Lab 3 - Data Engineering & EDA Submission

Student Name: Ali Cihan Ozdemir

Student ID: 9091405

Date: 2026-02-06

GitHub Repository

https://github.com/alicih4n/Lab3-DataEngineering.git

Project Summary

1. Cloud Database & Data Generation: Established a PostgreSQL connection on Neon. Created a Python script (lab3_sdg.py) to generate 500 synthetic employee records. Implemented 20% 'dirty data' logic (missing values, inconsistent casing, invalid dates) to simulate real-world data engineering challenges.

2. Data Wrangling & Cleaning: Developed a Jupyter Notebook to extract data from the cloud. Used Pandas to inspect quality, impute missing salaries using position-based medians, standardize job titles, and filter out logical date errors.

3. Feature Engineering & Scaling: Created 'start_year' and 'years_of_service' features. Applied Z-Score standardization (StandardScaler) to salary data to prepare it for potential Neural Network applications.

4. Visual Intelligence: Built advanced visualizations including a Grouped Bar Chart for salary trends and a FacetGrid Heatmap for departmental salary distributions using a synthetic SQL-style join.

Contribution Validation

Please note that this project was completed entirely by Ali Cihan Ozdemir.
Group partner Roshan was absent from class and did not contribute to this lab submission.