Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   There are total of 6 categorical variables in the dataset. Season, mnth, holiday, weekday, workingday and weathersit.

   - For season – Most of the bookings were happening with season 3 with median of around 5000 booking, followed by season 2 and season 4, This shows this variable can be a good predictor.
   - For mnth – Most of the bookings were happening in mid months, April may with 4000 average. So this also shows it's a good predictor variable.
   - Weathers it– Most of the booking were happening during, wathersit value is 1, which is when clear, few clouds or partly coulds, so this is also a good predictor variable.
   - Holiday – Most booking was happening when it a holiday. So its clearl;y biased, so not a good predictor
   - Weekday – shows the same trend on all weekday, so it's not of great use to prediction.
   - Workday – shows most booking on working day with 5000 booking on median. So could be a good predictor.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   To avoid dummy variable trap, we need to drop one of the dummy category so we drop the first dummy category variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Temp, atemp and cnt has the highest correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - By histplot of residuals, and checking for error terms normally distributed with mean 0
   - Linear relationship between x and y using pair plot
   - And checking for multicollinearity using VIF calculation.

   - Train R^2 :0.819
   - Train Adjusted R^2 :0.815
   - Test R^2 :0.78
   - Test Adjusted R^2 :0.77

Seems to be a good model.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Based on features to select param on my ref, the data was changing, tried with different no of features on ref initialisation and continued learning model till I got the better r-squared value and p-values for all the features.

Based on the observations, the following were the important features,

```
yr              0.231558
holiday        -0.077744
workingday      0.018088
temp            0.573845
windspeed      -0.161943
season_2        0.073553
season_4        0.128051
weathersit_2   -0.075772
weathersit_3   -0.276427

top 3 could be year, temp, weather situation.
```

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Linear regression, is a supervised machine learning method[which has labelled data and predicts output based on labelled input], which finds a linear equation, which describes the correlation between dependent variable and independent variable.

   We can have 2 types on linear regression.

   - With one independent variable – simple linear regression
   - With multiple independent variables – Multiple Linear regression

   Equation of the simple linear regression is given bt,

   y = b0 + b1. X

   where b0 is the intercept

   and b1 is the slope

   Equation of the simple linear regression is given bt,

   y = b0 + b1. X1 + b2.x2

   where b1,b2 are coefficients and x1, x2 are all dependent variables.

   Y is the independent variable.

   Also the sign of coffients designates, whether it is positively correlated or negatively correlated.

Linear regression helps in predicting the dependent variable and also it helps in telling how accurate the prediction is usinh rsquared and pvalues.

2. Explain the Anscombe's quartet in detail. (3 marks)

Sometimes it is possible that the data could be summary statistics could be same, but when done the exploratory graph analysis of these dataset there relation look totally different, Hence the Anscombe's quartet states that, it is not only enough to rely on summary statistics but it is also important to do the exploratory data analysis and understand the correlation.

The quartet is still often used to illustrate the importance of the looking at the data graphically, before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties

3. What is Pearson's R? (3 marks)

It is the measure of how close the observations are to best line fit. The pearson's correlation coefficient also tells you whether the slope of the line of the best fit is negative or positive. When the slope is negative, the r is negative, when the slope is positive the r is positive. When the r is 1 or -1, all the points fall exactly on the line of best fit.
If the r is greater than 0.5 or less than 0.5, then points are close to the line of best fit
If the r is between 0 and 0.3 or 0 and -0.,3 then points are far from the line of best fit.
If r is 0 then line of best fit is not helpful in describing the relationship between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is data preprocessing technique which is used to transform the features or variables in a dataset to a similar scale, The purpose the scaling is that all features contribute equally to the model and to avoid the domination of features with learge values.
Feature scaling becomes important when dealing with datasets containing features that have different ranges, units of measurements, or order of magnitude. In such cases ariation in feature values can lead to biased model performance or difficulties during the learning process.

Normalisation is a technique n which values are shifted and rescaled so that the values ranges between 0 and 1. It is also known as min max scaling.

Standardisation is a technique where are the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If all the independent variables are orthogonal to each other, then the VIF=0, If there is a perfect correlation then the VIF is infinity, A large value of VIF indicates that there is a correlation between the variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plots are also known as quantile- quantile plots, they plot the quantiles of the sample distribution again the quantiles of the theoretical distribution. Doing this helps us determine whether the datasets follows any particular type of probability distribution like normal, exponential or uniform.