Data Intake Report

Group Name: <NLP: Resume Extraction>
Name: <Tatiana Moteu Ngoli>
Email: <mtatiana@aimsammi.org>
Country: <Germany>
Data storage location: <https://github.com/TatianaMoteuN/Data-Glacier/tree/master/week10>

Tabular data details:

Total number of observations	<160>
Total number of files	<1>
Total number of features	</>
Base format of the file	<.json>
Size of the data	< 160 B >

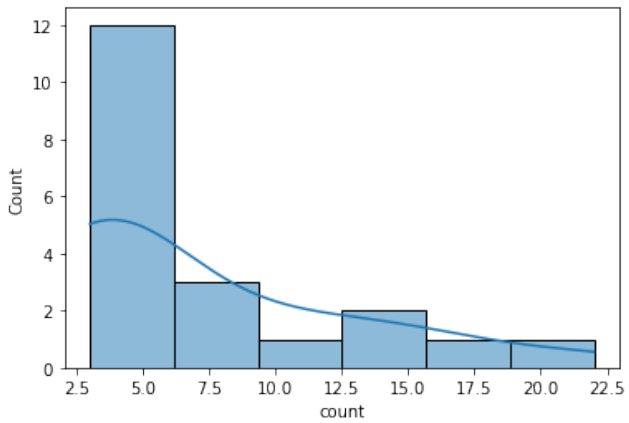
Note: Replicate same table with file name if you have more than one file.

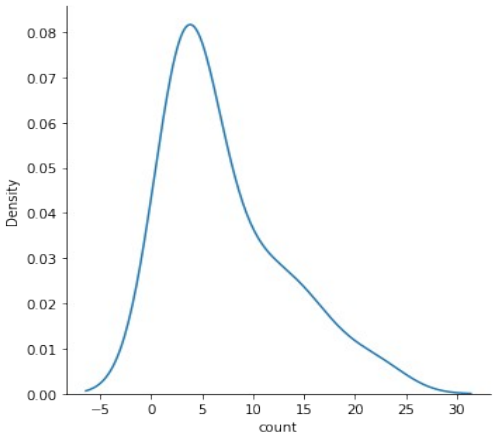
Problem description:
Resumes contain surfeit information that is not relevant for the HR/authority, and they have to manually process the resumes to shortlist the promising candidates for them. And, thus making the shortlisting task a herculean task for HR. By making use of the NER(Named Entity Recognition) model of NLP this problem can be solved by finding and classifying the entities that are present in each resume into predefined classes such as person name, college name, academics information, relevant experiences, skill set, etc.

EDA performed on the data:

We have plotted the distribution of some entities tags in order to have a general overview on the data.

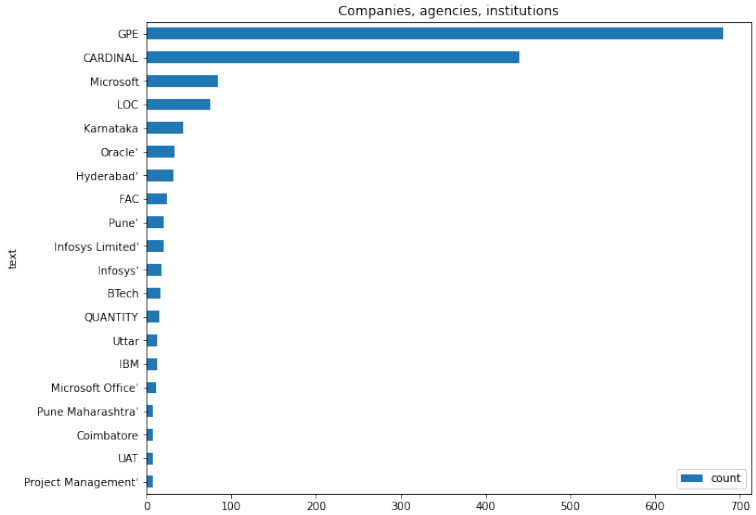
- **PERSON**: in the images below, we can see the person count regarding each person tags found in the text





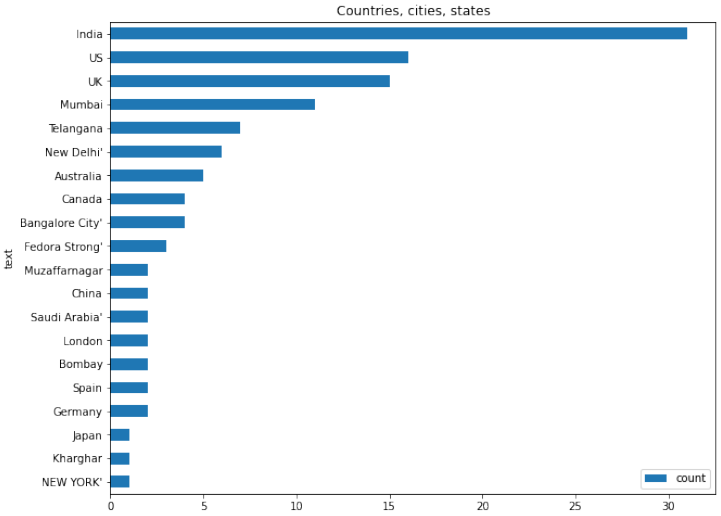
We can observe that count person between 2-6 appear the most in the data and also there is a name which comes 22 time in the data

- **ORG**: in the image below we have a list of companies, agencies, institutions that appear in the data



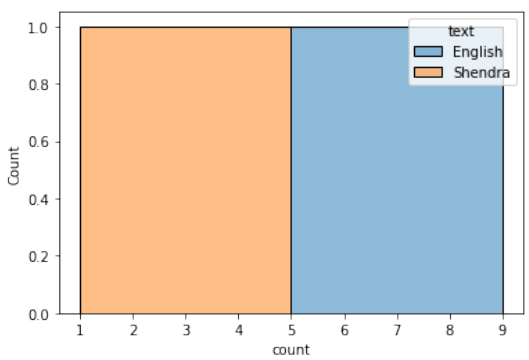
We can observe that GPE showed a high distribution and next come CARDINAL and MICROSOFT

- **GPE**: in the image below we have a list of countries, cities, states that appear in the data



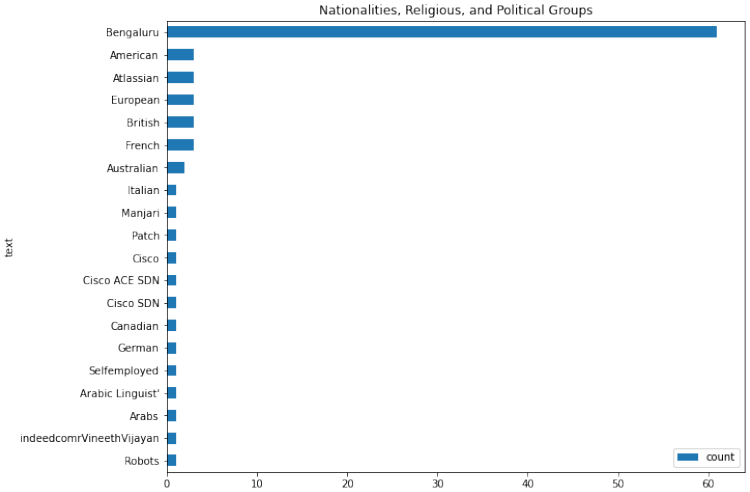
We can observe that the Indian country appear the most in the data which means that the resumes have been mostly submitted by people who leave in Indian. Then come US, UK and Mumbai.

- **LANGUAGE**: We realized that there are two languages appearing in the resumes and the image below showed us the distribution of those languages



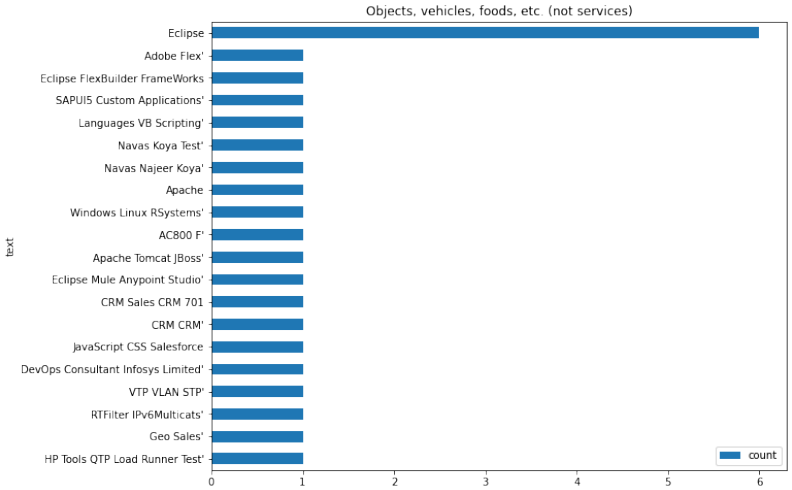
We can observe an equal distribution between the two languages (English and Shendra) found in the data

- **NORP**: the image below shows us a list nationalities, religious and political groups



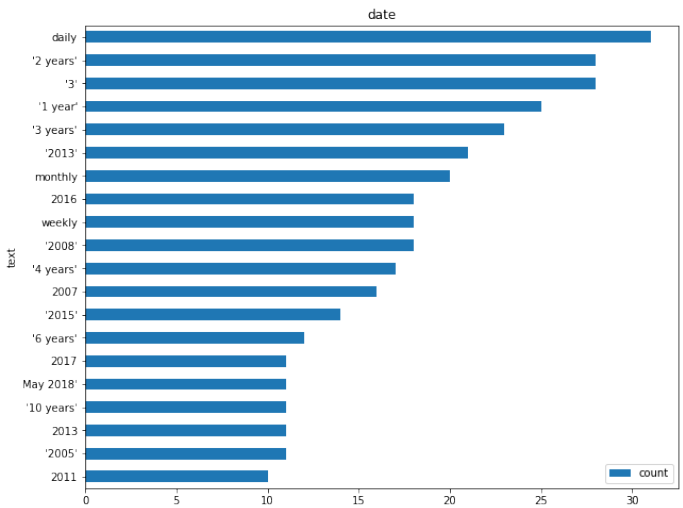
We can observe a higher distribution on Bengaluru which means that most of candidates come from Bengaluru. Then come American, Atlassian, European, British and French people.

- **PRODUCT**: In the image below we have a list of objects, vehicles, foods, etc.



We can observe that many candidates have used Eclipse as material in their past work

- DATE: in the image below we have absolute or relative dates or periods



We can observe that most of the candidates have done daily works and then most of them has 2-3 years of experiences in their fields

Final recommendation: based on these observations, we recommend to focus on the years of experiences of each candidates, the companies where the worked in the past, the materials that they have been used in their works in the selection process depending on the job description.