RESUME CLASSIFICATION

Project Members:

MS. DEVADHARSHINI

MS. HARSHITHA

MR. NAGARAM CHEEDELLA

MS. ASHA RANI

MS. PRACHI GABANI

MS. SRILATHA

Project Mentor:

MR. RAJASHEKAR

BUSINESS OBJECTIVE

The document classification solution should significantly reduce the manual human effort in the HRM. It should achieve a higher level of accuracy and automation with minimal human intervention

Abstract:

In the domain of online job recruitment, accurate job and resume classification is vital for both the seeker and the recruiter. We have built an automatic text classification system that utilizes various techniques like Term frequency-inverse document frequency with Machine Learning and Convolution Neural network for training the model with texts and classifying them into labels and finally to compare their results. Using resume data of applicants, we have categorized them into different categories.

We aim to compare the results obtained by various algorithms that are generated using the same data so that the efficiency of each algorithm can be evaluated. From the result, it is evident that RANDOM FOREST gives a better F1 score, test accuracy, Recall Score and precision, compared to other models.

INTRODUCTION

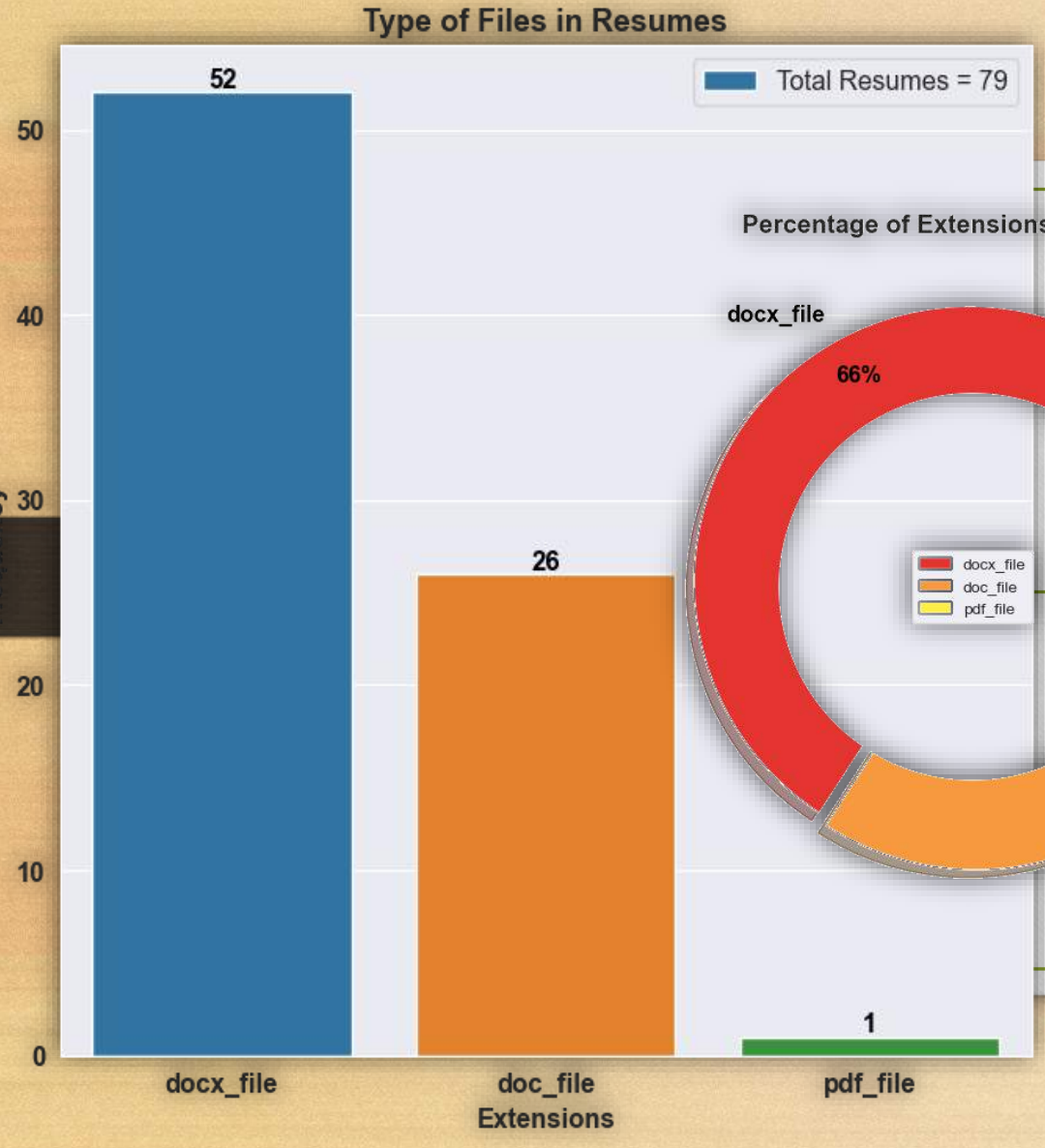
In today's fast-moving corporate industry, recruiters often need to go through vast amounts of resume to analyze the applicants reliably to decide upon the deserving candidates. But it is not possible to keep up with the pace today. So, automated classification of resumes is needed to ease out the process. To do the same, a bulk of labeled resume data is required, and job openings are divided into a certain number of predefined job categories.

In this project we dive into building a **Machine learning model** for **Resume Classification** using **Python** and basic **Natural language processing** techniques. We would be using Python's libraries to implement various NLP (natural language processing) techniques like t**okenization**, **lemmatization**, **parts of speech tagging**, etc

The basic **data analysis** process is performed such as **data collection**, **data cleaning**, **exploratory data analysis**, **data visualization**, and **model building**. The dataset consists of two columns, namely, **Role Applied** and **Resume**, where 'role applied' column is the domain field of the industry and 'resume' column consists of the text extracted from the resume document for each domain and industry.

The aim of this project is achieved by performing the various data analytical methods and using the **Machine Learning models** and **Natural Language Processing** which will help in classifying the categories of the resume and building the Resume Classification Model.

EXPLORATORY DATA ANALYSIS

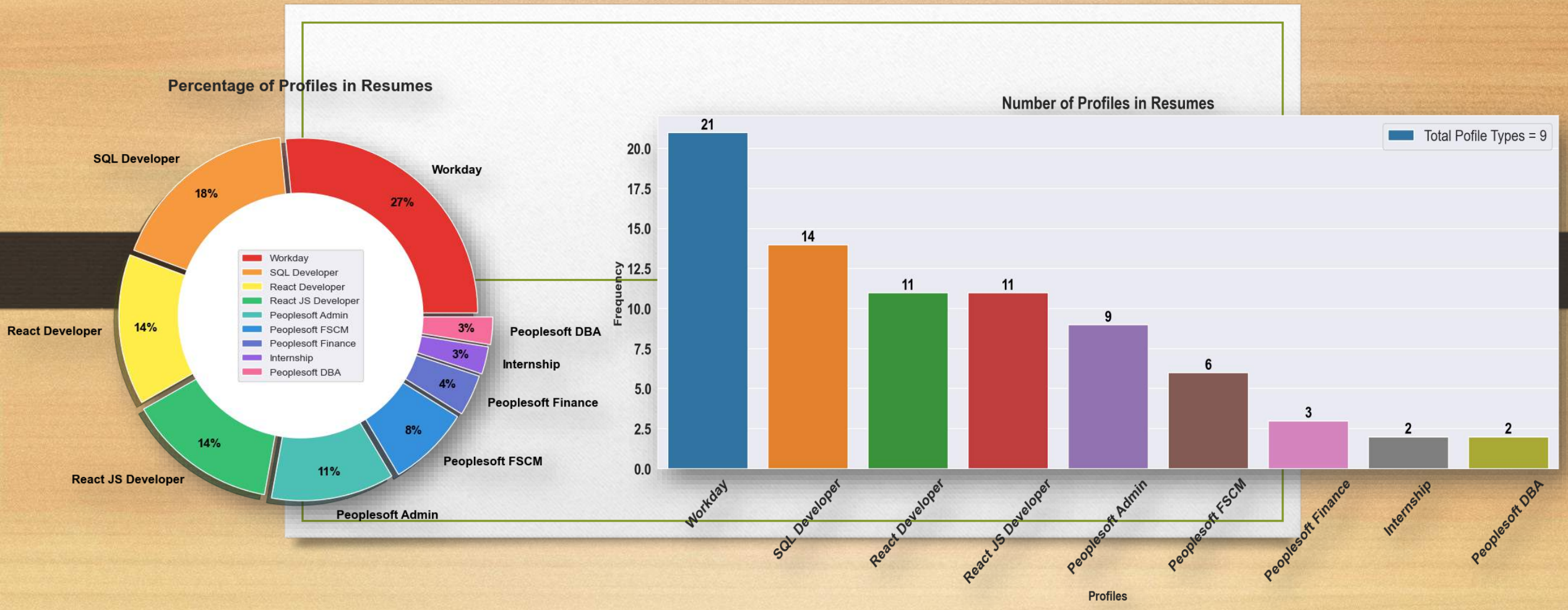


	Resumes	Profile
0	Internship_Ravali_Musquare Technologies (1).docx	Internship
1	Internship_Susovan Bag_Musquare Technologies.docx	Internship
2	Peoplesoft Admin_AnubhavSingh.docx	Peoplesoft Admin
3	Peoplesoft Admin_G Ananda Rayudu.docx	Peoplesoft Admin
4	Peoplesoft Admin_Gangareddy.docx	Peoplesoft Admin
...	...	...
74	Sri Krishna S_Hexaware.docx	Workday
75	Srikanth-Hexaware.docx	Workday
76	SSKumar_Hexaware.docx	Workday
77	Venkateswarlu B_Hexaware.docx	Workday
78	Vinay Kumar_Hexaware.docx	Workday

79 rows × 2 columns

EXPLORATORY DATA ANALYSIS

we have total **9** types of Profiles in the Resumes

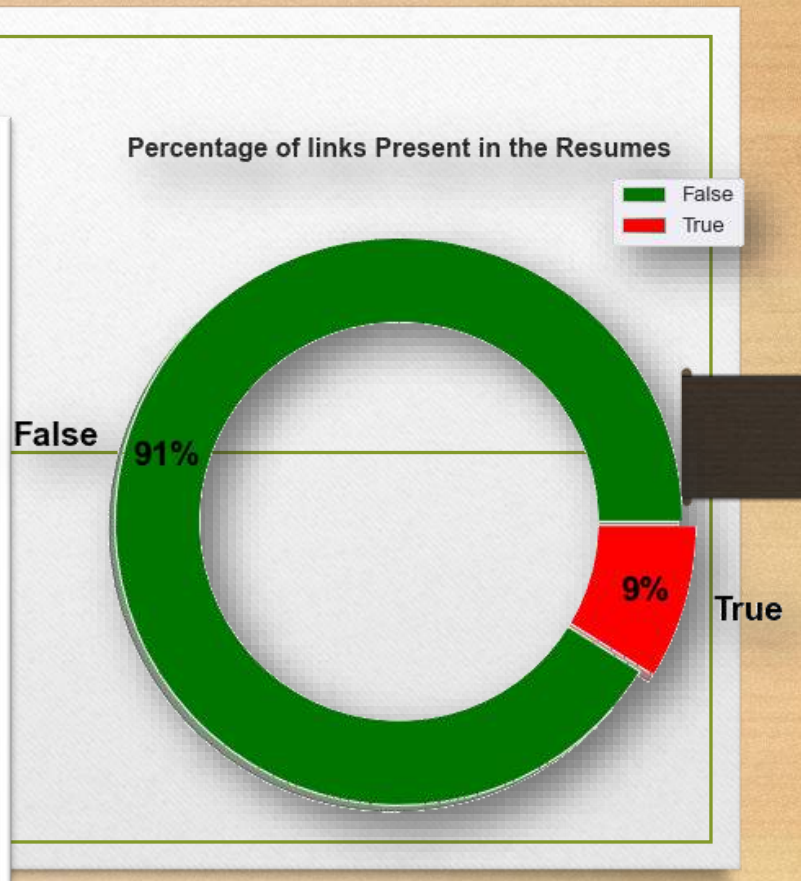


EXPLORATORY DATA ANALYSIS :

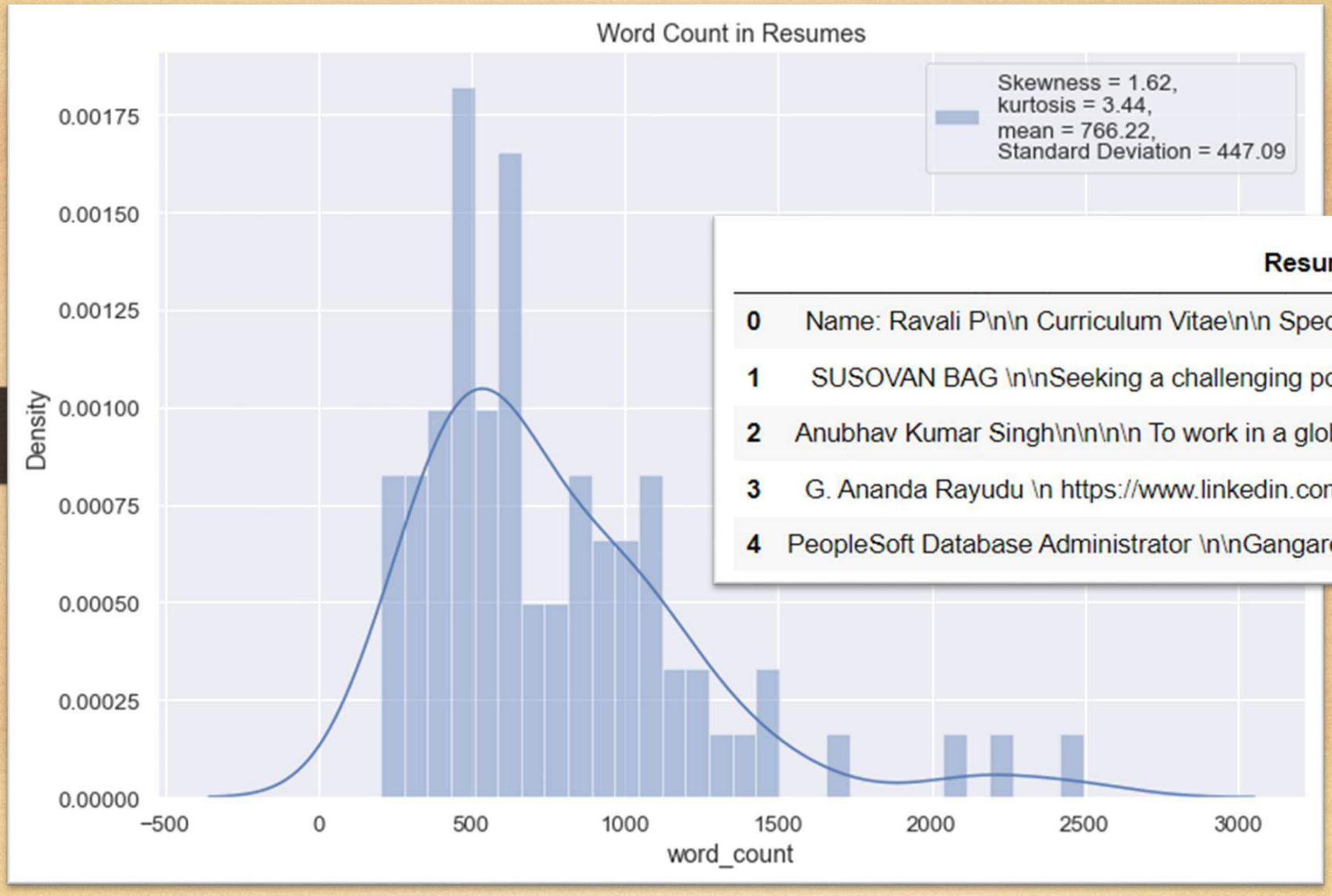
Extracting Text from different Resume files and creating a data-frame with Column of Text from Resumes And Profile for which each of it Applied for.

	Resumes	Profiles
0	Name: Ravali P\n\n Curriculum Vitae\n\n Specia...	Internship
1	SUSOVAN BAG \n\nSeeking a challenging posi...	Internship
2	Anubhav Kumar Singh\n\n\n\n To work in a globa...	Peoplesoft Admin
3	G. Ananda Rayudu \n https://www.linkedin.com/i...	Peoplesoft Admin
4	PeopleSoft Database Administrator \n\nGangared...	Peoplesoft Admin
...	...	...
74	Workday Integration Consultant \n\nName ...	Workday
75	S R I K A N T H (W O R K D A Y H C M C O N...	Workday
76	WORKDAY | HCM | FCM \n\nName Role \n\n: Kumar ...	Workday
77	Venkateswarlu.B \n\n\n\nWorkday Consultant\n\n...	Workday
78	VINAY KUMAR .V \nWORKDAY FUNCTIONAL CONSULTANT...	Workday

79 rows × 2 columns

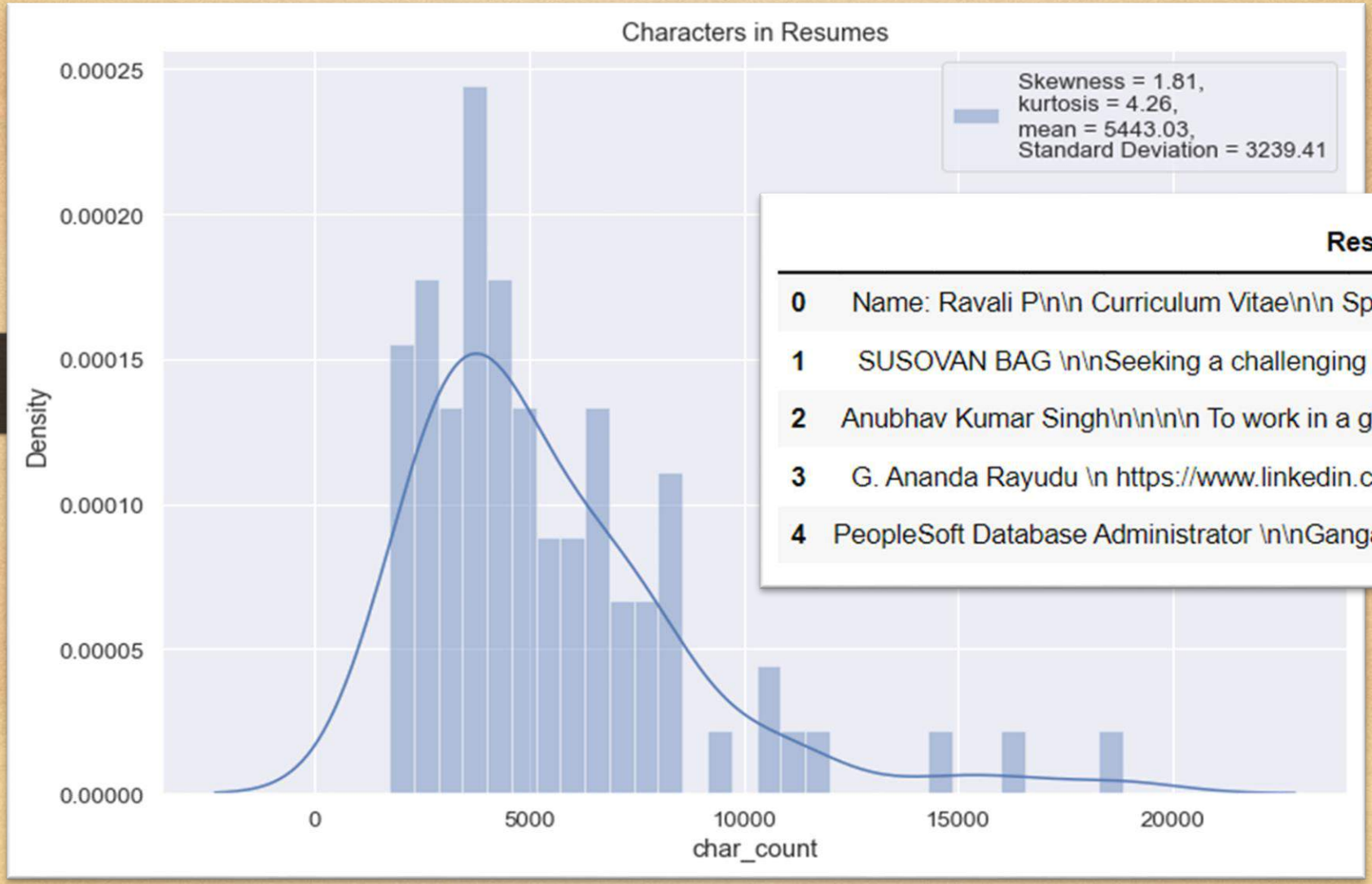


FEATURE ENGINEERING



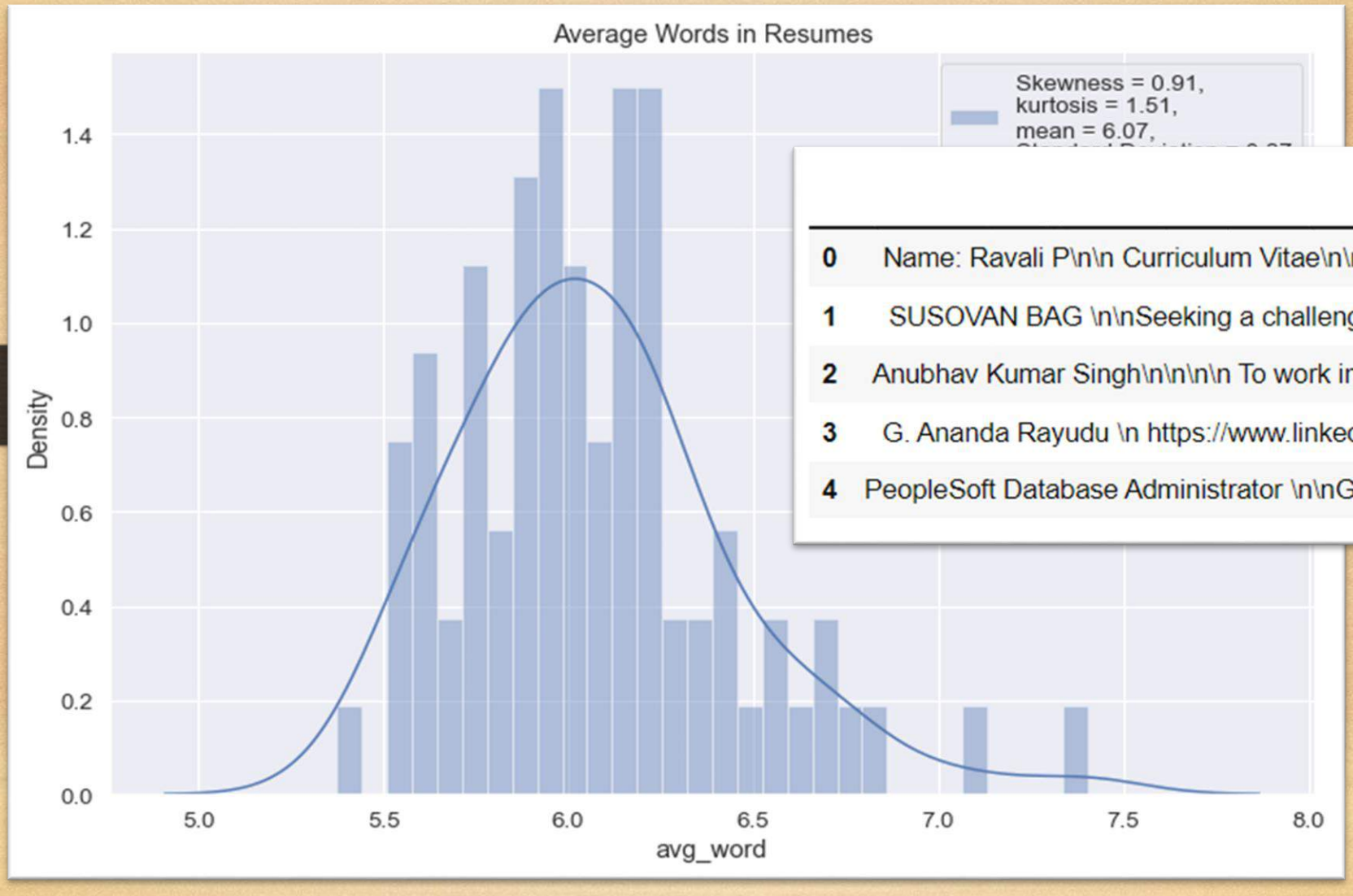
	Resumes	word_count
0	Name: Ravali P\n\n Curriculum Vitae\n\n Specia...	594
1	SUSOVAN BAG \n\nSeeking a challenging posi...	427
2	Anubhav Kumar Singh\n\n\n\n To work in a globa...	999
3	G. Ananda Rayudu \n https://www.linkedin.com/i...	1229
4	PeopleSoft Database Administrator \n\nGangared...	1115

FEATURE ENGINEERING



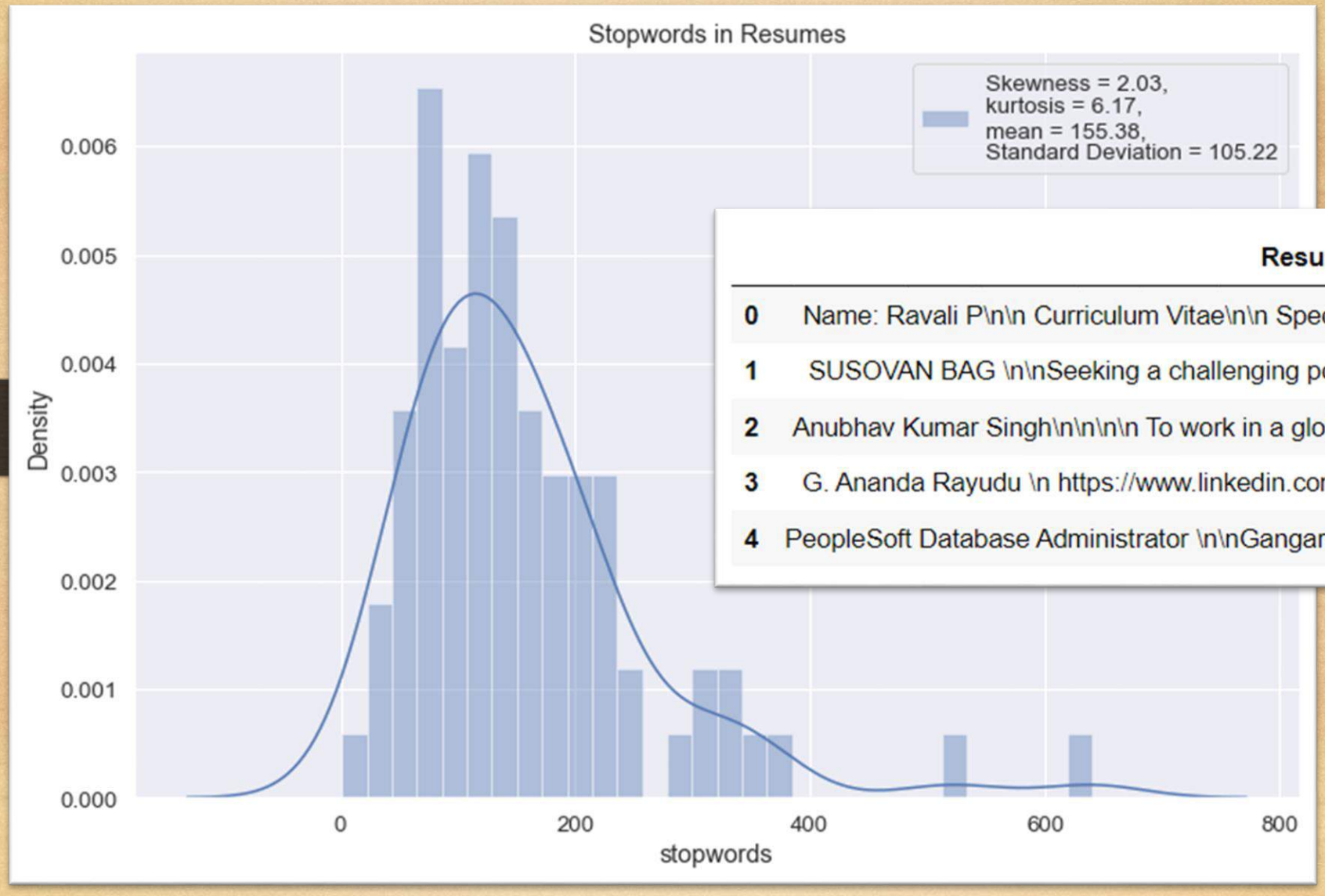
	Resumes	char_count
0	Name: Ravali P\n\n Curriculum Vitae\n\n Specia...	3541
1	SUSOVAN BAG \n\nSeeking a challenging posi...	1745
2	Anubhav Kumar Singh\n\n\n\n\n To work in a globa...	7749
3	G. Ananda Rayudu \n https://www.linkedin.com/i...	8063
4	PeopleSoft Database Administrator \n\nGangared...	6484

FEATURE ENGINEERING



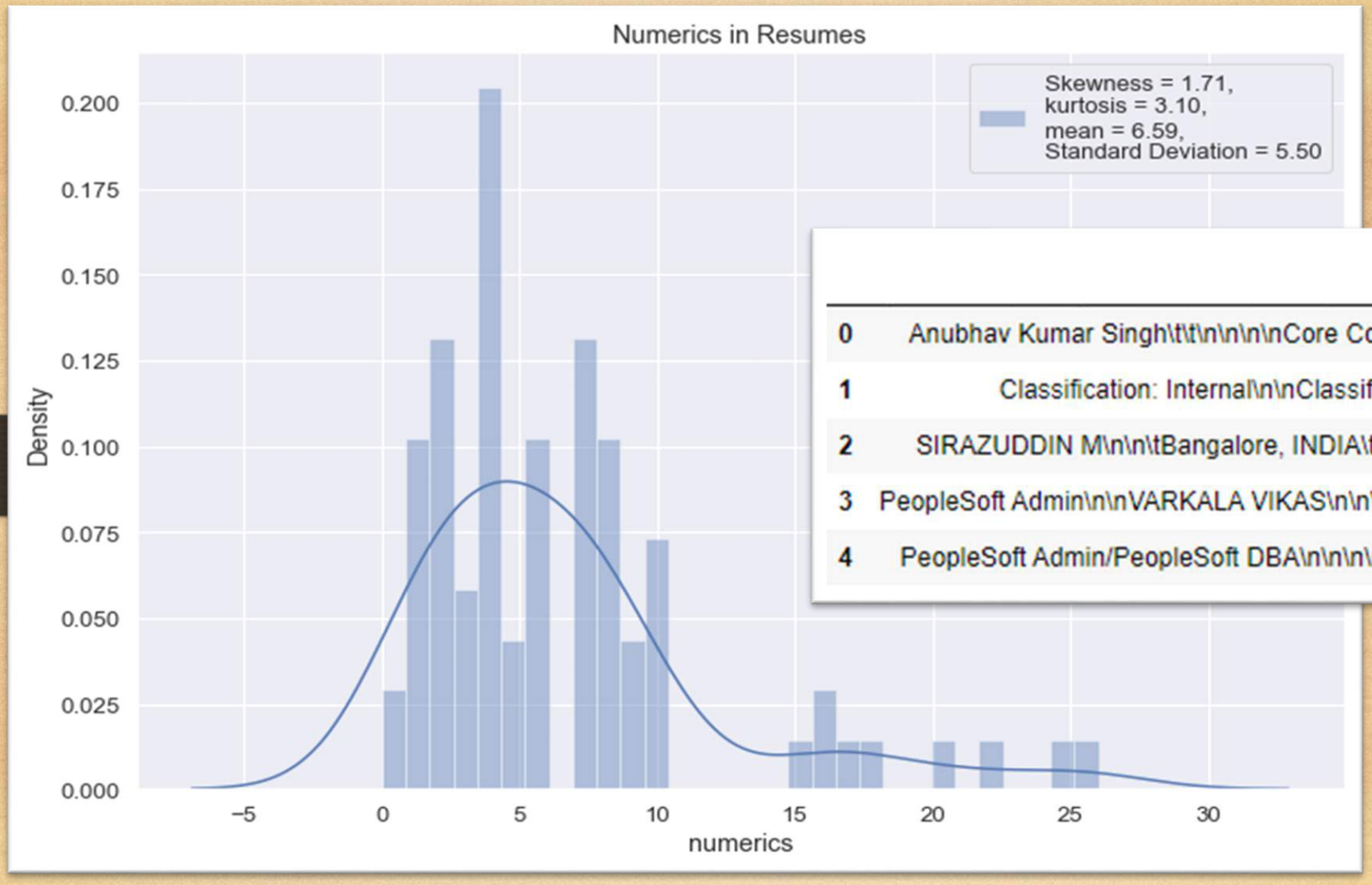
	Resumes	avg_word
0	Name: Ravali P\n\n Curriculum Vitae\n\n Specia...	7.080000
1	SUSOVAN BAG \n\nSeeking a challenging posi...	6.286432
2	Anubhav Kumar Singh\n\n\n\n To work in a globa...	6.085227
3	G. Ananda Rayudu \n https://www.linkedin.com/i...	6.114583
4	PeopleSoft Database Administrator \n\nGangared...	6.237636

FEATURE ENGINEERING



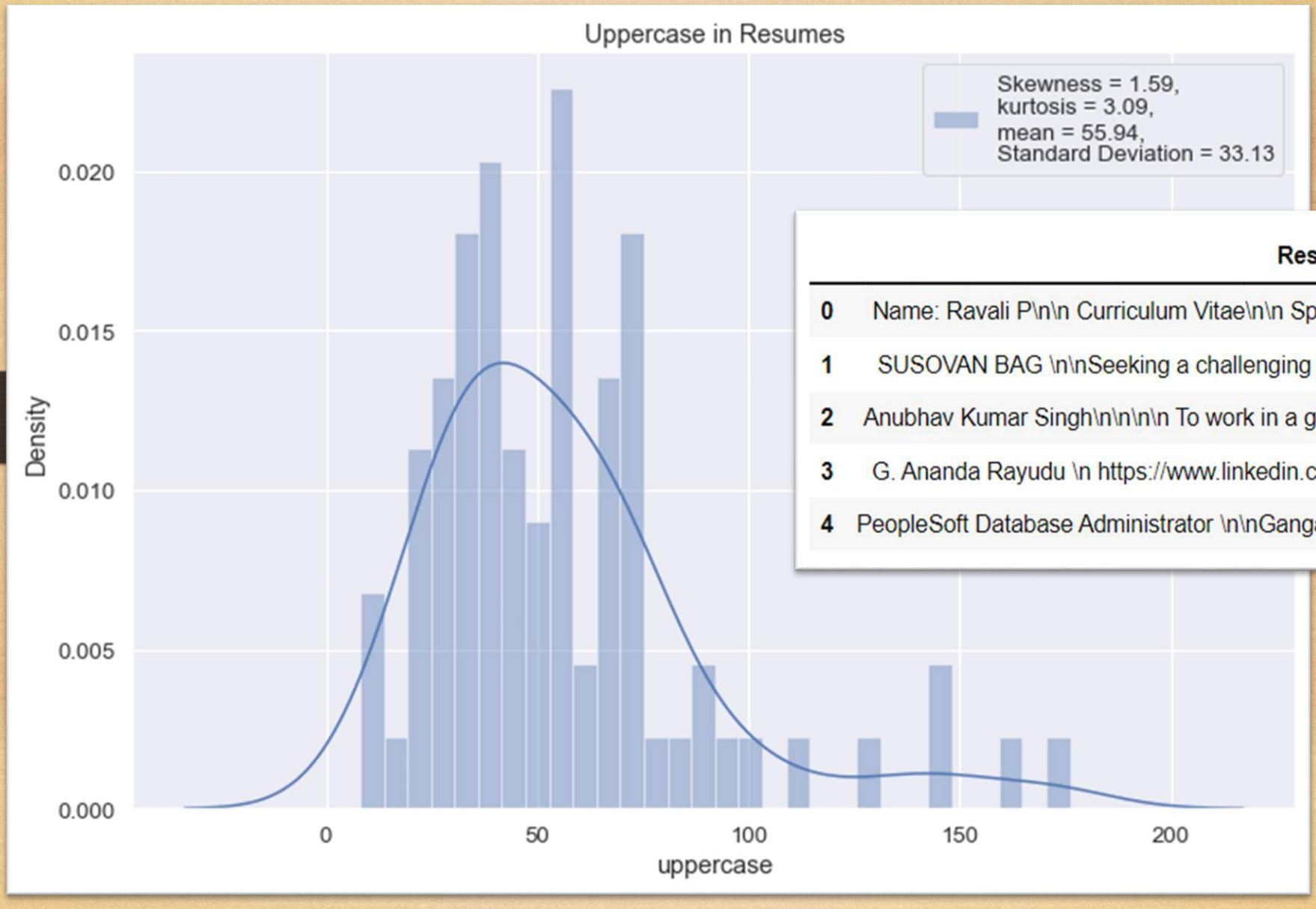
	Resumes	stopwords
0	Name: Ravali P\n\n Curriculum Vitae\n\n Specia...	65
1	SUSOVAN BAG \n\nSeeking a challenging posi...	39
2	Anubhav Kumar Singh\n\n\n\n To work in a globa...	190
3	G. Ananda Rayudu \n https://www.linkedin.com/i...	188
4	PeopleSoft Database Administrator \n\nGangared...	177

FEATURE ENGINEERING



	Resumes	numerics
0	Anubhav Kumar Singh\t\t\n\n\n\nCore Competenci...	4
1	Classification: Internal\n\nClassification: In...	7
2	SIRAZUDDIN M\n\n\tBangalore, INDIA\t\n\n\n\n\n...	15
3	PeopleSoft Admin\n\nVARKALA VIKAS\n\n\n\nCaree...	4
4	PeopleSoft Admin/PeopleSoft DBA\n\n\n\nGanesh ...	16

FEATURE ENGINEERING



	Resumes	uppercase
0	Name: Ravali P\n\n Curriculum Vitae\n\n Specia...	55
1	SUSOVAN BAG \n\nSeeking a challenging posi...	37
2	Anubhav Kumar Singh\n\n\n\n To work in a globa...	57
3	G. Ananda Rayudu \n https://www.linkedin.com/i...	72
4	PeopleSoft Database Administrator \n\nGangared...	72

FEATURE ENGINEERING

Converting Extracted Above Data into a Data-Frame
To use this as Features (Predictors, Attributes or Input) for Model to Predict the different Classes

	Profiles	word_count	char_count	avg_word	stopwords	numerics	uppercase
0	Peoplesoft	1023	8010	6.415929	190	4	60
1	Peoplesoft	558	4917	6.317817	114	7	53
2	Peoplesoft	455	3800	6.109162	86	15	77
3	Peoplesoft	918	7943	6.575787	225	4	58
4	Peoplesoft	1641	11121	6.429279	303	16	85
...	...	...	...	...	...	...	...
72	Workday	904	7030	6.442391	197	3	57
73	Workday	1063	6737	5.995570	194	9	85
74	Workday	1076	8329	6.030810	236	6	69
75	Workday	1031	6836	6.272418	193	7	46
76	Workday	702	5126	6.223201	150	2	34

77 rows × 7 columns

TEXT PRE-PROCESSING

Text pre-processing includes converting to lowercase, removing spaces, html links, emails, symbols, numbers, stop-words, tokenization and lemmatization.

Removing All Unwanted Character's

	Resumes	Profiles	Clean_Resumes
0	Name: Ravali P\n\n Curriculum Vitae\n\n Specia...	Internship	name ravali curriculum vitae specialization co...
1	SUSOVAN BAG \n\nSeeking a challenging posi...	Internship	susovan bag seeking challenging position field...
2	Anubhav Kumar Singh\n\n\n\n To work in a globa...	Peoplesoft Admin	anubhav kumar singh work globally competitive ...
3	G. Ananda Rayudu \n https://www.linkedin.com/i...	Peoplesoft Admin	ananda rayudu profile summary years experience...
4	PeopleSoft Database Administrator \n\nGangared...	Peoplesoft Admin	peoplesoft database administrator gangareddy p...
...	...	...	...
74	Workday Integration Consultant \n\nName ...	Workday	workday integration consultant name sri krishn...
75	S R I K A N T H (W O R K D A Y H C M C O N...	Workday	seeking suitable positions workday hcm techno ...
76	WORKDAY | HCM | FCM \n\nName Role \n\n: Kumar ...	Workday	workday hcm fcm name role kumar workday consul...
77	Venkateswarlu.B \n\n\n\nWorkday Consultant\n\n...	Workday	venkateswarlu workday consultant professional ...
78	VINAY KUMAR .V \nWORKDAY FUNCTIONAL CONSULTANT...	Workday	vinay kumar workday functional consultant expe...

79 rows × 3 columns

Word Tokenization - Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens.

Removing Stop-words - A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

TEXT PRE-PROCESSING :

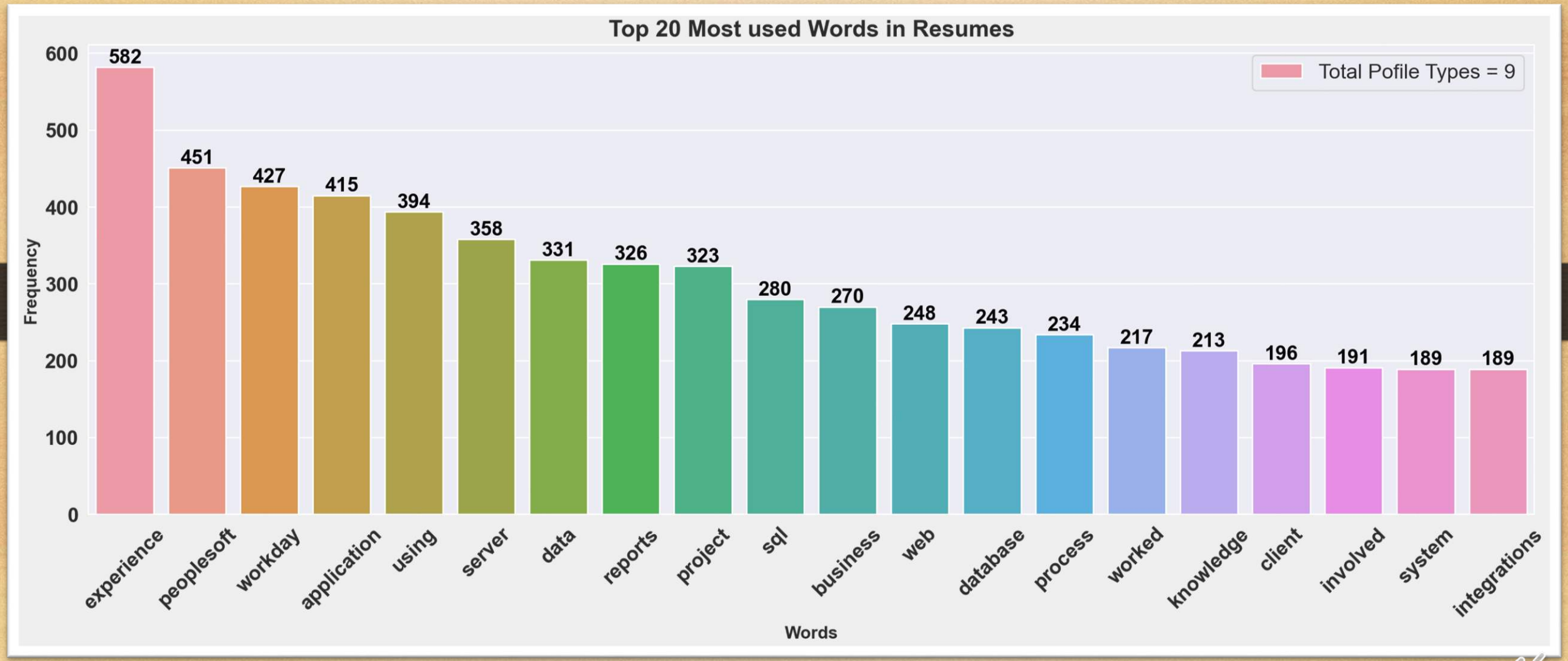
Before Text pre-processing

'G. Ananda Rayudu \n https://www.linkedin.com/in/anandgudapati\n\n\n\n\n\nProfile Summary: \n\n•\n\n•\n\n•\n\n•\n\n•\n\n7+ yea
rs of experience in implementing, upgrading, and supporting PeopleSoft database administration including Human Capital Manageme
nt (HCM), Financials, Campus Solutions and Portal IHUB \nExpertise in installation, configuration, setup, security and manageme
nt of the PeopleSoft Internet Architecture PIA environment. \n\nIn-depth experience in analysis, planning, development, and imp
lementation stages, including load testing, quality assurance, and tuning. \n\nGained extensive exposure in deploying PeopleSof
t Environments. \n\nExperienced in Troubleshooting PeopleSoft Components. \n\nSkilled with the capability to analyse & interpre
t unique problems with a combination of training experience & logical thinking to find the right solutions \n\n\n\nCore Compete
ncies: \n\n➢PeopleSoft I\n➢Troubleshooting ➢Project /Data\n➢Installation &\n➢PUM (DPK) \n➢Upgrades – P\n➢Expertise in a\n
➢Elastic Searc\n➢PeopleSoft R\n➢Integration B\n➢Maintaining \n➢PeopleSoft S\n➢Installing Mi\n➢Configuring \n➢PeopleSoft C
\n➢AWS-EC2. \n\n➢Cloud Manag\n\n‖ Work Experience: \n\nOrganization \n\nPROJECT #1 Multiple clients \n\n \n\npl
ementation. \n\n\ting \t PeopleSoft performance issues. \n\n\t Migration \n Configuration of PeopleSoft components. \n\nInstall
and configure. \n\neople tools. \n\n\t➢Expertise in applying Patches, and Updates via Change Assistant tool including Tax Updat
es. \th Install and configure. \n\nefreshes / Cloning. \n\nroker setup and configuration \n➢Maintaining workflow of PeopleSoft
users & monitoring log files and search bottleneck of servers. ➢PeopleSoft Security resetting passwords and locking/unlocking u
ser profiles. \n\n➢Installing Middle tier components and Oracle quarterly Security patches. \n\nChange Assistant to apply fixes
and patch sets. \n\nloud Infrastructure-IaaS. \n\ner-Lift and Shift of application to cloud. \n\n : IDC Techn
ologies Sol. (I) Pvt. Ltd. \n\n: Texas Department of Transportation(TxDOT) \n\nDURATION \nROLE \nENVIRONMENT \n\n\t:

After Text pre-processing

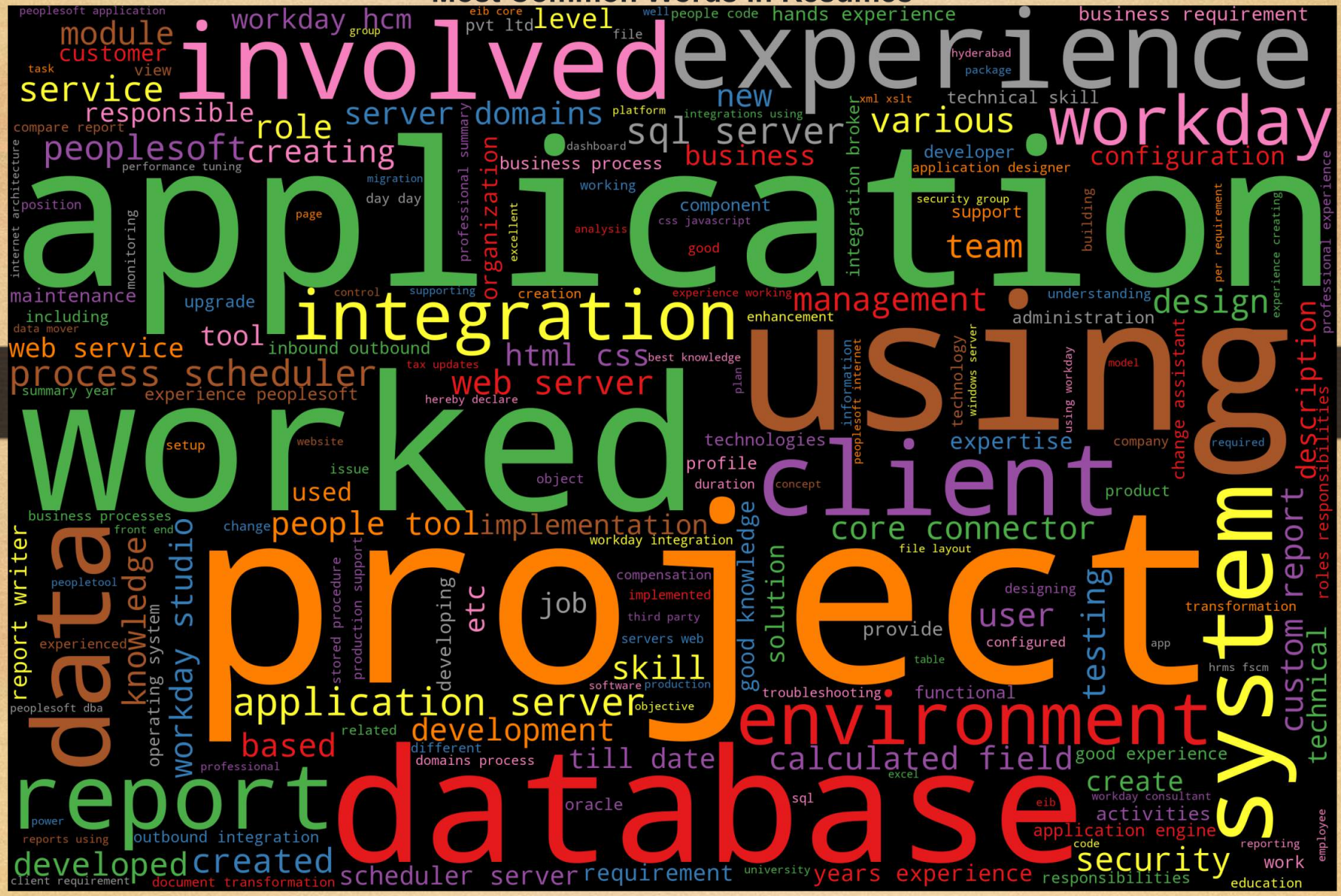
'ananda rayudu profile summary years experience implementing upgrading supporting peoplesoft database administration including
human capital management hcm financials campus solutions portal ihub expertise installation configuration setup security manage
ment peoplesoft internet architecture pia environment depth experience analysis planning development implementation stages incl
uding load testing quality assurance tuning gained extensive exposure deploying peoplesoft environments experienced troubleshoo
ting peoplesoft components skilled capability analyse interpret unique problems combination training experience logical thinkin
g find right solutions core competencies peoplesoft troubleshooting project data installation pum dpk upgrades expertise elasti
c searc peoplesoft integration maintaining peoplesoft installing configuring peoplesoft aws cloud manag work experience organiz
ation project multiple clients plementation ing peoplesoft performance issues migration configuration peoplesoft components ins
tall configure eople tools expertise applying patches updates via change assistant tool including tax updates install configure
efreshes cloning roker setup configuration maintaining workflow peoplesoft users monitoring log files search bottleneck servers
peoplesoft security resetting passwords locking unlocking user profiles installing middle tier components oracle quarterly secu
rity patches change assistant apply fixes patch sets loud infrastructure iaas lift shift application cloud idc technologies sol
pvt ltd texas department transportation txdot duration role environment aug till date peoplesoft dba people tools hrms fscm tux

EXPLORATORY DATA ANALYSIS



WORD CLOUD

Most Common Words in Resumes

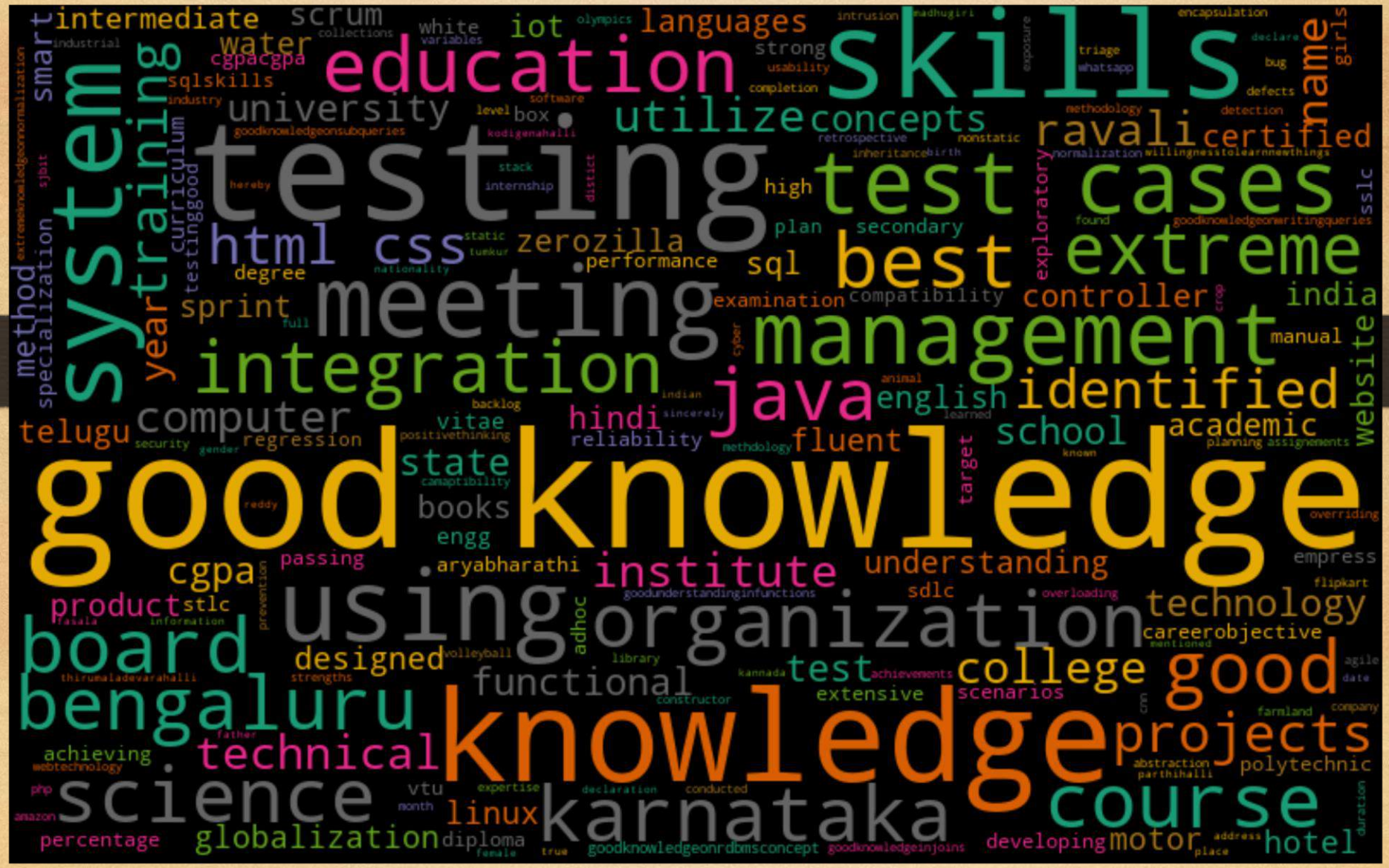


EXPLORATORY DATA ANALYSIS

10 most common words used in each Profile Resumes

EXPLORATORY DATA ANALYSIS

Most Common Words in Internship Resumes

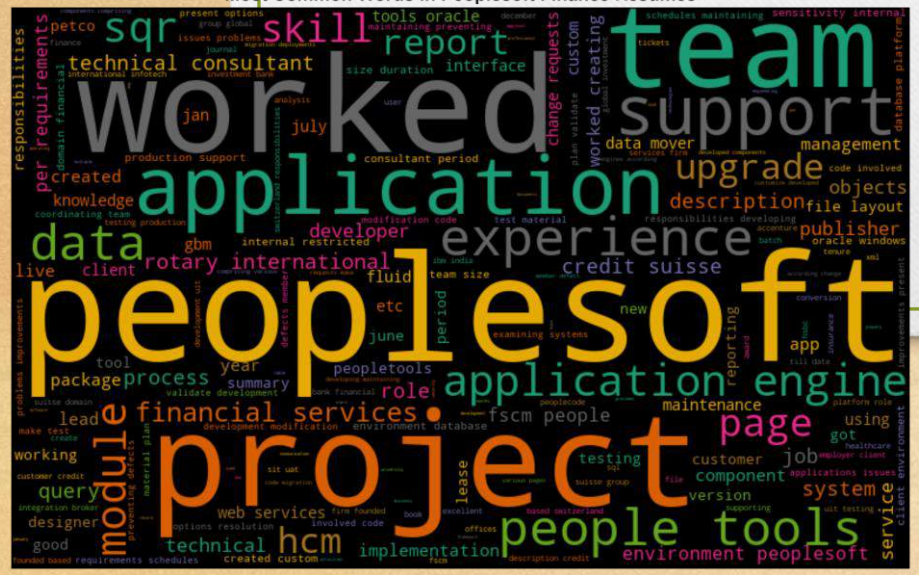


EXPLORATORY DATA ANALYSIS

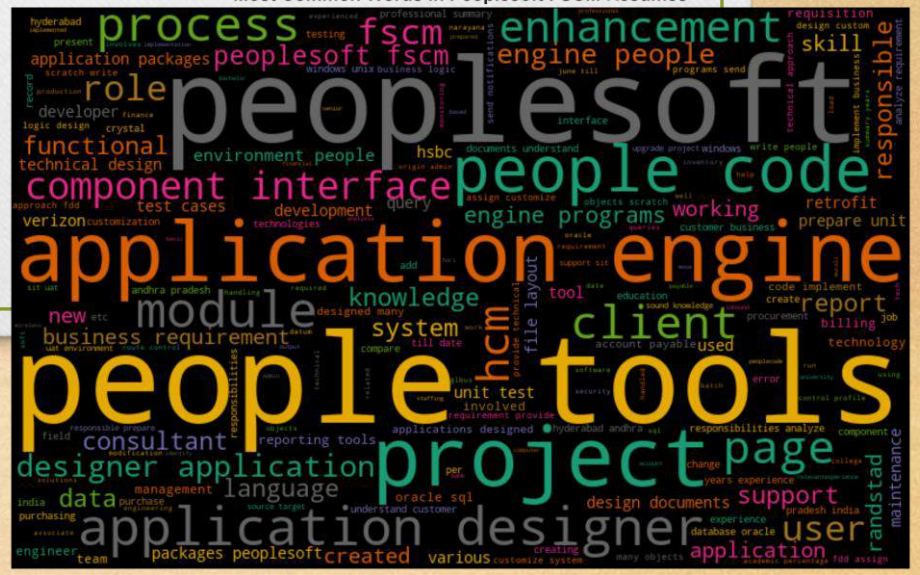
Most Common Words in Peoplesoft Admin Resumes

Most Common Words in Peoplesoft DBA Resumes

Most Common Words in Peoplesoft Finance Resumes



Most Common Words in Peoplesoft FSCM Resumes



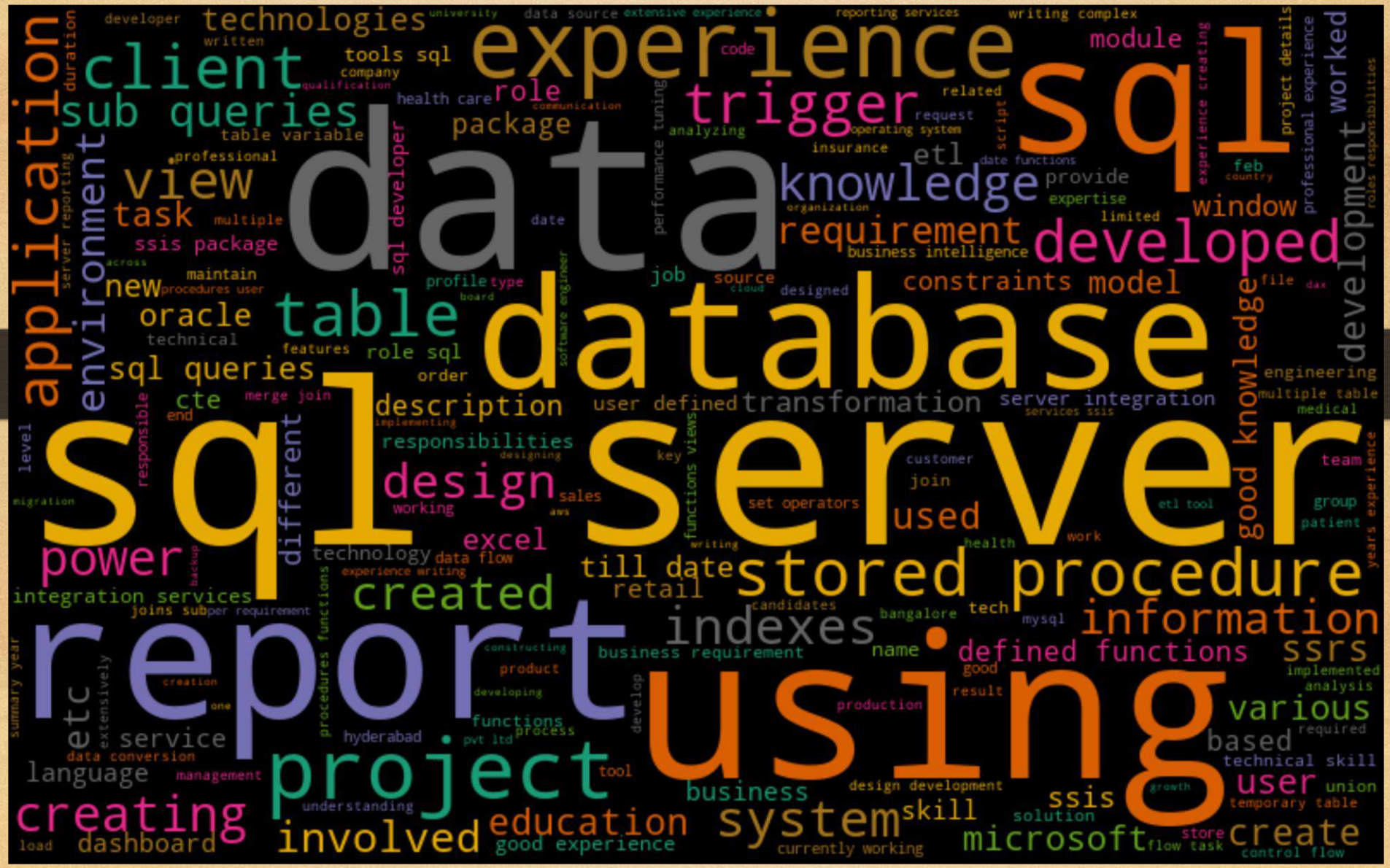
EXPLORATORY DATA ANALYSIS

Most Common Words in React Developer Resumes

Most Common Words in React JS Developer Resumes

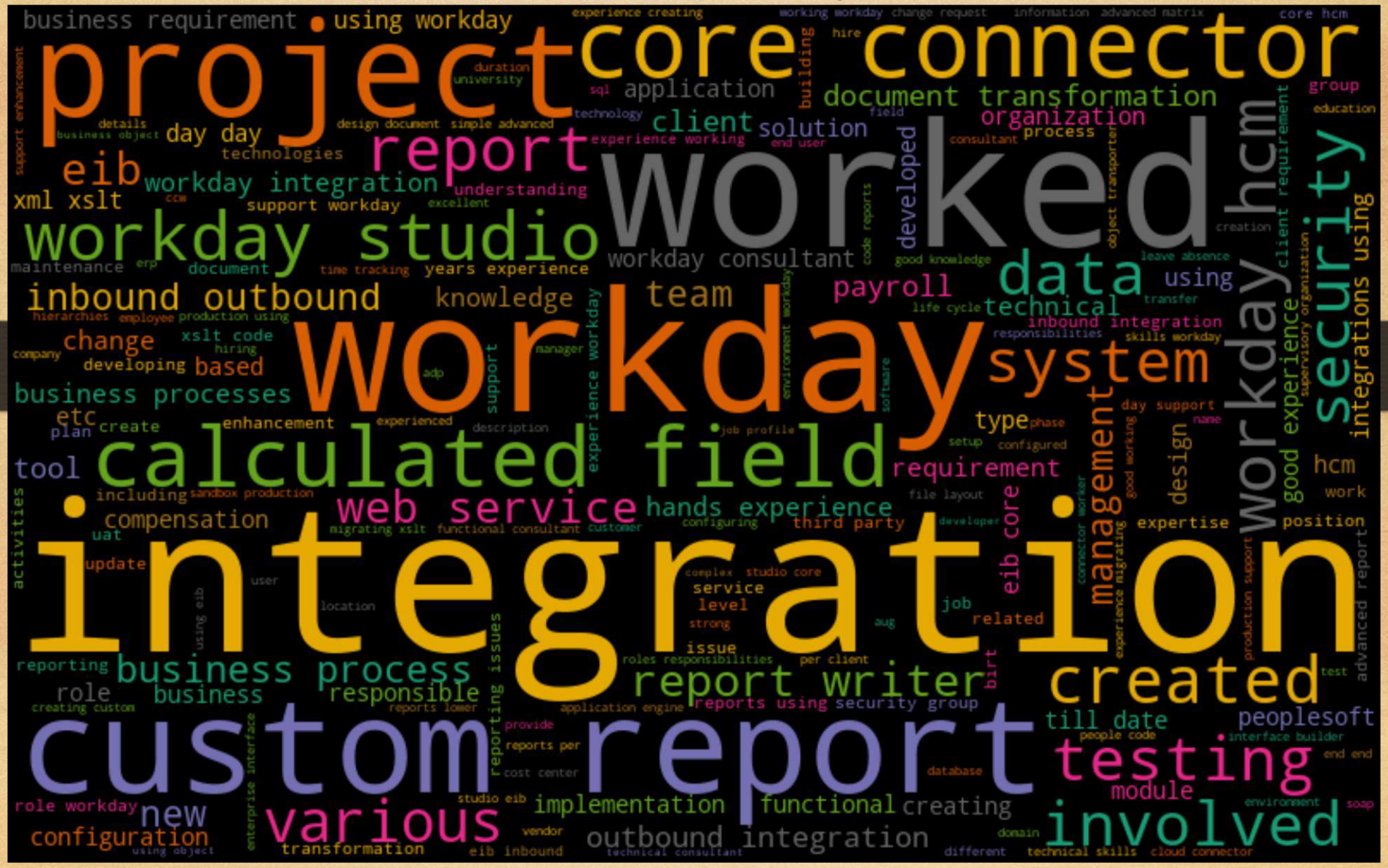
EXPLORATORY DATA ANALYSIS

Most Common Words in SQL Developer Resumes



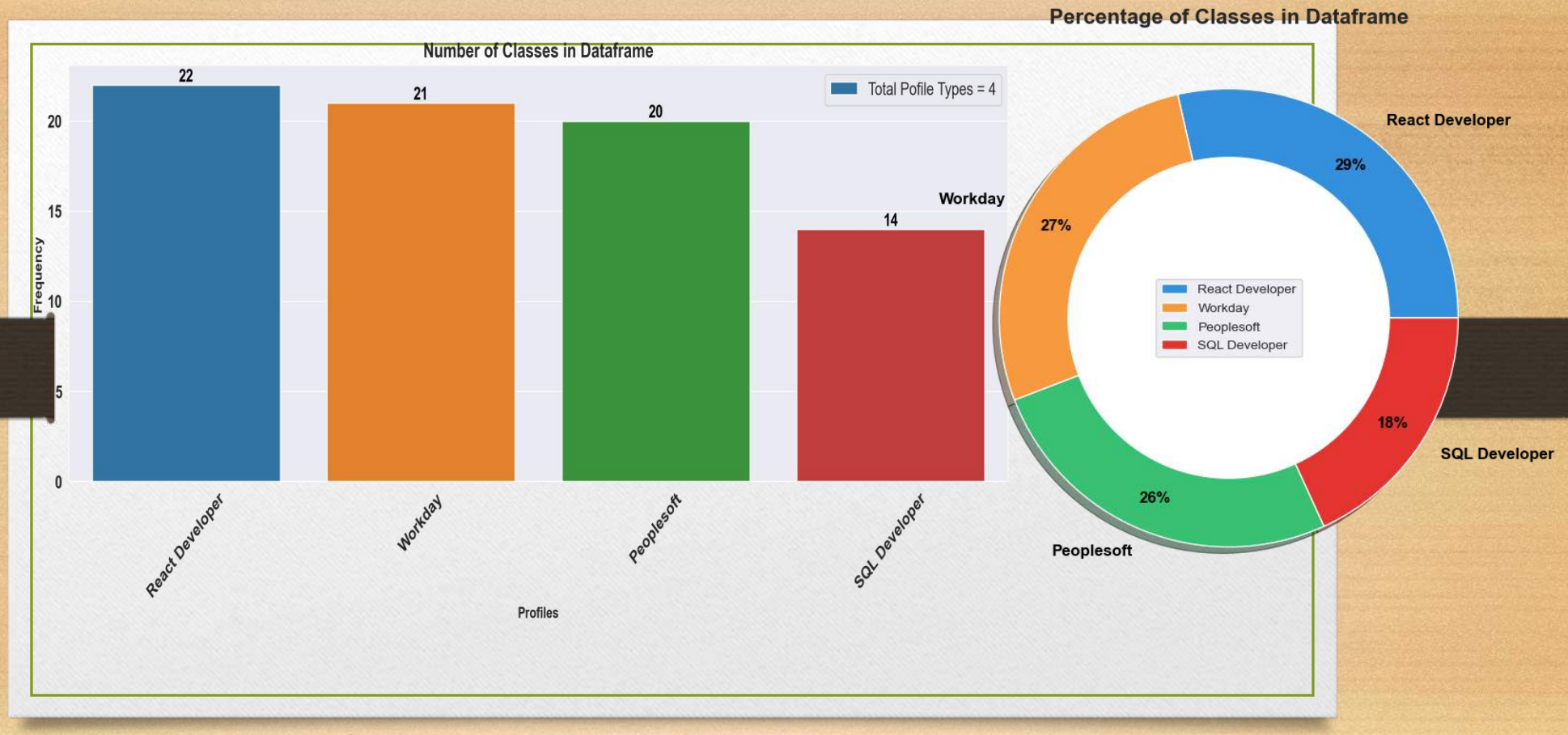
EXPLORATORY DATA ANALYSIS

Most Common Words in Workday Resumes



EXPLORATORY DATA ANALYSIS

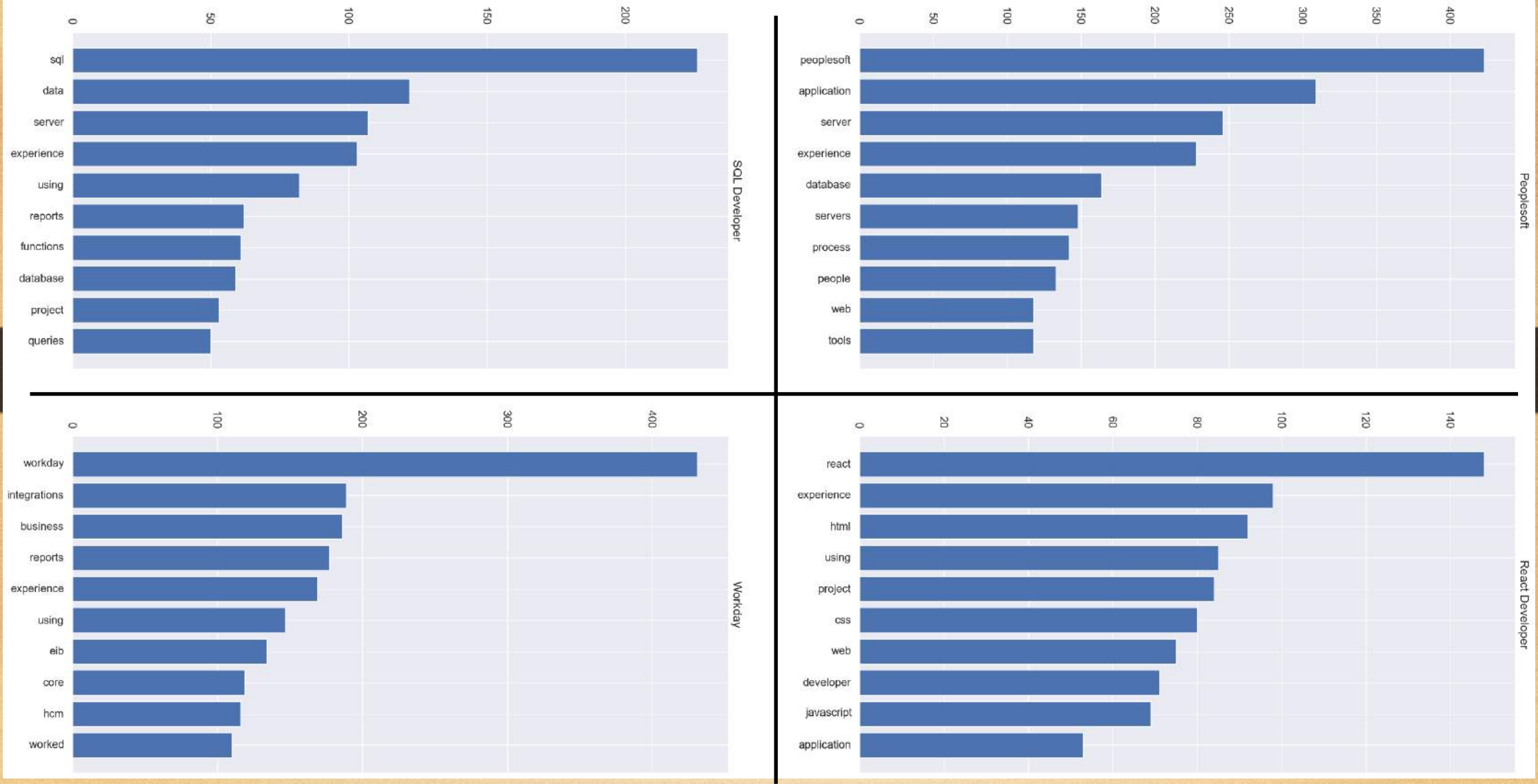
Classes in the Data-Frame
Plotting Classes for Insights



There are Total 4 Classes in the Data Frame which means this a Multiclass Classification Problem.
Imbalance found in the dataset we can use Oversampling Techniques.

EXPLORATORY DATA ANALYSIS

10 Most Common Words Used in Different Classes



FEATURE ENGINEERING

Count Vectorizer
with N-grams (Bigrams & Trigrams)

	abil	abil work	abil work team	absenc	absenc manag	academ	accentur	accept	accept level	accept level statu	...	year experi workday	year extens	year extens experi	year industri	year industri role	year month	year pass	year profession
0	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	1	0
1	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
74	1	1	1	2	1	0	0	1	0	0	...	0	0	0	0	0	0	0	1
75	1	0	0	7	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0
76	0	0	0	1	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0
77	1	0	0	2	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
78	1	1	1	1	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0

79 rows × 5000 columns

FEATURE ENGINEERING

TF-IDF Vectorizer
with N-grams (Bigrams & Trigrams)

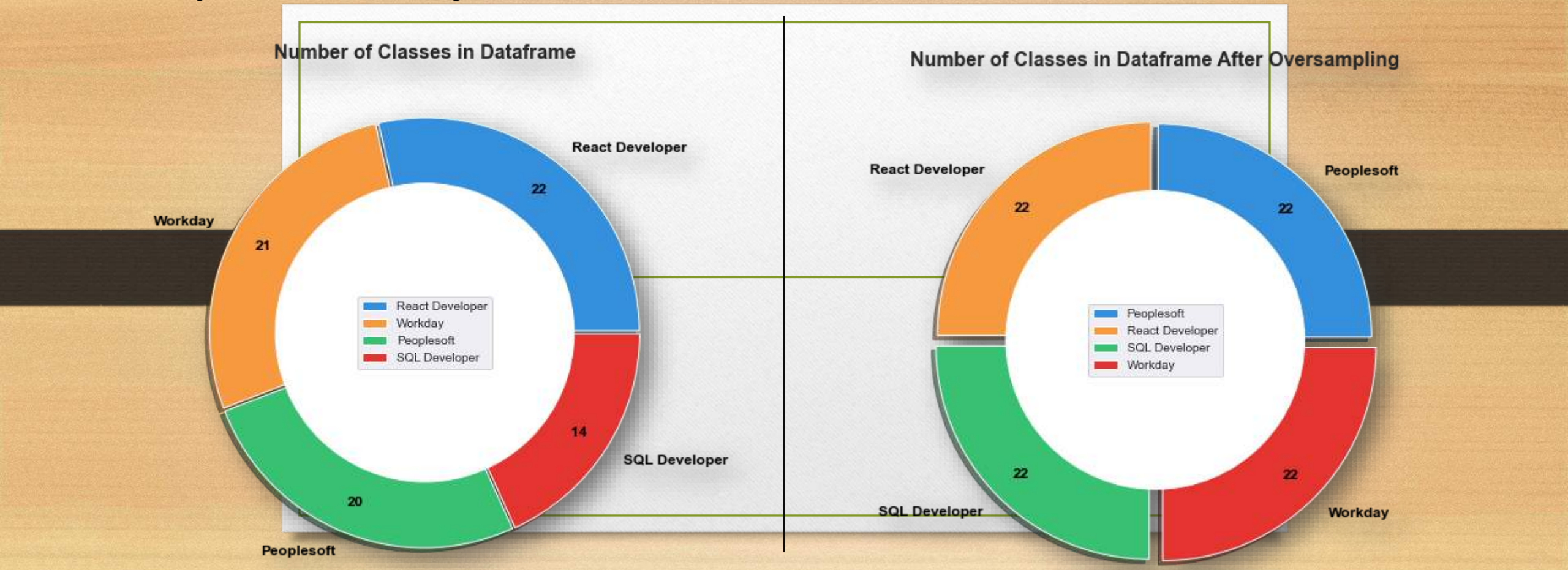
	abil	abil work	abil work team	absenc	absenc manag	academ	accentur	accept	accept level	accept level statu	...	year experi workday	year extens	year extens experi	year industri	year industri role	year month
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.038269	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.079691	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.018711	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
74	0.012106	0.016893	0.021707	0.036605	0.021707	0.000000	0.0	0.019454	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
75	0.014051	0.000000	0.000000	0.148696	0.000000	0.000000	0.0	0.022579	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
76	0.000000	0.000000	0.000000	0.015341	0.000000	0.000000	0.0	0.000000	0.0	0.0	...	0.016306	0.0	0.0	0.0	0.0	0.0
77	0.013449	0.000000	0.000000	0.040664	0.000000	0.000000	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
78	0.016005	0.022334	0.028699	0.024198	0.028699	0.000000	0.0	0.000000	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0

79 rows × 5000 columns

FEATURE ENGINEERING

SMOTE: Synthetic Minority Oversampling Technique

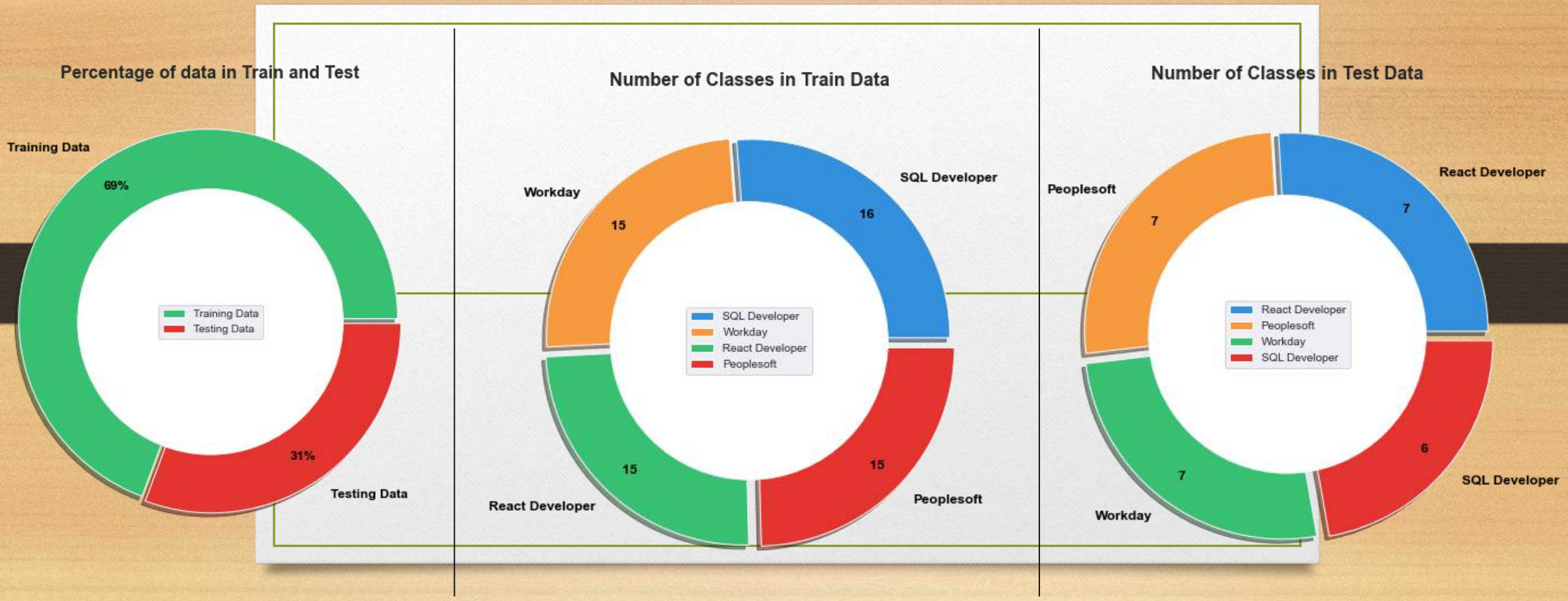
SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.



TRAIN TEST SPLIT

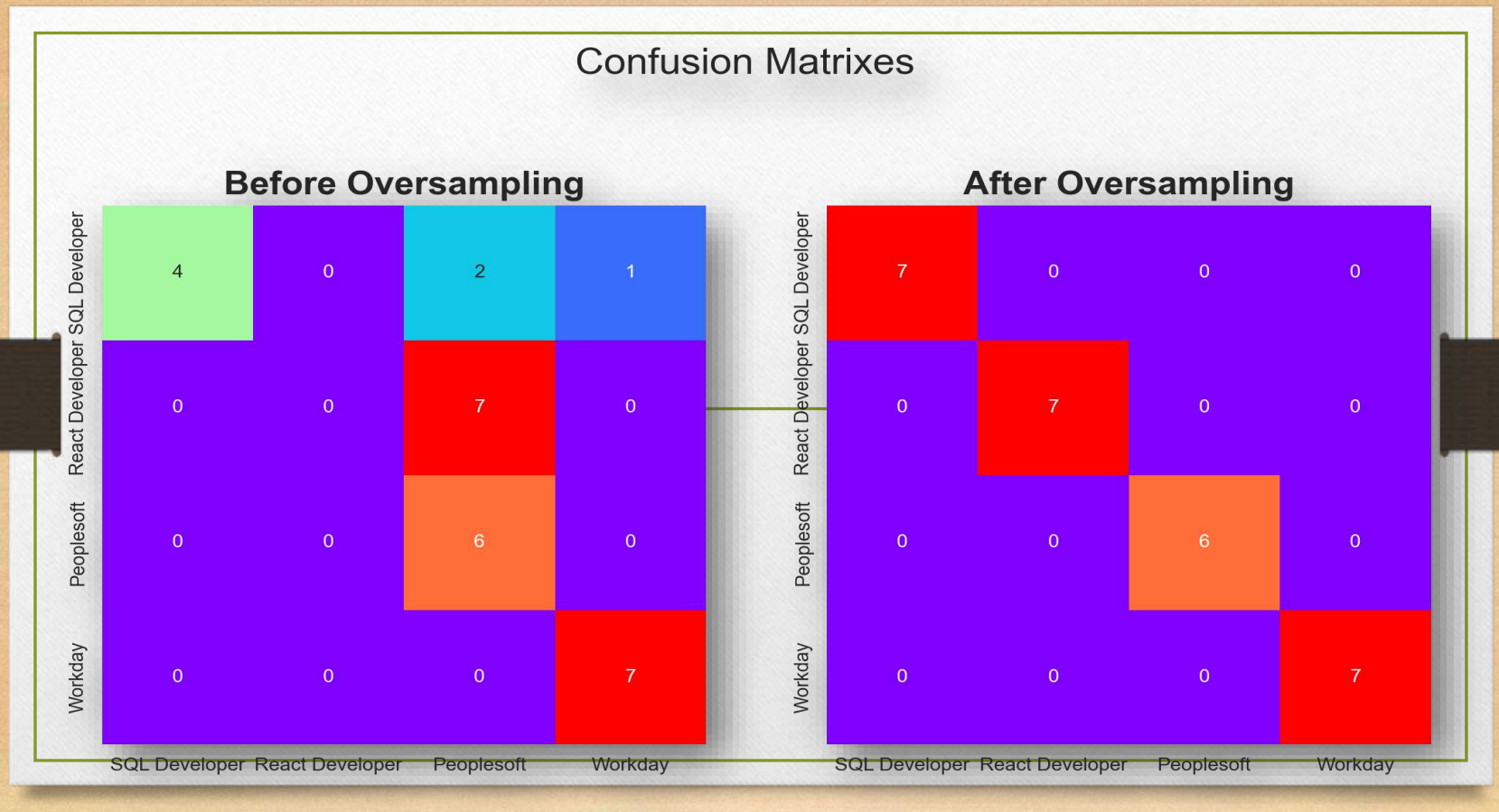
Stratified Sampling

In stratified Sampling the ratio of all the classes is maintained on both training and testing data thus this type of Split results in good accuracy and overall model building performance.



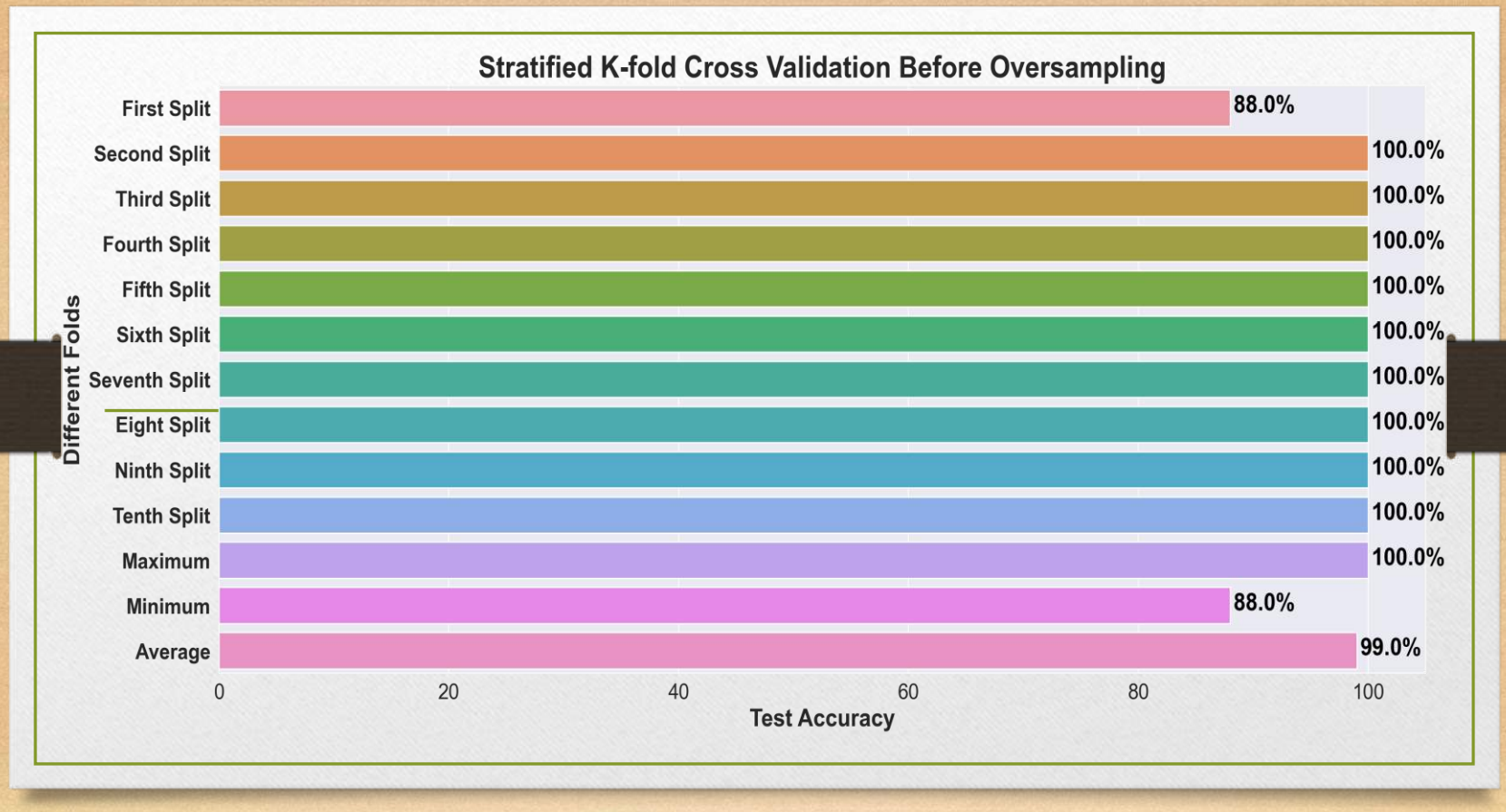
FEATURE ENGINEERING:

Sometimes when the records of a certain class are much more than the other class, our classifier may get biased towards the prediction. In this case, the confusion matrix for the classification problem shows how well our model classifies the target classes and we arrive at the accuracy of the model from the confusion matrix.



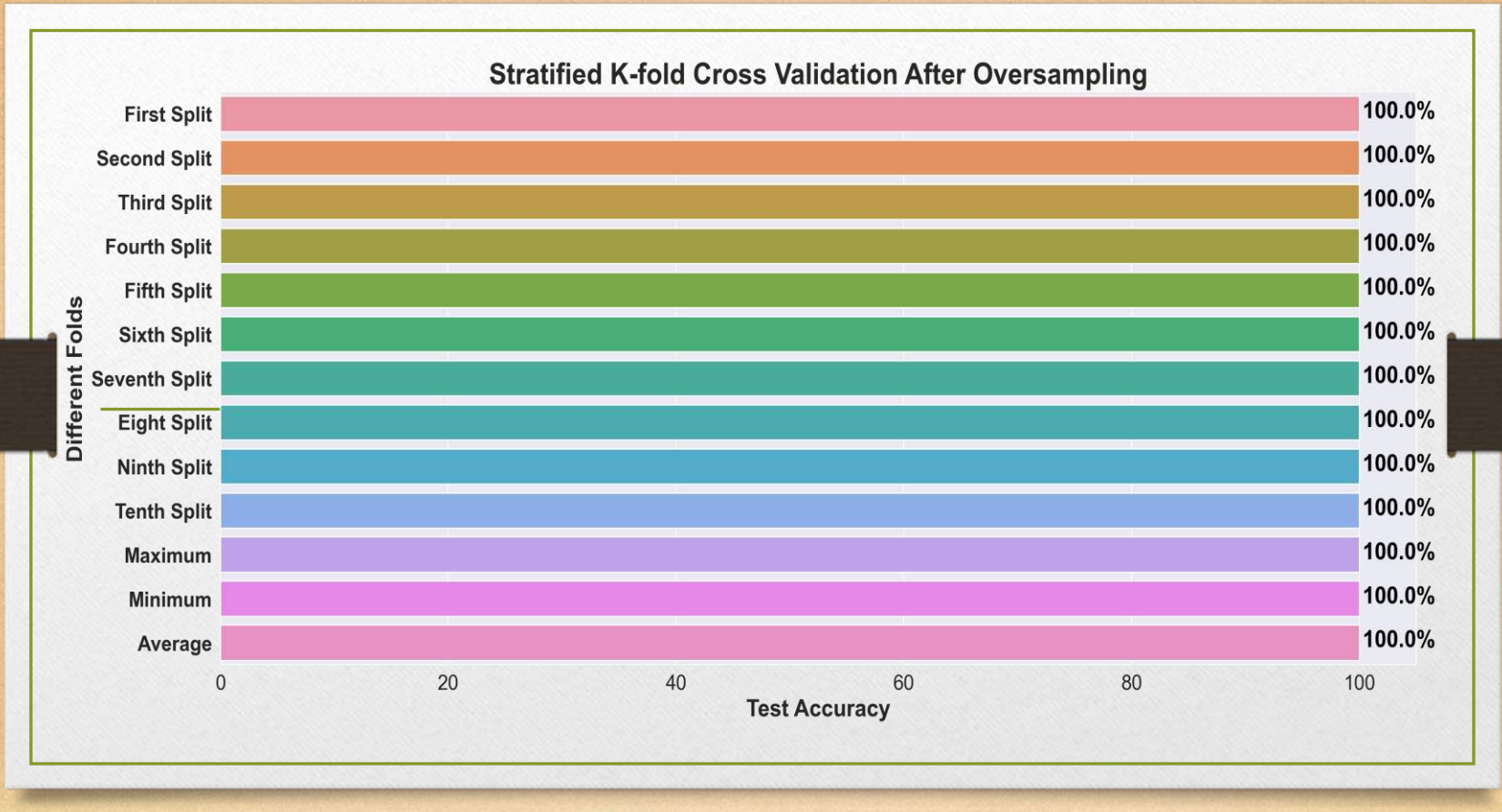
MODEL BUILDING

If we do random sampling to split the dataset into training set and test set. Then we might get a majority of one of the class in training and minority of other in testing. If we train our model obviously we will be getting bad evaluation scores.

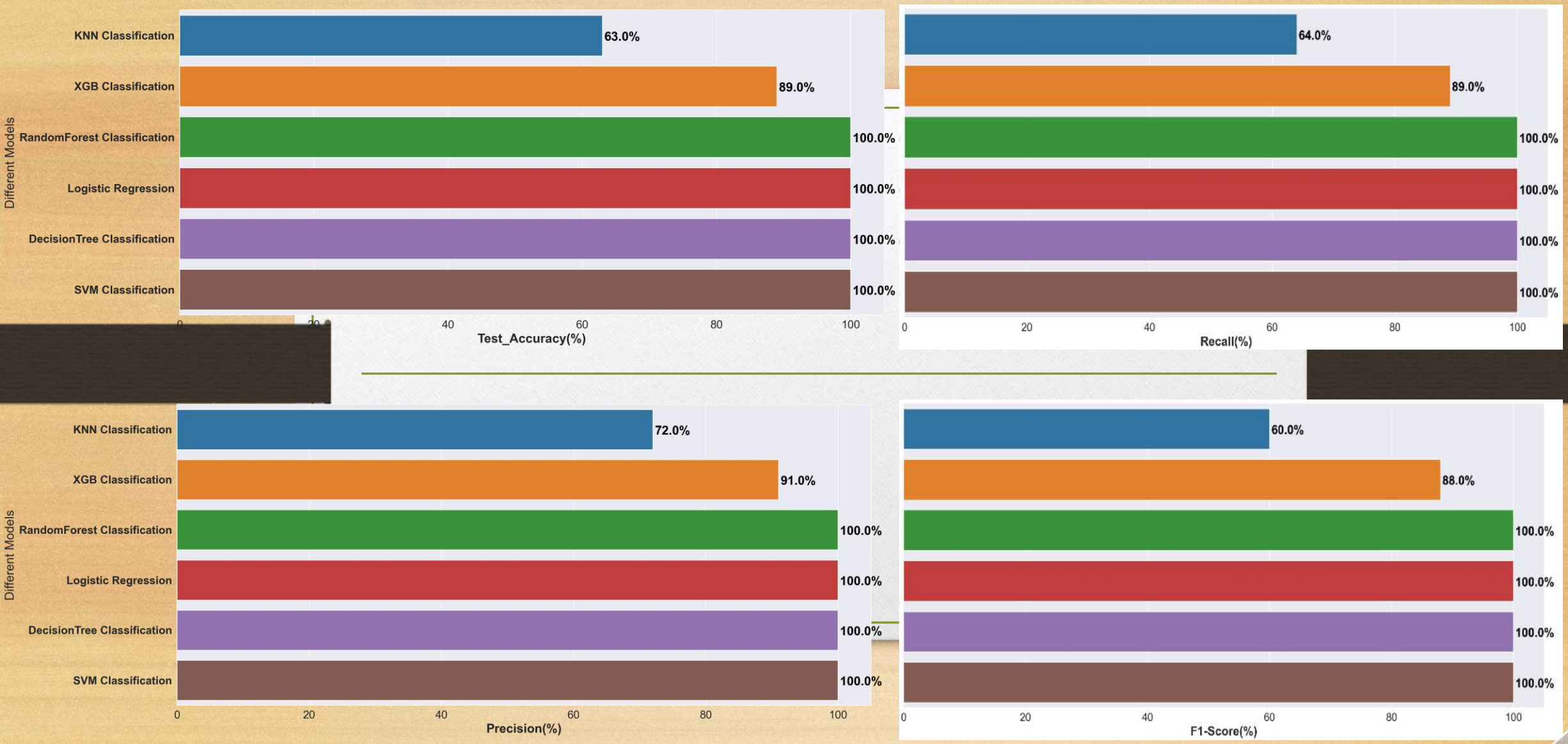


MODEL BUILDING

Stratified k-fold cross-validation is the same as just k-fold cross-validation, But Stratified k-fold cross-validation, it does stratified sampling instead of random sampling.



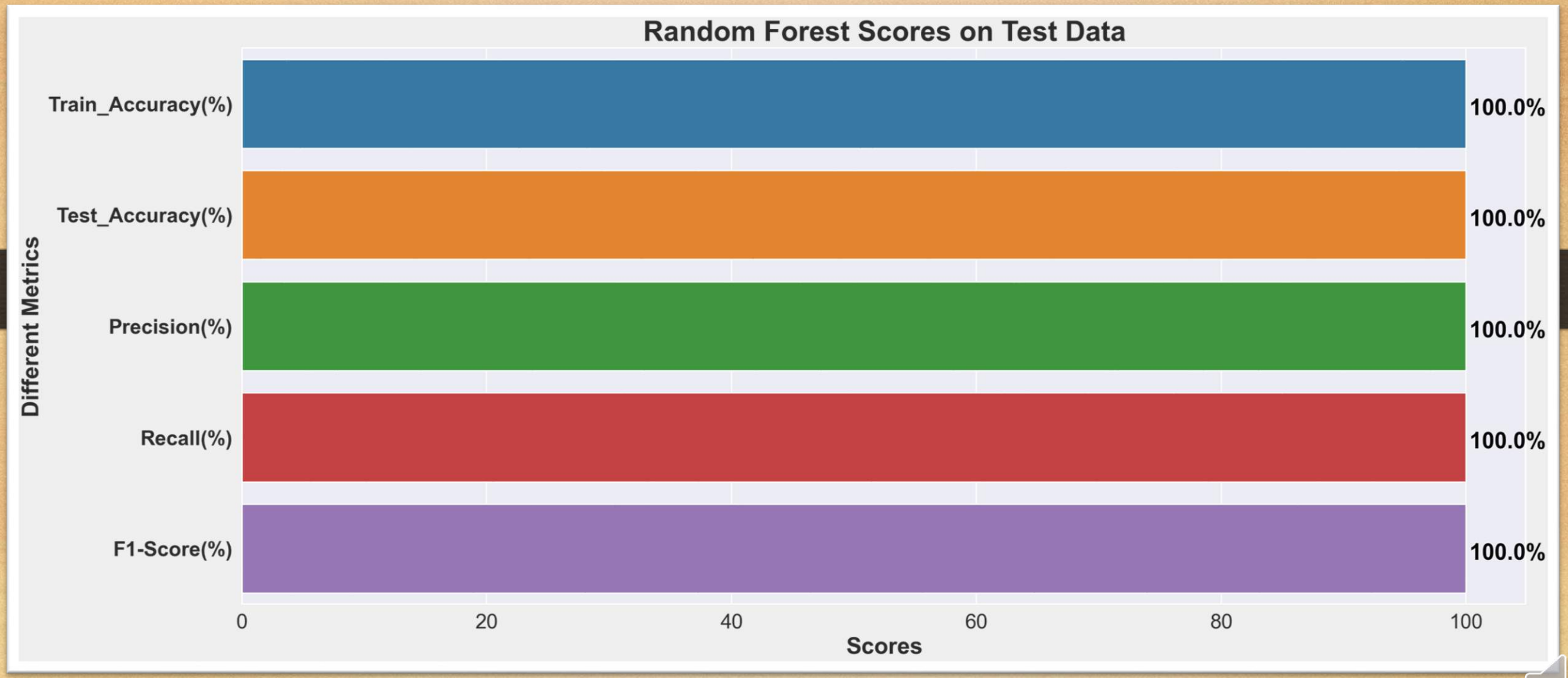
MODEL EVALUATION



MODEL EVALUATION

MODEL SELECTION

Random Forest Classification Model has 100% Accuracy on Test as well on Training Dataset.
0% Error . 100% Recall , good Precision and F1-Score. Here we can see, as there is no Overfitting.



DEPLOYMENT

RESUME CLASSIFICATION

Resume Details | Resume Classification | Model Evaluation | Data Analysis

RESUME Details

Upload Resume (Single File Accepted) 👇

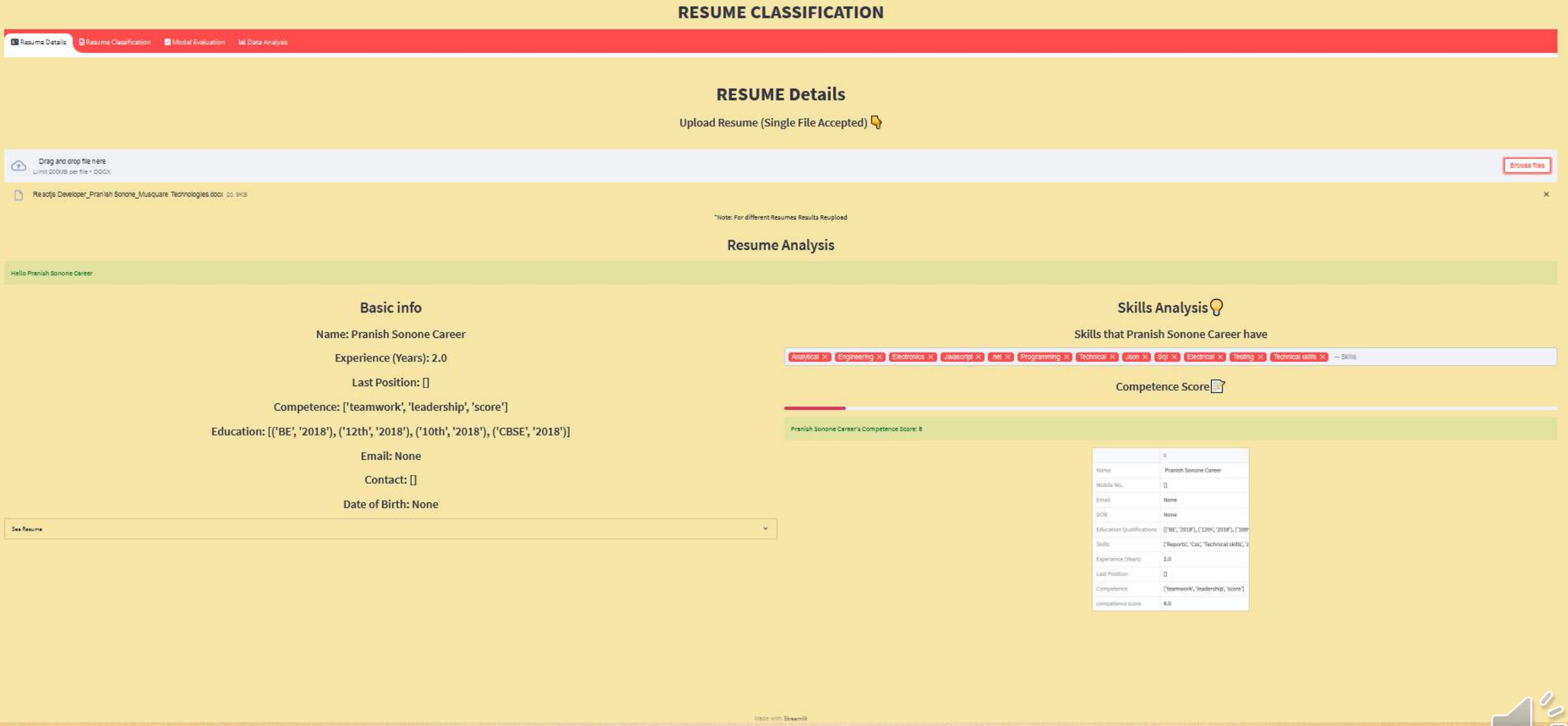
Drag and drop file here
Limit 200MB per file • DOCX

Browse files

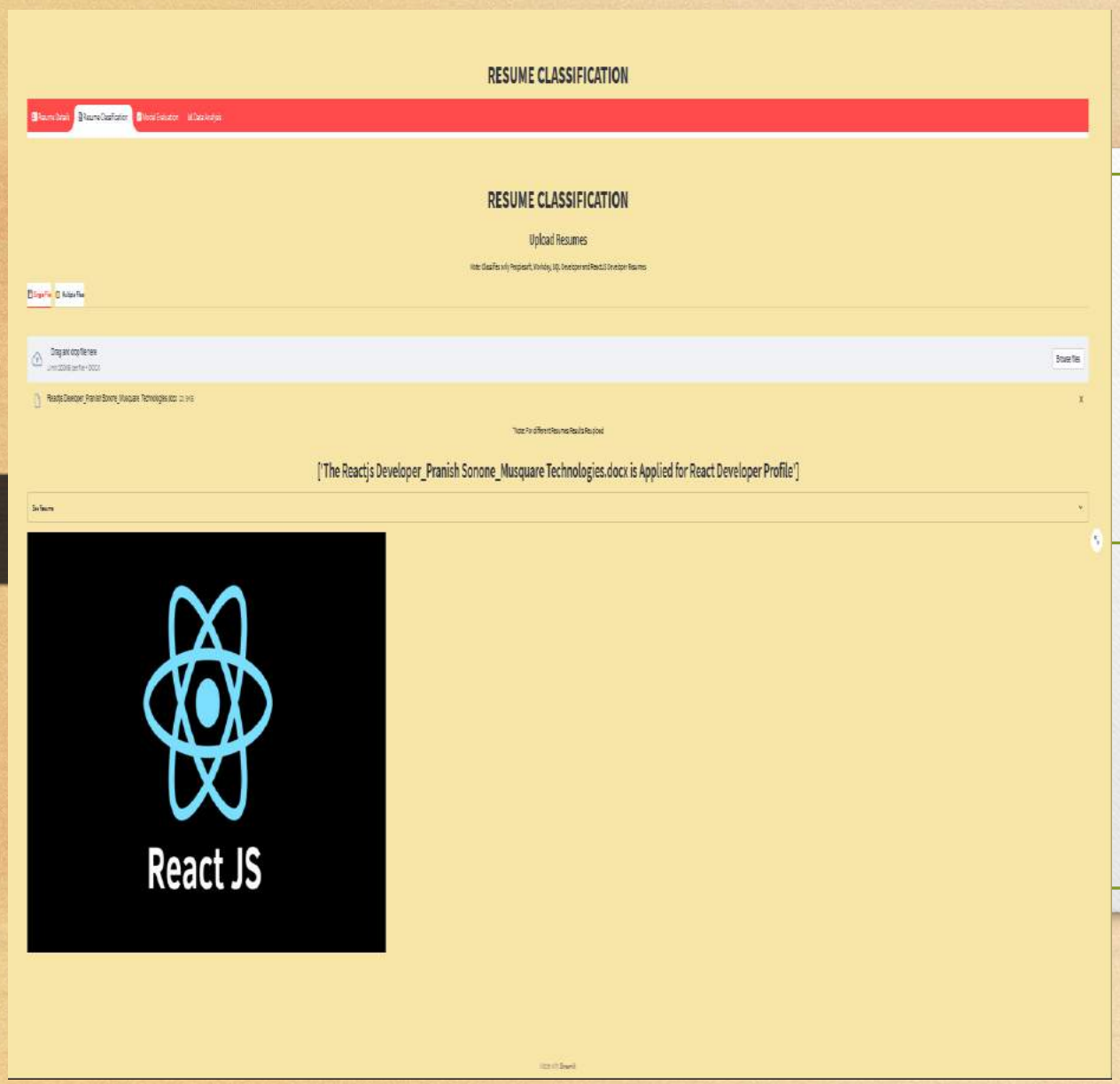
*Note: For different Resumes Results Reupload

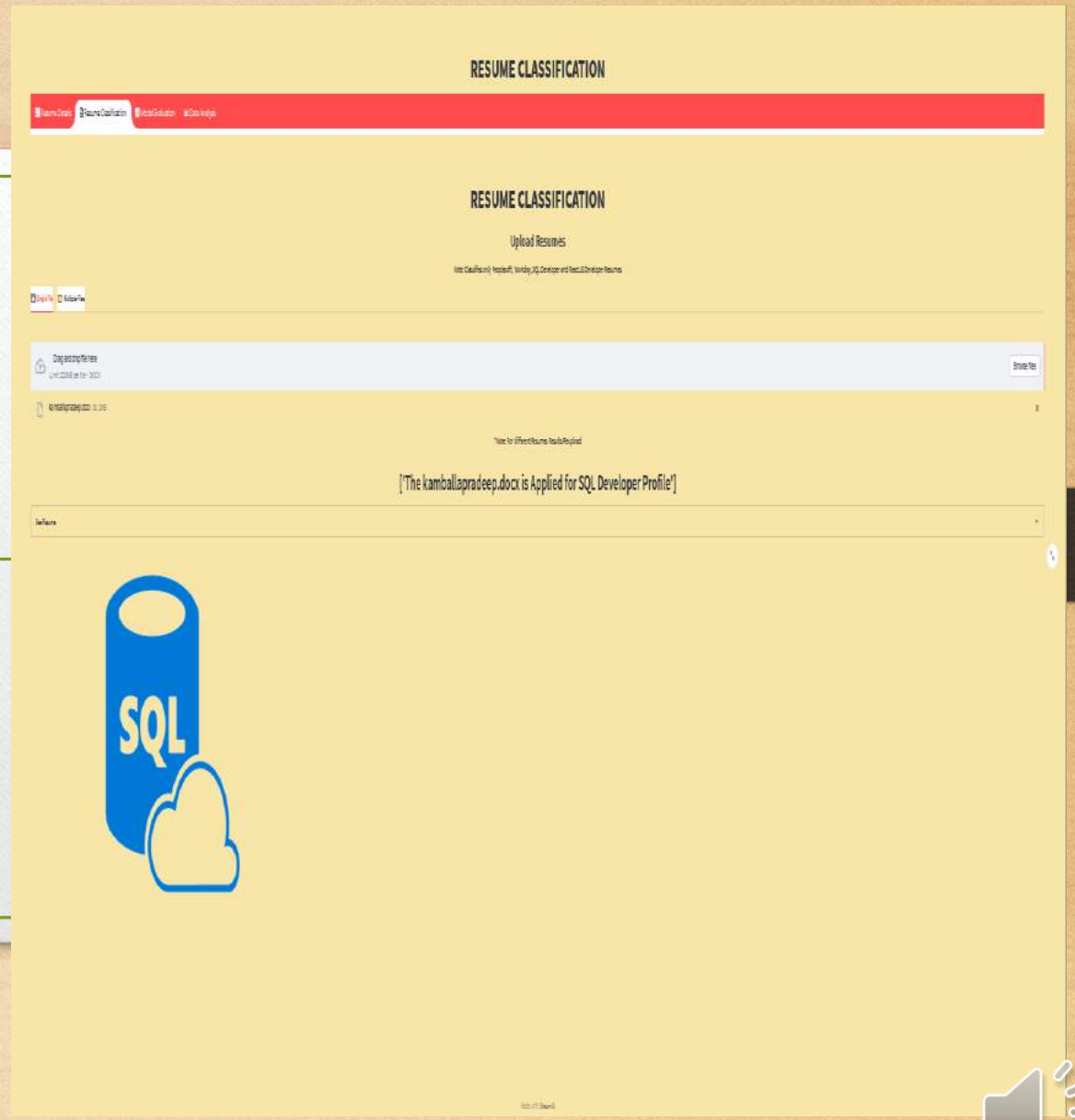
Made with Streamlit

D E P L O Y M E N T

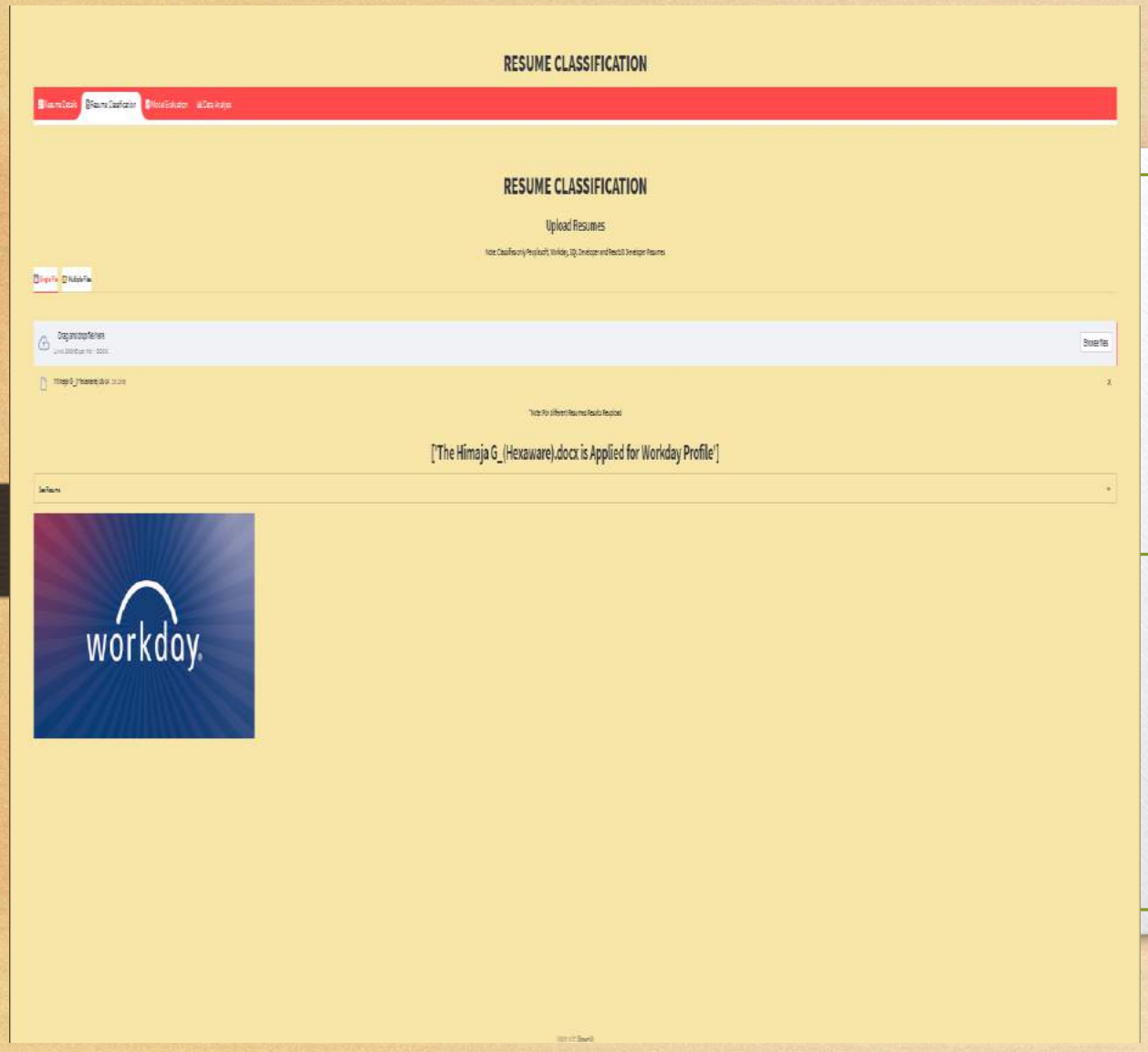

RESUME CLASSIFICATION
Resume Details
Resume Classification
Model Evaluation
Data Analysis
RESUME Details
Upload Resume (Single File Accepted)
Drag and drop file here
Limit 200MB per file • DOCX
Browse files
Reactjs Developer_Pranish Sonone_Musquare Technologies.docx 20.9KB
*Note: For different Resumes Results Reupload
Resume Analysis
Hello Pranish Sonone Career
Basic info
Name: Pranish Sonone Career
Experience (Years): 2.0
Last Position: []
Competence: ['teamwork', 'leadership', 'score']
Education: [('BE', '2018'), ('12th', '2018'), ('10th', '2018'), ('CBSE', '2018')]
Email: None
Contact: []
Date of Birth: None
See Resume
Skills Analysis
Skills that Pranish Sonone Career have
Analytical
Engineering
Electronics
Javascript
.net
Programming
Technical
Json
Sql
Electrical
Testing
Technical skills
Competence Score
Pranish Sonone Career's Competence Score: 8
Name
Pranish Sonone Career
Mobile No.
Email
None
DOB
None
Experience (Years)
2.0
Last Position
Competence
['teamwork', 'leadership', 'score']
competence score
8.0

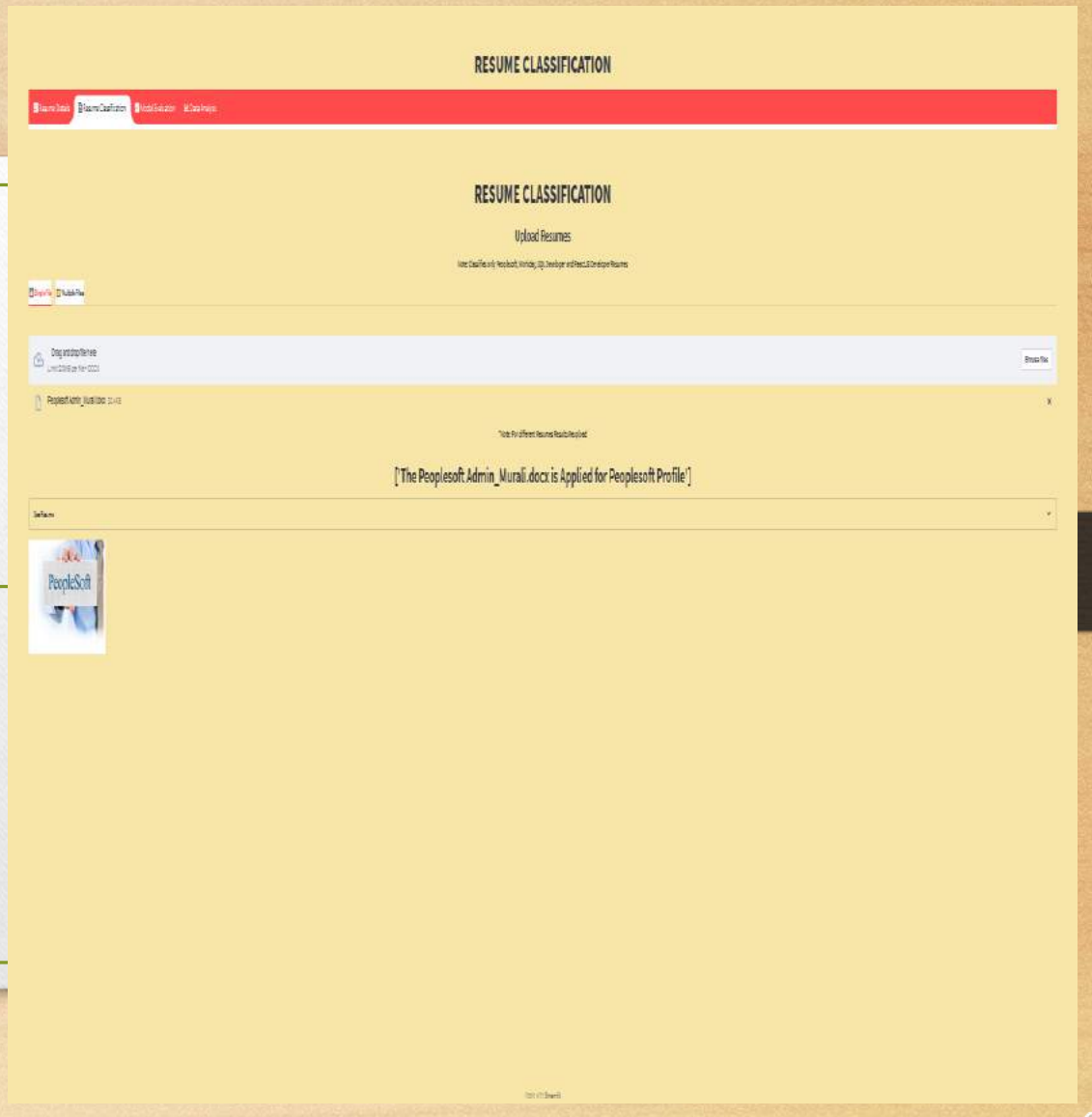
DEPLOYMENT:


RESUME CLASSIFICATION
RESUME CLASSIFICATION
Upload Resumes
['The Reactjs Developer_Pranish Sonone_Musquare Technologies.docx is Applied for React Developer Profile']
React JS


RESUME CLASSIFICATION
RESUME CLASSIFICATION
Upload Resumes
['The kamballapradeep.docx is Applied for SQL Developer Profile']
SQL

DEPLOYMENT


RESUME CLASSIFICATION
RESUME CLASSIFICATION
Upload Resumes
['The Himaja G_(Hexaware).docx is Applied for Workday Profile']
workday.

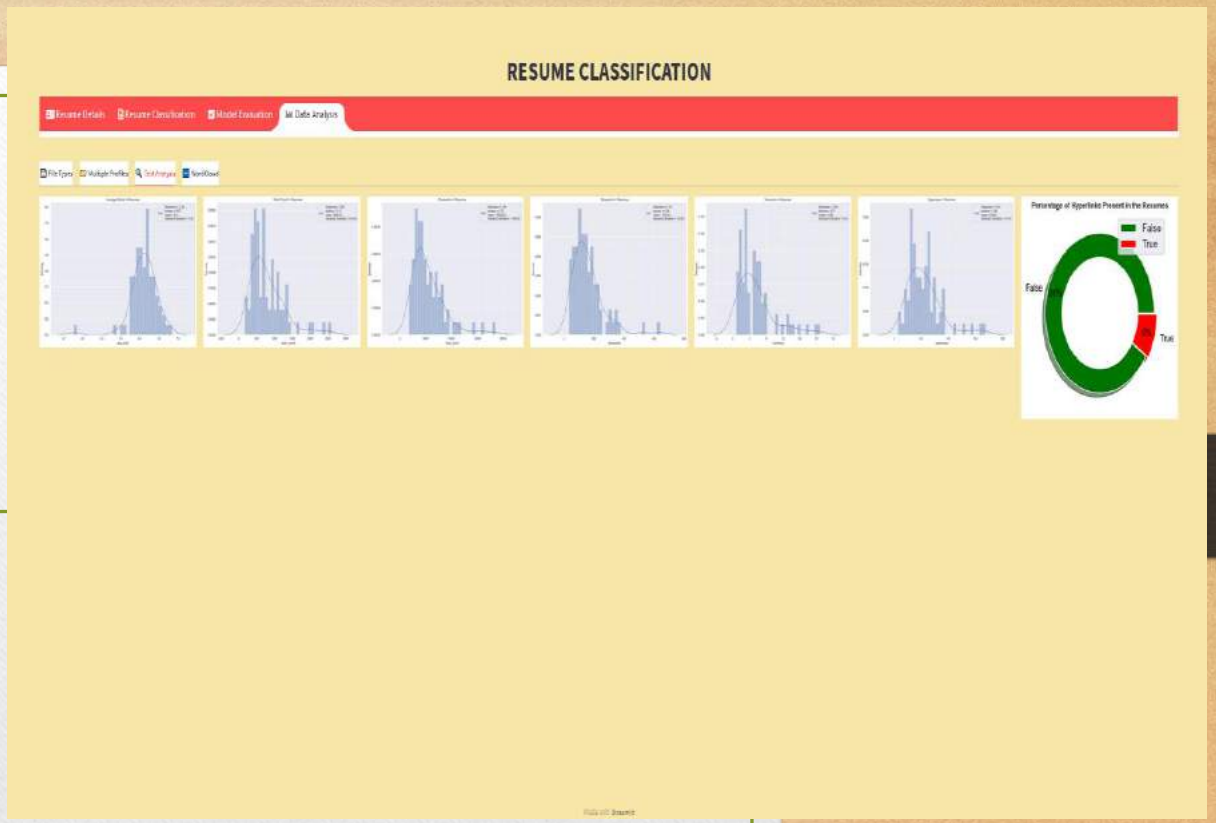

RESUME CLASSIFICATION
RESUME CLASSIFICATION
Upload Resumes
['The Peoplesoft Admin_Murali.docx is Applied for Peoplesoft Profile']
PeopleSoft

DEPLOYMENT

RESUME CLASSIFICATION

D E P L O Y M E N T


RESUME CLASSIFICATION
Resume Details
Resume Classification
Model Evaluation
Data Analysis
Percentage of data in Train and Test
Training Data
69%
31%
Testing Data
Training Data
Testing Data
Number of Classes in Train Data
React Developer
Peoplesoft
Workday
SQL Developer
15
14
14
10
Number of Classes in Test Data
React Developer
Workday
Peoplesoft
SQL Developer
7
7
6
4
Number of Classes in Dataframe After Oversampling


RESUME CLASSIFICATION
Resume Details
Resume Classification
Model Evaluation
Data Analysis
False
True

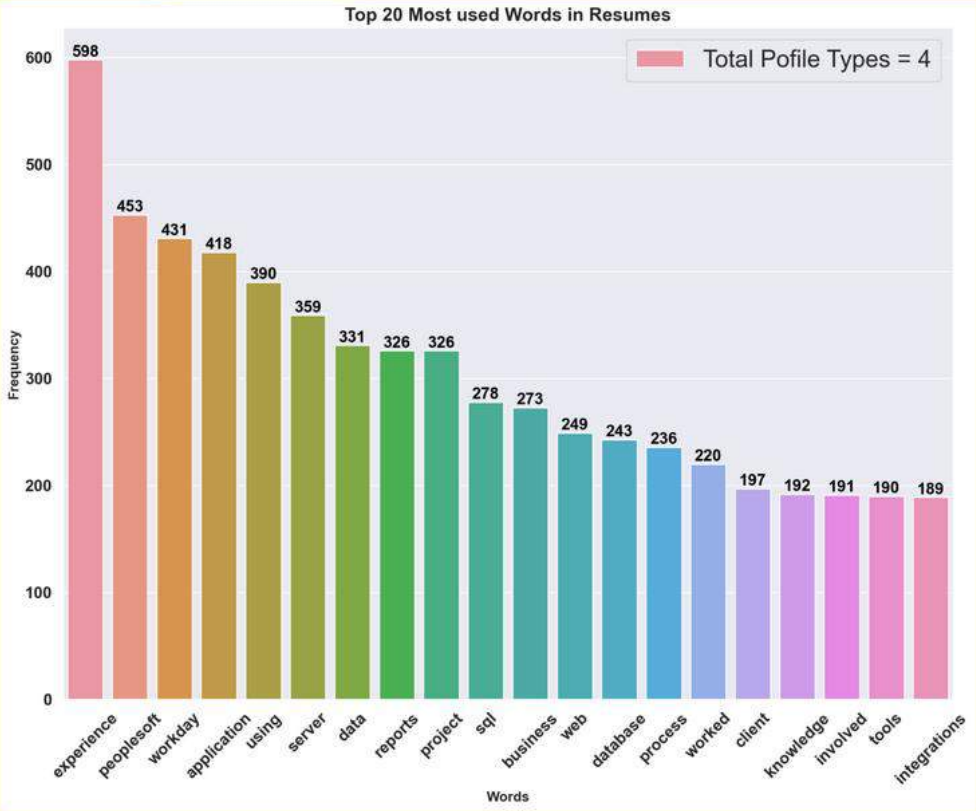
DEPLOYMENT

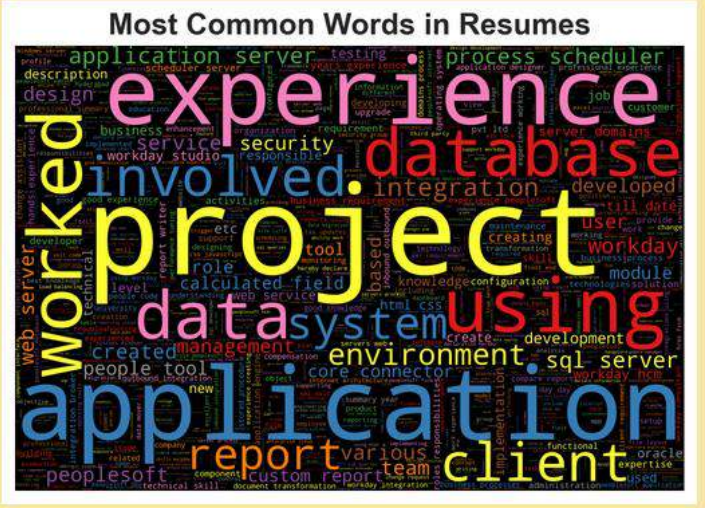
RESUME CLASSIFICATION

Resume Details | Resume Classification | Model Evaluation | Ist Data Analysis

File Types | Multiple Profiles | Text Analysis | WordCloud

Most Word Used and WordCloud | Profiles WordCloud and Word BarChart





CHALLENGES

- Faced various errors related to dependencies, libraries, and compatibility issues, which hindered the smooth deployment of the model and we overcame with the help of Chatgpt.
- We struggled a lot to do pre-processing as it took almost 2-3 weeks to get a detailed dataset from our imbalanced data.


thank
you