

## **CSCI 5502 — Milestone 0 Project Proposal**

**Group #: 7**

**Team Members:** Soorej S Nair, Anjana Anand, Nivid Pathak, Karan Cheemalapati

### **1) Project Description**

#### **Project Goal:**

Understand labor market structure and build models that predict and explain wage variation.

Our project investigates what factors drive wage differences across occupations in the United States using official labor statistics. We aim to build data mining and machine learning models that analyze how wages vary based on occupation type, industry, and geographic location. The main goal is to identify which variables most strongly influence salary levels and whether we can predict wages using these labor market characteristics.

This project matters because understanding wage patterns helps students, job seekers, policymakers, and organizations make informed decisions about careers, workforce planning, and economic trends. We will perform exploratory data analysis, regression modeling to predict wages, and clustering to identify groups of similar occupations.

### **2) Data Sources**

**Dataset:** Occupational Employment and Wage Statistics (OEWS)

**Publisher:** U.S. Bureau of Labor Statistics

**Link:** <https://www.bls.gov/oes/tables.htm>

#### **Estimated Size:**

~800 occupations × 20

- 40 variables depending on table

#### **Key Variables:**

- Occupation Title
- Industry Code
- Location (State / Metro)
- Mean Annual Wage
- Median Wage
- 10th / 25th / 75th / 90th Wage Percentiles
- Employment Level

#### **Access Constraints:**

Publicly available data, no login required. Some values may be missing or suppressed for privacy in small employment groups.

### **3) Potential Bias / Limitations**

#### **1. Wage Suppression Bias**

Some wages are suppressed or rounded in small employment groups. This may distort salary

predictions for niche occupations.

*Mitigation:* Remove or impute missing values and analyze only reliable groups.

## **2. Industry Aggregation Bias**

Industries are broad categories that may hide role-specific differences.

*Mitigation:* Use multiple features and clustering to capture sub-group patterns.

## **3. Geographic Cost-of-Living Differences**

High wages in certain states may reflect living costs rather than job value.

*Mitigation:* Normalize wages by region and compare relative trends.