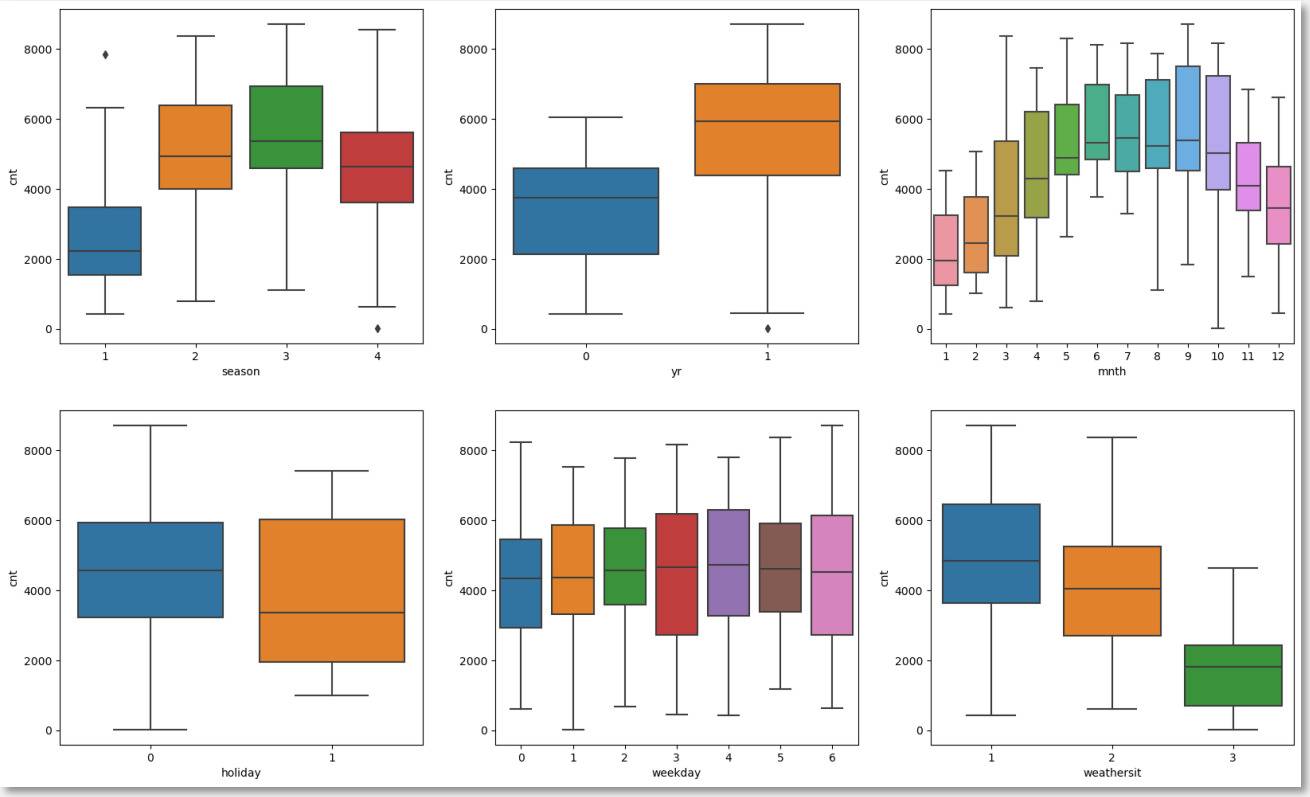
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   In total, we have 6 categorical variables in the dataset: season, yr, mnth, holiday, weekday and weathersit.

   Looking at the boxplots shown below we have below inferences: -

   - Out of the 4 seasons, fall has the most demand, followed by summer, winter and spring.
   - Looking at the demand in years 2018 and 2019, we can infer that the business has grown much higher in 2019 which is a very positive sign for boombikes.
   - Looking at the months, it is clearly visible that the demand of the rental bikes is higher during the mid months and it is probably due to the season changes.
   - For holiday, it is clearly evident that on a non-holiday the demand is more. The median is much higher for non-holiday compared to holiday.
   - For weekday, the median of demand is almost same irrespective of day of the week.
   - It is quite interesting to see prominent patterns in weathersit. There is absolutely no bike rental demand for category 4 days (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog). For other categories, the demand is in decreasing order of 1(Clear, Few clouds, partly cloudy, partly cloudy), 2(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) and 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds).



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

   pd.get_dummies is used to convert all the categorical values to dummy variables(columns). However, to optimize the number of columns, we need only N-1. If the value of all columns is False then it represents the dropped columns. It avoids multicollinearity.

Here is an example for the assignment. We can see that season has 4 values and when we create dummy variables for it with drop_first=True option, it does not create a dummy column for “fall”. When the value of other three dummy columns (summer, spring, winter) are False then it represents the value “fall”.

```
#convert season to categrical dummy variables
df['season'] = df['season'].map({1: 'spring', 2: 'summer', 3: 'fall', 4: 'winter'})
df.season.value_counts()
```

✓ 0.0s

```
season
fall      188
summer    184
spring    180
winter    178
Name: count, dtype: int64
```

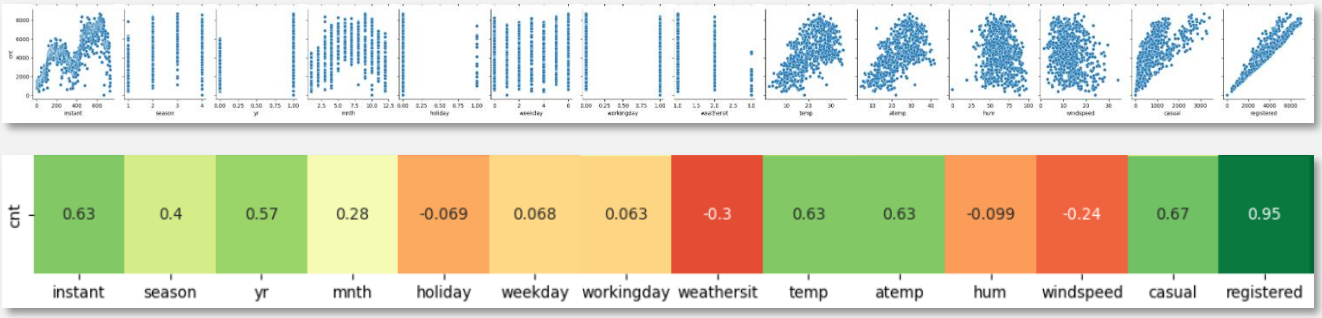
```
#create dummy features for season and drop first
season = pd.get_dummies(df['season'], drop_first = True)
season
```

✓ 0.0s

	spring	summer	winter
0	True	False	False
1	True	False	False
2	True	False	False
3	True	False	False
4	True	False	False

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

   Type of the users(registered, casual) has the highest correlation with cnt, followed by temperature.

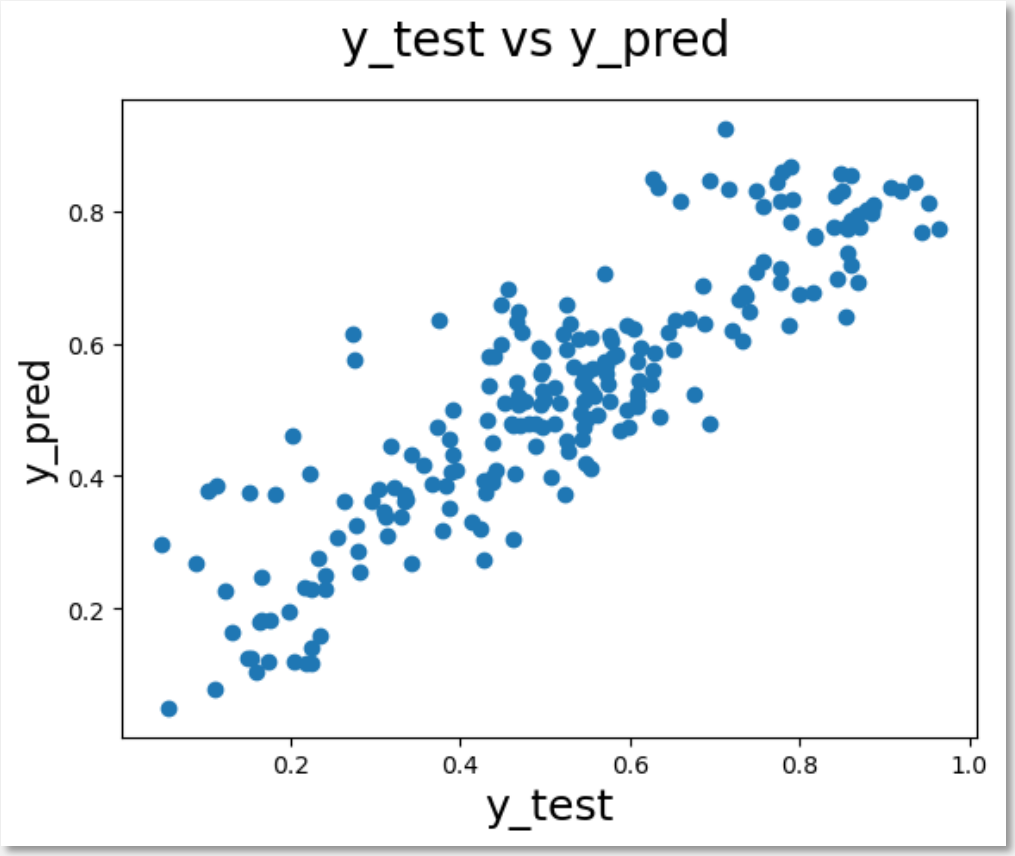


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

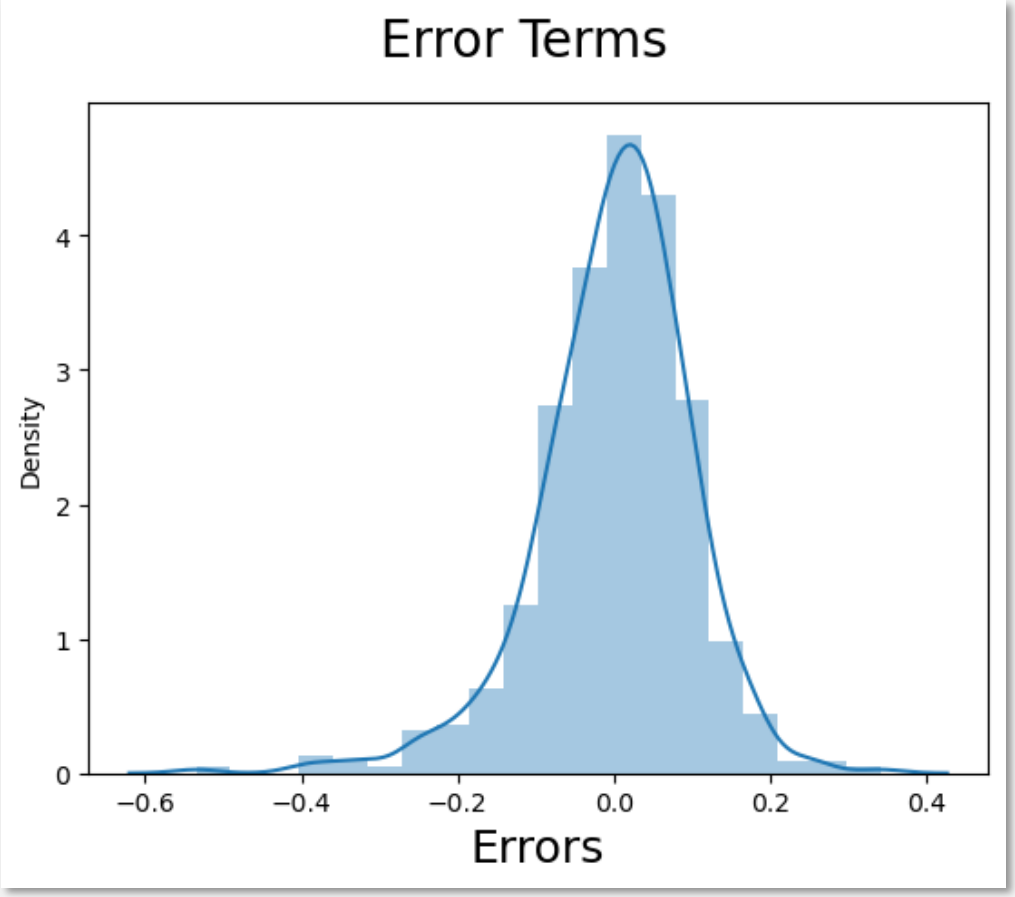
   There are 4 assumption of linear regression model as follows: -

   a. There is a linear relationship between X and Y

Validation: Upon plotting the actual vs. predicted values, the relationship appears as a straight line.

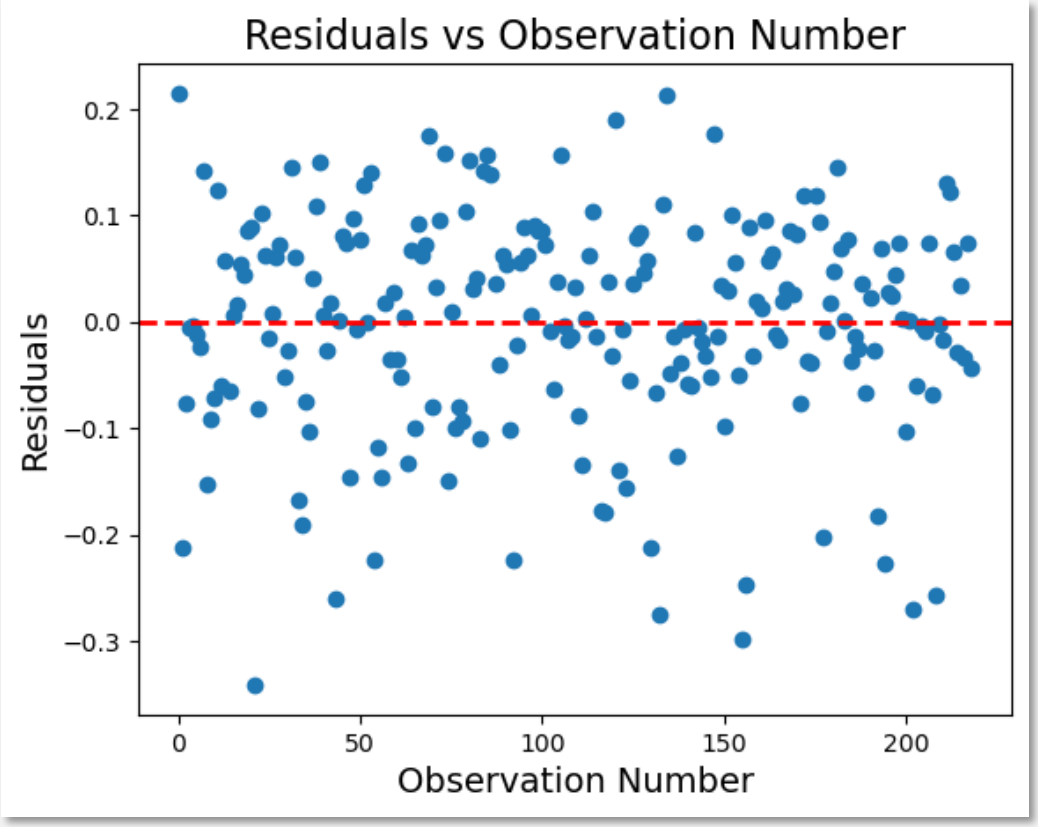


b. Error terms are normally distributed with mean zero (not X, Y)
Validation: Upon plotting the residuals, it seems well distributed across the axis and is not random.



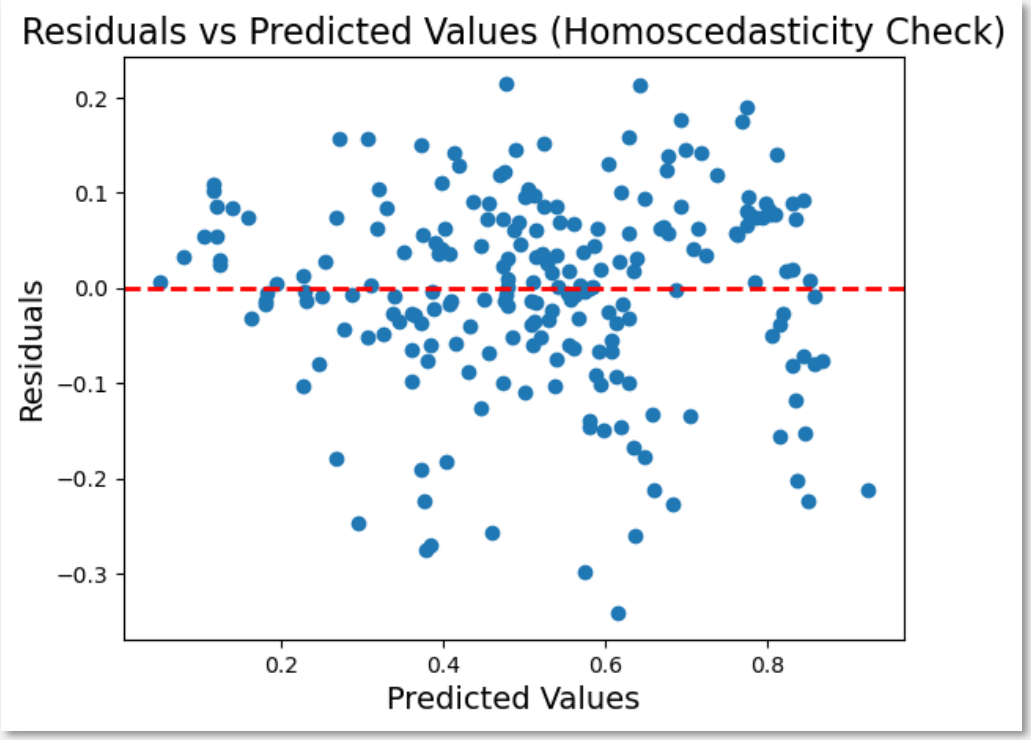
c. Error terms are independent of each other

Validation: Upon plotting scatter plot for residuals (error terms) against the number of observations we can see that the residuals (errors) from the regression model does not exhibit any systematic pattern or correlation. Hence, we can say that the error terms are independent of each other.



d. Error terms have constant variance (homoscedasticity)

Validation: Upon plotting residuals against predicted values. The spread of residuals is found to be roughly constant.



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

   The equation for the best fitted line as per the final model is as below:
   count of users = 0.037724 + 0.236765 x yr + 0.547294 x temp - 0.177205 x windspeed + 0.090213 x summer + 0.120757 x winter + 0.093519 x clear&fewcloud + 0.094501 x Sep
   So, we can say that the top 3 features contributing significantly towards explaining the demand are – temperature(temp), year(yr) and winter.

General Subjective Questions

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

   Linear regression is a machine learning technique. It analyses past/historic data to establish relationship/pattern between a set of input and one output, such that it can be represented by a formula y=mx+c, this can be used to predict future values of output based on new inputs.

   The goal is to find the best-fitting line that minimizes the sum of the squared differences between the observed and predicted values.

   Assumptions of linear regression are: -
   - There is a linear relationship between X and Y
   - Error terms are normally distributed with mean zero (not X, Y)
   - Error terms are independent of each other
   - Error terms have constant variance (homoscedasticity)
   -

   Limitations of linear regression are: -
   - Linear regression may not perform well when the relationship between the variables is highly non-linear.
   - It is sensitive to outliers, which can disproportionately influence the model.
   - The assumptions need to be carefully validated for the model to be reliable.

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

   Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

   The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

   Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

   In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear

correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.
The Pearson correlation coefficient is a dimensionless value that ranges from -1 to 1:

- r=1: Perfect positive linear correlation
- r=−1: Perfect negative linear correlation
- r=0: No linear correlation

The formula for Pearson's correlation coefficient between two variables X and Y with sample size n is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Here,

- X_i and Y_i are the individual data points.
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Scaling is the process of transforming data to a standard scale, ensuring that all variables or features have the same scale or magnitude. The purpose of scaling is to bring the numerical values of different features into a comparable range, preventing certain features from dominating others. Scaling is particularly important in machine learning algorithms that rely on distances or gradients, such as k-nearest neighbours, support vector machines, and gradient-based optimization algorithms.
   It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
   There are two major methods to scale the variables, i.e. standardisation and MinMax scaling. Standardisation basically brings all the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:

   - Standardisation: $x = \frac{x-mean(x)}{sd(x)}$
   - MinMax Scaling: $x = \frac{x-min(x)}{max(x)-min(x)}$

   Here is an example of minmax scaling from the assignment.

Before Scaling

```
df_train.head()
```

✓ 0.0s

	yr	holiday	workingday	temp	hum	windspeed	cnt	spring	summer	winter	...	May	Nov	Oct	Sep	Mon	Sat	Sun	Thu	Tue	Wed
653	1	0	1	19.201653	55.8333	12.208807	7534	0	0	1	...	0	0	1	0	0	0	0	1	0	0
576	1	0	1	29.246653	70.4167	11.083475	7216	0	0	0	...	0	0	0	0	0	0	0	1	0	0
426	1	0	0	16.980847	62.1250	10.792293	4066	1	0	0	...	0	0	0	0	1	0	0	0	0	0
728	1	0	0	10.489153	48.3333	23.500518	1796	1	0	0	...	0	0	0	0	0	0	0	0	1	0
482	1	0	0	15.443347	48.9583	8.708325	4220	0	1	0	...	0	0	0	0	1	0	0	0	0	0

5 rows × 29 columns

Scaling the Features using minmax scaler(normalised between 0 to 1)

```
#Instantiate an object
scaler = MinMaxScaler()

# Fit on data - Apply scaler() to only numeric variables
num_vars = ['temp', 'hum', 'windspeed', 'cnt']
df_train[num_vars] = scaler.fit_transform(df_train[num_vars]) #transform and fit the scaler on trainig dataset
df_train.head()
```

✓ 0.0s

	yr	holiday	workingday	temp	hum	windspeed	cnt	spring	summer	winter	...	May	Nov	Oct	Sep	Mon	Sat	Sun	Thu	Tue	Wed
653	1	0	1	0.509887	0.575354	0.300794	0.864243	0	0	1	...	0	0	1	0	0	0	0	1	0	0
576	1	0	1	0.815169	0.725633	0.264686	0.827658	0	0	0	...	0	0	0	0	0	0	0	1	0	0
426	1	0	0	0.442393	0.640189	0.255342	0.465255	1	0	0	...	0	0	0	0	1	0	0	0	0	0
728	1	0	0	0.245101	0.498067	0.663106	0.204096	1	0	0	...	0	0	0	0	0	0	0	0	1	0
482	1	0	0	0.395666	0.504508	0.188475	0.482973	0	1	0	...	0	0	0	0	1	0	0	0	0	0

5 rows × 29 columns

We can see that the values of the 4 columns got scaled between 0 and 1.

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

VIF calculates how well one independent variable is explained by all the other independent variables combined. VIF measures how well a predictor variable can be predicted using all other predictor variables. The VIF is given by:

$$VIF_i = \frac{1}{1-R_i^2}$$

where '*i*' refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. There are prominently two reasons for infinite values of VIF: -

- Perfect Linear Relationship: The presence of a perfect linear relationship between the i-th predictor and the other predictors in the model leads to an infinite VIF.
- Redundant Information: When one predictor can be exactly predicted by a linear combination of the other predictors, it introduces redundant information, making the model ill-posed.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.