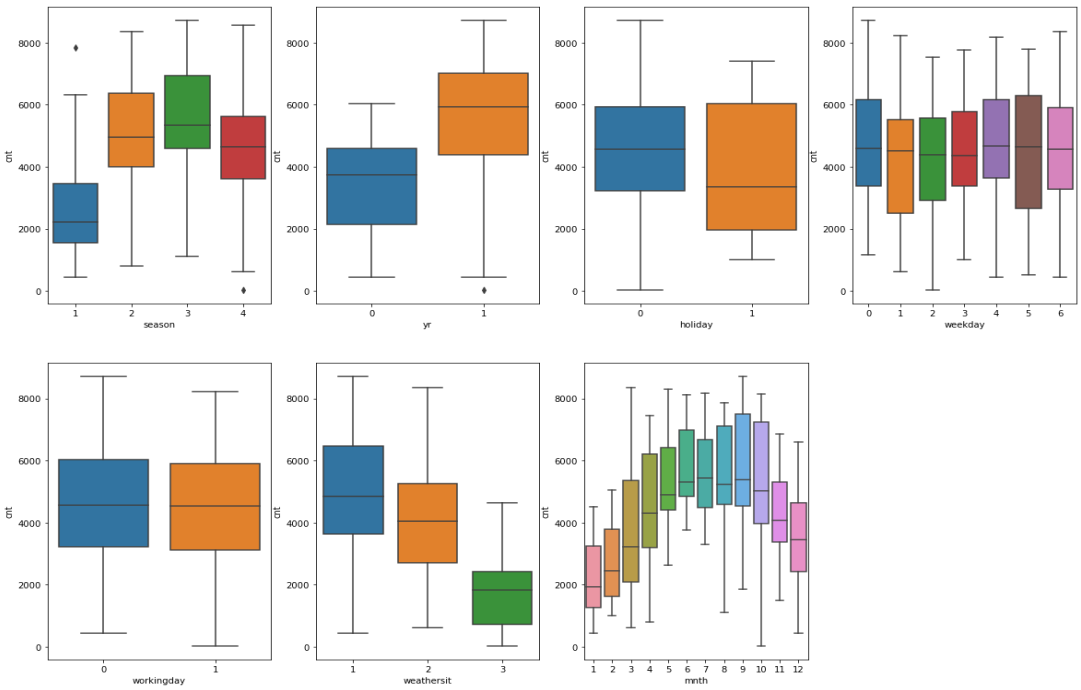
Assignment-based Subjective Questions :

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 variable in the dataset were season , yr , holiday, weekday ,workingday, and weathersit and mnth . These were visualized using a boxplot (Fig. attached) .

- These variables had the following effect on our dependant variable:-
  - Season - The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
  - Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable. Highest count was seen when the weathersit was' Clear, Partly Cloudy'.
  - Yr - The number of rentals in 2019 was more than 2018
  - Holiday - rentals reduced during holiday.
  - Mnth - September saw highest no of rentals while December saw least. This observation is in accordance with the observation made in weathersit. The weather situation in December is usually heavy snow due to which the rentals might have dropped.
  - Weekday - The count of rentals is almost even throughout the week
  - Workingday – The median count of users is constant almost throughout the week.

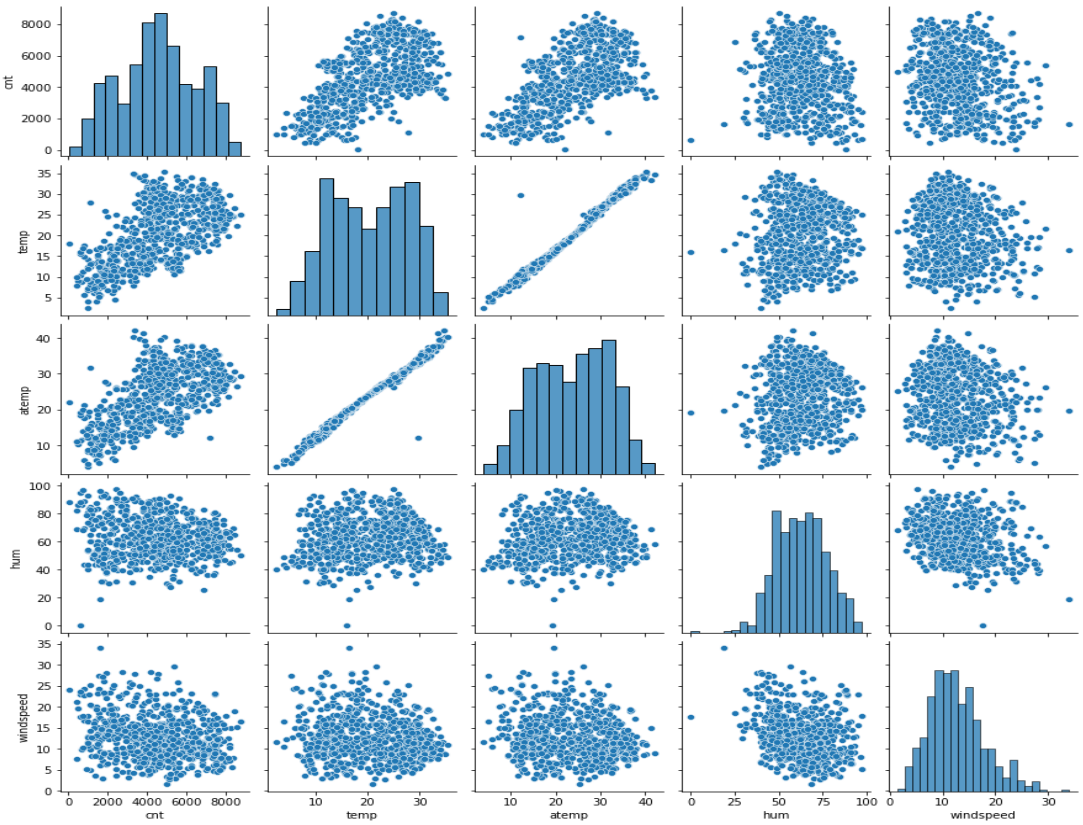
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- Drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then it is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

- Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

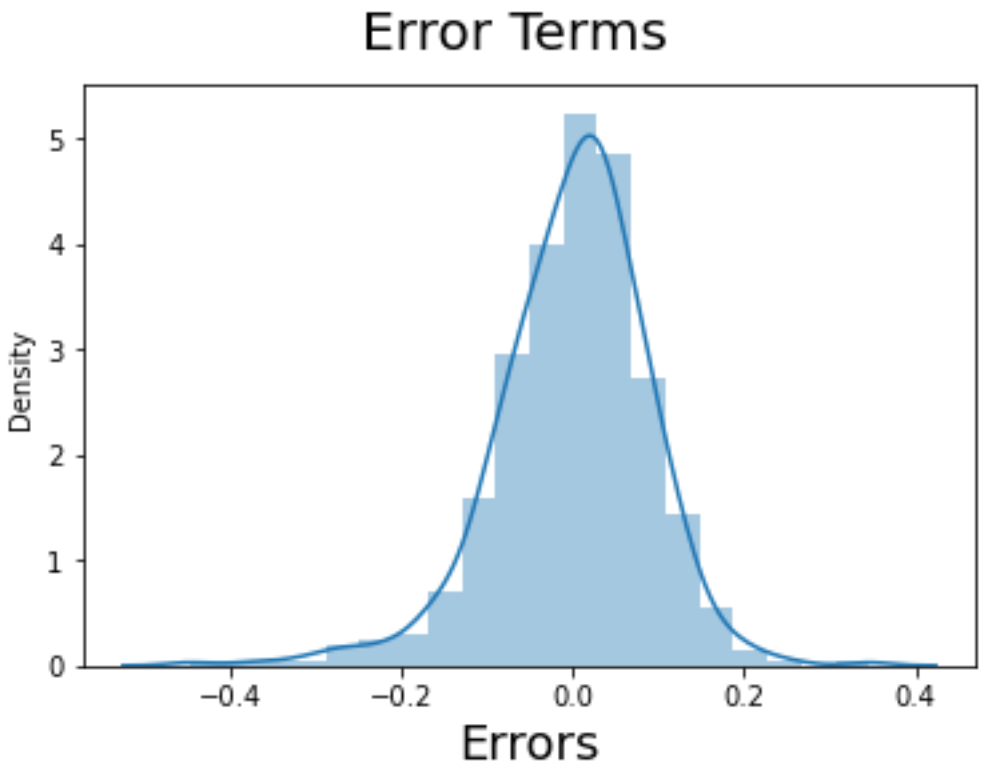


- Observing the above pairplot it can be seen that, "temp" and "atemp" are the two numerical variables which are highly correlated with the target variable "cnt".

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

The following tests were done to validate the assumption of linear regression:

- Linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not.
- Residual distribution should follow normal distribution and centred around 0 or mean = 0. We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.



- Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF to get the quantitative idea about how much the feature variables are correlated with each other in the new model.

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

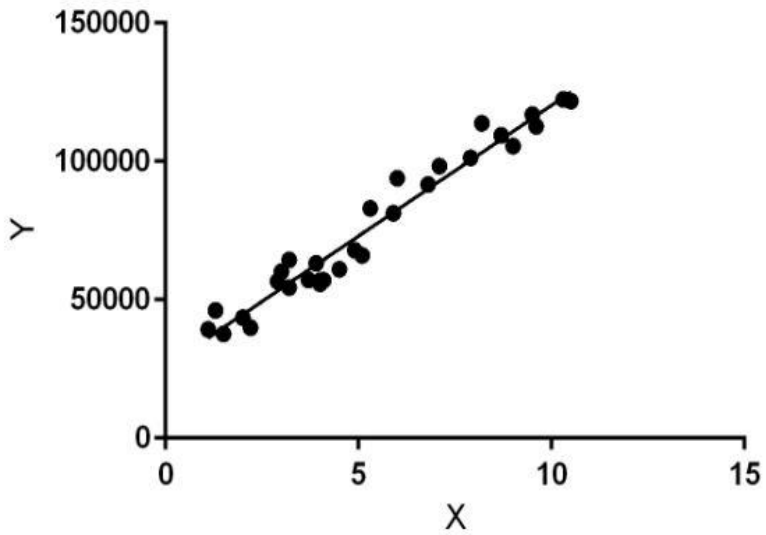
The top 3 features:

- temp – coefficient = 0.437655
- yr – coefficient = 0.234287
- weathersit_Light Snow & Rain – coefficient = -0.292892

General Subjective Questions

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

- 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.

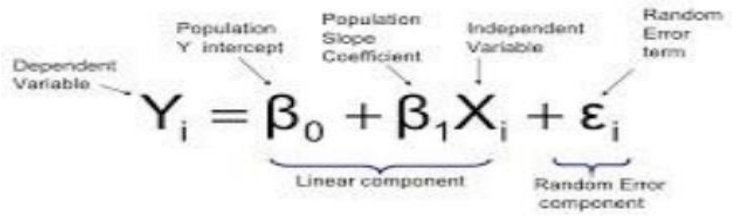


- Linear regression performs the task to predict a dependent variable value (Y) based on a given independent variable (X). So, this regression technique finds out a linear relationship between X (input) and Y (output). Hence the name is Linear Regression.
- In the figure above, X(input) is the work experience and Y(output) is the salary of a person. The regression line is the best fit line for out model.
- So basically Regression method tries to find the best fit line which shows the relationship between the dependent variable and independent variable (predictors) with least error.

  Now Regression is broadly divided into simple linear regression and multiple linear regression.

  - Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.

    The equation for SLR will be:



- Multiple Linear Regression : MLR is used when the dependent variable is predicted using multiple independent variables.

  The equation of MLR will be:

$$observed\ data \rightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon$$

$$predicted\ data \rightarrow \quad y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

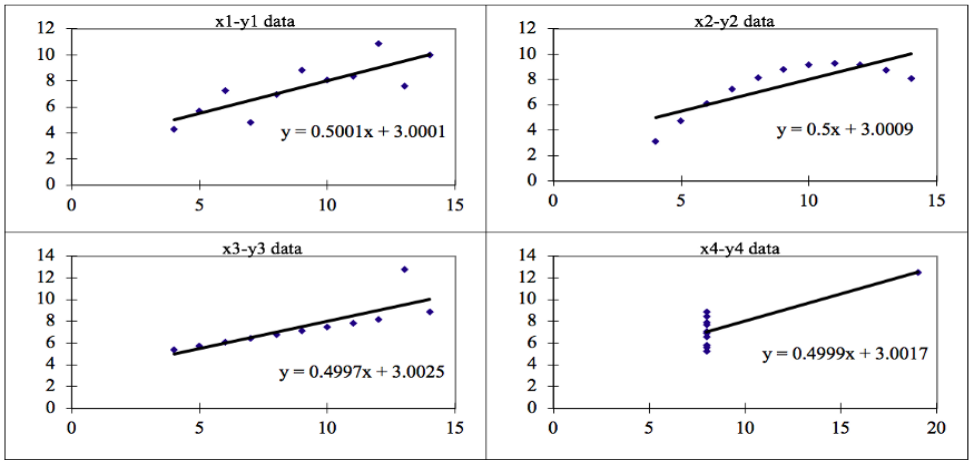
$$error \rightarrow \quad \varepsilon = y - y'$$

B1 = coefficient for x1 variable
B2 = coefficient for x2 variable
B3 = coefficient for x3 variable and so on….
B0 is the intercept (constant term)

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

- Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the Importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	



- Dataset 1 : This fits the linear regression model pretty well.
- Dataset 2 : This could not fit linear regression model on the data quite well as the data is not-linear.
- Dataset 3 : shows the outliers involved in the dataset which cannot be handled by linear regression model.
- Dataset 4 : shows the outliers involved in the dataset which cannot be handled by linear regression model

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

- Pearson's R is also reffered as the (Pearson's Correlation Coefficient). It is a statistic that measures the linear correlation between two variables. Like all correlation, it also has a numerical value that lies between (-1.0 to +1.0).
- It shows the linear relationship between two sets of data.
- In simple terms it tells us "can we draw a line graph to represent the data?"
- It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
- Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given 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

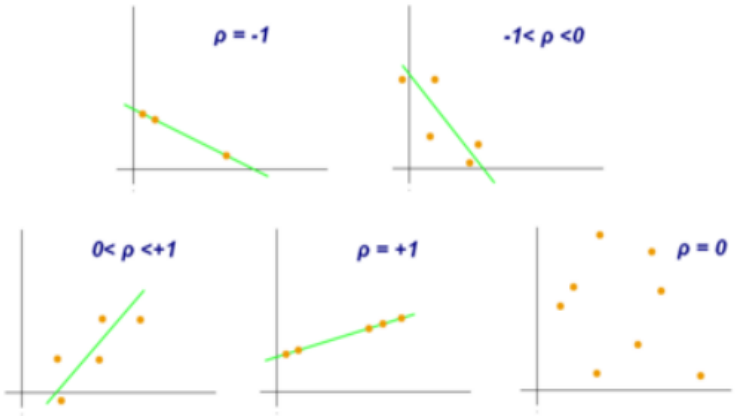
x_i = values of the x-variable in a sample

$\bar{x}$ = 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

- We can understand it better by using graph for example



As can be seen from the graph above,

- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association

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

- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S. NO.	Normalisation	Standardisation
1.	Minimum and maximum value of feature are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0,1] or [1,-1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called	Scikit-Learn provides a transformer called (StandardScaler) for standardization.

	(MinMaxScaler) for Normalization.	
6.	This transformation squishes the n-dimensional data into an n-dimentional unit hypercube.	It translets the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution.	It is useful when the feature distribution is normal or Gaussian.
8.	It is often called as scaling Normalization.	It is often called as Z-score Normalization.

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

- VIF = Variance Inflation Factor
- The VIF gives us how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

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

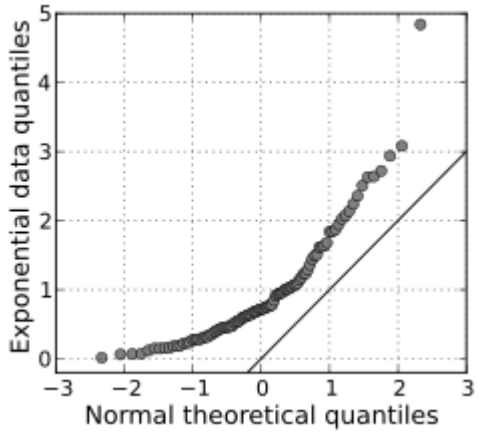
- Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will perfect correlation and it's R-squared value will be equal to 1.
- So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "Infinity".
- The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity. If there was no correlation with other predictors.

  Rule for interpreting the variance inflation factor:
  - 1 = not correlated.
  - Between 1 & 5 = moderately correlated.
  - Greater than 5 = highly correlated

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

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
- The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot. If the two data sets come from a common distribution, the points will fall on that reference line.

- A Q-Q plot showing the 45 degree reference line:



- If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line Y = X. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line Y = X.
- Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
- A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distribtions.