# 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)

# In this dataset there is not much correlation between the categorical variables and the nu,ber of users that use the service . Although I expected that the working days and the season might have an impact on the independent variable by looking at the use case itself

# 

**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 helps to create n-1 levels for the categorical variables that has more than 2 values in dummy variable creation

**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)

# The features temp and atemp had the highest correlation with the taret variable temp. year (yr) also had a higher correlation

**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)

**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)

# Based on the co-effeicients derived from the linera regression model Temp, year and weathersit contribute significantly towards the demand of the shared bikes

# 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 is a supervised ML algorithm that is used to compute linear relationship between the independent and dependent variables. Its basic principle is to fit the data into a line y = mx+c where x is the feature , m is the slope or co-efficient and c is a constant . The models equation defines clear coeffcients to understand the influence of each feature present in the data set to predict the target variable . This model is used for data that has numeric and contionous varaiables. The categorical variables aid to the decision by having 0 or 1 . The model is evaluated using Adjusted R2 and prob(F statistics) to determine hoe well the model fits the line and how well it can predict

**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 datasets, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quite similar

**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>

# R measures the strength and direction of correlation between the variables , It ranges between -1 and 1 and is also called as correlation co-efficient

**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>

# Since the linear regression model derives the best fit line and assigns co-efficients to the features , all the numeric features needs to be in the same scale to understand the strength of the correlation . otherwise the co-efficients would also be in different scales which would not help to infer the strong features that would predict the target variables

**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>

# VIF can be inifinite when there is multi collinearity or perfect correlation between the independent variables. This could mean that one variable is expressed as linear combination of others in the dataset and that variable with VIF as infinite could be dropped

**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>

# In statistics, a Q–Q plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point on the plot corresponds to one of the quantiles of the second distribution plotted against the same quantile of the first distribution.