Name: Siddharth Sujir

Subjective Questions:

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]:
   - Analysis on **Season** variable indicates; total bike usage is more in Fall and less in Winter.
   - Analysis on **weathersit** variable indicates that total bike usage is **more** when weather is Clear, few clouds, partly cloudy, partly cloudy, and **less** when there is light precipitation (light snow, light rain, thunderstorm).
   - Analysis on **weekday** variable shows that the usage is low beginning of the week and gradually increases during the week till weekend.

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

   [Ans]:
   During the dummy variable creation, we just need k-1 dummy variables. Keeping k variables makes one of them redundant.

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

   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?*

   In the distplot plotted with residual, the assumption of linear regression holds true.
   - With mean zero
   - Errors are following a normal distribution.
   - Some variables in X (temp) having a linear distribution with Y (cnt) based on pairplot.

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

   1. Temp
   2. Yr
   3. Season

General Subjective Questions

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

   [Ans]:

   Linear regression is used to predict a value of a variable based on another variable in the given dataset. The value used to predict is called the independent variable and value that is predicted is called the dependent variable.

   There are two types of regression.
   - Simple linear regression – One variable can be used to predict the value of the target variable. The equation of the best fit line is represented using y= B0 + B1X where B0 is the intercept and B1 is the coefficient.

   - Multiple linear regression – Multiple variables in the dataset can be used to predict the target variable. The equation of the best fit line can be represented using y=B0+B1x1+B2x2+……+Bnxn.

2. *Explain the Anscombe's quartet in detail.*
   Anscombe's quartet explains the importance of plotting the data before building the model. When we take four datasets that have nearly similar statistical observations like mean, variance etc. When it gets plotted, they look different. It helps us understand the importance of data visualization and how easy it is to fool regression algorithm.

3. *What is Pearson's R?*
   Pearson's R gives a numerical summary of strength of linear association between the variables.
   R=1 means data is perfectly linear with positive slope
   R=-1 means the data is perfectly linear with negative slope
   R=0 means there is no linear association

4. *What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?*
   - Scaling is process of converting the value of a variable to a normalized scale.
   - Scaling is performed for the following reasons:
     - Ease of interpretation.
     - Faster convergence for gradient descent.
   - Normalized scaling uses the min and max values to do the scaling. Standardized scaling used the mean of the variable and standard deviation of the variable to do the scaling.

5. *You might have observed that sometimes the value of VIF is infinite. Why does this happen?*
   VIF infinite, it means that there is perfect correlation between the independent variables. To solve it, we need to remove the variable that is causing the perfect collinearity.

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

   Q-Q plot, also called Quantile-Quantile plot, is a graphical method of determining if two samples of data came from the same population. It's a plot of quantile of first dataset against the quantile of second dataset.
   In linear regression, it helps us determine if two populations have the same distribution. Also, it helps us determine if residuals follow a normal distribution, which is one of the assumptions of linear regression.