ml

February 9, 2023

2_R_explore_practice.org

Getting the data

Download the raw file `usedcars.csv` from this location and put it into a dataframe `usedcars`: tinyurl.com/yp5r3kw7

```
## Did you check if the CSV file was actually CSV and had a header?
usedcars <- read.csv(file="https://tinyurl.com/yp5r3kw7")
str(usedcars)
```

```
'data.frame': 150 obs. of  6 variables:
 $ year        : int  2011 2011 2011 2011 2012 2010 2011 2010 2011 2010 ...
 $ model       : chr  "SEL" "SEL" "SEL" "SEL" ...
 $ price       : int  21992 20995 19995 17809 17500 17495 17000 16995 16995 16995 ...
 $ mileage     : int  7413 10926 7351 11613 8367 25125 27393 21026 32655 36116 ...
 $ color       : chr  "Yellow" "Gray" "Silver" "Gray" ...
 $ transmission: chr  "AUTO" "AUTO" "AUTO" "AUTO" ...
```

Exploring the structure of numerical data

1. Display common summary statistics for the `year` variable of the `usedcars` dataset to find out when the vehicle records commenced, and when they were posted.

   ```
   summary(usedcars$year)
   ```

   ```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2000    2008    2009    2009    2010    2012
   ```

2. Based on the `price`, what kind of cars dominated the listing?

```
summary(usedcars$price) # economy/mid-range cars dominate
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3800   10995   13592   12962   14904   21992
```

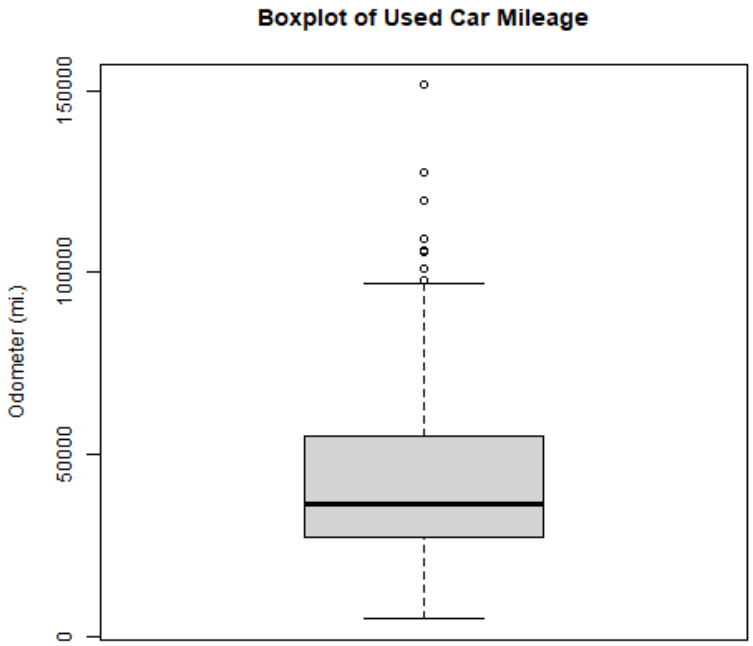
3. Based on the `mileage` of the listed cars, are there outliers?
   (a) compute the Inter-Quartile Range((`IQR`) * 1.5
   (b) how many values are above the outlier threshold?

```
outlier <- IQR(usedcars$mileage) * 1.5
paste("Outlier threshold:", outlier)
paste("There are",
      length(which(usedcars$mileage > outlier)),
      "outliers")
```

```
[1] "Outlier threshold: 41886.375"
[1] "There are 56 outliers"
```

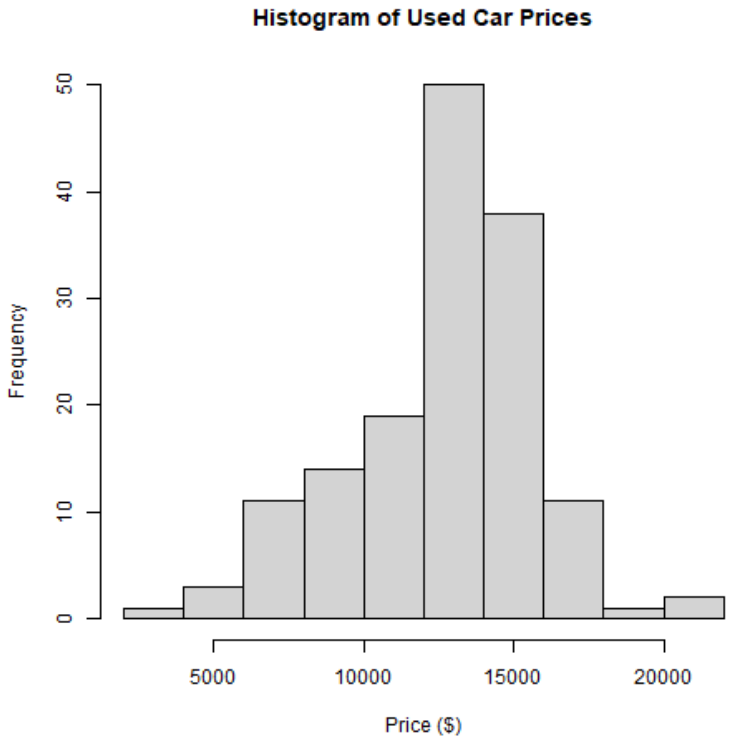
4. Visualize the spread of the data `mileage`.

```
boxplot(usedcars$mileage,
        main = "Boxplot of Used Car Mileage",
        ylab = "Odometer (mi.)")
```



5. Visualize frequency of `price`.

```
hist(usedcars$price,
     main = "Histogram of Used Car Prices",
     xlab = "Price ($)")
```



6. How many cars are priced between $12,000 and $14,000?

```
r <- sum(usedcars$price > 12000 & usedcars$price < 14000)
paste(r, "cars are priced in ($12,000,$14,000)")
paste("That's",
      format(100*r/nrow(usedcars), digits=3),
      "percent of all listed cars")
```

```
[1] "49 cars are priced in ($12,000,$14,000)"
[1] "That's 32.7 percent of all listed cars"
```

NEXT Exploring the structure of categorical data

1. Check which variables of the `usedcars` data frame are `numeric` and which are `categorical`.

```
str(usedcars)
```

```
'data.frame': 150 obs. of  6 variables:
 $ year        : int  2011 2011 2011 2011 2012 2010 2011 2010 2011 2010 ...
 $ model       : chr  "SEL" "SEL" "SEL" "SEL" ...
 $ price       : int  21992 20995 19995 17809 17500 17495 17000 16995 16995 16995
 $ mileage     : int  7413 10926 7351 11613 8367 25125 27393 21026 32655 36116 ..
 $ color       : chr  "Yellow" "Gray" "Silver" "Gray" ...
 $ transmission: chr  "AUTO" "AUTO" "AUTO" "AUTO" ...
```

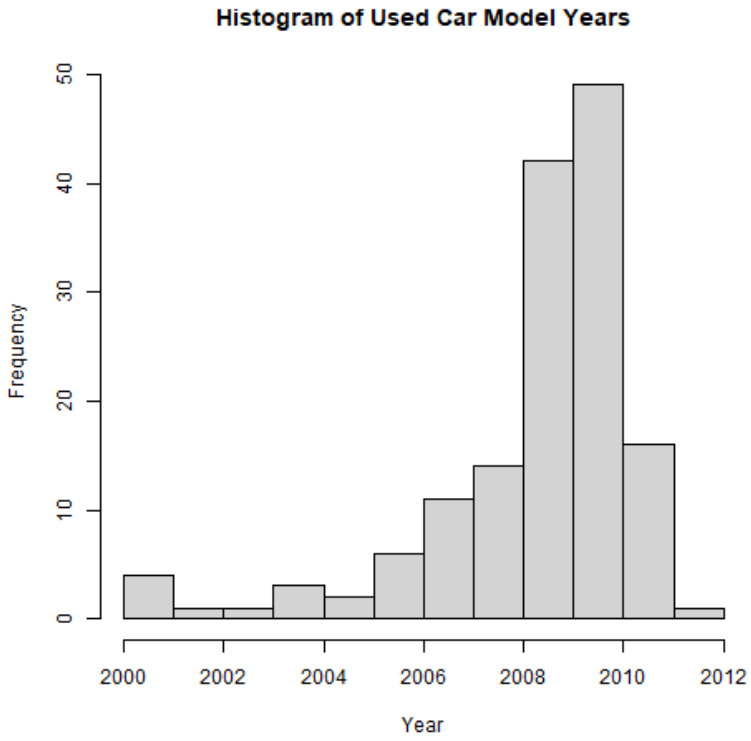
2. Display the car model `year` as a frequency table.

```
table(usedcars$year)
```

```
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
   3    1    1    1    3    2    6   11   14   42   49   16    1
```

3. Visualize the frequency table.

```
hist(usedcars$year,
     main = "Histogram of Used Car Model Years",
     xlab = "Year")
```



4. Display the proportions of the different car models:

```
prop.table(table(usedcars$model))
```

```
       SE       SEL       SES
0.5200000 0.1533333 0.3266667
```

5. Display the proportions of `color` with a single decimal place.

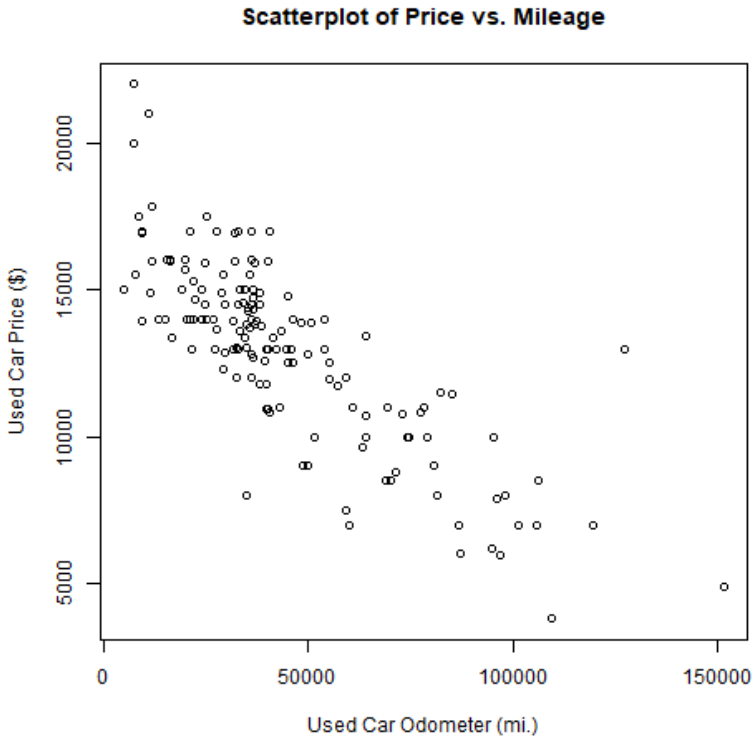
```
color_table <- table(usedcars$color)
color_pct <- prop.table(color_table) * 100
round(color_pct, digits=1)
```

```
Black   Blue   Gold   Gray  Green    Red Silver  White Yellow
 23.3   11.3    0.7   10.7    3.3   16.7   21.3   10.7    2.0
```

Exploring and visualizing relationships

1. Visualize the relationship between `price` and `mileage` in `usedcars`.

```
plot(x = usedcars$mileage,
     y = usedcars$price,
     main = "Scatterplot of Price vs. Mileage",
     xlab = "Used Car Odometer (mi.)",
     ylab = "Used Car Price ($)")
```



2. Create a cross table of `color` vs. `model`.

```
table(usedcars$color,usedcars$model)
```

```
       SE SEL SES
Black  19   3  13
```

```
Blue    9   3   5
Gold    1   0   0
Gray    7   5   4
Green   4   1   0
Red    12   2  11
Silver 11   7  14
White  14   1   1
Yellow  1   1   1
```

3. Visualize the cross table.

```
barplot(table(usedcars$color,usedcars$model),
        main = "Stacked Barplot of Used Car Model Colors",
        xlab = "Used car model types",
        ylab = "Frequency")
```