Project Overview Document

1. Introduction & Domain

This project focuses on the Human Resources (HR) analytics domain. The main motivation is to explore employee retention, performance, and workforce planning. The scope includes building a relational database from IBM HR data supplemented with synthetic data to cover missing attributes such as names, job titles, and performance reviews.

2. Data Description

The primary dataset is IBM HR analytics, supplemented with synthetic data generated through SQL and Python. Eight core tables were developed: employees, departments, jobs, performance_reviews, salaries, attrition, training, and projects. Key attributes include employee demographics, hire dates, salaries, attrition status, and performance data. Limitations: names and job titles were randomly generated and do not reflect real individuals.

3. Database Design & Implementation

The schema was normalized to reduce redundancy, with clear relationships among tables. Primary and foreign keys were implemented to enforce integrity. Data was seeded synthetically where IBM data was incomplete. For example, manager assignments were generated and reviewer IDs in performance reviews were linked to department managers.

4. SQL Queries

Fifteen queries were developed to showcase capabilities including joins, subqueries, aggregations, window functions, rollups, and case statements. These queries deliver business insights such as attrition by department, salary distribution, tenure analysis, and performance trends. They provide actionable insights to HR decision-makers.

5. Python/R Integration

PostgreSQL was integrated into Jupyter Notebook using psycopg2 and SQLAlchemy. Data was retrieved into pandas DataFrames for analysis. Analysis included descriptive statistics (mean, median, distribution of salaries, attrition rates) and inferential statistics such as correlations between tenure, performance, and attrition.

6. Data Visualization

Visualization was performed using matplotlib and seaborn. Key visuals include attrition rate by department, salary distribution histograms, tenure vs. attrition scatterplots, and performance score trends. These visualizations highlight workforce risk areas and compensation insights more effectively than raw tables.

7. Challenges & Resolutions

Challenges included handling missing job titles, generating realistic synthetic data, and resolving SQL insert mismatches. Resolutions involved using faker libraries for names, assigning managers by department, and adjusting schema/insert statements carefully.

8. Conclusion

The project demonstrates the ability to design a relational HR database, generate synthetic data, construct advanced queries, and integrate SQL with Python for analysis and visualization. Insights gained can help HR identify turnover risks, align salary structures, and monitor performance trends. Future work could include predictive analytics and deployment in BI tools like Tableau or Power BI.