Job Description dataset tranformation proposal:

Feature Name	Type	Transformation / Action	Unique Values / Notes	Value Example
Job Id	BIGINT	Drop	Unique for each row	1089843540111562
Experience	String	Parse to extract min and max experience as integers	Variable string patterns	"5 to 15 Years"
Qualifications	Categorical	Encode as One-Hot or Ordinal	~10 unique values	"M.Tech"
Salary Range	String	Already parsed into `salary_min`, `salary_max` (keep those)	Not needed after parsing	"$59K-$99K"
location	Categorical	Drop	214 unique values	"Douglas"
Country	Categorical	Drop	216 unique values	"Isle of Man"
Latitude	Float	Use directly or transform to polar	Continuous	54.2361
Longitude	Float	Use directly or transform to polar	Continuous	-4.5481
Work Type	Categorical	One-Hot Encode	5 unique values	"Intern"
Company Size	Integer	Keep as is	Continuous numeric	26801

Job Posting Date	Date	Convert to datetime, encode cyclical (sin/cos)	Continuous over time	"2022-04-24"
Preference	Categorical	One-Hot Encode	3 unique values	"Female"
Contact Person	String	Drop	High-cardinality, not predictive	"Brandon Cunningham"
Contact	String	Drop	Irrelevant	"001-381-930-7517x737"
Job Title	Categorical	Encode (One-Hot or Embedding)	147 unique values	"Digital Marketing Specialist"
Role	Categorical	Encode (One-Hot or Embedding)	376 unique values	"Social Media Manager"
Job Portal	Categorical	One-Hot Encode	16 unique values	"Snagajob"
Job Description	Text	NLP processing (TF-IDF, embeddings, keyword extraction)	Free text	"Oversee an organization’s social..."
Benefits	Text (List)	Tokenize and extract standard benefits into binary flags	Free text; extract top-N benefits	"Flexible Spending Accounts (FSAs), Relocation Assistance"

Skills	Text	Tokenize and extract known skill keywords; binary flags or frequency	Free text; extract top-N skills	"Social media platforms, Analytics"
Responsibilities	Text	NLP processing similar to `Job Description`	Free text	"Manage and grow social media..."
Company	Categorical	Encode (One-Hot or Embedding)	885 unique values	"Icahn Enterprises"
Company Profile	JSON String	Parse into structured fields	Only 1884 missing out of ~1.6M	'{"Sector":"Diversified",...}'

Company profile:

Field in Company Profile	Type	Unique Values	Transformation / Action	Value Example
Sector	Categorical	204	One-Hot or Ordinal encoding	"Diversified"
Industry	Categorical	204	One-Hot or Embedding	"Diversified Financials"
City	Categorical	344	Drop or encode	"Sunny Isles Beach"
State	Categorical	98	One-Hot or drop	"Florida"
Zip	String	497	Drop	"33160"

Website	URL String	881	Drop or extract domain (optional)	"www.ielp.com"
Ticker	String	819	Create binary flag `is_public`	"IEP"
CEO	String	836	Drop name, derive gender feature	"David Willetts"
CEO_gender (derived)	Categorical	3	One-Hot Encode (`male`, `female`, `unknown`)	"male"

CEO Gender Counts:

- **Male:** 1,222,568
- **Female:** 123,662
- **Unknown:** 262,388