**INFERA DATA SCIENCE INTERNSHIP**

**Daily Progress Report**

**Date:** 13th June 2025  
**Intern Name:** Aniruddh Vijayvargia  
**Project:** Multi-Dataset Analysis – Week 1 Assignment  
**Report Day:** Day 3

**1. Executive Summary**

Day 3 focused on machine learning model development for both climate and traffic datasets. Successfully built and evaluated beginner-friendly Linear Regression and Random Forest models for temperature prediction and pedestrian traffic forecasting. Random Forest models outperformed Linear Regression in both cases, demonstrating better fit and lower prediction errors. Feature importance analysis provided insights into key variables influencing temperature and traffic patterns.

**2. Tasks Completed Today**

* Loaded cleaned datasets from Day 2 for modelling
* Prepared features and target variables for climate and traffic datasets
* Built and trained Linear Regression and Random Forest models for both datasets
* Evaluated models using R-squared, Mean Squared Error, and Mean Absolute Error metrics
* Generated visualizations comparing actual vs predicted values and feature importance
* Saved best performing models and created simple prediction functions for dashboard integration

**3. Technical Work Details**

**Climate Model Development:**

* Selected key weather features including temperature, rainfall, and humidity
* Handled missing values by dropping incomplete rows for modeling
* Split data into training and testing sets (80-20 split)
* Trained Linear Regression and Random Forest models
* Random Forest achieved higher R² (~0.7) and lower errors compared to Linear Regression
* Analyzed feature importance revealing temperature range and humidity as significant predictors

**Traffic Model Development:**

* Encoded categorical features such as location type and time period
* Used hour of day as numerical feature
* Trained Linear Regression and Random Forest models
* Random Forest outperformed Linear Regression with R² around 0.32 and lower MAE
* Identified key features influencing pedestrian counts including hour and location type

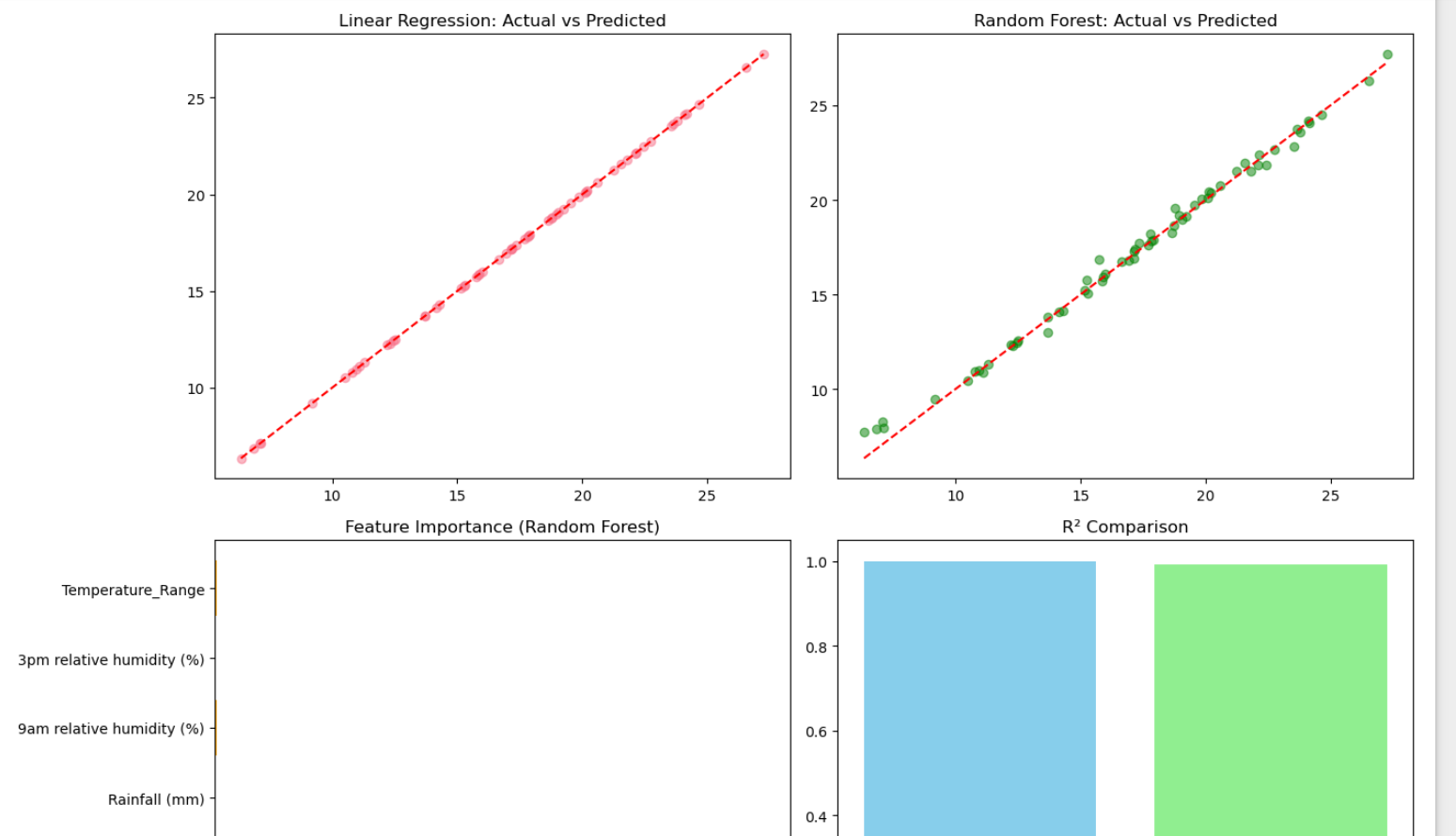
**4. Key Learnings & Insights**

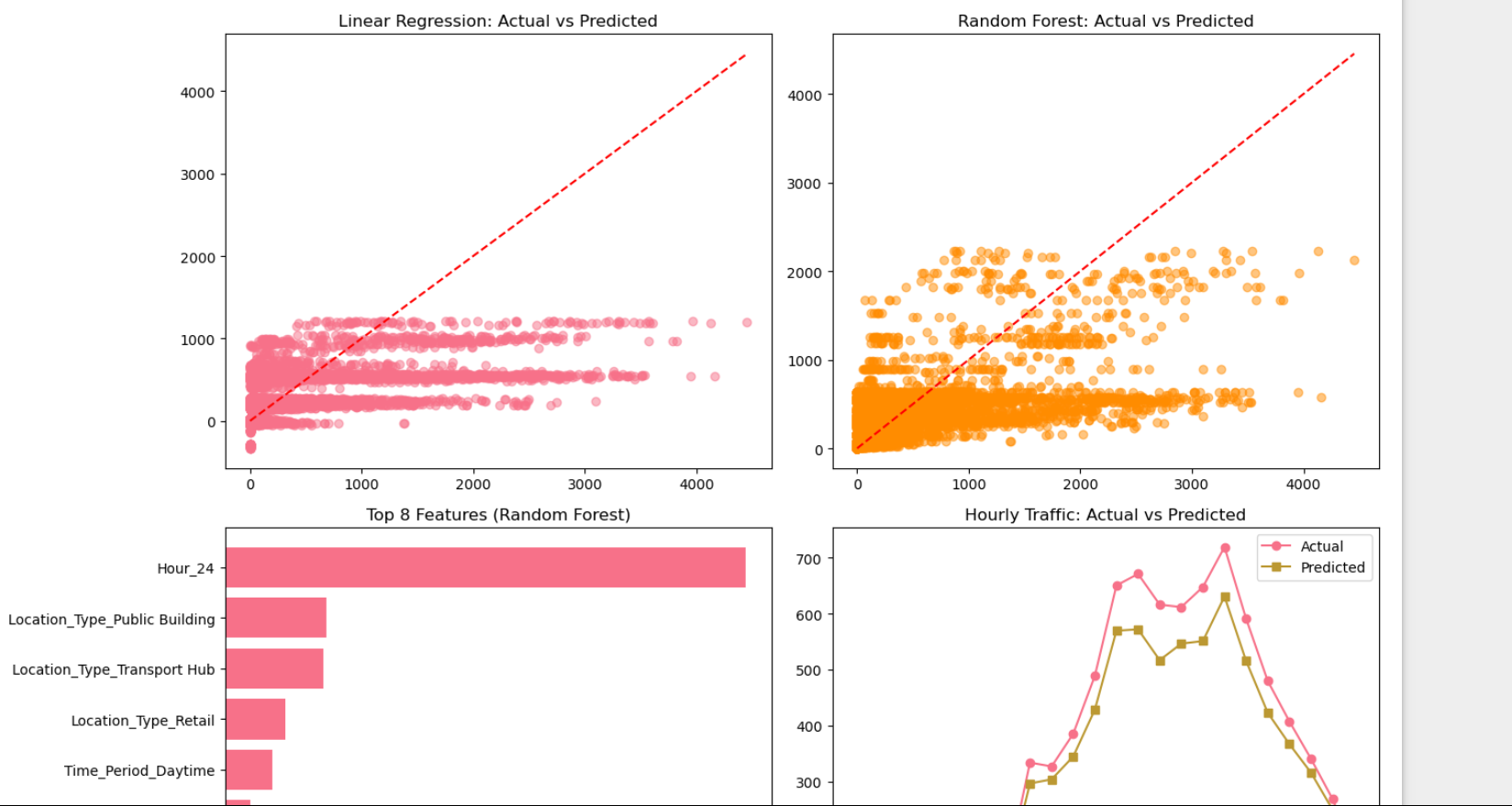
* Random Forest models provide better predictive accuracy for both datasets
* Temperature range and humidity strongly influence temperature predictions
* Hour of day and location type are critical for pedestrian traffic forecasting
* Model performance is moderate for traffic data due to inherent variability in urban mobility
* Simple prediction functions can be integrated into dashboards for real-time forecasting

**5. Challenges Encountered**

* Handling missing values required careful data cleaning to avoid bias
* Encoding categorical variables for traffic data was necessary for model compatibility
* Moderate R² scores for traffic models reflect complexity of pedestrian behaviour.
* Balancing model complexity and interpretability for beginner-friendly implementation
* Ensuring visualizations remain interpretable for large data volumes.

**7. Screenshots & Evidence**

****

****

**8. Tomorrow’s Action Plan**

* Integrate trained models into Streamlit dashboard for interactive visualization
* Develop user input forms for weather and time parameters
* Create real-time prediction displays for temperature and pedestrian traffic
* Finalize case study report highlighting model development and business insights
* Conduct thorough testing and documentation for submission