ASSIGNMENT - JOB MARKET ANALYSIS

1. PROJECT AIM

Analysis of job market data to explore the open positions related to the jobs in the field of data or machine learning in a particular region. While it is not compulsory to do so, you could use web scraping techniques introduced in this course to collect data from these websites. Also, note that you are not limited to the sites mentioned above. Identity one job that you aspire to transition into and identify the skills required for the job.

A. Objectives & Assignment Description:

- i. Collect and use job market data to explore data related positions or machine learning related positions in a particular region (e.g., US, India).
- ii. You will have to use appropriate keywords (for e.g., "data scientist", "data engineer", "ML engineer") to extract information on the positions available on websites of your choice.

2. PROCEDURES:

A. Data Collection Method

The data collection involves collecting the data from any one of the websites like Indeed, Glassdoor, Naukri, Monster. We also must use one keyword to extract the required jobs from website. For this assignment I choose the following.

- Website Indeed
- Keyword Machine Learning
- i. Robot text file from indeed does not block /jobs and /viewjob URLs for user agent
 *. We also inspected the page to find corresponding html tags, classes and other attributes to extract the data.
- ii. **Data discovery**: I have selected the indeed website for data extraction. Challenges faced are the following. The website uses Cloudflare for protection against DDoS mitigation, a Web Application Firewall, API protection, etc. But this also make sure we do not get the data by using simple requests library. I have also tries using special headers. But this was also not successful. Hence, I have

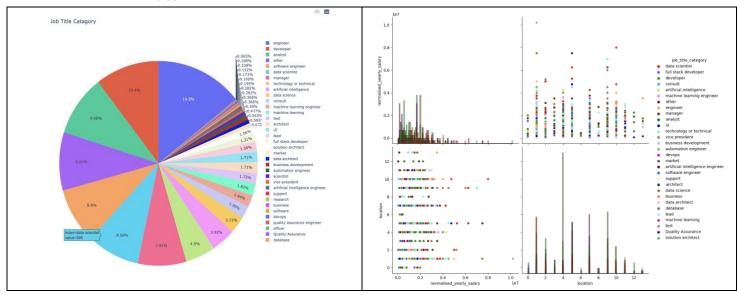
used selenium library to open the webpage using Chrome driver from selenium library. Each webpage opening takes around little more than 3 seconds to complete. Sometimes more than 3 seconds I have also used beautiful soup library as I found it easier to navigate the complex html page. I also found ways to do pagination and access pages of data. This makes up the discovering phase of the Data collection. This solution arises another issue of time taken for extraction of data. must python library concurrent use (https://docs.python.org/3/library/concurrent.futures.html) to achieve parallel execution of certain function to extract more data in less amount of time. For getting details about each job, we need to extract /viewjob page for each job entry in jobs page (index page with list of jobs). Pagination should also be done using start and limit parameter to extract the data. We also had filters like location, programming language, job type for extracting information in detail. The website shows a bug where we are not able to access the list beyond 1000. This issue is solved by using more filter to reduce the total result set.

- iii. **Data structuring**: For each job we extracted the details like, title, time of posting, company, location, metadata (includes job type, salary range and other special details), partial job description and link to full job details page. From individual page, we have extracted the programming languages and full description of the job.
- iv. **Data Cleaning**: Here we encoded the data for job type, programming languages into Boolean form where we have True (1) for presence and False (0) for absence. We have also removed the duplicates and aggregated the columns for the duplicate assuming same jobs have same title, company, date posted, company rating, full description. For presence column it was a simple Boolean and (sum) and for metadata it was concatenation. We also had to extract data again in case we were getting invalid description.
- v. Enriching Data: Enriching the data is done by using the using nltk to extract major programming language tokens. We also did the same for job type from metadata. We have full-time, contract, fresher, internship, part-time, temporary, permanent, freelance as job types and these categories can be overlapping. We also extracted salary from metadata and separate it into maximum, minimum and normalized salary along with intended. Here normalized salary means salary calculated for entire year. We also had to create category for jobs as job title can be very similar but different when be compare. We have identified major clusters using K-means for the title and used this as a reference to create our major category of jobs.

								_	_	_																							
	[60_100e	company	lecation	date	job_desc	mode ning	₩	Table Python	Ampèr 0	AT MAY NO	Hiberrate	Shell 1 Scripting	Nesis Nede je Besis	C RS	rierium Boris	trap Go React	HTMLS Part SQL	best and 200.	origit Co+ Cor	tract Fresher	inamatip fast- tine	Temporary	Description	sales, strip a	nenge_salary	anyan may	day saley_bepare	e normalised_peats_takey	/ Personni 1	Presiance	processes_jab_title	job_litte_category	description_clar
2248	M. Digher II	Turing	Begalan, Sambala	PodedPoded 10 days ago	Sold understanding of mustine learning sigori	Notice 688	htps://k.hded.com/c/SV /mClasCHSSSCAss.	to to	Table 15	be Table Table Table	Table	Table	False False	No. Tox	Table 1	ne law face	Table Table True	Table Table	Table Table	Talan Talan	Table Table	False	(rip Alous Turnigh Based in Pala Alba, Califor	Test	MI	ten	tel tel	1 16	i Talor	Tales	machine learning engineer is	mathine learning originals	turing lawed pais also saliforn furing world
803	Smir Baires, Iralyd	COVECE	Begiliru Gerataka	ProtesPrested 25 days ago	Build and deploy machine learning modes where	fultine 480	https://kindedoors/s/SK jkut/Sel/State/SL	tor tor	No 1	be Table Table Table	Table	False	False False	No tue	false 1	she false false	false false. The	Table Table	Tue Tabe	Talon Talon	Table Table	Table ⁴	About CHROD' (vir. Al CHROD), we are on a risk	Net	Set	Sell	No.	1 160	K False	False	business analyst	analyst	orindox coindox mission lea change bowed dec.
784	Data Scientist	Durf Analytics	Bengsiuni, Garrotaka	EngioyerActive S days ago	triticinis, porvellite seponaltie for deci	Springs of the Springs of the Color of the C	https://k.indexd.com/company/Curl- Analytics/	tu tu	Table 10	be false false false	Taba	False	False False	Table Table	False	also falso falso	Non-Table Table	Tabe Taba	Talue False	Talan Talan	Table Table	False	Company Name Curl Technologism Surgainmin	Description of Contract of Con	90000	NA 1602	36.00 per	w 90006.00	2 No	False	data scientist	data scientist	conpany name curl tech location bangalors webs
1876	Full Stack Developer	Dut Analytics	Sergituru. Karrataka	PodledPodled 10 days ago	Currients or developing Products & Solutions	PURSO-PURSON 470 yesful-time 470	110x3h:hdwdcorpi(55) j=8140x:78428	201 201	Tue N	te film film film	Tue	Table	False False	Table Table	Table 1	ine fine. The	No No Te	Tyle Tue	Tue false	Table Table	Table Table	Table	Company Name: Curl (H) Cocation: Bangation (H)	15,00,000 - 125,00,000 v year	100000.00	ADDRESS 2007	10.00 per	1000000	Table	Table	ful stack developer	ful stack developer	company name curl boats bangators website
1087	ara Science - ASAL with Big data Registed	Anlage Infords	Begilas Sambio	Employerácti w 29 days ago	As a Solution Architect, the candidate will be	Up to MEDICOTE IN proof of the CTS	htps://k.inded.com/companyik/age- htbs://	tu tu	No. 1	te fate fate fate	Taba	Talor	Tabe Tabe	to to	Table 1	ine fine fine	No False True	Table True	tue tue	Talan Talan	Table Table	False	As a Solution inchitect, the candidate will be	Spin-Michigan per	5000000.00	Net 5000	1000 per	r 500000.00	i No	Talse	data science artificial intelligence mactine L.	data architect	adulior architect candido responsible driv
10337	Data Scientist - Technology	Anlay Hitach	Depley, Smiths	Poste/hosed 30+ days ago	Operior in at least one of the blowing fig	164m 470	htps://h.hdwd.con/r/SSF /s=40c8803c686C	201 - 201	No. 1	te film film film	Table	Table	False False	No Tox	Table 1	ine Non Title	No No Ton	Table Tole	Tox False	Talon Talon	Table Table	False	(r) frieldes and responsibilities) Digitors (ar.,	Net	Set	Net	NAT NA	i 166	K Tabe	Tabe	dascintractoray	technology or technol	role responsibility explore larg defeats our
2903	MA Developer (Autoration Anystern)	The Automation Office	Begalani, Garrataka	EngloyerActive 3 days ago	As an end-to-end automation solidates partner,	POSCHE-POSCHER AND partial-transfer.	htps:/hinkedoorsjongang/NO-fre- Autmati.	te te	Table To	be False False False	Table	False	False False	Toe Non	False I	ate false false	false false. The	Table Title	Talce Talon	False False	Table Table	False	Ribric Proces Autoration Developin) (Autorat	#150,000 - #10,00,000 v year	9600000	7000000 10000	2000 per	900000	i lite	Table 1	a developer automation anywhere	developer	robotic process automatic developer automatic.
2904	MADevisor (IPVI) TAX	The Automation Office	Despites Carotako	EngloyerActive 3 (Bys ago	Diperience - I to 5 Year*/yWe are looking/b	75,0000-50,0000 4 yes/si trefferia. 450	Hpc/h.hdeconjonpay/50-7e- Admai.	74 74	Table To	be Table Table Table	Table	Table	False False	Tue Non	Table 1	ine fine fine	No to the	Tyle Tue	Note False	Table Table	Table Table	False	Organisation - TAC ALTONATION(s/Position - Oteo	P5,00,000- 100,00,000 e year	1900000	ACCUSED 2000	1000 per	THERE	Table	Table	rpa developer siyath	· ·	organisation too automatic position developer.
3433	Senior M. Engineer	Symphonyki	Bergaluru. Kamataka	ProtecProted 30s days ago	Synghoryal Summit is un ki- powered if and lime	fulfilms 450	https://kindedcom/s/dist ju-ktots/s/s/s/s.	34 34	Non 10	be Table Table Table	Table	Table	False False	Tie Tie	false	ine fine fine	Not false. The	rube Tue	tue tue	Talse Talse	Table Table	False	ly Introduction: (v)n Symphonyall Summit is —	Net	Set	161	No.	i ne	ri Fabe	False	nactice barries explane	mathine learning engineer	introduction symphonyal summ alpowered enter
5097	Serior Clata Scientist	Scienaptic Systems	Despites, Constato	PostedPosted 30+ days ago	An a Senior Cala Scientist you will play a pri_	745m 430	htps://h.hdwd.conjn/SST ja-654759655be5.	201 201	Table To	te false false false	Table	Table	False False	Table Table	Table 1	ine time time	No rate Tox	Table Table	Talox Talox	Talon Talon	Table Table	False 1	70111111111111111111111111111111111111	Net	MI	NA	NAT NA	d No	K Tabo	Talas	data scientist	data scientist	bergeluru kamataka India job typ Sultima c
429 /	zue Bi Enginer (Azue-Pouer B)	Mrit Reserch	Bergaluni, Gersabila	EmployerActive 22 - I days ago	In years of experience in Prover If Heuriconi	Nichteday 68 of 68	https://k.indexd.com/company/Mrid- Brown.	to to	No N	be false false false	Table	Table	False False	Table Table	Table 1	ane have have	Note Table. The	Table Table	Table Table	Table Table	Table Table	false	Designation: Asset Bi Engineer(Asset + Power	Net	MI	161	Net Ne	d No	K False	False	soure 5 engineeroursposer bi	espiner	designation acure bi engineeracur power bi es.
9879 S	micr Software Engineer - Socient	Digital Green	Begsiani, Sambia	PurticiPutad 10+ days ago	As a key member of the solved opy learn, the po	fulfers 640	htps://kindedconje/dif- je-fe/fe/Elbedb.	tu tu	Table To	in Tales Tales Tales	Taba	Taba	False False	Table Table	False I	also falso falso	No No Tue	Table True	Talue False	Talan Talan	Table Table	False	(Volkanataka, Bergalone, India (V) (VIII-III Tier	Net	MI	Net	No.	1 16	Table	Talas	software engineer backend	sabuse engineer	kerutaka tangelon inda full tin technologi

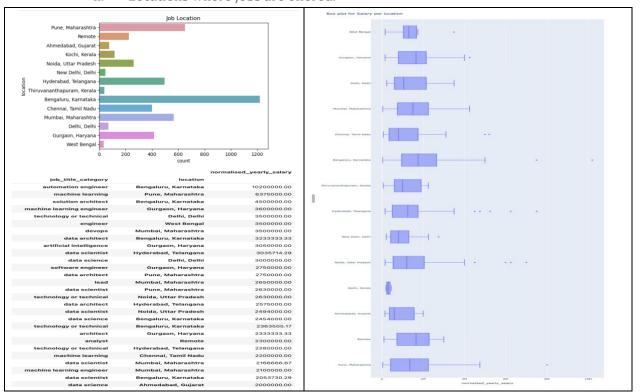
B. Market Data Visualization

i. Titles



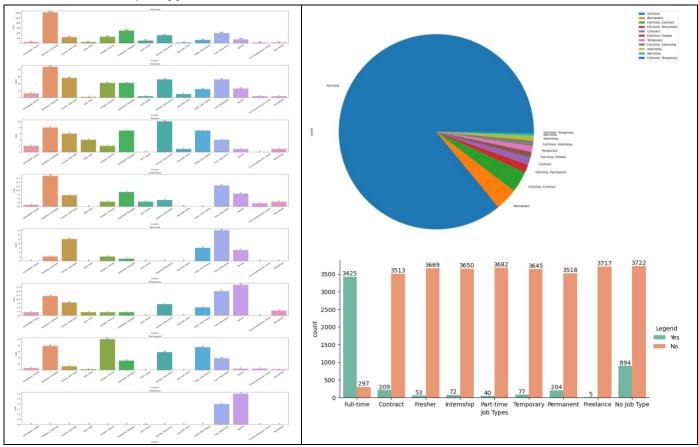
We can see that top categories are engineer, developer, analyst, software engineer, data scientist, manager, technology related, AI, Data science, consult, machine learning etc. In comparison to salary, other than positions like management and leadership, most of the jobs are well distributed in the category. But we can see some jobs in data scientist, machine learning and artificial intelligence as the highest salary job.

ii. Locations where jobs are offered.



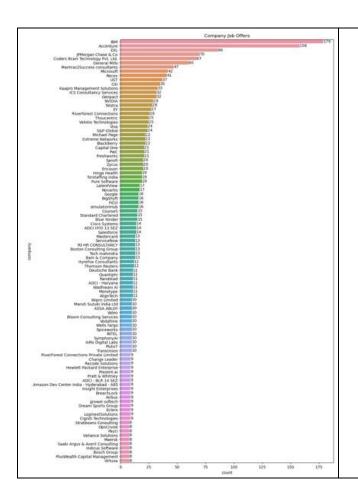
We can see most jobs have demand in Bangalore, Pune, Mumbai and Hyderabad. We can also see salary also follows the same order.





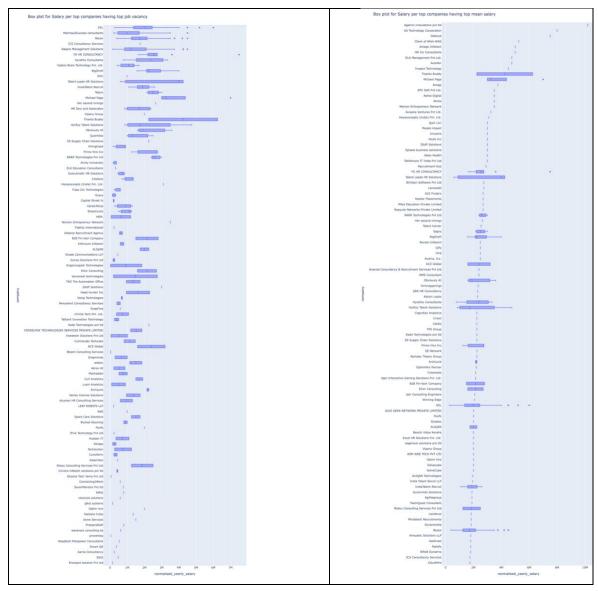
Note that pie chart only contains the top combination of the job types. We can see that full time employee is having highest demand. Part time is only having 10th position in demand.

iv. Company Offers.

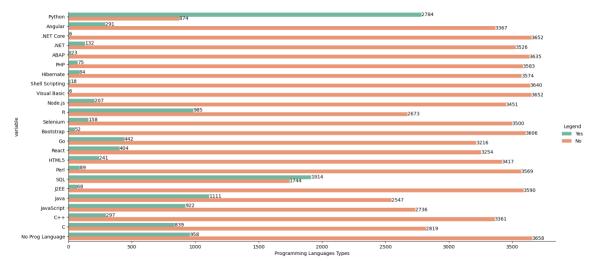


In the dataset, a significant number of companies are observed to have a limited job offering, often less than 10 positions. This trend persists as we narrow down our focus to the top 100 companies. It is noteworthy that there is a total of 1549 companies in the entire dataset out of which thousands of companies have only very few openings.

Furthermore, the salary distribution across different companies indicates a variation in compensation. Notably, only a few companies offer the highest-paying positions, and these tend to be specific roles with limited availability.



v. Skills



The top demanded skills are Python, SQL, Java, R, JavaScript, C, Go, React etc.

vi. Rating

:	job_title	company	location	date	job_desc	metadata	rating	href	Full- time	Python
2240	ML Engineer III	Turing	Bengaluru, Karnataka	PostedPosted 13 days ago	Solid understanding of machine learning algori	Full-time	4.80	https://in.indeed.com/rc/clk? jk=63ee2469b824ee	True	True
3173	Senior Business Analyst	CoinDCX	Bengaluru, Karnataka	PostedPosted 24 days ago	Build and deploy machine learning models where	Full-time	4.80	https://in.indeed.com/rc/clk? jk=df76e526b80126	True	True
1184	Data Scientist	Curl Analytics	Bengaluru, Karnataka	EmployerActive 5 days ago	In this role, you will be responsible for desi	Up to ₹19,03,336 a year Full-time +1 Monday to	4.70	https://in.indeed.com/company/Curl- Analytics/j	True	True
1676	Full Stack Developer	Curl Analytics	Bengaluru, Karnataka	PostedPosted 10 days ago	Curl works on developing Products & Solutions	₹5,00,000 - ₹25,00,000 a year Full-time	4.70	https://in.indeed.com/rc/clk? jk=9040bc7c84c209	True	True
1097	Data Science - AI/ML with Big data Architect	Anlage Infotech	Bengaluru, Karnataka	EmployerActive 29 days ago	As a Solution Architect, the candidate will be	Up to ₹50,00,000 a year Full-time +1	4.70	https://in.indeed.com/company/Anlage- Infotech/	True	True
1337	Data Scientist – Technology	Anlage Infotech	Bengaluru, Karnataka	PostedPosted 30+ days ago	Expertise in at least one of the following fie	Full-time	4.70	https://in.indeed.com/rc/clk? jk=60dd8833eb98d7	True	True
2933	RPA Developer (Automation Anywhere)	TAO The Automation Office	Bengaluru, Karnataka	EmployerActive 3 days ago	As an end-to- end automation solutions partner,	₹7,00,000 - ₹12,00,000 a year Full-time +1 Mon	4.50	https://in.indeed.com/company/TAO-The- Automati	True	True
2934	RPA Developer (UiPath)	TAO The Automation Office	Bengaluru, Karnataka	EmployerActive 3 days ago	Experience – 3 to 5 Years*:\nWe are looking fo	₹15,00,000 - ₹20,00,000 a year Full- time Monda	4.50	https://in.indeed.com/company/TAO-The- Automati	True	True
3433	Senior ML Engineer	SymphonyAl	Bengaluru, Karnataka	PostedPosted 30+ days ago	SymphonyAl Summit is an Al-powered IT and Ente	Full-time	4.50	https://in.indeed.com/rc/clk? jk=80f807307b7c48	True	True
3287	Senior Data Scientist	Scienaptic Systems	Bengaluru, Karnataka	PostedPosted 30+ days ago	As a Senior Data Scientist you will play a cri	Full-time	4.50	https://in.indeed.com/rc/clk? jk=6841f99657bee5	True	True
429	Azure BI Engineer(Azure+Power BI)	Infiniti Research	Bengaluru, Karnataka	EmployerActive 20 days ago	4+ years of experience in Power BI Visualizati	Full-time Day shift +1	4.40	https://in.indeed.com/company/Infiniti- Researc	True	True
3673	Senior Software Engineer - Backend	Digital Green	Bengaluru, Karnataka	PostedPosted 30+ days ago	As a key member of the technology team, the po	Full-time	4.40	https://in.indeed.com/rc/clk? jk=8c0f6d030bad8b	True	True

We can also see that the Turing, CoinDCX, Curl Analytics, Anlage Infotech, TAO The Automation Office, SymphonyAI, Scienaptic Systems, Infiniti Research, Digital Green, Yara, Sophos, Salesforce, Google, Elixir Consulting, EXCELUS LEARNING SOLUTIONS PVT LTD, Bain & Company, Adobe, SAP, NVIDIA, NEC Software Solutions and Microsoft is having the highest ratings from other users. This might show the trust these company have with the public.

C. My Ideal Job

- i. My ideal job has following preferences.
 - 1. High Salary
 - 2. Good Company High rating
 - 3. Better Location
 - 4. Scope for improvement.
 - 5. Skills Python, R, SQL, Go, Machine learning, etc.

I will prefer the Bengaluru as the preferred location as this location shows highest demand in job and highest variance in the salary. This indicates the potential for having

highest promotions and salary hikes which shows scope of improvement in the job position. The same location also has relatively cool climate and I personally prefer this place as my relatives are also in same city. I will conclude that the ideal job is Software Engineer, Machine Learning, Google Assistant (ID:3958) from Google. But If I ony consider the jobs for which salary is given, then I will choose Data Science - AI/ML with Big data Architect (ID: 1097) from Anlage Infotech company.