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)

   Categorical Variables:

   - **Season:** Summer and fall are the most favorable seasons for biking, suggesting higher targets and strategic advertising during these times. Spring sees significantly lower usage.
   - **Working Day:** Registered users tend to rent bikes on working days, while casual users prefer non-working days. This difference balances out in the total count. Tailoring strategies to these patterns could boost rentals.
   - **Weathersit:** Clear or lightly cloudy days are the most favorable for biking. Registered users still rent on light rainy days, likely for commuting.
   - **Weekday:** The total bike count shows no significant pattern across weekdays, but registered users rent more on working days, while casual users prefer weekends.
   - **Year:** Data shows an increase in bike rentals from 2018 to 2019.
   - **Holiday:** Casual users rent more on holidays compared to registered users.

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

   Using one-hot encoding, dummy variables are created to represent each category of a categorical variable, with values of 1 indicating the presence of a category and 0 indicating its absence. For a categorical variable with three categories, three dummy variables would typically be generated. However, by setting **drop_first=True**, the first (or reference) category is dropped. This step is crucial to avoid multicollinearity in the model, as including all dummy variables would make them linearly dependent. The reference category can still be identified in the data by a row where all other dummy variables for that category are 0.

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

   - The **registered** , **casual** and **temp** are the features with high positive correlation with cofficients **0.95 , 0.67 , 0.63** respectively but since variables are actually part of the target variable as values of these columns sum up to get the target variable, these features in further linear regression process.Hence after excluding these **temp** has strong positive correlation with target variable.
   - **windspeed** has strong negative correlation (-0.24) with target variable.

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

- **Linear relationship between independent and dependent variables**: The linearity is confirmed by examining the points that are symmetrically distributed around the diagonal line in the actual vs. predicted plot, as illustrated in the figure below.

- **Independence of error terms**: The absence of any specific pattern in the error terms relative to the predictions indicates that the error terms are independent of each other.

- **Normal distribution of error terms**: A histogram and distribution plot provide insight into the normal distribution of error terms, centered around a mean of 0. This is clearly shown in the figure below.

- **Constant variance of error terms (homoscedasticity)**: The error terms exhibit approximately constant variance, thereby satisfying the assumption of homoscedasticity

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

The top variables are:

- **Year** : The year-on-year growth appears to be organic, considering the geographical factors.
- **Season**: The winter season plays a crucial role in driving the demand for shared bikes.
- **Temperature** is the most significant factor positively impacting the business,
- **Weathersit**: Conditions like rain, humidity, wind speed, and cloudiness have a negative effect.

General Subjective Questions

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

Linear regression is a technique used to establish the best linear relationship between independent and dependent variables. The algorithm finds the best-fitting line that maps this relationship.

Types of Linear Regression:

1. **Simple Linear Regression (SLR)**: Uses a single independent variable. The line equation is $Y=\beta0+\beta1X$
2. **Multiple Linear Regression (MLR)**: Uses multiple independent variables. The line equation is $Y=\beta0+\beta1X1+\cdots+\beta pXp+\epsilon$

- **$\beta0$** : The Y-intercept, or the value of Y when X=0
- **$\beta1,\beta2,\ldots,\beta p$** : The slopes or gradients for the respective independent variables.

Cost Functions:

Cost functions determine the best values for β0,β1,…,βp to minimize prediction errors. This is done by minimizing the difference between the predicted values (Ypred) and actual values (Yi). The most common cost function is the sum of squared errors:

- **Cost function equation**: J(β0,β1)=Σ(Yi−(β0+β1Xi))^2

Minimization Approaches:

Two primary methods are used to minimize cost functions:

1. **Closed Form**: Provides an exact solution.
2. **Gradient Descent**: Iteratively adjusts β\betaβ values to find the minimum error.

Ordinary Least Squares (OLS):

OLS is a method used to minimize the residual sum of squares (RSS), which is the sum of the squared differences between the actual and predicted values. The error for each data point is given by ei=Yi−Ypred , and OLS minimizes the total e^2 to estimate the best-fitting line.

- **RSS Equation**: RSS=Σ(Yi−Ypred)^2

In summary, OLS is used to minimize RSS and estimate the optimal beta coefficients for the regression model.

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

Anscombe's Quartet consists of four datasets with nearly identical summary statistics, including mean, variance, correlation, and linear regression lines. Despite these similarities, the datasets reveal dramatically different patterns when plotted as scatter plots.

Key Insights:

Illustration: One of Anscombe's datasets shows that while descriptive statistics might appear similar, the relationships between data points can look entirely different when plotted.

Anscombe's Quartet: This set of datasets highlights that even with matching statistical measures, datasets can exhibit distinct patterns when visualized.

Outliers: The quartet also emphasizes the impact of outliers. Without them, the descriptive statistics would significantly differ.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. The coefficient ranges from -1 to 1:

-1 indicates a perfect negative linear relationship (as one variable increases, the other decreases).

0 indicates no linear relationship.
1 indicates a perfect positive linear relationship (as one variable increases, the other also increases).
Pearson's R is widely used to determine how closely two variables are related, with values closer to -1 or 1 indicating a stronger relationship.

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

Scaling is a data preparation step for regression models that adjusts the values of different variables to a specific range, making them comparable.

Data collected from various sources can have different units and ranges, leading to high variance. Without scaling, this variance can distort model coefficients and hinder accurate comparisons. Scaling ensures that all features contribute equally, improving the model's interpretability.

Normalization (Min-Max Scaling) vs. Standardization:

Normalization (Min-Max Scaling): Rescales data to a range between 0 and 1, which can help in normalizing outliers

$$MinMaxScaling: \quad \frac{x-\min(x)}{\max(x)-\min(x)}$$

Standardization: Converts data to a standard normal distribution with a mean of 0 and a standard deviation of 1

$$Standardization: \quad \frac{x-mean(x)}{sd(x)}$$

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

The value of VIF (Variance Inflation Factor) becomes infinite when R^2 is equal to 1. This occurs because the VIF formula is:

VIF=1/(1-R^2)

When R^2 is 1, it means there is perfect correlation between two independent variables, leading to the denominator becoming zero, and thus the VIF becomes infinite. This indicates severe multicollinearity, where one independent variable can be perfectly predicted by another, making it impossible to estimate the regression coefficients reliably.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distributions of two datasets or to assess if a dataset follows a theoretical distribution (such as normal, exponential, or uniform).

Uses in Linear Regression:

- **Distribution Check**: It helps determine if the training and test datasets come from the same distribution.
- **Normality Check**: It assesses if the residuals from a regression model are normally distributed.

Interpretations:

- **Similar Distribution**: Data points align along a 45-degree line, indicating similar distributions.
- **Y-values < X-values**: Y quantiles are lower than X quantiles.
- **X-values < Y-values**: X quantiles are lower than Y quantiles.
- **Different Distributions**: Data points deviate from the straight line, indicating different distributions.