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)**

Ans: The categorical variables in our dataset are as below:

- **Weekday :** The sale is throughout the week similar , no spike or dip seen .
- **Workingday :** The sale is throughout the week similar , no spike or dip seen .
- W**eathersit :** The sale seems to be high during the clear weather , while least during the light snow
- **Holiday :** Sales doesn't seem to be affected with this variable .
- **Season** : We can see the sales is highest in fall and least in spring .
- **Yr** : **'**yr**'** with two values 0 and 1 indicating the years 2018 and 2019 respectively with sales in 2019 is much higher showing gaining popularity.
- **mnth** : Slow sales count in starting months of year with gradually increasing with highest in september , then there is dip

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

Ans : It is important to use drop_first=True during dummy variable creation because if we don't drop the first column then your dummy variables will be correlated . This may affect some models adversely and the effect is stronger when the cardinality is smaller.

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

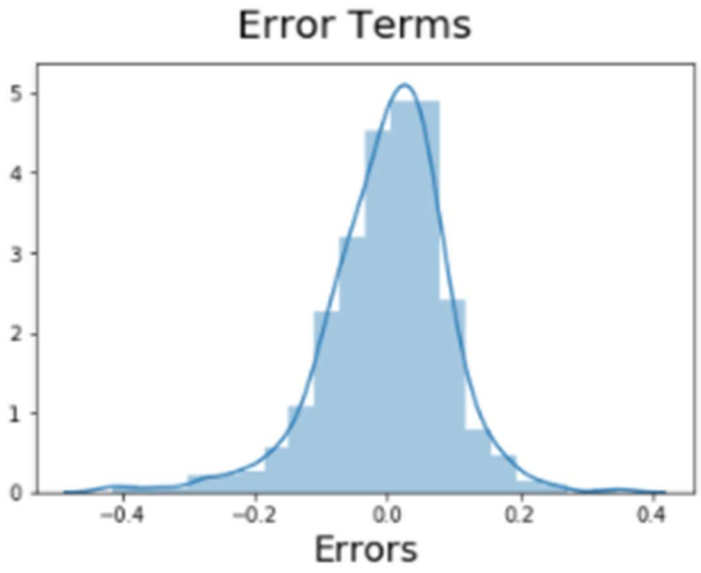
Ans : Temperature has highest correlation with the target variable , i.e, cnt

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

Ans : Assumptions of Linear regression :

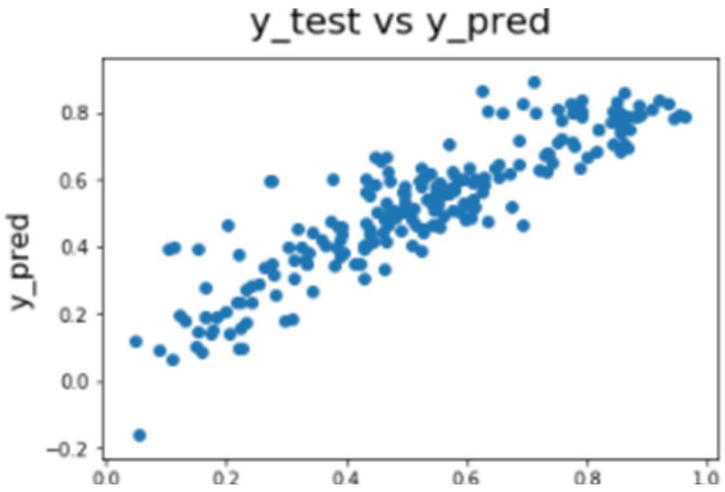
- Multicollinearity (error terms are independent of each other): VIF for all independent variables is less than 5 .
- Linear relationship between X and Y
- Error terms are normally distributed .
- Error terms have constant variance

In the below figure we can see for the model , we can see the error term are normally distributed



Caption

and residuals have mean value of zero .



Caption

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

Ans : Temperature , year and sept are contributing towards explaining the demand of shared bikes their coefficients are highest among all the variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans : In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables. In technical terms, linear regression is a machine learning algorithm that finds the best linear - fit relationship on any given data , between dependent and independent variables.
Independent variable is the predictor value and dependent variable is the output variable.

$$y = B0 + B1*x$$

y : dependent variable
x : independent variable
B0 : intercept
B1: slope

So with unit increase in x there would be B1 unit increase in y

Assumption of linear regression :

- Linear relationship between X and Y
- error terms are normally distributed
- error terms are independent of each other
- error terms have constant variance .

2. Explain the Anscombe's quartet in detail.

It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven *(x,y)* pairs .
We see the real relationships in the datasets start to emerge when we plot them. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?). Dataset III looks like a tight linear relationship between *x* and *y*, except for one large outlier. Dataset IV looks like *x* remains constant, except for one outlier as well.

Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

3.What is Pearson's R?

The most common measure of "correlation" or "predictability" is **Pearsons coefficient of correlation**, although there are certainly many others. Pearsons **r**, as it is often symbolised, can have a value anywhere between -1 and 1.

The larger r, ignoring sign, the stronger the association between the two variables and the more accurately you can predict one variable from knowledge of the other variable. At its extreme, a correlation of 1 or -1 means that the two variables are perfectly correlated, meaning that you can predict the values of one variable from the values of the other variable with perfect accuracy. At the other extreme, an r of zero implies an absence of a correlation

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

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Eucledian distance between two data points in their computations, this is a problem.

If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.
Normalization usually means to scale a variable to have a values **between** 0 and 1, while **standardization** transforms data to have a mean of zero and a standard deviation of 1.

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

The ***variance inflation factor*** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing *collinearity/multicollinearity*. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

The value of VIF is infinite shows a perfect correlation between two independent variable.
In the case of perfect correlation we get R2 =1 , which lead to 1/(1-R2) infinity.

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

Q-Q Plots (Quantile-Quantile plots) 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.

Q-Q plots is important to find out if two setsof 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.