Post01: Applications and Benefits of R in Industry

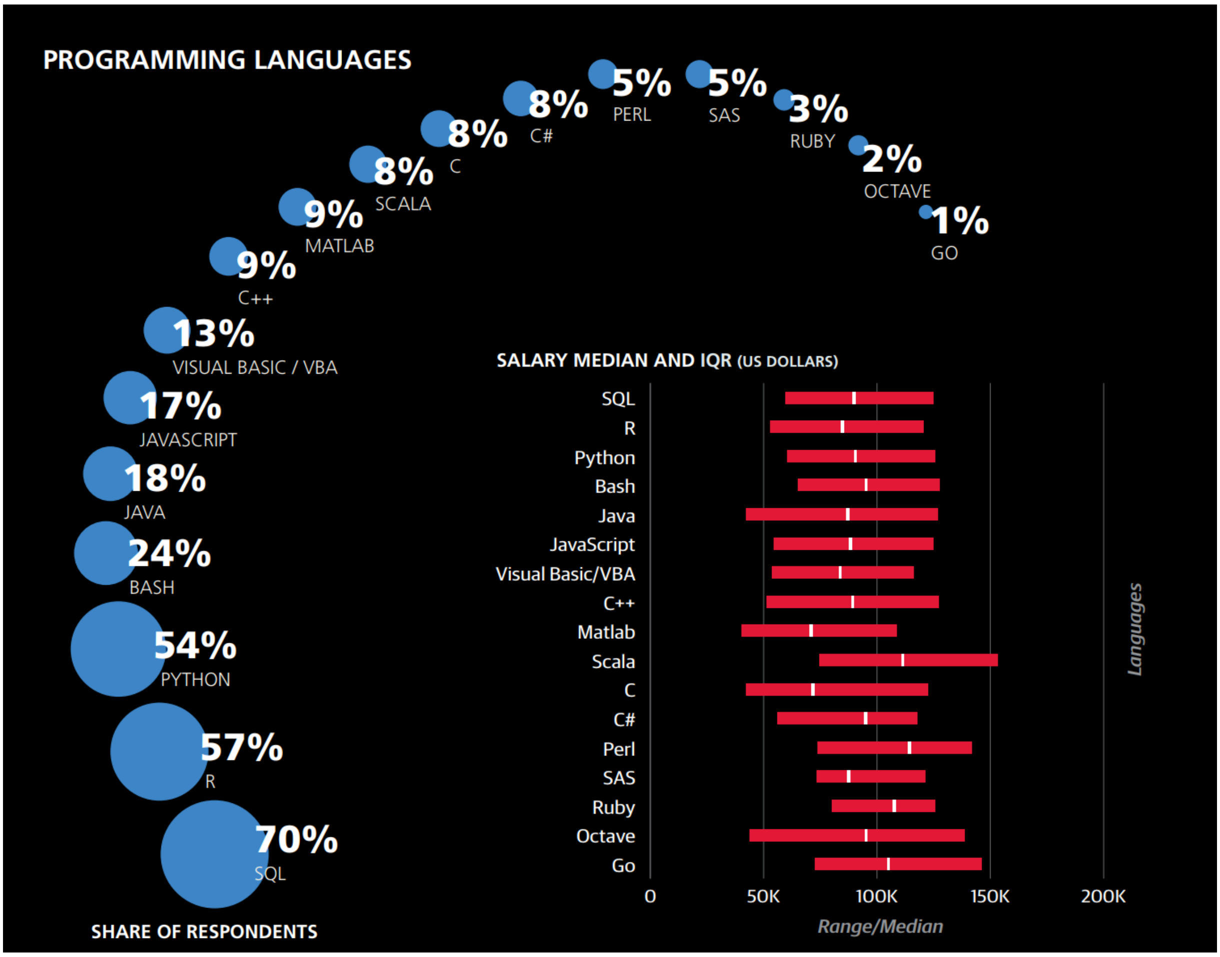
Yash Sanghrajka

October 30, 2017

Introduction

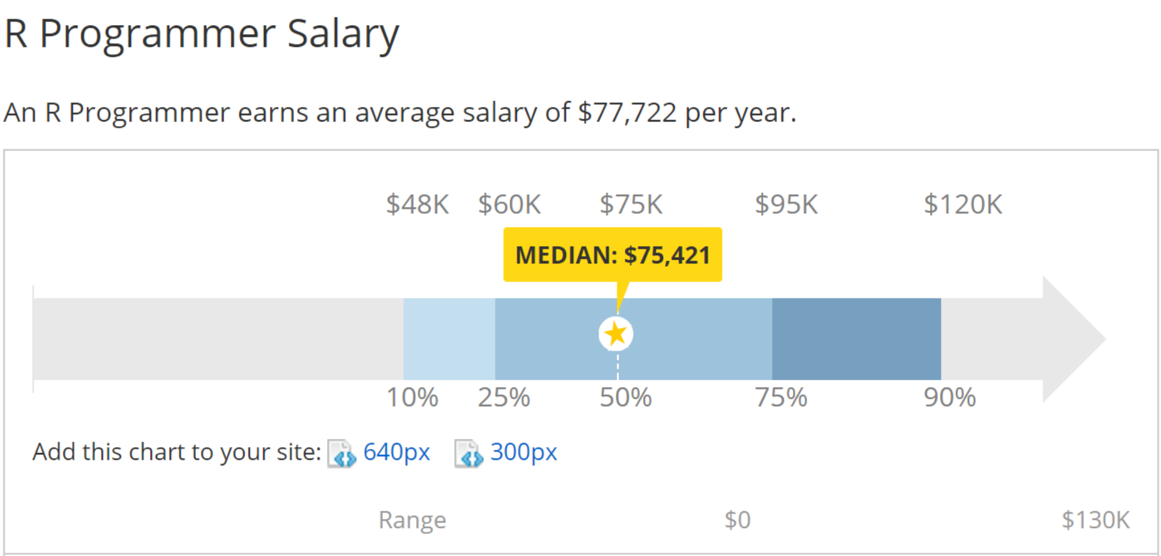
R, as a language, has been around for a long time. An implementation of the S language, R was created at the University of Auckland in New Zealand by Robert Gentleman and Ross Ihaka. The project was started in 1992, and was released in beta version in 2000.

However, recently, the appeal to use R has grown immensely. The 2016 Data Science Salary Survey states that for job postings in the field of Data Science, over 57% of postings list R as a requirement, making it the second most popular programming language behind SQL.



This visual shows the most popular programming languages in data science, along with their average salaries.

According to PayScale, the average salary for an R programmer is approximately $77000, much higher than the national average salary.



R Programmer Salary

An R Programmer earns an average salary of $77,722 per year.

Add this chart to your site: 640px 300px

Range $0 $130K

So, why is there such a demand for R programmers?

To answer that question, I will be discussing how companies are using R, how prominent R is in industry, and lastly, the benefits of R, along with its shortcomings and how they have been improved in recent years.

Audience

The target audience for this is not only Statistics 133 students, as they can learn about the applications of R in industry, along with what specific concepts are crucial for the current industrial age, but also people who are looking to learn more about data analytics and potentially become data analysts and programmers. This post will tell them why R is a crucial language to learn, and how R has been applied to numerous industries. An example how this can be potentially used for this class is the introduction. Using this post can show the benefits of learning R, and show it's applications to not only statistics majors, but people studying a wide variety of subjects.

Industry uses of R

The first question we want to answer is what industries R is becoming prominent in. According to David Smith, "We saw, way back in 2007, just how R had taken over academia." However, as time progressed, companies started using R on a wide scale for their own data analysis.

The first companies that realized the power of R were social media companies.

According to DataInformed , "Facebook, which processes more than 500 terabytes of data a day, uses R to understand how its users interact with the service. Exploratory data analysis helps Facebook understand what its users are doing throughout the day and how viral memes propagate through the social network. Data visualization is a big part of this work, and Facebook has shared its best practices in an online Udacity course, and even used a chart created with R in its IPO prospectus."

An example of R that was used in a collaborative project between Facebook and Stanford Researchers is:

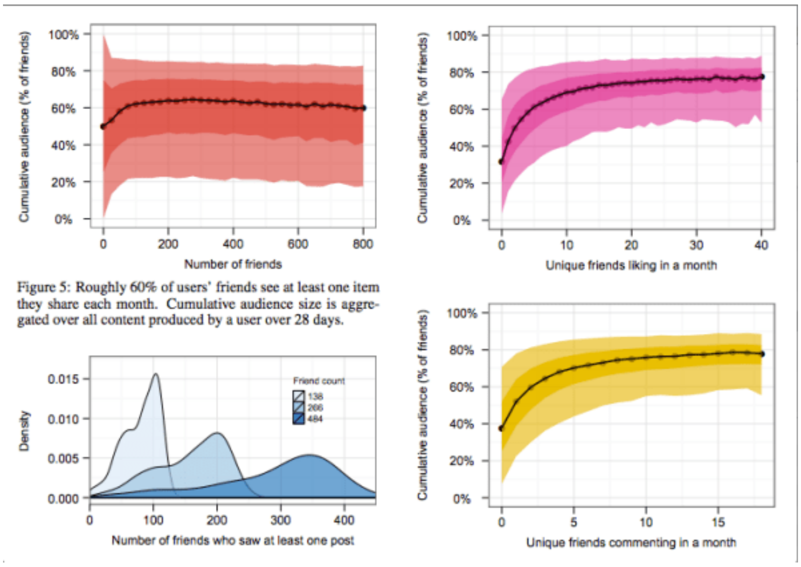


Figure 5: Roughly 60% of users' friends see at least one item they share each month. Cumulative audience size is aggregated over all content produced by a user over 28 days.

As we can see, social media organizations rely heavily on data to boost their performance, making R a clear favorite in the industry. By being able to analyze social trends, Facebook and other social media sites can use R to improve their product and grow their membership.

Another field that uses R very heavily is the Financial Services field.

According to Dan Butcher, "Shops that do high-frequency/low-latency trading often use the statistical language R, and they need people who can implement a systematic way of developing programs to analyze investment data and risk management. Important skills to learn include time-series, forecasting, portfolio selection, covariance clustering, prediction and derivative securities. Using R, we can simulate extreme events and their effect on prices, conduct portfolio analysis and visualize the covariances in a portfolio that we built," This application to the finance industry is disrupting the industry as a whole. Quantitative trading firms such as Citadel, Renaissance Technologies, and Two Sigma are using quantitative analysis through programs such as R to disrupt the industry and post returns that are much higher than traditional bulge bracket banks.

It is even used in sports analytics.

R is often used from everything from simple data cleaning and visualization to complex model building to analyze and predict player and team success. Even in our class alone, we have used NBA statistics to compute the efficiency of players, and have done an analysis of these players based on their game-time stats.

The applications of R are evident when looking at NBA statisticians. When looking at all the statisticians and analysts employed by NBA teams, we can see that R is a common skill.

The subsequent analysis was made off of this website, which provides a list of every NBA statistician and provides their LinkedIn details. I went through each LinkedIn/Personal Website, and saw whether they listed that they were skilled in R.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
## 
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```

```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
#Importing the Dplyr and ggplot2 libraries.
```

```
nba_analysts = read.csv("nba_analysts_R.csv",stringsAsFactors = FALSE)
#Importing CSV of Data filled with the Names, Teams, the Education Level, and whether they knew R.
head(nba_analysts)
```

```
##             Name               Team Education.Level R.Experience
## 1  Dan Rosenbaum      Atlanta Hawks             PhD           No
## 2   David Sparks     Boston Celtics             PhD          Yes
## 3   Glenn DuPaul      Brooklyn Nets       Bachelors          Yes
## 4    Rami Antoun      Brooklyn Nets         Masters          Yes
## 5 Logan MacPhail      Brooklyn Nets         Masters           No
## 6  Scott Bullock  Charlotte Hornets         Masters           No
```

```
tables_r_experience = table(nba_analysts$R.Experience)
#This shows how many NBA analysts described themselves are R experts.
tables_r_experience
```

```
## 
##   No  Yes 
##   38   21
```

```
as.integer((tables_r_experience)[2])/
  (as.integer(tables_r_experience)[2]+as.integer(tables_r_experience)[1])
```

```
## [1] 0.3559322
```

```
#This calculates the proportion of analysts who said that they were R experts.
```

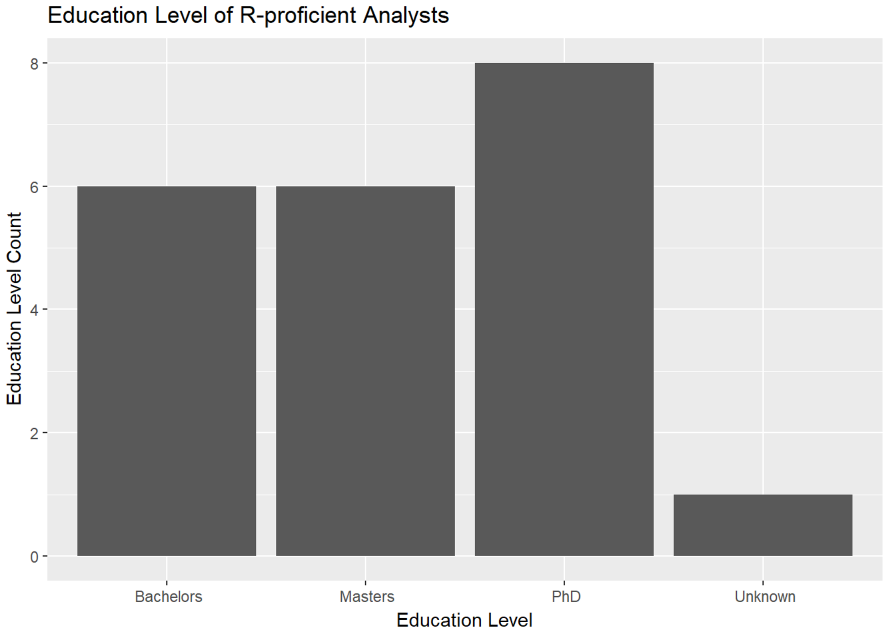
As we can see, 21 out of the 59 analysts consider themselves experts. In an industry that is growing, and becoming more and more important, the fact that over a third list R as a crucial skill shows how R has affected the industry.

When we narrow down the analysts who just list R, we can analyze their attributes, specificially, their education levels.

```
nba_analysts_with_r = nba_analysts[nba_analysts$R.Experience == " Yes",]
#This piece of code filters only analysts who have said that they are experienced in R.
head(nba_analysts_with_r)
```

```
##              Name                Team Education.Level R.Experience
## 2     David Sparks       Boston Celtics             PhD          Yes
## 3     Glenn DuPaul        Brooklyn Nets       Bachelors          Yes
## 4      Rami Antoun        Brooklyn Nets         Masters          Yes
## 7     David Kaplan    Charlotte Hornets       Bachelors          Yes
## 10     Jon Nichols  Cleveland Cavaliers         Masters          Yes
## 15 Tommy Balcetis       Denver Nuggets       Bachelors          Yes
```

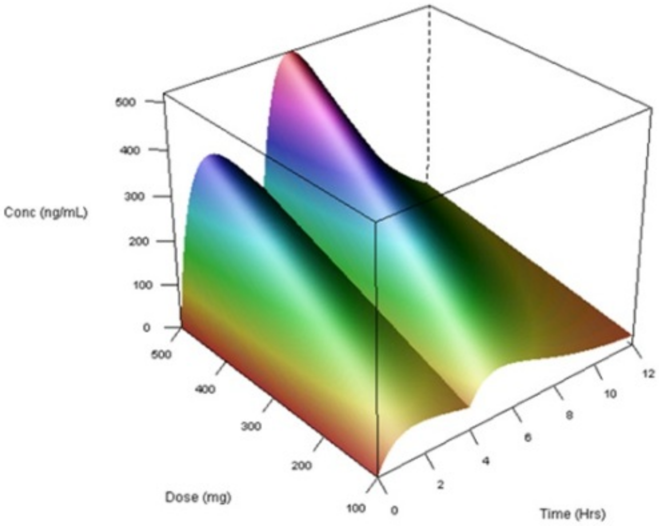
```
edu_level_plot = ggplot(nba_analysts_with_r,aes(x = Education.Level)) + geom_bar()
#This chunk of code uses ggplot to measure how many R programming analysts have a specific kind of education level
.
edu_level_plot + ggtitle("Education Level of R-proficient Analysts") + labs(x="Education Level", y= "Education Lev
el Count")
```



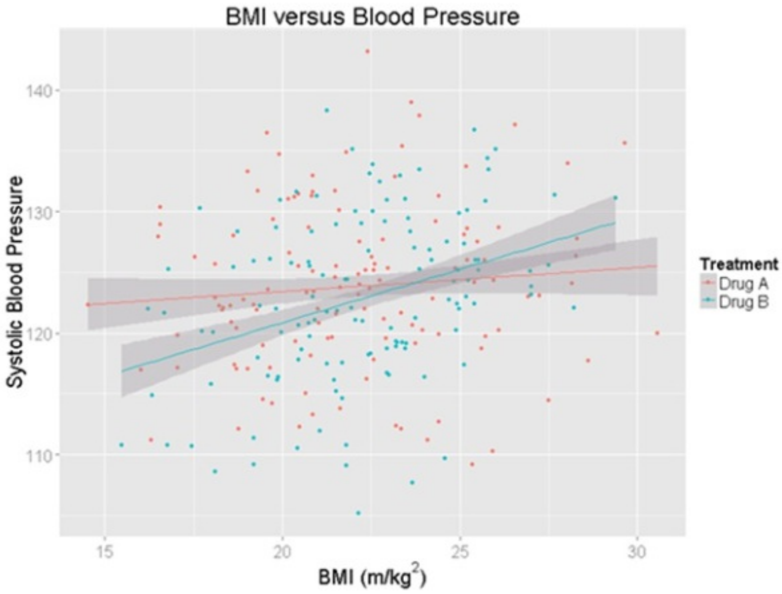
As we can see, there are very similar numbers of NBA analysts who have Bachelors, Master, and PhD's in the field. However, the reason I present this is to show how people from various skill levels and knowledge in R have gone on to become data analysts in the NBA.

Another field that it is used for is the pharmaceutical industry. The easy mapping techniques, along with the variety of different libraries, make it useful for pharmacists to use R as a way to measure drug use, effectiveness, and drug development.

These are two examples of graphics created by pharmacists in the field.



This graph uses the Lattice library to show the concentration of a dose over time.



This graph uses the ggplot library to compare the blood pressure of people on two different medicines based on BMI.

Section Wrap-Up

As we can see, there are numerous industries that use R. R's ubiquity makes it a really powerful language to learn, and can be applicable, no matter what you choose to study.

Benefits and Shortcomings

Now that we have covered some of the industries that using R, we can discuss the benefits to using R that clearly show when discussing why companies are shifting to R over other languages, along with the shortcomings that people have expressed about R, and how these are being alleviated over time.

Benefits

1. Free

According to InfoWorld, "At the time when it first came out, the biggest advantage was that it was free software. The source code and everything about it was available to look at."

This definitely provides a clear benefit to the user, as anybody from established corporations to amateur programmers can benefit from.

2. Use of Libraries

Quoting an industry expert, InfoWorld, states, "All R's graphics and charting capabilities, Adams says, are"unmatched." The dplyr and ggplot2 packages for data manipulation and plotting, respectively, "have literally improved my quality of life," he says."

The ggplot library is especially popular for its ease and style, and is among the most used libraries in R. Libraries in general can perform a wide variety of functions, throughout all forms of the data cycle, including data cleaning, analysis and presentation.

3. Machine Learning Capabilities

R has recently been used heavily for its machine learning capabilities. According to InfoWorld, ""Any new research in the field probably has an accompanying R package to go with it from the get-go. So in this respect, R stays at the cutting edge," he says. "The caret package also offers a pretty nifty way of doing machine learning in R through a relatively unified API." Peng also notes that a lot of popular machine learning algorithms are implemented in R." As companies are relying on automation further, the ability to perform advanced tasks in machine learning enhances R's appeal, and allows programmers to be able to automate and enhance the level of data analysis they provide.

4. Open-Source Community

R is an open-source software. This provides a much better community than closed-source software, as it provides more rigorous testing and support for programmers and developers. According to Graham Williams, R has been "reviewed by many internationally renowned statisticians and computational scientists", providing a better run community for people working in R.

5. Compatibility with other Softwares

According to Graham Williams, "R plays well with many other tools, importing data, for example, from CSV files, SAS, and SPSS, or directly from Microsoft Excel, Microsoft Access, Oracle, MySQL, and SQLite. It can also produce graphics output in PDF, JPG, PNG, and SVG formats, and table output for LATEX and HTML." It's compatibility with other software make it extremely applicable to industry. Being able to use much different software aims to provide ease to users, and also improves the functionality of those using R for their data analytics.

Now that we have covered the benefits of R, we can go into the shortcomings, and how they are being alleviated.

Shortcomings and Solutions

1. Security

The Security measures that are in other, comparable, languages are not as strong in R. However, recently, there have been steps taken to decrease security threats. According to InfoWorld,"The security issue, however, has been lessened by developments such as the use of virtual containers on the Amazon Web Services cloud platform, Peng says."

2. Memory issues

Memory issues pose a large problem, as companies are required to work with more and more data when conducting their data analysis. According to InfoWorld, “Peng says.”In that sense, it’s kind of an old technology in the way it was originally designed." The design of the language can sometimes pose problems in working with very large data sets, he says. Data has to be stored in physical memory. But as computers have gotten more memory, this has become less of an issue, Peng notes."

However, as stated before, there has been an improvement in the industry as computers have gotten more memory and certain software have improved memory space.

Examples of this are evident in industry, as R is now used to sift through terabytes of data.

3. Steep Learning Curve

One aspect that isn’t noted very often is the steep learning curve to learning R.

According to Graham Williams, “R is not so easy to use for the novice. There are several simple-to use graphical user interfaces (GUIs) for R that encompass point and-click interactions, but they generally do not have the polish of the commercial offerings.”

However, there are numerous work arounds to this that are allowing people to learn R effectively. In addition to classes such as Stat 133, there are numerous MOOC’s and videos online that are providing detailed learning plans for the R language.

Section Wrap-Up

This section delved into the benefits and shortcomings of the R language. The benefits include its price, libraries, machine learning capabilities, open-sourced community, and compatibility with other software and languages. Its shortcomings include its security measures, its memory issues and its steep learning curve, yet those three are being tackled and have gotten a lot better.

Conclusion

As I mentioned, there are numerous reasons why someone should learn R, and based on what you want to focus on, you can learn a wide variety of different techniques to tailor it to whatever industry interests you. Overall, some of the key benefits of R include the abundance of libraries, the ease of accessibility, and its powerful data analysis and visualization tools. It is evident that R is a language that will help shape the future of data analysis, and is a language that is worth learning.

References

http://www.oreilly.com/data/free/2016-data-science-salary-survey.csp

https://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html

http://data-informed.com/companies-use-r-compete-data-driven-world/

https://www.fastcompany.com/3030063/why-the-r-programming-language-is-good-for-business

https://news.efinancialcareers.com/us-en/244980/get-a-financial-analytics-and-data-science-job-by-learning-the-r-programming-language

https://www.quanticate.com/blog/bid/102741/using-the-statistical-programming-language-r-in-the-pharma-industry

https://www.nbastuffer.com/analytics101/nba-teams-that-have-analytics-department/

https://en.wikipedia.org/wiki/R_(programming_language)

https://www.payscale.com/research/US/Job=R_Programmer/Salary

http://analyticstrainings.com/?p=101