**AIT 580 – Data Analytics Research Project**

**Name: Pritham Mahajan  
 G-NUMBER: G01462517**

**Dataset: - NYC JOB POSTING**

**ABSTRACT**

In the age of technological advancement, the authorities are making use of digital platforms with great frequency to connect with the diverse pool of talent to facilitate the hiring process. This research finds the amount of data collected from the jobs which are posted on the official website for New York City. On analyzing the dataset carefully, it is seen that a combination of external and internal job postings opportunities is available for urban and regional employees, which provides a complete overview of the labor conditions within the city. The primary objective of this study is to identify the various changing patterns and gain valuable insights of the jobs that are posted on the online portals and gain complete knowledge on how the job market is varying in New York. The research uses advanced analytics and machine learning techniques to filter large amounts of data set and make use of helpful information that can further studies and used by various applicants, recruiters, and policymakers [2]. To begin with the analysis the job postings are grouped under various categories like department, job title, qualifications, work location and skills. The basic step is to look out for the chances that are distributed within the various sectors which direct to the potential demand positions available in the city. The aim is to look for the present requirements and focus on regional employment by analyzing the quantity and importance of a specific job role[4]. To conclude, this project gives a detailed extensive study of various job postings on the job websites for New York City, which gives insights about the employment opportunities in NYC. The study aim is to find talent and hiring strategies and policymaking. This provides the use of data to enhance urban areas hiring procedures in a changing city.[1]

**INTRODUCTION**

A major change in ways to recruit talent has recently been made ideal by the alignment of advances in technology and the role of municipalities in today's state of employee’s changes. Online networks have become crucial resources for professionals searching for a variety of skills as they act as networks for communication and knowledge sharing. This study seeks to find the complicated data covered up with the job ads on the corporate website of the City of NYC, a city popular for its active and lively job market [2]. This dataset provides an in-depth analysis of the complex network of both regional and urban employment by linking the postings for city employees with alternatives which are available to its citizens. With the approach of analyzing job postings based upon department, qualifications, location, and skills, the study aims to reveal the complex correlations among various factors that impact the job market [3]. This research is not only an educational activity; it also has real-world implications for candidates, employers, and lawmakers; offering an important basis for the improvement of urban hiring practices in the face of a rapidly evolving urban environment [1].

**Research Questions**

Our study aims to clarify significant aspects by analyzing the complicated relationship of salary and employment opportunities posted by different agencies on online portals for job seekers in in New York City, this research will explore the following research questions:

* **Which are the most frequent top 20 job postings in each category and how many openings are they offering?**
* **In what way do the distributions of job posting vary among the top 10 agencies?**
* **What are the differences between each of the agencies' full- and part-time employment distributions?**
* **Is there any noticeable relationship between the number of available positions at different companies and the corresponding ranges of wages they provide?**

The purpose of these research questions is to figure out how job postings are distributed among different agencies which post jobs by looking at the connection between salary range and number of postings.

**Literature Review**

The primary goal of the research is to gain insights about jobs posted in online job website, specifically in the instances where labeled training information is very limited. This research uses a two-staged semi-supervised approach which integrates the measures related to specific sector of job description with supervised and unsupervised machine learning approaches. The analysis displays an increase of 14% in accuracy categorizing sectors rather than matching every occupation by using tools like BERT, Word2Vec, and TFIDF. In the field of information technology, the paper shows the approach's possible across countries [2].

On the other hand, looking at how well the jobs are recommended by the system based on the search is another objective of this research. This research analyzes various other similarity metrics which include Euclidean, Cosine, Jaccard, and City Block. It also examines the cold-start problem and tests the outcome of different other distance-based similarity techniques. Additionally, the cosine metric produces 90-94% precise recommendations by prioritizing inefficiencies in low-complex vectors. With respect to the NYC job postings, the paper illustrates the advantages of customized job recommendations and highlights various other factors that affec6 the advertisement for jobs in NYC [3].

Finally, the objective of the suggested distributed timing job scheduling system is to enhance big data analysis in real time in a data analysis medium. The system is composed of components for managing jobs, scheduling task, managing task, managing logs and inspecting worker node. It operates on a centerless master-worker structure.  This system helps in achieving flexibility in carrying out jobs and collecting log in the real time for accurate analysis, which would further improve efficiency of workers and reduce the processing time for employment. The research shows how helpful these developments   are helpful to enhance performance in general and boost data analysis platforms [4].

**Materials and Methods**

Started this research by selecting dataset from the City of New York's official job posting site. This dataset includes both external postings that are open to everyone and internal postings that are only for city employees. Recognizing the fundamental value of this dataset, the study aims to extract insights from the job market in one of the world's fastest growing cities. After selecting and confirming the dataset, the first thing to do was carefully clean the data. This is an important step because it ensures the reliability as well as accuracy of the analyses. In the Excel environment, focused on organizing and identifying the messy and misplaced data in the "Level" column. Levels are initially displayed as 0, 1, 2, 3, 4, and 5. Each cell is reviewed and irregularities like A1, B1, and M1 are fixed to clean the levels. Meanwhile, the data in Full Time/Part Time column is organized by excluding various other expressions and filled blank spaces with "N/A" to ensure accuracy. A special challenge raised in the Hours/Shift column, where it required careful attempts to convert non-uniform timings and into precise working hours in order create clean dataset for visualization. Adding more, in Job Category column all the data was mixed, and each type of category was included in all the rows, manually cleaned those rows in order to make the dataset look organized.

Many other significant tasks are done in Python, carrying on with the data cleaning process. The accuracy of data is improved by removing "duplicate entries". Moreover, the columns like "business title, unnamed 20 are dropped to simplify and make the data more manageable and lot easier to analyze. The Civil Service column which consisted of extra open brackets which might result in undesired results were removed. To make the data more structured the Work Location 1 column has been refined by only using street\_address and dropping state/Zip\_Code.

Upon cleaning the dataset, SQL is used in the research where the cleaned dataset is loaded into the database and few basic queries were run to demonstrate knowledge and understanding. Besides SQL, RStudio is used in the research for data visualization. Framed a few research questions which would arise by looking at the data set which have been solved using visual representation of data. In particularly the research analyzes how the range of pay varies within different agencies and nuanced question upon the job opportunities was distribution of job postings among different agencies, these were solved through visualization by plotting graphs and charts and to get the idea for the employees to filter out their preferred options.

To sum up, this research requires careful assessing of data to organize it and obtain meaningful insights of from New York City dataset. Which is done using multiple tools like Excel, Python, SQL and RStudio. The overall goal of the research is obtained through data cleaning procedures, loading into database and visualizations, which all together work to bring out complex situations in the job market in New York City to provide accurate information of the job market to both policymakers and job seekers.

**Results**

**A graph of a job posting

Description automatically generated**

The plot shows how the top 20 job postings are distributed among various job categories. The length of the bars stands for the total number of positions open in each job category. The graphical representation gives an overview of the most highly demanded roles in each job category of New York City, providing information regarding the present trends in the job market. The plot is made clearer by using cyan color bars and grid lines to easily note the number of positions, overall easier to distinguish between and determine the varying significance of various job categories. Looking into the graph carefully to gain more insights it is seen that the highest number of job openings were posted for Social Services, making it one of the important fields in the job market in New York City. Moreover, a moderate number of positions were available in categories like Data and Innovation, Public Health and Safety and Constituent Services and Community. However, Building Operations and Maintenance, Finance and accounting were seen with comparatively lower demand. The job patterns among the top 20 positions in each category can be more clearly understood using this visualization, allowing job seekers, researchers, and readers to gain a deeper awareness of the distribution trends and specific areas of the labor market.

**A graph of various colored bars

Description automatically generated**

This plot depicts information on the number of job postings of top 10 agencies in New York City. The length of the bar indicates the total number of jobs posted on the online website by a particular agency. The agencies are listed on x-axis and Number of postings are listed on y-axis. To better understand the data in the graph the bars are colored in different colors and a legend is placed indicating the top 10 agencies who regularly post about jobs available in New York City.  To enhance the understanding of the plot is made using ggplot2 library and the horizontal bar chat is used. The x-axis coordinates are removed to make the graph look cleaner as the same coordinates were used in legend, by this the plot gets interactive and the viewers will engage with the plot and spend more time in exploring data. The aim of the plot is to present a clear picture of the title " top 10 agencies based on highest number of job postings" in a simple way.

Basically, when looked at the graph, the highest number of jobs posted by dept. social services, approximately 1150 postings in New York City. Followed by taxi and limousine commission agency with least amount with less than 300 job postings. With the help of graphical representation, it is more vibrant to understand the job postings of the agencies posted in NYC, which gives the viewers a complete brief about the agencies which post hiring opportunities in high volume. Further, with the help of these insights job applicants can filter out and apply for prominent roles and types of jobs according to their needs, as the employment opportunities are in bulk it might be easy for the applicants to get a success on their desired role.

**A graph with numbers and a number of positions

Description automatically generated**

The generated plot gives an easy understanding of the employment opportunities released by different agencies mainly focusing on the differences among full-time jobs and part-time jobs. Every bar in the plot denotes separate agency and the length of the bar shows the number of jobs posted by a particular agency. Adding more, the number of job types like part-time or full-time jobs which are posted by the agencies have been highlighted by carefully organizing the bars. To make the plot viewer friendly, the bars have been given different colors where the blue bar represents part-time jobs, and the red bars show the number of full-time jobs posted by a particular agency in New York City. To better understand the insights of the job postings in the city, along with the bar plots, making it easy for the viewers quickly understand what is being displayed. The title "Part-time / Full-time distribution by Agency" clearly conveys the purpose and job seekers can easily filter out the type of job they are seeking for.

In addition, a color-coded legend is used to make the understanding level of viewers feasible, with the legend viewers can easily identify the distribution of full-time and part-time positions which are posted by the agency in NYC. This graph gives a useful review of the different types of jobs available as per to the job seekers’ needs. The visualization helps to identify the differences in job descriptions, giving both job seekers and researchers a good understanding of job opportunities posted by the specific agencies. It's an effective tool for discovering trends and variations in hiring processes, which adds to an extensive comprehension of the complexities at work in the job market in New York City.

A graph with black dots and numbers

Description automatically generated

The relationship between open positions and ranges of salaries across different agencies are shown in a scattered plot. plotting each point indicates a particular job posting and offers a graphical illustration of the overall distribution of positions and related salary ranges. This visual research clears the way for a more nuanced comprehension of workforce patterns by providing insights into whether hiring agencies with a greater number of positions usually provide ranging salary ranges.

The research, that is conducted using reliable R data import and visualization methods, has significance for organizations, decision-makers, and job searchers. Understanding the intricate connection between salary ranges and job openings is crucial for making important choices, promoting workforce planning, hiring methods, and promoting transparency in the market. The methodology utilized ensures an in-depth review, and the findings support evidence-based strategies for establishing the constantly evolving work environment.

**Limitations and Further Research**

As there was not much time to deeply analyze the complete structure of the New York job market the study had to be limited to a certain limit. However, there are questions regarding the ability to apply the study's findings regarding the New York City job market due to limitations related to the accuracy of the data and transparency. The analysis may be biased by the dataset's possible excess of industries or job types, making results that might not accurately reflect the overall job market. That limitation might make it harder to apply the results of the study to the wide range of businesses and jobs found in New York City. Furthermore, opinions in the data could unintentionally maintain present disparities in job possibilities, which would make it more difficult to create beneficial initiatives and approaches.

The accuracy and completeness of the data and the constantly changing character of job markets are the primary concerns. An insufficient awareness of the complexities present in the labor market can result from missing or incorrect data in the dataset. The dynamic nature of job trends, which are affected by outside events and economic conditions, makes it difficult for the study to recognize unexpected shifts in the demand for positions or skills. As a result, the study might offer an overview that's does not complex enough to adjust hiring procedures or statutes in response to the constantly changing dynamics of the labor market. It is critical to tackle these barriers to guarantee the strength and significance of the research findings and to promote more informed decision-making concerning employment regulations and processes.

**Discussion and Conclusion**

In conclusion, by using data from the official job posting website of the City of New York, this research project provided valuable insight into the job market in New York City. A 14% increase in job sector precision for classification has been shown with the two-staged semi-supervised approach that uses machine learning techniques and includes measures associated with job sectors. The analysis also examined the efficacy of the recommendation system, analyzing the cold-start issue and evaluating different similarity metrics. A precise recommendation tool that works especially well in low-complexity vector scenarios is the cosine metric.

In addition, the study presented a networked timing job planning system aimed to improve big data analysis in real time. It has been noticed that this system, which uses a centerless master-worker structure, has a capacity to reduce processing times and boost efficiency in the analysis of employment data. Despite these improvements, it is essential to acknowledge the research's limitations, which include the deficiency of contextual and outcome data as well as incomplete job descriptions. Later studies may tackle these constraints and explore additional topics such as predicting skill requirements, tracking the dynamics of employment markets in real time, and improving the user interface on job boards. These advances would help create a more thorough understanding of the labor market, which would be beneficial to employers and job seekers equally.

**References**

**[1]** *Jobs NYC Postings - Catalog*. (2023b, November 3).

<https://catalog.data.gov/dataset/nyc-jobs>

**[2]** *Two stage job title identification system for online job advertisements*. (2023). IEEE Journals & Magazine | IEEE Xplore. <https://ieeexplore.ieee.org/document/10050019>

**[3]** *Combined application of various techniques for personalized job recommendation*. (2023, January 5). IEEE Conference Publication | IEEE Xplore. [**https://ieeexplore.ieee.org/document/10083944**](https://ieeexplore.ieee.org/document/10083944)

**[4]** *Design of distributed timing job scheduling system for data analysis platform*. (2023, April 1). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/10261027>