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?
    - In the Fall season bookings are more compared to other seasons and in each season the booking count shows increasing trends from 2018 to 2019.
    - Most of the bookings have been done during the 2nd & 3rd quarter of the year. The Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of the year. The booking count shows increasing trends from 2018 to 2019 for each month.
    - Clear weather attracts more bookings & booking increases from 2018 to 2019 for each weather.
    - Thu, Fir, Sat have a greater number of bookings as compared to the start of the week.
    - On the Holiday booking is less compared to non-holiday days.
    - Booking seemed to be almost equal either on working days or non-working days.
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
    - It is important for the order to achieve k-1 dummy variables as it can be used to delete extra columns while creating dummy variables.
    - For Example: We have three variables: Furnished, Semi-furnished & un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So, we can remove it
    - It is also used to reduce the collinearity between dummy variables.

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

    atemp and temp both have the same correlation with the target variable of 0.63 which is the highest among all numerical variables.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - Linearity of the relationship between independent variables & the dependent variables, we can validate it using residual plots.
    - Independence of errors, the residuals should be independent of each other (i.e. no correlation between residuals).
    - Homoscedasticity – The variance of residuals should remain constant across all levels of the independent variables.
    - Normality of Errors – The residuals should be normally distributed.
    - No Multicollinearity – The independent variables should not be highly correlated with each other. The VIF values greater than 5 or 10 indicate the presence of multicollinearity should be dropped.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    - temp
    - Light Snow Rain
    - sept

General Subjective Questions

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

Linear regression is one of the simplest and most widely used algorithms in statistics and machine learning for predicting a continuous dependent variable (target) based on one or more independent variables (predictors). The goal of linear regression is to find the best-fitting line that describes the relationship between the independent variables and the dependent variable.

There are two types of linear regression:

Simple Linear Regression: Involves one independent variable.

Multiple Linear Regression: Involves more than one independent variable.

The General Equation of Linear Regression

The model for linear regression can be described by the equation:

y = beta_0 + beta_1* x1 + beta_2 *x2 + ….. + beta_n* xn + epsilon

Where: y is the dependent variable (response variable).

Beta_0 is the intercept (the value of when all ’s are zero).

Beta_1,…Beta_n are the coefficients (or slopes) that measure how much the changes for a unit change in the corresponding -variable.

X1…..Xn are the independent variables (predictors).

Epsilon is the error term (residuals), representing the difference between the actual and predicted -values.

In simple linear regression (with only one predictor variable), the equation becomes:

y = Beta_0+ Beta_1 * x + epsilon

Steps of the Linear Regression Algorithm:

1. **Define the Hypothesis (Model)**The first step in linear regression is to define the hypothesis function, which in linear regression is a linear equation. This function is used to predict the value of the dependent variable based on the independent variable(s).

For a single variable:

Y(predicted) = beta_0 + beta_1* x

Where: y (predicted) is the predicted value of y

Beta_0 is the intercept, and Beta_1 is the slope or coefficient for the independent variable x.

2**. Estimate the Coefficients**

The next step is to find the values of the parameters Beta_0 and Beta_1 (or more generally, Beta_n) that best fit the data. The most commonly used method for estimating these coefficients is Ordinary Least Squares (OLS).

Ordinary Least Squares (OLS) is a technique that minimizes the sum of the squared differences between the actual and predicted values of y.

OLS minimizes this sum of squared errors by finding the optimal values for Beta_0 and Beta_1

3. Make Predictions

Once the parameters are estimated, the linear equation can be used to make predictions on new data points.

For a new data point x(new), the predicted value of y is:

Y(new) = beta_0 + beta_1* x(new)

In multiple linear regression, the equation extends to include more predictor variables:

Y(predicted) = beta_0 + beta_1* x1 + beta_2 *x2 + … + beta_n* xn

4. Evaluate the Model

Once the model is built, it is important to evaluate its performance using various metrics. Common evaluation metrics for linear regression are:

Mean Squared Error (MSE): Measures the average of the squared differences between actual and predicted values.

R-squared : Represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

Adjusted R Sqaured: Adjusts the R Sqaured for the number of predictors in the model.

Assumptions of Linear Regression

Linear regression makes several key assumptions:

1. Linearity: The relationship between the independent variables and the dependent variable is linear.

2. Independence: Observations are independent of each other.

3. Homoscedasticity: The variance of residuals is constant across all levels of the independent variables.

4. Normality of Residuals: The residuals (errors) should be normally distributed.

5. No Multicollinearity: The independent variables should not be highly correlated with each other.

If these assumptions are violated, the performance of the model can be affected, and remedies like transforming variables, adding interaction terms, or using regularization methods might be necessary.

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

Anscombe's Quartet is a set of four distinct datasets that were constructed by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before performing statistical analysis. The quartet consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression line), but when visualized, they reveal very different distributions and relationships between the variables.

The purpose of Anscombe's Quartet is to show that relying solely on summary statistics without visualizing the data can lead to misleading conclusions. Each dataset has the following characteristics:

1. Similar Summary Statistics:
   Mean of the x values.
   Mean of the y values.
   Variance of the x values.
   Variance of the y values.
   Correlation coefficient between x and y.
   Equation of the linear regression line (slope and intercept).
   R-squared (R^2) value of the regression.

2. Different Visual Patterns: Despite having the same summary statistics, each dataset exhibits very different relationships between x and y, which becomes evident only through graphical representation.

Why Is Anscombe's Quartet Important?

Anscombe's Quartet highlights the following key lessons:

1. The Limitations of Summary Statistics: Descriptive statistics like mean, variance, correlation, and regression coefficients can be identical across datasets but fail to capture the true structure of the data. This shows that blindly relying on these numbers without examining the data visually can lead to incorrect conclusions.

2. The Importance of Visualization: Graphical representations such as scatter plots provide critical insights that statistics alone might not reveal. Visualizing data can expose patterns, trends, non-linear relationships, outliers, and other anomalies that might go unnoticed otherwise.

3. Impact of Outliers and Leverage Points: Outliers and influential points can have a disproportionately large effect on regression models and other analyses. It's important to detect and understand such points to ensure accurate modeling.

Visualization of Anscombe's Quartet

When plotted, each of the datasets reveals drastically different patterns:

1. Dataset 1: A clean linear relationship.
2. Dataset 2: A curved, non-linear relationship.
3. Dataset 3: A linear relationship with a significant outlier.
4. Dataset 4: Most points are the same, with one extreme outlier driving the regression.

This is a powerful reminder that "seeing is believing" in data analysis. Visualizations allow you to detect relationships, patterns, and anomalies that raw numbers cannot convey.

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

Pearson's r, or Pearson correlation coefficient, is a statistical measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to +1:

- +1 indicates a perfect positive linear relationship (as one variable increases, the other also increases).
- -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases).
- 0 means no linear relationship.

Mathematically, it is calculated by dividing the covariance of the two variables by the product of their standard deviations.

Here's an example of how Pearson's r can be calculated:

Let's say you want to measure the correlation between hours studied and exam scores for 5 students.

Data:

Student	Hours Studied(X)	Exam Score(Y)
1	2	65
2	4	70
3	6	75
4	8	85
5	10	90

Steps to calculate Pearson's r:

1. Find the mean of X and Y:
   Mean of X (hours studied): X(mean) = 2+4+6+8+10/5 = 6

Mean of Y (exam scores): Y(mean) = 65+70+75+85+90/5 = 77

2. Calculate the covariance between X and Y = 26
3. Calculate the standard deviations of X and Y, SD of X is 2.8 and SD of Y is 9.27
4. Use the Pearson formula: r = cov(X,Y)/SD of x* SD of y

= 26/2.8*9.27 = 1

The final value of r will indicate the correlation strength. In this case, because hours studied and exam scores are likely to be positively correlated, r to be close to 1.

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

What is Scaling?

Scaling is a data preprocessing technique used in machine learning to adjust the range of the feature values so that they fall within a specific range. It's important because machine learning algorithms often work better or converge faster when features are on a similar scale.

Why is Scaling Performed?

- Improves model performance: Some models, like gradient-based models), can perform poorly if features are not scaled, because different ranges of feature values can lead to inefficient learning.
- Speeds up training: Scaling helps optimization algorithms like gradient descent converge faster.
- Prevents bias towards features with larger ranges: Features with larger values might dominate the learning process and bias the model if scaling is not applied.

Difference between Normalized & Standardized scaling

Normalization (Min-Max Scaling): It rescales the feature values to a specific range, typically [0, 1].

Formula:

X{norm} = {X – X(min)}/{X_max}- X_min}

Example: If the range of a feature is [10, 200], normalization will bring all values into the range [0, 1], where:10 becomes 0 & 200 becomes 1

Standardization (Z-score Scaling): It transforms the data to have a mean of 0 and a standard deviation of 1.

Formula:

X(std) = X – mu/sigma

Example: A feature with mean = 50 and standard deviation = 10 will be transformed to have a mean of 0 and standard deviation of 1.

Key Differences Between Normalization and Standardization

Aspect	Normalization Scaling	Standardization Scaling
Range of Values	Transforms data to a specific range, typically [0,1]	Centers data to mean 0 and scales to have unit variance
Use Case	When features have different ranges or scales(non-Gaussian data)	When data is Gaussian distributed or when algorithms assume this
Sensitive to outliers	Yes, sensitive to outliers	Less sensitive to outliers
Range after scaling	[0,1] (or another specified range)	No specific range, but data has mean 0 and standard deviation 1

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

   VIF is a measure used in regression analysis to detect the presence of multicollinearity between independent (predictor) variables. Multicollinearity occurs when one predictor variable is highly correlated with one or more other predictor variables.

   Formula for VIF: VIF = 1/(1 - R_i^2)

   is the R-squared value of the regression of the ith predictor on all other predictors.

   Why VIF Becomes Infinite?

   The VIF can become infinite (or extremely high) when:

   1. Perfect Multicollinearity: Perfect multicollinearity occurs when one predictor variable is a perfect linear combination of other predictor(s).
      In such a case, the R^2i for that predictor variable would be 1 (because the variable is perfectly predictable from the other variables).

      VIF(Xi) = 1/1 - 1 = 1/0 = infinity

   Examples of Situations Where VIF Becomes Infinite:

   Duplicate variables: If you include the exact same variable twice in the model (or one is a multiple of the other), the VIF will be infinite. For example, if you have a variable X1 and another variable X2 = 2*X1 , this will result in perfect multicollinearity.

   Dummy variable trap: When you include all categories of a categorical variable as dummy variables without leaving out one category (the reference category), this leads to perfect collinearity and infinite VIF. This is why in one-hot encoding, one category is typically excluded.

In summary, VIF becomes infinite when perfect multicollinearity exists, meaning one variable is exactly predictable from others. This should prompt you to reassess the model and address multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   A Quantile-Quantile (Q-Q) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, most commonly a normal (Gaussian) distribution. It helps determine whether a set of data follows a specific distribution.

   In a Q-Q plot:
   The x-axis represents the theoretical quantiles (i.e., quantiles of the theoretical distribution, like a normal distribution).
   The y-axis represents the quantiles of the actual data.

   If the data follows the theoretical distribution (e.g., a normal distribution), the points on the Q-Q plot will approximately lie along a 45-degree straight line.

   How to Interpret a Q-Q Plot:
   Straight line: If the points roughly form a straight line, the data is likely to follow the theoretical distribution.
   Upward or downward curve: If the points curve away from the line, this indicates deviations from the theoretical distribution.
   S-shaped curve: Indicates light tails (less variation in extremes).
   Inverted-S shape: Indicates heavy tails (more variation in extremes).

   Importance of Q-Q Plot in Linear Regression: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. The Q-Q plot helps check this assumption. Let's explore its role and importance:

   1. **Assessing Normality of Residuals:** Why it's important: In linear regression, normality of residuals ensures that hypothesis tests (like t-tests and F-tests) are valid. If the residuals are not normally distributed, the standard errors of the coefficients might be biased, leading to incorrect p-values and confidence intervals. How the Q-Q plot helps: By plotting the residuals of the regression model against a normal distribution, the Q-Q plot allows you to visually assess whether the residuals deviate from normality.
      - Residuals along a straight line: The residuals are approximately normally distributed, and you can trust the results of your linear regression.
      - Curved or scattered residuals: This indicates non-normality, suggesting that linear regression assumptions may be violated.

   2. Identifying Outliers or Heavy Tails: Why it's important: Outliers or heavy tails in the residuals can affect the accuracy of the model's coefficients. They can also cause issues like heteroscedasticity (non-constant variance of residuals).How the Q-Q plot helps: A Q-Q plot can

show if the tails of the residual distribution deviate from normality. Points deviating in the tails (upper right or lower left) indicate potential outliers or extreme values.

3. Diagnosing Model Fit: If the Q-Q plot shows significant deviations from a straight line, it might indicate that the model's assumptions (linearity, homoscedasticity, or normality) are violated. This could prompt a reassessment of the model, suggesting a transformation of variables or switching to a different model type (e.g., using robust regression).

Steps in Using a Q-Q Plot in Linear Regression:

1. Fit the regression model: Build a linear regression model using your data.

2. Extract residuals: After fitting the model, extract the residuals (the differences between the observed and predicted values).

3. Generate the Q-Q plot: Plot the residuals on the y-axis against the theoretical normal quantiles on the x-axis.

Importance of the Q-Q Plot:

1. Validity of Inference: Ensures that hypothesis tests, p-values, and confidence intervals from the regression are reliable by checking the normality of residuals.

2. Model Diagnostics: Helps diagnose potential issues with model fit, like outliers or skewed residuals.

3. Improves Model Choice: Helps in deciding whether you need to use a transformation (e.g., log transformation) or a different model type (e.g., non-linear regression).