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

NOTE: We can infer that the below points:

1) FALL -Season has a good demand for the bikes followed by Summer & Winter.

2) Year 2019 has very good amount of booking as compared to previous year.

3) The demand for bikes is increasing every month till June. from June to September the demand is steady having September month as highest demand and then there is a decline in demand for bikes in last three months.

4) Holiday has a increased demand for bikes.

5) Clear weather has the highest demand for bikes.

6) Weekday do not show any significant trend on the demand of the bikes. They may or may not be a good feature.

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

We should use drop_first=True because while creating dummy columns for n levels of data ., the actual columns created are n. the n levels can also be well explained by n-1 columns. so dropping the first column reduces the extra column and also reduces the correlations created amoung dummy variables.

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

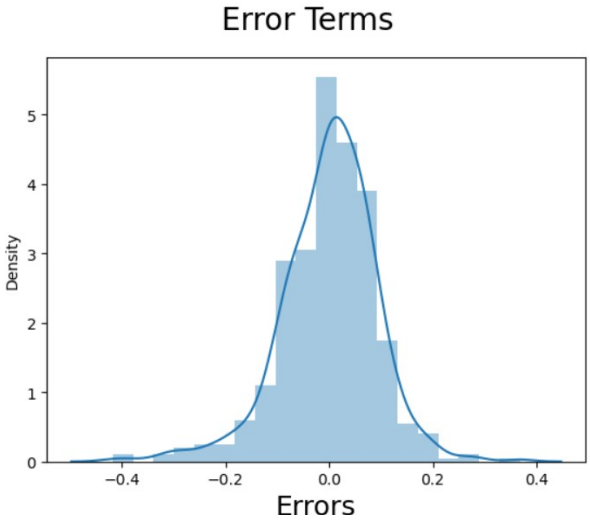
temp & atemp has the highest correlation with the target variable cnt. (0.63)

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

These are the assumptions for linear regression:

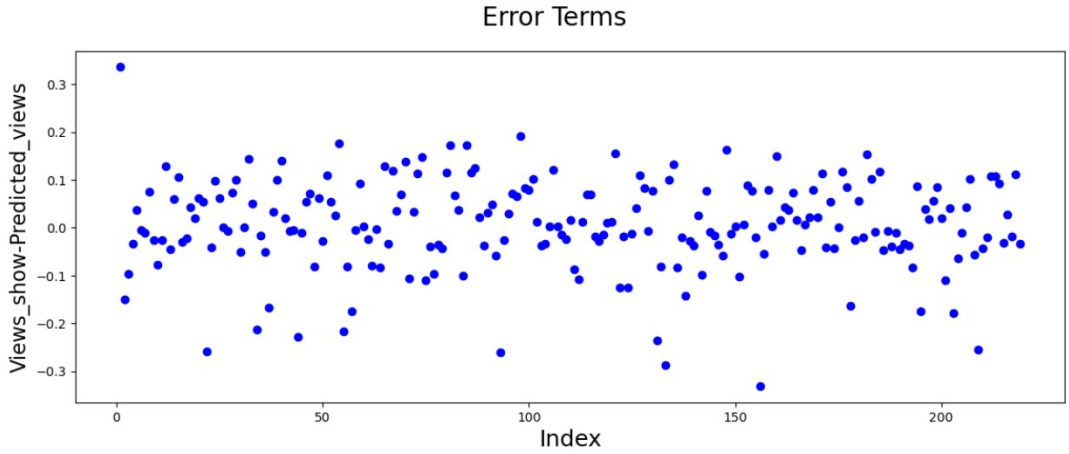
1. Linear relationship between X and Y

2. Error terms are normally distributed (not X, Y)



3. Error terms are independent of each other

4. Error terms have constant variance (homoscedasticity).



As we can see the error terms are randomly distributed and there is no pattern which means the output is explained well by the model and there are no other parameters that can explain the model better

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

1) Temperature (atemp) - A coefficient value of 0.46 indicates that a unit increase in atemp variable increases the bike demand by 0.46 units.

2) Weather Situation 3 (Light snow & rain) - A coefficient value of -0.28 indicates that, a unit increase in Weathersit 3 variable decreases the bike demand by 0.28 units.

3) Year (yr) - A coefficient value of 0.235 indicates that a unit increase in yr variable

increases the bike demand by 0.235 units.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form, i.e., numerical value.

Linear regression can be expressed mathematically as:

$y= \beta 0 + \beta 1x$

Here, Y= Dependent Variable

X= Independent Variable

β 0= intercept of the line

$\beta 1$ = Linear regression coefficient (slope of the line)

The relation between independent and dependent variables is a straight line with a slope.

The best fit line is selected among those minimizing the sum of the squares of those (observed) values that differ from the (predicted) values, which is known as the least squares formulation. This will lead to the line being as aligned to the data points as possible by minimizing any type of error occurring.

Q2. 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.This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data.

Q3. What is Pearson's R?

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

The formula for Perason's r :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the v-variable

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

Answer: Feature Scaling also known as data normalization is a data preprocessing step. It is a method used to normalize the range of independent variables or features of data. Independent variables are of different types. The numerical ones can be of type age(0-100

years), salary(in the range of 1000s), dimensions(decimal points), and many more. We don't want our machine learning model to confuse a feature with a larger magnitude as a better one. Feature scaling in Machine Learning would help all the independent variables to be in the same range, for example- centered around a particular number(0) or in the range (0,1), depending on the scaling technique.

Scaling is Imortant beacsue the ML model assign weights to the independent variables according to their data points and conclusions for output. In that case, if the difference between the data points is high, the model will need to provide more significant weight to the farther points, and in the final results, the model with a large weight value assigned to undeserving features is often unstable. This means the model can produce poor results or can perform poorly during learning.

Techniques to Perform Feature Scaling : **Min-Max Normalization (or Min-Max scaling)**

This technique rescales a feature or observation value with a distribution value between 0 and 1.

Formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, x is the feature, and min() and max() are the minima and maximum values of the feature, respectively.

Standardization

This technique rescales a feature value so that it has a distribution with 0 mean value and variance equal to 1.

Formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector.

Normalization transforms your data into a range between 0 and 1.

Standardization transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

Normalization/standardization are designed to achieve a similar goal, which is to create features that have similar ranges to each other. We want that so we can be sure we are capturing the true information in a feature, and that we dont over weigh a particular feature

just because its values are much larger than other features.

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

Answer: VIF is a measure of how well a predictor variable is correlated with all other variables, excluding target variable. So high VIF typically greater that 5 is considered a matter of concern. If there is perfect correlation, then VIF is close to infinity.

So VIF as infinity shows a perfect correlation. In the case of perfect correlation r2=1, which leads to 1/(1-r2) as infinity.

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

Answer 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.