# **Analytics Capstone**

# **Data Preparation Report**

Predictability of Career and Professional Excellence

Group 4

**Introduction**

In today's volatile work market, knowing the geography of potential possibilities and the variables impacting career predictability is critical. With the introduction of platforms such as LinkedIn, massive volumes of data are created every day, offering important insights into job listings, business profiles, and industry trends. This dataset, which includes over 33,000 job posts gathered over two consecutive time periods, provides a complete picture of the present employment market. Each ad includes key details such as title, description, pay information, and location, making it a valuable resource for research and analysis.

With this dataset, we hope to investigate numerous areas of career predictability and professional excellence, revealing trends, patterns, and insights that can help shape career decisions and organizational strategy. The opportunities for study are extensive, ranging from investigating the top paid occupations to projecting salary using natural language processing (NLP). Furthermore, the information allows for the examination of firm profiles, benefit offers, and workforce dynamics, which sheds insight on the changing nature of job options.

**Data Sets:**

The dataset LinkedIn Job Postings – 2023 will be utilized for modeling and analysis in the project.

**Dataset Description:**

The dataset used for this project consists of over 33,000 job postings sourced from LinkedIn, spanning two distinct time periods. Each job posting includes extensive attributes such as title, description, salary details, location, and application URLs. Supplementary datasets contain company profiles with information on company descriptions, sizes, headquarters location, and employee counts. These datasets enable comprehensive exploration of job market dynamics, including trends in compensation, benefits, and company demographics. Together, they form a rich source of information for modeling career predictability, professional excellence, and industry trends.

**Data Preparation:**

A diagram of a data flow

Description automatically generated

**Select data:**   
  
Include relevant features like job title, description, salary details, location, and application URLs. Exclude irrelevant characteristics or those with many missing values.   
Reasons: Specific characteristics are critical for studying job market trends and forecasting career outcomes. Excluded characteristics lack relevance or have poor data quality.

We will include the following columns for analysis:

job\_id: Unique identifier for each job posting.

title: Job title.

max\_salary, med\_salary, min\_salary: Salary information.

formatted\_work\_type: Type of work.

location: Job location.

formatted\_experience\_level: Required experience level.

skills\_desc: Description of required skills.

We will exclude columns such as **company\_id, applies,** **original\_listed\_time,** **remote\_allowed, views,** **job\_posting\_url,** **application\_url,** **application\_type, expiry,** **closed\_time,** **posting\_domain, sponsored, currency,** **compensation\_type, and scraped** as they are not directly relevant to the analysis goals or contain redundant information.

**Clean data:**   
  
**Actions:** Remove entries with missing values in essential characteristics, use median values to fill in missing data in salary columns, and fix formatting issues.

Data Cleaning Report:

Missing Values: We will handle missing values in salary information and skill descriptions by imputation or removing incomplete records.

Outliers: We will identify and address outliers in salary data if present.

Data Consistency: We will ensure consistency in formatting of categorical variables like work type and experience level.  
**Rationale:** Improve data quality for reliable analysis and modeling by assuring consistency and completeness.

**Construct Data:**   
  
**Derived characteristics:** Create new characteristics based on job descriptions, such as experience level, or dummy variables for different work kinds.   
**Generated Records:** To gather broad industry information and include data for organizations that do not have any public job openings.   
**Purpose:** Expand the dataset with new attributes to improve analysis and modeling capabilities.

**Integrate data.**   
  
**Merged Data:** Combine job advertisements with company profiles to include industry, size, and location information.   
**Aggregations:** Create summary statistics from job postings, such as the average income per industry or the distribution of job kinds by location.   
**Outcome:** Create a consistent dataset for extensive analysis, including data from many sources.

**Format Data:**   
  
**Reformatted Data:** To fulfill the modeling tool's criteria, ensure uniformity in attribute order and data format.   
**Syntactic Changes:** Standardize text fields by deleting special characters and guaranteeing consistent length.   
**Goal:** Prepare the dataset for easy incorporation into modeling methods, hence improving performance and accuracy.

**Conclusion:**

As we end our exploration of this vast dataset, it becomes clear that the field of career predictability and professional success is varied and dynamic. Analyzing job advertisements, business profiles, and industry trends yields vital data for job seekers, employers, and legislators alike. From detecting developing job titles to evaluating the influence of remote work on employment trends, the dataset presents a comprehensive picture of the current job market scenario.  
Future updates and research of this dataset are expected to provide even more detailed insights on time-based patterns, corporate growth trajectories, and shifts in demand for certain skills and credentials. By harnessing the richness of information included in this dataset, stakeholders may make educated decisions that promote career growth, organizational success, and economic prosperity. As we continue to navigate the changing terrain of work and employment, this dataset is a valuable tool for determining the future of careers and professional success.