Auteur: Arie Twigt

Druk: bravenewbooks.nl
Omslagontwerp: Rosemarijn Twigt
Vormgeving binnenwerk: Rosemarijn Twigt

Data Analyseren met R
ISBN 9789402132687

NUR 000

Inhoud

Voorwoord

Allereerst, wil ik je bedanken dat je dit boek erbij pakt om kennis te maken met de open source statistische programmeertaal R. Daarbij wil ik gelijk mijn vrouw bedanken die voor de illustraties heeft gezorgd in dit boek.

Ook wil ik **Dylan van Riel** en **Jonathan de Melker Worms** bedanken voor de uitgebreide review van dit boek. Dankzij hun tips en scherpe reviews durf ik dit boek meer met een gerust hart te publiceren.

Oké, en dan nu naar de inhoud. De programmeertaal R is de afgelopen jaren een van mijn favoriete tools geworden voor het analyseren van data. Voor het ontwikkelen van software kies ik tegenwoordig meestal voor Python, maar als het puur gaat om het uitgebreid analyseren en rapporteren van data vind ik dat R hier betere tools voor heeft. Daarbij is R oorspronkelijk gebouwd met het doel om data te analyseren en te bewerken.

Hoe ben ik ooit in aanraking gekomen met R en waarom ben ik er zo enthousiast over geworden? Het begon tijdens mijn studie **Business Administration - Information & Knowledge Management** op de Vrije Universiteit Amsterdam. Tijdens het vak **Business Intelligence** werden er onderwerpen behandeld zoals **Data Mining, Big Data, Machine Learning** en **Data Analytics**. Om deze onderwerpen samen te vatten: waardevolle informatie halen uit data.

Omdat ik tussen het bestuderen van papers ook graag de praktijk in wilde duiken, ben ik hiermee naar de professor gegaan, Frans Feldberg, die mij adviseerde om eens een kijkje te nemen naar de statistische programmeertaal **R**. Ik moet toegeven dat werken met R in het begin vrij uitdagend was. Uiteindelijk was ik er zo enthousiast over geworden dat ik er bijna dagelijks mee experimenteerde. Uit enthousiasme heb ik daar

uiteindelijk een handleiding voor geschreven op Google Blogger. Dat is niet deze handleiding, maar de handleiding die ik in 2013 heb geschreven: Handleiding R op **r-handleiding.blogspot.com**. Deze handleiding is door de komst van dit boek achterhaald, zeer achterhaald.

In de afgelopen jaren heb ik bij verschillende klanten oplossingen met R geïmplementeerd of de klant laten kennismaken met R. Hierbij heb ik gezien dat R over het algemeen positief wordt ontvangen en het zeer toepasbaar is in het bedrijfsleven. Voorbeelden zijn klanten die voor processen met complexe berekeningen overstappen van pakketten als MatLab, SPSS of Excel naar R. Het is makkelijk te implementeren, de mogelijkheden zijn eindeloos en het is gratis.

Het feit dat R open source is, zorgt af en toe voor een licht drempeltje bij klanten. Wat heeft bijgedragen om over deze drempel heen te stappen is dat steeds meer grote gerenommeerde bedrijven gebruik maken van Open Source Software, waarondre R als het om data gaat. Gelukkig heeft Microsoft sinds 2015 R in verschillende software geïntegreerd, bijvoorbeeld op **Microsoft Azure**, **Microsoft Visual Studio** (met **R tools for Visual Studio**) en **Microsoft SQL Server Management Studio 2016**. Ook heeft Microsoft een eigen versie van R, **Microsoft R open**. Deze ontwikkeling zorgde ervoor dat mensen meer ontvankelijk werden om R in te zetten voor data-gerelateerde projecten.

Nogmaals, ik wil je erg bedanken dat je dit boek erbij hebt gepakt. Met de opbouw en alle instructies wil ik de denkwijze die ik heb en de methoden die ik gebruik voor het analyseren van data naar jou overbrengen. Dit zijn in dit boek vooral de basics. Volg mijn Linkedin-pagina (https://www.linkedin.com/in/arietwigt/) voor mijn posts over experimenten in R die meer uitgebreide onderwerpen behandelen.

Natuurlijk zijn er mensen die dit op een andere manier aanpakken, dit zal je naar verloop van tijd zelf ook doen als je het aardig onder de knie hebt. In ieder geval zal dit boek je helpen met het onder de knie krijgen van de basics en zal je alle aspecten bezitten om uitgebreid data te kunnen analyseren met R. Ik ben ervan overtuigd dat je

met dit boek net zoveel plezier zult hebben als het plezier dat ik heb ervaren bij het samenstellen van dit boek.

Inleiding

R wordt over het algemeen een open source statistische programmeertaal genoemd. Deze beschrijving is een hele mond vol, maar bevat wel de belangrijkste aspecten van R:

- Open source;
- Statistisch;
- Programmeertaal.

Laten we deze 3 aspecten even uitgebreid behandelen.

Open source

R is een open source programmeertaal. Dit betekent o.a. dat niemand, geen organisatie en geen persoon, de eigenaar is van R en je het dus overal gratis voor kunt gebruiken. Daarbij geeft **open source** aan dat de broncode waarmee de programmeertaal is opgebouwd, **open** en **beschikbaar** is. Je kunt bijvoorbeeld kijken naar de broncode achter de functies in R. Een ander belangrijk aspect van open source in dit geval is dat je alle vrijheid hebt om onderdelen aan te passen en er extra mogelijkheden aan kunt toevoegen. De mogelijkheden hierin zijn eindeloos. Hiermee bedoel ik vooral dat er geen zaken zijn die bewust zijn afgesloten of dichtgetimmerd.

Een belangrijk voorbeeld hiervan zijn de verschillende **packages** die beschikbaar zijn in R. **Packages** zijn verzamelingen van extra functies die geschreven zijn door andere ontwikkelaars en over het algemeen vrij beschikbaar worden gesteld. Deze packages maken de programmeertaal zeer krachtig, flexibel en toepasbaar voor verschillende werkzaamheden, sectoren en branches. Zo zijn er bijvoorbeeld packages die het

makkelijk maken om **Twitter** data te importeren (het **twitteR** package), packages die formules van financiële calculaties bevatten (het **FinCal** package), packages waarmee je marketing analyses kunt uitvoeren door Google Analytics data te importeren in R (het officiële **RGoogleAnalytics** package van Google) of packages om op een prachtige manier data te visualiseren (bijvoorbeeld het **ggplot2** package). Dit maakt R een Zwitsers zakmes als het gaat om werken met data. Dit is voor een groot deel te danken is aan het feit dat het een open source programmeertaal is. Met een actieve community van ontwikkelaars die de programmeertaal R steeds verder verbeteren en uitbreiden, is er in R steeds meer mogelijk om te werken met data.

> **Packages zijn goed gedocumenteerd** en worden beheerd in de Comprehensive R Archive Network (CRAN). Dit is ook de plek waar over het algemeen de belangrijkste packages staan en worden beheerd door de community van R. Deze packages hebben aan een aantal belangrijke criteria voldaan zoals software tests. De documentaties van deze packages zijn over het algemeen te vinden in de vorm van een overzichtelijk pdf-bestand. In deze documentatie kun je alle functies van het package bekijken, contactgegevens vinden van de ontwikkelaar en met voorbeelden zien hoe functies van het package worden toegepast. Naast CRAN, zijn er ook steeds meer packages te vinden op **Github**. Deze werken gewoon op dezelfde manier als packages die op CRAN staan, de installatie is alleen iets anders. In dit boek wordt er ook verschillende keren gebruik gemaakt van packages. Je zult zien dat packages zeer eenvoudig te installeren en te gebruiken zijn.

Statistisch

R wordt een statistische programmeertaal genoemd omdat het standaard een uitgebreide collectie statistische functies bevat. Dit betekent absoluut niet dat R beperkt wordt tot het vakgebied statistiek of dat het minder handig is voor andere vakgebieden. De reden dat R deze statistische functies bevat, is omdat het oorspronkelijk ontwikkeld is om (statistisch) onderzoek mee uit te voeren. Je kunt het hierbij vergelijken met softwarepakketten als **SPSS** en **MatLab**. Je zou op dezelfde

manier kwantitatief onderzoek kunnen doen of je scriptie kunnen schrijven in R. Statistische toetsen, visualisaties of formules zijn zeer eenvoudig met functies toe te passen in R zonder dat je daar een speciaal **package** of bepaalde plug-in voor hoeft te installeren. Deze functies maken het data analyseren veel makkelijker. Naast de ingebakken statistische functies bevat R namelijk een breed scala aan functies om databewerkingen of analyses uit te voeren.

De veelzijdigheid van R komt goed van pas bij analyses waarbij uitgebreide of statistische toepassingen nodig zijn. Dit zorgt ervoor dat je een extreem krachtige tool in handen hebt om data te analyseren, zelfs als het een stuk ingewikkelder wordt. Daarom kun je het statische aspect van R eerder zien als een waardevolle toevoeging dan een beperking.

Programmeertaal

Het feit dat R een programmeertaal is, zorgt voor veel flexibiliteit. Het werk dat je met R hebt gedaan, is in de meeste gevallen een reeks aan R commands. Een R command is een regel code waarmee je een opdracht stuurt naar R om iets uit te voeren. Deze reeks aan commands kun je uiteindelijk samenvoegen tot een **script**. Uiteindelijk zelfs een hele verzameling aan scripts en andere bestanden. In een script worden de commands van boven naar beneden uitgevoerd, net zoals bij andere script-gebaseerde programmeertalen (**scripting languages**). Een script in R wordt een **R-script** genoemd en heeft de **.R** extensie (bijvoorbeeld: **analyse.R**). Je kunt ook de korte letter gebruiken zoals : **analyse.r**. Een R-script kun je overal mee naartoe nemen en in verschillende systemen toepassen. Zo kun je bijvoorbeeld R-code toepassen om data in een database aan te passen. Verschillende databases zijn namelijk in staat om R-code uit te voeren (bijvoorbeeld **Oracle** en **Microsoft SQL Server (vanaf 2016)**. Ook op

Big Data platforms en software zoals **Hadoop**, **Spark** kun je R code schrijven om data te bewerken.

Het werk dat je maakt in R is dus op veel verschillende plekken toepasbaar en kun je eenvoudig meenemen en delen. Laten we het bijvoorbeeld even vergelijken met Excel: iets dat in Excel is gemaakt blijft in Excel. De logica en code die je in Excel hebt gebouwd, kun je niet overhevelen naar een ander systeem (hooguit kun je de data delen), deze zitten namelijk in het Excel-bestand opgeslagen. Functies en handelingen die je bouwt in R kunnen op veel meer plekken worden gebruikt. Dit geldt trouwens voor de meeste programmeertalen, de code is los te trekken van de omgeving waar het in is gebouwd.

Programmeren klinkt voor sommigen misschien heel eng en code kan er soms heel ingewikkeld uitzien. Je moet je echter hierdoor niet laten afschrikken. Het is natuurlijk nodig dat je de syntax van R begrijpt en weet wanneer je welke code moet toepassen. Hier zit een leercurve in die bij R soms als vrij stijl wordt ervaren. Echter kun je deze overwinnen met genoeg oefening waardoor je een beter begrip krijgt van de syntax en de structuur. Hierdoor zal je voor jezelf een eigen manier zal vinden om R goed onder de knie te krijgen. Als je hebt gewerkt met formules in Excel, kun je sommige onderdelen van die kennis meenemen om R te leren begrijpen. Het is daarbij ook een verschil in mindset: R is geen software waarmee je op knopjes drukt en kan klikken en slepen, maar een programmeertaal waarmee aan de hand van commands een reeks handelingen kunt uitvoeren. Je gaat voornamelijk code schrijven.

R is flexibel door de vele mogelijkheden, de vele verschillende datatypes en de vele datastructuren. Maar deze flexibiliteit kan als complex en ingewikkeld worden ervaren als je er zomaar in duikt en begint zonder enig begrip van de structuur te hebben. Daarom wordt er in dit boek stap voor stap de belangrijkste structuren en datatypes in R uitgebreid behandeld.

De structuur van R

Het is de bedoeling dat je na het lezen en het volgen van de instructies in dit boek programmeren en data analyseren met R goed onder de knie hebt. Daarom beginnen we bij het begin, namelijk de bouwstenen van de structuur in R. Zonder goede kennis van de structuur in R loop je het gevaar dat je tijdens het analyseren van data in de knoop komt, bijvoorbeeld door dingen te willen doen die qua structuur helemaal niet kunnen.

Data bewerken

In bijna elk data-analyse project moet de data eerst worden bewerkt voordat het geanalyseerd kan worden. De instructies in dit boek geven je goede voorbeelden hoe je dit in R kunt doen. In veel gevallen kan data, na het ondergaan van een aantal bewerkingen, goed geanalyseerd worden. R heeft een breed scala aan mogelijkheden om data te bewerken. De bewerkingen in dit boek zijn qua niveau logisch opgebouwd van eenvoudig naar geavanceerd. Daarbij zijn veel instructies en voorbeelden van bewerkingen gebaseerd op situaties die ik tijdens projecten bij klanten ben tegengekomen. Ik zal je dus geen nutteloze trucjes of handelingen leren, in ieder geval zo weinig mogelijk. Wie weet kun je een bepaalde functie straks direct tijdens je werk toepassen.

Data analyseren

De mogelijkheden om data te analyseren zijn eindeloos. Zowel visueel of bijvoorbeeld statistisch. Ik heb nooit de illusie gehad om alle mogelijkheden in één boek te stoppen.

Echter zit er wel een rode lijn in dit boek op het gebied van data analyseren. Onderwerpen zijn bijvoorbeeld analyses door te **slicen en dicen** zoals de Business Intelligence wereld dit noemt, statistische functies, visualisaties en speciale R-functies. Ook deze heb ik toegepast bij klanten en hebben verschillende keren tot waardevolle inzichten geleid. Dit kan voor jou, of voor het bedrijf waar je werkt, ook al snel tot waardevolle inzichten leiden.

Geen big data

Dit boek behandelt **niet** het onderwerp **Big Data**. Dit is, even heel simpel uitgelegd, data die te groot is om door je eigen computer of laptop bewerkt te kunnen worden. Hiervoor heb je bijvoorbeeld geavanceerde software nodig om meerdere computers aan elkaar te koppelen voor meer rekenkracht of moet je gebruik maken van cloud oplossingen. Bekende software-omgevingen of tools om dit te doen zijn bijvoorbeeld **Apache Hadoop** en **Apache Spark**. Tegenwoordig zijn er ook verschillende Microsoft-producten beschikbaar zoals **Microsoft R** server of **Microsoft Azure Batch Services**. Het leuke aan deze software is dat je er R in kunt programmeren. Hierdoor wordt het mogelijk om R-bewerkingen toe te passen op gigantische datasets, datasets met tientallen of zelfs honderden miljoenen rijen aan data. Als je R onder de knie hebt, kun je altijd naar de websites gaan die deze **Big Data** oplossingen aanbieden. Je zult de R-code herkennen en daardoor de stap kunnen maken naar **Big Data.**

> In het artikel **Microsoft Azure Batch: Doorlooptijden terugbrengen van dagen naar uren**. (https://arietwigt.nl/2016/09/17/microsoft-azure-batch-doorlooptijden-terugbrengen-van-dagen-naar-uren/) leg ik uit hoe je bijvoorbeeld met Microsoft Azure Batch Services op een simpele manier R calculaties in parallel kunt laten uitvoeren. Dit is slechts een van de opties om veel en zware calculaties uit te voeren in R als je eigen computer het niet aan kan.

Geen uitgebreide statistiek of econometrie

Naast **Big Data** wordt er in dit boek ook geen uitgebreide uitleg gegeven over geavanceerde wiskundige modellen. Je krijgt echter wel voldoende informatie over de statistische functies die je kunt toepassen in R. Bijvoorbeeld voor het onderwerp **Lineaire regressieanalyse** wordt er in dit boek de statische betekenis en onderbouwing van de output van de functie uitgelegd. Hiermee kun je de uitkomsten van het model goed interpreteren. Een lineaire regressie is een zeer populair voorspellingsmodel, maar slechts een van de vele mogelijkheden van voorspellende analyses. Ondanks dit feit is het een mooi begin en zal goede kennis van R, die overgebracht wordt in dit boek, je helpen als je meer uitgebreide en complexe Machine Learning modellen wilt gebruiken.

Wat je leert

Dit boek kan jouw eerste kennismaking zijn met R. Als je al vaker met R hebt gewerkt, is dit boek alsnog een goede toevoeging aan jouw kennis over R. De opbouw en de structuur van R worden namelijk op een zo duidelijk en compleet mogelijke manier uitgelegd. In dit boek wordt er zoveel mogelijk een afweging gemaakt om de basics uit te leggen zonder dat het geestdodend eenvoudig en saai is. Daarbij zijn onderwerpen die te diep gaan en over het algemeen voor werkzaamheden niet relevant zijn, weggelaten. Het is de bedoeling dat je aan het einde van dit boek de belangrijkste onderdelen van de structuur van R onder de knie hebt. Met deze kennis zal je aan het einde van dit boek R goed in de vingers hebben en snelheid opbouwen bij het programmeren in R. Ook is dit boek een mooi naslagwerk waarmee je verschillende commands kunt opzoeken. Je hoeft er namelijk niet naar te streven om de hele syntax uit je hoofd te leren, zelfs de beste programmeurs gebruiken Google. Sterker nog…. zo goed als iedere programmeur gebruikt Google.

Wat je eraan hebt

R is een waardevolle toevoeging aan je skillset en CV. Vooral bij vacatures voor een rol als bijvoorbeeld Data Scientist staat R vaak op het lijstje van vereiste of gewenste skills. Andere tools als bijvoorbeeld Python kom je ook vaak tegen, de laatste tijd misschien iets vaker dan R. Python is ook een uitstekende tool om data te analyseren. Van oorsprong is Python meer een echte ontwikkel-taal en geeft het daarom meer mogelijkheden bij het ontwikkelen van data gedreven applicaties. Ook is Python wat meer geschikt dan R voor het ontwikkelen van Artificial Intelligence, omdat de community hier meer inzet in Python dan in R. Alhoewel je met R ook applicaties kunt bouwen of A.I. kunt ontwikkelen, zou je grofweg kunnen zeggen dat je voor de beste ervaring bij het uitgebreid analyseren van data kiest voor R en als je iets wilt ontwikkelen je kiest voor echte ontwikkel-talen zoals Python, Java of Scala. Daar staat wel tegenover dat steeds meer gerenommeerde software-pakketen kiezen voor de integratie met R (en tegelijkertijd ook met Python) zoals bijvoorbeeld Microsoft dat doet. Zelf heb ik bijvoorbeeld een opdracht gedaan met R bij een grote verzekeraar die over het algemeen een Microsoft IT-architectuur heeft. Omdat Microsoft Visual Studio en Microsoft SQL Server een uitgebreide R plugin hebben, was R de taal waarmee de uitgebreide calculaties geprogrammeerd werden. Om deze en nog veel meer andere redenen is programmeren in R een waardevolle skillset waar je bij bedrijven en in de toekomst veel aan zult hebben.

Samenvattingen en oefeningen

Ieder hoofdstuk eindigt met een samenvatting en oefeningen. Omdat het een boek is welke je kunt volgen door instructies uit te voeren in jouw eigen R-sessie, kan het afronden van een hoofdstuk langer duren dan als je het alleen leest. Een samenvatting geeft een mooie afronding van het hoofdstuk en vat mooi samen welke onderwerpen

er in het hoofdstuk zijn besproken. Naast de samenvattingen heeft ieder hoofdstuk ook twee of drie oefeningen. Deze oefeningen zijn gemaakt om je, los van de instructies, stukken R code te laten schrijven waardoor je wordt gedwongen je eigen opgedane kennis toe te passen. Hiervoor kun je hulp vinden in het betreffende hoofdstuk, maar voor sommige vragen moet je zelf een creatieve manier vinden of Google raadplegen. Hierdoor word je nog beter getraind en ervaar je hoe het is om ergens tegen aan te lopen en om dit vervolgens op te lossen. Mocht je er echt niet uit komen, kun je de antwoorden van de oefeningen vinden aan het einde van dit boek in het hoofdstuk **Antwoorden voor oefeningen**.

Termen

In dit boek zijn er een aantal termen die ik regelmatig gebruik. Ik zal hierbij van tevoren uitleggen wat ik met deze termen bedoel.

Variabele

Een variabele is een object dat een bepaalde waarde bevat. De waarde die aan een variabele wordt toegewezen, kan verschillende vormen hebben zoals een getal, tabel, visualisatie of zelfs voorspellingsmodel.

Object

Een object is een entiteit in R. Er zijn verschillende typen objecten in R zoals data frames (tabellen), matrixen, lijsten (lists) en vectoren. Ieder type object vervult zijn

eigen functie tijdens een R-sessie. Hierbij verschillen de objecten qua mogelijkheden om deze te bewerken. Je zou kunnen stellen dat eigenlijk alles in R een object is.

Working Directory

De **Working Directory** is een belangrijk onderdeel in R. Het is namelijk de map op jouw computer waar de R-sessie op zoek gaat naar bestanden als deze geïmporteerd moeten worden of waar R-bestanden naar exporteert, bijvoorbeeld databestanden of andere **R**-scripts. In programmeer-termen kun je zeggen dat de **Working Directory** een ander woord is voor de **root directory** van de huidige R-sessie. Je kunt het `getwd()` command gebruiken om te kijken wat de **working directory** voor de huidige R-sessie is. Met het `setwd()` command kun je de **working directory** veranderen.

Project

Een project is eigenlijk geen R-entiteit. Het is een bestand (**.rproj**) dat aangemaakt wordt door RStudio en kan daarom ook alleen door RStudio worden gebruikt. Een project is een handige **wrapper** voor de gehele omgeving van de R-sessie. Als je bijvoorbeeld bezig bent met een bepaalde analyse en je hebt hiervoor verschillende bestanden geopend, waarden aangemaakt en een working directory ingesteld, wordt dit allemaal bijgehouden en teruggehaald zodra je het project weer opent.

Functie

Een functie is een reeks aan handelingen en bewerkingen op waarden in R. Net zoals bij andere programmeertalen zijn functies een van de belangrijkste bouwstenen van R. R heeft standaard een breed scala aan functies waarmee je uitgebreid data kunt

analyseren. Daarbij kun je packages installeren die andere functies bevatten die R niet heeft. Deze packages zijn verzamelingen van functies die andere personen gemaakt hebben en beschikbaar stellen. Ten slotte kun je zelf functies schrijven en jouw verzameling aan functies samenbrengen onder een package welke je eventueel beschikbaar stelt voor andere personen. Op deze manier is de cirkel mooi rond en blijft de programmeertaal R bloeien.

Dataset

Een dataset is een verzameling gegevens die in een R-sessie kan worden ingeladen. Dit kan in de vorm van een tabel, matrix, of lijst zijn. Deze datasets kun je bewerken of gebruiken om er verschillende functies op toe te passen. Door dit op verschillende manieren te doen, kun je verschillende inzichten uit een dataset halen.

Legenda

Commentaar

In R is het mogelijk om commentaar te geven in het script. Deze regels worden door R overgeslagen tijdens het verwerken van de code. Commentaar is een handig middel om aan te geven wat er gebeurt bij verschillende plekken in het script, voor jezelf en ook voor anderen die het script lezen.

```
# dit is commentaar
```

Output uit de R-console

In sommige voorbeelden geeft R output terug. Deze output wordt aangegeven met ##.

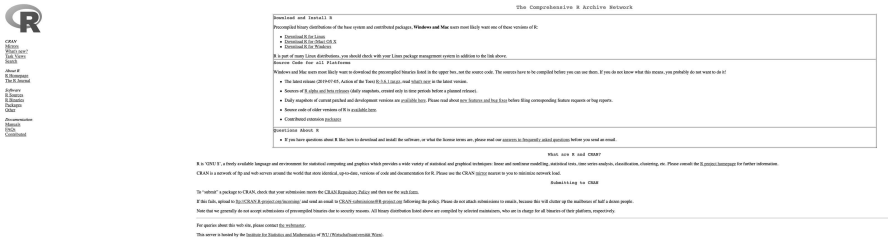
```
print("Dit is output :)")

## [1] "Dit is output :)"
```

Installatie software en bestanden

Ik adviseer je om RStudio pas te installeren als R succesvol is geïnstalleerd. Het is mogelijk om RStudio te installeren zonder dat je R hebt geïnstalleerd, alleen werkt het pas als R ook is geïnstalleerd. RStudio maakt namelijk gebruik van R. Als je RStudio installeert nadat je R hebt geïnstalleerd, zoekt RStudio automatisch naar de locatie van R op jouw computer. Dit hoef je dan later zelf niet meer te doen. Als je dit andersom doet, is er de kans dat je dit later in RStudio zelf moet configureren.

R installeren

Op de officiële website van R, **The Comprehensive R Archive Network**(CRAN) (https://cran.r-project.org), kun je R downloaden. Zoals je kunt zien is R voor verschillende besturingssystemen beschikbaar: Windows, Mac OSX en Linux.



Figuur 1 De homepage van CRAN

Voor Windows, kies je voor **Download R for Windows.** Op de volgende pagina kies je voor **install R for the first time**. Hiermee download je de normale **base** versie van R, de standaard R met alle standaard functies. Simpel gezegd, "hiermee download je R." Hierna volgt een wizard waarmee R wordt geïnstalleerd. Kies zo veel mogelijk voor de al standaard keuzes. Pas eventueel aan welke snelkoppelingen je wel en niet wilt hebben.

Als je een Mac OS-gebruiker bent, kies je voor **Download R for (Mac) OS X**. Hierdoor kom je op een pagina waar verschillende download packages beschikbaar zijn, kies voor de meest recente versie. Als het package eenmaal gedownload is, kun je de eenvoudige stappen van het installatieproces volgen. De meest gebruiksvriendelijke keuzes zijn al door de installatie wizard geselecteerd.

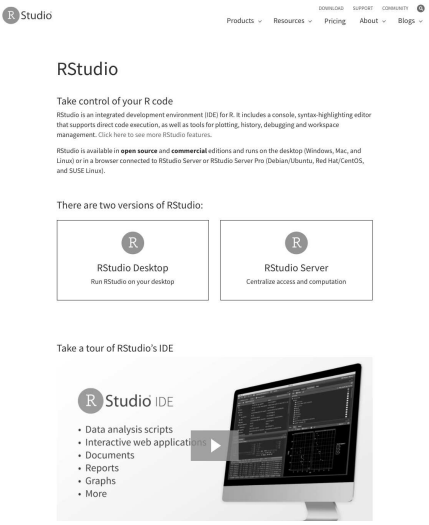
RStudio installeren

Om RStudio te installeren ga je naar:

https://www.rstudio.com/products/rstudio/#Desktop.

Kies voor de **Open Source Edition** door op **Download RStudio Desktop** te klikken. Hierdoor kom je op de pagina met alle beschikbare versies van RStudio voor verschillende platforms. Kies voor het platform waar je mee wilt werken. Volg daarna

de instructies van de wizard. Kies ook hier zoveel mogelijk voor de standaard mogelijkheden.



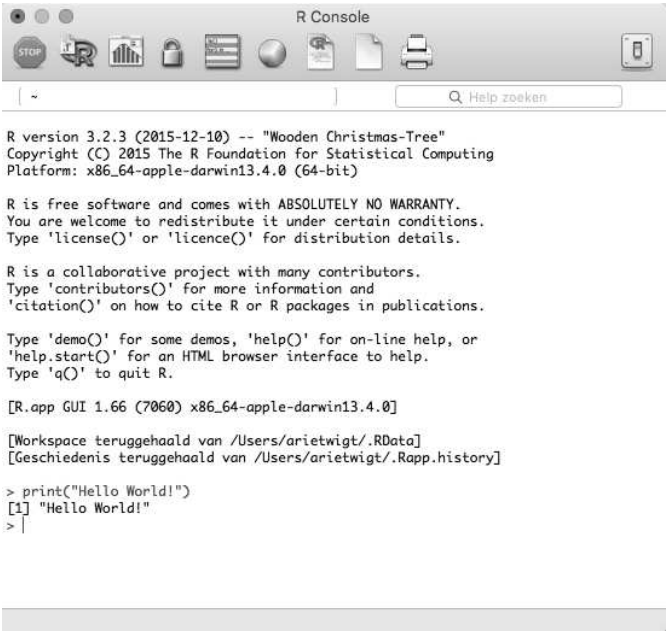
Figuur 2 De homepage van RStudio

R en RStudio

RStudio is een zeer uitgebreide en zeer gebruiksvriendelijke ontwikkelomgeving voor R. Het geeft je als gebruiker een veel beter beeld van waar je mee bezig bent tijdens een R-sessie. Daarbij heeft RStudio veel hulpmiddelen die je kunt gebruiken door erop te klikken. Ik adviseer je daarom om RStudio te gebruiken. Zo goed als iedere R-ontwikkelaar gebruikt RStudio en het is al een lange tijd met afstand de beste omgeving om met R te werken. Mocht je toch liever de normale R-omgeving gebruiken, staat niets je daarvoor in de weg. Ook dan kun je alle instructies in dit boek probleemloos volgen.

De R-console

De R-console is een venster waar je commands kunt invoeren. Het invoeren van commands is wat we in dit boek voor het gemak **programmeren** noemen. Eigenlijk spreken we pas van programmeren als we een script bouwen dat uit een verzameling commands bestaat en dit door een systeem wordt uitgevoerd.



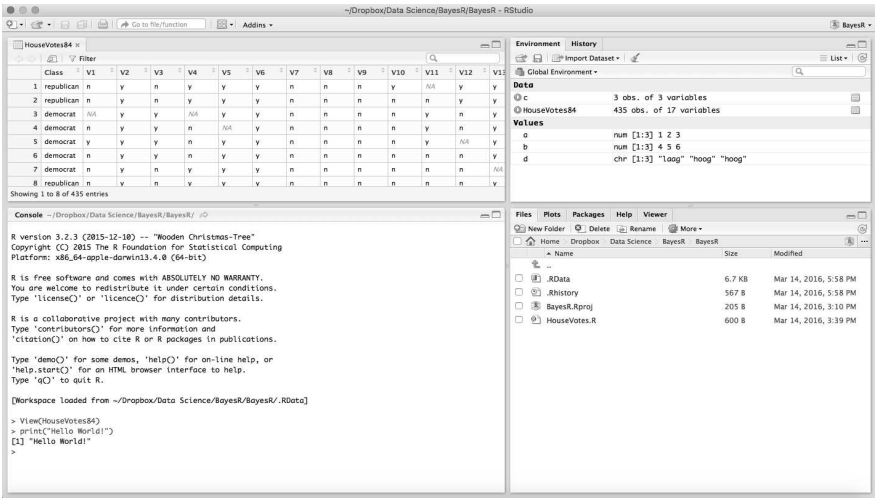
Figuur 3 De standaard R-console (niet RStudio)

De R-console reageert op ieder command dat we invoeren. Grofweg doet de R-console twee dingen: opslaan en reageren. Dit wordt een **Read Evaluate Print Loop** (REPL) genoemd. Deze R-console is the zien in **Figuur 3**. Als we een waarde aan een nieuwe variabele toewijzen, slaat R deze op en geeft hij pas deze waarde terug als we deze variabele weer oproepen. Als we echter een functie gebruiken of zelf een waarde invoeren, geeft R een reactie. Deze reactie is dezelfde waarde, of de waarde na de toepassing van een functie. In **Figuur 3** wordt de waarde `"Hello World!"` aan de

`print()` functie gegeven. R past de functie toe op deze waarde en geeft het resultaat terug.

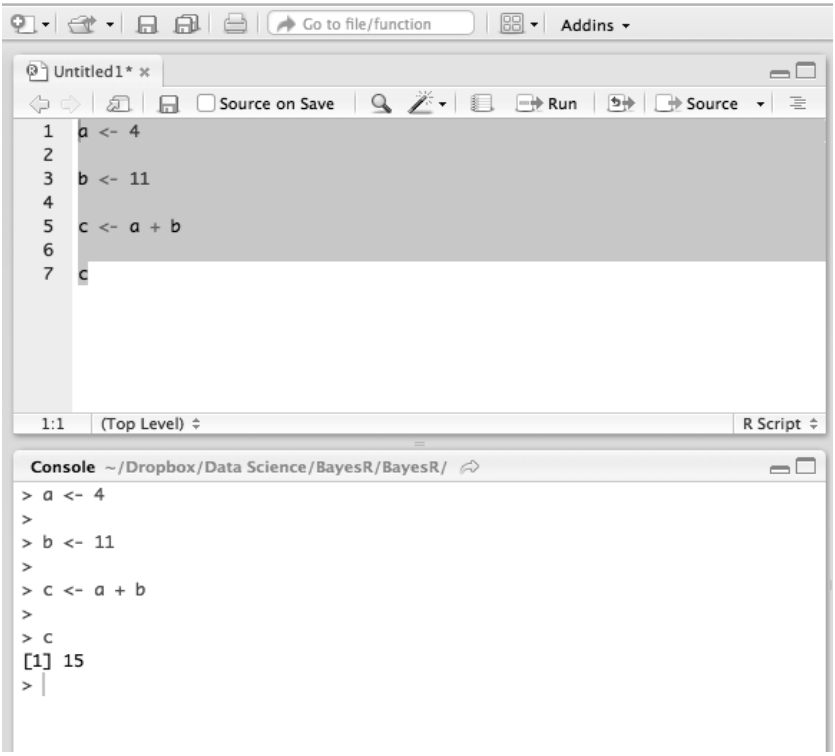
De R Studio omgeving

In de onderstaande afbeelding **Figuur 4** zie je een screenshot van de **RStudio IDE**. **IDE** staat voor **Intelligent Development Environment** en dat beschrijft ook precies de toegevoegde waarde van RStudio ten opzichte van de standaard R-console. De RStudio-omgeving is een stuk meer uitgebreid dan de simpele R-console alleen. Echter is de IDE zeker niet ingewikkeld. De extra vensters zijn namelijk visuele hulpmiddelen. Het venster links onderin bij **Figuur 4** zal je waarschijnlijk al herkennen. Het is namelijk exact dezelfde R-console die we in **Figuur 3** hebben gezien, alleen verwerkt in de IDE van RStudio. Deze werkt **exact** hetzelfde als de normale R-console. Een toevoeging in RStudio is dat hier de R-console **auto-completion** toepast op de commands. Als je bijvoorbeeld een functie invoert of een object oproept, geeft de R-console een hint voor het code dat je bedoelt.



Figuur 4 De R-console in RStudio

Links bovenin zie je een dataset. Omdat je logischerwijs met datasets werkt in R, wil je af en toe weten hoe de dataset eruitziet. Deze datasets worden dan op dit venster weergegeven. Daarbij zijn er handige mogelijkheden om data te sorteren en te selecteren. Dit venster is ook de plek waar jouw R-scripts verschijnen. Met het knopje links bovenin (het blaadje met het plusje) kun je een nieuw R-script toevoegen. **Ik adviseer je om een R-script te gebruiken.** Een klein voorbeeld van een script kun je vinden in **Figuur 5**. Zo heb je altijd een overzicht van de commands die in de R-sessie worden uitgevoerd en belangrijker nog, je kunt het opslaan als bestand en is daarmee een eindproduct van de R-sessie: een R-script door jou ontwikkeld.



Figuur 5 Een script ontwikkelen in RStudio

Script in R

In een script kun je de commands intypen en laten uitvoeren in de console. Dit doe je door de cursor aan het begin van de regel van het command te plaatsen en `Ctrl + Enter (Windows)` of `Cmd + Enter` (Mac) in te gebruiken. Je kunt ook de hele regel selecteren en `Ctrl + Enter` of `Cmd + Enter` gebruiken om een command uit te voeren.

Rechts bovenin heb je het overzicht van de objecten in de huidige R-sessie. Dit zijn bijvoorbeeld variabelen met de toegewezen waarden, functies en datasets. Daarbij heb je een tab met **History**, hiermee krijg je een handig overzicht van de commands die je in deze R-sessie hebt uitgevoerd.

Het venster rechts onderin bevat de meeste tabs van alle vensters. De **Files** tab is de verkenner waarmee je door de bestanden en mappen kunt bladeren. De **Plots** tab is de plek waar de grafieken en andere visualisaties worden weergegeven. Deze tab opent zich automatisch zodra je in R code invoert om een visualisatie weer te geven. De **packages** tab geeft een overzicht van de gedownloade en geopende packages in de huidige omgeving (jouw account op de laptop of computer). Packages zijn modules gemaakt door ontwikkelaars om extra functies toe te voegen in R, deze zullen we ook tegenkomen in dit boek.

Het **Viewer** venster is iets nieuwer dan het **Plots** venster en neemt de interactieve visualisaties zoals kaarten voor zijn rekening.

Deze vensters in RStudio geven samen een uitstekende ervaring voor het data analyseren met R. Ongetwijfeld zal je dit zelf ervaren. De zojuist behandelde onderdelen zijn slechts het topje van de ijsberg qua functies in RStudio. Voor het verdere verloop van dit boek zijn de belangrijkste onderwerpen in ieder geval behandeld.

Datasets

In sommige hoofdstukken in dit boek wordt er gebruik gemaakt van datasets. De meeste databestanden zijn in R al beschikbaar. Voor datasets die standaard al in R beschikbaar zijn, wordt in dit boek uitgelegd hoe je deze kunt gebruiken. Sommige databestanden worden in R geïmporteerd. Dit zijn de volgende databestanden welke eenvoudig op mijn (oude) blog kunt downloaden:

Databestanden:

- **Consumentenprijzen.csv**
- **Projecten.csv**
- **Projecten2.csv**

Voor een overzicht van alle datasets (ook van de oude tutorial) kun je naar de pagina van mijn oude R cursus gaan op http://r-handleiding.blogspot.nl/p/voorbeeld-databestanden-downloaden.html .

Referenties

Op sommige plekken in dit boek wordt er doorverwezen naar een bron. Dit zijn bronnen waar een bepaald onderwerp in dit boek extra uitgebreid wordt besproken. Het is nuttig om deze bronnen na te gaan als je, naast de rode lijn in dit boek, dieper wilt ingaan over een bepaald onderwerp. De nummers van de bronnen kun je makkelijk terugvinden in het hoofdstuk **Referenties** aan het einde van dit boek. Achter de nummers van de bronnen staat informatie om de bron te raadplegen.