# 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)

# Most of the categorical variables do seem to have an impact of the number of bikes are rented.

# The variables with no impact are : workingday and weekday.

# I made use of box plots to test this behaviour.

# 

**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)

# This is necessary because the default one-hot encodings produce random variables(bernoulli) such that all a sum of all variables lead to 1. This creates a situation of extreme multi-collinearity and therefore it becomes very important to drop one of the generated one of the binary flags(bernoulli) indicating the presence of a particular level of the categorical data..

**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 variable atemp has the highest correlation (0l.63) with the target variable.

# Though it is also highly correlated with the temp variable. So I drop the temp variables to avoid instability caused by multicollinearity.

**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)

# The top features that have an impact are :

# atemp

# yr\_2019

# humidity

# 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 statistical method used to model the relationship between a dependent variable (y) and one or more independent variables (x).

# The relationship is expressed as a straight line

# y=β0+β1x+ϵ

# Where

# β₀ is the intercept,

# β₁ is the slope,

# ε is the error term.

# The goal is to find the best-fitting line by minimizing the sum of squared errors (residuals) between observed and predicted values.

# In multiple linear regression, the model extends to multiple predictors:

# y=β0+β1x1+β2x2+⋯+βnxn+ϵ

# Linear regression assumes linearity, independence of observations, constant variance of errors (homoscedasticity), and normally distributed residuals

**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 identical statistical properties (mean, variance, correlation), but look very different when plotted.

# In general, it is used to understand the importance of plotting when it comes to EDA and highlights how the statistical summaries like mean, std etc can be misleading

# Despite having the same basic statistics, the datasets show different relationships, such as linear,

# non-linear, and ones with outliers or no variation.

**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 of correlation between two variables. Correlation is a metric that quantifies the relationship between two variables.

# Pearson’s correlation specifically measures the degree of linear dependance between to variables.

# The formula is as mentioned below:

# r = [n(Σxy) − ΣxΣy]/ √[n(Σx2) − (Σx)2][n(Σy2) − (Σy)2]

# The value of Pearson’s R is between [-1,1]. The actual magnitude of the value tell the strength of the relationship where as the sign of the value tells about the direction -> positive or negative correlation.

**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 performed in order to maintain the stability of the gradient descent algorithm. In a linear model, the weights(coefficients) attached to a an input variable scale to correct for the impact of the magnitude of the variable, but when performing gradient descent, it helps the convergence speed if the loss surface is symmetrical as it would help all variables converge at the same rate using the same learning rate.

Normalized Scaling: Also known as min-max scaling forces the transformed variable to be in the range of 0-1.

Standard Scaling: Allows the transformed output to take values from -inf to inf, theretically, but forces each variable to have a zero mean and unit variance.

**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>

# Looking at the formula for the VIF,

# VIF = 1/(1-R^2)

# Now in a scenario where a variable can be expressed as combination of all other variables of the model data, the R^2 = 1 which leads to the VIF -> inf.

**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 (Quantile-Quantile plot) is a graph used to compare the distribution of a dataset with a theoretical reference distribution, generally the normal distribution. In a Q-Q plot, the quantiles of the data are plotted against the quantiles of the reference distribution. If the data follows the reference distribution, the points will lie roughly along a straight line. Deviations from this line suggest that the data does not follow the assumed distribution. Q-Q plots are commonly used to visually assess normality.

# Any significant deviations from a straight line point towards existence of skew, thick tails etc.