IDS Project Report

Student Budget Analysis & Prediction Tool

Group Members:

Abdur Rahman Khan (242855)
Muhammad Ahmed Junaid (242843)
Abdul Rafey Arfan (242851)
Muhammad Arqam Aqeel (242874)

Subject: Introduction to Data Science

Instructor: Sir Muhammad Irfan ul Haq

Department of Creative Technology

BSDS-III-A

Submission Date: 24-12-2025

1. Abstract

This project presents a data-driven approach to analyzing student monthly expenditures. University students often struggle with budget management due to varying costs in transport, accommodation, and food. Using a dataset collected from students (Student Expenditure Survey.xlsx), we performed data cleaning, exploratory data analysis (EDA), and implemented a predictive 'Budget Calculator' tool. The system helps students estimate their expected monthly expenses based on parameters like commute type, accommodation status, and laundry preferences. The expected outcome is a clearer understanding of financial patterns among students and a practical tool for financial planning.

2. Introduction

Background: Financial literacy and management are crucial for university students, many of whom are managing their own finances for the first time. Expenses vary significantly based on lifestyle choices such as living in a hostel vs. being a day scholar.

Importance: Understanding these spending patterns helps in identifying major cost drivers (e.g., transport vs. food) and allows for better financial decision-making.

Basic Idea: The project analyzes survey data to identify correlations between lifestyle choices and expenses. It culminates in an interactive Python-based calculator that inputs a user's profile and outputs a predicted budget range.

3. Problem Statement

The Problem: Students often underestimate their monthly costs, leading to financial stress. There is a lack of localized data regarding how much a typical student spends on specific categories like 'Indrive' vs. 'University Bus'.

Issues in Existing System: Manual estimation is error-prone. Generic online calculators do not account for specific university contexts (e.g., local transport rates).

Proposed Improvements: A data-backed calculator that uses real peer data to provide accurate estimates.

4. Objectives

Main Objective: To analyze student spending behaviors and develop a prediction tool for monthly budgets.

- To clean and preprocess raw survey data (handling outliers and null values).
- To visualize spending distributions across different demographics (Gender, Year of Study).

- To analyze the impact of daily commute methods on overall transport costs.
- To develop an interactive function that estimates expenses based on user inputs.

5. Scope of the Project

Included:
Data Cleaning (IQR method for outliers).
Visualizations (Box plots, Bar charts, Pie charts).
Interactive Budget Calculator function.

Limitations:

The prediction is based on a specific sample size (approx. 120 rows).
Inflation or sudden price changes are not real-time updated.

Not Included:

A web-based interface or mobile application (currently runs in Python environment).

6. Tools & Technologies

Programming Language: Python

IDE/Environment: Google Colab

Libraries: Pandas (Data Manipulation), NumPy (Math), Matplotlib & Seaborn (Data Visualization)

7. Methodology / Working

1. Data Collection: Survey data loaded from 'Student Expenditure Survey.xlsx'.
2. Data Cleaning: Renamed columns for consistency. Handled outliers in 'distance' and 'transport cost' using IQR (Interquartile Range).
3. Data Transformation: Standardized categorical values (e.g., 'Walking' cost set to 0).
4. Exploratory Data Analysis (EDA): Generated charts to see relationships (e.g., Transport Cost vs. Commute Method).
5. Calculator Logic: Created a function 'interactive_budget_calculator()' that filters the dataset based on user input and calculates the mean expense.

8. Implementation / Code Explanation

Key Logic Modules:

1. Outlier Removal:

We used the IQR method to remove unrealistic data points (e.g., extremely high distances).
Q1 (25%) and Q3 (75%) were calculated to define lower and upper bounds.

2. Data Cleaning:

Ensured that students who selected 'Walking' had a Transport Cost of 0. Cleaned column names by stripping whitespace.

3. Budget Calculator Function:

The function accepts inputs like 'Accommodation Status' and 'Daily Commute'. It filters the cleaned DataFrame (df_cleaned) to find similar students and calculates the average total expenditure.

9. Results / Output

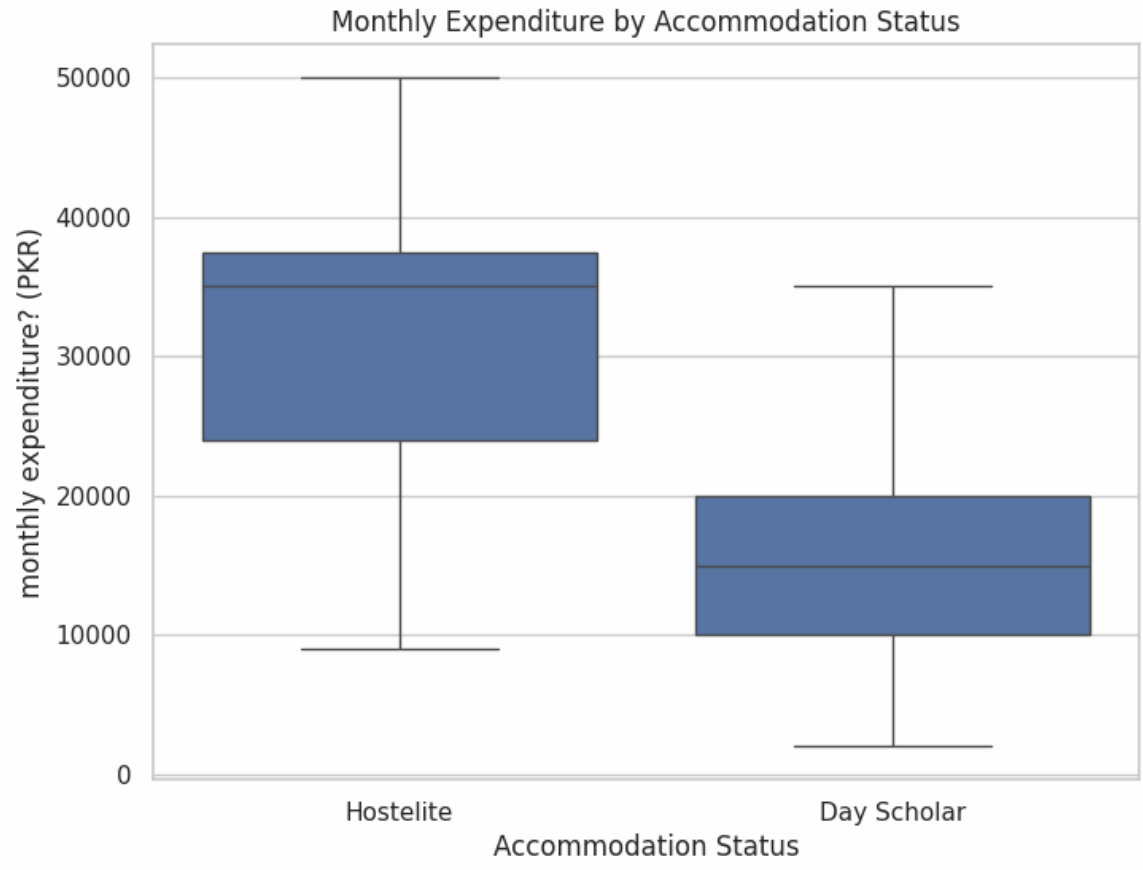
The analysis provided key insights, such as 'Private Transport' being significantly more expensive than 'University Bus'. Below are descriptions of the generated outputs:

Sample Calculator Output:

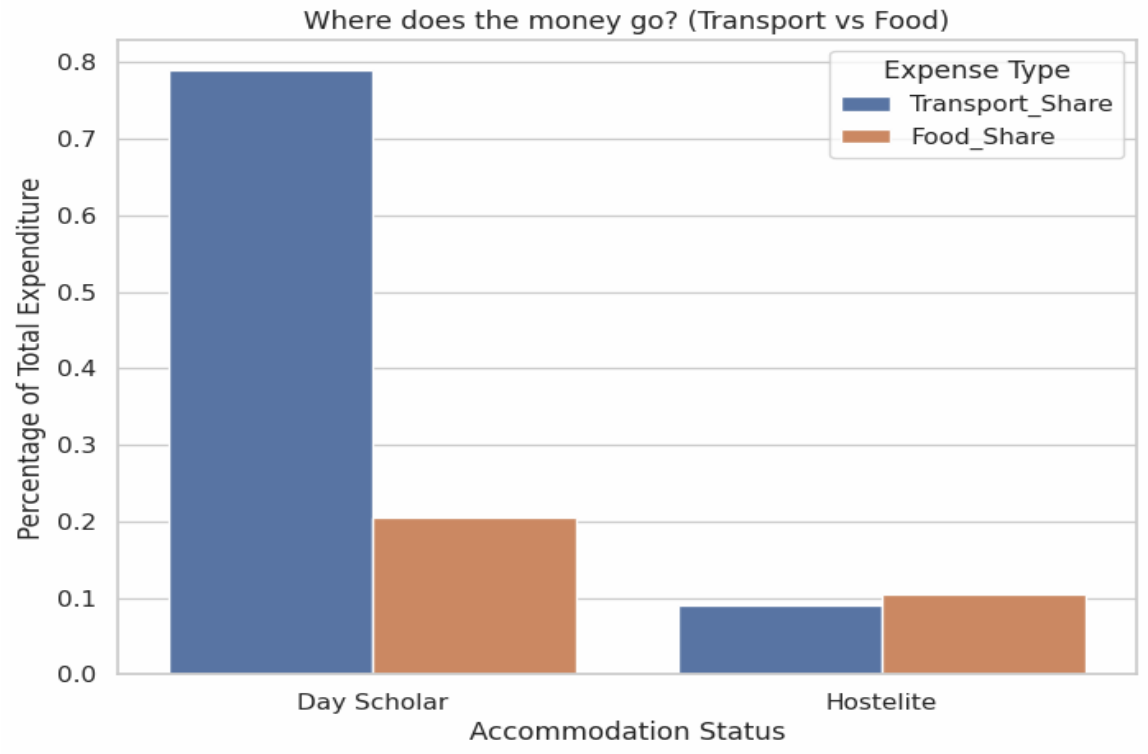
Input: Day Scholar, Private Transport, Sophomore.
Predicted Range: 35,000 PKR - 45,000 PKR (derived from average of matching records).

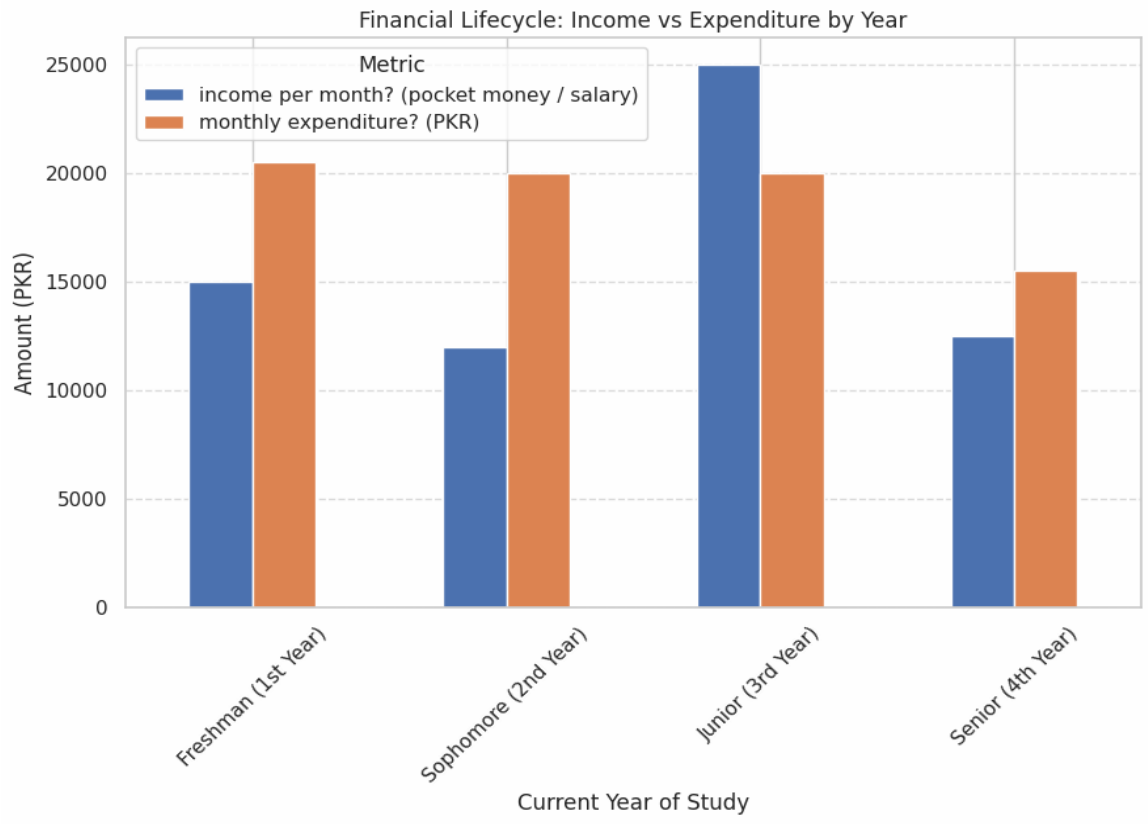
Visualizations Generated:

- **Box Plot:** Shows Distribution of Monthly Expenditure by Accommodation Status

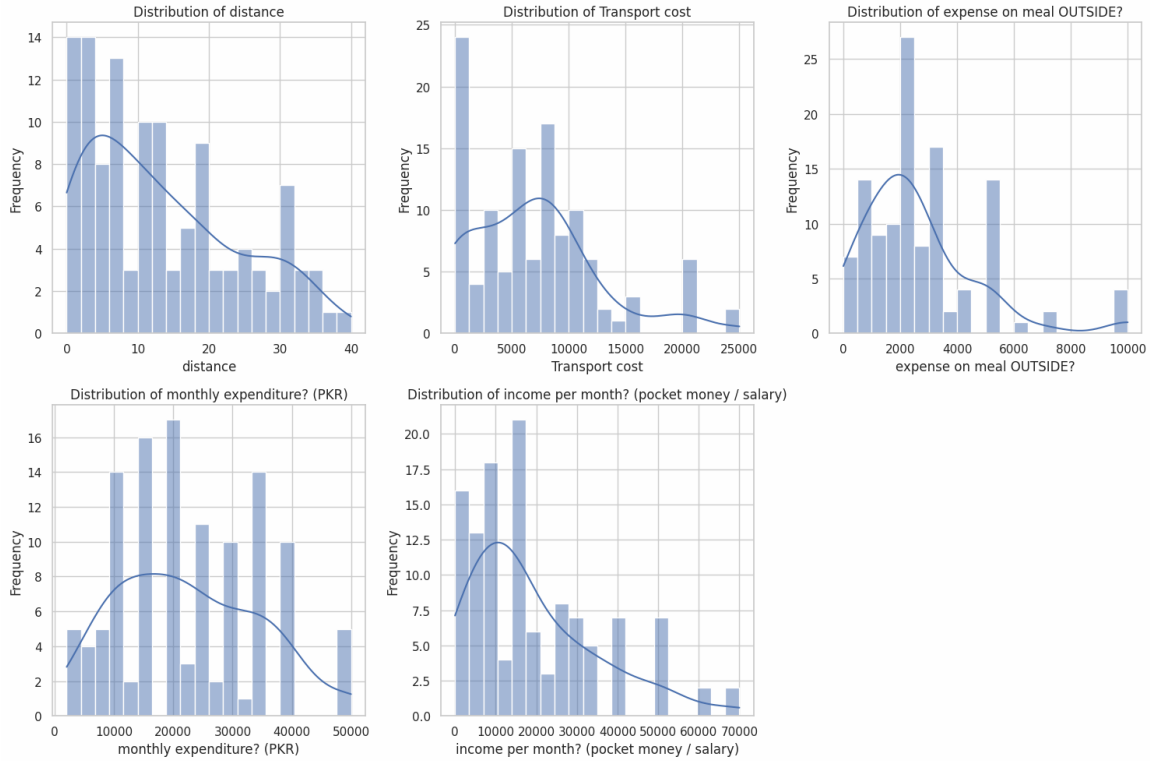


- **Bar Chart:** Expense Type and Year wise distribution of monthly expenditure.

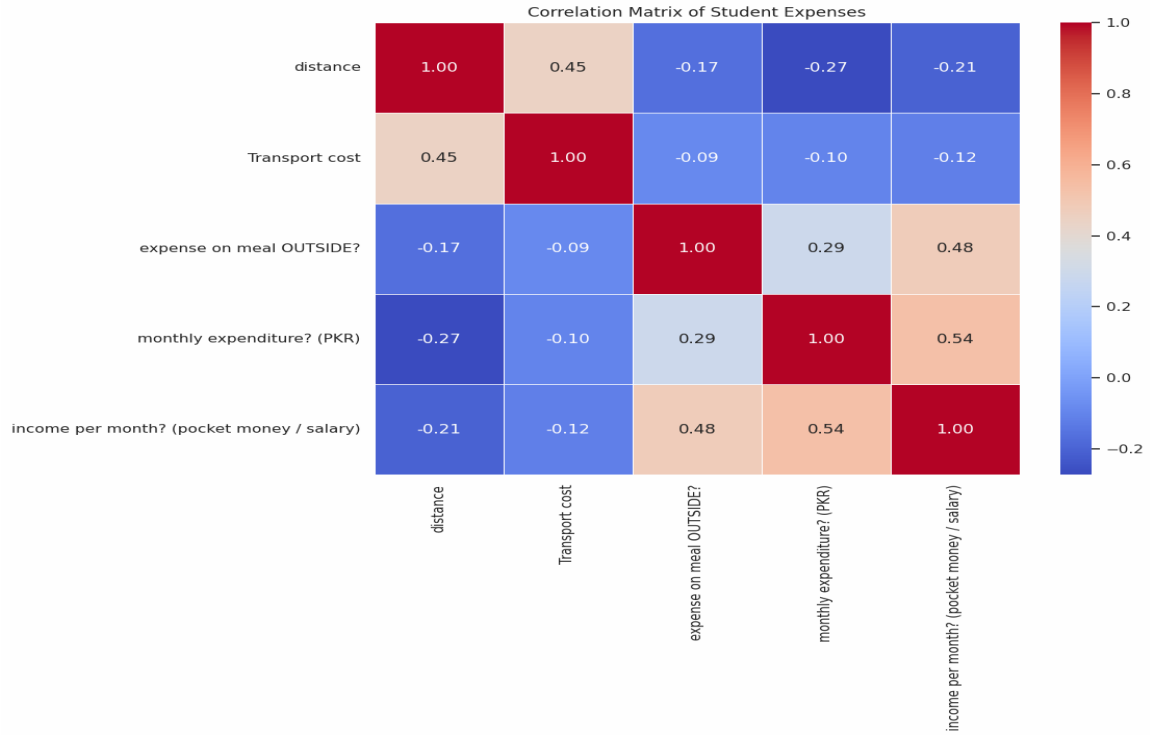




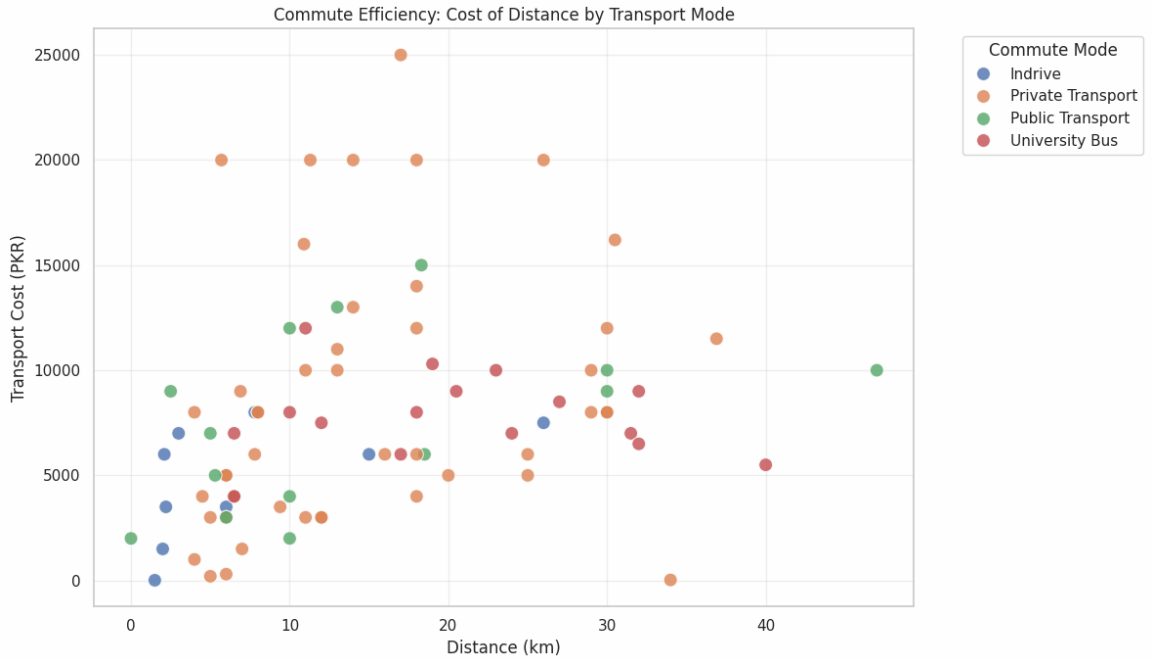
- **Histogram:** Show "Distribution" & "Frequency"



- **Heatmap:** Show "Correlation" & "Relationships"



- **Scatter Plot:** Compare Distance to University vs. Transport Cost



10. Insights

- **Commute Cost Disparities:** "Private Transport" is significantly more expensive (avg ~9,200 PKR) compared to the "University Bus" (avg ~7,800 PKR) or "Public Transport" (avg ~7,100 PKR). Students who walk naturally have zero transport costs, which had to be manually corrected in the data cleaning process.
- **Distance vs. Cost Anomalies:** There is not a strictly linear relationship between distance and cost for all students. Some students traveling short distances (e.g., via Indrive or Private Transport) pay disproportionately high fares compared to those traveling longer distances on subsidized transport like the University Bus.
- **Outlier Detection:** The data contained unrealistic outliers, such as students reporting extremely high distances (up to 500 km), which were removed to prevent skewed averages. Similarly, inconsistent transport costs were filtered out to improve prediction accuracy.
- **Spending Distribution:** The majority of students rely on "Pocket Money" rather than earned income. Monthly expenditure varies widely, with many students spending between 20,000 to 40,000 PKR, but there are significant outliers at both the lower and higher ends.
- **Demographic Split:** The dataset is heavily skewed towards younger students (Freshmen and Sophomores), meaning the insights primarily reflect the spending habits of 1st and 2nd-year students rather than seniors or post-grads.

10. Conclusion

The project successfully analyzed student expenditure patterns. By removing 7+ outliers and correcting data inconsistencies, we improved the accuracy of the dataset. The final Budget Calculator serves as a useful utility for students to plan their finances based on realistic peer data.

11. Future Enhancements

- Integrate Machine Learning (Regression) for more precise predictions.
- Develop a web interface (using Streamlit or Flask) for easier access.
- Expand the dataset to include more universities for broader applicability.

12. References

- Pandas Documentation (pandas.pydata.org)
- Seaborn Visualization Library (seaborn.pydata.org)
- Class Lectures: Introduction to Data Science by Sir Muhammad Irfan ul Haq