Managing and Understanding Data

Escribir vuestro nombre y apellidos

3 de marzo, 2021

Contents

R data structures

The R data structures used most frequently in machine learning are *vectors*, *factors*, *lists*, *arrays*, and *data frames*.

To find out more about statistical methods with R see (Teetor 2011; Hothorn and Everitt 2014; Baayen 2008).

Vectors

The fundamental R data structure is the **vector**, which stores an ordered set of values called **elements**. A vector can contain any number of elements. However, all the elements must be of the same type; for instance, a vector cannot contain both numbers and text.

There are several vector types commonly used in machine learning:`integer`(numbers without decimals), `numeric` (numbers with decimals), `character` (text data), or `logical` (`TRUE` or `FALSE` values). There are also two special values: `NULL`, which is used to indicate the absence of any value, and `NA`, which indicates a missing value.

Create vectors of data for three medical patients:

```
# create vectors of data for three medical patients
subject_name <- c("John Doe", "Jane Doe", "Steve Graves")
temperature <- c(98.1, 98.6, 101.4)
flu_status <- c(FALSE, FALSE, TRUE)
```

Access the second element in body temperature vector:

```r
# access the second element in body temperature vector
temperature[2]
```

```
## [1] 98.6
```

Examples of accessing items in vector include items in the range 2 to 3.

```r
## examples of accessing items in vector
# include items in the range 2 to 3
temperature[2:3]
```

```
## [1]  98.6 101.4
```

Exclude item 2 using the minus sign

```r
# exclude item 2 using the minus sign
temperature[-2]
```

```
## [1]  98.1 101.4
```

Use a vector to indicate whether to include item

```r
# use a vector to indicate whether to include item
temperature[c(TRUE, TRUE, FALSE)]
```

```
## [1] 98.1 98.6
```

Exploring and understanding data

After collecting data and loading it into R data structures, the next step in the machine learning process involves examining the data in detail. It is during this step that you will begin to explore the data's features and examples, and realize the peculiarities that make your data unique. The better you understand your data, the better you will be able to match a machine learning model to your learning problem. The best way to understand the process of data exploration is by example. In this section, we will explore the `usedcars.csv` dataset, which contains actual data about used cars recently advertised for sale on a popular U.S. website.

Since the dataset is stored in CSV form, we can use the `read.csv()` function to load the data into an R data frame:

```r
##### Exploring and understanding data --------------------

## data exploration example using used car data
usedcars <- read.csv("usedcars.csv", stringsAsFactors = FALSE)
```

Exploring the structure of data

One of the first questions to ask in your investigation should be about how data is organized. If you are fortunate, your source will provide a **data dictionary**, a document that describes the data's features. In our case, the used car data does not come with this documentation, so we'll need to create our own.

```
# get structure of used car data
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 numeric variables

```
## Exploring numeric variables -----

# summarize numeric variables
summary(usedcars$year)
```

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

```
summary(usedcars[c("price", "mileage")])
```

```
##      price          mileage
##  Min.   : 3800   Min.   :  4867
##  1st Qu.:10995   1st Qu.: 27200
##  Median :13592   Median : 36385
##  Mean   :12962   Mean   : 44261
##  3rd Qu.:14904   3rd Qu.: 55125
##  Max.   :21992   Max.   :151479
```

```
# calculate the mean income
(36000 + 44000 + 56000) / 3
```

```
## [1] 45333.33
```

```
mean(c(36000, 44000, 56000))
```

```
## [1] 45333.33
```

```
# the median income
median(c(36000, 44000, 56000))
```

```
## [1] 44000
```

```r
# the min/max of used car prices
range(usedcars$price)
```

```
## [1]  3800 21992
```

```r
# the difference of the range
diff(range(usedcars$price))
```

```
## [1] 18192
```

```r
# IQR for used car prices
IQR(usedcars$price)
```

```
## [1] 3909.5
```

```r
# use quantile to calculate five-number summary
quantile(usedcars$price)
```

```
##      0%     25%     50%     75%    100%
##  3800.0 10995.0 13591.5 14904.5 21992.0
```

```r
# the 99th percentile
quantile(usedcars$price, probs = c(0.01, 0.99))
```

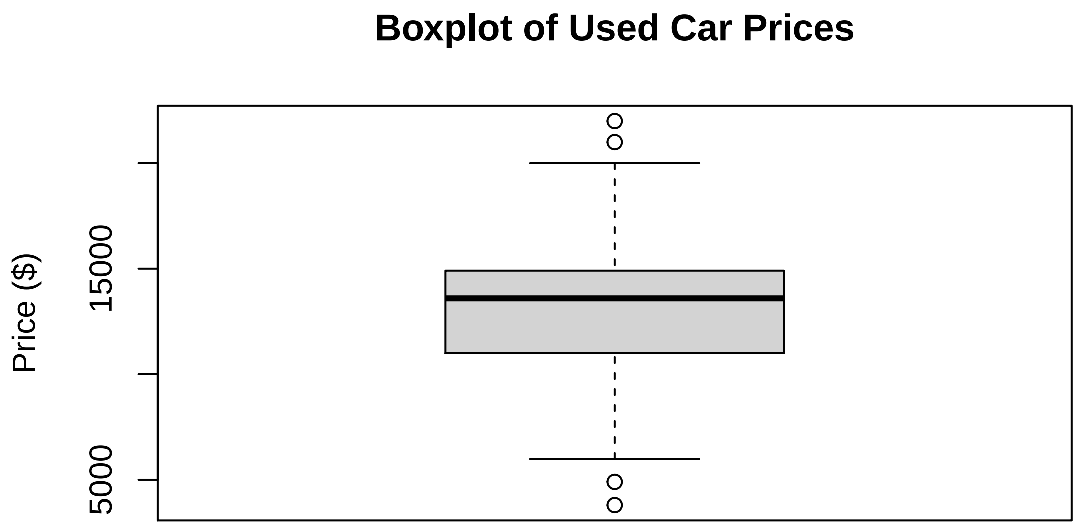
```
##       1%      99%
##  5428.69 20505.00
```

```r
# quintiles
quantile(usedcars$price, seq(from = 0, to = 1, by = 0.20))
```

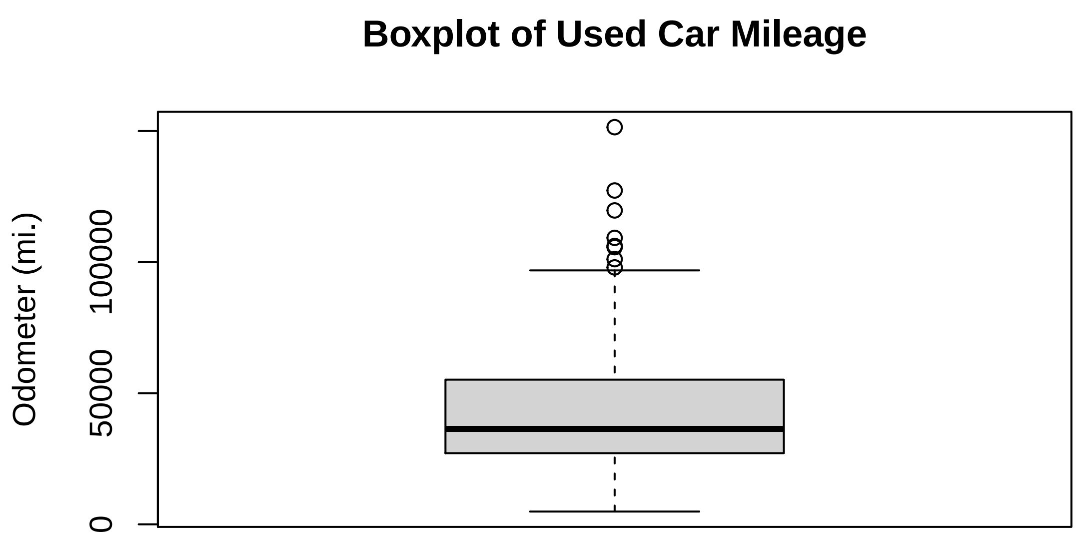
```
##      0%     20%     40%     60%     80%    100%
##  3800.0 10759.4 12993.8 13992.0 14999.0 21992.0
```

Visualizing numeric variables - boxplots

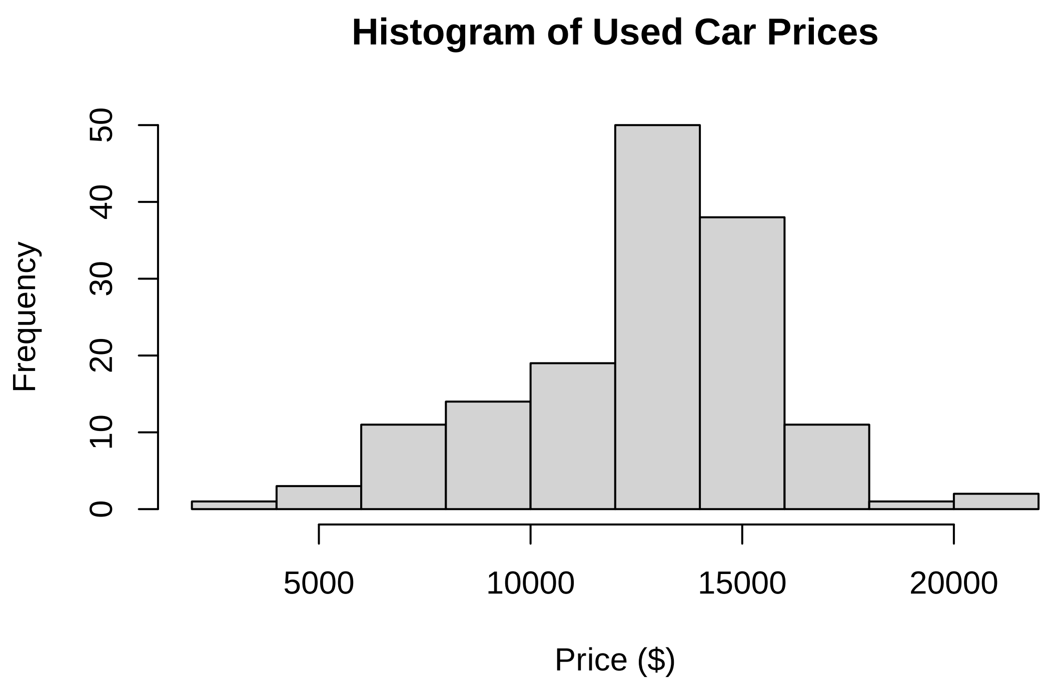
```r
# boxplot of used car prices and mileage
boxplot(usedcars$price, main="Boxplot of Used Car Prices",ylab="Price ($)")
```



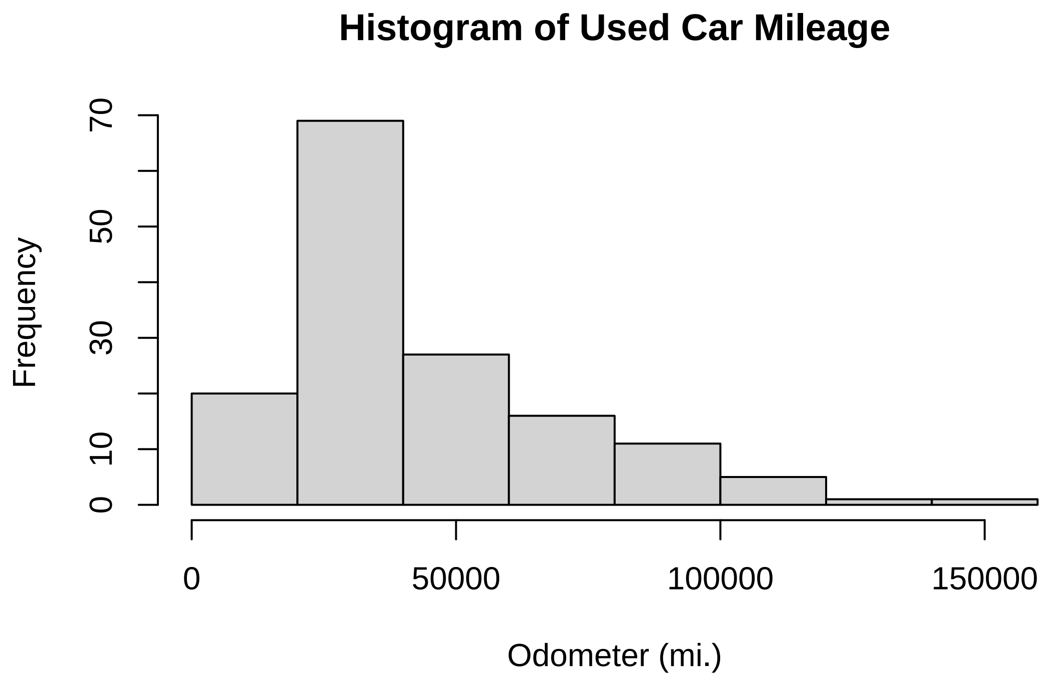

```
boxplot(usedcars$mileage, main="Boxplot of Used Car Mileage",
        ylab="Odometer (mi.)")
```



```r
# histograms of used car prices and mileage
hist(usedcars$price, main = "Histogram of Used Car Prices",
     xlab = "Price ($)")
```




```r
hist(usedcars$mileage, main = "Histogram of Used Car Mileage",
     xlab = "Odometer (mi.)")
```




```r
# variance and standard deviation of the used car data
var(usedcars$price)
```

```
## [1] 9749892
```

```r
sd(usedcars$price)
```

```
## [1] 3122.482
```

```r
var(usedcars$mileage)
```

```
## [1] 728033954
```

```r
sd(usedcars$mileage)
```

```
## [1] 26982.1
```

Measuring spread - quartiles and the five-number summary

The **five-number summary** is a set of five statistics that roughly depict the spread of a dataset. All five of the statistics are included in the output of the `summary()` function. Written in order, they are:

1. Minimum (`Min.`)
2. First quartile, or Q1 (`1st Qu.`)
3. Median, or Q2 (`Median`)
4. Third quartile, or Q3 (`3rd Qu.`)
5. Maximum (`Max.`)

Measuring spread - variance and standard deviation

In order to calculate the standard deviation, we must first obtain the **variance**, which is defined as the average of the squared differences between each value and the mean value. In mathematical notation, the variance of a set of `n` values of `x` is defined by the following formula. The Greek letter `mu` (μ) (similar in appearance to an m) denotes the mean of the values, and the variance itself is denoted by the Greek letter `sigma` (σ) squared (similar to a b turned sideways):

$$Var(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

The standard deviation is the square root of the variance, and is denoted by `sigma` as shown in the following formula:

$$StdDev(X) = \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Note. For more details on using mathematical expressions in Latex (R Markdown) see https://es.sharelatex.com/learn/Mathematical_expressions.

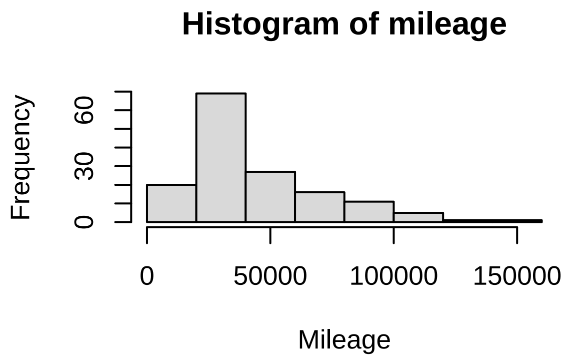
##Table with information about mileage and price

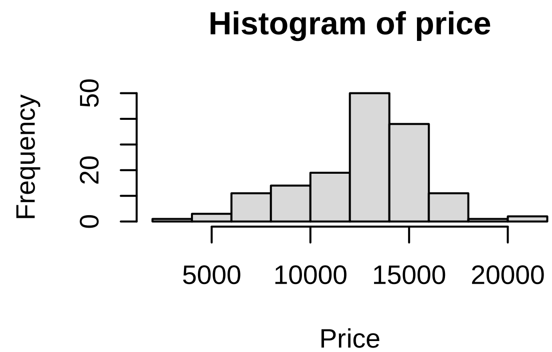
```
mileage_s<-summary(usedcars$mileage)
price_s<-summary(usedcars$price)
kable(rbind(mileage_s,price_s))
```

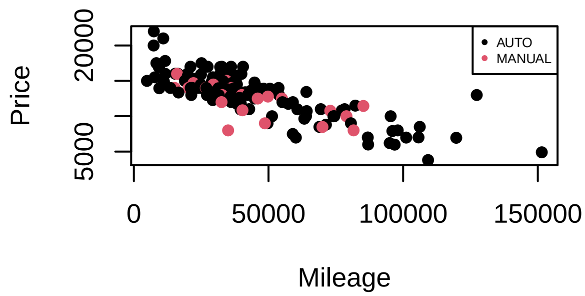
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
mileage_s	4867	27200.25	36385.0	44260.65	55124.5	151479
price_s	3800	10995.00	13591.5	12961.93	14904.5	21992

##Some descriptive graphics

```
par(mfrow=c(2,2))
hist(usedcars$mileage,xlab="Mileage",main="Histogram of mileage",col="grey85")
hist(usedcars$price,xlab="Price",main="Histogram of price",col="grey85")
usedcars$transmission<-factor(usedcars$transmission)
plot(usedcars$mileage,usedcars$price,pch=16,col=usedcars$transmission,xlab="Mileage",ylab="Price")
legend("topright",pch=16,c("AUTO","MANUAL"),col=1:2,cex=0.5)
```







#References

Baayen, R Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Hothorn, Torsten, and Brian S Everitt. 2014. *A Handbook of Statistical Analyses Using R*. CRC press.

Teetor, Paul. 2011. *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*. " O'Reilly Media, Inc.".