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. From the model created, following categorical variables are still retained after feature elimination and have significant impact –

a) The **weather situation** contributes to the sales only when it is case 4 -, i.e., **Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.** The coefficient obtained in this case is -0.2875, by which we can infer that the sales plummet by a factor of 0.2875 when the above conditions are active

b) **Weekdays** do not have a significant impact on the dependent variable, as the coefficients do not have a significant magnitude.

c) For the **months** categorical variable, we can find a significant increase in sales in the month of **September**, with the coefficient being 0.4203

d) Lastly, the **year** 2018 was not good for the bike sales as the coefficient observed is -0.2478

---

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

A. As we know, the categorical variables in linear regression have to be encoded with dummy variables to convert into numerical ones.

This is done by using one – hot encoding, where each state of the categorical variable is represented by a dummy variable with two states, on and off, represented by 0 and 1.

For example, the season variable in the assignment had 4 states – spring, summer, winter and fall.

The encoded dummy variables for this would be as follows:

	Spring	Summer	Winter	Fall
Spring	1	0	0	0
Summer	0	1	0	0
Winter	0	0	1	0
Fall	0	0	0	1

However, one observation can be made from the above table – if we treat fall as one state and the other 3 seasons as the second – it converts into a binary variable with only 2 states –

	Fall	Not Fall
Fall	1	0
Not Fall	0	1

Which means that the 4th variable is actually redundant, and the above information can be represented using only 3 variables – Spring, Summer and Winter, and when all of them are **off** or 0, it means the season is fall.

This means that any categorical variable with n states can be represented by (n-1) dummy variables.
This eliminates one feature from the final model for each categorical variable, which is very important as a higher number of redundant features will cause the final adjusted r squared for the model to reduce, as the model will be penalized one extra time for the nth state column we include.

---

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

A. Looking at the pair plots of all the variables with **cnt**, we can conclude that **temp** has the highest positive correlation with the target variable, as the relationship among them is almost linear.

---

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

A. The following methods were used to validate the assumptions:

   a) Seaborn pair plots to validate linear relationship of the variables with the target variable.

b) **Heatmap** using Pearson's coefficient and **Variance Inflation Factor** test after each iteration to determine collinearity of variables. Variables with high collinearity were detected and eliminated at each step.

c) Plotting the error terms against the mean of the dataset for homoscedasticity test. As the errors were equally randomly distributed across the mean, no visible relationship was found.

d) Histogram plot of the error variables in the train set to check for normal distribution. The error values were found to be almost normally distributed and centred around 0.

---

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

A. a) **Weather situation**, which has a significant negative impact on the final sales depending on the conditions.

b) **Windspeed,** which also has a negative impact on the sales, increased wind speed leads to reduced sales.

c) **Month** categorical variable, a significant increase is observed in the second half of the year, peaking in **September** with a positive effect on the bike sales.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A.