Many useful functions are available in many different R packages, many of the same functionalities also in different packages, so it all boils down to user preferences and work, that one decides to use particular package. From the perspective of a statistician and data scientist, I will cover the essential and major packages in sections. And by no means, this is not a definite list, and only a personal preference.

```
###################################
### 1. Loading and importing data
###################################

#1.1. Loading from binary files
# Reading from SAS and SPSS
install.packages("Hmisc", dependencies = TRUE)
# Reading from Stata, Systat and Weka
install.packages("foreign", dependencies = TRUE)
# Reading from KNIME
install.packages(c("protr","foreign"), dependencies = TRUE)
# Reading from EXCEL
install.packages(c("readxl","xlsx"), dependencies = TRUE)
# Reading from TXT, CSV
install.packages(c("csv","readr","tidyverse"), dependencies = TRUE)
# Reading from JSON
install.packages(c("jsonLite","rjson","RJSONIO","jsonvalidate"), dependencies = TRUE)
# Reading from AVRO
install.packages("sparkavro", dependencies = TRUE)
# Reading from Parquet file
install.packages("arrow", dependencies = TRUE)
devtools::install_github("apache/arrow/r")
```

1. Loading and importing data

Loading and read data into R environment is most likely one of the first steps if not the most important. Data is the fuel.

Breaking it into the further sections, reading data from binary files, from ODBC drivers and from SQL databases.

1.1. Importing from binary files

```
# Reading from SAS and SPSS
install.packages("Hmisc", dependencies = TRUE)
# Reading from Stata, Systat and Weka
install.packages("foreign", dependencies = TRUE)
# Reading from KNIME
install.packages(c("protr","foreign"), dependencies = TRUE)
# Reading from EXCEL
install.packages(c("readxl","xlsx"), dependencies = TRUE)
# Reading from TXT, CSV
install.packages(c("csv","readr","tidyverse"), dependencies = TRUE)
# Reading from JSON
install.packages(c("jsonLite","rjson","RJSONIO","jsonvalidate"), dependencies =
TRUE)
# Reading from AVRO
install.packages("sparkavro", dependencies = TRUE)
# Reading from Parquet file
install.packages("arrow", dependencies = TRUE)
devtools::install_github("apache/arrow/r")
# Reading from XML
install.packages("XML", dependencies = TRUE)
```

1.2. Importing from ODBC

This will cover most of the used work for ODBC drives:

```
install.packages(c("odbc", "RODBC"), dependencies = TRUE)
```

1.3. Importing from SQL Databases

Accessing SQL database with a particular package can also have great benefits when pulling data from database into R data frame. In addition, I have added some useful R packages that will help you query data in R much easier (RSQL) or even directly write SQL Statements (sqldf) and other great features.

```
#Microsoft MSSQL Server
install.packages(c("mssqlR", "RODBC"), dependencies = TRUE)
#MySQL
install.packages(c("RMySQL","dbConnect"), dependencies = TRUE)
#PostgreSQL
install.packages(c("postGIStools","RPostgreSQL"), dependencies = TRUE)
#Oracle
install.packages(c("ODBC"), dependencies = TRUE)
#Amazon
install.packages(c("RRedshiftSQL"), dependencies = TRUE)
#SQL Lite
install.packages(c("RSQLite","sqliter","dbflobr"), dependencies = TRUE)
#General SQL packages
install.packages(c("RSQL","sqldf","poplite","queryparser"), dependencies = TRUE)
```

2. Manipulating Data

Data Engineering, data copying, data wrangling and data manipulating data is the very next task in the journey.

2.1. Cleaning data

Data cleaning is essential for cleaning out all the outliers, NULL, N/A values, wrong values, doing imputation or replacing them, checking up frequencies and descriptive and applying different single- , bi-, and multi-variate statistical analysis to tackle this issue. The list is by no means the complete list, but can be a good starting point:

```
install.packages(c("janitor","outliers","missForest","frequency","Amelia",
                   "diffobj","mice","VIM","Bioconductor","mi",
                    "wrangle"), dependencies = TRUE)
```

2.2. Dealing with R data types and formats

Working with correct data types and knowing your ways around handling formatting of your data-set can be overlooked and yet important. List of the must have packages:

```
install.packages(c("stringr","lubridate","glue",
                   "scales","hablar","readr"), dependencies = TRUE)
```

2.3. Wrangling, subseting and aggregating data

There are many packages available to do the task of wrangling, engineering and aggregating, especially {base} R package should not be overlooked, since it offers a lot of great and powerful features. But following is a list of those most widely used in the R community and easy to maneuver data:

```
install.packages(c("dplyr","tidyverse","purr","magrittr",
                   "data.table","plyr","tidyr","tibble",
                   "reshape2"), dependencies = TRUE)
```

3. Statistical tests and Sampling Data

3.1. Statistical tests

Many of the statistical tests (Shapiro, T-test, Wilcox, equality, …) are available in base and stats package that are available with R engine. Which is great, because primarily R is a statistical language, and many of the tests are already included. But adding additional packages, that I have used:

```
install.packages(c("stats","ggpubr","lme4","MASS","car"),
                   dependencies = TRUE)
```

3.2. Data Sampling

Data sampling, working with samples and population, working with inference, weights, and type of statistical data sampling can be find in these brilliant packages, also including those that are great for surveying data.

```
install.packages(c("sampling","icarus","sampler","SamplingStrata",
                    "survey","laeken","stratification","simPop"),
                     dependencies = TRUE)
```

4. Statistical Analysis

Regarding of type of the variable, type of the analysis, and results a statistician wants to get, there are list of packages that should be part of daily R environment, when it comes to statistical analysis.

4.1. Regression Analysis

Frankly, one of the most important analysis

```
install.packages(c("stats","Lars","caret","survival","gam","glmnet",
                  "quantreg","sgd","BLR","MASS","car","mlogit","earth",
                  "faraway","nortest","lmtest","nlme","splines",
                  "sem","WLS","OLS","pls","2SLS","3SLS","tree","rpart"),
dependencies = TRUE)
```

4.2. Analysis of variance

Distribution and and data dispersion is core to understanding the data. Many of the tests for variance are already built-in in R engine (package stats), but here are also some, that might be useful for analyzing variance.

```
install.packages(c("caret","rio","car","MASS","FuzzyNumbers",
                   "stats","ez"), dependencies = TRUE)
```

4.3. Multivariate analysis

Using more than two variables is considered multi-variate analysis. Excluding regression analysis and analysis of variance (between 2+ variables), since it is introduced in section 4.1., covering statistical analysis with working on many variables like factor analysis, principal axis component, canonical analysis, discrete analysis, and others:

```
install.packages(c("psych","CCA","CCP","MASS","icapca","gvlma","smacof",
                  "MVN","rpca","gpca","EFA.MRFA","MFAg","MVar","fabMix",
                  "fad","spBFA","cate","mnlfa","CSFA","GFA","lmds","SPCALDA",
                  "semds", "superMDS", "vcd", "vcdExtra"),
 dependencies = TRUE)
```

4.4. Classification and Clustering

Based on different type of clustering and classification, there are many packages to cover both. Some of the essential packages for clustering:

```
install.packages(c("fpc","cluster","treeClust","e1071","NbClust","skmeans",
                 "kml","compHclust","protoclust","pvclust","genie", "tclust",
                 "ClusterR","dbscan","CEC","GMCM","EMCluster","randomLCA",
                 "MOCCA","factoextra",poLCA), dependencies = TRUE)
```

and for classification:

```
install.packages("tree", "e1071")
```

4.5. Analysis of Time-series

Analysing time series and time-serie type of data will be done easier with the following packages:

```
install.packages(c("ts","zoo","xts","timeSeries","tsModel", "TSMining",
               "TSA","fma","fpp2","fpp3","tsfa","TSdist","TSclust","feasts",
               "MTS", "dse","sazedR","kza","fable","forecast","tseries",
               "nnfor","quantmod"), dependencies = TRUE)
```

4.6. Network analysis

Analyzing networks is also part of statistical analysis. And some of the relevant packages:

```
install.packages(c("fastnet","tsna","sna","networkR","InteractiveIGraph",
                  "SemNeT","igraph","NetworkToolbox","dyads",
                   "staTools","CINNA"), dependencies = TRUE)
```

4.7. Analysis of text

Besides analyzing open text, once can analyse any kind of text, including the word corpus, the semantics and many more. Couple of starting packages:

```
install.packages(c("tm","tau","koRpus","lexicon","sylly","textir",
         "textmineR","MediaNews", "lsa","SemNeT","ngram","ngramrr",
         "corpustools","udpipe","textstem", "tidytext","text2vec"),
          dependencies = TRUE)
```

5. Machine Learning

R has variety of good machine learning packages that are powerfull and give you the full Machine Learning cycle. Breaking down the sections by it's natural way.

5.1. Building and validating the models

Once you build one or more models, after comparing the results of each models, it is also important to validate the models against the test or any other datasets. Here are powerfull packages to do model validation.

```
install.packages(c("tree", "e1071","crossval","caret","rpart","bcv",
                  "klaR","EnsembleCV","gencve","cvAUC","CVThresh",
                  "cvTools","dcv","cvms","blockCV"), dependencies = TRUE)
```

5.2. Random forests packages

sdfs

```
install.packages(c("randomForest","grf","ipred","party","randomForestSRC",
                  "grf","BART","Boruta","LTRCtrees","REEMtree","refr",
                  "binomialRF","superml"), dependencies = TRUE)
```

5.3. Regression type (regression, boosting, Gradient descent) algoritms packages

Regression type of machine learning algorithm are many, with additional boosting or gradient. Some of very usable packages:

```
install.packages(c("earth", "gbm","GAMBoost", "GMMBoost", "bst","superml",
                   "sboost"), dependencies = TRUE)
```

5.4. Classification algorithms

Classifying problems have many of the packages and many are also great for machine learning cases. Handful.

```
install.packages(c("rpart", "tree", "C50", "RWeka","klar", "e1071",
                   "kernlab","svmpath","superml","sboost"),
dependencies = TRUE)
```

5.5. Neural networks

There are many types of Neural networks and many of different packages will give you all types of NN. Only couple of very useful R packages to tackle the neural networks.

```
install.packages(c("nnet","gnn","rnn","spnn","brnn","RSNNS","AMORE",
                   "simpleNeural","ANN2","yap","yager","deep","neuralnet",
                   "nnfor","TeachNet"), dependencies = TRUE)
```

5.6. Deep Learning

R had embraced deep learning and many of the powerfull SDK and packages have been converted to R, making it very usable for R developers and R machine learning community.

```
install.packages(c("deepnet","RcppDL","tensorflow","h2o","kerasR",
                   "deepNN", "Buddle","automl"), dependencies = TRUE)
```

5.7. Reinforcement Learning

Reinforcement learning is gaining popularity and more and more packages are being developered in R as well. Some of the very userful packages:

```
devtools::install_github("nproellochs/ReinforcementLearning")
install.packages(c("RLT","ReinforcementLearning","MDPtoolbox"),
dependencies = TRUE)
```

5.8. Model interpretability and explainability

Results of machine learning models can be a black-box. Many of the packages are dealing to have black-box more like “glass box”, making the models more understandable, interpretable and explainable. Very powerfull packages to do just that for many different machine learning algorithms.

```
install.packages(c("lime","localModel","iml","EIX","flashlight",
                    "interpret","outliertree","breakDown"),
dependencies = TRUE)
```

6. Visualisation

Visualisation of the data is not only the final step to understanding the data, but can also bring clarity to interpretation and buidling the mental model around the data. Couple of packages, that will help boost the visualization:

```
install.packages(c("ggvis","htmlwidgets","maps","sunburstR", "lattice",
  "predict3d","rgl","rglwidget","plot3Drgl","ggmap","ggplot2","plotly",
  "RColorBrewer","dygraphs","canvasXpress","qgraph","moveVis","ggcharts",
  "igraph","visNetwork","visreg", "VIM", "sjPlot", "plotKML", "squash",
  "statVisual", "mlr3viz", "klaR","DiagrammeR","pavo","rasterVis",
  "timelineR","DataViz","d3r","d3heatmap","dashboard" "highcharter",
  "rbokeh"), dependencies = TRUE)
```

7. Web Scraping

Many R packages are specificly designed to scrape (harvest) data from particular website, API or archive. Here are only couple of very generic:

```
install.packages(c("rvest","Rcrawler","ralger","scrapeR"),
             dependencies = TRUE)
```

8. Documents and books organisation

Organizing your documents (file, code, packages, diagrams, pictures) in readable document and have it as a dashboard or book view, there are couple of packages for this purpose:

```
install.packages(c("devtools","usethis","roxygen2","knitr",
                    "rmarkdown","flexdashboard","Shiny",
                    "xtable","httr","profvis"), dependencies = TRUE)
```

Wrap up

The R script for loading and installing the packages is available at Github. Make sure to check the Github repository for latest list updates. And as always, feel free to fork the code or commit updates, add essentials packages to list, comment, improve and agree or disagree.

You can also run the following command to install all of the packages in a single run:

```
install.packages(c("Hmisc","foreign","protr","readxl","xlsx",
                 "csv","readr","tidyverse","jsonLite","rjson",
                 "RJSONIO","jsonvalidate","sparkavro","arrow","feather",
                 "XML","odbc","RODBC","mssqlR","RMySQL",
                 "dbConnect","postGIStools","RPostgreSQL","ODBC",
                 "RSQLite","sqliter","dbflobr","RSQL","sqldf",
                 "poplite","queryparser","influxdbr","janitor","outliers",
                 "missForest","frequency","Amelia","diffobj","mice",
                 "VIM","Bioconductor","mi","wrangle","mitools",
                 "stringr","lubridate","glue","scales","hablar",
                 "dplyr","purr","magrittr","data.table","plyr",
                 "tidyr","tibble","reshape2","stats","Lars",
                 "caret","survival","gam","glmnet","quantreg",
                 "sgd","BLR","MASS","car","mlogit","RRedshiftSQL",
                 "earth","faraway","nortest","lmtest","nlme",
                 "splines","sem","WLS","OLS","pls",
                 "2SLS","3SLS","tree","rpart","rio",
                 "FuzzyNumbers","ez","psych","CCA","CCP",
                 "icapca","gvlma","smacof","MVN","rpca",
                 "gpca","EFA.MRFA","MFAg","MVar","fabMix",
                 "fad","spBFA","cate","mnlfa","CSFA",
                 "GFA","lmds","SPCALDA","semds","superMDS",
                 "vcd","vcdExtra","ks","rrcov","eRm",
                 "MNP","bayesm","ltm","fpc","cluster",
                 "treeClust","e1071","NbClust","skmeans","kml",
                 "compHclust","protoclust","pvclust","genie","tclust",
                 "ClusterR","dbscan","CEC","GMCM","EMCluster",
                 "randomLCA","MOCCA","factoextra","poLCA","ts",
                 "zoo","xts","timeSeries","tsModel","TSMining",
                 "TSA","fma","fpp2","fpp3","tsfa",
                 "TSdist","TSclust","feasts","MTS","dse",
                 "sazedR","kza","fable","forecast","tseries",
                 "nnfor","quantmod","fastnet","tsna","sna",
                 "networkR","InteractiveIGraph","SemNeT","igraph",
                 "dyads","staTools","CINNA","tm","tau","NetworkToolbox"
                 "koRpus","lexicon","sylly","textir","textmineR",
                 "MediaNews","lsa","ngram","ngramrr","corpustools",
                 "udpipe","textstem","tidytext","text2vec","crossval",
                 "bcv","klaR","EnsembleCV","gencve","cvAUC",
                 "CVThresh","cvTools","dcv","cvms","blockCV",
                 "randomForest","grf","ipred","party","randomForestSRC",
                 "BART","Boruta","LTRCtrees","REEMtree","refr",
                 "binomialRF","superml","gbm","GAMBoost","GMMBoost",
                 "bst","sboost","C50","RWeka","klar",
                 "kernlab","svmpath","nnet","gnn","rnn",
                 "spnn","brnn","RSNNS","AMORE","simpleNeural",
                 "ANN2","yap","yager","deep","neuralnet",
                 "TeachNet","deepnet","RcppDL","tensorflow","h2o",
                 "kerasR","deepNN","Buddle","automl","RLT",
                 "ReinforcementLearning","MDPtoolbox","lime","localModel",
                 "iml","EIX","flashlight","interpret","outliertree",
                 "dockerfiler","azuremlsdk","sparklyr","cloudml","ggvis",
                 "htmlwidgets","maps","sunburstR","lattice","predict3d",
                 "rgl","rglwidget","plot3Drgl","ggmap","ggplot2",
                 "plotly","RColorBrewer","dygraphs","canvasXpress","qgraph",
                 "moveVis","ggcharts","visNetwork","visreg","sjPlot",
                 "plotKML","squash","statVisual","mlr3viz","DiagrammeR",
                 "pavo","rasterVis","timelineR","DataViz","d3r","breakDown",
                 "d3heatmap","dashboard","highcharter","rbokeh","rvest",
                 "Rcrawler","ralger","scrapeR","devtools","usethis",
                 "roxygen2","knitr","rmarkdown","flexdashboard","Shiny",
                 "xtable","httr","profvis"), dependencies = TRUE)
```

Happy R-ing.