1) Based off my observations I can infer the following

    aWe see more bookings during the Fall season. We also see a steep rise of booking in 2019, eg: Jan 2018 vs Jan 2019 we see a steep rise.
    More bookings seen between May to October, trend follows a bell curve
    When the weather is clear we see more bookings
    Wed, Thu, Fri, Sat saw more bookings
    Saw same bookings on working and non-working days
    Overall in 2019 we saw more bookings than 2018

2) Drop_first = True is important to use as it helps reduce the extra column created during dummy variable creation. Thus the reducing of the correlations among dummy variables.

   For example if we have three types of values in a Catagorical column then we can create dummy variable for that column. So, if a variable is not A or B the it obvious that it is C. We don't need to create a 3rd variable to identify it.

3) ‘temp’ variable has the closest correlation with the target variable

4) I have validated the Linear Regression Model based on the following five assumptions-

    Normality of error terms
    Multicollinearity check
    Linear relationship validation
    Homoscedasticity
    Independence of residuals

5) The top three features contributing significantly towards explaining the demand of the shared bikes are as follows -

    temp
    winter
    sep

General Subjective Answers

1) Linear regression is a fundamental statistical and machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. It's a method to find the best-fitting linear relationship between the input variables and the target variable. The goal of linear regression is to predict the value of the dependent variable based on the values of the independent variables.

Here's a detailed explanation of the linear regression algorithm:

- Assumptions: Linear regression is based on several assumptions:

    - Linearity: The relationship between the independent and dependent variables is assumed to be linear.
    - Independence: Observations are assumed to be independent of each other.
    - Homoscedasticity: The variance of errors (residuals) is constant across all levels of the independent variables.
    - Normality: The errors are normally distributed.

- Simple Linear Regression vs. Multiple Linear Regression:
    - Simple Linear Regression: This involves a single independent variable (feature) and one dependent variable (target). The goal is to find the best-fit line that minimizes the distance between the predicted values and the actual values.
    - Multiple Linear Regression: This extends the concept to more than one independent variable. The goal is to find the best-fit hyperplane in a higher-dimensional space.
- Model Evaluation: Once the model parameters are estimated, it's crucial to evaluate its performance. Common evaluation metrics include:
    - R-squared (Coefficient of Determination): Measures the proportion of the variance in the dependent variable that's predictable from the independent variables.
    - Root Mean Squared Error (RMSE): The square root of the average squared differences between predicted and actual values.
- Extensions and Variations: Linear regression can be extended in various ways, such as Ridge Regression and Lasso Regression, which introduce regularization to prevent overfitting. These methods add penalty terms to the cost function to constrain the magnitudes of the coefficients.
  Linear regression is a simple yet powerful algorithm that serves as a foundation for more complex modeling techniques and is widely used in various fields including economics, finance, biology, and machine learning

2) Anscombe's quartet is a set of four small datasets that have nearly identical statistical properties when analyzed using basic summary statistics like mean, variance, correlation, and linear regression parameters. However, when visualized, these datasets reveal dramatically different patterns and relationships, highlighting the importance of data visualization in understanding and interpreting data. The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the limitations of relying solely on summary statistics and the need for graphical exploration of data.

Here's a detailed explanation of Anscombe's quartet:

1. The Datasets: Each dataset in Anscombe's quartet consists of 11 paired (x, y) data points. The four datasets are labeled I, II, III, and IV. Despite having different values, they all share similar summary statistics:
    - Mean of x: 9.0
    - Mean of y: 7.5

  - Variance of x: 11.0
  - Variance of y: 4.125
  - Correlation coefficient: Approximately 0.816
  - Linear regression line:
  - y=3+0.5x
- Visual Exploration:

While the summary statistics are nearly identical, the datasets exhibit drastically different patterns when graphed:

  - Dataset I: Forms a clear linear relationship with a tight fit, perfectly following the linear regression line.
  - Dataset II: Appears as a curve with a slight upward trend, deviating from linearity.
  - Dataset III: Contains an outlier that strongly affects the linear regression line, despite other points forming a linear relationship.
  - Dataset IV: Has a strong nonlinear relationship, with an outlier affecting the correlation and regression parameters.
- Implications:

  Anscombe's quartet illustrates several crucial points:

  - Relying solely on summary statistics might not reveal the true nature of the data.
  - Visualization can uncover patterns, trends, and anomalies that statistical measures might overlook.
  - Summary statistics can be misleading, especially when dealing with complex datasets.
  - Linear regression parameters and correlation values can be heavily influenced by outliers.
- Importance of Data Visualization:

The quartet serves as a reminder that graphical exploration is an essential step in understanding data before drawing conclusions. Data visualization tools, such as scatter plots and line plots, allow researchers to identify patterns and relationships that might not be apparent through summary statistics alone.

- Teaching and Communication:

Anscombe's quartet is often used in statistics education to teach the importance of visualization and the limitations of summary statistics. It emphasizes critical thinking when interpreting data and highlights the potential pitfalls of relying solely on quantitative measures.

In summary, Anscombe's quartet is a powerful demonstration of how datasets with similar summary statistics can have vastly different visual patterns and relationships. It serves as a cautionary example of the need to combine statistical analysis with effective data visualization for a comprehensive understanding of data.

3) Pearson's correlation coefficient, often denoted as r or Pearson's
r , is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is one of the most commonly used correlation coefficients and is named after Karl Pearson, the statistician who introduced it.

The Pearsons correlation coefficient, r can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the variables, greater than 0 value indicates positive association and less than 0 means negative association.

4) Scaling is a preprocessing technique used in data analysis and machine learning to transform numerical variables into a specific range or distribution. The purpose of scaling is to ensure that all features or variables are on a similar scale, which can improve the performance of various algorithms and analyses. Scaling is particularly important when working with algorithms that are sensitive to the scale of input features.

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

1. Algorithm Performance: Many machine learning algorithms, such as gradient descent-based optimization algorithms (e.g., in linear regression or neural networks), work more effectively when features are on a similar scale. This can lead to faster convergence and better model performance.
2. Distance-Based Methods: Algorithms that rely on distance calculations, like k-nearest neighbors and support vector machines, can be heavily influenced by the scale of features. Scaling ensures that the distances are more meaningful and prevent one feature from dominating the others.
3. Regularization: Regularization techniques, like Lasso and Ridge regression, involve penalizing large coefficient values. Scaling helps ensure that all features have comparable influences on the regularization term.
4. Visualization: Scaling can make visualizations more interpretable by avoiding the domination of one variable over others due to scale differences.

Difference between Normalized Scaling and Standardized Scaling:

Normalized Scaling: Normalization scales features to a specific range, usually between 0 and 1. It preserves the relative relationships between data points while ensuring they fall within a uniform range.

Standardized Scaling (Z-score scaling): Standardization transforms features to have a mean of 0 and a standard deviation of 1. It centers the data around zero and scales it based on the variability.

5) The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a multiple linear regression model. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to determine the individual effects of each variable on the dependent variable. VIF helps quantify the extent to which multicollinearity affects the estimation of regression coefficients.

The occurrence of an infinite VIF value indicates the presence of perfect multicollinearity in the regression model, where one variable can be perfectly predicted from a linear combination of other variables. This issue needs to be addressed to ensure the reliability of the regression analysis.

6) A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess the distributional similarity between a sample dataset and a theoretical distribution, such as the normal distribution. It's a commonly used visualization technique to determine whether the data follows a specific distribution and to identify deviations from that distribution.
The Q-Q plot works by plotting the quantiles of the sample data against the quantiles of the theoretical distribution. If the data perfectly follows the theoretical distribution, the points on the plot will fall along a straight line (the 45-degree line). Deviations from this line indicate differences in distribution.
Here's how a Q-Q plot is used and its importance in linear regression:

1) Distribution Assessment
2) Identify Non-Normality
3) Model Validation
4) Residual Analysis
5) Outlier Detection