



PROJECT REPORT

DATA ANALYTICS  
  
ABSENCE OF INSIGHTS FOR THE RELATIONSHIP BETWEEN STUDENT'S ECONOMIC BACKGROUND, ACADEMIC PERFORMANCE, COMPETENCE & EXPECTED SALARY.

|  |  |  |  |
| --- | --- | --- | --- |
| **Created By:** | Shivohum Pandita | **Approved By:** | <Data Analytics> |
| **Created On:** | 12-07-2025 | **Approved On:** |  |

Page left blank intentionally

**INDEX**

[**1** **PROJECT DETAILS** 2](#_Toc143445375)

[**2** **SUMMARY** 2](#_Toc143445376)

[**3** **INTRODUCTION** 2](#_Toc143445377)

[3.1 Background 2](#_Toc143445378)

[3.2 Stakeholders 3](#_Toc143445379)

[3.3 Objectives 3](#_Toc143445380)

[**4** **METHODOLOGY** 4](#_Toc143445381)

[4.1 Considerations & Assumption 4](#_Toc143445382)

[4.2 Approach 5](#_Toc143445383)

[4.3 Activities 5](#_Toc143445384)

[**5** **TARGETTED V/S ACHIEVED OUTPUT** 6](#_Toc143445385)

[**6** **CONCLUSION** 7](#_Toc143445386)

[**7** **APPENDICES** 8](#_Toc143445387)

[7.1 Appendix A – Title 8](#_Toc143445388)

# **PROJECT DETAILS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name** | ABSENCE OF INSIGHTS FOR THE RELATIONSHIP BETWEEN STUDENT'S ECONOMIC BACKGROUND, ACADEMIC PERFORMANCE, COMPETENCE & EXPECTED SALARY. | | |
| **Project Sponsor** | Tushar Topale | | |
| **Project Manager** | Harshada Topale | | |
| **Start Date** | 25-06-2025 | **Completion Date** | 13-07-2025 |

# **SUMMARY**

This project involved analyzing student participation data across multiple events to derive actionable insights for event organizers and recruiters. The work began with **extensive data cleaning** using Python (Pandas) to handle inconsistencies like duplicate registrations, mismatched CGPA, college names, cities, and missing values. Steps included:

* Standardizing column names and data formats.
* Merging multiple knowledge source columns into a single **knowledge\_source** field.
* Normalizing values (CGPA clipped to 0–10, converting experience and salary fields to numeric).
* Removing duplicates and resolving conflicting entries by aggregating at the per-student level using median/mode logic.
* Creating **calculated columns** such as Email\_Norm for unique identification and Channel\_First to extract the primary promotion channel.

Once the dataset was cleaned (3,902 validated records), a range of **Power BI measures and calculated tables** were created for deep analysis:

* **is\_student flag:** Ensures only students (not professionals) are included.
* **Per Student Grad table:** Aggregates graduation year per student for consistent year-based metrics.
* **Channel First column:** Identifies the main promotion channel per student.
* **Python Months table:** Used to build experience distribution histograms.
* **ExpectedSalary\_Q3 measure:** Calculates average salary expectations of top 25% students by CGPA & Python experience.

# **INTRODUCTION**

## Background

This data analytics project set out to understand how factors like academic performance, technical skills, and student background influence career-related expectations such as salary. By analysing student registration data from various events, the project aimed to uncover patterns and insights that could help guide students, educators, and organizations in making more informed decisions.

In today’s competitive job market, knowing what influences student success is more important than ever. This project shows how data analytics can help uncover meaningful trends in the education and career space—empowering institutions and stakeholders to support students more effectively and plan for future opportunities.

## Stakeholders

The key stakeholders involved in this project include:

* **Students:** As the primary data contributors, their academic records, experiences, and expectations form the core of the analysis. The insights generated can help them understand how their skills and background impact career outcomes.
* **Educational Institutions and Faculties:** Colleges and professors can use the findings to better guide students on academic and skill development paths aligned with industry trends.
* **Event Organizers (e.g., Cloud Counselage):** The data was collected through event registrations. Organizers are interested in understanding student demographics, participation patterns, and how to improve outreach and engagement strategies.
* **Career Counselors and Recruiters:** Insights about expected salary, CGPA, and experience can assist them in aligning hiring practices and advice with student readiness.
* **Data Analysts and Project Team:** Responsible for cleaning, analyzing, and presenting the data. Their goal is to ensure accuracy, reliability, and usability of the information.

## Objectives

The main objective of the project, as outlined in the Project Charter, was:

* **To analyse student data to understand how academic performance, technical skills, and personal background relate to career expectations and event participation.**

Though the Main Objective remains the same , there are some more objectives since Project Charter was approved, that is

Detecting Fraud or Anomalies: Identify unusual patterns within datasets, helping to mitigate risks, ensure compliance, and protect against financial losses.  
  
Improving Decision Making: Provide decision-makers with timely, accurate, and actionable insights to support strategic planning, risk management, and resource allocation, leading to better-informed decisions.

# **METHODOLOGY**

## Considerations & Assumption

While carrying out this data analytics project, I made the following considerations and assumptions to ensure the analysis remained relevant, clear, and accurate:

* **Data Cleaning Requirements: The raw dataset had multiple issues — same students registered for multiple events with different CGPA, graduation year, city, and college name values.**
* **Deduplication & Conflict Resolution:**
  + - **Exact duplicates were removed.**
    - **Records were grouped by email + event to consolidate information.**
    - **Median was used for continuous fields (CGPA, salary, Python experience) and mode for categorical fields (college name, designation).**
* **Data Standardization:** 
  + - * **Column names were converted to lowercase and standardized.**
    - **Emails and event names were converted to lowercase for consistency, city names set to title case.**
    - **"nan", "None", and empty strings were replaced with nulls.**
* **Data Validation:**
* **CGPA clipped to 0–10 range.**
* **Negative experience/salary values removed.**
* **Leadership skills converted to boolean format.**
* **Scope: Focused only on students (using is\_student flag), excluding professionals.**

These considerations helped guide how I handled the data and ensured that the conclusions drawn were based on thoughtful assumptions and practical limitations.

## Approach

My approach began with loading the “All Events Data” sheet into Python using Pandas and performing a detailed review of its structure and content. I standardized all column names by converting them to lowercase, replacing spaces with underscores, and mapped verbose names like “Email ID” to more concise alternatives such as email. I cleaned string data by trimming whitespace, converting emails and event names to lowercase, and applying title case to cities. I also handled missing values by replacing empty strings, “nan”, and “None” with nulls.

To resolve conflicts where the same student had multiple entries with varying details, I grouped the data by email + event and applied **median aggregation** for continuous fields like CGPA and **mode** for categorical fields like college name and designation, with a median fallback for numeric fields where the mode was unavailable. Knowledge source information from how\_did\_you\_know and others\_specify was merged into a single knowledge\_source column. Additionally, I clipped CGPA values to the 0–10 range, removed negative Python experience values, converted leadership skills into boolean format, normalized family income, and ensured expected salary was in numeric format.

For analysis, I created new calculated fields such as Email\_Norm for consistent student identification and Channel\_First to extract the primary promotion channel from multiple entries. I then built per-student aggregated tables to ensure each student had one consolidated record, enabling accurate metrics like average CGPA, graduation year, and expected salary. This cleaned and transformed dataset formed the foundation for building interactive Power BI dashboards, allowing me to explore trends, correlations, and participation patterns effectively.

I visualized and analysed it using tools like Excel and Tableau. This structured approach helped ensure insights were based on consistent and reliable data.

## Activities

To deliver the project effectively, I carried out the following activities:

* Data Cleaning: Removed duplicates, normalized string formats, corrected data types.
* Transformation: Created per-student tables, aggregated fields, normalized income and salary data.
* Integrity Checks: Excluded invalid or inconsistent rows (untrusted data).
* Visualization: Created Power BI dashboards to analyze graduation year, CGPA distribution, Python experience, salary expectations, and promotion channel effectiveness.
* Insight Generation: Answered all key questions (Q1–Q18) including distribution trends, correlations, and participation analysis.
* Reporting: Documented conclusions and recommendations for event organizers and recruiters.

# **TARGETTED V/S ACHIEVED OUTPUT**

**Targeted Output:**

The goal of the data analytics project was to understand how academic performance, skills, and background influence student outcomes—especially expected salary and participation in events. The targeted outputs included:

* Reliable & Clean Data: Remove duplicates, handle missing values, and resolve conflicts (e.g., varying CGPA, graduation year, and college names for the same student).
* Per-Student Consolidation: Ensure each student is represented only once using aggregated data (median/mode logic) for accuracy.
* Actionable Insights: Identify trends in graduation years, CGPA distribution, Python experience, and salary expectations.
* Promotion Channel Analysis: Determine which communication channels were most effective for event participation.
* Visualization & Reporting: Build interactive dashboards in Power BI to present insights in a user-friendly way for stakeholders.

**Achieved Output:**

The project achieved its main goals:

• Clean Dataset: Produced a trusted dataset of 3,902 rows, representing 2,109 unique students, with standardized columns and validated numeric values.

• Per-Student Metrics: Created a shared Per Student Grad table and calculated measures (e.g., is\_student, Channel\_First, ExpectedSalary\_Q3) to support consistent analysis across all visuals.

• Key Insights:

Most students graduate in 2024–2025 (peak at 714 students in 2025).

Average CGPA is 8.04 and average family income is ₹1.29 Lakh.

Students with high CGPA (>9) and more Python experience expect higher salaries (~17.7 LPA).

WhatsApp (41%) and Email are the most effective promotion channels.

824 students attended Data Science–related events, indicating strong interest.

• Interactive Dashboards: Built visualizations that allow filtering by college, city, event, and graduation year, enabling stakeholders to explore data dynamically.

• Decision Support: Delivered clear, data-driven recommendations to improve outreach strategies, focus on top colleges, and target final-year students for recruitment drives.

# **CONCLUSION**

This project transformed messy, inconsistent registration data into a **trusted, analysis-ready dataset** and uncovered actionable insights.  
Key findings include:

* Most students graduate in **2024–2025**, forming the main recruitment pool.
* Students with **higher CGPA and Python experience expect significantly higher salaries** (avg. 16.32 LPA for top 25%).
* **WhatsApp** is the most effective channel for outreach.
* Data Science events attract the **largest share of student participation** (824 students).
* Colleges like MIT Academy of Engineering and KLE Society’s College of BCA drive the highest participation.

These insights empower stakeholders to make **better event planning, training, and hiring decisions**.

**Future Scope**

* Predictive Modelling: Extend the analysis to forecast expected salaries or participation likelihood using machine learning models.
* Additional Data Sources: Incorporate placement data, internship history, or LinkedIn outcomes for deeper career insights.
* Skill-Level Analysis: Expand beyond Python experience to include other technical skills (SQL, ML, Cloud) to understand employability better.
* Automated Data Pipelines: Develop an ETL process to continuously clean and update data for real-time dashboards.

# **APPENDICES**

## Appendix A – Raw Data Samples

* Selected rows from the original “All Events Data” sheet showing typical issues such as duplicate entries, missing values, and inconsistent graduation year, CGPA, and city information.

## Appendix B – Data Cleaning Scripts

Python (Pandas) code snippets used to:

* Load and preprocess data (standardizing column names, trimming spaces).
* Merge how\_did\_you\_know and others\_specify into knowledge\_source.
* Convert CGPA, Python experience, and salary fields to numeric values.
* Remove duplicates and resolve conflicting records using median/mode aggregation.

## Appendix C – Calculated Columns & Measures

* Email\_Norm: LOWER(TRIM(email)) – ensures unique student identification.
* Channel\_First: Extracts the first promotion channel from knowledge\_source.
* Per Student Grad: Calculated table grouping data by normalized email and selecting maximum graduation year for consistency.
* ExpectedSalary\_Q3: DAX measure that calculates average expected salary for students in the top 25% by CGPA and Python experience.

## Appendix D – Visualizations

* Graduation Year Distribution: Bar chart showing number of students per graduation year.
* Python Experience Histogram: Distribution of students by months of Python programming experience.
* Salary vs CGPA Scatter Plot: Showing correlation between CGPA and salary expectations.
* Promotion Channel Pie Chart: Highlights WhatsApp as the dominant channel.
* Top 5 Colleges Table: Displays colleges with highest number of participating students.

## Appendix E – Conflict Resolution Examples

* Before-and-after examples of records where conflicting CGPA or college names were resolved using median/mode logic.

## Appendix F – Dashboard Snapshots

* Screenshots of key Power BI dashboards used for interactive analysis, including:
* Graduation Year interactive card and slicers.
* City-wise and college-wise participation heatmaps.
* Data Science event participation insights.

## 

## 