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

A) Here are some of the inferences:
   a. Fall season seems to have attracted more booking.Booking count has increased drastically from 2018 to 2019.
   b. Most of the bookings has been done from may to oct.Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
   c. Clear weather attracted more booking which makes sense. And in comparison to previous year, i.e 2018, booking increased for each weather situation in 2019.
   d. Thu, Fri, Sat have more number of bookings as compared to the start of the week.

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

   A) Drop_first=True helps in reducing the extra column created during dummy variable creation
   B) It helps in reducing further collinations

3. Looking at the pair-plot among the numerical variable, which one has the highest correlation with the target variable?
   a. Variable 'atemp' has the highest correlation with 'cnt' followed by 'tmp'

4. How do you validate the assumptions of linear regression after building the model on training dataset.
   a. Assumptions of linear regression can be validated by plotting a displot of the residuals and checking if there is a normal distribution

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of shared bikes?
   a. Following are the top 3 features:
      i. Temp, weathersit and year.

General subjective questions

1) Explain the linear regression algorithm in detail

   Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables (predictors). It assumes a linear relationship between the independent variables and the dependent variable, which can be represented by a straight line equation. Linear regression is widely used for prediction and forecasting tasks in various fields such as economics, finance, and social sciences.

If only one independent variable, then it is a simple linear regression and if there are multiple independent variables, then it is a multiple linear regression

Linear regression is about finding the best-fitting line through your data points, so you can understand and predict how one thing changes as another thing changes.

2) Explain the Anscombe's quartet in detail?

Anscombe's quartet is a famous example in statistics that illustrates the importance of graphically visualizing data and the potential pitfalls of relying solely on summary statistics. It consists of four datasets that have nearly identical statistical properties but vastly different graphical representations.

3) What is Pearson's R?

Pearson's r, also known as the Pearson correlation coefficient or Pearson product-moment correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the relationship between two variables, indicating how much one variable changes when the other variable changes.

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

   a. Scaling is a preprocessing technique used in data analysis and machine learning to standardize or normalize the range of features or variables in a dataset. The main purpose of scaling is to bring all features into a similar scale or range to ensure fair comparison and to prevent features with larger magnitudes from dominating those with smaller magnitudes during model training or analysis.
   b. Normalized scaling squeezes values into a specific range, while standardized scaling centers the values around 0 and spreads them out evenly.

5) You have observed that sometimes VIF value is infinite. Why does this happen?
   a. The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, which can cause issues with the estimation of coefficients and lead to unreliable interpretations of the model.
      i. $\mathrm{VIF}(X_i)=1-R_{Xi}21$

b. A VIF value of infinity occurs when $\boxed{R_{X_i2}}$ is equal to 1, which implies that the predictor variable $\boxed{X_i}$ can be perfectly predicted by a linear combination of the other predictor variables in the model. In other words, $\boxed{X_i}$ is completely redundant given the other predictor variables, leading to infinite VIF.

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

   a. A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a given sample of data follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the sample data to the quantiles of a specified theoretical distribution, typically the normal distribution.

   b. Below are the uses of Q-Q plot

      i. Checking assumptions

      ii. Spotting patterns

      iii. Fixing issues

      iv. Ensuring reliable results