



PROJECT REPORT

PYTHON  
  
Fresher\_DataAnalysis\_Python

|  |  |  |  |
| --- | --- | --- | --- |
| **Created By:** | Aritra Sarkar | **Approved By:** | Harshada Topale |
| **Created On:** | 18-06-2024 | **Approved On:** | DD-MMM-YYYY |

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 2](#_Toc143445379)

[3.3 Objectives 2](#_Toc143445380)

[**4** **METHODOLOGY** 2](#_Toc143445381)

[4.1 Considerations & Assumption 3](#_Toc143445382)

[4.2 Approach 3](#_Toc143445383)

[4.3 Activities 3](#_Toc143445384)

[**5** **TARGETTED V/S ACHIEVED OUTPUT** 3](#_Toc143445385)

[**6** **CONCLUSION** 3](#_Toc143445386)

[**7** **APPENDICES** 4](#_Toc143445387)

[7.1 Appendix A – Data Visualization 4](#_Toc143445388)

**General Instructions for using the Live Project Report Template**

* This template and the subsequent document created using this template is a confidential document and is the intellectual property of Cloud Counselage Pvt. Ltd. Circulating it outside of the organisation without the consent of Cloud Counselage Pvt. Ltd. is the breach of company policies and will lead to legal actions
* This template is a guideline document to communicate the implementation of design ideas and the results of the work to the stakeholders.
* The **text between inequality (< >) is to be replaced** by relevant text
* Please **remove the yellow highlight on the Text** between the inequality (< >). This is done to help you notice the text to be changed/replaced
* The text in *italics* highlighted in grey is just for reference and should be removed after adding the relevant text

# **PROJECT DETAILS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name** | Fresher\_DataAnalysis\_Python | | |
| **Project Sponsor** | Tushar Topale | | |
| **Project Manager** | Harshada Topale | | |
| **Start Date** | 18-06-2024 | **Completion Date** | 09-07-2024 |

# **SUMMARY**

# **Project Deliverables:** The project was expected to provide insights into the relationship between students' economic backgrounds, academic performance, competence, and expected salary. Key deliverables included data visualizations, comprehensive documentation, Python code for data analysis, and a demo video summarizing the findings.

* **Need for the Project:** There was a clear gap in understanding how various factors such as economic background and programming experience impact students' academic and career outcomes. This project aimed to address this gap and provide valuable insights to improve the initial screening process for internships and job applications
* **Long-Term Benefits:** The insights gained from this project can help recruiters design more effective and equitable screening processes, leading to better job placements. Educational institutions can implement targeted interventions to support students from diverse backgrounds. Policymakers can use the findings to develop policies that ensure equal opportunities for all students. The project also serves as a foundation for future research and emphasizes the importance of programming skills and leadership in curriculum development.

# **INTRODUCTION**

## Background

Millions of students apply for internships and jobs each year, making the initial screening process crucial. Recruiters often reject over 70% of resumes due to a lack of detailed insights into applicants' backgrounds. This project addresses the need for insights into the relationship between students' economic background, academic performance, competence, and expected salary. By understanding these relationships, educational institutions and recruiters can make more informed decisions, leading to better support for students from diverse backgrounds and more effective recruitment processes. The dataset includes attributes such as GPA, family income, Python programming experience, and expected salary, which will be analysed to uncover key patterns and correlations.

## Stakeholders

* **Students:** They will benefit from insights into how their background, academic performance, and skills impact their career prospects.
* **Educational Institutions:** Schools and universities can use the findings to develop targeted support programs, improve curricula, and enhance student services.
* **Recruiters and Employers:** Companies can refine their recruitment processes, leading to more effective and equitable candidate selection.
* **Policymakers:** Insights will help in developing policies that support students from diverse economic backgrounds, promoting inclusivity.
* **Project Team:** The team, including data analysts and developers, will gain valuable experience and contribute to a high-impact project.

## Objectives

The objective of the project is to analysis comprehensively data of college students who applied for internships/jobs with help of python.

# **METHODOLOGY**

These conventions are all about the positions of line breaks, how many characters should go on a line, and everything in between.

## Considerations & Assumption

**Constraints:**

* **Data Availability:** The analysis depends on the availability and completeness of data provided by students and institutions. Incomplete or inaccurate data can limit the accuracy of the findings.
* **Data Privacy:** Ensuring the privacy and confidentiality of student data is a primary concern, which may restrict the type and amount of data that can be used.
* **Resource Limitations:** The project is constrained by the available resources, including time, technical tools, and expertise. This affects the scope and depth of the analysis.
* **Technological Limitations:** The project relies on the use of Python for data analysis, which may have limitations in handling large datasets or complex computations.

**Challenges:**

* **Data Quality:** Ensuring high-quality data is a challenge, as inconsistencies and lies can impact the reliability of the analysis.
* **Correlation vs. Causation:** Identifying relationships between variables does not imply causation, and drawing incorrect conclusions can mislead stakeholders.
* **Diverse Backgrounds:** Students come from diverse backgrounds, making it challenging to generalize findings across all demographics.

**Assumptions:**

* **Data Accuracy:** It is assumed that the data collected is accurate and represents the true attributes of the students.
* **Homogeneity of Data:** The analysis assumes that the data from different institutions and regions are comparable and can be analysed together.
* **Static Attributes:** The attributes such as family income and GPA are considered static for the duration of the analysis, even though they may change over time.
* **Student Participation:** It is assumed that the students included in the dataset have participated voluntarily and provided honest and accurate information.

## Approach

**1. Data Collection:** The project began with collecting a comprehensive dataset containing various attributes of the students, such as GPA, family income, Python programming experience, and expected salary. Ensuring the accuracy and completeness of this data was crucial for the subsequent analysis.

**2. Data Cleaning and Preprocessing:** Using Python, primarily with libraries like Pandas, the data was cleaned to handle missing values, remove duplicates, and correct inconsistencies. This step ensured that the dataset was ready for analysis.

**3. Exploratory Data Analysis (EDA):** We used Jupyter Notebook to perform EDA, utilizing visualization libraries such as Seaborn and Matplotlib. This helped in understanding the distribution of data, identifying trends, and uncovering any outliers or anomalies in the dataset.

**4. Feature Selection:** Key features relevant to the problem were selected. This included attributes like GPA, family income, Python programming experience, and leadership skills. These features were chosen based on their potential impact on academic performance and expected salary.

**5. Correlation Analysis:** We conducted correlation analysis to identify relationships between different attributes. Seaborn and Matplotlib were used to create heatmaps and scatter plots, visually representing these relationships. This step was crucial in identifying which factors most significantly impact students' outcomes.

**6. Model Building:** To quantify the relationships, linear regression models were built using Pandas and Scikit-Learn. These models helped in predicting the expected salary based on other attributes like GPA, family income, and Python experience. The choice of linear regression was due to its simplicity and effectiveness in establishing linear relationships between variables.

**7. Results Interpretation and Visualization:** The results from the regression models were interpreted to provide insights into how various factors influence the expected salary and academic performance. Visualizations were created using Seaborn and Matplotlib to present these findings in an easily understandable format.

**8. Reporting:** A comprehensive report was prepared, summarizing the methodology, analysis, and findings. This included visualizations, interpretations of the results, and recommendations based on the insights gained. The report was compiled in a standard industry format to ensure clarity and professionalism.

## Activities

**1.Requirement Gathering**: Gather detailed requirements from stakeholders, including what insights they seek from the dataset, specific questions to answer, and any constraints or preferences regarding tools and formats.

**2.Planning**: Create a project plan outlining tasks, timelines, resource allocation (such as team members or software tools), and milestones. This plan should also cover data acquisition, cleaning, analysis, and reporting phases.

**3**.**Data Acquisition**: Obtain the dataset containing information about student interns. Ensure data privacy and compliance with any applicable regulations (e.g., GDPR, HIPAA).

**4.Data Cleaning and Preparation**: Clean the dataset to handle missing values, outliers, and inconsistencies. Transform data into a format suitable for analysis, ensuring data quality and integrity.

**5**.**Data Analysis**: Perform exploratory data analysis (EDA) to understand distributions, correlations, and trends in the data. Apply statistical methods and machine learning techniques as necessary to answer specific questions.

**6.Visualization**: Create visualizations (e.g., plots, charts) to effectively communicate findings and insights from the data analysis. Use tools like matplotlib, seaborn, or Plotly in Python for this purpose.

**7**.**Interpretation and Insights**: Analyze results from the data to draw meaningful conclusions and insights related to the problem statement. Relate findings back to initial requirements and stakeholder expectations.

**8.Documentation**: Prepare comprehensive documentation including a project charter, project plan, risk assessment (RAID) logs, software requirements specification (SRS) document, project report summarizing findings, and any other relevant documentation.

**9**.**Presentation and Demo**: Present findings to stakeholders through a demo video and/or live presentation. Explain methodologies used, key insights discovered, and implications for decision-making.

**10**.**Feedback and Iteration**: Gather feedback from stakeholders and make any necessary revisions or refinements to the analysis or documentation based on feedback received.

# **TARGETTED V/S ACHIEVED OUTPUT**

### 🡺Targeted Outputs in Project Plan:

1. **Data Cleaning and Preparation**:
   * Target: Clean and prepare dataset to handle missing values, outliers, and inconsistencies.
   * Achieved: Successfully cleaned dataset, handled most missing values and outliers. Some inconsistencies required further attention.

**Reason for Deviation**: Initial assessment underestimated the complexity of data inconsistencies, requiring more time for thorough cleaning.

1. **Data Analysis**:
   * Target: Conduct comprehensive analysis including descriptive statistics, correlation analysis, and regression where applicable.
   * Achieved: Conducted thorough descriptive statistics and correlation analysis. Regression analysis partially completed, additional time needed for advanced models.

**Reason for Deviation**: Complexity of data relationships and need for more sophisticated models than initially anticipated.

1. **Visualization**:
   * Target: Create clear and informative visualizations using Python libraries (matplotlib, seaborn).
   * Achieved: Developed various plots and charts to visualize distributions, correlations, and trends effectively.

**Reason for Deviation**: Initially underestimated the time required for iterative refinement of visualizations to meet stakeholder expectations.

1. **Documentation**:
   * Target: Prepare comprehensive documentation including project charter, project plan, RAID logs, SRS document, and project report.
   * Achieved: Completed project charter, project plan, and SRS document. RAID logs and project report in progress.

**Reason for Deviation**: Balancing analysis and documentation timelines, prioritizing initial deliverables before final report completion.

1. **Presentation and Demo**:
   * Target: Create a demo video and prepare for live presentation to stakeholders.
   * Achieved: Demo video created, presentation slides drafted. Final polish and rehearsal ongoing.

**Reason for Deviation**: Iterative nature of data analysis and insights discovery required more time for preparation of cohesive presentation materials.

### 🡺Lessons Learned:

* **Time Estimation**: Better estimation of time required for data cleaning and advanced analysis techniques.
* **Iterative Process**: Acknowledgment of iterative nature of visualization and analysis refinement.
* **Documentation Planning**: Early integration of documentation milestones with analysis phases.

# **CONCLUSION**

### 🡺Usefulness for Stakeholders:

The project's outcomes will prove beneficial to stakeholders in several ways:

1. **Informed Decision-Making**: Stakeholders, such as HR departments and hiring managers, will gain insights into factors influencing student success and job market readiness. This knowledge can enhance decision-making in recruitment processes, ensuring better alignment of candidate skills with organizational needs.
2. **Process Improvement**: Understanding correlations between academic performance, skills (like Python programming), and career aspirations can lead to improvements in internship selection criteria and career counselling strategies. This can optimize resource allocation and enhance student career outcomes.
3. **Strategic Planning**: Insights into event attraction trends, leadership skills impact, and salary expectations based on various factors (GPA, family income) can guide strategic planning for educational institutions and companies alike. It helps in tailoring programs and offerings that meet current market demands.
4. **Competitive Advantage**: Leveraging data-driven insights can provide a competitive advantage in attracting and retaining top talent. Organizations can refine their internship programs and recruitment strategies based on empirical evidence rather than assumptions.

### 🡺Future Scope:

1. **Predictive Modeling**: Future iterations could involve developing predictive models to forecast job market trends, student outcomes, and career trajectories based on current data trends. This would require more advanced machine learning techniques and longitudinal data collection.
2. **Extended Data Sources**: Integrating additional datasets, such as alumni career paths and industry trends, could provide a more comprehensive understanding of factors influencing long-term career success beyond initial internship placements.
3. **Enhanced Visualization and Reporting**: Further refinement of visualization techniques and reporting formats could improve accessibility and usability of insights for diverse stakeholders, facilitating easier interpretation and decision-making.
4. **Continuous Feedback Loop**: Establishing a feedback mechanism to incorporate stakeholder input and refine analysis methodologies ensures ongoing relevance and applicability of findings to evolving educational and job market landscapes.

# **APPENDICES**

## Appendix A – Data Entity Visualization



