Multivariate Analysis of Career Preferences:
Insights from Kolkata's College Students

[Project Code: OP07]

Report submitted by:

SAMYABRATA ROY[1A*]

RISHAV KUMAR BANERJEE [2A]

SAUNAK DATTA[3A]

For the successful completion of the internship as

Data Science Intern

[Tenure: July 2024 - October 2024]

Under the supervision of

Debashis Ghosh

Head, Technology Business Development & Incubation,
IDEAS - Institute of Data Engineering, Analytics and Science Foundation,
Technology Innovation Hub, Indian Statistical Institute, Kolkata

Report submitted to :

IDEAS - Institute of Data Engineering, Analytics and Science Foundation
Technology Innovation Hub
Indian Statistical Institute
Kolkata, West Bengal, India

[*] Corresponding author

[A] BSc. (Hons.) Statistics from Sister Nivedita University [2022-2025]

[1] Email: 22f2001443@ds.study.iitm.ac.in
[2] Email: rishavbanerjee2004@gmail.com
[3] Email: saunakdatta.17@gmail.com

Declaration by the Students

We hereby declare that the work which is being presented in this project entitled "Multivariate Analysis of Career Preferences: Insights from Kolkata's College Students" in fulfilment of the requirements for the completion of the internship as a "Data Science Intern" is an authentic record of our own work carried out during the period from 6th July 2024 to 30th October 2024 under the supervision of Mr. Debashis Ghosh, Head, Technology Business Development & Incubation, IDEAS - Institute of Data Engineering, Analytics and Science Foundation, Technology Innovation Hub, Indian Statistical Institute, Kolkata.

We further declare that no portion of the project or its data will be published without the Institute's or supervisor's permission. We have not previously applied for any other degree or award using the topics and findings described in my dissertation.

Samyabtrata Roy : Samyabrata Roy 24/12/24.

Rishav Kumar Banerjee : RKB 24/12/24.

Saunak Datta : S Datta 24/12/24

(Signature of the students with Name and Date)

Acknowledgement

With immense pleasure and a deep sense of gratitude, we wish to express our sincere thanks to our supervisor, **Mr. Debashis Ghosh, Head, Technology Business Development & Incubation, IDEAS - Institute of Data Engineering, Analytics and Science Foundation, Technology Innovation Hub, Indian Statistical Institute, Kolkata**, for his motivation and continuous encouragement. Without his support, this project would not have been successfully completed.

We are grateful to the **Chief Executive Officer, Mr. Agnimitra Biswas** for providing us the facility to carry out this project at their institute.

We would like to convey a special note of thanks to **Mr. Diptendu Dutta**, for his insightful sessions, which played a crucial role in shaping the direction of this project. His willingness to help truly made a difference.

We also wish to extend my profound sense of gratitude to our parents for providing us with the moral support and encouragement whenever required.

Samyabrata Roy
Rishav Kumar Banerjee
Saunak Datta

Place: Kolkata
Date: 24/12/2024

Abstract

This project investigates the factors shaping academic and career preferences of college students in Kolkata, a critical issue in India’s evolving education-to-employment landscape shaped by the Fourth Industrial Revolution. The study is based on a primary dataset of 186 valid survey responses, capturing demographic, academic, financial, familial, and perceptual attributes of students aged 18–25. Exploratory Data Analysis revealed a predominance of middle-income households, balanced gender representation, and strong clustering of students in science and technology disciplines. Bivariate and multivariate analyses highlighted clear associations between degree choice and expected salary, moderate intergenerational effects of parental occupation on sectoral preferences, and consistent academic performance across school levels. Multicollinearity among academic indicators was diagnosed through VIF and resolved via Principal Component Analysis, while non-normality in perceptual variables was addressed with non-parametric approaches such as Kruskal–Wallis and Dunn’s post-hoc tests. Findings demonstrate that structural academic trajectories and parental sectoral backgrounds exert greater influence on career expectations than gender or income, while external influences such as peers and social media remain relatively uniform across groups. The study thus underscores the need for data-informed counselling and policy frameworks that strengthen student agency in academic decision-making and reduce institutional imbalances.

Keywords: Career Preferences, Multivariate Analysis, Non-parametric Analysis, Academic Discipline, Higher Education

Contents

Particular **Page No.**

Problem Statement

In India's rapidly evolving job market shaped by the **Fourth Industrial Revolution**, students **aged 18 to 25** must make academic decisions with long-term career implications earlier than ever. In Kolkata, a city known for its educational heritage and socio-cultural diversity, this process is often influenced by complex and sometimes conflicting factors such as parental expectations, peer influence, and perceived job market trends. While the **National Education Policy (NEP) 2020** expands access and flexibility, the core motivations behind student choices of academic disciplines remain underexplored. This study aims to address this gap by identifying the dominant factors—whether personal interest, societal pressure, or guidance—affecting students' subject selection in Kolkata, thereby informing more effective academic advising and policy-making.

Introduction

Career decision-making among young adults is becoming increasingly complex in the era of the Fourth Industrial Revolution, where automation, AI, and emerging technologies are rapidly reshaping the job market. In India, particularly among students aged 18 to 25, there is a growing urgency to align academic pathways with future employability, often under conditions of uncertainty and pressure. In metropolitan hubs like Kolkata—renowned for its intellectual legacy and socio-cultural plurality—the decision to pursue a particular academic discipline is rarely straightforward. Students must navigate a confluence of personal interest, family expectations, peer influence, and perceived job market trends, all while contending with evolving educational frameworks such as the National Education Policy (NEP) 2020. While the NEP has widened access and flexibility in higher education, the underlying drivers that shape student preferences for disciplines like science, commerce, humanities, or newer domains such as data science remain inadequately understood.

Existing research highlights the psychological and structural complexities of these choices. For example, a 2021 study on higher secondary students in Kolkata revealed that over two-thirds face substantial parental pressure during stream selection, often resulting in stress and decision fatigue. Complementing this, a 2024 study focusing on migrant graduates in the city found a strong link between academic mismatch and underemployment-induced anxiety. These findings point to an urgent need to systematically examine the real, dominant factors influencing subject choices in higher education. Understanding whether students are guided more by intrinsic interest, informed guidance, or socio-cultural coercion can provide valuable insights for designing better academic counselling frameworks and student-centred policy interventions.

In addressing questions of career choice and preference, methodological considerations play a critical role. Raw survey data, particularly when derived from self-reported responses, often contain inconsistencies, outliers, and overlapping variables. An initial step in such research typically involves Exploratory Data Analysis (EDA), which can provide insights into the structure of the dataset, highlight skewness in distributions, and uncover hidden patterns. Techniques such as histograms, density plots, and boxplots may assist in visualising variation, while correlation matrices and scatterplots could reveal the strength of associations or the potential presence of multicollinearity. These exploratory stages are not ends in themselves but serve as the groundwork upon which more advanced analyses are built.

When multiple factors are considered simultaneously, one recurring challenge is multicollinearity, where predictor variables are strongly correlated with each other. This situation can inflate standard errors in regression models, making it difficult to isolate the contribution of individual variables. In such contexts, approaches like the Variance Inflation Factor (VIF) may be used to diagnose the severity of the issue, while dimension reduction techniques such as Principal Component Analysis (PCA) can help condense redundant information into fewer uncorrelated components. By doing so, researchers maintain interpretability while reducing the instability that collinearity introduces.

Another layer of complexity arises from the presence of outliers and non-normal data distributions, particularly when responses are collected using ordinal measures like Likert scales. Conventional parametric tests often assume normality and equal variances, assumptions that may not always hold in real-world educational or social science data. In these cases, non-parametric methods—for instance, the

Kruskal–Wallis test for group comparisons or Dunn's post-hoc analysis—provide more robust alternatives that do not depend heavily on distributional assumptions. Similarly, techniques like Levene's test may be applied to examine the homogeneity of variances across groups, guiding the choice between parametric and non-parametric frameworks.

By situating the analysis within this methodological landscape, such projects move beyond descriptive accounts and into structured, theory-driven inquiry. The motivation lies in ensuring that findings are not merely artefacts of noisy data but instead reflect stable patterns. Whether through outlier management, correlation analysis, or dimensionality reduction, the methodological emphasis is on striking a balance between rigour and interpretability. This orientation not only enhances the robustness of the insights but also underscores their potential relevance for policymakers, educators, and guidance practitioners seeking to better understand the determinants of career preferences.

Objectives

The primary objective of this study is to systematically investigate the factors shaping students' academic and career-related choices and to derive actionable insights that can inform both counselling practices and policy interventions. Specifically, the project pursues the following aims:

➢ **Analysis of Association and Independence:**

To examine the interrelationships among parameters and understand how one factor may influence or remain independent of another. By applying statistical tests of association and visualisation methods, the study seeks to identify which variables affect each other most strongly. Such analysis helps to diagnose underlying dependencies, offering insights into how external pressures or personal choices interact to shape academic and career decisions.

➢ **Ranking of Influencing Factors:**

To identify and rank the relative importance of socio-demographic, academic, financial, and perceptual variables in shaping student preferences. Feature-ranking approaches provide a systematic way of prioritising factors, highlighting those with the greatest impact. This ranking can support targeted interventions by educators, parents, and policymakers, ensuring focus on the most decisive drivers of student outcomes.

➢ **Prediction of Academic Pathways:**

To explore predictive modelling techniques that can forecast students' degree and subject choices based on other parameters. Such models serve as diagnostic tools, capable of signalling potential mismatches between aspirations and academic choices, and guiding early interventions. Beyond prediction, this objective reflects the broader aim of developing frameworks that can help reduce dropout rates, minimise dissatisfaction, and support students in making decisions aligned with their genuine interests.

Taken together, these objectives emphasise not only descriptive insights but also predictive and diagnostic capacities. By ranking factors, building predictive models, and mapping associations, the study aspires to contribute toward reducing academic mismatch, minimising student dissatisfaction, and enabling young learners to make choices more aligned with their intrinsic interests rather than external pressures.

Data Collection

To carry out this investigation, primary data were collected from the field using a structured survey conducted in both online and offline modes simultaneously between **September 2024 and October 2024**. The survey was specifically designed to gather first-hand information from **students aged 18 to 25 pursuing undergraduate and postgraduate education in Kolkata.** The multi-modal survey approach ensured broader reach and inclusivity, capturing responses across diverse socio-economic and academic backgrounds within the Kolkata region.

➢ Determination of Parameters:

Based on the research objective, to understand the factors influencing students' academic discipline choices, these parameters were operationalised into survey questions to enable systematic analysis. The table below categorises each parameter with parameter group, type and description.

Parameter Group	Parameter	Type	Description
Demographic and Academic Profile	Age group	Categorical parameter	Respondent's age range (18–20, 20–22, 22–25)
	Gender	Categorical parameter	Gender identity (Male, Female, Others)
	Type of degree enrolled	Categorical parameter	Degree currently pursued (BA, BSc, BTech, etc.)
	Subject specialization	Textual Parameter [Attribute]	Subject field associated with the degree program
	Academic scores (Class X, Class XII, CGPA)	Ordinal/Numerical Parameters	Marks in Class X & XII (in percentage bands), CGPA scale from 0–10
Personal Interest & Alignment	Favourite subjects	Textual Parameter [Attribute]	Subjects students liked most in Class X and XII
	The profession the student wants to pursue	Textual Parameter [Attribute]	Career aspiration as written by the respondent
	Perceived alignment between degree and personal interest	Ordina Parameter (Likert)	0–5 scale on how well the degree matches personal interest
Financial and Family Background	Annual family income	Categorical parameter	Household income bracket in lakhs
	Financial constraint	Categorical parameter	Yes/No/Maybe

	perception		response on whether finances affected the choice
	Educational investment capacity	Numerical Parameter	Amount (in lakhs) the family can invest in student’s education
	Number of earning members	Numerical Parameter	Count of earners in the household
	Parent’s occupation	Textual Parameter	Open-ended field on parents’ job
	Parent’s qualification	Textual Parameter	Highest level of education completed by parents
External Influences	Parental influence	Ordina Parameter (Likert)	0–5 scale on how much parents influenced career choice
	Peer influence	Ordina Parameter (Likert)	Same as above, for friends/peers
	Teacher/Mentor influence	Ordina Parameter (Likert)	Influence of educators
	Socio-cultural influence	Ordina Parameter (Likert)	Influence of society/tradition/culture
	Social media influence	Ordina Parameter (Likert)	Impact of digital/social platforms on choices
Career Perception & Decision Confidence	Expected salary	Numerical Parameter	Anticipated annual income in Lakhs post-career establishment
	Preferred working sector	Textual Parameter	Respondent's desired industry or field
	Age of career settlement	Numerical Parameter	Age by which the respondent wishes to be career-settled

Table 1: Parameter Table

These parameters were drawn from relevant literature, including prior studies on academic stress and youth decision-making in Kolkata, as well as foundational theories of educational choice behaviour.

➢ Questionnaire Design:

The questionnaire was structured to capture both quantitative and qualitative responses through a mix of closed-ended and open-ended questions. It consisted of five main sections, each corresponding to the parameter groups identified earlier:

- Demographic and Academic Profile
- Personal Interest and Career Alignment
- Financial and Family Background
- External Influences
- Career Perception and Aspirations

A combination of categorical (e.g., gender, degree type), ordinal (e.g., Likert scales for parental influence), and ratio-type questions (e.g., expected salary) was used.

The full questionnaire is provided in Appendix I.

➢ Sampling Strategy:

The sample was drawn from the population of students aged 18 to 25 pursuing undergraduate and postgraduate education in Kolkata. A **Simple Random Sampling Without Replacement (SRSWOR)** technique was employed to ensure equal and independent selection of respondents, with no repetition. The survey was distributed across multiple educational institutions and online platforms to reach students from diverse academic streams, institutions, and socio-economic backgrounds, ensuring broad representation within the Kolkata region.

➢ Ethical Considerations:

To maintain ethical integrity throughout the research process, careful attention was given to protecting the rights and privacy of all participants. Although the questionnaire collected **personally identifiable information (PII) such as name and email ID** for the sake of completeness and authentication, this information was permanently removed during the data pre-processing stage. The final dataset used for analysis was fully anonymized, ensuring that no respondent could be personally identified.

In accordance with ethical research practices, explicit informed consent was obtained from each participant before submission. Respondents were asked:
“Do you consent to the use of the information provided in this form for analysis and research purposes? Your data will be handled confidentially and used solely for academic or research objectives.”

Only those who affirmatively provided consent were included in the study. Responses from participants who declined consent were deleted during the data pre-processing phase and were excluded from all stages of analysis.
This study is committed to the exclusive use of collected data for academic research purposes. No part of the dataset will be used for commercial, personal, or non-academic objectives under any circumstances.

Data Description

The final dataset used for further steps consists of **n = 186 valid responses**, drawn from an initial **192 survey submissions**. Responses from individuals who did not provide informed consent were excluded. To preserve respondent anonymity, **personally identifiable information (PII) such as name and email ID was removed before analysis**.

Each row in the dataset represents a unique participant, and each column corresponds to one of the **24 survey parameters**, organised under five thematic categories: Demographic and Academic Profile, Personal Interest and Alignment, Financial and Family Background, External Influences, and Career Perception & Decision Confidence.

The dataset was collected using **Google Forms**, stored in **Google Sheets**, and subsequently exported as a **CSV file** for analysis.

The raw dataset has not been published to avoid redundancy, as it primarily included unprocessed responses and formatting inconsistencies. Only the cleaned and standardised version, used for analysis, is provided in **Appendix II**.

Data Cleaning

Before conducting any statistical analysis, substantial data cleaning was carried out to ensure the accuracy, reliability, and consistency of the dataset. Given that the data was collected through self-reported survey responses, it contained various issues such as incomplete entries, inconsistent formatting, invalid or duplicate values, and missing responses. These anomalies had the potential to distort analysis and weaken the validity of insights.

A systematic cleaning process was implemented using **Python (primarily with Pandas, NumPy, and regex operations)** to detect, handle, and rectify such inconsistencies. The cleaning process involved multiple stages, including response filtering, column standardisation, datatype conversion, handling missing values, merging equivalent categories, and removing personally identifiable information. The cleaned dataset served as the basis for all further statistical and inferential procedures in this study.

The cleaning began with filtering responses based on consent. **Only those respondents who explicitly agreed to let their data be used for analysis were retained**. This step ensured compliance with ethical research practices and preserved the integrity of the study. Identifiers such as names, email addresses, and timestamps were also removed at this stage to protect respondent privacy and prevent any potential bias.

Next, column names were standardized and simplified for consistency. For example, long survey questions were renamed into concise labels such as "Percent_10" (percentage in Class 10) or "Family_income." This made the dataset easier to work with while ensuring clarity in subsequent analysis. Furthermore, the naming conventions were aligned with those presented in Section: Data Collection, Sub-section: Determination of Parameters (see Table 1). This ensured consistency between the conceptual framework of parameter selection and their operational use in the dataset, allowing the reader to easily cross-reference variables across sections.

Subsequently, categorical variables were cleaned and harmonized. A major effort went into standardizing responses where participants had used different spellings, abbreviations, or informal terms. For instance, college names such as "snu", "sister nivedita univ", and "Sister Nivedita University" were all mapped to a single uniform label. A similar process was followed for subjects of specialization, preferred working sectors, and intended professions. This harmonization was crucial to avoid treating equivalent categories as separate groups, which could otherwise distort frequency counts and statistical comparisons.

Numeric fields were also checked and corrected. Reported values like percentages, CGPA, expected salaries, or educational spending often appeared in inconsistent formats (ranges, words, or irregular expressions). These were systematically converted into numeric forms. For example, responses like "80–85%" were replaced with the midpoint (82.5), and phrases like "not a single penny more" were interpreted as zero spending. In addition, certain variables that were inherently numerical but had been collected as categorical ranges (e.g., family income brackets or age ranges) were converted into numerical entities by assigning the midpoint of the respective class. Such conversions ensured that all quantitative variables could be meaningfully analyzed and compared on a common scale.

In cases of missing or invalid responses, values were either treated as NaN or carefully recoded. For instance, vague entries like "NA," "nothing," or "engineering" (without further detail) in the specialization column were treated as missing, preventing misleading classifications.

Finally, the dataset was reorganized to maintain a logical structure, with related columns grouped together (e.g., academic performance, family background, personal choices, career preferences). This step not only improved readability but also facilitated smooth progression into the analysis phase.

Through this multi-step cleaning pipeline, the dataset was transformed from raw, inconsistent survey responses into a coherent, structured, and analysis-ready form. This foundation ensured that the patterns uncovered in subsequent statistical analysis reflect real insights rather than artifacts of messy data. The final cleaned dataset has been documented and can be found in **Appendix II** for reference.

Methodologies[1]

This project aims to identify the most dominant factors influencing students' academic discipline choices for higher education, rank those influences, and uncover patterns and correlations among various socio-demographic, academic, and perceptual variables. The analysis is based on a structured dataset obtained from a primary survey conducted among students aged 18 to 25 across institutions in Kolkata. Python was used as the primary analytical environment, incorporating libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data wrangling and visualisation; SciPy and Statsmodels for statistical testing, ANOVA, and regression modelling; Scikit-Posthocs for non-parametric post-hoc analysis; and Scikit-learn for dimensionality reduction through PCA.

Additionally, tools such as Pingouin were used for advanced statistical measures, while fuzzy matching, Google Sheets API, and other utility modules supported preprocessing and data extraction tasks. By applying a combination of descriptive statistics, inferential testing (e.g., ANOVA, Chi-square, Levene's test, Kruskal-Wallis), correlation analysis (e.g., Spearman's), and factor importance techniques, this methodology aims to derive interpretable, ranked insights into what drives students' academic and career decisions, ultimately contributing to research, guidance systems, and policy development.

➔ *Summary of the Dataset :*

i) Gender-wise Distribution:

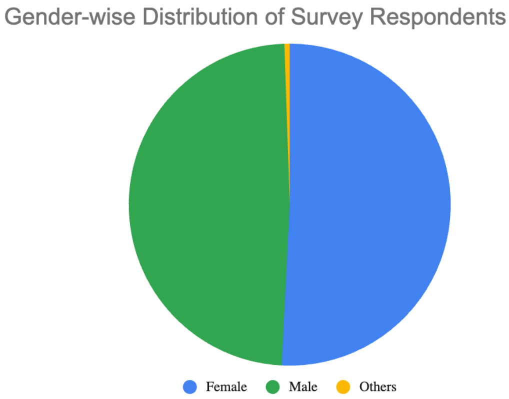


Diagram 1: Gender-wise Distribution of Survey Respondents

The survey responses include a fairly balanced distribution of gender identities, with Female and Male respondents each comprising nearly half of the total sample, and a very small percentage identifying as Others. This diversity adds value to the generalizability of gender-related insights in the study, with **Female: 50.8%, Male: 48.7%, and Others: 0.5%.**

ii) College-wise Distribution of Respondents:

College Name	Frequency
Sister Nivedita University	26
Institute of Engineering and Management	16
St Xaviers University	13

[1] For access to the datasets, code, and analysis details related to this project, visit the GitHub repository: GitHub Link

Jadavpur Universty	9
Techno International New Tow	8
Techno India University	7
St Xaviers College (Autonomous)	7
Presidency University	6
Maulana Azad College	6
Surendranath College	5
Asutosh College	5
University of Engineering and Management	4
The Bhawanipur Educational Society College	4
Amity University	4
University of Calcutta	4
Netaji Subhash Engineering College	3
Kusum Devi Sunderlal Dugar Jain Dental College	2
Sivnath Sastri College	2
Adamas University	2
The West Bengal National University Of Juridical Sciences	2
Acharya Jagadish Chandra Bose College	2
Gurudas College	2
St Pauls Cathedral Mission College	2
Tara Devi Harakh Chand Kankaria Jain College	2
Gokhale Memorial Girls College	2
Shri Shikshayatan College	2
Loreto College	2
Ramakrishna Mission Residential College	2
St Thomas College Of Engineering And Technology	2
Vivekananda College	2
Institute Of Business Management	2

Scottish Church College	2
Dr Sudhir Chandra Sur Institute Of Technology And Sports Complex	1
Calcutta Girls High College	1
IIEST Shibpur	1
Dayananda Sagar College Of Engineering	1
Narasinha Dutt College	1
Sister Nibedita Government General Degree College For Girls	1
The Heritage Academy	1
Indian Institute Of Science Education And Research Kolkata	1
Parul University	1
Jogesh Chandra Chowdhury Law College	1
Maharaja Manindra Chandra College	1
Swami Vivekanand Institute Of Science And Technology	1
Calcutta School Of Tropical Medicine	1
Seth Anandaram Jaipuria College	1
Vijaygarh Jyotish Ray College	1
Indian Institutes Of Technology Madras (Online Degree)	1
Heritage College	1
Dr Shyamaprasad Mukherjee College	1
Heritage Law College	1
NSHM Knowledge Campus	1
Coochbehar Government Engineering College	1
Govt College Of Engineering And Leather Technology	1
Brainware University	1
Jis University	1
Indian Association For The Cultivation Of Science	1
IPGMER	1
Vivekananda College Thakurpukur	1

Table 2: College-wise Distribution of Survey Respondents

The frequency distribution reveals that the 186 respondents in the dataset are spread across **58 different colleges**, highlighting the diverse institutional representation of the sample. While some colleges contributed a relatively larger number of responses, others were represented by only a handful of students. This broad coverage strengthens the study by incorporating perspectives from a variety of academic environments, reducing the risk of institutional bias. At the same time, the unequal distribution across colleges is something which may influence the generalizability of certain findings.

➔ *Exploratory Data Analysis :*

A comprehensive exploratory analysis of the dataset was performed using **YData Profiling**[2]. Rather than manually checking each column individually, YData Profiling generates an interactive dashboard-like report that provides a consolidated view of the dataset. This feature made it a more efficient and insightful choice compared to other options. The profile offers detailed summaries of variable distributions, correlations, and data quality checks.

From the analysis, certain aspects required special attention — notably the presence of extreme outliers and missing values in several numerical fields including: Excepted Salary, Age (by what age students are aiming to settle down) and Marks_10. To address the first issues i.e. potential presence of outliers, the Interquartile Range (IQR) method was applied, and values lying beyond the acceptable bounds were removed. This ensured that subsequent statistical models and interpretations were not disproportionately influenced by extreme data points.

Once the outlier removal was carried out using the IQR method, the dataset was reduced to a refined sample that differed from the original one. Since this essentially created a new sample, it was necessary to perform the exploratory data analysis again in order to reassess the distributions, correlations, and data quality of the updated dataset. This second round of EDA was duly conducted to ensure that the insights and patterns derived were reflective of the cleaned sample rather than the raw dataset.

After the outlier treatment, the final sample size was reduced from **186** to **172**. This adjustment was crucial for maintaining the robustness of the analysis, as it ensured that extreme values no longer distorted the results. The subsequent analyses, therefore, are based on this cleaned dataset.

i) Gender-wise Distribution:

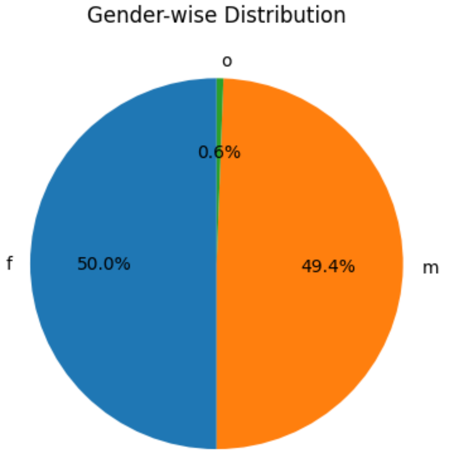


Diagram 2: Gender-wise Distribution of Survey Respondents (Post Outlier Removal)

[2] Given the interactive nature of the profile, the full **HTML report** is provided separately and can be accessed at https://github.com/samyaroy/Multivariate_Analysis_of_Career_Preference_in_Kolkata/blob/main/dataset/datasetCleanedDescription.html.

After removing outliers, the gender distribution of respondents remains fairly balanced, though with slight shifts in proportions. Female respondents now make up 50%, Male respondents account for 49.4%, and Others represent 0.6% of the sample. Despite these minor changes, the distribution continues to reflect a diverse set of gender identities, ensuring that gender-related insights drawn from the study remain broadly representative.

ii) Age-wise Distribution:

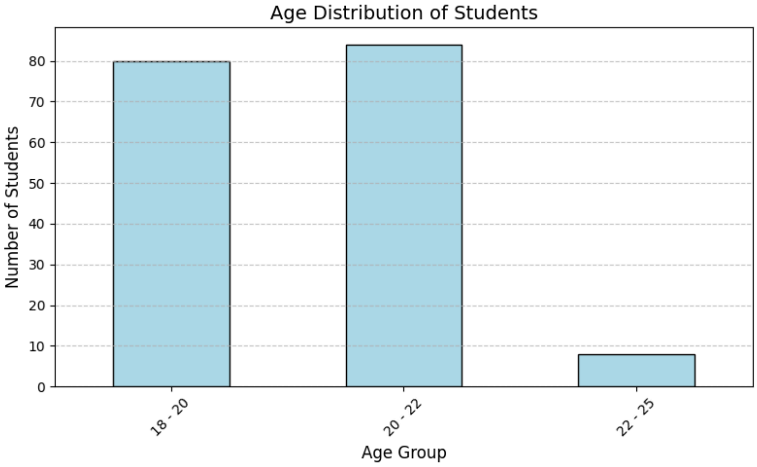


Diagram 3: Age-wise Distribution of Survey Respondents (Post Outlier Removal)

Following the removal of outliers, the age distribution of students is now concentrated across three distinct groups. The largest proportion of respondents fall within the **18–20** and **20–22** age brackets, together representing the majority of the sample. A comparatively smaller share of participants are in the **22–25** age group. This distribution reflects the typical age composition of undergraduate and early postgraduate students, which aligns well with the focus of the study.

iii) College-wise Distribution of Respondents:

College Names	Frequency
Sister Nivedita University	24
Institute of Engineering and Management	16
St Xaviers University	11
Jadavpur University	8
Techno International	7
Techno India University	7
Maulana Azad College	5
Asutosh College	5
Surendranath College	5
Presidency University	5
St Xaviers College (Autonomous)	5
University of Engineering and Management	4
Amity University	4
University of Calcutta	4
The Bhawanipur Educational Society College	4
Netaji Subhash Engineering College	3
Sivnath Sastri College	2

Tara Devi Harakh Chand Kankaria Jain College	2
Kusum Devi Sunderlal Dugar Jain Dental College	2
Acharya Jagadish Chandra Bose College	2
Vivekananda College	2
Scottish Church College	2
Gokhale Memorial Girls College	2
Shri Shikshayatan College	2
Adamas University	2
Gurudas College	2
St Thomas College of Engineering and Technology	2
St Pauls Cathedral Mission College	2
Ramakrishna Mission Residential College	2
Narasinha Dutt College	1
IIEST Shibpur	1
Dr Sudhir Chandra Sur Institute Of Technology And Sports Complex	1
Loreto College	1
Calcutta Girls High College	1
Dayananda Sagar College of Engineering	1
Parul University	1
The Heritage Academy	1
Vijaygarh Jyotish Ray College	1
Indian Institute of Science Education And Research Kolkata	1
Maharaja Manindra Chandra College	1
The West Bengal National University of Juridical Sciences	1
Jogesh Chandra Chowdhury Law College	1
Heritage College	1
Swami Vivekanand Institute Of Science And Technology	1
Calcutta School of Tropical Medicine	1
Seth Anandaram Jaipuria	1
Indian Institutes of Technology Madras (Online Degree)	1
Govt College of Engineering and Leather Technology	1
Heritage Law College	1
Dr Shyamaprasad Mukherjee College	1
Nshm Knowledge Campus	1
Coochbehar Government Engineering College	1
Institute of Business Management	1
Brainware University	1
JIS University	1
Indian Association For The Cultivation of Science	1
IPGMER	1
Vivekananda College Thakurpukur	1

Table 2: College-wise Distribution of Survey Respondents (Post Outlier Removal)

The updated frequency distribution, obtained after dropping outliers, shows that the **172 valid responses** are now distributed across **57 colleges**. Although the overall diversity of institutional representation remains evident, the reduction in sample size has slightly altered the balance of responses. A few colleges continue to contribute a comparatively larger share, while many others are represented by only a small number of students. This adjustment still ensures coverage from a wide range of academic settings, though the unequal distribution across institutions becomes more pronounced after outlier removal, which may have an impact on the generalizability of certain results.

iv) FamilyIncome Distribution across the Students:

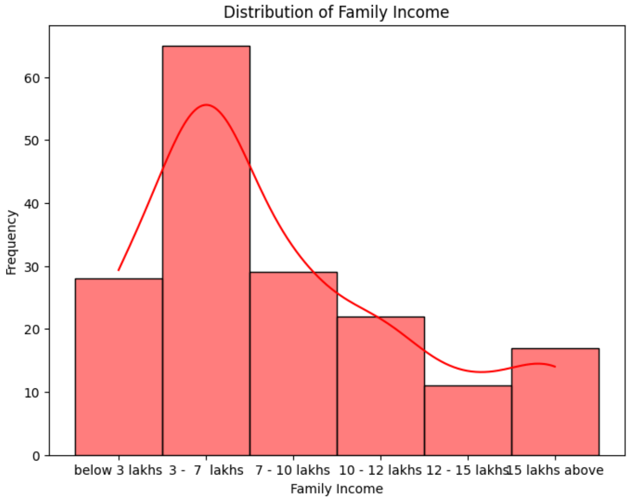


Diagram 4: Family Income Distribution of Survey Respondents (Post Outlier Removal)

The distribution of family income among the respondents shows a clear concentration in the middle-income brackets. The largest proportion of students belong to families earning between 3–7 lakhs annually, followed by those in the 7–10 lakhs range. A notable share of respondents also comes from families earning below 3 lakhs, while comparatively fewer students reported family incomes above 12 lakhs. This pattern highlights a predominance of middle-income households in the sample, with representation gradually declining as income levels increase.

iv) Distribution of Students across Degrees:

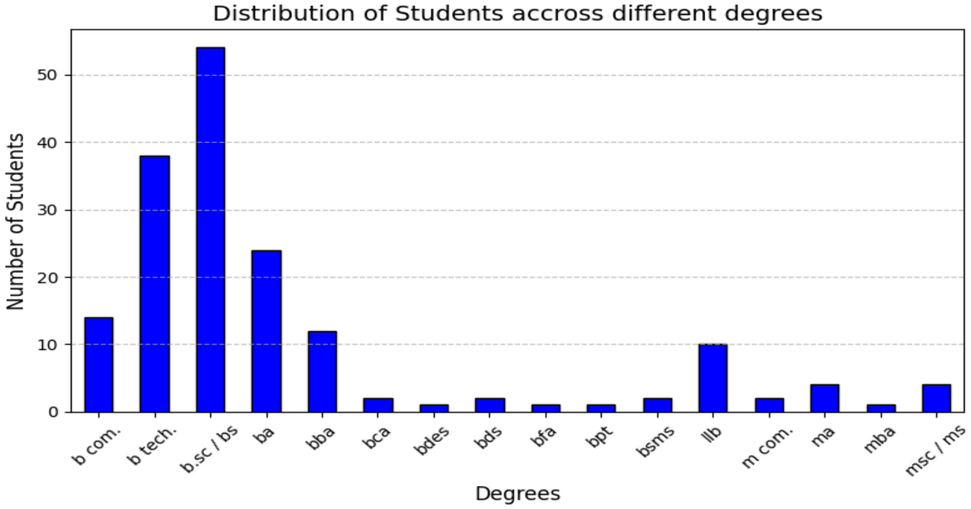


Diagram 5: Distribution of Students across Degrees (Post Outlier Removal)

The degree-wise distribution of students indicates that the largest share of respondents are enrolled in

B.Sc./BS programs, followed by B.Tech and BA students. Moderate representation is observed from B.Com and BBA, while other degrees such as BCA, B.Des, BDS, BFA, BPT, BSMS, LLB, M.Com, MA, MBA, and MSc/MS have comparatively fewer respondents. This pattern suggests that the sample is predominantly composed of students from science and technology backgrounds, with a smaller yet diverse representation from commerce, management, law, and other disciplines.

v) Distribution of Parent's Occupation of the Students across Sectors:

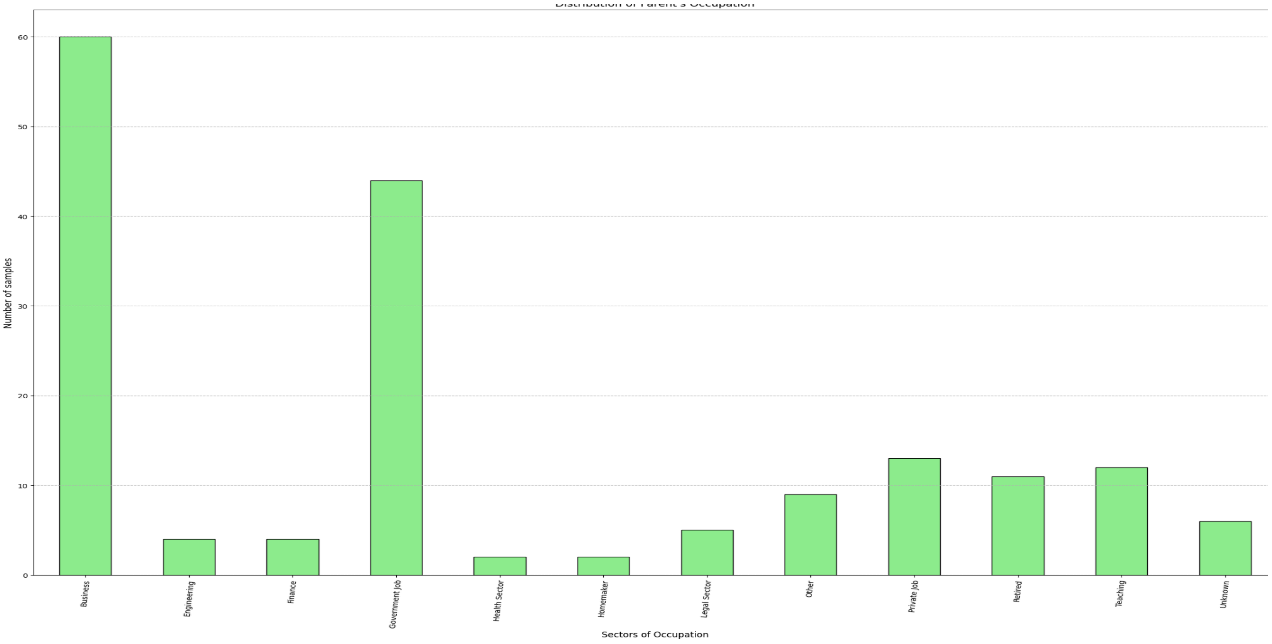


Diagram 6: Distribution of Parent's Occupation

The occupation-wise distribution reveals that the largest proportion of respondents' parents are engaged in Business, followed by those employed in Government jobs. Private jobs, teaching, and retired categories show moderate representation. Other sectors such as engineering, finance, legal sector, healthcare, homemaking, and miscellaneous occupations account for relatively fewer respondents. This distribution highlights that the sample is dominated by families with business and government service backgrounds, while still reflecting a diverse range of occupational sectors at smaller scales.

vi) Correlation between Degree chosen by students and Students' expected salary:

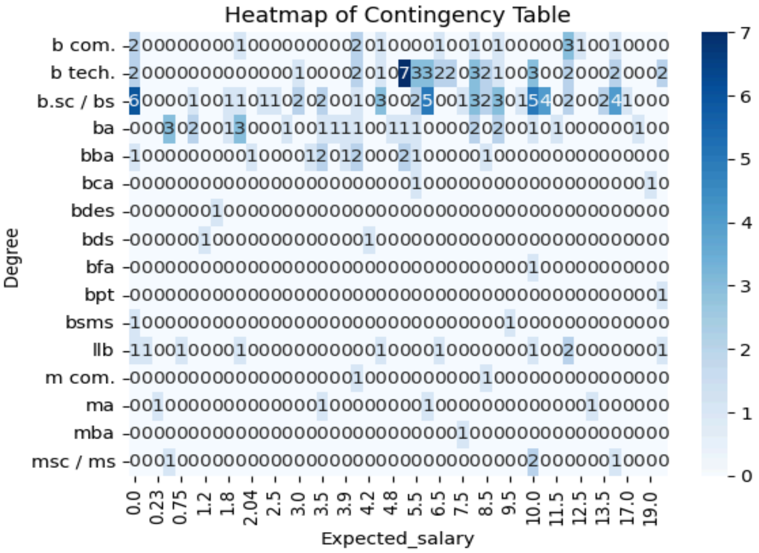


Diagram 7: Heatmap of Contingency Table (Degree vs. Expected Salary)

The contingency heatmap illustrates the relationship between students' degrees and their expected salaries. The highest concentrations are observed among B.Tech and B.Sc./BS students, particularly in the mid-range salary expectations, with notable peaks around 4.5 to 6.5 LPA. BA and B.Com students also show moderate representation across lower to mid salary ranges. Degrees such as BBA, BCA, B.Des, BDS, BFA, BPT, BSMS, LLB, M.Com, MA, MBA, and MSc/MS display comparatively sparse distributions, indicating fewer respondents in these categories with widely spread salary expectations. Overall, the heatmap reflects a dominance of science and technology students in shaping the salary expectation trends, with smaller but diverse contributions from other academic disciplines.

➔ *Inference :*

With the evidence and insights gathered from the Exploratory Data Analysis (EDA), the next step is to move forward with a more detailed analysis of the dataset. Before applying statistical methods, some data preparation is needed:

i) Data Preparation:

To enable the examination of relationships between variables, categorical data was systematically encoded into numeric formats to facilitate statistical testing. For instance, the *Gender* variable was transformed into a new feature, Gender_code, where male students were assigned a value of 1, female students a value of –1, and students identifying with other gender categories a value of 0. Similarly, responses to the question *"Do you feel financially constrained while choosing your career in higher studies?"* were encoded into the variable Financially_constrained_coded, with "Yes" represented as 1, "No" as 0, and "Maybe" as 2. These transformations ensured that categorical information could be meaningfully incorporated into subsequent analyses.

ii) Bivariate Analysis:

Bivariate analysis was conducted to investigate the relationships between pairs of variables in the dataset. This step helps to identify potential associations and dependencies that may influence students' career-related preferences. For categorical variables, the Chi-Square test of independence was applied to test associations, and where relevant, measures of association such as Cramér's V were used to assess the strength of these relationships. For continuous numerical variables, appropriate correlation coefficients were computed to evaluate the direction and magnitude of linear relationships. The following analyses present key findings from this stage.

A. Gender vs. Profession Aspired by Students:

Objective: To examine if gender biases exist in students' career aspirations.

Hypotheses for Chi-Square Test:

Null Hypothesis (H_{A0}): Gender does not influence the profession aspired to by students.

Alternative Hypothesis (H_{Aa}): Gender influences the profession aspired to by students.

Result:

- Chi-Square Statistic : 96.931
- p-value : 0.874
- Degrees of Freedom (df) : 114

Inference:

Since the p-value (0.874) is greater than 0.05, the null hypothesis cannot be rejected. This indicates that gender and profession aspired to are independent variables. In other words, there is no significant evidence of gender bias in career aspirations among the surveyed students.

B. Degree Chosen vs. Expected Salary as a Fresher:

Objective: To explore the relationship between the degree chosen by students and their expected starting salary.

Hypotheses for Chi-Square Test:

Null Hypothesis (H_{B0}): The degree chosen by students is independent of their expected starting salary.

Alternative Hypothesis (H_{Ba}): The degree chosen by students is dependent on their expected starting salary.

Result:

- Chi-Square Statistic: 1074.962
- p-value: 8.18×10^{-21}
- Degrees of Freedom (df): 675

To further assess the strength of this association, **Cramér's V test** was applied.

Hypotheses for Cramér's V:

Null Hypothesis (H_{Ba0}): The strength of association between the degree chosen and the expected starting salary is weak or negligible.

Alternative Hypothesis (H_{Ba1}): The strength of association between the degree chosen and the expected starting salary is moderate to strong.

Result:

Cramér's V Statistic: 0.645

Inference:

The p-value (8.18×10^{-21}) of the Chi-square test conducted is significantly less than 0.05, leading to the rejection of the null hypothesis. This indicates that the degree chosen by students and their expected starting salary are dependent variables. Furthermore, the Cramér's V statistic (0.645) demonstrates a strong association between the two variables, supporting the alternative hypothesis for Cramér's V.

C. Parents' Occupation vs. Profession Chosen by Students:

Objective: To determine whether the profession chosen by students is influenced by their parents' occupation.

Hypotheses for Chi-Square Test:

Null Hypothesis (H_{C0}): The profession chosen by students is independent of their parents' occupation.

Alternative Hypothesis (H_{Ca}): The profession chosen by students is not independent of their parents' occupation.

Result:

- Chi-Square Statistic: 673.721
- p-value: 0.096
- Degrees of Freedom (df): 627

Inference:

Since the p-value of the test is greater than 0.05, the null hypothesis cannot be rejected. This indicates that the profession chosen by students is independent of their parents' occupation. In other words, students' professional aspirations do not appear to be significantly shaped by parental occupation.

D. Preferred Sector vs. Parents' Occupation:

Objective: To validate the hypothesis that students' preferred work sector is influenced by their parents' occupation.

Hypotheses for Chi-Square Test:

Null Hypothesis (H_{D0}): The preferred sector of students is independent of their parents' occupation.

Alternative Hypothesis (H_{Da}): The preferred sector of students is dependent on their parents' occupation.

Result:

- Chi-Square Statistic: 294.448
- p-value: 1.51×10^{-12}
- Degrees of Freedom (df): 143

To measure the strength of this association, Cramér's V was computed.

Hypotheses for Cramér's V:

Null Hypothesis (H_{Da0}): The strength of the association between students' preferred sector and their parents' occupation is weak or negligible.

Alternative Hypothesis (H_{Da1}): The strength of the association between students' preferred sector and their parents' occupation is moderate to strong.

Result:

Cramér's V Statistic: 0.394

Inference:

The p-value (1.51 × 10^{-12}) is well below 0.05, leading to rejection of the null hypothesis for the Chi-Square test. This indicates that students' preferred work sectors are dependent on their parents' occupation. The Cramér's V statistic (0.394) suggests a moderate association, implying that while parental occupation does influence students' sectoral preferences, the relationship is not overwhelmingly strong.

E. **Family Income vs. Amount Students are Ready to Invest in Higher Education:**

Pearson's Correlation Coefficient (r): 0.331

Inference: There is a weak positive correlation, suggesting that as family income increases, students are slightly more willing to invest in higher education.

F. **Class 10 Average vs. Class 12 Average:**

Pearson's Correlation Coefficient (r): 0.475

Inference: A moderate positive correlation is observed, indicating that students who perform well in Class 10 tend to perform similarly in Class 12.

ii) Multivariate Analysis:

A key challenge in moving from bivariate to multivariate analysis is the problem of high dimensionality. When multiple predictors or explanatory variables are included in the dataset, there is an increased likelihood of multicollinearity, i.e., a situation where two or more variables are highly correlated with each other. Multicollinearity can distort statistical inferences by inflating the standard errors of coefficients, making it difficult to assess the true effect of individual predictors.

To diagnose the presence of multicollinearity in our dataset, we first computed the Variance Inflation Factor (VIF) scores across the numerical variables. The VIF quantifies how much the variance of an estimated regression coefficient is increased due to multicollinearity. Generally, a VIF value above 5 (or in stricter cases, above 10) is considered an indication of problematic multicollinearity. By examining the VIF values for the variables in this dataset, we establish whether corrective steps such as variable reduction or transformation are required before proceeding with further multivariate modeling.

A. **Examining VIF scores:**

Feature	VIF
Age_Group_code	95.98189
Gender_code	1.163013
college_code	6.123583

Percent_10	127.0164
Marks_10	40.64543
Percent_12	154.8993
Marks_12	39.96194
CGPA	9.56676
Expected_salary	3.365143
Spending	2.341092
parents_influence	8.771882
friend_influence	5.782541
mentor_influence	7.670473
social_expectations	6.210238
social_media_influence	5.774449
personal_choice	12.00577

Table 3: VIF scores among the numerical variables of the dataset

The results of the Variance Inflation Factor (VIF) analysis reveal the presence of significant multicollinearity among several variables. While most predictors such as Gender_code (VIF = 1.16), Expected_salary (VIF = 3.36), and Spending (VIF = 2.34) show acceptable levels of collinearity, certain variables exhibit extremely high VIF values. In particular, the cluster of academic performance indicators—Percent_10 (VIF = 127.02), Marks_10 (VIF = 40.65), Percent_12 (VIF = 154.90), and Marks_12 (VIF = 39.96)—stand out with VIF scores far exceeding conventional thresholds. This indicates severe redundancy and overlapping information among these variables, which poses a serious threat to the stability and interpretability of multivariate models.

To address this issue, a dimension reduction technique is warranted. Accordingly, Principal Component Analysis (PCA) is applied to this group of highly collinear academic variables in the next subsection. PCA will allow us to retain the underlying information while reducing redundancy, thereby ensuring a more reliable basis for subsequent multivariate modeling.

B. PCA on academic variables:

To address the issue of strong multicollinearity identified among the academic performance variables (Percent_10, Marks_10, Percent_12, and Marks_12), Principal Component Analysis (PCA) was applied. PCA transformed these four highly correlated features into two uncorrelated components, thereby reducing redundancy while preserving most of the underlying information. The resulting components, labeled Marks1 and Marks2, were retained as they jointly explained over 80% of the total variance in the original features. This dimensionality reduction ensured that the essential information from students' academic records was preserved, while mitigating the risks posed by multicollinearity in subsequent multivariate modeling. The explained variance ratio of these components is illustrated in the diagram below.

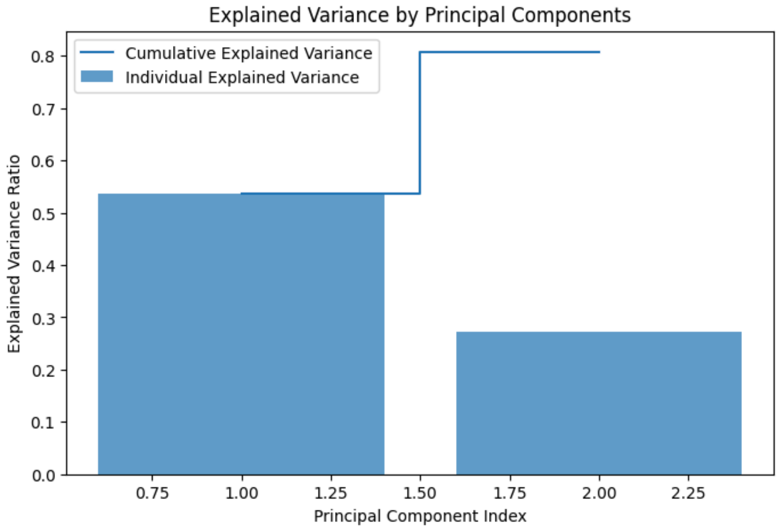


Diagram 8: Explained variance due to redundant columns by principal components

C. Correlation Analysis:

Following the dimensionality reduction step, a fresh correlation analysis was carried out on the dataset with the newly derived components (*Marks1* and *Marks2*) included in place of the original academic performance variables. This approach eliminates the redundancy and inflated correlations that were previously observed due to multicollinearity among *Percent_10, Marks_10, Percent_12,* and *Marks_12*. The updated correlation matrix now reflects more stable and reliable relationships across the variables, ensuring that subsequent multivariate models are not biased by overlapping information. By incorporating the PCA-transformed components, the analysis provides a clearer picture of how academic performance, alongside other socio-economic and personal factors, interacts to shape students' career-related decisions.

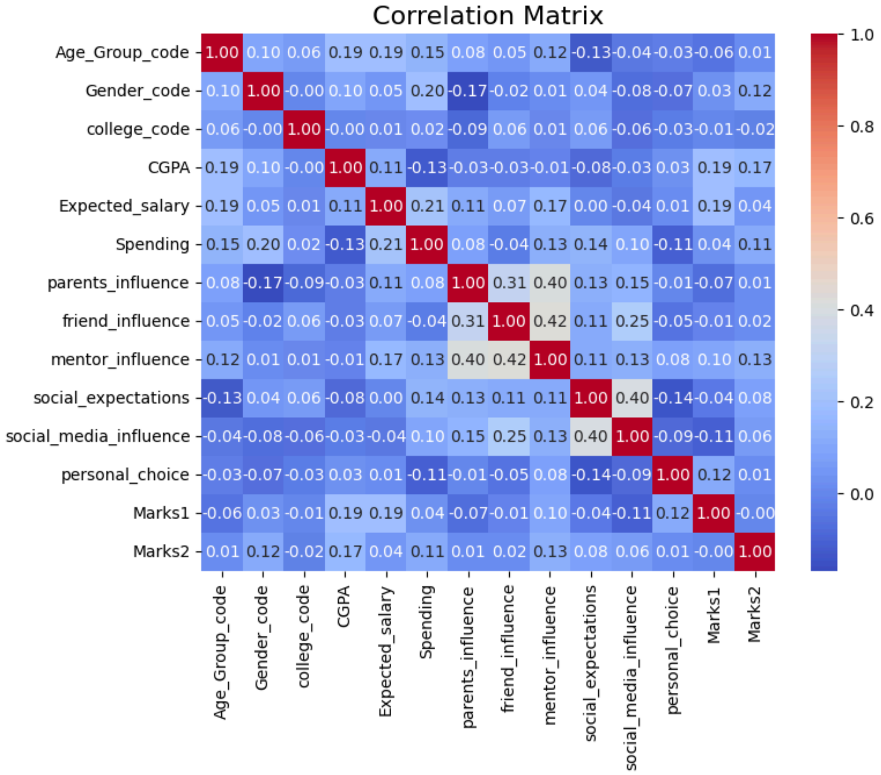


Diagram 9: Correlation Matrix

D. Inclusion of Likert Scale Analysis:

To extend the scope of the study, we included the evaluation of parameters measured using the Likert scale. These parameters capture students' perceptions and external influences on career decision-making, which go beyond socio-demographic and academic indicators. The chosen parameters for analysis were:

- Parents' Influence
- Friends' Influence
- Mentor's Influence
- Social Expectations
- Social Media Influence
- Personal Choice

Given the ordinal nature of Likert scale data, it was essential to test whether these variables follow a normal distribution before proceeding with advanced statistical modeling. Accordingly, two diagnostic approaches were applied:

1. Kolmogorov–Smirnov (KS) Test for Normality
2. Skewness Analysis

The KS test was used to formally assess the normality of each variable, while skewness values were calculated to further understand the symmetry of their distributions.

Hypotheses for the Kolmogorov–Smirnov Test [One-sample KS Test (Sample vs. Reference Distribution)]:

Null Hypothesis ($H_{\alpha 0}$): The sample comes from the specified reference distribution (i.e., the data follows a normal distribution).

$$H_0: F(x) = F_0(x), \text{ where } F_0(x) \sim N(\mu, \sigma^2)$$

Alternative Hypothesis ($H_{\alpha 1}$): The sample does not come from the specified reference distribution (i.e., the data does not follow a normal distribution).

$$H_1: F(x) \neq F_0(x)$$

Findings:

Parameter	KS Statistic	P-value	Skewness	Normality Rejected?
Parents' Influence	0.1769	3.54e-05	-0.4191	Yes
Friends' Influence	0.2275	2.65e-08	0.3995	Yes
Mentor's Influence	0.1719	6.44e-05	-0.0693	Yes
Social Expectations	0.1897	6.79e-06	0.3576	Yes
Social Media Influence	0.2470	9.87e-10	-0.9697	Yes
Personal Choice	0.2543	2.71e-10	-0.0835	Yes

Table 4: Kolmogorov–Smirnov Test Results

E. Group Analysis Based on Family Income:

To investigate whether students' perceptions and influences on career choices vary across different economic backgrounds, the dataset was spliced into six subgroups based on family income levels. This stratification allowed for a more structured comparison of responses across different socio-economic strata. The six groups represent households earning below ₹3 lakhs, ₹3–7 lakhs, ₹7–10 lakhs, ₹10–12 lakhs, ₹12–15 lakhs, and above ₹15 lakhs annually. Within each of these subgroups, students' ratings on the Likert-scale parameters: Parents' Influence, Friends' Influence, Mentor's Influence, Social Expectations, Social Media Influence, and Personal Choice—were retained for analysis. By slicing the data into these income-defined clusters, it becomes possible to evaluate whether economic background plays a significant role in shaping the strength or direction of these influences on students' career decision-making.

a. Testing Homogeneity of Variances (Levene's Test):

Before proceeding with group comparisons, it was necessary to assess whether the assumption of equal variances across groups was satisfied. Levene's Test was applied for this purpose.

Null Hypothesis ($H_{\beta 0}$): The variances of the subgroups are equal (homogeneity of variance).

$H_{\beta 0}$: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2$ **, where** σ_i^2 represents the variance of the i-th subgroup

Alternative Hypothesis ($H_{\beta a}$): At least one subgroup has a variance different from the others (heterogeneity of variance).

If the null hypothesis is rejected, it implies that not all income-based subgroups share the same variance, and thus homoscadasticity assumption can't be taken.

b. Kruskal–Wallis Test for Group Comparisons:

Since Likert scale data are ordinal in nature and homogeneity of variances may not always hold, the Kruskal–Wallis test, a non-parametric (assumption-free) alternative to one-way ANOVA, was employed to compare the distributions of the Likert-scale variables across the six income groups.

Null Hypothesis ($H_{\gamma 0}$): The distributions of the Likert-scale variable are the same across all income groups.

Alternative Hypothesis ($H_{\gamma a}$): At least one income group differs in its distribution of the Likert-scale variable.

This two-step approach—testing variance homogeneity first, and then applying the Kruskal–Wallis test—ensures that group-level differences in perceptions and influences are evaluated rigorously, while respecting the ordinal and non-normal nature of Likert scale data.

Findings:

Parameter	Levene's P-value	Kruskal-Wallis H Statistics	P Value of H Statistics	Conclusion
Parents' Influence	0.054383158798699704	3.9309217695599425	0.5594038625590363	Homogeneous but no significant differences
Friends' Influence	0.8781763904519573	0.7739475214711443	0.9786652063593133	Homogeneous but no significant differences
Mentor's Influence	0.23865267708633006	11.681621092909412	0.039420562593182024	Significant differences among the subgroups
Social Expectations	0.46217142321602656	3.054037537425287	0.6916549802590057	Homogeneous but no significant differences
Social Media Influence	0.10322488781436037	2.2181496394923412	0.8182100805476558	Homogeneous but no significant differences
Personal Choice	0.4512763171306099	7.404464096629299	0.1922551815222204	Homogeneous but no significant differences

Table 5: Results of Kruskal-Walis test done above

c. Post-Hoc Analysis Using Dunn's Test:

Objective: To identify specific pairwise differences among the six income-based subgroups after detecting a significant result in the Kruskal–Wallis H test.

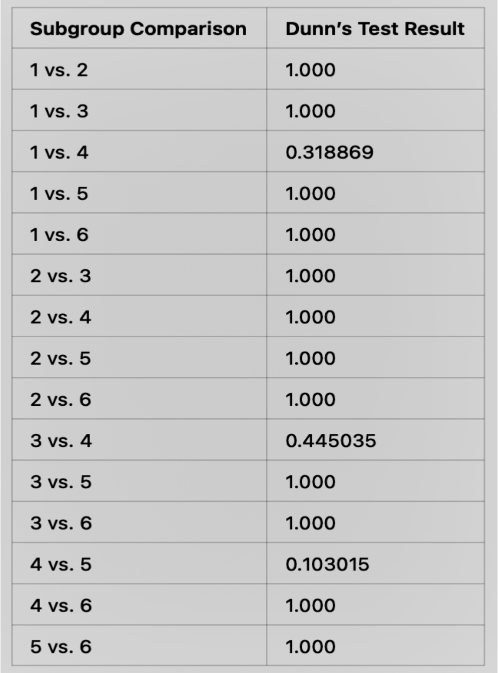
Hypotheses for Dunn's Test:

Null Hypothesis ($H_{\delta 0}$): There is no significant difference in the distribution of Likert-scale responses between the two income subgroups being compared.

Alternative Hypothesis ($H_{\delta a}$): There is a significant difference in the distribution of Likert-scale responses between the two income subgroups being compared.

Results:

The table below presents the pairwise comparisons between the six income groups, with the reported values representing the corresponding test statistics or p-values for each comparison.



Subgroup Comparison	Dunn's Test Result
1 vs. 2	1.000
1 vs. 3	1.000
1 vs. 4	0.318869
1 vs. 5	1.000
1 vs. 6	1.000
2 vs. 3	1.000
2 vs. 4	1.000
2 vs. 5	1.000
2 vs. 6	1.000
3 vs. 4	0.445035
3 vs. 5	1.000
3 vs. 6	1.000
4 vs. 5	0.103015
4 vs. 6	1.000
5 vs. 6	1.000

Table 6: Results of pair-wise Dunn's test (showing p-values only)

Findings:

The majority of the pairwise comparisons yielded p-values close to 1.000, indicating that most subgroup pairs did not exhibit statistically significant differences in their Likert-scale responses.

Comparisons involving Subgroup 4 (e.g., 1 vs. 4, 3 vs. 4, and 4 vs. 5) showed relatively lower p-values (e.g., 0.318869, 0.445035, and 0.103015, respectively). However, these values are still above the conventional significance threshold ($p < 0.05$).

Thus, although some subgroup comparisons hinted at moderate differences, the overall evidence suggests that no strong or statistically significant pairwise differences exist among the considered groups for the Likert-scale variables.

To summarize, the methodology employed in this study integrated a structured sequence of exploratory, bivariate, and multivariate analyses. Beginning with rigorous preprocessing steps such as outlier removal, encoding, and treatment of missing values, the dataset was refined to ensure robustness of subsequent results. The progression from EDA to bivariate associations (via χ^2 tests, Cramér's V, and correlation coefficients) allowed for the identification of key relationships between demographic, academic, and perceptual factors. Multicollinearity among academic performance indicators was systematically addressed through VIF diagnostics and PCA, ensuring dimensionality reduction without loss of substantive information. Finally, the inclusion of Likert-scale evaluations, followed by group-based comparisons across income subgroups using Levene's, K-W's H, and Dunn's tests, enabled a nuanced understanding of how personal perceptions and socio-economic contexts influence career-related choices. Together, these steps provide a comprehensive, multi-layered framework to uncover both direct and subtle drivers of students' academic and career decisions.

Findings

The findings of this study reveal several important insights into the factors shaping students' academic and career preferences. Beginning with the descriptive analysis, the sample exhibited a balanced gender distribution, with male and female respondents contributing almost equally, and a small fraction identifying as others. Age-wise, the majority of students fell within the 18–22 age group, reflecting the typical undergraduate population, while representation of older students was relatively limited. The college-wise distribution indicated broad institutional coverage across 57 colleges, though with noticeable concentration from certain institutions such as Sister Nivedita University and St. Xavier's University, which may have introduced some imbalance. Family background analysis revealed a predominance of middle-income households (₹3–7 lakhs annually) and parents working in business or government services, setting the socio-economic context in which career preferences are formed.

Building upon this descriptive backdrop, bivariate analysis offered deeper insights into relationships among key variables. First, the analysis of gender and profession aspired to found no significant association, suggesting that students' career aspirations are not shaped by gender bias in this sample. However, the degree chosen and expected starting salary were found to be strongly related: students from science and technology disciplines such as B.Sc. and B.Tech expected higher starting salaries compared to those from arts and commerce backgrounds. The strength of this relationship, confirmed through Cramér's V, underscores the role of academic discipline in shaping salary expectations.

Parental background also influenced certain career-related choices. While parents' occupation did not significantly affect the specific profession students aspired to, it did have a moderate association with the preferred sector of employment. For example, students whose parents were engaged in business or government services showed distinct patterns in their sectoral preferences, suggesting intergenerational transmission of career aspirations at the sector level rather than the occupation level.

Economic and academic indicators further added nuance. Family income was positively correlated with the amount students were willing to invest in higher education, though the strength of this relationship was weak, indicating that financial aspirations for education extend beyond household income levels. Academic performance showed stronger alignment, with a moderate correlation between Class 10 and Class 12 averages, suggesting consistency in students' academic achievement across school years. Interestingly, spending on higher education showed virtually no correlation with the number of colleges considered, highlighting that financial commitment and choice breadth are largely independent.

Together, these findings suggest that career and academic choices are shaped by a mix of educational background, economic resources, and parental influence, while gender exerts little effect in this context. The patterns highlight both structural constraints and individual agency, offering valuable implications for educators, policymakers, and career guidance professionals aiming to design equitable support systems for students.

n addition to these bivariate insights, the multivariate analysis provided further clarity. Variance Inflation Factor (VIF) testing revealed significant multicollinearity among academic performance indicators such as Class 10 and Class 12 marks and percentages. To address this, Principal Component Analysis (PCA) was applied, which successfully reduced these overlapping measures into two uncorrelated components that together explained over 80% of the variance. This ensured that the influence of academic performance could be analyzed without distortion, highlighting the consistency of students' academic achievement while minimizing redundancy.

The inclusion of perceptual variables measured through Likert scales (parents' influence, friends' influence, mentors' influence, social expectations, social media, and personal choice) added another dimension to the findings. Tests of normality (Kolmogorov–Smirnov) confirmed that these variables did not follow normal distributions, justifying the use of non-parametric approaches. When grouped by family income, the Kruskal–Wallis test revealed that most perceptual influences were consistent across income brackets, with the exception of mentors' influence, which showed statistically significant variation. However, post-hoc Dunn's tests suggested that these differences were not strong enough to indicate clear pairwise distinctions among specific income groups.

Taken together, the multivariate and group-level findings reinforce the earlier conclusions: structural academic pathways (such as chosen degree and consistent academic performance) and sectoral expectations (linked to parental background) are more decisive in shaping students' preferences than income or gender. External influences such as parents, peers, and social factors remain broadly uniform across socio-economic categories, with only slight variation in the role of mentors. These results underscore that while socio-economic context frames opportunities, students' career and academic decisions are more strongly driven by educational trajectories and perceived returns on specific disciplines.

Conclusions

The analysis of career preferences among college students in Kolkata highlights the central role of academic pathways in shaping economic expectations and sectoral aspirations. Rather than being primarily driven by demographic attributes such as gender or family income, students' choices appear to be structured by the degrees they pursue and the occupational backgrounds of their parents. This suggests that career trajectories are strongly embedded within institutional and familial contexts, reinforcing the view that education functions not only as a personal investment but also as a mechanism of social transmission.

At the same time, the study indicates that many commonly assumed determinants—such as gender-based bias or household financial capacity—may have less direct influence than often presumed. This is important because it reframes the discussion: the inequalities shaping career outcomes are not simply demographic but structural, linked to disciplinary opportunities, academic consistency, and sectoral signalling. In this sense, the findings invite a reconsideration of how guidance and counselling frameworks are designed. Instead of targeting broad demographic categories, interventions may be more effective if they focus on aligning students' intrinsic interests with the opportunities realistically associated with their chosen academic streams.

Although the study engages with a limited set of prior works, its contribution lies in empirically showing how academic performance trajectories and parental occupational influence interact to shape career expectations in an urban Indian setting. This insight extends the conversation on student decision-making beyond psychological stress or labour market mismatch, offering a perspective that foregrounds structural academic experiences. Future research with broader datasets and stronger integration with existing literature can refine these patterns further, but the present study offers a starting point for understanding how educational choices mediate the pathways from aspiration to opportunity.

Significance of the Project

The significance of this project lies in its ability to connect methodological objectives with concrete insights that can guide decision-making in education and policy.

➢ **Analysis of Association and Independence:**

One of the core aims was to explore how different factors relate to one another—whether they operate independently or influence each other. The results revealed clear patterns: while gender showed no significant link with career aspirations, academic discipline was strongly associated with salary expectations, and parental background shaped preferred sectors of employment. Such associations highlight where interventions may be most effective—for example, by addressing structural expectations around disciplines and salaries, or by recognising the subtle role of family influence in sectoral choices.

➢ **Ranking of Influencing Factors:**

The study also sought to prioritise the factors that matter most in shaping students' academic and career pathways. Analysis showed that structural variables—such as chosen degree, consistent academic performance, and parental occupation—carry more weight than demographic attributes like gender or family income. This ranking provides a sharper lens for counsellors and educators, allowing them to focus on the decisive drivers of student outcomes rather than dispersing attention across weaker influences.

➢ **Prediction of Academic Pathways:**

A further objective was to examine whether student trajectories could be predicted from the available parameters. The findings indicate that predictive modelling has strong potential, particularly when redundant academic variables are consolidated through techniques like PCA. The ability to forecast degree or subject choices can be extended into practical tools for diagnosing mismatches, reducing dropout tendencies, and supporting students in making choices that align with their intrinsic interests.

Beyond these methodological contributions, the project carries broader policy relevance. By clarifying associations, ranking key factors, and exploring predictive potential, it offers a foundation for creating better institutional infrastructure—one where students are systematically guided through evidence-based counselling. Policymakers could leverage such insights to design programmes ensuring that career advice comes from credible and experienced mentors within relevant fields, rather than from misaligned or uninformed sources. In doing so, the risk of students being swayed by unjustified external pressures could be mitigated, enabling them to pursue disciplines that match both their abilities and long-term aspirations.

Taken together, the project not only enriches academic understanding but also provides a roadmap for strengthening career guidance frameworks, reducing dissatisfaction, and fostering more informed, student-centred educational choices.

Limitations of the Project

1. Sample Size Constraints:

As the known population of college students aged 18 to 25 in Kolkata is certainly larger than 10,000, the required sample size for this study was estimated using the standard formula for large populations:

Required sample size $(n) \;=\; \frac{Z^2 . p . (1-p)}{e^2}$;

where $Z \;=\; 1.96$ (for a 95% confidence level),

$p \;=\; 0.5$ (maximum variability),

& $e \;=\; 0.5$ (margin of error)

This results in an ideal sample size of approximately **384**. However, due to practical constraints such as limited time and accessibility, the study proceeded with **187 valid responses**. **While sufficient for exploratory analysis, the reduced sample size may marginally affect the statistical generalizability of the findings to the wider population.**

2. Outlier Detection Approach:

In the *Exploratory Data Analysis* subsection under *Methodologies*, outliers were identified and removed using the **Interquartile Range (IQR)** method (applied across all numerical fields) and, in certain cases, the **Min–Max** approach. While this step was taken to ensure the reliability of subsequent analyses, it has two limitations. First, the removal of outliers may have inadvertently excluded valid data points, thereby altering the representativeness of the dataset. Second, given that the dataset is multivariate in nature, the IQR method may not be the most efficient approach for detecting outliers. More sophisticated methods, such as distance-based approaches (e.g., Euclidean distance from the mean) or the **Mahalanobis distance**, which accounts for correlations among variables, could have provided a more context-aware identification of multivariate outliers. Future work could consider these advanced techniques to enhance the robustness of the analysis.

3. Representation Imbalance in Subgroups:

The distribution of respondents across certain parameters shows imbalance that may affect representativeness. For instance, in the Age Group parameter, the majority of responses are concentrated in the first two categories (18–20 and 20–22 years), while the 22–25 years group is comparatively underrepresented. This may limit the generalizability of insights specific to older students.

Similarly, in the College-wise Distribution, most colleges contributed only 1–10 responses, whereas institutions such as Sister Nivedita University, St. Xavier's University, and the Institute of Engineering and Management had around 15 or more responses each. This uneven representation could introduce bias, as certain colleges disproportionately influence the dataset.

To mitigate these issues, several remedies could have been applied. At the data collection stage, **stratified sampling** (or quota-based sampling) would have ensured proportional or equal representation across both age and college groups. Post data collection, methods such as **re-sampling** (to balance subgroup sizes), **weighting** (to adjust the influence of underrepresented groups), and **normalization** (to standardize subgroup effects in statistical analysis) could have been employed. Incorporating these approaches would have improved the representativeness and reduced the potential bias in the study findings.

4. Missing Values Handling Approach:

In cases where missing values occurred in the dataset, the approach adopted was primarily simplistic: either replacing the missing entry with a reported placeholder value (zero in most of the cases) or, in many cases, dropping the entire row containing the missing field. While this ensured that the dataset remained usable for subsequent analysis, it may have reduced the overall efficiency of the study by discarding potentially valuable data. A more robust approach would have been to apply techniques such as interpolation, statistical imputation, or estimation based on similar cases. These methods would have preserved a larger portion of the dataset and likely enhanced the reliability of the results.

5. Absence of Predictive Modeling:

The study primarily focused on extracting insights through exploratory, bivariate, and multivariate analyses rather than building a predictive model. Although the dataset contained features suitable for predictive modeling—such as using academic, socio-economic, and perceptual factors to predict outcomes like Degree or Subject—the emphasis was placed on understanding underlying relationships and patterns instead. While this provided valuable in-depth insights, it also represents a limitation in that no predictive framework was developed. A future extension of this work could involve training predictive models (e.g., logistic regression, decision trees, or machine learning approaches) to forecast career choices, thereby complementing the interpretive insights with practical predictive utility.

6. Limited Literature Benchmarking:

While the study references selected prior works, it does not undertake an extensive literature benchmarking. As such, its contribution is primarily empirical, and future research could strengthen the theoretical positioning by situating these findings within broader national and international studies on career decision-making.

References

[1] W. J. Conover, **Practical Nonparametric Statistics**, *3rd ed., Hoboken, NJ: Wiley, 1999.*

[2] P. Sprent and N. C. Smeeton, **Applied Nonparametric Statistical Methods,** *4th ed., Boca Raton, FL: Chapman and Hall/CRC, 2007.*

[3] Biswas, M.M., Das, K.C. & Sheikh, I. **Psychological implications of unemployment among higher educated migrant youth in Kolkata City, India.** *Sci Rep 14, 10171 (2024). Available: https://doi.org/10.1038/s41598-024-60958-y*

[4] P. Basu and A. Chakrabartty, **Coping stress among the students at 12th grade in higher secondary schools in Kolkata,** *Health Outlook, Vol. 1, no: 1, pp: 17-24, 2019.*

[5] M. T. Borchert, **Career Choice Factors of High School Students,** *M.S. thesis, Univ. of Wisconsin–Stout, Menomonie, WI, Dec. 2002. [Online].*
Available: http://www.uwstout.edu/lib/thesis/2002/2002borchertm.pdf

Appendix I

Career Preference Survey Questionnaire

IDEAS - Institute of Data Engineering, Analytics and Science Foundation
Technology Innovation Hub, Indian Statistical Institute, Kolkata

This survey is part of an internship project titled 'Multivariate Analysis of Career Preferences: Insights from Kolkata's College Students **[Project Code: OP07]**. Your responses will remain confidential and will only be used for academic purposes. Please answer all questions honestly. This should take about 5–10 minutes to complete.

Section 0A: Consent of the Responder:

Do you consent to the use of the information provided in this form for analysis and research purposes?:

[Your data will be handled confidentially and used solely for academic or research objectives]

□ Yes   □ No

Section 0B: Personal Identifiable Information (PII):

[This section is collected for the sake of completeness, will be deleted before any study done on the dataset]

Email : .... [Verified, Auto collected through Google form]

Name : ....

Section 1: Demographic Factors

Age Group : □ 18 - 20   □ 20-22   □ 22-25

Gender : □ Male   □ Female   □ Others

College Name: .... [Write full name of your college]

Degree : □ BA   □ BSc/BS/BCA   □ B Tech.

□ B Com.   □ BBA   □ LLB

□ MA   □ MSc/MS   □ MTech

□ M Com.   □ Diploma   □ Other

Subject of Specialisation : .... [With which subject you are pursuing the degree]

[e.g : Mathematics, Animation, English etc.]

*In case of BBA mention your specialisation if any, or mention "Management"

*In case of LLB write "Law"

*In case of BCA write "Computer Application"

Your current CGPA : □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9 □ 10

Marks obtained in Class X: □ 90 - 100   □ 80 - 90   □ 70 - 80

[in percentage] □ 60-70   □ 50 - 60   □ below 50

Marks obtained in Class XII: □ 90 - 100   □ 80 - 90   □ 70 - 80

[in percentage] □ 60-70   □ 50 - 60   □ below 50

Do you feel that your chosen degree aligns with your personal choice?

Not at all □ 0   □ 1   □ 2   □ 3   □ 4   □ 5   A lot

Section 2: Personal Interest & Alignment

Favourite Subject(s) up to Class X and Marks Obtained:
[out of 100]

……………………………….…. - ……………………………………

……………………………….…. - ……………………………………

Favourite Subject(s) up to Class XII and Marks Obtained:
[out of 100]

……………………………….…. - ……………………………………

……………………………….…. - ……………………………………

Profession you would like to pursue in future: …………………………………………………

Section 3: Financial and Family Background

Annual Family Income: □ below 3 □ 3 - 7 □ 7 - 10
[in lakhs] □10 - 12 □ 12 - 15 □ above 15

Amount your family can invest in your higher education: ………………………………… [in lakhs]

Parents' occupation: ……………………………………………………………

Do you feel financially constrained while choosing your career in higher studies?

□ Yes □ No □ Maybe

Section 4: Career Perception & Decision Confidence

Your expected salary: …………………………………………………………… [in Lakhs Per Annum]

By the age you want to settle in your career: …………………………………… [in years]

Your preferred Working Sector: ………………………………………………….[e.g. Banking sector, IT sector]

Section 5: External Influences

How much did your parents influence you towards choosing your career path?

Not at all □0 □1 □2 □3 □4 □5 A lot

How much did your friends/peers influence you towards choosing your career path?

Not at all □ 0 □ 1 □2 □3 □ 4 □ 5 A lot

How much did your teachers/mentors influence you towards choosing your career path?

Not at all □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 A lot

How much of your choice of career path has been affected by social or cultural expectations?

Not at all □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 A lot

How much has social media affected your career choices in recent days?

Not at all □ 0 □ 1 □ 2 □ 3 □ 4 □ 5 A lot

Appendix II

Dataset (cleaned)

	Age Group	Gender	College Name	Degree	Subject	% in 10	Favorite Sub 10	Marks 10	% in 12	Favorite Subject 12	Marks 12	CGPA	profession	preffed sect	Expected salary	Age	Financially constrained	Family income	Parents' occupation	Spending	Parents influence	Friend influence	Mentor Influence	Social expectations	Social media influence	Personal choice
0	18 - 20	f	institute of engineering and management	bba	analytics	85	biology	97	75	biology	99	8	data analy			0	maybe	3 - 7 lakh	Unknown		3	1	1	1	1	5
1	20 - 22	f	institute of engineering and management	bba	finance	85	mathemat cs	99	75	economic	80	9	financial services	governmer / public sector	3.5	28	maybe	10 - 12 lakhs	Governmen Job	10	4	1	3	4	2	2
2	18 - 20	f	asutosh college	b.sc / bs	statistics	85	mathemat cs	98	85	mathema cs	86	6	chartered accountan	undecided	8	25	maybe	3 - 7 lakh	Business	2.5	5	2	4	3	1	4
3	18 - 20	f	sister nivedita university	ba	political science	95	socialscie ce	94	85	politicals ence	87	7	entreprene rship	corporate/ ivate secto	2	0	yes	3 - 7 lakh	Business		4	5	5	5	5	4
4	20 - 22	f	sister nivedita university	msc / ms	psychology	65	maths	60	65	accounts	65	7	chartered accountan	it secrtor	0.5	27	maybe	3 - 7 lakh	Business	10	3	1	1	4	3	5
5	20 - 22	f	sister nivedita university	ma	english	75	english	81	75	english	87	6	chartered accountan	n.q.	6	28	no	3 - 7 lakh	Finance	1	5	5	5	4	4	5
6	20 - 22	f	sister nivedita university	ma	english	55	english	82	65	english	72	8	professor	academia/ ducation	3.5	28.5	maybe	below 3 lakhs	Business		1	1	1	5	1	5
7	20 - 22	f	sister nivedita university	ma	english	75	english	82	85	english	88	6	professor	it secrtor	13	25	no	below 3 lakhs	Business	2	5	1	2	3	2	4
8	20 - 22	f	sister nivedita university	ba	english	85	science	80	75	english	80	8	professor	corporate/ ivate secto	3.5	25	yes	3 - 7 lakh	Other	2	5	1	2	5	2	3

9	22 - 25	f	university of calcutta	m com.	accountancy	85	science	87	85	accountan y	92	8	commerci undecided	governme /public sector	8.5	0	no	3 - 7 lakh	Business		2	1	3	1	1	5
10	18 - 20	f	gokhale memoria girls college	ba	political science	75	science		65			6	entreprene rship	governme /public sector	0.5	21	yes	below 3 lakhs	Business	0.5	4	1	1	1	2	3
11	18 - 20	f	techno international	b tech.	computer science and business system	85	bengali	95	85	chemistry	89	7	software developm nt enginee	corporate/ ivate secto	5	30	yes	10 - 12 lakhs	Governmen Job	6.5	4	3	2	3	3	3
12	18 - 20	f	techno international	b tech.	computer science and business system	85	math	80	95	chemistry	98	7	entreprene rship	it secrtor	5	22	maybe	below 3 lakhs	Business	2	4	4	4	4	3	3
13	20 - 22	f	techno international	b tech.	computer science engineering	75	geography	80	75	chemistry	85	6	software developm nt enginee	it secrtor	5.5	0	maybe	10 - 12 lakhs	Business	10	5	3	5	5	5	1
14	18 - 20	f	techno international	b tech.	computer	85	mathemat cs	81	85	mathema cs	89	7	cyber security	it secrtor	6	30	yes	3 - 7 lakh	Governmen Job	4	3	3	5	3	5	5
15	20 - 22	f	sister nivedita university	b.sc / bs	mathematic	85	mathemat cs	91	85	mathema cs	93	8	chartered accountan	corporate/ ivate secto	1.8	23.5	maybe	3 - 7 lakh	Retired	3.5	1	1	2	1	1	5
16	18 - 20	f	tara devi harakh chand kankaria jain college	b.sc / bs	microbiolog	95	biology	100	95	psycholo y	100	8	forensic microbiol gist	governme /public sector	2	24.5	yes	7 - 10 lakhs	Private Job	1.5	5	1	5	4	2	5
17	20 - 22	f	sister nivedita university	b.sc / bs	statistics	85	mathemat cs	91	85	mathema cs	85	6	data analy	it secrtor	5.5	29	maybe	3 - 7 lakh	Business	5.5	5	1	5	3	2	4
18	18 - 20	f	sister nivedita university	ba	english	75	english	68	65	english	65	5	professor	corporate/ ivate secto	2	23	maybe	3 - 7 lakh	Business	7.5	5	1	4	4	5	4

19	18 - 20	f	adamas universit	b tech.	biotechnolo y	95	mathemat cs	99	85	mathema cs	92	9	biotech industry	healthcare sector	3	25	yes	below 3 lakhs	Legal Secto	4	4	2	3	1	2	4
20	20 - 22	f	dr sudhir chandr sur institute of technology and sports complex	b tech.	computer science engineering	85	science	94	85	biology	96	8	web developer	it secrtor	5	24	maybe	3 - 7 lakh	Governmen Job	4	1	3	4	1	1	1
21	18 - 20	f	shri shikshayatar college	ba	journalism and mass communica on	85	english	85	85	english	89	7	journalist	corporate/ ivate secto	2	30	maybe	3 - 7 lakh	Governmen Job	1	1	1	1	1	1	5
22	20 - 22	f	iiest shibpur	b tech.	electrical engineering	95	mathemat cs	100	95	maths	100	8	entrepren rship	corporate/ ivate secto	20	28	yes	15 lakhs above	Business	15	5	2	5	5	3	1
23	20 - 22	f	maulana azad college	b.sc / bs	statistics	95	mathemat cs	92	95	mathema cs	97	7	data analy	governme /public sector	10	25	maybe	3 - 7 lakh	Governmen Job		5	1	5	1	1	5
24	20 - 22	f	surendranath college	b.sc / bs	statistics	95	mathemat cs		95			8	data analy	it secrtor	3.5	30	maybe	below 3 lakhs	Homemake	2.5	3	3	4	4	4	5

For brevity, only the first 25 rows of the cleaned dataset are presented here. To access the complete dataset, please refer to the supplementary materials provided separately (see https://github.com/samyaroy/Multivariate_Analysis_of_Career_Preference_in_Kolkata/blob/main/dataset/datasetCleaned.csv).

Dataset (outliers-free)

	Age Group	Gender	College Name	Degree	Subject	% in 10	Favorite Sub 10	Marks 10	% ir 12	Favorite Subject 12	Mark 12	CGPA	profession	preferred sector	Expected salary	Age	Financially constrained	Family income	Parents' occupation	Spending	Parents influence	Friend influence	Mentor Influence	Social expectatior	Social media influence	Personal choice
0	18 - 20	f	institute of engineering and management	bba	analytics	85	biology	97	75	biology	99	8	data analyst			0	maybe	3 - 7 lakhs	Unknown		3	1	1	1	1	5
1	20 - 22	f	institute of engineering and management	bba	finance	85	mathematics	99	75	economics	80	9	financial services	government/public sector	3.5	28	maybe	10 - 12 lakhs	Government Job	10	4	1	3	4	2	2
2	18 - 20	f	asutosh college	b.sc / bs	statistics	85	mathematics	98	85	mathematics	86	6	chartered accountant	undecided	8	25	maybe	3 - 7 lakhs	Business	2.5	5	2	4	3	1	4
3	18 - 20	f	sister nivedita university	ba	political science	95	socialscience	94	85	political science	87	7	entrepreneurship	corporate/private sector	2	0	yes	3 - 7 lakhs	Business		4	5	5	5	5	4
4	20 - 22	f	sister nivedita university	msc / ms	psychology	65	maths	60	65	accounts	65	7	chartered accountant	it secrtor	0.5	27	maybe	3 - 7 lakhs	Business	10	3	1	1	4	3	5
5	20 - 22	f	sister nivedita university	ma	english	75	english	81	75	english	87	6	chartered accountant	n.q.	6	28	no	3 - 7 lakhs	Finance	1	5	5	5	4	4	5

6	20 - 22	f	sister nivedita university	ma	english	55	english	82	65	english	72	8	professor	academia/education	3.5	28.5	maybe	below 3 lakhs	Business		1	1	1	5	1	5
7	20 - 22	f	sister nivedita university	ma	english	75	english	82	85	english	88	6	professor	it secrtor	13	25	no	below 3 lakhs	Business	2	5	1	2	3	2	4
8	20 - 22	f	sister nivedita university	ba	english	85	science	80	75	english	80	8	professor	corporate/private sector	3.5	25	yes	3 - 7 lakhs	Other	2	5	1	2	5	2	3
9	22 - 25	f	university of calcutta	m com.	accountancy	85	science	87	85	accountancy	92	8	commercial undecideds	government/public sector	8.5	0	no	3 - 7 lakhs	Business		2	1	3	1	1	5
10	18 - 20	f	gokhale memorial girls college	ba	political science	75	science		65			6	entrepreneurship	government/public sector	0.5	21	yes	below 3 lakhs	Business	0.5	4	1	1	1	2	3
11	18 - 20	f	techno international	b tech.	computer science and business system	85	bengali	95	85	chemistry	89	7	software development engineer	corporate/private sector	5	30	yes	10 - 12 lakhs	Government Job	6.5	4	3	2	3	3	3
12	18 - 20	f	techno international	b tech.	computer science	85	math	80	95	chemistry	98	7	entrepreneurship	it secrtor	5	22	maybe	below 3 lakhs	Business	2	4	4	4	4	3	3

					and business system																					
13	20 - 22	f	techno international	b tech.	computer science engineering	75	geography	80	75	chemistry	85	6	software development engineer	it secrtor	5.5	0	maybe	10 - 12 lakhs	Business	10	5	3	5	5	5	1
14	18 - 20	f	techno international	b tech.	computer	85	mathematics	81	85	mathematics	89	7	cyber security	it secrtor	6	30	yes	3 - 7 lakhs	Government Job	4	3	3	5	3	5	5
15	20 - 22	f	sister nivedita university	b.sc / bs	mathematics	85	mathematics	91	85	mathematics	93	8	chartered accountant	corporate/private sector	1.8	23.5	maybe	3 - 7 lakhs	Retired	3.5	1	1	2	1	1	5
16	18 - 20	f	tara devi harakh chand kankaria jain college	b.sc / bs	microbiology	95	biology	100	95	psychology	100	8	forensic microbiologist	government/public sector	2	24.5	yes	7 - 10 lakhs	Private Job	1.5	5	1	5	4	2	5
17	20 - 22	f	sister nivedita university	b.sc / bs	statistics	85	mathematics	91	85	mathematics	85	6	data analyst	it secrtor	5.5	29	maybe	3 - 7 lakhs	Business	5.5	5	1	5	3	2	4
18	18 - 20	f	sister nivedita university	ba	english	75	english	68	65	english	65	5	professor	corporate/private sector	2	23	maybe	3 - 7 lakhs	Business	7.5	5	1	4	4	5	4

19	18 - 20	f	adamas university	b tech.	biotechnology	95	mathematics	99	85	mathematics	92	9	biotech industry	healthcare sector	3	25	yes	below 3 lakhs	Legal Sector	4	4	2	3	1	2	4
20	20 - 22	f	dr sudhir chandra sur institute of technology and sports complex	b tech.	computer science engineering	85	science	94	85	biology	96	8	web developer	it secrtor	5	24	maybe	3 - 7 lakhs	Government Job	4	1	3	4	1	1	1
21	18 - 20	f	shri shikshayatan college	ba	journalism and mass communication	85	english	85	85	english	89	7	journalist	corporate/private sector	2	30	maybe	3 - 7 lakhs	Government Job	1	1	1	1	1	1	5
22	20 - 22	f	iiest shibpur	b tech.	electrical engineering	95	mathematics	100	95	maths	100	8	entrepreneurship	corporate/private sector	20	28	yes	15 lakhs above	Business	15	5	2	5	5	3	1
23	20 - 22	f	maulana azad college	b.sc / bs	statistics	95	mathematics	92	95	mathematics	97	7	data analyst	government/public sector	10	25	maybe	3 - 7 lakhs	Government Job		5	1	5	1	1	5
24	20 - 22	f	surendranath college	b.sc / bs	statistics	95	mathematics		95			8	data analyst	it secrtor	3.5	30	maybe	below 3 lakhs	Homemaker	2.5	3	3	4	4	4	5

For brevity, only the first 25 rows of the outliers-free cleaned dataset are presented here. To access the complete dataset, please refer to the supplementary materials provided separately (see https://github.com/samyaroy/Multivariate_Analysis_of_Career_Preference_in_Kolkata/blob/main/dataset/datasetCleaned_outliersFree.csv).

Appendix III

Python Code[3]

Importing necessary dependencies:

```python
import gspread
from google.colab import auth
from google.auth import default
from google.colab import drive
from google.colab import drive

from fuzzywuzzy import process, fuzz

import re
import subprocess
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

import pingouin as pg
import shutil
import torch
from fpdf import FPDF
from ast import keyword

import statistics

import scipy.stats as stats
from scipy.stats import f_oneway
from scipy.stats import chi2_contingency
from scipy.stats import levene
from scipy.stats import kruskal
from scipy.stats import spearmanr
from scipy.stats import kstest

import scikit_posthocs as sp

import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.outliers_influence import variance_inflation_factor

from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler

from ydata_profiling import ProfileReport
```

Authorizing Google User & Accessing Google sheet:

```python
auth.authenticate_user()
```

[3] GitHub repository: GitHub Link

```
drive.mount('/content/drive')
creds, _ = default()
gs = gspread.authorize(creds)
worksheetPrime = gs.open('Questionnaire 1  (Responses)').sheet1
```

Data Loading & Data Preprocessing:

```
rows = worksheetPrime.get_all_values()
df = pd.DataFrame(rows).reset_index(drop=True)
df.columns = df.iloc[0]
df = df[1:].reset_index(drop=True)
df.drop(columns=['Timestamp', 'Email address','Name\n'], inplace=True)

df2=df[df['Do you consent to the use of the information provided in this form for
analysis and research purposes? Your data will be handled confidentially and used solely
for academic or research objectives'] == 'Yes, I consent'].reset_index()

df2=df2.drop("Do you consent to the use of the information provided in this form for
analysis and research purposes? Your data will be handled confidentially and used solely
for academic or research objectives. [ Put it on the top]", axis=1)


df2 = df2.rename(columns={
'Subject of Specialisation [With which subject you are pursing the degree]': 'Subject',
 'Your current CGPA?' :'CGPA',
 'Percentage obtained in Class X' : 'Percent_10',
 'Percentage obtained in Class XII' : 'Percent_12',
 'Favourite Subject(s) till Class X and marks obtained in that subject' : 'fav_marks_X',
 'Favourite Subject(s) till Class XII and marks obtained in that subject' :
'fav_marks_XII',
 'Which profession you would like to pursue in future ? (e.g., data scientist ,
modelling, entrepreneurship) ' : 'profession',
 'Annual Family Income (appx)' : 'Family_income',
 'How much can you invest on your in education ? (in lakhs)' : 'Spending',
 'Do you feel financially constrained while choosing your career in higher studies? ' :
'Financially constrained',
 'What is your expected salary (in Lakhs Per Annum) ' : 'Expected_salary',
 'By what age do you want to settle in your career?' : 'Age',
 'What is your preferred Working Sector?(E.g.private sector or public sector or IT
sector e.t.c)' : 'preffed sector',
 'Which profession you would like to pursue in future ? (e.g., data scientist ,
modelling, entrepreneurship) ' : 'profession'
})

df2 = df2.rename(columns={
 'On scale 1-5 what do you think about the influence from your friends towards you for
choosing your career path' : 'friend_influence',
 'On scale 1-5 what do you think about the influence from your teacher/mentor towards
you for choosing your career path' : 'mentor_influence',
 'On scale 1-5 what do you think your career choice has been affected by social or
cultural expectations' : 'social_expectations',
 'On a scale of 1-5 how much social media has affected your career choices in recent
days.' : 'social_media_influence',
 'On scale 1-5 what do you think about the influence from your parents towards you for
choosing your career path' : 'parents_influence'})
```

Data Cleaning Section:

- Target Column: College name

```
df2['College Name'] = df2['College Name'].str.lower().str.strip().apply(lambda x:
re.sub(r'[^A-Za-z0-9 ]+', '', x))
college_mapping = {
    'iem': 'institute of engineering and management',
    'techno international new town': 'techno international',
    'techno international newtown':'techno international',
    'calcutta university':'university of calcutta',
    'snu': 'sister nivedita university',
    'university of engineering  management':'university of engineering and management',
    'thk jain':'Tara Devi Harakh Chand Kankaria Jain College'.lower(),
    'fiem':'Future Institute of Engineering and Management'.lower(),
    "st. xavier's college, kolkata" : "st. xavier's college",
    "st xaviers college kolkata" : "st xaviers college",
    "st xaviers autonomous kolkata" : "st xaviers college",
    'st xaviers universitykolkata':"st xaviers university",
    'techno india':'techno india university',
    'besc':'the bhawanipur educational society college',
    'the bhawanipur educational society':'the bhawanipur educational society college',
    "the bhawanipore education society":"the bhawanipur educational society college",
    'techno main salt lake':'techno india university',
    'the bhawanipur college':'the bhawanipur educational society college',
    'drsudhir chandra sur institute of technology and sports complex':'dr sudhir chandra
sur institute of technology and sports complex',
    'auk':'amity university'
}
df2['College Name'] = df2['College Name'].map(college_mapping).fillna(df2['College
Name'])
df2['College Name'] = df2['College Name'].replace(r'.*institute of engineering and
management.*', 'institute of engineering and management', regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*university of calcutta.*',
'university of calcutta', regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*xaviers.*college.*', 'st xaviers
college (autonomous)', regex=True)
```

```
df2['College Name'] = df2['College Name'].replace(r'.*xaviers university.*', 'st xaviers university', regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*iem.*', 'institute of engineering and management', regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*amity.*', 'amity university', regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*sivanth sastri*', 'sivnath sastri college', regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*west bengal national university of juridical sciences*', 'the west bengal national university of juridical sciences', regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*university of engineering and management .*', 'university of engineering and management', regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*rkmrc.*', 'Ramakrishna Mission Residential College'.lower(), regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*presidency university.*', 'presidency university'.lower(), regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*thomas college of engineering.*', 'st thomas college of engineering and technology'.lower(), regex=True)
df2['College Name'] = df2['College Name'].replace(r'.*kusum devi.*', 'Kusum Devi Sunderlal Dugar Jain Dental College'.lower(), regex=True)
df2 = df2[df2['College Name'] != 'drop out']  #drop out students are not taking under population so do in sample.
```

- Target Column: Age, Gender & Subjects of Specilization

```
df2['Age'] = df2['Age'].astype(str)
df2['Gender'] = df2['Gender'].apply(lambda x: "M" if x == 'Male' else "F" if x == 'Female' else "O")

df2['Subject'] = df2['Subject'].str.lower().str.strip().apply(lambda x: re.sub(r'[^A-Za-z0-9 ]+', '', x))
df2['Subject'] = df2['Subject'].replace(r'.*statistic.*', 'Statistics'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*computer sc.*', 'computer science engineering', regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*cse.*', 'computer science engineering', regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*pharmacology.*', 'pharmacology'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*physics.*', 'Physics'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*microbiology .*', 'MicroBiology '.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*mathematics.*', 'Mathematics'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*marketing.*', 'Marketing'.lower(), regex=True)
```

```python
df2['Subject'] = df2['Subject'].replace(r'.*life sciences.*', 'Life Sciences'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*aiml.*', 'Computer Science - Artificial intelligence and Machine Learning'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*data science.*', 'Computer Science - Data Science '.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*psychology.*', 'Psychology'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*journalism and mass communication.*', 'Journalism and Mass communication'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*information technology.*', 'Information Technology'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*laboratory technology.*', 'Medical Laboratory Technology'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*electronics and communication.*', 'Electronics and Communication'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*electrical.*', 'Electrical Engineering'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*hr.*', 'Human Resource'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*history.*', 'History'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*geography.*', 'Geography'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*fashion.*', 'Fashion Designing'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*economic.*', 'Economics'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*data visualization.*', 'Data Visualization'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*computer application.*', 'Computer Application'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*chemical.*','Chemical Engineering'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*csbs.*', 'Computer Science and Business System'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*biomedical.*', 'Biomedical Engineering'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*hospital management.*', 'Hospital Management'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*political science.*', 'Political Science'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*english.*', 'English'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*food.*', 'Food Technology'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*analytic.*', 'Analytics'.lower(), regex=True)
```

```
df2['Subject'] = df2['Subject'].replace(r'.*animation.*', 'Animation'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*agricul.*','Agriculture'.lower(),regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*account.*', 'Accountancy'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*constitutional.*','Constitutional Matters'.lower(),regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*finance.*', 'Finance'.lower(),regex=True)
df2['Subject'] = df2['Subject'].replace(r'.*management.*', 'Management'.lower(), regex=True)
df2['Subject'] = df2['Subject'].replace(r'medical', np.nan)
df2['Subject'] = df2['Subject'].replace(r'engineering', np.nan)
df2['Subject'] = df2['Subject'].replace(r'na', np.nan)
df2['Subject'] = df2['Subject'].replace(r'business',  'Computer Science and Business System'.lower())
df2['Subject'] = df2['Subject'].replace(r'.*computer.*', 'computer science engineering')
df2['Subject'] = df2['Subject'].replace(r'.*instrumentation.*', 'Electronics and Instrumentation Engineering'.lower(), regex=True)
null_list = np.array(['nothing','bcom','llb','law','bachelor of business administration'])
for i in range(len(null_list)):
 df2['Subject'] = df2['Subject'].replace(null_list[i], np.nan,regex=True)
```

- Target Column: Spending

```
df2['Spending'] = df2['Spending'].astype(str)
df2['Spending'] = df2['Spending'].str.lower().str.strip()
df2['Spending'] = df2['Spending'].replace(r'Not a single penny more','0',regex=True)
df2['Spending'] = df2['Spending'].replace(r'.*crore.*','100',regex=True)
df2['Spending'] = df2['Spending'].apply(process_spend)
```

- Target Column: Degree & Expected Salary

```
df2['Degree '] = df2['Degree '].replace(r'B.Sc / BS / BCA', 'B.Sc / BS')
df2['Degree '] = df2['Degree '].str.lower().str.strip()
df2['Degree '] = df2['Degree '].replace(r'.*bfa.*','bfa',regex=True)
df2['Degree '] = df2['Degree '].replace(r'.*llb.*','llb',regex=True)
df2['Degree '] = df2['Degree '].replace(r'.*bsms.*','bsms',regex=True)
df2['Degree '] = df2['Degree '].replace(r'.*bs-ms.*','bsms',regex=True)

df2.loc[df2['Subject'].str.lower().str.strip() == 'computer application', 'Degree '] = 'bca'

df2['Expected_salary'] = df2['Expected_salary'].astype(str)
df2['Expected_salary'] = df2['Expected_salary'].apply(clean_and_process_expected_salary)
```

- Target Column: Percent 10 & Percent 12

```
df2['Percent_10'] = df2['Percent_10'].astype(str)
df2['Percent_10'] = df2['Percent_10'].apply(average_range)
```

```
df2['Percent_12'] = df2['Percent_12'].astype(str)
df2['Percent_12'] = df2['Percent_12'].apply(average_range)
```

- Target Column: fav_marks_X & fav_marks_XII
  [Converted to **Favorite_Subject_10, Marks_10** & **Favorite_Subject_12, Marks_12** respectively]

```
# Extract the subject and marks from the 'fav_marks_XII' column
df2[['Favorite_Subject_10', 'Marks_10']] =
df2['fav_marks_X'].str.extract(r'([A-Za-z]+)[\s:,\-(]+(\d+)')
df2[['Favorite_Subject_12', 'Marks_12']] =
df2['fav_marks_XII'].str.extract(r'([A-Za-z]+)[\s:,\-(]+(\d+)')

df2['Favorite_Subject_10'] = df2.apply(lambda row: row['fav_marks_X'] if
has_no_numbers(row['fav_marks_X']) else row['Favorite_Subject_10'], axis=1)

# Convert 'Marks_12' to numeric, if necessary
df2['Marks_10'] = pd.to_numeric(df2['Marks_10'], errors='coerce')
df2['Marks_12'] = pd.to_numeric(df2['Marks_12'], errors='coerce')

df2=df2.drop('fav_marks_X', axis=1)
df2=df2.drop('fav_marks_XII', axis=1)

Subject_mapping2 = {
   'maths': 'mathematics',
   'math':'mathematics',
   'computerscience': 'computer science',
   }

df2['Favorite_Subject_10'] =
df2['Favorite_Subject_10'].map(college_mapping).fillna(df2['Favorite_Subject_10'])
df2['Favorite_Subject_12'] =
df2['Favorite_Subject_12'].map(college_mapping).fillna(df2['Favorite_Subject_12'])
df2 = df2.applymap(lambda x: x.lower() if isinstance(x, str) else x)
```

- Target Column: Parents' occupation

```
df2['profession'] = df2['profession'].str.lower().str.strip().apply(lambda x:
re.sub(r'[^A-Za-z0-9 ]+', '', x))
df2['profession'] = df2['profession'].replace(r'.*data analyst.*', 'data analyst',
regex=True)
df2['profession'] = df2['profession'].replace(r'.*msc.*','academia', regex=True)
df2['profession'] = df2['profession'].replace(r'.*sde.*','software development
engineer')
```

```
df2['profession'] = df2['profession'].replace(r'.*entrepreneur.*', 'entrepreneurship', regex=True)
df2['profession'] = df2['profession'].replace(r'.*data sc.*', 'data scientist', regex=True)
df2['profession'] = df2['profession'].replace(r'.*commerce related.*', 'commercial jobs', regex=True)
df2['profession'] = df2['profession'].replace(r'.*software developer.*', 'software development engineer', regex=True)
#df2['profession'] = df2['profession'].replace(r'.*.*', '', regex=True)
df2['profession'] = df2['profession'].replace(r'.*hr.*', 'human resource', regex=True)
df2['profession'] = df2['profession'].replace(r'.*doctor.*', 'doctor', regex=True)
df2['profession'] = df2['profession'].replace(r'.*cyber.*', 'cyber security', regex=True)
df2['profession'] = df2['profession'].replace(r'.*business.*', 'entrepreneurship', regex=True)
df2['profession'] = df2['profession'].replace(r'.*administrative.*', 'administration', regex=True)
df2['profession'] = df2['profession'].replace(r'.*advoca.*', 'advocate', regex=True)
df2['profession'] = df2['profession'].replace(r'.*ai.*', 'artificial intelligence', regex=True)
df2['profession'] = df2['profession'].replace(r'.*sde.*', 'software development engineer', regex=True)
df2['profession'] = df2['profession'].replace(r'.*software eng.*', 'software development engineer', regex=True)
df2['profession'] = df2['profession'].replace(r'.*sports.*', 'sports', regex=True)
df2['profession'] = df2['profession'].replace(r'.*professor.*', 'professor', regex=True)
df2['profession'] = df2['profession'].replace(r'.*research.*', 'researcher', regex=True)
df2['profession'] = df2['profession'].replace(r'.*scientist.*', 'researcher', regex=True)
df2['profession'] = df2['profession'].replace(r'.*teach.*', 'academia', regex=True)
df2['profession'] = df2['profession'].replace(r'.*web.*', 'web developer', regex=True)
df2['profession'] = df2['profession'].replace(r'.*sports.*', 'sports', regex=True)
df2['profession'] = df2['profession'].replace(r'.*law.*', 'lawyer', regex=True)
df2['profession'] = df2['profession'].replace(r'.*advoc.*', 'lawyer', regex=True)
df2['profession'] = df2['profession'].replace(r'.*judiciary.*', 'judiciary', regex=True)
df2['profession'] = df2['profession'].replace(r'.*trading.*', 'entrepreneurship', regex=True)
df2['profession'] = df2['profession'].replace(r'.*ca.*', 'chartered accountant', regex=True)
df2['profession'] = df2['profession'].replace(r'.*ca.*', 'chartered accountant', regex=True)
df2['profession'] = df2['profession'].replace(r'.*ca.*', 'chartered accountant', regex=True)

df2['profession'] = df2['profession'].str.lower().str.strip()
```

```
df2['profession'] = df2['profession'].replace(r'.*yet.*' , 'undecided', regex=True)
df2['profession'] = df2['profession'].replace(r'.*not sure.*' , 'undecided', regex=True)
df2['profession'] = df2['profession'].replace(r'havent thought about it yet',
'undecided', regex=True)
df2['profession'] = df2['profession'].replace(r'teacher or banking job or data
scientist', 'undecided', regex=True)
df2['profession'] = df2['profession'].replace(r'any statistician related role',
'undecided', regex=True)
df2['profession'] = df2['profession'].replace(r'job', 'undecided', regex=True)
df2['profession'] = df2['profession'].replace(r'ias officer or researcher', 'undecided',
regex=True)
df2['profession'] = df2['profession'].replace(r'business startups or govt jobs',
'undecided', regex=True)
```

- Target Column: Preferred Sector

```
# Apply the mapping to the preferred sector column
#df2['mapped_sector'] = df2[column_name].map(sector_mapping)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*it.*', 'it secrtor',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*private.*', 'corporate/private
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*public.*', 'government/public
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*admin.*', 'adminstative
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*bio.*', 'healthcare sector',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*any.*' , 'undecided',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*banking.*', 'finance/banking
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*goverment.*',
'government/public sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*corporate.*',
'corporate/private sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*government.*',
'government/public sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*research.*', 'education
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*engineering.*', 'engineering
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*medical.*', 'healthcare
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*healthcare.*', 'healthcare
sector', regex=True)
```

```
df2['preffed sector'] = df2['preffed sector'].replace(r'.*hospital.*', 'healthcare
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*education.*',
'academia/education', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*university.*',
'academia/education', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*academi.*',
'academia/education', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*sports.*', 'sports sector',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*arts.*', 'creative fields',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*content.*', 'creative fields',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*agriculture.*', 'agriculture
sector', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*sales.*', 'customer care',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*watchman.*', 'others',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*home.*', 'others', regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*nil.*', 'undecided',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*both.*', 'undecided',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*na.*', 'undecided',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*not.*', 'undecided',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*customer.*', 'others',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*court.*', 'judiciary',
regex=True)
df2['preffed sector'] = df2['preffed sector'].replace(r'.*business.*',
'corporate/private sector', regex=True)
```

- Target Column: Family Income & Parents’ Occupation

```
df2['Family_income']=df2['Family_income'].astype(str)

def clean_occupation(occupation):
   if pd.isna(occupation) or occupation.strip() == "":
       return "Unknown"
   occupation = occupation.lower()
   if "business" in occupation or "entrepreneur" in occupation or "buisness" in
occupation or "self-employed" in occupation:
       return "Business"
   elif any(keyword in occupation for keyword in ["retired","ex-serviceman","ex"]):
       return "Retired"
   elif "service" in occupation or "government" in occupation or "railway" in occupation
or "sbi" in occupation or "defence" in occupation or "govt" in occupation:
       return "Government Job"
   elif "private" in occupation or "cesc" in occupation or "corporate" in occupation or
"sales manager" in occupation:
       return "Private Job"
   elif "teacher" in occupation or "professor" in occupation or "teaching" in occupation
or "education" in occupation:
       return "Teaching"
   elif "engineer" in occupation:
       return "Engineering"
   elif "lawyer" in occupation or "judge" in occupation or "legal" in occupation or
"advocacy" in occupation:
       return "Legal Sector"
   elif any(keyword in occupation for keyword in ["doctor","health","physiotherapist"]):
       return "Health Sector"
   elif "ca" in occupation or "finance" in occupation or "accountant" in occupation:
       return "Finance"
   elif any(keyword in occupation for keyword in ["homemaker","house wife", "housewife",
"home maker"]):
       return "Homemaker"
   elif any(keyword in occupation for keyword in ["tax"]):
       return "Finance"
   elif any(keyword in occupation for keyword in ["nothing", "nothing", "na", "nil",
"unknown", "no idea"]):
       return "Unknown"
   else:
       return "Other"

# Apply function to clean data
df2["Parent's occupation"] = df2["Parent's occupation"].apply(clean_occupation)
```

- Target Column: Age [by ‘Age’, the age chosen by the student they are planning to settle by is referred]

```
data = df2

def clean_age_column2(age):
 if (clean_age_column(age)) and (clean_age_column(age)>100):
   dt=[]
   dt.append(clean_age_column(age)/10)
   dt.append(clean_age_column(age)%10)
   age=statistics.mean(dt)
 else:
   return clean_age_column(age)
```

```
def clean_age_column(age):
    # Remove words and keep only numbers or ranges
    age = re.sub(r'[^\d\-]', '', str(age))
    # Handle ranges like "20-22"
    if '-' in age:
        try:
            start, end = map(int, age.split('-'))
            return (start + end) / 2  # Return average of the range
        except ValueError:
            return None

    # Handle single numbers
    elif age.isdigit():
        return int(age)
    else:
        return None

# Apply the function to the Age column
df2['Age'] = data['Age'].apply(clean_age_column2).fillna(0)
```

- Target Columns: all parameters collected in the Likert Scale

```
# Example: Converting a column named 'column_name' to integers
list5=['parents_influence', 'friend_influence',
'mentor_influence','social_expectations', 'social_media_influence','personal_choice']
for i in range(len(list5)):
  df2[list5[i]] = df2[list5[i]].astype(int)
```

- Rearranging the columns:

```
print(df2.columns)
df2['CGPA']=df2['CGPA'].astype(float)
# Specify the new column order
new_column_order = ['index','Age Group', 'Gender', 'College Name', 'Degree ',
'Subject', 'Percent_10', 'Favorite_Subject_10', 'Marks_10','Percent_12',
'Favorite_Subject_12', 'Marks_12', 'CGPA', 'co-curricular activity', 'profession',
'preffed sector', 'Expected_salary', 'Age', 'Financially constrained',
'Family_income', "Parents’ occupation", 'Spending', 'parents_influence',
'friend_influence', 'mentor_influence', 'social_expectations',
'social_media_influence', 'personal_choice']
# Reorder DataFrame columns
df2 = df2[new_column_order]
```

Data Exporting & Uploading to Google Drive

```
df2.to_csv('/content/drive/MyDrive/pdf files/refined.csv', index=False)
```

Performing EDA on Cleaned Dtaset

```
profile = ProfileReport(df2, title="Career Preferences Data Report")
profile.to_file("datasetCleanedDescription.html")
```

Deleting Outliers with Help of IQR Method

```
# Copy of the original dataframe for reference
df_clean = df2.copy()

# Create a DataFrame to store eliminated data points
outliers = pd.DataFrame()

# Define a function to identify outliers based on IQR
def detect_outliers_iqr(data, column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return data[(data[column] < lower_bound) | (data[column] > upper_bound)]

# Define a function to detect outliers in numerical ranges
def detect_outliers_range(data, column, min_value, max_value):
    return data[(data[column] < min_value) | (data[column] > max_value)]

# Define the columns and conditions for outlier detection
columns_conditions = {
    "Percent_10": {"method": "range", "min": 0, "max": 100},
    "Percent_12": {"method": "range", "min": 0, "max": 100},
    "CGPA": {"method": "range", "min": 0, "max": 10},
    "Expected_salary": {"method": "iqr"},
    "Spending": {"method": "iqr"},
}

# Process each column for outlier detection
for column, condition in columns_conditions.items():
    if condition["method"] == "range":
        outliers_detected = detect_outliers_range(df_clean, column,
condition["min"], condition["max"])
    elif condition["method"] == "iqr":
        outliers_detected = detect_outliers_iqr(df_clean, column)

    # Add detected outliers to the outliers DataFrame
    outliers = pd.concat([outliers, outliers_detected])
```

```
   # Remove outliers from the clean DataFrame
   df_clean = df_clean.drop(outliers_detected.index)

# Remove duplicates from the outliers DataFrame
outliers = outliers.drop_duplicates()

# Print summary
print("Number of outliers removed:", len(outliers))
print("Cleaned data size:", len(df_clean))

df2=df_clean.reset_index(drop=True)

df2.to_csv('datasetCleaned_outliersFree.csv', index=False)
```

Modifing Dataset to Perform Analysis

```
df3=df2

df3['Gender_code'] = df3['Gender'].apply(lambda x: 1 if x == 'm' else -1 if x ==
'f' else 0)
df3['Financially constrained_coded'] = df2['Financially constrained'].apply(lambda
x: 0 if x == 'no' else 1 if x == 'yes' else 2)
df3['Age_Group_code'] = df3['Age Group'].apply(find_class_mid)
df3['college_code'] = df3['College Name'].astype('category').cat.codes + 1
df3['Family_income_coded'] = df2['Family_income'].apply(
   lambda x: 0 if x == 'below 3 lakhs' else
             1 if x == '3 -  7  lakhs' else
             2 if x == '7 - 10 lakhs' else
             3 if x == '10 - 12 lakhs' else
             4 if x == '12 - 15 lakhs' else 5)

new_column_order = ['df_index','Age_Group_code','Age Group','Gender_code',
'Gender','college_code', 'College Name',
                   'Degree ', 'Subject','Percent_10','Favorite_Subject_10',
'Marks_10','Percent_12','Favorite_Subject_12',
                   'Marks_12','CGPA','co-curricular activity','profession',
'preffed sector','Age','Expected_salary','Financially constrained_coded',
                   'Financially
constrained','Family_income_coded','Family_income', "Parent's occupation",
                   'Spending', 'parents_influence', 'friend_influence',
'mentor_influence',
                   'social_expectations', 'social_media_influence',
'personal_choice']
```

```
# Reorder DataFrame columns
df3 = df3[new_column_order]

df3.to_csv('refined_ForWork.csv', index=False)
```

Performing EDA on the Outliers-free Dataset

- Gender-wise Distribution:

```
Gender_counts = df3['Gender'].value_counts()
print(Gender_counts)
print("Sample Size: ",len(df3['Gender'].dropna()))
#------------------------------------------------------
Gender_counts.plot.pie(autopct='%1.1f%%', startangle=90)
plt.ylabel('')  # Hide the y-label
plt.title('Gender-wise Distribution')
plt.show()
```

- Family Income-wise Distribution:

```
print(df3['Family_income'].unique())
print(df3['Family_income'].value_counts())

income_order = ['below 3 lakhs', '3 -  7  lakhs',  '7 - 10 lakhs', '10 - 12
lakhs', '12 - 15 lakhs', '15 lakhs above']

# Convert the column to a categorical type with the defined order
df3['family_income'] = pd.Categorical(df3['Family_income'],
categories=income_order, ordered=True)

plt.figure(figsize=(8, 6))
sns.histplot(df3['family_income'], kde=True, bins=10, color='red')
plt.title('Distribution of Family Income')
plt.xlabel('Family Income')
plt.ylabel('Frequency')
plt.show()
```

- Age Group-wise Distribution:

```
age_group_counts = df3['Age Group'].value_counts()

# Plot a bar diagram
plt.figure(figsize=(8, 5))
age_group_counts.sort_index().plot(kind='bar', color='lightblue',
edgecolor='black')
```

```
# Add labels and title
plt.xlabel('Age Group', fontsize=12)
plt.ylabel('Number of Students', fontsize=12)
plt.title('Age Distribution of Students', fontsize=14)
plt.xticks(rotation=45, fontsize=10)
plt.yticks(fontsize=10)

# Show grid lines
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Show the plot
plt.tight_layout()
plt.show()
```

- Subject-wise Distribution:

```
age_group_counts = df3['Subject'].value_counts()

# Plot a bar diagram
plt.figure(figsize=(22, 15))
age_group_counts.sort_index().plot(kind='bar', color='blue', edgecolor='black')

# Add labels and title
plt.xlabel('Subjects', fontsize=12)
plt.ylabel('Number of Students', fontsize=12)
plt.title('Distribution of Students accross different Subjects', fontsize=14)
plt.xticks(rotation=85, fontsize=10)
plt.yticks(fontsize=10)

# Show grid lines
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Show the plot
plt.tight_layout()
plt.show()
```

- Distribution of Parent's Occupation among the responders:

```
age_group_counts = df3["Parent's occupation"].value_counts()

# Plot a bar diagram
plt.figure(figsize=(22, 15))
age_group_counts.sort_index().plot(kind='bar', color='lightgreen',
edgecolor='black')
```

```
# Add labels and title
plt.xlabel('Sectors of Occupation', fontsize=12)
plt.ylabel('Number of samples', fontsize=12)
plt.title("Distribution of Parent's Occupation", fontsize=14)
plt.xticks(rotation=85, fontsize=10)
plt.yticks(fontsize=10)

# Show grid lines
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Show the plot
plt.tight_layout()
plt.show()
```

- Distribution of Students' desired profession

```
age_group_counts = df3["profession"].value_counts()

# Plot a bar diagram
plt.figure(figsize=(22, 15))
age_group_counts.sort_index().plot(kind='bar', color='lightyellow',
edgecolor='black')

# Add labels and title
plt.xlabel('Desired profession', fontsize=12)
plt.ylabel('Number of samples', fontsize=12)
plt.title("Distribution of Students' desired profession", fontsize=14)
plt.xticks(rotation=85, fontsize=10)
plt.yticks(fontsize=10)

# Show grid lines
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Show the plot
plt.tight_layout()
plt.show()
```

Analysis:

- Association between. Profession Aspired by Students:

```
data = df3
# defining cramer’s V statistics
def cramers_v(chi2, n, rows, cols):
   return np.sqrt(chi2 / (n * (min(rows-1, cols-1))))
```

```
# Create a contingency table for Gender and Profession
contingency_table = pd.crosstab(data['Gender'], data['profession'])

# Perform the Chi-Square Test of Independence
chi2, p, dof, expected = chi2_contingency(contingency_table)

# Display the results
#print("Contingency Table:")
#print(contingency_table)
print("\nChi-Square Statistic:", chi2)
print("P-value:", p)
print("Degrees of Freedom:", dof)
#print("\nExpected Frequencies:")
#print(expected)
alpha = 0.05
if p < alpha:
 print("The test shows a significant relationship (reject the null hypothesis).")
else:
 print("The test does not show a significant relationship (fail to reject the null hypothesis).")
```

- Association between Degrree chosen by the students and their expected salary as a fresher:

```
# Create a contingency table
contingency_table = pd.crosstab(df3['Degree '], df3['Expected_salary'])

# Perform Chi-Square Test
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)

# Results
#print("Contingency Table:\n", contingency_table)
print("Chi-Square Statistic:", chi2)
print("p-value:", p)
print("Degrees of Freedom:", dof)
#print("Expected Frequencies:\n", expected)

# Interpretation
alpha = 0.05
if p <= alpha:
   print("Reject the null hypothesis: Degree and parents_influence are dependent.")
else:
   print("Fail to reject the null hypothesis: Degree and parents_influence are independent.")
```

- Association between Preffered Sector of Working by the Students and Their Parents’ Occupation:

```
# Create a contingency table
contingency_table = pd.crosstab(df3['preffed sector'], df3["Parent's occupation"])

# Perform Chi-Square Test
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)

# Results
#print("Contingency Table:\n", contingency_table)
print("Chi-Square Statistic:", chi2)
print("p-value:", p)
print("Degrees of Freedom:", dof)
#print("Expected Frequencies:\n", expected)

# Interpretation
alpha = 0.05
if p <= alpha:
    print("Reject the null hypothesis: Degree and parents_influence are
dependent.")
else:
    print("Fail to reject the null hypothesis: profession chosen and parent's
occupation are independent.")

# Assuming you have the Chi-Square test results
n = contingency_table.sum().sum()
rows, cols = contingency_table.shape
v = cramers_v(chi2, n, rows, cols)
print("Cramér's V:", v)
```

- Association between profession desired to be pursed by the student and Parent’s occupation:

```
# Create a contingency table
contingency_table = pd.crosstab(df3['profession'], df3["Parent's occupation"])
# Perform Chi-Square Test
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)
# Results
print("Chi-Square Statistic:", chi2)
print("p-value:", p)
print("Degrees of Freedom:", dof)
# Interpretation
```

```
alpha = 0.05

if p <= alpha:

   print("Reject the null hypothesis: Degree and parents_influence are
dependent.")

else:

   print("Fail to reject the null hypothesis: profession chosen and parent's
occupation are independent.")
```

- Association between Degrree chosen by the students and their expected salary as a fresher:

```
# Create a contingency table
contingency_table = pd.crosstab(df3['Degree '], df3['Expected_salary'])

# Perform Chi-Square Test
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)

# Results
#print("Contingency Table:\n", contingency_table)
print("Chi-Square Statistic:", chi2)
print("p-value:", p)
print("Degrees of Freedom:", dof)
#print("Expected Frequencies:\n", expected)

# Interpretation
alpha = 0.05
if p <= alpha:
   print("Reject the null hypothesis: Degree and parents_influence are
dependent.")
else:
   print("Fail to reject the null hypothesis: Degree and parents_influence are
independent.")

# Assuming we have the Chi-Square test results
n = contingency_table.sum().sum()
rows, cols = contingency_table.shape
v = cramers_v(chi2, n, rows, cols)
print("Cramér's V:", v)

sns.heatmap(contingency_table, annot=True, cmap="Blues")
plt.title("Heatmap of Contingency Table")
plt.show()
```

Performing VIF Test on Numarical Columns to Detect Multicolinearity

```
selected_columns = ['Age_Group_code','Gender_code', 'college_code',
'Percent_10','Marks_10','Percent_12','Marks_12','CGPA','Expected_salary','Spending
','parents_influence',  'friend_influence',
'mentor_influence','social_expectations',
'social_media_influence','personal_choice']  # Replace with your desired columns
for i in range(len(selected_columns)):
 #print(type(df3[selected_columns[i]][1]))
 # Replace NaN values in 'column_name' with 0
 df3[selected_columns[i]] = df3[selected_columns[i]].fillna(0)
print(type(selected_columns))
df3_selected = df3[selected_columns]
VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = df3_selected.columns

# Calculating VIF for each feature
vif_data["VIF"] = [
   variance_inflation_factor(df3_selected.values, i)
   for i in range(len(df3_selected.columns))
]

print(vif_data)
```

Performing PCA on Academic Variables:

```
df_new = df3[['Marks_10','Marks_12','Percent_10','Percent_12']]
print(df_new)

def perform_pca(data, n_components=None):
   """
   Performs PCA on the given dataset and returns explained variance ratios,
   transformed components, and the PCA model.

   Args:
       data (pd.DataFrame): The input dataset (numerical columns only).
       n_components (int or None): Number of principal components to keep.
                                    If None, keep all components.

   Returns:
       tuple: (explained variance ratios, transformed components, PCA model)
   """
   # Standardize the data (important for PCA)
   scaler = StandardScaler()
   data_scaled = scaler.fit_transform(data)
```

```
    # Perform PCA
    pca = PCA(n_components=n_components)
    principal_components = pca.fit_transform(data_scaled)

    # Explained variance ratio
    explained_variance_ratio = pca.explained_variance_ratio_

    return explained_variance_ratio, principal_components, pca


n_components = 2  # Specify the number of components
explained_variance, components, pca_model = perform_pca(df_new, 2)
df3 = pd.concat([df3, pd.DataFrame(components, columns=['Marks1', 'Marks2'])],
axis=1)
df3 = df3.drop(['Marks_10','Marks_12','Percent_10','Percent_12'], axis=1)

print("Explained Variance Ratio:", explained_variance)
print("Principal Components:\n", components)
print(type(components))
# Plot the explained variance
plt.figure(figsize=(8, 5))
plt.bar(range(1, len(explained_variance) + 1), explained_variance, alpha=0.7,
align='center', label='Individual Explained Variance')
plt.step(range(1, len(explained_variance) + 1), np.cumsum(explained_variance),
where='mid', label='Cumulative Explained Variance')
plt.xlabel('Principal Component Index')
plt.ylabel('Explained Variance Ratio')
plt.title('Explained Variance by Principal Components')
plt.legend(loc='best')
plt.show()
```

Construction of Correlation Matrix:

```
def plot_correlation_matrix(data, columns):
    # Compute the correlation matrix for the specified columns
    correlation_matrix = data[columns].corr()

    # Set up the figure and axis
    plt.figure(figsize=(8, 6))

    # Plot the heatmap
    sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f',
cbar=True)
```

```
   # Add title
   plt.title("Correlation Matrix", fontsize=16)

   # Show the plot
   plt.show()

selected_columns.append('Marks1')
selected_columns.append('Marks2')
selected_columns.remove('Marks_10')
selected_columns.remove('Marks_12')
selected_columns.remove('Percent_10')
selected_columns.remove('Percent_12')
plot_correlation_matrix(df3, selected_columns)
```

Checking for Normality among the Variables Collected in Likert's Scale

```
list = ['parents_influence', 'friend_influence', 'mentor_influence',
'social_expectations', 'social_media_influence', 'personal_choice']

for item in list:
 # Perform the KS test
 stat, p_value = kstest(df3[item].dropna(), 'norm',
args=(np.mean(df3[item].dropna()), np.std(df3[item].dropna())))

 # Display the result
 print(f"KS Statistic: {stat}")
 print(f"P-value: {p_value}")
 skew_value = stats.skew(df3[item].dropna())
 print(f"Skewness: {skew_value}")
 if skew_value > 0:
   stat2, p_value1 = kstest(np.log(df3[item].dropna()), 'norm',
args=(np.mean(np.log(df3[item].dropna())), np.std(np.log(df3[item].dropna()))))
   print(f"KS Statistic for Log-Transformed Data: {stat2}")
   print(f"P-value for Log-Transformed Data: {p_value1}")
 # Interpretation
 if p_value < 0.05:
   print(f"Reject the null hypothesis: The sample does not follow a normal
distribution.{item}")
 else:
   print(f"Fail to reject the null hypothesis: The sample appears to follow a
normal distribution.{item}")
```

Performing Kruskal-Wali's H Test:

```
for item in list:

 df4 = df3[['Family_income', item]]

#print(df4)

 new_df=[]

# Load the dataset

 data = df4

# Ensure the column 'Family_income' exists

 if 'Family_income' not in data.columns:

   raise ValueError("The dataset does not contain the 'Family_income' column.")

# Grouping the dataset based on `Family_income`

 income_groups = data.groupby("Family_income")

# Splitting the data into six subgroups

 subgroups = [group for _, group in income_groups]

# Check the number of subgroups

 if len(subgroups) > 6:

   print(f"More than 6 groups detected. Only first 6 will be used.")

   subgroups = subgroups[:6]

 print(item)

# Display the size of each subgroup

 for i, group in enumerate(subgroups, start=1):

   print(f"Subgroup {i}: Size = {len(group)}")


# If needed, save subgroups into separate variables

 for i, group in enumerate(subgroups, start=1):

   new_df.append(group[item].tolist())


 new_df2=pd.DataFrame(new_df)

# Example data

 data = new_df2.transpose()

# Check data structure
```

```
 print(data)
#print(data[0])
 df5=[data[0].dropna(),
                     data[1].dropna(),
                    data[2].dropna(),
                    data[3].dropna(),
                    data[4].dropna(),data[5].dropna()]
# Test Homogeneity of Variances
 stat, p_val = levene(data[0].dropna(),
                    data[1].dropna(),
                    data[2].dropna(),
                    data[3].dropna(),
                    data[4].dropna(),data[5].dropna())
 print("Levene's Test p-value:", p_val)
 if p_val < 0.05:
   print("Warning: Variances are not homogeneous.")
 stat, p_value = kruskal(data[0].dropna(),
                     data[1].dropna(),
                    data[2].dropna(),
                    data[3].dropna(),
                    data[4].dropna(),data[5].dropna())


# Display the results
 print(f"Kruskal-Wallis H Statistic: {stat}")
 print(f"P-value: {p_value}")


# Interpretation
 if p_value < 0.05:
   print("Reject the null hypothesis: There are significant differences between groups.")
   p_values1=sp.posthoc_dunn(df5, val_col=item, group_col='Family_income',
p_adjust='bonferroni')

   p_values2=sp.posthoc_dunn(df5, val_col=item, group_col='Family_income',
p_adjust='holm')

   print(p_values1)

   print(p_values2)

 else:

   print("Fail to reject the null hypothesis: No significant differences between
groups.")
```