KarthikReddy_COHORT_III
Linear Regression Subjective Questions

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

The categorical variables substantially influence daily bike demand. The year indicator (yr) shows a clear upward shift in average rentals in the later year, which captures a growth trend in usage and is one of the strongest predictors. Seasonality also matters: warmer seasons show higher demand and colder or shoulder seasons show lower demand, reflecting user behavior tied to weather and daylight. Weather condition (weathersit) has a pronounced effect , clear weather increases rentals while mist, rain or snow reduce demand considerably. Month, weekday, holiday and workingday dummies further refine the model by capturing within-season and weekly patterns such as weekend peaks or holiday dips.

2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Answer:

Using drop_first=True prevents the dummy variable trap, which is perfect multicollinearity that occurs when all k dummy columns for a categorical variable are included. Dropping one reference category keeps the design matrix full rank so OLS can compute unique coefficient estimates. It also makes the remaining coefficients interpretable as effects relative to the dropped baseline, whose effect is absorbed in the intercept.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

atemp (apparent temperature) has the highest correlation with cnt; it aligns closely with human comfort and therefore with the decision to use a bike outdoors.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

I validated assumptions using diagnostics on the model residuals and collinearity checks. Linearity and homoscedasticity were inspected with a residuals versus fitted-values plot , absence of a pattern and no funnel shape supports linearity and constant variance. Normality of errors was assessed with a Q–Q plot; close alignment to the 45-degree line indicates approximately normal residuals. Multicollinearity was measured using VIF scores; predictors with high VIF were reviewed and removed or combined (for example dropping temp when atemp captured the same signal). If any assumption showed a serious violation I tested simple remedies such as transformations, removing collinear predictors, or using robust inference.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

The top three significant features are:

- yr: captures year-on-year increase in demand and is the strongest positive driver.

- atemp: apparent temperature closely tracks rider comfort and is a key meteorological predictor.

- a weathersit dummy (for example Mist_Cloudy or Light_Rain_Snow): adverse weather dummies have large negative coefficients, indicating strong reductions in demand during poor weather.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression models the relationship between a continuous target and predictors by fitting a linear equation: y_hat = beta0 + beta1*x1 + ... + betap*xp + error. Ordinary Least Squares (OLS) finds coefficient estimates that minimize the sum of squared residuals. In matrix terms, when X'X is invertible the OLS solution is beta = (X'X)^{-1} X'y. Each coefficient represents the expected change in the target for a one-unit change in the corresponding predictor, holding others constant. Valid inference requires assumptions (linearity, independence, homoscedasticity, normality of errors and no perfect multicollinearity); diagnostics and corrective steps are used when assumptions are violated.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer:

Anscombe’s quartet consists of four datasets constructed to have nearly identical summary statistics , means, variances, correlation coefficients and even the same fitted regression line , yet the scatterplots for each are very different. The four plots demonstrate a well-behaved linear relationship, a nonlinear pattern, a dataset affected by a vertical outlier, and a relationship dominated by a single high-leverage point. The quartet teaches that summary numbers alone can be misleading and that visual inspection is essential before trusting statistical conclusions.

3. What is Pearson’s R? (3 marks)

Answer:

Pearson’s R is the Pearson product-moment correlation coefficient measuring the strength and direction of a linear relationship between two continuous variables. It ranges from -1 (perfect negative linear) to +1 (perfect positive linear), with 0 indicating no linear association. It is computed as the covariance of the variables divided by the product of their standard deviations. A high absolute value indicates strong linear association but does not imply causation and can miss non-linear relationships.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling rescales numeric features so they lie on a comparable range or distribution, which prevents features with large magnitudes from dominating model fitting or distance calculations and improves numerical stability and optimizer convergence. Normalization (Min-Max) rescales values to a fixed interval, typically [0,1], using (x − min) / (max − min) and is sensitive to outliers. Standardization (Z-score) centers data to mean zero and scales to unit variance using (x − mean) / std; it is less affected by extreme values and is preferred when features have different scales or when algorithms assume roughly Gaussian features.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

VIF is defined as 1 / (1 − R_i^2) where R_i^2 is the R-squared from regressing predictor i on the other predictors. VIF becomes infinite when R_i^2 equals 1, which means predictor i is perfectly predicted by the other predictors — perfect multicollinearity. Typical causes include the dummy variable trap, redundant or duplicate features, or exact linear combinations of variables. The remedy is to remove or combine redundant predictors, drop a reference dummy when encoding categories, or use dimensionality reduction.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

A Q-Q (quantile-quantile) plot compares quantiles of the sample residual distribution to quantiles of a theoretical distribution, typically the normal distribution. In regression, it checks whether residuals are approximately normal; points lying close to the 45-degree line support the normality assumption. Normal residuals are important because p-values and confidence intervals for OLS coefficients rely on approximate normality; large deviations suggest considering transformations, robust methods, or alternative models.