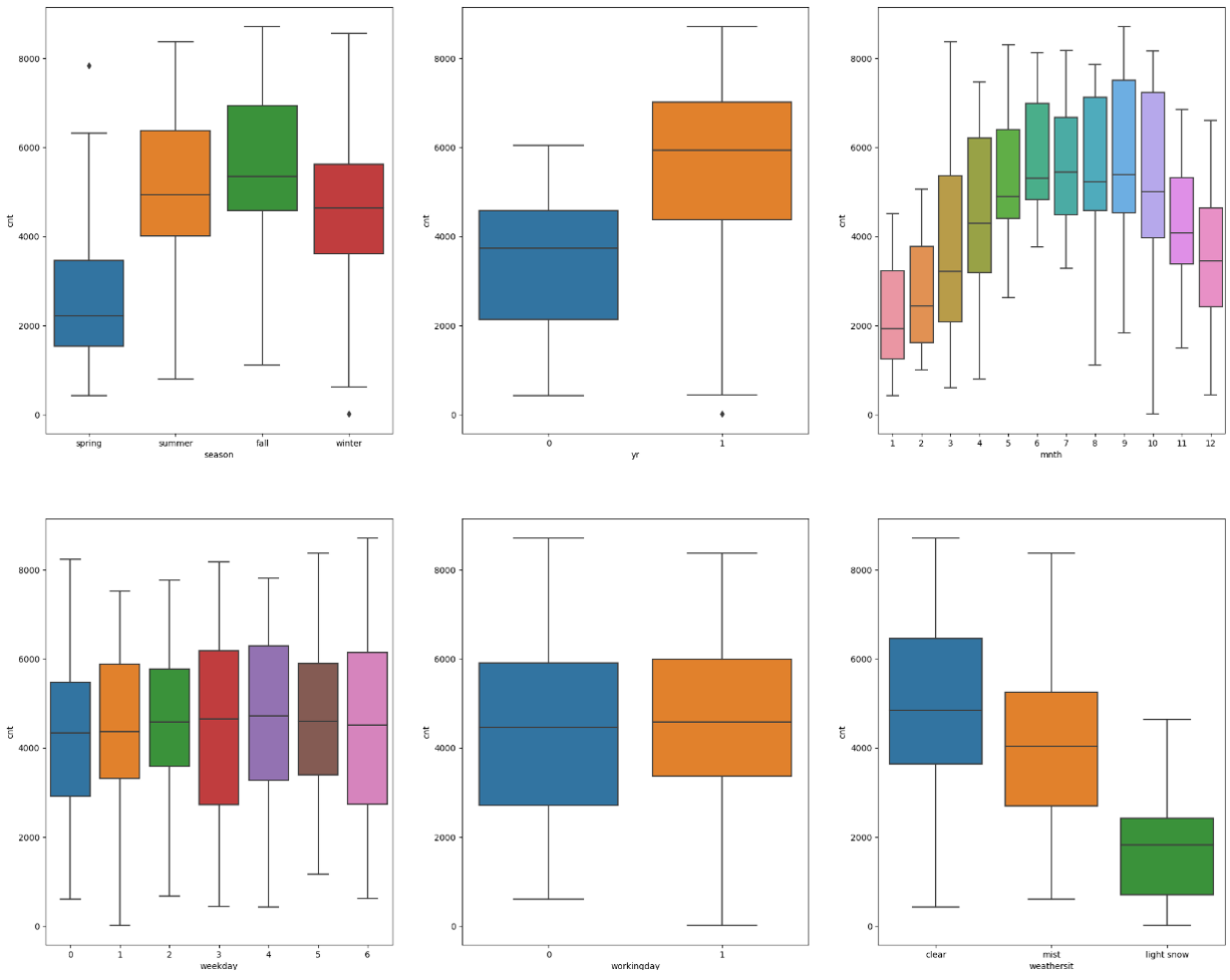
Assignment-based Subjective

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:

   Here the dependent variable we are trying to solve for is ‘cnt’. Variables likes Season, yr, month, weathersit and holiday have a strong effect on the dependent variable. Weekday and holiday don’t have a strong effect on the dependent variable. This can be inferred from seeing the plots for the variable:



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

Answer:

We use drop_first=True during dummy variable creation because the extra column is not needed and is redundant. By using drop_first=True, you essentially eliminate one of the dummy variables for each categorical feature, leaving only k-1 columns for a variable with k categories. This helps avoid perfect multicollinearity because you don't have a situation where the values in one column can be perfectly predicted from the values in the other columns. This can be verified from the below example from the assignment. We have four values in Season column, Spring, Summer, Fall and Winter On converting it to dummy variable:

	Keeping first Column					Removing First Column		
Season	**Spring**	**Summer**	**Fall**	**Winter**		**Summer**	**Fall**	**Winter**
Spring	1	0	0	0		0	0	0
Summer	0	1	0	0		1	0	0
Fall	0	0	1			0	1	0
Winter	0	0	0	1		0	0	1

The season column is converted to four columns when drop_first=True is not used. The season column is converted to three columns when drop_first=True is used. Checking for the rows in season column with the columns when first column is dropped Spring is marked as 0 0 0 Summer as 1 0 0, fall as 0 1 0 and winter as 0 0 1 When all the three columns have 0 value here, it is spring. Hence, the first column is redundant.

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

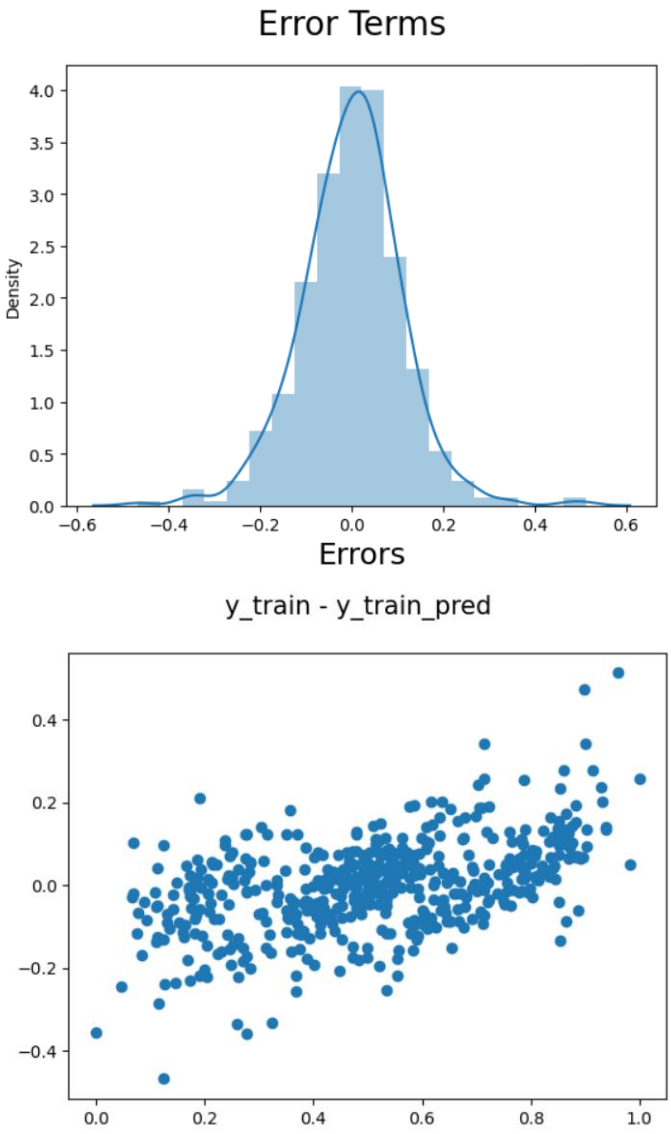
Answer:

Considering the overall dataset, ‘registered’ variable will have the highest correlation with the ‘cnt’ target variable.

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

Answer:

- Residual errors follow a normal distribution
- Mean for the error is centered at 0
- Variance of Errors doesn't follow any trends



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

Answers:

Top three feature contributing significantly to demand of shared bike are:

Var	Coeff
- Quarter_JulAugSep	0.2817 (this is a derived variable of mnth)
- light snow	-0.2680
- yr	0.2440

General Subjective Questions

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

Answer:

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The basic idea behind linear regression is to find the best-fitting straight li ne (or hyperplane in the case of multiple independent variables) that minimizes the sum of the squared differe nces between the observed and predicted values. It is represented as below:

Y= bx + C

Y is the estimated dependent variable
b is the regression coefficient
C is constant
x is the independent variable

The above equation tries to predict the value of Y depending upon the value of x. Linear regression is used to determine the value of b and C in this case, how does a unit change in X effect Y and what will be the value of Y if X is zero
For Multiple linear regression, the equation is updated as

Y = b1x1 + b2x2 + b3x3…… + C
Where b1, b2 are coefficient of x1,x2… and so on

In multiple linear regression, we have multiple independent variable and we try to find how change in one independent variable effects the dependent variable.
For a linear regression following assumptions are made:

1. There is linear relation between dependent and independent variables
2. The error terms are normally distributed
3. Mean for error term is 0
4 Assumption of Homoscedasticity The variance of the residuals (the differences between observed and predic ted values) should be constant across all levels of the independent variable(s).
5. Assumption of Independence: The residuals should be independent of each other, meaning there should be no pattern in the residuals.

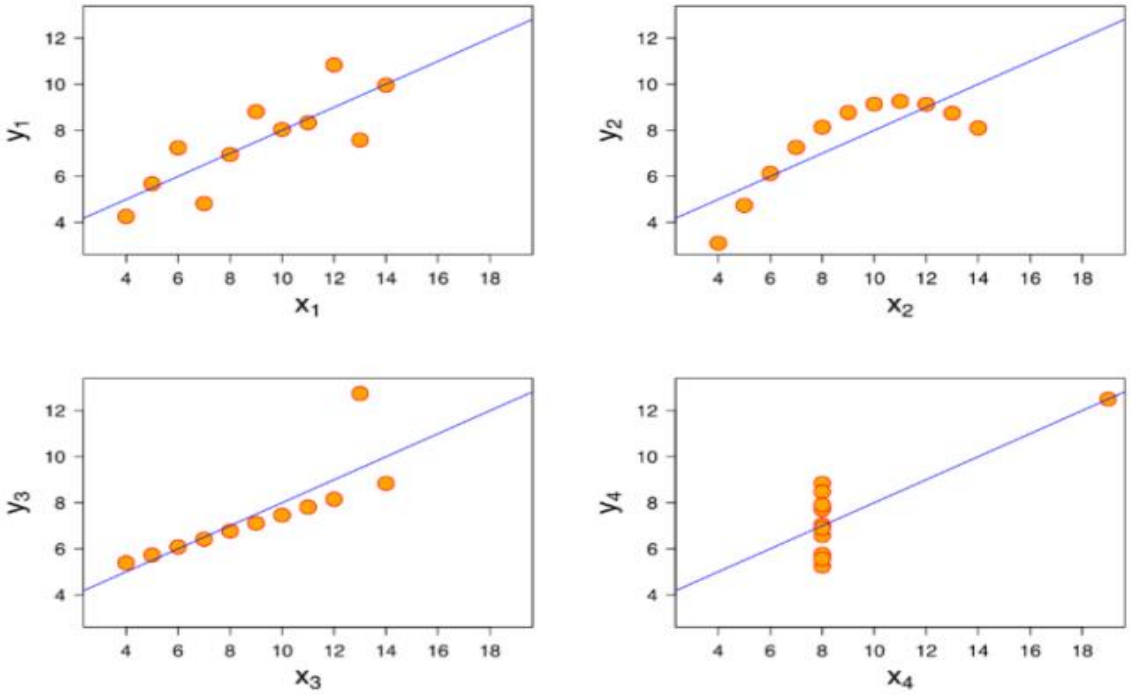
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet developed by a statistician Francis Anscombe highlights the importance of visually analyzing t he data. He created four datasets with nearly identical statistics but appear completely different when graphed. Th e idea was to stress the importance of visual analysis and counter the impression "numerical calculations are exact, but graphs are rough." Looking at the below table, we can observe, sum, average and standard deviation of all the four dataset are almost exactly same.
Looking at the below table, we can observe, sum, average and standard deviation of all the four datasets is almost exactly same.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats



Graph and table courtesy: https://medium.datadriveninvestor.com/anscombes-quartet-12649db7eac0?gi=a66b864878aa

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

Answer:

Pearson's R or Pearson product-moment correlation coefficient measures linear correlation between two variables X and Y. It is a measure of strength of the relation between two variables and their association with each other and explains the effect on one variable when other variable changes. Its value ranges from +1 to -1, +1 being total positive linear correlation, 0 meaning no linear correlation and -1 meaning total negative linear correlation. A positive correlation means that when X increases Y also increases whereas a negative correlation means when X decreases Y also decreases. It is represented as ρ for population and r for sample.

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

Where, N = number of pairs of data X and Y are the variables.

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

Answer:

Scaling is a method used to normalize/standardize the range or features of the data. Range of data for different variable may vary widely and that is the reason it is suggested to do scaling in the data pre-processing step when using a machine learning algorithm. When applying a machine learning algorithm, say linear regression, the gradient descent will take iterations to fit the line. If we have two variables whose range vary widely, gradient descent will be able to work it out in lesser number of iteration while will need a larger number of iteration for variable with a larger range. Hence, scaling is applied to bring down the cost function gradient descent.

Normalization is a scaling method in which the values are rescaled in such a way that they end up between 0 and 1. This is also known as Min-Max Scaling

$$X' = X- Xmin / Xmax – Xmin$$

Xmax = Maximum value of X
Xmin = Minimum value of X

Normalization is helpful in cases where data follows a Gaussian distribution. It subsidizes the effect of outliers as it has a bounding range.
Standardization is a method in which we rescale the value to be centred around mean with a unit standard deviation. Mean of the attributes become 0 and the distribution has a unit standard deviation.

$$X'= X-\mu/\sigma$$

μ is the mean of the feature variable
σ is the standard deviation of the feature variable

Standardization can be helpful for cases where data doesn't follow Gaussian distribution. It doesn't take care of ou tliers as it doesn't have a bounding range

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

Answer:

VIF or variance inflation factor is the measure of correlation between one predictor variable with rest of predictors in the model. It explains how well can a predictor variable be explained with the help of other predictor variables.

An infinite VIF indicates that the variable can be expressed exactly by a linear combination of other variables which show infinite VIF as well. This is due to a perfect correlation between the two variables, because of which we get R 2=1 and VIF = 1/1-R2 becomes infinity. To solve this issue, we need to drop one of the variable from the dataset w hich is causing multicollinearity.

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

Answer;

A Q-Q plot of quantile quantile plot is a probability plot to compare two probability distributions by plotting quantil e of one against another. It is used to compare properties such as scale, location and skewness is similar or differen t in the two datasets. When creating a linear regression model, we assume that the errors are normally distributed with mean 0. Also, that the errors are independent and Homoscedasticity. A QQ plot is plotted between Y_actual a nd Y_Pred to check and verify the assumptions made for linear regression.