# 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)

# There are several categorical variables, namely season, month, year, weekday, working day, and weather situation, which have a significant impact on the dependent variable 'cnt'. The figure below illustrates the correlation among these variables.

# 

# 

**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)

# The purpose of creating dummy variables is to convert a categorical variable with 'n' levels into 'n-1' binary columns, where each column indicates the presence (1) or absence (0) of a particular level. The drop\_first=True parameter is used to drop the first level, resulting in 'n-1' columns. This helps reduce multicollinearity among the dummy variables. For example, if there are 3 levels, setting drop\_first=True will eliminate the first column.

**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 exhibit the highest correlation with the target variable 'cnt' compared to the other variables.

**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 have validated the assumption of Linear Regression Model based on below 5 assumptions:

1. Normality of error terms: Error terms should be normally distributed

2. Multicollinearity check: There should be insignificant multicollinearity among variables.

3. Linear relationship validation: Linearity should be visible among variables

4. Homoscedasticity: There should be no visible pattern in residual values.

5. Independence of residuals: No 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)

# The top three features that have the most significant impact on explaining the demand for shared bikes are temperature, year, and Weather.

# 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 form of predictive modelling technique which tells us the

# relationship between the dependent (target variable) and independent variables

# (predictors). Since linear regression shows the linear relationship, which means it finds

# how the value of the dependent variable is changing according to the value of the

# independent variable. If there is a single input variable (x), such linear regression is

# called simple linear regression. And if there is more than one input variable, such linear

# regression is called multiple linear regression. The linear regression model gives a

# sloped straight line describing the relationship within the variables.

# A regression line can be a Positive Linear Relationship or a Negative Linear

# Relationship. The goal of the linear regression algorithm is to get the best values for a0

# and a1 to find the best fit line and the best fit line should have the least error.

# In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which

# helps to figure out the best possible values for a0 and a1, which provides the best fit line

# for the data points.

**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 can be defined as a group of four data sets which are nearly

# identical in simple descriptive statistics, but there are some peculiarities in the dataset

# that fools the regression model if built. They have very different distributions and appear

# differently when plotted on scatter plots. It was constructed to illustrate the importance of

# plotting the graphs before analysing and model building, and the effect of other

# observations on statistical properties. There are these four data set plots which have

# nearly same statistical observations, which provides same statistical information that

# involves variance, and mean of all x,y points in all four datasets.

# 

# 1st data set fits linear regression model as it seems to be linear relationship

# between X and y

# 2nd data set does not show a linear relationship between X and Y, which means

# it does not fit the linear regression model.

# 3rd data set shows some outliers present in the dataset which can’t be handled

# by a linear regression model.

# 4th data set has a high leverage point means it produces a high correlation

# coefficient.

# Its conclusion is that regression algorithms can be fooled so, it’s important to do data

# visualization before build machine learning model.

**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 numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

**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 means you are transforming your data so that it fits within a specific scale. It is one

# type of data pre-processing step where we will fit data in specific scale and speed up the

# calculations in an algorithm. Collected data contains features varying in magnitudes, units

# and range. If scaling is not performed than algorithm tends to weigh high values

# magnitudes and ignore other parameters which will result in incorrect modelling.

# Difference between Normalizing Scaling and Standardize Scaling:

# 1. In normalized scaling minimum and maximum value of features being used whereas in

# Standardize scaling mean and standard deviation is used for scaling.

# 2. Normalized scaling is used when features are of different scales whereas standardized

# scaling is used to ensure zero mean and unit standard deviation.

# 3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling

# is not having or is not bounded in a certain range.

# 4. Normalized scaling is affected by outliers whereas standardized scaling is not having

# any effect by outliers.

# 5. Normalized scaling is used when we don’t know about the distribution whereas

# standardized scaling is used when distribution is normal.

# 6. Normalized scaling is called as scaling normalization whereas standardized scaling is

# called as Z Score Normalization.

**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>

# If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

**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 (q-q) plot is a graphical technique for determining if two data sets come

# from populations with a common distribution.

# Use of Q-Q plot:

# A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second

# dataset. By a quantile, we mean the fraction (or percent) of points below the given value.

# That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and

# 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come

# from a population with the same distribution, the points should fall approximately along this

# reference line. The greater the departure from this reference line, the greater the evidence

# for the conclusion that the two data sets have come from populations with different

# distributions.

# Importance of Q-Q plot:

# When there are two data samples, it is often desirable to know if the assumption of a

# common distribution is justified. If so, then location and scale estimators can pool both data

# sets to obtain estimates of the common location and scale. If two samples do differ, it is also

# useful to gain some understanding of the differences. The q-q plot can provide more insight

# into the nature of the difference than analytical methods such as the chi-square and

# Kolmogorov-Smirnov 2-sample tests.