Data Science Language Analysis

Final Report

Nazli Ozum Kafaee
Prash Medirattaa
Avinash Prabhakaran

2018-04-15

1 Introduction

It is common to hear from people working with data that they have a clear choice between R and Python. Almost everyone who has worked with both R and Python have one or the other as their favorite. We were curious if there might be a specific reason underlying such choice. After some brainstroming, we came up with the hypothesis that a person's choice between R and Python might be due to their preference in a specific data science task such as data visualization, data wrangling and machine learning. This hypothesis was based solely on observations and personal experience, but we set the goal to explore such causal relationship (if there exists one) with data in our data analysis project.

2 Methodology

2.1 Data collection

Our primary data source was an online survey created on Google Forms and distributed to the MDS cohort, faculty and teaching assistants through Slack channels as well as to people in the authors' Linkedin and Whatsapp network. In the end, we managed to collect 85 responses.

One of the concerns we had to address was regarding the storage of the collected data. Since we used Google Forms for our survey, the data was hosted in the US. Assuming that a great majority of our respondents were Canadian residents, we made sure to inform them of the fact that the data being collected is hosted in the US and to get their consent before proceeding further in the survey. More information on this matter can be found in the UBC Office of Research Ethics - Using Online Surveys document.

2.2 Study Design

In our survey, we wanted repondents to answer two main questions:

- Which of the following programming languages do you prefer more?
  Possible answers: "R" / "Python"
- What is your favorite data science task?
  Possible answers: "Data wrangling" / "Data visualization" / "Machine Learning"

In the former, respondents were required to choose one of "R" or "Python" and in the latter, they could choose one of three options which were "Data wrangling", "Data visualization" and "Machine Learning". The answers to these two questions would provide us the information for the dependent and independent variables in our analysis, respectively.

In order to fully discover the causal relationship between task preference and language preference, we also collected data about factors that could have an effect on both of our dependent and independent variable. The primary goal in collecting information on possible confounding variables was to ensure that we can control for these in our analysis later on. We determined five possible confounding variables for which we asked the following questions:

- What is your academic background?
  Possible answers: "Computer Science/Computer Engineering" / "Mathematics/Statistics" / "Other"
- How many years of coding experience do you have prior to using Python/R?
  Possible answers: "Less than 1" / "1 to 5" / "More than 5"
- Do you enjoy/love coding?
  Possible answers: "Yes" / "No" / "Indifferent"
- Which programming language did you learn first?
  Possible answers: "Python" / "R" / "SAS" / "Matlab" / "C" / "Java" / "Other"
- How many programming languages do you use actively?
  Possible answers: "1" / "2" / "3" / "4" / "5 or more"

We thought that 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 people from Mathematics/Statistics degrees would have been introduced to R in general. However, we did not anticipate any bias towards R or Python by graduates of any other degrees. We also believed that the amount of coding experience could be a confounder as it can be indicate how open the user is in selecting a statistical programming language over a general-purpose programming language. However, we also realized 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. Another variable we wanted to collect information on was the user's attitude towards coding. The outlook towards 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. Again, a person's first programming language would be very influential as it dictates their style of coding and would also be a deciding factor in what they seek for in other languages. Some of the programming languages are more closely related to Python whereas some others are more related to R. The number of programming languages a person actively uses could be a deciding factor too as it can dictate how comfortable the user is in using different syntaxes and will also be indicative of how flexible the user.

Our survey can be accessed fully here.

2.3 Analysis Methods

The data collected as a result of our survey was downloaded as a `csv` file and imported into R for analysis. All data wrangling and visualization were done in the R computing environment. The code chunks that download data, apply wrangling and create plots can be found in read_data.R, clean_data.R and get_plots.R, respectively.

3 Exploratory Data Analysis

3.1 Wrangling

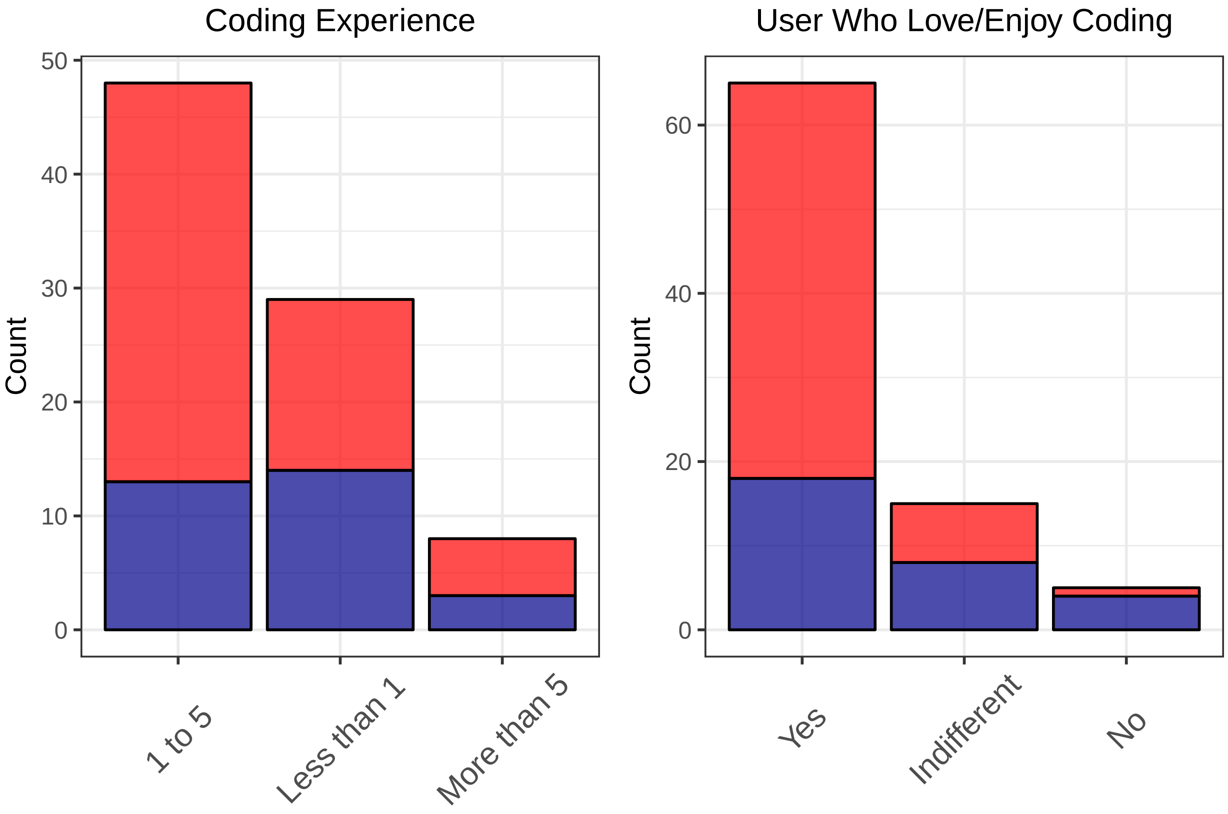
We had to do some initial wrangling to prepare the data collected for exploratory data analysis. Data wrangling was done primarily to capture the academic background information. The first question in our survey was "What is your academic background?". This question had three main options "Computer Science / 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”.

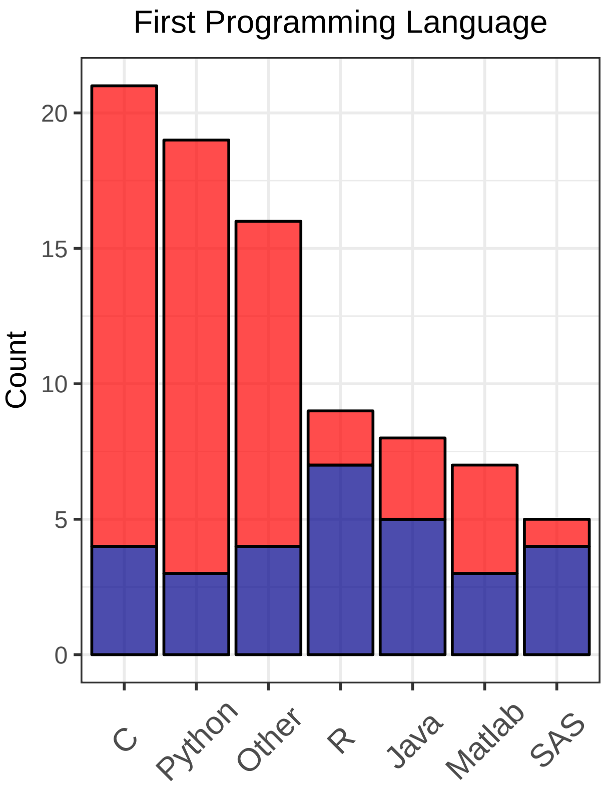
We were faced with a similar issue in one of our following questions. When we asked the respondents which programming language they learned first, we gave them six main options to choose from and an “Other” option to fill in if necessary. Again, they could freely type the name of their first programming 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.

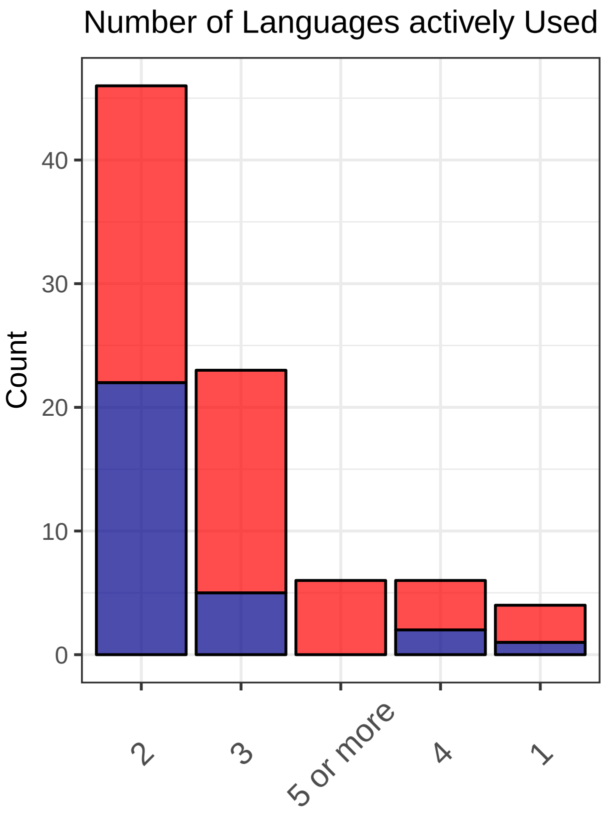
3.2 Visualizations

Explanations to be made



Explanations to be made





4 Statistical Analysis

Fitting a GLM without any confounding variables.

```r
model <- glm(binary ~ task, data = data)
```

Fitting GLM with all the confounding variables.

```r
model <- glm(binary ~ task + background + experience + attitude + first + active, data = data)
```

Final Model with only first language as confounder

```r
model <- glm(binary ~ task + first, data = data)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.4300646	0.1261721	3.4085568	0.0010476
taskData wrangling	0.1639973	0.1370335	1.1967679	0.2351170
taskMachine Learning	0.4494499	0.1055554	4.2579522	0.0000584
firstJava	-0.2797896	0.1683552	-1.6619008	0.1006523
firstMatlab	-0.0981138	0.1752994	-0.5596927	0.5773350
firstOther	0.0542111	0.1336807	0.4055268	0.6862297
firstPython	0.1109147	0.1258204	0.8815319	0.3808095
firstR	-0.3941029	0.1628568	-2.4199350	0.0179127
firstSAS	-0.3855535	0.2038596	-1.8912699	0.0623988

Releveling the data to obtain the comparison between Machine Learning and Data Wrangling.

```r
model <- glm(binary ~ task + first, data = data_relevel)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.8795145	0.0876497	10.0344249	0.0000000
taskData visualization	-0.4494499	0.1055554	-4.2579522	0.0000584
taskData wrangling	-0.2854526	0.1259172	-2.2669873	0.0262345
firstJava	-0.2797896	0.1683552	-1.6619008	0.1006523
firstMatlab	-0.0981138	0.1752994	-0.5596927	0.5773350
firstOther	0.0542111	0.1336807	0.4055268	0.6862297
firstPython	0.1109147	0.1258204	0.8815319	0.3808095
firstR	-0.3941029	0.1628568	-2.4199350	0.0179127
firstSAS	-0.3855535	0.2038596	-1.8912699	0.0623988

5 Results

6 Conclusion