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)**

   The categorical variables from the dataset are season, yr, mnth, holiday, weekday, workingday, weathersit.

   season: Most bike bookings were happening in season3, followed by season2 & season4.
   yr: by increase in year the bike booking count increases.
   mnth: Most bike bookings were happening in the months May to Oct.
   weathersit: There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable. Most of the bike bookings were happening during 'clear_clouds', followed by 'most_cloudy'.
   holiday: Most of the bike booking were happening when it is not a holiday.
   weekday: weekday variable shows very close trend on all days of the week.
   workingday: Most of the bike bookings were happening on 'workingday'.

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

   The importance of using **drop_first**=**True** is, it helps in reducing the redundant column created during dummy variable creation. If not the multicollinearity is created among dummy variables thus, a big issue when trying to interpret the model. Hence using it we can reduce the correlation among created dummy variables.

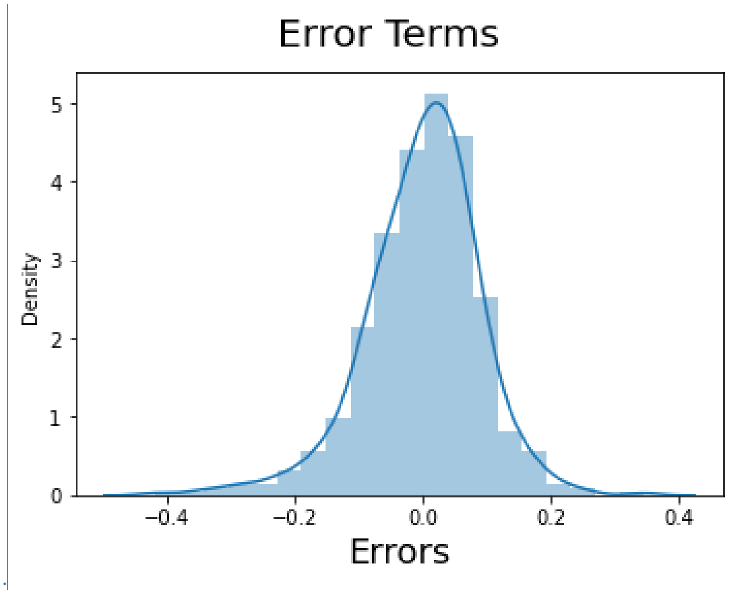
   Eg: workingday – there can be two dummy columns for this column non-workingday and workingday but if workingday is 1 then obviously non-workingday will be 0 and also if workingday is 0 then obviously non-workingday will be 1, so to reduce the redundant column drop_first=True is used.

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

   The numerical variables temp and atemp are strongly correlated with the target variable (cnt).

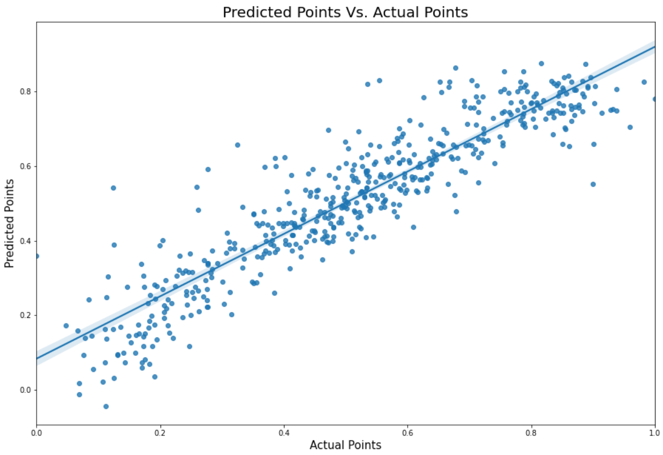
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Normality of Error



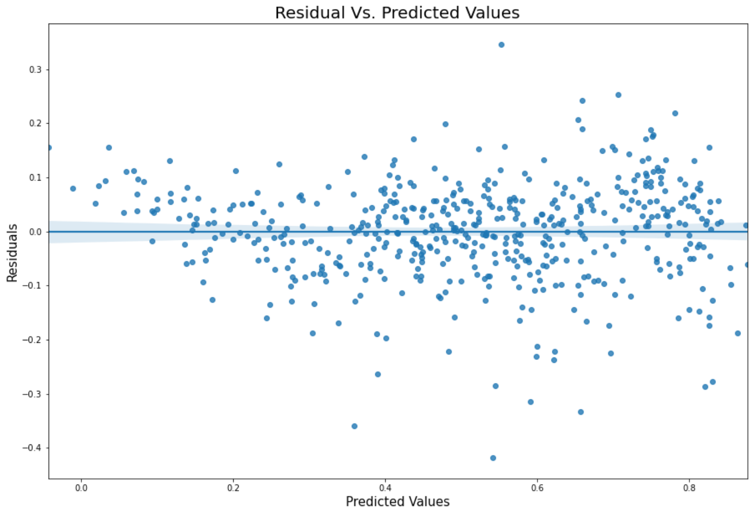
Observation: Error values are normally distributed.

Homoscedasticity



Observation: The probability of the errors has constant variance

Independence of Errors



Observation: Error values are statistically independent

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

The top 3 features:

- temp - coefficient: 0.473
- yr: coefficient : 0.234
- weathersit_light_snow– coefficient: -0.286377

General Subjective Questions

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

Linear Regression Algorithm is a type of supervised learning machine learning algorithm that is used for the prediction of numeric values. Linear regression is a part of regression analysis which is a technique of predictive modelling that helps finding out the relationship between Input and the target variable. In the case of linear regression the two variables which are on the x-axis and y-axis should be linearly correlated.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Types of Linear Regression

Linear regression further divided into two types of the algorithm:

- **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable.
- **Multiple Linear regression:** If more than one independent variable is used to predict the value of a numerical dependent variable.

Regression Line:

The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1.X$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
β_0 = intercept of the line (Gives an additional degree of freedom)
β_1 = Linear regression coefficient (scale factor to each input value).

The regression line of Multi Linear Regression for *p* explanatory variables $x_1, x_2, \ldots, x_p$ is defined to be

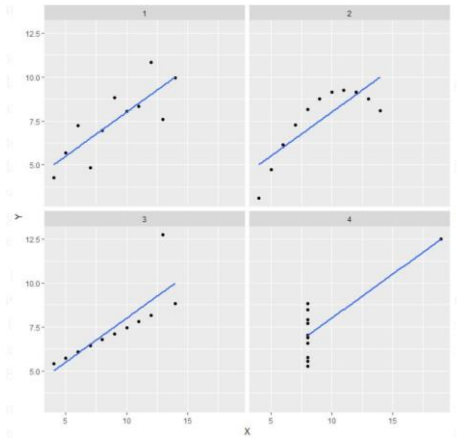
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

β1 = coefficient for X1 variable, β2 = coefficient for X2 variable
β3 = coefficient for X3 variable and so on… β0 is the intercept (constant term).

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

Anscombe's Quartet defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.



Explanation of this output:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.

- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

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

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

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
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization is technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

Here's the formula for normalization: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.
When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1
If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1.

Standardization is a technique where the values are centred on the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

Here’s the formula for standardization: $X' = \frac{X - \mu}{\sigma}$
μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/ (1-R2) infinity.
Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it’s R-squared value will be equal to 1.So, VIF = 1/ (1-1) which gives VIF = 1/0 which results in “infinity”**.**

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

(3 marks)

Quantile-Quantile (Q-Q) plot are plots of two quantiles_against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

The advantages of the Q-Q plot are:

1. The sample sizes do not need to be qual.
2. Many distribution aspects can be simultaneously tested.

The Q-Q plot is used to answer the following questions:

- Do two datasets come from population with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?