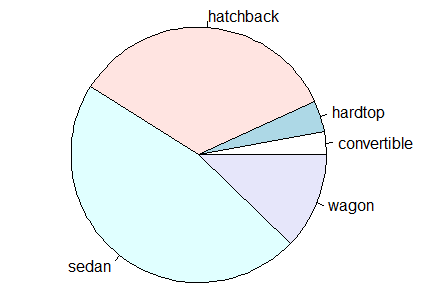
|  |
| --- |
| UMGC |
| Data Preprocessing in R |
| Week 2 Exercise |
|  |
| **DATA 630 – Data mining** |
|  |



|  |
| --- |
| In this exercise, you will use RStudio to load the cars data and to apply data preprocessing filters, including variables subset selection, discretization, and data rows subset selection. You will learn about the statistics the software provides for different attribute types, four unsupervised discretization methods and the differences among them, and attribute plots. |

# Data Preprocessing In R

[Data Preprocessing In R 1](#_Toc512331408)

[Why Preprocessing? 2](#_Toc512331409)

[Launch the Program 2](#_Toc512331410)

[Set the Working Directory 3](#_Toc512331411)

[Loading Data into R 4](#_Toc512331412)

[Exploring the Data 5](#_Toc512331413)

[Variable Filters 10](#_Toc512331414)

[Factor Function 10](#_Toc512331415)

[Discretization 11](#_Toc512331416)

[Cut Function 15](#_Toc512331417)

[Deleting a Variable 15](#_Toc512331418)

[Row filters 18](#_Toc512331419)

[Rows Range 18](#_Toc512331420)

[Handling Missing Values 18](#_Toc512331421)

[Remove Rows with Missing Values 18](#_Toc512331422)

[Display the Rows with Missing Values 19](#_Toc512331423)

[Number of Missing Values in a Data Row 20](#_Toc512331424)

[Number of Missing Values for Each Variable 21](#_Toc512331425)

[Replacing missing values with the mean 21](#_Toc512331426)

[Subset of Rows Matching an Equality Condition 22](#_Toc512331427)

[Sorting 23](#_Toc512331428)

[Saving Data to CSV File 23](#_Toc512331429)

[Data visualization 24](#_Toc512331430)

[Generic Plot function 24](#_Toc512331431)

[Copying a Plot to Clipboard 24](#_Toc512331432)

[Saving a Plot as an Image 25](#_Toc512331433)

[Saving a plot as a PDF file 26](#_Toc512331434)

[Additional Plot Examples 27](#_Toc512331435)

# Why Preprocessing?

Real-world data is usually inconsistent, missing values of interest, and has errors. Since data inconsistency and incompleteness affect the outcome of the data mining algorithm, data preprocessing is required.

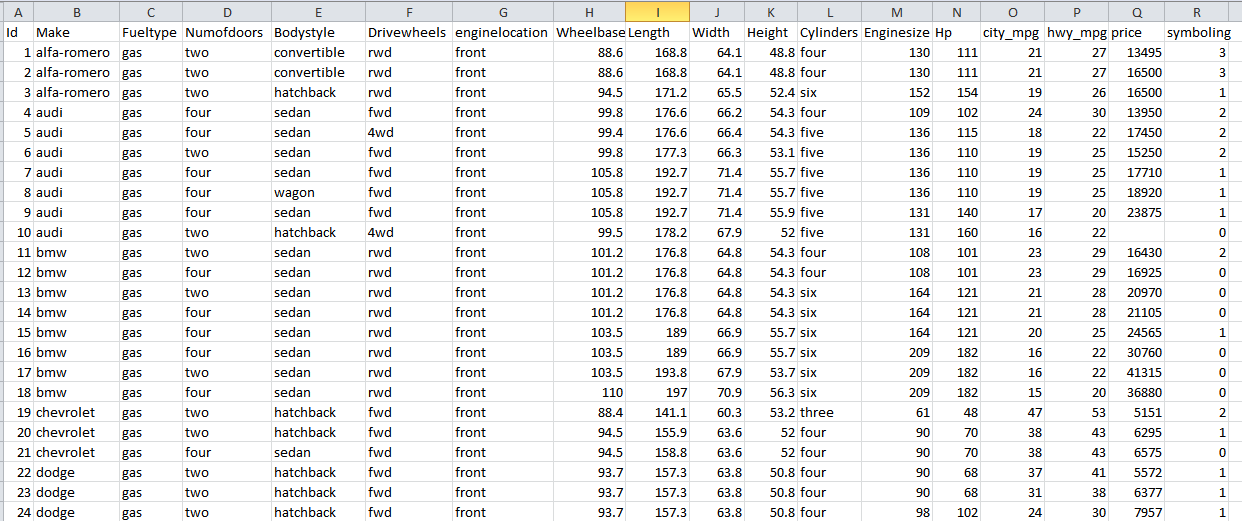
The data preprocessing operations include:

* Data cleansing – filling in the missing values, removing the outliers, and correcting any discrepancies in values,
* Data reduction – reducing the number of records, and/or removing irrelevant attributes, and
* Data discretization – converting numeric attributes to nominal by partitioning the values into bins.

In this exercise, we will load the cars data file in RStudio and run data pre-processing filters, including discretization.

Figure 1 shows the partial content of the cars.csv file. The column headings in the first row of the file are the attribute names called variables. The remaining 205 rows are the data, where each row is a single car record. If an attribute value is missing, the cell is left blank.

Column headers – attribute names / variables



Missing Price value

Data

(each row is a car record)

Figure 1: Partial Content of cars.csv file

# Launch the Program

Select RStudio from the Start menu in Figure 2 to launch the program. You may also double click on the RStudio Shortcut if you created one.

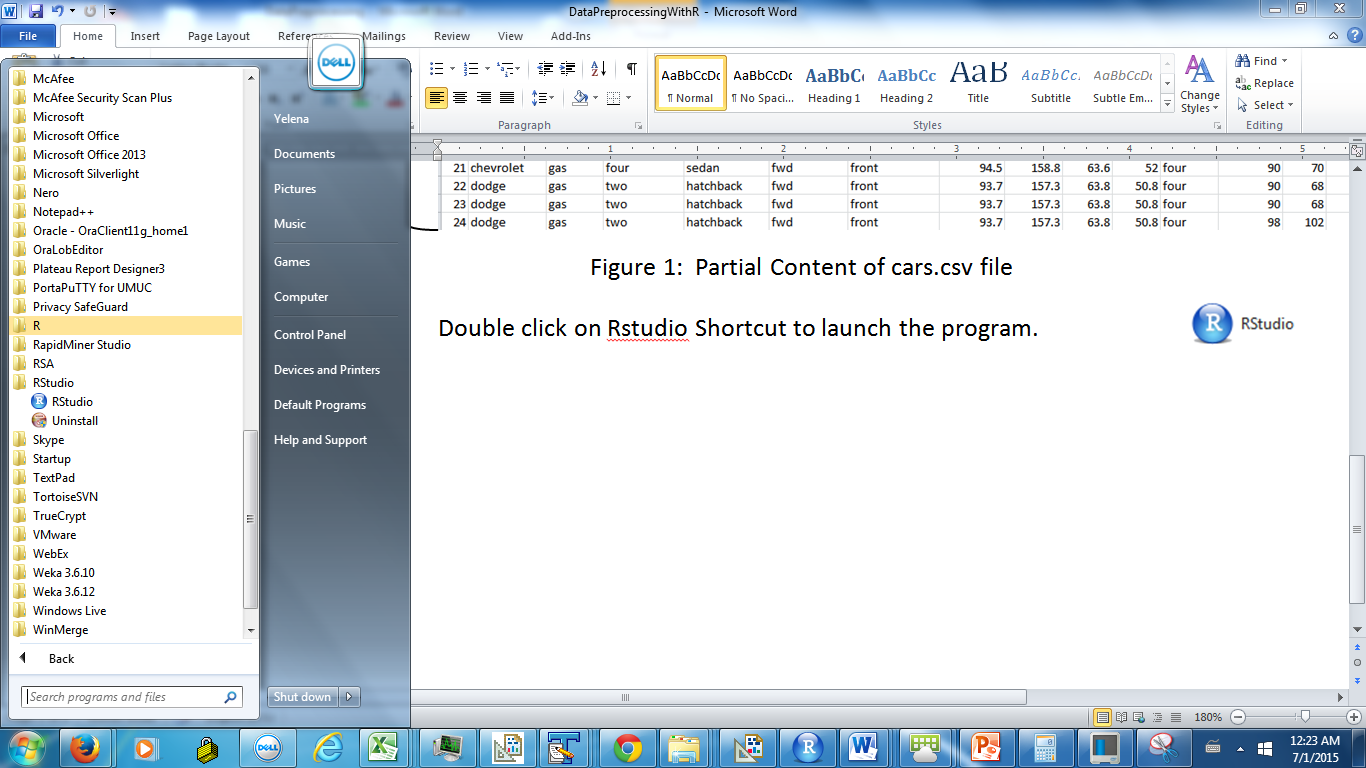
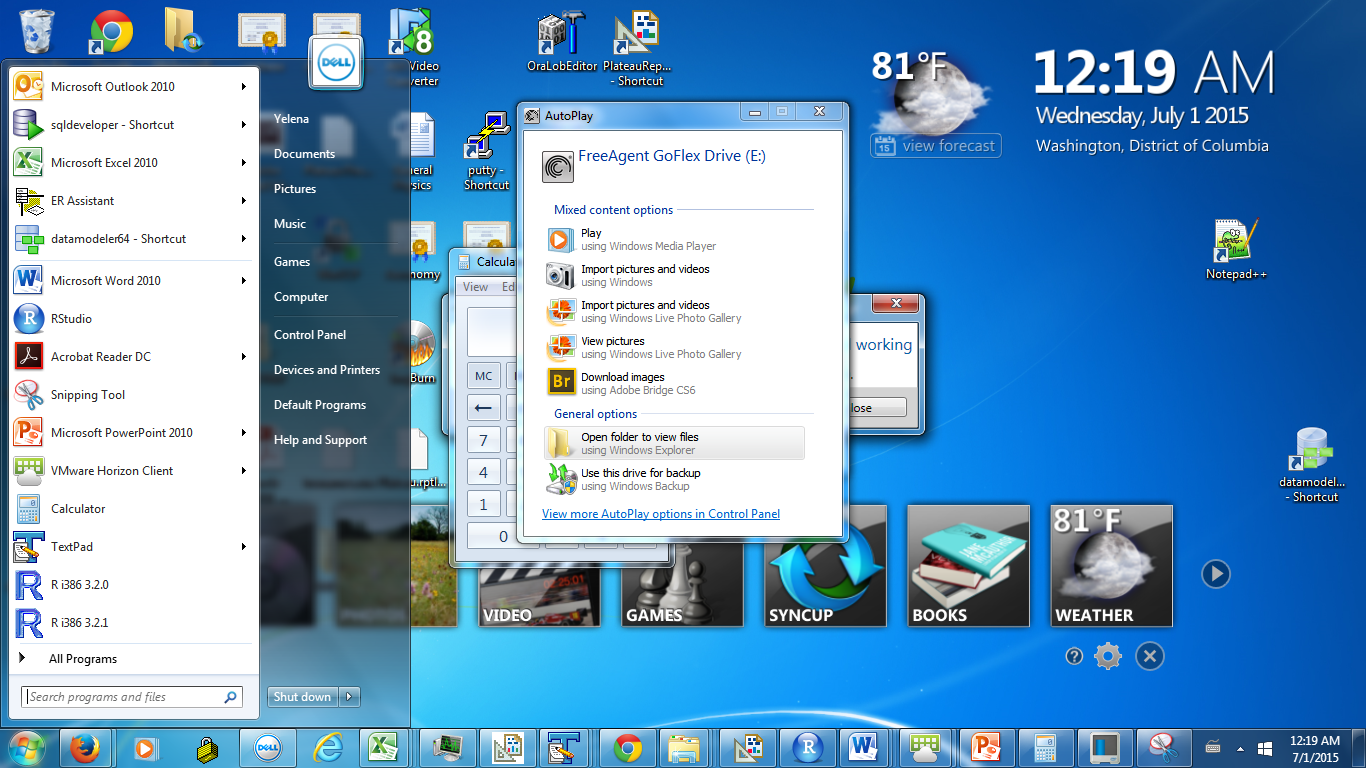
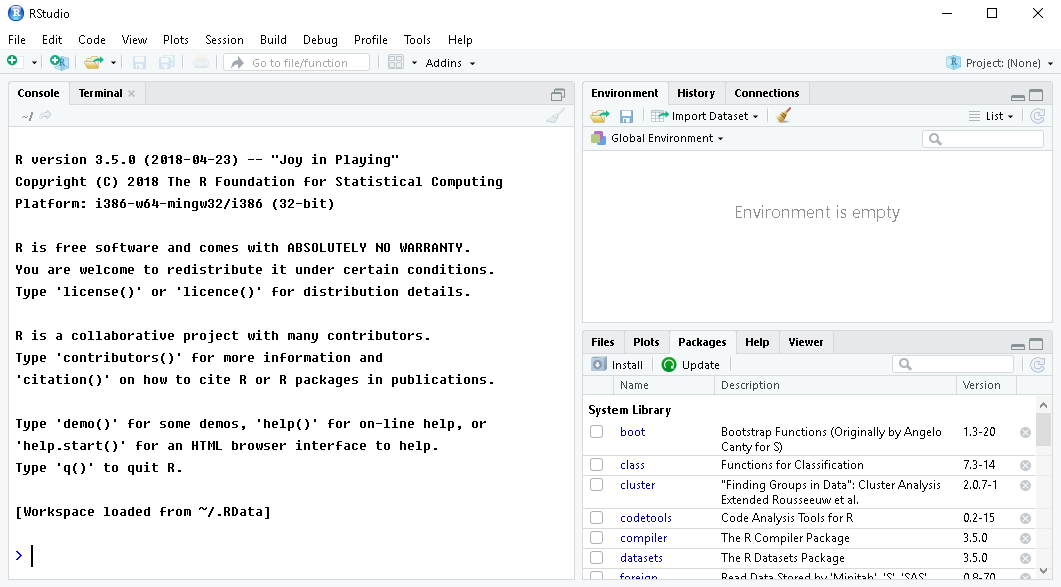


Figure 2: Select RStudio from the menu

An interface in Figure 3 will open. The left panel is a console window. You type the commands at the > prompt.

The environment tab at the top right window shows the variables and data in memory. The History tab allows viewing the commands you ran earlier. Each time you run a command, a new entry is added to the history.



The plots are displayed in the plots tab

Packages- software add-ons. Some packages are pre-installed, and some need to be installed.

Command Prompt

Files tab allows browsing the files on pc

R-studio built-in help

Data and Variables in memory – the panel is empty since we just launched the program.

Console- Frame where you enter the commands

R version

Session Commands History

Figure 3: RStudio Interface

The R language is **case sensitive**. For instance, Dir and dir have a different meaning.

# Set the Working Directory

A **working directory** is a hard drive location where R looks for a file that you attempt to open. Suppose the cars.csv file we want to load is in the E:/Datasets folder.

To set the working directory to E:/Datasets, enter the following setwd command in the console window and hit the Enter key. The directory path is specified in parentheses enclosed in double quotes.

setwd("E:/Datasets")

To verify that the working directory is set correctly, run the dir() command in Figure 4 to display the file names in the current working directory.



Figure 4: Files in the current directory

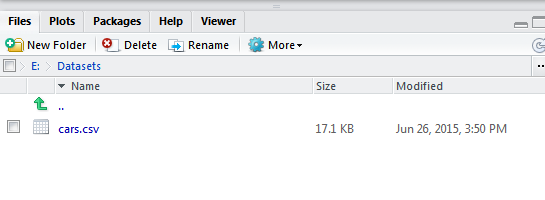
In addition, the title bar of the console window in Figure 5 displays the current working directory.



Working directory

Figure 5: Console Window Title Bar

An alternative way to view the files in the current working directory is to select the Files tab in the bottom right panel in Figure 6.



Current Working Directory

Files in the Current Working Directory

Files tab is active

Figure 6: Files Tab

# Loading Data into R

Use read.csv command to read the cars file content into a variable called cars. The type of variable that stores the content from the first sheet in a CSV file is called a **data frame**.

* The structure of a data frame resembles a matrix structure. However, the data types for variables in the data frame can be different from one another.
* The data frame can also be viewed as a set of variables with the same number of values.
* The variable name cannot start with a number and cannot contain special characters, including !, @, $, @, +, -. and \*
* The two character operator <- is an assignment operator.

The first input parameter for the read.csv function is the data file name enclosed in double quotes. The second parameter, head=TRUE, specifies that the first row in the file contains the column headers. The sep parameter is the column delimiter enclosed in double quotes. For example, sep=“,” means that the values in each data row are comma delimited.

Reads factor variables with char values as factors; prevents their conversion to chr type

The values delimiter

Command to Read from CSV file

Assignment Operator

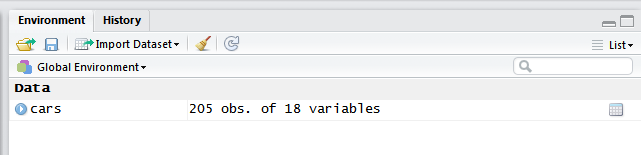
cars<-read.csv(file="cars.csv", head=TRUE, sep=",", as.is=FALSE)

Data frame name – stores data from the first sheet in CSV file

File Name

Read the column headings from the first row

Select Environment tab in the top right panel in Figure 7. Upon successful read.csv command execution, you will see the cars data frame information under the data heading. This information includes the number of observations and the number of variables.



Click to view the data in a tabular format

Cars Data Frame contains 205 observations and 18 variables

Data frame name

Environment tab is active

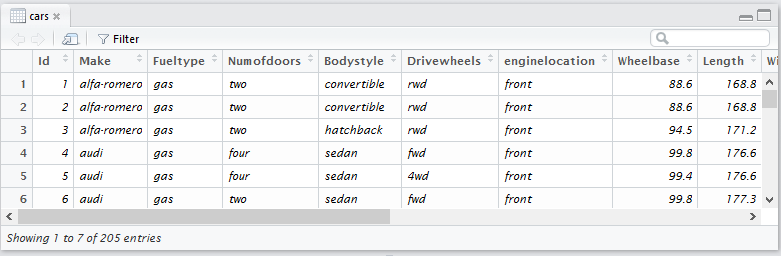
Figure 7: Cars has been added to Environment Tab

To preview the cars data in a tabular format, enter the View command at the prompt. Note, the first letter of the command is in upper case.

View(cars)

The first row in the preview panel shows the variable names, and the remaining rows are the data. The status bar at the bottom of the panel shows the number of data rows in the dataset. Make sure that the number matches the number of rows in the .csv file.

The arrows next to the column headings allow sorting the data rows by the corresponding variable values.



Headers – Variable names

Data frame name

Number of rows in the dataset

Use the arrows for sorting data

Figure 8: Data Preview

# Exploring the Data

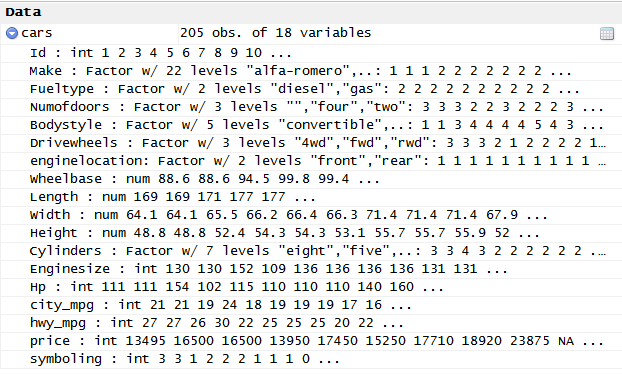
Click on the blue arrow to expand the data frame properties in Figure 9. In addition to the number of observations and the number of variables, you will see the properties for each variable, including the variable name and variable type. The variable types are:

**int** – The values for this variable are continuous integers.

**num** - The values for this variable are real numbers.

**Factor** – The values for this variable are discrete, pre-defined labels. For each factor variable, the pre-defined discrete values are called levels. For example, diesel and gas are the fueltype levels.

The data type of a variable determines what methods and operations we may use. For example, we cannot add or multiply categorical variables. On the other hand, we cannot use some methods such as association rules for continuous numeric variables.



Number of variables

Number of observations

Variable data type

Variable name

Figure 9: Cars Data in Environment Panel

Run the following **names** command to display the variable names we just loaded from the cars.csv file. The command takes the data frame name as an input parameter. The name is not enclosed in double quotes.

Make sure that the number of variables and the variables names in Figure 10 match the number of populated columns and the column headings in the csv file. (Note that R replaces the spaces in variable names with an underscore.)

names(cars)

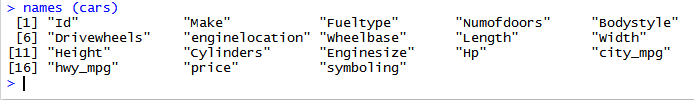
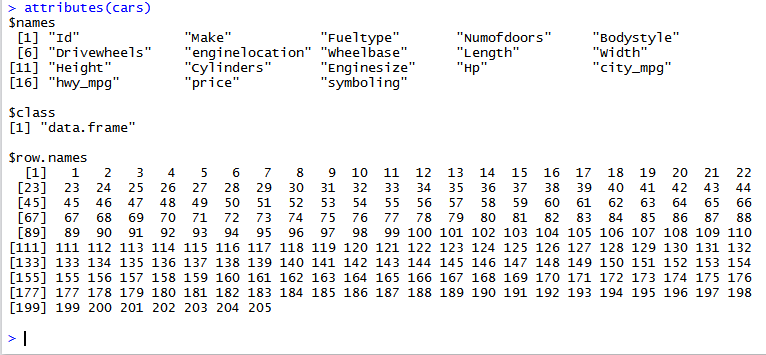


Figure 10: Attribute names

Each data row in the data frame is labelled with the corresponding value in the first column. To display the row labels in addition to the variable names on Figure 11, run the attributes command. The command takes the data frame name as an input.

attributes(cars)



Data row names

Data Frame name

Variable names

Figure 11: Variable names and row names

To display the size of the data frame, run the **dim** command. The command takes the name of the data frame as an input. The first number in the output in Figure 12 is the number of data rows, and the second number is the number of variables.

dim (cars)

18 Variables



205 Data Rows

Figure 12: Dim Command output

To display just the number of data rows, run the following **nrow** command. The command takes the data frame name as an input.

nrow(cars)

The output in Figure 13 shows that the cars data contains 205 rows.



Figure 13: Number of Rows

The following **dim** command also returns the number of rows. 1 in squared brackets means the first dimension of the data frame, or the number of rows.

dim(cars)[1]

To display the number of variables, run the following **ncol** command. The command takes the data frame name as an input.

ncol(cars)

The output in Figure 14 shows that the cars data has 18 variables.



Figure 14: Number of Variables

The following **length** and **dim** command also return the number of variables. [2] means the second dimension of the data frame, or number of columns.

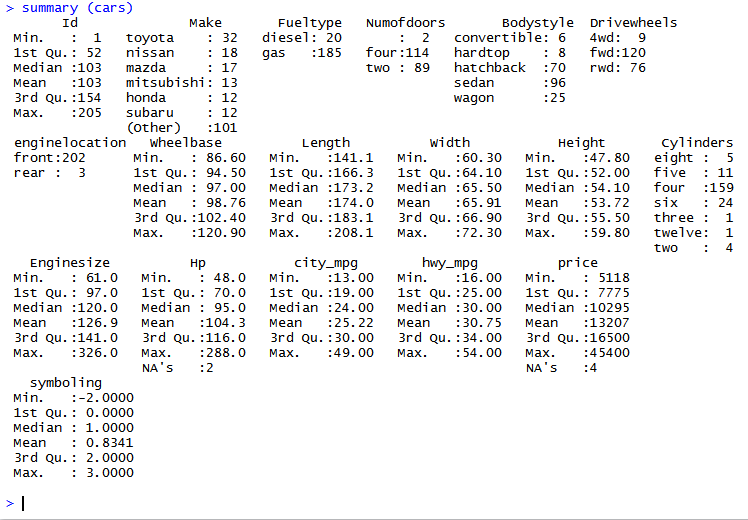
length(cars)

dim(cars)[2]

Run the **summary** command to display the descriptive statistics for all variables or for a single variable. To display the statistics for all variables, specify the data frame name as the input to the summary command.

summary (cars)

Figure 15 shows the output from the summary command. The output includes a table with statistics for each variable in the data frame.



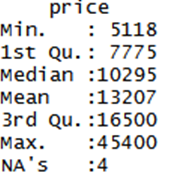
**NA stands for missing values.**

Factor Variable

Numeric Variable

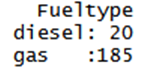
Figure 15: Descriptive Statistics for All Variables

The statistics for continuous numeric variables are the minimum value, maximum value, 1st and 3rd quartile, median, and mean. The statistics also include the number of instances with missing attribute value, if there are any.



For example, the variable price does not have a pre-defined set of discrete values. The values of price are continuous real numbers. The NA’s :4 indicates that 4 car instances are missing a price value.

A factor variable is a categorical variable with pre-defined discrete values called **levels**. Statistics for a factor variable include the levels and the number of data rows corresponding to each level. For example, the levels of the Fueltype variable are diesel and gas.



185 cars have a gas fuel type

20 cars have a diesel fuel type

Levels

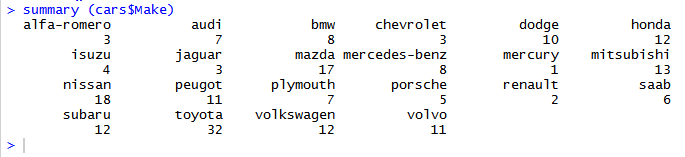
You may use summary command to display the descriptive statistics for a single variable. Append the $ and the variable name to the data frame name. For example, type the following command to display the descriptive statistics for the Make attribute.

summary(cars$Make)

Since Make is a factor variable, the statistics for Make are the levels and the number of values at each level.

The individual variable name

Data Frame name



Each car make value is a level

Figure 16: Descriptive statistics for Make attributes

You may also use the following lapply command to display the descriptive statistics for all variables. This command loops through the cars data frame variable and displays the statistics for each variable. Figure 17 shows the partial output of the lapply command.

The first input parameter is the data frame name, and the second parameter is the function to run for each variable.

lapply(cars, summary)

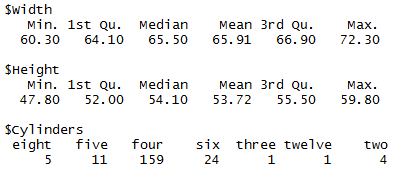


Figure 17: Partial Output from lapply command

To display the minimum and maximum value of the numeric variable, use the range function. The function takes the data frame name followed by $ followed by variable name as input.

range(cars$Width)

Figure 18 shows that the lowest car width value is 60.3 and the highest car width value is 72.3.



Highest value

Lowest value

Figure 18: Car Width Range

Notice that the output from the summary command excludes standard deviation statistics. To display the standard deviation for a numeric variable, run the sd command. The command takes variable name as an input.

sd(cars$Width)

The output in Figure 19 is the standard deviation of the car width variable.



Figure 19: Car Width Standard Deviation

What happens when we run the mean command on a variable with missing values? Run the following command to display the mean of a price variable.

mean(cars$price)

The output on Figure 20 shows that the mean value of a variable is unknown if at least one value is missing.



Figure 20: Mean of a Price Variable with Missing Values

The optional second parameter for a mean function enables ignoring the missing values in the mean computation. We set the parameter na.rm=TRUE to ignore the missing values. When the parameter is omitted, the default value is FALSE.

mean(cars$price, na.rm=TRUE)

The output in Figure 21 is the mean car price.



Figure 21: Mean of a Price Variable with Missing Values Ignored

# Variable Filters

Variable filters add variables, remove variables, and change variable values. The number of data rows remains the same.

## Factor Function

The factor function converts the variable with discrete numeric values into a categorical factor variable.

The last variable symbolling has the categorical values -2, -1, 0, 1, 2, 3. However, the last line in Figure 9 shows that the variable is numeric. In addition, the output from the summary command in Figure 22 includes the descriptive statistics for numeric variables (min, median, mean, etc.)

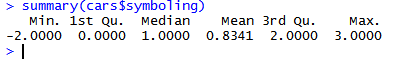


Figure 22: Symboling attribute statistics before factor

To convert the variable to a factor datatype, we run the factor function. The function takes a data frame variable as an input parameter. <- is an assignment operator.

cars$symboling<-factor(cars$symboling)

To validate if the variable has been converted from numeric to factor, we run the summary command in Figure 23. The statistics have changed from min, median, mean, etc. to the levels and the number of data rows with the value at each level.



27 cars have symboling level 3

The symboling values are the Levels

Figure 23: Symboling variable statistics after factor

## Discretization

**Discretization** is the process of converting numeric variables into categorical variables (factors) by dividing the numeric values into distinct ranges. Some algorithms such as association rules, covered later in the course, work only with factor variables.

In some textbooks and online documentation, the authors may refer to discretization as binning or recoding.

**Unsupervised discretization** does not consider the class of a data instance. The dependent variable can be either numeric, or integer, or factor.

To run unsupervised discretization, you need to install the arules package if you have not installed it before. Run the following command to install arules package.

install.packages("arules")

Upon successful installation, the arules package will be added to the user library as shown in Figure 24.

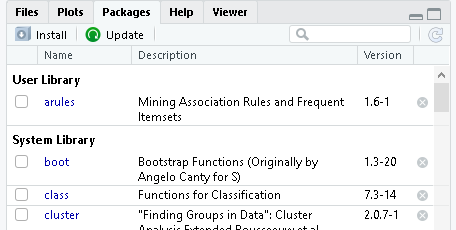


Figure 24: Arules Package added to the User Library

Run the library command to call the package. This command needs to be run in every session where you need to use the package.

library (arules)

Let’s explore different unsupervised discretization binning methods:

**Equal Interval width** is the default unsupervised discretization method. Data values are partitioned into intervals of equal width. This approach is sensitive to the extremely large and/or extremely small values called outliers. In addition, this approach does guarantee that the initial distribution of values remains.

We will run the equal interval width method on city\_mpg variable to partition the values into six ranges.

The output from the first command in Figure 25 is the statistics for the city\_mpg variable before running the discretization, including minimum value, median, mean, etc.

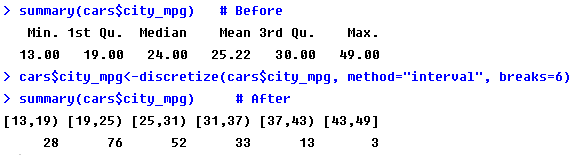
The second command takes the variable to be discretized as a first input, the discretization method as a second input, and the number of bins as a third input.

cars$city\_mpg<-discretize(cars$city\_mpg, method="interval", breaks=6)

The output from the summary command on the city\_mpg variable after discretization includes the value ranges (levels) and the number of cars that have a city\_mpg value in each range.

For example, 28 cars have a city\_mpg between 13 and 19. 76 cars have a city\_mpg between 19 and 25. The value ranges have approximately the same width.

The round bracket indicates that the boundary point does not belong to an interval. The square bracket indicates that the boundary point belongs to an interval.



Value ranges = levels

Variable Name

Discretization Method

Number of bins

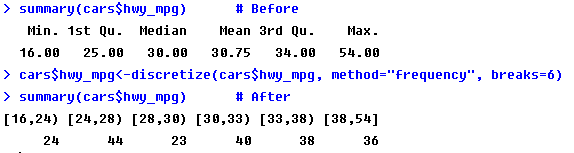
Figure 25: City\_mpg Statistics before and after Discretization

**Equal Frequency Method** – In this method, the interval length is based on the distribution of variable values. Approximately the same number of data instances should fall into each bin or data range.

cars$hwy\_mpg<-discretize(cars$hwy\_mpg, method="frequency", breaks=6)

Figure 26 shows the summary command output for hwy\_mpg before and after running the discretization.

For each interval, the round bracket indicates that the boundary point does not belong to the interval. The square bracket indicates that the boundary point belongs to the interval.



Variable Name

Number of bins

Discretization Method

Figure 26: hwy\_mpg Statistics before and after Discretization

**K-means clustering** discretization method is also known as a univariate k-means clustering method. This algorithm randomly selects the center points for each partition. Each value within a partition is closer to the partition’s center point than to the center points of other partitions. This discretization approach attempts to preserve the original values distribution. (The k-means clustering method will be covered later in the course).

The first parameter is a variable name. The second parameter is a discretization method, which is cluster. The third parameter is the number of bins.

Number of bins

Discretization Method

Variable Name

cars$Enginesize<-discretize(cars$Enginesize, method="cluster", breaks=6)

Figure 27 shows the summary command output for the Enginesize attribute before and after running the discretization.

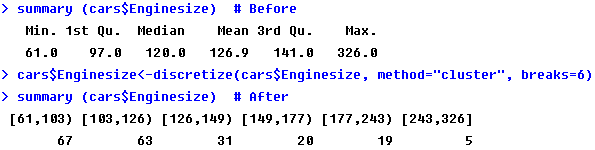


Figure 27: summary command output for Enginesize before and after Discretization

**NOTE: The above command may return different bin sizes subject to the version of R installed. For R version 4.0, you may see [61, 94.5) as the first bin containing 38 values.**

**User Specified Method** – a user specifies the boundary points for each interval.

Suppose that we want to divide the wheelbase attribute values into the following ranges:

Wheelbase below **90**

Wheelbase between **90** and ***100***

Wheelbase between ***100*** and ***110***

Wheelbase above ***110***

The first parameter is the attribute name. The second parameter is the discretization type, which is set to “fixed” here. The third parameter is a vector of the boundary points. –Inf stands for the negative infinity, and Inf stands for positive infinity.

The boundary points are negative infinity, 90, 100, 110, and positive infinity.

Interval Boundary Points

cars$Wheelbase<-discretize(cars$Wheelbase, method="fixed", breaks=c(-Inf, 90, 100, 110, Inf))

Figure 28 shows the summary command output for the Wheelbase variable before and after running the discretization filter.

The interval boundary points match the specified boundaries in the input vector. The round bracket indicates that the boundary point does not belong to an interval. The square bracket indicates that the boundary point belongs to an interval.

The most dominant value range for the wheelbase is between 90 and 100. Only 8 data rows have wheelbase below 90.

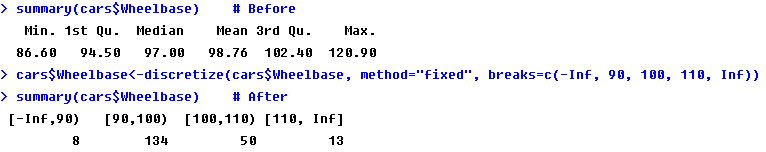


Figure 28: Wheelbase Variable before and after Discretization

## Cut Function

The Cut function is an alternative approach to perform unsupervised discretization with equal width binning. The function partitions the variable values into the specified equal width categories. The function takes a variable name and number of categories as input parameters.

Figure 29 shows the statistics for the Length variable before and after running the cut function. The statistics before running the cut function include minimum, median, mean, etc. The statistics after running the cut functions are 4 value ranges and the number of data rows with the Length value in each range.

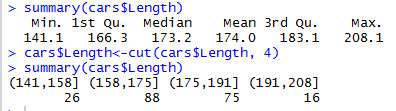
summary(cars$Length)

Number of ranges

Variable name

cars$Length<-cut(cars$Length, 4)

summary(cars$Length)

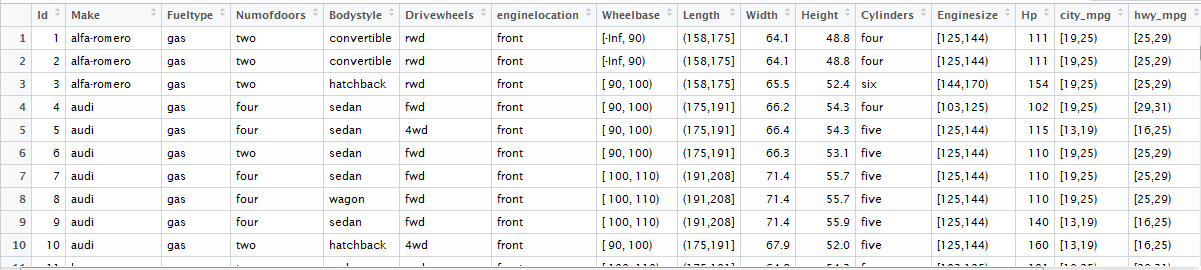


Length statistics before running cut function

Length statistics after running cut function

Figure 29: Cut Function Output

Figure 30 is a tabular view of the partial cars data after applying the above-demonstrated discretization. The Wheelbase, Length, Enginesize, city\_mpg, and hwy\_mpg values now have interval ranges.



The value are discrete value ranges

Figure 30: Discretized Values – Tabular View

## Deleting a Variable

The id variable uniquely identifies each data row. Since unique identifiers may affect the data mining performance, we may choose to drop them.

Run the following command to create a copy of the cars data frame with the id variable removed. A comma after an opening bracket indicates that we are removing a column. The number following the minus sign signifies the position of the variable to remove.

newcars<-cars[, -1]

The newcars data frame has been added to the Environment tab in the upper right panel on Figure 31. The number of variables changed from 18 to 17. The number of observations remains unchanged.

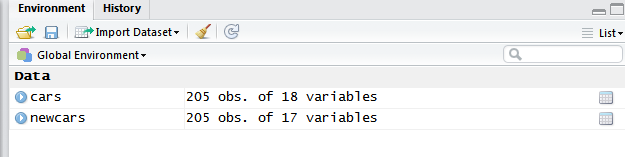


Figure 31: newcars data frame

Run the View command to preview the newcars in a tabular format on Figure 32. The Id variable has been removed, and Make is now the first variable.

View(newcars)

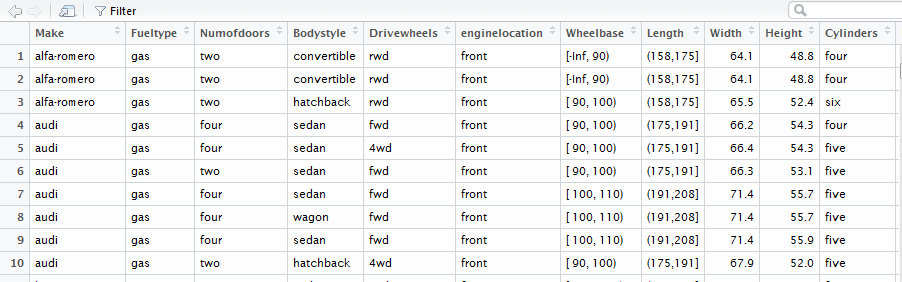


Figure 32: ID has Been Removed

To remove more than one variable in a single command, we specify the variable positions as a vector. Variables Id, FuelType, and Numofdoors have positions 1, 3, and 4 respectively.

To specify the positions of variables to remove, we need to create a vector with numeric values.

Function c(1, 3, 4) creates a numeric vector with the values 1, 3, 4. The minus sign preceding the vector means that the columns at the specified column positions are to be removed.

newcars<-cars [, - c(1, 3, 4)]

Variables Id, FuelType, and Numofdoors have now been removed. The Make variable has been moved to the first position, Body style has been moved to the second position, etc.

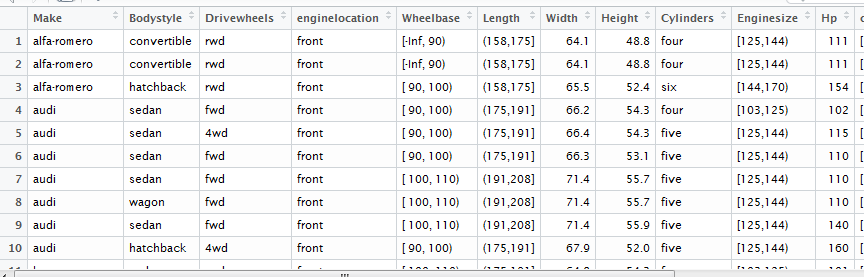


Figure 33: Id, FuelType, and Numofdoors Removed

Instead of specifying the variables to remove, you may specify the variables to keep. When the minus sign preceding the vector is omitted, the variables at the positions specified in the vector are kept, and the remaining variables are removed.

Suppose that we want to keep Id, Make, FuelType, Numofdoors, and Bodystyle. These variables are at positions 1, 2, 3, 4, and 5 respectively. The command takes a vector c(1, 2, 3, 4, 5) as an input.

newcars<-cars [, c(1, 2, 3, 4, 5)]

Figure 34 shows the tabular view of the partial newcars data frame. The variables in the data frame are Id, Make, FuelType, Numofdoors, and Bodystyle.

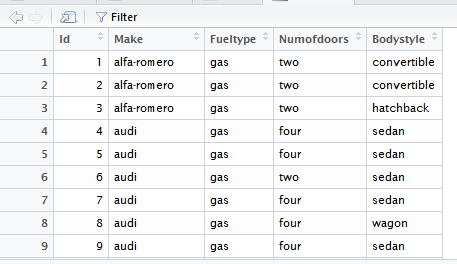


Figure 34: Only Specified Variables Remained

The following command is an alternative approach to keep the first 5 attributes. The numbers in the square bracket specify the range of variable positions.

newcars<-cars[1:5]

Another approach to remove a column is to set it to NULL. The first command copies all data from cars data frame into newcars. The second command removes the Id column.

newcars<-cars

NULL needs to be in the upper case.

newcars$Id<-NULL

# Row filters

Row filters are used to add rows, remove rows, and process rows. The number of variables remain the same.

## Rows Range

To select a subset of rows, specify the data frame name followed by square brackets. The first parameter in the brackets is the range of row numbers. The missing number after a comma implies that all variables are included.

newcars<-cars[1:10,]

The tabular view in Figure 35 shows that newcars now contains the first 10 rows and all variables from the cars data frame.

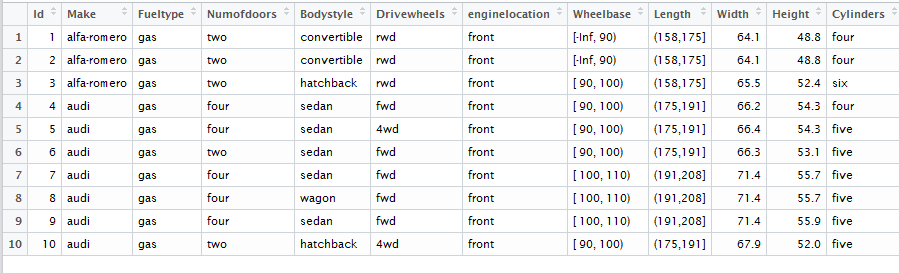


Figure 35: First 10 Rows

## Handling Missing Values

### Remove Rows with Missing Values

We may create a copy of the cars data frame removing those observations that are missing at least one value for at least one variable. Let us call the data frame without missing values newcars.

**Method 1 – Use the subset command.** This command takes the data frame name and a logical condition as an input. The rows that meet the condition are part of the subset. The complete.cases condition is true when the data row has a value for each attribute.

newcars<-subset(cars, complete.cases(cars))

**Method 2 – Use square brackets.** Instead of using the subset function, we may specify the row matching condition in the square brackets. If we do not specify the columns after the comma inside the brackets, the subset will include all columns.

newcars<-cars[complete.cases(cars),]

**Method 3 – Use na.omit command.** The command na.omit also removes data rows with missing values. The command takes the data frame name as an input.

newcars<-na.omit(cars)

Figure 36 shows the cars and newcars environment variables. The cars data frame contains 205 observations, and newcars contains 199 observations. Hence, at least 6 observations in the cars dataset are missing a value for at least one variable.

Cars and newcars have the same number of variables.

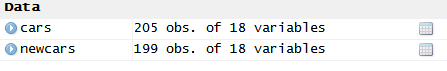
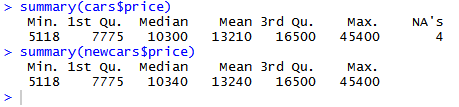


Figure 36: Number of Observations has Changed

We may also compare statistics for the individual variables. For example, the output from the summary command at the top of Figure 37 shows that 4 observations in the cars data frame are missing a price value (NA’s 4). The summary command at the bottom of the figure, shows that all observation in newcars data frame have a value. (The NA’s column is omitted)



Prices has 4 missing values.

Price does not have missing values.

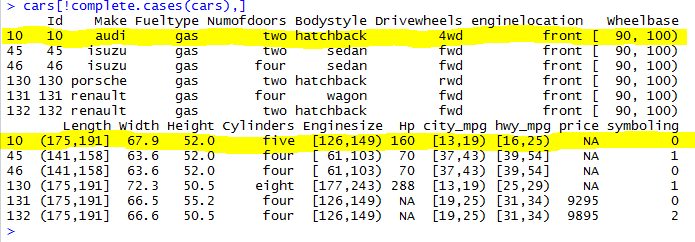
Figure 37: Price Variable Statistics Comparison

### Display the Rows with Missing Values

The following command returns the data rows missing a value for one or more variables. The ! is the logical operator for “not”. The selection condition for rows is incomplete cases. Since we did not specify the columns after the comma inside the brackets, the output in Figure 38 includes all columns.

cars[!complete.cases(cars),]

Since the variables do not fit on a single output line, the output has 2 lines for each data row. The highlighted lines in Figure 38 correspond to the same data row.



**Missing values**

Figure 38: Rows with Missing Values

If the dataset is large, we might not be interested in previewing data rows with missing values. Instead, we could display how many data rows are missing a value for at least one variable.

The nrow function returns the row count. This function takes a data frame or a subset of a data frame as input. Instead of specifying the data frame name, we may specify the command that generates the data frame or data frame subset.

Hence, the output of the cars[!complete.cases(cars),] command discussed above can be an input for nrow command.

nrow(cars[!complete.cases(cars),])

The output in Figure 39 shows that 6 data rows are missing at least one value.



Figure 39: Number of Rows with Missing Values

### Number of Missing Values in a Data Row

Whenwe remove the data rows with missing values, we risk losing data. An alternative is to fill in the missing values with the most common value for the categorical variables and with the mean value for numeric variables. However, this approach might not be the best option if values are missing for most of the variables.

Hence, it could be useful to **check the number of missing values in each data row**. The “apply” function loops through data rows one at a time, and applies the specified function to each row. The first parameter is the data frame name. To loop through rows, we specify 1 for the second parameter.

The third parameter is the function to apply. We may use a built -in function, or we may define an inline function. Keep in mind that unlike a regular function, an inline function is temporary, and available only within the current command.

For each value in a row, is.na returns true if the value is missing and false if the value is not missing. The sum counts the number of true values returned for a row. For instance, if the data row is missing 2 values, the sum will return 2.

Is.na () is a function we can use to check if the value is missing.

Data frame

Loop through rows

Inline function

apply(cars, 1, function (cars) sum(is.na(cars)))

Figure 40 shows that the first 9 rows do not have missing values. The 10th row has one missing value, etc.

10th row

First 9 rows

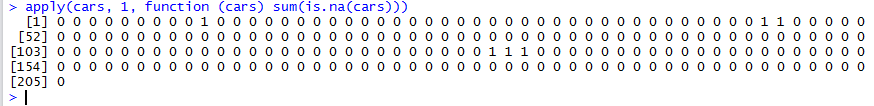


Figure 40: Number of Missing Values in Each Data Row

### Number of Missing Values for Each Variable

Removing rows with missing values and/or filling in missing values might not always be the best approach when a variable is missing values for most of the rows. It could be useful to check how many data rows are missing a value for each variable.

We use an apply function above and change the value of the second input parameter to 2. In another words, this apply function now loops through the variables or columns instead of data rows.

apply(cars, 2, function (cars) sum(is.na(cars)))

The output on Figure 41 shows the number of missing values for each variable. For example, four data rows are missing a price value. Two data rows are missing an Hp value. No data rows are missing Make.

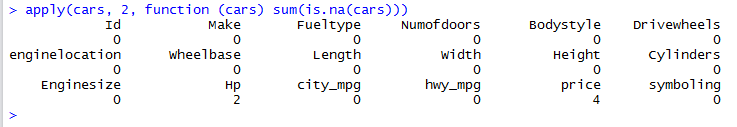


Figure 41: Number of Missing Values for Each Variable

### Replacing missing values with the mean

Since only two data rows are missing the Hp value and only four data rows are missing the price, we can replace the missing values with the mean value for that variable.

The following command finds the rows with missing values for the Hp variable and replaces the missing values with the attribute mean. In general, is.na returns true when the input value is missing. Hence, the function is.na on the left hand side returns the row numbers for where the Hp value is missing.

The mean function on the right hand side takes the data frame column name as a first input. The second input na.rm=TRUE means that R should ignore the missing values in the mean value computation.

cars$Hp[is.na(cars$Hp)]<-mean(cars$Hp, na.rm=TRUE)

Now run the following command to replace the missing price values with the price mean.

cars$price[is.na(cars$price)]<-mean(cars$price, na.rm=TRUE)

To validate that all missing values have been replaced, we rerun the command to check the number of missing values for each variable.

apply(cars, 2, function (cars) sum(is.na(cars)))

The output in Figure 42 shows that no variables have any missing values as all counts are confirmed to be 0.

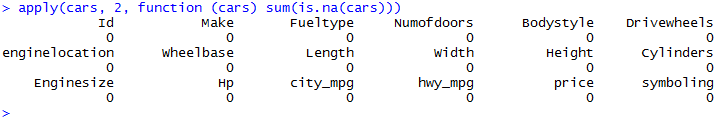


Figure 42: No More Missing Values

## Subset of Rows Matching an Equality Condition

We can take a subset of the data that matches the specified equality condition. For example, the following command takes the data rows where symboling is -2. The first input parameter is the data frame name, and the second input parameter is a condition. **The double equals sign “==” is used for testing equality**.

newcars<-subset.data.frame(cars, cars$symboling=="-2")

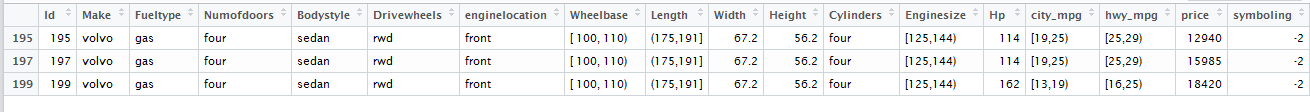


Figure 43: Tabular View – Observations with symboling =2

The following command is an alternative approach to take a subset of data rows where symboling is -2. The data frame name cars is followed by the square brackets. Inside the brackets, we specify the row selection condition followed by a comma. After the comma, we specify the variables we want to select. When the variables are unspecified, all variables are selected.

newcars<-cars [cars$symboling=="-2", ]

## Sorting

We may sort the data frame rows by values of the specified variable. The following command creates a new data frame called cars\_sorted, which is the copy of the cars data frame sorted by price ascending. We specify the name of the data frame followed by square brackets. The first number in the brackets specifies the rows to select. The nested function order(cars$price) returns an array of data row indices for data rows ordered by price value. Since the columns specification after the comma is omitted, the cars\_sorted data frame will contain all columns from cars data frame.

cars\_sorted<-cars[order(cars$price), ]

To validate that the cars\_sorted is ordered by the price descending, we run the following head command to preview the first 6 car prices on Figure 44.

By default, the head command displays the first 6 rows. When the variable name is specified, the command returns value only for the specified variable.

head(cars\_sorted$price)



Figure 44: First 6 Sorted Car Prices

# Saving Data to CSV File

To save the pre-processed version of the CSV file to the working directory on your hard drive, use the write.csv command. The first parameter is the data frame name. The second parameter is the file name. The third parameter instructs R not to add row names.

write.csv(cars, file="CarsD.csv", row.names=FALSE)

To validate if the file was saved, run the dir() command to list the files in the working directory. Figure 45 shows that the file CarsD.csv was saved into the working directory.

dir()



Figure 45: CarsD.csv File in in the Working Directory

Locate the file on a hard drive to preview the content. Figure 1 shows the partial content of the file. The first line contains the variable names, and the remaining lines are the data. The values for Wheelbase, enginesize, city\_mpg, and hwy\_mpg variables are value ranges as transformed previously.

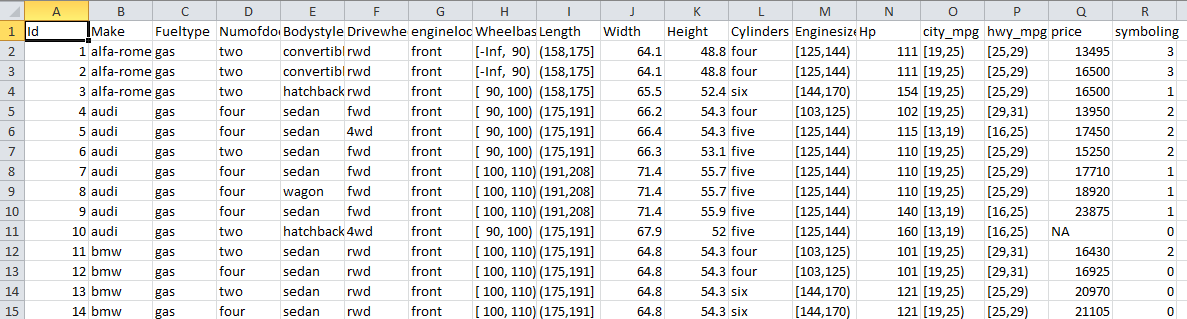


Figure 46: CarsD.csv File Partial Content

# Data visualization

## Generic Plot function

Plots facilitate an understanding of the data, including the central tendency and skewness.

In the plot below, the X-axis coordinates are the symbolling values. The values on the Y-axis are the number of observations that have each value. The chart indicates that 0 and 1 are the dominant symboling values because the bars are the highest at X=0 and then at X=1.

type = h gives us a vertical lines plot that looks similar to a histogram.

plot(table(cars$symboling), type="h", col="dark green")



Delete the current plot

Refresh the plot

Move back and forth between constructed plots

Click Zoom to open the plot in a new window

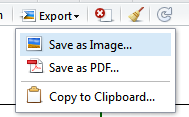
Delete all plots

Export menu allows copying the chart to the clipboard, and saving the chart as a PDF or as an image file.

Figure 47: Vertical Lines Plot

## Copying a Plot to Clipboard

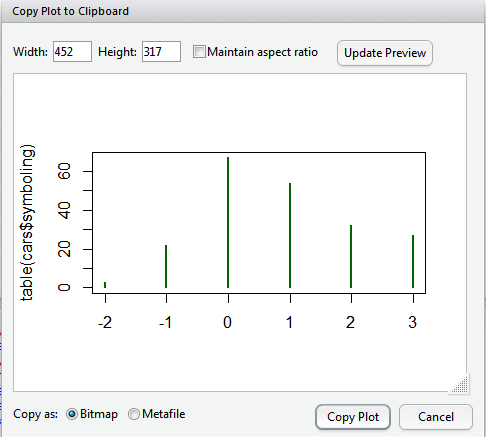
Click on the arrow beside the Export menu to expand it as shown in Figure 48. Select Copy to Clipboard.



Select Copy to Clipboard

Figure 48: Select Copy to Clipboard from Export Menu

The Copy Plot to Clipboard window shown in Figure 49 will open. You may adjust the width and height of the plot by changing the Width and Height value. Click on the Copy Plot button. The window will close. Return to your Word document and select paste from the Edit menu to paste the plot.



Click Copy plot to continue

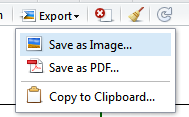
Update the Image preview after the size readjustment

You may adjust the width and Height.

Figure 49: Copy Plot to Clipboard Window

## Saving a Plot as an Image

Click on the arrow beside the Export menu to expand it as shown in Figure 50. Select Save as image.

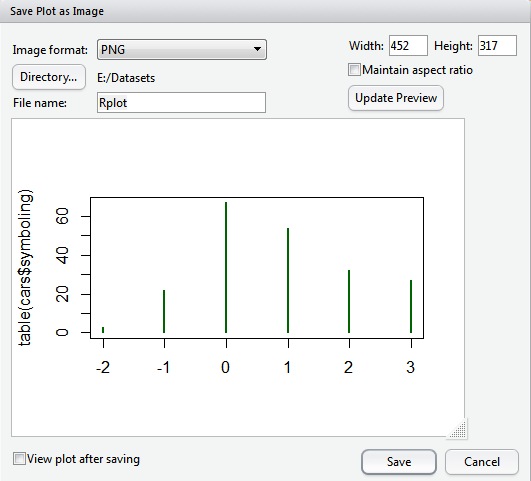


Select Save as Image

Figure 50: Select Save as Image from Export Menu

The Save as Image window in Figure 51 will open. The drop down at the top left allows selecting an image format. In addition to PNG, the supported formats include JPEG, BMP, and SVG.

Click on the Directory button to select the folder where you want to save an image. By default, the chart will be saved to the working directory specified earlier.

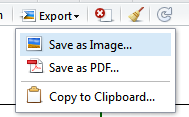


1. Readjust the image size if necessary
2. Click to Update Preview to preview the readjustments.
3. Click Save to continue
4. Choose the image format
5. Browse to the folder here you want to save the image (the default path is a working directory.
6. Enter the name or keep the default name

Figure 51: Save Plot as Image Window

## Saving a plot as a PDF file

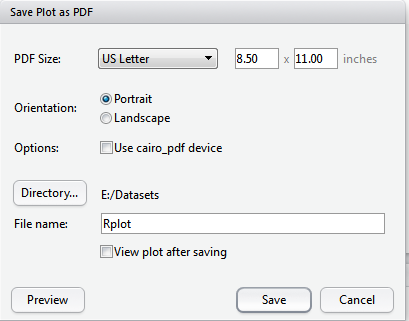
Click on the arrow beside the Export menu to expand it as shown in Figure 52. Select Save as PDF.



Select Save as PDF

Figure 52: Select Save as PDF

The Save Plot as PDF window in Figure 53 will open. Follow the steps to specify the page size, the file name, and the directory. Click Save to continue.



1. Select the Page Size
2. Select Page Orientation
3. Browse to the folder here you want to save the image (the default path is a working directory.
4. Enter the name or keep the default name
5. Click Save to continue.

Figure 53: Save Plot as PDF Window

## Additional Plot Examples

|  |  |
| --- | --- |
| Pie Chart – shows the frequency or relative frequency for the discrete values of a categorical variable.  pie(table (cars$Bodystyle)) |  |
| Histogram  hist(cars$Hp, col="blue")  x-axis attribute name  Data frame name  Bars color |  |
| Boxplot  boxplot(cars$Height, col="maroon") |  |
| Step plot  plot(table(cars$Cylinders), type="s", col="dark red") |  |
| Line plot  plot(table(cars$Bodystyle), type="l", col="dark orange") |  |
| Points Plot  plot(table(cars$Bodystyle), type="p", col="dark magenta") |  |
| Lines and Points Plot  plot(table(cars$Bodystyle), type="b", col="cyan") |  |