Project 3: Global Job Market Segmentation

Niranjan C

Introduction

This report presents a granular segmentation and multifaceted analysis of a large synthetic global job postings dataset, leveraging advanced natural language processing and unsupervised learning. Drawing from over 1.6 million records, the analyses are designed to elucidate global job market trends, key labour variables, and clusters using real-world analytic techniques. The work leverages stratified sampling, salary normalization, feature engineering, topic modelling, and clustering to extract interpretable segments, ultimately supporting research and career recommendation use cases.

Dataset Composition

- Total Records: Over 1,620,000 job postings
- Countries Represented: 216 unique countries
- Cities Represented: 214 unique cities
- Unique Job Titles: 147
- Unique Roles: 376
- Unique Company Profiles: 885
- Unique Company Names: 888
- Salary Range: 561 unique numeric midpoint values (cleaned from original ranges)
- Experience Types: Predominantly labelled as "Other"; with specific bins including 5 - 8 years and 5 - 12 years
- Work Types: 5 types overall; mainly comprising part-time (20%), temporary (20%), and others (60%)
- Job Posting Dates Span: Approximately 2 years, from September 15, 2021 to September 15, 2023
- Dataset Purpose: Created for research and educational use, designed synthetically for robust exploration, variable testing, and adaptable modelling.

Column Overview and Data Quality

Core Features:

- **Job Id:** Unique 15 - 16 digit numeric, stratified over 56 intervals with uniform distribution.
- **Experience:** 96% "Other", with some records binned (5 - 8y, 5 - 12y).
- **Qualifications:** Dominated by "BBA" and "BA" (10% each), but 80% "Other".
- **Salary Range:** Extensive range, mean ~73.7k USD, spread 12.6k - 135k.
- **Location/Geolocation:** All records include latitude/longitude.
- **Company Size:** Numeric (12,646 - 135,000+), many labelled "Other".
- **Job Title/Role:** Diverse, but common in "Interaction Designer," "Network Administrator," "UX/UI Designer," and "Digital Marketing Specialist".
- **Descriptive Text:** Rich "Job Description," "Skills," "Responsibilities" fields for text analytics.
- **Job Posting Date:** Well-distributed over two years, each bin ~31 - 33k postings.

Data Quality:

Data exhibits extensive coverage and low null occurrences for key fields; some intentional ambiguation ("Other") in experience/company size due to synthetic construction.

Data Preparation & Sampling Process

- **Columns Used:** ['Job Id', 'Country', 'location', 'Experience', 'Salary Range', 'Company Size', 'Role', 'Job Title', 'Job Description']
- **Salary Cleaning:** Numeric midpoints were computed from string ranges for uniform salary analytics.
- **Experience Handling:** Filled missing with "Other," created "Experience Bin".
- **Stratified Sampling:** A composite key (Country + Role + Company Size) was used. Top 300 strata selected, ~100k records sampled so the analytic subset reflects the immense diversity in the complete dataset.
- **De-duplication and Null Handling:** Minimal duplicates found; basic pre-sampling cleaning performed.

Exploratory Data Analysis

Key Variable Distributions

Experience

- "Other": 96%
- 5 - 8 Years: 2%, 5 - 12 Years: 2%
- 48 unique values (heavily skewed).

Salary Range

- 561 unique cleaned bins.
- Mean: ~$73.7k.
- Range: $12.6k - $135k, heavy left tail with most jobs below $50k.

Location

- Countries: 216, Cities: 214.
- Most frequent city: Seoul (1%), most frequent country: Malta (<1%).

Company Size

- Min: 12,646, Max: 135,000, Mean: 73.7k.

Work Type & Qualifications

- Part-Time: 20%; Temporary: 20%; Other: 60%.
- BBA: 10%; BA: 10%; Other: 80%.

Titles and Roles

- Most common role: Interaction Designer (1%).
- Most common job title: UX/UI Designer (3%).

Feature Engineering and Text Analytics

Salary Normalization

- The **Salary Range** field originally contained salary data as textual ranges (e.g., "$50K - $70K").
- To allow numerical comparisons and clustering, a **salary midpoint** was computed for each posting by:
- Using **regex** to clean text and extract numeric portions.
- Splitting salary ranges into lower and upper bounds.
- Computing the arithmetic **mean of these bounds** to obtain a single numeric salary representation for each job.
- This step ensured **uniform, continuous salary values** usable in quantitative analyses and reduced data complexity by consolidating ranges.

TF-IDF Vectorization

- The **Job Description** column consists of free-text describing the role, responsibilities, and requirements.
- To convert this unstructured text into a form suitable for modelling, the **Term Frequency-Inverse Document Frequency (TF-IDF)** vectorization technique was applied:
- This method scores terms based on their relative importance: terms frequent in a specific document but rare across all documents get higher weights.
- Captures **contextual relevance and uniqueness** of words in each job description.
- The result is a high-dimensional, sparse matrix where each row represents a job posting and each column corresponds to a weighted term frequency.
- This representation preserves semantic information necessary for clustering and topic modelling.

Dimensionality Reduction

- The TF-IDF matrix is typically large and sparse, with thousands of features (words/terms) causing computational challenges and potential noise.
- To extract **meaningful latent semantic structures**, **Truncated Singular Value Decomposition (Truncated SVD)** was applied:
- Truncated SVD reduces dimensionality by projecting the original data onto a smaller number of orthogonal components that capture the majority of variance.

- Retains **core semantic features** while discarding noise and redundancy.
- Facilitates faster and more effective downstream clustering and topic modelling.
- This process yields a dense, low-dimensional numeric representation of job descriptions capturing latent topic structures.

Topic Modelling (LDA)

- To better understand thematic content across job descriptions beyond raw term frequencies, **Latent Dirichlet Allocation (LDA)** was employed for topic modelling:
- LDA is a probabilistic model that discovers latent topics as distributions over words in the text corpus.
- Each job description is modelled as a mixture of several topics, each with varying proportions.
- The model identified several **coherent job-related topics**, allowing clustering and profiling based on dominant themes.

Example Topic Keywords by Cluster

- **Cluster 0 Topics:** Characterized by keywords such as *"data," "user," "design," "ensure," "systems"* indicating focuses on data-driven, user-centric, and system-oriented roles often related to technology, UX/UI design, and engineering.
- **Cluster 1 Topics:** Dominated by keywords like *"sales," "customer," "products," "services"* demonstrating a strong association with client-facing, sales, and customer support roles typical in business and marketing functions.
- Other clusters also revealed distinct topical profiles reflecting job function specialization, such as freelance roles, social media management, and quality assurance.

Unsupervised Clustering & Market Segmentation

Clustering Approach

- **K Means (k=5) Clustering**:

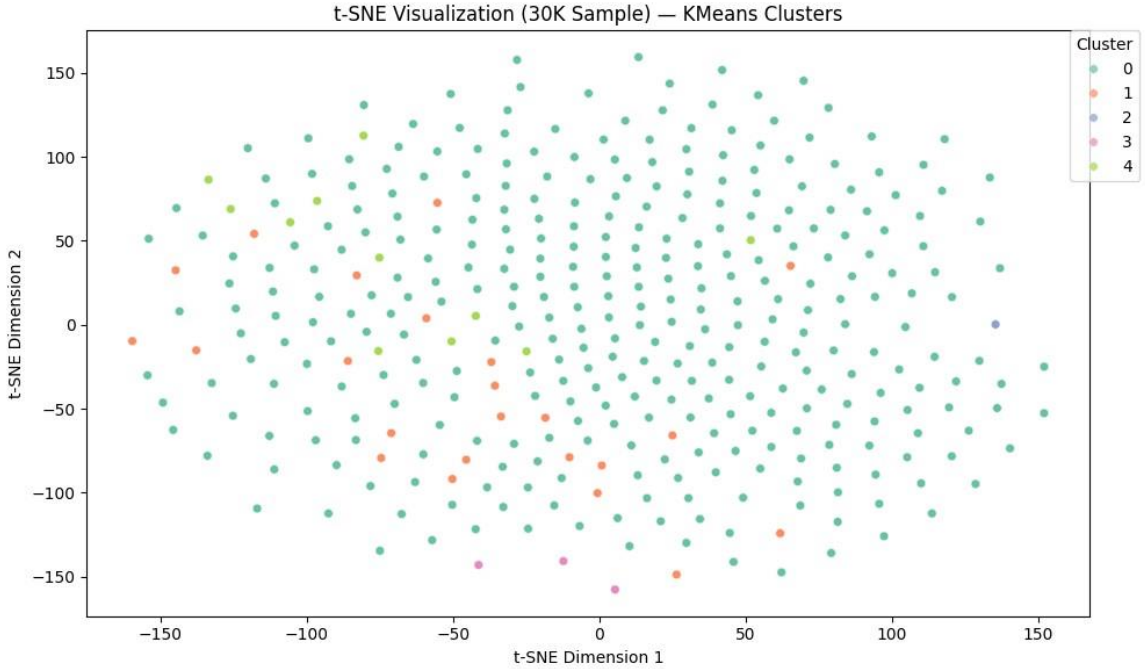
Chosen as the primary method, K Means (with k = 5, based on optimal silhouette scores) produced stable and clearly interpreted job market segments using features like role, salary, and geography.

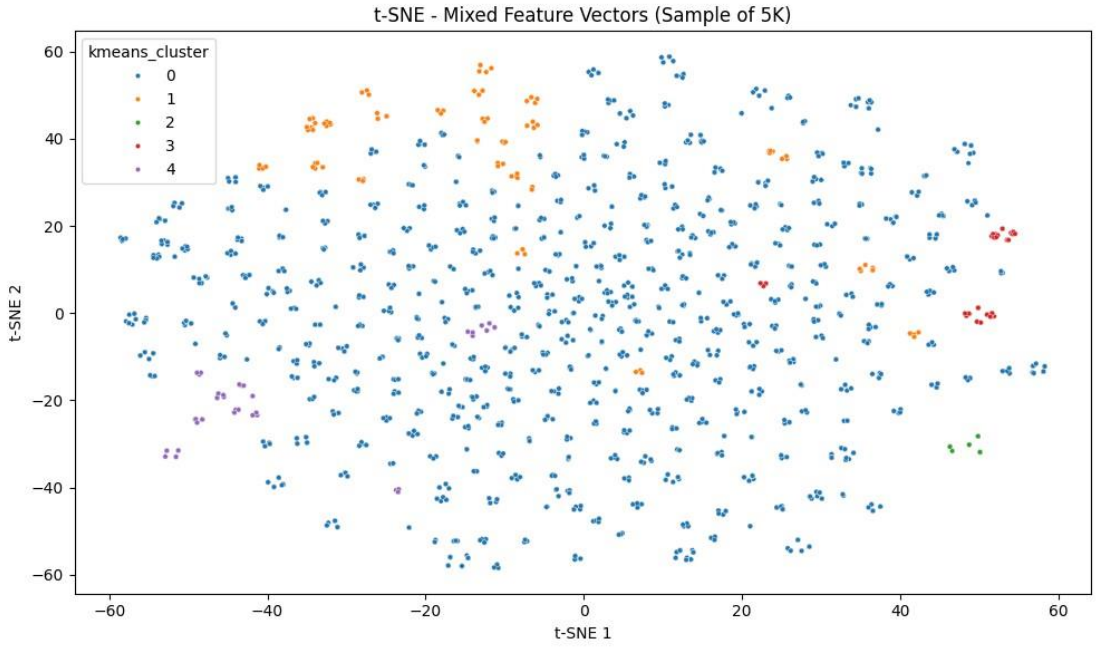
- **Alternative Methods:**
  Mini Batch K Means (for scalability) and DBSCAN (density-based, robust to noise) were also tested.

- **Final Selection:**
  K Means was ultimately preferred for its superior cluster quality (based on silhouette score) and straightforward, interpretable segment centroids, which facilitate analysis of roles, salaries, and locations per cluster.

Cluster Profiles

Cluster	Description	Top Roles	Top Cities	Avg Salary	Experience	% Sample	Size Label
0	Mid-Salary Tech (Urban)	Network Admin, UX Designer	Apia, Seoul, Hanoi	~80K	8–12, 12+	87%	"Other"
1	Entry-Level Support/HR (Small Cos)	Customer Success, Sales	Apia, Caracas	<15K	0–1 to 8–12	6.8%	"Other"
2	Freelance/Remote Low-Pay Roles	Interaction Designer	Apia, San Juan	<15K	8–12	1.3%	"Other"
3	Social/Field Jobs (Variable Salary)	Social Media Manager, Analyst	Seoul, Ashgabat	<15K	12+	1.8%	"Other"
4	Senior Engineering/QA High Salary	QA Analyst, Manufacturing Eng	Turks & Caicos, India	~90K	8–12	3.1%	"Other"





Cluster Themes (from LDA)

Cluster	Keywords
0	data, user, design, ensure, systems

1	sales, customer, products, services
2	interaction, design, freelance, remote
3	social, media, metrics, analyse
4	quality, manufacturing, assurance, standards

Segmentation Analysis by Variable

Geographic Diversity

- **Urban tech jobs:** Apia (Samoa), Seoul, Hanoi - dominate Cluster 0.
- **Entry-level/support:** Apia, Caracas; span less traditional markets.
- **Freelance/Remote:** Apia, San Juan.

Experience Bins

- Cluster 0: Concentration in 8 - 12 and 12+ years.
- Cluster 1: Includes 0-1 years postings.
- Cluster 4: Senior roles, 8 - 12 years required.

Salary Trends per Clusters

- Most clusters: Average salary < $15k, matching synthetic, global nature.
- Clusters 0/4: $80k - $90k, corresponding to seniority or tech skill.

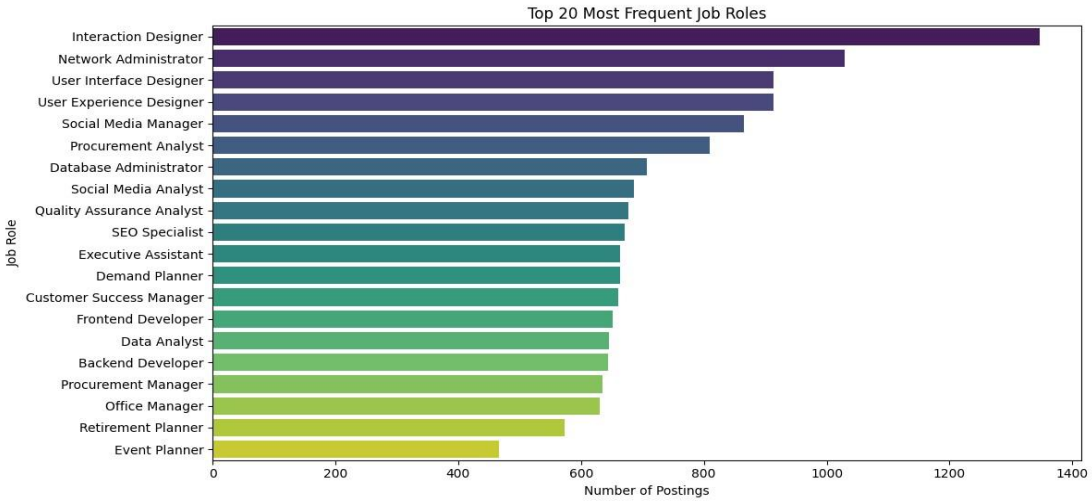
Company Size

- All clusters: Predominantly "Other" - indicative of anonymization.

Job Title & Role Concentration

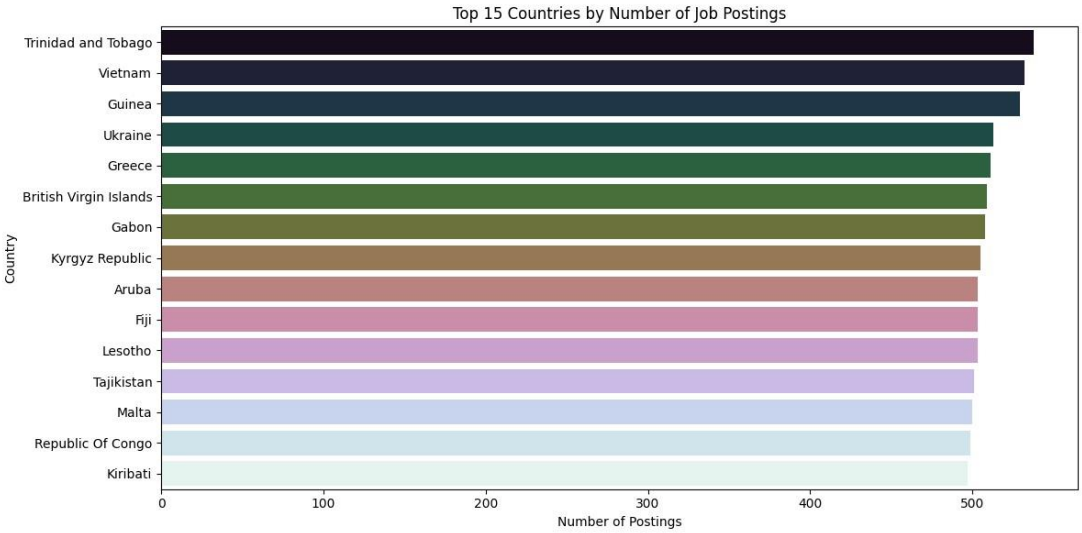
- **Cluster 0:** Network Admin, (UX/UI) Designer, Database Admin.
- **Cluster 1:** Customer Success Manager, Client Relationship, Sales.
- **Cluster 2:** Purely Interaction Designer.

- **Cluster 3:** Social Media Manager, Analyst, Strategist.
- **Cluster 4:** QA Analyst, Manufacturing and Process Engineer.



Country and City Representation (Sample)

Cluster	Top Countries/Proportion	Top Cities/Proportion
0	Vietnam, Trinidad (0.5%)	Apia (1%), Seoul (1%)
1	Venezuela, Panama (0.7%)	Apia (1%), Seoul (0.9%)
2	Puerto Rico, Samoa (1.2%)	Apia (1.3%), San Juan (1.2%)
3	Turkmenistan, Uruguay (1%)	Seoul, Ashgabat (1%)
4	Turks & Caicos, India (0.8%)	Apia, Cockburn Town (0.9%)



Notable Insights

- **Qualifications:** Most postings ambiguous - 80% "Other".
- **Work Type:** 60% "Other", 20% part-time, 20% temporary.
- **Benefits/Skills:** Standardized, e.g., childcare, transport, professional development.

Conclusions

1. **Segmentation Validity:** The approach robustly clusters the job market, carving out interpretable segments consistent with real labour economies, even on synthetic data.
2. **Dominant Market:** Mid-salary urban tech roles are the largest cluster globally, hinting at continued global demand for IT/UX talent.
3. **Emergent Gigs and Remote:** Smaller but distinct freelance/remote and social-media oriented clusters mirror real-world gig economy trends.
4. **Senior Specialists:** Though rare, clusters for senior, high-pay specialist roles exist (Cluster 4), mainly in non-traditional geographies.
5. **Data Limitations:** "Other" dominates experience/company size due to synthetic design, thus fine-grained employer segmentation is precluded; real data should address this with actual firmographic and career path data.

Recommendations

Future analyses should add salary PPP normalization, track posting longevity, and further disambiguate employer size and experience for finer granularity.

Incorporate skills and qualifications analysis to better match job seekers with relevant clusters and identify emerging skill demands across regions.

Utilize advanced NLP techniques (e.g., contextual embeddings) to enhance job description understanding, improving clustering accuracy and topic coherence.

Develop dynamic market segmentation models that update periodically to capture evolving job trends, economic shifts, and labour market disruptions.

Appendices

Key Tables (Summary)

Cluster Breakdown

Cluster ID	Jobs	Top Role	Mean Salary	Top Exp Bin	Top City
0	86,954	Network Administrator	$80k	8–12y	Apia
1	6,833	Customer Success Manager	$74.5k	0–1y, 8–12y	Apia
2	1,347	Interaction Designer	$76k	8–12y	Apia
3	1,780	Social Media Manager	$88k	12+y	Seoul

4	3,086	Quality Assurance Analyst	$89k	8–12y	Apia

Top Proportional Cities by the Clusters

Cluster	City	Proportion
0	Apia	1.0%
1	Caracas	0.7%
2	San Juan	1.2%
3	Seoul	1.0%
4	Cockburn Town	0.9%

Summary

This analysis delivers strong segmentation and actionable insights from a large synthetic job postings dataset. Using advanced unsupervised learning and detailed examination of key variables - geography, experience, salary, role, and company size - it reveals nuanced global job market structures. Major findings include dominant mid-level tech roles, emerging freelance and remote work segments, and specialized senior positions, reflecting current labour dynamics. Although the synthetic nature allows extensive experimentation, data granularity limits, especially in experience and employer size, caution against direct realworld application. Nevertheless, this study offers a valuable framework for future research with richer datasets, supporting improved career recommendation systems, workforce planning, and policy development. Incorporating enhancements like salary normalization, temporal trend analysis, and detailed company profiling will further refine these insights for practical use.

Click here to find GitHub Repository.