Data Analytics Capstone Topic Approval Form

Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information.

Student Name: Natosha Minto

Student ID: 011693982

Capstone Project Name: Data Analytice Career Insights: A Data-Driven Approach to Choosing the Right Path

Project Topic: When nearing the end of a Bachelor of Science in Data Analytics, 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, and Business Intelligence Analyst, 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 requirements 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.

Research Question: 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?

> **Hypothesis:** The field of Data Analytics is rapidly evolving, with organizations collecting increasing volumes of data daily. A key challenge lies in not just gathering data but utilizing it effectively for decision-making and innovation. It is hypothesized that Data Scientist roles are in highest demand due to their expertise in data wrangling, cleaning, and analysis, which are critical to helping organizations derive actionable insights from raw data.

Context: 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.

Data: To effectively analyze trends in the Data Analytics job market, several key data variables are essential. These include the job title, which identifies the specific roles in demand (Data Scientist, Data Engineer, Data Analyst), and the job category, which indicates the broader field or specialization (Data Science, Data Engineering, Business Intelligence). Both variables are critical for understanding the demand across different job sectors. Salary data is another important variable, offering insights into the financial compensation for these roles, with a focus on salary in USD for standard comparison. This is particularly useful for assessing the economic attractiveness of different job categories. Additionally, the work setting or remote ratio is crucial for evaluating the prevalence of remote, hybrid, or in-person roles in the job market. As many job seekers prioritize work flexibility, understanding these trends can guide career decisions. Experience level is another necessary variable, as it indicates the level of expertise required for various roles, helping to assess the opportunities available to individuals at different stages of their careers. The employment type (full-time, part-time, or contract) provides further context on the stability and nature of job offerings, which is important for long-term career planning. Company location provide geographical context, helping to explore trends in job availability and salary discrepancies across regions, while company size (categorized as small, medium, or large) can influence job roles and benefits. Understanding these aspects will allow for a more holistic view of the job market and guide decisions about where and what roles to target. For this analysis, two existing datasets will be used. The first dataset was sourced from Glassdoor, a platform offering detailed insights into various job categories within Data

Analytics, including roles like Data Engineering, Data Science, and Data Architecture. It also provides information on work settings, which is crucial for identifying trends in remote, hybrid, or in-person work environments. This dataset contains variables such as job title, salary, work setting, experience level, employment type, and company size, among others. It was downloaded manually from Glassdoor and processed into a DataFrame for analysis. The second dataset was sourced from LinkedIn and complements the Glassdoor dataset by providing additional perspectives on job titles, salaries, employment types, and remote work ratios. This dataset helps validate the findings from the Glassdoor data and ensures a broader understanding of the job market by incorporating data from another widely-used platform. It includes similar variables such as job title, salary, experience level, employment type, and company location. The data from LinkedIn was programmatically downloaded and processed into a DataFrame. By combining both datasets, this analysis aims to reduce potential biases and provide a more comprehensive view of the job market trends in Data Analytics.

Ownership and Permission: Data is obtained from Glassdor & LinkedIn

- **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.

Both datasets will be cleaned and prepared for analysis by ensuring consistency between variables, making them ready for comparison and insightful analysis. Given that the data is publicly available and used for educational purposes, permission to use it in this capstone project is granted.

Data Gathering: *Describe the data-gathering methodology you will use to collect data.* Glassdoor data is downloaded directly from their platform and LinkedIn data can also be downloaded directly from their platform or also wrangled using an API or Web scraping.

Data Analytics Tools and Techniques:
To analyze the data from the Glassdoor and LinkedIn datasets effectively, several data-analysis techniques can be applied. These techniques will help identify trends, patterns, and relationships between job titles, salaries, experience levels, work settings, and other variables. The key techniques for this analysis will include:

1. **Descriptive Statistics**: This technique will be used to summarize the main features of the datasets, including measures of central tendency (mean, median) and dispersion (standard deviation, range). Descriptive statistics will provide an overview of salary distributions, the frequency of remote or hybrid roles, and the spread of job titles and experience levels. For example, it will allow us to compare the average salaries across different job categories and experience levels, or identify the most common work settings.
2. **Exploratory Data Analysis (EDA)**: EDA is crucial for understanding the data before diving into more complex analyses. This includes visualizing distributions and relationships between variables using histograms, box plots, scatter plots, and bar charts. For example, a scatter plot could visualize the relationship between salary

and experience level, while bar charts could help compare the prevalence of remote work across job categories. EDA will also help identify any data quality issues (such as missing or inconsistent values) that need to be addressed.

3. **Correlation Analysis**: To understand the relationships between numeric variables, such as salary, experience level, and company size, correlation analysis will be employed. This technique will allow us to determine the strength and direction of the relationships between key variables. For example, we can assess whether higher experience levels correlate with higher salaries, or whether larger companies offer more remote positions. Pearson's correlation coefficient can be calculated to quantify these relationships.
4. **Comparative Analysis**: Since we are comparing two datasets (Glassdoor and LinkedIn), it is important to perform comparative analysis to identify similarities and differences in job market trends across platforms. For example, we can compare salary data for similar job titles across both platforms or assess the prevalence of remote work in different sectors or job categories. Statistical tests such as t-tests or ANOVA may be used to assess whether differences between job categories or datasets are statistically significant.
5. **Regression Analysis**: If there are hypotheses about predicting job salaries or other outcomes based on specific variables, linear regression analysis could be used. For instance, we can use regression to predict salary based on experience level, company size, and other factors. This will help in determining the relative importance of different predictors of salary and job settings.
6. **Cluster Analysis**: If the goal is to segment the job market into distinct categories based on job roles, salaries, or work environments, cluster analysis (such as k-means clustering) can be applied. This technique will group similar job roles together based on shared characteristics, helping to identify patterns or job categories that are closely related to each other.
7. **Data Visualization**: Throughout the analysis, data visualization tools, such as Matplotlib and Seaborn in Python, will be used to create clear and insightful charts and graphs. Visualizations will be key to conveying the findings of the analysis in an accessible and meaningful way, allowing for quick interpretation of trends and patterns.

Together, these data-analysis techniques will help uncover valuable insights about the Data Analytics job market, such as the most in-demand job roles, the salary ranges for different job titles, the impact of experience level, and trends in remote work. These insights will provide actionable information for those entering the industry, as well as help organizations assess where their hiring practices may align with broader market trends.

Justification of Tools/Techniques: The data-analysis techniques chosen are appropriate because they align with the goals of understanding job market trends, salary disparities, and work settings. Descriptive statistics provide a clear overview of the data, allowing us to identify central tendencies and variations in salary, job categories, and work environments. Exploratory Data Analysis (EDA) enables visual inspection of relationships between variables and helps identify outliers or data quality issues before further analysis. Correlation analysis helps quantify relationships between numeric variables like salary and experience, providing deeper insights into job market dynamics. Comparative analysis across two datasets allows for a broader perspective, minimizing platform-specific biases. Regression analysis helps predict salaries based on key factors, which is valuable for decision-making. Cluster analysis segments the job market into distinct categories, aiding in understanding which roles are similar or different. These techniques combined will ensure a comprehensive and data-driven approach to identifying actionable trends in the Data Analytics job market.

Application Type, if applicable (select one):

- ☐ mobile
- ☐ web
- ☒ stand-alone

Programming/Development Language(s), if applicable: Python

Operating System(s)/Platform(s), if applicable: N/A

Database Management System, if applicable: I plan to process data directly in Python

Project Outcomes: Key outcomes and deliverables of this project would 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: The 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. Data Visualization: The project will produce several visualizations, such as histograms, scatter 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. Predictive Insights: Utilizing regression models, the project will predict salary ranges based on key variables like experience, job category, and work setting. This will provide practical recommendations for professionals seeking to maximize their earning potential within the Data Analytics field. 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.

Projected Project End Date: 12/20/2024

Sources: Glassdoor & LinkedIn

Human Subjects or Proprietary Information

Does your project involve the potential use of human subjects? (Y/N): No
Does your project involve the potential use of proprietary company information? (Y/N): No

---

STUDENT SIGNATURE

Natosha Minto

By signing and submitting this form, you acknowledge that any cost associated with the development and execution of your data analytics solution will be your (the student) responsibility.

TO BE COMPLETED BY AN INSTRUCTOR

The capstone topic is approved by an instructor.

COURSE INSTRUCTOR SIGNATURE:

Jim Ashe, Ph.D. Mathematics
COURSE INSTRUCTOR APPROVAL DATE:

11/27/2024

Project Compliance with IRB (Y/N): Y