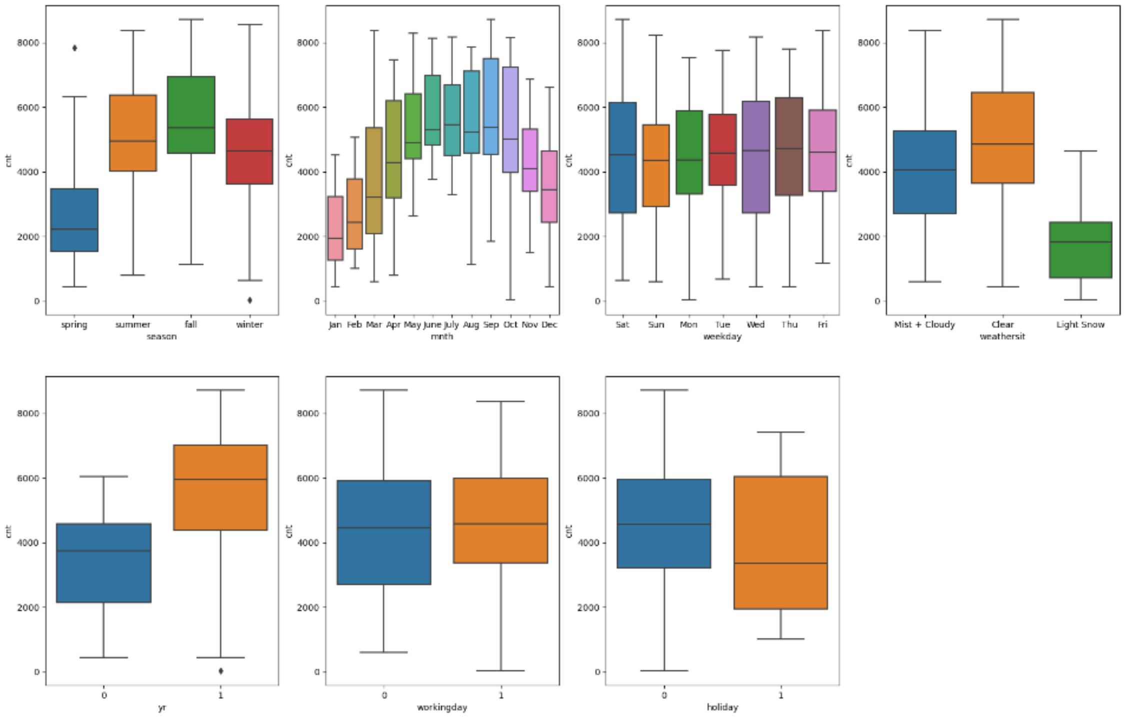
Assignment-based Subjective Questions

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 identified are

- season
- year
- month
- holiday
- weekday
- workingday
- weathersit



##### observations from the plots above

1) People are more likely to rent bikes in the summer and in the fall season

2) Bike rental rates are the most in the months of September and October

3) More bikes rented on Saturday, Wednesday and Thursday

4) Most bike rentals are when the weather is clear

5) More bikes were rented in 2019 than in 2018

6) Bike rental rates are higher on holidays

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

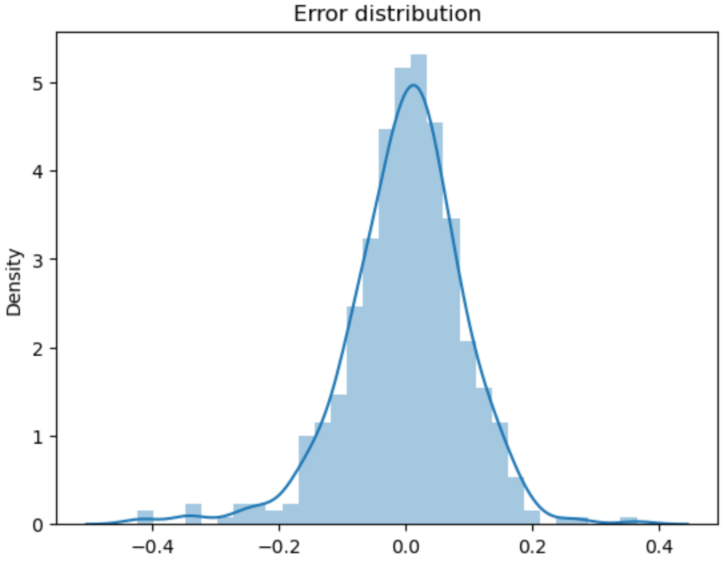
If a categorical variable has N levels it can be represented as N-1 dummy columns, drop_first = True enables the 1 redundant column to be dropped.

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

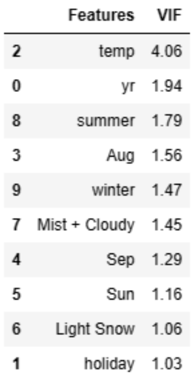
'temp' has the highest correlation with the target variable .

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

By the below plots :
1 – Normal Residuals

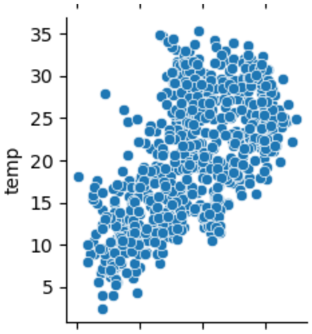


2 – No Multicollinearity (VIF < 5)



	Features	VIF
2	temp	4.06
0	yr	1.94
8	summer	1.79
3	Aug	1.56
9	winter	1.47
7	Mist + Cloudy	1.45
4	Sep	1.29
5	Sun	1.16
6	Light Snow	1.06
1	holiday	1.03

3- Linear relationship (Temp vs cnt is a linear relationship)



4 – No pattern in the residuals

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 contributing significantly towards explaining the demand of the shared bikes are temperature(temp), year (yr), weathersit(Lightsnow)

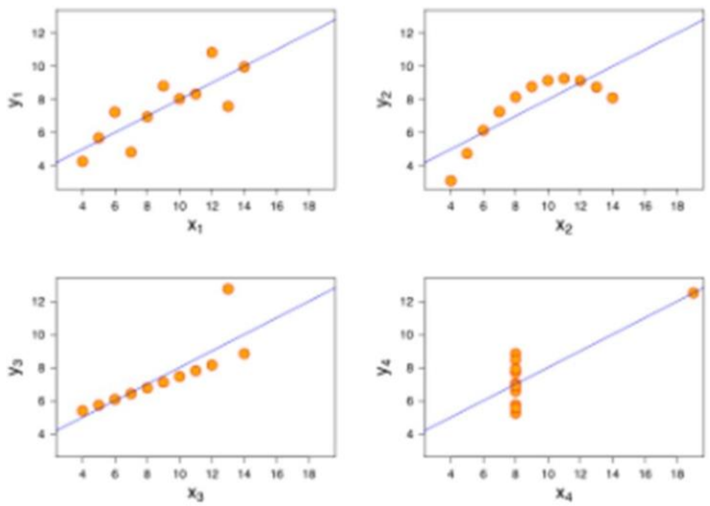
General Subjective Questions

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

Statistical Method used for predictive analysis, where we model to predict 1 dependent variable using 1 or more independent variables

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.



In the above plot, the first one seems to be doing a decent job, the second one clearly shows thatlinear regression can only model linear relationships and is incapable of handling any other kind of data. The thirs and fourth images showcase the linear regression models sensitive to outliers. If outliers are not there, we could have got the great line through the data points. So, we should not run a regression without having a good look at our data.

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

The Pearson correlation coefficient, r, is to understand the strength of linear relationship between 2 variables. It is the ratio between the covariance of two variables and the product of theirstandard deviations

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

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

Scaling is a pre-processing step, applied to independent variables before the model is build. It is use to normalize the data within a particular range.

In case of Multiple Linear Regression when there are lot many variables, many of them might be on very different scales. The model obtained will have varying coefficients which will be very difficult tointerpret. So, scaling is needed for **ease of interpretation**. Another reason for Scaling is **faster convergence of Gradient Descent methods**

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

VIF indicates collinearity between independent variables. If there is a perfect correlation between two independent variables then we get R2 = 1

And if we calculate VIF which is 1/(1-R2) then it will come as infinity.

To solve this, we need to drop one of the variables which is causing Multicollinearity.

This could also happen if the variable with infinite VIF value is expressed perfectly with linearcombination of other independent variables.

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

Q-Q Plot(Quantile-Quantile plot) is a plot of quantiles of 2 distribution against each other. The pattern of points in the plot is used to compare the distributions. If the 2 distributions are similar or linearly related then the points will lie on the line y=x, commonly called as 45-degree reference line. If the points are far from the reference line, the conclusion can be made that the datasets are from different distribution. Below is the q-q plot for Normal Distribution.

Use/Importance in LR – In Linear Regression, when Training and Test datasets are received from different source, we can use Q-Q plot to conclude that both of them are from Populations with samedistribution.