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? (3 marks)**

Ans 1: From the box plot screenshot below.

- **Seasonal Trends**: Bookings saw a notable increase during the fall season, with a significant rise observed each year from 2018 to 2019.

- **Monthly Patterns**: Bookings were at their peak during the months of May to October, showing a consistent trend of rising from the beginning to the middle of the year and declining towards the year's end.

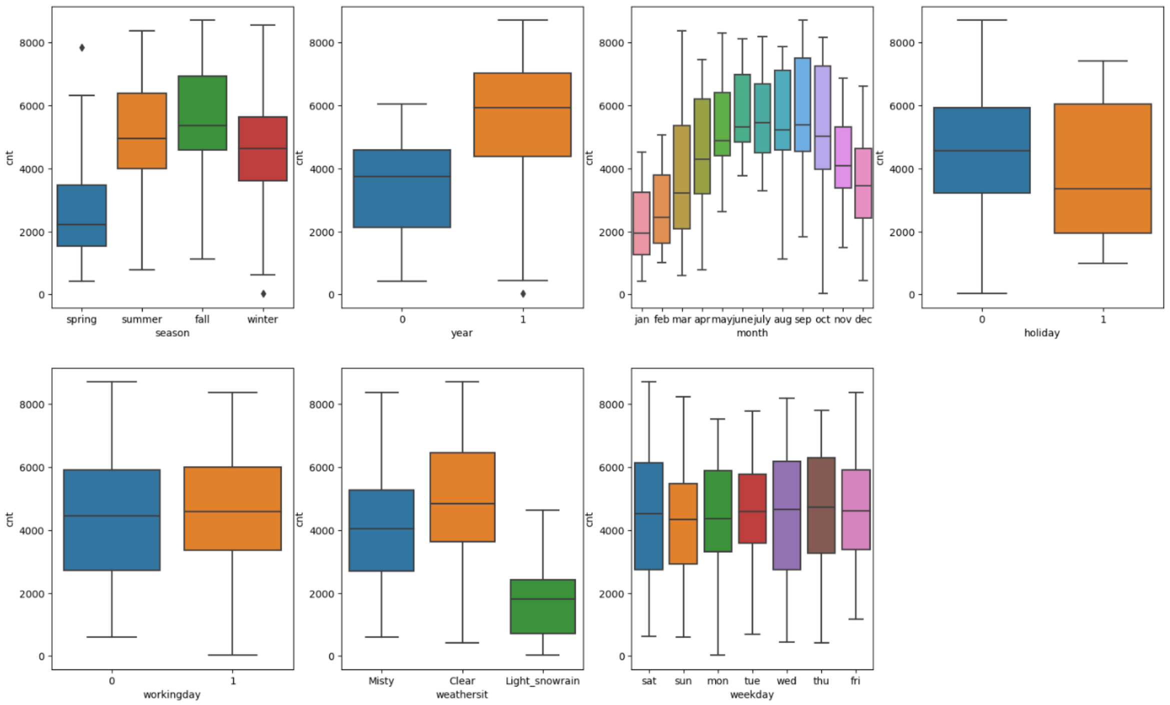
- **Weather Influence**: Bookings were higher on days with clear weather, which aligns with common expectations.

- **Day of the Week**: Thursdays, Fridays, Saturdays, and Sundays were the days with the highest number of bookings, surpassing weekdays' booking count.

- **Holiday Impact**: Bookings were generally lower on non-holiday days, suggesting that holidays influenced people to spend time at home with family.

- **Weekday vs. Weekend**: Bookings seemed balanced between working and non-working days, indicating that the day of the week didn't strongly impact the number of bookings.

- **Yearly Growth**: Bookings in 2019 showed significant growth compared to the previous year, reflecting positive progress in the business.



2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
Ans 2:
Utilizing `drop_first = True` is crucial since it serves to minimize the generation of surplus columns when creating dummy variables. This reduction effectively curtails the emergence of excessive correlations among the resulting dummy variables.

In terms of syntax, setting `drop_first` to `True` (the default being `False`) leads to the creation of k-1 dummies out of k categorical levels. Consequently, it omits the initial level when generating dummies.

For instance, when dealing with a categorical column featuring three distinct values, employing a dummy variable approach requires only two variables – A and B. The absence of A and B naturally signifies the presence of the third value, C. Thus, the inclusion of the third variable to represent C becomes redundant.

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

Ans 3: '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? (3 marks)**
Ans 4:

I have validated the assumption of Linear Regression Model based on below 4 assumptions

- Normality of Error Terms: The error terms should conform to a normal distribution.

- Multicollinearity Check: Variables should not exhibit significant multicollinearity, ensuring independence among them.

- Validation of Linear Relationship: Clear linear relationships should be evident among the variables.

- Homoscedasticity: Residual values should display uniform variance without any discernible patterns.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
Ans 5:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- winter
- sep

General Subjective Questions

1.**Explain the linear regression algorithm in detail. (4 marks)**
Ans 1:
Linear Regression Algorithm:

Linear Regression is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the independent variables (features) and the dependent variable (target).

Objective:
The primary goal of Linear Regression is to find the best-fitting straight line (regression line) that minimizes the difference between the predicted values and the actual values of the target variable. This line can then be used to make predictions for new data points.

Mathematical Representation:
The linear regression equation is typically represented as:

Y = b0 + b1*X1 + b2*X2 + ... + bn*Xn

- Y: Dependent variable (target)

- b0: Intercept (constant term)
- b1, b2, ..., bn: Coefficients for the independent variables (features) X1, X2, ..., Xn

Algorithm Steps:

1. Data Collection: Gather a dataset consisting of input features (independent variables) and corresponding target values (dependent variable).

2. Data Pre-processing: Handle missing values, outliers, and normalize/standardize the features if needed.

3. Splitting the Data: Divide the dataset into training and testing subsets. The training set is used to train the model, and the testing set is used to evaluate its performance.

4. Model Training:
   - Initialize the coefficients b0, b1, ..., bn with random values.
   - Calculate the predicted values using the linear regression equation.
   - Calculate the error (residual) between predicted and actual values.

5. Cost Function (Loss Function):
   - Calculate the mean squared error (MSE) or another appropriate cost function to quantify the error between predicted and actual values.

6. Gradient Descent:
   - Update the coefficients (b0, b1, ..., bn) iteratively using gradient descent to minimize the cost function.
   - Adjust coefficients in the direction that reduces the cost function gradient.

7. Model Evaluation:
   - Use the trained model to make predictions on the testing set.
   - Evaluate the model's performance using various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc.

8. Prediction:
   - Once the model is trained and evaluated, it can be used to make predictions on new, unseen data.

Types of Linear Regression:

1. Simple Linear Regression: Involves a single independent variable and a linear relationship between that variable and the target.

2. Multiple Linear Regression: Deals with multiple independent variables and their linear relationship with the target.

Assumptions:

Linear Regression assumes:
- Linearity: A linear relationship between features and target.
- Independence: Residual errors are independent of each other.
- Homoscedasticity: Uniform variance in residuals.
- Normality: Residual errors are normally distributed.
- No Multicollinearity: Minimal correlation among independent variables.

Advantages:
- Simplicity and interpretability.
- Well-suited for scenarios with a clear linear relationship.

Limitations:
- Assumes a linear relationship, which might not hold for all datasets.
- Sensitive to outliers.
- May not perform well with complex relationships.

In summary, Linear Regression is a fundamental algorithm for predictive modelling that aims to find the best-fitting linear relationship between input features and a continuous target variable. It serves as a building block for more complex regression techniques and is widely used in various fields.

2. **Explain the Anscombe's quartet in detail. (3 marks)**
Ans 2:
Anscombe's quartet is a famous statistical demonstration that emphasizes the importance of data visualization and the potential pitfalls of relying solely on summary statistics. It consists of four datasets that have nearly identical statistical properties when analyzed using basic summary measures (mean, variance, correlation, linear regression parameters), but they exhibit vastly different relationships when visualized.

Background:
The quartet was created by the statistician Francis Anscombe in 1973 to underscore the concept that examining data graphically can reveal patterns, trends, and anomalies that might go unnoticed when relying solely on numerical summaries. Anscombe's quartet is often used to emphasize the importance of data visualization in exploring and understanding datasets.

The Four Datasets:
The four datasets within Anscombe's quartet share these properties:
- Each dataset consists of 11 data points.
- There is one independent variable (X) and one dependent variable (Y) in each dataset.

However, when plotted, they reveal drastically different relationships:

1. Dataset I (Linear Relationship):
   - X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
   - Y values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

- When plotted, it closely resembles a linear relationship.

2. Dataset II (Non-linear Relationship):
   - X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
   - Y values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74
   - When plotted, it shows a curved relationship.

3. Dataset III (Linear Relationship with Outlier):
   - X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
   - Y values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
   - When plotted, it appears linear except for one outlier.

4. Dataset IV (Horizontal Line):
   - X values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
   - Y values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89
   - When plotted, it forms a horizontal line.

Implications:
Anscombe's quartet demonstrates that while summary statistics can provide useful insights, they are not sufficient for fully understanding the nature of data relationships. Visualizations play a crucial role in revealing patterns, trends, and anomalies that might not be evident from summary measures alone. The quartet highlights the importance of exploratory data analysis and the need to visualize data before drawing conclusions or making decisions based solely on numerical summaries.

3. **What is Pearson’s R? (3 marks)**
Ans 3:
Pearson's correlation coefficient, often denoted as "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:

- 1: Represents a perfect positive linear relationship, meaning as one variable increases, the other variable increases proportionally.
- 0: Implies no linear relationship between the variables.
- -1: Indicates a perfect negative linear relationship, meaning as one variable increases, the other variable decreases proportionally.

Pearson's correlation coefficient is sensitive to linear relationships but may not capture non-linear associations between variables. It's widely used in various fields, including statistics, data analysis, and scientific research, to assess how closely two variables vary together.

Mathematically, Pearson's correlation coefficient is calculated as:

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

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

Pearson's correlation coefficient is used to assess the strength and direction of linear relationships, and it's a fundamental tool for understanding associations between variables in statistics and data analysis.

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

Scaling is a pre-processing step in data preparation where the values of features (variables) in a dataset are transformed to a common scale. The goal is to ensure that all features have comparable magnitudes, which can help improve the performance of certain machine learning algorithms and make the optimization process more efficient.

Why Scaling is Performed:
- Many machine learning algorithms work better when features are on similar scales, as large differences in magnitudes between features can affect the model's ability to converge.
- Scaling prevents certain features from dominating others in terms of their impact on the algorithm's calculations.
- Distance-based algorithms, like k-nearest neighbors and clustering, are sensitive to the scale of features. Scaling ensures that the algorithm treats all features equally.
- Gradient descent optimization in many algorithms converges faster when features are scaled.

Normalized Scaling:
- In normalized scaling (also called Min-Max scaling), feature values are transformed to lie within a specific range, usually between 0 and 1.
- The formula for normalized scaling is: X_new = (X - X_min)/(X_max - X_min)

- This method is sensitive to outliers since it uses the minimum and maximum values to scale the data.

Standardized Scaling (Z-score normalization):
- In standardized scaling, feature values are transformed to have a mean of 0 and a standard deviation of 1.
- The formula for standardized scaling is: X_new = (X - mean)/Std

- This method is less sensitive to outliers compared to normalized scaling, as it uses the mean and standard deviation.

Key Differences:
- Range: Normalized scaling restricts values to a specific range (e.g., 0 to 1), while standardized scaling centers data around mean 0 and adjusts by standard deviation.
- Sensitivity to Outliers: Normalized scaling can be affected by outliers, while standardized scaling is more robust to outliers.
- Interpretation: Normalized scaling preserves the original distribution of the data, while standardized scaling transforms the data into a standard normal distribution.
- Usage: Normalized scaling is useful when the features have a defined range and you want to maintain that range. Standardized scaling is often used when the data distribution is not known or when there are concerns about the impact of outliers.

Both normalized and standardized scaling have their use cases, and the choice depends on the characteristics of your data and the requirements of your machine learning algorithm.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
Ans 5:
VIF (Variance Inflation Factor) is a measure used to detect multicollinearity in a regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, which can lead to unstable coefficient estimates and make it challenging to interpret the individual impact of each variable.

In the context of VIF, the occurrence of infinite VIF values usually indicates an extreme case of perfect multicollinearity. Perfect multicollinearity arises when one predictor variable can be exactly predicted by a linear combination of other predictor variables in the model. This situation creates a mathematical issue when calculating the VIF.

Here's why infinite VIF values might occur:

1. Redundant Variables: One predictor variable is a linear combination (exact or nearly exact) of other predictor variables. For example, if you have two variables that are perfectly correlated (correlation coefficient of 1), one can be expressed as a multiple of the other, causing perfect multicollinearity.

2. Dividing by Zero: In the VIF formula, division by zero can occur if the variance of the predictor variable is extremely close to zero, leading to an infinite VIF value.

3. Numerical Precision: Due to the limitations of numerical precision in computer calculations, very small numbers might be treated as zero, causing division errors when calculating VIF.

When dealing with infinite VIF values, consider the following steps:

- Check for correlated predictor variables: Investigate the correlation matrix of your predictor variables to identify highly correlated pairs that might be causing multicollinearity.

- Remove redundant variables: If you find variables that are perfectly correlated or nearly so, consider removing one of them from the model.

- Regularization techniques: If multicollinearity persists, consider using regularization techniques like Ridge Regression or Lasso Regression, which can help mitigate the impact of multicollinearity by introducing a penalty term on the coefficients.

- Data preprocessing: Check if there are data preprocessing issues, such as extremely small or zero variance in a variable, and address them.

Addressing multicollinearity is important for accurate and reliable regression modeling.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Ans 6:

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a certain theoretical distribution. It helps you check whether your data behaves like a specific distribution, such as a normal (Gaussian) distribution. In a Q-Q plot, you compare the quantiles (ordered values) of your dataset to the quantiles of the theoretical distribution you're testing against.

How a Q-Q Plot Works:

- Imagine you have a dataset of numbers. A Q-Q plot sorts these numbers and plots them against the quantiles of a theoretical distribution.
- If the points on the plot fall on or near a straight line, it suggests your data matches the theoretical distribution. If the line is roughly diagonal, your data might not match the theoretical distribution.

Use and Importance in Linear Regression:

In the context of linear regression, a Q-Q plot can help you assess the assumption of normality for the residuals (the differences between predicted and actual values). This is crucial because linear regression assumes that the residuals are normally distributed.

- Assuming Normality: If the Q-Q plot shows the residuals aligning well with a straight line, it indicates that the residuals are normally distributed. This means that the linear regression assumptions regarding the residuals' distribution are likely met.

- Detecting Deviations: If the points on the Q-Q plot deviate significantly from a straight line, it suggests that the residuals do not follow a normal distribution. This might indicate issues like outliers or other data irregularities.

- Model Reliability: Ensuring normality of residuals is important because if this assumption is violated, the reliability of your linear regression model's predictions and statistical inferences might be compromised.

In simple terms, a Q-Q plot is like comparing your data to a "normal" behavior pattern. If your data points follow the pattern, it suggests your data is behaving as expected. If not, it could signal some irregularities that might affect the accuracy of your linear regression results.