Data Science Language Analysis

EDA Report

Nazli Ozum Kafaee
Prash Medirattaa
Avinash Prabhakaran

2018-04-15

Contents

Introduction

Our data analysis project seeks to understand how the choice of programming language is affected by the preference of data science tasks like data wrangling, data visualization and machine learning. We chose to restrict the programming language to Python and R. We also restricted the data science task to data wrangling, data visualization and machine learning. Our hypothesis was that there is a significant difference in people favoring R or Python depending on their choice of data science task. Therefore, we designed our survey to primarily to test this hypothesis.

In our survey, we had to take into account some confounding variables too. Therefore we collected data on the user's academic background, their attitude towards coding, the first programming language they learned, the number of programming languages they actively used, and the duration of their programming experience in years.

Methodology and Tools

We created the survey in Google Forms. Data science community was the targeted response group. The data was hosted in US and we informed our respondents in Canada of the same at the outset. User consent was taken to proceed. The survey had seven simple questions. We rolled out this survey and succeeded in collecting 85 responses. Initially, the survey was given to current MDS cohort, faculties and TA's. Then the survey was shared on data science channels, WhatsApp groups, and LinkedIn groups.

Data Wrangling

We had to do some initial wrangling to prepare the data collected for exploratory data analysis. Data wrangling was done primarily to capture the academic background of respondents. The first question in our survey was “What is your academic background?”. This question had three options “Computer Science / Computer Engineering”, “Mathematics / Statistics” and “Others”. The “Others” option enabled the user to give their academic background if it was different from the first two categories. Once we analysed the data, we found that “Others” which was the second highest choice comprised of a variety of answers. We also saw that there were aggregate patterns in the data. We observed that engineering and business studies were recurring answers, so we decided to split “Others” to add two new categories “Engineering” and “Business / Economics” taking the options available in academic background to five.

We faced an issue with one of the other questions as well. When we asked the respondents their first programming language, they were given six specific options and “Others” for the rest. In “Others”, they could enter their first programming language if it was not one of the predetermined languages listed by us. We observed that none of the languages that we had listed represented a major group of people and highest number of respondents chose “Others”. Though “Others” comprised of many languages, each had only one or two respondents. Therefore, we had to leave the option as it was.

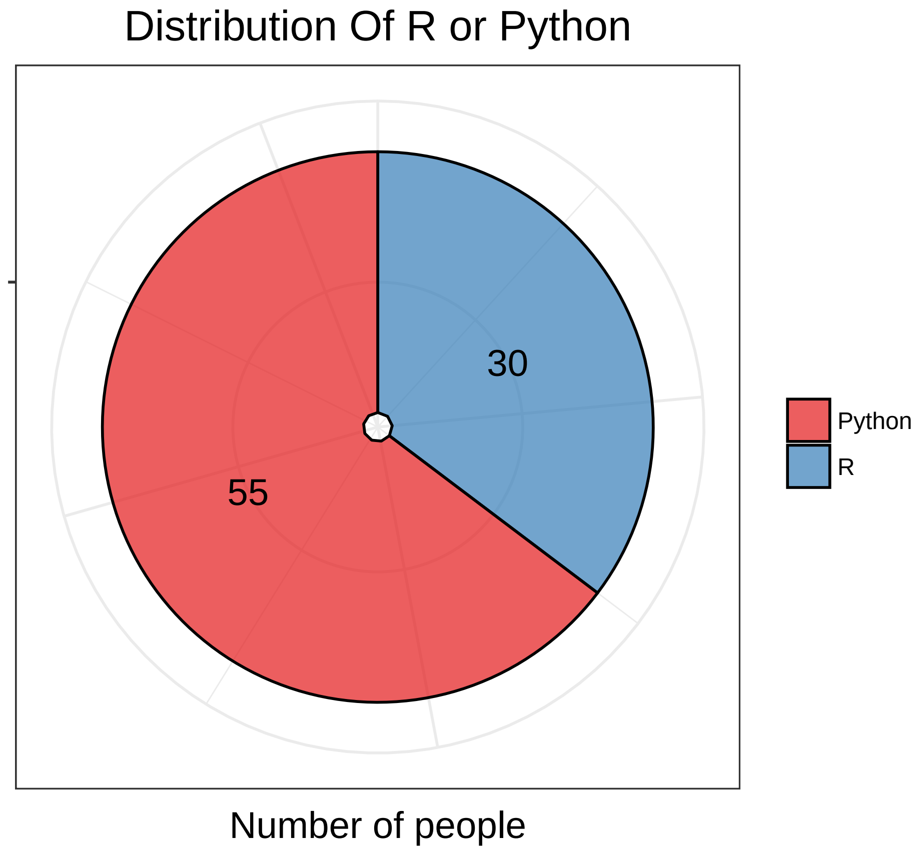
Our code for the data cleaning process described above can be found here.

Visualizations

Distribution of Primary Variables

Language Preference

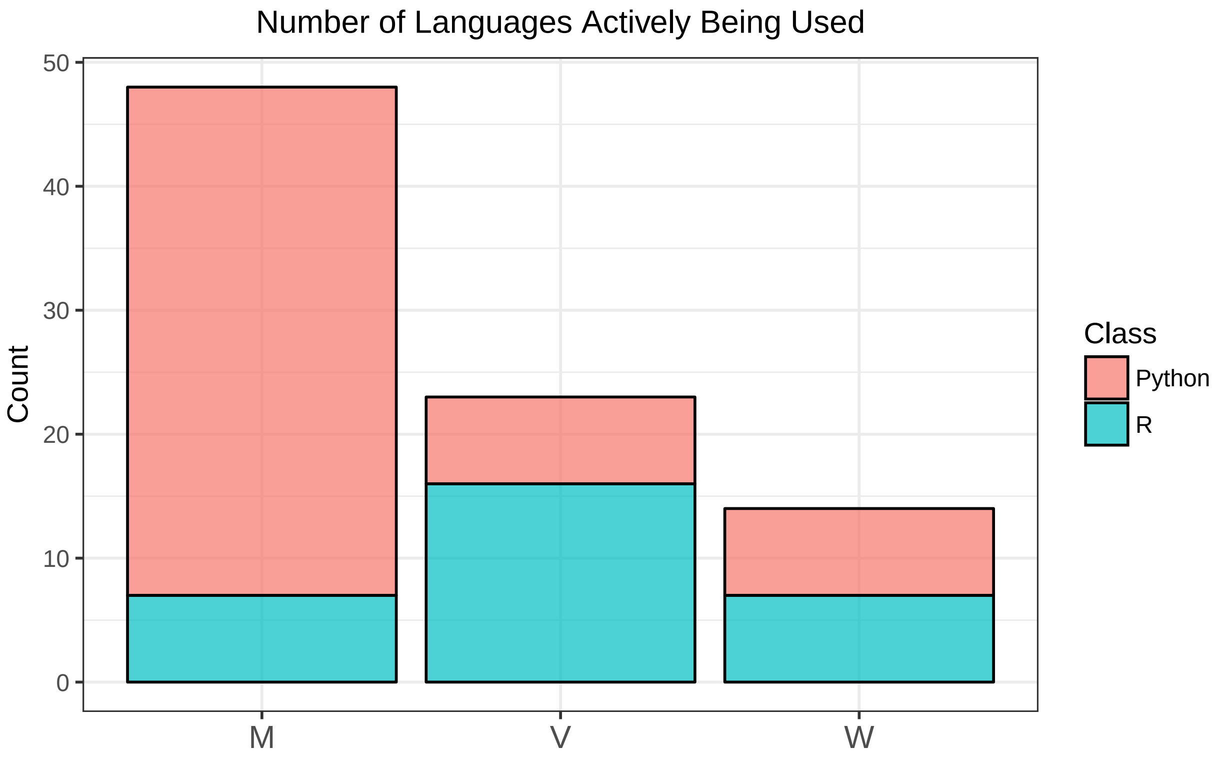
The plot below represents the basic split of language preference with the number of respondents. 55 people say they prefer Python over R and 30 say the reverse.



Even though this is not an even distribution, we can say that we have collected a good number of responses from people to get a good picture of preference for either of the languages.

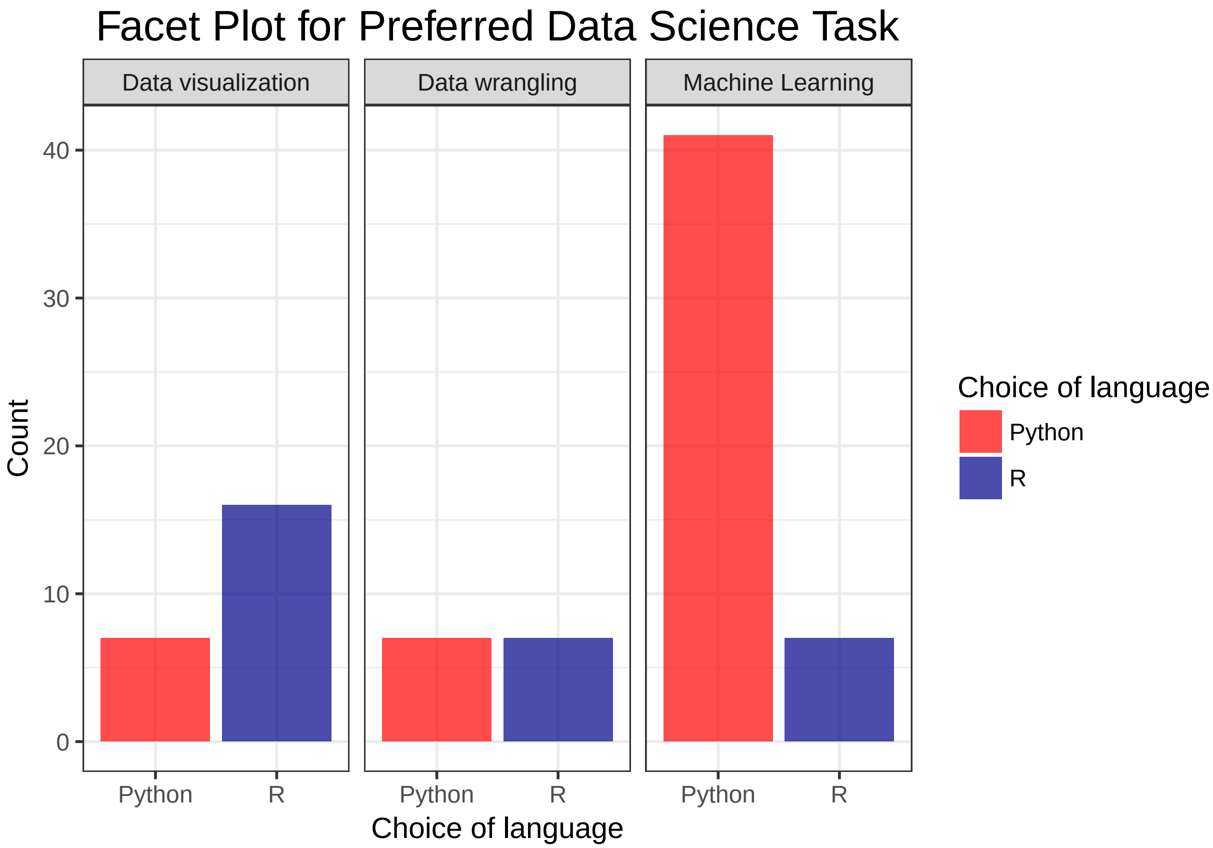
Favorite Data Science Task

The plot below shows the split of people with a different choice for their favorite data science task. We can see that the majority have chosen machine learning as their preferred data science task. Data wrangling seems to be the least favorite based on the numbers.



M: Machine Learning, V: Data visualization, W: Data wrangling

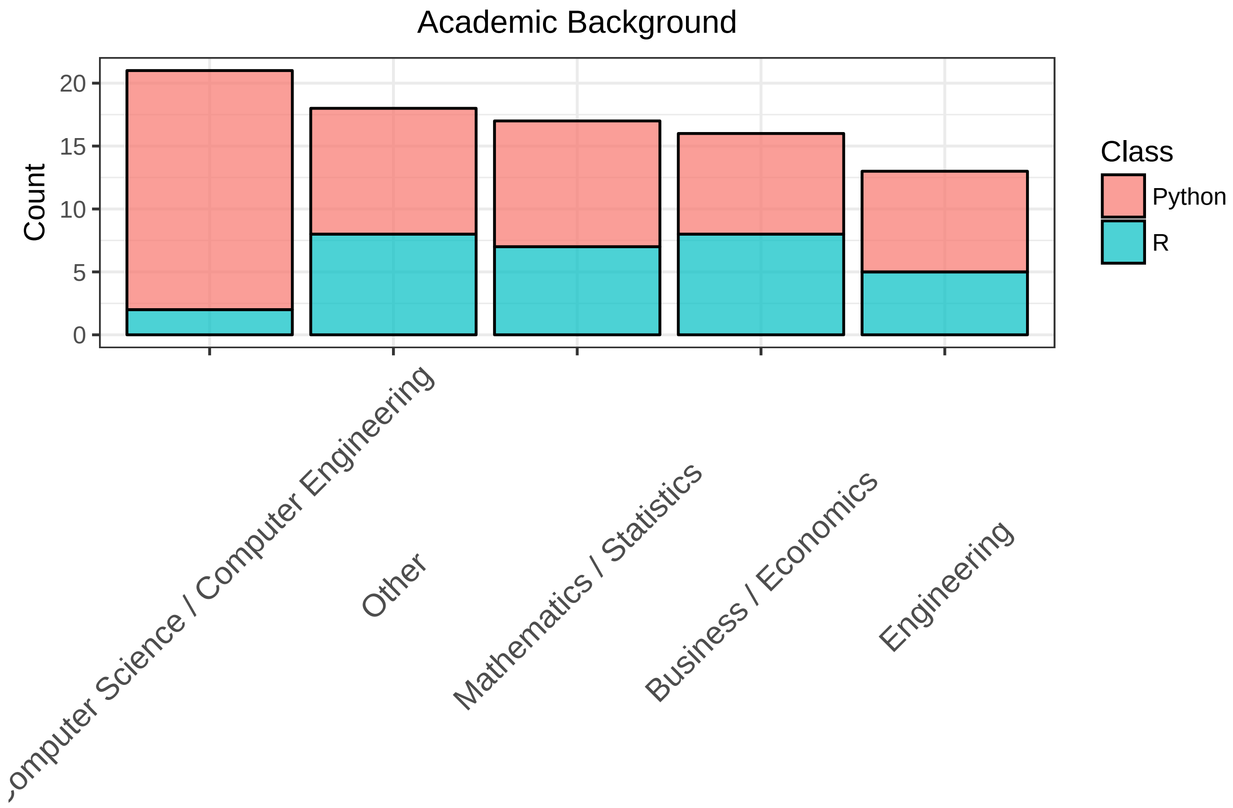
Of course, our main aim is to relate these tasks to the programming language people prefer more. For Machine learning, Python is the preferred language, but for Data visualization, R was the preferred language. Both these results are not surprising and went with our initial expectations. Interestingly, for wrangling, Python and R were equally preferred. This differed with our initial hypothesis where we expected R to be preferred when it came to data wrangling.



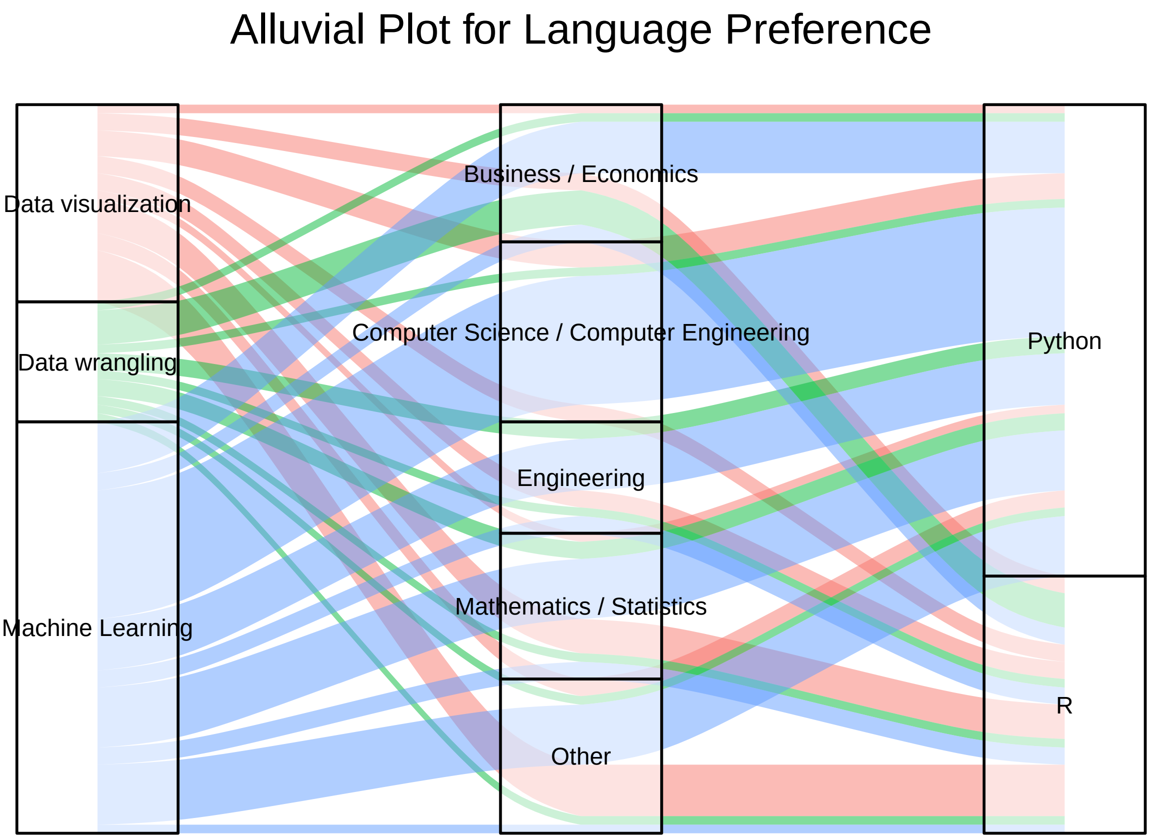
Distribution of Confounding Variables

Academic Background

We had thought that the academic background would be a confounding variable as people with Computer Science/Computer Engineering background would have been introduced to Python as part of their degree and R would have been introduced to students of Mathematics/Statistics degrees. However, we had not anticipated any bias towards R or Python by students of any other degrees.



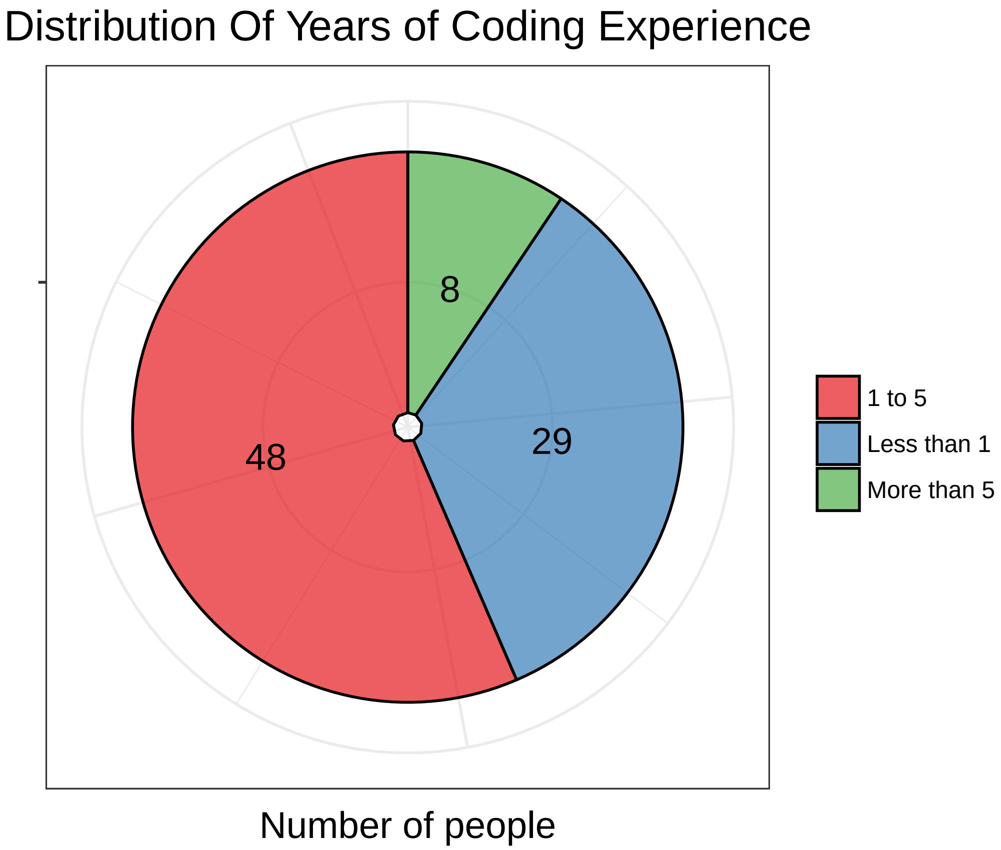
In our survey, we captured 85 responses in total. The maximum number of respondents were computer science or computer engineering graduates closely followed by mathematics and statistics graduates. As we can see above, our initial hypothesis about computer graduates leaning more towards python seems to be relevant. However, there is no significant difference in the preference between Python and R for other academic backgrounds.



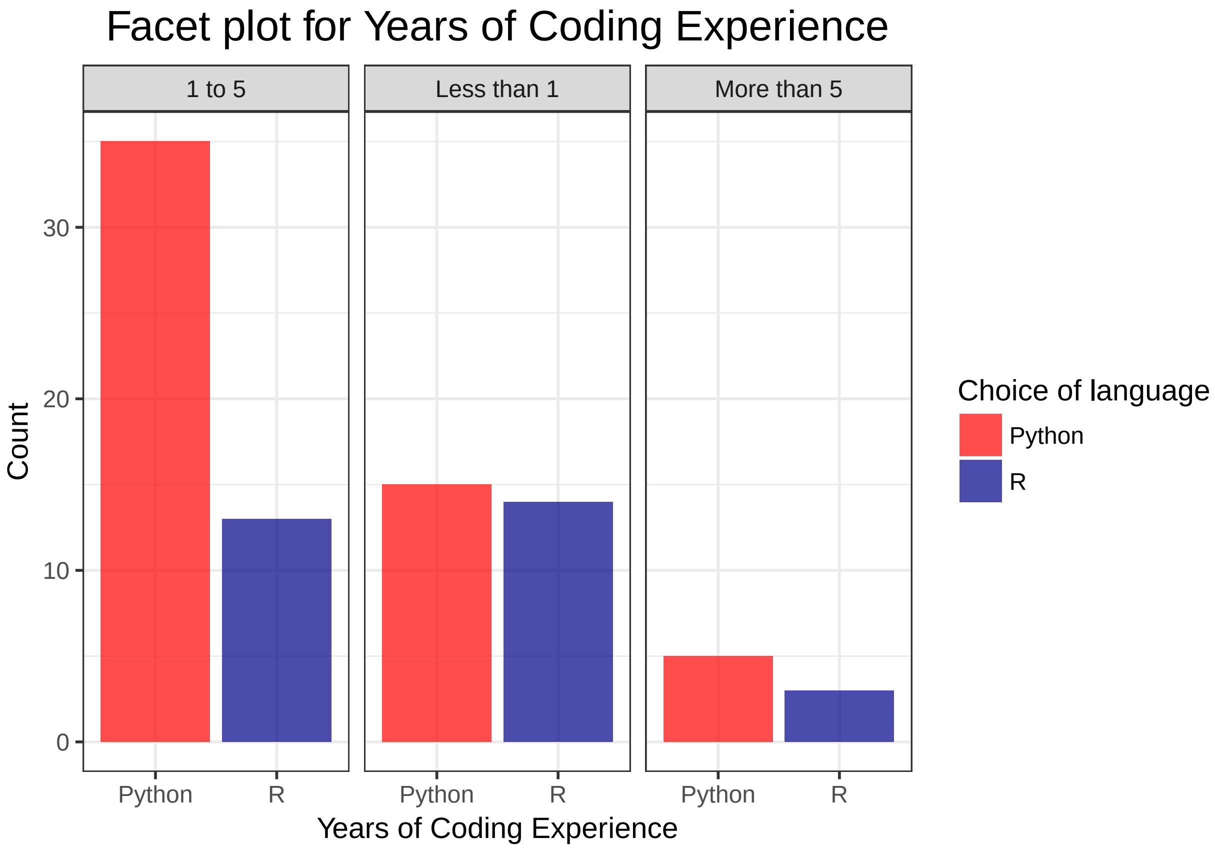
The graph above shows the relationship between the tasks `Data Viz` , `Data Wrangling` and `Machine Learning` and the preferred languages `Python` and `R`.

Coding Experience

We had the belief that the years of coding experience could be a confounder as it can be indicative of how open the user is in selecting a statistical programming language over a general-purpose programming language. However, we also realized that users could become highly opinionated when they had more experience, and they might prefer Python. Therefore, we included this variable in our survey as it would be interesting to analyze.



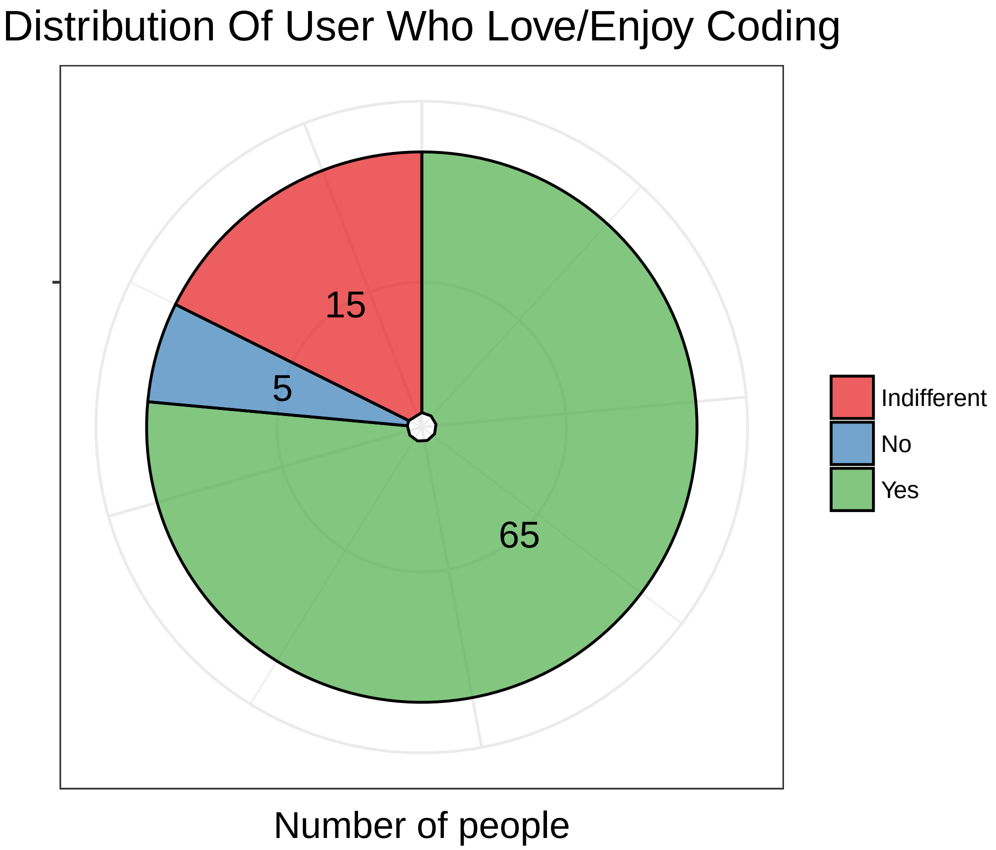
The plot above shows the distribution of coding experience in our survey responses. We can see that people mostly have an intermediate level of experience although the number of novice programmers is also quite high.



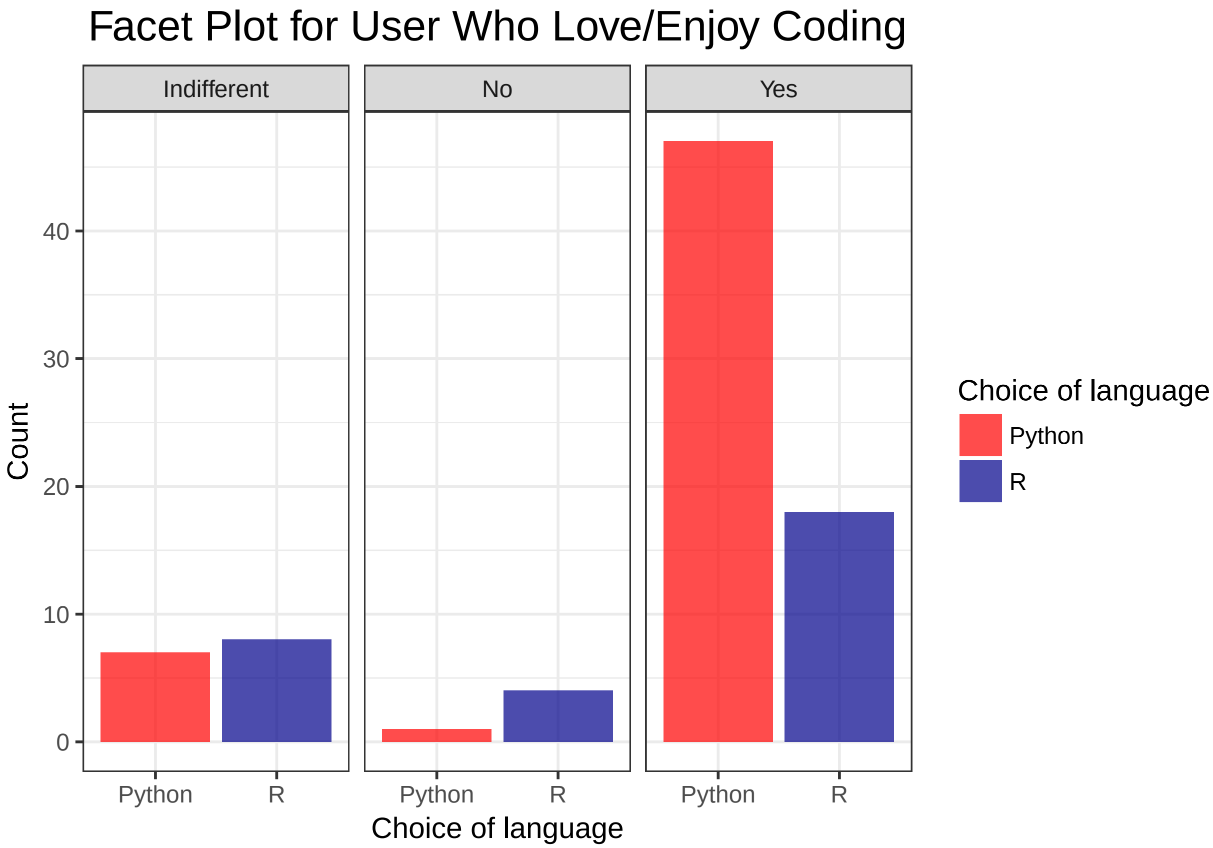
The visualization above is decoding the relationship between the number of years of coding experience and the choice between R and Python. We see Python was the clear choice for intermediate programmers, but the choice does not seem to be clear-cut in the other categories. In all categories just by looking at the numbers Python is preferred more.

Attitude Towards Coding

We thought that the user's outlook towards coding, i.e., love/enjoy coding could be a confounder as Python is a general-purpose programming language and it can be used in various areas and its application is not limited to Data Science/Statistics, whereas R is a statistical programming language and is mainly used only in the fields of Data Science and Statistics.

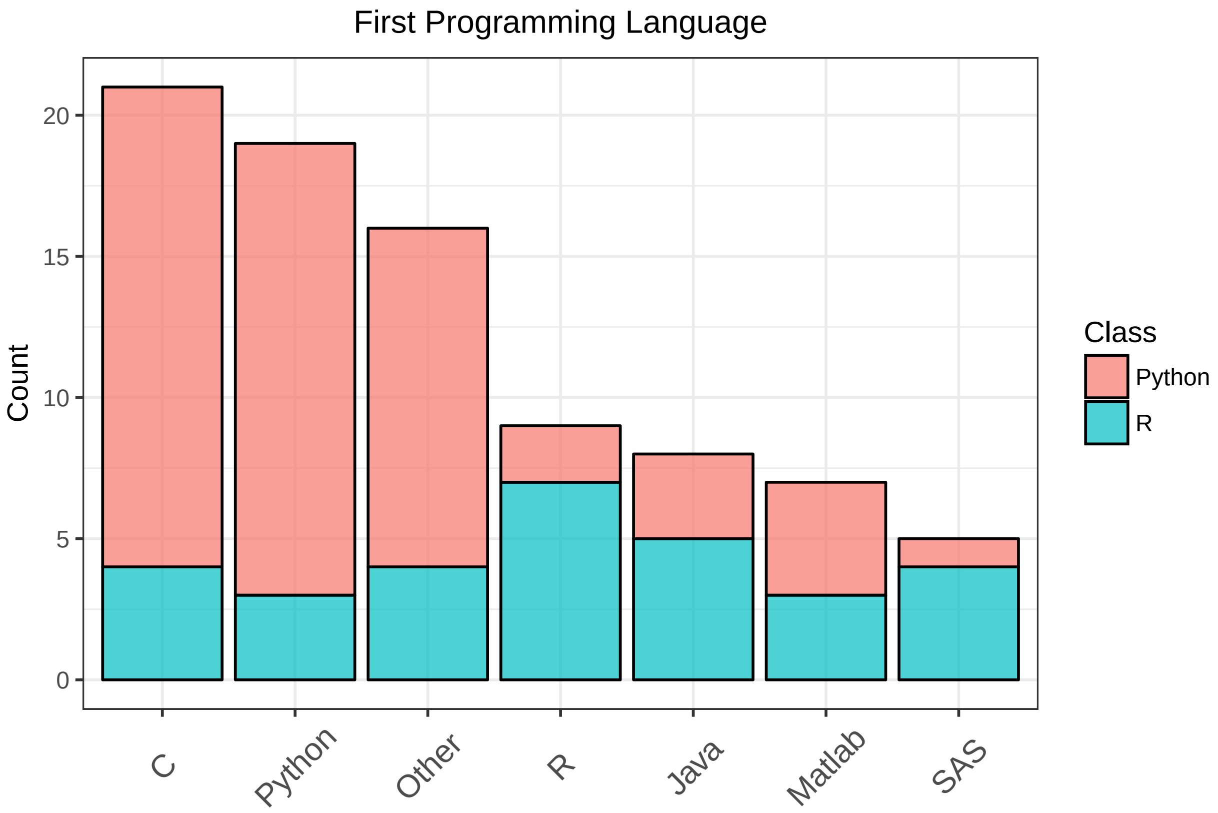


In the pie chart above, we can see that majority of respondents love coding. Coming to the relationship between R and Python, we can see that respondents who loved coding preferred Python while the rest who didn't like coding preferred language R.



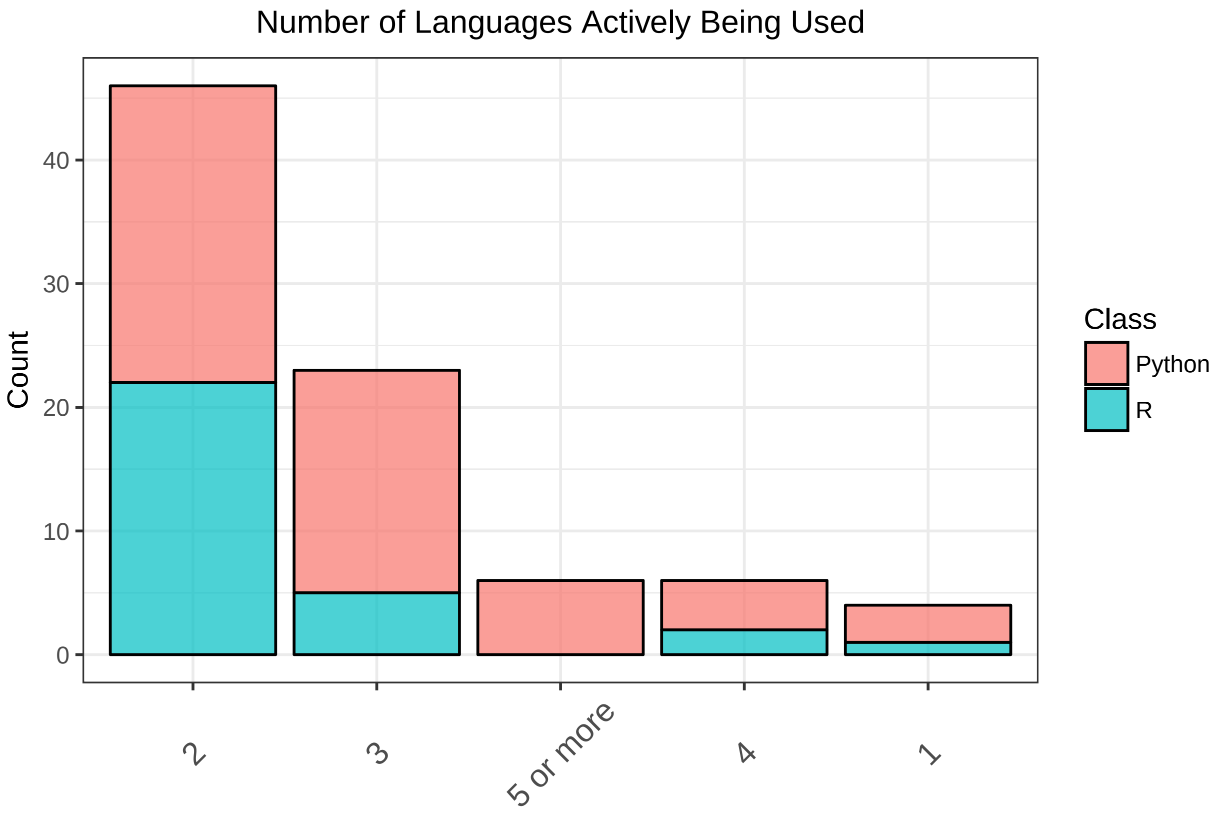
First Programming Language

We thought that a person's first programming language would be very influential as it dictates their style of coding and will also be a deciding factor in what they seek in other languages. Some of the languages, listed below are more closely related to Python whereas some others are more related to R.



Number of Languages Actively Used

The number of programming languages a person actively uses could be a deciding factor as it can dictate how comfortable the user is in using different syntaxes and will also be indicative of how flexible the user.



This graph depicts the results to our question "How many programming languages do you use actively?". The maximum number of respondents stated that they use two languages. Majority of people who use three languages actively list Python as their preferred language. However, based on the results in other categories, this variable does not seem to be a confounder for the preference of Python and R.

Conclusion

After looking at the plots above, we conclude that our data offers possibilities for further analysis. In our final analysis we will analyze whether a person's favorite data science task creates a meaningful effect in their choice between using R and Python. It appears that some of the variables such as academic background and the first programming language are in fact confounders as we had anticipated. Therefore, our analysis will take these into consideration.