1.From Your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- ***Fall season has more bike shared compared to other seasons***
- ***With respect to months Bike sharing shows an increasing trend in the first half and decreasing towards the end of the year due to approaching Holiday season***
- ***Clear weather has attracted more sharing of bikes.***
- ***Bike sharing increases from mid of week till Saturday.***
- ***Sunday sees the lowest number of Bike shares as it is holiday***
- ***Bike share is less in holidays***
- ***2019 has seen a boom increase in bike share indicating the Boombikes business growth***

---

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

Drop_first = True will remove extra categorical column created during dummy variable creation.

If there are n category levels, then n-1 is the columns are only required for representation.

To add more details here we have 4 category levels under the Weathersit column – Clear, Mist, Light Rain/Snow and Heavy Rain/Snow. In this case when the value for the Weathersit is represented as 0 or 1 for each of the category level dummy variables and the value of Clear/Light Rain/Mist is 0 then this indicates that the Heavy Rain is the category for the record which is obvious.

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

Temp variable has highest correlation of 0.65

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

- ***Error terms by comparing Testing and Trained model was normally distributed, its variance value is constant and independent***
- ***Feature selection***
- ***Multicollinearity should be insignificant***
- ***Getting rid of overfitting by verifying r2 for both test and train with minimal variation***

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

- ***Working day***
- ***Temperature***
- ***Windspeed***

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables.

Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

Y = mX + b

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0,Y would be equal to b.

the linear relationship can be positive or negative

---

2. Explain the Anscombe's quartet in detail. (3 marks)
   Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
   Analyze mean, variance, correlation, linear regression, and other metrics for each dataset within Anscombe's Quartet to showcase identical summary statistics.

---

3. What is Pearson's R? (3 marks)

Coefficients of correlation are generally used in statistics to measure a relationship between two variables. The correlation generally shows a specific value of the degree of a linear relationship between two variables, say X and Y. There are many types of correlation coefficients that are used in statistics. However, Karl Pearson's correlation (also known as Pearson's R) is the correlation coefficient that is most frequently used in linear regression.

- ***Positive Correlation (0 to +1)***
- ***Negative Correlation (0 to -1)***
- ***Zero Correlation (0)***

---

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. sklearn. preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn. preprocessing. Scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

---

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   VIF becomes infinite if there is perfect multicollinearity among the predictors. This occurs when one predictor variable is an exact linear function of another or a combination of other predictor variables. In mathematical terms, if the matrix of predictors has a determinant of zero, it indicates that the predictors are perfectly collinear, leading to an infinite VIF for the involved variables.

---

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Q-Q plot compares the quantiles of a dataset's distribution to the quantiles of a theoretical distribution (usually normal). It plots the observed quantiles against the expected quantiles from the theoretical distribution.
   By plotting the residuals of the regression model on a Q-Q plot, you can visually assess if these residuals approximately follow a normal distribution. If the residuals lie close to a straight line on the Q-Q plot, it indicates that they are approximately normally distributed.