# NEXT LEAP GRADUATION PROJECT: HIRING EFFICIENCY ANALYSIS OF JOB POSTINGS ON LINKEDIN

LinkedIn is the world's largest professional network, bringing together millions of job seekers, recruiters, and employers across industries. As the platform continues to scale, LinkedIn wants to go beyond increasing the number of job listings. The new focus is on improving **Hiring Efficiency** - helping employers attract suitable candidates faster, with the right mix of salary, role clarity, and visibility. The team has identified **Hiring Efficiency** as a key business metric. It reflects how effectively a job post converts visibility into actual applicant interest, given its title, location, company size, and compensation.

## QUESTION

How to improve **Hiring Efficiency** of the recruiters/recruiting company based on their job posts, and help LinkedIn enhance the recruiter success rates, candidate satisfaction, and overall platform engagement.

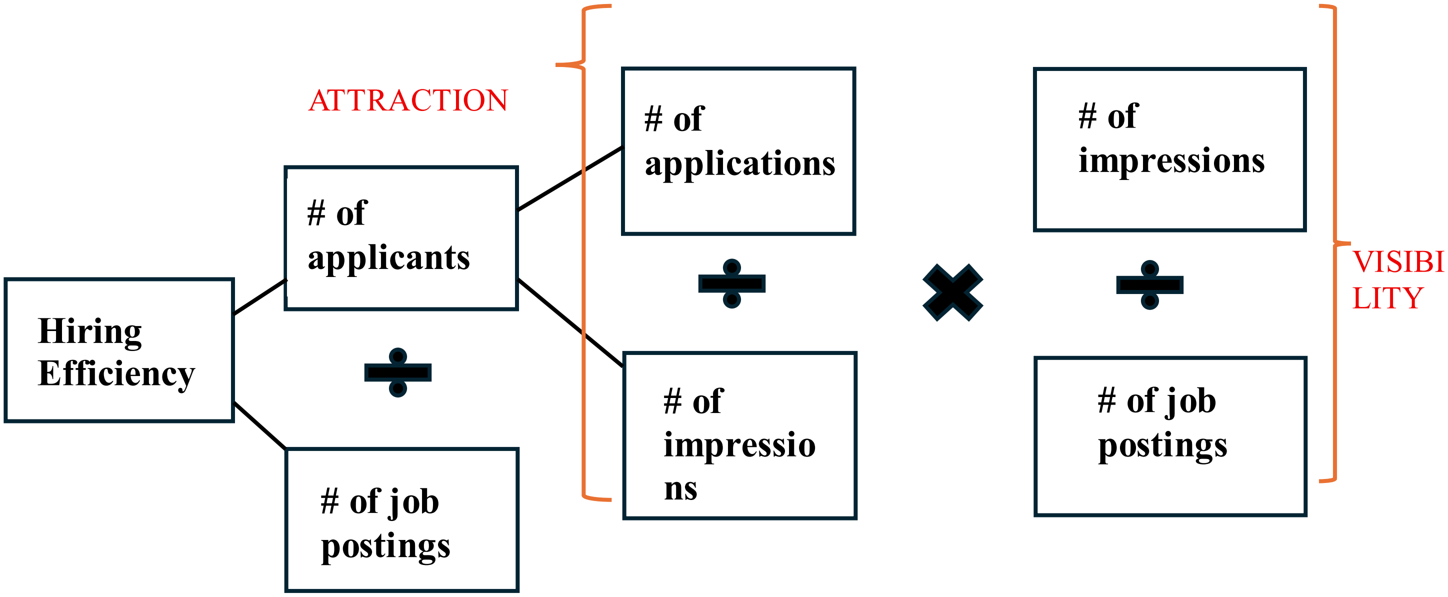
### KEY OBJECTIVES:

1. **Define and deconstruct** Hiring Efficiency as a metric across key dimensions like **title, salary, industry, company size, seniority, and location**.
2. **Identify which dimensions like roles or locations** attract the highest and lowest applicant efficiency.
3. **Analyze relationships** between the dimensions like salary, seniority level, and number of applicants to find optimal ranges for faster hiring.
4. **Evaluate how organization type, benefits, or employment type** impact applicant volume and efficiency.
5. **Present actionable insights** that can help recruiters design better job postings

## HOW TO SOLVE?

### Step 1: Defining the metrics

disintegrating the key metric **Hiring Efficiency** into several fragments like applications, impressions,.etc, which helps in deriving the answer.



**Hiring efficiency** = No. of applicants / No. of job postings.

Further disintegration gives **2 actionable metrics**,

1. **ATTRACTION:**
   Attraction = No of applications (applies)/ no of impressions (views)

   This metric explains how effectively the job posts convert **viewers into applicants**

   Most of the Viewers, see the job post but converting them into applicants is quiet difficult. Therefore, the solution largely relies on **how attractive the job post is**, specifically, which factors that might influence the viewers to apply for the job**.**

   **Factors that might affect the attraction:**

1. Salary disclosure (whether the job posts contains information about salary)
2. Salary package ( eg. $40k-80k)
3. Pay period (eg. Hourly, yearly…)
4. Job roles or Titles (eg. Data analyst)
5. Remote allowance (does the post contains the information regarding remote allowance)

2. **VISIBILITY:**

Visibility = no of impressions (views)/ no. of job postings

This metrics explains **how many people does the the post reaches.**

Finding out whether the job post is reaching the candidates is as vital as attraction of the post. There are factors that might influence the candidates to take a look at the job post

Factors that might affect the visibility of the post:

1. Job roles or Titles
2. Location (job location details)
3. Time (time when the recruiters post about the job)
4. Work type (eg. Contract, full time, part time)
5. Company (eg. Amazon, Microsoft)

3. **Other useful metrics**: Time influence

Few other factors like Time duration of the post might affect the applications and visibility of the post.
Eg. How long the post has been active?

**Step 2: Formulating the Hypothesis**

Formulate practical questions that can be analyzed and derive useful insights. The hypothesis should be based on the following metrics

ATTRACTION (how attractive is the post?)
For eg:

1. does salary disclosure percent impact the job application?
   useful variables: salary_disclosure (not given in the dataset) & applies
2. does low salary decrease the job applications?
   Useful variables: salary_package (not given in the dataset) & applies

VISIBILITY (how visible the posts are?)
For eg:

1. Is there specific job roles (among the popular job domains like IT, healthcare etc.) gets more views? Useful variables: title & views
2. certain locations where job roles in demand or does the viewership depends on the location?
   Useful variables: location & views

OTHER USEFUL METRICS: TIME INFLUENCE:

1. does post duration impact in impression of the post?
   Useful variables: post_duration_days (not given in the dataset) & views
2. does post duration impact applications?
   Useful variables: post_duration_days (not given in the dataset) & applies

**Step 3: Loading the data in Jupiter notebook:**

The LinkedIn dataset contains over 120,000 job postings, which was stored in Google Drive and opened in Jupyter notebook for data analysis. After loading the file, the Job posting data was then saved into a variable called **"data"**, which was then used for further analysis.

**Step 4: Clean the data:**

The data may contain lot of missing values, outliers and/or duplicates, which might affect the analysis. Therefore, those issues needs to be addressed first before jumping into the analysis

The process of data cleaning:

1. viewing the data and getting some basic information about the data

2. check whether the data has duplicates
3. check for missing values
4. deal with the missing values
5. check for outliers (unusual/abnormal values in the data)
6. visualize to see the presence of outliers.
7. deal with the outliers cleaning the data:

1. The data view:
   It gave the information about data like the data types (numeric or categorical or binomial), and basic description of the data for example average, SD (spread of the data), etc.

2. Removing the duplicates:
   The duplicate data were identified and removed.

3. Check for missing values:

   The missing values in data were identified by finding out the percentage of the missing values in each columns.
   The result showed that, huge amount of data seem to be missing for example, 75% of the data were missing in max_salary, 99% of the data were absent in closing_time etc.

4. Addressing missing values:

   The missing values in the data were analysed and handled carefully based on their percentage of the missing data by using practical approaches like imputing or dropping it and made sure meaningful information was retained.

   - company_name had <5% of the data missing, so the missing data were filled "unknown".
   - Pay_period, compensation_type, formatted_experience_level application_url had more than >20% data missing but it's still filled with "unknown" to preserve the data.
   - The salary fields like, normalized_salary, max_salary and min_salary and applies had >50% data missing but the missing data seems to be

non random so the data wasn’t dropped, instead kept it as NaN since it’s numerical value.

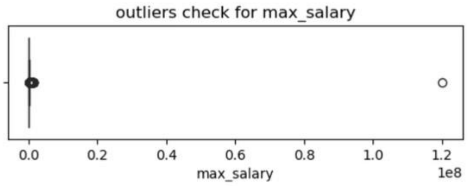
- The views data had <5% of the missing data, so the missing data were imputed with median.
- Fields like link closed_time and skills_desc had more than 90 % of missing data so dropping them seemed to be an ideal option.

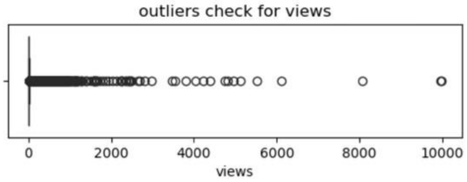
5. Checking for outliers or extreme values in numerical fields:

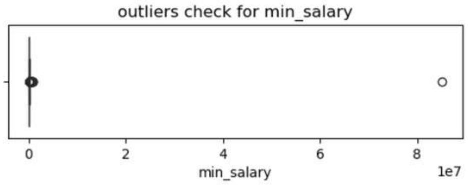
The outliers are unusual or extreme values that are abnormaly higher or lower than the rest of the data. The outliers were identified using IQR (inter quartile range) method.
The IQR method is applied only for the main numeric columns (normalized_salary, max_salary, min_salary, views and applies). Using the IQR method, the lower and upper (higher) extreme values were identified.
Result:

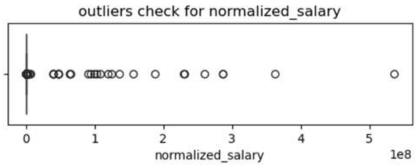
```
max_salary : 2739 outliers views
: 22830 outliers
min_salary : 2035 outliers
normalized_salary : 1664 outliers
applies_clean : 4086 outliers
(the applies_clean doesn’t have NaNs)
```

6. Outlier visualisation:
The detected outliers were visualized using boxplots.









Here the dots away from the “ | ” are outliers

7. Adressing the outliers:
   The detected outliers were eliminated carefully.
   Note: the detected outliers were not eliminated in Views and applies because it is common to have extreme values in those fields. For eg, big companies will get more views and applies but ot the case for small companies.

**Step 5: Data preprocessing:**

After cleaning the data, several additional transformations were applied to prepare dataset for analysis. These transformations helped to standardised the datasets.

1. flaging remote allowance in a new variable called remote status.

   The remote_allowed contains 0 and 1 and lot of missing values. To make it easier to analyse, A new variable **remote_status** was created with: Missing values were labelled as “unknown”, 1 as allowed and 0 as not allowed.

2. time format is unreadable so changing it to readable formata and formingnew labels like months and seasons.

   Several time-related fields like original_listed_time, expiry, listed_time (stored in milliseconds) were converted to standard date-time format. From these timestamps, Month was extracted and, **Season** was created using a custom function based on U.S. seasons (Winter, Spring, Summer, Fall).
3. Disintegerating location into states_and_countries and cities, making it easier to analyse

   Job location originally came as long strings (e.g., New York City Metropolitan Area). A cleaned version of the location was created by removing terms like Metropolitan Area, Bay, Area, region, urban..etc. The cleaned version is then split into 2 fields, states_and_countries and cities. Miss spellings and unwanted spaces ere then removed from those variables. Inconsistent values like Michigan were then converted to standard two letter code like MI

4. Creating salary range:

A new column salary_range was created If both **max_salary** and **min_salary** were available, salary_range = max_salary – min_salary. If values were missing, the range was set to **NaN**. This helped analyze how transparent companies were with salary information.

5. Grouping salaries to form salary_package:

   The normalized salary was grouped into meaningful salary bands like <40k, 40k-80k,..etc. This allowed cleaner visualization of salary distribution.

6. Grouping Views to form views_gps:

   The views column ranged widely, so it was grouped into bins like <50, "51100",..etc. This allowed cleaner visalization of views.

7. Grouping Title to form title_gps:

   The title column was widely ranged, so it was grouped into bins like if the title column contained "analysis" or "analyst" it is grouped as "analysts/analytics". This version of title gps helped to do analysis on popular job titles.

8. Salary disclosure:

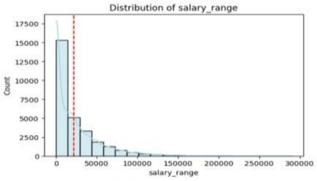
   The column is created to see whether the job posts mentioned salary details or not like max_salary, min_salary and normalized_salary (annual salary) or not (a binary variable, 0 or 1).

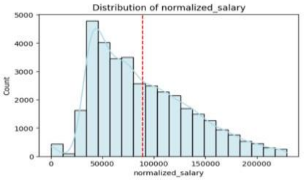
Step 6 Exploratory analysis

Numeric variable distribution:

Histograms were plotted for **salary_range and normalized_salary**.Both variables are **strongly right skewed**, meaning most job postings offer lower salary ranges, with only a few very high values.

The **views and applies** columns also show heavy skew, with most jobs receiving low engagement. Because of this, analyzing them as raw numbers is not meaningful, and grouping them into categories is more appropriate.





Categoric variable distribution:

1. **title_gps**

The dataset shows a high concentration of job postings in IT & Engineering, Sales, Healthcare, and Administrative roles. A large number of postings fell under the "Other" group, indicating job types that do not fit into the popular domains.

2. **States and Countries**

Most postings were from major U.S. states such as CA, TX, NY, FL, NC, and IL. A small number of international postings also appeared. The distribution says hiring concentration in major employment hubs like california.

3. **Remote Status**

The majority of postings fall under "Unknown" due to missing data. Among known values, "Available" remote jobs represent a smaller portion.

4. **Salary Package**

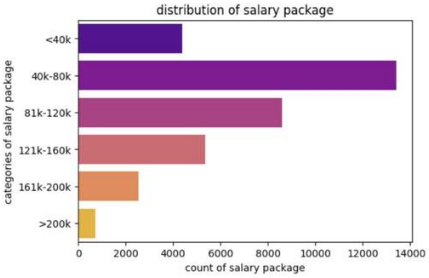
Most jobs fall into the 40K–80K, 81K–120K , 12k-160k and <40k categories.Very high salary categories like >200K form only a small portion of postings.

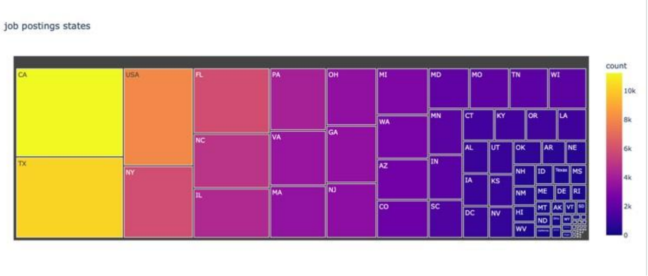
5. **Work Type / Experience Level**

Full time roles appear most frequently.Experience levels show a balanced distribution, with many jobs classified as Mid level, Entry level, or Unknown.

6. **Views gps:**

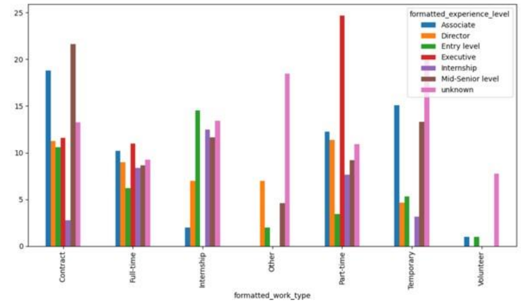
Most of the job posts had views less than 50 so the distribution chart isn't much reliable

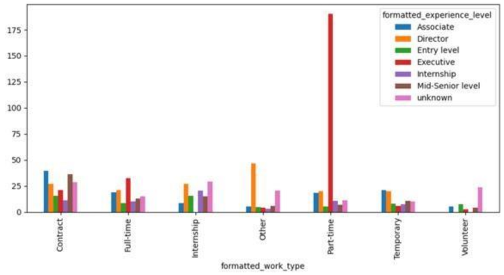




Multivariate analysis:

To see how job visibility, attraction changes across different job types, experience level and months. The results showed that Executive level part time roles received high no. of applies, Full time jobs received the highest views, especially in February.





## Step 7: Hypothesis testing:

The formulated hypothesises were tested using different statistical methods like correlation analysis and ANOVA, in jupyter notebook

ATTRACTION:

1. **does salary disclosure percent impacting the job application:**

   using functions in python, a new column was created named salary disclosure. This variable is grouped with the applies (job application) to see whether there is a correlation between 2 ( eg if the disclosure increases doeas the applies also increases) here spearmanr correlation was used since the salary disclosure was a binary (0 or 1). The results shown that it seems like there is a weak positive correlation between salary_disclosure and applies.

2. **does low salary decreases the job applications?**

   Here the salary package is grouped with the applies to see whether there is any correlation between 2. Here ANOVA was used to see whether there is a correlation, the results both are related to each othere but the question wasn't answered properly so to see high or low salary affects the applications, for that spearmanr was used it seem there is no positive correlation (it means it doesn't have anything to do with high or low salary.

3. **does pay period affect job applications?**

   In the pay period, some values like weekly, bi weekly has very low value so monthly, biweekly, weekly were labled as other, easier for analysis. Then the cleaned version of pay_period was then grouped with applies and did spearmanr correlation. The results showed that pay_period might have correlation between pay_period and applies and statisticaly significant. The barchart showed that hourly pay jobs gets more applications.

4. **is there any job roles in demand?**

   It seemed like removing other gives better insights on some of the popular and well know jobs, so the new variable called title_gp2 (without "other") was then grouped with applies to see are there any correlation. Here ANOVA was used because the title group was a nominal data and more than 2 groups. The result showed that tat the popular jobs (like sales, IT, analyst, etc..) has major impact in application which means that, it is highly correlated with applies and in barchart among the popular domains shown that, It and engineering, analysts/analytics domains received more applies than others.

5. **does remote allowance affect applications?**

   remote allowance was grouped with applies and did a spearmanr correlation test (because remote allowance was a binomial). The result shown that remote allowance has weak positive correlation towards job applications(applies) meaning, mentioning remote allowance in job posts has no impact in applies

VISIBILITY

1. **Is there specific job roles(among the popular job domains) gets more views** title_gp2 (without “other”) was then grouped with views to see are there any correlation. Here ANOVA was used because the title group was a nominal data and more than 2 groups. The result showed that tat the popular jobs (like sales, IT, analyst, etc..) has major impact in application which means that, it is highly correlated with the impression, in barchart among the popular domains shown that, marketing domain received more views than others.

2. **certain locations where job roles in demand.**

   states_or_countries was grouped with views to see which states gets more impressions. Here ANOVA was used because states_or_countries are categorical and not ordinal data (not ranking). The result showed, it is highly significant tat the there is high corelation between states and views. The treemap was formed to see which state recieves more views, it seems CA, TX NY received more impressions.

3. **certain months where applicants actively look for job.**

   Month was grouped with views to see which month gets more impressions. Here ANOVA was used because months are categorical and not ordinal data (not ranking). The result showed, it is highly significant tat the there is high corelation between month and views and the barchart explains that tha candidates actively look for job in february month

4. **certain seasons where applicants actively look for job**

   “Seasons” was grouped with views to see which season gets more impressions. Here ANOVA was used because seasons are categorical and not ordinal data (not ranking). The result showed, it is highly significant tat the there is high corelation between month and views and the barchart explains that tha candidates actively look for job in winter (according to U.S)

5. **does work type influence the impressions**

formatted_work_type was grouped with views to see which which work type gets more impressions. Here ANOVA was used because work type is categorical and not ordinal data. The result showed, it is highly significant tat the there is high corelation between work type and views and the barchart explains that the candidates actively look for contract based, internship based jobs.

TIME INFLUENCE:

Post_duration days was grouped with views and applies to see if there's any correlation between them. But it was significant that there is weak positive correaltion between them, meaning duration of post in Linkedin may or maynot affect the views or applications of the post.

Step 8 Conclusion and buisness level recommendations.

## CONCLUSION

This project contains analysis of linkedin dateset which contains more than 120,000 job postings to provide insights to the recruiters regarding improving the hiring efficiency by analysing VISIBILITY (views), ATTRACTION (applications), and characteristics such as salary, work type, experience level, and geography. After performing data cleaning, data preprocesing, EDA, and statistical testing, provided lots of meaningful buisness insights.

### Findings

**Salary disclosure has only a weak positive impact** on applications. Revealing salary does not significantly increase applicant behavior.

**Pay period** (hourly vs. monthly) shows some impact **hourly jobs receive more applications**, likely due to industry-specific demand.

**Job roles matters**: IT & Engineering, Analyst/Analytics, Sales, and Healthcare roles receive significantly more applications compared to others.

**Remote allowance does not strongly influence applications**, meaning remote availability alone is not responsible for of applicant behavior.

**Job titles strongly impact impressions**, Marketing, IT, and Analyst roles receive more views.

**Location matters:** States like **CA, TX, and NY** receive significantly more views than others, showing strong regional hiring activity.

**Month and Season influence visibility:** February and Winter show the highest job-seeker activity.

**Work type affects visibility:** Internship and Contract roles attract higher views, suggesting that early-career and flexible opportunities drive more browsing behavior.

**Post duration (days active) has only a weak correlation** with views and applies. A longer posting does *not* guarantee more visibility or more applications.

## Business Recommendations

Based on the insights, here are practical and actionable recommendations for hiring teams:

1. **Improve Visibility in Low-View States:** Since a handful of states dominate impressions (CA, TX, NY), employers in **low-visibility states** should, Increase **job ads in there.**

2. **post more jobs related to It and engineering, analysts/ analytics,** because these Job titles heavily influence both views and applications.

3. **Salary Disclosure alone won't influence applications:** Since salary disclosure has only has weak effect, companies may or may not need to mention the salary in the job posts

4. **Posting Strategy with job seeker timing:** Views spike during **February/Winter Season. Post more jobs during that time,** Avoid relying heavily on slow months for hiring

5. **Optimize job posts based on work type** Internship and Contract roles receive high visibility. companies can promote these roles more aggressively**,** Advertise multiple opportunities related to that

6. **Monitor Low Performer Job Categories:** Some job categories receive significantly fewer applications and views. Companies should, Reassess job description clarity, Improve job title specificity