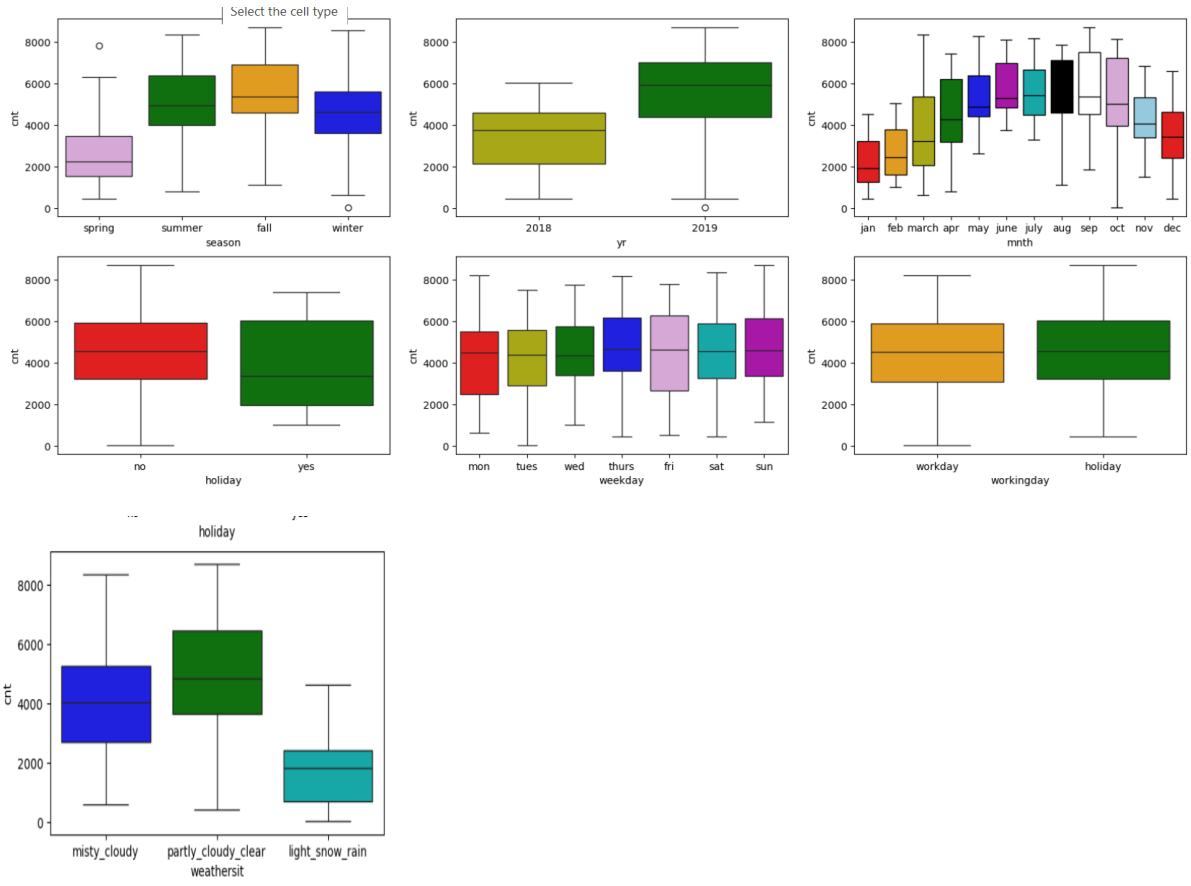
Assignment-based Subjective Questions

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



A: We can make the following inferences about the dataset:

- Season: The count of customers seems to be highest during the fall and summer season and lowest during the spring season.

- Year: The number of customers was higher in 2019 as compared to 2018.

- Month: The count of customer increased steadily from January to June, stabilized for a few months while peaking in September, then it dropped again in November and December.

- Holiday: The count of customers seems to higher on days which were not holidays than those which were holidays.

- Weekday: The count of customers was lower on Monday, Tuesday, Wednesday as compared to Thursday, Friday, Saturday and Sunday.

- Workingday: The count of customers seems to be almost equal to both holidays and working days.

- Weather situation: The count of customers is highest on clear/partly cloudy days and lowest on days with light rainy/snowy weather.

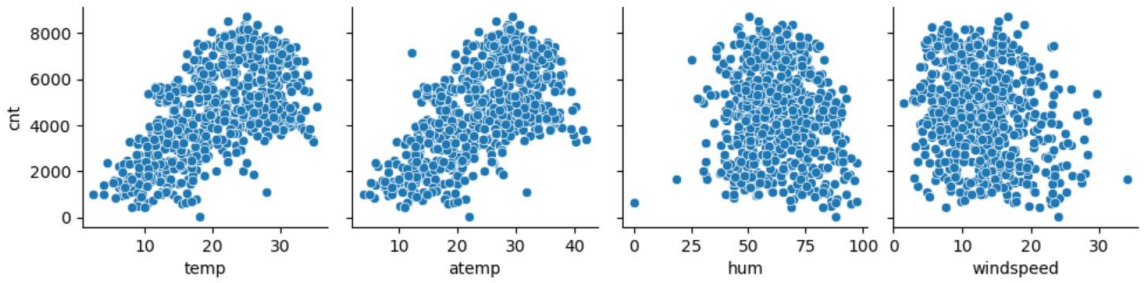
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

A: When dealing with dummy variable creation of categorical features, a common convention is to drop one of the new columns from each feature. The argument comes from statistics: without dropping a column, we know that the sum of all these columns will be 1 in every row. For example, encoding gender

as two variables, is_male and is_female, produces two features which are perfectly negatively correlated. This is called the dummy variable trap: perfect multicollinearity between the predictors.

If we decide to drop the first column, the algorithm drops the category value name that comes first alpha-numerically in the set. In the gender example, the first column would be Female since F comes before M. In this way, we can avoid the dummy variable trap and multicollinearity among predictor variables.

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



A: After plotting the pair-plot among the numerical variables, it was observed that temp and atemp variables have 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)

A: Assumptions for Linear Regression are:

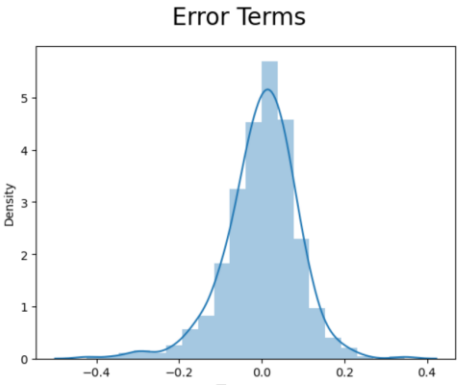
1. There is a linear relationship between X and Y:

X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.

This assumption was verified through a pairplot between the target (cnt) and predictor variables.

2. The error terms should be normally distributed with mean as zero.

This assumption was tested by plotting the residuals and observing their distribution around the mean.



3. The error terms are independent of each other.

This can be checked by:

- Checking multicollinearity among the predictor variables using the VIF (Variable Inflation Factor) values. VIF should be less than 5 for all predictor variables included in the model.
- Plotting scatterplots of residuals vs. predicted values to ensure no patterns exist.

4. The error terms have constant variance.

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

A: The top 3 features contributing significantly towards explaining the demand of the shared bikes are temp, 2019 (yr) and partly_cloudy_clear (weathersit).

General Subjective Questions

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

Machine Learning is a branch of Artificial intelligence that focuses on the development of algorithms and statistical models that can learn from and make predictions on data. **Linear regression** is also a type of machine-learning algorithm more specifically a **supervised machine-learning algorithm** that learns from the labelled datasets and maps the data points to the most optimized linear functions, which can be used for prediction on new datasets.

Supervised machine learning algorithms are a type of machine learning algorithm where the algorithm learns from labelled data. Labeled data means the dataset whose respective target value is already known. Supervised learning has two types:

- **Classification**: It predicts the class of the dataset based on the independent input variable. Class is the categorical or discrete values. like the image of an animal is a cat or dog?
- **Regression**: It predicts the continuous output variables based on the independent input variable. like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.

When there is only one independent feature, it is known as Simple Linear Regression, and when there is more than one feature, it is known as Multiple Linear Regression. A **simple linear regression** model (SLR) attempts to explain the relationship between a dependent variable and an independent one using a straight line.

The independent variable is also known as the predictor variable, and the dependent variables are also known as the output variables.

The primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

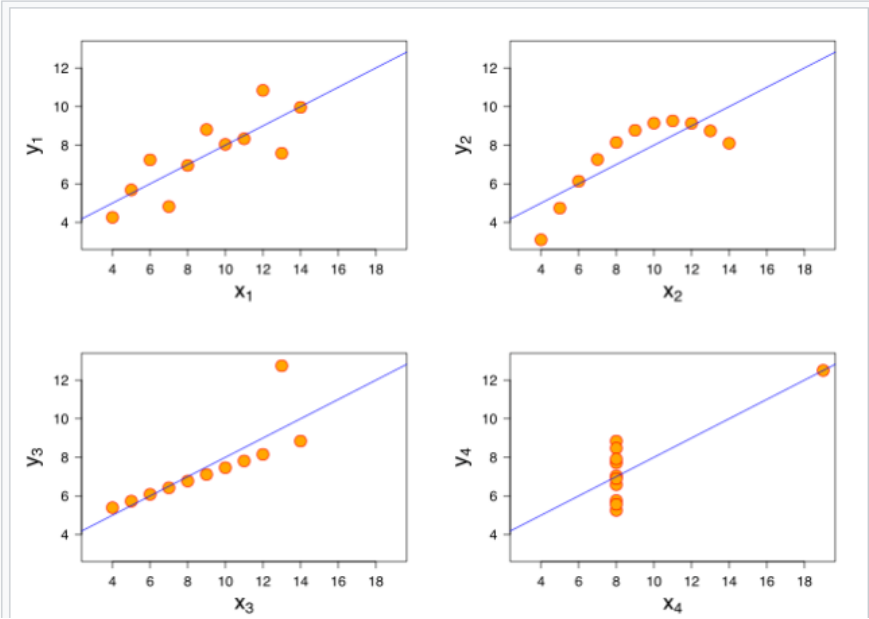
General equation of a straight line, which is fitted during simple linear regression is $\mathbf{y = \beta_o + \beta_1 x}$

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

Developed by statistician F.J. Anscombe in 1973, Anscombe's Quartet is a set of four datasets, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quire similar. However, after visualizing (plotting) the data, it becomes clear that the datasets are markedly different.

The effectiveness of Anscombe's Quartet is not due to simply having four different datasets which generate the same statistical properties, it is that four **clearly different** and **visually distinct** datasets are producing the same statistical properties.

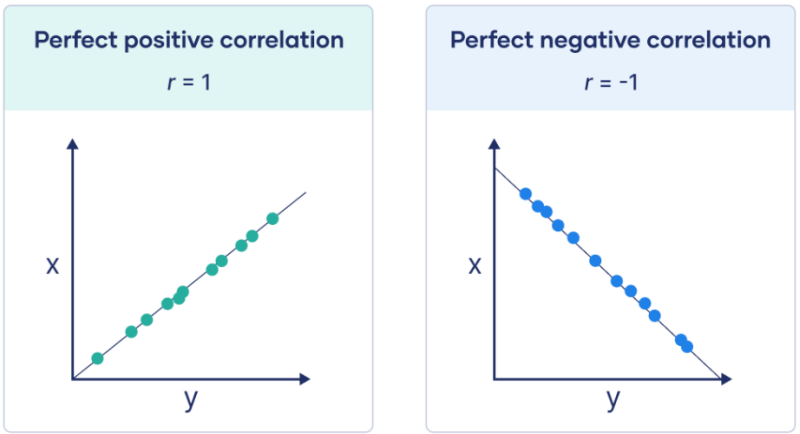
They were constructed by Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.



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

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- Correlation between 0 and 1 indicates positive correlation: When one variable changes, the other variable changes in the same direction. E.g. Baby length & weight (longer the baby, the heavier their weight).
- 0 correlation indicates no linear correlation: There is no linear relationship between the variables. E.g. Car price & width of windshield wipers (price of a car is not related to the width of its windshield wipers).
- Between 0 and –1 indicates negative correlation: When one variable changes, the other variable changes in the opposite direction. E.g. Elevation & air pressure (higher the elevation, the lower the air pressure).



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

A: Scaling 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.

It is of 2 types:

- Standardisation brings all the data into a standard normal distribution with mean 0 and standard deviation 1.
- MinMax scaling/ Normalization, on the other hand, brings all the data in the range of 0-1.

The formulae used in the background for each of these methods are as given below:

a) Standardisation: $x = x - \text{mean}(x) / \text{sd}(x)$

b) MinMax Scaling / Normalization: $x = x - \min(x) / \max(x) - \min(x)$

MinMax scaling /standardization are designed to achieve a similar goal, which is to create features that have similar ranges to each other. This is done to ensure we are capturing the true information in a feature, and that we don't over weigh a particular feature just because its values are much larger than other features.

If all of the features are within a similar range of each other, then there's no real need to standardize/normalize. If, however, some features take on values that are much larger/smaller than others then normalization/standardization is called for.

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

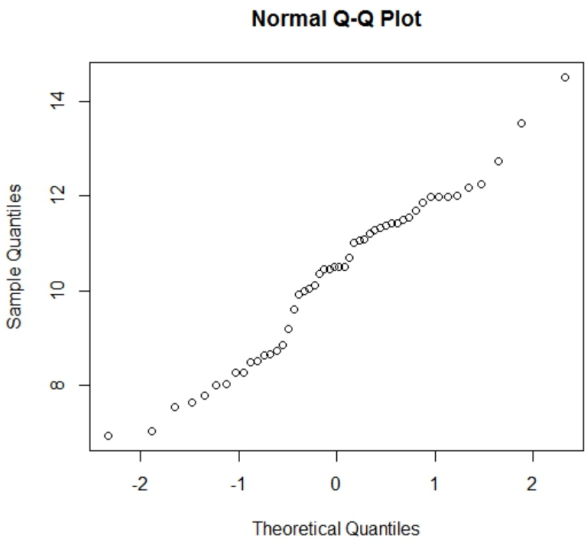
If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

An infinite VIF indicates that there is perfect collinearity and the concerned variables are completely redundant. For instance, including an indicator of maleness is completely redundant with an indicator of femaleness.

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

The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption.

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
8
10
12
14
-2
-1
0
1
2