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 in the dataset were season, holiday, weathersit, mnth, yr & weekday. These were visualized using boxplot. These variables had the following effects on our dependant variable:

   a) season:
   Fall season seems to have the highest median when compared to others, closesly followed by Summer and Winter.
   Spring season doesn't seem to attract many customers..

   b) weathersit:
   There was no users during heavy rain & snow indicating
   that this weather is extremely unfavourable. Highest count was seen
   when weather was clear & misty.

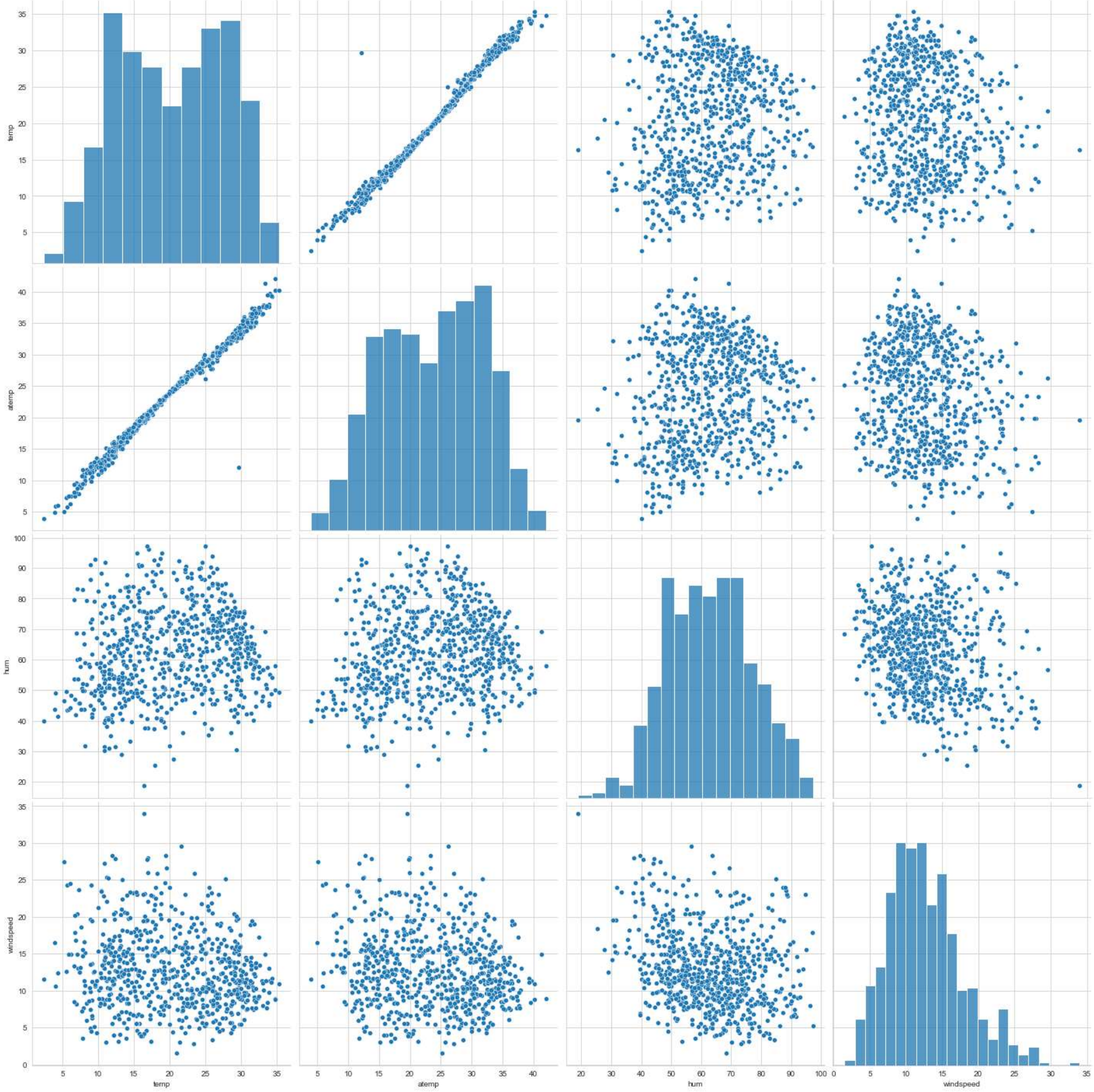
   c) holiday:
   During the holidays, less people seem to rent a bike since the median has dropped by 2000 approx. Another point to be noted is that there is a lot more variability in rentals during the holidays.

   d) mnth:
   September saw the highest number of rentals while December
   saw the least. In general, May-October Months seems to have more rentals than the rest of the year
   e) yr:
   The number of rentals in 2019 was more than 2018

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

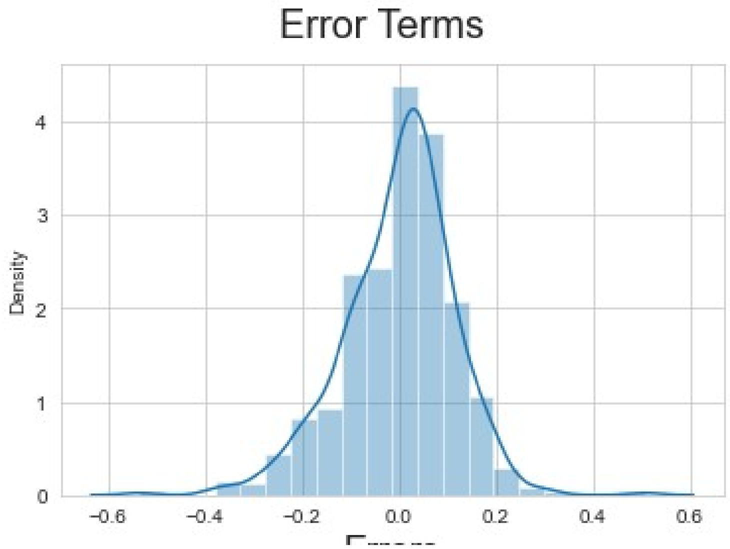
If you won't drop the first column then your dummy variable will be correlated(redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may trouble converging and lists of variable importance may be distorted. Another reason is, if we have all dummy variables it leads to multicollinearity between the dummy variables. To keep this under control, we drop one column

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

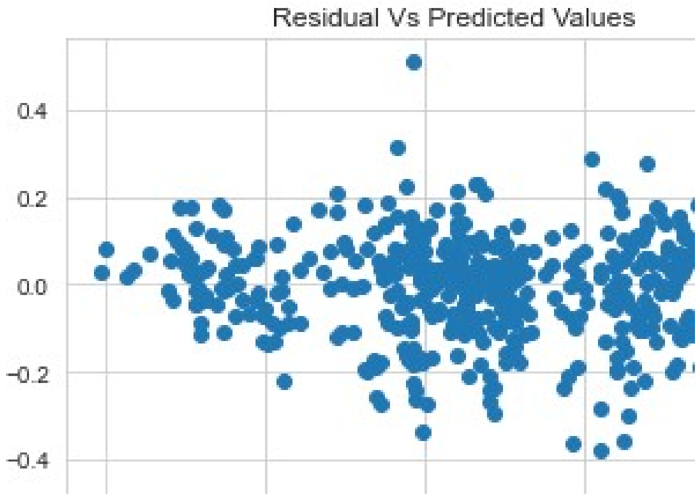


"temp" and "atemp" are the two variables which are highly correlated with the target variable(cnt)

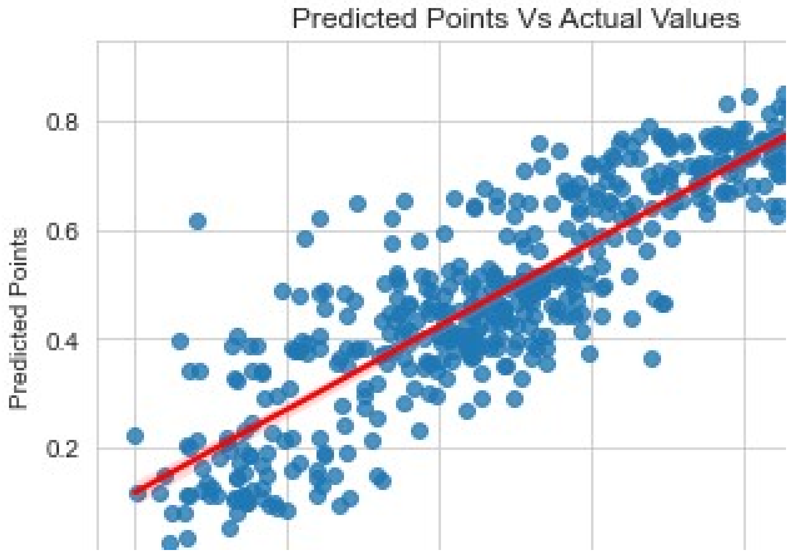
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
    a. *Residual distribution should follow normal distribution and centred around 0(mean = 0).*



    b. *Error terms are independent of each other.*



c. *Error Terms have approximately a constant variance. Hence follow Assumption of Homoscedasticity*



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

   Year(yr): A coefficient value of ‘0.257098’ indicates that year wise bike rental numbers are increasing.

   September: A coefficient value of ‘0.094174’ indicates that during September bike rentals increases

   Light Rain: A coefficient value of ‘-0.29725’ indicates that the light rain deters people from renting out bikes

General Subjective Questions

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

   Linear Regression (LR) is a supervised learning algorithm that uses a line as a function that approximately characterises all the data points in given set.

The LR graphically shows as a line that passes through / near all the data points in such a manner that the vertical distance between the data points and the line is minimum. If the number of variables in the data set other than dependent variable is one, then LR is called as Simple Linear Regression (SLR) else in case of multiple variables its called as Multiple Linear Regression (MLR).

Mathematically LR is defined as **y=ß0+ß1X**,

where y is the dependent variable and

X is an independent variable.

ß0 is referred to as the intercept of line or constant.

ß1 is called as the coefficient.

Since it is not possible for all the points to be fitted on a straight line, the aim of the Linear Regression is to minimise the Cost Function (CF).

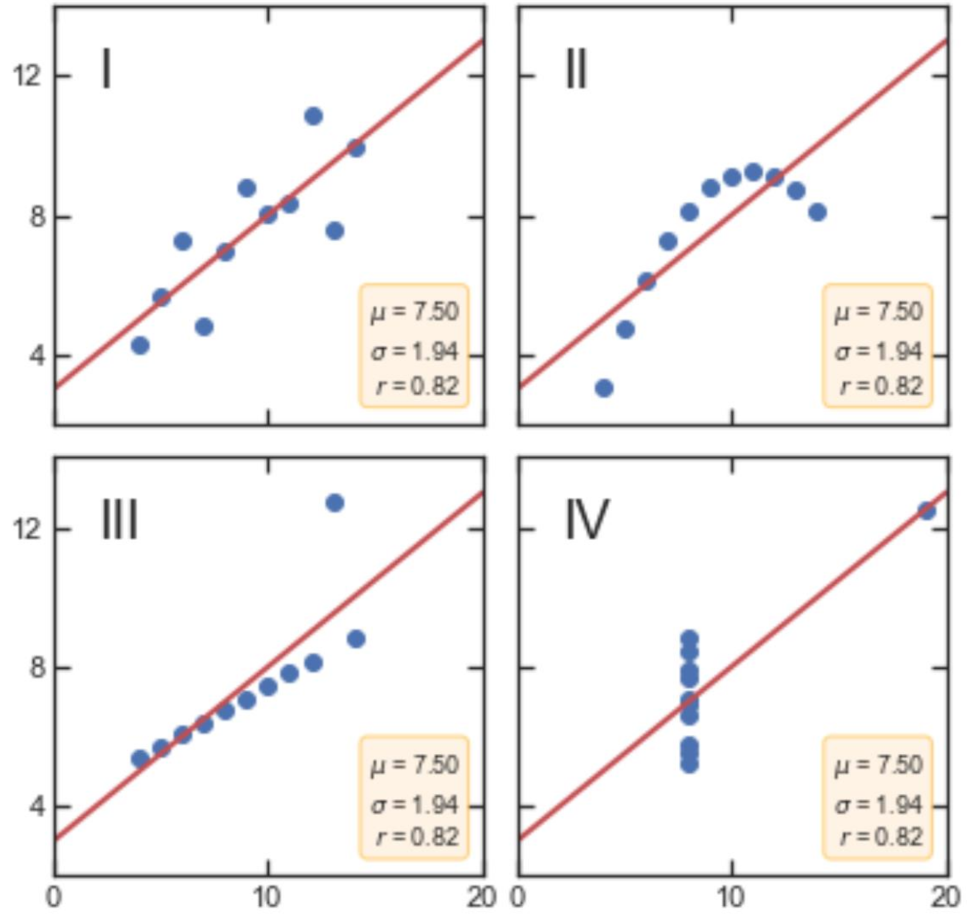
CF is the difference between actual value and the predicted value. In case of LR, CF is defined as Root Mean Squared Error (RMSE) or the mean Mean Squared Error (MSE). This method is also called as Ordinary Least Square (OLS) Method and has been applied in this instant study.

To minimise ß0 and ß1 values Gradient Descent (GD) method is used. In this method coefficient values are selected randomly and thereafter iteratively updated to arrive at the minimum value for the cost function. The GD algorithm calculates the next point by calculating the gradient at the current position and scales it by the learning rate and later subtracts it from the current position. The learning rate therefore determines the step size. Too less would imply a slowly iterating algorithm. A big step would lead to overstepping the minima, thereby giving false minima.

Once the minima is established, the value of the coefficients, ß0 and ß1, is calculated to deduce the predictor function. In case of MLR, there would be number of ß1 corresponding to the number of independent variable

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

Anscombe's quartet compromises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.



Explanation of this output:

a) In the 1st one (top left) if we look at the scatter plot we can observe that there seems to be a linear relationship between x & y.

b) In the 2nd one (top right) if we look at this figure we can conclude that there is a non-linear relationship between x & y.

c) In the 3rd one (bottom left) we can say that when there's a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated to be far away from that line.

d) Finally, the 4th one (bottom right) shows an example when one high- leverage point is enough to produce a high correlation coefficient.

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

Pearson R, also called as Pearson's Correlation Coefficient, is a numerical value that indicates the strength (numerical value) and direction (positive or negative) relationship between two variables.

1. The variables may be continuous or categorical.
2. The Pearson's correlation 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 negative slope (i.e., both variables tend to change in different directions).
   - r = 0 means there is no linear association
3. Based on the degree of correlation, an analyst can ascertain whether the given two variables are independent or not, which is a significant whilst undertaking regression modeling

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

Scaling of variables is a method to transform data over a defined scale (min-max range).

Since value of a variable has a large impact on the value of a coefficient, especially, in a linear model; in a multivariable data set with a large number of variables across different value range, retaining original values will lead to large swings in the coefficients associated with that variable. This will lead to difficulty in undertaking a realistic assessment of importance of a particular variable based on the size of its coefficient. Scaling compresses/expands variable data over the defined range and if the same scaling limits are applied to all variables, this will lead to generation of coefficients in a linear model that can be compared easily.

In normalised scaling the variable values are scaled between limits 0 and 1, commonly referred to as Min-Max scaling.

Whereas, in standardised scaling the variable values are scaled to have a mean of 0 and a standard deviation of 1

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

VIF: The variance inflation factor gives how much the variance of the coefficient estimate is being inflated by co linearity.

$$\mathrm{VIF} = \frac{1}{1 - R^2}$$

Where R is the R-square value of that independent variable which we want to check how well this independent variable can be explained perfectly by other independent variables, then it will have a perfect correlation and it’s Rsquared value will be equal to 1.

$$\mathrm{VIF} = \frac{1}{1 - 1^2}$$

which gives VIF =

$$\mathrm{VIF} = \frac{1}{0}$$

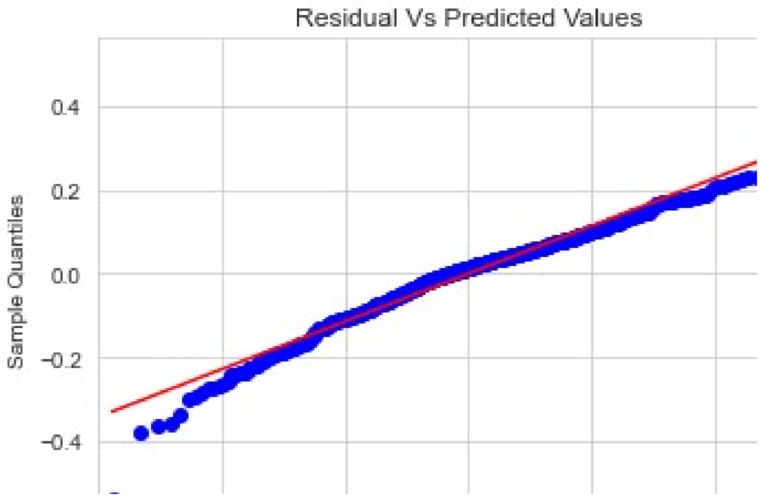
which results in 'Infinity'.

So, If there is perfect correlation, then VIF = Infinity.

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

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. It is used to compare the shape of distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.



If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

The Q-Q plot is used to answer the following questions:
a) Do two datasets come from populations with common distribution?
b) Do two datasets have common location and scale?
c) Do two datasets have similar distributional shapes?
d) Do two data sets have similar tail behavior

In a linear regression, the Q-Q plot is used to reconfirm the distribution of the residuals. They are also used to find the skewness and kurtosis of the distribution. It is also used to validate the basic assumption of a linear regression, i.e, its linear.