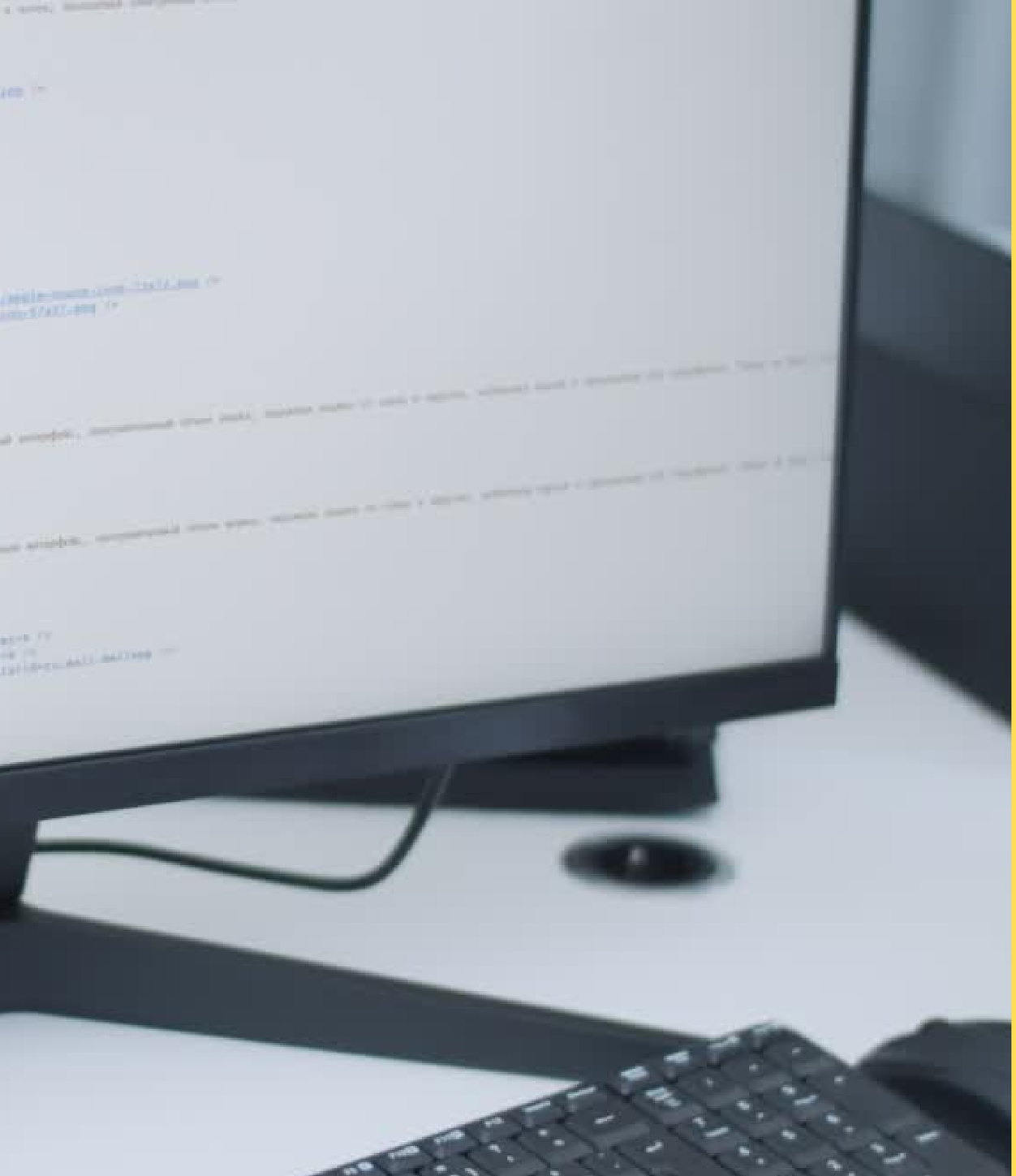


Anabelle Capois Espinal

KaggleX BIPOC Mentorship Program

FINAL PROJECT SHOWCASE

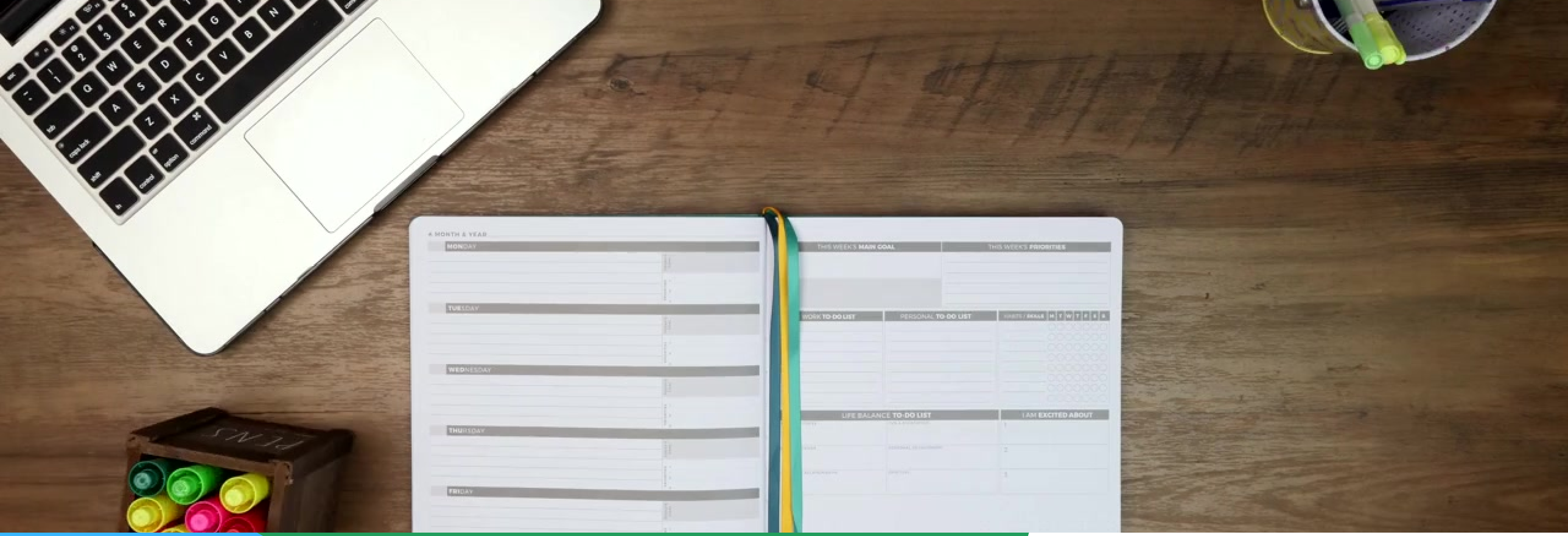
kaggle

MY BACKGROUND

- AFRO-LATINA (DOMINICAN REPUBLIC)
  - FIRST GENERATION
- NYC IS HOME
- EDUCATION:
  - BACHELOR IN BUSINESS & ECONOMICS, GERMAN
  - MBA IN STRATEGY & INTERNATIONAL BUSINESS
- ASPIRING DATA ANALYST
  - TRANSITIONING FROM THE NON-PROFIT SECTOR

JOURNEY INTO DATA

- **FEBRUARY - MAY**
  - TECH TALENT & STRATEGY DATA SCIENCE BOOTCAMP CERTIFICATE
- **MARCH - JULY**
  - GOOGLE X MLT TECH PREP CAREER COACHING PROGRAM
- **AUGUST - OCTOBER**
  - KAGGLEX BIPOC MENTORSHIP PROGRAM
- **SEPTEMBER**
  - ANITAB.ORG GRACE HOPPER CONFERENCE BOOTCAMP SCHOLAR
    - 1 YR MEMBERSHIP
- **OCTOBER - DECEMBER**
  - CODEFIRST GIRLS PYTHON COURSE
- **NOVEMBER**
  - AFROTECH
    - SPONSORED BY TECH EQUITY COLLECTIVE AS TECH PREP ALUMNA



ABOUT THE PROJECT

GETTING THE DATA

ISAAC WEN · UPDATED 2 YEARS AGO

47 | New Notebook | Download (41 kB)

GitHub Programming Languages Data

Statistics for Programming Languages used on GitHub

Data Card | Code (10) | Discussion (1)

About Dataset

Context

A common question for those new and familiar to computer science and software engineering is what is the most best and/or most popular programming language. It is very difficult to give a definitive answer, as there are a seemingly indefinite number of metrics that can define the 'best' or 'most popular' programming language.

One such metric that can be used to define a 'popular' programming language is the number of projects and files that are made using that programming language. As GitHub is the most popular public collaboration and file-sharing platform, analyzing the languages that are used for repositories, PRs, and issues on GitHub and be a good indicator for the popularity of a language.

Content

This dataset contains statistics about the programming languages used for repositories, PRs, and issues on GitHub. The data is from 2011 to 2021.

Source

This data was queried and aggregated from BigQuery's public github_repos and githubarchive datasets.

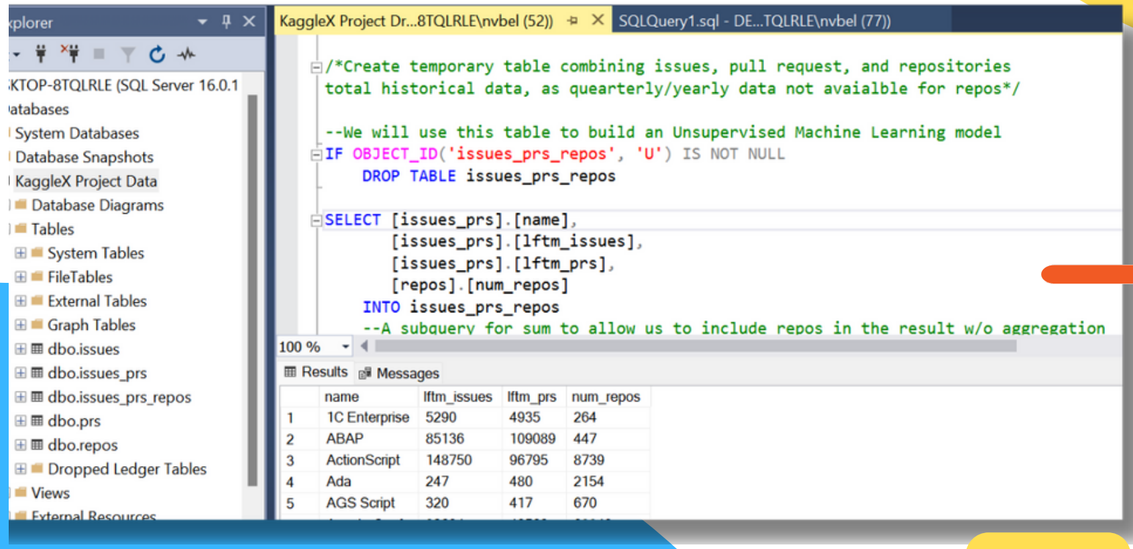
Usability 10.00

License Attribution 4.0 International (CC ...

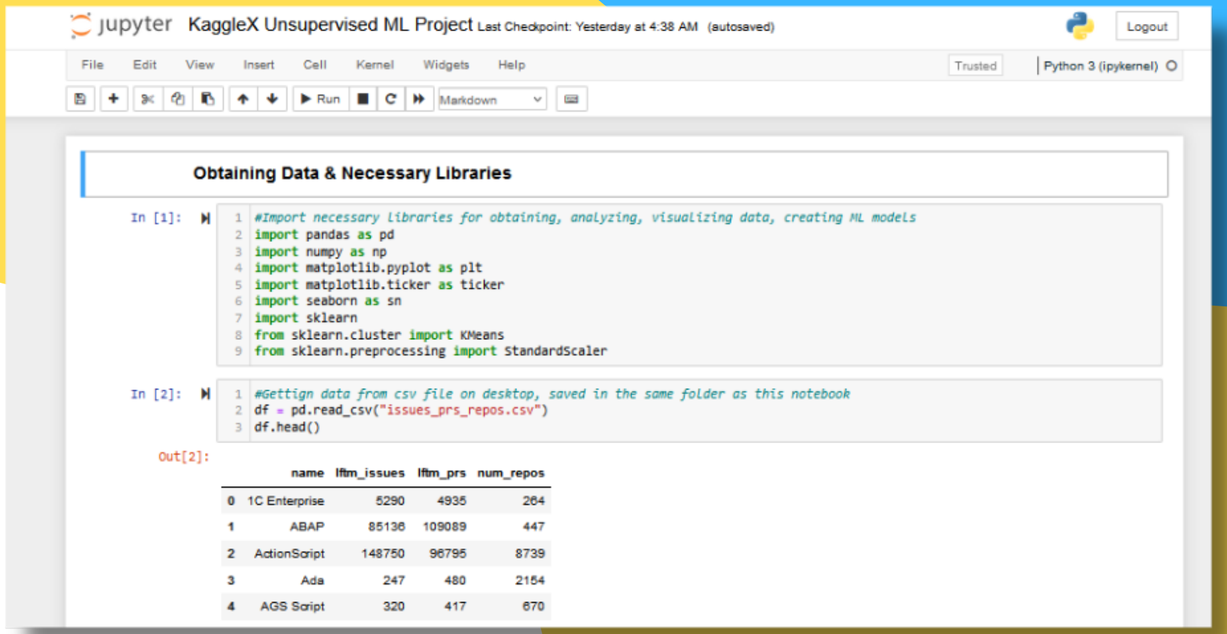
Expected update frequency Annually

Tags Computer Science | Software | Programming | Python | Bigquery

KAGGLE



SQL



PYTHON

UNDERSTANDING THE DATA

DESCRIPTIVE STATISTICS

	Issues	Pull Requests	Repositories
count	158.000	158.000	158.000
mean	8216556.544	10155053.709	48137.177
std	31919571.481	39013309.545	150351.314
min	100.000	107.000	7.000
25%	6523.500	5300.250	848.000
50%	118605.000	114340.500	4232.500
75%	1159712.250	1369640.500	16846.250
max	279509718.000	323026578.000	1100421.000

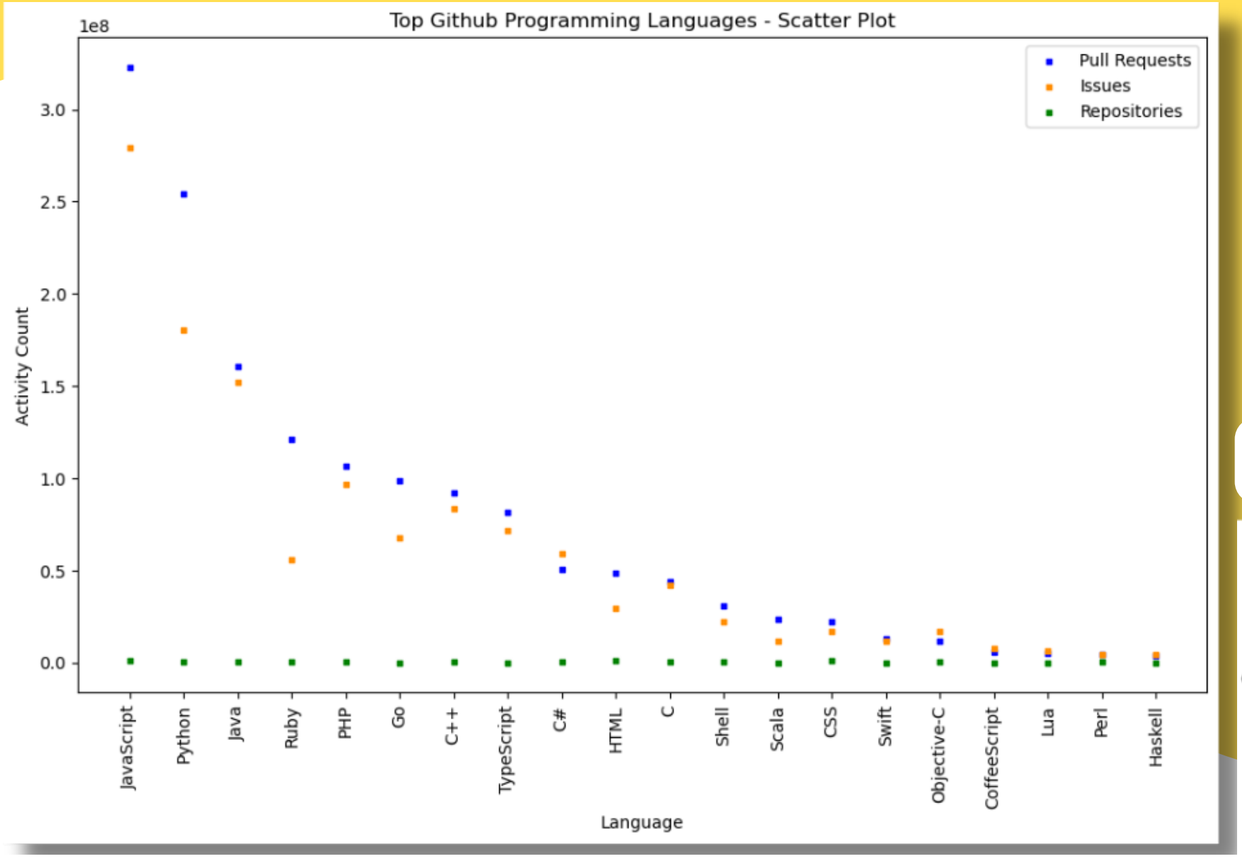
	Pull Requests	Issues	Repositories
count	20.000	20.000	20.000
mean	75070355.800	60922248.800	314200.800
std	86468292.363	71121886.130	311673.052
min	3831778.000	4105280.000	29898.000
25%	12730829.500	11451322.000	63207.750
50%	46285958.000	35883729.000	223036.500
75%	100461920.000	74419777.750	418319.000
max	323026578.000	279509718.000	1100421.000

FINAL DATA

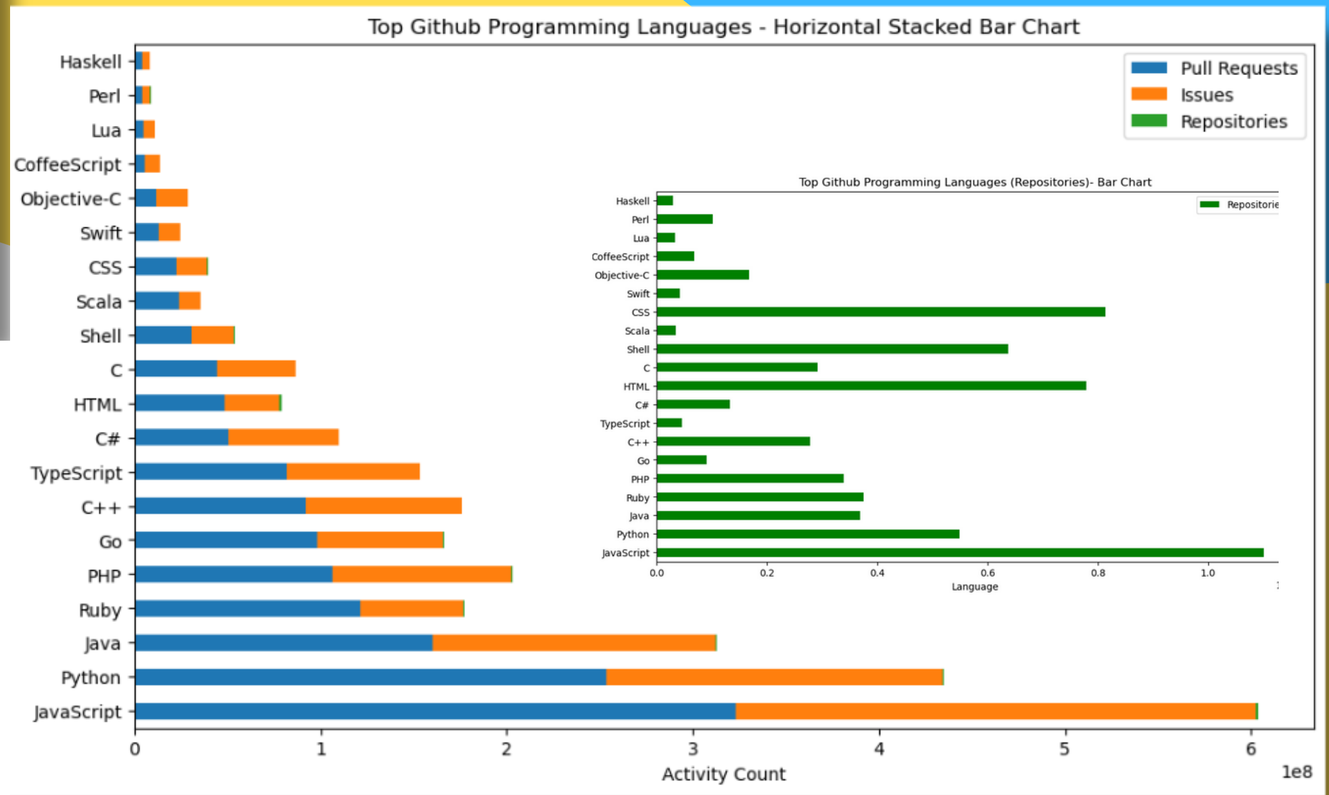
TOP 20 LANGUAGES ACROSS ALL THREE CATEGORIES

Language	Pull Requests	Issues	Repositories
JavaScript	323026578.0	279509718.0	1100421.0
Python	253883739.0	180276683.0	548870.0
Java	160463616.0	152051010.0	369282.0
Ruby	121277814.0	55582961.0	374802.0
PHP	106413734.0	96345155.0	339901.0
Go	98477982.0	67459678.0	91119.0
C++	92274560.0	83463172.0	278066.0
TypeScript	81755312.0	71405313.0	46332.0
C#	50816884.0	59130078.0	133013.0
HTML	48391480.0	29639106.0	779549.0
C	44180436.0	42128352.0	292000.0
Shell	30818060.0	22385664.0	638068.0
Scala	23737196.0	11485782.0	34501.0
CSS	22424786.0	16635720.0	813443.0
Swift	13062315.0	11347942.0	42372.0
Objective-C	11736373.0	16621072.0	168007.0
CoffeeScript	5697280.0	7929520.0	68833.0
Lua	4735302.0	6499600.0	34089.0
Perl	4401891.0	4105280.0	101450.0
Haskell	3831778.0	4443170.0	29898.0

SCATTER PLOT



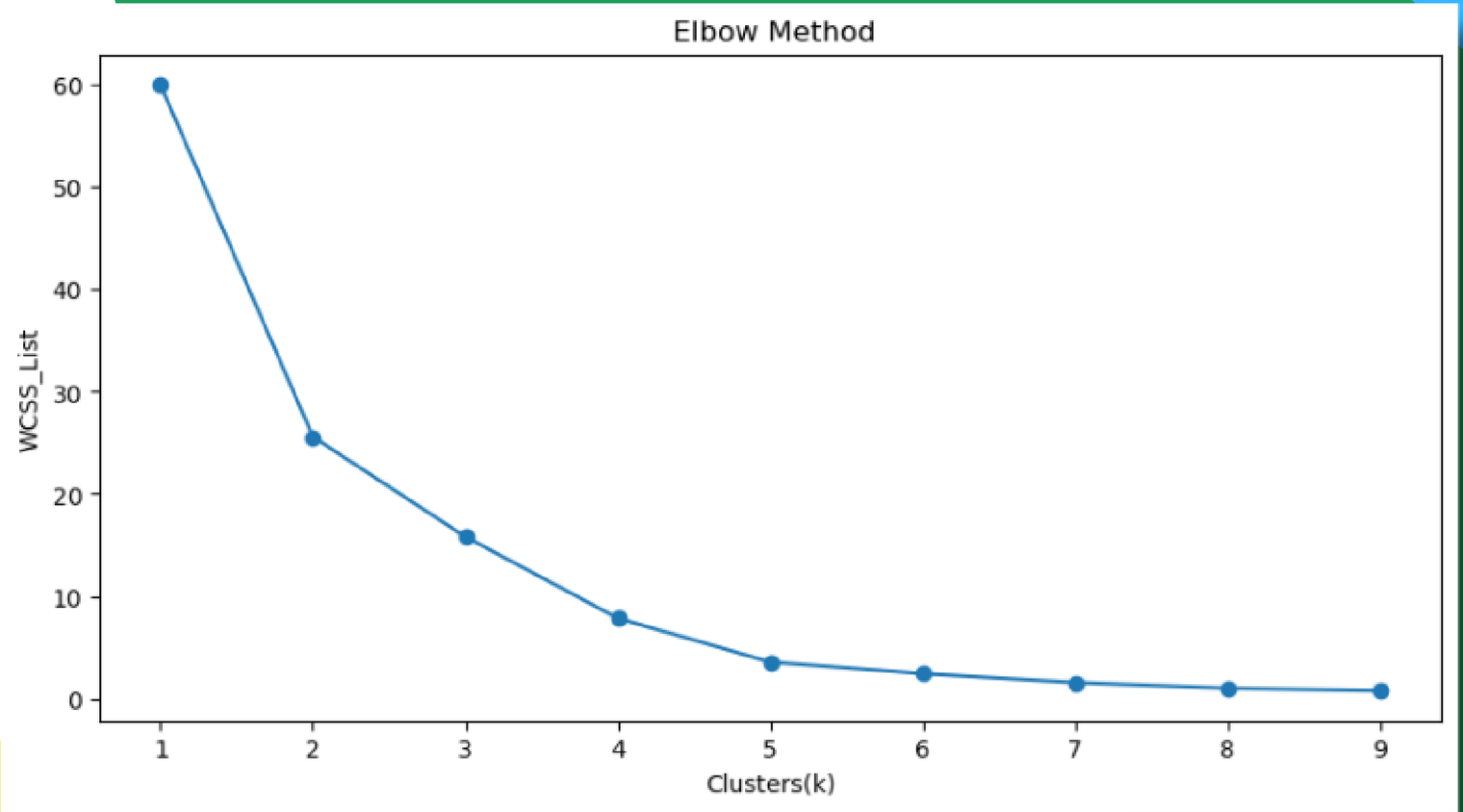
STACKED BAR CHART



VISUALIZATION

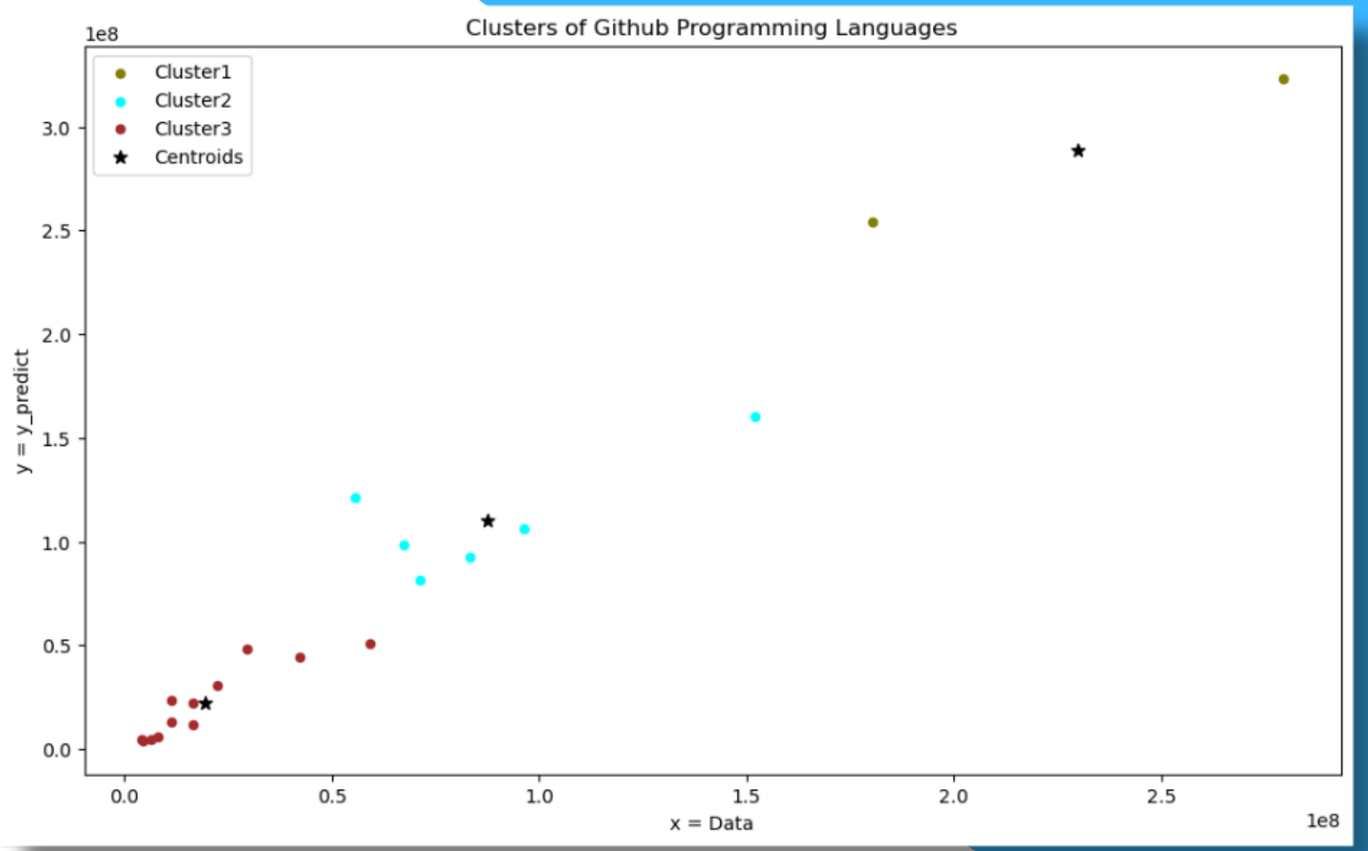
UNSUPERVISED MACHINE LEARNING

USING THE ELBOW METHOD FOR KMEANS CLUSTERING



CLUSTERS DATA & VISUALIZATION

Language	Pull Requests	Issues	Repositories	T_Issues	T_Pull Requests	T_Repositories	KMeans_3
JavaScript	323026578.0	279509718.0	1100421.0	3.153263	2.942093	2.588112	1
Python	253883739.0	180276683.0	548870.0	1.721763	2.121687	0.772494	1
Java	160463616.0	152051010.0	369282.0	1.314590	1.013223	0.181319	1
Ruby	121277814.0	55582961.0	374802.0	-0.077023	0.548269	0.199490	2
PHP	106413734.0	96345155.0	339901.0	0.510998	0.371901	0.084601	2
Go	98477982.0	67459678.0	91119.0	0.094307	0.277740	-0.734350	2
C++	92274560.0	83463172.0	278066.0	0.325167	0.204134	-0.118950	2
TypeScript	81755312.0	71405313.0	46332.0	0.151225	0.079319	-0.881782	2
C#	50816884.0	59130078.0	133013.0	-0.025853	-0.287776	-0.596442	2
HTML	48391480.0	29639106.0	779549.0	-0.451279	-0.316555	1.531853	0
C	44180436.0	42128352.0	292000.0	-0.271114	-0.366520	-0.073082	2
Shell	30818060.0	22385664.0	638068.0	-0.555915	-0.525070	1.066120	0
Scala	23737196.0	11485782.0	34501.0	-0.713152	-0.609087	-0.920727	2
CSS	22424786.0	16635720.0	813443.0	-0.638861	-0.624659	1.643426	0
Swift	13062315.0	11347942.0	42372.0	-0.715141	-0.735748	-0.894817	2
Objective-C	11736373.0	16621072.0	168007.0	-0.639073	-0.751481	-0.481247	2
CoffeeScript	5697280.0	7929520.0	68833.0	-0.764454	-0.823137	-0.807712	2
Lua	4735302.0	6499600.0	34089.0	-0.785081	-0.834551	-0.922084	2
Perl	4401891.0	4105280.0	101450.0	-0.819621	-0.838508	-0.700342	2
Haskell	3831778.0	4443170.0	29898.0	-0.814747	-0.845272	-0.935880	2



PROJECT SUMMARY

PERSONAL PROJECT GOAL:

IMPROVE SQL, PYTHON, MACHINE LEARNING KNOWLEDGE AND SKILLS

PROCESS:

2011-2021 GITHUB DATA OBTAINED FROM KAGGLE HAS BEEN CLEANED AND RESTRUCTURED USING SQL AND PYTHON. THE NEW DATA IS THEN USED TO VISUALIZE AND CREATE MACHINE LEARNING MODEL.

RESULTS:

- LIST OF THE TOP 20 LANGUAGES FOR GITHUB ACTIVITIES
- CREATED STACKED BAR CHART TO VISUALIZE DATA
- UNSUPERVISED MACHINE LEARNING METHOD OF K-MEANS CLUSTERING UTILIZED
- OPTIMAL CLUSTER COUNT IDENTIFIED WITH ELBOW METHOD
- THREE CLUSTERS USED, AS ANY CAN BE CHOSEN PAST THE OPTIMAL OF 2 GIVEN BY THE ELBOW METHOD GRAPH
- PYTHON AND JAVASCRIPT HAD THEIR OWN CLUSTER WITH UNPARALLELED ACTIVITY COUNT. FOR DATA ANALYSTS, PYTHON WOULD BE THE LANGUAGE TO FOCUS ON.
    - AFTER IMPROVING PYTHON SKILLS, BASED ON THIS DATA, MY NEXT FOCUS SHOULD BE RUBY AND THEN HTML

DATA SCIENCE CONCEPTS APPLIED:

- STATISTICS
- DATA STRUCTURE
- DATA CLEANING
- DATA ANALYSIS
- ALGORITHMS
- MACHINE LEARNING
- UNSUPERVISED ML

LEARNINGS AND TAKEAWAYS

- START SIMPLE, AND BUILD THE COMPLEXITY OF YOUR PROJECT AS TIME ALLOWS
- FIRST IDENTIFY THE DATA, THEN DETERMINE WHAT CAN BE DONE WITH THE DATA
  - THIS DATA DID NOT FIT WHAT I ORIGINALLY WANTED TO ACHIEVE FROM THE PROJECT
- IT'S OKAY TO PIVOT AND ADJUST YOUR PROJECT AS YOU COME ACROSS LIMITATIONS
- DURING YOUR LEARNING JOURNEY, YOU CAN ALWAYS COME BACK TO A PROJECT AFTER "COMPLETION" AND CONTINUE TO BUILD ON IT AS YOU BUILD YOUR EXPERTISE AND KNOWLEDGE
- THERE ARE MULTIPLE WAYS IN WHICH YOU CAN WRITE YOUR CODE AND GET THE SAME RESULT. YOUR GOAL IS TO ACHIEVE AS MUCH SIMPLICITY AS POSSIBLE FOR THE CODE TO BE EASIER FOR OTHERS TO READ AND UNDERSTAND

RESOURCES

PYPL POPULARITY OF PROGRAMMING LANGUAGE INDEX
HTTPS://PYPL.GITHUB.IO/PYPL.HTML?COUNTRY=

GITHUB PROGRAMMING LANGUAGES DATA (KAGGLE.COM)
HTTPS://WWW.KAGGLE.COM/DATASETS/ISAACWEN/GITHUB-PROGRAMMING-LANGUAGES-DATA?SELECT=REPOS.CSV

IMPORTING CSV FILE INTO SSMS #SQL #SQLSERVER #SSMS - YOUTUBE
HTTPS://WWW.YOUTUBE.COM/WATCH?V=K5_U6XRBL_S

SQL FOR DATA ANALYSIS. FOR THOSE WHO ARE STARTING THEIR... | BY ALEX SOUZA | BLOG DO ZOUZA | MEDIUM
HTTPS://MEDIUM.COM/BLOG-DO-ZOUZA/SQL-FOR-DATA-ANALYSIS-E8D0356ECD3C

PLOT MULTIPLE COLUMNS OF PANDAS DATAFRAME ON BAR CHART WITH MATPLOTLIB - GEEKSFORGEEKS
HTTPS://WWW.GEEKSFORGEEKS.ORG/PLOT-MULTIPLE-COLUMNS-OF-PANDAS-DATAFRAME-ON-BAR-CHART-WITH-MATPLOTLIB/

K-MEANS CLUSTERING MODEL IN 6 STEPS WITH PYTHON | BY SAMET GIRGIN | PURSUITOFDATA | MEDIUM
HTTPS://MEDIUM.COM/PURSUITNOTES/K-MEANS-CLUSTERING-MODEL-IN-6-STEPS-WITH-PYTHON-35B532CFA8AD

STATQUEST: K-MEANS CLUSTERING - YOUTUBE
HTTPS://WWW.YOUTUBE.COM/WATCH?V=4B5D3MUPQMA

K MEANS CLUSTERING | #DATA MINING | USING CSV FILE - YOUTUBE
HTTPS://WWW.YOUTUBE.COM/WATCH?V=UM6IZ6JJCPG

K-MEANS CLUSTERING ALGORITHM WITH PYTHON TUTORIAL - YOUTUBE
HTTPS://WWW.YOUTUBE.COM/WATCH?V=INLZ3IU5FFW

HOW TO PERFORM K MEANS CLUSTERING IN PYTHON(STEP BY STEP) - YOUTUBE
HTTPS://WWW.YOUTUBE.COM/WATCH?V=FLOPH6UQDIW

kaggle

HTTPS://WWW.LINKEDIN.COM/IN/ANABELLECAPOIS/

GitHub

HTTPS://GITHUB.COM/ANABELLECE

HTTPS://WWW.LINKEDIN.COM/IN/ANABELLECAPOIS/