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

There were 6 categorical variables in the dataset.

The inference that We could derive were for affect of these are as follows

- ***season****: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking.*
  Per final model
  season_4: - A coefficient value of '0.128744' indicated that w.r.t season_1, a unit increase in season_4 variable increases the bike hire numbers by 0.128744 units

- ***mnth****: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month.*
  Per final model
  mnth_9: *A coefficient value of '0.094743' indicated that w.r.t mnth_1, a unit increase in mnth_9 variable increases the bike hire numbers by 0.094743 units.*

- ***Year*** *:* Most sales were in 2019 then 2018 .
  Per final model
  A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

  Weather Situation 3 (weathersit_3) – ***Per final model***
  A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units

- ***holiday****: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. Which indicated, holiday CANNOT be a good predictor for the dependent variable.*

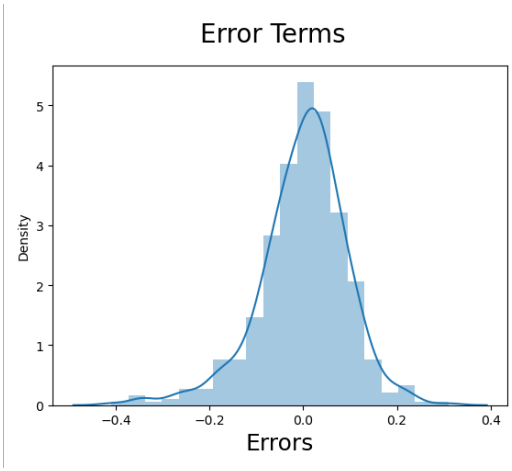
- ***weekday****: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor*
  Per final model
  weekday_6: A coefficient value of '0.056909' indicated that w.r.t weekday_1, a unit increase in weekday_6 variable increases the bike hire numbers by 0.056909 units.

- ***workingday****: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years).*
  weathersit_3:
  A coefficient value of '0.043203' indicated that, a unit increase in workingday variable increases the bike hire numbers by 0.043203 units.

2) **Why is it important to use drop_first=True during dummy variable creation?**
   We drop first dummy variable for each set of dummies created because it causes multicollinearity.
   When we create dummy variables for a categorical variable with n levels, by default we get "n" dummy variables, one for each category. However, this introduces redundancy, since values of n-1 dummy variables are known the value of the last one can always be inferred. This redundancy is what creates multicollinearity.
3) **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Target Variable was highly positively correlated with predictor variable 'temp' , 'atemp' and 'yr' with 'temp' and 'atemp' being highest at 0.63
4) **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Validation of assumption of Linear Regression was done by performing "Residual Analysis Of Training Data" i.e Error terms are normally distributed with mean zero.

Plot the histogram of error terms :

```
res = y_train-y_train_pred
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((res), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                    # Plot heading
plt.xlabel('Errors', fontsize = 18)
```



From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid

5) **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
As per our final Model, the top 3 predictor variables that influences the bike booking are:

Temperature (temp) - A coefficient value of '0.5636' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
Year (yr) - A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

General Subjective Questions

1) Explain the linear regression algorithm in detail

LinearRegression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. It is one of the simplest and most widely used machine learning models.

Linear regression has two primary purposes—understanding the relationships between variables and prediction.

The coefficients represent the estimated magnitude and direction (positive/negative) of the relationship between each independent variable and the dependent variable.
The equation allows you to predict the mean value of the dependent variable given the values of the independent variables that you specify.

Components of Linear Regression:

1. Dependent Variable (Target): This is the variable you want to predict.
2. Independent Variables (Predictors): These are the variables that are used to predict the dependent variable.
3. Coefficients (Weights): The linear regression model estimates coefficients that represent the impact of each independent variable on the dependent variable.
4. Intercept This is the value of (y) when all the predictor variables are 0. It represents the point where the regression line crosses the y-axis.

The Linear Regression Equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Types of Linear Regression:

1. Simple Linear Regression:
- There is only one independent variable.

2. Multiple Linear Regression:

- There are two or more independent variables.

Assumptions of Linear Regression:

1. Linearity: The relationship between the independent and dependent variables is linear.
2. Independence: The observations are independent of each other.
3. Homoscedasticity: The residuals (errors) have constant variance at all levels of the independent variables.
4. Normality: The residuals are normally distributed.
5. No Multicollinearity: Independent variables are not highly correlated with each other.

Example:

Suppose you want to predict the price of a house based on its size (in square feet) and number of bedrooms.

2. Explain the Anscombe's quartet in detail

Anscombe's Quartet is a set of four small datasets, each with nearly identical statistical properties, but which appear very different when graphed. It was created by the British statistician Francis Anscombe* sans the name, in 1973 to demonstrate the importance of graphing data before analyzing it. Anscombe wanted to show that relying solely on summary statistics (such as mean, variance, correlation, and linear regression coefficients) can be misleading without visualizing the data.

Key Characteristics of Anscombe's Quartet:

Despite their distinct appearances when graphed, all four datasets in the Anscombe's Quartet share very similar statistical properties. Specifically:

- Mean of X and Y: The mean of both the independent variable (X) and the dependent variable (Y) is nearly the same across all four datasets.
- Variance of X and Y: The variance (a measure of spread) of both X and Y is almost identical across the datasets.
- Correlation: The Pearson correlation coefficient between X and Y is almost the same (~0.816) for all four datasets.
- Linear regression line: Each dataset has a very similar linear regression equation, which leads to nearly the same slope and intercept values.
- Residuals: The residuals (differences between observed and predicted values) for the linear regression models are also very similar.

However, despite these nearly identical summary statistics, the datasets look very different when plotted, illustrating the limitations of relying solely on statistical summaries.

The Four Datasets

1. Dataset 1:

- This dataset resembles a typical set of points that are linearly related.
- The points are well distributed, and a linear regression line fits well.
- This dataset looks as you might expect given the summary statistics.

2. Dataset 2:

- The X values are all the same except for one outlier.

- Most of the points lie on a vertical line, but a single point influences the regression line dramatically.
- Despite having the same regression line and correlation as the first dataset, this dataset is clearly not suitable for a linear regression analysis.

3. Dataset 3:
- In this dataset, the Y values are nearly constant, except for one outlier.
- The regression line looks reasonable, but the single outlier distorts the correlation, giving a misleading impression of a strong relationship between X and Y.

4. Dataset 4:
- This dataset forms a perfect curve, with most points lying along a non-linear path.
- The linear regression line is a poor fit for this data because the relationship is clearly not linear.
- Nonetheless, the correlation and regression values are similar to those of the other datasets.

Visualization of Anscombe's Quartet:

When you plot these four datasets:
- Dataset 1 looks like a typical scatter plot with a clear linear relationship between X and Y.
- Dataset 2 has an extreme outlier that heavily influences the linear regression line.
- Dataset 3 is mostly a horizontal line, with one outlier skewing the correlation and regression results.
- Dataset 4 shows a non-linear relationship, where a quadratic or other nonlinear model would be more appropriate.

Inference:

1. The Danger of Relying Solely on Summary Statistics:
- The quartet demonstrates that summary statistics like means, variances, and correlations can be identical across datasets that are dramatically different in nature.
- Without visualization, these datasets could be misunderstood.

2. The Importance of Graphing Data:
- Anscombe's quartet reinforces the idea that it's crucial to visualize data before performing analysis. Graphs can reveal patterns, outliers, and non-linear relationships that statistics alone might obscure.

3. The Impact of Outliers:
- Outliers can have a significant impact on statistical analysis, such as correlation and linear regression. Outliers may skew results, as seen in Datasets 2 and 3, and graphs can help in identifying these outliers
4 The Limitations of Linear Regression:
- Linear regression assumes a linear relationship between variables, but not all relationships are linear, as shown in Dataset 4. Visualization can help assess whether a linear model is appropriate or if a different model is needed.

3) What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies both the strength and direction of the relationship, with values ranging from -1 to 1. Some underlying properties of same are :

1. Range:
- +1: Perfect positive linear relationship (as one variable increases, the other also increases).
- -1: Perfect negative linear relationship (as one variable increases, the other decreases).
- 0: No linear relationship between the variables.

2. Formula:
r = [n(Σxy) − ΣxΣy] / sqrt{n(Σx^2) − (Σx)^2][n(Σy^2) − (Σy)^2]}..

3. Interpretation:
- r > 0: Positive correlation (both variables move in the same direction).
- r < 0: Negative correlation (variables move in opposite directions).
- r = 0: No linear correlation (but other types of relationships may exist).

4. Strength of Correlation:
- 0.0 to 0.3 (or -0.3): Weak correlation.
- 0.3 to 0.7 (or -0.3 to -0.7): Moderate correlation.
- 0.7 to 1.0 (or -0.7 to -1.0): Strong correlation.

To further summarize
- Pearson's R only measures linear relationships. If the relationship is non-linear, Pearson's R may not capture it well.
- It is sensitive to outliers, which can distort the correlation.

Pearson's R is widely used in statistics to understand the association between two continuous variables.

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

Scaling is a technique used to adjust the range or distribution of data, typically by transforming numerical features in a dataset. It is done to ensure that features are on a similar scale, which can improve the performance of machine learning algorithms, especially those that rely on the distance between data points (like gradient-based algorithms or those based on distance metrics such as K-nearest neighbors and support vector machines).

Why is Scaling Performed?
Scaling is performed for several reasons:

1. Improves Algorithm Performance: Many machine learning algorithms (like linear regression, logistic regression, KNN, and neural networks) assume that the features are on similar scales. If not, the feature with a larger range might dominate the learning process, leading to skewed or inaccurate models.

2. Improves Convergence: Scaling can help algorithms like gradient descent converge faster by preventing features with larger ranges from dominating the gradient calculations.

3. Equal Contribution of Features: Features with larger ranges can disproportionately affect models that are sensitive to the magnitude of data values (e.g., SVM, KNN, and neural networks). Scaling ensures that all features contribute equally to the model.

4. Distance-based Models: Models that use distances (e.g., K-means clustering, KNN) need scaled data because unscaled features with large ranges can distort the distance measurements between data points.

3*. Difference Between Normalized Scaling and Standardized Scaling*

1. Normalization (Min-Max Scaling):
- Purpose: Normalization transforms the data to fit within a fixed range, usually [0, 1].
- Formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- When to Use: Normalization is useful when the data does not follow a normal distribution and you want to scale features to a fixed range (e.g., [0, 1]).
- **Use Case:** Often used in algorithms like KNN, neural networks, or in any situation where the algorithm depends on the magnitude of values.

Example: A feature with a range of values between 100 and 1000 will be scaled to the range [0, 1], preserving the relative distances between data points.

2. Standardization (Z-score scaling):
- Purpose: Standardization transforms the data to have a mean of 0 and a standard deviation of 1. It centers the data around the mean and scales it according to the standard deviation.

Where "mu" is the mean of the feature and \(\sigma \) is the standard deviation.
- When to Use: Standardization is useful when data follows a normal (Gaussian) distribution, and you want to ensure that the features have similar distributions.
- **Use Case:** Often used in algorithms like logistic regression, linear regression, support vector machines, and principal component analysis (PCA), which assume that the data is normally distributed.
Example: A feature with a mean of 50 and a standard deviation of 10 will be standardized 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?

VIF is a measure used to detect the degree of multicollinearity in a set of regression variables. It quantifies how much the variance of a regression coefficient is inflated due to multicollinearity.

When the Variance Inflation Factor (VIF) is infinite (or extremely large), it indicates a serious issue of perfect multicollinearity in your dataset. This means that one or more of the independent variables (features) are perfectly or almost perfectly correlated with each other.

- A low VIF (close to 1) suggests that the variable has little or no correlation with other independent variables.
- A high VIF (greater than 5 or 10) indicates significant multicollinearity, meaning the variable is highly correlated with other variables in the model.

Why VIF can be Infinite:
When VIF becomes infinite, it happens because:
1. Perfect Collinearity: One or more independent variables are perfect linear combinations of other variables. In other words, one feature is an exact replica or linear transformation of another. This makes the correlation between them 1 (or -1), and the regression model cannot compute unique solutions for the coefficients.

2. Division by Zero: VIF is calculated as:
$$VIF = 1/(1 - R^2)$$

Where (R^2) is the coefficient of determination from a regression of one independent variable on all other independent variables. If the \(R^2 \) value is 1 (indicating perfect multicollinearity), then \(1 - R^2 \) becomes 0, and the VIF calculation results in a division by zero, making VIF infinite.

Consequences:
- Unstable Coefficients: If multicollinearity exists, the coefficients of the regression model become unstable, leading to unreliable estimates.
- Model Inaccuracy: The presence of infinite or very high VIF values makes it hard for the model to discern the contribution of each variable, leading to potential overfitting or poor generalization to new data.

How to Address Infinite VIF:
1. Remove Highly Correlated Variables: If two or more variables are perfectly correlated, remove one of them from the model.

2. Combine Variables: If two variables are highly correlated but still useful, you might combine them into a single feature using techniques like Principal Component Analysis (PCA).

3. Regularization: Apply regularization techniques like Ridge Regression or Lasso to reduce the impact of multicollinearity by shrinking the coefficients of correlated variables.

Conclusion:
When VIF is infinite, it indicates perfect multicollinearity, meaning one variable is a perfect linear combination of others. This makes it impossible for the model to compute reliable estimates for the coefficients, and removing or combining highly correlated variables can help resolve this issue.

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

Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution, most commonly the normal distribution. It compares the quantiles of the dataset against the quantiles of a reference distribution (usually a normal distribution).

How Q-Q Plot Works:

- X-axis: The theoretical quantiles (expected quantiles from a specified distribution, such as the normal distribution).
- Y-axis: The quantiles from the sample data.

If the data follows the specified distribution (e.g., normal distribution), the points in the Q-Q plot should fall approximately along a 45-degree straight line.

Why is Normality of Residuals Important?

- Assumption of Linear Regression: Linear regression assumes that the residuals are normally distributed. This assumption is crucial for the accuracy of hypothesis tests (e.g., t-tests for coefficients) and confidence intervals in regression.
- Impact on Statistical Inference: If the residuals are not normally distributed, the standard errors, p-values, and confidence intervals generated by the regression model may be invalid or misleading.

Steps to Use a Q-Q Plot in Linear Regression:

1. Fit a Linear Regression Model: First, fit your linear regression model to the data.
2. Extract Residuals: Compute the residuals (the difference between the observed and predicted values).
3. Plot the Residuals: Generate a Q-Q plot of the residuals against the quantiles of a normal distribution.
4. Interpret the Q-Q Plot:
   - Points close to the straight line: If the points fall roughly along the 45-degree line, the residuals are approximately normally distributed, and the assumption of normality holds.
   - Points deviating from the straight line: Significant deviations (such as curves or clusters) indicate that the residuals are not normally distributed, which violates one of the linear regression assumptions.

Common Patterns in a Q-Q Plot and Their Meaning:

1. Straight line (normal distribution):
   - The residuals follow a normal distribution, and the assumption of normality holds.

2. S-shaped curve:
   - This suggests the presence of heavy tails (either too many high or low residuals). This could mean the residuals have more extreme values than a normal distribution, indicating non-normality.

3. U-shaped curve:
   - This suggests that the residuals are under-dispersed or over-dispersed. This means the distribution is skewed, indicating non-normal residuals.

4. Steep rise in tails:

- This indicates the presence of outliers, where extreme values deviate from the expected normal distribution, again pointing to non-normality.

Importance of Q-Q Plot in Linear Regression:

In the context of linear regression, a Q-Q plot is primarily used to check whether the residuals (the differences between the observed and predicted values) follow a normal distribution. This is important because one of the key assumptions of linear regression is that the residuals are normally distributed.

1. Assessing Model Fit: The normality of residuals is a key assumption in linear regression. A Q-Q plot helps to check this assumption visually, ensuring that statistical tests based on the model (like confidence intervals and p-values) are valid.

2. Detecting Outliers: A Q-Q plot can reveal if there are outliers in the residuals, as these will appear as points far from the reference line in the plot. This helps identify problematic data points that could affect the regression model.

3. Handling Skewed Residuals: If the Q-Q plot shows a systematic deviation (e.g., skewness or kurtosis), it may suggest that a transformation (like a log or square-root transformation) is needed to better meet the assumptions of linear regression.

4. Model Diagnostics: The Q-Q plot is an essential diagnostic tool for checking whether the model assumptions hold and whether adjustments or more complex model are needed.
