# 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)

Categorical variable were having different type of effect on dependent variable.

Seasons of Fall was indicating rise in bike rent more then other seasons. Weather condition of 1 category favors increase in bike rent. Month of June, july, August and September are having higher bike rent count. We can see increasing number of rent from to saturday and then we can we sudden drop on sunday(s).

# 

**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 = True is used in dummy variable creation to prevent Mulit colliearity which can lead to ‘Dummy Variable Trap’

**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)

# Numerical Variable of Temperature was having highest correlation with 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)

I plotted distribution plot of error and it was having normal distribution which conveyed that model was doing great work.

**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)

# Seasons , temperature and weather condition were primary factor in explaining 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 algorithm is study of linear relationship between independent variable and dependent variable. Equation of regression line is linear equation. We can distinguished linear regression into two type-

# 1. Simple linear regression : Aims to find linear relationship between one independent variable and one dependent variable.

2. Multiple Linear Regression: Aims to finds linear relationship between two or more independent variable and one dependent variable.

**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 that have the same summary statistics (mean, variance, correlation, and regreession line) but look very different when plotted. It shows that summary statistics alone can be misleading. The first dataset has a clear straight-line relationship between the variables... making it a good fit for linear regression. The second dataset has a curved relationship. The third dataset includes an outlier that affects the regression line, even though most points follow a straight-line trend. The fourth dataset has points that are mostly vertical with one outlier, which also impacts the model. annscombe's quartet teachees us the importance of looking at data plots to better understand the data and not just rely on summary statistics like correlation or regression coefficients.

**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 is a measure that shows the strength and direction of a linear relationship between two variables. It ranges from -1 to 1. A value of 1 means a perfect positive relationship, -1 means a perfect negative relationship, and 0 means 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)

# <Your answer for Question 9 goes here>

**Scaling** adjusts data to ensure all features have similar importance when used in machine learning. It helps models perform better, euqalize weights of all features, to optimize the backend math require when processing the algorithm.

Difference between normalization and standardization is that normalization rescale data between 0 and 1 while standardization center data to have mean 0 and standard deviation of 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>

The value of VIF becomes infinite when there is **perfect multicollinearity** between two or more independent variables in the dataset. It happen simply because two or more variable may be having multicollinearity between them.

**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 used to check if the data follows a specific distribution... It compares the quantiles of your data with the quantiles of a normal distribution.

In **linear regression**, a Q-Q plot is important for checking if the error of the model are normally distributed. If the Q-Q plot shows a straight line, it indicates the Error are normal. If the points deviate significantly from the line, it means that the error may not be normally distributed.