The dataset contains 318,477 job postings, each represented as a row with various features such as job title, company, salary range, location, and classification. The data was provided in CSV format and includes over 13 columns, consisting of both categorical and numerical data types.

According to the data type distribution, most of the columns are of type **object**, which includes textual or categorical data like Title, Company, Location, and Requirement. Only two columns, LowestSalary and HighestSalary, are of type \*\*int64`, representing numerical salary ranges.

A missing value analysis reveals that certain fields contain significant amounts of missing data. In particular:

* Area has over **190,000** missing values,
* Location, Classification, and SubClassification each have over **100,000** missing values,
* FullDescription, JobType, and Company have relatively smaller but still notable numbers of missing entries,
* Requirement is mostly complete with minimal missing data.

These visual insights justify the decision to retain only the most relevant and complete columns for core analysis: **Title, Company, Requirement, LowestSalary, HighestSalary**

Meanwhile, other fields such as Location, Area, Classification, SubClassification, FullDescription, and JobType are used only for descriptive summaries or optional visualization and are not included in modeling due to their high proportion of missing values or limited predictive value.

A graph with blue rectangular bars

AI-generated content may be incorrect.

A graph with blue bars

AI-generated content may be incorrect.