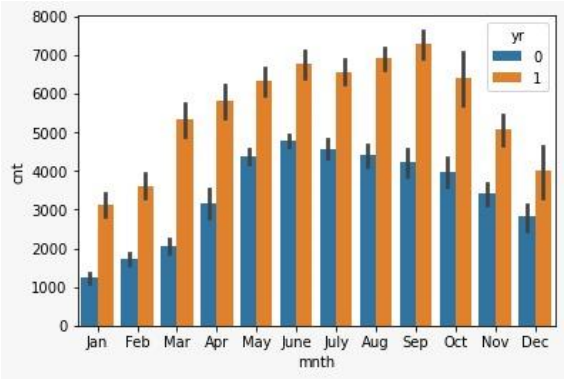
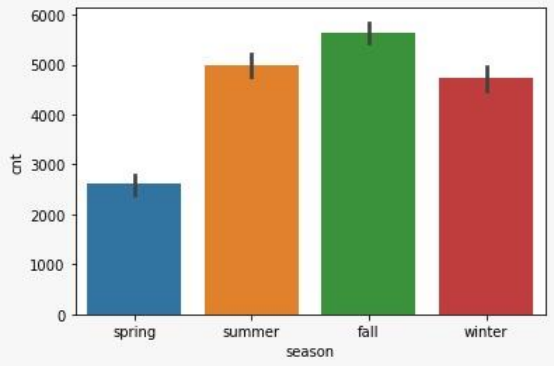
Assignment-based Subjective Questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: 1. In the year of 2019, months from April to October receives higher counts as compared to rest of the months.



2. Summer and fall receive higher count as compared to Spring and Winter.



Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: Drop first is essential when creating dummy variables because it helps to avoid the problem of data multicollinearity. If the first column is not removed, the model estimates will be unstable and biassed due to perfect correlation with the other dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

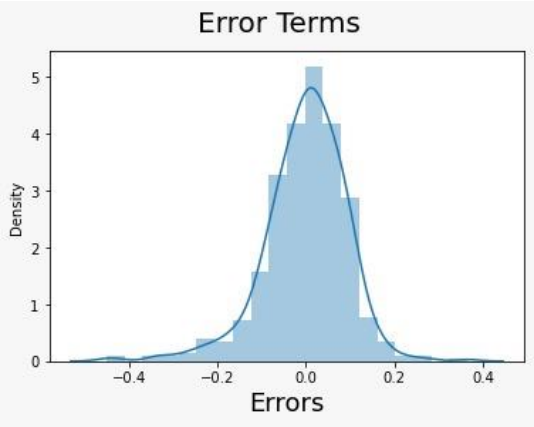
Ans: Looking at the pair plot, temperature(temp) and feeling temperature(atemp) has the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: Assumptions**:**

Linear relationship: With the help of scatter plot and pair plot, we can check the relationship between the independent and dependent variables.

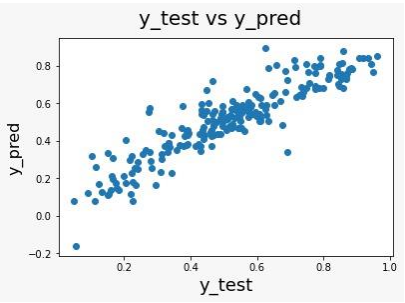
Multivariate normality: With the following graph, we can check this.



No or little multicollinearity: From the correlation matrix and VIFs

	Features	VIF
3	atemp	4.67
2	workingday	3.40
12	winter	2.59
11	spring	2.07
0	yr	2.05
7	Nov	1.79
5	Jan	1.66
10	Mist + Cloudy	1.52
4	Dec	1.46
6	July	1.35
8	Sep	1.20
1	holiday	1.13
9	Light Snow	1.07

Homoscedasticity: constant variance can be checked by scatter plot of y act – y pred



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Ans: Temp, atemp and winter are affecting the target variable majorily.

General Subjective Questions

Q: Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Q: Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a collection of four data sets that are nearly identical in simple descriptive statistics, but have some peculiarities that fool the regression model if built. They have very different distributions and show up differently on scatter plots. It was built in 1973 by statistician Francis Anscombe to demonstrate the significance of plotting graphs before analysing and modelling, as well as the impact of other observations on statistical properties. There are four data set plots that have nearly identical statistical observations and provide the same statistical information, which includes the variance and mean of all x,y points in all four datasets.

Q: What is Pearson's R?

The Pearson correlation coefficient is a descriptive statistic, which means it summarises a dataset's characteristics.

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson's r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)

- The correlation coefficient.

It describes the strength and direction of a linear relationship between two quantitative variables in particular. Although different disciplines have different interpretations of relationship strength (also known as effect size). In addition, the Pearson correlation coefficient is an inferential statistic, which means it can be used to test statistical hypotheses. We can specifically test for a significant relationship between two variables.

Q: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It is a data Pre-Processing step that is applied to independent variables in order to normalise the data within a specific range. It also aids in the speeding up of algorithm calculations.

Most of the time, the collected data set contains features with widely disparate magnitudes, units, and ranges. If scaling is not performed, the algorithm only considers magnitude rather than units, resulting in incorrect modelling. To solve this problem, we must scale all of the variables to the same magnitude level.

Normalized Scaling - It gathers all data between 0 and 1. sklearn.preprocessing. MinMaxScaler aids in the implementation of normalisation in Python. Values are replaced by their Z scores after standardisation.

Standardized Scaling - It transforms the data into a standard normal distribution with a mean () of zero and a standard deviation of one (). sklearn.preprocessing. Python's scale aids in the implementation of standardisation. One disadvantage of normalisation over standardisation is that it removes some data information, particularly about outliers.

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

The value of VIF is infinite when there is perfect multicollinearity in the data, meaning that the predictor variables are perfectly correlated with each other, making it impossible to determine the unique effect of each variable on the outcome.

It is important to note that scaling has no effect on the other parameters such as t-statistic, F-statistic, p-values, R-squared, and so on.

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

A Q-Q plot (Quantile-Quantile Plot) is a graphical method to check if a set of data is approximately normally distributed. It plots the sample quantiles against the theoretical quantiles of a normal distribution.

In linear regression, a Q-Q plot is used to check the assumption of normality of residuals, which is important for valid inference. If the residuals are not normally distributed, it can affect the validity of statistical tests and confidence intervals for the regression coefficients. A Q-Q plot helps to visually inspect if the residuals are close to a straight line, indicating normality, or if there are deviations, suggesting non-normality.