A]. 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?
Answer-1. For categorical variables, box plots and bar graphs were plotted and the following were the major effects on the count of bikes:

1. *Season*- The use of the bikes was lowest in the spring season with the median lying somewhere around 2100 and highest in the fall with a median of 5400 approx followed by the summer season with a median of 5000. The median for used bikes was 4700 in winter.
2. *Year*- A substantial rise in the use of BoomBikes was seen in 2019 compared to 2018. Approximately 2000 new users increased in the year 2019.
3. *Months*- The overall use of BoomBikes is high from July to October.
4. *Holiday* and Workingday- The day being a holiday or working day does not seem to have much impact on the target variable.
5. *Weekday*- Looks like not even the weekdays have much effect on the count of BoomBikes.
6. *Weather*- It is seen that when there is little snowfall, the usage of BoomBikes is very low, and in clear weather, the usage is very high. Also, there are a moderate amount of users riding BoomBikes in the mist climate as well.

Q2. Why is it important to use drop_first=True during dummy variable creation?
Answer-2. After creating dummy variables, we don't need all the dummy variables. We can infer from (m-1) dummy variables, where m is the level of the variable. This will eventually help in contributing to making our model lean and also help in reducing the correlation among the dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer-3. By looking at the numerical pair-plot created, the temperature seems to be the highest correlated numeric variable with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer-4.

1. *Assumption-1*: There should be a linear relationship between the independent and dependent variables. We can do this by looking at the pairplots and seeing the pattern of the scatterplot if they are straight or in a curvilinear pattern.
2. *Assumption-2*: No multicollinearity. We can do this by dropping the independent variables with high multicollinearity by looking at VIF values that we can calculate.
3. *Assumption-3*: Error terms are normally distributed. We can do this by plotting a distribution plot between y_train and y_train_pred.

4. *Assumption-4*: Homoscedasticity. We can do this by plotting a residual plot and seeing if the variance of the error terms is constant across the values of the dependent variable.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer-5. The top three features significantly contribute to the demand for the shared bikes:

1. Temperature(temp) with a coefficient of 0.3822. This means if the temperature rises by one unit, the demand for bikes will rise by 0.3822 units.
2. LightSnow with a coefficient of -0.2488. This means if the LightSnow rises by one unit, the demand for bikes will decrease by 0.2488 units.
3. Year(yr) with a coefficient of 0.2365. This means that every year, the demand for bikes will rise by 0.2365 units.

B]. General Subjective Questions:

Q1. Explain the linear regression algorithm in detail.
Answer-1. Regression is one of the machine learning algorithms in which the output variable that is to be predicted is a numeric variable. In the regression, there is past data that has labels and which is why it is called a supervised learning algorithm. The linear regression technique finds the linear relationship between the dependent variable and the independent variables. This relationship is explained by a line called the 'Best-fit line'. The equation of this straight line is given by [(Y)pred = B0 + B1.X] where (Y)pred=predicted value of output variable, B0=intercept, B1=slope, X=predictors. The best-fit line is the line that fits the plotted scatter plot(between the dependent variable as y and independent variables as x) in the best way possible. This line performs a regression task of predicting values of the target variable based on independent variables and can also forecast the trend of the target variables based on the past data of independent variables with a certain percentage of accuracy.

Q2. Explain the Anscombe's quartet in detail.
Answer-2. Anscombe's quartet includes 4 datasets that have the same statistical properties but each of them follows different patterns when plotted on graphs. A dataset for 4 different branches-(Mumbai, Pune, Delhi, and Kolkata) with say, target variable (Sales) and predictor (Discount) for some product is given to us for 12 months with different values in each of it. If calculated, the average sales and average discount might be the same for all 4 branches and the standard deviation of sales and the standard deviation of the discount will be some value that will be the same for all 4 branches as well. Now, this might look like all 4 branches are performing the same as they have the same average sales and discounts with the same corresponding standard deviations but when we plot a graph with x=Discounts and y=Sales, we will get a huge difference in the pattern followed by each of these 4 graphs. This is called Anscombe's quartet which tells us how one can be easily deceived by just looking at the numeric values or summary of the data that may look similar and why it is important to visualize the data to avoid misinterpreting the data.

Q3. What is Pearson's R?
Answer-3. Pearson's R is a type of correlation that is used to measure the intensity of association between two continuous variables. Also known as bivariate correlation, it measures the linear correlation between two variables on a scale of -1 to 1. Pearson's R can neither capture the non-linear relationship between two variables nor it can differentiate between dependent and independent variables.
There are several assumptions for Pearson's R:

1. Variables should be normally distributed.
2. Scale of measurement should be a ratio.
3. The relation should be linear.
4. The data should not contain outliers.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer-4. Scaling is essentially bringing all the features of the data to the same level of scale. There can be several features with varying values with different units and if we make any analysis on such data, the final result will be a biased and skewed one. Scaling avoids such situations and helps interpret and infer the data which is biased-free. In addition, scaling also helps in processing the data faster for a certain given task. For eg- Faster conversion of gradient descent.
There are 2 main types of scaling methods:

1. Normalization- Also known as MinMax scaling, it compresses the data between 0 and 1 hence, becomes easy to interpret and less work for the system also.
2. Standardization- it subtracts the mean and divides by SD(Standard Deviation) such that it is centered at 0 and has SD=1.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer-5. When the correlation between two independent variables is a perfect correlation, the VIF value that we get is infinity. The VIF is calculated by the formula VIF = ((1)/(1-R2)) and in the case of perfect correlation, the R2=1 which makes the VIF = 1/0 which is infinity. In such cases, it is a must to drop one of these two features having perfect correlation.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer-6. A Q-Q plot aka Quantile-Quantil plot is a plot of two quantiles against each other. It helps us to determine whether a set of data comes from a population with a common distribution or not. A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Various aspects such as the shift in scale, change in the symmetry, and presence of outliers can be detected by a Q-Q plot.