# داده‌کاوی کاربردی با R

**مولفان:**

**محمد مرادی**

**مجید جوادی**

**سهیلا مهرمولایی**

نشردانشگاهی کیان
Kian Publication

| | |
|---|---|
| سرشناسه | مرادی، محمد، ۱۳۶۶. |
| عنوان و نام پدیدآور | داده‌کاوی کاربردی R / مولفان محمد مرادی، مجید جوادی، سهیلا مهرمولایی. |
| مشخصات نشر | تهران: انتشارات دانشگاهی کیان، ۱۳۹۶. |
| مشخصات ظاهری | ۳۶۷ ص.: مصور. |
| شابک | ۹۷۸-۶۰۰-۳۰۷-۱۶۸-۱ |
| وضعیت فهرست‌نویسی | فیپا. |
| موضوع | داده کاوی. Data mining. |
| موضوع | داده کاوی -- نرم‌افزار. Data mining -- Software. |
| شناسه افزوده | جوادی، مجید، ۱۳۶۶. |
| شناسه افزوده | مهرمولایی، سهیلا، ۱۳۶۲. |
| رده بندی کنگره | QA ۷۶/۹/د۲م۴ ۱۳۹۶. |
| رده بندی دیویی | ۰۰۶/۳۱۲ |
| شماره کتابشناسی ملی | ۴۶۴۰۲۳۷ |

انتشارات دانشگاهی کیان

نشردانشگاهی کیان
Kian Publication

نام کتاب : داده‌کاوی کاربردی با R
مولفان : محمد مرادی - مجید جوادی - سهیلا مهرمولایی
ناظر فنی : علی محمودی
ویراستار : لیلا رفیعی
صفحه آرا : مرضیه امانت
طراح جلد : شیلان هوشیاری
چاپ اول : ۱۳۹۶
تیراژ : ۱۰۰۰
چاپ : ستاره سبز
صحافی : نمونه

شابک : ۹۷۸-۶۰۰-۳۰۷-۱۶۸-۱
ISBN : 978-600-307-168-1


خرید اینترنتی آسان از:
www.kianpub.com



kianpublication

برای دریافت اخبار و اطلاعات مفید و شرکت در قرعه‌کشی، ما را در این شبکه‌ها دنبال کنید.

# سخن ناشر

**«سپس، به کاتبان و نویسندگانت بنگر و بهترین آنها را بر کارهای خود بگمار...**
**کاتبان و نویسندگانی برگزین که قدر خود را بشناسند، چون کسی که به قدر خود شناخت ندارد، دیگران را هم نمی‌شناسد.»** **«گرفته‌شده از نامه‌ی ۵۳ نهج‌البلاغه به مالک‌اشتر»**

اگرچه نوشتن و پرداختن زکات علم از توصیه‌های اکید بزرگان و گواه بر کرامت اهل دانش است؛ اما امروزه پرداختن به انگیزه‌ها و اهداف نوشتن بیش‌تر جلوه می‌کند. بی‌شک این‌که چه کسی می‌نویسد مهم نیست؛ اما این‌که چرا و به چه پشتوانه‌ای می‌نویسد، درخور تأمل است.

ما معتقدیم که چاپ روزافزون کتاب‌های به اصطلاح «زرد» که خالی از هرگونه نوآوری و بی‌توجه به استانداردهای چاپ کتاب و نیازهای مخاطبان است، حاصل تفکر بازاری مستولی بر جامعه‌ی نشر است. بی‌پرده آن‌که عنوان پر زرق و برق، دستاویز قراردادن مضمون‌های نو با هدف فروش بالا و طویل‌کردن سیاهه‌ی سابقه‌ی علمی، نمی‌تواند دلیل محکمی برای چاپ و نشر کتابی باشد که خواننده‌ی مشتاق با صرف هزینه‌های نه چندان کم آن را تهیه می‌کند؛ به امید آن که چیزی از آن بیاموزد.

باید پذیرفت که انگیزه‌ی نوشتن کم از محتوای نوشته نیست و بین این دو رابطه‌ای مستقیم برقرار است. اگر انگیزه‌ی نوشتن، تولید دانش باشد، بی‌شک نویسنده از قلم بی‌محتوا وکم‌عمق پرهیز می‌کند و اگر دغدغه‌ی دانش و فرهنگ زخم‌خورده در میان باشد، ناشر تنها به عنوان پرطمطراق بسنده نمی‌کند.

و چقدر امروزه، فرهنگ و دانش این مرز و بوم که گرفتار آفت بی‌انگیزگی و زخم هوس است، نیازمند ناشران و نویسندگانی است که نیت‌شان کمک به رشد دانش و ارتقای فرهنگ جامعه است و به راستی که التیامی بر این درد نیست؛ مگر نویسندگانی که قدر خود و دیگران را می‌دانند و خوب می‌فهمند که کتاب، ابزار سودجویی‌های مغرضانه نیست و می‌کوشند تا خود را از هرگونه عطش نام و رسم و ثروت تهی کنند.

ما در انتشارات دانشگاهی کیان خود را بری از عیب و خطا نمی‌دانیم؛ اما همواره بیش از پیش می‌کوشیم تا در راستای تولید علم و نشر کتاب‌های پرمحتوا، دست نویسندگانی که انگیزه‌ی پاک دارند را بفشاریم و در کنارشان باشیم و از خداوند متعال می‌خواهیم که در این مسیر صعب و پرخطر در سایه‌ی لطف و عنایت خود از آن‌چه به عهده‌ی ما نهاده شده، سربلند و پیروز برآییم.

**انتشارات دانشگاهی کیان**

# مقدمه مولفان

در حالی که بیش از دو دهه از مطرح شدن اصول و مفاهیم داده‌کاوی می‌گذرد، به جرات می‌توان گفت که اکنون در دوران شکوفایی و اوج کاربردهای آن هستیم. در واقع، در سال‌های اخیر و در روبه‌رویی با حجم بسیار زیادی از داده‌ها در حوزه‌های مختلف کاری، استفاده از داده‌کاوی به‌عنوان یکی از الزامات فرایندهای علمی، پژوهشی و تجاری مطرح شده است. بر این اساس، به ندرت می‌توان فرایندها و پروژه‌های مختلف پژوهشی و تجاری را یافت که از مزیت‌های پرشمار داده‌کاوی بهره‌ای نجسته باشند.

در پی این اقبال عمومی به استفاده از روش‌های داده‌کاوی، یکی از مسایل و چالش‌های مهم، البته از دیدگاه فنی و کارشناسان علوم رایانه، چگونگی پیاده‌سازی الگوریتم‌ها و روش‌های داده‌کاوی و اعمال آن‌ها بر مجموعه‌های داده‌ای بوده و می‌باشد.

اگرچه بسیاری از زبان‌های برنامه‌نویسی و سیستم‌های بانک‌های اطلاعاتی، امکانات و تسهیلاتی را برای انجام فرایندهای داده‌کاوی فراهم می‌نمایند، اما به‌کارگیری ابزارهای اختصاصی که به‌صورت قابل قبولی از گام‌های مختلف فرایند داده‌کاوی پشتیبانی کند، یکی از خواسته‌های منطقی جامعه کاربری می‌باشد.

بر همین اساس، ابزارها و زبان‌های برنامه‌نویسی متعددی معرفی شده‌اند، ولی آنچه در طی سال‌های اخیر توجهات را به صورت گسترده‌ای به خود جلب کرده است، زبان R می‌باشد. آنچه این زبان برنامه‌نویسی را از سایر رقبایش متمایز می‌سازد، قابلیت‌های تعبیه شده در آن به عنوان زبانی برای انجام محاسبات و تحلیل‌های آماری است. از این‌رو، می‌توان این زبان برنامه‌نویسی را یک ابزار ایده‌آل برای انجام فرایندهای تحلیل داده و به‌صورت مشخص داده‌کاوی دانست. شاهد این ادعا نیز آمارهای منتشر شده از سوی منابع معتبر است که میزان اقبال بالای متخصصان به این زبان را تایید می‌کند.

با توجه به ویژگی‌ها و اهمیت این زبان، در این کتاب بر آن شدیم که مروری کلی بر چگونگی انجام فرایندهای تعامل با داده با استفاده از R داشته باشیم و دریچه‌ای نو بر روی کارشناسان و پژوهشگران عرصه تحلیل داده بگشاییم. در واقع هدف اصلی کتاب حاضر این است که قابلیت‌هایی بیشتر و انعطاف‌پذیرتر از ابزارها و بسته‌های نرم‌افزاری مربوط به داده‌کاوی که استفاده از آن‌ها در میان دانشجویان و پژوهشگران رایج است

را به مخاطبان ارایه کند و این مهم از طریق یادگیری زبان R و افزونه‌های مربوط به آن محقق می‌شود.

البته لازم به ذکر است که بررسی تمامی جوانب و نکات مربوط به استفاده از این زبان برای داده‌کاوی، احتیاج به چندین جلد کتاب دارد؛ از این رو در این کتاب تلاش بر این بوده است که بتوانیم ضمن پرداختن به مباحث متنوع و در عین حال مهم، امکان فراگیری ساده و به‌دور از پیچیدگی را نیز فراهم آوریم.

اگرچه پیش‌نیاز بهره‌برداری کامل از این کتاب، آشنایی اولیه با مفاهیم زبان R و نیز اصول داده‌کاوی می‌باشد، اما به‌منظور تسهیل فرایند آموزش، مقدمه‌ای بر مفاهیم ذکر شده در دو فصل ابتدایی ارایه شده است.

در انتها ضمن تشکر از مدیریت و دست‌اندرکاران نشر دانشگاهی کیان که در طی مراحل تکمیل این کتاب نهایت همکاری را با مولفان داشتند، از خوانندگان گرامی درخواست می‌کنیم که نظرات، پیشنهادات و انتقادات خود درباره کتاب حاضر را از طریق پست الکترونیکی RDataMining.96@gmail.com با ما در میان بگذارند.

## فهرست مطالب

## فهرست مطالب

## فهرست مطالب

## فهرست مطالب

# فصل ۱

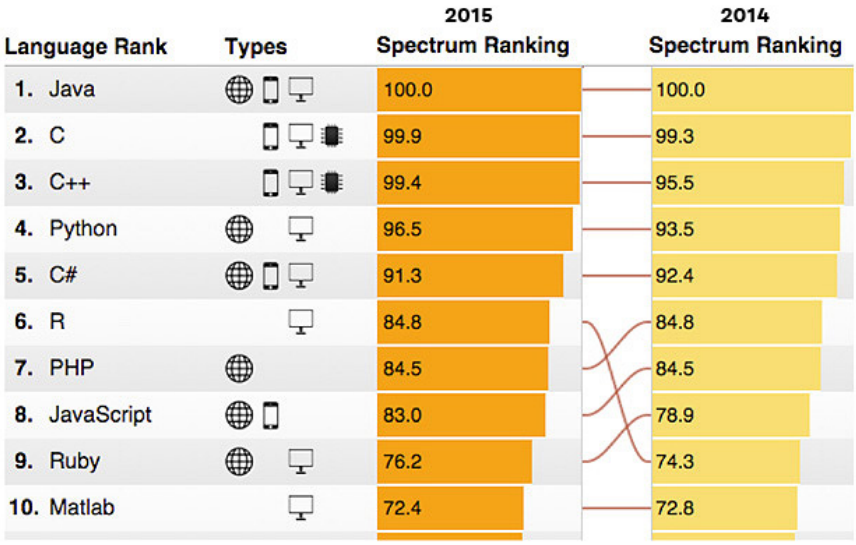
# مقدمه‌ای بر زبان R

دنیای فناوری اطلاعات سال‌هاست که شاهد معرفی زبان‌های برنامه‌نویسی به منظور انجام فعالیت‌ها و اهداف مختلف می‌باشد. شاید بر شمردن تمامی این زبان‌ها کاری بسیار دشوار به نظر برسد، با این حال از آنجایی که هر یک از آنها برای منظور مشخصی ایجاد شده‌اند، انتخاب گزینه‌ای مناسب برای هدفی معین از دشواری کمتری برخوردار است.

در کنار بسیاری از زبان‌های برنامه‌نویسی چند و همه‌منظوره؛ یکی از جریان‌های اصلی در حوزه توسعه زبان‌های برنامه‌نویسی، فعالیت برای تولید زبان‌های خاص منظوره بوده است. اگرچه، چنین زبان‌هایی نیز بسیار متنوع هستند و در انجام یک کار مشخص می‌توان در چنین زبان‌هایی به گزینه‌های متعددی برخورد و از ویژگی‌های آنها استفاده کرد.

یکی از زبان‌های خاص منظوره در حوزه تحلیل آماری داده‌ها که در طی چند سال گذشته بسیار مورد توجه قرار گرفته است، زبان R می‌باشد. شاهدی بر ادعای اهمیت این زبان نیز آمار رسمی ارایه شده به وسیله IEEE در رابطه با وضعیت زبان‌های برنامه‌نویسی محبوب و پراستفاده

در سال ۲۰۱۵م است. جایگاه R در این آمار نمایانگر کارایی و اهمیت آن است که سبب شده است تعداد درخور توجهی از توسعه‌دهندگان در سراسر دنیا به سمت آن متمایل شوند (تصویر ۱-۱).




**تصویر ۱-۱. زبان‌های برنامه‌نویسی محبوب در سال ۲۰۱۵م**[1]

علاوه بر ایده اصلی و اولیه تعبیه شده در R به منظور انجام فرایندهای تحلیلی و بازنمایی بصری داده‌ها؛ قابلیت‌های این زبان در انجام فرایند داده‌کاوی سبب شده است که به صورت گسترده به عنوان یکی از زبان‌ها (و ابزارهای) داده‌کاوی مورد استفاده قرار بگیرد. در واقع این مساله به این صورت توجیه‌پذیر است که داده‌کاوی به عنوان علم بررسی حجم زیادی از داده‌ها به منظور استخراج الگوهای پنهان آنها نیز به‌نوبه‌ی‌خود یک فرایند تحلیلی بر روی داده‌ها شناخته می‌شود. از این‌رو، بسیار منطقی است که از R به منظور کاوش داده‌ها به روش‌های مختلف استفاده شود.

علاوه بر این، زبان R قابلیت‌های جالب توجهی در زمینه بازنمایی بصری داده‌ها و ایجاد گرافیک‌های مناسب به جهت پشتیبانی از فرایندهای تحلیلی ارایه می‌کند.

بر این اساس و با توجه به اهمیت و کاربردهای زیاد زبان R، در این کتاب به صورت مشخص به چگونگی انجام فرایند داده‌کاوی با استفاده از R می‌پردازیم. در این راه، علاوه بر معرفی و استفاده از یک کتابخانه اختصاصی که برای داده‌کاوی در R معرفی شده است (Rattle)؛ فرایند

1. http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages

کاوش داده‌ها را تنها با استفاده از قابلیت‌های اولیه R و سایر ابزارهای موجود به انجام می‌رسانیم. همچنین، به موضوع گرافیک و بازنمایی بصری داده‌ها (با استفاده از واسط گرافیکی Rattle) در کنار برخی از مباحث پیشرفته داده‌کاوی، نظیر کاوش شبکه‌های اجتماعی، نیز پرداخته می‌شود. در پایان به منظور جمع‌بندی آموخته‌ها، سه پروژه داده‌کاوی به صورت کامل بررسی و عملیاتی می‌شوند.

شایان ذکر است، برای استفاده حداکثری از محتوای کتاب حاضر، مخاطبان گرامی باید با مباحث اولیه زبان R و نحوه استفاده از آن و نیز مفاهیم داده‌کاوی آشنایی داشته باشند. اگرچه، آگاهی از این موارد الزامی نخواهد بود، ولی عدم داشتن اطلاعات کافی در این زمینه ممکن است که مخاطبان را در فرایند یادگیری مباحث مطرح شده با مشکلاتی مواجه نماید.

با این حال، اگرچه پرداختن به جزییات مباحث گفته شده خارج از محدوده این کتاب است، ولی به منظور تسهیل فرایند یادگیری، دو فصل ابتدایی به ترتیب مقدماتی را درباره‌ی زبان R و داده‌کاوی فراهم می‌کنند. از این‌رو، خوانندگانی که با این مباحث آشنایی کافی دارند می‌توانند کتاب را از فصل سوم شروع نمایند.

در انتهای این مقدمه، یک بار دیگر تاکید می‌شود که هدف اصلی کتاب حاضر معرفی و بررسی فرایند داده‌کاوی با استفاده از زبان R می‌باشد. در این مسیر تلاش می‌شود تا جنبه‌های مختلف و مرتبط از زبان R، مانند مباحث پیشرفته داده‌کاوی نظیر کاوش شبکه‌های اجتماعی و بازنمایی بصری داده‌ها مورد بررسی قرار بگیرند.

## ۱-۱. معرفی R

R یک زبان و محیط برنامه‌نویسی برای انجام محاسبات و تحلیل‌های آماری است. از دیدگاه سراسری بودن، نسخه‌هایی از این زبان برای سیستم‌های عامل شناخته شده نظیر لینوکس، ویندوز و Mac OS و همچنین معماری‌های مختلف مانند اینتل[1] و اسپارک[2] ارایه شده است. از جنبه تاریخی، R در ابتدا توسط پژوهشگرانی از دانشگاه اوکلند نیوزلند در سال ۱۹۹۶م ابداع شد و در حال حاضر توسط گروهی از پژوهشگران و توسعه‌دهندگان از موسسات و دانشگاه‌های مختلف مدیریت شده و توسعه می‌یابد.

این زبان از دیدگاه تجاری و نحوه توسعه مبتنی بر فلسفه و مفاهیم نرم‌افزارهای کد منبع باز[3] ایجاد شده و بر همان اساس نیز توسعه یافته و می‌یابد. به‌صورت مشخص این مساله به این

1. Intel
2. Spark
3. Open Source

معناست که کد منبع هر یک از مولفه‌های[1] R به صورت رایگان در اختیار عموم قرار دارد و این امکان را برای کاربران فراهم می‌کند تا بتوانند میزان کیفیت و قابلیت اعتماد آنها را برای فعالیت‌ها و پروژه‌های موردنظرشان بررسی و آزمایش نمایند.

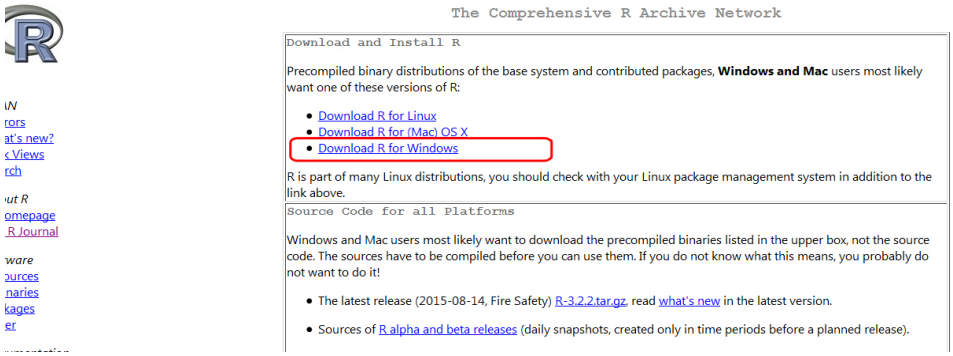
به‌صورت کلی، منتقدان رویکرد توسعه نرم‌افزار مبتنی بر کد منبع باز به یک دلیل مهم به این شیوه ایراداتی را وارد می‌کنند و آن عدم وجود جامعه کاربری فعال و پاسخ‌گو در زمینه راهنمایی کاربران برای استفاده از آن نرم‌افزار و یا زبان برنامه‌نویسی و یا سیستم می‌باشد. اگرچه این موضوع و انتقاد در پاره‌ای از موارد صحیح است، اما در رابطه با R اینچنین نیست؛ چراکه تعداد درخور توجهی از مستندات و فیلم‌های آموزشی، انجمن‌ها و کتاب‌های آموزشی به این زبان اختصاص پیدا کرده‌اند که تا حدود بسیار زیادی می‌توانند مشکلاتی را که کاربران با آن روبه‌رو می‌شوند، مرتفع نمایند.

البته در کنار تمام مزایایی که R به توسعه‌دهندگان ارایه می‌کند، مشکلاتی نظیر عدم توانایی مواجهه با مجموعه‌های داده‌ای بسیار بزرگ[2] را نیز می‌توان به عنوان یکی از معایب این زبان برشمرد. دلیل فنی این مساله آن است که تمامی محاسبات در R در حافظه اصلی[3] سیستم رایانه‌ای صورت می‌پذیرد و از آنجایی که این حافظه باید در اختیار دیگر برنامه‌های سیستم‌عامل نیز قرار بگیرد، امکان انجام حجم بالایی از محاسبات را از R سلب می‌کند. با این حال، این موضوع به این معنا نیست که هیچ‌گونه راه‌حلی برای رفع آن وجود ندارد. در مسیر حل این مشکل، واسط[4]های انعطاف‌پذیر ارتباط R با بانک‌های اطلاعاتی، امکان انجام محاسبات سنگین، نظیر فرایندهای تحلیلی داده‌ها و به‌صورت مشخص داده‌کاوی را فراهم می‌سازد.

## ۱-۲. نصب و راه‌اندازی

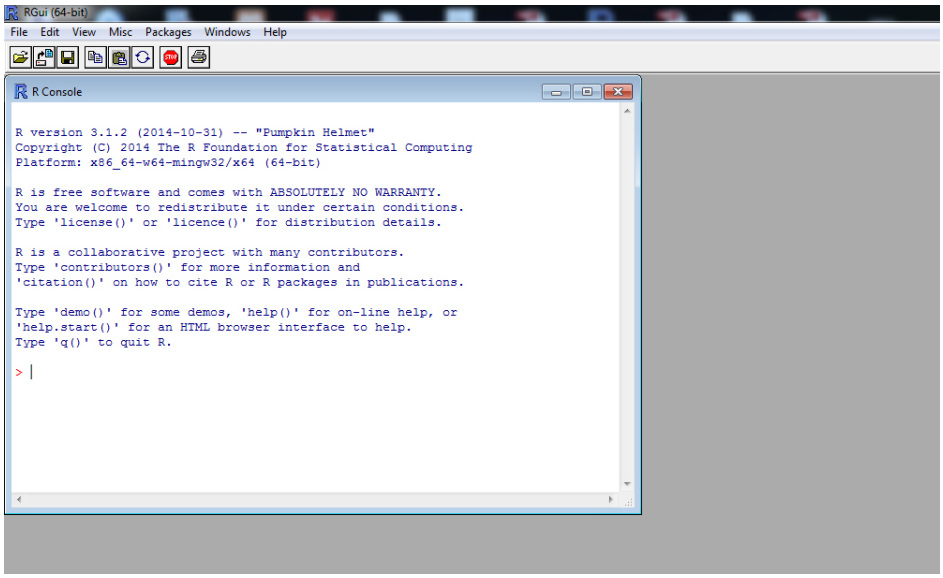
در ابتدا باید R را بر روی سیستم خود نصب و راه‌اندازی نمایید. ساده‌ترین راه برای این کار، دریافت R از وب‌سایت رسمی آن (cran.r-project.org) می‌باشد. سپس، دستورالعمل‌های ارایه شده به منظور نصب آن را به صورت مرحله‌به‌مرحله دنبال کنید (تصویر ۱-۲).

1. Component
2. Large Datasets
3. RAM
4. Interface




**تصویر ۱-۲. دریافت فایل‌های مربوط به نصب R**

در واقع نصب R همانند نصب دیگر برنامه‌های نرم‌افزاری است و مساله خاصی در رابطه با آن وجود ندارد. پس از نصب، با کلیک بر روی آیکون ایجاد شده به وسیله برنامه، کنسول R به نمایش درمی‌آید (تصویر ۱-۳).




**تصویر ۱-۳. نمایی از کنسول R**