**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)**

**Ans**:

We see the 'season', 'mnth', 'weatherist' and 'weekday' are categorical variables in the data set. Based on the derived model below analysis found:

- More bike rentals demanded on the winters as compared to the summer and spring.
- We had observed that the September month had higher use of rentals.
- In terms of days the maximum focus was on days like Wed, Thurs and Sat and more on working day.

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

**Ans:**

We will set this parameter for the following reason.

- If we create the dummy variables (n) for a categorical variable. There will be one extra column created and we can derive the same using the other variable(n-1). So that it reduces the correlation created among the dummy variables.
- This will Avoiding Multicollinearity.

For example, we have 3 below dummy categorical variables created for 'furnishingstatus' variable. We don't need three columns. We can drop the `furnished` column, as the type of furnishing can be identified with just the last two columns mentioned below.

- `00` will correspond to `furnished`
- `01` will correspond to `unfurnished`
- `10` will correspond to `semi-furnished`

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

**Ans:**

We see the highest correlation with the 'temp' variable, and it is more linearly dependent on the target 'cnt' variable. and the correlation value is 0.63. It indicates the strongest linear relation ship compared to the other variables.

**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

**Ans:**

We can validate the assumptions using the below parameters:

1. **Linearity**: The observed and predicted variables should show a linear relationship. The derived model 'cnt' is having the linear relationship with 'temp' variable.
2. **Normality**: The distribution plot of the residuals should show as normal distribution.
3. **Homoscedasticity**: The residuals plot showing the random scatter.
4. **No Multicollinearity**: VIF value should show below 5 for all the dependent variables. Hight VIF(>10) indicates the high multicollinearity.

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

**Ans**:

Based on the derived model the top 3 highest absolute coefficients values are:

1. **Temp**: With coefficient value of '0.412'. Which indicates the strong positive impact on the bike rentals.
2. **Yr –** With coefficient of '0236' . Which indicates the positive impact on the bike demand.
3. Light snow:- With coefficient of ' -0.29'.Which indicates the significant negative impact on the bike rentals.

## General Subjective Questions:

### 1. Explain the linear regression algorithm in detail.

**Ans**:

The Linear Regression is a statical method used to modelling the relationship between the dependent variables one or more of the independent variables. There are two types of linear regressions ,1. Simple linear and 2. Multiple linear regression.

Below steps for linear regression algorithm:

1. **Model Representation: Using the best fit line.**

   $Y = \beta_0 + \beta 1 . X_1 + \beta_2 . X_2 + \beta_3 . X_3 \ldots\ldots . B_n . X_n$

   Y = Dependent target variable
   $\beta_0$ ,, $\beta_{1..}$, $\beta_n$ = Coefficients of the model.
   $X_1$ ,$X_{2,..}$ $X_n$ = Independent variable.

   Objective of the Model is to minimize the sum of squared the difference of predicted and the actual value.

   $e_i = Y_i - Y_p$ = error term difference between the actual and the predicted value.
   RSS (Residual Sum of Squares) = = $e_{i1}^2 + e_2^2 \ldots\ldots\ldots\ldots e_n^2$

   $$RSS = \sum_{i=0}^{n} (Y_i - \beta i - \beta i . Xi)^2$$

2. **Model evaluation:**

   Linear regression model is assessed using the two metrics.

   1. **R-Squared:**

      R2= (1- RSS/TSS)
      RSS= Residual Sum of Squares, TSS= Sum of the error of the data from mean.
      R2 is between 0 to 1. The more the R2 score vale the better the model fits your data.

   2. **Residual Standard Error (RSE)**

      $$RSE = \sqrt{\frac{RSS}{(n-p-1)}}$$

      RSS= Residual Sum Of Squares,
      n number of observation in the dataset.
      p= number of predictors

      *As the smaller RSE the model is best fit the data set.*

There are few assumptions in linear regression model,

- Linear relationship between the X and Y.
- Error is distributed normally.
- Errors terms are independent of each other.
- Error terms have constant variance.

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

**Ans:**

Anscombe's quartet is contains four data sets with close statistical terms as mean, variance, R2, correlations and the linear regression lines but they are having the different representation when we scatter plots on a graphical representation.

The purpose of the Anscombe's quarter is used to illustrate the importance of the EDA and draw backs if we depend only on the statistics. Also detects the nuances, outliers, trends and diverse relationship in the data sets.

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

**Ans:**

Pearson's R also known as the Pearson Correlation Coefficient(r). It is a statistical measure the strength and direction of the of the linear relationship between the 2 continuous variables. It ranges from -1 to 1.

Formula

$$r = \frac{\sum xi-x^{-})(yi-y^{-})}{\sqrt{\sum (xi-x^{-})2(yi-y^{-})2}}$$

Where:

$xi$ and $yi$ are the individual sample points.

$x^{-}$x$^{-}$ and $y^{-}$y$^{-}$ are the means of the $x$ and $y$ samples, respectively.

Interpretation based on r values:

- **Pearson's value '1':** indicates the perfect positive linear relationship between two variables. If one variable increases, another variable also increases proportionally.
- **Pearson's value '**0' indicates the no relationship between two variables.
- **Pearson's value '-1'** indicates the perfect negative linear relationship between two variables. If one variable increases, another variable decreases proportionally.

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

**Ans:**

Scaling is the data process to normalize the data within range. It is used to bring the multiple variables with different data ranges to specific single range.

Scaling is performed to ensure uniformity, Improve Convergence Speed, Enhance the model performance and prevent the numerical instability. Enormous differences in feature magnitude can cause the instability in algorithms. So that we will follow the below approaches for scaling the dataset.

There are two types of scaling methods. One is Normalized approach, and another is standardized method.

| **Normalized Scaling** | **Standardized scaling** |
|---|---|
| It is also known as min-max scaling. It scales the data ranges between [0, 1] or [ -1 , 1] . | It is also known as Z-Score normalization. It rescales the data to have mean as '0' and standard deviation of '1' but does not impose any date ranges. |
| It is used when the Algorithm do not make any assumption about the data distribution. | It is used when Algorithm make assumptions about the data distribution. |
| Useful when we need features to within specific data range. | Does not bound the data to specific ranges. |
| It preserves the relationship in the actual data. | Data is centred around the zero with a unit of standard deviation. |
| It is sensitive to outliers | Less sensitive to outliers compared to min-max normalization. |

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

**Ans**:

VIF (Variance Influence Factor) value is used to detect the multicollinearity in the regression analysis. The formula for VIF for predictor Xi = $1/\{1-Ri^2\}$

The VIF becomes infinite when perfect correlation exists between two independent variables. We get the R2=1 and then VIF becomes infinite. To solve this problem, we need to drop the variable from the data set which is causing the multicollinearity.

It happens due to below reasons:

1. Perfect Multicollinearity. Predictor is exactly linear relationship with one or more predictors.
2. Duplicate columns: If 2 columns are duplicate both prediction values are same.
3. Constant or near-Constant columns.

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

**Ans:**

A Q-Q(Quantile-Quantile) plot is a graphical technique used to assess the data set come from population with a common distribution.

A Q-Q plot is a scattered plot created by plotting the two sets of quantiles against one another. If both se of quantiles coming from the same distribution, we should see the pints on the Q-Q plot will lie on the reference line.

Importance of Q-Q plot in the linear Regression:

1. Assessing the residuals normal distribution.
2. Detecting the outliers.
3. Evaluating the model Fit.
4. Q-Q model can provide more insight into the nature of the difference than analytical methods.
5. The sample sizes do not need to be equal.