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?
Ans: Since we have transformed the categorical variables into dummies variables with value 0 or 1, we can infer the effect of categorical variables on the dependent variable by examining the estimated coefficients for the corresponding dummy variables in the linear regression model.

2. Why is it important to use drop_first=True during dummy variable creation?
Ans: When creating dummy variables for a categorical variable with **k** categories, the convention is to create **k-1** dummy variables. Because the dummy variables are perfectly collinear, they can be perfectly predicted from one another. Thus, by adding "drop_first=True", we could avoid multicollinearity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: Column "**atemp**" is the one that has highest correlation with the target variable, with the value of 0.63. Even columns "**casual**" and "**registered**" have higher correlation with the target variable with values of 0.67 and 0.95 respectively, these 2 columns compose into the target variable and thus should be ignored/removed in the analysis.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans: In this analysis I used Residual analysis and R-squared to validate the assumption of linear regression after building the model on the training set.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans: Based on the OLS summary of the final model, there are 3 top features contributing significantly. Those are "yr_2019" (original column "yr" with value of 2019), "weathersit_light_rain_snow" (original column "weathersit" with value of 3 for Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds), and "season_spring" (original column "season" with value of 1 for Spring).

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans: Linear regression is a supervised machine learning algorithm that is used for predicting a continuous target variable based on one or more independent variables.

The basic idea behind linear regression algorithm is to find the best line that fits the data points in such a way that it minimizes the sum of the squared errors between the predicted values and the actual values. This line is called the regression line or the best-fit line.

The linear regression model can be expressed as:

y = b0 + b1 * x1 + b2 * x2 + ... + bn * xn

where y is the target variable, x1, x2, ..., xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients that determine the slope and intercept of the regression line.

Linear algorithm can be categorized further into 2 types:
+ Simple Linear Regression: When there is only one independent variable, thus it is called simple linear regression.
+ Multiple Linear Regression: When there are multiple independent variables, hence it is called multiple linear regression.

There are several steps to implement a linear regression algorithm:
+ Data collection: to collect data, including the target variable and the independent variables.
+ Data preprocessing: this step involves some tasks like cleaning, normalization, feature scaling.
+ Model fitting: this step is to fit a linear regression model to the data
+ Model evaluation: Using a set of metrics such as R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to determine how well the model fits the data.
+ Prediction: The trained model can be used to predict the target variable with new data.

2. Explain the Anscombe’s quartet in detail?
Ans: Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe to demonstrate the importance of visualizing data before applying further algorithms on dataset. Each dataset consists of 11 pairs of (x, y) values, and all four datasets have the same statistical summary, such as mean, variance, and correlation. However, when the datasets are plotted, they look different from each other, illustrating the need to explore the data visually.

Below is the details of four datasets in Anscombe's quartet:
+ Dataset **1**: This dataset has a linear relationship between x and y, with a slope of approximately 0.5. The correlation coefficient between x and y is 0.816, and the regression line has an R-squared value of 0.67.

+ Dataset **2**: This dataset has a non-linear relationship between x and y, with a quadratic curve. The correlation coefficient between x and y is 0.816, and the regression line has an R-squared value of 0.67.

+ Dataset **3**: This dataset has a linear relationship between x and y, but it is heavily influenced by an outlier. The correlation coefficient between x and y is 0.816, and the regression line has an R-squared value of 0.66.

+ Dataset **4**: This dataset has a perfect linear relationship between x and y, except for one point that is an extreme outlier. The correlation coefficient between x and y is 0.816, and the regression line has an R-squared value of 0.17.

In short, Anscombe's quartet illustrate the importance of visualizing data and exploring different models when analyzing data.
It also highlights the importance of exploring different models for fitting the data and using multiple metrics to evaluate the model's performance

3. What is Pearson’s R?
   Ans: Pearson’s R , or Pearson correlation coefficient, is a measure of the linear relationship between two variables. It is used to measure the strength and direction of the linear association between two variables, where a value of 1 represents a perfect positive correlation, a value of -1 represents a perfect negative correlation, and a value of 0 represents no correlation.
   The Pearson correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations.
   Below is the formula or Pearson R:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans: Scaling is a preprocessing step in machine learning that involves transforming the features of a dataset to be on a similar scale. Scaling is performed to ensure that each feature contributes equally to the model and to avoid bias towards certain features that have larger values than others.
   Reason to perform scaling: machine learning algorithms are sensitive to the scale of the input features. If the features have different scales, then some features may dominate the distance calculation, which would lead to incorrect results.

Normalization (or min-max scaling) is the process to rescale the values into a specific range, normally between 0 and 1. We can achieve this by subtracting the minimum value of the variable from each value and then dividing the result by the range of the variable (meaning: the difference between the maximum and minimum values).
Formula: $\mathbf{x_{new} = (x_i – x_{min}) / (x_{max} – x_{min})}$

Standardized scaling (also known as z-score scaling) involves the transformation of the values of a variable so that they have a mean of 0 and a standard deviation of 1. We can achieve subtracting the mean of the variable from each value and then dividing the result by the standard deviation of the variable. This results in a transformed variable where the mean is 0 and the standard deviation is 1.
Formula: $\mathbf{x_{new} = (x_i} – \mu) \mathbf{/ s}$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: When the value of VIF is infinite, it indicates that perfect multicollinearity exists among the predictor variables in the model. Perfect multicollinearity may occur when there's one or more predictor variables perfectly predicted by a linear combination of other predictor variables in the model. In other words, one predictor variable could be expressed as a linear combination of other predictor variables with perfect accuracy. This situation causes the estimation of the regression coefficients to be impossible.
To address this issue, it is essential to identify the source of multicollinearity and take appropriate measures, such as removing one of the correlated variables or transforming the variables.

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

Ans: A Q-Q plot, also known as a quantile-quantile plot, is a graphical tool that can be used to evaluate the distributional similarity between two sets of data. Q-Q plots can be used to validate the assumption of normality, which is a critical assumption for the validity of many statistical inference procedures, such as hypothesis testing and confidence intervals.

A Q-Q plot compares the quantiles of the two datasets by plotting them against each other. The x-axis of the plot illustrates the quantiles of one set of data, while the y-axis illustrate the quantiles of the other set of data. If the two sets of data have a similar distribution, the points on the Q-Q plot should fall along a straight line. Deviations from a straight line indicate a deviation from the normal distribution.

The use and importance of a Q-Q plot in linear regression:
+ Identifying departures from normality: The Q-Q plot can help identify any departures from normality in the residuals, such as skewness or heavy tails. This information can be used to decide whether the assumption of normality is reasonable.
+ Model selection: The Q-Q plot can also be used to compare the distribution of residuals across different linear regression models. The model with residuals that conforms a normal distribution is preferred.

+ Validating the assumption of normality: The Q-Q plot is used to check if the residuals in a linear regression model are normally distributed. In linear regression, normality of residuals is a key assumption, and violation of this assumption could affect the validity of the statistical inferences made from the model.