WESTERN GOVERNORS UNIVERSITY

Data Analytics Career Insights:

A Data Driven Approach to Choosing the Right Path

By: Natosha Minto

Data Analytics Career Insights
A Data Driven Approach to Choosing the right Path

Western Governors University

Table of Contents

A. Project Overview

When nearing the end of obtaining a Bachelor of Science in Data Analytics degree, choosing a specialization within the vast field of Data Analytics can feel both exciting and challenging. With roles such as Data Scientist, Data Engineer, Machine Learning Engineer, Business Intelligence Analyst, and many more, understanding which paths offer the best alignment with career goals, income expectations, and work environment preferences requires careful evaluation. This project seeks to analyze publicly available datasets to provide insights into job demand, salary trends, remote work opportunities, and skill experience across various roles. By leveraging these insights, the project aims to guide informed decision-making for career planning. This analysis not only benefits individual job seekers but can also be adapted from an organizational perspective. Companies aiming to build or expand their data departments can use this approach to assess trends in hiring, understand skill demands in the industry, and identify areas for growth by comparing hiring patterns across competing organizations.

A1. Research Question or Organizational Need:

What roles within the Data Analytics field are most in demand, and how do they compare in terms of salary, required skills and remote work availability?

A2. Context and background:

The decision of which career path to pursue within the Data Analytics field can be both daunting and critical, particularly as an individual nears the completion of a Data Analytics degree. As new graduates face an overwhelming number of career options, each with its own set of requirements, salary expectations, and work environments, it becomes vital to make data-driven decisions to align career choices with personal goals and market trends. Similarly, organizations seeking to build or expand their data departments face similar challenges in determining which roles are most essential for growth. By leveraging data analysis, both individuals and companies can gain valuable insights into current trends, demand for specific job titles, and the factors that influence successful career paths and department development.

Data analysis can provide clarity by systematically examining job market trends across different Data Analytics roles. Publicly available datasets, such as those from job listing sites like Glassdoor and LinkedIn, offer a wealth of information about job titles, salaries, company sizes, work environments, and skill requirements. Analyzing this data allows for a deeper understanding of which roles are growing in demand, the skills most sought after by employers, and the salary ranges associated with different positions. By quantifying these aspects, individuals can make more informed decisions about which career paths are most promising and align with their preferences for job stability, income, and work-life balance.

For example, it may reveal whether data science or data engineering roles offer higher salaries or if remote work opportunities are more prevalent in certain specializations. For organizations, understanding trends in the job market is crucial when determining where to invest in expanding their data teams. Companies can use similar datasets to assess which roles are becoming more prevalent in the market and identify areas where their teams may need to grow. For instance, if the analysis shows that data science roles are in high demand, organizations might prioritize hiring data scientists to handle growing volumes of data.

Additionally, analyzing industry hiring patterns can help organizations stay competitive by ensuring they are hiring for roles that align with market demands and trends. Furthermore, data analysis enables the

identification of trends that may not be immediately apparent through casual observation. For example, by examining the prevalence of remote work across different job titles, it becomes possible to see how the rise of remote work is reshaping the job market for Data Analytics professionals. Companies can use this data to inform their recruitment strategies, while individuals can factor in their preference for remote work when making career decisions.

Ultimately, data analysis is essential for uncovering insights that might otherwise go unnoticed, enabling both individuals and organizations to make strategic, data-driven decisions. Whether for personal career planning or organizational growth, understanding market trends through data analysis can provide a clearer path forward, ensuring that decisions are based on current and relevant information.

A3. Summary of Published Works:

The U.S. Bureau of Labor Statistics (BLS) provides detailed insights into the data science profession, including its job outlook, qualifications, and work environment. According to 2023 studies, data scientists typically require a bachelor's degree in fields like mathematics, statistics, or computer science, with some employers preferring advanced degrees. The role involves extracting meaningful insights from data, often in office settings with full-time schedules. The median annual wage for data scientists in May 2023 was $108,020, and the field is expected to grow by 36% from 2023 to 2033, much faster than average. With 20,800 job openings projected annually, this growth reflects the rising demand for data science skills across industries.

The 2021 O'Reilly Data & AI Salary Survey highlights compensation trends and professional development in data science and AI fields. The average salary for data and AI professionals was $146,000, with regional variations such as $176,000 in California. However, gender disparities persist, with women earning 84% of what men earn. Certifications in cloud platforms like AWS and Microsoft Azure significantly boost earning potential, emphasizing the value of continual upskilling. The survey also notes that 64% of respondents pursued training to enhance their skills or gain certifications, underlining the importance of flexibility and career growth in this field. These findings align with industry trends that encourage professionals to adapt to evolving job market demands while maintaining work-life balance.

A 2023 analysis from 365 Data Science offers an optimistic perspective on the data science job market, even amidst tech layoffs and the growing influence of AI. Data scientists accounted for only 3% of layoffs compared to 22% for software engineers, underscoring their crucial role in business and technology. The rise of AI further elevates the importance of skills in statistics, Python, machine learning, and APIs, as these are essential for driving innovation. Analyzing 1,000 job postings in 2024, the study found that nearly half of employers require a degree in data science, while Python (57%), machine learning (69%), and natural language processing (19%) remain highly sought-after skills. Cloud certifications are increasingly valuable, with 19.7% of roles requiring them, and salaries range from $160,000 to $200,000 annually. The study emphasizes that while degrees in data science or related fields can be advantageous, a mix of technical proficiency, programming knowledge, and analytical expertise is key to thriving in this rapidly evolving profession.

Together, these insights from the BLS, O'Reilly, and 365 Data Science highlight the growing demand, lucrative opportunities, and adaptability required to excel as a data scientist.

Links to Articles:

U.S. Bureau of Labor Statistics: https://www.bls.gov/ooh/math/data-scientists.htm

O'Reilly: https://www.oreilly.com/radar/2021-data-ai-salary-survey/

365 Data Science: https://365datascience.com/career-advice/data-scientist-job-market/

A3a. How A3's Published work informs project Development:

Each of the published works informs the development of the project by providing valuable insights into the current state of the data science job market, the skills required, and the trends shaping the profession's future.

The U.S. Bureau of Labor Statistics (BLS) study helps inform the project by offering foundational data on the demand for data scientists, their qualifications, and the expected job growth over the next decade. With projections of 36% growth from 2023 to 2033, the BLS study provides essential context for understanding the long-term stability and opportunities within the field. This aligns with the project's focus on the future of data science careers, confirming that there will be a significant need for skilled professionals in the coming years. The wage data and required qualifications from the BLS also guide the project's analysis of what it takes to succeed in the field, highlighting the importance of formal education and expertise in core data science skills.

The O'Reilly 2021 Data & AI Salary Survey contributes to the project by offering insights into the compensation landscape, revealing how regional salary variations and certifications influence earning potential. The report's emphasis on cloud certifications, particularly AWS and Microsoft Azure, informs the project's understanding of the growing value of these certifications in the job market. By highlighting the gender pay gap and the importance of career flexibility, the O'Reilly survey provides a broader view of the data science profession, which is crucial for the project's analysis of compensation trends and career growth. The survey also reinforces the idea that ongoing training and certification play a vital role in career advancement, which ties into the project's exploration of the skills needed to remain competitive in the data science field.

Finally, the 365 Data Science report offers an optimistic outlook on the data science profession, particularly in light of recent tech layoffs and the rise of AI technologies. The report's findings that data scientists are less affected by layoffs compared to other tech roles, and its insights into the increasing demand for skills in machine learning, natural language processing, and cloud certifications, directly inform the project's understanding of how data science professionals can thrive despite economic challenges. This study helps contextualize the impact of AI on the job market, reinforcing the idea that data scientists, with the right skill set, will remain essential in the AI-driven future. It also offers up-to-date data on job postings and skills demand, which are integral to the project's analysis of the evolving requirements for aspiring data scientists.

Together, these published works provide a comprehensive view of the data science job market, salary trends, and the skills needed to succeed, offering key data points and perspectives that inform the development of the project and guide its analysis of current and future opportunities in the field.

A4. Deliverables for data analytics solutions implemented:

The primary deliverables of this project will include:

1. **Graphs, Charts, and Plots**:
   Utilizing Python libraries (Matplotlib, Seaborn) to create visualizations such as bar charts, plots, and histograms, these will provide insights into the distribution of job roles, salary ranges, and the prevalence of remote work across Data Analytics positions.

2. **Salary Analysis Report**:
   A Summarizing salary trends, including average salaries across different Data Analytics roles, and variations based on experience levels.

3. **Work Environment Trends**:
   Visual and narrative analysis of the prevalence of remote, hybrid, and in-person roles across different Data Analytics job titles and industries.

4. **Comparative Dataset Analysis**:
   A comparative analysis of the job market trends using datasets from Glassdoor and LinkedIn, highlighting any differences or similarities in trends across both platforms.

A5. Benefits for data analytics solutions to support a decision- making process:

Key outcomes and deliverables of this project will include:

Career Path Insights: By analyzing job market trends in the Data Analytics field, the project will provide a comprehensive comparison of various job roles, including Data Science, Data Engineering, and Data Architecture. The primary outcome will be identifying the most in-demand roles, the required experience levels, and salary expectations, helping to guide career decisions.

Salary Analysis: A detailed analysis of salary trends across different job titles and experience levels, highlighting discrepancies and patterns in compensation. This will provide insights into which roles offer the best salary potential, considering factors such as location, company size, and work setting (remote, hybrid, in-person).

Work Environment Trends: This project will analyze the prevalence of remote, hybrid, and in-person roles across different job titles, company sizes, and industries, offering valuable insights into the work settings most commonly associated with Data Analytics roles. This outcome will help understand how work flexibility aligns with specific career paths.

Comparative Dataset Analysis: By leveraging datasets from Glassdoor and LinkedIn, the project will deliver a comparative analysis of the two data sources, ensuring that insights are based on a broader and more balanced perspective. The project will produce several visualizations, such as plots, and bar charts, to showcase trends in salary, experience, job titles, and work settings. These visualizations will be clear and actionable for both personal decision-making and broader industry trends.

Comprehensive Report: A final report will summarize all findings, including key trends, analysis techniques, and insights drawn from the datasets. The report will be structured to highlight actionable takeaways for both personal career decision-making and broader industry analysis.

Presentation of Results: A presentation summarizing the project's findings and visualizations will be delivered to demonstrate the project's outcomes clearly and effectively to a target audience, such as potential employers or industry professionals.

These outcomes will provide valuable, data-driven insights into the Data Analytics job market, equipping individuals with the knowledge needed to make informed career decisions and helping organizations assess the evolving demand for Data Analytics roles.

B. Data Analytics Project Plan

B1. Goals, Objectives and Deliverables

The goal of this project is to analyze publicly available datasets to uncover hiring trends, salary ranges, and work environment preferences within the Data Analytics field. By identifying in-demand positions, skills, and remote work opportunities, the project aims to help students entering the workforce make informed decisions about their career paths. Additionally, the analysis will offer insights into how companies experiencing growth can optimize their hiring strategies to identify roles most beneficial for organizational expansion. Ultimately, the project seeks to provide actionable recommendations tailored to entry-level professionals and companies looking to align their hiring practices with industry trends.

B2. Scope of Data Analytics Career Insights:

This project focuses on analyzing hiring trends, salary expectations, and work environment preferences within the Data Analytics field, specifically for roles such as Data Scientist, Data Engineer, Machine Learning Engineer, and Business Intelligence Analyst. It will leverage two publicly available datasets, combining them to explore variables such as job titles, salaries (including conversion to USD for standardization), experience levels, remote work ratios, and company size. The analysis aims to provide actionable insights for both individual job seekers, such as students entering the workforce, and organizations seeking to optimize their hiring strategies.

Key activities include:

- Cleaning and preparing the datasets to ensure consistency and alignment.
- Conducting descriptive and comparative statistical analyses to identify salary trends, remote work opportunities, and in-demand skills.
- Visualizing findings through charts and graphs to facilitate understanding and communication.

The project excludes advanced predictive modeling and focuses on current and historical data to provide insights relevant to entry-level professionals and organizations. Deliverables include a detailed analysis report, a presentation of key findings, and recommendations tailored to the target audiences. The project is limited to publicly available data and does not incorporate proprietary or sensitive information.

B3. Planning methodology:

CRISP-DM to Organize and Implement the Project:

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a well-suited methodology for this project as it provides a structured approach to organizing and implementing data analysis tasks. CRISP-DM consists of six phases, which align seamlessly with the project's objectives:

1. **Business Understanding:**
   In this phase, we define the project's goals and success criteria. For this project, the goal is to identify hiring trends, salary benchmarks, and job preferences for roles in Data Analytics to guide career and organizational decision-making. The key questions include: What are the most in-demand roles? What are the salary expectations for beginner-level positions?
2. **Data Understanding:**
   This phase involves exploring the datasets from Glassdoor and LinkedIn to assess their relevance and comprehensiveness. We examine variables such as job titles, salary, work environments, and company size to understand the data's potential to answer the research questions.
3. **Data Preparation:**
   Here, we clean and preprocess the data to ensure consistency and accuracy. This includes handling missing values, aligning variable names across datasets, converting salaries to USD, and standardizing work environment classifications. Combining datasets into a cohesive structure ensures alignment for analysis.
4. **Modeling:**
   Though predictive modeling is not the focus; statistical and comparative analyses will be performed to identify trends. Visualization tools like Matplotlib will be used to create graphs and charts that highlight key findings, such as salary distributions and remote work trends.
5. **Evaluation:**
   The results will be assessed to ensure they meet the project's goals. For example, are the insights actionable for students deciding on a career path? Do the findings provide valuable hiring strategy suggestions for organizations? Feedback loops will refine the analysis as necessary.
6. **Deployment:**
   The final deliverables will include a comprehensive report and a presentation summarizing the findings. Visualizations will be shared to communicate trends effectively, with tailored recommendations for both individual and organizational audiences.

CRISP-DM ensures the project remains focused and structured, with clear deliverables at each phase, making it ideal for organizing and implementing this project.

B4. Milestones Timeline:

This timeline will ensure that each phase of CRISP-DM is addressed, and that progress is tied to specific deliverables at key milestones.

Week 1: Business Understanding
- Start Date: 10/18/2024
- End Date: 11/24/2024
- Activities:
  - Define the project objectives and goals.
  - Identify key questions and the problem to be solved (e.g., understanding job trends in Data Analytics).
  - Document success criteria and deliverables.

Week 2: Data Understanding
- Start Date: 10/25/2024
- End Date: 10/31/2021
- Activities:
  - Locate and evaluate publicly available datasets related to the topic.
  - Explore datasets to assess relevance, structure, and data quality.
  - Identify gaps or additional data needs.

Week 3: Data Preparation
- Start Date: 11/1/2024
- End Date: 11/7/2024
- Activities:
    - Wrangle, clean, and preprocess data (handle missing values, standardize formats).
    - Merge datasets and ensure cohesion for analysis.
    - Document data preparation steps for reproducibility.

Week 4: Modeling
- Start Date: 11/8/2024
- End Date: 11/13/2024
- Activities:
    - Select and apply appropriate analysis techniques (visualization, comparisons).
    - Run statistical models to extract insights.
    - Validate the models and refine them based on results.

Week 5: Evaluation
- Start Date: 11/14/2024
- End Date: 11/20/2024
- Activities:
    - Assess findings to ensure they align with project objectives.
    - Review analysis outcomes for accuracy and relevance.
    - Begin compiling results for reporting.

Week 6: Deployment - Task 1 Submission
- Start Date: 11/21/2024
- End Date: 11/27/2024
- Activities:
    - Submit findings and insights for Task 1 based on business objectives.
    - Include key observations.

Week 7: Deployment - Task 2 Submission
- Start Date: 11/28/2024
- End Date: 12/5/2024
- Activities:
    - Align the results from Tasks 1 and 2.
    - Incorporate additional findings into a cohesive presentation for Task 3.

Week 8: Deployment - Task 3 Submission
- Start Date: 12/6/2024
- End Date: 12/15/2024
- Activities:
    - Finalize the comprehensive report and submit Task 3.
    - Reflect on project findings and their implications for decision-making.

B5. Cost and Resources:

This project incurred no additional costs as it utilizes publicly available datasets and open-source resources. The datasets used for analysis were gathered from Glassdoor.com and LinkedIn job listings, both offering free access to information on job roles, salaries, and company information. The data wrangling and analysis were performed using tools such as Python, Pandas, Matplotlib, and Jupyter Notebooks, all of which are open-source platforms. As a result, no paid third-party services, hardware, or specialized software were required for execution.

The main resources used included:

- **Data sources**: Glassdoor and LinkedIn (public datasets)
- **Software**: Python (Pandas, Matplotlib, Jupyter Notebooks)
- **Development environment**: Local machine
- **Work hours**: Estimated total work hours are around 120 hours, allocated across data cleaning, analysis, and report generation.

Since these resources are freely accessible, the overall cost of the project was minimal, aside from time and effort for data processing and analysis.

B6. Measurable criteria for Evaluating Project Execution:

1. Data Completeness and Quality
    - Target: 90% of data fields should be free of errors, inconsistencies, and missing values after the cleaning process.
    - Metric: Data validation checks (missing value identification, consistency across datasets).
    - Goal: Ensure that the datasets are usable for drawing actionable insights and answering the research questions.

2. Alignment with Objectives
    - Target: The analysis should successfully answer key questions regarding hiring trends, salary benchmarks, job demand, and remote work preferences within Data Analytics.
    - Metric: Specific alignment with business objectives, ensuring results guide both job seekers and companies in their decisions.
    - Goal: Provide valuable insights that help guide career decisions for students and hiring strategies for companies.

3. Timely Completion of Milestones
    - Target: All project milestones should be completed within the designated 8-week timeline.
    - Metric: Adherence to timeline for each phase (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment).
    - Goal: Ensure that the project progresses without delays and that key milestones are achieved on time.

4. Accuracy of Analysis
    - Target: Analysis and calculations must be free of significant errors. Statistical methods and trend analyses must be accurate.
    - Metric: Verification through cross-checking and ensuring data processing steps are correctly executed.
    - Goal: Ensure that the analysis provides valid and reliable results that can be used for decision-making.

5. Practical Insights and Recommendations
    - Target: The project should deliver clear, actionable insights regarding hiring trends, salary expectations, and remote work opportunities.
    - Metric: Feedback from stakeholders on the clarity and usefulness of the recommendations.
    - Goal: Help job seekers make informed decisions and assist organizations in optimizing their hiring practices.

6. Stakeholder Satisfaction
    - Target: Aim for at least 85% positive feedback from stakeholders (instructors, peers, or evaluators).
    - Metric: Evaluation rubric scores and qualitative feedback from stakeholders.
    - Goal: Ensure the project meets the expectations of all stakeholders and delivers value.

7. Documentation and Reproducibility
    - Target: Ensure that all data preparation steps, analysis processes, and findings are thoroughly documented and reproducible.
    - Metric: Clear, well-organized code, detailed project report, and a reproducible analysis workflow.
    - Goal: Provide transparency and allow others to replicate the analysis or build upon the work.

By meeting these measurable criteria, the project will be considered successful in terms of execution, quality of analysis, and achieving the intended goals.

C. Design of Data Analytics Solution

C1. Hypothesis:

The Data Analytics field is rapidly evolving, with organizations collecting increasing volumes of data daily. It is hypothesized that Data Scientist roles are in the highest demand due to their expertise in data wrangling, cleaning, and analysis—skills critical for transforming raw data into actionable insights. This hypothesis aims to uncover trends in job demand and salary distributions, particularly between Data Science and Data Engineering roles, to guide career and strategic decisions.

C2. Analytical Methods:

For this project, the following analytical methods will be implemented to analyze the Glassdoor and LinkedIn datasets effectively:

1. **Descriptive Analytics**

    This method will summarize key characteristics of the datasets, such as average salaries, the frequency of remote or hybrid roles, and job title distributions. For example, descriptive

statistics will identify central tendencies (mean, median) and variations (range, standard deviation) to understand salary ranges or the proportion of roles offering remote work.

2. **Exploratory Data Analysis (EDA)**
   EDA will involve visualizing data through histograms, scatter plots, box plots, and bar charts to uncover patterns, trends, and relationships. This method is critical for identifying outliers, inconsistencies, or missing data, as well as for understanding correlations between variables such as salary and experience level.
3. **Correlation Analysis**
   This method will measure relationships between numeric variables, such as salary and experience, or company size and work settings. For instance, Pearson's correlation coefficient will be used to quantify the strength and direction of these relationships, helping identify factors that influence job opportunities or salary differences.
4. **Comparative Analysis**
   Since data is sourced from Glassdoor and LinkedIn, comparative analysis will identify differences and similarities across the platforms. This includes comparing salary ranges, job title distributions, and work setting trends (remote vs. in-office roles) between the two datasets.
5. **Data Visualizations**
   Visualizations like bar charts, line graphs, and heatmaps will communicate insights clearly, making trends and relationships easily interpretable for stakeholders.

C2a. Justification of Analytical Methods:

These methods are appropriate because they directly address the project's goals of understanding job market trends, salary disparities, and work settings.

- **Descriptive Analytics** provides a foundational summary of key data characteristics, enabling quick identification of trends such as which roles are most common.
- **EDA** allows for deeper exploration of relationships and highlights potential data issues that could affect accuracy.
- **Correlation Analysis** quantifies key relationships, such as the link between experience level and salary, providing evidence-based insights.
- **Comparative Analysis** ensures a broader perspective by examining trends across platforms, reducing the risk of platform-specific biases.
- **Data Visualizations** help present findings in a clear, accessible format, making the results actionable for job seekers and organizations.

These methods ensure accurate, actionable insights aligned with the objective of understanding the Data Analytics job market.

C3. Tools and Environments:

The data analytics solution was developed using the following tools and environments.

Jupyter Notebook

- Jupyter Notebook served as the primary development environment for writing and executing Python code. Its interactive features made it ideal for exploring the data, running iterative analysis, and visualizing results.

Python

- Python, a versatile and widely used programming language for data analysis, was utilized for all data processing and analysis tasks.

Python Libraries

- **Pandas**: Used for data manipulation and cleaning, including loading datasets, handling missing values, and restructuring data for analysis.
- **NumPy**: Used for numerical operations and supporting data structures like arrays.
- **Matplotlib and Seaborn**: Employed to create visualizations such as histograms, plots, and bar charts to represent insights.
- **Scipy**: Used for statistical testing during comparative analysis.

Data Sources

- The project leveraged publicly available datasets from Glassdoor and LinkedIn. These datasets were loaded and processed within Jupyter Notebook for analysis.

Hardware/Software Environment

- The project was executed on a local machine running a Windows operating system.

Version Control

- GitHub: Used to manage version control during the project.

C4. Methods and Metrics to Evaluate Statistical Significance:

To evaluate the output of the data analytics solution, a combination of descriptive statistics and visual analytics methods were employed. Specifically, counts, means, medians, modes, and data visualizations such as plots and charts were utilized. These methods provided insights into the central tendencies, distributions, and patterns within the datasets gathered from Glassdoor and LinkedIn. The application of these methods ensured the identification of trends, anomalies, and outliers, enabling a comprehensive analysis of the data.

C4a. Justification for the Chosen Methods and Metrics:

The use of these methods and metrics is appropriate for evaluating the output of the data analytics solution for several reasons:

1. **Descriptive Statistics (Counts, Mean, Median, and Mode):**

   These measures summarize the key characteristics of the datasets in a clear and concise manner, which is essential for understanding underlying trends and behaviors in the data.
   - The **mean** identifies the average value, useful for evaluating typical trends across job salaries or demand indicators.
   - The **median** is a robust measure of central tendency, particularly effective in skewed datasets such as salary distributions where extreme values (outliers) can distort the mean.
   - The **mode** helps to detect the most frequently occurring values, offering insights into common patterns such as preferred job titles or locations.
2. **Visual Analytics (Plots and Charts):**
   - Visual representations of the data enhance interpretability by making complex relationships and trends apparent at a glance. For instance, bar charts and scatter plots can reveal correlations or distributions more effectively than raw numbers.
   - Visual tools also facilitate communication of insights to stakeholders who may not have a technical background, thereby supporting decision-making.

These chosen methods and metrics align with the goals of the analysis by ensuring clarity, accuracy, and accessibility of insights derived from the datasets. They provide a structured framework for evaluating the model's output in a way that supports actionable outcomes.

C5. Practical Significance:

To assess the practical significance of the data analytics solution, the focus will be on evaluating its ability to provide actionable insights, drive meaningful outcomes, and support decision-making processes. This assessment will begin by determining whether the insights derived from the analysis address the key questions related to salary trends and job demand across data analytics roles. A critical factor is ensuring the results align with the project's goal of providing valuable guidance for career planning and strategic decision-making. Additionally, the clarity and applicability of the insights will be evaluated to ensure they are actionable. For instance, findings such as identifying high-paying job locations or trends in remote work opportunities must be specific and practical enough to influence decisions effectively.

The solution's performance will also be measured against baseline metrics to establish whether it improves upon prior knowledge or decision-making capabilities. For example, the analysis must provide new perspectives on salary ranges or job availability that were previously unclear or unavailable. The visualizations and statistical summaries produced will be assessed for their effectiveness in guiding stakeholders, such as career planners or hiring managers, to make informed decisions. Stakeholder feedback will play a key role in validating that the analysis meets their needs and supports confident decision-making.

Furthermore, the efficiency and scalability of the solution will be considered, ensuring that it can be easily replicated or adapted to broader datasets or other job markets with minimal modifications. Finally, the impact of the analysis on its intended audience will be evaluated, specifically its ability to help individuals make informed career decisions or assist organizations in shaping hiring strategies. If the analysis provides actionable, relevant, and decision-supportive insights that stakeholders find valuable and applicable to real-world challenges, the solution will be deemed practically significant.

C6. Visual Communication:

By visually communicating the findings of the data analytics solution, a variety of tools and graphical representations will be utilized to present the insights clearly and effectively. The report will include multiple visualizations that provide a comprehensive view of the analyzed datasets, tailored to highlight key trends and comparisons. While some visualizations will be determined based on the findings during the analysis phase, specific charts and graphs have been identified as essential to the project.

Bar charts will be used to illustrate the number of job listings per job title and to highlight the availability of entry-level positions. These charts will also visually depict the median salaries for entry-level roles, making it easy to compare job opportunities and compensation across titles. Summary statistics will complement these visualizations by showcasing the minimum, maximum, and median salaries for each job title, providing a clear picture of salary ranges.

A doughnut chart will visualize the distribution of remote jobs across different job titles, offering insight into the prevalence of remote work opportunities in the analyzed roles. Additionally, a stacked horizontal bar chart will display the distribution of remote, onsite, and hybrid positions by job title, allowing stakeholders to easily understand the dynamics of workplace flexibility within the field.

These visualizations will be created using tools from Python libraries such as Matplotlib and Seaborn, ensuring high-quality and customizable outputs. The combination of these graphical representations will effectively communicate complex data trends, making the findings accessible to diverse audiences. This approach ensures that the report not only conveys findings effectively but also supports informed decision-making.

D. Description of Datasets:

D1. Source of Data:

The datasets used in this project are sourced from two well-known platforms: Glassdoor and LinkedIn. The Glassdoor dataset contains detailed job-related data, including variables such as job title, salary, work setting, experience level, employment type, company size, and location. The LinkedIn dataset, complements the Glassdoor data by providing similar variables while ensuring a broader and more diversified view of the Data Analytics job market.

D2. Appropriateness of the Data:

These datasets are highly suitable for analyzing trends in the Data Analytics job market. They include critical variables such as job title, salary (standardized in USD), remote ratio, experience level, and employment type, which align with the project's goals of understanding job demand, salary trends, and work settings in the field. The inclusion of geographical and company-size data allows for deeper insights into regional trends and company-specific job dynamics. By leveraging data from two major platforms, the analysis mitigates bias and ensures robust conclusions by capturing a wide spectrum of job postings.

D3. Data Collection Methods:

- Dataset 1 (Glassdoor): The dataset is publicly available on Glassdoor and was manually downloaded from their platform. Glassdoor provides job-related data that is open for analysis by users for non-commercial purposes. As this is publicly accessible data, it is permissible to use for the capstone project.

- Dataset 2 (LinkedIn): The dataset was collected from LinkedIn job postings via the LinkedIn API, a method commonly used for gathering publicly available data. LinkedIn provides data for educational and research purposes, as long as it complies with their terms of service and usage restrictions. As this data is publicly accessible for research use, it is appropriate to use for this project.

After both datasets were obtained, they were then processed into Pandas DataFrames for further cleaning and analysis, ensuring consistency across variables for comparison.

D4 Observations on Data Quality and Completeness:

Both datasets are of high quality, providing comprehensive job-related information. However, minor inconsistencies were observed, such as varying formats for salary data and differences in how job titles were categorized. These issues will be addressed during the data-cleaning process. Additionally, while both

datasets are rich in variables, there may be gaps in certain fields, such as incomplete company size data or missing salary information in specific postings. Overall, the datasets are sufficiently complete for a meaningful analysis, provided that appropriate cleaning and preprocessing steps are performed.

D5. Data Governance, Privacy, Security and Ethical Considerations:

- **Data Governance:** The data was sourced from publicly accessible platforms (Glassdoor and LinkedIn) and is being used for educational purposes, adhering to their terms of service. Both datasets will be securely stored and managed, ensuring compliance with best practices in data governance.
- **Data Privacy and Security:** As the datasets contain publicly available information, no personally identifiable information (PII) is involved. However, precautions will be taken to anonymize company-specific insights to avoid unintended breaches of confidentiality. The data will be stored in encrypted environments, and access will be restricted to authorized project contributors.
- **Ethical, Legal, and Regulatory Compliance:** The use of this data aligns with the ethical and legal frameworks of both platforms. Data from LinkedIn was gathered using their API, which complies with their guidelines for non-commercial, research purposes. Similarly, Glassdoor's terms allow data usage for educational projects.

D5a. Precautions: To uphold data governance and ethical standards, the analysis and presentation of findings will avoid disclosing sensitive information about specific companies or roles. Any visualizations or reports will aggregate data to maintain confidentiality. Additionally, the data will be used exclusively for this capstone project and not for any commercial endeavors, ensuring compliance with platform usage policies.

E. Sources and References

Planning methodology: CRISP-DM to Organize and Implement the Project:

https://www.datascience-pm.com/crisp-dm-2/

Data obtained from:

https://www.glassdoor.com/index.htm

https://www.linkedin.com/feed/

Related published works:

https://www.bls.gov/ooh/math/data-scientists.htm

https://www.oreilly.com/radar/2021-data-ai-salary-survey/

https://365datascience.com/career-advice/data-scientist-job-market/