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?

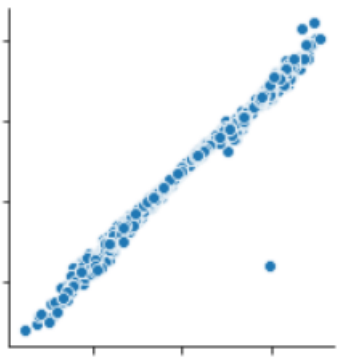
Ans: There are several categorical variables in the dataset like, 'season', 'mnth', 'weekday', 'weathersit'. And the other are the Binary categorical variables in the form of 1's and 0's. These variables unlike the continuous variables they form patterns on a particular no.of elements or we can say levels. These patterns on these levels are unable to identify mostly we can say they depend on the count. They cannot by entered into the regression equation just as they are. Instead, they need to be recoded into a series of variables which can then be entered into the regression model. So, We use these dummy variables created from these categorical variables to fit into the regression model.

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

Ans: The attribute **drop_first = True** is from the method get_dummies() from pandas Library which is used to create the dummy variables for the categorical variables. From the concept of creating the dummy variables we need only "n-1" no.of dummy variables where "n" is the no.of levels in the categorical variables. It reduces the no of dummy variables for each categorical variable and also helps understanding the data. **drop_first = True** helps dropping the first dummy variable created to generate "n-1" dummies.

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

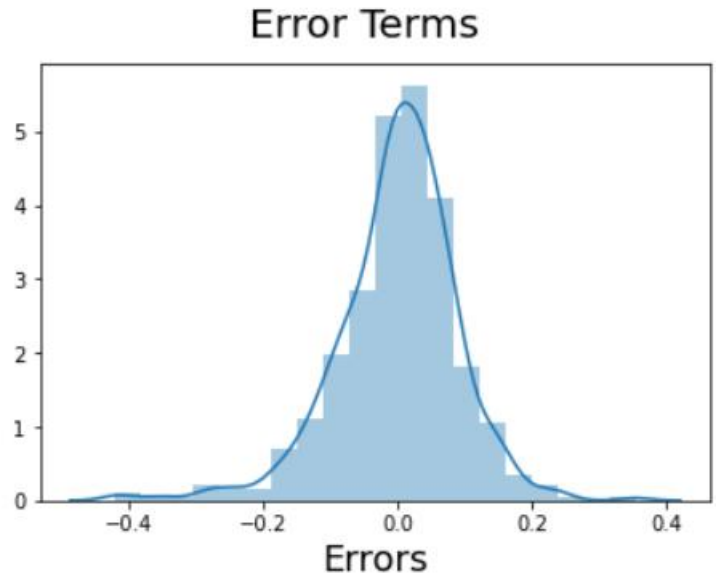
Ans : Looking at the pair plot from the given data we can clearly see that there is the highest correlation between the variables "atemp" and "tmp" as we can see from the plot and the heat map.



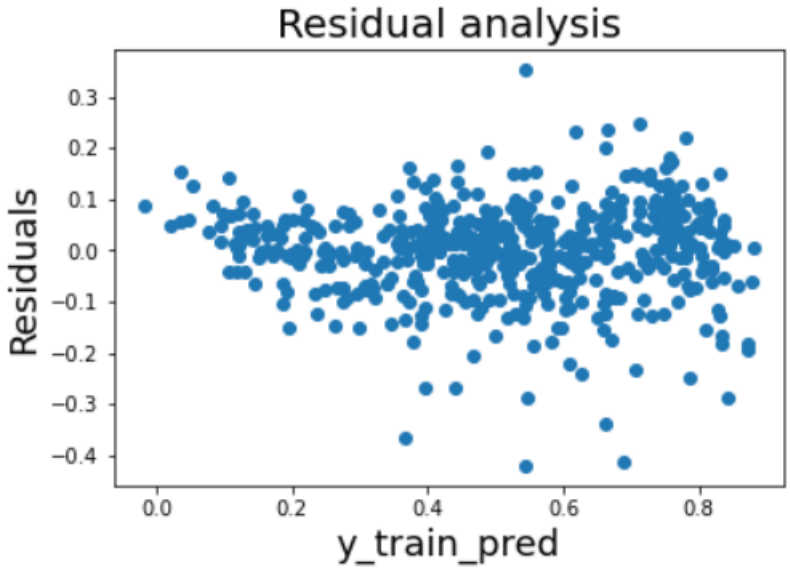
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans : We can validate the Linear regression model based on various factors. They are

Assumptions for Residual analysis



Error terms are normally distributed with mean zero



Error terms are independent of each other and Error terms have constant variance

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

Ans : The best 3 features that contributing significantly towards explaining the demand of the shared bikes

1. Yr
2. Tmp
3. weathersit_Light snow

These 3 variables have significantly highest coefficients. For the feature 'yr' the coefficient is 0.2338,

tmp has 0.4923 and finally

weathersit_Light snow has -0.2856.

That is for every one-unit change in the value of x in the equation

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

The y value changes by the coefficient value of that variable.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Regression:

- "Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable"
- The output variable to be predicted is a ***continuous variable***, e.g. scores of a student

Step 1: **Reading and Understanding the Data**

Let's start with the following steps:

1. Importing data using the pandas library

2. Understanding the structure of the data

Step 2: **Visualising the Data**

Step 3: Performing Simple Linear Regression

Equation of linear regression

$y=c + m_1 x_1+m_2 x_2 +...+m_n x_n$

y is the response

c is the intercept

$m1$ is the coefficient for the first feature

mn is the coefficient for the nth feature

Step 4: Residual analysis

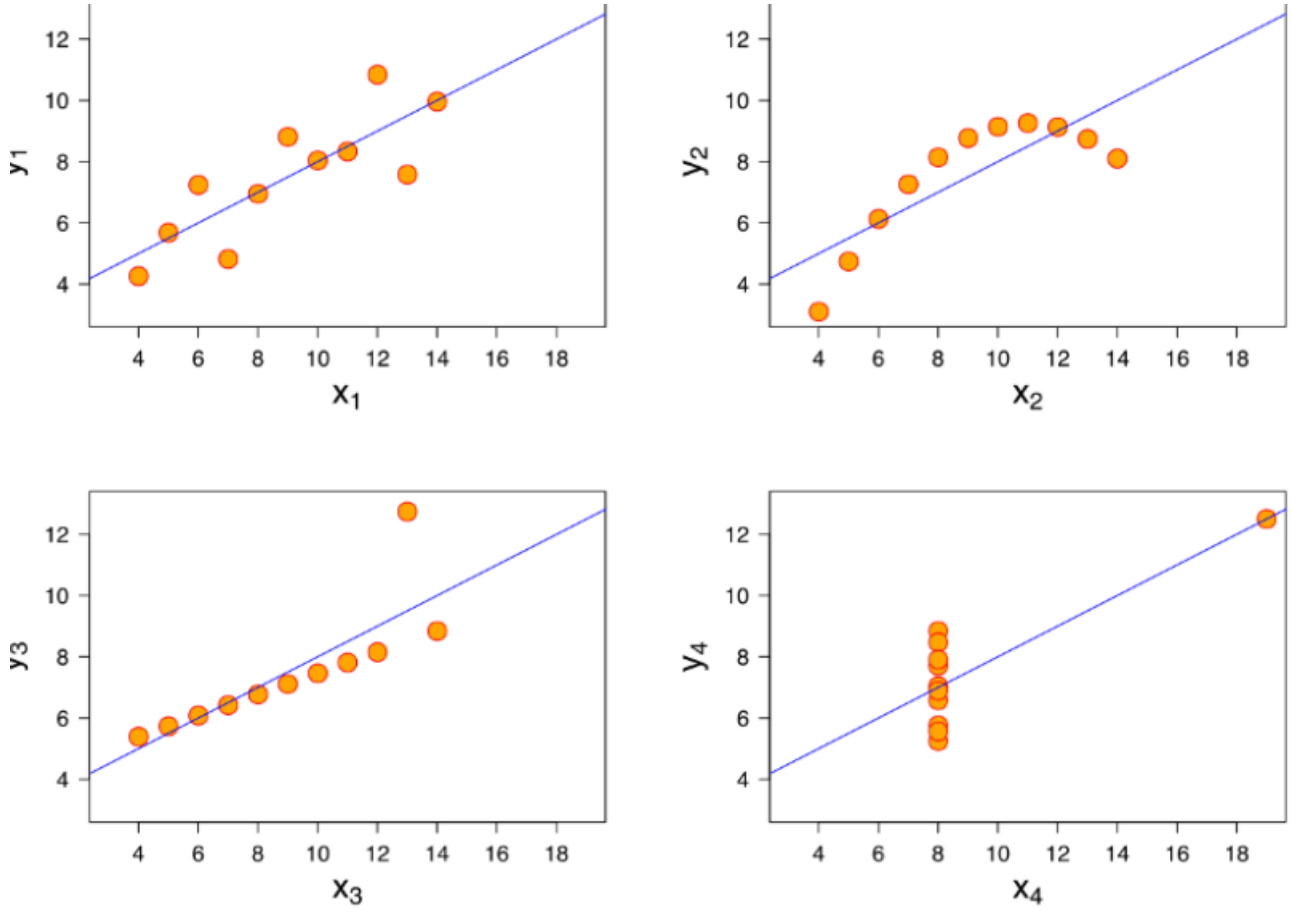
To validate assumptions of the model, and hence the reliability for inference

Step 5: **Predictions on the Test Set**

Now that you have fitted a regression line on your train dataset, it's time to make some predictions on the test data. For this, you first need to add a constant to the X_test data like you did for X_train and then you can simply go on and predict the y values corresponding to X_test using the predict attribute of the fitted regression line.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing **eleven** (x,y) pairs. ... The correlation coefficient between x and y is 0.816 for each dataset. They all have nearly identical variances, correlations, and regression lines. To demonstrate both the importance of **graphing data** before analysing it and the effect of outliers and other influential observations on statistical properties.



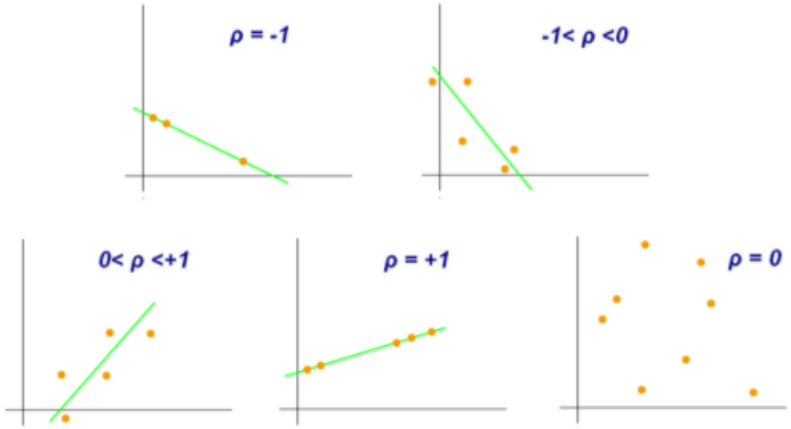
- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Ans: Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. It has a value between +1 and −1.

- **r = 1** means the data is perfectly linear with a **positive slope** (i.e., both variables tend to change in the same direction).
- **r = -1** means the data is perfectly linear with a **negative slope** (i.e., both variables tend to change in different directions).
- **r = 0** means there is **no linear association.**
- Examples plots.



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

Scalling: Scaling is a method of standardising the values of the variables to a comparable range to all other values of the variables. Scaling of variables is an important step because, some variables have a different scale with respect to all other numerical variables, which take very small values. Also, the categorical variables that may encoded earlier take either 0 or 1 as their values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.

Scaling is performed to

1. Ease of interpretation
2. Faster convergence for gradient descent methods

Difference between normalized scaling and standardized scaling

Standardizing

- Subtracting mean and dividing by standard deviation such that mean is 0 and has SD of 1.

It can be calculated as

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

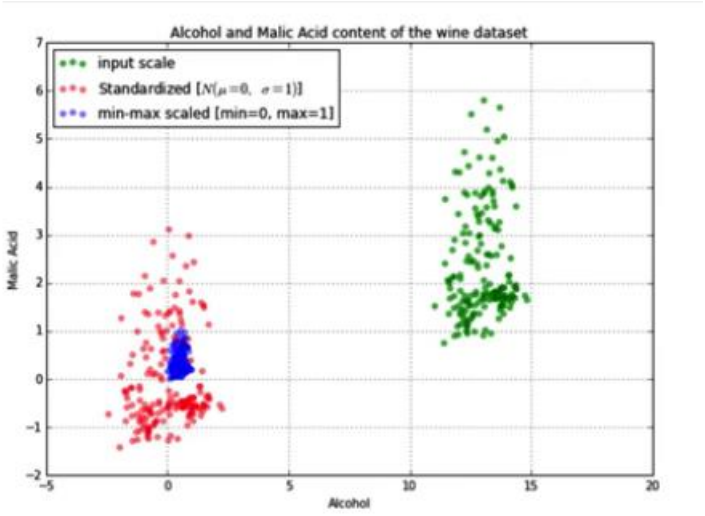
Minmax scaling

- It brings all of the data in the range of 0 and 1.

It can be calculated as

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

They can be plotted as -



Here the red dots represent the standardised, blue represents the Min Max and the green represents the actual data points.

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

Ans: As we know that the value of VIF can be calculated by the formula.

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

VIF value can only be infinite only when the value of the R^2 is equal to one. That means the Linear regression model that is built is exactly perfect.

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