LINEAR REGRESSION ASSIGNMENT

US Bike Sharing

ABSTRACT

A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.

AFZAL AHMAD

Data Science

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a supervised machine learning technique used to establish a relationship between a dependent variable (referred to as the target) and one or more independent variables, known as predictors. The primary objective is to create a linear equation that best explains this relationship. This equation comprises coefficients associated with the predictors and a constant term, all of which are estimated through an iterative process. The goal is to identify the line that minimizes prediction errors, and this line is known as the best fit line.

During the process of determining the best fit line, we calculate the sum of squared residuals (SSR) and aim to minimize this value. The optimization of SSR is achieved using the gradient descent algorithm, which iteratively adjusts the coefficients to reach a minimum SSR.

To assess the predictive performance of our model, we utilize a metric called R-squared (R^2). R^2 quantifies how well the model fits the data and can range from 0 to 1. A higher R^2 value, closer to 1, indicates a better-fitting model, signifying that a larger proportion of the variance in the target variable is explained by the predictors.

Q2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a collection of four small datasets that have nearly identical simple descriptive statistics, yet they exhibit vastly different characteristics when graphically visualized or analysed further. This dataset was created by the British statistician Frank Anscombe in 1973 to emphasize the importance of data visualization and the potential pitfalls of relying solely on summary statistics.

The four datasets in Anscombe's quartet have the following properties:

Dataset I:

X-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

Y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82

Dataset II:

X-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

Y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26

Dataset III:

X-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

Y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42

Dataset IV:

X-values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8

Y-values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91

Significance of these quartets:

1. When we calculate statistics like mean, variance, correlation coefficient, and linear regression for each of these datasets, it's found that all are similar or nearly identical.
2. When these points are plotted, each dataset exhibit a different pattern. Dataset I exhibit a roughly linear relationship, Dataset II shows a curved pattern, Dataset III demonstrates an outlier influence, and Dataset IV has a clear outlier that strongly influences the regression line.
3. These quartets remind us to give importance to EDA process and not to solely rely on summary statistics. Tools like scatter plots, histograms, and regression lines can provide deeper insights into the underlying patterns and outliers within the data.

Q3. What is Pearson's R?

Ans:

Pearson's correlation coefficient, often denoted as Pearson's R, is a statistical measure used to quantify the strength and direction of the linear association between two continuous variables. It provides insight into the degree of correlation between the variables, with values ranging between -1 and +1.

A Pearson's R of 1 signifies a perfect positive linear relationship, indicating that as one variable increases, the other also increases linearly. Conversely, a Pearson's R of -1 denotes a perfect negative linear relationship, where as one variable increases, the other decreases linearly. A value of 0 suggests no linear relationship; however, other types of relationships may still exist.

It's important to note that Pearson's R assumes the presence of a linear relationship between the variables and that the data adheres to a bivariate normal distribution. This method is sensitive to outliers and may not effectively capture non-linear relationships.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is a preprocessing step that involves adjusting the values of the different predictor variables so that all of them are in a similar scale.

Scaling is performed because:

1. When all are in uniform range, the impact different variables are easier to interpret.
2. The model reaches convergence faster.

Scaling just affects the coefficients of the predictors and not the statistical parameters.

Two types of scaling:

Normalization (Min-Max)	Standardization (Z-score Scaling)
It scales down the feature values within 0 to 1.	Variables are scaled such that they have zero mean and standard deviation of 1
$X = x - \min(x) \,/\, \max(x) - \min(x)$	$X = x - \text{mean}(x) \,/\, \text{sd}(x)$
Useful if needed to preserve the relationship between the data points	Useful when the algorithm is given a data that is normally distributed

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: The variable inflation factor (VIF) is a statistical metric used to assess the extent of multicollinearity among the predictors within a regression model. VIF quantifies the degree of multicollinearity, and it can become infinite in the case of perfect multicollinearity. Perfect multicollinearity arises when one variable can be precisely predicted as a linear combination of other independent variables. VIF can reach infinity under the following circumstances:

It's important to identify and address high VIF values, typically by removing or transforming variables that contribute to multicollinearity, to ensure the stability and interpretability of a regression model.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It helps in comparing the quantiles of the observed data to the quantiles of the expected theoretical distribution.

Importance of Q-Q plots:

1. Q-Q plots help assess whether the assumption by Linear regression models about the normal distribution of residuals holds true.
2. These plots help reveal the presence of outliers in the data points.
3. It helps determine the accuracy of the model by exhibiting the normal distribution of the residuals.
4. The plot indicates If the model deviates from the assumed distribution in which case it would be necessary to go for another regression model.

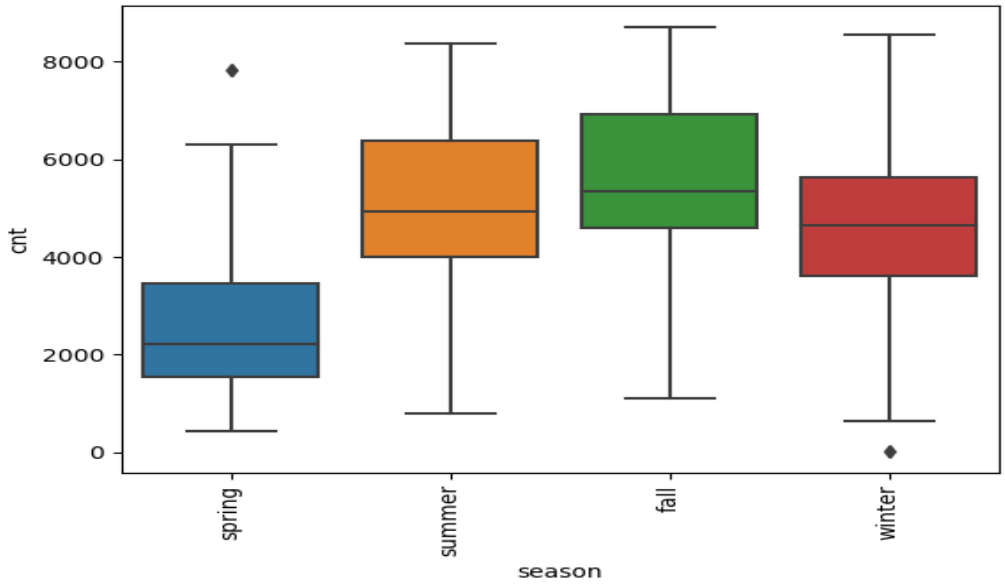
In conclusion, Q-Q plots help in assess normality assumption and ensures the model adequately fits the data.

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

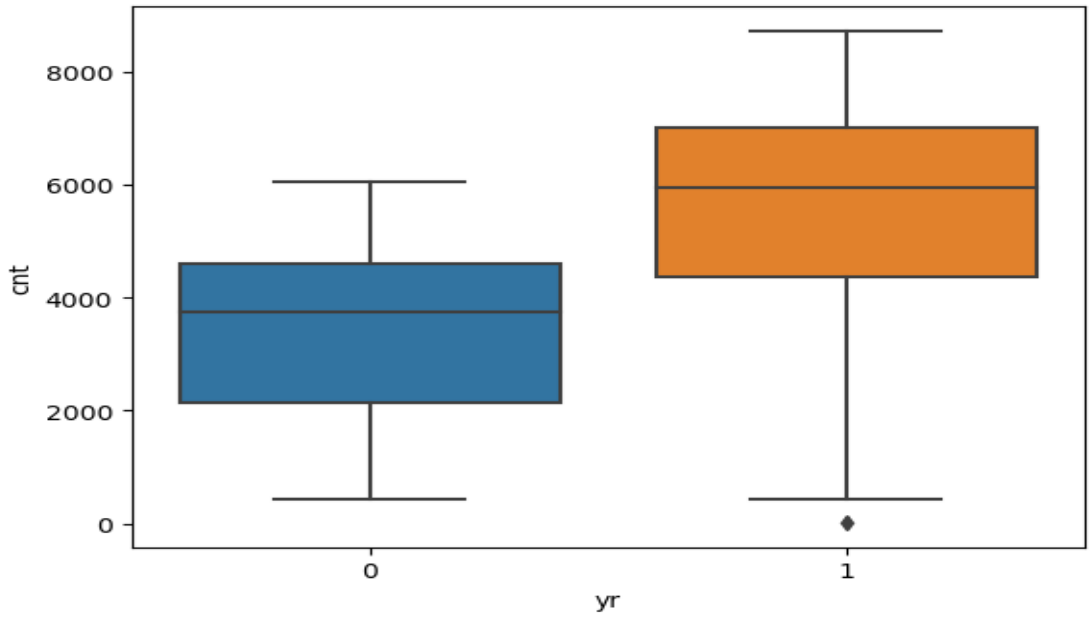
Ans. I have plotted the categorical variables with the target variables on boxplot and has inferred following effect on target:

1. Season 3: fall has highest demand for rental bikes

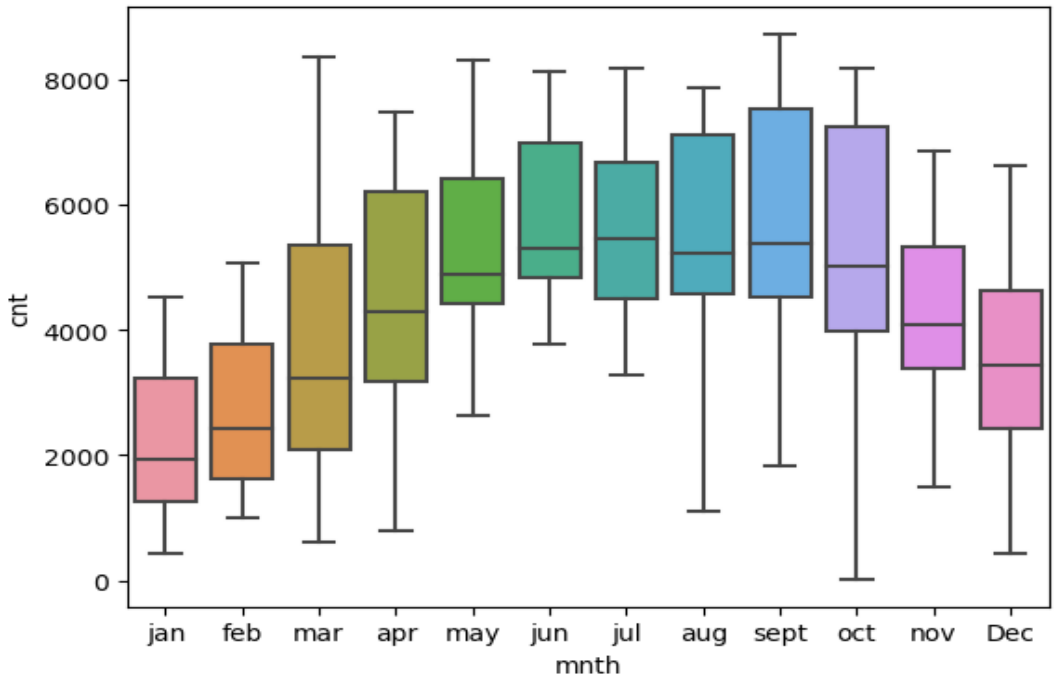


---

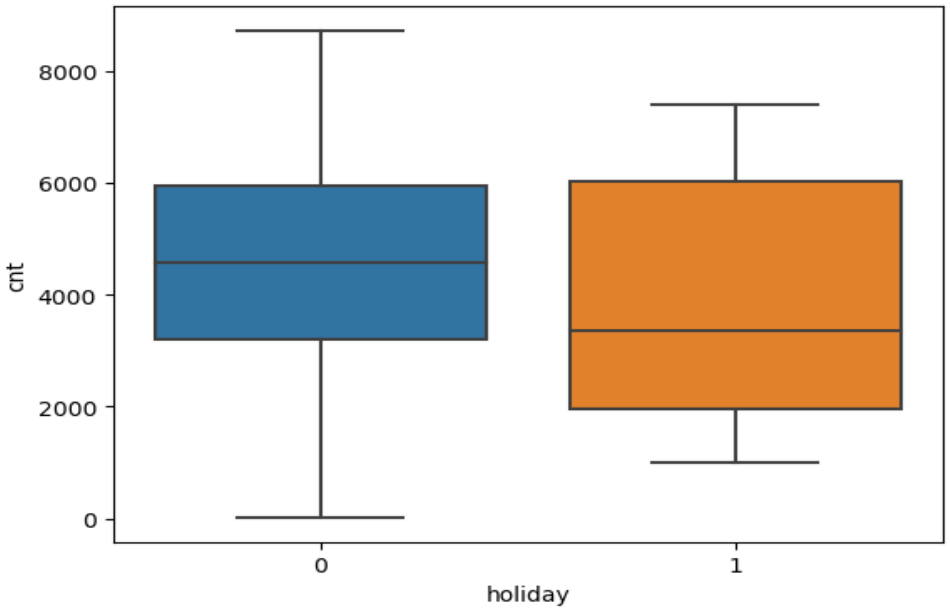
2. I observe that there has been an increase in demand for the upcoming year.



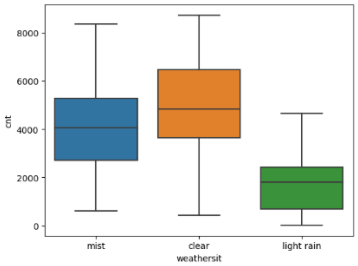
3. The demand exhibits continuous growth each month until June, with September experiencing the highest demand. Following September, there is a decline in demand.



4. On holidays, there is a decrease in demand.



5. Clear weather conditions are associated with the highest demand.

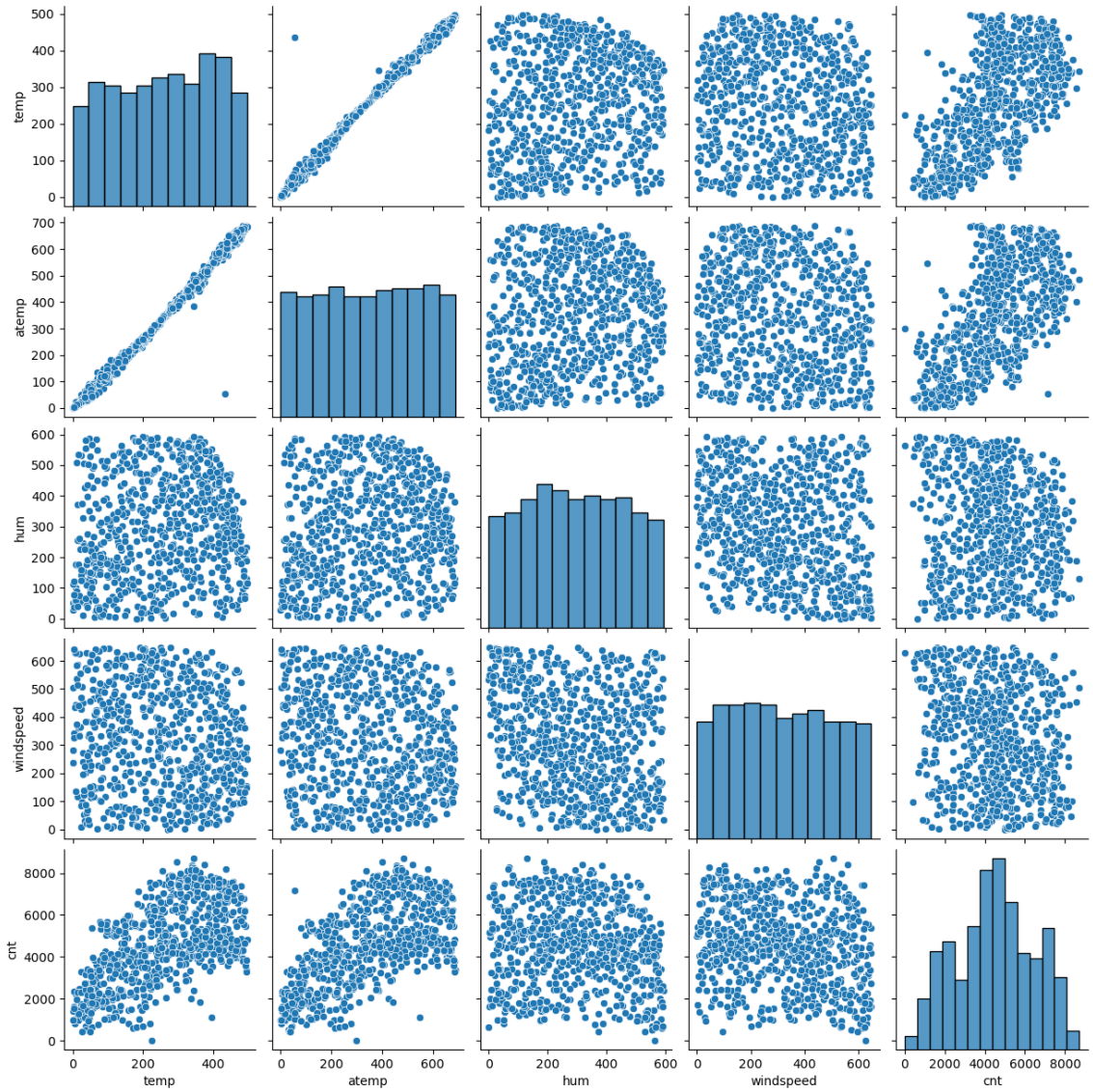


Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans: Using drop_first=True is crucial as it helps eliminate the redundant column generated during dummy variable creation, thereby mitigating correlations among the dummy variables. Failing to drop one of the dummy variables derived from a categorical variable introduces redundancy into the dataset, resulting in the presence of a constant variable (intercept), which, in turn, can lead to multicollinearity issues.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: The features "temp" and "atemp" exhibit the strongest correlation and demonstrate a clear linear relationship with the target variable "cnt."



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: I have verified the following assumptions:

- The error terms follow a normal distribution with a mean of 0.
- There are no discernible patterns in the error terms.
- Multicollinearity has been assessed using VIFs (Variable Inflation Factors).
- Linearity has been examined.
- Overfitting has been addressed by evaluating the R-squared (R^2) value and Adjusted R-squared (R^2) value.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Features “Year”, “temp” and month “sept” are highly related with target column, so these are top contributing features in model building.