1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
Ans.
We have the below variables as categorical variables,
'season', 'yr', 'mnth', 'holiday','weekday', 'workingday', 'weathersit'
We can Infer below for the above mentioned variables

Season

From the boxplot & barplot, Fall season has highest cnt value, and Spring has least cnt value

Yr

From the boxplot & barplot, 2019 has more cnt than 2018

Mnth

From the boxplot & barplot, September has highest cnt values, and January has least cnt values.

Holiday

From the boxplot & barplot, cnt values are lower in holidays

Weekday

From the boxplot & barplot, The cnt value increses from Sun to Sat, with high values on Saturday, Friday, and low values on Monday

Workingday

From the boxplot & barplot, there's very less difference between a working day and not a working day, however we can infer that working day has more cnt than the other

Weathersit

From the boxplot & barplot, There're no rentals on Heavy Snowfall day, and also less rentals on light snowfall day. Clear weather has high rentals followed misty weather.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans.
Without drop_first condition, we'll be creating n dummy variables for a categorical variable with n unique values. This might lead to multi-collinearity as they're highly correlated, and might also lead to unstable estimation, inflated variance, and misleading importance of variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation
with the target variable? (1 mark)
Ans.
Temp and atemp has highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans.
We know that the assumption is that the error terms or residuals are normally distributed and centered around 0, that is with mean 0. We did test that for our model for both train and test sets, and found that the residuals or error terms are centred around 0, and also normally distributed.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans.
Based on the final model, we have the top 3 features as temp, yr and weathersit_Light_snowrain(weathersit)

1. Explain the linear regression algorithm in detail. (4 marks)
Ans.
Linear Regression is a supervised Machine Learning algorithm used for regression tasks. We use this to predict the continuous numeric variable, which we call as target variable, based on input or independent features(predictors) which are already present. Linear Regression aims to find a relationship between the predictors and the target variable by fitting it into a linear equation. We'll come up with the best fit line equation, which has the predictors with coefficient and a constant too, that helps us to achieve at the target or outcome, which has very less error(difference between actual and predicted)

We achieve this by doing multiple iterations, by adding/removing the features as needed into the linear model as we build it. Finally the model which has the best significance, based on R2, adjusted R2, and p values we'll select the final best model.
We have two types of Linear Regression models, namely Simple Linear Regression and Multi Linear Regression.
Simple Linear Regression is the one which has only one Predictors, and Multi Linear Regression is the one which has more than one Predictors.
We have below assumptions for the LR,
The relationship between the predictor and the target variable is linear.
Their observation are independent of each other. Their Residuals are normally distributed with mean 0.

Simple Linear Regression is represented as below
$y = \beta_0 + \beta_1 * x + \epsilon$

Multi Linear Regression is represented as below
$y = \beta_0 + \beta_1 * x1 + \beta_2 * x2 + \beta_3 * x3 + \ldots + \beta_n * xn + \epsilon$

y is the target variable
X, x1,x2,…,xn are independent variables
β_0 is the intercept(constant)

*β1, β2, .., βn a*re the coefficients of the predictors
ϵ is the error terms or residuals

Our aim is to find the model which has the least residuals sum of squares, and we use techniques like Gradient Descent, Ordinary Least Squares (OLS) to achieve the minimized cost function model and best fitting line.

2. Explain the Anscombe's quartet in detail. (3 marks)
Ans.
Anscombe's quartet is developed by statistician Francis Anscombe in 1973. It emphasizes the importance of visualizing the data along with looking into the statistical features. It consists of four datasets which identical statistical features like mean, variance, etc. but when we visualize the data, the distribution of the data is not same for the four data sets.
Hence, we can say that we shouldn't always depend on the summary statis tics alone for the inferring or analyzing the data, we should also look/visualize the data. More balanced approach is to do both the summary statistics and visualization of data.

3. What is Pearson's R? (3 marks)
Ans.
Pearson's R is a statistical measure of the variable on the strength of their linear relationship. It ranges form -1 to 1.

- **Positive** *r*: Indicates a positive linear relationship between variables.
- **Negative** r: Indicates a negative linear relationship.
- **Magnitude**: Closer to 1 implies a stronger linear association, while values closer to 0 suggest a weaker correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Ans.
Scaling is a method we used to make sure all the numeric variables have same range of data. Because if we don't do this during our data prep before we build the model, our ML algorithm might weigh some variables higher and some variables lower, because of the range of the values associated to the variables. So it's better if we do Feature scaling before we build our LM.
Differences between Normalized scaling and Standardized scaling

1. Normalized scaling scales the feature/variable between 0 and 1, while Standardized scales the features with a standard deviation of 1 and centered around 0
2. Normalization preserves the original distribution while the Standardization modifies the distribution to have a mean of 0 and standard deviation of 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans.
High values in VIF indicates that the predictor variable are linearly related to each other, which causes multi-collinearity. So if it's infinite that indicates the the two predictor variables are strongly related to each other, and it's a case of strong multi-collinearity.
So we should be removing those variable which causes multi-collinearity to come up with a model which has predictor variables that has low vif value(<5v or <10), to build a LM.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)
Ans.
Q-Q plot stands for Quantile – Quantile plot, is a graphical technique used to compare two datasets if they follow any distribution like normal distribution.
We know that the quantiles are the values that divide the data, like first quartile is 25%, second or median is 50%, and third is 75%. We'll plot the quantiles of the both datsets, one aginst the each other and if their distribution they'll fit into a straight line, if not they'll not fit into a straight line.
If they fit into a straight line, they are normally distributed , if not they're not

Importance of Q-Q:

1. We use this to verify our assumption of Linear regression, that the residuals or error terms are normally distributed.
2. If they don't fit into a line, we use this to identify the outliers too.
3. We use this to evaluate the Linear Regression models fit.