Question 1 – Model Comparison Report (Polynomial Regression With and Without Interactions)

Problem Statement:

- The objective is to build and compare regression models to predict hourly bike rental count using weather, temporal, and categorical features.
- The dataset is split into an 80% training set and a 20% test set.
- Models evaluated include Linear Regression and Polynomial Regression of degrees 2, 3, and 4 without interaction terms, along with a quadratic model (degree 2) that includes interaction terms.
- Categorical variables were one-hot encoded, numerical variables standardized, and the target transformed using log1p with smearing correction applied during prediction.
- Model selection is based solely on test-set MSE.
- Data set Link: [Bike Sharing Demand | Kaggle]

Some Formulas and Model definitions Used for checking:

- Log Transformation: To stabilize variance and reduce skewness in the target variable, where 'y' is the

$$Y_{log} = \log(1 + y)$$

- Degree 2 Polynomial (No Interactions):

$$\hat{y} = w_0 + \Sigma_j (w_j x_j) + \Sigma_j (w_j^{(2)} x_j^2)$$

- Degree 2 Polynomial (No Interactions):

$$\hat{y} = w_0 + \Sigma_j (w_j x_j) + \Sigma_j (w_j^{(2)} x_j^2) + \Sigma_{j,k}(w_{jk} x_j x_k)$$

- Degree 3 Polynomial (No Interactions):

$$\hat{y} = w_0 + \Sigma_j (w_j x_j) + \Sigma_j (w_j^{(2)} x_j^2) + \Sigma_j (w_j^{(3)} x_j^3)$$

- Degree 4 Polynomial (General Form):

$$\hat{y} = w_0 + \Sigma_j (w_j x_j) + \Sigma_j (w_j^{(2)} x_j^2) + \Sigma_j (w_j^{(3)} x_j^3) + \Sigma_j (w_j^{(4)} x_j^4)$$

Solution Explanation:

- The dataset was chronologically sorted and split to prevent leakage.
- Feature engineering added time variables, cyclical encodings, and additional features such as temp_diff and humid_ws. Numerical features were standardized and categorical features one-hot encoded.
- Polynomial Features generated degree 2, 3, and 4 polynomial features without interactions by filtering cross terms. Additionally, a degree-2 polynomial model with interactions was evaluated by retaining all cross terms.
- Ridge CV was used for regularization with wide alpha ranges, and TruncatedSVD was optionally applied for higher-degree models.
- All models were evaluated on the transformed target using smearing correction to ensure unbiased predictions.

Model Performance Output:

MODEL PERFORMANCE (TEST SET)

Model	MSE	R^2	Alpha	Interactions
Linear Regression	39457.76	0.1667	N/A	No
Polynomail of Degree 2 (No Interactions)	32476.84	0.3141	24.244620	No
Polynomail of Degree 2 (With Interactions)	27515.75	0.4189	17.433288	Yes
Polynomail of Degree 3 (No Interactions)	27306.20	0.4233	62.355073	No
Polynomail of Degree 4 (No Interactions)	28613.59	0.3957	62.355073	No

Best Model:

BEST MODEL INTERPRETATION BASED ON MSE VALUES

Best Model : Polynomail of Degree 3 (No Interactions)
Test MSE : 27306.20
Test R^2 : 0.4233
Alpha : 62.355073
Interactions : No

- Among all evaluated models, the Degree 3 Polynomial Regression model without interaction terms achieved the lowest test-set MSE (27281.25) and the highest R^2 (0.4238).
- This model captures nonlinear relationships effectively while maintaining stability due to Ridge regularization and feature clipping.

- Degree 4 overfits slightly, and Linear Regression underfits the data. The degree-2 model with interactions provides useful comparison insights but does not outperform Degree 3.

Conclusion:

- The Degree 3 Polynomial Regression model (without interactions) is the optimal model for predicting bike rental count based on test-set performance.
- It balances flexibility and generalization while meeting all project constraints.
- The inclusion of interaction terms in the degree-2 model does not improve performance beyond that of the degree-3 model but is valuable for comparison.