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)

Effect of Categorical variables
weekday- cnt is highest on Tuesday and lowest on Sunday.

season- During fall cnt is highest followed by summer, winter and spring. Spring has lowest cnt.

yr- There is sigficant rise of cnt in 2019 compared to 2018

mnth- cnt has positive impact on the month Sep and Oct.

day- cnt is highest on Thursday and lowest on Monday and Tuesday. Though there is no significant difference between highest and lowest value.

holiday- cnt is less in holidays.

workingday - cnt is more in workingday.

weathersit - cnt is highest on Clear, Few clouds, Partly cloudy, Partly cloudy weather followed by Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist.cnt
is lowest during Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. We donot have any records for cnt during
Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog in our dataset.

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)
To reduce redundancy. If we have p level of categorical variable the number of dummy variables, we need is p-1. It is good for interpretation. Without that an extra effect will come with base state which negatively impact the regression model by causing instability in coefficient estimation.

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)

emp has the highest correlation with target variable cnt.

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)

Considered the below three values:
p-values – For all independent variables it is <0.05
R-Square- The value is 0.800(80%) which is pretty high
P(F-Statistics) – It is almost 0

IVF- For all features <5 which means multicollinearity has been removed from the model to the desired extent.

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)

Final model equation is as below:

cnt= 0.0414 + 0.2369 * yr - 0.0700 * holiday + 0.5912 * temp -0.0893 * windspeed + 0.0930 * season_summer + 0.1253 * season_winter -0.2383 * weathersit_Ligh_Snow + 0.0997 * Sep + 0.0561 * Oct -0.0362 * Monday -0.0351 * Tuesday -0.0383 * Wednesday

So based on the above equation we can say the top three features which have significant impact(positive) on the demand of shared bikes are

1) temp
2) yr – significant rise in 2019 compared to 2018
3) season (summer and winter)

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 and widely used machine learning model predicting a continuous outcome (dependent variable) based on one or more predictor variables (independent variables). It is based on supervised learning.

There are two types of linear regression-

Simple Linear Regression- Only one independent variable predicts the value of dependent variable.

Multiple Linear Regression- More than one independent variable predict outcome of dependent variable.

A linear line shows the relationship between dependent and independent variables. X axis

represent the value of independent variables and Y axis represent the value of independent variable. There can be both positive and negative relationship between dependent and independent variable.

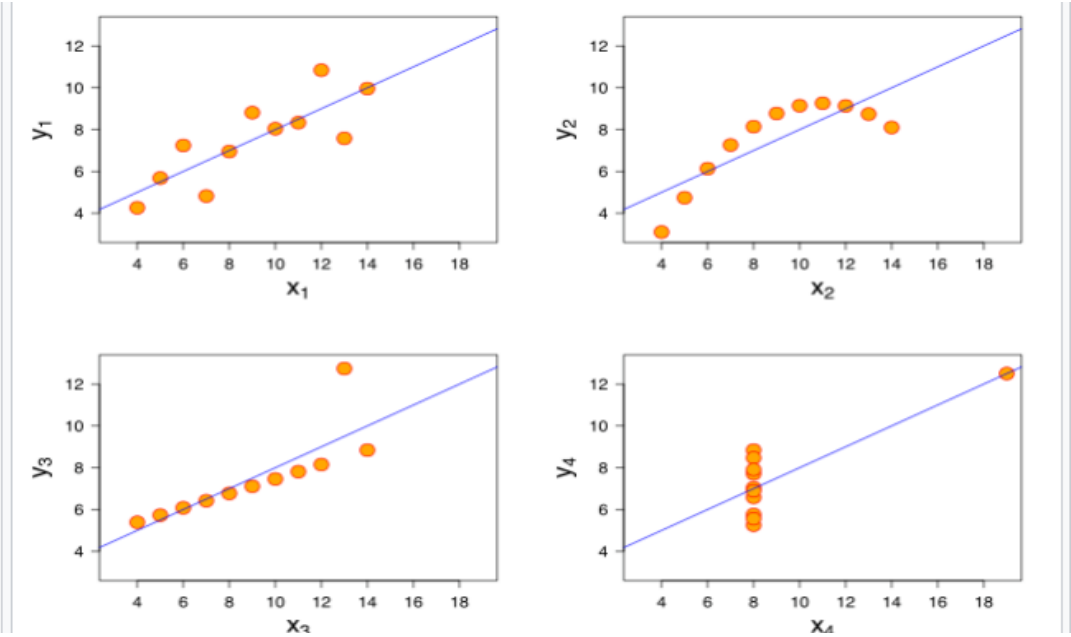
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet comprises of four datasets having same descriptive summary statistics (same mean, variance, R-squared, correlations and linear regression line) but having different representations when we scatter plots on a graph. Each of four datasets in Anscombe's quartet has 11 x-y pairs of data. When plotted each dataset seems to have unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

Anscombe's Quartet Four Datasets



Data Set 1: x and y fit the linear regression model pretty well.
Data Set 2: x and y have a clear relationship but that is not-linear. So, it cannot fit the linear regression model.
Data Set 3: Though x and y have a linear relationship but it shows single outlier involved in the

data set and outliers cannot be handled by the linear regression model.
Data Set 4: All the x values are exactly the same except for one point which is outlier. So, it cannot fit the linear regression 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, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It describes how well the linear relationship between two continuous variables can be measured by a straight line.

It has value between -1 to 1

1 -> Perfect positive linear correlation

-1->Perfect negative linear correlation

0->No linear correlation 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 method in data preprocessing which is applied to independent variables to normalize data within a particular range. It is useful for faster interpretation and speeding up calculations in algorithm.

When data in dataset is highly varies in magnitude, units and range without scaling algorithm will only take magnitude ignoring units. It is not correct for modelling. In this case scaling should be performed to bring data in same level of magnitude. Scaling only affects the coefficient not p-value and R-Square.

Normalized scaling- It bring all the data between 0 and 1.
Sklearn.preprocessing.MinMaxScaler helps to implement Normalization in python.
x=x-min(x)/max(x)-min(x)

Standardized scaling- It replaces all the data by their Z-score. It brings all the data into a standard normal distribution which has mean 0 and normal deviation of 1.
Sklearn.preprocessing.Scale helps to implement Standardize scaling in python.
x=x-mean(x)/sd(x)

There is a chance of losing data specially outliners in case of 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>

VIF is infinite meaning there is a perfect correlation between independent variables. Higher VIF means higher degree of multicollinearity. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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 the graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from same population or not. It is particularly helpful for assessing whether a dataset is normally distributed or it follows some other known distribution pattern.

In the context of linear regression, it's specifically used to check the normality of the *residuals* (the errors) of the model.
