Assignment-based Subjective Questions

Q.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 : My analysis have observed the following outputs:

- Peak bike demand: Fall season
- Decreased bike demand: Spring
- Bike demand (2019) > Bike demand (2018)
- Peak demand months: May to October
- High demand during: Clear, misty/cloudy weather
- Lower demand during: Light rain, light snow
- Consistent demand: Weekdays, unaffected by working day status

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

Ans : The importance to use drop_first=True during dummy creation :

- Helps achieve n-1 dummy variables by excluding an extra column during dummy variable creation.
- Example: In a scenario with three variables (e.g., Furnished, Semi-furnished, and Un-furnished), it ensures only 2 variables are used to avoid redundancy.
- Reduces collinearity between dummy variables, improving the model's stability and interpretability.

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

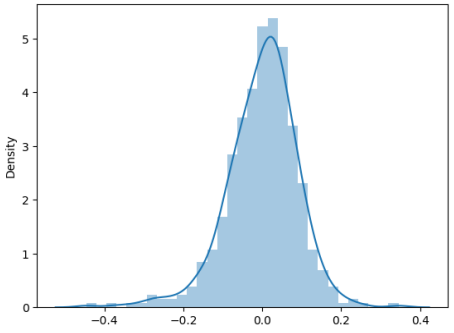
Ans: The variables ‘atemp’ and ‘temp’ both exhibit a strong correlation with the target variable which is the highest among all numerical variables.

Q.4 How did you validate the as
sumptions of Linear Regression after building the model on the training set? (3 marks)

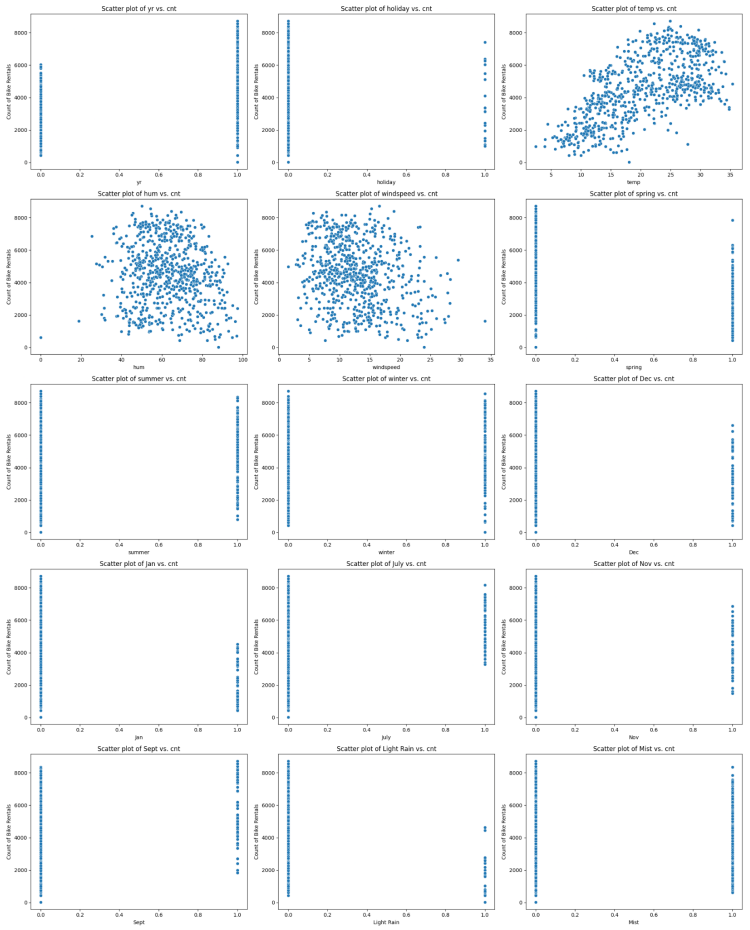
Ans: To validate the assumptions of Linear Regression, we create a distribution plot (distplot) of the residuals. This plot helps us determine if the residuals follow a normal distribution and have a mean of zero. The diagram confirms that the residuals exhibit a normal distribution with a mean equal to zero.

We have performed several other assumptions too like **Linearity** ,**Multicollinearity**.

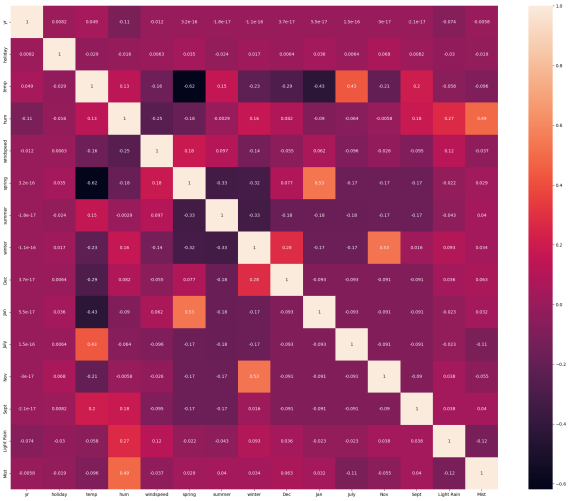
Residual Analysis



Linearity Check



Heatmap



Q.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: Based on the final model the top 3 features contributing significantly towards explaining the demand of the shared bikes are :

1. **'Temp'** : Temperature with coefficient of 0.438717 which indicates a unit increase in temp will increase the bike hire number by 0.438717 units.
2. '**Light Rain':** This variable has the coefficient of -0.291749 which means a unit increase in the 'Light Rain' variable will decrease the demand of the bike hiring by -0.291749. These make this feature significant towards explaining the demand of shared bikes.
3. **'yr' :** Year has a coefficient of 0.234455 which means a unit increase in this variable will increase the demand by 0.234455 units. This shows that there are many chances of demand for bike hiring next year.

General Subjective Questions

Note : We took the help of external sources in order to answer these questions and understand these topics better.

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

Ans: Linear regression is a fundamental machine learning algorithm used for predictive analysis. It's vital to grasp its mechanics to understand how it operates and apply it effectively.

Introduction:

Linear regression aims to model the relationship between a dependent variable (usually denoted as 'y') and one or more independent variables (often denoted as 'x'). The goal is to find a linear equation that best represents this relationship.

Mathematical Representation:

In its simplest form, the linear regression equation for a single independent variable can be represented as:

$y = mx + b$

where:

- y is the dependent variable we want to predict,
- x is the independent variable,
- m is the slope (or coefficient) of the line, and
- b is the intercept (the point where the line intersects the y-axis).

Objective:

The objective of linear regression is to find the best-fitting line (often referred to as the "regression line") that minimizes the difference between the predicted values from this line and the actual values of the dependent variable.

How the Algorithm Works:

1. Initialize:

Start with random values for m and b.

2. Calculate Predictions:

Use the current m and b to predict y for each x using the linear equation.

3. Calculate Error:

Calculate the error for each prediction by finding the difference between the predicted y and the actual y values.

4. Update Parameters:

Adjust m and b to minimise the error. Typically, this is done using optimization algorithms like gradient descent, where we iteratively adjust mm and bb based on the derivative of the error function.

5. Repeat:

Repeat steps 2-4 until the error is minimised (convergence is achieved) or a specified number of iterations is reached.

Key Concepts:

Cost Function:
A common cost function used is the Mean Squared Error (MSE), which measures the average squared differences between predicted and actual values.

Gradient Descent:
It's an optimization algorithm that helps adjust the parameters (mm and bb) to minimize the cost function, moving in the direction of the steepest decrease in the error.

Learning Rate:
A hyperparameter that controls the step size in the gradient descent algorithm. It's crucial to choose an appropriate learning rate to ensure convergence without overshooting or taking excessively small steps.

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

Ans: Anscombe's quartet comprises four distinct datasets, each with 11 data points, created by Francis Anscombe in 1973. Despite having nearly identical summary statistics such as means, variances, and correlation coefficients, these datasets illustrate dramatically different patterns when graphically visualized. Dataset I displays a straightforward linear relationship between x and y, while Dataset II showcases a non-linear pattern akin to a quadratic curve. Dataset III, although seemingly linear, is heavily impacted by an outlier, emphasizing the influence of atypical data points. Conversely, Dataset IV is dominated by a single outlier, making any discernible relationship between x and y essentially absent. Anscombe's quartet serves as a powerful reminder that summary statistics alone may not provide a complete understanding of data; visual representation is essential to grasp the underlying relationships and patterns within the dataset. This exemplifies the importance of incorporating data visualization into the analysis to gain a comprehensive insight into the nature of the data.

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

Ans: Pearson's correlation coefficient, denoted as r, is a statistical measure used to determine the strength and direction of a linear relationship between two continuous variables. It quantifies how well the variables' relationship can be described by a straight line. The coefficient ranges from -1 to 1:

- $r = 1$ indicates a perfect positive linear relationship,
- $r = -1$ signifies a perfect negative linear relationship, and
- $r = 0$ implies no linear relationship.

The calculation of r involves the covariance and standard deviations of the variables. It is widely used in various fields, including statistics, economics, psychology, and biology, to analyze relationships, build predictive models, and make informed decisions. Pearson's r is based on the assumption of a linear relationship between the variables, which might not always hold true for complex relationships. It's essential to interpret the correlation coefficient in conjunction with the context of the data and consider other forms of correlation (e.g., Spearman's rank correlation) for non-linear relationships. Understanding Pearson's r is fundamental in statistical analysis, aiding researchers and analysts in exploring and understanding relationships between variables.

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

Ans: Scaling in data preprocessing refers to the process of transforming the values of variables to a standardized range or distribution, facilitating better analysis and modeling. It's crucial in various machine learning algorithms and statistical techniques to ensure fair comparison and equal weighting of features.

Reasons for Scaling:

Equal Treatment of Features:

In many algorithms, features with larger numerical values may dominate the learning process. Scaling helps avoid this bias and treats all features equally.

Algorithm Sensitivity:

Some algorithms (e.g., k-nearest neighbors, support vector machines) are sensitive to the scale of the input features. Scaling ensures consistent behavior and performance.

Gradient Descent Convergence:

Algorithms using gradient descent for optimization converge faster when features are on a similar scale.

Distance-based Algorithms:

Scaling ensures distance-based metrics (e.g., Euclidean distance) are meaningful and not dominated by features with large values.

Difference between Normalized Scaling and Standardized Scaling:

- Range:
  - Normalized scaling scales features to a range between 0 and 1.
  - Standardized scaling does not confine values to a specific range and can result in negative values.
- Sensitivity to Outliers:
  - Normalized scaling is sensitive to outliers due to the range defined by minimum and maximum values.
  - Standardized scaling is less affected by outliers due to the use of mean and standard deviation.
- Distribution Shape:
  - Normalized scaling alters the shape of the original distribution, especially if the range is significantly different from the original range.
  - Standardized scaling maintains the shape of the original distribution.

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

Ans: The Variance Inflation Factor (VIF) is a measure used to quantify the extent of multicollinearity in a set of predictor variables in a regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it challenging to distinguish their individual effects on the dependent variable.

When the VIF is calculated as infinite, it usually indicates a perfect multicollinearity. Perfect multicollinearity occurs when a set of independent variables is perfectly linearly dependent, meaning one of the independent variables can be exactly predicted from the others using a linear combination. In such cases, the VIF is not defined because the correlation is perfect, and the calculation of VIF involves division by zero, resulting in an infinite value.

Perfect multicollinearity can stem from various reasons, such as:

Linear Dependence: One variable is a linear combination of other variables in the model, making them perfectly collinear.

Data Error or Duplicates: Data entry errors or duplication of observations can lead to identical values, causing perfect multicollinearity.

Perfect multicollinearity is problematic for regression analysis, making it difficult to estimate the coefficients accurately. It often necessitates addressing the issue by removing one of the collinear variables, combining them, or employing other methods to resolve the linear dependency. Detecting and resolving multicollinearity is crucial for obtaining reliable and interpretable regression results.

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

Ans : A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to assess whether a dataset follows a specified probability distribution, typically the normal distribution. It's a visual comparison of the observed quantiles of the data against the quantiles that would be expected if the data were normally distributed.

Components of a Q-Q Plot: A Q-Q plot consists of:

Sample Quantiles (Observed Data): Plotted on the x-axis, these are the data's quantiles, usually sorted in ascending order.

Theoretical Quantiles (Expected Data): Plotted on the y-axis, these are the quantiles that the data would have if it followed the specified distribution (e.g., normal distribution).

Use and Importance in Linear Regression: In linear regression, Q-Q plots serve several essential purposes:

- **Normality Assessment:** Q-Q plots are crucial for checking the normality assumption of the errors (residuals) in a linear regression model. If the residuals are normally distributed, the Q-Q plot will closely follow a straight line.
- **Identifying Skewness and Outliers:** Deviations from the straight line in a Q-Q plot can indicate skewness or outliers in the data. If the plot deviates significantly, it suggests that the data may not be normally distributed.
- **Validating Regression Assumptions:** Linear regression assumes that the residuals are normally distributed with a mean of zero and constant variance. A Q-Q plot helps validate this assumption.
- **Model Validation:** A Q-Q plot can also be used to compare the distribution of the predicted values from a model with the desired distribution. This aids in assessing the model's overall validity.
- **Comparing Different Distributions:** Besides normality, Q-Q plots can be used to compare the data's distribution to other distributions (e.g., log-normal, exponential) to determine the best-fit distribution.

Q-Q plots are a powerful visualization tool in statistics, particularly in linear regression, enabling the assessment of normality assumptions and the detection of outliers or deviations from the expected distribution. They provide valuable insights into the quality and validity of the regression model and help researchers make informed decisions regarding the appropriateness of the assumed data distribution.