**Assignment Report**

**Chanutda Manosorn (s5313405)**

**Jehyun Kim (s)**

Big Data Analysis – 3803ICT

[Part 1 – Data Preparation and Preprocessing 2](#_Toc197099216)

[(1) Describe the dataset 2](#_Toc197099217)

[(2) Describe the steps for data preparation and preprocessing 4](#_Toc197099218)

[(3) Hypotheses about the analysis outcome 6](#_Toc197099219)

[Part 2 – Data Analysis and Interpretation 7](#_Toc197099220)

[(1) Relevant information in job metadata 7](#_Toc197099221)

[(2) Job market by location 8](#_Toc197099222)

[(3) Job market by sector 11](#_Toc197099223)

[(4) Visualize the results on an interactive visualization 12](#_Toc197099224)

[Part 3 – Evaluation 13](#_Toc197099225)

[Part 4 – Case studies 14](#_Toc197099226)

# **Part 1 – Data Preparation and Preprocessing**

## **(1) Describe the dataset**

The dataset contains job advertisements with a variety of features including job titles, companies, locations, descriptions, and salary ranges. It appears to have been extracted from a job portal or employment listings. It contains 318,477 job postings, each represented as a row with various features such as job title, company, salary range, location, and classification (Figure 1). The data was provided in CSV format and includes over 13 columns, consisting of both categorical and numerical data types.

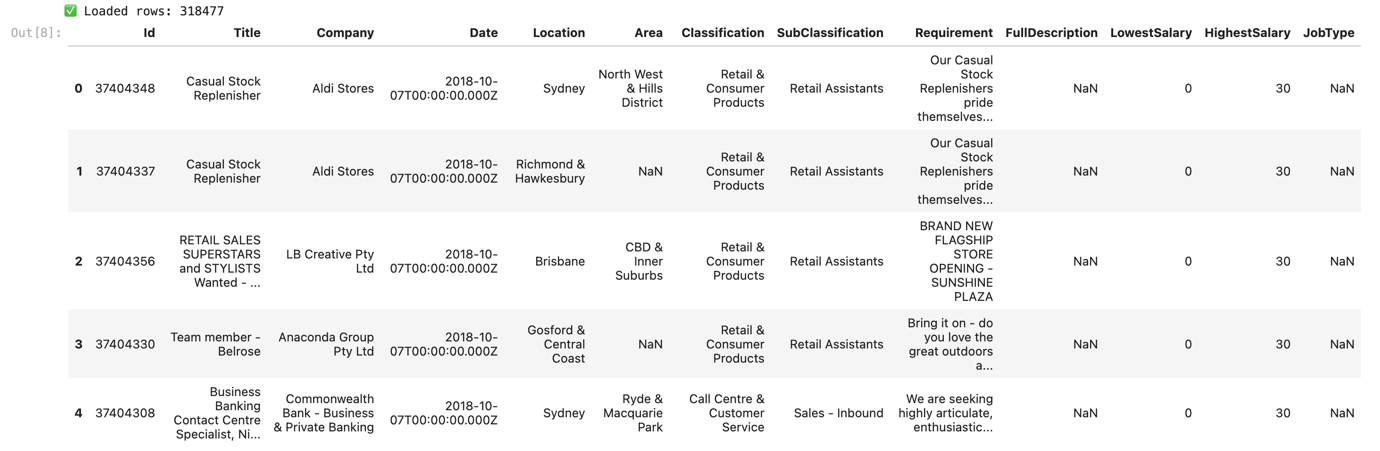


Figure 1 Sample Data Frame

According to the data type distribution (Figure 2), 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 graph with blue rectangular bars

AI-generated content may be incorrect.

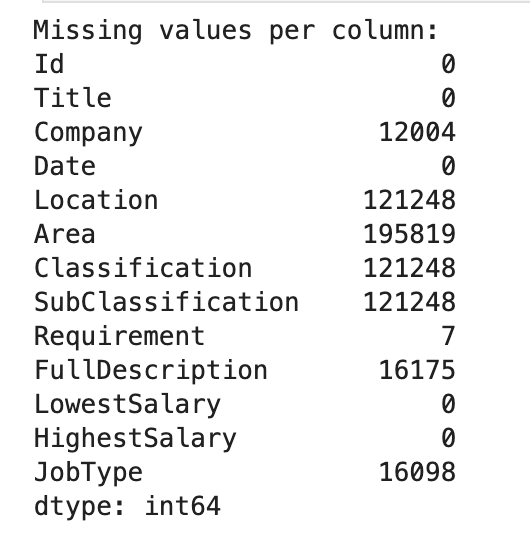
Figure 2 Numbers of data types

A missing value analysis reveals that certain fields contain significant amounts of missing data (Figure 3). 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.

The 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. However, there’s no need to drop those missing rows using df\_clean = df.dropna() as it would affect the overall data shape. Therefore, they can benefit in key analyses, but could benefit inference for some missing value in Company**.**

A graph with blue bars

AI-generated content may be incorrect.****

*Figure 2 Missing Values per Column*

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

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 3).

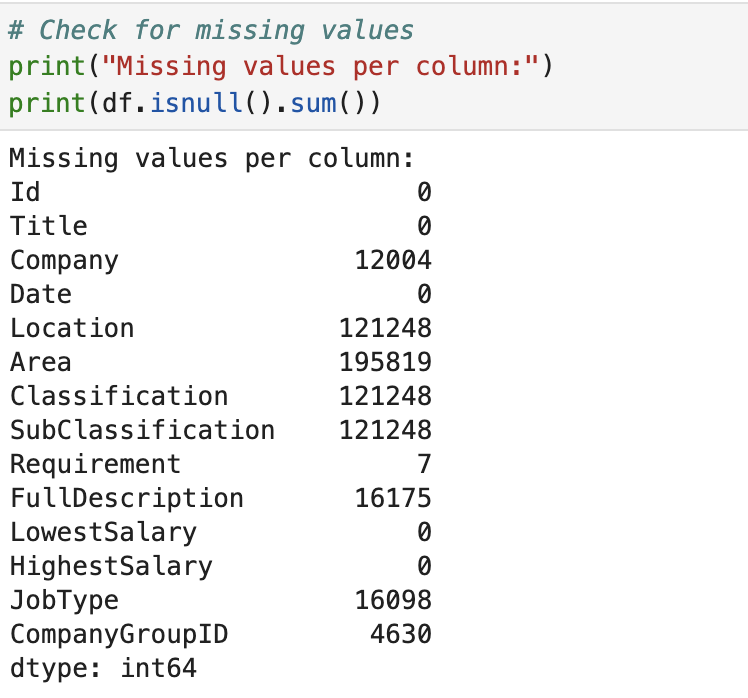


Figure 3 Numbers of missing values after infering CompanyGroupID

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 4). 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 pre-processed dataset was saved for use in Task 2.

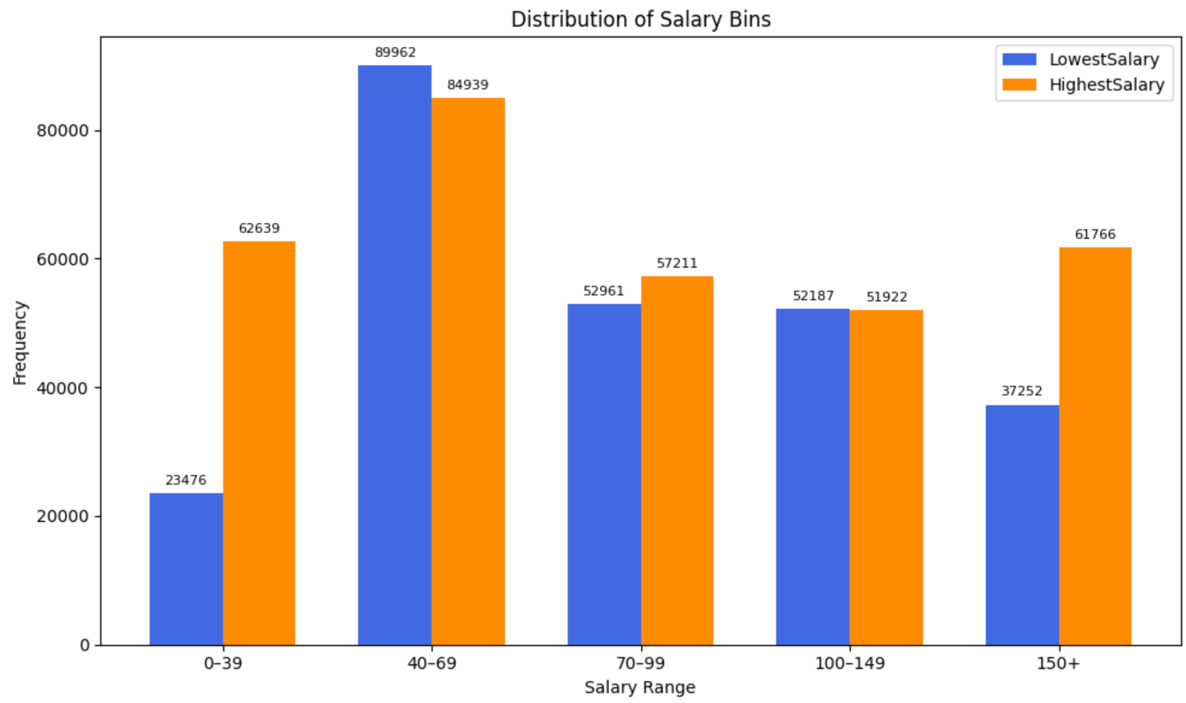


Figure 4 Skewed distribution in salary bins chart

## **(3) Hypotheses about the analysis outcome**

1. Job titles and company identities are expected to serve as significant predictors of salary ranges, reflecting industry-specific compensation standards and hierarchical job structures (e.g., higher salaries typically found in IT or Mining compared to Retail or Customer Service).
2. The Requirement field may reveal patterns of qualification expectations when analysed alongside job titles, potentially identifying skill-level groupings, educational preferences, or experience benchmarks commonly associated with certain roles or companies.
3. Despite high missingness in geographic fields, aggregate-level location trends (e.g., distribution of postings across states or regions) may still offer useful insights, particularly in identifying job market concentrations and regional disparities in employment offerings.

**Part 2 – Data Analysis and Interpretation**

**(1) Relevant information in job metadata**

**<Jeh>**

**(2) Job market by location**

The Australian job market displays a strong geographic concentration, particularly within the eastern states of New South Wales, Victoria, and Queensland. Figures 5 and 6 highlight that NSW alone recorded over 73,000 job postings, with Sydney accounting for the majority. Melbourne and Brisbane also serve as central hubs in VIC and QLD respectively. These three cities dominate their regional markets, while in contrast, states like WA and SA present a more distributed job structure. Despite 38% of the data labelled as "Unknown," the remaining dataset reliably captures consistent trends in market concentration, job centralization, and regional employer clustering.

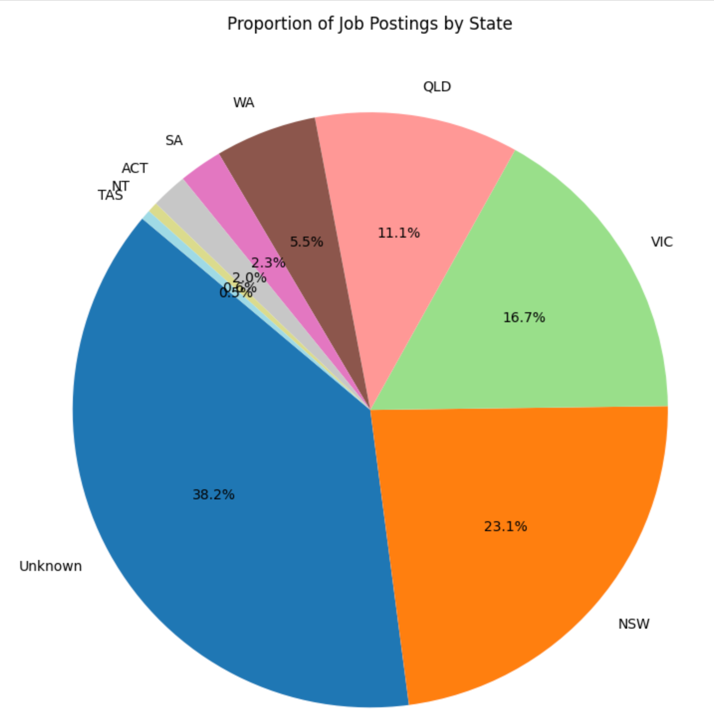


Figure 5 Proportion of jobs by state

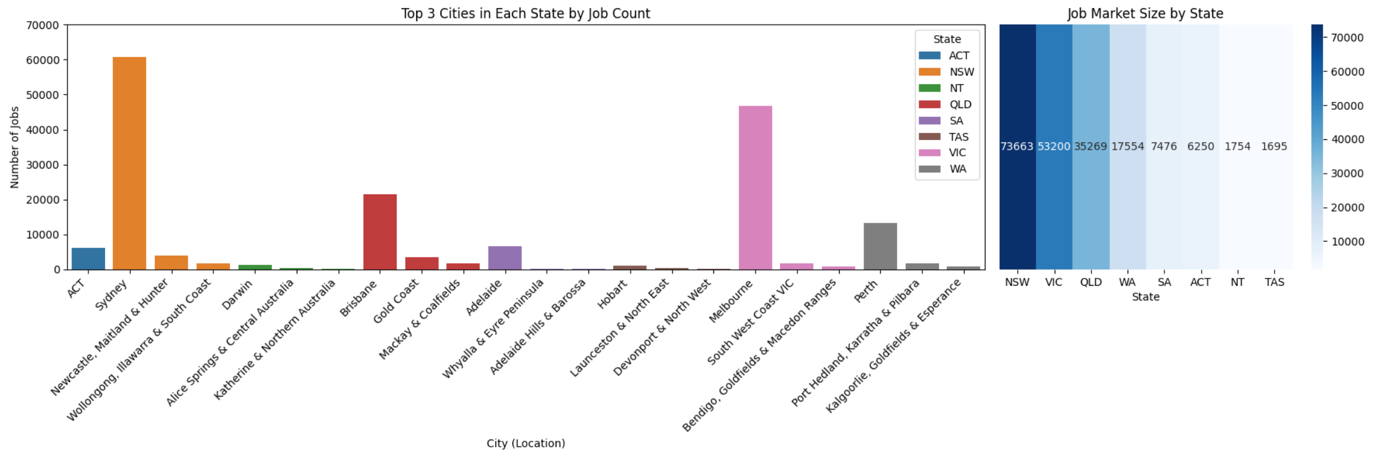


Figure 6 Top 3 jobs hub cities by state

Figures 7 explores salary differences by state and employer. While ACT shows the highest overall average salary when considering all job postings, Figure 4 reveals that NT features extreme outliers—most notably, Paxus—offering exceptionally high salaries likely due to specialized or high-demand roles. These outliers in NT contrast with ACT’s more consistent salary range across top companies. These findings reinforce that salary levels are influenced by both geographical context and specific employer practices.

A graph with different colored circles

AI-generated content may be incorrect.

Figure 7 Major city hubs by state

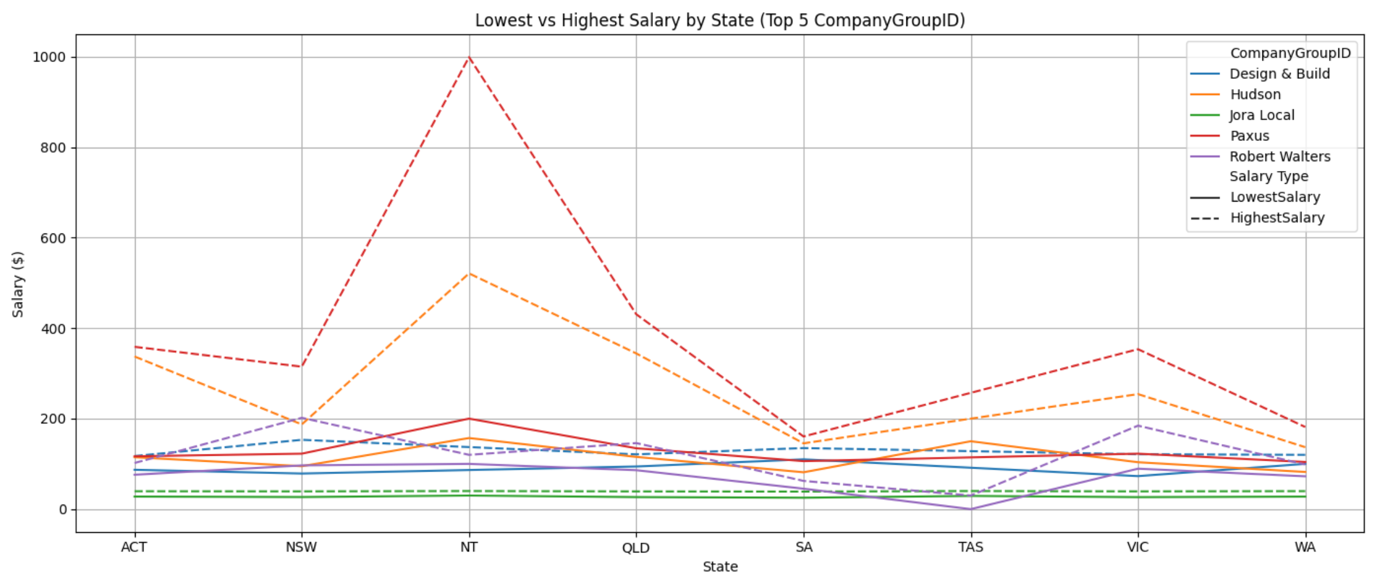


Figure 8 Lowest and Highest salaries comparison amongst top 5 companies

Figures 9 to 11 shift focus to salary distributions and requirement language. Salaries across states were compared using salary bins, revealing that high-paying roles (over 150K) consistently yield the highest averages, particularly in ACT and NT. These positions are often associated with government, healthcare, and project-based roles, as suggested by frequent keywords such as “contract,” “project,” and “health.” This aligns with earlier findings showing high salaries in NT being driven by specific companies like Paxus. Time-series analysis of job postings (Figure 10) shows consistent spikes at the beginning of each month (red dash-line), reflecting structured recruitment cycles. Finally, TF-IDF analysis of requirement keywords (Figure 11) indicates that terms like “team,” “opportunity,” and “role” are common nationwide, while high-salary bins tend to feature more specialized or formal terms, suggesting a strong link between job description language and salary expectations.

A graph of a salary

AI-generated content may be incorrect.

Figure 9 Average salary ranges by state

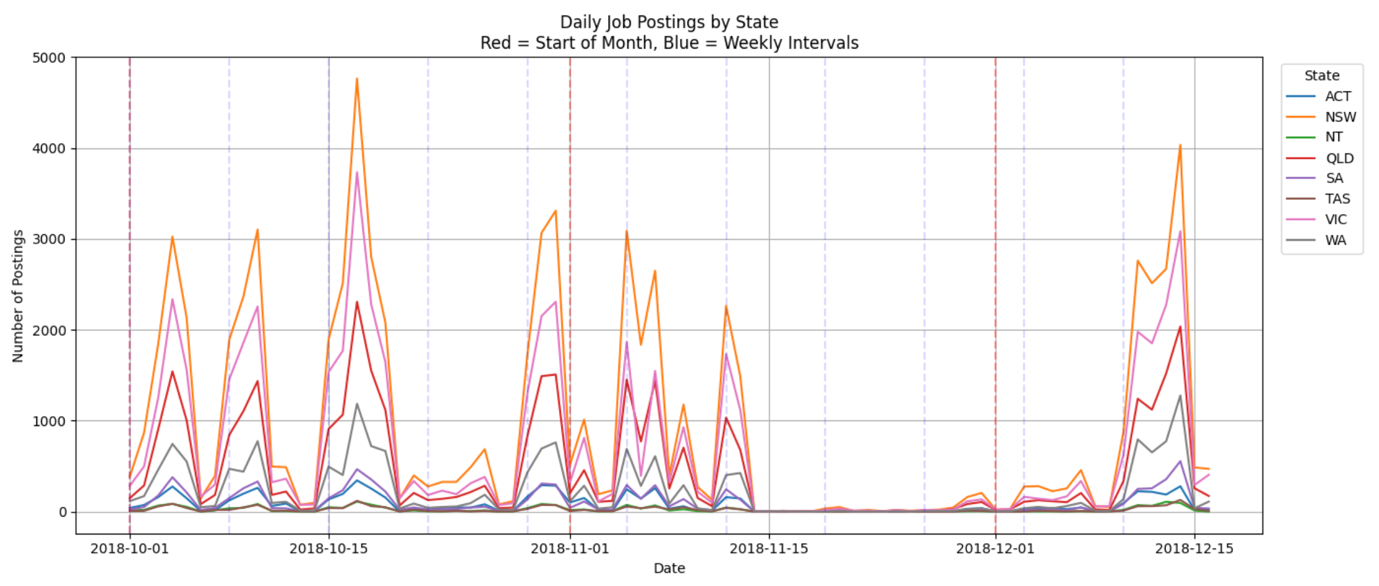


Figure 10 Job advertisement behaviors by state



A close-up of words

AI-generated content may be incorrect.

Figure 11 Top 20 requirement keywords by state and salary bin

**(3) Job market by sector**

**(4)** Visualize the results on an interactive visualization

# **Part 3 – Evaluation**

(1) Findings in the job market’s data

(2) Suggested actions for balancing the market

(3) Data analytics refinement

(4) Implications for employees and employers based on findings

# **Part 4 – Case studies**

(1) Case study 1

(2) Case study 2