SURVEY REPORT

PROGRAMMING LANGUAGE PREFERENCES AMONG DATA SCIENCE STUDENTS: PYTHON, R, AND JAVA

FACULTY OF COMPUTER AND DATA SCIENCE

18 APRIL, 2023

TEAM 11

SURVEY METHODOLOGY
Dr. Ebrahim Galal

SUPERVISOR
Eng. Passant Hamza

ABSTRACT

INTRODUCTION

Python, R and Java languages are the most popular programming languages used for various purposes. They have their own unique features and functionalities that make them suitable for different tasks.

This survey aims to determine the preferred programming language among computer and data science students. The survey will gather information on participants' experience with each language, as well as the factors that influence their choice.

Insights gained from the survey will inform educational institutions about the suitable programming language for their specifications and education level, leading to better educational outcomes and better-prepared graduates.

OBJECTIVES AND HYPOTHESIS

OBJECTIVES

- Which language is preferred by the students?
- Which language is most used by students depending on their specification?
- What are the functionalities that affect their choice?
- How do students find the ease of learning and using each language?
- Rating students' familiarity level and experience with each language?
- How do job market demand for each language influence the language choice for students?
- How do students choose between these three languages for specific data science tasks or projects?
- What is the importance of having knowledge of multiple languages in the field of data science?

HYPOTHESIS

- The preferred language between the students is Python.
- There's a correlation between age and preferred language
- There's a correlation between level and preferred language
- There's a correlation between familiarity and preferred language
- There's a correlation between the gender and preferred language

PARTICIPANTS AND DESIGN

PARTICIPANTS

Students that will fill this survey will be from the General Department of Computing and Data Science University. We will send this survey to about 700 students. We will be in contact with students to fill the survey through social media, and in the college face-to-face. The survey was conducted using Google Forms to gather the required information. The survey will be a mix of open-ended and closed-ended questions.

SURVEY DESIGN

The survey was consisted of closed-ended questions, as MCQ, ranking questions, and checklist questions having the option of "others" in case the answer the students' want is not given. We tested the survey on a small group first before distributing it. Every student had a view about the survey's purpose before filling it. Students will need about 5-10 minutes maximum to fill the survey and can skip any questions they want.

TIMELINE

TIMELINE

The survey project will be completed in four weeks, divided into seven phases:

APRIL 12
APRIL 17

Phase 1: *Proposal*
Outlining the importance and goals of the survey, specifying the main points of the whole process of making a survey, and determining why the survey is needed.

Phase 2: *Questionnaire*
The survey questions will be carefully crafted to ensure that they are relevant and clearly written. To collect information and data that will answer the research objectives of the survey.

APRIL 26
MAY 9

Phase 3: *Sampling and Pretests*
defining a sample group based on characteristics and conducting pre-tests to identify and solve problems with the survey before giving it to the larger group.

Phase 4: *Collect the Data*
collecting the survey responses and preprocessing the data to ensure completeness and accuracy and that it's ready for analysis.

Phase 5: *Expectation vs Real*
compare the expected outcomes to the actual results.
To evaluate the effectiveness and ensure that the survey project meets its objectives and provides valuable insights

MAY 10
MAY 20

Phase 6: *Data Analysis*
Analyze the survey data to interpret the survey data and provide insights. using different statistical techniques and tools.

Phase 7: *Report & Presentation*
Communicating the findings of the survey based on the results to be informative and accessible to the audiences.

BUDGET
ETHICS & CONFIDENTIALITY

Incentives for participants:
The survey was made in a Google form so this costed us nothing but to encourage the students to participate we provided a small gift in return.

```
import pandas as pd
```

```
ID = pd.read_csv(r"C:\Users\A\Downloads\Questionnaire - Python , R and Java.csv")
```

```
ID["please, put the last 6 number of your id \nFor random prize drawing"]
```

```
0       494864
1       452375
2       445293
3       367672
4       372760
        ...
267     478752
268    1459861
269     478733
270     461379
271     615417
Name: please, put the last 6 number of your id \nFor random prize drawing, Length: 272, dtype: object
```

```
id = ID["please, put the last 6 number of your id \nFor random prize drawing"]
```

```
id.info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 272 entries, 0 to 271
Series name: please, put the last 6 number of your id
For random prize drawing
Non-Null Count  Dtype
--------------  -----
231 non-null    object
dtypes: object(1)
memory usage: 2.2+ KB
```

```
id = id.dropna()
```

```
id = id.drop_duplicates()
```

```
choosen_id = id.sample(n=3,replace=False)
```

```
choosen_id
```

```
7      449693
209    613228
166    460097
Name: please, put the last 6 number of your id \nFor random prize drawing, dtype: object
```

As we promised the students that we are going to choose people randomly using there ID number in order to give them a prize, and those are 4 students that will get a gift.

ETHICS AND CONFIDENTIALITY

We will explain for the students the purpose of the survey before filling it up and they have the option to accept or refuse. They will assured that their responses will be unspecified and personal information will be stored securely and won't be accessed except by the research members only.

QUESTIONNAIRE

TYPES OF QUESTIONS USED:

Our questionnaire was overall 27 questions. It was a mix between ranking, multiple choice, and checklist questions. We tried to let the questions as simple as possible.

There won't be any personal questions asked and every student will know the purpose of the questionnaire before filling it up.

MAIN SURVEY QUESTIONS:

The main questions helped us to get a feedback from the students about which language from Python, R, and Java do they prefer. After filling the form, the percentage at the end will let us decide which language is mostly preferred by the students and the ease of using it.

SAMPLING PRETEST

INTRODUCTION

In this survey, the population is the students of the general department in the college (Data Science) from the first level to the fourth level, whose ages range from approximately 18-23 years, which studies in Faculty of Computers and Data Science in Alexandria. The survey was sent to about 700 people but was filled by 272 of different levels and ages.

DEFINING THE SAMPLE:

There are some crucial characteristics we need to define for ensuring the survey's results and accuracy:

- Age:

The age range we got was between 18-23 and this helped since knowing more languages differs as being in a different educational level. This helped us to gather different opinions and know how the students thinks of the programming languages.

- Major:

We used checklist questions in order to know which are the students in between Mobile Applications, Cyber Security, AI, Data Science, and Web Development.

- Programming Languages:

We gave the students a chance to let us know the importance of knowing more than one programming language.

COLLECTING THE DATA

INTRODUCTION

The importance of this phase is to obtain a clean dataset that doesn't have any careless responses, missing values and inconsistencies while being well-framed and easy to read. We also need to be sure that this data has no errors.

This will let us review the survey responses and remove any careless answers and missing values that occurs if the student skipped any of the questions. It is important to control the missing values properly to have an accurate and unbiased analysis.
At last, the data set should be organized to ensure that the insights gained from the analysis are clear and relevant to the research questions.

- Data Cleaning:

We used python in order to frame our data.
In order to clean tour we starting renaming the columns to be more clear.

- Data Transformation:

We starting downloading the open-ended questions to change their format to be easier to deal with.

EXPECTATION VS REAL

Here, we started comparing between the expected outputs and the actual results. We made this step in order to know how effective our survey is, and the it met the objectives we made.

First, lets get the the relations between the questions. We will test the "Familiarity_with_Python".

HYPOTHESIS:

- Null Hypothesis:

There is no significant difference in the mean familiarity with Python among the groups of Most_language_proficient_in.

- Alternative Hypothesis:

There is a significant difference in the mean familiarity with Python among the groups of Most_language_proficient_in.

- The P score we got was less than the significance value "0.5", so we will reject the null hypothesis, and accept our alternative hypothesis, which is (that the people who are most proficient in Python, Java, and R have different levels of familiarity with Python).
- To determine which groups are significantly different, we perform Tukey's post hoc test.
- The mean difference between Java and Python that we got was 1.2636 with a p-value of 0.0.

EXPECTATION VS REAL

- This suggests that people who are most proficient in Python have a higher "Familiarity with Python" score than those who are most proficient in Java.

- Similarly, the mean difference between the Python and R groups is -0.7354, with a p-value of 0.0277.

- This means that people who are most proficient in Python also have a higher "Familiarity with Python" score than those who are most proficient in R Language.

- These results confirm that people who are more proficient in python are more likely to have a more "Familiarity With Python" score than people who are more proficient in other languages.

JAVA:

- We will repeat this process with Java to find the familiarity with it.

- We will find that the mean difference between the Java and Python groups is -0.6716, with a p-value of 0.0. This suggests that people who are most proficient in Java have a higher "Familiarity with Java" score than those who are most proficient in Python.

- Similarly, the mean difference between the Java and R groups is -1.1158, with a p-value of 0.0001. This suggests that people who are most proficient in Java also have a higher "Familiarity with Java" score than those who are most proficient in R.

- These results suggest that people who are most proficient in Java are likely to have higher Familiarity_with_Java scores than those who are most proficient in other languages.

EXPECTATION VS REAL

R:

- We will repeat this process with R to find the familiarity with it.

- the table indicates that there is a significant difference in means between the Java and Python groups (reject = True), and between the Java and R groups (reject = True). However, there is no significant difference between the Python and R groups (reject = False)

- The mean difference between the Java and Python groups is 0.5114, with a p-value of 0.0017. This suggests that people who are most proficient in Python have a higher Familiarity_with_R score than those who are most proficient in Java

- Similarly, the mean difference between the Java and R groups is 1.057, with a p-value of 0.0004. This suggests that people who are most proficient in R also have a higher Familiarity_with_R score than those who are most proficient in Java

- Proficient in Python (more than java) -> familiar with R.
- Proficient in R (more than java) -> familiar with R.
- We can get an insight that people proficient in Java aren't so familiar with R as people who are proficient in Python.

BUT it doesn't indicate that people proficient in R should have a higher "Familiarity with R" than people who are most proficient in Python .

One possible reason is because the first year student did not learn python, which affected the survey results.

EXPECTATION VS REAL

Rules we can infer from these observations:-

- People who are proficient with Python are likely to be familiar with Python more or equal (java/r) and if not, there's a high chance there responses to be careless.
-
- People who are proficient with java are likely to be familiar with java more or equal (python/r) and if not, there's a high chance their responses to be careless.
-
- Note: Not the same goes for R like we concluded above

What we're going to do to find and exclude all Careless Responses:-

- We add a new column "mis" that consists 0 values, so we can use it later to know the careless responses.
- Then we converted our inferred rules into a code to find all the careless responses.
- We found that the careless responses we had were 29 responses.
- We then deleted then and got 223 responses.
- Next step, we started deleting the useless columns.
- At the end, we run the Cronbach's alpha test on the dependent scale columns and got only 4 columns.

EXPECTATION VS REAL

WHY DIDN'T WE DO CRONBACH'S ALPHA?

The data we have isn't normal.

WHY DIDN'T WE DO "AVERAGE INTER-ITEM CORRELATION" AND "AVERAGE ITEM-TOTAL CORRELATION?

We need to have more than one ranking question that is divided with the same concept.
For example, having about 5 questions that tells me how much does the student prefer using python by a scale from 1 to 5.

DISCRIMINANT VALIDITY:

We have many questions that has no relations between them so the correlation in that case is so low which is approximately zero.

EXPECTATION VS REAL

```
Student_age: correlation coefficient = 0.92
Student_gender: correlation coefficient = 1.00
Helps_the_most: correlation coefficient = 0.77
High_in_demand_language: correlation coefficient = 0.79
Prefer_to_use_familiar_language: correlation coefficient = 0.67
Preferred_language_for_projects: correlation coefficient = 0.68
Consider_trends_when_choosing_a_language: correlation coefficient = 0.66
Most_language_proficient_in: correlation coefficient = 0.57
Consider_requirements_when_choosing_a_language: correlation coefficient = 0.46
Best_community_support_and_resources: correlation coefficient = 0.39
Familiarity_with_Python: correlation coefficient = 0.68
Familiarity_with_R: correlation coefficient = 0.69
Familiarity_with_Java: correlation coefficient = 0.66
How_easy_learning_Python: correlation coefficient = 0.48
How_easy_learning_R: correlation coefficient = 0.50
How_easy_learning_Java: correlation coefficient = 0.35
Importance_degree_of_knowing_multiple_languages: correlation coefficient = -0.34
Do_language_features_influence_you: correlation coefficient = 0.65
Consider_requirements_when_choosing_a_language: correlation coefficient = 0.46
Mean correlation coefficient: 0.58
```

The image above gives me the results of the correlation. It clearly shows that the ranking questions have a very low correlation and most probably the reason is these questions are affected by how the students think while answering these questions. Any of the students in solving first time may think that he's better in a specific language and the other time while solving a specific language and than starts getting errors and is not able to solve it than refills the survey and chooses another language that he prefers depending on this situation.

ANALYSIS

INTRODUCTION

In this phase, we will pass through several steps to ensure that we drew accurate and relevant conclusions from the survey data.

1) DESCRIPTIVE STATISTICS:

	Student_level	Familiarity_with_Python	Familiarity_with_R	Familiarity_with_Java
count	223.000000	223.000000	223.000000	223.000000
mean	2.147982	3.573991	2.650224	3.349776
std	0.849262	1.274365	1.128496	1.104286
min	1.000000	1.000000	1.000000	1.000000
25%	2.000000	3.000000	2.000000	3.000000
50%	2.000000	4.000000	3.000000	3.000000
75%	2.000000	5.000000	3.000000	4.000000
max	4.000000	5.000000	5.000000	5.000000

Over here, we found all the statistics we can get for the familiarity of the students with each language.

Importance_degree_of_knowing_multiple_languages	Do_language_features_influence_you
223.000000	223.000000
3.627803	3.816143
1.069873	0.863321
1.000000	1.000000
3.000000	3.000000
4.000000	4.000000
4.500000	4.000000
5.000000	5.000000

ANALYSIS

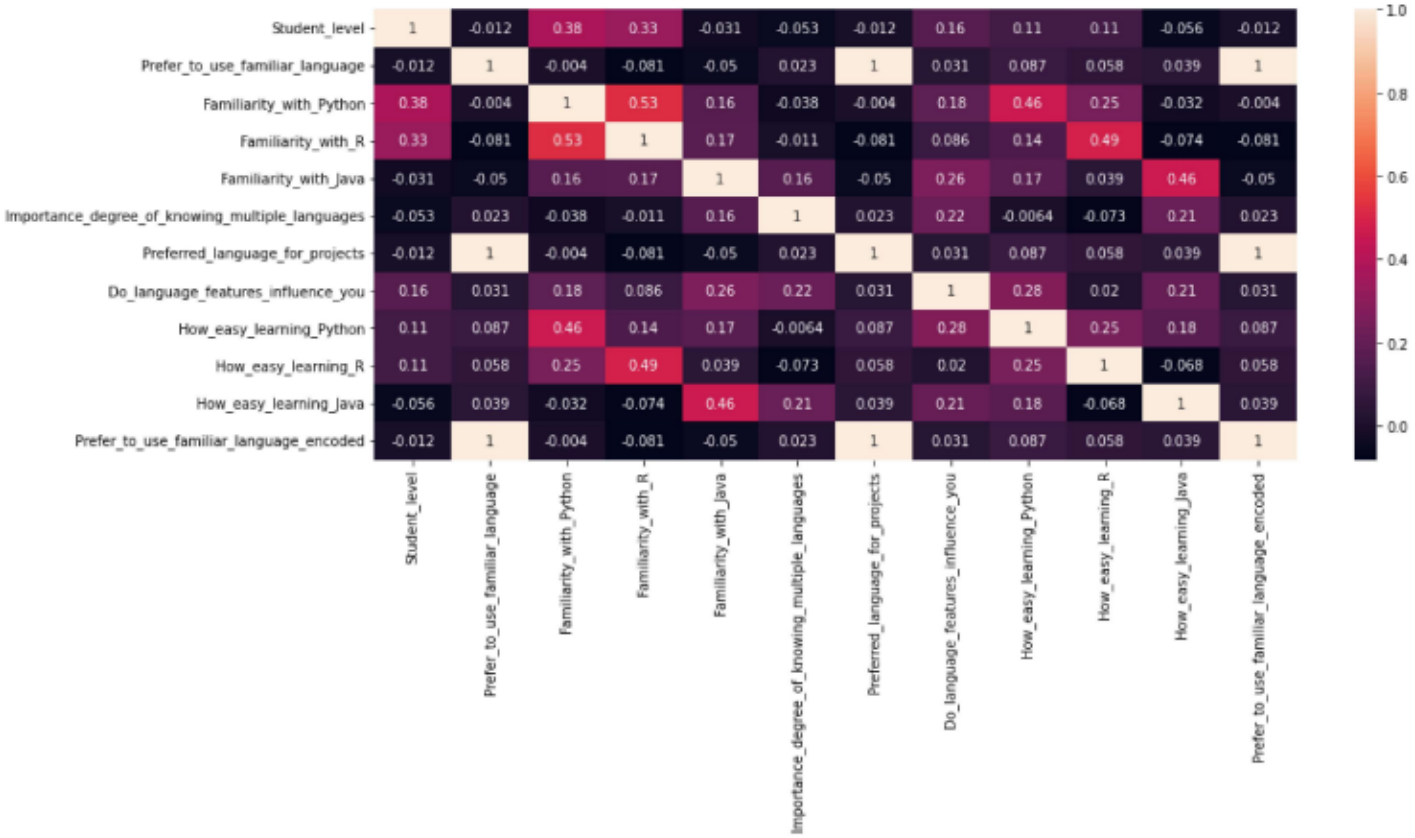
In the previous picture, we found the also the measures for how important is it to know more than one programming language and how does the features on each language influence you.

How_easy_learning_Python	How_easy_learning_R	How_easy_learning_Java
223.000000	223.000000	223.000000
3.860987	3.147982	3.089686
1.010509	1.082449	1.031500
1.000000	1.000000	1.000000
3.000000	2.000000	2.000000
4.000000	3.000000	3.000000
5.000000	4.000000	4.000000
5.000000	5.000000	5.000000

The last part, were the measures of how much is it easy learning each of the languages.

ANALYSIS

2) CORRELATION ANALYSIS:



This correlation matrix shows me the relations between the columns and as the color becomes lighter this means that the correlations increases.

3) REGRESSION:

```
Accuracy Score on Training Data: 95.51%
R-squared score Training Data: 95.51%
Accuracy Score on Test Data: 66.67%
R-squared score Test Data: 66.67%
```

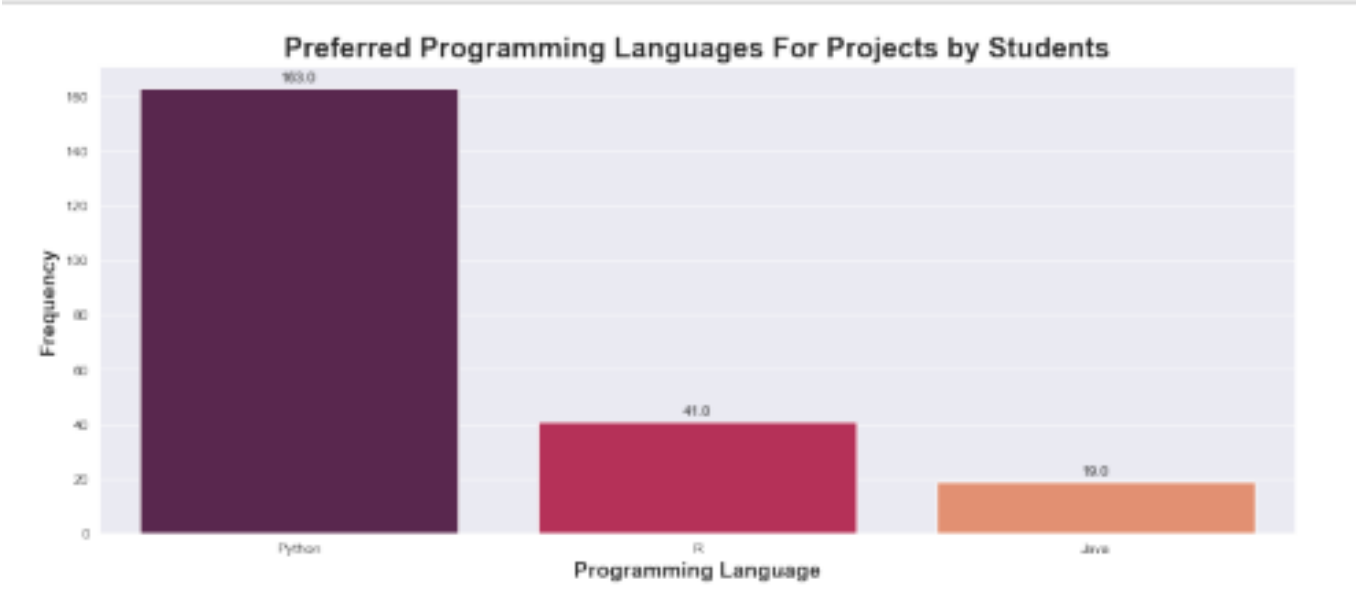
After we used the logistic regression, the accuracy percentage and R-square score were "95.51%" for the training data. For the test data, the accuracy and the R-squared were "66.67%".

ANALYSIS

```
T-statistic: 41.8805316325393
P-value: 2.3294438721448886e-107
Confidence Interval: (3.4067290089548203, 3.7412530538254485)
```

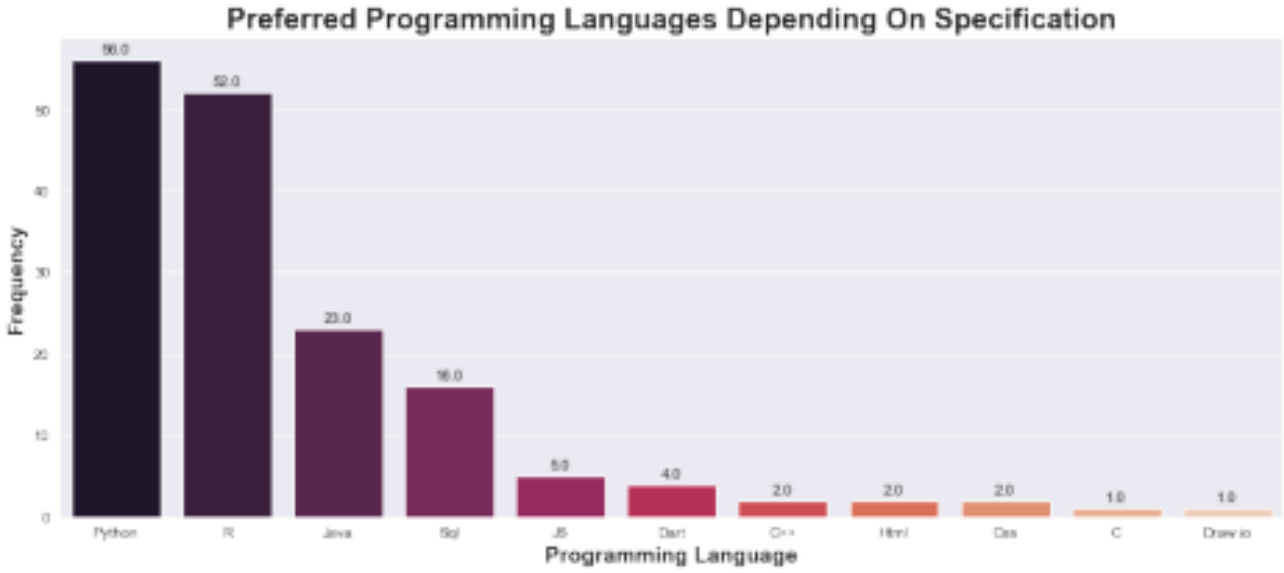
Last part in the logistic regression, we found that the test is 41.88 and the P-value is 2.329e-187 and the range of confidence interval I got is between 3.4067 and 3.7413.

4) DATA VISUALIZATION:



This bar plot gives me the frequency of which programming language does the students prefer the most and as we can see that it's "Python".

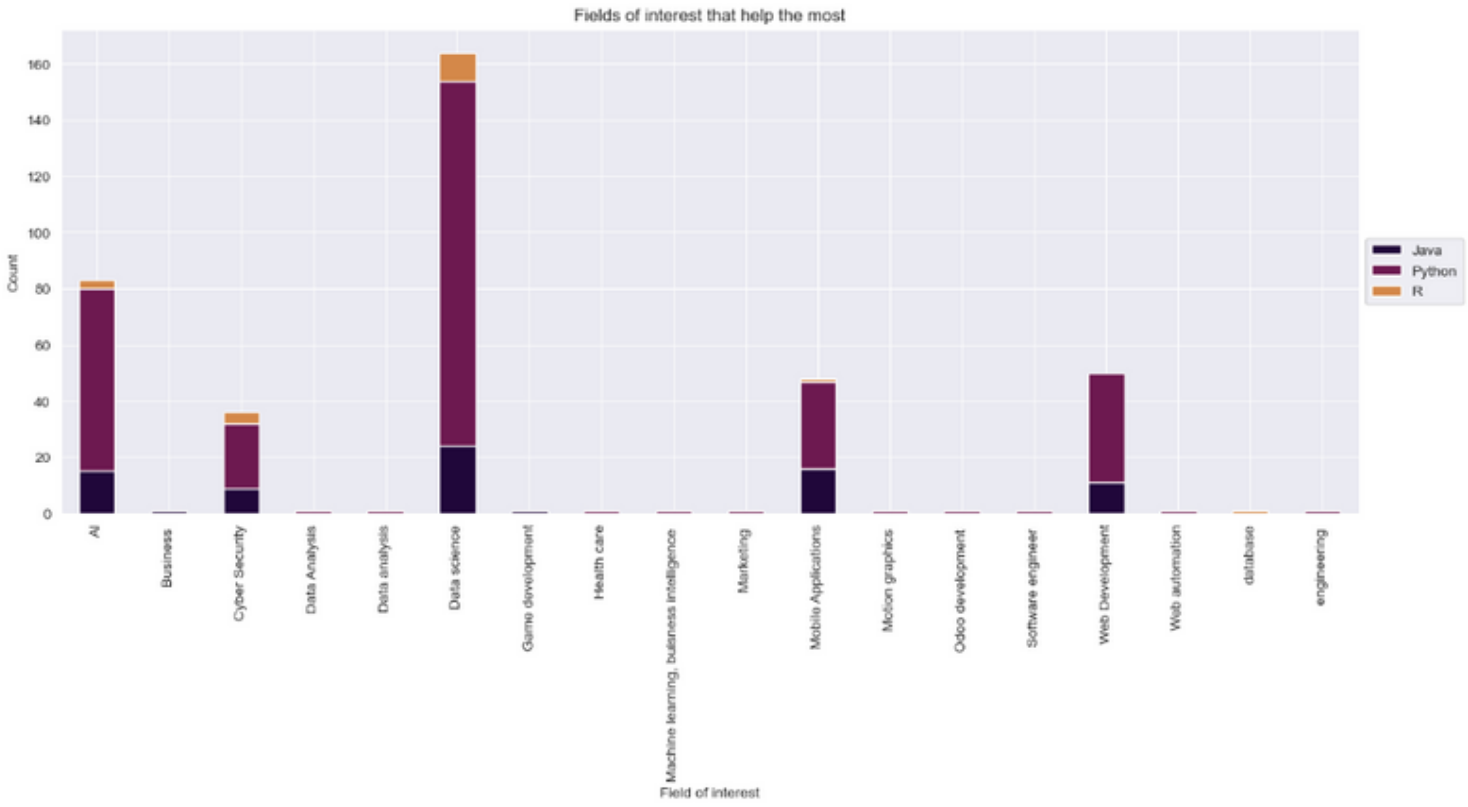
ANALYSIS



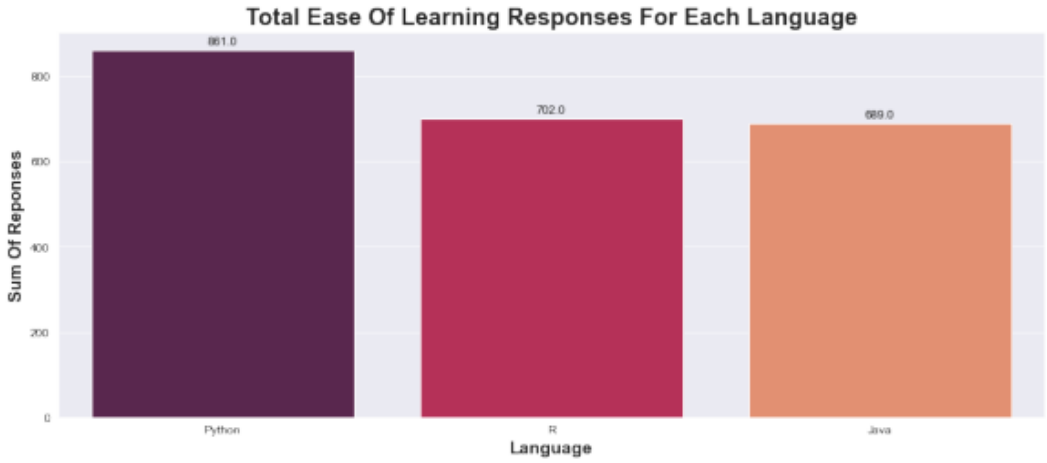
This bar plot gives me the frequency of the preferred programing language between the student depending on their specifications.

Preferred_language_for_projects	Java	Python	R
Language_used			
C	0	1	0
C++	0	2	0
Dart	0	4	0
Draw.io	0	1	0
Html;Css;JS	0	1	0
Html;Css;JS;Sql	0	0	1
JS	0	3	0
Java	8	9	5
Python	1	46	7
Python;Java	0	1	0
R	1	37	13
R;Python	0	1	0
Sql	0	8	7

ANALYSIS

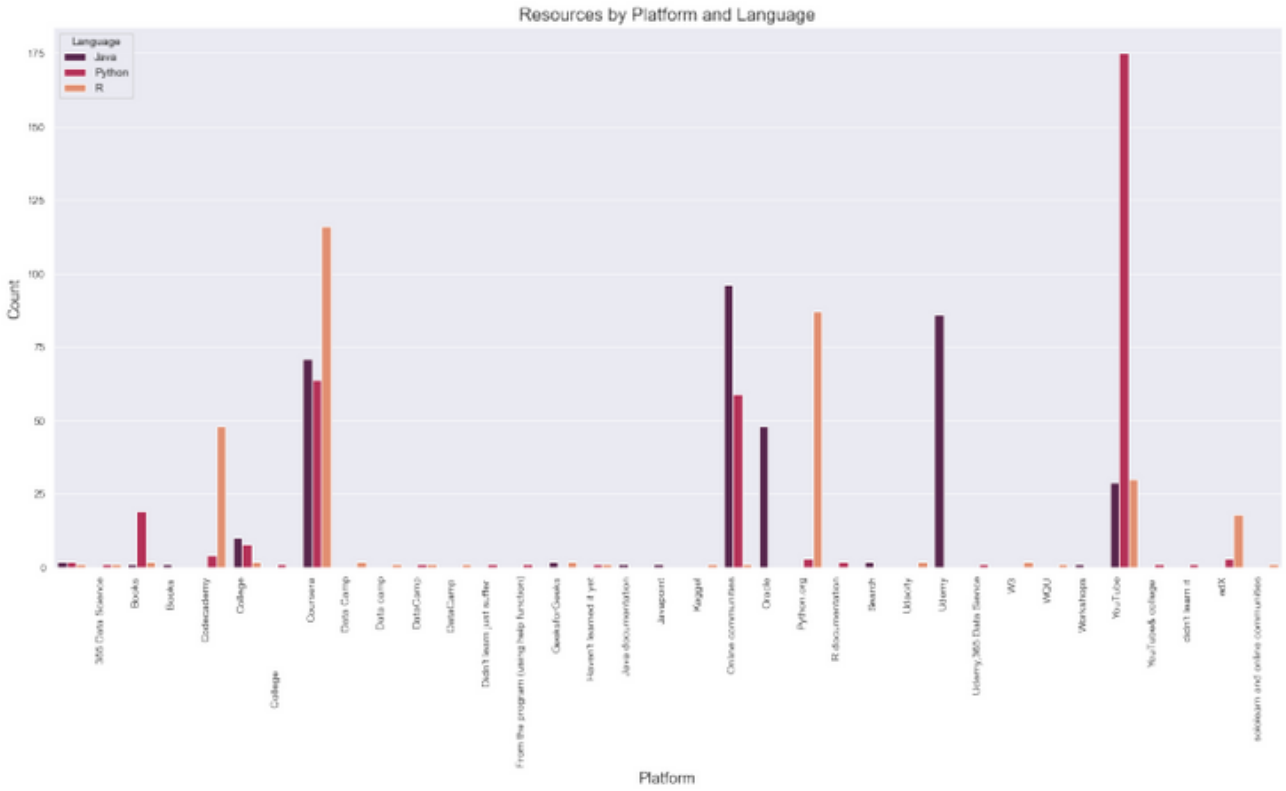


This output tells me which between Python, Java, and R helps them the most when dealing which each of those fields.

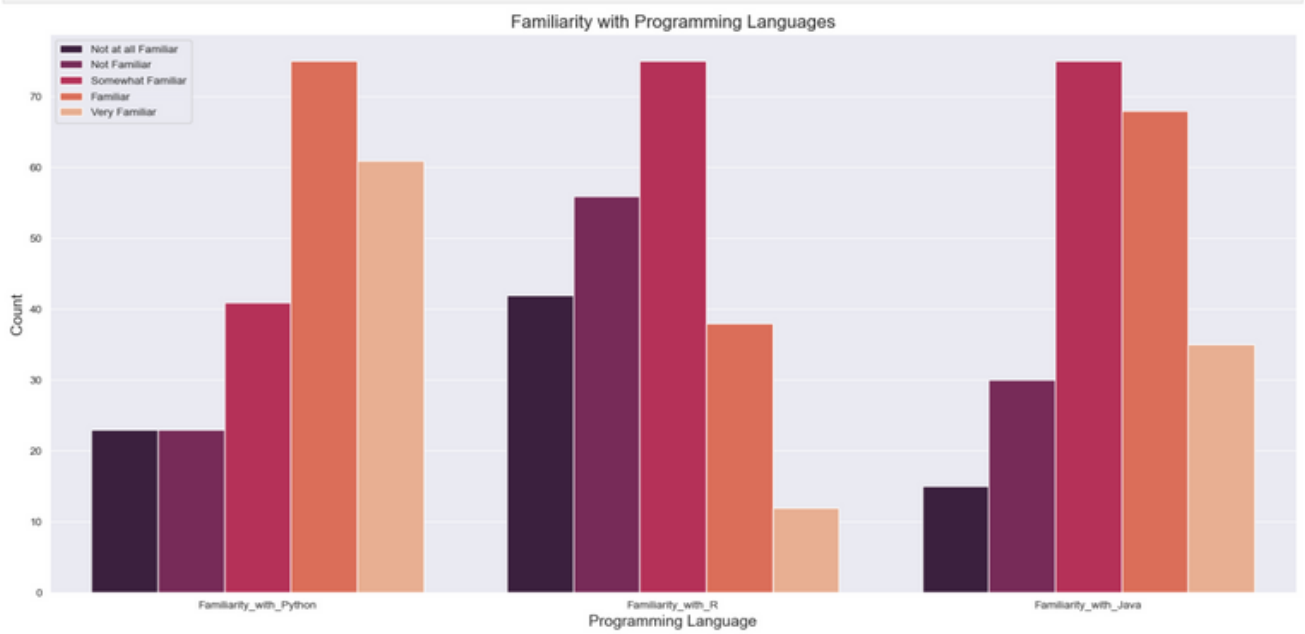


This bar plot shows me that the learning "Python" is easier than learning "R" and that "R" is easier than learning "Java".

ANALYSIS



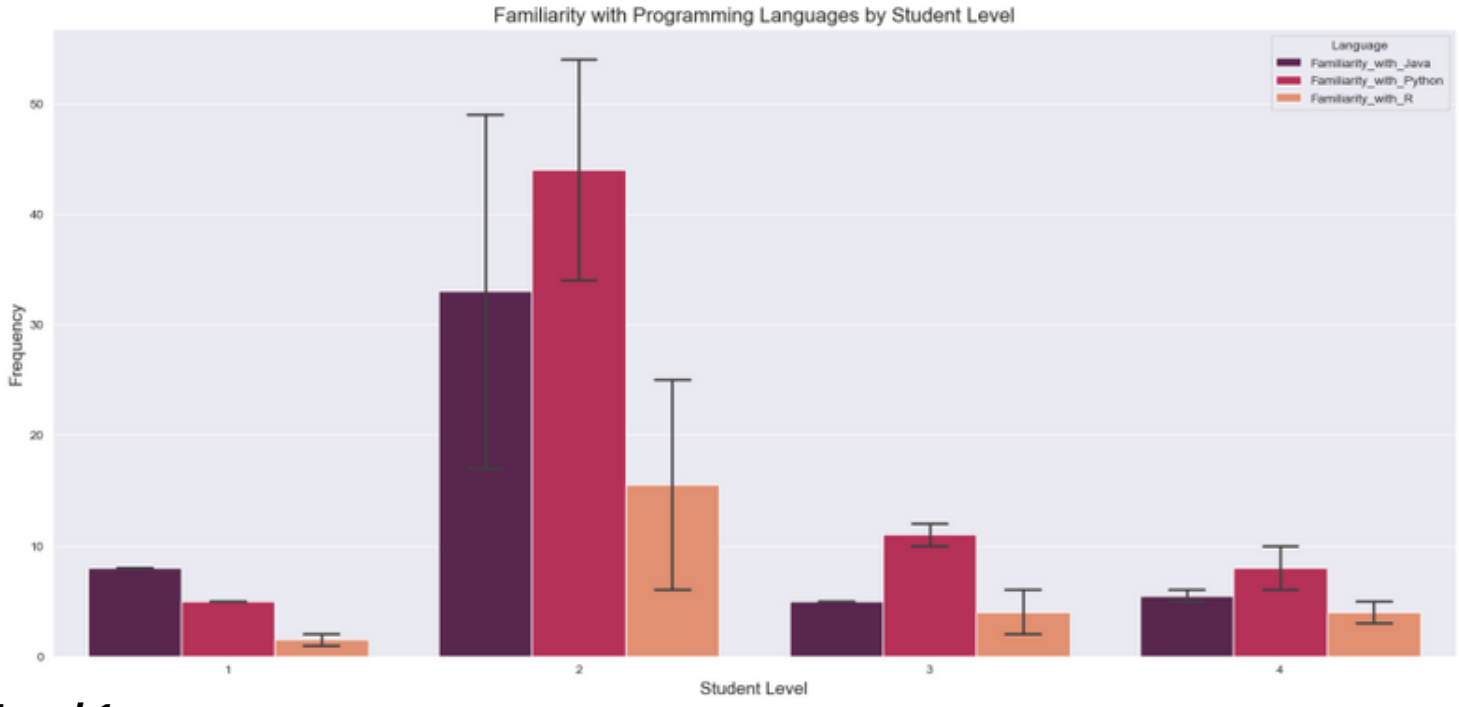
Finally, this plot shows me the resources preferred to learn those programming languages from.



Python and Java seem to be more popular among the students than R. Python and Java have a similar distribution of scores, with the majority of students reporting moderate to high levels of familiarity.

ANALYSIS

R, on the other hand, has a different distribution of scores, with the majority of students reporting a moderate level of familiarity followed by a relatively low level of familiarity, and a significant number of students reporting no familiarity with R.
Overall, it appears that the students are more familiar with Python and Java than R.



Level 1:

- Java is the most common language, this is because they have not
- studied any other language.
- Some students have gained significant familiarity with Python despite it
- not being studied at this level.

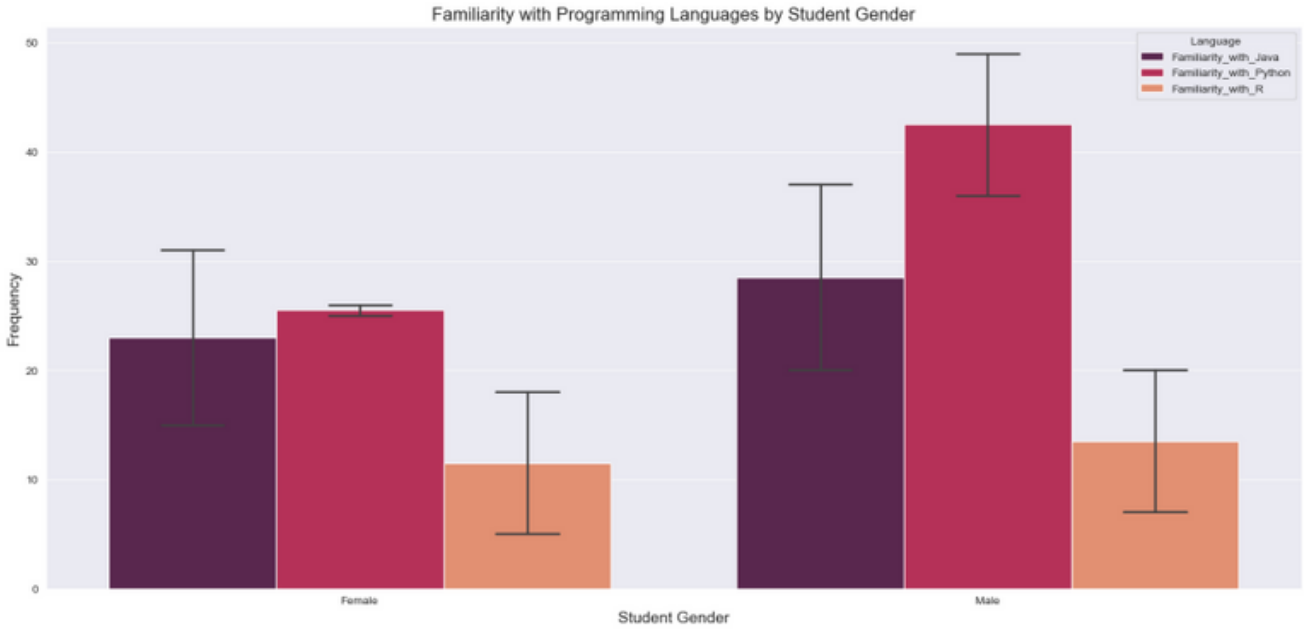
Level 2:

Python & Java are the most common languages, it means that these languages are gaining more prominence and use in the curriculum at this level.

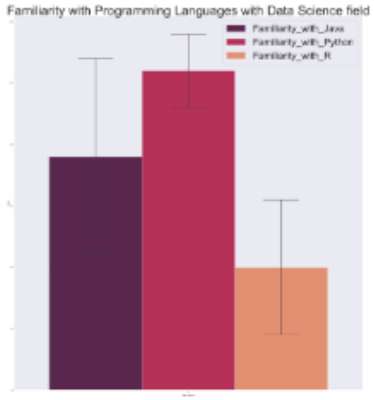
Level 3 & 4:

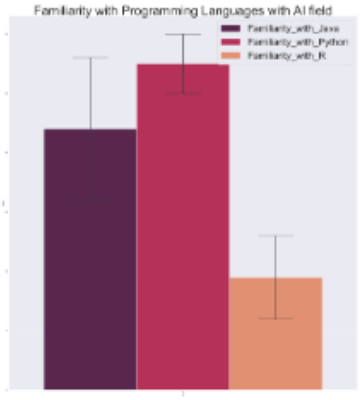
- Python is the most common language.
- Overall, it appears that the students are more familiar with Python and Java than R.
- The fact that Python has relatively high counts even at lower levels, possibly due to its usefulness in data science and other analytical fields, while R may be more specialized and taught to students in specific courses.

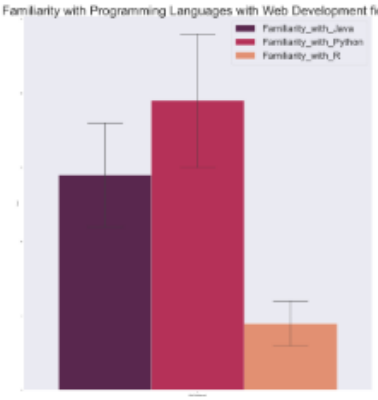
ANALYSIS

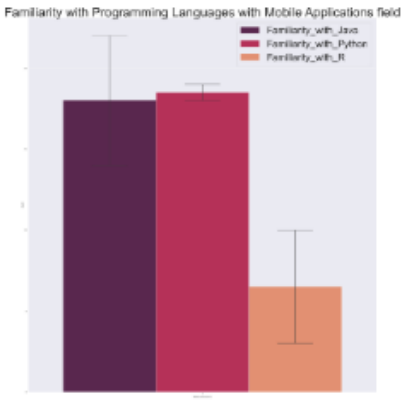


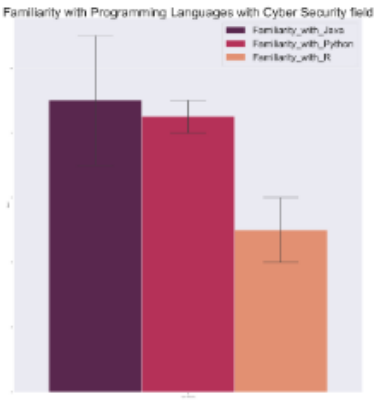
Those results tells me that males prefer using "Python" while women prefer using "Java" and that both prefer using "Python" and "java" more than "R".









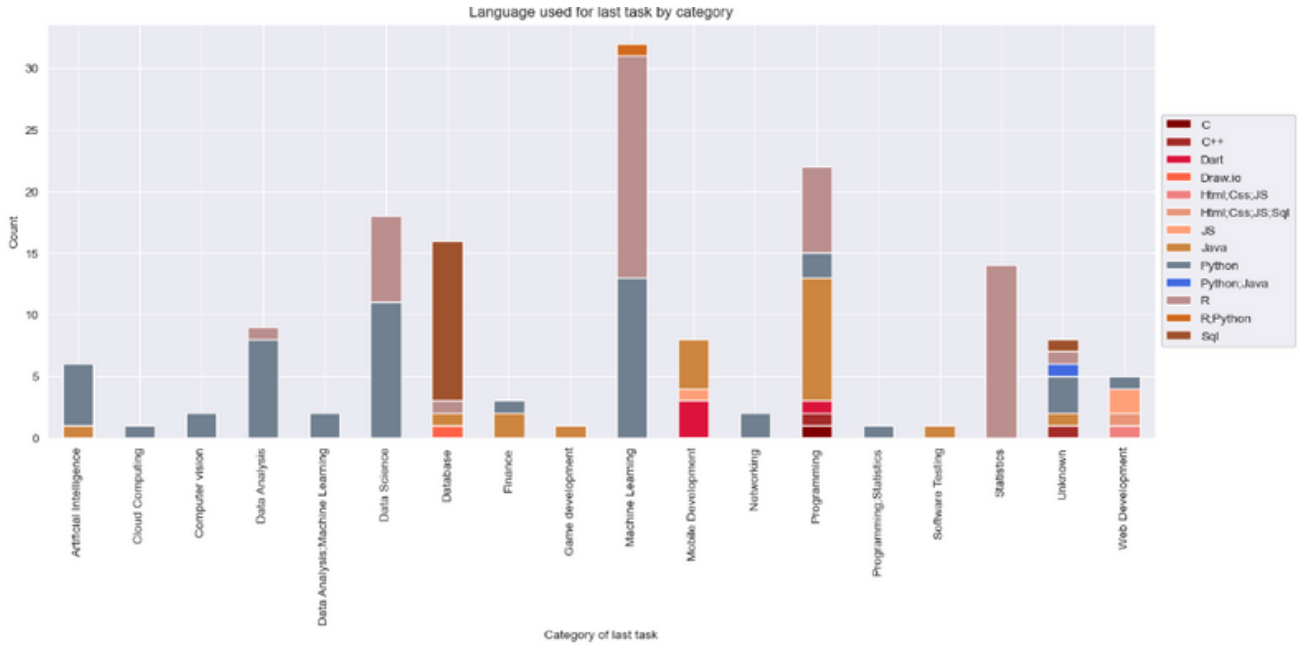


ANALYSIS

These 5 plots describes the familiarity of each "Python", "Java", and "R" with all the fields I have.

The feedback I got from those plots is that:

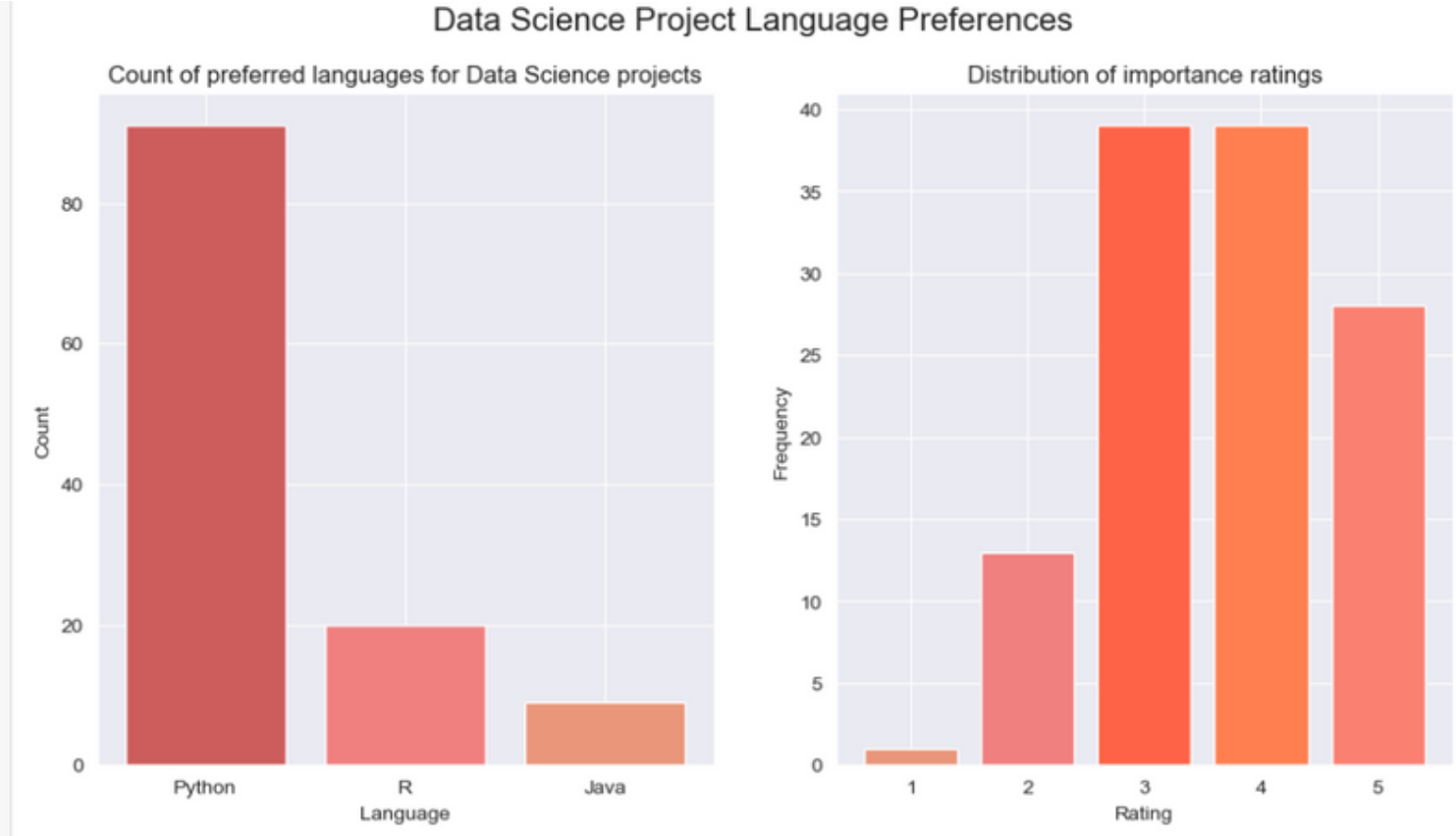
- Python is most familiar with students interested in Data Science, AI and Web Development fields.
- Java is most familiar with students interested in Mobile Applications and Cyber Security fields.
- R is slightly more familiar in Cyber Security field compared to other fields.
- Overall, Python and Java seem to be the most familiar programming languages among the students, while R is the least familiar.



- After we made analysis for the "Last_task_worked_on" and "Category_of_last_task_worked_on" columns to see the types of tasks or projects that the students have worked on in the past , we found that:
- Python is used in multiple fields like Artificial intelligence ,Data science ,Machine learning , computer vision ,cloud computing , Networking , statistics
- From the above , We can conclude that Python is used in various Fields

ANALYSIS

- **R** is used in statistics , machine learning , Data science , Data Analysis
- From the above, we also conclude that most of uses of R are in statistics and machine learning
- **Java** is used in Programming , Mobile development ,Game development , Finance , Artificial intelligence, Software testing
- Although there are many languages used in programming such as C and C++, but Java is the most used because we took our sample from our college which teaching a programming course using Java
- **SQL** is mainly used in database
- SQL appeared greatly in students' responses because a large number of respondents are currently studying Database
- Depending on the level and different studies, other languages appeared, such as **Dart** which used in Mobile Programing
- And also **HTML , JS and CSS** are used in Web development



- In data science, Python is the most preferred language and students find the need to learn multiple languages is (4) of importance.

KRUSKAL-WALLIS

```
Kruskal-Wallis test for Familiarity_with_Python:
Test statistic: 26.475291070287565
P-value: 7.583778506425359e-06
There is a significant difference in Familiarity_with_Python between at least two groups.

Kruskal-Wallis test for Familiarity_with_R:
Test statistic: 23.107419442251295
P-value: 3.8353647864454084e-05
There is a significant difference in Familiarity_with_R between at least two groups.

Kruskal-Wallis test for Familiarity_with_Java:
Test statistic: 2.037250407303049
P-value: 0.5647113512131181
There is no significant difference in Familiarity_with_Java between any groups.
```

- We used Kruskal-Wallis test for finding the p-value of "Familiarity_with_Python" and "Familiarity_with_R" and "Familiarity_with_Java"

	1	2	3	4
1	1.000000	0.000144	0.000160	0.000338
2	0.000144	1.000000	0.872095	1.000000
3	0.000160	0.872095	1.000000	1.000000
4	0.000338	1.000000	1.000000	1.000000

- Level of education has an impact on the familiarity with Python and R among students.

	1	2	3	4
1	1.000000	0.000228	0.052961	0.000113
2	0.000228	1.000000	1.000000	0.552537
3	0.052961	1.000000	1.000000	0.743875
4	0.000113	0.552537	0.743875	1.000000

- Level of education may not have a significant impact on the familiarity with Java among students.

TEST OF HYPOTHESIS

```
preferred_language = df['Preferred_language_for_projects']
contingency_table = pd.crosstab(preferred_language, df['Familiarity_with_Python'])
chi2, p_value, dof, expected = chi2_contingency(contingency_table)
# Set significance level
alpha = 0.05
# Interpret the results
if p_value < alpha:
    print("Reject the null hypothesis. Python is the most preferred language.")
else:
    print("Fail to reject the null hypothesis. Insufficient evidence to conclude that Python is the most preferred language.")
```

```
Reject the null hypothesis. Python is the most preferred language.
```

After applying hypothesis test with null hypothesis that python isn't the most preferred
language for projects and alternative hypothesis that python is the mostpreferred language for projects on the column of "Preferred_language_for_projects " , we rejected the null hypothesis and we found that python is the most preferred languages for projects.

```
student_age = df['Student_age']

# Create a contingency table
contingency_table = pd.crosstab(student_age, preferred_language)

# Perform chi-square test of independence
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

# Set significance level
alpha = 0.05

# Interpret the results
if p_value < alpha:
    print("Reject the null hypothesis. There is a correlation between student age and preferred language.")
else:
    print("Fail to reject the null hypothesis.Insufficient that there is a correlation between student age and preferred language
```

```
Reject the null hypothesis. There is a correlation between student age and preferred language.
```

After applying hypothesis test with null hypothesis that there isn't a correlation between student_age and the preferred language for projects and alternative hypothesis that there is a correlation between student_age and the preferred language for projects, we rejected the null hypothesis and found that there is a correlation between student_age and the preferred language for projects.

TEST OF HYPOTHESIS

```
student_gender = df["Student_gender"]
# Create a contingency table
contingency_table = pd.crosstab(index=student_gender, columns=preferred_language)

# Perform the chi-square test
chi2, p_value, dof, expected = chi2_contingency(contingency_table)
# Set the significance level
alpha = 0.05
# Compare the p-value with the significance level
if p_value < alpha:
    print("Reject the null hypothesis. There is a correlation between student gender and preferred language.")
else:
    print("Fail to reject the null hypothesis. There is no significant correlation between student gender and preferred language.
```

```
Fail to reject the null hypothesis. There is no significant correlation between student gender and preferred language.
```

After applying hypothesis test with null hypothesis that there isn't a correlation between student_gender and the preferred language for projects and alternative hypothesis that there is a correlation between student_gender and the preferred language for projects, we failed to reject the null hypothesis and found that there isn't a correlation between student_gender and the preferred language for projects.

CONCLUSION

CONCLUSION

In summary, the proposed survey to explore data science students' experience with Python, R, and Java programming languages is a well-designed approach to gain valuable insights.

The survey can help students make informed decisions about which language to learn based on their interests and specific data science tasks or projects.
And also it is critical in promoting better educational outcomes and preparing students for the job market in data science. As it can provide insights to educational institutions to improve their curricula.

Ultimately, this survey can help improve the field of data science by giving insights into what students need and what their experiences are.
This will better equip students with the necessary skills and knowledge to succeed in their future careers.