Final Report

Analytics Capstone

Predictability of Career and Professional Excellence

By Team 4

Saikeshav Motwani

Ayush Patel

Viraj Bhavsar

Mitesh Singh

Sherin Jacob

Vedant Pandya

Table of Content

Abstract

This project delves into the rich dataset provided by LinkedIn, comprising over 10,713 job postings and supplemental data on industries, companies, and skills. Through rigorous data analysis techniques including cleaning, exploratory data analysis, clustering, predictive modeling, trend analysis, and visualization, valuable insights are extracted to guide individuals, organizations, policymakers, and educational institutions in navigating the dynamic landscape of the modern workforce.

Key findings reveal geographical distribution of job postings, industry attractiveness, demand by work type and experience level, and salary disparities across various sectors. The project's significance lies in empowering stakeholders with actionable insights for career development, talent management, and workforce preparation, contributing to economic growth and societal resilience.

Tools Used in this Project:

Data Preparation & Cleaning: Microsoft Excel, Microsoft Access

Visualizations: Tableau

Modelling: IBM SPSS Modeler

Introduction

LinkedIn has redefined the landscape of talent acquisition and career development in the digital age by becoming the leading platform for professional networking and job recruitment. LinkedIn provides an endless supply of information with its enormous database of job posts and business profiles, providing unmatched insights into the dynamic nature of the labour market. Careers and professional development are changing all the time in an era of rapid technology improvement, shifting economic conditions, and changing social mores. As we stand on the threshold of a new decade, the ability to predict and adapt to these changes becomes increasingly vital for individuals and organizations alike.

Overview of the Dataset

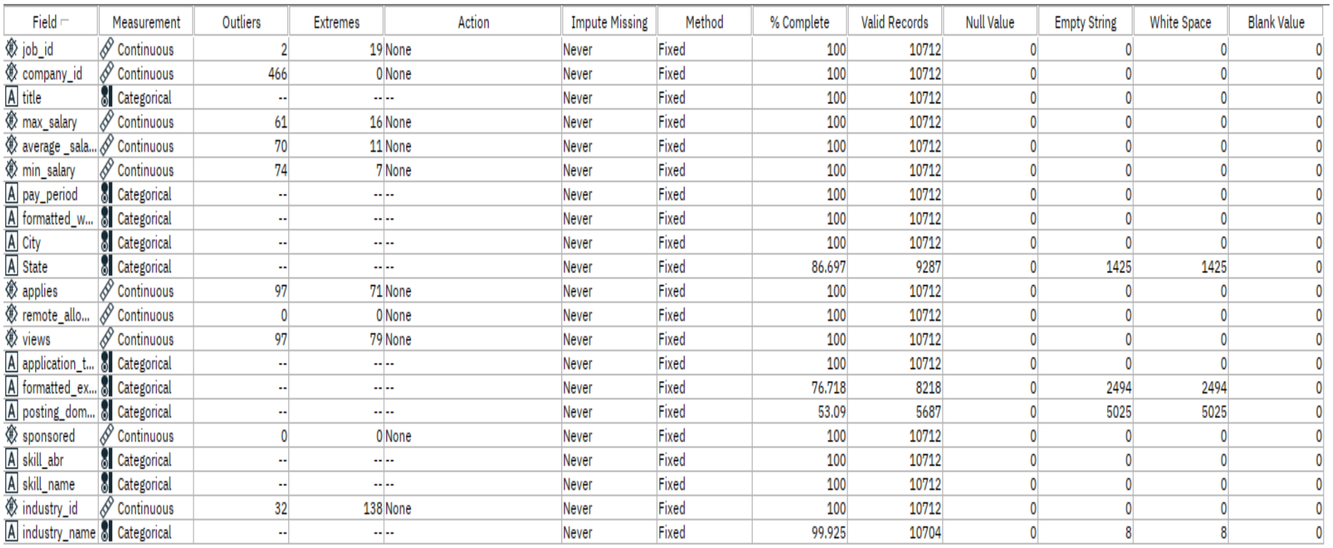
The dataset under consideration offers a near-comprehensive snapshot of over 33,000 job postings on LinkedIn, meticulously recorded over two distinct days, separated by months. Each posting is a treasure trove of information, enriched with 27 distinct attributes ranging from the job title and description to the specifics of salary and the nature of employment, be it remote, contractual, or otherwise. We also have some other datasets that are related to each other like industries, skills,

This data analytics project aims to delve deep into LinkedIn's dataset comprising over 33,000 job postings, each annotated with 27 detailed attributes, alongside separate datasets detailing company profiles, job benefits, required skills, and industry classifications. The objective is to harness this wealth of information to glean insights into the future of careers, identifying trends and patterns that could shape professional trajectories in the years to come. We decided to select some related 21 fields only. Here is the overview of the selected fields.

Selected Data from the Dataset:

Field	Type	Description
job_id	int	Related Job ID for each job posted
Company_id	int	Related Company Id for each job posted
title	String	Job Title
Max_salary	float	Maximum Salary
Average_salary	float	Average Salary
Min_salary	float	Minimum Salary
pay_period	String	Payment Period
formatted_work_type	string	Work Type
City	string	City of job
State	string	State of job
applies	int	No of application per job
remote_allowed	flag	Remote job allowed or no
views	int	Views per job
application_type	string	Application type
formatted_experience_level	string	Experience level
posting_domain	string	Job posted domain
sponsored	flag	Sponsored or no
skill_abr	string	Skills Attribute
skill_name	String	Skill name
industry_id	int	Industry Id
industry_name	string	Name

The project's analytical potential is further enhanced by additional datasets that include industry classifications, solid profiles, and necessary skills in addition to the job postings dataset. These supplementary datasets provide important context and complexity to the analysis, allowing scholars to investigate relationships, pinpoint important factors influencing employment demand, and detect hidden patterns within sectors or skill sets. The project intends to provide practical insights into the future of careers by utilising this wide range of data sources, enabling stakeholders to confidently navigate the changing labour market landscape and make well-informed decisions.



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
job_id	Continuous	2	19	None	Never	Fixed	100	10712	0	0	0	0
company_id	Continuous	466	0	None	Never	Fixed	100	10712	0	0	0	0
title	Categorical	--	--	--	Never	Fixed	100	10712	0	0	0	0
max_salary	Continuous	61	16	None	Never	Fixed	100	10712	0	0	0	0
average _sala...	Continuous	70	11	None	Never	Fixed	100	10712	0	0	0	0
min_salary	Continuous	74	7	None	Never	Fixed	100	10712	0	0	0	0
pay_period	Categorical	--	--	--	Never	Fixed	100	10712	0	0	0	0
formatted_w...	Categorical	--	--	--	Never	Fixed	100	10712	0	0	0	0
City	Categorical	--	--	--	Never	Fixed	100	10712	0	0	0	0
State	Categorical	--	--	--	Never	Fixed	86.697	9287	0	1425	1425	0
applies	Continuous	97	71	None	Never	Fixed	100	10712	0	0	0	0
remote_allo...	Continuous	0	0	None	Never	Fixed	100	10712	0	0	0	0
views	Continuous	97	79	None	Never	Fixed	100	10712	0	0	0	0
application_t...	Categorical	--	--	--	Never	Fixed	100	10712	0	0	0	0
formatted_ex...	Categorical	--	--	--	Never	Fixed	76.718	8218	0	2494	2494	0
posting_dom...	Categorical	--	--	--	Never	Fixed	53.09	5687	0	5025	5025	0
sponsored	Continuous	0	0	None	Never	Fixed	100	10712	0	0	0	0
skill_abr	Categorical	--	--	--	Never	Fixed	100	10712	0	0	0	0
skill_name	Categorical	--	--	--	Never	Fixed	100	10712	0	0	0	0
industry_id	Continuous	32	138	None	Never	Fixed	100	10712	0	0	0	0
industry_name	Categorical	--	--	--	Never	Fixed	99.925	10704	0	8	8	0

Fig.1-Data Audit on the Final Dataset used

Project Significance

For Individuals: Offers guidance on emerging career paths, skills in demand, and strategic locations for career development, enabling informed decision-making regarding education and professional growth.

For Organizations: Provides a benchmark for competitive compensation, identifies trending benefits to attract top talent, and highlights skill gaps in the market, informing talent acquisition and development strategies.

For Policy Makers and Educational Institutions: Sheds light on industry trends and the evolving demand for job roles, aiding in the formulation of educational curricula and policy decisions to prepare the workforce of tomorrow.

The project holds significant importance across various sectors due to its potential to revolutionize career development, talent management, and workforce preparation. Its significance can be understood in the following ways:

- **Empowering Individuals:** People often find it difficult to successfully navigate professional paths in the fast-paced employment market of today. The project gives people the power to make well-informed decisions about their education and professional development by providing advice on in-demand skills, new job routes, and key places for career development. Individual fulfilment and professional achievement are increased by this empowerment.

- **Increasing Organisational Competitiveness:** Success for organisations depends on attracting and maintaining great employees. Organisations can benefit from the project's insights regarding market skill gaps, trending benefits that attract top talent and competitive compensation benchmarks. Equipped with this data, establishments may maximise their approaches to attracting and nurturing talent, so increasing their capacity for innovation and competitiveness.
- **Informing Education and Policy:** To prepare the workforce of the future, policymakers and educational institutions are necessary. The research contributes to the development of educational curricula and policy decisions by providing insight into industry trends, changing employment positions, and skill requirements. By doing this, educational programmes are guaranteed to be in line with business needs, producing graduates who are more equipped and a more flexible workforce.
- **Driving Economic Growth:** In the end, the project's importance lies in its capacity to promote wealth and economic growth. The initiative enhances worker efficiency and productivity by streamlining talent management tactics, improving the match between candidates and job possibilities, and coordinating education with industry demands. Consequently, this promotes economic growth, innovation, and competitiveness at the local and national levels.

Methodology

Data Cleaning and Preprocessing: Standardize, clean, and prepare the dataset for analysis, including handling missing values and categorizing textual data.

For the Data cleaning, we have removed the records here the salaries were empty because every job a person applies would like to know the salaries. That is why we decided to remove the data which did not have these records. Thus, we have a dataset of 10713 job postings, moreover we added a calculated field named Average salary of the job listings using the minimum salary and maximum salary fields.

Next, using Microsoft Access we have created relationships between different datasets to form a final file of data. Using the Query function of Microsoft Access, we have created a Final Dataset containing all important and relevant fields.

Dataset Relationship:

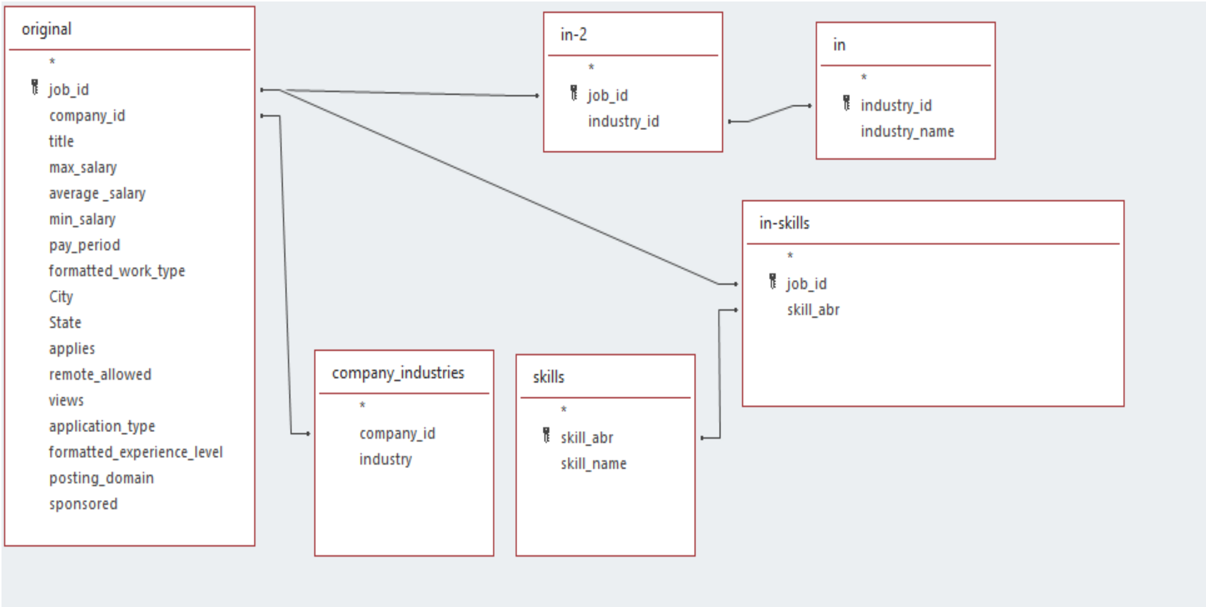


Fig. 2-Relationship between the Datasets

We have created the relationship with all these different datasets and form the final data using job_id, company_id, skill_abr, industry_id as foreign key in each dataset.

This is the Query Used to create the final Data:

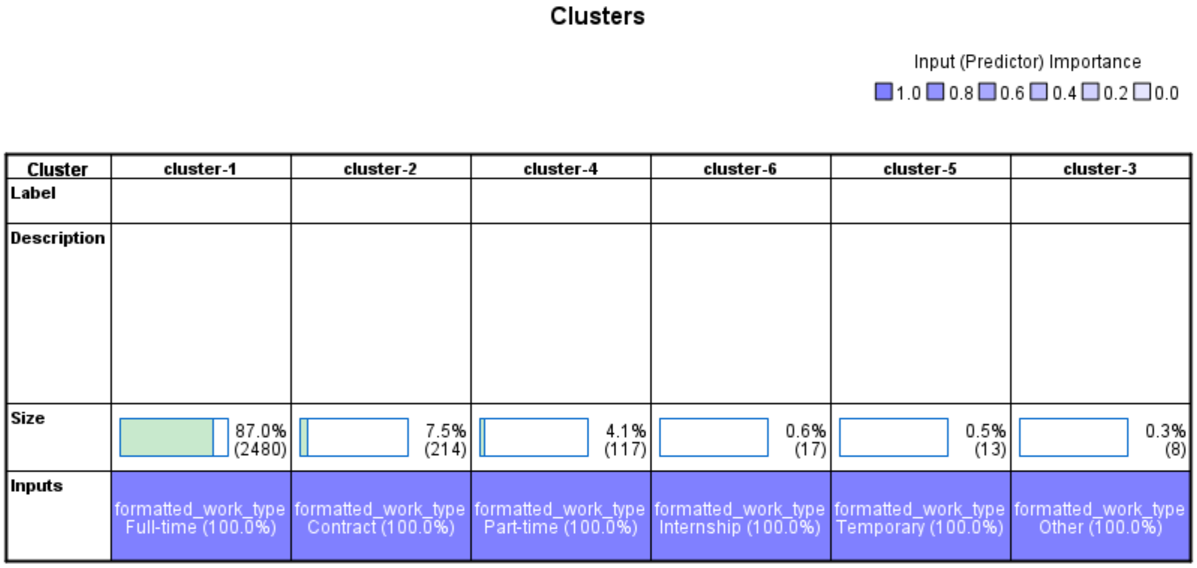
SELECT original.*, [in-skills].skill_abr, skills.skill_name, [in-2].industry_id, [in].industry_name

FROM ((original INNER JOIN ([in] INNER JOIN [in-2] ON [in].industry_id = [in-2].industry_id) ON original.job_id = [in-2].job_id) INNER JOIN company_industries ON original.company_id = company_industries.company_id) INNER JOIN ([in-skills] INNER JOIN skills ON [in-skills].skill_abr = skills.skill_abr) ON original.job_id = [in-skills].job_id;

Exploratory Data Analysis (EDA): Conduct an initial exploration to understand the distribution of key variables such as salaries, benefits, and job locations.

Clustering:

- Derived type node from the data node (job postings) and set formatted_work_type and application type as an input field.
- Then derive K-means unsupervised modeling node from the type node.
- The largest cluster was 2480 (87%) for work type full-time and smallest was 8 (0.3%) for work type other.



Clusters

Input (Predictor) Importance

■1.0 ■0.8 ■0.6 ■0.4 □0.2 □0.0

Cluster	cluster-1	cluster-2	cluster-4	cluster-6	cluster-5	cluster-3
Label						
Description						
Size	87.0% (2480)	7.5% (214)	4.1% (117)	0.6% (17)	0.5% (13)	0.3% (8)
Inputs	formatted_work_type Full-time (100.0%)	formatted_work_type Contract (100.0%)	formatted_work_type Part-time (100.0%)	formatted_work_type Internship (100.0%)	formatted_work_type Temporary (100.0%)	formatted_work_type Other (100.0%)

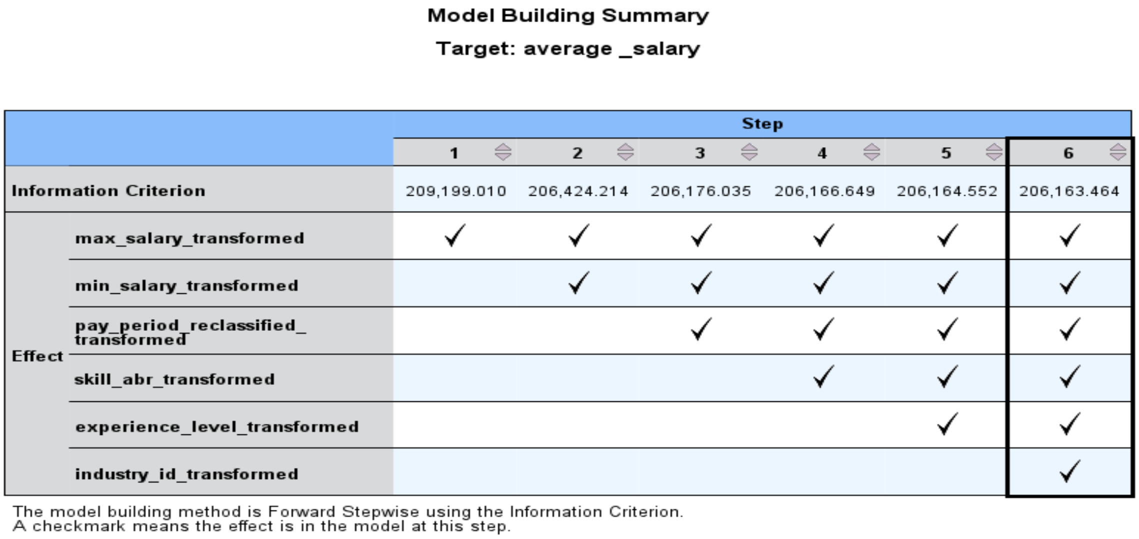
Fig.3-K-means Clustering

Predictive Modeling:

Employ machine learning algorithms to predict salaries and identify key factors influencing job benefits and skill requirements.

- Derive a type node from the data node (job postings) and set max_salary and min_salary as input fields and average salary as target field.
- Then derive a supervised regression model from the type node. It was a failure as there is no error in any versions.
- So, we separated the state and country from the location column to better analyze the data and created new data file (with location.xlsx).
- Then derived type node to add state as input field to get outputs based on location.
- Then derive a reclassify node for experience level and created a linear model 1 node for predicting average salary.
- Then derive reclassify node for pay period and then derive linear model 2 node and subsequently for work type create linear model 3.

- The third model is the Linear modelling model for predicting the average salary and we have three versions of it to predict the salary and the version 2.0 is better than other.
- Model 4 was CHAID model derived from work type reclassify node.
- Model 4 is for the prediction of the job to be allowed to work in remote based on experience level and work type.



Model Building Summary

Target: average _salary

		Step					
		1	2	3	4	5	6
Information Criterion		209,199.010	206,424.214	206,176.035	206,166.649	206,164.552	206,163.464
Effect	max_salary_transformed	✓	✓	✓	✓	✓	✓
	min_salary_transformed		✓	✓	✓	✓	✓
	pay_period_reclassified_transformed			✓	✓	✓	✓
	skill_abr_transformed				✓	✓	✓
	experience_level_transformed					✓	✓
	industry_id_transformed						✓

The model building method is Forward Stepwise using the Information Criterion.
A checkmark means the effect is in the model at this step.

Fig. 4-Steps to building the perfect Linear Modelling

The model building method used is Forward Stepwise Selection, which is a strategy for selecting features to add to a model one at a time. The target variable for the model is average_salary. The effects that have been added to the model are max_salary_transformed, min_salary_transformed, pay_period_reclassified, skill_abr_transformed, and experience_level_transformed.

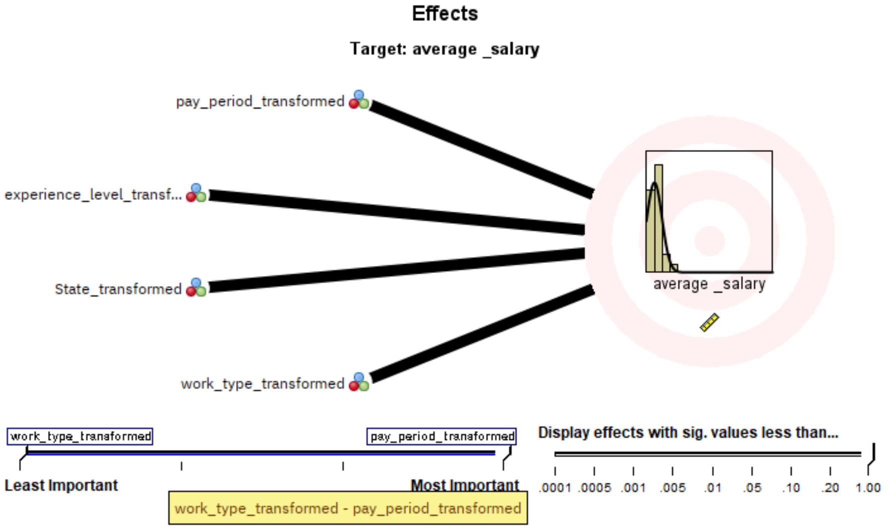


Fig.5-The Effect of Inputs on the Target Values

Effects of pay, person and transformed variables on a target variable, average_salary. The most key factors (with the lowest significance values) are listed on the right side of the table. The least key factors are on the left.

the purpose of this diagram is to summarize the findings of a statistical analysis on what factors most influence average salary.

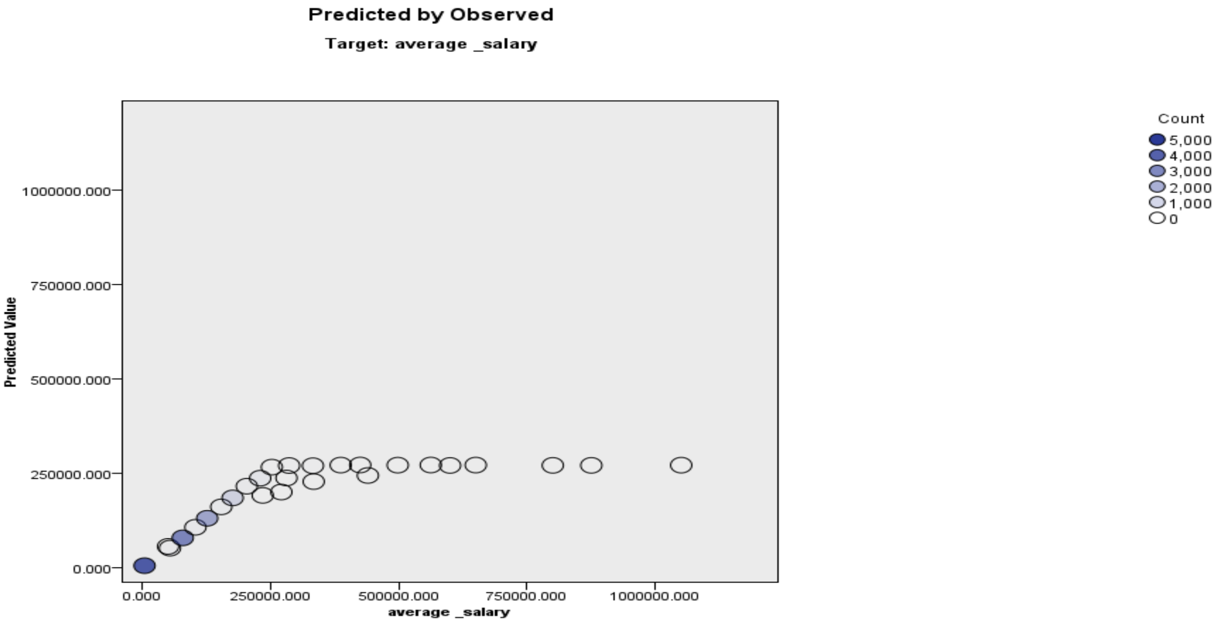
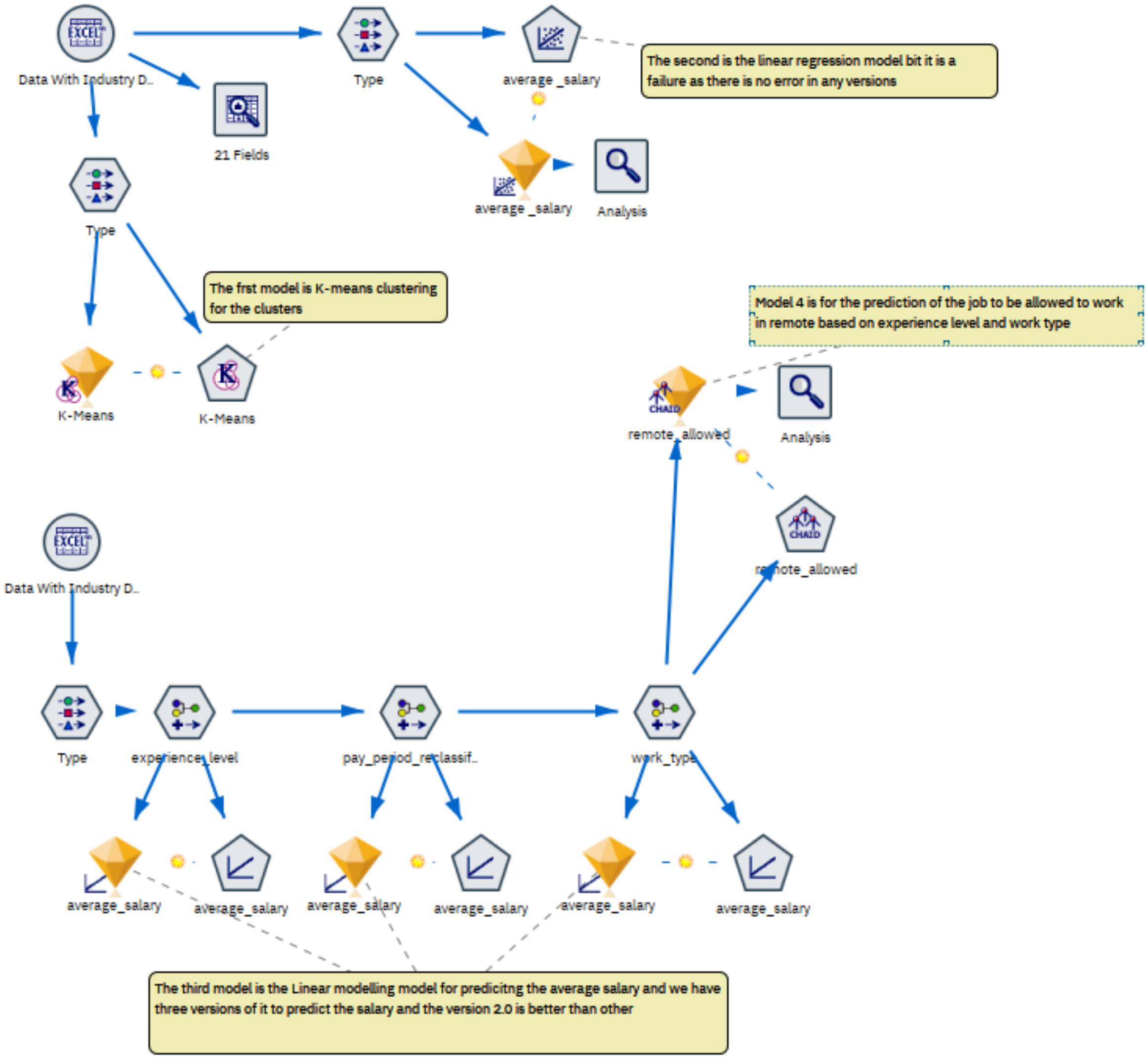


Fig.6 -Predictors observed over average salary

The graph would show a cluster of points concentrated around a diagonal line. This would indicate the model is accurately predicting average salaries.

Final Streamline for Modelling



The second is the linear regression model bit it is a failure as there is no error in any versions

The frst model is K-means clustering for the clusters

Model 4 is for the prediction of the job to be allowed to work in remote based on experience level and work type

The third model is the Linear modelling model for predicitng the average salary and we have three versions of it to predict the salary and the version 2.0 is better than other

Fig.7- Final Model Stream

We have created 4 Models to help us predict the key factors of a job

1. K-Means clustering: To predict the clusters of the work type that are important in a job. We decided to employ this model to delineate the spectrum of work types with respect to varying levels of experience.
2. Linear Regression: While evaluating the model we found that there was no error and while analyzing more in detail found that the model failed as it never gave an error or no less than 99% prediction. We decided ***not to*** use this model again as it is a failure.
3. Linear Modelling: As we can clearly see in the model summary 2.0 has the best accuracy among all three model to predict the average salary, but we also know that model 2 has more inputs that are related to the average salary the max and min salary inputs thus it has more accuracy. Thus, we are going to use the model 2.0 & 3.0 and test it to know which is better.
4. Chaid : While analyzing this model, we concluded it is good to predict the job to get allowed for remote or no. We will summarize the assessment results in terms of business success criteria, determining whether the project has successfully met its initial objectives. Approved models that meet the selected criteria will be identified for further consideration.

Key Findings

JOB POSTING BY STATE

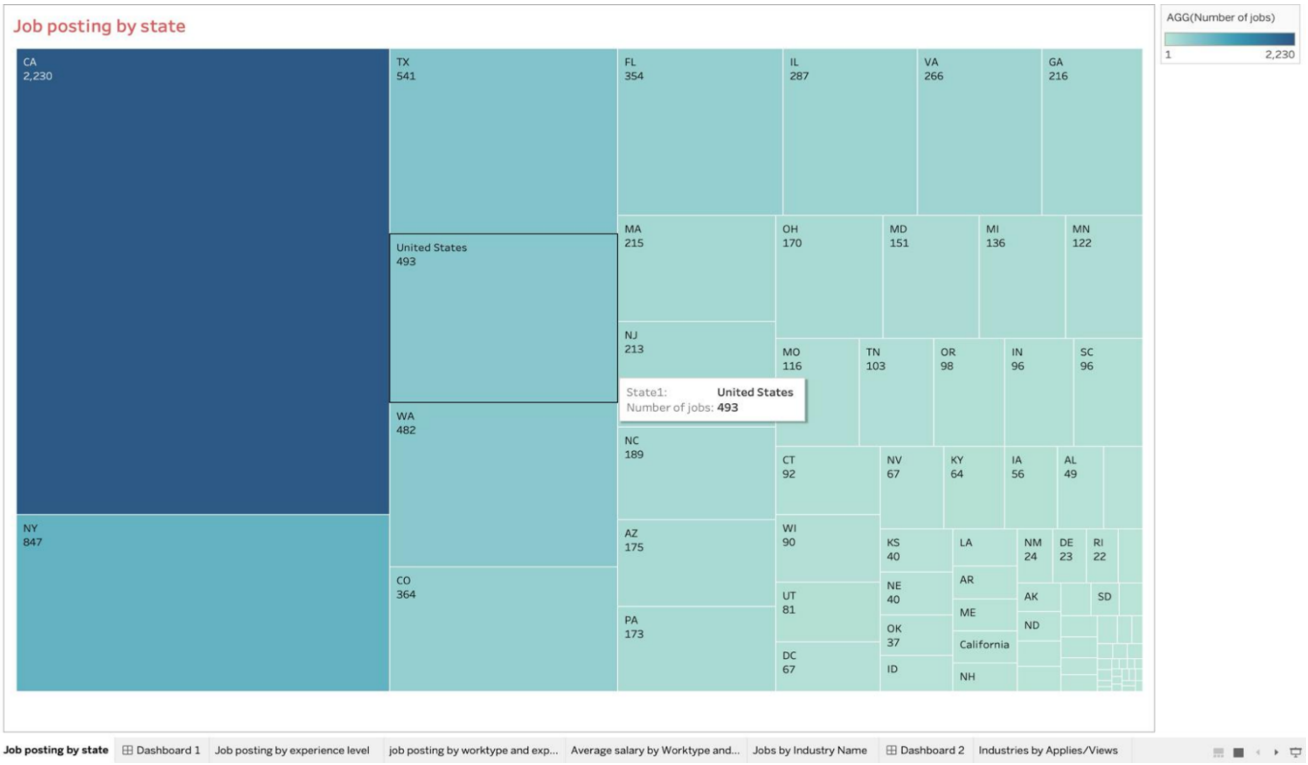


Fig.8-Job Posting by State

Understanding the geographical distribution of job postings across states provides valuable insights into regional job markets and economic activity. By examining the concentration of job opportunities, organizations can tailor their recruitment strategies and resource allocation to effectively target regions with high demand for skilled labor.

The analysis reveals the distribution of job postings across several states within the United States. Among the states analyzed, California (CA) boasts the highest number of job postings, with 2290 positions. Following closely behind is New York (NY) with 868 job postings, and Texas with 566 positions. The United States accounts for 514 job postings. Washington (WA) and Florida (FL) also exhibit notable numbers of job postings, with 494 and 378 positions, respectively.

Examining the relationship between states in terms of job postings may uncover interesting insights into economic dynamics and labor market trends. For example, neighboring states or those with similar economic profiles may exhibit comparable patterns in job postings. Conversely, differences in job postings between geographically close states could indicate unique regional factors influencing employment opportunities, such as industry composition, regulatory environments, or demographic trends. Further analysis of these relationships could provide deeper insights into regional economic interdependencies and opportunities for collaboration or competition among states.

Analyzing these trends can aid policymakers, businesses, and job seekers in making informed decisions regarding workforce development, relocation, and talent acquisition strategies.

Final Dashboard

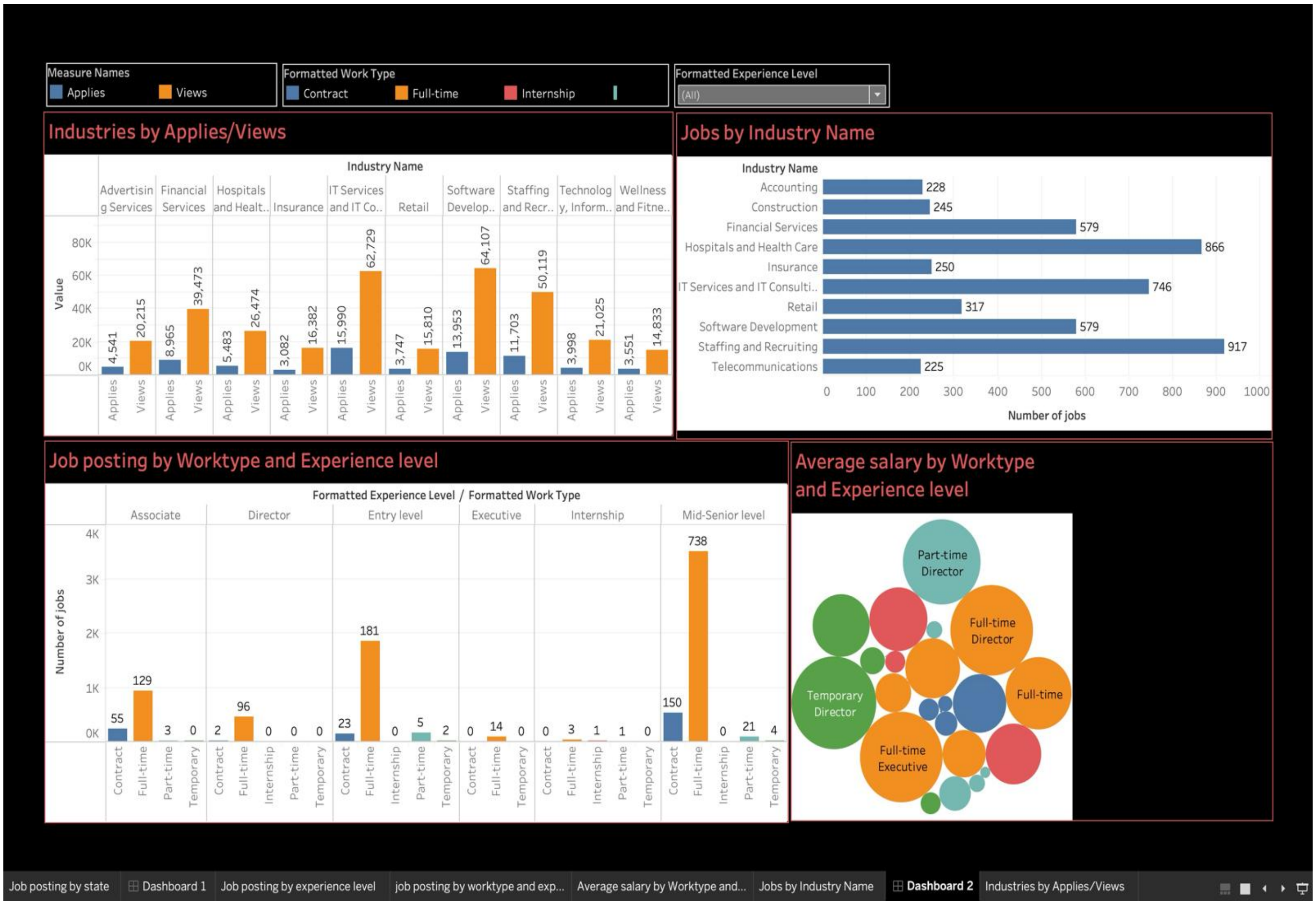


Fig.9- Final Dashboard

1.Industries by Applies/Views

This chart titled "Industries by applies/ views" shows the number of applications and/or views for job postings categorized by industry.

- Industry attractiveness: Industries with a high number of applications and views compared to others might be considered more attractive to job seekers. This could be due to factors like salary, work-life balance, or industry growth.
- Job seeker interest: Analyzing the application/view ratio can provide insights into the level of competition for jobs within an industry. High applications relative to views suggest a competitive landscape, while the opposite might indicate a lack of qualified candidates.

2. Jobs by Industry Name

The chart appears to be a horizontal bar chart titled “Jobs by Industry Name” from a Tableau dashboard. It shows the number of job postings for various industries. Here are the key findings we can glean from this chart:

- Distribution of Jobs by Industry: The chart provides a snapshot of the distribution of job postings across different industries. Industries with the highest number of postings are experiencing higher demand for workers. In this dashboard, we can determine the exact numbers of job postings per industry, as we can see Staffing and Recruiting, Hospitals and Health Care, and IT Services and IT Consulting have the most postings.
- Comparison for Specific Needs: This chart can be helpful for job seekers who are interested in understanding which industries are currently hiring the most. By looking at the industries with the most postings, they can target their job search efforts more effectively.
- Industry Demand by Experience: We can analyze which industries have the highest demand for workers at a particular experience level (e.g., Mid-Senior). This can be helpful for job seekers to identify industries where their experience is most valued.
- Identify Skill Gaps: By comparing the distribution of job postings across experience levels within an industry, we might identify potential skill gaps. For instance, an industry with a high number of Mid-Senior postings but limited Entry-Level opportunities might suggest a need for training programs to bridge the gap.

By analyzing both charts together, we can gain a deeper understanding of job seeker interest and the overall health of the job market within different industries. We can identify industries with high job growth, strong candidate interest, and potentially higher competition for top talent.

3. Demand by Worktype and Experience:

The chart titled "Job posting by Worktype and Experience level" reveals the distribution of job postings across different work types and experience levels. This can help identify:

- High demand worktypes: Worktypes with a significantly higher number of postings compared to others indicate high demand in the job market.
- Experience sweet spot: By analyzing which experience level (entry-level, mid-level, senior) has the most postings within a worktype, we can see where the current talent gap lies.

4. Salary by Worktype and Experience:

The chart titled "Average salary by Worktype and Experience level" displays the average salary offered for different worktypes and experience levels. This allows us to understand:

- Salary disparity: We can see how salaries vary across worktypes and how much experience influences earning potential.
- Competitive compensation: Worktypes with high average salaries for specific experience levels might be more competitive in attracting talent.

By analyzing both charts together, we can gain valuable insights into the job market. We can identify worktypes with high demand and the corresponding experience level employers seek. Additionally, we can understand how salaries are distributed across worktypes and how experience influences compensation.

Conclusion

To sum up, our data analytics effort on LinkedIn job posts provides a revolutionary perspective for comprehending and navigating the future of careers. It seeks to offer priceless insights for governments, employers, and job seekers alike through rigorous analysis and predictive modelling. Through identifying new employment opportunities, emphasising in-demand skills, and identifying key areas for professional development, the project enables people to make well-informed choices on their educational goals and future routes. In addition, it gives organisations the ability to pinpoint areas for skill development, maximise talent acquisition tactics, and maintain competitiveness in a labour market that is changing quickly. Beyond the scope of individual careers, the project's consequences also act as a stimulation for educational institutions and policymakers to support workforce development programmes that promote economic growth and societal resilience, as well as to align curriculum with industry demands. In the end, this project serves as a light of understanding, providing stakeholders with clear and insightful guidance through the shifting terrain of the workforce of the future.

References

- https://www.kaggle.com/datasets/arshkon/linkedin-job-postings?resource=download
- Anderson, B.A., Knestrick, J.M., & Barroso, R. (2015). Capstone Projects: Exemplars of Excellence in Practice: Springer Publishing Company.
- Heddle, N. M. (2007). The research questions. Transfusion, 47(1), 15-17.
- TopResume. https://www.topresume.com/career-advice/soft-skills-and-how-to-showcase-them-on-resume
- Tukey, John (1977), *Exploratory Data Analysis*, Addison-Wesley.
- License information: CC BY-SA 4.0