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?

- Fall season has attracted more booking
- Period of may,june,july,august and September had higher number of booking compared to other months.
- It seems like when there on non holiday days more bookings are done
- It seems like the end of the weekend like thu,fri and sat had more number of bookings than the start of the weekend
- We also found that the number of bookings are quite high when weather is clear
- 2019 had quite higher number of booking than 2018

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

drop_first =True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

temp variable has the highest correlation with the target variable.

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

Normality of error terms
Multicollinearity check
Linear relationship validation

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

temp
spring
December

1) Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical and machine learning technique used for modeling the relationship between a dependent variable (response) and one or more independent variables (features or predictors). It aims to find the best-fitting straight line (or hyperplane in higher dimensions) that represents the linear relationship between the variables.
Mathematically, the simple linear regression model for a single feature (x) and a single target variable (y) can be represented as:

y = mx+c.
y is the target variable (dependent variable) that we want to predict.
x is the independent variable (feature)
c is constant.
There are two types of Linear Regression:-

1. **Simple Linear Regression**: When there may be simplest one unbiased variable (x) used to are expecting the goal variable (y). The relationship among x and y is modeled as a directly line.

2. **Multiple Linear Regression**: When there are more than one independent variables (x1, x2, ..., xn) used to are expecting the goal variable (y). The relationship among y and the a couple of predictors is modeled as a hyperplane in n-dimensional space.

2) Explain the Anscombe's quartet in detail

Anscombe's Quartet can be defined as a set of four information units that are nearly equal in simple descriptive facts, however there are a few peculiarities inside the dataset that fools the regression version if constructed. They have very one of a kind distributions and seem otherwise whilst plotted on scatter plots.

This tells us about the importance of visualising the information before applying numerous algorithms obtainable to build models out of them which shows that the statistics features need to be plotted with a view to see the distribution of the samples that let you perceive the various anomalies gift within the records like

outliers, variety of the facts, linear separability of the facts, and so forth. Also, the Linear Regression may be best be considered a in shape for the records with linear relationships and is incapable of managing some other sort of datasets.

3) What is Pearson's R?

Pearson's correlation coefficient, frequently called Pearson's R is a statistical measure that quantifies the energy and direction of the linear dating among two non-stop variables. It changed into added by means of Karl Pearson in 1896 and is widely used in data and information evaluation to evaluate the degree of association between two variables.

Pearson's R is a cost between -1 and 1:

A high quality value (in the direction of 1) indicates a nice correlation, meaning that as one variable will increase, the other tends to growth as well.
A negative fee (closer to -1) shows a terrible correlation, meaning that as one variable will increase, the opposite tends to decrease.
A value of 0 suggests no linear correlation, that means there's no regular linear courting among the two variables.
Mathematically, Pearson's correlation coefficient (r) is calculated the usage of the following formula.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

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

Scaling is a preprocessing step in data guidance where the functions (independent variables) are transformed to a fashionable scale to make certain that every one capabilities have comparable tiers or magnitudes. It involves adjusting the values of the features so they fall inside a particular range or have 0 imply and unit variance.

Scaling is completed for numerous motives:

Feature Magnitude: Features with distinct scales may have a sizeable impact on the overall performance of certain gadget getting to know algorithms. For example, algorithms based on distance calculations (e.g., ok-nearest associates, help vector machines) may be touchy to the scale of the functions.

Convergence: Gradient-primarily based optimization algorithms, like the ones utilized in neural networks or linear regression, generally tend to converge quicker and greater reliably whilst features are scaled.

Interpretability: Scaling could make the model extra interpretable through placing all functions on the same scale, making it easier to compare the coefficients of the functions.

Regularization: Some regularization strategies, like L1 and L2 regularization, count on that functions are at the same scale, and scaling enables the regularization terms work successfully.

The common scaling techniques are normalized scaling and standardized scaling

Normalized Scaling (Min-Max Scaling):

Also referred to as Min-Max Scaling, it transforms the functions to a particular variety, usually between 0 and 1.
The formula to normalize a characteristic

$$x_{\text{normalized}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

The minimal cost of the feature is mapped to 0, and the most value is mapped to one, and all other values are linearly transformed hence.
Standardized Scaling (Z-Score Scaling or Standardization):

Standardization transforms the features to have 0 mean and unit variance.

The formula to standardize a function

$$x_{\text{standardized}} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

Here, the imply of the characteristic is subtracted from every data factor, and the end result is split by way of the usual deviation of the characteristic.
The key difference among the 2 scaling methods is the range of values they produce. Normalized scaling bounds the information within a specific variety (e.g., 0 to 1), while standardized scaling facilities the statistics round zero with a variance of 1

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

The incidence of an countless cost for the Variance Inflation Factor (VIF) is associated with the issue of perfect multicollinearity in a couple of linear regression version. Perfect multicollinearity happens when one or greater impartial variables can be exactly expected via a linear aggregate of other independent variables inside the version.

VIF is a measure used to come across multicollinearity many of the functions in a more than one linear regression model. Specifically, it quantifies how a lot the variance of an expected regression coefficient increases because of collinearity between a selected impartial variable and the other unbiased variables inside the version.

The formula to calculate the VIF for a given independent variable is:

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

Where Ri^2 is the R-squared value obtained by regressing the i-th independent variable against all the other independent variables in the model.

When perfect multicollinearity exists for a particular independent variable, the R-squared value Ri^2 becomes equal to 1. In that case, the numerator in the VIF formula becomes zero:

1-Ri2=1-1=0

Hence VIF becomes infinte
VIF=1/0=infinite

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

A Q-Q plot (Quantile-Quantile plot) is a graphical device used to evaluate whether a dataset follows a particular theoretical distribution, such as the everyday distribution. It helps to visually examine the quantiles of the sample records towards the quantiles of the theoretical distribution. If the records comes from the desired theoretical distribution, the points on the Q-Q plot will about fall on a directly line. Deviations from the straight line suggest departures from the assumed distribution.
Importance of Q-Q plot in linear regression:

Normality Check: The Q-Q plot is a visible tool to quickly test whether or not the residuals of the linear regression model observe a normal distribution. This is crucial due to the fact if the residuals are not generally disbursed, the statistical inferences made the usage of the model, which includes confidence intervals and hypothesis assessments, may not be legitimate.

Model Diagnostics: By analyzing the Q-Q plot, you could become aware of potential departures from normality within the residuals. If the points on the plot deviate from a instantly line, it shows that the residuals are not generally dispensed, which may require similarly research or model refinement.

Assumption Validation: The normality of residuals is one of the key assumptions of linear regression. Validating this assumption thru the Q-Q plot helps make certain the reliability and accuracy of the model's predictions and inferences.