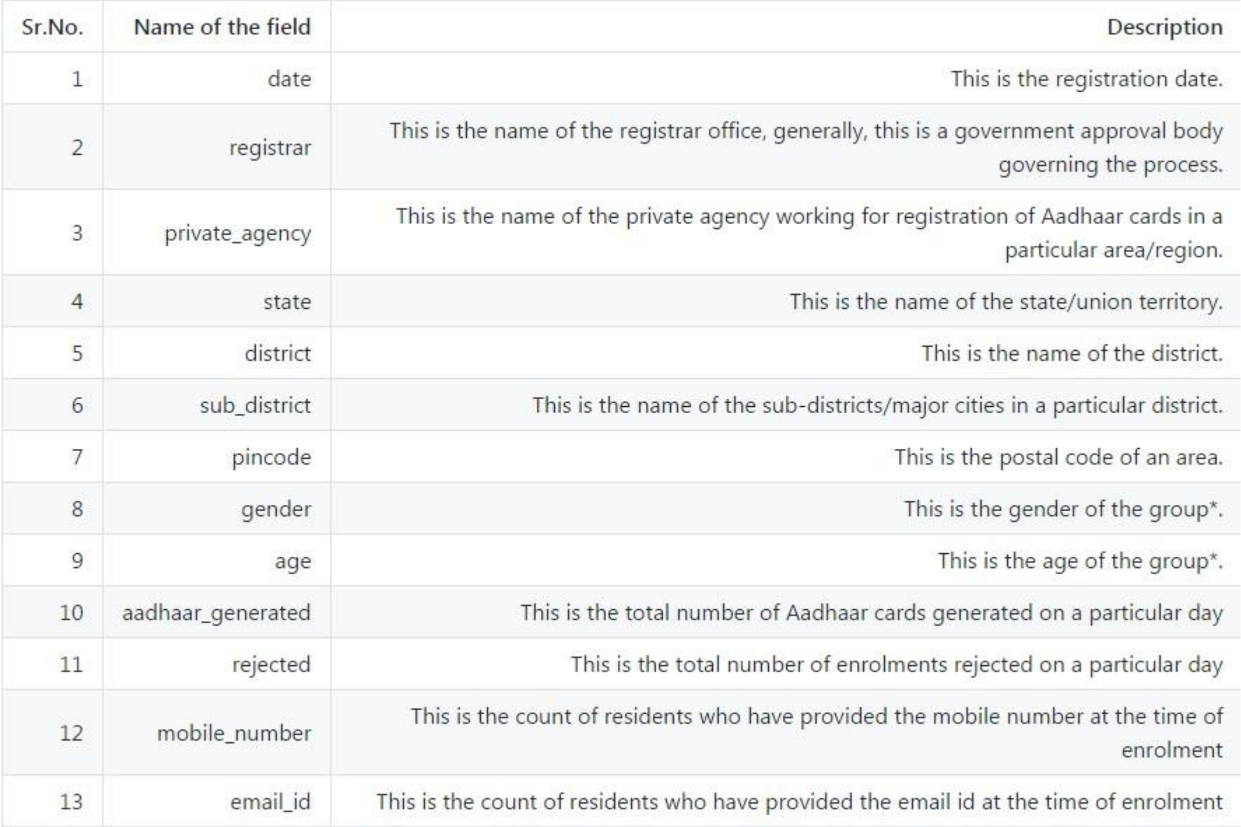
Problem Statement

You are working for the data analysis team and wish to analyse the data in hand for various demographic parameters. The analysis at hand involves basic data preparation, processing and understanding. Further, you also wish to forecast the effects of certain information on the overall Aadhaar number generation. The metadata/dictionary is provided below:

Metadata/Data Dictionary



Sr.No.	Name of the field	Description
1	date	This is the registration date.
2	registrar	This is the name of the registrar office, generally, this is a government approval body governing the process.
3	private_agency	This is the name of the private agency working for registration of Aadhaar cards in a particular area/region.
4	state	This is the name of the state/union territory.
5	district	This is the name of the district.
6	sub_district	This is the name of the sub-districts/major cities in a particular district.
7	pincode	This is the postal code of an area.
8	gender	This is the gender of the group*.
9	age	This is the age of the group*.
10	aadhaar_generated	This is the total number of Aadhaar cards generated on a particular day
11	rejected	This is the total number of enrolments rejected on a particular day
12	mobile_number	This is the count of residents who have provided the mobile number at the time of enrolment
13	email_id	This is the count of residents who have provided the email id at the time of enrolment

(*: explained in the example below).

Note: The dataset does not contain the headers. You should use the header names in the order as mentioned above.

You can understand the data dictionary better by the following example: A row with data - 20150420, Allahabad Bank, A-Onerealtors Pvt Ltd, Uttar Pradesh, Ambedkar Nagar, Akbarpur, 224155, F, 15, 5, 0, 0, 4 indicates that

- On 20 Apr 2014 (date), for A-Onerealtors Pvt Ltd (private_agency) registered with Allahabad Bank (registrar) at PIN code 224155, Akbarpur (sub_district), Ambedkar Nagar (district), Uttar Pradesh (state)
- Among the group of women aged 15
- There were 5 Aadhar numbers generated and 0 were rejected
- Out of the 5 that applied, none had an email ID and 4 had mobile numbers

Checkpoints

Checkpoint 1

Load the data into HDFS, Hive Managed table, Hive External table and Spark DataFrame.

1. Commit the screenshot of the view/result of the top 25 rows from each individual store (HDFS, Hive – Managed/External and Spark DataFrame).

Checkpoint 2

2. Describe the schema.
3. Find the count and names of registrars in the table.
4. Find the number of states, districts in each state and sub-districts in each district.
5. Find the number of males and females in each state from the table and display a suitable plot.
6. Find out the names of private agencies for each state.
7. Plot the number of private agencies for each state.

Checkpoint 3

8. Find top 3 states generating most number of Aadhaar cards?
9. Find top 3 private agencies generating the most number of Aadhar cards?
10. Find the number of residents providing email, mobile number? (Hint: consider non-zero values.)
11. Find top 3 districts where enrolment numbers are maximum?
12. Find the no. of Aadhaar cards generated in each state?

Checkpoint 4

13. Create a data frame using the file and provide its summary.
14. Write a command to see the correlation between “age” and “mobile_number”? (Hint: Consider the percentage of people who have provided the mobile number out of the total applicants)
15. Find the number of unique pincodes in the data?
16. Find the number of Aadhaar registrations rejected in Uttar Pradesh and Maharashtra?

Checkpoint 5

On the given dataset, perform EDA and find:

17. The top 3 states where the percentage of Aadhaar cards being generated for males is the highest.
18. In each of these 3 states, identify the top 3 districts where the percentage of Aadhaar cards being rejected for females is the highest.

19. The top 3 states where the percentage of Aadhaar cards being generated for females is the highest.
20. In each of these 3 states, identify the top 3 districts where the percentage of Aadhaar cards being rejected for males is the highest.
21. The summary of the acceptance percentage of all the Aadhaar cards applications by bucketing the age group into 10 buckets.