**(2) Describe the steps for data preparation and preprocessing**

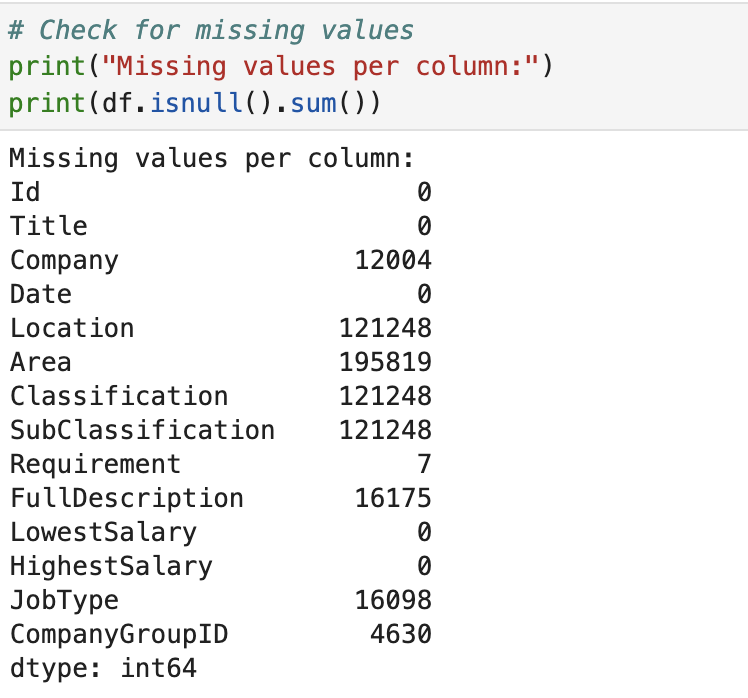


Figure 1 Numbers of missing values after infer\_company function is applied

The dataset was initially loaded using Pandas with error handling for malformed rows. Following initial exploration using .shape, .info(), and .describe(), the author adopted a flexible missing data strategy. Instead of discarding rows, important fields such as Title, Company, Requirement, and salary fields were retained even when incomplete. Meanwhile, columns with high missingness (like Area and SubClassification) were excluded from core analyses but preserved for potential inference. A key enhancement was the creation of the CompanyGroupID column. This was primarily derived from the Company field, and missing values were inferred using a custom function leveraging related columns (Requirement, Classification, and Location). Inference included state abbreviation mapping and general category labeling, effectively reducing missing rates for company-related information (Figure 1).

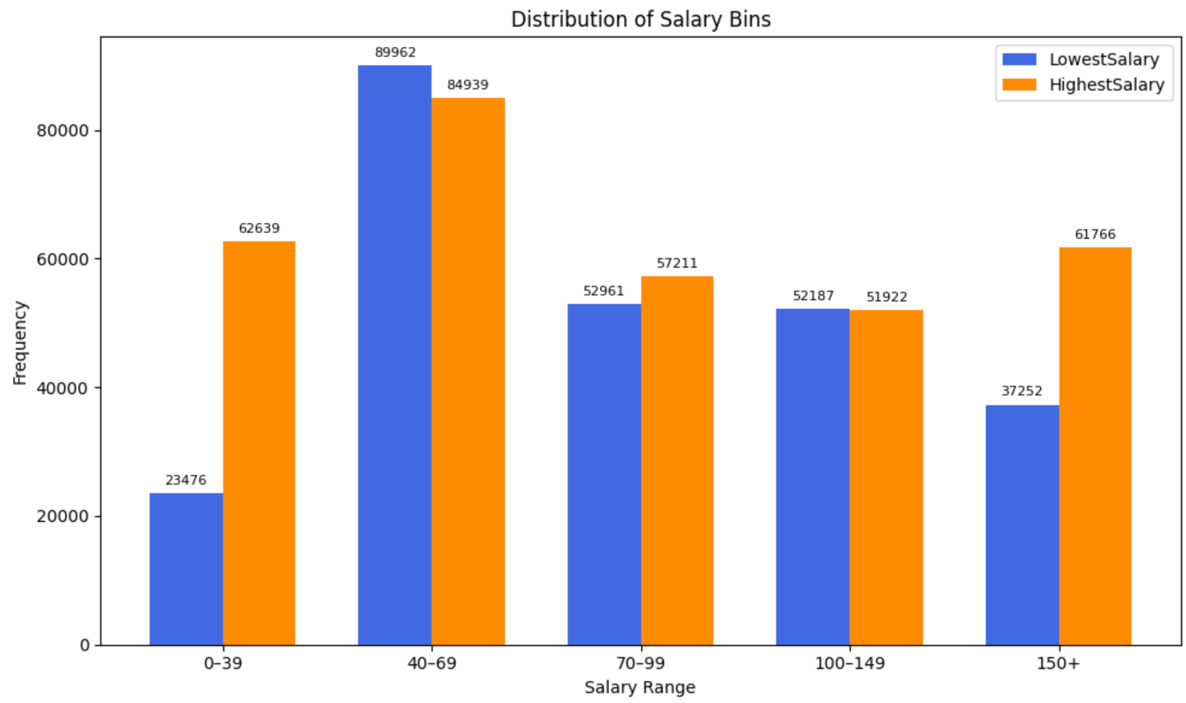


Figure 2 A salary bin chart shows a skewed distribution

Textual normalization was applied instead of numerical scaling. This included standardizing state names and binning salary ranges into five categories: 'Very Low' (0–39k), 'Low' (40–69k), 'Medium' (70–99k), 'High' (100–149k), and 'Very High' (150k+), to support clearer analysis and visualization (Figure 2). To verify the nature of the salary distribution, a Shapiro-Wilk test was conducted on the LowestSalary field. The extremely low p-value (< 0.05) confirmed non-normality, which aligned with the observed right-skewed distribution—characterized by a cluster between 30–120k and rare but extreme outliers near 999k. The preprocessed dataset was saved for use in Task 2.