Problem Set 1 - ARE 212 - Anna Cheyette

2024-01-23

```
library(haven)
```

```
## Warning: package 'haven' was built under R version 4.3.1
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.2      v readr     2.1.4
## v forcats   1.0.0      v stringr   1.5.0
## v ggplot2   3.4.2      v tibble    3.2.1
## v lubridate 1.9.2      v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
```

```r
# get directory of current file
current_directory <- dirname(dirname(rstudioapi::getSourceEditorContext()$path))

# read in data
carsdf <- read_dta(file.path(current_directory, "Data",
                             "pset1_2024.dta"))
```

Summarize data

The data in column qu has a very wide range. The minimum value is 294, while the max value is 351,477. This is a much wider range than any of the other variables. The data is also heavily right-skewed.

```r
# work with data
summary(carsdf)
```

```
##     domestic             qu            price         horsepower
##  Min.   :0.0000   Min.   :   294   Min.   : 7.938   Min.   : 13.00
##  1st Qu.:0.0000   1st Qu.:  1782   1st Qu.:14.539   1st Qu.: 34.12
##  Median :0.0000   Median :  4794   Median :19.004   Median : 51.50
##  Mean   :0.2544   Mean   : 23050   Mean   :21.032   Mean   : 53.43
##  3rd Qu.:0.7500   3rd Qu.: 20498   3rd Qu.:24.568   3rd Qu.: 66.75
##  Max.   :1.0000   Max.   :351477   Max.   :50.373   Max.   :118.00
##      fuel             width           height           weight
##  Min.   : 5.300   Min.   :132.0   Min.   :120.0   Min.   : 520.0
##  1st Qu.: 7.125   1st Qu.:153.0   1st Qu.:139.1   1st Qu.: 760.0
##  Median : 8.800   Median :159.0   Median :142.0   Median : 925.0
##  Mean   : 8.810   Mean   :160.0   Mean   :141.9   Mean   : 928.1
##  3rd Qu.:10.000   3rd Qu.:165.0   3rd Qu.:145.0   3rd Qu.:1063.8
##  Max.   :15.000   Max.   :180.5   Max.   :155.0   Max.   :1525.0
##      luxury
```
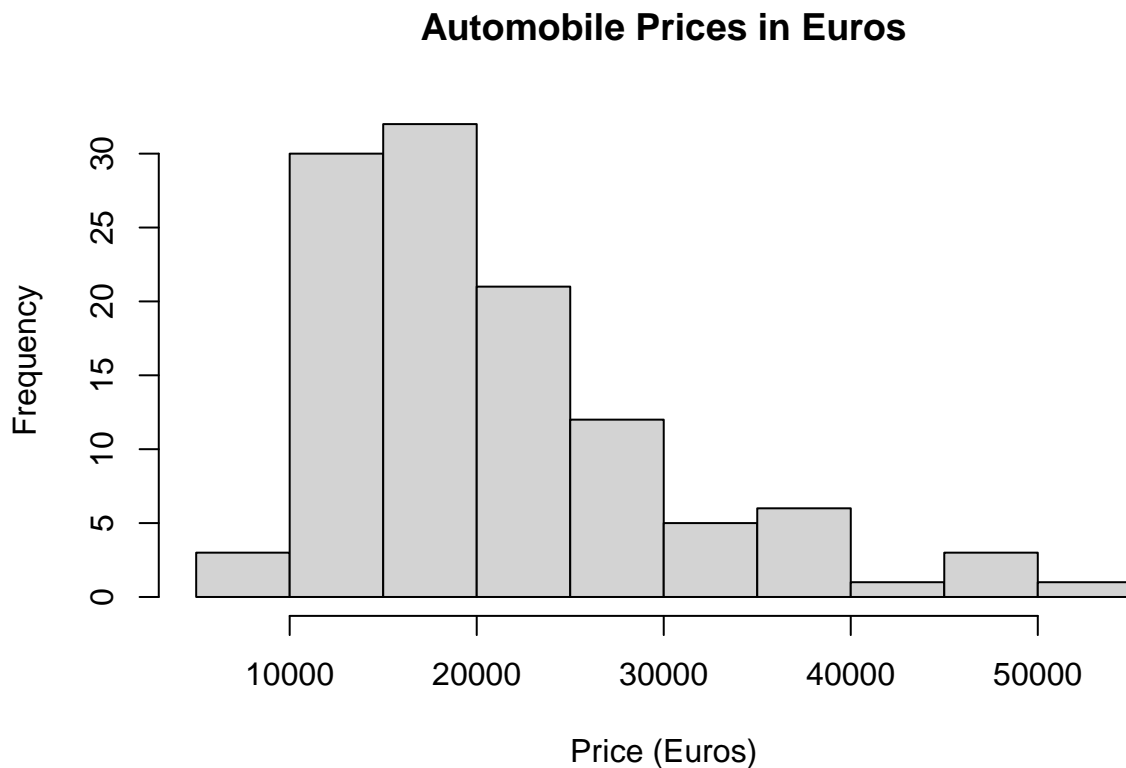
```
##  Min.    :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.05263
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

Create a new variable

```
carsdf <-
  carsdf %>%
  mutate(price_euros = price*1000)
```

Make histogram of prices

```
# histogram of price in euros
price_hist <- hist(carsdf$price_euros,
                   main = "Automobile Prices in Euros",
                   xlab = "Price (Euros)")
```

Automobile Prices in Euros


```
# Open a blank PDF
pdf(file.path(current_directory,
              "Output",
              "price2024khist.pdf"))
# Create a histogram (this will go into the blank PDF)
price_hist
```

```
## $breaks
##  [1]  5000 10000 15000 20000 25000 30000 35000 40000 45000 50000 55000
```

```
## 
## $counts
##  [1]  3 30 32 21 12  5  6  1  3  1
## 
## $density
##  [1] 5.263158e-06 5.263158e-05 5.614035e-05 3.684211e-05 2.105263e-05
##  [6] 8.771930e-06 1.052632e-05 1.754386e-06 5.263158e-06 1.754386e-06
## 
## $mids
##  [1]  7500 12500 17500 22500 27500 32500 37500 42500 47500 52500
## 
## $xname
## [1] "carsdf$price_euros"
## 
## $equidist
## [1] TRUE
## 
## attr(,"class")
## [1] "histogram"
```

```
# Close the PDF
dev.off()
```

```
## pdf
##   2
```

Make boxplot of weights

```
# boxplot of weight
weight_boxplot <- boxplot(carsdf$weight, main = "Car Weight Boxplot")
```

Car Weight Boxplot
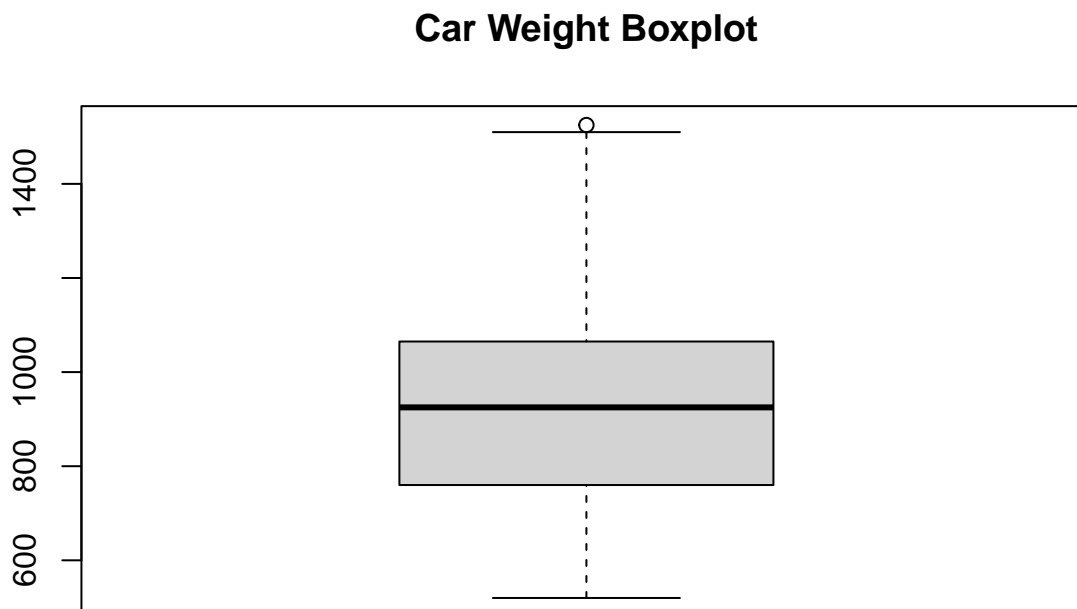
Write updated data to file

```
write.csv(carsdf,
          file.path(current_directory, "Data",
                    "temp2024.csv"))
```