

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)

Answer:

Dependent Variable (cnt) Demand

Demands increases in workingday, summer, winter, Saturday, August, September

Demand decreases in Light snow, Mist cloudy, December, February, January, November

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

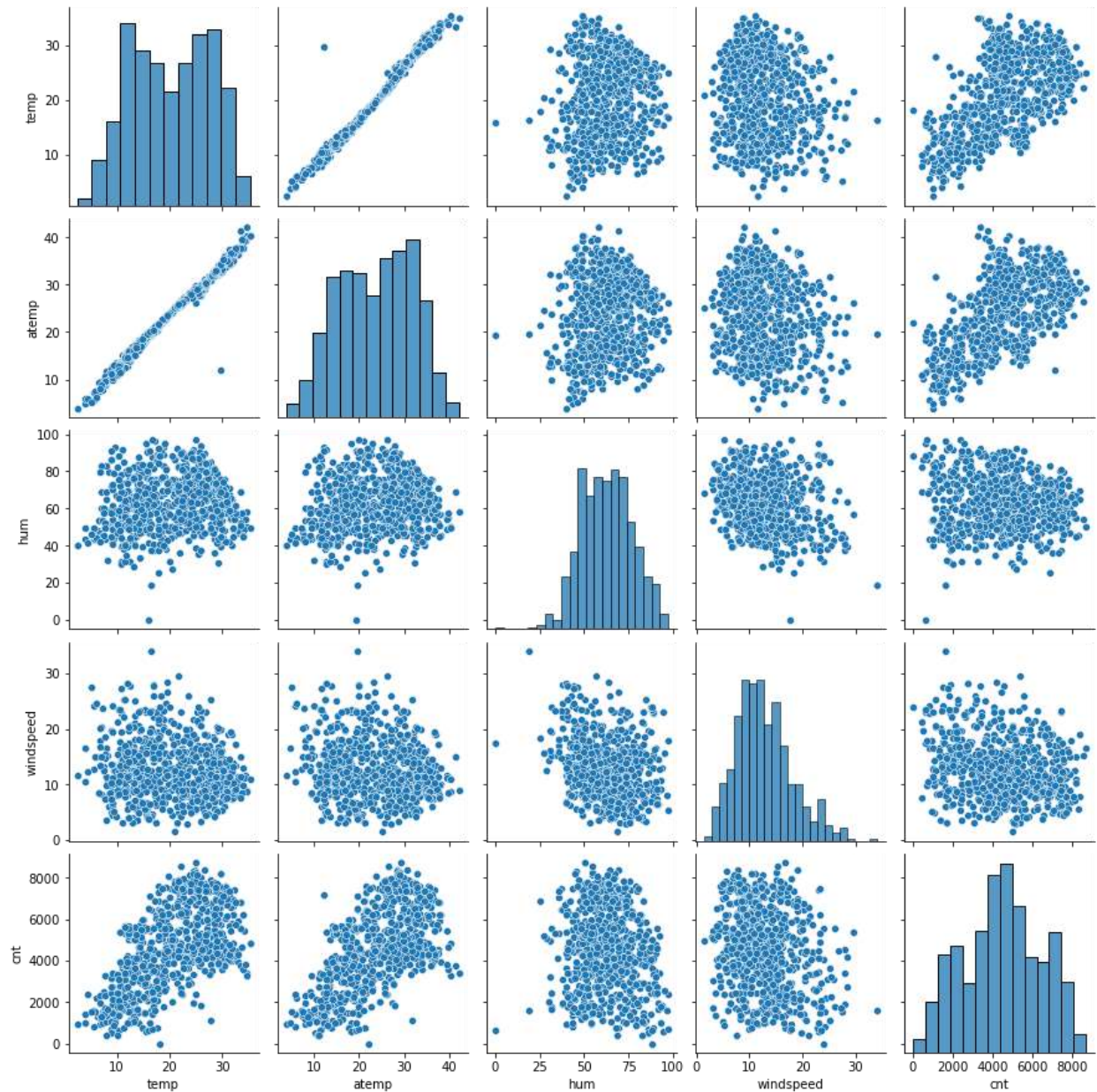
Answer:

When we have a categorical variable with K mutually exclusive categories, we actually only need K – 1 new dummy variables to encode the same information. This is because if all of the existing dummy variables equal 0, then we know that the value should be 1 for the remaining dummy variable.

So for example, if `region_North == 0`, and `region_South == 0`, and `region_West == 0`, then `region_East` must equal 1. This is implied by the existing 3 dummy variables, so we don't need the 4th. The extra dummy variable literally contains redundant information. So, it's a common convention to drop the dummy variable for the first level of the categorical variable that you're encoding.

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

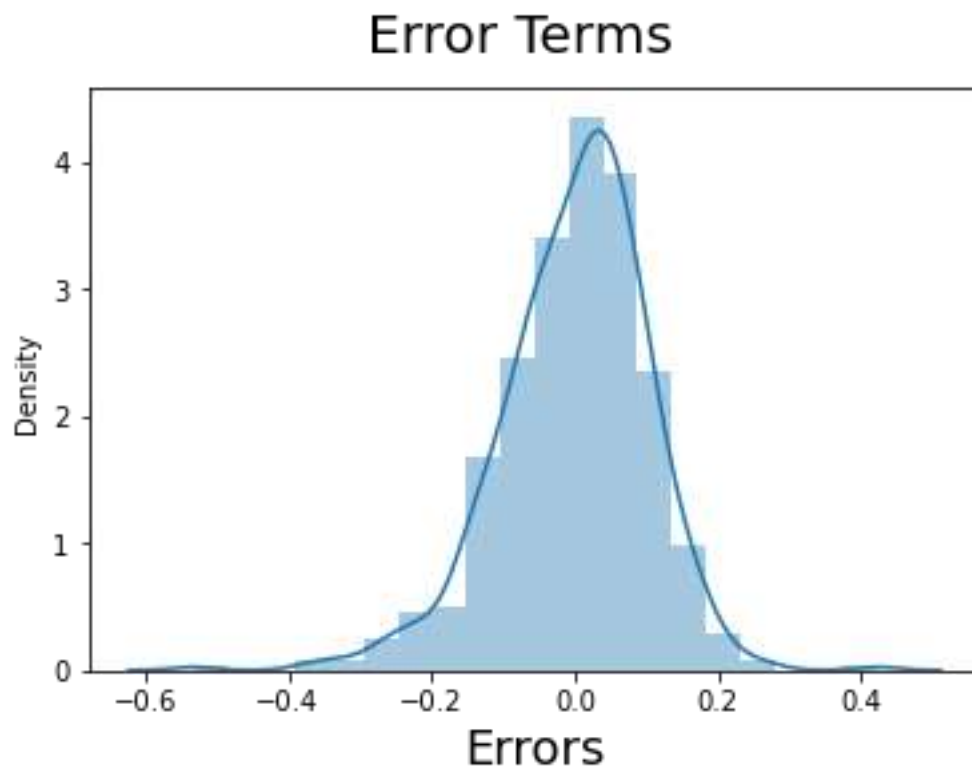
Answer:



From the figure, 'temp' and 'atemp' numerical variable have the highest correlation with target variable 'cnt' of 0.63 and 0.63 respectively

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

Answer:



From the above histogram plot, since the error terms are also normally distributed, which proves, the major assumptions of linear regression

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

top features explaining cnt are:

- | | |
|----------------------------------|---------|
| 1. weather situation(Light snow) | -0.3149 |
| 2. year (0: 2018, 1:2019) | 0.2462 |
| 3. windspeed | -0.2370 |

General Subjective Questions

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

Answer:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(\text{slope}) = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$a(\text{intercept}) = \frac{n \sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analysed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

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

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r , the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1 .

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

$r > 0 < 0.5$ means there is a weak association

$r > 0.5 < 0.8$ means there is a moderate association

$r > 0.8$ means there is a strong association

Pearson r Formula

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

Here,

- 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 y-variable

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

Answer:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

If there is perfect correlation, then $VIF = \infty$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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.

Few advantages:

- a) It can be used with sample sizes also
- b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

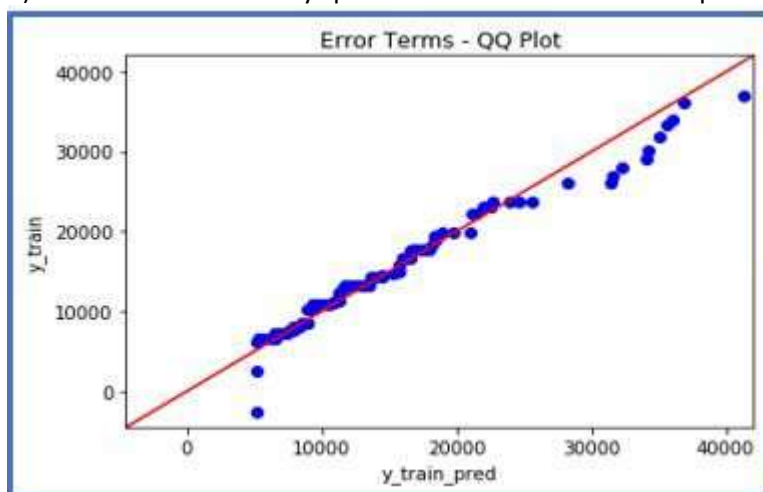
- i. come from populations with a common distribution
- ii. Have common location and scale
- iii. Have similar distributional shapes
- iv. Have similar tail behaviour

Interpretation:

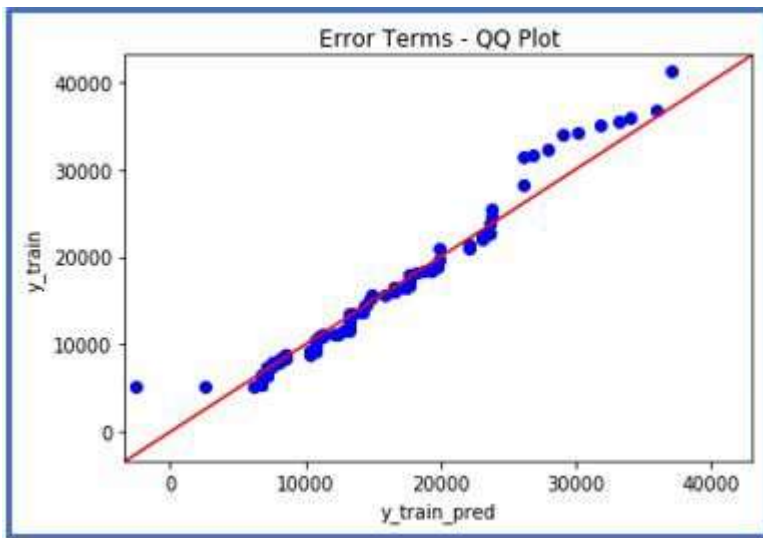
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis