Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. The results of the categorical variables analysis for the number of users is as follows:

- Weekday: The median of the number of users is highest for Thursdays. The maximum users occurred on a Sunday and Friday had the highest number of average users.
- Month: September had the highest summary statistics across median, maximum and average users.
- Year: 2019 had more users than 2018
- Season: Fall season had the best user statistics across median, average and maximum.
- Weather: The weathersit=1 category has the best user statistics
- Holiday?: Non-holidays has better user statistics
- Working day? : The metrics were similar for both the categories.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. The drop_first=True allows us to drop the additional redundant column. This results in reduced correlation among the created dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. Both the ‘temp’ and ‘atemp’ variables have similarly high correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. The assumptions were validated by plotting the distribution of the residuals and the scatter plot of the predicted v/s true target values.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. Top features:

1. atemp: Warmer temperatures seem to encourage more users
2. weathersit_3: Harsher weather discourages bicycle users
3. yr: More users used it in the second year, may suggest a trend of increasing users as the service becomes popular with each year.

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a supervised learning algorithm that tries to find the best fitting affine relationship between the target/dependent variable and a set of feature/independent variables. The following equation is used for establishing the target and feature variables.

$$y = a\_0 + a\_1 * x\_1 + a\_2 * x\_2 + \ldots \quad (1)$$

where y is the target variable and *x_1, x_2, …* are the feature variables. The aim of linear regression is to determine the coefficient values *a_0, a_1, a_2 …* such that the mean square error/some other error metric over the set of training data/observations is minimized. The linear regression algorithm has the following steps for a Gradient descent-based optimization.

1. The coefficient values are randomly initialized.
2. The relationship shown in Eq. (1) is used to determine the predicted values for each point in the training dataset.

3. The error in predicted target variable is computed for each of the training data point and accumulated in terms of metrics such as mean square error (MSE) or mean absolute error (MAE).
4. The partial derivative with respect to each of the coefficient are computed at each of the training data points and accumulated as the cost for each coefficient.
5. Each coefficient is updated as a difference between it earlier value and the product of a learning rate and the accumulated partial derivative cost for the variable
6. Go to step 2.

The above steps are repeated till the coefficient values converge to within a tolerance value.

Q2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet refers to a set of 4 datasets, each containing 11 data points, published by Francis Anscombe in 1973. All the 4 datasets have nearly identical descriptive statistics – sample mean and standard deviation, for x and y, correlation between x and y, linear regression coefficients and R-square. The purpose of the publication was to demonstrate the need for graphical/visual analysis of datasets and not just to rely on summary statistics for data analysis.

Q3. What is Pearson's R?

Ans. Pearson's R is used to quantify the correlation between two variables. It is determined as the ratio of the covariance of the two variables and the product of their standard deviations. This ensures the normalization of the covariance and the values range between [-1, 1]. The negative values indicate that one variable's value decreases linearly with the other. Similarly positive values indicate linear increase. Pearson's R indicates the existence of linear relationships only and ignores other types of relationships.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling?

Ans. Scaling is a process of transforming the feature and/or target variables to within a specified range or to a distribution. Scaling helps the modelling in multiple ways:

- It allows the joint analysis of feature variables irrespective of their units and order
  of magnitude.
- It allows the optimisation algorithms to consider the cost/penalty functions more uniformly irrespective of the features' range/scale.

Normalization	Standardization
Transforms values to a range of [0,1].	Transforms values such that the resulting data has zero mean and unit SD.
Assumes uniform distribution of the values.	Applicable to datasets where the underlying phenomenon that generates the points follows a Gaussian distribution
Outliers may reduce the resolution of the data values.	Lesser effect of outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. Infinite VIF indicates perfect correlation. It signifies that a subset of the other variables has manged to predict the variable with infinite VIF perfectly across all data points. It generally indicates redundant features such as *weight in kgs* and *weight in lbs*. Since the values are linearly related with a constant factor having both the variables in the model does not make sense.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. Q-Q or quantile-quantile plots compare the quantiles of a dataset against the quantiles of a desired/test distribution. It is used to compare the similarity of a dataset to a standard/test distribution. In linear regression, the Q-Q plot can be used to determine if the residuals obtained from the linear regression model follow a normal distribution. One of the primary assumptions of the linear regression modelling is the normal distribution of the errors, which indicates the residuals to be randomly distributed and validates the linear relationship.