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?**
   a. The categorical variables provide specific behavioural patterns of the customers of BikeSharing. Based on the outcome of the model, we can conclude that:
      i. The BikeSharing is showing a positive trend during the Seasons Summer and Winter.
      ii. The Weather Situations Mist (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) and Snow(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) are one of the causes for a decline in the BikeSharing for the day.
      iii. The BikeSharing is more used by the customers during the months of August and September.
      iv. The year 2019 shows an increased number of users for BikeSharing.

2. **Why is it important to use *drop_first=True* during dummy variable creation?**
   a. The concept of dummy variables is used to convert the categorical values to numeric values. This can also be done via encoding. However, during analysis, as the data may result in fractional numbers, the significance of the value is lost. To save that, we plan to create dummy columns which hold binary values representing each of the value in the categorical variable.

      The values of a categorical variable can be represented in N-1 ways (where N is the number of possible values in the data dictionary). A value of 0 in all the other columns infers that the value is set to the column representing it.

      In other words, if the values A,B,C are in categorical variables, we can create the binary columns B,C. When the value of B,C is 0,0 it will represent that the categorical variables' value is A. Using this way, we will preserve the data significance and also limit the number of dummy columns created.

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   a. Following are the assumptions for a linear regression:
      i. The Error terms are normally distributed: The Error is the difference between the actual given Y value and the value predicted by the model. For a model to be satisfied as a linear regression, this curve shall be normally distributed and the Mean value to be passing thru 0.0.
         1. A Histogram plotting the error differences (b/w y_given – y_predicted) will give the graph confirming our assumptions.
      ii. Multicollinearity: Each of the predictor variables shall be independent or not correlated with any other predictor variables. To validate this assumptions, we've used the heatmap for the data to look ahead for the correlations. Furthermore, we've computed the Variance Inflation Factor(VIF) for each of the independent variables included in the model. The values >5 indicate a correlation between other variables. This is handled by selecting the variables such that the VIF values are low. If any variables' VIF value is found to be high, they are removed from the model building process.
      iii. homoscedasticity: The Linear Regression model assumes the homoscedasticity. This is seen as a variance in the error terms. To handle this, we've plotted a regplot of the predicted Y values and the residuals (diff b/w y_actual – y_predicted).

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   a. Here are the top 3 features: temp (0.5173), year (0.2326), Winter/ Season

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
The Linear Regression algorithm is used to statistically analyse the past data and predict the target variable. It can predict continuous variables. It is used to show the relationship between a target/dependent variable with one or more independent variables. The idea behind a linear regression is to understand the existing dataset to identify the coefficients for a best fit line ($Y= B_0 + B_1X_1+ B_2X_2+…$).

The goal of the Linear Regression is to identify a best-fit line. Although multiple lines can be drawn for a given dataset, we need to identify the best fit line to minimize the error terms. To achieve this, we use the Mean Squared Error (MSE) cost function, which will provide the average of Squared Error (i.e., diff. b/w the predicted and actual value). The points plotted for a cost function are referred as Residuals; they are ideally the difference between the predicted and actual values for a target variable. The gradient descent is used to minimize the MSE cost function. A regression model updates the coefficient values using this gradient descent method.

To evaluate the Goodness of a Line (aka, best fit line), we use the statistical metric called R-Squared Method. R-Squared (also called as Coefficient of Determination) method determines the goodness of a fit by measuring the strength of the relationship between the dependent and independent variables(between 0-1).

Assumptions of LR:

- The Linear Regression assumes a Linear relationship between the independent variables and dependent variables.
- It shall ensure the independent variables have no/little correlation among themselves. This is referred as multicollinearity. We can identify the multicollinearity by plotting a heatmap comprising of all the independent variables. This shall give us a relationship between the numeric variables. To identify further, we rely on the VIF calculation during model creation. The VIF (Variable Inflation Factor) will give us the degree of correlation of an independent variable with the existing set of other independent variables.
- Homoscedasticity check is to ensure a homogeneous standard deviation in the error values. To verify this, we can plot a regplot against the y values and the residual errors. The line in this plot will pass thru 0.0 and, we observe the data points are equally scattered across both the axes.

Steps to Perform a LR:

1. Reading the data – Read the data from the given dataset.
2. Cleaning the data – Clean the data to ensure no missing values/null values. Also, ensure there are no duplicates.
3. Validate the data – At times, the given information can have data errors. For example, in the given case, the weekday field has errors starting 01-Mar-2019. So, to correct it, we have relied on the date field to extract the week information.
4. Identify the dependent/Target variable and the independent variables.
5. Perform an EDA – This analysis is to understand the relationships among the variables. These relationships show on how the data is related to each other columns including multi-collinearity.
6. Handle Categorical Variables – The categorical variables cannot be used to participate in the model building. So, these values are encoded to numbers. As, the sanctity of the variable shall not be lost, the best option is to create a dummy variable for each of the categorical value in the categorical column. In Total of N-1 dummy variables are created for N unique values of a categorical value.
    i. Once, the dummy variables are created, we no longer require the original column. Hence, this can be dropped.
7. Splitting the Data – The given dataset is to be used for both training and testing activities. Having a separate testing set gives us the first-hand information on how the model is behaving while dealing with the unknown data.

8. Scale the Data – Scaling the training will help us normalize the variables in accordance with the other variables in the dataset. This becomes handy while comparing the coefficients of variables in our model building process. The scaling happens in 2 steps.
    i. Perform fit the training set – This step will try to understand the existing dataset.
    ii. Perform a transform of the training dataset - This step will scale the variables as per the scaler method used.
    iii. The scaler instance is retained and the same is used to scale the test dataset. This is done to ensure that the scaler model, based on the learning of training data set(via the fit method), will now be used to transform the test dataset.
9. Building the model – Split the dependent variable into a separate series and rest all independent variables in another series.
    i. The statsmodel api used for model building assumes the constant value as 0 (ie., the line passes thru X=0). This is not the case; hence, we're adding a new column as constant so as the statsmodel api (sm) can fit assuming the constant is other than 0.
10. Steps of model building:
    i. Add a constant to the X training set. using the sm.add_contant(X_train)
    ii. The Model building uses OLS method (Ordinary Least Squares) to determine the relation between the dependent and independent variables.
    iii. The model is fit using the command lr.fit()
    iv. The fit data results in a summary(); The summary has following important fields:
        1. Coef – The coefficient for each of these columns including the added constant.
        2. The P value – Provides the probability. It shall be less than the significance value Alpha (5% - assumed here)
        3. R-Squared value – Gives the determination of the fit for the model.
        4. Adj. R-Squared value – Gives the value to compare across multiple versions of models in determining what's the best.
    v. Calculate the VIF (Variance Inflation Factor) to determine if the variable in the model has multicollinearity. I.e.., collinearity among the independent variables. Optimal values for this shall be <5.
    vi. Repeat the process by altering the independent variables a best value of R2 and P values and VIF are identified.
11. To identity the variables to be included in the model, we can go for different approaches. The Automated approach uses RFE (Recursive Feature Elimination) to shortlist the variables/features of that can help a best fit for the model.
12. However, a few of these variables can be removed manually based on the model performance and comparison.
13. Finally, when the model is identified to be good, we start applying on the test set.
    i. Scale the test set.
    ii. Split the dataset to X and y series.
    iii. Transform the model by using the same scalar of the training set.
    iv. Predict the y values using the model.
    v. Populate a residuals by y_actual - y_predicted.
    vi. Check if the residuals data is a normal distribution – to affirm our assumptions.
    vii. Check if the residuals and the y-actuals form a well distributed scatter plot to confirm homoscedasticity.
    viii. Calculate the R2 value to confirm the determination of the fit for the model.

2. **Explain the Anscombe's quartet in detail.**
    a. Anscombe's quartet is used to illustrate the importance of EDA and the drawbacks of depending only on the summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers and other crucial details that may not be found in the summary statistics alone.
    b. The Anscombe's quartet consists of 4 sets of datasets, having identical stats (R2, correlations, linear regression lines etc.) but have different linear regression lines(when plotted on a scatter plot).
    c. On Plotting different datasets with the coefficient values, we observe that each of the scatter plot has a different relationship between the line and the X, y points. Some sets are more inclined to the best fit line, some are scattered away etc.

3. **What is Pearson's R?**

The Pearson's correlation is a measure of the linear correlation (strength and direction) between two sets of data. It is the ratio between the covariance of the two variables and the product of their standard deviations. It is essentially a normalized measurement of covariance. The value of Pearson's R will always be between -1 and 1 (1 – an unrealistically perfect correlation).

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$\rho_{x,y}$ → between 0 & 1 -> Positive Correlation; when one variable changes, the other variable also changes in the same direction.

$\rho_{x,y}$ → 0 -> No correlation; There is no relationship between the variables.

$\rho_{x,y}$ → between 0 & -1 -> Negative correlation; When one variable changes, the other variable changes in the opposite direction.

Persons R is used when:

- Both the variables are quantitative.
- Both the variables are normally distributed.
- The data has no outliers.
- The relationship is linear.

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

Scaling/feature scaling is a method to normalize the range of independent variables. Feature scaling helps us bring all the other variables into a same range of data. This is useful during the linear regression model building to compare the coefficients of the variables. As, all of them are in the same range, it would be easy for comparison and evaluating the model. There are 2 types of scaling that can be performed. Normalized Scaling and Standardized Scaling.

Normalized Scaling: Also known as Min-Max Scaling. It will rescale the feature so as the range of the value lies in between 0 & 1. The formula of the normalization is given as:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

Standardized Scaling: Feature standardization makes the values of each feature in the data have zero mean and unit variance. The method of calculation is to ensure the Mean = 0 and Std-Dev = 1.

$$x' = \frac{x - \overline{x}}{\sigma}$$

The Normalized scaling is used when the data is not Normally Distributed. So, that it can confine all the variable data in the range of 0-1 without losing the relative difference among them.
The standardized scaling can be helpful when data is normally distributed. It will try to transform the data in a way where the mean=0 and std.dev=1. This process is not impacted by any outliers in the data.

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

The VIF value is an index that determines the increase of variance of an estimated regression coefficient due to multicollinearity. If all the independent variables are perfectly independent to each other, then the VIF value will be 1. And, if there is a perfect correlation across these independent variables, then the VIF value will be infinite. When all the independent variables the model summarizes the R2 values to be 1. This means that the model is a 100% fit for the given data set.

$$VIF = \frac{1}{(1-R^2)}$$

The VIF is infinite indicates that the value (1-R2) is 0. Which means that the R2 value is 1. This means that the model is 100% fit to the given data set.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
The Q-Q plot is a probability plot, a graphical technique for determining if 2 datasets plotting their quantiles against each other. The plot's X and Y axes have the quantile data plotted from first and second datasets respectively.
If the plot points are similar, the Q-Q plot will approximately lie on the identity line Y=X. The greater the departure from this reference line, the greater the evidence for the conclusion that the two datasets have come from the populations with different distributions.

The Q-Q plot can be used to plot the quantiles of a sample distribution against the quantiles of a theoretical distribution. Doing this helps us determine if the dataset follows any distributions (like, Normal, Uniform, exponential etc.)

Q-Q plots are useful to determine:

- If 2 populations are of the same distribution
- If the residuals follow a normal distribution.
- Skewness of the distribution.