Assignment 3

NGUYEN Ngoc Nhu Y

Task 1

```
#Read the dataset and replace "?" with NA
engine <- read.csv("Engine.csv", na.strings = "?")
automobile <- read.csv("Automobile.csv", na.strings = "?")
maintenance <- read.csv("Maintenance.csv", na.strings = "?")
```

```
#Inspect the structure of the dataset "Engine.csv"
str(engine)
```

```
'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  : int  111 154 102 115 110 140 160 101 121 121 ...
 $ FuelTypes   : chr  "gas" "gas" "gas" "gas" ...
 $ Aspiration  : chr  "std" "std" "std" "std" ...
```

```
#Inspect the structure of the dataset "Automobile.csv"
str(automobile)
```

```
'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 ...
```

```
#Inspect the structure of the dataset "Maintenance.csv"
str(maintenance)
```

```
'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" ...
```

```r
#Count total missing values after replacing "?" with NA
cat("Missing values in Engine:", sum(is.na(engine)), "\n")
```

```
Missing values in Engine: 6
```

```r
cat("Missing values in Automobile:", sum(is.na(automobile)), "\n")
```

```
Missing values in Automobile: 0
```

```r
cat("Missing values in Maintenance:", sum(is.na(maintenance)), "\n")
```

```
Missing values in Maintenance: 0
```

```r
# Check which columns contain NA in the Engine dataset
colSums(is.na(engine))
```

```
 EngineModel   EngineType NumCylinders   EngineSize   FuelSystem   Horsepower
           0            5            0            0            0            1
   FuelTypes   Aspiration
           0            0
```

```r
# Count number of rows in Engine affected by missing values
rows_with_na <- sum(rowSums(is.na(engine)) > 0)
cat("Rows with missing values in Engine:", rows_with_na, "\n")
```

```
Rows with missing values in Engine: 6
```

Notes:

Only Engine has missing values "?" (Engine = 6, Automobile & Maintenance = 0).

-> only Engine has affected rows (Missing values occur only in Horsepower)

-> only Engine needs cleaning

When inspecting the column, it is appeared that there are:

- 5 missing entries occur in EngineType
- 1 missing entry occurs in Horsepower

Replacing "?" with NA does not change the distribution because no actual values are modified; NA only marks missing entries.

```r
#Convert categorical variables to factors
automobile$BodyStyles <- as.factor(automobile$BodyStyles)
engine$FuelTypes <- as.factor(engine$FuelTypes)
maintenance$ErrorCodes <- as.factor(maintenance$ErrorCodes)

#Compute the median Horsepower for imputation
median_hp <- median(engine$Horsepower, na.rm = TRUE)

#Replace the missing Horsepower values with median
engine$Horsepower[is.na(engine$Horsepower)] <- median_hp
```

Horsepower is a numerical variable

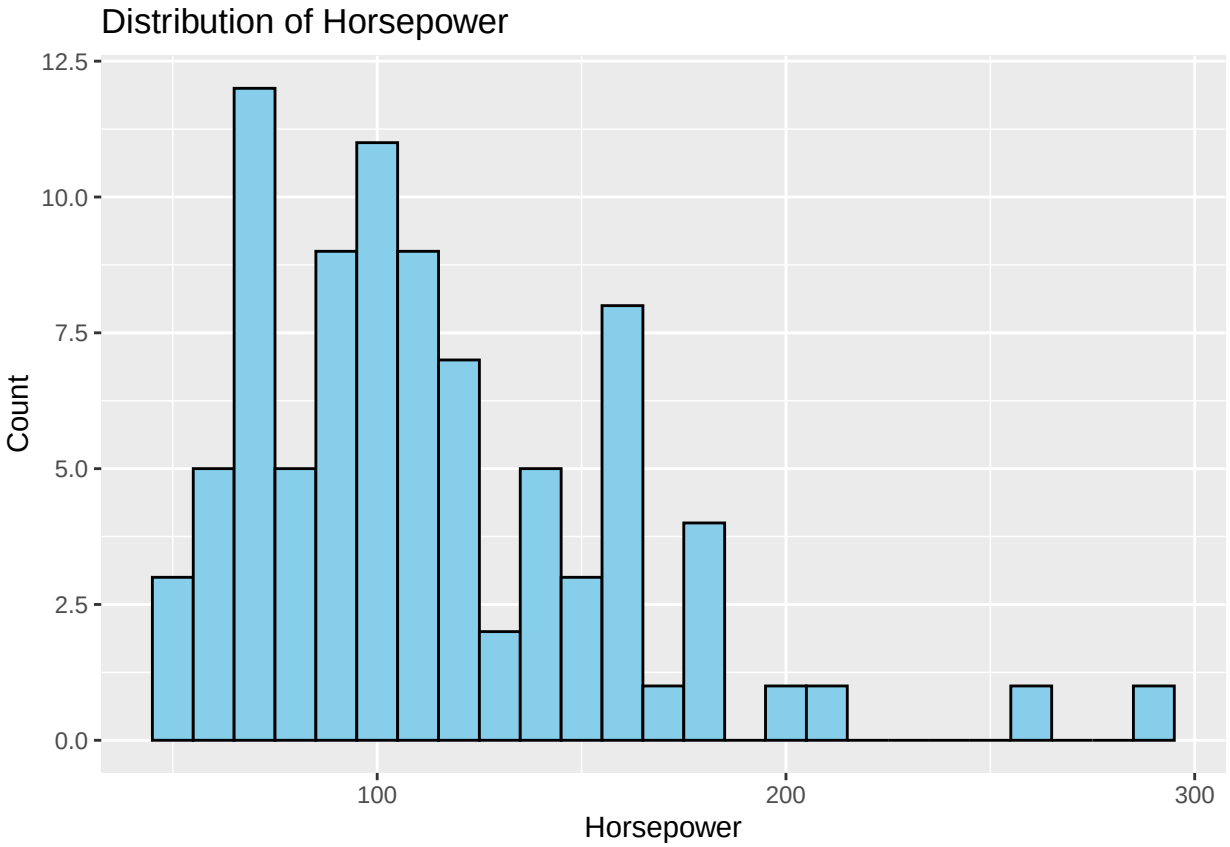
-> use median -> resistant to outliers and preserves the central tendency

Note on EngineType missing values: EngineType is categorical; missing entries were retained as NA and it is only required to replace missing Horsepower values.

Distribution of Horsepower Across Engine Types

```r
# Distribution of Horsepower Across Engine Types
library(ggplot2)

#Visualize Horsepower distribution
ggplot(engine, aes(x = Horsepower)) +
  geom_histogram(binwidth = 10, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Horsepower",
       x = "Horsepower",
       y = "Count")
```



The histogram indicates a strongly right-skewed distribution. The majority of Horsepower values are concentrated between 50 and 170, suggesting that most engines in the dataset belong to small to mid-sized vehicles. However, the long right tail indicates several high-horsepower engines, which act as outliers.

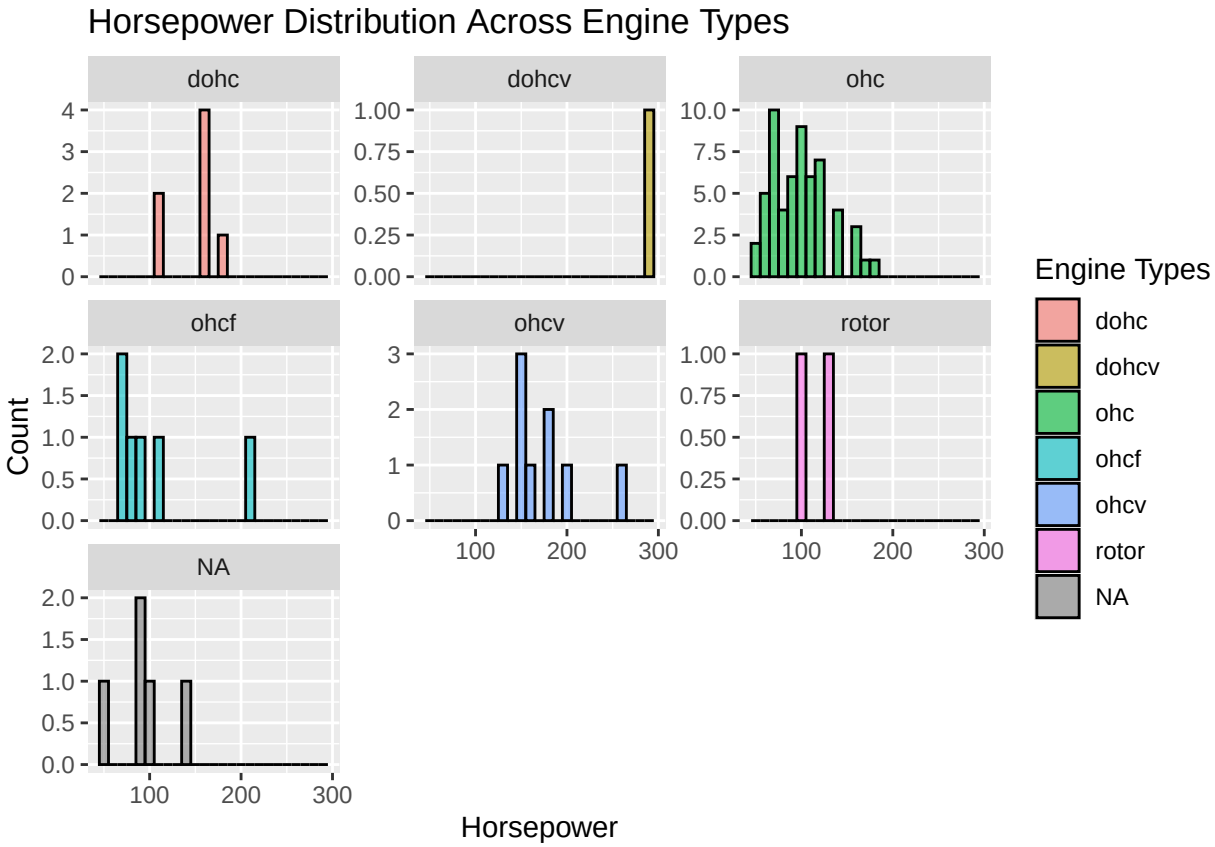
Since the missing values were replaced with the median horsepower, the distribution remains stable and does not shift. This is expected because the median is robust to extreme values and represents the central tendency of the dataset.

Task 2

Horsepower Distribution Across Engine Types

```
# Distribution of Horsepower Across Engine Types
library(ggplot2)

# Histogram of Horsepower for each engine type
ggplot(engine, aes(x = Horsepower, fill = EngineType)) +
  geom_histogram(binwidth = 10, alpha = 0.6, color = "black") +
  facet_wrap(~EngineType, scales="free_y") +
  labs(title = "Horsepower Distribution Across Engine Types",
       x = "Horsepower",
       y = "Count",
       fill = "Engine Types")
```



The distribution of horsepower varies noticeably across different engine types. OHC engines appear most frequently in the dataset and generally fall within the low- to mid-range horsepower levels, indicating that they are typically used in standard consumer vehicles. In contrast, DOHC and OHCV engines display a wider spread and extend into higher horsepower ranges, suggesting that these engine types are associated with stronger performance capabilities. Rotor engines are rare and exhibit a very narrow horsepower distribution, highlighting their limited use in this dataset.

=> The engine type is closely related to the achievable horsepower, with more advanced configurations tending to produce greater power.

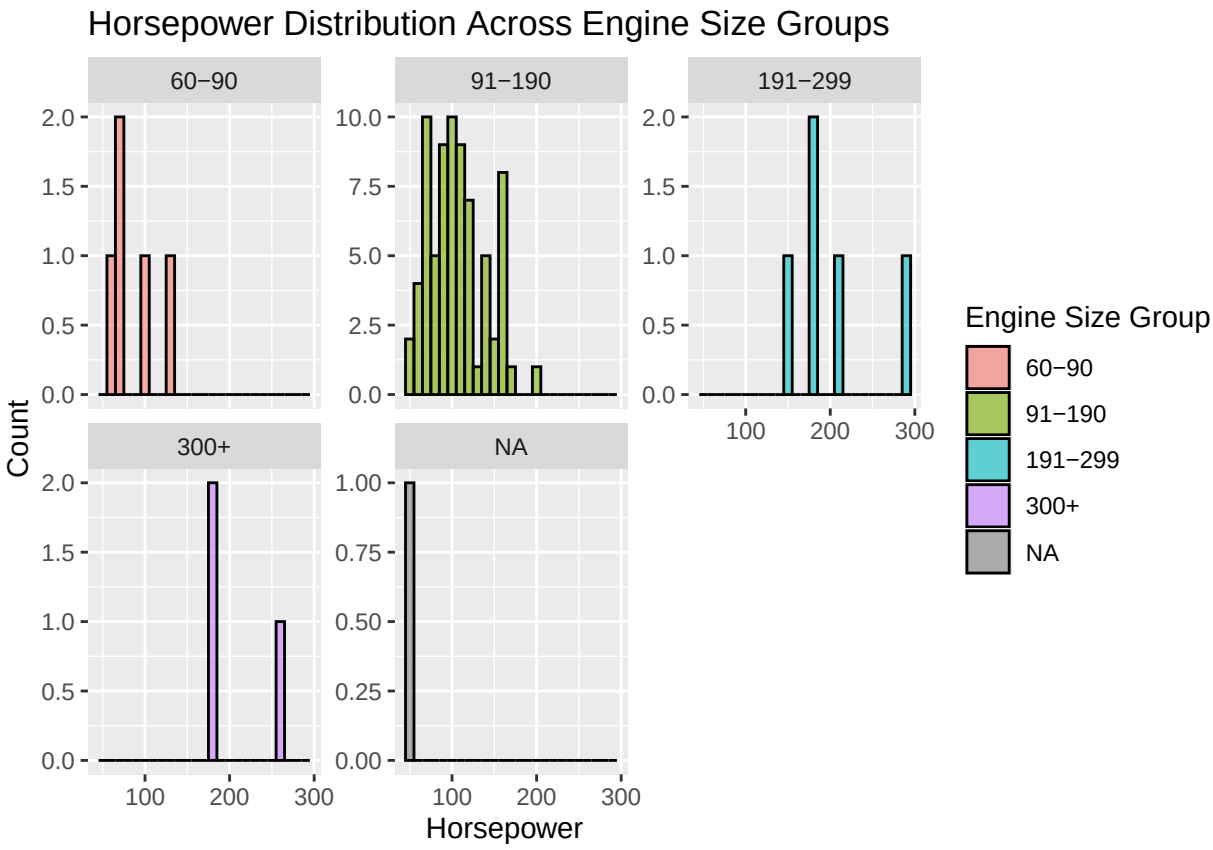
Horsepower Distribution Across Engine Size Groups

```
# Create Engine Size Groups
engine$SizeGroup <- cut(engine$EngineSize,
  breaks = c(60, 90, 190, 299, Inf),
  labels = c("60-90", "91-190", "191-299", "300+"))

#Create table for Engine Size Groups
table (engine$SizeGroup)
```

```
  60-90  91-190 191-299    300+
     5      74       5       3
```

```r
#Facet_wrap Histogram of Horsepower by Size Groups
ggplot(engine, aes(x = Horsepower, fill = SizeGroup)) +
  geom_histogram(binwidth = 10, alpha = 0.6, color = "black") +
  facet_wrap(~ SizeGroup, scales = "free_y") +
  labs(title = "Horsepower Distribution Across Engine Size Groups",
       x = "Horsepower",
       y = "Count",
       fill = "Engine Size Group")
```



When horsepower is examined across the engine size groups, the distribution reveals that the majority of engines belong to the 91-190 size range, which corresponds to mid-sized engines commonly found in everyday vehicles.These engines have a broad horsepower spread but remain mostly within standard performance levels.Smaller engines in the 60-90 group show notably lower horsepower values, reflecting their limited power output.Engines in the 191-299 group shift toward higher horsepower ranges but occur much less frequently, while engines in the 300+ group are extremely rare and represent high-performance or specialized models.

=> Clear relationship between engine size and horsepower, where larger engines tend to support higher power capacity.

Task 3

3.1 Do diesel cars have higher average CityMpg than gasoline cars?

```
t_test_fuel <- t.test(CityMpg ~ FuelTypes,
                      data = merge(automobile, engine, by = "EngineModel"))
t_test_fuel
```

```
    Welch Two Sample t-test

data:  CityMpg by FuelTypes
t = 3.9004, df = 22.592, p-value = 0.0007392
alternative hypothesis: true difference in means between group diesel and group gas is not equal to 0
95 percent confidence interval:
 2.824648 9.218138
sample estimates:
mean in group diesel    mean in group gas
            30.30000             24.27861
```

To determine whether diesel vehicles have higher average CityMpg than gasoline vehicles:

-> The results show a statistically significant difference between the two groups with:

- t = 3.90
- df = 22.6
- p = 0.00074

-> diesel: 30.30 MPG, gasoline: 24.28 MPG

-> Diesel cars have significantly higher CityMpg than gasoline cars

-> The 95% confidence interval (2.82 to 9.22 MPG) confirms that this difference is meaningful

-> Since diesel engines generally operate more efficiently at lower speeds and produce better fuel economy in city driving

Note: *check whether two groups have different averages -> t-test: best choice*

3.2 How does DriveWheels affect fuel efficiency (CityMpg and HighwayMpg)?

```
# Mean CityMpg by DriveWheels
city_means <- aggregate(CityMpg ~ DriveWheels, data = automobile, mean)
city_means
```

```
  DriveWheels  CityMpg
1         4wd 23.11111
2         fwd 28.32500
3         rwd 20.53333
```

```r
# One-way ANOVA for CityMpg
anova_city <- aov(CityMpg ~ DriveWheels, data = automobile)
summary (anova_city)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
DriveWheels   2   2844  1422.1   47.64 <2e-16 ***
Residuals   201   6000    29.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Mean HighwayMpg by DriveWheels
highway_means <- aggregate(HighwayMpg ~ DriveWheels, data = automobile, mean)
highway_means
```

```
  DriveWheels HighwayMpg
1         4wd   27.22222
2         fwd   34.23333
3         rwd   25.64000
```

```r
# One-way ANOVA for HighwayMpg
anova_highway <- aov(HighwayMpg ~ DriveWheels, data = automobile)
summary (anova_highway)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
DriveWheels   2   3526  1763.2   56.77 <2e-16 ***
Residuals   201   6242    31.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To examine whether DriveWheels affects fuel efficiency:

DriveWheels has a significant effect on both CityMpg and HighwayMpg

(p-value $<$ 2e-16 $<$ 0.001)

Besides,

- FWD cars use the **least** fuel, giving them the highest MPG
- 4WD cars use the **most** fuel, resulting in the lowest MPG
- RWD vehicles fall between the two

***Note**: DriveWheels strongly influences fuel economy, and ANOVA is the correct test for comparing three or more groups*

3.3 Filtering engines with trouble and Identifying Top 5 Trouble Types

```r
# Filter trouble-related cases
trouble <- maintenance[maintenance$ErrorCodes != 0, ]

# List top 5 most common troubles
top5_troubles <- sort(table(trouble$Troubles), decreasing = TRUE)[1:5]
top5_troubles
```

```
       Cylinders            Chassis Ignition (finding)    Noise (finding)
              38                    25                    22                    19
      Worn tires
              16
```

Cylinders are the most common troubles (38)

Then Chasis (25), Ignition Issues (22), Noise-related Issues (19), and Worn tires (16)

***Note**: Count how many times each category appears -> frequency table*

3.4 Association Between Engine Type and Trouble Type

```
trouble$HasTrouble <- ifelse(trouble$Troubles == "none" | is.na(trouble$Troubles), 0, 1)

table_simple <- table(trouble$ErrorCodes, trouble$HasTrouble)
chisq.test(table_simple)
```

```
    Chi-squared test for given probabilities

data:  table_simple
X-squared = 174.4, df = 2, p-value < 2.2e-16
```

- Chi-square test -> there is significant association (p-value $<$ 2.2e-16 $<$ 0.001)
- Vehicles with ErrorCodes # 0 -> more likely to have trouble
- ErrorCodes is a useful indicator of suspected issues

***Note**: two categorical variables are connected -> Chi-square test*

Task 4

4.1 Which error type (ErrorCodes) occurs most frequently

```
# Count frequency of each error type
error_counts <- table(maintenance$ErrorCodes)
error_counts
```

```
 -1   0   1
164  28 182
```

- ErrorCode = 1 (confirmed trouble): most frequent
- ErrorCode = –1 (suspected trouble): second
- ErrorCode =0 (no trouble): least frequent

4.2 Analyze the factors that might influence the Maintenance Methods

```
# Merge datasets and filter trouble cases
library(dplyr)

trouble_data <- merge(automobile, engine, by = "EngineModel") %>%
  merge(maintenance, by = "PlateNumber") %>%
  filter(ErrorCodes != 0)
```

```
# FuelTypes vs Methods
table_Fuel_Methods <- table(trouble_data$FuelTypes, trouble_data$Methods)
table_Fuel_Methods
```

```
        Adjustment Replacement Urgent care
  diesel          7          21           0
  gas           128         177          29
```

```
# Chi-squared test
chisq.test(table_Fuel_Methods)
```

```
    Pearson's Chi-squared test

data:  table_Fuel_Methods
X-squared = 5.9481, df = 2, p-value = 0.0511
```

```
#BodyStyles vs Methods
table_Body_Methods <- table(trouble_data$BodyStyles, trouble_data$Methods)
table_Body_Methods
```

```
             Adjustment Replacement Urgent care
  convertible          7           8           0
  hardtop              2           4           2
  hatchback           48          68           9
  sedan               63          94          15
  wagon               15          24           3
```

```
#Chi-squared test
chisq.test(table_Body_Methods)
```

```
    Pearson's Chi-squared test

data:  table_Body_Methods
X-squared = 5.1868, df = 8, p-value = 0.7374
```

```r
#Troubles vs Methods
table_Troubles_Methods <- table(trouble_data$Troubles, trouble_data$Methods)
table_Troubles_Methods
```

```
                        Adjustment Replacement Urgent care
  Air conditioner                9           0           0
  Bearing                        0           3           0
  Brake fluild                   0           0          13
  Break system                   0          10           0
  Cam shaft                      0          12           0
  Chassis                        0          25           0
  Crank shaft                    0           7           0
  Cylinders                      0          39           0
  ECU's power                    8           0           0
  Fans                           0          13           0
  Front axe                      0           3           0
  Gear box (finding)             0           4           0
  Ignition                       9           0           0
  Ignition (finding)            23           0           0
  Loss of driving ability        0           0          16
  Noise (finding)               20           0           0
  O2 sensors                     0           1           0
  Oil filter                     0           4           0
  Painting                       0          15           0
  Pedals                         7           0           0
  Pressure sensors               0          11           0
  Real axe                       0           9           0
  Side slip                     15           0           0
  Steering wheel                 0           9           0
  Stroke                         0           4           0
  Suspected battery             10           0           0
  Suspected clutch              13           0           0
  Suspension                     6           0           0
  Temperature sensors            0           8           0
  Transmission                   0           5           0
  Valve clearance               15           0           0
  Worn tires                     0          16           0
```

```r
#Chi-squared test
chisq.test(table_Troubles_Methods)
```

```
    Pearson's Chi-squared test

data:  table_Troubles_Methods
X-squared = 724, df = 62, p-value < 2.2e-16
```

I chose Troubles, FuelTypes, and BodyStyles to clearly examine how different factors may influence the Maintenance Methods applied to trouble vehicles

Troubles x Methods (Significant)

- Chi-square p-value < 2.2e-16 < 0.05 (strongest significant relationship)
- Clear pattern:
  - Severe issues (Cylinders, Chassis, Cam shaft, Break system) -> mostly Replacement.
  - Minor or diagnostic issues (Noise, Ignition finding, Side slip) -> mostly Adjustment.

=> Type of trouble strongly determines the Maintenance Methods

FuelTypes (Near-significant)

- Chi-square p-value = 0.0511 > 0.05 (but closest to significance among all non-significant factors)
- Clear trend:
  - Diesel vehicles -> higher proportion of Replacement.
  - Gasoline vehicles -> more balanced distribution across methods.

=> Fuel type hints at the underlying stress level on the engine, which affects the likelihood of needing Maintenance Methods (Replacement, Adjustment)

FuelTypes (Not significant)

- Chi-square p-value = 0.7374 > 0.005 -> no statistical relationship

=> The shape or design of the car body does not influence how the vehicle must be repaired