# **Analytics Capstone**

# **Data Understanding Report**

Predictability of Career and Professional Excellence

Group 4

# ­­

**Introduction:**

The first step in understanding information is to have a thorough overview of the dataset that comprises LinkedIn job ads. This dataset contains more than 33,250 job advertisements that were collected on two separate days, separated by many months. There are 28 features that come with every job posting on the platform, offering a comprehensive understanding of the many aspects of career possibilities that are accessible. These characteristics include basic details like the job title, description, pay, location, and application URL in addition to more detailed information like the kind of employment (e.g., contract or remote). Through an analysis of this dataset, stakeholders may obtain significant insights on current trends in the job market, corporate preferences, and skill needs. This information can then be used to guide strategic planning and decision-making related to talent acquisition and workforce development initiatives.

**Data Understanding:**

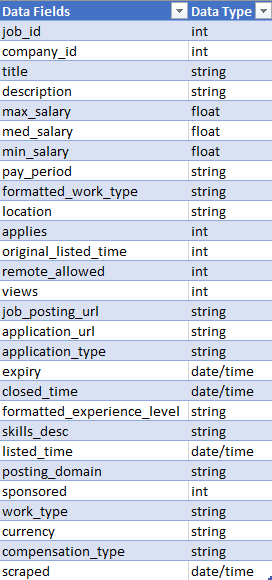
A diagram of data processing

Description automatically generated

**Collect Initial Data:**

**Task:** Retrieve the first dataset, which consists of more than 33,000 LinkedIn job posts and over 28 different data fields.

**Data:**



**Output:** Obtain a structured dataset containing 28 attributes for each job posting, including title, description, salary, location, application URL, work types, benefits, skills, industries, company details, and follower count.

**Describe Data:**   
  
**Task:** To summarize the properties of the dataset and conduct data profiling and descriptive statistics.

The dataset consists of the following columns:

**job\_id**: Unique identifier for each job posting.

**company\_id**: Identifier for the company posting the job.

**description**: has the description of the job

**title**: Job title.

**max\_salary,** **med\_salary,** **min\_salary**: Maximum, median, and minimum salaries.

**pay\_period**: Frequency of salary payment (e.g., hourly, monthly).

**formatted\_work\_type**: Type of work (e.g., full-time, contract).

**location:** Job location.

**applies:** Number of applications.

**original\_listed\_time:** Timestamp of when the job was originally listed.

**remote\_allowed:** Indicates if remote work is allowed.

**views**: Number of views.

**job\_posting\_url:** URL of the job posting.

**application\_url:** URL for applying to the job.

**application\_type**: Type of application.

**expiry,** **closed\_time**: Timestamps indicating job expiry and closure times.

**formatted\_experience\_level:** Required experience level.

**skills\_desc:** Description of required skills.

**listed\_time:** Timestamp of when the job was listed.

**posting\_domain:** Domain of the job posting.

**sponsored**: Indicates if the job posting is sponsored.

**currency**: Currency of the salary.

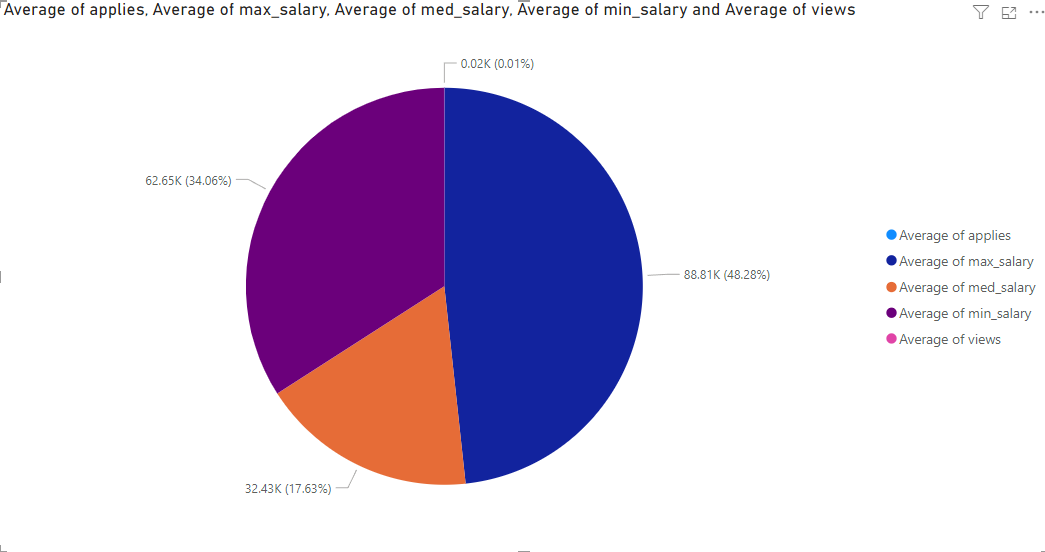
**compensation\_type**: Type of compensation.

**scraped**: Timestamp of when the data was scraped.

  
**Output:** Summary statistics for numerical characteristics, including the mean, median, mode, range, standard deviation, and percentiles. Furthermore, value ranges, frequency distributions, and unique value counts for categorical properties.

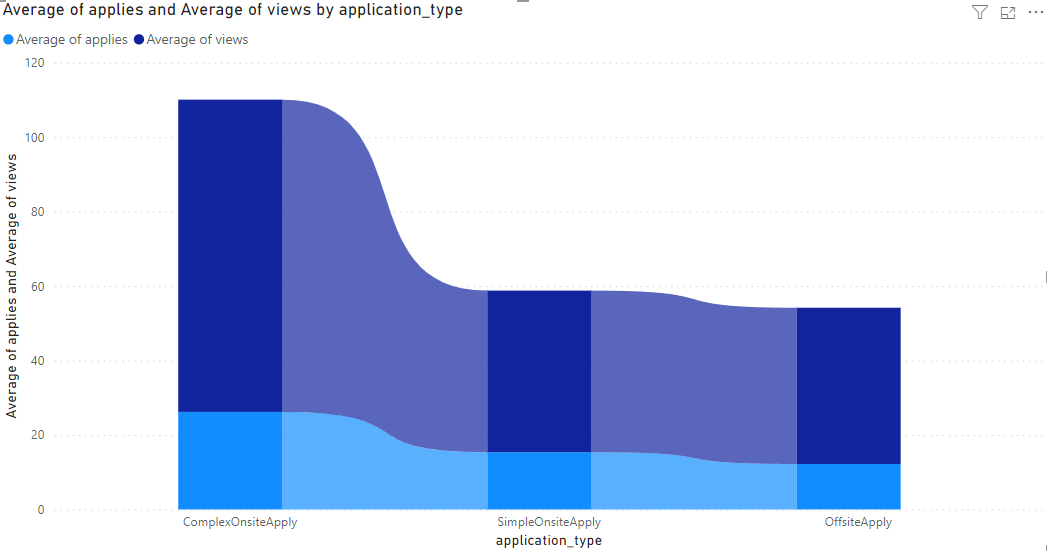
**Explore Data:**

**Task:** To learn more about the dataset's many components, including job titles, salaries, locations, industries, and firm profiles, do exploratory data analysis (EDA).



Here we can see the percentages of the average of views, salaries, applies

Also, we can see the average applies and average views with respect to the application types.



**Output:** Box plots, scatter plots, histograms, and bar graphs are used to visualize key features. Examine the data for correlations, outliers, trends, and patterns.

**Verify Data Quality:**

**Task:** Examine the dataset's correctness, consistency, dependability, and completeness to determine its quality.

**Note:** Although the quality of data is not that great but applying the data preparation methods to understand it properly will be a good next step.

**Output:** Detailed results of completeness, consistency, correctness, and reliability tests are included in the data quality assessment report, along with suggestions for data cleaning and preprocessing if needed.

**Conclusion:**

A thorough overview of the dataset containing job listings from LinkedIn has been supplied by the data comprehension phase, providing the groundwork for further analysis and insights. We have thoroughly examined the information to obtain a comprehensive grasp of its major qualities, structure, and insights into the modern work scene. With this information in hand, stakeholders may use the dataset's abundance of data to their advantage as they do more research and analysis to identify trends, patterns, and useful insights. We anticipate drawing important conclusions from the data as we go on to the following stages of research and interpretation. These conclusions will help us make strategic decisions and spur innovation in workforce development and talent acquisition strategies.