# Team A – Project Idea Introduction 02

Team Members: Gelareh Vakili, Simran Shah

Date: June 09, 2025

## Project Title: Salary Prediction Based on Developer Profiles

### Primary Dataset: Stack Overflow Developer Survey 2024

Dataset Source: Stack Overflow Developer Survey 2024  
Link to Dataset (Direct CSV Download): https://survey.stackoverflow.co/2024/survey\_results\_public.csv.zip  
  
Dataset Info:  
- Format: CSV (compressed in ZIP)  
- Size: Approximately 85MB uncompressed  
- Number of responses: ~89,000  
- Relevant Attributes:  
 • ConvertedCompYearly – Target variable: total yearly compensation  
 • Country, EdLevel, YearsCodePro, Employment, DevType – Key predictors  
 • Other features: tools, technologies used, remote work ratio, demographics  
  
Planned Analysis:  
We will perform a regression analysis to predict annual salary based on experience, education level, employment type, and developer type. The model will explore relationships between skillsets and compensation across countries.  
  
Analysis Techniques:  
- Preprocessing: Data cleaning, handling missing values, and one-hot encoding  
- Models: Linear Regression, Random Forest, and XGBoost Regressor  
- Visualizations: Correlation heatmaps, salary distribution plots, feature importance  
  
Performance Evaluation Metrics:  
- Root Mean Squared Error (RMSE)  
- Mean Absolute Error (MAE)  
- R² Score  
  
Why This Dataset?  
This dataset offers a comprehensive, recent snapshot of tech professionals globally and includes granular, well-labeled data relevant to compensation modeling. It is available in CSV format and does not originate from Kaggle, ensuring compliance with course policy.

### Backup Dataset: US Bureau of Labor Statistics (BLS) OEWS 2023

Dataset Source: U.S. Bureau of Labor Statistics – Occupational Employment and Wage Statistics (OEWS)  
Link: https://www.bls.gov/oes/tables.htm  
  
Dataset Info:  
- Format: XLSX and CSV by occupation and state  
- Attributes: Job title, mean/median salary, state, region, industry code  
  
Planned Use:  
We will use the dataset to predict or analyze salary trends based on job title, location, and industry sector. While more aggregated, this dataset complements our primary focus and is suitable for statistical modeling or clustering.  
  
Potential Analysis:  
- Clustering of roles based on salary and location  
- Regional comparisons and inflation-adjusted trends  
- Classification of roles by pay grade  
  
Performance Evaluation Metrics:  
- R² Score for regression-style tasks  
- Silhouette score or Davies–Bouldin Index for clustering  
  
This serves as a robust non-Kaggle alternative should the primary dataset prove too complex or incomplete for the required analysis.