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)

Based on the Bivariate analysis between “cnt”(Rental nikes count) and the categorical variable.

For Season variables the summer and fall seasons have higher effect on the dependent variable “cnt”. There is a higher number of “cnt” in the fall season.

For the Year variable bike rental count is significant higher on 2019 when compared to 2018

There is an increase in the bike rental count from Jan - June and after Sep the count declined

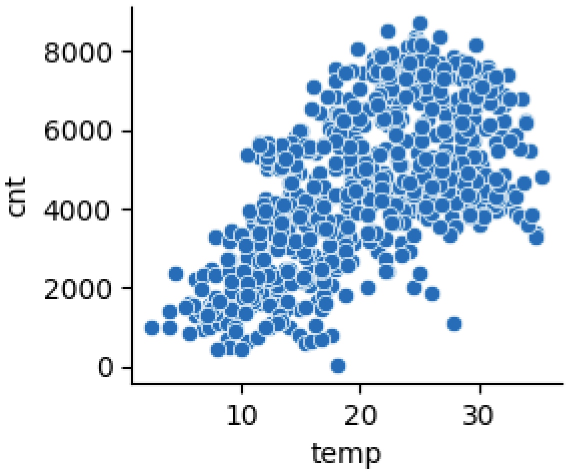
Higher bike rentals count in the clear weather when compared to Cloudy or Rainy season

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

The dummy variable is used to convert the categorical variable to as many 0/1 variables as there are different values. After converting to a dummy variable we can remove one column since it is represented by all other variables and it is redundant. The ‘*drop_first=True*’ is available as an argument for the pandas get_dummies function so we can use that directly.

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

The temp(temperature in Celsius) have the highest correlation



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

Linear relationship

The Relationship between the dependent variable and independent variable should be linear. We have to remove the outliers as a part of data preparation/cleaning. We can visualize the linear relationship by using the scatter plot.

Multivariate normality

The linear regression requires all variables to be multivariate normal

No or little multicollinearity

Multicollinearity exists when the independent variables are highly correlated between each other. We can use VIF to identify the multicollinearity and remove the variables one at a time which have high VIF(greater than 5).

No auto-correlation

Auto correlation occurs when the residuals are not independent of each other. We can use a scatter plot to check the auto-correlation.

Homoscedasticity

The disturbance between the independent variables and the dependent variable, noise, is the same across all the independent variables.

We can use the residual analysis to identify the Homoscedasticity

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

temp, windspeed, season_summer are the top three features.

General Subjective Questions

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

The Linear regression algorithm is used to identify linear relation between an independent variable and multiple dependent variable.

$y = mx + b$

m is the slope and b is the intercept

This relation helps to predict the dependent variable for the new independent variables.

Cost Function

The cost function of the linear regression is calculated as the average of squared error between y_pred and y

The cost function is used to calculate the optimum value for m and b with minimal cost function value.

Evaluation Metrics for Linear Regression

The strength of any linear regression model can be assessed using various evaluation metrics

R-Squared - Indicates the variation the model can capture

RSME - Describes how well the observed data points match the expected values.

Advantages

Linear regression is relative simple algorithm and it easy to implement

Linear regression can be used for real time applications because it can be implemented quickly and it is computationally efficient

It is observed that the outliers have minimal impact on the model performance.

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

The Anscombe’s quartet is a group of four data sets that are nearly identical in simple descriptive statistics.

It tells the importance of visualization of data before applying the algorithms to build the models. We need to identify various anomalies in the data like outliers.

3. What is Pearson’s R? (3 marks)

The Pearson correlation coefficient is for measuring a linear correlation.
The strength and direction of the correlation is measured by the number -1 and 1.

If the value is greater than 0 then the correlation is positive.
If the value is 0 or near 0 then there is no correlation.
If the value is less than 0 then negative correlation.

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

Scaling is an important part of the preprocessing step. We apply scaling to normalize the data within a particular range.

The data of each column can be of different units because of that the values within a row differ a lot, for example:- age and distance traveled in meters.

There are two methods to scale the data: Normalization and Standardization.

Normalization :- To brings all the data in the range of 0 and 1

$x = (x - min(x)\,) \,/\, (max(x) - min(x))$

Standardization :- Brings all the data into a standard normal distribution which have mean 0 and standard deviation 1

$x = (x - mean(x)/sd(x))$

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

The VIF is calculated using the formula

$$VIF = \frac{1}{1-Ri2}$$

If the two values are highly correlated the R will be approximately equal to 1. This leads to VIF becoming infinite.

To avoid such multicollinearity we should remove one of the columns.

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

The Q-Q plot is the graphical method to determine the two samples of data came from the same population or not.

By using the Q-Q plot we can determine the two samples have the same distribution.
The Linear regression performs better when features follow a normal distribution. The Q-Q plot helps to find that the residuals follow a normal distribution. Having a normal error is an assumption in the regression. The Q-Q plot helps to verify that assumption