Summary Of Data Cleaning & Filtering:

- The database design should include a unique identifier field.
- Consider only countries tagged as 'India' or 'unmarked.' (Note: Two Indian companies were erroneously tagged as 'ICELAND' which will be retagged as INDIA.
- Records where the PAN was mentioned as 'ZUMMY/PJDUM' or similar patterns (e.g., '*ZUMMY{Random numbers}*') should be removed.
- Consecutive double spaces, leading, and trailing spaces should be removed from the columns.
- PAN & PIN validation should be performed for foreign records, i.e., records tagged with "-" for the country.
- Test or dummy email addresses and those related to @about.in should be removed.
- Typographical errors, such as spaces, hyphens, and dots, should be rectified in the columns.
- Valid mobile numbers should be formatted in a uniform way, all should be 13 digits long.

	Stg_2	Stg_1	Total Records
Data Ingestion	**29236**	**32816**	**62052**
Foreign Country Removal,	**29027**	**32121**	**61148**
'@about' - Removal	**26944**	**32111**	**59055**
Test Record Removal	**26944**	**31217**	**58161**
ZUMMY/ PJDUM PAN Removal	**26887**	**29091**	**55978**
PAN & PIN Validation Foreign Records Removal*	**26767**	**27058**	**53825**
Total Records			**53825**

Attribute Wise Unique and Missing Values:

Attribute	Total	Missing Values	%	Count	%	Valid (against Regex)	%	Unique	%
org _pan	53825	443	0.8%	53382	99.17%	NA	NA	49390	91.76%
org_name	53825	0	0.0%	53825	100.00%	NA	NA	45578	84.67%
pan	53825	443	0.8%	53382	99.17%	53363	99.14%	48865	90.78%
gstn	53825	36133	67.1%	17692	32.86%	17210	31.97%	17214	31.98%
cin	53825	50605	94.0%	3220	5.98%	60	0.11%	101	0.18%
email	53825	3	0.0%	53822	99.99%	53497	99.39%	49365	91.71%
mobile	53825	1251	2.3%	52574	97.67%	51865	96.35%	47187	87.66%