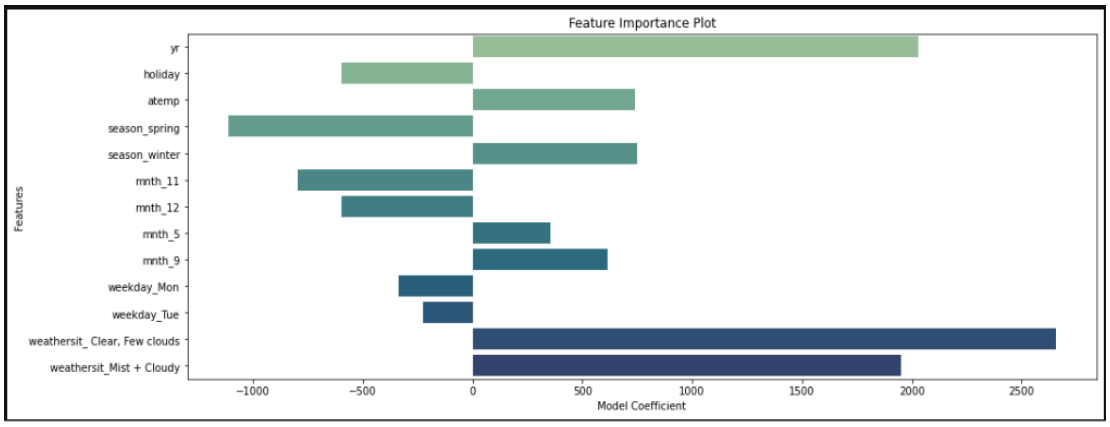
ASSIGNMENT BASED SUBJECTIVE QUESTIONS

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

A. The bike sharing dataset consists of multiple categorical variables which are as follows:

- Year
- Month
- Weekday
- Season
- Working Day
- Holiday
- Weather Situation

From the final model inference, we can see that apart from just one continuous column "atemp", all the other columns picked by the RFE + Filter based on P-value and VIF are categorical columns



- Among them, the 'weather situation: Clear' and 'weather situation: Mist' have the highest value of the coefficient, which also makes sense from a practical standpoint as the demand for bikes is ought to be higher on a clear day than on rainy days.
- The year column has a high positive value of coefficient indicating that demand has increased in 2019 as compared to 2018.
- The column 'season: Spring' has a high negative value of coefficient, which is evident from the time series analysis of the dependant variable, demand is usually higher during the winter months, which is reflected by the positive coefficient of 'season:Winter'.

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

A. If a categorical variable has n levels, then one-hot encoding using pd.get_dummies() will create n different columns. However, if we don't drop one of the columns, then there will be very high multicollinearity in the data as any of the n columns can be expressed as sum of the other n-1 columns (VIF will become infinite). This creates a dummy variable trap, which affects the model's p-values and the inference. Hence, it is important to drop one of the columns, **drop_first=True** drops the first column from that.

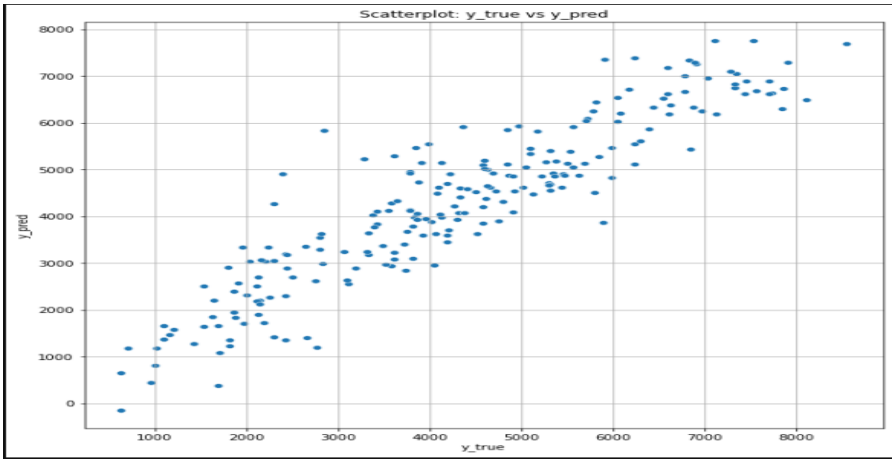
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

A. From the pair-plot, the column 'atemp' seems to have the highest correlation with the target variable, among all the other continuous variables.

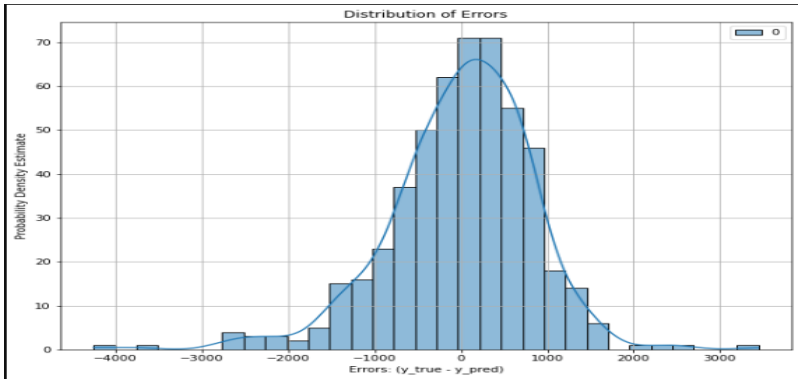
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

A. For validating the assumptions of Linear Regression, the following steps were performed:

   a. **Linear Relationship:** For validating the assumptions of the linear relationship in data, the scatterplot of y_true vs y_predicted was plotted. A linear pattern says that the relationship is linear

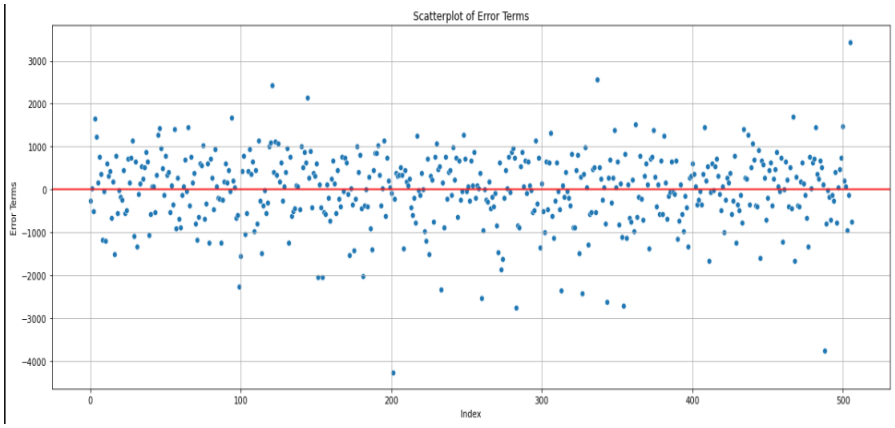


   b. **Errors are Normally Distributed:** The distribution plot for the errors was drawn. A normal distribution around mean zero confirms that the errors are infact, normally distributed



c. **Errors are Independent:** For assessing independence of errors, or to see if there is any serial autocorrelation, the Durbin-Watson statistical test was performed. A test statistic of 2 says zero autocorrelation (the model results 1.933, which confirm about the errors being independent.

d. **Homoscedasticity of Errors:** A scatterplot of the errors is plotted with a horizontal line about 0. The errors are evenly distributed about the line, which confirms about the homoscedasticity of errors.



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

A. Based on the final model, the top 3 features are:
- Weather situation: Clear, few clouds
- Weather situation: mist, cloudy
- Year: 2018 OR 2019

GENERAL SUBJECTIVE QUESTIONS

1. **Explain the linear regression algorithm in detail.**

A. Linear Regression is an approach in modelling a variable Y given one or multiple variables X, with the assumption of a linear relationship between them.

The model formulation is given by:

$Y = \beta_0 + \beta_1 X + €$,

Where,
Y: The dependent variable being modelled.
X: The independent variable (can be more than one) used to predict Y
β_0: The intercept of the equation (basically it is the value of Y when all X values are zero)
β_1: The coefficient of X. It tells by how much Y will change for unit change in X, considering all other factors remain constant.
€: Error Term not explained by the model.

The Linear Regression model finds the best fit line by minimising the sum of squared errors.

The Linear Regression Model has the following assumptions:

<u>For Fitting the Line:</u>

- There should be a linear relationship between X and Y

<u>For making Inference:</u>

- The residuals (the difference between actual Y vs predicted Y) should be normally distributed with mean around zero.
  - If the residuals are not normally distributed, then the p-values of the X obtained during the hypothesis test to determine the significance of the coefficients becomes unreliable.
- Residuals are independent of each other.
  - If the residuals are not independent, the error standard distribution will be biased and the inferences will not be proper.
- Residuals have a constant variance
  - If the residuals exhibit a certain pattern, or if the variance is not constant, then there is some information in the residuals which the model has not been able to capture, and the model can be improved further.
- [FOR MULTIPLE LINEAR REGRESSION] There should not be any multicollinearity between the independent variables.
  - If there is multicollinearity between the predictors, then the p-values of the predictors are not reliable, the signs may invert, the coefficients swing wildly.
  - Multicollinearity does not affect the predictions or the goodness of fit of the model.

2. Explain the Anscombe's quartet in detail

A. The Anscombe's quartet consists of four data sets containing 11 data points, each of which has a very similar descriptive statistics (mean, variance, correlation etc.) but when plotted on a chart, the distributions look completely different from each other. The Anscombe's quartet asserts the importance of presenting data through visual storytelling instead of showing in tabular form.

The following are the four datasets:

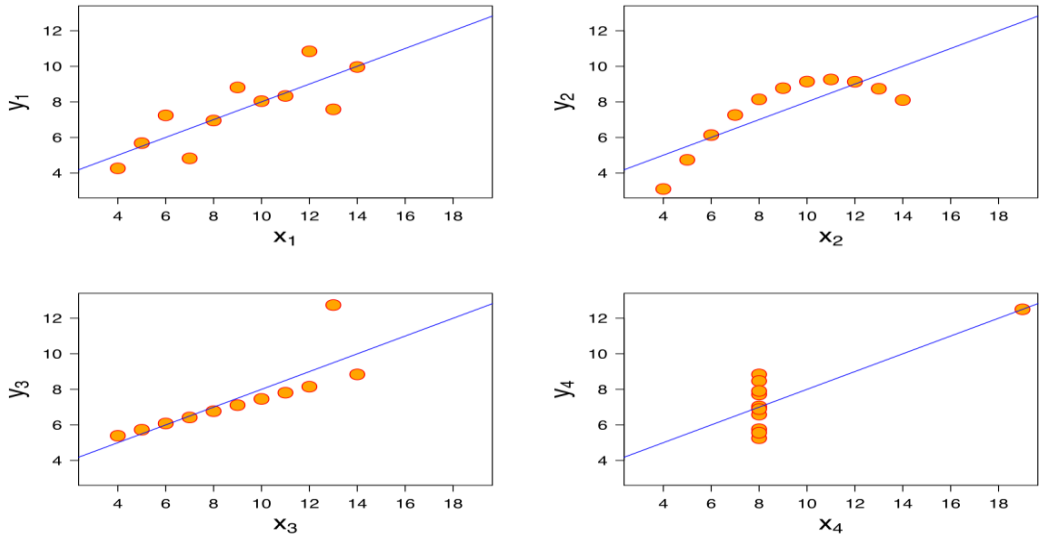
Anscombe's quartet

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

The Descriptive statistics of datasets are as follows:

Statistic	Value
Mean of X	9
Mean of Y	7.50 (same upto 2 decimal places)
Variance of X	11
Variance of Y	4.125 (+- 0.003)
Correlation of X and Y	0.816 (same upto 3 decimal places)
LR Best fit line	Y = 3.00 + 0.500X (same till 2 and 3 decimal places respectively)
R2 Score	0.67 (same upto 2 decimal places)

However, when plotted, the distributions look completely different, as shown below:



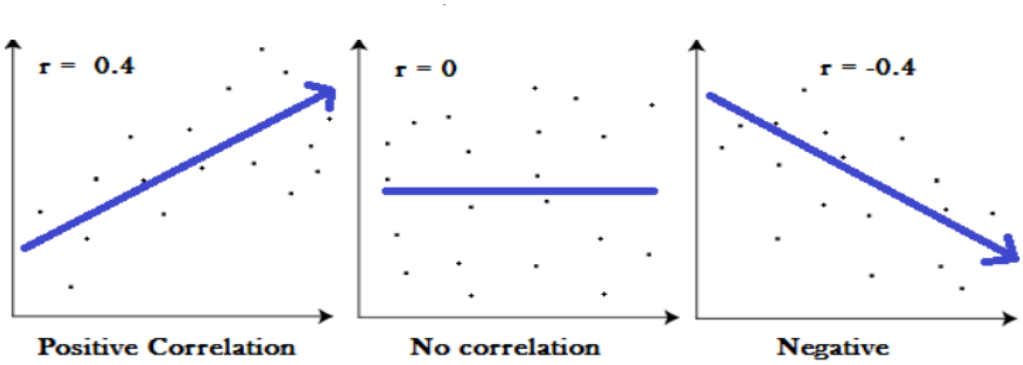
Observations:

- The first data appears to be a linear relationship between X and Y modelled by the given line.
- In the second graph, the relationship between X and Y is quadratic, hence the Pearson correlation coefficient is not valid. A GLM model is a better fit to it.
- The third graph has a very strong linear relationship, but due to the presence of outlier, the regression line is offset and is not the best-fit line for the data.
- The fourth graph shows that there is not relationship between X and Y, but due to one outlier, the descriptive statistic shows a correlation of 0.81.

3. What is Pearson's R?

A. Pearson's r, or Pearson's Correlation Coefficient, is a numerical measure which describes the strength of linear association between two variables. It varies between -1 and +1 where:

- r >0: positive correlation or a positive association between the two variables
- r<0 : negative correlation or a negative association between the two variables.
- r = +1: The association is perfectly linear with a positive slope
- r = -1: The association is perfectly linear with a negative slope
- r = 0: No correlation/association between the two variables.



The formula for Pearson's R is given by:

$$r = \mathbf{cov(X, Y) / \sigma_X \sigma_Y} = \sum(X_i - X')(Y_i - Y') / \left(\sqrt{\Sigma}(X_i - X')^2\right)\left(\sqrt{\sum}(Y_i - \bar{Y})^2\right)$$

where,

X, Y: The two variables/series being considered
Cov(X,Y): Covariance of X and Y
$\sigma_X \sigma_Y$: Product of Standard Deviation of X and Y
X', Y': Mean of X and Y respectively

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A. Scaling is the process of making transformations to the data to bring it under a desired range/magnitude or to centre it around the mean. There are two types of scaling usually performed as a part of data pre-processing:

- Normalization: It is the process in which the data is shifted and rescaled within the range [0,1].
- Standardization: It is the process of rescaling the data to be centred around zero mean and unit standard deviation.

Feature Scaling is performed as a part of the data pre-processing step. Feature scaling makes the flow of gradient descent easier and it helps the optimisation function to reach the global minima faster. Feature scaling is needed for Linear Regression models, distance-based algorithms as without scaling, the algorithm may get biased towards the feature with a higher magnitude. However, some algorithms like tree-based models do not explicitly need feature scaling.

The difference between normalized scaling and standardized scaling are as follows:

Normalized Scaling

- Normalized scaling rescales the data to within the range zero and one.
- Normalized scaling affects the outliers within the data.

Standardized Scaling

- Standardized scaling rescales the data to be centred around 0 mean and unit standard deviation.
- Standardized scaling does not affect the outliers within the data.

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

A. The Variance Inflation Factor, or VIF is a measure of the degree of multicollinearity between the independent features in a dataset. The VIF becomes infinite when there is a perfect relationship between one or more variables, i.e, when a combination of one or more variables perfectly describes the other independent variable.

The VIF of the i^{th} independent variable is given by:

$VIF_i = 1/(1-R_i^2)$

Where R_i^2 is the R2 score of the ith independent variable.
It is calculated by forming a regression model for each independent variable using the other independent variables as the regressors.

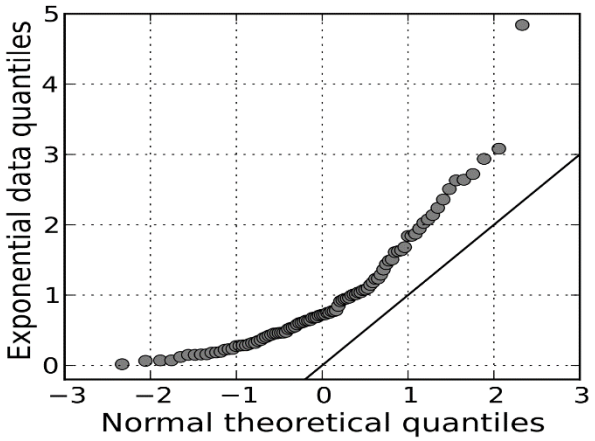
Now, if VIF_i = infinite
Then $1-R_i^2 = 0$
$=> R_i^2 = 1$

That is, in the regression model formed with ith variable as the target variable and other independent variables as the regressors, the regressors can explain 100% of the variance in the ith variable. A VIF of infinity means there is a definite multicollinearity between the variables. It can also be infinite during dummy variable trap, i.e, during Dummy Encoding the categorical variables, if drop_first = True is not supplied (or if one of the dummy variables is not removed manually)

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

A. The Q-Q Plot is a probability plot which describes if two datasets or series come from a same population with equal distribution.
The Q-Q plot is a graph of the quantiles of the first dataset plotted against the quantiles of the second dataset. A reference line at 45 degree is plotted on the graph. If the two distributions are coming from the same population with equal distribution, the points will lie approximately on the 45 degree reference line. The more the deviation from the reference line, the more evidence we get that the two data come from population with different distributions. It can also be used to compare if a particular dataset comes from a defined theoretical population distribution.

The Q-Q Plot is displayed as below:



The use and importance of Q-Q plots are as following:

- Q-Q plots are used to compare the shape and graphical properties like skewness, scale, location, presence of outliers in the two distributions.
- It is a more powerful approach to comparing two distributions rather than using histogram plots.
- It can also compare two distributions with different sample size.