# 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)

# Analysis of the categorical variables from shows that bike rental rates are likely to be higher in

# • Demand is higher in the summer and fall seasons, especially when the weather is clear and

# pleasant

# • Bike rental counts have increased from 2018 to 2019

# 

**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 creating dummy variables for categorical data, it's important to set drop\_first=True to prevent multicollinearity in regression models. It helps in reducing the extra column during dummy variable creation. If we have n categorical levels, then we need to use n-1 columns to represent dummy variables

**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 temp and atemp variables have 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)

1. Normalization
2. Linear Relationship
3. Multicollinearity
4. Homoscedasticity
5. No pattern or auto-correlation

**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: Temperature has the most significant positive impact on bike rental demand. As the temperature rises, the demand for bike rentals increases significantly

# Year(yr): The year has a positive impact on demand. Over time, there has been an increasing trend in bike rentals.

# Season

# 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 fundamental technique in statistics and machine learning that helps us understand and predict the relationship between variables.

# Linear Regression examines how one variable influences another. It helps us draw the best-fitting straight line through this data, allowing us to predict test scores based on study hours

# 

# Simple Linear Regression: Simple linear regression is used when there is only one independent variable that affects the dependent variable. The linear equation for simple linear regression is: Y= β0+β1X+ε

# Multiple Linear Regression: Multiple linear regression extends the concept to situations where there are multiple independent variables. The linear equation for multiple linear regression is Y=β0+β1X1+β2X2+...+βpXp+ε

# Assumptions:

# The relationship between the independent variables and the dependent variable is linear.

# The error terms (ε) are independent of each other.

# The variance of the error terms is constant across all levels of the independent variables.

# The error terms follow a normal distribution.

# Independent variables are not highly correlated with each other.

# Interpretation:

# Linear Regression allows for the interpretation of the coefficients.

# Evaluation: To asses the performance of a linear regression model, various metrics can be used i.e. R-squared, Adjusted R-squared

# Predications: Once the model is trained, it can be used to make predictions on new or unseen data by plugging in the values of the independent variables into linear equation

**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 collection of four datasets, each containing eleven (x, y) points, that were constructed by statistician Francis Anscombe in 1973. These datasets are designed to have nearly identical simple statistical properties—such as mean, variance, and correlation—but when visualized, they reveal strikingly different distributions and relationships between variables.

# Despite these nearly identical statistical summaries, plotting each dataset reveals distinct patterns:

# Dataset I: Displays a linear relationship between x and y, fitting the regression line well.

# Dataset II: Exhibits a clear nonlinear relationship, forming a curve.

# Dataset III: Contains an outlier that significantly influences the regression line.

# Dataset IV: Features a vertical line of points with one outlier, creating a perfect correlation that is 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, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Its value ranges from -1 to 1, where:

# +1 indicates a perfect positive linear relationship: as one variable increases, the other also increases proportionally.

# -1 indicates a perfect negative linear relationship: as one variable increases, the other decreases proportionally.

# 0 indicates no linear relationship between the variables.

**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 technique performed in pre-processing during building a machine learning model to

# standardize the independent feature variables in the dataset in a fixed range.

# The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

# The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

**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>

# A VIF value of infinity indicates perfect multicollinearity, meaning one predictor variable can be perfectly predicted by a linear combination of the other predictors in the model. This situation arises when the R² value, obtained by regressing one predictor against all others, equals 1, leading to division by zero.

# Perfect multicollinearity can occur if two or more variables are exact linear combinations of each other, such as when duplicate columns exist in the dataset. In such cases, the regression model cannot distinguish between the perfectly correlated variables, resulting in infinite VIF values.

# To address this issue, it's essential to identify and remove or combine the collinear variables, thereby reducing multicollinearity and ensuring the regression model can provide reliable coefficient estimates.

**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>

# The quantile-quantile plot is a graphical method for determining whether two samples of data

# came from the same population or not. A q-q plot is a plot of the quantiles of the first data set

# against the quantiles of the second data set.

# If the Q-Q plot shows a straight line, it provides evidence that the residuals are normally

# distributed, which is one of the key assumptions of linear regression.