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?

- Bookings are more in normal days than holidays
- Season 3 is having booking and then season 4 followed by season 2. Season 1 has very low bookings
- Booking happened during Weathersit 1 (Clear, Few clouds, Partly cloudy, Partly cloudy) with average of 4500
- during 5,6,7 mothns we can see consistent average booking of 4500 per month and in march it reached max booking of 8000 and in october it has no booking

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

- Drop first=True will drop the first dummy variable created for that group

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

- Registered column has the correlation with the target variable because cnt = registered + casual
- Temp and atemp variable are highly correlated with target variable cnt

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

- Since we could fit a linear line with 82% R2-score we could get a linear relationship between dependent and independent variables. Thus, assumption 1 is validated
- From the residual analysis graph, error terms are normally distributed, independent to each other and exhibit constant variance
- VIF should be in between 0-5
- P-value for any of the independent variables should be <0.05
- Correlation should be in the range of -0.7 to 0.7

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

- Temp
- Yr
- weathersit3

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Regression starts with fitting a line for 1 point in the data set. From this line it adds the new points to be fitted into the line while reducing its some of residual error. Thus after fitting all the data points it optimizes its total error by using algorithms like least square error method, gradient descent etc…

$$Y=m1x1+m2x2+m3x3+\ldots+c$$

Linear Regression model assumes a linear relationship between dependant and independent variables.

When dimensions are high, i.e., when we have multiple variables which are similar of a linear relationship in higher dimensions.

2. Explain the Anscombe's quartet in detail.

   This comprises four datasets that have nearly identical statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

   When graphed one dataset represents a nearly fitted regression line second data set represents a non linear curve third datasets represents perfectly fitted regression line with an outlier and the final data set represents a constant valued data point with an outlier

3. What is Pearson's R?

Pearson's R is a measure of linear correlation between two sets of data. It is the covariance of two variables divided by the product of their standard deviations. It is thus a normalized measurement of the covariance, such that the results are always between -1 and 1.

If R=

0; no correlation

0-0.5; week positive correlation

0.5-1; moderately positive correlation

1; highly positively correlated

-1; highly negatively correlated

-0.5 - 0; weekly inversely correlated

-1 – 0.5; moderately negative correlation

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

Scaling is a technique used to bring all the values into same range. We will perform scaling when we have small to very high range of values so that model accuracy will increase.

Differences:

Standard scaling makes the data into standard normal distribution format with mean = 0 and sd =1

Normalized scaling brings the data in between 0 and 1

Dummy variables will get impacted because of normalized scaling where as standard scaling doesn't impact.

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

   When there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

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

   A Q-Q plot represents 2 different quantile distributions of a given data.If the two distributions being compared are identical, the Q–Q plot follows the 45° line $y = x$. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line $y = x$. If the general trend of the Q–Q plot is flatter than the line $y = x$, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis.