# 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 –**

From the analysis from the categorical data, we can see that -

- For the variable season, we can clearly see that the **category 3: 'Fall',** has the highest median, which shows that the demand was high during this season and the least is **category 1: 'spring'**.
- The **year 2019** had a higher count of users as compared to the **year 2018**.
- Demand is continuously growing each month **till June** and reached its limit in September month, but demand started **declining after September**.
- When there is a **holiday**, demand has decreased.
- The count of rentals is almost even throughout the **week**.
- For **Working day,** we can see those maximum bookings happening between 4000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.
- There are **less users** when there is **heavy rain/ snow** indicating that this weather is quite adverse and **highest count** was seen when the weather situation was **Clear, Partly Cloudy.**

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

**Answer –**

When creating dummy variables for categorical features, the `drop_first=True` argument is important to avoid **multicollinearity**. Here's why:

- **Multicollinearity**: In regression models, multicollinearity occurs when one predictor variable can be predicted from another. If you create dummy variables for all categories of a categorical variable, the sum of these variables will always equal 1 (since one category is always present). This introduces perfect multicollinearity because the values of one dummy variable can be exactly inferred from the others.
- **Dummy Variable Trap**: This refers to the problem of having one redundant dummy variable that causes perfect collinearity. If all dummy variables are included, one of them is redundant because the sum of all the dummies equals 1.
- **Avoiding Redundancy**: By setting `drop_first=True`, one of the dummy variables (the first category) is dropped, eliminating redundancy and multicollinearity. The model can still make predictions correctly without this variable, as the dropped category can be inferred from the others.

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

**Answer –**

- After seeing the plots, we can see that the "**temp**" and "**atemp**" are highly correlated with each other.
- Secondly, the "**temp**" and "**atemp**" has the highest corelation with the target variable "**cnt**"

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

**Answer –**

To validate the assumptions of linear regression after building the model on the training set, a systematic approach is followed:

- **Linearity**: This assumes a linear relationship between the independent variables and the dependent variable. To check, plot residuals vs. predicted values. If the points are randomly scattered around the zero line without a distinct pattern, the assumption holds.
- **Independence of Errors**: Residuals should be independent. For time series data, plot residuals over time to detect patterns. Additionally, the Durbin-Watson test can identify autocorrelation, with values near 2 indicating independence.
- **Homoscedasticity**: The variance of residuals should remain constant across predicted values. To check, plot residuals vs. predicted values—a random spread suggests homoscedasticity.
- **Normality of Errors:** Residuals should follow a normal distribution. A Q-Q plot helps assess this, where points should lie on a straight line. Statistical tests like Shapiro-Wilk can also confirm normality.
- **Multicollinearity**: Independent variables shouldn't be highly correlated. Variance Inflation Factor (VIF) helps assess this, with VIF values greater than 5-10 suggesting multicollinearity.
- **Outliers and Leverage Points**: Check for influential data points using Cook's distance and leverage scores. Large values might indicate influential outliers that can distort the model.

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

- **Plan for High-Demand Periods**: Since warmer temperatures and non-holiday periods see higher demand, BoomBikes can increase fleet availability and marketing during these times.
- **Prepare for Weather Impact**: With reduced demand during bad weather (like rain or high wind), the company can adjust its operational strategy, perhaps offering promotions during low-demand periods.
- **Seasonal Promotions**: The higher demand during summer and winter suggests these periods could be targeted for seasonal campaigns, while strategies might be adjusted for spring, which sees a drop in usage.

## General Subjective Questions

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

Linear regression is a supervised learning algorithm used to predict a dependent variable (\(y\)) based on one or more independent variables (x). It assumes a linear relationship between the variables. In its simplest form, simple linear regression involves one independent variable and can be expressed as:

$$y=\beta0+\beta1x+\epsilon$$

Where y is the predicted value, x is the independent variable, beta_0 is the intercept, beta_1 is the slope, and ϵ is the error term.

For multiple linear regression, the model includes several independent variables:

$$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$$

**Algorithm Steps:**

- **Hypothesis Function**: The model assumes the dependent variable is a linear combination of the independent variables. The goal is to find the best-fit line (or plane) that minimizes prediction error.
- **Cost Function (Mean Squared Error):** To measure how well the model fits, the cost function calculates the difference between the predicted values and the actual values. The most common cost function is the Mean Squared Error (MSE):

$$MSE=n1i=1\sum n(yi-yi\hat{})2$$

- **Optimization (Gradient Descent):** The algorithm aims to minimize the cost function by adjusting the coefficients (β0,β1,…,βn). Gradient descent is commonly used for this, iteratively updating coefficients to reduce the error.
- **Model Evaluation**: Once the model is trained, it's evaluated using metrics like **R-squared** (explained variance) or RMSE (Root Mean Squared Error) to assess its performance and generalizability.

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

**Anscombe's quartet** is a set of four datasets created by statistician Francis Anscombe in 1973 to illustrate the importance of data visualization. Each dataset has nearly identical statistical properties—mean, variance, correlation, and linear regression line—yet they exhibit vastly different distributions when plotted.

**Key Features:**

- Mean of x-values: ~9

- Mean of y-values: ~7.5

- Variance of x-values: ~11

- Variance of y-values: ~4.1

- Correlation coefficient (r): ~0.816

- Linear regression equation: \(y = 3 + 0.5x\)

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

**Answer -**

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is widely used in statistics and data analysis to understand how two variables move in relation to each other.

Formula:

Pearson’s R is calculated as:

$$r=\sum(xi-\bar{x})2\sum(yi-\bar{y})2\sum(xi-\bar{x})(yi-\bar{y})$$

Where:

- \(x_i\) and \(y_i\) are the individual data points.
- \(\bar{x}\) and \(\bar{y}\) are the means of the respective variables.
- \(r\) ranges between -1 and 1.

Interpretation:

- r = 1: Perfect positive linear relationship (as one variable increases, the other increases).
- r = -1: Perfect negative linear relationship (as one variable increases, the other decreases).
- r = 0: No linear relationship between the variables.

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

**Answer -**

Scaling transforms features to a specific range for consistent analysis in machine learning. It’s performed to improve model performance, especially for algorithms sensitive to feature magnitude.

- Normalization: scales data to [0, 1].
- Standardization: centres data around mean 0, std. 1.

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

**Answer -**

A **Variance Inflation Factor (VIF)** becomes infinite when perfect multicollinearity exists between independent variables in a regression model. This happens when one predictor variable is an exact linear combination of one or more other predictors.

For example, if two variables are perfectly correlated (e.g., x1=2x2x_1 = 2x_2x1=2x2), VIF for one or both variables will be infinite. This is because VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity, and perfect multicollinearity means the variance is infinitely large.

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 plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, typically the normal distribution. The plot shows the quantiles of the sample data on the y-axis against the quantiles of the reference (theoretical) distribution on the x-axis.

**Use in Linear Regression:**

In linear regression, a key assumption is that the residuals (errors) are **normally distributed**. A Q-Q plot helps assess this assumption by comparing the distribution of the residuals to a normal distribution.