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

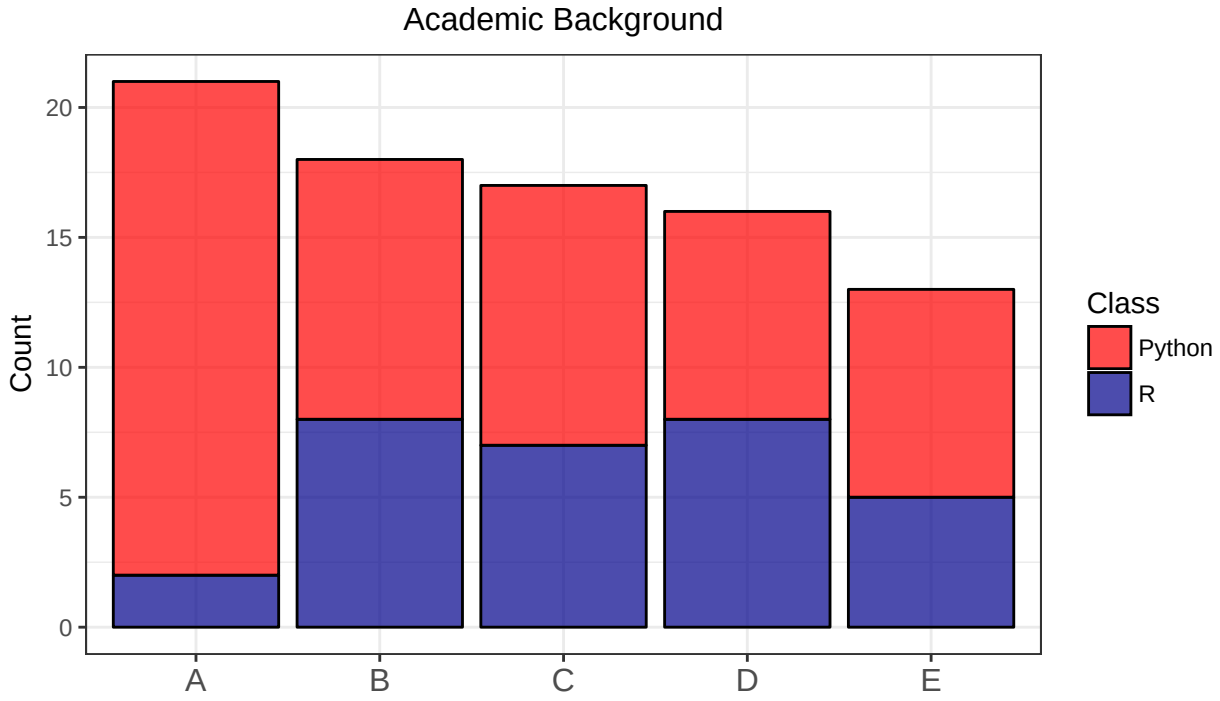
The data collected as a result of our survey was downloaded as a `csv` file and imported into R for analysis. All data wrangling and visualization were done in the R computing environment. The code chunks that download data, apply wrangling and create plots can be found in read_data.R, clean_data.R and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

3 Exploratory Data Analysis

3.1 Wrangling

The data collected from the survey required some initial wrangling in order to be prepared for exploratory and statistical data analysis. The main goal in wrangling was to organize answers that came from the "Other" answer option which enabled respondents to freely type their answer for a specific question if their answer did not correspond to any of the answer options we provided.

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". We saw that "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 some major categories that appeared in these answers but were types differently. Two major categories we observed were engineering and business studies. Therefore, we added "Engineering" and "Business/Economics" as new categories to the academic background variable and left the rest to "Other". Our final categorization of the academic background variable can be seen below together with the distribution of preference between R and Python in each category.



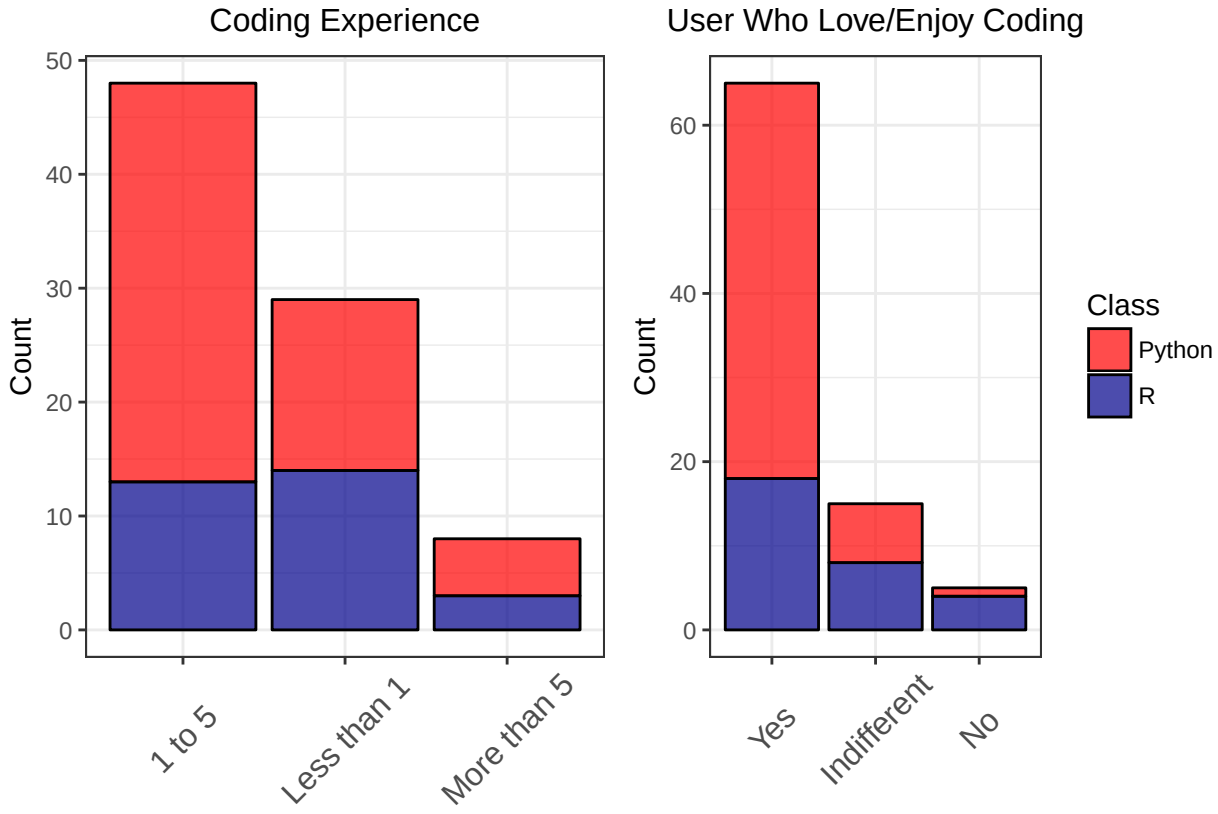
A: Computer Science / Computer Engineering, B: Other,
C: Mathematics / Statistics, D: Business / Economics,
E: Engineering

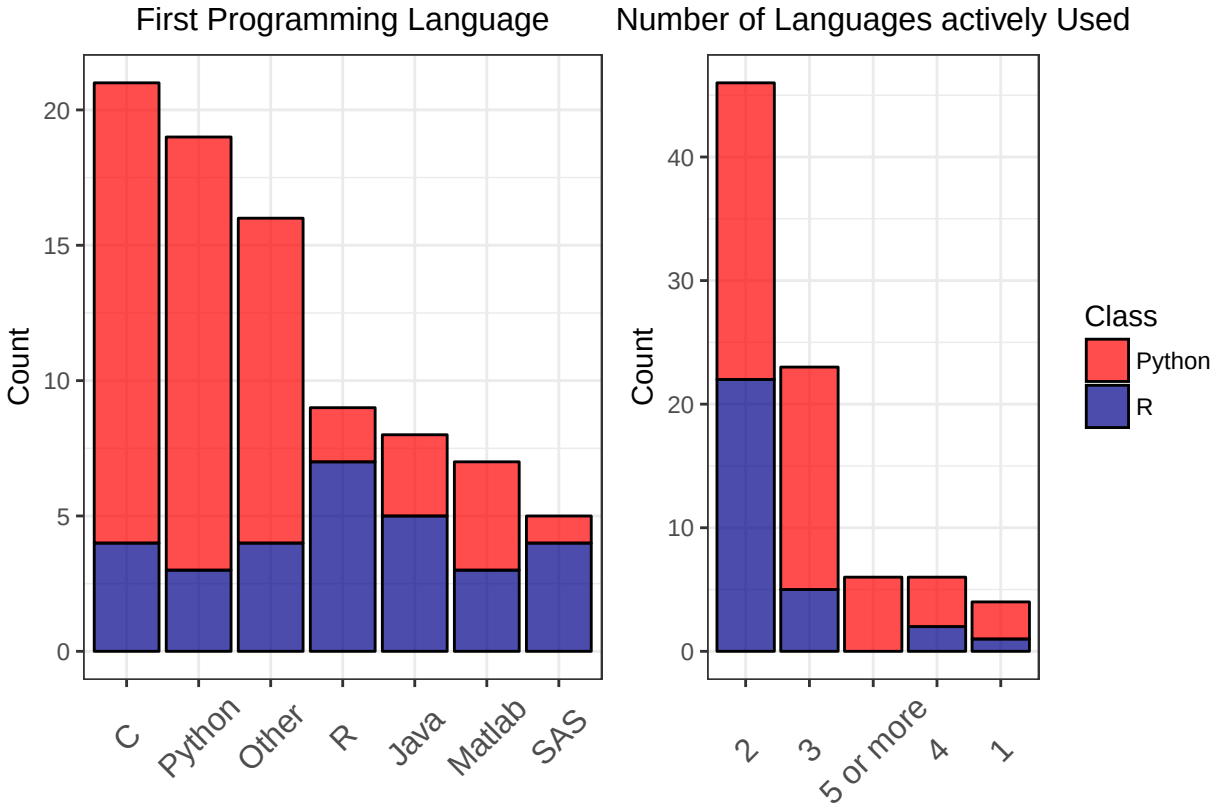
We were faced with a similar issue in one of our following questions. When we asked the respondents which programming language they learned first, we gave them six main options to choose from and an "Other" option to fill in if necessary. Again, they could freely type the name of their first programming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

We made use of stacked bar plots in order to see the difference in the preference between R and Python depending on each category in our possible confounding variables. We tried to observe if the variables we had thought to be confounding were really confounding variables.

As can be seen in the plots below, we have observed that the proportion between the two languages does not change much depending on a person's experience in coding. Again, even though proportions change based on the person's attitude towards coding, this change does not seem to be significant.


Coding Experience
Count
50
40
30
20
10
0
1 to 5
Less than 1
More than 5
User Who Love/Enjoy Coding
Count
60
40
20
0
Yes
Indifferent
No
Class
Python
R


First Programming Language
Count
20
15
10
5
0
C
Python
Other
R
Java
Matlab
SAS
Number of Languages actively Used
Count
40
30
20
10
0
2
3
5 or more
4
1
Class
Python
R

4 Statistical Analysis

```
#Fitting a GLM without any confounding variables.
model.1 <- glm(preference ~ task, family = binomial(link = 'logit'), data = data)
#Fitting GLM with all the confounding variables.
model.2 <- glm(preference ~ task + background + experience + attitude + first + active,
               family = binomial(link = 'logit'), data = data)
#Final Model with only first language as confounder
model.3 <- glm(preference ~ task + first, family = binomial(link = 'logit'), data = data)
#Releveling the data to obtain the comparison between Machine Learning and Data Wrangling.
model.4 <- glm(preference ~ task + first, family = binomial(link = 'logit'),
               data = data_relevel)
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.0771277	0.6589279	3.1522837	0.0016200
taskData visualization	-2.6285727	0.7241022	-3.6301126	0.0002833
taskData wrangling	-1.7508251	0.8015428	-2.1843188	0.0289388
firstJava	-1.5130241	1.0741419	-1.4085887	0.1589568
firstMatlab	-0.4368489	1.0927326	-0.3997766	0.6893210
firstOther	0.4308148	0.9279995	0.4642403	0.6424756
firstPython	0.9496340	0.9716656	0.9773259	0.3284078
firstR	-2.1295673	1.0916359	-1.9508037	0.0510804
firstSAS	-1.9691167	1.3912364	-1.4153718	0.1569595

```
sw_model <- step(model.2, direction = "both", trace=FALSE)
summary(sw_model)
```

```
##
## Call:
## glm(formula = preference ~ task + background + first + active,
##     family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.71388  -0.41985   0.05763   0.36812   2.86063
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                       -4.9757     2.9379
## taskData wrangling                                 2.1236     1.1197
## taskMachine Learning                               4.1142     1.1776
## backgroundComputer Science / Computer Engineering  3.8721     1.6800
## backgroundEngineering                             -0.3729     1.2800
## backgroundMathematics / Statistics                 0.4743     1.0138
## backgroundOther                                    0.4334     1.2456
## firstJava                                         -2.0701     1.4490
## firstMatlab                                       -0.6067     1.5278
## firstOther                                         0.6676     1.2053
## firstPython                                        3.0143     1.6049
## firstR                                            -0.9333     1.4231
## firstSAS                                          -0.6028     1.7588
## active2                                            1.4009     2.1620
```

```
## active3                                                3.8922     2.3830
## active4                                                2.7216     2.6835
## active5 or more                                       20.2089  2047.0820
##                                                      z value Pr(>|z|)
## (Intercept)                                           -1.694 0.090333 .
## taskData wrangling                                     1.897 0.057872 .
## taskMachine Learning                                   3.494 0.000476 ***
## backgroundComputer Science / Computer Engineering      2.305 0.021178 *
## backgroundEngineering                                 -0.291 0.770819
## backgroundMathematics / Statistics                     0.468 0.639855
## backgroundOther                                        0.348 0.727903
## firstJava                                             -1.429 0.153113
## firstMatlab                                           -0.397 0.691302
## firstOther                                             0.554 0.579665
## firstPython                                            1.878 0.060356 .
## firstR                                                -0.656 0.511941
## firstSAS                                              -0.343 0.731806
## active2                                                0.648 0.516992
## active3                                                1.633 0.102411
## active4                                                1.014 0.310493
## active5 or more                                        0.010 0.992123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.372  on 84  degrees of freedom
## Residual deviance:  52.694  on 68  degrees of freedom
## AIC: 86.694
##
## Number of Fisher Scoring iterations: 17
```

```
print(sw_model$formula)
```

```
## preference ~ task + background + first + active
```

5 Results

comparion	reference	estimate	std.error	p.value	odds-ratio	lowerCI	upperCI
Data wrangling	Data visualization	0.8777476	0.8199420	0.2843945	2.4054754	0.4666683	12.3991961
Machine Learning	Data visualization	2.6285727	0.7241022	0.0002833	13.8539814	3.2555726	58.9551588
Data wrangling	Machine learning	-1.7508251	0.8015428	0.0289388	0.1736306	0.0349474	0.8626559

The people whose favorite task is Data Wrangling are 2.4 times more likely to select Python as their favorite language compared to people whose favorite task is Data Visualization.

The people whose favorite task is Machine Learning are 13.9 times more likely to select Python as their favorite language compared to people whose favorite task is Data Visualization.

People whose favourite task is Data Wrangling are 5.8 times more likely to prefer R as their favourite language compared to people whose favorite task is Machine Learning.

6 Conclusion