Russell Chan
Ana Paula De Queiroz
Gabrielle Salamanca
Catarina Tegtmeier

Group Initial Data Exploration Summary Report

Initial Findings

The data set is quite large, with 167278 unique jobs, and 27 columns of related information for each item. Many however are incomplete, or contain outliers - for instance, only around ~5000 jobs contain information about the experiences required to apply, which implies other positions left this value blank. Salaries vary all the way from ~10k up to ~320k, either of which may be outliers that need to be removed. We also take out irrelevant job subgroups which include attorney, teacher, and assistant professors. We decided to keep the management consultant subgroup since we believe it is a data related jobs. We also decided to only focus on the variables that were important to our questions: employer name, work state, work city, paid wage per year, and job title subgroup. We may include further variables as needed as we develop further questions.

Important Details

As many questions include a consideration regarding cost-of-living standards, we decided to include another data set ('Cost of Living 2023.csv') which gives the cost index for every state. We will include this dataset in our repo so we can analyze the cost of living and potential other values.

The variation of the data turned out to be huge, so we moved on to do an outlier analysis. We plotted the distribution of the response variable and a box plot to look at the outliers. The bulk of the salary distribution resides between $66,000 and $104,000 and we see a highly skewed distribution with a minimum of $20,000 to a max of $1,400,000 per year. A more detailed analysis with graphs will be done in future.

5 fun facts:

- Distribution of the response variable is highly skewed to the right. We have outliers ranging from a salary of $20,000 year to $1,400,000 year within data related job titles.
- Of all jobs offered to foreign professionals, 79.52% are data related jobs while 20.48% are not.
- Software engineers & business analysts are the top 2 jobs in job titles and its subgroups, dwarfing the other categories.
- The US territories Puerto Rico and Guam are included in this dataset as states. They are the lowest paid locations.
- The top three job titles were software engineer nearly at $60,000, business analyst at about $20,000, and senior software engineer at about $14,000.