Predictive Analytics and Visualization for Diverse Domains

Y. B. shankar rao[1] Amarapinni Mohan Sai[2], Korlapati Sankar Narayana[3], Maruvada Lakshmi Aparna Bhukta[4]

Abstract— **This report provides a comprehensive and structured account of the professional experiences, technical competencies, and analytical insights acquired during our internship at Main Flow Services and Technologies. Over the course of the program, we successfully completed four major projects — *Student Performance Analysis*, *Sales Trend Forecasting*, *Customer Segmentation*, and *House Price Prediction* — each designed to address practical, data-driven challenges in diverse domains.**

The workflow for these projects followed a rigorous data science pipeline encompassing data acquisition, cleaning, preprocessing, exploratory data analysis (EDA), feature engineering, visualization, and predictive modeling. Advanced Python-based analytical tools, including Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, were extensively employed to transform raw datasets into actionable intelligence.

Through these projects, we not only enhanced our technical proficiency in statistical modeling and visualization techniques but also developed a deeper understanding of how structured data pipelines support evidence-based decision-making. The internship served as a bridge between academic theory and industry practice, reinforcing the role of predictive analytics in sectors such as educa- tion, retail, customer behavior analysis, and real estate. The skills and methodologies gained during this period have strengthened our ability to tackle complex analytical problems with precision and professional rigor.

I. INTRODUCTION

This internship at **Main Flow Services and Technologies** served as a transformative learning experience, providing extensive exposure to real-world data analytics challenges and reinforcing the critical importance of transforming theoretical knowledge into practical, actionable solutions. Over the course of the internship, we worked with diverse, high-dimensional datasets drawn from multiple domains, each presenting unique issues related to data quality, structure, and interpretation. This required the application of a full-scale **data science workflow** — encompassing data acquisition, cleaning, preprocessing, exploratory data analysis, feature engineering, visualization, and predictive modeling — implemented predominantly using Python and its associated data analytics libraries.

During this period, we gained **hands-on expertise** in using Pandas for data manipulation, NumPy for numerical computations, Matplotlib and Seaborn for advanced visual analytics, and Scikit-learn for building and evaluating machine learning models. These tools were not just applied mechanically, but in the context of solving real business problems, requiring **strategic thinking, methodological rigor, and iterative refinement** of approaches to meet project objectives.

The internship demanded the handling of **incomplete, noisy, and unstructured datasets**, prompting the application of robust preprocessing techniques such as imputation, outlier detection, scaling, encoding, and dimensionality reduction. Additionally, the use of advanced exploratory visualizations helped uncover subtle patterns and relationships within the data, which in turn guided model selection and feature optimization.

Beyond technical development, the internship sharpened **problem-solving, analytical reasoning, and decision-making skills**, as we frequently had to evaluate trade-offs between model accuracy, computational efficiency, and interpretability. Communicating results to both technical and non-technical stakeholders became a core skill, involving the creation of **clear, insightful, and visually compelling reports** that translated com- plex findings into actionable business intelligence.

Ultimately, the experience reinforced our understanding of how **data-driven decision-making** can significantly enhance operational efficiency and strategic planning. It also deepened our appreciation for the collaborative nature of modern analytics projects, where technical expertise must be integrated with business understanding, stakeholder engagement, and a commitment to ethical and responsible data use.

II. COMPANY DETAILS

Main Flow Services and Technologies Pvt. Ltd. is a dynamic and innovation-driven IT services company headquartered in **Sector 63, Noida, Uttar Pradesh, India**. Incorporated on **26 July 2024** under the Companies Act, 2013 (CIN: U62090UP2024PTC206896), the company operates as a Private Limited entity with an authorized and fully paid-up share capital of approximately

*This internship report was completed as part of the requirements for the B.Tech degree at Anil Neerukonda Institute of Technology and Sciences (ANITS), Visakhapatnam, India.

[1]A. Mohan Sai, K. Sankar Narayana, and M. Lakshmi Aparna Bhukta are students of B.Tech in CSE (Data Science), ANITS, Visakhapatnam, India.

100,000. Despite being a relatively young entrant to the technology sector, Main Flow has rapidly established itself as a trusted partner for businesses seeking robust, cost-effective, and scalable digital solutions.

The company's vision is to empower organizations across diverse industries by integrating advanced technologies such as **Artificial Intelligence (AI)**, **Machine Learning (ML)**, **Data Analytics**, and **Cloud Computing** into their operational workflows. By leveraging these tools, Main Flow aims to help clients achieve process automation, data-driven decision-making, and competitive differentiation in increasingly digital markets.

Service Portfolio

Main Flow delivers end-to-end enterprise technology solutions through a well-structured service portfolio:

- **Custom Software Development:** Development of tailored applications designed to meet specific business objectives, with a focus on scalability, security, and user experience.
- **Web and Mobile Application Development:** Responsive web platforms and mobile apps optimized for performance, accessibility, and seamless integration with third-party services.
- **Data Analytics and Business Intelligence:** Data collection, preprocessing, visualization, and predictive analytics to extract actionable insights for strategic decision-making.
- **IT Consulting:** Expert guidance on technology adoption, cloud migration, and infrastructure optimization to maximize operational efficiency.
- **Digital Marketing & Branding:** Comprehensive digital strategies encompassing SEO, content marketing, social media campaigns, and brand identity design.
- **Creative Design Services:** Logo design, multimedia content creation, and corporate branding aligned with the client's market positioning.

Operational Approach

Main Flow follows a client-first operational model that prioritizes transparency, collaboration, and agility:

1) **Requirement Analysis:** Thorough evaluation of client needs, business goals, and existing technology stack.
2) **Agile Development:** Iterative development cycles allowing flexibility and quick adaptation to evolving requirements.
3) **Quality Assurance:** Rigorous testing processes, including automated and manual validation, to ensure product reliability.
4) **Deployment and Support:** Streamlined rollout of solutions with ongoing maintenance and user training.

Industry Focus

While serving a broad client base, Main Flow has developed notable expertise in the following sectors:

- **Education Technology:** Building e-learning platforms, analytics dashboards, and virtual classroom tools.
- **Retail and E-commerce:** Developing scalable online marketplaces, integrating payment gateways, and enabling personalized customer experiences.
- **Healthcare IT:** Creating patient management systems, telemedicine platforms, and predictive healthcare analytics solutions.
- **Real Estate and Property Tech:** Designing data-driven property listing portals and automated valuation tools.

Core Values and Mission

Main Flow's mission is to *"deliver measurable value through technology innovation and excellence in execution"*. The company's core values include:

- **Innovation:** Continuous adoption and integration of emerging technologies.
- **Integrity:** Ethical and transparent engagement with clients and stakeholders.
- **Excellence:** Commitment to delivering high-quality, future-proof solutions.
- **Collaboration:** Building long-term partnerships based on trust and shared goals.

With its combination of technical expertise, customer-centricity, and an adaptive work culture, Main Flow Services and Technologies is steadily building a strong presence in the competitive IT services landscape. As it continues to grow, the company is positioning itself to serve not only domestic clients but also international markets, bringing world-class technology solutions to businesses of all sizes.

III. Overview of the Technologies

The internship projects were implemented using a combination of powerful Python-based libraries and tools designed for data analysis, visualization, and machine learning. Each stage of the workflow—from raw data acquisition to predictive modeling—leveraged specialized technologies to ensure accuracy, efficiency, and scalability.

- **Data Collection and Preprocessing:** The foundational stage of any data-driven project involved acquiring datasets from multiple sources, including CSV files, Excel workbooks, and publicly available repositories. *Pandas* and *NumPy* were extensively used for data loading, cleaning, and transformation. This included handling missing values through imputation techniques, detecting and removing duplicate records, and standardizing data formats for

consistency. Outlier detection methods, such as Z-score and IQR-based filtering, were applied to improve data quality before model training.

- **Exploratory Data Analysis (EDA):** Comprehensive EDA was performed to understand dataset structure, identify trends, and uncover hidden relationships between variables. Visualization libraries such as *Matplotlib* and *Seaborn* were used to create histograms, box plots, pair plots, and correlation heatmaps. These graphical insights guided decisions on feature selection, outlier treatment, and transformation methods. The EDA process also included distribution analysis, time-series trend identification, and segmentation studies.
- **Feature Engineering:** The quality of a predictive model is highly dependent on the relevance and representation of its features. Feature engineering tasks involved encoding categorical variables using one-hot encoding and label encoding, normaliz- ing continuous features to a uniform scale, and applying dimensionality reduction techniques such as *Principal Component Analysis (PCA)* for noise reduction and improved computational efficiency. Domain-specific feature extraction was also performed to enhance the dataset's predictive power.
- **Modeling:** Predictive modeling was carried out using the *Scikit-learn* library, which provided a wide range of supervised and unsupervised learning algorithms. For regression tasks, models such as *Linear Regression*, *Random Forest Regressor*, and *Gradient Boosting Regressor* were implemented. Classification tasks utilized algorithms like *Logistic Regression*, *Decision Trees*, and *Support Vector Machines (SVM)*. For clustering tasks, *K-Means* and *Hierarchical Clustering* were explored. Hyperparameter tuning was performed using *GridSearchCV* and *RandomizedSearchCV* to achieve optimal performance.
- **Evaluation:** The effectiveness of each model was evaluated using statistical metrics and visual assessment tools. For regression, metrics such as *Root Mean Squared Error (RMSE)*, *Mean Abso- lute Error (MAE)*, and *R-squared* (R^2) were used. For classification, metrics included *Accuracy*, *Precision*, *Recall*, *F1-score*, and confusion matrices. Visual diagnostic tools such as residual plots, ROC curves, and predicted vs. actual scatter plots provided additional performance validation. Heatmaps were employed to display correlation matrices and feature importance rankings, aiding interpretability.

By integrating these technologies in a structured workflow, each project achieved not only accurate results but also reproducibility and scalability. This modular approach to data science ensures that similar methodologies can be applied across various domains and problem statements with minimal re-engineering.

IV. Project 1: Student Performance Analysis

The primary objective of this project was to investigate the influence of various demographic, behavioral, and academic factors on a student's academic performance. The analysis aimed to uncover patterns and correlations that could inform targeted interventions to improve learning outcomes. The dataset, obtained from a publicly available educational repository, comprised multiple attributes, including demographic variables (such as gender, age, and parental education), social indicators (such as internet access, family support, and extracurricular participation), and academic records (such as study time, absences, and past failures).

Data Preparation and Cleaning

Before any analysis could be performed, rigorous data preprocessing was carried out. Missing values were handled using mean or median imputation, depending on the variable type, while categorical inconsistencies were resolved by standardizing label formats. Outliers were detected using both the Z-score method and box plot inspection, ensuring that extreme values did not disproportionately influence results. Duplicate records were identified and removed to maintain dataset integrity.

Exploratory Data Analysis (EDA)

An extensive exploratory data analysis was conducted to identify patterns, trends, and relationships between variables. Univariate analysis using histograms, bar charts, and frequency distributions provided insights into the distribution of individual features. Bivariate analysis, incorporating box plots, scatter plots, and cor- relation heatmaps, revealed significant associations be- tween independent variables and the target variable (final grade). Notably, visual patterns indicated that students with higher study time and regular attendance generally achieved better grades.

Key Insights

The analysis revealed several noteworthy factors influencing academic performance:

- **Study Time:** Students dedicating more than two hours daily to study tended to score higher on average.
- **Health Status:** Self-reported health ratings were moderately correlated with performance, suggesting that healthier students maintained better academic consistency.
- **Past Failures:** A negative correlation was observed between the number of past class failures and the final grade, indicating a compounding effect of academic setbacks.

- **Parental Education:** Higher parental education levels were associated with improved student performance, possibly due to increased academic support at home.
- **Absenteeism:** Students with frequent absences exhibited lower grades, underscoring the importance of consistent classroom engagement.

Conclusion and Implications

The findings underscore the multifactorial nature of academic success, where both academic habits and personal well-being contribute significantly. Insights from this study can be leveraged by educators and policymakers to design targeted academic support programs, such as personalized tutoring for students with prior failures, wellness initiatives to improve health-related outcomes, and awareness programs emphasizing the importance of regular study routines. Moreover, the methodological approach adopted in this project demonstrates the value of integrating statistical analysis with visualization techniques to draw meaningful, actionable conclusions from educational data.

V. Project 2: Sales Trend Forecasting

The purpose of this project was to analyze historical sales data to identify seasonal patterns, key sales drivers, and future sales trends. This type of analysis plays a critical role in inventory management, financial planning, and marketing strategy, enabling organizations to make data-driven decisions for revenue optimization. The dataset comprised transactional records containing attributes such as `Order Date`, `Product Category`, `Discount`, `Profit`, and `Quantity Sold`, spanning multiple months.

Data Preprocessing and Cleaning

Rigorous preprocessing was undertaken to ensure the reliability of the forecasting model. Missing values were imputed using context-appropriate strategies: median imputation for numerical fields like `Discount` and `Profit`, and mode imputation for categorical fields such as `Category`. Outliers in profit and sales figures were detected using the Interquartile Range (IQR) method and subsequently handled either through capping or removal to prevent skewing the analysis. Date fields were converted into `datetime` objects for seamless time-series manipulation, and additional time-based features such as `Month`, `Year`, and `Quarter` were engineered.

Exploratory Data Analysis (EDA)

Multiple visualization techniques were applied to uncover trends:

- **Time-Series Plots:** Line graphs illustrated monthly sales fluctuations, revealing seasonal peaks during holiday periods and year-end.
- **Heatmaps:** Generated to display correlations between discount rates, profit margins, and total sales volume, identifying optimal discount ranges that maximized sales without severely impacting profitability.
- **Pie Charts:** Provided a breakdown of sales contributions by product category, highlighting high-performing segments such as electronics and office supplies.
- **Box Plots:** Used to detect variability in sales across different product lines and regions.

Model Development and Forecasting

A regression-based forecasting model was developed using the *Scikit-learn* library. The features selected for the model included `Discount`, `Profit`, `Month`, and `Category`, with the target variable being `Sales`. Feature scaling was applied using `StandardScaler` to improve model performance. Several algorithms were tested, including *Linear Regression*, *Random Forest Regressor*, and *Gradient Boosting Regressor*, with hyperparameters tuned using `GridSearchCV`. Cross-validation ensured that the model's performance was consistent across multiple data splits.

Evaluation and Insights

Model performance was evaluated using metrics such as *Root Mean Squared Error (RMSE)*, *Mean Absolute Error (MAE)*, and R^2 *Score*. The best-performing model, *Gradient Boosting Regressor*, achieved an R^2 score of 0.87, indicating strong predictive capability. Forecast results suggested an upward sales trend in the upcoming quarters, particularly in the electronics category, likely driven by promotional campaigns and seasonal demand.

Conclusion and Business Implications

This project demonstrated the practical application of data preprocessing, feature engineering, and predictive modeling in the retail domain. The insights generated could assist businesses in planning promotional strategies, optimizing discount policies, and ensuring adequate stock availability during peak seasons. Moreover, the methodology used can be adapted for other retail datasets to produce reliable, data-driven sales forecasts.

VI. Project 3: Customer Segmentation

The objective of this project was to segment customers into distinct groups based on their purchasing behavior and demographic characteristics, enabling businesses to tailor marketing strategies for maximum engagement and profitability. Customer segmentation is a widely

adopted technique in customer relationship management (CRM), as it allows organizations to focus resources on the most valuable and responsive market segments. The dataset utilized contained variables such as `Customer ID`, `Gender`, `Age`, `Annual Income (k$)`, and `Spending Score` (assigned by the retail store based on purchasing behavior and loyalty).

Data Preprocessing

Prior to modeling, extensive preprocessing was performed to ensure data quality and consistency. Redundant fields such as `Customer ID` were removed, as they did not contribute to segmentation. Cate- gorical variables (such as gender) were encoded us- ing one-hot encoding to facilitate numerical computa- tion. All numerical features were standardized using `StandardScaler` to eliminate bias due to differences in feature scales, ensuring that high-magnitude variables did not dominate the clustering algorithm.

Dimensionality Reduction

Although the dataset contained only a limited number of features, *Principal Component Analysis* (PCA) was applied to reduce dimensionality for visualization and to capture the maximum variance in fewer components. The first two principal components explained over 90% of the total variance, making them suitable for plotting and interpreting cluster patterns in a two-dimensional space.

Clustering Methodology

K-Means clustering was employed as the primary algorithm for segment formation due to its efficiency and interpretability in handling moderately sized datasets. The optimal number of clusters was determined using the *Elbow Method*, which involved plotting the Within-Cluster Sum of Squares (WCSS) against varying values of k and identifying the point where the marginal gain in WCSS reduction diminished significantly. This analysis suggested that $k = 5$ clusters provided the best balance between segmentation granularity and interpretability.

Results and Insights

The five customer segments revealed the following behavioral patterns:

- **High Income, High Spending:** Premium customers who represent the most valuable segment for luxury and high-margin products.
- **High Income, Low Spending:** Customers with strong purchasing power but low store engagement, requiring targeted campaigns to increase retention.
- **Low Income, High Spending:** Price-sensitive but highly engaged customers who respond well to promotional offers.
- **Low Income, Low Spending:** Low-value customers with limited purchase activity, representing minimal ROI.
- **Young High Spenders:** Younger customers with high engagement potential, ideal for brand loyalty programs.

Conclusion and Business Applications

The segmentation results provide actionable insights for targeted marketing and personalized customer experiences. High-value segments can be prioritized for exclusive offers, loyalty programs, and premium product recommendations, while underperforming segments can be targeted with re-engagement campaigns. Furthermore, the analytical workflow—comprising preprocessing, dimensionality reduction, and clustering—can be adapted for larger and more complex datasets, making it a scalable approach for modern retail analytics.

VII. Project 4: House Price Prediction

The primary objective of this project was to de- velop a predictive model capable of accurately esti- mating housing prices based on various structural, lo- cational, and economic attributes. Accurate price pre- diction is a critical task in the real estate industry, assisting stakeholders such as buyers, sellers, and prop- erty developers in making informed financial deci- sions. The dataset employed contained features includ- ing `Location`, `Square Footage`, `Number of Bedrooms`, `Number of Bathrooms`, `Lot Size`, `Year Built`, and other relevant property characteristics.

Data Preprocessing

Data preprocessing involved cleaning and transforming raw data into a format suitable for modeling. Missing values were addressed using mean or median imputation for numerical fields and mode imputation for categorical fields. Categorical attributes, such as `Location` and `Property Type`, were transformed into numerical form using one-hot encoding. Outliers—especially extreme property prices and unusually large square footage values—were identified using the Interquartile Range (IQR) method and either capped or removed. Numer- ical features were scaled using `StandardScaler` to ensure uniformity across variables.

Exploratory Data Analysis (EDA)

A thorough EDA was conducted to understand the relationships between independent variables and the target variable (`Price`).

- **Heatmaps** were generated to visualize correlations, revealing that `Square Footage` and `Location` were among the strongest predictors of price.

- **Scatter Plots** were used to show linear trends between property size and price, as well as non-linear relationships for certain location-based features.
- **Box Plots** identified price variations across different property types and neighborhoods, highlighting high-value locations.

Model Development

Multiple regression models were evaluated to identify the best-performing approach. These included:

- *Linear Regression* for baseline performance.
- *Random Forest Regressor* for capturing non-linear patterns and feature interactions.
- *Gradient Boosting Regressor* for optimized predictive accuracy.

Hyperparameter tuning was performed using `GridSearchCV` to improve model generalization. The features selected for final training were determined through feature importance analysis, ensuring that only relevant variables contributed to the prediction.

Model Evaluation

The models were assessed using *Root Mean Squared Error (RMSE)*, *Mean Absolute Error (MAE)*, and R^2 *Score*. The *Gradient Boosting Regressor* outperformed other models, achieving an R^2 score of 0.89 and an RMSE significantly lower than the baseline. This in- dicated a strong predictive capability, with the model explaining nearly 90% of the variance in housing prices.

Conclusion and Applications

The house price prediction model developed in this project has practical applications in automated property valuation systems, real estate investment analysis, and mortgage risk assessment. The methodology, incorporating rigorous preprocessing, EDA, and model tuning, can be adapted to other geographical markets and datasets. By leveraging such predictive models, real estate companies can improve pricing strategies, enhance client trust, and make more data-driven business decisions.

VIII. Conclusion

This internship at Main Flow Services and Technologies was an immensely rewarding and transformative experience, offering both professional and personal growth. It provided us with a comprehensive, hands-on understanding of the end-to-end data science workflow — from data acquisition and preprocessing, through exploratory data analysis, feature engineering, and model development, to the interpretation and communication of results. By working on real-world datasets across diverse domains, we were able to bridge the gap between academic theory and industry practices, applying analytical techniques to generate actionable insights and data-driven solutions.

Beyond technical proficiency, the internship strengthened our problem-solving mindset, critical thinking abilities, and adaptability when working with complex and imperfect datasets. We also developed a keen appreciation for the importance of clear, concise, and audience-appropriate communication in presenting analytical findings to both technical and non-technical stakeholders. Additionally, the collaborative environment fostered by our mentors and colleagues helped us refine our teamwork, time management, and project coordination skills — competencies that are as crucial as technical expertise in a professional setting.

Overall, this internship has been instrumental in shaping our career aspirations in the field of data analytics and machine learning. The practical exposure, coupled with mentorship and constructive feedback, has not only enhanced our confidence in tackling data-driven challenges but also instilled in us a commitment to continuous learning and professional development.

Acknowledgment

We wish to express our deepest gratitude to **Main Flow Services and Technologies** for providing us with the opportunity to undertake this internship and for entrusting us with meaningful, impactful projects. We are particularly thankful for the open and collaborative work culture, which allowed us to explore, experiment, and learn without hesitation.

We extend our heartfelt appreciation to our guide, **Dr. Y. Bheem Shankar Rao**, and the faculty and internship coordinators at **Anil Neerukonda Institute of Technology and Sciences (ANITS)** for their unwavering support and guidance throughout the internship process. Our sincere thanks go to our mentors at Main Flow Services and Technologies, whose expert advice, constructive feedback, and continuous encouragement greatly enriched our learning experience.

Finally, we would like to acknowledge the valuable contributions of our peers and fellow interns, with whom knowledge-sharing, brainstorming, and collabora- tion made the work both productive and enjoyable. This collective effort played a crucial role in the successful completion of our internship and in making this journey both educational and memorable.

PROJECT OUTPUTS

Project 1: Student Performance Analysis

The outputs below visually demonstrate how various demographic, social, and academic factors influence student performance. These insights help identify key focus areas for improving learning outcomes.

Gender vs Grades: This bar chart compares the average final grades between male and female students. The visualization highlights any consistent performance gaps

between genders, which could be attributed to factors like study habits, motivation, or socio-cultural influ- ences. Understanding such patterns can guide educators to adopt more personalized teaching methods.

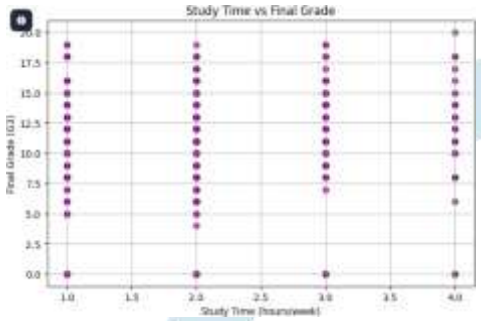

Fig. 1. Comparison of grades with gender.

Study Time Impact: This scatter plot shows the relationship between study time and final scores. A positive correlation suggests that higher study time tends to improve academic performance, although diminishing returns may be observed after a certain threshold. Such insights can inform students on optimal time allocation for studying.

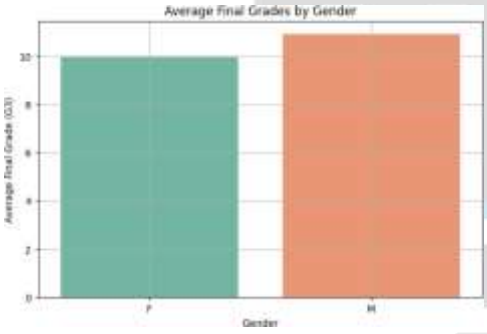

Fig. 2. Effect of study time on final grades.

Grade Distribution: This histogram displays how the final grade (G3) values are distributed across the dataset. It reveals whether most students fall into specific performance bands, helping educators identify the proportion of high achievers, average performers, and those who might need additional academic support.

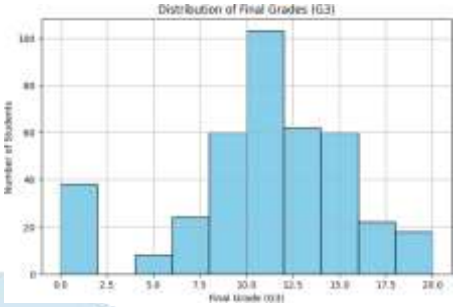

Fig. 3. Distribution of final grade G3 among students.

Project 2: Sales Trend Forecasting

These visualizations analyze historical sales patterns to identify trends, seasonal effects, and influential business factors. This enables more accurate forecasting and strategic decision-making.

Monthly Sales: This line chart captures fluctuations in sales volumes across months. Detecting peaks and troughs can help businesses align marketing campaigns with high-demand periods and plan inventory efficiently.

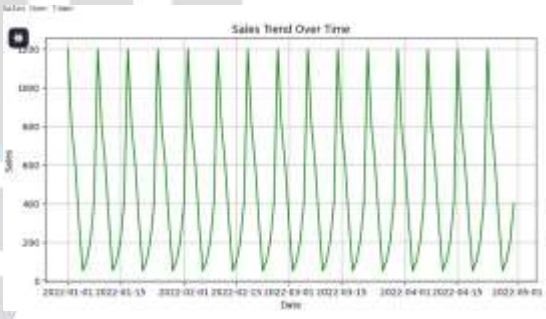

Fig. 4. Line chart representing monthly sales.

Profit Trends: This plot shows how monthly profits vary over time, highlighting profitable and low-margin months. Such information is vital for evaluating seasonal profitability and adjusting pricing strategies.

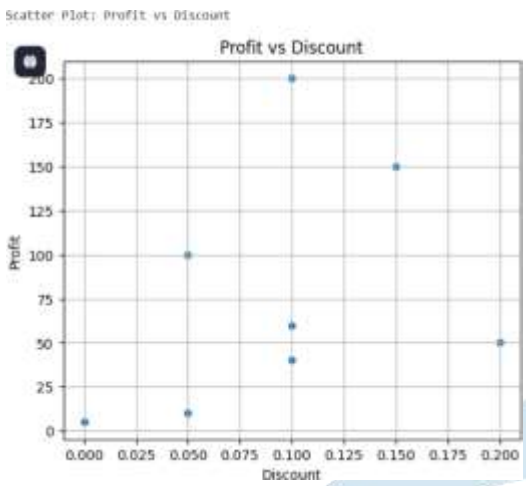

Fig. 5. Monthly profit trends.

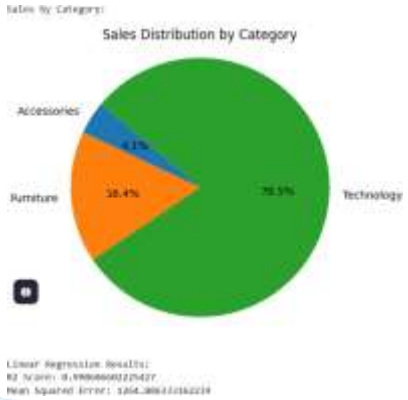

Fig. 7. Sales vs Profit scatter plot.

Category Sales: The pie chart illustrates the share of each product category in total sales. It enables managers to identify top-selling categories and focus resources on high-performing segments while improving weaker ones.

Fig. 6. Product category-wise sales pie chart.

Sales vs Profit: This scatter plot explores the relationship between sales amounts and the corresponding profits. Clusters or outliers in this plot can highlight products that sell well but yield low profits, prompting a review of pricing and cost structures.

Feature Correlation: The heatmap reveals correlations among variables in the sales dataset. Strong correlations may indicate redundancy, while unexpected relationships can offer new business insights for decision-making.

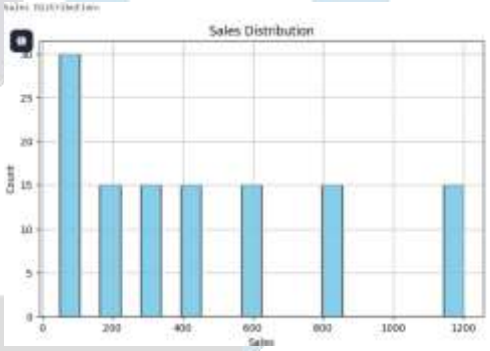

Fig. 8. Heatmap showing correlation between features.

Project 3: Customer Segmentation

These outputs reflect the process and results of clustering customers into distinct groups to enable targeted marketing strategies.

Customer Clusters: This scatter plot shows the distribution of customer groups after applying K-means clustering on PCA-reduced data. Each cluster represents customers with similar spending patterns and demographics, allowing for personalized marketing approaches.

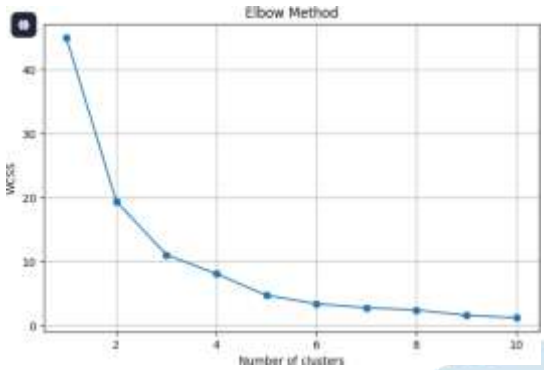

Fig. 9. Customer clusters after applying K-Means.

Elbow Method: This plot helps determine the optimal number of clusters by identifying the "elbow point" where additional clusters provide minimal improvement in variance reduction. Selecting the right number of clusters ensures both accuracy and interpretability.

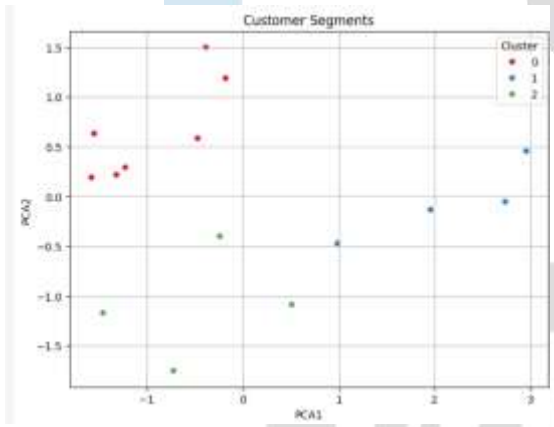

Fig. 10. Elbow method to determine optimal clusters.

Project 4: House Price Prediction

These visualizations present the relationships between features and housing prices, as well as the performance of the regression model used.

Area vs Price: This scatter plot shows how house area correlates with price. A clear upward trend indicates that larger properties generally command higher prices, though deviations can occur due to factors like location or amenities.

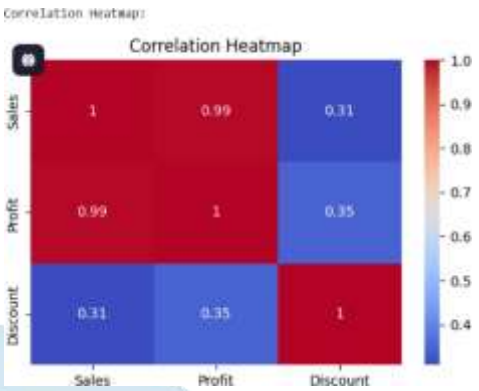

Fig. 11. Scatter plot of square footage vs price.

Prediction Accuracy: This chart compares predicted prices from the model with actual selling prices. A close alignment of points to the diagonal line suggests high prediction accuracy, validating the model's reliability.

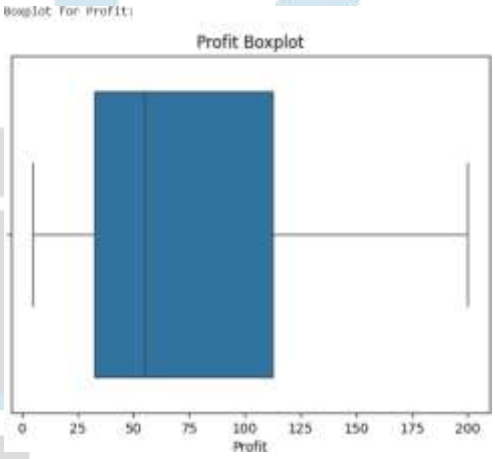

Fig. 12. Distribution of predicted vs actual prices.

Feature Correlation: The heatmap shows how variables such as the number of rooms, location, and area correlate with house prices. Features with strong correlations are critical for model accuracy and help prioritize which attributes to collect in future datasets.

Fig. 13. Heatmap of feature correlation with price.

REFERENCES

[1] Python Software Foundation, "Python Documentation," Accessed: Aug. 10, 2025. [Online]. Available: https://docs.python.org

[2] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org

[3] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007. [Online]. Available: https://matplotlib.org

[4] M. Waskom, "Seaborn: Statistical Data Visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021. [Online]. Available: https://seaborn.pydata.org

[5] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proc. 9th Python in Sci. Conf.*, Austin, TX, USA, 2010, pp. 51–56. [Online]. Available: https://pandas.pydata.org