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?

Answer:

Season (season), Month(mnth), Weekday(weekday), Weathersit(weathersit), Holiday(holiday), Workingday(workingday) and Year(yr) are the categorical variables in the dataset. Box plot and barplot were used for analyzing. Their effect on dependent variable can be inferred as follows:

- **Season (season):** Fall season would be good for business expansion as it attracted more bookings. However we can reserve the Spring season for bike maintenance as it has a low booking record.
- **Month(mnth):** September has the highest number of bookings. From the mid year i,e in the months of may, jun, jul, aug, sep and oct most the bookings were done.The bookings increase from start ot the year till september then gradually decreases.
- **Weekday(weekday):** Thursday, Friday, Saturday and Sunday have more bookings as compared to the start of the week. So we can infer that weekends have more bookings than weekdays
- **Weathersit(weathersit):** Clear weather brings in more bookings whereas thunderstorm, rains or other harsh weather does not attract bookings which indicates weather forecast consideration is important.
- **Holiday(holiday):** Bookings are less on holidays.
- **Workingday(workingday):** Be it working or non-working day the bookings don't seem to vary much.
- **Year(yr):** The demand for bikes has increased in the year 2019 as compared to 2018. Which indicates progress in business.

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

Answer:

drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

If we have categorical variables with n-levels, then we need to use n-1 columns to represent the dummy variables.

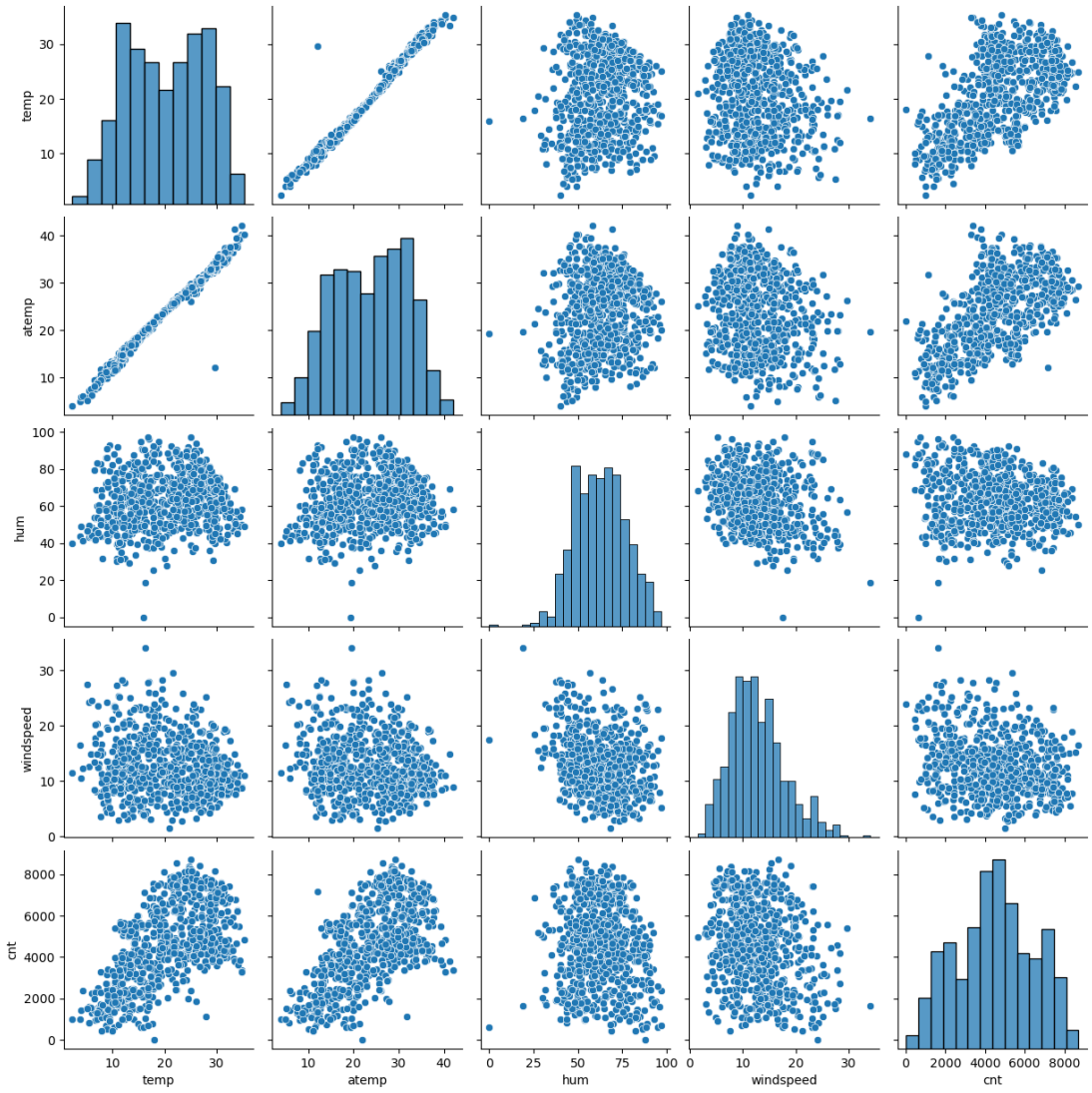
For example if we have 3 types of values in a Categorical column and we want to create a dummy variable for that column. If one variable is not either A orB, then it is obvious that it is C. So we do not actually need the 3rd variable to identify the C. A= 0 and B= 0 implies that it is C.

Categorical Variable	A	B
Dummy value for A	1	0
Dummy value for B	0	1
Dummy value for C	0	0

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

Answer:

'temp' and 'atemp' have the highest correlation with the target (cnt) variable.

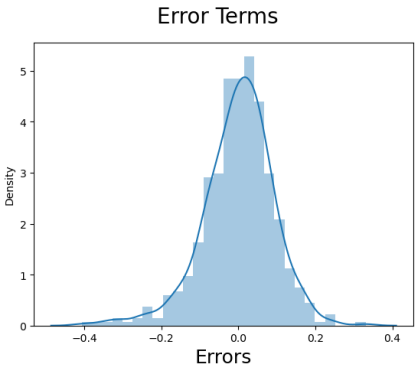


4.How did you validate the assumptions of Linear Regression after building the model on the training set?

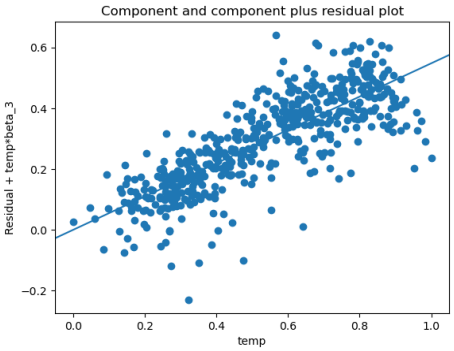
Answer:

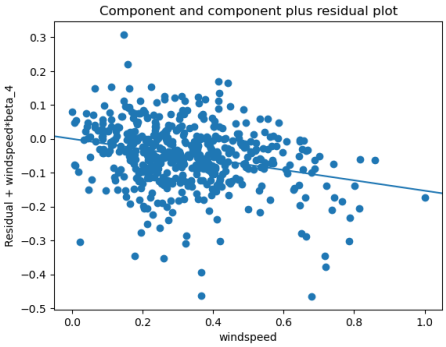
I validated the following assumptions of Linear Regression after building the model (lr_7) on the training set:

- Normality of Errors- Error terms should be normally distributed

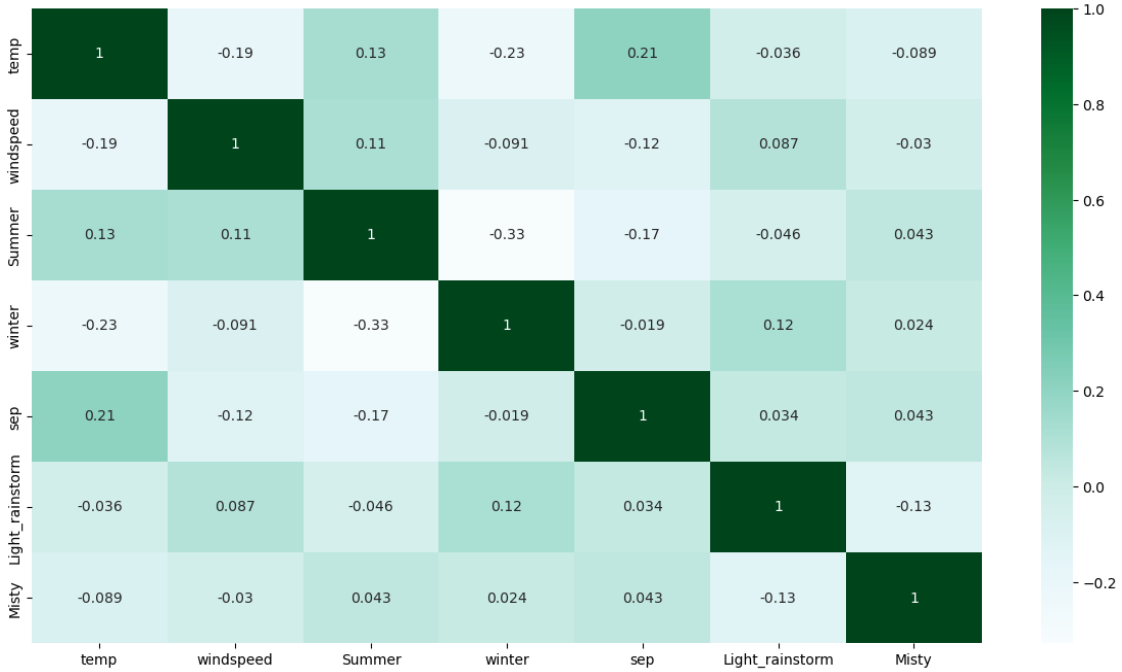


- Linear Relationship - Linearity should be visible among numeric variables

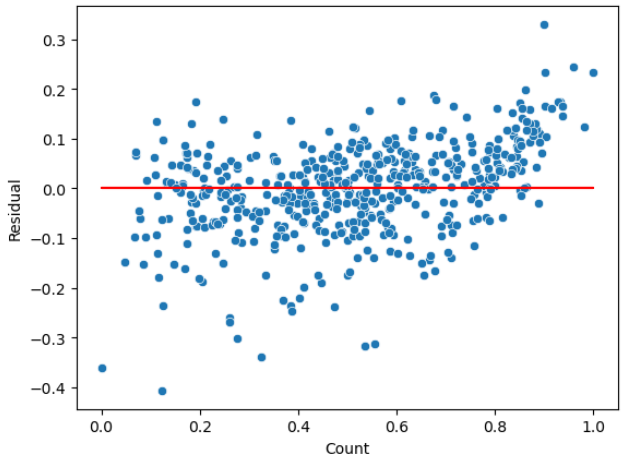




- Absence of Multicollinearity - There should be insignificant multicollinearity among variables.The heatmap shows there is no Multicollinearity as all the values are very low.



- Homoscedasticity - There should be no visible pattern in residual values.



- Independence of residuals - No auto-correlation. 2.097 durbin watson value indicates that there is no first order autocorrelation in the model.

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

Answer:

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Temperature (temp): temp has coefficient 0.548008. One unit increase in temperature increases the bike rentals by 0.548008 units.
2. Year (yr): Year has highest coefficient (0.232861). This indicates One unit increase in yr parameter increases the bike rentals by 0.232861.
3. Weather Situation (weathersit): The coefficient of Light_rainstorm is -0.282869 that means poor weather conditions lead to lower bike rentals.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Answer:

Linear regression algorithm is a supervised algorithm that computes the relationship between a dependent (target) variable and one or more independent features. It is used to find the best linear equation that can predict the value of the dependent variable based on the independent variables.

The Linear relationship between variables means that when the value of one or more independent variables changes, the value of the dependent variable also changes accordingly . It is given by the equation:

$$Y = mX + c$$

Where:

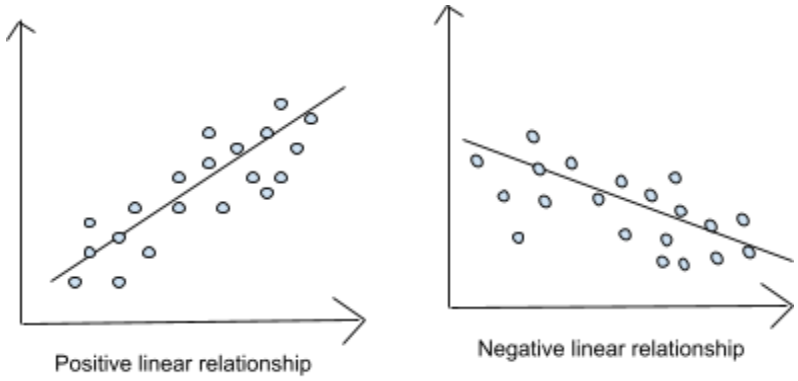
Y: Variable to predict.

X: Independent variable used to make predictions.

m: Slope the regression line which represents the effect X has on Y
c: Y-intercept.

The linear relationship can be positive or negative in nature:

1. Positive Linear Relationship: A linear relationship is called positive if both independent and dependent variables increase with respect to each other.
2. Negative Linear relationship: A linear relationship will be called negative if independent increases and dependent variable decreases.



Linear regression is of two types:

A. Simple Linear Regression: Used when the relationship is to be established between one dependent and one independent variable.

B. Multiple Linear Regression: Used when the relationship is to be established between one dependent and more than one independent variable(s).

A Linear Regression model is said to be valid if it does not violate the following assumptions:

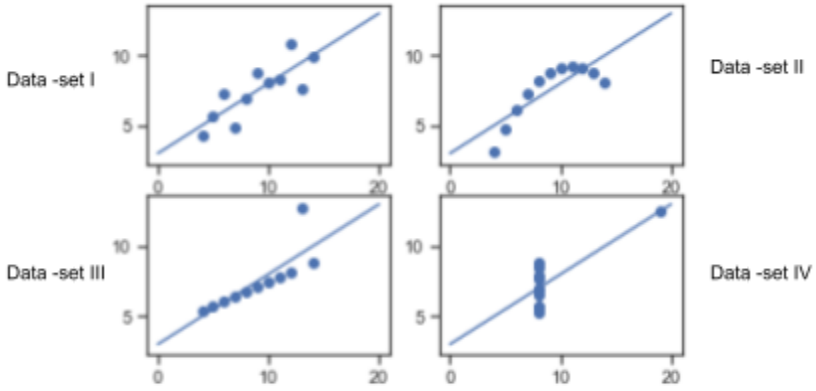
- Normality of Errors- Error terms should be normally distributed
- Linear Relationship - There should be a linear and additive relationship between dependent and independent variable(s). A linear relationship suggest that there a change in response Y (dependent variable) is due to one unit change in X (independent variable(s))
- Absence of Multicollinearity - The independent variables should not be correlated.
- Homoscedasticity - There should be no visible pattern in residual values. The error terms must have constant variance.
- Independence of residuals - No auto-correlation. There should be no correlation between the residual (error) terms. The Durbin Watson (DW) value should lie between 0 - 4.

2.Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the

regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It comprises four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.



Graphical representation of Anscombe's quartet

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.

Anscombe's Quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset

3.What is Pearson's R?

Answer:

Pearson R or Pearson's Correlation coefficients are used to measure how strong a relationship is between two variables. It is given by the formula:

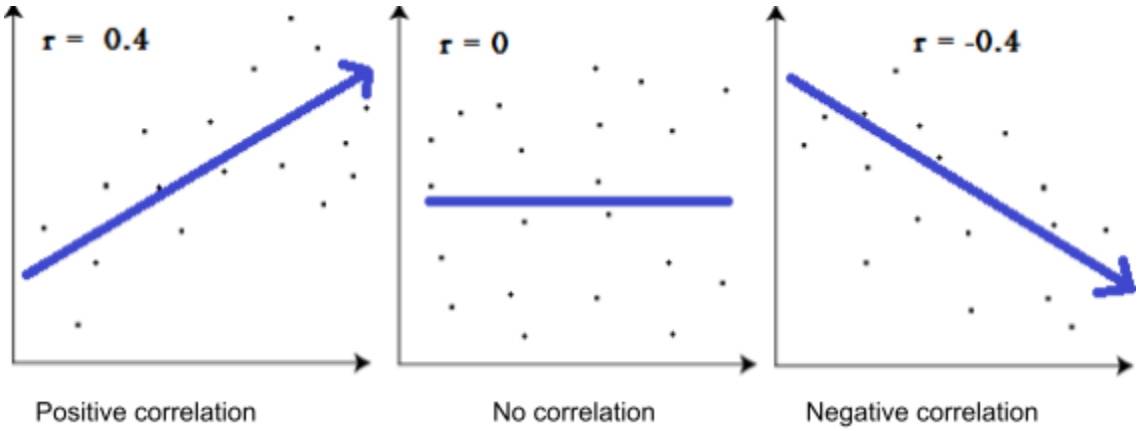
$$R= n(\sum xy) – (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$$

The Correlation coefficients ranges from - 1 to 1 where:

1 indicates strong positive relationship

-1 indicates strong negative relationship

0 indicates no relationship



The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example: |-75| = 75 has a stronger relationship than 65.

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

Answer:

a. Scaling is the process of normalizing the data within a particular range.
b. Since the data can have a large number of variables in different ranges, it is not feasible to compare them. Hence they need to be brought to a particular range/scale where their comparison becomes possible.
c. Scaling can be done using normalized scaling and standardized scaling:
    1. **Normalized scaling:** Normalized scaling is also known as min-max scaling. It Scales values between 0-1 and gets affected by outliers. The data is uniformly distributed across 0-1 (range of scaling). It is calculated by the formula:
    $$Normalization\ /\ MinMax\ scaling\ (x) = x - min(x) / max(x) - min(x)$$
    Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

    2. **Standardization:** Standardization brings all the data into a standard normal distribution with mean 0 and standard deviation 1.It is much less affected by outliers. It is not range bound. It is calculated by the formula:
    $$Standardization\ (x) = x - mean(x)/sd(x)$$
    Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer:

VIF = infinity means perfect correlation. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which yields:

VIF = 1/(1-R2) = 1/(1-1) = 1/0= infinity

To solve this issue we need to drop one of the variables from the dataset which is having perfect correlation.

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

Answer:

Q-Q plots are useful for checking whether a dataset follows a certain theoretical distribution, such as a normal distribution or a log-normal distribution. If the points on the Q-Q plot fall on a straight line, it indicates that the two datasets have the same distribution. If the points deviate from the straight line, it suggests that the two datasets do not have the same distribution. The degree and direction of deviation from the straight line can provide insights into the nature of the difference between the two datasets.

Following are the types of distribution that can occur in Q-Q plots:

1. **S shaped deviation:** An S-shaped deviation occurs when the points on the Q-Q plot form a curve that resembles the letter “S”. This indicates that one dataset has a heavier tail than the other dataset.
2. **J shaped deviation:** A J-shaped deviation occurs when the points on the Q-Q plot form a curve that resembles the letter “J”. This indicates that one dataset has a higher median than the other dataset.
3. **U shaped deviation:** A U-shaped deviation occurs when the points on the Q-Q plot form a curve that resembles the letter “U”. This indicates that one dataset has a lower median than the other dataset.

Advantages of Q-Q plots:

a. They are not affected by differences in sample size or scale, as long as the datasets have the same number of observations.
b. Also useful for identifying outliers or extreme values in a dataset.
c. They provide a clear and intuitive visual representation of how two datasets are compared in terms of their distributions.