Subjective Questions

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)

Based on the analysis following conclusions can be drawn about categorical variables:

1. The spring season has a moderate negative effect
2. If the weather is 'Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds' then the usage of bikes takes a significant hit

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.820
Model:	OLS	Adj. R-squared:	0.817
Method:	Least Squares	F-statistic:	285.9
Date:	Mon, 07 Aug 2023	Prob (F-statistic):	2.62e-181
Time:	12:15:59	Log-Likelihood:	476.25
No. Observations:	510	AIC:	-934.5
Df Residuals:	501	BIC:	-896.4
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2684	0.024	10.973	0.000	0.220	0.316
yr	0.2349	0.009	27.390	0.000	0.218	0.252
holiday	-0.0892	0.027	-3.291	0.001	-0.143	-0.036
temp	0.4214	0.030	14.248	0.000	0.363	0.479
windspeed	-0.1452	0.026	-5.605	0.000	-0.196	-0.094
spring	-0.1191	0.016	-7.578	0.000	-0.150	-0.088
winter	0.0469	0.013	3.651	0.000	0.022	0.072
Light	-0.2821	0.026	-10.965	0.000	-0.333	-0.232
Mist	-0.0759	0.009	-8.341	0.000	-0.094	-0.058

Omnibus:	63.227	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	156.197
Skew:	-0.641	Prob(JB):	1.21e-34
Kurtosis:	5.389	Cond. No.	13.5

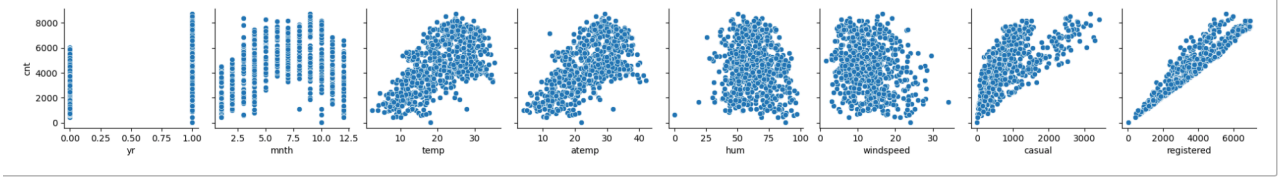
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

If we don't use drop_first = True, then we will be creating variables for all distinct values of that field. Since, these variables are correlated we will have multicollinearity problem.

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

The variable 'registered' has the highest correlation.

```
sns.pairplot(data=bikes, y_vars=['cnt'], x_vars=['yr','mnth','temp','atemp','hum','windspeed','casual','register
plt.show()
```

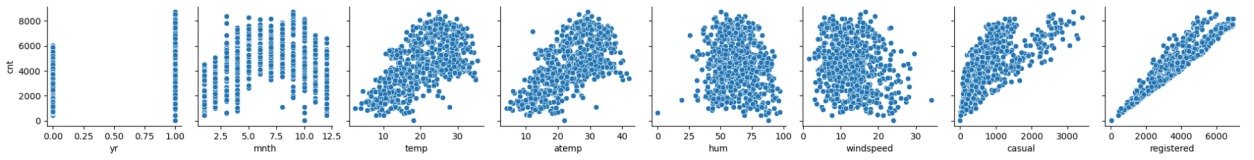


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

There are three assumptions:

1. There is linear relationship between independent variables and the target. This was validated using the pairplot. We can notice that for four variables there is linear relationship

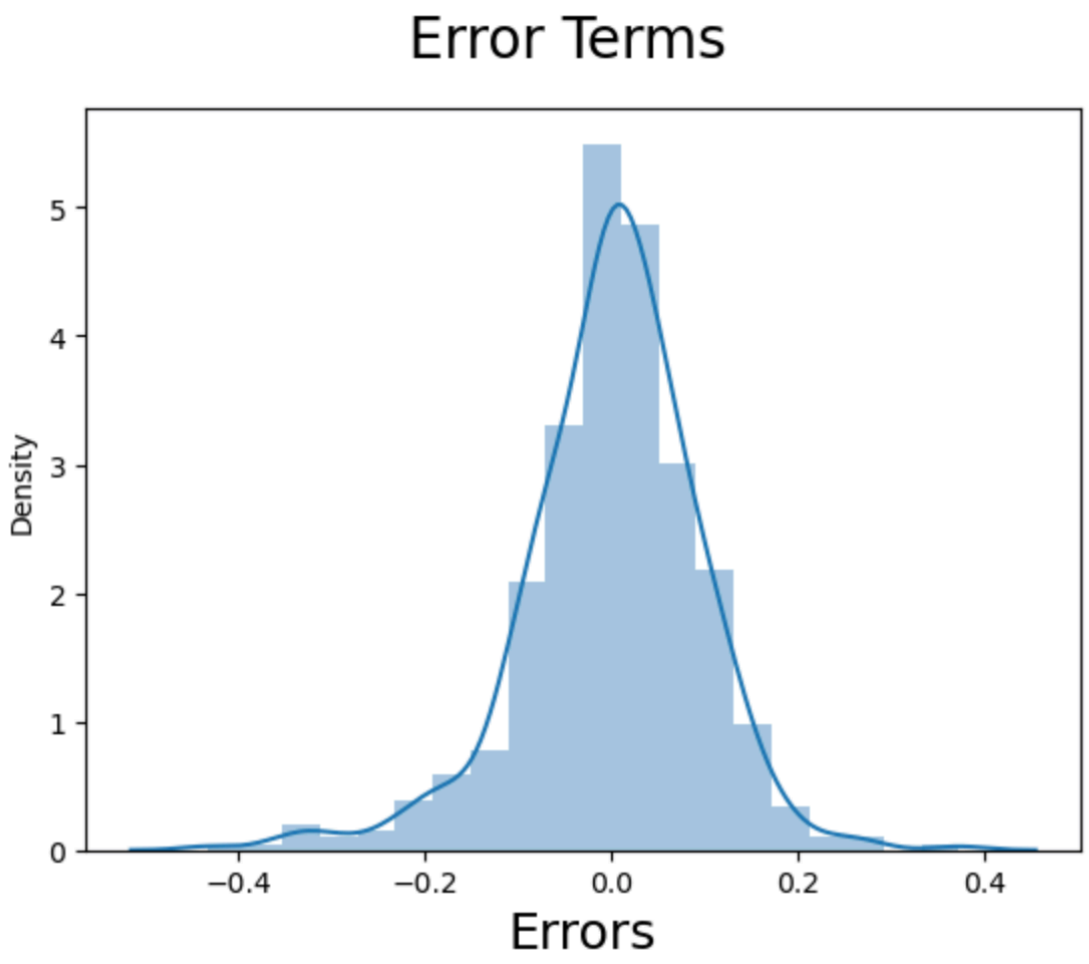

```
sns.pairplot(data=bikes, y_vars=['cnt'], x_vars=['yr','mnth','temp','atemp','hum','windspeed','casual','register
plt.show()
```

2. Error terms are normally distributed. This was validated by plotting the residual and noticing that it was normally distributed

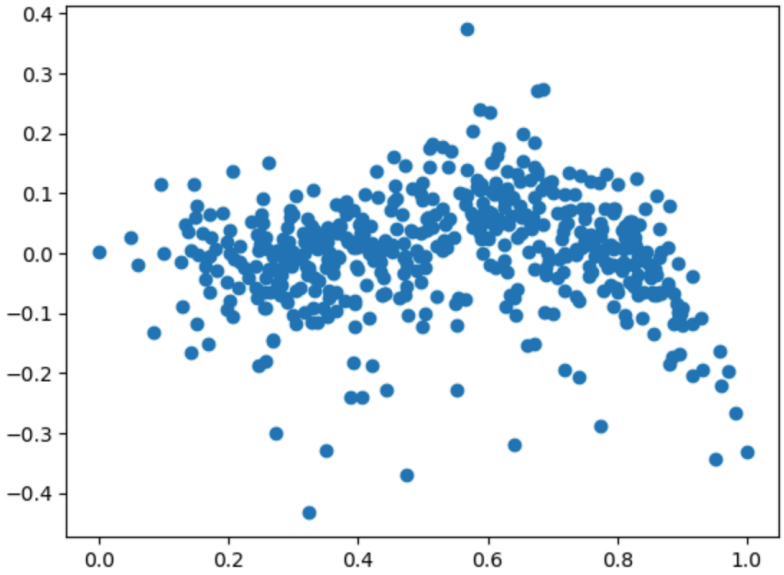
```
[ ]: 1 # Plot the histogram of the error terms
     2 fig = plt.figure()
     3 sns.distplot((y_train - y_train_cnt), bins = 20)
     4 fig.suptitle('Error Terms', fontsize = 20)
     5 plt.xlabel('Errors', fontsize = 18)
```

```
[ ]: Text(0.5, 0, 'Errors')
```



3. Error terms are independent. This was validated by doing scatter plot of residual with the most significant variable. It was noticed that the next residual value was not dependent on the previous one.

```
# We cannot scatter plot residual vs X_train as shapes are different. Hence, we will pick
# the vairable that has the strongest effect on the target variable and plot residual against it
plt.scatter(X_train_lm['temp'],y_train - y_train_cnt)
plt.show()
```



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

The top 3 features significantly contributing to demand of the bikes are:

1. temperature in Celsius
2. weather - ‘Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds’
3. yr - year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm that computes the linear relationship between a dependent variable and independent features. The algorithm's goal is to find the best linear equation that can predict the value of the target variable (y) based on the independent variables (X).

The linear regression tries to find the best fit line represented by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

where B0 is the intercept of the line and Bi's are the coefficient's of the independent variable.

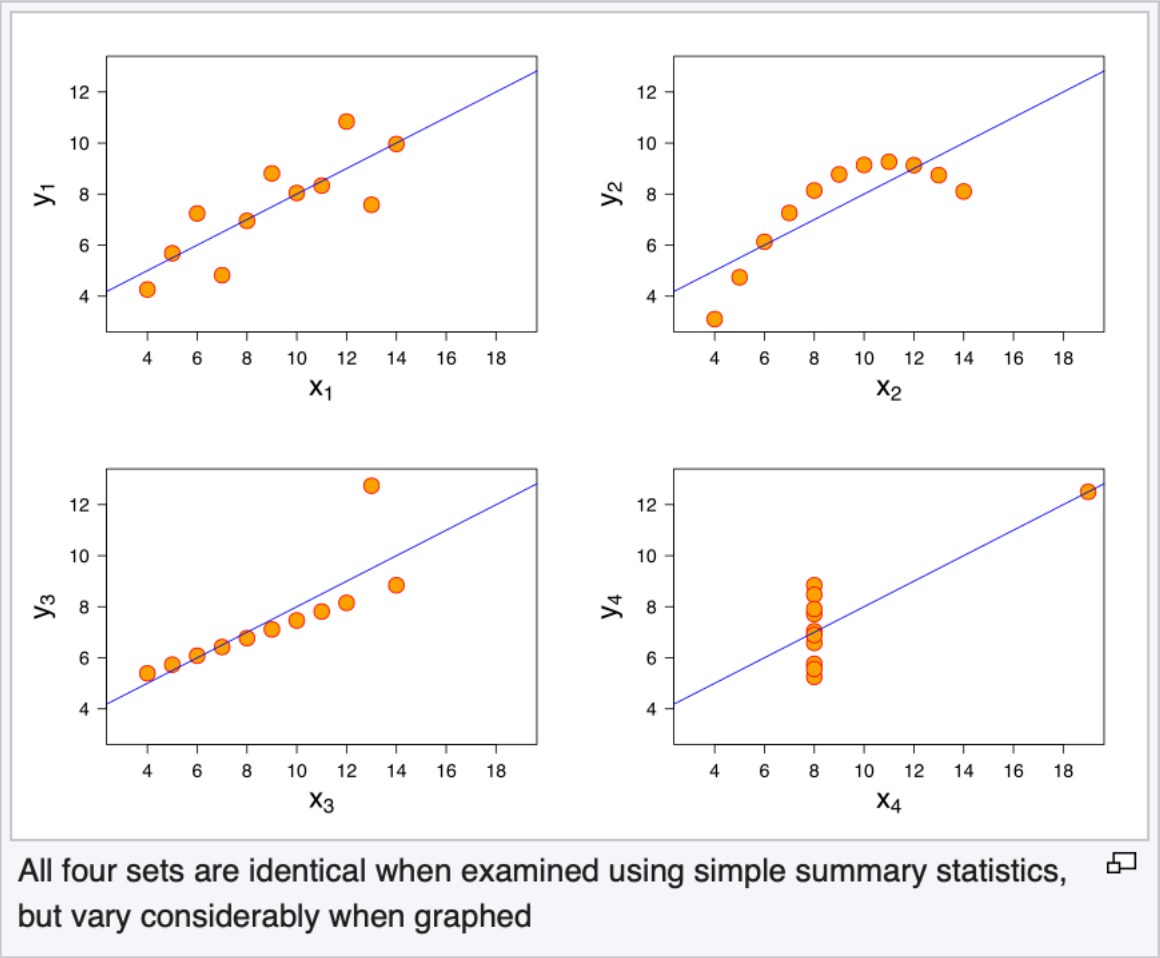
To find the best fit line, the model tries to minimize the residual i.e. the difference between the predicted value and the true value. This is achieved by using optimization algorithm such as gradient descent by iteratively modifying the model's parameters to reduce the mean squared error. The approach is to start with random coefficient values and then iteratively update the values to reach the minimum cost.

The steps involved in a linear regression modeling are:

1. Reading and understanding the data via plots to check for possible linear relationship
2. Data preparation including creating dummy variables for categorical variables
3. Splitting the data into training and testing sets
4. Rescaling the features
5. Using RFE (recursive feature elimination) to identify top n features
6. Building the model and dropping the features that are insignificant
7. Rebuilding the model
8. Use VIF to identify multicollinearity issues
9. Drop the top most feature with VIF > 5 and rebuild the model
10. Repeat step 9 until multicollinearity issue is resolved
11. Conduct residual analysis to validate normal distribution and independence between error terms
12. Make predications and evaluate the model

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet are four datasets that have identical descriptive statistics but have very different distribution. Case in point is the following image (source: Wikipedia)



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

The top left plot seems to be a linear relationship
The top right does not appear to have linear relationship
The bottom left seems to be linear but needs a different regression line than the one in the plot
The bottom right needs a line parallel to the y axis around the cluster of points (barring the outlier on top right)

Anscombe's Quartet was created by statistician Francis Anscombe (year 1973) to demonstrate the importance of plotting the graphs.

3. What is Pearson's R?

Pearson's coefficient measures linear association between two continuous variable. It gives the magnitude as well as the direction of association and is considered the best method of calculating the association.

Important details are:
The coefficient can range from +1 to -1. +1 indicates perfect positive, -1 indicates perfect negative and 0 indicates no association
It is independent of the unit of measurement

Interpretation:
Perfect - if +1 or -1
High - if lies between +/- 0.50 and +/- 1
Moderate - if lies between +/- 0.30 and +/- 0.49
Low - if lies between 0 and +/- 0.29 (not including 0)
No correlation - if it is 0

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

Feature scaling is the method to normalise the features of the model. Regression model works by minimising the cost function. In real world scenarios the model features are in different scale and hence the model will take more time to find the minimum of the cost function. Scaling optimises that process and hence is important. It also helps in interpreting the summary coefficients of the model because the features are on the same scale.

Normalised scaling - It is also known as min-max normalisation and it scales the feature in the range [0,1]. The formulae is:

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

Standardised scaling - It scales the feature to a standard normal distribution with mean 0 and standard deviation 1. The formulae is:

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

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

VIF is calculated as:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

where;

VIFi is ‘i’th variable represented as linear combination of other variables

Hence, VIF will be infinity when Ri squared is 1. that will happen when there is perfect collinearity. That would imply that ’i’th variables is equal to linear combination of other variables.

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

The QQ plot, quantile-quantile plot, helps us identify whether the data probably came from some distribution such normal. A QQ plot is a scatterplot created by plotting two sets of quantiles. If both sets came from the same distribution, we should see the points forming a somewhat straight line. In linear regression it can be used to analyse whether the residuals are normally distributed. To do that a standard normal distribution is plotted as theoretical quantile (red line in the below

image) along with the residuals (first ordered to generate quantile) as sample quantile. The sample quantile should roughly fit on the red line.