## Assignment-based Subjective Questions

**Question 1**. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
**Total Marks**: 3 marks (Do not edit)
**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

- Fall season having high no of count
- Count is higher in 2019 as compared to 2018
- Count is higher when it is clear weather
- Count is higher on Non holidays
- Less no of counts on Sunday

---

**Question 2.** Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first =True when creating dummy variables is important because it helps prevent multicollinearity. By dropping one category, we can avoid multicollinearity as the information from the dropped category is included in the remaining ones

---

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
**Total Marks:** 1 mark (Do not edit)
**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)

registered is having high correlation with target variable cnt

---

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

- Performed Residual analysis to validate the assumptions of Linear Regression
- Plotted the histogram of error terms and found that error is normally distributed across 0 which indicates that our model has handled the assumption of error normally distributed
- Homoscedasticity - we can see that variance is similar from both ends of the fitted line

---

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 5 goes below this line> (Do not edit)

Top 3 features -

- yr
- temp
- weathersit_Light Mist

---

## General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)
**Total Marks:** 4 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

Linear Regression is a type of Supervised Machine learning. Target variable is the continuous value.
There are two types of Linear Regression -

- Simple Linear Regression - No of predictors is one
    - eg y=b0+b1x where b0 is the interceptor and b1 is coefficient for slope x.
- Multiple Linear Regression - No of predictors are more than one
    - eg y =b0+b1x1+b2x2+b3x3+.....+bnxn where bo is the interceptor and b1, b2,b3,...,bn is coefficient/slopes of x1,x2,x3…xn predictors

Error is normally distributed across 0
Homoscedasticity should be applicable that variance is similar from both ends of the fitted line

---

**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It teaches us that relying solely on summary statistics can be misleading. Visualising data helps in understanding the true nature of the data and making better decisions based on it.

---

**Question 8.** What is Pearson's R? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

Pearson's r, also known as the Pearson coefficient, is a measure of the strength and direction of the linear relationship between two variables. It's a value between -1 and 1:

- 1 indicates a perfect positive linear relationship

- -1 indicates a perfect negative linear relationship
- 0 indicates no linear relationship

---

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

Scaling is a data preprocessing technique used in machine learning and data analysis to adjust the range of features in your data.

Why is it performed?

- Algorithms like gradient descent works faster when features are scaled
- Ensures that no single feature dominates due to its scale
- Makes it easier to compare and interpret the coefficients of the model

Normalized scaling vs Standardized scaling

- Normalized scaling
    - Rescales the data to a fixed range typically [0,1] or [-1,1]
    - Xnorm = X- Xmin/Xmax - Xmin
- Standardized scaling
    - Centers the data around the mean with a unit standard deviation
    - Xstd = X- u/ std

---

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

VIF is infinite when there is a perfect multicollinearity in the model. Perfect multicollinearity means that one or more independent variables can be exactly predicted from a linear combination of the other independent variables

---

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

The Q-Q plot is designed to help you visually compare the quantiles of your data to the quantiles of a theoretical distribution.It helps to evaluate the normality assumption, detect skewness and outliers, and guide model improvement if deviations are observed

Uses of Q-Q plot-

- To check if dataset follows a normal distribution
- It can compare the distribution of your data to any theoretical distribution
- Deviations from the expected line can help identify outliers or anomalies in the data
- Used to check the residuals distribution ensuring that the assumptions of the regression model are met

---