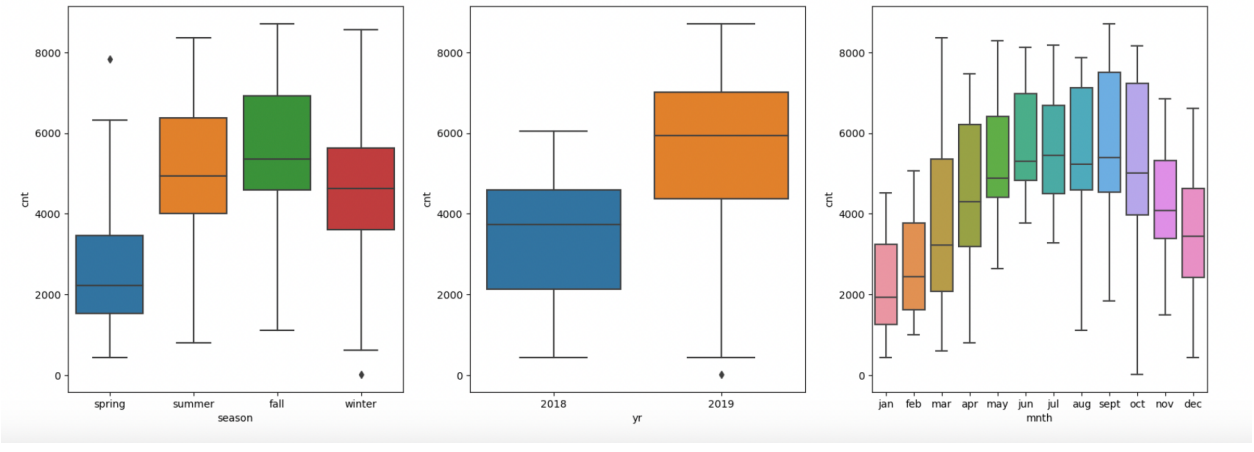
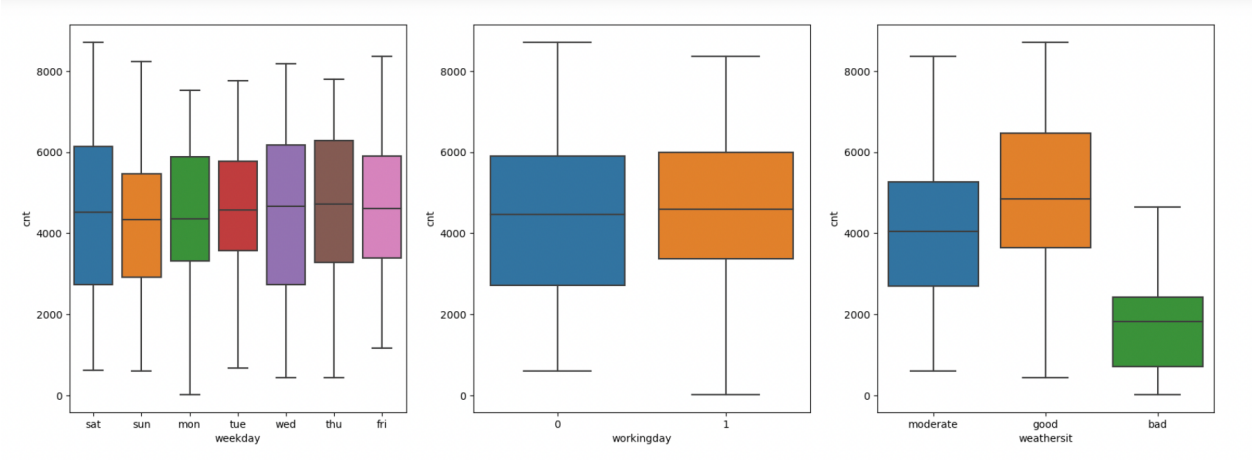
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

ANSWER:





Observations:

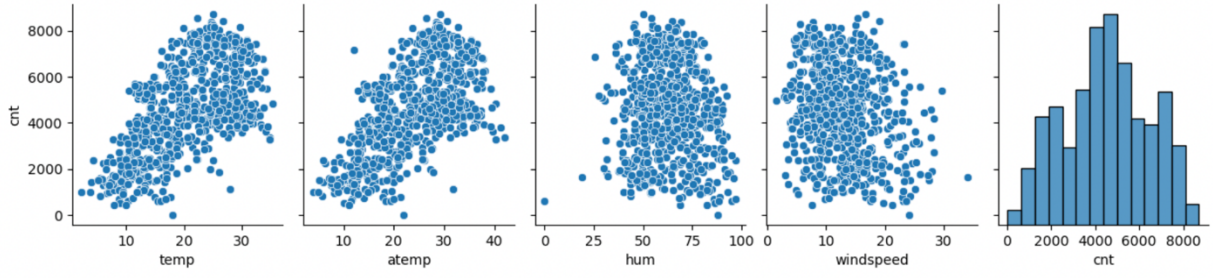
1. SEASON AND CNT: Seasons have a significant effect on the target variable Cnt.
   Cnt is highest in fall and lowest in spring
2. YEAR AND CNT: Cnt is more in 2019 than in 2018. It may be inferred that the demand is increasing with time.
3. MONTH AND CNT: Cnt is highest in sept, oct and lowest in jan, feb , march
4. WEEKDAYS AND CNT: No significant pattern observed.
5. WORKINGDAY AND CNT: More variation on non-working day but the median and highest demand is almost same.
6. WEATHERSIT AND CNT: CNT high on good-weathersit and low on bad-weathersit.

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

ANSWER: *The intention behind the dummy variable is that for a categorical variable with 'n' levels, we create 'n-1' new columns each indicating whether that level exists or not using a zero or one.*
Hence drop_first=True is used so that the resultant can match up n-1 levels.
Eg: If there are 3 levels, the drop_first will drop the first column.

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

ANSWER: temp and atemp have highest correlation with cnt



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

ANSWER: Multiple Linear Regression model assumptions

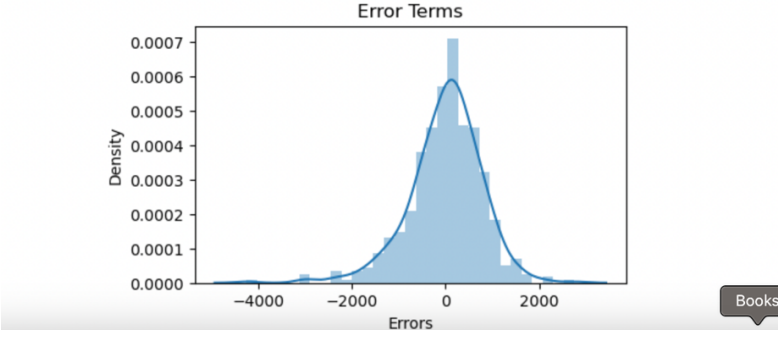
1. X and Y should have linear relationship
2. All independent variables should not be correlated with each other
3. Error terms should be normally distributed with mean at 0
4. Error terms should have constant variance

All independent variables should not be correlated with each other: This has been taken care by removing the variables with high VIF which signifies the correlation

Error terms are normally distributed has been verified during residual analysis:

```
In [55]: #Plot a histogram of the error terms
         def plot_res_dist(act, pred):
             sns.distplot(act-pred)
             plt.title('Error Terms')
             plt.xlabel('Errors')
```

```
In [69]: plt.figure(figsize=(5,3))
         plot_res_dist(y_train, y_train_pred)
```



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
ANSWER: The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. temp
2. yr
3. season_winter

General subjective questions:

Q1. Explain the linear regression algorithm in detail. (4 marks)

ANSWER: A liner regression is a type of supervised machine learning algorithm and it attempts to explain the relationship between a dependent and an independent variable using a straight line. The independent variable is also known as the predictor variable and the dependent variable is also known as the output variable. The relation is usually a straight line that best fits the different data points as close as possible. The output is a numerical variable like sales, age etc.

$y= \beta 0+ \beta\ 1x+ \varepsilon$

Here,

- Y= Dependent Variable
- X= Independent Variable
- β 0= intercept of the line
- $\beta 1$ = Linear regression coefficient (slope of the line)
- ε = random error

Types of Linear Regression

Linear Regression can be broadly classified into two types of algorithms:

1. Simple Linear Regression

    It has one independent and one dependent variable
    Equation is of the form: y = mx+c

2. Multiple Linear Regression

    When a number of independent variables more than one
    Equation is of the form: y= c+m1x1+m2x2… mnxn

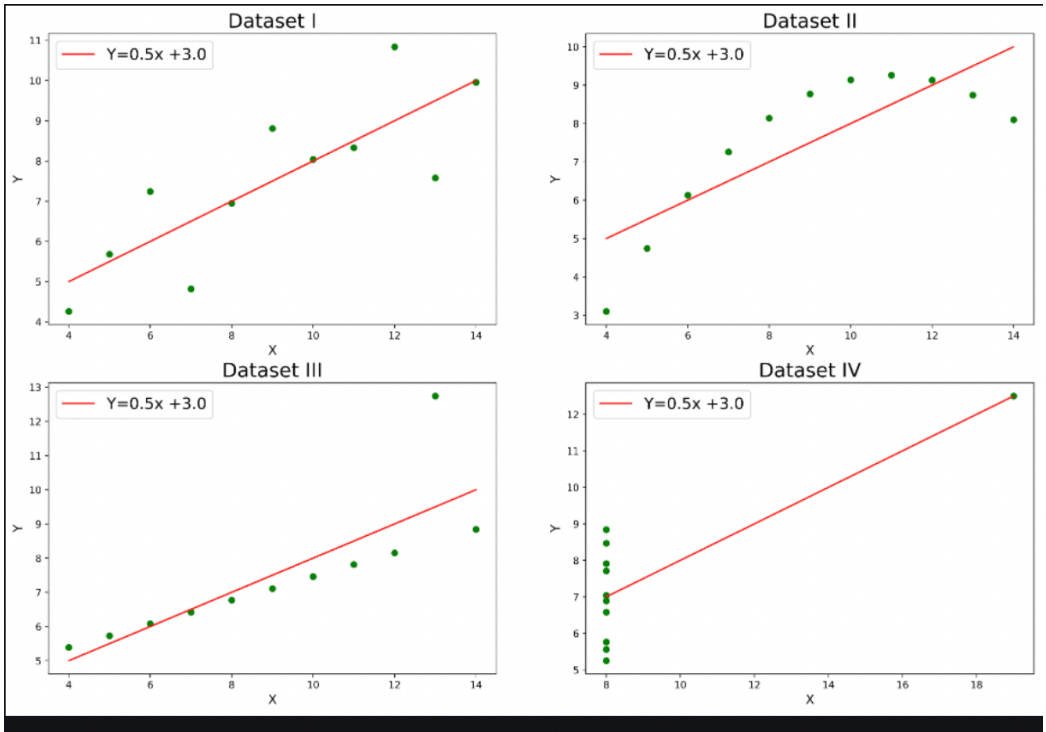
3. Non-Linear Regression

When the best fitting line is not a straight line but a curve, it is referred to as Non-Linear Regression.

Q2. Explain the Anscombe's quartet in detail. (3 marks)

ANSWER: Anscombe's quartet is used to explain the importance of Data visualisation. It contains a set of four dataset having identical descriptive analysis like mean, variance etc but showing very different trends when the data is plotted.

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.1	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.1	5.39	12.5
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89



Anscombe's quartet Plot

● 1st data set fits linear regression model as there seems to be linear relationship between X and y
● 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
● 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
● 4th data set shows some outliers present in the dataset which can't be handled by a linear regression model.

Conclusion: All the important features in the dataset must be visualised before applying any machine learning algorithm to create a good model.

Q3. What is Pearson's R? (3 marks)

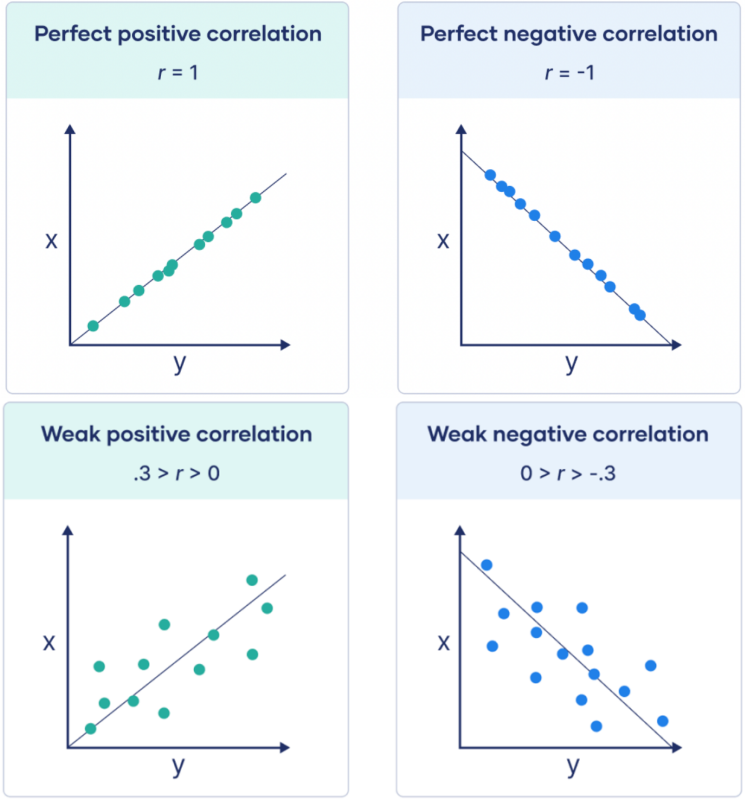
ANSWER: Pearson correlation coefficient (also known as Pearson's r) is a most common way of measuring linear correlation. It varies between -1 and 1 and measures the strength of relationship between two variables.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Pearson Coefficient Range	Correlation Type	Interpretation
Between 0 and 1	Positive correlation	When one variable increases, the other variable also increases and vice versa. Closer to 1 means higher correlation.
0	No correlation	The variables are not related to one another
Between -1 and 0	Negative correlation	When one variable increases, the other variable decreases and vice versa. Closer to -1 means higher negative correlation.



Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

ANSWER: In datasets we have features that have different range and magnitudes. In order for any machine learning algorithm to determine the importance of these variables without bias, the features should be on the same scale. Unscaled data can adversely affect the model's ability to make accurate predictions as it tends to weigh high on the features with higher values , therefore it is very important to perform the scaling while pre-processing the data.

Normalised scaling (Min Max Scaling) vs Standardised Scaling

Normalised scaling: It is a scaling technique where all values are rescaled into the range from 0 to 1

Standardised Scaling: It is a scaling technique where the values in a column as rescaled so that they have mean = 0 and variance = 1

Formulae:

Normalised scaling: $x = (x - \min(x))/(\max(x) - \min(x))$

Standardised scaling: $x = (x - \min(x))/ (\max(x) - \min(x))$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)

ANSWER: VIF(Variance Inflation Factor) measures the severity of multicollinearity in the ordinary least square regression analysis.
The formulation of VIF is given below:

$$VIF_i = \frac{1}{1 - R_i^2}$$

A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. A very high VIF value shows a perfect correlation between two independent variables.
In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To resolve this issue, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANSWER. Quantile-Quantile plot or Q-Q plot is a graphical tool which can help us analyse if two datasets come from populations with a common distribution. It is basically a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. A 45-degree reference line is also plotted, if the datasets are sampled from a common distribution, then the points should approximately fall on this line. The greater the distance of the points from this reference line, the greater the likelihood that the datasets come from a different distribution.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For ex, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For ex, if the two datasets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.