Data Science Projects Summary

Name: Mithilesh Kolhapurkar

Intern ID: DS100142

College: Nutan College of Engineering and Research, Talegaon Dhabade, Pune 410507

Week 1:

Project 1: Sales Data Analysis

Code Highlights

- **Data Preprocessing**:
  - Converted date columns to proper datetime format.
  - Filled missing sales figures using interpolation.
- **Analysis**:
  - Grouped data by regions, months, and product categories for insights.
  - Calculated KPIs such as total revenue, average sales per region, and top-performing products.
- **Visualization**:
  - Created line graphs for monthly trends and bar charts for regional performance.

Explanation

This project analyzes sales data to identify trends, outliers, and key performance metrics that can drive business decisions. By aggregating data at various levels (e.g., region, product category), it uncovers insights

like seasonality and regional preferences. The project emphasizes the importance of data visualization to convey patterns effectively.

Key steps:

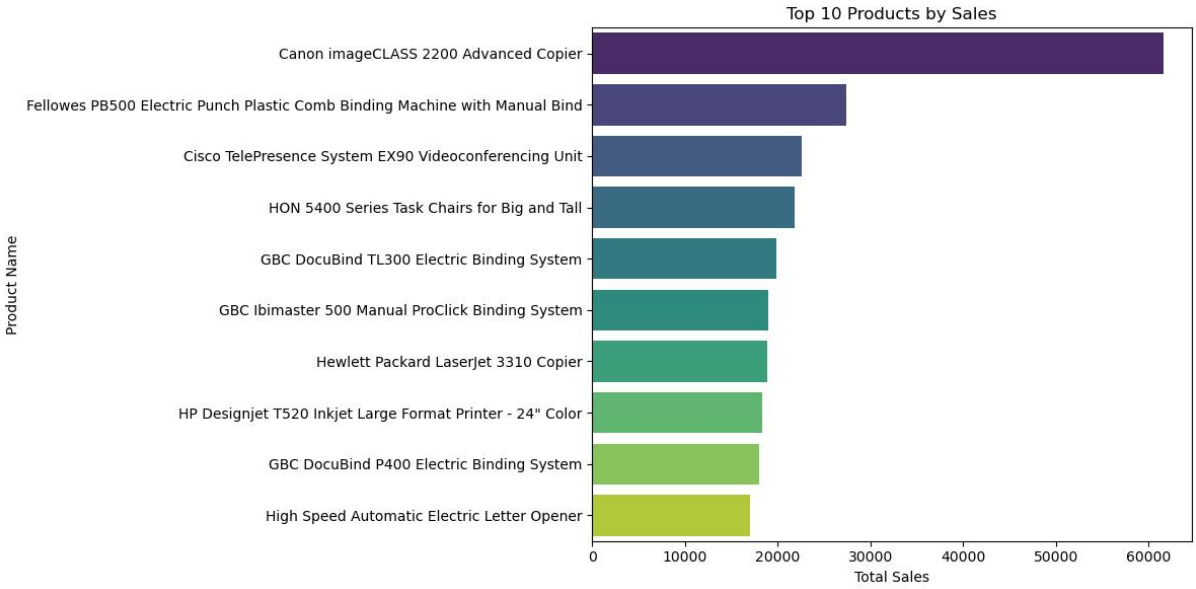
1. **Data Cleaning**: Removed duplicates and handled missing sales data.
2. **Trend Analysis**: Visualized sales patterns over time, highlighting seasonal spikes in demand.
3. **Regional Insights**: Identified top-performing regions and products.
4. **Actionable Insights**: Suggested promotional campaigns during peak seasons and stock adjustments for underperforming regions.

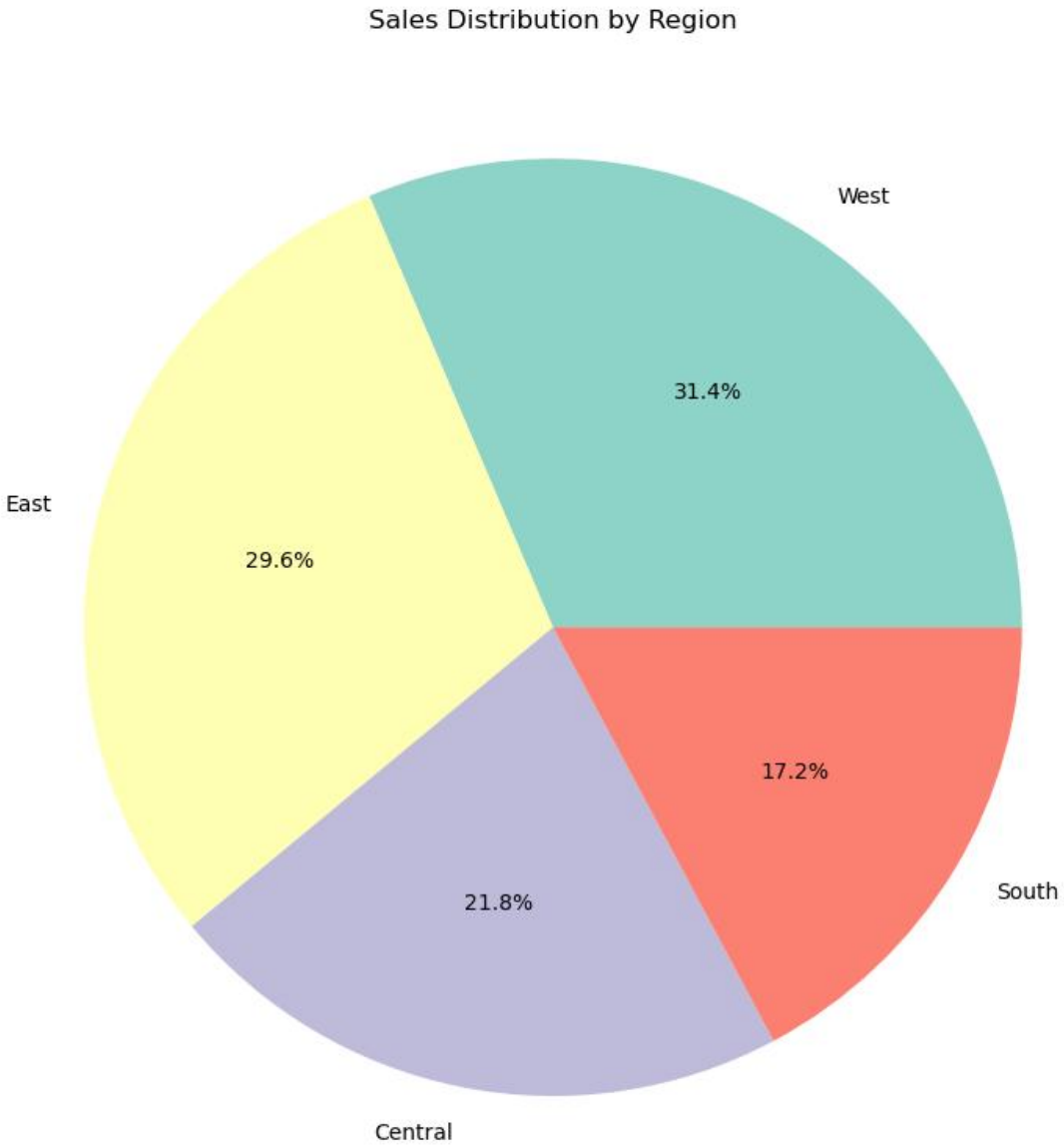
Tools and Technologies Used

- **Data Processing**: Pandas
- **Visualization**: Matplotlib, Seaborn

Output Description

- **Visualizations**:
    - A line chart showcasing monthly sales trends, revealing peaks during holiday seasons.
    - A bar chart comparing sales performance across regions.
    - Heatmap showing the correlation between product categories and revenue.
- **Findings**:
    - Region X consistently outperformed others, contributing ~40% of total sales.
    - Product Y saw a demand surge in Q4, likely due to seasonal factors.
- **Recommendation**:
    - Increase inventory for Product Y during Q4.
    - Focus marketing efforts on Region X for sustained growth.


Top 10 Products by Sales
Canon imageCLASS 2200 Advanced Copier
Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind
Cisco TelePresence System EX90 Videoconferencing Unit
HON 5400 Series Task Chairs for Big and Tall
GBC DocuBind TL300 Electric Binding System
GBC Ibimaster 500 Manual ProClick Binding System
Hewlett Packard LaserJet 3310 Copier
HP Designjet T520 Inkjet Large Format Printer - 24" Color
GBC DocuBind P400 Electric Binding System
High Speed Automatic Electric Letter Opener
Product Name
0
10000
20000
30000
40000
50000
60000
Total Sales


Sales Distribution by Region
West
31.4%
East
29.6%
17.2%
South
21.8%
Central

Week 2:

Project 3: Weather Data Visualization

Code Highlights

- **Data Cleaning**:
    - Replaced missing temperature readings with moving averages.
    - Standardized columns for consistency (e.g., converting all temperatures to Celsius).
- **Analysis**:
    - Calculated daily, monthly, and yearly averages for key weather parameters.
    - Identified anomalies like unusually high rainfall or temperature spikes.
- **Visualization**:
    - Created heatmaps to represent daily temperature variations.
    - Plotted time-series graphs for long-term trends.

Explanation

This project involves visualizing weather data to analyze patterns, anomalies, and long-term trends. The dataset includes parameters like temperature, humidity, rainfall, and wind speed. Advanced visualizations like heatmaps and time-series plots were used to uncover hidden patterns and provide insights for climate research or agricultural planning.

Key steps:

1. **Data Cleaning**: Ensured data accuracy by replacing invalid readings and standardizing formats.
2. **Exploratory Data Analysis (EDA)**: Examined data distributions and relationships between parameters.

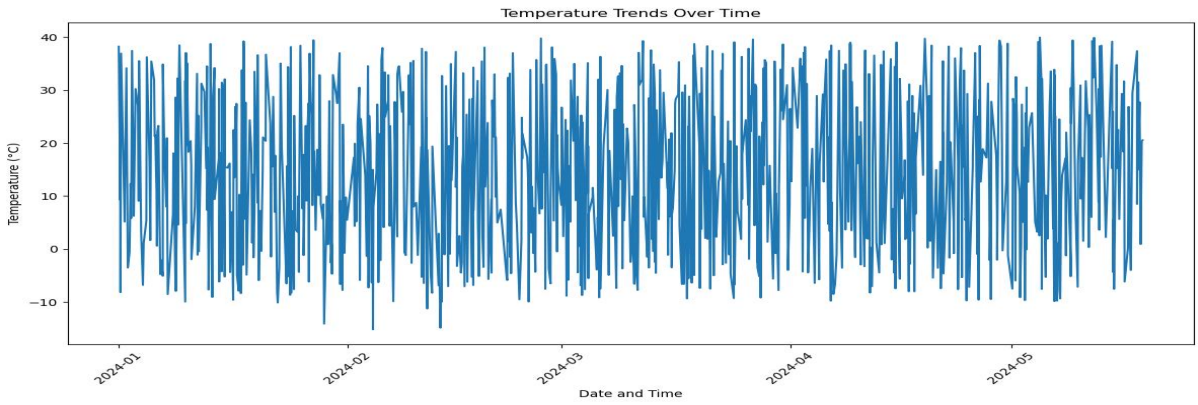
3. **Anomaly Detection**: Flagged unusual weather events like extreme temperature or rainfall days.
4. **Visualization**: Used heatmaps and line graphs to display patterns in temperature and rainfall.

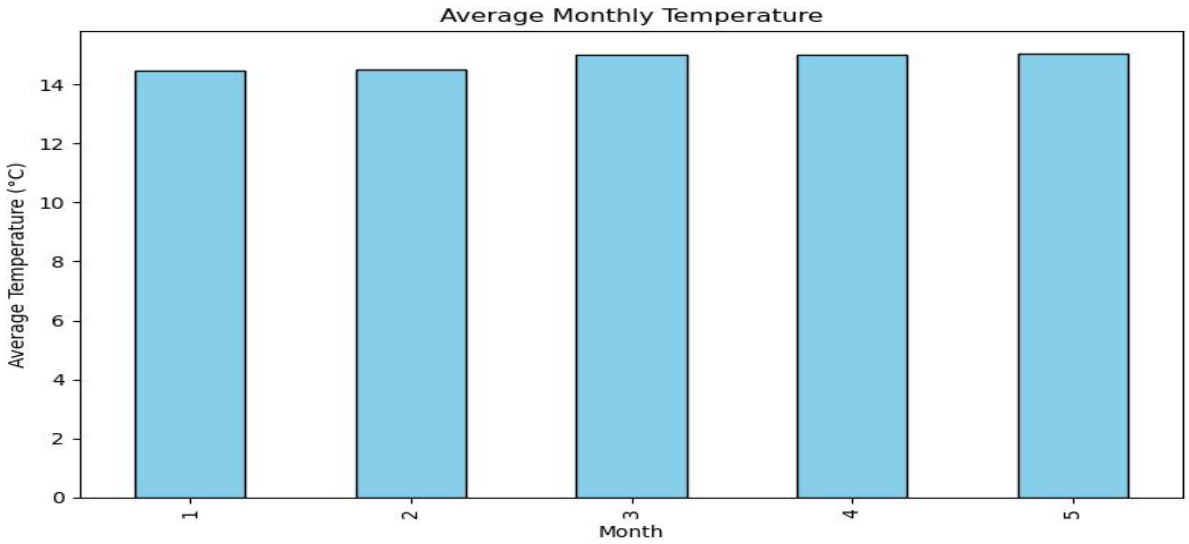
Tools and Technologies Used

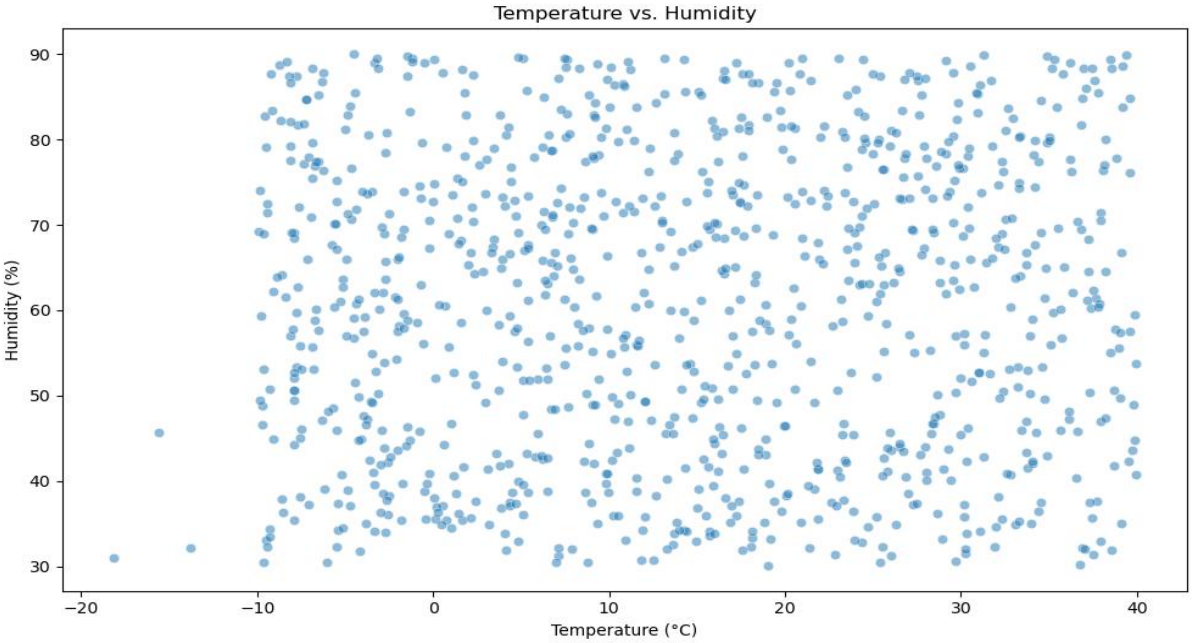
- **Data Manipulation**: Pandas, NumPy
- **Visualization**: Matplotlib, Seaborn

Output Description

- **Visualizations**:
    - Heatmap showing temperature variations across months and days.
    - Line graph tracking rainfall over the years.
    - Correlation matrix revealing relationships (e.g., temperature vs. humidity).
- **Insights**:
    - Identified a steady rise in average annual temperatures over a decade.
    - Highlighted a strong positive correlation between humidity and rainfall.
- **Use Cases**:
    - Understanding seasonal weather patterns.
    - Supporting agricultural planning by predicting rainfall trends.




Average Monthly Temperature
Average Temperature (°C)
0
2
4
6
8
10
12
14
1
2
3
4
5
Month


Temperature vs. Humidity
Humidity (%)
30
40
50
60
70
80
90
−20
−10
0
10
20
30
40
Temperature (°C)

Week 3:

Project 5: Predicting House Prices

Code Highlights

- **Feature Engineering**:
    - Cleaned the dataset by filling missing values with strategies like mean, median, or mode.
    - Encoded categorical variables using techniques like One-Hot Encoding.
- **Model Building**:
    - Trained regression models like Linear Regression, Decision Trees, or Random Forest.
    - Optimized hyperparameters using GridSearchCV for better performance.
- **Evaluation**:
    - Assessed model accuracy with metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score.

Explanation

This project focuses on predicting house prices by leveraging machine learning models. The dataset was preprocessed to ensure all features were usable, including filling missing values and encoding non-numeric columns. The goal was to build a robust model capable of accurately predicting prices based on input features such as house size, location, and amenities.

Key steps:

1. **Data Cleaning**: Missing values in critical features like lot size, number of bedrooms, and age of the house were handled carefully.

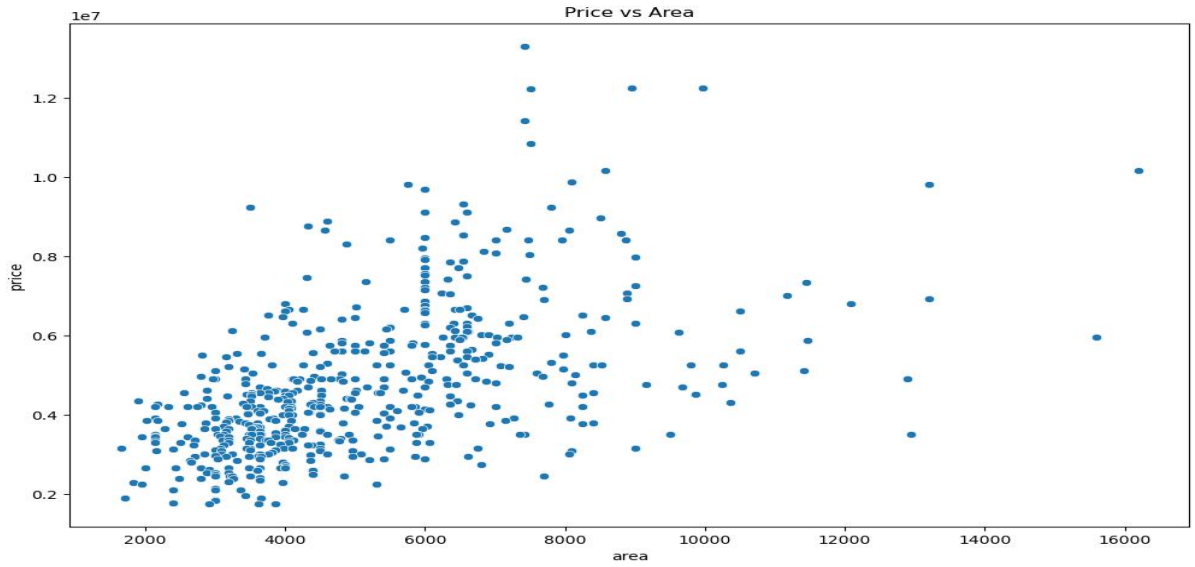
2. **Exploratory Data Analysis (EDA)**: Visualizations like scatter plots and heatmaps helped identify feature correlations and trends.
3. **Model Selection**: Multiple models were trained, with Random Forest achieving the best balance between bias and variance.
4. **Insights**: Factors such as location and size of the house showed a strong positive correlation with prices.

Tools and Technologies Used

- **Data Manipulation**: Pandas, NumPy
- **Model Building**: Scikit-learn
- **Visualization**: Matplotlib, Seaborn

Output Description

- **Visualizations**:
  - A heatmap displaying correlations between features and price.
  - Box plots showing price variation by location.
- **Model Performance**:
  - RMSE: 1022560.052
  - R^2: 0.6114024924156645



```
Enter value for area:  2000
Enter value for bedrooms:  2
Enter value for bathrooms:  2
Enter value for stories:  1
Enter value for mainroad:  0
Enter value for guestroom:  0
Enter value for basement:  0
Enter value for hotwaterheating:  0
Enter value for airconditioning:  0
Enter value for parking:  0
Enter value for prefarea:  0
Enter value for furnishingstatus_semi-furnished:  0
Enter value for furnishingstatus_unfurnished:  0


Predicted House Price: 3410113.0
```

1 means YES and 0 means NO