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:
I have done analysis on categorical columns using the boxplot. Below are the few points we can infer from the visualization –
• Season: Fall has highest demand for rental bikes
• Every month the demand for rental bike is increasing till June. September has the highest number of demand then it is decreasing.
• During weekdays and workingdays demands don't have that much of variation.
• Clear weathersit has highest demand
• When there is a holiday, demand has decreased.

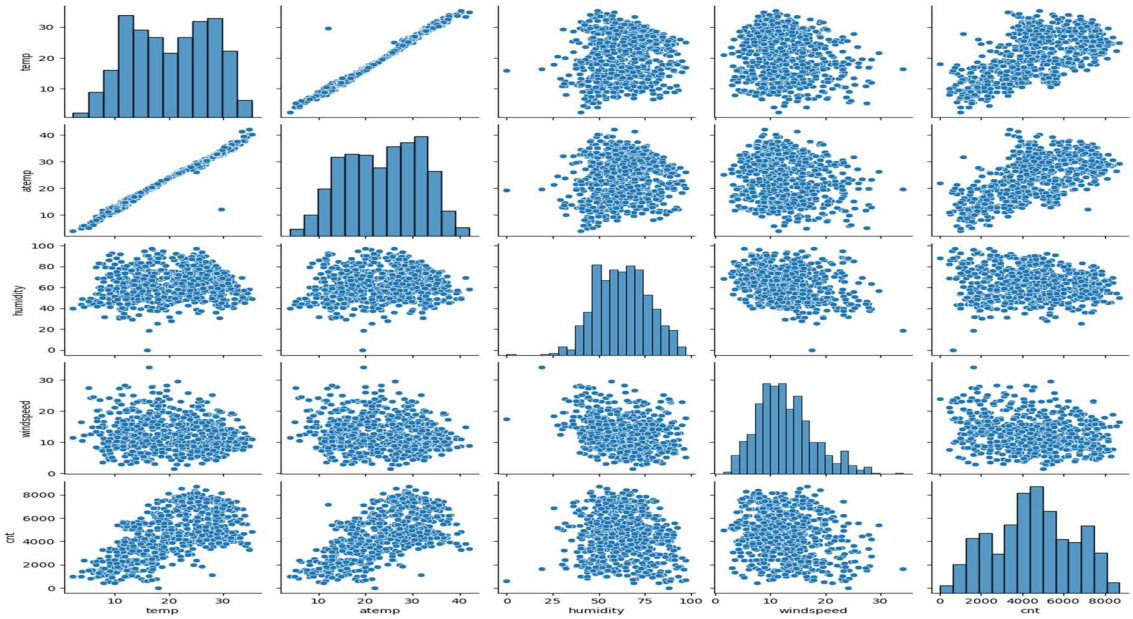
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Answer:
drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

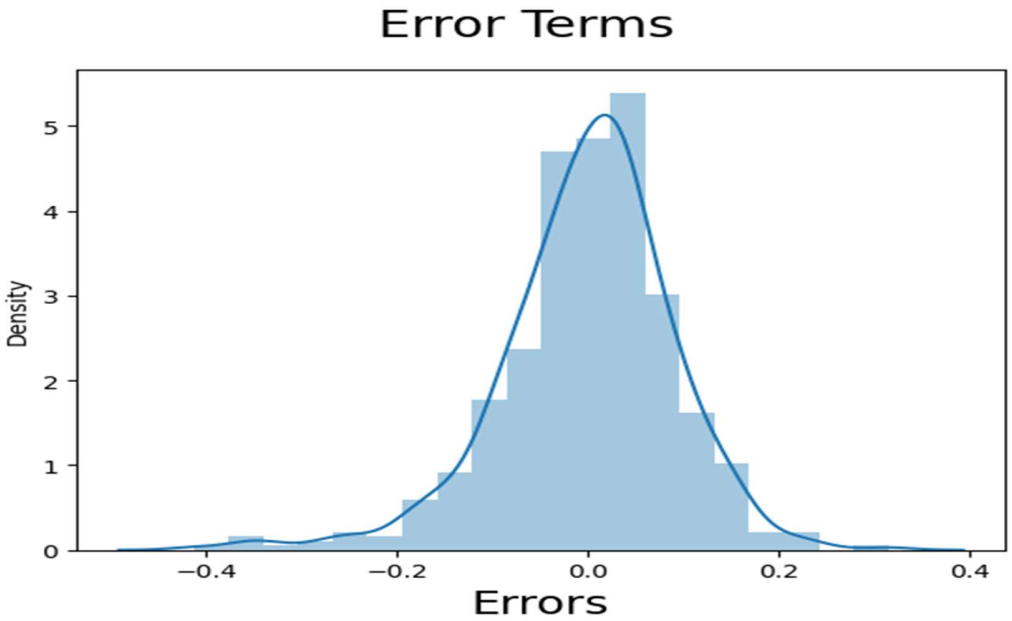
Answer:



From above, we can say that temp and atemp has highest correlation with the target variable 'cnt'. And it is also observed that temp and atemp are highly correlated with each other.

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

Answer:



The distribution of residuals should be normal and centred around 0. We test this assumptions of residuals by producing a distplot of residuals to see if they follow normal distribution or not.
In above diagram residuals are scattered around mean =0 as seen in the diagram above.

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

Answer:

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

temp , season_winter, month_sep.

General Subjective Questions

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

Answer:

Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.

The algorithm uses the best fitting line to map the association between independent variables with dependent variable.

There are 2 types of linear regression algorithms

a Simple Linear Regression – Single independent variable is used.
   i $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
b Multiple Linear Regression – Multiple independent variables are used.
   i $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p +$ ∈ is the line equation for MLR.
c $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
d $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$

Cost functions – The cost functions helps to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained**.

e Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
   i The straight-line equation is $Y = \beta_0 + \beta_1 X$
   ii The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as Y_i.
   iii $Now\ the\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum (y_i - \beta_1 x_i - \beta_0)^2$
f The unconstrained minimization are solved using 2 methods

i Closed form
ii Gradient descent

While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.

g $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
h OLS is used to minimize the total e^2 which is called as Residual sum of squares.
i $\text{RSS} = = \sum_{i=1}^{n} (y_i - y_{pred})^2$

Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

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

Answer:

Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great to for describing the general trends and aspects of the data.
Francis Anscombe realized in 1973 that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to illustrate the fact.

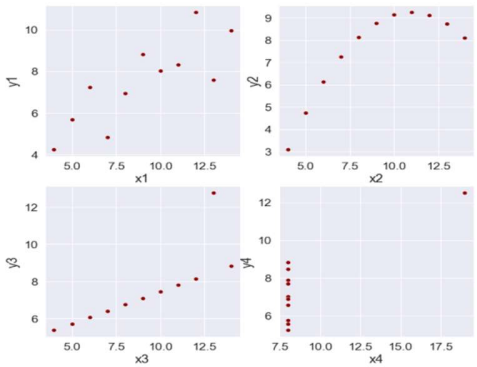
- Illustrations
  - One of the data set is as follows:

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.040000	9.140000	7.460000	6.580000
1	8	8	8	8	6.950000	8.140000	6.770000	5.760000
2	13	13	13	8	7.580000	8.740000	12.740000	7.710000
3	9	9	9	8	8.810000	8.770000	7.110000	8.840000
4	11	11	11	8	8.330000	9.260000	7.810000	8.470000
5	14	14	14	8	9.960000	8.100000	8.840000	7.040000
6	6	6	6	8	7.240000	6.130000	6.080000	5.250000
7	4	4	4	19	4.260000	3.100000	5.390000	12.500000
8	12	12	12	8	10.840000	9.130000	8.150000	5.560000
9	7	7	7	8	4.820000	7.260000	6.420000	7.910000
10	5	5	5	8	5.680000	4.740000	5.730000	6.890000

If the descriptive statistics are checked for above data set then all looks same:

	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	9.000000	9.000000	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.000000	4.000000	8.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.500000	6.500000	8.000000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	9.000000	9.000000	8.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	11.500000	11.500000	8.000000	8.570000	8.950000	7.980000	8.190000
max	14.000000	14.000000	14.000000	19.000000	10.840000	9.260000	12.740000	12.500000

However, when plotted these points, the relation looks completely different as depicted below.



- Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
- The dangers of outliers in data sets are warned by the quartet. Check the bottom 2 graphs. If those outliers would have not been there the descriptive stats would have been completely different in that case.
- Important points
    - Plotting the data is very important and a good practice before analysing the data.
    - Outliers should be removed while analysing the data.
    - Descriptive statistics do not fully depict the data set in its entirety.

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

Answer:

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:

- *-1 coefficient indicates strong inversely proportional relationship.*
- *0 coefficient indicates no relationship.*
- *1 coefficient indicates strong proportional relationship.*

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
$Σx^2$ = the sum of squared x scores
$Σy^2$ = the sum of squared y scores

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

Answer:

- What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
- Why – Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance.
  The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.
- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

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

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

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

Answer:

$$VIF = \frac{1}{1 - R^2}$$

The VIF formula clearly signifies when the VIF will be infinite. If the R^2 is 1 then the VIF is infinite. The reason for R^2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

Answer:

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

- Interpretations
    - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
    - Y values < X values: If y-values quantiles are lower than x-values quantiles.
    - X values < Y values: If x-values quantiles are lower than y-values quantiles.
    - Different distributions – If all the data points are lying away from the straight line.
- Advantages
    - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
    - The plot has a provision to mention the sample size as well.