Data Cleaning & Exploration

Preparing and Preprocessing the Job Market Dataset

October 8, 2025

title: “Introduction: Group 5” subtitle: “About The Team” —

**Team:** Connor Coulter, Wei Wang, Balqis Bevi Abdul Hannan Kanaga

**Topic:** AI vs. Non-AI Job Growth — Is AI taking over or creating more jobs?

**Course:** AD 688 – Cloud Analytics for Business

This site hosts our research rationale, intro, and literature review for Project Selection I, II and III.

# Import Data

Loaded dataset: (72498, 131)

/tmp/ipykernel\_11165/4082786718.py:32: FutureWarning:  
  
A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.  
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.  
  
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

# Data Cleaning & Preprocessing

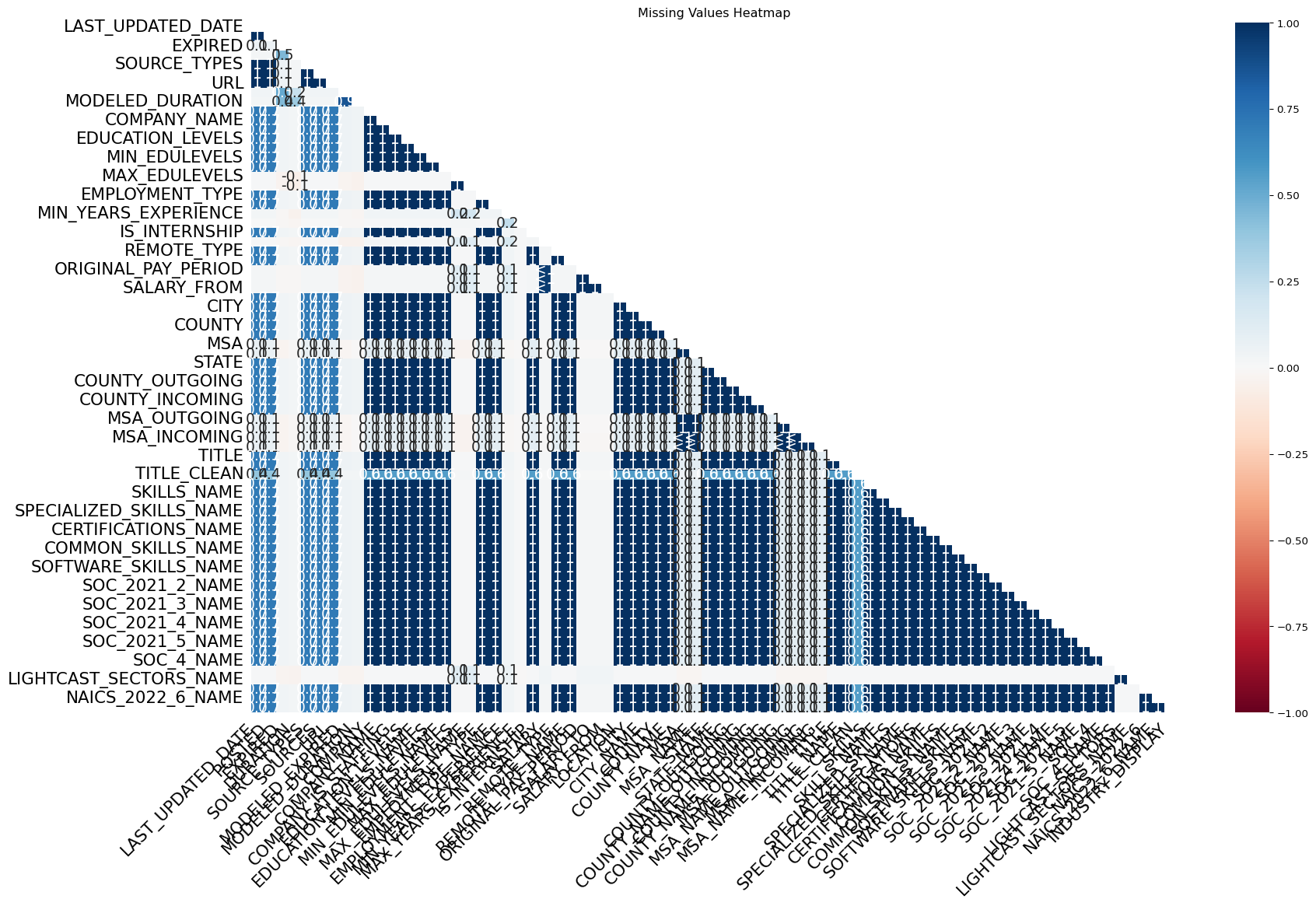
## Drop Unnecessary Columns

Derived non-null: {'INDUSTRY\_DISPLAY': np.int64(72454), 'SALARY\_DISPLAY': np.int64(72498)}

## Drop Unnecessary Columns

Remaining columns (first 30): ['LAST\_UPDATED\_DATE', 'POSTED', 'EXPIRED', 'DURATION', 'SOURCE\_TYPES', 'SOURCES', 'URL', 'MODELED\_EXPIRED', 'MODELED\_DURATION', 'COMPANY', 'COMPANY\_NAME', 'COMPANY\_IS\_STAFFING', 'EDUCATION\_LEVELS', 'EDUCATION\_LEVELS\_NAME', 'MIN\_EDULEVELS', 'MIN\_EDULEVELS\_NAME', 'MAX\_EDULEVELS', 'MAX\_EDULEVELS\_NAME', 'EMPLOYMENT\_TYPE', 'EMPLOYMENT\_TYPE\_NAME', 'MIN\_YEARS\_EXPERIENCE', 'MAX\_YEARS\_EXPERIENCE', 'IS\_INTERNSHIP', 'SALARY', 'REMOTE\_TYPE', 'REMOTE\_TYPE\_NAME', 'ORIGINAL\_PAY\_PERIOD', 'SALARY\_TO', 'SALARY\_FROM', 'LOCATION']

## Handle Missing Values



/tmp/ipykernel\_11165/2447604449.py:10: FutureWarning:  
  
A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.  
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.  
  
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.  
  
  
  
/tmp/ipykernel\_11165/2447604449.py:13: FutureWarning:  
  
A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.  
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.  
  
For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

## Remove Duplicates

Removed 3300 duplicates using ['TITLE', 'COMPANY\_NAME', 'LOCATION', 'POSTED']