



PROJECT REPORT

DATA ANALYTICS  
  
STUDENT SUCCESS METRICS: A COMPREHENSIVE DATA ANALYSIS

|  |  |  |  |
| --- | --- | --- | --- |
| **Created By:** | Narendra Patil | **Approved By:** | Suraj Mane |
| **Created On:** | 11-06-2024 | **Approved On:** | DD-MMM-YYYY |

**INDEX**

[**1** **PROJECT DETAILS** 1](#_Toc176867060)

[**2** **SUMMARY** 1](#_Toc176867061)

[**3** **INTRODUCTION** 1](#_Toc176867062)

[3.1 Background 1](#_Toc176867063)

[3.2 Stakeholders 2](#_Toc176867064)

[3.3 Objectives 2](#_Toc176867065)

[**4** **METHODOLOGY** 2](#_Toc176867066)

[4.1 Considerations & Assumption 2](#_Toc176867067)

[4.2 Approach 3](#_Toc176867068)

[4.3 Activities 4](#_Toc176867069)

[**5** **TARGETTED V/S ACHIEVED OUTPUT** 6](#_Toc176867070)

[**6** **CONCLUSION** 7](#_Toc176867071)

[**7** **APPENDICES** 8](#_Toc176867072)

[7.1 Appendix A – Metadata 8](#_Toc176867073)

[7.2 Appendix B – Tools and Technologies used 8](#_Toc176867074)

# **PROJECT DETAILS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name** | Student Success Metrics: A Comprehensive Data Analysis | | |
| **Project Sponsor** | Tushar Topale | | |
| **Project Manager** | Harshada Topale | | |
| **Start Date** | 11-06-2024 | **Completion Date** | 11-09-2024 |

# **SUMMARY**

The project aimed to conduct a detailed analysis of student interns to understand the relationship between academic performance, event participation, career aspirations, and other factors influencing their success. By analysing a dataset with attributes such as GPA, family income, event attendance, leadership skills, and career expectations, we sought to identify patterns and correlations that could help improve internship outcomes and student support.

Why the Project was Needed:

* To gain a deeper understanding of what drives student success, particularly in areas like career aspirations and academic performance.
* To evaluate how factors like family income, event participation, and leadership skills impact student performance and career expectations.
* To provide data-driven insights to tailor mentorship, support programs, and event planning for future students.

Expected Deliverables:

* Detailed exploration and analysis of the dataset to identify key trends.
* Visualization of relationships between different factors, such as family income and GPA, event participation and career expectations, etc.
* Insights that can help guide event promotion, curriculum improvement, and student support strategies.

Long-Term Benefits:

* The findings can be used to better align academic programs, career development activities, and events to student needs.
* Future event planning and career guidance initiatives can be informed by the data, potentially leading to improved student engagement and performance.
* The analysis may help institutions identify areas where students need more support, leading to better overall success rates and more targeted educational efforts.

# **INTRODUCTION**

## Background

The project aims to provide valuable insights into factors influencing student success. This analysis will help identify key areas for intervention and support, ensuring that students from all economic backgrounds have equal opportunities for success. The benefits include enhanced student support programs, better-informed career services, and improved job placement rates.

## Stakeholders

Project Owner: Harshada Topale

Key Stakeholders: Harshada Topale

End Customers: Cloud Counselage pvt ltd.

## Objectives

The Conduct a comprehensive analysis to uncover insights into the relationships between students' economic backgrounds, academic performance, competence, and expected salary.

# **METHODOLOGY**

## Considerations & Assumption

This section outlines the key risks, assumptions, and constraints considered while conducting the project.

Risks:

* Data Quality: The analysis heavily relies on the quality of the dataset. Incomplete, missing, or inaccurate data may affect the validity of insights derived from the analysis. Careful data cleaning and validation were required to mitigate this risk.
* Technical Challenges: The project involved handling potentially large datasets, which may have posed technical difficulties during processing and analysis. Issues such as slow performance or memory limitations could arise.
* Stakeholder Expectations: A risk of misalignment existed between what stakeholders expected from the analysis and what the data and timeframe could realistically deliver. This required careful communication throughout the project.

Assumptions:

* Comprehensive and Accurate Dataset: It was assumed that the dataset provided was complete and contained no significant outliers that could distort the analysis. Any anomalies in the data were expected to be minimal or handled during data cleaning.
* Stakeholder Availability: Regular feedback from stakeholders was expected during key project stages, ensuring that the analysis met their needs and expectations.
* Access to Necessary Tools: It was assumed that all required tools (e.g., Python libraries like Pandas, Seaborn, and Matplotlib) and software would be available throughout the project without any disruptions.

Constraints:

* Time: The project had a limited timeframe, meaning that the scope of the analysis and the depth of insights generated were constrained by the deadline.
* Resources: The analysis was limited to personal capacity, with no additional team members or external resources available to assist. Python tools and libraries were the primary resources for all aspects of data handling, transformation, and visualization.
* Technology: All stages of the project, from data cleaning to visualization, were exclusively performed using Python. No other technologies or programming languages were considered, limiting the flexibility of the toolset.

## Approach

To achieve the project’s objective of gaining insights into the relationship between students' academic performance, event participation, career aspirations, and influencing factors, a structured and systematic approach was followed. The project was divided into several key phases, each aimed at progressively refining the data and generating actionable insights. Here’s the conceptual outline of the approach adopted:

* Data Collection & Understanding: The first step involved acquiring a dataset that contained relevant attributes for each student. The dataset was carefully examined to understand its structure, key variables, and potential insights it could offer. This phase was crucial to ensure the dataset captured all essential factors related to students' performance and participation.
* Data Cleaning: In this phase, a thorough cleaning of the dataset was performed. This involved handling missing values, addressing inconsistencies, and removing irrelevant data. Special attention was given to outliers and anomalies that could skew the results. The goal was to ensure the dataset was accurate, consistent, and ready for analysis.
* Data Transformation & Manipulation: Following data cleaning, the dataset was transformed to derive new insights. For example, columns were created to calculate students’ total experience, categorize them by event participation, and track their performance metrics. This phase was aimed at enhancing the dataset by making it more analysable for specific metrics of interest.
* Exploratory Data Analysis (EDA): EDA was performed to uncover patterns, correlations, and trends in the data. Visualizations such as histograms, box plots, and bar charts were used to explore the relationships between academic performance (GPA), event participation, and other variables like family income and leadership skills. The objective was to identify significant factors influencing students' success.
* Focused Student Analysis: After understanding the overall trends, a deep dive into specific student attributes was conducted. This included identifying high-achieving students, understanding event preferences, and correlating academic performance with career aspirations. The analysis was extended to assess the impact of variables like family income on GPA and expected salary.
* Stakeholder Review & Feedback: Throughout the project, regular feedback loops with stakeholders were established. This ensured that the analysis was aligned with their expectations, and any gaps were identified early. Adjustments were made based on this feedback to focus on areas of highest relevance.
* Results Interpretation & Reporting: In the final phase, the findings were summarized and reported in a clear, actionable manner. The insights were structured to be easily interpreted by stakeholders, highlighting key factors that influenced student success and offering recommendations for future student development and event planning.

This structured approach was adopted to ensure the project was organized, efficient, and produced meaningful insights. Each phase built upon the previous one, ensuring a logical flow from data gathering to final recommendations, making it possible to address the problem systematically while minimizing risks.

## Activities

To successfully deliver the project and achieve the desired outcomes, a series of key activities were performed. Each activity contributed to the structured approach and ensured the project remained on track and aligned with stakeholder expectations. The major activities included:

1. Requirement Gathering:

* Collaborated with stakeholders to understand the project’s objective, scope, and desired insights.
* Identified the necessary attributes and variables to be included in the dataset, focusing on academic performance, event participation, career aspirations, and influencing factors like family income.
* Clarified the expectations regarding the depth of the analysis and the types of reports or insights that would be delivered.

1. Data Collection:

* Obtained the dataset containing attributes for each student, including GPA, events attended, family income, expected salary, leadership skills, and career aspirations.
* Ensured the dataset covered all relevant parameters for a holistic analysis.

1. Data Cleaning:

* Performed data validation checks to identify and address missing values, duplicates, and inconsistencies.
* Cleaned up erroneous entries and handled outliers that could distort the analysis.
* Dropped rows where key information (such as "College Name") was missing or irrelevant.

1. Data Transformation & Preparation:

* Transformed raw data into analysable formats by calculating additional fields like total experience with Python and leadership influence.
* Categorized students into different groups based on relevant metrics such as GPA and event participation for deeper analysis.
* Formatted the dataset for compatibility with Python tools like Pandas and Seaborn for analysis and visualization.

1. Exploratory Data Analysis (EDA):

* Conducted an exploratory analysis to discover key trends, correlations, and patterns across various dimensions like GPA, event participation, and family income.
* Utilized visualizations such as box plots, histograms, and bar charts to highlight relationships between variables.
* Examined the influence of leadership skills, family income, and other factors on academic performance and expected salary.

1. Student-Specific Analysis:

* Performed focused analyses on key student groups, including those with high CGPA and more experience in Python, to uncover specific insights about their career expectations and success factors.
* Analysed the distribution of students by event, promotion channel, and college participation to understand engagement trends.

1. Reporting & Visualization:

* Generated visual reports to represent findings clearly, including bar charts for student participation and box plots for GPA vs. family income, leadership skills, and expected salary.
* Summarized the insights in a clear, concise manner for stakeholder review, focusing on actionable recommendations.

1. Review & Feedback:

* Engaged with stakeholders throughout the project to ensure the analysis met their expectations.
* Incorporated feedback to refine and focus the analysis on areas of highest priority for the stakeholders.

1. Final Presentation:

* Delivered a final report summarizing all key findings, including the impact of events, leadership skills, and family income on student performance and career aspirations.
* Highlighted long-term benefits of the analysis for improving event planning and student development strategies.

# **TARGETTED V/S ACHIEVED OUTPUT**

Targeted Output:

The project plan aimed to answer 16 of the following 18 questions:

1. How many unique students are included in the dataset?
2. What is the average GPA of the students?
3. What is the distribution of students across different graduation years?
4. What is the distribution of students' experience with Python programming?
5. What is the average family income of the students?
6. How does the GPA vary between different colleges? (Show top 5 results only)
7. Are there any outliers in the quantity (number of courses completed) attribute?
8. What is the average GPA of students from each city? (Show top 5 results only)
9. Can we identify any relationship between family income and GPA?
10. How many students are from various cities? (Show top 5 results only)
11. How does the expected salary vary based on factors like 'GPA', 'Family income', 'Experience with Python (Months)'?
12. Which event tends to attract more students from specific fields of study? (Student distribution according to event)
13. Do students in leadership positions during their college years tend to have higher GPAs or better expected salary?
14. How many students are graduating by the end of 2024?
15. Which promotion channel brings in more student participation for the event?
16. Find the total number of students who attended the events related to Data Science?
17. Those students who have high CGPA & more experience in Python also have high expectations for salary? (Avg)
18. How many students know about the event from their colleges? Which of these top 5 colleges?

Achieved Output:

The project successfully addressed all 18 questions as outlined in the targeted output. Each question was thoroughly analysed and answered with the following insights:

* Unique students count.
* Average GPA of the students.
* Distribution of students across graduation years.
* Distribution of students' experience with Python.
* Average family income of the students.
* GPA variation among different colleges (top 5 results).
* Identification of outliers in the quantity of courses completed.
* Average GPA of students from various cities (top 5 results).
* Relationship between family income and GPA.
* Distribution of students across cities (top 5 results).
* Variation in expected salary based on GPA, family income, and Python experience.
* Distribution of students according to the type of event.
* Comparison of GPAs and expected salaries for students with and without leadership skills.
* Number of students graduating by the end of 2024.
* Promotion channels influencing student participation.
* Number of students attending Data Science-related events.
* Analysis of expected salaries for students with high CGPA and Python experience.
* Students who learned about the event from colleges and identification of top 5 colleges.

Reasons for Deviation (if any):

* Data Quality and Completeness: While all questions were addressed, the quality of some data (e.g., "College Name" entries) occasionally required additional cleaning, which could have affected the depth of analysis for certain queries.
* Technical Limitations: Handling large datasets and performing complex transformations could have posed technical challenges. While these were managed effectively, any limitations in computing resources might have affected the processing speed or efficiency.
* Stakeholder Expectations: Some deviations might have occurred due to evolving expectations or additional requests from stakeholders that were not initially planned. However, these were managed through regular updates and adjustments to ensure alignment with the project's goals.

Overall, the project met its objectives and delivered comprehensive insights, aligning with the initial plan and addressing all 18 questions successfully.

# **CONCLUSION**

This project successfully analysed the dataset of student interns, providing valuable insights into their academic performance, event participation, career aspirations, and influencing factors. By answering all 18 targeted questions, we gained a comprehensive understanding of student demographics, GPA distributions, and factors affecting their success.

Usefulness for Stakeholders:

The findings offer actionable insights for stakeholders to tailor educational programs, career guidance, and event planning based on students' needs and profiles. Understanding trends in GPA, income, and event participation helps in making data-driven decisions to enhance student engagement and career readiness.

Future Scope:

Future work could involve integrating additional data sources, such as real-time performance metrics or feedback, to further refine insights. Expanding the analysis to include longitudinal studies could also provide deeper understanding of student progress over time. Additionally, incorporating machine learning techniques could predict future trends and outcomes more accurately.

# **APPENDICES**

## Appendix A – Metadata

|  |  |
| --- | --- |
| Column Name | Explanation |
| First Name | Identify student individually |
| Quantity | Number of courses complete by each student during their time at the institute |
| Email | Mail id of student |
| College Name | College Name from which student graduates |
| Year of Graduation | Year in which student completes their graduation |
| City | City of residence or hometown of each student |
| GPA (Grade Point Average) | Represents the academic performance of each student based on their GPA |
| Experience with Python Programming (in months) | Indicates level of experience each student has with python programming |
| Family Income | Income level of each student’s family |
| Expected Salary | Salary expectations of each student |
| Leadership Skills | Denotes if student is in leadership position at college or not. |

## Appendix B – Tools and Technologies used

* Excel: For Primary study of data and Project Documentation.
* Python: Primary programming language used for data analysis and processing.
* Pandas: Library for data manipulation and analysis.
* NumPy: Library for numerical computations.
* Matplotlib: Used for creating static, animated, and interactive visualizations.
* Seaborn: Statistical data visualization library built on top of Matplotlib.
* Jupyter Notebook: Interactive environment for running and documenting code.
* Excel: For data input and some initial processing.
* Git/GitHub: Version control system used to track changes and manage collaboration.
* Word: Project Documentation