# 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)

# The Categorical Variables Weekday, Day (Inferred from dteday) were not part of the final model since they had a high p-value

# The Categorical Variables Summer, Workingday, month (inferred from dteday) were not part of the final since they had high p-value and high VIF and were removed as part of RFE

# The remaining categorical variables holiday, Sprint(inferred from season), Light Snow, Mist (Inferred from weathersit) have a negative impact on the final rental bike count

# Winter (inferred from Season) seems to have a positive implact on the final rental bike count

# 

**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)

# Drop First is important to use to reduce the total number of columns, since to represent n-categorical value only n-1 categorical values are required since the value 0,0,0….. 0 will also represent a value. Hence to simplify the amount of data that is fed to the model we use drop\_first = True. If one variable is not dropped then it can lead to perfect collinearity hence it is required to drop one column when you have n-categorical values

**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)

# Temp and atemp both seem to have the highest correlation with the target variable

**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)

# Checking the Adjust R2 value whose value was 0.819 which is good

# Checking F-statistic value which was very small

# Checking the R2 score for the trained and the test data with model get a similar score of 0.78

**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 contributing significantly towards explaining the demand

# Temperature, Year -> Positively impacts the demand

# Light Snow -> Negatively impacts the demand

# 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 statistical method for modeling the relationship between one or more independent variables (predictors) and a dependent variable (outcome). The goal is to find the best-fitting straight line (in the case of one predictor) or hyperplane (for multiple predictors) that predicts the outcome as accurately as possible.

Y=β0​+β1​X1​+β2​X2​+…+βn​Xn​+ϵ

The way to find the optimal line is to start with a line which intercepts with Y-axis. Then gradually move the line to towards an optimal line where the sum of the squared errors is the minimum

Assumptions of Linear Regression:

 **Linearity**: The relationship between predictors and outcome is linear.

 **Independence**: Observations are independent of each other.

 **Homoscedasticity**: The variance of residuals is constant across all levels of XXX.

 **Normality**: Residuals are normally distributed.

 **No Multicollinearity**: Predictors are not highly correlated with each other.

**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)

# <Your answer for Question 7 goes here>

**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 is a statistical measure that tells us about the linear relationship between 2 continuous variables. It ranges from -1 to 1, where r=1 means perfect positive relationship and r=-1 perfect negative relationship.

Formula forr= ​Cov(X,Y)​/ σX⋅σY

Where Cov(X,Y) : Covariance between variables X and Y

σX, σY: Standard deviation of X, Y

**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 when you transform the values of your dataset to a certain limited range.

# In normalized scaling the range of all the values is between 0 and 1. It is also called min-max scaling since the the formula is x – min / max .

# In standardized scaling has the mean as 0 and the standard deviation as 1.

# Normalized scaling is more sensitive to outliers since it uses min and max values whereas standardized scaling is less sensitive to outliers.

**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)

# Formula for VIF = 1/(1-R2)

# In the use case where one variable can be perfectly predicted by all the other variables the R2 (Coefficienct of determination) will be equal to 1.

# Hence the VIF = 1/(1-1) = 1/0 causes the denominator to be zero and hence the answer becomes infinite

**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)

A Q-Q plot compares the quantiles of observed data to a theoretical distribution to check if the data follows that distribution. If the points form a straight diagonal line, the data is close to the theoretical distribution.

In Linear regression Q-Q plot can be used to check if the data (error terms) is following a normal distribution since it is one of the assumptions made.

This can help improve the model and help indicating any skewness, outliers or model misfit