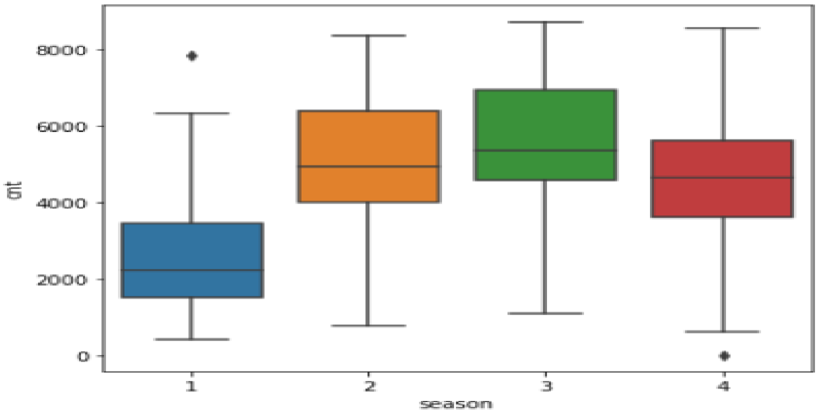
Assignment-based Subjective Questions

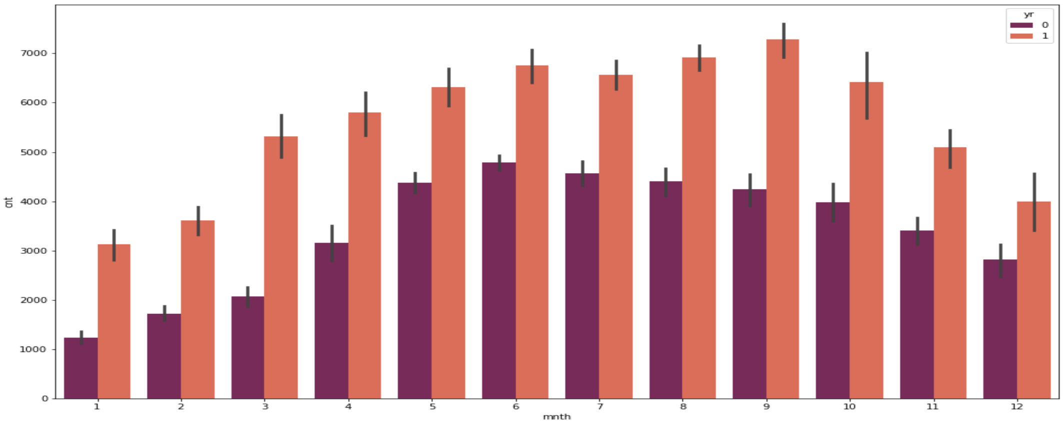
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Ans → In the Bike sharing dataset we have categorical variable like season, month, weekday and weathersit. As we can notice from below boxplot, categorical variables are having significant amount of effect on dependent variable 'cnt'.

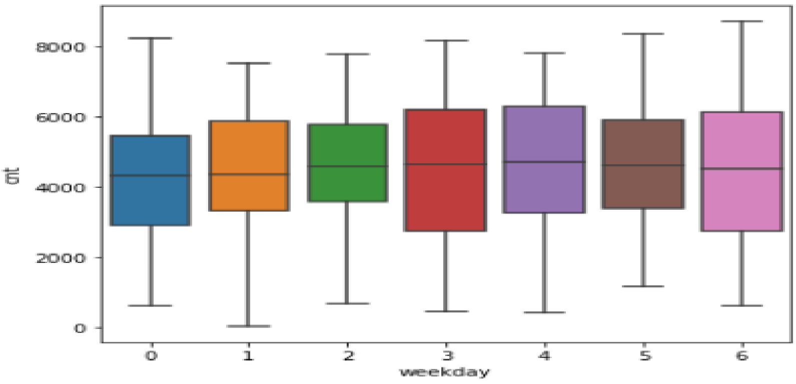
I. Season – We can notice that, Season_3 has very good amount of bike demand as compared to other seasons. And season_1 has very less amount of bike demand among all seasons.



II. Month(mnth) – From below visualisation, we can easily conclude that we have more no. demand on shared bike (dependent variable-cnt) in month_9 followed by month_8 in year 2019. And in 2018, we have more no. of demand on shared bike in month_6 followed by month_7.



III. Weekday – From the below boxplot we can conclude that dependent variable (cnt) has fair amount of bike rented on weekday_3 followed by weekday_6. One thing we can notice that on each day median is same for dependent variable.



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

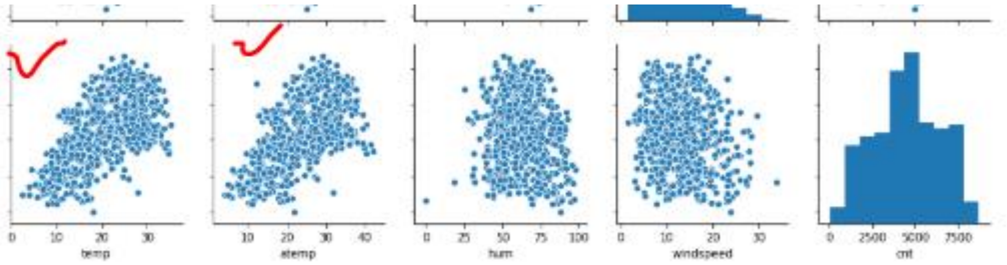
Ans → It's important to use drop_first=True during dummy variable creation to remove redundancies. Default value of drop_first is False. This parameter is used in get_dummies() method. If we use drop_first=False then we will get redundant features in the model.

For ex. If we have variable 'Gender' then we don't need to keep both the 'male' and 'female' category in dummy variable. We just need only 1 variable value to define the gender. If male=1 then the person gender is Male and if male=0 then person gender is Female.

In this way we can remove any one feature to make model redundant and keep only one feature and another one is just opposite to other feature.

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

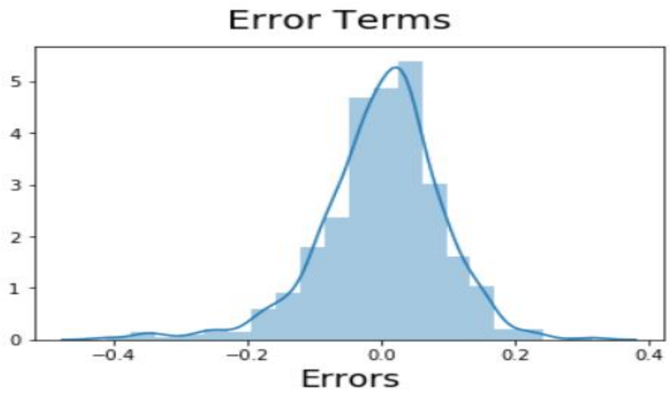
Ans → atemp and temp independent variables are the highest correlation with the target variable count(cnt), though we have 'temp' and 'atemp' are multicollinear in nature. 'atemp' variable is not selected in automatic feature selection (RFE) so the final model doesn't have any multicollinearity.



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

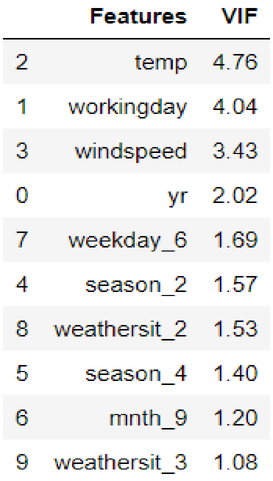
Ans → There is a linear relationship between independent and target variables in training dataset. Below are the assumptions of linear regression after building the model on training set.

i. Errors(residuals) follow a normal distribution with mean 0 in training set.



ii. There is a linear relationship between independent and dependent variables. For ex. In Bike sharing assignment, Temp is highly correlated with dependent variable count.

iii. No multicollinearity between the independent's variable. As we can see in linear model, we don't have any variables which VIF value is greater than 5 in training set.



	Features	VIF
2	temp	4.76
1	workingday	4.04
3	windspeed	3.43
0	yr	2.02
7	weekday_6	1.69
4	season_2	1.57
8	weathersit_2	1.53
5	season_4	1.40
6	mnth_9	1.20
9	weathersit_3	1.08

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

Ans → Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

i) **Temp** – The coefficient value of temp 0.5499 units indicated that if the value of temp variable increased then no. of bike booking also increased by 0.5499 units.

ii) **Year** – The coefficient value of year 0.2331 units indicated that if value of year variable increased then no. of bike booking also increased by 0.2331 units.

iii) **Season_4** – The coefficient value of season_4 0.1318 units indicated that season_4 having more no. of bike booking increased by 0.1318 units as compared to other seasons.

General Subjective Questions

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

Ans → Regression analysis is a form of predictive modelling techniques which defines the relationship between one dependent and one or many independent variables. Types of Regression are linear regression, logistics regression, polynomial regression and stepwise regression.

Linear regression is a machine learning algorithm based on supervised learning methods. It performs a regression task to compute the regression coefficients. It is one of the easiest statistical models. It’s used to predict values within a continuous range rather than trying to classify them into categories.

Linear regression performs a task to predict a dependent variable value (y) based on given independent variable (x). In this technique find out a linear relationship between input (x) and output (y). Equation of linear regression is y=mx+c, where y is dependent variable, x is independent/predictor variable, c is intercept and m is slope of best fit line. This is nothing but mathematical equation of straight line.

Linear regression is classified into 2 types namely Simple Linear Regression and Multiple Linear Regression.

a) **Simple Linear Regression** – There is a linear relationship between one independent variable (x) and one dependent variable (y).
y=β0+β1*x using this equation we need to find the best value for β0 and β1.

b) **Multiple Linear Regression** – There is a linear relationship between more than one variable (x1, x2…. xn) and dependent variable(y).
y=β0+β1*x1+ β2*x2+……+βn*xn+ε.

We need to update β0 and β1 values to get the best fit line.

Cost Function – The cost function helps us to figure out the best possible values for β0 and β1 which would provide the best fit line for data points. Since we need the best values for β0 and β1 so we need to convert search problem into minimisation problem we would like to minimise the error between predicted value and actual value.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

We choose the above formula to minimize. The difference between the predicted values and ground truth measures the error difference. We need to square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error (MSE) function. Now, using this MSE function we are going to change the values of β0 and β1 such that the MSE value settles at the minima.

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value and true y value.

Gradient Descent - The next important concept needed to understand linear regression is gradient descent. Gradient descent is a method of updating β0 and β1 to reduce the cost function (MSE). The idea is that we start with some values for β0 and β1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.

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

Ans → All four sets are identical when examined using simple summary statistics, but vary considerably when graphed. Anscombe's quartet comprises 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.

The datasets are as follows. The *x* values are the same for the first three datasets. It is not known how Anscombe created his datasets.

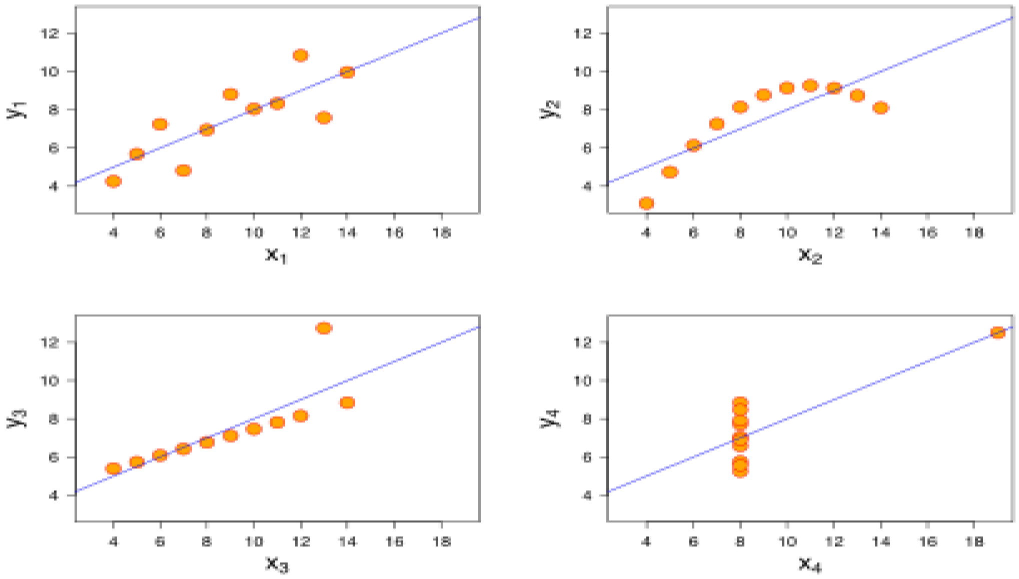
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

For all the four datasets, the summary statistics show that the means and the variances were identical for x and y across the groups.

Property	Value
Mean of *x* in each case:	9 (exact)
Variance of *x* in each case:	11 (exact)

Mean of *y* in each case:	7.50 (to 2 decimal places)
Variance of *y* in each case:	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case:	0.816 (to 3 decimal places)
Linear regression line in each case:	*y* = 3.00 + 0.500*x* (to 2 and 3 decimal places, respectively)

When plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story.



- The first scatter plot x1 (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph x2 (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph x3 (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph x4 (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

3. What is Pearson’s R? (3 marks)

Ans → 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
There are few assumptions on Pearson’s R correlation coefficient.

i) For the Pearson r correlation, both variables should be normally distributed. i.e. the normal distribution describes how the values of a variable are distributed. This is sometimes called the ‘Bell Curve’ or the ‘Gaussian Curve’.

ii) There should be no significant outliers. We all know what outliers are but we don’t know the effect of outliers on Pearson’s correlation coefficient, r. Pearson’s correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means, including outliers in your analysis can lead to misleading results.

iii) Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

iv) The two variables have a linear relationship. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric.

v) The observations are paired observations**.** That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable. For example, if you’re calculating the correlation between age and weight. If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks.

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

Ans → Scaling means that we need to transform data so that it fits within a specific scale, like 0–100 or 0–1. It is a method used to standardise the range of features of data. This step is required in data pre-processing while using machine learning algorithm.

We need to perform scaling for ease of interpretation of data and faster convergence for gradient descent method. Scaling the variables is an important step because we have noticed in bike sharing dataset, the variables ‘temp’, ’atemp’, ’cnt’, etc. are on different scale with respect to all other numerical variables which take very small value. Also, the categorical variables that we encoded earlier take either 0 or 1 as their values. Hence, it's important to have everything on the same scale for the model to be easily interpretable.

There are 2 popular scaling methods. (i) Normalized scaling and (ii) MinMax scaling

The major difference between standardized scaling over the normalized scaling is that it doesn't compress the data between particular range as in MinMax scaling. This is useful, especially if there is are extreme data points (outliers). Simply, the standardized scaling brings all the data into standard normal distribution with mean 0 and standard deviation 1, while MinMax scaling brings all the data in the range of 0 to 1.

MinMax Scaling x = (x–xmin)/(xmax–xmin)

Standardize Scaling x = (x–xmin)/(xmax–xmin)

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

Ans → VIF (Variance Inflation Factor) defines the correlation between one predictor variable and on the other predictor variable in the model. VIF calculates how well one independent variable is explained by all other independent variables combined. It's useful for solving the multicollinearity problem. The formula of VIF is VIF= 1/(1-R^2). The VIF value < 5 indicates the good sign and no need to eliminate the variable.

Sometimes we noticed that VIF value is infinite. This is indicating that the corresponding variable may be expressed exactly by a linear combination of other variables. Also, R^2 value is 1 i.e. there is perfect correlation between the variables. To solve this problem, we need to drop the variables and need go with RFE method.

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

   Ans → Definition - Q-Q (Quantile-Quantile) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as Normal, exponential. It also helps to determine if two data sets come from populations with a common distribution.

   Use - This plot useful in a scenario of linear regression when we have training and test dataset received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. Also, used to assess if residuals of model are normally distributed.

   Importance - Q-Q plots are used to visually check that your data meets the homoscedasticity and normality assumptions of linear regression. Q-Q plots let you check that the data meet the assumption of normality. They compare the distribution of your data to a normal distribution by plotting the quartiles of your data against the quartiles of a normal distribution. If your data are normally distributed then they should form an approximately straight line. Below are the possible interpretations for the two datasets.

   a. Similar distribution - If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.

b. X-value < Y-value - If x-quantiles are lower than the y-quantiles.

c. Y-value < X-value - If y-quantiles are lower than the x-quantiles.

d. Different distribution - If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis.