Assignment 1_Dinh Trong Hieu_22145799

2025-04-04

Student Surname: Dinh Trong
Student Firstname: Hieu
Student ID: 22145799
Subject Name: AP

ASSIGNMENT 1 CODE TRANSCRIPTION & EXPLAINATION

Task 1:

```
#Loading in neccessary Library
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
```

```
library(ggplot2)

#TASK 1:
#Data Input as well as replacing "?" appearance with "NA"

engine = read.csv("Engine.csv", na.strings = "?")
automobile = read.csv("Automobile.csv", na.strings = "?")
maintenance = read.csv("Maintenance.csv", na.strings = "?" )

#List out the all of the Loaded column and double check the data type
    str(engine)
```

```
## 'data.frame':    88 obs. of  8 variables:
##  $ EngineModel : chr  "E-0026" "E-0036" "E-0037" "E-0025" ...
##  $ EngineType  : chr  "ohcv" "ohcv" "ohcv" "dohc" ...
##  $ NumCylinders: chr  "twelve" "eight" "eight" "six" ...
##  $ EngineSize  : int  320 308 304 258 234 209 203 194 183 181 ...
##  $ FuelSystem  : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
##  $ Horsepower  : int  262 184 184 176 155 182 288 207 123 200 ...
##  $ FuelTypes   : chr  "gas" "gas" "gas" "gas" ...
##  $ Aspiration  : chr  "std" "std" "std" "std" ...
```

```
    str(automobile)
```

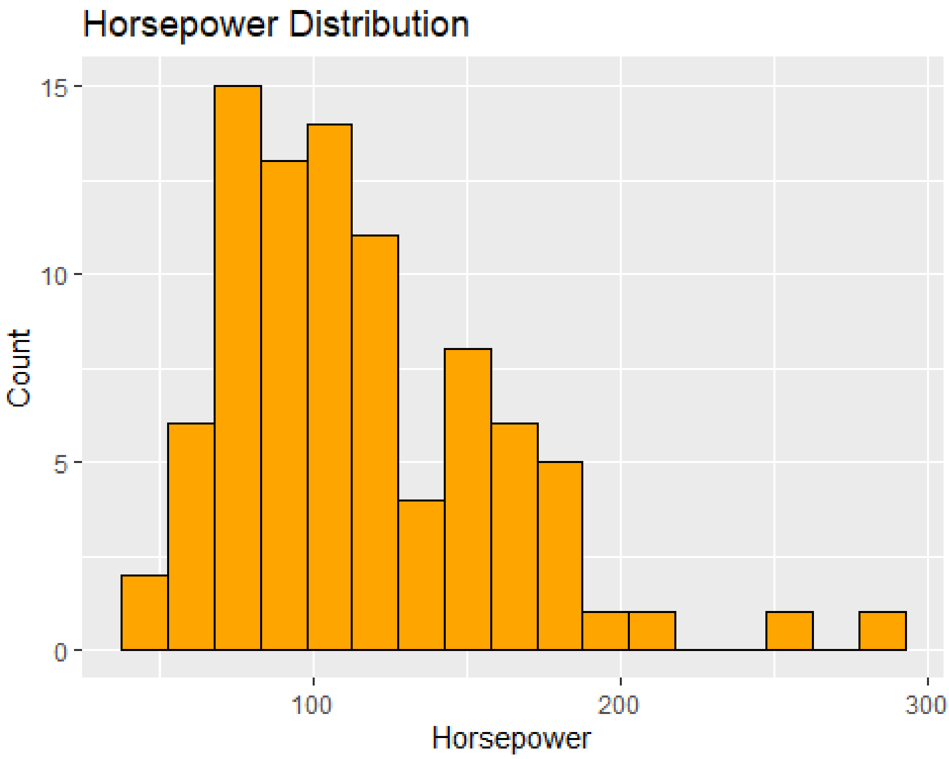
```
## 'data.frame':    204 obs. of  13 variables:
##  $ PlateNumber   : chr  "53N-001" "53N-002" "53N-003" "53N-004" ...
##  $ Manufactures  : chr  "Alfa-romero" "Alfa-romero" "Audi" "Audi" ...
##  $ BodyStyles    : chr  "convertible" "hatchback" "sedan" "sedan" ...
##  $ DriveWheels   : chr  "rwd" "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 ...
```

```
    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"
...
```

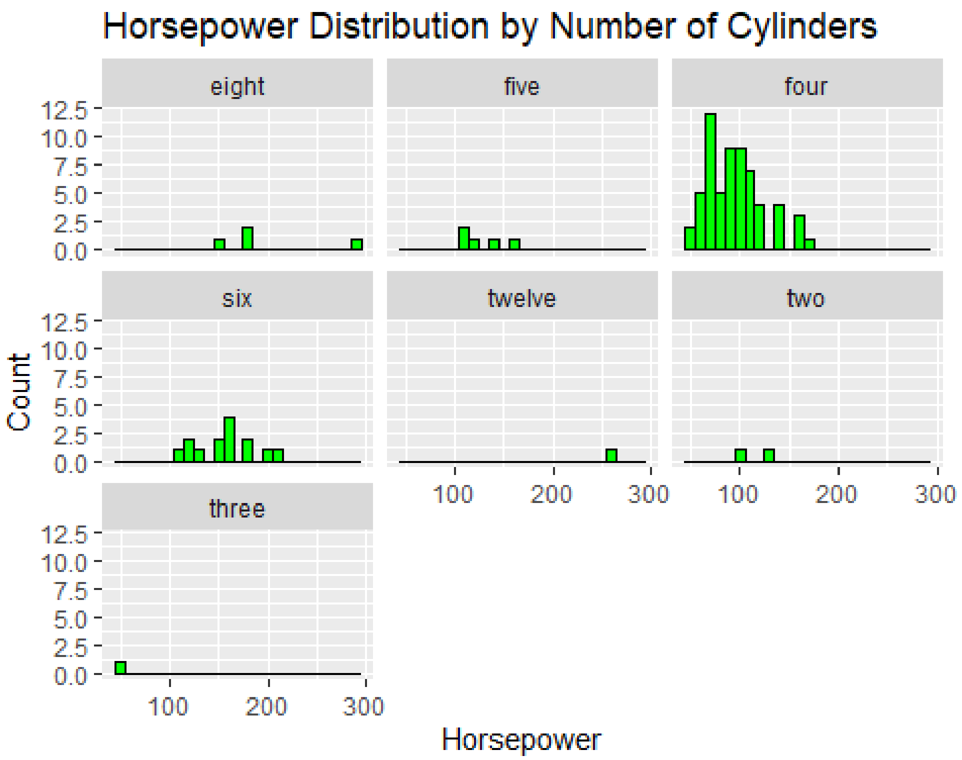
```
#Convert the catergorical variable as factors with dedicated levels
engine$FuelTypes = factor(engine$FuelTypes,
                          levels = c("diesel","gas"))

automobile$BodyStyles = factor(automobile$BodyStyles,
                          levels = c("hardtop","wagon","sedan", "hatchback",
"convertible") )

maintenance$ErrorCodes = factor(maintenance$ErrorCodes,
                        levels = c(1,0,-1),
                        labels = c("engine fails", "no error", "other vehicle
component fails")
                                )
```

```
#Replacing the missing values in the column Horsepower with the mean of
horsepower
mean_hp = mean(engine$Horsepower, na.rm = TRUE)
engine$Horsepower[is.na(engine$Horsepower)] = mean_hp

#illustrating the horsepower distribution through histogram:

ggplot(engine, aes(x = Horsepower)) +
  geom_histogram(binwidth = 15, fill = "orange", color = "black") +
  labs(title = "Horsepower Distribution", x = "Horsepower", y = "Count")
```

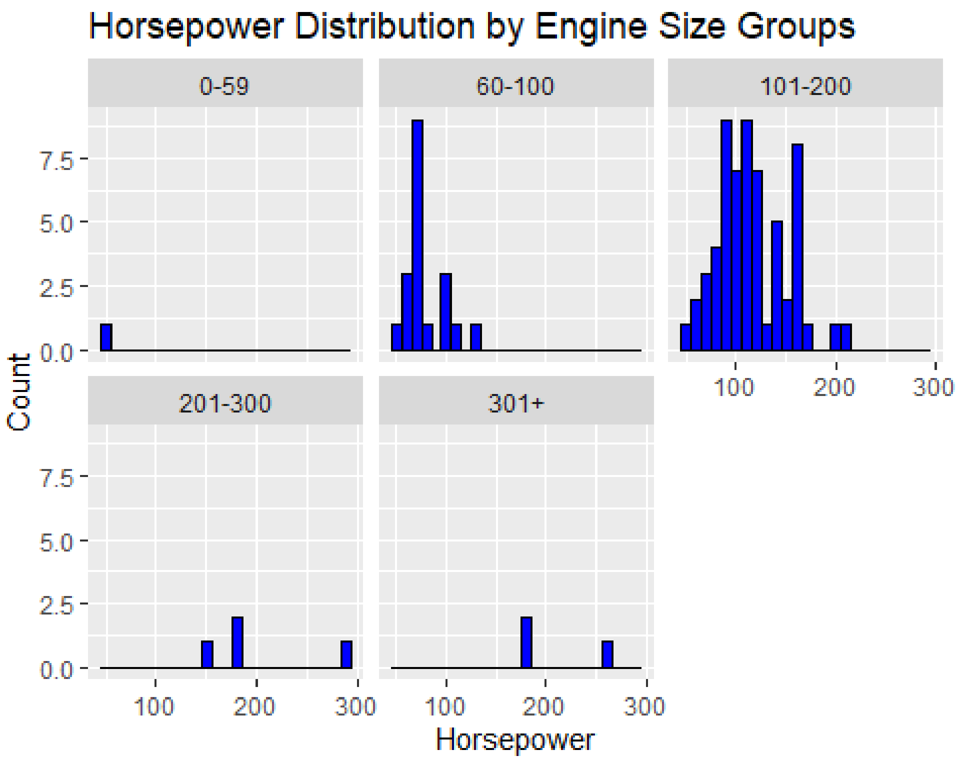


Task 2:

```
#Task 2
#Analyze the distribution of horsepower across the number of cylinders with
Histogram - faucet wrap type of illustration

ggplot(engine, aes(x = Horsepower)) +
  geom_histogram(binwidth = 10, fill = "green", color = "black") +
  facet_wrap(~ NumCylinders) +
  labs(title = "Horsepower Distribution by Number of Cylinders", x =
"Horsepower", y = "Count")
```



```
#Distribution of the horsepower across the groups of the engine sizes with
Histogram - faucet wrap type of illustration
engine$EngineSize <- cut(engine$EngineSize,
                               breaks = c(0,60, 100, 200, 300, Inf),
                               labels = c("0-59","60-100", "101-200", "201-
300", "301+"),
                                right = TRUE, include.lowest = TRUE)
engine$EngineSize <- factor(engine$EngineSize,
                                      levels = c("0-59","60-100", "101-200",
"201-300", "301+"),
                                      ordered = TRUE)
# Visualize horsepower distribution by engine size groups using facet wrap.
ggplot(engine, aes(x = Horsepower)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black") +
  facet_wrap(~ EngineSize) +
  labs(title = "Horsepower Distribution by Engine Size Groups", x =
"Horsepower", y = "Count")
```



Comments on the Findings of the Horsepower Distribution by Number of Cylinders by Histogram:

It could be seen that the variable of Four & Six Cylinders have the highest number counts, substantially higher than others. Furthermore, we can see a trend that the more cylinders that the engine have, the more horsepower that the engine can produce. Only when reaching the highest specs of high-performance vehicles (V8 and V12), there are the exception of a eight cylinders can beat the twelve cylinders

Comments on the Findings of the Horsepower Distribution by Engine Size Groups by Historgram:

Based on the engine sizes group distribution, we can see that most of the engine will have the engine size in the group of over 60 – under 200. Same thing can be seen just like on the horsepower distribution based on the number of cylinders, this histogram pointed out that the bigger the engine size of the vehicle is, it will likely produce bigger horsepower. This is understandable as most of the popular and more budget-friendly brand like “Mazda, Honda, Subaru, Mitsubishi, Chevrolet” are producing their vehicle in this range to prioritize efficiency and to make the car more affordable compared to those more luxury / high performance segment like “Porsche, Mercedes, Audi, BMW”. This current dataset suggested that the current customers of the center are being the mix of middle income (with the more budget-friendly/ family car) and both luxury vehicles, which the current lower-budget segment dominates. With that being said, the Maintenance Center should be based on these insights to have suitable strategic plans / promotion to promote the center to further attract the higher-end (special promotion, all rounded services for luxury cars) and keeping the current main segment develop (by promoting the quality services, authentic components...)

Task 3:

```r
#Task 3
#Filter list to define top 5 most common troubles related

# Merge Maintenance with Automobile data  using the key 'PlateNumber'
maint_auto <- merge(automobile, maintenance, by = "PlateNumber")

# Merge the combined data with Engines data using the key 'EngineModel'
maint_auto_engine <- merge(maint_auto, engine, by = "EngineModel")

# Filter records that report a trouble (which indicates by number "1" or "-1"
in the original dataset)
trouble_records <- maint_auto_engine %>%
  filter(ErrorCodes %in% c("engine fails","other vehicle component fails"))

# For diesel engines: Filter out to select the diesel type of engine then
sort out the top 5 with the highest count rates (n)
top5_diesel <- trouble_records %>%
  filter(ErrorCodes == "engine fails") %>%
  filter(FuelTypes == "diesel") %>%
  count(Troubles, sort = TRUE) %>%
  slice_head(n = 5)

# For gas engines: Filter out to select the gas type of engine then sort out
the top 5 with the highest count rates (n)
top5_gas <- trouble_records %>%
  filter(ErrorCodes == "engine fails") %>%
  filter(FuelTypes == "gas") %>%
  count(Troubles, sort = TRUE) %>%
  slice_head(n = 5)

# Display the results in tables
print("Top 5 Troubles for Diesel Engines:")
```

```
## [1] "Top 5 Troubles for Diesel Engines:"
```

```r
print(top5_diesel)
```

```
##      Troubles n
## 1   Cam shaft 3
## 2   Cylinders 3
## 3 Crank shaft 2
## 4      Stroke 2
## 5 ECU's power 1
```

```
print("Top 5 Troubles for Gas Engines:")

## [1] "Top 5 Troubles for Gas Engines:"

print(top5_gas)

##             Troubles  n
## 1          Cylinders 36
## 2 Ignition (finding) 22
## 3    Noise (finding) 19
## 4    Valve clearance 15
## 5               Fans 13
```

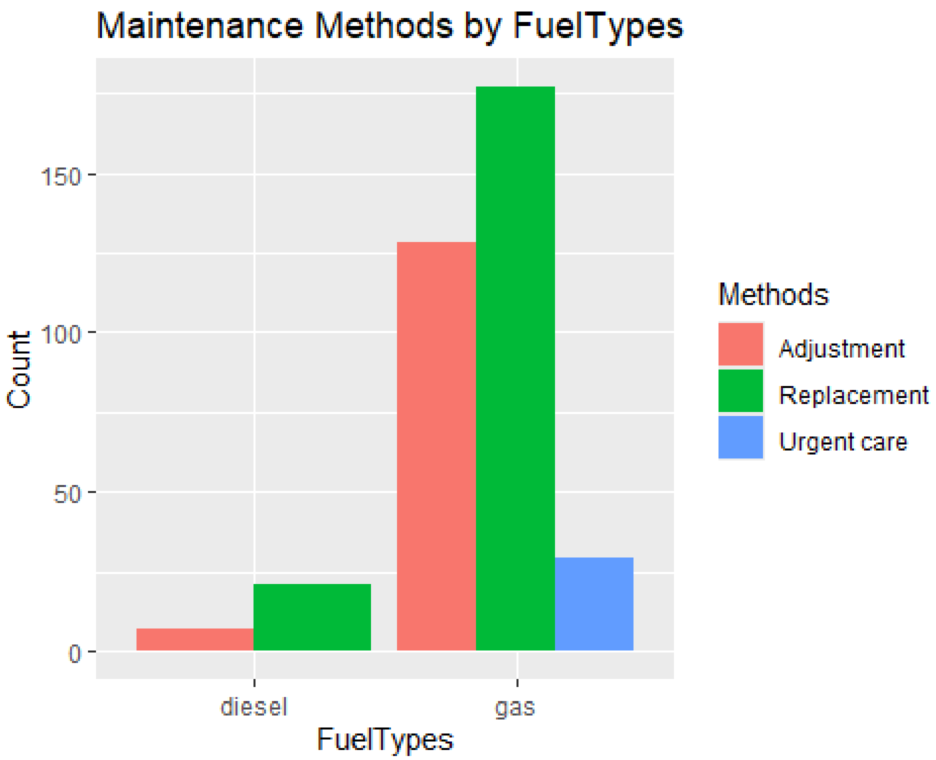
Comments on the Findings:

Looking at the result table of top 5 Troubles the diesel and gas variation both have the Cylinders Trouble in the top 5. This clearly shows that the Cylinder can be considered as the vulnerable part of the engine that can get into problems. Furthermore, we can see a substantial difference between the numbers of diesel and gas engines, pointing out that diesel engines are more durable and have a higher error-free rate. Furthermore, when looking at the distribution rate of gas and diesel, there are much more gas engine type than the diesel engine type. One final notable part is that in the top 2 and top 3 out of the top 5 highest troubles for Gas Engines are still in the finding stage, point out that there is the potential that gas engine might be harder to determine the errors when there is the suspect about errors on the vehicle.

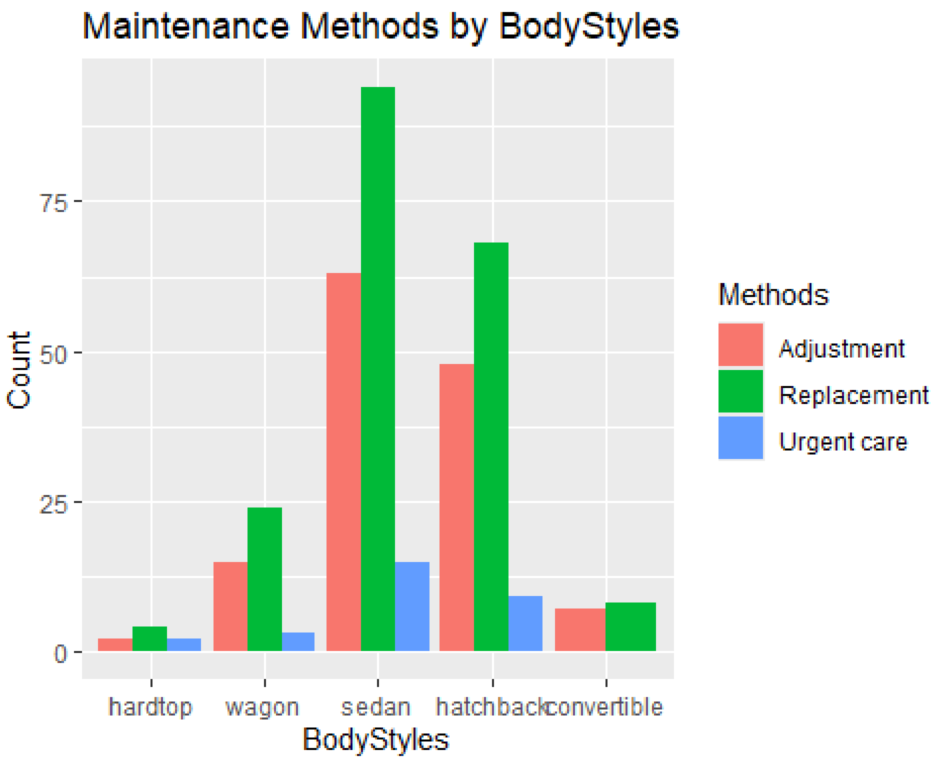
Task 4:

```
#Task 4:
# Filter the merged dataset for vehicles with reported troubles.
trouble_vehicles <- maint_auto_engine %>%
  filter(ErrorCodes %in% c("engine fails","other vehicle component fails"))

#The two selected factors are: FuelTypes and BodyStyles
# --- Factor 1: FuelTypes ---

# Establish a table between Maintenance Methods by FuelTypes
fuel_factor <- table(trouble_vehicles$Methods, trouble_vehicles$FuelTypes)
cat("Table: Maintenance Methods vs FuelTypes")

## Table: Maintenance Methods vs FuelTypes

print(fuel_factor)

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

```r
# Visualize the relationship with a bar plot.
ggplot(trouble_vehicles, aes(x = FuelTypes, fill = Methods)) +
  geom_bar(position = "dodge") +
  labs(title = "Maintenance Methods by FuelTypes", x = "FuelTypes", y =
"Count")
```



```r
# --- Factor 2: BodyStyles ---

# Establish a table between Maintenance Methods by BodyStyles
body_factor <- table(trouble_vehicles$Methods, trouble_vehicles$BodyStyles)
cat("Table: Maintenance Methods vs BodyStyles")

## Table: Maintenance Methods vs BodyStyles

print(body_factor)

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

# Visualize the relationship with a bar plot.
ggplot(trouble_vehicles, aes(x = BodyStyles, fill = Methods)) +
  geom_bar(position = "dodge") +
  labs(title = "Maintenance Methods by BodyStyles", x = "BodyStyles", y =
"Count")
```



Comments on the Findings of the Maintenance Methods counts by Fuel Type Factor:

As expected, the results for maintenance for Gas Engines are significantly higher in terms of numbers and urgent. Only gasoline engines need to be urgently fixed due to vital problems, while none were reported from the diesel engines. Also, the charts pointed out that the replacement for any engine is higher than those that just need to adjust only. Combining with the needs of urgent care for gasoline vehicles, this data suggested that the maintenance center should always prepare the components part ready to stock preparation for replacement scenarios or the need of urgent care.

Comments on the Findings of the Maintenance Methods counts by Fuel Type Factor:

The same key insights pointed out that all of the body styles vehicle have reported that the replacement rate is much higher than the vehicles that just need adjustment only. Also, 4 out of 5 body styles have been reported to have urgent care cases (except Convertible), so in stocks components for any body styles are vital for the maintenance center. Noticeably, the sedan and hatchback dominate the other body styles, further suggesting the center should further concentrate develop the process / suitable human resource training to sedan and hatchback body styles, while keeping the suitable investment for other body styles as well.