Exploratory Data Analysis (EDA)

Team Members

- Brenden Everitt
- Miliban Keyim
- Maninder Kohli
- Phuntsok Tseten

The goal of this analysis is to explore the survey data we collected from our MDS 2018-2019 Cohort, 554 TAs and lab instructor. We wanted to understand if undergraduate degree, size of analysis projects, and documentations of packages influenced preference for a particular computing language. We wanted to understand what factors influenced our response(i.e.the preference of a particular language) directly and what were the confounding factors.

Survey Questions Being Explored

- Q1 What is your preferred language to use when doing data analysis?
- Q2 What is the first computing language you learned?
- Q3 What was your undergraduate degree in?
- Q4 What is the typical size of your analysis projects? Small(10K rows), Medium(10-1M rows), or large (1m+ rows)?
- Q5 Which language do you think has better documentation when it comes to its data analysis packages?

Exploratory Data Analysis (EDA)

Importing Data and Data Wrangling

Analyzing Summary Data

Summary of all survey data

```
##       Q1                    Q2                    Q3          Q4
##  Python:36   C/C++/C#        :11   Business        : 7   Large : 6
##  R     :19   Other           :11   Computer Science: 4   Medium:20
##              Python          :11   Engineering     :12   Small :29
##              Java/JavaScript :10   Math/Statistics :14
##              R               : 5   Other           : 4
##              Visual Basic/VBA: 3   Other Science   :14
##              (Other)         : 4
##       Q5
##  Python:35
##  R     :20
##
##
##
##
##
```

Q1 What is your preffered language to use when doing data analysis?

```
##
## Python      R
##   0.65   0.35
```

65% of respondents preferred using Python vs R for data analysis.

Q2 What is the first computing language you learned?

```
##
##         C/C++/C#     HTML/XML/CSS  Java/JavaScript            Other
##             0.20             0.04             0.18             0.20
##           Python                R              SQL Visual Basic/VBA
##             0.20             0.09             0.04             0.05
```

Majority of the survey respondents were split between C/C++/C#, Other and Python at 20% as the first programming language used.

Q3 What was your undergraduate degree in?

```
##
##         Business Computer Science      Engineering  Math/Statistics
##             0.13             0.07             0.22             0.25
##            Other    Other Science
##             0.07             0.25
```

50% of the respondents had an undergraduate degree in either Math/Statstics or Other Science at 25% each.

Q4 What is the typical size of your analysis projects? Small(10K rows), Medium(10-1M rows), or large (1m+ rows)?

```
##
##  Large Medium  Small
##   0.11   0.36   0.53
```

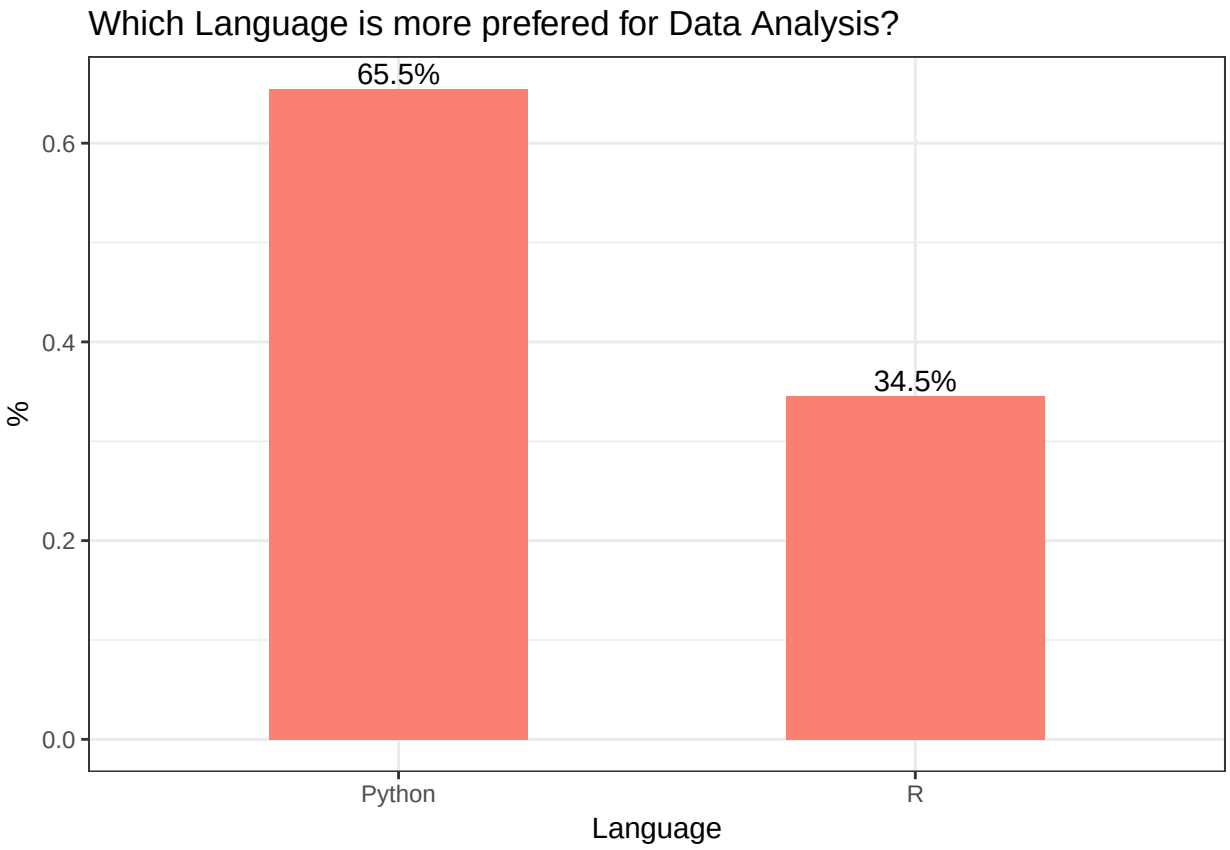
Over half of the respondents had experience with small analysis projects.

Q5 Which language do you think has better documentation when it comes to its data analysis packages?

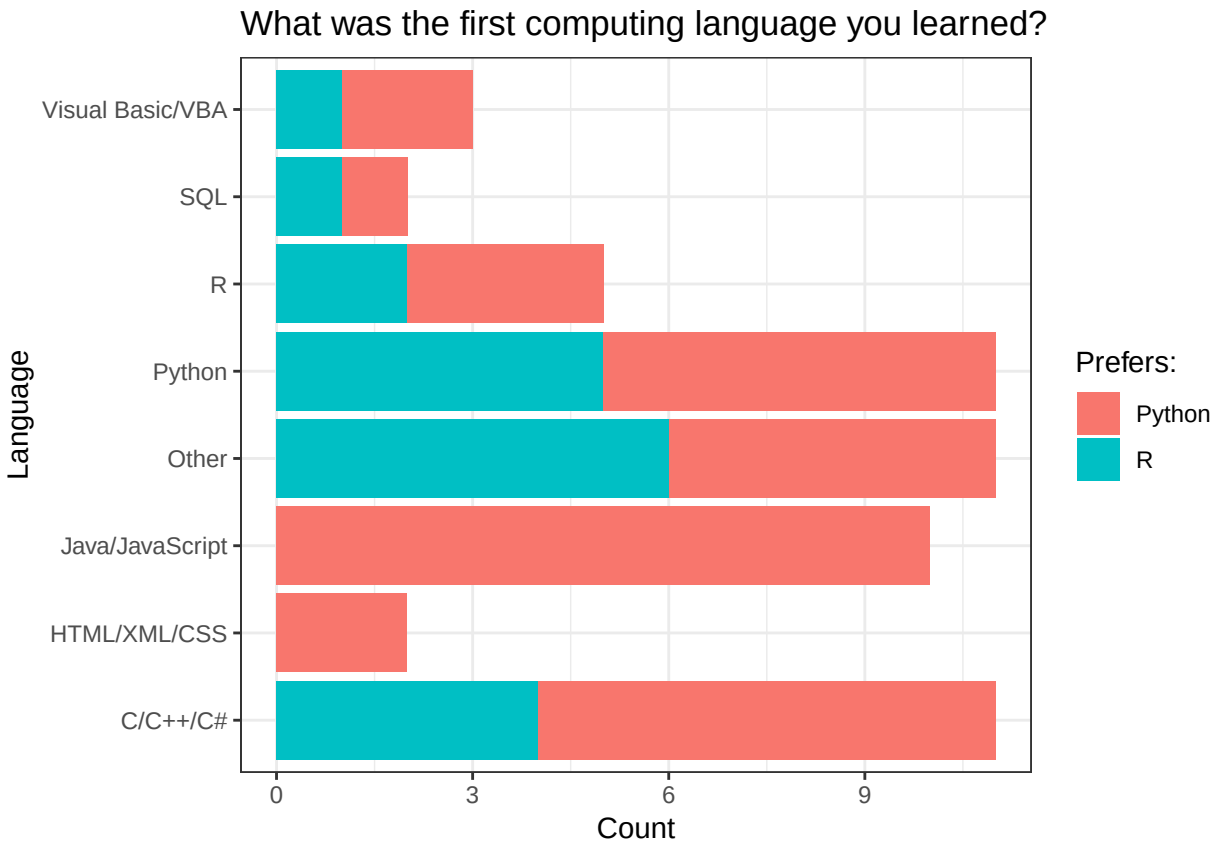
```
##
## Python      R
##   0.64   0.36
```

A large proportion of respondents (64%) believed Python had better documentation.

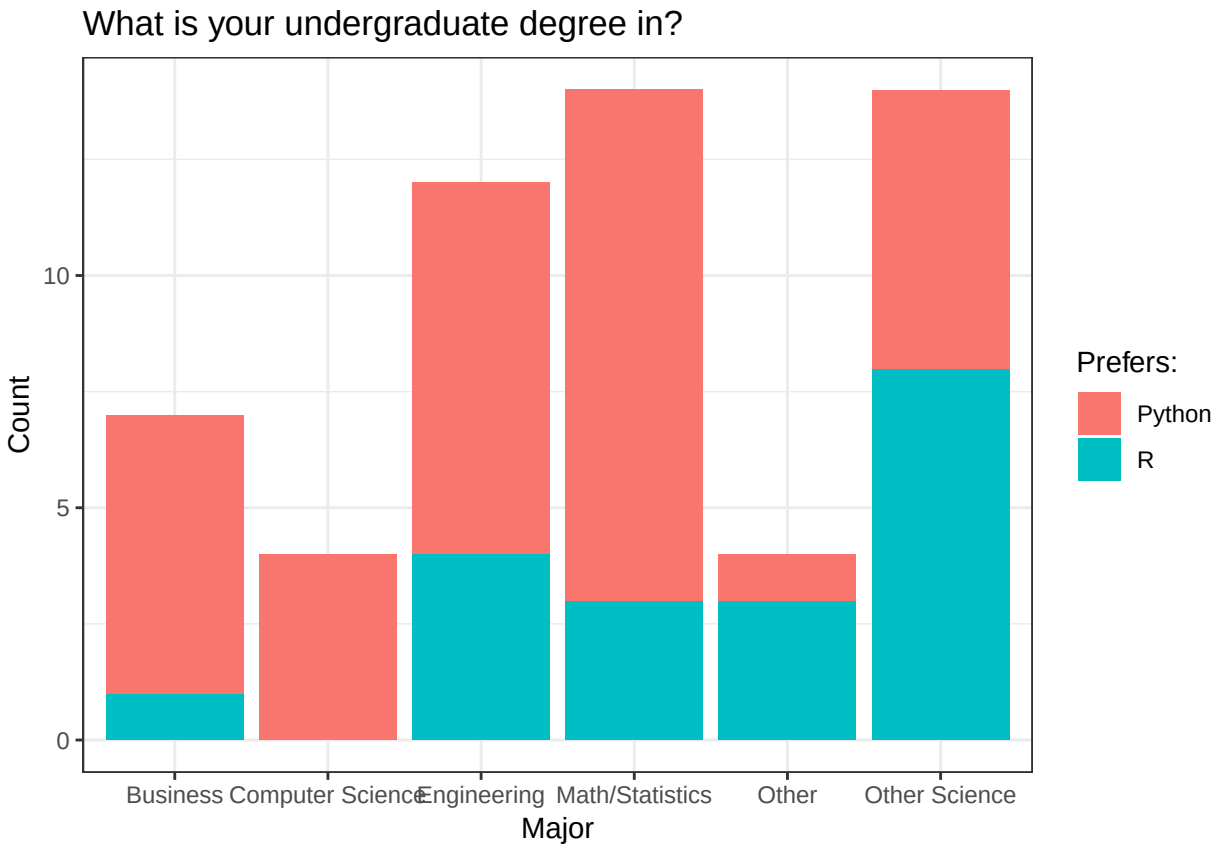
Plots



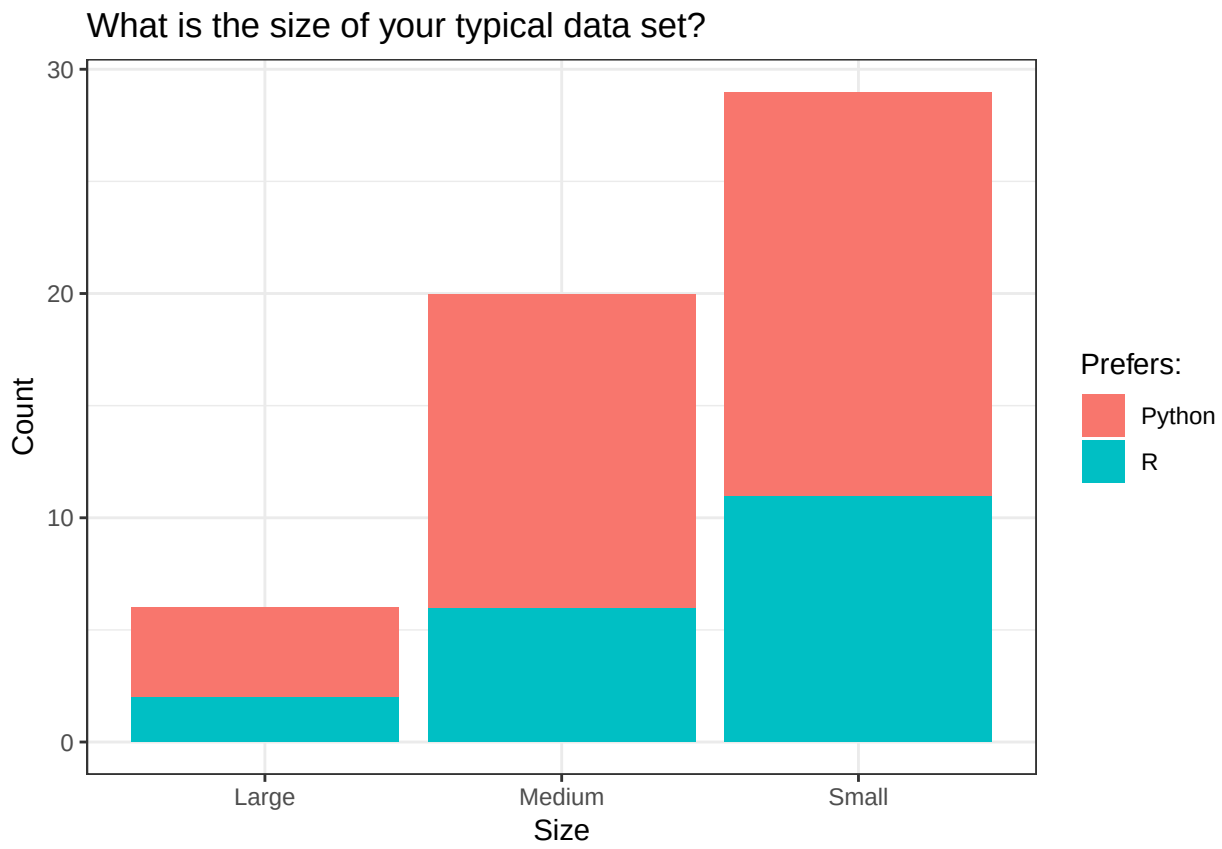
In plot1, we see that more people prefer python



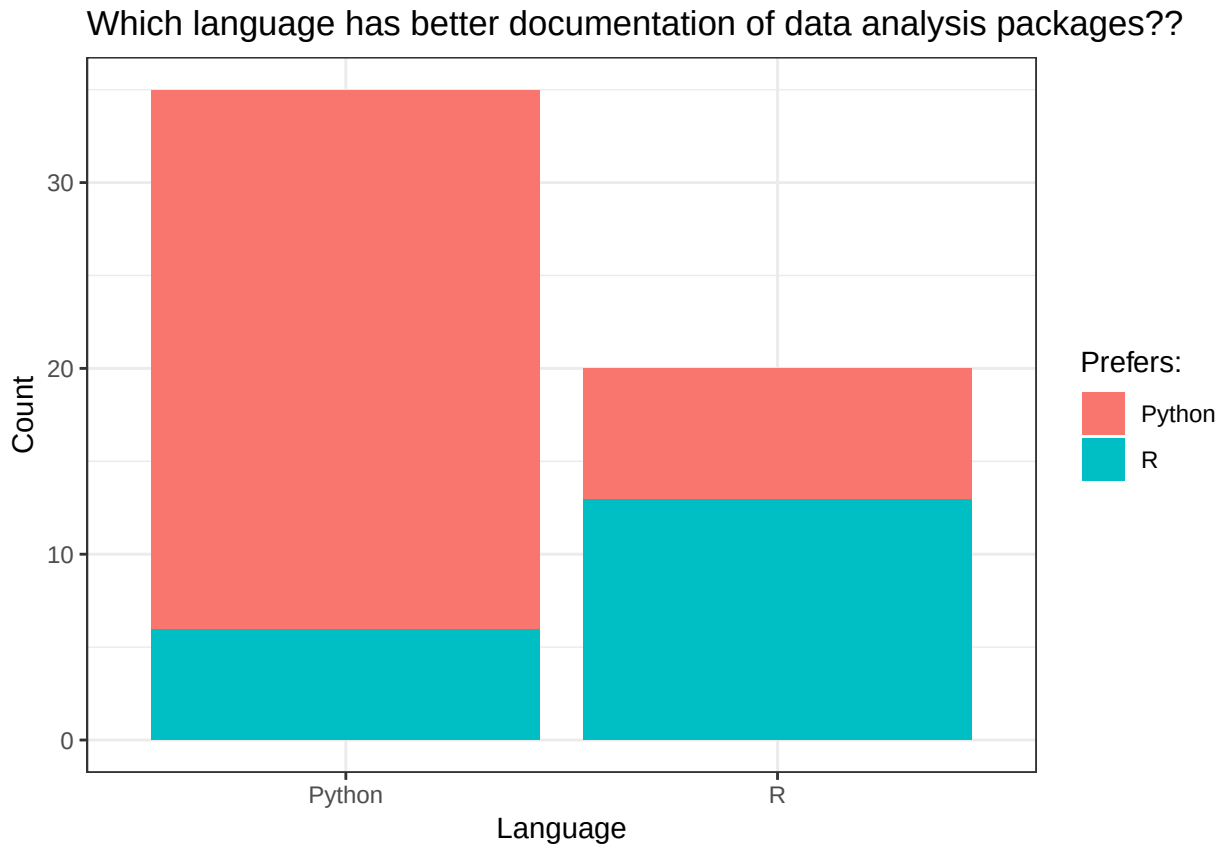
In plot2,those that learned Java/Javascript/HTML/CSS seem to all prefer python over R



In plot3,those with a CS background always prefer Python, those with Sciences/Other degrees seem to prefer R over Python.



In plot4,it seems more people prefer python no matter the size of the data set



In plot5,it seems highly correlated that people prefer to choose the language that they believe has better documentation.