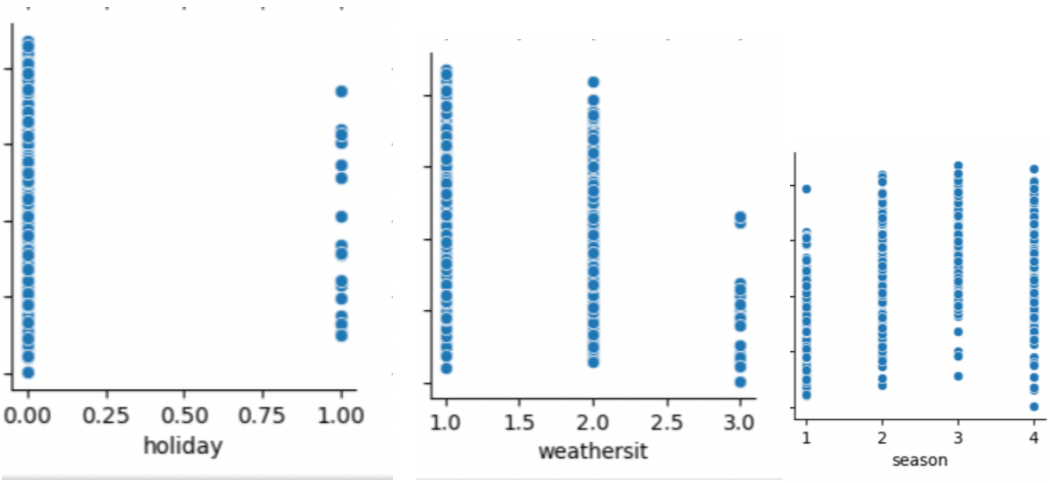
Name: Bharat Hegde

Executive PG Programme in Machine Learning & AI - January 2023

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)**



- **Holiday:** When there is holiday, commuters seem to use the Bike sharing service less.
- **Weather Situation:** When it is lightly rainy, commuters are clearly using the bike sharing service lesser.
- **Season:** In the spring, commuters are using the bike sharing lesser compared to other season. In the winter the distribution seems to have large variance.

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

If there is a categorical variable with N levels, then N-1 levels are sufficient to represent the information, otherwise the information will be redundant.

Example: We have the following dummy variables for Weather situation.

(Clear, Misty, LightRainyorSnow)
Note: for HeavyRainyOrSnow – there were no values. Hence not used.

100 – means Clear
010 – means Misty
001 – means Light Rain or snow

The variable Clear is redundant as it could be represented by (Misty, LightRainyorSnow) and when the variables are (00) it would mean that weather is clear.

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

The variable 'temp' seems to be highly correlated with the target variable (with correlation = 0. 63)

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

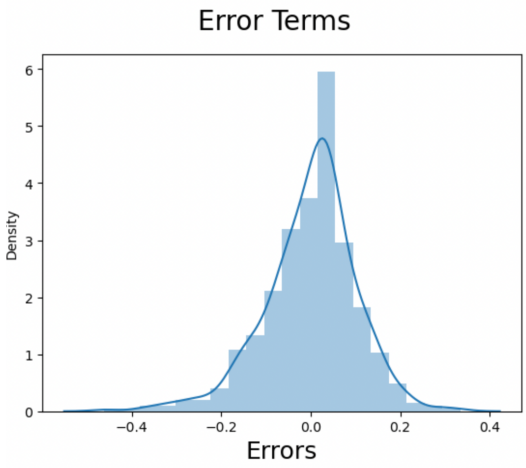
a) Sufficiently high R-squared value, and p values less than 0.05

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.790
Model:	OLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	315.3
Date:	Sat, 01 Apr 2023	Prob (F-statistic):	7.11e-167
Time:	09:41:30	Log-Likelihood:	436.45
No. Observations:	510	AIC:	-858.9
Df Residuals:	503	BIC:	-829.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2804	0.020	14.278	0.000	0.242	0.319
yr	0.2369	0.009	25.621	0.000	0.219	0.255
holiday	-0.0722	0.029	-2.476	0.014	-0.130	-0.015
temp	0.3830	0.026	14.786	0.000	0.332	0.434
windspeed	-0.1498	0.028	-5.400	0.000	-0.204	-0.095

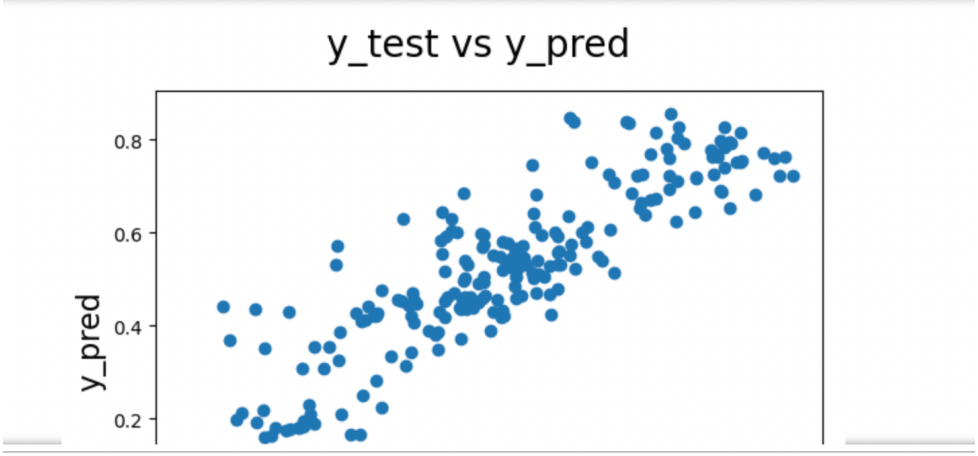
b) Residual Error values normally distributed around mean of 0.



c) VIF values less than 5.0

	Features	VIF
3	windspeed	3.64
2	temp	3.37
0	yr	2.02
5	spring	1.50
4	LightRainyOrSnow	1.05
1	holiday	1.03

d) Y_test vs Y_pred should linear pattern



e) R-squared value on test data (= 0.78) comparable to the R-squared value on the train data (0. 79)

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

   Based on the predicted co-efficients in the model, following features are contributing significantly.

   - temp (coefficient = 0.3830)
   - Weather situation – Light snow or rain (coefficient = -0.2453)
   - Year (coefficient = 0.2369)

General Subjective Questions

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

The goal of the linear regression is to express the relationship between the dependant variable (y) and independent variables (x_1, x_2, …., x_n). The idea is to find the best fit line, so that the error between the actual values and the predicted values is minimized.

The algorithm for a simple linear regression is explained below. (This method can be extended to Multi Linear Regression as well.)

y_i = actual dependant value
x_i = actual independent value
$a_0 + a_1.x_i$ = predicted value

The goal of the algorithm is to find out a_0 and a_1 such that Mean Least Squared Error (MSE) is minimized.

$MSE = 1/N \sum_{i=1 \text{ to } n} (y_i - (a_0 + a_1.x_i))^2$

- The algorithm starts off with a random values assigned to a_0 and a_1. At each iteration MSE is calculated.
- The method uses **Gradient Descent** method to proceed with updating the co-efficient while reducing the MSE.
- The algorithm proceeds till the MSE is minimized.

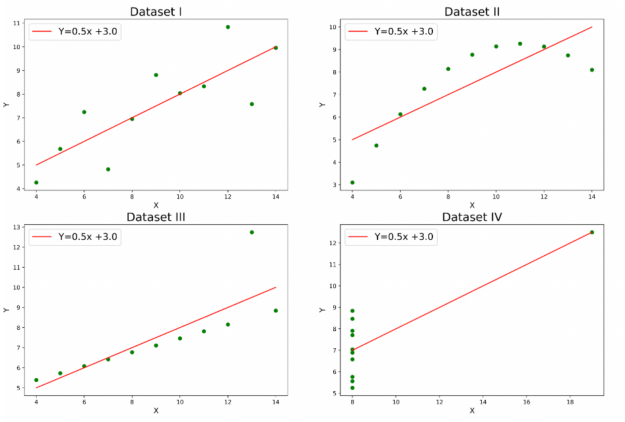
2. Explain the Anscombe's quartet in detail. (3 marks)

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. They have identical statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines. But they have different representations when scatter plotted.

Anscombe's quartet is used to emphasise the importance of graphically analysing the data via scatter plots, and inadequacy of basic statistical properties to describe the data sets.

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

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727



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

Pearson's R is a statistic which measures the correlation between 2 continuous variables, the values lies between -1 to 1. The value of 1 would mean perfect +ve correlation. And -1 would mean no correlation. And 0 means no correlation.

Pearson's R is the covariance of the two variables divided by the product of their standard deviations.

Pearson's R is given by:

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

- 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 y-variable

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

- Scaling is to change values of features so that they are comparable to others in the data set. Normally the values of the features are changed so that values are between 0 to 1 or between -1 to 1.
- Scaling needs to be done, so that co-efficient of different features in the model are comparable. This will help in better inference, i.e in the analysis of how much each variable is contributing to the change in the target variable.
- Standardized scaling basically brings all of the data into a standard normal distribution with mean 0 and standard deviation 1. Where as normalised scaling brings all the data between 0 and 1.

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

The formula for VIF is **1/ (1 – R2)**
When there is a perfect correlation between 2 independent variables, then R2 = 1. That will make VIF infinite.

To solve this problem we need to drop the variables which cause this perfect multicollinearity.

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

A Q-Q plot is a Quantile-Quantile plot, is a graphical plot used to compare the shape of distributions by plotting the quantiles against each other. A point (x,y) on the graph, corresponds to a particular quantile from one distribution plotted with the same quantile from the other distribution. When 2 distributions are similar, the points lie approximately on the identity line x=y.

This helps in the Linear Regression model, when training data and testing data are received separately, then Q-Q plot is used to ensure that both data have similar distribution. If points lie away from the x=y line, then the training and test data are from different distributions.