles) %>%
  #Group the unique values in FuelTypes and Troubles column
  summarise(count = n(), .groups = "drop") %>%
  #Count unique combination of FuelTypes and Troubles
  arrange(desc(count)) %>%
  #Arrange count column in descending order
  head(5) #Select the first 5 rows
CommonGasTroubles <- gasdf %>%
  filter(!is.na(FuelTypes) & !is.na(Troubles) & Troubles != "No error") %>%
  group_by(FuelTypes, Troubles) %>%
  summarise(count = n(), .groups = "drop") %>%
  arrange(desc(count)) %>%
  head(5)
```

FuelTypes	Troubles	count
diesel	Chassis	4
diesel	Cam shaft	3
diesel	Cylinders	3
diesel	Crank shaft	2
diesel	Steering wheel	2

FuelTypes	Troubles	count
gas	Cylinders	36
gas	Ignition (finding)	22
gas	Chassis	21
gas	Noise (finding)	19
gas	Loss of driving ability	16

```r
library("kableExtra")
kbl(CommonDieselTroubles) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

```r
kbl(CommonGasTroubles) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

Diesel and gas-fueled engines provide different issues. “Chassis” problems are the most common for diesel, followed by “Cam shaft” and “Cylinders,” indicating structural and component faults. Whereas gas engines show a wider issues, with “Cylinders” being the most common, followed by “Ignition (finding)” and “Chassis,” suggesting a combination of engine and vehicle parts issues. “Noise (finding)” and “Loss of driving ability” are specific to gas engines. Additionally, the top troubles for diesel vehicles are from sedan body only (figure 4), conversely to gas with five forms (figure 5).

Overall, the top troubles reveal that “Cylinders” and “Chassis” are prominent concerns across the vehicles. However, there will be different issues when considering fuel types. While the gas engine has the same troubles compared to the top 5 troubles, diesel has several specialized issues to consider, which are “Cam shaft”, “Crank shaft”, and “Steering wheel”. Diesels have lower frequency of problems than gasoline, and top issues are from sedan while gas engines have several issues occur in different body styles.

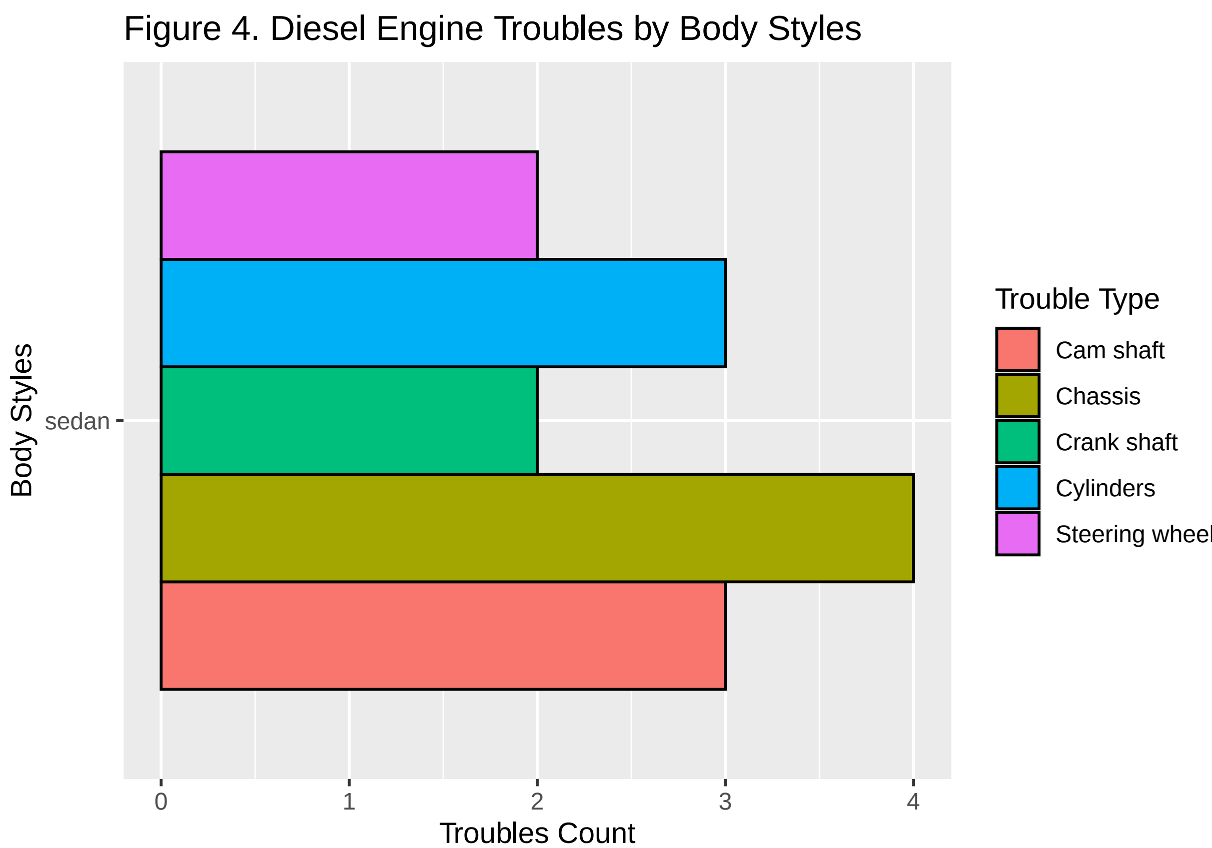
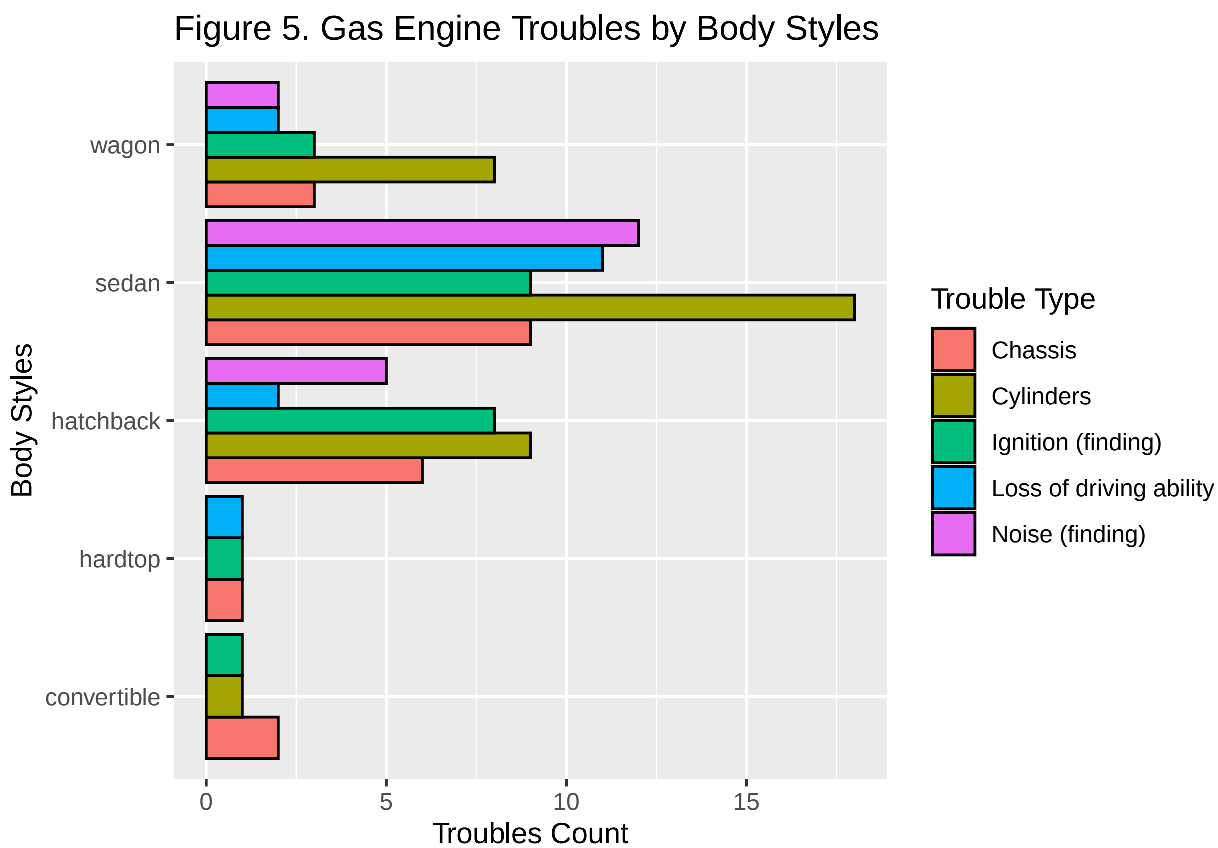


Figure 4. Diesel Engine Troubles by Body Styles
Body Styles
sedan
0
1
2
3
4
Troubles Count
Trouble Type
Cam shaft
Chassis
Crank shaft
Cylinders
Steering wheel

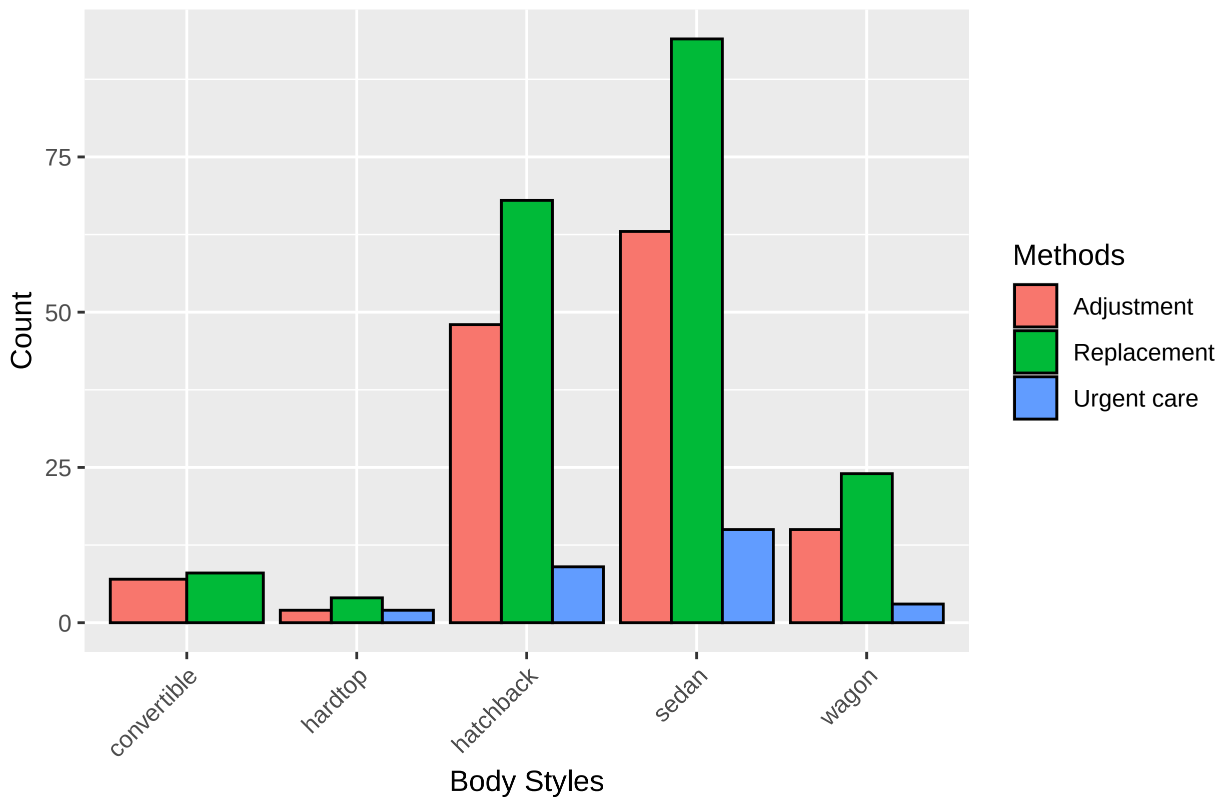


Task 4. Maintenance Methods

4.1. Maintenance by Body Styles

```
library(dplyr)
library("ggplot2")
Troubles_Methods <- TroubleFuels %>%
  filter(Troubles != "No error" & !is.na(Troubles) & !is.na(Methods))
  #Filter "No error" in TroubleFuels, no NA in Troubles and Methods data
BodyMethods <- Troubles_Methods %>%
  group_by(BodyStyles, Methods) %>%
  #Create groups for each specific body style and the maintenance method
  summarise(count = n(), .groups = "drop")
  #Calculate the number of rows (n())
  #.groups = "drop": remove the grouping structure from the data frame.
ggplot(BodyMethods, aes(x = BodyStyles, y = count, fill = Methods)) +
  geom_bar(stat = "identity", position = "dodge", colour = "black") +
  labs(title = "Figure 6. Maintenance Methods by Body Styles", x = "Body Styles", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

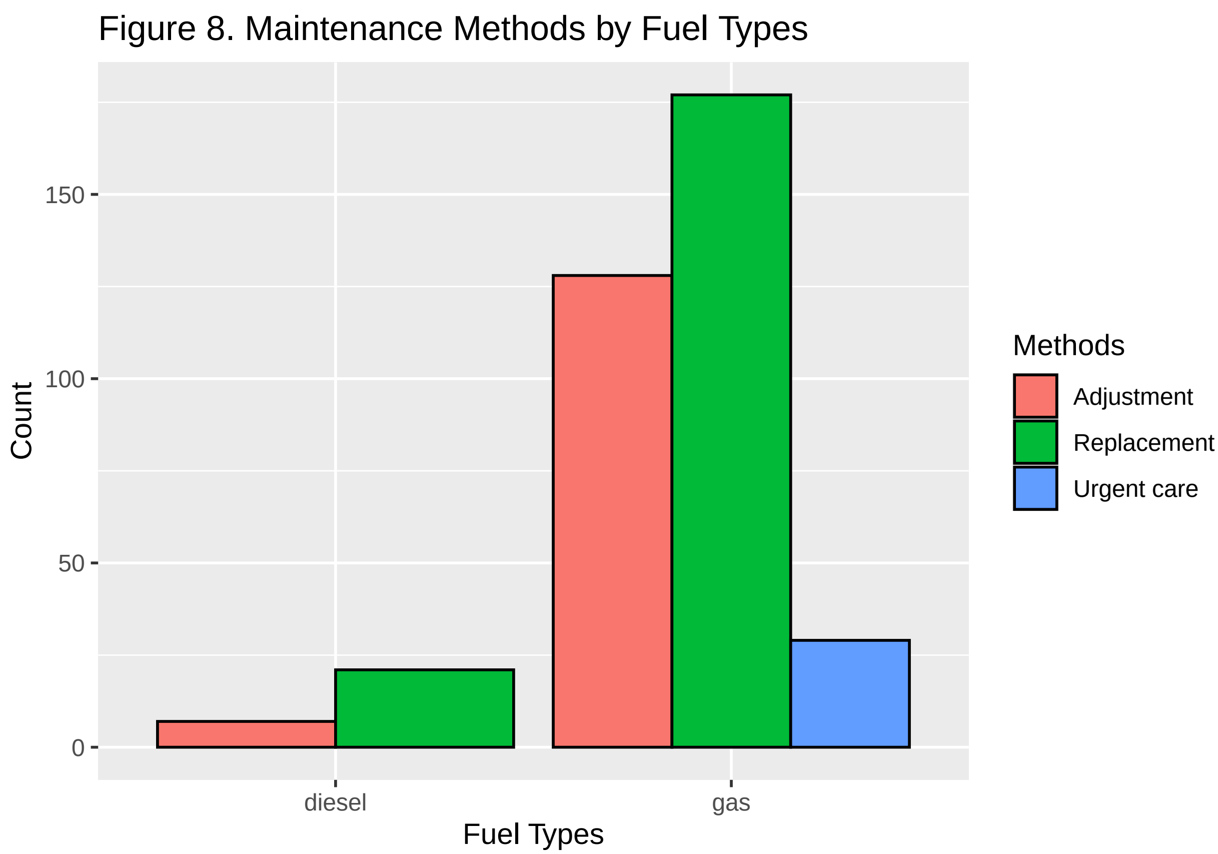
Figure 6. Maintenance Methods by Body Styles



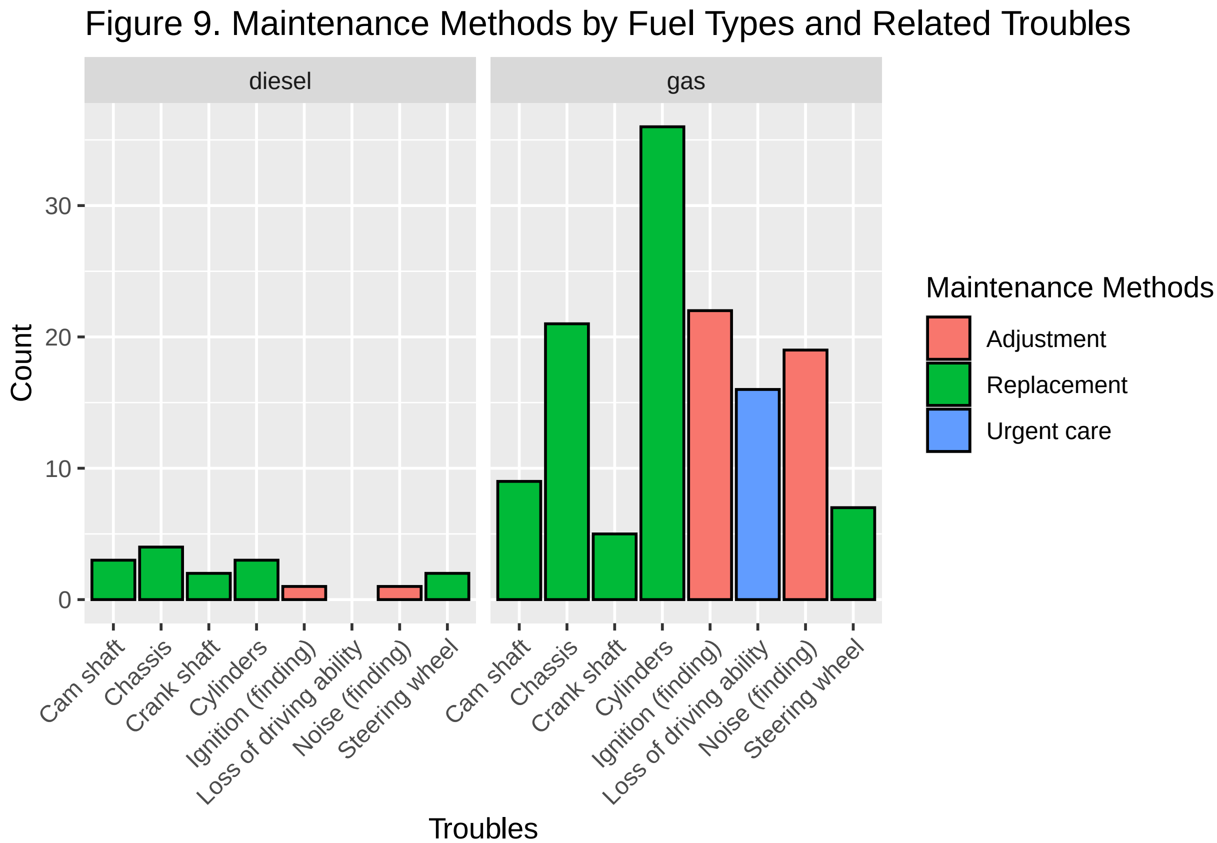
The bar chart illustrates that certain body styles tend to have a higher frequency of three maintenance methods compared to others. As popular car models with four cylinders and mid-range horsepower (between 100 and 125 hp), sedans and hatchbacks may require more frequent maintenance due to their widespread use and flexibility, requiring regular maintenance procedures such as replacements and adjustments. Hardtops and convertibles with lower trouble incidence have fewer maintenance procedures, also emphasizing on the two methods. This implies that the most common approaches often involves changing out and modifying damaged parts (figure 6). The urgent care method only applicable for vehicles that loss of driving ability (figure 7).

4.2. Maintenance by Fuel Types

```
library(dplyr)
library("ggplot2")
FuelMethods <- Troubles_Methods %>%
  group_by(FuelTypes, Methods) %>%
  summarise(count = n(), .groups = "drop")
ggplot(FuelMethods, aes(x = FuelTypes, y = count, fill = Methods)) +
  geom_bar(stat = "identity", position = "dodge", colour = "black") +
  labs(title = "Figure 8. Maintenance Methods by Fuel Types", x = "Fuel Types", y = "Count")
```




```
## ‘summarise()‘ has grouped output by ’FuelTypes’. You can override using the
## ‘.groups‘ argument.
```



"Replacement" and "Adjustment" methods are primarily needed for both gas and diesel engines; however, "Urgent care" methods are not necessary for diesel engines. Gas vehicles show a substantially higher count across all maintenance methods, and may require frequent changes to maintain maximum performance since their systems are created for versatility with daily use and higher speed. This type, often associated with the horsepower of 100 to 125 and 4 cylinders (figure 2), potentially has lower durability, which might experience more sudden and critical failures, such as loss of driving ability (figure 9), requiring immediate attention. In contrast, diesel engines, known for their durability, may require merely part replacements, and adjustments. This fuel type contains lower horsepower with 4 cylinders and below (figure 2), can be adjusted or replaced on a scheduled basis, with fewer unexpected problems requiring urgent care.