Categorical Data Analysis Laboratory Exercise 1

Garridos, Charlene P.

2024-01-16

Contents

1 Data Preparation

1.1 Load The Dataset

```
data(mtcars)
View(mtcars)
```

1.2 Convert Gear Column To A Categorical Variable

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
is.na(mtcars)
```

```
##                     mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear
## Mazda RX4           FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Mazda RX4 Wag       FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Datsun 710          FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Hornet 4 Drive      FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Hornet Sportabout   FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Valiant             FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Duster 360          FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Merc 240D           FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Merc 230            FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Merc 280            FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Merc 280C           FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Merc 450SE          FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Merc 450SL          FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Merc 450SLC         FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Cadillac Fleetwood  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Lincoln Continental FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Chrysler Imperial   FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Fiat 128            FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Honda Civic         FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Toyota Corolla      FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Toyota Corona       FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Dodge Challenger    FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
## AMC Javelin         FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Camaro Z28          FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Pontiac Firebird    FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Fiat X1-9           FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Porsche 914-2       FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Lotus Europa        FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Ford Pantera L      FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Ferrari Dino        FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Maserati Bora       FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## Volvo 142E          FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##                      carb
## Mazda RX4           FALSE
## Mazda RX4 Wag       FALSE
## Datsun 710          FALSE
## Hornet 4 Drive      FALSE
## Hornet Sportabout   FALSE
## Valiant             FALSE
## Duster 360          FALSE
## Merc 240D           FALSE
## Merc 230            FALSE
## Merc 280            FALSE
## Merc 280C           FALSE
## Merc 450SE          FALSE
## Merc 450SL          FALSE
## Merc 450SLC         FALSE
## Cadillac Fleetwood  FALSE
## Lincoln Continental FALSE
## Chrysler Imperial   FALSE
## Fiat 128            FALSE
## Honda Civic         FALSE
## Toyota Corolla      FALSE
## Toyota Corona       FALSE
## Dodge Challenger    FALSE
## AMC Javelin         FALSE
## Camaro Z28          FALSE
## Pontiac Firebird    FALSE
## Fiat X1-9           FALSE
## Porsche 914-2       FALSE
## Lotus Europa        FALSE
## Ford Pantera L      FALSE
## Ferrari Dino        FALSE
## Maserati Bora       FALSE
## Volvo 142E          FALSE
```

```
mtcars$gear
```

```
##  [1] 4 4 4 3 3 3 3 4 4 4 4 3 3 3 3 3 3 4 4 4 3 3 3 3 3 4 5 5 5 5 5 4
```

```
unique(mtcars$gear)
```

```
## [1] 4 3 5
```

```
as.factor(mtcars$gear)
```

```
##  [1] 4 4 4 3 3 3 3 4 4 4 4 3 3 3 3 3 3 4 4 4 3 3 3 3 3 4 5 5 5 5 5 4
## Levels: 3 4 5
```

```
mtcars$gear <- as.factor(mtcars$gear)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

2 Data Exploration

2.1 Summary of The Data

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am         gear        carb      
##  Min.   :0.0000   3:15   Min.   :1.000  
##  1st Qu.:0.0000   4:12   1st Qu.:2.000  
##  Median :0.0000   5: 5   Median :2.000  
##  Mean   :0.4062          Mean   :2.812  
##  3rd Qu.:1.0000          3rd Qu.:4.000  
##  Max.   :1.0000          Max.   :8.000
```

```r
Gear_number <- mtcars$gear

freq <- table(Gear_number)
relative_freq <- prop.table(freq)

result <- data.frame(Gear_number = names(freq), Frequency = as.vector(freq),
                     Relative_Frequency = as.vector(relative_freq))

result
```

```
##   Gear_number Frequency Relative_Frequency
## 1           3        15            0.46875
## 2           4        12            0.37500
## 3           5         5            0.15625
```

```r
mode <- names(freq[freq == max(freq)])
mode
```

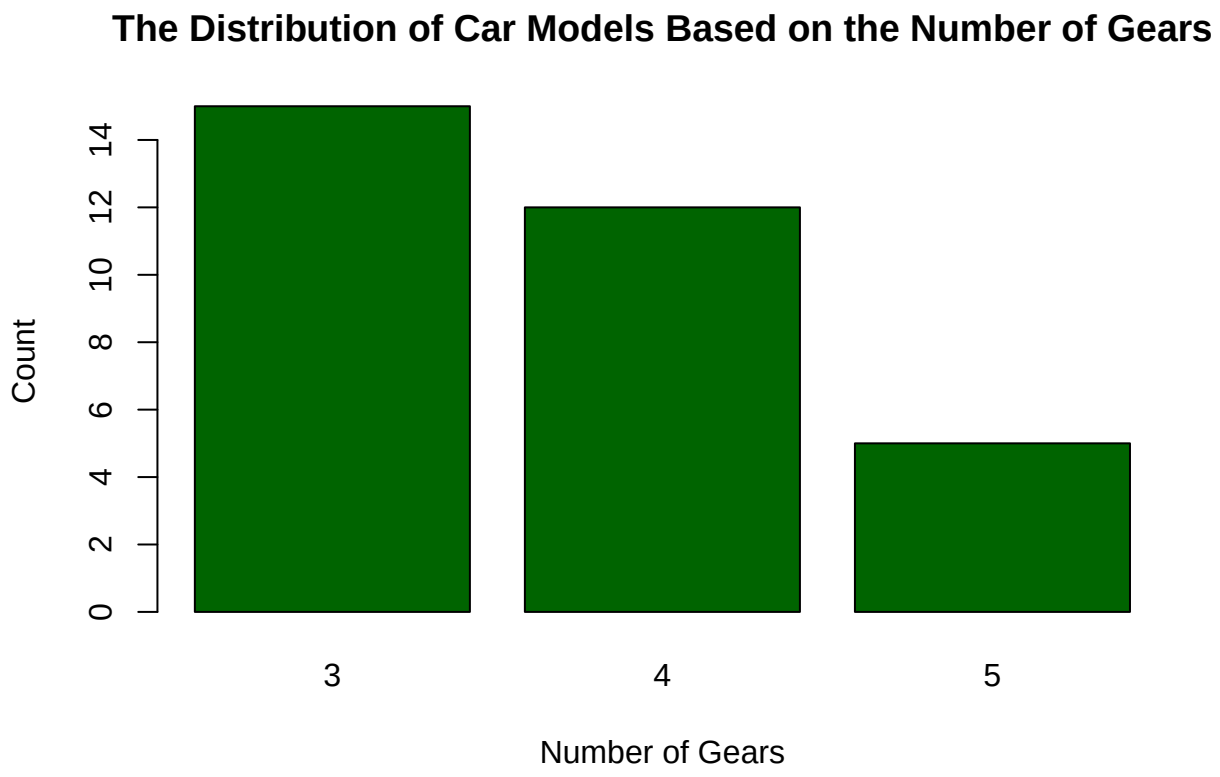
```
## [1] "3"
```

Interpretation

The results show that out of 32 car models, 15 have three gears. On the other hand, only 5 of the cars have 5 gears, and the rest have 4 gears. This means that the common (mode) number of gears is 3 gears.

2.2 Visualization

```
plot(Gear_number, xlab = "Number of Gears",
     ylab = "Count",
     main = "The Distribution of Car Models Based on the Number of Gears",
     col = "dark green")
```



Interpretation

The figure above shows the distribution of 32 automobiles based on the number of gears. There are three categories: '3', '4', and '5'. Fifteen car models have three gears. Out of 35, twelve automobiles have four gears. Lastly, only 5 cars have 5 gears.

3 Association Test

3.1 Convert Transmission Column To A Categorical Variable

```
mtcars$am <- as.factor(mtcars$am)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
```

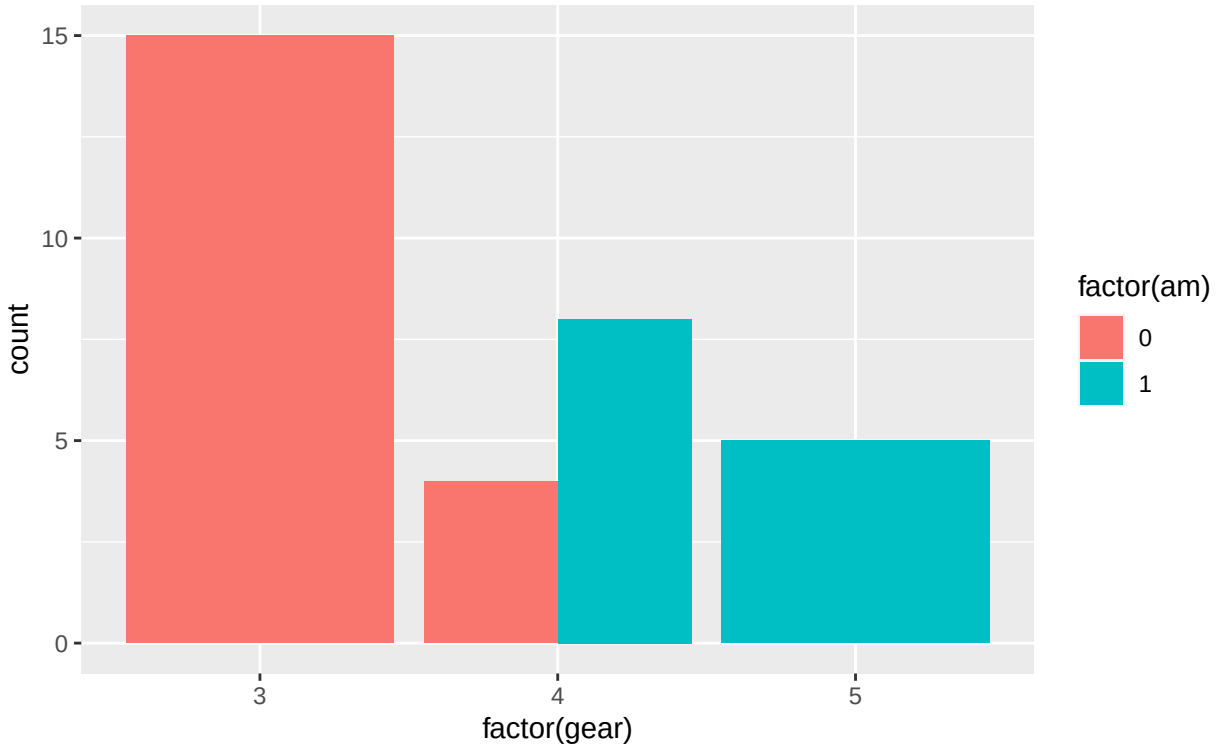
```
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

3.2 Contingency table

```
Gear_number <- mtcars$gear
Transmission_type <- mtcars$am
table(Gear_number, Transmission_type, dnn = c("Number of Gears", "Type of Transmission"))
```

```
##                Type of Transmission
## Number of Gears  0  1
##               3 15  0
##               4  4  8
##               5  0  5
```

```
library(ggplot2)
ggplot(mtcars, aes(factor(gear), fill = factor(am))) +
  geom_bar(position = "dodge")
```



Interpretation

The contingency table and the graph displays the number of automobiles based on two variables: transmission type and number of gears. As previously stated, the number of gears is classified into three groups. As for the transmission, there are two types, which are 0 for automatic and 1 for manual.

Fifteen cars are automatic and belong to the category of 3 gears. As for the cars that have 4 gears, out of 12, eight manual and four are automatic. In terms of having 5 gears, all five cars are manual.

3.3 Chi-Square Test for Independence

- H_O: There is no relationship between the type of gear and type of transmission.
- H_1: There is relationship between the number of gear and type of transmission.
- Level of Significance, $\alpha = 0.05$

```
Chisq_test <- chisq.test(Gear_number, Transmission_type)
Chisq_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  Gear_number and Transmission_type
## X-squared = 20.945, df = 2, p-value = 2.831e-05
```

Interpretation of the Results

The p-value is 2.831e-05, which is less than the significant level of 0.05. Hence, there is enough evidence to reject the null hypothesis and accept the alternative hypothesis. Therefore, this suggest a significant relationship between the number of gears and the type of transmission.

4 Data Interpretation

4.1 Significant Association

In the statistical output that is mentioned previously, fifteen cars that have three gears are all automatic. Out of the 12 automobiles with four gears, eight are manual and four are automatic. Additionally, all five cars that have five gears are all manual.

Based on the data, the association between the number of gears and the type of transmission is that if the car is automatic, a lesser number of gears are required. Given that it is the appropriate number for the appropriate amount of engine power that goes to the wheels to drive at any given speed and because automatic transmissions have a torque converter. Furthermore, if the type of transmission of the car is manual then it requires a larger number of gears. This explains why all the cars with 3 gears are all automatic and all the cars with 5 gears are all manual. Thus, there is significant association between the number of gears and the type of transmission.

4.2 Reflection

Aside from the fact that it tests the relation or association of one variable to another and how other variables affect each other, it is also beneficial as a guide in making decisions by understanding the behavior of the data. For instance, since the number of gears and the type of transmission are related, the company wants to know if a particular gear number is more suitable for certain transmission types. The company can adjust the production based on the results if the number of gears cannot suffice for this certain transmission type. To elaborate more, if decreasing the number of gears to two will result in a much faster speed than having three gears, then the particular gear number for a certain transmission type is useful for making production plans and manufacturing products with better features. This information is useful for advertising. Furthermore, the company can also expand its study to develop a much more efficient product that is sellable to buyers. Supported by science and strong analysis, the company can market their car products to the public, which will not only attract buyers but also investors.

5 Reference

- https://plotnine.readthedocs.io/en/stable/generated/plotnine.data.mtcars.html
- https://www.geeksforgeeks.org/contingency-tables-in-r-programming/
- https://statsandr.com/blog/chi-square-test-of-independence-in-r/