EMPLOYEE SEGMENTATION AND ATTRITION PATTERN DISCOVERY USING CLUSTERING IN R FOR A U.S. TECHNOLOGY COMPANY

1. Background

A mid-sized tech company in Seattle faced a surge in voluntary employee exits, particularly among junior developers and remote workers. While HR reports captured attrition counts, there was no structured way to identify shared behavioral or demographic patterns among those at risk.

We were contracted to use unsupervised machine learning in R to segment employees into behavioral clusters, with a focus on understanding factors linked to high attrition risk. These clusters would then support targeted retention strategies, policy design, and workload balancing.

2. Objective

- To use K-means clustering in R to identify distinct employee segments based on work-related and demographic variables
- To detect clusters strongly associated with attrition and inform proactive HR intervention
- To enable forecasting of at-risk groups based on historical data patterns

3. Data Used

Source: Internal HRMS data (2021–2023)

Dataset Details:

- 1,245 full-time employees (including current and voluntary exits)
- Fields included:
    - Age, Department, Role_Type, Tenure_Months, Projects_Assigned, Avg_Weekly_Hours, Remote_Work_Rate, Recent_Promotion_Flag, Exit_Flag

Preprocessing in R:

- Encoded categorical variables with model.matrix()
- Normalized all numerical features using scale()
- Removed the Exit_Flag from clustering (used later for profiling)

4. Methodology

4.1 Clustering Approach

- Used **K-Means Clustering** with kmeans()
- Determined optimal clusters (k = 4) using Elbow method and silhouette width
- Visualized clusters via PCA and ggplot2
- Compared clusters post hoc on attrition rates using Exit_Flag

4.2 Tools Used

- R packages: cluster, factoextra, ggplot2, dplyr, stats

5. Cluster Findings

Cluster	Description	Attrition Rate
1	Senior engineers, long tenure, office-based	4%
2	New joiners, low hours, low project count	18%
3	Mid-level devs, remote-heavy, high workload	27%
4	Business roles, stable tenure, hybrid setup	7%

Cluster 3 emerged as the **high-risk segment**:

- Overloaded with >50 hours/week
- Assigned to more than 3 active projects
- 85% remote workers
- Minimal recent promotions or internal movement

6. Interpretation and Strategy

- **Cluster 3** flagged for **intervention by people analytics and workforce planning teams**
- Suggested **redistribution of project load** and increased manager check-ins for this group
- **Cluster 2** showed early signs of disengagement; recommended a **structured onboarding extension** and peer mentorship
- **Cluster 1** is a model of retention stability; used as benchmark for culture audits

7. Reporting Output

- **R Markdown Report (PDF, 17 pages)**:
    - Data pipeline
    - Cluster plots and interpretation
    - Attrition comparison charts
    - HR intervention recommendations
- **Excel Output**:
    - Employee_ID + cluster label
    - Cluster-wise summaries for tenure, workload, promotions, and exit flags
    - Action plan grid for each segment
- **Shiny App (Optional)**:
    - Upload HR data and visualize segmentation in-browser
    - Includes sliders to test workload/promotion impact scenarios

8. Business Impact

- Cluster-based policies led to a **19% attrition drop in Cluster 3** over the next two quarters
- Project reassignment and mental health support introduced for high-risk roles
- Segmentation model adopted by HR analytics team for **quarterly attrition prediction**
- Framework reused for **diversity gap analysis** in performance appraisal review