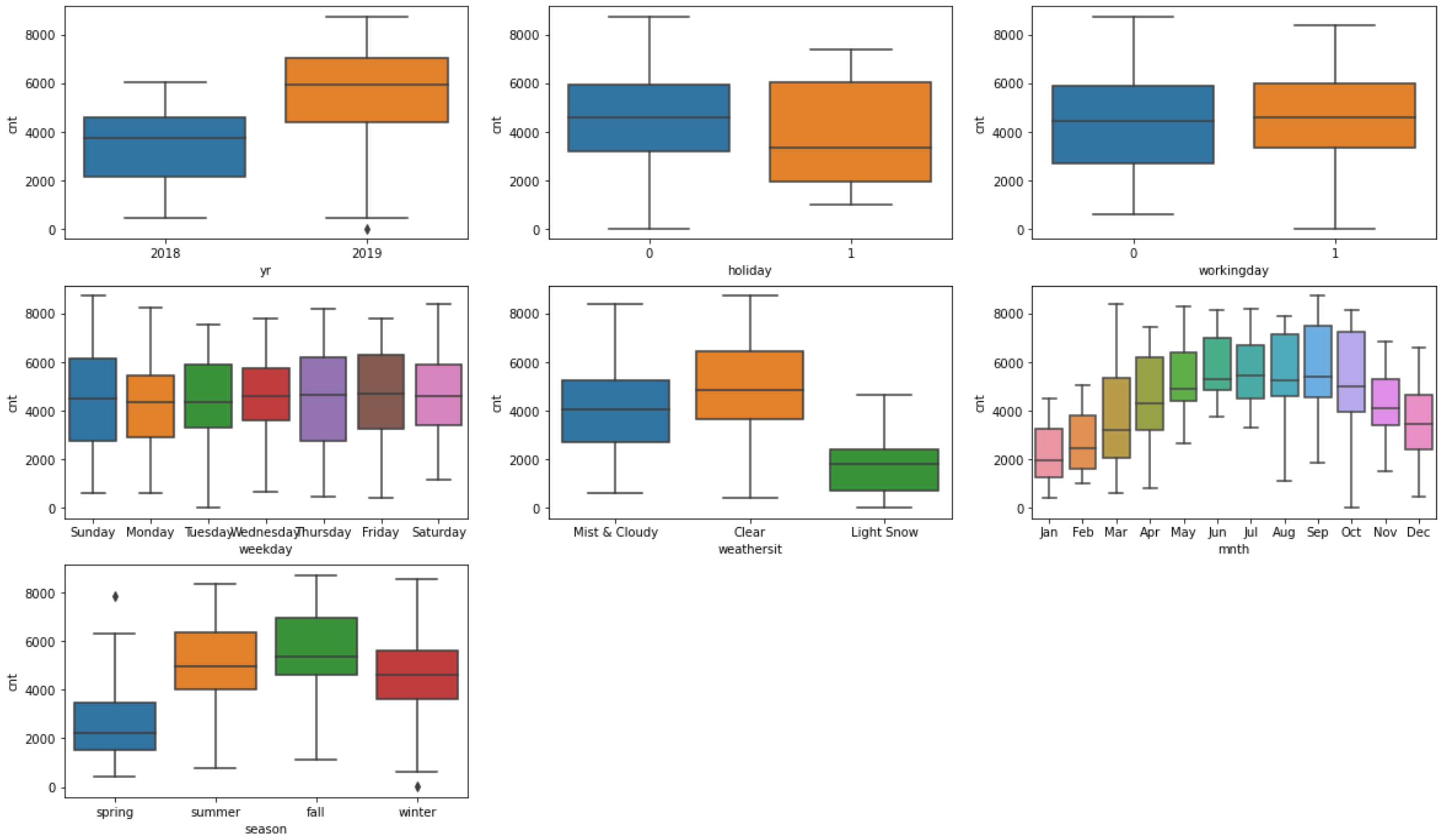
Assignment-based Subjective Questions

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: Effect of categorical variables on target variable Count of Rental bikes (cnt)

   1. **Year**: ‘yr’ Column has 2 binary values (2018 and 2019), below chart clearly represents that the target variable has increased in the year 2019 and the median is also high.
   2. **Holiday**: ‘holiday Column has 2 binary values (0 and 1), below chart represents that there is no significant difference on the target variable on the holiday or non holiday but the median of non-holiday is slightly higher.
   3. **Working day**: ‘workingday Column has 2 binary values (0 and 1), below chart represents that there is no significant difference on the target variable on the workingday or non workingday but the median of workingday is slightly higher.
   4. **Weekday**: ‘weekday’ Column has 7 categorical values (Sunday to Monday), below chart represents that there is no significant difference on the target variable.
   5. **Weather:** ‘weatherit Column has 3 categorical values, below chart represents that there is significant difference on the target variable. Clear weather has higher median of target variable.
   6. **Month**: ‘mnth’ Column has 12 categorical values (Jan to Dec), below chart represents that there is an increase in the count of rental bikes from May to Oct months.
   7. **Season**: ‘season’ Column has 4 categorical values, below chart represents that there is an increase in the count of rental bikes on summer and fall season. Fall has the highest count of rental bikes.



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

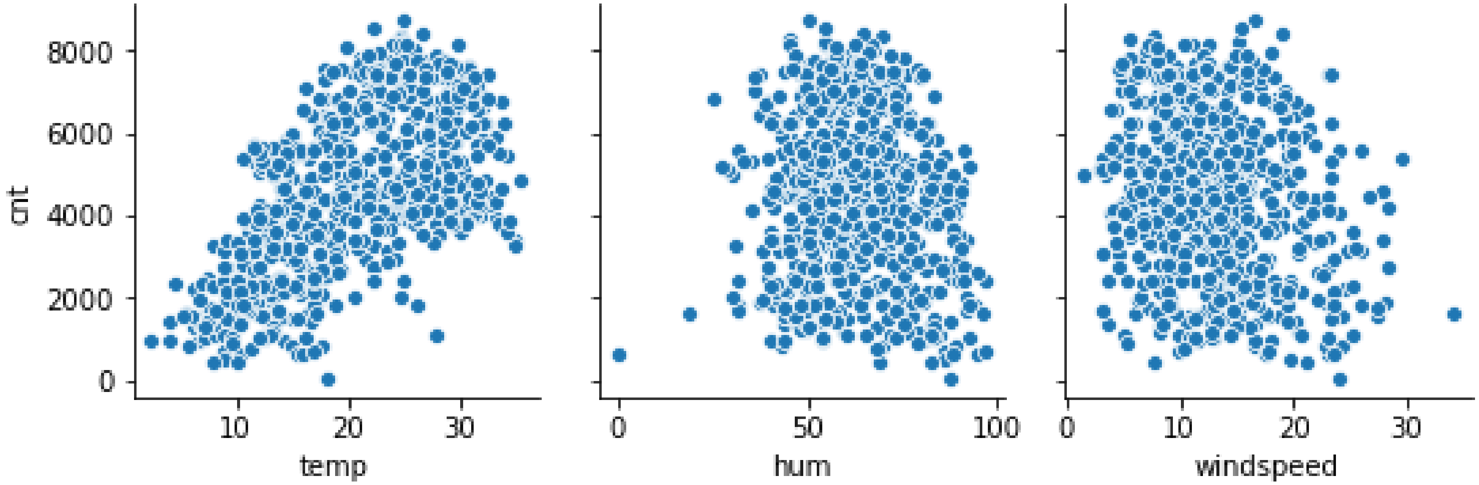
Ans:

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.

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

Ans: Pairplot was generated for 3 numerical variables, temp', 'hum', and 'windspeed' with the target variable 'cnt'.

Out of these 3, 'temp' feature is highly correlated with the target 'cnt' as can be seen clearly in below chart



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

Ans:

Assumptions were validated based on the below factors

P value: The P value of the variables should not be greater than 0.05. P value closer to zero shows that the variable is significant and the coefficient which has been calculated is not be default.

VIF value: The VIF values of the variables should not be greater than 5. Anything below 5 is ok. The variables with VIF values higher than 5 were rejected

Combination of P value and VIF:

We could have :

- High P value and High VIF (Remove straightaway)
- -High-low
  - High P , low VIF (Remove these first)
  - Low P, high VIF (Remove after the above once are removed)

- Low P Low VIF (Keep these variables)

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: the top 3 features contributing significantly towards explaining the demand of the shared bikes

1. Spring
2. Light snow
3. 2019

Final Equation is:¶

cnt = - 0.1723 X windspeed - 0.2988 X spring - 0.0405 X summer - 0.0766 X winter¶

+ 0.2476 X 2019 + 0.0705 X sep - 0.0455 X Monday - 0.2964 X Light Snow - 0.0892 X Mist & Cloudy + 0.585¶

General Subjective Questions

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

   Ans:

   Machine learning is all about predictions – a machine learning, thinking and predicting what's next.
   – what will a machine learn, how will a machine analyze, what will it predict.

   In any machine learning model once gas to understand two terms

   - Data
   - Algorithm

   Majority of the machine learning algorithms fall under the supervised learning category. It is the process where an algorithm is used to predict a result based on the previously entered values and the results generated from them.

Suppose we have an input variable ‘x’ and an output variable ‘y’ where y is a function of x (y=f{x}). Supervised learning reads the value of entered variable ‘x’ and the resulting variable ‘y’ so that it can use those results to later predict a highly accurate output data of ‘y’ from the entered value of ‘x’.

A regression problem is when the resulting variable contains a real or a continuous value. It tries to draw the line of best fit from the data gathered from a number of points.

1. Simple Linear Regression : The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straightline. The straight line is plotted on the scatter plot of these two points.

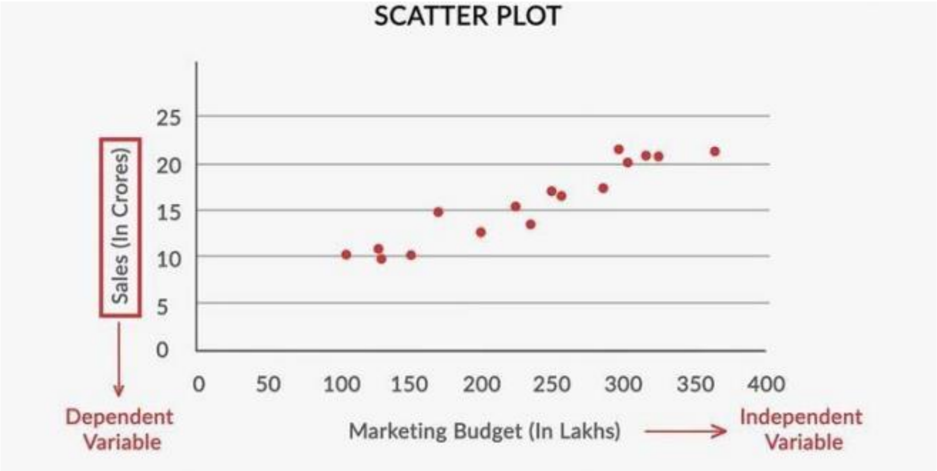


Figure 2 - Scatter plot

A simple linear regression model 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 variables are also known as the **output variables**. The equation of simple linear regression is given by:

$\hat{Y}=b_0+b_1X_1$where Y cap is the Predicted output variable, b zero is the intercept and b1 is the slope or coefficient

Multiple Linear Regression : Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

The equation of simple linear regression is given by:

$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$where Y cap is the Predicted output variable, b zero is the intercept and b1 and b2 are the slopes or coefficient of different independent variables.

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

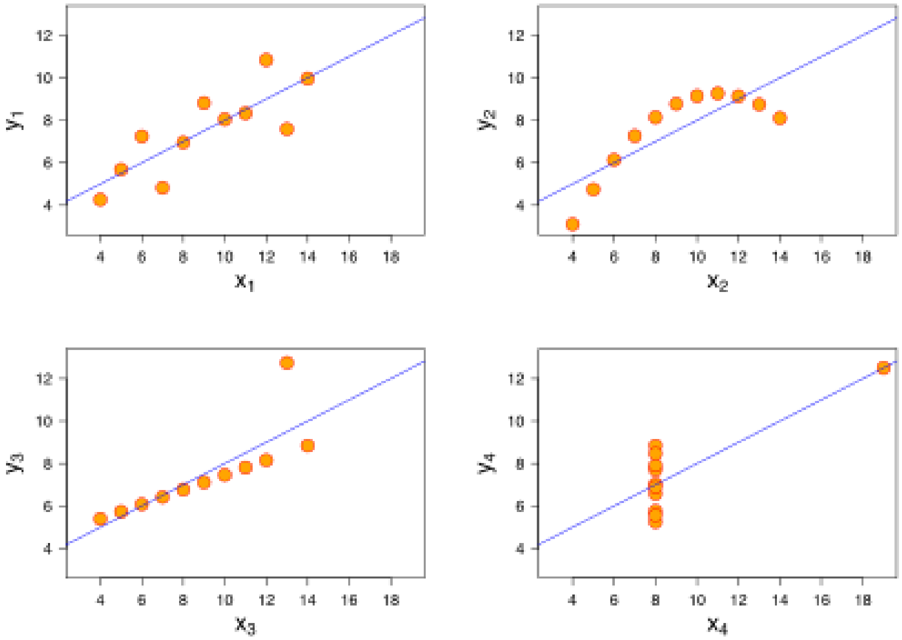
   Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

   Data for all 4 data sets:

Property	**Value**	**Accuracy**
Mean of *x*	9	exact
Sample variance of *x* :	11	exact
Mean of *y*	7.50	to 2 decimal places

Sample variance of *y* :	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



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

Ans: Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity.

The relationship between two variables is generally considered strong when their r value is larger than 0.7. The correlation r measures the strength of the linear relationship between two quantitative variables. Pearson r: ... The extreme values r = -1 and r = 1 occur only in the case of a perfect linear relationship.

Pearson's r is usually used to express the correlation between two quantities. ... You could calculate Pearson's r to evaluate whether the two quantities are correlated. R^2 is usually used to evaluate the quality of fit of a model on data

A Pearson's correlation is used when you want to find a linear relationship between two variables. It can be used in a causal as well as a associative research hypothesis but it can't be used with a attributive RH because it is univariate.

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

Why Scaling is performed:

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.

Difference between normalized scaling and standardized scaling:

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

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

Ans:
In general one starts with the selection of all variables, and proceeds by repeatedly deselecting variables showing a high VIF. ...
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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.

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

(3 marks)

Ans:

In statistics, a Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen.

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.

Importance of a Q-Q plot in linear regression:

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