

Assignment-1

Execute the following cell to load the tidyverse library:

```
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2    3.5.1      v tibble    3.2.1
## v lubridate  1.9.3      v tidyr     1.3.1
## v purrr      1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Execute the following cell to load the data. Refer to this website <http://archive.ics.uci.edu/ml/datasets/Auto+MPG> for details on the dataset:

```
autompg = read.table(
  "http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",
  quote = "\"",
  comment.char = "",
  stringsAsFactors = FALSE)
head(autompg,20)
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9
## 1	18	8	307	130.0	3504	12.0	70	1	chevrolet chevelle malibu
## 2	15	8	350	165.0	3693	11.5	70	1	buick skylark 320
## 3	18	8	318	150.0	3436	11.0	70	1	plymouth satellite
## 4	16	8	304	150.0	3433	12.0	70	1	amc rebel sst
## 5	17	8	302	140.0	3449	10.5	70	1	ford torino
## 6	15	8	429	198.0	4341	10.0	70	1	ford galaxie 500
## 7	14	8	454	220.0	4354	9.0	70	1	chevrolet impala
## 8	14	8	440	215.0	4312	8.5	70	1	plymouth fury iii
## 9	14	8	455	225.0	4425	10.0	70	1	pontiac catalina
## 10	15	8	390	190.0	3850	8.5	70	1	amc ambassador dpl
## 11	15	8	383	170.0	3563	10.0	70	1	dodge challenger se
## 12	14	8	340	160.0	3609	8.0	70	1	plymouth 'cuda 340
## 13	15	8	400	150.0	3761	9.5	70	1	chevrolet monte carlo
## 14	14	8	455	225.0	3086	10.0	70	1	buick estate wagon (sw)
## 15	24	4	113	95.00	2372	15.0	70	3	toyota corona mark ii
## 16	22	6	198	95.00	2833	15.5	70	1	plymouth duster
## 17	18	6	199	97.00	2774	15.5	70	1	amc hornet
## 18	21	6	200	85.00	2587	16.0	70	1	ford maverick
## 19	27	4	97	88.00	2130	14.5	70	3	datsum pl510
## 20	26	4	97	46.00	1835	20.5	70	2	volkswagen 1131 deluxe sedan

Question 1.1: print the structure of the unedited data set. How many samples and features are there? Ans
- 398 samples and 9 features.

```
str(autopmg)
```

```
## 'data.frame':   398 obs. of  9 variables:
## $ V1: num  18 15 18 16 17 15 14 14 14 15 ...
## $ V2: int   8 8 8 8 8 8 8 8 8 8 ...
## $ V3: num  307 350 318 304 302 429 454 440 455 390 ...
## $ V4: chr  "130.0" "165.0" "150.0" "150.0" ...
## $ V5: num  3504 3693 3436 3433 3449 ...
## $ V6: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ V7: int  70 70 70 70 70 70 70 70 70 70 ...
## $ V8: int   1 1 1 1 1 1 1 1 1 1 ...
## $ V9: chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel sst" ...
```

Execute the following cell to assign names to the columns of the dataframe:

```
colnames(autopmg) = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year", "origin", "name")
str(autopmg) # Validate the result
```

```
## 'data.frame':   398 obs. of  9 variables:
## $ mpg   : num  18 15 18 16 17 15 14 14 14 15 ...
## $ cyl   : int   8 8 8 8 8 8 8 8 8 8 ...
## $ disp  : num  307 350 318 304 302 429 454 440 455 390 ...
## $ hp    : chr  "130.0" "165.0" "150.0" "150.0" ...
## $ wt    : num  3504 3693 3436 3433 3449 ...
## $ acc   : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ year  : int  70 70 70 70 70 70 70 70 70 70 ...
## $ origin: int   1 1 1 1 1 1 1 1 1 1 ...
## $ name  : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel sst"
```

Question 1.2: complete the code segment below to remove samples with missing horsepower (hp) values represented as a “?” in the dataset.

```
autopmg = autopmg %>% filter(!(hp == '?'))
autopmg %>% filter(hp == '?') # Validate the result
```

```
## [1] mpg   cyl   disp  hp    wt     acc   year   origin name
## <0 rows> (or 0-length row.names)
```

Question 1.3: complete the code segment below to remove samples with the name “plymouth reliant”

```
autopmg = autopmg %>% filter(!(name == 'plymouth reliant'))
autopmg %>% filter(name == 'plymouth reliant') # Validate the result
```

```
## [1] mpg   cyl   disp  hp    wt     acc   year   origin name
## <0 rows> (or 0-length row.names)
```

Question 2.1: complete the code segment below to select all features except ‘name’

```

autompg = autompg %>% select(-name)
str(autompg) # Validate the result

```

```

## 'data.frame': 390 obs. of 8 variables:
## $ mpg : num 18 15 18 16 17 15 14 14 14 15 ...
## $ cyl : int 8 8 8 8 8 8 8 8 8 8 ...
## $ disp : num 307 350 318 304 302 429 454 440 455 390 ...
## $ hp : chr "130.0" "165.0" "150.0" "150.0" ...
## $ wt : num 3504 3693 3436 3433 3449 ...
## $ acc : num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ year : int 70 70 70 70 70 70 70 70 70 70 ...
## $ origin: int 1 1 1 1 1 1 1 1 1 1 ...

```

Execute the following cell to change the type of hp values from character to numeric:

```

autompg$hp = as.numeric(autompg$hp)
str(autompg) # Validate that type has changed from chr to num for hp

```

```

## 'data.frame': 390 obs. of 8 variables:
## $ mpg : num 18 15 18 16 17 15 14 14 14 15 ...
## $ cyl : int 8 8 8 8 8 8 8 8 8 8 ...
## $ disp : num 307 350 318 304 302 429 454 440 455 390 ...
## $ hp : num 130 165 150 150 140 198 220 215 225 190 ...
## $ wt : num 3504 3693 3436 3433 3449 ...
## $ acc : num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ year : int 70 70 70 70 70 70 70 70 70 70 ...
## $ origin: int 1 1 1 1 1 1 1 1 1 1 ...

```

Question 2.2: complete the code cell to modify 'origin' column to reflect local (1) and international models (0)

```

autompg = autompg %>% mutate(origin = ifelse(origin == 1, 'local', 'international'))
head(autompg, 20)

```

```

##   mpg cyl disp  hp  wt  acc year    origin
## 1   18   8  307 130 3504 12.0   70    local
## 2   15   8  350 165 3693 11.5   70    local
## 3   18   8  318 150 3436 11.0   70    local
## 4   16   8  304 150 3433 12.0   70    local
## 5   17   8  302 140 3449 10.5   70    local
## 6   15   8  429 198 4341 10.0   70    local
## 7   14   8  454 220 4354   9.0   70    local
## 8   14   8  440 215 4312   8.5   70    local
## 9   14   8  455 225 4425 10.0   70    local
## 10  15   8  390 190 3850   8.5   70    local
## 11  15   8  383 170 3563 10.0   70    local
## 12  14   8  340 160 3609   8.0   70    local
## 13  15   8  400 150 3761   9.5   70    local
## 14  14   8  455 225 3086 10.0   70    local
## 15  24   4  113  95 2372 15.0   70 international
## 16  22   6  198  95 2833 15.5   70    local

```

```
## 17 18 6 199 97 2774 15.5 70 local
## 18 21 6 200 85 2587 16.0 70 local
## 19 27 4 97 88 2130 14.5 70 international
## 20 26 4 97 46 1835 20.5 70 international
```

Question 2.3: print the structure of the dataframe. What types are the columns 'cyl' and 'origin'? Ans - 'cyl' is type int and 'origin' is type chr.

```
str(autompg)
```

```
## 'data.frame': 390 obs. of 8 variables:
## $ mpg : num 18 15 18 16 17 15 14 14 14 15 ...
## $ cyl : int 8 8 8 8 8 8 8 8 8 8 ...
## $ disp : num 307 350 318 304 302 429 454 440 455 390 ...
## $ hp : num 130 165 150 150 140 198 220 215 225 190 ...
## $ wt : num 3504 3693 3436 3433 3449 ...
## $ acc : num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ year : int 70 70 70 70 70 70 70 70 70 70 ...
## $ origin: chr "local" "local" "local" "local" ...
```

Question 3.1: complete the code segment below to change the types of 'cyl' and 'origin' columns to factor

```
catcols = c('cyl', 'origin')
autompg[catcols] = lapply(autompg[catcols], as.factor)
str(autompg) # Validate the change
```

```
## 'data.frame': 390 obs. of 8 variables:
## $ mpg : num 18 15 18 16 17 15 14 14 14 15 ...
## $ cyl : Factor w/ 5 levels "3","4","5","6",...: 5 5 5 5 5 5 5 5 5 5 ...
## $ disp : num 307 350 318 304 302 429 454 440 455 390 ...
## $ hp : num 130 165 150 150 140 198 220 215 225 190 ...
## $ wt : num 3504 3693 3436 3433 3449 ...
## $ acc : num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ year : int 70 70 70 70 70 70 70 70 70 70 ...
## $ origin: Factor w/ 2 levels "international",...: 2 2 2 2 2 2 2 2 2 2 ...
```

Question 3.2: complete the code segment below to create a scatter plot of mpg vs. displacement by color coding the points according to the origin (local or international). Add axes labels and title for the plot. Comment on what you observe: mpg and displacement are inversely proportional. As displacement increases, miles per gallon decreases and vice-versa.

```
p = ggplot(data = autompg, aes(x = mpg, y = disp, color = wt)) +
  geom_point() +
  labs(x = 'Miles Per Gallon', y = 'Displacement', title = 'Miles Per Gallon vs Displacement')
p
```

Miles Per Gallon vs Displacement

