PREDICTING DATA SCIENCE JOB SALARIES

By

Abhay Aanabathula

OVERVIEW

- Problem statement
- Background
- My goal
- Data
- EDA
- Modeling and predictions

PROBLEM STATEMENT

- Salary varies
  - By state
  - By job level/ position/ entailments
  - On average, range of $87-130,000
- With any job, it's hard to answer this question:
  - What should we pay you?
- Getting a salary estimate based on your skills would make it a lot easier!

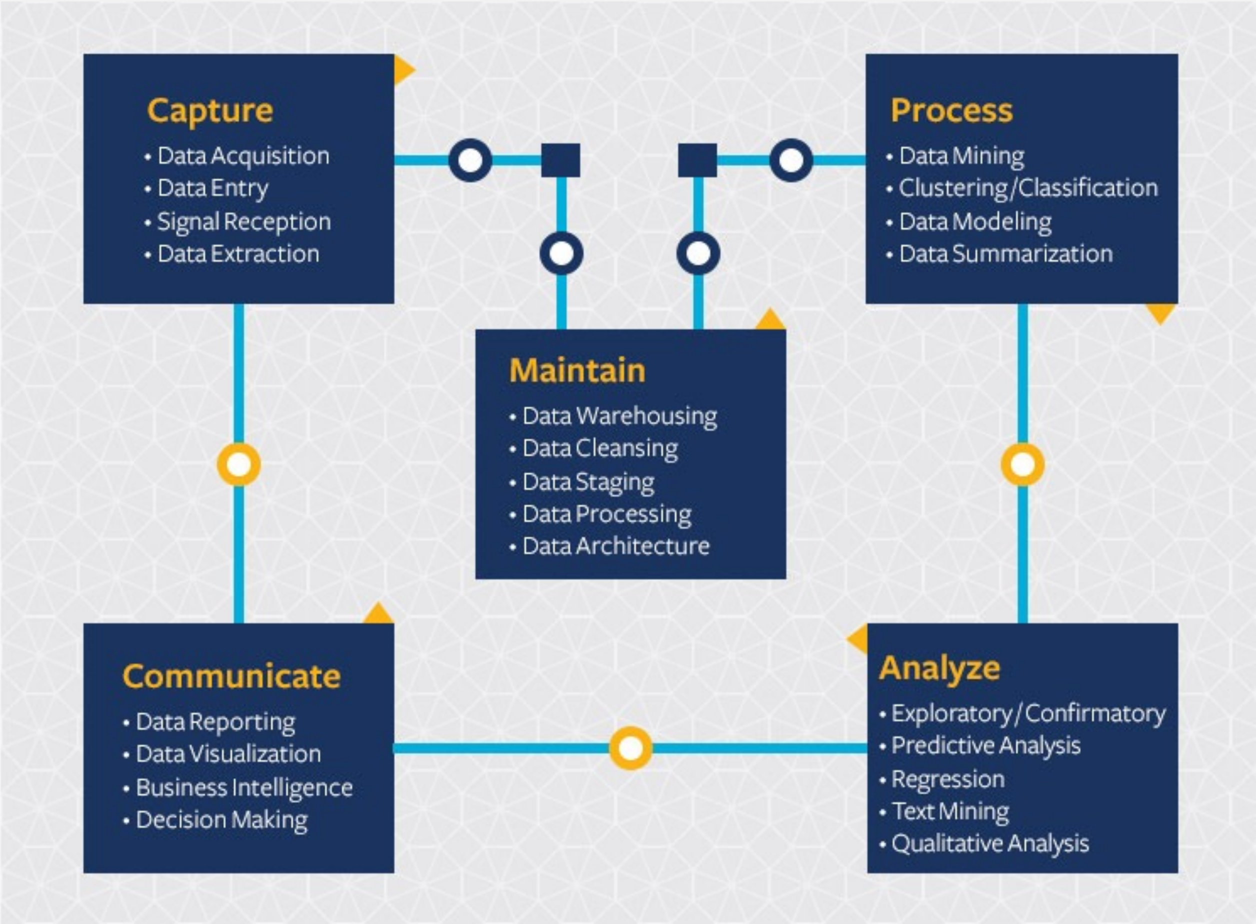
https://www.ziprecruiter.com/Salaries/What-Is-the-Average-Data-Scientist-Salary-by-State
https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm

WHAT IS DATA SCIENCE?

- Organizing and analyzing massive amounts of data
- Successful data scientists can:
  - Identify relevant questions
  - Collect data from a multitude of different data sources
  - Organize the information
  - Translate results into solutions
  - Communicate the findings in a way that positively affects decisions
- Needed in all industries

https://ischoolonline.berkeley.edu/data-science/what-is-data-science/

LIFE CYCLE OF A DATA SCIENTIST



Capture

- Data Acquisition
- Data Entry
- Signal Reception
- Data Extraction

Process

- Data Mining
- Clustering/Classification
- Data Modeling
- Data Summarization

Maintain

- Data Warehousing
- Data Cleansing
- Data Staging
- Data Processing
- Data Architecture

Communicate

- Data Reporting
- Data Visualization
- Business Intelligence
- Decision Making

Analyze

- Exploratory/Confirmatory
- Predictive Analysis
- Regression
- Text Mining
- Qualitative Analysis

From https://ischoolonline.berkeley.edu/data-science/what-is-data-science/

> “The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”
>
> – Hal Varian, chief economist at Google and UC Berkeley professor of information sciences business, and economics

From https://ischoolonline.berkeley.edu/data-science/what-is-data-science/

MY GOAL

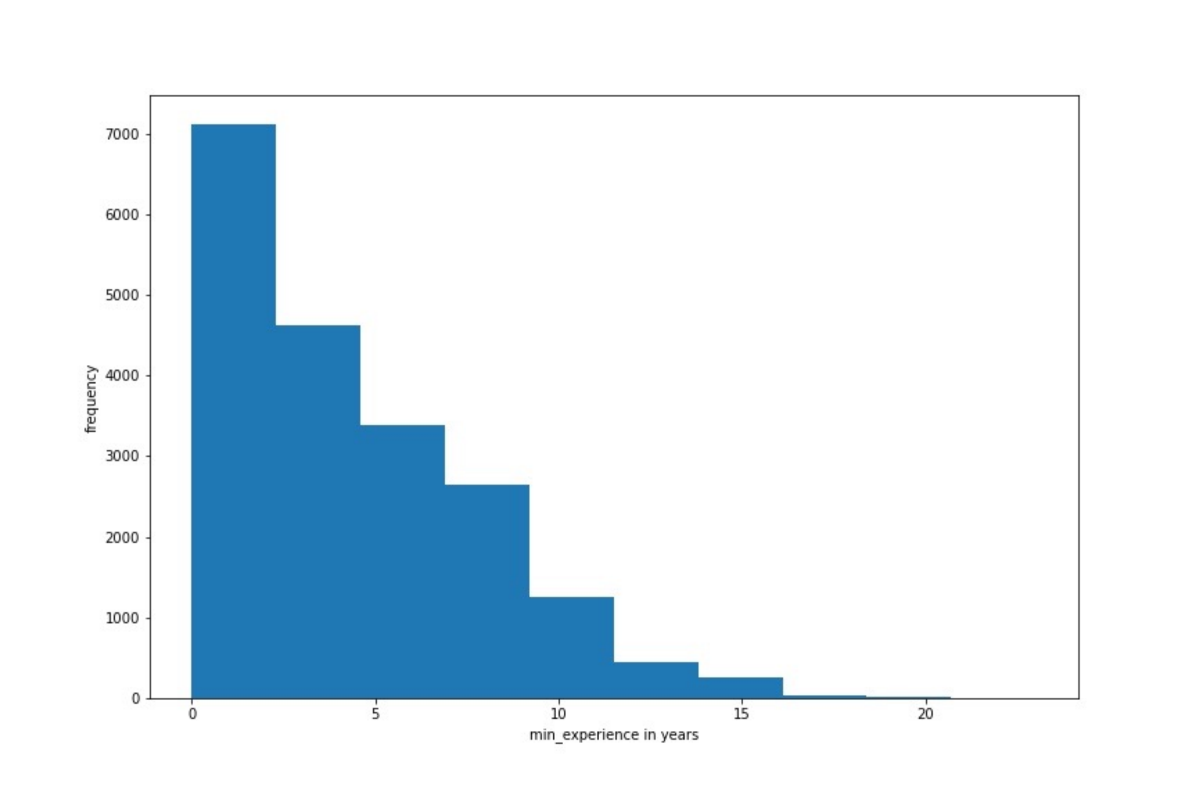
- Create models that accurately predict the average salary of a job based on experience, the job description, the job title, location, and key skills
- Create an interactive application (streamlit) for individuals to input their skills and be given a salary estimate
- Stretch goals:
  - Webscrape from job searching webpages like LinkedIn, Indeed, and Glassdoor
  - Apply same logic to other job titles and in other sectors of the job market

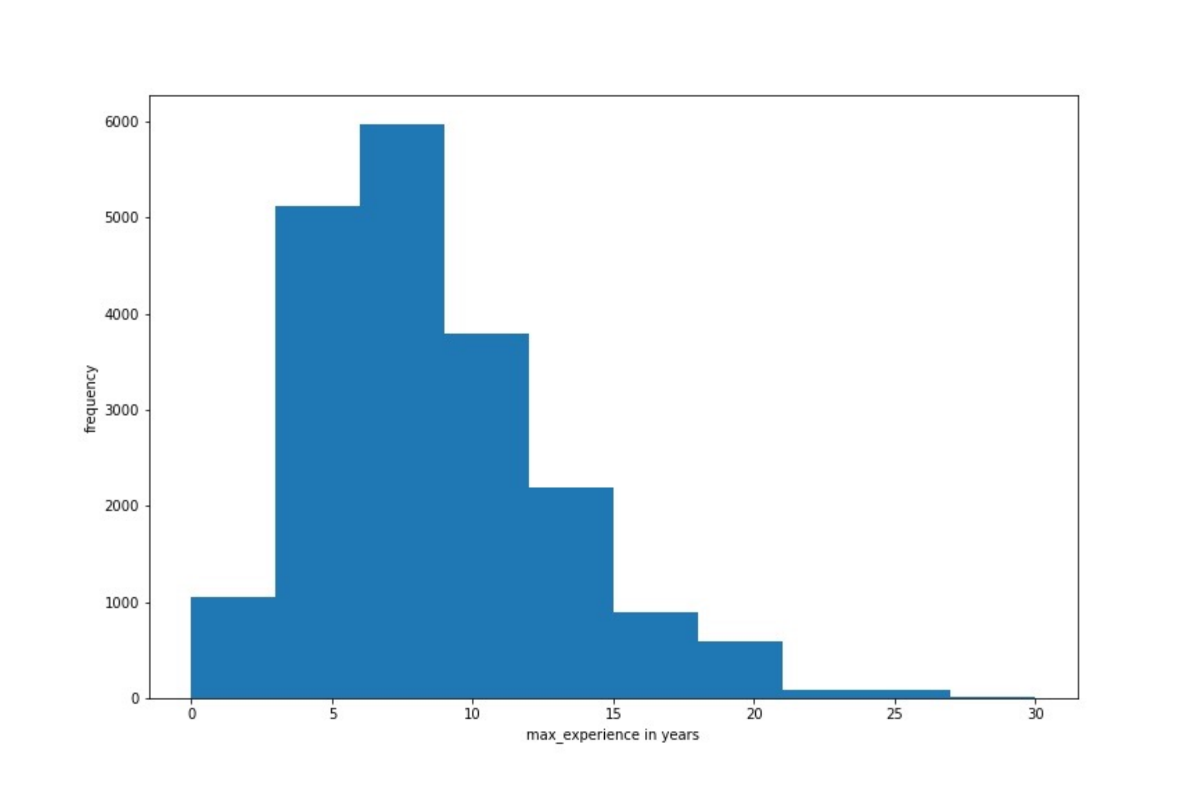
THE DATA

- From a Kaggle challenge
  - https://www.kaggle.com/ankitkalauni/predict-the-data-scientists-salary-in-india
- Variables: experience, job_description, job_desig, job_type, key_skills, location, and salary_range
- Transformations/ restructuring/ reformatting:
  - Dropped job_type
  - NLP for job_description (tokenize, lemmatize, stop_words)
  - Experience → min_experience and max_experience
  - Salary_range → average_salary
  - Count Vectorize all string data together

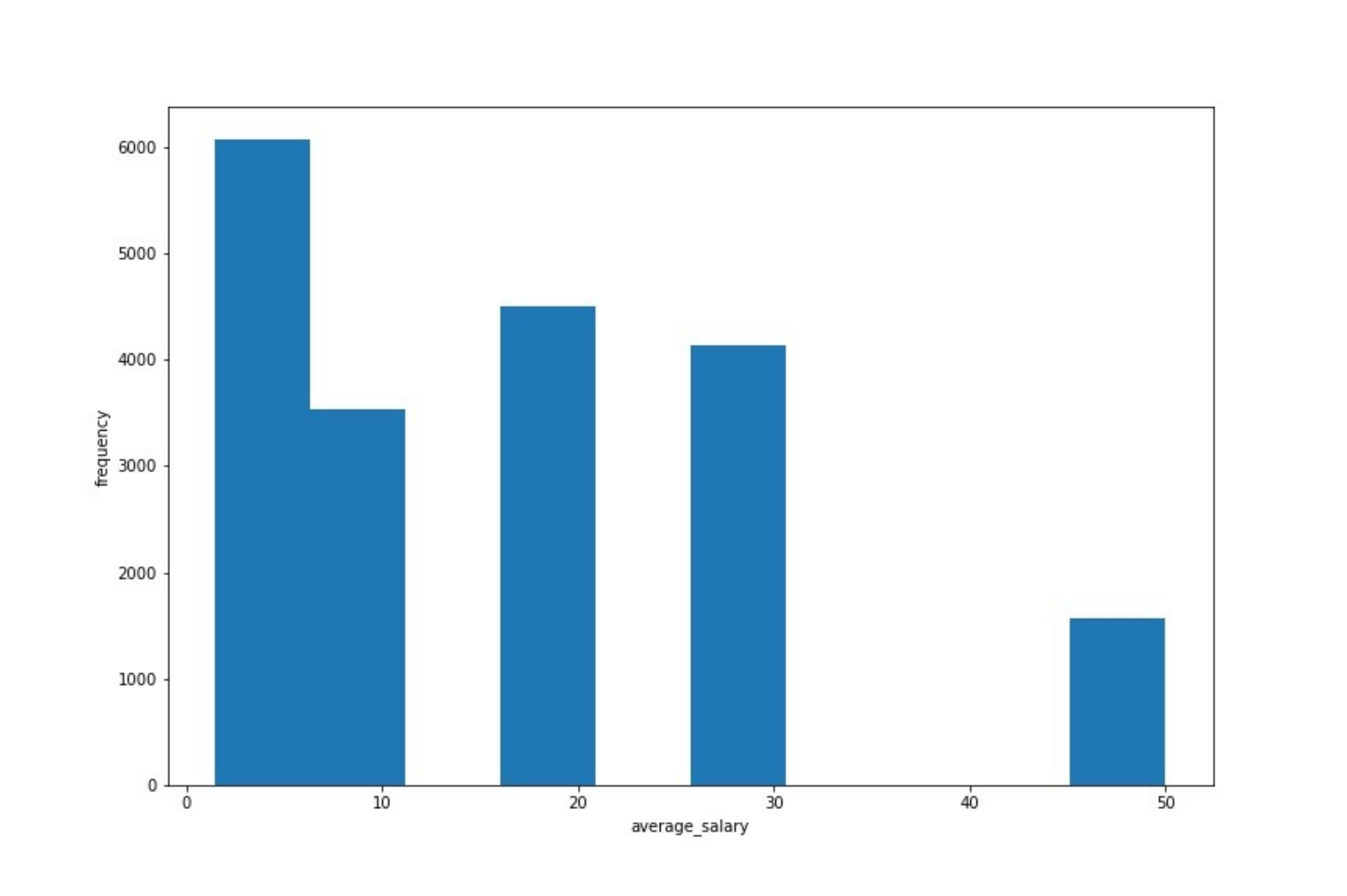
EDA

MIN AND MAX EXPERIENCE DISTRIBUTION

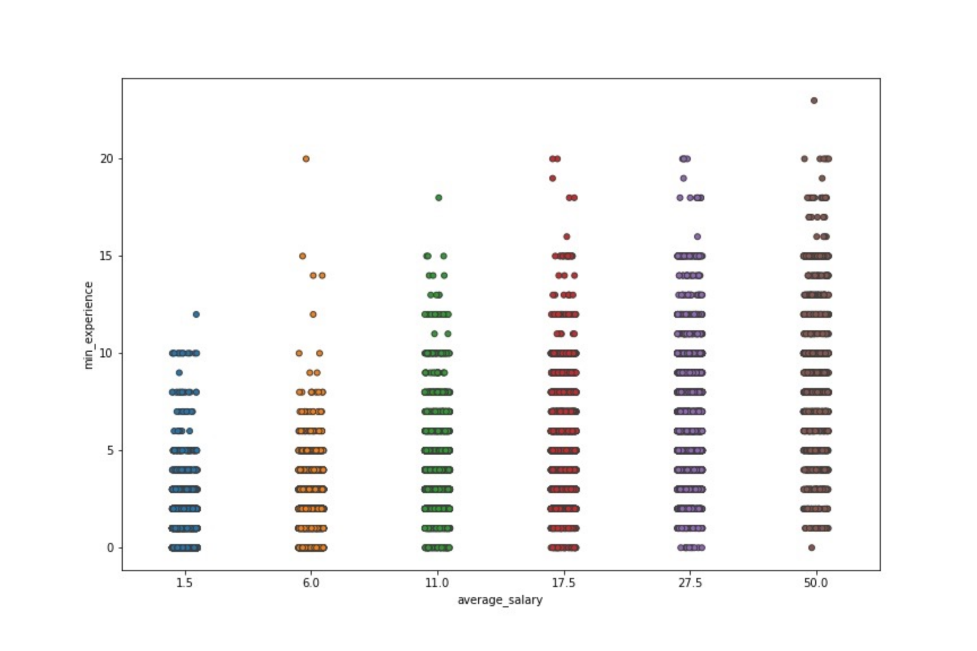


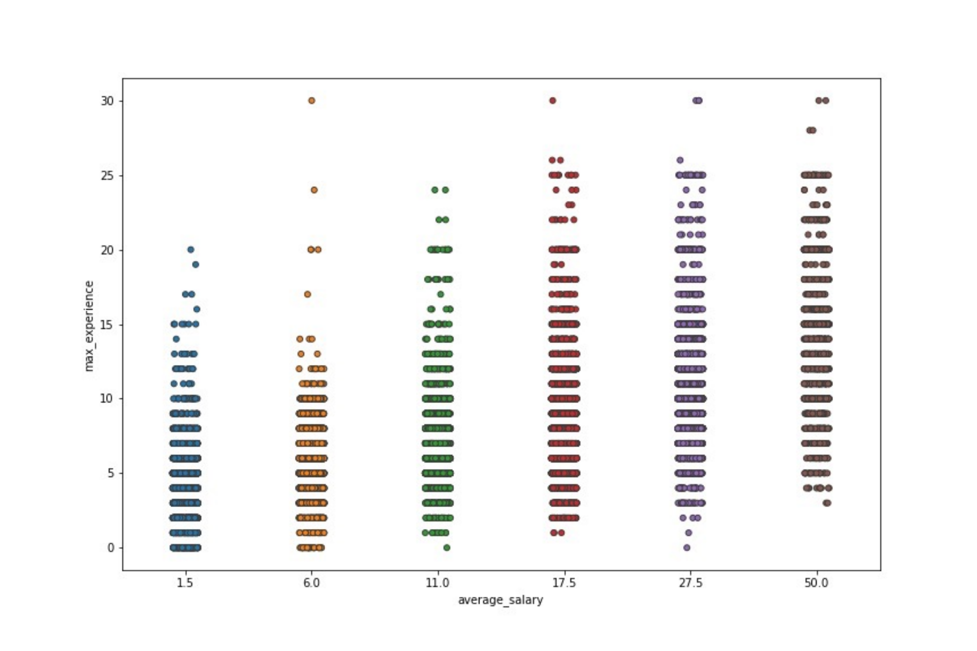


AVERAGE_SALARY DISTRIBUTION

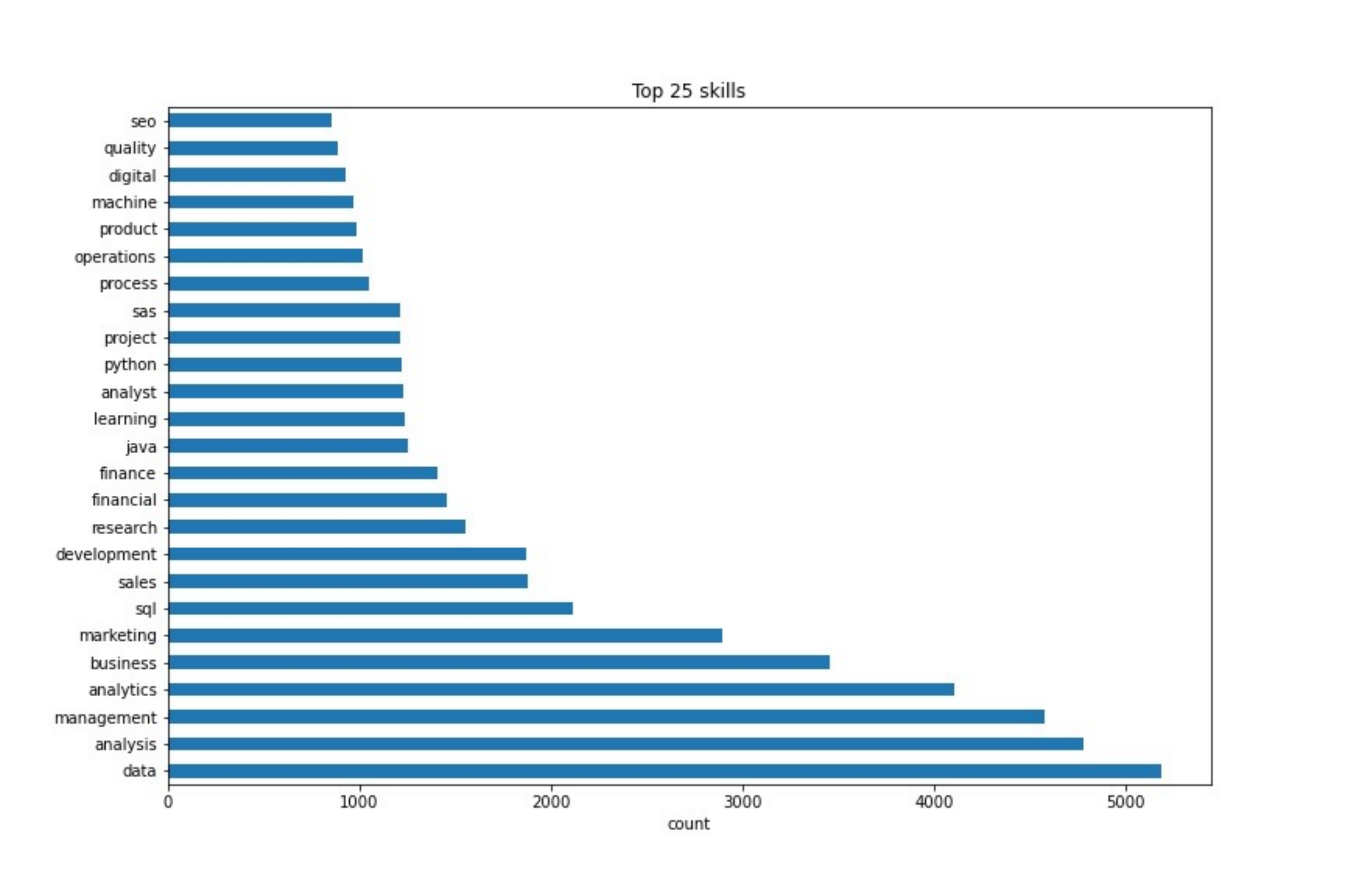


EXPERIENCE VS. AVERAGE_SALARY

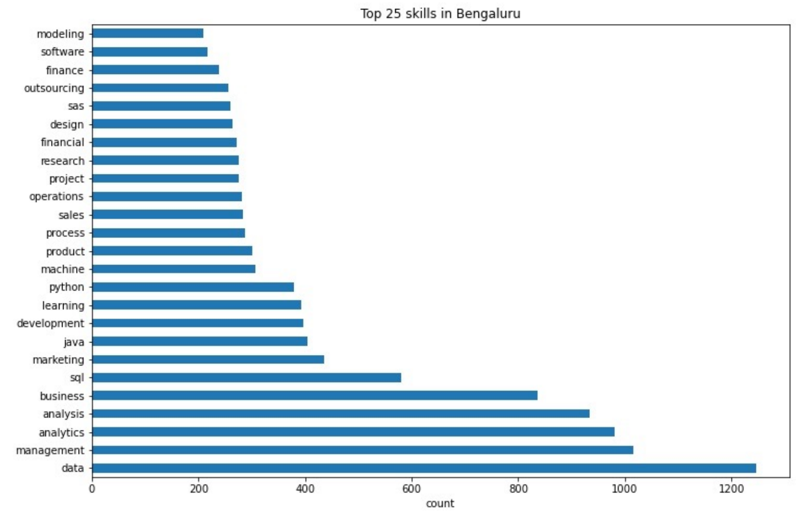


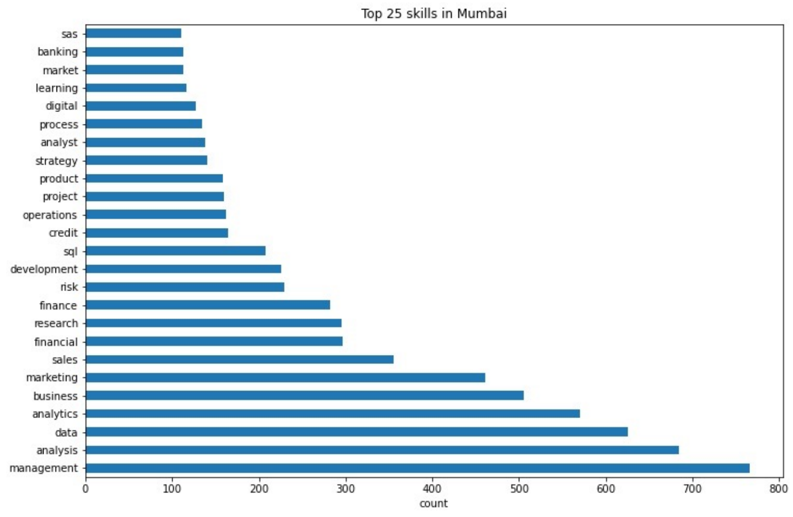


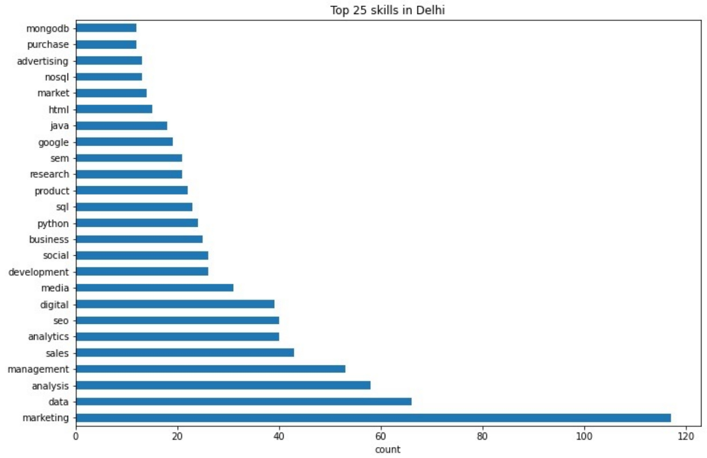
TOP 25 SKILLS OVERALL



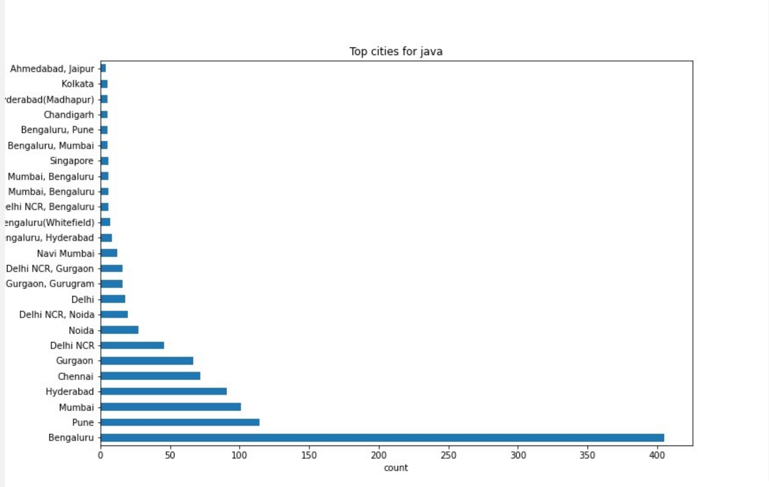
TOP SKILLS IN MOST POPULATED CITIES

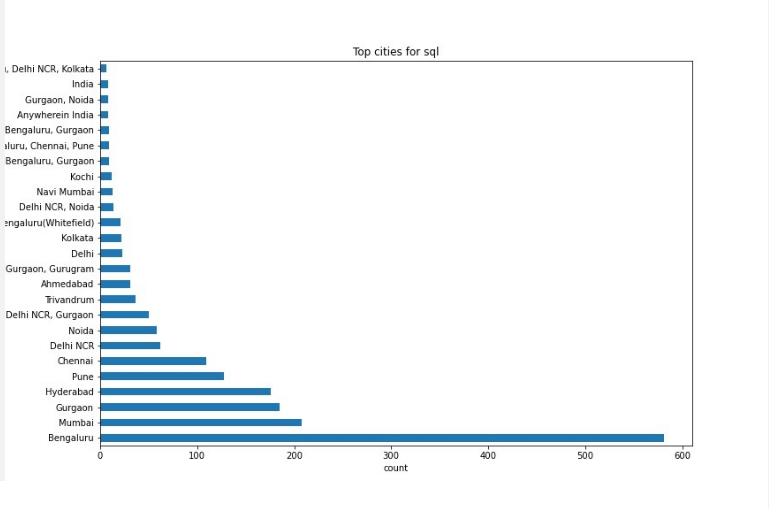


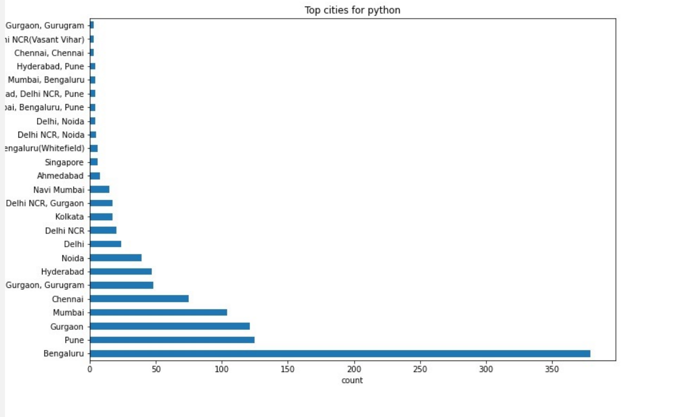




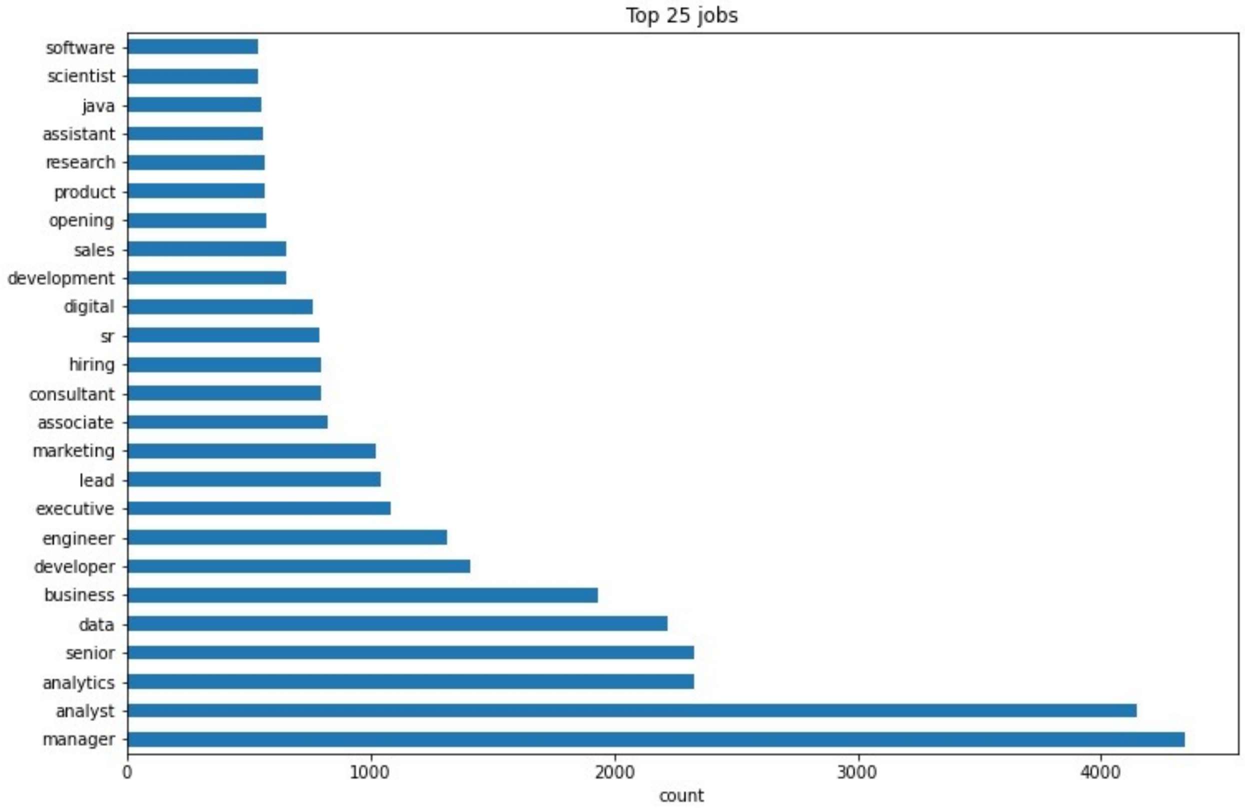
HARD SKILLS IN CITIES



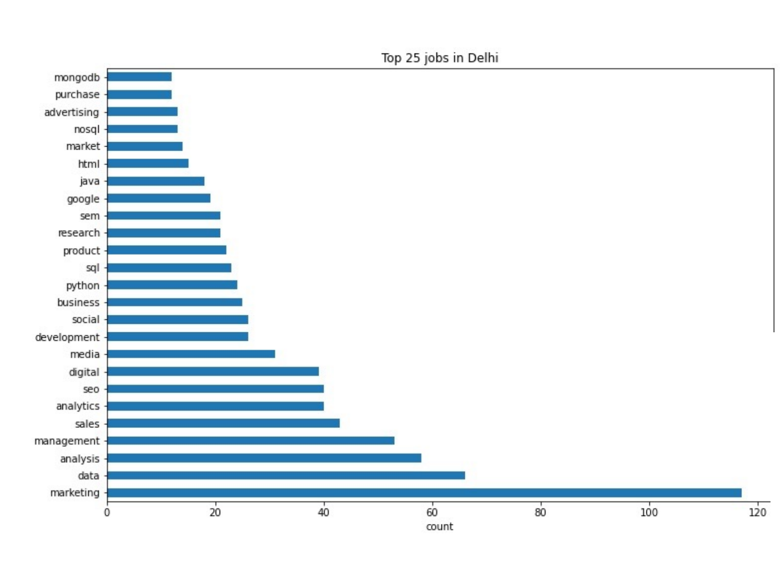


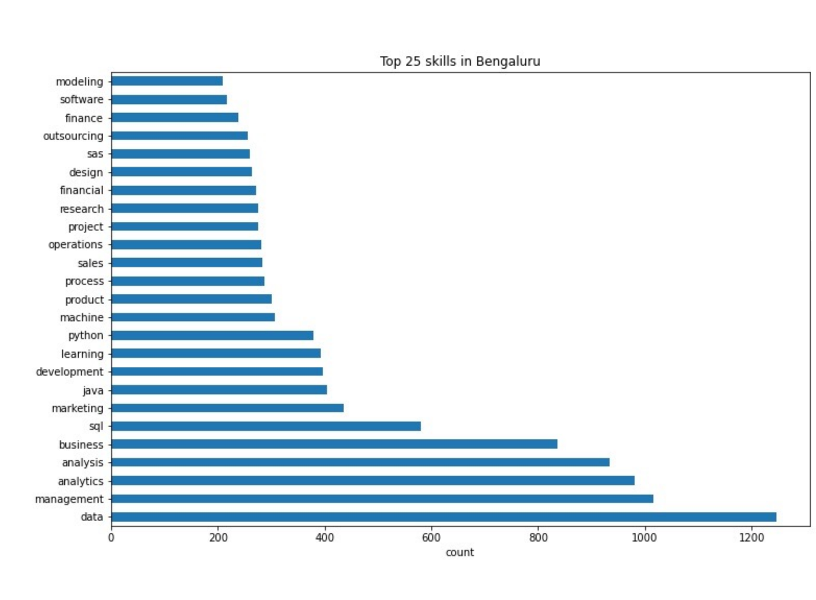


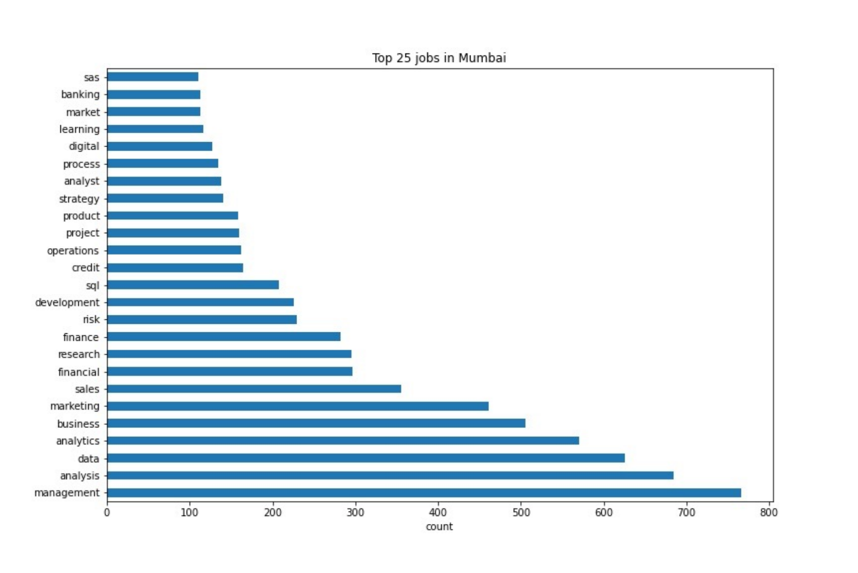
TOP 25 JOB DESIGNATIONS/TITLES



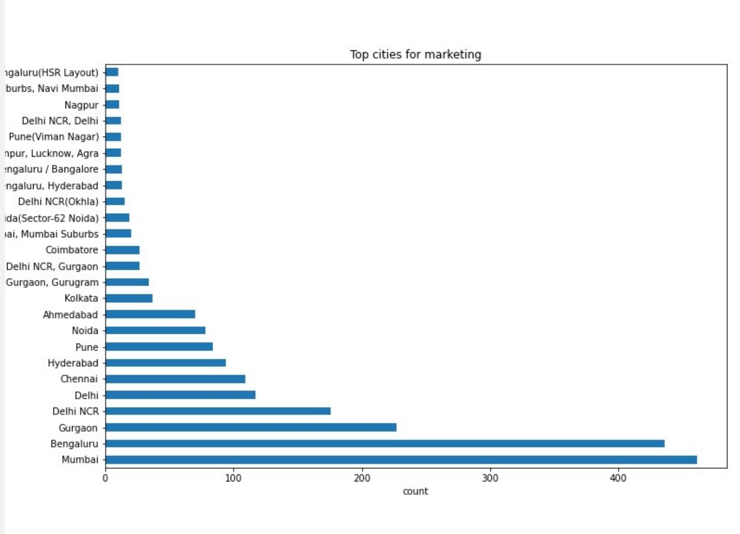
TOP JOB DESIGNATIONS IN MOST POPULATED CITIES

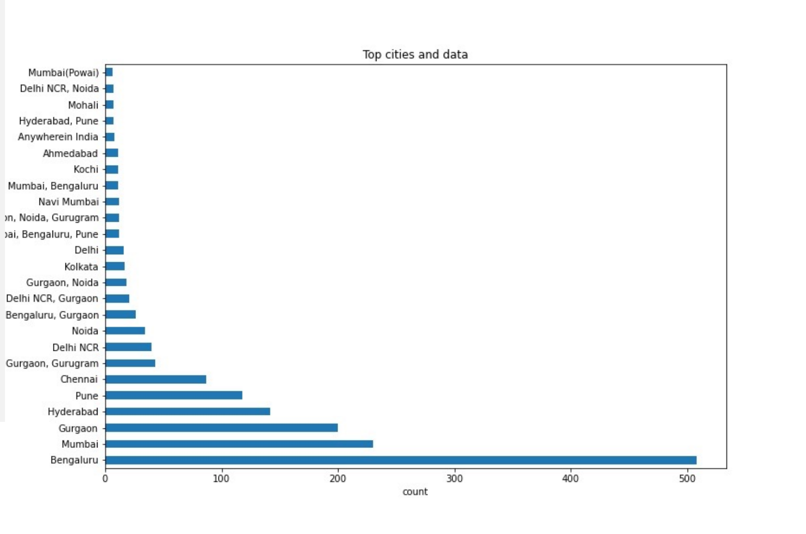


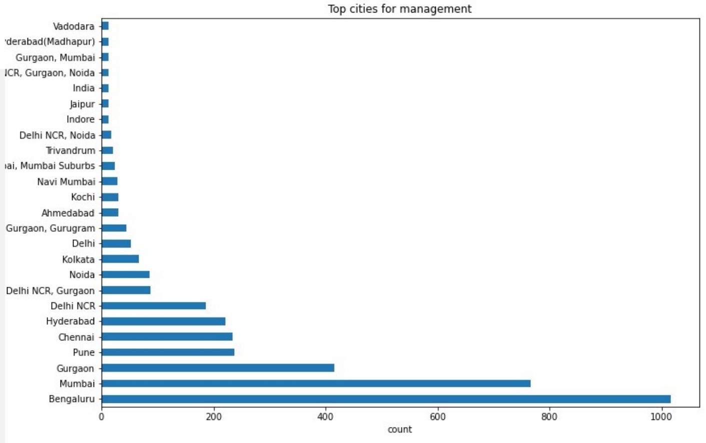




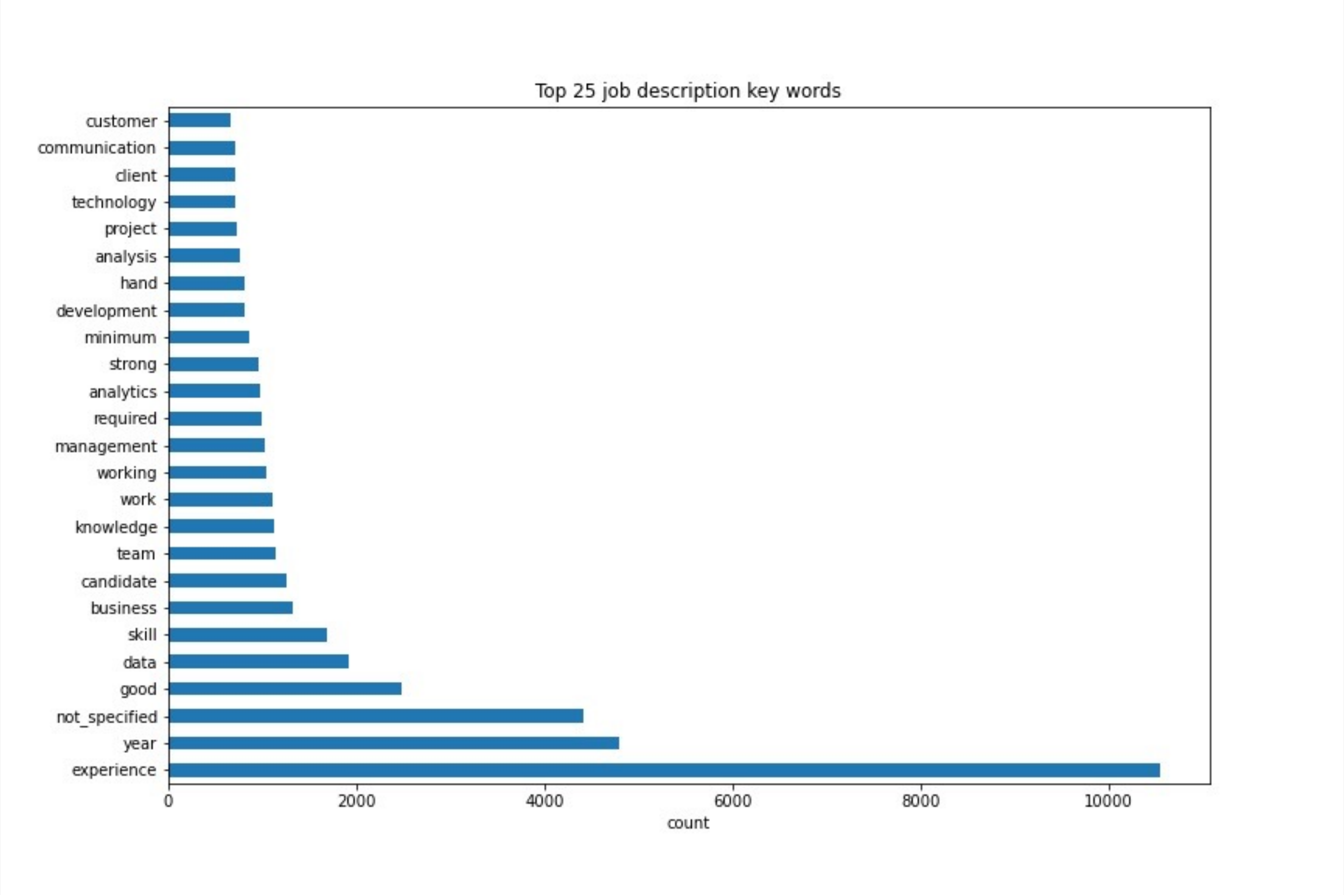
JOB DESIGNATIONS IN CITIES



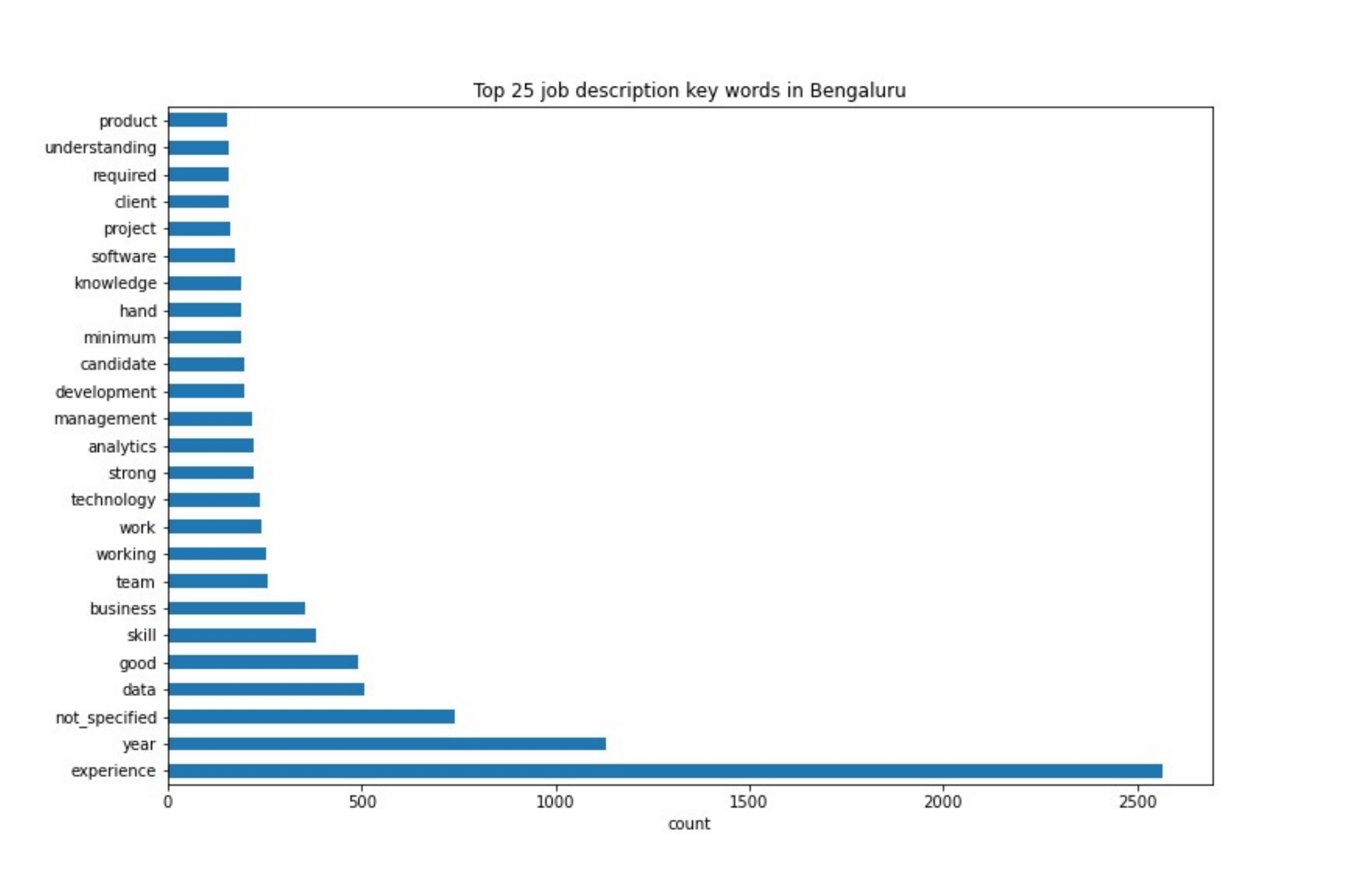




TOP 25 JOB DESCRIPTION KEY WORDS

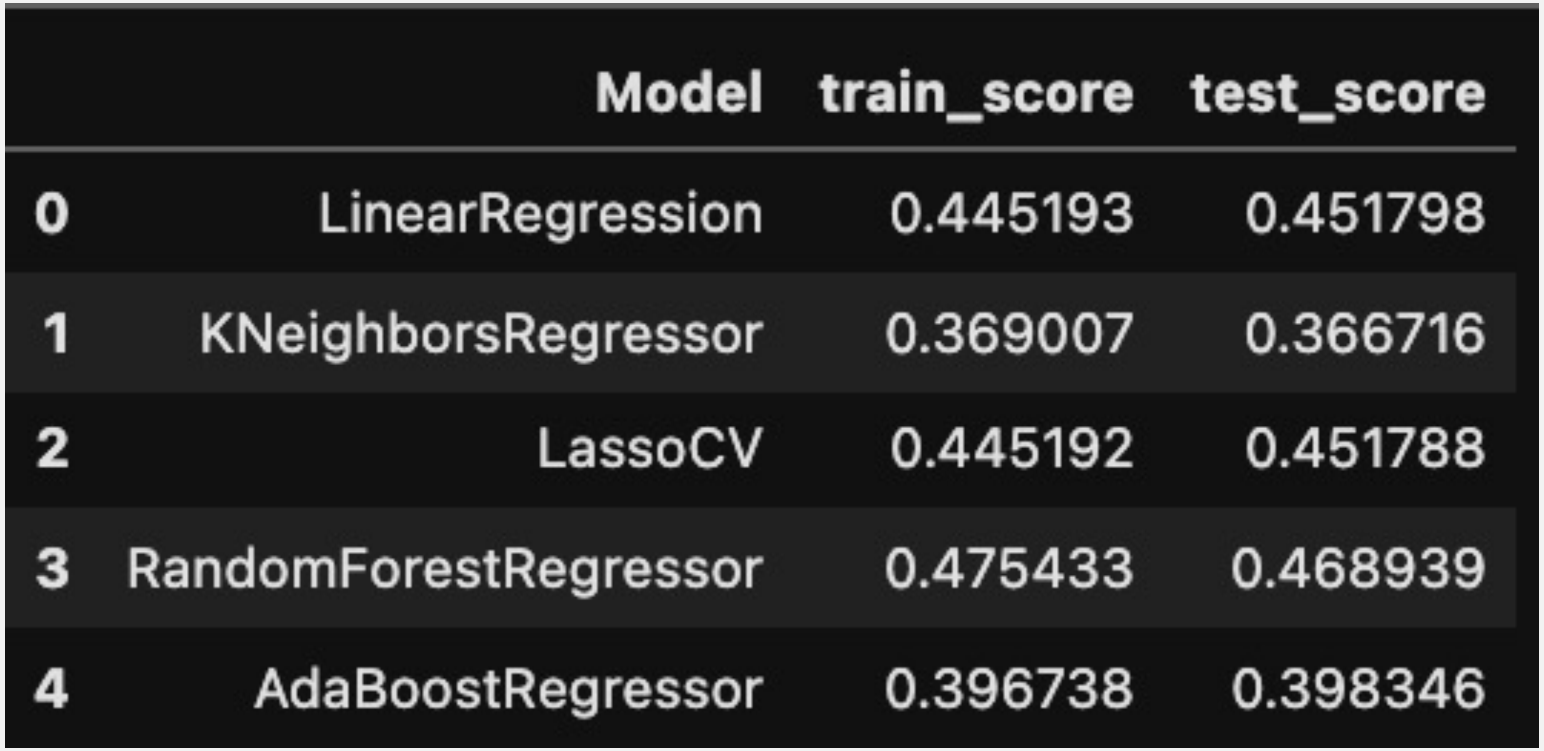


JOB DESCRIPTIONS IN MOST POPULATED CITIES



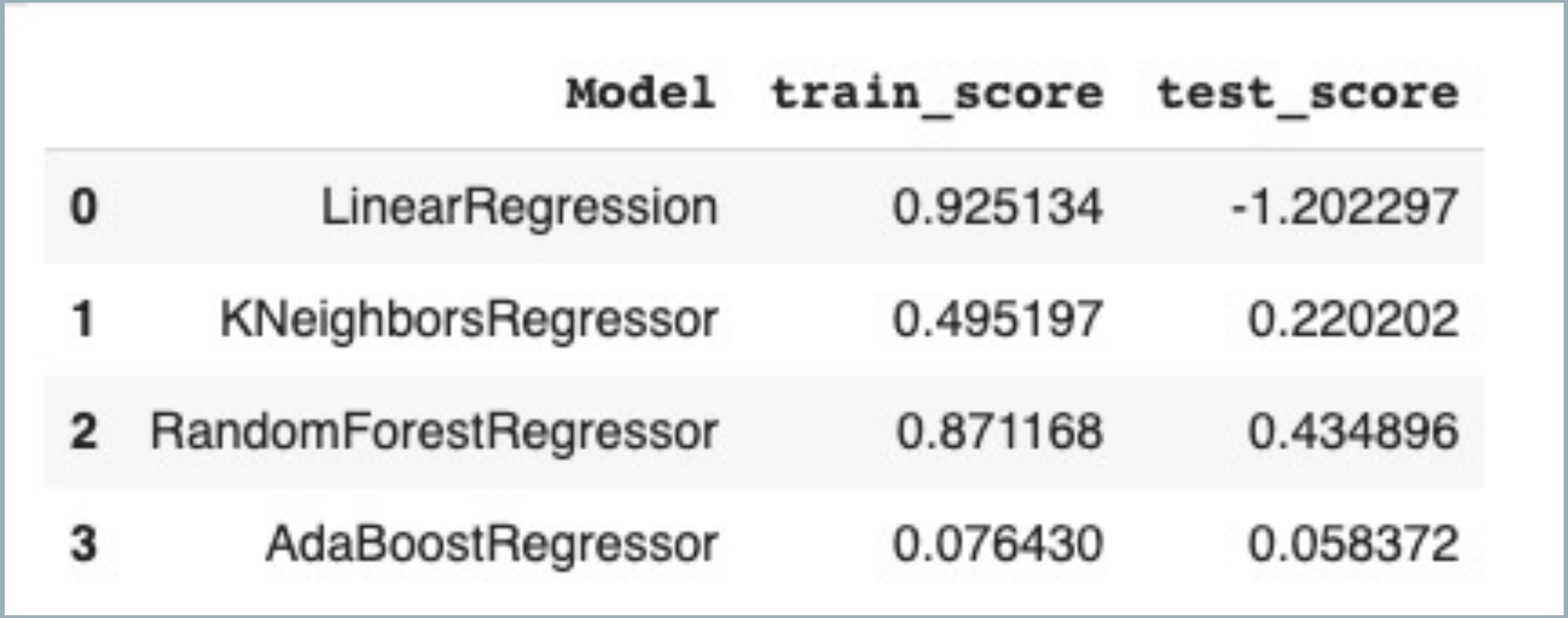
MODELING AND PREDICTIONS

QUANTITATIVE VARIABLE ANALYSIS



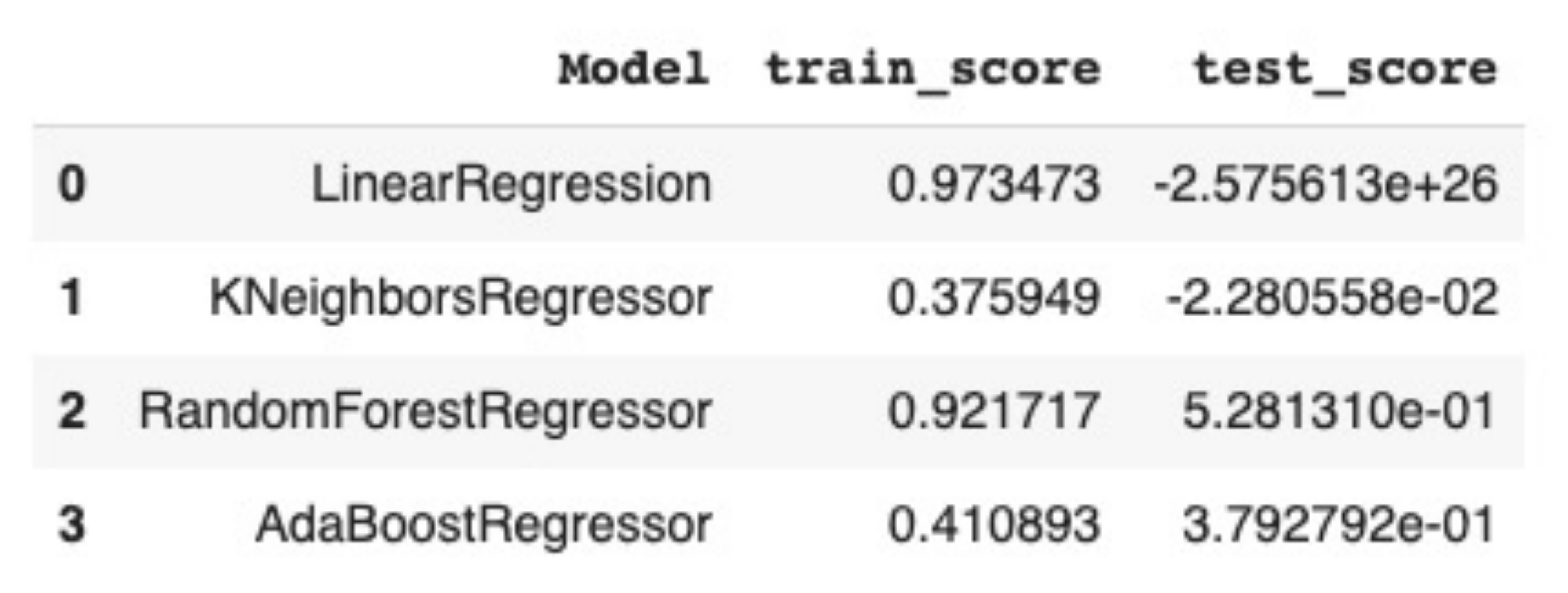
	Model	train_score	test_score
0	LinearRegression	0.445193	0.451798
1	KNeighborsRegressor	0.369007	0.366716
2	LassoCV	0.445192	0.451788
3	RandomForestRegressor	0.475433	0.468939
4	AdaBoostRegressor	0.396738	0.398346

NLP ANALYSIS



	Model	train_score	test_score
0	LinearRegression	0.925134	-1.202297
1	KNeighborsRegressor	0.495197	0.220202
2	RandomForestRegressor	0.871168	0.434896
3	AdaBoostRegressor	0.076430	0.058372

FULL ANALYSIS (ALL VARIABLES INCLUDED)

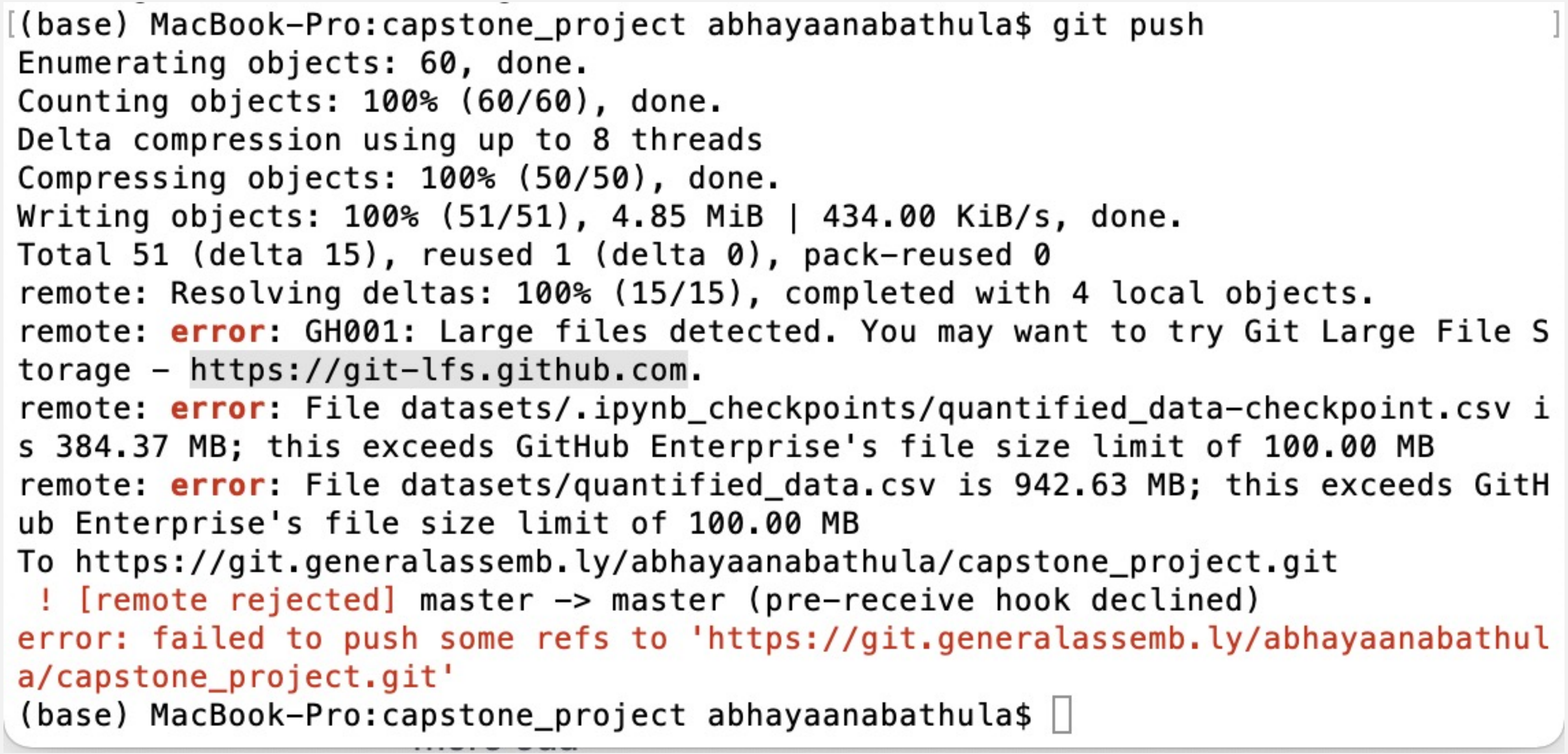


	Model	train_score	test_score
0	LinearRegression	0.973473	-2.575613e+26
1	KNeighborsRegressor	0.375949	-2.280558e-02
2	RandomForestRegressor	0.921717	5.281310e-01
3	AdaBoostRegressor	0.410893	3.792792e-01

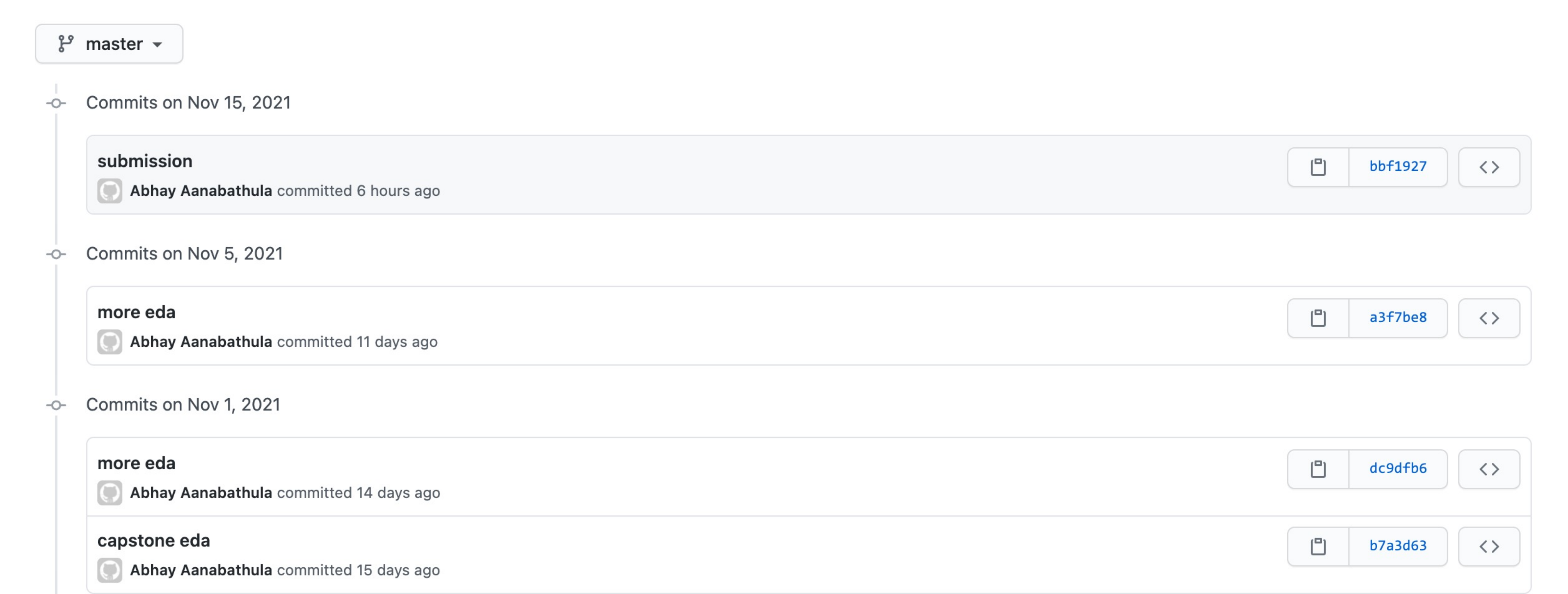
NOW THINGS FELL APART

PROBLEMS WITH FULL ANALYSIS

- Dataframe created too large
- Session kept crashing
- Used all available RAM



```
[(base) MacBook-Pro:capstone_project abhayaanabathula$ git push                    ]
Enumerating objects: 60, done.
Counting objects: 100% (60/60), done.
Delta compression using up to 8 threads
Compressing objects: 100% (50/50), done.
Writing objects: 100% (51/51), 4.85 MiB | 434.00 KiB/s, done.
Total 51 (delta 15), reused 1 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (15/15), completed with 4 local objects.
remote: error: GH001: Large files detected. You may want to try Git Large File S
torage - https://git-lfs.github.com.
remote: error: File datasets/.ipynb_checkpoints/quantified_data-checkpoint.csv i
s 384.37 MB; this exceeds GitHub Enterprise's file size limit of 100.00 MB
remote: error: File datasets/quantified_data.csv is 942.63 MB; this exceeds GitH
ub Enterprise's file size limit of 100.00 MB
To https://git.generalassemb.ly/abhayaanabathula/capstone_project.git
 ! [remote rejected] master -> master (pre-receive hook declined)
error: failed to push some refs to 'https://git.generalassemb.ly/abhayaanabathul
a/capstone_project.git'
(base) MacBook-Pro:capstone_project abhayaanabathula$ 
```



Commits on Nov 15, 2021

submission
Abhay Aanabathula committed 6 hours ago — bbf1927

Commits on Nov 5, 2021

more eda
Abhay Aanabathula committed 11 days ago — a3f7be8

Commits on Nov 1, 2021

more eda
Abhay Aanabathula committed 14 days ago — dc9dfb6

capstone eda
Abhay Aanabathula committed 15 days ago — b7a3d63

WHAT’S NEXT?

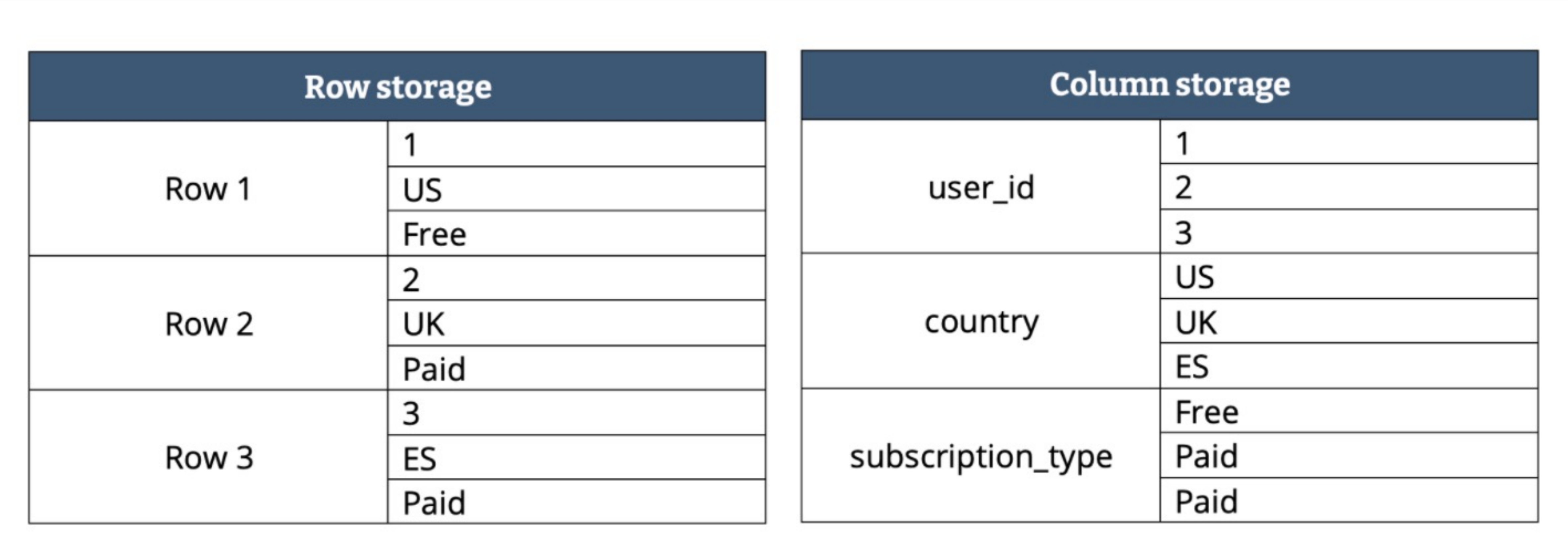
DEALING WITH LARGE DATASETS

- Allocate more memory
- Work with a smaller sample
- Use a computer with more memory
- Change the data format
- Stream data or use progressive loading
- Use a relational database
- Use a big data platform

https://machinelearningmastery.com/large-data-files-machine-learning/

MY PLAN

- Figure out sampling
- Look into other data formats maybe
  - Pickle - stream data
  - Parquet- column storage
  - Feather – memory allocation
- Relational database – mySQL, SQLite
- Big Data Platform – AWS
- Achieve my stretch goals



Row storage	
Row 1	1
	US
	Free
Row 2	2
	UK
	Paid
Row 3	3
	ES
	Paid

Column storage	
user_id	1
	2
	3
country	US
	UK
	ES
subscription_type	Free
	Paid
	Paid

THE END BUT NOT THE END