1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

1. Fall season has more bookings. All seasons, the booking is increased from 2018 to 2019.
2. Comparatively better bookings on May, June, July, August, September and October. I could see increasing tren from start and mid of the year and decreasing trend mid to end of the year
3. More bookings on clear weather
4. Towards Mid of the week to end of the week (Thursday to Sunday), bookings are increasing
5. Booking seems same both for non-working and working days
6. Bookings are increased overall in 2019

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

drop_first=True is an important parameter to consider. It serves to avoid the issue of multicollinearity, which is a common problem in statistical models.

By setting drop_first=True, one level of each categorical variable is dropped when creating the dummy variables. This means that one category becomes the reference or baseline category, and the remaining categories are represented by the dummy variables. Dropping the first category helps to avoid the perfect multicollinearity issue.

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

We can clearly understand that TEMP and ATEMP are having high correlation, there is a linear relationship between TEMP and ATEMP.

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

After building the model with Training Data set, we can build linear regression model with below code

```
lr = sm.OLS(y_train,X_train_lm_6).fit()
print(lr.summary())
```

1. Residual Plot: Plotting the residuals against the predicted values can help identify patterns or deviations from linearity. Ideally, the residuals should be randomly scattered around zero without any discernible patterns.

2. Linearity: Evaluating the linearity assumption involves examining the relationship between the independent variables and the dependent variable.

3. Multicollinearity: Multicollinearity occurs when independent variables are highly correlated with each other

    a. This is done using VIF

using F-Statistics value of 230.0 (which is greater than 1) and the R2 values i.e almost equals to zero, states that the overall model is significant

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

Below are the Top3 features contributing significantly towards explaining the demand of the shared bikes

I. temp

II. windspeed

III. season_summer

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression models can be classified into two types depending upon the number of independent variables:

b. Simple linear regression: When the number of independent variables is 1
c. Multiple linear regression: When the number of independent variables is more than 1

The equation of the **best fit regression line $Y = \beta_0 + \beta_1 X$** can be found by minimising the cost function

It assumes a linear relationship between the input variables and the target variable. The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted and actual values of the target variable.
The algorithm works by estimating the coefficients of the linear equation that represents the line. The equation takes the form:
y = b0 + b1x1 + b2x2 + ... + bn*xn
Where:
y is the target variable
x1, x2, ..., xn are the input variables
b0, b1, b2, ..., bn are the coefficients to be estimated
The coefficient b0 represents the y-intercept, which is the value of y when all input variables are zero. The coefficients b1, b2, ..., bn represent the slopes of the line for each input variable, indicating how the target variable changes with respect to each input variable.
The linear regression algorithm estimates the coefficients using a method called ordinary least squares (OLS). OLS aims to minimize the sum of the squared differences between the predicted and actual values of the target variable. This involves finding the values of b0, b1, b2, ..., bn that minimize the following cost function:
Cost = Σ(y_pred - y_actual)^2
Once the coefficients are estimated, the model can be used to make predictions by plugging in the values of the input variables into the equation. The resulting predicted value represents the expected value of the target variable.

Linear regression has several assumptions:

Linearity: The relationship between the input variables and the target variable is linear.
Independence: The observations are independent of each other.
Homoscedasticity: The variance of the errors is constant across all levels of the input variables.
Normality: The errors are normally distributed with a mean of zero.
No multicollinearity: The input variables are not highly correlated with each other.
If these assumptions are violated, the accuracy and interpretability of the model may be compromised, and other regression techniques or data transformations may be more appropriate.
In summary, linear regression is a simple yet powerful algorithm for predicting a continuous target variable based on one or more input variables. It estimates the coefficients of a linear equation that represents the relationship between the variables, using the OLS method to minimize the squared differences between predicted and actual values.

The strength of the linear regression model can be assessed using 2 metrics: 1. R^2 or Coefficient of Determination 2. Residual Standard Error (RSE)

$R^2 = 1 - (RSS / TSS)$

In Linear regression, at each X, finds the best estimate for Y ● At each X, there is a distribution on the values of Y Model predicts a single value, therefore there is a distribution of error terms at each of these values

1. t statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. F statistic: Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be
3. R-squared: After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values **independent** variables in X

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, but exhibit vastly different graphical representations. It demonstrates the importance of visualizing data and not relying solely on summary statistics.

The quartet consists of four subsets of data, each containing 11 (x, y) pairs:

Dataset I: In this dataset, the relationship between x and y is approximately linear, with a slight positive slope. The scatter plot of the data points resembles a simple linear relationship.

Dataset II: This dataset is also linear but with an outlier point that significantly influences the regression line and correlation coefficient. The outlier creates a non-representative summary statistic for the relationship.

Dataset III: In this dataset, the relationship between x and y is non-linear, specifically quadratic. However, the summary statistics, such as the correlation coefficient, are the same as in Dataset I, leading to an incorrect interpretation if only summary statistics are considered.

Dataset IV: This dataset contains an apparent relationship between x and y when divided into two distinct groups. However, each group has the same summary statistics as the other datasets, highlighting the risk of making assumptions without visual inspection.

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

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges between -1 and 1.

Pearson's R is calculated using the following formula:

r = (Σ((x - mean(x))(y - mean(y)))) / (sqrt(Σ(x - mean(x))^2) * sqrt(Σ(y - mean(y))^2))

Where:

x and y are the two variables being analyzed.

mean(x) and mean(y) represent the means of x and y, respectively.

Σ denotes the sum over all the data points.

The value of Pearson's R indicates the strength and direction of the relationship:

A positive value (between 0 and 1) indicates a positive linear relationship, meaning that as one variable increases, the other tends to increase as well. The closer the value is to 1, the stronger the positive relationship.

A negative value (between -1 and 0) indicates a negative linear relationship, meaning that as one variable increases, the other tends to decrease. The closer the value is to -1, the stronger the negative relationship.

A value of 0 indicates no linear relationship between the variables. However, it is important to note that non-linear **relationships** may still exist.

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

Scaling refers to the process of transforming variables to a common scale to facilitate better comparison and analysis. It involves adjusting the values of variables so that they fall within a specific range or distribution.

Scaling is performed for several reasons:

Comparison: Scaling allows for a fair and meaningful comparison between variables. When variables have different units or magnitudes, it becomes challenging to directly compare their values. Scaling brings them to a common scale, enabling more accurate comparisons.

Model Performance: Scaling can improve the performance of certain machine learning algorithms. Many algorithms, such as those based on distance or gradient descent, are sensitive to the scale of variables. Unequal scales can lead to biased models or dominance by variables with larger values. Scaling helps mitigate these issues and ensures that all variables contribute equally to the model's performance.

Normalized scaling and standardized scaling are two common scaling techniques:

Normalized Scaling (or Min-Max Scaling): In normalized scaling, variables are transformed to a specific range, typically between 0 and 1. The formula for normalized scaling is: scaled_value = (value - min) / (max - min)

Standardized Scaling (or Z-score Scaling): In standardized scaling, variables are transformed to have a mean of 0 and a standard deviation of 1.

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

The VIF (Variance Inflation Factor) is a measure used to assess multicollinearity in regression models. It quantifies how much the variance of the estimated regression coefficients is inflated due to collinearity among the predictor variables. A VIF value of infinity typically occurs when perfect multicollinearity exists in the model.

Perfect multicollinearity means that one or more predictor variables in the model can be perfectly predicted by a linear combination of other predictor variables. In other words, there is an exact linear relationship among the variables, making it impossible to estimate the coefficients accurately.

To handle perfect multicollinearity, it is necessary to identify and remove one or more of the collinear variables from the model. This can be achieved by carefully examining the variables, their relationships, and the context of the problem being analyzed. It is important to understand the underlying causes of multicollinearity and consider the implications for the interpretation and validity of the regression results.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular distribution. It compares the observed quantiles of a variable against the quantiles of a theoretical distribution, typically the standard normal distribution (mean = 0, standard deviation = 1).

In linear regression, Q-Q plots are important for several reasons:

Normality Assumption: One of the key assumptions of linear regression is that the errors (residuals) follow a normal distribution. By examining the Q-Q plot of the residuals, you can assess whether this assumption is met. If the residuals fall along a straight line in the Q-Q plot, it suggests that the errors are normally distributed, which is important for obtaining reliable parameter estimates and valid hypothesis tests.

Outlier Detection: Q-Q plots can help identify outliers or heavy-tailed distributions. If the points in the Q-Q plot deviate significantly from the straight line, it indicates potential departures from normality.