# 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)

# Basically, categorical variables are redundant in nature. So after encoding them we can remove some of the columns based on the significance factor which p value.

# Also not all categorical variables influence target variable. For example, we have working\_day and holiday variables, but also, we have 7 variables from Monday to Sunday which shows high collinearity. So, removing them is obvious.

# 

**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)

# After creating new columns for categorical variables, we have N columns for N categories. But if we look into them, one of them is obviously redundant and not necessary. So, the argument “drop\_first” makes one of the categorical variable to drop.

**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” is highly correlated 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)

We can validate the LR model by following checks

1. Residual histogram plot should be gaussian

2. Mean value should be zero

3. Variance should be constant

**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 3 features are – “temp”, “weather\_sit Partly cloudy”, “wind\_speed”

# 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 model which works on adjusting a line based on the distance between line and points. Mathematically, it is a line equation trying to fit in the feature plane. But there are some assumptions whether we can use linear regression or not. Those assumptions explained in 4th question.

# Basically this adjustment of line or model according to data is based on residual score. Always the model focuses on increasing r2 score, ultimately it decides the line coefficients.

# And “p” value, VIF value used in deciding whether a variable/feature is significant or not.

**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>

# It is a set of four datasets that have nearly identical statistical properties but reveal very different distributions and relationships when plotted.

# Each dataset in the quartet has similar values for mean, variance, correlation, and linear regression line.

**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>

# It is also known as the Person correlation coefficient, is a measure of the linear correlation between two variable. It quantifies the degree and direction of relationship between two variables with a value ranging from -1 to 1.

# -1 indicates a perfect negative relation

# +1 indicates a perfect positive relation

# 0 indicates not linear relation

**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 process where we adjust the range and distribution of numerical values, so the they fall into specific range or have particular statistical properties.

# Scaling improves model performance, avoid dominance of features, making data comparable.

# Normalized scaling - Min Max scaling

# Standardized scaling – mean = 0, SD = 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>

# It represents multi-collinearity. If two features are perfectly collinear, then VIF becomes infinite. It is a problem because this makes regression coefficients unstable and can lead to overfitting, large standard errors and high sensitivity to small changes in data.

# So, always remove highly collinear features

**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>

# Q-Q is a Quantile-Quantile plot, a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically normal distribution. It calculates the quantiles, plot the points on a scatter plot of each quantile and interpret the linearity