# 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)

# Temperature (temp) – A coefficient value of ‘0.5499’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5499 units.

# Year (yr) – A coefficient value of ‘0.2331’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2331 units.

# Weather (weathersit\_3) – A coefficient value of ‘-0.2880’ indicated that a unit increase in weathersit3 variable increases the bike hire numbers by 0.2880 units.

# Hence these variables should be considered on priority basis for achieving maximum bike rentals.

# 

**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)

# When we create the dummy variable, for n categorical variable we will be using n dummy variable. However out of n variable, 1 variable can be predicted from the other variables which will lead to multi-collinearity. Hence to avoid this issue drop\_first=True is important to use.

**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)

# “atemp” has 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)

# Train R^2 :0.836

# Train Adjusted R^2 :0.832

# Test R^2 :0.796

# Test Adjusted R^2 :0.785

# Hence the model is the good model.

**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)

# Temperature (temp)

# Year (yr)

# Weather(weathersit\_3)

# 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)

# <Your answer for Question 6 goes here>

# Linear regression model is to build the relationship between a dependent variable (y) and one (or more) independent variable (x) by fitting a linear equation. The linear equation is:

# y=b0+b1x+ϵ

# Steps in linear regression –

# Data collection and preparation

# EDA

# Model Building

# Model evaluation

# R\_squared

# Adjusted r-squared

# Mean square Error

# Root mean square error

**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>

# Anscombe's quartet is a set of four different datasets that have nearly identical simple descriptive statistics yet appear very different when graphed. It was to illustrate the importance of graphing data before analyzing it and to show how relying solely on statistics can be misleading.

**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)

# <Your answer for Question 8 goes here>

# Pearson’s R measures the strength and direction of the linear relationship between 2 continuous variables. It is denoted by ‘r’ and ranges from -1 to 1.

**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)

# <Your answer for Question 9 goes here>

# Scaling is a pre-processing technique to adjust the range of independent variable. This helps to fit the data to specific range so that it can improve the performance and accuracy of the machine learning algorithm.

# Scaling performed for following reasons –

# To improve the algorithm performance.

# Handling different units.

# Normalization transfers the data within the range 0 to 1.

# Standardization ensures the mean to be 0 and standard deviation as 1.

**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)

# <Your answer for Question 10 goes here>

# An infinite value of VIF indicates the perfect multi-collinearity among the independent variable in LR model. This occurs when one independent variable can be perfectly predicted by 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)

# <Your answer for Question 11 goes here>

# A Q-Q plot is created by plotting the quantiles of the observed data against the quantiles of a specified theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line.