Data Science Language Analysis

EDA Report

Nazli Ozum Kafaee
Prash Medirattaa
Avinash Prabhakaran

2018-04-15

Contents

Introduction

In our data analysis project, we wanted to understand how the choice of programming language is affected by the preference of a data science task, i.e., data wrangling, data visualization and machine learning. We chose to restrict the programming language into 2 options as Python and R. We also restricted the data science task to data wrangling, data visualization and machine learning. Our hypothesis was that we would observe a significant difference in a people favoring R or Python depending on their choice of data science task. Therefore, we designed our survey to primarily understand if this hypothesis is correct.

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 use, and their experience in programming in years.

Methodology and Tools

We created the survey on Google Forms. The data is hosted in the US, but we made sure to inform our respondents in Canada of this issue at the beginning of the survey. Users were required to provide their consent to proceed. The survey had 7 easy to follow the question. We successfully rolled out this survey and were managed to collect 85 responses. The audience targeted were specifically from the data science community. Initially, the survey was given to current MDS cohort, faculty and TA's. Then the survey was shared on the various 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 information. The first question in our survey was "What is your academic background?". This question had three main options "Computer Science / Computer Engineering", "Mathematics / Statistics" and "Other". The "Others" option enabled the user to freely type their academic background if it did not fit in the main two categories listed previously. We saw that in the end, "Other" comprised a lot of different answers and made the second highest in terms of share. We decided to split "Other" category and create new categories as we saw that there were aggregate patterns in the data. We observed that engineering and business studies were recurring answers in the results, so we decided to create new categories for these and leave the rest to "Others". Therefore, we added "Engineering" and "Business / Economics" as new categories and remained the rest to "Other".

We were faced with a similar issue in one of our following questions. When we asked the respondents which programming language they learned first, we gave them six main options to choose from and an "Other" option to fill in if necessary. Again, they could freely type the name of their first programming language if it was not one of the predetermined languages listed by us. We observed that the "Other" option comprised of varying languages but each answer held one or two people and none of the languages that we had not listed represented a major group of people. Therefore, we aggregated all answers in "Other" and kept them together.

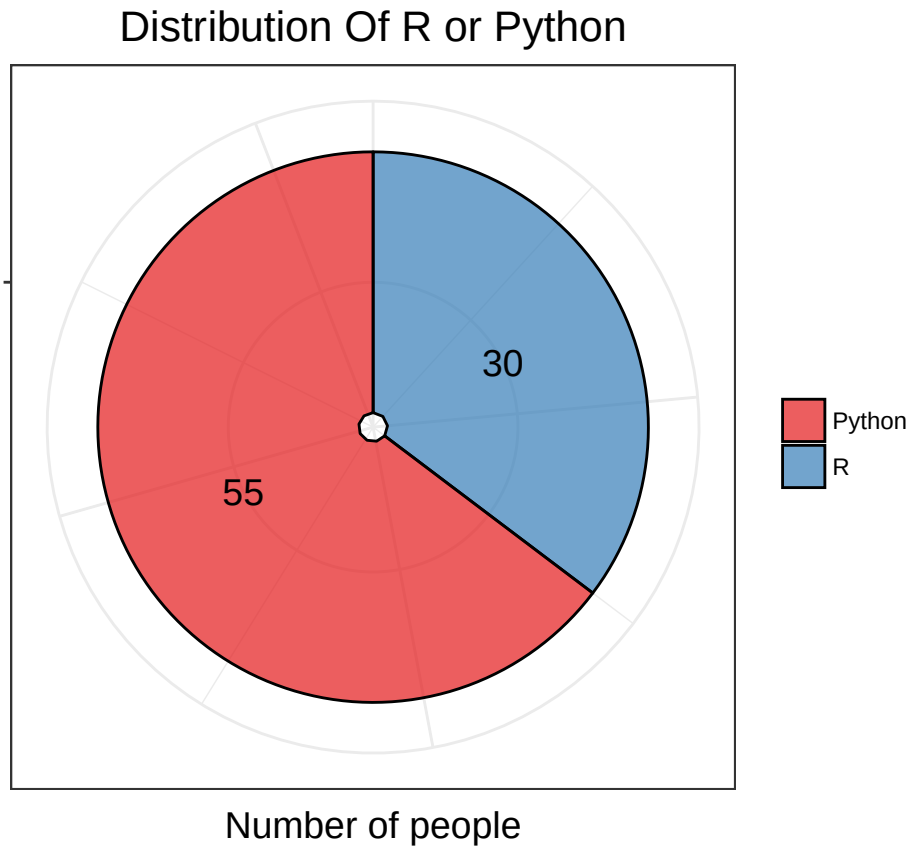
Our code for the data cleaning process described above can be found here.

Visualizations

Distribution of Primary Variables

Language Preference

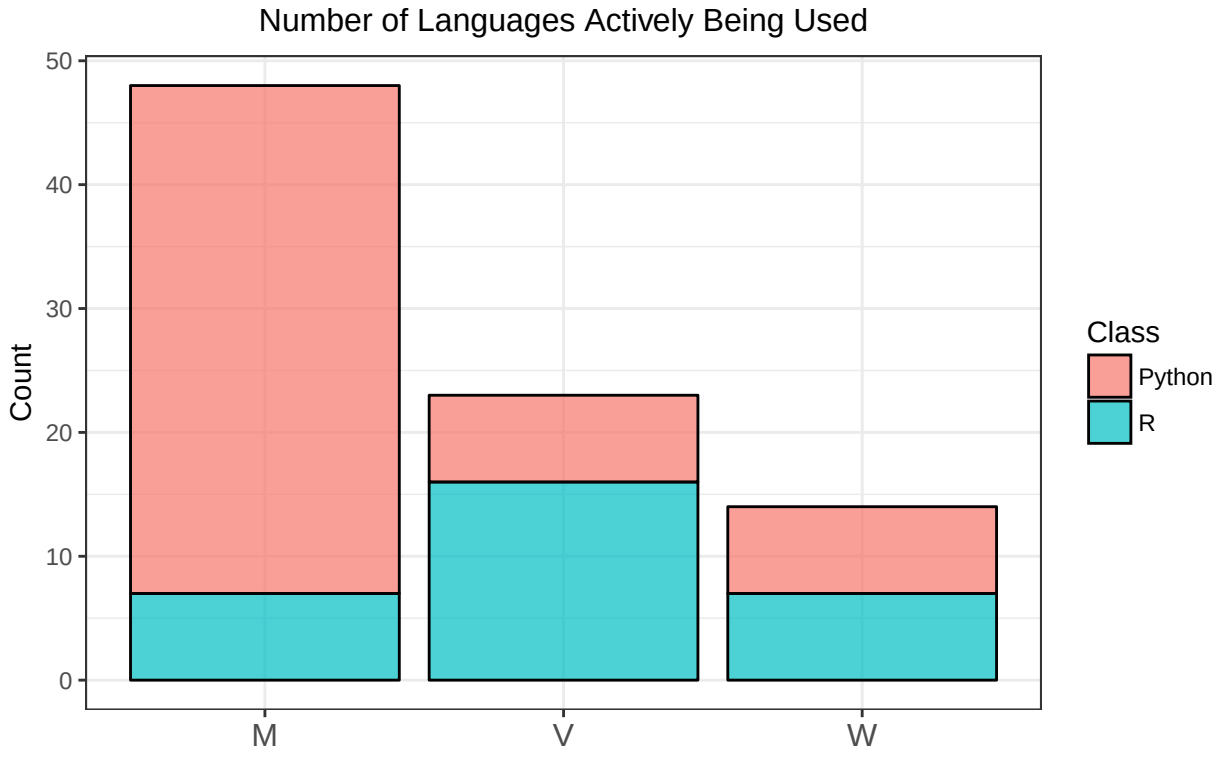
The plot below represents the basic split of the language preference with the number of respondents. 55 people say they prefer Python over R and 30 say the reverse.



Even though there is not an even distribution, we can say that we have collected a good number of responses from people preferring either of the languages.

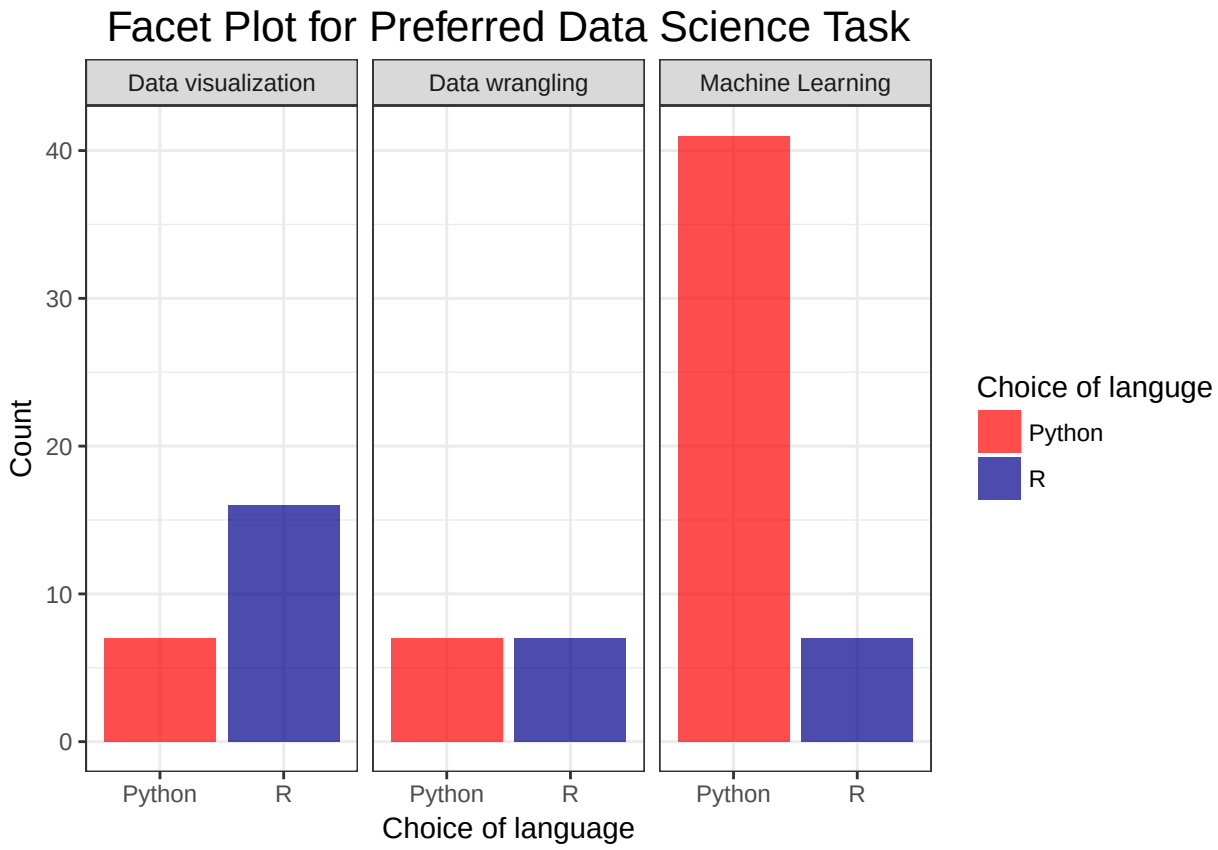
Favorite Data Science Task

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



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

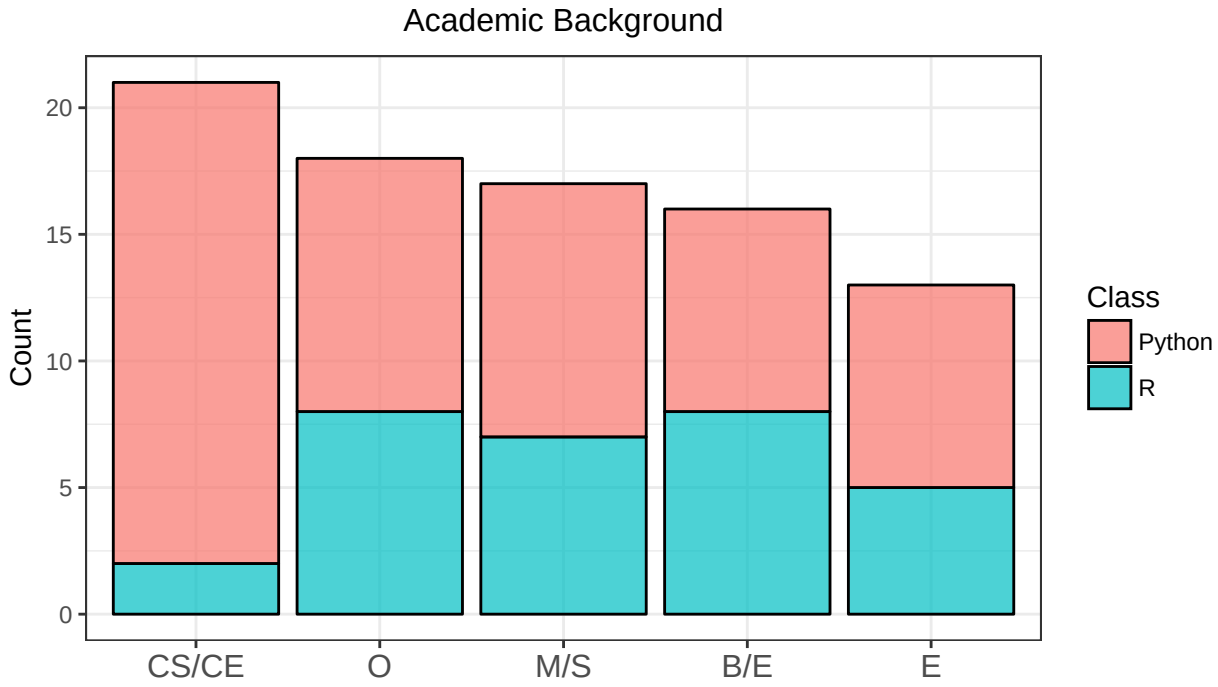
Of course, our main aim is to relate these tasks to which programming language people prefer more. For Machine learning, Python is the preferred language, but for Data visualization, R seems to be the preferred language. Both these results are not surprising given our initial expectations. However, interestingly for wrangling Python and R were equally preferred. Our initial hypothesis was to see R preferred more when it came to data wrangling.



Distribution of Confounding Variables

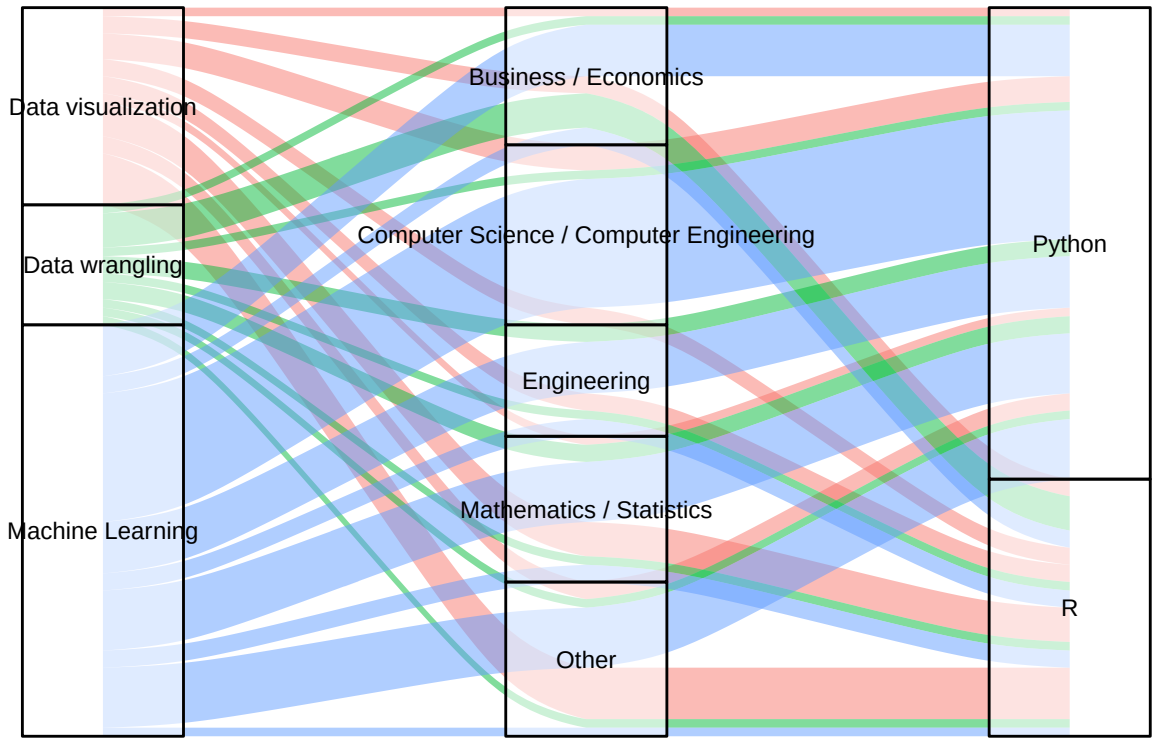
Academic Background

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.



CS/CE: Computer Science/Computer Engineering, O: Other,
M/S: Mathematics/Statistics, B/E: Business/Economics,
E: Engineering

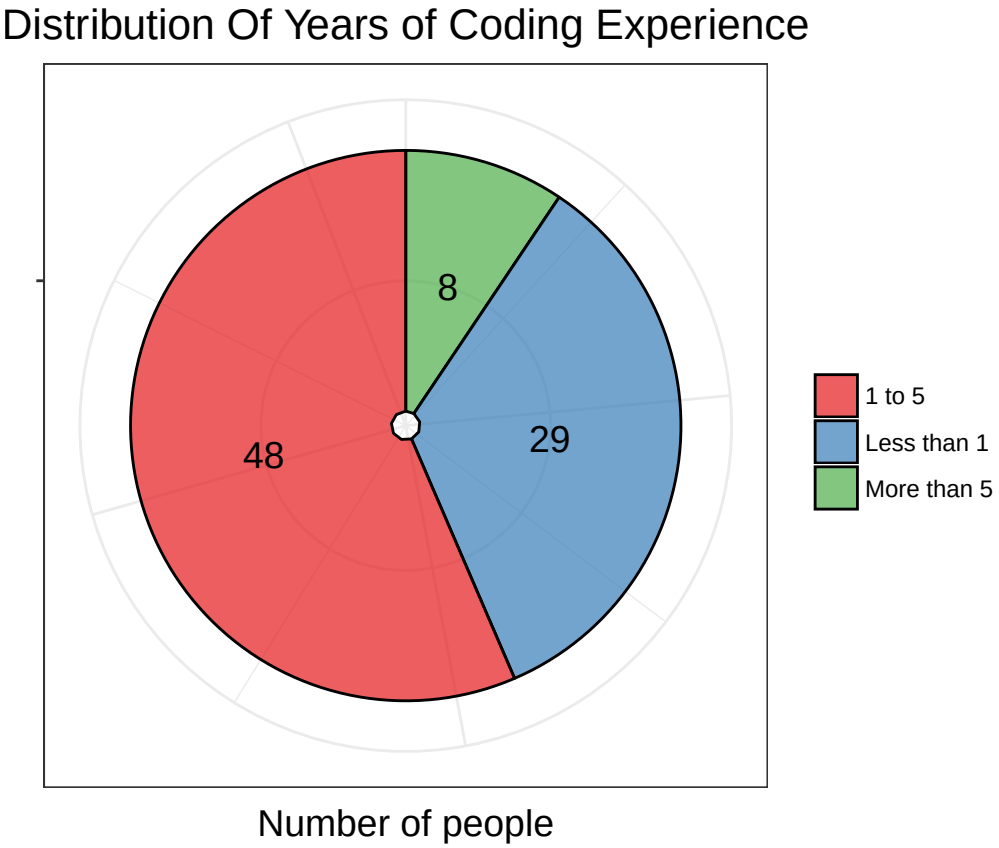
In our survey, we captured 85 responses in total. The maximum number of respondents were computer science or computer engineering graduates closely followed by mathematics and statistics graduates. As we can see above, our initial hypothesis about computer graduates leaning more towards python seems to be relevant. However, there did not seem to be a significant difference in the preference between Python and R given other academic backgrounds.



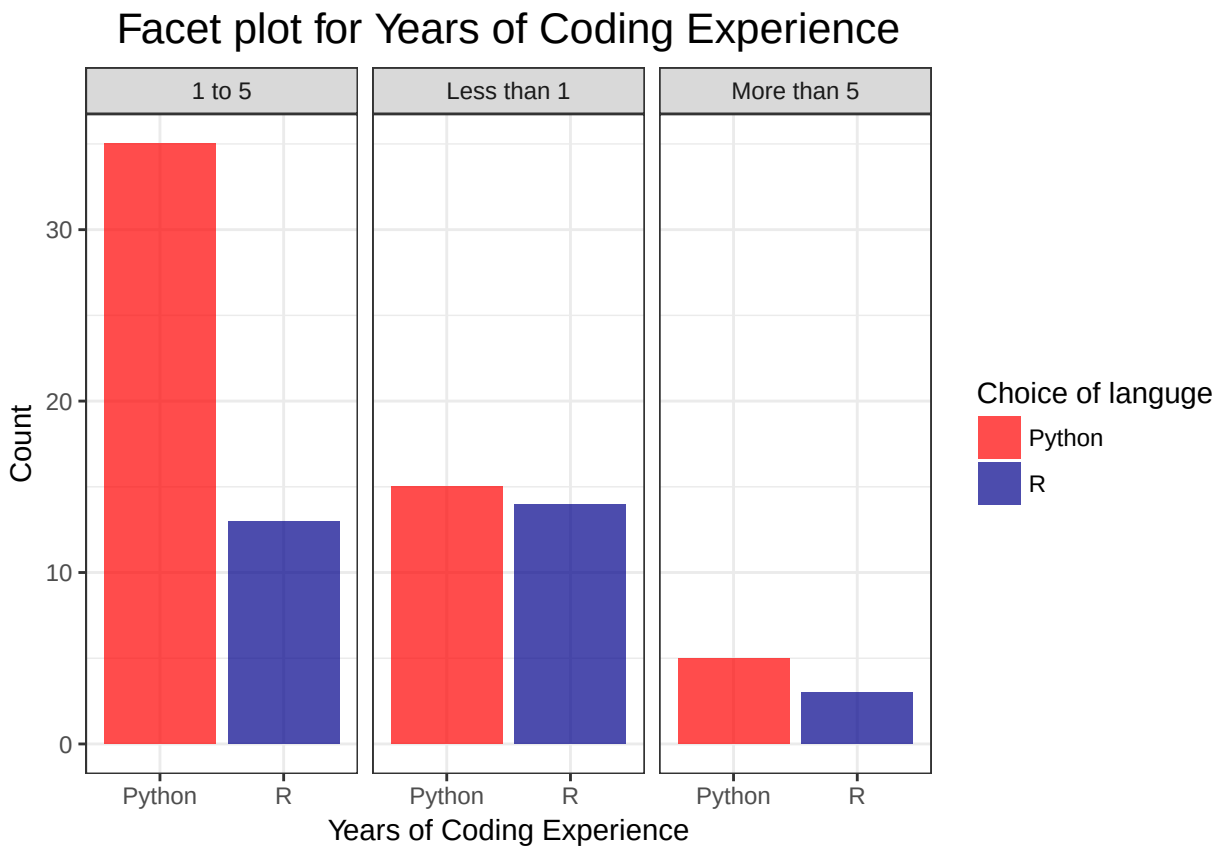
The graph above shows the relationship between the tasks `Data Viz` , `Data Wrangling` and `Machine Learning` and the preferred languages `Python` and `R`.

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 realize that it is possible that a user can become highly opinionated when they have greater experience, and they might prefer Python. Therefore, we wanted to include this variable in our survey as it would be interesting to analyze.



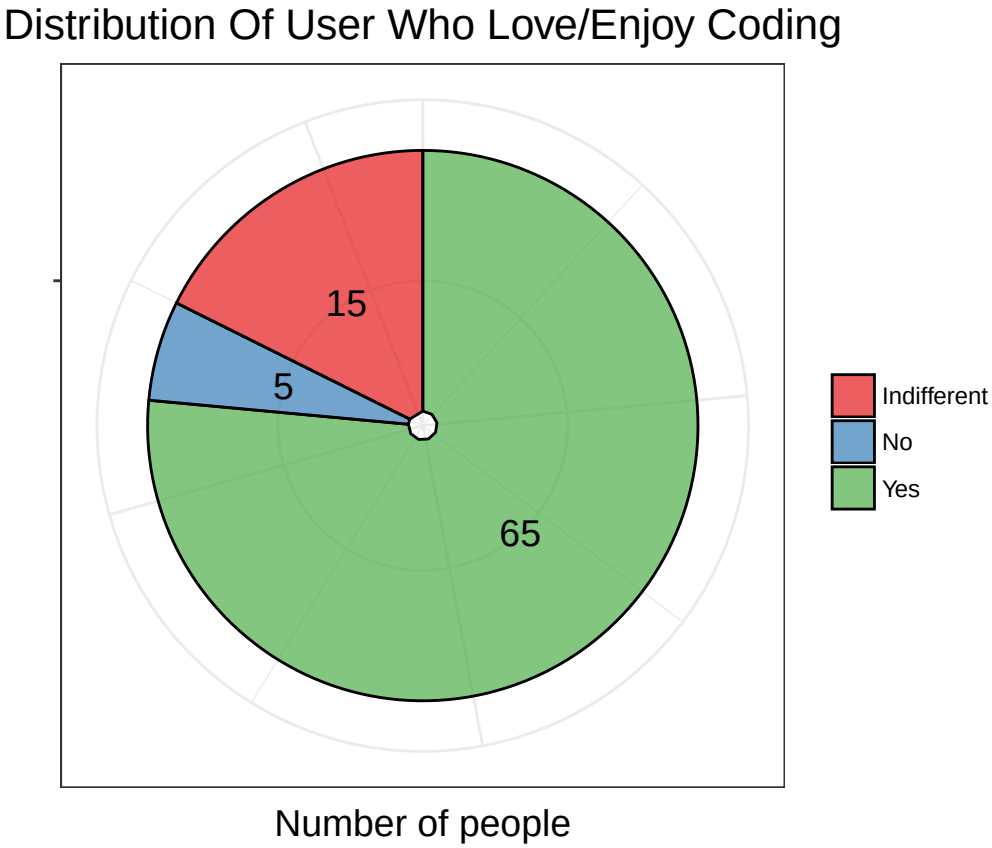
The plot above shows the distribution of coding experience in our survey responses. We can see that people mostly have an intermediate level of experience although the number of novice programmers is also quite high.



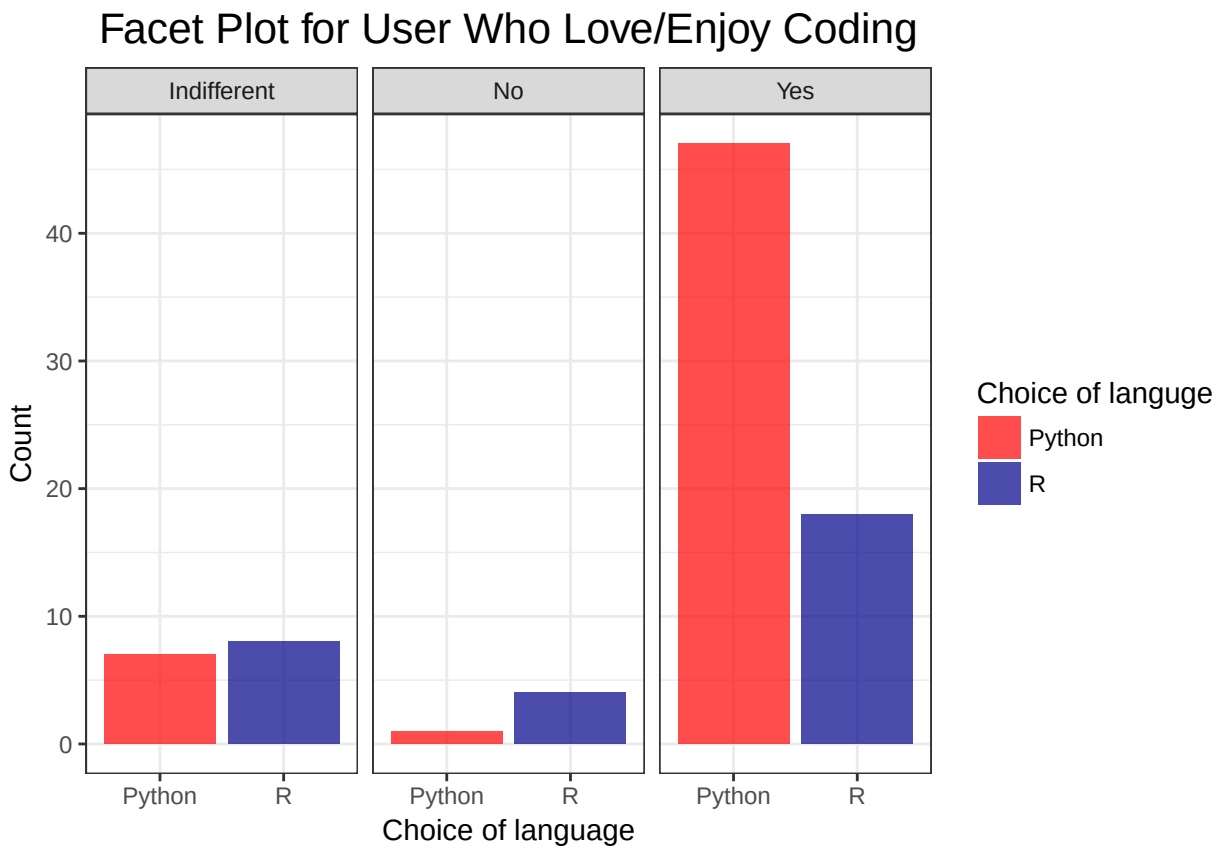
The visualization above is decoding the relationship between the number of years of coding experience and the choice between R and Python. We see Python was the clear choice for intermediate programmers, but the choice does not seem to clear-cut in the other categories. In all categories, Python is preferred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

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.

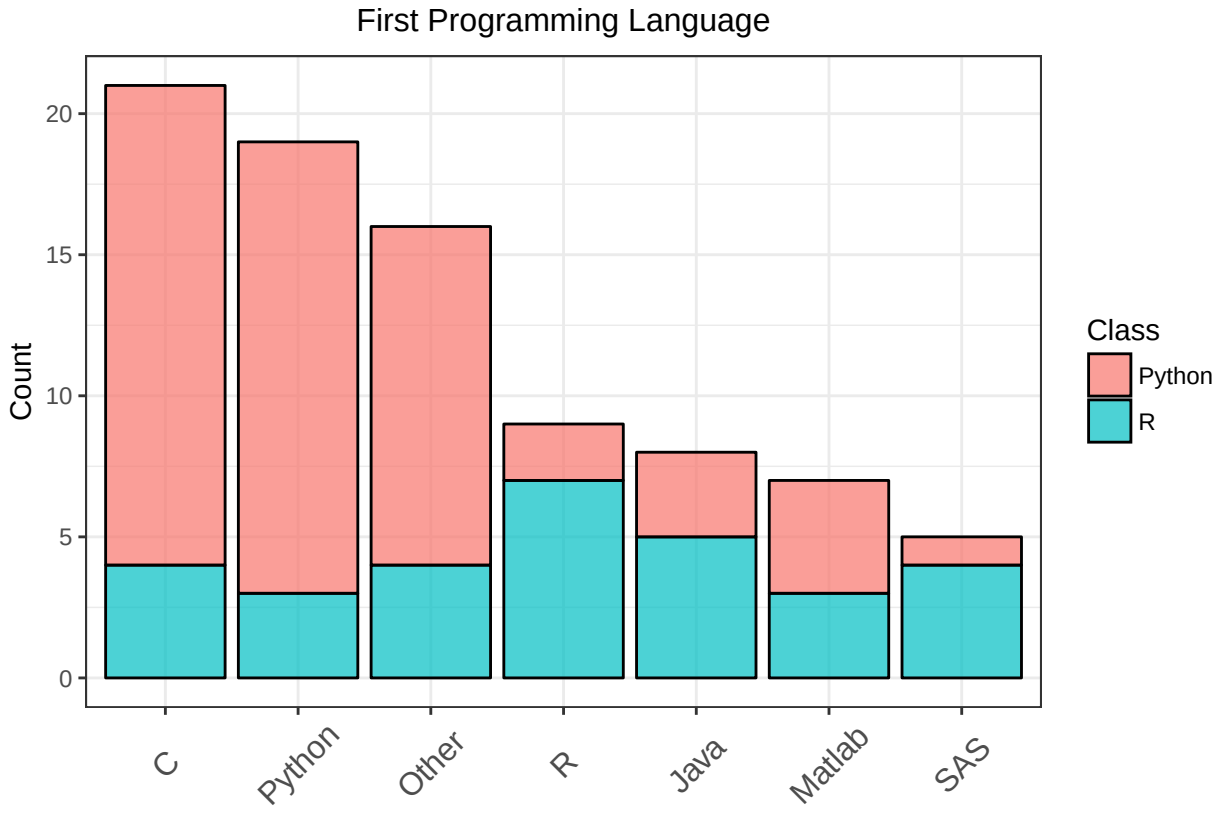


On the pie chart above, we can see the majority of people love coding from the respondents. Now we tried to see what can be the relationship between R and Python. We can see respondents who loved coding preferred Python, and for the people who didn't like coding their preferred language is R.



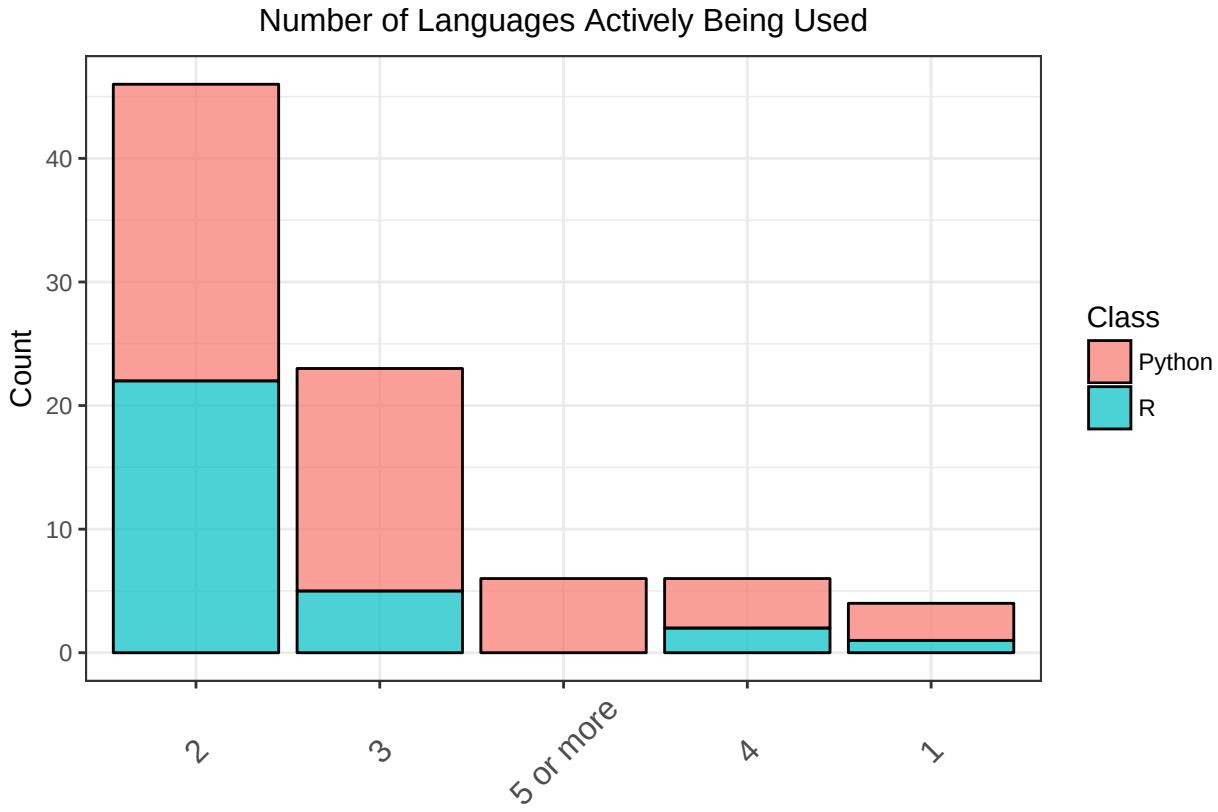
First Programming Language

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 for 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 appears to be promising for further analysis for our final analysis where we will analyze whether a person's favorite data science task creates a meaningful effect in their choice between using R and using 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.