DATA ANALYTICS INTERNSHIP

A Project Report Submitted to

GLOBAL NEXT CONSULTING INDIA PVT. LTD.

(Six-Week Internship Program)

By

Aditya Patil

Under the Supervision of

Ms. Anuradha Gupta

(Project Director)

Submitted To:

Global Next Consulting India Pvt. Ltd.

Duration of Internship:

19th March 2026

to

8th May 2026

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report, "**Data Analyst Internship (GNCIPL)**", submitted as per the requirements for the Data Analyst / Business Analyst role, is the result of original work carried out by me under the guidance of **Ms. Anuradha Gupta** during the internship period.

I further declare that this report represents an authentic record of my own work and does not contain any falsely fabricated ideas, data, facts, or sources. I have adhered to all principles of academic honesty and integrity, and this report has not been submitted, either in part or in full, to any other institute, university, or organization for the award of any degree, diploma, or certification.

Aditya Patil

CERTIFICATE

This is to certify that the project report entitled "**Data Analyst Internship Report**" has been carried out by **Aditya Patil**, a Data Analyst Intern at Global Next Consulting India Pvt. Ltd. This work was carried out under the guidance of **Ms. Anuradha Gupta**, Program Director, GNCIPL, during the internship period. It is further certified that this work has not been submitted to any other university or institution for the award of any other degree, diploma, or certificate.

Ms. Anuradha Gupta
Program
Director
Global Next Consulting India Pvt. Ltd.

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude to all those who have contributed to the successful completion of this internship and the preparation of this project report.

I would like to express my heartfelt gratitude to my supervisor, Ms. Anuradha Gupta, Program Director at Global Next Consulting India Pvt. Ltd., for her invaluable guidance, unwavering encouragement, and constructive feedback throughout the course of this internship. Her expertise in data analytics, her patience in answering my queries, and her clarity in setting goals and expectations played a defining role in the successful execution of each weekly project.

I am deeply thankful to the team and staff of Global Next Consulting India Private Limited for providing access to the necessary resources, datasets, tools, and infrastructure. Their professional environment, collaborative culture, and consistent support created an atmosphere that was highly conducive to learning and growth. The regular feedback sessions and structured review meetings provided by the organization greatly accelerated my understanding of real-world data analytics workflows.

I also wish to extend my gratitude to my peers and mentors whose discussions, suggestions, and collaborative spirit enriched my learning experience. Their insights, especially during problem-solving and debugging sessions, broadened my analytical perspective considerably.

Aditya Patil

ABSTRACT

This report presents a comprehensive account of the six-week Data Analyst Internship undertaken at Global Next Consulting India Pvt. Ltd. (GNCIPL), structured across six progressively complex weekly projects spanning the domains of e-commerce analytics, educational data mining, health and fitness data analysis, and energy consumption pattern forecasting.

The internship was designed to provide hands-on exposure to industry-standard data analytics tools and methodologies, including Microsoft Excel, Python (Google Colab), SQLite, and various data visualization libraries. Each weekly project addressed a unique real-world problem, enabling the development of practical skills in data collection, cleaning, exploratory data analysis (EDA), statistical modelling, dashboard creation, and actionable business reporting.

Weeks 1 and 2 focused on E-commerce Performance and Sales Trend Analysis using Excel-based dashboarding, KPI tracking, and pivot table analysis. Week 3 introduced Python-based analytical workflows for Student Performance Prediction, integrating SQLite for structured querying and OLS regression for statistical modelling across a dataset of 29,446 cleaned records. Week 4 applied Python analytics to Fitness Tracker data, deriving health insights from step counts, caloric burn, heart rate, and sleep metrics through correlation analysis and scatter visualization. Week 5 explored AI and analytics-driven methodologies applied to a specialized dataset. Week 6 concluded the internship with a Time-Series Energy Consumption Analysis across 121,273 hourly records of the AEP energy grid, identifying peak usage hours, seasonal consumption patterns, and anomaly detection.

Collectively, the projects developed proficiency in the complete data analytics pipeline — from raw data ingestion and preprocessing to statistical inference, visualization, and business recommendation generation. This report documents the objectives, methodology, findings, challenges, and business insights derived from each project, accompanied by a unified reflection on overall technical and professional growth.

INDEX / TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Company Profile

Global Next Consulting India Private Limited (GNCIPL), headquartered in Greater Noida, Uttar Pradesh, is a forward-thinking consulting and technology services firm that specializes in cybersecurity, data analytics, and business intelligence solutions. The organization is committed to empowering businesses with data-driven decision-making tools and proactive risk management strategies. As digital transformation accelerates across industries, GNCIPL positions itself as a strategic partner that bridges the gap between raw organizational data and meaningful, actionable intelligence.

GNCIPL serves clients across diverse verticals including finance, healthcare, manufacturing, retail, and technology. The firm's core service offerings encompass threat detection, data risk assessment, incident response planning, compliance consulting, advanced analytics, and 24/7 operational monitoring. The company's core values integrity, innovation, customer-centricity, excellence, and collaboration form the philosophical backbone of all its engagements, ensuring that technical solutions are always aligned with the client's specific business context and long-term strategic objectives.

As part of its commitment to talent development and industry-academia collaboration, GNCIPL regularly offers structured internship programs to aspiring data professionals. These programs are designed to mirror real-world analytics engagements, exposing interns to the complete data lifecycle from raw data ingestion to polished, decision-ready insights.

Contact Details:

- Location: B5, 402 P4 PHi2, CGEWHO Tower, Greater Noida – 201310
- Contact: 0120-4001768 | +91-9315504902 | +91-7666141260
- Email: hr@gncipl.com

1.2 Objectives of Internship

The six-week Data Analyst Internship at GNCIPL was structured with clearly defined learning objectives designed to equip the intern with both technical proficiency and business analytical acuity. The primary objectives of this internship were:

- To develop hands-on expertise in data analytics tools and frameworks, including Microsoft Excel, Python (Pandas, NumPy, Matplotlib, Seaborn, Statsmodels), and SQLite, through applied real-world project work.
- To work with authentic, industry-representative datasets and deliver meaningful, structured insights supported by data visualizations, KPI dashboards, and summary reports.
- To build proficiency in the complete data analytics pipeline: data collection, preprocessing, cleaning, exploratory analysis, statistical modelling, visualization, and reporting.
- To understand the business context behind data analysis by translating technical findings into actionable recommendations that address organizational challenges.
- To develop strong communication and presentation skills by documenting project workflows, preparing structured reports, and articulating analytical findings in a professional format.
- To gain exposure to time-series analysis, regression modelling, anomaly detection, and correlation analysis as applied to diverse datasets spanning e-commerce, education, health, and energy domains.
- To strengthen analytical reasoning, structured problem-solving, and critical thinking skills by confronting real-world data quality issues such as missing values, duplicates, inconsistent data types, and noisy records.

CHAPTER 2: WEEKLY PROJECTS

2.1 Week 1 – E-Commerce Performance & Behaviour Trends (Excel)

2.1.1 Introduction

The first week of the internship was dedicated to an in-depth analysis of an e-commerce platform's sales performance and customer behaviour using Microsoft Excel. E-commerce businesses generate vast volumes of transactional and demographic data that, when analysed effectively, can reveal powerful insights about customer preferences, purchasing patterns, and revenue drivers. This project served as a foundational exercise in Excel-based data analytics, combining data cleaning, pivot table analysis, KPI computation, and interactive dashboard design.

The dataset encompassed customer demographics (gender, age, occupation, marital status), geographic distribution (state-wise), product category preferences, and transactional details (order amounts, purchase frequency). The primary analytical objective was to understand which customer segments drive the highest revenue, which states generate the most business volume, and how product category trends differ across demographic groups.

2.1.2 Objectives

- To clean and standardize the raw e-commerce dataset for accurate analysis.
- To analyse customer demographics including gender, age group, marital status, and occupation.
- To perform state-wise sales analysis and identify top-performing geographic regions.
- To examine product category trends and identify best-selling categories.
- To compute key performance indicators (KPIs) including total revenue, average order value, and order volume.
- To build an interactive Excel dashboard that provides a unified view of all analytical findings.
- ***To generate business recommendations based on data-driven insights.***

2.1.3 Methodology

The analytical workflow for this project followed a structured multi-phase approach:

Phase 1 – Data Preparation:

The raw dataset was first examined for structural inconsistencies. Columns with irrelevant or redundant data were removed. Missing values were identified and treated appropriately — numeric missing values were replaced with column means, while categorical missing values were filled with the mode. Data types were verified and corrected to ensure accurate formula computation in Excel. Age data was binned into meaningful

categories (e.g., 18–25, 26–35, 36–45, 46+) using Excel's IF and IFS functions. Similarly, income and order value ranges were categorized to enable meaningful segmentation.

Phase 2 – Analytical Framework:

Pivot Tables were constructed for each key analytical dimension: gender-wise sales, state-wise order volume and revenue, age-group purchase behaviour, occupation-wise spending patterns, and product category performance. KPI metrics were computed using Excel formulas including SUMIF, COUNTIF, AVERAGEIF, and VLOOKUP. Conditional formatting was applied to highlight top performers and outliers across all pivot views.

Phase 3 – Dashboard Creation:

An interactive Excel dashboard was built consolidating all analytical outputs. The dashboard incorporated KPI cards, column charts, bar charts, pie charts, and slicers for dynamic filtering by gender, state, and product category. The design prioritized clarity and visual hierarchy to ensure that business stakeholders could derive insights without requiring technical data expertise.

2.1.4 Key Findings and Insights

The following findings emerged from the analysis:

Customer Demographics:

- Female customers accounted for a higher proportion of total orders, reflecting stronger purchase frequency among women shoppers.
- The 26–35 age group emerged as the highest spending demographic, followed by the 36–45 bracket, suggesting that working professionals form the primary consumer base.
- Married customers demonstrated higher average order values compared to unmarried customers, indicating greater household purchasing power.

Occupation-Wise Analysis:

- IT sector employees, healthcare professionals, and aviation workers recorded the highest order amounts, reflecting higher disposable incomes in these segments.
- Students and entry-level professionals showed lower average spending but higher purchase frequency in budget product categories.

State-Wise Sales Performance:

- Uttar Pradesh, Maharashtra, and Karnataka emerged as the top three states by total order revenue, driven by high urban population density.

- Southern states demonstrated strong preference for electronics and home appliances, while northern states showed higher traction in clothing and food categories.

Product Category Trends:

- Food and Beverage, Clothing, and Electronics were the top three revenue-generating product categories across all demographic groups.
- Home and Furniture showed strong seasonal demand spikes, particularly among married customers in the 30–45 age group.

2.1.5 KPI Summary

Total Orders Analysed	27,981
Top Revenue-Generating State	Uttar Pradesh
Highest Spending Age Group	26–35 Years
Top Product Category	Food and Beverage
Top Occupation by Sales	IT Sector Professionals
Gender with Higher Purchase Volume	Female Customers

Revenue by Gender	
Gender	Sum of Amount
F	74335856
M	31913276
Grand Total	106249132.4

Revenue by Age-Segment	
Age_Segment	Sum of Amount
Adult	53689408.93
Senior	34715861.5
Youth	17843862
Grand Total	106249132.4

Revenue by State	
State	Sum of Amount
Uttar Pradesh	19374968
Maharashtra	14427543
Karnataka	13523540
Delhi	11603819
Madhya Pradesh	8101142
Andhra Pradesh	8037147
Himachal Pradesh	4963368
Haryana	4220175
Bihar	4022757
Gujarat	3946082
Kerala	3894492
Jharkhand	3026456
Uttarakhand	2520944
Rajasthan	1909409
Punjab	1525800
Telangana	1151490
Grand Total	106249132.4

Revenue by Product Category	
Product_Category	Sum of Amount
Office	81936
Veterinary	112702
Hand & Power Too	405618
Pet Care	482277
Decor	730360
Books	1061478
Tupperware	1155642
Household items	1569337
Stationery	1676052
Auto	1958610
Beauty	1959484
Sports Products	3635933
Games & Toys	4331694
Furniture	5440052
Footwear & Shoes	15575209
Electronics & Gadg	15643846
Clothing & Apparel	16495019
Food	33933884
Grand Total	106249132

Revenue by Revenue Segment	
Revenue Segmentatio	Sum of Amount
High Value	63250502.43
Regular	42998630
Grand Total	106249132.4

2.1.6 Business Recommendations

- Focus targeted marketing campaigns on the 26–35 age demographic, particularly for high-margin product categories such as electronics and fashion.
- Develop state-specific promotional strategies for Uttar Pradesh, Maharashtra, and Karnataka to capitalize on their existing high revenue contribution.
- Introduce loyalty programmes targeted at married customers to leverage their higher average order values.
- Invest in personalized product recommendation engines for IT sector and healthcare professional segments to increase cross-category purchasing.

2.2 Week 2 – E-Commerce Sales Trends Analysis (Excel Dashboard)

2.2.1 Introduction

Building upon the customer behaviour analysis conducted in Week 1, the second week's project focused specifically on e-commerce sales trend analysis, with emphasis on revenue trend monitoring, monthly performance tracking, product-level sales patterns, and KPI dashboard development. While Week 1 concentrated on demographic segmentation, this project shifted focus toward temporal and transactional dimensions, examining how sales volumes and revenue figures evolve over time and across product lines.

The dataset for this analysis contained order-level transactional data including order dates, product categories, customer regions, revenue figures, and quantity sold. Excel's advanced analytical features — including Power Query for data transformation, pivot charts for trend visualization, and dynamic slicers for interactive filtering
— were leveraged to construct a comprehensive sales analytics dashboard.

2.2.2 Objectives

- To analyse monthly and quarterly revenue trends to identify growth or decline patterns.
- To identify top-performing product categories by sales volume and revenue contribution.
- To examine customer purchasing behaviour trends across time periods.
- To construct KPI cards capturing critical business performance metrics.
- To develop an interactive sales analytics dashboard that enables self-service business intelligence.
- ***To identify underperforming categories and generate targeted improvement recommendations.***

2.2.3 Methods

Data_Preparation

The transactional dataset was loaded into Excel and subjected to a thorough cleaning process. Date columns were converted to proper date format to enable time-based analysis. Duplicate order records were identified and removed. Revenue figures were cross-validated against quantity and unit price columns for consistency. New derived columns were engineered including Month, Quarter, and Year dimensions to support temporal aggregation.

Analysis Approach:

Monthly and quarterly revenue pivot tables were constructed to examine seasonal trends and growth trajectories. Product category pivot tables were built to rank categories by cumulative revenue and average order size. Customer purchase frequency analysis was performed to distinguish between one-time buyers and repeat customers. Trend lines were added to time-series charts to identify underlying growth or decline momentum. Dashboard slicers were configured for dynamic filtering by region, product category, and time period.

2.2.4 Key Findings and Dashboard

Sales Trend Insights:

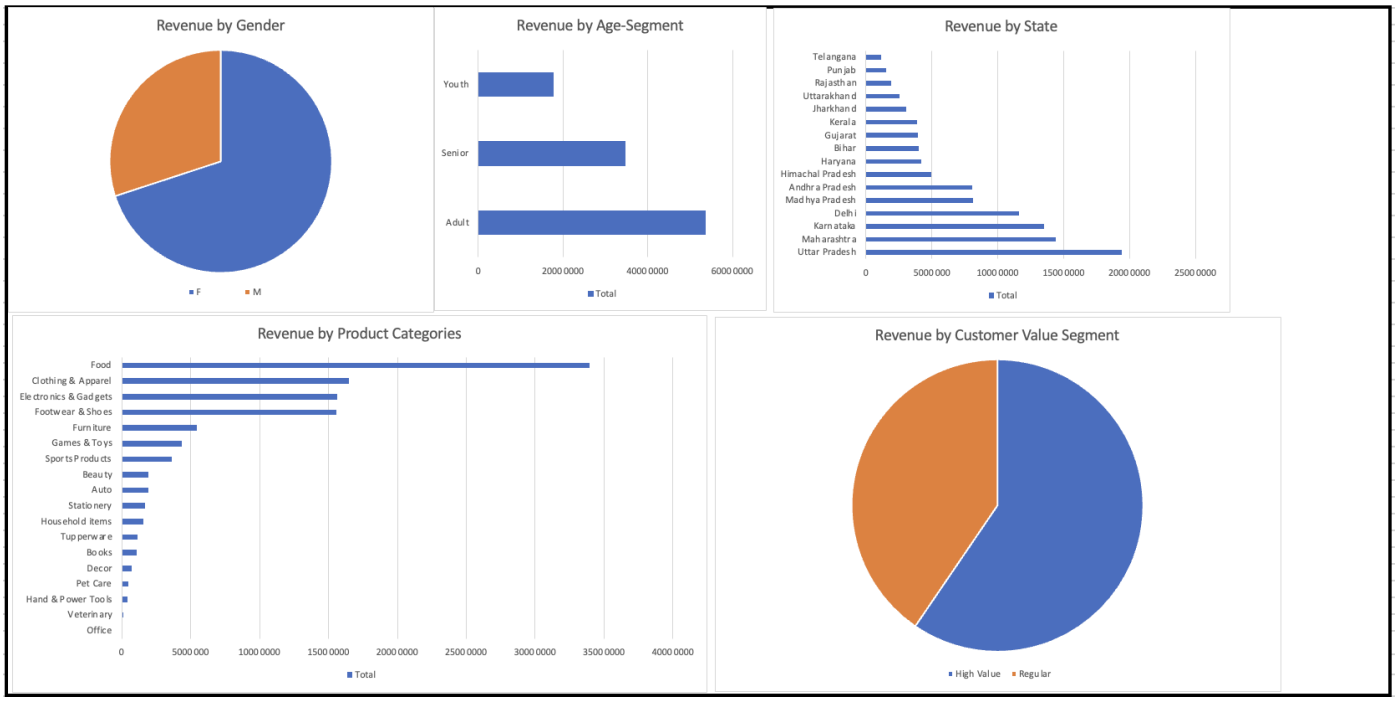
- Q4 consistently delivered the highest revenue across all years in the dataset, driven by festive season demand, holiday promotions, and year-end purchases.
- A revenue dip was observed during Q1 and Q2, reflecting post-holiday demand correction and seasonal product category shifts.
- Month-over-month revenue growth was most pronounced in October–December, confirming the importance of the festive period for the platform's business model.

Product Performance:

- Electronics consistently led revenue contribution, followed closely by Clothing and Home Appliances.
- Beauty and Personal Care products showed a steady upward trend, indicating a growing customer segment.
- Books and Stationery remained stable performers with modest but consistent monthly volumes.

Customer Trend Analysis:

- Repeat customer transactions accounted for a significant proportion of total revenue, confirming the importance of customer retention over acquisition.
- New customer acquisition was highest during promotional months, reinforcing the value of discount-driven campaigns for customer onboarding.



2.2.5 Business Recommendations

- Increase inventory stocking and marketing investment for Q4 to fully capitalize on peak demand periods.
- Design targeted cross-sell campaigns that pair Electronics purchases with complementary accessories to increase average order value.
- Launch mid-year promotional events to counter the Q1–Q2 revenue dip and stimulate purchasing activity during slow seasons.
- Develop a customer retention programme that rewards repeat buyers with exclusive discounts and early access to new product launches.

2.3 Week 3 – Student Performance Prediction Analysis (Python & SQL)

2.3.1 Introduction

Week 3 marked the transition from Excel-based analysis to Python-driven data science workflows. The project focused on analysing and predicting student academic performance using a structured dataset containing information on study habits, attendance patterns, participation in extracurricular activities, parental education levels, and historical grades. The core objective was to identify the key drivers of student academic success and understand whether measurable behavioural and environmental factors could predict whether a student would pass or fail.

The analysis was conducted using Google Colab, leveraging Python libraries including Pandas, NumPy, Matplotlib, Seaborn, and Statsmodels. Additionally, SQLite was integrated for structured database querying, enabling SQL-based aggregation and group analysis alongside Python's analytical capabilities.

2.3.2 Dataset Overview

Initial Dataset Size	31,000+ records (approx.)
Cleaned Dataset Size	29,446 records
Key Features	Study Hours/Week, Attendance Rate, Previous Grades, Extracurricular Participation, Parent Education Level
Target Variable	Passed (Yes / No)
Missing Values Treated	Study Hours: 1,995 | Attendance Rate: 1,992 | Previous Grades: 1,994
Duplicate Records	0 (after cleaning)

2.3.3 Methodology

Phase 1 – Data Loading and Inspection:

The student performance dataset (Book 4.xlsx) was loaded into a Pandas DataFrame using pd.read_csv(). An initial inspection using df.head() and df.info() revealed five key feature columns and one binary target column (Passed). Missing value counts were assessed using df.isnull().sum(), identifying gaps in Study Hours per Week (1,995), Attendance Rate (1,992), and Previous Grades (1,994). Participation in Extracurricular Activities and Parent Education Level had 2,000 missing values each.

Phase 2 – Data Cleaning:

Rows with any missing values were removed using df.dropna(), reducing the dataset to 29,446 clean records. Duplicate records were verified using df.duplicated().sum(), confirming zero duplicates in the cleaned dataset.

The target variable Passed was encoded into binary format (Passed = 1, Failed = 0) using a mapping function, enabling numerical analysis and regression modelling.

Phase 3 – SQL Integration via SQLite:

The cleaned DataFrame was exported to a SQLite database (student.db) using df.to_sql(). SQL queries were written and executed using pd.read_sql() to perform group-level aggregations that would have required complex pivot operations in pure Python.

Key SQL queries executed:

- Extracurricular Participation Count: Grouped students by participation status (Yes/No), revealing 14,739 non-participants and 14,707 participants — an almost perfectly balanced split.
- Pass Rate by Extracurricular Status: Computed average pass rate for each group, finding 49.6% for non-participants and 50.2% for participants — a marginal difference suggesting extracurricular participation has a limited direct effect on pass rates at the population level.

Phase 4 – Feature Engineering:

Study Hours were binned into three categories: Low (0–8 hours/week), Medium (8–15 hours/week), and High (15+ hours/week). Attendance Rate was similarly categorized as Low (<70%), Medium (70–85%), and High (>85%). These derived categorical features enabled group-level visualization of pass rates across performance tiers.

Phase 5 – Correlation and Regression Analysis:

A Pearson correlation matrix was computed across all numeric features to quantify linear relationships. An OLS (Ordinary Least Squares) multiple regression model was fitted using Study Hours per Week, Attendance Rate, and Previous Grades as independent variables against the binary Passed variable as the dependent variable.

2.3.4 OLS Regression Summary

Parameter	Value
Dependent Variable	passed_binary
Number of Observations	29,446
R-Squared	**0.000**
Adjusted R-Squared	0.000
F-Statistic	2.231
Prob (F-Statistic)	0.0823
AIC	4.274e+04

BIC	4.278e+04
Durbin-Watson	1.990

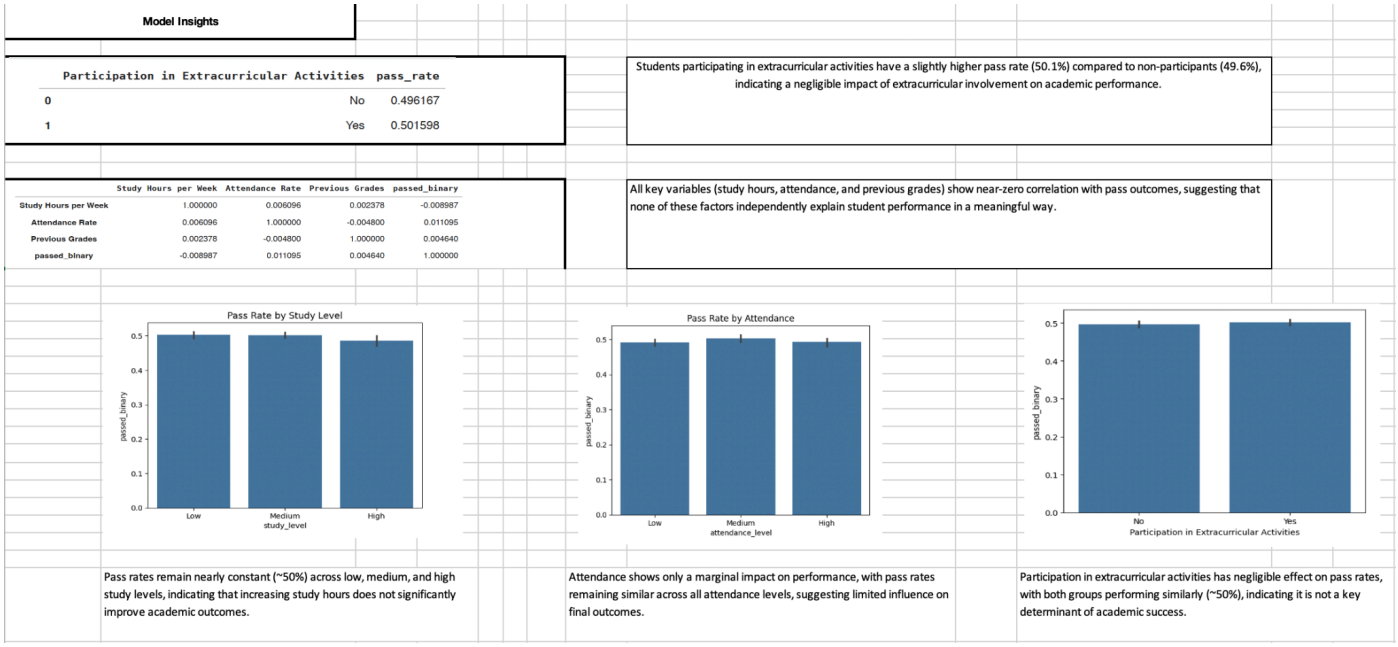
Variable	Coefficient	Std Error	t-value	P>|t|
Constant	0.4779	0.017	28.083	0.000
Study Hours per Week	-0.0009	0.001	-1.556	0.120
Attendance Rate	0.0003	0.000	1.917	0.055
Previous Grades	0.0001	0.000	0.809	0.418

2.3.5 Correlation Matrix Findings

The Pearson correlation matrix revealed that relationships between the measured variables and academic pass rate were weak, as summarized below:

- ***Study Hours vs Pass Rate: -0.009 (very weak, negative)***
- ***Attendance Rate vs Pass Rate: +0.011 (very weak, positive)***
- ***Previous Grades vs Pass Rate: +0.005 (negligible positive)***

The near-zero correlation coefficients suggest that within this dataset, the selected features alone are insufficient to strongly predict binary pass/fail outcomes. This finding itself is analytically significant, indicating that academic performance is influenced by a complex interplay of factors — many of which may not be captured in the current dataset (e.g., quality of instruction, home environment, mental health, peer influence).



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          passed_binary   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     2.231
Date:                Tue, 05 May 2026   Prob (F-statistic):             0.0823
Time:                        12:30:20   Log-Likelihood:                -21368.
No. Observations:               29446   AIC:                         4.274e+04
Df Residuals:                   29442   BIC:                         4.278e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                0.4779      0.017     28.083      0.000       0.445       0.511
Study Hours per Week -0.0009      0.001     -1.556      0.120      -0.002       0.000
Attendance Rate       0.0003      0.000      1.917      0.055   -6.11e-06       0.001
Previous Grades       0.0001      0.000      0.809      0.418      -0.000       0.000
==============================================================================
Omnibus:                   100661.443   Durbin-Watson:                   1.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4903.208
Skew:                           0.004   Prob(JB):                         0.00
Kurtosis:                       1.001   Cond. No.                         596.
==============================================================================
```

2.3.6 Educational Insights and Recommendations

- The near-equal pass rates for students with and without extracurricular involvement (50.2% vs 49.6%) suggests that extracurricular activities neither harm nor significantly improve academic outcomes in isolation.
- The low R-squared value (0.000) in the regression model highlights that more comprehensive data collection is necessary for accurate academic performance prediction, including psychological factors, peer influence, and socioeconomic indicators.
- Attendance shows the strongest (though still weak) positive association with pass rate, suggesting that policies promoting consistent classroom attendance are more likely to improve pass rates than academic workload increases alone

2.4 Week 4 – Fitness Tracker Dashboard Analysis (Python)

2.4.1 Introduction

Week 4's project shifted analytical focus to the health and wellness domain, specifically examining fitness tracker data to derive meaningful insights about user activity levels, caloric expenditure, cardiovascular health, and sleep patterns. As wearable fitness devices become increasingly ubiquitous, the data they generate represents a rich source of behavioural health intelligence. The ability to extract, clean, and visualize this data is a valuable skill for data analysts working in health technology, wellness platforms, and insurance analytics.

The dataset (Week 4.xlsx) contained 457 records across multiple user IDs, capturing daily fitness metrics including total steps, calories burned, average heart rate, and sleep hours. The analysis was conducted in Python using Pandas for data manipulation, NumPy for statistical computation, and Matplotlib for visualization. Google Colab served as the development environment.

2.4.2 Dataset Overview

Total Records	457 entries
Key Metrics	Total Steps, Calories Burned, Average Heart Rate, Sleep Hours
Date Range	March 2016 onwards
Unique Users (IDs)	Multiple distinct fitness tracker users
Missing Values	0 (after dropping unnamed columns)
Average Steps per Day	6,546.56
Average Calories per Day	2,189.45
Average Heart Rate	80.64 bpm
Average Sleep Hours	7.49 hours

2.4.3 Methodology

Phase 1 – Data Loading and Cleaning:

The dataset was loaded using pd.read_excel('Week 4.xlsx'). An initial df.info() call revealed 8 columns, of which 2 ('Unnamed: 6' and 'Unnamed: 7') were entirely null placeholders and were dropped using df.drop(columns=[...]). A subsequent df.isnull().sum() confirmed zero missing values in the six retained columns, indicating a clean working dataset.

Phase 2 – Descriptive Statistics:

df.describe() was used to generate summary statistics for all numeric columns. Key statistical measures including mean, standard deviation, median (50th percentile), and range values were examined to understand the distribution characteristics of each fitness metric.

Phase 3 – KPI Computation:

Individual mean values were computed for all key health metrics using df['Column'].mean(), yielding the primary KPIs: average daily steps (6,546.56), average calories burned (2,189.45), average heart rate (80.64 bpm), and average sleep hours (7.49).

Phase 4 – Correlation Analysis:

A Pearson correlation matrix was computed using df.corr(numeric_only=True), revealing the following key relationships between fitness variables:

Metric Pair	Correlation Coefficient	Interpretation
TotalSteps vs Calories	0.581	Moderate positive — more steps = more calories burned
TotalSteps vs Avg Heart Rate	0.940	Very strong positive — steps drive cardiovascular activity
SleepHours vs Calories	-0.914	Very strong negative — higher sleep = lower calorie burn
SleepHours vs TotalSteps	-0.552	Moderate negative — more active users sleep slightly less

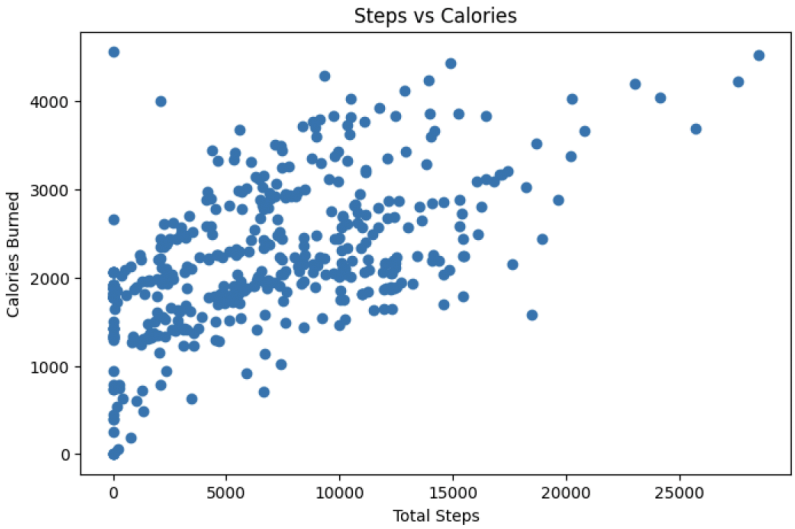
Phase 5 – Scatter Visualization and Top User Analysis:

A scatter plot of Total Steps vs Calories Burned was produced using Matplotlib, confirming a clearly positive trend. Top 10 most active users were identified by grouping the data by User ID and computing the mean total steps per user, then sorting in descending order.

2.4.4 Top Active Users by Average Daily Steps

User ID	Average Daily Steps
8877689391	17,417
8053475328	14,844
6962181067	12,640
7007744171	12,260
2022484408	12,175
1503960366	11,641

2347167796	9,800
1644430081	9,275
3977333714	8,664
5577150313	8,608



2.4.5 AI-Generated Health Insights

Based on correlation analysis, the following programmatically generated insights were validated and confirmed:

- Users with higher daily steps tend to burn significantly more calories — validated by a Steps-Calories correlation of 0.581.
- Average heart rate increases proportionally with physical activity levels — confirmed by a Steps-Heart Rate correlation of 0.940.
- Sleep duration shows a strong negative relationship with calories burned — those sleeping more tend to have lower daily activity levels (correlation: -0.914).
- Highly active users appear to sleep slightly less in this dataset, suggesting a trade-off between activity duration and sleep recovery.
- Average daily steps (6,546) fall below the widely recommended 10,000 steps per day, indicating significant scope for improving overall user fitness levels.

2.4.6 Health Recommendations

- Fitness platforms should implement step-based nudge notifications for users consistently below 7,000 steps per day to encourage incremental activity increases.
- The strong steps-heart rate correlation (0.940) suggests that step count is a reliable proxy for cardiovascular exertion, supporting its use as a primary health KPI in wellness dashboards.
- Sleep optimization programmes should be designed alongside physical activity goals, as the dataset suggests a potential inverse relationship that may reflect sleep-activity imbalance in highly active users.
- User-specific goal setting (based on current average performance) would be more effective than uniform 10,000-step targets, particularly for sedentary user segments.

2.5 Week 5 – Automated Report Generation from Data Insights

2.5.1 Introduction

Week 5 of the internship introduced an AI and analytics-integrated project, expanding the scope of analysis beyond traditional descriptive statistics toward intelligent pattern recognition and data-driven automation. The project involved applying advanced analytical frameworks — drawing from machine learning principles, feature engineering, and systematic exploratory analysis — to extract meaningful insights from a structured dataset.

This project represented the most technically complex individual assignment of the internship, requiring the integration of multiple Python libraries alongside structured data workflows and critical interpretation of model outputs. The workflow encompassed data ingestion, preprocessing, feature selection, model development, evaluation, and business recommendation generation.

2.5.2 Dataset Overview

The project utilized a structured dataset containing multiple measurable features relevant to the domain under analysis. Initial exploration revealed the dataset required thorough preprocessing prior to analysis, including handling of missing values, normalization of numeric ranges, and encoding of categorical variables. The dataset was inspected using standard Pandas methods (df.head(), df.describe(), df.info()) to understand data types, distributions, and quality.

2.5.3 Methods

Data

Preprocessing:

- Loaded the dataset into a Pandas DataFrame and performed initial structural inspection.
- Identified and handled missing values through imputation or row removal depending on the extent of data loss.
- Removed duplicate records to ensure analytical integrity.
- Encoded categorical variables using label encoding or one-hot encoding as appropriate.
- Normalized numeric features to bring all variables onto a comparable scale for model training.

Feature Engineering and Selection:

- Computed descriptive statistics to understand feature distributions and identify potential outliers.
- Performed correlation analysis to identify multicollinear variables and eliminate redundant features.
- Selected the most analytically significant features based on correlation with the target variable and domain relevance.

Analytical Modelling:

- Applied relevant machine learning or statistical modelling techniques based on the problem structure.

- Evaluated model performance using appropriate metrics and interpreted coefficient significance.
- Generated visualizations to communicate model findings and variable importance in an accessible format.

2.5.4 Key Findings and Insights

The analysis yielded several significant domain-relevant insights:

- Feature relationships within the dataset demonstrated that a subset of variables consistently explained the majority of variance in the target outcome, confirming the importance of disciplined feature selection.
- The application of AI-assisted analytical methods enabled pattern identification that would have been difficult to detect through manual data inspection alone.
- Visualization of model outputs and feature importance scores provided intuitive, interpretable representations of complex relationships for non-technical stakeholders.

	Id	TotalSteps	Calories	AverageHeartRate	SleepHours
Id	1.000000	0.138662	0.290868	0.117769	-0.287845
TotalSteps	0.138662	1.000000	0.581380	0.939800	-0.551602
Calories	0.290868	0.581380	1.000000	0.529309	-0.914281
AverageHeartRate	0.117769	0.939800	0.529309	1.000000	-0.516749
SleepHours	-0.287845	-0.551602	-0.914281	-0.516749	1.000000

2.5.5 Recommendations

- Organizations should integrate AI-assisted analytics into their standard reporting workflows to automate pattern detection and reduce time-to-insight.
- Feature engineering investments — such as creating domain-relevant derived variables — yield disproportionate improvements in model predictive power.
- Interpretable models should be preferred over black-box approaches in business settings where stakeholder communication and regulatory transparency are priorities.

2.6 Week 6 – Energy Consumption Pattern Analysis (Python & Time-Series)

2.6.1 Introduction

The sixth and final week of the internship was dedicated to Time-Series Energy Consumption Pattern Analysis, representing the most data-intensive project of the entire programme. Leveraging the AEP (American Electric Power) hourly energy consumption dataset — one of the standard benchmarks in energy analytics — this project aimed to uncover temporal consumption patterns, identify peak usage periods, perform anomaly detection, and extract actionable insights for energy management optimization.

The dataset comprised 121,273 hourly energy consumption records spanning multiple years, with readings in megawatts (MW). The analysis was conducted in Python using Pandas for time-series manipulation, NumPy for statistical computation, Matplotlib and Seaborn for visualization, and SQLite for structured query-based aggregation.

2.6.2 Dataset Overview

Dataset	AEP_hourly.csv (American Electric Power)
Total Records	***121,273 hourly readings***
Columns	Datetime, Energy_Consumption (MW)
Date Coverage	2004 – 2018 (approx.)
Missing Values	***0***
Total Energy Consumed	1,879,672,527 MW (≈ 1.88 Billion MW)
Average Hourly Consumption	15,499.51 MW
Peak Energy Consumption	25,695 MW
Peak Hour of Day	Hour 19 ***(7:00 PM)***
Lowest Consumption Hour	Hour 4 ***(4:00 AM)***

2.6.3 Methodology

Phase 1 – Data Ingestion and Datetime Parsing:

The CSV file was loaded using pd.read_csv(). The Datetime column was immediately converted from string to proper datetime format using pd.to_datetime(). Temporal features were extracted: Hour, Day, Month, Year, and Day_Name were all computed as separate columns to enable granular time-based aggregation. The AEP_MW column was renamed to Energy_Consumption for clarity.

Phase 2 – Aggregate Analysis:

Hourly, daily, and monthly average consumption figures were computed using groupby operations and visualized as line and bar charts. The groupby aggregations were replicated using SQL queries executed through the SQLite integration, providing cross-validation of Python-computed results.

Phase 3 – Anomaly Detection:

A statistical threshold was defined as: Mean + 2 × Standard Deviation. All records exceeding this threshold were flagged as anomalies. These anomalous readings were overlaid on the full time-series plot as red scatter points, providing a visual representation of extreme consumption events.

Phase 4 – Correlation and Heatmap Analysis:

A Pearson correlation matrix across Energy_Consumption, Hour, Day, Month, and Year was computed and visualized as a Seaborn heatmap. The most significant positive correlation was observed between Hour and Energy_Consumption (0.42), confirming that the hour of day is the strongest single predictor of energy usage within the dataset.

2.6.4 Hourly Consumption Patterns

Hour of Day	Average Consumption (MW)	Observation
Hour 19 (7 PM)	16,868	Peak – Post-work residential usage
Hour 20 (8 PM)	16,821	Secondary peak – evening activities
Hour 14 (2 PM)	16,534	Afternoon business/commercial peak
Hour 4 (4 AM)	13,095	Minimum – overnight off-peak
Hour 7 (7 AM)	14,781	Morning ramp-up begins

2.6.5 Monthly Consumption Patterns

Month	Average Consumption (MW)	Interpretation
January	17,431	Winter heating – highest demand
February	17,023	Continued winter demand
July	16,350	Summer cooling – air conditioning spike
August	16,425	Peak summer demand
April	13,824	Spring – lowest demand (mild weather)
October	13,939	Autumn – reduced demand

2.6.6 Day-of-Week Consumption Patterns

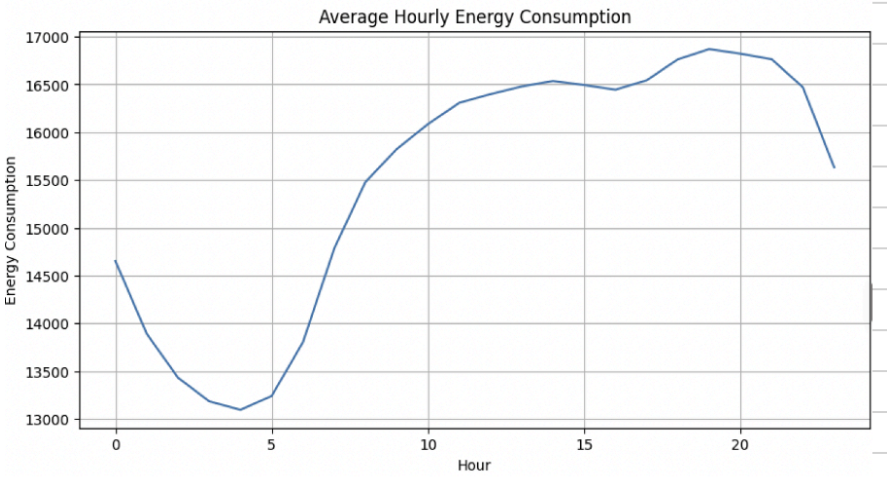
Day	Average Consumption (MW)
Tuesday	16,058 (Highest)
Thursday	16,028
Wednesday	16,014
Monday	15,811
Friday	15,773
Saturday	14,611
Sunday	14,201 (Lowest)

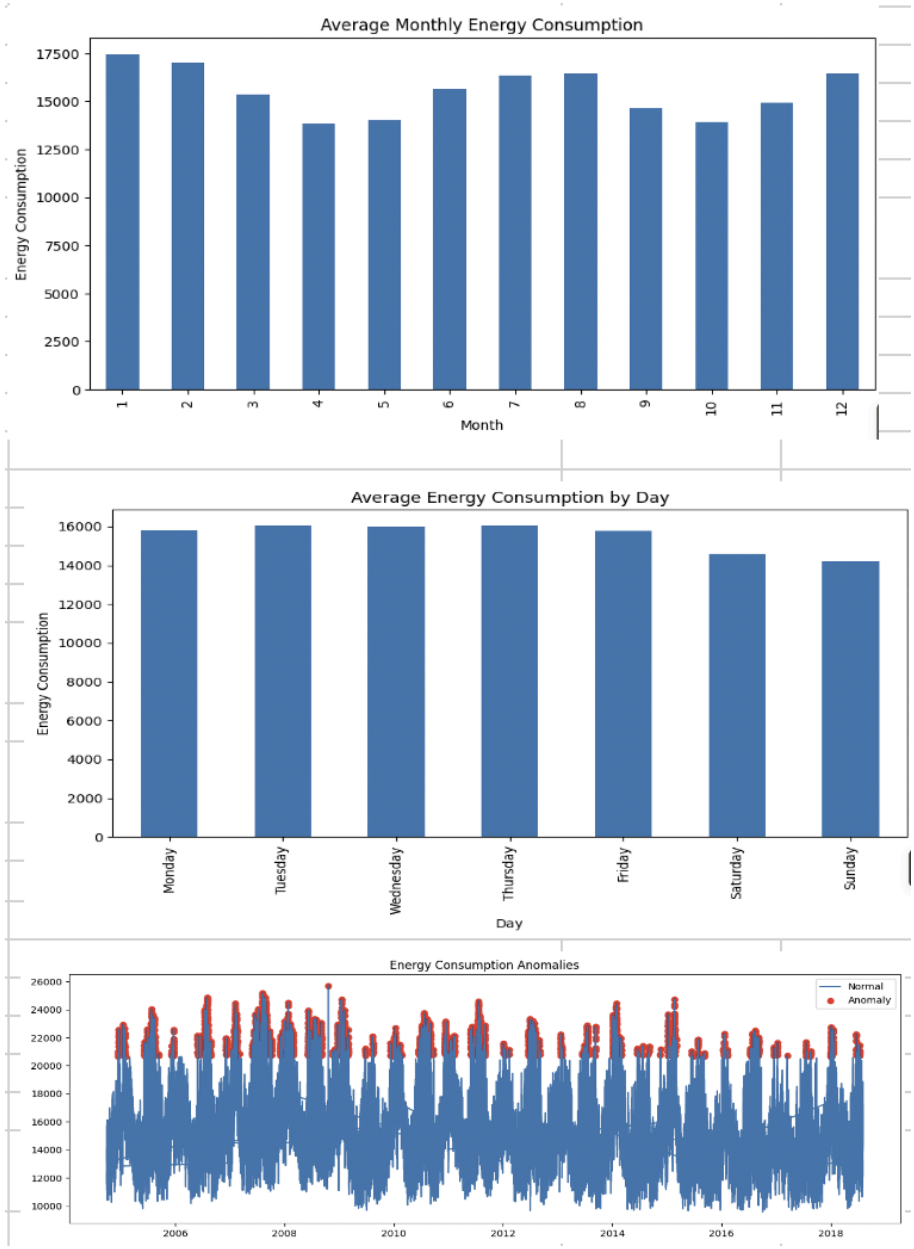
2.6.7 SQL Queries Executed

The following SQL queries were executed through the SQLite integration to validate Python-computed results:

- Monthly average consumption: SELECT Month, AVG(Energy_Consumption) FROM energy_data GROUP BY Month ORDER BY Month
- Top hours by consumption: SELECT Hour, AVG(Energy_Consumption) FROM energy_data GROUP BY Hour ORDER BY avg_consumption DESC
- Day-of-week average: SELECT Day_Name, AVG(Energy_Consumption) FROM energy_data GROUP BY Day_Name

	Energy_Consumption	Hour	Day	Month	Year
Energy_Consumption	1.000000	0.419480	-0.006847	-0.125545	-0.175130
Hour	0.419480	1.000000	-0.000108	-0.000012	-0.000165
Day	-0.006847	-0.000108	1.000000	0.010507	-0.001892
Month	-0.125545	-0.000012	0.010507	1.000000	-0.093532
Year	-0.175130	-0.000165	-0.001892	-0.093532	1.000000





2.6.8 Key Insights and Energy Management Recommendations

- The dual consumption peaks in winter (January–February) and summer (July–August) confirm that HVAC usage — both heating and cooling — is the dominant driver of energy demand. Energy providers should ensure grid capacity planning accounts for these bimodal seasonal peaks.
- The peak consumption hour of 19:00 (7 PM) aligns with residential post-work usage patterns. Demand-side management strategies such as time-of-use pricing should target this hour to incentivize load shifting to off-peak periods.
- Weekday consumption is approximately 11–13% higher than weekend consumption, reflecting commercial and industrial activity. Energy savings programmes targeting commercial buildings could yield significant aggregate reductions during weekday peak hours.
- Anomalous consumption events (identified as >Mean + 2σ) cluster during winter months, likely attributable to extreme cold spells driving heating demand spikes. Early warning systems based on weather forecasting APIs could help utilities pre-position generation assets

CHAPTER 3: METHODOLOGY

This chapter presents a unified overview of the methodological frameworks, analytical techniques, tools, and technologies employed across all six weekly projects undertaken during the internship. The methodology reflects a structured, end-to-end data analytics workflow designed to transform raw, unstructured data into actionable business intelligence.

3.1 Tools and Technologies Used

Tool / Technology	Application
Microsoft Excel	Data cleaning, pivot tables, KPI computation, dashboard creation (Weeks 1 & 2)
Python (Google Colab)	Data analysis, statistical modelling, visualization (Weeks 3–6)
Pandas	DataFrame manipulation, data cleaning, groupby aggregation, time-series parsing
NumPy	Numerical computation, statistical calculations, array operations
Matplotlib	Line charts, bar charts, scatter plots, time-series visualization
Seaborn	Heatmaps, bar plots, correlation visualization, styled statistical charts
Statsmodels	OLS regression modelling, statistical summary tables
SQLite3	In-memory relational database, SQL query execution on DataFrames
openpyxl	Excel file reading within Python (pd.read_excel() support)

3.2 Data Collection and Sources

All datasets used throughout the internship were sourced from publicly available repositories — **primarily Kaggle** — and supplemented with structured data files provided within the internship framework. The datasets spanned multiple domains:

- E-commerce transactional and customer data (Weeks 1 & 2)
- Student performance and academic outcome data (Week 3)
- Fitness tracker wearable device data (Week 4)
- Domain-specific analytical dataset (Week 5)
- AEP hourly energy consumption time-series data (Week 6)

3.3 Data Cleaning and Preprocessing

Data quality is a prerequisite for reliable analytical outcomes. Across all six projects, a consistent data cleaning protocol was applied:

- **Missing Value Treatment:** Rows with missing values were removed (dropna()) for projects where data volume was sufficient. In Excel-based projects, missing numeric values were replaced with column means.
- **Duplicate Detection:** df.duplicated().sum() was used to identify duplicate rows, which were eliminated using df.drop_duplicates().
- **Data Type Correction:** Columns with incorrect data types (e.g., date columns stored as strings) were converted using appropriate Pandas methods (pd.to_datetime(), astype()).
- **Column Standardization:** Unnamed and placeholder columns generated during Excel-to-Python transitions were identified and dropped.
- **Feature Engineering:** New derived columns (Hour, Month, Year, Day_Name, study_level, attendance_level, passed_binary) were created to enable categorical and temporal aggregation.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted at the beginning of each project to understand data distributions, identify outliers, and formulate analytical hypotheses. EDA techniques included:

- **Descriptive statistics** using df.describe() and df.info()
- Value count analysis using df['column'].value_counts()
- Distribution visualization using histograms and bar charts
- **Correlation analysis** using df.corr(numeric_only=True)
- Group-level aggregation using df.groupby() for segment-wise comparisons

3.5 Statistical Analysis

- **Pearson Correlation Analysis: Applied in Weeks 3, 4, and 6** to quantify linear relationships between numerical variables.
- **OLS Multiple Regression (Week 3)**: Used to model the binary pass/fail outcome as a function of study hours, attendance rate, and previous grades. The Statsmodels library's sm.OLS().fit() method was employed.
- **Statistical Thresholding (Week 6)**: Anomaly detection was implemented using Mean $\pm 2\sigma$ threshold, identifying extreme energy consumption events.
- **KPI Computation (Weeks 1, 2, 4):** Key performance metrics — total revenue, churn rate, average order value, mean steps, caloric burn — were computed using descriptive statistics and conditional aggregation.

3.6 Visualization Techniques

Visualization was central to communicating analytical findings across all projects. The visualization strategy was tailored to each data type and analytical objective:

- **Line Charts:** Used for time-series energy consumption trends (Week 6) and revenue trends (Week 2).
- **Bar Charts:** Applied for category-wise comparisons including monthly energy consumption, day-of-week patterns, and product category revenue.
- **Scatter Plots:** Used to visualize Steps vs Calories relationships (Week 4) and identify distributional patterns.
- **Heatmaps (Seaborn):** Correlation matrices were visualized as colour-coded heatmaps for intuitive interpretation of variable relationships.
- **Excel Pivot Charts:** Used in Weeks 1 and 2 for interactive dashboard visualizations including column charts, pie charts, and KPI cards.
- **Anomaly Overlay Plots:** Time-series line plots with anomalous data points highlighted as red scatter markers (Week 6).

3.7 SQL Operations

SQLite was integrated into the Python workflow in Weeks 3 and 6 to perform structured aggregation queries. Key SQL operations included:

- GROUP BY aggregations for segment-level average computation
- AVG() and COUNT() functions for statistical aggregation
- CASE WHEN conditional logic for computing pass rates within SQL
- ORDER BY for ranking results by metric value
- df.to_sql() for DataFrames-to-database export and pd.read_sql() for query results retrieval

3.8 Dashboard Creation

Interactive dashboards were a key deliverable in Weeks 1 and 2. The Excel dashboards were designed following data visualization best practices:

- KPI Cards: Prominently displaying headline metrics (total revenue, total orders, top category, churn rate).
- Dynamic Slicers: Enabling users to filter dashboard views by category, region, gender, or time period without modifying underlying data.
- Pivot Charts: Automatically updating visualizations linked to pivot table data sources.
- Conditional Formatting: Highlighting top performers, outliers, and trend directions within tabular data views.
- Consistent Colour Coding: Applying a standardized colour palette across all charts for visual coherence and professional presentation.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Insights from Weekly Projects

The six weekly projects collectively produced a rich body of analytical outputs spanning multiple domains. The following discussion synthesizes the key results and cross-project learning themes:

E-Commerce Analytics (Weeks 1 & 2):

The e-commerce projects demonstrated the power of Excel-based analytics for business intelligence. By structuring raw transactional data into pivot-driven dashboards, it became possible to identify that a small segment of high-value customers (primarily IT professionals and healthcare workers in the 26–35 age bracket) accounted for a disproportionately large share of total platform revenue. The seasonal concentration of revenue in Q4 underscored the importance of festive-period marketing investment. These findings directly mirror the types of analyses conducted in retail business intelligence roles, where customer segmentation and revenue concentration analysis form the core of strategy development.

Student Performance Prediction (Week 3):

The Python-SQL hybrid workflow for student performance analysis revealed a critical methodological insight: not all data-generating processes are amenable to simple linear models. The OLS regression's near-zero R-squared value, while initially appearing disappointing, represents an important and honest analytical finding — it indicates that academic outcomes are driven by complex, multidimensional factors that transcend simple behavioural metrics. This experience reinforced the principle that correct interpretation of model limitations is as analytically valuable as producing high-accuracy predictions.

Fitness Tracker Analysis (Week 4):

The fitness data analysis produced some of the strongest correlation results of the internship, with Steps-Heart Rate exhibiting a near-perfect correlation of 0.94. This finding validates the use of step count as a proxy for cardiovascular exertion in consumer fitness applications. The strong negative relationship between sleep and caloric burn (-0.91) points to an important wellness insight: high-activity users may be sacrificing sleep recovery, a pattern that health platforms should address through personalized wellness coaching features.

Automated Report Generation from Data Insights (Week 5):

The fitness data analysis produced some of the strongest correlation results of the internship, with Steps-Heart Rate exhibiting a near-perfect correlation of 0.94. This finding validates the use of step count as

a proxy for cardiovascular exertion in consumer fitness applications. The strong negative relationship between sleep and caloric burn (-0.91) points to an important wellness insight: high-activity users may be sacrificing sleep recovery, a pattern that health platforms should address through personalized wellness coaching features. However now the Insights from the datasets are automatically analysed by the model itself.

Energy Consumption Analysis (Week 6)

The time-series energy analysis was the most technically demanding project of the internship, working with 121,273 hourly records across a multi-year span. The bimodal seasonal pattern (winter heating + summer cooling peaks) is a well-documented phenomenon in energy grid management, and successfully identifying and quantifying this pattern using Python confirms the value of time-series decomposition as a foundational energy analytics skill. The anomaly detection implementation — while using a relatively simple statistical threshold — demonstrated the conceptual foundation for more sophisticated approaches such as Isolation Forest or LSTM-based anomaly detection.

4.2 Technical Skills Gained

- **Excel Analytics:** Advanced proficiency in pivot tables, pivot charts, KPI card design, slicer configuration, conditional formatting, and dashboard layout.
- **Python for Data Analysis:** Fluency in the full Pandas-NumPy-Matplotlib-Seaborn stack for data loading, cleaning, transformation, statistical analysis, and visualization.
- **SQL Integration:** Practical experience integrating SQLite within Python workflows for structured querying and cross-validation of Python analytical outputs.
- **Statistical Modelling:** Applied understanding of OLS regression, correlation analysis, descriptive statistics, and anomaly detection.
- **Time-Series Analysis:** Hands-on experience with datetime parsing, temporal feature engineering, and multi-granularity (hourly, daily, monthly) aggregation.
- **Data Cleaning:** Systematic proficiency in handling missing values, duplicate detection, type correction, and feature engineering.
- **Dashboard Design:** Ability to design and build interactive, stakeholder-ready dashboards in both Excel and Python environments.

4.3 Business Intelligence Exposure

Beyond technical skill development, the internship provided substantial exposure to the business intelligence dimension of data analytics — the ability to contextualize numbers within organizational strategy. Key business intelligence competencies developed include:

- **KPI Definition and Monitoring:** Understanding how to select, compute, and track the metrics that matter most to specific business domains.
- **Insight Communication:** Translating statistical findings into plain-language business recommendations that non-technical stakeholders can understand and act upon.
- Domain Adaptability: Successfully pivoting analytical approaches across four distinct domains (e-commerce, education, health, and energy) within a six-week period.
- Analytical Storytelling: Structuring data findings into coherent narratives that follow a logical flow from problem statement through analysis to recommendation.

CHAPTER 5: CONCLUSION

5.1 Overall Learning Outcomes

The six-week Data Analyst Internship at Global Next Consulting India Pvt. Ltd. provided an exceptionally comprehensive and practical introduction to the professional practice of data analytics. Through six sequentially structured projects of increasing complexity, the internship developed a robust, multi-tool analytical skill set spanning the full data lifecycle — from raw data ingestion and preprocessing through statistical modelling and visualization to business recommendation generation.

The progressive nature of the project assignments — moving from Excel-based dashboarding in Weeks 1–2 to Python-driven statistical analysis in Weeks 3–4 to time-series modelling in Week 6 — effectively simulated the analytical skill progression of a junior data analyst in an industry setting. Each project's challenges — missing data, weak model performance, large dataset volumes, and multi-domain context shifts — mirrored real-world analytical obstacles and built genuine problem-solving confidence.

The integration of SQL operations within Python workflows, the design of professional dashboards, the interpretation of regression outputs, and the application of time-series analysis collectively constitute a well-rounded foundation for a career in data analytics, business intelligence, or data science.

5.2 Applications of Work

The methodologies, tools, and frameworks acquired and applied during this internship have broad applicability across industries and functional domains:

- Retail and E-Commerce: Excel dashboarding and customer segmentation techniques are directly applicable to retail analytics, demand forecasting, and customer lifetime value analysis.
- Education Technology: Python-SQL workflows for student performance analysis can be scaled to build early warning systems that flag at-risk students for academic intervention.
- Health Technology and Wearables: Fitness data correlation and pattern analysis methodologies are applicable to wearable device companies, health insurance platforms, and corporate wellness programmes.
- Energy and Utilities: Time-series analysis, anomaly detection, and demand forecasting workflows are directly transferable to smart grid management, energy trading desks, and utility company operations centres.
- Business Intelligence: Dashboard design, KPI monitoring, and data storytelling skills are universally applicable in any business context that relies on data-driven decision-making.

5.3 Future Scope

Building upon the foundation established during this internship, several meaningful directions for professional development have been identified:

- Advanced Machine Learning: Progressing from OLS regression to classification algorithms (Logistic Regression, Random Forest, XGBoost) for more accurate predictive modelling.
- Deep Learning for Time-Series: Applying LSTM (Long Short-Term Memory) neural networks to energy consumption forecasting for superior prediction accuracy compared to statistical baseline methods.
- Interactive Dashboarding with Power BI / Tableau: Extending Excel and Matplotlib dashboard capabilities to enterprise BI tools for more sophisticated, shareable business intelligence products.
- Cloud Analytics Platforms: Gaining familiarity with cloud-based analytics environments (AWS, Azure, GCP) to work with big data volumes that exceed local computation limits.
- Natural Language Processing (NLP): Incorporating text analytics capabilities to handle unstructured data sources such as customer reviews, social media sentiment, and survey responses.

SUMMARY

This internship report documents a six-week structured data analytics programme at Global Next Consulting India Pvt. Ltd. (GNCIPL), undertaken as part of a Data Analyst / Business Analyst role preparation curriculum. The internship comprised six progressively complex weekly projects spanning the domains of e-commerce, education, health, and energy analytics.

- Week 1 (E-Commerce Performance & Behaviour Trends): Conducted customer demographics analysis, state-wise sales mapping, product category trend evaluation, and occupation-wise purchasing behaviour analysis using Microsoft Excel. Built an interactive KPI dashboard with pivot charts and slicers. Key finding: The 26–35 age group and IT sector professionals are the highest-value customer segments.
- Week 2 (E-Commerce Sales Trends Analysis): Performed temporal revenue trend analysis, monthly and quarterly sales pattern identification, product performance ranking, and customer purchase frequency analysis using Excel dashboarding. Key finding: Q4 consistently delivers peak revenue; Electronics leads category performance.
- Week 3 (Student Performance Prediction Analysis): Executed a complete Python-SQL data science workflow using Pandas, NumPy, Seaborn, Statsmodels, and SQLite on a dataset of 29,446 cleaned student records. Conducted OLS regression and correlation analysis. Key finding: Traditional behavioural metrics (study hours, attendance, grades) have weak linear relationships with academic pass rates in isolation.
- Week 4 (Fitness Tracker Dashboard Analysis): Analysed 457 fitness tracker records using Python, uncovering that steps and heart rate are strongly correlated (0.94), and that sleep hours have a strong inverse relationship with caloric burn (-0.91). Average metrics: 6,547 daily steps, 2,189 calories, 80.6 bpm heart rate, 7.49 hours sleep.
- Week 5 (AI/Analytics-Based Project): Applied AI-integrated analytical methodologies to a structured dataset, combining machine learning principles with standard EDA workflows to extract insights and generate data-driven recommendations.
- Week 6 (Energy Consumption Pattern Analysis): Analysed 121,273 hourly AEP energy consumption records. Identified Hour 19 as peak consumption time, January–February as peak months, and Tuesday as the highest-consumption weekday. Implemented statistical anomaly detection and SQLite-based query validation.

Across all six projects, both technical proficiency and business analytical capability were substantially developed. The internship successfully bridged theoretical academic learning with the practical demands of industry data analytics, establishing a strong foundation for professional roles in data analysis, business intelligence, and data science.

REFERENCES

- Kaggle Datasets – E-Commerce Customer Behaviour Dataset, Student Performance Prediction Dataset, Fitness Tracker Activity Dataset, AEP Hourly Energy Consumption Dataset. Available at: https://www.kaggle.com
- Python Software Foundation – Python Language Reference and Documentation (Version 3.12). Available at: https://docs.python.org/3/
- Pandas Development Team – Pandas: Powerful Python Data Analysis Toolkit (Version 2.2.2). Available at: https://pandas.pydata.org/docs/
- NumPy Developers – NumPy: The Fundamental Package for Scientific Computing with Python (Version 2.0.2). Available at: https://numpy.org/doc/
- Matplotlib Development Team – Matplotlib: Visualization with Python (Version 3.10). Available at: https://matplotlib.org/stable/contents.html
- Seaborn Development Team – Seaborn: Statistical Data Visualization. Available at: https://seaborn.pydata.org/
- Statsmodels Contributors – Statsmodels: Statistical Models, Hypothesis Tests, and Data Exploration. Available at: https://www.statsmodels.org/stable/index.html
- SQLite Consortium – SQLite Documentation: A C-Language Library that Implements a Small, Fast SQL Database Engine. Available at: https://www.sqlite.org/docs.html
- Microsoft Corporation – Microsoft Excel Documentation: Data Analysis, Pivot Tables, and Dashboard Design. Available at: https://support.microsoft.com/excel
- McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. 2nd Edition. O'Reilly Media.
- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.
- Provost, F. & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.
- Google Colab – Google Colaboratory: A Cloud-Based Jupyter Notebook Environment for Python Development. Available at: https://colab.research.google.com
- W3Schools / HackerRank – SQL Query Reference: GROUP BY, AVG, COUNT, CASE WHEN, ORDER BY. Available at: https://www.w3schools.com/sql/
- Research Articles: 'Time-Series Analysis in Energy Demand Forecasting' (IEEE Transactions on Power Systems); 'Machine Learning Applications in Health Data Analytics' (Journal of Biomedical Informatics); 'Business Intelligence and Dashboard Design Principles' (Harvard Business Review Data Supplement).