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:

- The number of rides taken are the highest in fall, but has seen a significant dip in Spring
- The number of bikes taken are increased in 2019 compared to 2018
- The number of bikes taken are increasing from January to july and then constantly decreasing to the year-end. Less bikes are taken in the starting months of the year.
- Lesser bikes are booked on holidays
- Above 4000 bikes are booked on all the weekdays
- Almost the same number of bikes are rented on both working-day and weekends/holidays
- No bikes are booked when there is heavy rain type of weather. There is a severe dip in bikes rented when there is light snow/rain type of weather.
- Most rides are booked when the weather is clear.

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

Ans:

- If drop_first= True is not used, there will be more unnecessary columns added during the dummy variable creation.
  For example, if encoding has to be done for 3 values [Furnished, Semi-Furnished and Not furnished]

- **Without** using drop_first = True, the encoding looks like
  Furnished - 100
  Semi-Furnished - 010
  Not furnished - 001

- **With** drop_first = True, the encoding looks like
  Furnished - 10
  Semi-Furnished - 10
  Not furnished - 00

  Encoding for N levels can be done using (N-1) variables, which can be achieved using drop_first=True during dummy variable creation.

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

Ans:

- Variable 'temp' has the highest (0.63) correlation with the target 'cnt' carriable.

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

Ans:

- Initially, Linearity between the dependent variable and independent variables is checked using a scatter plot and such relation is found.
- During residual analysis, the Error terms are checked if it is normally distributed and has a mean=0.
- Collinearity is visualized using heatmaps and eliminated in the model design phase as VIF is included. Validations included checking if the VIF is less than 5 for every feature included in the model.
- Homoscedasticity is verified by plotting a scatter plot between the residuals and the predicted values 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:

- Spring season with a negative coefficient of 0.29
- Year with positive coefficient of 0.25
- Windspeed with a negative coefficient of 0.2

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans:

a) Linear Regression is a Machine Learning algorithm which performs regression task and is based on Supervised Learning.
b) It predicts a target variable based on one or several predictor variables and is used to find relationships between variables

Steps involved in Linear regression:

- Data preparation and understanding:
    - Data is loaded from a file into a data frame
    - Data is cleaned, missing values in the data are handled and only necessary columns for the analysis are sent to the next step
    - Outliers are also handled to make the analysis effective.
- Data Visualization:
    - Data of continuous variables is visualized using scatterplots to check linearity

    - Categorical data is visualized using box-plots and the dependency with the target variable is checked.
- Visualize Correlation:
    - The features with more correlation are redundant and will effect the model. So, these are visualized here.
- Dummy Variable Creation:
    - For categorical variables dummy variables are created. For 'N' levels in the data, (N-1) dummy variables are defined.
- Train-test split:
    - The total data is split into training and testing data with a ratio 70:30 or 80:20.
    - Scaling of the data is also done either using Normalization or Standardization techniques.
- Building a model:
    - By Feature Selection: A linear model can be built by identifying the most correlated predictor with the target variable and continuing in the same process.
    - By Feature elimination: Selecting all the predictors initially and manually eliminating the impact-less features on the model
    - By Automated Approach: Using Recursive Feature Elimination (RFE) by selecting Top 'N' features.
- Residual Analysis:
    - Using the built model, predict values within the training set and compute error as we already have the actual value.
- Making predictions:
    - Using the same model, predict the target variable in the test set that is separated previously.
- Model Evaluation:
    - Compare the R-Squared value obtained using the model in both training and test datasets and check if they are close/approximately same.
    - The R-Squared value Is the variance which the model can explain using the predictors.

2. Explain the Anscombe's quartet in detail.

Ans:

- Anscombe's quartet consists of four datasets that have nearly identical descriptive statistics, but have different distributions that can be visualized when plotted on a scatter plot.
- Each dataset consists of 11 points.
- The four datasets generates completely different plots that no regression algorithm can predict.

- When plotted, the quartet can be analysed as:
    - Dataset1: That fits a Linear regression model pretty well.
    - Dataset2: Has Curvilinear structure and cannot fit a Linear regression model well

    - Dataset3: There is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line
    - Dataset4: This shows an example when one high-leverage point is enough to produce a high correlation coefficient.
- This is used to describe the importance of visualizing data graphically before analyzing according to the relationship of the datapoints.

3. What is Pearson's R?

Ans:

- Pearson product-moment correlation coefficient, or Pearson's R is a statistical measure of linear correlation between two variables.
- It has correlations ranging between -1.0 and +1.0.
- Pearson's R is the covariance of the two variables divided by the product of their standard deviations
- Mathematically,

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

- Requirements to apply Pearson's R:
    - Scale of measurement should be a interval/ ratio
    - Variables are approximated to be normally distributed
    - Variables should have linear relationship
    - No outliers should be present in the data.

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

Ans:

- Scaling is done on the data for better interpretation if its on a pre-defined range.
- For comparability, all the variables should be scaled into a particular range as it is difficult to compare multiple variables with multiple ranges.
- Though scaling does not affect the model, the coefficients will have a great impact as the measure is completely different.
- So, for interpretation of relationships between variables, scaling is done.
- Scaling can either be done by Normalization or Standardization techniques.
- **Normalization**:
    - Also known as Min-Max Scaling
    - It brings all the datapoints to a range of 0 and 1
    - As the range is pre-defined, interpretation is easier during evaluation.
    - It brings outliers too within the range so they are handled.
    - Widely used technique.
- **Standardization**:
    - It replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
    - Values are distributed around 0.
    - As there is no particular range like normalization, interpretation is relatively difficult.
    - Outliers are not handled.

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

Ans:

- VIF is given by the formula,

  VIF = 1 / (1 – R-Squared)

- R-Squared has the range of 1 to infinite.
- So there is a chance of R-Squared value being equal to '1' (i.e., the target and predictor variables are absolutely correlated)

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

Ans:

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
- Quantile points are a fraction of the data in the dataset.
- Quantile information can be obtained like 50% Quantile is the median of the data

Use:

- Q-Q plot can be used even when the sample sizes of the datasets are not equal.
- Validates whether two datasets come from a common distribution.
- To check whether two datasets have a similar distribution shape and tail behavior.