Linear regression subjective questions

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

Ans: - It is observed to be demand for bikes was more in 2019 than 2018, so just focus as there is increase in 2019 and might be facing dips in their revenues due to the ongoing Corona pandemic and by the time it reduces the things will be better.

- Can focus more on Summer & Winter season, August, September month, Weekends, Working days as they have good influence on bike rentals.

- spring season has negative coefficients and negatively correlated to bike rentals. So we can give some offers there to increase the demand

- weathersit variable, it is observed that negative coefficients for Mist +cloudy and Lightsnow weather... And yes we can give offers

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

Ans: Using `drop_first=True` prevents multicollinearity by removing one dummy variable, making models more stable and interpretable. It ensures the reference category is implied, avoiding redundant information.

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

Ans: Among the numerical variables Count is highly corelated with temperature

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans: By performing Residual analysis on predicted value and I have plotted distplot between y_train and y_train_pred (predicted values) to find out normal distribution which resulted in normal distribution shows error terms distributed normally.

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

Ans: Features like Temperature, Seasons especially in Summer, Year mainly 2019 has raise in bike demands

Also Feature like Workingday's are most demanding for the bike use.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: The model assumes a linear relationship, represented as: y = beta_0x + beta_1x + beta_2x + C

Dependent variable (predicted outcome).

- β0\beta_0: Intercept (value of yy when predictors are 0).
- β1,β2,…\beta_1, \beta_2, Coefficients that represent the effect of predictors (x1,x2,…x_1, x_2,)
- ϵ\epsilon: Error term accounting for unexplained variability.

2. **Types of Linear Regression**:

- **Simple Linear Regression**: Models one independent variable.
- **Multiple Linear Regression**: Models two or more independent variables.

Algorithm Steps

1. **Data Preparation**: Clean, preprocess, and split data into training and testing sets.
2. **Model Assumptions**:
   - Linearity: Relationship between predictors and target must be linear.
   - Normality: Residuals (errors) should follow a normal distribution.
   - Homoscedasticity: Constant variance of residuals across predictions.
   - Independence: Observations and residuals are independent.
   - No Multicollinearity: Predictors shouldn't be highly correlated.
3. **Fitting the Model**:
   - The algorithm minimizes the **sum of squared errors (SSE)**:
   - It uses methods like **Ordinary Least Squares (OLS)** to estimate the coefficients (β\beta) by minimizing SSE.
4. **Evaluation Metrics**:
   - R2R^2: Proportion of variance in the dependent variable explained by the model.
   - Mean Squared Error (MSE) or Root Mean Squared Error (RMSE): Measures prediction accuracy.
   - p-values: Test the significance of individual predictors.
5. **Model Validation**: Use diagnostic plots (residual vs. predicted, Q-Q plot, etc.) to validate assumptions.

2. Explain the Anscombe's quartet in detail.
Ans: Anscombe's quartet contains four datasets that share similar statistical properties but differ significantly in their visual patterns. It emphasizes the importance of data visualization in identifying anomalies and trends, rather than relying solely on summary statistics.

3. What is Pearson's R?

Ans: Pearson's correlation coefficient (rr) measures the strength and direction of a linear relationship between two variables. It ranges from -1 to +1, where:

- +1: Perfect positive correlation, -1: Perfect negative correlation, 0: No correlation.

4. What is scaling? Why is it performed? Difference between normalized and standardized scaling

- *Scaling*: Adjusts feature values to be comparable.
- *Why*: Helps improve model performance, especially in algorithms sensitive to magnitude.
- *Normalized Scaling*: Scales values to [0, 1], focusing on proportions.
- *Standardized Scaling*: Centers values around 0 with a standard deviation of 1, retaining data distribution.

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

Variance Inflation Factor (VIF) becomes infinite when perfect collinearity exists between variables, meaning one variable can be expressed as an exact linear combination of others, leading to unstable coefficients.

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

Ans: A Q-Q plot compares residuals from a model to a theoretical normal distribution.

- *Use*: Checks whether residuals follow a normal distribution.
- *Importance*: Ensures the linear regression assumption of normality, enhancing model validity.